



Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

JC : Oct 21,2022

Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing

ACM Transactions on Computing for Healthcare, Vol. 3, No. 1, Article 2. Publication date: October 2021.

YU GU, ROBERT TINN, HAO CHENG, MICHAEL LUCAS, NAOTO USUYAMA, XIAODONG LIU, TRISTAN NAUMANN,
JIANFENG GAO, and HOIFUNG POON,
Microsoft Research

Presenter

Ekapob Sangariyavanich MD.

Ph.D. student year 2020 in Data Science for Healthcare and Clinical Informatics program



Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

Topic

- Introduction
- BERT
- Methods
- Results
- Discussion
- Conclusion



Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

Introduction



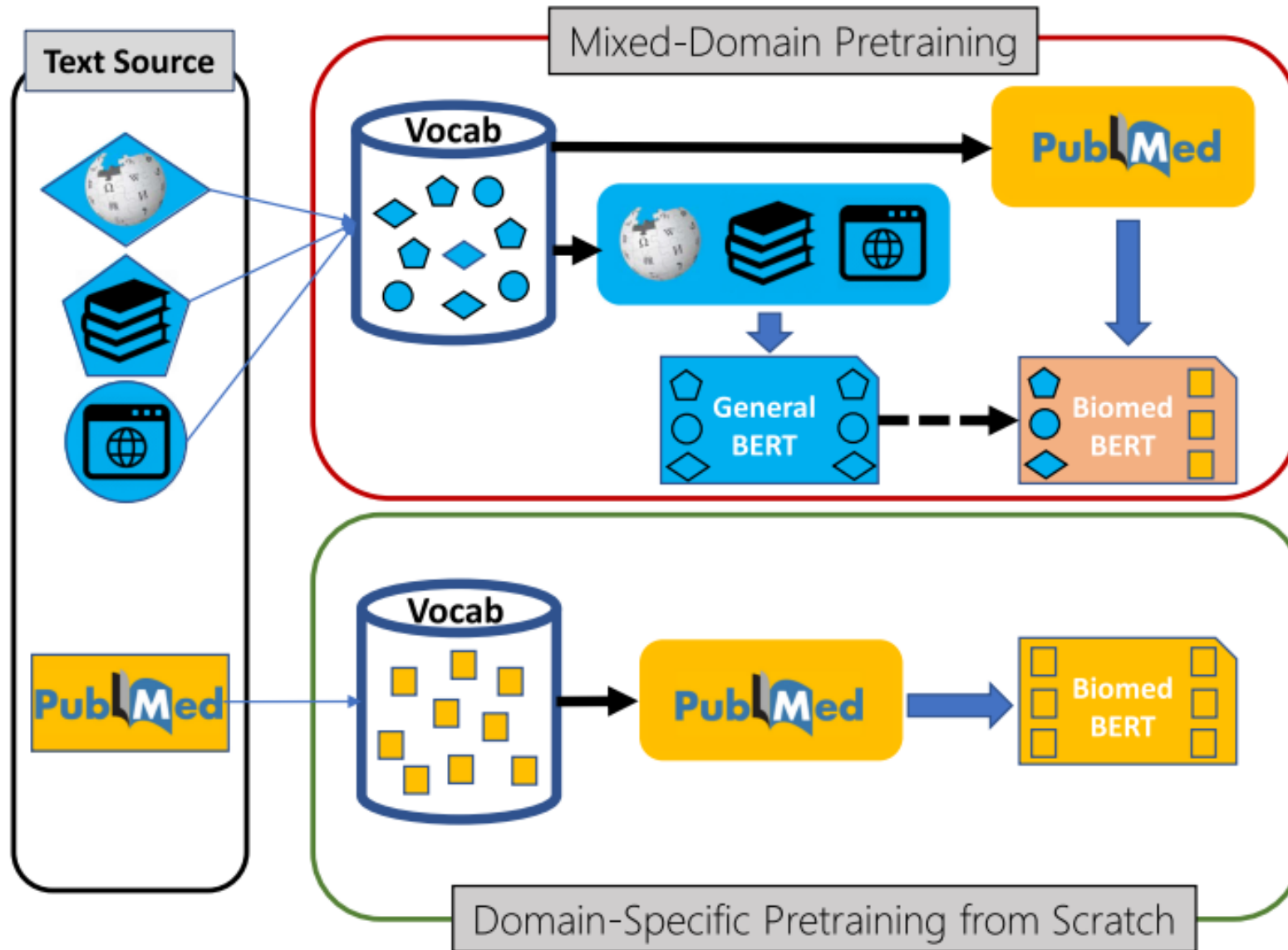
- Bidirectional Encoder Representations from Transformers (BERT) has become a standard building block for training task-specific NLP models.
- The original BERT model was trained on Wikipedia and BookCorpus.
- In specialized domains like biomedicine, past work has shown that using **in-domain** text can provide additional gains over **general-domain** language models.



Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics





Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

- A prevailing assumption is that **out-domain** text is still helpful and previous work typically adopts a mixed-domain approach, such as by starting domain-specific pretraining from an existing general-domain language model.

"We question this assumption"

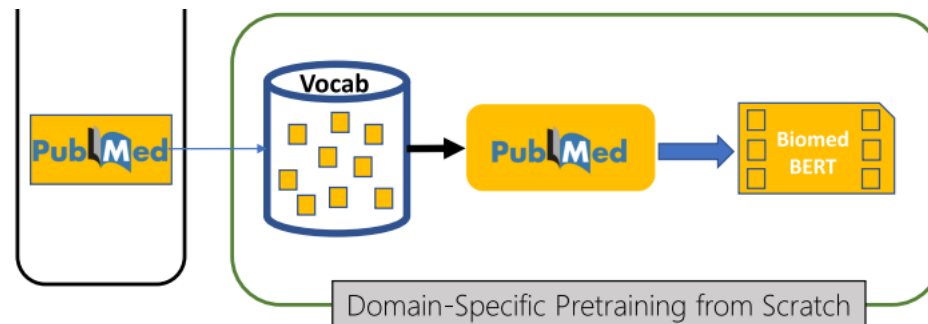


- Successful transfer learning occurs when the target data is limit and the source domain is highly relevant to the target one.
- In fact, the majority of general domain text is substantively different from biomedical text, raising the prospect of negative transfer that actually hinders the target performance.
- Current biomedical NLP benchmark has limit coverage of biomedical applications.



• Contributions

- 1) Domain-specific pretraining from scratch substantially outperforms continual pretraining of generic language models (PubMedBERT)



- 2) A comprehensive biomedical NLP benchmark from publicly-available datasets



Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

Overview of BERT



Text representation

- Conversion raw text to a numerical form (numeric vector)

One hot encoding

The cat sat on the mat

The: [0 1 0 0 0 0]

cat: [0 0 1 0 0 0]

sat: [0 0 0 1 0 0]

on: [0 0 0 0 1 0]

the: [0 0 0 0 0 1]

mat: [0 0 0 0 0 1]

Count vectorizer

```
corpus = ['The sky is blue and beautiful',  
          'The king is old and the queen is beautiful',  
          'Love this beautiful blue sky',  
          'The beautiful queen and the old king']
```

	and	beautiful	blue	is	king	love	old	queen	sky	the	this
0	1	1	1	1	0	0	0	0	1	1	0
1	1	1	0	2	1	0	1	1	0	2	0
2	0	1	1	0	0	1	0	0	1	0	1
3	1	1	0	0	1	0	1	1	0	2	0



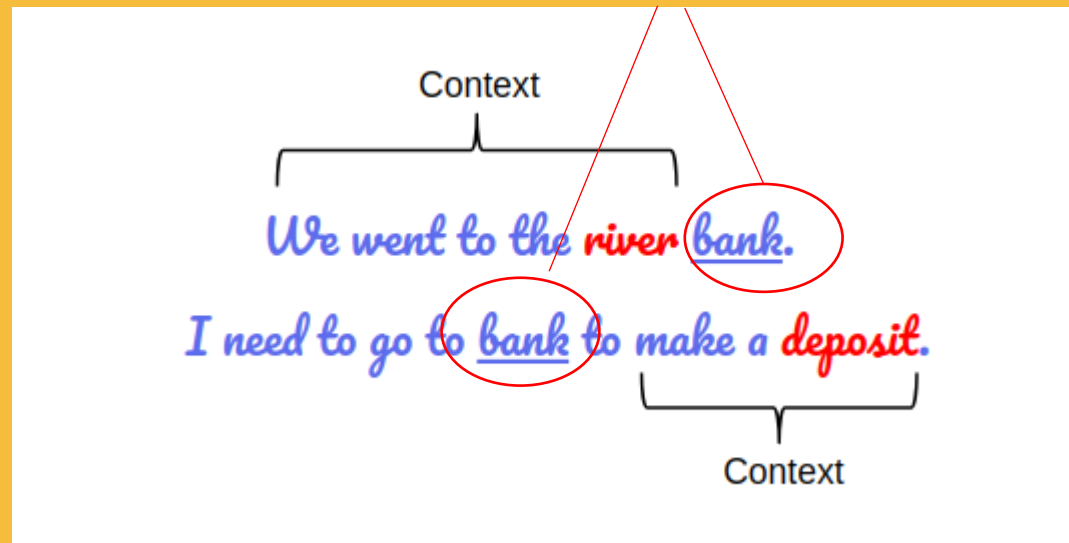
Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

- 2013 : Pre-trained word embeddings -> **Word2Vec**
- Map a single word as a single vector
- Out of vocabulary problem
- Do not take into account the word position
- **No contextual representation**

Word2Vec gives
the same vector.



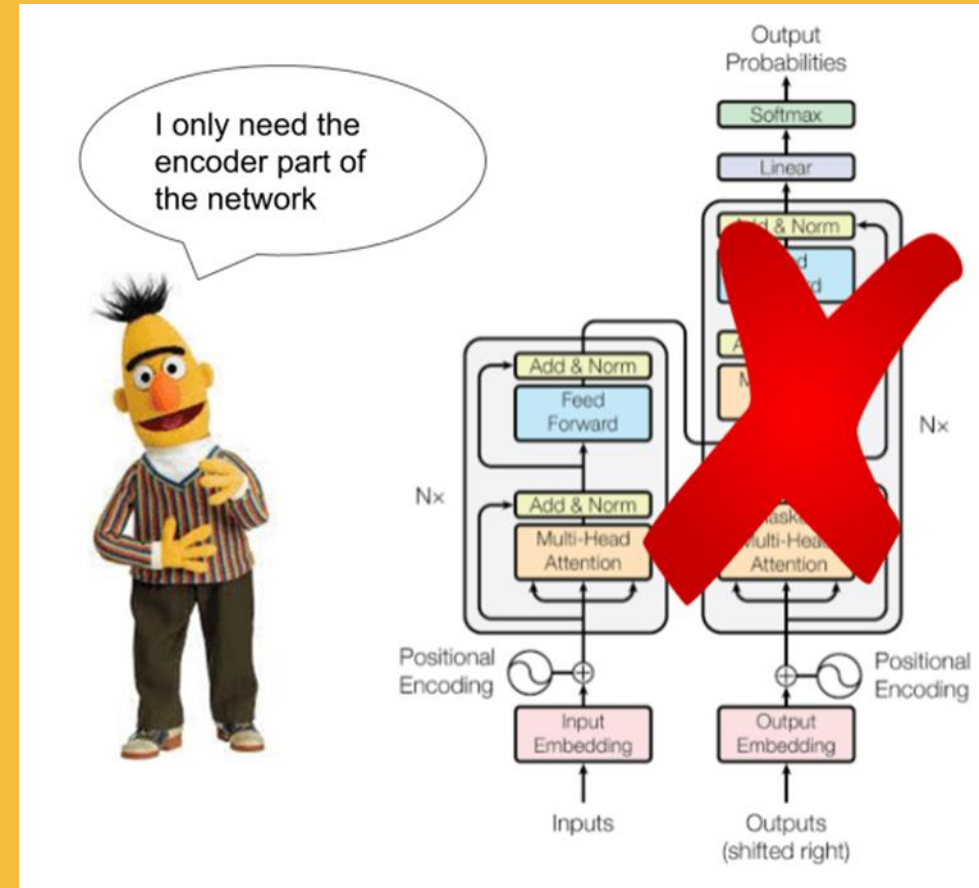
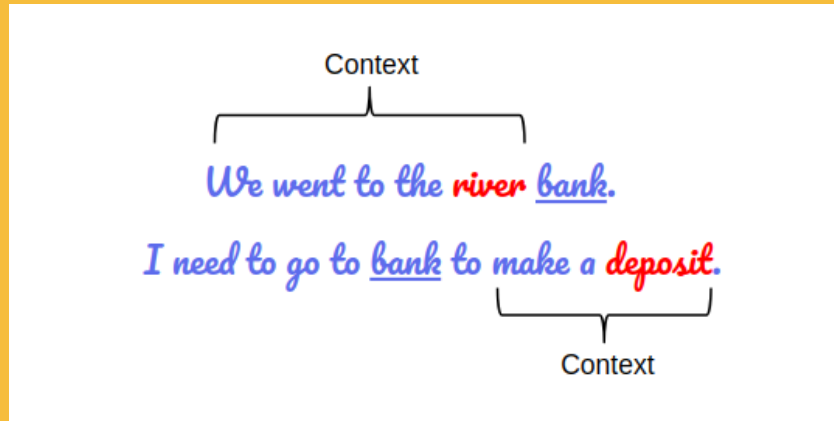


Mahidol University

Faculty of Medicine Ramathibodi Hospital

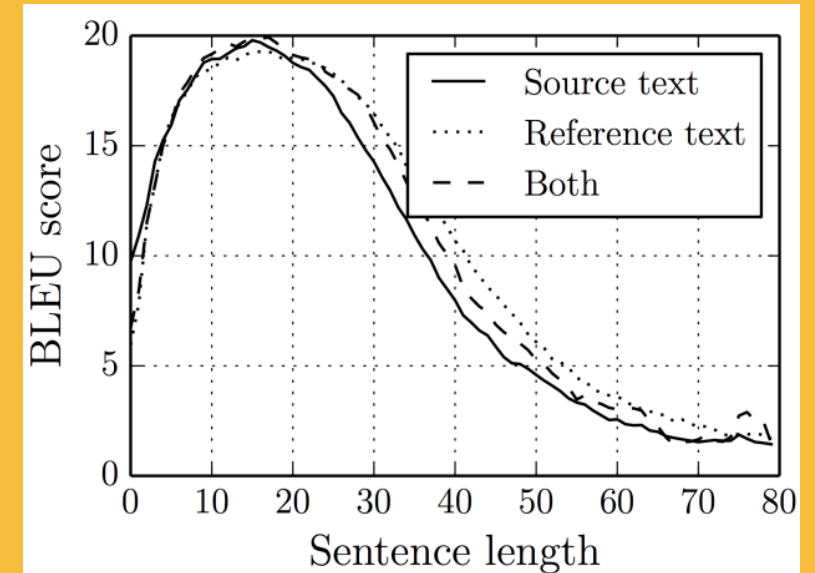
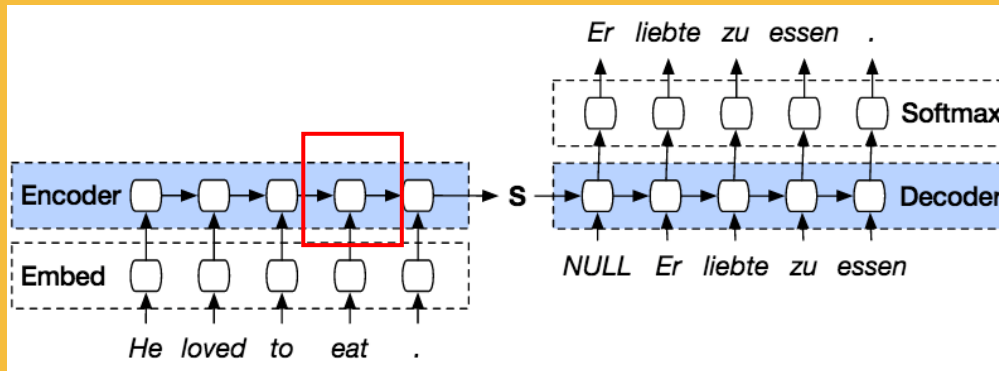
Department of Clinical Epidemiology and Biostatistics

- BERT
- Transformer base –self attention
- No CNN , RNN
- Contextual representation





- Why the transformer ?
 - sequence to sequence model problem :



- Fixed-length of context vector
- Model could only prioritize the important of words that were most recently processed.
 - loss information in the long sequences (long range dependencies)

“The cat ran away when the dog chased it down the street”.



- Attention mechanism :
selectively concentrating on a few relevant things.

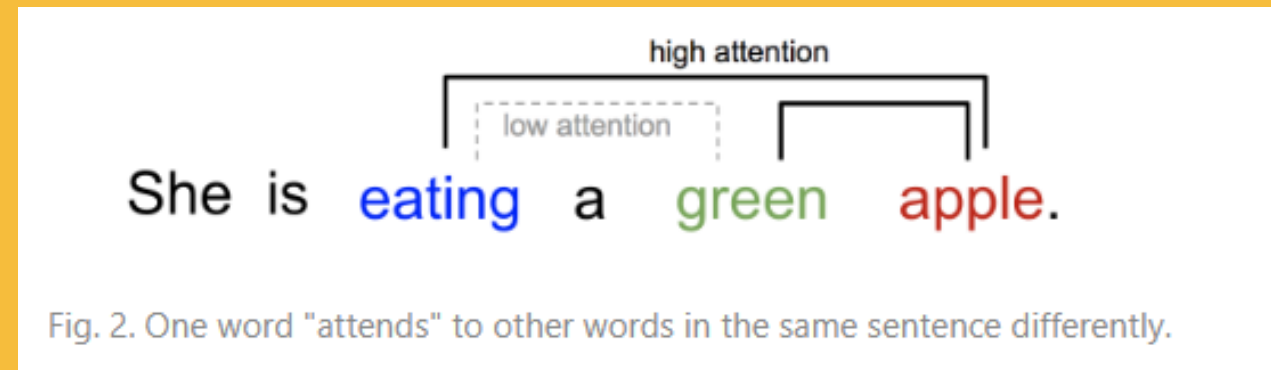
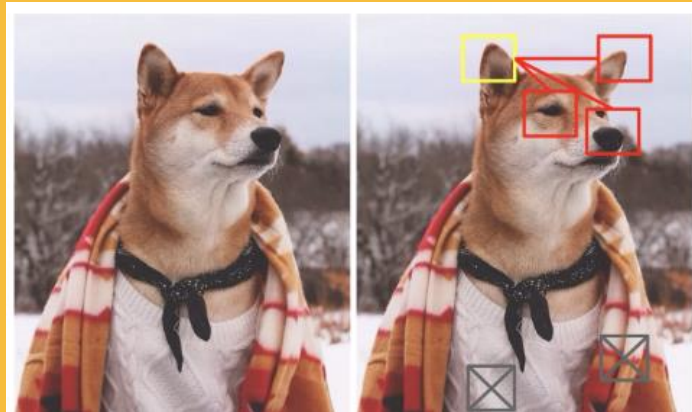


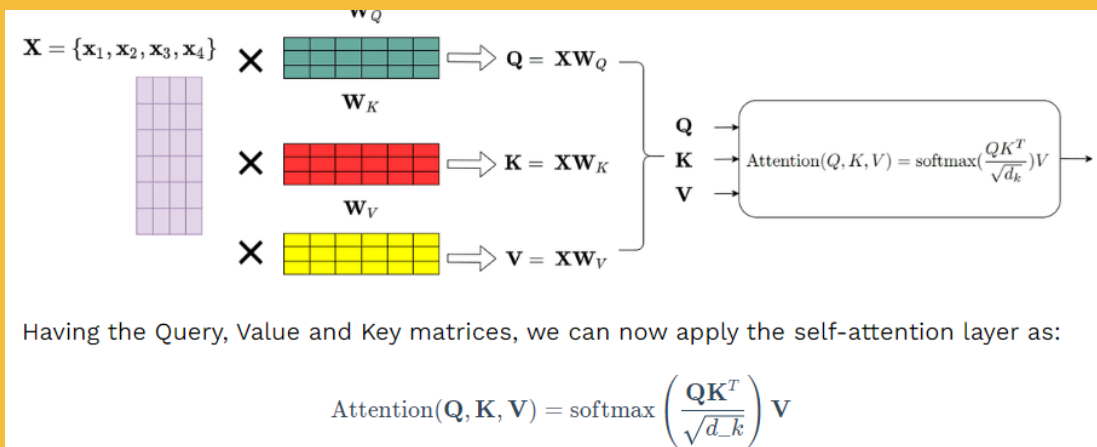
Fig. 2. One word "attends" to other words in the same sentence differently.

Attention allows the model to focus on the relevant parts of the input sequence as needed.



• Self attention

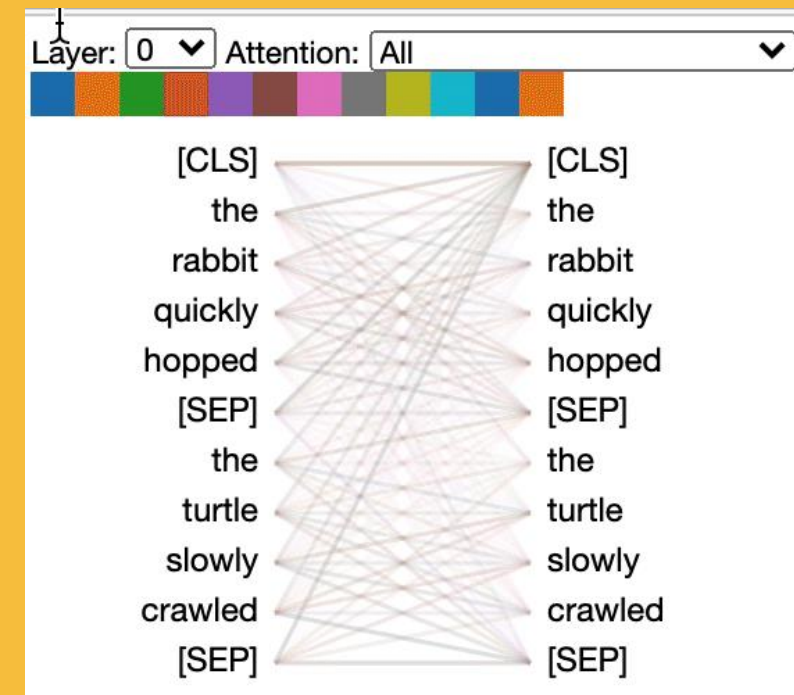
- attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence.
- Create Query , Key , Value vectors for attention score calculation



Self-attention
Probability score matrix

	Hello	I	love	you
Hello	0.8	0.1	0.05	0.05
I	0.1	0.6	0.2	0.1
love	0.05	0.2	0.65	0.1
you	0.2	0.1	0.1	0.6

Softmax(Attention)
equation





- if I can give this restaurant a 0 I will we be just ask our waitress leave because someone with a reservation be wait for our table my father and father-in-law be still finish up their coffee and we have not yet finish our dessert I have never be so humiliated do not go to this restaurant their food be mediocre at best if you want excellent Italian in a small intimate restaurant go to dish on the South Side I will not be go back
- this place suck the food be gross and taste like grease I will never go here again ever sure the entrance look cool and the waiter can be very nice but the food simply be gross taste like cheap 99cent food do not go here the food shot out of me quick then it go in
- everything be pre cook and dry its crazy most Filipino people be used to very cheap ingredient and they do not know quality the food be disgusting I have eat at least 20 different Filipino family home this not even mediocre
- seriously f*** this place disgust food and shitty service ambience be great if you like dine in a hot cellar engulf in stagnate air truly it be over rate over price and they just under deliver forget try order a drink here it will take forever get and when it finally do arrive you will be ready pass out from heat exhaustion and lack of oxygen how be that a head change you do not even have pay for it I will not disgust you with the detailed review of everything I have try here but make it simple it all suck and after you get the bill you will be walk out with a sore ass save your money and spare your self the disappointment
- i be so angry about my horrible experience at Medusa today my previous visit be amaze 5/5 however my go to out of town and I land an appointment with Stephanie I go in with a picture of roughly what I want and come out look absolutely nothing like it my hair be a horrible ashy blonde not anywhere close to the platinum blonde I request she will not do any of the pop of colour I want and even after specifically tell her I do not like blunt cut my hair have lot of straight edge she do not listen to a single thing I want and when I tell her I be unhappy with the colour she basically tell me I be wrong and I have do it this way no no I do not if I can go from Little Mermaid red to golden blonde in 1 sitting that leave my hair fine I shall be able go from golden blonde to a shade of platinum blonde in 1 sitting thanks for ruin my New Year's with 1 the bad hair job I have ever have

(a) 1 star reviews

- I really enjoy Ashley and Ami salon she do a great job be friendly and professional I usually get my hair do when I go to MI because of the quality of the highlight and the price the price be very affordable the highlight fantastic thank Ashley i highly recommend you and ill be back
- love this place it really be my favorite restaurant in Charlotte they use charcoal for their grill and you can taste it steak with chimichurri be always perfect Fried yucca cilantro rice pork sandwich and the good tres lech I have had.The desert be all incredible if you do not like it you be a mutant if you will like diabeetus try the Inca Cola
- this place be so much fun I have never go at night because it seem a little too busy for my taste but that just prove how great this restaurant be they have amazing food and the staff definitely remember us every time we be in town I love when a waitress or waiter come over and ask if you want the cab or the Pinot even when there be a rush and the staff be run around like crazy whenever I grab someone they instantly smile acknowlegde us the food be also killer I love when everyone know the special and can tell you they have try them all and what they pair well with this be a first last stop whenever we be in Charlotte and I highly recommend them
- great food and good service what else can you ask for everything that I have ever try here have be great
- first off I hardly remember waiter name because its rare you have an unforgettable experience the day I go I be celebrate my birthday and let me say I leave feel extra special our waiter be the best ever Carlos and the staff as well I be with a party of 4 and we order the potato salad shrimp cocktail lobster amongst other thing and boy be the food great the lobster be the good lobster I have ever eat if you eat a dessert I will recommend the cheese cake that be also the good I have ever have it be expensive but so worth every penny I will definitely be back there go again for the second time in a week and it be even good this place be amazing

(b) 5 star reviews

Figure 2: Heatmap of Yelp reviews with the two extreme score.



- To address the problem of out-of-vocabulary words → **Byte-Pair encoding (BPE)**
(Word piece tokenizer)

"Here is the sentence I want embeddings for."

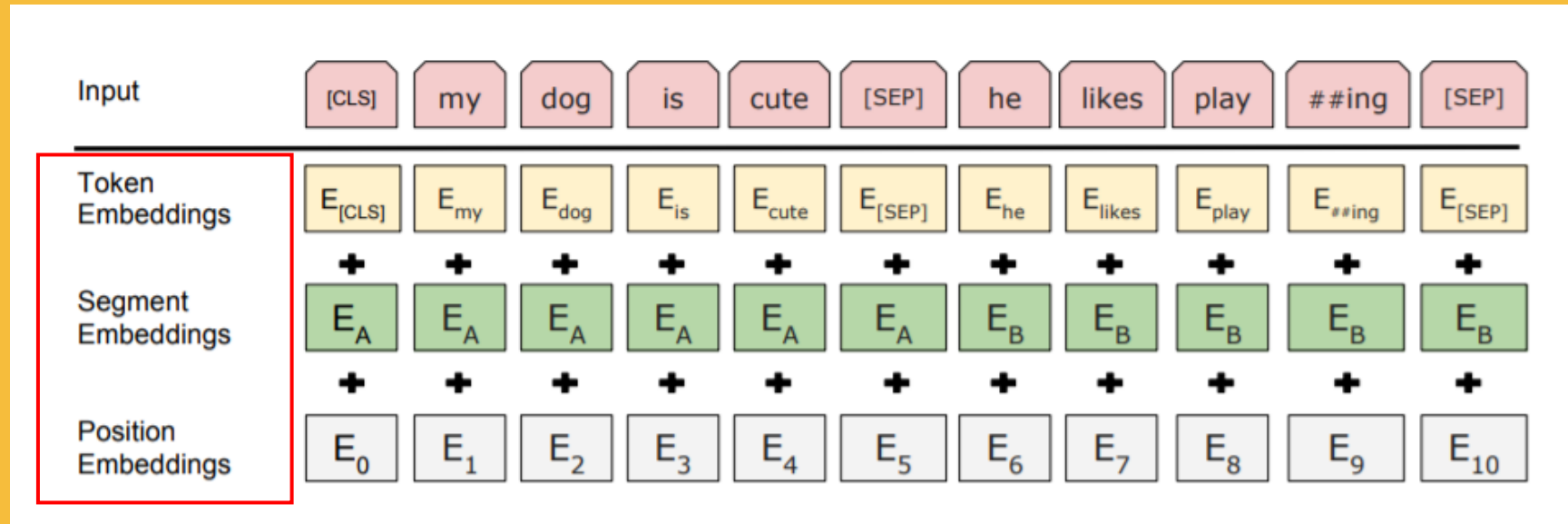


BERT Tokenizer

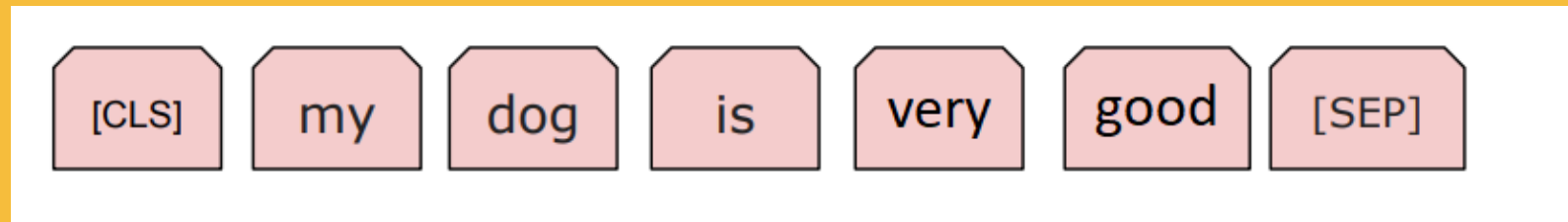
['[CLS]', 'here', 'is', 'the', 'sentence', 'i', 'want', 'em', '##bed', '##ding', '##s', 'for', '.', '[SEP]']



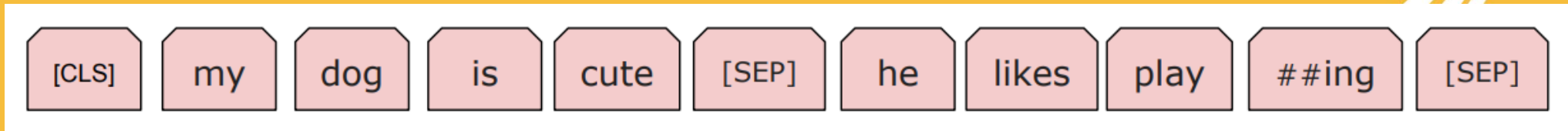
• Input embedding



Single sentence task



Sentences pair task

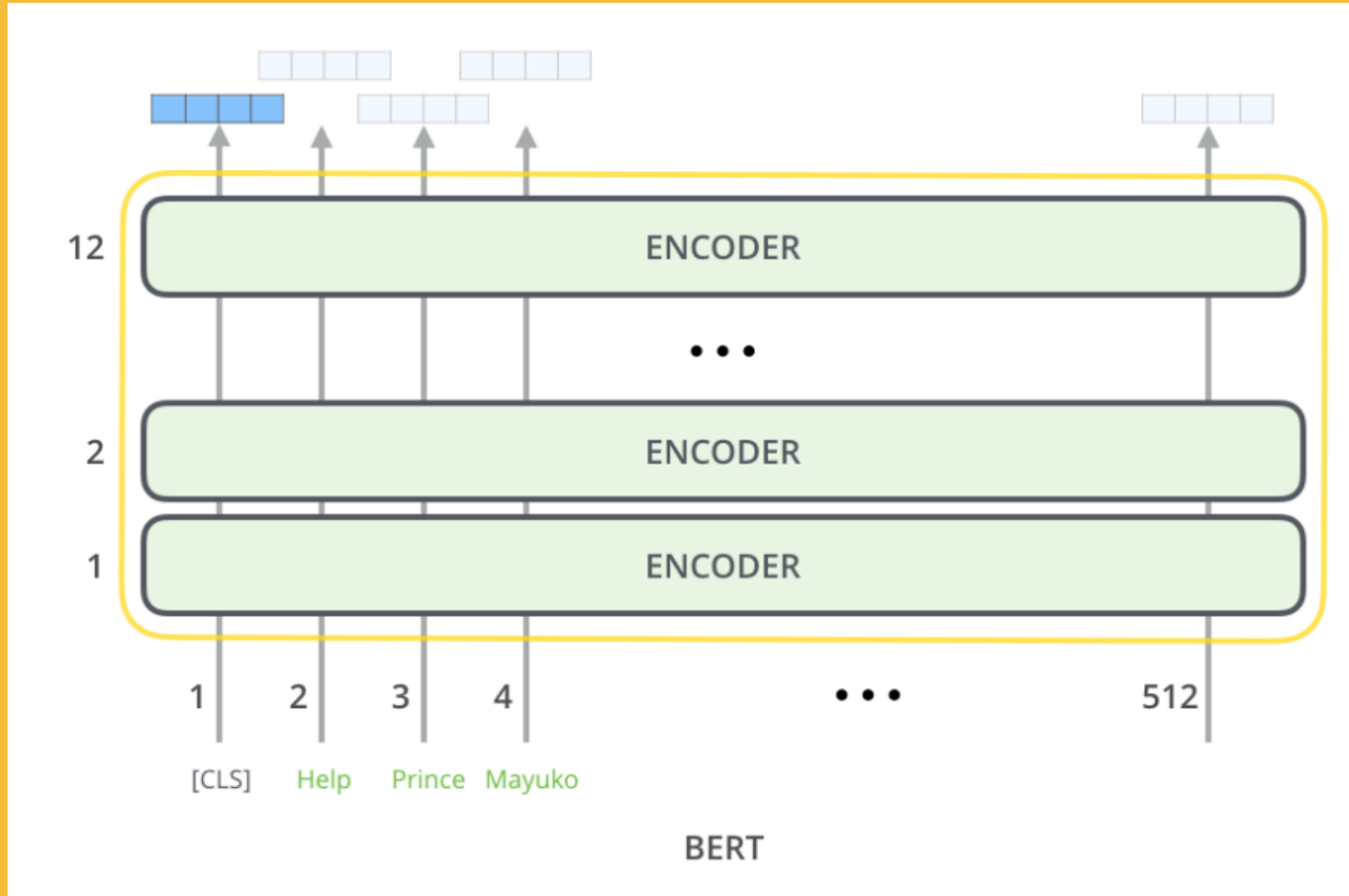




Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics



BERT Base:

12 layers (transformer blocks)

12 attention heads

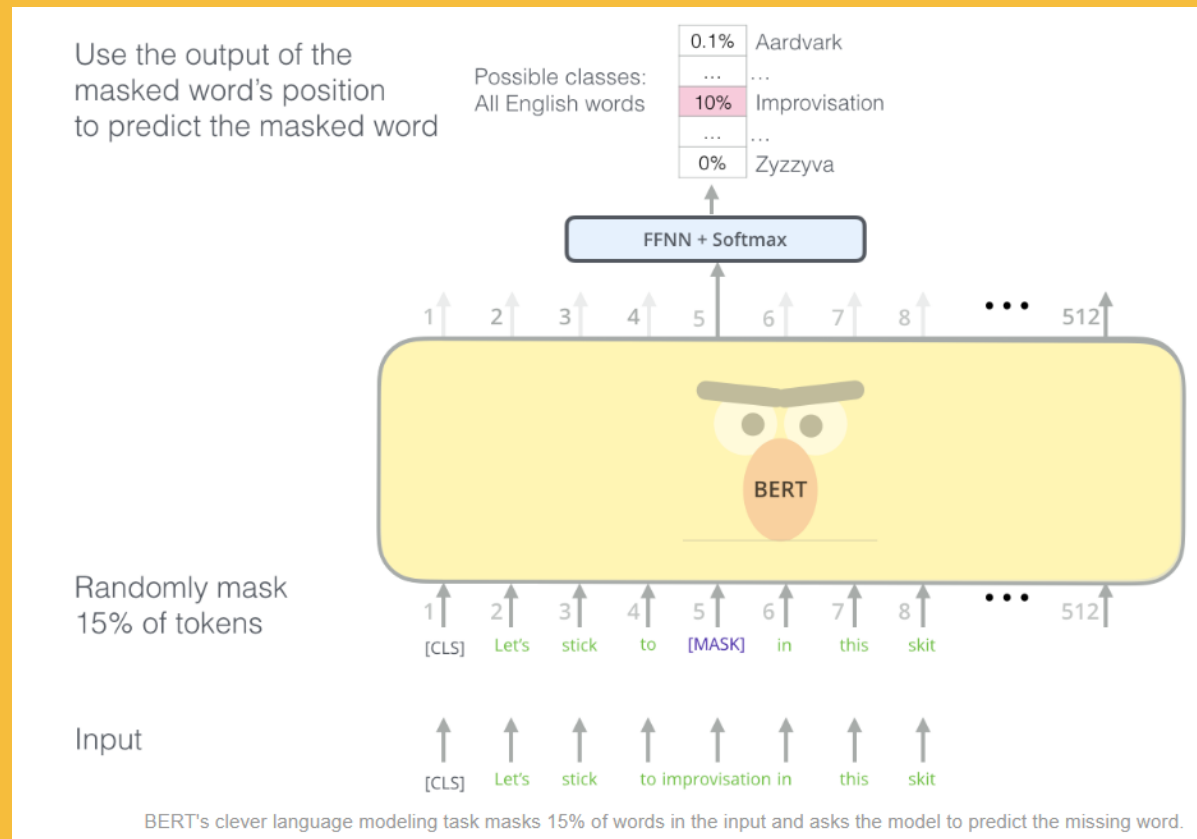
110 million parameters



- **BERT Pretraining**

1. Masked Language Modeling (MLM)

- randomly masked 15% of the words





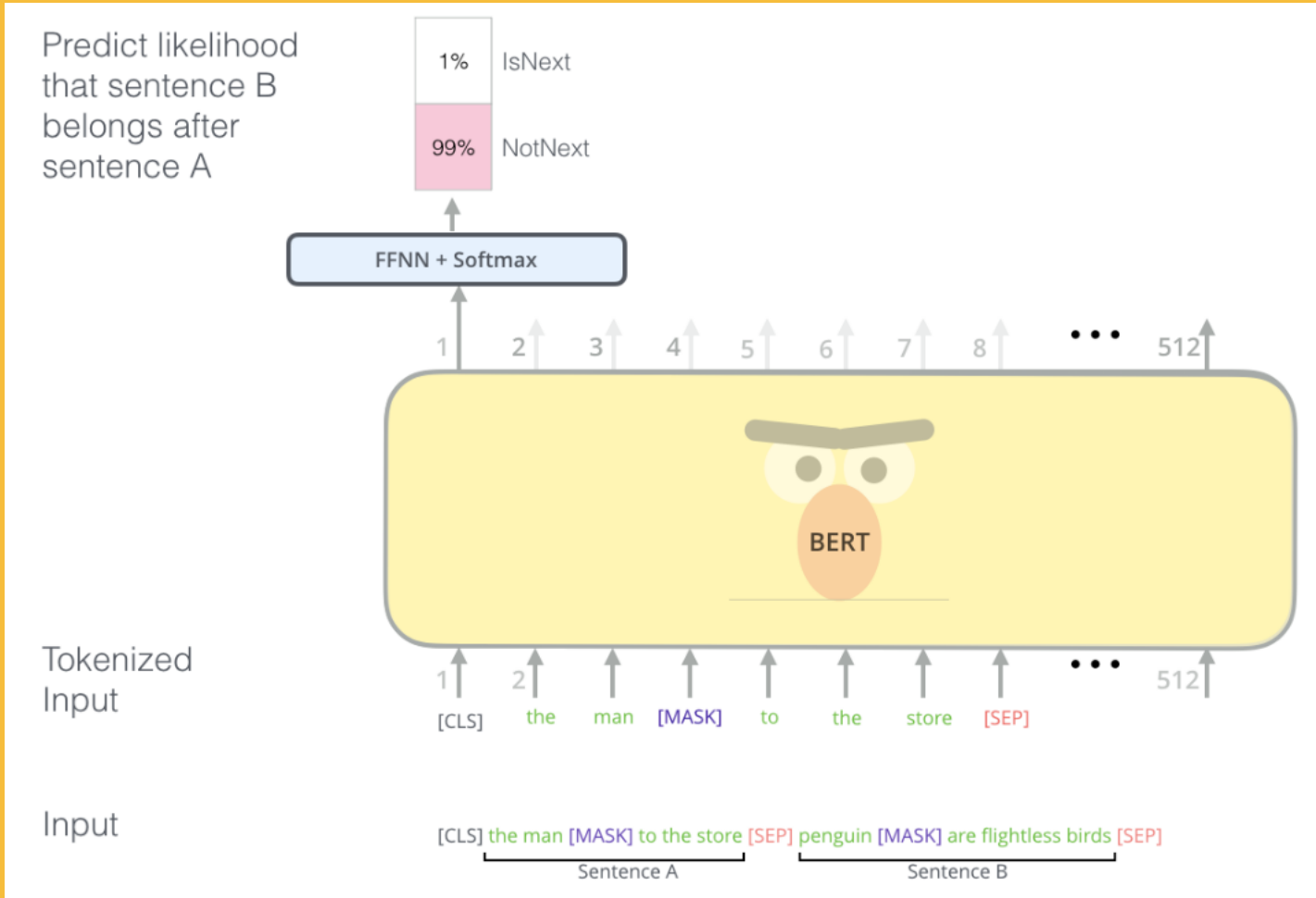
Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

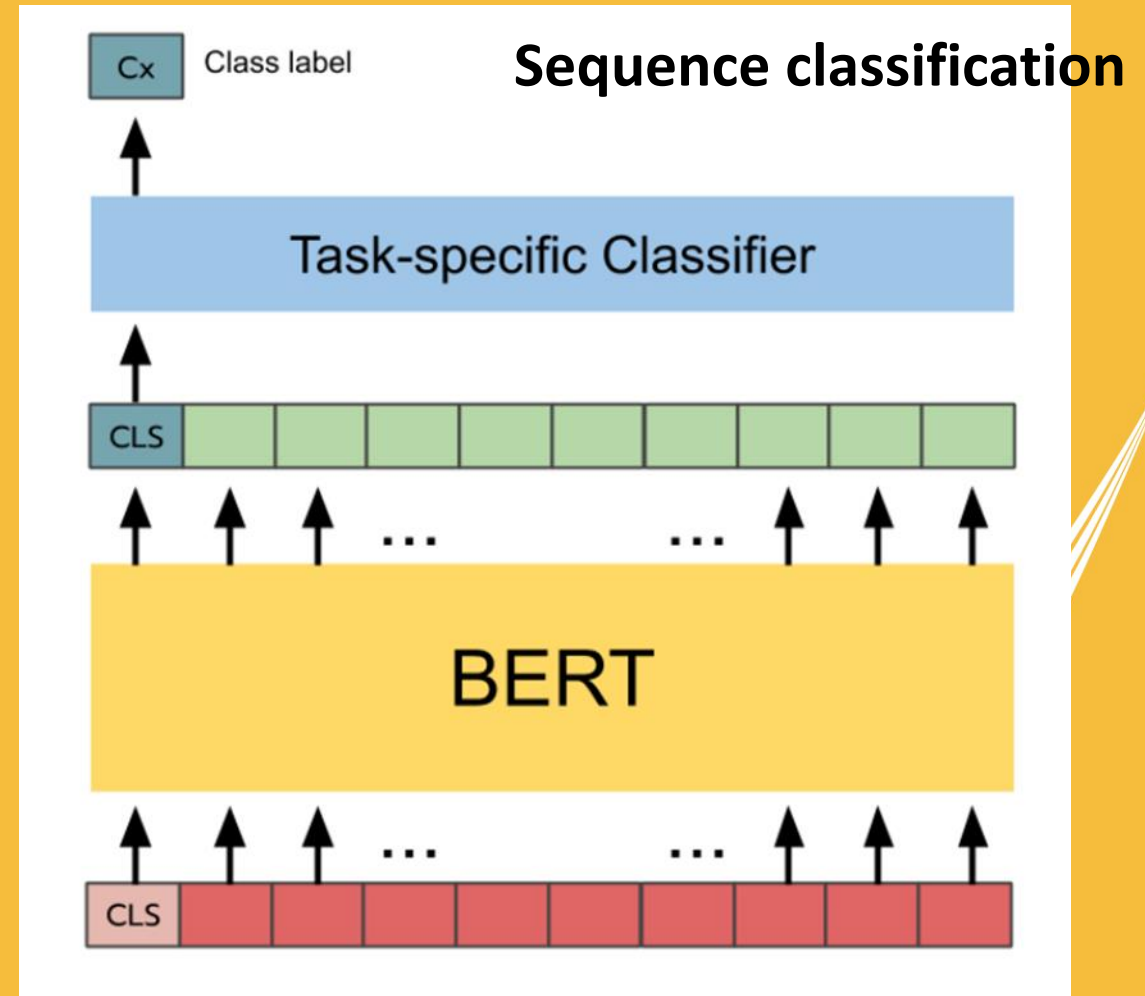
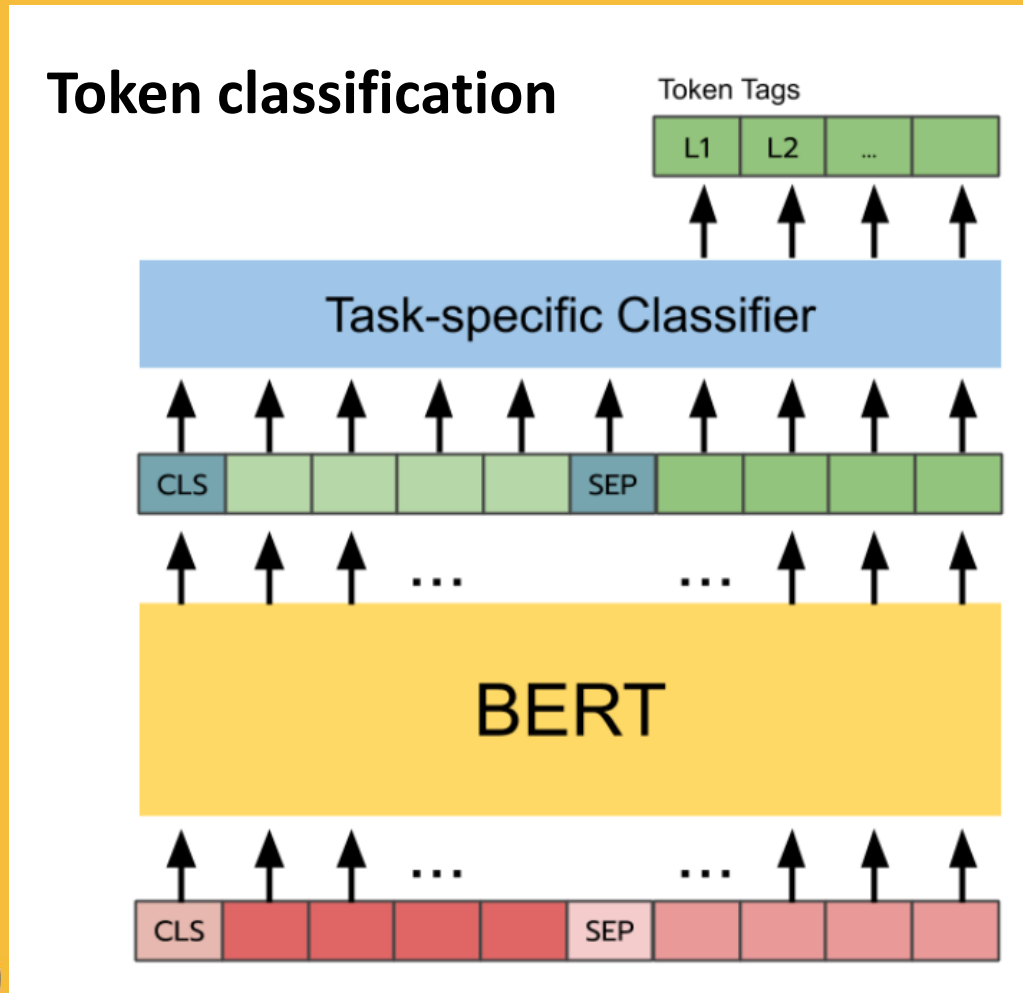
2. Next sentence prediction (NSP)

some task require the model to work with 2 sentences : QA , sentence similarity





Downstream task





Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

Methods



1. Biomedical Language Model Pretraining

1) Mixed-Domain Pretraining.

BioBERT: PubMed abstracts and PubMed Central (PMC) full-text articles

Blue BERT: both PubMed text and de-identified clinical notes from MIMIC-III

Sci-BERT: a random sample of 1.14M papers from Semantic Scholar (full text)

- Note that in the continual pretraining approach, the vocabulary is the same as the original BERT model.



2) Domain-Specific Pretraining from Scratch

PubMedBERT (abstract only / abstract + full text)

Table 1. Comparison of Common Biomedical Terms in Vocabularies Used by the Standard BERT, SciBERT, and PubMedBERT (Ours)

Biomedical Term	Category	BERT	SciBERT	PubMedBERT (Ours)
diabetes	disease	✓	✓	✓
leukemia	disease	✓	✓	✓
lithium	drug	✓	✓	✓
insulin	drug	✓	✓	✓
DNA	gene	✓	✓	✓
promoter	gene	✓	✓	✓
hypertension	disease	hyper-tension	✓	✓
nephropathy	disease	ne-ph-rop-athy	✓	✓
lymphoma	disease	l-ym-ph-oma	✓	✓
lidocaine	drug	lid-oca-ine]	✓	✓
oropharyngeal	organ	oro-pha-ryn-ge-al	or-opharyngeal	✓
cardiomyocyte	cell	card-iom-yo-cy-te	cardiomy-ocyte	✓
chloramphenicol	drug	ch-lor-amp-hen-ico-l	chlor-amp-hen-icol	✓
RecA	gene	Rec-A	Rec-A	✓
acetyltransferase	gene	ace-ty-lt-ran-sf-eras-e	acetyl-transferase	✓
clonidine	drug	cl-oni-dine	clon-idine	✓
naloxone	drug	na-lo-xon-e	nal-oxo-ne	✓

A ✓ indicates the biomedical term appears in the corresponding vocabulary; otherwise, the term will be broken into word pieces separated by a hyphen. These word pieces often have no biomedical relevance and may hinder learning in downstream tasks.



2.A Comprehensive Benchmark for Biomedical NLP

- General domain NLP benchmark
 - GLUE (The General Language Understanding Evaluation)
- Biomedical domain NLP benchmark
 - BLUE (The Biomedical Language Understanding Evaluation benchmark)
 - cover only 5 tasks : lack QA task
 - Mix biomedical and clinical application



- **BLURB (Biomedical Language Understanding & Reasoning Benchmark)**

- Focus on PubMed-based biomedical applications and leave the exploration of the clinical domain, and other high-value verticals, to future work.

Table 3. Datasets Used in the BLURB Biomedical NLP Benchmark

Dataset	Task	Train	Dev	Test	Evaluation Metrics
BC5-chem	NER	5,203	5,347	5,385	F1 entity-level
BC5-disease	NER	4,182	4,244	4,424	F1 entity-level
NCBI-disease	NER	5,134	787	960	F1 entity-level
BC2GM	NER	15,197	3,061	6,325	F1 entity-level
JNLPBA	NER	46,750	4,551	8,662	F1 entity-level
EBM PICO	PICO	339,167	85,321	16,364	Macro F1 word-level
ChemProt	Relation Extraction	18,035	11,268	15,745	Micro F1
DDI	Relation Extraction	25,296	2,496	5,716	Micro F1
GAD	Relation Extraction	4,261	535	534	Micro F1
BIOSSES	Sentence Similarity	64	16	20	Pearson
HoC	Document Classification	1,295	186	371	Micro F1
PubMedQA	Question Answering	450	50	500	Accuracy
BioASQ	Question Answering	670	75	140	Accuracy

6 NLP tasks

Note: We list the numbers of instances in train, dev, and test (e.g., entity mentions in NER and PICO elements in evidence-based medical information extraction).



3. Dataset and Tasks in BLURB

3.1 Named Entity Recognition (NER)

classify named entities mentioned in unstructured text into pre-defined categories

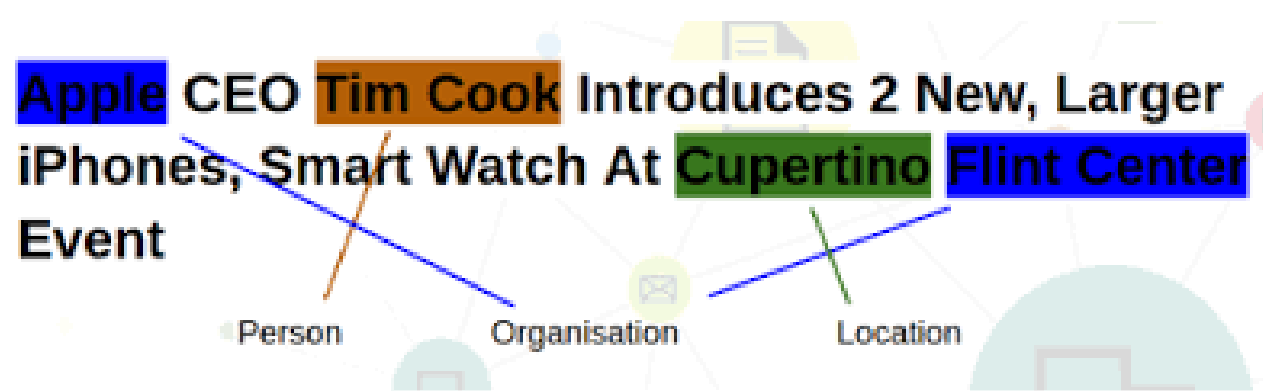


Figure 1: An example of NER application on an example text



- Token classification problem

BC5-chem : detect chemical (drugs) entity 1 = inside entity 0 = outside

tokens (sequence)	ner_tags (sequence)
["Torsade", "de", "pointes", "ventricular", "tachycardia", "during", "low", "dose", "intermittent", "dobutamine", "treatment", "in", "a", ...]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

BC5-disease: detect disease entity

["Prolonged", "left", "ventricular", "dysfunction", "occurs", "in", "patients", "with", "coronary", "artery", "disease", "after", "both", "dobutamine", "and", "exercise", "induced", "myocardial", "ischaemia", "."]	[0, 1, 2, 2, 0, 0, 0, 0, 0, 1, 2, 2, 0, 0, 0, 0, 0, 0, 1, 2, 0]
---	---



- NCBI-disease : detect disease entity
- BC2GM-disease: detect gene entity

```
[ "It", "was", "found", "that", "Nopp140", "binds", "primarily", "to",  
"the", "CK2", "regulatory", "subunit", ",", ",", "beta", "." ]
```

```
[ 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 2, 2, 2, 2, 0 ]
```

- JNLBPA : detect molecular biology entity

```
[ "Presence", "of", "beta", "2-M", "was", "analyzed", "by",  
"immunohistochemistry", "." ]
```

```
[ 0, 0, 1, 2, 0, 0, 0, 0, 0 ]
```



3.2 Evidence-Based Medical Information Extraction (PICO)

- Token classification problem
- Token will be classified as PIO element

⚡ Hosted inference API ⓘ

🔗 Token Classification Examples ▾

Those in the aspirin group experienced reduced duration of headache compared to those in the placebo arm (P<0.05)

Compute

Computation time on cpu: cached

Those in the **aspirin** **Intervention** group experienced reduced **duration** of headache **Outcome** compared to those in the **placebo** **Intervention** arm (P<0.05)

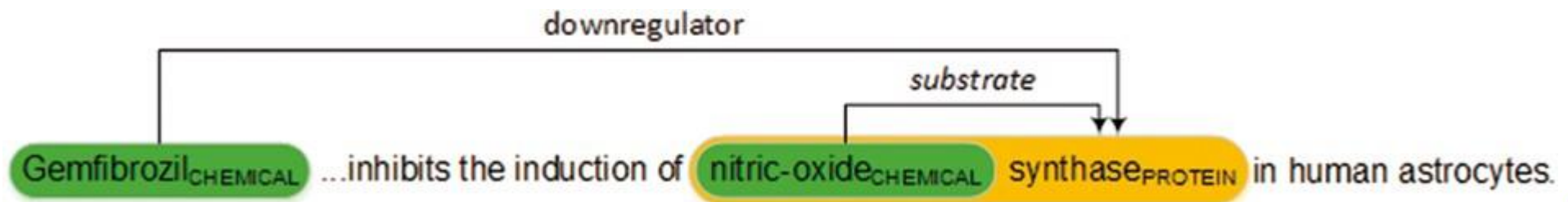
<https://huggingface.co/kamalkraj/BioELECTRA-PICO?text=Those+in+the+aspirin+group+experienced+reduced+duration+of+headache+compared+to+those+in+the+placebo+arm+%28P%3C0.05%29>



3.3 Relation Extraction (RE)

- Sequence classification
- Chempro : interactions between chemical and protein entities.

UPREGULATOR (CPR : 3), DOWNREGULATOR (CPR : 4), AGONIST (CPR : 5), ANTAGONIST (CPR : 6), SUBSTRATE (CPR : 9)—as well as everything else.





- DDI : drug-drug interaction

extract 4 types of relationship

The DDI corpus: ddis

mechanism

Lansoprazole may decrease the absorption of enoxacin.

effect

Additive CNS depression may occur when antihistamines are administered with barbiturates.

advice

Patients taking isoniazid and disulfiram concomitantly should closely monitored.

int

Clopidogrel interacts with omeprazol.

- GAD : gene associated with disease



3.4 Sentence Similarity

- Sentence regression task
- Biosses dataset : 100 pairs of PubMed sentences
- Score 0 (no relation) to 4 (equivalent meanings)

sentence1 (string)	sentence2 (string)	score (float32)
"In PC9 cells, loss of GATA6 and/or HOPX did not alter cell growth, whereas reduction of GATA2 and EGFR inhibited cell viability as..."	"Aurora-A is required for the correct localisation and function of centrosomal components like centrosomin, NDEL1, LATS and TACC..."	0.2
"It has recently been shown that Craf is essential for Kras G12D-induced NSCLC."	"It has recently become evident that Craf is essential for the onset of Kras-driven non-small cell lung cancer."	4



3.5 Document Classification

- Sequence classification problem
- HOC dataset : 10 classes of cancer hallmark → multilabel classification

Hallmark

Sustaining proliferative signaling
Evading growth suppressors
Resisting cell death
Enabling replicative immortality
Inducing angiogenesis
Activating invasion and metastasis
Genomic instability and mutation
Tumor promoting inflammation
Cellular energetics
Avoiding immune destruction



3.6 Question Answering (QA)

- Sequence classification problem
- PubMedQA : set of QA about research questions

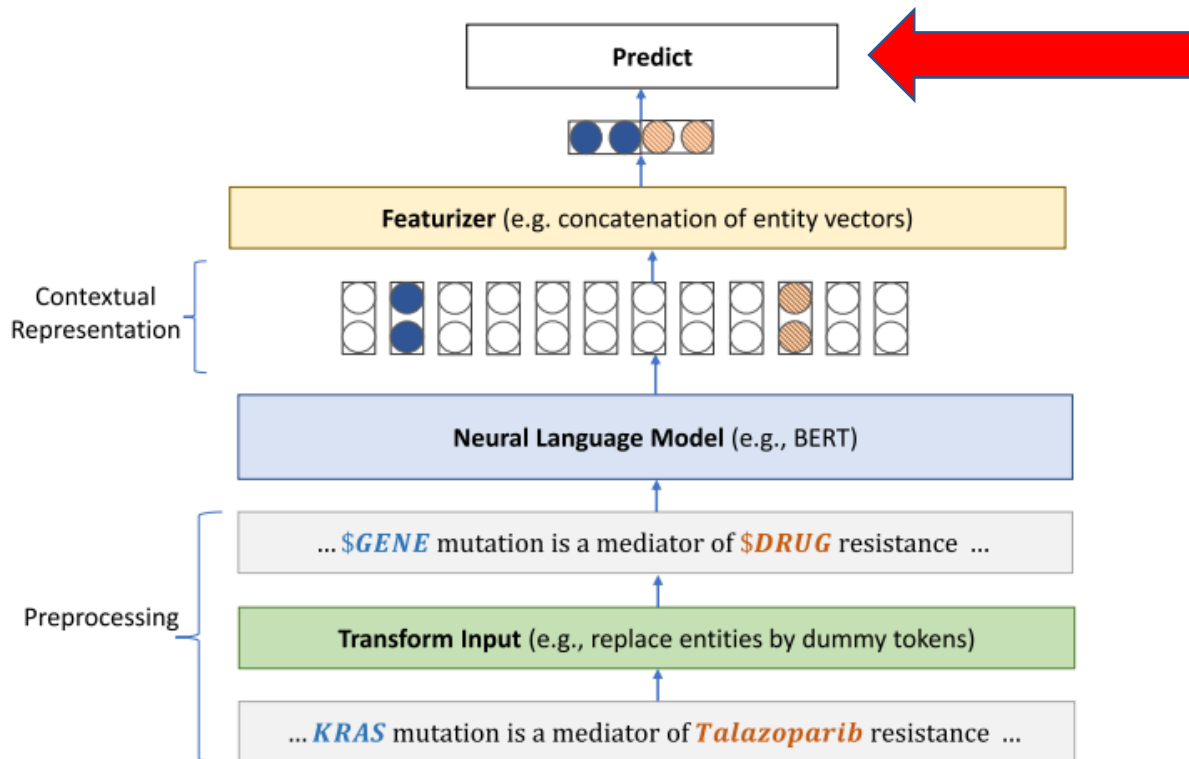
label : yes/no/maybe

question (string)	context (sequence)	long_answer (string)	final_decision (string)
"Do mitochondria play a role in remodelling lace plant leaves during..."	{ "contexts": ["Programmed cell death (PCD) is the regulated death of cells..."	"Results depicted mitochondrial dynamics in vivo as PCD progresses within the lace..."	"yes"
"Landolt C and snellen e acuity: differences in strabismus amblyopia?"	{ "contexts": ["Assessment of visual acuity depends on the optotypes used for..."	"Using the charts described, there was only a slight overestimation of visual acuity b..."	"no"
"Syncope during bathing in infants, a pediatric form of water-induced..."	{ "contexts": ["Apparent life-threatening events in infants are a..."	"Aquagenic maladies" could be a pediatric form of the aquagenic urticaria."	"yes"

- BioASQ : question from a PubMed abstract
 answer: Yes /No



4. Task-Specific Fine-Tuning



- Cross-entropy loss for classification tasks
- Mean square error for regression tasks
- Hyperparameter search using the development set based on task-specific metrics.
- Fine-tune the parameters of the task-specific prediction layer as well as the underlying neural language model.

Fig. 2. A general architecture for task-specific fine-tuning of neural language models, with a relation-extraction example. Note that the input goes through additional processing such as word-piece tokenization in the neural language model module.



5. Experimental Settings

- We generate the vocabulary and conduct pretraining using the latest collection of PubMed5 abstracts: 14 million abstracts, 3.2 billion words, 21 GB.
- Training is done for 62,500 steps with batch size of 8,192
- The training takes about 5 days on one DGX-2 machine with 16 V100 GPUs
- We use whole word masking , with a masking rate of 15%.
- For comparison, we use the public releases of BERT , RoBERTa , BioBERT , SciBERT , ClinicalBERT , and BlueBERT .



Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

BERT	BERT-base 12 transformer layers and 100 million parameters
RoBERTa	Same architecture as BERT, but uses a byte-level BPE as a tokenizer (same as GPT-2) and uses a different pretraining scheme.
BioBERT	BERT + continual training with PubMed abstract, PMC fulltext
SciBERT	Train from full text of papers from the corpus of semanticscholar.org.
ClinicalBERT	continual pretraining from BioBERT
BlueBERT	BERT + continual training with PubMed text and de-identified clinical notes from MIMIC-III



Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

RESULTS



1. Domain-Specific Pretraining vs Mixed-Domain Pretraining

Table 6. Comparison of Pretrained Language Models on the BLURB Biomedical NLP Benchmark

	BERT		RoBERTa	BioBERT	SciBERT		ClinicalBERT	BlueBERT	PubMedBERT
	Uncased	Cased	Cased	Cased	Uncased	Cased	Cased	Cased	Uncased
BC5-chem	89.25	89.99	89.43	92.85	92.49	92.51	90.80	91.19	93.33
BC5-disease	81.44	79.92	80.65	84.70	84.54	84.70	83.04	83.69	85.62
NCBI-disease	85.67	85.87	86.62	89.13	88.10	88.25	86.32	88.04	87.82
BC2GM	80.90	81.23	80.90	83.82	83.36	83.36	81.71	81.87	84.52
JNLPBA	77.69	77.51	77.86	78.55	78.68	78.51	78.07	77.71	79.10
EBM PICO	72.34	71.70	73.02	73.18	73.12	73.06	72.06	72.54	73.38
ChemProt	71.86	71.54	72.98	76.14	75.24	75.00	72.04	71.46	77.24
DDI	80.04	79.34	79.52	80.88	81.06	81.22	78.20	77.78	82.36
GAD	80.41	79.61	80.63	82.36	82.38	81.34	80.48	79.15	83.96
BIOSSES	82.68	81.40	81.25	89.52	86.25	87.15	91.23	85.38	92.30
HoC	80.20	80.12	79.66	81.54	80.66	81.16	80.74	80.48	82.32
PubMedQA	51.62	49.96	52.84	60.24	57.38	51.40	49.08	48.44	55.84
BioASQ	70.36	74.44	75.20	84.14	78.86	74.22	68.50	68.71	87.56
BLURB score	76.11	75.86	76.46	80.34	78.86	78.14	77.29	76.27	81.16

The standard task-specific models are used in the same fine-tuning process for all BERT models. The BLURB score is the macro average of average test results for each of the six tasks (NER, PICO, relation extraction, sentence similarity, document classification, question answering). See Table 3 for the evaluation metric used in each task.



2. Ablation Study on Pretraining Techniques

Effect of vocabulary and WWM

Table 7. Evaluation of the Impact of Vocabulary and WWM on the Performance of PubMedBERT on BLURB

	Wiki + Books		PubMed	
	Word Piece	Whole Word	Word Piece	Whole Word
BC5-chem	93.20	93.31	92.96	93.33
BC5-disease	85.00	85.28	84.72	85.62
NCBI-disease	88.39	88.53	87.26	87.82
BC2GM	83.65	83.93	83.19	84.52
JNLPBA	78.83	78.77	78.63	79.10
EBM PICO	73.30	73.52	73.44	73.38
ChemProt	75.04	76.70	75.72	77.24
DDI	81.30	82.60	80.84	82.36
GAD	83.02	82.42	81.74	83.96
BIOSSES	91.36	91.79	92.45	92.30
HoC	81.76	81.74	80.38	82.32
PubMedQA	52.20	55.92	54.76	55.84
BioASQ	73.69	76.41	78.51	87.56
BLURB score	79.16	79.96	79.62	81.16





Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

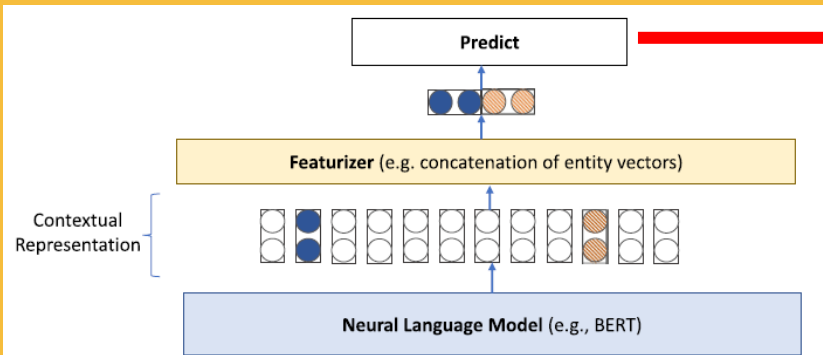
Table 9. Evaluation of the Impact of Pretraining Corpora and Time on the Performance on BLURB

Pretraining Vocab	Wiki + Books → PubMed		PubMed (half time)	PubMed
	Wiki + Books	PubMed	PubMed	PubMed
BC5-chem	92.85	93.41	93.05	93.33
BC5-disease	84.70	85.43	85.02	85.62
NCBI-disease	89.13	87.60	87.77	87.82
BC2GM	83.82	84.03	84.11	84.52
JNLPBA	78.55	79.01	78.98	79.10
EBM PICO	73.18	73.80	73.74	73.38
ChemProt	76.14	77.05	76.69	77.24
DDI	80.88	81.21	81.96	82.36
GAD	82.36	82.47	82.80	83.96
BIOSSES	89.52	89.93	92.12	92.30
HoC	81.54	83.14	82.13	82.32
PubMedQA	60.24	54.84	55.28	55.84
BioASQ	84.14	79.00	79.43	87.56
BLURB score	80.34	80.03 	80.23	81.16 

In the first two columns, pretraining was first conducted on Wiki & Books, then on PubMed abstracts. All use the same amount of compute (twice as long as original BERT pretraining), except for the third column, which only uses half (same as original BERT pretraining).



3. Ablation Study on Fine-Tuning methods



The top layer in the BERT already captures many non-linear dependencies across the entire text span.

Table 12. Comparison of Linear Layers vs Recurrent Neural Networks for Task-Specific Fine Tuning in NER (Entity-level F1) and Relation Extraction (Micro F1), All Using the Standard PubMedBERT

Task-Specific Model	Linear Layer	Bi-LSTM
BC5-chem	93.33	93.12
BC5-disease	85.62	85.64
JNLPBA	79.10	79.10
ChemProt	77.24	75.40
DDI	82.36	81.70
GAD	83.96	83.42



Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

DISCUSSION & CONCLUSION



- The prevailing assumption is that such mixed-domain pretraining is advantageous.
- We show that this type of transfer learning may not be applicable when there is a sufficient amount of in-domain text, as is the case in biomedicine.
- Comparing clinical BERTs with PubMedBERT on biomedical NLP tasks show that even related text such as clinical notes may not be helpful, since we already have abundant biomedical text from PubMed.
- Our results show that we should distinguish different types of transfer learning and separately assess their utility in various situations.



Mahidol University









Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

- Comprehensive benchmarks and leaderboards are available for the general domains (e.g., GLUE and SuperGLUE), they are still a rarity in biomedical NLP.
- In this article, we create the first leaderboard for biomedical NLP, BLURB—a comprehensive benchmark containing 13 datasets for six tasks
- To accelerate research in biomedical NLP, we release our state-of-the-art biomedical BERT models and set up a leaderboard based on BLURB.
- Future directions include further exploration of domain-specific pretraining strategies, incorporating more tasks in biomedical NLP, and extension of the BLURB benchmark to clinical and other high-value domains.



BULRB Leaderboard

BLURB									
Leaderboard Paper Models Tasks Submit News									
The Overall score is calculated as the macro-average performance over tasks. Details can be found within our publication .									
Show <input type="text" value="100"/> entries									
Rank	Model	BLURB Score (Macro Avg.)	Micro Avg.	NER	PICO	RE	SS	Class.	QA
1	BioLinkBERT-Large — Stanford  	84.30	84.80	86.89	74.19	82.74	93.63	84.88	83.50
2	BioM-ELECTRA-Large — University of Delaware  	83.81	84.67	86.88	73.67	83.17	91.09	84.03	84.00
3	BioLinkBERT-Base — Stanford  	83.39	83.84	86.39	73.97	81.56	93.27	84.35	80.81
4	PubMedBERT-LARGE (fine-tuning stabilization; uncased; abstracts) — Microsoft Research  	82.91	83.58	86.28	73.61	81.77	92.73	82.70	80.37



Mahidol University

Faculty of Medicine Ramathibodi Hospital

Department of Clinical Epidemiology and Biostatistics

THANK YOU