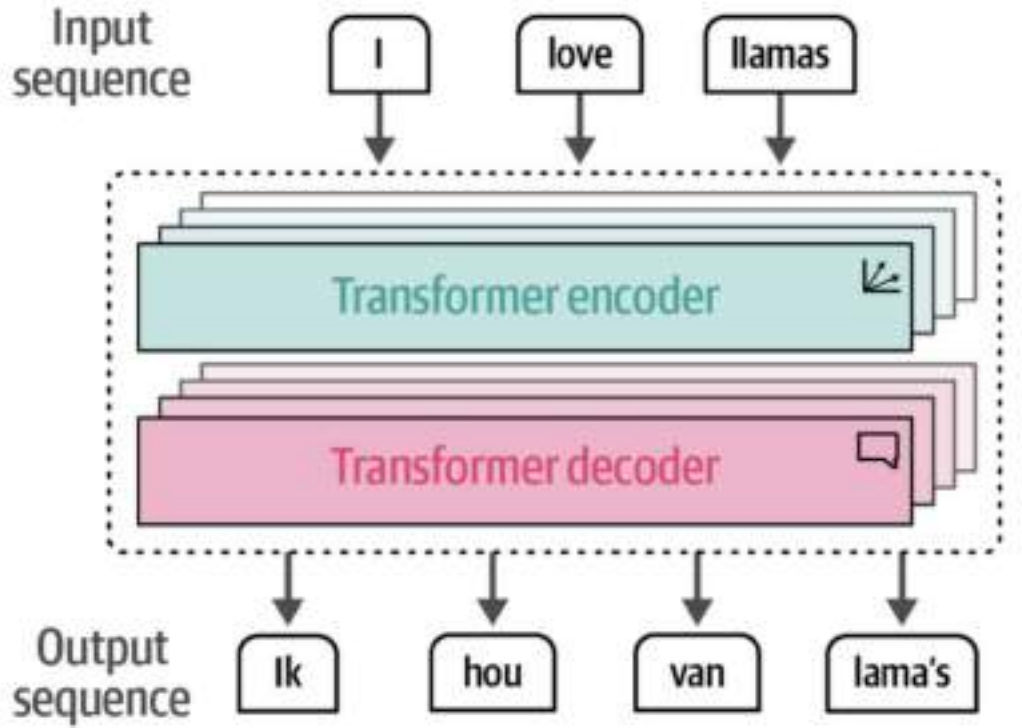
Computing for Medicine

BERT

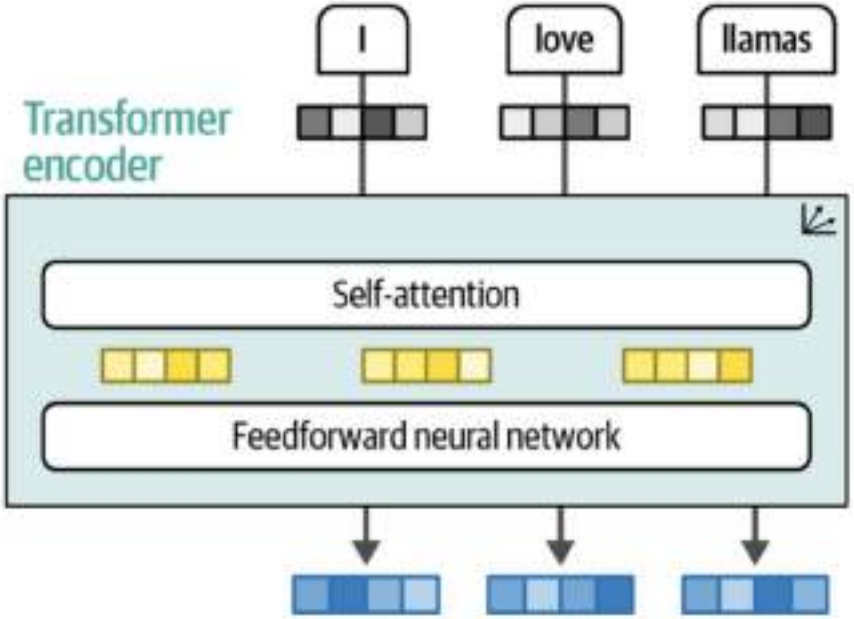
@Tavpritesh
#Computing4Medicine

Recap

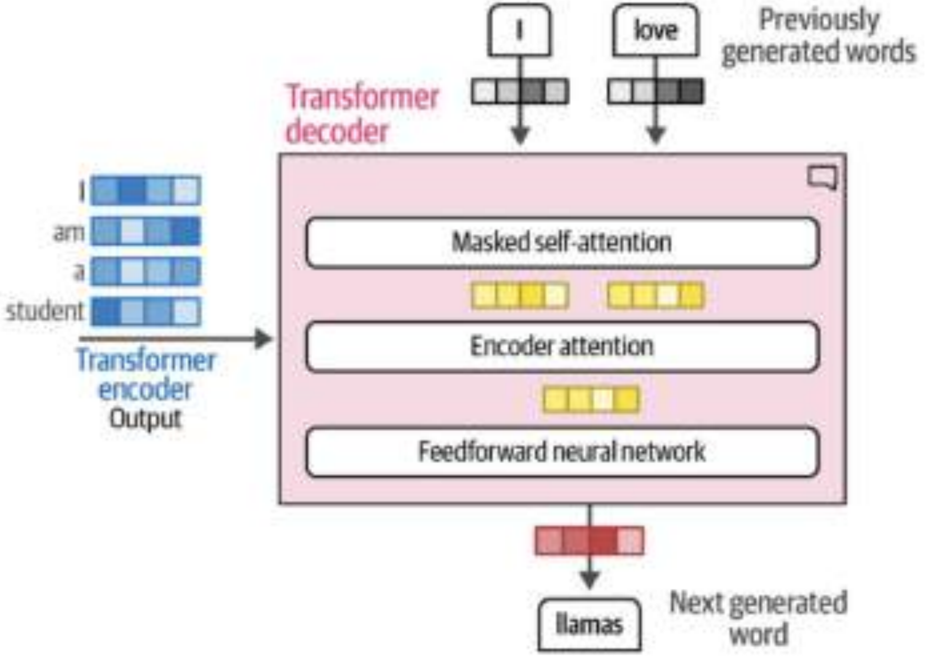


Input sequence
I
love
llamas
Transformer encoder
Transformer decoder
Output sequence
Ik
hou
van
lama's

Encoder and Decoder Side by Side



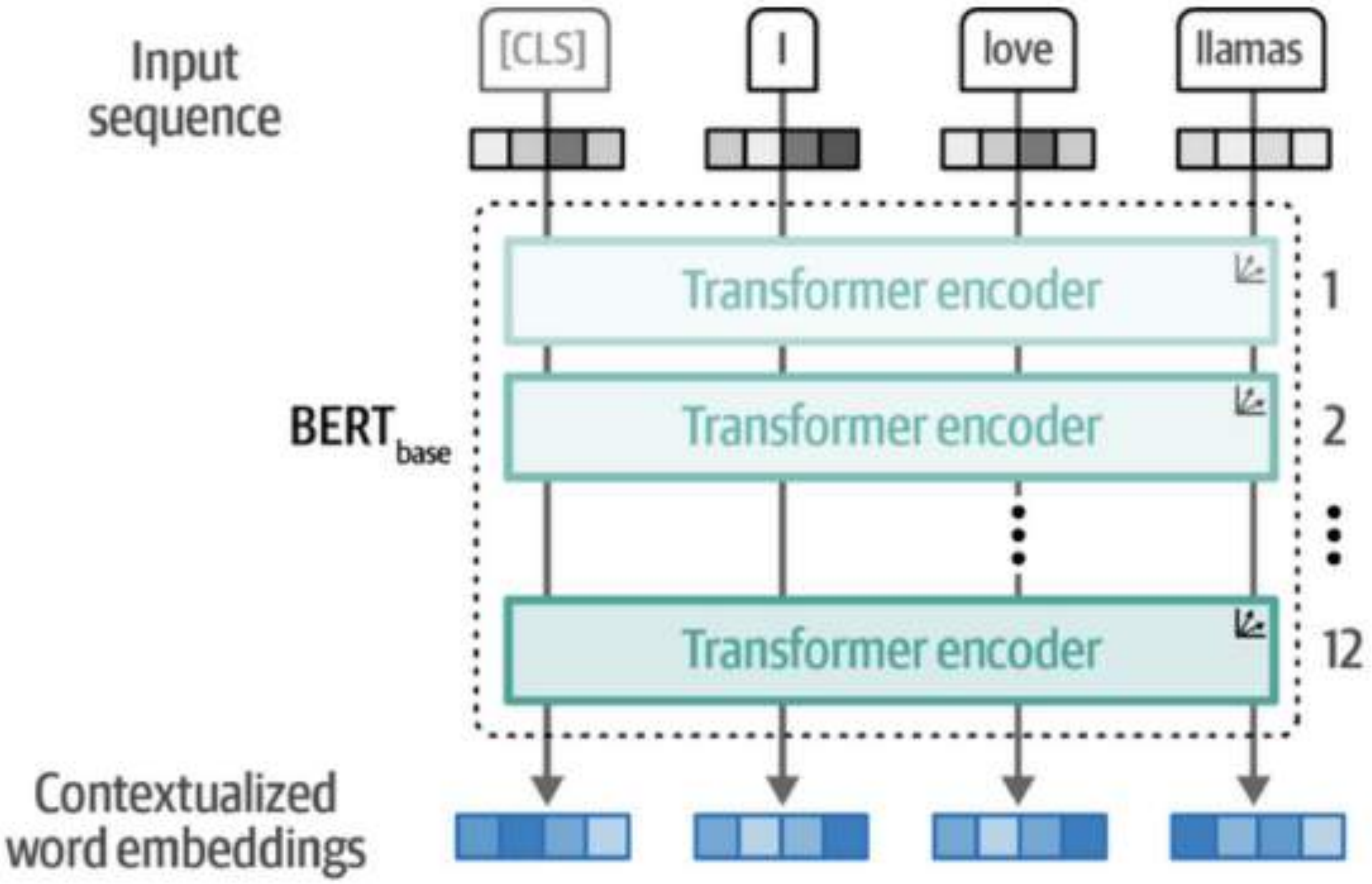
Encoder

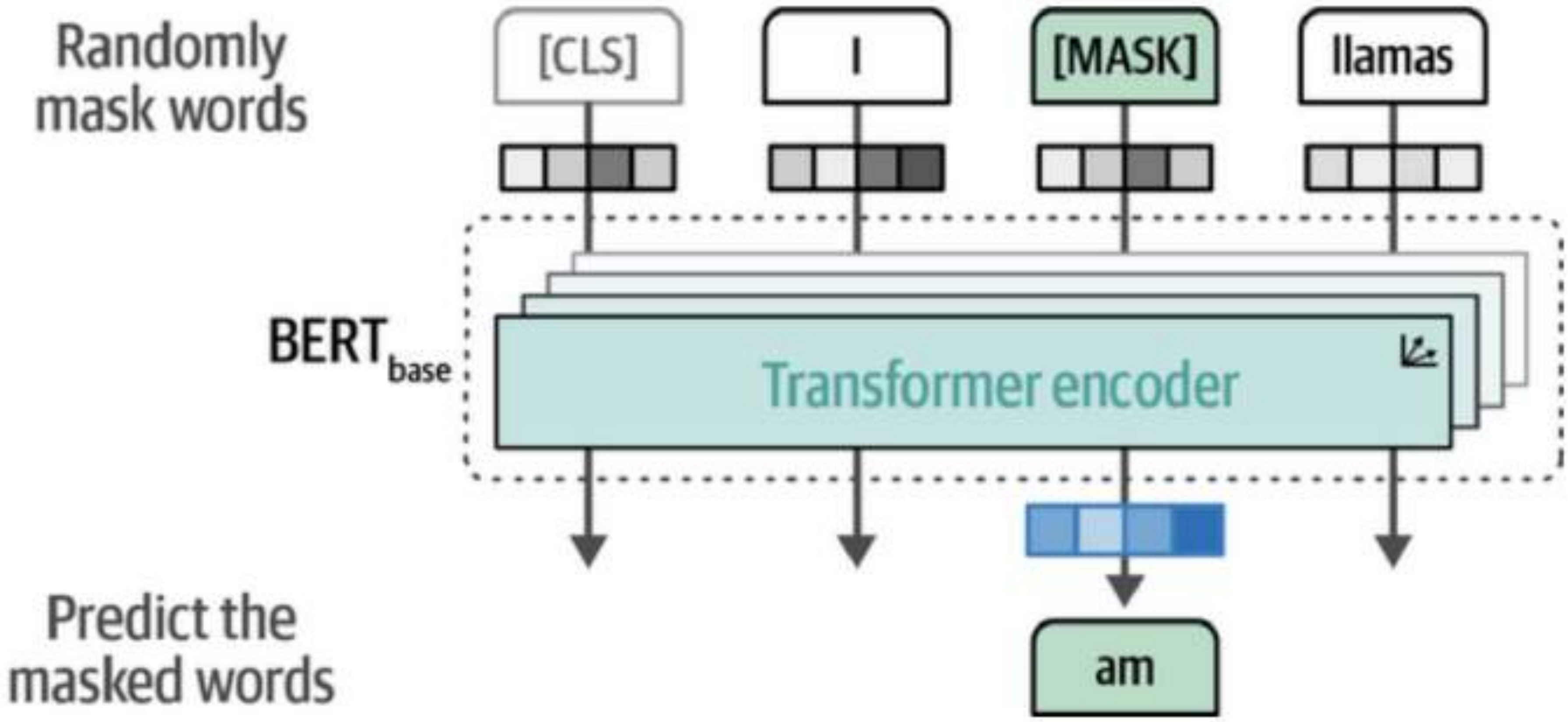


Decoder

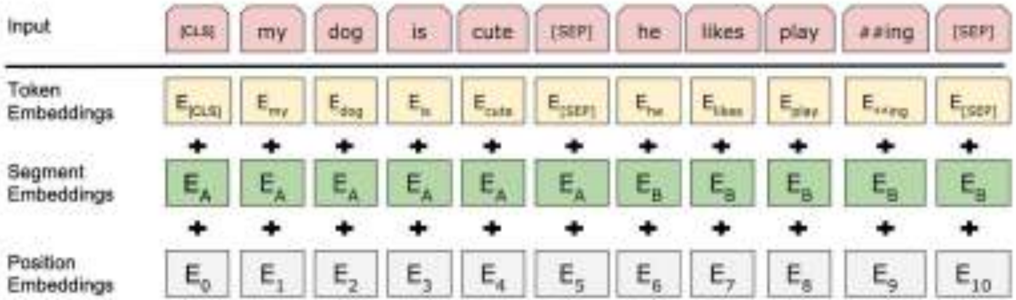
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

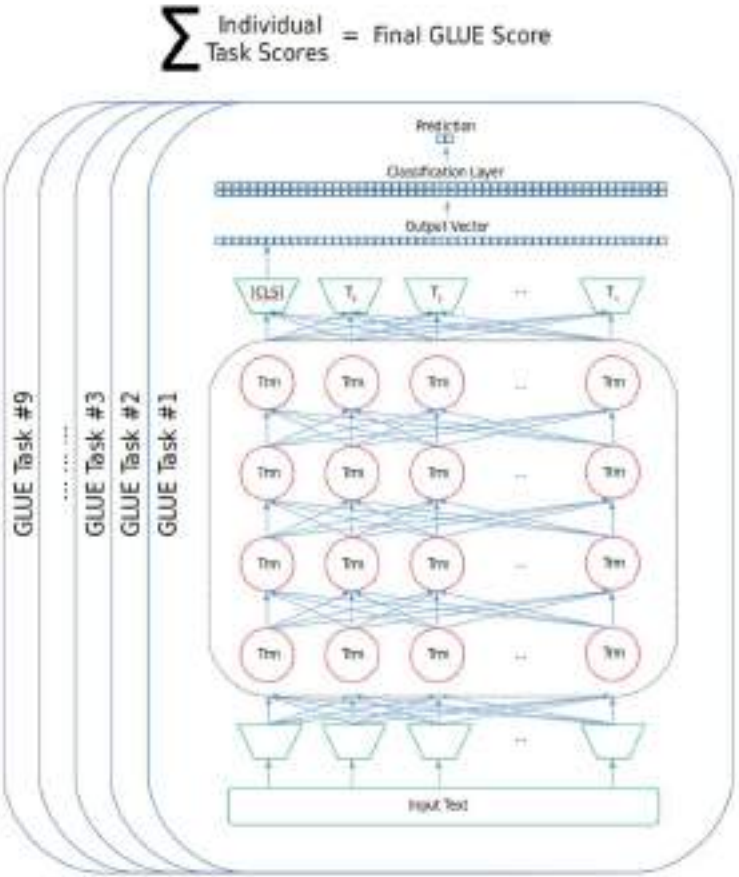
Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

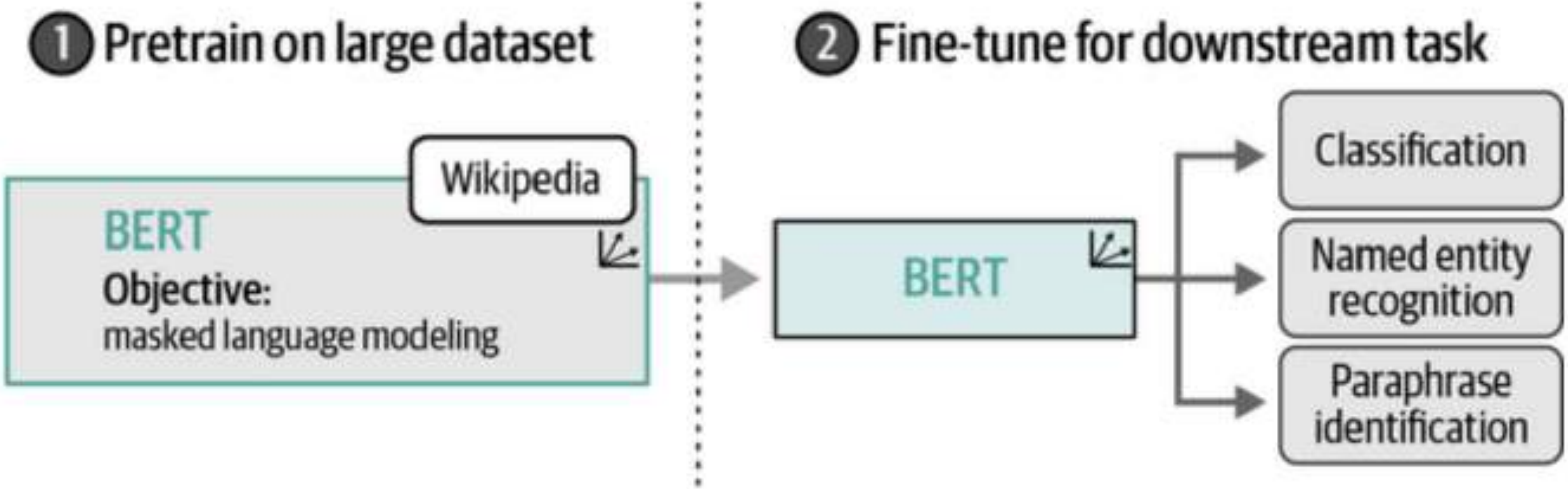

Input sequence
[CLS]
I
love
llamas
BERT base
Transformer encoder
1
Transformer encoder
2
Transformer encoder
12
Contextualized word embeddings


Randomly mask words
[CLS]
I
[MASK]
llamas
BERT$_{base}$
Transformer encoder
Predict the masked words
am

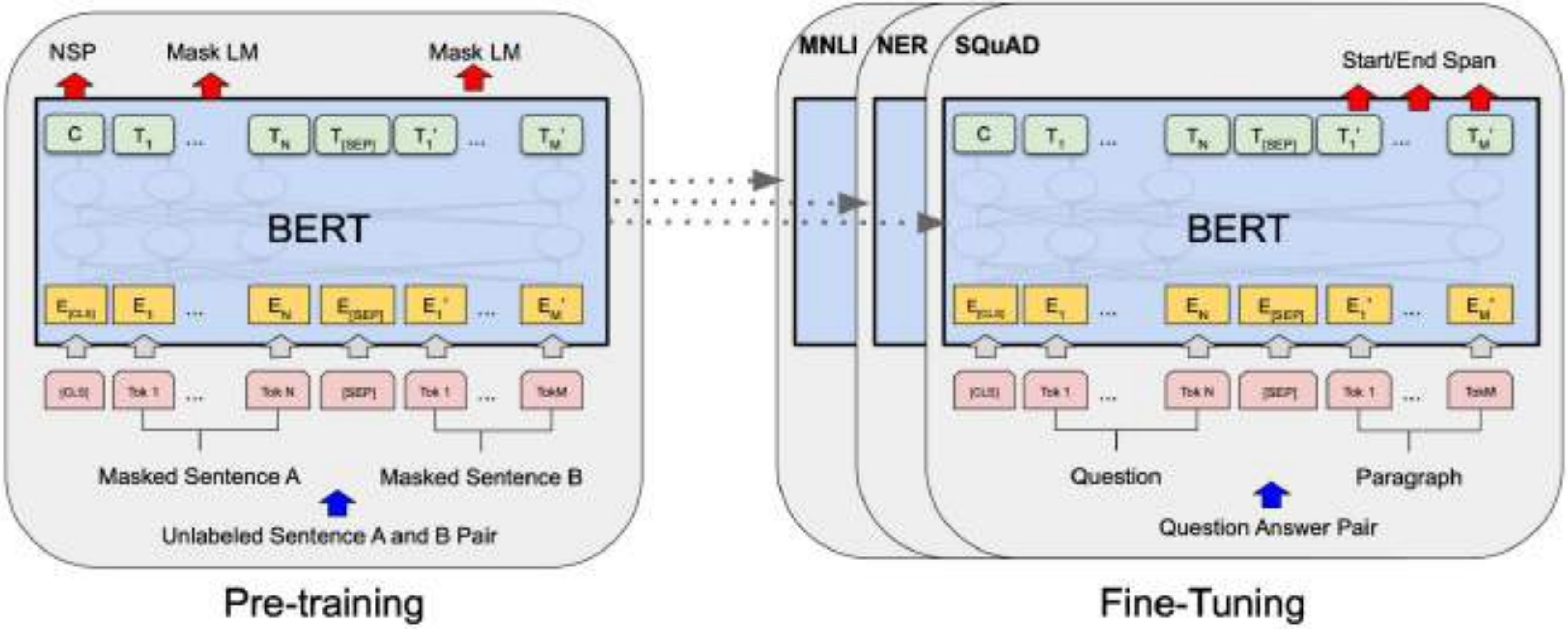
BERT and General Language Understanding Evaluation (GLUE)





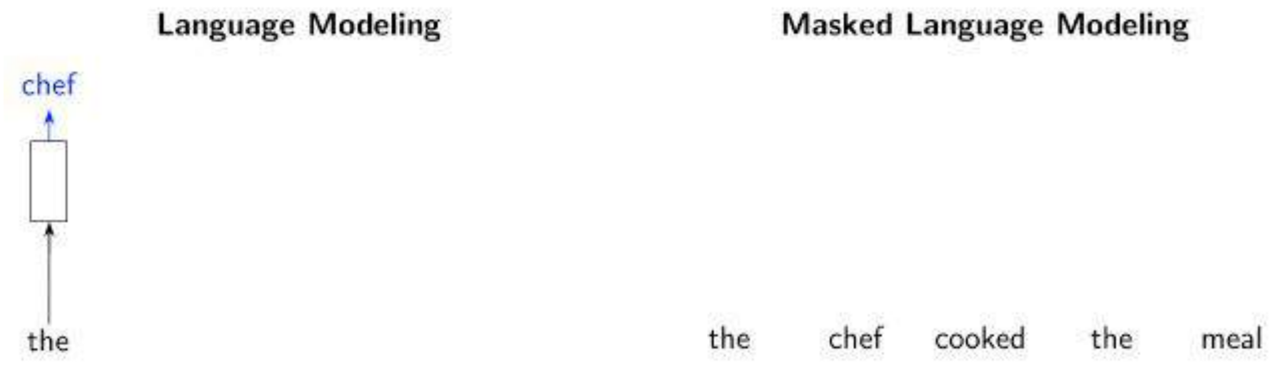

1 Pretrain on large dataset
2 Fine-tune for downstream task
Wikipedia
BERT
Objective:
masked language modeling
BERT
Classification
Named entity recognition
Paraphrase identification

BERT



NSP
Mask LM
Mask LM
C
BERT
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair
Pre-training
MNLI
NER
SQuAD
Start/End Span
BERT
Question
Paragraph
Question Answer Pair
Fine-Tuning

Masked Language Model



https://ai.googleblog.com/2020/06/pegasus-state-of-art-model-for.html

Bioinformatics, 2019, 1–7
doi: 10.1093/bioinformatics/btz682
Advance Access Publication Date: 10 September 2019
Original Paper

Data and text mining

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

Jinhyuk Lee [1,†], Wonjin Yoon [1,†], Sungdong Kim [2], Donghyeon Kim [1], Sunkyu Kim [1], Chan Ho So [3] and Jaewoo Kang [1,3,*]

[1]Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea, [2]Clova AI Research, Naver Corp, Seong-Nam 13561, Korea and [3]Interdisciplinary Graduate Program in Bioinformatics, Korea University, Seoul 02841, Korea

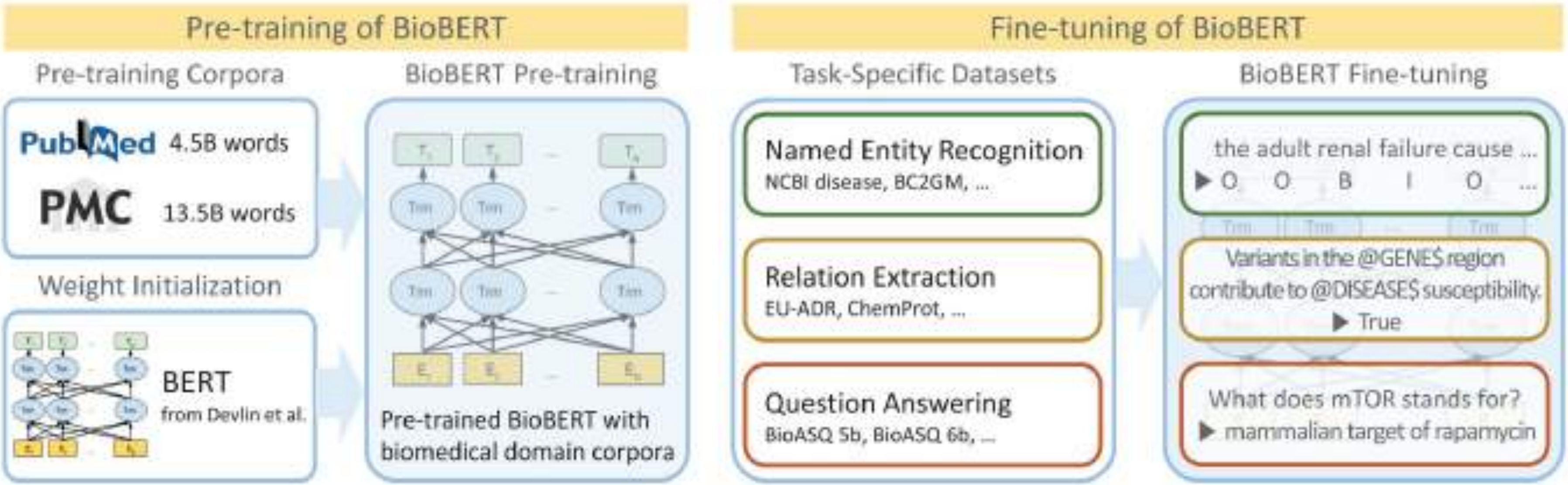
*To whom correspondence should be addressed.
[†]The authors wish it to be known that the first two authors contributed equally.

Associate Editor: Jonathan Wren

Received on May 16, 2019; revised on July 29, 2019; editorial decision on August 25, 2019; accepted on September 5, 2019

BioBERT

BERT: Bidirectional Encoder Representations from Transformers



Training

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

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

Publicly Available Clinical BERT Embeddings

Emily Alsentzer
Harvard-MIT
Cambridge, MA
emilya@mit.edu

John R. Murphy
MIT CSAIL
Cambridge, MA
jrmurphy@mit.edu

Willie Boag
MIT CSAIL
Cambridge, MA
wboag@mit.edu

Wei-Hung Weng
MIT CSAIL
Cambridge, MA
ckbjimmy@mit.edu

Di Jin
MIT CSAIL
Cambridge, MA
jindi15@mit.edu

Tristan Naumann
Microsoft Research
Redmond, WA
tristan@microsoft.com

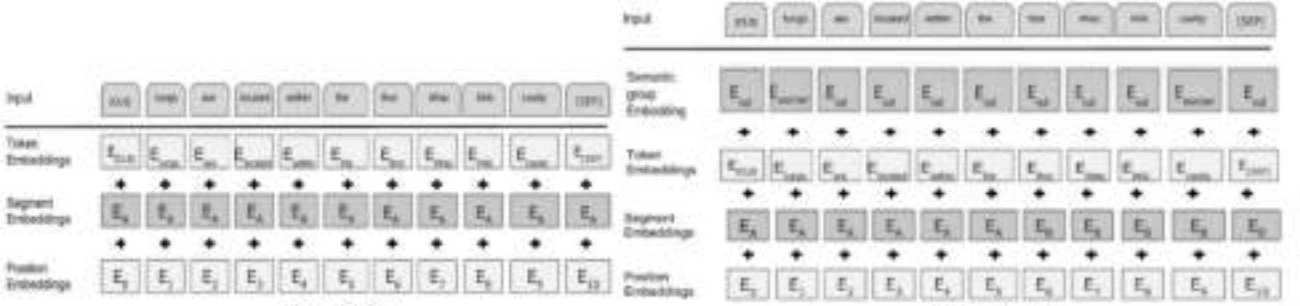
Matthew B. A. McDermott
MIT CSAIL
Cambridge, MA
mmd@mit.edu

Model	MedNLI	i2b2 2006	i2b2 2010	i2b2 2012	i2b2 2014
BERT	77.6%	93.9	83.5	75.9	92.8
BioBERT	80.8%	**94.8**	86.5	78.9	**93.0**
Clinical BERT	80.8%	91.5	86.4	78.5	92.6
Discharge Summary BERT	80.6%	91.9	86.4	78.4	92.8
Bio+Clinical BERT	**82.7%**	94.7	87.2	**78.9**	92.5
Bio+Discharge Summary BERT	**82.7%**	94.8	**87.8**	78.9	92.7

UMLS BERT

UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus

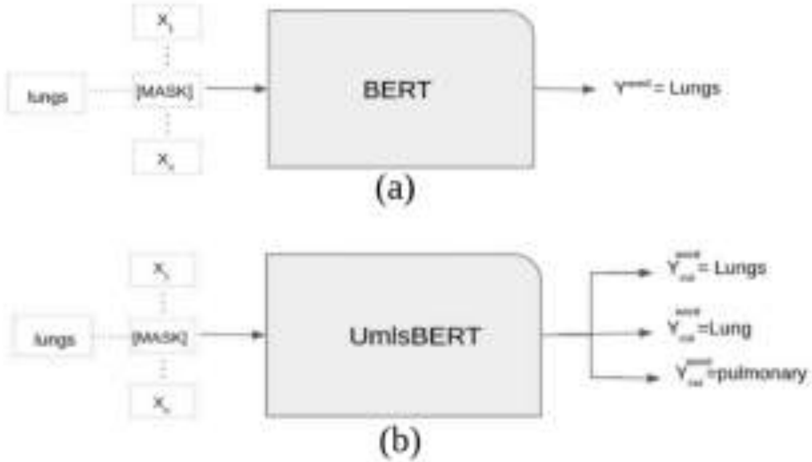
George Michalopoulos[1], Yuanxin Wang[1], Hussam Kaka[1], Helen Chen[1], Alex Wong[1]
[1]University of Waterloo
{gmichalo, yuanxin.wang, hussam.kaka, helen.chen, alexander.wong}@uwaterloo.ca



(a) BERT (b) UmlsBERT

UMLS BERT

Semantic group embeddings Firstly, we introduced a new embedding matrix called $SG \in \mathbb{R}^{d \times D_s}$ into the input embedding of the BERT model, where d is BERT's transformer hidden dimension and $D_s = 6$ is the number of unique UMLS semantic groups that could be identified in the vocabulary of our model. In particular, in this matrix, each row represents the unique semantic group in UMLS that a word can be identified with (for example the word 'heart' is associated with the semantic group 'Anatomy' in UMLS).



Publicly Available Clinical BERT Embeddings

Emily Alsentzer
Harvard-MIT
Cambridge, MA
emilya@mit.edu

John R. Murphy
MIT CSAIL
Cambridge, MA
jrmurphy@mit.edu

Willie Boag
MIT CSAIL
Cambridge, MA
wboag@mit.edu

Wei-Hung Weng
MIT CSAIL
Cambridge, MA
ckbjimmy@mit.edu

Di Jin
MIT CSAIL
Cambridge, MA
jindi15@mit.edu

Tristan Naumann
Microsoft Research
Redmond, WA
tristan@microsoft.com

Matthew B. A. McDermott
MIT CSAIL
Cambridge, MA
mmd@mit.edu

Model	MedNLI	i2b2 2006	i2b2 2010	i2b2 2012	i2b2 2014
BERT	77.6%	93.9	83.5	75.9	92.8
BioBERT	80.8%	**94.8**	86.5	78.9	**93.0**
Clinical BERT	80.8%	91.5	86.4	78.5	92.6
Discharge Summary BERT	80.6%	91.9	86.4	78.4	92.8
Bio+Clinical BERT	**82.7%**	94.7	87.2	**78.9**	92.5
Bio+Discharge Summary BERT	**82.7%**	94.8	**87.8**	78.9	92.7

From NLP to Natural Language Understanding

PEGASUS

PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization

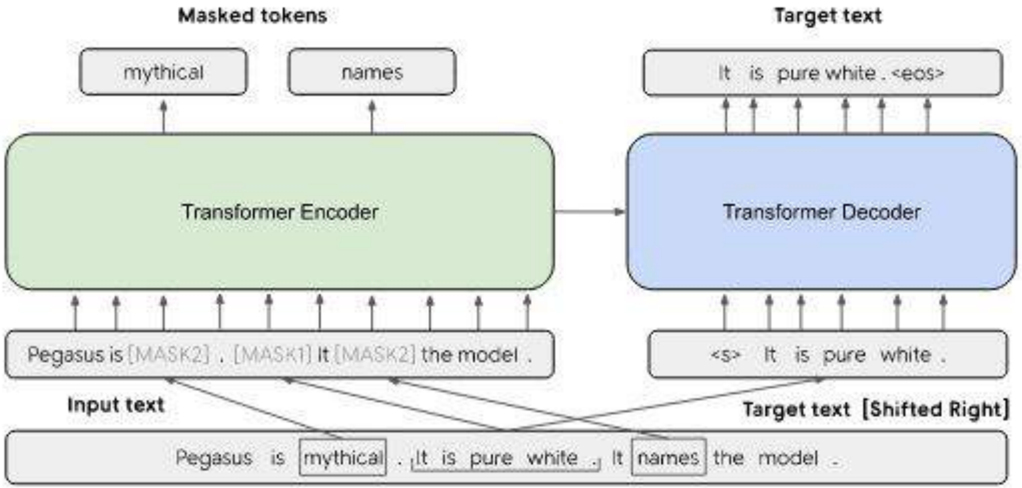
Jingqing Zhang *[1] Yao Zhao *[2] Mohammad Saleh [2] Peter J. Liu [2]

PEGASUS

Core idea: Gap Sentence Generation
Uses self-supervised pre-training objective; complete sentences masked
Seq2seq transformer architecture

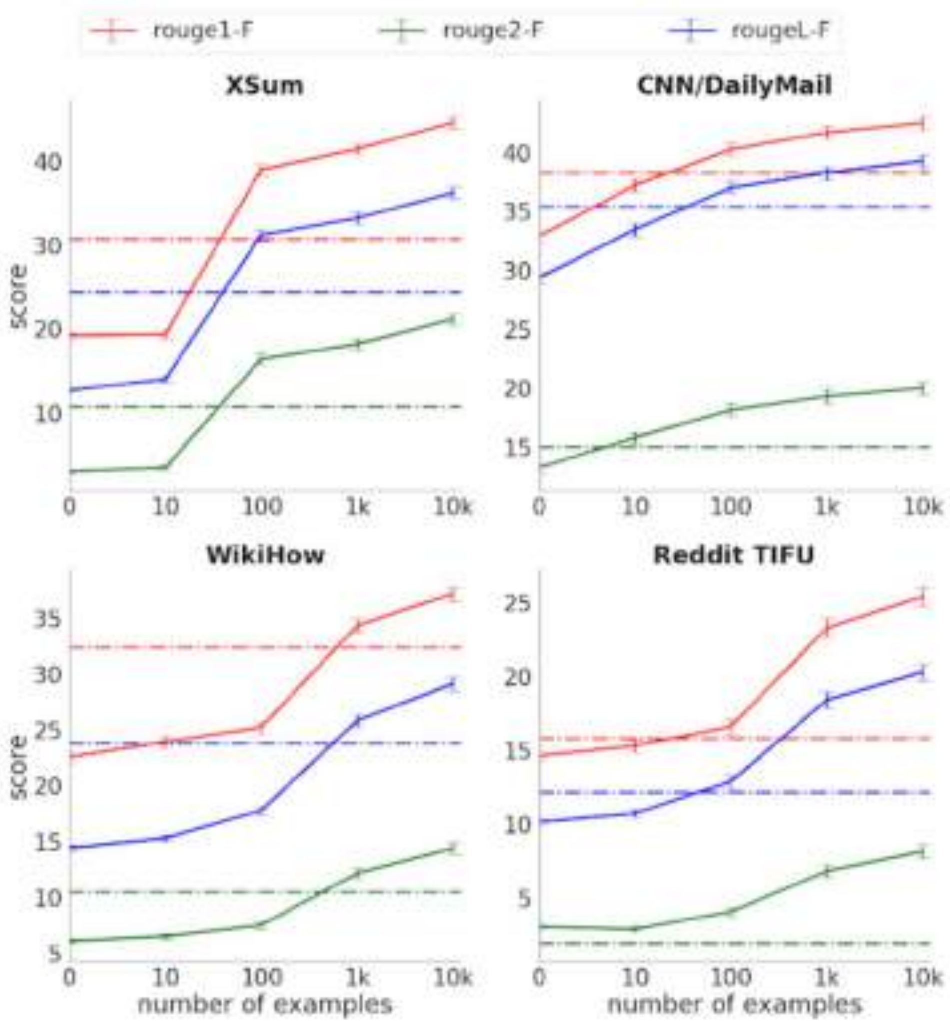
https://ai.googleblog.com/2020/06/pegasus-state-of-art-model-for.html

Combine MLM with GSG



https://ai.googleblog.com/2020/06/pegasus-state-of-art-model-for.html

Performance



Navy frigates in Portsmouth 'to be sunk or scrapped'

5 February 2013

Four ships | Five ships | Two ships | Three ships | Six ships

The decommissioned Type 22 frigates
HMS Cumberland, HMS Campbeltown, HMS Chatham and HMS Cornwall
are currently moored in Portsmouth Harbour.
Bidders had until 23 January to register an interest in the former Devonport-based ships. The BBC understands no proposals to preserve the ships have been submitted. Those who have registered an interest are finalising their bids with viewings set to take place in late February and March. A final decision is not expected until the spring. The government's Disposal Services Authority, which is handling the sale, wants to award at least one of the frigates to a UK ship recycler to determine the capacity of the UK's industry in the field. Penny Mordaunt, Conservative MP for Portsmouth North, said it was important UK recyclers had the chance to prove themselves in the field but she was also keen to see at least one of them saved from the scrapyard. She added: "For anyone that has served on a ship it's your home, you've literally been through the wars with it... and you want them to have a noble second life. "My preference is to go for the reef and diving attraction. "We've got to get best value for the budget but a reef would also generate income for part of the country through tourism." The Ministry of Defence has previously said it will "consider all options" for the frigates to ensure "best financial return for the taxpayer". A spokeswoman would not comment on the number or nature of the bids received due to "commercial sensitivity". Originally designed as a specialist anti-submarine ship, the Type 22 frigate evolved into a powerful surface combatant with substantial anti-surface, anti-submarine and anti-aircraft weapons systems. They were also known for having excellent command and control, and communication facilities, making them ideal flagships on deployments, with a complement of about 280 crew. Last year, the aircraft carrier HMS Ark Royal was sold as scrap for £3m.

Model Summary: No proposals have been submitted to preserve four Royal Navy frigates for reuse, the BBC has learned.

PEGASUS correctly abstracted "four Royal Navy frigates" from an article that mentioned HMS Cumberland, HMS Campbeltown, HMS Chatham and HMS Cornwall…!!!

Reusable Embeddings

Available embeddings for clinical data and concepts. Since ELMo models use character information and BERT models use sub-word information, they can generate a representation for any concept.

Name	Model	Data/Concepts	Terms	Dim.
PubMed-w2v.bin[a]	word2vec	PubMed	2.4 M	200
PMC-w2v.bin[b]	word2vec	PubMed Central	2.5 M	200
PubMed-and-PMC-w2v.bin[c]	word2vec	PubMed, PubMed Central	4.1 M	200
wikipedia-pubmed-and-PMC-w2v.bin[d]	word2vec	PubMed, PubMed Central, Wikipedia	5.5 M	200
drug word embeddings[e]	word2vec	PubMed, DrugBank	553,195	420
AWE-CM [49]	word2vec	UMLS CUI (concepts)	265 M	300
claims_codes_hs_300 [55]	word2vec	ICD-9 codes (concepts)	51,327	300
claims_cuis_hs_300 [55]	word2vec	UMLS CUI (concepts)	14,852	300
cui2vec [56]	word2vec/GloVe	UMLS CUI (concepts)	108,477	500
concept embeddings [58]	AiTextML	MeSH ID (concepts)	26,103	100
word embeddings [58]	AiTextML	PubMed	513,196	100
ELMo (PubMed model) [11]	ELMo	PubMed	NA	1024
BioBERT [15]	BERT	PubMed	NA	768/1024
ClinicalBERT [16,17]	BERT	MIMIC III	NA	768

[a] http://evexdb.org/pmresources/ngrams/PubMed/.
[b] http://evexdb.org/pmresources/ngrams/PMC/.
[c] http://evexdb.org/pmresources/vec-space-models/wikipedia-pubmed-and-PMC-w2v.bin.
[d] http://evexdb.org/pmresources/vec-space-models/wikipedia-pubmed-and-PMC-w2v.bin.
[e] https://github.com/chop-dbhi/drug_word_embeddings.

Evaluation Tasks

	Word embeddings		Evaluation	
Paper	Corpora	Model	Intrinsic	Extrinsic
De Vine et al., 2015 [30]		word2vec		clinical information extraction
Chiu et al., 2016 [72]	PMC, Pubmed, PMC + Pubmed	word2vec	relatedness, similarity	NER
Dubois et al., 2017 [32]	text notes, OHSUMED	GloVe	–	disease prediction, mortality prediction
Wang et al., 2018 [29]	Mayo Clinic text notes, PubMed, Wikipedia, Google News	GloVe	similarity (qualitative),	clinical information extraction, reaction extraction
Huang et al., 2016 [46]	MedHelp online forum, PubMed, Wikipedia		cluster quality evaluation	
Choi et al., 2016 [55]	OHSUMED, medical claims	word2vec	conceptual similarity, medical relatedness	–
Yu et al., 2016 [52])	UMLS and MeSH terms	LDA	UMLS-Similarity	–
Mencia et al., 2016 [58]	BioASQ, PubMed	All-in-text	MiniMayoSRS, UMNSRS similarity/relatedness	–
Bong et al., 2017 [49]	MIMIC-III	word2vec	similarity	–
Patel et al., 2017 [35]	PubMed, medical claims	word2vec	medical term similarity	medical coding review
Beam et al., 2018 [56]	PubMed, medical claims, UMLS semantic types	word2vec, Glove	comorbidity, causative, and drug-conditions relations and UMNSRS similarity/relatedness	
Zhao et al., 2018 [43]	PubMed, DrugBank	word2vec	UMNSRS similarity/relatedness	drug name recognition/classification
Craig et al., 2017 [37]	discharge notes	word2vec (Skipgram)	—	30-day unplanned prediction
Nguyen et al., 2017 [38]	hospital patient records	random init, word2vec	cluster evaluation	unplanned readmission within 6 months prediction
Pham et al., 2016 [39]	hospital patient records	random init + advanced techniques	—	unplanned readmission prediction, high-risk patient prediction
Escudié et al., 2018 [33]	electronic health records from hospital	three-layer stack of denoising autoencoders	—	disease prediction
Gehrmann et al., 2018 [36]	discharge summaries from MIMIC-III	word2vec	word similarity	phenotype classification
Wang et al., 2018 [29]	clinical notes from hospital, PMC, News, Wikipedia + Gigaword	word2vec (Skip-gram)	word/semantic similarity	information extraction, smoking status prediction, fracture detection,
Kholgi et al., 2016 [34]	i2b2/VA 2010 [81], ShARe/CLEF 2013 eHealth Evaluation Lab [82]	word2vec (Skip-gram)	—	disease prediction, term extraction