Publicly Available Clinical BERT Embeddings

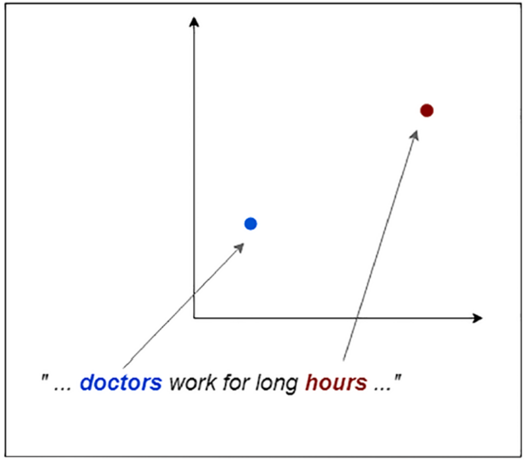
Ruyi Pan
Master student, Department of Statistical Sciences
University of Toronto
July 29, 2021

Publicly Available Clinical BERT Embeddings

- Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, Matthew B. A. McDermott

- arXiv, 20 June 2019, version 3

Preparation 1: Word Embeddings



- **Definition**

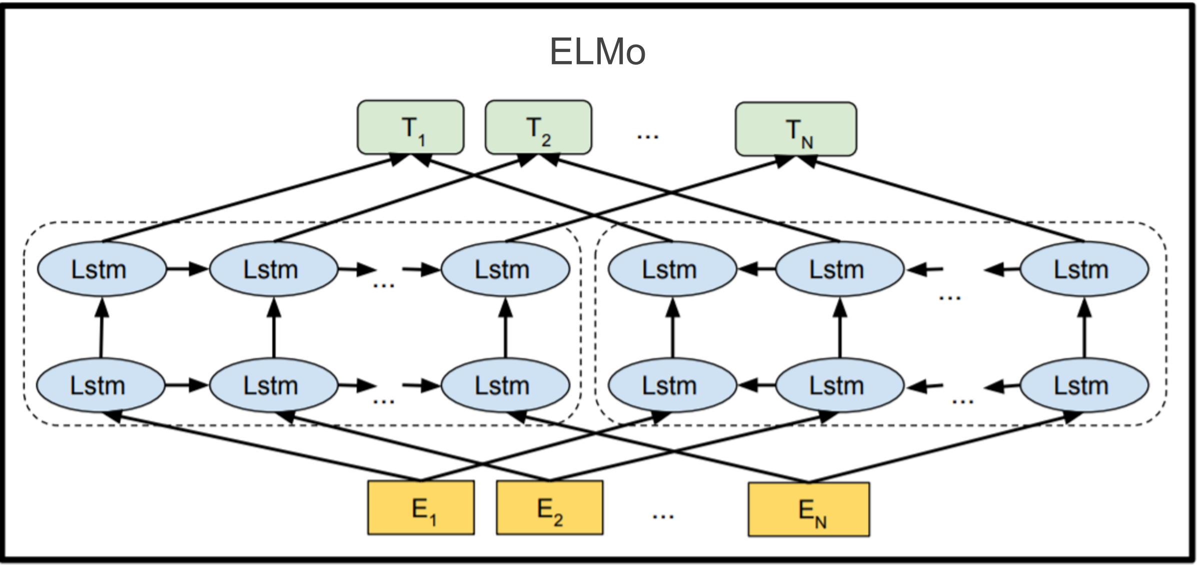
  A real-valued vector represents word

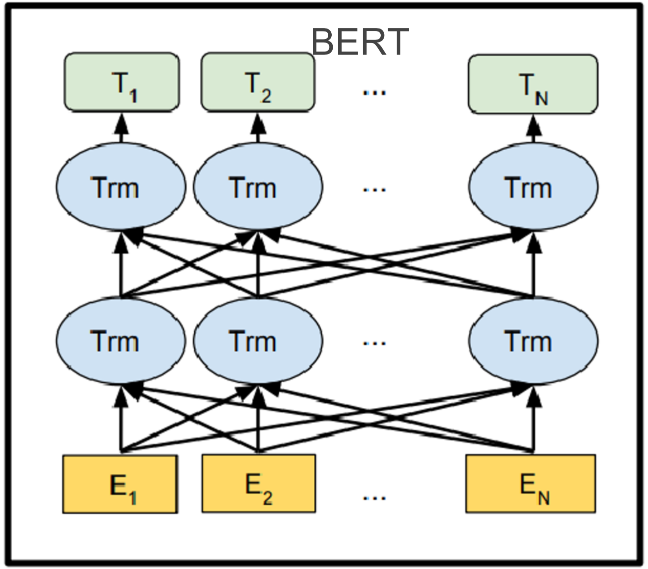
- **Type**
  - **Word2vec**
  - **Global Vectors (GloVe)**
  - **FastText**
  - **Embedding from Language Models (ELMo)**
  - **Bidirectional Encoder Representations from Transformers (BERT)**

Embeddings Comparison

Model Name	Context Sensitive	Particularities
Word2Vec	NO	**Prediction-based** model continuous bag-of-words CBOW/skip-gram(SG) (small neural networks), embedding dimensions, the length of context window
GloVe	NO	**Count-based** model, co-occurrence matrix
FastText	NO	Extending the word2vec SG model with internal **sub-word** information
ELMo	YES	**Contextualized** word embedding, look entire sentence, bidirectional, RNN
BERT	YES	Deep bidirectional representations, both **left and right context in all layers**

ELMo vs. BERT





BERT, in general, been found to be superior to ELMo and far superior to non-contextual embeddings.

Preparation 2: BERTS

- **General BERT**
  - BooksCorpus dataset (800M words), text passages of English Wikipedia
  - **BERT-Base** and **BERT-Large**
- **Domain-specific BERT**
  - **BioBERT**
    - PubMed abstracts and PMC full-text articles.
  - **SciBERT**
    - Random sample of 1.14M full-text papers from Semantic Scholar(18% computer science papers, 82% biomedical papers)
  - **ClinicalBERT**

Goal of the Paper

- Train and release Clinical BERT Models

- Examine the performance

Train Model

- **Data**

  Clinical text, 2 million notes, MIMIC-III database

  - Freely-available, de-identified health-related data, > 40,000 patients, Beth Israel Deaconess Medical Center (2001-2012)

- **Models**

Model	Text	Initialized from
Clinical BERT	All note types	BERT-Base
Discharge Summary BERT	Discharge Summary	BERT-Base
Bio+Clinical BERT	All note types	BioBERT
Bio+Discharge Summary BERT	Discharge Summary	BioBERT

Examine the performance

- 2 **Named-entity recognition** (NER) tasks
  - i2b2 2010
  - i2b2 2012
- 1 **Medical natural language inference** task
  - MedNLI
- 2 **De-identification** (de-ID) tasks
  - i2b2 2006
  - i2b2 2014

Dataset	Metric	Dim	# Sentences Train	Dev	Test
MedNLI	Accuracy	3	11232	1395	1422
i2b2 2006	Exact F1	17	44392	5547	18095
i2b2 2010	Exact F1	7	14504	1809	27624
i2b2 2012	Exact F1	13	6624	820	5664
i2b2 2014	Exact F1	43	45232	5648	32586

Results: Clinical BioBERT Win on 3 tasks

Clinically fine-tuned BioBERT shows improvements over **BioBERT or general BERT** on MedNLI, i2b2 2010, and i2b2 2012.

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

Results: Clinical BERT Lose on 2 tasks

Clinical BERT offers **no improvements** over **BioBERT or general BERT** on two de-ID tasks: i2b2 2006 and i2b2 2014

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

Results 3: Qualitative Embedding Comparisons

Clinical BERT retains **greater cohesion** around medical or clinical operations relevant terms than **BioBERT**

Model	Disease			Operations			Generic		
	Glucose	Seizure	Pneumonia	Transfer	Admitted	Discharge	Beach	Newspaper	Table
BioBERT	insulin exhaustion dioxide	episode appetite attack	vaccine infection plague	drainage division transplant	admission sinking hospital	admission wave sight	coast rock reef	news official industry	tables row dinner
Clinical	potassium sodium sugar	headache stroke agitation	consolidation tuberculosis infection	transferred admitted arrival	admission transferred admit	disposition transfer transferred	shore ocean land	publication organization publicity	scenario compilation technology

Nearest neighbors for 3 sentinel words for each of 3 categories.

Limitations

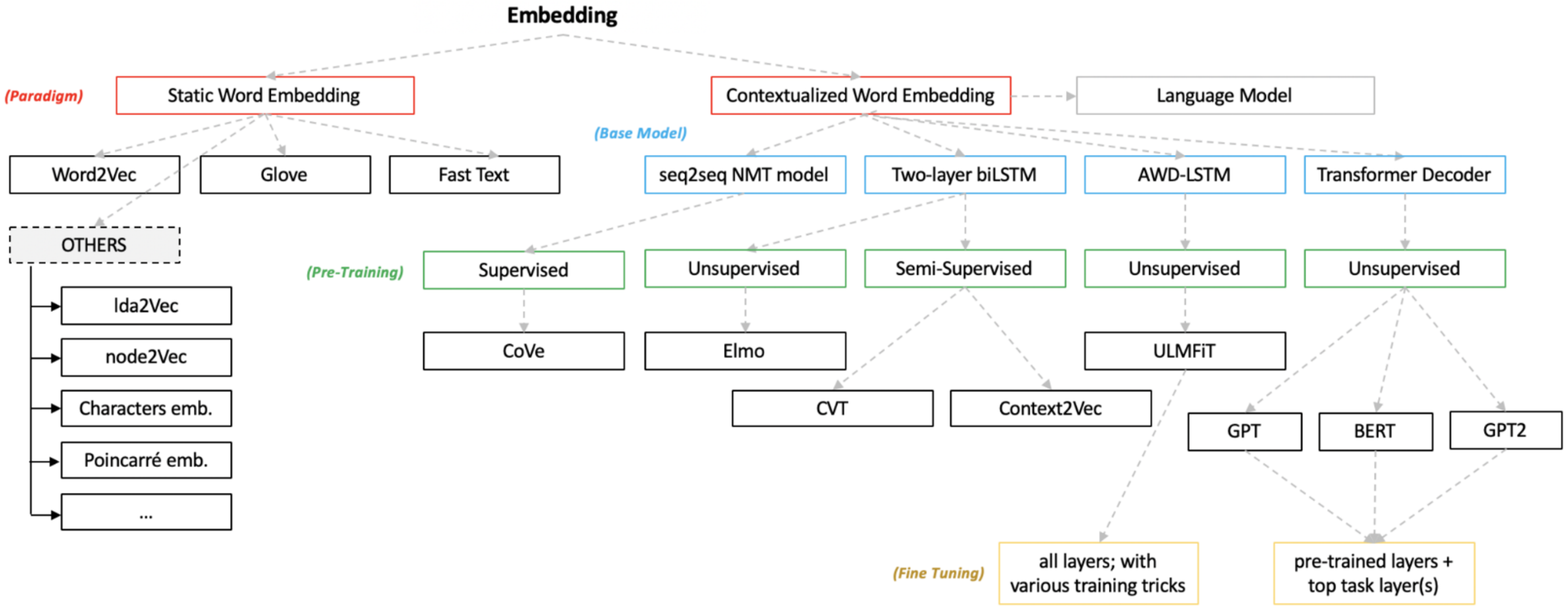
- Do not experiment with any more advanced model architectures
- MIMIC only contains notes from the intensive care unit of a single healthcare institution
- No improvements for two de-ID tasks

Summary

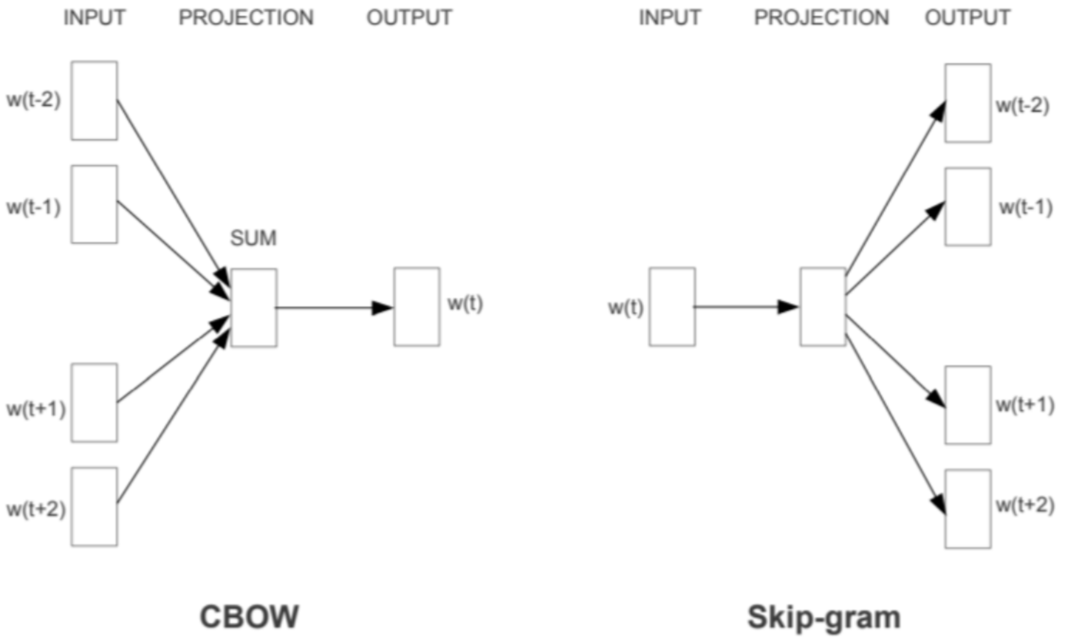
- Pretrain and **release** clinically oriented BERT models
- Clinical embeddings are **superior** to general or Bio- BERT specific embeddings for non de-ID tasks
- Clinical BERT shows **no improvements** to general or Bio- BERT on de-ID tasks

Extra Slides

More Embeddings



Word2Vec


INPUT
PROJECTION
OUTPUT
w(t-2)
w(t-1)
SUM
w(t)
w(t+1)
w(t+2)
CBOW
INPUT
PROJECTION
OUTPUT
w(t)
w(t-2)
w(t-1)
w(t+1)
w(t+2)
Skip-gram

MedNLI

#	Premise	Hypothesis	Label
1	ALT , AST , and lactate were elevated as noted above	patient has abnormal lfts	entailment
2	Chest x-ray showed mild congestive heart failure	The patient complains of cough	neutral
3	During hospitalization , patient became progressively more dyspnic requiring BiPAP and then a NRB	The patient is on room air	contradiction
4	She was not able to speak , but appeared to comprehend well	Patient had aphasia	entailment
5	T1DM : x 7yrs , h/o DKA x 6 attributed to poor medication compliance , last A1c [** 3-23 **] : 13.3 % 2	The patient maintains strict glucose control	contradiction
6	Had an ultimately negative esophagogastroduodenoscopy and colonoscopy	Patient has no pain	neutral
7	Aorta is mildly tortuous and calcified .	the aorta is normal	contradiction