LREC 2022

Re-train or Train from Scratch? Comparing Pre-training Strategies of BERT in the Medical Domain

Authors: Hicham El Boukkouri [1], Olivier Ferret [2], Thomas Lavergne [1], Pierre Zweigenbaum [1]

[1] Université Paris-Saclay, CNRS, LISN, Orsay, France || {elboukkouri, lavergne, pz}@limsi.fr
[2] Université Paris-Saclay, CEA List, F-91120, Palaiseau, France || olivier.ferret@cea.fr

Introduction

General Context & Motivation

Transfer Learning

Main Idea: knowledge acquired on a first task can be successfully transferred to other tasks, improving final performance

- Some Important Milestones
  - ULMFiT *(Howard et Ruder, 2018)*
    ⇒ re-adapt a *Language Model* to multiple NLP tasks
  - ELMo *(Peters et al., 2018)*
    ⇒ *Contextualized Embeddings* from a *Language Model* + feature extraction
  - BERT *(Devlin et al., 2018)*
    ⇒ end-to-end fine-tuning, Transformer-based, improved performance

Specialized Domains

Technical domains (e.g. medical domain) that often have specific vocabularies, writing styles, etc. Require specialized models.

- BERT in Specialized Domains
  - Off-the-shelf pre-trained models (often, general-domain)
  - Adapted by re-training on specialized corpora (e.g. biomedical literature)
  - Keep the original vocabulary (often, general-domain)

⇒ Possible mismatch with target domain ⇐

Methodology

Objective: analyze the impact of using general-domain WordPieces w/ BERT in specialized domains

- Proposed Approach
  - Given BERT's original general-domain WordPieces
  - Given a specialized corpus in the target domain

    ⇒ *Learn specialized WordPieces*
    → *Compare against BERT's original WordPieces*

    ⇒ *Train BERT from scratch using the specialized WordPieces*
    → *Compare against a model re-trained on specialized texts*

BERT & the WordPieces

Effect on the Tokenization

How Does BERT Tokenize?

BERT's tokenization system (simple + WordPiece tokenization)

- A simple tokenization (e.g. handling punctuation)

  Hello, my name's Hicham. → Hello | , | my | name | ' | s | Hicham | .

- A WordPiece tokenization (i.e. handling Out-Of-Vocabulary tokens)

  Hello, my name's **Hicham**. → Hello | , | my | name | ' | s | **Hic** | **##ham** | .

Experimental Setup

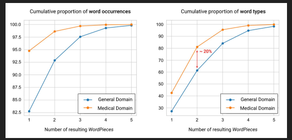
- We use the default general-domain WordPieces from BERT (base-uncased)
  → training corpora: English Wikipedia + BooksCorpus
- We also train a set of medical-domain WordPieces
  → training corpora: MIMIC-III + PMC OA abstracts

Domain	Corpora	# of documents	# of words
Medical	MIMIC-III	4.17 million	0.5 billion
	PMC OA abstracts	4.65 million	0.5 billion

Medical corpora used for learning WordPiece units

Results: Number of Splits

- We run WordPiece tokenization on specialized texts…



Results: Quality of Splits

- …as well as on a few specialized terms from the medical domain

Token	Medical Tokenization	General Tokenization
paracetamol	[paracetamol]	[para, ce, tam, ol]
choledocholithiasis	[choledoch, olithiasis]	[cho, led, och, oli, thi, asi, s]
borborygmi	[bor, bor, yg, mi]	[bo, rb, ory, gm, i]

Tokenization of specialized terms using WordPiece vocabularies from different domains

⇒ The medical vocabulary splits these tokens into fewer WordPieces

⇒The resulting WordPieces also seem to be more meaningful*

BERT & the WordPieces

Effect on Downstream Performance

Experimental Setup

- Comparison of two different scenarios
  - Training a model from scratch using specialized WordPieces
  - Re-training a model from a general-domain checkpoint
    ⇒ i.e. inheriting the general WordPiece vocabulary

NOTE: for a fair comparison of all models, we train all of them ourselves (including general BERT) under the exact same conditions

Experimental Setup

- We use the following corpora for pretraining

Domain	Corpora	# of documents	# of words
General	WIKIPEDIA (EN)	11.9 million	2.14 billion
	OPENWEBTEXT	3.15 million	1.28 billion
Medical	MIMIC-III	4.17 million	0.5 billion
	PMC OA abstracts	4.65 million	0.5 billion

Statistics for the used pre-training corpora

⇒ For general BERT, we replace *BooksCorpus* with *OpenWebText*

Experimental Setup

- We train the following configurations

Re-training on a Specialized Corpus

- (V = **general**, C_1 = **general**, C_2 = ∅)
  Trained on a general corpus using a general vocabulary (equiv. BERT)
- (V = **general**, C_1 = **general**, C_2 = **medical**)
  Same as before, plus re-training on a medical corpus

Training from Scratch

- (V = **medical**, C_1 = **medical**, C_2 = ∅)
  Trained from scratch on a medical corpus using a medical vocabulary
- (V = **medical**, C_1 = **medical**, C_2 = **medical**)
  Same as before, plus re-training on a medical corpus

Downstream Tasks

- And evaluate on various downstream tasks, namely

i2b2 2010 Clinical Concept Extraction (*sequence labeling*)

The patient had headache that was relieved only with oxycodone. A CT scan of the head showed microvascular ischemic changes. A followup MRI which also showed similar changes. This was most likely due to her multiple myeloma with hyperviscosity.

Clinical Concept Types

Problem Treatment Test

Number of examples: ~28,000 training / ~45,000 test

Downstream Tasks

- And evaluate on various downstream tasks, namely

MedNLI Clinical Natural Language Inference (*pair classification*)

Sentence 1 : Labs were notable for Cr 1.7 (baseline 0.5 per old records) and lactate 2.4.
Sentence 2 : Patient has normal Cr.
Contradiction

Sentence 1 : Nystagmus and twiching of R arm was noted.
Sentence 2 : The patient had abnormal neuro exam.
Entailment

Number of examples: ~12,000 training / ~1,000 test

Downstream Tasks

- And evaluate on various downstream tasks, namely

DDI & ChemProt Biomedical Relation Extraction (*text classification*)

Chemprot

Mitiglinide (**@CHEMICAL$**), a new anti-diabetic drug, is thought to stimulate **@GENE$** secretion by closing the ATP-sensitive K+ (K(ATP)) channels in pancreatic beta-cells.
Active (CPR:3)

DDI

@DRUG$** should be administered with caution to patients receiving **@DRUG$ (disulfiram, Wyeth-Ayerst Laboratories).
Conseil (DDI-advise)

Number of examples: | DDI ~4,000/1,000 | ChemProt ~6,500/3,500 |

Evaluation Method

- We use the following approach
  - Multiple seeds: 10 different seeds for each fine-tuning
  - Model selection: epoch w/ best performance on a on dev. set
  - Final scores: compute mean +/- std using the 10 different seeds
- External baselines
  - BERT: original general-domain model (bert-base-uncased)
  - BlueBERT: a medical version of BERT (clinical + biomedical)

Experiment Results

- **Pre-training Sanity Check**
  We are able to reach relatively similar performance with our own general BERT
- **General BERT < Medical BERT**
  The general-domain versions of BERT (G, G, ∅) & original BERT perform worse than versions trained on medical corpora
- **Is Training from Scratch Really Better?**
  While the fully medical model (M, M, M) is the overall best configuration, the difference with the retrained model (G, G, M) is not large enough to warrant the additional training & computational cost

V	C_1	C_2	I2B2	MEDNLI	CHEMPROT	DDI
G	G	∅	85.66 ± 0.18	77.31 ± 0.71	67.47 ± 0.99	75.81 ± 1.02
G	G	M	89.00 ± 0.17	**84.91** ± 0.46	72.29 ± 0.58	78.82 ± 1.11
M	M	∅	88.80 ± 0.10	83.54 ± 0.43	71.30 ± 0.51	79.40 ± 1.15
M	M	M	**89.20** ± 0.20	84.32 ± 0.73	**72.97** ± 0.46	**80.11** ± 0.79
Baselines						
BERT			86.42 ± 0.31	77.85 ± 0.63	69.22 ± 0.56	77.89 ± 0.92
BLUEBERT			88.70 ± 0.21	84.53 ± 0.76	68.35 ± 0.61	77.89 ± 0.65

20/25 BERT & the WordPieces Effect on Downstream Performance

Experiment Results

Pre-training Sanity Check
We are able to reach relatively similar performance with our own general BERT
General BERT < Medical BERT
The general-domain versions of BERT (G, G, ∅) & original BERT perform worse than versions trained on medical corpora
Is Training from Scratch Really Better?
While the fully medical model (M, M, M) is the overall best configuration, the difference with the retrained model (G, G, M) is not large enough to warrant the additional training & computational cost

21/25 BERT & the WordPieces Effect on Downstream Performance

Experiment Results

Pre-training Sanity Check
We are able to reach relatively similar performance with our own general BERT
General BERT < Medical BERT
The general-domain versions of BERT (G, G, ∅) & original BERT perform worse than versions trained on medical corpora
Is Training from Scratch Really Better?
While the fully medical model (M, M, M) is the overall best configuration, the difference with the retrained model (G, G, M) is not large enough to warrant the additional training & computational cost

22/25 BERT & the WordPieces Effect on Downstream Performance

Experiment Results

Pre-training Sanity Check
We are able to reach relatively similar performance with our own general BERT
General BERT < Medical BERT
The general-domain versions of BERT (G, G, ∅) & original BERT perform worse than versions trained on medical corpora
Is Training from Scratch Really Better?
While the fully medical model (M, M, M) is the overall best configuration, the difference with the retrained model (G, G, M) is not large enough to warrant the additional training & computational cost

23/25 BERT & the WordPieces Effect on Downstream Performance

Experiment Results

Pre-training Sanity Check
We are able to reach relatively similar performance with our own general BERT
General BERT < Medical BERT
The general-domain versions of BERT (G, G, ∅) & original BERT perform worse than versions trained on medical corpora
Is Training from Scratch Really Better?
While the fully medical model (M, M, M) is the overall best configuration, the difference with the retrained model (G, G, M) is not large enough to warrant the additional training & computational cost

Conclusion

Main Takeaways

Findings & Limitations

- Mismatch when using General WordPieces in Specialized Domains
  ⇒ unknown words are split into many subwords
  ⇒ resulting WordPieces are less meaningful
- Training from scratch using a specialized vocabulary/corpus
  ⇒ overall better than re-training from general BERT
  ⇒ may not be worth the additional pre-training cost
- Limitations: experiments conducted for a single domain & language and missing experiments (V=*medical*, C1=*general*, C2= •)