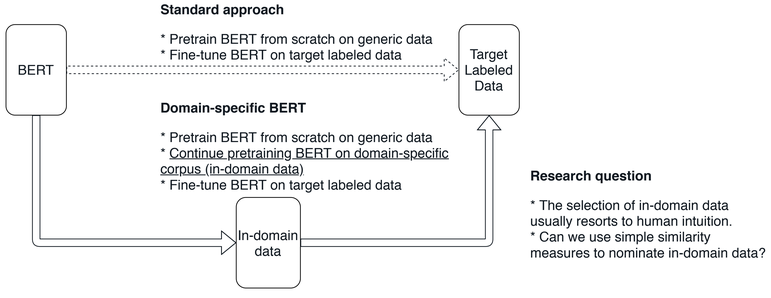
Cost-effective Selection of Pretraining Data: A Case Study of Pretraining BERT on Social Media

Xiang Dai[♣♠] and Sarvnaz Karimi[♣] and Ben Hachey[♦] and Cecile Paris[♣]

♣ CSIRO Data61 ♠ University of Sydney ♦ Harrison AI

Motivation

- Recent studies on domain-specific BERT models show that, when **in-domain** data is used for pretraining, target task performance can be improved.
- However, the selection of in-domain data usually resorts to human intuition.



We aim to use simple similarity measures to nominate in-domain data, so we

- conduct a case study of pretraining BERT on social media text which has very different tenor from existing domain-specific BERT models.
- release two pretrained BERT models trained on tweets and forum text.
- investigate the correlation of source-target similarity and task accuracy using different domain-specific BERT models.

Resources

- **Paper**: Dai, Xiang, Karimi, Sarvnaz, Hachey, Ben, Paris, Cecile: Cost-effective Selection of Pretraining Data: A Case Study of Pretraining BERT on Social Media. In: Findings of the Association for Computational Linguistics: EMNLP 2020; Online: 1675–1681.
- **Models**: https://bit.ly/35RpTf0

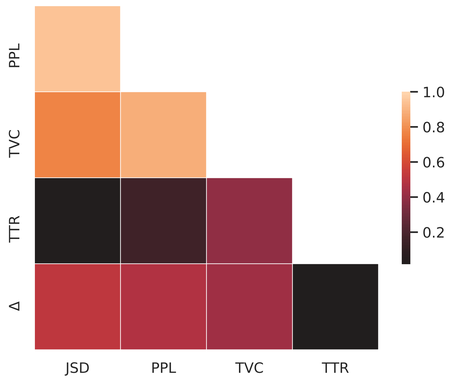
Effectiveness of our Pretrained BERT Models

Target Text type	Corpus	BERT	Bio	Clinical	Sci	Twitter	Forum
Tweets	Airline (*C*)	80.5± 0.3	79.0± 0.5	78.8± 0.8	78.8± 0.9	80.8± 0.6	**81.6± 0.5**
	BTC (*N*)	78.0± 0.5	75.2± 0.3	76.9± 0.5	77.4± 0.4	**79.0± 0.5**	77.0± 0.4
	SMM4H-18 task3 (*C*)	76.5± 0.9	75.4± 1.1	75.6± 0.7	75.4± 1.0	77.0± 1.0	**77.2± 1.3**
	SMM4H-18 task4 (*C*)	89.4± 0.5	87.7± 0.4	88.1± 0.8	88.7± 0.8	90.3± 0.3	**91.1± 0.6**
Forum	CADEC (*N*)	71.9± 0.6	72.1± 0.6	72.1± 0.8	**73.2± 0.4**	72.1± 1.0	72.9± 0.6
	SemEval-14 laptop (*N*)	81.1± 0.8	79.3± 0.3	78.5± 0.4	**81.6± 1.1**	81.3± 0.6	81.4± 1.1
	SemEval-14 restaurant (*N*)	87.5± 0.6	84.9± 0.3	85.5± 0.7	86.7± 0.5	87.4± 0.7	**89.3± 0.5**
	SST-2 (*C*)	92.4± 0.2	91.1± 0.5	90.4± 0.3	91.4± 0.4	92.3± 0.4	**93.4± 0.4**
Non-social media	EBM (*N*)	41.5± 0.5	42.1± 0.2	41.1± 0.5	**42.4± 0.7**	40.5± 0.5	41.5± 0.5
	i2b2-10 (*N*)	85.8± 0.1	**87.4± 0.2**	**87.4± 0.1**	87.3± 0.2	84.8± 0.2	85.2± 0.1
	JNLPBA (*N*)	72.5± 0.3	**74.2± 0.2**	71.9± 0.1	73.6± 0.3	72.2± 0.2	72.5± 0.2
	Paper Field (*C*)	74.5± 0.1	74.3± 0.1	73.3± 0.1	**75.1± 0.1**	74.1± 0.1	73.3± 0.2

- Effectiveness of different BERT models, evaluated on downstream tasks. *C*: Classification task, for which we report macro-F1; *N*: NER task, for which we report span-level micro-F1. underline: the best result is significantly better than the second best result (paired student's t-test, p: 0.05).

Similarity measures can be used to nominate in-domain pretraining data

- Three measures of the similarity between source and target data
  - *Language model perplexity (PPL)*: construct Kneser-Ney smoothed 3-gram models on source data and use the perplexity of target data relative to these language models as the similarity
  - *Jensen-Shannon divergence (JSD)*: measure the probability of each term (up to 3-gram) in source and target data, separately. Then use the Jensen-Shannon divergence between these two probability distributions as the similarity
  - *Target vocabulary covered (TVC)*: measures the percentage of the target vocabulary present in the source data, where only content words (nouns, verbs, adjectives) are counted



- Diversity measure: type token ratio (*TTR*, $\frac{\#\text{unique tokens}}{\#\text{tokens}}$), that measures the lexical diversity of the source data.
- Employ the Pearson correlation analysis to find out the relationships between improvements due to domain-specific BERT models and similarity between source and target data.
- **Results show that JSD has the strongest correlation (0.519) with the improvement due to domain-specific models, while the other two measures also have modest correlation (**0.481 **for PPL and** 0.436 **for TVC).**