Cross-Domain Language Modeling: An Empirical Investigation

Vincent Nguyen[1,2]**, Sarvnaz Karimi**[2]**, Maciek Rybinski**[2]**,Zhenchang Xing**[1]

Australian National University[1] and CSIRO's Data61[2]

Vincent.Nguyen@anu.edu.au

https://ngu.vin

Background

Motivation

- Language models are confined to fixed-sized vocabulary at training-time.
- Vocabulary sets are often incompatible with downstream tasks, especially considering a cross-domain scenario, due to sub-word overlap.

Datasets

- We consider the cross-domain scenario and use a proxy pretraining dataset for the cross-domain between the GENERAL and BIOMEDICAL domains.
- For task evaluation, we evaluate on two benchmark datsets: *GLUE* and *BLUE*.

Tasks

Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy

GLUE Benchmark, picture from https://mccormickml.com/2019/11/05/GLUE/

Corpus	Train	Dev	Test	Task	Metrics	Domain	Avg sent len
BIOSSES, sentence pairs	64	16	20	Sentence similarity	Pearson	Biomedical	22.9
DDI, relations	2937	1004	979	Relation extraction	micro F1	Biomedical	41.7
ChemProt, relations	4154	2416	3458	Relation extraction	micro F1	Biomedical	34.3
i2b2 2010, relations	3110	11	6293	Relation extraction	F1	Clinical	24.8
HoC, documents	1108	157	315	Document classification	F1	Biomedical	25.3
MedNLI, pairs	11232	1395	1422	Inference	accuracy	Clinical	11.9

BLUE Benchmark statistics

Methodology

Main Problems

- Contemporary transformer-based language models use a one-size-fits all vocabulary resulting in morpheme conflation at the sub-word level when used cross-domain.
- **Proposal:** Expand the vocabulary size such that such conflation is minimal.
- **Compromise:** Pretraining must be repeated and vocabulary is a hyper-parameter that must be tuned to a target cross-domain.

Experiments

We vary (1) the vocabulary size (5000—100,000) and (2) pretraining data in a transformer to determine the effect of cross-domain overlap on downstream benchmark tasks.

$\|V\|$	Jaccard Similarity	Num. Overlaps	% Vocab Used	Num. Tokens used in GLUE tasks	Num. Tokens used in BLUE tasks
5000	94.6	4708	99.5	4970	4713
10000	87.8	8733	99.5	9893	8786
20000	73.8	14609	99.0	19418	14989
30000	62.8	18490	98.2	28457	19498
40000	54.6	21193	97.1	37057	22980
50000	48.2	23083	95.8	45239	25726
60000	43.0	24359	94.4	53109	27888
70000	38.9	25226	92.7	60545	29549
80000	35.6	25858	90.8	67563	30961
90000	32.8	26287	89.1	74369	32118
100000	30.4	26593	87.3	80842	33095

Jaccard Index and Overlap Proportion.

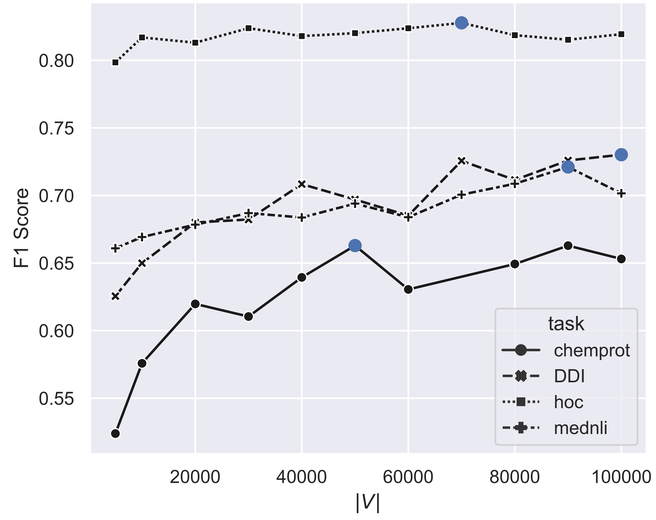
Results

Benchmark	Pretraining Corpora	Effectiveness
BLUE (F1)	Wiki	0.6973
	PubMed	0.6706
	PubMed+Wiki	**0.7186**†
GLUE (Acc)	Wiki	0.7090
	PubMed	0.7060
	PubMed+Wiki	0.6906

Average downstream (F1 for BLUE, Acc. for GLUE) benchmark performance

Pretrained Language Model performance of large (L) vocabulary sizes (greater than 50,000) and smaller (S) vocabulary sizes on downstream task.

Domain	Task	S	L	L-S
General Domain	CoLA	14.3	14.7	+0.40
	MNLI	69.5	71.1†	+1.60
	MRPC	79.4	79.6	+0.20
	QNLI	73.6	79.3	+5.70
	QQP	79.7	80.6	+0.90
	RTE	53.8	53.5	-0.50
	SST-2	81.5	84.0†	+2.50
	STS-B	36.7	28.8	-7.90
	WNLI	46.8	47.5	+0.70
Biomedical	biosses	13.6	19.0	+5.40
	chemprot	59.4	65.2	+5.80
	DDI	66.9	71.2†	+4.30
	HoC	81.4	82.1	+0.70
	MedNLI	67.6	70.2†	+2.60



Evaluation of biomedical tasks against varied $|V|$.

Key Findings

- The biomedical domain benefits from combined pretraining data while the general domain sees a slight performance reduction.
- Vocabulary size is a hyper-parameter that should be tuned carefully in a cross-domain scenario.
- Biomedical (specialised) task benefits more from vocabulary separation than more general tasks.

Future Work

- It is difficult to separate performance from sub-word overlap reduction and an increase in model parameters. A fixed model size with a variation in cross-doamin sub-word can further verify these results.

Acknowledgments

This research is supported by the Australian Research Training Program and the CSIRO Postgraduate Scholarship and CSIRO's Future Science platform for Precision Health.