Using Similarity Measures to Select Pretraining Data for NER

Xiang Dai, Sarvnaz Karimi, Ben Hachey, Cecile Paris

2019-March-22

Knowledge transfer

WHAT

Word representation, entity representation ...

HOW

Pretrain+transfer, multi-task ...

WHY

More data, good performance ...

Knowledge transfer

WHAT

Word representation, entity representation ...

HOW

Pretrain+transfer, multi-task ...

WHY

More data, good performance ...

WHERE

...

Given a NER dataset, nominate the most suitable source data to pretrain word vectors or language models (LMs)

- The more similar the source data is to the target data, the better the pretrained models are, all other aspects (such as source data size) being equal.
- However, what is similar source data is still left to intuition.

Pretraining word representations

- Word vectors and Language Models pretrained on a large amount of unlabelled data can dramatically improve various NLP tasks.
  - Word2Vec, GloVe, ELMo, BERT.
- Most previous work has focused on different pretraining architectures, objectives and adaptation strategies.
  - LSTM, Transformer.
  - Autoencoder, LM, Skip-thought, NLI.
  - Fine-tuning, feature extraction.
- Our work focuses on the impact of similarity between pretraining data and target task data.

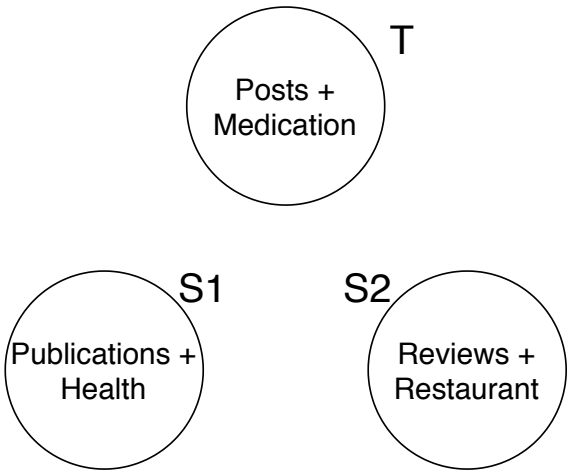
Text Similarity

- Word-level
  - Functional similarity (school versus college)
  - Associative similarity (school versus teacher)
- Sentence-level
  - Paraphrase identification.
  - Natural language inference.
  - Word embedding distance, stylistic variation, etc.
- We extend the study of similarity to corpus-level, and focus on its implication on unsupervised pretraining.

Selecting strategies in the literature

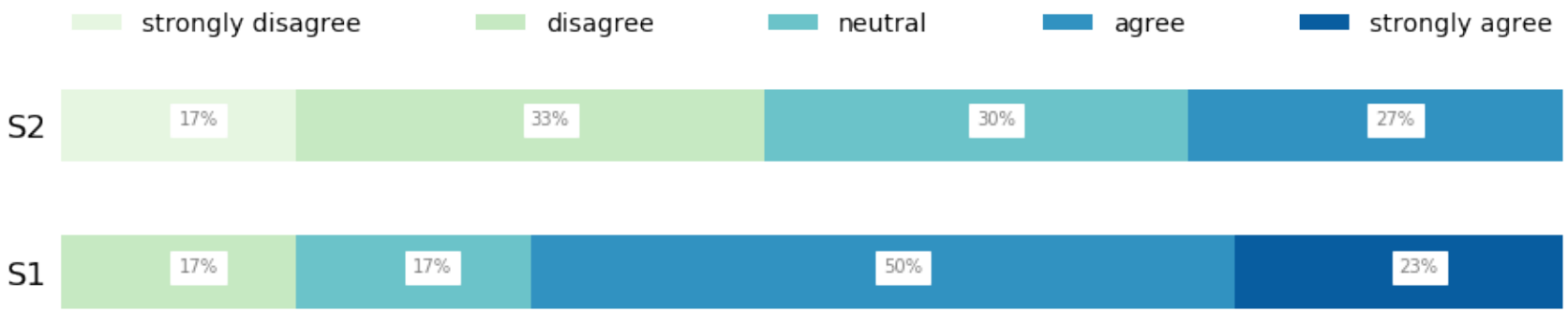
- Collecting a large amount of generic data, e.g., web crawl.
- Selecting data from a similar field (the subject matter of the content being discussed), e.g., biology.
- Selecting data from a similar tenor (the participants in the discourse, their relationships to each other, and their purposes), e.g., Twitter, or online forums.

A survey of NLPer intuition



- Corpora description
  - T: Online forum posts about medications.
  - S1: Research papers about biology and health.
  - S2: Online reviews about restaurants, hotels, barbers, mechanics, etc.
- Questions
  - Do you think unsupervised pretraining models on S1 (S2) would be useful for supervised named entity learning on T?

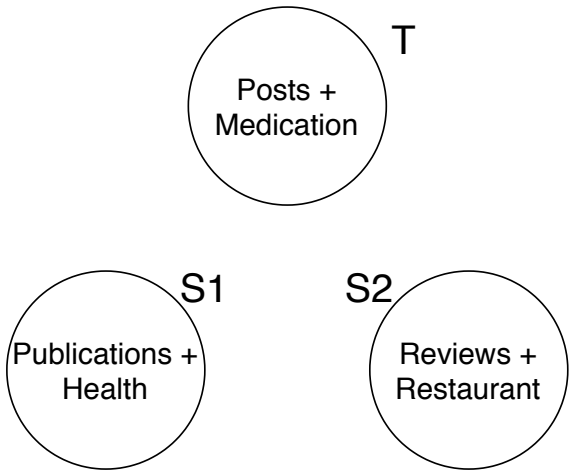
The human intuition is to prioritise field over tenor



- T: Online forum posts about medications.
- S1: Research papers about biology and health.
- S2: Online reviews about restaurants, hotels, barbers, mechanics, etc.
- 30 responses received.
- A Wilcoxon signed-rank test indicates that scores are significantly higher for S1 than for S2 ($Z = 43.0, p < 0.001$).

Our explanations of the human intuition

- Researchers who select pretraining data from a similar field believe that, data with similar field tend to share similar vocabulary.
  - Some additional comments from participants and reviewers: depends on what NE you want to detect; if the term doesn't exist in the pretraining data, it cannot help anyway.
- Those who select pretraining data from a similar tenor believe that tenor may impact the writing style of text.



Three measures based on these intuitions

- Vocabulary intersection rate: $VIR(D_S, D_T) = \frac{|V_{D_S} \cap V_{D_T}|}{|V_{D_T}|}$
  - A variant considering only content words (nouns, verbs, adjectives)
- Language model perplexity: Kneser-Ney smoothed 5-gram models, trained on the source data, then evaluated on the target data.
- Word vector variance
  - train word vectors on the source data, denoted as $WS \in \mathbb{R}^{|V_S| \times d}$
  - use WS as initial weights of a new model, continue training on the target data, denoted as WT.
  - $WVV(D_S, D_T) = \frac{1}{|V_S|} \frac{1}{d} \sum_i^{|V_S|} \sum_j^d (WS_i^j - WT_i^j)^2$

Some unsuccessful attempts

- Adversarial autoencoder
  - Train the model on source data, a classifier is trained to predict whether the input sentence is a real sentence from the dataset or a fake sentence from the generator.
  - Use the trained classifier to assign probabilities to target sentences, a proxy to measure similarity between source and target data.
- Universal sentence encoder
  - Use an off-the-shelf sentence encoder to generate sentence representation of the dataset: $\mathbb{R}^{m \times d}$.
  - Calculate the distance of two matrices, which represent source data and target data, respectively.

Datasets: Source datasets

1BWB	News Crawl data.
MIMIC	A clinical database.
PubMed	Biomedical literature covering the fields of biomedical and health.
Wiki	WikiText-103, released by [Merity et al., 2016] and selected based on the writing quality.
Yelp	Online reviews about local businesses.

- All source datasets are samples to similar size (500M).

Datasets: Target NER datasets

CADEC	Posts taken from AskaPatient, where consumers can discuss their experiences with medications.
CoNLL2003	Newswire from the Reuters RCV1 corpus.
CRAFT	Full-length, open-access biomedical journal articles.
JNLPBA	Abstract of biomedical journal articles.
ScienceIE	Computer Science, Material Sciences and Physics journal articles.
Wetlab	Protocols written by researchers on conducting biology and chemistry experiments.

PPL and WVV capture tenor more than field

Target	Source	Similarity			
		PPL	WVV	VIR (%)	VcIR (%)
CADEC	None	–	–	–	–
	1BWB	307.4	1.137	**81.73**	**82.94**
	MIMIC	1007.0	1.134	78.19	81.69
	PubMed	927.4	1.195	78.81	79.79
	Wiki	519.8	1.196	79.74	76.71
	Yelp	**291.1**	**1.104**	80.76	82.28
CoNLL2003	None	–	–	–	–
	1BWB	**480.6**	**1.020**	**75.64**	**87.35**
	MIMIC	2945.0	1.542	34.47	39.55
	PubMed	3143.1	1.356	53.29	68.41
	Wiki	650.4	1.159	66.21	80.87
	Yelp	2025.5	1.399	53.92	68.95
CRAFT	None	–	–	–	–
	1BWB	1328.1	2.073	59.07	62.98
	MIMIC	2427.5	2.390	48.73	50.03
	PubMed	**360.3**	**1.838**	**76.29**	**80.69**
	Wiki	974.7	2.075	63.66	63.12
	Yelp	2085.7	2.187	48.01	50.85
JNLPBA	None	–	–	–	–
	1BWB	1190.8	2.000	39.90	53.54
	MIMIC	2533.4	2.172	36.95	50.04
	PubMed	**205.9**	**1.597**	**58.87**	**80.17**
	Wiki	717.9	2.036	42.34	53.05
	Yelp	2134.4	2.155	30.78	41.41
ScienceIE	None	–	–	–	–
	1BWB	884.6	1.197	71.50	76.78
	MIMIC	2706.7	1.461	54.29	59.34
	PubMed	**345.6**	**1.037**	**83.25**	**87.01**
	Wiki	684.2	1.127	76.99	78.01
	Yelp	1562.2	1.347	62.32	66.42
WetLab	None	–	–	–	–
	1BWB	1526.0	2.167	59.67	61.47
	MIMIC	3046.1	2.393	53.83	55.31
	PubMed	**1104.7**	**2.078**	**71.39**	**74.46**
	Wiki	1617.8	2.158	61.02	60.31
	Yelp	1784.5	2.240	54.16	54.96

- PPL and WWV: $\Downarrow$
- VIR: $\Uparrow$
- Yelp is more similar to CADEC than PubMed and MIMIC from both PPL and WVV perspectives.
- The language model trained on PubMed achieves lower PPL on ScienceIE.
- All sources are measured against WetLab with relatively high PPL and WVV values.

Different similarity measures can reach a consensus

- Different measures can lead to almost the same answer regarding which source is the most similar one to a given target.
- Given a target and two sources, do similarity measures make the same conclusion as to which source is more similar?
  - 60 binary comparisons. For example, given WetLab, is 1BWB a more similar source than Wiki? PPL shows that 1BWB is more similar, while WVV gives an opposite answer.
  - Fleiss's kappa, 0.733, shows a high agreement between conclusions inferred using different measures.

Can these similarity measures predict the effectiveness of pretrained models for NER tasks?

- BiLSTM-CRF: supervised model for the target NER task.
- Baseline model: word embedding weights are randomly initialized, no pretrained LMs are used.
- In different experiments, we replace pretrained models and observe the improvement of $F1$ score on target task.

Experimental setup for NER

- Word2vec with its default hyper-parameter setting [Mikolov et al., 2013].
- Replace the word embedding weights initialized by word vectors pretrained on different sources, then make these weights trained jointly with other model parameters.
- Language models pretrained using the architecture proposed by [Jozefowicz et al., 2016].
- The contextualized representation of each word in the target data set is generated using the outputs of the pretrained LMs, then injected to the input of the second BiLSTM layer of the supervised model [Peters et al., 2018].

Our proposed similarity measures are good predictors of the pretrained models for NER

Target	Source	Similarity				NER F_1 Score			
						Pretrained word vectors		Pretrained LMs	
		PPL	WVV	VIR (%)	VcIR (%)	F_1 score	Δ	F_1 score	Δ
CADEC	None	–	–	–	–	66.14 (± 0.53)	–	66.14 (± 0.53)	–
	1BWB	307.4	1.137	**81.73**	**82.94**	69.44 (± 0.52)	3.30	70.08 (± 0.43)	3.94
	MIMIC	1007.0	1.134	78.19	81.69	69.65 (± 0.43)	3.51	70.11 (± 0.48)	3.97
	PubMed	927.4	1.195	78.81	79.79	69.84 (± 0.55)	3.70	70.15 (± 0.50)	4.01
	Wiki	519.8	1.196	79.74	76.71	69.62 (± 0.15)	3.48	69.32 (± 0.65)	3.18
	Yelp	**291.1**	**1.104**	80.76	82.28	**70.27 (± 0.34)**	**4.13**	**70.46 (± 0.52)**	**4.32**
CoNLL2003	None	–	–	–	–	82.08 (± 0.38)	–	82.08 (± 0.38)	–
	1BWB	**480.6**	**1.020**	**75.64**	**87.35**	**86.36 (± 0.29)**	**4.28**	**89.78 (± 0.12)**	**7.70**
	MIMIC	2945.0	1.542	34.47	39.55	84.94 (± 0.35)	2.86	83.68 (± 0.30)	1.60
	PubMed	3143.1	1.356	53.29	68.41	85.56 (± 0.46)	3.48	84.15 (± 0.22)	2.07
	Wiki	650.4	1.159	66.21	80.87	86.32 (± 0.28)	4.24	89.11 (± 0.23)	7.03
	Yelp	2025.5	1.399	53.92	68.95	85.58 (± 0.26)	3.50	85.19 (± 0.38)	3.11
CRAFT	None	–	–	–	–	69.17 (± 0.64)	–	69.17 (± 0.64)	–
	1BWB	1328.1	2.073	59.07	62.98	73.97 (± 0.06)	4.80	71.23 (± 0.81)	2.06
	MIMIC	2427.5	2.390	48.73	50.03	73.01 (± 0.22)	3.84	71.90 (± 0.26)	2.73
	PubMed	**360.3**	**1.838**	**76.29**	**80.69**	**75.45 (± 0.28)**	**6.28**	**75.45 (± 0.09)**	**6.28**
	Wiki	974.7	2.075	63.66	63.12	74.07 (± 0.40)	4.90	69.75 (± 0.09)	0.58
	Yelp	2085.7	2.187	48.01	50.85	72.48 (± 0.13)	3.31	72.75 (± 0.26)	3.58
JNLPBA	None	–	–	–	–	70.45 (± 0.21)	–	70.45 (± 0.21)	–
	1BWB	1190.8	2.000	39.90	53.54	72.39 (± 0.23)	1.94	72.54 (± 0.34)	2.09
	MIMIC	2533.4	2.172	36.95	50.04	**73.24 (± 0.29)**	**2.79**	71.76 (± 0.13)	1.31
	PubMed	**205.9**	**1.597**	**58.87**	**80.17**	72.77 (± 0.65)	2.32	**74.29 (± 0.40)**	**3.84**
	Wiki	717.9	2.036	42.34	53.05	72.77 (± 0.27)	2.32	72.42 (± 0.23)	1.97
	Yelp	2134.4	2.155	30.78	41.41	72.53 (± 0.18)	2.08	72.51 (± 0.21)	2.06
ScienceIE	None	–	–	–	–	26.85 (± 0.17)	–	26.85 (± 0.17)	–
	1BWB	884.6	1.197	71.50	76.78	34.40 (± 0.50)	7.55	38.10 (± 0.31)	11.25
	MIMIC	2706.7	1.461	54.29	59.34	31.23 (± 0.15)	4.38	35.27 (± 0.43)	8.42
	PubMed	**345.6**	**1.037**	**83.25**	**87.01**	**37.91 (± 0.12)**	**11.06**	**42.07 (± 0.03)**	**15.22**
	Wiki	684.2	1.127	76.99	78.01	36.15 (± 0.11)	9.30	40.39 (± 0.05)	13.54
	Yelp	1562.2	1.347	62.32	66.42	33.92 (± 0.14)	7.07	36.05 (± 0.02)	9.20
WetLab	None	–	–	–	–	76.91 (± 0.10)	–	76.91 (± 0.10)	–
	1BWB	1526.0	2.167	59.67	61.47	78.66 (± 0.35)	1.75	78.94 (± 0.05)	2.26
	MIMIC	3046.1	2.393	53.83	55.31	78.68 (± 0.14)	1.13	78.65 (± 0.13)	1.74
	PubMed	**1104.7**	**2.078**	**71.39**	**74.46**	**78.93 (± 0.28)**	**2.02**	**79.62 (± 0.07)**	**2.71**
	Wiki	1617.8	2.158	61.02	60.31	78.45 (± 0.20)	1.54	79.05 (± 0.21)	2.14
	Yelp	1784.5	2.240	54.16	54.96	78.48 (± 0.15)	1.57	79.04 (± 0.19)	2.13

LM performance is more predictable using our proposed measures

- Pearson correlation analysis: the relationships between improvement due to pretrained models and VIR, VcIR, PPL and WVV.
- 1: positive linear correlation; 0: no linear correlation; and -1: negative linear correlation.
- VcIR is the most informative factor in predicting the effectiveness of pretrained word vectors.
- LM performance has a stronger correlation with similarity measures than the one of word vectors.

	Word vectors	**LMs**
VIR	0.454	0.666
VcIR	0.469	0.739
PPL	-0.398	-0.618
WVV	-0.406	-0.747

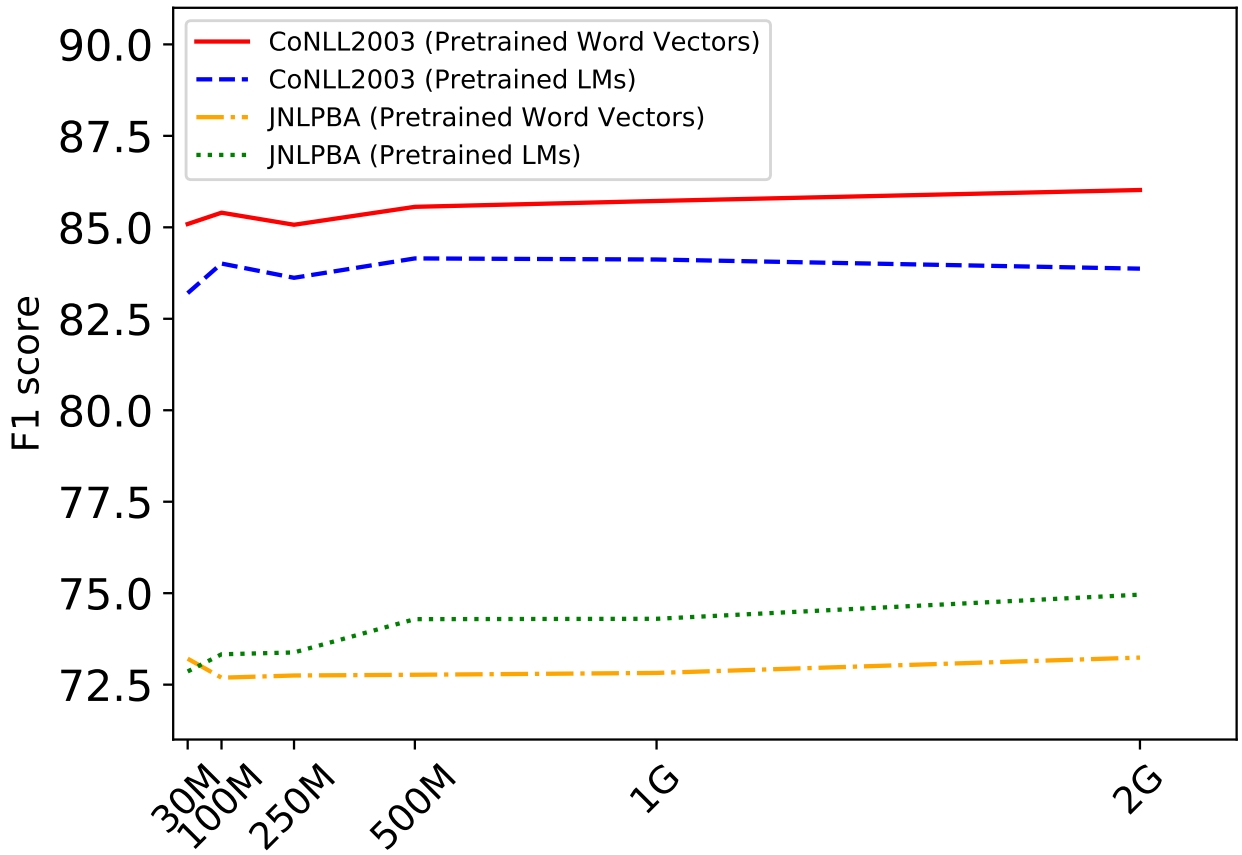
Small similar pretraining data can outperform large generic data

- Publicly available word vectors (GloVe) and LMs (ELMo) pretrained on much larger corpora (6 billion and 5.5 billion tokens respectively, comparing to our 1 million tokens).
- This is especially important in practice, because collecting data and pretraining models are expensive.

	Word vectors		**LMs**	
	GloVe	Ours	ELMo	Ours
CADEC	**70.30**	70.27	**71.91**	70.46
CoNLL2003	**90.25**	86.36	**91.34**	89.78
CRAFT	74.22	**75.45**	**75.77**	75.45
JNLPBA	73.19	**73.24**	73.65	**74.29**
ScienceIE	37.10	**37.91**	41.15	**42.07**
WetLab	**79.15**	78.93	79.57	**79.62**

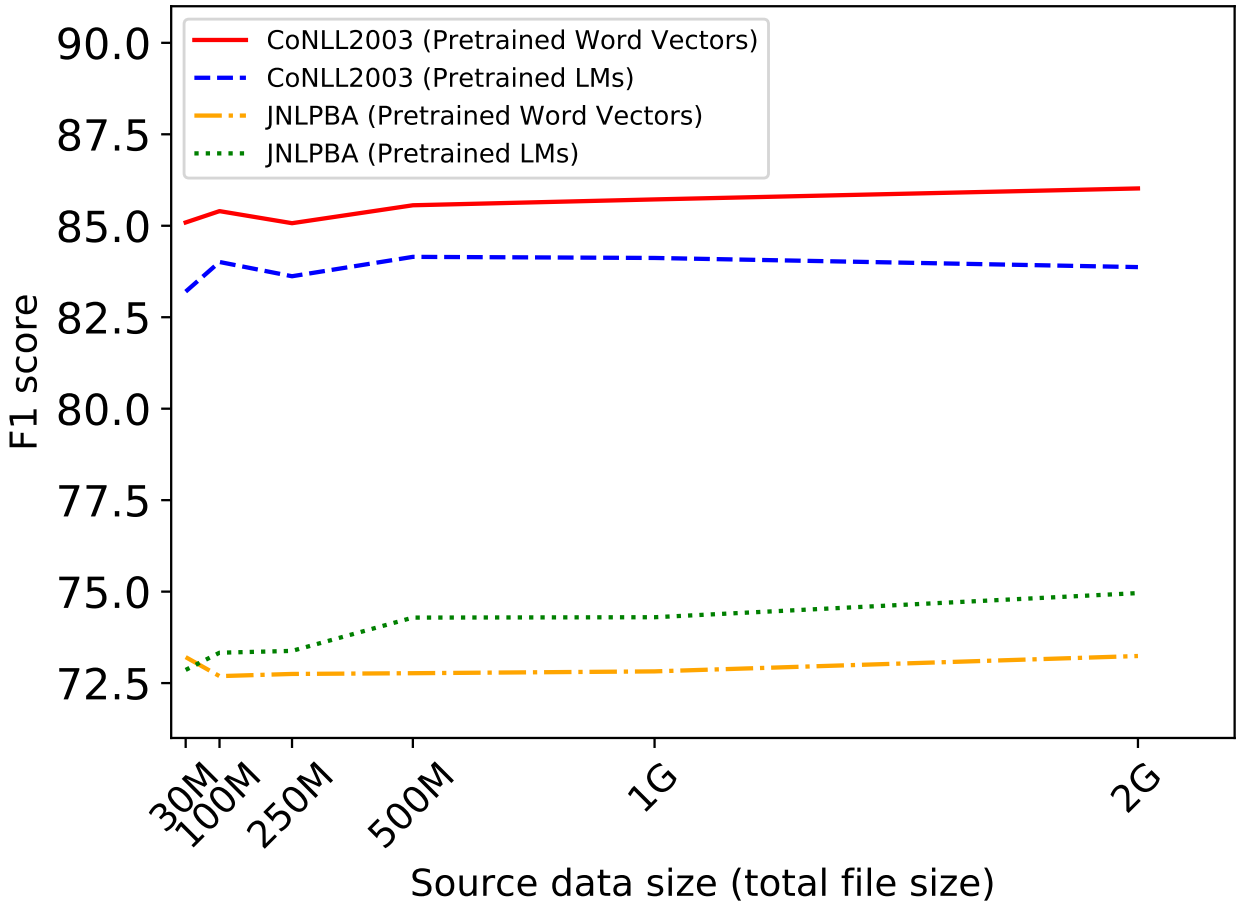
Big data is good, but is it worthy?

- Sample six PubMed subsets of different size as pretraining data, CoNLL2003 and JNLPBA as target data.
- 500 MB of pretraining data appears to be sufficient to calculate similarity, and factor out the impact of size.



If similar source data is available, pretrain LMs. Otherwise, pretrain word vectors.

- PubMed is dissimilar to CoNLL2003, pretrained word vectors is better.
- PubMed is similar to JNLPBA, so pretrain LMs.



Contributions

- Three cost-effective similarity measures: Vocabulary Intersection Rate, Language Model Perplexity, and Word Vector Variance. (Good predictors of NER performance.)
- Consider tenor, as well as field. (In contrast to human intuition.)
- Models pretrained on a modest amount of similar data outperform the ones pretrained on very large generic data.
- Our pretrained word vectors and LMs are publicly available.


Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016).
Exploring the limits of language modeling.
CoRR abs/1602.02410.


Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016).
Pointer sentinel mixture models.
CoRR abs/1609.07843.


Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013).
Efficient estimation of word representations in vector space.
CoRR abs/1301.3781.


Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018).
Deep contextualized word representations.
In *NAACL*, pages 2227–2237, New Orleans, Louisiana.