USING SIMILARITY MEASURES TO SELECT PRETRAINING DATA FOR NER

Xiang Dai, Sarvnaz Karimi, Ben Hachey, Cecile Paris

CSIRO Data61 and University of Sydney

Motivation

- Word vectors and Language Models pretrained on a large amount of unlabelled data can dramatically improve various NLP tasks. However, the measure and impact of similarity between pretraining data and target task data are left to intuition.

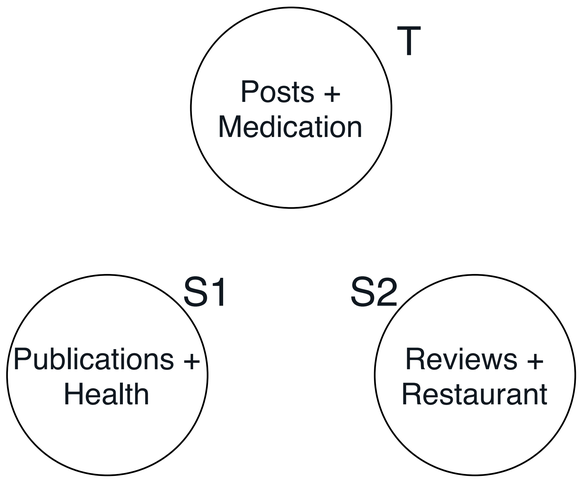

Fig. 1: Do you think unsupervised pretraining on S would be useful for supervised NER learning on T?

- We conducted a survey on human intuition regarding selection of pretraining data across 30 NLP or machine learning practitioners.
- Participants were provided short descriptions of the target data set T, and two possible source data sets S1 and S2 as
  - T: Online forum posts about medications;
  - S1: Research papers about biology and health;
  - S2: Online reviews about restaurants, hotels, barbers, mechanics, etc.

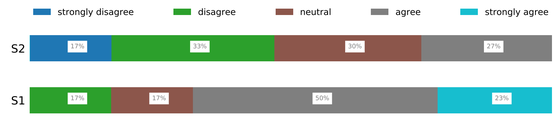

Fig. 2: Likert scale ratings from NLP and ML practitioners for the survey.

- Intuition varies across practitioners.
- Practitioners favour **field** (the subject matter of the content being discussed) over **tenor** (the participants in the discourse, their relationships to each other, and their purposes) when selecting pretraining data.

Methods and Results

- Goal: To develop a **cost-effective** approach that, given a NER data set, nominates the most suitable source data to pretrain word vectors or LMs from several options.
- We propose three simple similarly measures: Target Vocabulary Covered (TVC), Language Model Perplexity (PPL), Word Vector Variance (WVV).
- Calculate source-target similarity values using proposed similarity measures.
- Pretrain word vectors [2] and language models [1] using different sources seperately.
- Replace the word embedding weights initialized by different pretrained word vectors.
- Incorporate different pretrained LMs, following the approach proposed by [3].
- Find out the relationships between improvement due to pretrained models and TVC, TVcC, PPL and WVV.

		Similarity				NER F_1 Score			
						Pretrained word vectors		Pretrained LMs	
Target	Source	PPL	WVV	TVC (%)	TVcC (%)	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

Fig. 3: Similarity between source and target data sets (left), and the effectiveness of word vectors and LMs pretrained using different sources for NER (right). Lower PPL or WVV values indicate higher similarity between source and target, while higher TVC and TVcC values indicate higher similarity. *None* rows refer to the models that word embedding weights are randomly initialized with no pretrained LMs. Δ shows absolute improvement.

Tenor vs. Field

- Tenor is **reflected more** than Field by our proposed measures.
- Yelp (reviews about local businesses) is more similar to CADEC (online posts about medications) than PubMed (publications about biomedical and health) and MIMIC (notes about clinical).
- PubMed is similar to ScienceIE (publications about computer science, material and physics)
- All sources are measured against WetLab (protocols about biology experiments) with relatively high PPL and WVV values, although WetLab has a similar field (biology) with PubMed.

Practical guidelines

- Our proposed similarity measures can reach high level of consensus (inter-method agreement).
- LM performance has a stronger correlation with similarity measures than word vectors (Pearson correlation analysis).
- VCcR is the most informative factor in predicting the effectiveness of pretrained word vectors given a target data set.
- Word vectors and LMs pretrained on small similar sources can achieve competitive or even better performance than the ones pretrained on larger sources (publicly available GloVe and ELMo).
- If source and target data are dissimilar, pretrained word vectors is a better option than pretrained LMs, no matter how large source data is.
- Pretrained LMs outperform pretrained word vectors, if source is similar to target.
- Hyper-parameter tuning can overall produce better performance, but our observation that similar sources generate better pretrained models can still hold.

References

[1] Rafal Jozefowicz et al. “Exploring the Limits of Language Modeling”. In: *CoRR abs/1602.02410* (2016).

[2] Tomas Mikolov et al. “Efficient Estimation of Word Representations in Vector Space”. In: *CoRR abs/1301.3781* (2013).

[3] Matthew Peters et al. “Deep Contextualized Word Representations”. In: *NAACL*. New Orleans, Louisiana, 2018, pp. 2227–2237.

Contact

- Due to visa issues, the first author cannot physically attend the conference. If you are interested in our work, please feel free to contact: Xiang Dai (Email: dai.dai@csiro.au, personal website: https://daixiangau.github.io/).