



Cross-Lingual Transfer Learning for POS Tagging without Cross-Lingual Resources

WeiYang

weiyang@godweiyang.com www.godweiyang.com

East China Normal University

Department of Computer Science and Technology

2018.01.17





Outline

Outline

Introduction

Model

Experiments

Conclusions

References







Motivations

- Models with ancillary resources such as parallel corpora greatly dependent on the quantity and quality of the resources.
- Given no linguistic resources between the source language and the target language, transfer learning methods can be utilized instead.







General Model

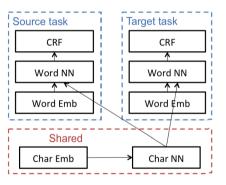


Figure: Transfer model used for cross-lingual transfer. [Yang et al, 2017]





Model

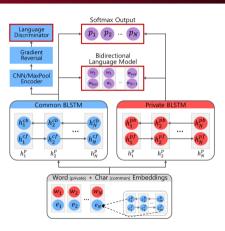


Figure: Model architecture. [Kim et al,. 2017]



Loss Functions

Sequence Tagging Loss

$$\mathcal{L}_p = -\sum_{i=1}^S \sum_{j=1}^N p_{i,j} \log(\hat{p}_{i,j})$$

Language Classifier Loss

$$\mathcal{L}_a = -\sum_{i=1}^{S} l_i \log(\hat{l}_i)$$

Bidirectional Language Model Loss

$$\mathcal{L}_{l} = -\sum_{i=1}^{S} \sum_{j=1}^{N} \log(P(w_{j+1}|f_{j})) + \log(P(w_{j-1}|b_{j}))$$



Total Loss Functions

$$\mathcal{L} = w_s(\mathcal{L}_p + \lambda \mathcal{L}_a + \lambda \mathcal{L}_l)$$

- \bullet λ from 0 to 1.
- w_s is equal to 1 when training target language.
- w_s is equal to $Size_t/Size_s$ when training source language.





Language-Adversarial Training

- Pass the common Bi-LSTM output to three CNN filters using max pooling and tanh activation function.
- Concatenate the three vectors and forward them to the language discriminator through the gradient reversal layer.
- The discriminator is implemented as a fully-connected neural network whose activation function is Leaky ReLU.
- The gradient from the language classifier is negated so that the bottom layers are trained to be language-agnostic.







	Target only		Source (English) → Target				
Language	С	c,l	c,l	p,l	c,p,l	c,l+a	c,p,l+a
Swedish	93.26	94.31	94.36	94.39	94.51	94.38	94.63
Danish	92.13	93.41	93.34	93.76	94.05	93.74	94.26
Dutch	83.24	84.73	85.20	84.92	84.85	84.99	85.83
German	89.27	90.69	90.06	90.40	90.01	90.14	90.71
Avg	⁻ 89.47	⁻ 90.78 ⁻	⁻ 90.74	⁻ 90.87	90.86	⁻ 90.82	91.36
Slovenian	93.06	93.79	93.83	94.06	94.20	93.93	94.06
Polish	91.30	91.30	91.69	92.11	91.86	91.77	92.11
Slovak	86.53	89.56	90.11	89.88	89.98	90.40	90.01
Bulgarian	93.45	95.27	95.33	95.50	95.52	95.25	95.65
Avg	91.09	92.48	92.74	92.89	92.89	92.84	92.95
Romanian	93.20	94.09	94.22	94.17	94.05	93.91	94.20
Portuguese	94.23	95.18	95.42	95.15	95.55	95.36	95.51
Italian	93.80	95.95	95.79	95.61	95.84	95.70	95.92
Spanish	91.94	93.34	93.34	93.31	93.29	92.94	93.44
Avg	93.29	94.64	94.69	94.56	94.68	94.48	94.77
Persian	93.91	94.63	94.68	94.79	94.78	94.49	94.83
Hungarian	93.20	93.27	94.40	94.66	94.69	94.29	94.45
Total Avg	91.61	92.82	92.98	93.05	93.08	92.95	93.26
	Swedish Danish Dutch German Ävg Slovenian Polish Slovak Bulgarian Ävg Romanian Portuguese Italian Spanish Ävg Persian Hungarian	Language c Swedish 93.26 Danish 92.13 Duch 83.24 German 89.27 Avg 89.47 Slovenian 93.06 Polish 91.30 Slovak 86.53 Bulgarian 93.49 Avg 91.90 Portuguese 94.23 Italian 93.90 Spanish 93.91 Avg 93.29 Persian 93.29 Hungarian 93.20	Language c c. Swedish 39.26 93.1 Danish 92.13 93.41 Dutch 83.24 94.73 German 89.27 90.68 Siloweinia 91.30 91.30 Polish 91.30 91.30 Slowak 86.53 89.56 Bulgarian 93.49 95.27 Avg 91.09 92.48 Romanian 93.20 94.09 Portuguese 94.23 95.18 Italian 93.80 95.95 Spunish 91.94 93.34 Avg 93.29 93.64 Hungarian 93.20 93.29 Hungarian 93.20 93.29	Language c c.l c.l Swedish 93.26 94.31 94.36 Danish 92.13 93.41 93.34 Dutch 83.24 84.73 85.20 German 89.27 90.69 90.06 Avg 93.06 93.79 93.43 Polish 91.30 91.30 91.50 Slowak 86.53 89.56 90.11 Bulgarian 93.45 95.27 95.33 Avg 91.09 94.82 92.74 Portuguese 94.23 95.18 95.49 Portuguese 94.23 95.18 95.29 Spanish 91.94 93.24 93.54 Avg 91.94 93.34 93.34 Avg 93.29 93.29 93.29 Avg 93.29 93.29 93.29 Hungarian 93.20 93.29 93.29	Language c c.l b.l 9.4 9.43 Swedish 93.26 94.31 94.36 94.39 Danish 92.13 93.41 93.34 93.76 Dutch 83.24 84.73 85.20 84.92 German 89.27 90.69 90.69 90.40 Avg 89.07 90.78 90.74 90.87 Polish 91.30 91.30 91.09 92.11 Slowak 86.53 89.56 90.11 88.63 89.56 90.11 88.60 89.79 93.33 95.50 Romanian 93.20 94.09 94.22 94.79 92.74 92.87 94.09 94.22 94.77 Portiguese 93.89 95.95 95.79 95.61 Sparish 93.49 93.44 93.34 93.31 93.61 Sparish 93.49 94.69 94.22 94.79 94.25 94.79 94.56 94.79 94.25 94.79 94.69 94.25	Language c c.l p.l c.p.l Swedish 39.26 94.31 94.36 94.39 94.51 Danish 92.13 93.41 93.46 93.76 94.05 Dutch 83.24 94.73 85.20 84.92 84.85 German 89.27 90.68 90.06 90.40 90.01 Avg 91.30 91.39 91.39 93.33 94.09 94.21 Polish 91.30 91.30 91.09 92.11 91.88 89.98 Bulgarian 93.45 95.27 95.33 95.09 92.89 Romanian 93.20 94.09 94.22 94.17 94.05 Portuguese 94.23 95.18 95.27 95.31 95.05 95.89 Romanian 93.20 94.09 94.22 94.17 94.05 Spuinsh 91.99 94.22 95.18 95.55 95.79 95.61 95.84 Spuinsh 91.99 </td <td>Language c c.l c.l p.l c.p.d c.l-a Swedish 93.26 94.31 94.36 94.39 94.51 93.44 Danish 92.13 93.41 93.34 93.76 94.05 93.74 Dutch 83.24 84.73 85.20 84.92 94.09 90.01 90.10 90.11 90.11 91.02 90.11 91.02 90.19 91.01 91.01 93.39 91.09 94.20 93.93 91.09 92.11 91.86 91.79 91.03 91.09 91.01 93.03 93.09 91.70 93.03 93.09 91.70 93.03 93.09 91.70 93.03 93.09 91.70 93.03 93.09 93.29 93.09 93.29 93.09 93.93 93.00 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.24 93.24 93.24 93.24</td>	Language c c.l c.l p.l c.p.d c.l-a Swedish 93.26 94.31 94.36 94.39 94.51 93.44 Danish 92.13 93.41 93.34 93.76 94.05 93.74 Dutch 83.24 84.73 85.20 84.92 94.09 90.01 90.10 90.11 90.11 91.02 90.11 91.02 90.19 91.01 91.01 93.39 91.09 94.20 93.93 91.09 92.11 91.86 91.79 91.03 91.09 91.01 93.03 93.09 91.70 93.03 93.09 91.70 93.03 93.09 91.70 93.03 93.09 91.70 93.03 93.09 93.29 93.09 93.29 93.09 93.93 93.00 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.70 93.24 93.24 93.24 93.24

Figure: POS tagging accuracies with 1280 tag-labeled training examples for each target language.

[Kim et al,. 2017]





<u>Performance</u>

		Target only		Source (English) → Target				t
Language Family	Language	р	p,l	p,l	c,l	p,c,l	c,l+a	p,c,l+a
Germanic	Swedish	87.43	90.49	91.02	90.45	90.48	90.72	90.70
	Danish	86.42	90.00	90.74	90.69	90.02	90.16	90.79
	Dutch	76.76	82.24	82.61	82.46	82.10	82.58	82.15
	German	86.25	88.95	89.10	88.69	88.93	88.08	89.68
	Avg	84.22	⁻ 87.92	88.37	^{88.07}	87.88	87.88	88.33
Slavic	Slovenian	87.02	89.97	90.29	90.00	90.32	89.58	90.59
	Polish	82.10	84.13	85.21	85.41	85.30	85.46	85.50
	Slovak	76.22	81.03	82.95	83.40	82.68	82.70	83.17
	Bulgarian	87.32	92.81	92.68	92.07	92.30	92.20	92.39
	Avg	83.16	86.98	87.78	87.72	87.65	87.48	87.91
Romance	Romanian	88.67	91.44	91.44	90.87	91.22	90.85	91.37
	Portuguese	90.66	93.73	93.55	93.90	93.81	93.58	94.20
	Italian	89.78	93.99	93.82	93.27	93.46	93.51	94.00
	Spanish	85.91	91.07	90.59	90.59	91.07	90.17	90.88
	Avg	88.76	92.56	92.35	92.16	92.39	92.03	92.61
Indo-Iranian	Persian	90.64	92.40	91.98	91.97	92.12	92.18	91.83
Uralic	Hungarian	89.14	90.65	91.45	91.48	90.91	91.52	90.72
	Total Avg	86.02	89.49	89.82	89.66	89.62	89.52	89.86

Figure: POS tagging accuracies with 320 tag-labeled training examples for each target language. [Kim et al,. 2017]





Conclusions

- Transferring languages in the same families can gain higher accuracies.
- When the tag-labeled training sentences are extremely little or abundant, just utilizing private Bi-LSTMs shows better accuracies.
- Utilizing multiple languages as the source languages can gain higher performance.





Comparison

Extra Resource	Index & Model	F ₁ score		
Extra resource	macx ac model	Type	Value (±std)	
gazetteers	0) Collobert et al. 2011 [†]	reported	89.59	
gazetteers	 Chiu et al. 2016[†] 	reported	91.62±0.33	
AIDA dataset	2) Luo et al. 2015	reported	91.20	
CoNLL 2000 / PTB-POS dataset	3) Yang et al. 2017 [†]	reported	91.26	
1B Word dataset & 4096-8192-1024	4) Peters et al. 2017†‡	reported	91.93±0.19	
1B Word dataset	 Peters et al. 2017^{†‡} 	reported	91.62±0.23	
	 Collobert et al. 2011[†] 	reported	88.67	
	7) Luo et al. 2015	reported	89.90	
	8) Chiu et al. 2016 [†]	reported	90.91±0.20	
	 Yang et al. 2017[†] 	reported	91.20	
Transfer Learning 4	10) Peters et al. 2017 [†]	reported	90.87±0.13	
	11) Peters et al. 2017 ^{†‡}	reported	90.79±0.15	
None Di		mean	87.38±0.36	
	12) Rei 2017 †‡	max	87.94	
		reported	86.26	
		mean	90.76±0.08	
	13) Lample et al. 2016 [†]	max	91.14	
		reported	90.94	
		mean	91.37±0.17	
	14) Ma et al. 2016 [†]	max	91.67	
		reported	91.21	
	15) LM-LSTM-CRF †‡	mean	91.71±0.10	
	15) Em Estim-Citi	max	91.85	

Figure: F1 score on the CoNLL03 NER dataset. [Liu et al., 2018]





Comparison

Semi-supervised Learning vs Transfer Learning

- It seems that semi-supervised learning is better than transfer learning on some tasks.
- Semi-supervised learning is not always useful for the lack of unlabeled data in the same domain.
- Andrew Ng had said that transfer learning is an important research direction in the next five years.

Future

- Semi-supervised learning and transfer learning can be combined to increase performance.
- Other methods like active learning can be added.





References



Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, Eric Fosler-Lussier. (2017).

Processing, pages 2822–2828, Copenhagen, Denmark, September 7–11, 2017.

Cross-Lingual Transfer Learning for POS Tagging without Cross-Lingual Resources. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language

Zhilin Yang, Ruslan Salakhutdinov, William W. Cohen. (2017).

Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks.

In ICLR 2017.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank F. Xu, Huan Gui, Jian Peng, Jiawei Han. (2018).

Empower Sequence Labeling with Task-Aware Neural Language Model.

In AAAI, arXiv:1709.04109, 2018.

