Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition

Joey Tianyi Zhou[1*], Hao Zhang[1*], Di Jin[2], Hongyuan Zhu[1], Meng Fang[3], Rick Siow Mong Goh[1], Kenneth Kwok[1]

* equal contribution

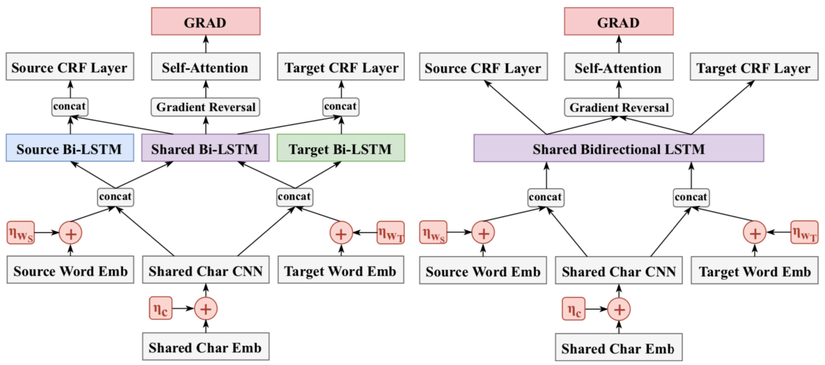
[1]Agency for Science, Technology and Research, [2]Massachusetts Institute of Technology, [3]Tencent Robotics X

Motivation

- Deep learning based methods for named entity recognition usually require mass of training data for better generalization abilities.
- When annotated corpus is small, those methods degrade significantly, since hidden features cannot be learned adequately. *Transfer learning* is a way to overcome such obstacle by borrowing knowledge from other resources.
- Although the existing transfer-based methods show promising performance in low-resource settings. There are two issues deserved to be further investigated on:
  - **Representation Difference**: They did not consider the representation difference across source and target and enforced them to be shared across languages/domains.
  - **Resource Data Imbalance**: the training size of high-resource is usually much larger than the of low-resource.

Almost all existing methods neglect the difference in their models, thus resulting in poor generalization.

Approach



- To consider the resource representation difference issue, we introduce two variants of DATNet, termed DATNet-P and DATNet-F.
  - DATNet-F: all the units of BiLSTM are shared by both resources.
  - DATNet-P: BiLSTM units are decomposed into shared component and the resource related one.
- To handle the resource data imbalance issue, we propose General Resource-Adversarial Discriminator (GRAD) to impose the resource weight to pay more attention to low-resource and hard samples.

Generalized Resource-Adversarial Discriminator (GRAD):

$$\ell_{GRAD} = -\sum_i \{I_{i\in D_S}\alpha(1-r_i)^\gamma \log r_i + I_{i\in D_T}(1-\alpha) r_i^\gamma \log(1-r_i)\}$$

where r_i is a scalar, $I_{i\in D_S}$, $I_{i\in D_T}$ are the identity functions to denote whether r_i is from high- or low-resources.

α is a weighting factor to balance the loss contribution from high and low resource.

$(1-r_i)^\gamma$ (or r_i^γ) controls the loss contribution from individual samples by measuring the discrepancy between prediction and true label (easy samples have smaller contribution).

γ is a factor that smoothly adjusts the rate at which easy examples are down-weighted.

Experiments

Datasets

Benchmark	Resource	Language	# Training Tokens (# Entities)	# Dev Tokens (# Entities)	# Test Tokens (# Entities)
CoNLL-2003	Source	English	204,567 (23,499)	51,578 (5,942)	46,666 (5,648)
Cross-language NER					
CoNLL-2002	Target	Spanish	207,484 (18,797)	51,645 (4,351)	52,098 (3,558)
CoNLL-2002	Target	Dutch	202,931 (13,344)	37,761 (2,616)	68,994 (3,941)
Cross-domain NER					
WNUT-2016	Target	English	46,469 (2,462)	16,261 (1,128)	61,908 (5,955)
WNUT-2017	Target	English	62,730 (3,160)	15,733 (1,250)	23,394 (1,740)

- CoNLL-2003 dataset is treated as high-resource (i.e. source) in our experiments.
- CoNLL-2002 datasets as the low-resource (i.e. target) data for cross-language transfer settings.
- WNUT-2016/2017 datasets as the low-resource (i.e. target) data for cross-domain transfer settings.

Comparison with State-of-the-Art Results

Mode	Methods		Additional Features: POS	Gazetteers	Orthographic	CoNLL Datasets: Spanish	Dutch	WNUT Datasets: WNUT-2016	WNUT-2017
Mono-language /domain	(Gillick et al., 2016)		×	×	×	82.59	82.84	-	-
	(Lample et al., 2016)		×	√	×	85.75	81.74	-	-
	(Partalas et al., 2016)		√	√	√	-	-	46.16	-
	(Limsopatham and Collier, 2016)		×	×	√	-	-	52.41	-
	(Lin et al., 2017a)		√	√	×	-	-	-	40.42
	Our Base Model	Best	×	×	×	85.53	85.55	44.96	35.20
		Mean & Std				85.35±0.15	85.24±0.21	44.37±0.31	34.67±0.34
Cross-language /domain	(Yang et al., 2017)		×	√	×	85.77	85.19	-	-
	(Lin et al., 2018)		×	√	×	85.88	86.55	-	-
	(Feng et al., 2018)		√	×	×	86.42	**88.39**	-	-
	(von Däniken and Cieliebak, 2017)		×	√	×	-	-	-	40.78
	(Aguilar et al., 2017)		√	×	√	-	-	-	41.86
	DATNet-P	Best	×	×	×	**88.16**	88.32	50.85	41.12
		Mean & Std				87.89±0.18	88.09±0.13	50.41±0.32	40.52±0.38
	DATNet-F	Best	×	×	×	87.04	87.77	**53.43**	**42.83**
		Mean & Std				86.79±0.20	87.52±0.19	53.03±0.24	42.32±0.32

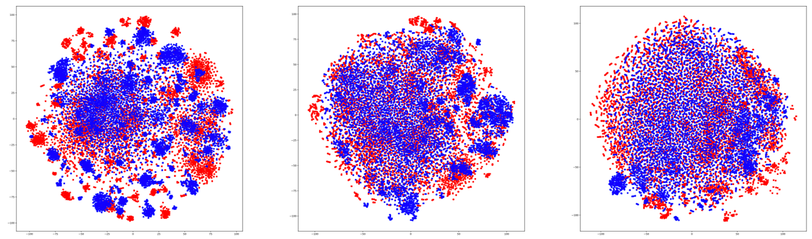
Transfer Learning Performance

Tasks	CoNLL-2002 Spanish NER						WNUT-2016 Twitter NER					
# Target train sentences	10	50	100	200	500	1000	10	50	100	200	500	1000
Base	21.53	42.18	48.35	63.66	68.83	76.69	3.80	14.07	17.99	26.20	31.78	36.99
+ AT	19.23	41.01	50.46	64.83	70.85	77.91	4.34	16.87	18.43	26.32	35.68	41.69
+ P-Transfer	29.78	61.09	64.78	66.54	72.94	78.49	7.71	16.17	20.43	29.20	34.90	41.20
+ F-Transfer	39.72	63.00	63.36	66.39	72.88	78.04	15.26	20.04	26.60	32.22	38.35	44.81
DATNet-P	39.52	62.57	64.05	**68.95**	**75.19**	**79.46**	9.94	17.09	25.39	30.71	36.05	42.30
DATNet-F	**44.52**	**63.89**	**66.67**	68.35	74.24	78.56	**17.14**	**22.59**	**28.41**	**32.48**	**39.20**	**45.25**

1. DATNet-F outperforms DATNet-P on cross-language transfer when the target resource is extremely low, however, this situation is reversed when the target dataset size is large enough (e.g., more than 100 sentences);
2. DATNet-F is generally superior to DATNet-P on cross-domain transfer.

Feature Visualization

- CoNLL-2003 English examples
- CoNLL-2002 Spanish examples



(a) no AD (b) AD (c) GRAD

The visualization of extracted features from shared bidirectional-LSTM layer. GRAD makes the distribution of extracted features from the source and target datasets much more similar by considering the data imbalance, which indicates that ***the outputs of BiLSTM are resource-invariant***.

Effects of Different Components

CoNLL-2002 Spanish NER: Model	F1-score	Model	F1-score	WNUT-2016 Twitter NER: Model	F1-score	Model	F1-score
Base	85.35	+AT	86.12	Base	44.37	+AT	47.41
+P-T (no AD)	86.15	+AT +P-T (no AD)	86.90	+P-T (no AD)	47.66	+AT +P-T (no AD)	48.44
+F-T (no AD)	85.46	+AT +F-T (no AD)	86.17	+F-T (no AD)	49.79	+AT +F-T (no AD)	50.93
+P-T (AD)	86.32	+AT +P-T (AD)	87.19	+P-T (AD)	48.14	+AT +P-T (AD)	49.41
+F-T (AD)	85.58	+AT +F-T (AD)	86.38	+F-T (AD)	50.48	+AT +F-T (AD)	51.84
+P-T (GRAD)	86.93	+AT +P-T (GRAD) (*DATNet-P*)	**88.16**	+P-T (GRAD)	48.91	+AT +P-T (GRAD) (*DATNet-P*)	50.85
+F-T (GRAD)	85.91	+AT +F-T (GRAD) (*DATNet-F*)	87.04	+F-T (GRAD)	51.31	+AT +F-T (GRAD) (*DATNet-F*)	**53.43**

* AT: Adversarial Training; P-T: P-Transfer; F-T: F-Transfer; AD: Adversarial Discriminator; GRAD: Generalized Resource-Adversarial Discriminator.

1. GRAD shows the stable superiority over the normal AD regardless of other components.
2. The DATNet-P architecture is more suitable to cross-language transfer whereas DATNet-F is more suitable to cross-domain transfer.

Conclusion

In this paper we develop a transfer learning model DATNet for low-resource NER, which aims at addressing representation difference and resource data imbalance problems. We introduce two variants, DATNet-F and DATNet-P, which can be chosen according to cross-language/domain user case and target dataset size. To improve model generalization, we propose dual adversarial learning strategies, i.e., AT and GRAD. Extensive experiments show the superiority of DATNet over existing models and it achieves significant improvements on CoNLL and WNUT NER benchmark datasets.

Acknowledgement

This paper is supported by the Singapore Government's Research, Innovation and Enterprise 2020 Plan, Advanced Manufacturing and Engineering domain (Programmatic Grant No. A1687b0033, A18A1b0045) and the Agency for Science, Technology and Research, under the AME Programmatic Funding Scheme (Project No. A18A2b0046, A1718g0048).