Yale

Robust Multilingual Part-of-Speech Tagging via Adversarial Training

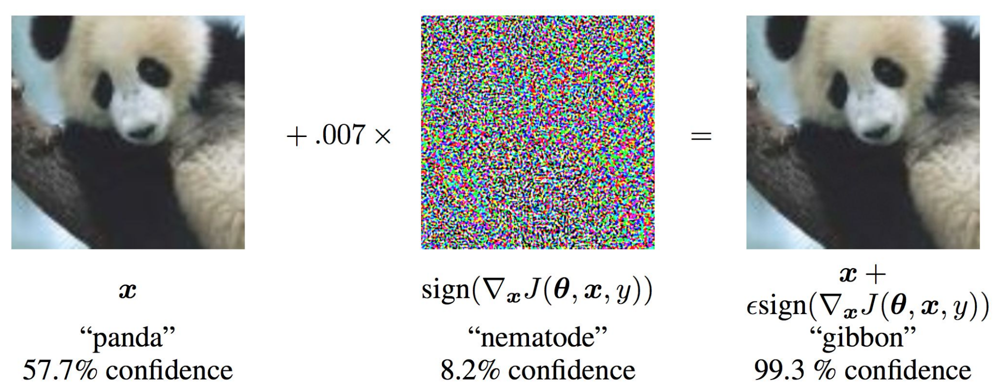
Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev
Department of Computer Science, Yale University

LILY Lab

Introduction

Adversarial examples: very close to original inputs but are likely to be misclassified by the current model

[Goodfellow, et al 2015]



Adversarial training (AT) aims to improve robustness to input perturbations by training on both clean examples and adversarial examples.
Yet, the specific effects of the robustness obtained from AT are still unclear in the context of **NLP**, e.g.,

- how to interpret perturbations on natural language input?
- Is AT language/task dependent?

This paper proposes and analyzes a neural POS tagging model that exploits AT. In our experiments on PTB-WSJ and the Universal Dependencies (UD) dataset (27 languages), we not only find that AT improves the overall tagging accuracy, but also obtain the following insights into AT in the context of NLP:

1) AT prevents over-fitting well in low resource languages
2) AT boosts tagging accuracy for rare/unseen words
3) the improved tagging performance by AT contributes to downstream tasks, e.g., dependency parsing
4) AT helps the model to learn cleaner word representations

Thus, AT can be interpreted from the perspective of natural language. We also find:

5) AT is generally effective in different languages and different sequence labeling tasks.

These positive results motivate further use of AT in NLP.

Tagging Models

1. Baseline: BiLSTM-CRF

- Character-level BiLSTM
- Word-level BiLSTM
- Conditional random field (CRF) for global inference of tags

Loss function:

$$L(\theta; s, y) = -\log p(y \mid s; \theta)$$

2. Adversarial Training (AT)

At each training step, we first generate adversarial examples by adding small perturbations to the inputs in the direction that significantly increases the loss function.
Then, the model is trained on the mixture of clean examples and adversarial examples.

Generating adversarial examples. Given a sentence

$$s = [w_1, w_2, \ldots, c_1, c_2, \ldots]$$

define adversarial perturbations on its word/character embeddings:

$$\eta = \epsilon\, g / \|g\|_2 \quad \text{where } g = \nabla_s L(\hat{\theta}; s, y)$$

Adversarial example:

$$s_{\mathrm{adv}} = s + \eta$$

Note:
- Normalize embeddings [Miyato et al., 2017]
- Set small perturbation norm ϵ to be $\alpha\sqrt{D}$ (i.e., proportional to $\sqrt{D}$, where $s \in \mathbb{R}^D$

Training. Minimize adversarial loss:

$$\tilde{L} = \gamma L(\theta; s, y) + (1-\gamma) L(\theta; s_{\mathrm{adv}}, y)$$

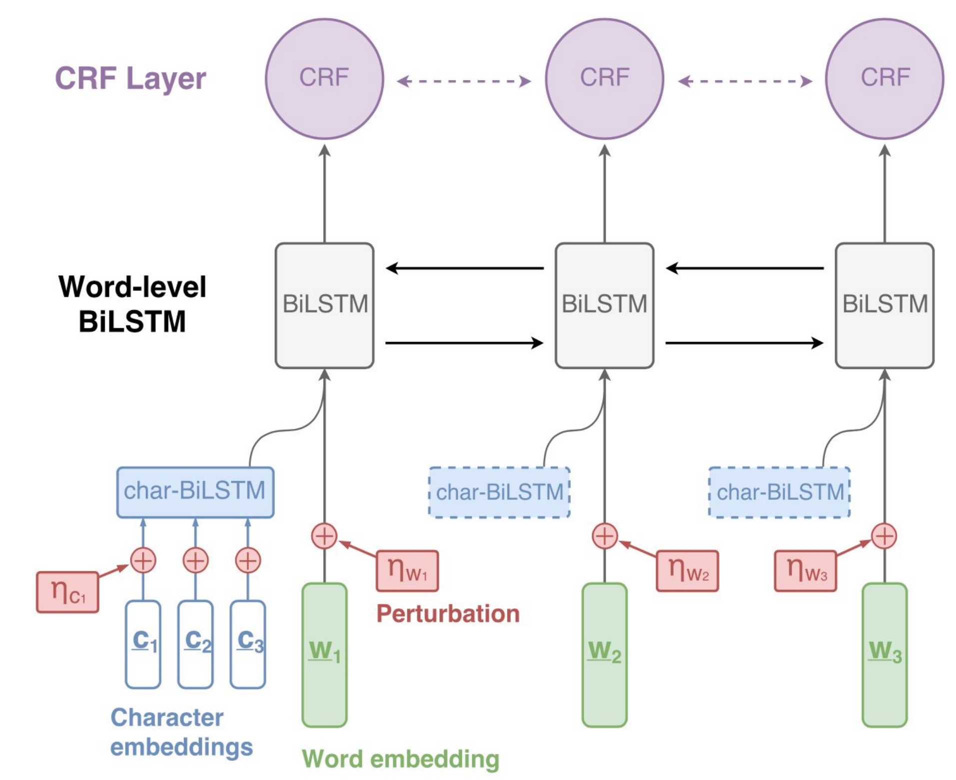


Figure 1. Illustration of our BiLSTM-CRF-AT model.

Experiments & Results

1. Dataset

- PTB-WSJ (English)
- UD v1.2 (27 languages)

for POS tagging

2. Results

PTB-WSJ. Tagging accuracy:
97.54 (baseline) → 97.58 (AT)
outperforming most existing works.

UD (27 languages).
Improvements by AT on all languages.

- 21 resource-rich: 96.45 → 96.65 (0.20% up on average)
- 6 resource-poor: 91.20 → 91.55 (0.35% up on average)

Followed the definition of resource rich/poor in [Plank et al., 2016].

=> AT prevents overfitting especially well in low-resource languages.

AT's regularization is generally effective across different languages.

AT is a data augmentation technique: we generate and train with new examples the current model is particularly vulnerable to.

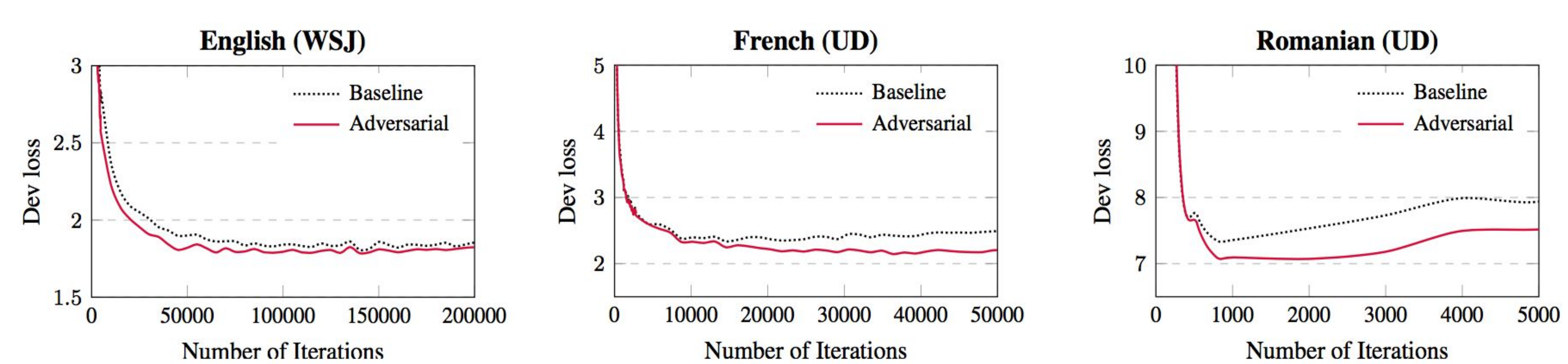


Figure 2. Learning curves for three representative languages (Romanian is low resource)

Analysis

1. Word-level Analysis

Motivation: poor tagging accuracy on rare/unseen words is a bottleneck in existing POS taggers. Does AT help for this issue?

- Tagging accuracy on words categorized by the frequency of occurrence in training.

English (WSJ)

Word Frequency	0	1-10	10-100	100-	Total
# Tokens	3240	7687	20908	97819	129654
Baseline	**92.25**	95.36	96.03	98.19	97.53
Adversarial	92.01	**95.52**	96.10	98.23	97.57

French (UD)

Word Frequency	0	1-10	10-100	100-	Total
# Tokens	356	839	1492	4523	7210
Baseline	87.64	94.05	94.03	98.43	96.48
Adversarial	87.92	**94.88**	94.03	98.50	96.63

- Tagging accuracy on neighbor words

English (WSJ)

Word Frequency	0	1-10	10-100	100-	Total
# Tokens	6480	15374	41815	195637	259306
Baseline	97.76	97.71	97.80	97.45	97.53
Adversarial	**98.06**	97.71	97.89	97.47	97.57

French (UD)

Word Frequency	0	1-10	10-100	100-	Total
# Tokens	712	1678	2983	9045	14418
Baseline	95.08	97.08	97.58	96.11	96.48
Adversarial	**95.37**	97.26	97.79	96.23	96.63

=> Notable improvements on rare words and neighbors of unseen words

2. Sentence-level Analysis

Sentence-level accuracy & downstream dependency parsing performance

English (WSJ)

	Sentence-level Acc.	Stanford Parser UAS	Stanford Parser LAS	Parsey McParseface UAS	Parsey McParseface LAS
Baseline	59.08	91.53	89.30	91.68	87.92
Adversarial	**59.61**	**91.57**	**89.35**	**91.73**	**87.97**
(w/ gold tags)	–	(92.07)	(90.63)	(91.98)	(88.60)

French (UD)

	Sentence-level Acc.	Parsey Universal UAS	Parsey Universal LAS
Baseline	52.35	84.85	80.36
Adversarial	**53.36**	**85.01**	**80.55**
(w/ gold tags)	–	(85.05)	(80.75)

- Robustness to rare/unseen words enhances sentence-level accuracy
- POS tags predicted by the AT model also improve downstream dependency parsing. Sentence-level accuracy is important for downstream tasks.

3. Word Representation Learning

Motivation: does AT help to learn more robust word embeddings?

- Cluster words based on POS tags, and measure the tightness of word vector distribution within each cluster (using cosine similarity metric)

English (WSJ)

POS Cluster	NN	VB	JJ	RB	Avg.
1) Initial (GloVe)	0.243	0.426	0.220	0.549	0.359
2) Baseline	0.280	0.431	**0.309**	0.667	0.422
3) Adversarial	**0.281**	**0.436**	0.306	**0.675**	**0.424**

French (UD)

POS Cluster	NOUN	VERB	ADJ	ADV	Avg.
1) Initial (polyglot)	0.215	0.233	0.210	0.540	0.299
2) Baseline	0.258	0.271	0.262	0.701	0.373
3) Adversarial	**0.263**	**0.272**	**0.263**	**0.720**	**0.379**

=> AT learns cleaner embeddings (stronger correlation with POS tags)

4. Other Sequence Labeling Tasks

Motivation: does this AT POS tagging model generalize to other sequence labeling tasks?

Chunking (PTB-WSJ). F1 score:
95.18 (baseline) → 95.25 (AT)

Named entity recognition (CoNLL-2003). F1 score: 91.22 (baseline) → 91.56 (AT)

=> The proposed AT model is generally effective across different tasks.

Conclusion

- Interpreted the effects of AT from NLP perspective
- Confirmed the general applicability and efficacy of AT