



# Sequence Tagging with Little Labeled Data

WeiYang

10142130214

[weiyang@godweiyang.com](mailto:weiyang@godweiyang.com)

[www.godweiyang.com](http://www.godweiyang.com)

East China Normal University  
Department of Computer Science and Technology

2018.01.19



# Outline

Outline

Sequence Tagging

Semi-supervised Learning

Transfer Learning

Conclusions

References





# Introduction

## Definition

Sequence tagging is a type of pattern recognition task that involves the algorithmic assignment of a categorical label to each member of a sequence of observed values.

## Significance

Sequence tagging is one of the first stages in most natural language processing applications, such as part-of-speech tagging, chunking and named entity recognition.

## Approaches

- Traditional models
  - Hidden Markov Models
  - Conditional Random Fields
- Neural network models
  - RNN, LSTM, GRU



# Neural Network Model

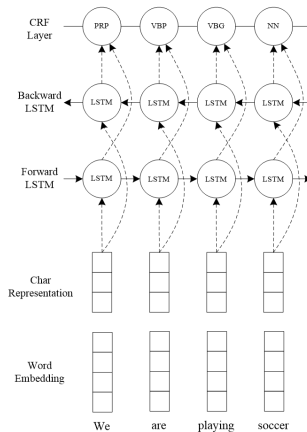


Figure: The main architecture of the neural network. [Ma et al., 2016]



# Results

Model	POS		NER					
	Dev	Test	Dev			Test		
	Acc.	Acc.	Prec.	Recall	F1	Prec.	Recall	F1
BRNN	96.56	96.76	92.04	89.13	90.56	87.05	83.88	85.44
BLSTM	96.88	96.93	92.31	90.85	91.57	87.77	86.23	87.00
BLSTM-CNN	97.34	97.33	92.52	93.64	93.07	88.53	90.21	89.36
BRNN-CNN-CRF	97.46	97.55	94.85	94.63	94.74	91.35	91.06	91.21

Figure: Performance on POS and NER. [Ma et al., 2016]



# Sequence Tagging with Little Labeled Data

## Backgrounds

Although recent neural networks obtain state-of-the-art performance on several sequence tagging tasks, they can't be used for tasks with little labeled data.

## Approaches

- Self-taught learning
- Active learning
- Transductive learning
- Semi-supervised learning
- Transfer learning



# References

## Language Models Added

- **[ACL17]** Semi-supervised Multitask Learning for Sequence Labeling. (Marek Rei)
- **[ACL17]** Semi-supervised Sequence Tagging with Bidirectional Language Models. (Matthew et al.)

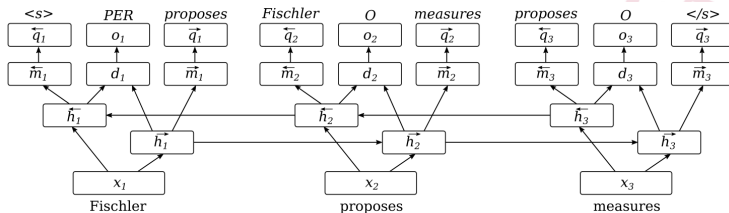
## Graph-based

- **[EMNLP10]** Efficient Graph-Based Semi-Supervised Learning of Structured Tagging Models. (Subramanya et al.)
- **[EMNLP17]** Scientific Information Extraction with Semi-supervised Neural Tagging. (Luan et al.)
- **[IEEE SLT14]** Graph-based Semi-supervised Acoustic Modeling in DNN-based Speech Recognition. (Liu et al.)



# Language Models Added

## Paper: Semi-supervised Multitask Learning for Sequence Labeling



**Figure:** The sequence tagging model with an additional LM objective. [Marek Rei, 2017]





# Language Modeling Objective

$$\begin{aligned}
 \vec{m}_t &= \tanh(\vec{W}_m \vec{h}_t) \\
 \overleftarrow{m}_t &= \tanh(\overleftarrow{W}_m \overleftarrow{h}_t) \\
 P(w_{t+1} | \vec{m}_t) &= \text{softmax}(\vec{W}_q \vec{m}_t) \\
 P(w_{t-1} | \overleftarrow{m}_t) &= \text{softmax}(\overleftarrow{W}_q \overleftarrow{m}_t) \\
 \vec{E} &= - \sum_{t=1}^{T-1} \log(P(w_{t+1} | \vec{m}_t)) \\
 \overleftarrow{E} &= - \sum_{t=2}^T \log(P(w_{t-1} | \overleftarrow{m}_t)) \\
 E &= E + \gamma(\vec{E} + \overleftarrow{E})
 \end{aligned}
 \tag{1}$$



# Results

	CoNLL-00		CoNLL-03		CHEMDNER		JNLPBA	
	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST
Baseline	92.92	92.67	90.85	85.63	83.63	84.51	77.13	72.79
+ dropout	93.40	93.15	91.14	86.00	84.78	85.67	77.61	73.16
+ LMcost	<b>94.22</b>	<b>93.88</b>	<b>91.48</b>	<b>86.26</b>	<b>85.45</b>	<b>86.27</b>	<b>78.51</b>	<b>73.83</b>

**Figure:** Performance on NER and chunking. [Marek Rei, 2017]



# Results

	GENIA-POS		PTB-POS		UD-ES		UD-FI	
	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST
Baseline	98.69	98.61	97.23	97.24	96.38	95.99	95.02	94.80
+ dropout	98.79	98.71	97.36	97.30	96.51	96.16	95.88	95.60
+ LMcost	<b>98.89</b>	<b>98.81</b>	<b>97.48</b>	<b>97.43</b>	<b>96.62</b>	<b>96.21</b>	<b>96.14</b>	<b>95.88</b>

Figure: Performance on POS. [Marek Rei, 2017]



# Language Models Added

## Paper: Semi-supervised Sequence Tagging with Bidirectional Language Models

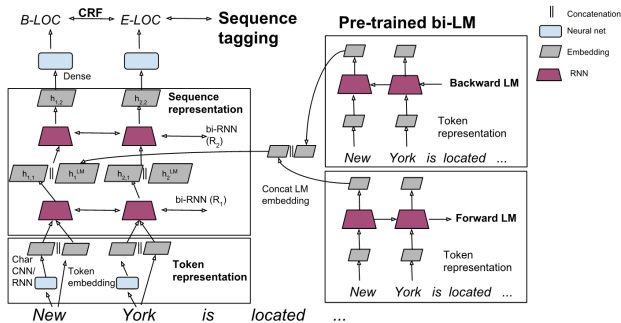


Figure: The language model augmented sequence taggers. [Matthew et al., 2017]



# Bidirectional Language Model

$$\begin{aligned} h_k^{LM} &= [\overrightarrow{h_k^{LM}}; \overleftarrow{h_k^{LM}}] \\ h_{k,1} &= [\overrightarrow{h_{k,1}}; \overleftarrow{h_{k,1}}; h_k^{LM}] \end{aligned} \quad (2)$$

## Alternative

- Replace (3) with  $f(\overrightarrow{h_{k,1}}; \overleftarrow{h_{k,1}}; h_k^{LM})$ .
- Concatenate the LM embeddings at different locations in the baseline sequence tagger.
- Decrease the number of parameters in the second RNN layer.



# Results

Model	$F_1 \pm \text{std}$
Chiu and Nichols (2016)	$90.91 \pm 0.20$
Lample et al. (2016)	90.94
Ma and Hovy (2016)	91.37
Our baseline without LM	$90.87 \pm 0.13$
TagLM	<b><math>91.93 \pm 0.19</math></b>

**Figure:** Performance on NER using CoNLL 2003 data and unlabeled text. [Matthew et al., 2017]



# Results

Model	$F_1 \pm \text{std}$
Yang et al. (2017)	94.66
Hashimoto et al. (2016)	95.02
Søgaard and Goldberg (2016)	95.28
Our baseline without LM	$95.00 \pm 0.08$
TagLM	<b><math>96.37 \pm 0.05</math></b>

**Figure:** Performance on chunking using CoNLL 2000 data and unlabeled text. [Matthew et al., 2017]



# Conclusions

- The language model transfer across domains.
- The model is robust even when trained on a large number of labeled data.
- Training the sequence tagging model and language model together increases performance.





# Graph-based

## **Paper:** Scientific Information Extraction with Semi-supervised Neural Tagging Steps

- Construct a graph of tokens based on their semantic similarity.
- Use the CRF marginal as a regularization term to do label propagation on the graph.
- The smoothed posterior is then used to either interpolate with the CRF marginal or as an additional feature to the neural network.



# Graph-based

## Graph Construction

$$w_{uv} = d_e(u, v) \text{ if } v \in K(u) \text{ or } u \in K(v).$$

## Label Propagation

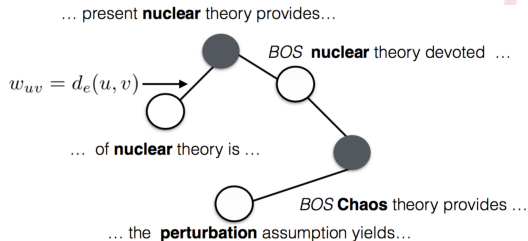


Figure: Label Propagation. [Luan et al., 2017]

# Graph-based

## Uncertain Label Marginalizing

$$\mathcal{Y}(x_t) = \begin{cases} \{y_t\} & \text{if } p(y_t|x; \theta) > \eta \\ \text{All label types} & \text{otherwise} \end{cases} \quad (3)$$

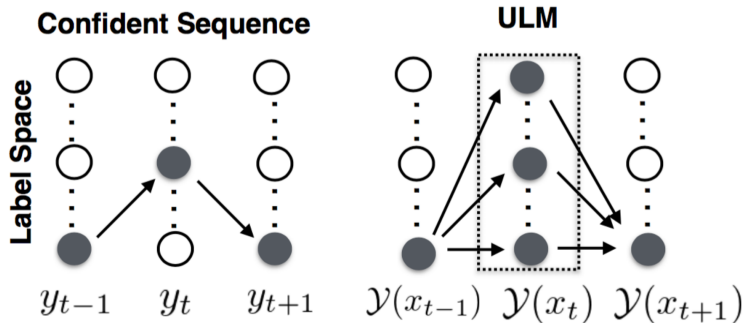
## Score

$$\phi(y; x, \theta) = \sum_{t=0}^n T_{y_t, y_{t+1}} + \sum_{t=1}^n P_{t, y_t} \quad (4)$$

## Probability

$$p_{\theta}(\mathcal{Y}(x^k)|x^k) = \frac{\sum_{y^k \in \mathcal{Y}(x^k)} \exp(\phi(y^k; x^k, \theta))}{\sum_{y' \in Y} \exp(\phi(y'; x, \theta))} \quad (5)$$

# Graph-based



**Figure:** Lattice representation of UML. [Luan et al., 2017]



# Results

Posterior	Training	Dev	Test
-	-	50.2	42.9
-	ULM	51.3	44.4
GRAPHINTERP	-	50.9	43.3
GRAPHINTERP	ULM	<b>51.9</b>	<b>45.3</b>
GRAPHINTERP*	-	50.7	44.0
GRAPHINTERP*	ULM	51.8	45.7
GRAPHFEAT*	-	51.4	44.9
GRAPHFEAT*	ULM	<b>52.1</b>	<b>46.6</b>

Figure: Results. [Luan et al., 2017]



# Conclusions

- In-domain data performs better than cross-domain data.
- The combination of in-domain data and ULM algorithms performs well.
- We can add language models into the model in the future to capture the context information.



# References

## Cross-domain Transfer

- **[ICLR17]** Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. (Yang et al.)
- **[ACL16]** Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning. (Peng et al.)
- **[Workshop17]** Multi-task Domain Adaptation for Sequence Tagging. (Peng et al.)

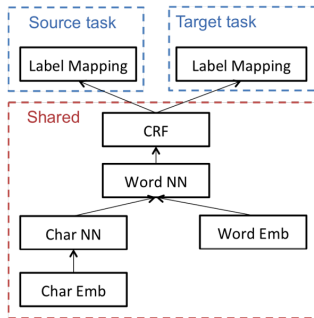
## Cross-lingual Transfer

- **[ICLR17]** Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. (Yang et al.)
- **[EMNLP17]** Cross-Lingual Transfer Learning for POS Tagging without Cross-Lingual Resources. (Kim et al.)

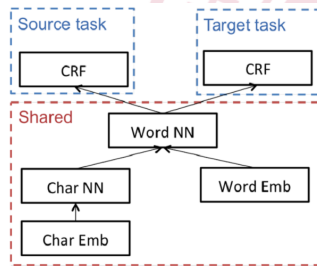


# Cross-domain Transfer

**Paper:** Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks



(a) Label mapping exist.



(b) Disparate label sets.

**Figure:** Two cross-domain transfer model. [Yang et al., 2017]





# Cross-domain Transfer

**Paper:** Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning.

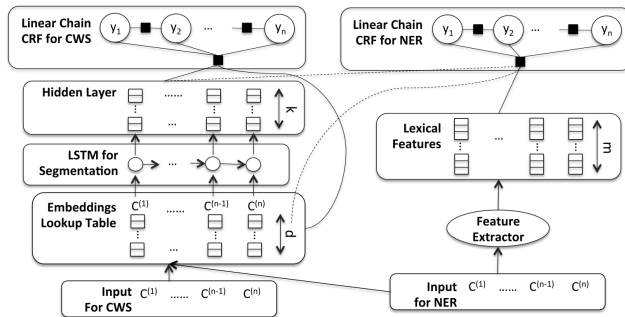


Figure: The joint model for CWS and NER. [Peng et al., 2016]



# Cross-domain Transfer

**Paper:** Multi-task Domain Adaptation for Sequence Tagging.

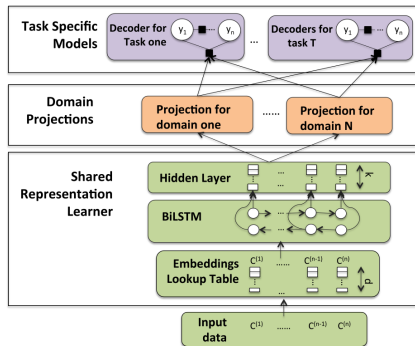


Figure: The Multi-task DA model. [Peng et al., 2017]



# Domain Projections

## Domain Masks

$$m_1 = [\vec{1}, \vec{1}, \vec{0}], m_2 = [\vec{1}, \vec{0}, \vec{1}]$$
$$\hat{h} = m_d \odot h$$

## Linear Projection

$$\hat{h} = T_d h$$

(6)

(7)



# Results

Settings	Datasets	CWS			NER		
	Methods	Prec	Recall	F1	Prec	Recall	F1
Baseline	Separate	86.2	85.7	86.0	57.2	42.1	48.5
	Mix	87.0	86.1	86.5	60.9	44.0	51.1
Domain Adapt	Domain Mask	88.7	87.1	87.9	68.2	48.6	56.8
	Linear Projection	88.0	87.5	87.7	73.3	45.8	56.4
Multi-task DA	Domain Mask	<b>89.7</b>	88.3	<b>89.0</b>	60.2	<b>52.3</b>	<b>59.9</b>
	Linear Projection	89.1	<b>88.6</b>	88.9	<b>68.6</b>	49.5	57.5

Figure: Results for CWS and NER. [Peng et al., 2017]



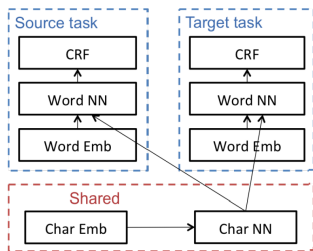
# Conclusions

- Multi-task learning can help domain adaptation.
- The number of shared parameters has great impact on the performance.
- We may use other domain adaptation methods besides parameter sharing and representation learning.



# Cross-lingual Transfer

**Paper:** Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks



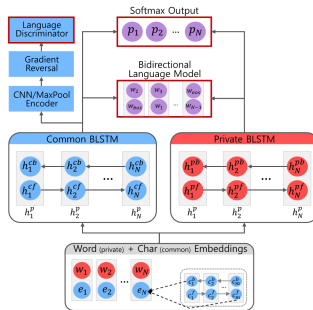
**Figure:** The cross-lingual transfer model. [Yang et al., 2017]

**Shortcoming:** The model isn't applicable when the input spaces differ.



# Cross-lingual Transfer

**Paper:** Cross-Lingual Transfer Learning for POS Tagging without Cross-Lingual Resources



**Figure:** The cross-lingual transfer model. [Kim et al., 2017]



# Cross-lingual Transfer

## Sequence Tagging Loss

$$\mathcal{L}_p = - \sum_{i=1}^S \sum_{j=1}^N p_{i,j} \log(\hat{p}_{i,j}) \quad (8)$$

## Language Classifier Loss

$$\mathcal{L}_a = - \sum_{i=1}^S l_i \log(\hat{l}_i) \quad (9)$$

## Bidirectional Language Model Loss

$$\mathcal{L}_l = - \sum_{i=1}^S \sum_{j=1}^N \log(P(w_{j+1}|f_j)) + \log(P(w_{j-1}|b_j)) \quad (10)$$





# Cross-lingual Transfer

## Total Loss

$$\mathcal{L} = w_s(\mathcal{L}_p + \lambda\mathcal{L}_a + \lambda\mathcal{L}_l) \quad (11)$$



# Results

Language Family	Language	Target only		Source (English) → Target				
		c	c,l	c,l	p,l	c,p,l	c,l+a	c,p,l+a
Germanic	Swedish	93.26	94.31	94.36	94.39	94.51	94.38	<b>94.63</b>
	Danish	92.13	93.41	93.34	93.76	94.05	93.74	<b>94.26</b>
	Dutch	83.24	84.73	85.20	84.92	84.85	84.99	<b>85.83</b>
	German	89.27	90.69	90.06	90.40	90.01	90.14	<b>90.71</b>
	Avg	89.47	90.78	90.74	90.87	90.86	90.82	<b>91.36</b>
Slavic	Slovenian	93.06	93.79	93.83	94.06	<b>94.20</b>	93.93	94.06
	Polish	91.30	91.30	91.69	<b>92.11</b>	91.86	91.77	<b>92.11</b>
	Slovak	86.53	89.56	90.11	89.88	89.98	<b>90.40</b>	90.01
	Bulgarian	93.45	95.27	95.33	95.50	95.52	95.25	<b>95.65</b>
	Avg	91.09	92.48	92.74	92.89	92.89	92.84	<b>92.95</b>
Romance	Romanian	93.20	94.09	<b>94.22</b>	94.17	94.05	93.91	94.20
	Portuguese	94.23	95.18	95.42	95.15	<b>95.55</b>	95.36	95.51
	Italian	93.80	<b>95.95</b>	95.79	95.61	95.84	95.70	95.92
	Spanish	91.94	93.34	93.34	93.31	93.29	92.94	<b>93.44</b>
	Avg	93.29	94.64	94.69	94.56	94.68	94.48	<b>94.77</b>
Indo-Iranian	Persian	93.91	94.63	94.68	94.79	94.78	94.49	<b>94.83</b>
Uralic	Hungarian	93.20	93.27	94.40	94.66	<b>94.69</b>	94.29	94.45
	Total Avg	91.61	92.82	92.98	93.05	93.08	92.95	<b>93.26</b>

Figure: POS tagging accuracies. [Kim et al., 2017]



# Conclusions

- The language classifier can train the common LSTM to be language-agnostic.
- Either too many or too little labeled data decrease the performance.
- Multiple source languages can be used to increase the performance.



# Conclusions

## 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



Xuezhe Ma and Eduard Hovy. (2016).

## **End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF.**

In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1064–1074, Berlin, Germany, August 7-12, 2016.



Marek Rei. (2017).

## **Semi-supervised Multitask Learning for Sequence Labeling.**

In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 2121–2130, Vancouver, Canada, July 30 - August 4, 2017.



# References

 Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, Russell Power. (2017).

**Semi-supervised Sequence Tagging with Bidirectional Language Models.**

In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1756–1765, Vancouver, Canada, July 30 - August 4, 2017.



 Yi Luan, Mari Ostendorf, Hannaneh Hajishirzi. (2017).

**Scientific Information Extraction with Semi-supervised Neural Tagging.**

In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2631–2641, Copenhagen, Denmark, September 7–11, 2017.



# References

-  Zhilin Yang, Ruslan Salakhutdinov, William W. Cohen. (2017).  
**Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks.**  
In ICLR 2017.
-  Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, Eric Fosler-Lussier. (2017).  
**Cross-Lingual Transfer Learning for POS Tagging without Cross-Lingual Resources.**  
In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2822–2828, Copenhagen, Denmark, September 7–11, 2017.
-  Nanyun Peng, Mark Dredze. (2017).  
**Multi-task Domain Adaptation for Sequence Tagging.**  
In Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 91–100, Vancouver, Canada, August 3, 2017.



# References

-  Amarnag Subramanya, Slav Petrov, Fernando Pereira. (2010).  
**Efficient Graph-Based Semi-Supervised Learning of Structured Tagging Models.**  
In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 167–176, MIT, Massachusetts, USA, 9-11 October 2010.
-  Yuzong Liu, Katrin Kirchhoff. (2014).  
**Graph-based Semi-supervised Acoustic Modeling in DNN-based Speech Recognition.**  
In IEEE SLT 2014.





# References



Nanyun Peng, Mark Dredze. (2016).

**Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning.**

In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 149–155, Berlin, Germany, August 7-12, 2016.