Semi-supervised sequence tagging with bidirectional language models

张健

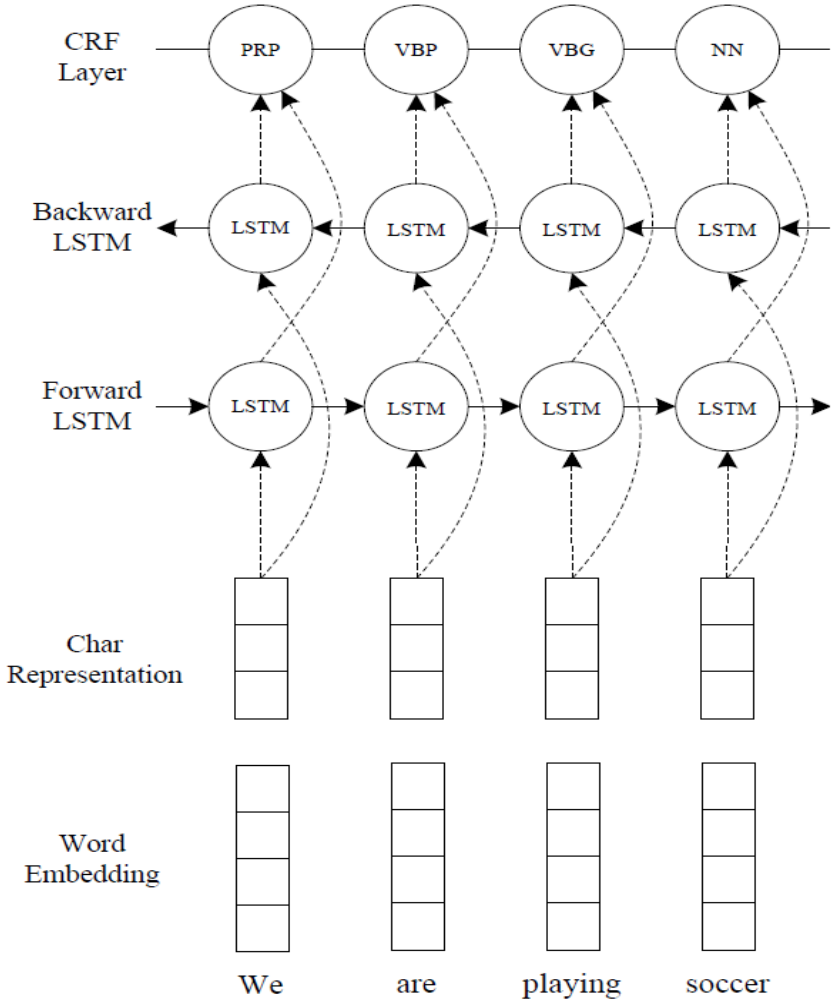
2018.10.18

Introduction

- The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

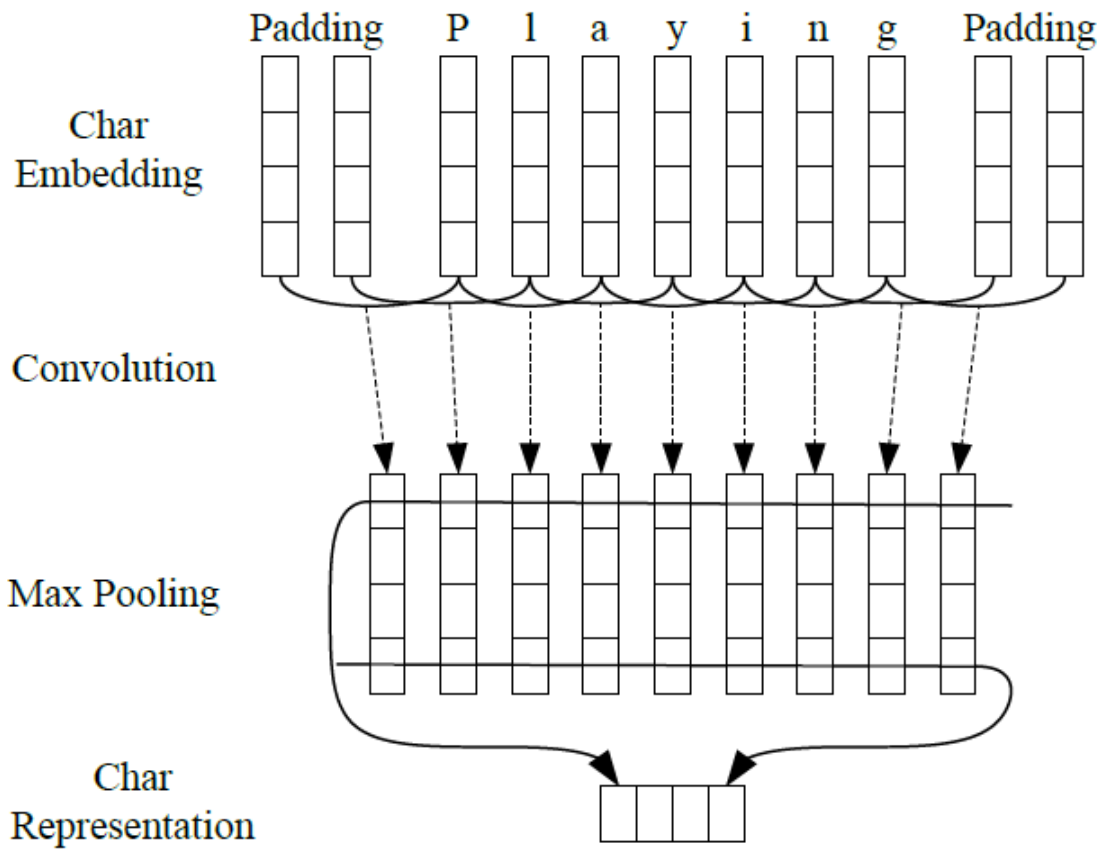
Person
Date
Location
Organi-
zation

Basic model

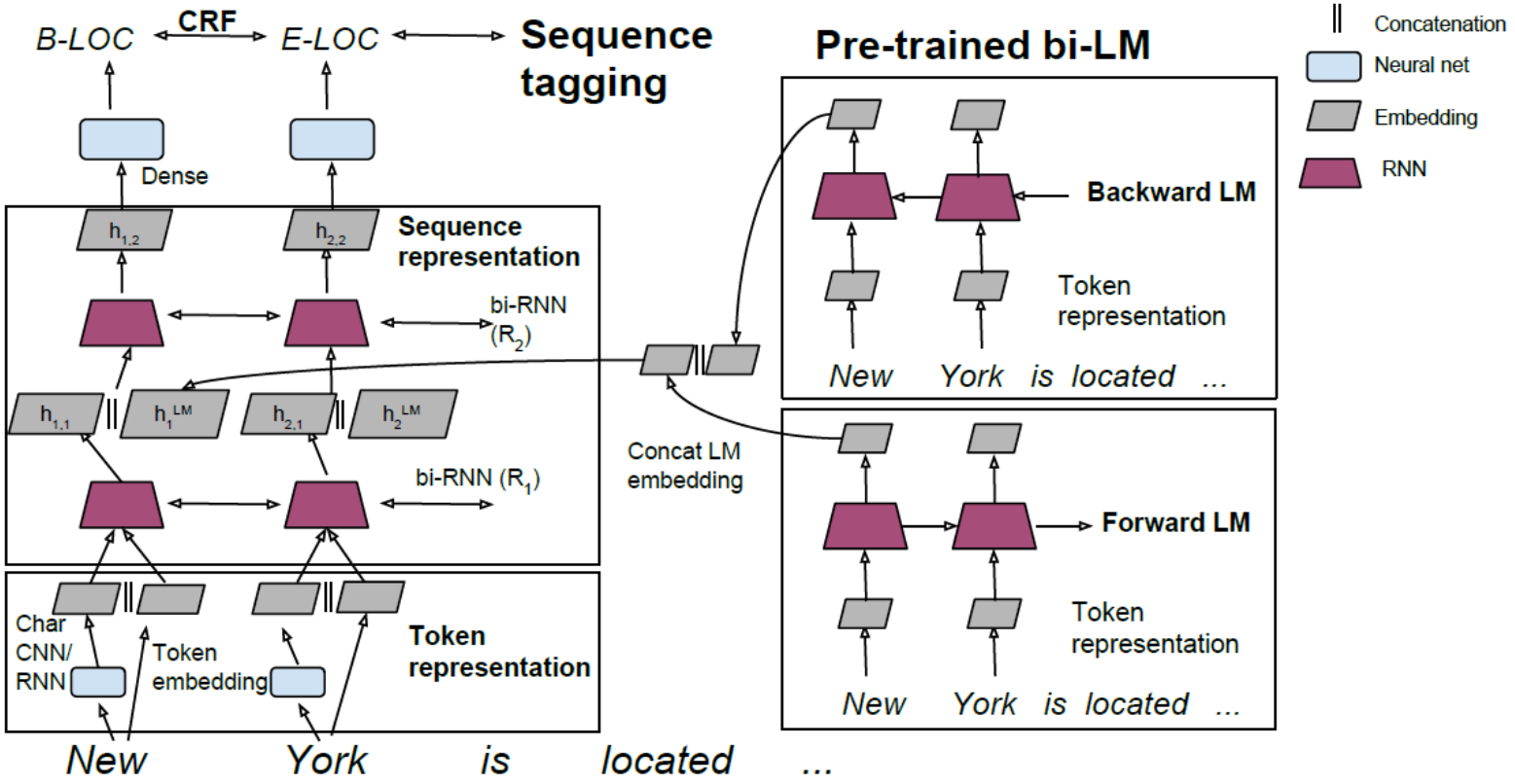


$$c_k = C(t_k; \Theta_c)$$
$$w_k = E(t_k; \Theta_w)$$
$$x_k = [c_k; w_k]$$

Basic model


Padding
P
l
a
y
i
n
g
Padding
Char Embedding
Convolution
Max Pooling
Char Representation

Language model augmented sequence taggers (TagLM)



Notations

- $\overrightarrow{h_{k,1}} = \overrightarrow{R_1}(x_k, \overrightarrow{h_{k-1,1}}; \Theta_{\overrightarrow{R_1}})$
- $\overleftarrow{h_{k,1}} = \overleftarrow{R_1}\ (x_k, \overleftarrow{h_{k+1,1}}; \Theta_{\overleftarrow{R_1}})$
- $h_{k,1} = [\overrightarrow{h_{k,1}}; \overleftarrow{h_{k,1}}; h_k^{LM}]$
- $h_{k,1} = f([\overrightarrow{h_{k,1}}; \overleftarrow{h_{k,1}}; h_k^{LM}])$

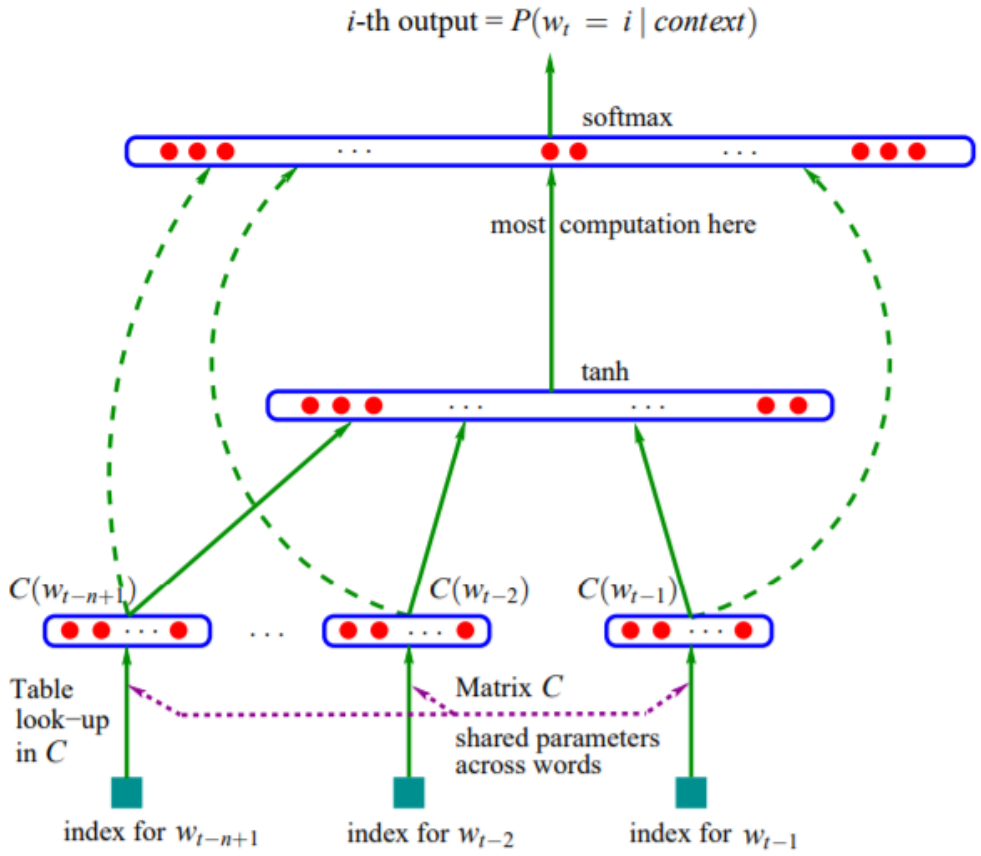
Language model

- $p(t_1, t_2, \dots, t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2, \dots, t_{k-1})$

- $p(t_1, t_2, \dots, t_N) = \prod_{k=1}^{N} p(t_k | t_{k-1})$
- $p(t_k | t_{k-1}) = \frac{p(t_k, t_{k-1})}{p(t_{k-1})} \approx \frac{count(t_k, t_{k-1})}{count(t_{k-1})}$

Neural language model

$$f(w_t, \ldots, w_{t-n+1}) = \hat{p}(w_t | t_1^{n-1})$$



A Neural Probabilistic Language Model (Bengio, et al., 2003)

Pre-trained language models

- Exploring the Limits of Language Modeling (Józefowicz et al. 2016)
- Trained on the 1B Word Benchmark
- Single best model took three weeks to train on 32 GPUs and achieved 30.0 test perplexity. It uses a character CNN with 4096 filters for input, followed by two stacked LSTMs, each with 8192 hidden units and a 1024 dimensional projection layer

Bidirectional LM

- Forward LSTM-8192-1024
- Backward LSTM-2048-512
- The forward and backward LMs are independent, without any shared parameters

Experiments

Model	$F_1 \pm$ **std**
Chiu and Nichols (2016)	90.91 ± 0.20
Lample et al. (2016)	90.94
Ma and Hovy (2016)	91.37
Our baseline without LM	90.87 ± 0.13
TagLM	$\mathbf{91.93 \pm 0.19}$

Overall system results

Experiments

Model	External resources	F_1 Without	F_1 With	Δ
Yang et al. (2017)	transfer from CoNLL 2000/PTB-POS	91.2	91.26	+0.06
Chiu and Nichols (2016)	with gazetteers	90.91	91.62	+0.71
Collobert et al. (2011)	with gazetteers	88.67	89.59	+0.92
Luo et al. (2015)	joint with entity linking	89.9	91.2	**+1.3**
Ours	no LM vs TagLM *unlabeled data only*	90.87	**91.93**	+1.06

Compare to include additional labeled data or gazetteers.

Experiments

Forward language model	Backward language model	LM perplexity		$F_1 \pm$ std
		Fwd	Bwd	
—	—	N/A	N/A	90.87 ± 0.13
LSTM-512-256*	LSTM-512-256*	106.9	104.2	90.79 ± 0.15
LSTM-2048-512	—	47.7	N/A	91.40 ± 0.18
LSTM-2048-512	LSTM-2048-512	47.7	47.3	91.62 ± 0.23
CNN-BIG-LSTM	—	30.0	N/A	91.66 ± 0.13
CNN-BIG-LSTM	LSTM-2048-512	30.0	47.3	$\mathbf{91.93 \pm 0.19}$

How to use LM embeddings?

Use LM embeddings at	$F_1 \pm$ std
input to the first RNN layer	91.55 ± 0.21
output of the first RNN layer	$\mathbf{91.93 \pm 0.19}$
output of the second RNN layer	91.72 ± 0.13

- augment the input of the first RNN layer;

$$x_k = [c_k; w_k; h_k^{LM}]$$

- augment the output of the first RNN layer;

$$h_{k,1} = [\overrightarrow{h_{k,1}}; \overleftarrow{h_{k,1}}; h_k^{LM}]$$

- augment the output of the second RNN layer;

$$h_{k,2} = [\overrightarrow{h_{k,2}}; \overleftarrow{h_{k,2}}; h_k^{LM}]$$

Dataset size

- A hypothesis, addition of LM embeddings to be most beneficial in cases where the task specific annotated datasets are small
- Samples 1% of the CoNLL 2003 training set
- Test F1 increased 3.35% (from 67.66 to 71.01%)

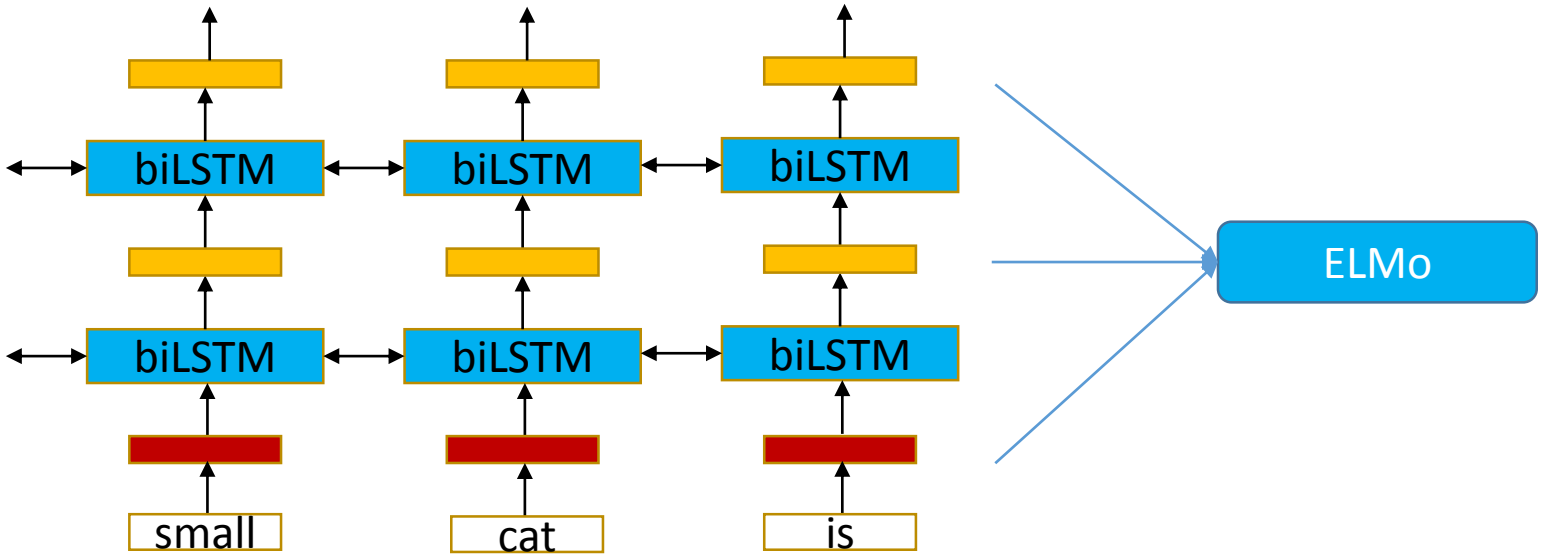
Does the LM transfer across domains?

- Both CoNLL 2003 and 1 Billion Word Benchmark are derived from news articles
- Applied TagLM SemEval 2017 Shared Task 10, ScienceIE
- Requires end-to-end joint entity and relationship extraction
- Scientific publications across computer science, material sciences, and physics. Defines three broad entity types (Task, Material and Process)
- TagLM increased F1 on the development set by 4.12% (from 49.93 to to 54.05%) ranked first

ELMo(Embeddings from Language Models)

Deep contextualized word representations(NAACL 2018)

$$\text{ELMo}_k^{task} = E\left(R_k; \Theta^{task}\right) = \Upsilon^{task} \sum_{j=0}^{L} s_j^{task} h_{k,j}^{LM}$$



ELMo(Embeddings from Language Models)

Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

Conclusions

Contributions:

1. Significantly outperforms current state of the art models in two popular datasets for NER and Chunking
2. biLM layers efficiently encode different types of syntactic and semantic information
3. Improves performance when labeled datasets are small

Conclusions

Shortages:

1. The model is well complicated
2. It is a language dependent method

Improvements:

1. Simply concatenate the LM embeddings with sequence model
2. Applying LM embedding to other NLP tasks

THANK YOU