Named Entity Recognition with Bidirectional LSTM-CNNs

Zhen Cheng

Named Entity Recognition

- Input: x

  sentence/list of words
- Output: y*

  list of tags(e.g. BIOES, BIO, YN, etc)
- Definition:

$$y^* = \underset{y \in \text{GEN}(x)}{\arg\max} \left(\sum_{i=1}^{n} \text{score}(y_i | y_1, \cdots, y_{i-1}) \right)$$

Named Entity Recognition

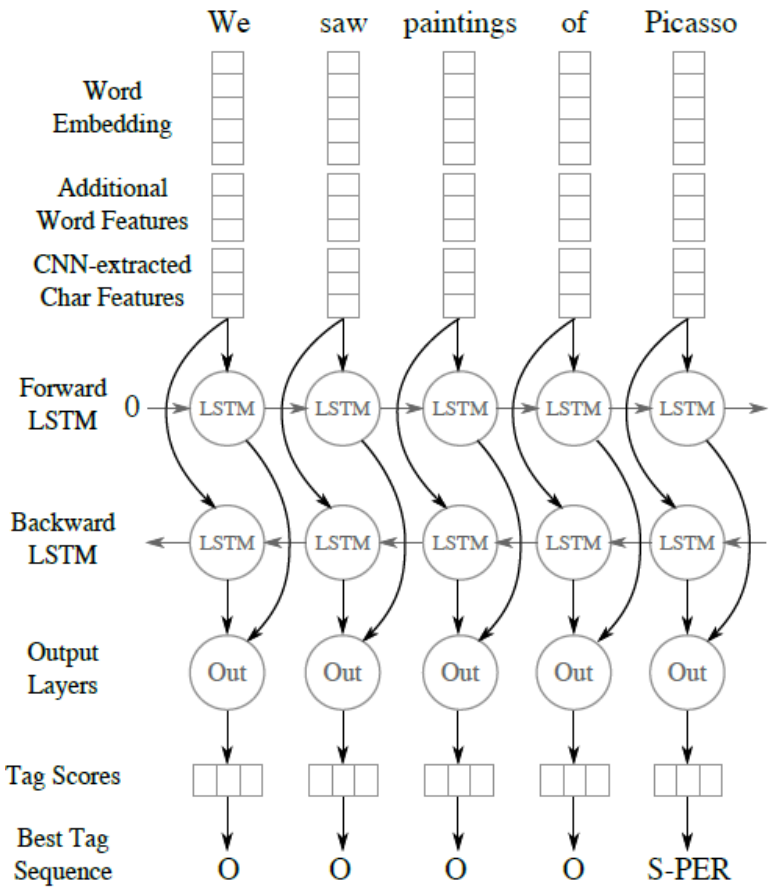
Previous approaches:

- Traditional methods:
  - SVM
  - Perceptron
  - Conditional Random Fields (Lafferty et al., 2001)

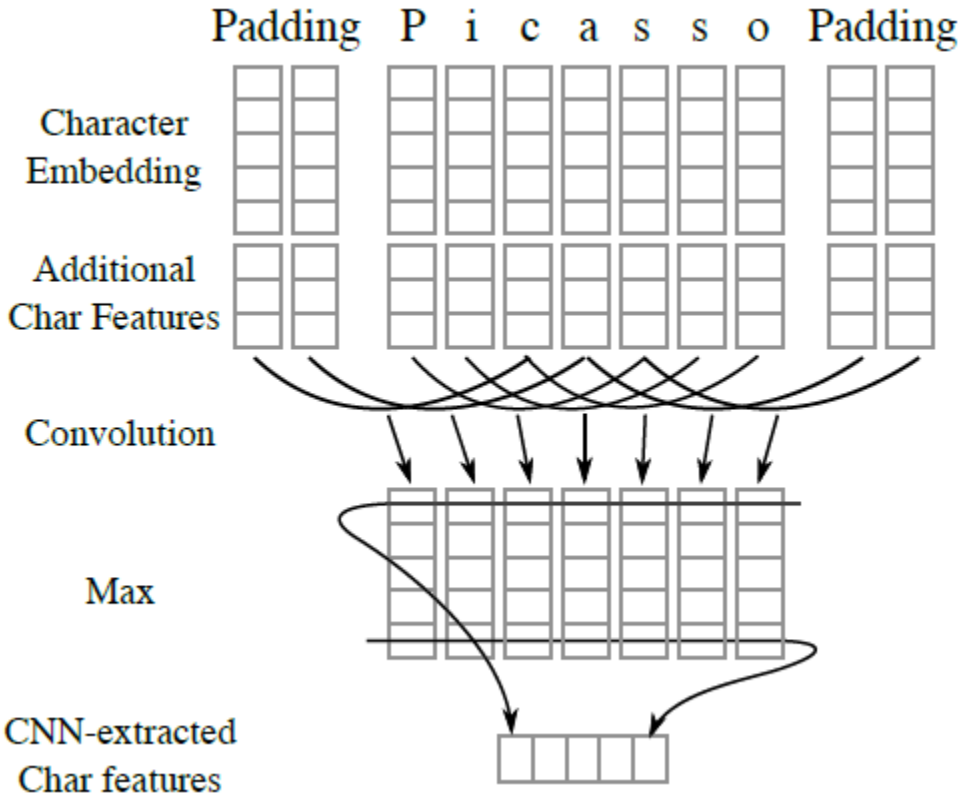
  Disadvantage: depends on the choice of features
- Previous neural network model:
  - Depends on word embedding
  - Fixed size local windows

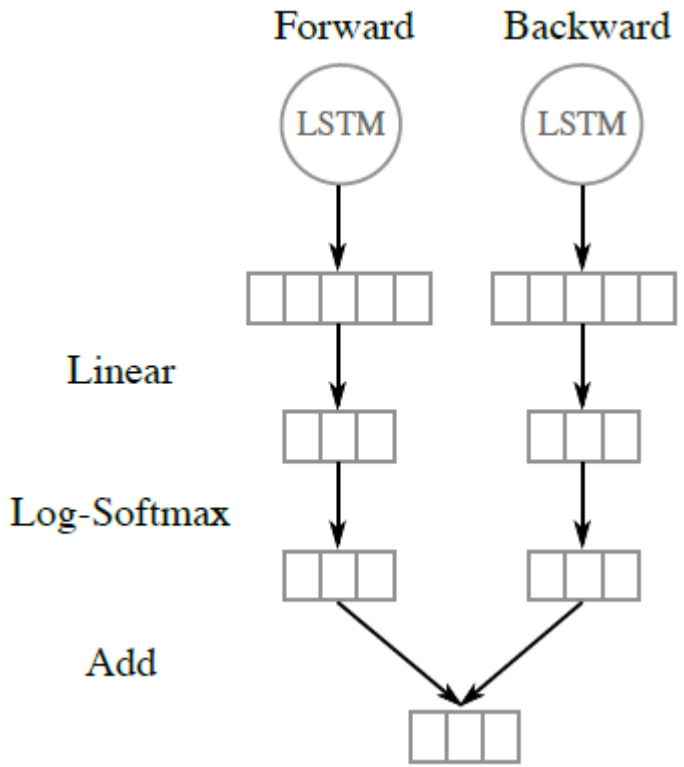
  Disadvantages:
  - Unable to capture the character level features
  - Only contextual information

BiLSTM-CNN Model



BiLSTM-CNN Model





BiLSTM-CNN Model

- Contributions:
  - Automatically detect both words and character features
  - Propose a new method of encoding partial lexicons
- Features:
  - Word embedding using Collobert et al. (2011b)
  - Add capitalization feature & Lexicons
  - Extracting character features using a CNN
  - Sequence-labeling with BiLSTM
- Dataset:
  - CoNLL-2003
  - OntoNotes 5.0

Word Embedding

- Using 50-dimensional word embedding by Collobert et al. (2011b)
  - Stanford's GloVe embeddings
  - Google's word2vec embeddings - case sensitive and fewer punctuations
- All words are lower-cased before passing through the lookup table
- The pre-trained embeddings are allowed to be modified during training

Word Embeddings	CoNLL-2003	OntoNotes
Random 50d	87.77 (± 0.29)	83.82 (± 0.19)
Random 300d	87.84 (± 0.23)	83.76 (± 0.37)
GloVe 6B 50d	91.09 (± 0.15)	86.25 (± 0.24)
GloVe 6B 300d	90.71 (± 0.21)	86.26 (± 0.30)
Google 100B 300d	90.60 (± 0.23)	85.34 (± 0.25)
Collobert 50d	**91.62** (± 0.33)	**86.28** (± 0.26)
Our GloVe 50d	91.41 (± 0.21)	86.24 (± 0.35)
Our Skip-gram 50d	90.76 (± 0.23)	85.70 (± 0.29)

Add capitalization feature

- Using a separate lookup table to add a capitalization feature
  - ALLCAPS
  - Upperinitial
  - lowercase
  - MixedCaps
  - Noinfo
- Compared with character type features & character-level CNN

Lexicons

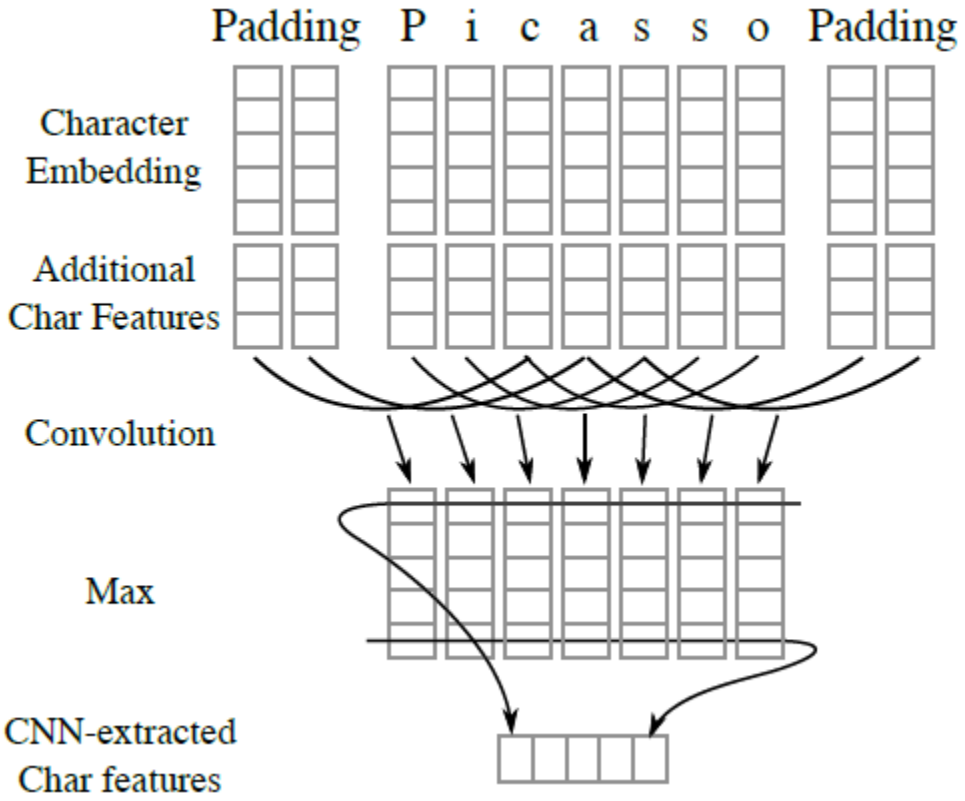
- SENNA lexicons & DBpedia lexicons

Category	SENNA	DBpedia
Location	36,697	709,772
Miscellaneous	4,722	328,575
Organization	6,440	231,868
Person	123,283	1,074,363
Total	171,142	2,344,578

Text	Hayao	Tada	,	commander	of	the	Japanese	North	China	Area	Army
LOC	–	–	–	–	–	B	I	–	S	–	–
MISC	–	–	–	S	B	B	I	S	S	S	S
ORG	–	–	–	–	–	B	I	B	I	I	E
PERS	B	E	–	–	–	–	–	–	S	–	–

Extracting character features using a CNN

- 25-dimensions character embedding
  - Padding both sides of the words
  - Randomly initialize using uniform distribution in [-0.5, 0.5]
  - Add PADDING & UNKNOWN tokens
- Character type
  - Upper case
  - Lower case
  - Punctuation
  - other



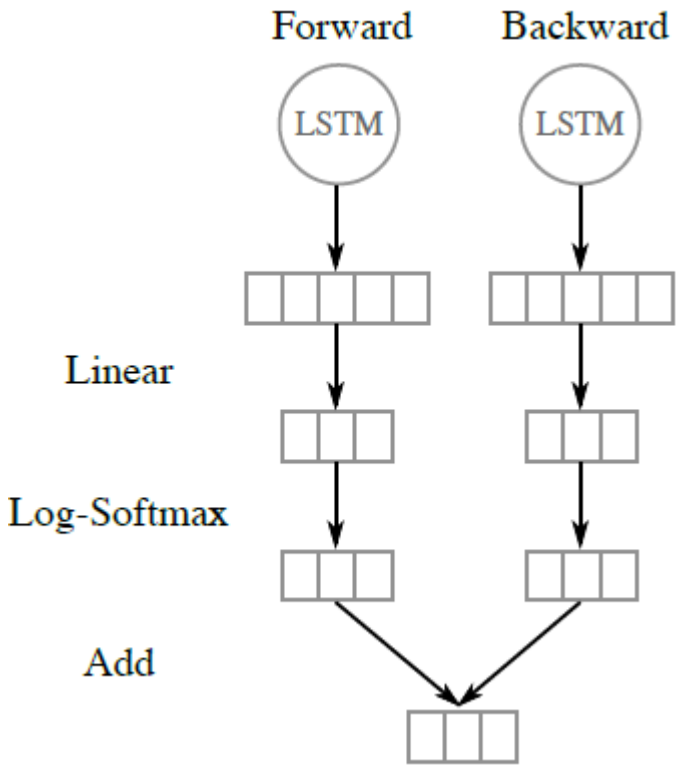
Extracting character features using a CNN

- Character-level BiLSTM is not better than CNN and has much more time consuming
- Character-level CNNs can replace hand-crafted character features in some cases
- But systems with weak lexicons may benefit from character features

Features	BLSTM		BLSTM-CNN		BLSTM-CNN + lex	
	CoNLL	OntoNotes	CoNLL	OntoNotes	CoNLL	OntoNotes
none	76.29 (± 0.29)	77.77 (± 0.37)	83.38 (± 0.20)	82.53 (± 0.40)	87.77 (± 0.29)	83.82 (± 0.19)
emb	88.23 (± 0.23)	82.72 (± 0.23)	90.91 (± 0.20)	86.17 (± 0.22)	**91.62** (± 0.33)	86.28 (± 0.26)
emb + caps	90.67 (± 0.16)	86.19 (± 0.25)	90.98 (± 0.18)	86.35 (± 0.28)	91.55 (± 0.19)*	86.28 (± 0.32)*
emb + caps + lex	**91.43** (± 0.17)	**86.21** (± 0.16)	**91.55** (± 0.19)*	86.28 (± 0.32)*	91.55 (± 0.19)*	86.28 (± 0.32)*
emb + char	-	-	90.88 (± 0.48)	86.08 (± 0.40)	91.44 (± 0.23)	**86.34** (± 0.18)
emb + char + caps	-	-	90.88 (± 0.31)	**86.41** (± 0.22)	91.48 (± 0.23)	86.33 (± 0.26)

Sequence-labeling with BiLSTM

- A BiLSTM model can take into account an effectively infinite amount of context on both sides of a word.



Training

- Preprocessing
  - Digit sequences -> 0
  - Group sentences by word length
  - Shuffle
- Hyper-parameter Optimization(Two rounds)
  - Perform random search
  - Perform Optunity's implementation of particle swarm

Hyper-parameter	CoNLL-2003 (Round 2)		OntoNotes 5.0 (Round 1)	
	Final	Range	Final	Range
Convolution width	**3**	$[3, 7]$	**3**	$[3, 9]$
CNN output size	**53**	$[15, 84]$	**20**	$[15, 100]$
LSTM state size	**275**	$[100, 500]$	**200**	$[100, 400]$[10]
LSTM layers	**1**	$[1, 4]$	**2**	$[2, 4]$
Learning rate	**0.0105**	$[10^{-3}, 10^{-1.8}]$	**0.008**	$[10^{-3.5}, 10^{-1.5}]$
Epochs[11]	**80**	-	**18**	-
Dropout[12]	**0.68**	$[0.25, 0.75]$	**0.63**	$[0, 1]$
Mini-batch size	**9**	-[13]	**9**	$[5, 14]$

Round	CoNLL-2003	OntoNotes 5.0
1	93.82 ($\pm$ 0.15)	**84.57** ($\pm$ 0.27)
2	**94.03** ($\pm$ 0.23)	84.47 ($\pm$ 0.29)

Training

- Labeling score:

$$S([x]_1^T, [i]_1^T, \theta') = \sum_{t=1}^{T} \left(A_{[i]_{t-1},[i]_t} + [f_\theta]_{[i]_t,t}\right)$$

- Objective function:

$$\log P([y]_1^T \mid [x]_1^T, \theta')$$
$$= S([x]_1^T, [y]_1^T, \theta') - \log \sum_{\forall [j]_1^T} e^{S([x]_1^T, [j]_1^T, \theta')}$$

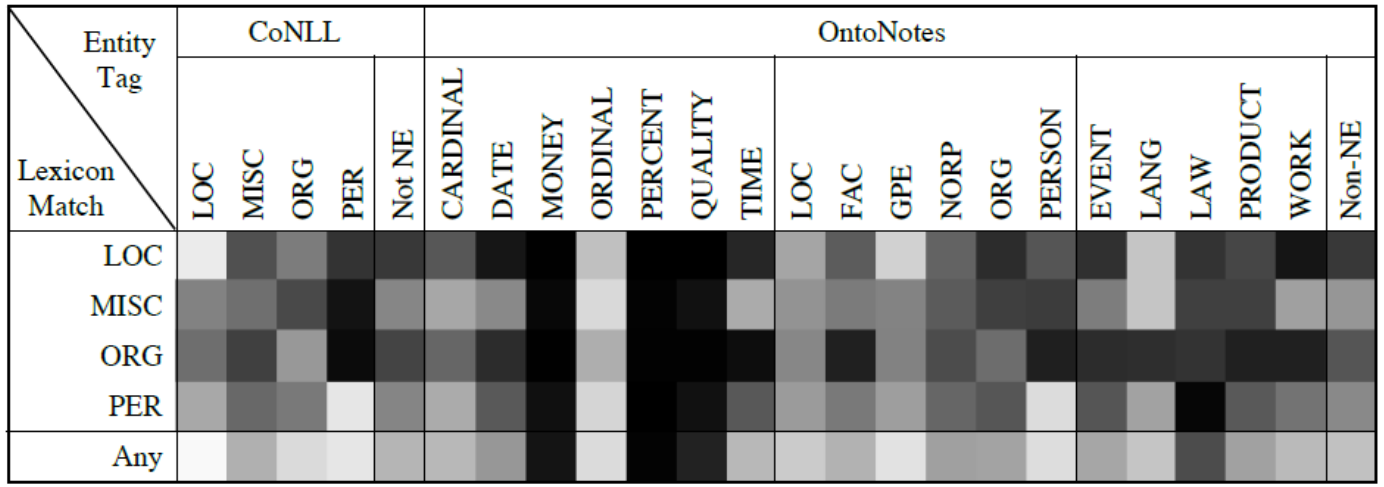
- Inference: Viterbi
- Method: SGD + dropout
- Parameters update: fixed learning rate

Dropout	CoNLL-2003		OntoNotes 5.0	
	Dev	Test	Dev	Test
-	93.72 (± 0.10)	90.76 (± 0.22)	82.02 (± 0.49)	84.06 (± 0.50)
0.10	93.85 (± 0.18)	90.87 (± 0.31)	83.01 (± 0.39)	84.94 (± 0.25)
0.30	94.08 (± 0.17)	91.09 (± 0.18)	83.61 (± 0.32)	85.44 (± 0.33)
0.50	94.19 (± 0.18)	91.14 (± 0.35)	84.35 (± 0.23)	**86.36** (± 0.28)
0.63	-	-	84.47 (± 0.23)	86.29 (± 0.25)
0.68	**94.31** (± 0.15)	**91.23** (± 0.16)	-	-
0.70	**94.31** (± 0.24)	91.17 (± 0.37)	**84.56** (± 0.40)	86.17 (± 0.25)
0.90	94.17 (± 0.17)	90.67 (± 0.17)	81.38 (± 0.19)	82.16 (± 0.18)

BiLSTM-CNN Exception

- According to different feature set trials failed:
  - CoNLL: 5-10%
  - OntoNotes: 1.5%
- Lower learning rate reduces failure rate

Performance



Lexicon	Matching	Encoding	CoNLL-2003	OntoNotes
No lexicon	-	-	83.38 (± 0.20)	82.53 (± 0.40)
SENNA	Exact	YN	86.21 (± 0.39)	83.24 (± 0.33)
	Exact	BIOES	86.14 (± 0.48)	83.01 (± 0.52)
DBpedia	Exact	YN	84.93 (± 0.30)	83.15 (± 0.26)
	Exact	BIOES	85.02 (± 0.23)	83.39 (± 0.39)
	Partial	YN	85.72 (± 0.45)	83.25 (± 0.33)
	Partial	BIOES	86.18 (± 0.56)	**83.97** (± 0.38)
	Collobert's method		85.01 (± 0.31)	83.24 (± 0.26)
Both	Best combination		**87.77** (± 0.29)	83.82 (± 0.19)

Performance

Model	CoNLL-2003			OntoNotes 5.0		
	Prec.	Recall	F1	Prec.	Recall	F1
FFNN + emb + caps + lex	89.54	89.80	89.67 (± 0.24)	74.28	73.61	73.94 (± 0.43)
BLSTM	80.14	72.81	76.29 (± 0.29)	79.68	75.97	77.77 (± 0.37)
BLSTM-CNN	83.48	83.28	83.38 (± 0.20)	82.58	82.49	82.53 (± 0.40)
BLSTM-CNN + emb	90.75	91.08	90.91 (± 0.20)	85.99	86.36	86.17 (± 0.22)
BLSTM-CNN + emb + lex	91.39	**91.85**	**91.62** (± 0.33)	**86.04**	**86.53**	**86.28** (± 0.26)
Collobert et al. (2011b)	-	-	88.67	-	-	-
Collobert et al. (2011b) + lexicon	-	-	89.59	-	-	-
Huang et al. (2015)	-	-	90.10	-	-	-
Ratinov and Roth (2009)[18]	91.20	90.50	90.80	82.00	84.95	83.45
Lin and Wu (2009)	-	-	90.90	-	-	-
Finkel and Manning (2009)[19]	-	-	-	84.04	80.86	82.42
Suzuki et al. (2011)	-	-	91.02	-	-	-
Passos et al. (2014)[20]	-	-	90.90	-	-	82.24
Durrett and Klein (2014)	-	-	-	85.22	82.89	84.04
Luo et al. (2015)[21]	**91.50**	91.40	91.20	-	-	-