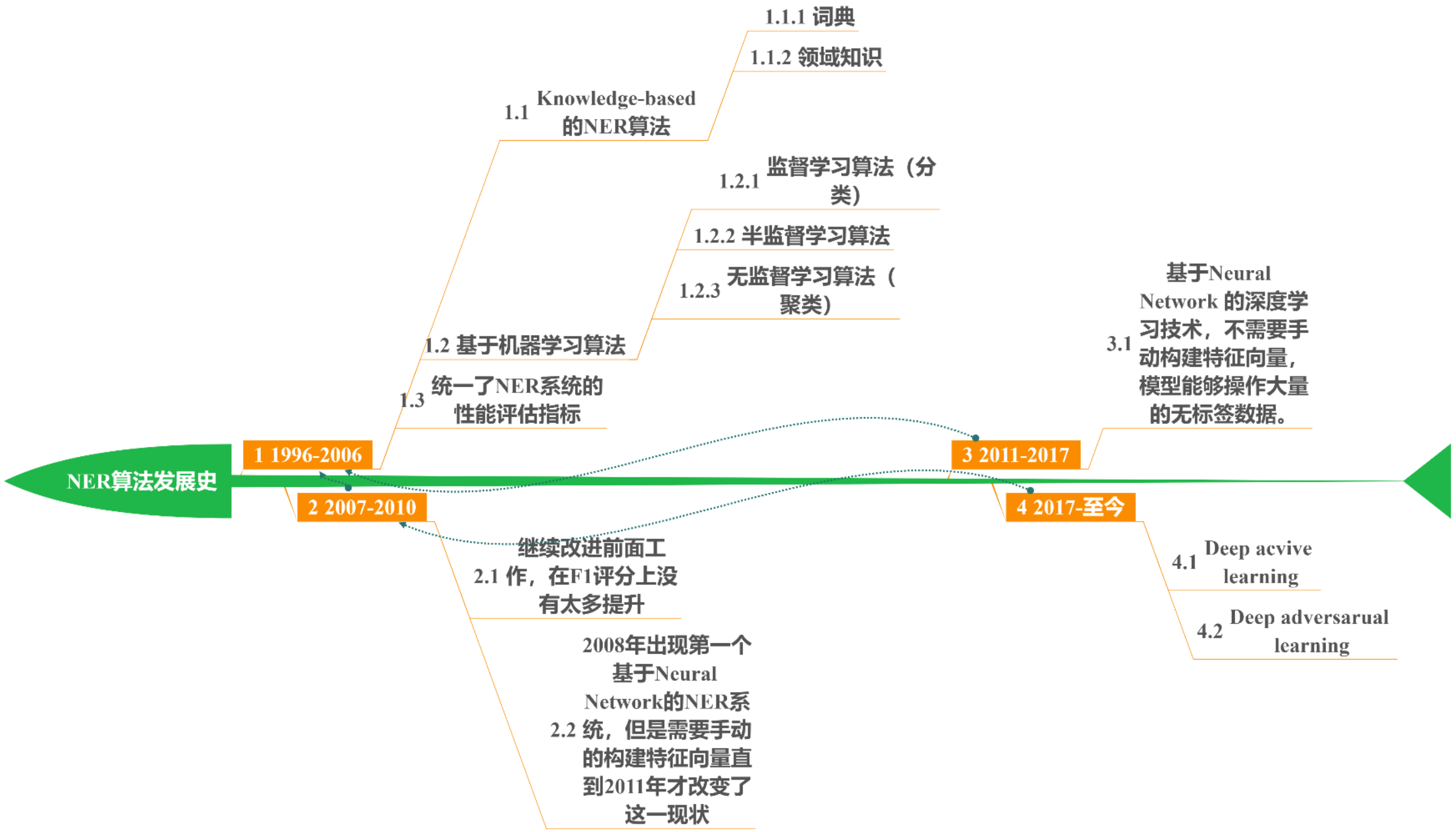
A Survey in the Named Entity Recognition Focus on Deep Neural Network, Active Learning and Adversarial Learning

- **Presented by: Junpeng Zhu**
- Jan 2, 2019
- **ID：52184506004**

OUTLINE

- Introduction
- Methodology
- Models
- Conclusions
- Opinions

INTRODUCTION



NER算法发展史

1 1996-2006

1.1 Knowledge-based 的NER算法

1.1.1 词典

1.1.2 领域知识

1.2 基于机器学习算法

1.2.1 监督学习算法（分类）

1.2.2 半监督学习算法

1.2.3 无监督学习算法（聚类）

1.3 统一了NER系统的性能评估指标

2 2007-2010

2.1 继续改进前面工作，在F1评分上没有太多提升

2.2 2008年出现第一个基于Neural Network的NER系统，但是需要手动的构建特征向量直到2011年才改变了这一现状

3 2011-2017

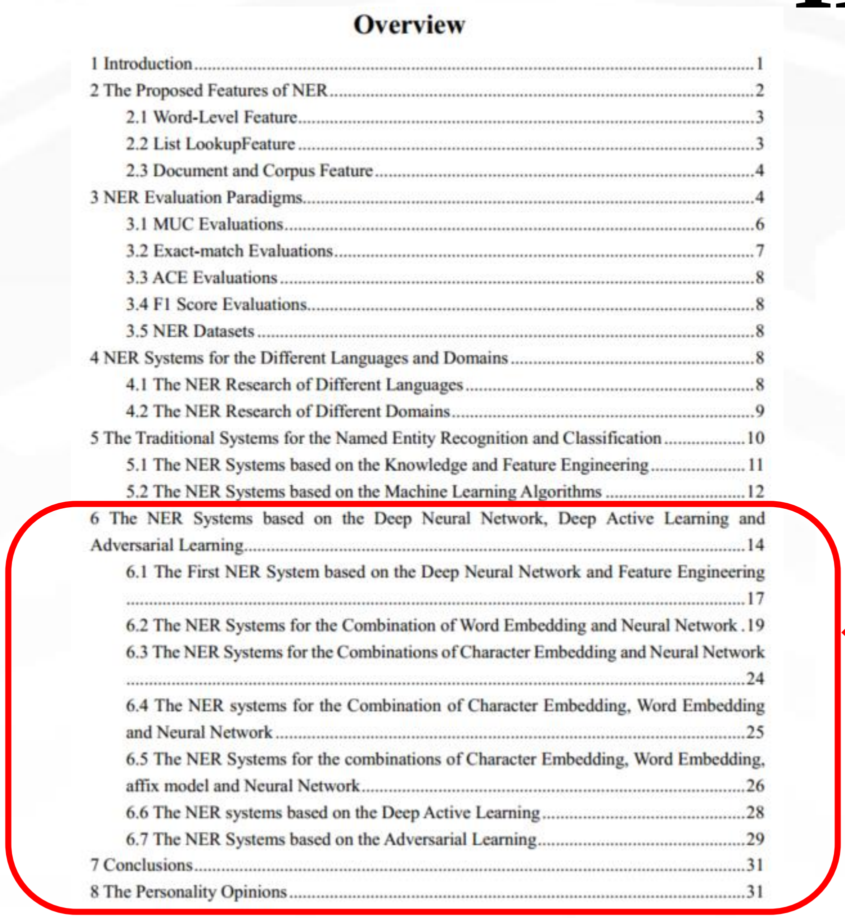
3.1 基于Neural Network 的深度学习技术，不需要手动构建特征向量，模型能够操作大量的无标签数据。

4 2017-至今

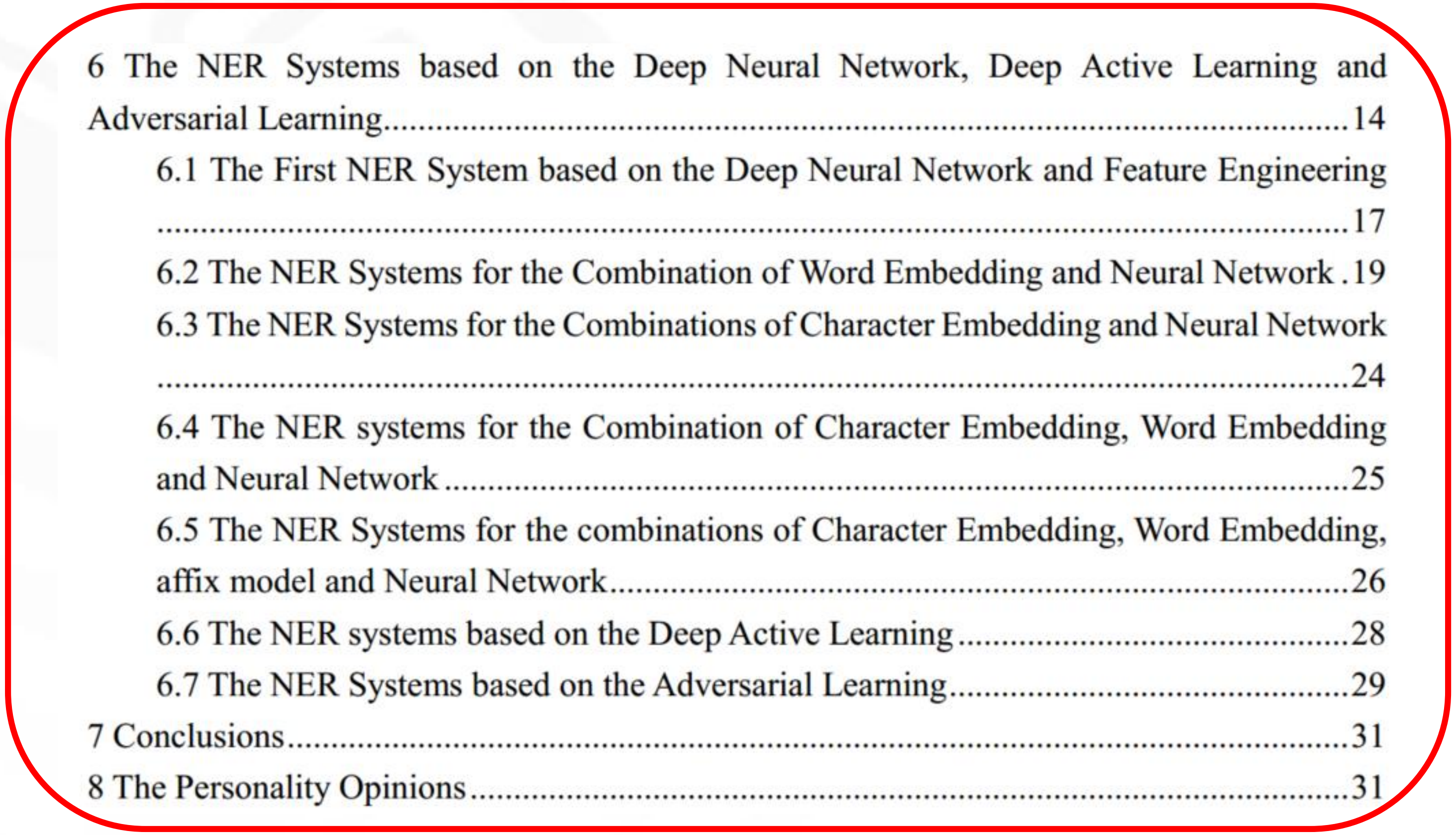
4.1 Deep acvive learning

4.2 Deep adversarual learning

INTRODUCTION



Overview



6 The NER Systems based on the Deep Neural Network, Deep Active Learning and Adversarial Learning 14
6.1 The First NER System based on the Deep Neural Network and Feature Engineering 17
6.2 The NER Systems for the Combination of Word Embedding and Neural Network . 19
6.3 The NER Systems for the Combinations of Character Embedding and Neural Network 24
6.4 The NER systems for the Combination of Character Embedding, Word Embedding and Neural Network 25
6.5 The NER Systems for the combinations of Character Embedding, Word Embedding, affix model and Neural Network 26
6.6 The NER systems based on the Deep Active Learning 28
6.7 The NER Systems based on the Adversarial Learning 29
7 Conclusions 31
8 The Personality Opinions 31

INTRODUCTION

- **Named Entity Recognition** (NER) is **a key component** in **NLP systems**
- Accurate systems **using deep neural networks** (DNN) have only been **introduced in the last few years**
- We present a comprehensive survey of **deep neural network** , **deep active learning** and **adversarial learning** architectures, experimental results, advantages and disadvantages for NER systems

OUTLINE

- Introduction
- Methodology
- Models
- Conclusions
- Opinions

METHODOLOGY

- **named entity recognition**, **neural network** named entity recognition models, **deep learning models** for named entity recognition**, deep active learning** for the named entity recognition and **deep adversarial learning** for the named entity recognition
- In total, **44 articles** were reviewed and were selected for the survey.

OUTLINE

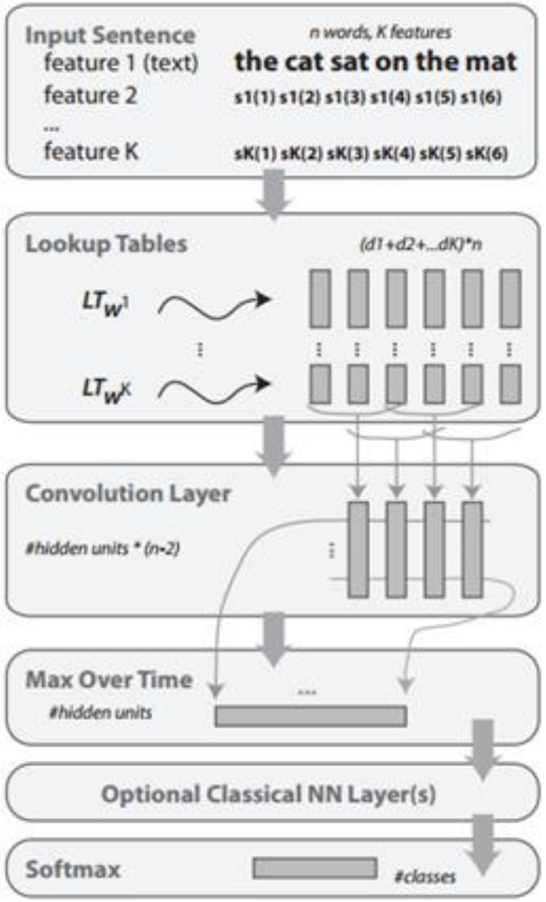
- Introduction
- Methodology
- Models
- Conclusions
- Opinions

MODELS

- The First NER System based on the **Deep Neural Network** and **Feature Engineering**
- The NER Systems for the Combination of **Word Embedding** and **Neural Network**
- The NER Systems for the Combinations **of Character Embedding** and **Neural Network**
- The NER systems for the Combination of **Character Embedding**, **Word Embedding** and **Neural Network**
- The NER Systems for the combinations of **Character Embedding**, **Word Embedding**, **affix** model and **Neural Network**
- The NER systems based on the **Deep Active Learning**
- The NER Systems based on the **Adversarial Learning**

The First NER System based on the Deep Neural Network and Feature Engineering

- The first layer extracts features for each word. The first layer has to map words into real-valued vectors.
- The second layer extracts features from the sentence treating it as a sequence with local and global structure (i.e., it is not treated like a bag of words).
- The following layers are classical NN layers. A general deep NN architecture for NLP. Given an input sentence, the NN outputs class probabilities for one chosen word.



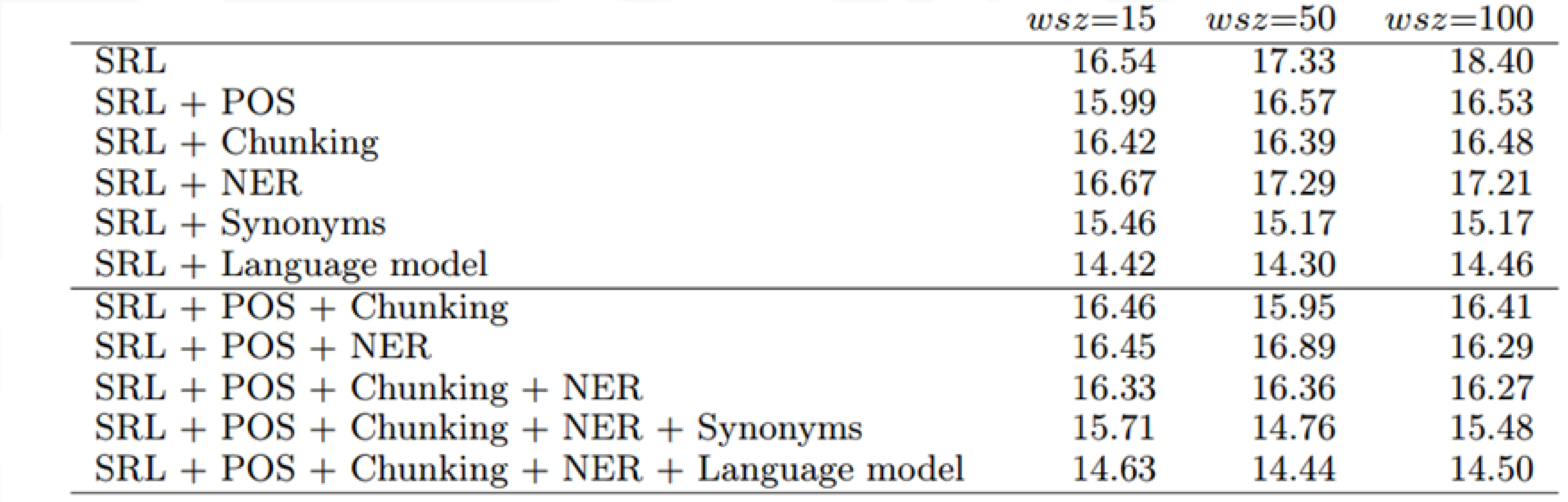
Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on Machine learning （ICML'08） . ACM, 2008: 160-167.

The First NER System based on the Deep Neural Network and Feature Engineering

- ER labeled data was obtained by running the Stanford Named Entity Recognizer over the PropBank dataset（https://propbank.github.io/） version 1 (about 1 million words). It uses the dictionary of the 30, 000 most common words from Wikipedia, converted to lower case. Other words were considered as unknown and mapped to a special word.

Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on Machine learning （ICML'08）. ACM, 2008: 160-167.

The First NER System based on the Deep Neural Network and Feature Engineering



	wsz=15	wsz=50	wsz=100
SRL	16.54	17.33	18.40
SRL + POS	15.99	16.57	16.53
SRL + Chunking	16.42	16.39	16.48
SRL + NER	16.67	17.29	17.21
SRL + Synonyms	15.46	15.17	15.17
SRL + Language model	14.42	14.30	14.46
SRL + POS + Chunking	16.46	15.95	16.41
SRL + POS + NER	16.45	16.89	16.29
SRL + POS + Chunking + NER	16.33	16.36	16.27
SRL + POS + Chunking + NER + Synonyms	15.71	14.76	15.48
SRL + POS + Chunking + NER + Language model	14.63	14.44	14.50

Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on Machine learning （ICML'08） . ACM, 2008: 160-167.

The First NER System based on the Deep Neural Network and Feature Engineering

Advantages

- Making full use of huge data sets
- The model is the better for generalization than others which based on the combination of feature-engineering and machine learning algorithms
- This is an important result, given that the NLP community considers syntax as a mandatory feature for semantic extraction

Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on Machine learning （ICML'08） . ACM, 2008: 160-167.

The First NER System based on the Deep Neural Network and Feature Engineering

Disadvantages

- The model is not avoid to use the hand-constructed feature vector which based on the feature-engineer, which manually constructed feature vectors from orthographic features (e.g., capitalization of the first character), dictionaries and lexicons.
- There is no F1 score results for the combination of NER and deep neural network
- A large number of hand-labeled data is necessary for the model, which is limited.

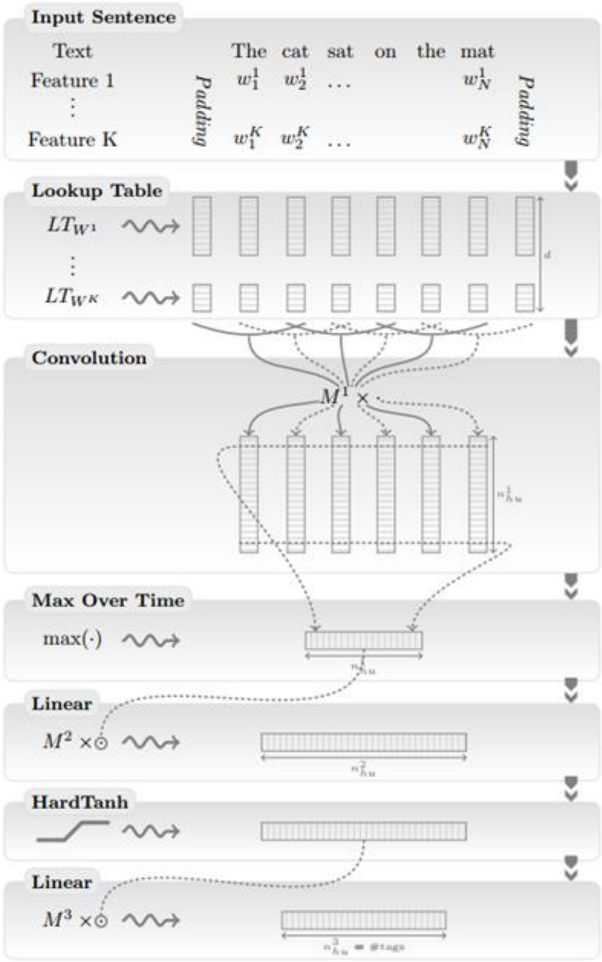
Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on Machine learning （ICML'08）. ACM, 2008: 160-167.

MODELS

- The First NER System based on the **Deep Neural Network** and **Feature Engineering**
- The NER Systems for the Combination of **Word embedding** and **Neural Network**
- The NER Systems for the Combinations **of Character Embedding** and **Neural Network**
- The NER systems for the Combination of **Character Embedding**, **Word Embedding** and **Neural Network**
- The NER Systems for the combinations of **Character Embedding**, **Word Embedding**, **affix** model and **Neural Network**
- The NER systems based on the **Deep Active Learning**
- The NER Systems based on the **Adversarial Learning**

The NER Systems for the Combination of Word embedding and Neural Network

- The first layer extracts features for each word.
- The second layer extracts features from a window of words or from the whole sentence, treating it as a sequence with local and global structure (i.e., it is not treated like a bag of words).
- The following layers are standard Neural Network layers.



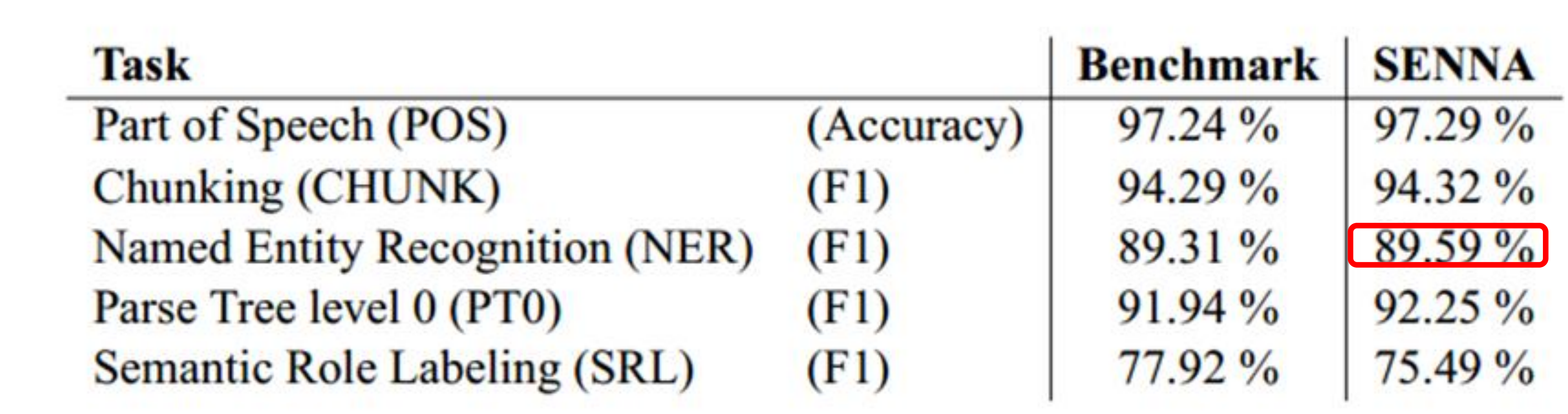
Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research （JMLR'11）, 2011, 12(Aug): 2493-2537.

The NER Systems for the Combination of Word embedding and Neural Network

- Our **first English corpus** is the entire English Wikipedia. We have removed all paragraphs containing non-roman characters and all MediaWiki markups. The resulting text was tokenized using the Penn Treebank tokenizer script.14 The resulting data set contains about 631 million words. As in our previous experiments, we use a dictionary containing the 100,000 most common words in WSJ, with the same processing of capitals and numbers. Again, words outside the dictionary were replaced by the special "RARE" word.
- Our **second English corpus** is composed by adding an extra 221 million words extracted from the Reuters RCV1 (Lewis et al., 2004) data set.15 We also extended the dictionary to 130,000 words by adding the 30,000 most common words in Reuters. This is useful in order to determine whether improvements can be achieved by further increasing the unlabeled data set size.

Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research （JMLR'11）, 2011, 12(Aug): 2493-2537.

The NER Systems for the Combination of Word embedding and Neural Network



Task		Benchmark	SENNA
Part of Speech (POS)	(Accuracy)	97.24 %	97.29 %
Chunking (CHUNK)	(F1)	94.29 %	94.32 %
Named Entity Recognition (NER)	(F1)	89.31 %	89.59 %
Parse Tree level 0 (PT0)	(F1)	91.94 %	92.25 %
Semantic Role Labeling (SRL)	(F1)	77.92 %	75.49 %

Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research （JMLR'11）, 2011, 12(Aug): 2493-2537.

The NER Systems for the Combination of Word embedding and Neural Network

- The authors achieved **89.59% F1 score** on English CoNLL 2003 dataset by including gazetteers and SENNA embeddings.
- The paper rely on large **unlabeled data sets** and let the training algorithm discover internal representations that prove useful for all the tasks of interest.

Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research （JMLR'11）, 2011, 12(Aug): 2493-2537.

The NER Systems for the Combination of Word embedding and Neural Network

Advantages

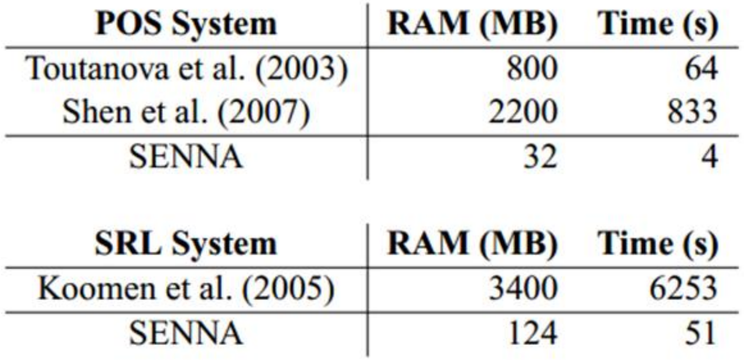
- The model make full use of huge data set and unlabeled data, which is almost from scratch.
- The model is the better for generalization than others.
- Using the word embeddings vector instead of hand-constructed feature vectors, which is represented by n-dimension vector space.

Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research （JMLR'11）, 2011, 12(Aug): 2493-2537.

The NER Systems for the Combination of Word embedding and Neural Network

Advantages

- The RAM and computational speed are most fast between systems based on the word embeddings, which is minimal computational requirements.



POS System	RAM (MB)	Time (s)
Toutanova et al. (2003)	800	64
Shen et al. (2007)	2200	833
SENNA	32	4

SRL System	RAM (MB)	Time (s)
Koomen et al. (2005)	3400	6253
SENNA	124	51

Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research （JMLR'11）, 2011, 12(Aug): 2493-2537.

The NER Systems for the Combination of Word embedding and Neural Network

Advantages

- avoid task-specific engineering and disregarding a lot of prior knowledge

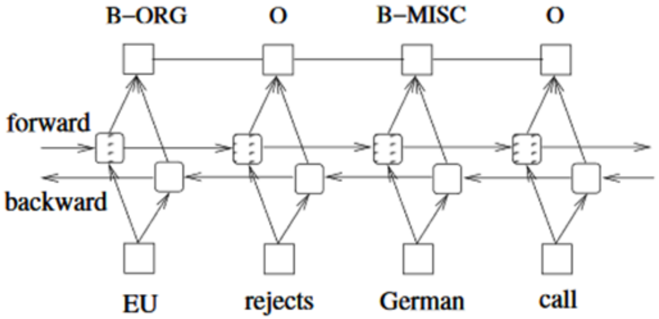
Disadvantages

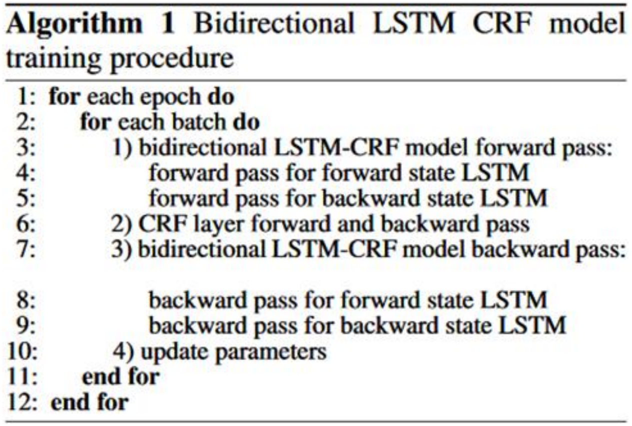
- It is necessary to add others languages new corpora for F1 score.
- The model is dependent to the word embedding.
- It is necessary to **resort the word embedding**.

Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research （JMLR'11）, 2011, 12(Aug): 2493-2537.

The NER Systems for the Combination of Word embedding and Neural Network

- In each epoch, we divide the whole training data to batches and process one batch at a time. Each batch contains a list of sentences which is determined by the parameter of batch size. For each batch, we first run bidirectional LSTM-CRF model forward pass which includes the forward pass for both forward state and backward state of LSTM. As a result, we get the output score for all tags at all positions.





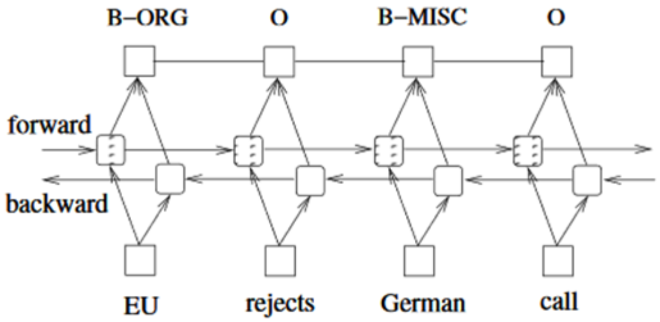
Algorithm 1 Bidirectional LSTM CRF model training procedure

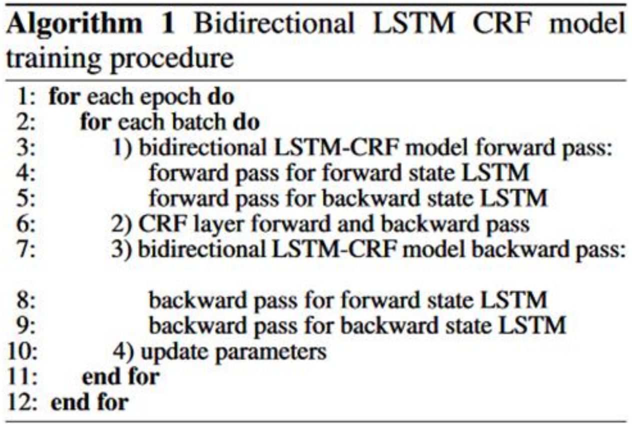
```
for each epoch do
    for each batch do
        1) bidirectional LSTM-CRF model forward pass:
            forward pass for forward state LSTM
            forward pass for backward state LSTM
        2) CRF layer forward and backward pass
        3) bidirectional LSTM-CRF model backward pass:

            backward pass for forward state LSTM
            backward pass for backward state LSTM
        4) update parameters
    end for
end for
```

Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

The NER Systems for the Combination of Word embedding and Neural Network

- Then, the model runs CRF layer forward and backward pass to compute gradients for network output and state transition edges. After that, we can back propagate the errors from the output to the input, which includes the backward pass for both forward and backward states of LSTM. 3） Finally we update the network parameters which include the state transition matrix, and the original bidirectional LSTM parameters.




Algorithm 1 Bidirectional LSTM CRF model training procedure

```
for each epoch do
    for each batch do
        1) bidirectional LSTM-CRF model forward pass:
            forward pass for forward state LSTM
            forward pass for backward state LSTM
        2) CRF layer forward and backward pass
        3) bidirectional LSTM-CRF model backward pass:
            backward pass for forward state LSTM
            backward pass for backward state LSTM
        4) update parameters
    end for
end for
```

Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

The NER Systems for the Combination of Word embedding and Neural Network

- **Datasets and Experimental Results** 84.26% F1 score on English CoNLL 2003 dataset
- **Conclusions** The model **is the first work** of applying a **BI-LSTM-CRF** model to NLP benchmark sequence tagging data. The model is **robust** and it has **less dependence on word embedding** as compared to the observation in (Collobert et.al., 2011). It can achieve accurate tagging accuracy **without resorting to word embedding**

Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

The NER Systems for the Combination of Word embedding and Neural Network

Advantages

- The experiments show that BI-LSTM-CRF model is robust
- It has less dependence on word embedding as compared to previous observations (Collobert et al., 2011). It can produce accurate tagging performance without resorting to word embedding.
- The first work of applying a BI-LSTM-CRF model to NLP benchmark sequence tagging data
- Without resorting to word embedding

Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

The NER Systems for the Combination of Word embedding and Neural Network

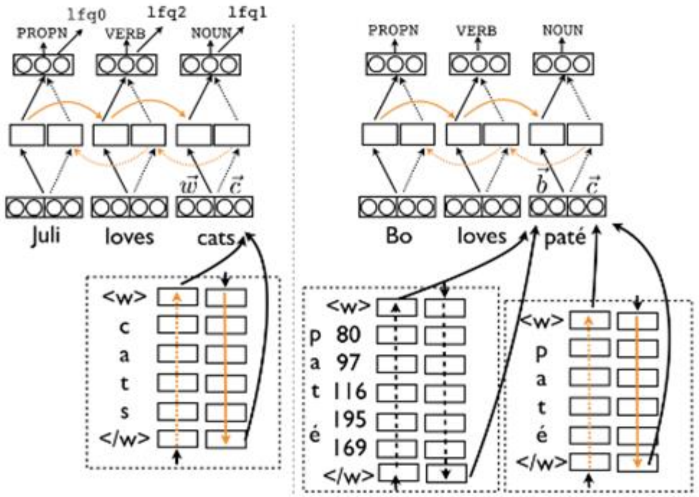
Disadvantages

- The F1 score is less than previous models, which is necessary improved.
- It is not enough for experiments which use English corpus only. The experimental results should add the new corpora.

Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

The NER Systems for the Combination of Word embedding and Neural Network

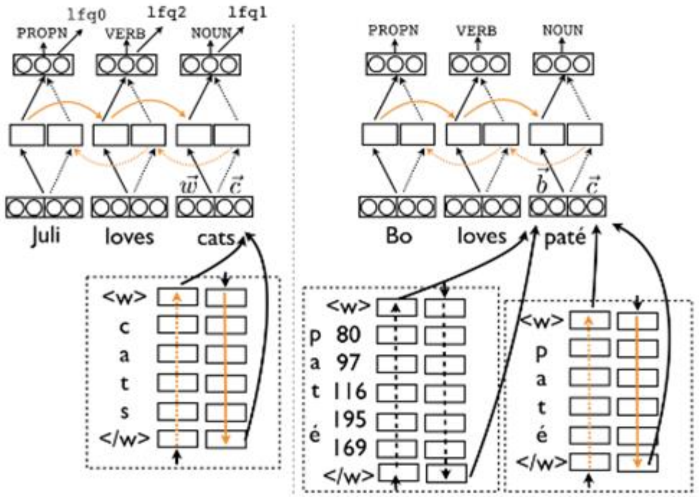
- evaluate the effectiveness of different representations in bi-LSTMs
- compare these models across a large set of languages and under varying conditions(data size, label noise)
- propose a novel bi-LSTM model with auxiliary loss， whicl combines the POS tagging loss function with an auxiliary loss function that accounts for rare words.



Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[J]. arXiv preprint arXiv:1604.05529, 2016. 引用：127

The NER Systems for the Combination of Word embedding and Neural Network

- evaluate the effectiveness of different representations in bi-LSTMs
- compare these models across a large set of languages and under varying conditions(data size, label noise)
- propose a novel bi-LSTM model with auxiliary loss， which combines the POS tagging loss function with an auxiliary loss function that accounts for rare words.



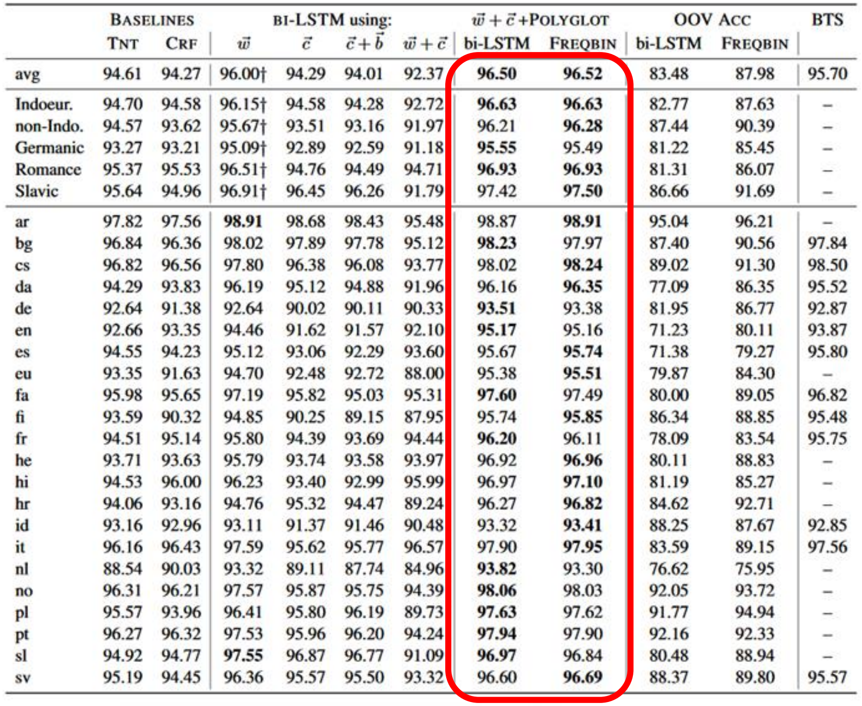
Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[J]. arXiv preprint arXiv:1604.05529, 2016. 引用：127

The NER Systems for the Combination of Word embedding and Neural Network

- **Datasets** For the multilingual experiments, we use the data from the Universal Dependencies project v1.2 (Nivre et al., 2015) (17 POS) with the canonical data splits. We consider all languages that have at least 60k tokens and are distributed with word forms, resulting in 22 languages. We also report accuracies on WSJ (45 POS) using the standard splits (Collins, 2002; Manning, 2011).

Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[J]. arXiv preprint arXiv:1604.05529, 2016. 引用：127

The NER Systems for the Combination of Word embedding and Neural Network



	BASELINES		BI-LSTM using:				$\vec{w}+\vec{c}$+POLYGLOT		OOV ACC		BTS
	TNT	CRF	$\vec{w}$	$\vec{c}$	$\vec{c}+\vec{b}$	$\vec{w}+\vec{c}$	bi-LSTM	FREQBIN	bi-LSTM	FREQBIN	
avg	94.61	94.27	96.00†	94.29	94.01	92.37	**96.50**	**96.52**	83.48	87.98	95.70
Indoeur.	94.70	94.58	96.15†	94.58	94.28	92.72	**96.63**	**96.63**	82.77	87.63	–
non-Indo.	94.57	93.62	95.67†	93.51	93.16	91.97	96.21	**96.28**	87.44	90.39	–
Germanic	93.27	93.21	95.09†	92.89	92.59	91.18	**95.55**	95.49	81.22	85.45	–
Romance	95.37	95.53	96.51†	94.76	94.49	94.71	**96.93**	**96.93**	81.31	86.07	–
Slavic	95.64	94.96	96.91†	96.45	96.26	91.79	97.42	**97.50**	86.66	91.69	–
ar	97.82	97.56	**98.91**	98.68	98.43	95.48	98.87	**98.91**	95.04	96.21	–
bg	96.84	96.36	98.02	97.89	97.78	95.12	**98.23**	97.97	87.40	90.56	97.84
cs	96.82	96.56	97.80	96.38	96.08	93.77	98.02	**98.24**	89.02	91.30	98.50
da	94.29	93.83	96.19	95.12	94.88	91.96	96.16	**96.35**	77.09	86.35	95.52
de	92.64	91.38	92.64	90.02	90.11	90.33	**93.51**	93.38	81.95	86.77	92.87
en	92.66	93.35	94.46	91.62	91.57	92.10	**95.17**	95.16	71.23	80.11	93.87
es	94.55	94.23	95.12	93.06	92.29	93.60	95.67	**95.74**	71.38	79.27	95.80
eu	93.35	91.63	94.70	92.48	92.72	88.00	95.38	**95.51**	79.87	84.30	–
fa	95.98	95.65	97.19	95.82	95.03	95.31	**97.60**	97.49	80.00	89.05	96.82
fi	93.59	90.32	94.85	90.25	89.15	87.95	95.74	**95.85**	86.34	88.85	95.48
fr	94.51	95.14	95.80	94.39	93.69	94.44	**96.20**	96.11	78.09	83.54	95.75
he	93.71	93.63	95.79	93.74	93.58	93.97	96.92	**96.96**	80.11	88.83	–
hi	94.53	96.00	96.23	93.40	92.99	95.99	96.97	**97.10**	81.19	85.27	–
hr	94.06	93.16	94.76	95.32	94.47	89.24	96.27	**96.82**	84.62	92.71	–
id	93.16	92.96	93.11	91.37	91.46	90.48	93.32	**93.41**	88.25	87.67	92.85
it	96.16	96.43	97.59	95.62	95.77	96.57	97.90	**97.95**	83.59	89.15	97.56
nl	88.54	90.03	93.32	89.11	87.74	84.96	**93.82**	93.30	76.62	75.95	–
no	96.31	96.21	97.57	95.87	95.75	94.39	**98.06**	98.03	92.05	93.72	–
pl	95.57	93.96	96.41	95.80	96.19	89.73	**97.63**	97.62	91.77	94.94	–
pt	96.27	96.32	97.53	95.96	96.20	94.24	**97.94**	97.90	92.16	92.33	–
sl	94.92	94.77	**97.55**	96.87	96.77	91.09	**96.97**	96.84	80.48	88.94	–
sv	95.19	94.45	96.36	95.57	95.50	93.32	96.60	**96.69**	88.37	89.80	95.57

Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[J]. arXiv preprint arXiv:1604.05529, 2016. 引用：127

The NER Systems for the Combination of Word embedding and Neural Network

Conclusions

- The paper evaluated token and sub-token-level representations for neural network-based part-of-speech tagging across **22 languages** and **proposed a novel multi-tasks bi-LSTM with auxiliary loss**.
- The auxiliary loss is effective at improving the accuracy of rare words.

Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[J]. arXiv preprint arXiv:1604.05529, 2016. 引用：127

The NER Systems for the Combination of Word embedding and Neural Network

Advantages

- It is not sensitive to data set size and label noise
- Across 22 languages, and the F1 score is better than previous systems.
- It works especially well for morphologically complex languages

Disadvantages

- The model is increasingly complex , but the change of F1 score is not much obvious.

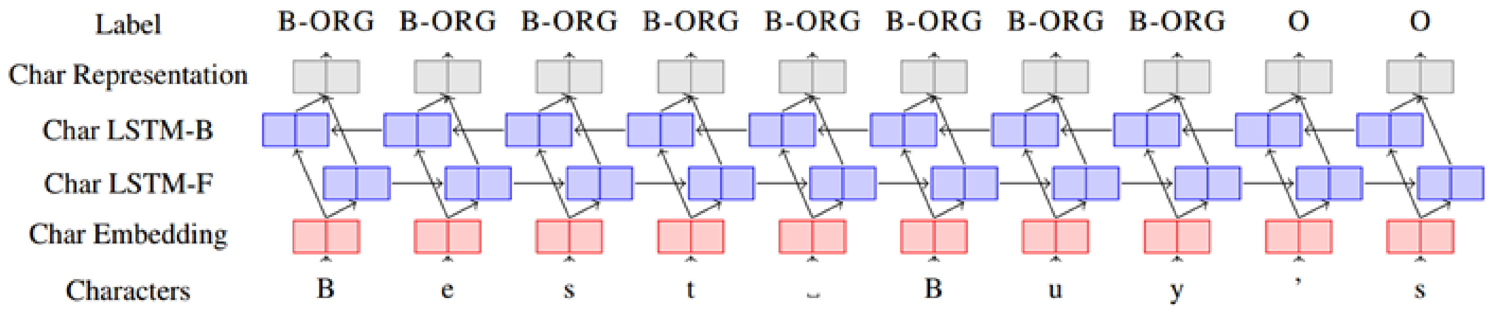
Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[J]. arXiv preprint arXiv:1604.05529, 2016. 引用：127

MODELS

- The First NER System based on the **Deep Neural Network** and **Feature Engineering**
- The NER Systems for the Combination of **Word Embedding** and **Neural Network**
- The NER Systems for the Combinations **of Character Embedding** and **Neural Network**
- The NER systems for the Combination of **Character Embedding**, **Word Embedding** and **Neural Network**
- The NER Systems for the combinations of **Character Embedding**, **Word Embedding**, **affix** model and **Neural Network**
- The NER systems based on the **Deep Active Learning**
- The NER Systems based on the **Adversarial Learning**

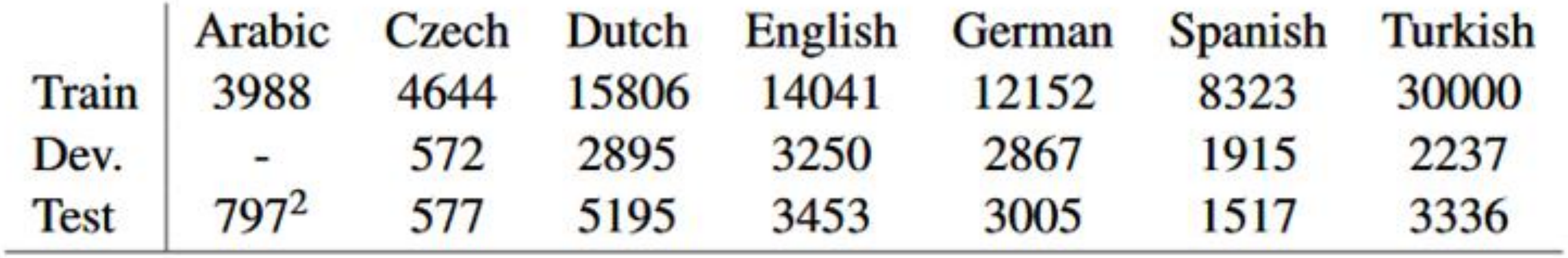
The NER Systems for the Combinations of Character Embedding and Neural Network

- Sentence is represented as a sequence of **characters**.

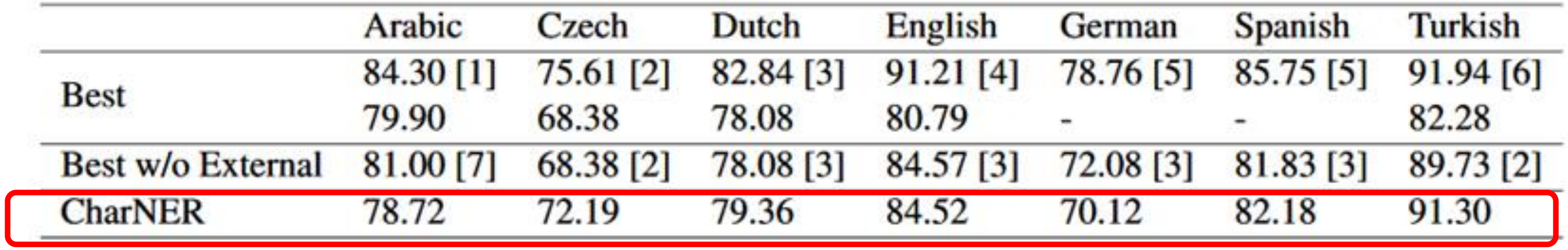


Kuru O, Can O A, Yuret D. Charner: Character-level named entity recognition[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics （COLING'16） : Technical Papers. 2016: 911-921.

The NER Systems for the Combinations of Character Embedding and Neural Network



	Arabic	Czech	Dutch	English	German	Spanish	Turkish
Train	3988	4644	15806	14041	12152	8323	30000
Dev.	-	572	2895	3250	2867	1915	2237
Test	797[2]	577	5195	3453	3005	1517	3336



	Arabic	Czech	Dutch	English	German	Spanish	Turkish
Best	84.30 [1]	75.61 [2]	82.84 [3]	91.21 [4]	78.76 [5]	85.75 [5]	91.94 [6]
	79.90	68.38	78.08	80.79	-	-	82.28
Best w/o External	81.00 [7]	68.38 [2]	78.08 [3]	84.57 [3]	72.08 [3]	81.83 [3]	89.73 [2]
CharNER	78.72	72.19	79.36	84.52	70.12	82.18	91.30

Kuru O, Can O A, Yuret D. Charner: Character-level named entity recognition[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics （COLING'16）: Technical Papers. 2016: 911-921.

The NER Systems for the Combination of Word embedding and Neural Network

Advantages

- character-level model. Taking characters as the primary representation is superior to considering words as the basic input unit.
- The main contribution is to show that the same deep character level model is able to achieve good performance on multiple languages without hand engineered features or language specific external resources.

Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[J]. arXiv preprint arXiv:1604.05529, 2016. 引用：127

The NER Systems for the Combination of Word embedding and Neural Network

Disadvantages

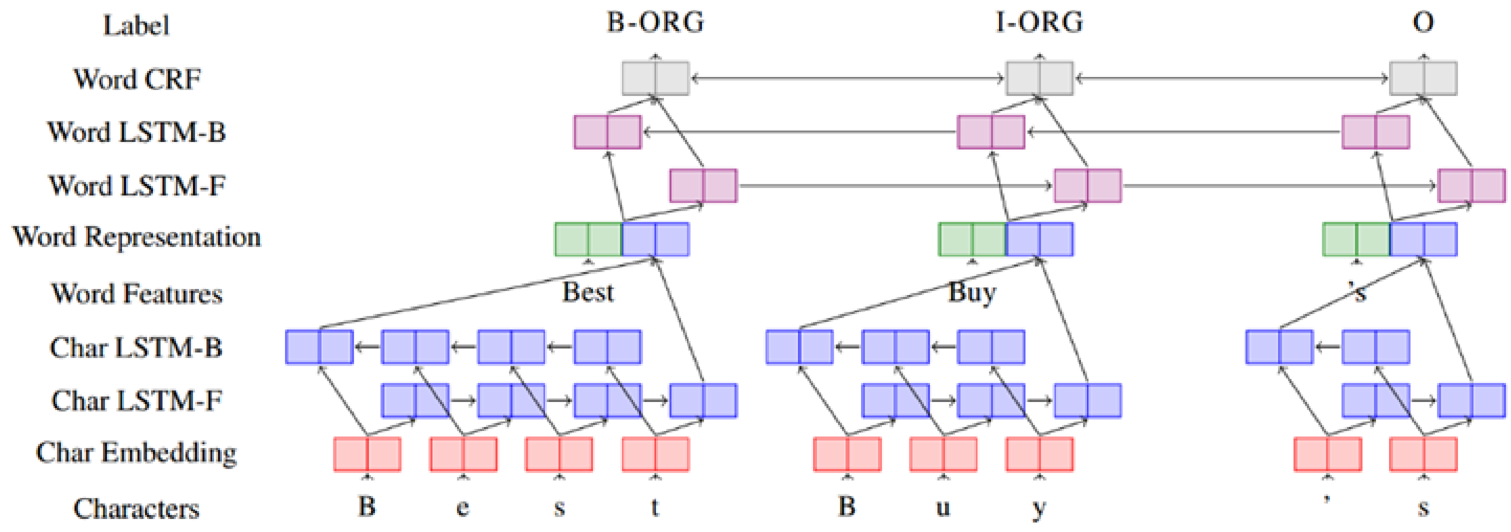
- There is nothing specific to NER in the model. It should be evaluate on other tasks such as part-of-speech tagging and shallow parsing. (Multi-tasks)
- The robust experiment is lack.

Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[J]. arXiv preprint arXiv:1604.05529, 2016. 引用：127

MODELS

- The First NER System based on the **Deep Neural Network** and **Feature Engineering**
- The NER Systems for the Combination of **Word Embedding** and **Neural Network**
- The NER Systems for the Combinations **of Character Embedding** and **Neural Network**
- The NER systems for the Combination of **Character Embedding**, **Word Embedding** and **Neural Network**
- The NER Systems for the combinations of **Character Embedding**, **Word Embedding**, **affix** model and **Neural Network**
- The NER systems based on the **Deep Active Learning**
- The NER Systems based on the **Adversarial Learning**

The NER systems for the Combination of Character Embedding, Word Embedding and Neural Network



Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv preprint arXiv:1603.01354, 2016. 引用：435

The NER Systems for the Combination of Word embedding and Neural Network

- **Dataset and Experimental Results** The F1 score of the model achieves 91.21% using **CoNLL2003**
- **Conclusions and Advantages** It is a truly end-to-end model relying on no task-specific resources, feature engineering or data pre-processing**.**

Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[J]. arXiv preprint arXiv:1604.05529, 2016. 引用：127

The NER Systems for the Combination of Word embedding and Neural Network

Disadvantages

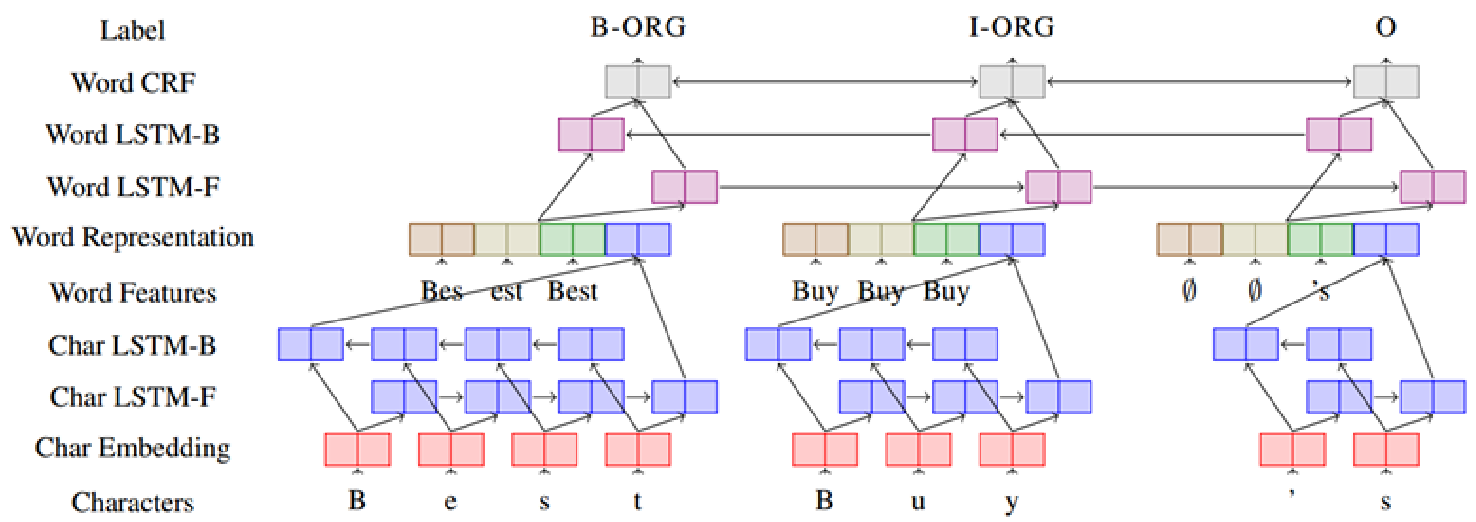
- The model can be further improved by exploring multi-tasks learning approaches to combine more useful and correlated information. In a word, it is not support for multi-tasks learning.
- The model may be further explored in the different application areas such as bioinformatics, medical.
- It is necessary to add the new corpora for supporting the F1 score.

Plank B, Søgaard A, Goldberg Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss[J]. arXiv preprint arXiv:1604.05529, 2016. 引用：127

MODELS

- The First NER System based on the **Deep Neural Network** and **Feature Engineering**
- The NER Systems for the Combination of **Word Embedding** and **Neural Network**
- The NER Systems for the Combinations **of Character Embedding** and **Neural Network**
- The NER systems for the Combination of **Character Embedding**, **Word Embedding** and **Neural Network**
- The NER Systems for the combinations of **Character Embedding**, **Word Embedding**, **affix** model and **Neural Network**
- The NER systems based on the **Deep Active Learning**
- The NER Systems based on the **Adversarial Learning**

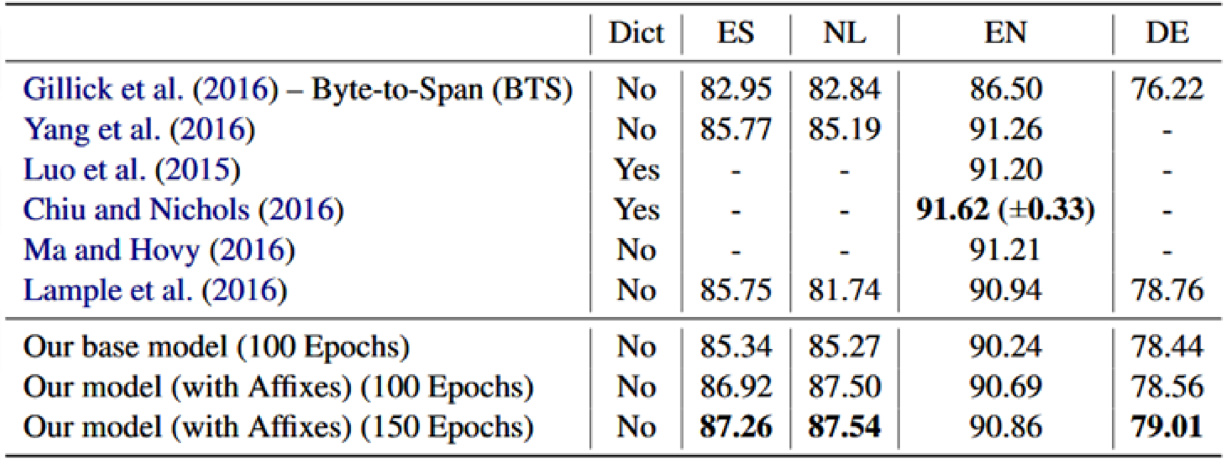
The NER Systems for the combinations of Character Embedding, Word Embedding, affix model and Neural Network



Yadav V, Sharp R, Bethard S. Deep Affix Features Improve Neural Named Entity Recognizers[C]//Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. 2018: 167-172

The NER Systems for the combinations of Character Embedding, Word Embedding, affix model and Neural Network

- **Datasets and Experimental Results**



	Dict	ES	NL	EN	DE
Gillick et al. (2016) – Byte-to-Span (BTS)	No	82.95	82.84	86.50	76.22
Yang et al. (2016)	No	85.77	85.19	91.26	-
Luo et al. (2015)	Yes	-	-	91.20	-
Chiu and Nichols (2016)	Yes	-	-	**91.62 (±0.33)**	-
Ma and Hovy (2016)	No	-	-	91.21	-
Lample et al. (2016)	No	85.75	81.74	90.94	78.76
Our base model (100 Epochs)	No	85.34	85.27	90.24	78.44
Our model (with Affixes) (100 Epochs)	No	86.92	87.50	90.69	78.56
Our model (with Affixes) (150 Epochs)	No	**87.26**	**87.54**	90.86	**79.01**

Yadav V, Sharp R, Bethard S. Deep Affix Features Improve Neural Named Entity Recognizers[C]//Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. 2018: 167-172

The NER Systems for the combinations of Character Embedding, Word Embedding, affix model and Neural Network

- **Conclusions** Straight-forward and language-independent approach shows performance gains compared to other neural systems for NER, achieving a new state of the art on Spanish, Dutch, and German NER as well as the MedLine portion of DrugNER

Yadav V, Sharp R, Bethard S. Deep Affix Features Improve Neural Named Entity Recognizers[C]//Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. 2018: 167-172

The NER Systems for the combinations of Character Embedding, Word Embedding, affix model and Neural Network

Advantages

- The model is tested in multi-languages, which shows up the ability of generalization.

Disadvantages

- The model adds the most successful features from feature-engineering approaches： affixes, which is necessary to the labeled datasets. The pro-process is different to us.

Yadav V, Sharp R, Bethard S. Deep Affix Features Improve Neural Named Entity Recognizers[C]//Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics. 2018: 167-172

MODELS

- The First NER System based on the **Deep Neural Network** and **Feature Engineering**
- The NER Systems for the Combination of **Word Embedding** and **Neural Network**
- The NER Systems for the Combinations **of Character Embedding** and **Neural Network**
- The NER systems for the Combination of **Character Embedding**, **Word Embedding** and **Neural Network**
- The NER Systems for the combinations of **Character Embedding**, **Word Embedding**, **affix** model and **Neural Network**
- The NER systems based on the **Deep Active Learning**
- The NER Systems based on the **Adversarial Learning**

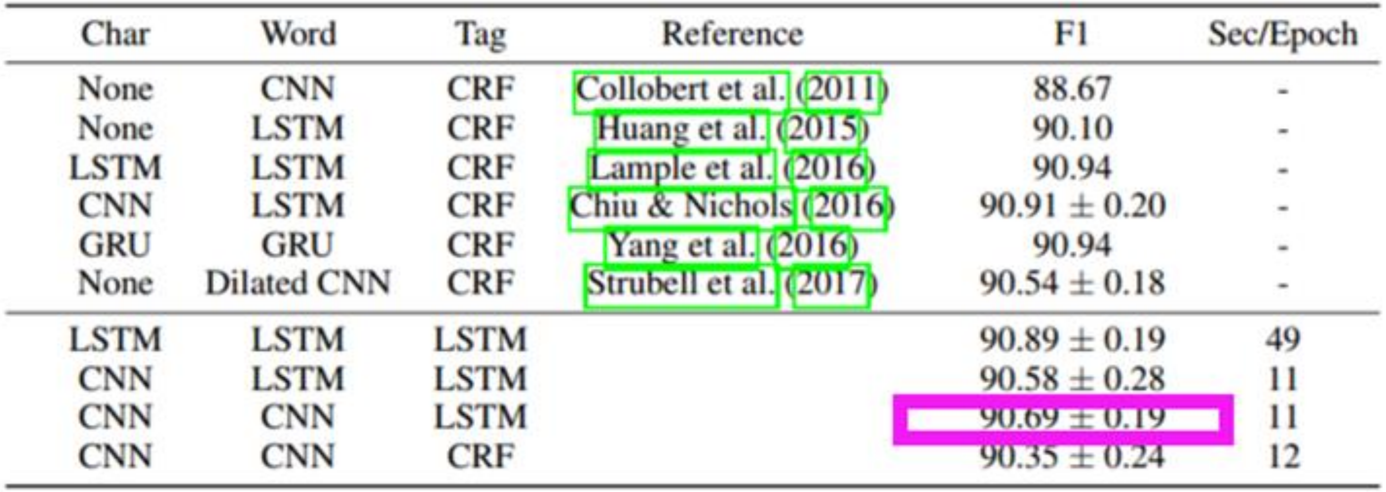
The NER systems based on the Deep Active Learning

- In this work, the combination of deep learning and active learning drastically reduced the number of labeled data. The CNN-CNN-LSTM model consisting of convolutional character and word encoders and a long short term memory (LSTM) tag decoder. The model achieves nearly state-of-the-art performance on standard datasets for the task while being computationally much more efficient than best performing models
- 主动学习：通过“选择策略”主动从未标注的样本集中挑选部分（1个或N个）样本让相关领域的专家进行标注；然后将标注过的样本增加到训练数据集给“学习模块”进行训练；当“学习模块”满足终止条件时即可结束程序，否则不断重复上述步骤获得更多的标注样本进行训练。

Shen Y, Yun H, Lipton Z C, et al. Deep Active Learning for Named Entity Recognition[J]. arXiv preprint arXiv:1707.05928, 2017. 引用：21

The NER systems based on the Deep Active Learning

- **Dataset** On the CoNLL-2003 English dataset



Char	Word	Tag	Reference	F1	Sec/Epoch
None	CNN	CRF	Collobert et al. (2011)	88.67	-
None	LSTM	CRF	Huang et al. (2015)	90.10	-
LSTM	LSTM	CRF	Lample et al. (2016)	90.94	-
CNN	LSTM	CRF	Chiu & Nichols (2016)	90.91 ± 0.20	-
GRU	GRU	CRF	Yang et al. (2016)	90.94	-
None	Dilated CNN	CRF	Strubell et al. (2017)	90.54 ± 0.18	-
LSTM	LSTM	LSTM		90.89 ± 0.19	49
CNN	LSTM	LSTM		90.58 ± 0.28	11
CNN	CNN	LSTM		90.69 ± 0.19	11
CNN	CNN	CRF		90.35 ± 0.24	12

Shen Y, Yun H, Lipton Z C, et al. Deep Active Learning for Named Entity Recognition[J]. arXiv preprint arXiv:1707.05928, 2017. 引用：21

The NER systems based on the Deep Active Learning

Advantages

- The model achieves **the incremental training** with each batch of new labels: mix newly annotated samples with the older ones, and update the neural network weights for a small number of epochs, before querying for labels in a new round. This modification **drastically reduces the computational requirements** of active methods and makes it practical to deploy them.
- The model contains convolutional character-level encoder, convolutional word-level encoder, and long short term memory (LSTM) tag decoder, which **trains much faster than other deep models**.

Shen Y, Yun H, Lipton Z C, et al. Deep Active Learning for Named Entity Recognition[J]. arXiv preprint arXiv:1707.05928, 2017. 引用：21

The NER systems based on the Deep Active Learning

Disadvantages

- The model is necessary to the labeled data sets is used in training steps, but it is necessary to sample a little of data only.
- It is necessary to add the new corpora for supporting the F1 score.

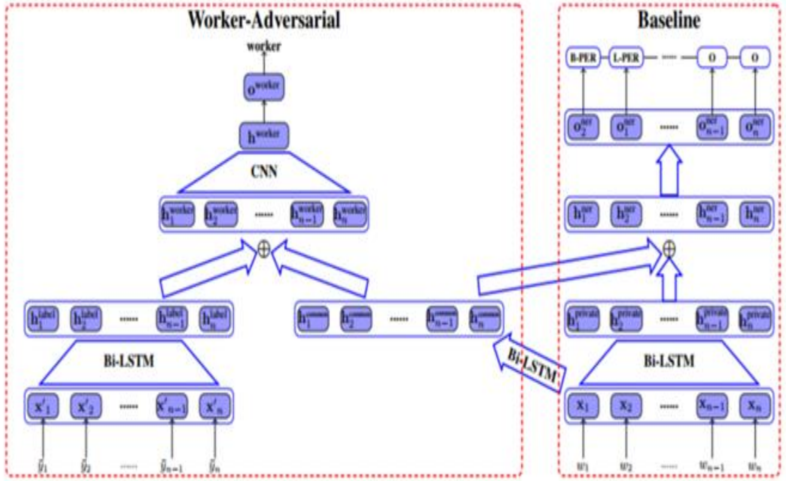
Shen Y, Yun H, Lipton Z C, et al. Deep Active Learning for Named Entity Recognition[J]. arXiv preprint arXiv:1707.05928, 2017. 引用：21

MODELS

- The First NER System based on the **Deep Neural Network** and **Feature Engineering**
- The NER Systems for the Combination of **Word Embedding** and **Neural Network**
- The NER Systems for the Combinations **of Character Embedding** and **Neural Network**
- The NER systems for the Combination of **Character Embedding**, **Word Embedding** and **Neural Network**
- The NER Systems for the combinations of **Character Embedding**, **Word Embedding**, **affix** model and **Neural Network**
- The NER systems based on the **Deep Active Learning**
- The NER Systems based on the **Adversarial Learning**

The NER Systems based on the Adversarial Learning

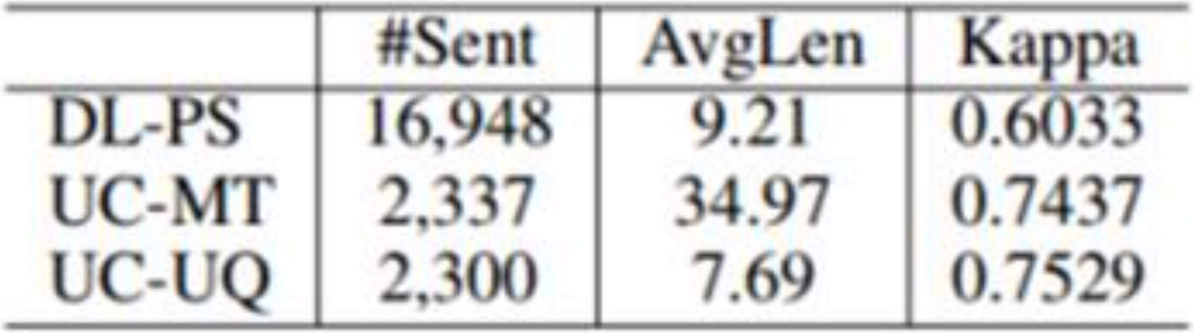
- To quickly obtain new labeled data, we can choose **crowd-sourcing** as an alternative way at lower cost in a short time. But as an exchange, **crowd annotations from non-experts may be of lower quality than those from experts**. To make full use of noisy sequence labels, the following model is proposed. Inspired by adversarial learning, the model uses a common Bi-LSTM and a private Bi-LSTM for representing annotator-generic and -specific information.


Worker-Adversarial
worker
o^{worker}
h^{worker}
CNN
h_1^{worker} h_2^{worker} ...... h_{n-1}^{worker} h_n^{worker}
h_1^{label} h_2^{label} ...... h_{n-1}^{label} h_n^{label}
Bi-LSTM
x'_1 x'_2 ...... x'_{n-1} x'_n
$\bar{y}_1$ $\bar{y}_2$ ...... $\bar{y}_{n-1}$ $\bar{y}_n$
h_1^{common} h_2^{common} ...... h_{n-1}^{common} h_n^{common}
Bi-LSTM
Baseline
B-PER L-PER ...... O O
o_2^{ner} o_1^{ner} ...... o_{n-1}^{ner} o_n^{ner}
h_1^{ner} h_2^{ner} ...... h_{n-1}^{ner} h_n^{ner}
$h_1^{private}$ $h_2^{private}$ ...... $h_{n-1}^{private}$ $h_n^{private}$
Bi-LSTM
x_1 x_2 ...... x_{n-1} x_n
w_1 w_2 ...... w_{n-1} w_n

Yang Y S, Zhang M, Chen W, et al. Adversarial Learning for Chinese NER from Crowd Annotations[C]// Proceedings of the AAAI Conference on Artificial Intelligence (AAAI’18), 2018.

The NER Systems based on the Adversarial Learning

- **Datasets** The datasets include labeled datasets and unlabeled datasets.

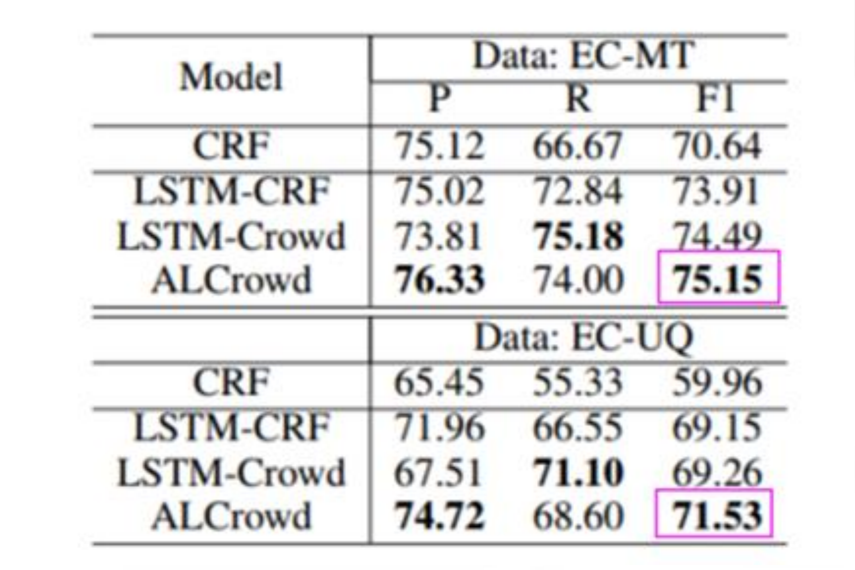


	#Sent	AvgLen	Kappa
DL-PS	16,948	9.21	0.6033
UC-MT	2,337	34.97	0.7437
UC-UQ	2,300	7.69	0.7529

Yang Y S, Zhang M, Chen W, et al. Adversarial Learning for Chinese NER from Crowd Annotations[C]// Proceedings of the AAAI Conference on Artificial Intelligence (AAAI’18), 2018.

The NER Systems based on the Adversarial Learning

- **Experimental Results**



Model	Data: EC-MT		
	P	R	F1
CRF	75.12	66.67	70.64
LSTM-CRF	75.02	72.84	73.91
LSTM-Crowd	73.81	**75.18**	74.49
ALCrowd	**76.33**	74.00	**75.15**
	Data: EC-UQ		
CRF	65.45	55.33	59.96
LSTM-CRF	71.96	66.55	69.15
LSTM-Crowd	67.51	**71.10**	69.26
ALCrowd	**74.72**	68.60	**71.53**

Yang Y S, Zhang M, Chen W, et al. Adversarial Learning for Chinese NER from Crowd Annotations[C]// Proceedings of the AAAI Conference on Artificial Intelligence (AAAI’18), 2018.

The NER Systems based on the Adversarial Learning

- **Conclusions** The experimental results show that the proposed approach outperforms strong baseline systems.
- **Advantages** The adversarial learning makes full use of **noisy sequence labels**
- **Disadvantages** The model cannot handle entities separated by other entities or non-entity words.

Yang Y S, Zhang M, Chen W, et al. Adversarial Learning for Chinese NER from Crowd Annotations[C]// Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'18), 2018.

OUTLINE

- Introduction
- Methodology
- Models
- Conclusions
- Opinions

CONCOLUTIONS

- Neural network models generally outperform feature-engineered models.
- The combination of character and word hybrid neural networks generally outperform other representational choices.
- The current methods make full use of unlabeled, large number of data sets.

CONCOLUTIONS

- Another interesting direction is to apply models to data from other domains such as social media (Twitter and Weibo).
- Can other supervised learning methods replace affixes of feature engineering methods in the Section 6.5?
- There is still interesting progress to be made by incorporating key features of past feature-engineered models into modern Neural Network architectures.

CONCOLUTIONS

- The combination of LSTM and CRF is still the focused research topic, especially in the different ways of using small corpus and training sets.
- The adversarial learning makes full use of noisy sequence labels.
- It should be explored for transfer learning, active learning and joint learning in NER systems.

OUTLINE

- **Introduction**
- **Methodology**
- **Models**
- **Conclusions**
- **Opinions**

OPINIONS

- With the evolution of the computer sciences technologies, such as knowledge engineer, machine learning, deep neural network, active learning and adversarial learning, the progress of the NER systems.
- It is trends to combine NER systems and focused methods.
- It is weak to the NER systems for the specific domains, Which should be explored for us.

OPINIONS

- It may be future work for the NER systems for the combinations of **character embedding**, **word embedding**, and **neural network** adds the **unsupervised learning**, **supervised learning** or **semi-supervised learning algorithms**.