Boundary Smoothing for Named Entity Recognition

The 60th Annual Meeting of the ACL

Enwei Zhu Jinpeng Li

HwaMei Hospital, UCAS
Ningbo Institute of Life and Health Industry, UCAS

May 2022, Dublin, Ireland

Outline

Named Entity Recognition: Task Definition

Input text:

- The White House is in Washington D.C.

Named entities in text:

- The $\boxed{\text{White House}}_{\texttt{ORG}}$ is in $\boxed{\text{Washington D.C.}}_{\texttt{LOC}}$

Named entities as a set of tuples:

- { (ORG, 1, 2), (LOC, 5, 6) }

Named Entity Recognition: Task Definition

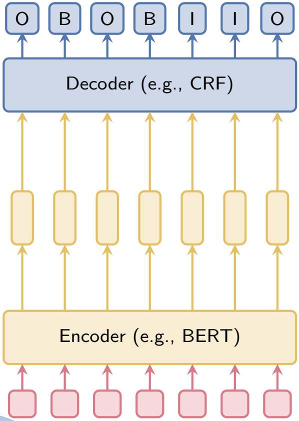
Input: A piece of raw text

- A T-length sequence of tokens: $x_1, x_2, \ldots, x_T$

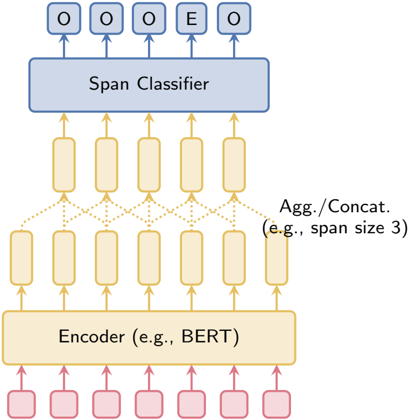
Output: A set of entities

- $\{(type_i, start_i, end_i) \mid type_i \in S, 0 \leq start_i \leq end_i < T\}$, where S is the set of entity types
- Each entity is specified by its type and boundaries (start and end positions)

Named Entity Recognition: Existing Methods



Sequence Tagging



Span Classification

Boundary Annotation Is Ambiguous

Take CoNLL 2003 Annotation Guidelines as the example:

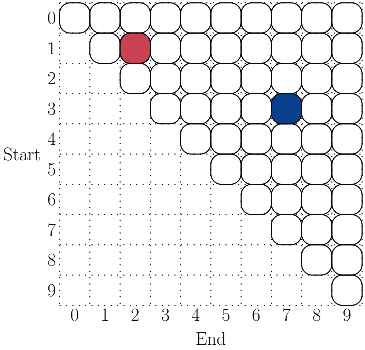
- Entity types are clear and easily distinguishable
  - PER, LOC, ORG, MISC
- Entity boundaries may be ambiguous for "boundary words"

Text	Boundary words
[The [White House]ORG]ORG	Article
[The [Godfather]PER]PER	Article
[[Clinton]PER government]ORG	Modifier
[Mr. [Harry Schearer]PER]PER	Person title
[[John Doe]PER, Jr.]PER	Name appositive

Sharpness in Classification Targets

The fitting targets of neural span-based NER models

- The annotated spans are assigned with full probability
- All other spans are assigned with zero probability



Over-Confidence

Over-confidence: The confidence of a predicted entity is much higher than its correctness probability

A manifestation: Disconnect between dev. loss and F_1 score

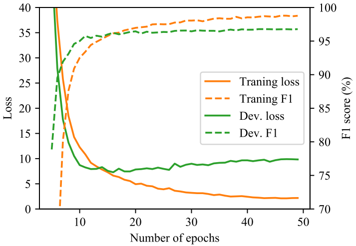


Figure: Cross entropy loss

Over-Confidence

Over-confidence: The confidence of a predicted entity is much higher than its correctness probability

A manifestation: Disconnect between dev. loss and F_1 score

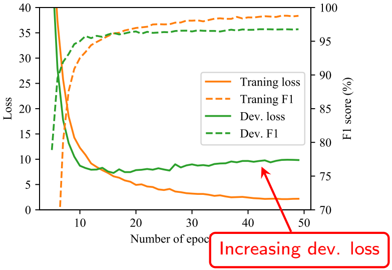


Figure: Cross entropy loss

Outline

Boundary Smoothing: Motivation

Our observations

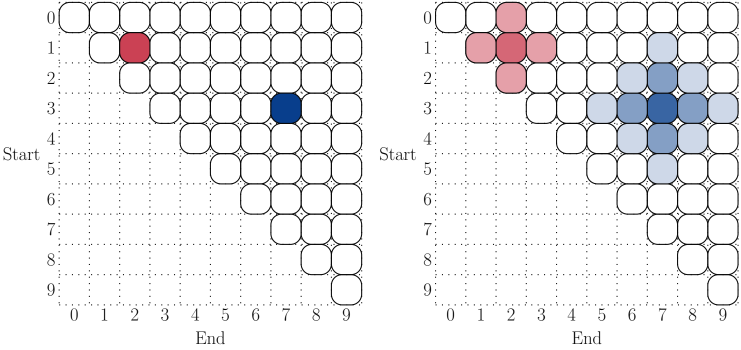
- Entity boundary annotation may be ambiguous
- Sharpness exists in the targets of span-based NER models (which should not, given the ambiguity of boundaries)
- Span-based NER models encounter over-confidence issue

Our Solution: **Boundary Smoothing**

- Re-allocate entity probabilities from annotated spans to the surrounding ones
- A regularization technique
- Inspired by label smoothing [Szegedy et al., 2016]

Boundary Smoothing

Re-allocate entity probabilities from annotated spans to the surrounding ones



(a) Hard boundary (b) Smoothed boundary

Outline

Experimental Settings

Backbone

- RoBERTa-base (English); BERT-base-wwm (Chinese)
- BiLSTM

Decoder

- Biaffine [Yu et al., 2020]
- Fuse representations of start and end tokens
- Simple but powerful

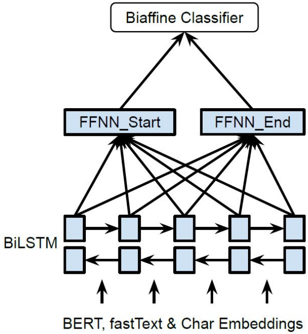


Figure: Biaffine Decoder.
Source: [Yu et al., 2020].

Experimental Settings

Eight datasets

- English corpora with flat entities
  - CoNLL 2003; OntoNotes 5
- English corpora with nested entities
  - ACE 2004; ACE 2005
- Chinese corpora with flat entities
  - OntoNotes 4; MSRA; Resume NER; Weibo NER

Evaluation

- Exact match of entity type and boundaries
- Precision, recall, F_1 score

Outline

Main Results: English Datasets

Compared with baseline: +0.2% to +0.6% F_1 score

CoNLL 2003			
Model	Prec.	Rec.	F1
Lample et al. (2016)	–	–	90.94
Chiu and Nichols (2016)†	91.39	91.85	91.62
Peters et al. (2018)	–	–	92.22
Akbik et al. (2018)†	–	–	93.07
Devlin et al. (2019)	–	–	92.8
Straková et al. (2019)†	–	–	93.38
Wang et al. (2019)†	–	–	93.43
Li et al. (2020b)	92.33	94.61	93.04
Yu et al. (2020)†	93.7	93.3	93.5
Baseline	92.93	94.03	93.48
Baseline + BS	93.61	93.68	**93.65**

+0.17%

OntoNotes 5			
Model	Prec.	Rec.	F1
Chiu and Nichols (2016)	86.04	86.53	86.28
Li et al. (2020b)	92.98	89.95	91.11
Yu et al. (2020)	91.1	91.5	91.3
Baseline	90.31	92.13	91.21
Baseline + BS	91.75	91.74	**91.74**

+0.53%

ACE 2004			
Model	Prec.	Rec.	F1
Katiyar and Cardie (2018)	73.6	71.8	72.7
Straková et al. (2019)†	–	–	84.40
Li et al. (2020b)	85.05	86.32	85.98
Yu et al. (2020)	87.3	86.0	86.7
Shen et al. (2021)	87.44	87.38	87.41
Baseline	86.67	88.42	87.54
Baseline + BS	88.43	87.53	**87.98**

+0.44%

ACE 2005			
Model	Prec.	Rec.	F1
Katiyar and Cardie (2018)	70.6	70.4	70.5
Straková et al. (2019)†	–	–	84.33
Li et al. (2020b)	87.16	86.59	86.88
Yu et al. (2020)	85.2	85.6	85.4
Shen et al. (2021)	86.09	87.27	86.67
Baseline	84.29	88.97	86.56
Baseline + BS	86.25	88.07	**87.15**

+0.59%

Main Results: English Datasets

Compared with previous SOTA: +0.2% to +0.6% F_1 score

CoNLL 2003			
Model	Prec.	Rec.	F1
Lample et al. (2016)	–	–	90.94
Chiu and Nichols (2016)†	91.39	91.85	91.62
Peters et al. (2018)	–	–	92.22
Akbik et al. (2018)†	–	–	93.07
Devlin et al. (2019)	–	–	92.8
Straková et al. (2019)†	–	–	93.38
Wang et al. (2019)†	–	–	93.43
Li et al. (2020b)	92.33	94.61	93.04
Yu et al. (2020)†	93.7	93.3	93.5
Baseline	92.93	94.03	93.48
Baseline + BS	93.61	93.68	**93.65**

+0.15%

OntoNotes 5			
Model	Prec.	Rec.	F1
Chiu and Nichols (2016)	86.04	86.53	86.28
Li et al. (2020b)	92.98	89.95	91.11
Yu et al. (2020)	91.1	91.5	91.3
Baseline	90.31	92.13	91.21
Baseline + BS	91.75	91.74	**91.74**

+0.44%

ACE 2004			
Model	Prec.	Rec.	F1
Katiyar and Cardie (2018)	73.6	71.8	72.7
Straková et al. (2019)†	–	–	84.40
Li et al. (2020b)	85.05	86.32	85.98
Yu et al. (2020)	87.3	86.0	86.7
Shen et al. (2021)	87.44	87.38	87.41
Baseline	86.67	88.42	87.54
Baseline + BS	88.43	87.53	**87.98**

+0.57%

ACE 2005			
Model	Prec.	Rec.	F1
Katiyar and Cardie (2018)	70.6	70.4	70.5
Straková et al. (2019)†	–	–	84.33
Li et al. (2020b)	87.16	86.59	86.88
Yu et al. (2020)	85.2	85.6	85.4
Shen et al. (2021)	86.09	87.27	86.67
Baseline	84.29	88.97	86.56
Baseline + BS	86.25	88.07	**87.15**

+0.27%

Main Results: Chinese Datasets

Compared with baseline: +0.3% to +1.2% F_1 score

OntoNotes 4			
Model	Prec.	Rec.	F1
Zhang and Yang (2018)	76.35	71.56	73.88
Ma et al. (2020)	83.41	82.21	82.81
Li et al. (2020a)	–	–	81.82
Li et al. (2020b)	82.98	81.25	82.11
Chen and Kong (2021)	79.25	80.66	79.95
Wu et al. (2021)	–	–	82.57
Baseline	82.79	81.27	82.03
Baseline + BS	81.65	84.03	**82.83**

+0.80%

MSRA			
Model	Prec.	Rec.	F1
Zhang and Yang (2018)	93.57	92.79	93.18
Ma et al. (2020)	95.75	95.10	95.42
Li et al. (2020a)	–	–	96.09
Li et al. (2020b)	96.18	95.12	95.75
Wu et al. (2021)	–	–	96.24
Baseline	95.82	95.78	95.80
Baseline + BS	96.37	96.15	**96.26**

+0.46%

Weibo NER			
Model	Prec.	Rec.	F1
Zhang and Yang (2018)	–	–	58.79
Ma et al. (2020)	–	–	70.50
Li et al. (2020a)	–	–	68.55
Shen et al. (2021)	70.11	68.12	69.16
Chen and Kong (2021)	–	–	70.14
Wu et al. (2021)	–	–	70.43
Baseline	68.65	74.40	71.41
Baseline + BS	70.16	75.36	**72.66**

+1.25%

Resume NER			
Model	Prec.	Rec.	F1
Zhang and Yang (2018)	94.81	94.11	94.46
Ma et al. (2020)	96.08	96.13	96.11
Li et al. (2020a)	–	–	95.86
Wu et al. (2021)	–	–	95.98
Baseline	95.81	96.87	96.34
Baseline + BS	96.63	96.69	**96.66**

+0.32%

Main Results: Chinese Datasets

Compared with previous SOTA: +0.02% to +2.1% F_1 score

OntoNotes 4				
Model	Prec.	Rec.	F1	
Zhang and Yang (2018)	76.35	71.56	73.88	
Ma et al. (2020)	83.41	82.21	82.81	
Li et al. (2020a)	–	–	81.82	
Li et al. (2020b)	82.98	81.25	82.11	
Chen and Kong (2021)	79.25	80.66	79.95	
Wu et al. (2021)	–	–	82.57	
Baseline	82.79	81.27	82.03	
Baseline + BS	81.65	84.03	**82.83**	+0.02%

MSRA				
Model	Prec.	Rec.	F1	
Zhang and Yang (2018)	93.57	92.79	93.18	
Ma et al. (2020)	95.75	95.10	95.42	
Li et al. (2020a)	–	–	96.09	
Li et al. (2020b)	96.18	95.12	95.75	
Wu et al. (2021)	–	–	96.24	
Baseline	95.82	95.78	95.80	
Baseline + BS	96.37	96.15	**96.26**	+0.02%

Weibo NER				
Model	Prec.	Rec.	F1	
Zhang and Yang (2018)	–	–	58.79	
Ma et al. (2020)	–	–	70.50	
Li et al. (2020a)	–	–	68.55	
Shen et al. (2021)	70.11	68.12	69.16	
Chen and Kong (2021)	–	–	70.14	
Wu et al. (2021)	–	–	70.43	
Baseline	68.65	74.40	71.41	
Baseline + BS	70.16	75.36	**72.66**	+2.16%

Resume NER				
Model	Prec.	Rec.	F1	
Zhang and Yang (2018)	94.81	94.11	94.46	
Ma et al. (2020)	96.08	96.13	96.11	
Li et al. (2020a)	–	–	95.86	
Wu et al. (2021)	–	–	95.98	
Baseline	95.81	96.87	96.34	
Baseline + BS	96.63	96.69	**96.66**	+0.55%

Outline

Ablation Study: Smoothing Parameters

The effect of boundary smoothing remains robust
Standard label smoothing does not have such effect

	CoNLL 2003	ACE 2005	Resume NER
Baseline	93.48	86.56	96.34
BS ($\epsilon = 0.1$, $D = 1$)	93.50	86.65	96.63
BS ($\epsilon = 0.2$, $D = 1$)	93.56	86.96	**96.66**
BS ($\epsilon = 0.3$, $D = 1$)	**93.65**	86.81	96.50
BS ($\epsilon = 0.1$, $D = 2$)	93.45	**87.15**	96.33
BS ($\epsilon = 0.2$, $D = 2$)	93.39	86.99	96.62
BS ($\epsilon = 0.3$, $D = 2$)	93.57	86.71	96.28
LS ($\alpha = 0.1$)	93.43	86.31	96.31
LS ($\alpha = 0.2$)	93.37	86.17	96.38
LS ($\alpha = 0.3$)	93.26	85.65	96.26

Ablation Study: Backbone

Boundary smoothing works regardless of PLM and BiLSTM
RoBERTa outperforms original BERT on English NER

- RoBERTa is trained on much more data
- RoBERTa focuses on MLM by removing NSP

	CoNLL 2003	ACE 2005	Resume NER
Baseline	93.48	86.56	96.34
+ BS	**93.65**	**87.15**	**96.66**
Baseline w/ BERT-base	91.84	84.51	
+ BS	**92.05**	**84.95**	
Baseline w/ BERT-large	92.92	85.83	
+ BS	**93.08**	**86.33**	
Baseline w/ RoBERTa-large	93.66	87.82	
+ BS	**93.77**	**88.02**	
Baseline w/ MacBERT-base			96.41
+ BS			**96.75**
Baseline w/ MacBERT-large			96.46
+ BS			**96.75**
Baseline w/o BiLSTM	93.13	86.22	96.24
+ BS	**93.30**	**86.58**	**96.56**

Outline

Outline

Model Calibration

Calibration: Do prediction confidences well reflect accuracy?

	Confidence	Correctness Probability
Well calibrated	80% (−)	80%
Over confident	90% (↑)	80%
Under confident	60% (↓)	80%

How to measure model calibration?

- Expected calibration error (ECE) [Guo et al., 2017]
  - Group samples into confidence bins $I_m = \left(\frac{m-1}{M}, \frac{m}{M}\right]$
  - Lower ECE suggests better calibration

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

Model Calibration

How to measure model calibration?

- Reliability diagram [Guo et al., 2017]
  - Plot accuracy against confidence
  - Being closer to the diagonal suggests better calibration

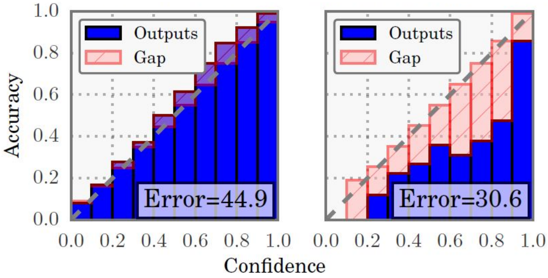
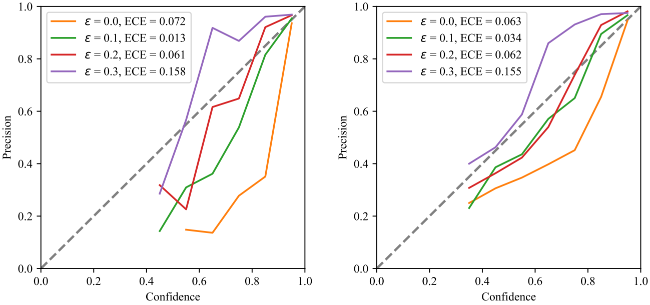


Figure: Reliability diagrams. Source: [Guo et al., 2017].

Calibration of Entities

Entity calibration: How well do the entity confidence reflect its probability to be a true entity?

- Models encounter over-confidence without BS
- BS improves calibration; $\epsilon = 0.1$ achieves the lowest ECE

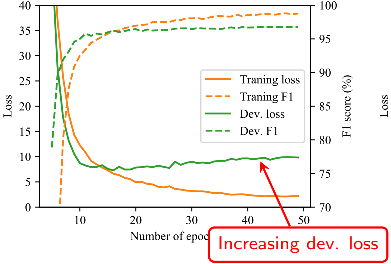


(a) CoNLL 2003 (b) OntoNotes 5

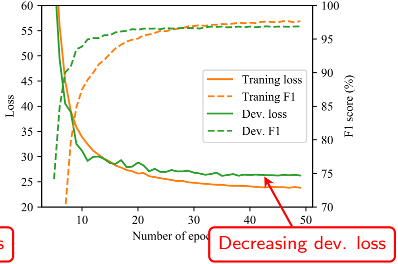
Over-Confidence

Over-confidence: The confidence of a predicted entity is much higher than its correctness probability

Boundary smoothing alleviates the disconnect between dev. loss and F_1 score



(a) Cross entropy loss



(b) Boundary smoothing loss

Outline

Why Does It Improve Performance?

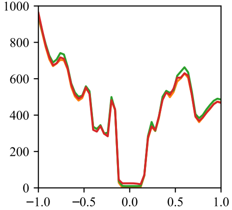
How does boundary smoothing improve the performance?

- Recall: Sharpness exists in the targets of span-based NER models
- Neural networks may prefer continuous solutions [Hornik et al., 1989]
- It is intuitively hard to optimize the neural models given the sharpness
- Boundary smoothing mitigates the sharpness

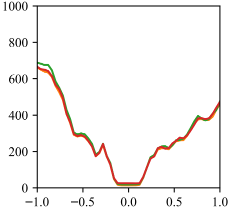
Loss Landscape Visualization

- Flatter minima and less chaotic loss landscapes result in better generalization and trainability [Hochreiter and Schmidhuber, 1997, Li et al., 2018]
- Many techniques improve loss landscapes, e.g., residual connection, small batch size [Li et al., 2018]

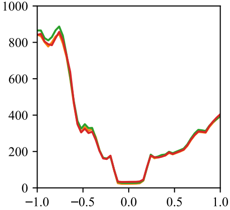
Loss Landscape Visualization



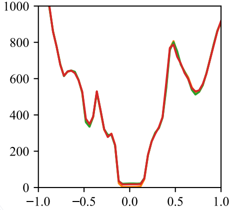
(a) CoNLL 2003



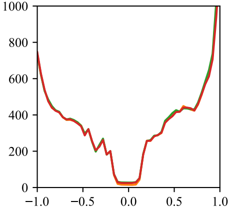
(b) CoNLL 2003 ($\epsilon = 0.1$)



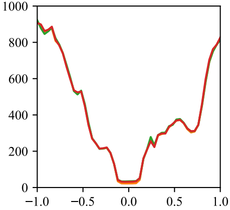
(c) CoNLL 2003 ($\epsilon = 0.2$)



(d) OntoNotes 5



(e) OntoNotes 5 ($\epsilon = 0.1$)



(f) OntoNotes 5 ($\epsilon = 0.2$)

Conclusions

We propose boundary smoothing for span-based neural NER models

Boundary smoothing re-assigns entity probabilities from annotated spans to the surrounding ones

With the help of boundary smoothing, our model:

- Achieves SOTA performance on eight well-known NER benchmarks
- Presents better entity calibration, less over-confidence
- Arrives at flatter neural minima and more smoothed loss landscapes

References I

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017).
On calibration of modern neural networks.
In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Hochreiter, S. and Schmidhuber, J. (1997).
Flat minima.
Neural Computation, 9(1):1–42.

Hornik, K., Stinchcombe, M., and White, H. (1989).
Multilayer feedforward networks are universal approximators.
Neural networks, 2(5):359–366.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018).
Visualizing the loss landscape of neural nets.
In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 6391–6401.

References II

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016).
Rethinking the inception architecture for computer vision.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Yu, J., Bohnet, B., and Poesio, M. (2020).
Named entity recognition as dependency parsing.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476, Online. Association for Computational Linguistics.