Boundary Smoothing for Named Entity Recognition

Enwei Zhu[1,2], Jinpeng Li[1,2]

[1]HwaMei Hospital, University of Chinese Academy of Sciences

[2]Ningbo Institute of Life and Health Industry, University of Chinese Academy of Sciences

Background

Named entity recognition (NER) is a fundamental NLP task. As a common setting, an entity is regarded as correctly recognized only if its type and two boundaries exactly match the ground truth. Recently, span-based models have gained much popularity in NER studies, and achieved SOTA results [1, 2].

In NER engineering, we observe:

1. The annotation of boundaries is more *ambiguous* than entity types. In particular, "boundary words" (e.g., articles or modifiers) may cause the ambiguity in the boundaries of an entity mention.
2. The fitting targets of span-based neural NER models have noticeable *sharpness* between adjacent spans. The positive spans (i.e., annotated entities) are assigned with full probability to be an entity, whereas all other spans are assigned with zero probability.
3. Span-based neural NER models easily encounter the *over-confidence* issue, i.e., the confidence of a predicted entity is much higher than its correctness probability.

Method: Boundary Smoothing

Span-based NER models enumerate all candidate spans and classify them into entity types. For any span (i, j) in a T-length sequence, a span-based model typically yields predicted probabilities $\hat{y}_{ij}$, and computes the cross entropy loss against the ground truth y_{ij}:

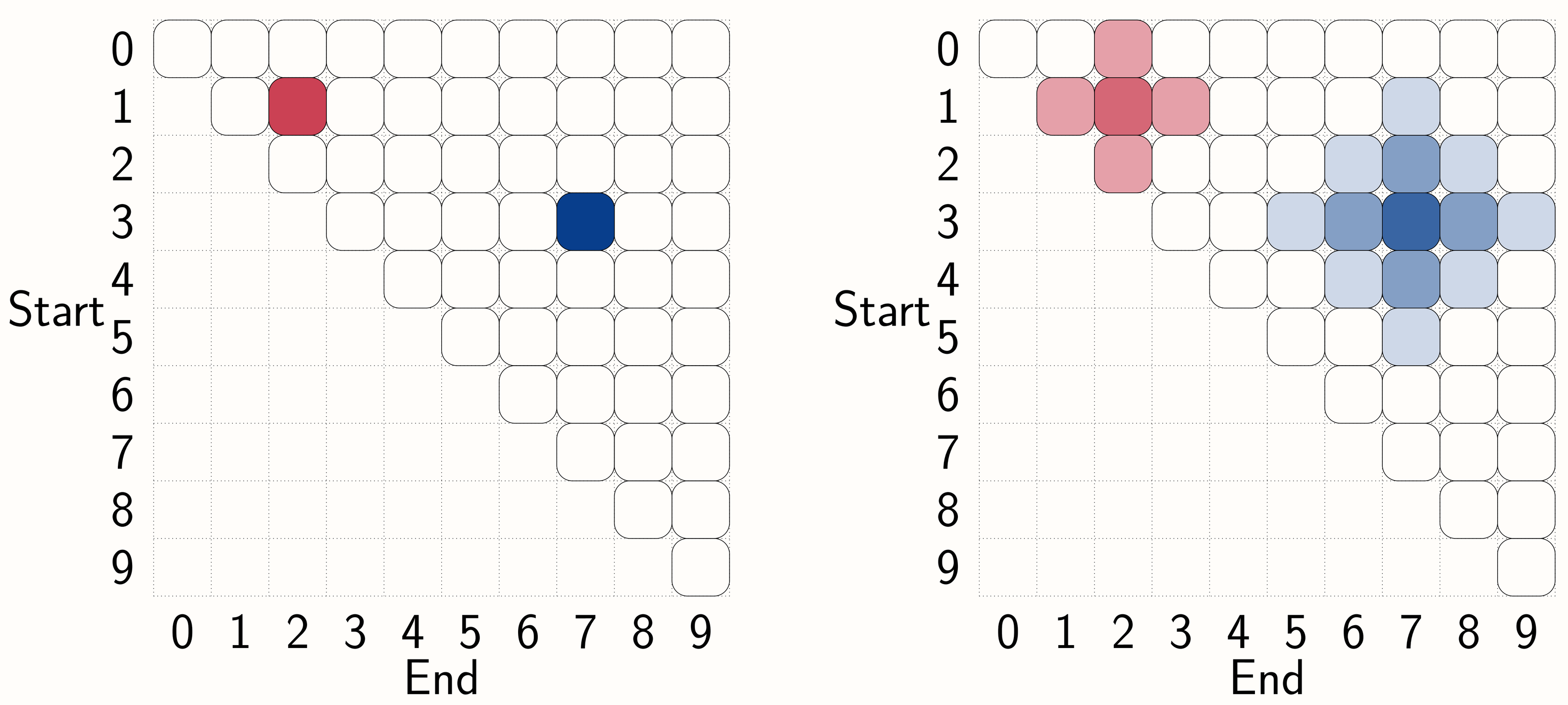
$$\mathcal{L}_{\mathrm{CE}} = -\sum_{0 \le i \le j < T} y_{ij}^{\mathsf{T}} \log(\hat{y}_{ij}).$$

Figure 1a visualizes the ground truth y_{ij} for an example sentence with two entities. The positive spans are assigned with full entity probability, whereas all other spans are assigned with zero probability. This creates sharpness between the fitting targets of adjacent spans.

Due to the ambiguity in boundary annotation, the spans surrounding an annotated one deserve a small entity probability. Inspired by label smoothing [3], we propose *boundary smoothing*, which re-allocates a portion of entity probabilities from the annotated spans to the surrounding ones. Figure 1b visualizes the boundary smoothing targets $\tilde{y}_{ij}$. Hence, the boundary smoothing loss is:

$$\mathcal{L}_{\mathrm{BS}} = -\sum_{0 \le i \le j < T} \tilde{y}_{ij}^{\mathsf{T}} \log(\hat{y}_{ij}).$$

Boundary smoothing explicitly reduces the sharpness, resulting in more continuous targets across spans. Boundary smoothing also helps alleviate over-confidence, because it prevents the model from concentrating all probability mass on the scarce positive samples.



(a) Hard boundary (b) Smoothed boundary

Figure 1: An example of hard and smoothed boundaries. The example sentence has ten tokens and two entities of spans (1, 2) and (3, 7), which are smoothed by sizes of 1 and 2 in the right subfigure, respectively.

Experimental Results

Our baseline includes a base-size pretrained language model (RoBERTa-base for English, BERT-base-wwm for Chinese) and the biaffine decoder [2]. Boundary smoothing consistently improves the performance, and achieves results better than or competitive with previous SOTA systems on eight well-known NER benchmarks.

	CoNLL 2003	OntoNotes 5	ACE 2004	ACE 2005
Previous SOTA	93.50	91.30	87.41	86.88
Baseline	93.48	91.21	87.54	86.56
Baseline + BS	**93.65**	**91.74**	**87.98**	**87.15**
	OntoNotes 4	MSRA	Weibo NER	Resume NER
Previous SOTA	82.81	96.24	70.50	96.11
Baseline	82.03	95.80	71.41	96.34
Baseline + BS	**82.83**	**96.26**	**72.66**	**96.66**

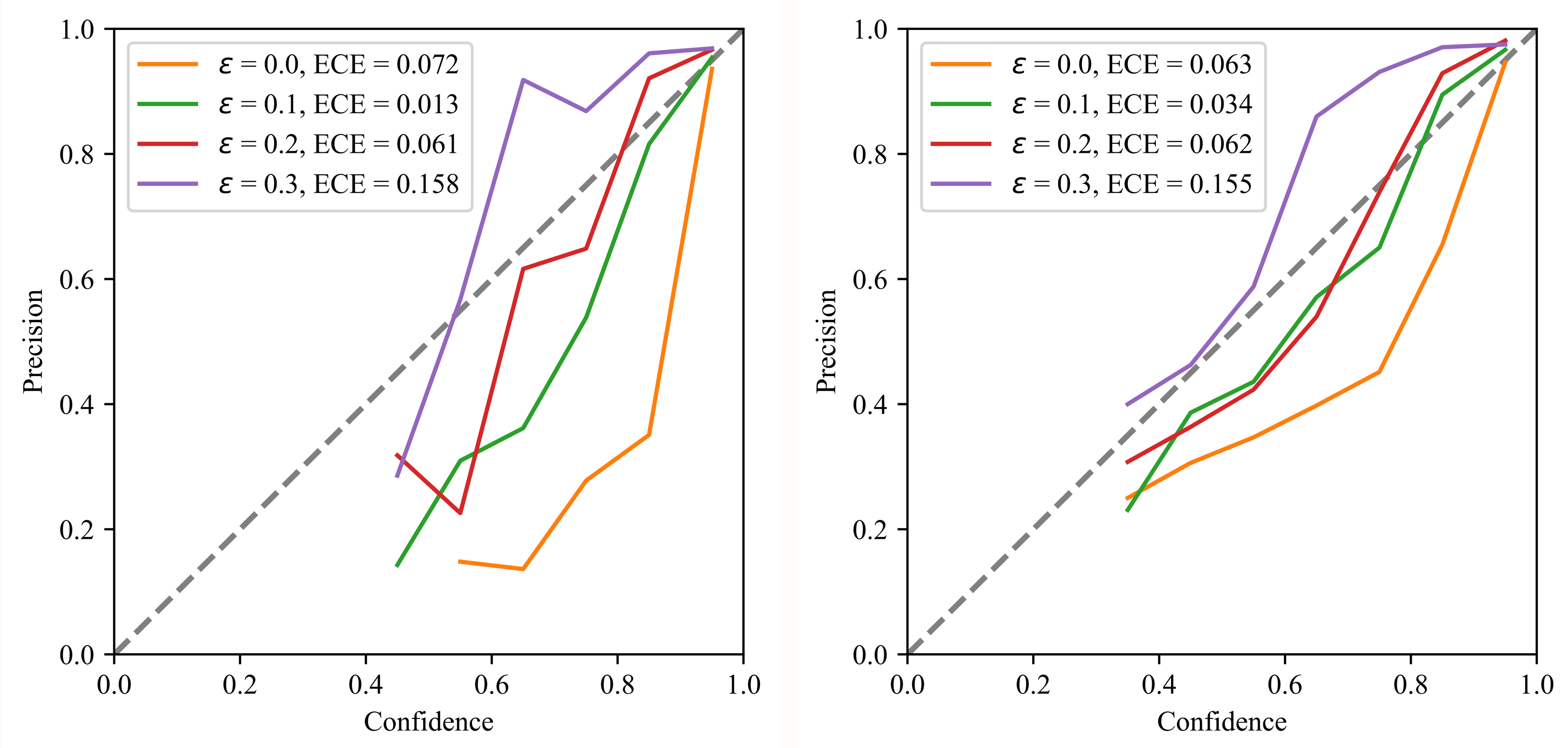
Table 1: Results on NER benchmarks. BS means boundary smoothing.

Further In-Depth Analysis

Over-Confidence and Entity Calibration

Modern neural networks are poorly calibrated; they produce confidences much higher than the corresponding correctness probabilities [4]. We follow their approach to evaluate *entity calibration*: How well can the entity confidence represent the precision rate?

Specifically, we group all the predicted entities by confidence into ten bins, and then calculate the precision rate for each bin. For well calibrated models, the precision rate should be close to the confidence level for each bin. As shown in Figure 2, the baseline models present large expected calibration errors (ECEs), with precision rates much lower than confidences, suggesting significant over-confidence. Boundary smoothing can effectively reduce ECEs and improve entity calibration.

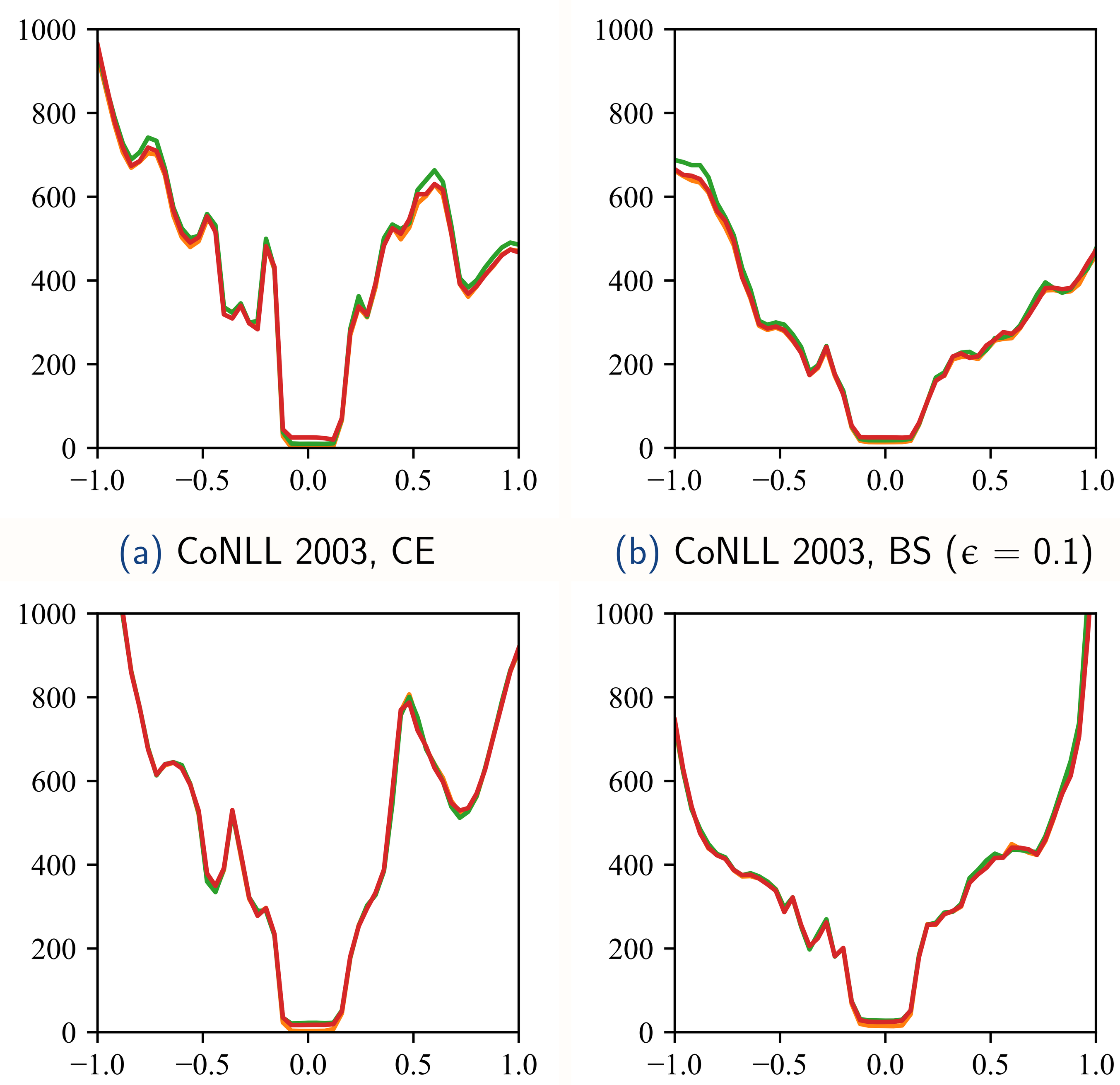


(a) CoNLL 2003 (b) OntoNotes 5

Figure 2: Reliability diagram of recognized entities on CoNLL 2003 and OntoNotes 5.

Loss Landscape Visualization

Figure 3 visualizes the loss landscapes for our models. Compared with the baseline, boundary smoothing brings less chaotic loss landscapes, and helps arrive at flatter minima. These properties are associated with better trainability and generalization [5], and thus explain the performance gain.



(a) CoNLL 2003, CE (b) CoNLL 2003, BS ($\epsilon = 0.1$)

(c) OntoNotes 5, CE (d) OntoNotes 5, BS ($\epsilon = 0.1$)

Figure 3: Visualization of loss landscapes on CoNLL 2003 and OntoNotes 5.

References

[1] M. G. Sohraby, M. Miway. 2018. Deep exhaustive model for nested named entity recognition. *EMNLP*.

[2] J. Yu, B. Bohnet, M. Poesio. 2020. Named entity recognition as dependency parsing. *ACL*.

[3] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna. 2016. Rethinking the inception architecture for computer vision. *CVPR*.

[4] C. Guo, G. Pleiss, Y. Sun, K. Q Weinberger. 2017. On calibration of modern neural networks. *ICML*.

[5] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein. 2018. Visualizing the loss landscape of neural nets. *NeurIPS*.

60th Annual Meeting of the Association for Computational Linguistics, 22nd - 27th May 2022, Dublin, Ireland