Debiasing Generative Named Entity Recognition by Calibrating Sequence Likelihood

Yu Xia[1], Yongwei Zhao[2], Wenhao Wu[1], Sujian Li[1]
[1]Key Laboratory of Computational Linguistics, Peking University, MOE, China
[2]SKL of Processors, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China
{yuxia, waynewu, lisujian}@pku.edu.cn, zhaoyongwei@ict.ac.cn

Abstract

Recognizing flat, overlapped and discontinuous entities uniformly has been paid increasing attention. Among these works, Seq2Seq formulation prevails for its flexibility and effectiveness. It arranges the output entities into a specific target sequence. However, it introduces bias by assigning all the probability mass to the observed sequence. To alleviate the bias, previous works either augment the data with possible sequences or resort to other formulations. In this paper, we stick to the Seq2Seq formulation and propose a reranking-based approach. It redistributes the likelihood among candidate sequences depending on their performance via a contrastive loss. Extensive experiments show that our simple yet effective method consistently boosts the baseline, and yields competitive or better results compared with the state-of-the-art methods on 8 widely-used datasets for Named Entity Recognition.

1 Introduction

Recently, recognizing flat, overlapped and discontinuous entities in a unified manner has been paid increasing attention. Among the existing works for unified Named Entity Recognition (NER), Seq2Seq formulation prevails for its flexibility and effectiveness in unified modeling (Yan et al., 2021; Lu et al., 2022; Ye et al., 2022). Typically, it arranges the output entities into a fixed order to form a target sequence, and trains the generative model by maximum likelihood estimation (MLE).

However, this estimation introduces bias by assuming a deterministic target distribution, where the model learns to assign all the probability mass to the observed target sequence. The biased estimation hurts the performance during decoding where predicted sequence likelihoods often do not accurately rank the performance of the generated sequences. To alleviate the bias, (Zhang et al., 2022) propose two data augmentation methods that sample possible sequences from the target space. Others resort to other formulations, *e.g.*, W^{2}NER (Li et al., 2022) reformulates NER as a word-word relation classification. In this study, we stick to the Seq2Seq formulation and explore how to mitigate the bias from another perspective orthogonal to (Zhang et al., 2022).

topK/B	CoNLL03	OntoNotes5.0	ACE04	ACE05
1/5	93.14	90.27	86.85	84.76
5/5	96.58	96.43	93.14	92.26
10/10	97.20	97.09	94.38	93.24
topK/B	**GENIA**	**CADEC**	**ShARe13**	**ShARe14**
1/5	78.93	70.53	79.69	80.35
5/5	89.66	81.17	89.36	90.68
10/10	91.64	83.01	91.11	91.87

Table 1: Oracle F1, *i.e.*, maximum F1 over topK candidates, on NER datasets based on BARTNER (Yan et al., 2021). topK/B denotes picking topK candidates out of candidates generated by beam search with beam size B.

Beam search decoding algorithms maintain B candidates in descending likelihoods and output the highest one. However, the rest candidates could contain predictions with better performance. We measure this phenomenon with oracle scores. As shown in Table 1, the beam candidates contain predictions with up to 8.1 points higher F1 over the outputted one, averaged on eight datasets. Doubling the beam size further increases the advantage to 9.38 points.

Recently, reranking-based methods proposed for the abstractive summarization task offer a potential technique (Liu and Liu, 2021; Ravaut et al., 2022). They train a discriminator on the candidates to predict a score for picking out the best candidate. For example, SimCLS (Liu and Liu, 2021) regards the cosine similarity between the input and candidate representations as the score. However, when applying reranking-based methods to our task, we find a challenge originating from the nature of information extraction. Candidates of the same input share most of the words and the discriminators trained from scratch have difficulty differentiating them

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1137–1148
July 9-14, 2023 ©2023 Association for Computational Linguistics

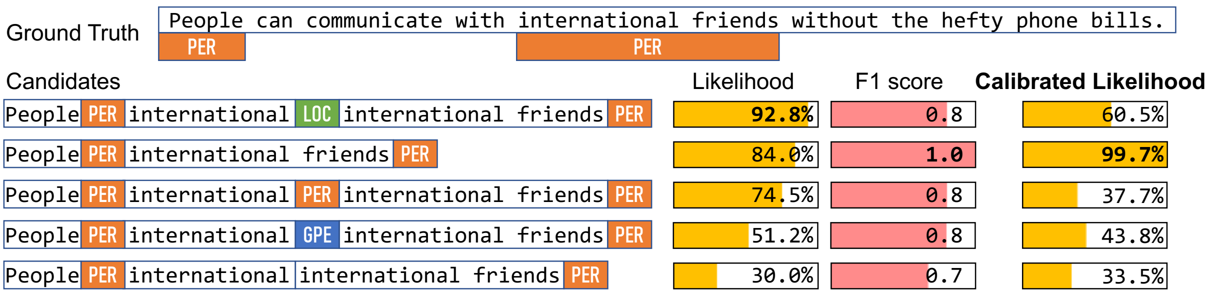
Ground Truth
People can communicate with international friends without the hefty phone bills.
PER
PER
Candidates
Likelihood
F1 score
Calibrated Likelihood
People PER international LOC international friends PER
92.8%
0.8
60.5%
People PER international friends PER
84.0%
1.0
99.7%
People PER international PER international friends PER
74.5%
0.8
37.7%
People PER international GPE international friends PER
51.2%
0.8
43.8%
People PER international international friends PER
30.0%
0.7
33.5%

Figure 1: Illustration of Sequence Likelihood Calibration. After guiding the estimated sequence likelihood by F1 score, the likelihood is more consistent with the F1 score. More cases can be found in Appendix 8.

(detailed in Sec. 3.3).

To address the above issue, we propose RerankNER to debias generative NER based on a reranking framework adapted for the task. Specifically, we first train the generative model in the standard way, resulting in a biased model. Then, we generate several candidates for each input with beam search. Instead of training a separated discriminator on the candidates sharing most of the words, we calibrate the generative model with a contrastive loss defined on the candidates. The contrastive loss aims to make the estimated sequence likelihoods consistent with their relative task performance as shown in Figure 1. This objective softens the target distribution and thus alleviates the bias.

Our contributions are summarized as follows:

1. To the best of our knowledge, we are the first to explore reranking-based methods in the field of generative information extraction (Ye et al., 2022).
2. We propose a method for generative NER tackling the bias problem.
3. Experimental results show that our method consistently boosts the baseline, and yields competitive results compared with the state-of-the-art methods on 8 widely-used datasets for NER.

2 Method

2.1 Task Formulation

We unify three NER subtasks (*i.e.* the flat, overlapped, and discontinuous NER) as follows. Given an input sentence of n tokens $X = x_1 x_2 \dots x_n$, the m output entities are arranged into a target sequence $Y = E_1 E_2 \dots E_m$, $E_i = y_i^1 y_i^2 \dots y_i^{j-1} y_i^j l_i$, where $y_i^1, ..., y_i^j$ denotes the tokens in the i-th entity and l_i denotes the label of the i-th entity. Our goal is to model the conditional probability $P(Y|X)$, which is factorized auto-regressively into $\prod_{t=0}^{|Y|} P(y_t|X, Y_{<t})$.

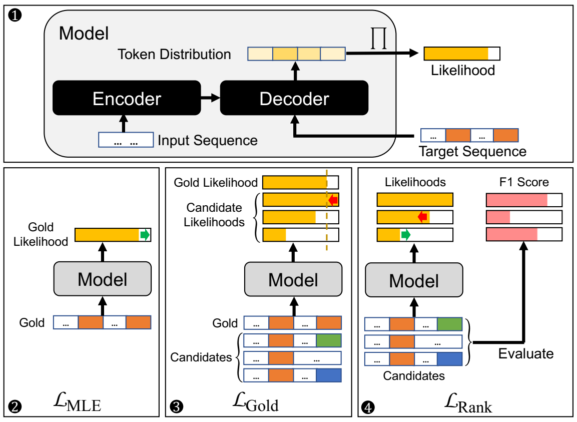

Figure 2: Illustration of the optimization objectives. ❶ The model estimates a sequence likelihood for each given target sequence. ❷ $\mathcal{L}_{\text{MLE}}$ maximizes the likelihood of the gold sequence. ❸ $\mathcal{L}_{\text{Gold}}$ penalizes any candidate with higher likelihood than the gold. ❹ $\mathcal{L}_{\text{Rank}}$ coordinates the likelihoods of the candidates with their ordering of F1 scores.

2.2 Overview

Given a generative NER model trained on the target sequences with the standard MLE, we perform sequence likelihood calibration to alleviate the bias. First, we generate several candidates for each input with beam search and evaluate their task performance (F1 score is used). Then, we continue training the model with the contrastive loss to make the estimated sequence likelihoods consistent with their relative task performance. Finally, we generate the answer with the standard beam search by the calibrated model.

2.3 Sequence Likelihood Calibration

The contrastive loss depicted in Figure 2 is composed of three terms $\mathcal{L}_{\text{MLE}}, \mathcal{L}_{\text{Rank}}, \mathcal{L}_{\text{Gold}}$.

$\mathcal{L}_{\text{MLE}}$ is identical to the standard MLE used in the first training stage. It maintains the generating ability of the model during the calibration process.

$\mathcal{L}_{\text{MLE}}$ maximizes the sequence likelihood of the gold target sequence Y, where the sequence likelihood is calculated as the product of token-level likelihood:

$$\mathcal{L}_{\text{MLE}} = -S(Y)$$
$$S(Y) = \sum_t \log P_\theta(y_t|X, Y_{<t})$$

and θ denotes model parameters.

$\mathcal{L}_{\text{Rank}}$ improves the consistency between the estimated sequence likelihoods and the task performance of the candidate sequences. We adopt the margin ranking loss (Hopkins and May, 2011) for this term, *i.e.*,

$$\mathcal{L}_{\text{Rank}} = \sum_{i,j} \max\left(0, S(\hat{Y}_j) - S(\hat{Y}_i) + \lambda\right)$$

where $\hat{Y}_i, \hat{Y}_j$ is a pair of candidates generated by beam search, provided that $\hat{Y}_i$ has a higher F1 score than $\hat{Y}_j$. λ denotes the margin, a hyper-parameter.

Apart from the supervision of relative order in the candidates, we utilize the supervision of the gold sequence as well. $\mathcal{L}_{\text{Gold}}$ ensures the sequence likelihoods of the generated candidates do not overstep the likelihood of the gold.

$$\mathcal{L}_{\text{Gold}} = \sum_{i} \max\left(0, S(\hat{Y}_i) - S(Y) + \lambda\right)$$

where $\hat{Y}_i$ denotes a candidate sequence, provided that it is not an equivalent of the gold.

The contrastive loss is the sum of the terms:

$$\mathcal{L} = \mathcal{L}_{\text{MLE}} + \alpha\mathcal{L}_{\text{Rank}} + \bar{\alpha}\mathcal{L}_{\text{Gold}}$$

where α and $\bar{\alpha}$ are coefficients.

3 Experiments

3.1 Main Results

We conduct experiments on eight datasets of three NER subtasks in total. Precision (P), Recall (R) and Micro F1 score (F1) are reported as previous works. We use BART-large as our backbone. For fair comparison, we reproduce BARTNER (Yan et al., 2021) using the public code [1] and get similar results reported in the paper. We compare our model principally with SOTA generative NER models, including (Yan et al., 2021; Zhang et al., 2022; Lu et al., 2022). Performances of SOTA discriminative NER models (Li et al., 2022) are also listed for reference. Refer to Appendix A for more details.

The results for flat, overlapped and discontinuous NER are shown in Table 2, Table 3 and Table 4 respectively. On eight datasets, our proposed sequence calibration consistently boosts the baseline. It achieves SOTA performance among the generative methods. Noting that our method gets competitive results even compared with discriminative methods that use extra embedding and domain pretrained model, which shows the potential of generative models.

3.2 Analysis of Improvement

We manually analyze the predictions corrected by the calibration. Apart from reranking the correct candidate to the top beam, RerankNER can generate new candidates with boundary or type corrected. More cases can be found in Appendix B.

In addition to manually observing examples, we also quantitatively analyze the sources of gain. We find that the gain mostly comes from samples with low likelihood, which means sequence likelihood calibration is more effective for samples with higher difficulty. Specifically, we group the samples in the test set into ten groups according to their original sequence likelihood and evaluate their performance before (colored in orange) and after (colored in blue) calibration. It can be seen from Figure 3 that the F1 scores of most groups get improved after calibration, and the improvement is greater for samples with lower likelihoods.

We also conduct the hit@top-k evaluation. Specifically, we iterate over the test samples and increase the number of hits when a gold answer exists among the top-k candidates. Table 5 shows that calibration slightly increase the hit@top-k across various datasets.

3.3 Variants of Reranker

As stated in Section 1, we observe that previous methods have difficulty capturing the subtle nuance among the candidates. We have investigated three variants: (1) SimCLS (Liu and Liu, 2021). (2) SimCLS with our modification which concatenates the input and the candidate representation and projects it to a score to replace the cosine similarity. (3) Picking out the best candidate based on the estimated likelihood of our model. Overall, we find their training losses fluctuate and their performance consistently lower than the baseline which selects the top beam with the highest likelihood. Future work could investigate this phenomenon in more depth.

[1] https://github.com/yhcc/BARTNER/

	Model	CoNLL03			OntoNotes5.0		
		P	R	F1	P	R	F1
Discriminative	(Akbik et al., 2019)[1] [BERT-Large]	-	-	92.86	-	-	-
	(Li et al., 2020)[3] [BERT-Large]	92.47	93.27	92.87	91.34	88.39	89.84
	(Shen et al., 2021)[2] [BERT-Large]	92.13	93.73	92.94	-	-	-
	(Wang et al., 2021a)[1] [BERT-Large]	-	-	93.21	-	-	-
	(Li et al., 2022) [BERT-Large]	92.71	93.44	93.07	90.03	90.97	90.50
Generative	(Straková et al., 2019)[1] [BERT-Large]	-	-	93.07	-	-	-
	(Zhang et al., 2022) [T5-Base]	92.78	93.51	93.14	89.77	91.07	90.42
	(Lu et al., 2022) [UIE (T5-Large)]	-	-	92.99	-	-	-
	(Yan et al., 2021) [BART-Large]	92.61	**93.87**	93.24	89.99	90.77	90.38
	Ours [BART-Large]	**93.26**	93.69	**93.48**	**90.03**	**91.24**	**90.63**

Table 2: Results on flat NER datasets. [1] means using extra embedding (*e.g.* character embedding and POS embedding). [2] means using extra context. [3] means reproduction from (Yan et al., 2021).

	Model	ACE04			ACE05			Genia		
		P	R	F1	P	R	F1	P	R	F1
Discriminative	(Yu et al., 2020)[2] [BERT-Large]	87.3	86.0	86.7	85.2	85.6	85.4	81.8	79.3	80.5
	(Li et al., 2020)[4] [BERT-Large]	85.83	85.77	85.80	85.01	84.13	84.57	81.25	76.36	78.72
	(Xu et al., 2021) [BERT-Large]	86.9	85.8	86.3	85.7	85.2	85.4	80.3	78.9	79.6
	(Shen et al., 2021)[2] [BERT-Large]	87.44	87.38	87.41	86.09	87.27	86.67	80.19	80.89	80.54
	(Li et al., 2022)[3] [BERT-Large]	87.33	87.71	87.52	85.03	88.62	86.79	83.10	79.76	81.39
Generative	(Straková et al., 2019) [BERT-Large]	-	-	84.40	-	-	84.33	-	-	78.31
	(Zhang et al., 2022) [T5-Base]	86.36	84.54	85.44	82.92	87.05	84.93	**81.04**	77.21	79.08
	(Lu et al., 2022) [UIE (T5-Large)]	-	-	86.89	-	-	85.78	-	-	-
	(Yan et al., 2021) [BART-Large]	87.27	86.41	86.84	83.16	86.38	84.74	78.57	79.3	78.93
	Ours [BART-Large]	**87.64**	**87.61**	**87.63**	**85.01**	**87.47**	**86.22**	79.51	**79.48**	**79.49**

Table 3: Results on overlapped NER datasets. [1] means using extra embedding. [2] means using extra context. [3] means using domain pretrained model (*e.g.* ClinicalBERT and BioBERT). [4] means reproduction from (Yan et al., 2021)

	Model	CADEC			ShARe13			ShARe14		
		P	R	F1	P	R	F1	P	R	F1
Discriminative	(Tang et al., 2018)	67.80	64.99	66.36	-	-	-	-	-	-
	(Dai et al., 2020) [ELMO]	68.90	69.00	69.00	80.50	75.00	77.70	78.10	81.20	79.60
	(Li et al., 2020) [BERT-large]	-	-	69.90	-	-	82.50	-	-	-
	(Wang et al., 2021b)[1] [BERT-Large]	70.50	72.50	71.50	84.30	78.20	81.20	78.20	84.70	81.30
	(Li et al., 2022)[1] [BERT-Large]	74.09	72.35	73.21	85.57	79.68	82.52	79.88	83.71	81.75
Generative	(Zhang et al., 2022) [T5-Base]	71.35	**71.86**	71.60	81.09	78.13	79.58	77.88	**83.77**	80.72
	(Yan et al., 2021) [BART-Large]	70.08	71.21	70.64	**82.09**	77.42	79.69	77.2	83.75	80.34
	Ours [BART-Large]	**72.33**	71.01	**71.66**	81.86	**78.48**	**80.14**	**78.68**	83.63	**81.01**

Table 4: Results on discontinuous NER datasets. [1] means using domain pretrained model (*e.g.* ClinicalBERT and BioBERT).

	CoNLL03	OntoNotes5.0	ACE04	ACE05
hit@3	3196/3119	7732/7734	559/566	759/779
hit@5	3240/3138	7858/7869	582/586	786/797
	GENIA	CADEC	ShARe13	ShARe14
hit@3	1135/1161	981/962	8046/8077	14405/14578
hit@5	1245/1254	1005/980	8085/8124	14481/14659

Table 5: Hit@top-k evaluation. Each element in the table denotes the hit count among top-k candidates before/after calibration.

4 Related Work

Named Entity Recognition The existing methods for NER can be broadly classified into sequence labeling formulation, span-based formulation and generative-based formulation. A majority of initial works adopt sequence labeling formulation which assigns each token a tag from a predefined tagging scheme (Huang et al., 2015; Lample et al., 2016). Then, the span-based formulation is proposed which enumerates all possible spans and

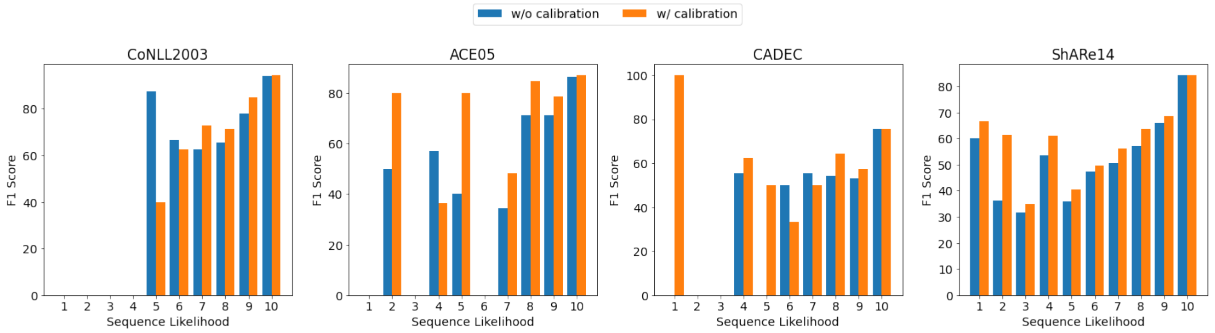

Figure 3: Distribution of F1 scores for samples with different sequence likelihood before (blue) and after (orange) calibration. X-axis shows 10 groups of samples categorized by their sequence likelihood before calibration. Y-axis shows the F1 score of each group.

performs classification at the span-level (Wang and Lu, 2019). Recently, researchers have grown more interest in tackling the three subtasks uniformly, i.e., flat NER, overlapped NER and discontinuous NER. We refer to them as unified NER in the rest of the passage. The above two formulations have major drawbacks in modeling unified NER. For example, sequence labeling methods need to design different tagging schemas for each subtask (Dai et al., 2020). While span-based methods have to trade-off between maximal span length and computation efficiency due to the enumeration operation (Luan et al., 2019). Generative-based formulation prevails in unified NER for its flexibility in generating variable-length entities (Lu et al., 2022; Yan et al., 2021). In this paper, we adopt BART-NER (Yan et al., 2021) as our backbone generative model.

Bias in Generative NER Since the generative model generates outputs in an autoregressive manner which differs largely from the extraction objective of NER, it introduces incorrect biases during training. (Zhang et al., 2022) analyze these biases from the causality perspective and attribute them to two confounders namely pre-context confounder (the model may be biased to pre-generated words which have no causal relation with the word to be generated) and entity-order confounder. They propose two data augmentation methods to address them respectively. (Tan et al., 2021) observe that overlapped NER is essentially an unordered task and propose a sequence-to-set network to predict entity spans simultaneously in a non-autoregressive manner. W^{2}NER (Li et al., 2022) abandons the generative-based formulation and model unified NER as a word-word relation classification based on the proposed relation schema. In this paper, we improve the generative-based method by exploiting the candidate information and get comparable or better results.

Reranking Reranking has been explored in various tasks of Natural Language Processing for long. In question answering, passage reranking is used as the first stage to retrieve relevant passages where the answer might locate and reorder them according to their scores. Similarly, answer reranking is used as the last stage to refine the answer selection. In neural machine translation, (Bhattacharyya et al., 2021) apply an energy-based model on the top of BERT to reorder candidates according to their BLEU scores. In abstractive summarization, SimCLS (Liu and Liu, 2021) trains a separate second-stage model with discriminative ranking loss to select the best summary candidate. BRIO (Liu et al., 2022) optimizes the autoregressive language model by a contrastive loss over the discrete space of the generated texts. SummaReranker (Ravaut et al., 2022) adopts a mixture-of-expert architecture as the reranker to measure the quality of the candidates with multiple metrics. To the best of our knowledge, there is no work exploring reranking methods on generative IE.

5 Conclusion

Through pilot experiments, we find the decoded candidates provide potential supervision. Based on this finding, we propose RerankNER to debias generative NER based on a reranking framework adapted for the task. It consistently boosts the baseline and achieves competitive results with state-of-the-art generative methods on eight NER datasets, which verifies the effectiveness of candidate order supervision. Future work could consider extending this method to other generative IE tasks. Another meaningful direction is to consider incorporating Large Language Models into the reranking process.

Limitations

RerankNER conducts calibration after the regular training, which introduces extra computational overhead. This drives us to further improve the overall efficiency of our method. Recent works find that few-shot learning serves as an effective finetuning method of pretrained language models. It is reasonable to investigate our model under few-shot learning to reduce the overhead. Although we get competitive results with the state-of-the-art methods, there is still a gap between the oracle score and the best results. We leave them as our future work.

Acknowledgement

We thank the anonymous reviewers for their helpful comments on this paper. This work was partially supported by National Key R&D Program of China (No. 2022YFC3600402) and National Social Science Foundation Project of China (21&ZD287). The corresponding author of this paper is Sujian Li.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 724–728. Association for Computational Linguistics.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4528–4537, Online. Association for Computational Linguistics.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An effective transition-based model for discontinuous NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870, Online. Association for Computational Linguistics.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel, and Ralph M. Weischedel. 2004. The automatic content extraction (ACE) program - tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal*. European Language Resources Association.

Mark Hopkins and Jonathan May. 2011. Tuning as ranking. In *Conference on Empirical Methods in Natural Language Processing*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5849–5859. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3036–3046, Minneapolis, Minnesota. Association for Computational Linguistics.

Danielle L. Mowery, Sumithra Velupillai, Brett R. South, Lee M. Christensen, David Martínez, Liadh Kelly, Lorraine Goeuriot, Noémie Elhadad, Sameer Pradhan, Guergana K. Savova, and Wendy W. Chapman. 2013a. Task 1: Share/clef ehealth evaluation lab 2013. In *Conference and Labs of the Evaluation Forum*.

Danielle L. Mowery, Sumithra Velupillai, Brett R. South, Lee M. Christensen, David Martínez, Liadh Kelly, Lorraine Goeuriot, Noémie Elhadad, Sameer S. Pradhan, Guergana K. Savova, and Wendy W. Chapman. 2013b. Task 2 : Share/clef ehealth evaluation lab 2014.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 143–152. ACL.

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pages 142–147. ACL.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and label: A two-stage identifier for nested named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2782–2794, Online. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural architectures for nested NER through linearization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.

Zeqi Tan, Yongliang Shen, Shuai Zhang, Weiming Lu, and Yueting Zhuang. 2021. A sequence-to-set network for nested named entity recognition. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI-21*.

Buzhou Tang, Jianglu Hu, Xiaolong Wang, and Qingcai Chen. 2018. Recognizing continuous and discontinuous adverse drug reaction mentions from social media using LSTM-CRF. *Wirel. Commun. Mob. Comput.*, 2018.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus LDC2006T06.

Bailin Wang and Wei Lu. 2019. Combining spans into entities: A neural two-stage approach for recognizing discontiguous entities. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6216–6224, Hong Kong, China. Association for Computational Linguistics.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021a. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1800–1812. Association for Computational Linguistics.

Yucheng Wang, Bowen Yu, Hongsong Zhu, Tingwen Liu, Nan Yu, and Limin Sun. 2021b. Discontinuous named entity recognition as maximal clique discovery. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 764–774. Association for Computational Linguistics.

Yongxiu Xu, Heyan Huang, Chong Feng, and Yue Hu. 2021. A supervised multi-head self-attention network for nested named entity recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14185–14193. AAAI Press.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822, Online. Association for Computational Linguistics.

Hongbin Ye, Ningyu Zhang, Hui Chen, and Huajun Chen. 2022. Generative knowledge graph construction: A review. *arXiv preprint arXiv:2210.12714*.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 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.

Shuai Zhang, Yongliang Shen, Zeqi Tan, Yiquan Wu, and Weiming Lu. 2022. De-bias for generative extraction in unified ner task. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 808–818.

A Details

A.1 Dataset Statistics

The statistics of the datasets are listed in Table 6.

Flat NER subtask We conduct experiments on CoNLL-2003 (Sang and Meulder, 2003) and OntoNotes5.0 (Pradhan et al., 2013) in English. We follow the experimental settings as previous works (Lample et al., 2016; Yan et al., 2021).

Overlapped NER subtask We conduct experiments on ACE 2004 (Doddington et al., 2004), ACE 2005 (Walker et al., 2006), and GENIA (Kim et al., 2003). For ACE 2004 and ACE 2005, we shuffle and split the documents into training, development, and testing in a ratio of 8:1:1 following (Yu et al., 2020). For GENIA, the ratio is set to 8.1:0.9:1.0 following (Yan et al., 2021).

Discontinuous NER subtask We conduct experiments on CADEC (Karimi et al., 2015), ShARe13 (Mowery et al., 2013a), and ShARe14 (Mowery et al., 2013b). These datasets contains approximately 10% discontinuous entities. We follow the experimental settings from (Dai et al., 2020).

A.2 Implementation Details

For the fine-tuning stage, we use the code, the hyper-parameters, the package version from (Yan et al., 2021) and get comparable results on all datasets reported in the paper. We set the max epoch as 30 with early stop (patience=5). We use AdamW optimizer with the same learning rate as (Yan et al., 2021). Linear learning rate scheduling is employed. For all subtasks, we do predictions on the word-level, i.e., only the position index of the first BPE of each entity word is used.

For the calibration training, we use the standard beam search to generate 5 candidates for each input sentence. We adopt the hyper-parameters as the fine-tuning stage except for the newly added ones. We implement both the fixed margin and the linear margin. The linear margin $\lambda = \bar{\lambda}(j - i)$ denotes the linear margin depending on the order difference of the candidates, and $\bar{\lambda}$ is a hyper-parameter. We search the value of the margin $\bar{\lambda}$ within [0.01, 0.1]. We search the value of coefficient α within [0.1, 1]. Table 7 “mask out tie” means whether we mask out the comparison between candidates with the same F1 score in the contrastive loss. Effects of “add $\mathcal{L}_{Gold}$” and “mask out tie” differs across 8 datasets, so we view them as hyper-parameters. All experiments are conducted on the NVIDIA RTX 3090 GPU with 24G memory.

A.3 Baselines

The following methods can adapt to all NER subtasks. Please refer to the original papers for the other methods designed specifically for a certain NER subtask.

BERT-MRC (Li et al., 2020) reformulates NER as a machine reading comprehension (MRC) task and extract entities by answering questions such as "find locations in the text".

UIE (Lu et al., 2022) represents various information structures with a structured extraction language and tackles general information extraction tasks with a unified text-to-structure generation framework.

(Zhang et al., 2022) analyzes incorrect biases in the generative NER models from the causality perspective and proposes two data augmentation methods to address them. Note that T5-Base they use has the same number of Transformer layers as BART-Large.

W^{2}NER (Li et al., 2022) reformulates unified NER as a word-word relation classification task based on the proposed relation schema.

		Sentence					Mention			
		#All	#Train	#Dev	#Test	Avg.Len	#All	#Ovlp.	#Dis.	Avg.Len
Flat	CoNLL2003	20744	17291	-	3453	14.38	35089	-	-	1.45
	OntoNotes5.0	76714	59924	8528	8262	18.11	104151	-	-	1.83
Ovlp.	GENIA	18546	15023	1669	1854	25.41	56015	10263	-	1.97
	ACE04	8512	6802	813	897	20.12	27604	12626	-	2.50
	ACE05	9697	7606	1002	1089	17.77	30711	12404	-	2.28
Dis.	CADEC	7597	5340	1097	1160	16.18	6316	920	679	2.72
	ShARe13	18767	8508	1250	9009	14.86	11148	663	1088	1.82
	ShARe14	34614	17404	1360	15850	15.06	19070	1058	1656	1.74

Table 6: Dataset Statistics. "Ovlp." and "Dis." denote overlapped and discontinuous mentions respectively.

Hyper-parameter	Value
epoch	30
warmup step	0.01
learning rate	[1e-5, 2e-5, 4e-5]
batch size	[16, 24, 32]
beam size	5
margin $\bar{\lambda}$	[0.01, 0.1]
coefficient $\alpha = \bar{\alpha}$	[0.1, 1.0, 5.0]
add $\mathcal{L}_{Gold}$	[Yes, No]
mask out tie	[Yes, No]

Table 7: Hyper-parameter settings.

B Case Study

Table 8 shows some examples corrected by the sequence likelihood calibration.

C Generative Model

Our method is agnostic to the generative model. In this study, we adopt BARTNER (Yan et al., 2021), an Encoder-Decoder framework with pointer mechanism, to model the probability $P(Y|X)$:

Encoder encodes the input sentence X into vectors H^{Enc}, which can be denoted as:

$$H^{\text{Enc}} = \text{Encoder}(X) \tag{1}$$

where $H^{\text{Enc}} \in R^{n \times d}$ and d is the dimension of the hidden state.

Decoder predicts the index probability distribution step-by-step according to $P(y_t|X, Y_{<t})$. Since $Y_{<t}$ consists of the indices of the pointers and tags, it needs to be mapped to the vocabulary indices before inputted to the Decoder. We get the hidden state at the t-th step by:

$$h_t^{\text{Dec}} = \text{Decoder}(H^{\text{Enc}}; \hat{Y}_{<t}) \tag{2}$$

Finally, we get the index probability distribution P_t by:

$$\begin{aligned} G^{\text{Dec}} &= \text{Embed}(G) \\ E^{\text{Enc}} &= \text{Embed}(X) \\ \hat{H}^{\text{Enc}} &= \alpha * H^{\text{Enc}} + (1-\alpha) * E^{\text{Enc}} \\ P(y_t|X, Y_{<t}) &= \text{Softmax}([\hat{H}^{\text{Enc}} \otimes h_t^{\text{Dec}}; G^{\text{Dec}} \otimes h_t^{\text{Dec}}]) \end{aligned} \tag{3}$$

where $\text{Embed}(\cdot)$ is the embedding layer shared between the Encoder and Decoder, G denotes the label token while X denotes the entity words. $\hat{H}^{\text{Enc}}$ denotes the input representation. $\otimes$ denotes the dot product. For training, we use the cross-entropy loss with teacher forcing. During inference, we generate the target sequence auto-regressively.

CADEC (2=ADR)	
Muscle twitching, stiff neck, constant lightheadedness, always worrying about a brain tumor or something.	
0.50,-0.05,Muscle twitching 2 stiff neck 2 constant lightheadedness 2 always worrying about a brain 2 0.50,-0.12,Muscle twitching 2 stiff neck 2 constant lightheadedness 2 always worrying about a brain tumor 2 0.75,-0.15,Muscle twitching 2 stiff neck 2 constant lightheadedness 2 always worrying 2 0.50,-0.18,Muscle twitching 2 stiff neck 2 constant lightheadedness 2 always worrying about a 2 0.50,-0.23,Muscle twitching 2 stiff neck 2 constant lightheadedness 2 always worrying about a tumor 2	
0.86,-0.07,Muscle twitching 2 stiff neck 2 lightheadedness 2 0.75,-0.22,Muscle twitching 2 stiff neck 2 lightheadedness 2 always worrying about a brain tumor 2 0.50,-0.33,Muscle twitching 2 stiff neck 2 constant lightheadedness 2 always worrying about a brain tumor 2 0.75,-0.34,Muscle twitching 2 stiff neck 2 lightheadedness 2 always worrying about a brain tumor or something 2 0.75,-0.39,Muscle twitching 2 stiff neck 2 stiff neck 2 lightheadedness 2 always worrying about a brain tumor 2	
Possibly diarrhea and stomach pain, but most likely none because I am taking this with a nasty antibiotic for a sinus infection that definitely causes diarrhea and nausea.	I stopped taking it the next day and within 72 hours my swelling decreased significantly, muscle aches and joint pain disappeared, memory loss is not as severe, breathing is easier, stamina is back etc.
0.67,-0.05,diarrhea 2 stomach pain 2 1.00,-0.14,diarrhea 2 stomach pain 2 diarrhea 2 nausea 2 0.86,-0.49,diarrhea 2 stomach pain 2 nausea 2 0.88,-0.57,diarrhea 2 stomach pain 2 diarrhea 2 nausea 2 0.86-0.76,diarrhea 2 stomach pain 2 diarrhea 2 nausea 2	0.80,-0.004,swelling 2 muscle aches 2 joint pain 2 memory loss 2 breathing is 2 stamina is 2 0.80,-0.16,swelling 2 muscle aches 2 joint pain 2 memory loss 2 breathing is 2 stamina is 2 0.89,-0.22,swelling 2 muscle aches 2 joint pain 2 memory loss 2 breathing is 2 stamina is 2 0.89,-0.30,swelling 2 muscle aches 2 joint pain 2 memory loss 2 breathing is 2 stamina is 2 0.89,-0.41,swelling 2 muscle aches 2 joint pain 2 memory loss 2 breathing is 2 stamina is 2
1.00,-0.08,diarrhea 2 stomach pain 2 diarrhea 2 nausea 2 0.86,-0.23,diarrhea 2 stomach pain 2 diarrhea 2 nausea 2 0.86,-0.29,diarrhea 2 stomach pain 2 diarrhea 2 nausea 2 1.00,-0.44,diarrhea 2 stomach pain 2 diarrhea 2 nausea 2 1.00,-0.53,diarrhea 2 stomach pain 2 diarrhea 2 nausea 2 stomach pain 2	1.00,-0.09,swelling 2 muscle aches 2 joint pain 2 memory loss 2 0.80,-0.23,swelling 2 muscle aches 2 joint pain 2 memory loss 2 breathing is 2 stamina is 2 0.80,-0.25,swelling 2 muscle aches 2 joint pain 2 memory loss 2 breathing is 2 stamina is 2 0.89,-0.26,swelling 2 muscle aches 2 joint pain 2 memory loss 2 breathing is 2 stamina is 2 0.80-0.27,swelling 2 muscle aches 2 joint pain 2 memory loss 2 breathing is 2 stamina 2
CONLL2003 (2=LOC,3=PER,4=ORG,5=MISC)	
POLAND GOT MONEY FROM POST-WAR SWISS ACCOUNTS.	Mike Cito, 17, was expelled from St Pius X High School in Albuquerque after an October game in which he used the sharpened chin strap buckles to injure two opposing players and the referee.
0.80,-0.08,POLAND 2 POST-WAR 5 SWISS 5 1.0,-0.23,POLAND 2 SWISS 5 0.50,-0.37,POLAND 2 POST-WAR SWISS 5 1.0,-1.33,POLAND 2 POST-WAR POST-WAR 5 SWISS 5 0.80,-1.41,POLAND 2 POST-WAR ACCOUNTS 5 SWISS 5	0.67,-0.01,Mike Cito 3 St Pius X 4 Albuquerque 2 1.0,-0.19,Mike Cito 3 St Pius X High School 4 Albuquerque 2 0.67,-0.48,Mike Cito 3 St Pius X School 4 Albuquerque 2 0.67,-0.58,Mike Cito 3 St Pius X 3 Albuquerque 2 0.67,-0.61,Mike Cito 3 St Pius X 2 Albuquerque 2
1.0,-0.1.0,POLAND 2 SWISS 5 0.80,-0.25,POLAND 2 POST-WAR 5 SWISS 5 0.80,-0.74,POLAND 2 POST-WAR 5 SWISS 5 SWISS 5 1.0,-0.76,POLAND 2 SWISS 5 SWISS 5 0.80,-0.78,POLAND 2 POST-WAR 5 SWISS 5 POST-WAR 5	1.0,-0.11,Mike Cito 3 St Pius X High School 4 Albuquerque 2 0.67,-0.14,Mike Cito 3 St Pius X 4 Albuquerque 2 0.80,-0.41,Mike Cito 3 St Cito X High School 4 Albuquerque 2 0.67,-0.44,Mike Cito 3 St Pius X School 4 Albuquerque 2 0.86,-0.48,Mike Cito 3 St Pius X High School 4 Albuquerque 2 St Pius X 4
There is the international prestige Singapore would enjoy, but "more importantly there is a genuine national interest in fostering better global free trade and an open market", said Tan Kong Yam, head of Business Policy at the National University of Singapore.	
0.86,-0.04,Singapore 2 Tan Kong Yam 3 Business Policy 4 National University of Singapore 4 1.0,-0.07,Singapore 2 Tan Kong Yam 3 National University of Singapore 4 0.86,-0.36,Singapore 2 Tan Kong Yam 3 Business Policy 5 National University of Singapore 4 0.67,-0.68,Singapore 2 Tan Kong Yam 3 Business Policy 4 National University of Singapore 4 0.80,-0.68,Singapore 2 Tan Kong Yam 3 Business Policy of National University of Singapore 4	
1.0,-0.06,Singapore 2 Tan Kong Yam 3 National University of Singapore 4 1.0,-0.34,Singapore 2 Tan Kong Yam 3 National University of Singapore 4 National University of Singapore 4 1.0,-0.44,Singapore 2 Tan Kong Yam 3 National University of Singapore 4 Tan Kong Yam 3 0.86,-0.44,Singapore 2 Tan Kong Yam 3 National University of Singapore 4 Singapore 2 0.80,-0.45,Singapore 2 Tan Kong Yam 3 National University of Singapore 4	
ACE04 (2=LOC,3=GPE,4=WEA,5=VEH,6=PER,7=ORG,8=FAC)	
I believe our issues do relate directly to the appointing of electors for the state of Florida.	
0.67,-0.06,I 6 our 3 electors for the state of Florida 6 the state of Florida 3 0.80,-0.31,I 6 our 3 electors for the state of Florida 6 the state of Florida 3 Florida 3 0.89,-0.32,I 6 our 6 electors for the state of Florida 6 the state of Florida 3 0.50,-0.32,I 6 our 3 electors for the state of Florida 6 the state of Florida 3 0.50,-0.33,I 6 our 3 electors for the state of Florida 6 the state of Florida 3	
1.0,-0.01,I 6 our 6 electors for the state of Florida 6 the state of Florida 3 Florida 3 0.89,-0.10,I 6 electors for the state of Florida 6 the state of Florida 3 Florida 3 0.80,-0.20,I 6 our 3 electors for the state of Florida 6 the state of Florida 3 Florida 3 0.91,-0.47,I 6 our 6 electors for the state of Florida 6 the state of Florida 3 state 3 Florida 3 0.89,-0.48,I 6 our 6 electors for the state of Florida 6 the state of Florida 3	
One hundred South Koreans will be in the northern capital Pyongyang, to meet their North Korean relatives.	
0.83,-0.08,One hundred South Koreans 6 the northern capital 3 the northern capital Pyongyang 3 their 6 their North Korean relatives 6 North Korean 3 0.77,-0.09,One hundred South Koreans 6 South 3 the northern capital 3 the northern capital Pyongyang 3 their 6 their North Korean relatives 6 North Korean 3 0.77,-0.17,One hundred South Koreans 6 South 2 the northern capital 3 the northern capital Pyongyang 3 their 6 their North Korean relatives 6 North Korean 3 0.67,-0.20,One hundred South Koreans 6 the northern capital 3 the northern capital Pyongyang 3 their 3 their North Korean relatives 6 North Korean 3 0.62,-0.20,One hundred South Koreans 6 South 3 the northern capital 3 the northern capital Pyongyang 3 their 3 their North Korean relatives 6 North Korean 3	
0.92,-0.03,One hundred South Koreans 6 South 3 the northern capital 3 Pyongyang 3 their 6 their North Korean relatives 6 North Korean 3 1.0,-0.03,One hundred South Koreans 6 the northern capital 3 Pyongyang 3 their 6 their North Korean relatives 6 North Korean 3 0.92,-0.12,One hundred South Koreans 6 South Koreans 6 the northern capital 3 Pyongyang 3 their 6 their North Korean relatives 6 North Korean 3 0.77,-0.17,One hundred South Koreans 6 South Koreans 6 the northern capital 3 the northern capital Pyongyang 3 their 6 their North Korean relatives 6 North Korean 3 0.92,-0.23,One hundred South Koreans 6 South Koreans 3 the northern capital 3 Pyongyang 3 their 6 their North Korean relatives 6 North Korean 3	
Netanyahu supporters are calling either for a change in the law or for simultaneous elections for the Knesset and Prime Minister, which would allow their candidate to run.	
0.73,-0.07,Netanyahu 6 Netanyahu supporters 6 the Knesset and Prime Minister 6 their 6 their candidate 6 1.0,-0.17,Netanyahu 6 Netanyahu supporters 6 the Knesset 7 Prime Minister 6 their 6 their candidate 6 0.5454545454544859,-0.347797691822052,Netanyahu 6 Netanyahu candidate 6 the Knesset and Prime Minister 6 their 6 their candidate 6 0.7999999999999359,-0.3507174551486969,Netanyahu 6 Netanyahu supporters 6 the Knesset and Prime Prime Minister 6 their 6 their candidate 6 0.8333333333332694,-0.35162267088890076,Netanyahu 6 Netanyahu supporters 6 Knesset 7 Prime Minister 6 their 6 their candidate 6	
0.9999999999999332,-0.010539926588535309,Netanyahu 6 Netanyahu supporters 6 the Knesset 7 Prime Minister 6 their 6 their candidate 6 0.7272727272726645,-0.10700102150440216,Netanyahu 6 Netanyahu supporters 6 the Knesset and Prime Minister 6 their 6 their candidate 6 0.7272727272726645,-0.37671196460723877,Netanyahu 6 Netanyahu supporters 6 the Knesset and Prime Minister 7 their 6 their candidate 6 0.7272727272726645,-0.6056239604949951,Netanyahu 6 Netanyahu supporters 6 the Knesset and Prime Minister candidate 6 their 6 their candidate 6 0.8333333333332694,-0.6295873522758484,Netanyahu 6 Netanyahu supporters 6 the Knesset 6 Prime Minister 6 their 6 their candidate 6	

Table 8: Case Study. Candidates before (upper) and after (lower) calibration. Each candidate is formatted as "F1, log-probability, target sequence". The number denotes the corresponding entity type.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
After Conclusion.

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
The first.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Appendix A

☑ B1. Did you cite the creators of artifacts you used?
Appendix A

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We obtained proper licencing and will not distribute the artifacts.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We stick to the intended use only.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix A

C ☑ Did you run computational experiments?

3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
A.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
A.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
A.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
A.2

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.