NameTag 3: A Tool and a Service for Multilingual/Multitagset NER

Jana Straková and Milan Straka
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské nám. 25, Prague, Czech Republic
{strakova,straka}@ufal.mff.cuni.cz

Abstract

We introduce NameTag 3, an open-source tool and cloud-based web service for multilingual, multidataset, and multitagset named entity recognition (NER), supporting both flat and nested entities. NameTag 3 achieves state-of-the-art results on 21 test datasets in 15 languages and remains competitive on the rest, even against larger models. It is available as a command-line tool and as a cloud-based service, enabling use without local installation. NameTag 3 web service currently provides flat NER for 17 languages, trained on 21 corpora and three NE tagsets, all powered by a single 355M-parameter fine-tuned model; and nested NER for Czech, powered by a 126M fine-tuned model. The source code is licensed under open-source MPL 2.0, while the models are distributed under non-commercial CC BY-NC-SA 4.0. Documentation is available at https://ufal.mff.cuni.cz/nametag, source code at https://github.com/ufal/nametag3, and trained models via https://lindat.cz. The REST service and the web application can be found at https://lindat.mff.cuni.cz/services/nametag/. A demonstration video is available at https://www.youtube.com/watch?v=-gaGnP0IV8A.

1 Introduction

Named entity recognition (NER), the task of identifying proper names such as persons, locations, and organizations in natural text, is a fundamental preprocessing step in many natural language processing (NLP) and knowledge extraction systems. While both flat and nested (embedded) NER have been extensively researched, particularly for English, many other languages still lack off-the-shelf, open-source NER tools that can be easily integrated into academic and research workflows.

We introduce NameTag 3, an open-source tool, web application, and web service for both flat and nested named entity recognition. NameTag 3 achieves state-of-the-art performance on 21 test datasets across 15 languages: Cebuano, Chinese, Croatian, Czech, Danish, English, Norwegian Bokmål, Norwegian Nynorsk, Portuguese, Russian, Serbian, Slovak, Swedish, Tagalog, and Ukrainian. Additionally, it delivers competitive results on Arabic, Dutch, German, Maghrebi, and Spanish.

The key characteristics of NameTag 3 are:

- open-source NER tool,
- support for both flat and nested NER,
- availability as command-line tool, web application, or cloud-based REST API webservice, allowing use without installation,
- an open-source MPL 2.0 license for code,
- a non-commercial CC BY-NC-SA 4.0 license for models,
- trained models,
- support for training custom models,
- modestly-sized models (126M or 355M),
- SOTA on 21 datasets in 15 languages.

Lastly, given the recent accomplishments of large language models, we also perform zero-shot and few-shot evaluations of DeepSeek-R1, demonstrating that when training data are available, NameTag 3 undoubtedly delivers substantially better performance while requiring several orders of magnitude fewer resources.

2 Related Work

One of the most well-known NLP pipelines for NER is Stanza (Qi et al., 2020), a neural-based framework developed by the Stanford NLP Group. Stanza provides pre-trained models for multiple languages.[1] This pipeline is based on pre-BERT, frozen contextual character-level word embeddings (Akbik et al., 2018) with Bi-LSTM and CRF

[1] https://stanfordnlp.github.io/stanza/ner_models.html

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 31–39
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

	NameTag 3	Stanza	SpaCy
Languages	17	29	24
Architecture	fine-tuned PLM	frozen Flair embeddings, Bi-LSTM + CRF	fine-tuned PLM or CNN
Flat NER	✓	✓	✓
Nested NER	✓	✗	✗
Single multilingual model	✓	✗	✓
Cross-lingual transfer	✓	✗	✓
Cloud-based service running	✓	✗	✗

Table 1: High-level technical and architectural overview of NameTag 3, Stanza, and SpaCy.

(Huang et al., 2015) layers on top.

Another known NLP pipeline is SpaCy (Honnibal and Montani, 2017). SpaCy is a free, open-source library for advanced Natural Language Processing (NLP) in Python. SpaCy uses multitask learning with pretrained transformers like BERT in its newer models, and CNNs in its older models.

Since 2014, NameTag has provided NER for Czech and English in academic settings as NameTag 1 (Straková et al., 2014). In 2019, NameTag 2 (Straková et al., 2019) expanded to six languages — English, German, Dutch, Spanish, Czech, and Ukrainian — each with a separately trained model.

This publication introduces NameTag 3, which surpasses its predecessors by improving F1 scores and further expands the number of languages available. Unlike NameTag 2, which used a Bi-LSTM layer over frozen multilingual BERT embeddings, NameTag 3 fine-tunes pre-trained models with either a softmax head for flat NER or a seq2seq head for nested NER, and adds multitagset learning.

Compared to Stanza, NameTag 3 so far supports fewer languages overall but includes some that Stanza does not cover. While Stanza employs a Bi-LSTM over frozen contextualized embeddings and trains separate models for each language, NameTag 3 takes a different approach. It is a fine-tuned PLM trained as a single joint model across multiple languages, datasets, and tagsets, enabling cross-lingual transfer even for languages not present in the training data. Additionally, NameTag 3 supports nested NER and provides a cloud-based web service.

A high-level technical and architectural overview of NameTag 3, Stanza, and SpaCy is available in Table 1, and the performance evaluation in F1 is presented in Table 3.

3 Data

3.1 Flat NE Datasets

We utilized the following flat NE datasets, adhering to their official train/dev/test splits for training, tuning, and evaluation, respectively. All UNER corpora were released under the UniversalNER v1 (UNER) initiative (Mayhew et al., 2024).[2] All OntoNotes 5.0 corpora follow the CoNLL-2012 train/dev/test split (Pradhan et al., 2012) over the original OntoNotes 5.0 data.[3]

- **Arabic OntoNotes 5.0**,
- **Chinese OntoNotes 5.0**,
- **Chinese UNER GSDSIMP**,
- **Chinese UNER GSD**,
- **Croatian UNER SET**,
- **Czech CNEC 2.0 CoNLL** — In order to train and serve the Czech Named Entity Corpus 2.0 (Ševčíková et al., 2007) jointly within a large multilingual model, the original annotation of the CNEC 2.0 has been harmonized to the standard 4-label tagset with `PER`, `ORG`, `LOC`, and `MISC`, resulting in an extensive simplification of the original annotation and flattening of the original nested entities.
- **Danish UNER DDT**,
- **Dutch CoNLL-2002** (Tjong Kim Sang, 2002),
- **English OntoNotes 5.0**,
- **English UNER EWT**,
- **English CoNLL-2003** (Tjong Kim Sang and De Meulder, 2003),
- **German CoNLL-2003** (Tjong Kim Sang and De Meulder, 2003),

[2] https://www.universalner.org/
[3] https://catalog.ldc.upenn.edu/LDC2013T19

	Flat Mono & Multi	Nested ACE 2004	Nested ACE 2005	Nested CNEC 2.0
Encoder	XLM-R Large	RoBERTa Large	RoBERTa Large	RobeCzech Base
Frozen epochs	0	20	20	20
Frozen learning rate	–	1e-3	1e-3	1e-3
Epochs	30	60	50	20
Batch size	8	8	16	4
Peak learning rate	2e-5	2e-5	2e-5	2e-5
Warmup epochs	1	1	1	1
Learning rate decay	cosine	cosine	cosine	cosine

Table 2: Training hyperparameters.

- **Maghrebi Arabic French UNER Arabizi**,
- **Norwegian Bokmål UNER NDT**,
- **Norwegian Nynorsk UNER NDT**,
- **Portuguese UNER Bosque**,
- **Serbian UNER SET**,
- **Slovak UNER SNK**,
- **Spanish CoNLL-2002** (Tjong Kim Sang, 2002),
- **Swedish UNER Talbanken**,
- **Ukrainian Lang-uk** — Ukrainian Lang-uk NER corpus[4] based on the Lang-uk initiative.[5] The corpus uses four classes `PER`, `ORG`, `LOC`, and `MISC`. (Please note that we harmonized the original `PERS` to a more common `PER`.)

For cross-lingual/out-of-domain evaluation on unseen languages/datasets, respectively, we used the following UNER (Mayhew et al., 2024) test datasets: **Cebuano UNER GJA**, **Chinese UNER PUD**, **Portuguese UNER PUD**, **Russian UNER PUD**, **Swedish UNER**, **Tagalog UNER TRG**, and **Tagalog UNER Ugnayan**.

3.2 Nested NE Datasets

We evaluate NameTag 3 on the following nested NE corpora:

- English **ACE-2004**, (Doddington et al., 2004).[6] We reuse the train/dev/test split used by most previous authors (Lu and Roth, 2015; Muis and Lu, 2017; Wang and Lu, 2018).
- English **ACE-2005**.[7] Again, we use the train/dev/test split by Lu and Roth (2015); Muis and Lu (2017); Wang and Lu (2018).
- Czech **CNEC 2.0** — Czech Named Entity Corpus 2.0 (Ševčíková et al., 2007). We use the official evaluation script distributed with the dataset, which evaluates 46 fine-grained entity types and 4 entity containers.

4 Methodology

All NameTag 3 models are fine-tuned pre-trained language models of either Large (355M) or Base (126M) size. For flat NER, we apply a classification softmax head on top of the language model, while for nested NER, we use a seq2seq decoding head instead (Straková et al., 2019). Both flat and nested NameTag 3 models support training on a collection of datasets, potentially in different languages. However, only NameTag 3 allows training on multiple tagsets with differing label sets.

4.1 Flat NER

For flat NER, NameTag 3 enables multitagset learning by assigning a separate classification head to each tagset and jointly training the encoder and all classification heads. During inference, the classification head corresponding to the requested tagset is used, ensuring that only valid tags are predicted, see visualization in Fig. 2.

The currently supported tagsets are:

- `conll`: The CoNLL-2002 and CoNLL-2003 (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003) tagset,
- `uner`: The Universal NER v1 (Mayhew et al., 2024) tagset,
- `onto`: The OntoNotes 5.0 tagset.

The NameTag 3 multilingual flat NER model was trained on the training portions of the flat NER datasets described in Sec. 3.1. Training batches were constructed using square root temperature

[4] https://github.com/lang-uk/ner-uk
[5] https://lang.org.ua/en/
[6] https://catalog.ldc.upenn.edu/LDC2005T09
[7] https://catalog.ldc.upenn.edu/LDC2006T06

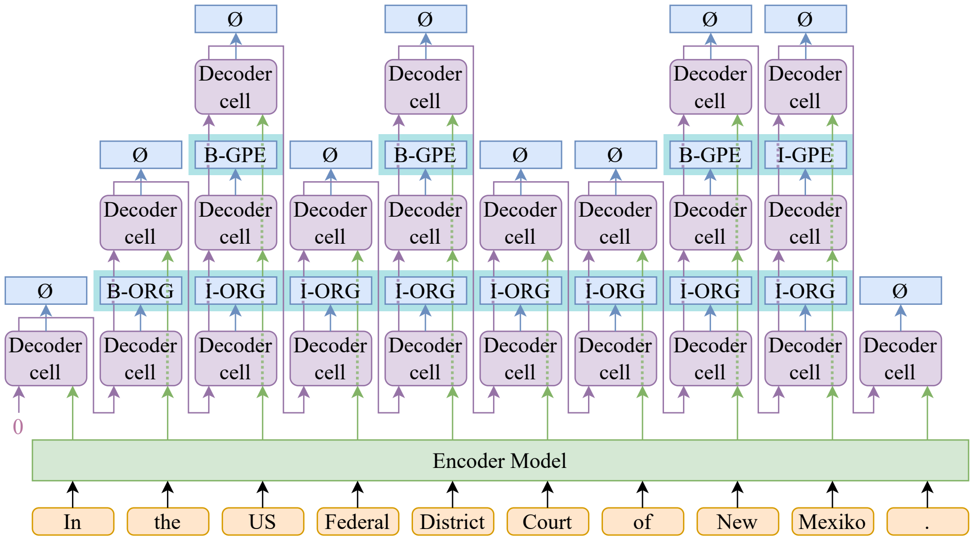

Figure 1: Visualization of the nested NER seq2seq decoder with hard attention on the current token. The example sentence is taken from ACE-2004 (Doddington et al., 2004).

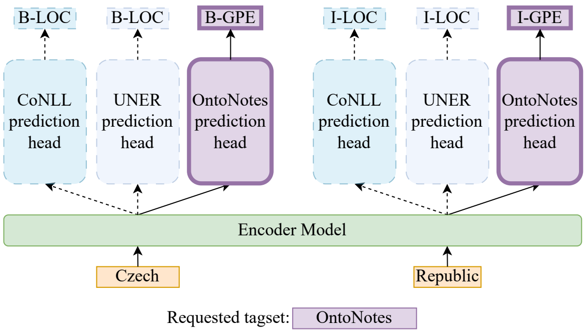

Figure 2: Visualization of the flat NER classification heads for multiple tagsets.

sampling, in which the examples from the corpora are sampled into training batches proportionally to the square root of the number of their sentences, similarly to van der Goot et al. (2021). This approach effectively downsamples the largest corpora while upsampling the smallest ones. To achieve balanced performance across all datasets, we use a macro span-based F1 score with uniform weighting as our evaluation objective. The training hyperparameters are described in Table 2.

4.2 Nested NER

For nested named entity recognition, we replace the flat softmax classification head with a sequence-to-sequence (seq2seq) decoder head (Straková et al., 2019), see visualization in Figure 1. This decoder generates a sequence of linearized (flattened) nested output labels for each input token embedded by the pre-trained LM encoder. The Transformer encoder and seq2seq decoder weights are fine-tuned jointly. Before fine-tuning, we perform a few pre-training epochs with frozen Transformer encoder weights to allow the seq2seq decoder to adjust to them. This helps ensure a smoother transition into fine-tuning. The training hyperparameters are described in Table 2.

5 Results

5.1 Flat NER

Table 3 presents NameTag 3 span-based micro F1 with the monolingual (Mono) models and the multilingual (Multi) model of 355M params.

Alongside our results, we report the highest F1 scores from the respective leaderboards on `https://paperswithcode.com/` where available, and/or the current state-of-the-art academic baselines; many of these models originate from academic research and do not provide ready-to-use tools, and/or often rely on significantly larger model capacities in terms of parameter count.

Apart from the state-of-the-art models, we also compare NameTag 3 to popular NLP toolkits supporting named entity recognition: Stanza (Qi et al., 2020) and SpaCy (Honnibal and Montani, 2017). Our system surpasses both these toolkits on all the datasets where pretrained models are available.[8]

Table 7 presents out-of-domain evaluation on unseen languages/datasets by cross-lingual transfer. The accompanying previous SOTA results are from Mayhew et al. (2024).

[8]Both Stanza and SpaCy provide models for more languages, but trained on different datasets with possibly different tag sets, preventing direct comparison on more languages.

Corpus	Mono F1	Multi F1	Stanza F1	SpaCy F1	SOTA F1	SOTA Ref.	SOTA Params
Arabic OntoNotes v5	75.50	74.20	—	—	**76.40**	Aloraini et al. (2020)	136M
Chinese OntoNotes v5	**81.76**	81.63	79.2♡	—	80.20	Li et al. (2023)	147M
Chinese UNER GSDSIMP	88.99	**90.99**	—	—	89.40	Mayhew et al. (2024)‡	355M
Chinese UNER GSD	90.14	**91.53**	—	—	89.50	Mayhew et al. (2024)‡	355M
Croatian UNER SET	94.08	**95.55**	—	—	95.00	Mayhew et al. (2024)‡	355M
Czech CNEC 2.0 CoNLL	85.31	**86.24**	—	—	—	—	—
Danish UNER DDT	87.21	**89.75**	—	—	88.10	Mayhew et al. (2024)‡	355M
Dutch CoNLL-2002	95.16	94.93	89.2♡	—	**95.70**	Wang et al. (2021)	1117M†
English OntoNotes v5	90.22	90.19	88.8♡	89.8◇	**92.07**	Li et al. (2020)	336M
English UNER EWT	86.27	**87.03**	—	—	85.80	Mayhew et al. (2024)‡	355M
English CoNLL-2003	93.80	94.09	92.1♡	91.6◇	**94.60**	Wang et al. (2021)	1853M†
German CoNLL-2003	87.77	87.48	81.9♡	—	**88.38**	Wang et al. (2021)	1108M†
Maghrebi UNER Arabizi	72.77	84.49	—	—	**86.20**	Mayhew et al. (2024)‡	355M
Norw. Bokmål UNER NDT	93.97	**95.83**	—	—	—	—	—
Norw. Nynorsk UNER NDT	93.71	**94.51**	—	—	—	—	—
Portuguese UNER Bosque	**91.18**	90.89	—	—	90.40	Mayhew et al. (2024)‡	355M
Serbian UNER SET	94.85	**97.10**	—	—	96.60	Mayhew et al. (2024)‡	355M
Slovak UNER SNK	86.79	**88.46**	—	—	85.50	Mayhew et al. (2024)‡	355M
Spanish CoNLL-2002	88.95	90.29	88.1♡	—	**90.40**	Wang et al. (2021)	1105M†
Swedish UNER Talbanken	90.73	**91.79**	—	—	88.30	Mayhew et al. (2024)‡	355M
Ukrainian Lang-uk	90.45	**92.88**	86.1♡	—	88.73	NameTag 2	110M

Table 3: NameTag 3 flat NER span-based micro F1 with the monolingual (Mono) models and the multilingual (Multi) model of 355M params. We report the highest F1 scores from the respective leaderboards on `https://paperswithcode.com/` where available. †Wang et al. (2021) use a concatenation of multiple embeddings, incl. several Base and Large. ‡For Mayhew et al. (2024), we report the better result from the "in-language" (Table 4) and "all" (Table 5). ♡ `https://stanfordnlp.github.io/stanza/ner_models.html`. ◇ `https://spacy.io/usage/facts-figures`.

Model	F1
ChatGPT 3.5 zero-shot (Xie et al., 2024)	68.97†
ChatGPT 3.5 ICL with self-annotated demonstrations (Xie et al., 2024)	74.99†
DeepSeek R1 32B zero-shot	64.33
DeepSeek R1 32B 5-shot	74.26
DeepSeek R1 70B zero-shot	67.97
DeepSeek R1 70B 5-shot	74.00
NameTag 3	**94.09**

Table 4: Comparison of NameTag 3 with NER performed by prompting LLMs on the (entire) English CoNLL-2003 test dataset (3 684 sentences). †Xie et al. (2024) report the mean of two samples of 300 sentences.

LLM Evaluation We include comparison of NameTag 3 with LLMs in Table 4 to demonstrate that fine-tuning "smaller" models (355M vs. 70B parameters) is still worthwhile even in the era of generative AI. We prompt DeepSeek-R1 70B (DeepSeek-AI et al., 2025), currently one of the best available open-source sub-100B LLMs,[9] in zero-shot and 5-shot settings, and we also reprint similar prompting experiments on ChatGPT 3.5 reported in literature (Xie et al., 2024).

[9] Our goal was to evaluate the best available replicable model that can run without enormous resources in order to be a viable NER system alternative.

Model	GPU	Batch	Sentences per sec.	Time
DeepSeek R1 70B zero-shot	AMD MI210	1	0.05	23h
DeepSeek R1 70B 5-shot	AMD MI210	1	0.04	25h
DeepSeek R1 32B zero-shot	AMD MI210	1	0.08	13h
DeepSeek R1 32B 5-shot	AMD MI210	1	0.06	16h
NameTag 3	AMD MI210	1	801	4.6s
NameTag 3	AMD MI210	8	784	4.7s
NameTag 3	NVIDIA A30	1	646	5.7s
NameTag 3	NVIDIA A30	8	801	4.6s

Table 5: Sentence throughput in sentences per second of the NameTag 3 REST API and Deep Seek REST API by predicting the (entire) English CoNLL-2003 test dataset (3 684 sentences).

Corpus	F1	SOTA F1	SOTA Ref.	SOTA Params.
ACE-2004	88.39	**88.72**	Shen et al. (2023)	345M
ACE-2005	87.21	**88.83**	Yuan et al. (2022)	223M
CNEC 2.0	**86.39**	83.44	NameTag 2	110M

Table 6: NameTag 3 nested NER span-based micro F1. CNEC 2.0 is the only corpus modeled with a Base-sized monolingual Czech encoder RobeCzech Base (126M). The ACE models are based on RoBERTa Large (355M).

Corpus	F1	SOTA F1
Cebuano UNER GJA	**96.97**	82.2
Chinese UNER PUD	**89.35**	86.0
Portuguese UNER PUD	**91.77**	87.5
Russian UNER PUD	**75.51**	73.6
Swedish UNER PUD	**91.27**	88.0
Tagalog UNER TRG	**97.78**	83.7
Tagalog UNER Ugnayan	75.00	**76.1**

Table 7: Cross-lingual/out-of-domain evaluation on unseen languages/datasets predicted by cross-lingual transfer with the NameTag 3 multilingual flat model of 355M parameters. The metric is flat NER span-based micro F1. Previous SOTA F1 are from Mayhew et al. (2024), whose multilingual model is also of 355M.

NameTag 3, a fine-tuned 355M model, achieves 20 percent points higher F1 score while being more than 10,000 times faster, as demonstrated in performance measurements Tab 5. Therefore, when training data are available, NameTag 3 constitutes a much more accessible and practical system, allowing users to keep processed data private using only a single consumer-grade GPU. The complete script for LLM evaluation including the used prompts and few-shot example selection is available at `https://github.com/ufal/nametag3/tree/acl2025/llm_baseline`.

5.2 Nested NER

Table 6 shows the NameTag 3 nested NER results, evaluated as span-based micro F1. NameTag 3 with the seq2seq head for nested NER achieves state-of-the-art results on the canonical Czech nested corpus with 46 entity types and 4 containers, while reaching near-SOTA results for English nested corpora.

6 Conclusions

We introduced NameTag 3, a multilingual, open-source named entity recognition tool for both flat and nested NER. It is available as a command-line tool (`https://github.com/ufal/nametag3`) and as a web application with a cloud-based REST API (`https://lindat.mff.cuni.cz/services/nametag`). NameTag 3 includes pre-trained models and supports custom training.

NameTag 3 demonstrates state-of-the-art performance on 21 test datasets across 15 languages: Cebuano, Chinese, Croatian, Czech, Danish, English, Norwegian Bokmål, Norwegian Nynorsk, Portuguese, Russian, Serbian, Slovak, Swedish, Tagalog, and Ukrainian, while also performing well in Arabic, Dutch, German, Maghrebi, and Spanish.

The tool is released under the open-source MPL

2.0 license, with models distributed under non-commercial CC BY-NC-SA 4.0.

We hope NameTag 3 will be particularly valuable for the academic community and researchers working with multilingual NLP and non-English texts.

Limitations

Since NameTag 3 classifies into a predefined set of named entity classes, it is not susceptible to issues generally associated with generative AI, such as hallucinations or the production of misleading or harmful information.

By jointly training on 21 datasets across 17 languages, NameTag 3 is less prone to biases that typically affect monolingual or culturally homogeneous models. We hope that this multilingual approach helps mitigate issues like overrepresentation of Western-centric names and gender imbalances in named entity distributions.

However, most of our training datasets are written in Latin scripts, with the exception of Chinese (three datasets), Arabic (two datasets), and Ukrainian (one dataset). We recognize the need to further improve coverage by incorporating additional languages.

This brings us to an important limitation: As a supervised, fine-tuned model, NameTag 3 relies on gold-standard, manually annotated training data. Expanding the diversity and volume of such data is crucial for further improving performance across languages and domains.

In future work, we plan to expand our set of manually annotated training data while also exploring silver-standard, semi-automated data to further increase the volume of training material.

Acknowledgments

This research was supported by the Johannes Amos Comenius Programme (P JAC) project No. CZ.02.01.01/00/22_008/0004605, Natural and anthropogenic georisks, and it has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ. The work described herein uses resources hosted by the LINDAT/CLARIAH-CZ Research Infrastructure (projects LM2018101 and LM2023062, supported by the Ministry of Education, Youth and Sports of the Czech Republic).

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Abdulrahman Aloraini, Juntao Yu, and Massimo Poesio. 2020. Neural coreference resolution for Arabic. In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 99–110, Barcelona, Spain (online). Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) program-tasks, data, and evaluation. *Proceedings of LREC*, 2.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *Preprint*, arXiv:1508.01991.

Hongjun Li, Mingzhe Cheng, Zelin Yang, Liqun Yang, and Yansong Chua. 2023. Named entity recognition for chinese based on global pointer and adversarial training. *Scientific Reports*, 13(1):3242.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

Wei Lu and Dan Roth. 2015. Joint mention extraction and classification with mention hypergraphs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 857–867. Association for Computational Linguistics.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4322–4337, Mexico City, Mexico. Association for Computational Linguistics.

Aldrian Obaja Muis and Wei Lu. 2017. Labeling Gaps Between Words: Recognizing Overlapping Mentions with Mention Separators. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2608–2618. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Named Entities in Czech: Annotating Data and Developing NE Tagger. In *Lecture Notes in Artificial Intelligence, Proceedings of the 10th International Conference on Text, Speech and Dialogue*, volume 4629 of *Lecture Notes in Computer Science*, pages 188–195, Berlin / Heidelberg. Springer.

Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. PromptNER: Prompt locating and typing for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12492–12507, Toronto, Canada. Association for Computational Linguistics.

Jana Straková, Milan Straka, and Jan Hajič. 2014. Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. 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.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-independent Named Entity Recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. 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 CoNLL-2003*, pages 142–147. Edmonton, Canada.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Bailin Wang and Wei Lu. 2018. Neural segmental hypergraphs for overlapping mention recognition. In *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing, pages 204–214. Association for Computational Linguistics.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Automated concatenation of embeddings for structured prediction. 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 2643–2660, Online. Association for Computational Linguistics.

Tingyu Xie, Qi Li, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2024. Self-improving for zero-shot named entity recognition with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 583–593, Mexico City, Mexico. Association for Computational Linguistics.

Zheng Yuan, Chuanqi Tan, Songfang Huang, and Fei Huang. 2022. Fusing heterogeneous factors with triaffine mechanism for nested named entity recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3174–3186, Dublin, Ireland. Association for Computational Linguistics.