Token-Free Cross-Lingual Named Entity Recognition for Classical Chinese

Zhongqing JIANG†, Zengqing WU†, and Chuan XIAO†,††

† Graduate School of Information Science and Technology, Osaka University
1–5 Yamadaoka, Suita, Osaka, 565–0871 Japan
†† Graduate School of Informatics, Nagoya University
Furo–cho, Chikusa–ku, Nagoya, 464–0814 Japan
E-mail: †{jiang.zhongqing,chuanx}@ist.osaka-u.ac.jp, ††wuzengqing@hus.osaka-u.ac.jp

Abstract With the support of cross-lingual transfer learning techniques, the performance of named entity recognition (NER) for low-resource languages has recently received considerable improvement. However, the performance of existing methods still falls short on token-free low-resource target languages, such as Classical Chinese. In this paper, we target NER for Classical Chinese, and propose a method for token-free cross-lingual NER zero-shot transfer based on the ByT5 model. We build a Modern Chinese-Classical Chinese entity alignment model with the ByT5-based Modern Chinese NER classifier to obtain the predicted labels of Classical Chinese corpus. Following, we train a pre-trained language model for Classical Chinese and fine-tune it using supervised learning methods in combination with the prediction labels obtained. We evaluate our approach on a widely used Classical Chinese NER dataset C-CLUE and obtain better results than various baseline methods in the token-free condition and reduce the computational requirements for training the model.

Key words Named Entity Recognition, Natural Language Processing, Transfer Learning, Classical Chinese

1 Introduction

Natural language processing (NLP) for low resource languages where labeled data is scarce has always been a challenge in the field, and the advent of pre-trained language models has improved the situation for solving this problem. The latest large-scale pre-trained language models (LLM), such as GPT-3 [1], BERT [6], T5 [18] and other self-supervised models that are pre-trained and fine-tuned on large-scale corpora have substantially improved the performance of NLP tasks in low-resource languages since they do not rely on labeled data for training [15]. Building on this, the development of cross-lingual transfer learning techniques has led to significant performance improvements for many NLP tasks, such as named entity recognition (NER), especially for those languages with low resources [8]. A number of transfer learning methods provide some pseudo-labels for low-resource target languages by using pre-trained cross-lingual language models of the source language, thus allowing the target language model to learn in a zero-shot configuration [5], [8], [20]. This enables the training of models for low-resource languages no longer relies on expert annotations and large-scale labeled datasets. In terms of specific applications, taking the Classical Chinese language, a low resource language, studied in this paper as an example, the study [12] points out that there are a large number of unlabeled Classical Chinese texts that have not been included as part of the language model training, and with the help of LLM, these corpus sets are able to be used.

However, despite the fact that the aforementioned methods have brought a considerable improvement in the performance of NLP tasks, restricted by the limited corpus of the target language and the feature differences between the target language and the source language, there is still much space for improvement in the performance of existing methods on unlabeled, low-resource target languages. Existing studies argue that the fact of large pre-trained language models trained on general-domain texts makes them difficult to transfer to some domain-specific texts [15], [22]. General-domain texts, such as English and Modern Chinese, differ significantly from the linguistic features of Classical Chinese [12], [22]. On the other hand, the small size of the extant corpus of Classical Chinese, compared to other languages, limits the performance of LLM self-supervised learning.

Further, aside from the problem of limited dataset size, Classical Chinese has additional word separation requirements compared to most modern languages. In contrast to European languages such as English and German, Chinese and Japanese do not use whitespace as a boundary between words, and Classical Chinese goes further by not having markers as separators between sentences, which makes word separation more challenging [9]. The Classical Chinese corpus with expert annotations for punctuation is very limited, and expert annotations are subject to certain errors. Existing studies have shown that the accuracy of Classical Chinese word separation affects performance on other downstream tasks like NER [2]. The performance of existing attempts to pre-train and transfer learning directly on Classical Chinese without word separation and markers is limited.

Therefore, in this paper, we focus on exploring how to better implement the NER task for token-free Classical Chinese. We argue that the transfer learning approach via Modern Chinese-Classical Chinese entity alignment on the unlabeled corpus can perform better than the traditional automated word separation combined with supervised learning methods and direct training using LLM in a zero-shot setting.

We propose a token-free cross-lingual zero-shot transfer learning model based on the ByT5 model [23] to process the Classical Chinese NER task. The method is based on a Modern Chinese NER classifier based on ByT5, a pre-trained token-free byte-level language model that directly processes the raw text, combined with the entity alignment method to obtain prediction labels for corpus of Classical Chinese. This is followed by training a pre-trained language model for Classical Chinese and fine-tuning it using supervised learning methods with the obtained prediction labels.

We evaluate our model on the open-source Classical Chinese text dataset C-CLUE [10] and compare it with existing baseline methods. It turns out that our model achieves the best performance in the Classical Chinese NER task. We also conduct ablation experiments to analyze how the components of our model affect its performance.

2 Related Works

2. 1 Chinese NER Task

A lot of existing work has been done on the Chinese NER task. Early work trained language models based on word separation, e.g., LSTM-CRF [16], Latice-LSTM [25], etc. Subsequently, some LLM-based work and the development of transfer learning techniques have driven the progress of Chinese NER tasks. For example, the BERT-wwm model [4], which uses BERT as the framework and incorporates certain whole word masking strategies, and the RoBERTa model [14], which adjusts the selection and experimental configuration of its hyperparameters based on BERT, are widely used as baselines for Classical Chinese NER problems. Also based on BERT, some studies that trained on large scale Classical Chinese corpus, such as AnchiBERT and SikuBERT achieved better results compared to traditional models [7], [22]. Some other models, such as deep neural networks, have also been applied in Classical Chinese NER problems [24]. However, as mentioned earlier, still limited by the scale of the original corpus of Classical Chinese, the performance of LLM on Classical Chinese is still deficient compared to its performance on other modern languages.

Several transfer learning and contrast learning methods have been used to train Classical Chinese language models. One research uses a contrastive learning approach to enhance machine reading comprehension of both languages by contrastive learning of Classical Chinese and Modern Chinese [13]. The AT-CCNER model uses an adversarial transfer approach to transfer the word segmentation features learned by the language model from large-scale translation to Classical Chinese, obtaining better performance than existing approaches [17]. It is evident from these studies that it is difficult to apply LLM directly to the training of a low-resource language like Classical Chinese, and that with some combination of learning from a broader corpus, such as transfer from Modern Chinese to Classical Chinese, can usefully enhance the performance of the experiments.

2. 2 Token-Free Language Models

The shortcomings of current research on token-free languages have been noted in some existing studies. For example, the ByT5 model, a byte-level sequence-to-sequence model, is based on the T5 architecture to design a Transformer architecture that processes raw text (i.e., bytes) directly, rather than language processing on sequences of words or subwords, on language-independent, token-independent NLP tasks [23]. This approach gives a significant performance boost for tasks in token-free languages, and is the reason why this study employs ByT5 as the basis of our language model. Classical Chinese, as a language that does not delimit words and sentences with punctuation and whitespace, makes the application of models with tokenizer on it difficult, because these steps of dividing the raw text into tokens are usually based on whitespace and markers. In addition, since all e-texts in Classical Chinese are entered manually, there is a certain amount of noise and mis-entries in them, which can negatively affect the results of both the word segmentation and the non-token-free methods. Comparatively, ByT5 excels in processing noisy text data due to features such as direct processing of raw text [19]. For these reasons, this paper argues that the ByT5 model is a good match for the linguistic properties of Classical Chinese, it avoids the interference of errors in intermediate steps such as word separation on the final results, and weakens the disadvantages of the absence of markers that make the language

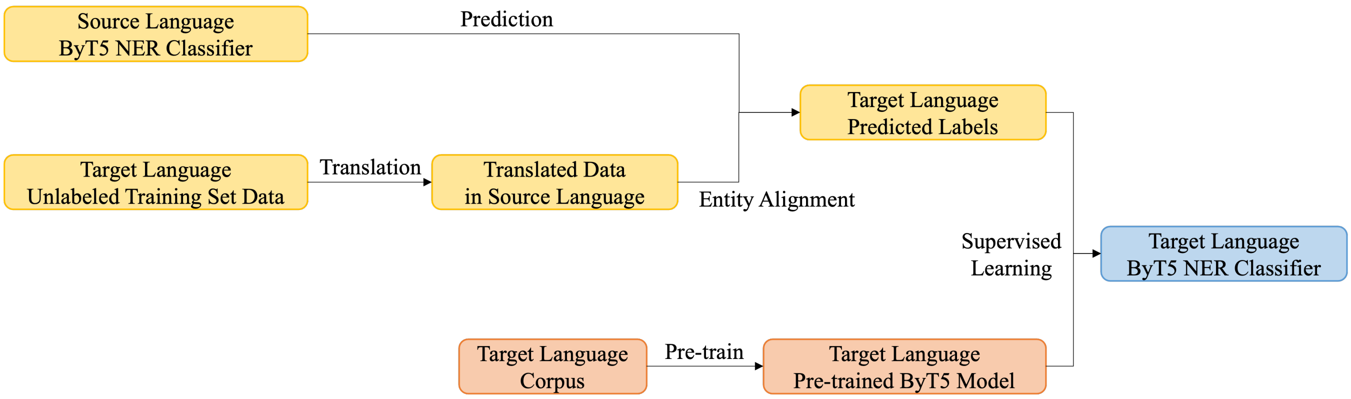

Figure 1 Framework

model difficult to apply directly.

Nevertheless, the reason we did not directly train the Classical Chinese language model using ByT5 is that the model only shows more prominent performance with limited model size and its performance varies on languages with different token compression rates [23]. Relatively, we use the Modern Chinese NER classifier parameters of ByT5 as a source of transfer learning to achieve better performance by fine-tuning. To the best of our knowledge, this paper is the first to use a token-free language model to deal with the classical Chinese NER problem.

2.3 Cross-Lingual Pre-training Models

The most widely used cross-lingual pre-training model in existing studies is the multilingual version of BERT, which has achieved leading results in various NLP tasks. Whereas for low-resource languages, some studies have trained cross-lingual models based on BERT on large amounts of data, such as XLM-RoBERTa, showing the conclusion that fine-tuning for low-resource languages can further improve model performance [3]. In addition, some existing researches enhance the effect of fine-tuning on NER task by entity alignment from source to target language in pre-training phase [11], [26]. Therefore, this work try to improve the effectiveness of ByT5 on Classical Chinese by using transfer learning, entity alignment and fine-tuning from Modern Chinese parameters of ByT5 model to Classical Chinese.

3 Methodology

3.1 Model Architecture

Figure 1 illustrates the framework of our model. The model is divided into three main components, which are (1) using ByT5 [23]'s token-free NER classifier for the source language (Modern Chinese) to label the training set corpus of the target language (Classical Chinese) with prediction labels after building a Modern Chinese - Classical Chinese entity alignment during the translation procedure; (2) using ByT5's architecture to train a preliminary pre-trained language model for the target language; and finally (3) fine-tune the pre-trained language model for the target language using supervised learning methods in combination with the prediction labels obtained. Eventually, the model will result in a zero-shot token-free NER classifier for the target language.

3.2 Entity Alignment

The target language corpus is limited by the presence of relatively frequent missing or difficult-to-read parts, and the existence of a range of heterogeneous or rare Chinese characters, which results in a much larger character set size compared to other languages and impacts the effectiveness of a range of NLP downstream tasks, including NER. Hence, we use entity alignment to address this issue. To achieve entity alignment, we first translate the unlabelled corpus for the target language, align the translated source language with the target language for entity alignment, and label it using the NER classifier with ByT5 for the source language. The entity alignment is performed using the exact match method.

3.3 Pre-training Based on ByT5

We adapt the ByT5 architecture to pre-train the corpus of the target language. Since T5 architecture is capable of implementing major NLP downstream tasks, the trained ByT5 model is able to perform NER task after a simple fine-tuning. The corpus used for training is token-free.

3.4 Cross-Lingual Fine-tuning

We obtain prediction labels for the target language training in the zero-shot configuration by the entity alignment method and ByT5 NER classifier, and we apply these labels to the initially trained target language ByT5 model for supervised learning to obtain a zero-shot token-free NER classifier for the target language.

4 Experiments

4.1 Dataset

We evaluate our model using the Classical Chinese NER dataset C-CLUE [10]. The dataset includes six named entity categories including person (PER), location (LOC), organization (ORG), etc. Separately, limited by the size of the C-CLUE dataset, we use Daizhige [1], a Classical Chinese text corpus including 15 thousand and more texts from various Classical Chinese literature, as the source of training set for the model. We preprocessed this training set, including removing invalid tokens from non-Classical Chinese texts and non-UTF-8 characters that cannot be used as model inputs. Then, to match the use of the ByT5 model, we convert the raw text characters from Traditional Chinese to Simplified Chinese.

4.2 Pretraining Setup

Considering the size of the dataset, our model uses the same configuration as ByT5-small [23]. The model uses a transformer encoding-decoding model with 12 encoding layers and 4 hidden layers, and 300M parameters. We pre-train the ByT5 model for 1 million steps on token-free text with a maximum sequence length of 1024 bytes and a batch size of 2^{20} tokens. The experiments are run on Google Colab platform with Google Cloud TPU v2-8. In all experiments, the same network layers are set with the same parameters and perform in the same experimental environment.

For NER task, we fine-tune ByT5 for 1000 steps with the settings of a constant learning rate of 0.001 and a dropout rate of 0.1.

4.3 Evaluation Settings

The NER task uses Precision (P), Recall (R), and F1-score as metrics. As the indicator that reflects the model performance in a balanced way, F1 will be used as the primary indicator to measure the results of our experiments.

Table 1 Experimental Results on C-CLUE Dataset

Method	P	R	F1
BERT-Base	29.82	35.59	32.12
BERT-wwm	32.98	43.82	35.40
RoBERTa-zh	28.28	34.93	31.09
BiLSTM-CRF	49.76	54.77	52.15
Lattice-LSTM	51.96	56.92	54.33
Our Model	**52.21**	**56.94**	**54.47**

4.4 Baselines

We compare the performance of our model with the models commonly used in Classical Chinese NER task, including BERT [6], BERT-wwm [4], RoBERTa-zh [14], Lattice-LSTM [25], etc.

5 Results

Table 1 shows the experimental results of different models for the NER task on the C-CLUE dataset.

The results show that our token-free model obtains the best experimental performance compared to the above models. This reflects that our model can better handle the characteristics of Classical Chinese which is difficult to be segmented and the raw text has no punctuation marks. The main reason for this result is that the token-free model matches better with the Classical Chinese language in terms of word segmentation, amount of tokens, grammatical patterns, and other factors.

First, considering that Classical Chinese is featured by the absence of punctuation, traditional methods, such as LSTM-CRF, need to perform word segmentation first before implementing the NER task. Existing studies show that the quality of word segmentation directly affects the performance of the NER task. Compared with languages such as Modern Chinese and Classical Japanese, word segmentation in Classical Chinese is more difficult. The inevitable errors in word segmentation eventually degrade the performance of the NER task. Therefore, this paper uses the token-free approach to avoid the problem of word segmentation. This paper further argues that the language model based on token-free architecture is also applicable to other downstream tasks that are affected by word segmentation.

On the other hand, this paper argues that the advantage of applying a token-free model on Classical Chinese is not only the avoidance of the word segmentation process. Considering that there are many variants of Classical Chinese vocabulary and that the text has a lot of noise (e.g., human input errors, unreadable parts of the original text, etc.), the dataset carries noise, and the total number of tokens is large, all these factors decreases the performance of the pre-trained model. In contrast, the token-free model is thought to reduce the interference of these problems because it does not deal with characters/words, but directly with bytes. Further empirical studies are still needed to explore whether token-free models may also have better results in downstream tasks in other languages without punctuation.

Finally the transfer learning approach is able to deal with the limited corpus of Classical Chinese as a low-resource language. Since the corpus set of Classical Chinese is considerably small in size even if it is unlabeled. This paper shows

[1]：Available at: https://github.com/garychowcmu/daizhigev20

that the transfer learning and entity alignment methods can effectively improve the NER task performance of Classical Chinese.

6 Ablation Study

To analyze the effect of different factors on the validity of our model, we conducted ablation experiments on the model, as shown in Table 2.

Table 2 Results of Ablation Experiments

Method	P	R	F1
Our method [ByT5+Transfer]	52.21	56.94	54.47
-Cross-lingual Fine-tuning [ByT5]	49.28	55.18	52.06

Effectiveness of transfer learning: To test the effectiveness of transfer learning, we perform experiments after removing the part of transfer learning from the model. The results show that our model with transfer learning included has better performance compared to the ByT5 model trained on the Classical Chinese corpus only.

7 Conclusion

In this paper, we propose a token-free cross-lingual zero-shot transfer learning model for the Classical Chinese NER task. Through a series of experiments, we show that our model can effectively improve the performance of the NER task. We find that the token-free model, such as ByT5, can handle the linguistic features of Classical Chinese well. In addition, the transfer learning approach can effectively address the problem of limited size of the target language dataset, and the performance of the NLP task in the target language can be improved by transfer learning from the source language to the target language.

In future research, we will compare the models with more baselines. Further, we will perform more experiments to validate the model, including validation of entity alignment methods, comparison of using architectures (e.g., replacing ByT5 with BERT, mT5, other token-free models like Charformer [21], etc.). In addition, the generalization ability about the findings of this paper in other languages as well as in other NLP downstream tasks can be further discussed.

8 Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22H03903. We thank Dr. Ruoming Pang, Mr. Run Peng, Prof. Keishi Tajima, and Prof. Hiroki Nomiya for their instructive suggestions.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. Adversarial transfer learning for chinese named entity recognition with self-attention mechanism. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 182–192, 2018.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[4] Y Cui, W Che, T Liu, B Qin, Z Yang, S Wang, and G Hu. Pre-training with whole word masking for chinese bert. arxiv preprint arxiv: 190608101. 2019.

[5] Wietse de Vries, Martijn Wieling, and Malvina Nissim. Make the best of cross-lingual transfer: Evidence from pos tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, 2022.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Wang Dongbo, Liu Chang, Zhu Zihe, Liu Jiangfeng, Hu Haotian, Shen Si, and Bin Li. Sikubert and sikuroberta: Research on the construction and application of pre-training model of sikuquanshu for digital humanities. *Library Tribune*, pages 1–14, 2021.

[8] Julian Martin Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kardas, Sylvain Gugger, and Jeremy Howard. Multifit: Efficient multi-lingual language model fine-tuning. *arXiv preprint arXiv:1909.04761*, 2019.

[9] Shohei Higashiyama, Masao Utiyama, Eiichiro Sumita, Masao Ideuchi, Yoshiaki Oida, Yohei Sakamoto, Isaac Okada, and Yuji Matsumoto. Character-to-word attention for word segmentation. *Journal of Natural Language Processing*, 27(3):499–530, 2020.

[10] Zijing Ji, Yuxin Shen, Yining Sun, Tian Yu, and Xin Wang. C-clue: A benchmark of classical chinese based on a crowdsourcing system for knowledge graph construction. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction: 6th China Conference, CCKS 2021, Guangzhou, China, November 4-7, 2021, Proceedings 6*, pages 295–301. Springer, 2021.

[11] Bing Li, Yujie He, and Wenjin Xu. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *arXiv preprint arXiv:2101.11112*, 2021.

[12] Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. Ancient–modern chinese translation with a new large training dataset. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(1):1–13, 2019.

[13] Maofu Liu, Junyi Xiang, Xu Xia, and Huijun Hu. Contrastive learning between classical and modern chinese for classical chinese machine reading comprehension. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–22, 2022.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[15] Qiuhao Lu, Dejing Dou, and Thien Nguyen. Clinicalt5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP*

2022, pages 5436–5443, 2022.

[16] Nanyun Peng and Mark Dredze. Improving named entity recognition for chinese social media with word segmentation representation learning. *arXiv preprint arXiv:1603.00786*, 2016.

[17] Yongjie Qi, Hongchao Ma, Lulu Shi, Hongying Zan, and Qinglei Zhou. Adversarial transfer for classical chinese ner with translation word segmentation. In *Natural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I*, pages 298–310. Springer, 2022.

[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[19] David Samuel and Milan Straka. \'ufal at multilexnorm 2021: Improving multilingual lexical normalization by fine-tuning byt5. *arXiv preprint arXiv:2110.15248*, 2021.

[20] Henning Schäfer, Ahmad Idrissi-Yaghir, Peter Horn, and Christoph Friedrich. Cross-language transfer of high-quality annotations: Combining neural machine translation with cross-linguistic span alignment to apply ner to clinical texts in a low-resource language. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 53–62, 2022.

[21] Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*, 2021.

[22] Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. Anchibert: A pre-trained model for ancient chinese language understanding and generation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[23] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

[24] Chengxi Yan, Qi Su, and Jun Wang. Mogcn: Mixture of gated convolutional neural network for named entity recognition of chinese historical texts. *IEEE Access*, 8:181629–181639, 2020.

[25] Yue Zhang and Jie Yang. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564, 2018.

[26] Ran Zhou, Xin Li, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. Conner: Consistency training for cross-lingual named entity recognition. *arXiv preprint arXiv:2211.09394*, 2022.