Journal of Electronic Research and Application, 2024, Volume 8, Issue 5
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

A Novel Optimization Scheme for Named Entity Recognition with Pre-trained Language Models

Shuanglong Li[1], Xulong Zhang[2]*, Jianzong Wang[2]

[1]Shenzhen University, Shenzhen 518000, Guangdong Province, China
[2]Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518000, Guangdong Province, China

***Corresponding author:** Xulong Zhang, zhangxulong@ieee.org

Abstract: Named Entity Recognition (NER) is crucial for extracting structured information from text. While traditional methods rely on rules, Conditional Random Fields (CRFs), or deep learning, the advent of large-scale Pre-trained Language Models (PLMs) offers new possibilities. PLMs excel at contextual learning, potentially simplifying many natural language processing tasks. However, their application to NER remains underexplored. This paper investigates leveraging the GPT-3 PLM for NER without fine-tuning. We propose a novel scheme that utilizes carefully crafted templates and context examples selected based on semantic similarity. Our experimental results demonstrate the feasibility of this approach, suggesting a promising direction for harnessing PLMs in NER.

Keywords: GPT-3; Named Entity Recognition; Sentence-BERT model; In-context example

Online publication: September 30, 2024

1. Introduction

Named Entity Recognition (NER), also referred to as "moniker recognition," is the task of identifying and classifying entities within text that hold specific meanings, such as names of people, locations, organizations, and other proper nouns. For instance, in the sentence "Guangxi and other places have more cloudy and rainy days," "Guangxi" is a named entity that would be recognized as a location. The term NER was first introduced in the MUC-6 conference, and it has since garnered significant attention from both academia and industry due to its wide-ranging applications [1–3]. In search engines, accurate NER can enhance the precision of information retrieval, while in question-answering systems, it enables more accurate responses.

Named Entity Recognition (NER), the task of identifying and classifying key information units like names, organizations, and locations within text, has long been a cornerstone of natural language processing. Early NER systems relied heavily on handcrafted rules and dictionaries, often requiring extensive domain expertise and laborious maintenance to achieve reasonable performance. The advent of machine learning techniques offered a more adaptable solution, allowing models to learn patterns from annotated data and generalize to unseen examples.

However, it was the rise of deep learning that truly revolutionized the field. Neural network architectures, particularly Recurrent Neural Networks (RNNs) coupled with Conditional Random Fields (CRFs), demonstrated the ability to capture complex contextual dependencies in text, leading to significant improvements in NER accuracy [4]. Subsequent advancements, such as the incorporation of Convolutional Neural Networks (CNNs) for local feature extraction and Long Short-Term Memory (LSTM) units for improved handling of long-range dependencies, further pushed the boundaries of what was possible [5,6]. The integration of attention mechanisms, enabling models to focus on the most relevant parts of the input sequence, proved especially impactful, allowing for even finer-grained entity recognition and classification [7].

The Generative Pre-trained Transformer (GPT) is a powerful natural language processing model that leverages a self-attention mechanism to capture relationships within input sequences [8]. Its extensive pre-training enables rich language representation and contextual understanding, facilitating in-context learning without parameter fine-tuning [9]. This capability allows GPT to perform diverse natural language processing tasks by utilizing prompts and context, drawing upon its pre-trained linguistic knowledge. Motivated by this, we investigate the application of GPT to Named Entity Recognition (NER) in this paper. We propose a novel approach that employs context examples selected based on a principle of utterance distance, and we empirically validate the feasibility of this method. In essence, this paper's key findings can be encapsulated as follows:

(1) We investigate the feasibility of utilizing the pre-trained large language model GPT-3 for NER tasks, particularly focusing on scenarios where fine-tuning is not required.

(2) We propose a novel NER scheme that leverages carefully designed templates and the selection of context examples based on semantic similarity, thereby eliminating the need for model fine-tuning.

(3) We perform empirical evaluations to verify the effectiveness of our suggested approach, highlighting its promise for practical NER deployments.

2. Methodology

2.1. Utilizing GPT-3

GPT-3, a pre-trained large language model developed by OpenAI, is built upon the Transformer architecture, incorporating multi-layer self-attention mechanisms and feed-forward neural networks. Through extensive pre-training on a diverse corpus, GPT-3 acquires a generalized language representation, enabling it to adapt to various natural language processing tasks. Notably, GPT-3 exhibits a degree of zero-shot learning capability, performing new tasks without task-specific training data through in-context learning. The availability of OpenAI's Application Programming Interface (API) for GPT-3 further facilitates its use in downstream applications, making it the chosen base model for this research.

2.2. Sentence-BERT model

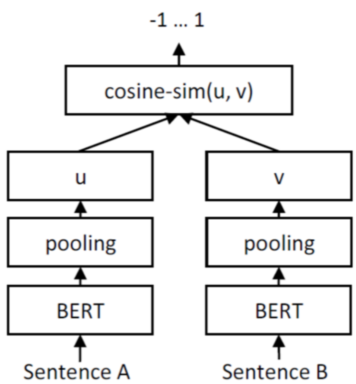

Figure 1. Model structure of Sentence-BERT

The Bidirectional Encoder Representations from Transformers (BERT) model, has proven to be highly effective for tasks that involve assessing the semantic similarity between pairs of sentences, as exemplified by its exceptional performance on the Semantic Textual Similarity (STS) benchmark. However, BERT's architecture necessitates the simultaneous processing of sentence pairs, incurring substantial computational costs. As an example, the computationally intensive process of finding the most semantically similar pair among 10,000 sentences requires roughly 50 million inference operations, translating to about 65 hours of processing time on a V100 GPU. This inherent computational burden renders BERT impractical for large-scale semantic similarity searches.

Sentence-BERT addresses BERT's computational limitations by adopting a Siamese network architecture with shared BERT parameters. This facilitates the generation of semantically rich sentence embeddings via contrastive learning. The proximity of embeddings for semantically similar sentences enables efficient similarity search using metrics like cosine similarity. By drastically reducing the computational time compared to BERT (e.g., from 65 hours to 5 seconds for finding the most similar pair among 10,000 sentences), Sentence-BERT is well-suited for tasks such as clustering and semantic-based information retrieval. Consequently, we leverage pre-trained Sentence-BERT in this study to derive sentence embeddings.

2.3. The proposed NER method

In this study, we adopt a strategy of selecting in-context examples based on their semantic similarity to the target sentence. The semantic similarity between sentences is quantified using the cosine similarity of their Sentence-BERT embeddings. We select the "N" most semantically similar sentences from the training set as in-context examples, which are then incorporated into the template and provided as prompts to GPT-3.

It's important to note that selecting the "N" most similar sentences may inadvertently exclude entity types present in the target sentence but absent in the chosen examples. To mitigate this, we remove one sentence from the candidate set and include a sentence that encompasses all entity types. This straightforward approach is feasible in our scenario due to the limited number of entity types (four) and the abundance of sentences containing all types in our dataset. However, more sophisticated strategies may be necessary for complex datasets with various entity types.

Furthermore, the choice of template significantly impacts the performance of our proposed scheme. The subsequent experimental section explores various template designs. Additionally, we conduct experiments to demonstrate the efficacy of selecting examples based on semantic similarity and to analyze the influence of the number of examples on the final results.

3. Experiments and results

3.1. Dataset

The CoNLL-2003 named entity dataset, derived from the Reuters Corpus and comprising news stories from August 1996 to August 1997, was utilized for the experiments. We focused solely on the English portion of the dataset, excluding the development set and unlabeled data. The dataset employs the BIO tagging scheme, marking the beginning (B) and inside (I) of entities, with "O" denoting non-entity tokens. The entity types within this dataset include four categories: person names (PER), locations (LOC), organizations (ORG), and miscellaneous entities (MISC).

The experiments utilized the OpenAI API for GPT-3, specifically the "davinci-instruct-beta-v3" engine with temperature set to 0 and top-p set to 0.9. A pre-trained Sentence-BERT model ("all-mpnet-base-v2") was employed without fine-tuning, generating 768-dimensional embeddings. Cosine similarity between these embeddings served as the semantic similarity measure between sentences, with higher values indicating greater similarity.

3.2. GPT-3 template selection

The selection of an appropriate template is crucial for the effective utilization of GPT-3 in the NER task. The initial and most intuitive approach is to directly employ the standard Begin, Inside, Outside (BIO) annotation method, as depicted in **Figure 2**. This method aims to directly prompt GPT-3 to produce the corresponding NER labels. The BIO scheme is a widely adopted labeling convention in entity extraction tasks. The "B" tag signifies the beginning of a named entity, with the first word of the entity being labeled as "B-" followed by the entity category. The "I" tag denotes words within the named entity, while "O" represents tokens that are not part of any named entity. To illustrate, in the sentence "Adam Bentley, who works at Microsoft, lives in Seattle," the BIO labels would be "B-PER I-PER O O O O O B-ORG O O O B-LOC." Our first template directly implements this standard annotation method, where the first line of the template contains the sentence to be processed, and the second line presents its corresponding BIO labels.

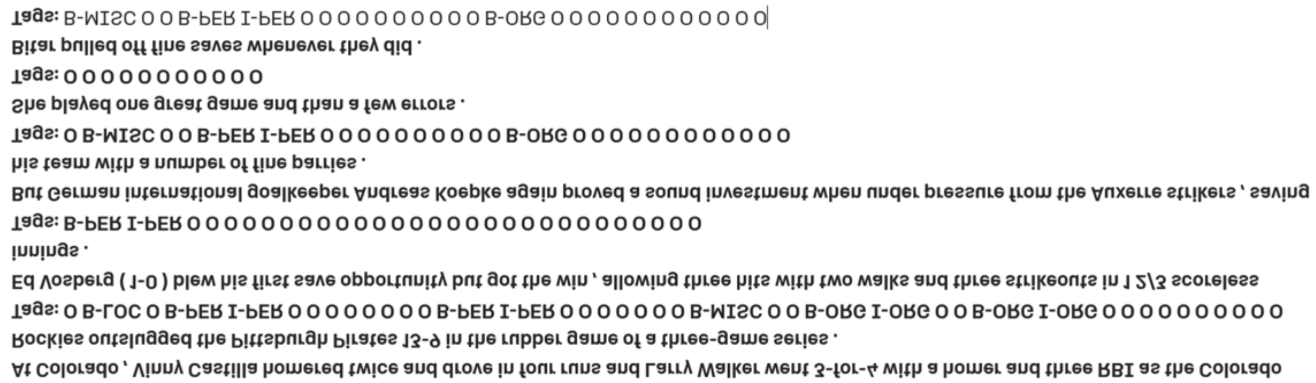

Figure 2. The "BIO-tag" template

While the BIO labeling method is a standard in entity extraction tasks, its direct application as a template for GPT-3 presents challenges. The abstract nature of the BIO labels makes it difficult for the language model to discern the relationship between these labels and the corresponding utterances, especially when provided with a limited number of in-context examples. This can lead to ambiguous or even erroneous outputs, including instances where the number of predicted labels misaligns with the number of tokens in the input sentence, as observed in **Figure 2**. The use of abbreviations in the labels (e.g., B-PER for Begin-Person) further complicates the matter, requiring the model to infer their meaning from the context. Non-fine-tuned language models, lacking extensive task-specific knowledge, struggle with such semantic reasoning, potentially resulting in suboptimal predictions. Consequently, a template that aligns more closely with natural language semantics is necessary to enhance GPT-3's performance on this task.

For our second proposed template, we employ a format that more closely resembles natural language expressions. Specifically, the first line of this template presents the sentence requiring entity extraction. The subsequent four lines explicitly list the four entity types targeted for extraction, along with their corresponding entity names identified within the sentence. This structure is illustrated in **Figure 3**.

In practical testing, this second template demonstrated significantly higher usability compared to the first template, successfully performing partial entity extraction. This indicates that, based on the provided context, the language model partially comprehends both the content and the objective of the entity extraction task. However, employing this template for entity extraction still resulted in certain errors. In some instances, the language model failed to recognize the need to extract all four entity types, leading to over-extraction or under-extraction of specific entity types. Moreover, certain entities were erroneously assigned to multiple entity types. As exemplified in **Figure 3**, the person entity "Nadim Ladki" was simultaneously recognized as both a person and a location, highlighting an incorrect assignment of a single entity to two distinct types. This suggests that the template itself might introduce interference in the model's reasoning process.

The primary cause of these issues lies in the complexity of the template. Its highly structured output necessitates the language model to learn the task definition from the context, reason about the task itself, and simultaneously understand the structure of the expected output. Accomplishing these three tasks concurrently proves challenging for the language model, particularly when provided with limited in-context examples. Consequently, this template can yield unstable results and incorrect outputs.

During his visit to Slovenia, Kwasniewski is also scheduled to meet Prime Minister Janez Drnovsek, representatives of Slovenian political parties and representatives of the Chamber of Economy.
Location: Slovenia
Person: Kwasniewski, Janez Drnovsek
Organization: Chamber of Economy
Misc: Slovenian

DNEVNI AVAZ
Location:
Person:
Organization: DNEVNI AVAZ
Misc:

Petra Drnovice 1 Slovan Liberec 3
Location:
Person:
Organization: Petra Drnovice, Slovan Liberec
Misc:

(Legia Warsaw), Ryszard Wieczorek (Odra Wodzislaw)
Location:
Person: Ryszard Wieczorek
Organization: Legia Warsaw, Odra Wodzislaw
Misc:

Nadim Ladki
Location:
Person: Nadim Ladki
Organization: Nadim Ladki
Misc:

Figure 3. The "multi-line tag + entity" template

Building upon the second template, we propose a third template that further streamlines the process. In this refined template, each example consists of two lines: the first line presents the sentence for entity extraction, and the second line directly lists all the entities and their corresponding types, as demonstrated in **Figure 4**. The key advantage of this template lies in its simplified structure, requiring only a single line of output. This reduction in complexity alleviates the burden on the language model, eliminating the need for intricate reasoning about the output format. Consequently, the model can focus more directly on the core task of entity extraction, leading to improved consistency and accuracy. Moreover, as evidenced in **Figure 4**, this template exhibits superior performance in handling longer sentences, suggesting that the language model can more effectively infer the task objective from the context and accurately classify entity names when presented with a less convoluted structure.

The preliminary empirical analysis revealed that the third template consistently outperformed the other two in terms of accuracy. Consequently, the third template was adopted for subsequent metric computations and result summarization throughout the remainder of this investigation.

Fiorentina will be without three suspended players -- defenders Daniele Carnasciali and Lorenzo Amoruso and midfielder Emiliano Bigica -- for a difficult home match against unpredictable, attack-oriented Perugia led by in-form Croat striker Milan Rapajic and the experienced Fausto Pizzi.
Roma, now coached by Argentine Carlos Bianchi and watched by Italian national coach Arrigo Sacchi, lost 3-1 to Cesena -- another repeat of last season when the Rome club also went out at the first hurdle.
Entities: "Roma": organization, "Argentine": misc, "Carlos Bianchi": person, "Italian": misc, "Arrigo Sacchi": person, "Cesena": organization, "Rome": location
Wealthy Parma, now coached by the former Italian international Carlo Ancelotti, were without new striker Enrico Chiesa and went down 3-1 at serie B club Pescara in their second round clash.
Entities: "Parma": organization, "Italian": misc, "Carlo Ancelotti": person, "Enrico Chiesa": person, "Pescara": organization
Milan 's player of the year George Weah missed a good first half opportunity but otherwise looked a little rusty while Italian team mate Roberto Baggio did not play due to injury.
Entities: "Milan": organization, "George Weah": person, "Italian": misc, "Roberto Baggio": person
Cosenza 1 Fiorentina 3
Entities: "Cosenza": organization, "Fiorentina": organization
Fiorentina will be without three suspended players -- defenders Daniele Carnasciali and Lorenzo Amoruso and midfielder Emiliano Bigica -- for a difficult home match against unpredictable, attack-oriented Perugia led by in-form Croat striker Milan Rapajic and the experienced Fausto Pizzi.
Entities: "Fiorentina": organization, "Daniele Carnasciali": person, "Lorenzo Amoruso": person, "Emiliano Bigica": person, "Perugia": organization, "Milan Rapajic": person, "Fausto Pizzi": person

Figure 4. The "entity + tag" template

3.3. Evaluating the role of in-context example selection in NER

The present study introduces a novel strategy for selecting in-context examples based on the principle of utterance similarity. To assess the efficacy of this approach, a comparative analysis was undertaken, contrasting it against a baseline method involving the random selection of in-context examples. The experimental setup for both methodologies incorporated the utilization of four in-context examples. Notably, within the randomly selected examples, a deliberate inclusion of at least one statement encompassing all entity types was ensured. This measure aimed to mitigate the potential issue of encountering entities in the target statement that were absent from the in-context examples. The comparative performance of these two strategies was evaluated using the first 100 statements from the CoNLL-2003 training set, with the outcomes summarized in **Table 1**.

Table 1. Experiments on the strategy of choosing in-context examples

Selection strategy	F1	F1 (case sensitive)
Randomized selection	0.548	0.580
In this paper, we propose	0.706	0.721

The effectiveness of entity extraction was evaluated using the F1 score, a balanced measure that combines precision and recall. True Positives (TP) indicate entities accurately identified by the model, False Positives (FP) signify entities incorrectly predicted, and False Negatives (FN) represent entities overlooked by the model. The F1 score is calculated as the harmonic mean of precision (P) and recall (R), as shown below:

$$F1 = 2 \times (P \times R) / (P + R)$$

Where:

$$P = TP / (TP + FP)$$

$$R = TP / (TP + FN)$$

The F1 score, which balances precision and recall, provides a comprehensive view of the model's effectiveness. A higher F1 score means the model strikes a good balance between avoiding false positives (identifying things that aren't entities) and false negatives (missing actual entities), demonstrating its ability to extract entities accurately.

The empirical results presented in **Table 1** unequivocally demonstrate that the proposed in-context example selection methodology, grounded in semantic similarity, yields a substantial enhancement in the efficacy of Named Entity Recognition (NER), as evidenced by the markedly superior F1 scores. The underlying rationale for this improvement can be attributed to the inherent nature of the selection process. By prioritizing semantically similar examples, the model is furnished with context that is likely to exhibit structural and entity-type congruencies with the target utterance. This facilitates a more informed and accurate prediction process. In contrast, the random selection of examples lacks this inherent advantage, potentially leading to the presentation of contextually disparate information that may hinder the model's ability to discern relevant entities.

3.4. The influence of varying in-context example numbers on NER efficacy

The influence of the quantity of in-context examples on the outcomes was also investigated. Experiments were conducted employing 2, 4, 8, and 16 in-context examples, each selected based on semantic proximity to the target statement within the training set. The evaluation was performed on the initial 100 statements of the test set, and the results are presented in **Table 2**.

Table 2. Experiments on the strategy of choosing in-context examples

Number of examples	F1	F1 (case sensitive)
N = 2	0.652	0.656
N = 4	0.706	0.721
N = 8	0.813	0.818
N = 16	0.894	0.901

The results delineated in **Table 2** underscore the positive correlation between the quantity of in-context examples and the efficacy of NER. When furnished with a mere two examples, the model's performance, as measured by precision and recall, was notably suboptimal. Furthermore, the emergence of extraneous entity labels beyond the predefined set (e.g., "event") suggests that an insufficient number of in-context examples may impede GPT-3's ability to accurately delineate the task's scope and boundaries. The model, in essence, struggles to grasp the precise nature of the expected output when provided with limited contextual cues.

The progressive increment in the number of in-context examples to four led to a substantial improvement in accuracy, albeit with the persistence of occasional redundant label assignments. Further augmentation of the in-context examples to eight resulted in an additional enhancement of both accuracy and recall. The most pronounced improvement was observed with the inclusion of 16 in-context examples, where the F1 score approached an impressive 0.9. This empirical evidence strongly suggests that a larger pool of in-context examples facilitates a more refined and effective in-context learning process within GPT-3. However, it is crucial to acknowledge the trade-off between performance gains and computational overhead. The computational demands escalate exponentially with the increase in in-context examples. Consequently, a balanced approach was adopted, wherein N = 4 was selected as the optimal number of in-context examples for the proposed methodology. This configuration strikes a judicious equilibrium between computational efficiency and NER performance.

3.5. Comparative analysis with existing NER approaches

To establish a comparative benchmark against prevailing methodologies, a selection of prominent NER models was incorporated into the evaluation, encompassing Lattice LSTM + CRF, LR-CNN + CRF, LGN + CRF, FLAT, and NFLAT [10–14]. The experimental setup was meticulously maintained across all models, ensuring uniformity in dataset partitioning and environmental conditions. The outcomes of this comparative assessment are comprehensively detailed in **Table 3**.

Table 3. Comparison with related NER methods

Methods	F1	F1(case sensitive)
Lattice + LSTM + CRF	0.582	0.587
LR-CNN + CRF	0.589	0.592
LGN + CRF	0.596	0.601
FLAT	0.603	0.608
NFLAT	0.658	0.664
Proposed method	0.706	0.721

Note: Long Short-Term Memory (LTSM), Conditional Random Field (CRF), Logistic Regression (LR), Convolutional Neural Network (CNN), Local Graph Network (LGN), Flat-Lattice Transformer (FLAT), Nested Flat-Lattice Transformer (NFLAT)

The empirical evidence presented in **Table 3** unequivocally establishes the superiority of the proposed methodology in comparison to both the FLAT and NFLAT models. Furthermore, it exhibits a performance advantage over the traditional Lattice, LR, and LGN methods. The observed outcomes underscore the efficacy of leveraging pre-trained large language models in conjunction with contextual samples to enhance the comprehension of semantic content within the context of Named Entity Recognition (NER). The inherent capacity of these models to capture intricate linguistic nuances and contextual dependencies contributes to their superior performance in discerning and classifying named entities within textual data.

4. Conclusion and outlook

The empirical investigations conducted in this study have unequivocally demonstrated the advantages of the proposed strategy for selecting in-context examples, highlighting its pivotal role in enhancing the efficacy of Named Entity Recognition (NER) within the GPT-3 framework. The findings further corroborate the notion that a more extensive repertoire of in-context examples serves to streamline and enhance the inferential process inherent to in-context learning. In summary, this research endeavor has successfully introduced and validated a schema for the selection of in-context examples, predicated upon semantic similarity, to populate templates for GPT-3-driven NER. The empirical evidence garnered through rigorous experimentation lends credence to the viability and effectiveness of this schema.

The current methodology, while effective, does exhibit a limitation in its capacity to pinpoint the precise location or span of identified entities within the utterance. This capability could prove instrumental in downstream NER applications, facilitating tasks such as information extraction and knowledge graph construction. Future research endeavors could explore the refinement of the existing template to enable GPT-3 to furnish entity positional information or the integration of supplementary techniques for entity localization. Such advancements would undoubtedly further amplify the utility and impact of this NER paradigm, solidifying its position as a valuable tool in the realm of natural language processing.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Grishman R, Sundheim B, 1996, Message Understanding Conference: A Brief History. 16th International Conference on Computational Linguistics, 1996: 466–471.

[2] Qu X, Gu Y, Xia Q, et al., 2024, A Survey on Arabic Named Entity Recognition: Past, Recent Advances, and Future Trends. IEEE Transactions on Knowledge and Data Engineering, 36(3): 943–959.

[3] Hu Z, Hou W, Liu X, 2024, Deep Learning for Named Entity Recognition: A Survey. Neural Computing and Applications, 36(16): 8995–9022.

[4] Jiang Y, Hu C, Xiao T, et al., 2019, Improved Differentiable Architecture Search for Language Modeling and Named Entity Recognition. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 3583–3588.

[5] Strubell E, Verga P, Belanger D, et al., 2017, Fast And Accurate Entity Recognition with Iterated Dilated Convolutions. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017: 2670–2680.

[6] Ronran C, Lee S, 2020, Effect of Character and Word Features in Bidirectional LSTM-CRF for NER. 2020 IEEE International Conference on Big Data and Smart Computing, 2020: 613–616.

[7] Baevski A, Edunov S, Liu Y, et al., 2019, Cloze-driven Pretraining of Self-Attention Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 5359–5368.

[8] Brown TB, Mann B, Ryder N, et al., 2020, Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems, 2020: 1–75.

[9] Wies N, Levine Y, Shashua A, 2023, The Learnability of In-context Learning. Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, 2023: 1–15.

[10] Zhang Y, Yang J, 2018, Chinese NER using lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1554–1564.

[11] Gui T, Ma R, Zhang Q, et al., 2019, CNN-based Chinese NER with Lexicon Rethinking. Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019: 4982–4988.

[12] Gui T, Zou Y, Zhang Q, et al., 2019, A Lexicon-based Graph Neural Network for Chinese NER. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 1040–1050.

[13] Li X, Yan H, Qiu X, et al., 2020, FLAT: Chinese NER using Flat-lattice Transformer. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020: 6836–6842.

[14] Wu S, Song X, Feng ZH, et al., 2021, NFLAT: Non-flat-lattice Transformer for Chinese Named Entity Recognition. Journal of Latex Class Files 14(8): 1–13. https://doi.org/10.48550/arXiv.2205.05832