TURNER: The Uncertainty-based Retrieval Framework for Chinese NER

Zhichao Geng, Hang Yan, Zhangyue Yin, Chenxin An, Xipeng Qiu*, Xuanjing Huang
School of Computer Science, Fudan University
Key Laboratory of Intelligent Information Processing, Fudan University
{zcgeng20, hyan19, xpqiu, xjhuang}@fudan.edu.cn

Abstract
Chinese NER is a difficult undertaking due to the ambiguity of Chinese characters and the absence of word boundaries. Previous work on Chinese NER focus on lexicon-based methods to introduce boundary information and reduce out-of-vocabulary (OOV) cases during prediction. However, it is expensive to obtain and dynamically maintain high-quality lexicons in specific domains, which motivates us to utilize more general knowledge resources, e.g., search engines. In this paper, we propose TURNER: The Uncertainty-based Retrieval framework for Chinese NER. The idea behind TURNER is to imitate human behavior: we frequently retrieve auxiliary knowledge as assistance when encountering an unknown or uncertain entity. To improve the efficiency and effectiveness of retrieval, we first propose two types of uncertainty sampling methods for selecting the most ambiguous entity-level uncertain components of the input text. Then, the Knowledge Fusion Model re-predict the uncertain samples by combining retrieved knowledge. Experiments on four benchmark datasets demonstrate TURNER’s effectiveness. TURNER outperforms existing lexicon-based approaches and achieves the new SOTA.

1 Introduction
Named Entity Recognition (NER) plays an essential part in natural language processing. The performance of the NER task has improved dramatically as a result of the recent advances in deep learning and pre-trained models. In comparison to English NER, Chinese NER is more challenging due to the character ambiguity and absence of word boundaries.

In recent years, lexicon-based approaches have dominated the Chinese NER field [Zhang and Yang, 2018; Li et al., 2020; Liu et al., 2021; Wang et al., 2021a]. Due to their ability to reduce OOV occurrence during prediction, lexicon-based approaches have become the necessary element for obtaining SOTA results in Chinese NER. These approaches begin by matching the input sentence with the lexicon and then incorporating the matched items’ boundary and semantic information into the model via a variety of methods. However, the lexicon-based approaches have several drawbacks. First, model performance is highly dependent on the quality of the lexicon, yet obtaining and dynamically maintaining a high-quality lexicon in a particular application scenario is costly. Second, existing methods are incapable of utilizing dynamically changing lexicons while introducing semantic knowledge and vice versa. A line of previous work [Li et al., 2020; Liu et al., 2021] utilizes word embeddings to represent matched items, but updating the lexicon requires retraining the model. Another tributary of previous work [Wang et al., 2021a] leverages only the word boundary information, which enables the usage of a dynamically updated lexicon, but this method discards the semantic information of words.

Considering the limits of lexicon-based methods, retrieval
methods, e.g. retrieving with search engines, are more generic resources of knowledge in the NER task. The retrieved knowledge can help eliminate ambiguity in the input text and enhance performance. Additionally, retrieved knowledge is presented as text, allowing the query-knowledge base to be dynamically updated. However, existing retrieval methods [Torisawa and others, 2007; Wang et al., 2021b] in the NER task retrieve the knowledge using traversal strategies, which is inefficient, especially in the case of online retrieval. Moreover, existing pre-trained models are sufficiently powerful to handle the majority of NER cases. Retrieving knowledge for simple samples that the model could have dealt with correctly is ineffective.

We notice that in Chinese NER, the model can produce diverse predictions about some input components, i.e., the model is uncertain about these components. Furthermore, our statistics demonstrate that the uncertain components are the bottleneck of performance. Therefore, retrieving knowledge for the uncertain components can provide the maximum support for the prediction. In this paper, we propose TURNER: The Uncertainty Retrieval framework for Chinese NER. By leveraging Monte-Carlo Dropout [Gal and Ghahramani, 2016] or parsing the top-K label sequences predicted by the base NER model, we can obtain the entity-level uncertain components with a high degree of complexity and uncertainty for the base NER model. Then by querying on the uncertain components, knowledge can be retrieved efficiently and effectively, eliminating uncertainty, as shown in Figure 1. Afterward, the Knowledge Fusion Model re-predicts the uncertain samples by integrating the original input text, the retrieved knowledge, and the output of the base NER model. We conduct experiments on 4 Chinese NER datasets to evaluate the effectiveness of our method. The results illustrate TURNER significantly outperforms the base NER model on all datasets, boosting the F1 score by 1.1 on average. TURNER also outperforms existing lexicon-based methods on Chinese NER, achieving new SOTA results. To the best of our knowledge, we are the first to combine uncertainty sampling and retrieval methods in the NER task.

2 Related Work and Background

2.1 Chinese Named Entity Recognition

With the advent of deep learning, neural networks have become a dominant solution to the NER task [Ma and Hovy, 2016]. In recent years, lexicon-based methods have become widespread for Chinese NER. Zhang and Yang [2018] propose lattice LSTM to encode both lexicon information and character information. Li [2020] propose the Flat-Lattice Transformer as a way to incorporate lexicon information into the transformer model without any loss of information. Liu [2021] make use of the BERT Adapter [Houlsby et al., 2019] to introduce lexicon information to the bottom layers of BERT. Wang [2021a] propose DyLex, which utilizes only the boundary information of words, allowing models to use dynamically updating lexicons.

2.2 Retrieval Methods for Neural Models

Retrieving knowledge from the database to enhance neural models has been applied in several NLP tasks [Qiu et al., 2014; Hashimoto et al., 2018; Gu et al., 2018]. They either use the retrieved text to guide the generation process or expand the training set automatically. In the NER task, some works introduce knowledge from the knowledge graph or Wikipedia [Torisawa and others, 2007; He et al., 2020]. They retrieve knowledge by traversing matches all possible sub-sequences, then encode the knowledge to incorporate it into the model. Wang [2021b] chunk the sentence into sub-sentences based on punctuation, then utilizes the sub-sentences as queries in the search engine to retrieve external context. The knowledge is chosen from the retrieved items that have the highest association with the original text.

2.3 Uncertainty

The predictive result in the classification task can be uncertain even with a high softmax output [Gal and Ghahramani, 2016]. Monte-Carlo Dropout [Gal and Ghahramani, 2016] (MC Dropout), a Bayesian approximation of the Gaussian process, is a general approach to estimate uncertainty. MC Dropout keeps dropout active in the prediction phase and evaluates the uncertainty according to the difference of several forward passes. In the sequence labeling task, Gui [2020] employ MC Dropout to create draft labels and locate the uncertain labels. Furthermore, they utilize two-stream attention to model long-term dependencies between draft labels and refine them, thereby getting rid of Viterbi decoding.

3 Approach

In this section, we will first introduce our uncertainty sampling methods for sequence labeling NER models, followed by our framework for Chinese NER utilizing the uncertainty.

3.1 Uncertainty Sampling and Retrieving

Sampling Method

MC Dropout MC Dropout [Gal and Ghahramani, 2016] is a general approach to obtain the uncertain components. Given an input sequence for the sequence labeling NER model, we first obtain the model’s prediction as the provisional result. Then, we utilize MC dropout to keep dropout active and generate $K$ candidate label sequences with Viterbi decoding in $K$ forward passes. As illustrated in Table 1, the difference between the predicted entity set of each candidate label sequence and the provisional result can be considered uncertain entities as the model predict them under mild perturbation. The uncertain components can then be obtained by merging all adjacent and overlapping uncertain entities. We sample uncertain entities rather than position-level labels because querying at the entity level improves retrieval precision.

Top-K Label Sequences MC dropout demands more GPU resources as it requires multiple forward passes. As a result, we also give an alternative strategy for sampling uncertainty in NER tasks. Given the label probability distribution of an input sequence, we can obtain $K$ legal label sequences with top-$K$ scores based on the variant of the Viterbi algorithm [Brown and Golod, 2010]. Previous works rerank the
Stage One: Provisional Result and Uncertainty Sampling

In the first stage, we use a base NER model to predict on the input sequence to get provisional results and uncertain components. The base NER model can be any sequence labeling model, such as BERT+MLP.

Formally, given input sequence \( X = [c_1, c_2, ..., c_n] \), we apply the base NER model to obtain the provisional results \( L_P \) and candidate label sequences \( L_C \) with dropout active. If there is no uncertain component in all candidates, the base NER model is highly confident about the current sample, then \( L_P \) is the final result. Otherwise, we take \( L_P \) as the provisional result and obtain the uncertain component \( U = [c_1, c_2+1, ..., c_{t+0}] \) using the method in Section 3.1. Afterward, we can utilize \( U \) as query to retrieve the knowl-
Stage Two: Knowledge Fusion Prediction

In the second stage, we employ $K$ as auxiliary knowledge to re-predict the sample with uncertain components $U$ using the Transformer-based Knowledge Fusion Model. To introduce the retrieved knowledge without losing any information, we concatenate $K$ behind $X$ to obtain the knowledge-enhanced input sequence $\tilde{X} = [c_1, c_2, ..., c_n, \{SEP\}, k_1, k_2, ..., k_m]$.

Considering the base NER model’s excellent accuracy in the confident components, we incorporate the label information of the provisional result to support the Knowledge Fusion Model. To avoid confusion or misleading, we mask the uncertain labels in the provisional result and obtain the modified provisional result $\tilde{L}_P$ as follows:

$$\tilde{l}_i = \begin{cases} l_i & \text{if } i \leq n \text{ and } c_i \notin U \\ \text{[mask]} & \text{if } c_i \in U \\ \text{[pad]} & \text{if } i > n \end{cases},$$

$$\tilde{L}_P = [\tilde{l}_1, \tilde{l}_2, ..., \tilde{l}_n, \tilde{l}_{n+1}, ..., \tilde{l}_{n+m+1}],$$

where $l_i$ is the i-th label of the provisional result and $\tilde{l}_i$ is the i-th label in $\tilde{L}_P$. Afterward, we incorporate $\tilde{X}$ and $\tilde{L}_P$ using the Transformer-based Knowledge Fusion Model to obtain the knowledge-enhanced probability distribution $\tilde{D}$:

$$H_{\tilde{L}_P} = \text{Label Embedding}(\tilde{L}_P),$$

$$H_\tilde{X} = \text{Character Embedding}(\tilde{X}),$$

$$\tilde{D} = \text{Transformer Encoder}(H_{\tilde{L}_P} + H_\tilde{X}),$$

where label embedding and character embedding are trainable parameters. The self-attention mechanism ensures that all information is fully integrated and the knowledge is efficiently utilized. Ultimately, we decode $\tilde{D}$ using the Viterbi algorithm to obtain the final result $\tilde{L}$. If there are multiple uncertain components in one input sequence, we process them separately in the second stage and sum all obtained $\tilde{D}$ before Viterbi decoding.

3.3 Training of TURNER

Since the parameters of the base NER model and the Knowledge Fusion Model are independent, the two-stage training procedure can be undertaken independently. The training of the base NER model is trivial, and we will focus on the training of the Knowledge Fusion Model.

To minimize the gap between training and prediction, the uncertainty of training data must be sampled. Therefore, we conduct N-fold jackknifing to divide the training data into N equal parts. In each step, we train the base NER model with N-1 pieces of data and then use the base NER model to sample uncertainty for the remaining one piece of data. Additionally, we conduct data augmentation by leveraging various checkpoints of the base NER model when generating training data for the second stage. For a given training sample, we append the sampled components of all checkpoints sequentially and discard the components that are significantly overlapped with existing uncertain components.

However, the data augmentation strategy will result in the repetition of some training samples, which may lead to overfitting on several input sequences. Therefore, we introduce the position-wised weighted average for the cross-entropy loss. By decreasing the weight assigned to the confident component, the model can concentrate on the uncertain components rather than overfitting the simple cases. Formally, we calculate the loss $\mathcal{L}$ as follows:

$$\lambda_i = \begin{cases} 1 & \text{if } c_i \in U \\ \alpha & \text{if } c_i \notin U \end{cases},$$

$$\mathcal{L} = \sum_{i=1}^{n} \lambda_i \cdot \text{loss}_i,$$

where $\lambda_i$ is the weight coefficient at position $i$; loss$_i$ is the cross-entropy loss at position $i$; $\alpha$ is a hyper parameter ranges $[0, 1]$. And we do not calculate loss for the knowledge text.

4 Experiments

We conduct comprehensive experiments to evaluate the effectiveness of TURNER. We use the BIESO label set and employ the standard F1-score as the evaluation metric. All experiments are conducted on 8 GeForce RTX 3090.

4.1 Datasets

We conduct experiments on four benchmark Chinese NER datasets to evaluate TURNER, including MSRA [Levow, 2006], Ontonotes 4.0 [Weischedel et al., 2011], Resume [Zhang and Yang, 2018] and Weibo [Peng and Dredze, 2015]. MSRA and Ontonotes are annotated using data from the news domain; Weibo is annotated from the Internet blogs;
Resume is annotated from the resumes. We follow the same train, dev, test split of the official version and previous work [Li et al., 2020]. The statistical information of each dataset is shown in Table 3.

| Type   | Train | Validation | Test  |
|--------|-------|------------|-------|
| MSRA   | Size  | 46364      | -     | 43.65 |
|        | Length$_{avg}$ | 46.8      | -     | 39.5  |
| Ontonotes | Size  | 15724      | 4301  | 4346  |
|        | Length$_{avg}$ | 31.3      | 46.6  | 47.9  |
| Weibo  | Size  | 1350       | 270   | 270   |
|        | Length$_{avg}$ | 54.7      | 53.7  | 55    |
| Resume | Size  | 3821       | 463   | 477   |
|        | Length$_{avg}$ | 32.5      | 30    | 31.7  |

Table 3: Dataset statistics. "Size" denotes the number of samples in the sub-set. "Length$_{avg}$" denotes the average length of samples.

4.2 Settings

Models and Training. We implement a strong base NER model using BERT$^2$+MLP in the first stage of TURNER. We do not use CRF because we observe that the same performance can be obtained by relying just on the Viterbi algorithm to constrain that there is no illegal transition in the predictive result. The Knowledge Fusion Model is also initialized using BERT. We search the learning rate in $\{2e^{-5}, 3e^{-5}\}$ and the weight coefficient $\alpha$ in $\{0.1, 1\}$, and other hyperparameters used are listed in Appendix A. We use the validation set to select the best checkpoint and report its performance on the test set.

Uncertainty and Retrieval. During prediction, we generate eight candidates for MC Dropout and four candidates for the top-$K$ method. To generate training data for the second stage of TURNER, we use MC Dropout for uncertainty sampling. The statistical results of the uncertainty sampling on four datasets are listed in Appendix B. For retrieving, we use Baidu, which is a widely used Chinese search engine. We prioritize the retrieved items belonging to the encyclopedia category and do not change the order of other retrieved items. Furthermore, each retrieved item is limited to a title and up to 50 characters of content, and all items are concatenated in order. The retrieved knowledge is limited to 400 characters.

4.3 Overall Performance

Table 4 shows the overall performance of TURNER in four benchmark datasets of Chinese NER. In the top part of the table we show the performance of lexicon-free baselines, including BiLSTM+CRF [Huang et al., 2015], TENER [Yan et al., 2019] and ERNIE [Sun et al., 2019]. In the second part of the table we provide several latest strong lexicon-based baselines, including Lattice-LSTM [Zhang and Yang, 2018], FLAT+BERT [Li et al., 2020], LEBERT [Liu et al., 2021], DyLex+BERT [Wang et al., 2021a]. In the bottom part, we show the performance of TURNER with two uncertainty sampling methods. The results demonstrate that TURNER significantly enhances the base NER model’s performance, boosting the average F1 score by 1.1 with both two uncertainty sampling methods and achieving new SOTA results. Although lexicon-based baselines are obviously superior to lexicon-free baselines, TURNER outperforms existing lexicon-based methods in each dataset, proving the potential of lexicon-free methods in Chinese NER.

TURNER makes more substantial advancements in MSRA and Ontonotes, which belong to the news domain, because entities referenced in the news are more likely to be retrieved by search engines. The improvement on Weibo is relatively limited, as many entities are usernames or nicknames, making it more challenging to retrieve related information using the search engine. It can be inferred that the effectiveness of TURNER is directly proportional to the relevance of the retrieved knowledge.

4.4 Ablation Study

We conduct ablation experiments to verify the contribution of each component of TURNER. The settings of the ablation experiment are shown as follows: (a) W/o retrieve: We do not retrieve knowledge for the uncertain components, i.e., the input for the Knowledge Fusion Model $\tilde{X}$ equals $X$. (b) W/o uncertainty: We abandon the two-stage architecture and do not sample the uncertainty. As a result, we use sentence-level queries to retrieve knowledge for each sample like Wang[2021b]. The knowledge is also concatenated behind the original input text. (c) W/o LE: We do not use the label embedding to provide the provisional results to the Knowledge Fusion Model. (c) W/o WL: We use the unweighted average method to calculate the loss for the Knowledge Fusion Model, which is equivalent to setting $\alpha$ to 1.

As shown in Table 5, the two-stage uncertain samples re-predicting architecture can be regarded as a particular way of model ensembling, which can boost the performance a bit without knowledge. Without uncertainty sampling, the
The contribution of context retrieved via sentence-level queries is limited. By combining uncertainty sampling and retrieval, TURNER can boost the performance to the greatest extent. The label embedding and weighted-average loss also contribute much to the Knowledge Fusion Model.

5 Discussion

5.1 TURNER With Offline Retrieval

The time cost of retrieval from search engines is still too high in some application scenarios. As a result, we conduct experiments with an offline knowledge graph to test TURNER’s generalization capabilities. We use the owntink knowledge graph, which contains more than 140 million (Subject, Predicate, Object) triplets about general knowledge. For each subject in the knowledge graph, we merge all related triplets containing it to generate a descriptive document. Then, for each query, we use the BM25 algorithm to retrieve the three most relevant descriptive documents as auxiliary knowledge in TURNER.

|                | MSRA   | Ontonotes | Weibo  | Resume |
|----------------|--------|-----------|--------|--------|
| TURNER$^{\text{MC}}$ | 96.85  | 83.56     | 70.78  | 96.36  |
| w/o retrieve  | 96.33  | 81.93     | 70.15  | 96.11  |
| w/o uncertainty | 95.85  | 82.45     | 70.27  | 95.87  |
| w/o LE        | 96.70  | 82.72     | 69.8   | 96.0   |
| w/o WL        | 96.85  | 82.60     | 70.78  | 96.30  |

Table 6: The performance of TURNER with offline retrieval.

As illustrated in Table 6, TURNER with offline retrieval outperforms existing SOTA lexicon-based methods in MSRA, Ontonotes, and Resume, which can be considered in-domain datasets for the knowledge graph. Because many internet phrases are out-of-domain for the knowledge graph, the offline retrieval does not operate effectively in Weibo. The results indicate the generalization ability of TURNER, which can work with various knowledge bases. The performance of offline retrieval is not as good as the search engines because search engines have broader knowledge and more precise retrieval.

5.2 MC Dropout vs. Top-$K$ Label Sequences

Because the two uncertainty sampling approaches are on par in performance, we undertake a more detailed analysis to compare them. We use two metrics to measure the quality of the uncertainty sampling process: (a) SAR: since most candidates are duplicate and will be filtered out, we use Sampling Acceptance Ratio (SAR) to denote the ratio of candidates that can be utilized in uncertainty sampling rather than being filtered out; (b) VSR: we use Valuable Sampling Ratio (VSR) to denote the ratio of candidates that yield a better F1 score than the provisional results after the filter. The higher the VSR, the better the uncertainty sampling recall ability. And the VSR/A SR ratio reflects how much noise is present in the sampling results. As illustrated in Table 7, the top-$K$ method has stronger ability to recognize uncertain components and yield better oracle scores, but there is also more noise, resulting in more retrieval times.

|                | MSRA   | Ontonotes |
|----------------|--------|-----------|
| SAR            | MC     | Top-$K$   |
|                | 0.013  | 0.096     |
| VSR            | 0.003  | 0.009     |
| Avg Retrieval  | 0.089  | 0.162     |
| Oracle F1      | 98.09  | 98.41     |

Table 7: The statistical results of two uncertainty sampling methods. "Avg Retrieval" denotes the average retrieval times for each sample.

From the perspective of computational efficiency, the additional GPU cost of two sampling methods can be calculated as follows:

\[
COST_M = O((k + \beta_1(1 + \gamma)^2) \cdot C),
\]

\[
COST_{Top-K} = O(\beta_2(1 + \gamma)^2 \cdot C),
\]

where $k$ is the number of candidates, $\beta$ is the retrieval ratio, $\gamma$ is the ratio of the length of the knowledge text to the original input text, and $C$ is the cost of the base NER model, i.e., BERT+MLP. MC Dropout demands more GPU resources in the first stage, while the top-$K$ method results in more retrieval times and consumes more GPU resources in the second stage because it samples more uncertain components. The ultimate choice should be based on actual details.

6 Conclusion and Future Work

In this paper, we propose TURNER, a framework for Chinese NER that integrates uncertainty sampling and knowledge retrieval. TURNER applies MC Dropout or top-$K$ label sequences to identify the uncertain components that lead to ambiguity with a base NER model. Afterward, the uncertain components can be utilized as queries to retrieve knowledge efficiently and effectively. Furthermore, the Knowledge Fusion Model re-predict the uncertain samples based on the retrieved knowledge. Comprehensive experiments on four datasets illustrate TURNER significantly outperforms existing lexicon-based methods and achieves new SOTA results.

For existing NLP technologies, it is impossible to dynamically maintain all required knowledge in neural models. The retrieval methods utilizing external knowledge bases can be solutions to this issue, while our uncertainty-based retrieval paradigm makes retrieval methods more effective and efficient. In the future, we will try to broadcast the uncertainty-based retrieval paradigm to more NLP tasks and applications.
References

[Brown and Golod, 2010] Daniel G Brown and Daniil Golod. Decoding hmms using the k best paths: algorithms and applications. BMC bioinformatics, 11(1):1–7, 2010.

[Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In International Conference on Machine Learning, 2016.

[Gu et al., 2018] Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. Search engine guided neural machine translation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[Gui et al., 2020] Tao Gui, Jiacheng Ye, Qi Zhang, Zhengyan Li, Zichu Fei, Yeyun Gong, and Xuanjing Huang. Uncertainty-aware label refinement for sequence labeling. In Empirical Methods in Natural Language Processing, 2020.

[Hashimoto et al., 2018] Tatsunori B. Hashimoto, Kelvin Guu, Yonatan Oren, and Percy Liang. A retrieve-and-edit framework for predicting structured outputs. In Neural Information Processing Systems, 2018.

[He et al., 2020] Qizhen He, Liang Wu, Yida Yin, and Heming Cai. Knowledge-graph augmented word representations for named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 7919–7926, 2020.

[Houlsby et al., 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzbski, Bruna Halila Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning, 2019.

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

[Levow, 2006] Gina-Anne Levow. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 108–117, 2006.

[Li et al., 2020] Xiaonian Li, Hang Yan, Xipeng Qiu, and Xuan-Jing Huang. Flat: Chinese ner using flat-lattice transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6836–6842, 2020.

[Liu et al., 2021] Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. Lexicon enhanced Chinese sequence labeling using BERT adapter. 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 5847–5858, Online, August 2021. Association for Computational Linguistics.

[Ma and Hovy, 2016] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In Meeting of the Association for Computational Linguistics, 2016.

[Peng and Dredze, 2015] Nanyun Peng and Mark Dredze. Named entity recognition for chinese social media with jointly trained embeddings. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 548–554, 2015.

[Qiu et al., 2014] Xipeng Qiu, Chaochao Huang, and Xuanjing Huang. Automatic corpus expansion for chinese word segmentation by exploiting the redundancy of web information. In International Conference on Computational Linguistics, 2014.

[Sun et al., 2019] Yu Sun, Shuhuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223, 2019.

[Torisawa and others, 2007] Kentaro Torisawa et al. Exploiting wikipedia as external knowledge for named entity recognition. In Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), pages 698–707, 2007.

[Wang et al., 2021a] Baojun Wang, Zhao Zhang, Kun Xu, Guang-Yuan Hao, Yuyang Zhang, Lifeng Shang, Linlin Li, Xiao Chen, Xin Jiang, and Qun Liu. Dylex: Incorporating dynamic lexicons into bert for sequence labeling. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2679–2693, 2021.

[Wang et al., 2021b] Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 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 (Volume 1: Long Papers), pages 1800–1812, Online, August 2021. Association for Computational Linguistics.

[Weischedel et al., 2011] Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. Ontonotes release 4.0. LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium, 2011.

[Yan et al., 2019] Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. Tener: adapting transformer encoder for named entity recognition. arXiv preprint arXiv:1911.04474, 2019.

[Yang et al., 2017] Jie Yang, Yue Zhang, and Fei Dong. Neural reranking for named entity recognition. In Recent Advances in Natural Language Processing, 2017.

[Zhengyan Li, Zichu Fei, Yeyun Gong, and Xuanjing Huang. Uncertainty-aware label refinement for sequence labeling. In Empirical Methods in Natural Language Processing, 2020.

[Hashimoto et al., 2018] Tatsunori B. Hashimoto, Kelvin Guu, Yonatan Oren, and Percy Liang. A retrieve-and-edit framework for predicting structured outputs. In Neural Information Processing Systems, 2018.

[He et al., 2020] Qizhen He, Liang Wu, Yida Yin, and Heming Cai. Knowledge-graph augmented word representations for named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 7919–7926, 2020.

[Houlsby et al., 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzbski, Bruna Halila Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In International Conference on Machine Learning, 2019.

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

[Levow, 2006] Gina-Anne Levow. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 108–117, 2006.

[Li et al., 2020] Xiaonian Li, Hang Yan, Xipeng Qiu, and Xuan-Jing Huang. Flat: Chinese ner using flat-lattice transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6836–6842, 2020.

[Liu et al., 2021] Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. Lexicon enhanced Chinese sequence labeling using BERT adapter. 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 5847–5858, Online, August 2021. Association for Computational Linguistics.

[Ma and Hovy, 2016] Xuezhe Ma and Eduard Hovy. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In Meeting of the Association for Computational Linguistics, 2016.
A Hyperparameters

| Hyperparameter | Baseline Model | Knowledge Fusion Model |
|----------------|----------------|------------------------|
| Epochs         | 20             | 10                     |
| Batch Size     | 32             | 32                     |
| Weight Decay   | 0              | 0                      |
| Dropout        | 0.1            | 0.1                    |
| Learning Rate  | 2e-5           | (2e-5,3e-5)            |
| Optimizer      | AdamW          | AdamW                  |
| Warm Up Ratio  | 0.1            | 0.1                    |
| Max Seq_Len    | 128            | 512                    |
| $\alpha$       | -              | {0.1,1}                |

Table 8: The hyperparameters we used for training. Other hyperparameters are the same as BERT. For the baseline model, we use same hyperparameters in all datasets. For the Knowledge Fusion Model, the search range is the same in all datasets.

B Uncertainty Sampling Results

| Type       | MC Dropout | Train | Validation | Test |
|------------|------------|-------|------------|------|
| MSRA       |            |       |            |      |
| Size $U$   | 10748      | -     | 335        |      |
| Num $U/C$  | 13748      | -     | 388        |      |
| Ontonotes  |            |       |            |      |
| Size $U$   | 4471       | 1133  | 1229       |      |
| Num $U/C$  | 6403       | 1464  | 1501       |      |
| Weibo      |            |       |            |      |
| Size $U$   | 614        | 86    | 93         |      |
| Num $U/C$  | 1027       | 133   | 137        |      |
| Resume     |            |       |            |      |
| Size $U$   | 1309       | 50    | 38         |      |
| Num $U/C$  | 1798       | 53    | 53         |      |

Table 9: Statistical results of two uncertainty sampling methods. "Size $U$" denotes the number of samples that contain at least one uncertain component. "Num $U/C$" denotes the total number of uncertain components in all samples.

C Case Study

In order to better illustrate the effectiveness of retrieved knowledge, we give a case of retrieved knowledge in Table 10.

Original Text

实践越来越证明，"亚太经合组织方式"符合本地区的实际，有利于各国成员在不同程度和要求得到较好的平衡，有助于发挥各国成员的能力，促进共同发展。

Uncertain Component

亚太经合组织

Knowledge w/ Search Engine

亚太经合组织的百度百科简介：亚太经合组织一般指亚太经济合作组织。亚太经济合作组织（英文：asia-pacificeco;亚太经合组织；asia-pacificeconomiccooperat...apecministers'meeting08-09November2021viri;亚太经合组织超载实现悉尼林业目标,政務澎湃新闻，thepaper,11月9日，亚太经合组织部长级会议联合发布部长声明，对亚太区域森林面积增长2790万公顷，如期实现，习近平在亚太经合组织工商领导人峰会上的主旨演讲(全文);坚持可持续发展共建亚太命运共同体在亚太经合组织工商领导人峰会上的主旨演讲(2021年11月1日);亚太经合组织概况，中华人民共和国外交部:【成立】亚太经合组织(asia-pacificeconomiccooperation,apel)习近平在亚太经合组织第二十八次领导人非正式会议上的讲话;很高兴同大家见面。首先，我感谢阿德恩总理和新西兰政府为本次会议作出的努力。本次会议以推动疫后经济复苏增加2650万公顷!中国为亚太地区森林增长贡献巨大!亚太经合组织超载实现悉尼林业目标!澳媒社论：亚太经济合作迈入关键一步!悉尼林业目标超载实现,中国森林面积13年增2650...携手开创亚太经济合作新篇章

Knowledge w/ Knowledge Graph

亚太经合组织第二十八次领导人非正式会议发言。北京同各国领导人：构建融合、创新、互联的亚太经合组织第二十八次领导人非正式会议发言。亚太经合组织各成员领导人聚首北京雁栖湖畔，举行亚太经合组织第二十八次领导人非正式会议。「中文名:亚太经合组织第二十八次首脑亚太经合组织部长级会议」，首脑亚太经合组织部长级会议。在人民大会堂开幕；中文名:亚太经合组织第二十八次首脑亚太经合组织林业部长级会议」，首脑亚太经合组织林业部长级会议，亚太经合组织领导人非正式会议。「中文名:亚太经合组织林业部长级会议」，首脑亚太经合组织林业部长级会议，亚太经合组织领导人非正式会议。「中文名:亚太经合组织林业部长级会议」，首脑亚太经合组织林业部长级会议，亚太经合组织领导人非正式会议。「中文名:亚太经合组织林业部长级会议」，首脑亚太经合组织林业部长级会议，亚太经合组织领导人非正式会议。「中文名:亚太经合组织林业部长级会议」，首脑亚太经合组织林业部长级会议，亚太经合组织领导人非正式会议。外文名:thearnualapeconomicleaders'meetings,成立时间:1992年4月。首脑

Table 10: An example of retrieved knowledge.
D Analysis on Uncertainty Sampling Methods

We give an investigation about the influencing factors of uncertainty sampling using the testset of MSRA, and the results are shown in Table 11. For all influencing factors, the performance shows a trend of first rising and then falling, indicating that there are local optimal choices.

| MC Dropout, Candidate Num = 8 |
|-----------------------------|
| Drop Rate | 0.05 | 0.1 | 0.2 | 0.4 |
| SAR          | 0.008 | 0.013 | 0.031 | 0.221 |
| VSR          | 0.002 | 0.003 | 0.005 | 0.002 |
| F1 Score   | 96.74 | 96.85 | 96.66 | 96.29 |

| MC Dropout, Drop Rate = 0.1 |
|------------------------------|
| Candidate Num | 2   | 4   | 8   | 16  |
| SAR             | 0.023 | 0.018 | 0.013 | 0.009 |
| VSR             | 0.006 | 0.005 | 0.003 | 0.002 |
| F1 Score       | 96.56 | 96.79 | 96.85 | 96.75 |

| Top-K Method |
|--------------|
| Candidate Num | 1   | 2   | 4   | 8   |
| SAR           | 0.118 | 0.104 | 0.096 | 0.088 |
| VSR           | 0.019 | 0.014 | 0.009 | 0.006 |
| F1 Score      | 96.42 | 96.50 | 96.54 | 96.46 |

Table 11: The results on MSRA datasets of our investigation.

For MC Dropout, as the dropout rate increases, the model’s capacity decreases significantly and more noisy candidates are generated. Therefore, with enough computational resources, finding the optimal forward times is a more general solution. The Top-K method generates more valuable candidates as well as more noisy candidates.