Exploiting Multiple Embeddings for Chinese Named Entity Recognition

Canwen Xu∗
Wuhan University
School of Computer Science
xucanwen@whu.edu.cn

Feiyang Wang∗
Wuhan University
School of Computer Science
wangfeiyang@whu.edu.cn

Jialong Han
Tencent AI Lab
jialonghan@gmail.com

Chenliang Li†
Wuhan University
School of Cyber Science and Engineering
cllee@whu.edu.cn

ABSTRACT

Identifying the named entities mentioned in text would enrich many semantic applications at the downstream level. However, due to the predominant usage of colloquial language in microblogs, the named entity recognition (NER) in Chinese microblogs experience significant performance deterioration, compared with performing NER in formal Chinese corpus. In this paper, we propose a simple yet effective neural framework to derive the character-level embeddings for NER in Chinese text, named ME-CNER. A character embedding is derived with rich semantic information harnessed at multiple granularities, ranging from radical, character to word levels. The experimental results demonstrate that the proposed approach achieves a large performance improvement on Weibo dataset and comparable performance on MSRA news dataset with lower computational cost against the existing state-of-the-art alternatives.

CCS CONCEPTS

• Computing methodologies → Information extraction.

KEYWORDS

Named Entity Recognition; Multiple Embeddings; Chinese Computing; Convolutional Gated Recurrent Unit

1 INTRODUCTION

As a fundamental task in NLP, Named Entity Recognition (NER) has been studied extensively in the past decade. Formally, NER is a task that locates the mentions of named entities and classifies their types like person, organization, location in unstructured text. Nowadays, the sources of text are various. Except for formal text like news, user-generated content like micro-blogging has played a central role in people’s digital life. Performing NER over this real-time information task of high volume would breed many downstream semantic applications.

Compared with the English language, NER on Chinese text faces more challenges. It is known that Asian languages like Chinese are naturally logographic. There are no conventional linguistic features (like capitalization) available for NER. Moreover, Chinese characters and words hold complex relations. Given no delimiter between characters, a word can be comprised of a single character or multi-character (i.e., n-char). Also, the meaning of a Chinese word can be inferred by its constituent characters. For instance, the meaning of ‘自信’ (confidence) can be derived by ‘自’ (self) and ‘信’ (belief). Furthermore, Chinese characters often have hieroglyphic forms. The same radical often indicates similar semantics. Besides, to build a new large NER dataset reflecting new words and memes over Chinese text is often expensive and labor-intensive. The available dataset is often dated or limited in size, which calls for a low-cost mechanism to harness the background knowledge from large unlabeled resources.

To this end, in this paper, we propose a simple yet effective neural framework for NER in Chinese text, named ME-CNER. Specifically, we derive a character representation based on multiple embeddings in different granularities, ranging from radical, character to word levels. We utilize a convolutional operation to derive the character representation based on the local radical-level context information. Also, a Conv-GRU network is designed to capture semantic representation for a character based on its local context and long-range dependence. To harness the background knowledge, we choose to initialize the embeddings at character- and word-levels with pre-trained embeddings over a large external corpus, and fine-tune them in a task-driven fashion. As illustrated in Figure 1, these character representations in different perspectives are concatenated and fed into a BiGRU-CRF (Bidirectional Gated Recurrent Unit - Conditional
To summarize, we make the following contributions: (1) We propose a multiple-embedding-driven neural framework for Chinese NER. Various semantic information is well exploited in different granularities. (2) Our method significantly outperforms the previous state-of-the-art with a large margin on Weibo dataset and achieves comparable performance on MSRA News dataset with a lower computational cost. Either Conv-GRU or radical embedding is revealed to be beneficial for the task. Our code is available at https://github.com/WHUIR/ME-CNER.

3 RELATED WORK

Many Deep Learning techniques have been proposed for NER task. Huang et al. [6] proposed a BiLSTM (Bidirectional Long Short-Term Memory)-CRF (Conditional Random Field) tagger for sequence labeling. The experimental results over a standard benchmark Weibo (Chinese Microblog) dataset show that more than 7.6% relative performance improvement is obtained against the existing state-of-the-art solutions. Also, we provide results on a large and formal dataset (contrary to Weibo), MSRA News, to confirm the effectiveness of our proposed method. To summarize, we make the following contributions: (1) We propose a multiple-embedding-driven neural framework for Chinese NER. Various semantic information is well exploited in different granularities. (2) Our method significantly outperforms the previous state-of-the-art with a large margin on Weibo dataset and achieves comparable performance on MSRA News dataset with a lower computational cost. Either Conv-GRU or radical embedding is revealed to be beneficial for the task. Our code is available at https://github.com/WHUIR/ME-CNER.

Figure 1: A live example for the proposed architecture. “班长” is a nominal mention, while “潘伟柏” is a named entity.

Random Field) tagger for sequence labeling. The experimental results over a standard benchmark Weibo (Chinese Microblog) dataset show that more than 7.6% relative performance improvement is obtained against the existing state-of-the-art solutions. Also, we provide results on a large and formal dataset (contrary to Weibo), MSRA News, to confirm the effectiveness of our proposed method. To summarize, we make the following contributions: (1) We propose a multiple-embedding-driven neural framework for Chinese NER. Various semantic information is well exploited in different granularities. (2) Our method significantly outperforms the previous state-of-the-art with a large margin on Weibo dataset and achieves comparable performance on MSRA News dataset with a lower computational cost. Either Conv-GRU or radical embedding is revealed to be beneficial for the task. Our code is available at https://github.com/WHUIR/ME-CNER.

2 RELATED WORK

Many Deep Learning techniques have been proposed for NER task. Huang et al. [6] proposed a BiLSTM (Bidirectional Long Short-Term Memory)-CRF (Conditional Random Field) network as NER tagger. Lample et al. [7] used a BiLSTM to extract word representations over character-level embeddings. Ma and Hovy [10] used a CNN (Convolutional Neural Network)-BiLSTM-CRF network to utilize both word and character level representations.

Chinese NER is more challenging compared to English. Many sophisticated methods have been proposed in this line. Wang et al. [16] proposed a gated CNN model. E and Xiang [4] utilized character and word mixed embeddings. Zhang and Yang [18] proposed Lattice LSTM model. As to NER in social media, Peng and Dredze [11] used the positional character embedding and jointly trained it with word embedding. Cao et al. [1] used adversarial transfer learning. Peng and Dredze [12] utilized word segmentation neural network to facilitate NER. Sun and He [15] proposed an integrated method to train on both F-score and label accuracy. Note that either character and word level embeddings are exploited in these works. However, ME-CNER is the first work to jointly consider radical, character and word level embeddings for Chinese NER.

3 MODEL STRUCTURE AND MULTIPLE EMBEDDINGS

In this section, we introduce the multiple embeddings from finer to coarse level of granularity. The overall structure of ME-CNER is shown in Figure 2.

3.1 Radical Embedding

Chinese characters are hieroglyphic in nature. The same parts (i.e., radical) in different characters often share the same meaning [5]. The motivation for the use of radicals for NER is intuitive yet cultural. Chinese speakers have a strong inclination to use “good characters” when naming their businesses and children. For example, characters with radicals like “火” (metal) and “水” (water) never appear in a name of a person or a business. On the contrary, characters with radicals including “木” (wood), “火” (fire) and “水” (water) are very common in a person’s name to meet the Wu-Xing theory (a Chinese folk belief).

We use a CNN network to extract local context features of the radical sequence. The use of radicals also enables us to better infer the semantics of the characters that only appear in the test set but not in the training set. That is, we can teach the model to generalize well for rare characters correctly.

3.2 Character Embedding

Characters are the elementary units in Chinese. Note that each Chinese character has its own meaning and can compose a word itself. Also, the meaning of a Chinese word can be inferred by considering its constituent characters. Thus, it is critical to exploit the rich semantics contained by characters.

Previous studies use character-level CNN model for NER tasks [10, 16]. However, CNN emphasizes the local n-char features within a specific window and is not capable of capturing long-range dependence. To deal with this problem, we propose Convolutional Gated Recurrent Unit (Conv-GRU) Network which is composed of a GRU layer and a convolution layer. First, the character embedding $c_i$ is fed into the GRU layer.

$$x_i = GRU(c_1, ..., c_i)$$ (1)

Then the output $X = [x_1, ..., x_I]$ is fed into a convolutional layer padded to the same length as the input, where $I$ is the length of the microblog.

$$Y = Conv(X)$$ (2)

Finally, the output of the convolutional layer is concatenated with the output of the GRU layer to form the final representation for each character.

$$z_i = x_i \oplus y_i$$ (3)

In this way, we can combine the semantic knowledge from both local context and long-term dependency together.

3.3 Word Embedding

As a high-level representation, word embedding is the most common way to exploit semantics in NER. However, a Chinese word
we concatenate the character embeddings in radical, character and word levels to form the final character representation. Then, we conduct experiments on a formal text dataset, MSRA News dataset [8], which only includes named entities. The statistics of the two datasets we use is shown in Table 1.

Table 1: Statistics of the two datasets used in our experiments.

| Dataset | Type     | Train   | Valid  | Test   |
|---------|----------|---------|--------|--------|
| Weibo   | Sentence | 1.4k    | 0.27k  | 0.27k  |
|         | Character| 73.3k   | 14.5k  | 14.8k  |
| MSRA    | Sentence | 46.4k   | -      | 4.4k   |
|         | Character| 2169.9k | -      | 172.6k |

would be a multi-characters. As shown in Figure 1, to align the segmented words and the characters, we duplicate the word embedding for its constituent characters. For example, for the example demonstrated in Figure 1, both component character “长” (class) and “校” (president) are aligned with a shared word embedding of “校长” (class president). If a word is not in the vocabulary of the word embedding, we initialize its embedding with random values.

3.4 BiGRU-CRF Tagger

We concatenate the character embeddings in radical, character and word levels to form the final character representation. Then, we utilize the BiRNN-CRF [6] tagger to label each sentence. BiGRU-CRF is composed of forward and backward GRU layers and a CRF layer on top of the former. Here, we initialize the word embeddings and character embeddings with pre-trained embedding provided by Tencent AI Lab [14] (https://ai.tencent.com/labs/nlp/embedding.html).

4 EXPERIMENTS

In this section, we evaluate the proposed ME-CNER against the existing state-of-the-art models. Then we carry out ablation studies to verify each design choice in our model.

4.1 Experimental Settings

Dataset. We utilize a standard Weibo NER dataset provided by [11]. It consists of both named entities and nominal mentions. Also, we conduct experiments on a formal text dataset, MSRA News dataset [8], which only includes named entities. The statistics of the two datasets we use is shown in Table 1.

Following the related works [4, 11], we perform Chinese word segmentation with jieba toolkit (https://github.com/fxsjy/jieba). The character-level radicals are extracted according to a look-up table provided in [17].

Parameter Setting and Evaluation Metrics. We set the size of the kernel of the convolutional layers to 3 for both radical and character level embeddings. In the Conv-GRU framework, the dimension of GRU is set to 150. The embedding size is fixed to be 200. We add dropout with a keep probability of 0.8. We run each experiment five times, then report average precision, recall and F1 scores. For BiGRU tagger, the BIO scheme [13] is utilized.

4.2 Performance Comparison

Table 2 summarizes the results of different models. When evaluated on Weibo dataset, ME-CNER improves 22.13 on precision (Pr), 2.14 on recall (Re), and 10.14 in terms of F1, against the best baseline (i.e., Lattice [18]). We would like to attribute the improvement to the following enhancements. First, our use of multiple embeddings could encode semantic information towards NER from different perspectives. Second, the Conv-GRU networks work well in exploiting character-level information, which is extremely critical in Chinese NER task. Finally, the pre-trained embeddings are learned over large-scale news, crawled web pages and online novels, which provide further background knowledge. Note that all baselines also utilize the word and character embeddings pre-trained over a large external corpus. To examine this factor further, we train Lattice with Tencent embeddings. The results show that Tencent embeddings could further enhance the NER performance for Lattice. However, our proposed ME-CNER still outperforms it by 4.84 in terms of F1.

On MSRA news dataset, our model performs slightly poorer than Lattice but outperforms all other baselines. To analyze, firstly, on this much larger dataset (see Table 1) the relatively simple architecture of our proposed model may suffer from a lack of learning capability. Also, the advantage of pre-trained embeddings fades with more supervised data. Moreover, on a small and noisy dataset, radicals can help generalize (to be detailed in Section 4.3) but such advantage again vanishes in a formal and much larger dataset. However, we would like to highlight the lower computational cost (< 0.5x training time of Lattice) and better parallelism of our model brought by GRU and CNN, compared to Lattice, a complex modified LSTM network, which cannot implement batch training so far.

4.3 Ablation Study

We further conduct ablation study for ME-CNER by removing radical embeddings and replacing Conv-GRU network with either a CNN or a BiLSTM network. Table 2 demonstrates the performance comparison.

Radical Embedding. It is obvious that ME-CNER experiences performance degradation to some extent without using radical level embeddings. As mentioned previously, radicals would abstract the meaning of similar words. We analyze its effect for nominal mentions and named entities by showing actual cases.

For nominal mentions, the word consisting of a single character “妈 (mother)”, means “mother” in older ages, and is gradually replaced by another character “娘” (mother) in modern days. This character is especially rare in social media, including our Weibo
We would like to thank the anonymous reviewers for their great insights. This research was supported by National Natural Science Foundation of China (No. 61872278). Chenliang Li is the corresponding author.

| Method                  | Weibo Pr | Weibo Re | Weibo F1 | MSRA News Pr | MSRA News Re | MSRA News F1 |
|-------------------------|----------|----------|----------|--------------|--------------|--------------|
| CPM (Chen et al. [2])   | -        | -        | -        | 91.22        | 81.71        | 86.20        |
| Joint(ep) (Peng and Dredze [11]) | 63.84    | 29.45    | 40.38    | -            | -            | -            |
| JTE (Peng and Dredze [12]) | 61.64    | 38.55    | 47.43    | -            | -            | -            |
| CiLin (Lu et al. [9])   | -        | -        | -        | -            | -            | -            |
| BLSTM-CRF+pretrain (Dong et al. [3]) | -        | -        | -        | 91.28        | 90.62        | 90.95        |
| CWME (E and Xiang [4])  | 65.29    | 39.71    | 49.47    | -            | -            | -            |
| F-Score Driven (Sun and He [15]) | -        | -        | 54.82    | -            | -            | -            |
| ATL (Cao et al. [11])   | 55.72    | 50.68    | 53.08    | -            | -            | -            |
| Lattice (Zhang and Yang [18]) | 53.04    | 62.25    | 58.79    | 93.57        | 92.79        | 93.18        |
| Lattice with Tencent Embedding | 69.97    | 59.11    | 64.09    | 92.84        | 91.68        | 92.26        |
| ME-CNER (ours)          | 75.17    | 64.39    | 68.93    | **91.57**    | **91.33**    | **91.45**    |
| - radical               | 73.00    | 63.46    | 67.90    | 91.52        | 90.52        | 91.02        |
| - radical, Conv-GRU → CNN | 71.15    | 62.36    | 66.46    | 90.73        | 89.37        | 90.05        |
| - radical, Conv-GRU → BiLSTM | 71.12    | 60.69    | 65.78    | 91.43        | 90.04        | 90.73        |

Table 2: Experiment results and ablation study on Weibo NER and MSRA datasets. The best and second-best performance is indicated with black boldface and blue italics respectively. For the ablation study, we first remove radical embedding (marked as “-radical”), then replace Conv-GRU with CNN or BiLSTM (marked as “Conv-GRU→CNN” and “Conv-GRU→BiLSTM” respectively).

5 CONCLUSION

In this paper, we exploit multiple embeddings for Chinese NER. Also, we propose Conv-GRU to better exploit character embedding. Our proposed ME-CNER gains a prominent improvement against state-of-the-art alternatives when testing on a real-world microblog dataset and achieves comparable performance on a large formal news dataset with a lower computational cost. Our work shows the effectiveness of exploiting multiple embeddings together in different granularities.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their great insights. This research was supported by National Natural Science Foundation of China (No. 61872278). Chenliang Li is the corresponding author.