Using Chinese Glyphs for Named Entity Recognition (Student Abstract)

Chan Hee Song
University of Notre Dame
Notre Dame, Indiana, USA
csong1@nd.edu

Arijit Sehanobish
Yale University, School of Medicine
New Haven, Connecticut, USA
arijit.sehanobish@yale.edu

Abstract

Most Named Entity Recognition (NER) systems use additional features like part-of-speech (POS) tags, shallow parsing, gazetteers, etc. Adding these external features to NER systems have been shown to have a positive impact. However, creating gazetteers or taggers can take a lot of time and may require extensive data cleaning. In this work instead of using these traditional features we use lexicographic features of Chinese characters. Chinese characters are composed of graphical components called radicals and these components often have some semantic indicators. We propose CNN based models that incorporate this semantic information and use them for NER. Our models show an improvement over the baseline BERT-BiLSTM-CRF model. We present one of the first studies on Chinese OntoNotes v5.0 and show an improvement of $+.64$ F1 score over the baseline. We present a state-of-the-art (SOTA) F1 score of 71.81 on the Weibo dataset, show a competitive improvement of $+0.72$ over baseline on the ResumeNER dataset, and a SOTA F1 score of 96.49 on the MSRA dataset.

1 Introduction

Augmenting named entity recognition (NER) systems with additional features like gazetteers, bag of words, or character level information is very common and have shown to improve the NER systems. Chinese is a logographic language and many Chinese characters have evolved from pictures. We incorporate this semantic information from the pictures of the Chinese characters to use it as an added feature for Chinese NER systems. Only recently, Meng et al. (2019) presented a complex glyph (character image) reinforced model that concatenates the glyph embeddings with BERT (Devlin et al. 2018) embeddings. In this paper, we present a new kind of glyph-augmented architecture that requires less data, has fewer parameters, more robust. We show that our models have a significant improvement over our baseline on 4 datasets, Chinese OntoNotes v5.0 (Pradhan et al. 2013), Weibo (Peng and Dredze 2015), ResumeNER (Zhang and Yang 2018), and MSRA (Levow 2006).

We approach the problem of incorporating Chinese glyphs as an image classification problem and present two convolutional neural networks (CNNs) which we call “strided” and “GLYNN”. We treat this encoding problem purely in terms of computer vision, i.e. to extract “meaningful” features from the image, instead of a specialized CNN that encapsulates the subtle radicals. This way, our CNNs are agnostic to the subtleties of Chinese characters. Both CNNs are used to encode the glyphs and these encoded images are then used as an added feature for our NER system. We also use an autoencoder to pretrain GLYNN and compare the results. Since we treat our problem as an image classification problem, our models are easier to train and implement. Furthermore, unlike Meng et al. (2019) who used an extensive dataset of Chinese glyphs gathered from different time periods, we use only about 4500 grayscale 64 x 64 Chinese characters in Hei Ti font. Hei Ti font, similar to sans-serif, is widely used and is easily available online. These characters are all the Chinese characters found in Chinese BERT vocabulary. Even though there are over $20,000$ CJK characters, we only have about a hundred out-of-vocabulary (oov) characters in our datasets. This enables us to use less data to train our model than the other glyph-based approaches and thus have fewer parameters in our models. We achieve state-of-the-art F1 score on the Weibo and MSRA dataset and report one of the first results of the OntoNotes v5.0 to the best of the authors' knowledge. Our code can be found in https://github.com/arijitthegame/BERT-CNN-models. The main strength of our model is as follows:

- Requires less amount of glyph data to train
- Fewer parameters than similar glyph models
- Robust to non-Chinese languages in the dataset

2 Architecture of our Glyph Models

Our model consists of the following parts: pre-trained BERT embeddings and the (pretrained) CNN embeddings. Chinese character in the dataset and in our vocabulary are matched with the corresponding image and is feed into the CNN. If the character is non Chinese (or Chinese and oov), we feed in a 64 x 64 array of 0s (or 1's). We concatenate the last four layers of BERT and the CNN vectors which are our new “character” embeddings. We then feed these

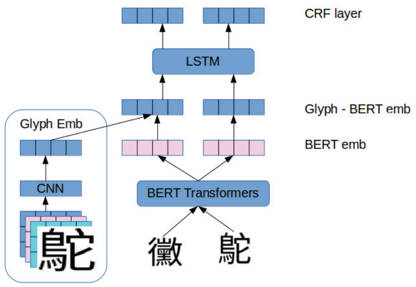

Figure 1: Architecture of our Glyph Models

character embeddings to the BiLSTM layer which are decoded via a CRF layer. The CNN-LSTM-CRF is then trained end-to-end while we keep BERT frozen. Thus we take advantage of BERT's large pre-scale training and the information from the glyphs encoded by our CNN's. We present 2 CNN models and 1 autoencoder for the CNN.

Strided CNN: Consists of 4x 2D Convolution layers with strides 2, filter size of 64, kernel size of 3 and activation leaky ReLU followed by a Flatten and a Dense layer of output dimension 64. Furthermore we normalize the final output by using Layer Normalization.

GLYNN CNN: Consists of 2 sets of Convolution, Batch Normalization and Maxpooling layers. We add the Dropout layers to prevent overfitting. Then we have a Flatten and a Batch Normalization layer, followed by a Dense layer of output dimension 256.

Autoencoder: Consists of a sequence of N encoder layers followed by N decoder layers used to pretrain GLYNN.

3 Experiments and Results

Models	OntoNotes v5.0	Weibo	Weibo NAM	ResumeNER	MSRA
BERT-BiLSTM-CRF (baseline)	79.09	70.79	71.80	95.72	95.3
GLYNN	79.47	**71.81**	**73.90**	96.44	**96.49**
Strided	**79.73**	70.70	73.35	96.38	95.63
Meng et al. (2019)	N/A	67.60	N/A	**96.54**	95.54
Zhang and Yang (2018)	N/A	58.79	53.04	94.46	93.18

Table 1: Experimental Results

Dataset: We used Chinese OntoNotes v5.0 dataset composed of 18 different tag sets compiled for CoNLL-2013 shared task and follow the standard train/dev/test split. We chose OntoNotes v5.0 because it includes all previous versions of OntoNotes and contains an extra genre; telephone communications, which makes the dataset more representative of the real world. We also use Weibo (both Named and Nominal Entity Mentions), ResumeNER, and MSRA dataset which has 4, 8 and 3 entity types respectively. We use the train, dev and test set for training, validation and testing respectively. All the results are from the test set.

Experimental Results: Table 1 shows our experimental results with these 4 datasets. We ran our models 20, 40, 15, 12 times on OntoNotes, Weibo, ResumeNER and MSRA respectively and we reported the best scores on the test sets. We have found that our results are a significant improvement over the baseline by performing a 2-sample t-test. We present a new baseline score for OntoNotes v5.0 and show a $+0.64$ improvement by our glyph models. We achieve state-of-the-art F1 score of 71.81 and 96.49 for the full Weibo and the MSRA dataset respectively. We present a competitive score on ResumeNER dataset, only falling 0.1 F1 score on the current SOTA. We didn't employ a full grid-search of the hyperparameters, so we hope a full grid-search could overcome that gap. Since some datasets contain a fair amount of non Chinese characters ($\frac{1}{9}$ of OntoNotes and $\frac{1}{6}$ of Weibo are not Chinese), we experimented with feeding in the image of non Chinese characters to the CNN instead of a blank image. In our experiments, we found that those results are not a significant improvement over the baseline. We believe the reason that feeding in the blank image showed better performance is due to the fact that the non Chinese characters do not have any semantic meaning.

4 Conclusion and Future Work

Using two very different CNNs with and without an autoencoder, we have shown gains over the baseline system on the 4 most commonly used Chinese NER datasets and have achieved state-of-the-art F1 score on the Weibo and MSRA dataset. The novelty of our approach lies in 3 salient features: a) very little data to augment our system, b) fewer parameters than similar glyph models, and c) robust.

We are excited by the future of glyphs in NLP and would like to use them for other NLU tasks. Finally we would like to thank Johns Hopkins University and all the mentors in the SCALE workshop for their hospitality and for facilitating an excellent atmosphere for conducting this research.

References

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805. preprint.

Levow, G.-A. 2006. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 108–117.

Meng, Y.; Wu, W.; Wang, F.; Li, X.; Nie, P.; Yin, F.; Li, M.; Han, Q.; Sun, X.; and Li, J. 2019. Glyce: Glyph-vectors for chinese character representations. http://arxiv.org/abs/1901.10125v3. preprint.

Peng, N., and Dredze, M. 2015. Named entity recognition for Chinese social media with jointly trained embeddings. In *EMNLP*, 548–554.

Pradhan, S.; Moschitti, A.; Xue, N.; Ng, H. T.; Bjorkelund, A.; Uryupina, O.; Zhang, Y.; and Zhong, Z. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on CoNLL.*

Zhang, Y., and Yang, J. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers)*, 1554–1564.