MONPA: Multi-objective Named-entity and Part-of-speech Annotator for Chinese using Recurrent Neural Network

Yu-Lun Hsieh
SNHCC, TIGP, Academia Sinica, and
National Cheng Chi University, Taiwan
morphe@iis.sinica.edu.tw

Yung-Chun Chang
Graduate Institute of Data Science,
Taipei Medical University, Taiwan
changyc@tmu.edu.tw

Yi-Jie Huang, Shu-Hao Yeh, Chun-Hung Chen, and **Wen-Lian Hsu**
IIS, Academia Sinica, Taiwan
{aszx4510,night,hep_chen,hsu}@iis.sinica.edu.tw

Abstract

Part-of-speech (POS) tagging and named entity recognition (NER) are crucial steps in natural language processing. In addition, the difficulty of word segmentation places extra burden on those who deal with languages such as Chinese, and pipelined systems often suffer from error propagation. This work proposes an end-to-end model using character-based recurrent neural network (RNN) to jointly accomplish segmentation, POS tagging and NER of a Chinese sentence. Experiments on previous word segmentation and NER competition datasets show that a single joint model using the proposed architecture is comparable to those trained specifically for each task, and outperforms freely-available softwares. Moreover, we provide a web-based interface for the public to easily access this resource.

1 Introduction

Natural language processing (NLP) tasks often rely on accurate part-of-speech (POS) labels and named entity recognition (NER). Moreover, for languages that do not have an obvious word boundary such as Chinese and Japanese, segmentation is another major issue. Approaches that attempt to jointly resolve two of these tasks have received much attention in recent years. For example, Ferraro et al. (2013) proposed that joint solutions usually lead to the improvement in accuracy over pipelined systems by exploiting POS information to assist word segmentation and avoiding error propagation. Recent researches (Sun, 2011; Qian and Liu, 2012; Zheng et al., 2013; Zeng et al., 2013; Qian et al., 2015) also focus on the development of a joint model to perform Chinese word segmentation, POS tagging, and/or informal word detection.

However, to the best of our knowledge, no existing system can perform word segmentation, POS tagging, and NER simultaneously. In addition, even though there are methods that achieved high performances in previous competitions hosted by the Special Interest Group on Chinese Language Processing (SIGHAN)[1], there is no off-the-shelf NLP tools for Traditional Chinese NER but only two systems for word segmentation and POS tagging, which poses a significant obstacle for processing text in Traditional Chinese. These problems motivate us to devise a unified model that serves as a steppingstone for future Chinese NLP research.

In light of the recent success in applying neural networks to NLP tasks (Sutskever et al., 2014; Lample et al., 2016), we propose an end-to-end model that utilizes bidirectional RNNs to jointly perform segmentation, POS tagging, and NER in Chinese. This work makes the following major contributions. First, the proposed model conducts multi-objective annotation that not only handles word segmentation and POS tagging, but also can recognize named entities in a sentence simultaneously. We also show that these tasks can be effectively performed by the proposed model, achieving competitive performances to state-of-the-art methods on word segmentation and NE recognition of previous SIGHAN shared tasks. Moreover, our system not only outperforms off-the-shelf NLP tools, but also provides accurate NER results. Lastly, we provide an accessible online API[2] that has been utilized by several research groups.

[1] http://sighan.cs.uchicago.edu/

[2] Please visit http://monpa.iis.sinica.edu.tw:9000/chunk

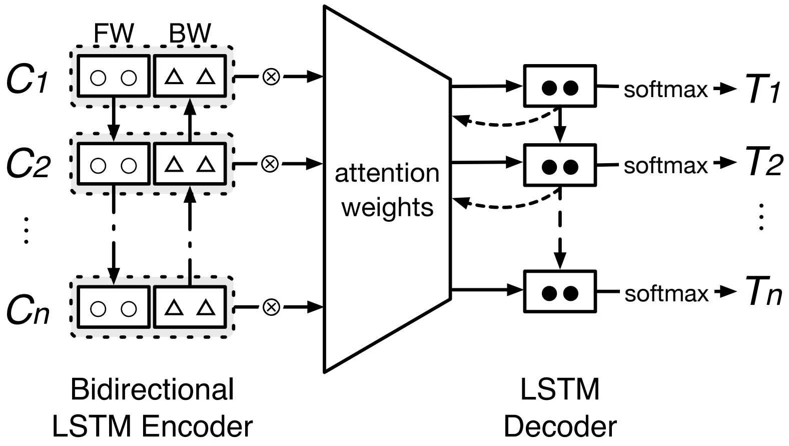

Figure 1: Overview of the encoder-decoder model with attention mechanism. Character embeddings C_1 to C_n of the input sentence is sequentially fed into the bidirectional LSTM, and the concatenated output is multiplied by attention weights and sent to the decoder for predicting the tag sequence T_1 to T_n. For simplicity, multiple layers of encoder and decoder as well as dropout layers between them are omitted.

2 Methods

Figure 1 illustrates the overview of our model, which in essence is an encoder-decoder (Sutskever et al., 2014) with attention mechanism (Luong et al., 2015). The input is a sequence of Chinese characters that may contain named entities, and the output is a sequence of POS tags and possibly NEs in the form of BIES tags. Our model mainly consists of: embedding layer, recurrent encoder layers, attention layer, and decoder layers. Detailed description of these layers are as follows.

Embedding Layer converts characters into embeddings (Mikolov et al., 2013), which are dense, low-dimensional, and real-valued vectors. They capture syntactic and semantic information provided by its neighboring characters. In this work, we utilize pre-trained embeddings using `word2vec` and over 1 million online news articles. **Recurrent Encoder Layers** use LSTM, or Long Short-Term Memory (Hochreiter and Schmidhuber, 1997), cells which have been shown to capture long-term dependencies (Greff et al., 2017). An LSTM cell contains a "memory" cell $\mathbf{c}_t$ and three "gates", *i.e.*, input, forget, and output. The input gate modulates the current input and previous output. The forget gate tunes the content from previous memory to the current. Finally, the output gate regulates the output from the memory. Specifically, let $\mathbf{x}_t$ be the input at time t, and $\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t$ correspond to input, forget, and output gates, respectively. $\mathbf{c}_t$ denotes the memory cell and $\mathbf{h}_t$ is the output. The learnable parameters include $W_{i,f,o,c}$ and $U_{i,f,o,c}$. They are defined as:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(W_i\mathbf{x}_t + U_i\mathbf{h}_{t-1}) \\
\mathbf{f}_t &= \sigma(W_f\mathbf{x}_t + U_f\mathbf{h}_{t-1}) \\
\mathbf{o}_t &= \sigma(W_o\mathbf{x}_t + U_o\mathbf{h}_{t-1}) \\
\tilde{\mathbf{c}}_t &= \tanh(W_c\mathbf{x}_t + U_c\mathbf{h}_{t-1}) \\
\mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \mathbf{i}_t \circ \tilde{\mathbf{c}}_t \\
\mathbf{h}_t &= \mathbf{o}_t \circ \tanh(\mathbf{c}_t)
\end{aligned}$$

where "$\circ$" denotes the element-wise product of vectors and σ represents the sigmoid function. We employ a straightforward extension named Bidirectional RNN (Graves et al., 2005), which encodes sequential information in both directions (forward and backward) and concatenate the final outputs. In this way, the output of one time step will contain information from its left and right neighbors. For tasks such as POS and NER where the label of one character can be determined by its context, bidirectional learning can be beneficial. **Attention Layer** is proposed by Luong et al. (2015) in an attempt to tackle the problem of finding corresponding words in the source and target languages when conducting machine translation. It computes a weighted average of all the output from the encoder based on the current decoded symbol, which is why it is also named "Global Attention." We consider it to be useful for the current tasks based on the same reasoning as using bidirectional encoding. Finally, **Recurrent Decoder Layers** take the sequence of output from the attention layer and project them onto a V-dimensional vector where V equals the number of possible POS and NE tags. The loss of the model is defined as the averaged cross-entropy between a output sequence and true label sequence.

3 Experiments

Test corpora from five previous SIGHAN shared tasks, which have been widely adopted for Traditional Chinese word segmentation and NER, were used to evaluate the proposed system. Besides the participating systems in the above shared tasks, we also compare with existing word segmentation toolkits Jieba and CKIP (Hsieh et al., 2012). The word segmentation datasets were taken from SIGHAN shared tasks of years 2003–2008, and NER dataset is from 2006. We follow the standard train/test split of the provided data, where 10,000 sentences of the training set are used as the validation set. Details of the word segmentation and NER datasets are shown in Table 1 and 2, respectively. Three metrics are used for evaluation: precision (P), recall (R) and F_1-score (F), defined by

$$F = \frac{2 \times P \times R}{P + R}$$

For word segmentation, a token is considered to be correct if both the left and right boundaries match those of a word in the gold standard. For the NER task, both the boundaries and the NE type must be correctly identified.

Table 1: Statistics of the word segmentation datasets (Number of words).

Year	AS		CityU	
	#Train	#Test	#Train	#Test
2003	5.8M	12K	240K	35K
2005	5.45M	122K	1.46M	41K
2006	5.5M	91K	1.6M	220K
2008	1.5M	91K	-	-

Table 2: Statistics of the 2006 NER dataset (Number of words).

#Train/Test Words		
Person	Location	Organization
36K / 8K	48K / 7K	28K / 4K

3.1 Experimental Setup

In order to obtain multi-objective labels of the training data, we first merge datasets from the 2006 SIGHAN word segmentation and NER shared tasks. Since rich context information is able to benefit deep learning-based approach, we augment the training set by collecting online news articles[3]. There are three steps for annotating the newly-created dataset. We first collect a list of NEs from Wikipedia and use it to search for NEs in the corpus, where longer NEs have higher priorities. Then, an NER tool (Wu et al., 2006) is utilized to label NEs. Finally, CKIP is utilized to segment and label the remaining words with POS tags. Three variants of the proposed model are tested, labeled as RNN_{CU06}, RNN_{YA}, and $\text{RNN}_{\text{CU06+YA}}$. RNN_{CU06} is trained using only word segmentation and NER datasets from the 2006 City University (CU) corpus; RNN_{YA} is trained using only online news corpus, and $\text{RNN}_{\text{CU06+YA}}$ is trained on a combination of the above corpora.

We implemented the RNN model using `pytorch`[4]. The maximum sentence length is set to 80, where longer sentences were truncated and shorter sentences were padded with zeros. The forward and backward RNN each has a dimension of 300, identical to that of word embeddings. There are three layers for both encoder and decoder. Dropout layers exist between each of the recurrent layers. The training lasts for at most 100 epochs or when the accuracy of the validation set starts to drop.

4 Results and Discussion

Note that since we combined external resources, performances of the compared methods are from the open track of the shared tasks. Table 3a lists the results of the RNN-based models and top-performing systems for the word segmentation subtask on the Academia Sinica (AS) dataset. First of all, RNNs exhibit consistent capabilities in handling data from different years and is comparable to the best systems in the competition. In addition, it is not surprising that the RNN_{YA} model perform better than RNN_{CU}. Nevertheless, our method can be further improved by integrating the CU06 corpus, demonstrated by the results from

[3] News articles are collected from the Yahoo News website and contains about 3M words.

[4] https://github.com/pytorch/pytorch

Table 3: Results for word segmentation on the Academia Sinica (AS) and City University (CU) datasets from different years of SIGHAN shared tasks. Bold numbers indicate the best performance in that column.

(a) AS dataset, open track

System	F-score			
	2003	2005	2006	2008
Gao et al. (2005)	95.8			
Yang et al. (2003)	90.4			
Low et al. (2005)		**95.6**		
Chen et al. (2005)		94.8		
Zhao et al. (2006)			**95.9**	
Jacobs and Wong (2006)			95.7	
Wang et al. (2006)			95.3	
Chan and Chong (2008)				**95.6**
Mao et al. (2008)				93.6
Jieba	83.0	80.9	81.3	81.8
CKIP	**96.6**	94.2	94.6	94.9
$\mathrm{RNN}_{\mathrm{CU06}}$	88.4	86.8	87.1	87.4
$\mathrm{RNN}_{\mathrm{YA}}$	94.4	92.8	93.0	93.3
$\mathrm{RNN}_{\mathrm{CU06+YA}}$	*94.6*	*93.2*	*93.6*	*93.8*

(b) CU dataset, open track

System	F-score		
	2003	2005	2006
Ma and Chen (2003)	**95.6**		
Gao et al. (2005)	95.4		
Peng et al. (2004)	94.6		
Yang et al. (2003)	87.9		
Low et al. (2005)		**96.2**	
Chen et al. (2005)		94.5	
Zhao et al. (2006)			**97.7**
Wang et al. (2006)			97.7
Jacobs and Wong (2006)			97.4
Jieba	80.3	81.2	82.4
CKIP	89.7	89.0	89.8
$\mathrm{RNN}_{\mathrm{CU06}}$	87.6	85.8	87.8
$\mathrm{RNN}_{\mathrm{YA}}$	88.0	87.2	88.5
$\mathrm{RNN}_{\mathrm{CU06+YA}}$	*91.5*	*90.1*	*91.7*

the $\mathrm{RNN}_{\mathrm{CU06+YA}}$ model. This indicates that RNN can easily adapt to different domains with data augmentation, which is an outstanding feature of end-to-end models. As for the CU dataset listed in Table 3b, all of the RNN models show considerable decrease in F-score. We postulate that it may be due to the training data, which is processed using an external tool focused on texts from a different linguistic context. It is also reported by (Wu et al., 2006) that segmentation criteria in AS and CU datasets are not very consistent. However, by fusing two corpora, the $\mathrm{RNN}_{\mathrm{CU06+YA}}$ can even surpass the performances of CKIP. Finally, comparison with Jieba validates that the RNN model can serve as a very effective toolkit for NLP researchers as well as the general public.

Table 4 lists the performances of proposed models and the only system that participated in the open track of the 2006 SIGHAN NER shared task. We can see that $\mathrm{RNN}_{\mathrm{CU06}}$ outperforms the model from Yu et al. (2006), confirming RNN's capability on jointly learning to segment and recognize NEs. Interestingly, $\mathrm{RNN}_{\mathrm{YA}}$ obtains a much lower F-score for all NE types. And $\mathrm{RNN}_{\mathrm{CU06+YA}}$ can only obtain a slightly better F-score for person recognition but not the overall performance of $\mathrm{RNN}_{\mathrm{CU06}}$, even with the combined corpus. We believe that boundary mismatch may be a major contributing factor here. We also observe that there are a large number of one-character NEs such as abbreviated country names, which can not be easily identified using solely character features.

Table 4: Results from the 2006 SIGHAN NER shared task (open track). Bold numbers indicate the best performance in that column.

System	F-score			
	PER	LOC	ORG	Overall
Yu et al. (2006)	80.98	86.04	68.01	80.51
$\mathrm{RNN}_{\mathrm{CU06}}$	81.13	**86.92**	**68.77**	**80.68**
$\mathrm{RNN}_{\mathrm{YA}}$	70.54	67.80	31.35	52.62
$\mathrm{RNN}_{\mathrm{CU06+YA}}$	**83.01**	82.46	54.57	75.28

5 Conclusions

We propose an end-to-end model to jointly conduct segmentation, POS and NE labeling in Chinese. Experimental results on past word segmentation and NER datasets show that the proposed model is comparable to those trained specifically for each task, and outperforms freely-available toolkits. Additionally, we implement a web interface for easy access. In the future, we will integrate existing knowledge bases, in order to provide a more advanced tool for the NLP community.

Acknowledgments

We are grateful for the constructive comments from three anonymous reviewers. This work was supported by grant MOST106-3114-E-001-002 from the Ministry of Science and Technology, Taiwan.

References

Samuel WK Chan and Mickey WC Chong. 2008. An agent-based approach to Chinese word segmentation. In *IJCNLP*. pages 112–114.

Aitao Chen, Yiping Zhou, Anne Zhang, and Gordon Sun. 2005. Unigram language model for Chinese word segmentation. In *Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics Jeju Island, Korea, pages 138–141.

Jeffrey P Ferraro, Hal Daumé III, Scott L DuVall, Wendy W Chapman, Henk Harkema, and Peter J Haug. 2013. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association* 20(5):931–939.

Jianfeng Gao, Mu Li, Andi Wu, and Chang-Ning Huang. 2005. Chinese word segmentation and named entity recognition: A pragmatic approach. *Computational Linguistics* 31(4):531–574.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. *Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005* pages 753–753.

Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2017. LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems* PP(99):1–11.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*. pages 473–479.

Yu-Ming Hsieh, Ming-Hong Bai, Jason S Chang, and Keh-Jiann Chen. 2012. Improving PCFG chinese parsing with context-dependent probability reestimation. *CLP 2012* page 216.

Aaron J Jacobs and Yuk Wah Wong. 2006. Maximum entropy word segmentation of Chinese text. In *COLING* ACL 2006*. page 185.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*. pages 260–270.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to Chinese word segmentation. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*. volume 1612164, pages 448–455.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1412–1421. http://aclweb.org/anthology/D15-1166.

Wei-Yun Ma and Keh-Jiann Chen. 2003. Introduction to ckip Chinese word segmentation system for the first international Chinese word segmentation bakeoff. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*. Association for Computational Linguistics, pages 168–171.

Xinnian Mao, Yuan Dong, Saike He, Sencheng Bao, and Haila Wang. 2008. Chinese word segmentation and named entity recognition based on conditional random fields. In *IJCNLP*. pages 90–93.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, page 562.

Tao Qian, Yue Zhang, Meishan Zhang, Yafeng Ren, and Donghong Ji. 2015. A transition-based model for joint segmentation, pos-tagging and normalization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1837–1846. http://aclweb.org/anthology/D15-1211.

Xian Qian and Yang Liu. 2012. Joint Chinese word segmentation, pos tagging and parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, pages 501–511. http://www.aclweb.org/anthology/D12-1046.

Weiwei Sun. 2011. A stacked sub-word model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 1385–1394. http://www.aclweb.org/anthology/P11-1139.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Xinhao Wang, Xiaojun Lin, Dianhai Yu, Hao Tian, and Xihong Wu. 2006. Chinese word segmentation with maximum entropy and n-gram language model. In *COLING* ACL 2006*. page 138.

Chia-Wei Wu, Shyh-Yi Jan, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2006. On using ensemble methods for Chinese named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Association for Computational Linguistics, Sydney, Australia, pages 142–145. http://www.aclweb.org/anthology/W/W06/W06-0122.

Jin Yang, Jean Senellart, and Remi Zajac. 2003. Systran's Chinese word segmentation. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*. Association for Computational Linguistics, pages 180–183.

Xiaofeng Yu, Marine Carpuat, and Dekai Wu. 2006. Boosting for Chinese named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. pages 150–153.

Xiaodong Zeng, Derek F. Wong, Lidia S. Chao, and Isabel Trancoso. 2013. Graph-based semi-supervised model for joint chinese word segmentation and part-of-speech tagging. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 770–779. http://www.aclweb.org/anthology/P13-1076.

Hai Zhao, Chang-Ning Huang, Mu Li, et al. 2006. An improved Chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*. Sydney: July, volume 1082117.

Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. 2013. Deep learning for Chinese word segmentation and POS tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, pages 647–657. http://www.aclweb.org/anthology/D13-1061.