Chinese Organization Name Recognition Based on Deep Learning: A Character-Level Recurrent Neural Network Method (Postprint)

Authors: Zhu Danhao, Yang Lei, Wang Dongbo

Date: 2017-11-08T00:00:00+00:00

Abstract

Abstract

Purpose: Chinese institution names feature complex structures and numerous rare words, making their recognition challenging. Accurate recognition is crucial for subsequent informatics tasks such as information extraction, information retrieval, knowledge mining, and institutional scientific research evaluation.

Method: Based on deep learning-based Recurrent Neural Network (RNN) methods, and considering the characteristics of Chinese characters and words, we redefine the input and output for institution name tagging and propose a character-level recurrent network tagging model.

Results: Using the word-level recurrent neural network method as a baseline, the character-level model proposed in this paper demonstrates significant improvements in precision, recall, and F-value for Chinese institution name recognition, with the F-value increasing by 1.54%. The improvement is more substantial when rare words are present, with the F-value increasing by 11.05%.

Limitations: The decoding process directly employs a greedy strategy, which is prone to falling into local optima. Modeling with Conditional Random Field algorithms could potentially yield globally optimal results.

Conclusion: The proposed method features a simple architecture and can leverage character-level features for modeling, achieving superior results compared to using only word-level features.

Full Text

Recognizing Chinese Organization Names Based on Deep Learning: A Character-Level Recurrent Neural Network Approach

Zhu Danhao[1],[2], **Yang Lei**[3], **Wang Dongbo**[4]
[1](Library of Jiangsu Police Institute, Nanjing 210031, China)
[2](Department of Computer Science and Technology, Nanjing University, Nanjing 210093, China)
[3](Department of Higher Education, Nanjing Traffic Technician College, Nanjing 210049, China)
[4](College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095, China)

Abstract

[**Objective**] Chinese organization names exhibit complex structures and contain numerous rare words, making their automatic recognition particularly challenging. Accurate identification of these names is crucial for subsequent information science tasks including information extraction, information retrieval, knowledge mining, and institutional research evaluation. [**Methods**] Based on deep learning recurrent neural network (RNN) methods and considering the characteristics of Chinese characters and words, we redefined the input and output for organization name tagging and proposed a character-level recurrent network tagging model. [**Results**] Using the word-level recurrent neural network method as a baseline, the proposed character-level model achieved significant improvements in precision, recall, and F-value for Chinese organization name recognition, with the F-value increasing by 1.54%. The improvement was even more pronounced for names containing rare words, where the F-value increased by 11.05%. [**Limitations**] The decoding process directly employs a greedy strategy, which is prone to falling into local optima; using conditional random fields for modeling might yield globally optimal results. [**Conclusions**] The proposed method features a simple architecture that leverages character-level features for modeling, achieving better results than approaches using only word-level features.

Keywords: Organization name recognition; Recurrent neural network; Deep learning

---

1 Introduction

Organizations generally refer to agencies, groups, or other enterprises and institutions, including universities, public and private companies, government departments, religious organizations, research institutions, international organizations, sports teams, music groups, military units, and more [1]. The effectiveness of organization name recognition significantly impacts subsequent tasks such as information extraction, information retrieval, knowledge mining, and institutional research evaluation. However, Chinese organization names contain many rare words and exhibit complex structures with substantial variation across different institutions, posing considerable challenges for accurate recognition.

Chinese organization name recognition can be framed as a sequence labeling problem. Models based on manual feature templates have been the primary approach for such tasks, employing algorithms including Conditional Random Fields [2], Support Vector Machines [3], and Maximum Entropy models [4]. These methods manually design discriminative feature templates targeting internal and external characteristics of Chinese organization names, then apply a powerful sequence labeling model to achieve good recognition performance. However, such approaches rely heavily on expert domain knowledge and are difficult to transfer and generalize across different types of corpora. In recent years, deep learning strategies based on recurrent neural networks have achieved considerable success in English sequence labeling tasks, including part-of-speech tagging, Chinese word segmentation, chunking, named entity recognition, and semantic role labeling [5-6]. Recurrent neural networks do not require specially crafted manual rules and can automatically learn features from distributed word vectors for tagging purposes, gradually becoming a research hotspot.

The primary input to recurrent neural networks is word vectors, whose quality directly determines system performance. For rare words, the model cannot obtain sufficient contextual information, resulting in poor word vector quality. Some studies have used complex rules to extract information from Chinese characters to enrich word vectors. For instance, Chen et al. used each character within a word to enhance Chinese word vectors; to address character ambiguity, they first clustered characters and used different character vectors for characters in different clusters [7]. Sun et al. used radicals to enhance Chinese word vectors, achieving improvements in character similarity comparison and Chinese word segmentation tasks [8].

However, simple architecture and easy generalization are the main advantages of recurrent neural networks. Although these complex word vector enhancement methods can alleviate the problem of sparse word vector information to some extent, their complex rules and difficult implementation weaken the advantages of RNNs. To address these issues, this paper proposes a completely character-based Chinese organization name recognition method that redefines the model' s input and output. The input consists of Chinese characters and spaces, while the output uses a new set of organization name tags. This method features a simple structure, is easy to implement, and requires no manual rules or external resources.

The main contributions of this paper are twofold: First, it applies recurrent neural networks to Chinese organization name recognition tasks, verifying the effectiveness of deep learning for this task; Second, it improves the tagging model according to the characteristics of Chinese characters and words, achieving better tagging performance.

2 Related Work

As a classic sequence labeling task, organization name recognition has long been a focus of information science research. In recent years, deep learning methods centered on recurrent neural networks have made new progress in sequence recognition. This section reviews relevant research from two directions: named entity recognition and recurrent neural networks.

2.1 Named Entity Recognition Research

Named entity recognition strategies have primarily revolved around rule-based and statistical methods, with statistical approaches being dominant. Representative methods include: Sun and Wang provided a systematic and detailed discussion of named entity research from technical and evaluation perspectives [9]; Pan Zhenggao proposed a probability-based named entity recognition strategy built upon internal and external rules [10]; Lu et al. completed product named entity recognition based on Conditional Random Field models [11]; Wu et al. presented a translation-weighted named entity strategy from the perspective of cross-language information retrieval [12]; Wang and Wang constructed a multi-feature knowledge-based named entity recognition model for various entities in project proposals by statistical analysis of features [13]; Chen et al. combined part-of-speech and external semantic knowledge from HowNet with Conditional Random Fields to automatically recognize theoretical entities in academic journals [14]; Yu et al. proposed a Chinese organization name automatic recognition method based on role tagging, using the Viterbi algorithm to assign roles to segmentation results according to their function in organization name recognition, then performing string recognition on the role sequence to ultimately achieve Chinese organization name recognition [15]; Guan et al. proposed a method for automatically constructing training corpora for organization names from user query logs to address the scarcity of such resources [16]; using the concept of cohesion to address information asymmetry and combining contextual information, they adopted Conditional Random Field models for organization name recognition. While statistical methods can effectively identify organization names across different corpora, they depend on rules and feature templates designed by experts for specific corpora, making them complex and difficult to transfer.

2.2 Recurrent Neural Networks

Recurrent neural networks have demonstrated strong tagging capabilities in many English sequence labeling tasks. Within the RNN framework, using Long Short-Term Memory (LSTM) modules instead of basic TANH modules yields better results. Huang et al. used bidirectional LSTM for sequence labeling and applied Conditional Random Fields (CRF) in the output layer for decoding, verifying on multiple datasets for part-of-speech tagging, chunking analysis, and named entity recognition tasks that the method achieved best performance when combined with manual rules and pre-trained word vectors [17]. Ma and Hovy used a bidirectional LSTM-CNNs-CRF model for end-to-end sequence labeling, employing Convolutional Neural Networks (CNNs) to learn character-level vectors for each word, then concatenating character-level and word vectors into an enhanced vector input to the bidirectional LSTM model, finally using CRF for decoding, validating the method on English part-of-speech tagging and named entity recognition tasks [18]. While English research has begun exploring the addition of character information in RNNs, similar studies are lacking for Chinese. English and Chinese differ significantly in characters and words; this paper designs a new algorithm to utilize Chinese character information based on these characteristics.

3 Methodology

3.1 System Framework

[Figure 1: see original paper] illustrates the tagging system framework, which consists of four layers. The bottom layer is the input layer, where the original model inputs words while our proposed character model inputs characters and word segmentation markers. The second layer is the vector mapping layer, converting the first layer' s input into corresponding distributed representation vectors. The third layer is the recurrent neural network layer, shown as a two-layer LSTM network in the figure. The topmost layer is the output layer, where RNN results are converted to output tags.

This paper improves the first input layer and the top output layer. For brevity, we directly use "LSTM" to refer to "LSTM recurrent network," as LSTM nodes can currently only be applied within recurrent networks.

3.2 Recurrent Neural Network Models

Recurrent Neural Network (RNN) is a neural network model particularly suitable for sequence labeling. In RNNs, at time step t, an input vector x_t is combined with the previous hidden state vector h_{t-1} to generate the current hidden state, as shown in equation (1):

$$h_t = f(Wx_t + Uh_{t-1} + b)$$

where $W \in \mathbb{R}^{n\times m}$, $U \in \mathbb{R}^{n\times n}$, and $b \in \mathbb{R}^n$ are coefficient matrices in the model, and f is the activation function. Finally, a Softmax layer can be added above the hidden state layer for classification tasks. Therefore, RNN can be understood as taking x as input and producing h as output.

Theoretically, RNNs can retain long-distance memory, but in practice, due to vanishing and exploding gradient phenomena, original RNN models struggle to achieve this. Hochreiter et al. and Sutskever et al. improved upon the original RNN by proposing Long Short-Term Memory (LSTM) modules, which solve the long-distance memory problem by adding memory cells and control gates to RNNs [19-20]. A standard LSTM module [20] operates as follows:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$

$$g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

$$h_t = o_t \odot \tanh(c_t)$$

At step t, the current hidden state h_t and memory c_t are computed from the previous memory cell c_{t-1}, previous hidden state h_{t-1}, and current input x_t. $\sigma(\cdot)$ and $\tanh(\cdot)$ are the Sigmoid and tangent functions, respectively. i_t, f_t, o_t, and g_t serve as control gates using the previous state and current input to regulate model input/output and memory transfer/preservation. Since memory cell transfer uses addition operations, the vanishing and exploding gradient phenomena caused by matrix multiplication are resolved during backpropagation.

In LSTM networks, stacking multiple hidden layers where lower layer outputs serve as higher layer inputs forms Deep LSTM. Simple LSTM networks compute left-to-right sequentially; computing hidden states simultaneously from right-to-left is called Bidirectional LSTM. Unless otherwise specified, all LSTM-based sequence labeling methods below use Bidirectional LSTM.

3.3 Word-Based Organization Name Tagging Model

Using LSTM for word-based organization name tagging is straightforward. [Figure 2: see original paper] provides a tagging example. The bottom layer is the input layer where each word belongs to a finite vocabulary set V. The dashed box contains the vector mapping layer and LSTM network. The top layer is the output layer producing corresponding tags from a finite tag set S. This paper uses a three-tag set $\{B\text{-ORG}, I\text{-ORG}, S\}$, where B-ORG indicates the first word of an organization name, I-ORG indicates subsequent words, and S indicates words not part of an organization name.

LSTM' s input at time t is vector x_t, so input word $v_t \in V$ must be converted to vector x_t, called v_t' s word vector. Let L be a $k \times n$ dimensional dense vector matrix where k is the vocabulary size, with each column corresponding to a word in V. Converting input word v_t to vector x_t simply involves looking up v_t' s index in V within L.

The current tag $s_t \in S$ must be computed from hidden state h_t using a simple Softmax function:

$$p(s_t = k|h_t) = \frac{\exp(f_k(h_t))}{\sum_{k' \in S} \exp(f_{k'}(h_t))}$$

where $f_k(h_t) = w_k^T h_t + b_k$ linearly transforms state h_t to a real number, with w_k as an n-dimensional coefficient vector and b_k as a bias term. Cross-entropy is used to compute the loss function. The loss at time t is:

$$L_t = -\sum_{k \in S} y_t(k) \log(p(s_t = k))$$

where $y_t(k) = 1$ if the true tag at step t is k, otherwise 0. The total loss function is the sum of each step' s loss:

$$L = \sum_t L_t = -\sum_t \sum_{k \in S} y_t(k) \log(p(s_t = k))$$

Model parameters include LSTM parameters, word vector matrix L, and parameters w_k, b_k for each tag k.

3.4 Character-Based Organization Name Tagging Model

[Figure 3: see original paper] illustrates the character-based tagging example. In the input layer, instead of words, individual Chinese characters and segmentation symbols `<GO>` are input. `<GO>` indicates that the next input character and previous input character do not belong to the same word. The LSTM layer is identical to the word model. In the output layer, tags are only produced at positions corresponding to `<GO>`.

This model inputs characters that must first be converted to character vectors via a lookup table before LSTM input. The loss function is only computed at `<GO>` positions. Therefore, we propose a new total loss function:

$$L = -\sum_{t \text{ where } i(t)=\text{GO}} \sum_{k \in S} y_t(k) \log(p(s_t = k))$$

where $i(t)$ denotes the input character at step t.

Regarding experimental setups for Chinese organization name recognition, there are two approaches. The first inputs raw text to build an integrated word segmentation and organization name recognition model, as in Zhou et al.' s research [2]. The second performs organization name recognition on pre-segmented corpora, as in Pan Zhenggao' s study [10]. This paper follows the second approach. Both setups may utilize character features, presented as feature templates, but our usage is fundamentally different—we directly use characters as the basic input unit.

4 Experiments

4.1 Dataset and Evaluation Metrics

Model performance was tested using the People's Daily corpus from the first half of 1998, released by the Institute of Computational Linguistics at Peking University. The corpus is pre-segmented and annotated with organization names marked as "nt" . The character model' s performance was evaluated against the word model baseline. In experiments, February 1998 data served as the test set, while January, March, April, May, and June 1998 data formed the training set. For the word model, the vocabulary built from training data contained 40,000 entries, including 39,999 high-frequency words and one rare word token "RAREWORD" . Words not appearing in the vocabulary during testing were marked as rare words. Similarly, a character table was built containing 5,500 entries: 5,498 high-frequency characters, one rare character token "RARECHAR" , and one segmentation symbol "GO" . shows statistics for words, characters, and rare items in the corpus. In the test set, rare words account for 3.18%, while rare characters are only 0.01%—almost negligible—demonstrating that character-based methods encounter fewer out-of-vocabulary issues.

Three metrics evaluate organization name tagging: precision P, recall R, and F-value, calculated as:

$$P = \frac{T}{N} \times 100\%$$

$$R = \frac{T}{M} \times 100\%$$

$$F = \frac{2PR}{P+R} \times 100\%$$

where T is the number of correctly tagged organizations, M is the total number of organizations in the test set, and N is the number of entities tagged by the model.

4.2 Parameter Settings

Mini-batch stochastic gradient descent was used for backpropagation with batch size 20 and initial learning rate 1.0, decreasing by a factor of 0.8 starting from epoch 5, for a total of 13 epochs. The word model' s maximum backpropagation steps were set to 35; since the character model requires more steps, its maximum was set to 55. All parameters were initialized to random values between -0.1 and 0.1. To prevent gradient explosion, gradient clipping [21] was applied with a threshold of 5.0. To mitigate overfitting, Dropout [22] was used with a rate of 0.8.

4.3 Experimental Results

presents recognition results with 2 hidden layers of dimension 650. "Overall" indicates performance on all organization names, while "Containing rare words" refers to cases where organization names include one or more rare words. Overall, the character model' s precision is 1.23% higher than the baseline, recall is 1.82% higher, and F-value is 1.54% higher. Performance on rare words is particularly outstanding, with precision 8.87% higher, recall 12.37% higher, and F-value 11.05% higher. The high metrics for rare words indicate that our method offers significant advantages when transferring to different corpora.

Error analysis reveals two main cases where the character model misses organization names (reducing recall). First, corpus annotation omissions or controversial cases: for example, in "电气化局三处是 [铁道部] 首家通过……", the corpus omits the tag but the algorithm correctly identifies it; similar cases include "[国家森林管理局]" and "[曲靖电厂]". Controversial cases include "图为新东安市场 [中安天平图书中心] 一角", where the model identifies "图为 [新东安市场中安天平图书中心] 一角". In our view, the character model' s conclusion is also reasonable, as using location as an organization name prefix is acceptable. Second, rare place names are sometimes misidentified as organization names because locations often constitute part of organization names, such as "委内瑞拉" (Venezuela) and "瑞士" (Switzerland). Observation shows the first case is the primary factor reducing recall.

The character model incorrectly tags non-organization names (reducing precision) in several situations. Most commonly, illegal overall tagging patterns occur: a legal organization name tag should begin with B-ORG followed by I-ORG tags (e.g., "[中俄总理定期会晤委员会]" → "B-ORG I-ORG I-ORG I-ORG I-ORG I-ORG"), but the model might tag "B-ORG S I-ORG I-ORG I-ORG I-ORG". Second, insufficient training data causes errors, such as with "[绵阳国家级高新技术产业开发区]", which is frequently misidentified because "绵阳" appears rarely in the training corpus and seldom in organization names, preventing correct learning.

In summary, recall errors mainly stem from corpus annotation omissions or controversial annotations, while precision errors primarily arise from insufficient data and illegal tagging patterns. For insufficient data, future work will use large-scale unannotated corpora to train character vectors and introduce multi-task learning techniques. For illegal tagging, CRF models can constrain outputs for more precise results.

5 Conclusion

This paper proposes a character-level Chinese organization name tagging model based on bidirectional deep LSTM, considering Chinese character and word characteristics. Compared to the baseline word-level model, the character model shows clear improvement in recognition capability, particularly for tagging rare words, demonstrating significant advantages when transferring to new corpora. Benefiting from deep learning, our model is fully end-to-end and no longer relies on manually designed rules, making it simpler and more user-friendly than traditional feature template methods. Future work will explore other deep learning methods for Chinese sequence labeling and attempt new approaches to further improve tagging performance.

References

[1] Shen Jiayi, Li Fang, Xu Feiyu, et al. Recognition of Chinese Organization Names and Abbreviations [J]. Journal of Chinese Information Processing, 2007, 21(6): 17-21.

[2] Zhou Junsheng, Dai Xinyu, Yin Cunyan, et al. Automatic Recognition of Chinese Organization Name Based on Cascaded Conditional Random Fields[J]. Acta Electronica Sinica, 2006, 34(5): 804-809.

[3] Huang Degen, Li Zezhong, Wan Ru. Chinese Organization Name Recognition Using Cascaded Model Based on SVM and CRF [J]. Journal of Dalian University of Technology, 2010, 50(5): 782-787.

[4] Teng Qingqing, Ji Jiuming, Zheng Yongting, et al. Applicability Analysis of Chinese Named Entity Recognition Method Based on Literatures [J]. Journal of Intelligence, 2010, 29(9): 157-161.

[5] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging [OL]. arXiv Preprint. arXiv: 1508.01991.

[6] Chen X, Qiu X, Zhu C, et al. Gated Recursive Neural Network for Chinese Word Segmentation [C]. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. 2015.

[7] Chen X, Xu L, Liu Z, et al. Joint Learning of Character and Word Embeddings [C]. In: Proceedings of the 24th International Conference on Artificial Intelligence. 2015.

[8] Sun Y, Lin L, Yang N, et al. Radical-enhanced Chinese Character Embedding [C]. In: Proceedings of International Conference on Neural Information Processing. Springer International Publishing, 2014: 279-286.

[9] Sun Zhen, Wang Huilin. Overview on the Advance of the Research on Named Entity Recognition [J]. New Technology of Library and Information Service, 2010(6): 42-47.

[10] Pan Zhenggao. Research on the Recognition of Chinese Named Entity Based on Rules and Statistics [J]. Information Science, 2012, 30(5): 708-712.

[11] Lu Wei, Ju Yuan, Zhang Xiaojuan, et al. Research on Product Named Entity Feature Selection and Recognition [J]. Document, Information & Knowledge, 2012(3): 4-12.

[12] Wu Dan, He Daqing, Lu Wei. The Extraction and Translation of Named Entity in Cross Language Information Retrieval [J]. Document, Information & Knowledge, 2012(3): 13-19.

[13] Wang Wenlong, Wang Dongbo. Project Application-oriented Named Entity Extraction Model Construction [J]. Information and Documentation Services, 2015(1): 30-34.

[14] Chen Feng, Zhai Yujia, Wang Fang. Automatic Theory Recognition in Academic Journals Based on CRF [J]. Library and Information Service, 2016, 60(2): 122-128.

[15] Yu Hongkui, Zhang Huaping, Liu Qun. Recognition of Chinese Organization Name Based on Role Tagging [C]. In: Proceedings of the 20th International Conference on Computer Processing of Oriental Languages. 2003: 79-87.

[16] Guan Xiaoda, Lv Xueqiang, Li Zhuo, et al. Chinese Organization Name Recognition in User Query Log [J]. New Technology of Library and Information Service, 2014(1): 72-78.

[17] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging [OL]. arXiv: 1508.01991.

[18] Ma X, Hovy E. End-to-End Sequence Labeling via Bi-directional LSTM-CNNs-CRF [OL]. arXiv Preprint. arXiv: 1603.01354.

[19] Hochreiter S, Schmidhuber J. Long Short-term Memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[20] Sutskever I, Vinyals O, Le Q V. Sequence to Sequence Learning with Neural Networks [A]. //Advances in Neural Information Processing Systems [M]. 2014: 3104-3112.

[21] Pascanu R, Mikolov T, Bengio Y. On the Difficulty of Training Recurrent Neural Networks [J]. Journal of Machine Learning Research, 2013, 28(3): 1310-1318.

[22] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

Author Contributions

Zhu Danhao: Conceived research idea, designed and conducted experiments, wrote manuscript;
Yang Lei: Assisted in research design, responsible for data preprocessing;
Wang Dongbo: Revised manuscript.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

The supporting data is self-archived by authors, E-mail: jisuanyuyan@163.com.
[1] Zhu Danhao, Yang Lei, Wang Dongbo. data cleaning programming.zip. People' s Daily corpus preprocessing program.

[2] Zhu Danhao, Yang Lei, Wang Dongbo. organization recognition model.zip. Character-based entity extraction model using RNN.

Dates

Received: 2016-08-01
Revised: 2016-10-26

ACRL Launches Information Literacy Framework Sandbox

The Association of College and Research Libraries (ACRL) Framework Advisory Board (FAB) recently announced the launch of the ACRL Framework for Information Literacy sandbox at sandbox.acrl.org.

The sandbox is a publicly accessible platform and repository that helps librarians and their educational partners discover, share, collect, and use ongoing work related to the ACRL Framework for Information Literacy in Higher Education for practice and professional development. The sandbox is a dynamic resource with content created by contributors engaged with the Framework.

ACRL President Irene M.H. Herold stated: "ACRL has launched this innovative resource to support the needs of librarians engaging with the Framework in various academic environments. By providing opportunities to discover and share teaching and professional development resources related to the Framework, the sandbox will help librarians promote the integration of information literacy into student learning. The sandbox will be members-only, so we encourage everyone to participate and contribute."

In this platform, visitors can browse and contribute by searching for materials that meet their needs and share their own materials with others. The sandbox will facilitate collaboration when librarians discover cases applicable to their libraries or find others researching similar topics.

For information on how to make the most of the sandbox, please refer to the sandbox help center.

(Compiled from: http://acrl.ala.org/framework/?p=332)
(News from this journal)

Note: Figure translations are in progress. See original paper for figures.