



A SURVEY ON CHINESE NAMED ENTITY RECOGNITION

Yaxiong Han

College of Computer Science and Software, Tianjin Polytechnic University, Tianjin, China.

ABSTRACT

Internet has become more and more connected with people's work and life, and the resources of online information has become increasingly diverse. We need to obtain useful information from a huge amount of data, which puts forward a great challenge to information extraction. Therefore, information extraction have drawn more attention. As a kind of information extraction technology, there are differences between Chinese Named Entity Recognition (NER) and Named Entity Recognition in other languages. This paper discusses the existing Chinese NER methods, features, models and evaluation methods and give a view of the future of NER.

KEY WORDS: Chinese named entity recognition, information extraction, model evaluation

1. INTRODUCTION:

The "Named Entity Recognition"(NER) was first proposed at the MUC-6 conference[1]. MUC-6 is the sixth in a series of Message Understanding Conferences, focused on information extraction tasks. Information extraction aims to extract the structured data and specific data relationships from text composed of natural language. For example, to extract specific information about company activities and international news from a piece of text. In the course of extracting information, some entity information with special meaning is indispensable for understanding the text, such as a person's name, organization name, location name, time, number and other information with specific meaning. These special information is called "Named Entity", and the task of recognising these entities with specific meaning in the text is called "Named Entity Recognition"(NER). Most of the early NER tasks aims to identify some special entities[2], mainly about three categories: person's name, location and institution name. As research continues, people have divided these three categories into more detailed categories, such as refining locations into countries, cities, streets etc[3] and refining the organization name into company name, government agency and so on. Moreover, the NER technology is also extended to some specialized domains. For example, in some biomedical corpora, the named entities mainly include protein names, virus names and cell names. In the corpus of botany, the named entities mainly include the subject, name and geographical location of the plant. In these domains, NER technology plays an active role in the information extraction[4]. However, in the Chinese information extraction task, Chinese characters usually cannot express a specific meaning. The words usually considered as the smallest unit to contain a specific meaning in text. In structured Chinese text, words usually do not have obvious boundary features, and there are many possibility of segmentation in a sequence. Therefore, the primary task of Chinese information extraction is to perform word segmentation, and the effect of word segmentation will also affect to the information extraction.

2. CHINESE WORD SEGMENTATION:

After decades of development, there are many excellent methods for Chinese word segmentation. The main results include three methods: dictionary-based segmentation[5,6], statistical-based segmentation[7,8] and understanding-based segmentation[9-11].

2.1 Dictionary-based Word Segmentation:

Dictionary-based word segmentation requires manually-build dictionary, and the quality of dictionary directly affects to the segmentation. Its main idea is to match the candidate string. If the candidate string successfully matched with the string in dictionary, then next segmentation will be performed. If the matching fails, the candidate string will be adjusted according to the matching strategy for the next matching.

This method relies heavily on a word dictionary and it is simple to implement, but it does not work well for ambiguous segmentation and cannot effectively process unregistered words which are not in the word dictionary.

2.2 Statistical-based Word Segmentation:

In Chinese, the credibility of words can be reflected according to the probability that a character and another are adjacent. The greater the probability of co-occurrence between characters, the higher of the probability that these characters will form words. Therefore, the word frequency information in the corpus can be statistically analyzed, and the segmentation result can be filtered by setting a threshold.

Statistical segmentation does not require a word dictionary. It has good stability

and cross-domain capabilities, but requires a large number of annotated text to train the model. In addition, the algorithm of the statistical model is complex and due to the limited tagged corpus, it cannot well cover the various situations that may occur.

2.3 Understanding-based Word Segmentation:

The understanding-based segmentation method mainly trains the neural network on a tagged corpus and obtains the word segmentation model. In the process of model training, the neural network can learn the grammar and semantic information included in the text, and store the learned information in the trained model. Then continues to train the model to reach a convergence state. The trained model is then used to segment the undivided corpus.

3. TECHNICAL METHODS:

In the information extraction system, it is quite important to recognize the unknown information. Early research on NER mainly consists of artificially constructing classification rules, which are triggered by unique rules related to positive and negative examples. Nowadays, the method of supervised learning is adopted, and the system automatically constructs rules for sequence labeling tasks. The idea of supervised learning is to learn the characteristics of positive and negative examples in named entities on a large number of annotated documents and to design rules that contain the essence of the corresponding examples. Furthermore, due to the lack of corpora, unsupervised learning can be employed to replace supervised learning.

3.1 Supervised Learning:

The supervised learning method considers the NER as a sequence labeling problem, which is also the mainstream recognition technology. The methods of supervised learning include: Hidden Markov Models (HMM)[12], Maximum Entropy Markov Models (MEMM), Conditional Random Fields (CRF)[13] and so on. The most commonly used model is the conditional random field model. Fig 1 shows the general structure of a Conditional Random Field model.

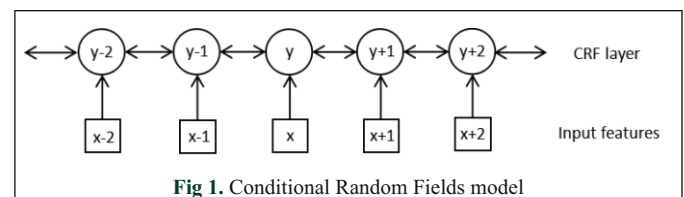


Fig 1. Conditional Random Fields model

The Conditional Random Field model transforms the conditional probability in the Maximum Entropy Markov Model into a form of eigenfunction, which is decomposed into two parts: the transfer feature and the state feature. The values of the two features are obtained by training. In the recognition process, a Viterbi decoding algorithm is usually adopted. The Conditional Random Field effectively solves the mark bias problems existing in Hidden Markov Model.

3.2 Unsupervised Method:

Due to the lack of available annotated corpora, the researchers further proposed an unsupervised approach. The most common method of this technique is clustering, which gathers together different named entities in similar contexts. There are also methods based on external resources, for example, when there is not enough available corpus in a particular domain, the migration learning can be employed. For example, taking the WordNet[14] as an external resource, the specific methods are as follows. 1) First, assign an entity type to synonym in

wordNet by co-occurrence of words in large-scale corpus. 2) Then assign an entity type to the word by comparing a certain length of context for a given word. 3) Based on the point mutual information[15], the point mutual information is used as a feature to classify the given words to determine which feature to input.

4. EXPERIMENTAL COMPARISON:

In order to experimentally test the effect of the current mainstream recognition model, we manually choose a Conditional Random Field model[16] and a deep neural network model which is based on BiLSTM-CRF[17] to perform an experiment of NER. The Conditional Random Field model is shown in Figure 1 while the structure of the deep neural network model is shown in Fig 2. We tested the recognition effects of these two models through experiments, and compared their recognition results.

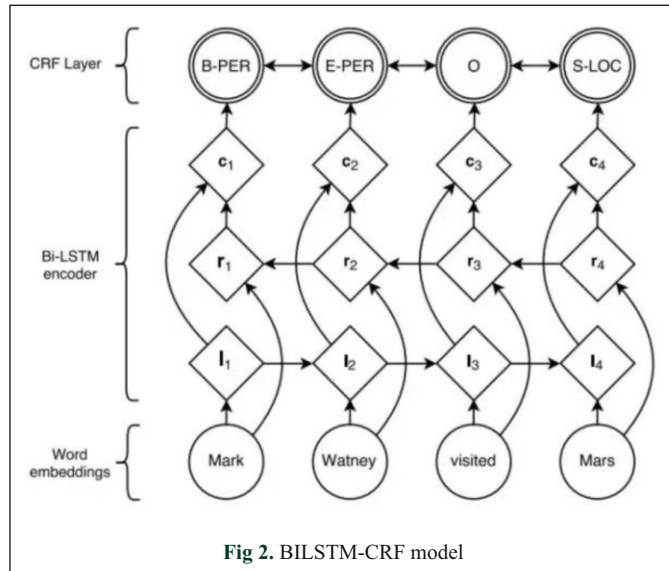


Fig 2. BiLSTM-CRF model

4.1 Corpus:

Bakeoff-3 is the third international Chinese processing competition organized by SIGHAN. We selected the MSRA Chinese NER corpus provided by this competition as the experimental corpus and divided this corpus into training set, development set and test set. The details of the three collections are listed in Table 1.

Table 1. Corpora statistics for different data sets

	Training Set	Development Set	Test Set
SENTENCE	35157	3772	7432
TOKEN	1627032	180106	362741

The corpus contains three entity tags, they are PERSON, LOCATION and ORGANIZATION. We used the BIOES tagging scheme in our experiment, where B represents the first character of an entity mention, I represents a non-starting and non-end character in an entity mention, E represents an ending character, O represents a character that does not consist of an entity, and S represents a character that constitutes the entity alone. For example, B-PER represents the first character of a PERSON type entity mention, and E-LOC represents the ending character of the LOCATION entity mention.

4.2 Features:

Features are used to describe the various attributes of named entity, and the features used are different for different systems. Features are generally represented by vectors. Each dimension of the vector can represent a feature's value, this value can be boolean or numeric. The entire vector represents all attributes of the word under the assumptions. Features are generally divided into three categories: 1) word-level features, including the context of the words, part of speech and other information. 2) dictionary-level features, to determine whether the current word belongs to a dictionary. 3) global features.

In the experiment, we selected the word-level features for the Conditional Random Field model. The details of the features are shown in Table 2. We used character vector as the feature for training the BiLSTM-CRF model, which was obtained by Word2vec training the Chinese Wikipedia corpus. We obtain a total of 16115 character vectors with 100 dimensions for each of these vectors. A lookup table is built using these vectors. In addition, we use the word segmentation feature of the input sequence as additional auxiliary training feature. Jieba system is used as the word segmentation tool. The hyperparameters used for training the neural model are shown in Table 3.

Table 2. Features used to train CRF Model

Description	Feature Template
Character Unigram	$w_i, -1 \leq i \leq 1$
Character Bigram	$w_i, w_j, 0 \leq i \leq 1$
Word Seg Unigram	$\text{Seg}(w_0)$
Word Seg Bigram	$\text{Seg}(w_{i-1})\text{Seg}(w_i), 0 \leq i \leq 1$
Word Seg Trigram	$\text{Seg}(w_{i-1})\text{Seg}(w_i)\text{Seg}(w_{i+1})$

Table 3. Hyperparameters used to train the neural Model

Parameter	Value	Parameter	Value
Character Dim	100	Dropout Rate	0.5
Hidden States	100	Batch Size	20
Segmentation Dim	20	Learning Rate	0.001
Peepholes	No	Gradient Clip	5.0

4.3 Experimental Results:

The NER model generally uses the following evaluation indicators: Precision, Recall, F1. According to the definition in Table 4, their respective calculation formulas are as follows:

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$r = \frac{TP}{TP + FN} \quad (2)$$

$$F = \frac{2Pr}{P + r} \quad (3)$$

Where P represents Precision, r represents Recall and F represents F1.

Table 4. Symbols for examples

	Recognized As Positive	Recognized As Negative
Actual Positive Example	TP	FN
Actual Negative Example	FP	TN

We show the results of the two models for the test set in Table 5 from three aspects of Precision, Recall and F1. From the results listed in Table 5, both models can achieve a good recognition results. It should be noted that the model based on deep learning has become the mainstream method in the field of Chinese information extraction. We can also see from the experimental results that the BiLSTM-CRF model has better performance than the Conditional Random Field model.

Table 5. Performance of two NER systems

Model	Precision	Recall	F1
CRF Model	81.57	75.12	78.21
BiLSTM-CRF baseline	90.41	90.54	90.48

5. CONCLUSION:

As a sub-task of information extraction, NER has received a generous concern and become a research focus. It has achieved remarkable results since the beginning of the research. This paper discusses the basic concepts and significance of NER and summarizes the existing Chinese NER techniques, features, and evaluation methods. Then a simple comparison experiment was conducted to compare the effects of the existing model. At present, the research on NER in some domains such as news is mature. How to apply mature technological methods to some emerging domains such as bio-agricultural science and biomedicine is a trend of future development of Chinese NER system.

Acknowledgement:

Thanks to SIGHAN bakeoff-3 and MSRA for providing the relevant experimental corpus.

REFERENCES:

1. Grishman R. Message understanding conference 6 : A brief history[J]. Proc. COLING-96, 1996.
2. Thielen C. An Approach to Proper Name Tagging for German[J]. 1995.
3. Lee S, Lee G G. Heuristic Methods for Reducing Errors of Geographic Named Entities Learned by Bootstrapping[M]// Natural Language Processing – IJCNLP 2005. Springer Berlin Heidelberg, 2005.
4. Settles B. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets[J]. In Proceedings of COLING 2004, International Joint Workshop On Natural Language Processing in Biomedicine and its Applications (NLPBA, 2004:104--107.
5. 黄德根, 朱和合, 王昆仑,等. 基于最长次长匹配的汉语自动分词[J]. 大连理工大学学报, 1999(6):831-835.
6. 周俊, 郑中华, 张炜. 基于改进最大匹配算法的中文分词粗分方法[J]. 计算机工程与应用, 2014, 50(2):124-128.
7. 费洪晓, 康松林, 朱小娟,等. 基于词频统计的中文分词的研究[J]. 计算机工程与应用, 2005, 41(7):67-68.
8. 张梅山, 邓知龙, 车万翔,等. 统计与词典相结合的领域自适应中文分词[J]. 中文信息学报, 2012, 26(2):8-13.
9. Chen X, Qiu X, Zhu C, et al. Gated Recursive Neural Network for Chinese Word Segmentation[C]. international joint conference on natural language processing, 2015: 1744-1753.
10. Zheng X, Chen H, Xu T. Deep learning for Chinese word segmentation and POS tagging[C]// Conference on Empirical Methods in Natural Language Processing. 2013.
11. Chen X, Qiu X, Zhu C, et al. Long Short-Term Memory Neural Networks for Chinese Word Segmentation[C]// Conference on Empirical Methods in Natural Language Processing. 2015:1197-1206.
12. Bikel, D. M., Miller, S., Schwartz, R., & Weischedel, R. (1998). Nymble: a high-performance learning name-finder. Anlp, 194--201.
13. Mccallum A, Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]// Conference on Natural Language Learning at Hlt-Naacl. Association for Computational Linguistics, 2003:188-191.
14. Alfonseca E, Man S. An Unsupervised Method for General Named Entity Recognition And Automated Concept Discovery[C]// 1 St International Conference on General Wordnet. 2002.
15. Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the Web: An experimental study [J]. Artificial Intelligence, 2005, 165(1):91-134.
16. Lafferty J D, Mccallum A, Pereira F C N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]// Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2001:282-289.
17. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging[J]. Computer Science, 2015.