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 an overview 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 usually called extraction information. 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. In Chinese, the credibility of words can be reflected according to the probability of 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.

2. CHINESE WORD SEGMENTATION:
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 unidivided 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.

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[4] 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 a 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 a 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.

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.

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 a ending character, O represents a character that does not consists 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 superparameters used for training the neural model are shown in Table 3.

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:

\[
P = \frac{TP}{TP + FP}
\]

\[
R = \frac{TP}{TP + FN}
\]

\[
F = \frac{2Pr}{P + r}
\]

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

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.

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:
2. Thielen C. An Approach to Proper Name Tagging for German[J]. 1995.