Chinese Named Entity Recognition with Conditional Probabilistic Models

Aitao Chen
Yahoo
701 First Avenue
Sunnyvale, CA 94089
aitao@yahoo-inc.com

Fuchun Peng
Yahoo
701 First Avenue
Sunnyvale, CA 94089
fuchun@yahoo-inc.com

Roy Shan
Yahoo
701 First Avenue
Sunnyvale, CA 94089
rshan@yahoo-inc.com

Gordon Sun
Yahoo
701 First Avenue
Sunnyvale, CA 94089
gzsun@yahoo-inc.com

Abstract

This paper describes the work on Chinese named entity recognition performed by Yahoo team at the third International Chinese Language Processing Bakeoff. We used two conditional probabilistic models for this task, including conditional random fields (CRFs) and maximum entropy models. In particular, we trained two conditional random field recognizers and one maximum entropy recognizer for identifying names of people, places, and organizations in un-segmented Chinese texts. Our best performance is 86.2% F-score on MSRA dataset, and 88.53% on CITYU dataset.

1 Introduction

At the third International Chinese Language Processing Bakeoff, we participated in the closed test in the Named Entity Recognition (NER) task using the MSRA corpus and the CITYU corpus. The named entity types include person, place, and organization. The training data consist of texts that are segmented into words with names of people, places, and organizations labeled. And the testing data consist of un-segmented Chinese texts, one sentence per line.

There are many well known models for English named recognition, among which Conditional Random Fields (Lafferty et al. 2001) and maximum entropy models (Berger et al. 2001) have achieved good performance in English in CoNLL NER tasks. To understand the performance of these two models on Chinese, we both models to Chinese NER task on MSRA data and CITYU data.

2 Named Entity Recognizer

2.1 Models

We trained two named entity recognizers based on conditional random field and one based on maximum entropy model. Both conditional random field and maximum entropy models are capable of modeling arbitrary features of the input, thus are well suit for many language processing tasks. However, there exist significant differences between these two models. To apply a maximum entropy model to NER task, we have to first train a maximum entropy classifier to classify each individual word and then build a dynamic programming for sequence decoding. While in CRFs, these two steps are integrated together. Thus, in theory, CRFs are superior to maximum entropy models in sequence modeling problem and this will also confirmed in our Chinese NER experiments. The superiority of CRFs on Chinese information processing was also demonstrated in word segmentation (Peng et al. 2004). However, the training speed of CRFs is much slower than that of maximum entropy models since training CRFs requires expensive forward-backward algorithm to compute the partition function.

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 173–176,
Sydney, July 2006.

We used Taku's CRF package[1] to train the first CRF recognizer, and the MALLET[2] package with BFGS optimization to train the second CRF recognizer. We used a C++ implementation[3] of maximum entropy modeling and wrote our own second order dynamic programming for decoding.

2.2 Features

The first CRF recognizer used the features C_{-2}, C_{-1}, C_0, C_{-1}, C_2, $C_{-2}C_{-1}$, $C_{-1}C_0$, C_0C_{-1}, C_1C_2, and $C_{-1}C_1$, where C_0 is the current character, C_1 the next character, C_2 the second character after C_0, C_{-1} the character preceding C_0, and C_{-2} the second character before C_0.

The second CRF recognizer used the same set of basic features but the feature C_2. In addition, the first CRF recognizer used the tag bigram feature, and the second CRF recognizer used word and character cluster features, obtained automatically from the training data only with distributional word clustering (Tishby and Lee, 1993).

The maximum entropy recognizer used the following unigram, bigram features, and type features: C_{-2}, C_{-1}, C_0, C_1, C_2, $C_{-4}C_{-3}$, $C_{-3}C_{-2}$, $C_{-2}C_{-1}$, $C_{-1}C_0$, C_0C_1, C_1C_2, C_2C_3, C_3C_4, and $T_{-2}T_{-1}$.

When using the first CRF package, we found the labeling scheme OBIE performs better than the OBIE scheme. In the OBI scheme, the first character of a named entity is labeled as "B", the remaining characters, including the last character, are all labeled as "I". And any character that is not part of a named entity is labeled as "O". In the OBIE scheme, the last character of a named entity is labeled as "E". The other characters are labeled in the same way as in OBIE scheme. The first CRF recognizer used the OBIE labeling scheme, and the second CRF recognizer used the OBI scheme.

We tried a window size of seven characters (three characters preceding the current character and three characters following the current character) with almost no difference in performance from using the window size of five characters.

When a named entity occurs frequently in the training data, there is a very good chance that it will be recognized when appearing in the testing data. However, for entity names of rare occurrence, they are much harder to recognize in the testing data. Thus it may be beneficial to examine the testing data to identify the named entities that occur in the training data, and assign them the same label as in the training data. From the training data, we extracted the person names of at least three characters, the place names of at least four characters, and the organization names of at least four characters. We removed from the dictionary the named entities that are also common words. We did not include the short names in the dictionary because they may be part of long names. We produced a run first using one of the NER recognizers, and then replaced the labels of a named entity assigned by a recognizer with the labels of the same named entity in the training data without considering the contexts.

3 Results

Run ID	Precision	Recall	F-Score
msra_a	91.22%	81.71%	86.20
msra_b	88.43%	82.88%	85.56
msra_f	88.45%	79.31%	83.63
msra_g	86.61%	80.32%	83.35
msra_r	87.48%	71.68%	78.80

Table 1: Official results in the closed test of the NER task on MSRA corpus.

Table 1 presents the official results of five runs in the closed test of the NER task on MSRA corpus. The first two runs, msra_a and msra_b, are produced using the first CRF recognizer; the next two runs, msra_f and msra_g, are produced using the second CRF recognizer which used randomly selected 90% of the MSRA training data. When we retrained the second CRF recognizer with the whole set of the MSRA training data, the overall F-Score is 85.00, precision 90.28%, and recall 80.31%. The last run, msra_r, is produced using the MaxEnt recognizer.

The msra_a run used the set of basic features with a window size of five characters. Slightly over eight millions features are generated from the MSRA training data, excluding features occurred only once. The training took 321 iterations to complete. The msra_b run is produced from the msra_a run by substituting the labels assigned by the recognizer to a named entity with the labels of the named entity in the training data if it occurs in the training data. For example, in the MSRA training data, the text 毕加索故居 in the sentence 我还到毕加索故居去瞻仰 is tagged as a place name. The same entity also appeared in MSRA testing data set. The first CRF recognizer failed to mark the text 毕加索故居 as

[1] Available from http://chasen.org/~taku/software/CRF++
[2] Available at http://mallet.cs.umass.edu
[3] Available at http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.htm

a place name instead it tagged 毕加索 as a person name. In post-processing, the text 毕加索故居 in the testing data is re-tagged as a place name. As another example, the person name 章念生 appears both in the training data and in the testing data. The first CRF recognizer failed to recognize it as a person name. In post-processing the text 章念生 is tagged as a person name because it appears in the training data as a person name. The text "全国人大香港特别行政区筹备委员会" was correctly tagged as an organization name. It is not in the training data, but the texts "全国人大", "香港特别行政区", and "筹备委员会" are present in the training data and are all labeled as organization names. In our post-processing, the correctly tagged organization name is re-tagged incorrectly as three organization names. This is the main reason why the performance of the organization name got much worse than that without post-processing.

	Precision	Recall	F-score
LOC	94.19%	87.14%	90.53
ORG	83.59%	80.39%	81.96
PER	92.35%	74.66%	82.57

Table 2: The performance of the msra_a run broken down by entity type.

	Precision	Recall	F-score
LOC	93.09%	87.35%	90.13
ORG	75.51%	78.51	76.98
PER	91.52	79.27	84.95

Table 3: The performance of the msra_b run broken down by entity type.

Table 2 presents the performance of the msra_a run by entity type. Table 3 shows the performance of the msra_b run by entity type. While the post-processing improved the performance of person name recognition, but it degraded the performance of organization name recognition. Overall the performance was worse than that without post-processing. In our development testing, we saw large improvement in organization name recognition with post-processing.

Run ID	Precision	Recall	F-Score
cityu_a	92.66%	84.75%	88.53
cityu_b	92.42%	84.91%	88.50
cityu_f	91.88%	82.31%	86.83
cityu_g	91.64%	82.46%	86.81

Table 4: Official results in the closed test of the NER task on CITYU corpus.

Table 4 presents the official results of four runs in the closed test of the NER task on CITYU corpus. The first two runs, msra_a and msra_b, are produced using the first CRF recognizer; the next two runs, msra_f and msra_g, are produced using the second CRF recognizer. The system configurations are the same as used on the MSRA corpus. The cityu_b run is produced from cityu_a run with post-processing, and the cityu_g run produced from cityu_f run with post-processing. We used the whole set of CITYU to train the first CRF model, and 80% of the CITYU training data to train the second CRF model. No results on full training data are available at the time of submission.

All the runs we submitted are based characters. We tried word-based approach but found it was not as effective as character-based approach.

4 Discussions

Table 4 is shows the confusion matrix of the labels. The rows are the true labels and the columns are the predicated labels. An entry at row *x* and column *y* in the table is the number of characters that are predicated as *y* while the true label is *x*. Ideally, all entries except the diagonal should be zero.

The table was obtained from the result of our development dataset for MSRA data, which are the last 9,364 sentences of the MSRA training data (we used the first 37,000 sentences for training in the model developing phase). As we can see, most of the errors lie in the first column, indicating many of the entities labels are predicated as O. This resulted low recall for entities. Another major error is on detecting the beginning of ORG (B-O). Many of them are mislabeled as O and beginning of location (B-L), resulting low recall and low precision for ORG.

	O	B-L	I-L	B-O	I-O	B-P	I-P
O	406798	86	196	213	973	46	111
B-L	463	5185	54	73	29	19	7
I-L	852	25	6836	0	197	1	44
B-O	464	141	3	2693	62	17	0
I-O	1861	28	276	55	12626	2	39
B-P	472	16	2	22	3	2998	8
I-P	618	0	14	1	49	10	5502

Table 4: Confusion matrix of on the MSRA development dataset

A second interesting thing to notice is the numbers presented in Table 2. They may suggest that person name recognition is more difficult

than location name recognition, which is contrary to what we believe, since Chinese person names are short and have strict structure and they should be easier to recognize than both location and organization names. We examined the MSRA testing data and found out that 617 out 1,973 person names occur in a single sentence as a list of person names. In this case, simple rule may be more effective. When we excluded the sentence with 617 person names, for person name recognition of our msra_a run, the F-score is 90.74, precision 93.44%, and recall 88.20%. Out of the 500 person names that were not recognized in our msra_a run, 340 occurred on the same line of 617 person names.

5 Conclusions

We applied Conditional Random Fields and maximum entropy models to Chinese NER tasks and achieved satisfying performance. Three systems with different implementations and different features are reported. Overall, CRFs are superior to maximum entropy models in Chinese NER tasks. Useful features include using BIOE tags instead of BIO tags and word and character clustering features.

References

Adam Berger, Stephen Della Pietra, and Vincent Della Pietra, A Maximum Entropy Approach to Natural Language Processing, *Computational Linguistics*, 22 (1)

John Lafferty, Andrew McCallum, and Fernando Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proc. 18th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA (2001) 282–289

Fuchun Peng, Fangfang Feng, and Andrew McCallum, Chinese Segmentation and New Word Detection using Conditional Random Fields, In *Proceedings of The 20th International Conference on Computational Linguistics (COLING 2004)* , pages 562-568, August 23-27, 2004, Geneva, Switzerland

Naftali Tishby and Lillian Lee, Distributional Clustering of English Words, In *Proceedings of the 31st Annual Conference of Association for Computational Linguistics*, pp 183--190, 1993.