A Two-Stage Approach to Chinese Part-of-Speech Tagging

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

Ya Zhang
Yahoo! Inc.
701 First Avenue
Sunnyvale, CA 94089
yazhang@yahoo-inc.com

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

Abstract

This paper describes a Chinese part-of-speech tagging system based on the maximum entropy model. It presents a novel two-stage approach to using the part-of-speech tags of the words on both sides of the current word in Chinese part-of-speech tagging. The system is evaluated on four corpora at the Fourth SIGHAN Bakeoff in the close track of the Chinese part-of-speech tagging task.

1 Introduction

A part-of-speech tagger typically assigns a tag to each word in a sentence sequentially from left to right or in reverse order. When the words are tagged from left to right, the part-of-speech tags assigned to the previous words are available to the tagging of the current word, but not the tags of the following words. And when words are tagged from right to left, only the tags of the words on the right side are available to the tagging of the current word. We expect the use of the tags of the words on both sides of the current word should improve the tagging of the current word. In this paper, we present a novel two-stage approach to using the tags of the words on both sides of the current word in tagging the current word. We train two maximum entropy part-of-speech taggers on the same training data. The difference between the two taggers is that the second tagger uses features involving the tags of the words on both sides of the current word, while the first tagger uses the tags of only the previous words. Both taggers assign tags to words from left to right. In tagging a new sentence, the first tagger is applied to the testing data, and then the second tagger is applied to the output of the first tagger to produce the final results.

We participated in the Chinese part-of-speech tagging task at the Fourth International Chinese Language Processing Bakeoff. Our Chinese part-of-speech taggers were trained only on the training data provided to the participants, and evaluated on four corpora in the close track of the part-of-speech tagging task. The words in both the training and testing data sets are already segmented into words.

2 Maximum Entropy POS Tagger

Maximum entropy model is a machine learning algorithm that has been applied to a range of natural language processing tasks, including part-of-speech tagging (Ratnaparkhi, 1996). Our Chinese part-of-speech taggers are based on the maximum entropy model.

2.1 Maximum Entropy Model

The conditional maximum entropy model (Berger, et. al., 1996) has the form

$$p(y \mid x) = \tfrac{1}{Z(x)} \exp(\sum_k \lambda_k f_k(x, y))$$

where $Z(x) = \sum_y p(y \mid x)$ is a normalization factor, and λ_k is a weight parameter associated with feature $f_k(x, y)$. In the context of part-of-speech tagging, y is the POS tag assigned to a word, and x represents the contextual information regarding the word in consideration, such as the surrounding words. A feature is a real-valued, typically binary, function. For example, we may define a binary feature which takes the value 1 if the current word of X is ‘story’ and its POS tag is ‘NNS’; and 0 otherwise. Given a set of training examples, the log likelihood of the model with Gaussian prior (Chen and Rosenfeld, 1999) has the form

$$L(\lambda) = \sum_i \log p(y^{(i)} \mid x^{(i)}) - \sum_k \frac{\lambda_k^2}{2\sigma^2} + const$$

Malouf (2002) compared iterative procedures such as *Generalized Iterative Scaling (GIS)* and *Improved Iterative Scaling (IIS)* with numerical optimization techniques like limited-memory BFGS (L-BFGS) for estimating the maximum entropy model parameters and found that L-BFGS outperforms the other methods. The use of L-BFGS requires the computation of the gradient of the log likelihood function. The first derivative with respect to parameter λ_k is given by

$$\frac{\partial L(\lambda)}{\partial \lambda_k} = E_{\tilde{p}} f_k(x,y) - E_p f_k(x,y) - \frac{\lambda_k}{\sigma^2}$$

where the first term $E_{\tilde{p}} f_k$ is the feature expectation with the empirical model, and the second term $E_p f_k$ is the feature expectation with respect to the model. In our model training, we used L-BFGS to estimate the model parameters by maximizing $L(\lambda)$ on the training data.

2.2 Features

The feature templates used in our part-of-speech taggers are presented in Table 1 and Table 2.

Word	$w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}$ $w_{i-2}w_{i-1}, w_{i-1}w_i, w_iw_{i+1}, w_{i+1}w_{i+2,}$ $w_{i-1}w_{i+1}, w_{i-1}w_iw_{i+1}$
Tag	$t_{i-1}, t_{i-2}t_{i-1}$
Word/Tag	$t_{i-1}w_i, t_{i-2}w_i$
Special	*FirstChar*, *LastChar*, *Length*, *ForeighWord*

Table 1: Feature templates used in the first stage POS tagger.

Tag	$t_{i+1}, t_{i+1}t_{i+2}, t_{i-1}t_{i+1}$
Word/Tag	w_it_{i+1}, w_it_{i+2}

Table 2: Additional feature templates used in the second stage POS tagger.

The features are grouped into four categories. The first category contains features involving word tokens only; the second category consists of features involving tags only; the third category has features involving both word tokens and tags. And the last category has four special features. In the feature templates, w_i denotes the current word, w_{i-2} the second word to the left, w_{i-1} the previous word, w_{i+1} the next word, w_{i+2} the second word to the right of the current word, and t_i denotes the part-of-speech tag assigned to the word w_i. The *FirstChar* refers to the initial character of a word, and the *LastChar* the final character of a word. The *Length* denotes the length of a word in terms of byte. And the feature *ForeignWord* indicates whether or not a word is a foreign word. Table 2 shows additional feature templates involving the part-of-speech tags of the following one or two words. The features involving the tags of the words in the right contexts are used only in the second maximum entropy POS tagger. Features are generated from the training data according to the feature templates presented in Table 1 and Table 2.

2.3 Training Models

The four training corpora we received for the Chinese part-of-speech tagging task include the Academia Sinica corpus (**CKIP**), the City University of Hong Kong corpus (**CityU**), the National Chinese Corpus (**NCC**), and the Peking University corpus (**PKU**). The CKIP corpus and the CityU corpus contain texts in traditional Chinese, while the NCC corpus and the PKU corpus contain texts in simplified Chinese. The texts in all four training corpora are segmented into words according to different word segmentation guidelines. And the words in all training corpora are labeled with part-of-speech tags using different tag sets.

Two maximum entropy POS taggers were trained on each of the four corpora using our own implementation of the maximum entropy model. The first-stage POS tagger was trained with only the feature templates presented in Table 1, while the second-stage POS tagger with the feature templates presented in both Table 1 and Table 2.

All the first-stage POS taggers, one for each corpus, were trained with the same feature templates shown in Table 1, and all the second-stage POS taggers were trained with the same feature templates shown in Table 1 and Table 2. The feature templates are not necessarily optimal for each individual corpus. For simplicity, we chose to apply the same feature templates to all four corpora.

The same parameter settings were applied in the training of all eight POS taggers. More specifically, no feature selection was performed. All features, including features occurring just once in the training data, were retained. The sigma square σ^2 was set to 5.0. And the training process was terminated when the ratio of the likelihood difference between the current iteration and the previous iteration over the likelihood of the current iteration is below the pre-defined threshold or the maximum number of iterations, which was set to 400, is reached. Both the first-stage POS tagger and the second-stage POS tagger were trained on the same corpus.

2.4 Testing the Models

The POS tagger assigns a part-of-speech tag to each word in a new sentence such that the tag sequence maximizes the probability *p(Y|X)*, where X is the input sentence, and Y the POS tags assigned to X. The decoder implements the beam search procedure described in (Ratnaparkhi, 1996). At each word position, the decoder keeps the top n best tag sequences up to that position. The decoder also uses a word/tag dictionary, consisting of the words in the training data and the tags assigned to each word in the training data. During the decoding phase, if a word in the new sentence is found in the training data, only the tags that are assigned to that word in the training corpus are considered. Otherwise, all the tags in the tag set are considered for a new word. So the tagger will not assign to a word, found in the training data, a tag that is never assigned to that word in the training data, even if that word should be assigned a new tag that was never assigned to the word in the training data. A word/tag dictionary is automatically built by collecting all the words in the training corpus and the tags assigned to every word in the training corpus.

The final output is produced in two steps. The first-stage POS tagger is applied on the testing data, and then the second-stage POS tagger is applied on the output of the first POS tagger. The second-stage tagger uses features involving POS tags of the following one or two words. The features involving the tags of following one or two words may be erroneous, since the tags assigned to the following one or two words by the first-stage tagger may be incorrect.

3 Evaluation Results

Five corpora are provided for the Chinese part-of-speech tagging task at the forth SIGHAN bakeoff. We selected four corpora, two in simplified Chinese and two in traditional Chinese.

Corpus	Training size (tokens)	Tagset size	No. of tags per token type
CityU	1,092,687	44	1.2587
CKIP	721,551	60	1.1086
NCC	535,023	60	1.0658
PKU	1,116,754	103	1.1194

Table 3: Training corpus size.

Table 3 shows the training corpus size, the tagset size, and the average number of tags per token type. The NCC tagset has 60 tags, but nine of the tags occurred only once in the training corpus. In all four corpora, most of the unique tokens have only a single tag. The percentage of token types having single tag is 83.29% in CityU corpus; 91.09 in CKIP corpus; 94.67 in NCC corpus; and 90.27% in PKU corpus. The proportion of token types having single tag in CityU corpus is much lower than in NCC corpus. In the NCC corpus, the organization names, location names, and a sequence of English words are all treated as single token, and these long single tokens are not ambiguous and are assigned to a single part-of-speech tag in the corpus.

corpus	Baseline	Testing size	Token/tag OOV-R
CityU	0.8433	184,314	0.0921
CKIP	0.8865	91,071	0.0897
NCC	0.9159	102,344	0.0527
PKU	0.8805	156,407	0.0594

Table 4: The testing data size and the baseline performance.

The baseline performance is computed by assigning the most likely tag to each word in the testing data. When a word in the testing data is found in the training corpus, it is assigned the tag that is most frequently assigned to that word in the training corpus. A new word in the testing data is assigned the most frequent tag found in the training corpus, which is the common noun in all four corpora. The baseline performances of the four testing data sets are presented in Table 4, which also shows the percentage of new token/tag in the testing data sets.

Our POS taggers are evaluated on four testing data sets, one corresponding to each training corpus. We trained eight POS taggers, two on each training corpus, and submitted eight runs in total on the Chinese part-of-speech tagging task, two runs on each testing data set. The first run, labeled ‘a’ in Table 5, is produced using the first-stage tagger, and the second run, labeled ‘b’ in Table 5, is the output of the second-stage tagger, which is applied to the output of the first tagger. For all of our runs, only the provided training data are used. Table 5 shows the official evaluation results of the eight runs we submitted in the close track. The third column, labeled ‘Total-A’, shows the accuracy of the eight runs. The accuracy is the proportion of correctly tagged words in a testing data set. Only one tag is assigned to every word in the testing data set. The remaining three labels, ‘IV-R’, ‘OOV-R’, and ‘MT-R’, may be defined in The Fourth SIGHAN Bakeoff overview paper.

Corpus	Run ID	Total-A	IV-R	OOV-R	MT-R
CityU	a	0.8929	0.9367	0.4608	0.8705
CityU	b	0.8951	0.9389	0.4637	0.8745
CKIP	a	0.9286	0.9618	0.5875	0.9099
CKIP	b	0.9295	0.9629	0.5869	0.9123
NCC	a	0.9525	0.9717	0.6059	0.9135
NCC	b	0.9541	0.9738	0.5998	0.9195
PKU	a	0.9420	0.9648	0.5813	0.9148
PKU	b	0.9450	0.9679	0.5818	0.9252

Table 5: Official evaluation results of eight runs in the close track of the Chinese part-of-speech tagging task.

4 Discussions

A Chinese verb can function as a noun, and vice versa, without suffix change. In PKU corpus, a verb is labeled with the tag ‘v’, and a verb that functions as a noun is labeled with the tag ‘vn’. In the PKU-b run, almost half of the incorrectly tagged verbs (v) were tagged as verbal noun (vn), and slightly more than half of the incorrectly tagged verbal nouns (vn) were tagged as verb (v).

The accuracy of our best runs on all four corpora is much higher than the baseline performance. On the PKU corpus, the accuracy is increased from the baseline performance of 0.8805 to 0.9450, an improvement of 7.33% over the baseline. The second-stage tagging increased the accuracy on all four corpora. On the PKU corpus, the accuracy is increased by about 0.32% over the first-stage tagging. The improvement may not seem to be large; however, it corresponds to an error reduction by 5.4%.

That the accuracy on the CityU corpus is the lowest among all four corpora is not surprising, given that the CityU testing data set has the highest out-of-vocabulary rate, and the CityU training corpus has the highest average number of tags assigned to each token type. Furthermore, the CityU training corpus has the lowest percentage of tokens with only one tag. The POS tagging task on CityU corpus seems to be most challenging among the four corpora.

5 Conclusions

We have described a Chinese part-of-speech tagger with maximum entropy modeling. The tagger with rich lexical and morphological features significantly outperforms the baseline system which assigns to a word the most likely tag assigned to that word in the training corpus. The use of features involving the part-of-speech tags of the following words further improves the performance of the tagger.

References

Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics,* 22(1):39-71.

Stanley F. Chen, and Ronald Rosenfeld. 1999. *A Gaussion Prior for Smoothing Maximum Entropy Models*, Technical Report CMU-CS-99-108, Carnegie Mellon University.

Rober Malouf. 2002. *A Comparison of Algorithms for Maximum Entropy Parameter Estimation*, Proceedings of CoNLL-2002.

Adwait Ratnaparkhi. 1996. *A Maximum Entropy Model for Part-of-Speech Tagging*, Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp 133-142.