A Maxent NER for Chinese Based on Ratnaparkhi's POS Tagger

Jack Zhao
Department of Linguistics
University of Illinois at Urbana-Champaign

Abstract

In this experiment, the maximum entropy model of Ratnaparkhi (1996) for part-of-speech tagging was applied to a named entity recognition task of Chinese. A named entity corpus was created from an LDC part-of-speech tagged Chinese corpus using a rule-based approach. Maximum entropy models were trained on this corpus using different sets of features. The results show that personal titles, location suffixes and verbs that roughly mean *verbally express* (*say*-verbs here after) seem to help increase the precision, but will lower the recall. The results also seem to suggest that if only history features, i.e., features before (and including) the current word, were used, personal titles, location suffixes and *say*-verbs do not seem to help that much.

1. Introduction

This experiment applies the maximum entropy model of Ratnaparkhi (1996) for English part-of-speech tagging on a named entity recognition task for Chinese. His feature set included words or tags that were two or one position preceding the current word or words that are two or one position after the current word. Because the behavior of sparsely occurring features was considered unreliable, he had a rare word feature, in which case, only the property of this rare word itself is examined. These features may be helpful for named entity recognition. So I decided to apply his model for Chinese named entity recognition. It will be seen later in this article that I have a similar feature set for the named entity recognition task for Chinese text. In addition, I tested the maximum entropy model with features that are unique to named entities. Readers are referred to his 1996 paper for mathematical details of the maximum entropy model and his set of features.

2. The corpus

In order to establish maximum entropy models for the named entity recognition task, a named entity-tagged corpus is a necessity. No such corpus was available for this experiment, but one LDC part-of-speech tagged Chinese corpus was made available through the UIUC Linguistics Department. So, I decided to convert this corpus through a rule-based approach. Up to this point, this experiment only classified personal names from location names. Organization name recognition is left for future experiment.

This is a relatively small corpus. The entire corpus has news reports between February 1996 and December 1998 of Xinhua News Agency of China and Central News Agency of Taiwan. Personal names and locations are actually identified in this corpus, but

they are not distinguished from each other. They are both tagged as NR, e. g., 张建松_NR, 中国_NR. Organizations are not tagged as such. Each component of an organization is tagged independently, e.g., 巴勒斯坦_NR 解放_NN 组织_NN. The reason why the current experiment does not deal with organizations is just because it is much harder to identify organizations from this corpus. This current experiment also does not consider numeric expressions or temporal expressions.

There are errors and inconsistencies in this corpus. For example, 新华社 is tagged both as NR and as NN for many times. The name of the former chairman of Kuo Ming Tang (the Nationalist Party) of Taiwan, 连战, is tagged as NN once, although it should have been tagged as NR.

3. The rules

Now that only personal names and location names are to be classified, we can focus on those tokens tagged as NR in the LDC part-of-speech corpus. After close inspections of the corpus, 87 rules were manually created personal name identification. These rules reflect the syntactic and morphological patterns which the context of personal names in Chinese news reports frequently follow. For example, tokens following a title such as 先生 or 夫人 are most likely personal names. Therefore the rule set includes following patterns.

(1) 先生. \s. *_NR
(2) 总统\s. *_NR
(3) . *_NR\s 说_VV
(4) 种子_NN\s. *_NR

where (1) means anything tagged as NR right after the title word 先生 is a personal name, (2) means that anything tagged as NR right after the word 总统 is a personal name, (3) means that anything tagged as NR preceding a verb 说 is a personal name and (4) means that anything tagged as NR following the regular word 种子 is a personal name.

Similarly, 63 rules were also created for location name identification. These rules make use of the frequent location suffixes that indicate rivers, lakes, seas, streets, mountains, and administrative hierarchies such as province, city and county.

3. Lexicons

The syntactic and morphological rules are usually too greedy in that they may capture an NR-tagged token as a personal name when it is an organization, or as a location when it is actually a person. For example, it is not uncommon to have structures like *Google says* in a news report. The kind of syntactic or morphological rules outlined above may mistakenly classify *Google* as a personal name whereas it is really an organization. It is certainly OK to get rid of such rules to avoid such errors, but doing so would leave many personal names unidentified. So, a few lexicons were created to narrow it down.

Location names were extracted from the dateline of a 1 Gigabytes raw corpus (i.e., neither segmented nor tagged) of news reports of Xinhua News Agency from December 1990 to September 2002. This corpus is selected because, although it is not segmented or tagged, the dateline of most of the news reports follows the following pattern, which is easy to parse.

(5) 新华社马德里 10 月 1 日电

The location names collected from such datelines were put into a location lexicon. But quite many datelines do not follow this pattern, the lexicon is manually checked and sequences that are apparently not locations were removed. However, typos are not corrected because they might be helpful in classifying mistyped words in the corpus. Dozens of more entries were manually entered into the lexicon later. At the time of this writing, the location lexicon has a total of 2398 entries. Each entry does not necessarily represent a unique location, phrases such as 纽约 and 纽约市 may both occur in the lexicon, although the former could also be picked out by a rule that uses the location suffix 市.

A lexicon of frequently used Chinese surnames was also created by using resources from the Web. Currently, 433 monosyllabic surnames and 77 disyllabic surnames (a total of 510 entries) are collected in this lexicon.

Although most surname characters are also used in regular Chinese words, some surnames are only used as personal surnames. For example, it is hard to imagine how such surnames as 冯, 廖 and 欧阳 may be used in regular Chinese phrases. One can be relatively confident that words that start with these characters are most likely personal names. So, such surnames were also put into a separate lexicon. But of course, these characters can also appear in location names such as 冯家湾.

A very small lexicon of organizations was also created. It has only 31 entries.

4. Creation of the training corpus

The NR-tagged tokens in the part-of-speech tagged Chinese corpus were classified through the set of rules and lexicons. Each target NR-tagged word is checked against the rule sets. If it matches a personal name identification rule, it suggests that it is most likely a personal name. But as I said above that these rules might be too greedy, so the target NR-tagged word is also checked against the lexicons to narrow it down.

In the first run, the classified NR-tagged proper nouns were put into two dynamic lexicons respectively, one for personal names and one for locations. These dynamic lexicons were used in the second run to classify those NR-tagged words that were not captured by the set of rules or lexicons, but do appear in the dynamic lexicons. In other words, if one mention of a token is identified as a personal name, then other mentions of the same token in the corpus are most likely personal name. The same is true for mentions of locations (Mikheev 1999).

Besides the rules and lexicons, I also made use of an empirical fact about native Chinese names (i. e. , not including transliterated foreign names). A Chinese name does not contain more than 3 characters if the surname is monosyllabic or it does not exceed 4 characters if the surname is disyllabic. And a full Chinese name has absolutely more than 1 character, whereas the length of the abbreviation of a location is quite often only 1 character long. So, the length of a token is also measured before it is classified.

Personal and location names of all 839 documents of this corpus were classified with these rules and lexicons. The revised corpus was manually inspected and errors were corrected. 40 of them were taken out as the gold standard.

5. Training maxent models

The features used for the current experiment are similar to those of Ratnaparkhi (1996) for part-of-speech tagging. Additionally, this experiment checks if the previous word is a personal title or if the next word is a *say*-verb. For rare words (a frequency of 5 by default of the application), the prefix and suffix of the word are considered and also considered is the length of this word.

The following table shows part of the features used in this named entity recognition task.

Table 1. Some features used in the maximum entropy model.

wi	The current word.
wi-1	The previous word.
wi-2	The word before the previous word.
wi-1, wi	The bigram of the previous and the current word
first-char(w_i)	The first character of the current word
last_char(w_i)	The last character of the current word
last2chars(w_i)	The last two characters of the current word.
first_char(w_{i-1})	The first character of the previous word.
last_char(w_{i-1})	The last character of the previous word.
last2chars(*wi-1*)	The last 2 characters of the previous word.
wi, wi+1	The bigram of the current word and the next word
wi+2	The word after the next word.
tagi-2, tagi-1	The tag bigram of two tokens to the left of the current word.
tagi-1, wi	The bigram of previous tag and the current word.
prevTitle	The previous word is a title.
hasLocSfx	The current word has a location suffix.
nextIsSay	The next word is a *say*-verb.

A python maximum entropy package by Zhang Le of the University of Edinburgh was used for the training task. In order to compare the effects of different sets of features, models were trained respectively on features that fall in the following sets:

Feature set 1:
Everything in Table 1.

Feature set 2:
Everything in Table 1 except personal titles, location suffixes and *say*-verbs.

Feature set 3:
No personal titles, no location suffixes and no *say*-verbs. Only features before and including the current word.

Feature set 4:
Everything before and including the current word, including personal titles, location suffixes and *say*-verbs.

Feature set 5:
No personal titles, no location suffixes, no *say*-verbs. Only features after and including the current word.

These models were trained on 799 documents of the corpus. Different numbers of iterations were also tried in an attempt to find out the optimal

number of iterations. The tags of the 40 documents of the gold standard were stripped to create the test data. All of these trained models were used to tag the test data. The models tagged every segmented word in the test data, but the evaluation only considers the personal names and location names.

6. Evaluation and discussion.

The test results of the maximum entropy models trained on different sets of features and with different numbers of iterations are shown in Table 2 on the next page. The results show that the model that uses all features (Model 1) and trained with 50 iterations (The default parameter estimation algorithm is limited-memory BFGS) seems to have the best performance in terms of overall precision (93.08%). In comparison, Model 2 used every feature except personal titles, location suffixes and *say*-verbs. None of the 3 iterations (50, 150, 500) yielded models of similar performance of Model 1 with 50 iterations. So, it looks like that these features unique to named entity recognition, i.e., personal titles, location suffixes and *say*-verbs do help quite a bit in improving the precision. But they also seem to lower the recall. This is understandable, since more features will necessarily make the conditions more stringent, and thus capture tokens with better accuracy. In the mean time, they will miss many other tokens which are named entities, but do not happen to meet the stringent conditions.

It is interesting to observe that if the models were trained on features before and including the current word, personal titles, location suffixes and *say*-verbs did not seem to contribute that much to the models. Compare Model 3 and Model 4 in Table 2, and we will see that for iteration number 50, 150 and 500, Model 4, which was trained with personal titles, location suffixes and *say*-verbs, is only trivially better than Model 3 in terms of overall precision. More tests need to be run in order to see if such features unique to named entities will only contribute to the performance of the models when both history and future features are considered.

Model 2 and Model 5 have similar performance level. It is interesting to ponder on the causes. Recall that Model 2s were trained with all features except personal titles, location suffixes and *say*-verbs, yet, Model 5s were trained without such information and trained only on features after and including the current word. Plus, future features do not contain tag information, in other words, Model 5s only had future word information available and were trained with much less information than Model 2s. Their performances were so close, maybe it is because they have roughly hit the local minimum of the performance curve.

In summary, personal titles, location suffixes, and *say*-verbs do seem to boost the performance the maximum entropy models for named entity recognition. Future work should include named entity of organizations and it is interesting to compare these maximum entropy models with, for example, Hidden Markov Models.

Reference:

Berger, Adam L. , Pietra, S. A. D. , Pietra, V. J. D. *A Maximum Entropy Approach to Natural Language Processing*.

Liu, D. C. and J. Nocedal (1989), *On the limited memory BFGS method for large-scale optimization*, Math. Programming 45 (1989), pp. 503--528.

Mikheev, Andrei, Marc Moens, Claire Grover (1999)*Named Entity Recognition without Gazetteers*, Proceedings of EACL 1999

Ratnaparkhi, Adwait (1996)*A Maximum Entropy Model for Part-Of-Speech Tagging* in Proceedings of the EMNLP (Upenn).

Zhangm Le, Maximum Entropy Modeling Toolkit for Python and C++, http://homepages.inf.ed.ac.uk/s0450736/pmwiki/pmwiki.php/Main/HomePage

Table 2. Overall percision, error rate and recall in percentage of different maxent models trained on different set of features and with different number of iterations.

	Model 1			Model 2			Model 3			Model 4			Model 5		
Iterations	Precision	Error rate	Recall	Precision	Error rate	Recall	Precision	Error rate	Recall	Precision	Error rate	Recall	Precision	Error rate	Recall
30	89.85	10.15	68.43												
45	91.84	8.16	74.29												
48	91.18	8.82	74.53												
50	93.08	6.92	71.86	87.82	12.18	75.45	89.56	10.44	75.54	90.21	9.80	74.90	87.12	12.88	74.07
52	92.88	7.12	72.05												
55	92.15	7.85	73.46												
150	88.42	11.58	77.75	88.14	11.86	77.32	91.06	8.93	75.33	91.71	8.29	75.05	85.07	14.93	75.33
500	86.84	13.16	77.36	86.51	13.49	76.16	90.08	9.92	74.65	90.86	9.14	74.75	84.17	15.83	75.02

Model 1: All features

Model 2: All features except personal titles, location suffixes and say-verbs.

Model 3: No personal titles, no location suffixes and no say-verbs. Only features before and including the current word.

Model 4: Everything before and including the current word, including personal titles, location suffixes and say-verbs. No features after the current word.

Model 5: No personal titles, no location suffixes, no say-verbs. Only features after and including the current word.