Tagging Complex NEs with MaxEnt Models: Layered Structures Versus Extended Tagset

Deyi Xiong[1,2], Hongkui Yu[1,4], and Qun Liu[1,3]

[1] Institute of Computing Technology, the Chinese Academy of Sciences, Beijing
dyxiong@ict.ac.cn
[2] Graduate School of the Chinese Academy of Sciences
[3] Inst. of Computational Linguistics, Peking University, Beijing
[4] Information Science & Technology College, Beijing University of Chemical Technology, Beijing

Abstract. The paper discusses two policies for recognizing NEs with complex structures by maximum entropy models. One policy is to develop cascaded MaxEnt models at different levels. The other is to design more detailed tags with human knowledge in order to represent complex structures. The experiments on Chinese organization names recognition indicate that layered structures result in more accurate models while extended tags can not lead to positive results as expected. We empirically prove that the {*start, continue, end, unique, other*} tag set is the best tag set for NE recognition with MaxEnt models.

1 Introduction

In recent years, MaxEnt models (see [1]) or MaxEnt-based models are widely used for many NLP tasks especially for tagging sequential data (see [2] and [3]). These models have great advantages over traditional HMM models. One of these advantages, which is often emphasized, is that MaxEnt models can incorporate richer features in a well-founded fashion than HMMs do.

When tagging NEs with MaxEnt models, the common problem that NE taggers have to face is how to improve the performance of the recognition of NEs with complex structures. For many NE tagging models, good results are gained for recognizing NEs with simple structures (e.g. person names) while bad results for the recognition of those with complex structures (e.g. organization names). We think the key is to efficiently represent multilevel structures of complex NEs. To address this problem, we take special measures for complex NEs recognition from two different perspectives.

We find that complex NEs are often constituted with other simple NEs. This directly inspires us to build a multi-layer cascaded MaxEnt model as our first way to represent multilevel structures. We put simple NEs at lower levels and train a MaxEnt model for tagging them. And complex NEs are put at higher levels and another model is trained for them. Then we firstly run the model at lower levels and later the model at higher levels.

K.-Y. Su et al. (Eds.): IJCNLP 2004, LNAI 3248, pp. 537–544, 2005.

Multi-layer models, such as cascaded models, hierarchical models[1] are very popular in NLP because of their sound fitness for many hierarchically structured tasks. Many multi-layer models are developed, such as finite state cascades for partial parsing (see [10]), hierarchical HMMs for information extraction (see [6] and [14]). However, to our knowledge, this is the first attempt to use cascaded MaxEnt models for complex NEs tagging. The first reason for this may be that most other MaxEnt taggers (e.g. taggers in [2] and [3]) focus on designing complex features to gain rich knowledge since features can represent attributes at different levels of granularity of observations. However, too complex features will result in other problems such as model consistency and data sparseness (see [2]). In our cascaded MaxEnt models, features are kept simple at each level. The other reason is the error propagation existing in multi-layer models. However, in our experiments, even without any measures to control this error propagation, a significant performance improvement is gained compared with MaxEnt models without cascades.

The other method to represent hierarchical structures is to design detailed tags with human knowledge. For many tasks, such as part-of-speech tagging, parsing, fine-grained tags will lead to more accurate models (see [13]). We think traditional tagset for NEs tagging is coarse-grained to some extent, so we design more detailed tags for different classes of elements which occur in complex NEs. Our intuition was that we would see a performance gain from these extended tags. However, the results are quite the contrary. This lead to an extensive but empirical discussion of designing an appropriate tag set for a particular task which Yu et al. (see [9]) think is worthy of further investigation.

The paper is organized as follows. In Section 2 we simply introduce how a MaxEnt NE tagger works. In Section 3 we discuss the construction of muliti-layer MaxEnt models and their application in the complex NEs recognition. In Section 4 we introduce the extended tag set for Chinese organization names recognition. In the last section, we present our conclusions and future work.

2 Tagging NEs with Maximum Entropy Models

NE tagging is the problem of learning a function that maps a sequence of observations $o = (o_1, o_2, ..., o_T)$ to a NE tag sequence $t = (t_1, t_2, ...t_T)$, where each $t_i \in T$, the set of individual tags which constitute NEs or non-NEs. A MaxEnt NE tagger is constructed on the set of events $H \times T$, where H is the set of possible observation and tag contexts, or "histories", and T is the set of allowable tags. The probability of a tag t conditioned on a history h is given as follows by MaxEnt models.

$$P(t|h) = \frac{1}{Z(h)} exp(\sum_i \lambda_i f_i(h,t)) \ . \tag{1}$$

[1] Cascaded models and hierarchical models are two different typical styles of multi-layer models. The differences are discussed in Section 3.

where the functions $f_i \in \{0, 1\}$ are model features and the λ_i are model parameters or weights of the model features, and the denominator $Z(h)$ is a normalization constant which is defined as:

$$Z(h) = \sum_t exp(\sum_i \lambda_i f_i(h, t)) \ . \tag{2}$$

This exponential form of MaxEnt models can be derived by choosing a unique model with maximum entropy which satisfies a set of constraints imposed by the empirical distribution q and the model distribution p with the following form:

$$E_p[f] = E_q[f] \ . \tag{3}$$

where E represents the expectation of f under a certain distribution.

Given training data $\Omega = \{h^i, t^i\}_{i=1}^N$, MaxEnt taggers are also trained to maximize the likelihood of the training data using the model distribution p:

$$L(p) = \prod_{i=1}^{N} P(t_i|h_i) \ . \tag{4}$$

The model parameters for the distribution p can be obtained by *Generalized Iterative Scaling* (see [4]) or *Improved Iterative Scaling* (see [7]). Given an observation sequence $(o_1, o_2, ..., o_T)$, a tag sequence candidate $(t_1, t_2, ..., t_T)$ has conditional probability:

$$P(o_1, o_2, ..., o_T|t_1, t_2, ..., t_T) = \prod_{i=1}^{T} P(t_i|h_i) \ . \tag{5}$$

The best tag sequence can be found by Viterbi search or beam search introduced by Ratnaparkni (see [2]).

3 Multi-layer MaxEnt Models

The fact that complex multilevel structures often appear in NLP tasks such as sequence labeling and parsing is the direct motivation for many different multi-layer models to be developed (see [10], [11] and [14]).

3.1 Cascaded Models vs. Hierarchical Models

There are two different methods frequently used to build multi-layer models. One way is to build models layer by layer, first simple structures of level one, then a little more complex structures of level two, and so forth. The models hypothesizing complex structures are cascaded on models for simple structures. We call multi-layer models constructed by this way cascaded models. The other way for the construction of multi-layer models is to build models recursively;

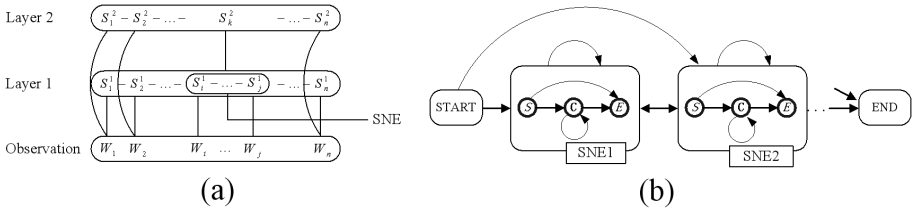

Fig. 1. Cascaded models (a) vs. hierarchical models (b). SNE means simple named entity predicted at the lower level.

bottom level models are embedded as sub-models in top level models. The multi-layer models built by this way have a hierarchical tree structure and therefore we name them hierarchical models.

Figure 1 gives a double-level cascaded model as well as a double-level hierarchical model for complex NEs recognition. The coupling between top levels and bottom levels in cascaded models is laxer than that of hierarchical models and this makes that cascaded models at different levels can be built separately and therefore flexible. Because of this flexibility, we choose cascaded MaxEnt models as our multi-layer models to represent multilevel structures in complex NEs.

3.2 Cascaded Double-Level MaxEnt Models for Chinese Organization Names Recognition

Chinese ORGs often include one or more PERs and/or LOCs on their left. For example, "中国长城工业总公司"(from MET-2 test corpus) is an organization name where "中国" and "长城" both are location names. We find that there are 698 different structures of ORG in six-month's China's People Daily (CPD) corpus(Figure 2). The most common structure is a location name followed by a noun, which accounts for 33.9% in CPD corpus. The statistical data indicate that ORGs should be recognized at a higher level than LOCs.

We design a cascaded double-level MaxEnt model (figure 1.a) for ORG recognition which has two MaxEnt models separately working at the bottom and top level. The bottom model predicts location and person names, and then the top model predicts organization names.

More specifically, Our cascaded double-level MaxEnt model works in the following steps:

1. Train a PER-LOC tagger as the bottom model with the training set where only PERs and LOCs are labeled.
2. Train a ORG tagger as the top model with the same training set where all PERs and LOCs are replaced by two tokens, separately, "*未##人*" and "*未##地*", and all ORGs are tagged.
3. When tested, the PER-LOC tagger firstly works and labels PERs as "*未##人*" and LOCs as "*未##地*".
4. Then the PER-LOC tagger takes as its input the output of the step 3 and labels ORGs in the test data.

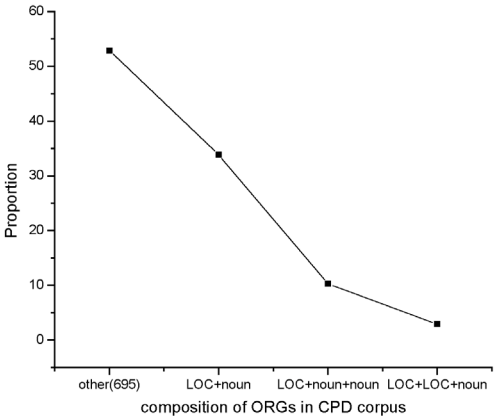

Fig. 2. Different Structures of ORG in CPD corpus and their proportions.

Of course, it is unrealistic to assume that true PER and LOC labels are provided by PER-LOC tagger, so there are some errors in the bottom model and they will be propagated to the top model. We think, however, multilevel structure knowledge represented by cascaded MaxEnt models has a greater positive effect on the performance than error propagation has the negative effect on it. Gao et al. (see [12]) used a similar approach to Chinese organization name recognition, but they didn't use Maxent models. And the other difference is that the way they treat the LOC level and ORG level is more like hierarchical models than cascaded models.

To compare our cascaded double-level MaxEnt model with the single-level MaxEnt model, we train them on one-month's (January, 1998) CPD news corpus (roughly 1,140,000 words) tagged with NER tags by Peking University, and test them on another month's (June, 1998) CPD news corpus (roughly 1,266,000 words). Both the training data and the test data are pre-segmented and converted to the annotation style used in [3]. We just use lexical features within a five-token context window and the *prevtag* features for both models so that the comparison can not be disturbed by introducing other factors. The *prevtag* features are binary functions on the previous tag and the current tag. An example of a *prevtag* feature is

$$f_j(t_i - 1, t_i) = \begin{cases} 1, \text{ if } t_{i-1} = start \text{ and } t_i = continue \\ 0, \qquad\qquad \text{otherwise} \end{cases} . \tag{6}$$

For both models, features that occur less than five times in the training data are not used, and the GIS algorithm is run for 1500 iterations. When having assigned the proper weight (λ value) to each of features, a beam search (see [2]) is used to find the best NE tag sequence for a new sentence.

The results of the two models are shown in Table 1. To assess the significance of the improvement according to F1, we use the paired samples t-test, and divide the test data into ten groups (each group contains three-day's CPD news). Each

Table 1. Performance of Single-Level Model and Double-Level Model When Tested on CPD Corpus.

Model	Total ORGs	Found ORGs	Correct ORGs	Prec.	Recall	F1
Single-Level Model	13453	8437	6893	0.817	0.512	63.0
Cascaded MaxEnt Model	13453	8808	7231	0.821	0.538	65.0

group is tested by both the single-level model and the double-level model. The p-value is less than 0.01. This indicates that layered structures are important for complex NEs tagging which are often neglected by MaxEnt taggers with a single level. Furthermore, the cascaded MaxEnt models are easy to be constructed. Taggers at higher levels can be trained on the output of lower taggers by rotated training used by [3], or directly trained on the corpus which is tagged with higher level tags.

4 Extended NE Tags

Traditional NE tags include *start*, *continue*, *end* and *unique* for each kind of NE and *other* for non-NE (see [3]). At the first, we doubt their ability to represent complex structures. Yu et al. (see [8]) design richer roles for Chinese organization names recognition with HMM-based models. In fact, these roles are special NE tags. Here we select some of Yu's roles as our extended ORG tags which are shown in Table 2.

Table 2. The extended tag set for ORG recognition.

Tags	Remarks	Examples
A	Prefix context	参与亚太经合组织的活动
C	Common tokens	北京*电影*学院
F	Translated terms	美国*摩托罗拉*公司
G	Location names	交通银行*北京*分行
I	Special tokens	*中央*电视台
J	Abbreviations	*巴*政府
D	Ends of ORGs	国务院侨务办公室
T	Unique ORGs	*新华社*
Z	Other (non-ORG)	

These tags are designed by incorporating human knowledge. For examples, the tag G indicates that its corresponding observation is a location name and the tag I shows that its observation is a token which is used to constitute ORGs very frequently. Therefore, in our intuition, we expect a performance gain from these extended tags designed with human knowledge.

We train two MaxEnt models; one is trained on the corpus labeled with traditional tag set, while the other is trained on the same corpus but labeled

with the extended tag set. Both models just incorporate lexical features within a five-token context window. This time we train the two models on CPD news corpus in June and test them on CPD news corpus in January. The results are out of our expectation, which are shown in Table 3. After careful comparison of the two tag set, we find the five tags (C, F, G, I, and J) in ORG of the extended tag set can be equivalent to *start* and *continue* in the traditional tag set in a finite-state automata. That is to say, the extended tag set is redundant for ORG recognition and therefore the probability mass of a candidate tag sequence might be lessened across redundant state transitions. Although these tags are designed with human knowledge, they violate the second criterion of Yu et al. (see [9]), in other words, the extended tag set is not efficient compared with the traditional one.

Table 3. Performance of Traditional-Tagset Model, Extended-Tagset Model & Cut-Tagset Model When Tested on CPD Corpus.

Tagset	Prec.	Recall	F1
Traditional tagset	0.865	0.651	74.3
Extended tagset	0.867	0.625	72.6
Cut tagset	0.767	0.618	68.5

Then we check the efficiency of the traditional tag set. We combine *start* and *continue* states into one state – *not-end* state because we find there are a very few features indicating the start of ORGs. The results, however, show that the cut tag set is not sufficient for ORG recognition. Furthermore, we find insufficient tag set result in a larger accuracy decrease of taggers than inefficient tag set does. All of these show that designing an appropriate (efficient and sufficient) tag set is important enough for modelers to consider.

5 Conclusions and Further Work

We think the recognition of complex NEs is one of the most difficult problems of NE tagging. However, most MaxEnt taggers do not distinguish complex NEs from simple ones and thereby not take any special measures for the recognition of complex NEs. We have shown that MaxEnt taggers can greatly benefit from layered structures when tagging complex NEs and that the traditional tag set is sufficient and efficient even for the recognition of NEs with complicated structures. The experience for designing an appropriate NE tagset is also helpful for other tagging tasks. All of our experiments are made on a large-scale test corpus and this ensures that the improvement is important.

We plan to design more representative cascaded MaxEnt models for Chinese ORGs recognition by using the probabilities of PERs or LOCs predicted by the bottom model as features on the top model in order to control the error propagation. And we want to make choosing a tag set for particular tasks automatic by incorporating the choosing mechanism into taggers.

References

1. Adam Berger, Stephen Della Pietra, and Vincent Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39-71.
2. Adwait Ratnaparkhi. 1998. A maximum entropy part-of-speech tagger. In *Proceedings of the EMNLP Conference*, pages 133-142, Philadelphia, PA.
3. Andrew Borthwick. 1999. A Maximum Entropy Approach to Named Entity Recognition. Ph.D. thesis, Computer Science Department, New York University.
4. Darroch, J.N., & Ratcliff, D. 1972. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5), 1470-1480.
5. Daniel M. Bikel, Richard Schwartz, and Ralph M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*, 34(1/2/3):211–231.
6. S. Fine, Y. Singer, and N. Tishby. 1998. The hierarchical hidden Markov model: Analysis and applications. *Machine learning*, 32:41-62, 1998.
7. Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. 1997. *Inducing features of random fields. IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:380-393.
8. YU Hong-Kui, ZHANG Hua-Ping, LIU Qun. 2003. Recognition of Chinese Organization Name Based on Role Tagging, In *Proceedings of 20th International Conference on Computer Processing of Oriental Languages*, pages 79-87, ShenYang
9. Yu Shihong, Bai Shuanhu and Wu Paul. 1998. Description of the Kent Ridge Digital Labs System Used for MUC-7. *In Proceedings of the MUC-7.*
10. Steven Abney. 1996. Partial Parsing via Finite-state Cascades. In Proceedings of the ESSLLI '96 Robust Parsing Workshop.
11. Thorsten Brants. 1999. Cascaded Markov Models. In Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics EACL-99. Bergen, Norway, 1999.
12. Jianfeng Gao, Mu Li and Chang-Ning Huang. 2003. Improved source-channel models for Chinese word segmentation. In *ACL-2003*. Sapporo, Japan, 7-12, July, 2003.
13. Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. ACL 2003, pp. 423-430.
14. M. Skounakis, M. Craven & S. Ray (2003). Hierarchical Hidden Markov Models for Information Extraction. Proceedings of the 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico. Morgan Kaufmann.