AUTOMATIC MODEL REFINEMENT — with an application to tagging

Yi-Chung Lin, Tung-Hui Chiang and Keh-Yih Su

Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan 300, Republic of China

ABSTRACT

Statistical NLP models usually only consider coarse information and very restricted context to make the estimation of parameters feasible. To reduce the modeling error introduced by a simplified probabilistic model, the Classification and Regression Tree (CART) method was adopted in this paper to select more discriminative features for automatic model refinement. Because the features are adopted dependently during splitting the classification tree in CART, the number of training data in each terminal node is small, which makes the labeling process of terminal nodes not robust. This over-tuning phenomenon cannot be completely removed by cross-validation process (i.e., pruning process). A probabilistic classification model based on the selected discriminative features is thus proposed to use the training data more efficiently. In tagging the Brown Corpus, our probabilistic classification model reduces the error rate of the top 10 error dominant words from 5.71% to 4.35%, which shows 23.82% improvement over the unrefined model.

1. INTRODUCTION

To automatically acquire knowledge from corpora, statistical methods are widely used recently (Church, 1989; Chiang, Lin & Su, 1992; Su, Chang & Lin, 1992). The performance of a probabilistic model is affected by the *estimation error* due to insufficient training data and the *modeling error* due to lacking complete knowledge of the problem to be conquered. In the literature, several smoothing methods (Good, 1953; Katz, 1987) have been used to effectively reduce the estimation error. On the contrary, the problem of reducing modeling error is less studied.

Probabilistic models are usually simplified to make the estimation of parameters feasible. However, some important information may be lost while simplifying a model. For example, using the contextual words, instead of contextual parts of speech, enhances the prediction power for tagging parts of speech. But, unfortunately, reducing the modeling error by increasing the degree of model granularity is usually accompanied by a large estimation error if there is not enough training data.

However, if only the discriminative features are involved (i.e., only those important parameters are used), modeling error could be significantly reduced without using a large corpus. Those discriminative features usually vary for different words, and it would be very time-consuming to induce such features from the corpus manually. An algorithm for automatically extracting the discriminative features from a corpus is thus highly demanded. In this paper, the *Classification and Regression Tree* (CART) method (Breiman, Friedman, Olshen & Stone, 1984) is first used to extract the discriminative features. However, CART basically regards all selected features as jointly dependent. Nodes in different branches are trained with different sets of data, and the available training data of a node becomes less and less while CART asks more and more questions. Therefore, CART can easily split and prune the classification tree to fit the training data and the cross-validation data respectively. The refinement model built by CART tends to be over-tuned and its performance is consequently not robust. A probabilistic classification model is, therefore, proposed to construct a more robust classification model. The experimental results show that this proposed model reduces the error rate of the top 10 error dominant words from 5.71% to 4.35% (23.82%

error reduction rate) while CART only reduces the error rate to 4.67% (18.21% error reduction rate).

2. PROBABILISTIC TAGGER

Since part of speech tagging plays an important role in the field of natural language processing (Church, 1989), it is used to evaluate the performance of various approaches in this paper. Tagging problem can be formulated (Church, 1989; Lin, Chiang & Su, 1992) as

$$\widehat{c}_1^n \approx \underset{c_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} P(w_i|c_i)P(c_i|c_{i-2}, c_{i-1}), \tag{1}$$

where $\widehat{c}_1^n$ is the category sequence selected by the tagging model, w_i is the i-th word, c_i is the possible corresponding category for the i-th word and c_1^n is the short-hand notation of the category sequence $c_1, c_2, \cdots, c_n$.

The Brown Corpus is used as the test bed for tagging in this paper. After preprocessing the Brown Corpus, a corpus of 1,050,004 words in 50,000 sentences is constructed. It contains 54,031 different words and 83 different tags (ignoring the four designator tags "FW," "HL," "NC" and "TL" (Francis & Kučera, 1982)). To train and test the model, the whole corpus is divided into the training set and the testing set. The *v-fold cross-validation* method (Breiman *et al.*, 1984), where v is set to 10 in this paper, is adopted to reduce the error in performance evaluation. The average number of words in the training sets and the testing sets are 945,004 (in 45,000 sentences) and 105,000 (in 5,000 sentences) respectively.

After applying back-off smoothing (Katz 1987) and robust learning (Lin *et al.*, 1992) on Equation (1) to reduce the estimation error, a tagger with 1.87% error rate in the testing set is then obtained. Although the error rate of overall testing set is small, many words are still with high error rates. For instance, the error rate of the word "that" is 9.08% and the error rate of the word "out" is 21.09%. To effectively improve accuracy over these words, it is suggested in this paper that the tagging model should be refined.

3. MODEL REFINEMENT

For not having enough training data, usually only coarse information and rather limited context are used in probabilistic models. Some discriminative features, therefore, may be sacrificed to make the estimation of parameters feasible. For example, compared to the tag-level contextual information used in a bigram or a trigram model, the word-level contextual information provides more prediction power for tagging parts of speech. However, even the simplest word-level contextual information (i.e., word bigram) requires a large number of parameters (about 3 billion in our task). Estimating such a large number of parameters requires a very huge corpus and is far beyond the size of the Brown Corpus. Thus, the word-level contextual information is usually abandoned.

To reduce the modeling error introduced by a simplified probabilistic model, one appealing approach is to extract only the discriminative features for those error dominant words. In this way, one can reduce the error rate without enlarging the corpus size. Different error dominant words, however, might be associated with different sets of discriminative features. To induce those discriminative features for each word from a corpus by hand is very time-consuming. Automatically acquiring those features directly from a corpus is thus highly desirable. In this section, the *Classification and Regression Tree* (CART) method (Breiman *et al.*, 1984) is adopted to automatically extract the discriminative features and resolve the lexical ambiguity.

CART, however, requires a large amount of training data and validation data, because it regards all those selected features as jointly dependent. The characteristic of being jointly dependent comes from the splitting process, which splits those children nodes only based on the data of their parent nodes. As a result, CART is easily tuned to fit the training data and validation data. Its performance is thus not robust. A probabilistic classification approach is therefore proposed to build robust refinement models with limited training data.

Table 1. Some statistics of the top 10 error dominant words.

Word	Frequency (%)	Error rate (%)	Proportion to overall errors (%)
that	0.895	9.08	4.33
out	0.170	21.09	1.91
to	2.259	1.43	1.72
as	0.603	4.97	1.60
than	0.160	17.17	1.46
more	0.188	12.74	1.28
about	0.147	15.24	1.20
for	0.785	2.77	1.17
one	0.252	8.26	1.11
little	0.068	29.30	1.06
TOTAL	5.526	5.71	16.84

3.1. The error dominant words

To select those words which are worth for model refinement, the top 10 error dominant words are ordered according to their contribution to overall errors, as listed in Table 1. The second column shows their relative frequencies in the Brown Corpus. The third column shows the error rates of those words tagged by the probabilistic tagger described in section 2. The last column shows the contribution of the errors of each word to the overall errors. The last row indicates that the top 10 error dominant words constitute 5.53% of the testing corpus and contribute 16.84% of the errors in the testing corpus. Their averaged error rate is 5.71% (i.e., the ratio of the total errors of these words to their total occurrence times in the testing corpus).

3.2. Feature selection

To reduce modeling error, more discriminative information should be incorporated in tagging. In addition to the trigram context information of lexical category, the features in Table 2 are considered to be potentially discriminative for choosing the correct lexical category of a given word.

Since the size of the parameters will be huge if all the features in Table 2 are jointly considered, it is not suitable to incorporate all of them. Actually only some of the listed features are really discriminative for a particular word. For instance, when we want to tag the word "out," we do not care whether the word behind it (i.e., the right-1 word) is "book," "money" or "win-

Table 2. The potentially discriminative features.

- The left-2, left-1, right-1 and right-2 categories (denoted as $L_{\text{catg}}(2)$, $L_{\text{catg}}(1)$, $R_{\text{catg}}(1)$ and $R_{\text{catg}}(2)$)
- The left-1 and right-1 words (denoted as $L_{\text{word}}(1)$ and $R_{\text{word}}(1)$)
- The distance from the left period (L_{period})
- The distance to the right period (R_{period})
- The distance from the nearest left noun (L_{noun})
- The distance to the nearest right noun (R_{noun})
- The distance from the nearest left verb (L_{verb})
- The distance to the nearest right verb (R_{verb})

Table 3. The improvement in the testing set after using CART as the refined word model. Value in parenthesis indicates the error rate of the validation set.

Word	Error rate of the 1st stage (%)	Error rate of using CART (%)	Reduction rate (%)
that	9.08	8.69 (7.47)	4.30 (17.73)
out	21.09	8.04 (7.13)	61.88 (66.19)
to	1.43	1.36 (1.12)	4.90 (21.68)
as	4.97	3.33 (2.82)	33.00 (43.26)
than	17.17	13.83 (11.24)	19.45 (34.54)
more	12.74	10.96 (9.35)	13.97 (26.61)
about	15.24	11.33 (9.94)	25.66 (34.78)
for	2.77	2.55 (2.28)	7.94 (17.69)
one	8.26	6.48 (5.94)	21.55 (28.09)
little	29.30	30.00 (25.16)	-2.39 (14.13)
TOTAL	5.71	4.67 (4.00)	18.21 (29.95)

dow;" we only care whether the right-1 word is "of." Thus, in this section, the CART (Breiman *et al.*, 1984) method is used to extract the really discriminative features from the feature set. The error rate criterion is adopted to measure the impurity of a node in the classification tree. For every error dominant word, its 4/5 training tokens are used to split the classification tree; the remaining 1/5 training tokens (not the testing tokens) are used to prune that tree. Then, all the questions asked by the pruned tree are considered to be the discriminative features.

3.3. CART classification model

In our task, a two-stage approach is adopted to tag parts of speech. The first stage is the probabilistic tagger described in section 2, which provides the most likely category sequence of the input sentence. The second stage consists of the refined word models of the error dominant words. In this stage, the pruned classification tree is used to re-tag the part of speech. The results in the testing set are shown in Table 3. In the table, the second column gives the error rates of the error dominant words in the first stage. The third column gives the error rates after using CART to re-tag those words, and the last column gives the corresponding reduction rates. In parenthesis it gives the performance in the validation set. The last row in Table 3 shows that the refined models built by CART can reduce the 18.21% of error rate for the 10 error dominant words. Only the performance of the word "little" deteriorated . This is due to the robustness problem between the cross-validation data and the testing data, which is induced by the rare occurrence of the discriminative features.

3.4. Probabilistic classification model

Because discriminative features are adopted dependently, CART can easily classify the training data and usually introduce the problem of over-tuning. Besides, due to the variation between the validation data and the testing data, the pruning process cannot effectively diminish the problem of over-tuning introduced while growing the classification tree. Thus, a probabilistic classification model, which uses all the features selected by CART in an independent way, is proposed in this section to robustly re-tag the lexical categories of the error dominant words.

Table 4. The 11 questions asked by the classification tree for the word "than."

- $Q_{1,1}$: $L_{\text{catg}}(2)$ = "RB" ?
- $Q_{1,2}$: $L_{\text{catg}}(2)$ = "IN" ?
- $Q_{2,1}$: $L_{\text{catg}}(1)$ = "AP" ?
- $Q_{3,1}$: $R_{\text{catg}}(1)$ = "CD" ?
- $Q_{3,2}$: $R_{\text{catg}}(1)$ = "JJ" ?
- $Q_{4,1}$: $R_{\text{catg}}(2)$ = "JJ" ?
- $Q_{5,1}$: $L_{\text{word}}(1)$ = "rather" ?
- $Q_{6,1}$: $R_{\text{word}}(1)$ = "the" ?
- $Q_{6,2}$: $R_{\text{word}}(1)$ = "with" ?
- $Q_{7,1}$: $L_{\text{period}} \leq 2$?
- $Q_{8,1}$: $L_{\text{period}} \leq 6$?

To use the probabilistic classification model, feature vectors are first constructed according to the questions asked by the pruned classification tree. Assume that the 11 questions in Table 4 are asked by the classification tree for the word "than." Every occurrence of "than" in the corpus is then accompanied by an 8-dimensional feature vector, $\vec{F} = [f_1, \ldots, f_8]$. The elements of the feature vector are obtained by the following rule.

$$f_i = \begin{cases} j, & \text{if } Q_{i,j} \text{ is true;} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Notice that $Q_{1,1}$ and $Q_{1,2}$ are merged into the same random variable because both of them ask about what the left-2 category is.

After constructing the feature vectors, the problem becomes to find a most probable category according to the given feature vector and it can be formulated as

$$\widehat{c} = \underset{c}{\operatorname{argmax}}\, P(c|f_1, \cdots, f_n), \tag{3}$$

where c is a possible tag for the word to be retagged. Assume that

$$\begin{aligned} & P(c|f_1, \cdots, f_n) \\ &= P(f_1, \cdots, f_n|c) \cdot \frac{P(c)}{P(f_1, \cdots, f_n)} \\ &\approx \prod_{i=1}^{n} P(f_i|c) \cdot \frac{P(c)}{P(f_1, \cdots, f_n)}. \end{aligned} \tag{4}$$

The probabilistic classification model (PCM) is then defined as

$$\widehat{c} \approx \underset{c}{\operatorname{argmax}} \prod_{i=1}^{n} P(f_i|c) \cdot P(c). \tag{5}$$

The estimation and learning processes of the PCM approach are generally more robust. As stated before, CART regards all selected features as jointly dependent. The available training data for a node become less as more questions are asked. On the contrary, due to the conditional independent assumption for $P(f_1, \cdots, f_n|c)$ in Equation (4), every parameter of PCM can be trained by the whole training data, and therefore, the estimation and learning processes are more robust.

Furthermore, every feature of PCM should be weighted to reflect its discriminant power because PCM regards all features of different branches in a tree as conditionally independent. Directly using these features without weighting cannot lead to good results. The weighting effect can be implicitly achieved by adaptive learning.

4. RESULTS AND DISCUSSION

After learning the model parameters (Amari, 1967; Lin *et al.*, 1992), the results of using the probabilistic classification model (PCM) are listed in Table 5. As shown in the last rows of Tables 3 and 5, the error rate of PCM is smaller than that of CART in the testing set while their error rates in the validation set are almost the same. The last row of Table 5 shows that the error rate of the 10 error dominant words is reduced from 5.71% to 4.35% (23.82% reduction rate) by refining the word models with the PCM approach.

In summary, due to dividing the features into independent groups, PCM can use the whole

Table 5. Improvement of using probabilistic classification model (PCM) as the refined word model. Value in parenthesis indicates the error rate of the validation set.

Word	Error rate of the 1st stage (%)	Error rate of using PCM (%)	Reduction rate (%)
that	9.08	7.98 (7.36)	12.11 (18.94)
out	21.09	7.60 (7.43)	63.96 (64.77)
to	1.43	1.33 (1.12)	6.99 (21.68)
as	4.97	2.69 (2.49)	45.88 (49.90)
than	17.17	13.49 (12.25)	21.43 (28.65)
more	12.74	10.15 (9.16)	20.33 (28.10)
about	15.24	10.60 (9.57)	30.45 (37.20)
for	2.77	2.39 (2.34)	13.72 (15.52)
one	8.26	6.19 (5.88)	25.06 (28.81)
little	29.30	28.66 (27.63)	2.18 (5.70)
TOTAL	5.71	4.35 (4.01)	23.82 (29.77)

training data to train every feature and hence construct a more robust refinement model. It is believed that this proposed probabilistic classification model (i.e., Equation (5)) can also be applied to other problems attacked by CART, such as voiced/voiceless stop classification and end-of-sentence detection, etc. (Riley 1989).

5. REFERENCE

Amari, S. (1967). A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, **16**, 299–307.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth Inc., Pacific Grove, California, USA.

Chiang, T. H., Lin, Y. C. & Su, K.-Y. (1992). Syntactic ambiguity resolution using a discrimination and robustness oriented adaptive learning algorithm. In *Proc. of COLING-92*, pp. 352–358. Aug 23–28 1992, Nantes, France.

Church, K. W. (1989). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the IEEE 1989 International Conference on Acoustics, Speech, and Signal Processing*, pp. 695–698. May 23–26 1989, Glasgow, U.K..

Francis, W. N. & Kučera, H. (1982). *Frequency analysis of English usage*. Houghton Mifflin Company.

Good, I. J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

Katz, S. M. (1987) Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **35**, 400–401.

Lin, Y.-C., Chiang, T.-H. & Su, K.-Y. (1992) Discrimination oriented probabilistic tagging. In *Proceedings of ROCLING V*, pp. 87–96. Sep 18–20 1992, Taipei, Taiwan, R.O.C.

Riley, M. D. (1989). Some applications of tree-based modeling to speech and language. In *Proceedings of the Speech and Natural Language Workshop*, pp. 339–352. Oct 15–18 1989, Cape Cod, Massachusetts, USA.

Su, K. Y., Chang, J. S. & Lin, Y. C. (1992). A discriminative approach for ambiguity resolution based on a semantic score function. In *Proc. of 1992 International Conference on Spoken Language Processing*, pp. 149–152. Oct 12–16 1992, Banff, Alberta, Canada.