Internal Use Only (非公開)

002

TR-S-0002

An Evaluation of Two Methods and Their Combination for Part-of-Speech Tagging

Weidong Qu Ruiqiang Zhang

Yoshinori Sagisaka

August 24, 2000

In this report, we give our comparison of two models for part-of-speech (POS) tagging on ATR treebank. One is the N-gram model and another is Maximum Entropy (ME) model. We investigate the feasibility of ME model when using different trigger types and their contribution to the model. We also give some results about their combination and some future work on this research.

Contents

1. Introduction

A large number of current languages processing systems use a part-of-speech tagger for pre-processing. The tagger assigns a (unique or ambiguous) part-of-speech tag each token in the input and passes its output to the next processing level, usually a parser. Due to the availability of large corpora, which have been manually annotated with part-of-speech information, many taggers use annotated text to train and "learn" either the probability distributions or rules and used the trained models to automatically assign part-of-speech to unseen text.

There are many approaches have been proposed to attack the problem of tagging text. Such as rule-based method, neural networks methods, finite state, and memory-based or statistical approaches. Among them, the Maximum Entropy framework and N-gram model have very strong position. According to current tagger comparisons, the ME framework seems to be the only other approach yielding comparable results to the N-grams model. It is a very interesting topic to determine the advantage of either of those models, to find their high accuracies, and to find a good combination of both.

The aim of this paper is to give a comparison of these two models on ATR treebank which are over 3000 tags in ATR Tagset, far more than the rudimentary, Upenn Tagset, and investigate the feasibility of ME model when using different trigger types and their contribution to the model. We also want to investigate the feasibility of combining the two methods to get an improving performance.

2. N-gram Models (Markov models)

The N-gram models (markov models) we use here are second order models for part-of-speech tagging. The states of the models represent tags; outputs represent the words. Transition probabilities depend on the states, thus pairs of tags. Output probabilities only depend on the most recent category. The underlying model is as fowling form:

$$\underset{t_1 .. t_T}{\arg\max} \left[\prod_{i=1}^{T} P\left(t_i \middle| t_{i-1}, t_{i-2}\right) P\left(w_i \middle| t_i\right) \right] P\left(t_{T+1} \middle| t_T\right)$$

For a give sequence of words W_1, W_2, ..., W_T of length T. $t_1,t_2,\ldots t_T$ are elements of tagset, the additional tags t_{-1},t_0,and t_{T+1} are beginning of sentence and the end of sentence markers. Using these additional markers can slightly improve tagging results.

The transition and output probabilities are estimated from a tagged corpus. We use maximum likelihood probabilities P which are derived from the relative frequencies:

$$P(t_3|t_2)=\frac{f(t_2,t_3)}{f(t_2)}$$

$$P(t_3|t_1,t_2)=\frac{f(t_1,t_2,t_3)}{f(t_1,t_2)}$$

$$P(w_3|t_3)=\frac{f(w_3,t_3)}{f(t_3)}$$

$$P(t_3)=\frac{f(t_3)}{N}$$

3. Maximum Entropy Model

The maximum Entropy Model is defined over H×T, where H is the set of possible word and tag contexts, or histories, and T is the set of allowable tags. The model's probability of a history h together with a tag t is as following form:

$$P(t|h)=\gamma\prod_{k=0}^{K}\alpha_k^{f_k(h,t)}p_0$$

Where:

t is tag we are predicting;

h is the history of **t;**

γ is a normalization coefficient

α is the weight of trigger **f**

f is trigger functions have value 1 or 0;

P0 is the default-tagging model

The trigger types we use are as following form; we here in table 1 just list 5 types, and in our model there are 18 trigger types.

#	Triggering word or tag	Triggered tag
1	WR1	t
2	W0WR1	t
3	TL1	t
4	WL1	t
5	WL1WL0	t

Table 1: Some important local trigger types.

4. Combining two methods

Several methods seem currently to be in use for combing information source. One method is justly simple using the product of the probabilities of the event given each feature as the probabilities of hypothesis T by the combined evidence:

$$P(T \mid M,S) = P(T \mid M) * P(T \mid S) \qquad (1)$$

In our experiments we use this method as to combine the different information sources. From the experiments it seems that this method just improved the performance slightly. This may be caused by the simple multiply two probabilities as their combining probability. If the information sources are independent give the hypothesis T, using the Bayes' rule, from (2) and (3) we can get (4).

$$P(M,S) = P(M) * P(S) \qquad (2)$$

$$P(M,S \mid T) = P(M \mid T) * P(S \mid T) \qquad (3)$$

$$P(T \mid M,S) = \frac{P(T \mid M) * P(T \mid S)}{P(T)} \qquad (4)$$

This means that to get the correct figure we should also divide P(T|M) * P(T|S) by P(T). Even if these assumptions are not valid, the above formula (1) shows that intuitively, an extra factor proportional to the probability P(T).should be considered.

Another method to combine the information source is use a linear interpolation of the probabilities:

$$P(T \mid M,S) = \lambda_1 P(T \mid M) + \lambda_2 P(T \mid S) \qquad \lambda_1 + \lambda_2 = 1.0 \qquad (5)$$

An issue that occurred in this method is that the weights are static, and not dependent on the relative predictive power of the two information sources under some condition.

A method to attack above shortcomings or problems is that when we estimate the probability P(T|S,M) of the hypothesis T given the evidence M and S , we go by way of the posterior odds which has following form:

$$P(T\mid M,S)=\frac{O(T\mid M,S)}{1+O(T\mid M,S)} \qquad (6)$$

$$\text{Where :}\quad \mathrm{O(T\mid M,S)}=\frac{O(T\mid M)*O(T\mid S)}{O(T)}$$

$$O(A\mid B)=\frac{P(A\mid B)}{P(\neg A\mid B)}=\frac{P(A\mid B)}{1-P(A\mid B)}$$

(Pearl 1988, pp34 - 39)

The advantages of this method are firstly, it is exact under the independence assumptions. Secondly, using the odds has a stabilizing effect when none of the independence assumptions are valid. Thirdly, the impact of each of the sources of information is allowed to change dynamically on the how much distinctive power they carry.

5. Experimental results

The main objective of this work is to compare two methods and evaluation the performance of their combination. So we run our experiments under different conditions to investigate the N-gram and ME model's advantages and how about their combination.

5.1 N-gram models

Our experiment using n-gram models obtained 77.6% accurate. We also use this result as our baseline performance. Latter, we will compare their results with this performance.

5.2 ME model: Using single trigger type.

In these experiment we just use default trigger model and one type trigger to investigate contribution of the different trigger type. The results are showed in figure 1. From the figure we can see that almost all triggers can improve the performance of ME model.

5.3 ME model: Using multi-trigger types.

In these experiments, we add the trigger type one by one to investigate the impact of these trigger type. The experiment results are showed in figure 2. From the figurel, figure 2 and table 1, we can see that most useful triggers are those with word information near the word that will be predicted. These seem that the adjacent words of the being predicted word contribute more information to the model.

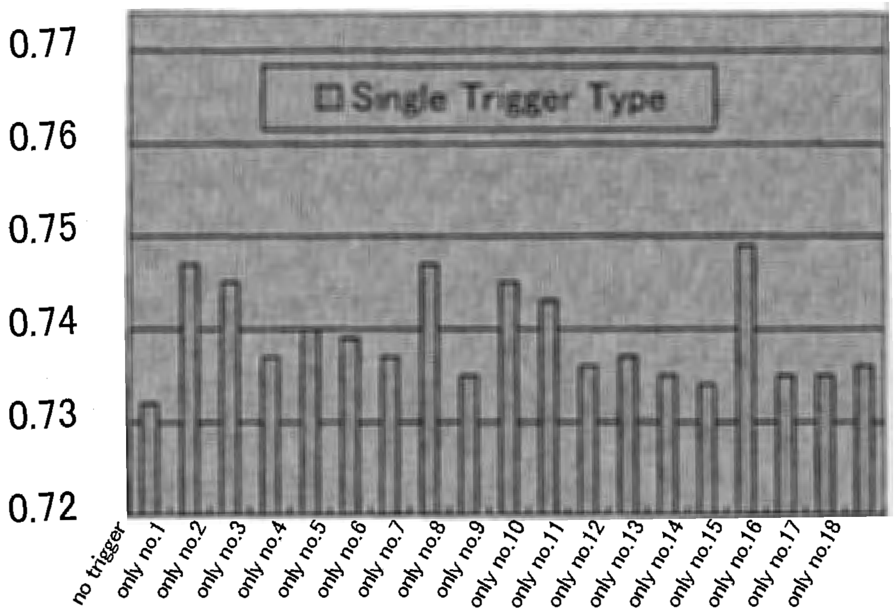

Figure 1: Using single trigger type

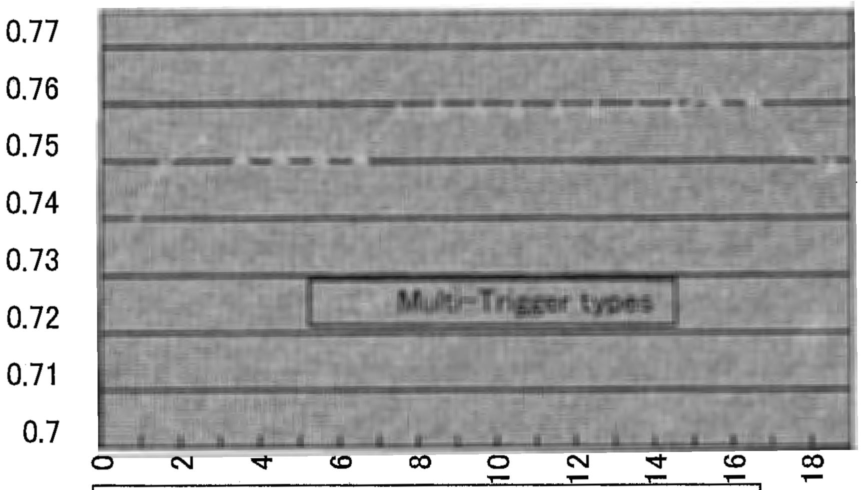

Figure 2: Using multi-trigger types

5.3 Combining the two models

In our third experiments, we investigated the feasibility of using combining methods do POS tagging. The results are as followings:

Using all trigger types, the accurate is about: 78.2 %

Using the top five trigger types according to their important, the accurate is about: 78.5 %.

These results show the combining methods are obviously improved the performance.

6. Conclusion and future work

In this report we have compare two main models of POS tagging and also investigate the feasibility of combining these two models to obtain a better accuracy. The experiments show that the two models are comparable to each other. The n-gram models are at least as well as ME framework's performance.

For the ME model, we investigate the different trigger types which contribution to the model. Our experiments showed only some local triggers improved the performance clearly; others have small contribution to the model. Long history triggers even degraded the performance of whole model.

Our combining experiments showed that combining the two methods obviously improved the accurate, but the degree seems not as well as we hopes.

From the analysis of experiments and other people's works, it seems the more accurate estimate of probability the better performance. So, future work we can use some method to improve the accurate of probability estimate.

7. Acknowledgements

The work here is on my one-month summer intern in ATR ITL. Many people help me so much during my stay here. First, many thanks go to my two supervisor Dr. Ruiqiang Zhang and Dr. Shuwu Zhang. From them, I was first introduced to the POS tagging. They supported me every step of the way. Their advice and friendship will continue to guide me in future work. I would like to thank Dr. Yoshinori Sagisaka, head of Dept. 2. His scientific intuition is the continuous subject of my admiration. I am grateful to him for advice and support through this one-month stay. Thanks to Kanokkiat Kanokphara, we enjoy many pleasure time during those days. I also received a

number of assistance from many people of ATR to whom I want to say thank you very much.

8. References

① Bernard Merialdo (1994). Tagging English Text with a Probabilistic Model. Computational Linguistics. 155-167 Volume 20, Number 2, 1994.

② Ruiqiang Zhang Ezra Black Andrew Finch Yoshinori Sagisaka. Using Detailed Contextual Models of Part-of-speech Tagging and Language Models of Speech Recognition By The Maximum Entropy Approach. ATR Technical Report. TR-IT-0334, February 14, 2000

③ Adwait Ratnaparkhi. (1996). A Maximum Entropy Model for Part-of-Speech Tagging. In Proceedings of the Conference on Empirical Methods in Nature Language Proceeding EMNLP-96, Philadelphia, PA.

④ Christer Samuelsson. 1993. Morphological tagging based entirely on Bayesian inference. In 9th Nordic Conference on Computational Linguistics NODALIDA-93, Stockholm University, Stockholm, Sweden.