2012 International Workshop on Information and Electronics Engineering (IWIEE)

Confirmation Based Self-Learning Algorithm in LVCSR's Semi-supervised Incremental Learning

Haifeng Li\(^a\)*, Tian Zhang\(^a\), Lin Ma\(^a\).

\(^a\)School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, 150001, China

Abstract

A new data selection algorithm is proposed in this paper for semi-supervised incremental learning of large vocabulary continuous speech recognition (LVCSR) system named Confirmation Based Self-Learning (CBSL). The CBSL algorithm first selects sentence level training corpus via the calculation of confidence measure, and then introduces the confirmation criterion to select word level corpus avoiding further calculation of confidence measure. It is proved that the proposed algorithm can improve the performance of acoustic model training with the highest raise by 4.42% of the recognition correctness rate in comparison with traditional single level and double level confidence measure based data selection algorithms. Besides, considering the characteristic of the distribution of high and low confidence measure data, both kinds of data are used for system training and a 1.41% increase of correctness rate is achieved by adding low confidence measure data.

© 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.

Keywords: Semi-supervised incremental learning; Data selection; Confirmation Based Self-Learning; High and low confidence measure

1. Introduction

In the field of LVCSR it is difficult to obtain large quantities of labeled data while relatively easy to get unlabeled ones. Semi-supervised learning has been proven to be an effective way against such kind of problems [1]. The algorithms for semi-supervised learning are usually classified into several classes, eg: Expectation Maximization (EM) [2], co-training [3], and incremental learning [4]. Of various kinds of semi-supervised learning algorithms incremental learning is a very useful and practical one. Some
researchers have adopted semi-supervised learning algorithms for LVCSR most of which are incremental learning ones and used confidence measure for data selection, eg: the word lattice based posterior probability and the N-best list based word posterior probability are calculated and used for confidence measure in [5] and [6] respectively, and in [7] Gillick L. et al. use linear combination of various confidence measures. All of the above algorithms are for sentence level data selection, however, in [8] Wessel F. and Ney H. employ word level filtering besides normal sentence level data selection, which unfortunately results in unstability sometimes because it will reduce the amount of training utterances compared with using sentence level confidence measure alone. This paper proposes one semi-supervised data selection method for LVCSR: the Confirmation based Self-Learning algorithm. It discards the calculation of word level confidence measure threshold which normally needs be set manually and tested for several times for a proper value, and meanwhile outperforms the sentence level algorithm. The experiments in this paper prove the validity and effectiveness of the proposed algorithm through comparing them with single level and double level confidence measure based algorithms. Moreover, considering the mismatch between real samples' distribution and model hypothesis, solely relying on the high confidence score will enlarge rather than reduce the deviation, probably to result in the decrease of the recognition correctness rate [9]. As a result high and low confidence scores are joined together to train the system in the experiments showing that adding data with low confidence score raises the correctness rate by 1.41% in comparison with only using high score data.

The rest of the paper is organized as follows: Section 2 introduces single level and double level confidence measure based data selection algorithms; the CBSL algorithm is illustrated in Section 3; and the experiments are explained in Section 4; finally the conclusion is given in Section 5.

2. Traditional Semi-Supervised Data Selection Algorithms

2.1. Confidence measure

Confidence measure is usually used for semi-supervised data selection. It is a measurement of the probability of correctness. And in speech recognition confidence measure is defined as a function evaluating the degree of matching between the model and the observation data, and the values of the function are comparable with each other. Specifically, in this paper we assume that the observation sequence is \( O = \{o_1, \ldots, o_t, \ldots, o_T\} \), the state at time \( t \) is \( s_t \), the beginning frame of the involved sentence (or word) is \( o_b \) at time \( t_b \), and the ending is \( o_e \) at time \( t_e \), and then the normalized acoustic confidence measure \( CM(O_{b,e}) \) is calculated using Equation (1):

\[
CM(O_{b,e}) = \frac{1}{t_e - t_b + 1} \sum_{t=t_b}^{t_e} P(o_t | o_t)
\]  

2.2. Single Level and Double Level Confidence Measure Based Data Selection Algorithms

Incremental learning, as a representative of traditional semi-supervised learning, usually augments the training set gradually with unlabeled data via confidence measure based data selection algorithms. For single level confidence measure based data selection, the average acoustic scores are calculated using Equation (1) and sorted in order, and then the recognized results with the highest scores exceeding some threshold are chosen. The chosen corpus is regarded as reliable, and used for model re-training iteratively. The iteration continues until no unlabeled data are left or some pre-defined number of times is reached. For double level confidence measure based data selection, the only difference with single level one is that the average acoustic scores in the word level are calculated after in the sentence level.
3. Confirmation Based Self-Learning Algorithm

The implementation of confidence measure based data selection algorithms relies heavily on the choice of the threshold, especially after adding the computation of the second level confidence score. Normally a satisfied value can only be achieved via many times of trials. Against this problem this paper proposes a new data selection algorithm: CBSL algorithm. This algorithm borrows the idea of sentence level data selection from confidence measure based algorithms for its first level selection; and the second level is based on the confirmation criterion which is defined and explained below.

**Confirmation criterion:** when lattice based N-best decoding is proceeding, multiple candidate recognition results are generated after Viterbi alignment; for the top \( n \) candidates of the N-best results, if \( n - 1 \) candidates own the same label, the label is regarded as "correct" and appended to the training data set, otherwise the label is discarded.

The algorithm is described in detail below (see Figure 1 for the flowchart of the algorithm):

**CBSL Algorithm:**

Let \( L_i \) be the present labeled training data set, \( U_i \) the unlabeled data set, \( L'_i \) the training corpus selected by the sentence level data selection algorithm, \( M_i \) the number of selected utterances, \( N_{i,m} \) the number of chosen words in the \( m \)th chosen sentence in the \( i \)th iteration, \( L''_i \) the training corpus selected by the confirmation criterion, where \( i \) is the number of iteration and \( i = 0 \) for initialisation.

- Train the system \( S_i \) with \( L_i \);
- Recognize \( U_i \) with \( S_i \) to get the decoding results;
- Calculate the confidence measure per frame of each sentence in \( U_i \) and sort them in order;
- Choose \( M_i \) sentences with the highest score to form \( L'_i \);
- Calculate the confidence measure per frame of each word within the chosen sentences and sort them in order;
- Choose \( N_{i,m} \) words in each chosen sentences of \( L'_i \) via the confirmation criterion to constitute \( L''_i \);
- \( L_{i+1} = L_i \cup L''_i \), \( U_{i+1} = U_i \setminus L''_i \); \( i = i + 1 \);
- Repeat from the first step until for some \( i \), \( U_i = \emptyset \), or the value of \( i \) exceeds some pre-set threshold.

![Fig. 1. Flowchart of CBSL algorithm](image)

Confirmation criterion is a voting method assuring to some extent the accuracy of chosen labels. Besides, the criterion makes the selection of words within sentences unrelated with the confidence measure threshold, preventing from the calculation of the second level confidence measure.
4. Experiments and Analysis

4.1. Experimental Platform and Data Preparation

In the experiment the MyPlayer system [10] is adopted as the platform. Ph97 phone set is used as the basic acoustic modeling unit [11]. Each unit is modeled by the left-to-right HMM without crossing with 3 exiting states, and the state output distribution is modeled using Gaussian mixture densities. And the 39-dimensional MFCC is adopted as the feature vectors. The training procedure is as in [12]. Token-passing algorithm [13] is adopted to generate the lattice based N-best results for decoding. The word recognition correctness rate is used for the evaluation of recognition results.

The training data are randomly selected from 863 corpus database [14]: 5100 sentences of which 5000 sentences are for training and the rest 100 for test. From the 5000 training utterances 500 randomly selected ones are used as the labeled data and for the initialization of the acoustic model. To handle the problem of the sparsity of training data and the nonuniformity of the distribution of the basic acoustic units, the chosen 500 sentences should fulfill the coverage of phones. In fact, in the experiment 87 of the chosen 500 sentences contain vM, erM and ehM which are rare in normal situation.

4.2. Experimental Results And Analysis

The 500 labeled sentences are used to train the initial acoustic model. And then semi-supervised incremental training for mandarin LVCSR is proceeding using the single level confidence measure based data selection algorithm, the double level one, and the CBSL algorithm respectively. And the performance of the systems trained with the corresponding labeled data is regarded as the upper bound of the system. The ending condition of the experiments is that the number of the chosen sentences is less than 50. Here take single level confidence measure as an example: let the training stage using the 500 labeled data be the first phase, and in the second phase 920 sentences are chosen, 292 in the third, 114 in the fourth and 66 in the fifth; the experiment ends then. As a result in total 1892 sentences are chosen for training. The experimental results are listed in Table 1 and Figure 2.

| Methods          | Phase 1 | Phase 2 | Phase 3 | Phase 4 | Phase 5 |
|------------------|---------|---------|---------|---------|---------|
| Single Level CM  | 60.65%  | 62.30%  | 62.43%  | 64.17%  | 64.55%  |
| Double Level CM  | 60.65%  | 59.87%  | 62.65%  | 61.91%  | 62.09%  |
| CBSL             | 60.65%  | 63.43%  | 63.70%  | 64.26%  | 65.07%  |
| Corresponding L  | 60.65%  | 66.83%  | 68.26%  | 68.91%  | 69.52%  |

Fig. 2. Recognition correctness rate comparison
As in Table 1 and Figure 2, the performance of double level confidence measure based algorithm fluctuates along different phases. Sometimes the correctness rate decreases. However, CBSL algorithm maintains higher correctness rate even than single level confidence measure algorithm while reducing the amount of training materials as well as double level confidence measure algorithm, which proves that CBSL algorithm is able to find the better balance point between the quantity and quality of the training data. The best result improves the recognition correctness rate by 4.42% compared with the system trained by the original 500 labeled data.

Considering the concentration of the distribution of the data with high confidence scores, which will enlarge the deviation between the model assumption and real data distribution, data with high and low confidence score are combined together to train the system together with the confirmation criterion. A series of experiments are executed and all of them express the same phenomenon, so here only lists a representative example. Train the system with 500 labeled data, and recognize the rest 4500 sentences. Choose 920 sentences of the data with the highest confidence score, and then add 300 sentences with lower score and 300 with the lowest score into the training data set respectively. Use the two data sets to train two acoustic models, and then four groups of experiments are done: 1) test set (Unseen Test); 2) 500 labeled data (L Data); 3) new added unlabeled data (New UnL Data); 4) the rest of unlabeled data (Rest UnL Data). The results are listed below:

| Training Set      | Test Set L Data | New UnL Data | Rest UnL Data | Unseen Test |
|-------------------|----------------|--------------|---------------|-------------|
| Only Use High CM  | 98.83%         | 71.82%       | 61.60%        | 62.26%      |
| High + Low CM     | 98.81%         | 66.89%       | 63.41%        | 63.67%      |

The results show that under the same condition, although the system trained with data with high and low confidence scores performs a little worse for the seen data (labeled data and new added unlabeled data) than the system trained with high confidence score data only, it does outperform for the unseen data (test set and the rest unlabeled data). As a consequence the addition of low confidence score data is helpful to improve the ability of system generalization.

5. Conclusions

A data selection strategy named CBSL algorithm is proposed which to some extent figures out the decline of the system performance caused by the dependency on the word level confidence measure, at the same time eliminates the negative effect of the in-sentence confidence threshold. The algorithm can effectively improve the performance of the acoustic model with the highest raise of the recognition correctness rate by 4.42%. As can be concluded the proposed algorithm is easy and effective to implement in real applications. Besides, through preliminary experiments it is proved that the addition of low confidence score data will raise the correctness rate over test set compared with only using high confidence score data. This reflects the potential of low confidence score data for improving the system performance and generalization, whereas how to delicately use the low confidence score data still remains an open question. The next step of work is to analyze the mathematical principle behind the algorithm. Also the underlying rules of low confidence score data are worth studying.

Acknowledgements
Our thanks to supports from the National Natural Science Foundation of China (60772076), Key Laboratory Opening Funding of MOE-Microsoft Key Laboratory of Natural Language Processing and Speech (HIT.KLOF.2009015) and Research Fund for the Doctoral Program of Higher Education No.20050213032.

References

[1] Zhu X. J.. Semi-supervised learning literature survey, Madison: Computer Sciences TR 1530 University of Wisconsin, 2007.
[2] Nigam K., McCallum A. K., S. Thrun, Mitchel T.. Text classification from labeled and unlabeled documents using EM, Machine Learning, vol.39, no. 2, pp.103–134, 2000.
[3] Wang W., Zhou Z. H.. A new analysis on co-training, In: Proceedings of the 27th International Conference on Machine Learning (ICML’10), Haifa, Israel, pp.1135-1142, 2010.
[4] Yarowsky D.. Unsupervised word sense disambiguation rivaling supervised methods, In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, pp.189-196, 1995.
[5] Fu Y. W., Chen G. P., Liu H. J.. Confidence measures based on word lattice for speech recognition (in Chinese), Computer Engineering and Applications, vol.42, no.36, pp.51-54, 2006.
[6] Rueber B.. Obtaining confidence measures from sentence probabilities, In Fifth Euro. Conf. on Speech Communication Technology, Rhode, Greece, pp.739-742, 1997.
[7] Gillick L., Ito Y., Young J.. A probabilistic approach to confidence measure estimation and evaluation, In Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Munich, Germany, pp.879-882, 1997.
[8] Wessel F., Ney H.. Unsupervised training of acoustic modelings for large vocabulary continuous speech recognition, IEEE Transactions on Speech and Audio Processing, vol.13, no.1, pp.23-31, 2005.
[9] Zhang R., Rudnicky A.. A new data selection approach for semi-supervised acoustic modeling, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2006, Pittsburgh, USA, pp.421-424, 2006.
[10] Zhang T., Li S., Gao C., Qiu R. F., Li H. F.. Audio-based digital multimedia content analysis and its visualization (in Chinese), Journal of Yanshan University, vol.34, no.2, pp.100-105, 2010.
[11] Huang C., Shi Y., Zhou J. L., Chu M., Wang T., Chang E.. Segmental tonal modeling for phone set design in Mandarin LVCSR, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.901-904, 2004.
[12] Wang H. L.. Research on confusion network and side information for speech recognition (in Chinese), Doctoral Dissertation in Harbin Institute of Technology, 2007.
[13] Young S. J., Russell N. H., Thornton J. H. S.. Token passing: a simple conceptual model for connected speech recognition systems, Cambridge University Engineering Department Technical Report CUED/F-INFENG/TR.38, July 1989.
[14] Qian Y. L, Lin S. X., Liu Q., Liu H.. A review on 2005 HTRDP (863) evaluation on Chinese information processing and intelligent human-machine interface (in Chinese), Journal of Chinese Information Processing, vol.20, Supplement, pp.1-6, 2006.