ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

ERROR IDENTIFICATION FOR LARGE VOCABULARY SPEECH RECOGNITION

Zheng-Yu ZHOU and Helen MENG

Human-Computer Communications Laboratory
Department of System Engineering & Engineering Management,
the Chinese University of Hong Kong, Hong Kong SAR, China
{zyzhou, hmmeng}@se.cuhk.edu.hk

ABSTRACT

This paper proposes two methods for identifying recognition error. The first method is a two-level schema [1]-- given the recognition hypothesis of an utterance, an utterance classifier (UC) is first applied to decide if the hypothesis is error-free or erroneous; followed by a word classifier (WC) which is applied to each word hypothesis in the erroneous utterance to decide if the word hypothesis is a misrecognition. The second method is a one-level schema in which a word classifier is applied directly to all word hypotheses to detect word recognition errors. We compare the two methods at both word and utterance levels. Experimental results show that the two methods are comparable in terms of word error detection. However, the two-level schema is very effective in filtering out error-free utterance hypotheses, which offers a key advantage to economize on word error detection.

1. INTRODUCTION

The currently prevalent language models in large-vocabulary continuous speech recognizers (LVCSR) are N-gram language models (LM) [2], partly because of its simplicity and efficiency. However, to further improve speech recognition performance, more sophisticated LM that incorporates higher level linguistic knowledge (including syntax and semantics) should be utilized [2-4], but at the expense of greater complexity and lower computational efficiency. In order to strike a balance between complexity and efficiency, we attempt to make increasing use of linguistic knowledge. We conceive of a multi-pass recognition framework: the first pass uses N-gram LM to generate *N*-best recognition hypotheses efficiently; the second pass detects possible recognition errors in the hypotheses; and a final pass applies more complex and expensive LM to error correction. This paper explores the feasibility of error detection in the second pass of the framework. Related previous work includes the rejection of erroneous word/utterance hypothesis prior to speech understanding [5-6] as well as confidence annotation in LVCSR to predict separate types of word errors [8].

This work proposes and compares a two-level schema [1] and a one-level schema to identify recognition errors in terms of both utterance and word levels. The two-level schema involves an utterance classifier (UC) in the first level and a word classifier (WC) in the second. The UC is applied to decide if the recognition hypothesis for every utterance is error-free or erroneous. In the latter case, the utterance is passed on to WC to decide whether or not it contains misrecognitions. The one-level schema directly uses the WC to identify erroneous utterances and words. How the two schemas will serve the final pass in the multi-pass recognition framework is similar: only those utterances labeled as error-containing need further processing, and efforts will be focuses on the erroneous regions pointed out by words labeled as wrong. Experiment results show that these two schemas perform similarly to find erroneous word hypotheses. However, the two-level schema outperforms one-level schema significantly in identifying error-free utterances.

The rest of the paper is organized as follows: Section 2 describes the LVCSR system with bigram LM we developed to provide N-best recognition hypotheses. Section 3 presents the utterance and word classifiers, which will be used in the error-identification schemas. Section 4 proposes the two schemas, together with experimental results and performance analysis. The conclusion is given in Section 5.

2. LVCSR

2.1. Recognizer Development

We developed a Mandarin LVCSR to generate N-best recognition hypotheses to support the current work on error identification. We first train a bigram LM by the CMU LM toolkit [9], using a 44,402-word dictionary and the Mandarin Chinese News Text corpus from LDC. This corpus includes news text from three sources, and we divide it into training/testing data sets as table 1.

Table 1: Content sources and training/testing sets

Content Source	Amount	Data Set
People Daily (news text)	282M	Training
Xinhua News (news text)	60.2M	Training
China Radio International (radio scripts)	218M	Randomly select 1M as testing data

After pruning all the bigrams with less than five occurrences, we obtained a final LM containing 267,172 bigrams and 38,483 unigrams. The test set perplexity is 233.85.

Then, we directly use the context-dependent triphone models contained in the Speech-Lab-In-A-Box [10] resource as acoustic models, and develop a word recognizer for Mandarin by the use of the HTK toolkit to combine the acoustic models with the word bigram LM.

2.2. Baseline Recognition Performance

We use the test set included in the Speech-Lab-In-A-Box (SLB) to evaluate our recognizer's performance. This test set includes 500 utterances, which are spoken by 25 speakers, with each speaker recording 20 utterances. Our recognizer achieves a test-set character accuracy of 82.1%, with 1,530 substitution errors, 159 deletion errors and 25 insertion errors.

Analysis of errors in the recognition outputs suggests that by utilizing additional linguistic knowledge, it is possible to correct the errors. For example, the utterance "虽然双方还存在着分歧…"(Although there is disagreement between the two parties…) was wrongly recognized as "孙双方还存在的分歧…" due to acoustical similarity. However, the correct recognition output was among the *N*-best hypotheses generated by the speech recognizer, and is possible to be picked out since it is the only grammatical and sensible utterance in these hypotheses. We set *N* to be 20 in all our experiments. Compared to the setting of *N*=10, using *N*=20 increases the chance of including the correct recognition hypothesis, while maintaining an acceptable computational speed. We believe that recognition errors can be corrected by utilizing more sophisticated linguistic knowledge that enforces appropriate syntactic and semantic constraints.

3. CLASSIFICATION OF UTTERANCE AND WORD RECOGNITION ERRORS

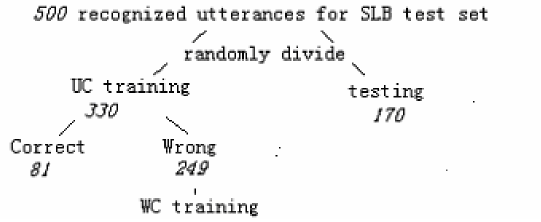
In this section, we will first introduce the organization of the data sets used to train and test the UC and WC. Then we present the utterance classifier and word classifier, which are used in the proposed error identification schemas.

3.1. Experimental Corpora

We perform all the classification experiments on the 500 test utterances in Speech-Lab-In-A-Lab (SLB test set). The SLB training utterances are not involved because we need to evaluate the classifiers' performance on unseen data. We manually label the recognition outputs for all 500 utterances at both the utterance and word levels. Utterance-level recognition hypotheses are labeled as either correct (error-free) or wrong (erroneous). Word hypotheses are labeled in the following way – if it is a substitution or insertion error, the word hypotheses is labeled wrong; if there is a deletion error, the two neighboring words are both labeled as wrong, because a deletion error may influence the statistical properties of the former word, the latter word, or both; the remaining words are labeled correct.

To train the classifiers, we randomly select 66% of the 500 utterances (330 utterances) to provide the training data, and use the remaining as the test data. All the 330 utterances are utilized to train the UC (UC training). And among the UC training utterances, those marked "wrong" are used as training data for the WC (i.e. the WC training set). The data set organization is described in Figure 1.

Figure 1: Organization of data sets for training and testing. Abbreviations include: Speech-Lab-In-A-Box (SLB), utterance classification (UC), word classification (WC).



3.2. Utterance/Word Classifiers

The objective for utterance classifier is to divide the *recognized utterances* into two classes: (i) erroneous utterances and (ii) error-free utterances, while the word classifier is to decide whether a word hypothesis is a misrecognition. For both utterance classifier and word classifier, we adopted the Support Vector Machine (SVM), not only because SVM is one of the best-performing classification algorithms provided in WEKA [11]; but also because it can be transferred into a simple linear projection model as follows:

$$r = \vec{p}^T \bullet \vec{f} + c$$

where $\vec{f}$ is the normalized feature vector, $\vec{p}$ is the projection vector, c is the threshold, and r is the confidence score. $r > 0$ either implies that an utterance contains no recognition errors or that a word hypothesis is correct. $r < 0$ implies that errors are present. This confidence score should be convenient to be incorporated into a recognition system.

The feature selection procedures for the two classifiers are similar. We first considered a set of candidate features such as acoustic scores, LM scores, combined scores, range of scores and the differences in scores between the top two recognition hypotheses at the utterance/word level. Then, we proceeded to apply a data-driven approach to refine this feature set as follows: We divided the training data (as depicted in Figure 1) into ten equal portions and conducted ten-fold cross-validation experiments. We deleted each feature one by one to see if the deletion has effect on the classification performance. If the performance is unchanged or improved, the feature will be removed from the existing feature (sub-)set.

After feature selection, the final feature set for utterance classifier is composed of 10 features, such as *Min top-choice N-best purity* (The minimum value of the *N*-best purity for each word in the top-scoring recognition hypothesis. The *N*-best purity for a word is the fraction of the *N*-best paths in which that word appears in the same position of the path), *High N-best Purity for top-scoring hypotheses* (The percentage of words in the top-scoring hypothesis with *N*-best purity above 75%.), *Mean LM score of top-scoring hypothesis* (The average value of the LM scores for the words in the top-scoring hypothesis.), and *Acoustic score span for top-scoring hypothesis* (The difference between the maximum and minimum acoustic scores of the words in the top-scoring hypothesis.[1]).

And the final feature set for word classifier has 8 features, including *N-best Purity of the word, Min LM score* (The minimum LM score among all the hypothesized words in the same position in the *N*-best hypotheses.), *Standard deviation of LM scores* (The standard derivation of LM scores across all hypothesized words in the same position in the *N*-best hypotheses.), *Number of observations* (The number of different word hypotheses appearing in the same position in the *N*-best hypotheses.), *Max Acoustic score* (The maximum acoustic score among all hypothesized words in the same position in the *N*-best hypotheses.), and so on.

4. ERROR-DETECTION SCHEMAS

4.1. Two-level Schema

The two-level schema first uses UC to filter out error-free recognized utterances from further processing, then applies WC to the rest error-containing utterances to identify erroneous word hypotheses. The basic idea of the two-level schema is that advanced linguistic knowledge such as grammar should only applied to the error-containing utterances, and efforts will be focused on the localized regions with erroneous word hypotheses, as detected by the utterance and word classifiers.

We test the two-level schema on the 170 testing utterances, and analyze the results at both utterance and word levels. We define the detection error rate as:

$$\text{detection error rate} = \frac{\text{number of incorrectly classified instances}}{\text{number of total instances}}$$

For utterance classification, an instance refers to an utterance and we obtained a 16.5% detection error rate. For word classification, an instance refers to a word and the error-rate is 16.8%. Details about the utterance and word classification are listed in Table 2. A noteworthy point is that the results presented for word classification is the overall result across all word hypotheses in the testing data. We obtain the overall word classification performance by combining the UC and WC, assuming all the words in those utterances classified as error-free by UC are classified as correct word hypotheses.

Table 2: Classification Performance in terms of *P* (Precision), *R* (Recall) and *F* (F-measure).

Kind	True Class	Classified as		*P*	*R*	*F*
		√	×			
Classify Utterances	√	25	22	0.806	0.532	0.641
	×	6	117	0.842	0.951	0.893
Classify Words	√	1499	121	0.874	0.925	0.899
	×	217	174	0.56	0.445	0.496

4.2. One-level Schema

The one-level schema directly applies the WC to all word hypotheses and does not involve UC at all. If all the word hypotheses in an utterance are classified as correct, the utterance will be labeled as error-free. Thus we can compare the two schemas at both utterance and word levels. We apply the one-level schema to all the 170 testing utterances, and label *error-free* on those utterances in which all the word hypotheses are classified as correct. The classification results for both utterances and words are presented in Table 3.

Table 3: Classification Performance in terms of *P* (Precision), *R* (Recall) and *F* (F-measure).

Kind	True Class	Classified as		*P*	*R*	*F*
		√	×			
Classify Utterances	√	16	31	0.727	0.340	0.463
	×	6	117	0.791	0.951	0.864
Classify Words	√	1489	131	0.875	0.919	0.896
	×	212	179	0.577	0.458	0.511

In this case, the utterance detection error-rate is 21.8%; while the word detection error-rate is 17.1%.

4.3. Comparison and Analysis

From Table 2 and Table 3, we can see that for word hypothesis classification, the two schemas perform similarly. The difference between the word detection error rates is only 0.3%. However, the one-level schema loses 36.1% recall rate in detecting error-free utterances when compared to the two-level schema. This suggests that the two-level schema is computationally more efficient in LVCSR.

Besides, the recall rates of erroneous word are low in both cases, only around 45%, possibly because of data

[1] We use the *normalized* acoustic score for each word, i.e. the raw acoustic score divided by the duration (in frames) of the word segment. This applies to all listed features in section 3.

sparseness. Among all the 3811 words in the training data, only 826 words are misrecognitions. In comparing with utterance classification, word classification is a more difficult task since a single word contains much less information than an utterance to make decision. We also tried to mix the utterance-level features and word-level features to train the word classifier, and found that adding utterance-level features only hurt the word classification, because the confusion the utterance-level features bring in outweighs the benefit.

We envision that in a multi-pass recognition framework, increasingly advanced linguistic knowledge will be applied in subsequent passes to correct errors detected in earlier passes. Hence the use of utterance classification helps focus successive computation on erroneous utterance and word hypotheses. This renders the two-level schema more favorable. We should also point out, however, that utterance classification is imperfect, i.e. the utterances labeled error-free may actually contain recognition errors. However, based on our experimental corpora, we found that among the 31 utterances labeled error-free by UC, there are only 8 erroneous word hypotheses and the overall character accuracy among these 31 utterances is as high as 98.1%.

An example of the usage of the two-level schema is as follows: the recognizer output "我国在信息资源的开发**商**相对落后" contains a recognition error (boldfaced). The single-character word "商" should be "上". The first UC level decided that the recognition hypothesis for this utterance contained error(s). It also means that the hypothesis will be further processed by more advanced linguistic knowledge. Then this utterance was passed to the second level in our schema, which involved the WC. The WC located that the recognition error occurred for the hypothesized word "商", due to its low value for *N-best purity* (Among the top twenty recognition hypotheses, ten include "商" and eight include "上".) This more detailed erroneous region information can be utilized when applying advanced knowledge to do error correction.

5. CONCLUSIONS

This paper advocates a multi-pass framework for LVCSR in which an increasing amount of linguistic knowledge is applied in successive passes to achieve an overall high recognition performance. As an initial step, we develop methods for detecting recognition errors in interim passes, so as to localize regions in which successive passes should dedicate computing resources for processing. We describe a two-level schema that involves an utterance classifier (UC) that attempts to detect errors in the recognition hypothesis for an input utterance. The UC is implemented with a support vector machine (SVM). An utterance-level hypothesis that is deemed erroneous will be further processed by a word classifier (WC), also based on SVM. The WC examines each word hypothesis in the utterance in an attempt to detect word-level errors. As a basis for comparison, we have also implemented a one-level schema that applies WC only and UC is bypassed. Instead, an utterance is deemed error-free if the WC does not detect any errors in the word hypotheses of the utterance. We performed experiments based on the Speech-Lab-In-A-Box corpora from Microsoft Research Asia. Results show that the two-level schema has a detection error rate of 16.5% for utterance-level misrecognitions. UC locates potentially problematic regions such that computational resources can be dedicated towards error correction in subsequent passes, e.g. detection of word-level misrecognition. Our word classifier has a detection error rate of 16.8%. In comparison, the one-level schema has a detection error rate of 21.8% for utterance-level misrecognitions and 17.1% for word-level misrecognitions. Hence the two-level schema is favored over the one-level schema.

6. ACKNOWLEDGMENTS

This work is partially supported by the Central Allocation Grant of the Research Grants Council of the Hong Kong Special Administrative Region (Project No.CUHK 1/02C). We also wish to thank Dr. Eric Chang and Microsoft Research Asia for providing us the Speech-Lab-In-A-Box data set.

7. REFERENCES

[1] Zhou, Z. & H.Meng, "A two-level schema for detecting recognition errors", *Proc. ICSLP* 2004

[2] S. Young, "Statistical modeling in continuous speech recognition", *Proc. UAI*, 2001.

[3] C. Chelba & F. Jelinek, "Structured language modeling", *Computer Speech and Language* (2000) 14, pp.283-332.

[4] S. Khudanpur & J. Wu, "Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling", *Computer Speech and Language* (2000) 14, pp. 355-372.

[5] Y. He & S. Young, "A Data-Driven Spoken Language Understanding System", *Proc. ASRU*, 2003.

[6] T. J. Hazen, S. Seneff & J. Polifroni, "Recognition confidence scoring and its use in speech understanding systems", *Computer Speech and Language* (2002) 16, pp. 49-67.

[7] C. Pao, P. Schmid & J. Glass, "Confidence scoring for speech understanding", *Proc. ICSLP*, 1998.

[8] L. Chase, "Word and acoustic confidence annotation for large vocabulary speech recognition", *Proc. Eurospeech* 1997.

[9] P. Clarkson & R. Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit", *Proc. Eurospeech* 1997.

[10] E. Chang, Y. Shi, J. Zhou & C. Huang, "Speech Lab in a Box: A Mandarin speech toolbox to jumpstart speech related research", *Proc. Eurospeech* 2001.

[11] I. Witten & E. Frank, *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.