A Two-Level Schema for Detecting Recognition Errors

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 a two-level schema for the automatic detection of possible errors in speech recognition hypotheses. Given the recognition hypothesis of an utterance, the first level in our schema applies an utterance classifier (UC) to decide if the hypothesis is error-free or erroneous. In the latter case, the utterance is passed on to the second level in our schema for further processing. A word classifier (WC) is applied to each of the word hypotheses in the utterance to decide whether or not it is a misrecognition. Hence the two-level schema can locate error-containing regions in the recognition hypotheses. These are the target regions to which we can apply more sophisticated and expensive language models for error correction as a next step. We have developed UC and WC based on Support Vector Machines (SVM). Experiments on Mandarin Chinese speech recognition using the Speech-Lab-In-A-Box corpora showed that the UC has a detection error rate of 16.5% for misrecognized utterances; the WC has a detection error rate of 19.8% for erroneous word hypotheses; and the overall two-level schema can catch 44.5% of the erroneous word hypotheses.

1. Introduction

Statistical language models [1] (especially *N*-grams) are prevalent in large-vocabulary continuous speech recognizers (LVCSR) due to the model's simplicity and computational efficiency. However, the N-gram model only captures limited linguistic constraints within a localized context. The use of more sophisticated language models that incorporate higher level linguistic knowledge (including syntax and semantics) should benefit speech recognition performance [2-4], but at the expense of greater complexity and lower computational efficiency. In order to strike a balance between complexity and efficiency, we conceive of a multi-pass recognition framework in which the first pass generates *N*-best recognition hypotheses efficiently; the second pass attempts to detect possible recognition errors in the hypotheses; and a final pass applies more complex and expensive language models to error correction. The overall objective of this framework is to make increasing use of linguistic knowledge to achieve improved speech recognition performance. This paper begins by exploring the feasibility of automatic detection of errors in recognition hypotheses based on confidence scores in recognition. Related previous work includes the utilization of confidence features to accept/reject word/utterance hypothesis prior to speech understanding [5-6] as well as confidence annotation in LVCSR for the prediction of separate types of word errors [7].

The current work proposes a two-level schema that involves an utterance classifier (UC) in the first level and a word classifier (WC) in the second. An efficient Mandarin Chinese speech recognizer is developed with a word bigram to generate *N*-best hypotheses for every test utterance. The UC is used to decide if the recognition hypothesis for every utterance is error-free or erroneous. In the latter case, the utterance is passed on to the second level for further processing by the WC. This level evaluates each word hypothesis in the utterance to decide whether or not it is a misrecognition. Hence the two-level schema can locate error-containing regions in the recognition hypotheses. These are the target regions to which we can apply more sophisticated and expensive language models for error correction as a next step.

The rest of the paper is organized as follows: Section 2 describes the LVCSR system. Section 3 presents the two-level schema for automatic error detection in recognition hypotheses, together with experimental results and performance analysis. Section 4 provides a conclusion.

2. LVCSR

In this section, we describe the LVCSR for Mandarin we built, report the experiment results, and analyze the performance of this recognizer.

2.1 Language Modeling

We use the Mandarin Chinese News Text corpus provided by LDC to build a word bigram LM. This corpus includes news text from three sources. We describe the content sources and the training/testing data sets for language modeling in the following table:

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

We performed word tokenization of the entire corpus using free LDC resources – the Chinese Segmenter and frequency dictionary. The dictionary contains 44,402 Chinese words with their corresponding frequencies and pronunciations. Then, we trained a word bigram LM using the CMU LM toolkit [8]. 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.

2.2 Recognizer Development

Context-dependent triphone models contained in the Speech-Lab-In-A-Box [9] resource is directly used as acoustic models in our LVCSR. We developed a word recognizer for Mandarin by the use of the HTK toolkit to combine the acoustic models with the word bigram LM.

2.3 Recognition Experiments

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. The given reference corresponding to these utterances are syllable strings, not character strings. We manually transcribed the utterances into character strings by listening to the waveforms and referring to the syllable strings. This generates reference character strings to evaluate our word-based recognizer.

Our recognizer achieves a test-set character accuracy of 82.1%. There are 1,530 substitution errors, 159 deletion errors and 25 insertion errors. In addition, we mapped the recognition outputs in terms of character stings into base syllable strings and tonal syllable strings respectively by using the pronunciation dictionary. This enables us to compare our recognizer's performance with other recognizers reported in [9]. These are syllable-based recognizers that use the same acoustic models and are tested on the same test data. The only difference is in the language models. One former recognizer did not utilize LM at all. The other incorporated a syllable bigram LM that is trained on the syllable strings of the SLB training utterances (which were also used for acoustic modeling). Our recognizer uses a word bigram LM. The comparison is presented in Table 2, which shows that the word bigram gave the best performance, corresponding to a character accuracy of 82.1%.

Table 2: Comparison of recognition results in terms of base syllable and tonal syllable accuracies.

LM	Recognition %Corr	
	Base Syl.	Tonal Syl.
None (from [9])	74.8	51.3
Syllable Bigram (from [9])	77.3	67.6
Word bigram (current work)	81.4	75.2

2.4 Error Analysis

Analysis of errors in the recognition outputs of our recognizer suggests that it is possible to correct the errors by utilizing additional linguistic knowledge. 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…" The recognition error "孙"is acoustically similar to "虽然". However, recognized output for the entire utterance is ungrammatical and nonsensical. The correct recognition output was among the *N*-best hypotheses generated by the speech recognizer. 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 hypotheses, while the computational speech is still acceptable. Hence we believe that recognition errors can be corrected by utilizing more sophisticated linguistic knowledge that enforces appropriate syntactic and semantic constraints.

3. Framework for Error Detection

We use a two-level framework to identify recognition errors. In the first level, an utterance classifier (UC) is applied to decide whether a transcribed utterance contains recognition errors. Only the error-containing utterances need to be further processed by a word classifier (WC), which serves to identify regions containing erroneous word hypotheses.

In this section, we will first introduce the organization of the data sets used to train and test the UC and WC. We will also present results in utterance classification, word classification as well as the overall performance of the two-level schema for recognition error detection.

3.1 Data

We use the recognition outputs for the 500 test utterances in Speech-Lab-In-A-Lab (SLB test set) to perform all the classification experiments. The SLB training utterances are not involved because we need to evaluate the classifiers' performance on unseen data. Recognition outputs for all 500 utterances were manually labeled at both the utterance and word levels. Utterance-level recognition hypotheses are labeled as either correct or wrong. Word hypotheses are labeled in the following way – if it is a substitution or insertion error, the word hypotheses is labeled wrong; but 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. Labeling for recognition errors is illustrated as follows:

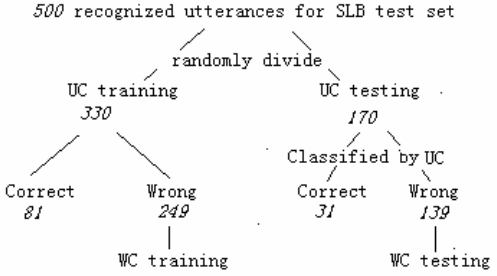
1.Substitution Error: Reference: …|信息|垃圾|的|产生|…
Hypothesis: …|信息|打击|的|产生|…
Manual labels …| √| ×|√| √|…

2.Insertion Error: Reference: …|基本|矛盾|依然|存在|…
Hypothesis: …|基本|的|矛盾|依然|存在|…
Manual labels …| √|×| √| √| √|…

3.Deletion Error. Reference: …|政治|关系|与|经济|关系|…
Hypothesis: …|政治|关系|经济|关系|…
Manual labels …| √| ×| ×| √|…

To train the utterance classifier (UC), we randomly selected 66% of the 500 utterances as the training data (UC training set), and use the remaining as the test data (UC testing set). Recall that the WC serves to identify word recognition errors in misrecognized utterances. Hence utterances in the UC training set that are marked "wrong" are used as training data for the WC (i.e. the WC training set). Similarly, utterances in the UC test set that are marked "wrong" are used as the testing data for the WC (i.e. WC test set). Organization of the various data sets 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). The numbers represent counts of utterances (in italics).



3.2 Classification Algorithm

For both the utterance and word classifiers, we adopted the Support Vector Machine (SVM) due to two reasons: (1) SVM is one of the best-performing classification algorithms provided in WEKA [10]; and (2) the classification procedure of SVM 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. It should be convenient to utilize this confident score for further processing.

3.3 Automatic Utterance Classification – Erroneous versus Error-free Recognition Hypotheses for Utterances

The first level of the proposed scheme involves automatic binary classification of *recognized utterances* into (i) erroneous utterances and (ii) error-free utterances. This utterance classifier (UC) applies the SVM algorithm. Inputs to the classifier are utterance-level features derived from the *N*-best outputs of the speech recognizer. We considered candidate features such as acoustic scores, LM scores, combined scores, range of scores and the differences in scores between the top two recognition hypotheses. We began with 16 input features that have been observed to be indicative of the absence/presence of recognition errors in a transcribed utterance. We proceeded to apply a data-driven approach to refine this input features set as follows: We divided the UC 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. This procedure resulted in 10 remaining features as listed below:

1. *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.
2. *High N-best Purity for all paths*: The percentage of words in all *N*-best paths with *N*-best purity above 75%.
3. *High N-best Purity for top-scoring hypotheses*: The percentage of words in the top-scoring hypothesis with *N*-best purity above 75%.
4. *Mean LM score of top-scoring hypothesis*: the average value of the LM scores for the words in the top-scoring hypothesis.
5. *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]
6. *Min LM score in top-scoring hypothesis:* the minimum LM score among all words in the top-scoring hypothesis.
7. *Max LM score in top-scoring hypothesis:* the maximum LM score among all words in the top-scoring hypothesis.
8. *Min acoustic score in top-scoring hypothesis:* the minimum acoustic score among all words in the top-scoring hypothesis.
9. *Total score drop*: the drop in the total score between the top two recognition hypotheses. The total score of a hypothesis is obtained by summing all acoustic and LM scores (in the log domain).
10. *Standard deviation of acoustic scores in top-scoring hypothesis*: Standard deviation across acoustic scores for all words in the top hypothesis.

Based on this final features set, we trained and tested the UC using the UC training set (330 utterances) and UC test set (170 utterances) respectively (see Figure 1). We obtained a 16.5% detection error rate for misrecognized utterances. The detection error rate is calculated 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; while for word classification, an instance refers to a word hypothesis. Table 3 presents details about the UC results.

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

True Class	Classified as		*P*	*R*	*F*
	√	×			
√	25	22	0.806	0.532	0.641
×	6	117	0.842	0.951	0.893

As a point of reference, we evaluated utterance classification results based on a single input feature. The best performance is 22.9%, obtained with *Min top-choice N-best purity* (i.e. first item in the above list). In other words, the use of a combined feature set brings a 27.9% error-rate reduction for utterance classification when compared to the use of a single feature. Table 3 shows that the UC can recall over 95% of the incorrectly recognized utterances for further processing. Utterance classification performance on the correctly recognized utterances is lower, possibly because of data sparseness.

3.4 Automatic Word Hypotheses Classification

The second level of the proposed scheme is to further process utterance-level recognition hypotheses that are labeled erroneous by the UC, in order to identify the location of misrecognized words in each utterance. Development of the word classifier (WC) is similar to that of the UC, in terms of adopting the SVM algorithm as well as the input feature selection process. 8 word-level features are selected:

1. *N-best Purity of the word.*
2. *Min LM score*: the minimum LM score among all the hypothesized words in the same position in the *N*-best hypotheses.
3. *Standard deviation of LM scores*: the standard derivation of LM scores across all hypothesized words in the same position in the *N*-best hypotheses.
4. *Mean LM score*: the mean LM score of all the hypothesized words in the same position in the *N*-best hypotheses.
5. *LM score span:* the difference between the maximum and minimum LM scores for all hypothesized words in

[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.

the same position in the *N*-best hypotheses.

6. *Number of observations*: the number of different word hypotheses appearing in the same position in the *N*-best hypotheses.
7. *Max Acoustic score:* the maximum acoustic score among all hypothesized words in the same position in the *N*-best hypotheses.
8. *Mean Acoustic score:* the mean acoustic score among all hypothesized words in the same position in the *N*-best hypotheses.

Table 4: Word Hypotheses Classification Performance in terms of *P* (Precision), *R* (Recall) and *F* (F-measure).

True Class	Classified as		P	R	F
	√	×			
√	1160	121	0.847	0.906	0.875
×	209	174	0.59	0.454	0.513

The detection error-rate for misrecognitions in word hypotheses is 19.8%. Detailed results are shown in Table 4. We evaluated word hypotheses classification results based on a single input feature. The best performance is 20.4%, obtained with *N-best Purity*. In other words, the use of a combined feature set brings a 4% error-rate reduction for utterance classification when compared to the use of a single feature. It is somewhat disappointing that the recall rate of misrecognized words is less than 50% (see Table 4), possibly due to data sparseness.

3.5 Performance Analysis for the Two-Level Schema

Our two-level schema detects recognition errors by two classifiers that examine utterance-level features and word-level features respectively. For example, the recognizer output "我国在信息资源的开发**商**相对落后" contains a recognition error (boldfaced), i.e., the single-character word "商", which should be "上". The two words are highly confusable – among the top twenty recognition hypotheses, ten include "商" and eight include "上". The first level of our schema invoked the UC, decided that the recognition hypothesis for this utterance contained error(s) due to the low value for *Min top-choice N-best Purity*. Hence 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*.

Recall from Section 3.1 that we held out 34% of the SLB test data (170 utterances) for evaluating the two-level schema. There were 31 utterances whose recognition outputs were labeled as error-free by the UC and hence no further processing was rendered. The character accuracy for these 31 utterances was 98.1%. This compares with the overall character accuracy of 82.1% for the entire SLB dataset as mentioned in Section 2.3.

Referring to Figure 1, there are 8 erroneous word hypotheses in the set of 31 utterances that are labeled correct by UC. Hence the total number of erroneous word hypotheses in the UC testing set is 391 (see Table 4). Among these errors, the two-level schema (with combined UC and WC) can catch 174 errors. Hence the two-level schema, in its entirety, can detect 44.5% of the erroneous word hypotheses for further linguistic processing and error correction.

4. Conclusions

This paper proposes a two-level schema for automatically detecting erroneous word hypotheses in the recognition outputs for Chinese utterances. In the first level, an SVM-based utterance classifier (UC) is used to decide if the recognition hypothesis for an input utterance is error-free or erroneous. In the latter case, the utterance is further processed by the second level, where an SVM-based word classifier (WC) is used to locate possibly misrecognized words in the utterance. Experiments on Mandarin Chinese speech recognition using the Speech-Lab-In-A-Box corpora showed that the UC has a detection error-rate of 16.5% for misrecognized utterances; the WC has a detection error rate of 19.8% for erroneous word hypotheses; and the overall two-level schema can catch 44.5% of the erroneous word hypotheses. These will be the target regions within which we can apply more sophisticated linguistic knowledge for error correction

5. Acknowledgments

We wish to thank Dr. Eric Chang and Microsoft Research Asia for providing the Speech-Lab-In-A-Box data set in support of this work. 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).

6. References

[1] S. Young, "Statistical Modeling in continuous speech recognition", Proc. UAI, 2001.

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

[3] 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.

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

[5] 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.*

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

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

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

[9] 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.

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