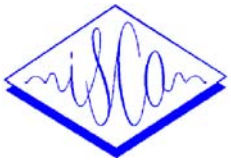

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

MIS-RECOGNIZED UTTERANCE DETECTION USING HIERARCHICAL LANGUAGE MODEL

Hirofumi Yamamoto, Genichiro Kikui and Yoshinori Sagisaka

ATR Spoken Language Translation Research Labs.
GITI Waseda Univ. 1-3-10 Nishi-Waseda Shinjuku-ku Tokyo
{hirofumi.yamamoto, genichiro.kikui, yoshinori.sagisaka}@atr.jp

ABSTRACT

In this paper, a mis-recognized utterance detection and modification scheme is proposed to recover speech recognition errors in speech translation. In a speech recognition stage, mis-recognition is frequently observed. The most of mis-recognitions result from mis-match of acoustic models and out-of-vocabulary (OOV) words. To cope with both acoustic model mis-match and OOVs, we adopt a hierarchical language model to identify them. A hierarchical language model can generate both hypotheses with and without OOVs (or acoustic mis-matched words). Likelihood difference of these hypotheses is used as utterance confidence measure. To confirm the possibility of this scheme, as a first experiment, we have conducted speech recognition experiments and mis-recognized utterance detection. Experiment results showed 99% detection rate for utterances with OOVs. This rate is considerably higher than 94% of a conventional detection method using a-posteriori probability. The rate of 80%, which is comparable to a conventional method were obtained for the utterances without OOVs. These results support the possibility of the proposed error detection and modification scheme.

1. INTRODUCTION

Recently, LVCSR based on statistical HMM acoustic models and trigram language models shows good recognition performance. However, still mis-recognition cannot be avoided for out-layered speech input. In speech translation, errors are frequently observed at the speech recognition stage, even though average word recognition accuracy is not low. Unfortunately, even one mis-recognized word in an utterance sometimes causes a fatal translation error. Considerable amount of studies have been carried out on the detection of word mis-recognition. However, most of these models use some index showing local acoustical inconsistencies such as confidence measures and do not consider linguistic mis-matches explicitly. Because of this, conventional methods for mis-recognized word detection cannot be applied when OOVs exist in input utterances. In this paper, we propose a new utterance detection and modification scheme where an elaborated language model is used to cope with mis-recognition more robustly even when OOVs exist.

2. MIS-RECOGNITION DETECTION SCHEME

We propose the following mis-recognition detection scheme. It consists of the following steps.

1. A mis-recognized utterance is detected to avoid fatal mis-translation.
2. Mis-recognized parts in the mis-recognized utterance are identified.
3. Mis-recognized parts are classified into OOVs and acoustically mis-matched words.
4. The OOVs and acoustically mis-matched words are independently processed using individual schemes.
5. The renewed recognition result is translated.

As shown in this flow, the proposed scheme tries to detect an utterance containing OOVs or acoustically mis-matched words without distinction in the first step. Since OOVs and acoustically mis-matched words cannot be separated automatically, they should be classified in the latter process that will not be discussed in this paper. Though conventional detection schemes try to find mis-matched word one by one simply looking at local acoustical inconsistencies, this scheme detect an utterance first. In the second step, mis-recognized words detection is applied sequentially. We adopt this scheme to cover mis-recognition parts that cannot be selected by word-based detection. There are quite many mis-recognitions where a succession of words showing relatively low confidence but any word cannot give sufficiently lower confidence that the detection scheme can point out. In most of these cases, at least one mis-recognized word exists in the utterance. The proposed scheme enables the detection of this type of mis-recognition, though mis-recognized parts must be identified in the latter step. The other advantage of the proposed method is in cost calculation. In this scheme, Costly precise mis-recognized word detection

should be carried only for mis-recognized utterances. In the third and fourth steps, mis-recognized words classified into OOVs and acoustic mis-match words, are differently processed, since different knowledge is required to modify them. In the final step, modified result used for speech translation.

In the proposed scheme, if mis-recognized utterance detection is fails, no modification is applied. Therefore, the first step is the most important in this scheme. In the following sections, a mis-recognized utterance detection method used in the first step and its evaluation results are described. In the first step, not only acoustic mismatched utterance but also with OOVs are detected using a hierarchical language model[1][2][3]. A hierarchical language model consists of two sub models. The first model represents OOVs' positional constraints within utterance. The second model represents OOVs' phoneme sequence constraints. As a result, OOVs are recognized in valid position in utterance with valid phoneme sequence as word. Furthermore, it can be expected that acoustic mis-matched words can be recognized as OOVs with a phoneme sequence represented by the second model in a hierarchical language model.

In the following section 3, we describe the OOV detection mechanism in a hierarchical language model and confidence measure using this model. In section 4, comparison results with conventional a-posteriori probability based method are shown.

3. UTTERANCE CONFIDENCE MEASURE USING HIERARCHICAL LANGUAGE MODEL

3.1. Hierarchical language model

In a hierarchical language model, two models are used to give constraints to OOV words as illustrated in Fig.1. One is an inter-word model that represents from in lexicon word to OOV (or from OOV to in lexicon word) transition probabilities. The other is an intra-word model that gives constraints to OOV's phoneme sequence. These models output "my / name / is / JH+OW+N+Z(OOV)" from input speech "my name is Jones", when "Jones" is OOV.

Inter-word constraints

For OOV inter-word model, class based bigrams[4] shown in next equation are used.

$$p(OOV|w_{i-1}) = p(c_{OOV}|w_{i-1})p(OOV|c_{OOV}) \quad (1)$$

In the above equation, OOV represents a particular OOV, c_{OOV} represents an OOV word class that include all of OOV.

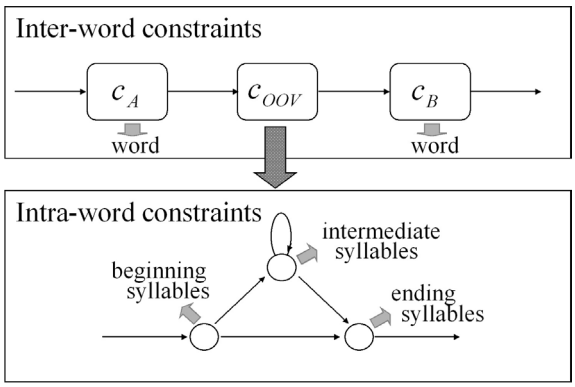

Fig. 1. A hierarchical language model
Output symbols in the upper layer and the lower layer represent word sequences and syllable sequences, respectively.

Intra-word constraints

For OOV intra-word phone sequence characteristics models, we employed a probabilistic FSA (Finite State Automaton) with three states are used. In this model, the output symbols for each state are sub-words (phonemes or syllables) or sub-word sequences.

Using this model, the occurrence probability of an OOV consisting of L-sub-word sequences $S^L = (s_1, s_2, ..., s_L)$ can be approximated by the following equation as a bigram:

$$p(OOV) = \begin{cases} p(s_{2,E}|s_{1,B}) & (\text{if } L = 2) \\ p(s_{1,B})p(s_{2,I}|s_{1,B}) & \\ \quad \prod_{i=2}^{L-1} p(s_{i,I}|s_{i-1,I})p(s_{L,E}|s_{L-1,I}) & (\text{if } L > 2), \end{cases} \quad (2)$$

where $s_{i,B}$, $s_{i,I}$ and $s_{i,E}$ represent the i-th beginning, intermediate and ending sub-words, respectively.

3.2. Mis-recognized utterance detection using a hierarchical language model

To detect mis-recognized utterance, word lattice calculated using a hierarchical language model is used. In word lattice, sometimes OOVs are included, if first best pass (recognized utterance) does not include OOV. In this case, path with OOV word exist as alternative hypotheses in the word lattice. If likelihood difference between the recognized utterance and the alternative hypothesis with OOV is small, the recognized utterance is not reliable. Therefore, this likelihood difference is used as utterance confidence measure. When the confidence measure is 0 (utterance is recognized with OOVs), the recognized utterance is the most unreliable. In fact, likelihood in alternative hypotheses can be easily calculated using forward-backward algorithm. In the example in subsection 2.1, if "my / name / is / job" is recognized as first best with likelihood 1,500, and "my / name / is

/ JH+OW+N+Z" has likelihood 1,200, utterance confidence measure of this utterance is 300 (1,500 - 1,200).

4. EXPERIMENTS

4.1. Experimental corpus

To create evaluation environments with OOVs, we used a small corpus that cannot cover large vocabulary for evaluations. This corpus is a Japanese appointment conversation corpus[6] consisting of 8,020 sentences. Furthermore, to avoid data sparseness problem in small evaluation corpus, a Japanese travel conversation corpus[5] comprising 432,639 sentences are used for language model adaptation source data. The Japanese appointment conversation corpus was split into two subsets: 5,480 sentences for language model adaptation target data and 2,540 sentences for evaluation.

The evaluation data of 2,540 sentences were also split into two subsets: with OOVs / without OOV. The set with OOVs consists of 663 sentences; each sentence contains at least one OOV for a total of 897 OOVs. The set without OOV consists of 626 sentences, where all words of a sentence are included in the vocabulary.

The with-OOVs set is used to evaluate with OOVs utterance detection performance. The without OOV set is used to confirm that the proposed method gives no bad effects to detect acoustic mis-matched utterance without OOV.

4.2. Experimental setup

Inter-word model

A travel conversation corpus consisting of 432,639 sentences was used as an adaptation source data, and the 5,480 sentences of the appointment conversation corpus was used as an adaptation target data. In creating the task dependent language model, an occurrence of word w $O_{adapt}(w)$ is estimated from $O_{source}(w)$ and $O_{target}(w)$, which are occurrences of the adaptation source data and the adaptation target data, respectively. Compensating for the data size imbalance between two corpora, occurrences in the adaptation target data are weighted by the balancing factor α (in this experiment, $\alpha = 50$). The estimated occurrence of word w is given by the following equation:

$$O_{adapt}(w) = O_{source}(w) + \alpha O_{target}(w). \qquad (3)$$

From in-lexicon-word to in-lexicon-word transition probabilities are calculated using above occurrence.

For calculating OOV's transition probabilities, we assume that all OOVs are noun. From in lexicon word to an OOV word class transition probability is given by next equation.

$$p(c_{OOV}|w_i) = \lambda \sum_{x \in noun} p(x|w_i). \qquad (4)$$

Table 1. Experimental conditions

Analysis	sampling rate: 16 kHz frame length/shift: 25 ms / 10 ms feature vector:12 MFCC+ 12 ΔMFCC+Δpower
Acoustic model	HMnet by ML-SSS[7][8] 1,400 states, 5 mixture components, gender-dependent models
Decoder[9]	1st pass: frame-synchronous Viterbi search 2nd pass: word lattice construction using FSA and rescoring

λ is calculated from the OOV rate in the evaluation set. In this experiment, λ is set to 0.12.

Intra-word model

The intra-word model consists of three states. Japanese syllables are used for sub-words and are output from each state. Furthermore, syllable sequences with occur more than 200 times in the training data are added as new output symbols. The numbers of the output symbols are 367, 253 and 409 in the beginning, intermediate and ending state, respectively. For calculation of from state to state transition probabilities, syllable sequence of nouns with less than 200 token counts in the adaptation target corpus are used as training data.

Other conditions

The other experimental conditions are shown in Table 1. On the first decoding pass, we used a hierarchical language model after converting intra-word FSA model to bigrams. Viterbi search is carried out with a simple combination of inter-word class based bigrams and an intra-word model[10]. On the second pass, FSA was used to prune the unreasonable transitions described above. Since, the N-gram implementation on the decoder, back-off smoothing gives non-zero probabilities of unreasonable transitions, such as a transition from a word to an intermediate subword.

4.3. Experimental result

Evaluation with OOVs utterance

The performance of the proposed method is compared with conventional a-posteriori probability based method[11] that shows one of the best performance in the current status. First, we evaluate utterance with OOVs detection rate in both methods. In both methods, utterances are regarded as mis-recognized utterance comparing their confidence measures to given thresholds. We set thresholds to give a best performance in the without OOV evaluation set. The result

in proposed method shows 99% detection rate, that is higher that 94% in the conventional method.

Evaluation without OOV utterance

Next, we evaluate performance in acoustic mis-matched utterance without OOV. We used two measures for evaluation. The first is correct utterance detection rate with thresholds that gives equal error rate (recall rate = precision rate). The second is mis-recognized utterance detection rate with thresholds that gives equal error rate. Baseline utterance accuracy is 54.4%. Therefore, when correct utterance recall rate is 100%, precision rate is 54.4%. When recall rate for mis-recognized utterance is 100%, precision rate is 45.6%. In the equal error rate for correct utterance detection, the proposed method results in 83% (about 28% improvement), the conventional method results in 84%. In the mis-recognized utterance detection, the proposed method results in 80%, and the conventional method results in 81%. The performance in the proposed method is almost the same as the conventional method. It is confirmed that the proposed method gives almost no bad effects for detecting mis-recognized utterance without OOV.

5. CONCLUSIONS

In this paper, a mis-recognized utterance detection and modification scheme is proposed to recover speech recognition errors in speech translation. In this scheme, not only with acoustic mis-matched words but also with OOVs utterances are detected. For this detection, hierarchical language model is used. The hierarchical language model consists of two models that represent acoustic mis-matched words' and OOVs' positional constraints in utterance and their phoneme sequence constraints. A hierarchical language model gives 1-best hypothesis and alternative hypothesis with OOVs. Likelihood difference between them is used for utterance confidence measure.

To confirm the possibility of this scheme, as a first experiment, we have conducted speech recognition experiments and mis-recognized utterance detection. Experiment results showed 99% detection rate for utterances including OOVs which is considerably higher than 94% in the conventional detection method using a-posteriori probability and 80% for the utterances without OOVs witch is comparable with the conventional method. These results support the possibility of the proposed error detection and modification scheme.

6. ACKNOWLEDMENTS

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus". This research was also partially supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research, 14380168, 2003.

7. REFERENCES

[1] K.Tanigaki, H.Yamamoto and Y.Sagisaka, "A hierarchical language model incorporating class-dependent word models for OOV words recognition," Proc. ICSLP2000, pp.123-126, 2000.

[2] S.Onishi, H.Yamamoto, Y.Sagisaka, "Structured language model for class identification of out-of-vocabulary words arising from multiple word-classes," Proc. Eurospeech2001, pp.693-696, 2001.

[3] Y.Ogawa, H.Yamamoto, Y.Sagisaka, G.Kikui, "Word class modeling for speech recognition with out-of-task words using a hierarchical language model," Proc. Eurospeech2003, 2003.

[4] S.Bai, H.Li, B.Yuan, "Building class-based language models with contextual statistics," Proc. ICASSP98, pp.173-176, 1998.

[5] T.Takezawa, E.Sumita, F.Sugaya, H.Yamamoto and S.Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world," Proc. 3rd International Conference on Language Resources and Evaluation, pp.147-152, 2002.

[6] A.Nakamura, S.Matsunaga, T.Shimizu, M.Tonomura and Y.Sagisaka, "Japanese speech database for robust speech recognition," Proc. ICSLP96, pp.2199-2202, 1996.

[7] J.Takami and S.Sagayama, "A successive state state splitting algorithm for efficient allophone modeling," Proc. ICASSP92 pp.573-576, 1992.

[8] M.Ostendorf and H.Singer, "HMM topology design using maximum likelihood successive state splitting," Computer Speech and Language,11(1), pp.17-41, 1997.

[9] T.Shimizu, H.Yamamoto, H.Masataki, S.Matsunaga and Y.Sagisaka, "Spontaneous dialogue speech recognition using cross-word context constrained word graphs," Proc. ICASSP96, pp.17-41, 1996.

[10] H. Yamamoto, S. Isogai and Y. Sagisaka. "Multi-class composite N-gram language model for spoken language processing using multiple word clusters". "Proc. of the 39th Annual meeting of the Ass. for Comp. Linguistics, pp. 531-538, 2001.

[11] F. Soong, W. Lo and S. Nakamura, "Generalized word posterior probability (GWPP) for measuring reliability of recognized words", Proc. of the 2004 Special Workshop In Maui, Hawaii, 2004.