論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Word-class models for unsupervised language model adaptation applied to spontaneous speech recognition
著者(和文)	ウィッタッカー エドワード, 古井 貞熙
Authors(English)	Luc Lussier, Edward Whittaker, Sadaoki Furui
出典(和文)	日本音響学会 2004年春季講演論文集, Vol. , No. 2-8-5, pp. 69-70
Citation(English)	, Vol. , No. 2-8-5, pp. 69-70
発行日 / Pub. date	2004, 3

Word-class models for unsupervised language model adaptation applied to spontaneous speech recognition *

◎Luc Lussier, Edward W. D. Whittaker and Sadaoki Furui (Tokyo Institute of Technology)

1 Introduction

The word error rate in spontaneous speech recognition tasks is still greater than that of read or dictation style speech and one of the reasons for this situation is the limited amount of training data available to create appropriate language models. For example, the "Corpus of Spontaneous Japanese" (CSJ) [4], one of the biggest spontaneous speech corpora at this time, contains about 7 million words of training data while the DARPA Hub 4 project has 130 million words of Broadcast News speech [1].

In order to deal with this limited amount of training data, various methods using word-class language models have been proposed by, among others, Moore and Young [5] as well as Yokoyama et al. [8]. The relatively small amount of computational overhead and good reduction in word error rate of the latter method has prompted us to further investigate its mechanism in order to better understand the source of the improvement it displays, see if it would be possible to improve on its results and propose a new method from our findings.

2 Language model adaptation

The method proposed here is based on the one used by Yokoyama et al. [8] and involves the combination of two n-gram based models, a general word n-gram language model built on the whole training corpus $\mathcal{T}$ and a word-class model whose components come from both the training corpus and the output from a first transcription hypothesis $\mathcal{H}$ of the presentation currently being recognized. The combination is performed by linear interpolation as illustrated in the following formula for a word w with history h:

$$p\,(w|h) = (1-\lambda)\cdot p_g\,(w|h) + \lambda\cdot p_{wc}\,(w|h) \qquad (1)$$

where λ is the interpolation weight between the two models, p_g is the general language model (G-LM) and p_{wc} is the word-class model (WC-LM).

The word-class model is made from 3 elements: a word-class definition, a class n-gram and a word-given-class component. This composition will be made explicit for each experiment by specifying the source of each component. For example, the model used in [8] is characterized by $(\mathcal{T}|\mathcal{H}|\mathcal{H})$ implying that the word-class definition is built from the training data and that both class n-gram and word-given-class components are computed from the hypothesis.

3 Experimental conditions

3.1 Acoustic model

The acoustic features used for the experiments are 25 dimension vectors consisting of 12 MFCC, their delta as well as the delta log energy. All the models used are gender dependent triphone HMMs with 3000 shared states and 16 Gaussian mixtures. Academic only models are used for the first and second test set and models containing both academic and extemporaneous presentations are used for the third test set. Cepstral mean subtraction is applied to each utterance.

3.2 Baseline language model

The baseline language model is built from the transcribed content of about 2590 presentations providing almost 7.5 million words of training data with a vocabulary size of 30678 words. Since the concept of word boundary is not clearly defined in Japanese, the term "word" refers to a morpheme as defined by Shinozaki and Furui [7]. Smoothing for all language models is performed using a variation of the technique developed by Kneser and Ney introduced by Goodman [2]. Also, based on empirical results obtained in [8] every word-class n-gram language models used in our experiments use 130 word-classes.

3.3 Development and evaluation test sets

We use the first of the three test sets defined in the CSJ benchmark paper by Kawahara et al. [3] as a development set and use the last two for evaluation. Each test set contains 10 presentations, test set one and two contain only academic presentations while test set three is made of extemporaneous presentations and finally, test set one contains only presentations made by male speakers as opposed to test set two and three which contain both female and male speakers in equal proportion.

4 Experimental results

Because the method involves the combination of a word and word-class n-gram and that such combination described in [6] by Ney et al. is generally expected to reduce the word error rate, a baseline that is also based on word-class models should also be used in the initial phase of our experiments in order to better evaluate the specific contribution of the different components of the studied models. Table 1 gives results for the general language model as well as for several candidate baseline results. The interpolation weight was first found using the EM algorithm for each presentation, then using the av-

---

* 話し言葉音声認識のための教師無し言語モデル適応における単語クラスモデルの検討
ルク ルシエ, エドワード ウィッテイカー, 古井 貞熙 (東工大)

Table 1. Average word error rate (%) on test set 1

Model	WER
G-LM ($\lambda = 0$)	27.67
G-LM + WC 3-gram $(\mathcal{T}\|\mathcal{T}\|\mathcal{T})$ $(\lambda = f(EM))$	27.08
G-LM + WC 3-gram $(\mathcal{T}\|\mathcal{T}\|\mathcal{T})$ $(\lambda = 0.15)$	27.12
G-LM + WC 3-gram $(\mathcal{T}\|\mathcal{T}\|\mathcal{T})$ $(\lambda = 0.30)$	**27.02**

erage of all weight values given by the EM algorithm for all presentations and then according to the value used in [8]. This last value, obtained empirically, albeit in a slightly different context, also gives the best performance in this case. The relative improvement compared to this baseline will be given in other tables in addition to the absolute word error rate.

We were then interested in verifying if the improvement simply came from interpolating the general language model with a uni-gram or word-class uni-gram built from the hypothesis. Those results are shown in Table 2 where both approaches perform only slightly better than the word-class baseline. Then, experiments were conducted with word-class tri-grams, for which results are given in Table 3, where except for the model built with the $(\mathcal{T}|\mathcal{H}|\mathcal{T})$ parameters, the two models give a good improvement.

Finally, as it appears that the methods using either $(\mathcal{T}|\mathcal{H}|\mathcal{H})$, presented in [8] or $(\mathcal{T}|\mathcal{T}|\mathcal{H})$ give the best results on the development sets, the results on test set 2 and 3 were verified. Table 4 gives the results of the general language model and of the 2 best performing methods on all test sets.

5 Conclusion

Our experiments have shown that using only a uni-gram model built on the hypothesis is not as good as combining this information, in the form of a word given class probability, with a word-class n-gram model.

Furthermore, using all of the training corpus instead of the transcription hypothesis to build the word-class n-gram component has proven beneficial on the development set but detrimental on one of the test sets which suggest a weight adjustment issue that is yet to be resolved. On the other hand, using this proposed method leads to a reduction of the required computational overhead because the word-class n-gram component only has to be computed once instead of every time a new recognition is performed.

Table 2. Average word error rate (%) for test set 1 with relative improvement from word-class baseline

G-LM + ($\lambda = 0.30$)	WER	Rel. Imp. (%)
WC 1-gram $(\mathcal{T}\|\mathcal{H}\|\mathcal{H})$	26.90	0.44
1-gram $(\mathcal{H})$	**26.74**	1.04

Table 3. Average word error rate (%) for test set 1 with relative improvement from word-class baseline

G-LM + ($\lambda = 0.30$)	WER	Rel. Imp. (%)
WC 3-gram $(\mathcal{T}\|\mathcal{H}\|\mathcal{H})$	25.22	6.66
WC 3-gram $(\mathcal{T}\|\mathcal{H}\|\mathcal{T})$	27.14	-0.44
WC 3-gram $(\mathcal{T}\|\mathcal{T}\|\mathcal{H})$	**24.89**	7.88

Table 4. Average word error rate (%) on all test sets for the G-LM and the best performing methods

Test set	G-LM	$(\mathcal{T}\|\mathcal{H}\|\mathcal{H})$	$(\mathcal{T}\|\mathcal{T}\|\mathcal{H})$
1 (dev)	27.67	25.22	24.89
2	27.05	24.81	24.80
3	25.78	24.48	24.79

Future work will involve attempts to combine the method presented in the current paper with more than one language model and also trying to find a way of automatically estimating the weight λ between the models.

References

[1] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In Aravind Joshi and Martha Palmer, editors, *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318, San Francisco, 1996. Morgan Kaufmann Publishers.

[2] J.T. Goodman. A bit of progress in language modeling. Technical report, Microsoft Research, 2001.

[3] Tatsuya Kawahara, Hiroaki Nanjo, Takahiro Shinozaki, and Sadaoki Furui. Benchmark test for speech recognition using the corpus of spontaneous Japanese. In *Proceedings SSPR*, pages 135–138, Tokyo, Japan, 2003.

[4] K. Maekawa, H. Koiso, Sadaoki Furui, and H. Isahara. Spontaneous speech corpus of Japanese. In *Proceedings of LREC*, volume 2, pages 947–952, Athens, Greece, 2000.

[5] Gareth Moore and Steve Young. Class-based language model adaptation using mixtures of word-class weights. In *Proceedings ICSLP*, 2000.

[6] H. Ney and S. Martin and F. Wessel *Corpus-based methods in language and speech processing*, chapter 6, pages 174–207. Kluwer Academic, 1997.

[7] Takahiro Shinozaki and Sadaoki Furui. Analysis on individual differences in automatic transcription of spontaneous presentations. In *Proceedings ICASSP*, volume 1, pages 729–732, 2002.

[8] Tadasuke Yokoyama, Takahiro Shinozaki, Koji Iwano, and Sadaoki Furui. Unsupervised language model adaptation using word classes for spontaneous speech recognition. In *Proceedings SSPR*, pages 71–74, Tokyo, Japan, 2003.