UNSUPERVISED LANGUAGE MODEL ADAPTATION FOR LECTURE SPEECH TRANSCRIPTION

Thomas Niesler
Department of Electrical Engineering
University of Stellenbosch
Stellenbosch, South Africa
trn@dsp.sun.ac.za

Daniel Willett
Speech Open Lab
NTT Communication Science Laboratories
NTT Corporation, Kyoto, Japan
willett@cslab.kecl.ntt.co.jp

ABSTRACT

Unsupervised adaptation methods have been applied successfully to the acoustic models of speech recognition systems for some time. Relatively little work has been carried out in the area of unsupervised language model adaptation however. The work presented here uses the output of a speech recogniser to adapt the backoff n-gram language model used in the decoding process. We report results for two different methods of language model adaptation, and find that best results are obtained when these two are used in conjunction with one another. The adaptation methods are applied to a Japanese large vocabulary transcription task, for which improvements both in perplexity and word error-rate are achieved.

1. INTRODUCTION

Speech recognition systems make use of an acoustic and a language model component. The former estimates the likelihood of the speech data given a hypothesis of the uttered word sequence, while the latter estimates the probability of the word sequence itself. Both of these components are commonly exposed to a wide variety of data during training in order to ensure good performance under a variety of test conditions. However, although the test conditions may be unknown in advance, they normally remain constant for some significant length of time. During this period both the acoustic and the language model can benefit from adaptation. For example, the speaker and the topic of discussion may remain unchanged for the length of a conversation. Hence the acoustic models can be adapted to more closely match the characteristics of the speakers voice, and the language model adapted to better model the subject and style of the conversation.

When some data from the target domain is available *a-priori*, such as a few recorded and transcribed sentences uttered by the target speaker, supervised adaptation may be performed. This style of adaptation has been shown to be successful both when applied to the acoustic model as well as the language model. When no such data is available *a-priori*, unsupervised adaptation may be appropriate. While adaptation algorithms such as MLLR have been successfully applied to the unsupervised adaptation of acoustic models, the unsupervised adaptation of the language model component has received little attention to date. Some exceptions to this may be found in [10], [11] and [13].

This paper deals with the experimental evaluation of unsupervised language model adaptation using two approaches that have been shown to perform well for supervised adaptation.

2. THE TASK

Experiments were conducted on a recognition system for recorded Japanese lecture speeches. The speech data and their transcription were provided by the Japanese national research project on Spontaneous Speech [9]. All speeches were recorded at conferences concerning speech, acoustics, linguistics and the Japanese language. In this respect the topic variety is quite limited. The acoustic data used in this work consists of 158 speeches, spoken by male and female speakers and with an approximate average length of 15 minutes. We set aside 7 speeches as a development test[1] set (dev-test) and another 7 as an evaluation test[2] set (eval-test). The specification of these sets is given in table 1.

Data set	#Speeches	Total length	#Words
Development	7	2.0h	23K
Evaluation	7	3.2h	36K
Training	144	38h	413K

Table 1: Development, evaluation and training data.

Note that the transcription of the 144 training speeches was the only source of language modeling data used in this work.

3. BASELINE LANGUAGE MODEL

The 413K words present in the reference transcription of the 144 training speeches were used to train a backoff trigram language model [3] using the CMU language modelling toolkit [1]. All experiments employed a closed vocabulary comprising the approximately 13K distinct words found in all 158 transcriptions in the lecture speech task. The trigram language model included all bigrams but excluded trigrams occurring only once. A minimum count of 12 was specified for unigrams. These parameters were determined by approximately optimising the perplexity on the reference transcription of the development-test (dev-test) set. The resulting model contains 109K bigrams and 45K trigrams, and gives a perplexity of 130.14 on the dev-test and 122.68 on the eval-test set reference transcription.

4. ACOUSTIC MODELS

A preexisting set of acoustic models that had been trained on Japanese read speech was available for the purposes of this research. Baseline acoustic models were obtained by retraining these preexisting models on the 144 speeches in the training set. This resulted in a set of tree-based state-clustered speaker-independent cross-word triphone models with 2000 states, 16 Gaussian mixtures per state and diagonal covariance matrices. The acoustic parameterisation consisted of 12 MFCCs, energy, and deltas, resulting in 26-dimensional feature vectors.

[1]The particular development test speeches are: a01f0090, a01m0070, a02f0082, a03m0045, a04m0121, a05f0039 and a06f0006.

[2]The particular evaluation test speeches are: a01m0007, a01m0035, a01m0074, a02m0117, a03m0100, a05m0031 and a06m0134.

5. SPEECH RECOGNITION ENGINE

Decoding was performed with a time-synchronous beam-search decoder that performs the Token-Passing procedure in a composition of Weighted Finite State Transducers [12]. The search is performed on each complete lecture speech in a single time-synchronous Viterbi-decoding run without incorporation of other means of segmentation.

The decoder makes use of a precompiled search network that includes the HMM structure, dictionary and the baseline unigram language model. The respective trigram deviation language models are composed on-the-fly (see [12]). In this respect, on-line transducer composition offers a convenient approach to decoding with modified language models that does not require expensive precomputation of the resulting transducer composition.

6. LANGUAGE MODEL ADAPTATION

Language model adaptation is normally performed in a supervised manner. This assumes the availability of a well-trained *background* language model together with a small amount of adaptation text from the target domain. The goal is to use this text to adapt the background model so that it will exhibit better performance on further material from the target domain. Good results have been achieved for this mode of adaptation using for example Bayes and MAP adaptation [2], linear interpolation [5] and minimum discriminative estimation [6].

Supervised adaptation requires the target domain and adaptation text to be available *a-priori*. In situations where this is not possible, unsupervised adaptation becomes attractive. Sections 6.1 and 6.2 describe the adaptation techniques used in this work.

6.1. Text selection

In order to obtain a language model more focused on the target domain, we may try to identify a subset of the training material that is in some sense closest to the target domain, and then adapt the background language model using this subset. We will achieve this by selecting from the 144 speeches in the training corpus a set of speeches judged most similar in character to the current recognition hypothesis. In order to measure the similarity between two speeches, we use an information retrieval measure known as *term frequency inverse document frequency* (tf-idf) [8]. Let there be D speeches (documents) in the training set. Denote the words of the training set vocabulary by $\{w_1, w_2, \ldots, w_V\}$, where V is the size of the vocabulary. Define the *term frequency* $tf(d_i, w_j)$ as the number of times the word w_j occurs in document d_i. Finally define the *inverse document frequency* $idf(w_j)$ to be:

$$idf(w_j) = \frac{D}{\text{number of documents containing } w_j}$$

Hence the inverse document frequency is large when the word w_j occurs in few documents. The tf-idf $\mathcal{T}(d_i, w_j)$ of document d_i and word w_j is defined by:

$$\mathcal{T}(d_i, w_j) = tf(d_i, w_j) \cdot \log\left(idf(w_j)\right)$$

The term frequency is large for frequent words, while the inverse document frequency is large for words occurring in few documents. Hence $\mathcal{T}(d_i, w_j)$ will be large when w_j occurs often in d_i but does not occur in many other documents. Such words may be expected to be good characteristics of the document d_i.

A measure of similarity $\mathcal{S}(d_i, d_k)$ of two documents d_i and d_k can now be defined:

$$\mathcal{S}(d_i, d_k) = \frac{\sum_{j=1}^{V}\left(\mathcal{T}(d_i, w_j) \cdot \mathcal{T}(d_k, w_j)\right)}{\sqrt{\left(\sum_{j=1}^{V}\mathcal{T}(d_i, w_j)^2\right) \cdot \left(\sum_{j=1}^{V}\mathcal{T}(d_k, w_j)^2\right)}}$$

If we define the vector $\mathbf{t}$ as follows:

$$\mathbf{t}(d_i) = \{\mathcal{T}(d_i, w_1), \mathcal{T}(d_i, w_2), \ldots, \mathcal{T}(d_i, w_V)\}$$

we see that the similarity may be expressed as the cosine of the angle between the vectors $\mathbf{t}(d_i)$ and $\mathbf{t}(d_k)$:

$$\mathcal{S}(d_i, d_k) = \frac{\mathbf{t}(d_i) \bullet \mathbf{t}(d_k)}{||\mathbf{t}(d_i)|| \cdot ||\mathbf{t}(d_k)||}$$

where the "$\bullet$" operator in the numerator is the vector dot product. Two documents will therefore be judged similar when corresponding words exhibit a high tf-idf. For such documents the vectors $\mathbf{t}(d_i)$ and $\mathbf{t}(d_k)$ will be directed in a similar direction, and hence the cosine of the angle between them will be close to 1.

Since $\mathcal{T}(d_i, w_j)$ is positive or zero, $\mathcal{S}(d_i, d_k)$ varies between 0 (for unrelated documents) and 1 (for highly related documents).

In order to identify the documents most closely related to the recognition hypothesis d_x, the similarity $\mathcal{S}(d_i, d_x)$ is calculated for each document d_i, $i = 1, 2, \ldots, D$. All documents for which:

$$\mathcal{S}(d_i, d_x) > \gamma \cdot S_{max} \tag{1}$$

are selected as adaptation material for the language model, where

$$S_{max} = \max_i \mathcal{S}(d_i, d_x)$$

and $0 \leq \gamma \leq 1$. When $\gamma < 1$, at least one document will be selected for use as adaptation material.

Related work has been carried out in applying text-selection methods to language model adaptation [7] and training set optimisation [4].

Linear interpolation

Once a subset of the training set has been selected by means of the tf-idf measure, this data must be used to adapt the background language model. This was achieved by building an n-gram language model from the adaptation data, and then obtaining a linear interpolation.

$$P_a(w|h) = \lambda(h) \cdot P_b(w|h) + \left(1 - \lambda(h)\right)\hat{P}_a(w|h)$$

In the above equation w indicates the word for which the probability is sought, and h the context upon which the language model will base its estimate of the probability. Then $P_b(w|h)$ is the background model, $\hat{P}_a(w|h)$ is the language model obtained from the adaptation data, and $P_a(w|h)$ is the adapted language model. The interpolation parameters $\lambda(h)$ were determined by means of the EM algorithm [5]. Since there are generally few adaptation data,

it is not possible to train interpolation parameters for each history h. Hence the set of histories were clustered according to their occurrence counts [2].

This form of linear interpolation has been shown to be a well-performing variant of MAP adaptation [2].

6.2. MDE adaptation

Minimum discriminant estimation (MDE) has been applied to supervised language model adaptation in [6]. The adaptation data is used to estimate a unigram distribution $P_a(w)$. The MDE method then finds an adapted language model $P_a(w|h)$ that is as close as possible (in the Kullback-Leibler sense) to the background language model $P_b(w|h)$ while maintaining $P_a(w)$ as its marginal distribution, i.e.:

$$\sum_h P_a(w|h) \cdot P_a(h) = P_a(w) \qquad \forall w$$

Since a closed-form solution to this problem is not available, it is normally determined iteratively by means of the Generalised Iterative Scaling (GIS) algorithm. An approximate solution is presented in [6]:

$$P_a(w|h) = \frac{\alpha(w) \cdot P_b(w|h)}{\sum_w \alpha(w) \cdot P_b(w|h)}$$

where

$$\alpha(w) = \left(\frac{P_a(w)}{P_b(w)} \right)^\beta$$

This can be shown to correspond to an approximate single iteration of the GIS algorithm. A value of $\beta = 0.5$ was taken for all our experiments, as recommended in [6].

7. EXPERIMENTAL RESULTS

The two techniques described in section 6 were applied to the lecture speech task introduced in section 2. Experimental results are presented in this section.

7.1. Adaptation by text selection

In order to evaluate language model adaptation by text selection, the algorithm described in section 6.1 was used to identify lecture speeches in the training set similar to the recognition hypothesis dev-rec0 obtained by decoding the dev-test set with the baseline language model LM0. The speeches identified in this way were used to adapt the baseline language model by means of linear interpolation as also described in section 6.1. Table 2 shows the perplexity of the adapted language model LM1 measured both on the development-test reference transcription (dev-ref) as well as the recognition hypothesis (dev-rec0) for a number of different choices of the parameter γ used in equation 1. The table shows a minimum at $\gamma = 0.35$ for the perplexity measured both on dev-ref and on dev-rec0. This strong correlation is remarkable, particularly since the high word error-rate implies that dev-ref and dev-rec0 differ significantly. The minimum is quite shallow and therefore the exact value of γ does not appear to be critical.

Table 3 shows the recognition results using $\gamma = 0.35$ as determined from table 2. Adaptation has lead to a 2.1% relative reduction in word-error rate and a 6.4% relative reduction in perplexity measured on the reference transcription (dev-ref).

Threshold	Perplexity	
γ	dev-ref	dev-rec0
0.1	124.98	102.21
0.15	123.70	101.59
0.25	122.43	101.02
0.35	**121.83**	**100.73**
0.50	122.78	101.05

Table 2: Optimisation of the threshold γ.

Language	Perplexity		WER
model	dev-ref	dev-rec0	%
LM0	130.14	107.16	33.5
LM1	121.83	100.73	32.8

Table 3: Adaptation by text selection (dev-test).

7.2. MDE adaptation

In this case, the baseline language model LM0 is adapted by MDE as described in section 6.2 using the recognition hypothesis dev-rec0 obtained from a recognition pass with LM0. This results in a new language model LM2. A further recognition experiment using LM2 yields a new recognition hypothesis dev-rec1 which is used to perform a second iteration of MDE to produce LM3. The results of this process are presented in table 4.

Language	Perplexity			WER
model	dev-ref	dev-rec0	dev-rec1	%
LM0	130.14	107.16	-	33.5
LM2	89.12	66.96	65.41	31.8
LM3	87.95	67.26	63.27	31.7

Table 4: Adaptation by MDE (dev-test).

From table 4 we see that a single iteration of MDE achieves a 5.1% relative decrease in the word error-rate and a 31.5% relative decrease in perplexity measured on the reference transcription (dev-ref). Hence the improvements are much larger than for text selection as presented in section 7.1. The second iteration of MDE adaptation achieves much smaller improvements.

7.3. Combined adaptation

Table 5 presents perplexity and recognition results when performing text selection and MDE adaptation in succession. Text selection is performed first to update the baseline language model LM0 using the recognition hypothesis dev-rec0, as in table 3. The resultant language model LM1 is then adapted by MDE, again using dev-rec0, to yield a new language model LM4. The perplexity of 86.30 and word error-rate of 31.7% are slightly better than those achieved in tables 3 and 4 by applying just one of the adaptation methods.

Another two iterations of combined adaptation were performed, and the results are included in table 5 (refer to table 7 for a key to the abbreviations used). The second iteration of text-selection followed by MDE leads to significant further improvements, while the third iteration shows no significant further gains.

Language model	Perplexity				WER %
	dev-ref	dev-rec0	dev-rec2	dev-rec3	
LM0	130.14	107.16	-		33.5
LM1	121.83	100.73	-	-	32.8
LM4	86.30	64.78	64.52	-	31.7
LM5	86.28	-	64.50	-	-
LM6	79.18	-	54.36	55.18	31.2
LM7	79.06	-	-	55.18	-
LM8	78.22	-	-	51.28	31.2

Table 5: Adaptation by text-selection and MDE (dev-test).

Overall the development-test word error rate has been improved by 6.9% relative.

Finally, table 6 shows the corresponding set of experiments applied to the evaluation-test set. Improvements are smaller than for the development-test set but show a similar tendency.

Language model	Perplexity				WER %
	eval-ref	eval-rec0	eval-rec2	eval-rec3	
LM0	122.68	92.11	-		36.9
LM9	113.62	86.54	-	-	-
LM10	89.85	62.40	62.14	-	35.8
LM11	89.72	-	62.12	-	-
LM12	86.62	-	55.23	56.23	35.7
LM13	86.30	-	-	56.16	-
LM14	88.18	-	-	53.34	35.7

Table 6: Adaptation by text-selection and MDE (eval-test).

8. SUMMARY AND CONCLUSIONS

We have evaluated two methods of unsupervised language model adaptation. Both methods were able to reduce language model perplexity as well as the recognition word error-rate for a Japanese large vocabulary transcription task. When used in conjunction with one another, further improvements were achieved.

These results are promising, especially in view of the small amount of language model training data that was available. They demonstrate the successful adaptation of the language model to the topic and style of each speaker in an unsupervised manner. The extension of these methods to larger text corpora, the incorporation of confidence measures and the combination with unsupervised acoustic model adaptation remains the subject of ongoing work.

9. ACKNOWLEDGMENT

We thank Dr. Shigeru Katagiri for supporting the authors' corporative work. The lecture speech data and transcription were provided by the Japanese Science and Technology Agency Priority Program "Spontaneous Speech: Corpus and Processing Technology".

Label	Description
LM0	Baseline (background) trigram language model.
dev-ref	Reference transcription for dev-test set.
dev-rec0	Dev recognition hypothesis using LM0.
LM1	LM0 adapted by text selection on dev-rec0.
LM2	LM0 adapted by MDE on dev-rec0.
dev-rec1	Dev recognition hypothesis using LM2.
LM3	LM2 adapted by MDE on dev-rec1.
LM4	LM1 adapted by MDE on dev-rec0.
dev-rec2	Dev recognition hypothesis using LM4.
LM5	LM4 adapted by text selection on dev-rec2.
LM6	LM5 adapted by MDE on dev-rec2.
dev-rec3	Dev Recognition hypothesis using LM6.
LM7	LM6 adapted by text selection on dev-rec3.
LM8	LM7 adapted by MDE on dev-rec3.
eval-ref	Reference transcription for eval-test set.
eval-rec0	Eval Recognition hypothesis using LM0.
LM9	LM0 adapted by text selection on eval-rec0.
LM10	LM9 adapted by MDE on eval-rec0.
eval-rec2	Eval Recognition hypothesis using LM10.
LM11	LM10 adapted by text selection on eval-rec2.
LM12	LM11 adapted by MDE on eval-rec2.
eval-rec3	Eval Recognition hypothesis using LM12.
LM13	LM12 adapted by text selection on eval-rec3.
LM14	LM13 adapted by MDE on eval-rec3.

Table 7: Legend for labels used in tables 3 to 6.

10. REFERENCES

[1] Clarkson, P; Rosenfeld, R; *Statistical language modelling using the CMU-Cambridge toolkit*, Proc. Eurospeech, Rodos, Greece, 1997.

[2] Federico, M; *Bayesian estimation methods for n-gram language models*, Proc. ICSLP, Philadelphia, 1996, pp. 240-243.

[3] Katz, S. *Estimation of probabilities from sparse data for the language model component of a speech recogniser*; IEEE Trans. ASSP, vol. 35, no. 3, pp. 400-401, March 1987.

[4] Klakow, D; *Selecting articles from the language model training corpus*, Proc. ICASSP, Istanbul, Turkey, 2000.

[5] Kneser, R; Steinbiss, V; *On the dynamic adaptation of stochastic language models*, Proc. ICASSP, Minneapolis, 1993, pp. 586-589.

[6] Kneser, R; Peters, J; Klakow, D; *Language model adaptation using dynamic marginals*, Proc. Eurospeech, Rodos, Greece, 1997, pp. 1971-1974.

[7] Mahajan, M; Beeferman, D; Huang, X. D; *Improved topic-dependent language modelling using information retrieval techniques*, Proc. ICASSP, Phoenix, Arizona, 1999.

[8] Salton, G; *Developments in automatic text retrieval*, Science, vol. 253, 1991, pp. 974-980.

[9] Shinozaki, T; Hori, C; Furui, S; *Toward automatic transcription of spontaneous speech*, Proc. Eurospeech, Aalborg, Denmark, 2001, pp. 491-494.

[10] Seymore, K; Rosenfeld, R; *Using story topics for language model adaptation*, Proc. Eurospeech, Rodos, Greece, 1997.

[11] Souvignier, B; Kellner, A; *Online adaptation for language models in spoken dialogue systems*, Proc. ICSLP, Sydney, 1998.

[12] Willett, D; Katagiri, S; *Recent advances in efficient decoding combining online transducer composition and smoothed language model incorporation*, Proc. ICASSP, Orlando, 2002.

[13] Zhu, X; Rosenfeld, R; *Improving trigram language modelling with the world wide web*, Proc. ICASSP, Salt Lake City, Utah, 2001.