Unsupervised Language Model Adaptation for Speech Recognition with no Extra Resources

Karel Beneš[1,2], Kazuki Irie[2,3], Eugen Beck[2,3], Ralf Schlüter[2], Hermann Ney[2,3]
[1] *BUT Speech@FIT, Brno, Czechia, Email: ibenes@fit.vutbr.cz*
[2] *i6-RWTH Aachen, Aachen, Germany, Email: {irie,beck,schlueter,ney}@cs.rwth-aachen.de*
[3] *AppTek, McLean, USA, http://www.apptek.com*

Abstract

Classically, automatic speech recognition (ASR) models are decomposed into acoustic models and language models (LM). LMs usually exploit the linguistic structure on a purely textual level and usually contribute strongly to an ASR systems performance. LMs are estimated on large amounts of textual data covering the target domain. However, most utterances cover more specific topics, e.g. influencing the vocabulary used. Therefore, it's desirable to have the LM adjusted to an utterance's topic. Previous work achieves this by crawling extra data from the web or by using significant amounts of previous speech data to train topic-specific LM on. We propose a way of adapting the LM directly using the target utterance to be recognized. The corresponding adaptation needs to be done in an unsupervised or automatically supervised way based on the speech input. To deal with corresponding errors robustly, we employ topic encodings from the recently proposed Subspace Multinomial Model. This model also avoids any need of explicit topic labelling during training or recognition, making the proposed method straight-forward to use. We demonstrate the performance of the method on the Librispeech corpus, which consists of read fiction books, and we discuss it's behaviour qualitatively.

Introduction

In a typical ASR system, there is a single language model estimating the prior probability of hypothesised word sequences. The parameters of such LM are optimized on a vast amount of text data, hopefully capturing as much of linguistic structure as possible. However, it can be argued that while there are global characteristics of a language, some properties of text are local to an utterance or a set of utterances. These local variations may capture speaking style, genre or topic. In this work, we focus on exploiting the topic of an utterance.

There are several prior works in the literature that focus on adapting language models: Oftentimes, features describing the context are provided as an additional input to a neural LM [3, 13, 1]. For large scale neural models, a mixture-of-experts approach was proposed recently [7]. When adapting to very small quantities of target data, a unigram description of topic of the target data is proposed to be combined with an n-gram background model [10]. In context of conversational agents, a DNN-driven mixture of n-gram LMs was proposed [15].

In most of these works, topic is considered to be a discrete category, so separate language models are trained for different topics. Every test utterance is then assigned to a topic or a mixture of topics, determining the *utterance-specific* language model to be used for recognition. In this work, we consider topic to be a continuous quantity which is intrinsic to every utterance. Therefore, we *construct* the utterance-specific language model, rather than *selecting* or *combining* it from a prepared set of LMs.

Following the practice of [10], we construct these utterance-specific LMs as unigram models. Since utterances typically consist of only tens of words, such LMs can be rather precise despite their low order. Through a series of oracle experiments, we show that this translates to very good recognition results and further improvements can be achieved by interpolation with a background model. Building on these promising results, we proceed to construct utterance-specific LMs from automatic transcriptions. After applying a topic classification model [11] inspired smoothing, we achieve a moderate, yet consistent 2 % relative improvement in word error rate on the Librispeech dataset.

Subspace Multinomial Model

Since we propose to adapt language models based on an automatic transcription, we need to mitigate the effect of errors made during the first-pass decoding. To achieve that, we propose to smooth the resulting utterance-specific LM through the Subspace Multinomial Model (SMM) [11].

An SMM models a unigram description $P_u(\cdot)$ of an utterance u in log-space:

$$P_u(w_j) = \frac{e^{\eta_{uj}}}{\sum_i e^{\eta_{ui}}} \tag{1}$$

where w_j is the j-th word of the vocabulary considered and η_{u_j} is the (un-normalized) log-probability of the word w_j.

These log-probabilities are then modelled in a low dimensional subspace:

$$\eta_u = \mathbf{m} + \mathbf{T}\mathbf{i}_u \tag{2}$$

Here, $\mathbf{m}$ is a global mean vector which can be learnt by simply computing the relative frequencies of individual words in the training corpus. Matrix $\mathbf{T}$ defines the subspace. Learning it is the core problem of training an

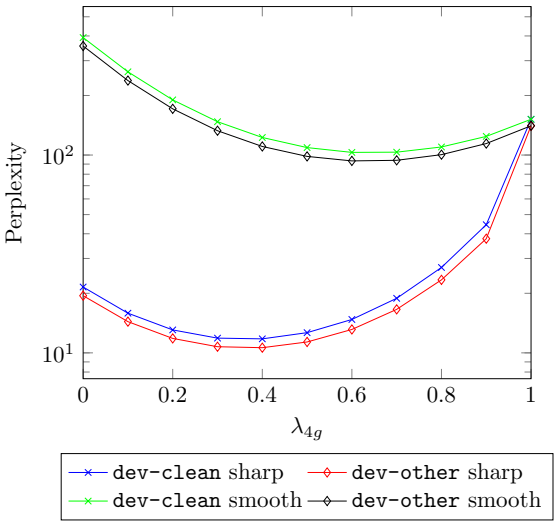

Figure 1: Perplexity of language models obtained by log-linear interpolation between the background and utterance-specific LM obtained from reference transcriptions of the individual utterances. Lower is better. The λ_{4g} term denotes the weight of the background LM, thus the convergence towards the right edge.

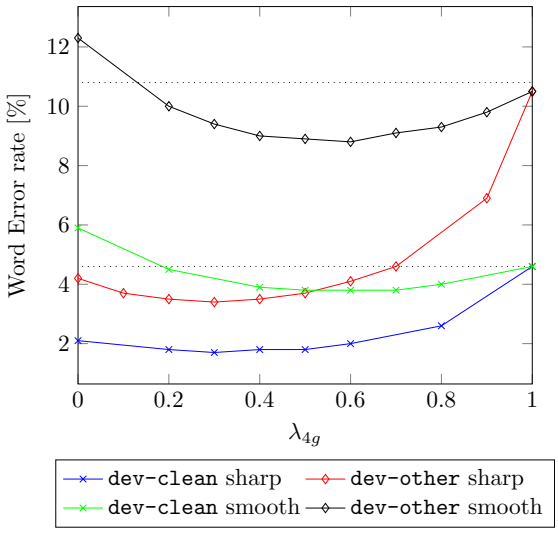

Figure 2: Word Error Rate of systems using an interpolation between the background LM and an oracle utterance-specific one. The λ_{4g} term denotes the weight of the background LM, thus the convergence towards the right edge. Interpolation between the background model and an utterance-specific one always brings improvements.

SMM; we use the l_1-SMM training procedure [8] to estimate it. Finally, $\mathbf{i}_u$ is a low dimensional description of η_u in this subspace.

We use a trained SMM to smooth an utterance-specific language model as follows: For a given unigram distribution P_f describing the first pass transcription, we find the low-dimensional representation $\mathbf{i}_f$, which fits it the best. Then the smoothed model P_s is given by composition of (1) and (2) with $\mathbf{i}_u = \mathbf{i}_f$. This way, only those patterns from P_f are retained, which can be reconstructed by the SMM.

Utterance specific language models

In our experiments, we work with two different sources of utterance-specific language models:

First are oracle transcriptions. Therefore, the results achieved with these language models cannot be directly compared with the baseline. However, these oracle experiments provide an estimate of potential of the approach in the optimal case, where the underlying transcriptions would be fully correct.

The second source are automatic transcriptions. We obtain these transcriptions by decoding individual utterances with a background language model. Therefore, these transcriptions do contain errors.

So overall, there are four utterance-specific LMs used, combining its source (oracle vs. automatic transcription) and whether we apply smoothing through SMM (*smooth*) or not (*sharp*).

We hope that the utterance-specific model will capture different patterns than those modeled by the background n-gram model. Therefore, we combine the utterance-specific unigram model $P_u(w_n)$ with the background model $P_{4g}(w_n|h)$ by means of log-linear combination:

$$\log P_c(w_n|h) = \lambda_{4g} \log P_{4g}(w_n|h) + (1 - \lambda_{4g}) \log P_u(w_n)$$

Table 1: Recognition results with different language models. First two are estimated on 20M lines consisting of 800M tokens, the last one is utterance-specific, thus always estimated on ca. 20 words in average.

LM	dev-clean	dev-other
4-g	4.6 %	10.7 %
LSTM	3.0 %	7.8 %
oracle 1-g	1.9 %	5.5 %

Librispeech ASR system

The Librispeech dataset [14] consists of ca. 1000 hours of read English speech. The audio comes from audiobooks of the LibriVox project[1]; it is sampled at 16 kHz. The dataset is split into training (960 hours), development (`dev`, 10.7 hours) and testing (`test`, 10.5 hours) parts. The development and testing parts are further split by speakers into easier (`-clean`) and more difficult parts (`-other`), approximately same size. We follow the practice of reporting results separately for these subsets.

The dataset also defines a standard vocabulary of 200k most common words to be used during recognition.

We use a 6-layer bidirectional LSTM RNN [5], 2×1000 units per layer. Output of the acoustic model is given by 12k of CART labels [2]. The RNN was trained [12] on the 960 hours `train` portion using Adam [9] with Nesterov momentum [4]. To prevent overfitting and improve generalization, dropout [6] with strength 0.2 was

[1] `https://librivox.org`

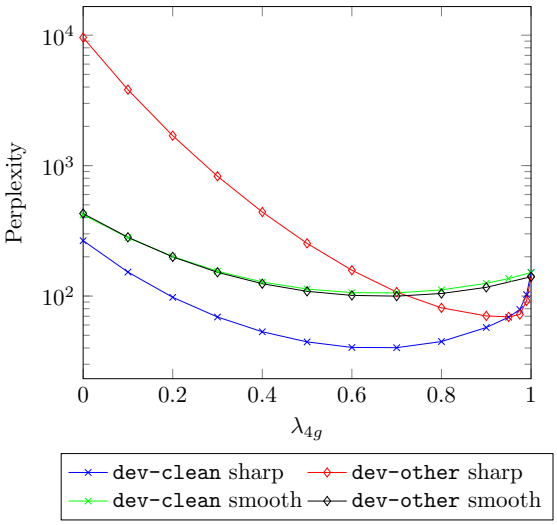

Figure 3: Perplexity of log-linear interpolation between the background LM and an utterance-specific one obtained from the first pass transcription obtained by decoding with the background LM only. Lower is better. The λ_{4g} term denotes the weight of the background LM, thus the convergence towards the right edge.

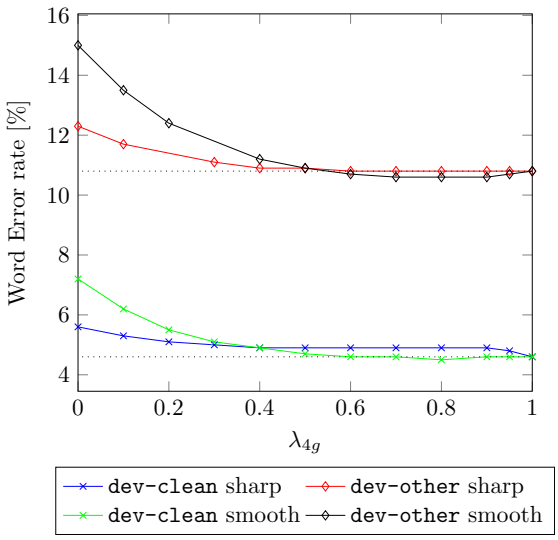

Figure 4: Word Error Rate of second pass decoding with interpolation between the background model and an utterance-specific LM obtained from first pass transcription by decoding with the background LM only. The λ_{4g} denotes weight of the background LM, thus the convergence towards the right edge. No improvement is achieved by interpolation with utterance-specific LMs without smoothing.

applied between layers. Additionally, L2 regularization with $\beta = 0.01$ was applied and noise with zero mean and variance 10^{-4} was being added to the gradients.

There are also several n-gram language models provided with the Librispeech corpus. We use the strongest one of them, a 4-gram model, as the background model. This model has perplexity of 146.7 on the transcriptions of `dev` portion of the corpus.

Baseline results of our system are summarized in Table 1. Despite being oblivious to any local context, the oracle utterance-specific LM alone has significantly better results than both the baseline models.

Oracle experiments

We begin our experiments with oracle transcriptions as source for the utterance-specific LMs.

We start by looking at the perplexity of the combined LM (Figure 1). As expected, there is no principal difference between `dev-clear` and `dev-other`, because acoustic signal has not been taken into account yet. Comparing the sharp LMs with the smooth ones, we can see that the perplexity is significantly deteriorated by the smoothing. In both cases, the interpolation improves significantly over the better of the two models combined. And in both cases, the overall behaviour is very smooth w.r.t. the interpolation coefficient λ_{4g}.

Moving on to the recognition results (Figure 2), we can see the overall properties transfer very well: Interpolation between an utterance-specific model and the background model always brings significant improvements over the baseline. Sharp models perform better, the difference is more pronounced on `dev-other`, as there the language model is more relied upon. The optimal background weight λ_{4g} is consistent with the values for best perplexity: 0.3 for the sharp models and 0.6 for the smoothed ones.

Second pass experiments

We start the discussion of utterance-specific language models derived from automatic transcription by investigating their perplexity (Figure 3). Since the utterance-specific language models contain errors from the first pass recognition, it is natural that there is a significant difference in the perplexity of sharp models between the `dev-clean` and `dev-other` portions. On the other hand, there is hardly any difference for the smooth models, suggesting that the SMM has been successful in smoothing the errors away. Note that while the interpolation of the sharp model with the background 4-gram model has achieved the best result also on `dev-other`, it only improves over the baseline in a narrow region of the interpolation weight where the background LM is dominant. Therefore, it is not expected that the second-pass recognition with this combined LM will behave differently than the first-pass one.

Then, we run the recognition with these LMs (Figure 4). In this case, there are no improvements when using sharp utterance-specific LMs. When the utterance-specific LM is smoothed through an SMM, there are improvements of around 2 % relative on both `dev-clean` and `dev-other`. Unlike with the oracle LMs, the optimal weight of the background model is higher ($\lambda_{4g} = 0.8$) than in case of tuning perplexity ($\lambda_{4g} = 0.6$). This can be explained by the fact that utterance-specific LM is favoured when optimizing for perplexity, because it already summarizes information from the acoustic signal. This information is readily available again during the second pass decoding,

Table 2: Comparison of Word Error Rate on the `dev` and `test` portions of Librispeech. No test results included for the sharp models, because they have not achieved any improvement on the `dev` data.

LM	λ_{4g}	dev-clean	dev-other	test-clean	test-other
4-g		4.6 %	10.8 %	5.0 %	11.5 %
+ sharp 1-g	1.0	4.6 %	10.8 %	—	—
+ smoothed 1-g	0.8	4.5 %	10.6 %	4.9 %	11.3 %

so more weight is put back on the background model.

Finally, we transfer the optimal interpolation coefficients to the test set; the results are reported in Table 2. The improvements do transfer to the fully unseen data.

Conclusion

In this work, we have been investigating the possibility of using utterance-specific language models. Through experiments on the Librispeech corpus, we have shown that a unigram description of a recording is potentially a very powerful model. When used in a two-pass decoding scheme, it was necessary to smooth the first-pass transcription to reduce the impact of recognition errors. Then, a consistent improvement of 2 % relative was achieved on all four evaluation portions of the dataset.

Our future work in this direction is to remove the information bottleneck of describing an utterance only by a single automatic transcription and a bag-of-words representation thereof.

Acknowledgements

The authors would like to thank Christoph M. Lüscher for providing a state-of-the-art acoustic model. The work was supported by Czech Ministry of Education, Youth and Sports from Project No. CZ.02.2.69/0.0/0.0/16_027/0008371.

References

[1] K. Beneš, S. Kesiraju, and L. Burget. i-vectors in language modeling: An efficient way of domain adaptation for feed-forward models. In *Proc. Interspeech 2018*, pages 3383–3387, 2018.

[2] K. Beulen, E. Bransch, and H. Ney. State tying for context dependent phoneme models. In *EUROSPEECH*. ISCA, 1997.

[3] X. Chen, T. Tan, X. Liu, P. Lanchantin, M. Wan, M. J. F. Gales, and P. C. Woodland. Recurrent neural network language model adaptation for multi-genre broadcast speech recognition. In *INTERSPEECH*, 2015.

[4] T. Dozat. Incorporating Nesterov Momentum into Adam. In *ICLR*, 2016.

[5] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *NEURAL NETWORKS*, pages 5–6, 2005.

[6] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[7] K. Irie, S. Kumar, M. Nirschl, and H. Liao. RADMM: recurrent adaptive mixture model with applications to domain robust language modeling. In *ICASSP*, pages 6079–6083. IEEE, 2018.

[8] S. Kesiraju, L. Burget, I. Szőke, and J. Černocký. Learning document representations using subspace multinomial model. In *Proceedings of Interspeech 2016*, pages 700–704. International Speech Communication Association, 2016.

[9] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[10] D. Klakow. Language model adaptation for tiny adaptation corpora. In *INTERSPEECH*. ISCA, 2006.

[11] M. Kockmann, L. Burget, O. Glembek, L. Ferrer, and J. Cernocký. Prosodic speaker verification using subspace multinomial models with intersession compensation. In Takao Kobayashi, Keikichi Hirose, and Satoshi Nakamura, editors, *INTERSPEECH*, pages 1061–1064. ISCA, 2010.

[12] C. Lüscher, W. Michel, K. Irie, E. Beck A. Zeyer, R. Schlüter M. Kitza, and H. Ney. RWTH ASR Systems for Librispeech: Hybrid vs. Attention. Submitted to *Interspeech*, Graz, Austria, Sep. 2019.

[13] T. Mikolov and G. Zweig. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 234–239, Dec 2012.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books. pages 5206–5210, 04 2015.

[15] A. Raju, B. Hedayatnia, L. Liu, A. Gandhe, C. Khatri, A. Metallinou, A. Venkatesh, and A. Rastrow. Contextual language model adaptation for conversational agents. In *Interspeech*, pages 3333–3337. ISCA, 2018.