Chameleon: A Language Model Adaptation Toolkit for Automatic Speech Recognition of Conversational Speech

Yuanfeng Song[1,2], **Di Jiang**[2], **Weiwei Zhao**[2]
Qian Xu[2], **Raymond Chi-Wing Wong**[1], **Qiang Yang**[1,2]
[1]Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Hong Kong SAR, China
[2]AI Group, WeBank Co., Ltd, Shenzhen, China
{songyf, raywong, qyang}@cse.ust.hk
{dijiang, davezhao, qianxu}@webank.com

Abstract

Language model is a vital component in modern automatic speech recognition (ASR) systems. Since "one-size-fits-all" language model works suboptimally for conversational speeches, language model adaptation (LMA) is considered as a promising solution for solving this problem. In order to compare the state-of-the-art LMA techniques and systematically demonstrate their effect in conversational speech recognition, we develop a novel toolkit named Chameleon, which includes the state-of-the-art *cache-based* and *topic-based* LMA techniques. This demonstration does not only vividly visualize underlying working mechanisms of a variety of the state-of-the-art LMA models but also provide an interface for the user to customize the hyperparameters of them. With this demonstration, the audience can experience the effect of LMA in an interactive and real-time fashion. We wish this demonstration would inspire more research on better language model techniques for ASR.

1 Introduction

In recent years, conversational speech recognition attracts much research attention in both academia and industry, since it is the very premise of building intelligent conversational applications. In contemporary ASR systems, language model plays an essential role of guiding the search among the word candidates and has a decisive effect on the quality of results (Jurafsky, 2000; Xu et al., 2018). However, most commercial ASR products simply rely on a "one-size-fits-all" language model. The mismatch between training and testing scenarios becomes a huge obstacle to high-quality ASR of conversational speeches in practice. Despite its simplicity and reliability, the widely used n-gram language model suffers the drawback of limited capacity of capturing the long-distance dependencies and richer semantic information, which greatly motivates the development of LMA techniques (Gandhe et al., 2018).

Although LMA techniques are increasingly considered as promising solutions for aforementioned limitation of n-gram language model, their effectiveness in real-life ASR tasks has never been systematically investigated so far. In this demonstration, we categorize the existing LMA models as two paradigms: the cache-based and the topic-based. The cache-based paradigm exploits historical observation by caching the previously used words in recent history, and then increases the probability of these words when predicting new words (Kuhn and De Mori, 1990; Lau et al., 1993; Chen, 2017). The topic models-based paradigm relies on the latent semantics discovered by probabilistic topic models, which are known for their ability to capture the semantic correlation between words and proven promising performance when applied to ASR systems (Chen et al., 2010; Wintrode and Khudanpur, 2014). Besides ASR, the topic models are also widely used in various applications such as word analysis (Kennedy et al., 2017), RFID data modeling (Kennedy et al., 2017), urban perception (de Oliveira Capela and Ramirez-Marquez, 2019) etc. We include a wide range of LMA techniques in Chameleon and some of them are specialized for ASR system. In Chameleon, the LMA techniques from the above two paradigms are implemented by conforming to the same APIs. Hence, the users could seamlessly switch between different LMA techniques and observe their real-time impact on ASR results. The ultimate goal of our demonstration is to provide a unique opportunity for the users to customize and experience the working mechanisms of a variety of the state-of-the-art LMA techniques in a vivid and interactive approach. We wish it will inspire more research on LMA in the field of ASR.

The rest of this paper is organized as fol-

Proceedings of the 2019 EMNLP and the 9th IJCNLP (System Demonstrations), pages 37–42
Hong Kong, China, November 3 – 7, 2019.

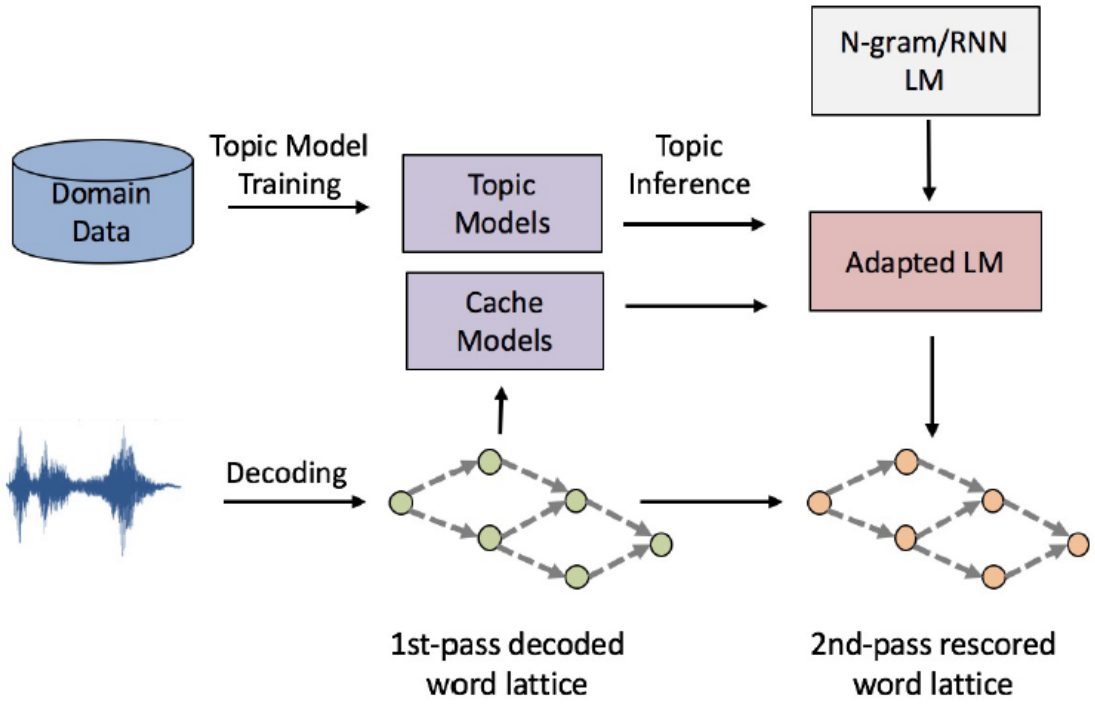


Figure 1: The Pipeline of Language Model Adaptation for ASR

lows. In Section 2, we describe the details of the Chameleon toolkit. In Section 3, we quantitatively demonstrate the performance of Chameleon, followed by the demonstration description in Section 4. Finally, we conclude this paper in Section 5.

2 Toolkit Description

In this section, we first describe the pipeline of LMA used in Chameleon, then we introduce the cache-based LMA paradigm and the topic-based LMA paradigm respectively.

2.1 Language Model Adaptation Pipeline

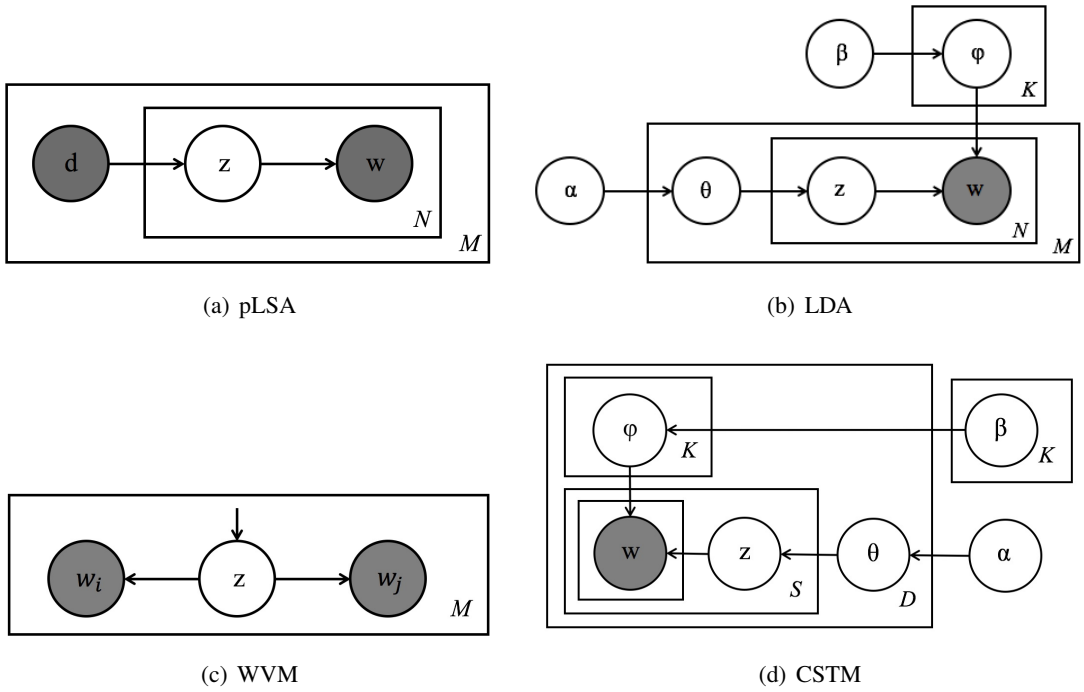
The pipeline of LMA used in Chameleon is illustrated in Figure 1. In the 1st-pass decoding, the system generates a word lattice containing the candidate results, which are further digested by LMA. Then the adapted language model rescores the word lattice and generate the 2nd-pass word lattice, in which the final decoding result can be straightforwardly obtained.

2.2 Cache-based Paradigm

There widely exists a phenomena named "word burstiness" in natural language such as conversational speech: if a word appears once, the same word and its semantically related words tend to appear again in the same speech (Madsen et al., 2005). Compared with basic n-gram model, the cache-based paradigm stores the recent historical information constructed by the 1st-pass decoded word lattices. Hence, it has the ability to emphasize the local context and boost the probabilities of recently seen words. The cache-based paradigm is widely used in language model, e.g. the cache model (Kuhn and De Mori, 1990) and the self-trigger models (Lau et al., 1993). In Chameleon, we implement the Trigger-based Discriminative Language Model (DLM) proposed in (Singh-Miller and Collins, 2007), which aims to find the optimal string $\mathbf{w}^*$ for a given acoustic input, denoted as $\mathbf{a}$, by the following equation:

$$\begin{aligned}\mathbf{w}^* = \arg\max(&\alpha \log P_{LM}(\mathbf{w}) \\ &+ \log P_{AM}(\mathbf{a}|\mathbf{w}) + \langle \beta, \phi(\mathbf{a}, \mathbf{w}, \mathbf{h}) \rangle)\end{aligned} \tag{1}$$

where P_{LM} represents a back-off n-gram language model, P_{AM} is an acoustic model, $\phi(\mathbf{a}, \mathbf{w}, \mathbf{h})$ maps the tuple $(\mathbf{a}, \mathbf{w}, \mathbf{h})$ into a feature-vector, and $\mathbf{h}$ is the history of $\mathbf{a}$ (represented as $\mathbf{h} = \{\mathbf{v_1}, \cdots, \mathbf{v_{i-1}}\}$). The parameter β is estimated using discriminative method such as perception. By caching the history of $\mathbf{a}$ and the trigger features such as the times word $\mathbf{w}$ appears in history $\mathbf{h}$, the trigger-based DLM aims to make full



(a) pLSA (b) LDA

(c) WVM (d) CSTM

Figure 2: The Graphical Models of Topic Models Included in Chameleon

use of the local context for ASR decoding.

2.3 Topic-based Paradigm

The topic-based paradigm is mathematically defined as below:

$$P(w|c) = \sum_{z} P(w|z)P(z|c) \quad (2)$$

where z is the latent topic, $P(w|z)$ is word probability given the topic and $P(z|c)$ is topic probability given the context c. Compared with the basic n-gram language models and cache-based models, the topic-based adaptation is able to predict word probability based on a long-term history and capture the long dependencies from the semantic perspective. A variety of topic models are included in Chameleon and their corresponding graphical models are illustrated in Figure 2:

- PLSA (Hofmann, 1999)
- LDA (Blei et al., 2003)
- Word Vicinity Model (WVM) (Chen et al., 2010)
- Conversational Speech Topic Model (CSTM) (Song et al., 2019)

The CSTM model is a newly designed topic model that is specialized for conversational speech. From the graphical model of CSTM in Figure 2(d), we can see that CSTM represents the words in a speech dialogue corpus D as mixtures of K "topics" and each "topic" is represented as a multinomial distribution over vocabulary of size V with Dirichlet prior β. The topic distribution θ for each speech dialogue is multinomial from a Dirichlet prior with parameter α, and each word w in a speech dialogue is drawn from a multinomial distribution of topic assignment z of the sentence it belongs to. Compared with traditional topic models such as LDA, CSTM has the ability to capture the utterance boundaries by constraining all the words in the same sentence sharing the same topic. In addition, CSTM explicitly models the "word burstiness" phenomena by allowing the word probability in the same topic varies in different documents, which is quite different from traditional topic models since their word probability in the same topic is usually fixed. The model learning process of CSTM is inspired by the inductive transfer learning mechanism (Pan and Yang, 2010) to make use of currently parallel training frameworks for LDA (Yuan et al., 2015) and well-trained open-sourced topic models (Jiang et al., 2018).

We conduct a linear interpolation between the conventional n-gram language model and the un-

igram model produced by the topic-based or the cache-based LMA techniques as below:

$$p_d(w|C) = \lambda P_{TM}(w|c)||P_{Cache}(w|c) + (1-\lambda)P_{LM}(w|c) \quad (3)$$

where $P_{LM}(w|c)$, $P_{TM}(w|c)$ and $P_{Cache}(w|c)$ are the probability given by n-gram language model, topic-based language model and cache-based language model respectively. λ is a trade-off parameter and empirically set by users. More sophisticated interpolation method such as (Della Pietra et al., 1992) can also be adopted for better performance.

3 Performance of Chameleon

In this section, we briefly describe the performance of some the aforementioned LMA techniques in terms of perplexity and Word Error Rate (WER). We use a custom service dataset in Mandarin Chinese with around 1000 hours dialogue speech in the experiments. 80% of speech data is used to train a full-fledged ASR system using Kaldi "chain" model, and the rest 20% of data is reserved for development and testing.

3.1 Perplexity

Figure 3 compares the perplexity (PPL) of the LMA techniques in Chameleon on testing data. In order to ensure the fairness of the comparison, all methods under study are trained based upon the transcript of the training data. We further adapt the n-gram language model with PLSA, LDA, Trigger-based DLM, WVM and CSTM respectively, which results in the following adapted language models: n-gram+PLSA, n-gram+LDA, n-gram+Trigger-based DLM, n-gram+WVM, n-gram + CSTM.

From Figure 3, we can observe that all the LMA techniques in Chameleon are effective at reducing the perplexity of testing data, indicating that they are helpful in predicting the words in testing data. Among all LMA techniques, CSTM achieves much lower perplexity than the other topic-based methods. This confirms the assumption that the latent topics discovered by CSTM provides valuable long-range dependency information of words. The superiority of CSTM over LDA shows that CSTM provides better fit for conversational data.

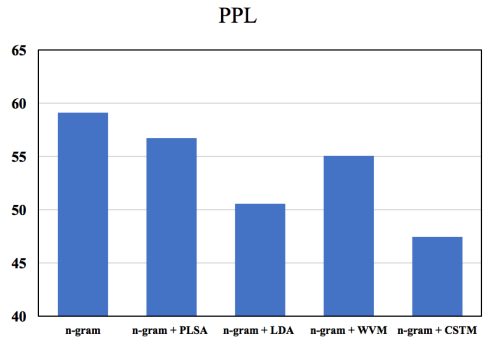


Figure 3: Perplexity Evaluation

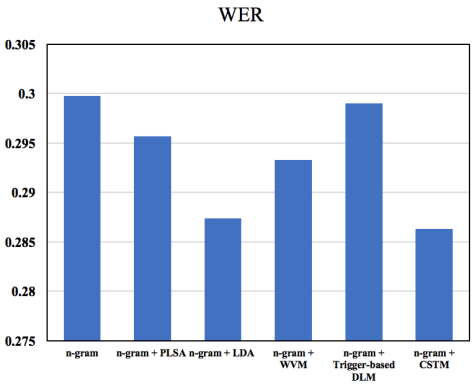


Figure 4: WER Evaluation

3.2 Lattice-rescoring

Since the ultimate goal of LMA is to improve ASR results, we further examine and compare the effectiveness of the LMA techniques in Chameleon in term of WER. Figure 4 presents the WER of all LMA techniques in Chameleon. This result shows that the LMA techniques in Chameleon is effective to reduce the errors in ASR results, indicating that utilizing the long-distance dependencies and richer semantic information is critical for ASR systems.

4 Demonstration Description

In this section, we describe the testbed ASR system and user interface of this demonstration. The goal of the demonstration is to provide a interactive approach for the users to experience the working mechanisms of a variety of the state-of-the-art LMA techniques mentioned above.

4.1 Testbed ASR System

The whole system is deployed on a machine with 314GB memory, 72 Intel Core Processor (Xeon), Tesla K80 GPU and CentOS. We trained a full-fledged ASR system based on conversational

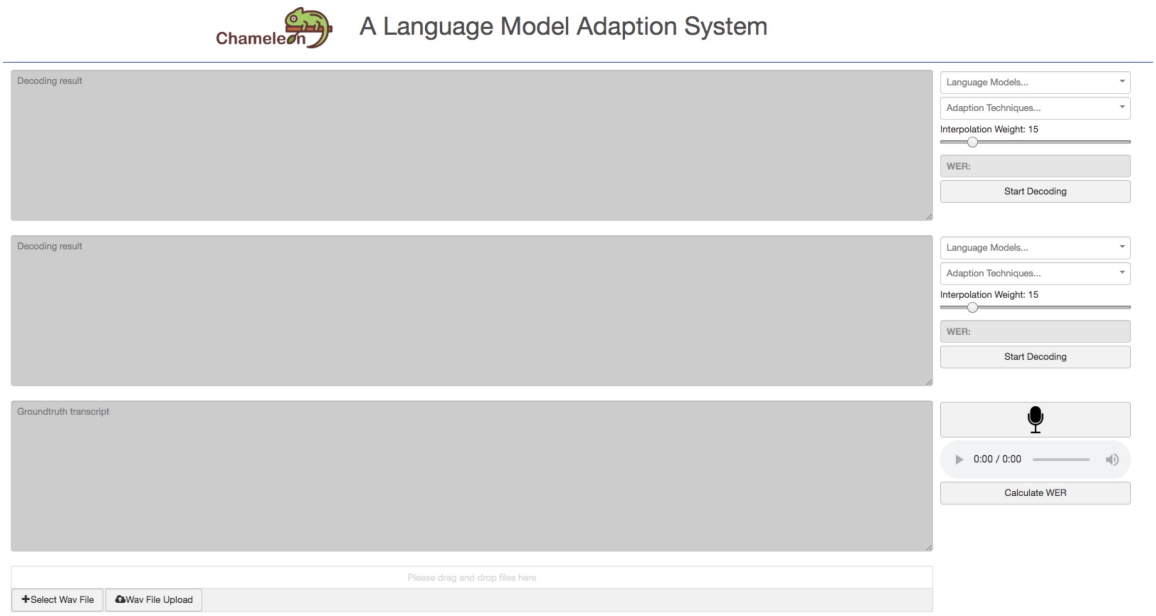


Figure 5: The User Interface of Chameleon

speeches collected from real-life customer service in Mandarin Chinese using the Kaldi toolkit[1]. The Kaldi "chain" model is used for the acoustic model. As for conventional language models, the back-off n-gram language models are trained by SRI Language Modeling Toolkit (SRILM) (Stolcke, 2002).

4.2 User Interface

We proceed to exhibit the three steps of using Chameleon with a screenshot of the user interface illustrated in Figure 5.

Step 1: The users can either upload recorded audio files or record conversational speech in real-time through the microphone provided by our system. Optionally, the groundtruth transcript can be provided by the user for the system to evaluate the WER of different LMA techniques.

Step 2: The baseline n-gram language model together with various LMA techniques described in Section 2 can be freely chosen by the users. A horizontal slider is also provided for the users to customize the interpolation weight λ. In order to facilitate the comparison of WER and decoding results of different LMA techniques, Chameleon supports applying two LMA techniques and presents their results simultaneously in a side-by-side fashion.

Step 3: When the user clicks the "Start Decoding" button, the decoding process starts. The decoded results will be presented to the corresponding text area after decoding completes. If the groundtruth transcript is provided and the "Calculate WER" button is clicked, the WER of the decoded results will be calculated and presented in the interface.

[1] http://kaldi-asr.org/

5 Conclusion and Future Work

In this demonstration, we show a novel language model adaptation toolkit named Chameleon that reveals the the effectiveness and differences of the state-of-the-art LMA techniques. Through this demonstration, the audience will have a unique journey of experiencing how LMA improves the ASR performance. In the future, we plan to include more LMA techniques and investigate new topic models dedicated for conversional speech recognition. In addition, the hyperparameter tuning step can be combined with current Automatic Machine Learning (AutoML) techniques (Quanming et al., 2018) to achieve better performance and user experience.

6 Acknowledgements

This research is partially supported by HKRGC GRF 16219816. We are grateful to the anonymous reviewers for their constructive comments on this paper.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Kuan-Yu Chen, Hsuan-Sheng Chiu, and Berlin Chen. 2010. Latent topic modeling of word vicinity information for speech recognition. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5394–5397. IEEE.

Xie Chen. 2017. Scalable recurrent neural network language models for speech recognition. In *Thesis*, pages 0–186. Cambridge.

Stephen Della Pietra, Vincent Della Pietra, Robert L Mercer, and Salim Roukos. 1992. Adaptive language modeling using minimum discriminant estimation. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 633–636. IEEE.

Ankur Gandhe, Ariya Rastrow, and Bjorn Hoffmeister. 2018. Scalable language model adaptation for spoken dialogue systems. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 907–912. IEEE.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Di Jiang, Yuanfeng Song, Rongzhong Lian, Siqi Bao, Jinhua Peng, Huang He, and Hua Wu. 2018. Familia: A configurable topic modeling framework for industrial text engineering. *arXiv preprint arXiv:1808.03733*.

Dan Jurafsky. 2000. *Speech & language processing*. Pearson Education India.

TF Kennedy, Robert S Provence, James L Broyan, Patrick W Fink, Phong H Ngo, and Lazaro D Rodriguez. 2017. Topic models for rfid data modeling and localization. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1438–1446. IEEE.

Roland Kuhn and Renato De Mori. 1990. A cache-based natural language model for speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 12(6):570–583.

Raymond Lau, Ronald Rosenfeld, and Salim Roukos. 1993. Trigger-based language models: A maximum entropy approach. In *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 45–48. IEEE.

Rasmus E Madsen, David Kauchak, and Charles Elkan. 2005. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd international conference on Machine learning*, pages 545–552. ACM.

Fernanda de Oliveira Capela and Jose Emmanuel Ramirez-Marquez. 2019. Detecting urban identity perception via newspaper topic modeling. *Cities*, 93:72–83.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Yao Quanming, Wang Mengshuo, Jair Escalante Hugo, Guyon Isabelle, Hu Yi-Qi, Li Yu-Feng, Tu Wei-Wei, Yang Qiang, and Yu Yang. 2018. Taking human out of learning applications: A survey on automated machine learning. *arXiv preprint arXiv:1810.13306*.

Natasha Singh-Miller and Michael Collins. 2007. Trigger-based language modeling using a loss-sensitive perceptron algorithm. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–25. IEEE.

Yuanfeng Song, Di Jiang, Xueyang Wu, Qian Xu, Raymond Chi-Wing Wong, and Qiang Yang. 2019. Topic-aware dialogue speech recognition with transfer learning. In *Interspeech, Austria*.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*.

Jonathan Wintrode and Sanjeev Khudanpur. 2014. Combining local and broad topic context to improve term detection. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 442–447. IEEE.

Hainan Xu, Tongfei Chen, Dongji Gao, Yiming Wang, Ke Li, Nagendra Goel, Yishay Carmiel, Daniel Povey, and Sanjeev Khudanpur. 2018. A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5929–5933. IEEE.

Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. 2015. Lightlda: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1351–1361. International World Wide Web Conferences Steering Committee.