INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Topic-Aware Dialogue Speech Recognition with Transfer Learning

Yuanfeng Song[1,2], *Di Jiang*[2], *Xueyang Wu*[1*], *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
[2]AI Group, WeBank Co., Ltd, Shenzhen, China

[1]{songyf, xwuba, raywong, qyang}@cse.ust.hk, [2]{dijiang, qianxu}@webank.com

Abstract

Dialogue speech widely exists in scenarios such as chitchat, meeting and customer service. General-purpose speech recognition systems usually neglect the topic information in the context of dialogue speech, which has great potential for improving the performance of speech recognition. In this paper, we propose a transfer learning mechanism to conduct topic-aware recognition for dialogue speech. We first propose a new probabilistic topic model named *Dialogue Speech Topic Model* (DSTM) that is specialized for modeling the context of dialogue speech. We further propose a novel transfer learning mechanism for DSTM to significantly reduce its training cost while preserving its effectiveness for accurate topic inference. The experiment results demonstrate that proposed techniques in language model adaptation effectively improve the performance of the state-of-the-art Automatic Speech Recognition (ASR) system.

Index Terms: automatic speech recognition, topic models, language modeling, transfer learning

1. Introduction

In ASR pipeline, the language model plays a vital role of guiding the search for the interpretation of acoustic features and measuring the overall acceptability of the decoding results. For many years, back-off n-gram language models have been prominently used in ASR due to simplicity and reliability [1]. However, they are limited in their ability of modeling long-range dependencies. In [2, 3], topic models are utilized for n-gram language model adaptation by introducing long-range dependencies and demonstrate promising performance in large vocabulary continuous speech recognition (LVCSR) tasks.

With their merits, existing topic models have severe drawbacks for being applied in dialogue speech recognition. First, they were not originally proposed for dialogue speech. Hence, some important linguistic structures in dialogue speech such as the utterance boundaries are completely neglected by them. However, utterance boundaries are effective for identifying latent topics, since each utterance is of limited length and the words within an utterance usually share the same topic. Second, they fail to capture the phenomenon of word burstiness, which widely exists in dialogue speech. If a word is used once in a dialogue, the same word and its semantically related words are much more likely to be used again. For example, if the word "movie" appears in a dialogue, the same word and semantically related words such as "actor" are more likely to appear again. The above characters of dialogue speech have not been properly modeled in existing topic models so far.

In this work, we first propose a novel topic model named *Dialogue Speech Topic Model* (DSTM) to simultaneously capture utterance boundaries and word burstiness in dialogue speech. With superior ability to capture the latent structure of dialogue speech data, DSTM is effective for improving the performance of topic-based language model adaptation by providing long dependencies from the perspective of semantics. However, the better performance brought by DSTM comes at a cost - training DSTM is demanding in terms of time consumption and hardware requirements. Meanwhile, although existing topic models suffer from the aforementioned drawbacks and have limited capability when applied in dialogue speech recognition, they still contain some high-level knowledge such as the word distribution for each topic, which can be used to facilitate topic inference of DSTM. Inspired by recent work in inductive transfer learning [4], we propose a novel mechanism to transfer the knowledge carried by Latent Dirichlet Allocation (LDA) to DSTM. Due to the wide availability of parallel training frameworks for LDA [5, 6, 7] and the abundance of open-sourced LDA models [8], the proposed mechanism effectively shifts the laborious training workload of DSTM to a much cheaper counterpart while preserves DSTM's superiority in accurate topic inference for dialogue speech. To the best of our knowledge, this paper is the first one discussing how to conduct transfer learning on topic models for ASR. Extensive experiments show that the proposed method outperforms several strong baselines and effectively improves the performance of ASR systems.

The rest of the paper is organized as follows. In Section 2, we briefly review the related work. In Section 3, we discuss the technical details of DSTM. In Section 4, we describe the transfer learning mechanism for topic models. In Section 5, we discuss how to incorporate DSTM with ASR systems. In Section 6, we present the experimental settings and results. Finally, we conclude this paper in Section 7.

2. Related Work

The present work is related to research fields such as topic-based language model adaptation and transfer learning. In the following subsections, we survey the most related works from the two fields.

2.1. Topic-based Language Model Adaptation

Topic models are known for effectively improving the performance of Automatic Speech Recognition (ASR) systems through providing richer contextual information for the decoding phrase [2, 3]. LDA [9] and PLSA [10] are widely used for generating document-specific language models. More recently, Word Vicinity Model (WVM) [2] was proposed to explore the words co-occurrence phenomenon as well as the latent topical context for ASR. Compared with PLSA and LDA,

[*] Work done when he worked as an intern at AI Group, Webank Co., Ltd.

http://dx.doi.org/10.21437/Interspeech.2019-1694

which describes the "word-document" co-occurrence, WVM attempts to discover the "word-word" co-occurrence dependence by latent topics. With the popularity of neural networks, neural language modeling such as recurrent neural network language model (RNNLM) was proposed in recent works [11]. Li et al. [12] further proposed two adaptation models (a cache model and a DNN-based model) for RNNLM to capture the topic information and the long-distance triggers in ASR.

2.2. Transfer Learning

In the past decade, transfer learning, as defined in [4], has been extensively studied to leverage the knowledge gained from one task (or domain) and apply it to help another. There are basically four kinds of transfer learning methods based on what part of the data and model is transferred: instance-based transfer learning [13, 14], feature-based transfer learning [15, 16], model-based transfer learning [17, 18] and relation-based transfer learning [19, 20]. Specifically, deep neural network-based model was recently applied to transfer learning [21, 22] to transfer knowledge from a well-tuned network. In the topic modeling area, the authors in [23] proposed to mine prior knowledge dynamically in the modeling process, and then a new topic model to use the knowledge to guide the model inference. Our work is related to model-based unsupervised transfer learning, and also inspired by recent work in inductive transfer learning [4], which usually shares parameters or prior distributions of hyperparameters between different models. We first train a LDA model using the corpus as a source model that is computationally efficient, and then some parameters in the LDA model is transferred to the DSTM model in order to circumvent the laborious training of DSTM.

3. Dialogue Speech Topic Model

In order to facilitate the discussion thereafter, we clarify some concepts in DSTM. *Dialogue transcript* refers to the transcript of a dialogue speech. *Utterance transcript* refers to the the transcript of an utterance and the utterance boundaries are obtained by speaker diarization [24]. We assume that the training corpus is composed of some dialogue transcripts and each dialogue transcript contains several utterance transcripts.

In the subsequent discussion, α and β are used to represent the Dirichlet hyperparameters. Specifically, we use $\alpha_{\cdot}$ to emphasize that α is a vector and $\beta_{\cdot\cdot}$ to emphasize that β is a two-dimensional matrix. The generative process of DSTM is presented in Algorithm 1. For each dialogue transcript d in the corpus D, DSTM draws a topic distribution θ_d from a Dirichlet prior parametrized by α (Line 3). Then, for each topic k in the topic set K and each d, a Multinomial distribution φ_{dk} is drawn from a Dirichlet prior with parameter β_k (Line 4 $\sim$ 6). This assumption models word burstiness by allowing for variations in the probability of each word in the same topic in different dialogue speeches. For each utterance transcript s in the utterance transcript set S_d belonging to d, DSTM draws a topic z_{ds} from θ_d (Line 8). Each word w in the word set W_s is drawn from the distribution $\varphi_{dz_{ds}}$ (Line 10), where W_s composes of all the words in transcript s. By constraining that the words in s should share the same topic, DSTM has the ability to capture utterance boundary information in dialogue speech. We proceed to discuss how to estimate the parameters of interest. The joint likelihood of $\mathbf{w}$ and $\mathbf{z}$ can be factored as follows:

$$P(\mathbf{w}, \mathbf{z}|\alpha, \beta) = P(\mathbf{w}|\mathbf{z}, \beta)P(\mathbf{z}|\alpha) \tag{1}$$

ALGORITHM 1: Generative Process of Dialogue Speech Topic Model

```
1  begin
2      for each dialogue transcript d ∈ D do
3          draw topic distribution θ_d ~ Dirichlet(α)
4          for each topic k ∈ K do
5              draw a word distribution
                 φ_dk ~ Dirichlet(β_k)
6          end
7          for each utterance transcript s ∈ S_d in d do
8              draw a topic z_ds ~ θ_d
9              for each position in utterance transcript s
                 do
10                 draw a word
                     w ~ Multinomial(φ_{dz_ds})
11             end
12         end
13     end
14 end
```

The first component $P(\mathbf{w}|\mathbf{z}, \beta)$ is an average over all possible φ:

$$\begin{aligned} P(\mathbf{w}|\mathbf{z}, \beta) &= \int_{\varphi} P(\varphi|\beta)P(\mathbf{w}|\varphi)d\varphi \\ &= \prod_{d\in D}\prod_{k\in K}\frac{B(n_{dk\cdot} + \beta_{k\cdot})}{B(\beta_{k\cdot})} \end{aligned} \tag{2}$$

where $B(\cdot)$ is the multidimensional Beta function, and n_{dkw} refers to the times that word w is assigned to topic k in dialogue d. The second component $P(\mathbf{z}|\alpha)$ is the topic assignment of the dialogue, which follows the Dirichlet distribution with hyperparameter α:

$$P(\mathbf{z}|\alpha) = \prod_{d\in D}\frac{B(n_{d\cdot\cdot} + \alpha_{\cdot})}{B(\alpha_{\cdot})} \tag{3}$$

Combining Equation (2) and (3) together yields the complete likelihood:

$$P(\mathbf{w}, \mathbf{z}|\alpha, \beta) = \prod_{d\in D}\left[\frac{B(n_{d\cdot\cdot} + \alpha)}{B(\alpha_{\cdot})}\prod_{k\in K}\frac{B(n_{dk\cdot} + \beta_{k\cdot})}{B(\beta_{k\cdot})}\right] \tag{4}$$

To perform Gibbs sampling, we further calculate $p(z_s|\mathbf{z}_{-\mathbf{s}}, \mathbf{w})$, where $\mathbf{z}_{-\mathbf{s}}$ is the set of topic assignment to all utterance transcripts except s. The conditional probability for Gibbs sampling is as follows,

$$\begin{aligned} P(z_s = k|\mathbf{z}_{-\mathbf{s}}, \mathbf{w}) &= \frac{P(\mathbf{w}, \mathbf{z})}{P(\mathbf{w}, \mathbf{z}_{-\mathbf{s}})} \\ &\approx (n'_{dk\cdot} + \alpha_{\cdot})\prod_{w\in W_s}\frac{n'_{dkw} + \beta_{k,w} - 1}{\sum_{v\in V} n'_{dkv} + \beta_{k,v} - 1} \end{aligned} \tag{5}$$

where n'_{dkw} is the number of the times word w is assigned to topic k in dialogue d without s. After the Gibbs sampling reaches steady-state, we can estimate each dialogue transcript d's topic distribution θ_{dk} and kth topic's word distribution φ_{dkw} by:

$$\theta_{dk} = \frac{n'_{dk\cdot} + \alpha_k}{\sum_{k\in K}(n'_{dk\cdot} + \alpha_k)}, \tag{6}$$

$$\varphi_{dkw} = \frac{n'_{dkw} + \beta_{kw}}{\sum_{v \in V}(n'_{dkv} + \beta_{kv})} \quad (7)$$

Since β in DSTM carries the prior information of topic-word distribution which varies according to topics, it should be non-uniform and need to be optimized throughout the corpus. Specifically, we utilized Monte Carlo EM [25] to optimize α and β. In the E-step, we choose the topic assignment for each word by running the Gibbs sampling according to Equation (5) until it reaches steady-state. In the the M-step, given the current topic assignments, we find the optimal α and β by maximizing the likelihood $P(\mathbf{w}, \mathbf{z}|\alpha, \beta)$ in Equation (4). The E-step and M-steps repeat until α and β converge.

4. Transfer Learning for DSTM Topic Inference

Based upon the discussion in Section 3, it is easy to see that training DSTM is costly in terms of both time and memory consumption. The time complexity of training DSTM is $O(I_E I_M |N||K|)$, where I_E is the number of EM iterations and I_M is the number of Markov Chain Monte Carlo iterations. $|N|$ is the number of words in the corpus and $|K|$ is the number of topics. The memory consumption of storing a DSTM model is $O(|D||K||V|)$, where $|D|$ is the number of dialogue transcripts in the corpus and $|V|$ is the size of the vocabulary. In contrast, with the state-of-the-art LDA training framework such as [5, 8], the time complexity of training a LDA model can be reduced to $O(I_M|N|)$. The memory consumption of storing a LDA model is only $O(|K||V|)$.

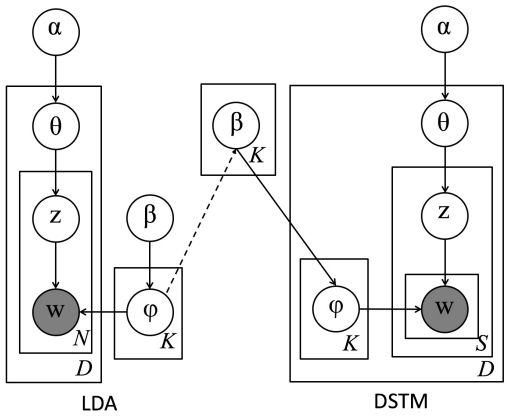

Figure 1: *Schematic diagram of transfer learning for DSTM*

Due to the great discrepancy of the training cost of LDA and DSTM, we discuss how to employ inductive transfer learning to circumvent the heavy burden of training a DSTM model. As shown in Figure 1, the hyperparameter $\beta_{k\cdot}^{DSTM}$ in DSTM is transferred from the parameter $\phi_{k\cdot}^{LDA}$ in LDA. After transfer learning, the Gibbs sampling formula $P(z_s = k|\mathbf{z_{-s}}, \mathbf{w})$ is updated to:

$$P(z_s = k|\mathbf{z_{-s}}, \mathbf{w}) \approx (n'_{dk\cdot} + \alpha_\cdot) \times \prod_{w \in W_s} \frac{n'_{dkw} + \phi_{k,w}^{LDA} - 1}{\sum_{v \in V} n'_{dkv} + \phi_{k,v}^{LDA} - 1} \quad (8)$$

When applying DSTM to the candidate transcript d of a new dialogue speech, we sample topics for the words in the candidate transcript according to Equation (8). The above procedure uses updated document-topic count n'_{dk} and topic-word count n'_{dkw} and fixed $\beta_{k\cdot}^{DSTM}$. Based on the sampling results, we can apply Equation (6) to estimate parameters such as θ_{dk} and φ_{dkw}, which are further utilized for n-gram language model adaptation.

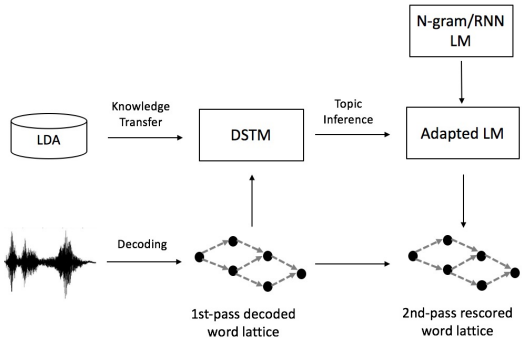

Figure 2: *Topic-Aware ASR with DSTM*

5. Topic-Aware ASR

Figure 2 shows the pipeline of topic-aware ASR. The topics discovered by DSTM are utilized to complement the n-gram language model by interpolation. For each dialogue speech, the transcript in the lattice of the first-pass decoding is considered as a dialogue transcript d. We further infer the topic distribution θ_d and φ_{dkw} according to Equations (8) and (7). Based on these parameters, we compute a document-specific unigram model by $P_{DSTM}(w|\theta_d) = \sum_{k \in K} \varphi_{dkw}\theta_{dk}$ and adapt the n-gram language model as follows:

$$p_d(w|C) = \lambda P_{DSTM}(w|\theta_d) + (1 - \lambda)P_{LM}(w|C) \quad (9)$$

where C is the context information on previous words, λ is a trade-off coefficient and $P_{LM}(w|C)$ is the probability given by n-gram language model. The adapted language model is further utilized for rescoring the lattice.

6. Experiments

In Section 6.1, we describe the experimental setup. In Section 6.2, we present experimental results in terms of perplexity, Word Error Rate (WER) and efficiency.

6.1. Experimental Setup

The corpus used for our experiments consists of about 1000 hours of dialogue speech collected from real-life customer service in Mandarin Chinese. We utilized 80% of the data for training a ASR system with the Kaldi Toolkit[1] and reserved the rest 20% for development (10%) and testing (10%). Specifically, we trained a Kaldi "chain" model based on the training data. The baseline n-gram language model is trained by the SRI Language Modeling Toolkit (SRILM) [26]. Parameters such as topic number $|K|$ and coefficient λ are tuned on the development dataset. All experiments were conducted on a machine with 314GB memory, 72 Intel Core Processor (Xeon), Tesla K80 GPU and CentOS.

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

6.2. Experimental Result

6.2.1. Perplexity

Table 1 compares the perplexity (PPL) of different language models. The n-gram language model, LDA, WVM and DSTM, are all trained on the transcript of the training data. We adapt the n-gram language model by RNNLM [27], LDA, Cache [12], WVM and DSTM by interpolation, which results in the following language models: n-gram+RNNLM, n-gram+LDA, n-gram+Cache, n-gram+WVM, n-gram + DSTM-S, n-gram + DSTM-TL and n-gram + DSTM-TL-E. Specifically, DSTM-S means DSTM trained from scratch, DSTM-TL means DSTM trained with transfer learning by training a LDA model from scratch and DSTM-TL-E means DSTM trained with transfer learning from an existing LDA model.

As we can see from Table 1, the methods based on DSTM achieve much lower perplexities than the other methods. This confirms our assumptions that the topics discovered by DSTM provides valuable long-range dependency information of words. The superiority of DSTM over LDA and Cache shows that DSTM provides better fit for the dialogue data. What's more, by comparing different DSTM-based methods, we observe that DSTM-TL and DSTM-TL-E obtain similar performance as DSTM-S, indicating that transfer learning only causes mild performance degrade in perplexity.

Table 1: *Perplexity of Different Language Models*

Model	PPL
n-gram	59.10
n-gram + RNNLM	48.01
n-gram + LDA	50.54
n-gram + WVM	55.03
n-gram + Cache	50.35
n-gram + DSTM-S	46.62
n-gram + DSTM-TL	47.41
n-gram + DSTM-TL-E	47.41

6.2.2. Lattice-rescoring

Since our ultimate goal is to improve ASR, we further examine the effectiveness of DSTM in term of WER. Table 2 lists WER of different language model adaptation methods. RNNLM, LDA, WVM, Cache, DSTM-S, DSTM-TL and DSTM-TL-E have respectively 1.28%, 1.24%, 0.65%, 0.79%, 1.48%, 1.35%, 1.35% improvement over the baseline n-gram language model. The methods based on DSTM achieve lower WER than the other methods. Furthermore, we observe that the performances of DSTM-TL, DSTM-TL-E and DSTM-S are quite close. This result indicates that the information loss caused by transferring from LDA to DSTM is little and validates the effectiveness of the proposed transfer learning mechanism. What's more, compared with LDA, DSTM produces more accurate topic inference results, which further help to reduce the WER.

6.2.3. Efficiency Analysis

We empirically examine the efficiency of DSTM and the results are shown in Table 3. We can see that DSTM-S suffers from low efficiency since training DSTM from scratch involves multiple iterations of Monte Carlo EM, which is quite time-consuming. Comparing DSTM-S and DSTM-TL, we observe that the proposed transfer learning mechanism reduces the training time of DSTM to a level similar to LDA. The time consumption of DSTM-TL-E indicates that the cost of transferring knowledge from an existing LDA model to DSTM is negligible. These property of DSTM is of great value in practice and we can apply it to improve ASR with little compromise on efficiency.

Table 2: *WER of Different Language Models*

Model	WER
n-gram	29.98%
n-gram + RNNLM	28.70%
n-gram + LDA	28.74%
n-gram + WVM	29.33%
n-gram + Cache	29.19%
n-gram + DSTM-S	28.50%
n-gram + DSTM-TL	28.63%
n-gram + DSTM-TL-E	28.63%

Table 3: *Time Consumption (In seconds) of the Model Training and Inference for Different Topic Models*

#Topic	LDA	DSTM-S	DSTM-TL	DSTM-TL-E
50	407.18	4512.19	406.72	0.89
100	449.77	10717.36	449.41	0.94
200	498.65	18364.31	498.18	1.03
300	536.74	34482.59	536.36	1.12
400	550.03	49730.21	550.17	1.18

7. Conclusions

In this paper, we propose a new framework for topic-based language model adaptation in dialogue speech ASR. We first propose a novel topic model named DSTM which simultaneously captures the utterance boundaries and word burstiness in natural dialogue speech. To relieve the heavy burden of training DSTM, we further design a transfer learning mechanism to transfer the knowledge carried by LDA to DSTM. Experimental results show that the proposed approach is able to improve the performance of the state-of-the-art ASR system. In the future work, we plan to investigate more about the intrinsic structures of dialogue speech and explore new topic models for language model adaptation.

8. Acknowledgements

This research is partially supported by HKRGC GRF 14205117. We are grateful to the anonymous reviewers for their constructive comments on this paper.

9. References

[1] J. R. Bellegarda, "Statistical language model adaptation: review and perspectives," *Speech communication*, vol. 42, no. 1, pp. 93–108, 2004.

[2] K.-Y. Chen, H.-S. Chiu, and B. Chen, "Latent topic modeling of word vicinity information for speech recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5394–5397.

[3] J. Wintrode and S. Khudanpur, "Combining local and broad topic context to improve term detection," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 442–447.

[4] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[5] J. Yuan, F. Gao, Q. Ho, W. Dai, J. Wei, X. Zheng, E. P. Xing, T.-Y. Liu, and W.-Y. Ma, "Lightlda: Big topic models on modest computer clusters," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1351–1361.

[6] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks*. Valletta, Malta: University of Malta, 2010, pp. 46–50.

[7] Z. Liu, Y. Zhang, E. Y. Chang, and M. Sun, "Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 26, 2011.

[8] D. Jiang, Y. Song, R. Lian, S. Bao, J. Peng, H. He, and H. Wu, "Familia: A configurable topic modeling framework for industrial text engineering," *arXiv preprint arXiv:1808.03733*, 2018.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[10] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[11] H. Xu, K. Li, Y. Wang, J. Wang, S. Kang, X. Chen, D. Povey, and S. Khudanpur, "Neural network language modeling with letter-based features and importance sampling," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6109–6113.

[12] K. Li, H. Xu, Y. Wang, D. Povey, and S. Khudanpur, "Recurrent neural network language model adaptation for conversational speech recognition," *INTERSPEECH, Hyderabad*, pp. 1–5, 2018.

[13] M. N. A. Khan and D. R. Heisterkamp, "Adapting instance weights for unsupervised domain adaptation using quadratic mutual information and subspace learning," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 1560–1565.

[14] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Boosting for transfer learning," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, pp. 193–200. [Online]. Available: http://doi.acm.org/10.1145/1273496.1273521

[15] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *CVPR 2011*. IEEE, 2011, pp. 3209–3216.

[16] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer joint matching for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1410–1417.

[17] Z. Zhao, Y. Chen, J. Liu, Z. Shen, and M. Liu, "Cross-people mobile-phone based activity recognition," in *Twenty-second international joint conference on artificial intelligence*, 2011.

[18] S. J. Pan, D. Shen, Q. Yang, and J. T. Kwok, "Transferring localization models across space." in *AAAI*, 2008, pp. 1383–1388.

[19] L. Mihalkova and R. J. Mooney, "Transfer learning by mapping with minimal target data," in *Proceedings of the AAAI-08 workshop on transfer learning for complex tasks*, 2008.

[20] J. Davis and P. Domingos, "Deep transfer via second-order markov logic," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 217–224.

[21] M. Long, J. Wang, Y. Cao, J. Sun, and P. S. Yu, "Deep learning of transferable representation for scalable domain adaptation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2027–2040, 2016.

[22] Z. Luo, Y. Zou, J. Hoffman, and F.-F. Li, "Label efficient learning of transferable representations acrosss domains and tasks," in *Advances in Neural Information Processing Systems*, 2017, pp. 165–177.

[23] Z. Chen and B. Liu, "Topic modeling using topics from many domains, lifelong learning and big data," in *International Conference on Machine Learning*, 2014, pp. 703–711.

[24] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[25] G. Celeux, D. Chauveau, and J. Diebolt, "Stochastic versions of the em algorithm: an experimental study in the mixture case," *Journal of Statistical Computation and Simulation*, vol. 55, no. 4, pp. 287–314, 1996.

[26] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.

[27] H. Xu, T. Chen, D. Gao, Y. Wang, K. Li, N. Goel, Y. Carmiel, D. Povey, and S. Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5929–5933.

[28] D. Klakow and J. Peters, "Testing the correlation of word error rate and perplexity," *Speech Communication*, vol. 38, no. 1-2, pp. 19–28, 2002.