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

Latent Dirichlet Allocation Based Acoustic Data Selection for Automatic Speech Recognition

Mortaza (Morrie) Doulaty[1,*], *Thomas Hain*[2]

[1]Microsoft, Germany
[2]University of Sheffield, UK

mdoulaty@microsoft.com, t.hain@sheffield.ac.uk

Abstract

Selecting in-domain data from a large pool of diverse and out-of-domain data is a non-trivial problem. In most cases simply using all of the available data will lead to sub-optimal and in some cases even worse performance compared to carefully selecting a matching set. This is true even for data-inefficient neural models. Acoustic Latent Dirichlet Allocation (aLDA) is shown to be useful in a variety of speech technology related tasks, including domain adaptation of acoustic models for automatic speech recognition and entity labeling for information retrieval. In this paper we propose to use aLDA as a data similarity criterion in a data selection framework. Given a large pool of out-of-domain and potentially mismatched data, the task is to select the best-matching training data to a set of representative utterances sampled from a target domain. Our target data consists of around 32 hours of meeting data (both far-field and close-talk) and the pool contains 2k hours of meeting, talks, voice search, dictation, command-and-control, audio books, lectures, generic media and telephony speech data. The proposed technique for training data selection, significantly outperforms random selection, posterior-based selection as well as using all of the available data.

Index Terms: Acoustic Latent Dirichlet Allocation, data selection, speech recognition

1. Introduction

Bootstrapping an speech recognition system for a new domain is a common practical problem. A typical scenario is to have some limited in-domain data from a target domain that ASR system is being built for and a pool of out-of-domain data, often containing a diverse set of potentially mismatched data. Using all of the available data is not a good choice in some cases, especially when the pooled data contains a lot of mismatched data to your target domain. There are two main concerns about using all of the available data. Some times the performance is sub-optimal compared to carefully selecting a matching set and in some cases the performance can be even worse [1, 2]. The other concern is the amount of computation needed to train the models. If a comparable or ideally a better model can be trained with a fraction of the available data, then it would be more computationally efficient to train with the smaller set. In these cases data selection becomes a crucial problem. The same problem is applicable for adaptation data selection as well, where the aim is to select data for adapting acoustic model using a limited in-domain dataset.

[*]Core part of this work was performed while the author was studying at the University of Sheffield

In this paper we propose to use acoustic Latent Dirichlet Allocation (aLDA) for matching acoustically similar data to the limited in-domain data from a pool of diverse data. aLDA is already applied for domain discovery [3] and domain adaptation [4] in automatic speech recognition as well as media entity recognition, such as show and genre identification in information retrieval systems for media archives [5, 6, 7].

Next section briefly discusses LDA and aLDA. Section 3 describes the experimental setup and how aLDA data selection technique works, followed by the conclusion in section 4 and references.

2. Acoustic Latent Dirichlet Allocation

As shown in our previous works [3, 4, 5], aLDA domain posteriors have a unique distribution across different domains that can be used to characterise the acoustic scenery. In this work we make use of aLDA domain posterior features as a basis of acoustic similarity in a data selection problem. The idea is that using acoustically similar data to a target domain for training acoustic models should improve the ASR accuracy on that domain. While using all of the available data which does not necessarily match the target domain could potentially harm the accuracy.

LDA is an unsupervised probabilistic generative model for collections of discrete data. Since speech observations are continuous data, first it needs to be represented by some discrete symbols, here called acoustic words. A GMM with N mixture components is employed for this purpose. The index of Gaussian component with the highest posterior probability is then used to represent each frame with a discrete symbol. Frames of every acoustic document of length T, $\mathbf{d}_i = \{\mathbf{u}_1, ..., \mathbf{u}_t, ..., \mathbf{u}_T\}$ are represented as:

$$v_t = \arg\max_n P(G_n|\mathbf{u}_t), \quad 1 \leq n \leq N \tag{1}$$

where G_n is a Gaussian component from a mixture of N components. With this new representation, document $\mathbf{d}_i$ is represented as $\tilde{\mathbf{d}}_i = \{v_1, ..., v_t, ..., v_T\}$. For each acoustic word v_t in each acoustic document $\tilde{\mathbf{d}}_i$, term frequency-inverse document frequency (tf-idf) can be computed as:

$$w_t = \text{tfidf}(v_t, \tilde{\mathbf{d}}_i, \tilde{\mathbf{D}}) = \text{tf}(v_t, \tilde{\mathbf{d}}_i)\ \text{idf}(v_t, \tilde{\mathbf{D}}) \tag{2}$$

where $\tilde{\mathbf{D}}$ is the set of all acoustic documents represented with acoustic words. With each document now represented with tf-idf scores as $\bar{\mathbf{d}}_i = \{w_1, ..., w_t, ..., w_T\}$, the LDA models can be trained.

A graphical representation of the LDA model is shown at Figure 1, as a three-level hierarchical Bayesian model. In this

http://dx.doi.org/10.21437/Interspeech.2019-1797

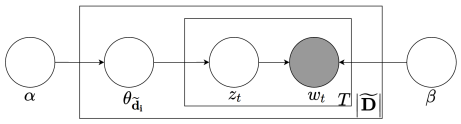

Figure 1: *Graphical model representation of LDA*

model, the only observed variables are w_t's. α and β are dataset level parameters, $\theta_{\tilde{\mathbf{d}}_i}$ is a document level variable and z_t is a latent variable indicating the domain from which w_t was drawn. The following joint distribution is the result of the generative process of LDA:

$$p(\theta, \mathbf{z}, \bar{\mathbf{d}}|\alpha, \beta) = p(\theta|\alpha) \prod_{t=1}^{T} p(z_t|\theta) p(w_t|z_t, \beta) \quad (3)$$

The posterior distribution of the latent variables given the acoustic document and α and β parameters is:

$$p(\theta, \mathbf{z}|\bar{\mathbf{d}}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \bar{\mathbf{d}}|\alpha, \beta)}{p(\bar{\mathbf{d}}|\alpha, \beta)} \quad (4)$$

Computing $p(\bar{\mathbf{d}}|\alpha, \beta)$ requires some intractable integrals. A reasonable approximate can be acquired using variational approximation, which is shown to work reasonably well in various applications [8]. The approximated posterior distribution is:

$$q(\theta, \mathbf{z}|\gamma, \phi) = q(\theta|\gamma) \prod_{t=1}^{T} q(z_t|\phi_t) \quad (5)$$

where γ is the Dirichlet parameter that determines θ and ϕ is the parameter for the multinomial that generates the latent variables.

Training minimises the Kullback-Leiber Divergence between the real and the approximated joint probabilities (equations 4 and 5) [8]:

$$\arg\min_{\gamma,\phi} \text{KLD}\left(q(\theta, \mathbf{z}|\gamma, \phi) \;||\; p(\theta, \mathbf{z}|\bar{\mathbf{d}}, \alpha, \beta)\right) \quad (6)$$

The posterior Dirichlet parameter $\gamma(\bar{\mathbf{d}})$ can be used as features representing the acoustic conditions. These features are used in different tasks, for example for genre and show entity identification and classification tasks [5, 6, 7, 9] or for domain discovery and adaptation in speech recognition [3, 4].

3. Experimental Setup

To evaluate the effectiveness of aLDA for data-selection in ASR, we are trying to solve this practical problem: given a small set of in-domain data and a large pool of out-of-domain and potentially mismatched data, what's the best set of data that can be selected from the pool to train a model for the in-domain data.

3.1. Data

The in-domain dataset consists of 32 hours of meeting data. Meeting participants used a wide-variety of devices to join the online meetings, including different headsets, earphones with microphones, laptop/table/phone microphone in a far-field setting (arm-length distance) and table-top meeting microphones. Essentially the data is a mixture of far-field and close-talking in different environments. Table 1 summarises some statistics of the in-domain dataset.

Table 1: *Statistics of the in-domain dataset*

Characterisitc	Notes
Gender	37% female / 63% male
Nativeness	77% native / 23% non-native
Device	53% laptop computer 11% desktop computer 19% mobile phone 9% tablet 8% other devices
Distance to microphone	27% far-field / 73% close-talk

Table 2: *Statistics of the out-of-domain dataset*

Domain	Duration (hours)	Percentage
Generic media	782	39.1%
Audio books	339	17.0%
Meeting	228	11.4%
Telephony speech	218	10.9%
Talks	172	8.6%
Command and Control	112	5.6%
Lectures	111	5.6%
Dictation	38	1.9%
Total	**2000**	**100%**

Meetings are mostly real discussions about IT-related topics and there was no control on the participants' recording and environmental conditions. From this in-domain set, 10 hours is used as the dev set and 22 hours as the test set.

The pool of out-of-domain dataset consists of 2000 hours of diverse and multi-domain data. Table 2 summarises the amount of data for each domain.

Around 39% of the pooled data belongs to the generic media domain which includes professional and amateur media recordings from radio, TV, pod-casts and YouTube. Meeting data (which is considered to be the best matching data for our in-domain data) is only 11% of the pooled data and they were not a part of the in-domain meeting recordings.

The data used for language modelling is fixed in all experiments and includes around 200 million words from Wikipedia, TedTalks, YouTube subtitles and e-books with a vocabulary of size 300 thousand words [10, 11]. For the lexicon, a base CMU dictionary was used and for the OOVs, a seq-to-seq g2p model was trained on the base lexicon and used to generate the missing pronunciations [12].

3.2. Baseline

The purpose of this study is to show how aLDA data selection can improve ASR accuracy of a target domain and for that reason all of the model architectures are the same in all of the experiments and the only difference is the amount of data used for training the acoustic models. For acoustic models, TDNN-LSTM model with 3 layers and 1024 nodes in each layer was trained using the lattice-free MMI objective function [13] in Kaldi toolkit [14]. During decoding a pruned 3-gram language model was used to generate lattices and the lattices were then rescored using a 5-gram language model. Table 3 presents the

Table 3: *Baseline results*

Model	WER		
	Overall	Far-field	Close-talk
Baseline with all data	29.4	53.4	20.9

WER for the test set and for its far-field and close-talk subsets. WER for the far-field subset of the test set is very high and that shows how challenging this dataset is.

3.3. aLDA Data Selection

All of the in-domain data was used for training the aLDA model with the procedure described in section 2 using a vocabulary size of 1024 (number of Gaussian mixture components) and 2048 latent domains. Both these values were selected based on our previous experiments [4, 5]. The trained aLDA model was then used to get the posterior Dirichlet parameter γ for all of the utterances in the training, dev and test set. The posterior vectors from the dev set were then clustered into 512 clusters using k-means clustering algorithm and the centroid of each cluster was used to represent each cluster.

An iterative approach was used to select the matching data from the pool of out-of-domain data. For each γ_i (centroid of the cluster i) the distance to all of the utterances in the training set is computed as:

$$\Phi(\gamma_i, \gamma_j), \forall \gamma_j \in \mathbf{S^{trn}} \tag{7}$$

where $\mathbf{S^{trn}}$ is the set of all Dirichlet posterior vectors of the training set and Φ is the cosine distance between the two vectors defined as:

$$\Phi(\gamma_i, \gamma_j) = 1 - \frac{\gamma_i \gamma_j}{||\gamma_i||_2 \; ||\gamma_j||_2} \tag{8}$$

The closest utterance (in terms of cosine distance between the Dirichlet posteriors and cluster centroid) that was smaller than a λ threshold was added to the selection and was removed from the pool. This iterative process continued until either the minimum distance criterion could not be met for all of the γ_i or the pool was depleted. Algorithm 1 shows this iterative process.

Tuning the λ threshold requires exploration of a range of values. In our experiments we found that this threshold value is not very sensitive and values in the range of 0.1 to 0.25 resulted in sensible amounts of data. In the final experiment a threshold value of 0.2 was used. This threshold value can also be used to control the amount of data being selected as well if there is budget on the amount of data.

3.4. Combining Text LDA with aLDA

Text-based LDA (tLDA) can also be used to further improve the aLDA data selection. The idea is that aLDA captures acoustic similarities in the data and tLDA can further help with linguistic content's similarity. tLDA is already shown to improve classification accuracy in LDA based acoustic information retrieval [5] as well as language modelling tasks [15, 16, 17, 18, 19]. Training tLDA models followed a similar procedure to aLDA and a comparable number of latent topics and vocabulary size was used. In our experiments tLDA on its own was not outperforming the baseline and hence those results are not included in this paper. An explanation for it could be the fact that pure linguistic similarity does not necessarily mean that the acoustic conditions are similar as well and thus cannot compensate for the acoustic mismatch.

Different approaches for combining aLDA and tLDA scores were examined. Including but not limited to linear combination of posteriors, two level hierarchical search and two independent search followed by union. At the end using two approaches independently and then combining the selected data resulted in the best performance.

Algorithm 1 Data-selection based on Dirichlet posterior

Input: Training data $\mathbf{S}^{trn}$ of M utterances,
Training set Dirichlet posteriors $\{\gamma_1^{trn}, \ldots, \gamma_M^{trn}\}$,
Dev set posterior centroids $\{\gamma_1^{dev}, \ldots, \gamma_N^{dev}\}$,
Distance threshold λ

Initialize: $\mathbf{S}^{new} = \{\}$; $count = 0$;

while $\mathbf{S}^{trn} \neq \varnothing$ **do**
  $count = 0$
  for All $\gamma_i^{dev} \in \{\gamma_1^{dev}, \ldots, \gamma_N^{dev}\}$ **do**
    $d = \min \; \Phi(\gamma_i^{dev}, \gamma_j^{trn}) \; \forall \gamma_j^{trn} \in \{\gamma_1^{trn}, \ldots, \gamma_M^{trn}\}$
    if $d < \lambda$ **then**
      $j^* = \arg\min_j \Phi(\gamma_i^{dev}, \gamma_j^{trn})$
      Remove $\gamma_{j^*}^{trn}$ from $\{\gamma_1^{trn}, \ldots, \gamma_M^{trn}\}$ set
      $\mathbf{S}^{trn} = \mathbf{S}^{trn} \setminus \{s_{j^*}^{trn}\}$
      $\mathbf{S}^{new} = \mathbf{S}^{new} \cup \{s_{j^*}^{trn}\}$
      $count = count + 1$
    end if
  end for
  if $count == 0$ **then**
    break
  end if
end while

Output: $\mathbf{S}^{new}$

3.5. Results and Discussion

In this section LDA based data selection is compared against random selection, using all of the available data (2000 hours) and phone-posterior based data selection [20]. Table 5 summerises the results of the experiments. For the random selection two budgets of 500 and 1000 hours are used and each experiment is repeated 2 times and an average value plus the standard deviation of the runs are provided (due to the data size and computation time this experiment could not be repeated more). Using all of the available data, the WER on the test set is 29.4. Phone-posterior based selection with a predefined budget of 1000 hours yields a WER of 29.0 which is slightly better than using all of the available data, but savings on computation time is massive (50% less data used for training the model). aLDA method selects 49.7% of the data and brings down the error by

Table 4: *WER and amount of data for different data selection methods*

Method	Amount of data (hours)	WER
Random selection	500	31.5 ($\pm$2.00)
	1000	30.1 ($\pm$0.98)
All of data	2000	29.4
Phone-posterior	1000	29.0
aLDA	995.4	28.5
aLDA + tLDA	1103.9	28.3

Table 5: *Amount of selected data by aLDA*

Component	Duration (hours)	Percentage of domain
Generic media	317.5	40.6%
Meeting	205.5	90.1%
Audio books	147.6	43.5%
Talks	136.9	79.6%
Lectures	94.6	85.2%
Telephony speech	84.3	38.7%
Dictation	6.3	16.6%
Command and Control	2.7	2.4%
Total	**995.4**	n/a

0.9% absolute. Combing aLDA with tLDA further reduces the error to 28.3% while selecting only 108.5 hours more data.

The results presented in table 5 show the effectiveness of the proposed aLDA data selection and how it can be further improved by using tLDA.

3.6. Analysis of the Selected Data

From the pool of 2000 hours, aLDA technique selected 995.4 hours. In this section the selected data is analysed to understand which parts of the training data was found to be the best match to the target in-domain data.

The training pool consists of data from 8 domains: audio books, command and control, dictation, generic media, lectures, meetings, talks and telephony speech. From these domains only the meeting data seems to be the best match, at least in terms of the domain tags associated with each component. As mentioned in section 3.1, the meeting data in our training set is not a part of the test set recordings, but rather some generic and diverse meeting data. It includes data from the AMI [21] and ICSI [22] projects as well as some other internal and external sources and in that sense it's not considered as an strictly in-domain data.

The majority of the selected data belongs to the generic media domain (which was the predominant class in our pool), also almost all of the available meeting data was selected showing that it was a very good match to our in-domain meeting data, at least compared to other data sources. Other interesting observation is the amount of data from dictation and command and control domains, where in total only 8 hours is selected. Checking those data, they are very clean audio. Command and control data set has a lot of very short utterances (single words) and that could contribute to the LDA domains posterior mismatch and not being selected.

In the previous section it was shown that including the data from tLDA selection improves the ASR performence while adding only 108 extra hours. Inspecting those extra data reveals that most of them are selected from talks and telephone speech (35h and 65h respectively). Suggesting that the textual similarities of those domains was picked up by the tLDA and we end up using all of the available talks data in the training of the aLDA+tLDA model. Those extra data improves the accruacy by 0.2% absolute.

4. Conclusions

Selecting matching data to a small set of in-domain data from a large pool of out-of-domain and mismatched data is a non-trivial problem. This problem arises in many practical applications of speech recognition where the task is to build an ASR system for a new target domain where there is a very limited amount of data is available. Often using all of the potentially mismatched data results in sub-optimal and poor performance compared to carefully selecting a matching subset.

In this paper aLDA based data selection is proposed for the first time and its effectiveness is experimented on a large dataset. Our in-domain dataset contains 32 hours of meeting data (mixed far-field and close-talking) and the pool of out-of-domain data consists of 2000 hours of data from very diverse domains. Using all of the available data, the baseline WER is 29.4%. Using the proposed iterative data selection technique and with slightly less than half of the training data the overal WER on our 20-hour test set is 0.9% absolute better than using all of the available data. Combining aLDA with tLDA further reduces the WER to 28.3%.

Future work can include automatic distance threshold finding, exploring the effectiveness of aLDA data selection with data augmentation, finding better ways to combine aLDA and tLDA and further analysis of the selected data by aLDA+tLDA.

5. Acknowledgements

The first author would like to thank Trevor Francis for supporting parts of this work.

6. References

[1] K. Wei, Y. Liu, K. Kirchhoff, and J. Bilmes, "Unsupervised submodular subset selection for speech data," in *Proc. of ICASSP*, Florence, Italy, 2014.

[2] M. Doulaty, O. Saz, and T. Hain, "Data-selective transfer learning for multi-domain speech recognition," in *Proc. of Interspeech*, Dresden, Germany, 2015.

[3] ——, "Unsupervised domain discovery using latent dirichlet allocation for acoustic modelling in speech recognition," in *Proc. of Interspeech*, Dresden, Germany, 2015.

[4] M. Doulaty, O. Saz, R. W. M. Ng, and T. Hain, "Latent Dirichlet Allocation Based Organisation of Broadcast Media Archives for Deep Neural Network Adaptation," in *Proc. of ASRU*, Arizona, USA, 2015.

[5] ——, "Automatic Genre and Show Identification of Broadcast Media," in *Proc. of Interspeech*, California, USA, 2016.

[6] S. Kim, S. Narayanan, and S. Sundaram, "Acoustic topic model for audio information retrieval," in *Proc. of WASPAA*, New Paltz NY, USA, 2009.

[7] S. Kim, P. Georgiou, and S. Narayanan, "On-line genre classification of TV programs using audio content," in *Proc. of ICASSP*, Vancouver, Canada, 2013.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] M. Rouvier, D. Matrouf, and G. Linares, "Factor analysis for audio-based video genre classification." in *Proc. of Interspeech*, Brighton, UK, 2009.

[10] K. Heafield, I. Pouzyrevsky, J. H. Clark, and P. Koehn, "Scalable modified kneser-ney language model estimation," in *Proc. of ACL*, Sofia, Bulgaria, 2013.

[11] O. Tange, "Gnu parallel - the command-line power tool," *;login: The USENIX Magazine*, vol. 36, no. 1, pp. 42–47, Feb 2011. [Online]. Available: http://www.gnu.org/s/parallel

[12] "CMU Sequence-to-Sequence G2P toolkit," https://github.com/cmusphinx/g2p-seq2seq.

[13] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Proc. of Interspeech*, California, USA, 2016.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget *et al.*, "The kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[15] O. Saz, M. Doulaty, S. Deena, R. Milner, R. W. M. Ng, M. Hasan, Y. Liu, and T. Hain, "The 2015 Sheffield system for transcription of multi-genre broadcast media," in *Proc. of ASRU*, Arizona, USA, 2015.

[16] S. Deena, M. Hasan, M. Doulaty, O. Saz, and T. Hain, "Combining feature and model-based adaptation of rnnlms for multi-genre broadcast speech recognition," in *Proc. of Interspeech*, California, USA, 2016.

[17] S. Deena, R. W. Ng, P. Madhyashta, L. Specia, and T. Hain, "Semi-supervised adaptation of RNNLMs by fine-tuning with domain-specific auxiliary features," in *Proc. of Interspeech*, Stockholm, Sweden, 2017.

[18] S. Deena, M. Hasan, M. Doulaty, O. Saz, and T. Hain, "Recurrent neural network language model adaptation for multi-genre broadcast speech recognition and alignment," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 27, no. 3, pp. 572–582, 2019.

[19] O. Saz, S. Deena, M. Doulaty, M. Hasan, B. Khaliq, R. Milner, R. W. M. Ng, J. Olcoz, and T. Hain, "Lightly supervised alignment of subtitles on multi-genre broadcasts," *Multimedia Tools and Applications*, vol. 77, no. 23, pp. 30 533–30 550, 2018.

[20] M. Doulaty, R. Rose, and O. Siohan, "Automatic optimization of data perturbation distributions for multi-style training in speech recognition," in *Proc. of SLT*, California, USA, 2016.

[21] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, W. Karaiskos, Vasilis Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. of MLMI*, Bethesda, USA, 2006.

[22] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. of ICASSP*, Hong Kong, 2003.