Unsupervised Domain Discovery using Latent Dirichlet Allocation for Acoustic Modelling in Speech Recognition

Mortaza Doulaty, Oscar Saz and Thomas Hain

Speech and Hearing Group, Department of Computer Science, University of Sheffield

{mortaza.doulaty, o.saztorralba, t.hain}@sheffield.ac.uk

Motivation

- New applications and domains are becoming the target of research in automatic speech recognition
- New domains can be "found data", such as media and historical audio archives
- Domain for some recording is hard to assess, e.g. YouTube recordings
- Loss of accuracy would be large due to wrong modelling decision
- Expressing data as a mixture of domains can be a better solution
- Aims of this study:
  - Unsupervised discovery of domains
  - Trying to find the relation of latent domains with existing manually labelled domains and meta–data
  - Building/adapting latent domain models

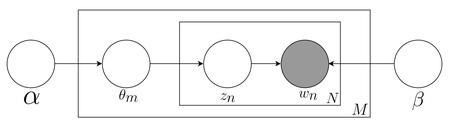
Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised probabilistic generative model for collections of discrete data

- Aims to describe how every item within the collection is generated
  - assuming there is a latent set of topics
  - each item is modelled as a finite mixture over those latent topics
- Mostly used for topic modelling of text corpora
- It can be applied to other types of data:
  - object categorisation and localisation in image processing (Sivic et al., 2005)
  - automatic harmonic analysis in music processing (Hu et al., 2009)
  - acoustic information retrieval in unstructured audio analysis (Kim et al., 2009)
- Joint distribution over the topic mixtures θ, set of N topics $\mathbf{z}$ and document $\mathbf{w}$:

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta)$$

- Graphical model representation of the joint distribution for the entire corpus:

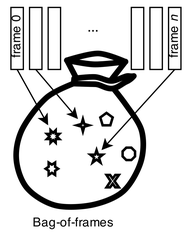


Training:

- Posterior distribution of the latent topic variables have some intractable integrals
- Variational approximation (approximated posterior)
- Markov–Chain Monte–Carlo based approaches

Unsupervised Domain Discovery

- Speech analogy of LDA:
  - words: audio frames
  - documents: segments
- Every frame needs to be represented by some discrete symbols, called acoustic "words"
- For a dictionary of size V, each frame is represented by $\hat{j}_t \in \{1...V\}$
- A Gaussian Mixture Model (GMM) is trained using Expectation Maximisation (EM) and mix–up procedure to reach the size V, enforcing the co–variance to be identity (equivalent to LBG-VQ)



- Means of Gaussian components used to created the codebook
- Assignment of frame x_t to codebook index j :

$$\hat{j}_t = \underset{j}{argmin} \, ||\mathbf{x}_t - \mathbf{m}_j||, \; j \in \{1...V\}$$

where $\mathbf{m}_j$ is the jth mixture component's mean vector

- Process:
  - Acoustic training data is used to train the LDA model
  - Same model can be used to infer the topics of test set segments
  - With latent domains, training/adaptation can be performed

Data Set Definitions

- 60 hours of speech from 6 domains were selected (10h each):
  - Radio (RD): BBCRadio4 broadcasts on February 2009
  - Television (TV): Broadcasts from BBC on May 2008
  - Telephone speech (CT): From the Fisher corpus
  - Meetings (MT): From AMI and ICSI corpora
  - Lectures (TK): From TedTalks
  - Read speech (RS): From the WSJCAM0 corpus
- Test set contains 1h from each of the domains (total 6h)
- Two types of features were used:
  - 13 dimensional PLPs plus Δ and $\Delta\Delta$
  - 26 dimensional bottleneck features extracted from a 4–hidden–layer DNN concatenated with 39D PLPs to form the 65D features
- Interpolated 3–gram language model contained 50k words used for decoding

Baseline

Features	Model	RS	RD	TK	CT	MT	TV	Total
PLP	ML	17.3	18.4	34.1	46.6	44.0	51.1	**36.0**
	ML Domain	16.9	19.1	35.1	44.4	44.0	52.9	**36.3**
	MAP	14.6	16.8	31.8	43.5	40.4	49.6	**33.6**
PLP+BN	ML	13.0	13.3	23.5	33.5	32.2	42.0	**26.8**
	ML Domain	12.6	14.0	25.0	34.3	33.2	44.0	**27.9**
	MAP	12.1	12.8	23.1	32.5	30.6	41.5	**26.2**

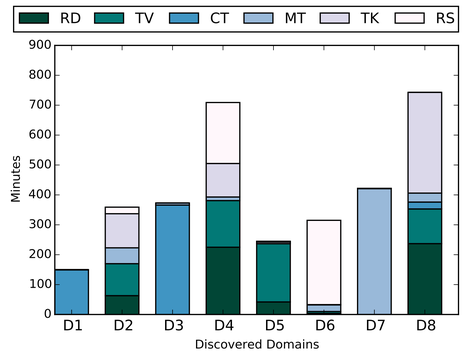
- MAP adapted models performed better than in–domain models
- MAP was chosen as a preferred setup for domain adaptation

Experiments

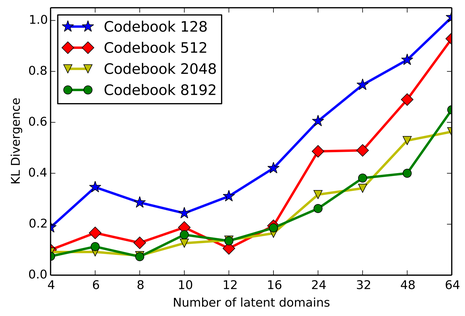
LDA Experiments

- Tuning model parameters:
  - K: number of domains
  - V: size of codebook
  - Codebooks of size 128 up to 8,192 were used
  - Domains of count 4 to 64 were used

Latent Domains' Data Distribution



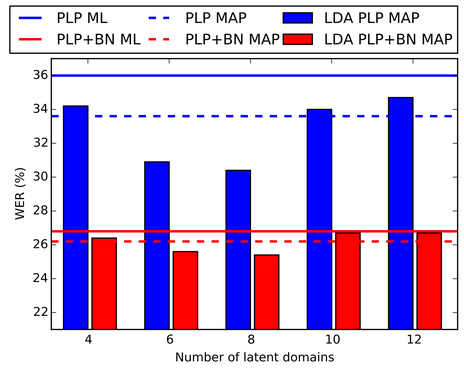
$K = 8$ and Codebook size = 2048



- KL divergence of latent domain distribution across-manually labelled domains in both training and test sets is used to measure the consistency of the latent domains discovered by LDA and how the distributions of data differs

Domain Adaptation

- Domain specific models are obtained using MAP adaptation



Error Variation Across Domains

- Lowest WER values with 30.4% and 25.4% for PLP and PLP+BN for $K = 8$
  - 16% and 5% relative improvements over the ML baseline
  - 10% and 3% relative improvements over the MAP (to human labelled domains)

Features	Model	RS	RD	TK	CT	MT	TV	Total
PLP	MAP	14.6	16.8	31.8	43.5	40.4	49.6	**33.6**
	LDA MAP	12.5	15.3	29.1	38.2	38.5	44.7	**30.4**
PLP+BN	MAP	12.1	12.8	23.1	32.5	30.6	41.5	**26.2**
	LDA MAP	11.9	12.8	22.3	31.1	31.0	41.0	**25.4**

- WER of LDA MAP models in latent discovered domains with both types of features

Features	D1	D2	D3	D4	D5	D6	D7	D8	Total
PLP	37.3	34.9	39.7	39.2	24.6	17.1	38.7	22.9	**30.4**
PLP+BN	33.9	29.2	30.4	32.8	19.7	12.6	30.9	19.2	**25.4**

Conclusion

- New unsupervised technique based on Latent Dirichlet Allocation proposed to discover the latent domains in highly diverse speech data
- The consistency of latent domains against manually labelled domains were studied
- Adapting to the latent domains were experimented and improvements of up to 16% over the baseline ML models and up to 10% over the MAP adapted models were achieved

Acknowledgement

This work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).