Language and the Shifting Sands of Domain, Space and Time

Timothy Baldwin

THE UNIVERSITY OF MELBOURNE

Talk Outline

Background

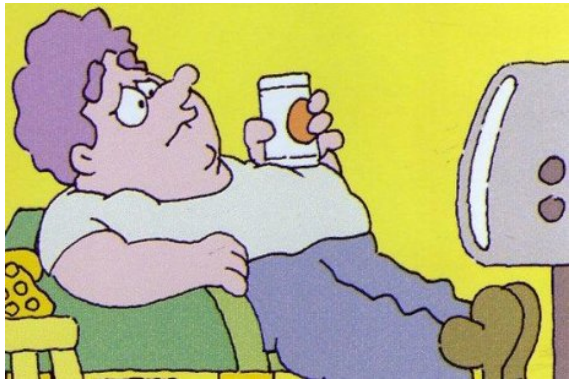
- Development/evolution of dialects/language varieties a fascinating process, encompassing:
  - politics (colonisation/imperialism/invasion, government structure/policy, armed conflict, ...)
  - anthropology (intra- and inter-community social structure, nature of agriculture, ...)
  - geography (impact of geography on language contact, cf. Japan, Switzerland, ...)
  - techno-sociology (media/nature of communication locally and internationally, ...)
- Research on dialects/language varieties a growing area of NLP (in large part thanks to VarDial!)

Trends in Dialectology

- In traditional dialectology, social factors were deliberately excluded, with the ideal consultant being [Wells, 1973, p45]:

  > *a man of seventy or so, still mentally alert and with an excellent memory for the days of his youth, a broad speaker, with an agricultural background, born in the village of native parents, married to a wife who is herself a native of the locality*

  = "NORMs": non-mobile, older, rural, male consultants

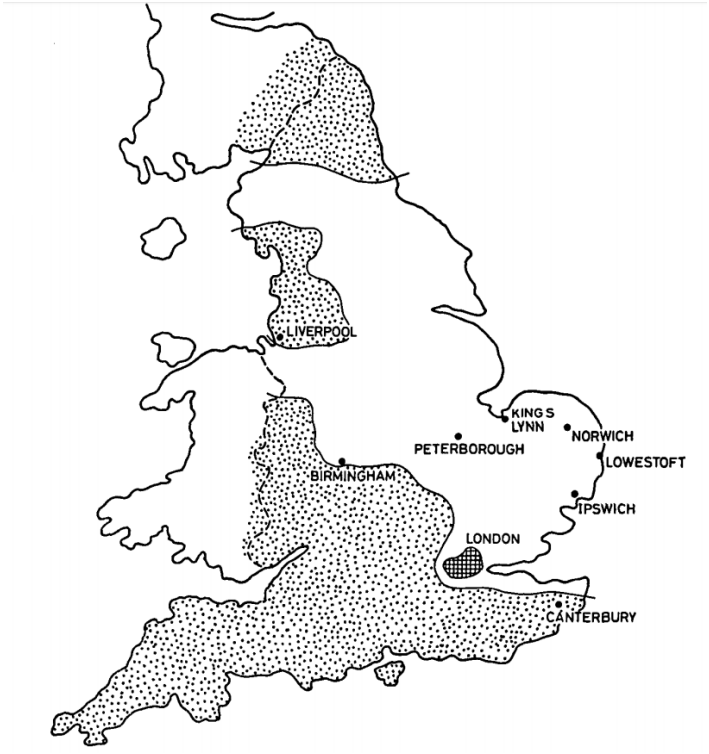


Trends in Dialectology

- Increasingly, however, the field has become predominantly quantitative [Reed and Spicer, 1952], focusing on the determination/impact of social variables [Labov, 1966], influence between dialects [Trudgill, 1986], "dialect acquisition" of individuals [Sibata, 1958, Payne, 1980, Chambers, 1992], and explanatory models of diachronic change [Trudgill, 1974]

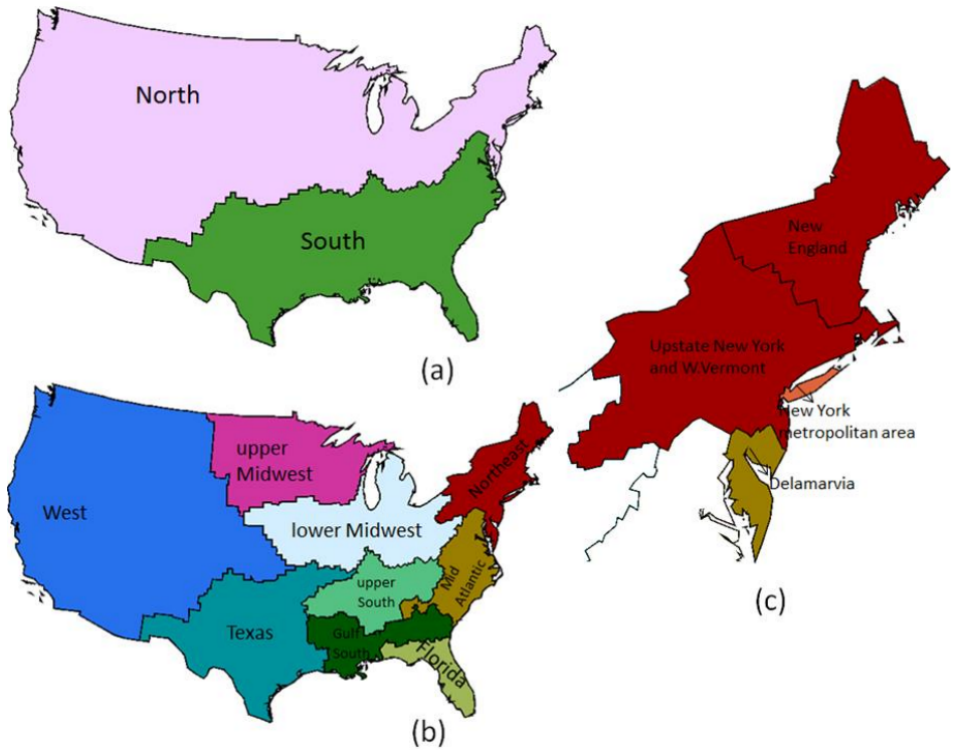
From Dialectology to Dialectometry

- In dialectology, "a pattern is assumed to exist, and then an isogloss is plotted" [Grieve et al., 2011, p2], e.g. in case of post-vocalic **/r/** [Orton et al., 1962], as rendered by Trudgill [1974]:



From Dialectology to Dialectometry

- Dialectometry, on the other hand, is data-driven, but similar in that it tends to be based on pre-identified sets of minimal pairs [Goebl, 2007, Nerbonne, 2006, Grieve et al., 2011]



VarDial and beyond

- Very healthy trend towards greater VarDial language and task variety (text, speech, variety-specific preprocessing, surprise varieties,...)
- ... but despite the fantastic work that has been done, do we understand the different dialects any better as a result of it/are we modelling dialect in all its social dimensions?

VarDial and beyond

- As with many NLP tasks, question of how much the models are learning about dialect/language variety vs. domain anomalies in training data, and how well the models generalise

> The Australian men's cricket team was bowled out for a low total

vs.

> A howler by the Aussie blokes #howzat #cmonaussie

- Related to this, more to be done in terms of modelling/evaluating the certainty of a given document being in a given dialect

VarDial and beyond

- Fascinating opportunities for jointly learning lexical alternation sets in dialectometry, and predicting dialectal shifts/influence
- Task carried out in social isolation, despite dialect/language variety being an inherently social construct (esp. when the focus is text)
- Little overlap between VarDial, work on user/message-level geolocation [Eisenstein et al., 2010, Roller et al., 2012, Han et al., 2014], computational sociolinguistics [Eisenstein et al., 2011, Bamman et al., 2014, Nguyen et al., 2016], diachronic language change [Cook et al., 2014, Frermann and Lapata, 2016, Rosenfeld and Erk, 2018], and dialectometry ... lots of mutual opportunities for cross-fertilisation

Overview of this Talk

- Patchwork of recent work on the seemingly disparate topics of:
  - domain debiasing
  - user geolocation
  - dialect map generation

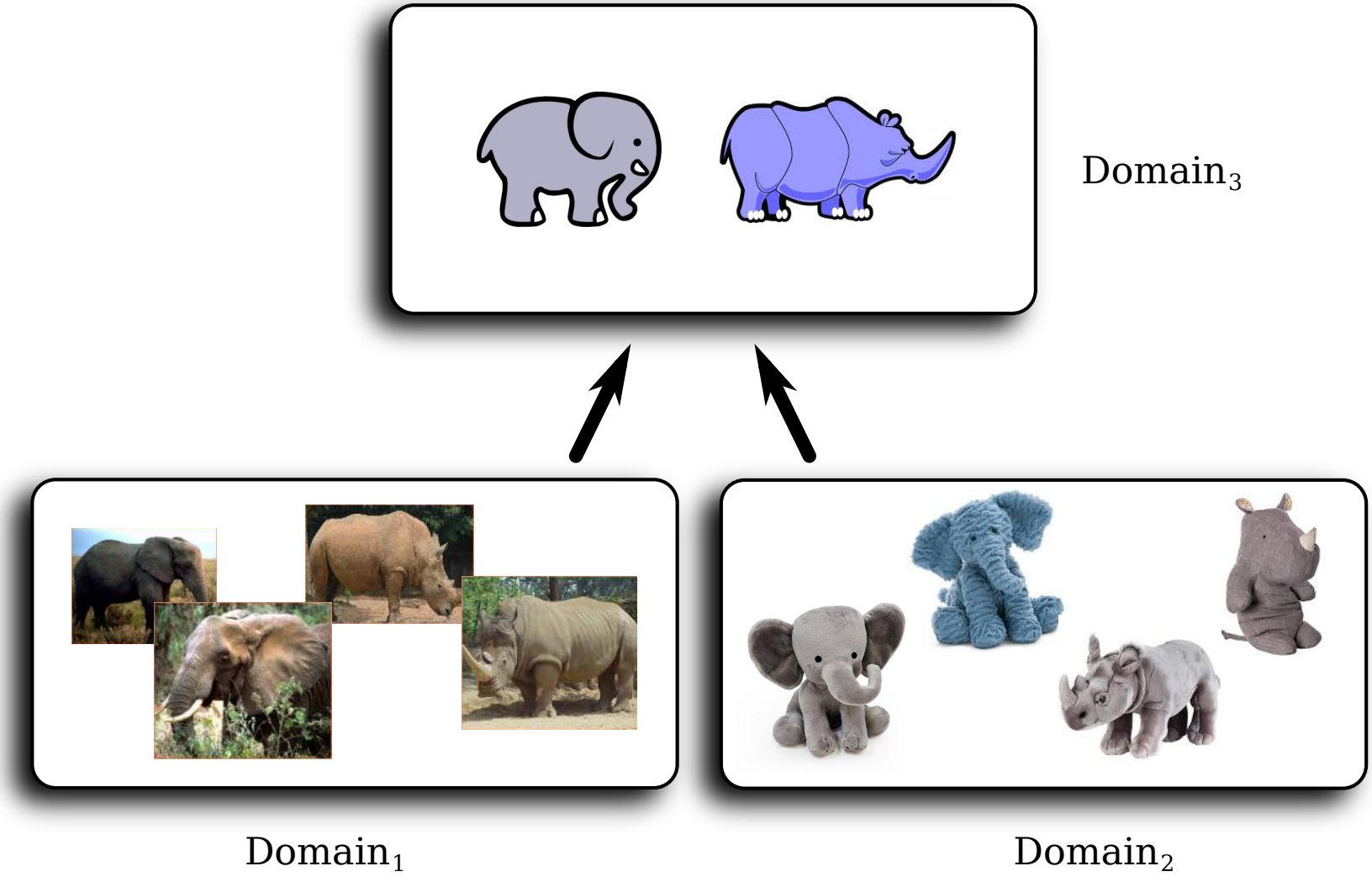
  all of which are individually directly relevant to the broader task of dialect/language variety identification

Talk Outline

Introduction

- **Background:** real-world language problems require learning from heterogeneous corpora
- **Aim:** learn robust models that generalise both *in-domain* and *out-of-domain*
- **Experimental setup:** train models on several domains, and test on unknown heldout domains, which we do not have prior knowledge of

Data Setting: Multiple Source Domains



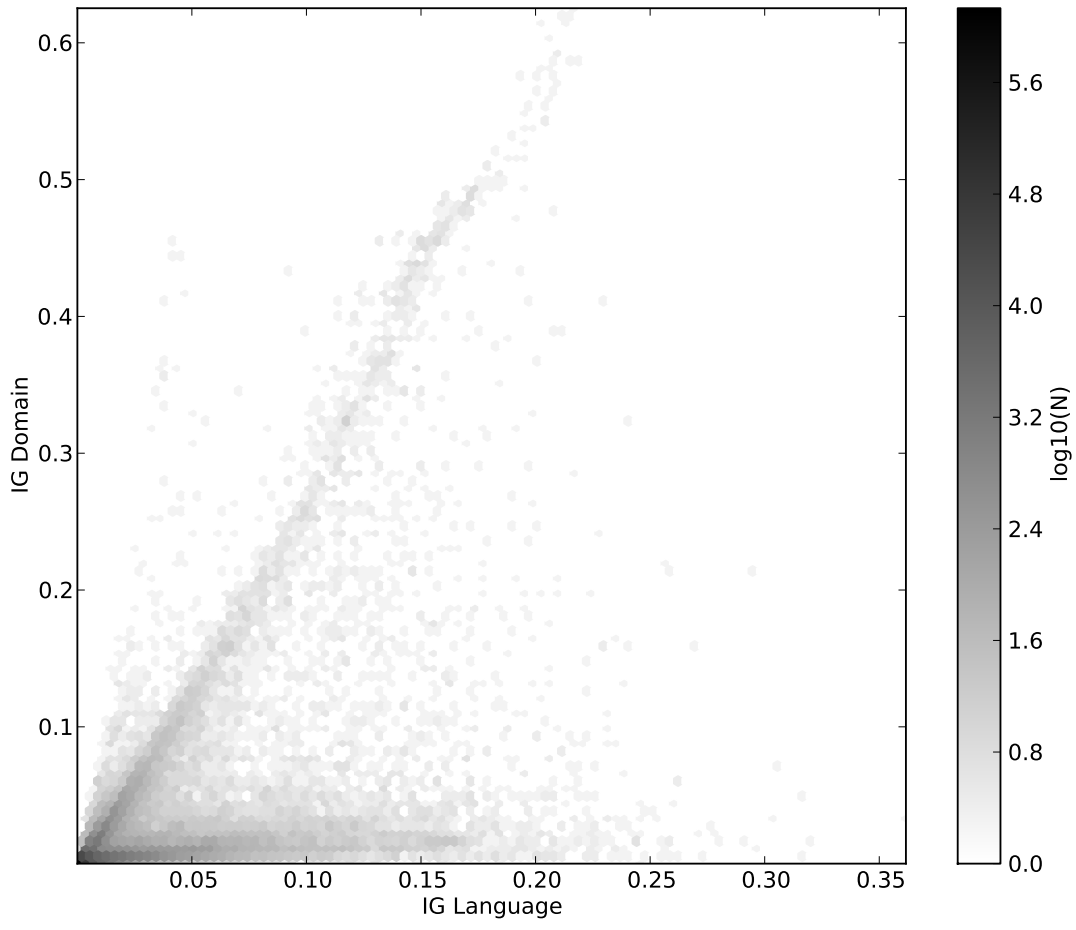
Approach

- In training, jointly optimise accuracy over primary task, and *lack of* accuracy at discriminating the source domain

  $\Rightarrow$ force model to generalise the document representation across domains, rather than learn idiosyncrasies of individual domains

- Similar data setup to that used in `langid.py` for explicit feature selection ...

`langid.py`: Language- vs. Domain-based IG



`langid.py`: What does this Mean?

- There are two distinct groups of features: (1) $\mathcal{IG}$ for language is strongly correlated with that for domain; and (2) $\mathcal{IG}$ for language is largely independent of that for domain

the second of these is what we are interested in

- Automatically detect language- (and not domain-) associated features:

$$\mathcal{LD}^{all}(t) = \mathcal{IG}^{all}_{lang}(t) - \mathcal{IG}_{domain}(t)$$

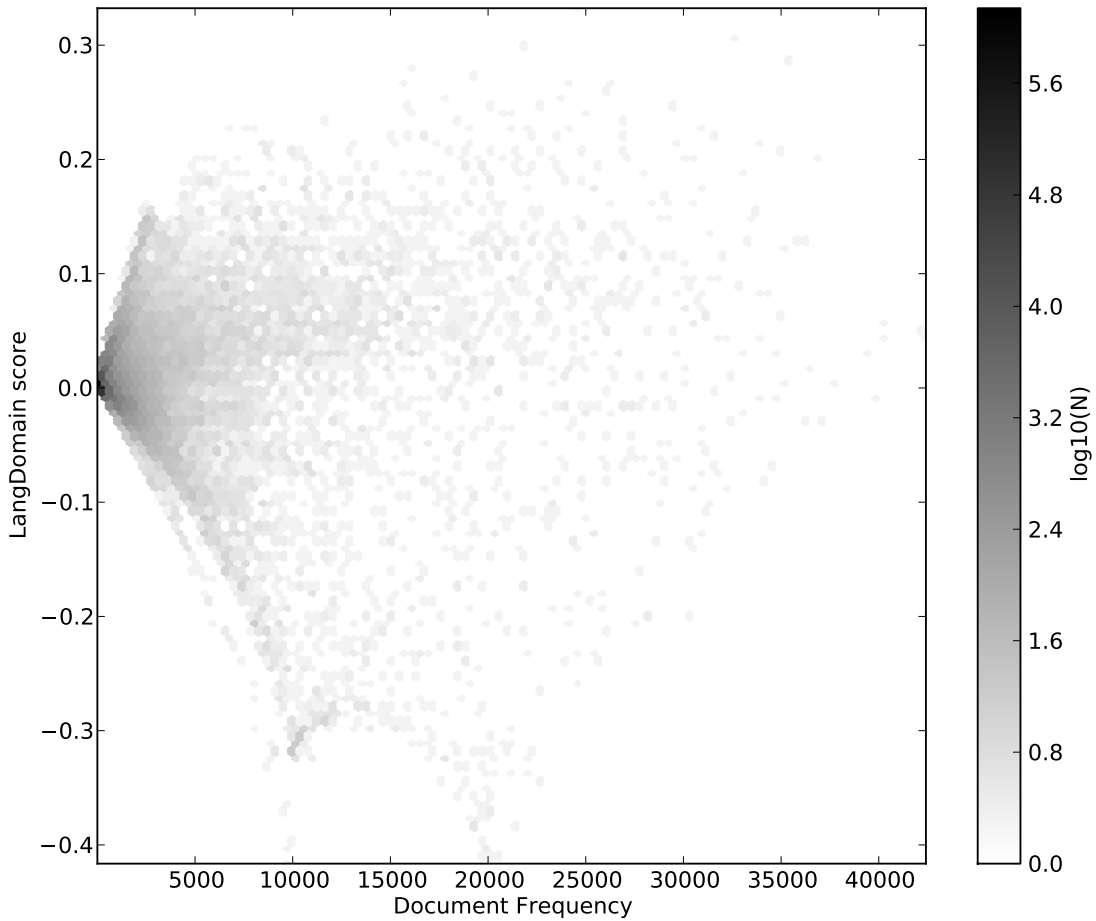
Source(s): Lui and Baldwin [2011, 2012]

`langid.py`: Computational Bottleneck

- Calculating the $\mathcal{IG}$ for low-order n-grams is fine, but it quickly becomes intractable for larger values of n
- Ideally, we want a method which scales to (very) large numbers of features/high n-gram orders, with little computational overhead

Source(s): Lui and Baldwin [2011, 2012]

`langid.py`: DF vs. $\mathcal{LD}$



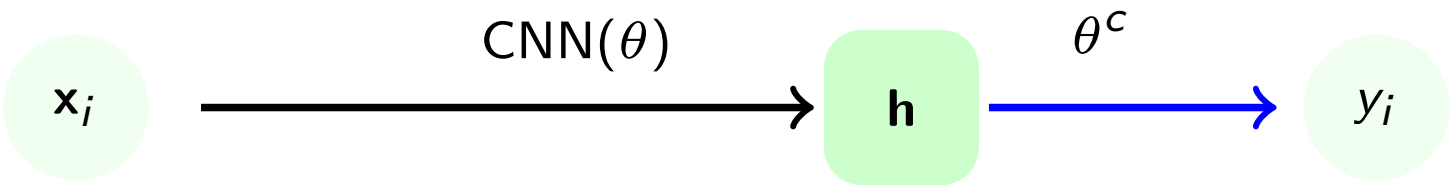
`langid.py`: Observation

- Low DF is a good predictor of low $\mathcal{LD}$ (but not vice versa)
- As $\mathcal{DF}$ is much cheaper to compute, we first identify the 15000 features with highest $\mathcal{DF}$ for a given n-gram order, and assign a $\mathcal{LD}$ score of 0 to all features outside this set
- The final feature representation is a combination of the top-N features for each a predefined set of n-grams

Source(s): Lui and Baldwin [2011, 2012]

New Approach: Baseline

- Baseline model = straight CNN [Kim, 2014]

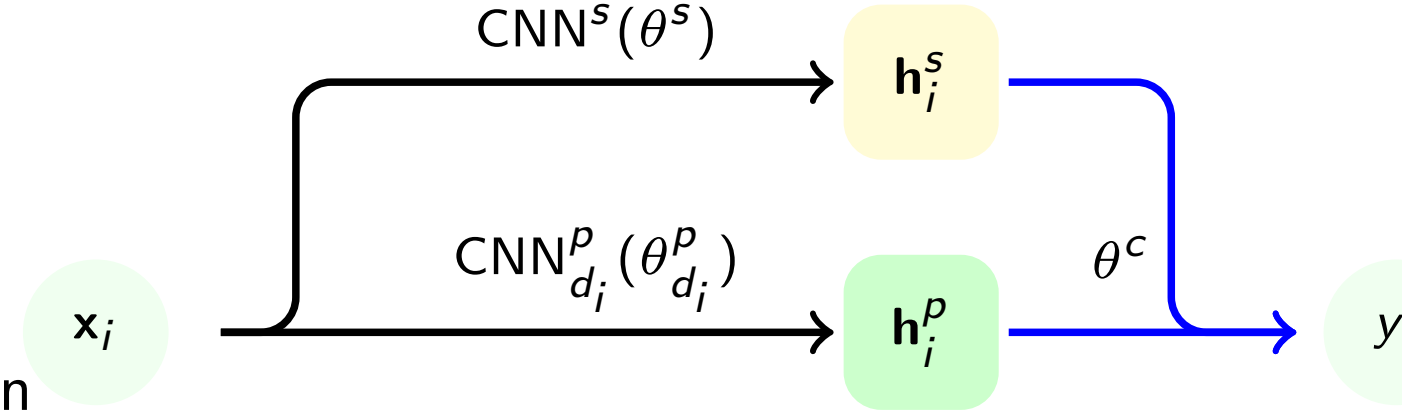


Source(s): Li et al. [2018]

New Approach 1: Domain-conditional Model ("Cond")

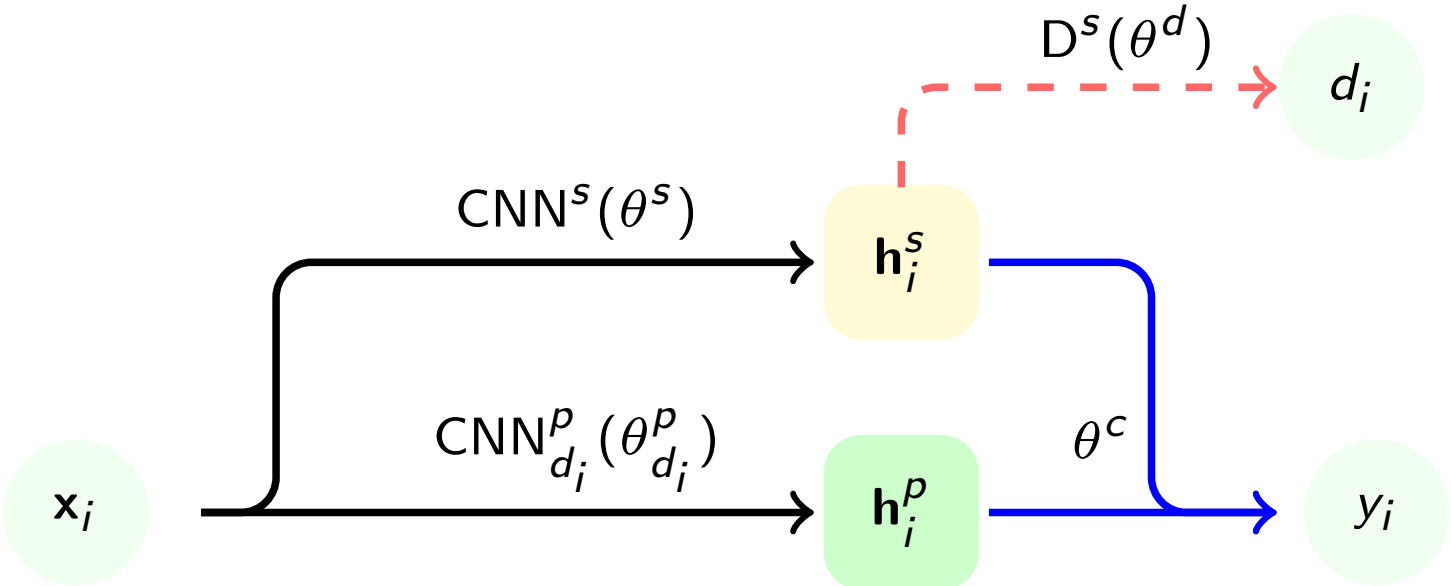
- Basic intuition: take inspiration from Daumé III [2007] in learning two representations of each instance $\mathbf{x}$:
  1. shared representation $\mathbf{h}_i^s$, using a shared CNN^s
  2. private representation $\mathbf{h}_i^p$ conditioned on domain identifier d_i of $\mathbf{x}$

  and concatenate the two to generate overall document representation



New Approach 1: Domain-conditional Model ("Cond")

- In order to avoid contamination of the shared representation with domain-specific concepts, optionally add adversarial discriminator [Goodfellow et al., 2014, Ganin et al., 2016] to force generalisation:



New Approach 1: Domain-conditional Model ("COND")

- Overall training objective:

$$\mathcal{L}^{\text{COND}} = \min_{\theta^c, \theta^s, \{\theta^p_\cdot\}} \max_{\theta^d} \mathcal{X}(\mathbf{y}|\mathbf{H}^s, \mathbf{H}^p, \mathbf{d}; \theta^c) \underbrace{-\lambda_d \mathcal{X}(\mathbf{d}|\mathbf{H}^s; \theta^d)}_{d}$$

where:

- $\mathbf{H}^s = \{\mathbf{h}^s_i(\mathbf{x}_i)\}_{i=1}^n$ = the shared representations for all instances
- $\mathbf{H}^p = \{\mathbf{h}^p_i(\mathbf{x}_i, d_i)\}_{i=1}^n$ = the private representations for all instances

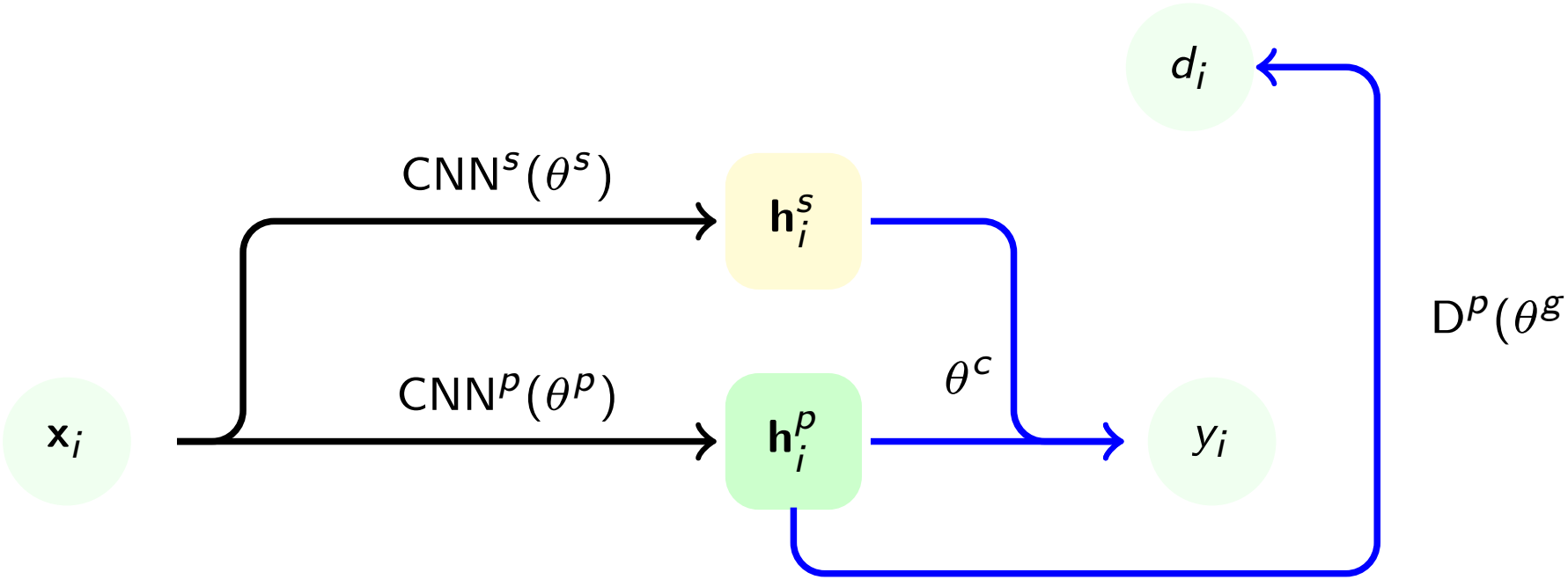
New Approach 1: Domain-conditional Model ("Cond")

- Train discriminator to be maximally accurate wrt θ^d, and maximally *inaccurate* wrt $\mathbf{H}^s$, based on gradient reversal during backpropagation [Ganin et al., 2016].
- At test time, select domain with lowest entropy wrt test instance

New Approach 2: Domain-generative Model ("GEN")

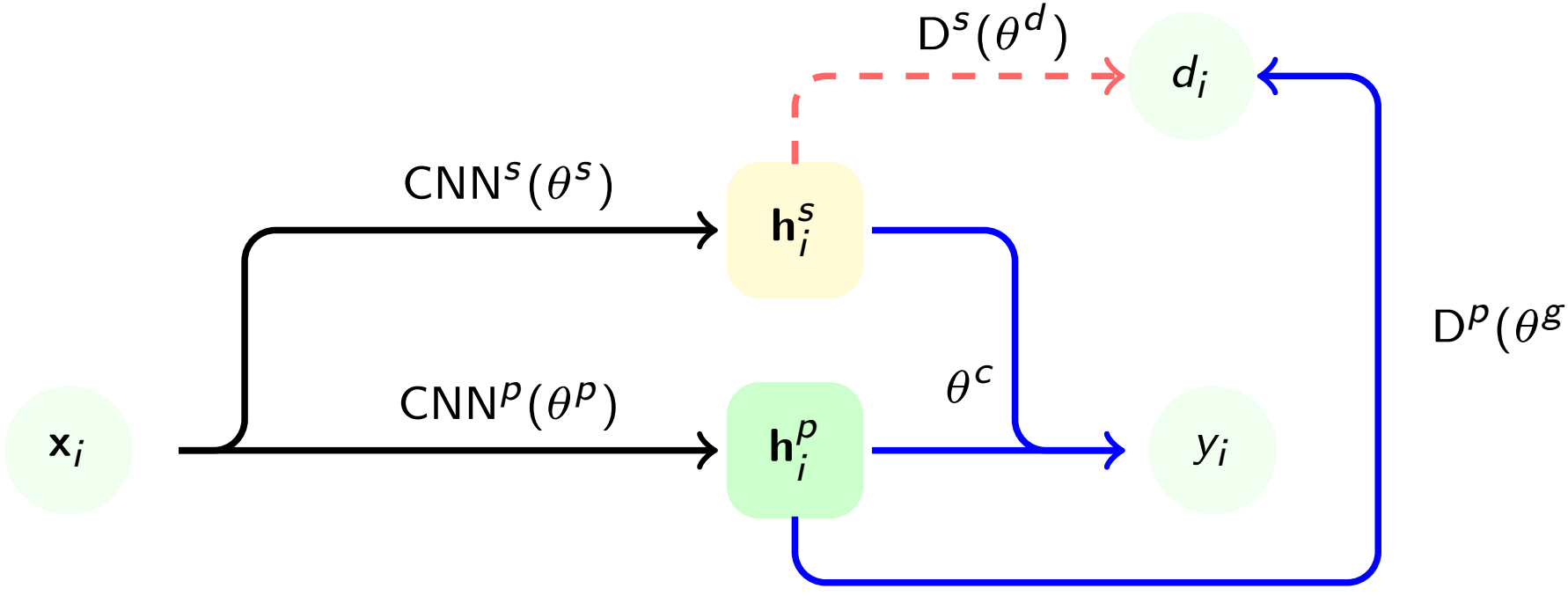
- Basic intuition: largely the same as Approach 2, but *generate* the domain (based on multi-task learning) rather than conditioning on it, by:
  1. computing $\mathbf{h}^p$ using a *single* CNN^p rather than several domain-specific CNNs
  2. using the private representation to predict the domain, encouraging differentiation between the domain-general and domain-specific representations

New Approach 2: Domain-generative Model ("GEN")



New Approach 2: Domain-generative Model ("GEN")

- Similarly to COND, optionally add an adversarial discriminator:



New Approach 2: Domain-generative Model ("GEN")

- Overall training objective:

$$\mathcal{L}^{\text{GEN}} = \min_{\theta^c,\theta^s,\theta^p,\theta^g} \max_{\theta^d} \mathcal{X}(\mathbf{y}|\mathbf{H}^s,\mathbf{H}^p;\theta^c) - \lambda_d \mathcal{X}(\mathbf{d}|\mathbf{H}^s;\theta^d) \underbrace{+\lambda_g \mathcal{X}(\mathbf{d}|\mathbf{H}^p;\theta^g)}_{g}$$

where:

- $\mathbf{H}^s = \{\mathbf{h}^s_i(\mathbf{x}_i)\}_{i=1}^n$ = the shared representations
- $\mathbf{H}^p = \{\mathbf{h}^p_i(\mathbf{x}_i)\}_{i=1}^n$ = the private representations

Experimental Setup

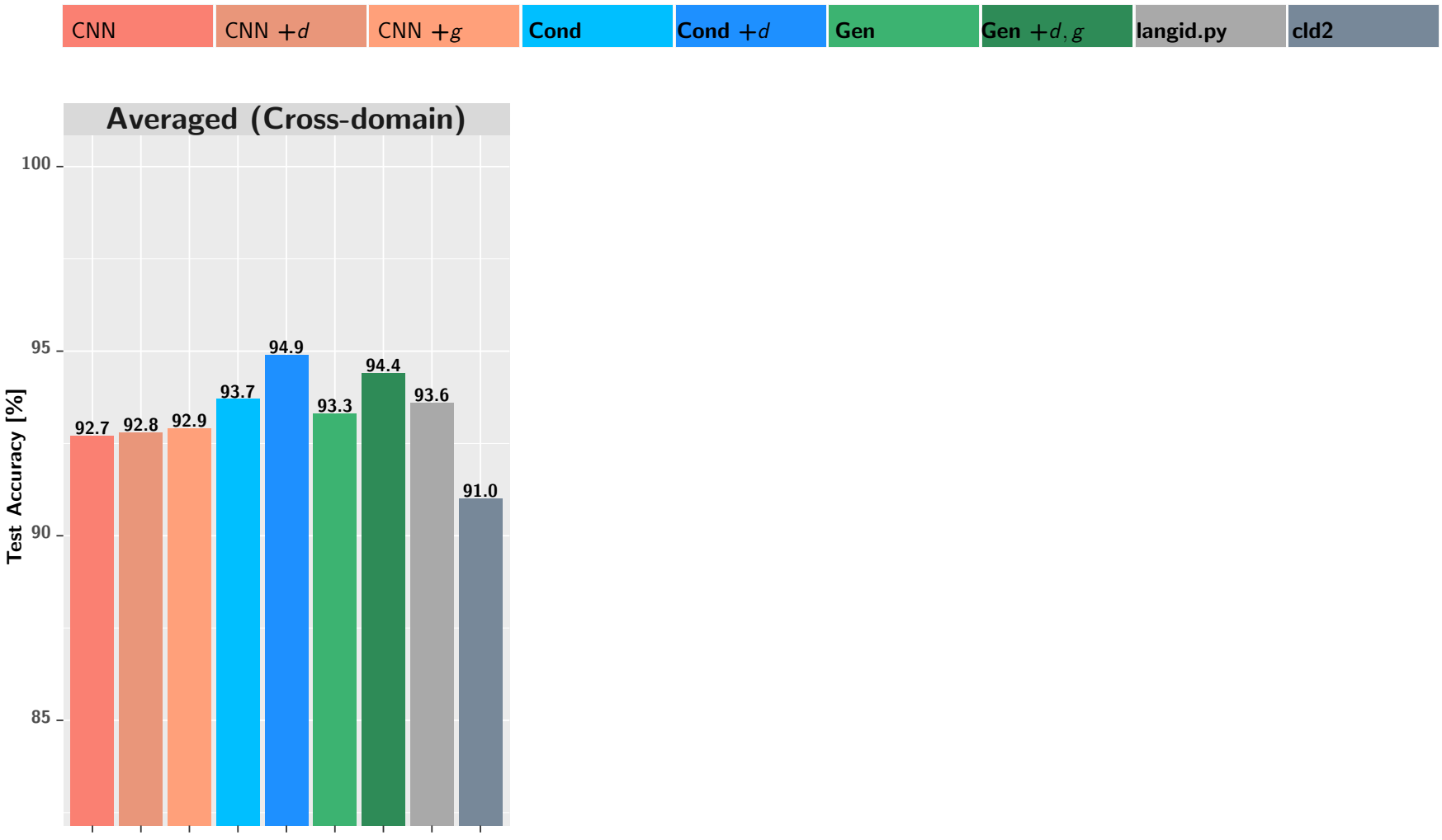
Task: document-level language identification

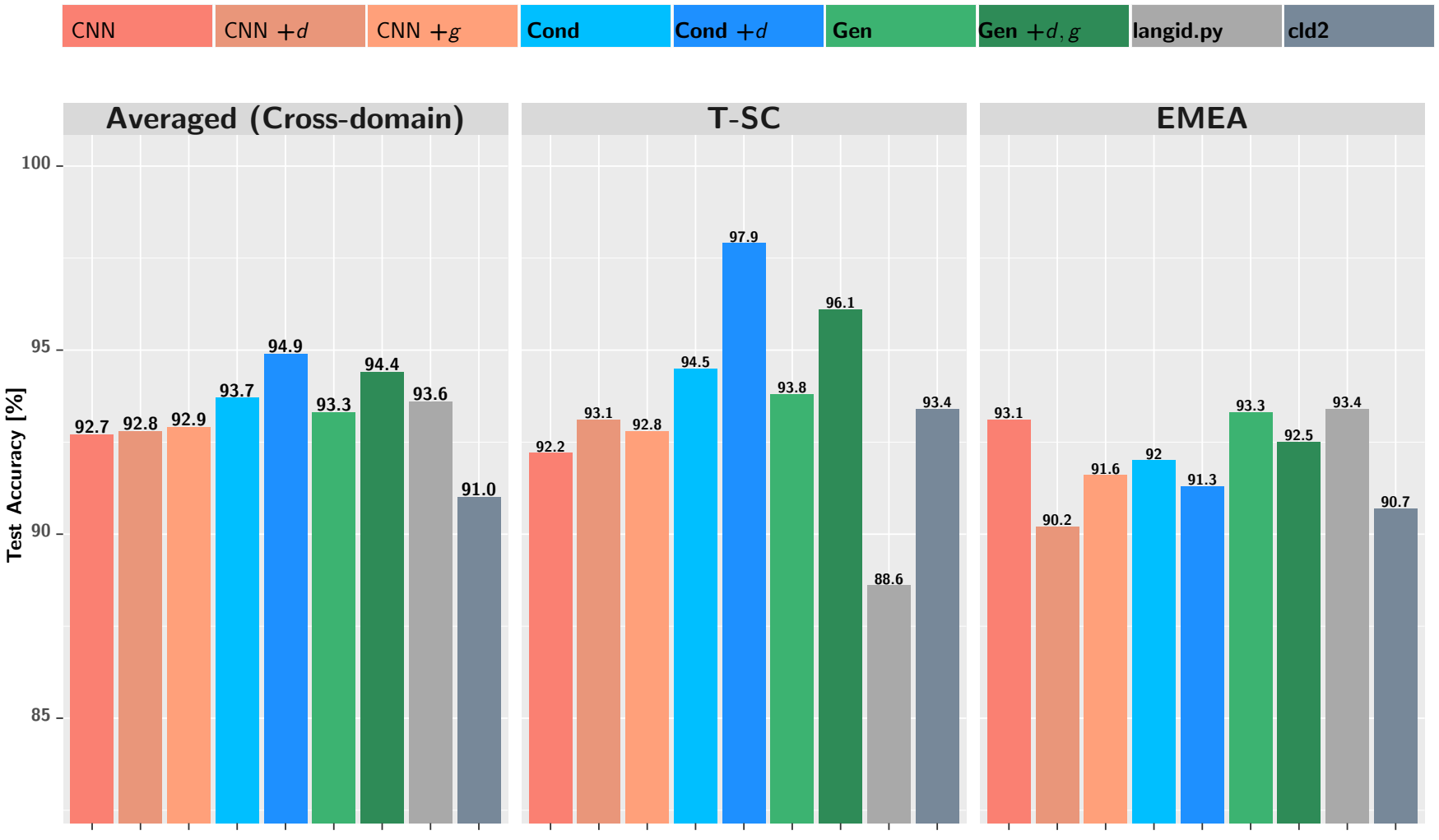
Target: 97 languages

Model: byte-level CNN (up to 1k bytes)

Datasets:
- **5** training domains [Lui and Baldwin, 2011]
- **7** heldout test domains

Evaluation: accuracy for both in-domain and cross-domain settings


CNN
CNN +d
CNN +g
Cond
Cond +d
Gen
Gen +d, g
langid.py
cld2
Averaged (Cross-domain)
Test Accuracy [%]
100
95
90
85
92.7
92.8
92.9
93.7
94.9
93.3
94.4
93.6
91.0


CNN
CNN +d
CNN +g
Cond
Cond +d
Gen
Gen +d, g
langid.py
cld2
Averaged (Cross-domain)
T-SC
EMEA
Test Accuracy [%]
100
95
90
85
92.7
92.8
92.9
93.7
94.9
93.3
94.4
93.6
91.0
92.2
93.1
92.8
94.5
97.9
93.8
96.1
88.6
93.4
93.1
90.2
91.6
92
91.3
93.3
92.5
93.4
90.7

Summary

- Methods for multi-domain generalisation, taking the domain as either an input (COND) or output (GEN), optionally with adversarial training over private domain representation
- In all cases, adversarial loss leads to large gains, esp. in terms of out-of-domain performance

Talk Outline

VarDial and User Geolocation

- User geolocation is the task of predicting the "home" location of a given individual based on their posted content, and optionally social interactions
- VarDial under a different name? Yes and no ...

To Infinity — Classification [Rahimi et al., 2017a]

Geolocation as Regression

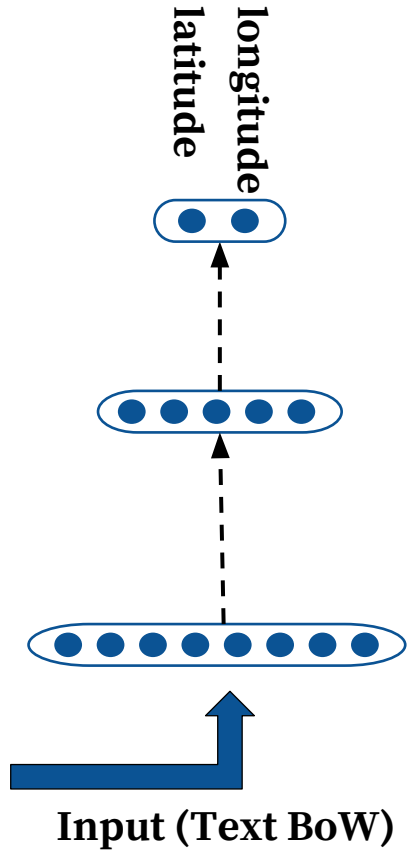
Concatenated User Tweets

Russell Peters @therealrussellp · May 31
Soooo... My daughter drew a picture of a "Scientist Potion" 👩‍🔬👩‍🔬👩‍🔬😂😂😂😂
… instagram.com/p/BUx9Q9eAP8B/
31 24 209

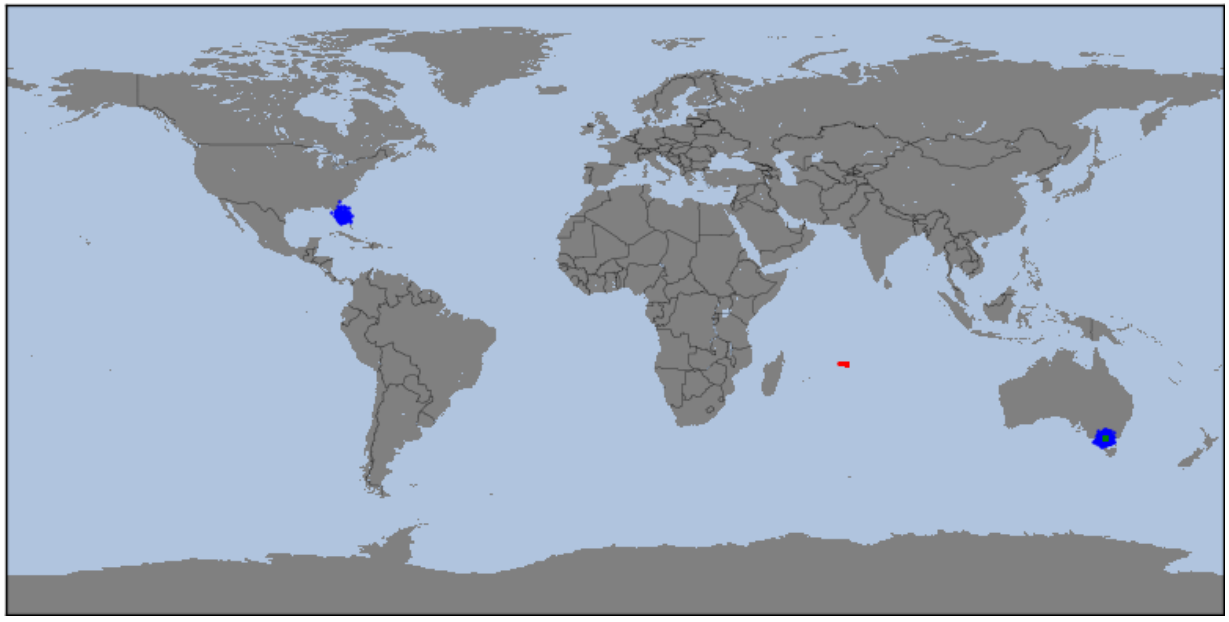
Russell Peters @therealrussellp · May 28
I'm trying to lose weight and get results but this might be too much change!…
instagram.com/p/BUo-I_hgFnt/
19 14 122

Russell Peters @therealrussellp · 30 Aug 2016
Just introduced my friend and actual musical genius @nilerodgers in Calgary!
instagram.com/p/BJwWsGfgV92/
6 6 76

Russell Peters @therealrussellp · 13 Jun 2016
Why does my coffee look scared?! #McCafe #DontWorryIllBeGentle
instagram.com/p/BGnQuZWoNDG/
9 10 59



Issues with Mean Squared Error Regression



- training samples
- predictions (MSE regression)
- predictions (classification)

From regression
with continuous labels

To classification
with discrete labels

Geolocation as Classification

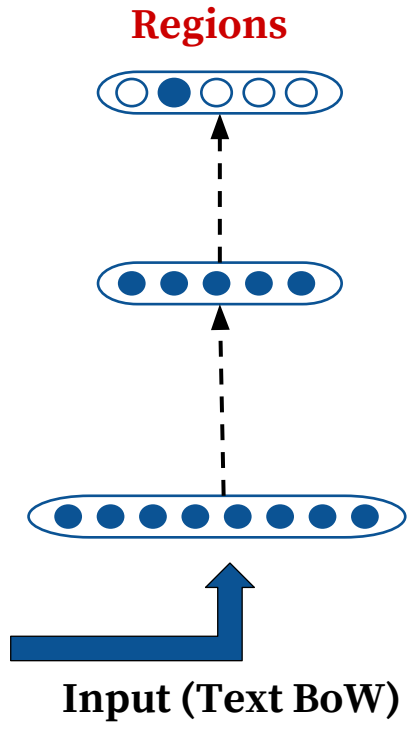
Concatenated User Tweets

Russell Peters @therealrussellp · May 31
Soooo... My daughter drew a picture of a "Scientist Potion" 👩‍🔬👩‍🔬👩‍🔬😂😂😂😂
… instagram.com/p/BUx9Q9eAP8B/
31 24 209

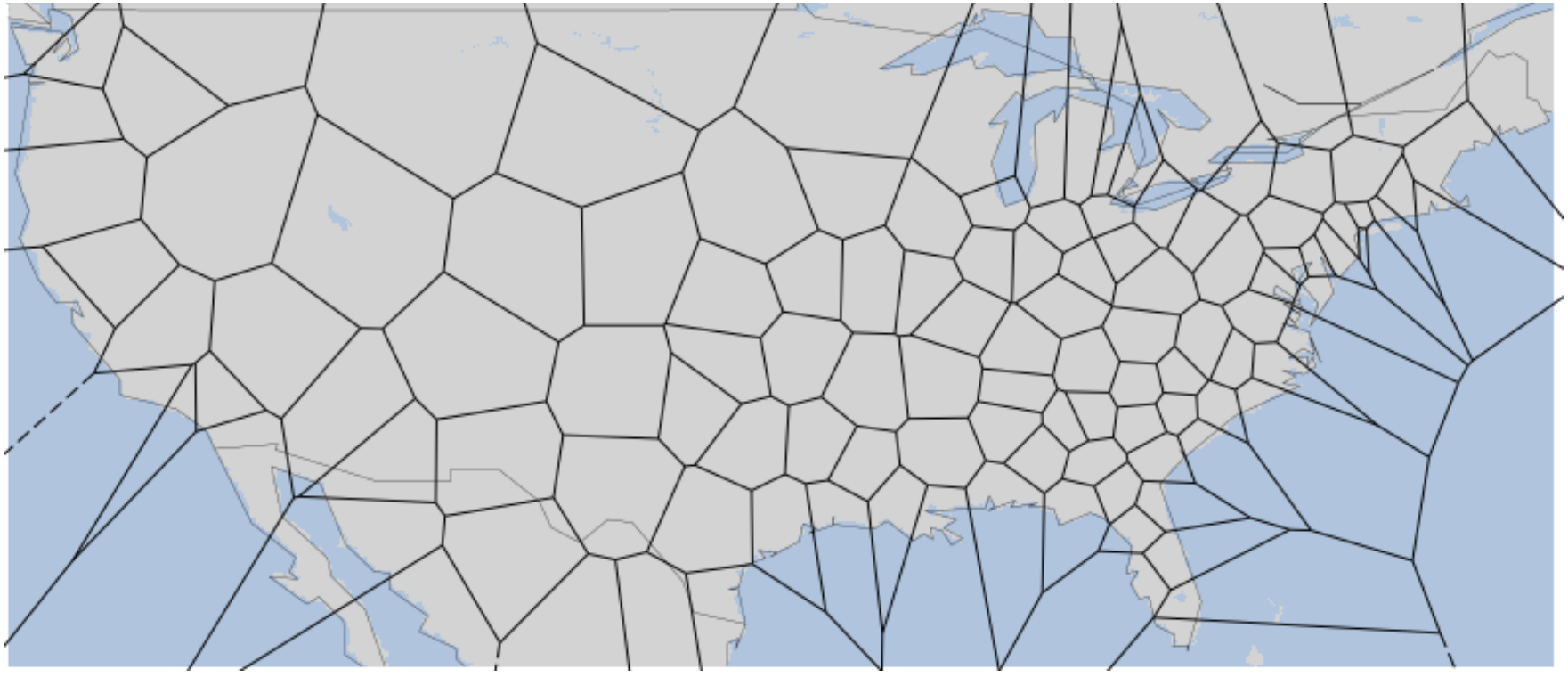
Russell Peters @therealrussellp · May 28
I'm trying to lose weight and get results but this might be too much change!…
instagram.com/p/BUo-I_hgFnt/
19 14 122

Russell Peters @therealrussellp · 30 Aug 2016
Just introduced my friend and actual musical genius @nilerodgers in Calgary!
instagram.com/p/BJwWsGfgV92/
6 6 76

Russell Peters @therealrussellp · 13 Jun 2016
Why does my coffee look scared?! #McCafe #DontWorryIllBeGentle
instagram.com/p/BGnQuZWoNDG/
9 10 59



Discrete Location Representation: *k*-means Clustering

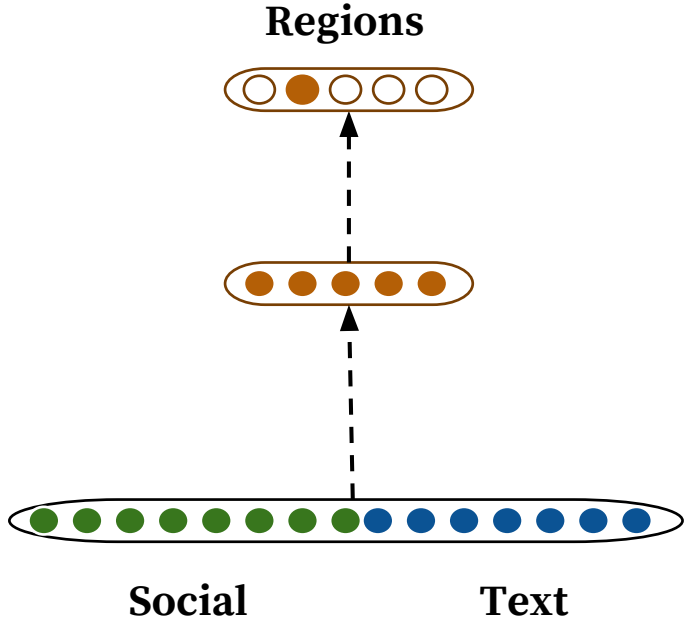


Slightly better median error than *k*-d tree [Roller et al., 2012]

Combining Text and Social Information: Feature Concatenation

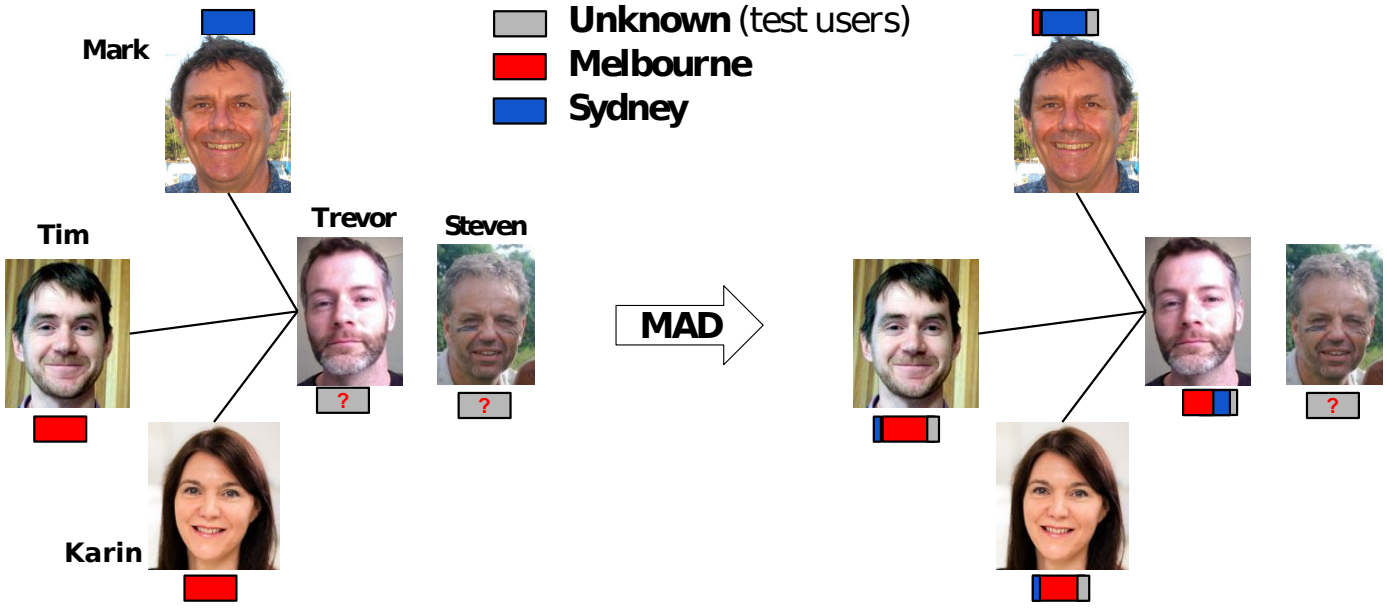
Social features: @-mentions

Text features: words + hashtags



Feature Concatenation $\Rightarrow$ Suboptimal Performance

Social Graph-based Model: Modified Adsorption (MAD)

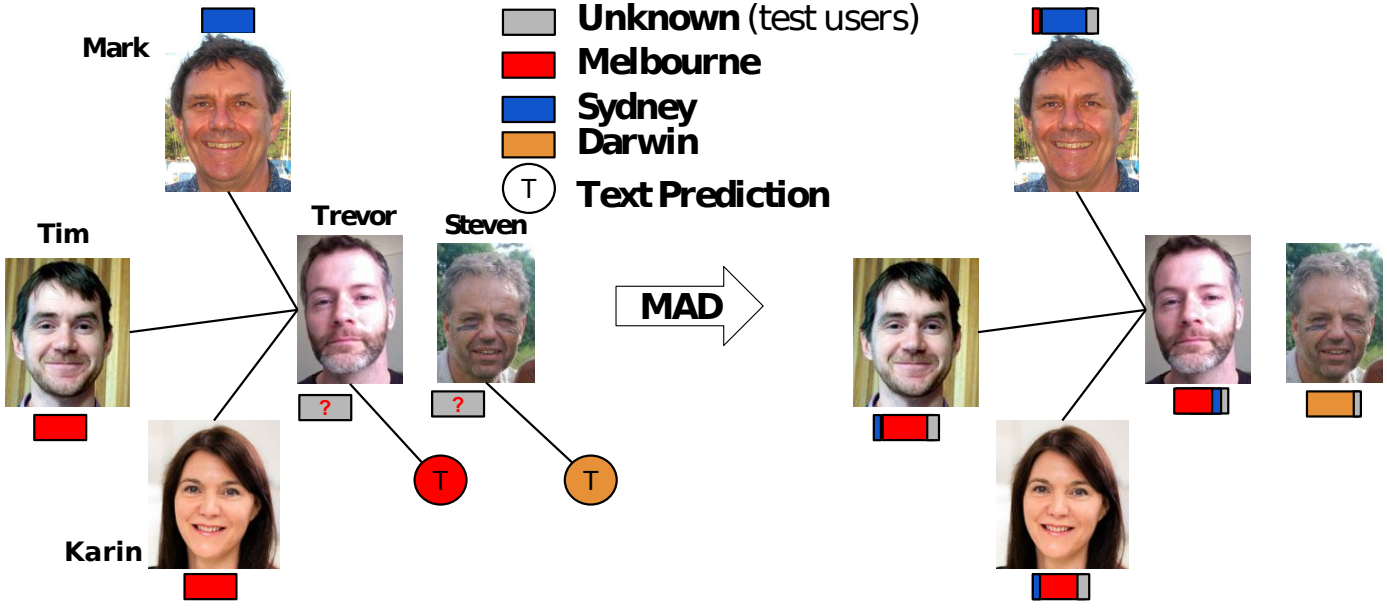


@-mention graph of training and test users

You are where your friends are.

Modified Adsorption ("MAD": Talukdar and Crammer [2009])

Text+Social

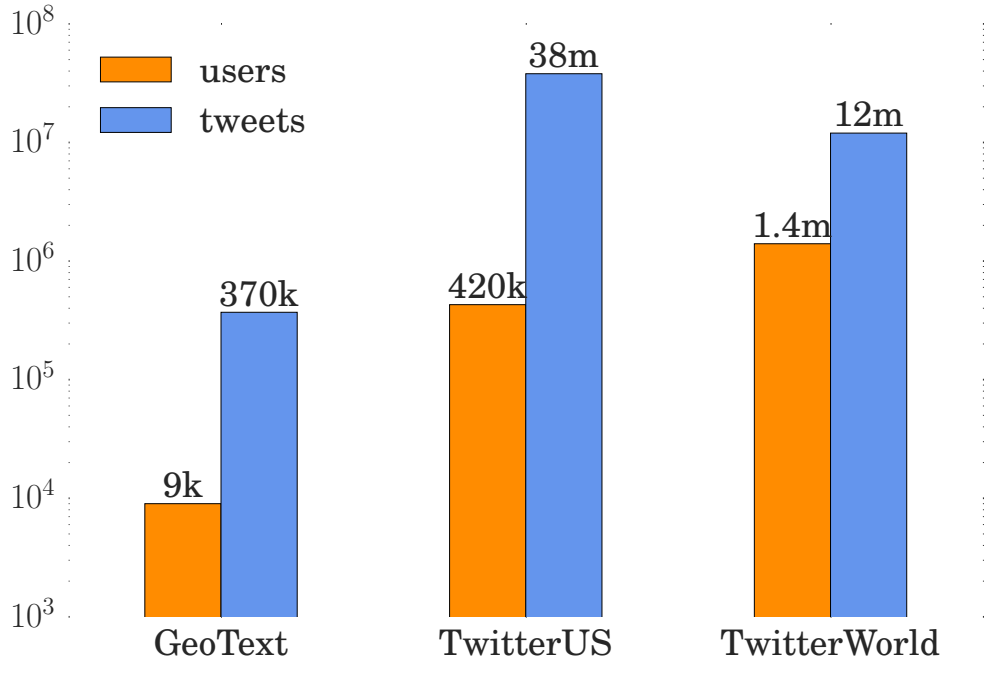


@-mention graph with added text dongles

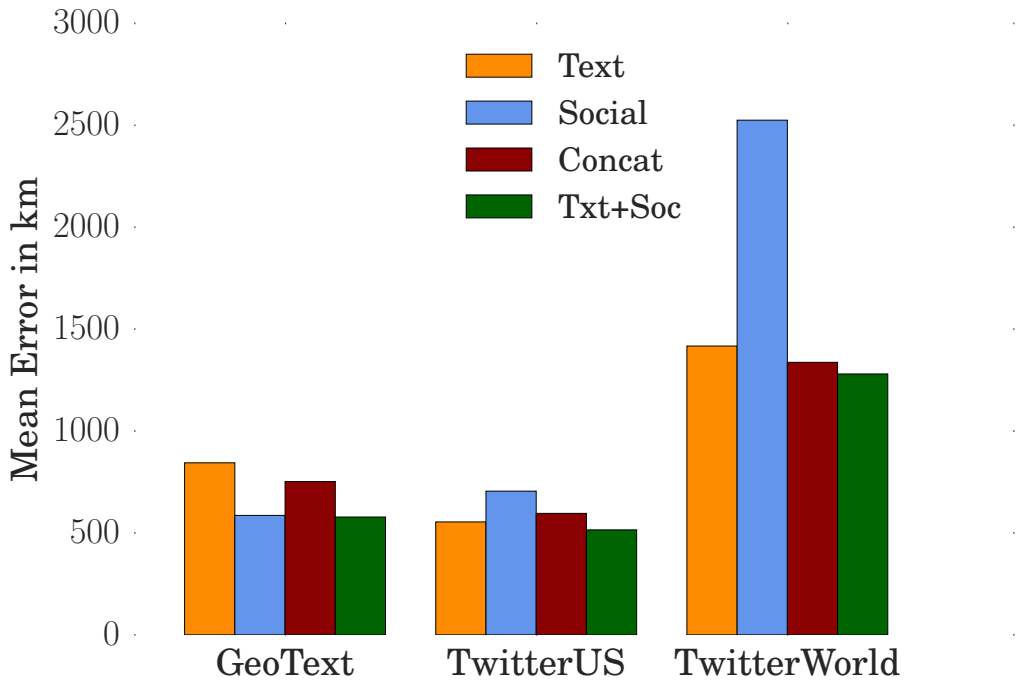
What to do with isolated users?

Connect the text-based predictions to test users before label propagation.

Twitter Geolocation Datasets

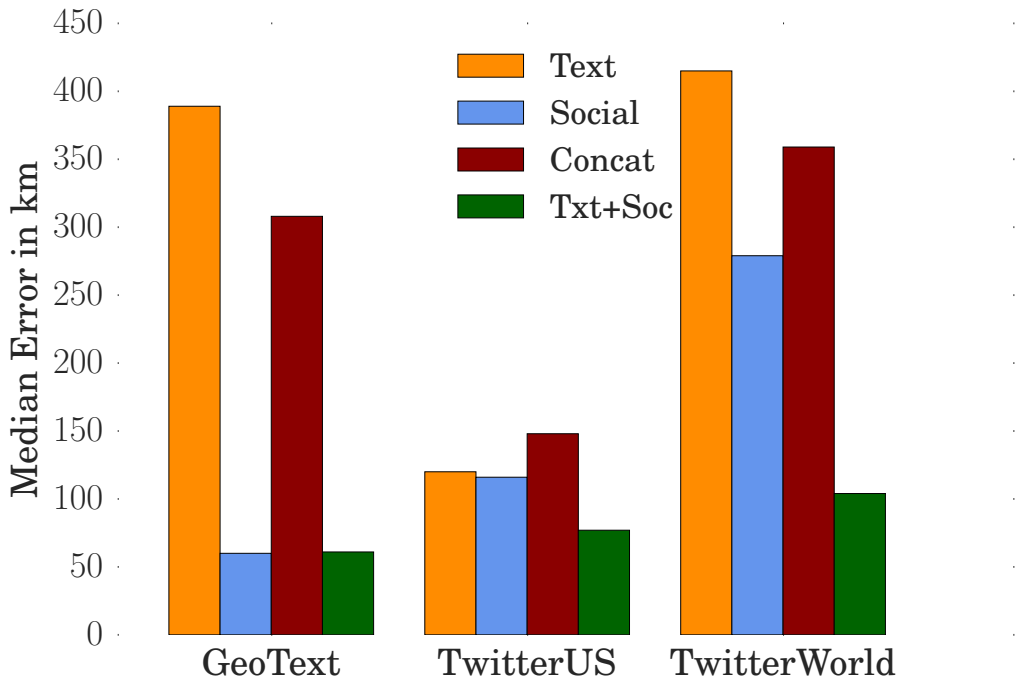


Results: Mean Error (km)



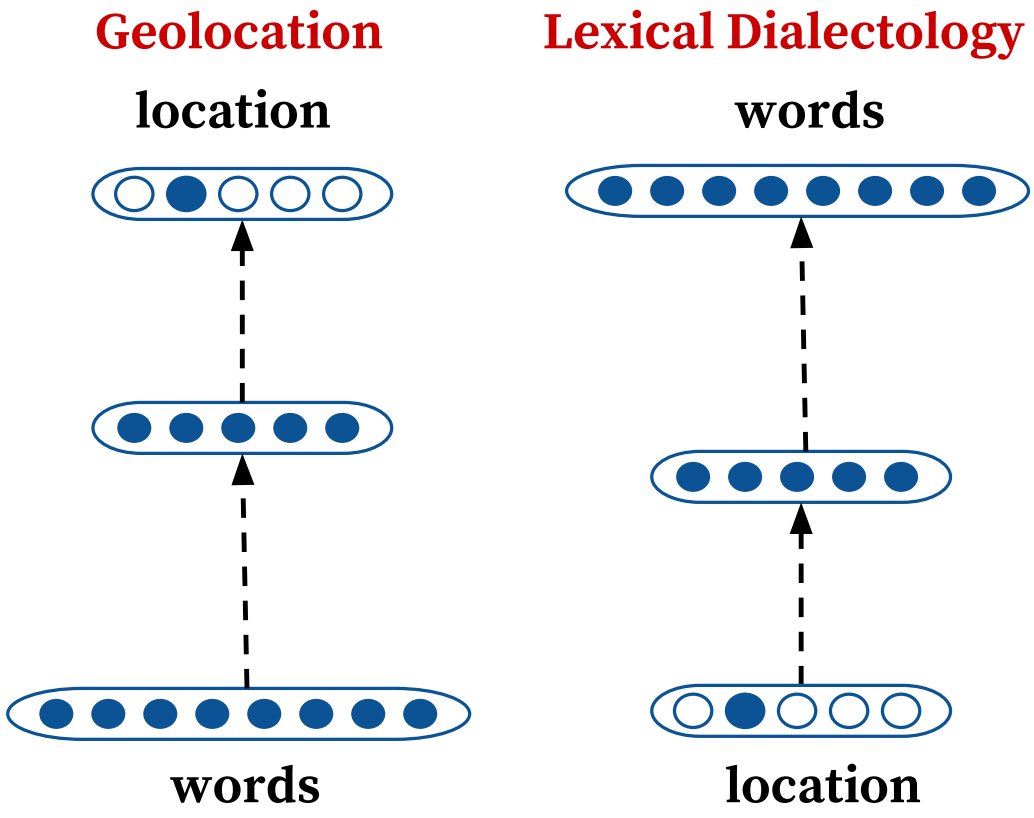
Network-based methods have higher mean error because of the isolated users.

Results: Median Error (km)



Network-based methods outperform text-based methods in median error.

Geolocation vs. Lexical Dialectology



DARE Dictionary

"DARE" [Cassidy, 1985] ... not machine readable!

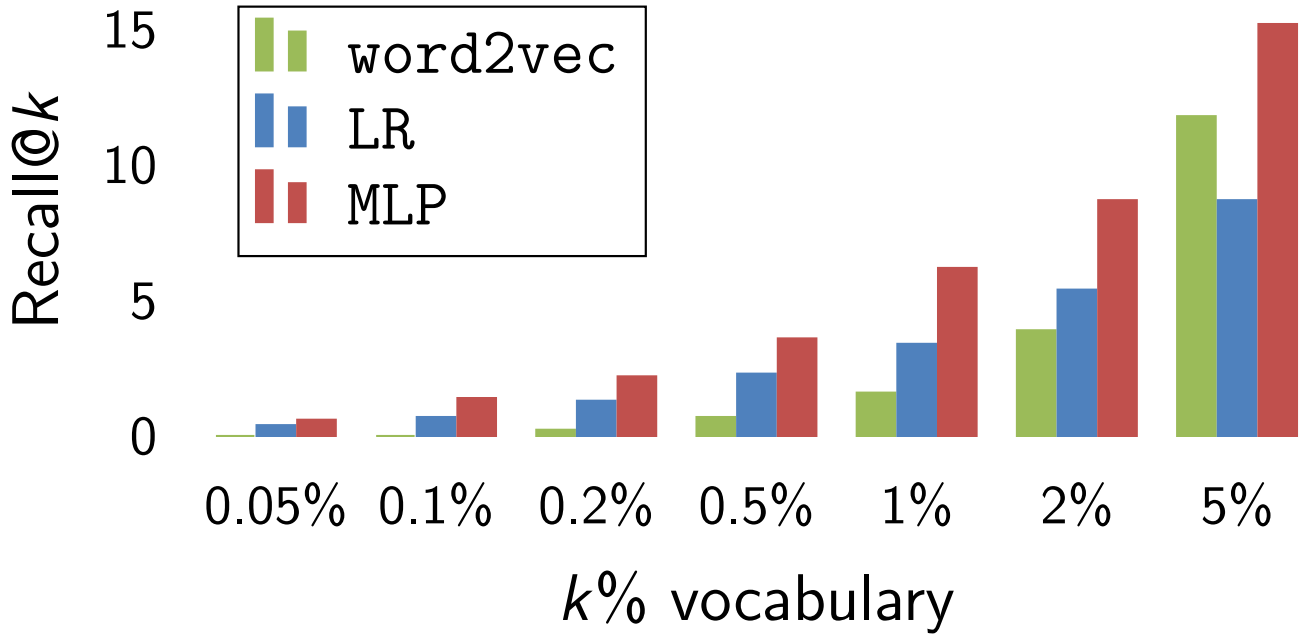
DAREDS


```
{
"dialect":  "inland north",
"dialect subregions":  "seattle, portland, salem,
eugene, boise, idaho, montana, dakota, bismarck,
sioux falls, sheridan, minneapolis, madison,
milwaukee, chicago, detroit, columbus, albany,
buffalo, nyc, newark",
"word":  "crimanetly"
}
```


github.com/afshinrahimi/acl2017/

crimanetly is used to express surprise as in:

Oh crimanetly, it's Robin Hood.



Micro-averaged Recall@k results for retrieving dialect terms given the dialect region text representation.
(N.B. TWITTER-US dataset covers only 1k out of 4.3k DARE words)

Summary

- Combining text and social information using network inference improves geolocation performance; pure concatenation doesn't
  - more recent work where we jointly learn text and network embeddings, to greater effect [Rahimi et al., 2018]
- Flipped text-based geolocation models can be used in lexical dialectology ... with further room for improvement ...

... And Beyond! — Regression [Rahimi et al., 2017b]

Geolocation as Regression: Redux

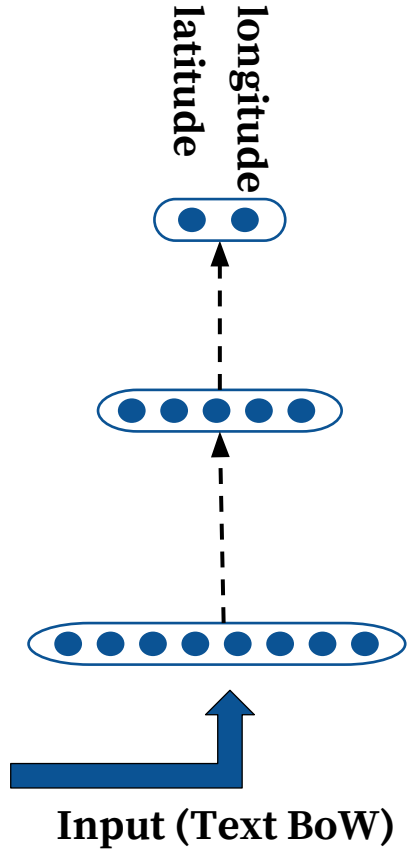
Concatenated User Tweets

Russell Peters @therealrussellp · May 31
Soooo... My daughter drew a picture of a "Scientist Potion" 👩‍🔬👩‍🔬👩‍🔬😂😂😂😂
… instagram.com/p/BUx9Q9eAP8B/
31 24 209

Russell Peters @therealrussellp · May 28
I'm trying to lose weight and get results but this might be too much change!…
instagram.com/p/BUo-I_hgFnt/
19 14 122

Russell Peters @therealrussellp · 30 Aug 2016
Just introduced my friend and actual musical genius @nilerodgers in Calgary!
instagram.com/p/BJwWsGfgV92/
6 6 76

Russell Peters @therealrussellp · 13 Jun 2016
Why does my coffee look scared?! #McCafe #DontWorryIllBeGentle
instagram.com/p/BGnQuZWoNDG/
9 10 59



Geolocation as Regression: Redux

- Single coordinate makes (some) sense for large documents/combinations of tweets
- But what about more fine-grained inputs?
  - *yinz*

Geolocation as Regression: Redux

- Single coordinate makes (some) sense for large documents/combinations of tweets
- But what about more fine-grained inputs?
  - *yinz*
  - *Over yonder*

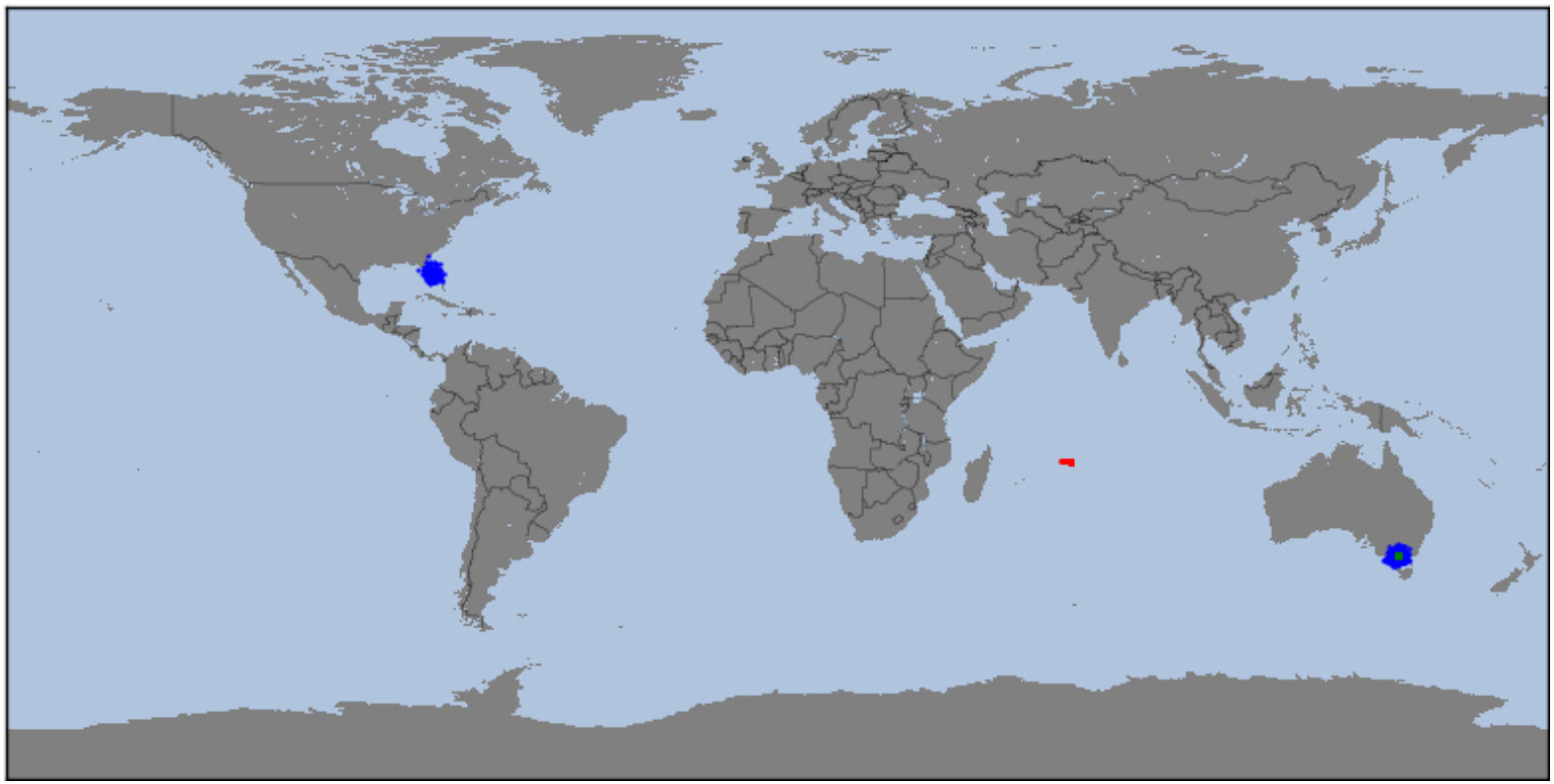
Geolocation as Regression: Redux

- Single coordinate makes (some) sense for large documents/combinations of tweets
- But what about more fine-grained inputs?
  - *yinz*
  - *Over yonder*
  - *Springfield*


DOH!


Washington
Montana
North Dakota
Minnesota
Ottawa
Montreal
Toronto
Michigan
South Dakota
Idaho
Wyoming
Nebraska
Iowa
Illinois
Indiana
Ohio
Nevada
Utah
United States
Kansas
Missouri
Virginia
New Jersey
Oklahoma
North Carolina
Maryland
Los Angeles
Arizona
New Mexico
Phoenix
San Diego
Dallas
Mississippi
Texas
San Antonio
Houston
Florida
Gulf of California
Monterrey
Gulf of Mexico

Beware the *Melbourne* Problem ...



Least Squared Regression

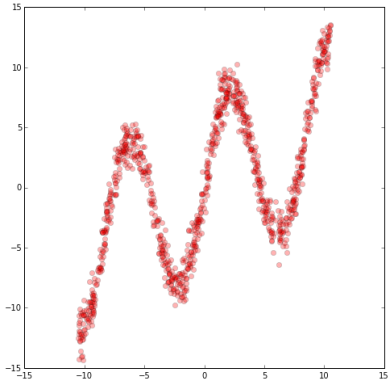
- The issue:

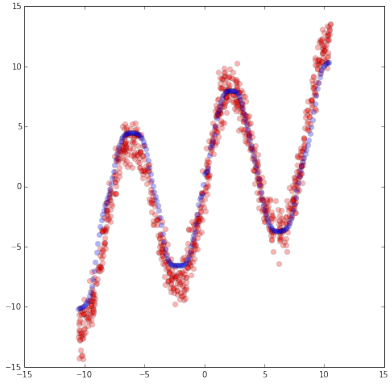
$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

... errors assumed to be normally distributed, but the data is multimodal

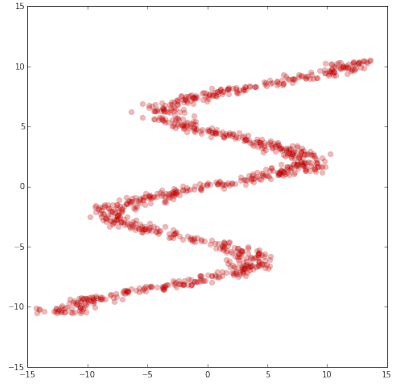
Worse Still: The Inverse Problem

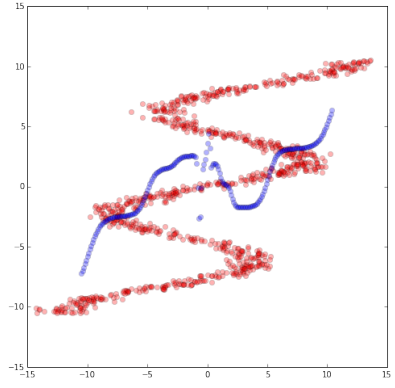
Where regression works:





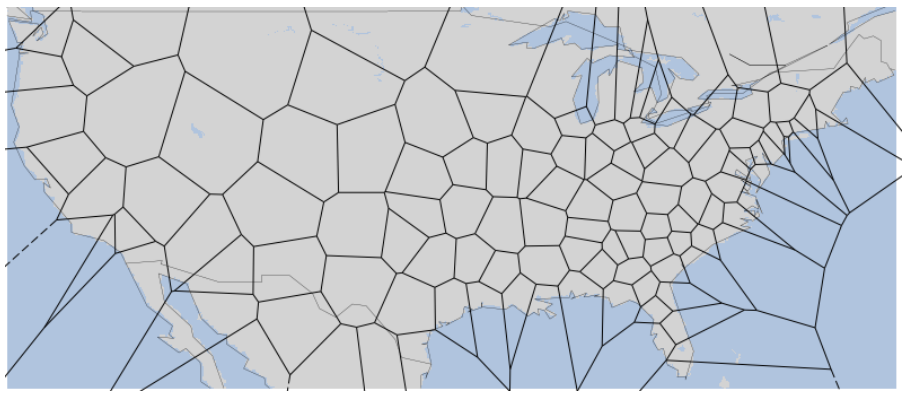
Where it doesn't:





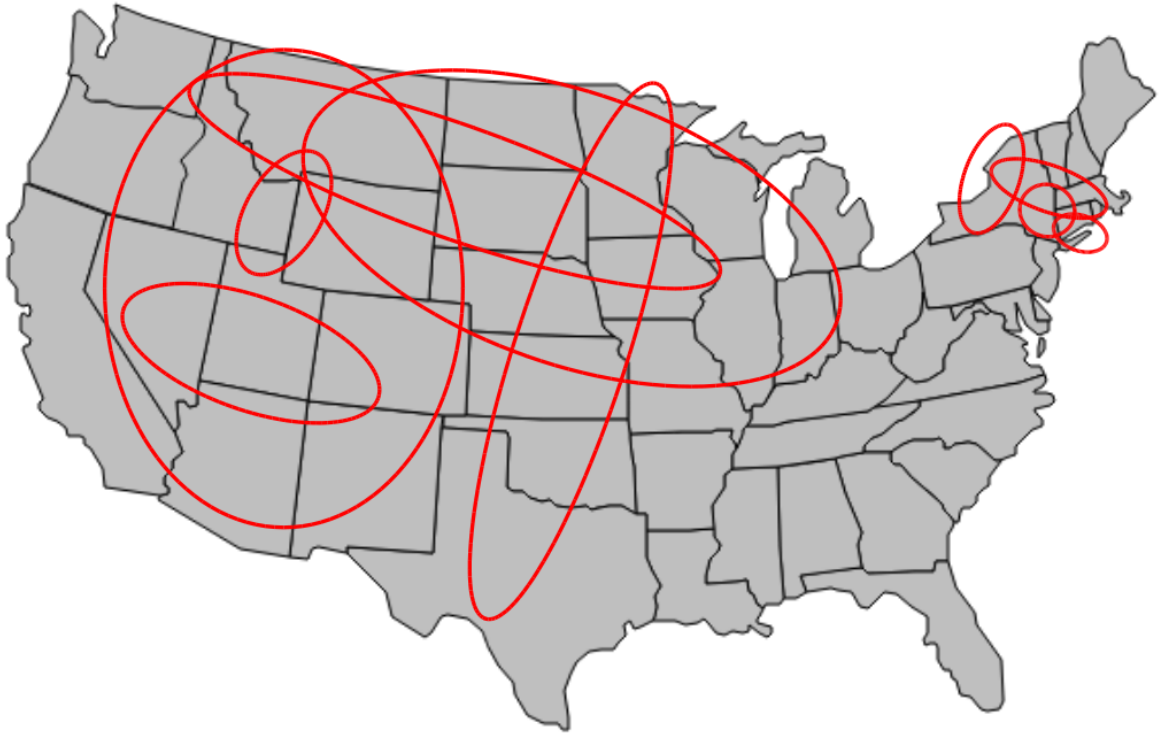
So we are Stuck with Geolocation Classification?

- The issue with classification:
  - over arbitrary inputs, the label set is often not predefined, giving rise to unsupervised discretisation methods, such as k-means, which encode arbitrary assumptions about the data:



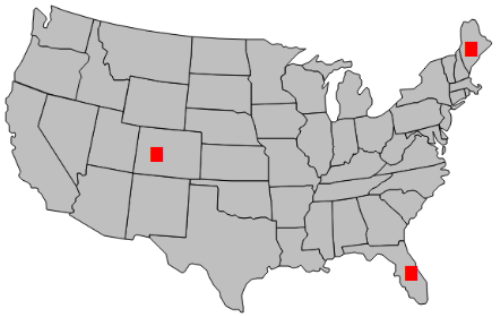
  - fine if output can be regions, but what about points?
  - and what if there are shifts in the underlying distribution over time?

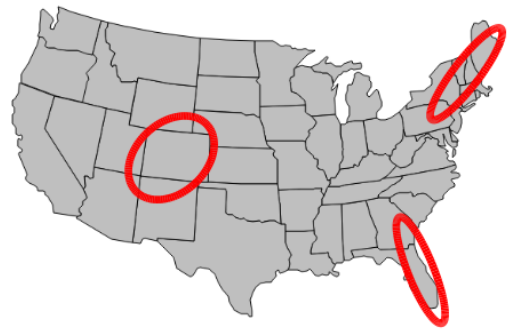
Representing Points as Gaussian Mixtures



$$\mathcal{P}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \Sigma_k)$$

2D Gaussians





$\boldsymbol{\mu} = \{\mu_x, \mu_y\}$ $\boldsymbol{\sigma} = \{\sigma_x.\sigma_y\}$ ρ

$$\mathcal{P}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \Sigma_k) \text{ where } \Sigma_k = \{\boldsymbol{\sigma}_k, \rho_k\}$$

Mixture Density Models ("MDN")

- Mixture density models ("MDNs": Bishop [1994]) offer a way of training a 2D Gaussian mixture model to predict the geolocation y of a given text input x:

$$\mathcal{P}(y|x) = \sum_{k=1}^{K} \pi_k(x)\mathcal{N}\big(y|\mu_k(x), \Sigma_k(x)\big)$$

- The parameters of the K Gaussians ($\times 6$ each) can be learned through a neural network, where the outputs are the

Mixture Density Models ("MDN")

Gaussian parameters, with constraints to ensure the parameters are well-defined:

$$\sigma \sim \text{SoftPlus}(\sigma') = \log(\exp(\sigma') + 1) \in (0, +\infty)$$
$$\pi \sim \text{SoftMax}(\pi')$$
$$\rho \sim \text{SoftSign}(\rho') = \frac{\rho'}{1 + |\rho'|} \in [-1, 1]$$

Mixture Density Models ("MDN")

- Training via negative log likelihood loss of each sample x given a 2d coordinate label y over:

$$\mathcal{L}(y|x) = -\log\left\{\sum_{k=1}^{K}\pi_k(x)\mathcal{N}\left(y|\mu_k(x),\Sigma_k(x)\right)\right\}$$

- Inference via:

$$\sum_{k=1}^{K}\pi_k\mathcal{N}(\mu_i|\mu_k,\Sigma_k)$$

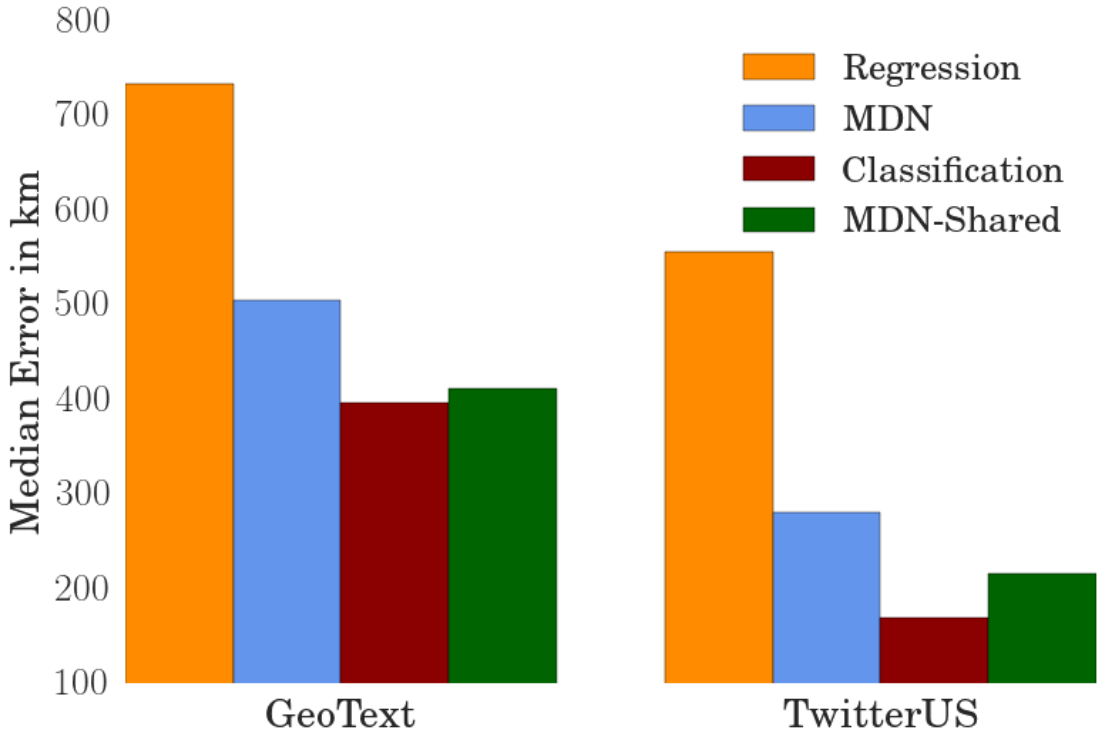
MDN with Shared Parameters (MDN-SHARED)

- Given the difficulty in training the model (esp. for each Σ_k), we also propose a variant where μ and Σ are shared among all samples as parameters of the output layer, and only π is predicted
- Initialise μ via K-means clustering; initialise Σ randomly between 0 and 10
- Training using the same cost function as the standard MDN, and also perform inference the same way

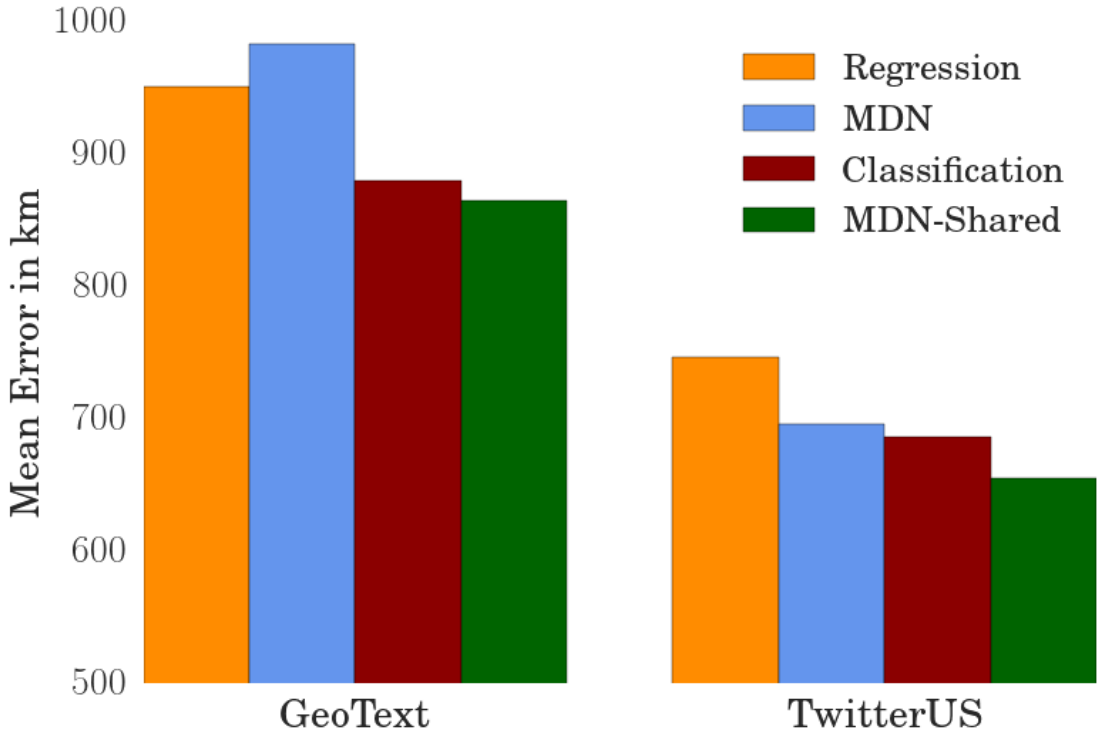
Geolocation Experiments

- Evaluation over GeoText and Twitter-US
- Compare `MDN` and `MDN-SHARED` with:
  - simple 2D linear output layer MLP with tanh hidden layer ("`Regression`")
  - state-of-the-art classification model for the respective datasets [Rahimi et al., 2015]
- Train neural network models with Adam optimiser, using early stopping and drop-out over the hidden layer and elastic net regularisation (with equal l_1 and l_2 shares)

Geolocation Results: Median Error (km)



Geolocation Results: Mean Error (km)



Lexical Dialectology

- Returning to the **dialectal term prediction task** based on DAREDS, we model the task by taking a lat–long pair as input, and outputting a (unigram) probability distribution over the vocabulary
- Motivated by MDNs, we adopt an architecture based on an RBF layer:
  - RBF neurons with μ and Σ parameters, which produce a probability as their activation function
  - additional tanh layer, and final SoftMax layer over the vocabulary

Visualisation of *hella*

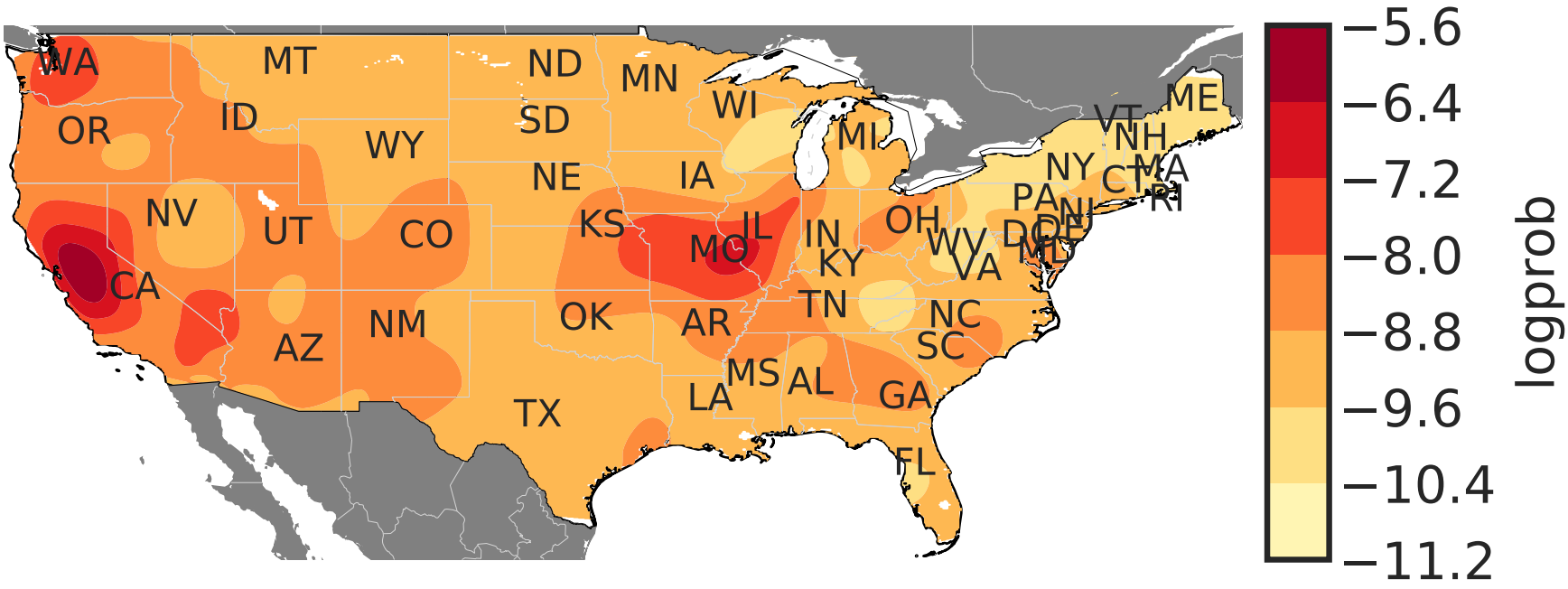


Figure: Log probabilities of *hella* (an intensifier mostly used in Northern CA, also the name of a company in IL) in continental U.S.

Visualisation of *yall*

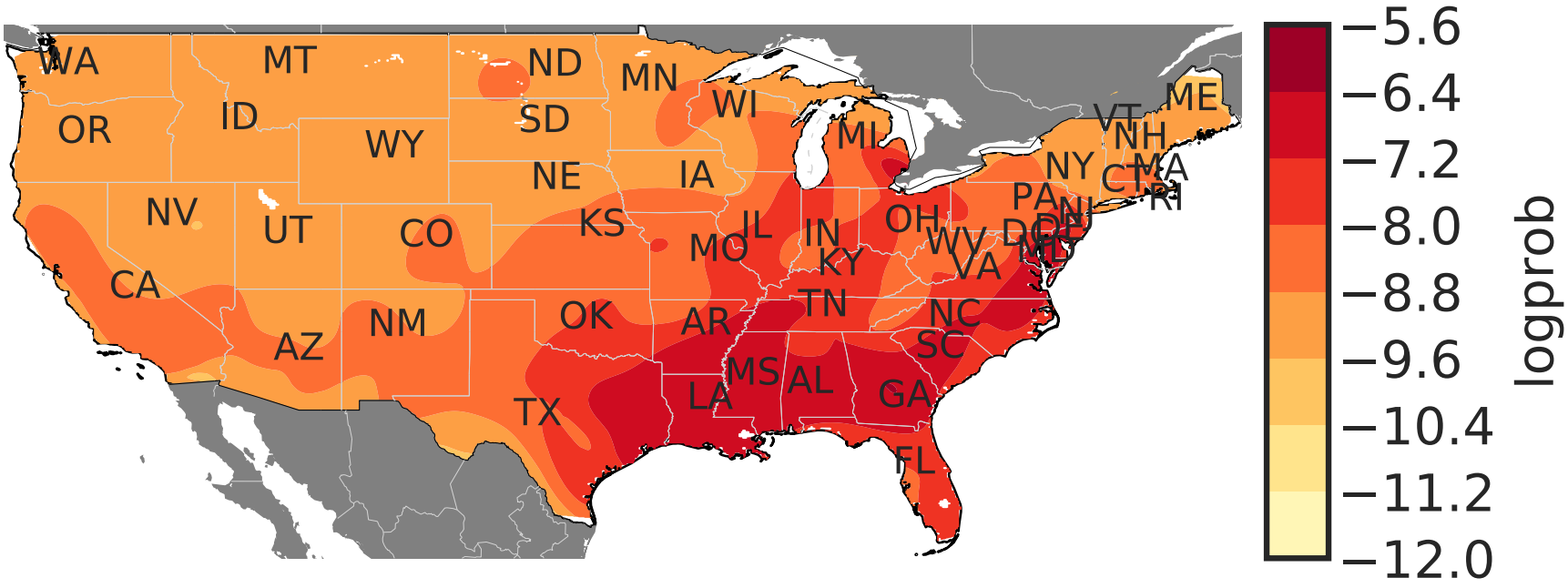


Figure: Log probabilities of *yall* (meaning "you all", used broadly in Southern U.S.) in continental U.S.

Visualisation of *yinz*

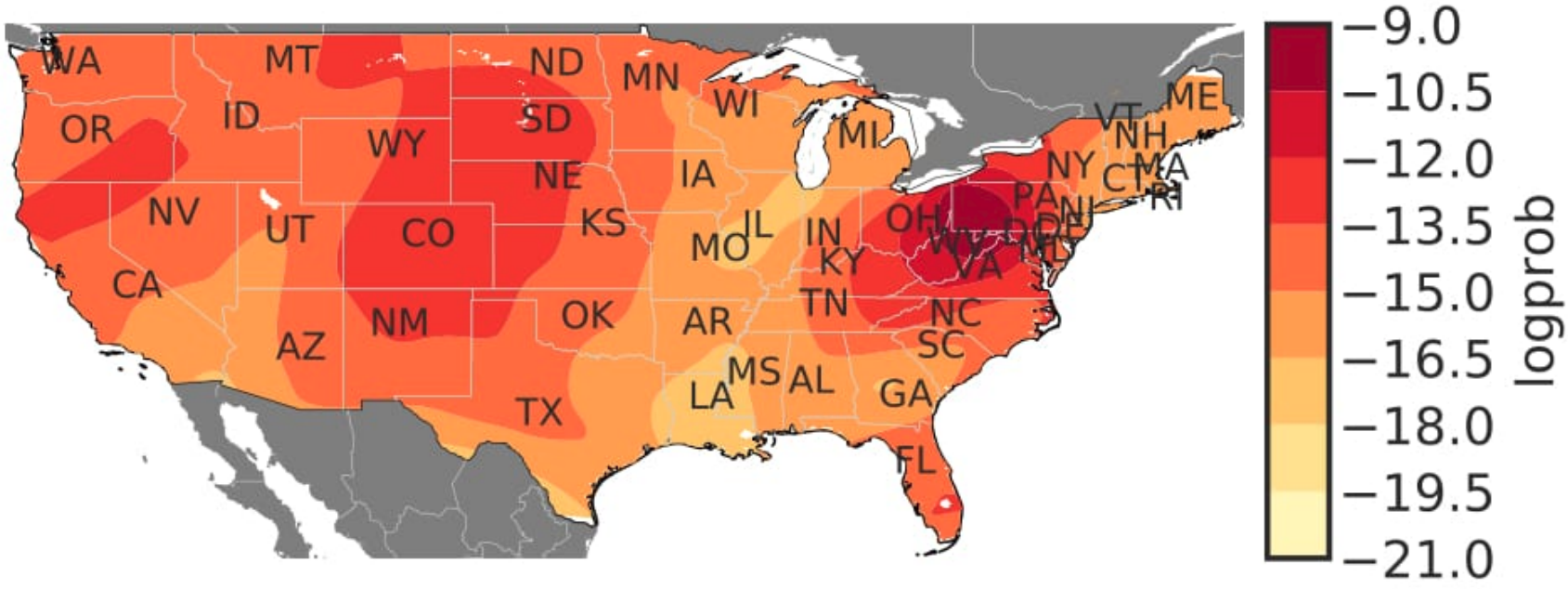


Figure: Log probabilities of *yinz* (meaning "you all", used in Western PA) in continental U.S.

Summary

- MDNs are an effective way of modelling the geolocation task
- If your multimodal problem has global structure you can share the parameters in MDN
- RBF layers are a cool alternative to ReLU/tanh/sigmoid, with fun applications in lexical dialectology

Implications for Analysis of Language Varieties/Dialects

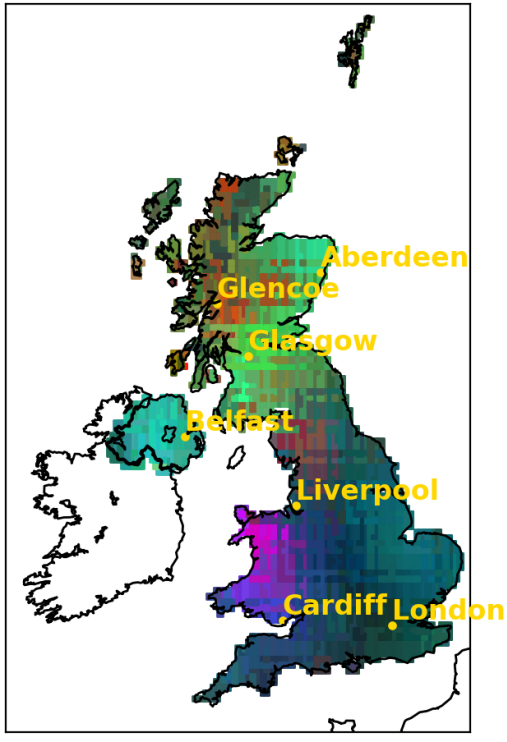
- Geolocation classification models are directly applicable to VarDial tasks (with or without interaction features) ... e.g. when applied to a set of around 8k training Twitter users split between Germany, Austria and Switzerland using the model of Rahimi et al. [2018], absolute improvement of 2–3% when (sparse!) interaction network added (based on 10 tweets only), and 8–9% for denser interaction network (based on 100 tweets)
- MDNs provide valuable geospatial visualisation tool for understanding the geographic bias of terms
- As we are dynamically learning the component Gaussians as part of the training of the modelling (unlike k-means etc.), possibility of adding a time dimension to dynamically model temporal shifts in the data

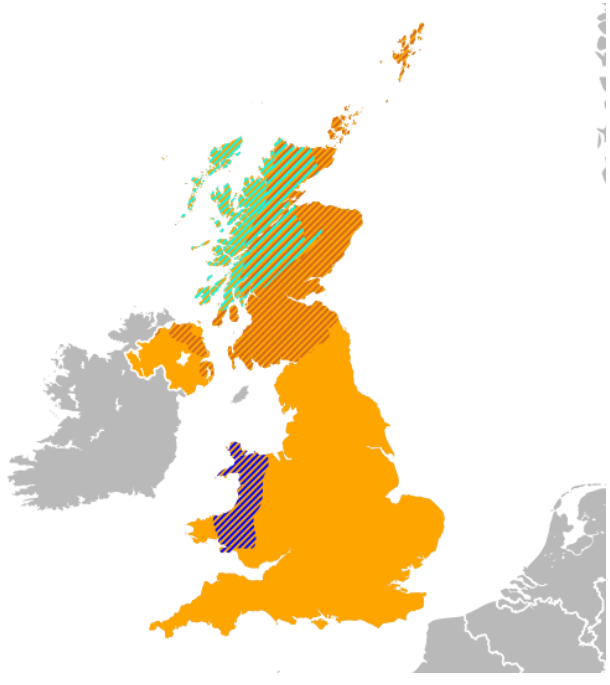
Talk Outline

Dialect Map Generation

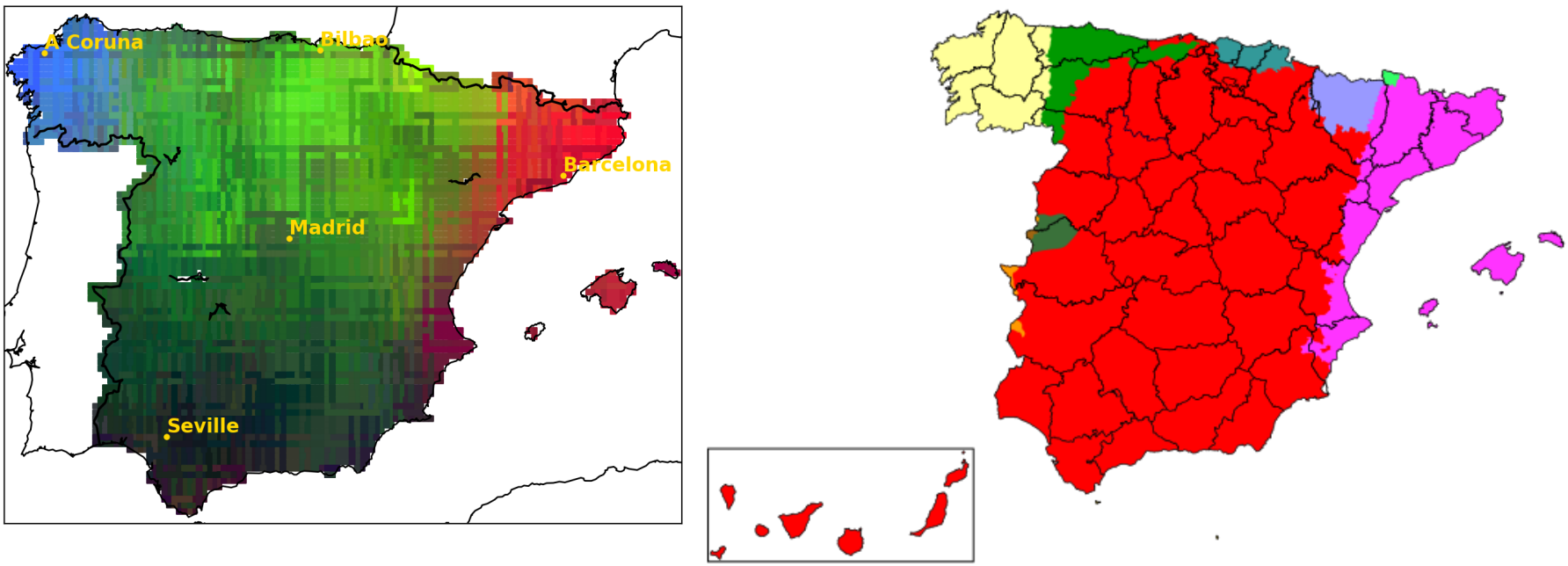
- As a final teaser, we (in collaboration with Dirk Hovy; to appear) have been playing around with "extreme dialectometry", in taking a large crawl of Twitter data from Europe and generating dialectal maps, by:
  - partitioning the map into uniform-sized grid cells
  - removing any English tweets with `langid.py`
  - training doc2vec [Le and Mikolov, 2014, Lau and Baldwin, 2016] over the concatenated tweets in each cell (document = cell id), with smoothing over local neighbourhoods
  - reducing the vector representations down to 3D using PCA, and interpreting the resulting vectors as RGB values

Dialect Map of UK (doc2vec vs. Wikipedia)

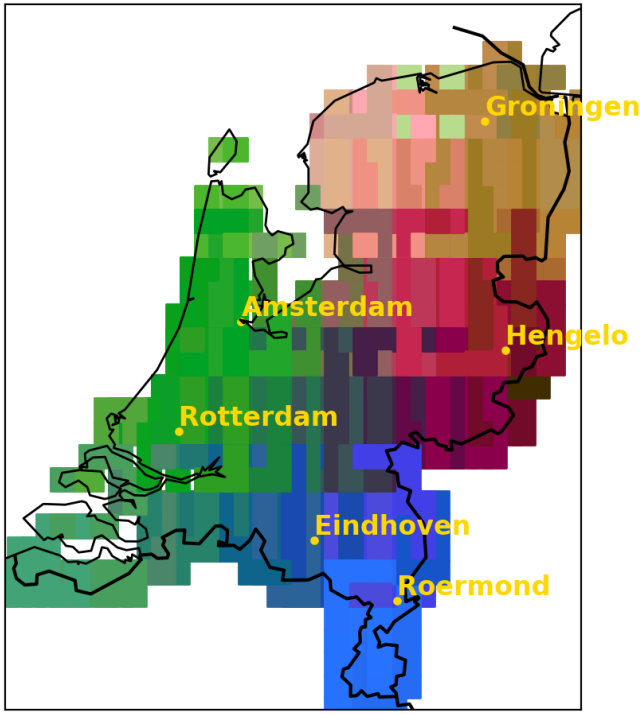


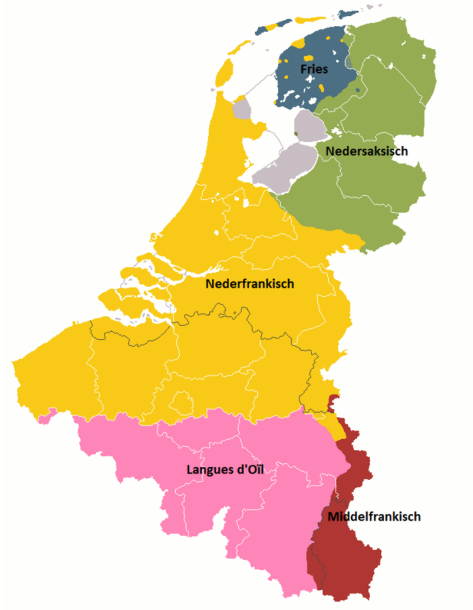


Dialect Map of Spain (doc2vec vs. Wikipedia)



Dialect Map of Netherlands (doc2vec vs. Wikipedia)





Talk Outline

Overall Summary

- VarDial is a fantastic effort which has galvanised the NLP community to work on dialects and language variants, but possibilities for cross-fertilisation with many other fields, in terms of modelling the social and spatio-temporal aspects of dialect change/interaction, and mitigating the structural biases
- Strong encouragement to the community to keep doing what it's doing, in addition to looking out for opportunities to bridge with other communities, and in particular to add to the collective understanding of the nature and formation/evolution of dialects/language varieties

Acknowledgements

- Joint work with Trevor Cohn, Dirk Hovy, Yitong Li, Afshin Rahimi
- This work was supported by the Australian Research Council and Xerox Research

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160, 2014.

Christopher Bishop. Mixture density networks. Technical report, Aston University, 1994. URL `https://www.microsoft.com/en-us/research/publication/mixture-density-networks/`.

Frederic Gomes Cassidy. *Dictionary of American Regional English*. Belknap Press of Harvard University Press, 1985.

J. K. Chambers. Dialect acquisition. *Language*, 68(4):673–705, 1992.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. Novel word-sense identification. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1624–1635, Dublin, Ireland, 2014.

Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, 2007.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1277–1287, Cambridge, USA, 2010. URL `http://www.aclweb.org/anthology/D10-1124`.

References

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 1365–1374, Portland, USA, 2011. URL `http://www.aclweb.org/anthology/P11-1137`.

Lea Frermann and Mirella Lapata. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45, 2016.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:59:1–59:35, 2016.

Hans Goebl. On the geolinguistic change in Northern France between 1300 and 1900: A dialectometrical inquiry. In *Computing and Historical Phonology: Proceedings of the Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 75–83, Prague, Czech Republic, 2007.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

References

Jack Grieve, Dirk Speelman, and Dirk Geeraerts. A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23:1–29, 2011.

Bo Han, Paul Cook, and Timothy Baldwin. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, 2014.

William Labov. *The social stratification of English in New York City*. Center for Applied Linguistics, Washington D.C., USA, 1966.

Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 78–86, Berlin, Germany, 2016.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*, pages 1188–1196, Beijing, China, 2014.

References

Yitong Li, Trevor Cohn, and Timothy Baldwin. What's in a domain? learning domain-robust text representations using adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies (NAACL HLT 2018)*, pages 474–479, New Orleans, USA, 2018.

Marco Lui and Timothy Baldwin. Cross-domain feature selection for language identification. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 553–561, Chiang Mai, Thailand, 2011.

Marco Lui and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea, 2012.

John Nerbonne. Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing*, 21:463–476, 2006.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593, 2016.

H. Orton, M. V. Barry, and W. J. Halliday. *Survey of English dialects. Vol. I: the six northern counties and the Isle of Man*. Arnold, Leeds, UK, 1962.

References

Arvilla C. Payne. Factors controlling the acquisition of the Philadelphia dialect by out-of-state children. In William Labov, editor, *Locating language in time and space*, pages 143–178. Academic Press, New York, USA, 1980.

Afshin Rahimi, Duy Vu, Trevor Cohn, and Timothy Baldwin. Exploiting text and network context for geolocation of social media users. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics — Human Language Technologies (NAACL HLT 2015)*, Denver, USA, 2015.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. A neural model for user geolocation and lexical dialectology. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 207–216, Vancouver, Canada, 2017a.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 167–176, Copenhagen, Denmark, 2017b.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. Semi-supervised user geolocation via graph convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 2009–2019, Melbourne, Australia, 2018.

References

David Reed and John L. Spicer. Correlation methods of comparing dialects in a transition area. *Language*, 28:348–359, 1952.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 1500–1510, Jeju Island, Korea, 2012. URL `http://www.aclweb.org/anthology/D12-1137`.

Alex Rosenfeld and Katrin Erk. Deep neural models of semantic shift. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 474–484, New Orleans, USA, 2018.

Takeshi Sibata. *Nihon no Hōgen [The dialects of Japan]*. Iwanami Shoten, Tokyo, Japan, 1958.

Partha Pratim Talukdar and Koby Crammer. New regularized algorithms for transductive learning. In *Proceedings of the European Conference on Machine Learning (ECML-PKDD 2009)*, pages 442–457, Bled, Slovenia, 2009.

Peter Trudgill. Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Language in Society*, 2:215–246, 1974.

References

Peter Trudgill. *Dialects in contact*. Blackwell, Oxford, UK, 1986.

J. C. Wells. *Jamaican pronunciation in London*, volume 25 of *Publications of the Philological Society*. Blackwell, Oxford, UK, 1973.