# Robust, Unbiased Natural Language Processing

Not Trevor Cohn ... but Timothy Baldwin

(joint work with Yitong Li)

THE UNIVERSITY OF MELBOURNE

# Talk Outline

1. Introduction
2. Robustness through Data Augmentation
3. Robustness through Cross-domain Debiasing
4. Robustness and Privacy through Author-demographic Debiasing
5. Summary

# Background

- NLP systems are notoriously domain-brittle, and generally rely on explicit transfer learning or (re-)training in target domain
    - off-the-shelf CoreNLP NER = 0.04 F-score at recognising geospatial NEs in highly localised data [Liu et al., 2014]; in case of Twitter data, F-score = 0.44 [Ritter et al., 2011]

# Background

- NLP systems are notoriously domain-brittle, and generally rely on explicit transfer learning or (re-)training in target domain
  - off-the-shelf CoreNLP NER = 0.04 F-score at recognising geospatial NEs in highly localised data [Liu et al., 2014]; in case of Twitter data, F-score = 0.44 [Ritter et al., 2011]
- Growing awareness of bias in our trained NLP models, often accentuated wrt the bias in our training datasets [Zhao et al., 2017]

# Background

- NLP systems are notoriously domain-brittle, and generally rely on explicit transfer learning or (re-)training in target domain
    - off-the-shelf CoreNLP NER = 0.04 F-score at recognising geospatial NEs in highly localised data [Liu et al., 2014]; in case of Twitter data, F-score = 0.44 [Ritter et al., 2011]
- Growing awareness of bias in our trained NLP models, often accentuated wrt the bias in our training datasets [Zhao et al., 2017]
- **Aim:** develop methods for training models that: (a) are robust to domain shift without sacrificing in-domain accuracy; and (b) generalise away from any explicit demographic biases in our training data

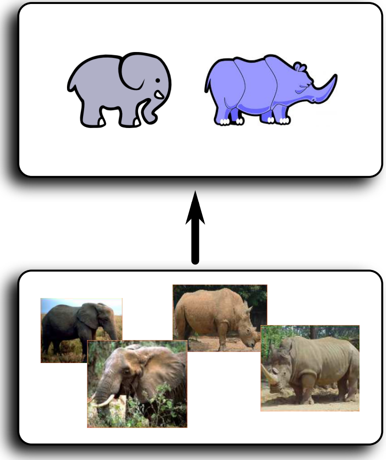
# Outline

- Three approaches to robustness, two of which are based on explicit debiasing:
  1. robustness through linguistically-motivated data augmentation [Li et al., 2017]
  2. robustness through cross-domain debiasing [Li et al., 2018b]
  3. robustness and privacy through author-demographic debiasing [Li et al., 2018a]
- In all cases, assume no access to target domain at training time
- Primary focus on document categorisation, but also some results for structured classification (and methods designed to generalise to other tasks)

# Talk Outline

1. Introduction
2. Robustness through Data Augmentation
3. Robustness through Cross-domain Debiasing
4. Robustness and Privacy through Author-demographic Debiasing
5. Summary

# Data Setting 1: Single Source Domain



Domain$_2$

Domain$_1$

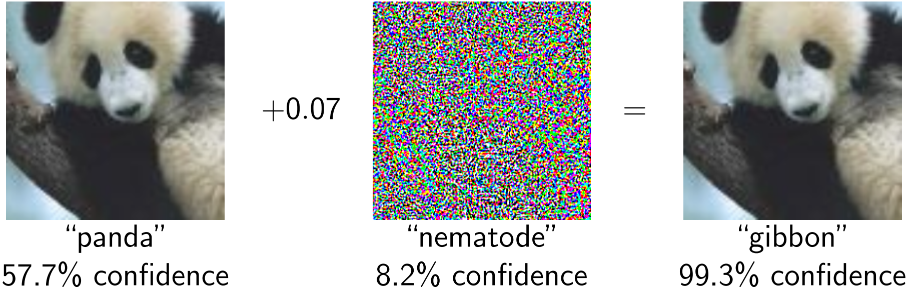
# Background

- Deep learning has achieved state-of-the-art results across many tasks, however, the resulting models are notoriously susceptible to overfitting, and suffer from a lack of generalisation and robustness
- Methods of training robust NNs:
  - variational approaches
  - model regularization
  - data augmentation

    e.g. adding noise to the layers: Gaussian Noise, dropout

# Adversarial Examples

Our approach is inspired by adversarial examples:



**Source(s):** Szegedy et al. [2014]

# Can We Generate "Adversarial" Noise over Text?

- Text is not continuous
- Embeddings are not a true (or human-intuitive/sufficiently expressive/...) representation of human language

# Can We Generate “Adversarial” Noise over Text?

- Text is not continuous
- Embeddings are not a true (or human-intuitive/sufficiently expressive/...) representation of human language
- **Idea:** possible to linguistically perturb training instances (while preserving felicity of labelling), to generate extra training data with greater variation?

# Generating Text Noise

Syntactic Noise: making syntactic changes

- **paraphrasing**: English Resource Grammar (“ERG”: Copestake and Flickinger [2000])
- **sentence compression** (“COMP”: Knight and Marcu [2000])

Semantic Noise: substitute near-synonyms of words

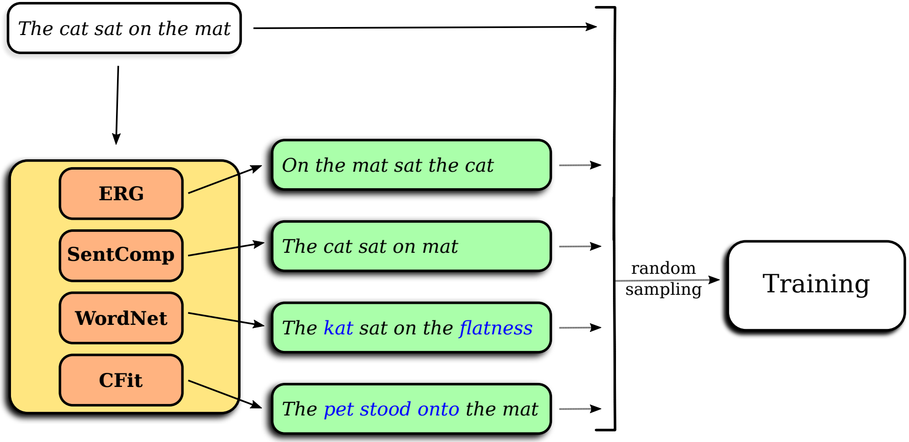
- two synonym resources:
  1. **WordNet** (“WN”: Miller et al. [1990])
  2. **“counter-fitted” word embeddings** (“CFIT”: Mrkšić et al. [2016])
- Use a language model to ensure the output is plausible/fluent in each case

# Noised Text Examples

| Method | Example |
|---|---|
| Original | The cat sat on the mat . |
| ERG | On the mat sat the cat . |
| Comp | The cat sat on ◇ mat ◇ |
| WN | The kat sat on the flatness . |
| CFit | The pet stood onto the mat . |

Table: Examples of generated sentences across four proposed methods. Modified words are marked by underwave, and elided words are denoted with a "◇".

# Model Training



# Evaluation Objectives

- Test the “noising” approach under two scenarios:
  - **Generalisation:** application to standard in-domain testing scenario; does it work like an implicit regularizer?
  - **Robustness:** application to very different testing data, e.g., cross-domain, can it handle domain-shifted inputs?

# Experimental Settings

**Task:** sentence-level classification

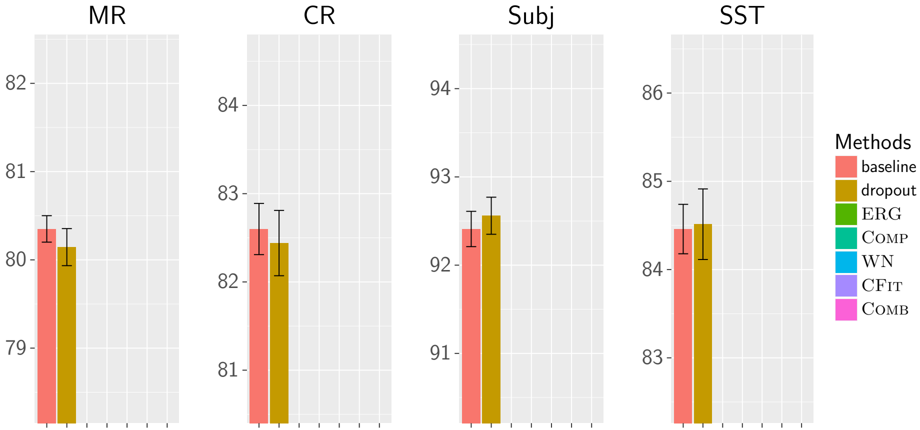
**Model:** convolutional neural network [Kim, 2014]

**Datasets:**

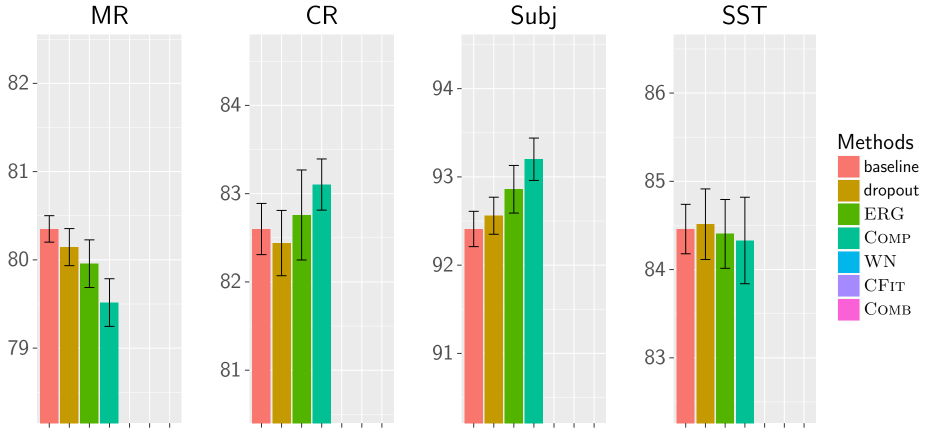
- MR: movie review sentence polarity dataset [Pang and Lee, 2008]
- CR: customer review dataset [Hu and Liu, 2004]
- Subj: subjectivity dataset [Pang and Lee, 2005]
- SST: Stanford Sentiment Treebank, using the 2-class configuration [Socher et al., 2013]

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

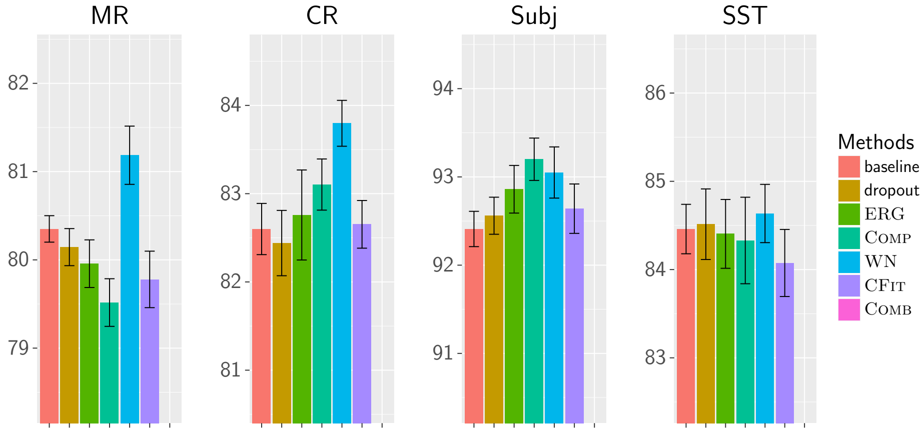
# In-domain Accuracy[%]



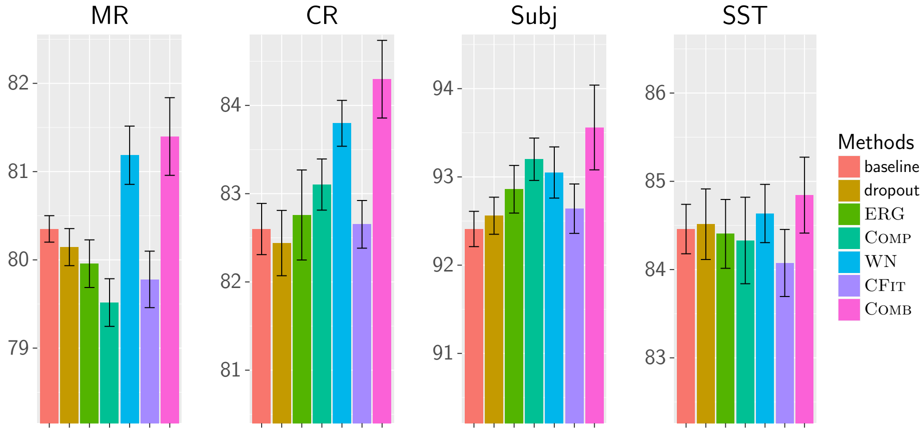
# In-domain Accuracy[%]



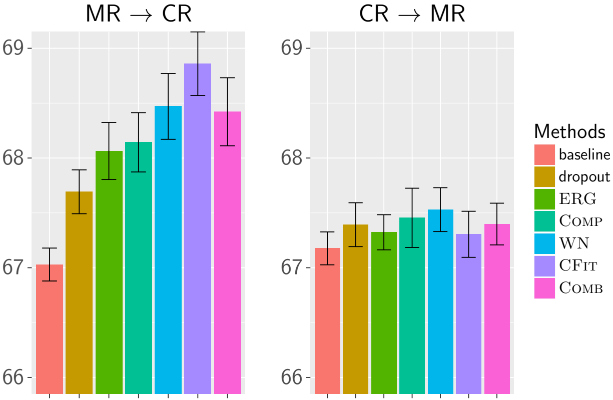
# In-domain Accuracy[%]



# In-domain Accuracy[%]



# Cross-domain Accuracy[%]



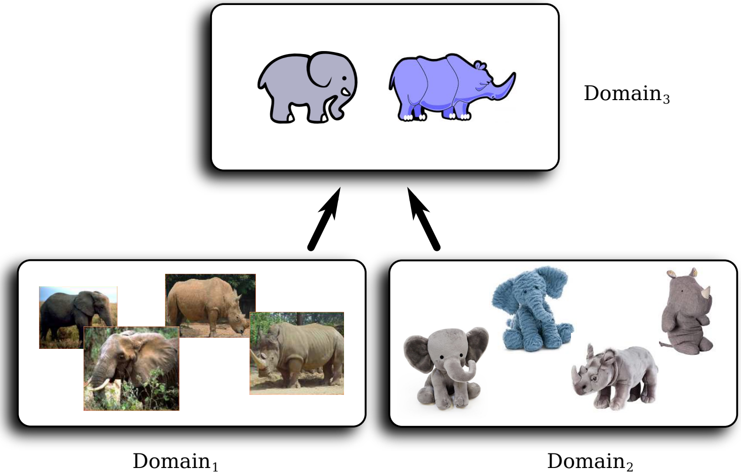
# Summary

- Linguistically-motivated method for training robust models, based on explicit linguistic “noising” through data augmentation
- Method outperforms standard training and dropout, and is generalisable to other NLP models/tasks

# Talk Outline

1. Introduction
2. Robustness through Data Augmentation
3. Robustness through Cross-domain Debiasing
4. Robustness and Privacy through Author-demographic Debiasing
5. Summary

# Data Setting 2: Multiple Source Domains



Domain$_3$

Domain$_1$

Domain$_2$

# 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

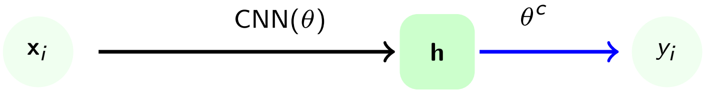
# Approach

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

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

# Approach 1: Baseline

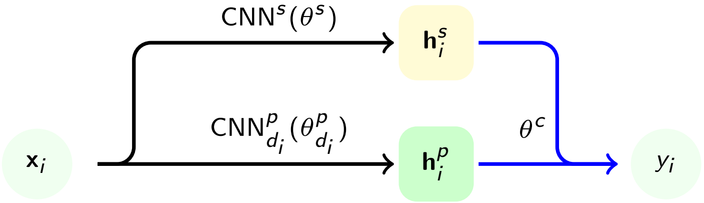
- Baseline model = straight CNN [Kim, 2014]



# Approach 2: Domain-conditional Model (“COND”)

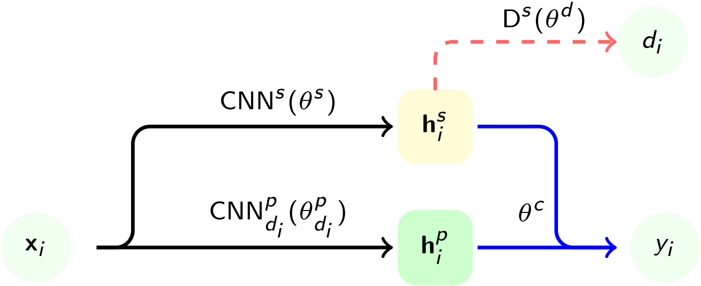
- Basic intuition: take inspiration from Daumé III [2007] in learning two representations of each instance $\mathbf{x}$:
  1. shared representation $\mathbf{h}_i^s$, using a shared CNN$^s$
  2. private representation $\mathbf{h}_i^p$ conditioned on domain identifier $d_i$ of $\mathbf{x}$

  and concatenate the two to generate overall document representation



# Approach 2: 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:



# Approach 2: Domain-conditional Model (“Cond”)

- Overall training objective:

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

  where:

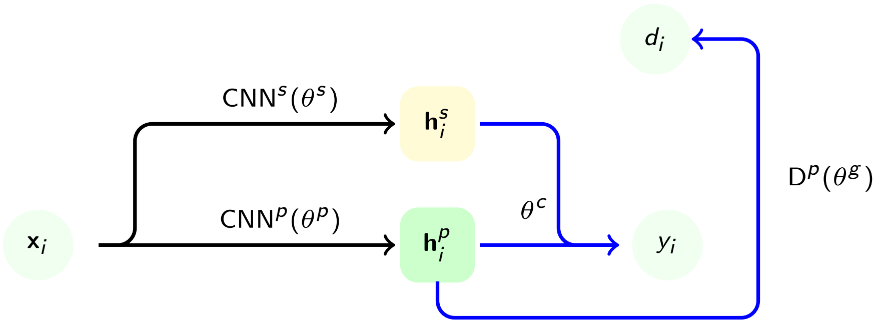
  - $\mathbf{H}^s = \{\mathbf{h}_i^s(\mathbf{x}_i)\}_{i=1}^n$ = the shared representations for all instances
  - $\mathbf{H}^p = \{\mathbf{h}_i^p(\mathbf{x}_i, d_i)\}_{i=1}^n$ = the private representations for all instances

- Train discriminator to be maximally accurate wrt $\theta^d$, 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

# Approach 3: Domain-generative Model ("Gen")

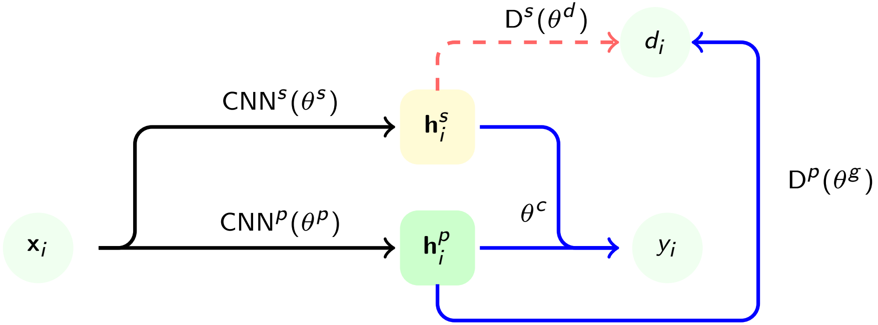
- Basic intuition: largely the same as Approach 2, but *generate* the domain (based on multi-task learning) rather than conditioning on it, by:
  1. computing $\mathbf{h}^p$ using a *single* CNN$^p$ rather than several domain-specific CNNs
  2. using the private representation to predict the domain,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

# Approach 3: Domain-generative Model (“Gen”)



# Approach 3: Domain-generative Model ("Gen")

- Similarly to Cond, optionally add an adversarial discriminator:



# Approach 3: Domain-generative Model (“GEN”)

- Overall training objective:

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

where:

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

# Experiment 1: Language Identification

Task: document-level language identification

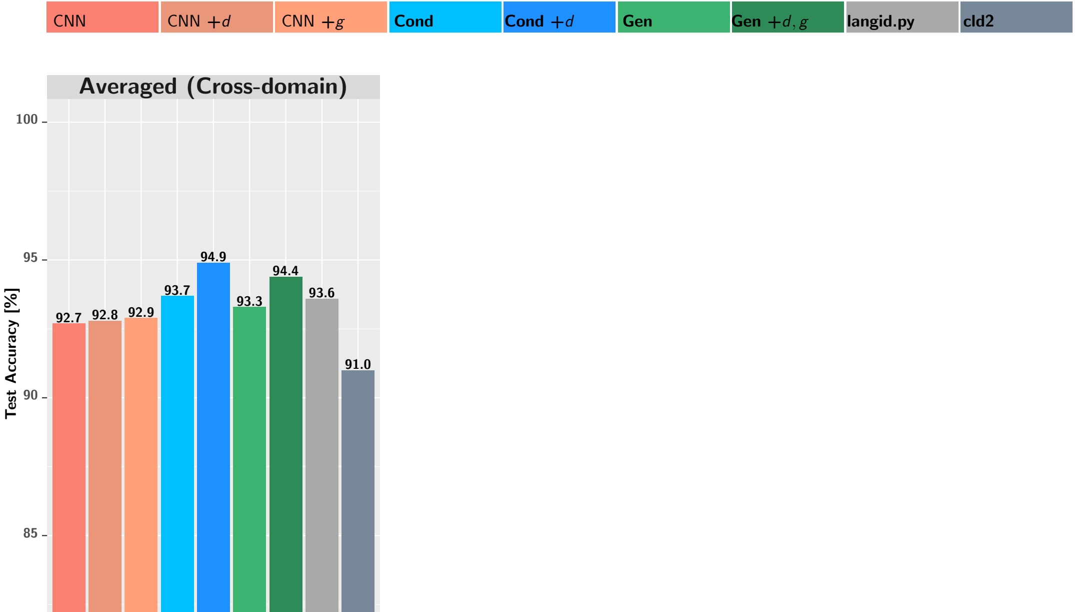
Target: 97 languages

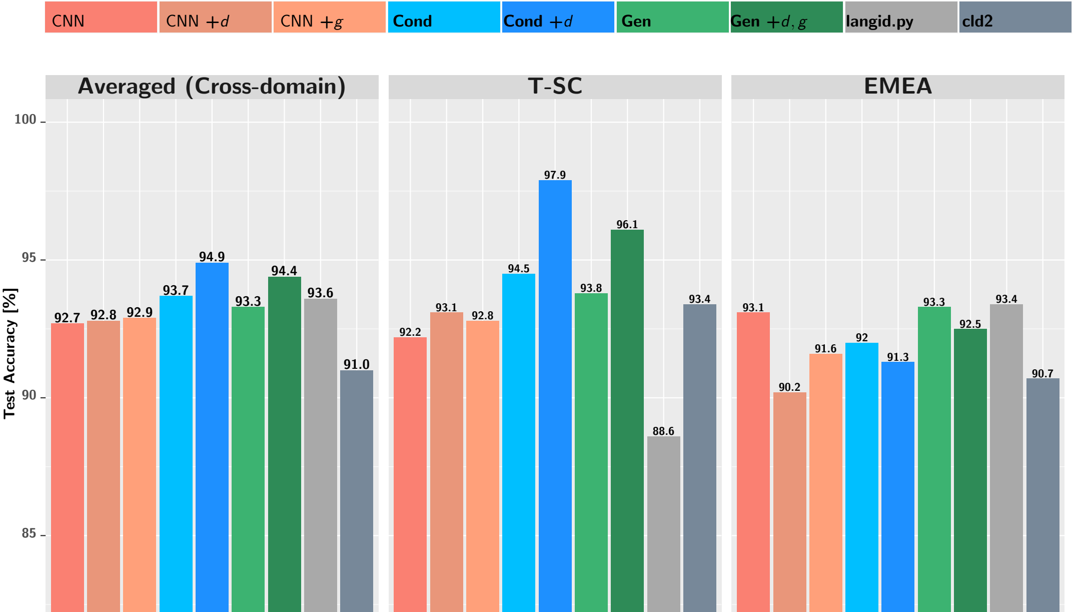
Model: byte-level CNN (up to 1k bytes)

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

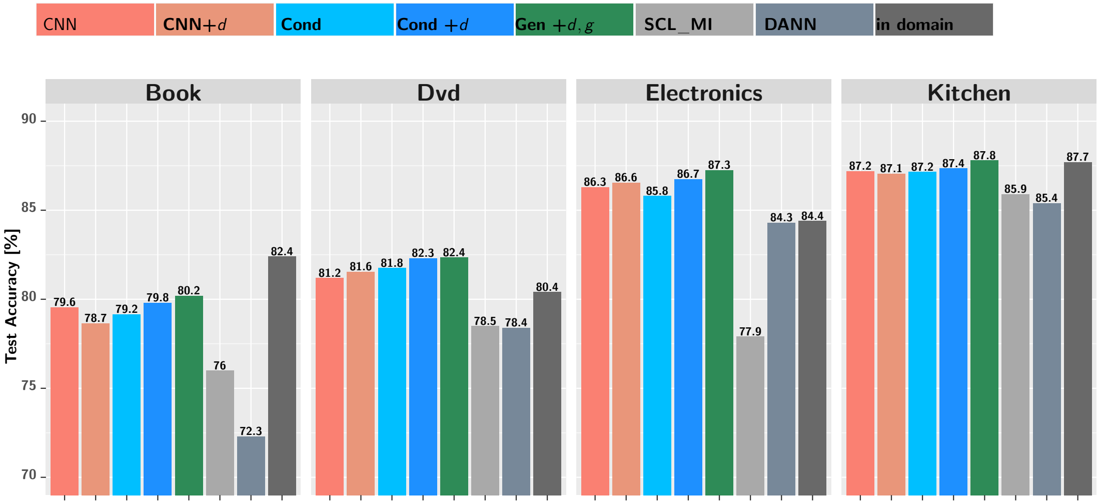
# Experiment 2: Sentiment Classification

Task: document-level sentiment classification (pos vs. neg)

Model: word-level CNN

Dataset: Multi-Domain Sentiment Dataset [Blitzer et al., 2007]:

- **16** training domains
- **4** heldout test domains


CNN
CNN+d
Cond
Cond +d
Gen +d, g
SCL_MI
DANN
in domain
Test Accuracy [%]
90
85
80
75
70
Book
79.6
78.7
79.2
79.8
80.2
76
72.3
82.4
Dvd
81.2
81.6
81.8
82.3
82.4
78.5
78.4
80.4
Electronics
86.3
86.6
85.8
86.7
87.3
77.9
84.3
84.4
Kitchen
87.2
87.1
87.2
87.4
87.8
85.9
85.4
87.7

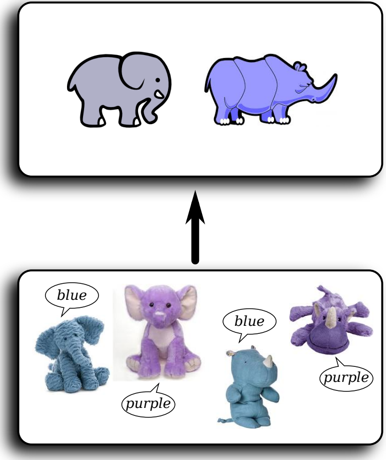
# Summary

- 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

1. Introduction
2. Robustness through Data Augmentation
3. Robustness through Cross-domain Debiasing
4. Robustness and Privacy through Author-demographic Debiasing
5. Summary

# Data Setting 3: Single Source Domain with Side Information



Domain$_2$

Domain$_1$

# Introduction

- There is growing awareness of the fact that deep learning is particularly susceptible to dataset bias, esp. in terms of demographic bias underlying standard datasets (e.g. women cook; doctors are men; English language writers are white, middle-aged, US males)
- The demographic biases "baked in" to many of our datasets tend to be implicitly learned by our models, and often accentuated [Hovy, 2015, Rabinovich et al., 2017]
- Much work left to be done on training unbiased models without sacrificing aggregate accuracy [Zhao et al., 2017], but equally, the interface between domain-robustness and demographic bias is not well understood
- Additionally, if our models are learning biased representations, there are potential privacy implications, in terms of the ability to regenerate training data from biases latent in our models

# Research Focus

- If we have access to demographic variables associated with training instances, can we explicitly debias our models such that:
  - they do not reflect those biases at test time
  - aggregate in-domain performance is not hurt (or ideally improved!)
  - cross-domain performance is potentially enhanced

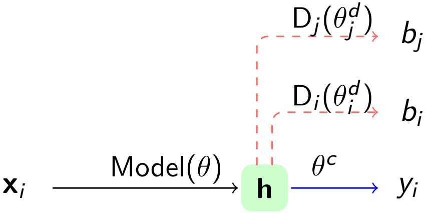
# Approach

- Similar to Li et al. [2018b], want to maximise target variable accuracy, while minimising accuracy over demographic variables, so adopt a similar approach with an adversarial discriminator per "private channel" (based on the individual demographic variables this time):

$$\hat{\theta} = \min_{\theta_M} \max_{\{\theta_{\mathrm{D}^i}\}_{i=1}^{N}} \mathcal{X}(\hat{\mathbf{y}}(\mathbf{x};\theta_M),\mathbf{y}) - \sum_{i=1}^{N} \left( \lambda_i \cdot \mathcal{X}(\hat{b}(\mathbf{x};\theta_{\mathrm{D}^i}), b_i) \right)$$

# Approach

- Architecture:



where $(\mathbf{x}_i, y_i)$ is a training instance with protected attributes $b_i$ and $b_j$, and D indicates a discriminator

# Experiment 1: POS Tagging

**Task:** POS tagging (based on Google Universal POS tagset)

**Model:** biLSTM; adversarial discriminator = single feed-forward layer applied to final hidden representation ($[\mathbf{h}_n; \mathbf{h}'_0]$)

**Datasets:**

- training domain = English Web Treebank for pre-training [Bies et al., 2012], and TrustPilot for fine-tuning [Hovy and Søgaard, 2015]
- test domains = TrustPilot + AAVE POS dataset [Jørgensen et al., 2016]

**Demographic variables:**

- age (under-35 vs. over-45)
- gender (male vs. female)

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

# Experiment 1: POS Tagging

- POS accuracy [%] over Trustpilot test set, stratified by SEX and AGE:

| | SEX | | | AGE | | |
|---|---|---|---|---|---|---|
| | F | M | $\Delta$ | O45 | U35 | $\Delta$ |
| BASELINE | 90.9 | 91.1 | 0.2 | 91.4 | 89.9 | 1.5 |
| ADV | **92.2** | **92.1** | **0.1** | **92.3** | **92.0** | **0.3** |

# Experiment 1: POS Tagging

- POS accuracy [%] over Trustpilot test set, stratified by SEX and AGE:

| | SEX | | | AGE | | |
|---|---|---|---|---|---|---|
| | F | M | Δ | O45 | U35 | Δ |
| BASELINE | 90.9 | 91.1 | 0.2 | 91.4 | 89.9 | 1.5 |
| ADV | **92.2** | **92.1** | **0.1** | **92.3** | **92.0** | **0.3** |

- POS accuracy [%] over AAVE dataset:

| | LYRICS | SUBTITLES | TWEETS | Average |
|---|---|---|---|---|
| BASELINE | 73.7 | 81.4 | 59.9 | 71.7 |
| ADV | **80.5** | **85.8** | **65.4** | **77.0** |

# Experiment 2: Sentiment Analysis

**Task:** (English) sentiment classification (5-way)

**Model:** CNN; adversarial discriminator = single feed-forward layer applied to final hidden representation

**Dataset:** TrustPilot (cross-validation, with dev partition)

**Demographic variables:**

- age (under-35 vs. over-45)
- gender (male vs. female)
- location (US, UK, Germany, Denmark, and France)

**Evaluation:** micro-averaged F-score

# Experiment 2: Sentiment Analysis

| | $F_1$ | | Discrimination [%] | | |
|---|---|---|---|---|---|
| | dev | test | AGE | SEX | LOC |
| Majority class | | | 57.8 | 62.3 | 20.0 |
| BASELINE | 41.9 | 40.1 | 65.3 | 66.9 | 53.4 |
| ADV-AGE | **42.7** | 40.1 | **61.1** | 65.6 | 41.0 |
| ADV-SEX | 42.4 | 39.9 | 61.8 | 62.9 | 42.7 |
| ADV-LOC | 42.0 | **40.2** | 62.2 | 66.8 | **22.1** |
| ADV-all | 42.0 | **40.2** | 61.8 | **62.5** | 28.1 |

# Findings

- Largely similar in-domain results, but considerably better balance across demographic variables
- Greatly improved cross-domain accuracy for POS tagging(!)
- Much greater preservation of privacy in hidden representations for test users

# Summary

- Adversarial learning method, as means of obfuscating demographic information of training users
- In-domain, we are able to preserve accuracy while debiasing the model to particular demographic traits
- Intriguing by-product of much better “out of demography” results for adversarially-trained method

# Talk Outline

1. Introduction
2. Robustness through Data Augmentation
3. Robustness through Cross-domain Debiasing
4. Robustness and Privacy through Author-demographic Debiasing
5. Summary

# Overall Summary

- Three approaches to robustness, two of which are based on explicit debiasing:
  1. robustness through linguistically-motivated data augmentation [Li et al., 2017]
  2. robustness through cross-domain debiasing [Li et al., 2018b]
  3. robustness and privacy through author-demographic debiasing [Li et al., 2018a]
- In each case, we were able to boost cross-domain robustness (without any retraining to new domains), and also able to expose less user demographic details with the final method

# Acknowledgements

- Joint work with Yitong Li and Trevor Cohn
- This work was supported by the Australian Research Council

# References

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. English Web Treebank. *Linguistic Data Consortium, Philadelphia, USA*, 2012.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, 2007.

Ann Copestake and Dan Flickinger. An open source grammar development environment and broad-coverage english grammar using HPSG. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.

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.

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.

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

Dirk Hovy. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 752–762, 2015.

Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 483–488, 2015.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177, Seattle, USA, 2004.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, 2016.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar, 2014. doi: 10.3115/v1/D14-1181.

Kevin Knight and Daniel Marcu. Statistics-based summarization-step one: Sentence compression. In *Proceedings of the 18th Annual Conference on Artificial Intelligence*, pages 703–710, Austin, USA, 2000.

# References

Yitong Li, Trevor Cohn, and Timothy Baldwin. Robust training under linguistic adversity. In *Proceedings of the 15th Conference of the EACL (EACL 2017)*, pages 21–27, Valencia, Spain, 2017.

Yitong Li, Timothy Baldwin, and Trevor Cohn. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, Melbourne, Australia, 2018a.

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, 2018b.

Fei Liu, Maria Vasardani, and Timothy Baldwin. Automatic identification of locative expressions from social media text: A comparative analysis. In *Proceedings of the 4th International Workshop on Location and the Web (LocWeb 2014)*, pages 9–16, Shanghai, China, 2014.

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.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990. URL `ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps`.

# References

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148, San Diego, USA, 2016.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 115–124, Ann Arbor, USA, 2005.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, 2017.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, pages 1524–1534, Edinburgh, UK, 2011.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, USA, 2013.

# References

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, Banff, Canada, 2014.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 2979–2989, 2017.