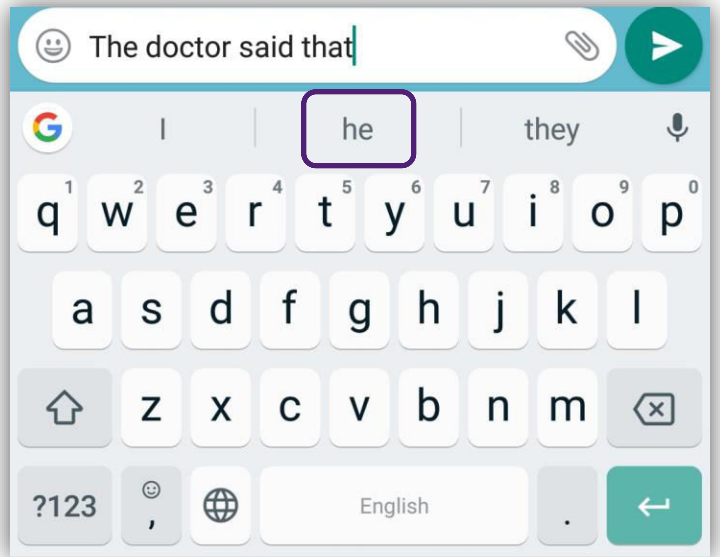
Identifying and Reducing Gender Bias in Word-Level Language Models

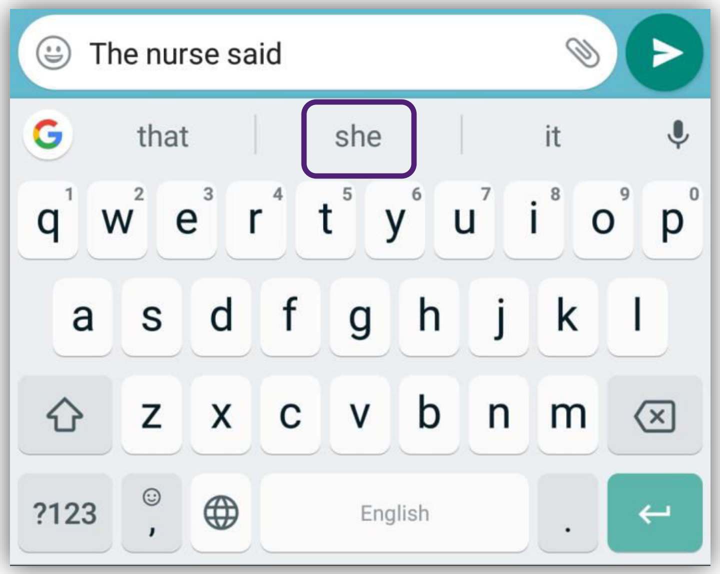
Shikha Bordia

Samuel R. Bowman

Word Prediction Models



The doctor said that

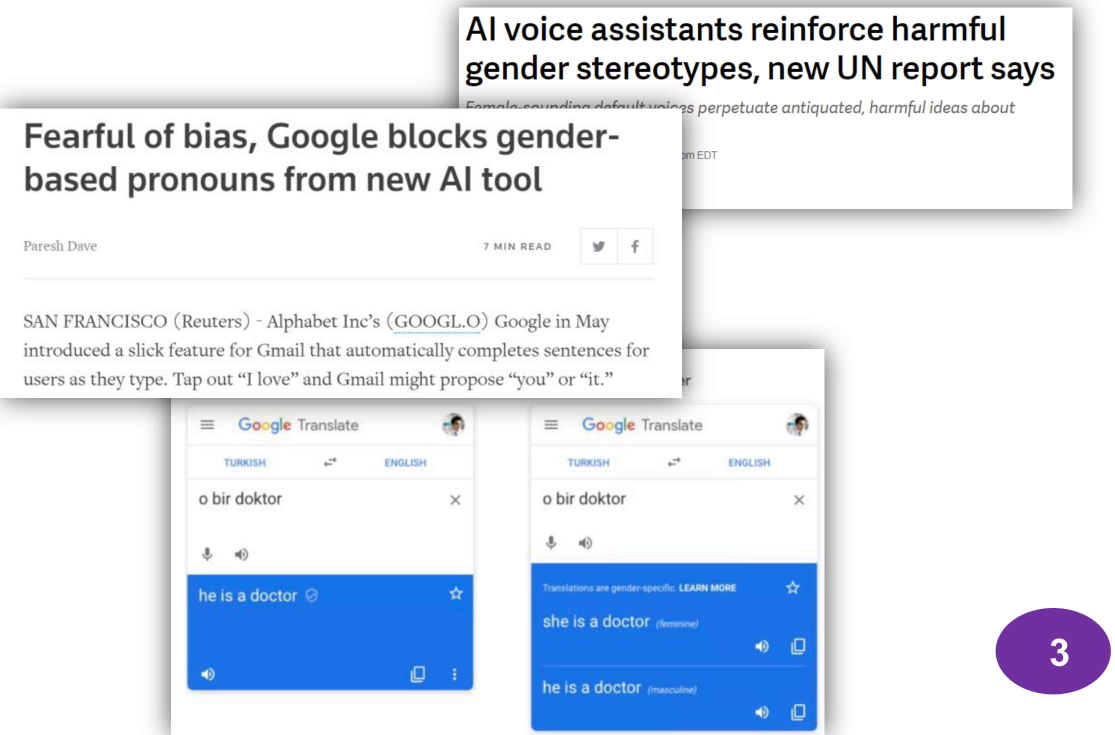


The nurse said

Source: Whatsapp

Bias in Language models

- Machine Translation
- Image Captioning
- Chatbots
- Text Summarization



AI voice assistants reinforce harmful gender stereotypes, new UN report says

Female-sounding default voices perpetuate antiquated, harmful ideas about

pm EDT

Fearful of bias, Google blocks gender-based pronouns from new AI tool

Paresh Dave

7 MIN READ

SAN FRANCISCO (Reuters) - Alphabet Inc's (GOOGL.O) Google in May introduced a slick feature for Gmail that automatically completes sentences for users as they type. Tap out "I love" and Gmail might propose "you" or "it."

o bir doktor

he is a doctor

o bir doktor

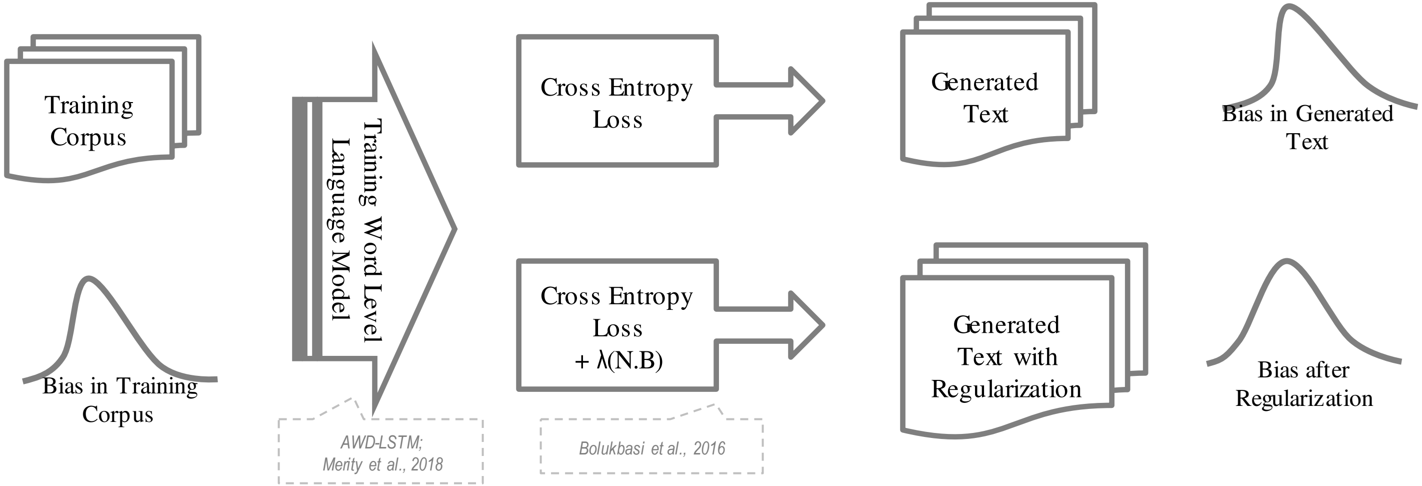
Translations are gender-specific. LEARN MORE

she is a doctor (feminine)

he is a doctor (masculine)

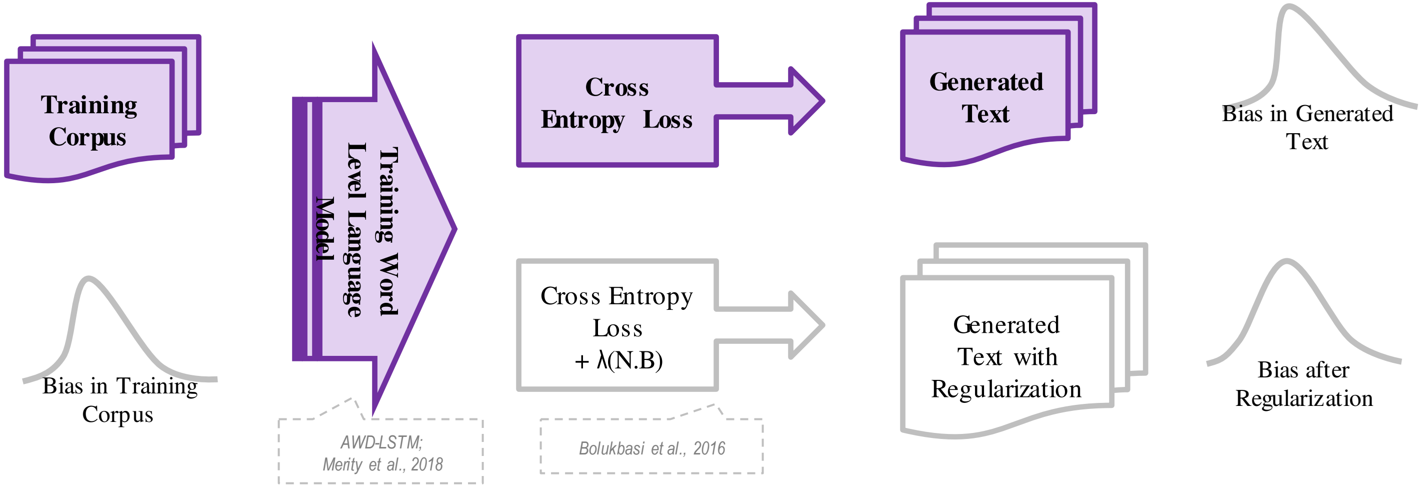
Source: Reuters, Verge

Overview


Training Corpus
Bias in Training Corpus
Training Word Level Language Model
AWD-LSTM; Merity et al., 2018
Cross Entropy Loss
Cross Entropy Loss + λ(N.B)
Bolukbasi et al., 2016
Generated Text
Bias in Generated Text
Generated Text with Regularization
Bias after Regularization

1. Propose a Bias Metric
2. Measure Bias at Corpus Level
3. Propose a Regularization Term
4. Evaluate Efficacy of Proposed Method

OVERVIEW



Training Corpus

Bias in Training Corpus

Training Word Level Language Model

AWD-LSTM;
Merity et al., 2018

Cross Entropy Loss

Cross Entropy Loss
+ λ(N.B)

Bolukbasi et al., 2016

Generated Text

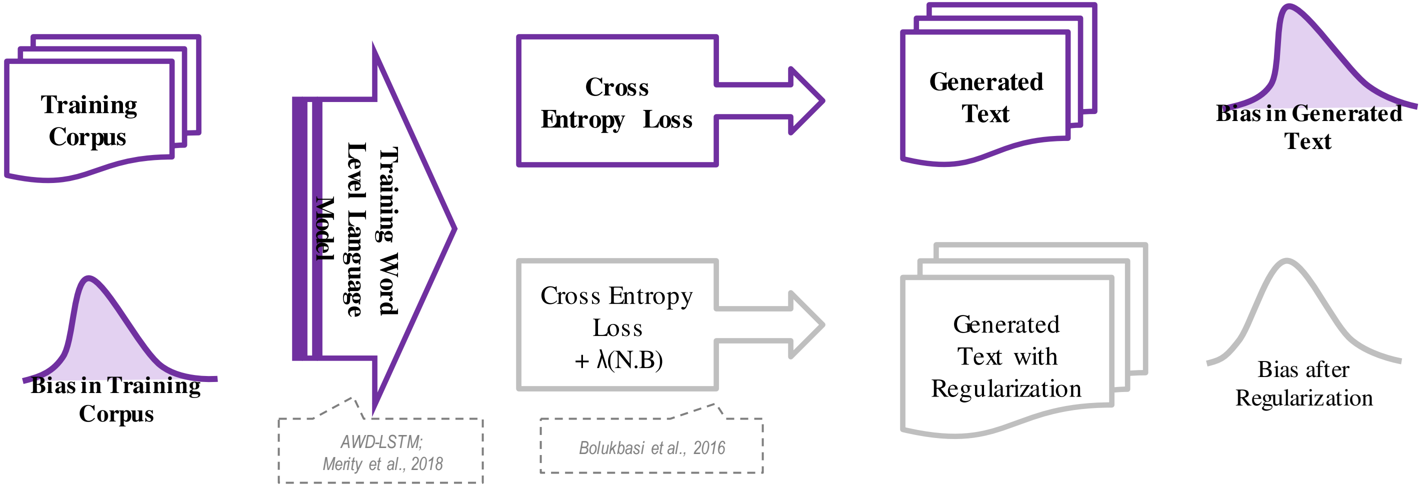
Generated Text with Regularization

Bias in Generated Text

Bias after Regularization

1. Propose a Bias Metric
2. Measure Bias at Corpus Level
3. Propose a Regularization Term
4. Evaluate Efficacy of Proposed Method

OVERVIEW


Training Corpus
Bias in Training Corpus
Training Word Level Language Model
AWD-LSTM; Merity et al., 2018
Cross Entropy Loss
Cross Entropy Loss + λ(N.B)
Bolukbasi et al., 2016
Generated Text
Generated Text with Regularization
Bias in Generated Text
Bias after Regularization

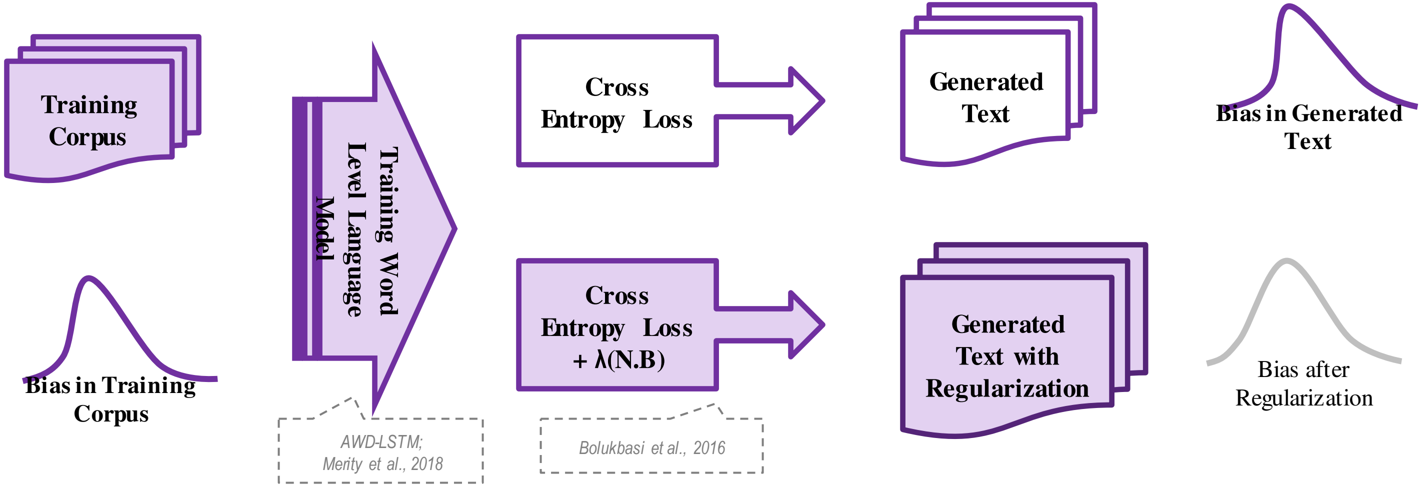
1. Propose a Bias Metric

2. Measure Bias at Corpus Level

3. Propose a Regularization Term

4. Evaluate Efficacy of Proposed Method

OVERVIEW



Training Corpus

Bias in Training Corpus

Training Word Level Language Model

AWD-LSTM; Merity et al., 2018

Cross Entropy Loss

Cross Entropy Loss + λ(N.B)

Bolukbasi et al., 2016

Generated Text

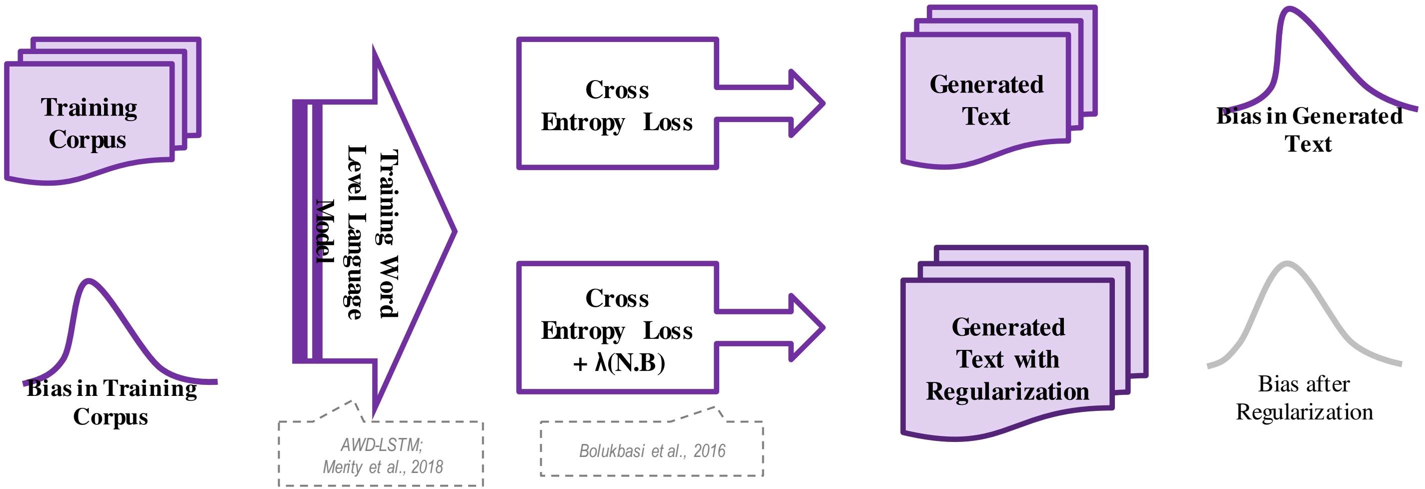
Generated Text with Regularization

Bias in Generated Text

Bias after Regularization

1. Propose a Bias Metric
2. Measure Bias at Corpus Level
3. Propose a Regularization Term
4. Evaluate Efficacy of Proposed Method

Overview



Training Corpus

Bias in Training Corpus

Training Word Level Language Model

AWD-LSTM; Merity et al., 2018

Cross Entropy Loss

Cross Entropy Loss + λ(N.B)

Bolukbasi et al., 2016

Generated Text

Generated Text with Regularization

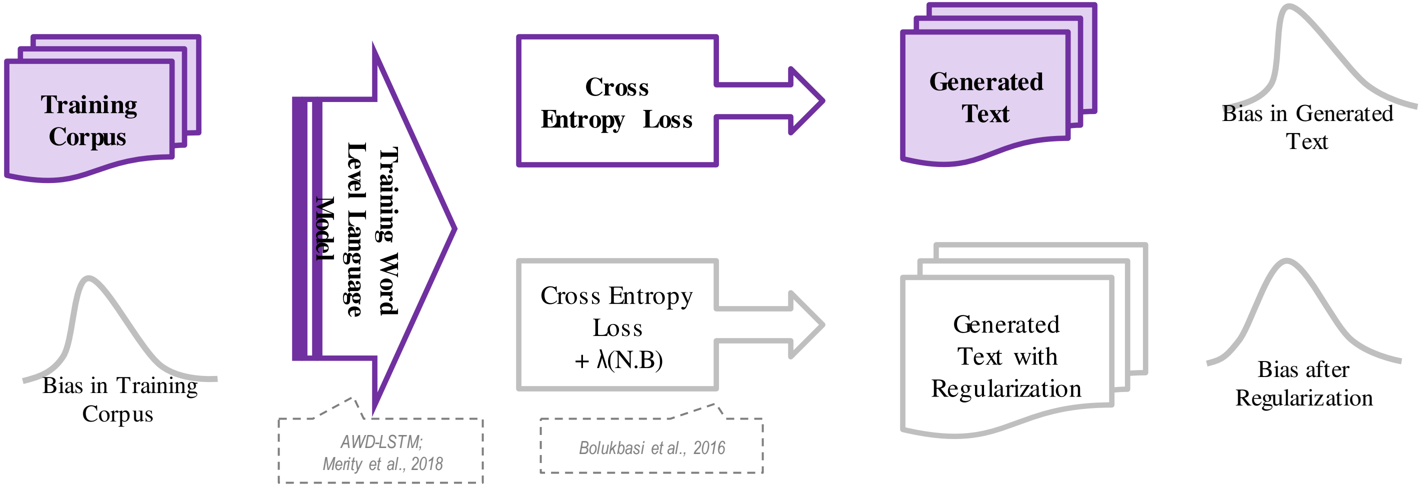
Bias in Generated Text

Bias after Regularization

1. Propose a Bias Metric
2. Measure Bias at Corpus Level
3. Propose a Regularization Term
4. Evaluate Efficacy of Proposed Method

Detailed Discussion

Detailed Discussion



Training Corpus

Bias in Training Corpus

Training Word Level Language Model

AWD-LSTM; Merity et al., 2018

Cross Entropy Loss

Generated Text

Bias in Generated Text

Cross Entropy Loss + λ(N.B)

Bolukbasi et al., 2016

Generated Text with Regularization

Bias after Regularization

1. Propose a Bias Metric

2. Measure Bias at Corpus Level

3. Propose a Regularization Term

4. Evaluate Efficacy of Proposed Method

Quantifying Bias: Gender Words

Male	← →	Female
Man	← →	Woman
Husband	← →	Wife
He	← →	She
Brother	← →	Sister
Father	← →	Mother
Uncle	← →	Aunt
Nephew	← →	Niece
Grandfather	← →	Grandmother
Actor	← →	Actress
…	← →	…

Quantifying Bias

- Probability of a word occurring in context with gendered words

$$P(w|g) = \frac{c(w,g)/\Sigma_i c(w_i,g)}{c(g)/\Sigma_i c(w_i)}$$

- Bias Score Definition

$$bias(w) = \log\left(\frac{P(w|f)}{P(w|m)}\right)$$

Measuring Bias: An Example

Sample Text: Training Corpus

….. location for the village as well as his medical career . dr farrer with his wife joan and children john peter and <unk> leaving australia in <unk> . the late *doctor* s son dr farrer pictured said the clock stopping was a nice touch as his father was so dedicated to it . born in sydney australia in <unk> his family later moved to melbourne and he was educated at <unk> grammar one of australia s oldest public schools . later he went to medical school and trained as a *doctor* . while at the alfred hospital in melbourne he met joan an operating theatre nurse and they were married in <unk> . in the early <unk> a <unk> arrived to say that his uncle roland farrer had died in england and the *doctor* was faced with the choice of taking over the yorkshire estate that had been in the family since the <unk> . he and his family took up residence in november <unk> where he worked until he retired . the *doctor* became ill in november <unk> and after a period in hospital returned to his home of <unk>

MEASURING BIAS: AN EXAMPLE

Sample Text: Identify Target Word – "Doctor"

….. location for the village as well as his medical career . dr farrer with his wife joan and children john peter and <unk> leaving australia in <unk> . the late ***doctor*** s son dr farrer pictured said the clock stopping was a nice touch as his father was so dedicated to it . born in sydney australia in <unk> his family later moved to melbourne and he was educated at <unk> grammar one of australia s oldest public schools . later he went to medical school and trained as a ***doctor*** . while at the alfred hospital in melbourne he met joan an operating theatre nurse and they were married in <unk> . in the early <unk> a <unk> arrived to say that his uncle roland farrer had died in england and the ***doctor*** was faced with the choice of taking over the yorkshire estate that had been in the family since the <unk> . he and his family took up residence in november <unk> where he worked until he retired . the ***doctor*** became ill in november <unk> and after a period in hospital returned to his home of <unk>

Measuring Bias: An Example

Sample Text: Identify Words in Context Window

….. location for the village as well as his medical career . dr farrer with his wife joan and children john peter and <unk> leaving australia in <unk> . the late ***doctor*** s son dr farrer pictured said the clock stopping was a nice touch as his father was so dedicated to it . born in sydney australia in <unk> his family later moved to melbourne and he was educated at <unk> grammar one of australia s oldest public schools . later he went to medical school and trained as a ***doctor*** . while at the alfred hospital in melbourne he met joan an operating theatre nurse and they were married in <unk> . in the early <unk> a <unk> arrived to say that his uncle roland farrer had died in england and the ***doctor*** was faced with the choice of taking over the yorkshire estate that had been in the family since the <unk> . he and his family took up residence in november <unk> where he worked until he retired . the ***doctor*** became ill in november <unk> and after a period in hospital returned to his home of <unk

Measuring Bias: An Example

Sample Text: Identify Gender Words in Context

john peter and <unk> leaving australia in <unk> . the late *doctor* s **son** dr farrer pictured said the clock stopping

later **he** went to medical school and trained as a *doctor* . while at the alfred hospital in melbourne **he** met joan

his uncle roland farrer had died in england and the *doctor* was faced with the choice of taking over the yorkshire

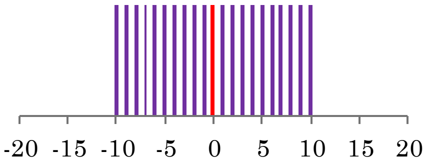
in november <unk> where **he** worked until **he** retired . the *doctor* became ill in november <unk> and after a period in

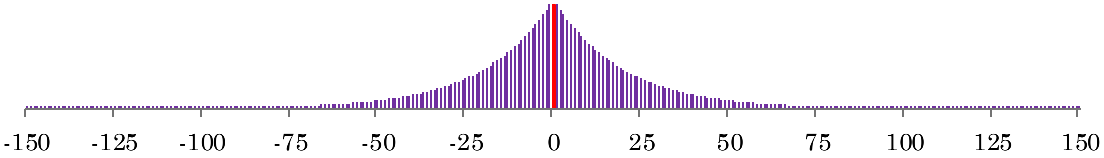
Measuring Bias: An Example

Sample Text: Large Context Window captures more

….. location for the village as well as **his** medical career . dr farrer with **his wife** joan and children john peter and <unk> leaving australia in <unk> . the late *doctor* s son dr farrer pictured said the clock stopping was a nice touch as **his father** was so dedicated to it . born in sydney australia in <unk> **his** family later moved to melbourne and **he** was educated at <unk> grammar one of australia s oldest public schools . later he went to medical school and trained as a *doctor* . while at the alfred hospital in melbourne he met joan an operating theatre nurse and they were married in <unk> . in the early <unk> a <unk> arrived to say that his uncle roland farrer had died in england and the *doctor* was faced with the choice of taking over the yorkshire estate that had been in the family since the <unk> . **he** and **his** family took up residence in november <unk> where he worked until he retired . the *doctor* became ill in november <unk> and after a period in hospital returned to **his** home of <unk>

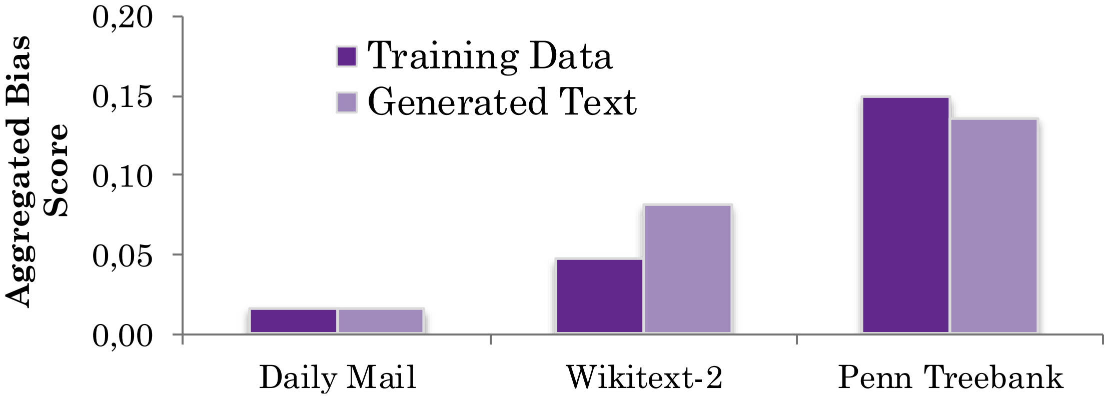
Measuring Bias: Defining Context


Small Context With Uniform Weights


Bigger Context With Exponentially Decaying Weights

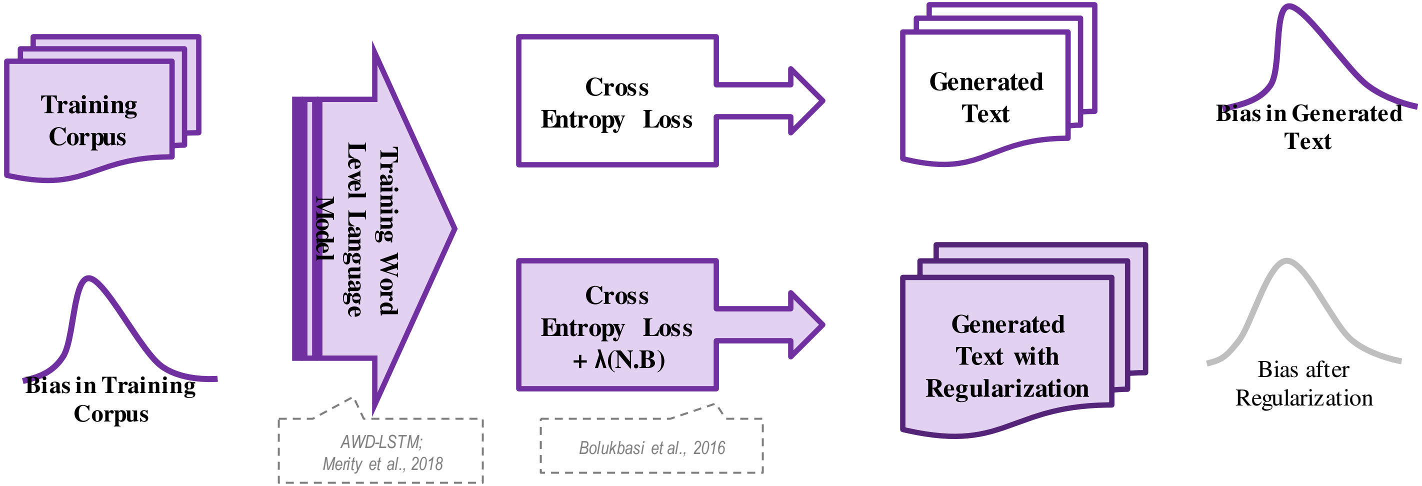
Bias Measure: Comparison of Datasets

- Relative Bias Scores of Datasets:
  Daily Mail < Wikitext-2 < Penn Treebank



1. Propose Metric | **2. Measure Bias** | 3. Propose Regularization | 4. Evaluate Efficacy

DETAILED DISCUSSION

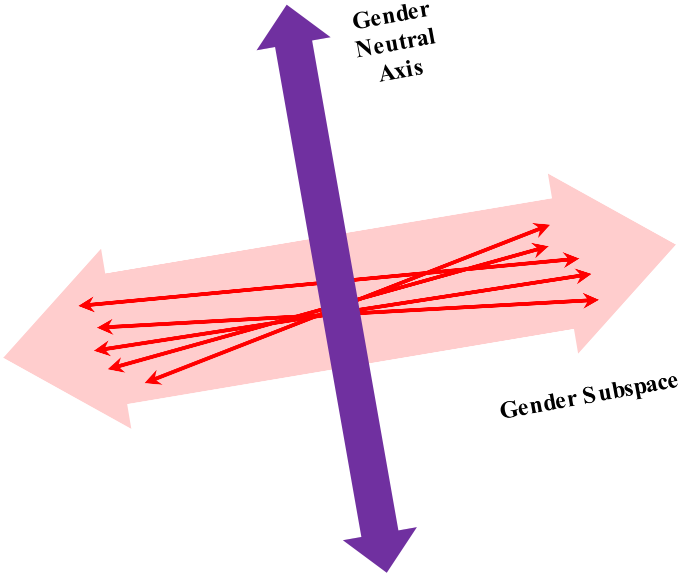


Training Corpus
Bias in Training Corpus
Training Word Level Language Model
AWD-LSTM; Merity et al., 2018
Cross Entropy Loss
Cross Entropy Loss + λ(N.B)
Bolukbasi et al., 2016
Generated Text
Generated Text with Regularization
Bias in Generated Text
Bias after Regularization

1. Propose a Bias Metric
2. Measure Bias at Corpus Level
3. Propose a Regularization Term
4. Evaluate Efficacy of Proposed Method

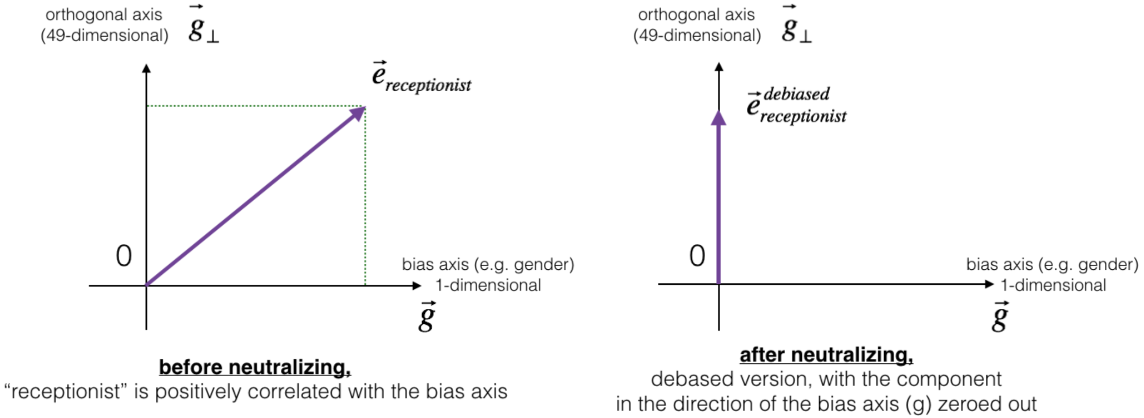
Regularization: Gender Subspace

Male	← →	Female
Man	← →	Woman
Husband	← →	Wife
He	← →	She
Brother	← →	Sister
Father	← →	Mother
Uncle	← →	Aunt
Nephew	← →	Niece
Grandfather	← →	Grandmother
Actor	← →	Actress
…	← →	…



Regularization: Loss Term

$$\mathcal{L}_B = \lambda \|NB\|_F^2$$

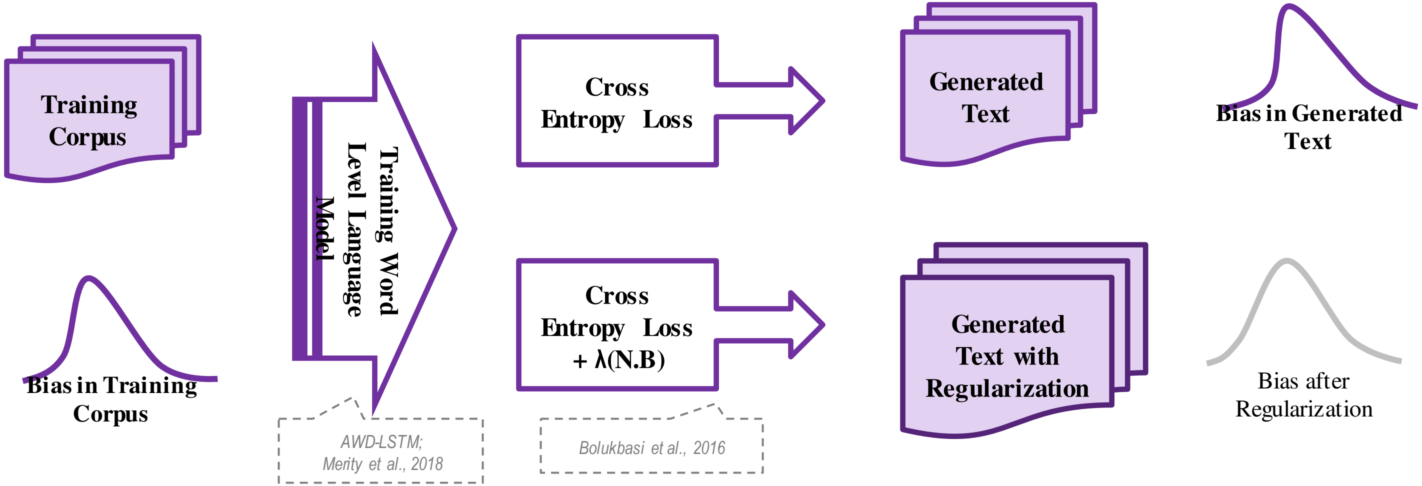


before neutralizing,
"receptionist" is positively correlated with the bias axis

after neutralizing,
debased version, with the component
in the direction of the bias axis (g) zeroed out

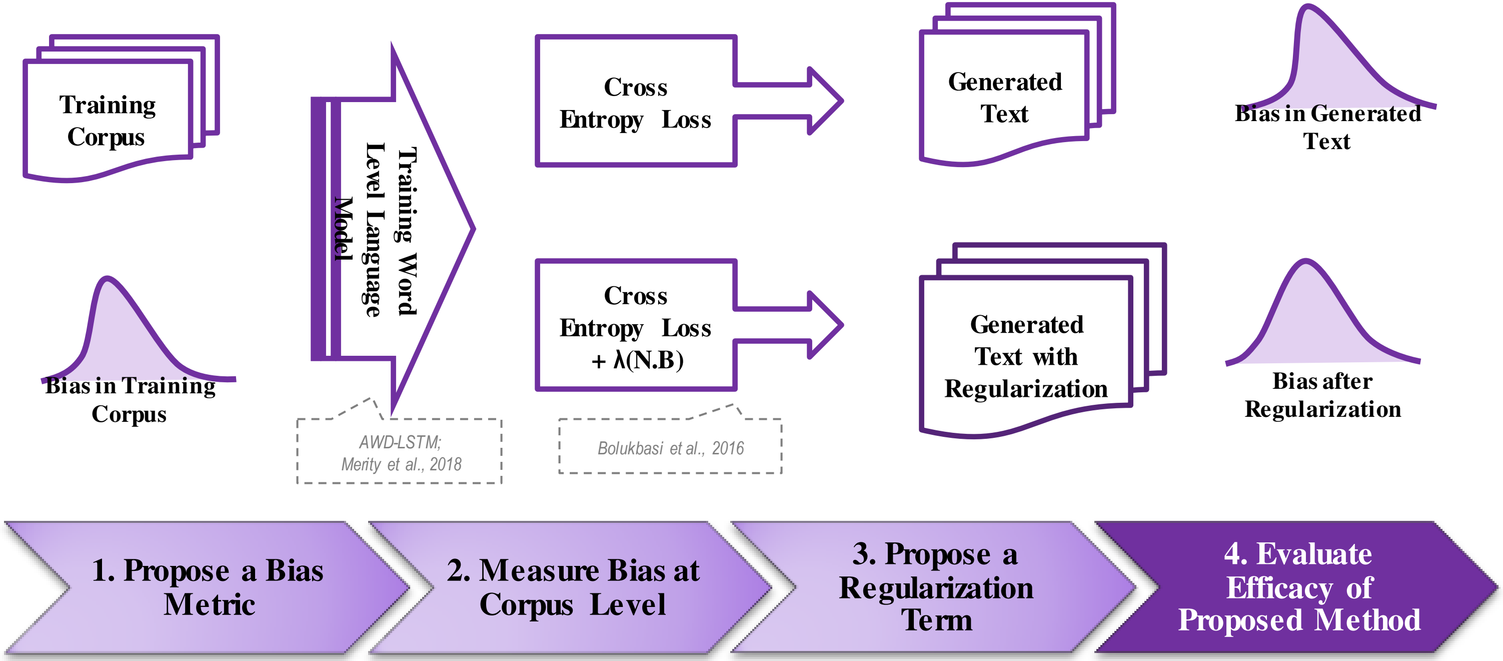
λ controls the importance of minimizing bias in the embedding matrix

Detailed Discussion


Training Corpus
Bias in Training Corpus
Training Word Level Language Model
AWD-LSTM; Merity et al., 2018
Cross Entropy Loss
Cross Entropy Loss + λ(N.B)
Bolukbasi et al., 2016
Generated Text
Generated Text with Regularization
Bias in Generated Text
Bias after Regularization

1. Propose a Bias Metric
2. Measure Bias at Corpus Level
3. Propose a Regularization Term
4. Evaluate Efficacy of Proposed Method

Detailed Discussion



Training Corpus

Bias in Training Corpus

Training Word Level Language Model

AWD-LSTM; Merity et al., 2018

Cross Entropy Loss

Cross Entropy Loss + λ(N.B)

Bolukbasi et al., 2016

Generated Text

Generated Text with Regularization

Bias in Generated Text

Bias after Regularization

1. Propose a Bias Metric

2. Measure Bias at Corpus Level

3. Propose a Regularization Term

4. Evaluate Efficacy of Proposed Method

Measure the Effect of Debiasing

- Distribution of bias

$$\mu_\lambda = mean(abs(bias_\lambda))$$
$$\sigma_\lambda = stdev(bias_\lambda)$$

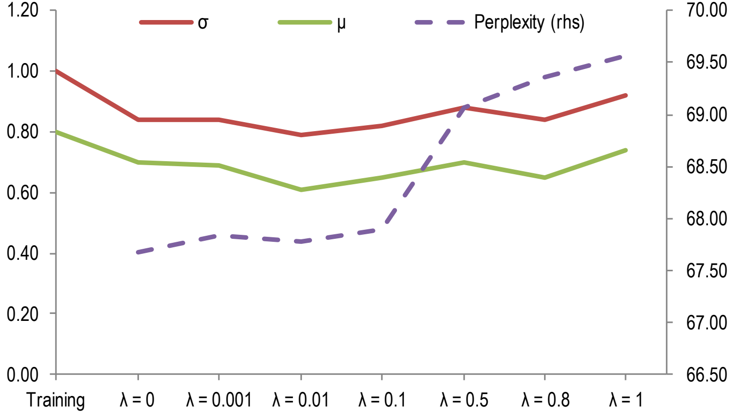
- Amplification of bias

$$bias_\lambda(w) = \beta * bias_{train}(w) + c$$

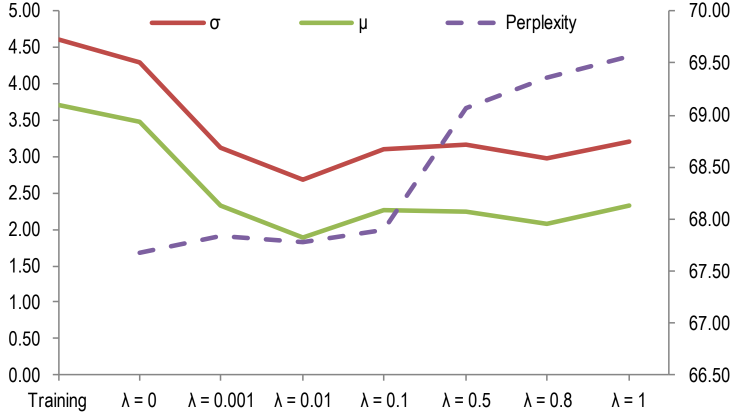
EFFECT OF DEBIASING: WIKITEXT2

Perplexity-bias trade off

Fixed Context



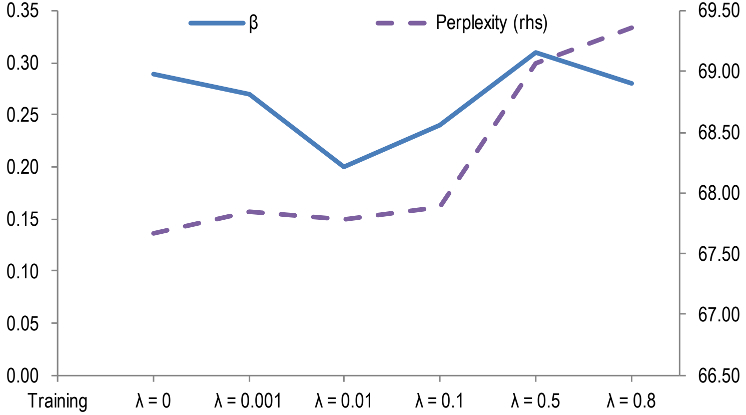
Exponentially Decay Context



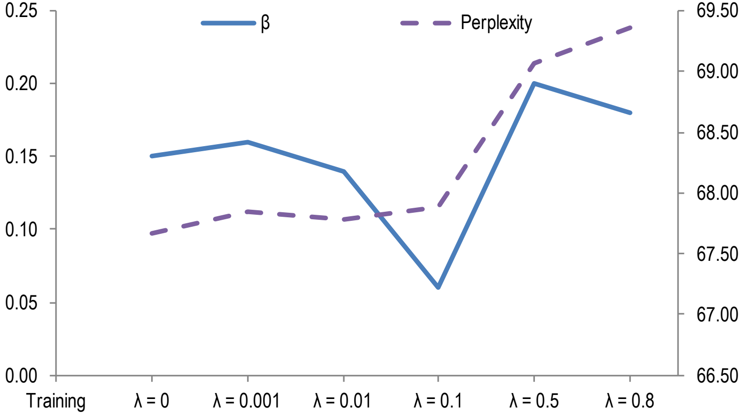
EFFECT OF DEBIASING: WIKITEXT2

Perplexity-bias trade off

Fixed Context



Exponentially Decay Context



EXAMPLES – GENERATED TEXT

	Crying	Prisoner
No Regularization	**she** was put on **her** own machine to raise money for **her** own wedding route which saw **her** **crying** and down a programme today . effects began by bottom of **her** marrow the ”	**his** legs and allegedly killed **himself** by suspicious points . in the latest case after an online page **he** left **prisoner** in **his** home in near manhattan on saturday when **he** was struck in **his** car operating in bay smoking and when **he** had”
High Regularization	**he** discovered peaceful facebook remains when **he** was caught **crying** officers but was arrested after they found the crash hire a **man** brown shocked **his brother** over	the ankle follows a worker **her** **prisoner** **she** died this year before now an profile which clear **her** eye borrowed for **her** organ own role . it was a huge accident after the drugs **she** had

Thank You

- sb6416@nyu.edu
- bowman@nyu.edu