It’s Morphin’ Time!

Combating Linguistic Discrimination with Inflectional Perturbations

Samson Tan, Shafiq Joty, Min-Yen Kan, Richard Socher

@samsontmr | samson.tan@salesforce.com

ACL 2020

Language technology is increasingly ubiquitous

Search

Google brings in BERT to improve its search results (TechCrunch)

Translation

Facebook adds 24 new languages to its automated translation service (VentureBeat)

Chatbots

Collaborating chatbots to form a digital workforce (Forbes)

BUT

salesforce

State of the art models are trained on only Standard English (often U.S. English)

WIKIPEDIA

BERT

What color are her eyes?
What is the mustache made of?

SQuAD2.0
The Stanford Question Answering Dataset

VQA Visual Question Answering

https://wikipedia.org; https://visualqa.org; https://rajpurkar.github.io/SQuAD-explorer
http://assets.gcstatic.com/u/apps/asset_manager/uploaded/2017/11/bert-sesame-street-twitter-profile-1489575494-custom-0.png

BUT

State of the art models are trained on only Standard English (often U.S. English)

Identical Train-Test Distributions Assumption:

All users speak error-free Standard (U.S.) English

BUT

salesforce

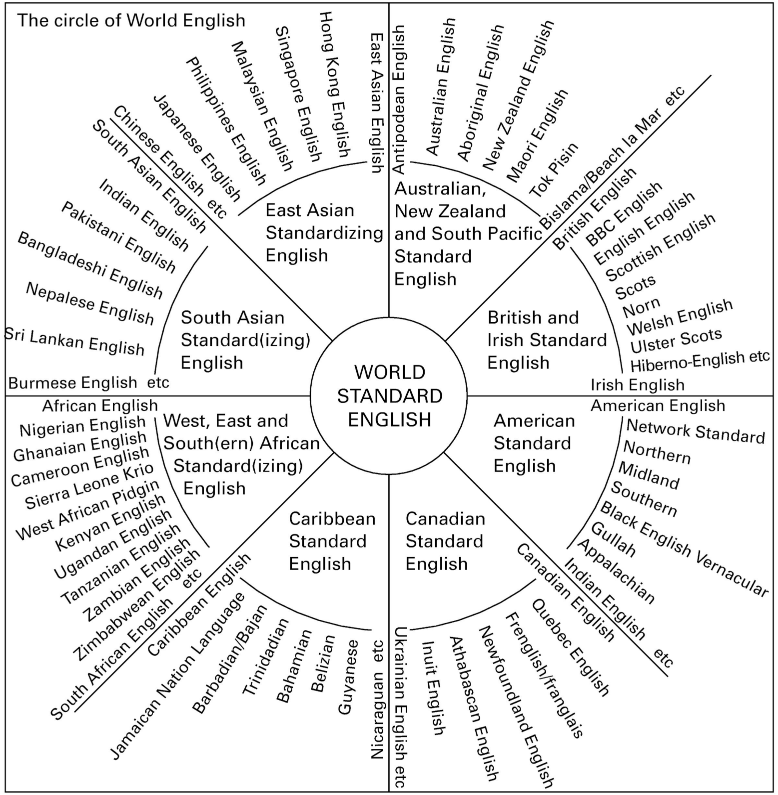
State of the art models are trained on only Standard English (often U.S. English)

↓

Identical Train-Test Distributions Assumption:
All users speak error-free Standard (U.S.) English

English == Standard U.S. English??

Not everyone speaks Standard U.S. English



MacArthur's Circle of World English; image source: :http://singlish.it/english-language/world-englishes

Not everyone speaks English perfectly

Ethnologue 2019:

Rank	Language	L1 speakers	L2 speakers	L2 Rank	Total
1	English	379.0 million	753.3 million	1	1.132 billion[5]

⅔ speak English as a second language

Image source: https://wikipedia.org

Ethical Implications: Linguistic Discrimination

The Guardian

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

▲ Facebook's machine translation mix-up sees man questioned over innocuous post confused with attack threat. Photograph: Thibault Camus/AP

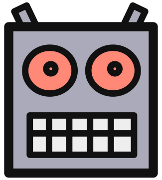
Facebook has apologised after an error in its machine-translation service saw Israeli police arrest a Palestinian man for posting "good morning" on his social media profile.

The man, a construction worker in the West Bank settlement of Beitar Illit,

Discrimination against speakers of non-standard Englishes

- Not understanding/misinterpreting them
- In some countries: likely to be ethnic minorities

Sorry, I didn't get that..



https://commons.wikimedia.org/wiki/File:Robot_icon.svg
https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

Are English NLP models biased against non-standard English speakers?

Inflectional Morphology

- Inflections indicate the tense, quantity, etc. of content words
- Many World Englishes exhibit inflectional variation
- Morphological acquisition is challenging for L2 learners

Part of Speech	Grammatical Category	Inflection	Examples
Noun	Number	-s, -es	Flower → Flowers Glass → Glasses
Noun, Pronoun	Case (Genitive)	-'s, -', -s	Paul → Paul's Francis → Francis' It → Its
Pronoun	Case (Reflexive)	-self, -selves	Him → Himself Them → Themselves
Verb	Aspect (Progressive)	-ing	Run → Running
Verb	Aspect (Perfect)	-en, -ed	Fall → (Has) fallen Finish → (Has) finished
Verb	Tense (Past)	-ed	Open → Opened
Verb	Tense (Present)	-s	Open → Opens
Adjective	Degree of Comparison (Comparative)	-er	Smart → Smarter
Adjective	Degree of Comparison (Superlative)	-est	Smart → Smartest

https://www.thoughtco.com/inflection-grammar-term-1691168

Example

When are the suspended team schedule to returned?

vs

When is the suspended team scheduled to return?

How robust are English NLP models to non-standard inflections?

Adv. Examples (Question Answering)

When are the suspended team schedule to returned?

Answer: 2018 → no answer

Who did BSkyB had an operating licenses from?

Answer: Ofcom → no answer

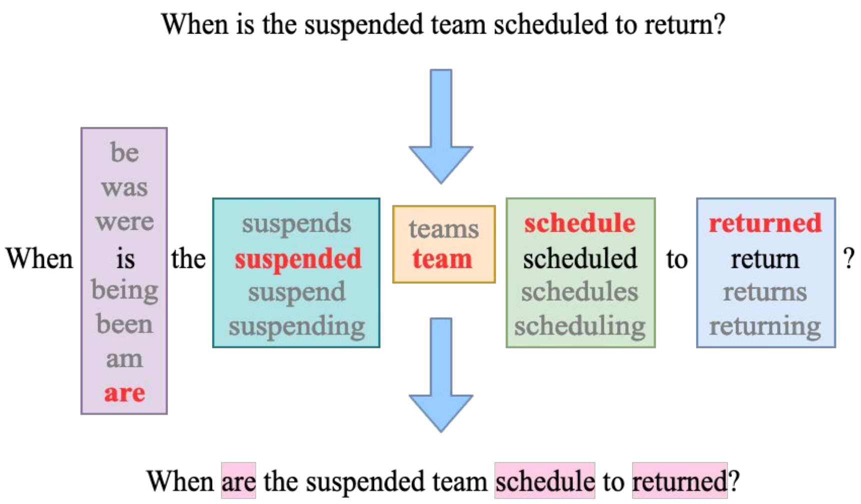
Intractable problem lacking polynomial time solutions necessarily negate the practical efficacy of what type of algorithm?

Answer: Exponential-time algorithms → polynomial time

Morpheus

Key idea: Perturb inflectional morphology of content words

- Find the inflections that maximize the model's loss (adversarial example)
- Only needs black-box access



Algorithm 1 Morpheus

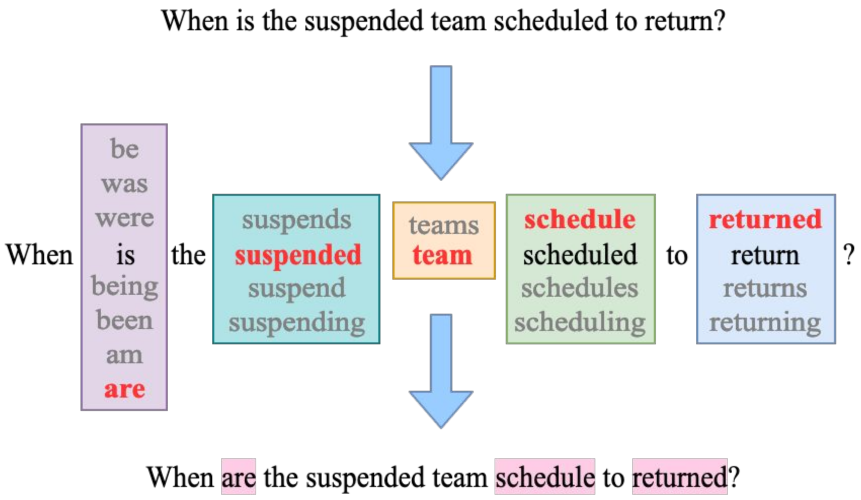
Require: Original instance x, Label y, Model f
Ensure: Adversarial example $\hat{x}$

$T \leftarrow \text{TOKENIZE}(x)$
for all $t_i \in T$ **do**
  if $\text{POS}(t_i) \in \{\text{NOUN}, \text{VERB}, \text{ADJ}\}$ **then**
    $I \leftarrow \text{GETINFLECTIONS}(t_i)$
    $t_i \leftarrow \text{MAXINFLECTED}(I, y, f)$
  end if
end for
$\hat{x} \leftarrow \text{DETOKENIZE}(T)$

Morpheus

Preserves semantics

- Base forms + word order remains unchanged
- Generates **plausible** + **semantically similar** adversarial examples



Algorithm 1 Morpheus

Require: Original instance x, Label y, Model f
Ensure: Adversarial example $\hat{x}$
  $T \leftarrow \text{TOKENIZE}(x)$
  for all $t_i \in T$ **do**
    if $\text{POS}(t_i) \in \{\text{NOUN}, \text{VERB}, \text{ADJ}\}$ **then**
      $I \leftarrow \text{GETINFLECTIONS}(t_i)$
      $t_i \leftarrow \text{MAXINFLECTED}(I, y, f)$
    end if
  end for
  $\hat{x} \leftarrow \text{DETOKENIZE}(T)$

Experiments

Tasks

Extractive Question Answering

- SQuAD 2.0
  - Answerable Questions (SQuAD 1.1)
  - Unanswerable Questions
- Models
  - BiDAF
  - ELMo-BiDAF
  - BERT
  - SpanBERT

Machine Translation

- WMT’14 English-French
- Models
  - Convolutional Seq2Seq
  - Transformer-big

SQuAD 2.0 models are significantly more brittle

Dataset	Model	Original	MORPHEUS
SQuAD 2.0 Answerable Questions (F_1)	GloVe-BiDAF	78.67	**53.94 (−31.43%)**
	ELMo-BiDAF	80.90	62.17 (−23.15%)
	$\text{BERT}_{\text{SQuAD 1.1}}$	93.14	82.79 (−11.11%)
	$\text{SpanBERT}_{\text{SQuAD 1.1}}$	91.88	82.86 (−9.81%)
	$\text{BERT}_{\text{SQuAD 2}}$	81.19	**57.47 (−29.21%)**
	$\text{SpanBERT}_{\text{SQuAD 2}}$	88.52	69.47 (−21.52%)

SQuAD 2.0 models are significantly more brittle

SQuAD 2.0 Answerable Questions (F_1)			
Original	Transfer	Clean	MORPHEUS
GloVe-BiDAF	$\text{BERT}_{\text{SQuAD 1.1}}$	93.14	89.67
	$\text{SpanBERT}_{\text{SQuAD 1.1}}$	91.88	90.75
	$\text{BERT}_{\text{SQuAD 2}}$	81.19	72.21
	$\text{SpanBERT}_{\text{SQuAD 2}}$	88.52	81.95
$\text{BERT}_{\text{SQuAD 1.1}}$	GloVe-BiDAF	78.67	71.33
	$\text{SpanBERT}_{\text{SQuAD 1.1}}$	91.88	88.68
	$\text{BERT}_{\text{SQuAD 2}}$	81.19	69.68
	$\text{SpanBERT}_{\text{SQuAD 2}}$	88.52	80.11
$\text{SpanBERT}_{\text{SQuAD 1.1}}$	GloVe-BiDAF	78.67	71.41
	$\text{BERT}_{\text{SQuAD 1.1}}$	93.14	87.48
	$\text{BERT}_{\text{SQuAD 2}}$	81.19	70.05
	$\text{SpanBERT}_{\text{SQuAD 2}}$	88.52	77.89

Example (Machine Translation)

salesforce

Original
The announcement came as fighting raged Thursday in the town of Safira, which experts say is home to a chemical weapons production facility as well as storage sites, reported the Britain-based Syrian Observatory for Human Rights.

Adversarial Example
The announcements coming as fight rage Thursday in the towns of Safira, which expert say is home to a chemical weapons production facility as well as storage site, reporting the Britain-based Syrian Observatory for Human Rights.

Original Translation
L'annonce a été faite alors que les combats faisaient rage jeudi dans la ville de Safira, qui, selon les experts, abrite une usine de fabrication d'armes chimiques ainsi que des sites de stockage, a indiqué l'Observatoire syrien des droits de l'homme, basé au Royaume-Uni.

Adversarial Example Translation:
Le président de la République, Nicolas Sarkozy, a annoncé jeudi que le président de la République, Nicolas Sarkozy, s'était rendu jeudi dans la capitale du pays, Nicolas Sarkozy.

[The President of the Republic, Nicolas Sarkozy, announced Thursday that the President of the Republic, Nicolas Sarkozy, had traveled Thursday in the capital of the country, Nicolas Sarkozy.]

Improving Robustness to Inflectional Perturbations

Adversarial Fine-Tuning

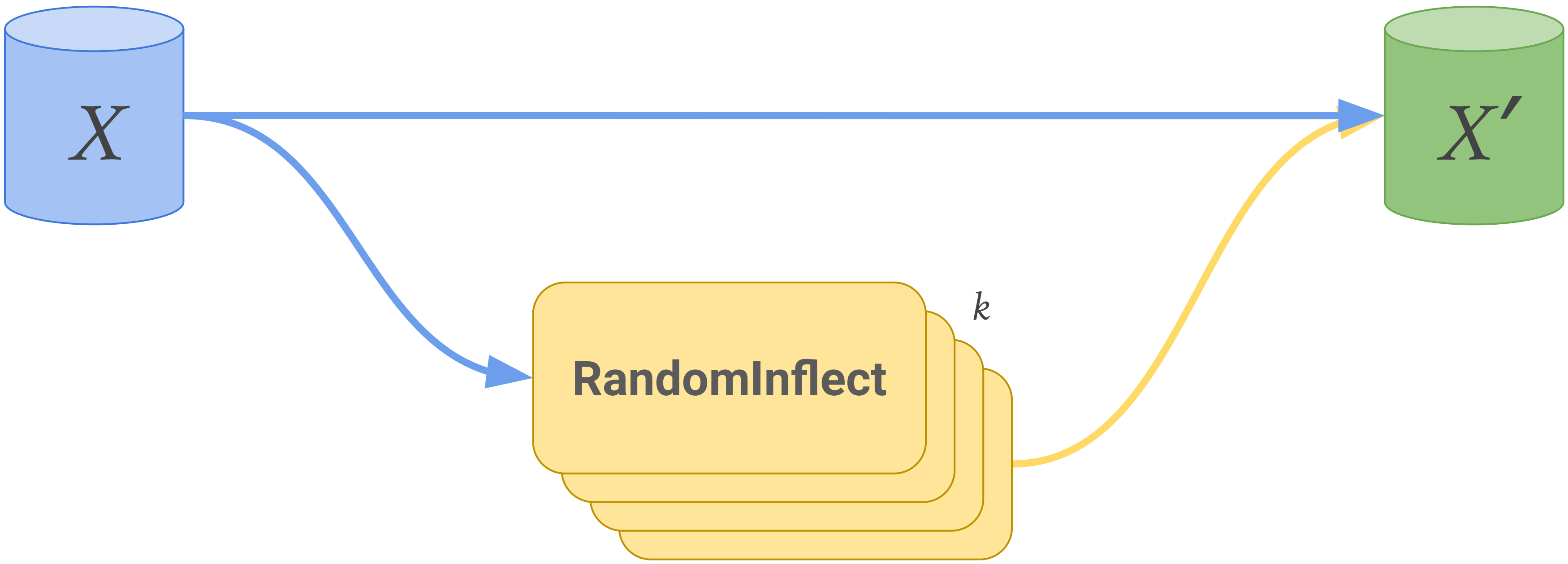
Key idea: ***Fine-tune*** trained models on adv. training set for 1 epoch

- Data generated via weighted random sampling using adversarial inflection distribution

Existing work ***retrains*** model on adversarial examples from scratch

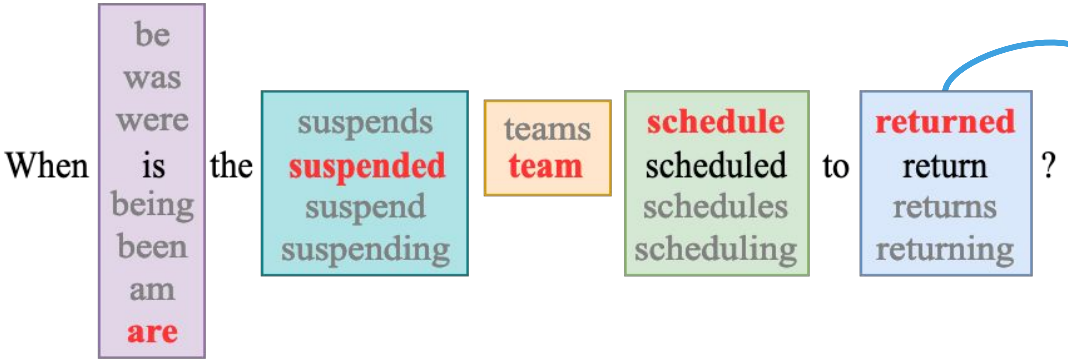
- Computationally expensive

Generating the Adversarial Training Set



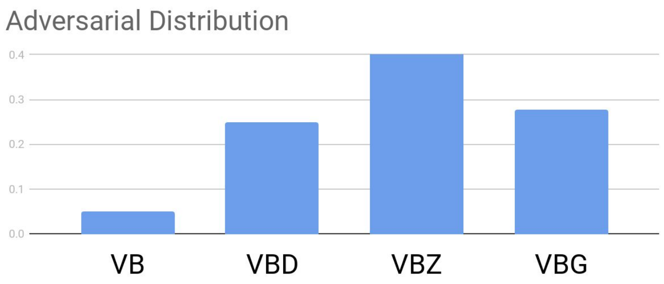
Generating the Adversarial Training Set

RandomInflect



$$x' = T_1 \ldots T_{|T|}$$

$$T_i \sim D_{adv}$$



Adversarial Fine-Tuning Improves Robustness!

Results

SpanBERT$_{\text{SQuAD 2}}$ (F$_1$)						
	Original			Adversarially Fine-tuned		
Dataset	Clean	MORPHEUS	Epoch	Clean	MORPHEUS$_{\text{orig}}$	MORPHEUS$_{\text{adv}}$
SQuAD 2.0 Ans	88.52	69.47 (−21.52%)	1	**86.80**	85.17 (−1.87%)	82.76 (−4.65%)
			4	86.15	**84.93 (−1.41%)**	**82.92 (−3.74%)**
SQuAD 2.0 All	87.71	73.26 (−16.47%)	1	86.00	84.72 (−1.48%)	82.41 (−4.17%)
			4	**87.08**	**85.93 (−1.32%)**	**84.71 (−2.72%)**
Transformer-big (BLEU)						
	Original			Adversarially Fine-tuned		
Dataset	Clean	MORPHEUS	Epoch	Clean	MORPHEUS$_{\text{orig}}$	MORPHEUS$_{\text{adv}}$
newstest2014	43.16	20.57 (−56.25%)	1	39.84	**31.79 (−20.20%)**	**31.43 (−21.10%)**
			4	**40.60**	31.99 (−21.20%)	30.82 (−24.08%)

Adversarial Fine-Tuning Improves Robustness!

Results

SpanBERT$_{\text{SQuAD 2}}$ (F_1)						
	Original			Adversarially Fine-tuned		
Dataset	Clean	MORPHEUS	Epoch	Clean	MORPHEUS$_{\text{orig}}$	MORPHEUS$_{\text{adv}}$
SQuAD 2.0 Ans	88.52	69.47 (−21.52%)	1	**86.80**	85.17 (−1.87%)	82.76 (−4.65%)
			4	86.15	**84.93 (−1.41%)**	**82.92 (−3.74%)**
SQuAD 2.0 All	87.71	73.26 (−16.47%)	1	86.00	84.72 (−1.48%)	82.41 (−4.17%)
			4	**87.08**	**85.93 (−1.32%)**	**84.71 (−2.72%)**
Transformer-big (BLEU)						
	Original			Adversarially Fine-tuned		
Dataset	Clean	MORPHEUS	Epoch	Clean	MORPHEUS$_{\text{orig}}$	MORPHEUS$_{\text{adv}}$
newstest2014	43.16	20.57 (−56.25%)	1	39.84	**31.79 (−20.20%)**	**31.43 (−21.10%)**
			4	**40.60**	31.99 (−21.20%)	30.82 (−24.08%)

Adversarial Fine-Tuning Improves Robustness!

Results

SpanBERT$_{\text{SQuAD 2}}$ (F_1)						
	Original			Adversarially Fine-tuned		
Dataset	Clean	MORPHEUS	Epoch	Clean	MORPHEUS$_{\text{orig}}$	MORPHEUS$_{\text{adv}}$
SQuAD 2.0 Ans	88.52	69.47 (−21.52%)	1	**86.80**	85.17 (−1.87%)	82.76 (−4.65%)
			4	86.15	**84.93 (−1.41%)**	**82.92 (−3.74%)**
SQuAD 2.0 All	87.71	73.26 (−16.47%)	1	86.00	84.72 (−1.48%)	82.41 (−4.17%)
			4	**87.08**	**85.93 (−1.32%)**	**84.71 (−2.72%)**
Transformer-big (BLEU)						
	Original			Adversarially Fine-tuned		
Dataset	Clean	MORPHEUS	Epoch	Clean	MORPHEUS$_{\text{orig}}$	MORPHEUS$_{\text{adv}}$
newstest2014	43.16	20.57 (−56.25%)	1	39.84	**31.79 (−20.20%)**	**31.43 (−21.10%)**
			4	**40.60**	31.99 (−21.20%)	30.82 (−24.08%)

Summary

salesforce

- Current models are trained on error-free, Standard (often U.S.) English
- Predisposes them to discriminate against non-standard dialect/L2 speakers
  - Known as **linguistic discrimination** or **linguicism**
- Adversarial examples targeting inflectional morphology expose this flaw
- Morpheus produces plausible + semantically similar adversaries
- Fine-tuning for a **single** epoch on an adv. training set improves robustness

Code: github.com/salesforce/morpheus


thank
you
BLAZE
YOUR
TRAIL
salesforce

Future Work

- Extend to other languages, in particular morphologically-rich languages
- Directly model L2/dialectal distributions
- Harden models without increasing dataset size

Examples (NMT)

Caused Transformer-big to output English

The first nine episode of Sheriffs Callie's Wild West will be available from November 24 on the site watchdisneyjunior.com or via its application for mobile phone and tablet.

Cue story about passport controls at Berwick and a barbed wires borders along Hadrian's Walls.

Cutting to the present are a rude awakenings, like snapped out of a dream.

Human Evaluation

Please choose the most suitable option for each question.

Who was this sentence likely written by?

Who upon arrive give the original viking settler a common identities?

- Native English speaker
- Someone who speaks English as a second language
- Beginner English learner or young child
- Not a human

What is the likelihood that the below sentences mean the same thing?

Who upon arrive give the original viking settler a common identities?

Who upon arriving gave the original viking settlers a common identity?

- Highly likely
- Likely
- Somewhat likely
- Somewhat unlikely
- Unlikely
- Highly unlikely

Submit

Human Evaluation

Results

Plausibility				
	Native U.S. English Speakers		Unrestricted	
	SQuAD 2.0	newstest2014	SQuAD 2.0	newstest2014
Native	11.58%	25.64%	22.82%	32.56%
L2 Speaker	**42.82%**	**42.30%**	**53.58%**	**52.82%**
Beginner	31.79%	23.33%	17.17%	10.25%
Non-human	13.84%	8.71%	6.41%	4.35%

Semantic Equivalence				
	Native U.S. English Speakers		Unrestricted	
	SQuAD 2.0	newstest2014	SQuAD 2.0	newstest2014
Highly Likely	**52.82%**	**62.30%**	33.84%	**40.76%**
Likely	20.51%	18.71%	**36.15%**	33.84%
Somewhat Likely	11.02%	7.94%	22.82%	19.48%
Somewhat Unlikely	6.92%	6.15%	5.38%	4.35%
Unlikely	3.58%	3.07%	1.53%	1.28%
Highly Unlikely	5.12%	1.79%	0.25%	0.25%

Ethical Implications

Discrimination against L2/nonstandard English speakers

- Not understanding/misinterpreting them
- In U.S. context: likely to be ethnic minorities

The Guardian

Facebook translates 'good morning' into 'attack them', leading to arrest

Palestinian man questioned by Israeli police after embarrassing mistranslation of caption under photo of him leaning against bulldozer

Related Work

Fairness in NLP

- Primarily focused on gender and racial biases
  (Bolukbasi et al., 2016; Rudinger et al., 2018; May et al., 2019; Bordia and Bowman, 2019)

Adversarial Attacks in NLP

- Character/word shuffling/insertion/deletion, synonym swapping
  (Jia and Liang, 2017; Belinkov and Bisk, 2018; Ebrahimi et al., 2018; Ribeiro et al., 2018; etc)
- Often changes the expected output of the model (with some exceptions)
- Does not make use of linguistic concepts like morphology

Adversarial Robustness

- Adversarial training: Computationally expensive
- Embedding averaging: Clean data performance affected
  (Belinkov and Bisk, 2018)