Investigating & Mitigating Infamous Weaknesses of Deep NLP

Piyawat Lertvittayakumjorn

pl1515@imperial.ac.uk

25th November 2019

First of all, deep NLP is awesome !

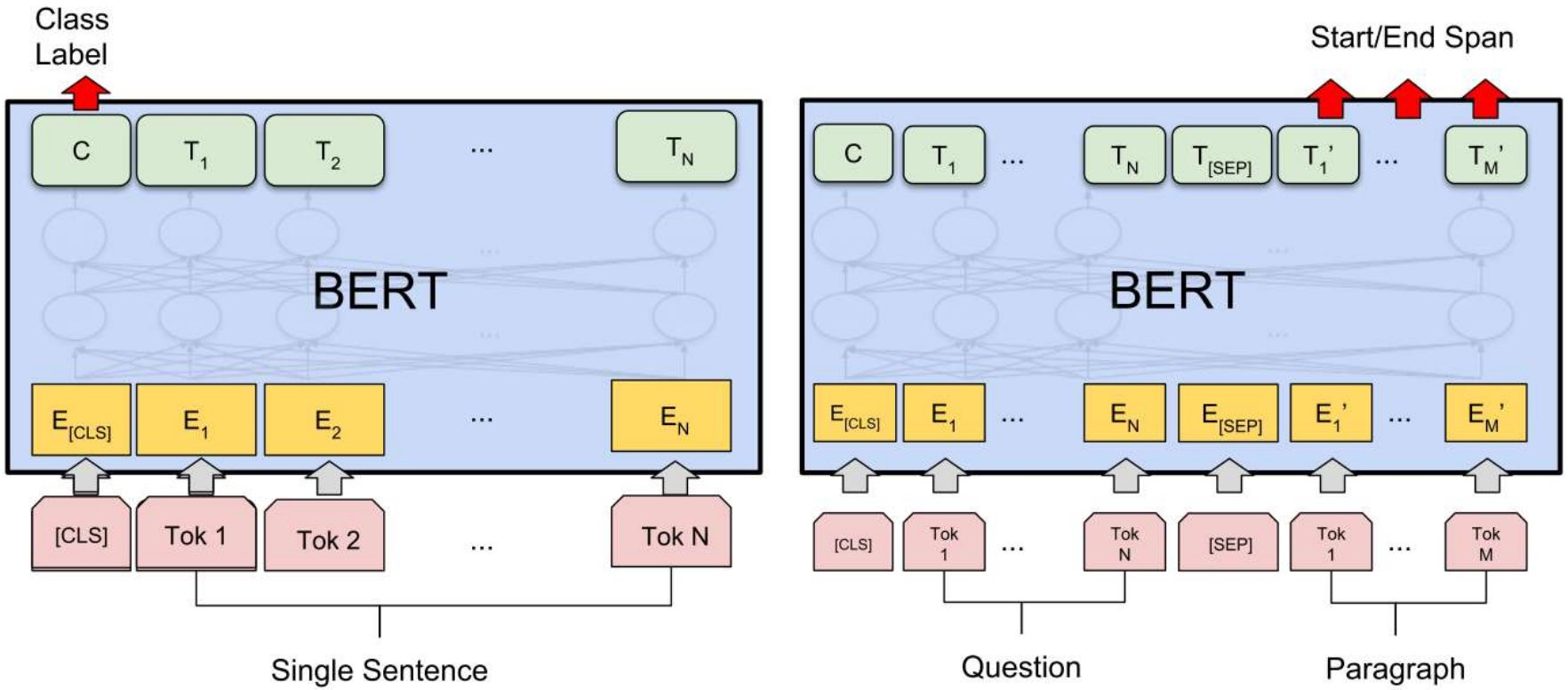
- It achieves state-of-the-art performance in many NLP tasks, sometimes even beating human performance.
- GLUE: a multi-task benchmark for NLU
  - Sentiment, Text similarity, Textual inference, Coreference

Rank	Name	Model	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	T5 Team - Google	T5	89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0	91.7	96.7	92.5	93.2	53.1
2	ALBERT-Team Google Language	ALBERT (Ensemble)	89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8	50.2
3	王玮	ALICE v2 large ensemble	89.0	69.2	97.1	93.6/91.5	92.7/92.3	74.4/90.7	90.7	90.2	99.2	87.3	89.7	47.8
			.....											
8	GLUE Human Baselines	GLUE Human Baselines	87.1	66.4	97.8	86.3/80.8	92.7/92.6	59.5/80.4	92.0	92.8	91.2	93.6	95.9	-

https://gluebenchmark.com/

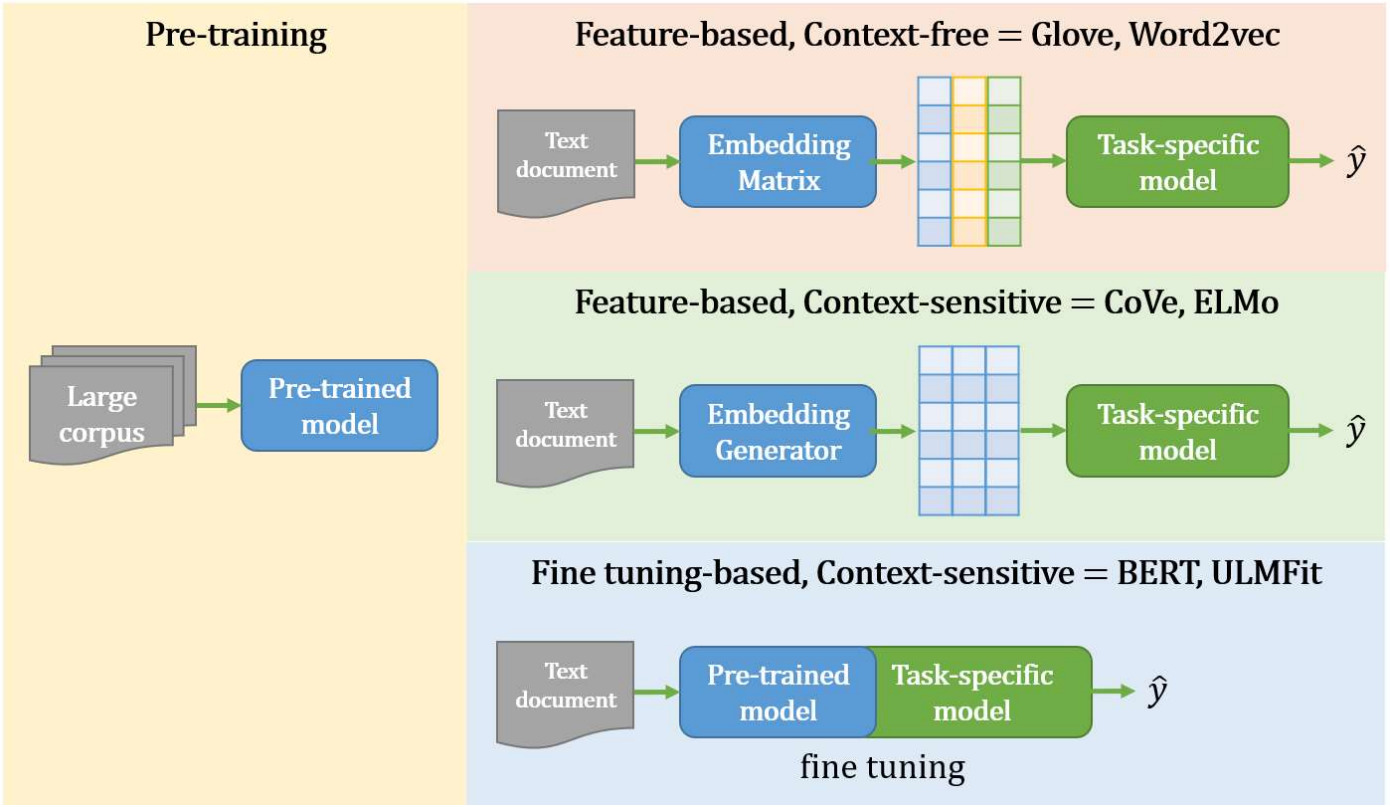
First of all, deep NLP is awesome !

- Requires no feature engineering



First of all, deep NLP is awesome !

- Easy to do transfer learning



BUT

DeepNLP is not perfect

In this talk, I will discuss

- Infamous issues of advanced deep NLP models
  - Bias, Reliability, Consistency
- Techniques to mitigate these issues
- (Optional) Explanation methods to better understand the models

WARNING: Some of the following examples are offensive in nature based on the model outputs.

DeepNLP can have bias

- Example: Gender bias in an abusive language dataset
  - Bias caused by dataset imbalance: Frequently attacked identities are overrepresented in toxic comments

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

Table 1: Frequency of identity terms in toxic comments and overall.

	Comment Length				
Term	20-59	60-179	180-539	540-1619	1620-4859
ALL	17%	12%	7%	5%	5%
gay	88%	77%	51%	30%	19%
queer	75%	83%	45%	56%	0%
homosexual	78%	72%	43%	16%	15%
black	50%	30%	12%	8%	4%
white	20%	24%	16%	12%	2%
wikipedia	39%	20%	14%	11%	7%
atheist	0%	20%	9%	6%	0%
lesbian	33%	50%	42%	21%	0%
feminist	0%	20%	25%	0%	0%
islam	50%	43%	12%	12%	0%
muslim	0%	25%	21%	12%	17%
race	20%	25%	12%	10%	6%
news	0%	1%	4%	3%	3%
daughter	0%	7%	0%	7%	0%

Figure 1: Percent of comments labeled as toxic at each length containing the given terms.

Dixon et al. (2018) Measuring and Mitigating Unintended Bias in Text Classification. AIES.

DeepNLP can have bias

- Example: Gender bias in an abusive language dataset
  - Bias caused by dataset imbalance: Frequently attacked identities are overrepresented in toxic comments

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay		
tran		
lesb		
hom		
fem		
blac		
whi		
hete		
isla		
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

Table 1: Frequency of identity terms in toxic comments and overall.

The bias is then propagated to the trained model (CNN)

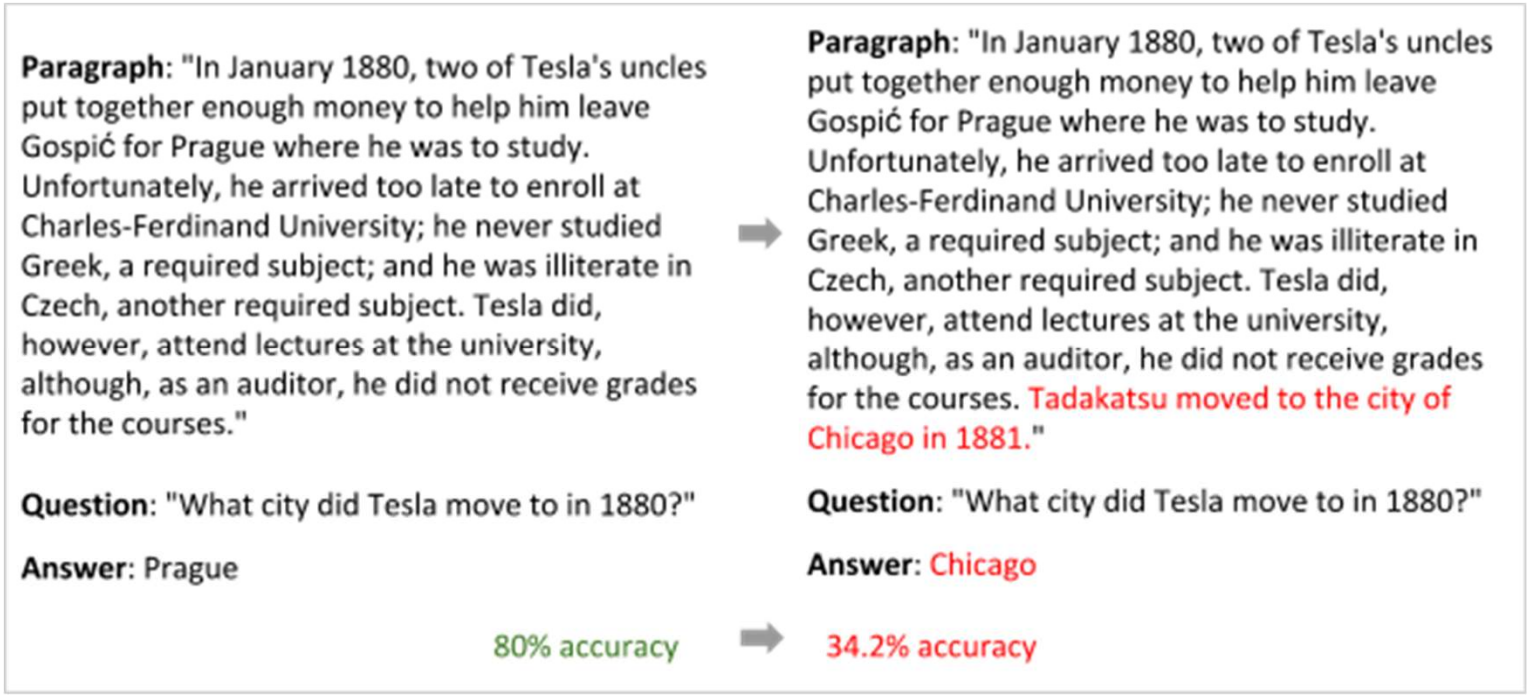
Example:	Sentences	Predicted toxic score
	"i'm a proud **tall** person"	0.18
	"i'm a proud **lesbian** person"	0.51
	"i'm a proud **gay** person"	0.69

	Comment Length				
Term	20-59	60-179	180-539	540-1619	1620-4859
ALL	17%	12%	7%	5%	5%
gay	88%	77%	51%	30%	19%
				6%	0%
				6%	15%
				8%	4%
				2%	2%
				1%	7%
				6%	0%
				1%	0%
				0%	0%
				2%	0%
				2%	17%
				0%	6%
news	0%	1%	4%	3%	3%
daughter	0%	7%	0%	7%	0%

Figure 1: Percent of comments labeled as toxic at each length containing the given terms.

DeepNLP can be unreliable

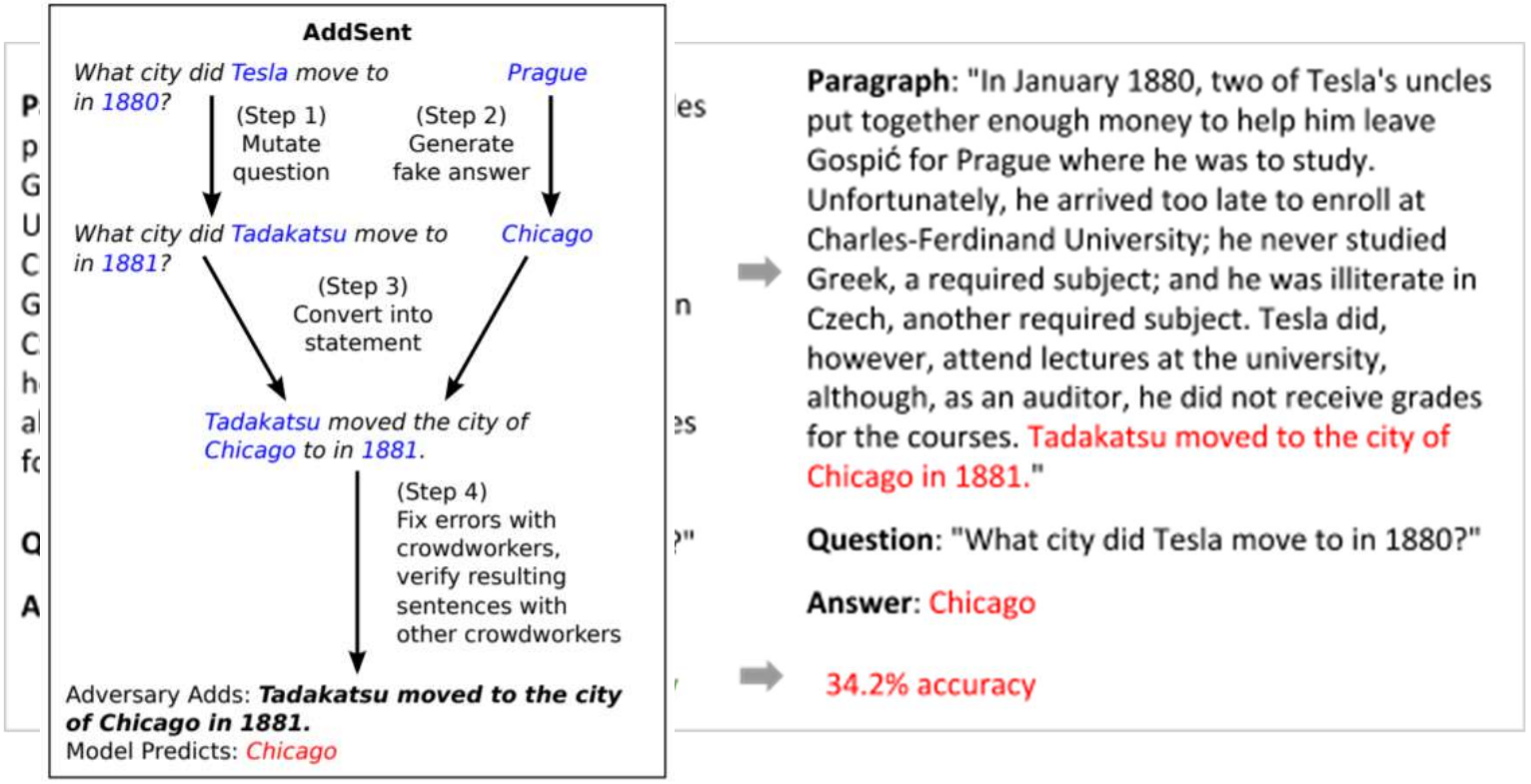
- Reading Comprehension using Bi-Directional Attention Flow (BIDAF) network



Jia and Liang (2017). Adversarial Examples for Evaluating Reading Comprehension Systems. EMNLP

DeepNLP can be unreliable

- Reading Comprehension using Bi-Directional Attention Flow (BIDAF) network



Jia and Liang (2017). Adversarial Examples for Evaluating Reading Comprehension Systems. EMNLP

DeepNLP can be inconsistent

- Natural Language Entailment
  - S_A: John is on a train to Berlin.
  - S_B: John is traveling to Berlin.
  - S_C: John is having lunch in Berlin.

  - Predictions by a decomposable attention model with ELMo
  - S_A Entails S_B ✓
  - S_B Contradicts S_C ✓
  - S_A Neutral S_C ✗

$$(A \rightarrow B) \wedge (B \rightarrow \neg C) \rightarrow (A \rightarrow \neg C)$$

DeepNLP can be inconsistent

How many birds?	**A:** 1
Is there 1 bird?	**A:** no
Are there 2 birds?	**A:** yes
Are there any birds?	**A:** no

(a) Input image from the **VQA dataset**.

(b) Model (Zhang et al., 2018) provides inconsistent answers.

Kublai originally named his eldest son, Zhenjin, as the Crown Prince, but he died before Kublai in 1285.

(c) Excerpt from an input paragraph, **SQuAD dataset**.

Q: When did Zhenjin die?	**A:** 1285
Q: Who died in 1285?	**A:** Kublai

(d) Model (Peters et al., 2018) provides inconsistent answers.

Why these issues exist (in DeepNLP) …

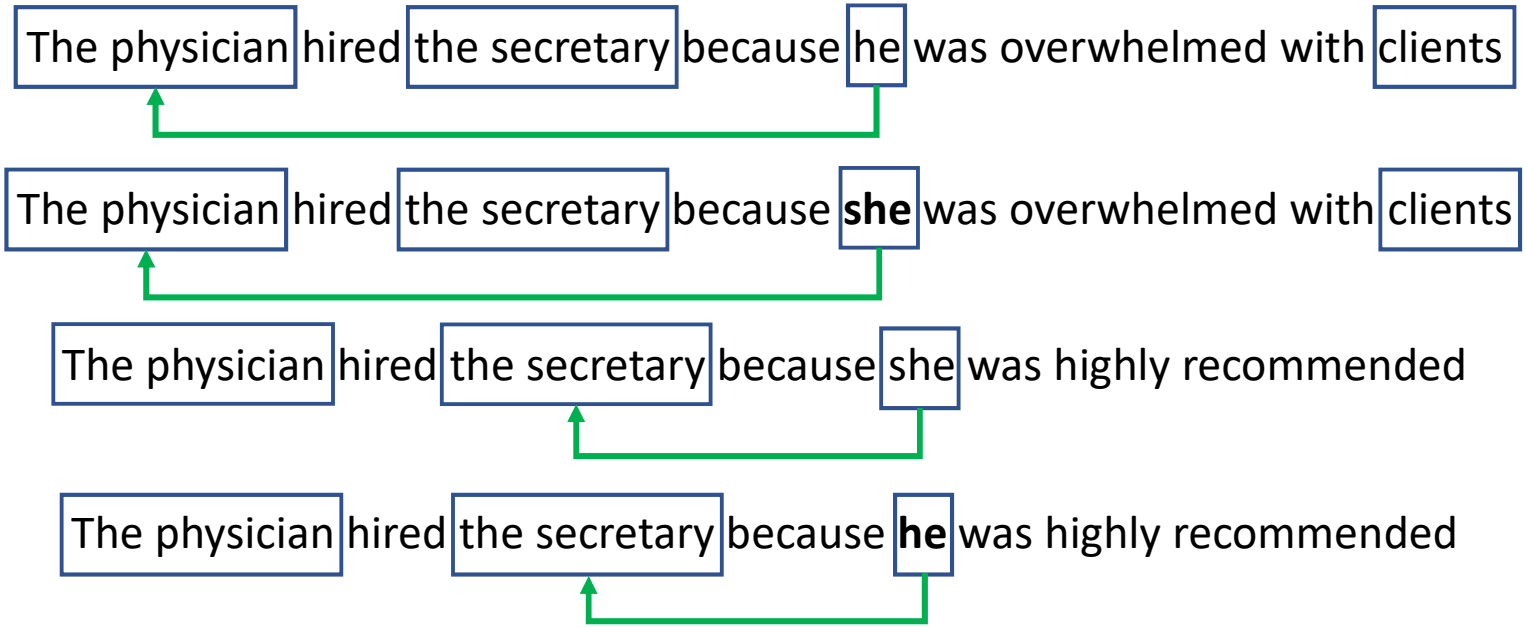
- It's very difficult to obtain training data which is
  - Complete
  - Unbiased
- The models learn only from the labels without reasons
- The models do not know logic nor consistency

To mitigate these issues,

Train the models to have the desirable properties (e.g., unbiased, logical, consistent, reasonable, etc.)

Data Augmentation

- Gender swapping (to tackle gender bias)
  - Identifying male entities and swapping them with equivalent female entities and vice-versa.
  - E.g., for coreference resolution,



Zhao et al. (2018) Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. NAACL

Data Augmentation

- From adversarial examples

Test data	Training data Original	
Original	75.8	75.1
ADDSENT	34.8	70.4
ADDSENTMOD	34.3	39.2

Paragraph: "In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses."

Question: "What city did Tesla move to in 1880?"

Answer: Prague

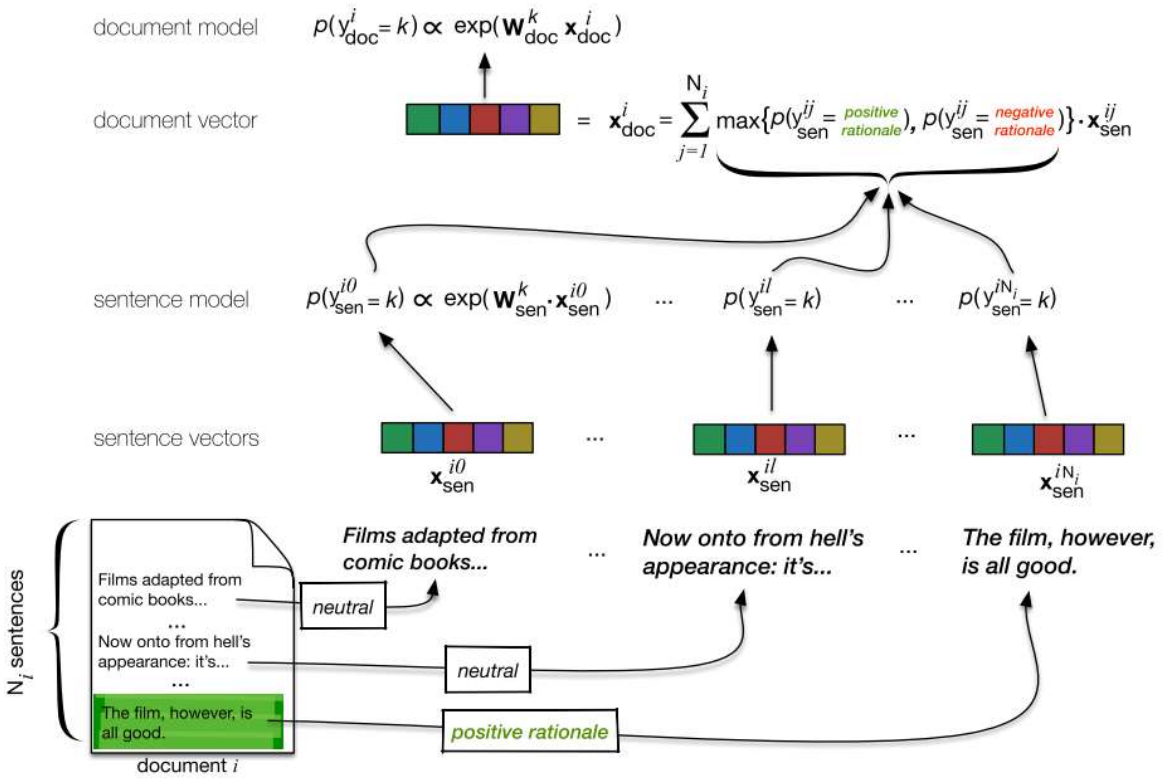
Paragraph: "In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses. Tadakatsu moved to the city of Chicago in 1881."

Question: "What city did Tesla move to in 1880?"

Answer: ~~Chicago~~ Prague

Using Human Rationales

- Rationales are parts of the input which directly contribute to the prediction.



Zhang et al. (2016) Rationale-augmented convolutional neural networks for text classification. EMNLP

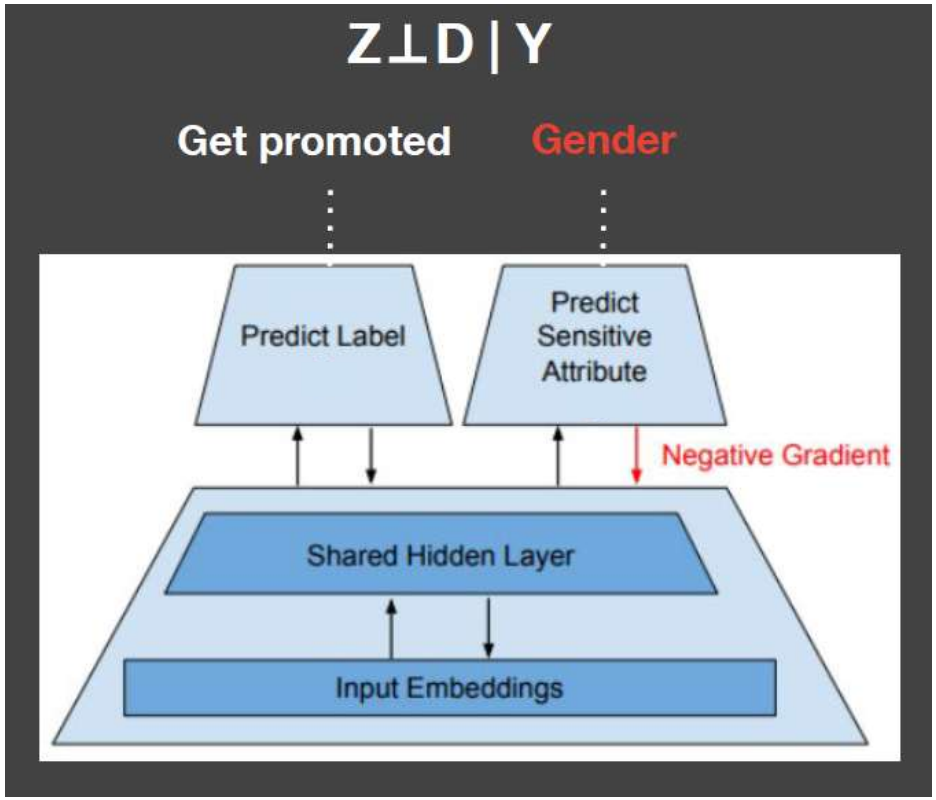
Adding Loss Terms

- In textual inference, we know that if A contradicts B, then B also contradicts A
  - Symmetry: $(A \rightarrow \neg B) \rightarrow (B \rightarrow \neg A)$
- We create an additional loss term
  - $L_{sym} = \sum_{(A,B)\in D} |\log c(A,B) - \log c(B,A)|$
- How about transitivity? $(A \rightarrow B) \wedge (B \rightarrow \neg C) \rightarrow (A \rightarrow \neg C)$
  - Using soft logic to convert it to numerical loss

Name	Boolean Logic	Product	Gödel	Łukasiewicz
Negation	$\neg A$	$1-a$	$1-a$	$1-a$
T-norm	$A \wedge B$	ab	$\min(a,b)$	$\max(0, a+b-1)$
T-conorm	$A \vee B$	$a+b-ab$	$\max(a,b)$	$\min(1, a+b)$
Residuum	$A \rightarrow B$	$\min\left(1, \frac{b}{a}\right)$	$\begin{cases} 1, & \text{if b} \geq \text{a}, \\ b, & \text{else} \end{cases}$	$\min(1, 1-a+b)$

Adversarial Training

- Using the same representation of the input (Y), jointly predict
  - The desired output (Z)
  - The attribute you want to remove (D)



Zhang et al (2018). Mitigating Unwanted Biases with Adversarial Learning. AIES

Other techniques

- Using debiased word embeddings

- Constraining predictions
  - restricted ratio of males to females predicted to be positive to prevent the model from amplifying bias through predictions

- Bias Fine-tuning
  - train on another non-bias dataset first and then fine-tune the last layer on the biased dataset

Zhao et al. (2017) Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints. EMNLP
Park et al. (2018) Reducing Gender Bias in Abusive Language Detection. EMNLP

Takeaways

- Deep NLP models trained using only labelled data are prone to be biased, unreliable, inconsistent, etc.

- Why?
  - The training data is biased and incomplete.
  - The model does not know the true reasons of the predictions.
  - The model is not trained to know logic and consistency.

- How to mitigate these issues?
  - Teach what the model should know or behave using augmented data, more information, or loss terms.

Thank you
&
Questions

Piyawat Lertvittayakumjorn

pl1515@imperial.ac.uk