

Paraphrase generation: adversarial examples / data augmentation

CS 685, Fall 2020

Advanced Natural Language Processing

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

stuff from last time...

- HW1 released, start early!
- Exam will be Nov 5-6

adversarial examples



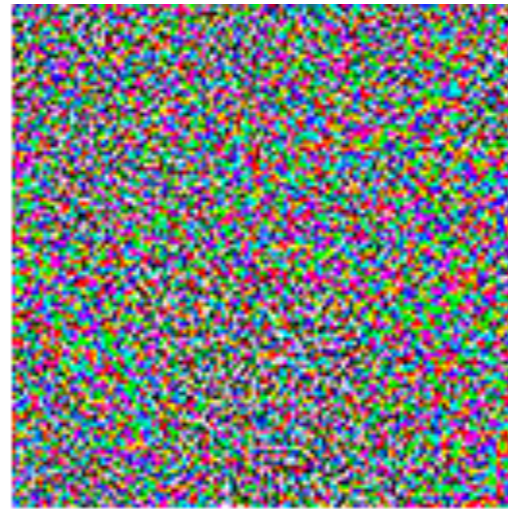
panda

57.7% confidence

adversarial examples



+ ϵ



=



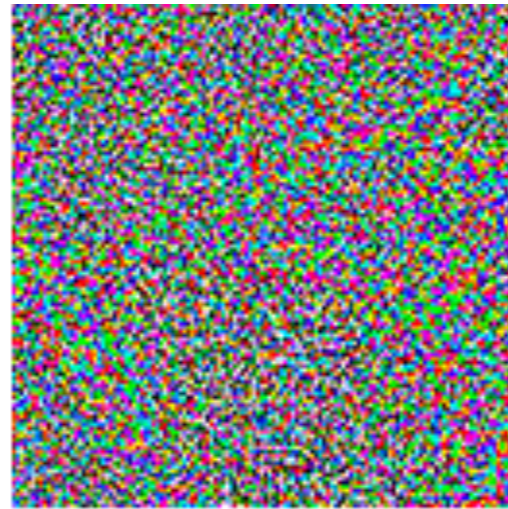
panda
57.7% confidence

gibbon
99.3% confidence

adversarial examples



+ ϵ



=

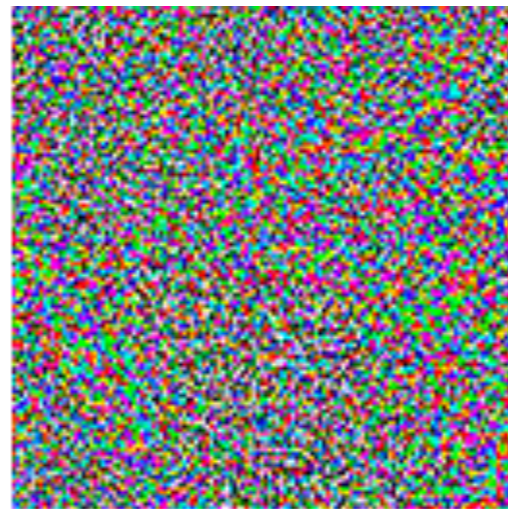


panda
57.7% confidence

gibbon
99.3% confidence

The movie was
very bad.

+ ϵ



=

???

Textual Entailment is the task of predicting whether, for a pair of sentences, the facts in the first sentence necessarily imply the facts in the second.

Premise

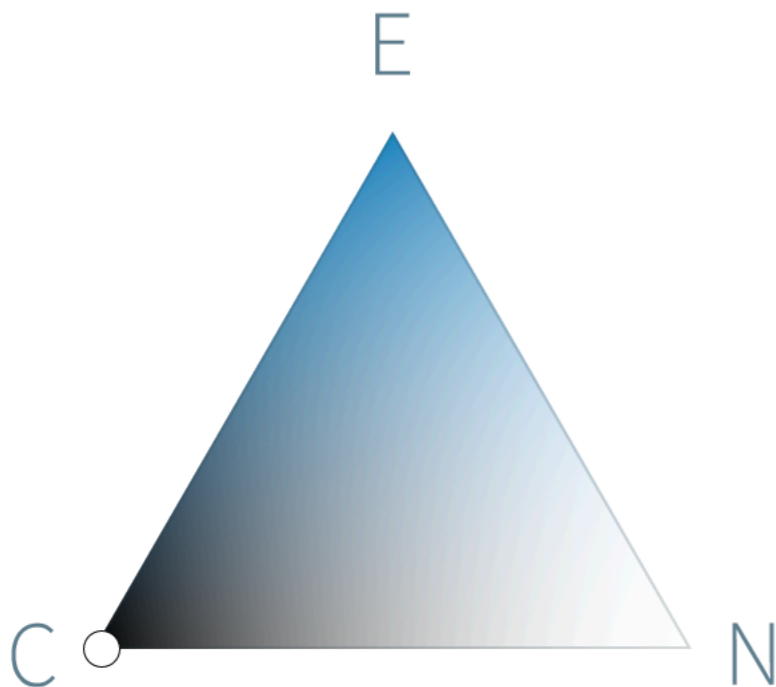
Two women are wandering along the shore drinking iced tea.

Hypothesis

Two women are sitting on a blanket near some rocks talking about politics.

Summary

It is **very likely** that the premise **contradicts** the hypothesis.



Judgment	Probability
Entailment	0%
Contradiction	98.8%
Neutral	1.2%

Premise

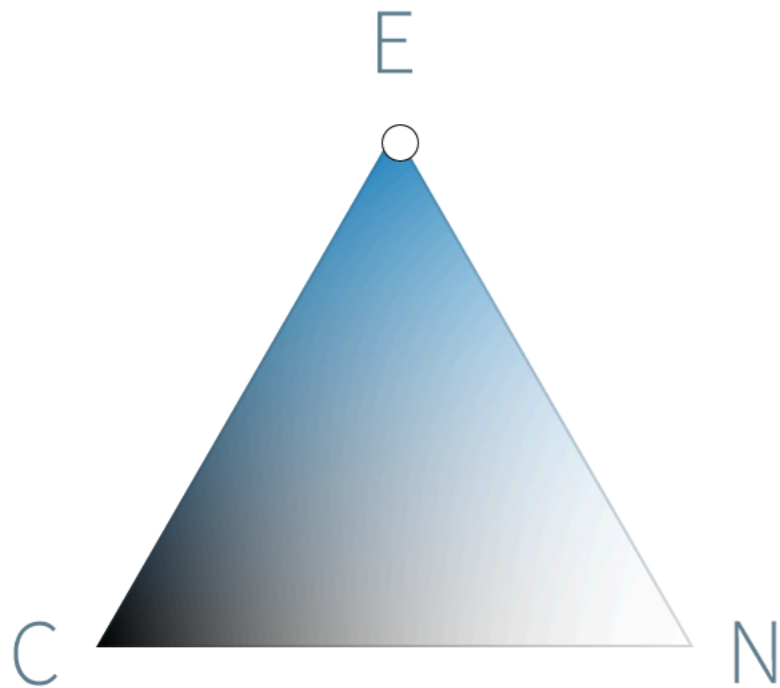
the dog ate all of the chickens

Hypothesis

chickens

Summary

It is **very likely** that the premise **entails** the hypothesis.



Judgment	Probability
Entailment	97.6%
Contradiction	0.3%
Neutral	2.1%

Premise

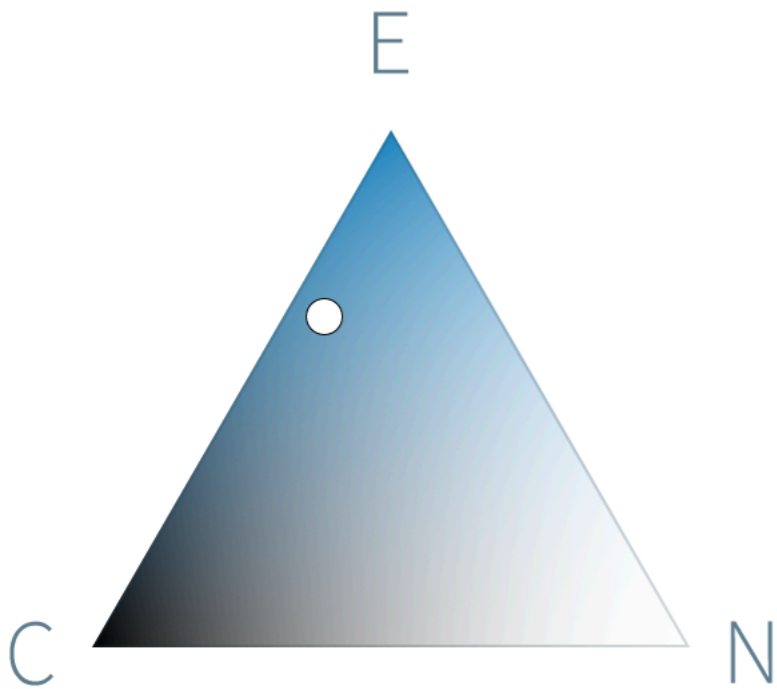
the red box is in the blue box

Hypothesis

red is blue

Summary

It is **somewhat likely** that the premise **entails** the hypothesis.



Judgment	Probability
Entailment	64%
Contradiction	29.2%
Neutral	6.7%

adversarial examples for NLP

the *build-it-break-it* workshop at EMNLP 2017
challenged humans to “break” existing systems by
coming up with linguistically-adversarial examples

“iid development data is unlikely to exhibit all the linguistic phenomena that we might be interested in testing”

“NLP systems are quite brittle in the face of infrequent linguistic phenomena, a characteristic which stands in stark contrast to human language users.”

lexical adversaries

create by word replacement
using thesaurus, WordNet,
word embedding similarity

(e.g., Jia et al., ACL 2017)

Input sentence

Exactly the kind of *unexpected delight* one
hopes for every time the lights go down

Exactly the kind of *thrill* one hopes
for every time the lights go down

Model prediction

positive

negative

syntactic adversaries

Input sentence

American drama doesn't get any more
meaty and muscular than this.

Doesn't get any more meaty and muscular
than this American drama.

Model prediction

positive

negative

how do we automatically
create such examples?
can we use a *paraphrase*
generation system?

an ideal syntactic paraphraser...

- produces grammatically-correct paraphrases that retain the meaning of the original sentence
- minimizes lexical differences between input sentence and paraphrase
- generates many diverse syntactic paraphrases from the same input

syntactic paraphrase generation

Usually you require inventory only when you plan to sell your assets .

example paraphrases


1. Usually, you required the inventory only if you were planning to sell the assets.
2. When you plan to sell your assets, you usually require inventory.
3. You need inventory when you plan to sell your assets.
4. Do the inventory when you plan to sell your assets.

syntactic paraphrase generation

Usually you require inventory only when you plan to sell your assets .

example paraphrases

1. Usually, you required the inventory only if you were planning to sell the assets.
2. When you plan to sell your assets, you usually require inventory.
3. You need inventory when you plan to sell your assets.
4. Do the inventory when you plan to sell your assets.



grammatical
preserve input semantics
minimal lexical substitution
high syntactic diversity

Long history of work on paraphrasing!

- *rule / template-based* syntactic paraphrasing
(e.g., McKeown, 1983; Carl et al., 2005)
 - high grammaticality, but very low diversity
- *translation-based* uncontrolled paraphrasing that rely on parallel text to apply machine translation methods
(e.g., Bannard & Callison-Burch, 2005; Quirk et al., 2004)
 - high diversity, but low grammaticality and no syntactic control
- *deep learning-based* controlled language generation with conditional encoder/decoder architectures
(e.g., Fidler & Goldberg, 2017; Shen et al., 2017)
 - grammatical, but low diversity and no paraphrase constraint

syntactically controlled paraphrase networks (SCPNS)

1. acquire millions of sentential paraphrase pairs through neural backtranslation
2. automatically label these pairs with descriptive syntactic transformations
3. train a supervised encoder/decoder model on this labeled data to produce a paraphrase given the original sentence and a target syntactic form

training data via *backtranslation*

isn't that more a topic for your priest ?



translate to Czech

není to více téma pro tvého kněze?



translate back to English

are you sure that's not a topic for you to discuss with your priest ?

training data via *backtranslation*

isn't that more a topic for your priest ?



translate to Czech

není to více téma pro tvého kněze?



translate back to English

are you sure that's not a topic for you to discuss with your priest ?

backtranslate the CzEng parallel corpus (Bojar et al., 2016) using a state-of-the-art NMT system, which yields ~50 million paraphrase pairs

through neural backtranslation, we can
generate *uncontrolled* paraphrases.

how can we achieve syntactic control?

labeling paraphrase pairs with descriptive syntactic transformations

- first experiment: *rule-based* labels
 - She drives home. She is driven home. active > passive
- Easy to write these rules, but low syntactic variance between the paraphrase pairs

using linearized syntactic parses as labels

S₁ isn't that more a topic for your priest ?

p₁

```
( ROOT ( S ( VP ( VBZ ) ( RB ) ( SBARQ ( IN ) ( NP ( NP ( JJR ) ( NP ( NP ( DT ) ( NN ) ) ( PP ( IN ) ( NP ( PRP$ ) ( NN ) ) ) ) ) ) ) ) ) ) ( . ) )
```

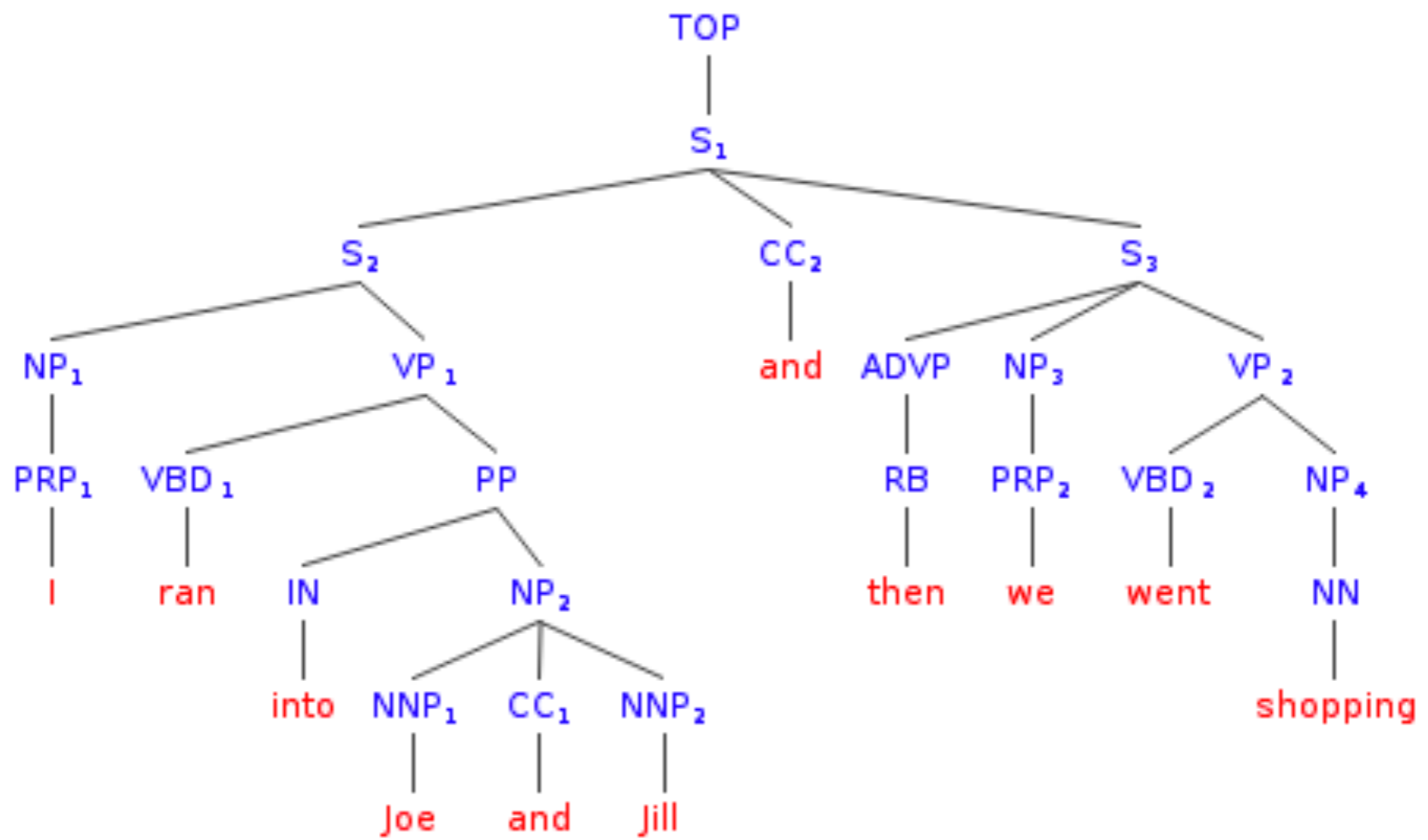


S₂

are you sure that's not a topic for you to discuss with your priest ?

p₂

```
( ROOT ( SBARQ ( SQ ( VBP ) ( NP ( PRP ) ) ( ADJP ( JJ ) ( SBAR ( S ( NP ( DT ) ) ( VP ( VBZ ) ( RB ) ( NP ( DT ) ( NN ) ) ( SBAR ( IN ) ( S ( NP ( PRP ) ) ( VP ( TO ) ( VP ( VB ) ( PRT ( RP ) ) ( PP ( IN ) ( NP ( PRP$ ) ( NN ) ) ) ) ) ) ) ) ) ) ) ) ) ) ) ( . ) )
```

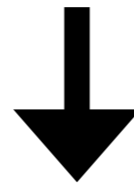


input to our model

S₁ isn't that more a topic for your priest ?

p₂

```
(ROOT (SBARQ (SQ (VBP) (NP (PRP)) (ADJP (JJ) (SBAR (S (NP (DT)) (VP (VBZ) (RB) (NP (DT) (NN)) (SBAR (IN) (S (NP (PRP)) (VP (TO) (VP (VB) (PRT (RP)) (PP (IN) (NP (PRP$) (NN))))))))))))) (.)))
```



S₂

are you sure that's not a topic for you to discuss with your priest ?

SCPN architecture

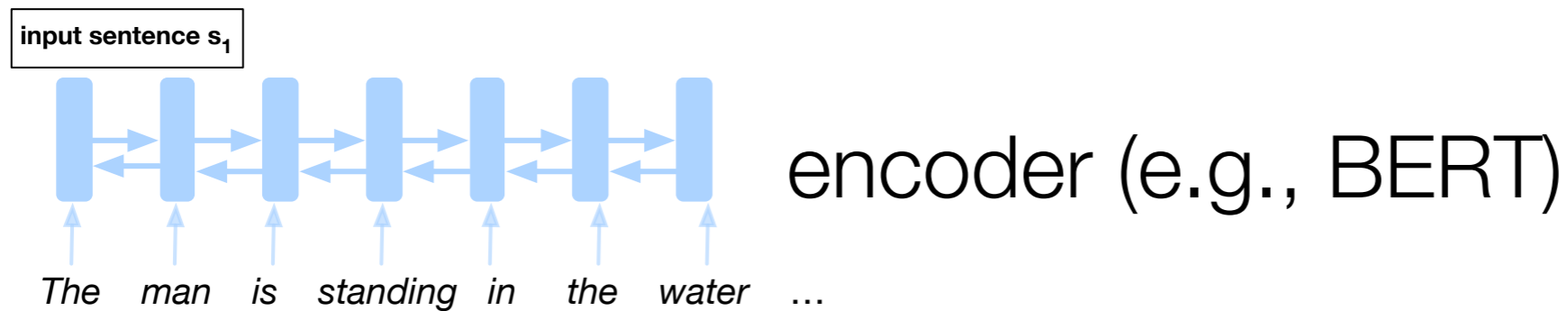
The man is standing in the water at the base of a waterfall

The man, at the base of the waterfall, is standing in the water

SCPN architecture

The man is standing in the water at the base of a waterfall

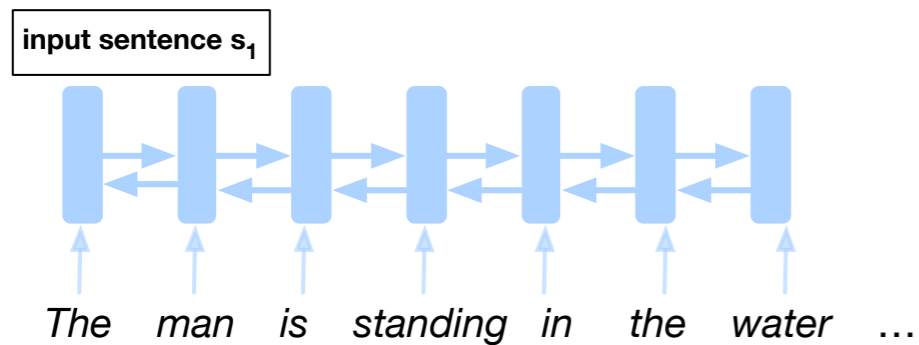
The man, at the base of the waterfall, is standing in the water



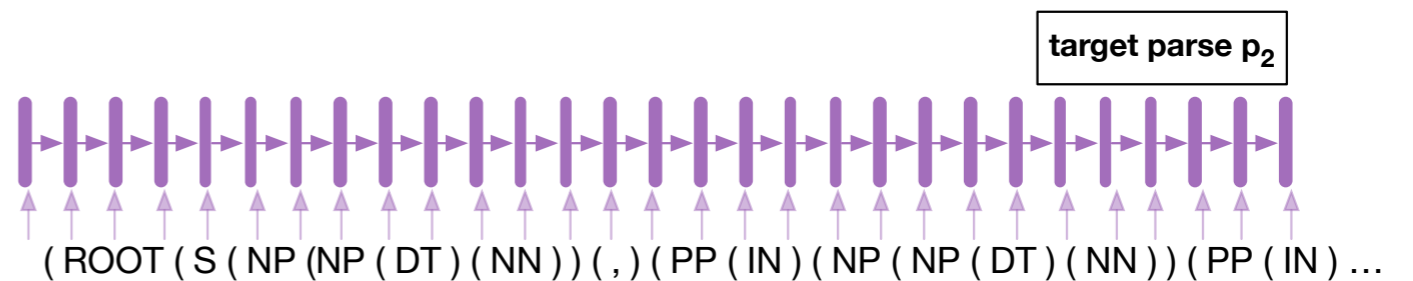
SCPN architecture

The man is standing in the water at the base of a waterfall

The man, at the base of the waterfall, is standing in the water



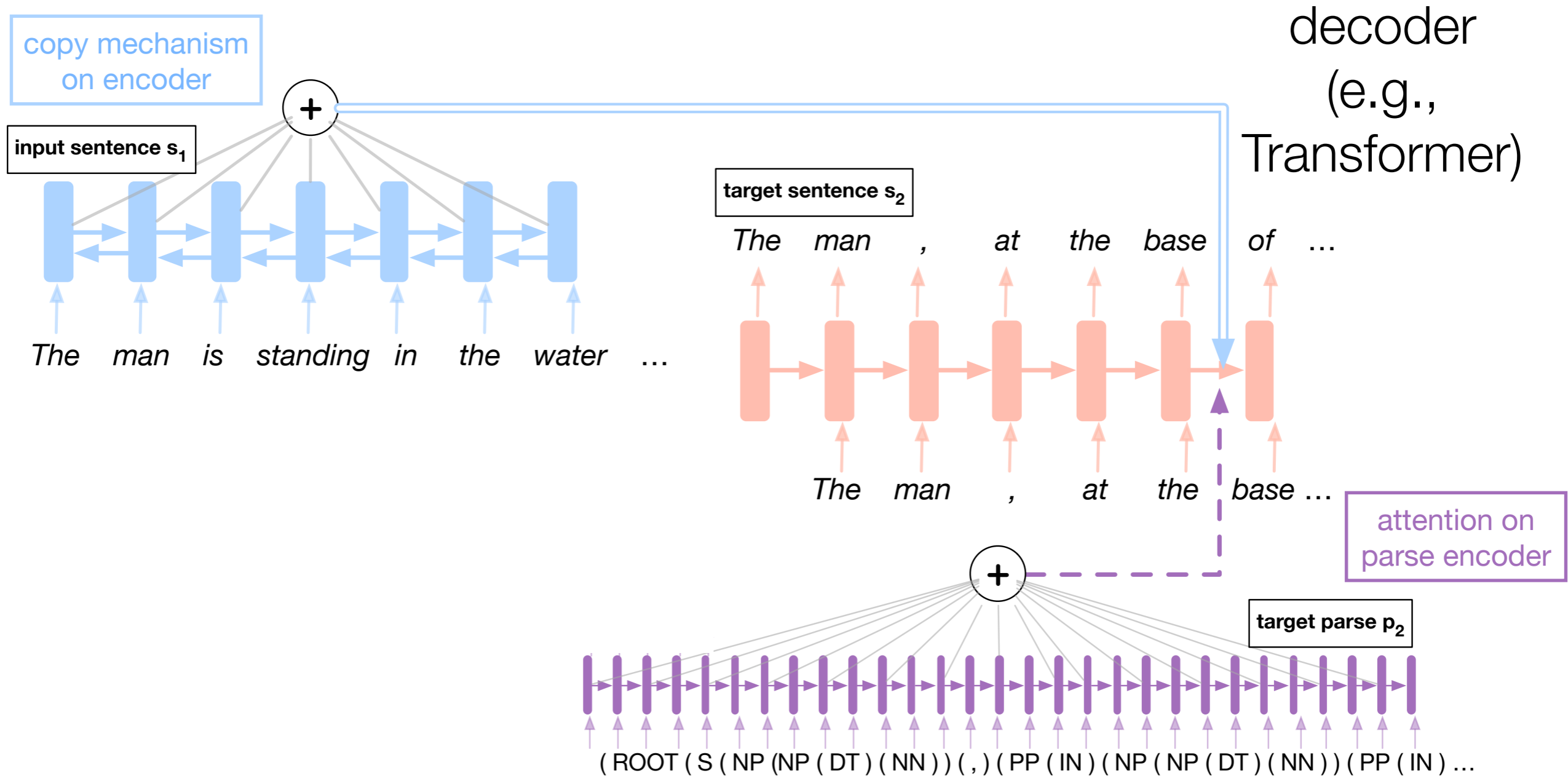
parse encoder (fine-tuned BERT?)



SCPN architecture

The man is standing in the water at the base of a waterfall

The man, at the base of the waterfall, is standing in the water



specifying a full target parse is unwieldy

we use the top two levels of the **linearized
parse tree** as a **parse template**

She drove home.

(S (NP (PRP)) (VP (VBD) (NP (NN)))) (.)

template: $S \rightarrow NP VP .$

paraphrase quality

- crowdsourced task, workers rate a paraphrase pair on a three point scale (Kok and Brockett, 2010)
 - 0 = no paraphrase
 - 1 = ungrammatical paraphrase
 - 2 = grammatical paraphrase

paraphrase quality

- crowdsourced task, workers rate a paraphrase pair on a three point scale (Kok and Brockett, 2010)

0 = no paraphrase

1 = ungrammatical paraphrase

2 = grammatical paraphrase

Model	2	1	0
SCPN w/ full parses	63.7	14.0	22.3
SCPN w/ templates	62.3	19.3	18.3
NMT-BT	65.0	17.3	17.7

} no significant quality loss despite adding syntactic control

adversarial evaluations

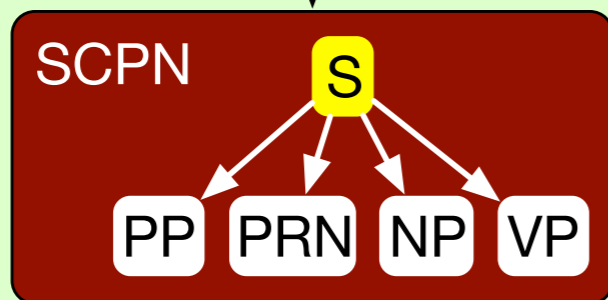
- how many held-out examples can we “break”?
 - a development example x is “broken” if the original prediction y_x is correct, but the prediction y_{x^*} for at least one paraphrase x^* is incorrect.
- this is only a valid measure if the paraphrase that breaks x actually has the same label as x
 - we conduct a crowdsourced evaluation to determine if the adversarial examples actually preserve the original label

two tasks

- sentiment analysis (Stanford Sentiment Treebank)
 - binary classification of sentences (0 = negative, 1 = positive)
 - many long sentences with high syntactic variance
- textual entailment (SICK)
 - 3-way classification of sentence pairs (0 = contradiction, 1 = neutral, 2 = entailment)
 - almost exclusively short, simple sentences

I'd have to say the star
and director are the big
problems here

negative

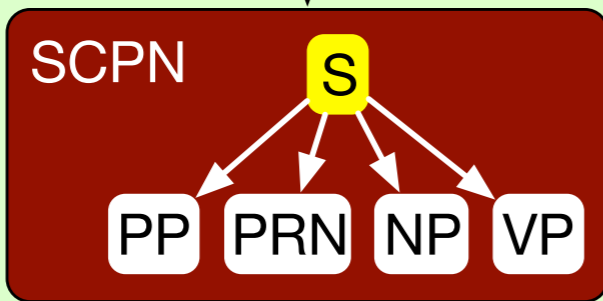


By the way, you know, the
star and director are the
big problems

~~positive~~

I'd have to say the star and director are the big problems here

negative



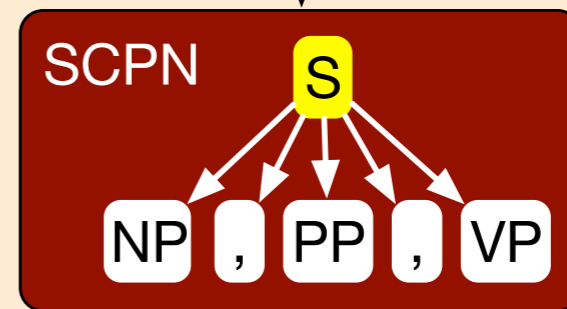
By the way, you know, the star and director are the big problems

~~positive~~

The man is standing in the water at the base of a waterfall

A man is standing in the water at the base of a waterfall

entailment



The man, at the base of the waterfall, is standing in the water

A man is standing in the water at the base of a waterfall

~~neutral~~

SCPN VS NMT

sentiment analysis

Model	Validity	% Dev Broken
SCPN	77.1	41.8
NMT-BT	68.1	20.2

SCPN VS NMT

sentiment analysis

Model	Validity	% Dev Broken
SCPN	77.1	41.8
NMT-BT	68.1	20.2

textual entailment

Model	Validity	% Dev Broken
SCPN	77.7	33.8
NMT-BT	81.0	20.4

improving robustness to adversaries

when we augment the training data with **SCPN** paraphrases, we are able to decrease the proportion of “broken” development examples without decreasing performance on original test data

sentiment analysis

Model	No augmentation		With augmentation	
	Test Acc	% dev broken	Test Acc	% dev broken
SCPN	83.1	41.8	83.0	31.4
NMT-BT	83.1	20.2	82.3	20.0

improving robustness to adversaries

when we augment the training data with **SCPN** paraphrases, we are able to decrease the proportion of “broken” development examples without decreasing performance on original test data

sentiment analysis

Model	No augmentation		With augmentation	
	Test Acc	% dev broken	Test Acc	% dev broken
SCPN	83.1	41.8	83.0	31.4
NMT-BT	83.1	20.2	82.3	20.0

textual entailment

Model	No augmentation		With augmentation	
	Test Acc	% dev broken	Test Acc	% dev broken
SCPN	82.1	33.8	82.7	19.8
NMT-BT	82.1	20.4	82.0	11.2

syntactic manipulation examples

Template

Paraphrase

GOLD

you seem to be an excellent burglar when the time comes.

(S (SBAR) (,) (NP) (VP))

when the time comes, you'll be a great thief.

(S (") (UCP) (") (NP) (VP))

"you seem to be a great burglar, when the time comes", you said.

(SQ (MD) (SBARQ))

can i get a good burglar when the time comes?

(S (NP) (IN) (NP) (NP) (VP))

look at the time the thief comes.

syntactic manipulation examples

Template

Paraphrase

GOLD

with the help of captain picard, the borg will be prepared for everything.

(SBARQ (ADVP) (,) (S) (,) (SQ))

now, the borg will be prepared by picard, will it?

(S (NP) (ADVP) (VP))

the borg here will be prepared for everything.

(S (S) (,) (CC) (S) (:) (FRAG))

with the help of captain picard, the borg will be prepared, and the borg will be prepared for everything... for everything.

(FRAG (INTJ) (,) (S) (,) (NP))

oh, come on captain picard, the borg line for everything.

SCPN adversarial sentiment examples

Template	Original	Paraphrase
(S (ADVP) (NP) (VP))	moody, heartbreaking, and filmed in a natural, unforced style that makes its characters seem entirely convincing even when its script is not.	so he's filmed in a natural, unforced style that makes his characters seem convincing when his script is not.
(S (PP) (,) (NP) (VP))	there is no pleasure in watching a child suffer.	in watching the child suffer, there is no pleasure.
(S (S) (,) (CC) (S))	the characters are interesting and often very creatively constructed from figure to backstory .	the characters are interesting, and they are often built from memory to backstory.

NMT adversarial sentiment examples

Original

Paraphrase

every nanosecond of the new guy reminds you that you could be doing something else far more **pleasurable**.

each nanosecond from the new guy reminds you that you could do something else much more **enjoyable**.

harris commands the screen, using his **frailty** to suggest the ravages of a life of corruption and **ruthlessness**.

harris commands the screen, using his **weakness** to suggest the ravages of life of corruption and **recklessness** .

Can we perform *style transfer* using
paraphrase generation models?

Style transfer: given an input sentence, modify its stylistic properties while preserving its semantics

“Style” is impossible to precisely define, and in some fields (e.g., sociolinguistics) it’s considered inseparable from semantics.

Here, we’ll consider “style” to loosely represent lexical and syntactic choice.

Shakespeare

"To be, or not to be: that is the question:
Whether 'tis nobler in the mind to suffer
The slings and arrows of outrageous fortune,
Or to take arms against a sea of troubles,
And by opposing end them. To die: to sleep..."

Twitter

Are yall okay? Like do you need my help??
I dont wanna talk to him abt that
Bron haters in shambles they want him to
retire so bad Imfaoooo

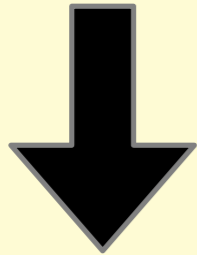
style transfer applications

- Data augmentation (see HW1 :)
- Text simplification
- Writing assistance
- Author obfuscation
- Adversarial example generation
- Components in automatic evaluations for text generation systems

Style transfer via paraphrasing (STRAP)

1. collect datasets of sentences from different styles (e.g., crawl Twitter, Project Gutenberg, etc)
2. generate a paraphrase for each sentence in these datasets by leveraging neural backtranslation
3. fine-tune a large-scale pretrained LM (e.g., GPT2) to perform the task of *inverse paraphrasing* for each style

**Why, uncle,
'tis a shame**

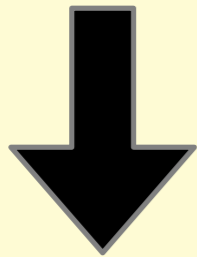


**it's a shame,
uncle**

Step 1:
*diverse
paraphrasing*

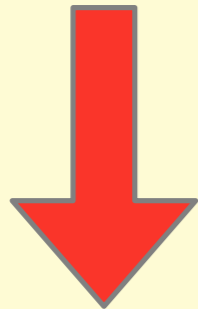
Use an uncontrolled
paraphraser trained on
backtranslated data
(fine-tuned LM #1)

Why, uncle,
'tis a shame



Step 1:
*diverse
paraphrasing*

it's a shame,
uncle



Step 2:
*inverse
paraphrasing
(Shakespeare,
Twitter)*

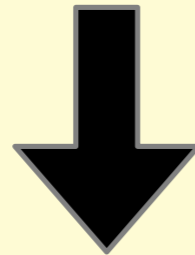
Why, uncle,
'tis a shame

Use an uncontrolled
paraphraser trained on
backtranslated data
(fine-tuned LM #1)

Train *inverse
paraphraser* to
reconstruct the
original sentence
(fine-tuned LM #2)

Training time

Why, uncle,
'tis a shame



it's a shame,
uncle

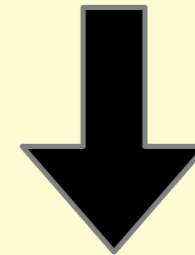


Why, uncle,
'tis a shame

Step 1:
*diverse
paraphrasing*

Step 2:
*inverse
paraphrasing
(Shakespeare,
Twitter)*

No lie... I would
jump in



I'd jump in there,
no doubt

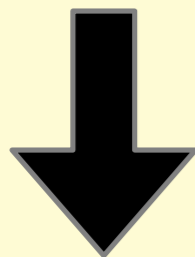


No lie... I would
jump in

At test-time, switch out a different style's inverse paraphraser to perform style transfer

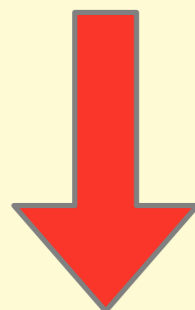
Training time

Why, uncle,
'tis a shame



*Step 1:
diverse
paraphrasing*

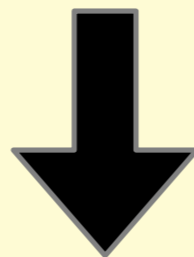
it's a shame,
uncle



*Step 2:
inverse
paraphrasing
(Shakespeare,
Twitter)*

Why, uncle,
'tis a shame

No lie... I would
jump in



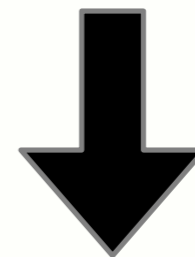
I'd jump in there,
no doubt



No lie... I would
jump in

Test time

O, wilt thou leave
me so unsatisfied?



Oh, you're gonna leave
me unsatisfied, right?



Ooh yall will leave
me unhappy lol

Direction	Input	Output
Tweet → Shak.	yall kissing before marriage?	And you kiss'd before your nuptial?
Lyrics → AAE	It's a good thing you don't have bus fare	It's a good thing u aint gettin no ticket
Tweet → Poetry	Its so disrespectful I hate ittttt	For 'tis so rude, I have a sick regard,
Shak. → Tweet	Need you my help?	Are yall okay? Like do you need my help??
AAE → Shak.	Aint got nooooo holiday spirit frfr	No spirit of this holiday, gentle man.
Poetry → Bible	For the foul thief is just at your gate.	the wicked thief art at thy door.
1990. → Tweet	Now I can't talk to him about it.	I dont wanna talk to him abt that
1990. → Poetry	I don't know why, but I sensed there was something wrong.	Something felt wrong; I knew not why,
Tweet → Bible	when ur going thru an episode and cant cope anymore	when thou art in an episode, and cannot stand it any more;
1890. → AAE	I was just thinking the same thing that you have expressed.	u said the same thing i thought lol
1990. → 1810.	He was being terrorized into making a statement by the same means as the other so-called "witnesses."	Terror had been employed in the same manner with the other witnesses, to compel him to make a declaration.
AAE → Shak.	If I got a dollar every time one of my friends told me they hate me, I'd be rich	I would have been rich, had I but a dollar for every friend that hath said they hate me.
Joyce → Bible	I appeal for clemency in the name of the most sacred word our vocal organs have ever been called upon	I beseech thee in the name of the most holy word which is in our lips, forgive us our trespasses.