Lecture 18: Adversarial NLP

"If it's not broken …", well, you're probably wrong. It's broken.

Harvard IACS

Chris Tanner

ADVERSARIAL NLP

ANNOUNCEMENTS

- HW4 is due tonight @ 11:59
- HW3 and Quiz 6 and Phase 3 are being graded
- Research Project Phase 4 will be a soft-assessment

Many of today's slides are based on, inspired by, or directly from Jack Morris (Cornell Tech), Mohit Iyyer (UMass Amherst), Graham Neubig (CMU)

Outline

- Introduction
- Paraphrasing
- Workshop time
- Modern approaches

Outline

- Introduction
- Paraphrasing
- Workshop time
- Modern approaches

Motivation

+

=

Classified as a
stop sign

Classified as a
70 mph speed limit sign

Goodfellow et al., 2014

Motivation

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

Motivation

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

Does this count?

Example

Could add random noise at the character level.

Input, x	Perturbation, x'
"True Grit" was the best movie I've seen since I was a small boy.	"True Grit" was the best **moive** I've seen **snice** I was a small boy.
Prediction: positive	Prediction: negative

Example

Could add random noise at the character level.

Input, x

"True Grit" was the best movie I've seen since I was a small boy.

Prediction: positive

Perturbation, x'

"True Grit" was the best **moive** I've seen **snice** I was a small boy.

Prediction: negative

This is easy to defend against, right? How?

Example

Input, x

Hi Enrique,

Did you get the photos that I sent from our hangout?

Prediction: not spam

Perturbation, x'

Hi Enrique,

Did u get the photoz that I sent from our hangout?

Prediction: spam

Example

Could

- train an RNN to identify and correct typos
- use a spellchecker to auto-correct the input

Adversarial perturbations can be useful for augmenting training data

Pruthi et al., Combating Adversarial Misspellings with Robust Word Recognition (2019)

Example

Could replace at the word level

Input, x

"True Grit" was the best movie I've seen since I was a <u>small boy</u>.

Prediction: positive

Perturbation, x'

"True Grit" was the best movie I've seen since I was a <u>tiny lad</u>.

Prediction: negative

Entailment

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.

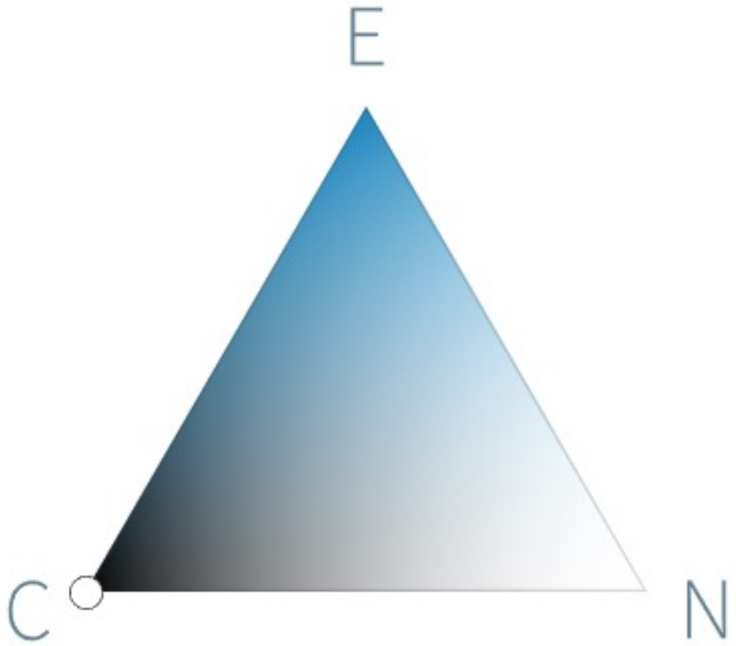
Premise

Two women are wandering along the shore drinking iced tea.

Hypothesis

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

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



Judgement	Probability
Entailment	0%
Contradiction	99.8%
Neutral	0.2%

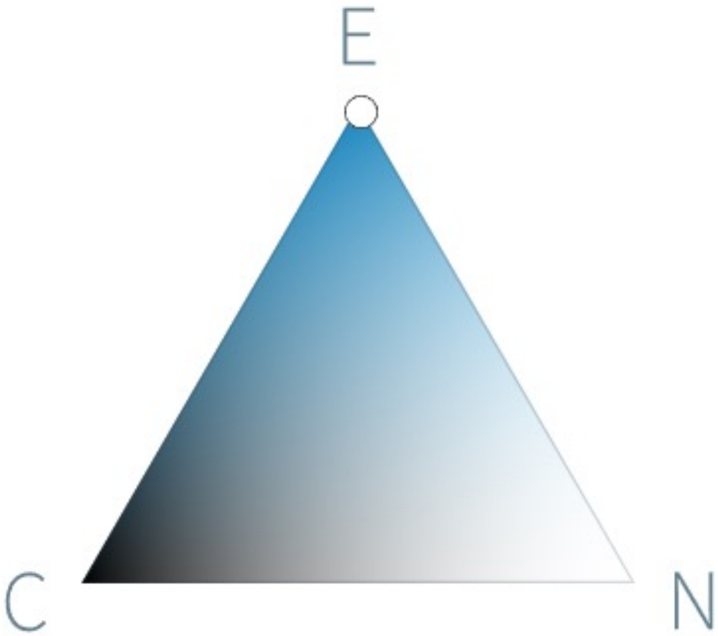
Premise

The dog ate all of the chickens

Hypothesis

chickens

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



Judgement	Probability
Entailment	98.6%
Contradiction	0.1%
Neutral	1.3%

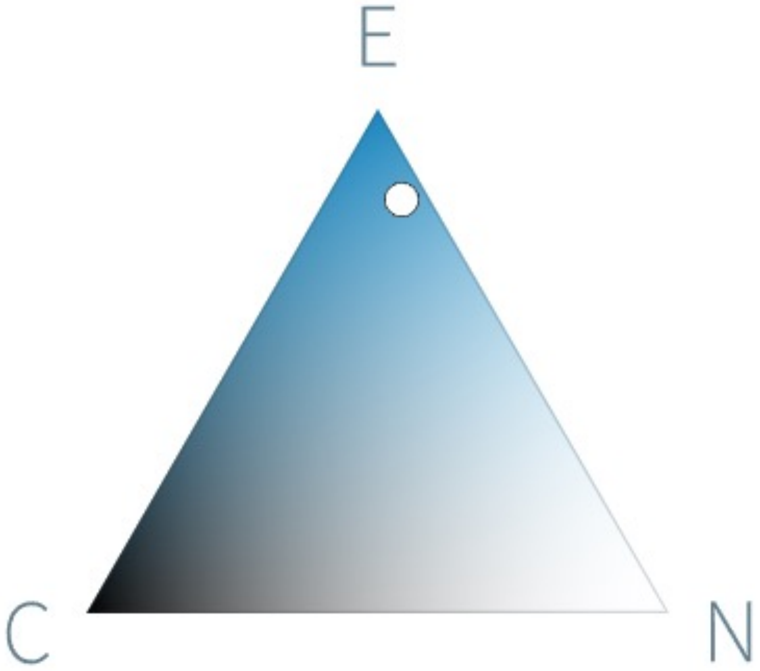
Premise

The red box is in the blue box

Hypothesis

red is blue

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



Judgement	Probability
Entailment	81.8%
Contradiction	5%
Neutral	13.1%

Adversarial?

EMNLP 2017 had a “build-it-break-it” workshop that challenged humans to break existing systems by creating linguistic-based adversarial examples

“i.i.d. training data is unlikely to exhibit all the linguistic phenomena that we might see at testing time”

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

Adversarial?

How do we differentiate between an **adversarial attack** versus a model that is just bad?

Adversarial

An **adversarial attack** should slightly alter the input in a way that is **semantically equivalent to humans** but yields an incorrect, adverse change in the model's output

Adversarial

Let (x, y) be an input, output pair

Let x' be an altered version of x, which yields output y'

A successful attack will minimize $|x - x'|$ while maximizing $|y - y'|$, such that $|y - y'| > \tau$ or class(y) $\neq$ class(y')

Outline

- Introduction
- Paraphrasing
- Workshop time
- Modern approaches

Outline

- Introduction
- Paraphrasing
- Workshop time
- Modern approaches

Paraphrasing

PROBLEM

How can we change text while preserving its meaning?

Paraphrasing

Word-level substitutions

(aka lexical adversaries)

- **Embeddings**: search for nearest-neighbors in the embedding space
- **Thesaurus**: look up the word in a thesaurus, WordNet, or PPDB
- **Hybrid**: search for nearest-neighbors in the counter-fitted embedding space (Mrkšić et al, 2016)

Mrkšić et al., Counter-fitting Word Vectors to Linguistic Constraints (2016)

Paraphrasing

Word-level substitutions

Counter-fitted embeddings inject antonymy and synonymy constraints into vector space representations to help separate **conceptual association** from **semantic similarity**

	east	expensive	British
Before	west	pricey	American
	north	cheaper	Australian
	south	costly	Britain
	southeast	overpriced	European
	northeast	inexpensive	England
After	eastward	costly	Brits
	eastern	pricy	London
	easterly	overpriced	BBC
	-	pricey	UK
	-	afford	Britain

Table 1: Nearest neighbours for target words using GloVe vectors before and after counter-fitting

Paraphrasing

Word-level substitutions are difficult to craft

(aka lexical adversaries)

- How we can determine if a word swap is “acceptable” or not?
- This can be approximated by, or includes, word sense disambiguation (WSD) and language modelling (LM)
- Thus, can’t craft perfectly valid word substitutions all the time, but can do reasonably well

Paraphrasing

Sentence-level substitutions

(aka syntactic adversaries)

INPUT SENTENCE	MODEL PREDICTION
American drama doesn't get any more meaty and muscular than this.	positive
Doesn't get any more meaty and muscular than this American drama.	negative

Example from Ettinger et al., 2017

Paraphrasing

How can we create these **syntactic adversaries**

(aka sentence-level substitutions) automatically?

Paraphrasing

Sentence-level substitutions

(aka syntactic adversaries)

- Cosine similarity between sentence embeddings of x and x' (e.g., based on a Universal Sentence Encoder)
- Substitute many phrases (e.g., PPDB 2.0)
- Perform machine translation

Pavlick et al., PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification (2015)

Ideal syntactic paraphraser

- Produces grammatically-correct paraphrases the retain the meaning of the original sentence
- Minimizes the lexical differences between the input sentence x and paraphrase x'
- Generates many diverse syntactic paraphrases from the same input

Syntactic paraphrase generation

ORIGINAL

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

PARAPHRASES

- Usually, you required the inventory only if you were planning to sell the assets
- When you plan to sell your assets, you usually require inventory
- You need inventory when you plan to sell your assets
- Do the inventory when you plan to sell your assets

Syntactic paraphrase generation

These are grammatical, preserve input semantics, have minimal lexical substitution, and high syntactic diversity

ORIGINAL

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

PARAPHRASES

- Usually, you required the inventory only if you were planning to sell the assets
- When you plan to sell your assets, you usually require inventory
- You need inventory when you plan to sell your assets
- Do the inventory when you plan to sell your assets

Long history of paraphrase work

1 • *rule / template-based* syntactic paraphrasing
(e.g., McKeown, 1983; Carl et al., 2005)
  - high grammaticality, but very low diversity

2 • *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

3 • *deep learning-based* controlled language generation with conditional encoder/decoder architectures
(e.g., Ficler & Goldberg, 2017; Shen et al., 2017)
  - grammatical, but low diversity and no paraphrase constraint

MT

1. Paraphrasing with descriptive syntactic transformations

- first experiment: *rule-based* labels
  - She drives home. She is driven home. active > passive

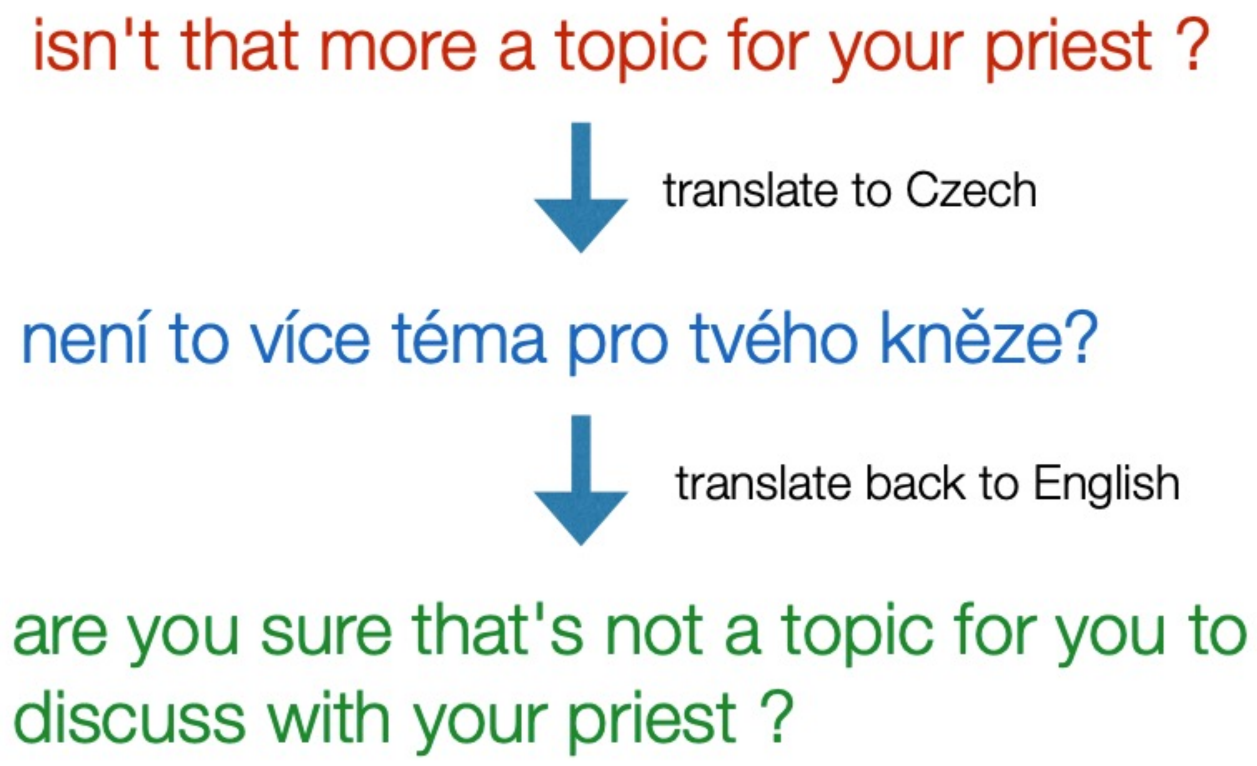
- Easy to write these rules, but low syntactic variance between the paraphrase pairs

MT

2. Translation-based uncontrolled paraphrasing

BACK-TRANSLATION

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



MT

2. Translation-based uncontrolled paraphrasing

- Could use several intermediate languages for backtranslation
- There is no control over how wild these sentences may get
- Limited research on the diversity and quality, due to many moving parts (data available, metrics, required human annotation)

3. Translation-based controlled paraphrasing

Step 1: Generate back-translation sentences from original sentences

Step 2: Run parser (e.g., constituency parsing) on the back-translation

Step 3: Train a new model that generates new text, conditioned on the original sentence and the parsed back-translation

3. Translation-based controlled paraphrasing

Step 1

s_1 isn't that more a topic for your priest ?

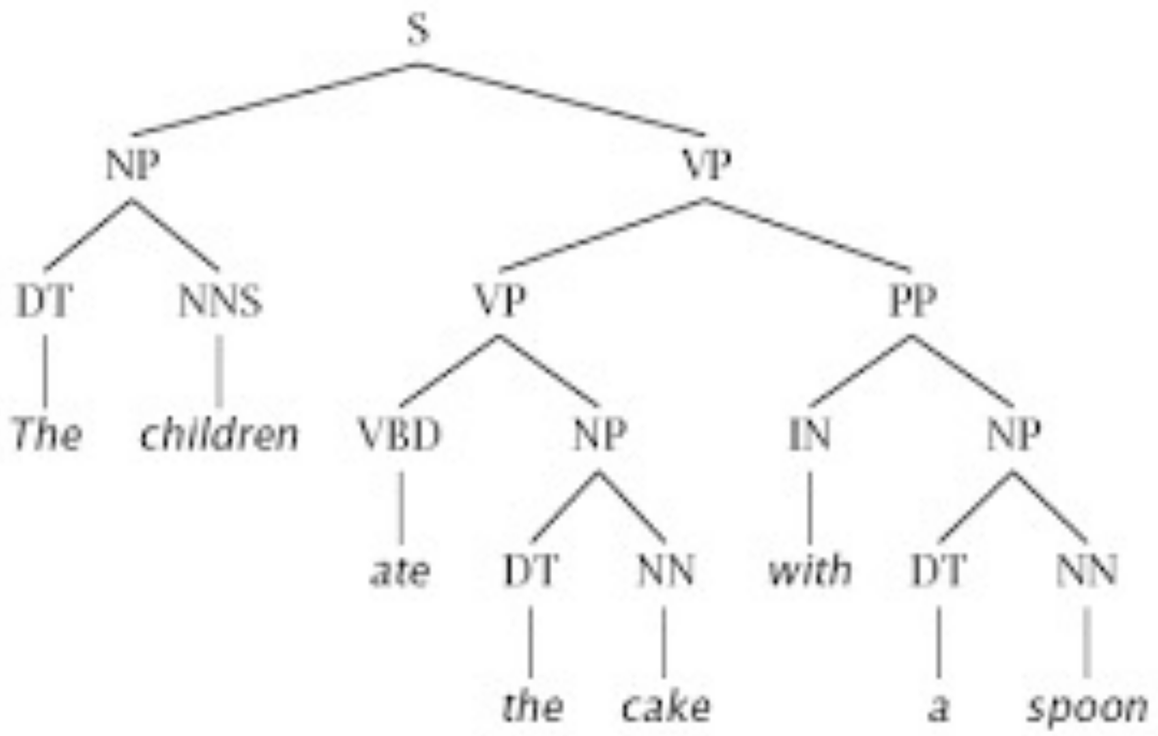
p_1 (ROOT (S (VP (VBZ) (RB) (SBARQ (IN) (NP (NP (JJR) (NP (NP (DT) (NN)) (PP (IN) (NP (PRP$) (NN))))))))) (.)))

s_2 are you sure that's not a topic for you to discuss with your priest ?

p_2 (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))))))))))))) (.)))

3. Translation-based controlled paraphrasing

Step 2



3. Translation-based controlled paraphrasing

Step 3

s_1 isn't that more a topic for your priest ?

p_2 (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))))))))))))) (.)))

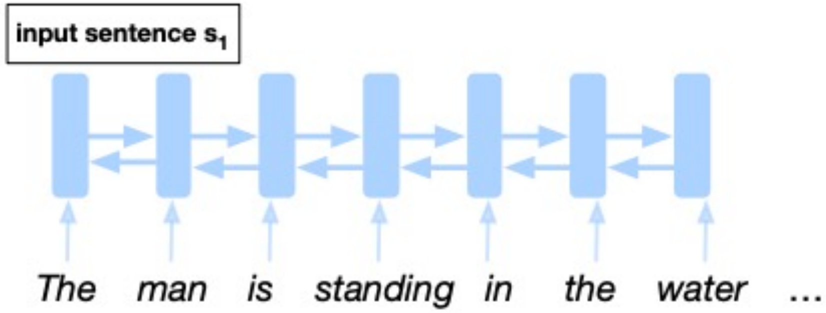
↓

s_2 are you sure that's not a topic for you to discuss with your priest ?

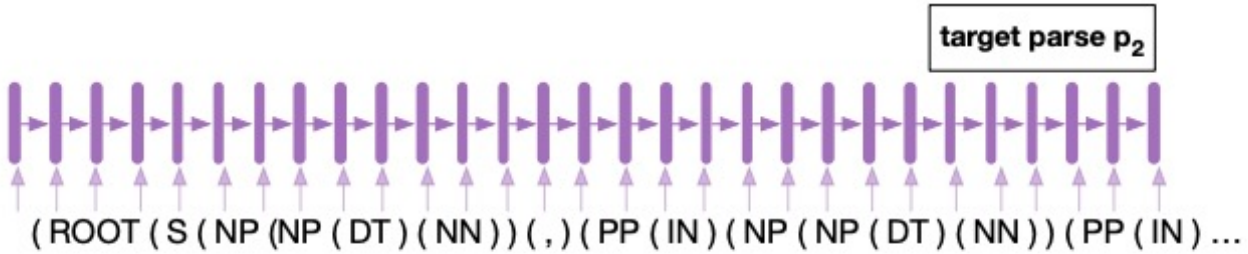
Step 3 (SCPN system)

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



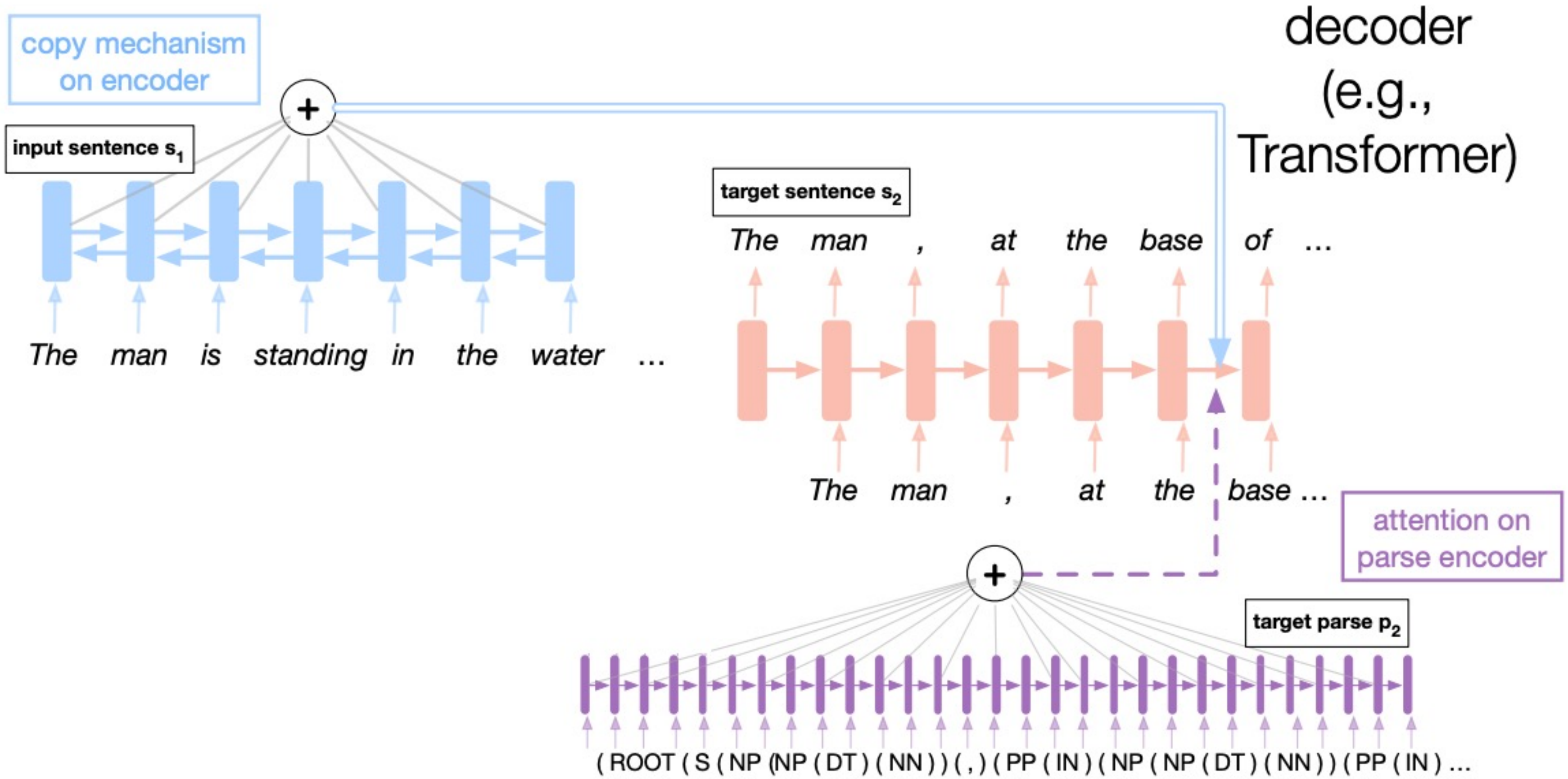
parse encoder (fine-tuned BERT?)



Step 3 (SCPN system)

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



3. Translation-based controlled paraphrasing

- Conditioning on the full target parse could be too rigorous and demanding
- Could "back-off" and use a pruned version of the parse

She drove home.

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

template: S →NP VP .

3. Translation-based controlled paraphrasing

SCPN system

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.

Iyyer et al., Adversarial Example Generation with Syntactically Controlled Paraphrase Networks (2018)

Outline

- Introduction
- Paraphrasing
- Workshop time
- Modern approaches

Outline

- Introduction
- Paraphrasing
- Workshop time
- Modern approaches

WORKSHOP TIME!

- How would you analyze how vulnerable your model is to adversarial attacks?
- How would you defend against such?

Outline

- Introduction
- Paraphrasing
- Workshop time
- Modern approaches

Outline

- Introduction
- Paraphrasing
- Workshop time
- Modern approaches

The field is budding (aka Wild West)

- Automated methods for determining semantic preservation are really hacky
- No agreed-upon definition of NLP adversarial examples

- We tried to lay out a theoretical definition for adversarial examples
  - TLDR: there are different definitions, depending on use case
    some adversarial examples wrt semantics, some wrt edit distance...

Input, x: "Shall I compare thee to a summer's day?" – William Shakespeare, Sonnet XVIII

Constraint	Perturbation, $\mathbf{x}_{adv}$	Explanation
Semantics	Shall I compare thee to a winter's day?	$\mathbf{x}_{adv}$ has a different meaning than $\mathbf{x}$.
Grammaticality	Shall I compares thee to a summer's day?	$\mathbf{x}_{adv}$ is less grammatically correct than $\mathbf{x}$.
Edit Distance	Sha1l i conpp$haaare thee to a 5umm3r's day?	$\mathbf{x}$ and $\mathbf{x}_{adv}$ have a large edit distance.
Non-suspicion	Am I gonna compare thee to a summer's day?	A human reader may suspect this sentence to have been modified. [1]

[1] Shakespeare never used the word "gonna". Its first recorded usage wasn't until 1806, and it didn't become popular until the 20th century.

Table 1: **Adversarial Constraints and Violations.** For each of the four proposed constraints, we show an example for which violates the specified constraint.

Morris et al., Reevaluating Adversarial Examples in Natural Language (2020)

Slide based on, inspired by, or directly from Jack Morris

The field is budding (aka Wild West)

Did human studies on a lot of generated adversarial examples from two NLP attacks

- original human studies said: these adversarial examples don't really preserve meaning, or grammar
- and **if we increase the threshold so that meaning is preserved, attack success rate drops a lot**

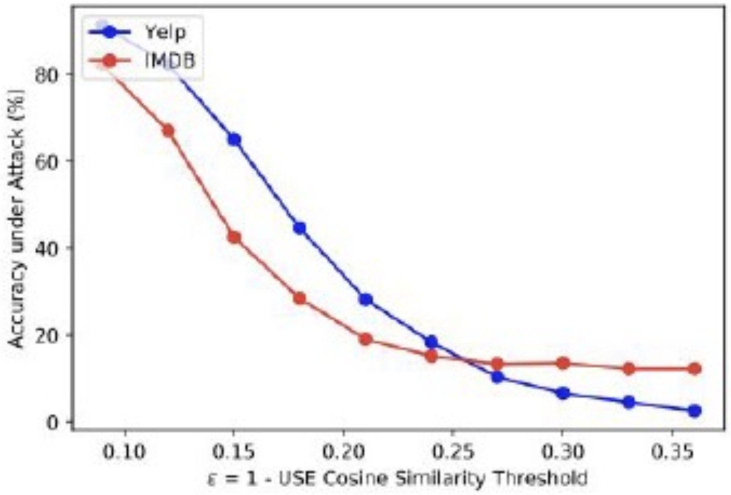


Figure 3: ϵ = 1 - USE Similarity Threshold vs. Accuracy under attack

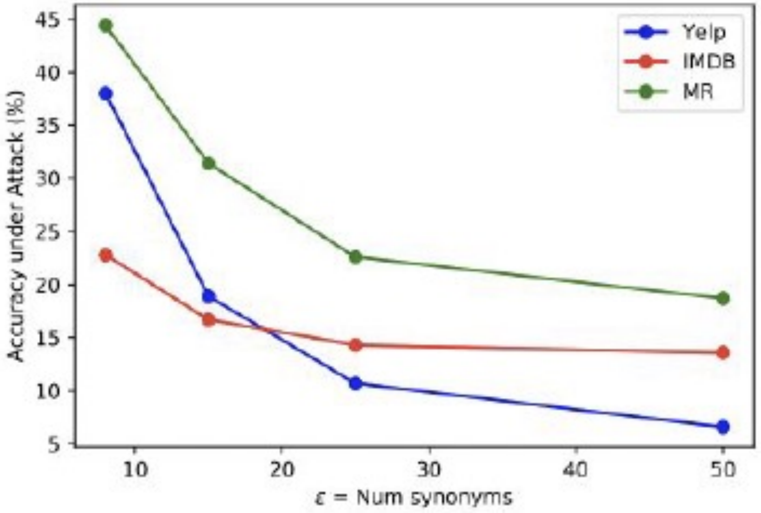


Figure 4: ϵ = Num synonyms vs. Accuracy under attack

Morris et al., Reevaluating Adversarial Examples in Natural Language (2020)

Slide based on, inspired by, or directly from Jack Morris

Examples of poor adversarial perturbations

Input, x:

"True Grit" was the **best movie** I've seen since I was a small boy.

Prediction: **Positive ✓**

Perturbation, x_{adv}:

"True Grit" was the **worst film** I've seen since I was a small boy.

"True Grit" was the best movie I've seen since I **were boy small**.

"True Grit" was the best movie I've seen since I was a **miniscule youngster**.

Prediction: **Negative ✗**

Morris et al., Reevaluating Adversarial Examples in Natural Language (2020)

Slide based on, inspired by, or directly from Jack Morris

Examples of poor adversarial perturbations

Perturbation, x_{adv}:

different **semantics** than original input	“True Grit” was the **worst film** I’ve seen since I was a small boy.
violates **grammar** (unlike the original input)	“True Grit” was the best movie I’ve seen since I **were boy small**.
this is just **suspicious** – nobody talks like that!	“True Grit” was the best movie I’ve seen since I was a **miniscule youngster**.

Prediction: **Negative ✗**

Morris et al., Reevaluating Adversarial Examples in Natural Language (2020)

Slide based on, inspired by, or directly from Jack Morris

Trends within adversarial literature

1. Attacks take the same overall approach, but modify one or two things
   - ex: use a greedy search instead of genetic algorithm
   - ex: use BERT word substitution instead of a thesaurus

2. Attacks compare success rates to each other but don't share code or models
   - minor implementation difference can make a massive impact on attack success rate

Overview

- A framework for Adversarial Attacks, Data Augmentation, and Adversarial Training
- Attacks consists of 4 main components

Morris et al., TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP (2020)

Overview

1. A task-specific **goal function** that determines whether the attack is successful in terms of the model outputs.
Examples: untargeted classification, targeted classification, non-overlapping output, minimum BLEU score.

2. A set of **constraints** that determine if a perturbation is valid with respect to the original input.
Examples: maximum word embedding distance, part-of-speech consistency, grammar checker, minimum sentence encoding cosine similarity.

3. A **transformation** that, given an input, generates a set of potential perturbations.
Examples: word embedding word swap, thesaurus word swap, homoglyph character substitution.

4. A **search method** that successively queries the model and selects promising perturbations from a set of transformations.
Examples: greedy with word importance ranking, beam search, genetic algorithm.

Morris et al., TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP (2020)

TextAttack Example

Is BERT Really Robust? (Jin, 2019)

Algorithm 1 Adversarial Attack by TEXTFOOLER

Input: Sentence example $X = \{w_1, w_2, ..., w_n\}$, the corresponding ground truth label Y, target model F, sentence similarity function $\mathrm{Sim}(\cdot)$, sentence similarity threshold ϵ, word embeddings Emb over the vocabulary Vocab.

Output: Adversarial example X_{adv}

Initialization: $X_{\text{adv}} \leftarrow X$
for each word w_i in X **do**
    Compute the importance score I_{w_i} via Eq. (2)
end for

Create a set W of all words $w_i \in X$ sorted by the descending order of their importance score I_{w_i}.
Filter out the stop words in W.
for each word w_j in W **do**
    Initiate the set of candidates CANDIDATES by extracting the top N synonyms using $\mathrm{CosSim}(\mathrm{Emb}_{w_j}, \mathrm{Emb}_{\text{word}})$ for each word in Vocab.
    CANDIDATES $\leftarrow$ POSFilter(CANDIDATES)
    FINCANDIDATES $\leftarrow \{\ \}$
    for c_k in CANDIDATES **do**
        $X' \leftarrow$ Replace w_j with c_k in X_{adv}
        if $\mathrm{Sim}(X', X_{\text{adv}}) > \epsilon$ **then**
            Add c_k to the set FINCANDIDATES
            $Y_k \leftarrow F(X')$
            $P_k \leftarrow F_{Y_k}(X')$
        end if
    end for
    if there exists c_k whose prediction result $Y_k \neq Y$ **then**
        In FINCANDIDATES, only keep the candidates c_k whose prediction result $Y_k \neq Y$
        $c^* \leftarrow \underset{c \in \text{FINCANDIDATES}}{\mathrm{argmax}} \ \mathrm{Sim}(X, X'_{w_j \rightarrow c})$
        $X_{\text{adv}} \leftarrow$ Replace w_j with c^* in X_{adv}
        return X_{adv}
    else if $P_{Y_k}(X_{\text{adv}}) > \underset{c_k \in \text{FINCANDIDATES}}{\min} P_k$ **then**
        $c^* \leftarrow \underset{c_k \in \text{FINCANDIDATES}}{\mathrm{argmin}} \ P_k$
        $X_{\text{adv}} \leftarrow$ Replace w_j with c^* in X_{adv}
    end if
end for
return None

Morris et al., TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP (2020)

TextAttack Example

Is BERT Really Robust? (Jin, 2019)

for each word w_i in X **do**
  Compute the importance score I_{w_i} via Eq. (2)
end for

Create a set W of all words $w_i \in X$ sorted by the descending order of their importance score I_{w_i}.
Filter out the stop words in W.
for each word w_j in W **do**
  Initiate the set of candidates CANDIDATES by extracting the top N synonyms using $\text{CosSim}(\text{Emb}_{w_j}, \text{Emb}_{\text{word}})$ for each word in Vocab.
  CANDIDATES $\leftarrow$ POSFilter(CANDIDATES)
  FINCANDIDATES $\leftarrow \{\}$
  for c_k in CANDIDATES **do**
    $X' \leftarrow$ Replace w_j with c_k in X_{adv}
    if $\text{Sim}(X', X_{\text{adv}}) > \epsilon$ **then**
      Add c_k to the set FINCANDIDATES
      $Y_k \leftarrow F(X')$
      $P_k \leftarrow F_{Y_k}(X')$
    end if
  end for
  if there exists c_k whose prediction result $Y_k \neq Y$ **then**
    In FINCANDIDATES, only keep the candidates c_k whose prediction result $Y_k \neq Y$
    $c^* \leftarrow \underset{c \in \text{FINCANDIDATES}}{\text{argmax}} \; \text{Sim}(X, X'_{w_j \to c})$
    $X_{\text{adv}} \leftarrow$ Replace w_j with c^* in X_{adv}
    return X_{adv}
  else if $P_{Y_k}(X_{\text{adv}}) > \underset{c_k \in \text{FINCANDIDATES}}{\min} P_k$ **then**

Search method: Greedy with "Word Importance Ranking"

Transformation: Counter-fitted embedding word swap

Constraint #1: Word embedding cosine similarity

Constraint #2: Word part-of-speech consistency

Constraint #1: Sentence embedding cosine similarity

Goal function: Untargeted classification

Morris et al., TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP (2020)

Better practices

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Marco Tulio Ribeiro[1] **Tongshuang Wu**[2] **Carlos Guestrin**[2] **Sameer Singh**[3]
[1]Microsoft Research [2]University of Washington [3]University of California, Irvine
marcotcr@gmail.com {wtshuang,guestrin}@cs.uw.edu sameer@uci.edu

- Motivation: dev accuracy is very short-sighted and tends to over-estimate performance
- This paper is inspired by principles of behavioral testing in software engineering
- New evaluation methodology and accompanying tool for comprehensive behavioral testing of NLP models
    - Guides users in what to test, by providing a list of linguistic capabilities, which are applicable to most tasks.

Conclusion

- Adversarial NLP is relatively new and still forming as a field
- Touches on software testing, data augmentation, robustness, learning theory, etc
- All systems can break; it's highly informative to be aware of this and understand how your model breaks
- One may want to analyze, defend, or attack.