On Evaluation of Adversarial Perturbations for Sequence-to-Sequence Models

Paul Michel, Xian Li,
Graham Neubig, Juan Pino

Carnegie Mellon University
Language Technologies Institute

NEULAB

facebook Artificial Intelligence

Adversarial Attacks/Perturbations

- Apply a **small** (indistinguishable) perturbation to the **input** that elicit **large** changes in the **output**

Adversarial Attacks/Perturbations

- Apply a **small** (indistinguishable) perturbation to the **input** that elicit **large** changes in the **output**

"panda"
57.7% confidence

Figure from Goodfellow et al. (2014)

Adversarial Attacks/Perturbations

- Apply a **small** (indistinguishable) perturbation to the **input** that elicit **large** changes in the **output**

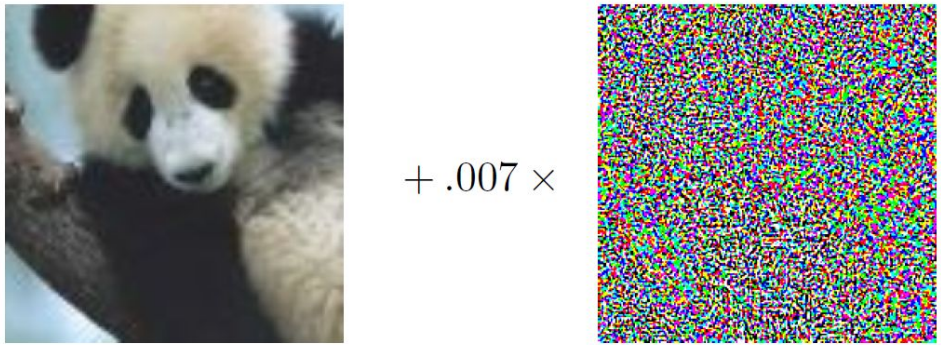


Figure from Goodfellow et al. (2014)

Adversarial Attacks/Perturbations

- Apply a **small** (indistinguishable) perturbation to the **input** that elicit **large** changes in the **output**

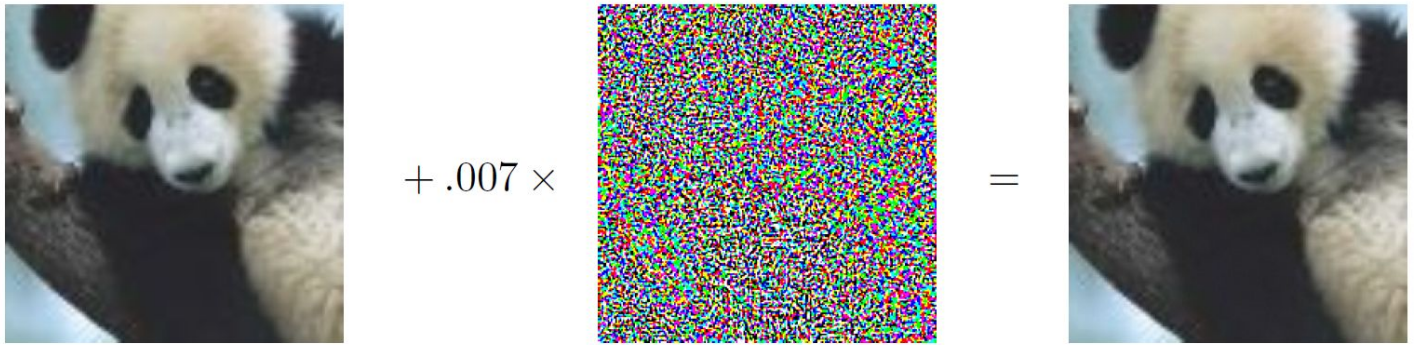


Figure from Goodfellow et al. (2014)

Adversarial Attacks/Perturbations

- Apply a **small** (indistinguishable) perturbation to the **input** that elicit **large** changes in the **output**

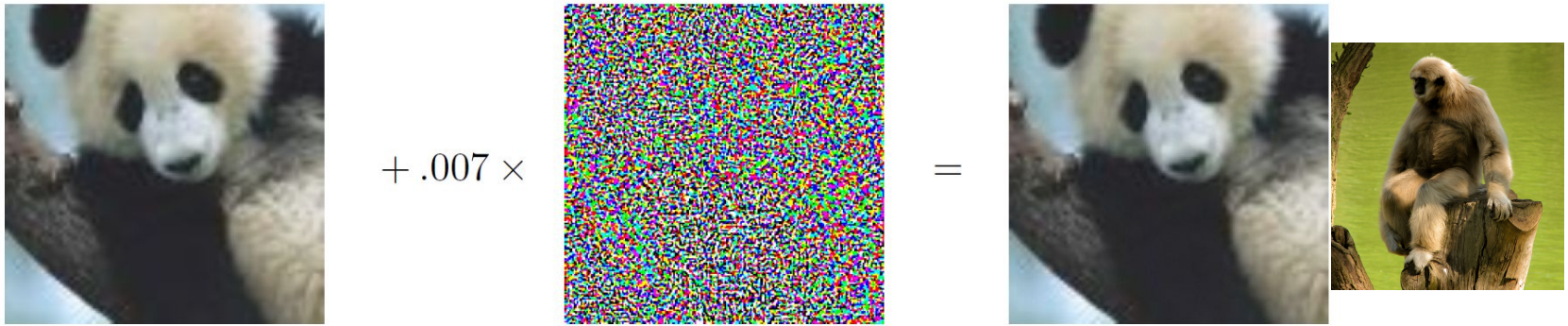
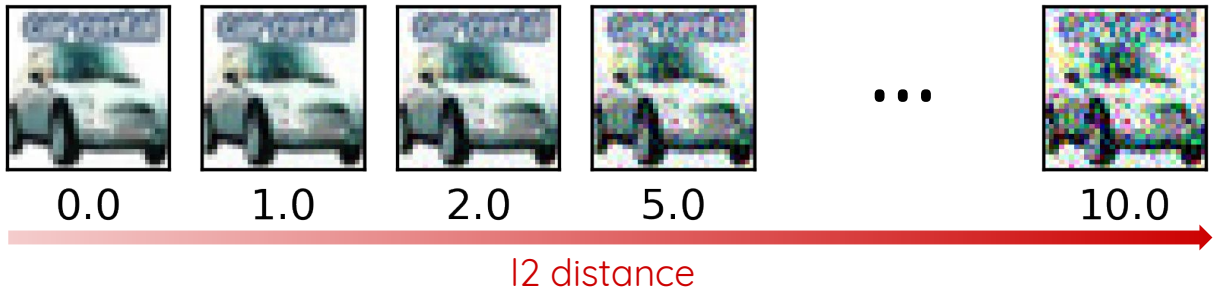


Figure from Goodfellow et al. (2014)

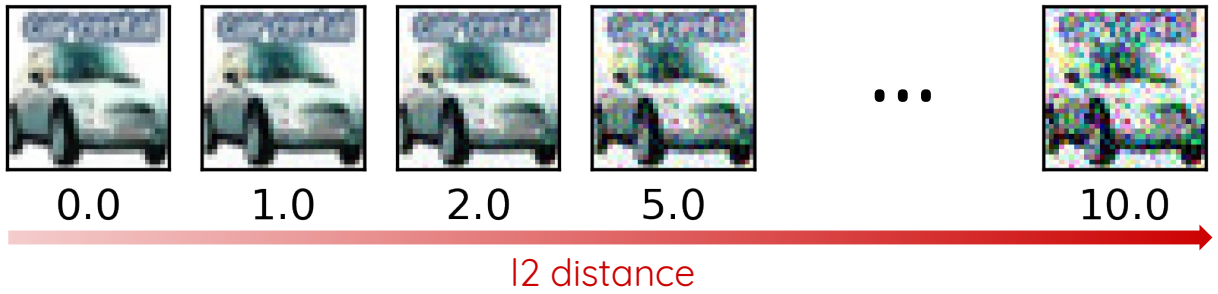
Indistinguishable Perturbations

- **Small** perturbations are well defined in **vision**
  - Small l2 ~= indistinguishable to the human eye



Indistinguishable Perturbations

- **Small** perturbations are well defined in **vision**
  - Small l2 ~= indistinguishable to the human eye



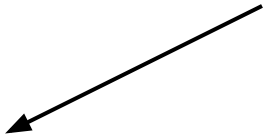
- What about **text**?

Not all Text Perturbations are Equal

He’s very friendly

Not all Text Perturbations are Equal

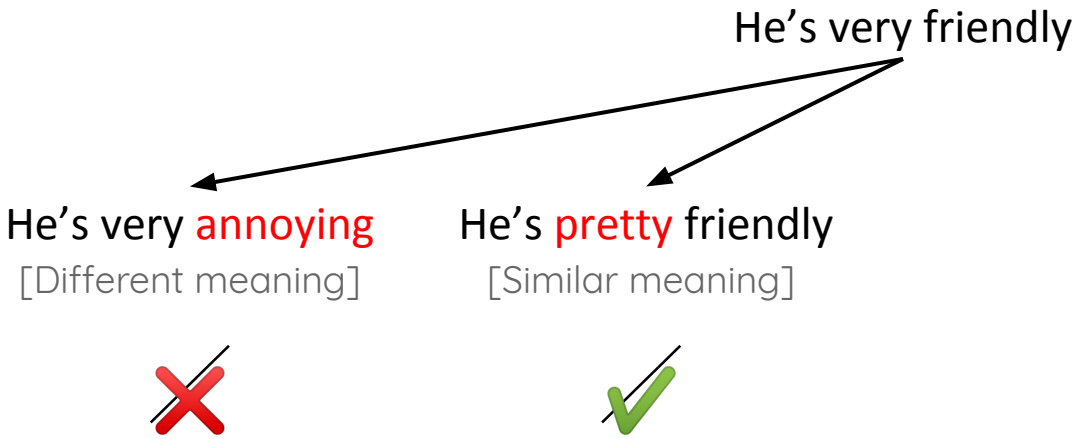
He’s very friendly



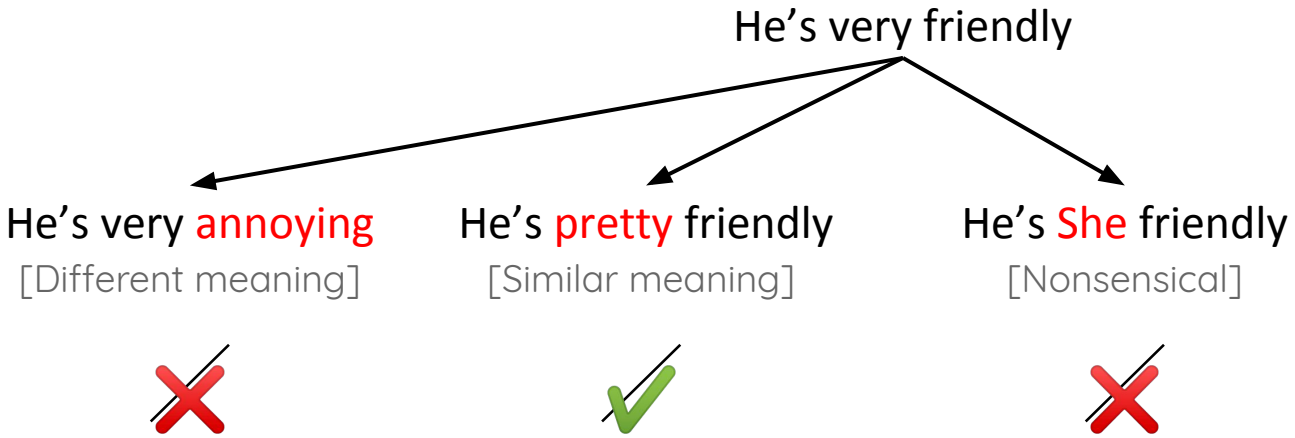
He’s pretty friendly

[Similar meaning]

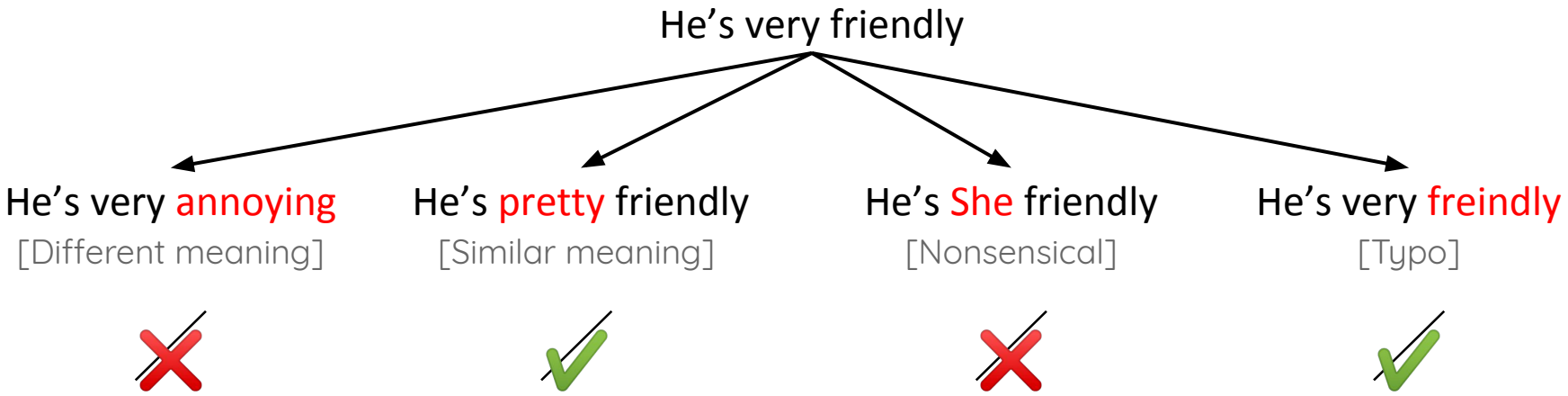
Not all Text Perturbations are Equal



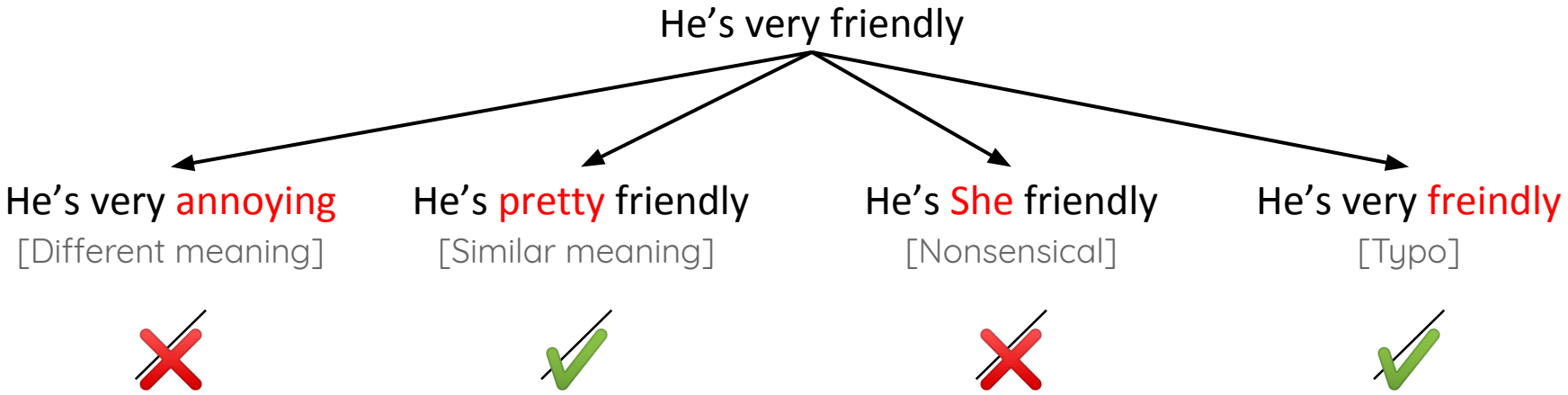
Not all Text Perturbations are Equal



Not all Text Perturbations are Equal

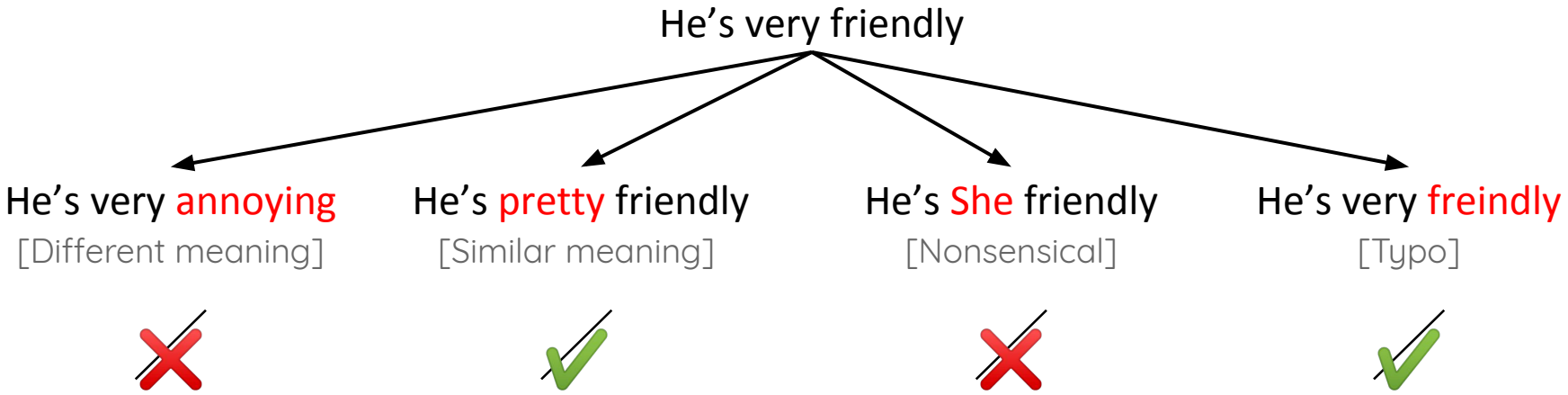


Not all Text Perturbations are Equal



⇒Can't expect the model to output the same output!

Not all Text Perturbations are Equal



⇒Can't expect the model to output the same output!

This paper:

Why and **How** you should evaluate adversarial perturbations

A Framework for Evaluating Adversarial Attacks

Problem Definition

Reference y

They plow it right back into filing more troll lawsuits.

Original x

Ils le réinvestissent directement en engageant plus de procès.

Problem Definition

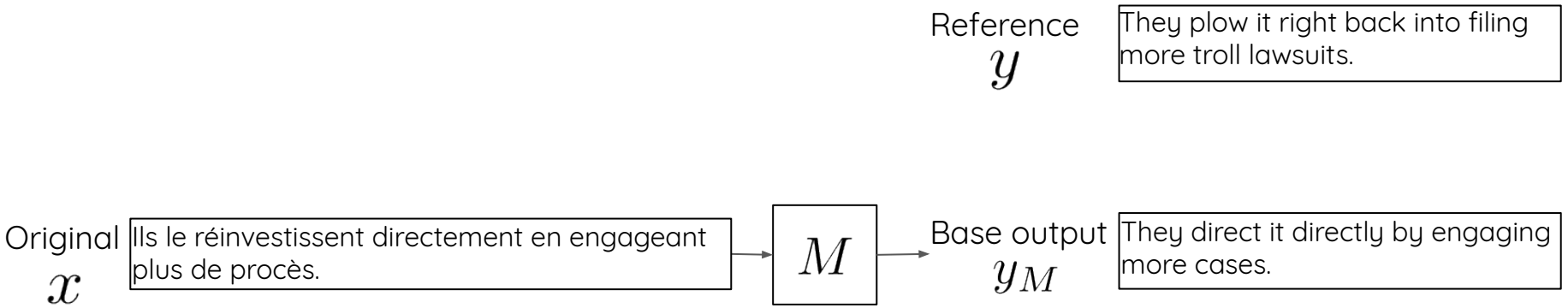
Reference
y

They plow it right back into filing more troll lawsuits.

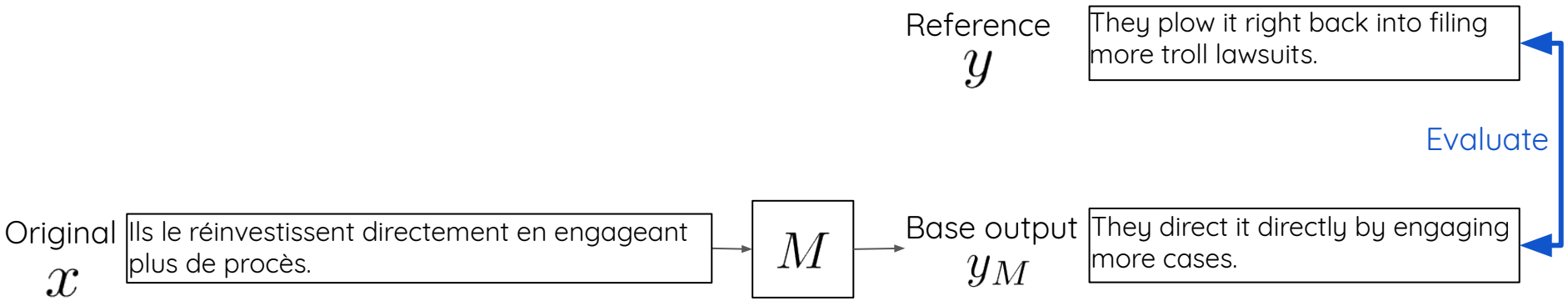
Original
x

Ils le réinvestissent directement en engageant plus de procès.

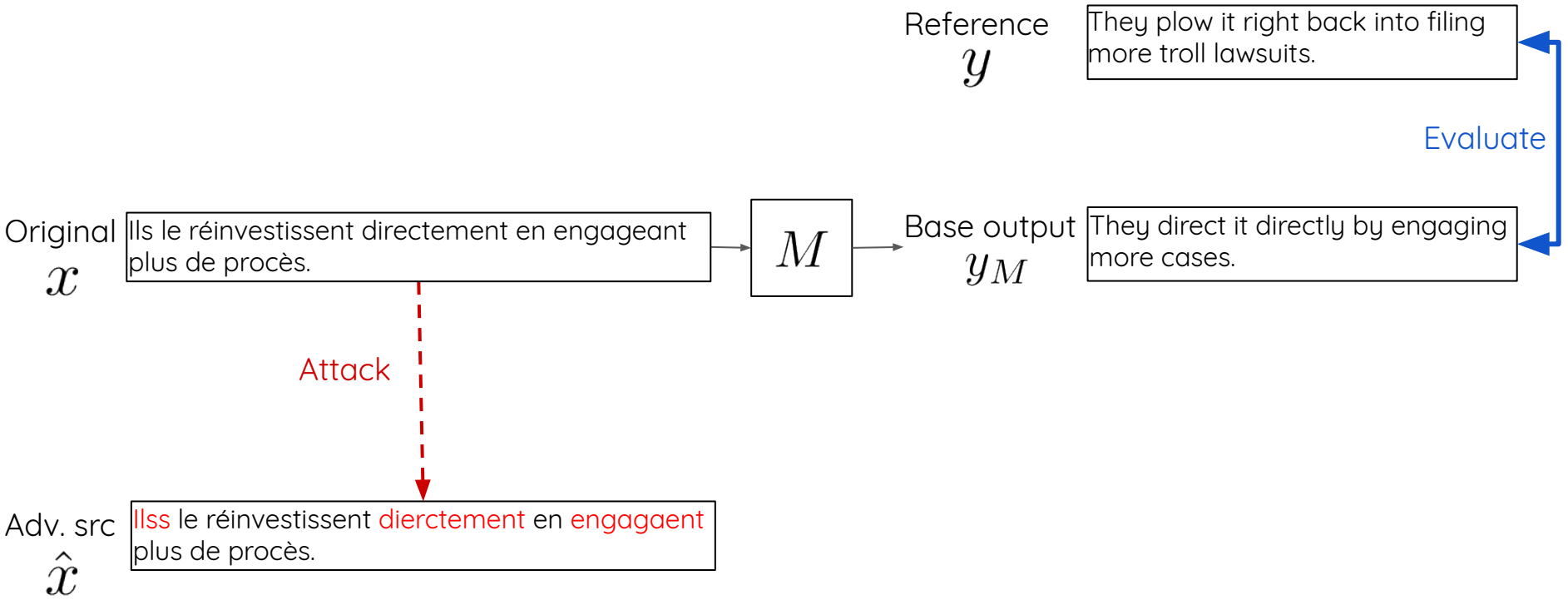
Problem Definition



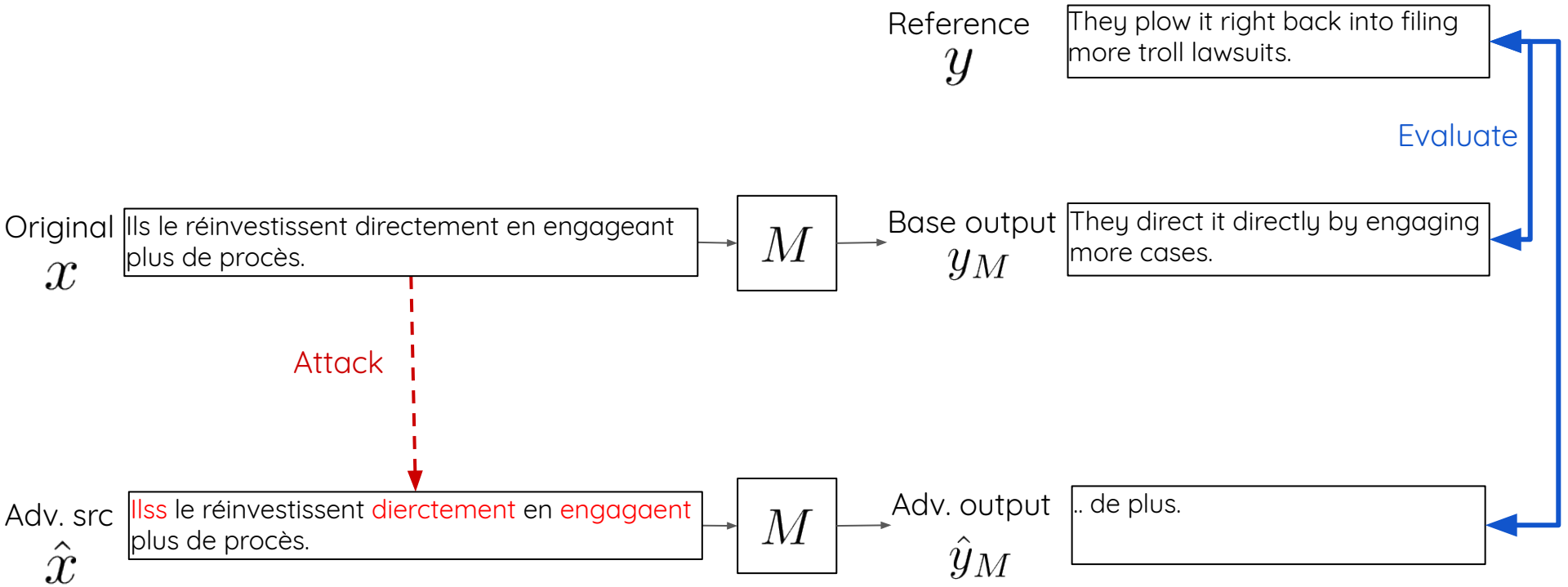
Problem Definition



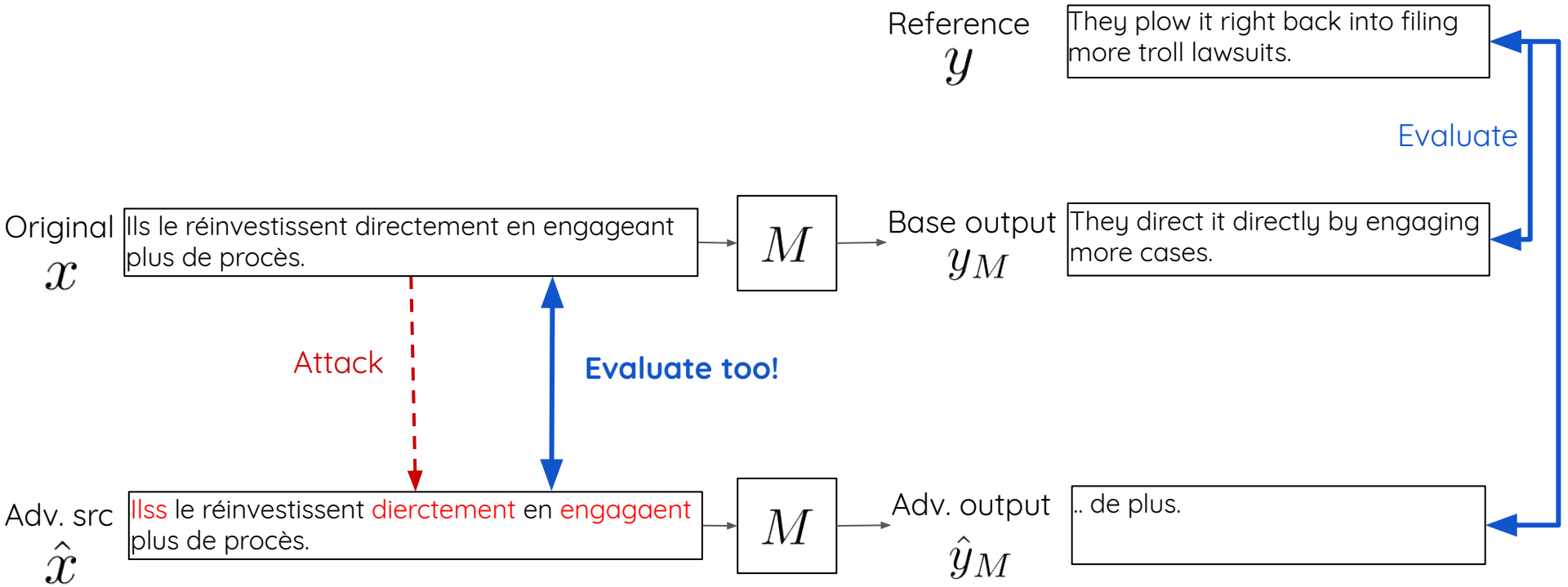
Problem Definition



Problem Definition



Problem Definition



Source Side Evaluation

- Evaluate **meaning preservation** on the source side

$$s_{src}(x, \hat{x})$$

- Where s_{src} is a **similarity metric** such that

s_{src}(He's very friendly , He's pretty friendly) > s_{src}(He's very friendly , He's very annoying)

s_{src}(He's very friendly , He's pretty friendly) > s_{src}(He's very friendly , He's She friendly)

[…]

Target Side Evaluation

- Given s_{tgt}, a similarity metric on the target side

Target Side Evaluation

- Given s_{tgt}, a similarity metric on the target side
- Evaluate **relative meaning destruction** on the target side

$$d_{tgt}(y, y_M, \hat{y}_M) = \begin{cases} \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} \end{cases}$$

Target Side Evaluation

- Given s_{tgt}, a similarity metric on the target side
- Evaluate **relative meaning destruction** on the target side

$$d_{tgt}(y, y_M, \hat{y}_M) = \left\{ \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} \right.$$

Target Side Evaluation

- Given s_{tgt}, a similarity metric on the target side
- Evaluate **relative meaning destruction** on the target side

$$d_{tgt}(y, y_M, \hat{y}_M) = \left\{ \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} \right.$$

Target Side Evaluation

- Given s_{tgt}, a similarity metric on the target side
- Evaluate **relative meaning destruction** on the target side

$$d_{tgt}(y, y_M, \hat{y}_M) = \begin{cases} 0 \text{ if } s_{tgt}(y, \hat{y}_M) \geq s_{tgt}(y, y_M) \\ \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} \end{cases}$$

Target Side Evaluation

- Given s_{tgt}, a similarity metric on the target side
- Evaluate **relative meaning destruction** on the target side

$$d_{tgt}(y, y_M, \hat{y}_M) = \begin{cases} 0 \text{ if } s_{tgt}(y, \hat{y}_M) \geq s_{tgt}(y, y_M) \\ \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} \end{cases}$$

Target Side Evaluation

- Given s_{tgt}, a similarity metric on the target side
- Evaluate **relative meaning destruction** on the target side

$$d_{tgt}(y, y_M, \hat{y}_M) = \begin{cases} 0 \text{ if } s_{tgt}(y, \hat{y}_M) \geq s_{tgt}(y, y_M) \\ \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} \text{ otherwise} \end{cases}$$

Successful Adversarial Attacks

- Ensure that:

$$1 - s_{src}(x, \hat{x}) < d_{tgt}(y, y_M, \hat{y}_M)$$

Successful Adversarial Attacks

- Ensure that:

$$\boxed{1 - s_{src}(x, \hat{x})} < d_{tgt}(y, y_M, \hat{y}_M)$$

Source meaning destruction

Successful Adversarial Attacks

- Ensure that:

$$1 - s_{src}(x, \hat{x}) < d_{tgt}(y, y_M, \hat{y}_M)$$

Source meaning destruction

Target meaning destruction

Successful Adversarial Attacks

- Ensure that:

$$1 - s_{src}(x, \hat{x}) < d_{tgt}(y, y_M, \hat{y}_M)$$

Source meaning destruction **Target meaning destruction**

- Destroy the meaning on the target side more than on the source side

Which similarity metric to use?

- Human evaluation
  - 6 point scale, details in paper

"How would you rate the similarity between the meaning of these two sentences?"

0. The meaning is completely different or one of the sentences is meaningless
1. The topic is the same but the meaning is different
2. Some key information is different
3. The key information is the same but the details differ
4. Meaning is essentially the same but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed [Language]

Which similarity metric to use?

- Human evaluation
  - 6 point scale, details in paper

"How would you rate the similarity between the meaning of these two sentences?"

0. The meaning is completely different or one of the sentences is meaningless
1. The topic is the same but the meaning is different
2. Some key information is different
3. The key information is the same but the details differ
4. Meaning is essentially the same but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed [Language]

- BLEU [Papineni et al., 2002]
  - Geometric mean of n-gram precision + length penalty

Which similarity metric to use?

- Human evaluation
  - 6 point scale, details in paper

"How would you rate the similarity between the meaning of these two sentences?"

0. The meaning is completely different or one of the sentences is meaningless
1. The topic is the same but the meaning is different
2. Some key information is different
3. The key information is the same but the details differ
4. Meaning is essentially the same but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed [Language]

- BLEU [Papineni et al., 2002]
  - Geometric mean of n-gram precision + length penalty

- METEOR [Banerjee and Lavie, 2005]
  - Word matching taking into account stemming, synonyms, paraphrases...

Which similarity metric to use?

- Human evaluation
  - 6 point scale, details in paper

"How would you rate the similarity between the meaning of these two sentences?"

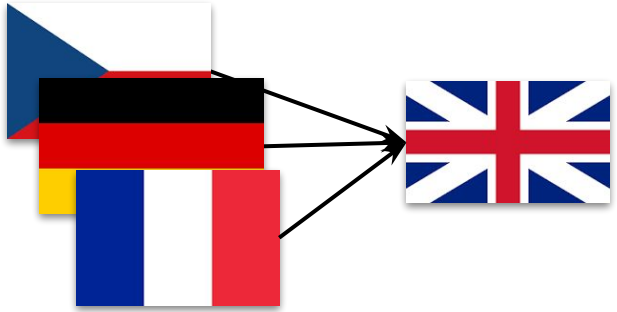
0. The meaning is completely different or one of the sentences is meaningless
1. The topic is the same but the meaning is different
2. Some key information is different
3. The key information is the same but the details differ
4. Meaning is essentially the same but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed [Language]

- BLEU [Papineni et al., 2002]
  - Geometric mean of n-gram precision + length penalty
- METEOR [Banerjee and Lavie, 2005]
  - Word matching taking into account stemming, synonyms, paraphrases...
- chrF [Popović, 2015]
  - Character n-gram F-score

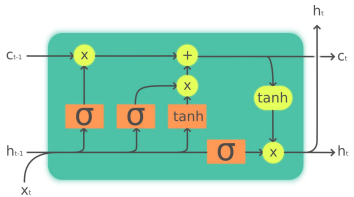
Experimental Setting

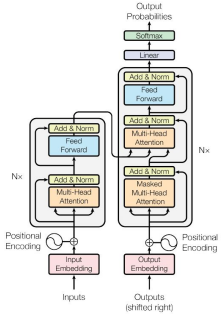
Data and Models

- Data
  - IWSLT 2016 dataset
  - {Czech, German, French} → English



- Models
  - LSTM based model
  - Transformer based model
  - Both word and sub-word based models





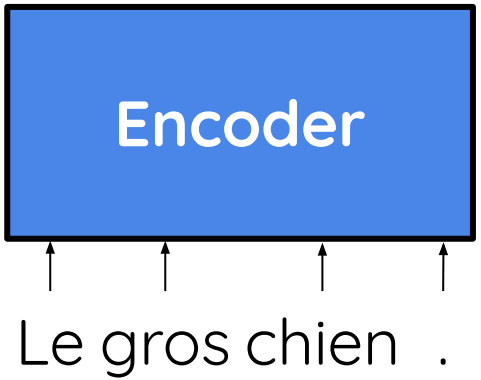
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions

Le gros chien .

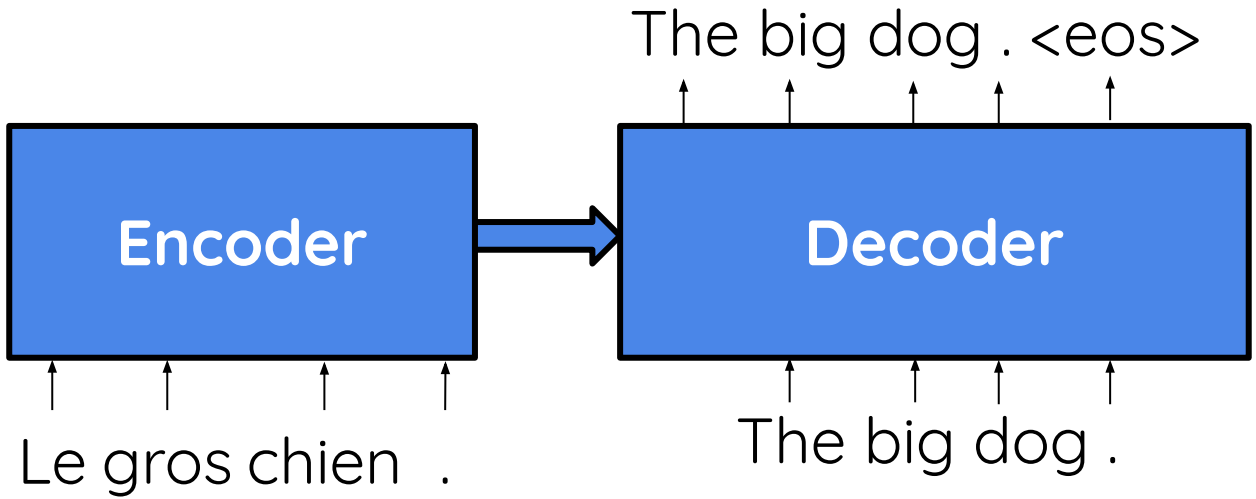
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions



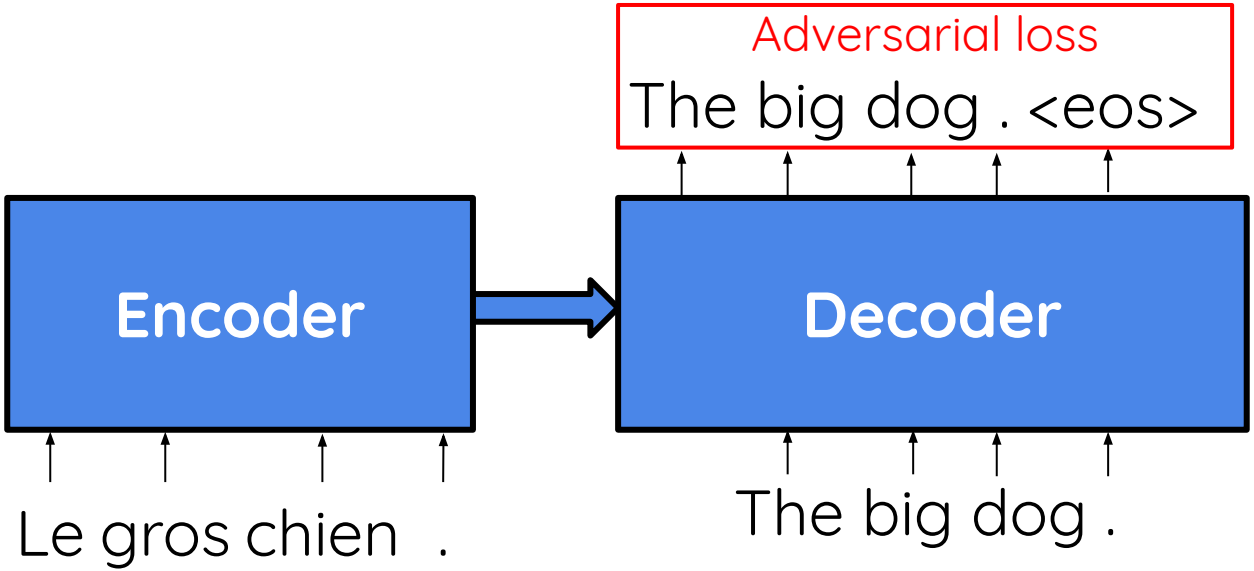
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions



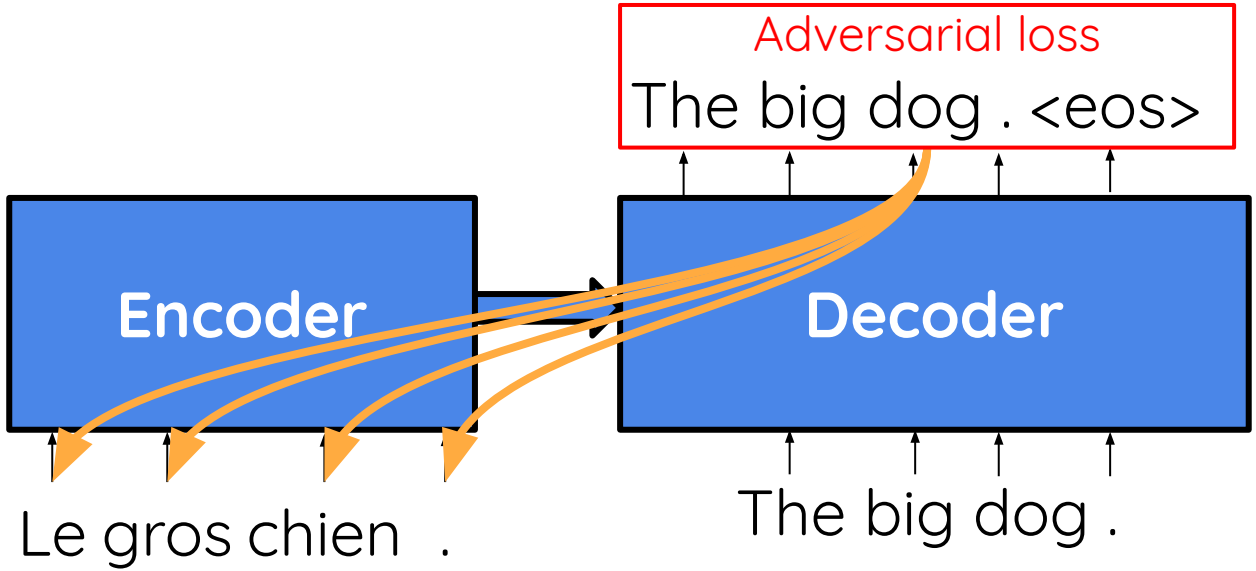
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions



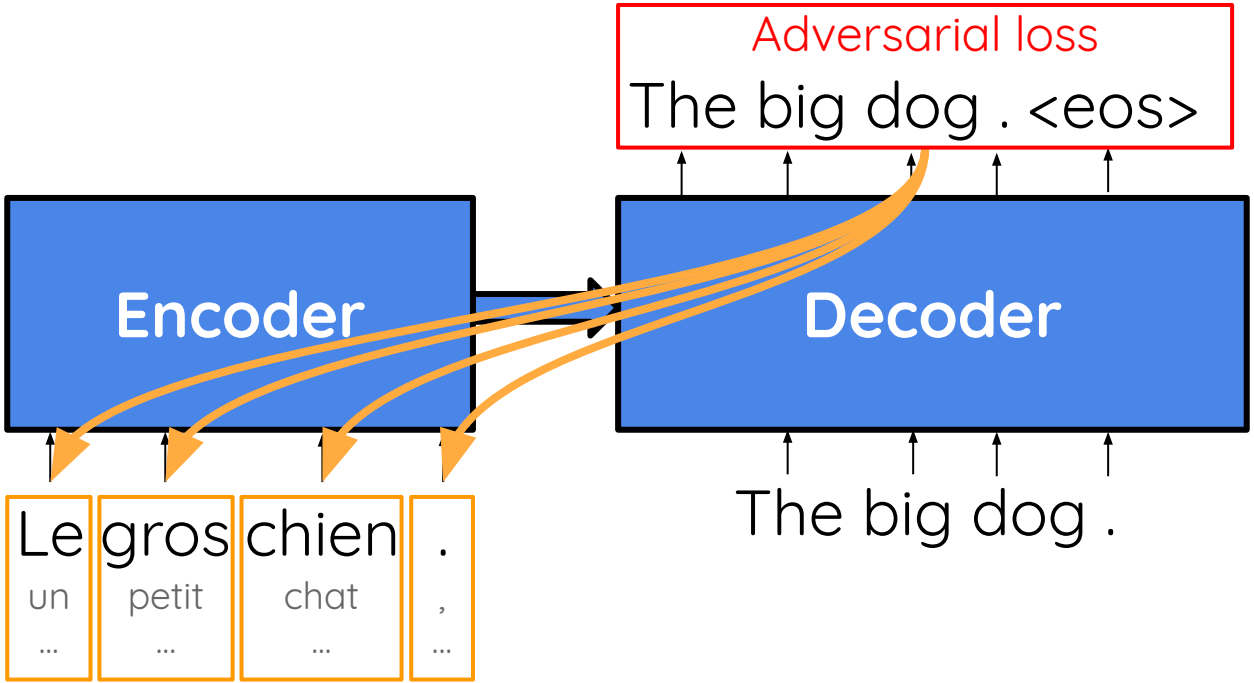
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions



Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions



Constrained Adversarial Attacks

Constrained Adversarial Attacks: kNN

- Only replace words with 10 nearest neighbors in embedding space

Example from our fr→en Transformer source embeddings

- grand (tall SING+MASC)
  - grands (tall PL+MASC)
  - grande (tall SING+FEM)
  - grandes (tall PL+FEM)
  - gros (fat SING+MASC)
  - grosse (fat SING+FEM)
- math (math)
  - maths (maths)
  - mathématique (mathematic)
  - mathématiques (mathematics)
  - objective (objective [ADJ] SING+FEM)

Constrained Adversarial Attacks: CharSwap

- Only swap word internal characters to get OOVs
  - grand → gr**na**d
  - adversarial → adv**re**sarial
  - […]
- If that’s impossible, repeat the last character
  - he → he**eeeeee**

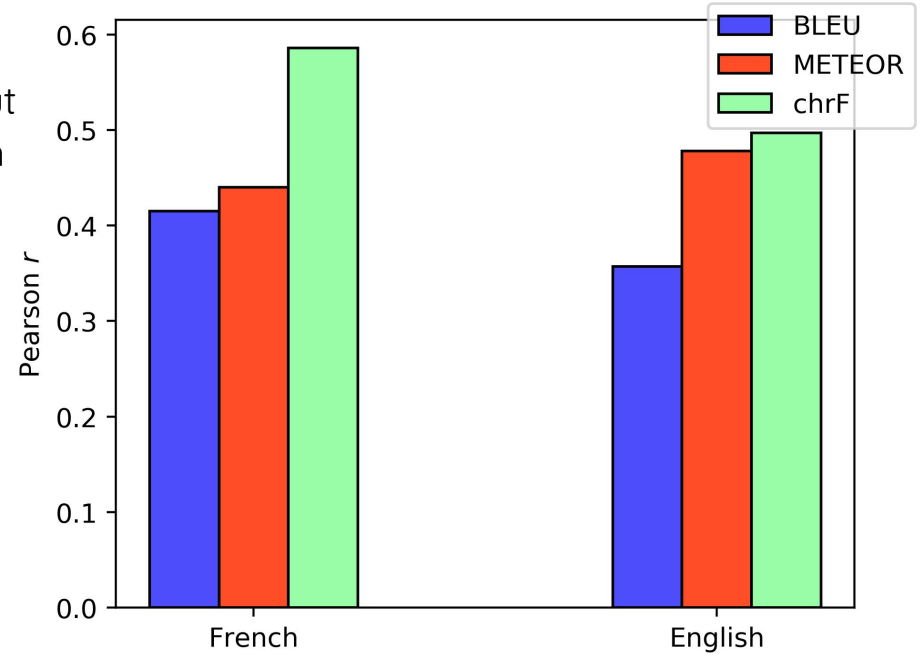
⇒ Realistic typos

Constrained Adversarial Attacks

Original	**Pourquoi** faire cela ?
English gloss	**Why** do this?
Unconstrained	**construisant** (English: building) faire cela ?
kNN	**interrogez** (English: to question VB.2nd.PL) faire cela ?
CharSwap	**Puorquoi** (typo) faire cela ?
Original	Si seulement je pouvais me muscler **aussi** rapidement.
English gloss	If only I could build my muscle **this** fast.
Unconstrained	Si seulement je pouvais me muscler **etc** rapidement.
kNN	Si seulement je pouvais me muscler **plsu** (typo for "more") rapidement.
CharSwap	Si seulement je pouvais me muscler **asusi** (typo) rapidement.

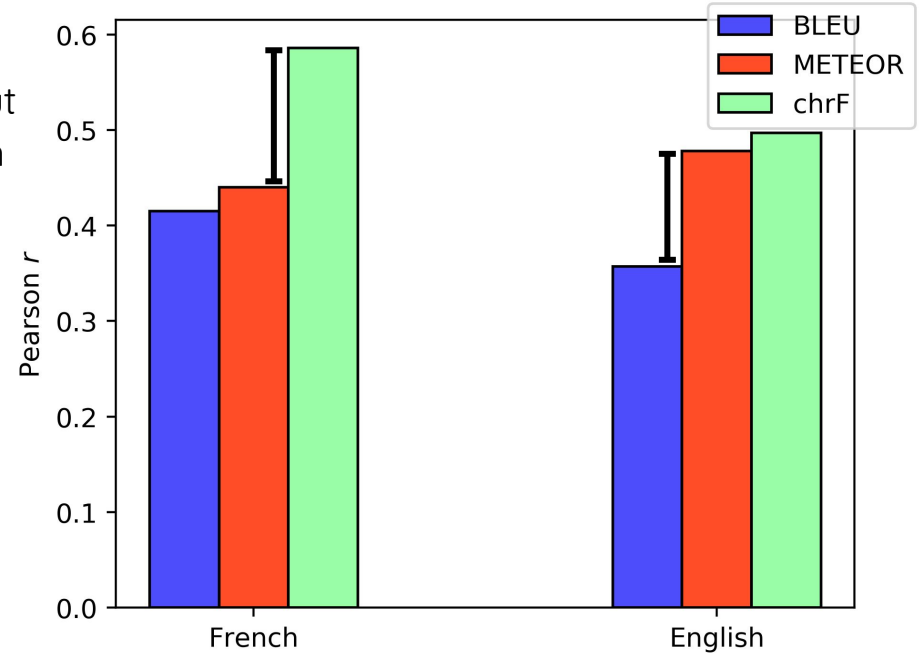
Choosing an Similarity Metric

- Human vs automatic (pearson r):
  - Humans score original/adversarial input
  - Humans score original/adversarial output
  - Compare scores to automatic metric with Pearson correlation



Choosing an Similarity Metric

- Human vs automatic (pearson r):
  - Humans score original/adversarial input
  - Humans score original/adversarial output
  - Compare scores to automatic metric with Pearson correlation



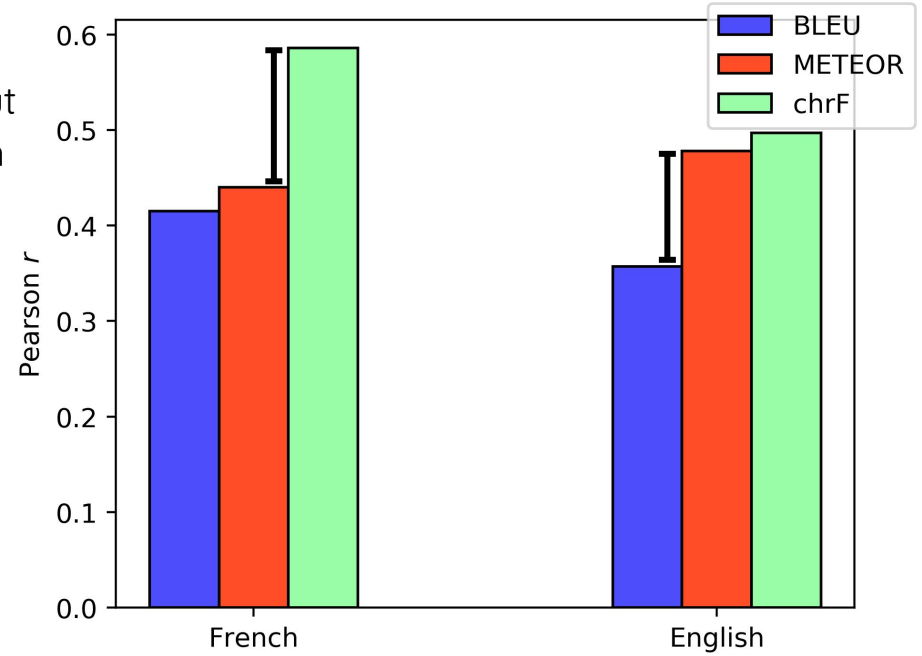
Choosing an Similarity Metric

- Human vs automatic (pearson r):
  - Humans score original/adversarial input
  - Humans score original/adversarial output
  - Compare scores to automatic metric with Pearson correlation

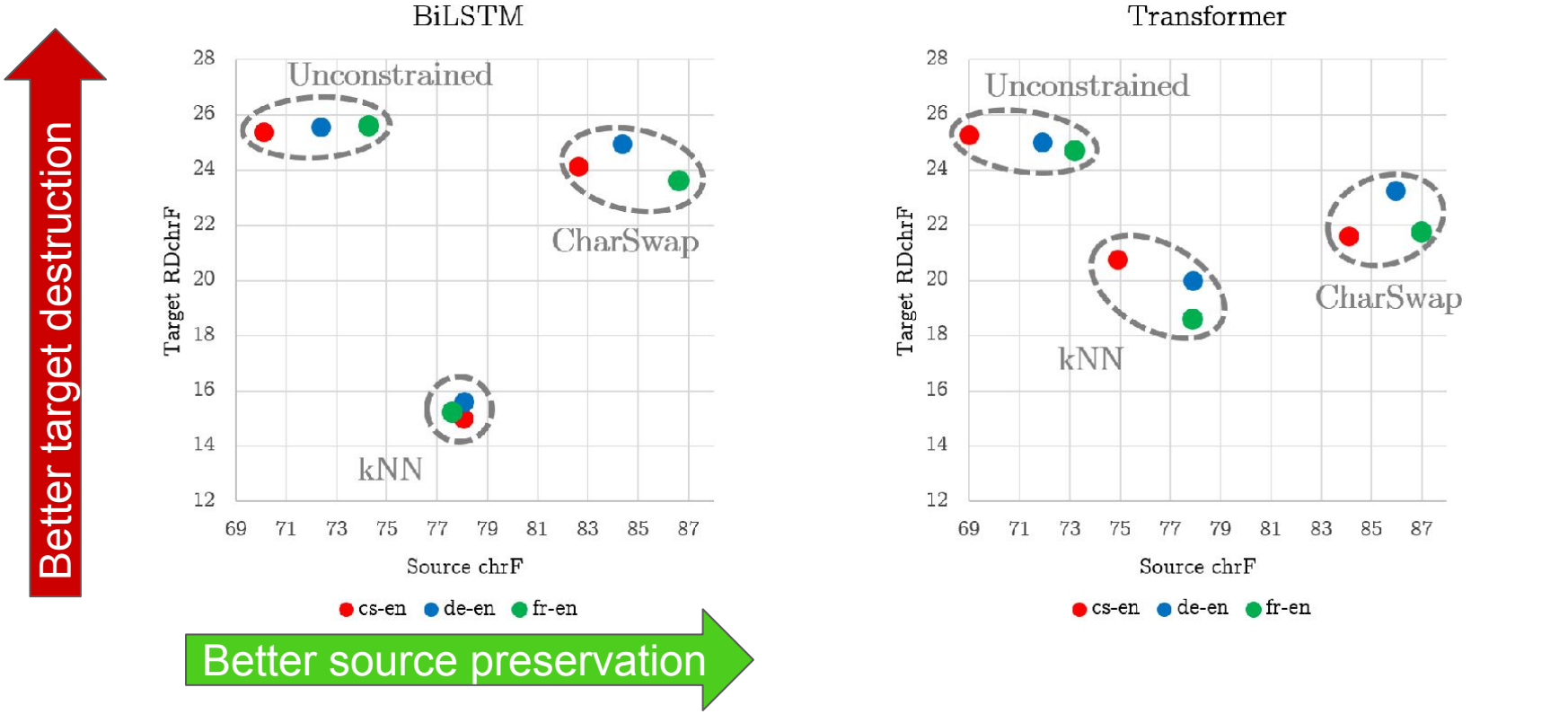
- chrF better

  ⇒ $s_{src} = s_{tgt}$:= **chrF**

  ⇒ d_{tgt} := **RDchrF**
  (**R**elative **D**ecrease in **chrF**)



Effect of Constraints on Evaluation



Effect of Constraints on Adversarial Training

Effect of Constraints on Adversarial Training

- Adversarial training ≈ training with adversarial examples

$$\mathcal{L}'(x, y) = (1 - \alpha)NLL(x, y) + \alpha NLL(\hat{x}, y)$$

Standard input

Adversarial input

  - α = 0: Standard training
  - α = 1 : Training only on adversarial examples

Effect of Constraints on Adversarial Training

- Adversarial training ≈ training with adversarial examples

$$\mathcal{L}'(x, y) = (1 - \alpha)NLL(x, y) + \alpha NLL(\hat{x}, y)$$

Standard input

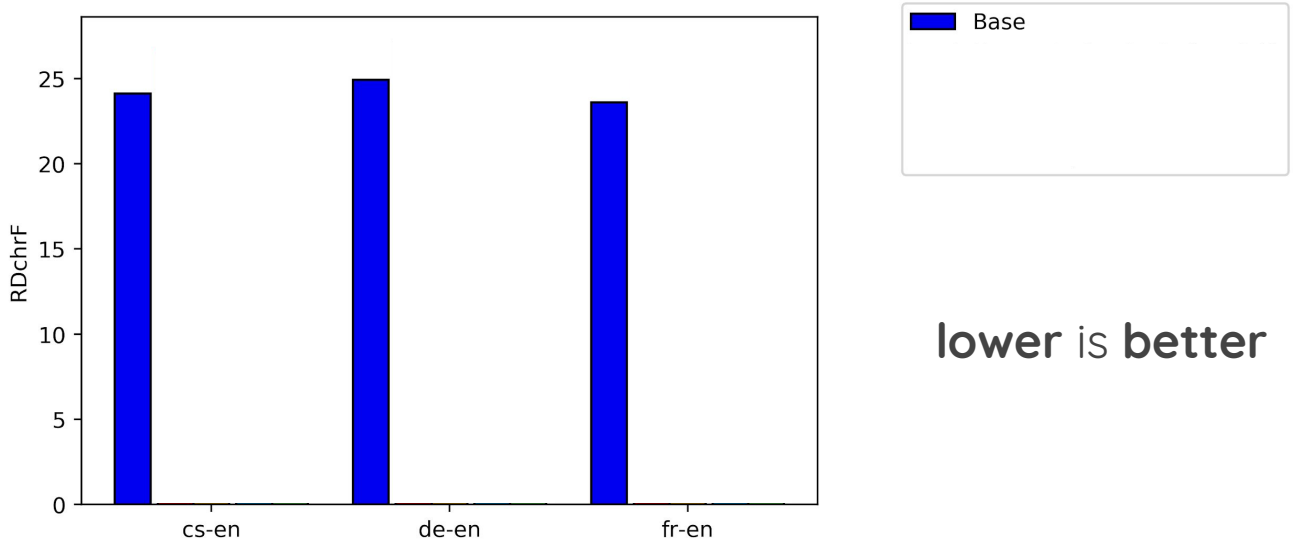
Adversarial input

  - α = 0: Standard training
  - α = 1 : Training only on adversarial examples

- Training with **Unconstrained** attacks vs **CharSwap** attacks
- Evaluate on
  - robustness to **CharSwap** attacks
  - Accuracy on **non-adversarial** data

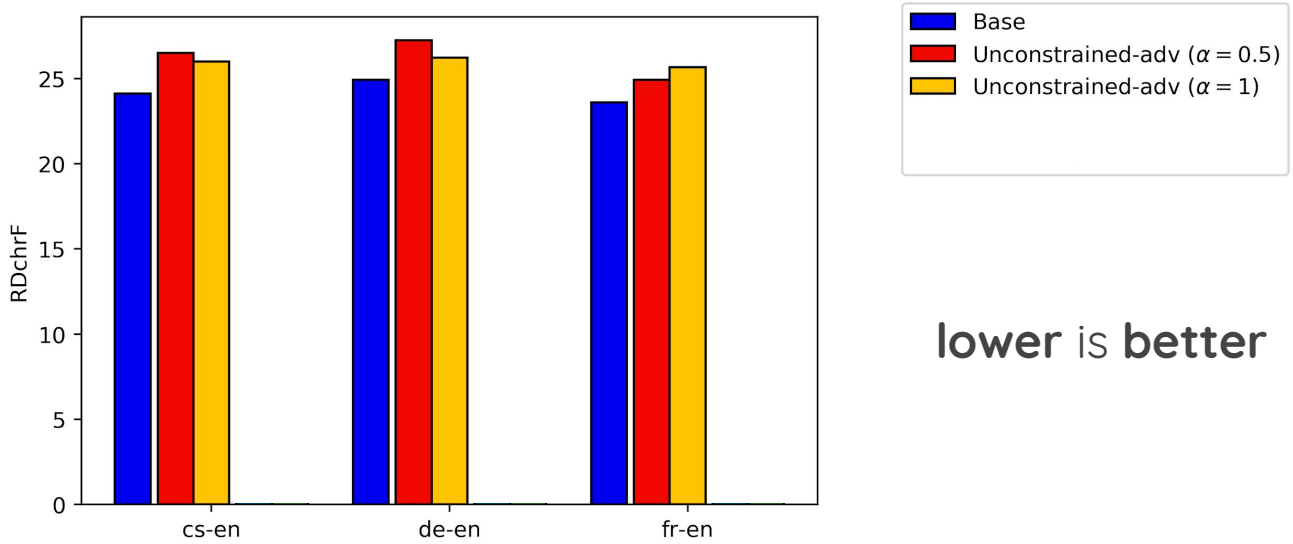
Effect of Constraints on Adversarial Training: Adversarial Robustness

- Robustness to CharSwap attacks on the validation set



Effect of Constraints on Adversarial Training: Adversarial Robustness

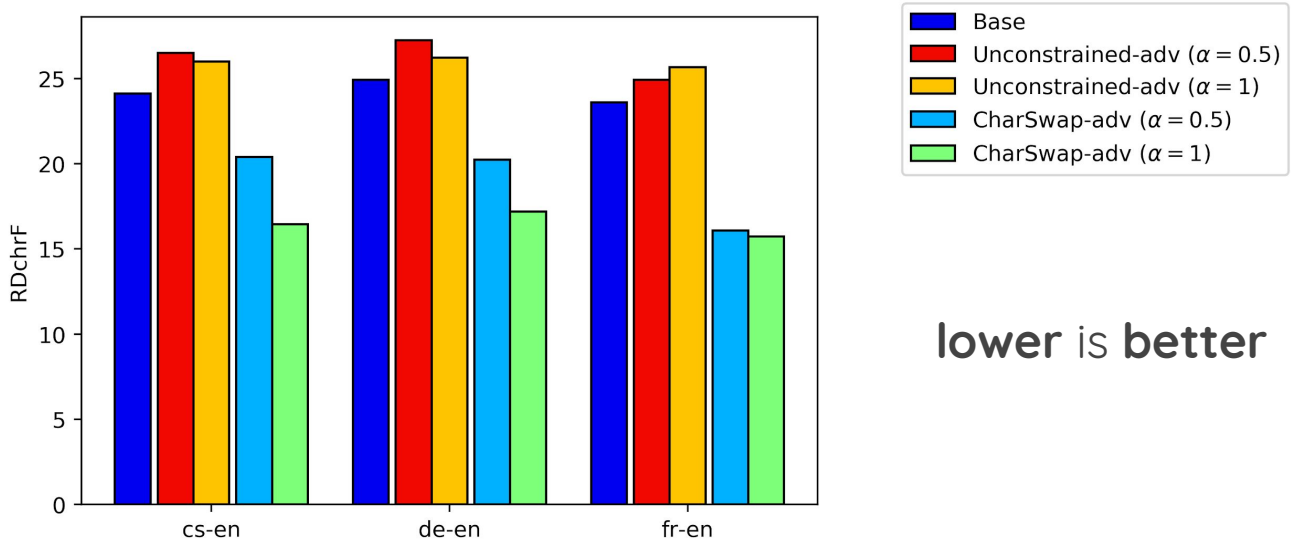
- Robustness to CharSwap attacks on the validation set



lower is **better**

Effect of Constraints on Adversarial Training: Adversarial Robustness

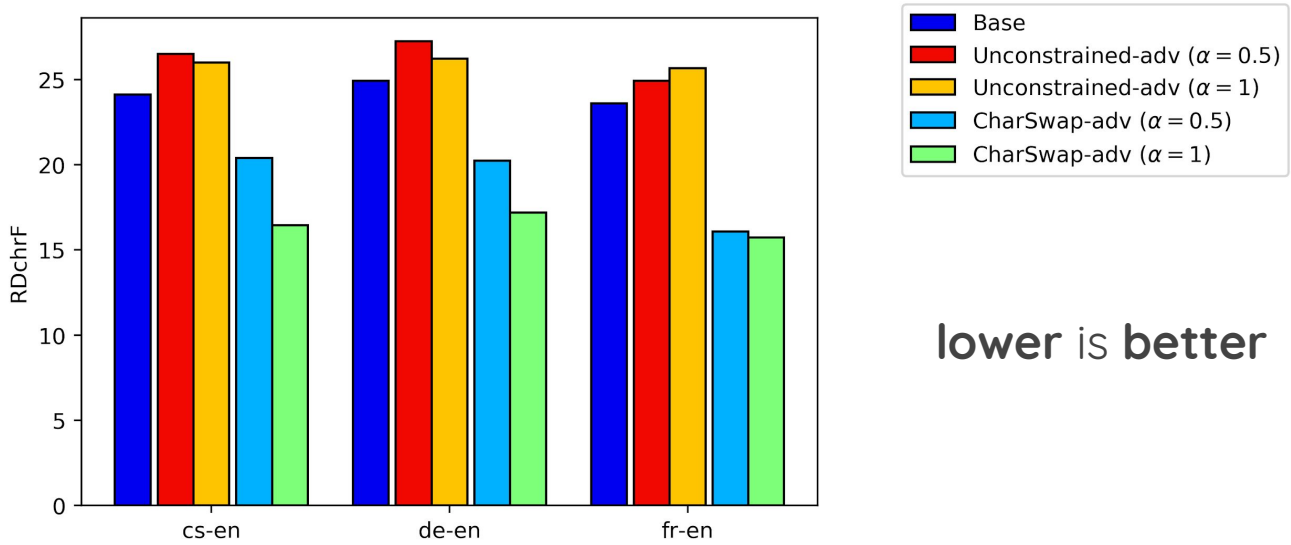
- Robustness to CharSwap attacks on the validation set



lower is **better**

Effect of Constraints on Adversarial Training: Adversarial Robustness

- Robustness to CharSwap attacks on the validation set

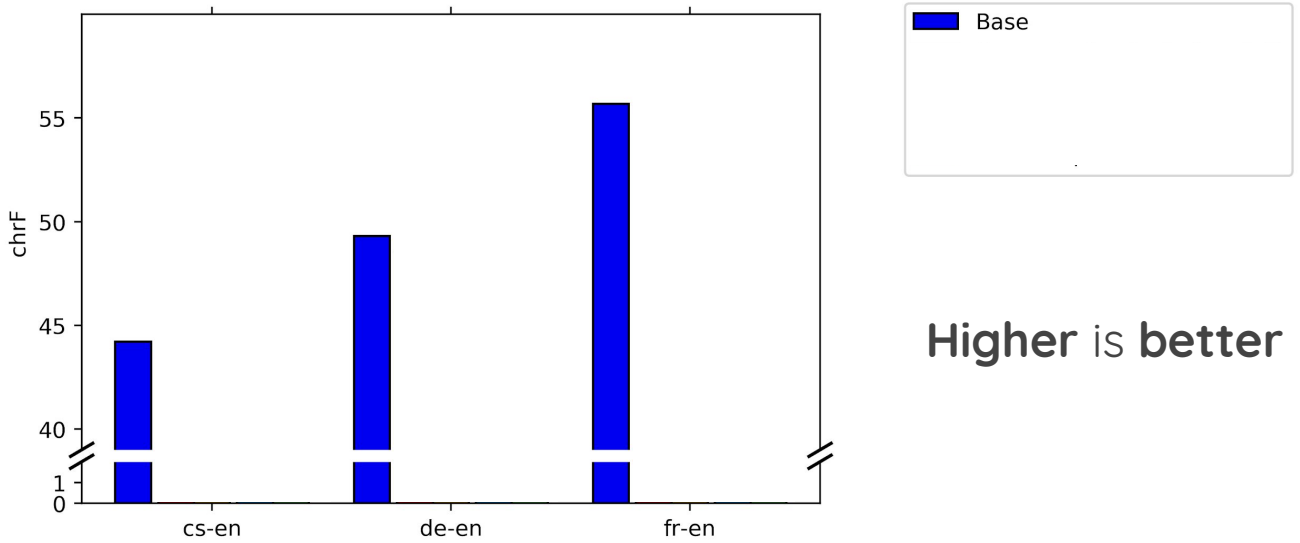


lower is **better**

- Adversarial training ⇒ **better robustness**

Effect of Constraints on Adversarial Training: Accuracy on Non-Adversarial Input

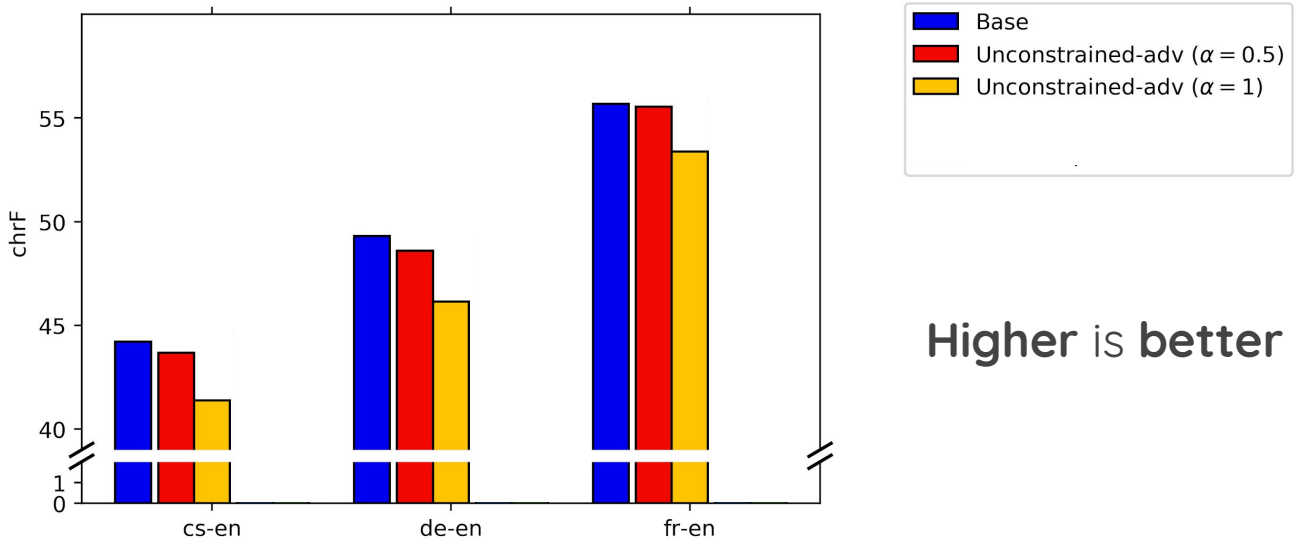
- Target chrF on the original test set



Higher is **better**

Effect of Constraints on Adversarial Training: Accuracy on Non-Adversarial Input

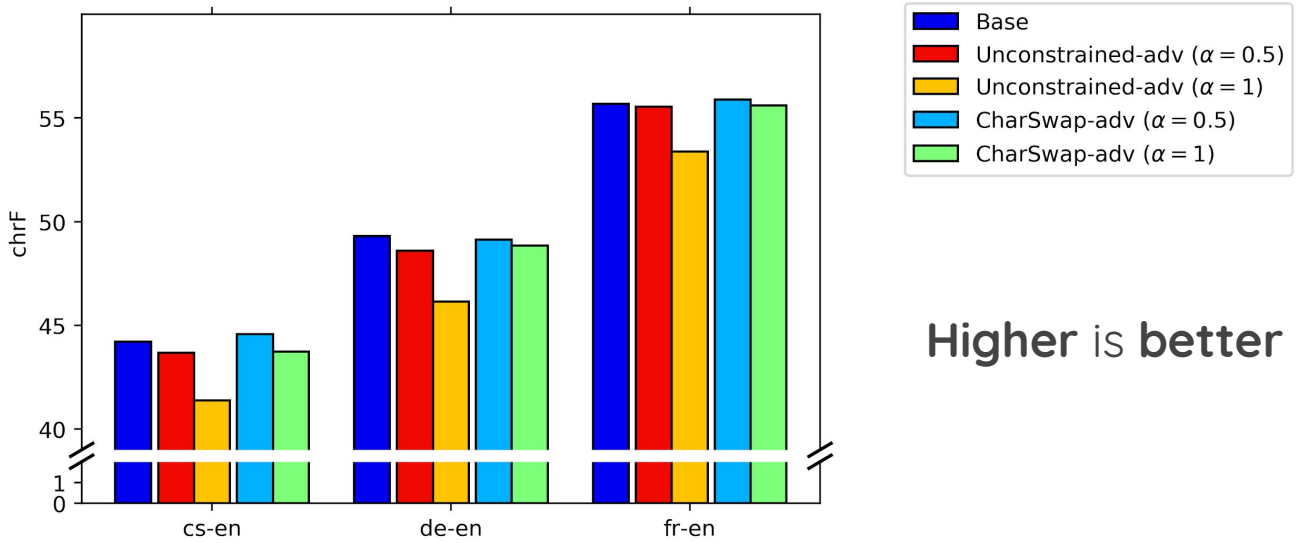
- Target chrF on the original test set



Higher is **better**

Effect of Constraints on Adversarial Training: Accuracy on Non-Adversarial Input

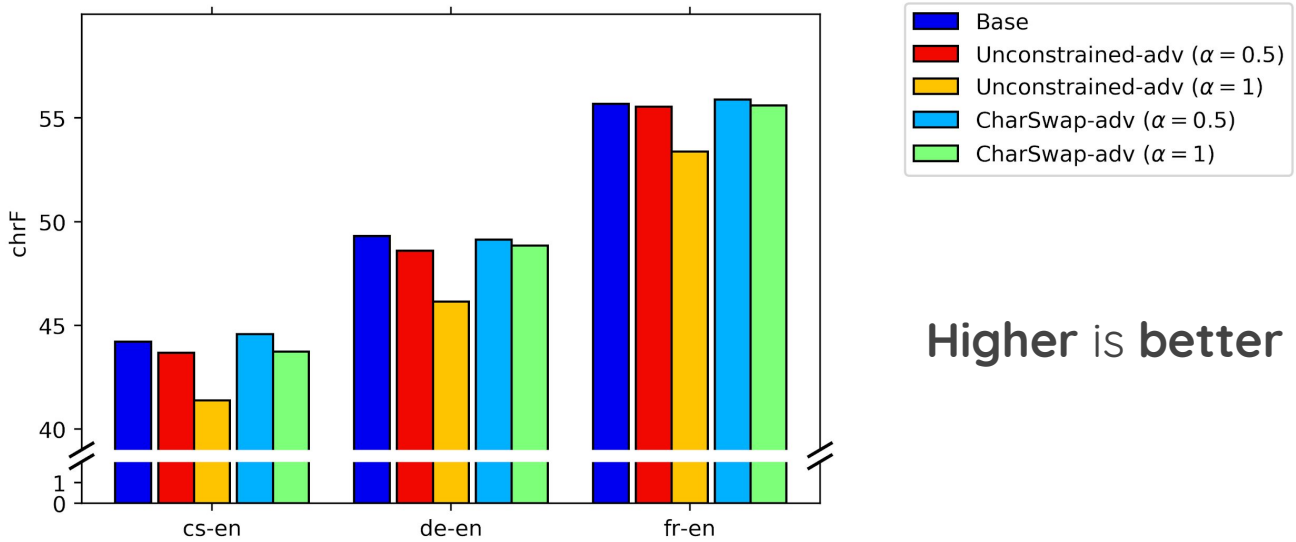
- Target chrF on the original test set



Higher is **better**

Effect of Constraints on Adversarial Training: Accuracy on Non-Adversarial Input

- Target chrF on the original test set

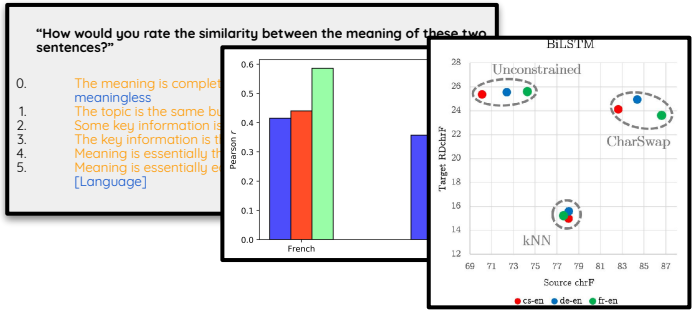


Higher is **better**

- Unconstrained attacks ⇒ **hurts accuracy**

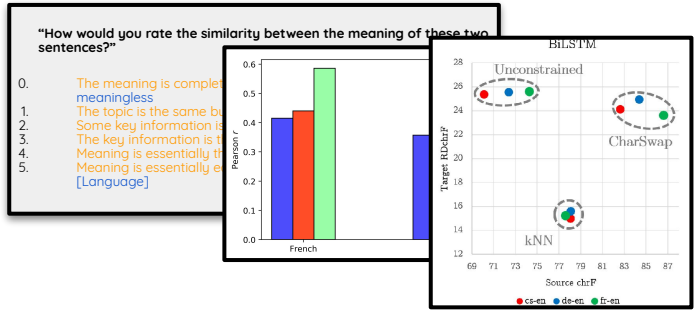
Takeway

- When doing adversarial **attacks**
  - Evaluate meaning preservation on the source side

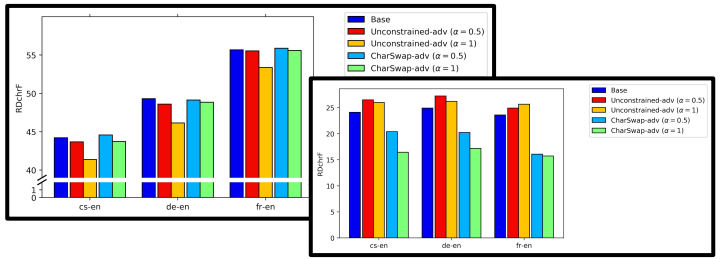


Takeway

- When doing adversarial **attacks**
  - Evaluate meaning preservation on the source side

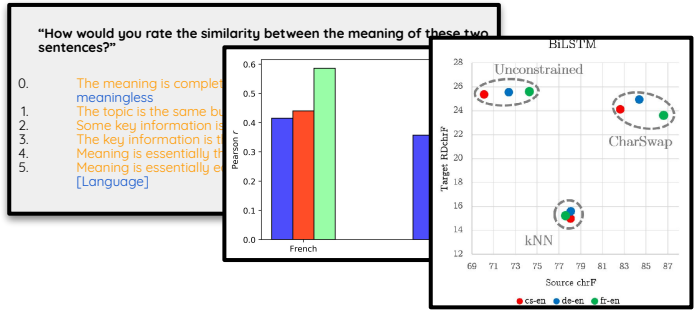


- When doing adversarial **training**
  - Consider adding constraints to your attacks

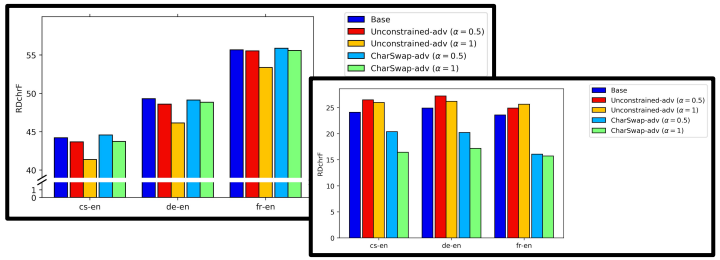


Takeway

- When doing adversarial **attacks**
  - Evaluate meaning preservation on the source side



- When doing adversarial **training**
  - Consider adding constraints to your attacks



- Not only true for seq2seq!
  - Easily transposed to classification, etc..
  - Just adapt s_{src} and s_{tgt} accordingly

TEAPOT

- Tool implementing our evaluation framework
- `pip install teapot-nlp`
- github.com/pmichel31415/teapot-nlp

```
teapot \
  --src examples/MT/src.fr \
  --adv-src examples/MT/adv.charswap.fr \
  --out examples/MT/base.en \
  --adv-out examples/MT/adv.charswap.en \
  --ref examples/MT/ref.en
```

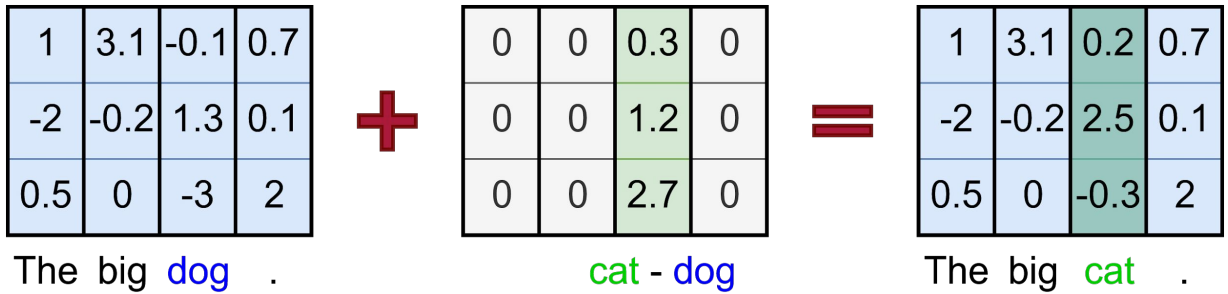
will output:

```
Source side preservation (ChrF):
Mean:   86.908
Std:    11.622
5%-95%: 64.109-97.683
----------------------------------------------
Target side degradation (ChrF):
Mean:   21.085
Std:    22.106
5%-95%: 0.000-67.162
----------------------------------------------
Success percentage: 65.20 %
```

Questions

Gradient Based Adversarial Attacks on Text

- Idea: Word substitution ⇔ Adding word vector difference



- Use the 1st order approximation to maximize the loss

$$\text{argmax}_w \mathcal{L}(x_i = v_w) - \mathcal{L}(x_i = v_{\text{dog}}) \approx \nabla_{x_i} \mathcal{L}^\intercal [v_w - v_{\text{dog}}]$$

Human Evaluation: the Gold Standard

Check for **semantic similarity** *and* **fluency**

"How would you rate the similarity between the meaning of these two sentences?"

0. The meaning is completely different or one of the sentences is meaningless
1. The topic is the same but the meaning is different
2. Some key information is different
3. The key information is the same but the details differ
4. Meaning is essentially the same but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed [Language]

Example of a Successful Attack

(source chrF = **80.89**, target RDchrF = **84.06**)

Original	Ils le réinvestissent directement en engageant plus de procès.
Adv. src.	Ilss le réinvestissent dierctement en engagaent plus de procès.
Ref.	They plow it right back into filing more troll lawsuits.
Base output	They direct it directly by engaging more cases.
Adv. output	.. de plus.

Example of an Unsuccessful Attack

(source chrF = **54.46**, target RDchrF = **0.00**)

Original	C’était en Juillet 1969.
Adv. src.	C’étiat en Jiullet 1969.
Ref.	This is from July, 1969.
Base output	This was in July 1969.
Adv. output	This is. in 1969.