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

Paul Michel, Xian Li, Graham Neubig, Juan Miguel Pino
Language Technologies Institute, Carnegie Mellon University
Facebook AI

Presentation by Jack Morris

https://qdata.github.io/deep2Read/

Background

- Adversarial attacks for NLP are not well defined

  “In discrete spaces such as natural language sentences, the situation is problematic; even a flip of a single word or character is generally perceptible by a human reader. Thus, most of the mathematical framework in previous work is not directly applicable to discrete text data.”

Background

- Adversarial attacks for NLP are not well defined

  “Moreover, there is no canonical distance metric for textual data like the `p norm in real-valued vector spaces such as images, and evaluating the level of semantic similarity between two sentences is a field of research of its own.”

This elicits a natural question: **what does the term “*adversarial perturbation*” mean in the context of natural language processing**?

Their proposal for NLP adversarial perturbations

"adversarial examples should be meaning-preserving on the source side, but meaning-destroying on the target side"

Idea: we should balance "meaning preserving" with "meaning destroying"

Basically: any meaning-preserving perturbation that results in the model output changing drastically highlights a fault of the model.

First, how can we compare similarity?

$$\mathrm{d}_{\mathrm{tgt}}(y, y_M, \hat{y}_M) = \begin{cases} 0 \text{ if } \mathrm{s}_{\mathrm{tgt}}(y, \hat{y}_M) \geq \mathrm{s}_{\mathrm{tgt}}(y, y_M) \\ \frac{\mathrm{s}_{\mathrm{tgt}}(y, y_M) - \mathrm{s}_{\mathrm{tgt}}(y, \hat{y}_M)}{\mathrm{s}_{\mathrm{tgt}}(y, y_M)} \text{ otherwise} \end{cases} \quad (1)$$

The "**target relative score decrease**", calculated as a function of the original translation "meaning similarity" and the new translation

Then measure attacker success like this:

$S := s_{src} + d_{tgt}$

Success = (amount source meaning preserved) + (amount target meaning destroyed)

TLDR: incorporate meaning preservation into loss function

So now we know how to calculate similarity.

Human evaluation is the best way to evaluate similarity in NLP. However, using human evaluation for all samples is not time or cost effective.

There are different computable scores for sentence similarity, like **BLEU** and **METEOR**.

But which should we use?

Idea

Give humans a quiz on meaning similarity.

Then, see which metric scores most similarly to human judgement.

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 equal but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed English[a]

[a] Or the language of interest.

chrF is best

Researchers compared BLEU, METEOR, and chrF scores as they correlated with human judgement.

chrF won, and it wasn't particularly close.

Language	BLEU	METEOR	chrF
French	0.415	0.440	**0.586***
English	0.357	0.478*	**0.497**

A general framework for adversarial attacks

$$\underset{1 \leq i \leq n, \hat{w} \in \mathcal{V}}{\arg\max} \; \mathcal{L}_{\text{adv}}(w_0, \ldots, w_{i-1}, \hat{w}, w_{i+1}, \ldots, w_n)$$

$$\mathcal{L}_{\text{adv}}(\hat{x}, y) = \sum_{t=1}^{|y|} \log(1 - p(y_t \mid \hat{x}, y_1, \ldots, y_{t-1}))$$

A general framework for adversarial attacks

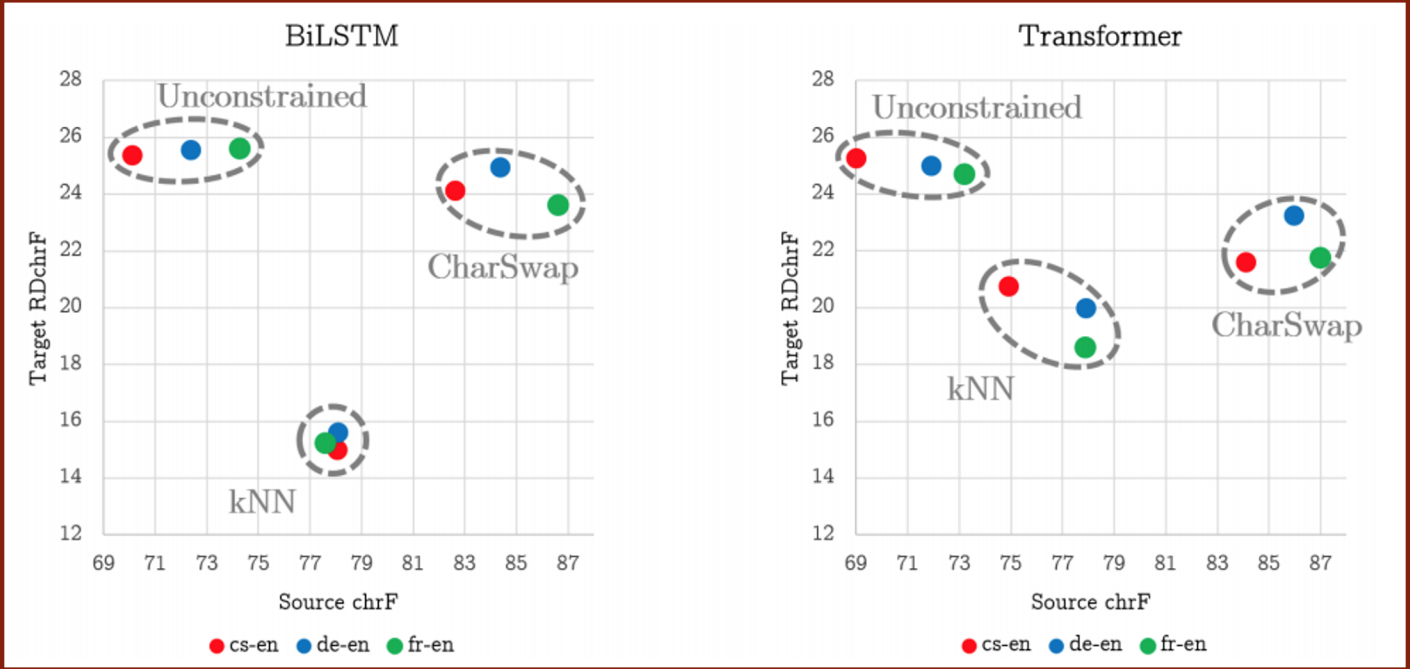
Three methods to test:

- Brute force. Turns out that this isn't as slow as you'd think. It's O(n|V|) where n is the number of words in the sentence and V is the vocabulary size. Not too bad for calculating a single example.

- K-Nearest-Neighbor. This is very similar to our work. They took the 10 closest words to each word in the embedding space and tried those.

- CharSwap. Also similar to our work! Swap two characters in a word such that the new word is out-of-vocabulary. (If you cannot find such a word, repeat the last character until you have found one.)

Results


BiLSTM
Unconstrained
CharSwap
kNN
Target RDchrF
Source chrF
28
26
24
22
20
18
16
14
12
69
71
73
75
77
79
81
83
85
87
cs-en
de-en
fr-en
Transformer
Unconstrained
kNN
CharSwap
Target RDchrF
Source chrF
cs-en
de-en
fr-en

Bonus:
Adversarial Training with Meaning-Preserving Attacks

- Use adversarial loss function (Goodfellow, 2014):

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

- Previous research (Ebrahimi, 2018) suggested that adversarial training improves robustness but hurts test performance

- They tested using “unconstrained” adversarial examples, and with CharSwap

- Both increased robustness. Unconstrained AEs did decrease performance on test set. But CharSwap test set performance increased!