# Robustness in NLP

## -- A short survey

**Presenter:** Shilin HE

2018-11-27

# Motivation

(a) Input 1 (b) Input 2 (darker version of 1)

Robustness of Neural Networks in Computer Vision is a hot topic!

People from AI/Security/SE all rush into this area, e.g., NIPS, ICML, ICLR, CVPR, CCS, IEEE S&P, NDSS, SOSP, ICSE, FSE

*Robustness are not well studied in NLP*

# Motivating Example

**Palestinian man is arrested by police after posting 'Good morning' in Arabic on Facebook which was wrongly translated as 'attack them'**

- Man uploaded the picture of himself leaning against the bulldozer in West Bank
- In the caption on Facebook, he wrote an Arabic term meaning 'good morning'
- A malfunction translated it to 'attack them' in Hebrew or 'hurt them' in English
- Police believed he was plotting an attack, so they swooped in and arrested him

By GARETH DAVIES FOR MAILONLINE
**PUBLISHED:** 11:03 GMT, 22 October 2017 | **UPDATED:** 11:35 GMT, 25 October 2017

Share 43 shares 15 View comments

Israeli police mistakenly arrested a Palestinian who posted 'good morning' in Arabic online which Facebook wrongly translated as 'attack them'.

The man uploaded a picture of himself leaning against a bulldozer at the Israeli settlement of Beitar Ilit, where he works, in the occupied West Bank.

Facebook 翻譯錯誤導致一名建築工人被抓，機器翻譯到底有多脆弱？

2017/11/20



# Motivation

Robustness of Neural Networks in NLP is not well-studied

- Semantic Preserving is **not clear,** i.e., paraphrase?
- Discrete Representation **fails** gradient based methods
- Difficulty of Generation Task

# Overview

## Attack

**Modify** a given input or **Generate** from noise

i.e., Optimisation Methods, Sensitive Features, Geometric Transformations

**Non-targeted** random, miss-classification or **Targeted** miss-classification

**white-box**, **grey-box** and **black-box** attacks

## Defense

**Reactive defences**

Detection of adversarial examples and input transformations (domain-specific)

**Obfuscation defences**

e.g., Obfuscate sensitive features

**Proactive defences**

Build natively robust models, e.g., **adversarial training**

# Papers so far

[EMNLP17] Adversarial examples for evaluating reading comprehension systems
[Arxiv18] Robust Neural Machine Translation with Joint Textual and Phonetic Embedding
[Arxiv18] Improving the Robustness of Speech Translation
[ICLR18] Synthetic and natural noise both break neural machine translation
[EMNLP18] Generating natural language adversarial examples
[ICLR18_Reject] Adversarial Examples for Natural Language Classification Problems
[ACL18] Towards Robust Neural Machine Translation
[NAACL18] Adversarial Example Generation with Syntactically Controlled Paraphrase Networks
[ACL18] HotFlip: White-Box Adversarial Examples for Text Classification [Short]
[Arxiv18] Detecting egregious responses in neural sequence-to-sequence models
[ACL18] Did the model understand the question
[ACL18] Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates
[ACL18]Trick Me If You Can: Adversarial Writing of Trivia Challenge Questions [Student Research Workshop]
[IJCAI18] Interpretable Adversarial Perturbation in Input Embedding Space for Text
[COLING18] On Adversarial Examples for Character-Level Neural Machine Translation
[EMNLP18] SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation
[ACL18] Know what you don't know: understanding questions for SQuAD
[CONLL18] Adversarial Over-Sensitivity and Over-Stability Strategies for Dialogue Models
[ACL17] adversarial learning for neural dialogue generation
[ICLR17] Adversarial Training Methods for Semi-Supervised Text Classification
[Arxiv17] Towards Crafting Text Adversarial Samples
[MILCOM16] Crafting adversarial input sequences for recurrent neural networks

# Summary

- Most methods are black-box
- Heuristic-based rules dominate this area, very simple
- Attack and Defences are half-half, I think defence is more
- Most defence methods are adversarial training
- Most methods show that adversarial training does not help on clean test data
- More difficulties in NLP than CV due to DISCRETE

# Translation Related

**[ICLR18] Synthetic and natural noise both break neural machine translation**

**[Arxiv18] Robust Neural Machine Translation with Joint Textual and Phonetic Embedding**

**[Arxiv18] Improving the Robustness of Speech Translation**

**[ACL18] Towards Robust Neural Machine Translation**

[COLING18] On Adversarial Examples for Character-Level Neural Machine Translation

[EMNLP18] SwitchOut: an Efficient Data Augmentation Algorithm for Neural Machine Translation

# Approach

[ICLR18] Synthetic and natural noise both break neural machine translation

**Yonatan Belinkov***
Computer Science and
Artificial Intelligence Laboratory,
Massachusetts Institute of Technology
belinkov@mit.edu

**Yonatan Bisk***
Paul G. Allen School
of Computer Science & Engineering,
University of Washington
ybisk@cs.washington.edu

"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."

↓ Google Translate
(German -> English)

"After being stubbornly defiant, it is clear to kenie Rlloe in which Reiehnfogle is advancing the boulders in a Wrot that is integral to Sahce, as the utterance and the lukewarm boorstbaen stmimt."

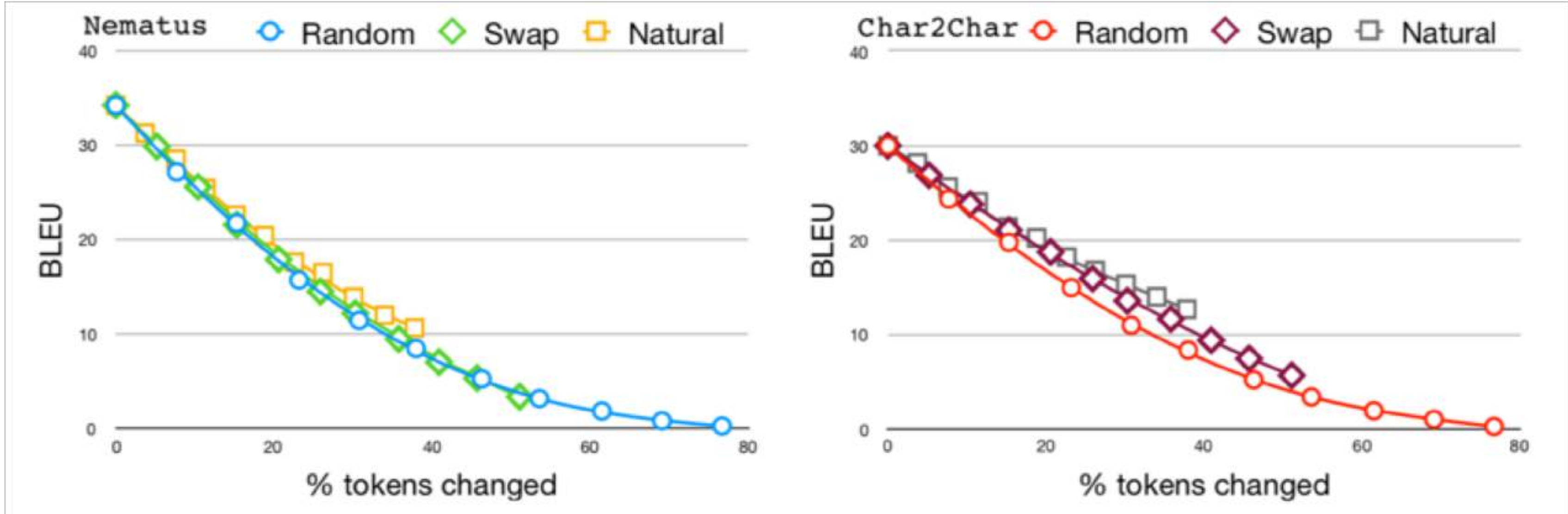



Figure 1: Degradation of Nematus (Sennrich et al., 2017) and char2char (Lee et al., 2017) performance as noise increases.

Char-based NMT, propose two types of black-box attack methods:

1. Natural Noise
2. Synthetic Noise

- **Natural Noise** --- naturally occurring errors (typos, misspellings, etc.)

  Wiki edit history, replace words in source sentences

- **Synthetic Noise** ---- Swap, Middle Random, Fully Random, and Keyboard Typo

1. Swap

   Swap two letters except the first and last letters (noise -> nosie)
2. Middle Random

   Randomize all the letters in a word except for the first and last (noise→nisoe)
3. Fully Random

   Completely randomized words, does not consider first/last letter (*noise*→*iones*)
4. Keyboard Typo

   Replace one letter in each word with an adjacent key (*noise*→*noide*).

Table 3: The effect of Natural (`Nat`) and synthetic noise (Swap `swap`, Middle Random `Mid`, Fully Random `Rand`, and Keyboard Typo `Key`) on models trained on clean (Vanilla) texts.

| | | | Synthetic | | | | |
|---|---|---|---|---|---|---|---|
| | | Vanilla | `Swap` | `Mid` | `Rand` | `Key` | `Nat` |
| French | `charCNN` | 42.54 | 10.52 | 9.71 | 1.71 | 8.26 | 17.42 |
| German | `charCNN` | 34.79 | 9.25 | 8.37 | 1.02 | 6.40 | 14.02 |
| | `char2char` | 29.97 | 5.68 | 5.46 | 0.28 | 2.96 | 12.68 |
| | `Nematus` | 34.22 | 3.39 | 5.16 | 0.29 | 0.61 | 10.68 |
| Czech | `charCNN` | 25.99 | 6.56 | 6.67 | 1.50 | 7.13 | 10.20 |
| | `char2char` | 25.71 | 3.90 | 4.24 | 0.25 | 2.88 | 11.42 |
| | `Nematus` | 29.65 | 2.94 | 4.09 | 0.66 | 1.41 | 11.88 |

Can we use spell checkers to fix the problem?

Table 5: Google Translate's performance with natural errors and the gains from using spell checking.

| French | | | German | | | Czech | | |
|---|---|---|---|---|---|---|---|---|
| Vanilla | `Nat` | Spelling | Vanilla | `Nat` | Spelling | Vanilla | `Nat` | Spelling |
| 43.3 | 16.7 | 21.4 | 38.7 | 18.6 | 25.0 | 26.5 | 12.3 | 11.2 |

natural noise cannot be easily addressed by existing tools

## ❖ Structure Invariant Representation

take the average character embedding as a word representation
for character scrambling (Swap, Mid, and Rand)

meanChar:

1. generate a word representation by averaging character embeddings
2. proceeds with a word-level encoder similar to the charCNN model.

## ❖ Black-box Adversarial Training

replace the original training set with a noisy training set, exactly same size

Table 6: Results of meanChar models trained and tested on different noise conditions: Scrambled (Scr), Keyboard Typo (Key), and Natural (Nat).

| Train \ Test | French Scr | French Key | French Nat | German Scr | German Key | German Nat | Czech Scr | Czech Key | Czech Nat |
|---|---|---|---|---|---|---|---|---|---|
| Vanilla | 34.26 | 4.27 | 12.58 | 27.53 | 3.34 | 9.41 | 3.73 | 2.06 | 3.25 |
| Key | 31.88 | 29.75 | 13.16 | 10.04 | 8.84 | 4.45 | 2.03 | 1.9 | 1.42 |
| Nat | 26.94 | 5.30 | 27.49 | 15.65 | 3.06 | 26.26 | 1.66 | 1.52 | 1.58 |
| Rand + Key | 13.60 | 11.09 | 6.12 | 26.59 | 22.41 | 11.07 | 9.97 | 7.48 | 4.21 |
| Rand + Nat | 28.28 | 5.10 | 20.40 | 13.87 | 3.73 | 12.74 | 4.89 | 2.82 | 3.42 |
| Key + Nat | 31.30 | 26.94 | 24.24 | 6.62 | 5.41 | 5.75 | 1.62 | 1.68 | 1.58 |
| Rand + Key + Nat | 3.10 | 3.28 | 2.76 | 8.02 | 5.79 | 6.36 | 1.73 | 1.74 | 1.66 |

Table 7: Results of charCNN models trained and tested on different noise conditions.

| | Train \ Test | Vanilla | Swap | Mid | Rand | Key | Nat | Ave |
|---|---|---|---|---|---|---|---|---|
| French | Swap | 39.01 | **42.56** | 33.64 | 2.72 | 4.85 | 16.43 | 23.20 |
| | Mid | **42.46** | 42.19 | **42.17** | 3.36 | 6.20 | 18.22 | 25.77 |
| | Rand | 39.53 | 39.46 | 39.13 | **39.73** | 3.11 | 16.63 | 29.60 |
| | Key | 38.49 | 10.56 | 8.69 | 1.08 | **38.88** | 16.86 | 19.10 |
| | Nat | 28.77 | 12.45 | 8.39 | 1.03 | 6.61 | **36.00** | 15.54 |
| | Rand + Key | 39.23 | 38.85 | 38.89 | 39.13 | 38.22 | 18.71 | 35.51 |
| | Rand + Nat | 36.86 | 38.95 | 38.44 | 38.63 | 6.67 | 33.89 | 32.24 |
| | Key + Nat | 38.47 | 17.33 | 10.54 | 1.52 | 38.62 | 34.66 | 23.52 |
| | Rand + Key + Nat | 36.97 | 36.92 | 36.65 | 36.64 | 35.25 | 31.77 | **35.70** |

**One important conclusion:**

**Training on noisy data** does not necessarily improve the **accuracy on clean test data**

**One Question:**

Why not append noisy training data to the original training set?

# Arxiv 18 (homophone)

**Robust Neural Machine Translation with Joint Textual and Phonetic Embedding**

**Hairong Liu**[1] **Mingbo Ma**[1,3] **Liang Huang**[1,3] **Hao Xiong**[2] **Zhongjun He**[2]
[1]Baidu Research, Sunnyvale, CA, USA
[2]Baidu, Inc., Beijing, China
[3]Oregon State University, Corvallis, OR, USA
{liuhairong, mingboma, lianghuang, xionghao05, hezhongjun}@baidu.com

Oct 15, 2018

**Improving the Robustness of Speech Translation**

**Xiang Li**[1,2*], **Haiyang Xue**[1,2,*†], **Wei Chen**[3], **Yang Liu**[4], **Yang Feng**[1,2], **Qun Liu**[5]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Voice Interaction Technology Center, Sogou Inc., Beijing
[4]Department of Computer Science and Technology, Tsinghua University, Beijing
[5]Huawei Noah's Ark Lab, Huawei Technologies, Hong Kong
lixiang@ict.ac.cn, xuehaiyang@ict.ac.cn, chenweibj8871@sogou-inc.com,
liuyang2011@tsinghua.edu.cn, fengyang@ict.ac.cn, qun.liu@huawei.com

Nov 2, 2018

# Arxiv 18 (homophone)

| | |
|---|---|
| Clean Input | 目前已发现有109人死亡, 另有57人获救 |
| Output of Transformer | at present, 109 people have been found dead and 57 have been rescued |
| Noisy Input | 目前已发现又109人死亡, 另有57人获救 |
| Output of Transformer | the hpv has been found dead so far and 57 have been saved |
| Output of Our Method | so far, 109 people have been found dead and 57 others have been rescued |

| | |
|---|---|
| **Speech** | zhè fèn lǐ wù bǎo hán yī fèn shēn qíng<br>这 份 礼 物 饱 含 一 份 深 情 |
| **ASR** | zhè fèn lǐ wù bǎo hán yī fèn shēn qǐng<br>这 份 礼 物 饱 含 一 份 申 请 |
| **Reference** | This gift is full of affection |
| **NMT** | This gift contains an application |

# Arxiv 18 (homophone)

**Baidu's work:**

1. Embedding both words and pronunciation units
2. Average all l pronunciation units
3. Weighted sum both textual Phonetic

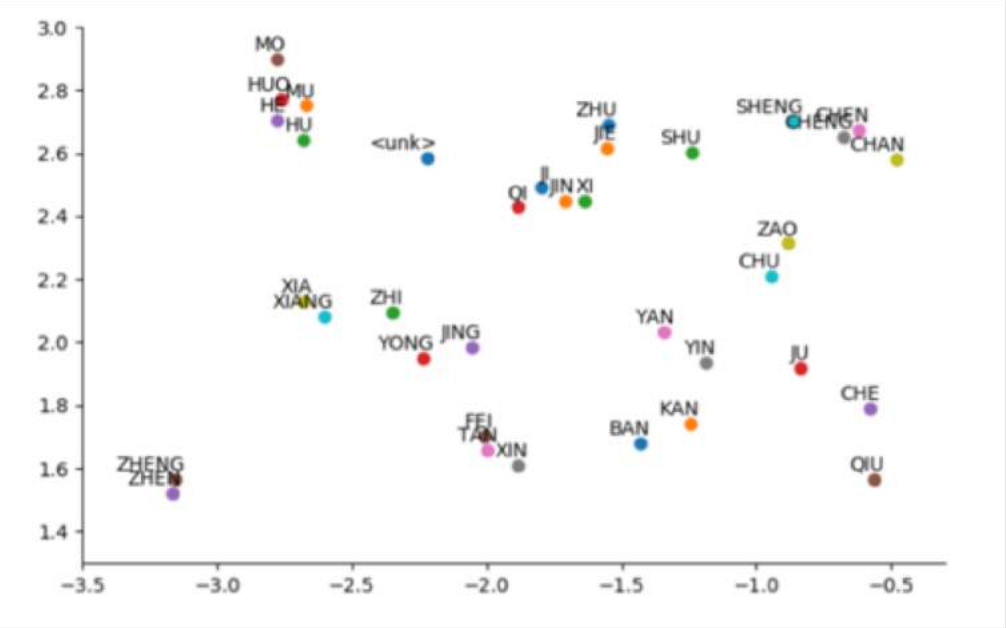
**Data Augmentation:**

Replace words in source sentences by homophones

# Arxiv 18 (homophone)

| Models | NIST06 (Dev Set) | NIST02 | NIST03 | NIST04 | NIST08 |
|---|---|---|---|---|---|
| Transformer-base | 45.97 | 47.40 | 46.01 | 47.25 | 41.71 |
| $\beta = 0.2$ | 47.14 | 48.63 | 47.82 | 48.63 | 43.77 |
| $\beta = 0.4$ | 48.56 | 49.41 | 48.73 | 50.53 | **45.16** |
| $\beta = 0.6$ | 48.32 | 48.83 | 48.82 | 49.86 | 44.17 |
| $\beta = 0.8$ | 48.15 | **49.42** | 49.44 | 49.98 | 44.86 |
| $\beta = 0.95$ | 48.91 | 49.33 | **50.46** | **50.57** | 44.83 |
| $\beta = 1.0$ | 45.6 | 47.04 | 46.42 | 47.65 | 40.27 |

Table 2: Translation results on NIST Mandarin-English test sets




**Do not need so many textual words?**

| Models | Before Augmentation | | | After Augmentation | | |
|---|---|---|---|---|---|---|
| | NIST06 | NoisySet1 | NoisySet2 | NIST06 | NoisySet1 | NoisySet2 |
| Transformer-base | 45.97 | 41.33 | 37.11 | 43.94 | 42.61 | 41.33 |
| $\beta = 0.95$ | **48.91** | **45.71** | **42.66** | **48.06** | **47.37** | **46.47** |

# Arxiv 18 (homophone)

**Sogou’s work: 4 types of noises**

1. Placeholder-based Substitution
2. Uniform Distribution-based Substitution
3. Frequency-based Substitution
4. Homophone-based Substitution

| Methods | Noise Example | | | |
|---|---|---|---|---|
| Placeholder | yǔ<br>语 | <SUB> | fān<br>翻 | yì<br>译 |
| Uniform | yǔ<br>语 | tāo<br>饕 | fān<br>翻 | yì<br>译 |
| Frequency | yǔ<br>语 | hǎo<br>好 | fān<br>翻 | yì<br>译 |
| Homophone | yǔ<br>语 | yīn<br>因 | fān<br>翻 | yì<br>译 |

**Similar conclusions**

# Insights

Hard code: too many manual efforts, very heuristic

Recall the challenges:

- Semantic Preserving is **not clear,** i.e., paraphrase?
- Discrete Representation **fails** gradient based methods
- Difficulty of Generation Task

More model-based methods will be introduced next week!

# THANKS