Evaluating Robustness to Input Perturbations for Neural Machine Translation

Xing Niu, Prashant Mathur, Georgiana Dinu, Yaser Al-Onaizan
Amazon AI

What is the Problem?

- NMT models are brittle to small perturbations in the input.
  - An example of NMT English translations for a Finnish input and its one-letter misspelled version.

Original input	Se kyllä tuntuu sangen luultavalta.
Translation	It certainly seems very likely.
Perturbed input	Se kyllä tumtuu sangen luultavalta.
Translation	It will probably darken quite probably.
Reference	It certainly seems probable.

  - This model is not very **robust** to input perturbations (e.g., misspelling)

How to Evaluate Robustness?

- Previous work

Original input	Se kyllä tuntuu sangen luultavalta.
Translation	It certainly seems very likely.
Perturbed input	Se kyllä tumtuu sangen luultavalta.
Translation	It will probably darken quite probably.
Reference	It certainly seems probable.

Noisy Input

Score absolute model performance

- This is an appropriate measure for noisy domain evaluation.
- But it does not disentangle model quality from the relative degradation under added noise.

How to Evaluate Robustness?

- This work
  - We propose two additional measures for robustness.

Original input	Se kyllä tuntuu sangen luultavalta.
Translation	It certainly seems very likely.
Perturbed input	Se kyllä tumtuu sangen luultavalta.
Translation	It will probably darken quite probably.
Reference	It certainly seems probable.

1. relative changes in translation quality (reference needed)
2. consistency in translation output (reference free)

Evaluation Metrics

- **Robustness**

Original input		Se kyllä tuntuu sangen luultavalta.
Translation	y'	It certainly seems very likely.
Perturbed input		Se kyllä tumtuu sangen luultavalta.
Translation	y^*	It will probably darken quite probably.
Reference	y	It certainly seems probable.

TQ: translation quality, e.g., BLEU

$\mathtt{TQ}(y', y)$

$\mathtt{TQ}(y^*, y)$

$$\text{ROBUST} = \frac{\mathtt{TQ}(y^*, y)}{\mathtt{TQ}(y', y)}$$

Evaluation Metrics

- **Consistency**
  - estimating robustness without the reference

Original input	Se kyllä tuntuu sangen luultavalta.
Translation y'	It certainly seems very likely.
Perturbed input	Se kyllä tumtuu sangen luultavalta.
Translation y^*	It will probably darken quite probably.
Reference	It certainly seems probable.

Sim can be any symmetric measure of similarity, e.g., symmetric BLEU

$$\mathtt{Sim}(y^*, y')$$

$$\mathtt{CONSIS} = \mathtt{Sim}(y^*, y')$$

Set-Up

- Models to be compared -- (stochastic) subword segmentation strategies
  - BPE (Sennrich et al., 2016)
  - BPE-Dropout (Provilkov et al., 2019)
  - SentencePiece (Kudo, 2018)

subword regularization

- Perturbations:
  - Synthetic misspelling
  - Letter case changing

- Data:
  - General domains: perturbations are applied to test sets of WMT etc.
  - Noisy domains: MTNT (Michel and Neubig, 2018) and 4SQ (Berard et al., 2019)

*see our paper for details

Results (General Domains)

Model	EN→DE		DE→EN		EN→FR		FR→EN		EN→FI		FI→EN		EN→JA		JA→EN	
BPE	39.70	2	40.01	3	41.47	1	39.24	1	20.43	2	24.31	3	24.28	1	22.80	2
BPE-Dropout	39.65	3	40.16	2	40.72	3	39.22	2	20.01	3	24.51	2	24.11	2	22.21	3
SentencePiece	39.85	1	40.25	1	41.05	2	39.14	3	20.63	1	24.67	1	22.63	3	22.99	1

- There is no clear winner among the three subword segmentation models based on BLEU scores.
  - No input perturbations yet

Results (General Domains)

	Model	BLEU	ROBUST	CONSIS	BLEU	ROBUST	CONSIS
		EN→DE (newstest2019)			DE→EN (newstest2019)		
original	BPE	39.70	–	–	40.01	–	–
	BPE-Dropout	39.65	–	–	40.16	–	–
	SentencePiece	39.85	–	–	40.25	–	–
+ misspelling	BPE	29.38	74.01	60.59	33.48	83.69	71.51
	BPE-Dropout	33.13	83.55	70.74	35.97	89.58	78.33
	SentencePiece	31.87	79.99	66.40	35.26	87.61	74.09
+ case-changing	BPE	31.61	79.63	73.26	33.72	84.27	73.19
	BPE-Dropout	35.04	88.37	80.04	36.34	90.48	78.96
	SentencePiece	33.49	84.05	76.24	34.48	85.65	74.55

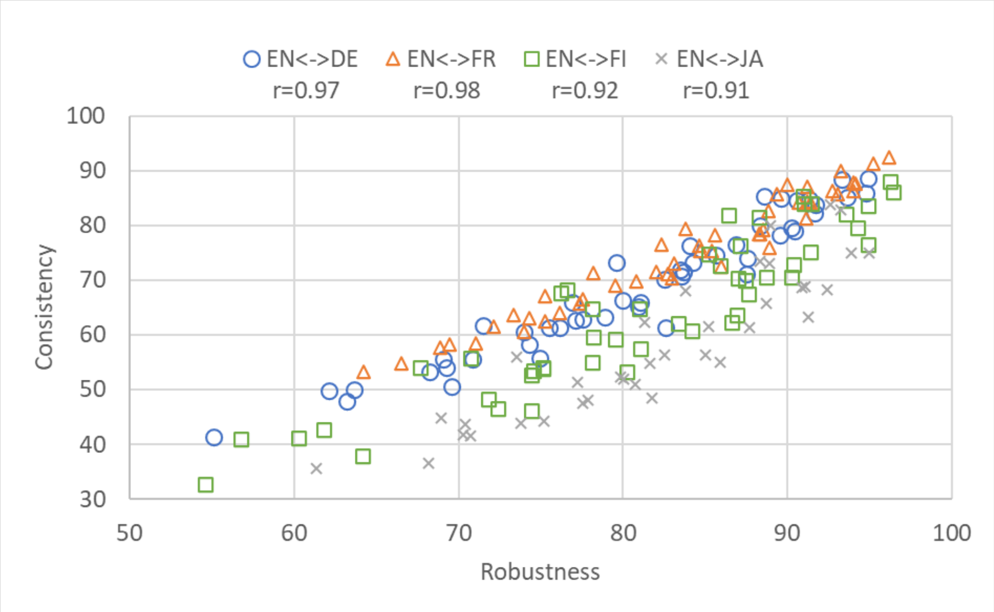
- ROBUST and CONSIS show clear and the same trend of models' robustness to input perturbations*
  - BPE-Dropout > SentencePiece > BPE

* across all languages we tested: EN<->DE, EN<->FR, EN<->FI, EN<-> JA. Please refer to the paper for complete results.

Robustness Versus Consistency

- Can we use **consistency** as a **robustness** proxy when the reference is unavailable?
  - Yes, at least for this class of models.

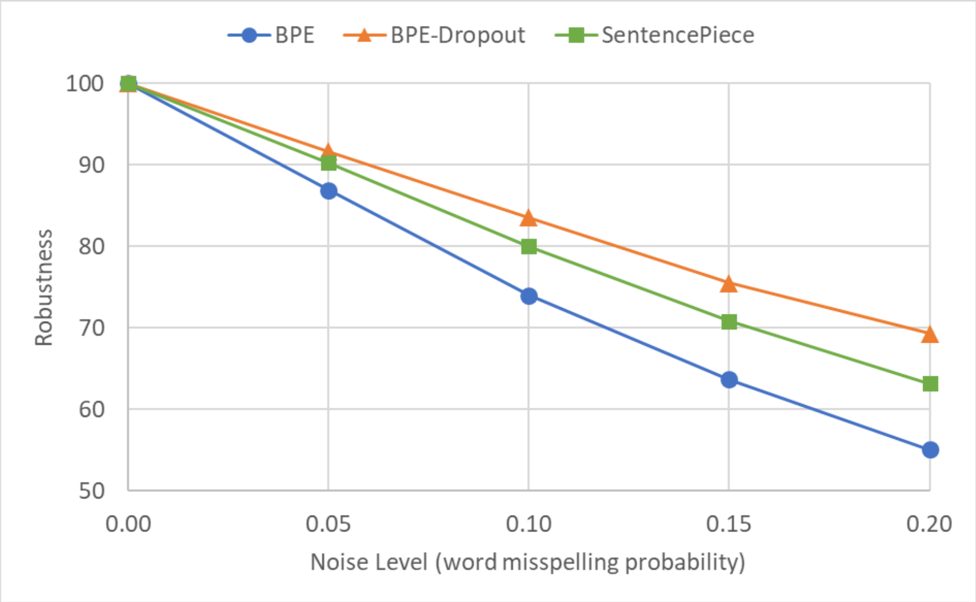
- **Consistency** strongly correlates with **Robustness**.
  - Data points are collected by varying the noise level of both perturbations.



Robustness Versus Noise Level

- Does the model ranking depend on the noisy level?
  - No.

- Varying the word misspelling probability does not change the ranking.
  - This observation applies to all language pairs and perturbations we investigated.



EN->DE

Summary

- We proposed two additional measures for NMT robustness.
  - Robustness: relative degradation in translation quality
  - Consistency: variation in translation output irrespective of reference translations

- We tested two popular subword regularization techniques.
  - Subword regularization is much more robust to synthetic input perturbations than standard BPE.
  - But it is unclear if subword regularization can help translating real-world noisy input. *see our paper for details

- We identified a strong correlation between robustness and consistency in these models.
  - Consistency can be used to estimate robustness on data sets or domains lacking reference translations.

Thank you!

Contact

Xing Niu

xingniu@amazon.com

Results (Noisy Domains)

	Model	MTNT (mtnt2019) EN→JA	JA→EN	EN→FR	FR→EN	4SQ FR→EN
baseline	BPE	$10.75_{\pm 0.49}$	$9.68_{\pm 0.59}$	$34.15_{\pm 0.93}$	$45.84_{\pm 0.89}$	$30.96_{\pm 0.85}$
	BPE-Dropout	$10.76_{\pm 0.47}$	$9.26_{\pm 0.64}$	$33.39_{\pm 0.95}$	$45.84_{\pm 0.90}$	$31.28_{\pm 0.84}$
	SentencePiece	$10.52_{\pm 0.51}$	$9.52_{\pm 0.68}$	$33.75_{\pm 0.91}$	$45.94_{\pm 0.92}$	$31.44_{\pm 0.85}$
fine-tuning*	BPE	$14.88_{\pm 0.52}$	$10.47_{\pm 0.69}$	$35.11_{\pm 0.95}$	$46.49_{\pm 0.90}$	$34.83_{\pm 0.86}$
	BPE-Dropout	$15.26_{\pm 0.53}$	$11.13_{\pm 0.68}$	$34.80_{\pm 0.93}$	$46.88_{\pm 0.88}$	$34.72_{\pm 0.84}$
	SentencePiece	$14.68_{\pm 0.53}$	$11.19_{\pm 0.72}$	$34.71_{\pm 0.93}$	$46.89_{\pm 0.90}$	$34.59_{\pm 0.86}$

- It is unclear if subword regularization can help translating real-world noisy input.

* fine-tuning: continue training baseline models with corresponding MTNT/4SQ training data