A Study of Morphological Robustness of NMT

Sai Muralidhar Jayanthi, Adithya Pratapa
Language Technologies Institute
Carnegie Mellon University

Carnegie Mellon University

SIGMORPHON @ ACL 2021

An Example

Die Schießereien haben nicht aufgehort. —de → en→ The shootings have not stopped.

0.852

Die Schießereien habe nicht aufgehort. —de → en→ The shootings did not stop, he said.

0.513

Should a morphological perturbation lead to significant change of output translation?

Overview

- We propose a methodology to create adversarial perturbations of source language sentences
- Specific focus on **morphological** perturbations
- We study robustness of NMT systems on 11 language pairs from various languages families, with varying amount of parallel data

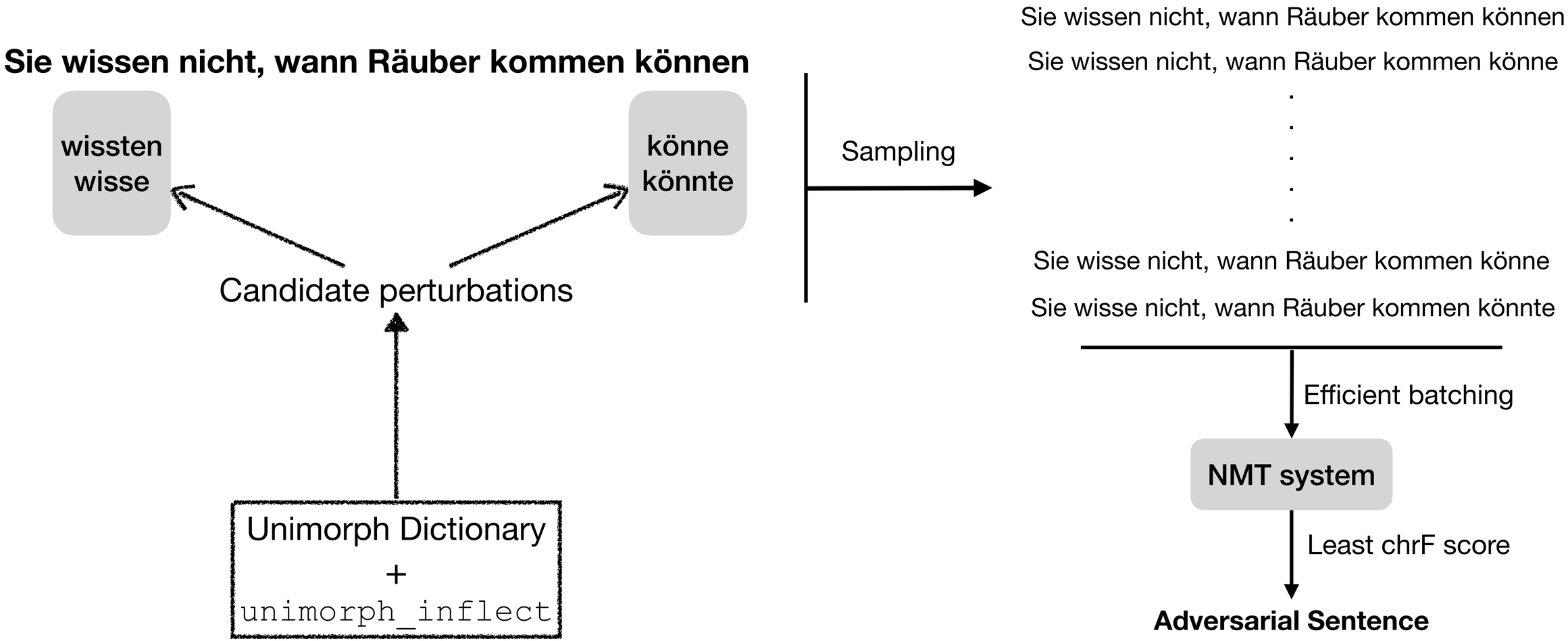
Prior Work

- Morpheus (Tan et al., 2020)
- Perturbs the inflectional morphology of source English tokens to craft adversarial sentences for English→French translation systems
- Utilizes `LemmInflect` tool to generate candidate inflectional forms

Extending to Multilingual NMT

- Need morphological reinflection models that work across many languages
- UniMorph (McCarthy et al., 2020): morphological data for >100 languages under a universal schema
- UniMorph has limited set of paradigms → `unimorph_inflect` to expand to more paradigms

Morpheus-*Multilingual*



Limitations

- Inflectional perturbations can lead to significant semantic change for morphologically-rich languages

Тренер полностью поддержал игрока. —ru → en→ The coach fully supported the player.

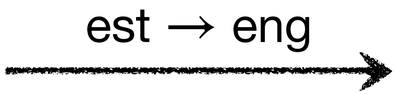
Тренера полностью поддержал **игрок**. —ru → en→ The coach was fully supported by the player.

Limitations

- Incorrect perturbations

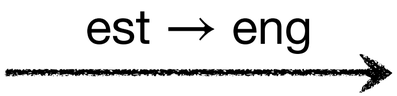
And stealing our children's future will one day be a crime.

Ning meie laste tuleviku varastamine saab ühel päeval kuriteoks.



And our children’s going to be the future of our own day.

Ning meie **laptegs** tuleviku varastamine saab ühel päeval kuriteoks.

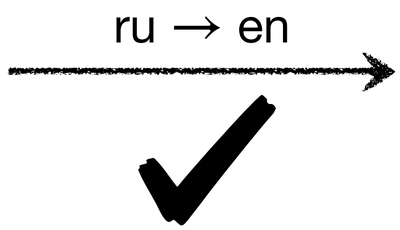


And our future is about the future of the future.

Limitations

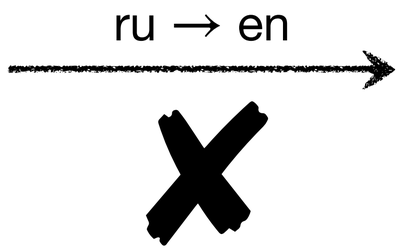
- No constraint imposed on the # perturbations per source sentence

Мой дедушка был необычайным человеком того времени.



My grandfather was an extraordinary man at that time.

Мой дедушка **будё необычайна человеков** того **времи**.



My grandfather is incredibly harmful.

Experiments

- Three experimental settings
  - Pretrained NMT systems (rus,deu)
  - models trained on TED corpus
  - translating learners' text (rus,deu)
- 11 language pairs
- Models trained used `fairseq`

Language	Family	Resource level
heb	Semetic	High
rus	Slavic	High
tur	Turkic	High
deu	Germanic	High
ukr	Slavic	High
Ces	Slavic	High
swe	Germanic	Medium
lit	Baltic	Medium
slv	Slavic	Low
kat	Kartvelian	Low
est	Uralic	Low

Evaluation Metrics

- Standard translation metrics: **BLEU**, **chrF** (character n-gram F score)
  - Limitation: perturbations in source can lead to valid proportional changes in the target
  - E.g. changing plurality of nouns in source
- Additional metric: **Noise Ratio**, $\mathrm{NR}(s, t, \tilde{s}, \tilde{t}) = \dfrac{100 - \mathrm{BLEU}(t, \tilde{t})}{100 - \mathrm{BLEU}(s, \tilde{s})}$

  (Anastasopoulos, 2019)

Pretrained NMT models

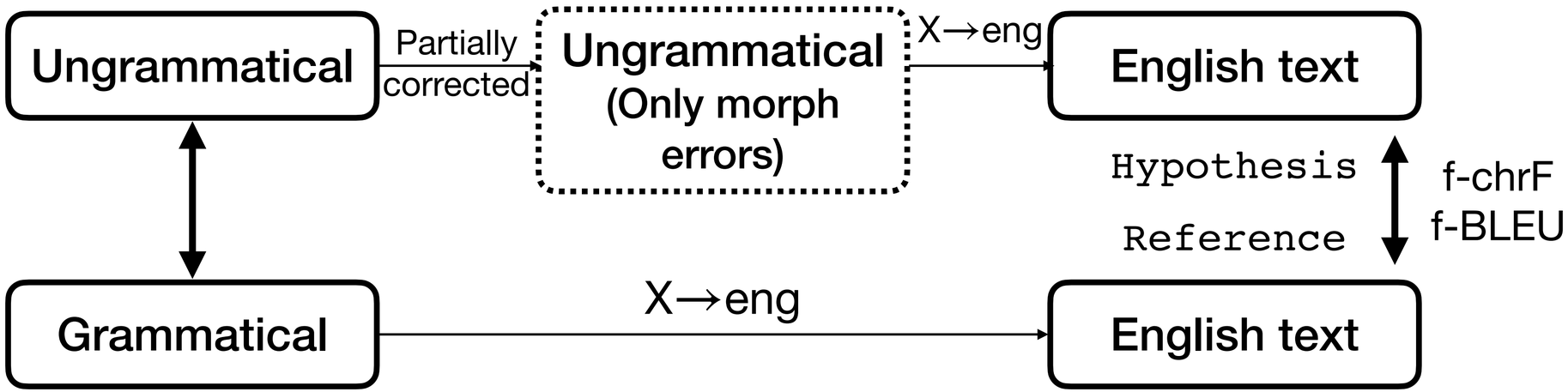
- Evaluating best-performing models on WMT19 news translation shared task (Ott et al., 2019, `fairseq`)
- Adversarial evaluation on newstest2018 (de→en, ru→en)
- BLEU ↓ chrF ↓ 
- NR ≈ 1 

X-en	Baseline		Adversarial		
	BLEU	chrF	BLEU	chrF	NR
ru-en	38.33	0.63	18.50	0.47	0.81
de-en	48.40	0.70	33.43	0.59	1.00

Learners' Text

- We evaluate robustness of pretrained NMT systems towards morphology-related errors in learners' text

- Below schematic to use Russian and German GEC



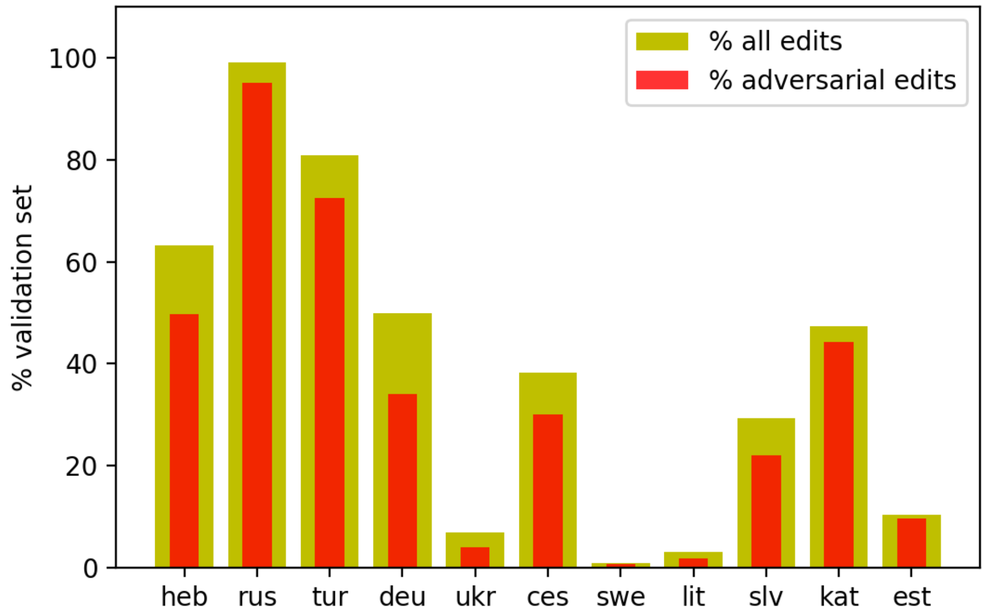
Learners' Text

- Evaluated using pretrained NMT systems (Ott et al., 2019)
- Metric: *faux*-BLEU, *faux*-chrF computed on the pseudo-reference
- Observation: both Russian and German models are robust to morphology-related errors

Dataset	f-BLEU	f-chrF
Russian GEC	85.77	91.56
German GEC	89.60	93.95

TED Corpus

- Multilingual TED corpus (Qi et al., 2018) provides parallel data for >50 language pairs
- Only chose languages with >80% accuracy with `unimorph_inflect`
- Perturbations on dev split
- transformer_iwslt_de_en architecture from `fairseq`

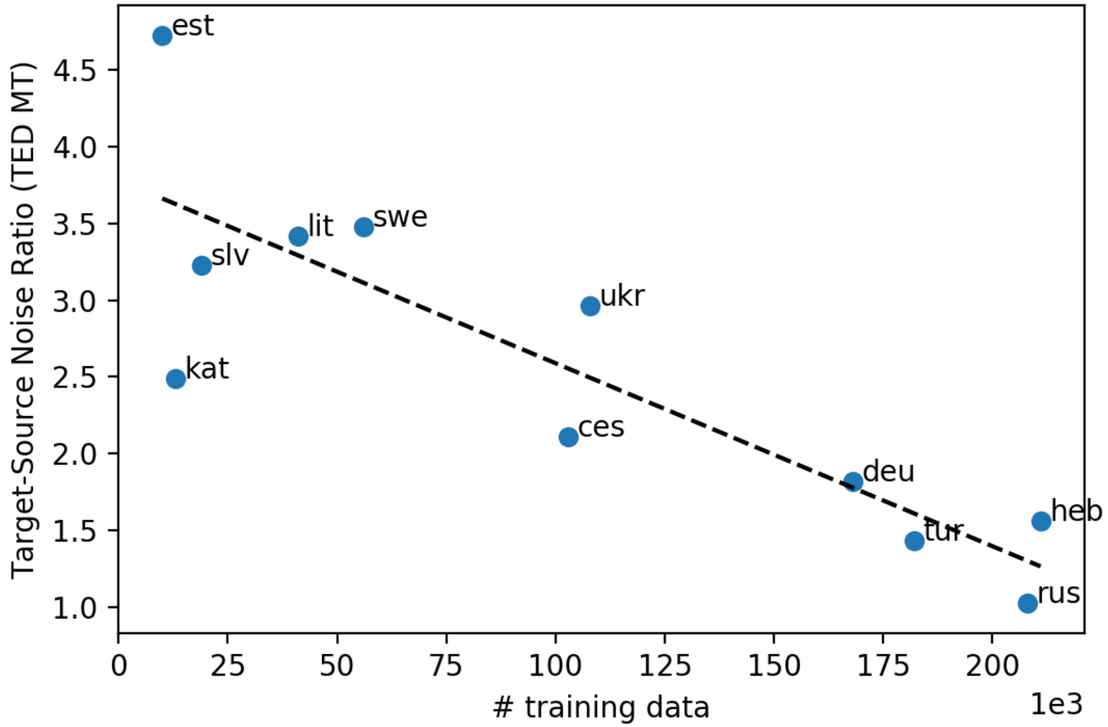


TED Corpus

X-en	# train	Baseline		Adversarial		
		BLEU	chrF	BLEU	chrF	NR
Hebrew	211k	40.1	0.59	33.9	0.54	1.56
Russian	208k	25.6	0.48	11.7	0.35	1.03
Turkish	182k	27.8	0.50	18.9	0.41	1.43
German	168k	34.2	0.56	31.3	0.54	1.82
Ukrainian	108k	25.8	0.47	25.7	0.47	2.96
Czech	103k	29.4	0.51	26.6	0.49	2.11
Swedish	56k	36.9	0.56	36.8	0.56	3.48
Lithuanian	41k	18.9	0.40	18.8	0.39	3.42
Slovenian	19k	11.5	0.32	10.5	0.31	3.23
Georgian	13k	5.8	0.25	4.9	0.21	2.49
Estonian	10k	6.7	0.26	6.5	0.25	4.72

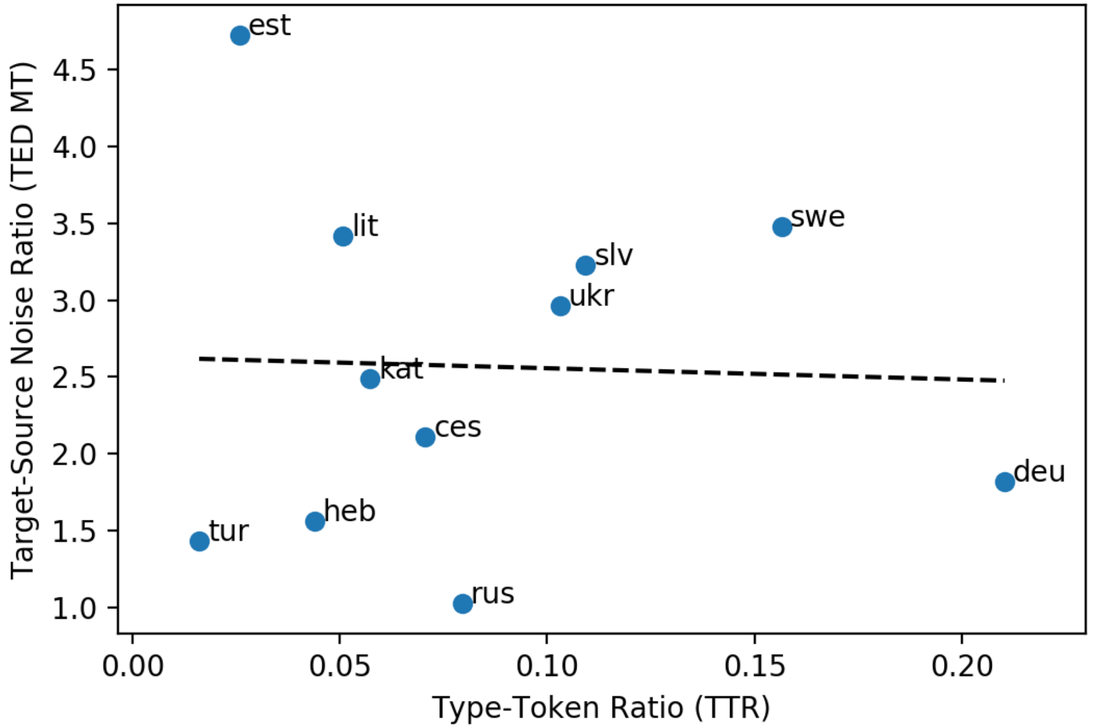
Effect of Resource Size

- # train data *vs* Noise Ratio
- Models for high-resource languages tend to be more robust



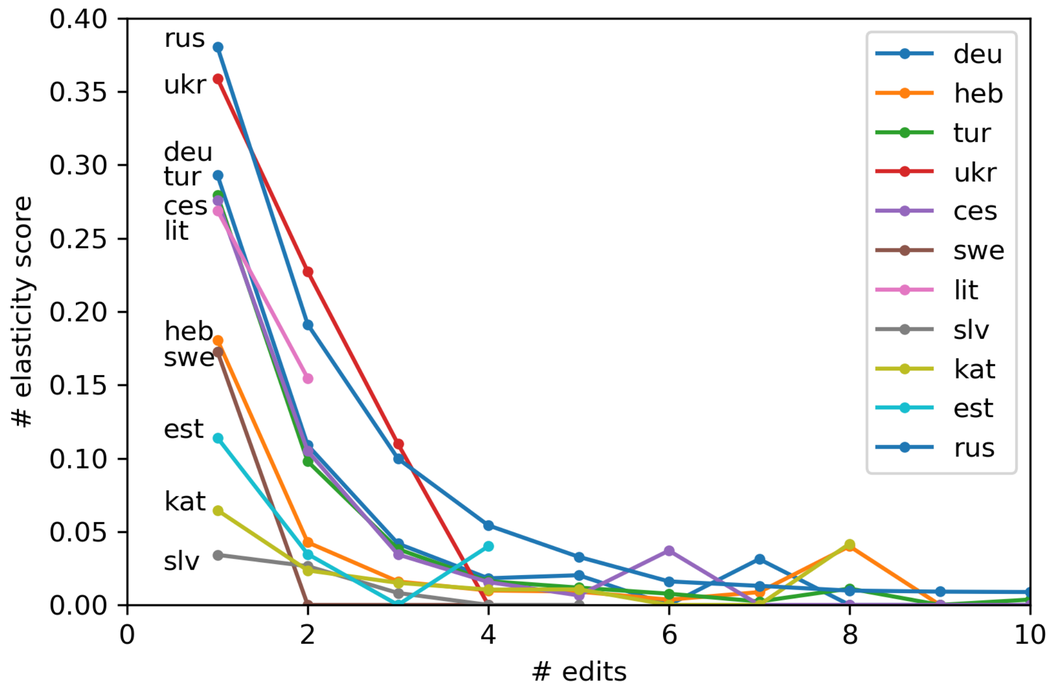
Effect of Morphological Richness

- Morphological richness measured by TTR on UniMorph
- TTR of language is not correlated with Noise Ratio



Analysis: Elasticity

- Perturbations in source are usually reflected in target
- *Elasticity*: fraction of perturbed sentences that lead to same target (translation) as the unperturbed source
- Russian models seem more elastic than Ukrainian models



Conclusions

- Robustness of an NMT model seems to depend on the resource size, and not on the morphological richness of the source language
- Perturbation method can be improved further to reduce significant semantic change, and limit number of edits
- Need further analysis of large NMT models for more language pairs