An Analysis of source-side Grammatical Errors in Neural Machine Translation

Antonis Anastasopoulos
aanastas@andrew.cmu.edu

NEULAB

Carnegie Mellon University
Language Technologies Institute

Highlights

Non-native English speakers outnumber native ones by 3:1, and they often produce ungrammatical utterances.

Grammatical Errors are Great Adversarial Examples!

Grammatical errors are **natural, meaning-preserving perturbations** (in most cases) but a NMT system trained on clean data cannot properly handle such input.

Robustness definition: a NMT system is robust, if it produces the same translation for a clean source $\tilde{\mathbf{x}}$ and a noisy version $\mathbf{x}$. Thus, we treat the output of the clean source as reference $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} \approx \mathrm{MT}^{actual}(\tilde{\mathbf{x}}) = \tilde{\mathbf{y}}.$$

We define the **Target-Source Noise Ratio** given a distance metric (e.g. BLEU) as follows:

$$\mathrm{NR}(\mathbf{x}, \tilde{\mathbf{x}}, \mathbf{y}, \tilde{\mathbf{y}}) = \frac{d(\mathbf{y}, \tilde{\mathbf{y}})}{d(\mathbf{x}, \tilde{\mathbf{x}})} = \frac{100 - \mathrm{BLEU}(\mathbf{y}, \tilde{\mathbf{y}})}{100 - \mathrm{BLEU}(\mathbf{x}, \tilde{\mathbf{x}})}.$$

This also functions as a reference-less criterion (variation to Michel et al. 2019) for evaluating adversarial attacks.

We evaluate 6 Grammar Error Correction corpora on Eng-Deu translation, using the SOTA model of from fairseq.

Dataset	Robust %	NR
WI+Loc. A	17.7	2.1
WI+Loc. B	21.2	2.4
WI+Loc. C	29.1	2.7
WI+Loc. N	28.8	3.2
NUCLE	20.7	2.9
FCE	20.5	2.4
JFLEG	12.4	2.2
Lang8	16.1	2.2
average	17.6	2.6

Fluency

Takeaways:

1) Robustness increases with fluency, and the datasets have different properties. (e.g. note WI+Loc, where fluency A < B < C < N)

2) The MT system generally magnifies the input noise (because NR > 1 in all cases)

Single Error Analysis

Focusing on single-error sentences, we find that different errors behave differently.

The average error is recoverable about 11% of the time.

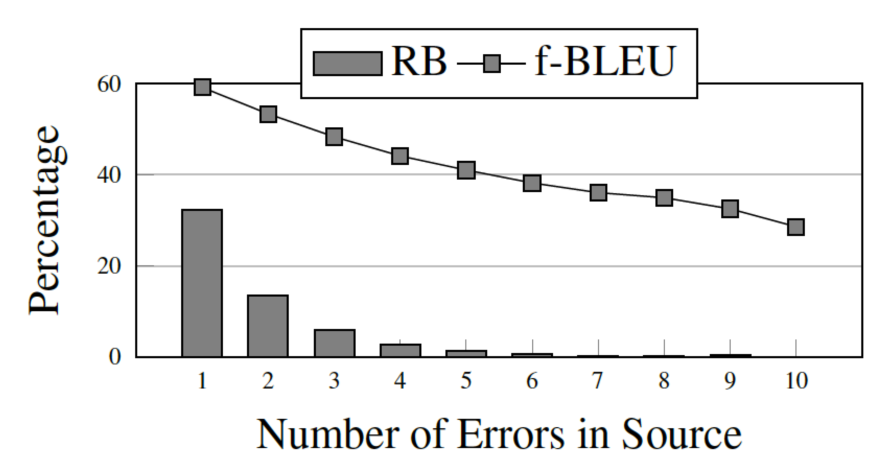
Morphological errors are more often recoverable, as well as orthography and word order errors.

Conjunctions and other misused words are typically translated verbatim, leading to incorrect translations.

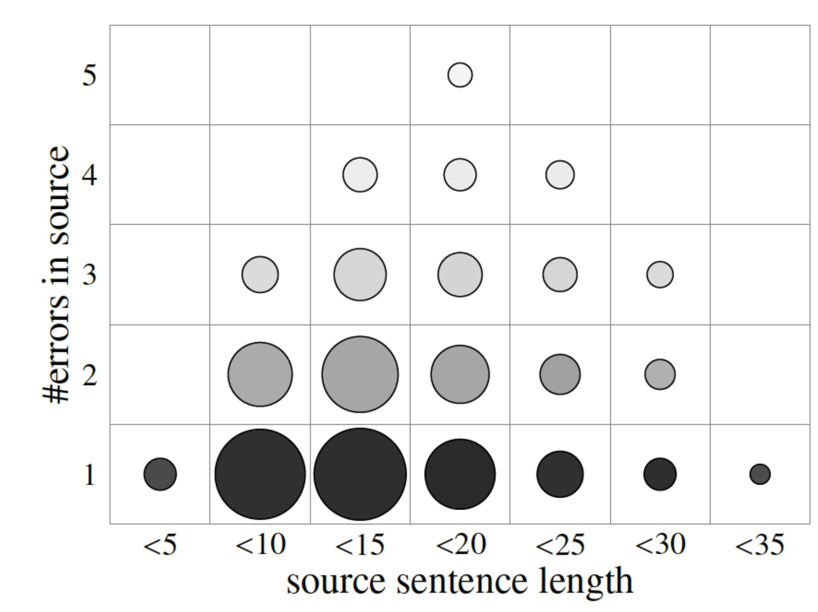
Recoverable Error	RB	Non-recoverable Error	RB
VERB-INFL	22%	CONJ	3%
VERB-SVA	22%	OTHER	5%
ORTH	19%	NOUN	6%
VERB-FORM	17%	ADV	7%
WO	17%	VERB	7%

Sentence Level Analysis

1. Increased amounts of noise in the source degrade translation performance.



2. Sentence length correlates with number of errors. Robustness (shown by opaqueness) correlates with both.

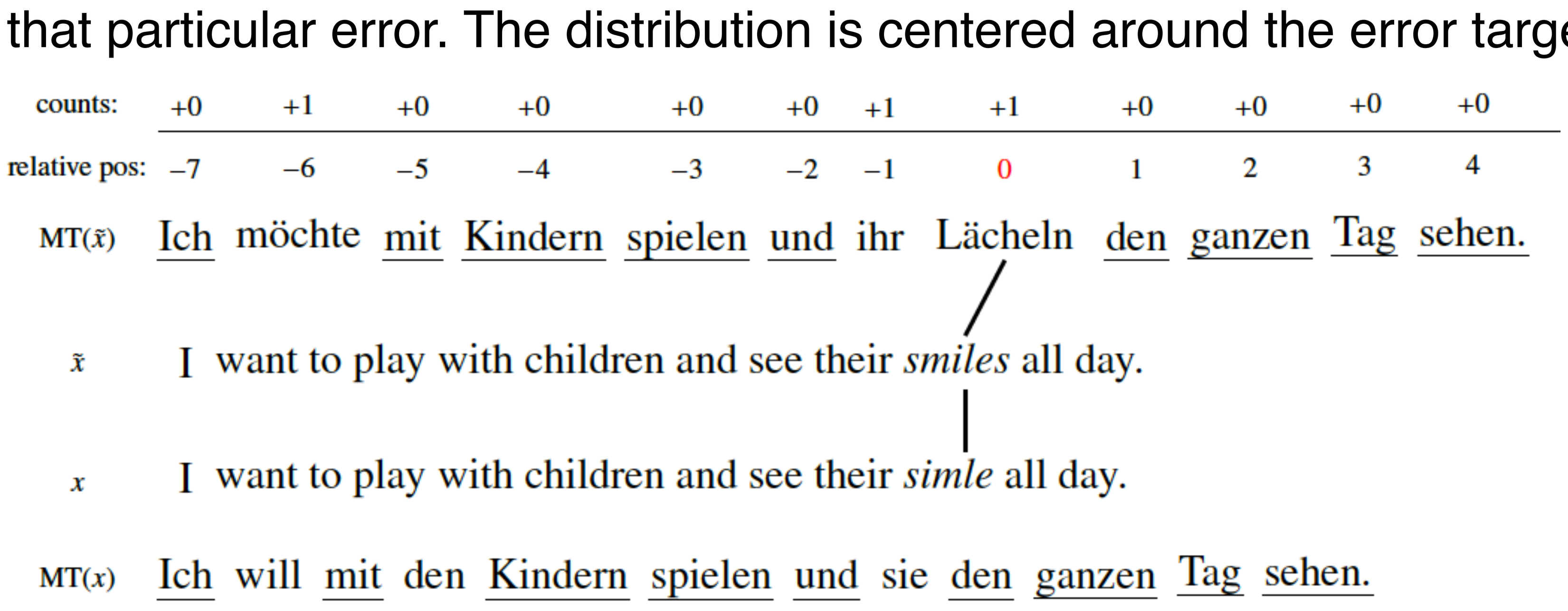


Divergence

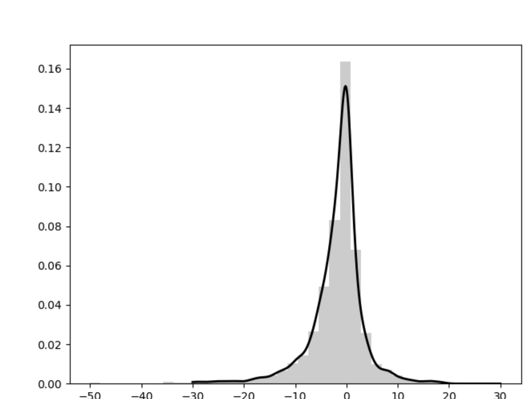
To visualize the effect of source errors on the output, we define a **divergence distribution** for each error type, over single-error sentences with that particular error. The distribution is centered around the error target position (found using alignments).

In the example to the right, *"simle"* is the error. Its correction, *"smiles"* is aligned to "Lächeln" which denotes the center of the divergence distribution.

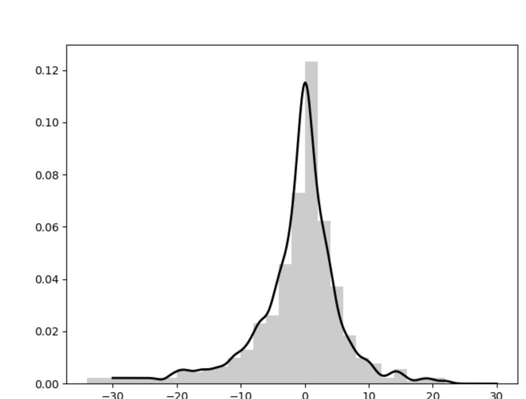
The non-underlined words are divergent and contribute to the divergence counts.



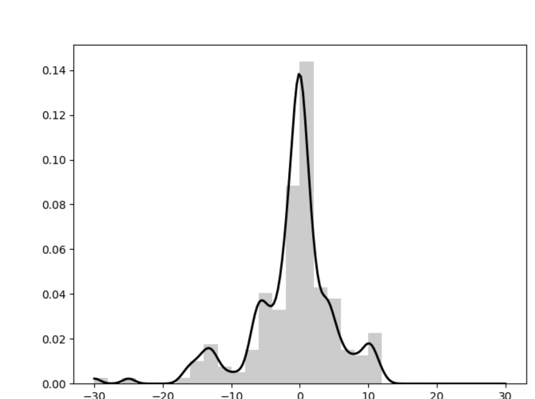
Finding: specific errors have notably different divergence distributions. Some examples here:



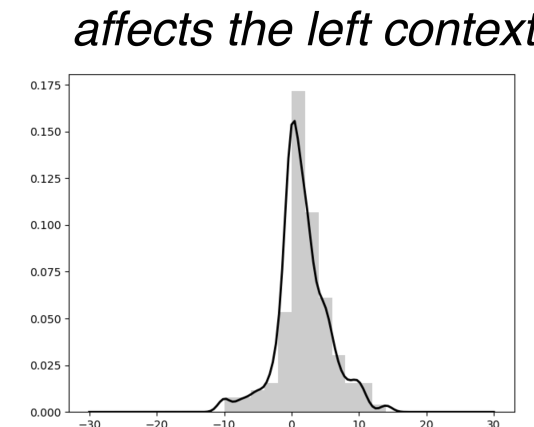
Orthography
(low mean, skew negative)
affects the left context



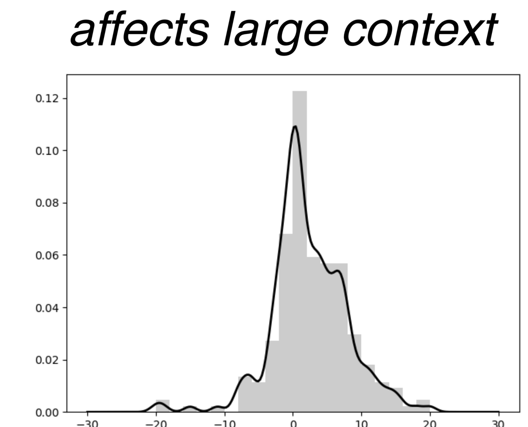
Verb Form Error
(low mean, large variance)
affects large context



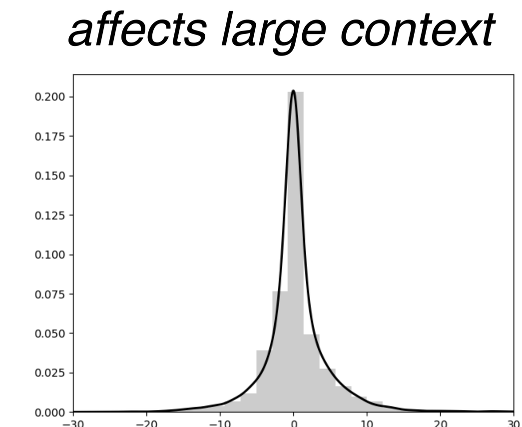
Adjective Error
(low mean, large variance)
affects large context



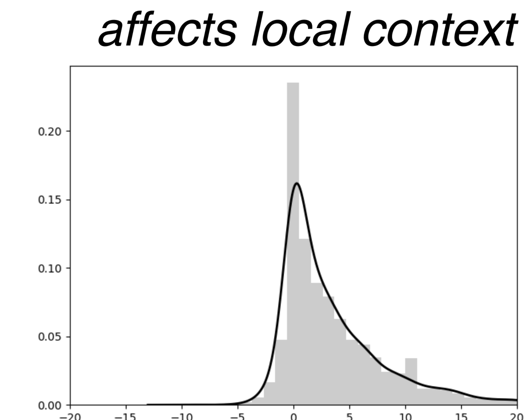
Contraction
(low mean, small variance)
affects local context



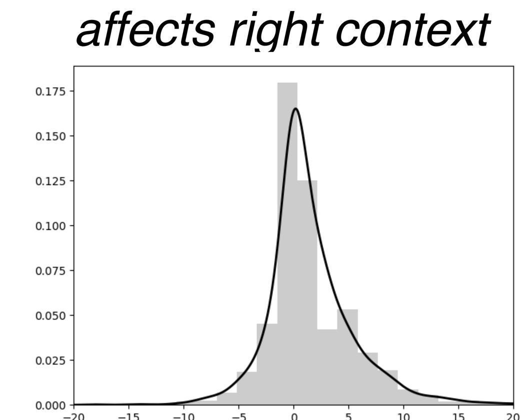
Conjunction
(largest mean)
affects right context



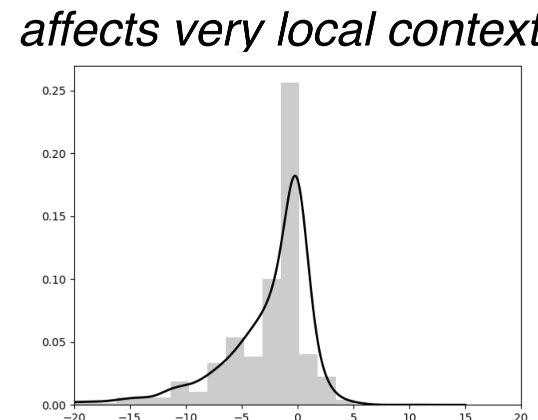
Spelling Mistake
(zero mean, skew positive)
affects very local context



sentence Beginning
affects right context



sentence Middle
affects right context



sentence End
affects local left context