An Analysis of Source-Side Grammatical Errors in NMT

Antonios Anastasopoulos
Language Technologies Institute
Carnegie Mellon University
aanastas@cs.cmu.edu

Abstract

The quality of Neural Machine Translation (NMT) has been shown to significantly degrade when confronted with source-side noise. We present the first large-scale study of state-of-the-art English-to-German NMT on real grammatical noise, by evaluating on several Grammar Correction corpora. We present methods for evaluating NMT robustness without true references, and we use them for extensive analysis of the effects that different grammatical errors have on the NMT output. We also introduce a technique for visualizing the divergence distribution caused by a source-side error, which allows for additional insights.

1 Introduction

Neural Machine Translation (NMT) has become the *de facto* option for industrial systems in high-resource settings (Wu et al., 2016; Hassan Awadalla et al., 2018; Crego et al., 2016) while dominating public benchmarks (Bojar et al., 2018). However, as several works have shown, it has a notable shortcoming (among others, see Koehn and Knowles (2017) for relevant discussion) in dealing with source-side noise, during both training and inference.

Heigold et al. (2018) as well as Belinkov and Bisk (2018) pointed out the degraded performance of character- and subword-level NMT models when confronted with synthetic character-level noise –like swaps and scrambling– on French, German, and Czech to English MT. Belinkov and Bisk (2018) and Cheng et al. (2018) also studied synthetic errors from word swaps extracted from Wikipedia edits. Anastasopoulos et al. (2019) focused on a small subset of grammatical errors (article, preposition, noun number, and subject-verb agreement) and evaluated on English-to-Spanish synthetic and natural data.

However, no previous work has extensively studied the behavior of a state-of-the-art (SOTA) model on natural occurring data. Belinkov and Bisk (2018) only trained their systems on about 200K parallel instances, while Heigold et al. (2018) and Anastasopoulos et al. (2019) trained on about 2M parallel sentences from the WMT'16 data. Importantly, though, none of them utilized vast monolingual resources through back-translation, a technique that has been consistently shown to lead to impressively better results (Sennrich et al., 2016a).

In this work, we perform an extensive analysis of the performance of a *state-of-the-art* English-German NMT system, with regards to its robustness against real grammatical noise. We propose a method for robustness evaluation without gold-standard translation references, and perform experiments and extensive analysis on all available English Grammar Error Correction (GEC) corpora. Finally, we introduce a visualization technique for performing further analysis.

2 Data and Experimental Settings

To our knowledge, there are six publicly available corpora of non-native or erroneous English that are annotated with corrections and which have been widely used for research in GEC.

The NUS Corpus of Learner English (NUCLE) contains essays written by students at the National University of Singapore (Dahlmeier et al., 2013). It has become the main benchmark for GEC, as it was used in the CoNLL GEC Shared Tasks (Ng et al., 2013, 2014). The Cambridge Learner Corpus First Certificate in English FCE corpus[1] (Yannakoudakis et al., 2011) consists of essays collected from learners taking the Cambridge Assessment's English as a Second or Other Language

[1] We use the publicly available portion.

(ESOL) exams.[2] The LANG-8 corpus (Tajiri et al., 2012) was harvested from user-provided corrections in an online learner forum. Both have also been widely used for the GEC Shared Tasks. Another small corpus developed for evaluation purposes is the JHU FLuency-Extended GUG corpus (JFLEG) (Napoles et al., 2017) with correction annotations that include extended fluency edits rather than just minimal grammatical ones. The Cambridge English Write & Improve (W&I) corpus (Andersen et al., 2013) is collected from an online platform where English learners submit text and professional annotators correct them, also assigning a CEFR level of proficiency (of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, 2001). Lastly, we use a portion of the LOCNESS corpus,[3] a collection of essays written by *native* English speakers. 50 essays from LOCNESS were annotated by W&I annotators for grammatical errors, so we will jointly refer to these two corpora as WI+LOC.

All datasets were consistently annotated for errors with ERRANT (Bryant et al., 2017), an automatic tool that categorizes correction edits.[3] This allows us to consistently aggregate results and analysis across all datasets.

2.1 Notation and Experimental Settings

Throughout this work, we use the following notations:

- $\mathbf{x}$: the original, noisy, potentially ungrammatical English sentence. Its tokens will be denoted as x_i.
- $\tilde{\mathbf{x}}$: the English sentence with the correction annotations applied to the original sentence x, which is deemed fluent and grammatical. Again, its tokens will be denoted as $\tilde{x}_i$.
- $\mathbf{y}$: the output of the NMT system when $\mathbf{x}$ is provided as input (tokens: y_j).
- $\tilde{\mathbf{y}}$: the output of the NMT system when $\tilde{\mathbf{x}}$ is provided as input (tokens: $\tilde{y}_j$).

For the sake of readability, we use the terms grammatical errors, noise, or edits interchangeably. In the context of this work, they will all denote the annotated grammatical errors in the source sentences ($\mathbf{x}$). We also define the number of errors, or the amount of noise in the source, to be equivalent to the number of annotated necessary edits that the source $\mathbf{x}$ requires to be deemed grammatical ($\tilde{\mathbf{x}}$), as per standard GEC literature.

The main focus of our work is the performance analysis of the NMT system, so our experimental design is fairly simple. We use the SOTA NMT system of Edunov et al. (2018) for translating both the original and the corrected English sentences for all our GEC corpora.[4] The system achieved the best performance in the WMT 2018 evaluation campaign, using an ensemble of 6 deep transformer models trained with slightly different back-translated data.[5]

3 Evaluating NMT Robustness without References

When not using human judgments on output fluency and adequacy, Machine Translation is typically evaluated against gold-standard reference translations with automated metrics like BLEU (Papineni et al., 2002) or METEOR (Banerjee and Lavie, 2005). However, in the case of GEC corpora, we do not have access to translations – only monolingual data (potentially with ungrammaticalities) and correction annotations.[6] Quality Estimation for MT also operates in a reference-less setting (see Specia et al. (2018) for definitions and an overview of the field) and is hence very related to our work, but is more aimed towards towards *predicting* the quality of the translation. Our goal instead, is to analyze the behavior of the MT system when confronted with ungrammatical input. Reference-less evaluation has also been proposed for text simplification (Martin et al., 2018) and GEC (Napoles et al., 2016), while the grammaticality of MT systems' outputs has been evaluated with target-side contrastive pairs (Sennrich, 2017).

In this work, the core of our evaluation of a system's robustness lies in the following observation: **a perfectly robust-to-noise MT system would produce the exact same output for the clean and erroneous versions of the same input sentence.**

[2] https://www.cambridgeenglish.org/

[3] NUCLE, LANG8, FCE, and WI+LOC are pre-annotated with ERRANT for the BEA 2019 GEC Shared Task. We also annotated JFLEG.

[4] We use all data, concatenating train, dev, and test splits. We sample 150K sentences from LANG8.

[5] Refer to (Edunov et al., 2018) for further system details.

[6] The ideal way to potentially obtain such references of noisy text is debatable, and the extent to which humans are able to translate ungrammatical text is unknown. A well-crafted investigation could ideally elicit translations of both original and (the multiple versions of) corrected texts from multiple translators in order to study this issue. Although we highly encourage such a study, we could not conduct one due to budgetary constraints.

Denoting a perfect MT system as a function $\text{MT}^{perfect}(\cdot)$ over input sentences to the correct output sentences $\hat{\mathbf{y}}$, then both input sentences $\mathbf{x}$ and $\tilde{\mathbf{x}}$ would yield the same output:

$$\hat{\mathbf{y}} = \text{MT}^{perfect}(\mathbf{x}) = \text{MT}^{perfect}(\tilde{\mathbf{x}}).$$

In our case, $\hat{\mathbf{y}}$ is unknown and we only have access to a very good (but still imperfect) system $\text{MT}^{actual}(\cdot)$. We propose, therefore, to treat the system's output of the cleaned input ($\tilde{\mathbf{y}}$) as reference. Our assumption is that $\tilde{\mathbf{y}}$ is a good approximation of the correct translation $\hat{\mathbf{y}}$:

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

Under this assumption, we can now evaluate our system's robustness by comparing $\mathbf{y}$ and $\tilde{\mathbf{y}}$ using automated metrics at the corpus or sentence level. Here we list the metrics that we use and briefly discuss their potential shortcomings.

Robustness Percentage (RB): Given a GEC corpus $\{X, \tilde{X}\}$, this corpus-level metric evaluates the percentage at which the system outputs agree at the sentence level:

$$RB = \frac{\sum_{\mathbf{x},\tilde{\mathbf{x}} \in \{X,\tilde{X}\}} c_{agree}(\text{MT}(\mathbf{x}), \text{MT}(\tilde{\mathbf{x}}))}{|X|},$$

$$c_{agree}(\mathbf{y}, \tilde{\mathbf{y}}) = \begin{cases} 1 & \text{if } \mathbf{y} = \tilde{\mathbf{y}}, \\ 0 & \text{otherwise.} \end{cases}$$

f-BLEU: BLEU is the most standard MT evaluation metric, combining n-gram overlap accuracy with a brevity penalty. We calculate sentence- and corpus-level BLEU-4 scores for every $\mathbf{y}$ with $\tilde{\mathbf{y}}$ as the reference. Note that the BLEU scores that we obtain in our experiments are not comparable with any previous work (as we do not use real references) so we denote our metric as faux BLEU (f-BLEU) to avoid confusion.[7]

f-METEOR: Same as above, we define faux-METEOR using the METEOR MT metric (Denkowski and Lavie, 2014) which is more semantically nuanced than BLEU.

Target-Source Noise Ratio (NR): A notable drawback of all the previously discussed metrics is that they do not take into account the source sentences $\mathbf{x}$ and $\tilde{\mathbf{x}}$ or their distance. However, it is expected that minimal perturbations of the input (e.g. some missing punctuation) will also be minimally reflected in the difference of the outputs, while more distant inputs (which means higher levels of noise in the uncorrected source) would lead to more divergent outputs. To account for this observation, we propose Target-Source Noise Ratio (NR) which factors the distance of the two source sentences into a metric. The distance of two sentences can be measured by any metric like BLEU, METEOR, etc. We simply use BLEU:

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

If the average (corpus-level) Noise Ratio score is smaller than 1 ($\text{NR}(X, \tilde{X}, Y, \tilde{Y}) < 1$) then we can infer that the MT system reduces the relative amount of noise, as there is higher relative n-gram overlap between the outputs than the inputs. On the other hand, if it is larger than 1, then the MT system must have introduced even more noise.[8]

Recently, Michel et al. (2019) proposed a criterion for evaluating adversarial attacks, which requires also having access to the correct translation $\hat{\mathbf{y}}$. Using a similarity function $s(\cdot)$, they declare an adversarial attack to be successful when:

$$s(\mathbf{x}, \tilde{\mathbf{x}}) + \frac{s(\mathbf{y}, \hat{\mathbf{y}}) - s(\tilde{\mathbf{y}}, \hat{\mathbf{y}})}{s(\mathbf{y}, \hat{\mathbf{y}})} > 1$$

In our reference-less setting, assuming $\hat{\mathbf{y}} \approx \tilde{\mathbf{y}}$ leads to $s(\tilde{\mathbf{y}}, \hat{\mathbf{y}}) = 1$. Finally, representing the similarity function with a distance function $s(\cdot) = 1 - d(\cdot)$ and simple equation manipulation, the criterion becomes exactly our Target-Source Noise Ratio. We have, hence, arrived at a reference-less criterion for evaluating any kind of adversarial attacks.[9]

4 Analysis

We first review the aggregate results across all datasets (§4.1) and with all metrics. We also

[7] In absolute numbers, we obtain higher scores than the scores of a MT system compared against actual references: the best English-German system from the WMT 2018 evaluation (Edunov et al., 2018) obtained a BLEU score of 46.5; our f-BLEU scores are in the [37-65] range, but we consider them informative only when viewed relative to other f-BLEU scores.

[8] As presented, the NR metric assumes that the length of the input and target sentences are comparable. In the English-German case, this is more or less correct. A more general implementation could include a discount term based on the average sentence length ratio of the two languages.

[9] Indeed, grammatical noise is nothing more than natural occurring adversarial noise.

dataset		number of sentences	average #corr/sent.	RB	over non-robust sent f-BLEU	over non-robust sent f-METEOR	NR
WI+LOC	A	9K	3.4	17.77	46.75	65.29	2.12
	B	10K	2.6	21.17	54.72	70.80	2.39
	C	5.9K	1.8	29.07	63.46	76.63	2.73
	N	500	1.8	28.80	64.79	77.35	3.23
NUCLE		21.3K	2.0	20.69	59.97	74.6	2.92
FCE		20.7K	2.4	20.48	50.45	67.49	2.43
JFLEG		1.3K	3.8	12.42	42.05	61.99	2.18
LANG8		149.5K	2.4	16.06	37.15	58.89	2.20
ALL\LANG8		69K	2.4	20.94	54.65	70.64	2.55
ALL		218.5K	2.4	17.60	42.65	62.59	2.55

Table 1: Aggregate results across all datasets. As expected, the NMT system's performance deteriorates as input noise increases. For all metrics except NR, higher scores are better.

present findings based on sentence-level analysis (§4.2). We investigate the specific types of errors that contribute to robustness as well as those that increase undesired behavior in §4.3. Finally, in Section §4.4 we introduce the more fine-grained notion of divergence that allows us to perform interesting analysis and visualizations over the datasets.

4.1 Aggregate Results

Table 1 presents the general picture of our experiments, summarizing the translation robustness across all datasets with all the metrics that we examined, and also providing basic dataset statistics. Note that the aggregate f-BLEU and f-METEOR scores in Table 1 are calculated excluding sentences where the system exhibits robustness. We made this choice in order to tease apart the differences across the datasets by focusing on the problematic instances; having between 17% and 29% of the scores be perfect 100 f-BLEU points would obscure the analysis. We also report average scores across all datasets (last row) as well as scores without including LANG8, since the LANG8 dataset is significantly larger than the others.

Takeaway 1: Increased amounts of noise in the source degrade translation performance. The first takeaway confirms the previous results in the literature (Belinkov and Bisk, 2018; Anastasopoulos et al., 2019). The average number of corrections per sentence and the robustness percentage (RB) column have a Pearson's correlation coefficient $\rho = -0.82$, while both f-BLEU and f-METEOR have lower $\rho = -0.71$.

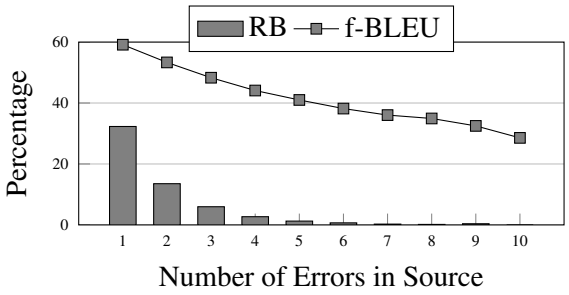

Figure 1: Effect of the number of errors on robustness. Robustness Percentage more than halves for each additional input sentence error, while f-BLEU on the non-robust sentences reduces linearly.

This is further outlined by the results on the WI+LOC datasets. The English proficiency of the students increases from the A to B to C subsets, and the N subset is written by native English speakers. An increase in English proficiency manifests as a lower number of errors, higher robustness percentage, and larger f-BLEU scores.

Takeaway 2: The MT system generally magnifies the input noise. This is denoted by the NR column which is larger than 1 across the board. This means that the MT system exacerbated the input noise by a factor of about 2.5. This effect is more visible when the source noise levels are low, as in the WI+LOC C and N or the NUCLE datasets.

4.2 Sentence-level Findings

We continue our analysis focusing on instance or sentence-level factors, presenting results combining all datasets.

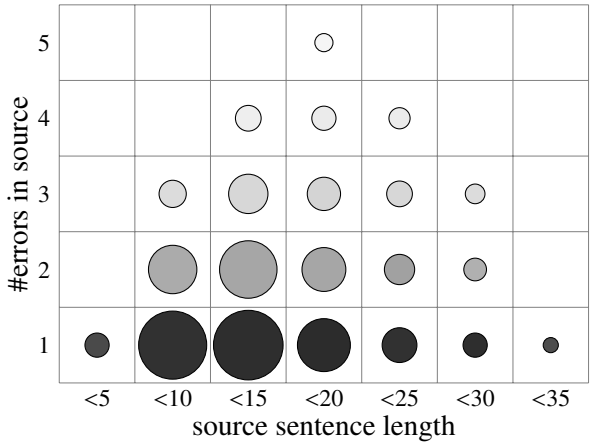

Figure 2: Robustness Percentage broken down by sentence length and number of source errors. The radius of each circle is proportional to the number of sentences and the opacity corresponds to RB score (darkest: RB=33%, lowest: RB≈2%). The model is more robust with few errors regardless of sentence length.

Effect of the input number of errors: Figure 1 clearly shows the compounding effect of source side errors. Each additional error reduces overall robustness by more than 50%: from robust behavior in about 32% of the 1-error instances, to 13% for the 2-errors instances, to 6% on instances with 3 errors; and so forth, to the point that the model is robust in less than 1% of instances with more than 5 source-side errors. The robustness drop when computed with f-BLEU is practically linear, starting from about 59 f-BLEU points when a single error is present, falling to about 28 when the source has more than 9 errors.

Effect of input length: One factor related to the number of input errors is the effect of the source sentence length. We find that there is a negative correlation between the input length and the model's robustness. This is to be expected, as input length and the number of errors are also correlated: longer sentences are more likely to more errors, and inversely, short sentences cannot have a large number of errors.

Figure 2 presents the RB score across these two factors. We bin the input sentences based on their sentence lengths and based on the number of errors in the source. We only plot bins that have a RB score of more than 1% (reflected in the opacity of the plot). It is clear that more errors in a source sentence lead to reduced robustness, while the sentence length is not as significant a factor.

A closer look at sentences with a single error reveals that the system is robust about 30% of the time regardless of their length, with a slight increase in accuracy as the length increases. Longer sentences provide more context, which presumably aids in dealing with the source noise. This pattern is similar across all rows in Figure 2.

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

Table 2: Some errors are easier to translate correctly than others. The average error has an RB score of 11%. We present the errors that fall out of the $[\mu \pm 2\sigma]$ range.

4.3 Error-level Analysis

In this section we aim to study and identify the error types from which the NMT system is able to recover, or not. To avoid the compounding effects of multiple source-side errors, we restrict this analysis to sentences that have a single error.

We have already discussed in Section 4.1 how the NMT system is robust on about 20% of the instances across all corpora. By selecting those instances and computing basic error statistics on them, we find that the average error is recoverable about 11% of the time ($\mu = 0.11$). Table 2 presents the errors that are harder or easier to translate correctly. We choose to present the errors that are at the bottom and top, respectively, of the ranking of the errors, based on the average RB score that their corresponding test instances receive.

The non-recoverable errors on the right side of Table 2 are mostly semantic in nature: all five of them correspond to instances where a semantically wrong *word* was used.[10] Correcting and even identifying these types of errors is difficult even in a monolingual setting as world knowledge and/or larger (document/discourse) context is needed. One could argue, in fact, that such errors are not *grammatical*, i.e. the source sentence is fluent. Furthermore, one could form a solid argument for not wanting/expecting an *MT* system to alter the semantics of the source. The MT system's job is exactly to accurately convey the semantics of the

[10] We refer the reader to Bryant et al. (2017) for a complete list of the error type abbreviations.

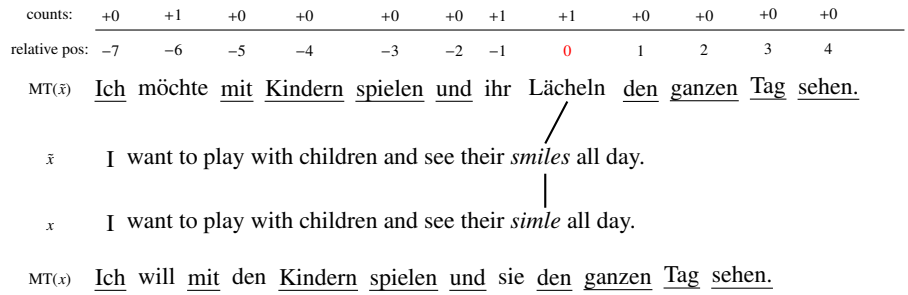

Figure 3: The procedure of computing *divergence* over a quadruple $(\mathbf{x}, \mathbf{y}, \tilde{\mathbf{x}}, \tilde{\mathbf{y}})$. Each token in output y not in the desired output $\tilde{y}$ is considered a divergent token (underlined=matching). The x-axis is centered around the token $\tilde{y}_k$ that aligns to the edit $x_i{*}\!\rightarrow \tilde{x}_j$. The counts describe the caused divergence relative to the expected error's position.

source sentence in the target language.

However, there are errors where the intended meaning is clear but ungrammatically executed, as in Table 2's left-side errors. There are three plausible (likely orthogonal, but we leave such analysis for future work) reasons why these errors are easier than average to correctly translate:

1. Self-attention. The encoder's final representations are computed through multiple self-attention layers, resulting in a representation heavily informed by the whole source context. The VERB-INFL, VERB-SVA, and VERB-FORM error categories (all related to morphology and syntactic constraints) apply to edits that subword modeling combined with self-attention would alleviate. Consider the example of the verb inflection (VERB-INFL) error *danceing*/dancing*. The segmentation in the erroneous and the corrected version is *dance|ing* and *danc|ing* respectively. In both cases, the morpheme that denotes the inflection is the same. Verb form (VERB-FORM) errors, on the other hand, typically involve infinitive, to-infinitive, gerund, or participle forms. It seems that in those cases the self-attention component is able to use the context to recover, especially because, as in the VERB-INFL example, the stem of the verb will most likely be the same.

Also, apart from the positional embeddings, no other explicit word order information is encoded by the architecture (unlike recurrent architectures focused on by all previous work, which by construction keep track of word order). We suggest that the self-attention architecture makes word order errors (WO errors are strictly defined as exact match tokens wrongly ordered, e.g. *know already*/already know*) easier to recover from.

2. The extensive use of back-translation. The SOTA model that we use has been trained on massive amounts of back-translated data, where German monolingual data have been translated into English. The integral part is that English sources were *sampled* from the De-En model, instead of using beam-search to generate the most likely output. This means that the model was already trained on a fair amount of source-side noise, which would make it more robust to such perturbations (Belinkov and Bisk, 2018; Anastasopoulos et al., 2019; Singh et al., 2019).

Although we do not have access to the back-translated parallel data that Edunov et al. (2018) used, we suspect that translation errors are fairly common and therefore more prevalent in the final training bitext, making the model more robust to such noise. Current English-to-German SOTA systems might not have issues with translating noun phrases, coordinated verbs, or pronoun number, but they still struggle with compound generation, coreference, and verb-future generation (Bojar et al., 2018).

3. Data preprocessing and subword-level modeling. It is worth noting that ERRANT limits the orthography (ORTH) error category to refer to edits involving casing (lower↔upper) and whitespace changes. Our model, as most of the SOTA NMT models, is trained and operates at the subword level, using heuristic segmentation algorithms like Byte Pair Encoding (BPE) (Sennrich et al., 2016b), that are learned on clean truecased data. Truecasing is also a standard preprocessing step at inference time, hence dealing with casing errors. The BPE segmentation also has the capacity to deal with whitespace errors. For example, the incorrect token "weatherrelated" gets segmented to *we|a|ther|related*. Although imperfect (the segment's segmentation with proper whitespacing is *we|a|ther related*), the two seg-

mentations agree for 3/4 tokens. Most previous work e.g. (Belinkov and Bisk, 2018) has focused on character-level modeling using compositional functions across characters to represent tokens, which are by construction more vulnerable to such errors.

4.4 Divergence

We introduce a method for computing a *divergence distribution*. Computing *divergence* requires a quadruple of $(\mathbf{x}, \tilde{\mathbf{x}}, \mathbf{y}, \tilde{\mathbf{y}})$. We will focus on instances where $\mathbf{x}$ and $\tilde{\mathbf{x}}$ differ only with a single edit, as a simple working example.

Process: Given a source side sentence pair $\mathbf{x}$ and $\tilde{\mathbf{x}}$ with a single grammatical error, it is trivial to identify the position i^* of the correction in $\tilde{\mathbf{x}}$, since we work on corpora pre-annotated with grammatical edits at the token level. Also, using traditional methods like the IBM Models (Brown et al., 1993) and the GIZA++ tool, makes it easy to obtain an alignment between $\mathbf{x}$ and $\mathbf{y}$, as well as between $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$. We use the alignment variable $\alpha_j = i$ to denote that the target word y_j is aligned to source position i, and equivalently the variables $\tilde{\alpha}$ for the corrected source pair. We denote as k^* the target position that aligns to the source-side correction, such that $\tilde{\alpha}_{k^*} = i^*$.

We define the set of divergent tokens $\mathcal{Y}*$ as the set of tokens of $\mathbf{y}$ that do not appear in $\tilde{\mathbf{y}}$:

$$\mathcal{Y}^* = \{y_j \mid y_j \notin \tilde{\mathbf{y}}\}.$$

Now, we use all the previous definitions to define the set $\mathcal{P}$ of target divergent positions for a quadruple $(\mathbf{x}, \tilde{\mathbf{x}}, \mathbf{y}, \tilde{\mathbf{y}})$ as the set of target-side positions of the tokens that are different between $\mathbf{y}$ and $\tilde{\mathbf{y}}$, but *relative* to the position of the target-side token that aligns to the source-side correction:

$$\mathcal{P}(\mathbf{x}, \tilde{\mathbf{x}}, \mathbf{y}, \tilde{\mathbf{y}}) = \{j - k*, \forall y_j \in \mathcal{Y}^*\}.$$

We provide an illustration of this process for a single-error example in Figure 3. The correction *simle*/smiles* is aligned to the word y_7 (Lächeln) in the reference target, so the center of the distribution is moved to $k^* = 7$. For the rest of the positions in the target reference $\tilde{\mathbf{y}}$, we simply update the counts based on whether the word $\tilde{y}_j$ is present in $\mathbf{y}$. The final step is collecting counts across all instances for all the relative divergent positions and analyzing the effect of a source-side error on the target sentence.

Essentially, we expect some source-side errors to have a very local effect on the translation output, which would translate in divergence distributions with low variance (since we center the distribution around k^*). Other source-side errors might cause larger divergence as they might affect the whole structure of the target sentence.

In the Figure 3 example, the only difference between $\mathbf{x}$ and $\tilde{\mathbf{x}}$ is a single word towards the end of the sentence, but the outputs $\mathbf{y}$ and $\tilde{\mathbf{y}}$ diverge on three words. One of them is 6 words away (before) from where we would have expected the divergence to happen (in relative position 0).

After collecting divergence counts for each instance, we can visualize their distribution and compute their descriptive statistics. We focus on the mean μ, standard deviation σ, and the skewness of the distribution as measured by Pearson's definition, using the third standardised moment, defined as:

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^3\right].$$

Across all datasets and errors, the distribution of the divergence caused by single errors in the source has a mean μ_{all}=0.7, standard deviation σ_{all}=5.1, and a slight positive skewness with $\gamma_{1_{all}}$=0.8. This means that the average error affects its general right context, in a ±5 word neighborhood.

In Figure 4 we present several of the errors with the most interesting divergence statistics. Some errors heavily affect their left (e.g. R:ORTH) or right context (U:CONJ). Also, some errors affect a small translation neighborhood as denoted by the low variance of their divergence distribution (e.g. U:CONTR). On the other hand, verb form errors (M:VERB:FORM) have the potential to affect a larger neighborhood: this is expected because English auxiliary verb constructions (e.g. "have eaten X") often get translated to German V2 constructions with an auxiliary verb separated from a final, non-finite main verb (e.g. "habe X gegessen").

In Figure 5 we present the divergence distributions across the sentence quartiles where the error appears. We find that errors in the sentence beginning (1st quartile) severely affect their right context. Errors towards the end of the sentence (4th quartile) affect their left context. Interestingly, we observe that mid-sentence errors (2nd, 3rd quartiles) exhibit much lower divergence variance than errors towards the sentence's edges.

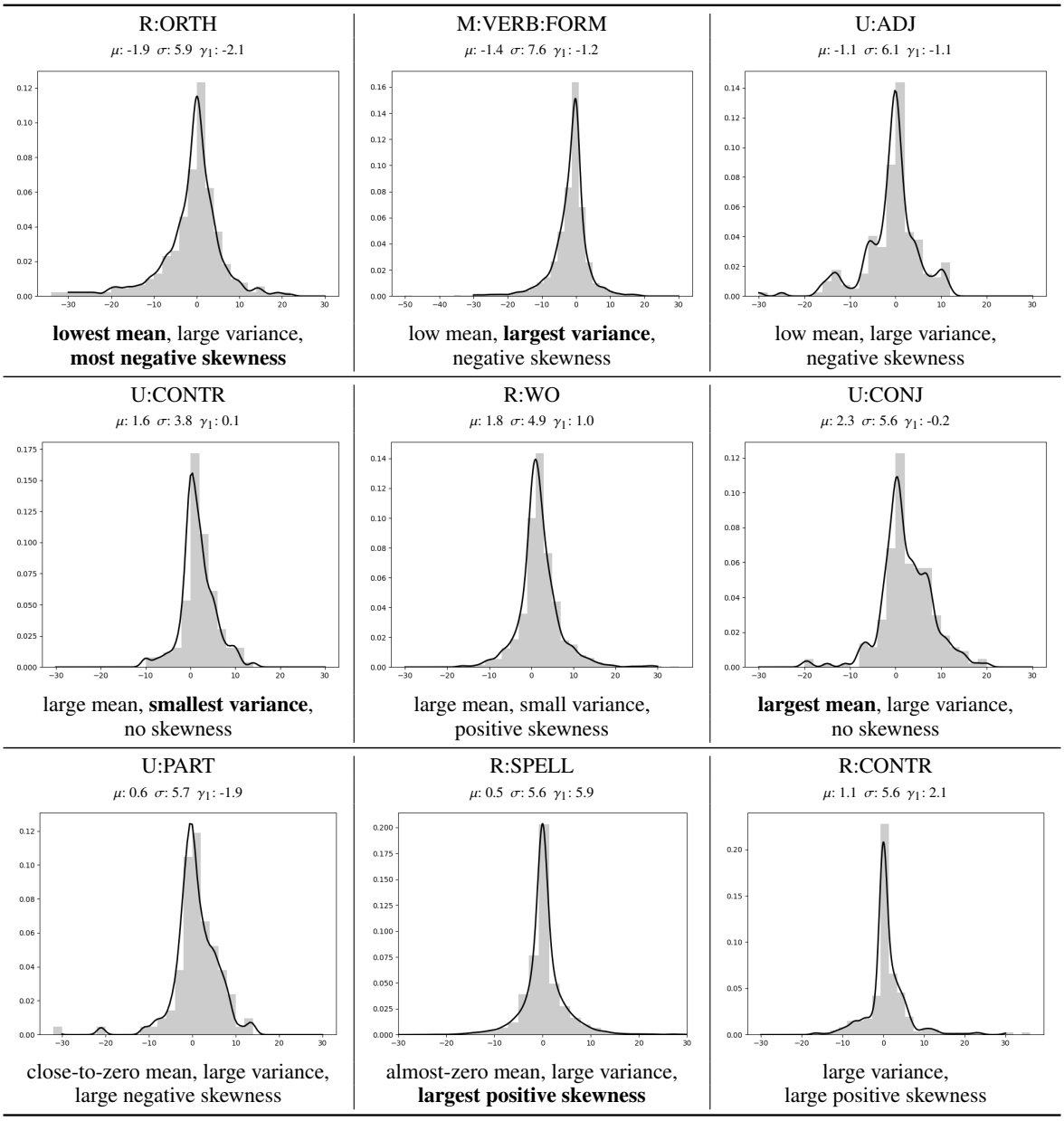

Figure 4: Some interesting errors with statistics on their *divergence distribution*. Some errors (negative mean and skewness: R:ORTH, M:VERB:FORM, U:ADJ) affect the left context of their translation more, while others affect their right translation context (positive mean and skewness R:WO, U:CONJ). Errors might affect a small neighborhood (low variance: U:CONTR, R:WO) or a larger part of the translation (high variance: M:VERB:FORM, U:ADJ, R:CONTR).

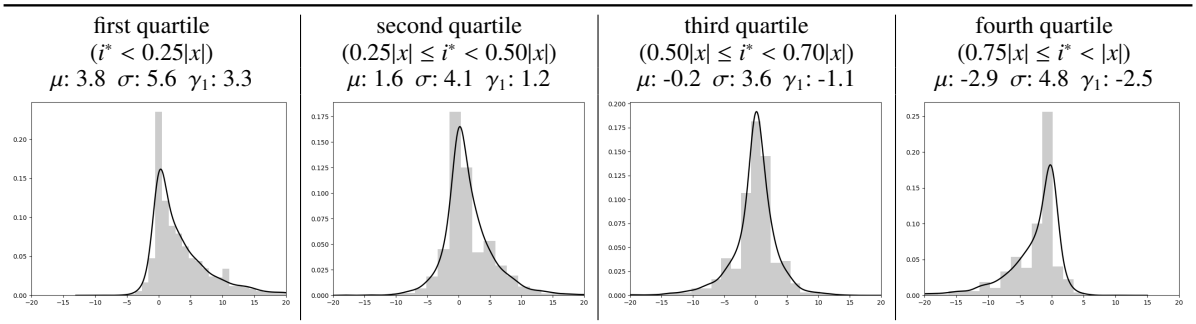

Figure 5: Divergence Distributions for single source error instances per the error's location quartile.

5 Limitations and Extensions

A major limitation of our analysis is the narrow scope of our experiments. We solely focused on a single language pair using a single MT system. Whether different neural architectures over other languages would lead to different conclusions remains an open question. The necessary resources for answering these questions at scale, however, are not yet available. We were limited to English as our source side language, as the majority of the datasets and research works in GEC are entirely English-centric. Perhaps small-scale GEC datasets in Estonian (Rummo and Praakli, 2017) and Latvian (Deksne and Skadina, 2014) could provide a non-English testbed. One would then need error labels for the grammatical edits, so if such annotations are not available, an extension of a tool like ERRANT to these languages would also be required. One should also be careful in the decision of what (N)MT system to test, as using a low-quality translation system would not produce meaningful insights.

Another limitation is that our metrics do not capture whether the changes in the output are actually grammatical errors or not. In the example in Figure 3: the German words "möchte" and "will" that we identified as divergent are practically interchangeable. Therefore, the NMT model is technically not wrong outputting either of them and it is indeed generally possible that differences between y and $\tilde{y}$ are just surface-level ones. The inclusion of f-METEOR as a robustness metric could partially deal with this issue, as it would not penalize such differences. We do believe it is still interesting, though, that a single source error can cause large perturbations in the output, as in the case of errors with large variance in their divergence distribution. Nevertheless, an extension of our study focusing on the grammatical qualities of the MT output would be exciting and automated tools for such analysis would be invaluable (i.e. MT error labeling and analysis tools extending the works of Zeman et al. (2011), Logacheva et al. (2016), Popović (2018), or Neubig et al. (2019)).

A natural next research direction is investigating how to use our reference-less evaluation metrics in order to create a more robust MT system. For instance, one could optimize for f-BLEU or any of the other reference-less measures that we proposed, in the same way that an MT system is optimized for BLEU (either by explicitly using their scores through reinforcement learning or by simply using the metric as an early stopping criterion over a development set). Cheng et al. (2018) recently proposed an approach for training more robust MT systems, albeit in a supervised setting where noise is injected on parallel data, and the proposed solutions of Belinkov and Bisk (2018) and Anastasopoulos et al. (2019) fall within the same category. However, no approach has, to our knowledge, used GEC corpora for training MT systems robust to grammatical errors. In any case, special care should be taken so that any improvements on translating ungrammatical data do not worsen performance on clean ones.

6 Conclusion

In this work, we studied the effects of grammatical errors in NMT. We expanded on findings from previous work, performing analysis on real data with grammatical errors using a SOTA system. With our analysis we were able to identify classes of grammatical errors that are recoverable or irrecoverable. Additionally, we presented ways to evaluate a MT system's robustness to noise without access to gold references, as well as a method for visualizing the effect of source-side errors to the output translation. Finally, we discussed the limitations of our study and outlined avenues for further investigations towards building more robust NMT systems.

Acknowledgements

The author is grateful to the anonymous reviewers, Kenton Murray, and Graham Neubig for their constructive and insightful comments, as well as to Gabriela Weigel for her invaluable help with editing and proofreading the final version of this paper. This material is based upon work generously supported by the National Science Foundation under grant 1761548.

References

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. *Proc. NAACL-HLT*.

Øistein E Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proc. BEA-NLP*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proc. ICLR*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, et al. 2018. Proceedings of the third conference on machine translation. In *Proc. WMT*.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proc. ACL*.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proc. ACL*.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, et al. 2016. Systran's pure neural machine translation systems. arXiv:1610.05540.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proc. BEA-NLP*.

Daiga Deksne and Inguna Skadina. 2014. Error-annotated corpus of latvian. In *Proc. Baltic HLT*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proc. WMT*.

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. 2001. *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proc. EMNLP*.

Hany Hassan Awadalla, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, Will Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation.

Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef Genabith. 2018. How robust are character-based word embeddings in tagging and mt against wrod scramlbing or randdm nouse? In *Proc. AMTA*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proc. WNMT*.

Varvara Logacheva, Chris Hokamp, and Lucia Specia. 2016. Marmot: A toolkit for translation quality estimation at the word level. In *Proc. LREC*.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric De La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proc. ATA*.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proc. NAACL-HLT*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2016. There's no comparison: Reference-less evaluation metrics in grammatical error correction. In *Proc. EMNLP*.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. Jfleg: A fluency corpus and benchmark for grammatical error correction. In *Proc. EACL*, Valencia, Spain.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proc. NAACL-HLT*.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The conll-2014 shared task on grammatical error correction. In *Proc. CoNLL*.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. 2013. The conll-2013 shared task on grammatical error correction. In *Proc. CoNLL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. ACL*.

Maja Popović. 2018. Error classification and analysis for machine translation quality assessment. In *Translation Quality Assessment*, pages 129–158. Springer.

Ingrid Rummo and Kristiina Praakli. 2017. TÜ eesti keele (võõrkeelena) osakonna õppijakeele tekstikorpus [the language learners corpus of the department of estonian language of the university of tartu]. In *Proc EAAL*.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In *Proc. EACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proc. ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proc. ACL*.

Sumeet Singh, Craig Stewart, Graham Neubig, et al. 2019. Improving robustness of machine translation with synthetic noise. arXiv:1902.09508.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for esl learners using global context. In *Proc. ACL*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv:1609.08144.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proc. ACL-HLT*.

Daniel Zeman, Mark Fishel, Jan Berka, and Ondřej Bojar. 2011. Addicter: what is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96:79–88.