Neural Machine Translation of Text from Non-Native Speakers

Antonios Anastasopoulos[†,1] **Alison Lui**[†,2] **Toan Q. Nguyen**[2] **David Chiang**[2]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Department of Computer Science and Engineering, University of Notre Dame
aanastas@cs.cmu.edu {alui,tnguye28,dchiang}@nd.edu

Abstract

Neural Machine Translation (NMT) systems are known to degrade when confronted with noisy data, especially when the system is trained only on clean data. In this paper, we show that augmenting training data with sentences containing artificially-introduced grammatical errors can make the system more robust to such errors. In combination with an automatic grammar error correction system, we can recover 1.0 BLEU out of 2.4 BLEU lost due to grammatical errors. We also present a set of Spanish translations of the JFLEG grammar error correction corpus, which allows for testing NMT robustness to real grammatical errors.

1 Introduction

Neural Machine Translation (NMT) is undeniably a success story: public benchmarks (Bojar et al., 2016) are dominated by neural systems, and neural approaches are the *de facto* option for industrial systems (Wu et al., 2016; Hassan Awadalla et al., 2018; Crego et al., 2016; Hieber et al., 2018). Even under low-resource conditions, neural models were recently shown to outperform traditional statistical approaches (Nguyen and Chiang, 2018).

However, there are still several shortcomings of NMT that need to be addressed: a (non-exhaustive) list of six challenges is discussed by Koehn and Knowles (2017), including out-of-domain testing, rare word handling, the wide-beam problem, and the large amount of data needed for learning. An additional challenge is robustness to noise, both during training and at inference time.

In this paper, we study the effect of a specific type of noise in NMT: grammatical errors. We primarily focus on errors that are made by non-native source-language speakers (as opposed to dialectal language, SMS or Twitter language). Not only is this linguistically important, but we believe that it would potentially have great social impact.

Our contributions are three-fold. First, we confirm that NMT is vulnerable to source-side noise when trained on clean data, losing up to 3.6 BLEU on our test set. This is consistent with previous work, yet orthogonal to it, since we use more realistic noise for our experiments. Second, we explore training methods that can deal with noise, and show that including noisy synthetic data in the training data makes NMT more robust to handling similar types of errors in test data. Combining this simple method with an automatic grammar correction system, we find that we can recover 1.5 BLEU. Third, we release Spanish translations of the JFLEG corpus,[1] a standard benchmark for English Grammar Error Correction (GEC) systems. We also release all other data and code used in this paper.

Our additional annotations on both the JFLEG corpus and the English WMT data will enable the evaluation of the robustness of NMT systems on realistic, natural noise: a robust system would ideally produce the same output when presented with either the original or the noisy source sentence. We hope that our datasets will become a benchmark for noise-robust NMT, because we believe that deployed systems should also be able to handle source-side noise.

2 Data

We focus on NMT from English to Spanish. We choose English to be our source-side language because there exist English corpora annotated with grammar corrections, which we can use as a

[†]Equal contribution. Work performed at the University of Notre Dame.

[1]Freely available at https://bitbucket.com/antonis/nmt-grammar-noise

Proceedings of NAACL-HLT 2019, pages 3070–3080
Minneapolis, Minnesota, June 2 - June 7, 2019.

source of natural noise. Moreover, since English is probably the most commonly spoken non-native language (Lewis et al., 2009), our work could be directly applicable to several translation applications. Our choice of Spanish as a target language enables us to have access to existing parallel data and easily create new parallel corpora (see below, §2.3).

For all experiments, we use the Europarl English-Spanish dataset (Koehn, 2005) as our training set. In the synthetic experiments of Section §2.2, we use the newstest2012 and newstest2013 as dev and test sets, respectively. Furthermore, to test our translation methods on real grammatical errors, we introduce a new collection of Spanish translations of the JFLEG corpus (§2.3).

2.1 Grammar Error Correction Corpora

To our knowledge, there are five publicly available corpora of non-native English that are annotated with corrections, which have been widely used for research in Grammar Error Correction (GEC). The NUS Corpus of Learner English (NUCLE) contains essays written by students at the National University of Singapore, corrected by two annotators using 27 error codes (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). Other corpora include the Cambridge Learner Corpus First Certificate in English FCE corpus (Yannakoudakis et al., 2011), which is only partially public, the LANG-8 corpus (Tajiri et al., 2012), which was harvested from online corrections, and the AESW 2016 Shared Task corpus, which contains corrections on texts from scientific journals.

The last corpus is the JHU FLuency-Extended GUG corpus (JFLEG) (Napoles et al., 2017). This corpus covers a wider range of English proficiency levels on the source side, and its correction annotations include extended fluency edits rather than just minimal grammatical ones. That way, the corrected sentence is not just grammatical, but also guaranteed to be fluent.

2.2 Synthetic grammar errors

Ideally, we would train a translation model to translate grammatically noisy language by training it on parallel data with grammatically noisy language. Since, to our knowledge, no such data exist in the quantities that would be needed, an alternative is to add synthetic grammatical noise to clean data. An advantage of this approach is that controlled introduction of errors allows for fine-grained analysis.

Error Type	Confusion Set
ART	{a, an, the, ∅}
PREP	{on, in, at, from, for, under, over, with, into, during, until, against, among, throughout,of, to, by, about, like, before, after, since, across, behind, but, out, up, down, off, ∅}
NN	{SG, PL}
SVA	{3SG, not 3SG, 2SG-Past, not 2SG-Past}

Table 1: Confusion sets for each grammar error type. The ART and PREP sets include an empty token (∅) allowing for insertions and deletions. SG, PL, 2SG, and 3SG stand for singular, plural, second-person and third-person singular respectively.

This is a two-step process, similar to the methods used in the GEC literature for creating synthetic data based on confusion matrices (Rozovskaya et al., 2014; Rozovskaya and Roth, 2010; Xie et al., 2016; Sperber et al., 2017). First, we mimic the distribution of errors found in real data, and then introduce errors by applying rule-based transformations on automatic parse trees.

The first step involves collecting error statistics on real data. Conveniently, the NUCLE corpus has all corrections annotated with 27 error codes. We focus on five types of errors, with the last four being the most common in the NUCLE corpus:

- DROP: randomly deleting one character from the sentence.[2]
- ART: article/determiner errors
- PREP: preposition errors
- NN: noun number errors
- SVA: subject-verb agreement errors

Using the annotated training set of the NUCLE corpus, we compute error distribution statistics, resulting in confusion matrices for the cases outlined in Table 1. For ART and PREP errors, we obtain probability distributions that an article, determiner, or preposition is deleted, substituted with another member of the confusion set, or inserted in the beginning of a noun phrase. For NN errors,

[2]This error is not part of the NUCLE error list.

Dataset	Percentage of Errors		
	Train	Dev	Test
sentences	2M	3K	3K
words	55M	74K	73K
DROP	100%	100%	100%
ART	96.4%	98.4%	99.8%
PREP	95.7%	95.9%	98.4%
NN	94.5%	91.0%	98.6%
SVA	93.1%	81.9%	82.0%
CLEAN+DROP	50%	50%	–
CLEAN+ART	48.2%	49.1%	–
CLEAN+PREP	47.8%	47.9%	–
CLEAN+NN	47.3%	45.5%	–
CLEAN+SVA	46.5%	41.0%	–
MIX-ALL	79.9%	77.8%	–

Table 2: Statistics on the original and synthetic En-Es datasets. Each (synthetic) sentence has exactly one introduced error, wherever possible. CLEAN+[ERROR] is the concatenation of the [ERROR] with the original clean dataset, while MIX-ALL includes six versions of each training sentence, one without errors and one for each error.

we obtain the probability of a noun being replaced with its singular or plural form. For SVA errors, the probability that a present tense verb is replaced with its third-person-singular (3SG) or not-3SG form. An additional SVA error that we included is the confusion between the appropriate form for the verb 'to be' in the past tense ('was' and 'were').

The second step involves applying the noise-inducing transformations using our collected statistics as a prior. We obtained parses for each sentence using the Berkeley parser (Petrov et al., 2006). The parse tree allows us to identify candidate error positions in each sentence (for example, the beginning of a noun phrase without a determiner, were one could be inserted). For each error type we introduced exactly *one* error per sentence, wherever possible, which we believe matches more realistic scenarios than previous work. It also allows for controlled analysis of the behaviour of the NMT system (see Section 4).

For each error and each sentence, we first identify candidate positions (based on the error type and the parse tree) and sample one of them based on the specific error distribution statistics. Then, we sample and introduce a specific error using the corresponding probability distribution from the confusion matrix. (In the case of DROP, NN, and SVA errors, we only need to sample the position and only insert/substitute the corresponding error.) If no candidate positions are found (for example, a sentence doesn't have a verb that can be substituted to produce a SVA error) then the sentence remains unchanged.

Following the above procedure, we added errors in our training, dev, and test set (henceforth referred to as [ERROR]). Basic statistics on our produced datasets can be found in Table 2, while example sentences are shown in Table 3. Furthermore, we created training and dev sets that mix clean and noisy data. The CLEAN+[ERROR] training sets are the concatenation of each [ERROR] with the clean data, effectively including a clean and a noisy version of each sentence pair.

We also created a training and dev dataset with mixed error types, in our attempt to study the effect of including all noise types during training. The MIX-ALL dataset includes each training pair six times: once with the original (clean) sentence as the source, and once for every possible error. We experimented with a mixed dataset that included each training sentence once, with the number of noisy sentences being proportional to the real error distributions of the NUCLE dataset, but obtained results similar to the [ERROR] datasets.

2.3 JFLEG-ES: Spanish translations of JFLEG

The JFLEG corpus consists of a dev and test set (no training set), with 747 and 754 English sentences, respectively, collected from non-native English speakers. Each sentence is annotated with four different corrections, resulting in four (fluent and grammatical) reference sentences. About 14% of the sentences do not include any type of error, with the source and references being equivalent.

We created translations of the JFLEG corpus that allow us to evaluate how well NMT fares compared to a human translator, when presented with noisy input. We will refer to the augmented JFLEG corpus as JFLEG-ES.

Two professional translators were tasked with producing translations for the dev and the test set, respectively. The translators were presented only with the original erroneous sentences; they did not have access to the correction annotations. They were asked to produce fluent, grammatical translations in European Spanish (to match the

Error Type	Example
ART	In October , Tymoshenko was sentenced to seven years in prison for entering into what was reported to be *a*/*∅ disadvantageous gas deal with Russia. Its ratification would require ∅/**the* 226 votes. It is *a*/**the* good result, which nevertheless involves a certain risk.
PREP	[. . .] the motion to revoke an article based *on*/**in* which the opposition leader , Yulia Tymoshenko , was sentenced. Its ratification would require ∅/**for* 226 votes.
NN	Its ratification would require 226 *votes*/**vote*. The *verdict*/**verdicts* is not yet final ; the court will hear Tymoshenko 's appeal in December.
SVA	As a rule, Islamists *win*/**wins* in the country; the question is whether they are the moderate or the radical ones. This cultural signature *accompanies*/**accompany* the development of Moleskine;

Table 3: Example grammatical errors that were introduced in the En-Es WMT test set.

Spanish used in the Europarl corpus). There exist cases where a translator might choose to preserve a source-side error when producing the translation, such as translation of literary works where it's possible that grammar or fluency errors are intentional; however, our translators were explicitly asked not to do that. The exact instructions were as follows:

> Please translate the following sentences. Note that some sentences will have grammatical errors or typos in English. Don't try to translate the sentences word for word (e.g. replicate the error in Spanish). Instead, try to translate it as if it was a grammatical sentence, and produce a fluent grammatical Spanish sentence that captures its meaning.

3 Experiments

In this section, we provide implementation details and the results of our NMT experiments. For convenience, we will refer to each model with the same name as the dataset it was trained on; e.g. the MIX-ALL model will refer to the model trained on the MIX-ALL dataset.

3.1 Implementation Details

All data are tokenized, truecased, and split into subwords using Byte Pair Encoding (BPE) with 32,000 operations (Sennrich et al., 2016). We filter the training set to only contain sentences up to 80 words.

Our LSTM models are implemented using DyNet (Neubig et al., 2017), and our transformer models using PyTorch (Paszke et al., 2017). The transformer model uses 6 layers, 8 attention heads, the dimension for embeddings and positional feed-forward are 512 and 2048 respectively . The sub-layer computation sequence follows the guidelines from Chen et al. (2018). Dropout probability is set to 0.2 (also in the source embeddings, following Sperber et al. (2017)). We use the learning rate schedule in Vaswani et al. (2017) with warm-up steps of 24000 but only decay the learning rate until it reaches 10^{-5} as inspired by Chen et al. (2018). For testing, we select the model with the best performance on the dev set corresponding to the test set. At inference time, we use a beam size of 4 with length normalization (Wu et al., 2016) with a weight of 0.6.

3.2 Results

We report the results obtained with the transformer model, as they were consistently better than the LSTM one. All the result tables for the LSTM models can be found in the Appendix.

The performance of our systems on the synthetic WMT test sets, as measured by detokenized BLEU (Papineni et al., 2002), is summarized in Table 4. When the system is trained only on clean data (first row) and tested on noisy data, it unsurprisingly exhibits degraded performance. We observe significant drops in the range of 1.0–3.6 BLEU.

WMT Training Set	En-Es WMT Test Set						
	CLEAN	DROP	ART	PREP	NN	SVA	AVERAGE ± STDEV
CLEAN	**33.0**	29.6	31.3	**32.0**	29.3	**32.1**	31.2 ± 1.5
DROP	31	30.2	30.0	30.0	28.3	30.6	30.0 ± 0.9
ART	31.2	28.4	30.8	30.2	27.7	30.8	29.8 ± 1.4
PREP	30.4	27.8	29.3	30.3	27.4	29.9	29.2 ± 1.3
NN	30.4	27.9	28.9	29.5	29.8	29.8	29.4 ± 0.8
SVA	31.2	28.7	30.2	30.3	28.2	30.9	29.9 ± 1.2
CLEAN+DROP	**32.9**	**31.4**	**31.4**	31.8	29.5	**32.0**	31.5 ± 1.2
CLEAN+ART	**32.7**	29.7	**31.7**	31.7	28.8	**32.1**	31.1 ± 1.5
CLEAN+PREP	**32.7**	29.6	31.2	**32.2**	29.0	31.8	31.1 ± 1.5
CLEAN+NN	32.5	29.4	30.7	31.4	**31.0**	31.6	31.1 ± 1.0
CLEAN+SVA	32.5	29.6	31.2	31.5	29.0	**31.9**	30.9 ± 1.4
MIX-ALL	**32.7**	30.9	**31.4**	**32.0**	**30.6**	**32.0**	**31.6 ± 0.7**

Table 4: BLEU scores on the WMT test set without (CLEAN) and with synthetic grammar errors. The best performing models for each test set are **highlighted**. When training and test match (highlighted) we generally observe higher results. However, including all clean and noisy data in the training set (MIX-ALL) yields the best results across almost all datasets, with the highest average BLEU *and* the lowest variance.

The largest drop (more than 3.5 BLEU) is observed with NN errors in the source sentence. This is not unreasonable: nouns almost always carry content significant for translation. Especially when translating into Spanish, a noun number change can, and apparently does, also affect the rest of the sentence significantly, for example, by influencing the conjugation of a subsequent verb. The second-largest drop (more than 3.0 BLEU points) is observed in the case of DROP errors. This is also to be expected; typos produce out-of-vocabulary (OOV) words, which in the case of BPE are usually segmented to a most likely rarer subword sequence than the original correct word.

We find that a training regime that includes both clean and noisy sentences ([CLEAN+ERROR) results in better systems across the board. Importantly, these models manage to perform en par with the CLEAN model on the CLEAN test set. Since the original training set is part of the [CLEAN+ERROR training sets, this behavior is expected. We conclude, thus, that including the full clean dataset during training is important for performance on clean data – one cannot just train on noisy data.

The [CLEAN+ERROR] systems exhibit a notable pattern: their BLEU scores are generally similar to the CLEAN system on all test sets, except for the test set that matches their training set errors (highlighted in Table 4), where they generally obtain the best performance.

The MIX-ALL model is our best system on all test sets (except DROP) and on average. Unlike the [CLEAN+ERROR] systems, it outperforms the CLEAN model on *all* noisy test sets and not only on a specific one. On average, using the MIX-ALL training set leads to an improvement of 0.4 BLEU over the CLEAN model and 0.1 – 0.7 BLEU over the [CLEAN+ERROR] models. Furthermore, the MIX-ALL model exhibits the smallest performance standard deviation of all models, averaging over all test sets. This is another indication that our system is more robust to multiple source-side variations. We further explore this intuition in Section 4.

On the more realistic JFLEG-ES dev and test sets, we observe same trends but at a smaller scale, as shown in Table 5. Our MIX-ALL model generally achieves comparable results when presented with each of the four reference corrections of the test set (corX columns). However, when we use the noisy source sentence as input (No CORR column) our MIX-ALL model obtains 1.4 BLEU improvements over the CLEAN model. The *difference* between the performance of the models when presented with clean and noisy input is another indicator for robustness. On the JFLEG-ES test set, the noisy source results in a −3.1 BLEU point drop for the CLEAN model, while the drop for our MIX-ALL model is smaller, at −1.7 BLEU points.

In addition, we experimented with using an automatic error-corrected source as input to our sys-

JFLEG-ES Dev

Training	Manual correction					No corr.	Auto corr.
	COR0	COR1	COR2	COR3	avg.		
CLEAN	32.1	31.5	32.5	33.3	32.4	31.1	31.2
MIX-ALL	31.9	31.4	32.2	**32.9**	**32.1**	**32.2**	**31.6**

JFLEG-ES Test

Training	Manual correction					No corr.	Auto corr.
	COR0	COR1	COR2	COR3	avg.		
CLEAN	**28.4**	**28.8**	**29.1**	**28.2**	**28.6**	26.2	**27.0**
MIX-ALL	27.7	28.1	28.1	27.5	27.8	**26.8**	**26.7**

Table 5: BLEU scores on the JFLEG-ES dev and test datasets. Our proposed MIX-ALL model is slightly behind the CLEAN model on manually corrected input (COR[0–3]). On noisy input (No corr.) the MIX-ALL outperforms the CLEAN model (26.8 > 26.2). Preprocessing the noisy input with a GEC model (Auto corr.) slightly improves results.

tem (column AUTO CORR of Table 5). We used the publicly available JFLEG outputs of the (almost) state-of-the-art model of Junczys-Dowmunt and Grundkiewicz (2016) as inputs to our NMT system.[3] This experiment envisions a pipeline where the noisy source is first automatically corrected and then translated. As expected, this helps the CLEAN model (by +1.1 BLEU), but our MIX-ALL training helps even further (by another +0.8 BLEU). Interestingly, the automatic GEC system only helps in the test set, while there are no improvements in the dev set. Naturally, since automatic GEC systems are imperfect, the performance of this pipeline still lags behind translating on clean data.

4 Analysis

We attempt an in-depth analysis of the impact of the different source-side error types on the behavior of our NMT system, when trained on clean data and tested on the artificial noisy data that we created.

ART Errors Table 6 shows the difference of the BLEU scores obtained on the sentences, broken down by the type of article error that was introduced. The first observation is that in all cases the difference is negative, meaning that we get higher BLEU scores when testing on clean data. Encouragingly, there is practically no difference when we substitute 'a' with 'an' or 'an' with 'a'; the model seems to have learned very similar representations for the two indefinite articles, and as a result such an error has no impact on the produced output. However, we observe larger performance drops when substituting indefinite articles with the definite one and vice versa; since the target language makes the same article distinction as the source language, any article source error is propagated to the produced translation.

PREP Errors Due to the large number of prepositions, we cannot present a full analysis of preposition errors, but highlights are shown in Table 7. Deleting a correct preposition or inserting a wrong one leads to performance drops of 1.2 and 0.8 BLEU points for the CLEAN model, but drops of 0.4 and 0.7 for the MIX-ALL model.

NN and SVA Errors We found no significant performance difference between the different NN errors. Incorrectly pluralizing a noun has the same adverse effect as singularizing it, leading to performance reductions of over 4.0 and 3.5 BLEU points respectively. We observe a similar behavior with SVA errors: each error type leads to roughly the same performance degradation.

5 Related Work

The effect of noise in NMT was recently studied by Khayrallah and Koehn (2018), who explored noisy situations during training due to web-crawled data. This type of noise includes misaligned, mistranslated, or untranslated sentences which, when used during training, significantly degrades the performance of NMT. Unlike our

[3] This model has been recently surpassed by other systems, e.g. (Junczys-Dowmunt et al., 2018), but their outputs are not available online.

Correct article	Substituted article				
	a	an	the	∅	*all*
a	–	0	−2.0	−2.1	−2.1
an	0	–	−5.7	−7.3	−6.3
the	−4.1	−2.2	–	−1.7	−1.8
∅	−3.1	−3.7	−1.5	–	−1.7
all	−3.8	−3.4	−1.5	−1.8	−1.7

Table 6: Effect of article substitutions in test data (ART) relative to clean test data (CLEAN), broken down by substitution type. Different article substitutions have very different impacts on BLEU; changing an indefinite article to definite is especially damaging.

Substitution	model BLEU difference	
	CLEAN	MIX-ALL
in→with	−6.7	−1.7
on→for	−6.0	−0.1
to→on	−2.9	−0.5
in→ ∅	−1.8	−1.9
∅→for	−1.6	−0.6
∅→*any*	−1.2	−0.4
any→∅	−0.8	−0.7

Table 7: Effect of selected preposition substitutions in test data (PREP) relative to clean test data (CLEAN), for the CLEAN and MIX-ALL models. The MIX-ALL model handles most errors more efficiently.

work, they primarily focus on a setting where the training set is noisy but the test set is clean.

In addition, Heigold et al. (2018) evaluated the robustness of word embeddings against word scrambling noise, and showed that performance in downstream tasks like POS-tagging and MT is especially hurt. Sakaguchi et al. (2017a) studied word scrambling and the *Cmabrigde Uinervtisy (Cambridge University) effect*, where humans are able to understand the meaning of sentences with scrambled words, performing word recognition (word level spelling correction) with a semi-character RNN system.

Focusing only on character-level NMT models, Belinkov and Bisk (2018) showed that they exhibit degraded performance when presented with noisy test examples (both artificial and natural occurring noise). In line with our findings, they also showed that slightly better performance can be achieved by training on data artificially induced with the same kind of noise as the test set.

Sperber et al. (2017) proposed a noise-introduction system reminiscent of WER, based on insertions, deletions, and substitutions. An NMT system tested on correct transcriptions achieves a BLEU score of 55 (4 references), but tested on the ASR transcriptions it only achieves a BLEU score of 35.7. By introducing similar noise in the training data, they were able to make the NMT system slightly more robust. Interestingly, they found that the optimal amount of noise on the training data is smaller than the amount of noise on the test data.

The notion of linguistically plausible corruption is also explored by Li et al. (2017), who created adversarial examples with syntactic and semantic noise (reordering and word substitutions respectively). When training with these noisy datasets, they obtained better performance on several text classification tasks. Furthermore, in accordance with our results, their best system is the one that combines different types of noise.

We present a summary of relevant previous work in Table 8. *Synthetic* errors refer to noise introduced according an artificially created distribution, and *natural* errors refer to actual errorful text produced by humans. As for *semi-natural*, it refers to either noise introduced according to a distribution learned from data (as in our work), or to errors that are learned from data but introduced according to an artificial distribution (as is part of the work of Belinkov and Bisk (2018)).

We consider our work to be complementary to the works of Heigold et al. (2018); Belinkov and Bisk (2018), and Sperber et al. (2017). However, there are several important differences:

1. Belinkov and Bisk (2018) and Sperber et al. (2017) train their NMT systems on fairly small datasets: 235K (Fr-En), 210K (De-En), 122K (Cz-En), and 138K sentences (Es-En) respectively. Even though they use systems like Nematus (Sennrich et al., 2017) or XNMT (Neubig et al., 2018) which generally achieve nearly SOTA results, it is unclear whether their results generalize to larger training data. In contrast, we train our system on almost 2M sentences.
2. All three systems introduce somewhat unrealistic amounts of noise in the data. The natural noise of Belinkov and Bisk (2018) consists of word substitutions based on Wikipedia errors or corrected essays (in the

Work	Errors	Noise Types	NMT level	Languages
(Heigold et al., 2018)	synthetic	character swaps, character flips, word scrambling	char, BPE	De→En
(Sperber et al., 2017)	synthetic	ASR errors	word	Es→En
(Belinkov and Bisk, 2018)	synthetic	character swap, middle scramble, full scramble, keyboard typo	char, BPE	Fr,De,Cz→En
	semi-natural	word substitutions		
this work	semi-natural	grammar errors: article, preposition, noun number, verb agreement	BPE	En→Es
	natural	JFLEG corpus		

Table 8: Previous work on evaluating the effect of noise in NMT systems. Character swaps refer to neighboring character reordering (e.g. noise→nosie), while character flips refer to character substitutions (e.g. noise→noiwe).

Czech case) but they substitute all possible correct words with their erroneous version, ending up with datasets with more than 40% of the tokens being noisy. For that reason, we refer to it as *semi-natural* noise in Table 8. Meanwhile, Sperber et al. (2017) test on the outputs of an ASR system that has a WER of 41.3%. For comparison, in the JFLEG datasets, we calculated that only about 3.5%–5% of the tokens are noisy – the average Levenshtein distance of a corrected reference and its noisy source is 13 characters.

3. The word scrambling noise, albeit interesting, could not be claimed to be applicable to realistic scenarios, especially when applied to all words in a sentence. The solution Belinkov and Bisk (2018) suggested and Sperber et al. (2017) discussed is a character- or spelling-aware model for producing word- or subword-level embeddings. We suspect that such a solution would indeed be appropriate for dealing with typos and other character-level noise, but not for more general grammatical noise. Our method could potentially be combined with GloVe (Pennington et al., 2014) or fastText (Bojanowski et al., 2017) embeddings that can deal with slight spelling variations, but we leave this for future work.

On the other side, Grammar Error Correction has been extensively studied, with significant incremental advances made recently by treating GEC as an MT task: among others, Junczys-Dowmunt and Grundkiewicz (2016) used phrased-based MT, Ji et al. (2017) used hybrid character-word neural sequence-to-sequence systems, Sakaguchi et al. (2017b) used reinforcement learning, and Junczys-Dowmunt et al. (2018) combined several techniques with NMT to achieve the current state-of-the-art. Synthetic errors for training GEC systems have also been studied and applied with mixed success (Rozovskaya and Roth, 2010; Rozovskaya et al., 2014; Xie et al., 2016), while more recently Xie et al. (2018) used backtranslation techniques to add synthetic noise for GEC.

6 Conclusion

In this work, we studied the effect of grammatical errors in NMT. We not only confirmed previous findings, but also expanded on them, showing that *realistic human-like* noise in the form of specific grammatical errors also leads to degraded performance. We added synthetic errors on the English WMT training, dev, and test data (including dev and test sets for all WMT 18 evaluation pairs), and have released them along with the scripts necessary for reproducing them. We also produced Spanish translations of the JFLEG corpus, so that future NMT systems can be properly evaluated on real noisy data.

Acknowledgments

This material is based upon work generously supported by the National Science Foundation under grants 1464553 and 1761548. We are grateful to the anonymous reviewers for their useful comments.

References

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 Conference on Machine Translation. In *Proc. WMT*, pages 131–198.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, et al. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proc. ACL*, pages 76–86.

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*, pages 22–31.

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. ArXiv:1803.05567.

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*, volume 1, pages 68–80.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The SOCKEYE neural machine translation toolkit at AMTA 2018. *Proc. AMTA*, page 200.

Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao. 2017. A nested attention neural hybrid model for grammatical error correction. In *Proc. ACL*, pages 753–762.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In *Proc. EMNLP*, pages 1546–1556.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018. Approaching neural grammatical error correction as a low-resource machine translation task. In *Proc. NAACL-HLT*, pages 595–606.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proc. WNMT*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. MT Summit*, pages 79–86.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proc. WNMT*, pages 28–39.

M Paul Lewis, Gary F Simons, Charles D Fennig, et al. 2009. *Ethnologue: Languages of the world*, volume 16. SIL International.

Yitong Li, Trevor Cohn, and Timothy Baldwin. 2017. Robust training under linguistic adversity. In *Proc. EACL*, pages 21–27.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proc. EACL*, pages 229–234.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, et al. 2017. DyNet: The dynamic neural network toolkit. arXiv:1701.03980.

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Singh Sachan, Philip Arthur, Pierre Godard, et al. 2018. XNMT: The eXtensible Neural Machine Translation toolkit. arXiv: 1803.00188.

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*, pages 1–14.

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*, pages 1–12.

Toan Nguyen and David Chiang. 2018. Improving lexical choice in neural machine translation. In *Proc. NAACL HLT*, pages 334–343.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Proc. NeurIPS Autodiff Workshop*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proc. EMNLP*, pages 1532–1543.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proc. COLING-ACL*, pages 433–440.

Alla Rozovskaya, Kai-Wei Chang, Mark Sammons, Dan Roth, and Nizar Habash. 2014. The Illinois-Columbia system in the CoNLL-2014 shared task. In *Proc. CoNLL*, pages 34–42.

Alla Rozovskaya and Dan Roth. 2010. Generating confusion sets for context-sensitive error correction. In *Proc. EMNLP*, pages 961–970.

Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2017a. Robsut wrod reocginiton via semi-character recurrent neural network. In *Proc. AAAI*, pages 3281–3287.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2017b. Grammatical error correction with neural reinforcement learning. In *Proc. IJCNLP*, pages 366–372.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, et al. 2017. Nematus: a toolkit for neural machine translation. In *Proc. EACL*, pages 65–68.

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

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proc. IWSLT*.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for ESL learners using global context. In *Proc. ACL*, pages 198–202.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proc. NeurIPS*, pages 5998–6008.

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.

Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, and Andrew Y. Ng. 2016. Neural language correction with character-based attention. arXiv: 1603.09727.

Ziang Xie, Guillaume Genthial, Stanley Xie, Andrew Ng, and Dan Jurafsky. 2018. Noising and denoising natural language: Diverse backtranslation for grammar correction. In *Proc. NAACL HLT*, pages 619–628.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proc. ACL-HLT*, pages 180–189.

A Results with LSTM models

Training Set	En-Es WMT Test Set						
	CLEAN	DROP	ART	PREP	NN	SVA	AVERAGE ± STDEV
CLEAN	**26.62**	24.08	25.35	25.63	23.34	**26.06**	25.18 ± 1.24
DROP	25.10	24.21	24.24	24.00	22.26	19.58	23.23 ± 2.02
ART	25.49	23.26	24.78	24.35	22.42	25.59	24.31 ± 1.26
PREP	25.49	22.99	24.39	25.22	22.78	25.07	24.32 ± 1.17
NN	25.35	23.04	23.06	24.15	24.73	24.61	24.16 ± 0.94
SVA	25.77	23.49	24.68	24.62	23.22	25.41	24.53 ± 1.01
CLEAN+DROP	**26.45**	**25.37**	25.59	25.59	23.64	25.92	25.43 ± 0.95
CLEAN+ART	**26.64**	24.60	**26.35**	26.08	23.69	26.48	**25.64 ± 1.21**
CLEAN+PREP	**26.60**	24.31	25.12	**26.30**	23.27	**26.14**	25.29 ± 1.31
CLEAN+NN	26.23	23.86	24.75	25.52	**25.20**	25.66	25.20 ± 0.82
CLEAN+SVA	**26.62**	24.22	25.49	25.86	23.79	**26.24**	25.37 ± 1.13
MIX-ALL	**26.60**	24.90	25.52	25.80	24.68	**26.03**	**25.59 ± 0.72**

Table 9: BLEU scores on the WMT test set without (CLEAN) and with synthetic grammar errors using an LSTM encoder-decoder model.

JFLEG-ES Dev

Training	Manual correction					No corr.	Auto corr.
	COR0	COR1	COR2	COR3	avg.		
CLEAN	**28.3**	27.3	28.4	28.2	28.0	27.1	27.7
MIX-ALL	28.2	**27.5**	**28.8**	**29.1**	**28.4**	**27.4**	**28.2**

JFLEG-ES Test

Training	Manual correction					No corr.	Auto corr.
	COR0	COR1	COR2	COR3	avg.		
CLEAN	**24.9**	**25.1**	**25.6**	**25.1**	**25.2**	22.8	23.5
MIX-ALL	24.8	25.0	25.3	25.0	25.0	**23.1**	**24.3**

Table 10: BLEU scores on the JFLEG-ES dev and test datasets with the LSTM encoder-decoder model.