Learning from Mistakes: Towards Robust Neural Machine Translation for Disfluent L2 Sentences

Shuyue Stella Li sli136@jhu.edu
Philipp Koehn phi@jhu.edu
Johns Hopkins University, Baltimore, MD 21218, USA

Abstract

We study the sentences written by second-language (L2) learners to improve the robustness of current neural machine translation (NMT) models on this type of data. Current large datasets used to train NMT systems are mostly Wikipedia or government documents written by highly competent speakers of that language, especially English. However, given that English is the most common second language, it is crucial that machine translation systems are robust against the large number of sentences written by L2 learners of English. By studying the difficulties faced by humans in their L2 acquisition process, we are able to transfer such insights to machine translation systems to recover from source-side fluency variations. In this work, we create additional training data with artificial errors similar to mistakes made by L2 learners of various fluency levels to improve the quality of the machine translation system. We test our method in zero-shot settings on the JFLEG-es (English→Spanish) dataset. The quality of our machine translation system on disfluent sentences outperforms the baseline by 1.8 BLEU scores.

1 Introduction

Neural machine translation (NMT) is a supervised learning problem that has been widely studied and achieved great success in numerous benchmarks (Koehn, 2020; Stahlberg, 2020; Bahdanau et al., 2014). Its power comes from learning high-level representations of meaning, which often relies on massive amounts of clean, parallel data. However, tiny perturbations of the data result in cascading degradation in the performance of the NMT model (Belinkov and Bisk, 2018; Cheng et al., 2018). Unlike humans who are able to ignore small discrepancies in trivial spelling and grammar errors, NMT systems still need to solve this crucial problem.

The noise in the data can come from various sources. The particular type of noise that we investigate in this work is when the source sentences of an NMT system are written by L2 learners of a language. Since the largest parallel corpora are mostly Wikipedia or government documents written by fluent speakers of that language, the L2 sentences are different from the ones seen by most machine translation models. Second, L2 learners come from different first language (L1) environments, bringing their own unique linguistic habits and cultural references into composing L2 sentences. Third, the collection and annotation of such data are difficult due to the linguistic diversity of the sentences. Therefore, the main challenges of translating L2 sentences are that they are not fluent, out-of-domain, and extremely low-resource.

When translating from low-flouency source sentences, NMT systems are especially prone to fail in the presence of highly noisy data. It might be trivial for humans to understand a sentence with grammar or spelling errors. But higher-level mistakes, such as unconventional usage of

Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track, pages 235–247
September 4–8, 2023, Macau SAR, China.

phrases, could be hard to understand even for humans. As shown in Figure 1, a good NMT system should ideally learn such disfluencies and ambiguities so that it can help both the L2 speaker better express themselves and the listener better understand the output.

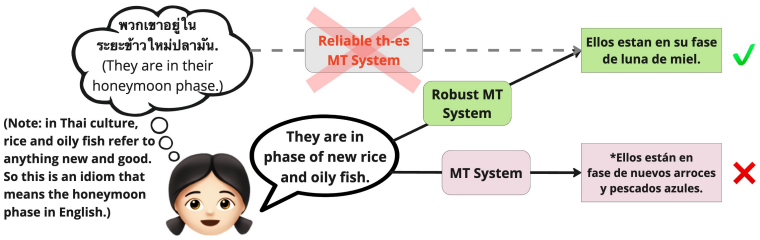


Figure 1: Robust NMT System from Disfluent English to Spanish. The L1 of the user is Thai. A robust English→Spanish NMT system is needed in the absence of a reliable Thai→Spanish system. The L2 English sentence she produces contains several L2 errors (e.g. missing article & phrase misuse). Thus, the goal of the robust NMT system is to recover the original meaning that the speaker *meant* to express.

We attempt to build an NMT model robust to disfluencies by first studying the mistakes made by the L2 learners on the word, phrase, and sentence levels. We compute detailed L2 error statistics that realistically resemble a cognitively-grounded second language acquisition (SLA) process. Then, we artificially inject the observed common errors into the clean training data to create a synthetic L2 training dataset. Our proposed artificial error augmentation method does not need any gold translations (except for the test set during evaluation) and can therefore be applied to extremely low-resource settings. In this work, our contributions include:

- We extensively analyze the writing errors produced by real L2 learners to study the second language acquisition process. We introduce a framework for creating written second language acquisition modeling that could be useful for a variety of applications.
- We propose a realistic error augmentation approach that incorporates low-level to high-level L2 errors and is target-language-agnostic. The data augmentation is able to improve the generalizability and robustness of the model even without labeled disfluent training data.
- Our experimental results show that error augmentation is extremely helpful. We observe an increase in the 1.8 BLEU score in the English→Spanish direction. We make our code and the generated silver dataset publicly available[1].

2 Related Work

2.1 Robust Machine Translation

Robust machine translation with noisy data has been a challenging research problem in the field of natural language processing (Belinkov and Bisk, 2018). The WMT Shared Tasks on Machine Translation Robustness (Li et al., 2019; Specia et al., 2020) aim to develop NMT models that can successfully handle real-world noises. One line of research focuses on using data augmentation techniques to generate additional training data. Some approaches add synthetic noise to the training data (Berard et al., 2019; Abdul Rauf et al., 2020). Several studies have explored the use of unsupervised and semi-supervised learning techniques for robust machine translation (Lample et al., 2018; Artetxe et al., 2018; Cheng and Cheng, 2019). Back-translation is commonly used as a bootstrapping method to augment training data and thus improve machine translation quality (Sennrich et al., 2015; Chauhan et al., 2022). Iterative methods can also be used to improve the quality of the back-translation (Hoang et al., 2018). Adversarial inputs have been widely used as a data augmentation approach for robust NMT and other NLP problems (Cheng et al., 2019; Hsu

[1] https://github.com/stellali7/L2MT

et al., 2022). These methods aim to leverage the vast amounts of unlabeled data available for machine translation and reduce the reliance on annotated data. However, one problem with many data augmentation methods is that it relies on an existing machine translation system, and might not be realistic and specific to the target domain of interest.

2.2 L2 Language Processing

Disfluent and ungrammatical sentences written by second-language learners can also be considered a source of noise. This is because most datasets do not contain sentences like these, and cannot generalize to irregular sentence formations that are determined by the L2 competency level and the L1 of the speakers. Existing work on disfluent sentences involves training parsing models on ungrammatical data (Hashemi and Hwa, 2016), or jointly training on a combination of clean and synthetic ungrammatical sentences (Anastasopoulos et al., 2019). Another approach incorporates the explicit syntactic and semantic structures into the NMT models to better handle disfluent sentences (Liu et al., 2021; Chen et al., 2017; Zhang et al., 2019a).

Most existing methods regarding disfluent or L2 data focus on creating low-level grammatical errors and have made minimal efforts on cognitive modeling of the actual SLA process. The Duolingo Second Language Acquisition Modeling Challenge (Settles et al., 2018) aims to combine knowledge of cognitive science, linguistics, and machine learning, but its scope is limited to token-level prediction and beginner-level language learner data. Even in the cognitive science literature, second language acquisition is largely studied in terms of spoken utterances rather than written sentences and focuses on individual case studies (Krashen, 1981). In our work, we propose cognitively-grounded errors to better model the SLA process.

2.3 Grammatical Error Correction

Grammatical error correction is closely related to disfluent sentence processing. Several robust machine translation approaches for disfluent sentences use a cascaded system to first correct the grammatical errors in the source sentence and then translate them into the target language (Anastasopoulos et al., 2019). There exist a number of publicly available corpora for grammatical error correction, including the NUCLE dataset of Singaporean English learners (Dahlmeier et al., 2013), the CoNLL-2014 GEC shared task dataset (Ng et al., 2014), and the ErAConD dialogue GEC dataset (Yuan et al., 2022). However, they mostly focus on error-coding rule-based grammatical errors.

In summary, unconstrained L2 generation and processing remain relatively underexplored, and our work attempts to study this topic in order to build a robust NMT system. Our work is inspired by the work done by Belinkov and Bisk (2018) and Anastasopoulos et al. (2019), but we focus on identifying more realistic and higher-level errors to model written SLA and thus create an artificial error augmentation corpus that closely resembles real L2 data.

3 Methods

3.1 Overview

With the main goal of training an NMT system robust to disfluent sentences, our work focuses on artificially generating data that are similar in distribution to the disfluent sentences. Through such artificial error augmentation, our NMT models can learn to be more robust to input sentences of various qualities. First, we study the types of mistakes and inconsistencies that contribute to disfluency. To do this, we manually annotate the fluency corrections that rewrite disfluent L2 sentences into native-sounding sentences. We learn an L2 error distribution from categorizing the corrections. Then we use this learned distribution to generate disfluent sentences from well-formed sentences by injecting syntactic, semantic, and sentence-level errors. Since the goal of the translation is to recover the speaker intent in the target language despite disfluency errors, we use the translation of the well-formed sentence as the pseudo-translation of the transformed

erroneous sentence. Finally, we train an NMT system on a combination of the clean (fluent) parallel corpus and our artificial disfluent pseudo-parallel corpus.

3.2 Disfluency Error Analysis

We study the SLA process, including the common errors made by L2 learners, by analyzing the disfluent sentences and their correction references. We summarize the holistic fluency rewrites into different error types according to the underlying cognitive discontinuity in the L1-L2 switching process (e.g. grammar, semantic, or usage differences between L1 and L2). Figure 2 shows two real examples of disfluent L2 sentences, their revisions, and the errors contained in each sentence. By modeling the mapping of disfluent sentences to their corrections, we construct a written SLA model containing information and the likelihood of occurrence of each type of L2 error.

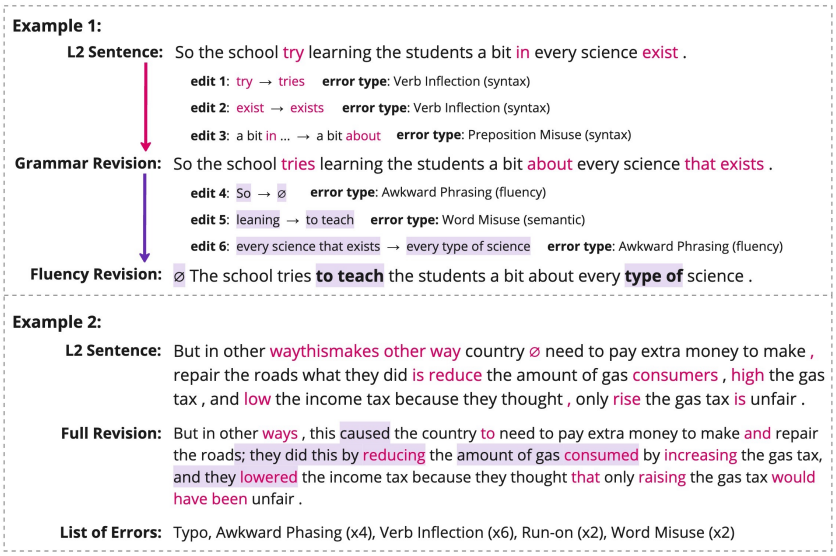


Figure 2: Examples of L2 Learner Errors that Contribute to Disfluency

We randomly select 100 sentences from the JFLEG dataset to perform manual error annotations (Napoles et al., 2017). This dataset contains disfluent sentences composed by L2 learners of English with a broad range of proficiency levels and various (but unspecified) native languages. For each disfluent sentence, it contains 4 corrections independently written by qualified fluent English speakers. Although the dataset is originally developed for grammatical error correction, the reference sentences not only correct the grammar mistakes but also make fluency rewrites that holistically improve the fluency and readability of the sentences. For example, some L2 learners tend to concatenate word-level translations when translating a phrase or idiom in their L1 into the target language, resulting in *awkward* (non-native-sounding) phrases. In the JFLEG annotations, the *awkward* words and phrases are rewritten to sound more *natural* (native-sounding) with the interpretation of the annotators. These annotations are extremely crucial in studying errors in the second language acquisition process, as it entails mistakes beyond being able to simply follow syntactic rules and grammar constructs.

Upon preliminary inspection, we summarize the major errors that contribute to the disfluency of sentences into the three main categories - *Grammar errors, Semantic errors and Fluency errors* - with subcategories for each type of error as shown in Table 1. First, grammar errors are low-level mistakes that violate grammatical rules. Beginner L2 learners tend to make these mistakes due to not being familiar with the grammar of the L2. Although the L1 of the writers of our dataset is various and unknown, this type of error is particularly common when the L1

Error Type	Subtype	Description
Grammar Errors	Verb Inflection	Incorrect tense/form or subject-verb agreement.
	Preposition Misuse	Wrong/missing preposition (typically in verb phrases).
	Article Misuse	Wrong or missing articles.
	Noun Form	Singular vs. Plural forms nouns, including special cases such as teeth→tooth.
Semantic Errors	Word Misuse	Made-up/wrong word or phrase, usually a synonym in their L1 but changes meaning in L2.
	Typo	Swapping of two adjacent letters; substitution of a letter with another; injection or deletion of a letter.
Fluency Errors	Awkward Phrasing	Uncommon usage that sounds unnatural to native speakers, but grammatically and semantically correct.
	Run-on/Long Sent.	Long and confusing sentences that would typically be broken into shorter sentences by native speakers.

Table 1: Common Error Types in Disfluent L2 Sentences.

of the learners is less *marked* ("easier") than the L2, resulting in a negative transfer (Eckman, 1977; Benson, 1986). For example, articles ('the', 'a', 'an') do not exist in Russian but exist in English, making articles less marked in Russian than in English. Therefore, it is difficult for Russian native speakers to use articles correctly when producing English sentences. Hence in our written SLA model, we attempt to determine the probability distribution of the grammar errors by approximating the markedness of the set of first languages of the L2 learners and English on four linguistic phenomena: verb inflection, preposition misuse, article misuse, and noun form.

The second category of errors is semantic errors, where the word or phrase of interest alters the meaning of the sentence. This is due to the lack of knowledge of the L2 language and its vocabulary usage. When there are multiple valid literal translations of a word from the source to the target language, the L2 learner might choose one arbitrarily without knowing the common combinations of phrases. Often, distinguishing which word to use out of a set of synonyms is a harder challenge than being familiar with the grammar rules, because it requires a more subtle understanding of the L2 language rather than a rigid memorization of rules. For example, such a word misuse, although grammatically acceptable, causes confusion even for humans (Edit 5 in Example 1 of Figure 2). Thus, we consider semantic errors higher-level than grammar errors. Additionally, we also classify typos as semantic errors. Although the cause of such errors is not the lack of semantic knowledge, typos can change the semantics of the sentence drastically for neural networks as it causes OOVs during tokenization (Belinkov and Bisk, 2018).

The last but arguably the most important category of errors is fluency errors. Although most grammatical error correction models can solve most of the above errors, revising fluency errors requires a more in-depth understanding of the language. During our annotation process, sentences with revisions that involve long-span word rearrangement or rewriting and sentences whose revisions differ largely from the source are labeled with "Awkward Phasing" errors. Note that awkward here means non-native sounding and is not related to the semantics of the sentence content. Furthermore, another common trend of the L2 sentences is that a number of the corrections broke down run-on or long sentences into shorter segments to make the sentence less confusing. To model the fluency errors, we compute the average number of sentences that each run-on or long L2 sentence breaks into shorter sentences and the percentage of lines marked with the Awkward Phrasing error.

3.2.1 Disfluency Error Analysis Results

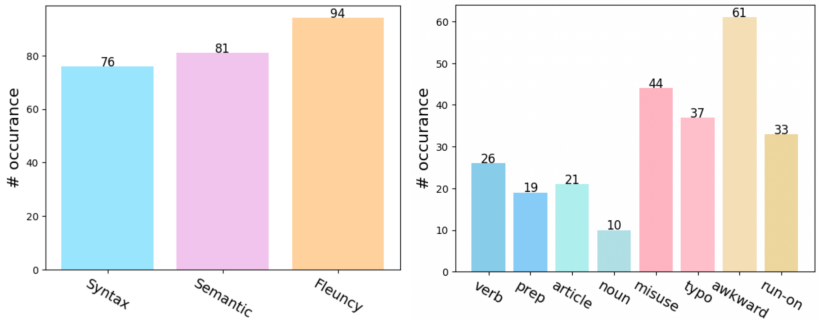
Table 2 summarizes the error distributions by each subtype. 90% of the sentences have at least one error, and each sentence has 2.5 errors on average. The percentage of tokens for each error is used in the error augmentation process to create realistic error distributions, and the percentage of lines in which each error occurs is used to include multiple errors in one synthetic dataset by

single-error mixture or in-line compounding (discussed in detail in Section 3.4).

Error Subtype	Total #	% of tokens*	% of lines
Verb Inflection	26	3.6	23
Preposition Misuse	19	4.4	18
Article Misuse	21	5.5	17
Noun Form	10	7.4	10
Word Misuse	44	3.7	35
Typo	37	4.2	31
Awkward Phrasing	61	3.7	47
Run-on/Long Sent.	33	3.3	27

Table 2: Disfluency Error Statistics. *Percent of tokens is calculated out of all the sentences with the error of interest. The values are used to generate realistic errors from clean texts.

The disfluency error distribution is visualized in Figure 3, which plots the frequency of manually annotated errors across 100 randomly sampled sentences from the development split of the JFLEG dataset. L2 sentences have 76 grammar errors, 81 semantic errors, and 94 fluency errors as plotted in Figure 3a. The more advanced an L2 learner is in their language study, the less likely they will make low-level errors, yet beginner-level learners tend to make all types of errors. Therefore, it is not surprising that fluency errors have the highest number of occurrences. Additionally, the distribution of error subtypes in Figure 3b provides a detailed breakdown of the errors that make the sentence grammatically incorrect or non-native sounding. Word Misuse and Awkward Phrasing errors are particularly common. This is partially because of the lack of familiarity and exposure to the proper or natural way to use their L2. However, it can also be attributed to the error coding method, which labels an alternative usage of words in the correction rewrites as a Word Misuse error and labels longer range rearrangement/rewrites as Awkward Phrasing errors. Since the rewrites have minimal constraints, a higher degree of freedom would cause more diverse rewrites, and thus more errors.



(a) Error Type Distribution (b) Error Subtype Distribution

Figure 3: Disfluency Error Distributions

3.3 Error Generation

3.3.1 Grammar Error Augmentation

After learning the distribution of disfluency errors in L2 sentences, we design algorithms to recreate the errors and inject them into clean, well-formed English sentences. For each type of sub-error, we first parse the clean sentences with the Berkeley Parser (Petrov et al., 2006) to determine the potential error injection sites. We create one grammar error per line using the script provided by Anastasopoulos et al. (2019). Lastly, the relative ratios of the grammatical errors as shown in the error subtype distribution plot in Figure 3b are used as weights to randomly sample the erroneous lines to form the synthetic Grammatical Error dataset.

3.3.2 Rule-based Typo Error Augmentation

We simulate typos with character-based swapping, substitution, insertion, and deletion edits. Swapping edits switch the order of two adjacent characters within a word. Substitution errors are generated using the statistics of most frequently switched letters on the English QWERTY keyboard Berry (2012). Lastly, insertion and deletion errors are generated at random. We only insert the typos in sentences with at least 4 words and words with at least 6 letters to avoid making the augmentation unnecessarily noisy. From Table 2, 4.2% of the tokens in a sentence contain typos, so we set the probability of randomly selecting a token to insert typos to match the real distribution.

3.3.3 Run-on Error Augmentation

Note that Run-on errors refer to both grammatically sound but confusing long sentences and ungrammatical run-on sentences with more than one main verb. We generate run-on or long sentences by first converting lines in the clean data with multiple sentences into one sentence joined by ',', ';', or ", and'". This is the most natural way to create run-on sentences, as the topic of each sentence in a line is similar. However, there is a limited number of lines with multiple sentences. So we perform a more aggressive data augmentation where each initial sentence S_i is appended by n randomly selected sentences from the same batch of size B. The number of sentences, n, to append to the initial sentence is determined as follows:

$$P(N = n \,|\, p, B) = \frac{p^n}{\sum_{i=1}^{B} p^i}, \tag{1}$$

where p is computed to match the realistic error distribution in Section 3.2.1. Specifically, if a sentence contains Run-on/Long Sentence errors, it has 1.22 of them on average, meaning that 1.22 extra sentences should be appended to the initial sentence.

3.3.4 Embedding-based Fluency Error Augmentation

Lastly, we group Word Misuse errors and Awkward Phrasing errors together, because it is often a fine line to determine when a phrase is "misused" or just not natural sounding enough. We use the Parrot[2] utterance augmentation framework to create paraphrases of the clean text in the same overall semantic space to simulate Word Misuse and Awkward Phrasing errors (Damodaran, 2021). Parrot is based on T5 and fine-tuned on paraphrase datasets. In the Parrot framework, the levels of adequacy and fluency can be adjusted to fit the goal of paraphrase generation. Adequacy is the degree to which the meaning of the sentence is preserved; whereas fluency measures how well-formed is the paraphrased sentence. Through preliminary experimentation, we set the adequacy and fluency thresholds to (0.3, 0.6) to generate Word Misuse errors and (0.7, 0.3) to generate Awkward Phrasing errors. For both generation tasks, we set Diversity to true in order to lessen the constraints on the generation. Parrot outputs the generated paraphrases in descending order of "diversity scores." We generate five paraphrases for each clean source sentence and randomly select one as the final output in order to introduce more nondeterministic variations in the generation process and encourage data diversity.

3.4 Error Combination

We propose two methods to combine different errors into one synthetic dataset. The first method simply generates one type of error per line and combines the lines into one dataset so that we have a multi-error dataset with single-error lines. The ratio of lines of each error type is determined by *% of lines* in Table 2. We call this method *Single-Error Mixture*. The second method models how L2 learners create errors more realistically. It compounds different errors in one line. We refer to this method as the *In-line Compounding*. Both combination schemes are explored in our experiments in different scenarios.

[2] https://github.com/PrithivirajDamodaran/Parrot_Paraphraser

4 Experimental Setup

4.1 Datasets

When choosing the source and target languages, the practical utility of our system is considered. English is the most common second language (Alemi, 2016), and Spanish is one of the most commonly translated languages[3]. Therefore, our work focuses on the English→Spanish translation direction. We use the English-Spanish Europarl dataset (Koehn, 2005) as the raw data to which we inject artificial errors. It contains 2,012,343 parallel sentence pairs.

To evaluate the robustness of the NMT model, we take 1,501 parallel sentences from the JFLEG corpus (28,106 words) (Napoles et al., 2017) and the JFLEG-es corpus (25,685 words) (Anastasopoulos et al., 2019). The JFLEG corpus is a selection of the GUG corpus, which is composed by L2 learners with a broad range of English proficiency levels and first languages, where the first languages of the writers are not disclosed (Heilman et al., 2014). In the JFLEG dataset, each disfluent sentence is annotated with four holistic fluency rewrites, making the JFLEG dataset unique as it corrects the disfluent sentences not only to void the grammar mistakes but also to make them natural-sounding. Anastasopoulos et al. (2019) extends the JFLEG into JFLEG-es dataset by manually translating the L2 sentences into Spanish, providing limited but valuable gold-standard translations.

4.2 Training Setup

Preprocessing To preprocess the data, we remove extra white spaces, preserve the casing, and tokenize with the SentencePiece[4] tokenizer into Byte-Pair-Encoding (BPE) with a vocab size of 50k (Sennrich et al., 2016). In each experiment, the L2 (disfluent) and correction (fluent) sentences are tokenized with the BPE model trained on the same training dataset used to train the NMT model. Then, the standard Fairseq preprocessing routine is used to further preprocess and binarize the data (Ott et al., 2019).

Training We use a simple transformer architecture with 4 encoder layers, 4 decoder layers, an embedding dimension of 512, a feed-forward dimension of 2048, 4 encoder attention heads, and 4 decoder attention heads. We apply a dropout of 0.3 and use the Adam optimizer with an epsilon value of 1e-6, betas of 0.9 and 0.98 (Kingma and Ba, 2015). We use the inverse square root learning rate scheduler following Vaswani et al. (2017) with an initial learning rate of 1e-7, 8000 warm-up steps to reach the target learning rate 4e-4. We train for 200000 steps and 8192 tokens per batch with an early stopping if the dev metric (BLEU score) does not improve in 4 epochs. During decoding, we use a beam size of 5.

Evaluation The evaluation results are measured with multiple metrics in order to present a more comprehensive set of comparisons. We use the detokenized BLEU score (Papineni et al., 2002) provided by SacreBLEU (Post, 2018), translation edit rate (TER), which measures the amount of editing required to match the reference (Snover et al., 2006), and BERTScore, which measures the contextualized embedding-based similarity (Zhang et al., 2019b). We evaluate the models on the disfluent L2 data and the fluency rewrite data from the JFLEG dataset (Napoles et al., 2017). The reference Spanish sentences are from the JFLEG-es dataset, which is a manual translation of the L2 sentences with the goal to recover L2 errors (Anastasopoulos et al., 2019).

4.3 Experiments

The baseline of our experiment is an NMT model trained on the clean Europarl English-Spanish data without error augmentation (exp #0 in Table 3).

To evaluate the effect of the grammatical error augmentation, we combine the 4 subtypes of grammar errors (exp #1 in Table 3) according to the distribution learned in Section 3.2.1 with the

[3] www.focusfwd.com/10-most-translated-languages

[4] https://github.com/google/sentencepiece

Single-Error Mixture method, as it allows us to explore different error ratios without repeating runs of the error generation script. The final error ratio reported in the results section is 5:4:4:2 for Verb Inflection, Preposition Misuse, Article Misuse, and Noun Form errors, respectively, which closely resembles the percentage of lines containing each error type as shown in Table 2.

In Experiments #2 and #3, we study the effect of Typo and Run-on errors generated following Section 3.3.2 and 3.3.3, respectively. The paraphrase dataset is used in Experiment #4, which combines the Word Misuse Error and the Awkward Phrasing Error types in a Single-Error Mixture fashion, as they are both line-level errors and cannot be easily compounded.

In Experiments #5 through #8, the '&' operator denotes errors combined with In-line Compounding, and '+' denotes errors combined with Single-Error Mixture. Using both methods during the error augmentation process imitates the realistic L2 learning process of compounding mistakes in one sentence but also allows for the efficient reuse of generated errors and avoids overloading too many errors in one sentence. For all error augmentation configurations, we add the clean data to the synthetic disfluent data to control for the noise contained in the training set inspired by Ye et al. (2022). Lastly, in Experiment #9, we sample 2M sentence pairs from the error combination of Experiment #8 to match the data size used in the baseline Clean model and run a controlled study to evaluate the effect of training dataset size.

5 Results & Analysis

Table 3 shows the detokenized BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and BERTScore (Zhang et al., 2019b) of the model trained with different error augmentation methods. Overall, all models have better performance on the "easier" Fluent test set, while the Disfluent test set posts a harder challenge on the models.

#	Error Desc.	Size	Fluent			Disfluent			Δ
			BLEU	TER	BERTScore	BLEU	TER	BERTScore	BLEU
0	Clean	2.0M	**27.4**	60.8	**0.868**	25.4	62.5	0.858	2.0
1	Grammar (G)	5.0M	26.5	60.9*	0.866	25.7	61.8*	0.859	0.8
2	Typo (T)	3.9M	26.9*	62.3	0.867*	25.9*	62.4	0.861*	1.0
3	Run-on (R)	3.5M	26.8	63.2	0.865	25.7	64.1	0.858	1.1
4	Paraphrase (P)	3.0M	26.4	61.0	0.865	25.6	62.1	0.857	0.9
5	T & R	4.0M	26.7	62.8	0.866	26.2	62.2	0.862	0.5
6	P + G	6.0M	27.0	60.7	0.867	26.5	61.3	0.862	0.5
7	T & R + G	7.0M	27.2	60.6	**0.868**	27.0	**60.8**	**0.864**	**0.2**
8	T & R + G + P	9.0M	**27.4**	**60.4**	**0.868**	**27.2**	**60.8**	0.863	**0.2**
9	TRGP Control	2.0M	27.3	61.1	0.867	26.4	61.9	0.859	0.9

Table 3: NMT Performance on Fluent (manually corrected) and Disfluent (L2) sentences. Bold values mark best overall performance; '*' marks best results from single-error augmentation.

Baseline Clean Model The model trained on clean data (exp #1) achieves the best performance on the Fluent test set. This is because the sentences in the clean data are the most similar to the fluent manual corrections. When evaluated on the Disfluent L2 test set, however, the performance of the Clean model drops by 2 BLEU scores. This suggests that the L2 errors in disfluent data cause performance degradation when the model has not seen any noisy or out-of-domain data.

Single-Error Augmentation All models trained with one type of synthetic error outperform the Clean model on the Disfluent dataset, while generally performing not as well on the Fluent test set. Out of the four models trained with single error augmentation, the Typo model recovers the most from L2 noises, outperforming the Clean model by 0.5 BLEU. This behavior can be explained by the introduced typo words, generated from a realistic distribution described in Section 3.3.2, creating more tokens in the vocabulary and alleviating the performance degradation caused by OOV tokens. The poor performance on the Fluent set might be due to the single-source errors changing the data distribution drastically, causing the model to overfit to a one error type.

Error Diversity The models trained with a combination of several errors (exps #5–7) perform better than the other models trained on only one type of error. This suggests that diversifying the error type improves the robustness of the model. Note that although the Grammar model (exp #1) is trained with a combination of four types of grammar errors, the error subtypes are relatively simple. Thus, the combination of the four subtypes is not as diverse as, for example, Typo & Run-on Errors in exp #5, and definitely not as diverse as its superset: Typo & Run-on + Grammar Errors in exp #7. Lastly, we can see that although increasing the training data size will improve performance (2M in exp #9 vs. 9M in exp #8), it does not dictate the quality of the trained MT system, as exp #9 outperforms #0 by a large margin with the same amount of data.

Robustness to Disfluency The relative performance of each model on the Fluent and Disfluent test sets is also an informative measure of robustness. Ideally, a model robust to disfluent sentences should achieve the same performance on noisy, disfluent data as it does on clean, fluent data. As shown in Table 3, the lower the value of ΔBLEU, the smaller the performance drop with noisy data and thus the more robust the model. Single error models in experiments #1 through #4 show stronger robustness (ΔBLEU=0.95) than the Clean model (ΔBLEU=2.0). The combined error augmentation models are the most robust models with ΔBLEU=0.35.

Overall The combination of 'Typo & Run-on + Grammar + Paraphrase' errors (exp #8) not only outperforms all other models on the Disfluent dataset but also has comparable results to the Clean model evaluated on the Fluent dataset. It is able to recover most of the noise and degradation of the NMT system caused by L2 disfluency without sacrificing performance on regular fluent data. The model in exp #9 has the highest diversity and replicates the gold-standard L2 error distribution obtained in Section 3.2. Although other error combinations improve robustness, they do not contain full coverage of error types and deviate from the error distributions, thus resulting in less optimal performance. Therefore, accurately representing L2 errors contributes to the development of high-quality synthetic datasets, suggesting the potential for cognitively-motivated studies of human-generated corpora to better understand the process of L2 error formation.

6 Conclusion

In conclusion, our study shows that by specifically targeting the challenges faced by second-language learners of English, we can improve the robustness of neural machine translation models to disfluent data. We first created a realistic L2 error distribution and then produced synthetic data using the learned distributions to resemble real L2 errors. Our method of creating artificial errors similar to those made by L2 learners proved to be effective in improving the quality of the machine translation system, even without gold-labeled training data. This approach can be extended to other language pairs and used to improve the performance of machine translation systems for other language learners as well. Overall, this work opens up exciting avenues for future research in combining cognitive science theories to improve the robustness of NLP systems to disfluent L2 data.

7 Limitations

While our study shows promising results in improving the robustness of neural machine translation models to disfluent data, there are several limitations that should be acknowledged. Firstly, our method of creating artificial errors is based on a limited set of patterns observed in L2 data. It is possible that there are other patterns of disfluencies in L2 data that our method does not capture. This motivates an extensive study on written second language acquisition, which is out of scope for the current project but would be of great value to both the research community and potential users. Secondly, our study only focuses on the English→Spanish language pair. Although we are currently creating an L2 dataset in other typologically diverse languages, it is unclear how well our approach would generalize to other L2 and target languages.

References

Abdul Rauf, S., Rosales Núñez, J. C., Pham, M. Q., and Yvon, F. (2020). LIMSI @ WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 803–812, Online. Association for Computational Linguistics.

Alemi, M. (2016). General impacts of integrating advanced and modern technologies on teaching english as a foreign language. *International Journal on Integrating Technology in Education*, 5(1):13–26.

Anastasopoulos, A., Lui, A., Nguyen, T. Q., and Chiang, D. (2019). Neural machine translation of text from non-native speakers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3070–3080, Minneapolis, Minnesota. Association for Computational Linguistics.

Artetxe, M., Labaka, G., Agirre, E., and Cho, K. (2018). Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Belinkov, Y. and Bisk, Y. (2018). Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Benson, B. (1986). The markedness differential hypothesis: Implications for vietnamese speakers of english. *Markedness*, pages 271–289.

Berard, A., Calapodescu, I., and Roux, C. (2019). Naver labs europe's systems for the wmt19 machine translation robustness task. *arXiv preprint arXiv:1907.06488*.

Berry, N. (2012). Sloppy typing, fat fingers and atomic typos.

Chauhan, S., Saxena, S., and Daniel, P. (2022). Improved unsupervised neural machine translation with semantically weighted back translation for morphologically rich and low resource languages. *Neural Processing Letters*, 54(3):1707–1726.

Chen, H., Huang, S., Chiang, D., and Chen, J. (2017). Improved neural machine translation with a syntax-aware encoder and decoder. *arXiv preprint arXiv:1707.05436*.

Cheng, Y. and Cheng, Y. (2019). Semi-supervised learning for neural machine translation. *Joint training for neural machine translation*, pages 25–40.

Cheng, Y., Jiang, L., and Macherey, W. (2019). Robust neural machine translation with doubly adversarial inputs. *arXiv preprint arXiv:1906.02443*.

Cheng, Y., Tu, Z., Meng, F., Zhai, J., and Liu, Y. (2018). Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia. Association for Computational Linguistics.

Dahlmeier, D., Ng, H. T., and Wu, S. M. (2013). Building a large annotated corpus of learner english: The nus corpus of learner english. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 22–31.

Damodaran, P. (2021). Parrot: Paraphrase generation for nlu.

Eckman, F. R. (1977). Markedness and the contrastive analysis hypothesis. *Language learning*, 27(2):315–330.

Hashemi, H. B. and Hwa, R. (2016). An evaluation of parser robustness for ungrammatical sentences. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1765–1774.

Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., and Tetreault, J. (2014). Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180.

Hoang, V. C. D., Koehn, P., Haffari, G., and Cohn, T. (2018). Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Hsu, C.-Y., Chen, P.-Y., Lu, S., Liu, S., and Yu, C.-M. (2022). Adversarial examples can be effective data augmentation for unsupervised machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6926–6934.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. iclr. 2015. *arXiv preprint arXiv:1412.6980*, 9.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Koehn, P. (2020). *Neural machine translation*. Cambridge University Press.

Krashen, S. (1981). Second language acquisition. *Second Language Learning*, 3(7):19–39.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Li, X., Michel, P., Anastasopoulos, A., Belinkov, Y., Durrani, N., Firat, O., Koehn, P., Neubig, G., Pino, J., and Sajjad, H. (2019). Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102, Florence, Italy. Association for Computational Linguistics.

Liu, Y., Wan, Y., Zhang, J.-G., Zhao, W., and Yu, P. S. (2021). Enriching non-autoregressive transformer with syntactic and semanticstructures for neural machine translation. *arXiv preprint arXiv:2101.08942*.

Napoles, C., Sakaguchi, K., and Tetreault, J. (2017). JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., and Bryant, C. (2014). The conll-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440.

Post, M. (2018). A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., and Madnani, N. (2018). Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Specia, L., Li, Z., Pino, J., Chaudhary, V., Guzmán, F., Neubig, G., Durrani, N., Belinkov, Y., Koehn, P., Sajjad, H., Michel, P., and Li, X. (2020). Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

Stahlberg, F. (2020). Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Ye, R., Wang, M., and Li, L. (2022). Cross-modal contrastive learning for speech translation. *arXiv preprint arXiv:2205.02444*.

Yuan, X., Pham, D., Davidson, S., and Yu, Z. (2022). ErAConD: Error annotated conversational dialog dataset for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 76–84, Seattle, United States. Association for Computational Linguistics.

Zhang, M., Li, Z., Fu, G., and Zhang, M. (2019a). Syntax-enhanced neural machine translation with syntax-aware word representations. *arXiv preprint arXiv:1905.02878*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019b). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.