Rare but Severe Neural Machine Translation Errors Induced by Minimal Deletion: An Empirical Study on Chinese and English

Ruikang Shi and **Alvin Grissom II** and **Duc Minh Trinh**
Haverford College
Haverford, PA, USA

Abstract

We examine the inducement of rare but severe errors in English-Chinese and Chinese-English in-domain neural machine translation by minimal deletion of the source text with character-based models. By deleting a single character, we can induce severe translation errors. We categorize these errors and compare the results of deleting single characters and single words. We also examine the effect of training data size on the number and types of pathological cases induced by these minimal perturbations, finding significant variation. We find that deleting a word hurts overall translation score more than deleting a character, but certain errors are more likely to occur when deleting characters, with language direction also influencing the effect.

1 Introduction

Pathological machine translation (MT) errors have been a problem since the field's inception, and they have been analyzed and categorized in the context of both statistical (SMT) and neural machine translation (NMT). Recent work examines pathologies in NLP models on classification problems: cases in which the models make wildly inaccurate predictions, often confidently, when input tokens are removed (Feng et al., 2018). Identifying these enriches our understanding of neural models and their points of failure. MT pathologies take the form of severe translation errors, the worst being hallucinations (Lee et al., 2019). These rare errors are difficult to study precisely because they are rare. In this paper, we examine severe errors induced by minimal deletions by automatically extracting translations with severe errors and manually categorizing them.

Previous work taxonomizes SMT errors (Vilar et al., 2006) and analyzes their effects on translation quality (Federico et al., 2014). More recently, Guerreiro et al. (2022) propose a taxonomy of MT pathologies, of which hallucinations are a category.[1] They note the shortcomings of current automatic detection methods, e.g., those based on quality estimation and heuristics, and look for critical errors in naturalistic settings. They also propose DEHALLUCINATOR, which flags problematic translations and replaces them with re-ranking.

Other work on Chinese-English (Zh-En) SMT examines tense errors caused by incorrectly translating 了 (*le*) (Liu et al., 2011) and syntactic failures caused by 的 (*de*). More recent work uses input perturbation to argue that NMT models, including those based on transformers (Vaswani et al., 2017), are brittle: Belinkov and Bisk (2018) examine the effect on NMT systems of several kinds of randomized perturbations by adding tokens, and Niu et al. (2020) study subword regularization to increase robustness to randomized perturbations. Raunak et al. (2021) argue that memorized training examples are more likely to hallucinate, and Voita et al. (2020) examine the contribution of source and target tokens to errors. Also related, Sun et al. (2020) suggest that BERT is less robust to misspellings than other kinds of noise, which can occur naturalistically or through other errors (e.g., encoding).

While we expect targeted adversarial examples—those explicitly designed to cause a system to fail (Jia and Liang, 2017; Ebrahimi et al., 2018)—to cause serious errors, we focus on the ostensibly more benign case of in-domain En↔Zh NMT with minimal deletions. Adding valid words introduces distractors with which the MT system must cope, while deleting words more often *removes* information without explicitly introducing lexical distractors. Both are noise, but the latter is more naturally framed as requiring recovery from missing information, while the former introduces irrelevant and misleading information. At the character level, this distinction is less clear, since both adding and re-

[1] Guerreiro et al. (2022) note that the term "hallucination" is overloaded and inconsistent; for this reason, we generally avoid the use of this term here.

Proceedings of the 29th International Conference on Computational Linguistics, pages 5175–5180
October 12–17, 2022.

moving characters requires that the model translate despite unseen input substrings—minimally corrupted inputs. Are minimal word or character corruptions more harmful to a purely character-based NMT model? The answer is not obvious.

While most prior work examines western European languages, we examine translation between Chinese and English, building upon work identifying errors by observing change in BLEU (Papineni et al., 2002) after perturbation (Lee et al., 2019). But in contrast this prior work, which adds tokens, we focus exclusively on single deletions to examine **minimal conditions**—i.e., a missing character or word, as in a typo or corruption—under which **severe errors** are newly induced. For our purposes, a severe error leads to a translation in which the original meaning is unrecoverable, but there are others, as well (Vilar et al., 2006). For our purposes, we use WORD CHANGING to cover these cases.

2 Finding Candidates

We now describe the training of our NMT model, method for extracting severe error candidates (**enumerations**), and the results of this extraction. For our extraction experiments, we begin by examining character deletion before repeating the same experiments with word deletion. All experiments are done in both directions and for two different training data sizes (1M and 10M sentences), allowing us to observe the effect of training data size, translation direction, and deletion type.

2.1 Data and Models

We train character-based En↔Zh models on the UN Parallel Corpus 1.0 (Ziemski et al., 2016) of sentence-aligned UN parliamentary documents.

We train two models in each direction with Sockeye 2 (Hieber et al., 2020)—the first on the first 1M sentences and the second on 10M—to observe the effect of training data size on severe errors. We use the final 8,041 sentences as validation and test data; the first 2,000 are test data.[2]

[2]We use a six-layer transformer with eight attention heads and a feed-forward network of 2,048 hidden units, trained on one 16GB Quadro P5000. Batch size is 256 and learning rate is .0002, reduced by a factor of .9 after 8 unimproving checkpoints. Training ceases when validation perplexity quiesces for 20 checkpoints of 4,000 updates. While BPE has been shown to have higher BLEU on several datasets, this is not always the case (Cherry et al., 2018), and it can sometimes cause anomalies in translation itself (Ataman et al., 2017; Huck et al., 2017). We want to analyze the effect of deletion under simple conditions without this added complexity.

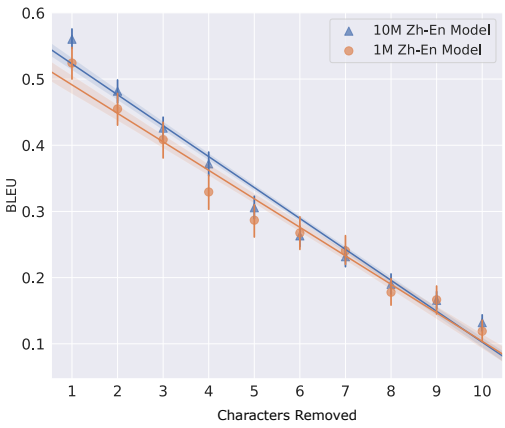

Figure 1: Zh-En BLEU as function of characters removed on valid sentences with 95% confidence intervals. There is a linear relationship, with average BLEU converging as more tokens are removed.

2.2 Identifying Error Candidates

On translated test sentences, if sentence-level BLEU is above 0.5, the translation is considered **valid**.[3] We translate valid sentences with one token missing, exhaustively trying every possible deleted token in every sentence. Perturbed sentences' translations are called **enumerations**. If an enumeration's sentence-level BLEU is less than .1, it is a candidate error, as these precipitous drops are outliers in the linear decline in BLEU as tokens are removed (Figure 1). For a more detailed specification of this process, see Algorithm 1, which is written for clarity rather than efficiency.[4]

3 Experiments and Results

We now discuss our experiments and the results of our enumeration extraction and the errors contained therein. All results are summarized in Table 2, with results on the same 2,000 test sentences.

3.1 Error Categorization

We manually categorize errors into four types in our analysis: WORD CHANGING, INABILITY, MISSING PARTS, and IRRELEVANT. Examples and descriptions are in Table 1.[5] What they have in common is

[3]We choose this because it is well above average BLEU for all models, resulting in an enumeration set with high average BLEU (Table 2), and few perturbed sentences reach this score (Figure 1).

[4]Code is hosted on GitHub.

[5]MISSING PARTS differs from INABILITY in that what *is* translated for MISSING PARTS is correct. IRRELEVANT translations are readable but unrelated to the source, while

Error Type	Example	Description
WORD CHANGING	**Source:** Occupational health and occupational risks. **Perturbed Source:** Occupational heath and occupational risks **Reference:** 职业 健康 与 职业 风险 zhíyè jiànkāng yǔ zhíyè fēngxiǎn occupational health and occupational risks **Translation:** 职业 道德 和 职业 危险 zhíyè dàodé hé zhíyè wéixiǎn occupational ethics with occupational dangers	The model only mistranslates the perturbed word, leading to a simple error in which *health* has been swapped with the unrelated word *ethics* (which is also orthographically distant in the source text).
INABILITY	**Source :** Christian Peace Action Groups. **Perturbed Source:** Christian PeaceAction Groups. **Reference:** 基督教 和平 行动 组织 jīdūjiào hépíng xíngdòng zǔzhī Christian Peace Action Groups **Translation:** Christian Peaction Groups	Instead of outputting Chinese, the model copies English characters, including the nonsense word *Peaction*.
MISSING PARTS	**Source:** Residential institutions: services for children. **Perturbed Source:** esidential institutions: services for children. **Reference:** 寄宿 机构 : 为 儿童 提供 服务 jìsù jīgòu : wèi értóng tígōng fúwù Residential institutions : for children provide services **Translation:** 对 儿童 的 服务 duì er tóng de fúwù for children 'de' services	Only some of the text is translated. In this example, though the translation is interpretable, a substantial portion of the text is entirely untranslated.
IRRELEVANT	**Source:** Maternal breastfeeding. **Perturbed Source:** aternal breastfeeding. **Reference:** 母乳 喂养 mǔrǔ wèiyǎng maternal breastfeeding **Translation:** 联合国 维持 和平 行动 经费 的 筹措 liánhéguó wéichí hépíng xíngdòng jīngfèi de chóucuò UN keep peace operation funding 'de' raise	This output is entirely hallucinated and has no apparent relationship to the input.

Table 1: Examples and descriptions of triggers and error types found in low-scoring enumerations.

that the original meaning is unrecoverable, though simple WORD CHANGING is not considered *a priori* severe in our analysis.

3.2 En-Zh 1M Training Sentence Results

There are 96 candidate severe errors among 14,722 enumerations: ten INABILITY, three IRRELEVANT and five MISSING PARTS. The rest are WORD CHANGING. We have 18 errors (.12%). [6]

One possible reason for these errors is that the model has insufficient training data to generalize. We investigate by training on ten times the data.

3.3 En-Zh Model Trained on 10M Sentences

We use the same corpus and architecture but use the first 10M instead of 1M parallel sentences to train (En-Zh-10M). Validation perplexity is nearly halved to 6.0 vs. the 1M model's 11.5 Likewise, BLEU on the test data increases by .08 to .4 (Table 2), as expected. Unexpectedly, BLEU on enumerations drops by .16 with more training data, much more than the .11 drop with 1M training sentences, suggesting more training data counterintuitively *increases* sensitivity to minimal character deletions, despite initial BLEU being higher.

There are 119 candidates among the 30,079 enumerations: 33 INABILITY and no MISSING PARTS or IRRELEVANT, giving a 0.11% probability of severe errors, approximately the same as the 1M model (0.12%).

The distribution of error types differs considerably when training on more data: INABILITY errors triple. We find that this is due to untranslated

INABILITY indicates a failure to generate readable output. It is possible in principle to have many types of errors in one bad translation, but we did not observe this.

[6] We also try removing sentences with English characters on the Chinese side, leaving 831,941 sentences on which to train. Translating these yields no INABILITY errors and leaves BLEU largely unchanged, suggesting that the untranslated named entities in the training data indeed cause INABILITY. There are three MISSING PARTS and two IRRELEVANT out of 63 potential hallucinations. Test BLEU is largely unchanged, and valid BLEU decreases only slightly.

Model	BLEU	Deletion	Valid	BLEU (Valid)	Enum.	BLEU (Enum.)	ΔBLEU	In.	MP	Irr.	Total Errors
En-Zh-1M	.32	Char	351	.77	14,722	.66	-.11 (-14.2%)	10	5	3	18 (0.12%)
En-Zh-10M	.40	Char	506	.80	30,079	.64	-.16 (-20.0%)	33	0	0	33 (0.11%)
Zh-En-1M	.39	Char	602	.73	11,093	.62	-.11 (-15.0%)	0	5	1	6 (0.05%)
Zh-En-10M	.42	Char	714	.78	14,031	.67	-.11 (-14.1%)	0	1	0	1 (0.007%)
En-Zh-1M	.32	Word	351	.77	2,521	.48	-.29 (-37.6%)	3	0	5	8 (0.32%)
En-Zh-10M	.40	Word	506	.80	4,945	.54	-.26 (-32.5%)	7	0	2	9 (0.18%)
Zh-En-1M	.39	Word	602	.74	6,666	.54	-.20 (-27.0%)	0	2	6	8 (0.12%)
Zh-En-10M	.42	Word	724	.78	8,461	.58	-.20 (-25.6%)	0	1	9	10 (0.11%)

Table 2: Results of candidate extraction for minimal deletion, BLEU for each extracted set of sentences, and error statistics in models, broken down into INABILITY (**In.**), MISSING PARTS (**MP**), and IRRELEVANT (**Irr.**). **Valid** sentences with BLEU > 0.5 are extracted to create minimally perturbed **enumerations**; from these candidates, bad translations are extracted based on BLEU decline post-perturbation (ΔBLEU). Despite character deletion introducing nonsense words into the input, word removal causes more of these severe errors. Surprisingly, despite Chinese characters containing more information, English deletion causes substantially higher decline in BLEU.

words in the training data, all of which are named entities.[7] Since more training data contains more untranslated named entities, INABILITY is more likely in models trained on more data. We therefore train a model on the data where no English appears in the references.

3.4 Zh-En Experiments

We examine Zh-En MT under the same character deletion conditions as En-Zh. Since Chinese characters contain more information than English letters, we expect greater sensitivity to deletions on Zh-En, but we do not find this (Table 2). Perturbing En-Zh leads to consistently steeper declines in BLEU, as seen in the valid vs. enumeration scores.

On the Zh-En model trained on 1M sentences, BLEU drops by .11, from .73 for the 602 sentences to .62 for the enumerations, whereas when trained on 10M sentences, we have .67 BLEU on enumerations, which is higher than that of the smaller model. This is, notably, the opposite of the En-Zh results, where more data decreased enumeration BLEU. Both Zh-En experiments decrease by .11 BLEU on enumerations, suggesting that the model with more training data is similarly robust to this perturbation as the smaller model, unlike the En-Zh case, in which the model trained on more data is more sensitive to character perturbations. As before, training models with more data decreases Zh-En errors: on Zh-En model trained on 1M sentences, we have 1 IRRELEVANT and 5 MISSING PARTS (.05%) errors, while on when trained on10M sentences, we have 1 MISSING PARTS (.007%). The remaining errors are WORD CHANGING.

There are no INABILITY errors in the two Zh-En experiments, which accords with the results from En-Zh, suggesting that INABILITY is due to the untranslated words in the training data. Since there are no untranslated Chinese words on the English side in the training data, we expect no INABILITY for a Zh-En model.

3.5 Minimal Word Deletion

We now examine *word* deletion as a basis of comparison. Does the character NMT model better handle the corrupted words caused by minimal character deletion, or is it more robust to whole word deletion, which leaves coherent words but removes more characters?[8] We find that, in all cases, deleting words leads to *substantially* lower BLEU than deleting characters, and though still rare, confirmed severe error rates also increase.

For En-Zh trained on 1M sentences, for instance, BLEU for enumerations drops to 0.48 in comparison to 0.66 when deleting characters, and these stark differences in BLEU persist.

On En-Zh trained on 1M sentences, we have 3 INABILITY and 5 IRRELEVANT (.32% severe errors). As expected, error rate increases considerably vs. character removal (.12%).

On En-Zh trained on 10M sentences, we have 7 INABILITY and 2 IRRELEVANT. .18% of 4,945 enumerations are severe errors, also more likely than with character deletion.

[7] By convention, sometimes named entities from English are not translated into Chinese. Ugawa et al. (2018) attempted to improve NMT with named entity tags to better handle compound and ambiguous words, and other previous work showed that contamination by another language (Khayrallah and Koehn, 2018) and copies of source sentences in the target training data can degrade NMT performance.

[8] We use THULAC (Sun et al., 2016) `fast` for Chinese tokenization.

Algorithm 1 Algorithm for Finding Candidates. We describe the logic in three distinct steps for clarity, though there is obvious potential for optimization.

```
function RUN(test_sents)
    valid ← find_valid(test_sents)
            ▷ Find valid sentences with BLEU ≥ 0.5.
    generate_enumerations(valid)
            ▷ Iterate over every character in every
              valid sentence and try removing one
              character at a time. These perturbed sen-
              tences are enumerations.
    candidates ← find_candidates(valid)
            ▷ Keep the enumerations that have
              sentence-level BLEU ≤ 0.1. We then
              manually identify those with actual se-
              vere errors and categorize them.
end function

function FIND_VALID(test_sentences)
    for each sentence s in test_sentences do
        s.bleu ← bleu(s)
        if s.bleu ≥ 0.5 then
            valid.add(s)
        end if
    end for
    return valid
end function

function GENERATE_ENUMERATIONS(valid)
    for each valid sentence s in valid do
        for each char index i in s do
            enum ← s.delete_char(i)
            s.enums.add(enum)
        end for
    end for
end function

function FIND_CANDIDATES(valid)
    for each sentence s in valid do
        for each enum in s.enums do
            new_bleu ← bleu(enum)
            if new_bleu ≤ 0.1 then
                candidates.add(enum)
            end if
        end for
    end for
    return candidates
end function
```

As with character deletion, increasing training size increases INABILITY errors but decreases overall error probability. There are no MISSING PARTS errors when deleting words on En-Zh.

3.6 Summary

We see substantial variation in errors, depending on the kind of deletion and translation direction, with INABILITY occurring exclusively on En-Zh. We expect more BLEU decline on Zh-En, since Chinese characters contain more semantic content and source sentences are shorter, but we find the opposite of this with word deletion. We also find that while the models are more sensitive to word deletion in terms of overall BLEU, this does not lead to drastic increases in severe errors, suggesting that these severe errors are unrelated to typical MT errors, in line with arguments that hallucinations should be considered separately from the typical MT errors (Guerreiro et al., 2022), due to the unique patterns of heuristic-based methods when attempting to detect them.

4 Conclusion and Future Work

We examine the effect of minimal deletions on rare but severe MT errors on Chinese and English, using outlier changes in BLEU after deletion to find candidates.

We find that the error rate for the model with a larger dataset is always lower, suggesting more data can improve models' performance against severe errors. Removing single words is more likely to cause severe errors but less likely to cause MISSING PARTS in our models, despite character deletion introducing invalid words. On En→Zh, we observe none when removing words. With the important caveat that these errors are already rare, limiting the conclusions we can make, this may suggest that Zh↔En models are better able to recover when characters are missing, even if the substrings themselves have never been observed, despite not having been trained with such noise. This is not obvious for a character-based model. Nor is it obvious that Zh→En models will be more robust to perturbations than En→Zh, but this is what we find, especially for words, perhaps because English words are simply longer. Furthermore, that ΔBLEU is not predictive of significantly more severe errors suggests that these errors are a different phenomenon from typical MT shortcomings.

Further research is needed to determine the effect various variables on robustness with targeted probes; future work can also determine how findings generalize across more language pairs (potentially typologies), tokenization schemes, and architectures. Training models with missing source *words* may increase robustness. For detection, unusually large disparities in length between source and target could signal INABILITY or MISSING PARTS errors, and vocabulary or semantic distance checks could flag bad translations (e.g., WORD CHANGING, INABILITY). It would also be instructive to examine the extent to which NMT robustness to noise mirrors that of humans.

Acknowledgements

We thank John Dougherty and Sorelle Friedler for their helpful feedback.

References

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331.

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

Colin Cherry, George F. Foster, Ankur Bapna, Orhan Firat, and Wolfgang Macherey. 2018. Revisiting character-based neural machine translation with capacity and compression. In *Proceedings of EMNLP*.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of COLING*, pages 653–663.

Marcello Federico, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2014. Assessing the impact of translation errors on machine translation quality with mixed-effects models. In *Proceedings of EMNLP*, pages 1643–1653.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of EMNLP*, pages 3719–3728.

Nuno M Guerreiro, Elena Voita, and André FT Martins. 2022. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. *arXiv preprint arXiv:2208.05309*.

Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. 2020. Sockeye 2: A toolkit for neural machine translation. In *Proceedings of EACL*, pages 457–458, Lisboa, Portugal.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of EMNLP*, pages 2021–2031.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2019. Hallucinations in neural machine translation. In *Interpretability and Robustness for Audio, Speech and Language Workshop*. Proceedings of NeurIPS.

Feifan Liu, Fei Liu, and Yang Liu. 2011. Learning from Chinese-English parallel data for Chinese tense prediction. In *Proceedings of IJCNLP*, pages 1116–1124.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of ACL*, pages 8538–8544.

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

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. *Proceedings of NAACL*.

Lichao Sun, Kazuma Hashimoto, Wenpeng Yin, Akari Asai, Jia Li, Philip Yu, and Caiming Xiong. 2020. Adv-BERT: BERT is not robust on misspellings! generating nature adversarial samples on BERT. *arXiv preprint arXiv:2003.04985*.

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. THULAC: An efficient lexical analyzer for Chinese.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of COLING*, pages 3240–3250.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, page 6000–6010.

David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *Proceedings of LREC*, Genoa, Italy.

Elena Voita, Rico Sennrich, and Ivan Titov. 2020. Analyzing the source and target contributions to predictions in neural machine translation. *arXiv preprint arXiv:2010.10907*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of LREC*, pages 3530–3534.