Rare but Severe Errors Induced by Minimal Deletions in English-Chinese Neural Machine Translation

Anonymous ACL submission

Abstract

We examine the inducement of rare but severe errors in English-Chinese and Chinese-English Transformer-based neural machine translation by minimal deletion of the source. We also examine the effect of training data size on the number and types of pathological cases induced by these perturbations, finding significant variation. We find that one type of hallucination can be remedied through data preprocessing.

1 Introduction

Pathological machine translation 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 inputs tokens are removed (Feng et al., 2018). Identifying these enriches our understanding of neural models and their points of failure. MT pathologies can take the form of severe translation errors, the worst being **hallucinations** (Lee et al., 2019), uninterpretable or irrelevant "translations". These rare errors are difficult to study precisely because they are rare.

Previous work taxonomizes SMT errors (Vilar et al., 2006) and analyzes their effects on translation quality (Federico et al., 2014). 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 Transformers (Vaswani et al., 2017), are brittle: Belinkov and Bisk (2018) examine the effect on NMT systems of several kinds of randomized perturbations (adding tokens), and Niu et al. (2020) study subword regularization to increase robustness to randomized perturbations. Raunak et al. (2021) argue memorized training examples are more likely to hallucinate. Sun et al. (2020) suggest BERT is less robust to misspellings than other kinds of noise, which can occur naturalistically or through other kinds of errors (e.g., encoding).

While it is intuitive to expect targeted adversarial examples (Jia and Liang, 2017; Ebrahimi et al., 2018) to cause serious errors, we focus on 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 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 removing characters requires the model to translate despite unseen input substrings—minimally corrupted inputs. Are minimal word or character corruptions more harmful to a character NMT model? The answer is not obvious and may vary between models.

Most prior work examines western languages; we focus on En↔Zh, building upon work identifying errors by observing change in BLEU (Papineni et al., 2002) (Δ BLEU) after perturbation (Lee et al., 2019). But in contrast this 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 such errors are newly induced.

2 Finding Candidates

We now describe the training of our NMT model, method for extracting hallucination candidates (**enumerations**), and the results of this extraction.

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 zhiye jiankang yu zhiye fengxian occupational health and occupational risks **Translation:** 职业 道德 和 职业 危险 zhiye daode he zhiye weixian occupational ethics with occupational dangers	Translation only mistranslates perturbed word. A simple error in which *health* has been swapped with the unrelated word *ethics*. This error is not a hallucination because it is interpretable.
INABILITY	**Source :** Christian Peace Action Groups. **Perturbed Source:** Christian PeaceAction Groups. **Reference :** 基督教 和平 行动 组织 jidujiao heping xingdong zuzhi Christian Peace Action Groups **Translation:** Christian Peaction Groups	Instead of outputting the correct Chinese translation, the model hallucinates English, including the nonsense word *Peaction* in thi example. This is a hallucination because it is unreadable.
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 jisu jigou : wei ertong tigong fuwu Residential institutions : for children provide services **Translation:** 对 儿童 的 服务 dui ertong de fuwu for children 'de' services	Only some parts of the sentence arae translated. While the resultant translation is interpretable, a substantial portion of the text is entirely untranslated. It is a hallucination because it is unreadable.
IRRELEVANT	**Source:** Maternal breastfeeding. **Perturbed Source:** aternal breastfeeding. **Reference:** 母 乳喂养 mu ruweiyang maternal breastfeeding **Translation:** 联合国 维持 和平 行动 经费 的 筹措 lianheguo weichi heping xingdong jingfei de cuochou UN keep peace operation funding 'de' raise	This output is entirely hallucinated and has no apparent relationship to the input, making it a catastrophic error.

Table 1: Error types and minimal triggers found in our analysis of low-scoring enumerations. IRRELEVANT and INABILITY are hallucinations.

2.1 Data and Models

We train character-based En↔Zh models on the UN Parallel Corpus 1.0 (Ziemski et al., 2016), consisting of sentence-aligned UN parliamentary documents and records from 1990 to 2004.

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

[1] We use a 6-layer Transformer with 8 attention heads and a feed-forward network of 2,048 hidden units, trained on one Nvidia Quadro P5000. Batch size is 256 and learning rate is .0002. Learning rate is reduced by a factor of .9 after 8 unimproving checkpoints. Training stops when validation perplexity quiesces for 20 checkpoints of 4,000 updates.

2.2 Identifying Error Candidates

On translated test sentences, if sentence-level BLEU is above 0.5, the translation is considered **valid**. We translate valid sentences with one character missing (for each character in the sentence). These perturbed sentences' translations are called **enumerations**. If an enumeration's sentence-level BLEU is less than 0.1, it is a candidate hallucination, as these precipitous drops are outliers in the linear decline in BLEU as tokens are removed (Figure 1).

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.

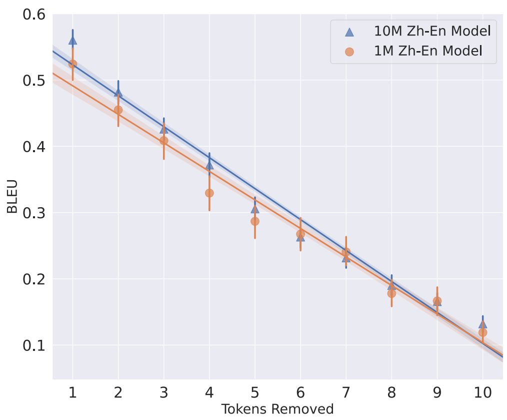

Figure 1: Zh-En BLEU as function of characters removed on valid sentences with 95% confidence intervals. There is a linear relationship, with average BLEU converging as more tokens are removed. The same is true on hallucinations (not shown).

3.1 Error Categorization

We manually categorize errors into four types in our analysis: WORD CHANGING, INABILITY, MISSING PARTS, and IRRELEVANT. Examples and descriptions of these are in Table 1.

3.2 En-Zh 1M Training Sentence Results

There are 96 candidate hallucinations among the 14,722 enumerations: ten INABILITY, three IRRELEVANT and five MISSING PARTS. The rest are all WORD CHANGING, which are errors, not hallucinations. We have 18 true hallucinations enumerations (0.12%). A possible reason for these hallucinations is that the model has insufficient training data to generalize. We investigate by training with ten times more 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 0.4 (Table 2), as expected. Unexpectedly, BLEU on enumerations drops by 0.16 with more training data, much more than the 0.11 drop with 1M training sentences, suggesting training data counterintuitively *increases* sensitivity to minimal character deletions, despite initial BLEU being higher.

The distribution of hallucination types differs significantly when training on more data: INABILITY triples. We find that this is due to untranslated words in the training data, all of which are named entities.[2] 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 En-Zh Model without Untranslated Words

We filter the 1M sentences, removing sentences with English characters on the Chinese side, leaving 831,941 sentences on which to train. Translating these yields no INABILITY errors, suggesting that the untranslated named entities in the training data indeed cause INABILITY. Test BLEU is largely unchanged, and valid BLEU decreases only slightly.

3.5 Zh-En Experiments

We examine Zh-En MT under the same character deletion conditions as En-Zh.

On Zh-En-1M, BLEU drops by 0.11, from 0.73 for the 602 sentences to 0.62 for the enumerations, whereas on Zh-En-10M, we have 0.67 BLEU on enumerations, which is higher than that of Zh-En-1M. This is, notably, the opposite of the En-Zh results, where more data decreased enumeration BLEU. Both Zh-En experiments decrease by 0.11 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 hallucinations: On Zh-En-1M, there are 91 possible hallucinations, of which we have 1 IRRELEVANT and 5 MISSING PARTS. 0.05% are hallucinations.

On Zh-En-10M, there are 90 possible hallucinations, among which we have 1 MISSING PARTS. Only 0.007% of enumerations are hallucinations.

There are no INABILITY errors in the two Zh-En experiments, which accords with the results from En-Zh, which suggest INABILITY is due to the untranslated words in the test data. Since there are no untranslated Chinese words on the English

[2] By convention, sometimes named entities from English not translated into Chinese. Previous work (Ugawa et al., 2018) has attempted to improve NMT with named entity tags to better handle compound and ambiguous words.

Model	BLEU	Deletion	Valid	BLEU (Valid)	Enum.	BLEU (Enum.)	In.	MP	Irr.	Total Hall.
En-Zh-1M	0.32	Char	351	0.77	14,722	0.66	10	5	3	18 (0.12%)
En-Zh-10M	0.40	Char	506	0.80	30,079	0.64	33	0	0	33 (0.11%)
Zh-En-1M	0.39	Char	602	0.73	11,093	0.62	0	5	1	6 (0.05%)
Zh-En-10M	0.42	Char	714	0.78	14,031	0.67	0	1	0	1 (0.007%)
En-Zh-1M	0.32	Word	351	0.77	2,521	0.48	3	0	5	8 (0.32%)
En-Zh-10M	0.40	Word	506	0.80	4,945	0.54	7	0	2	9 (0.18%)
Zh-En-1M	0.39	Word	602	0.74	6,666	0.54	0	2	6	8 (0.12%)
Zh-En-10M	0.42	Word	724	0.78	8,461	0.58	0	1	9	10 (0.11%)

Table 2: Results of candidate extraction for minimal deletion, BLEU for each extracted set of sentences, and hallucination 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, hallucinations are extracted based on BLEU decline. Despite character deletion introducing nonsense words into the input, word removal causes more hallucinations.

side in the training data, we expect no INABILITY for a Zh-En model.

3.6 Minimal Word Deletion

We now examine *word* deletion as a basis of comparison. Does the character NMT model better handle corrupted words (minimal character deletion) or whole word deletion, which leaves coherent words but removes more characters?[3] We find that, in all cases, deleting words leads to *significantly* lower BLEU than deleting characters, and though still rare, confirmed hallucination rates also increase.

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

on En-Zh-1M, out of 97 hallucination candidates, we have 3 INABILITY and 5 IRRELEVANT (0.32% hallucination). Hallucination likelihood increases significantly versus character removal (0.12%).

On En-Zh-10M, out of 114 candidate hallucinations, we have 7 INABILITY and 2 IRRELEVANT. 0.18% of 4,945 enumerations are hallucinations, also more likely than with character deletion.

As with character deletion, increasing training size increases the number of INABILITY but decreases overall hallucination probability. There are no MISSING PARTS errors when deleting words, suggesting that MISSING PARTS is caused by invalid words induced by character but not word deletion.

3.7 Summary

In all, we see substantial variation in hallucination patterns depending on the kind of deletion and the direction of translation, with INABILITY occurring exclusively on En-Zh. 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 *hallucinations*.

4 Conclusion

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 untranslated English words are a source of hallucinations and removing all training examples with words from the source language in the target examples eliminates INABILITY.

Both minimal character and word deletions induce hallucinations. The hallucination rate for the model with a larger dataset is always lower, suggesting that more data can improve the models' performance against hallucination. Experiments suggest that removing single words is more likely to cause hallucinations but less likely to cause MISSING PARTS errors in our character-based models, despite character deletion introducing invalid words.

More generally, removing words has more of a deleterious effect on translations than removing single characters, despite the the latter introducing nonsense words, suggesting that our character-based models are better able to recover when fewer characters are missing, even if the substrings themselves have never been observed, despite not having been trained with such noise. Further researach is needed to determine the nature of this apparent robustness with more targeted probes.

[3] We use THULAC (Sun et al., 2016) for Chinese segmentation.

References

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

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.

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.

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

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, undefinedukasz 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.

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