Machine Translation Robustness to Natural Asemantic Variation

Jacob Bremerman, Xiang Ren, Jonathan May
University of Southern California
Information Sciences Institute
jbrem@usc.edu

Abstract

We introduce and formalize an under-studied linguistic phenomenon we call Natural Asemantic Variation (NAV) and investigate it in the context of Machine Translation (MT) robustness. Standard MT models are shown to be less robust to rarer, nuanced language forms, and current robustness techniques do not account for this kind of perturbation despite their prevalence in 'real world' data. Experiment results provide more insight into the nature of NAV and we demonstrate strategies to improve performance on NAV. We also show that NAV robustness can be transferred across languages and fine that synthetic perturbations can achieve some but not all of the benefits of human-generated NAV data.

1 Introduction

We have arguably reached a point in Machine Translation (MT) where models achieve satisfactory performance under "sufficient" conditions (enough computer, data, etc.). Because of this, several directions in current MT research involve improving performance in less ideal conditions. Such directions include topics like low-resource MT, transfer learning, and robustness.

In this paper, we focus on robustness. Often, robustness is framed as the ability of a model to perform well on noisy input data. In the "real world" humans are able to process information even when data is presented imperfectly while models often struggle with this. Even slight, imperceptible changes to input data can cause drastic changes in model behavior (Goodfellow et al., 2015). In this way, robustness can also be viewed from the perspective of human ability.

There exist several varieties and extents of perturbations on standard input that do not affect a human's ability to perform a task on that input. A properly robust model should emulate this indiscrimination. Some classes of perturbations to which humans exhibit robustness include speling errors or ttypos, slang, CaSinG, v15sua11y-51m1ll4r ch4r4ct3r5, and synonym ~~replacement~~ substitution. There is much prior research in MT robustness which addresses one or more of these classes of perturbation. However, we focus on a less-studied class we call Natural Asemantic Variation (NAV).

Whereas aforementioned classes tend towards perturbations that generate agrammatical, nonstandard, or unnatural sentences, NAV perturbations represent the extra-semantic linguistic properties of a language that allow for subtle changes in nuance while expressing the same core meaning. Specifically, NAV is determined in context of two languages whereby a NAV perturbation is a kind of paraphrase in a source language that would not affect its translation in the target language. As this topic is under-studied, it is also not well-defined. We offer an example of NAV perturbations in Japanese in Table 1 and will provide further definitions and examples throughout the paper to help solidify the reader's understanding.

From Table 1, we see how slight modifications of a sentence in one language can be 'large' enough to convey specific nuanced differences in that language yet 'small' enough to not warrant a change to its translation in another language. A common challenge in MT has to do with the fact that there are almost always several correct answers, and we can imagine 'translationally-equivalent' mappings between large sets of similar sentences in different languages. Often we must resort to imperfect training and evaluating using subsets of these imagined sets due to data limitations. However, leveraging the STAPLE dataset (Mayhew et al., 2020), we have a unique opportunity to investigate NAV phenomena more directly with access to several (often hundreds) of high-quality NAV perturbations.

In this paper, we contribute a formalization for a linguistically-rich class of perturbations and a framing of NAV examples which help concretize

Perturbation	Asemantic Variation
彼女は私に本を返した	informal verb conjugation
彼女は俺に本を返しました	masculine "me" pronoun
彼女は私に本を返してくれました	“favor” nuance

Table 1: Examples of NAV perturbations for a Japanese sentence, 彼女は私に本を返しました (she returned the book to me). These natural variations, which cause nuance differences in Japanese, cannot be translated succinctly into English. Therefore, all perturbations can be translated reasonably to the same English sentence. A robust MT model (or human) should be able to extract the same semantic meaning regardless of the variation provided.

the concept. We also contribute an evaluation setup to measure NAV robustness in the form of a repurposed test set and simple metrics. We perform experiments that demonstrate improvements in NAV robustness and analyses that further describe the behavior of NAV-robustification, laying groundwork for future research in this area. Additional analyses reveal capabilities for NAV-robustness language-transfer and comparisons to synthetic data augmentation strategies.

2 NAV Robustness

In robustness, it is difficult to determine improvements without considering both what types of input models are meant to be robust *to* and what robustness *looks like*. For example, for an image classification task, a researcher may define robustness as the ability to predict the same class for an image after it is perturbed by introducing a Gaussian blur. This can be easily measured by performing these perturbations and comparing accuracy. If results improve, the researcher can conclude that their model is ‘robust’. However, claiming general robustness would not be as precise and accurate as claiming “robustness to Gaussian blur perturbations when evaluated on accuracy.”

Here, we do not claim to investigate ‘robustness’ as a blanket statement but rather a form of robustness, based on specific perturbations and specific evaluation metrics.

2.1 NAV Perturbations

To formally define Natural Asemantic Variation, we consider a corpus, C in two languages $\mathcal{X}, \mathcal{Y}$ that contain c pairs of translationally-equivalent sets of sentences:

$$C = \{(X_i, Y_i)\} \;\; \forall i \in \{1, ..., c\};$$
$$X_i = \{x_i^1, ..., x_i^n\}; \;\; Y_i = \{y_i^1, ..., y_i^m\};$$
$$\forall j \in \{1, ..., n\}, \; \forall k \in \{1, ..., m\}, \; x_i^j \Leftrightarrow y_i^k,$$

where $\Leftrightarrow$ represents translational equivalence, meaning the two sentences can be reasonably translated to each other, such as all the Japanese examples in Table 1 with “*she returned the book to me*” in English.

All $x \in X_i$ can then be considered NAV$_{\mathcal{X},\mathcal{Y}}$[1] perturbations of each other, provided they are also naturally-occurring, grammatical sentences (as opposed to perturbations including spelling errors, case changes, etc.).[2]

In this work, we further limit the corpus:

$$\forall i \; |X_i| > 1, \; |Y_i| = 1,$$

meaning only one reference translation is provided (if $|X_i| = 1$ as well, then C has the form of most standard MT datasets). To ground these definitions, we offer specific examples to distinguish NAV perturbations from non-NAV perturbations.

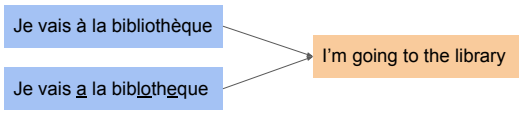

Figure 1: A non-NAV perturbation example. While the second French sentence does have the same translation as the first, it is not a naturally-occurring, grammatical sentence. Perturbations including orthographic errors as shown here are commonly considered in other MT robustness works but do not qualify as NAV.

Figures 1 and 2 show non-examples of NAV perturbation. The modifications are either nonstandard or have too large of an impact on semantics. Figure

[1]Note: a NAV perturbation must be considered in context of the language pair. x_i^1 and x_i^2 may be NAV$_{\mathcal{X},\mathcal{Y}}$ perturbations of each other but not NAV$_{\mathcal{X},\mathcal{Z}}$ perturbations

[2]Note: A paraphrase of a sentence cannot be assumed to be a NAV perturbation on its own because it may or may not cause a change to its translation in the relevant target language.

Figure 2: A non-NAV perturbation example. While the second French sentence is a naturally-occurring, grammatical sentence, the perturbation causes a significant semantic change that can be encoded succinctly in English, so these sentences would not be in the same X_i.

3 shows a simple example of a NAV perturbation. Figure 4 demonstrates the language dependence of NAV by using the same perturbation from Figure 3, highlighting the complexity of this problem. With a different target language, the two sentences would no longer exist in the same X_i.

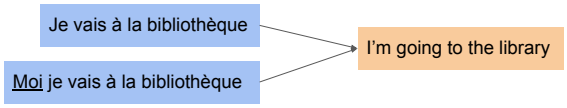

Figure 3: A NAV perturbation example. The second French sentence is both a natural sentence with additional nuance and maintains translational equivalence. The second French example places emphasis on the subject of the sentence, which is not easily expressed naturally in written English.

Figure 4: A non-NAV perturbation example. Spanish is able to succinctly and naturally express the same type of subject-emphasis nuance as in French. Thus, two sentences can only be evaluated as NAV perturbations of each other in the context of a specific target language.

2.2 NAV Robustness

In order to evaluate models for their robustness to NAV perturbations, we consider two desiderata, quality and consistency. In many tasks, these two properties are highly entangled, so robustness can be measured with standard performance metrics (quality implies consistency if there is only one correct prediction option). With MT, we can imagine how these properties could be separate. There are often several high-quality answer options in MT, so two models could both produce high-quality translations but differ greatly in their consistency. For our purposes, we posit that a more robust MT model should not only have increased or maintained translation quality but also consistency in output. Consistency in output would help ensure overall system robustness if we imagine the MT system being part of a larger NLP pipeline.

For translation quality, we use BLEU (Papineni et al., 2002) with SacreBLEU (Post, 2018). For translation consistency we define a measure, CONSIST, which rewards a model for exhibiting less variation in its outputs among perturbed inputs with equivalent semantics. For a pair $(X_i, \hat{Y}_i)$:

$$CONSIST_i = \frac{1}{|X_i|} \sum_{j=1}^{|\hat{Y}_i|} \frac{|\hat{y}_i^j|}{j},$$

where $\hat{Y}_i = [\hat{y}_i^1, ..., \hat{y}_i^m]$ sorted descending by $|\hat{y}_i^j|$ and $|\hat{y}_i^j|$ is the number of times the model outputs hypothesis $\hat{y}_i^j$ among the $|X_i|$ perturbations. We then take the macro average across all example pairs for the final system CONSIST metric.[3]

3 Analysis Setup

In this work, we seek to provide answers to several analysis questions related to NAV perturbations and MT models' robustness to them. Using our quality and consistency metrics we investigate the following:

- How can NAV perturbation data be used to improve an MT model's NAV robustness?
- Can a model's improved NAV robustness in one language pair be transferred to another language pair?
- How do synthetic perturbations compare with organic NAV data w.r.t. NAV robustness?
- What behavior changes does a 'NAV-robustified' MT model exhibit in other MT contexts?

3.1 Baseline Models

We work with two main classes of baseline models: mono-pair (can translate one specific language to one specific other language, e.g. ja-en) and multi-pair translation models (can translate between several languages, e.g. {hu,ja,pt}-{hu,ja,pt}). All models are transformer models (Vaswani et al., 2017) using the fairseq toolkit (Ott et al., 2019). Our principal models are multi-pair since many of

[3]This metric may break down in long-output tasks but works well for our domain of short, simple sentences.

our analysis questions relate to transfer, but we also perform preliminary experiments with mono-pair models for completeness.

We pre-train mono-pair models with the Tatoeba corpus for hu-en, ja-en and pt-en. This corpus is fairly close to the STAPLE domain and allows for reasonable baseline models. We obtain subwords and tokenize using SentencePiece (Kudo and Richardson, 2018).

For multi-pair model experiments, we use the M2M-100 model (Fan et al., 2020), which is trained on CC-Matrix (Schwenk et al., 2019) and CC-Aligned (El-Kishky et al., 2019). This model comes with a predetermined vocabulary and tokenizer.

3.2 NAV Data

This research is made possible largely by the unique dataset publicly released by Duolingo for the STAPLE shared task. For each English sentence, several (average 300) accepted translations of that sentence in one of five languages were sourced from Duolingo users and further annotated with frequency scores based on how often they used that specific translation. We use the enumerated valid translations as source-side NAV perturbations translating to the same target English sentence. Table 1 includes a real example from the training split for Japanese-English.

For NAV fine-tuning, we use different subsets of the STAPLE training corpus in three language pairs {hu,ja,pt}-en. For each "many-to-one" pair, we could use all of the many NAV perturbations or subselect. When subselecting, we consider three strategies. Since we also have frequency scores for each perturbation, we could select a number of the **most** frequent perturbations, **least** frequent perturbations, or uniformly sample for a **random** subset.

For NAV robustness evaluation, we use the full many-to-1 pairs of the STAPLE test corpus. On top of the {hu,ja,pt}-en sets, we also hold out {ko,vi}-en test sets to evaluate 0-shot language transfer using multilingual models.

3.3 Synthetic Perturbation

Obtaining NAV perturbations for other languages and domains is expensive, requiring human annotation by bilingual language experts to generate natural variations that don't affect translation equivalence. We consider applying synthetic perturbations to the STAPLE data to compare potential NAV robustness gains with a cheaper strategy.

We have the organic, human-generated variations in the STAPLE data, so we create a comparable synthetic dataset by taking the 1-most frequent sentence from each translation pair and perturbing it synthetically 9 times for a total of 10 variations.

For our roman script languages {hu,pt}-en, we use common noising techniques involving random casing changes and character substitutions, insertions and deletions as in (Niu et al., 2020). These perturbations do not resemble NAV perturbations, but they represent a common and easy way to add noisy data.

We also add synthetic NAV perturbations for the ja-en split. Roman script noising strategies are not perfectly translatable to ja-en. Also, ja-en arguably has the easiest rule-based modifications that can be programmed to automatically generate NAV perturbations. By no means are these methods exhaustive, but a combination of simple insertions of emphasis-related particles, substitutions of pronouns based on gender identity, and dropping unnecessary pronouns serve as a basic technique for synthetic NAV noising. Rules were implemented after scanning training data and observing patterns of NAV perturbation.

3.4 Additional Evaluations

We use other existing MT datasets to evaluate our models in other contexts such as performance on out-of-domain (OOD) data (7.1) and robustness-transfer to other types of noisy input (7.2). We use test splits from Tatoeba, OPUS-100 (Zhang et al., 2020) and MTNT (Michel and Neubig, 2018).

We also create additional evaluation sets from the STAPLE data to analyze in-domain model behavior. In addition to the 'all perturbations' test sets, we also filter out '1-most' and '1-least' test sets. These provide a more standard MT evaluation framework that relies on BLEU to measure in-domain translation performance on common ('1-most') and robustness to uncommon, NAV-perturbed ('1-least') sentence pairs.

4 NAV Robustness Experiments

For our first experiments, we fine-tune our mono-pair transformer models pre-trained on Tatoeba using strategies for subselecting STAPLE data discussed in Section 3.2 and evaluate using metrics from Section 2.2. After initially experimenting with several combinations of number of perturba-

tions per set pair and selection strategy,[4] we report on 4 conditions, representative of "real-world" MT scenarios:

- baseline: off-the-shelf MT model, no fine-tuning
- + 1-most: simple domain adaptation using parallel text in target domain (one "typical" translation per sentence).
- + 10-random: preferred "NAV robustness" data condition, having a small yet diverse set of NAV perturbations per translation pair
- + all: "throw in all the data" approach

Our results from these experiments are shown in Figure 5. The main purpose of these preliminary experiments is to provide evidence that NAV robustness improvement is possible on smaller scale models. We test various configurations to justify design choices for our future, more comprehensive experiments with large multi-pair models. We find that the '+ 10-random' condition results in the best trade-off of BLEU and CONSIST.

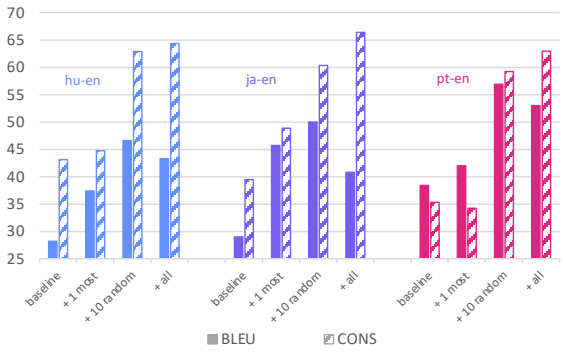

Figure 5: Results from mono-pair model experiments. The '+ 1 most' condition shows improvements over a baseline in a typical MT domain adaptation scenario where some target domain data is available. The '+ 10 random' condition shows how exposing the model to NAV perturbations further increases both BLEU and CONSIST scores beyond domain adaptation with frequent input forms. While including all NAV perturbations ('+ all') continues to improve CONSIST, we see a decrease in BLEU.

We repeat these experiments using a large multi-pair model. We find similar patterns in the results, which we plot in Figure 6. Here, the baseline is the M2M-100 model and models are fine-tuned and evaluated on only the same language pair, e.g. BLEU-ja + 1 most uses the baseline m2m model and fine-tunes on STAPLE ja-en 1-most pairs. [5]

[4] full results in appendix

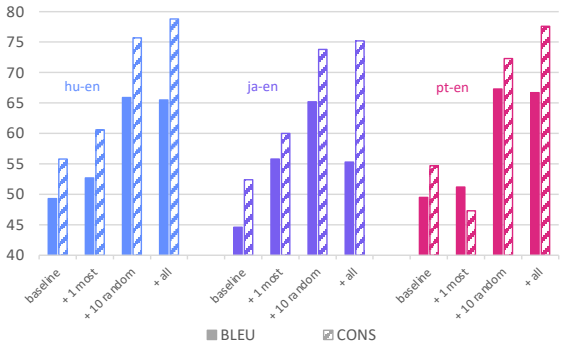

Figure 6: As in the mono-pair translation experiments, we see how '+ 1 most' generally shows improvements over a baseline and '+ 10 random' shows how exposing the model to NAV perturbations further increases both BLEU and CONSIST scores. While including all NAV perturbations continues to improve CONSIST, we see a trade-off in BLEU.

5 Transfer Experiments

With a large multi-pair model, we can consider zero-shot language-transfer. First, we take the M2M-100 baseline models and fine-tune them on one of the STAPLE fine-tuning language-pairs. We then evaluate on our held-out language test sets (unseen pairs during fine-tuning, ko,vi-en). Our results are shown in Figure 7. From previous experiments, we see sufficient evidence that " + 10 random " is appropriate for NAV robustifying, so we mainly report on this setting.

Results suggest zero-shot transfer of NAV robustness is possible, with 10-random NAV perturbations per pair showing larger robustness improvement than simply fine-tuning on 1-most. We also observe language differences. ko-en BLEU improves more from ja-en fine-tuning while vi-en improves more from hu-en.

Finally, we perform multilingual fine-tuning experiments in which we combine all three training sets together {hu,ja,pt}-en with the '+ 10 random' strategy and evaluate on all 5 test sets. We compare the multilingual fine-tuning results to the previous best results (according to BLEU) from our monolingual fine-tuning experiments. Results are shown

[5] To help validate our results, we run 3 different seeds for the randomization of data in 'random-10.' Standard deviation for BLEU scores is between 0.4 and 0.6. Standard deviation for CONSIST scores is between 0.7 and 1. The variation is small enough to not affect conclusions.

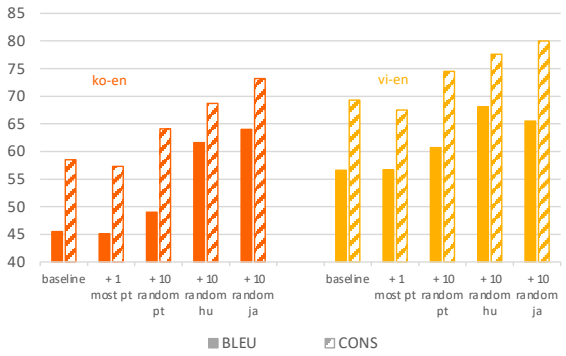

Figure 7: As in mono-pair experiments, including 10 NAV perturbations per pair shows larger robustness improvement than simply fine-tuning on 1-most.

in Figure 8. We see that using all 3 test sets for NAV fine-tuning is consistently a better option over any given single set.

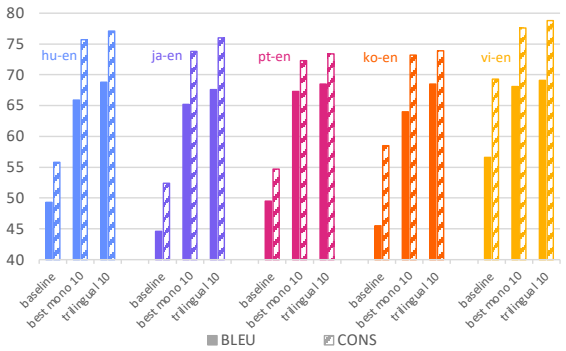

Figure 8: Multilingual (trilingual) fine-tuning improves both BLEU and CONSIST for every evaluation language pair compared to that pair's previous best model using monolingual fine-tuning.

6 Synthetic Experiments

We repeat experiments from the previous section using our synthetic fine-tuning sets (Section 3.3) and compare to the organic 10-random sets. We also combine all of the synthetic sets {hu,ja,pt}-en for multilingual fine-tuning and evaluate on our NAV robustness evaluation sets. Results are shown in Figure 9.

Our synthetic fine-tuning sets do improve robustness compared to a baseline model, but they are far from achieving the same improvements as the organically generated data. We also see how our NAV-oriented synthetic data (ja-en) more closely approximates organic NAV data gains compared to our non-NAV synthetic data ({hu,pt}-en).

We also test synthetic fine-tuning performance in zero-shot language transfer. Results are shown in

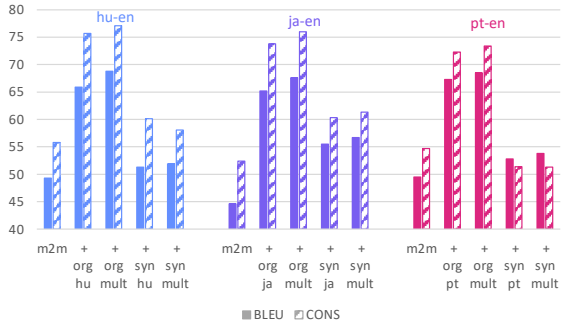

Figure 9: BLEU and CONSIST scores for organic vs. synthetic perturbations during fine-tuning. Synthetic perturbations improve robustness over baselines but organic perturbations are much more useful.

Figure 10. We continue to see that synthetic data is not able to provide the same NAV robustness gains as organically-generated data. However, in zero-shot language transfer, multilingual synthetic data more closely approaches organic-data performance than it can with in-language fine-tuning. This suggests in-language organic NAV data is most useful even if some NAV robustness can be improved with transfer.

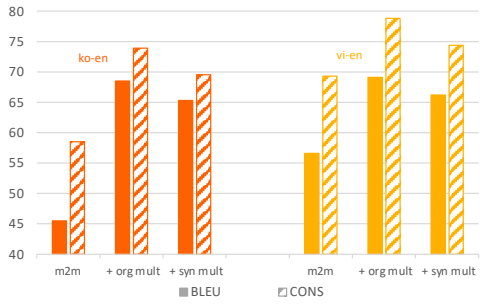

Figure 10: BLEU and CONSIST scores for organic vs. synthetic perturbations during fine-tuning in zero-shot language transfer. Synthetic perturbations improve robustness over baselines but organic perturbations are much more useful. However, this gap is smaller than that shown with in-language experiments.

7 Additional Evaluations

Results from the previous sections suggest that NAV robustness can be effectively learned (even in zero-shot scenarios) by exposing models to a variety of several NAV perturbations per translation example. However, these findings raise new questions about what specifically the models are learning, how transferable that learning is and how else the model's behavior changes.

In our next set of experiments, we take our best models from Sections 4 and 5, which we designate

as our 'NAV-robust' models and evaluate them compared to baseline MT models in other evaluation settings.

7.1 Out-of-Domain MT

To further investigate behavior of NAV-robust models, we run experiments to see the effect on performance in OOD tasks. For {hu,ja,pt}-en we use a held-out test split of the Tatoeba corpus. Our results are shown in Figure 11. For {ko,vi}-en we use the official OPUS-100 test set with results shown in Figure 12.

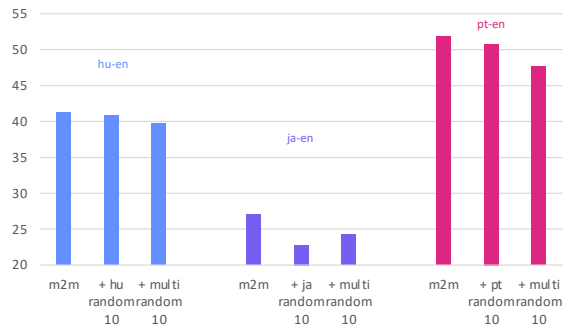

Figure 11: BLEU scores on held-out Tatoeba test sets. Baseline models perform better than NAV-robust models.

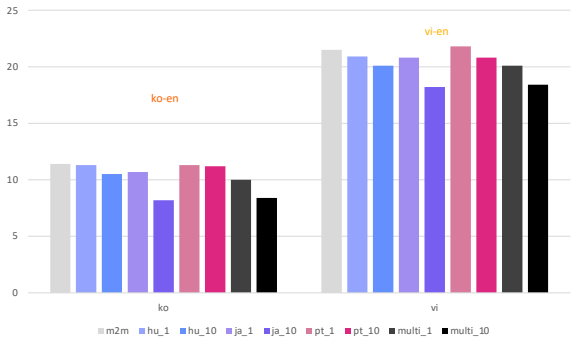

Figure 12: BLEU scores on held-out OPUS-100 test sets. Baseline models perform better than most NAV-robust models.

From these experiments, we see that generally NAV-robustification slightly worsens OOD performance. While these results appear negative, they are not too surprising considering common trends in fine-tuning and robustification. Often, fine-tuning can cause some forgetting and models with higher robustness perform worse when evaluated on original, un-noisy input. Correcting this widespread behavior is beyond the scope of this work.

While it's clear that improving NAV robustness has shown a decrease in 'regular' MT BLEU, there are many differences between our NAV evaluation settings and our experiment settings in this section. For example, these experiments are not just 1) out of domain, they also 2) no longer have a robustness component to them because we aren't evaluating the models on several perturbations (NAV or otherwise) of input sentences. The decrease in BLEU may be due to either or both of these properties, which our next subsections attempt to disentangle.

7.2 Robustness Transfer

Our next set of experiments in this section looks at measuring NAV-robustness transfer. By this we seek to answer the question *does a model fine-tuned deliberately for NAV robustness exhibit robustness to non-NAV-related noise?* For this, we use an existing MT test set designed to include noisy input text to challenge models' robustness, MTNT (Machine Translation of Noisy Text) (Michel and Neubig, 2018).

We simply evaluate our baseline and select NAV-tuned models on the ja-en test split of MTNT. Our results are shown in Figure 13. We see that our baseline performs better than our NAV-tuned models, suggesting zero-shot transfer of robustness may not be possible in this way. The NAV perturbations which our models were trained to be robust to do not overlap with many of the types of noise in MTNT, which resembles less-standard 'internet-speech'.

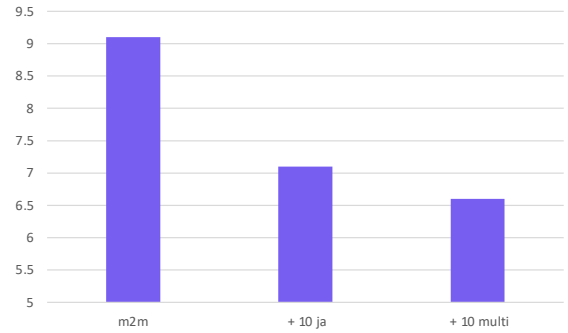

Figure 13: NAV-robust models perform worse than a baseline model in MTNT. MTNT is a difficult task (highlighted by the very low BLEU scores) and includes greater and more varied noise than within the NAV boundaries.

Again, we have another problem with this evaluation in that the models are not only tested against new classes of perturbations but also in a new domain. One way of testing performance on new

classes of perturbation but staying in-domain is to use our synthetic augmentation scripts to create a synthetic robustness test set.

We use the same scripts but apply them to the test splits of {hu,ja,pt} STAPLE. In this way, we can probe transferability of NAV robustness by evaluating our NAV-robust models on these synthetic test sets. We also evaluate the models fine-tuned on synthetic data for comparison. Results are shown in Figure 14.

In general, the synthetic-tuned models perform better on these synthetic test evaluations. However, the organic NAV-tuned models do exhibit some improved robustness compared to a baseline model. The NAV perturbations these models are fine-tuned on rarely overlap with the types of synthetic perturbations performed on the new {hu,pt}-en sets, explaining the only slight transferability. Organic NAV perturbations help more on our synthetic ja-en test set, likely because the ja-en test set attempts to synthetically create NAV perturbations, thus imitating the organic fine-tuning data more closely.

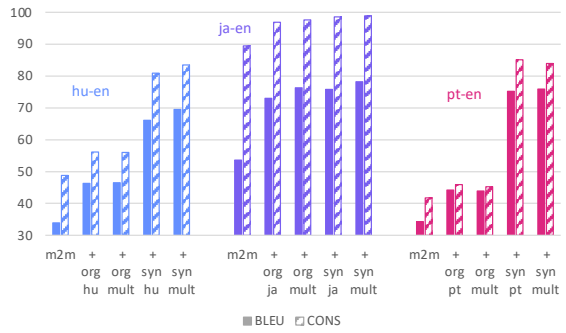

Figure 14: BLEU and CONSIST scores for organic vs. synthetic perturbations during fine-tuning on synthetic STAPLE test sets. Synthetic fine-tuning unsurprisingly improves synthetic test performance, but we also see how organic NAV fine-tuning performs fairly well on rule-based synthetic perturbation which aims to emulate NAV (ja-en) compared to random character-alteration synthetic test sets ({hu,pt}-en).

7.3 In-Domain MT

One possible source of confusion in our NAV robustness evaluation could be the fact that it is abnormal to have hundreds of test examples with the same reference translation. Perhaps BLEU is not as reliable in such conditions. To account for this, we calculate BLEU on in-domain data by using our '1-most' and '1-least' STAPLE test sets. This returns the task to single-pair examples while focusing on common and uncommon examples, respectively.

The results for '1-most' are shown in Figures 15 and 16. In our NAV experiments from Section 4, the "10 random" models outperform the "1 most" models, but here we see the opposite. This is not surprising as the '1-most' fine-tuning is most similar to this '1-most' evaluation. Also, this evaluation no longer contains a robustness aspect as each example pair uses a commonly-occurring source sentence.

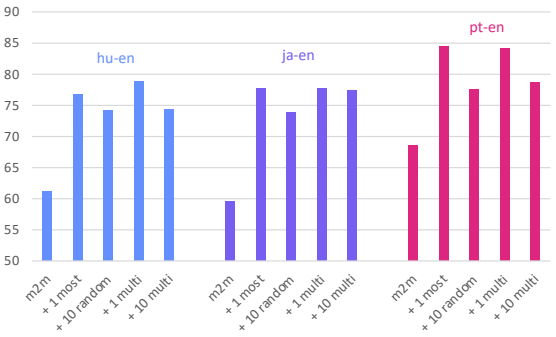

Figure 15: BLEU scores evaluating on 1-most STAPLE test sets. 1-most fine-tuning improves performance the most.

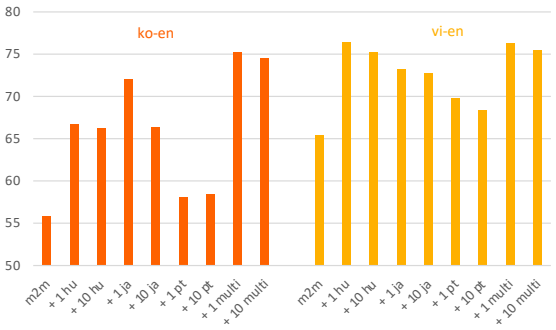

Figure 16: BLEU scores evaluating on 1-most STAPLE test sets for 0-shot language transfer. 1-most fine-tuning improves performance over baseline and 10-random models. We see similar language differences with ja-en improving ko-en more compared to hu-en with vi-en.

This raises the concern that it could be, in fact, that 10-random has higher BLEU on our robustness experiments because of the property of having hundreds of variations for each example. To disentangle this possible explanation, we also evaluate our models on a "1-least" STAPLE set. We source the *least* common source sentence for each example. This helps ensure the test inputs will exhibit more difficult NAV perturbations, thus serving as a better NAV-robustness test set without the added property of having several variations from each translation pair. The results are shown in Figures

17 and 18.

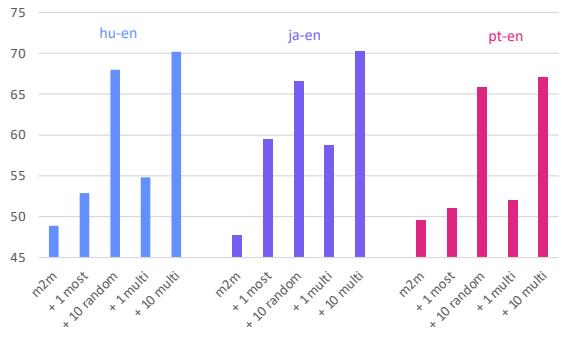

Figure 17: BLEU scores evaluating on 1-least STAPLE test sets. 10-random fine-tuning improves performance the most.

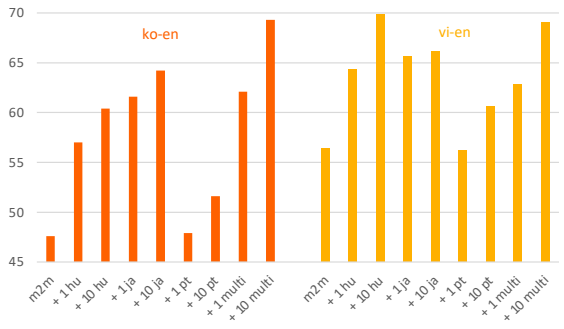

Figure 18: BLEU scores evaluating on 1-least STAPLE test sets for 0-shot language transfer. 10-random fine-tuning improves performance over baseline and 1-most models. We see language differences with ja-en improving ko-en more compared to hu-en with vi-en.

Our results now reflect those from our robustness experiments with "10-random" fine-tuning consistently performing better than "1-most". This suggests that the models showing highest NAV robustness by our evaluations also show higher robustness in a more standard MT evaluation on NAV-noisy data.

8 Discussion

Overall our work raises several questions about NAV phenomena and how to address them in MT. We are able to provide some answers to a few research questions and open pathways for future work. NAV describes a very subtle yet substantial mark of language that stands as an appropriate obstacle for current NLP research as we push the boundaries of AI.

Addressing NAV in a systematic way is an incredible challenge due to the vastness of possible variation. The STAPLE dataset provides a suitable way to begin investigating solutions to NAV-related problems in MT. We are able to show how NAV robustness can be improved using STAPLE data and how those improvements can transfer across languages. There seems to be language dependence whereby robustness learned in one language can have varying effects depending on the language transferred to.

We find that NAV-robust models can perform worse in non-NAV settings, which does not differ from several other findings in robustness work in which a more robust model may not perform better on original, non-noised input. We also address scaling issues by synthetically generating NAV and non-NAV examples. The synthetic NAV examples seem to help more than the non-NAV examples, but none of them are able to obtain the same improvements as human-created NAV data.

9 Related Work

Many other works investigate robustness in MT by considering different classes of perturbations and developing strategies to improve performance at test time. Niu et al. (2020) perform misspelling- and casing-related perturbations on MT test input and evaluate different models on their robustness to these. Salesky et al. (2021) address several classes of perturbations by replacing a standard unicode-based encoder with a visual encoder, which demonstrates higher robustness to perturbations associated with visual appearance of language, such as 1337speak. Zhang et al. (2020) generate additional clean and noisy data using back-translation and different datasets to improve performance on noisy text.

10 Conclusion

We present the reader with an under-studied subject in MT and contribute more formalized definitions and simple evaluation metrics for the phenomenon of natural asemantic variation. We also perform experiments, showing how NAV perturbations can be used during fine-tuning to improve robustness of MT models, even when evaluating on a different language-pair. Several questions remain to be explored to further formalize NAV, investigate its role in NLP modeling and improve techniques to increase nuanced understanding in AI.

References

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2019. A massive collection of cross-lingual web-document pairs. *arXiv preprint arXiv:1911.06154*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint*.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Stephen Mayhew, Klinton Bicknell, Chris Brust, Bill McDowell, Will Monroe, and Burr Settles. 2020. Simultaneous translation and paraphrase for language education. In *Proceedings of the ACL Workshop on Neural Generation and Translation (WNGT)*. ACL.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. *CoRR*, abs/2005.00580.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Elizabeth Salesky, David Etter, and Matt Post. 2021. Robust open-vocabulary translation from visual text representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7235–7252, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

A Appendix

A.1 Full Mono-Pair Results

Table 2 shows our full results from the mono-pair model experiments. In addition to CONSIST and BLEU, we also obtain a MATCH score, based on average percentage the output sentence matches the reference and NUM, the average number of different translations generated per set of semantically-equivalent NAV perturbations.

hu-en		Quality		Consistency	
cond	#	BLEU	MAT.	CONS.	NUM
base	0	28.2	11.1	43.1	21.3
	1	37.4	22.3	44.8	20.5
most	2	42.3	26.7	51.3	16.5
	10	45.4	29.6	58.0	12.4
	1	46.3	29.8	56.3	12.9
rand	2	46.9	32.0	58.3	11.9
	10	46.6	**32.3**	62.9	10.0
	1	47.2	30.6	56.4	13.0
least	2	**48.1**	31.4	58.8	11.3
	10	46.9	31.2	62.8	9.9
all	all	43.3	26.4	**64.4**	**8.4**

ja-en		Quality		Consistency	
cond	#	BLEU	MAT.	CONS.	NUM
base	0	29.0	14.5	39.5	46.8
	1	45.7	27.3	48.9	29.4
most	2	45.6	28.5	49.6	28.1
	10	50.5	**33.6**	60.5	16.2
	1	47.9	30.8	54.3	21.6
rand	2	**51.0**	33.2	57.3	18.1
	10	50.0	33.0	60.3	15.2
	1	49.4	31.7	55.1	21.2
least	2	48.8	32.2	56.8	18.6
	10	50.1	33.1	61.7	13.9
all	all	40.8	22.6	**66.4**	**9.1**

pt-en		Quality		Consistency	
cond	#	BLEU	MAT.	CONS.	NUM
base	0	38.4	12.0	35.3	34.3
	1	42.0	14.1	34.2	34.0
most	2	43.4	16.2	37.5	29.8
	10	52.8	26.1	53.3	16.9
	1	52.4	26.3	51.6	18.5
rand	2	54.8	26.3	55.8	15.7
	10	**56.9**	29.0	59.2	13.0
	1	53.8	27.9	52.8	18.2
least	2	55.7	29.4	55.3	15.3
	10	56.5	**30.0**	59.3	12.8
all	all	53.0	26.9	**63.0**	**10.0**

Table 2: Results for translation quality and consistency for three language directions under different fine-tuning conditions. Including additional perturbations increases test consistency consistently.