An Empirical study of Unsupervised Neural Machine Translation: analyzing NMT output, model's behavior and sentences' contribution

Anonymous ACL submission

Abstract

Unsupervised Neural Machine Translation (UNMT) focuses on improving NMT results under the assumption there is no human translated parallel data, yet little work has been done so far in highlighting its advantages compared to supervised methods and analyzing its output in aspects other than translation accuracy. We focus on three very diverse languages, French, Gujarati, and Kazakh, and train bilingual NMT models, to and from English, with various levels of supervision, in high- and low- resource setups, measure quality of the NMT output and compare the generated sequences word order and semantic similarity to source and reference sentences. We also use Layer-wise Relevance Propagation to analyze the model's behavior during training, and evaluate the source and target sentences' contribution to the NMT result, expanding the findings of previous works to the UNMT paradigm.

1 Introduction

Unsupervised Neural Machine Translation (UNMT) has been widely studied (Wang and Zhao, 2021; Marchisio et al., 2020; Kim et al., 2020; Lample et al., 2017; Artetxe et al., 2019; Su et al., 2019), in an effort to create efficient and trustworthy NMT models of excellent performance not relying on the existence of parallel data. Obtaining high-quality parallel corpora is expensive and time-consuming, especially for less-common language pairs. Unsupervised NMT hence aims to circumvent this limitation. NMT in general significantly aids in preserving indigenous languages by making global information accessible, supports migrants in overcoming language barriers to essential services, and enables the globalization of local news from smaller countries. UNMT particularly has broad applicability, especially in addressing linguistic diversity and information accessibility challenges.

However, there has been little effort on analyzing, apart from the quality of the output, the model behavior during UNMT, and the models' inner workings and the effects of various setups on hypotheses and generated translations' quality, monotonicity and semantic similarity, as well as model robustness and consistency. We analyze and compare UNMT approaches for two very diverse languages, French and Gujarati, translating to and from English. We research into the existence of different stages in UNMT, analyze source and target sentence tokens' contributions to the result (Bach et al., 2015) evaluate the quality and word alignment of generated translations, and Robustness and Consistency of our model to perturbed inputs. Our paper follows up closely on the work of Voita et al. (2020, 2021); Marchisio et al. (2022), and examines the following questions:

- Do the distinct stages of transformer-based NMT analyzed in previous works exist in Unsupervised, and joint Supervised & Unsupervised NMT?
- How does output quality, word alignment, semantic similarity, as well as source and target sentences' token contributions to the NMT output behave across the aforementioned stages?
- How Robust and Consistent are NMT models throughout training?

Our findings confirm the existence of NMT stages regardless of the level of training supervision, and show that Unsupervised methods produce translations more similar to source sentences in terms of word order, yet more

semantically distant. UNMT models tend to show higher Robustness and Consistency, and can more easily recover from sentence perturbations. We also observe that in reduced training data experiments, there is a heavy reliance on the source sentence for generating translations. Our focus is not on outperforming NMT state-of-the-art results, but rather on training and examining the behavior of bilingual models in various setups. We will be making our experiments' code public upon acceptance.

The paper is structured as follows: in Section 2 we present related work in the topics of UNMT, NMT analysis and other metrics analyzed in our work. In Section 3 we analyze the methods proposed for NMT analysis and the experiments conducted, while in Section 4 we present and discuss our findings. Finally, in Sections 5, 6 and 7 we conclude our work and highlight certain limitations and ethical considerations, respectively. We present additional experiments and results on the Robustness of the models and Semantic Similarity of input and output sentences in the supplementary material attached to the submission.

2 Related Work

Unsupervised Neural Machine Translation

UNMT aims to make NMT work in the absence of parallel data. Most common approaches have focused on cross- or multilingual initialization of a language model either through an alignment of monolingual embeddings (Artetxe et al., 2017; Lample et al., 2018; Conneau et al., 2017; Lample et al., 2017) or by model pretraining and fine-tuning (Lample and Conneau, 2019; Song et al., 2019; Liu et al., 2020). Back-Translation (BT) (Sennrich et al., 2015) translates monolingual data between languages, creating pseudo-parallel training corpora (Artetxe et al., 2017; Lample et al., 2017). Marchisio et al. (2022) first systematically examine the naturalness and diversity of the UNMT output, comparing it to similar quality human translations, and proposing a way to leverage UNMT to improve a classical supervised NMT system. In more recent works, Liu et al. (2022) introduce a flow-adapter architecture to separately model the distributions of source and target languages, and He et al. (2022) identify and mitigate a training and inference style and content gap between back-translated data and natural source sentences. Garcia et al. (2020a) expand the paradigm to multilingual UNMT, while Garcia et al. (2020b) use offline BT synthetic data to improve multilingual En-xx UNMT for low-resource languages xx.

Layer-wise Relevance Propagation (LRP)

LRP (Bach et al., 2015) measures relevance of the input components, or the neurons of a network, to the next layers' output, and is directly applicable to layer-wise architectures. Wu and Ong (2021) use LRP as an attribution method for sequence classification tasks. We extend its usage to the Transformer, and measure the relevance of source and target sentences to the NMT output.

Neural Machine Translation analysis

Voita et al. (2020) examine the source and target sentences' tokens' relative contributions to NMT output, adapting LRP to a Transformer, and experimenting with different training objectives, training data amounts and types of target sentence prefixes, and their effect on NMT output quality and monotonicity. Following up to that work, Voita et al. (2021) analyze NMT stages, drawing parallels to distinct SMT stages. Their findings include decomposing NMT into three phases, and using the key learning advantages of each stage to improve non-autoregressive NMT. We examine and identify if those stages exist in UNMT.

Robustness & Consistency

Previous works examine Robustness in NLP (Yu et al., 2022; Wang et al., 2021; La Malfa and Kwiatkowska, 2022), measuring and improving NLP models' performance against perturbed or unseen input. Specifically for NMT, Niu et al. (2020) propose two metrics, Robustness and Consistency to measure sensitivity of a model to input perturbations.

3 Method & Experiments

3.1 Model

We use a 6-layers 8-heads transformer-based model, XLM (Lample and Conneau, 2019), following the training configurations and hyperparameters suggested by the authors. We use Byte Pair Encoding to extract a 60k vocabulary, an embedding layer size of 1024, a dropout value and an attention layer dropout value of 0.1, and a sequence length of 256. We measure the quality of the Language Model (LM) with perplexity, and quality of the NMT output with BLEU both used as training stopping criteria,

when there is no improvement over 10 epochs. We first pre-train a LM in each language with the MLM objective, and use it to initialize the encoder and decoder of the NMT model. We then train NMT models, using Back-Translation (BT) and denoising auto-encoding (AE) with the monolingual data used for LM pretraining for UNMT, the Machine Translation (MT) objective for the Supervised NMT model, and BT-MT for the joint Unsupervised and Supervised approach.

3.2 Datasets

The languages we work with are English, French, Gujarati, and Kazakh and we're translating in all directions, English–French (En–Fr), French–English (Fr–En), English–Gujarati (En–Gu), Gujarati–English (Gu–En), English–Kazakh (En–Kk), Kazakh–English (Kk–En). For English and French, we use 5 million News Crawl 2007-2008 monolingual sentences for each language, and 23 million WMT14 parallel sentences. For Gujarati, we have 1.4 million sentences and for Kazakh we have 9.5M monolingual sentences, collected for both languages from Wikipedia, WMT 2018, 2019 and Leipzig Corpora (2016)[1]. As parallel data, we have 22k sentences from the WMT 2019 News Translation Task[2] for Gu–En and Kk–En, respectively/ As development and test sets, we use newstest2013 and newstest2014, respectively, for En–Fr and Fr–En, WMT19 for En–Gu and Gu–En and En–Kk and Kk–En.

3.3 Layer-wise Relevance Propagation

Voita et al. (2020) explain how LRP calculation in a Transformer is confusing due to the non-clear layered nature of the model. We follow their setup, with LRP to be propagated first inversely through the decoder and the encoder, up to the input model layer, and without assuming the conservation principle holds per layer, but only across processed tokens. LRP is the relevance of input neurons to the top-1 logit predicted by the model, and token contribution is the sum of the input neurons' relevance. Total source and target sentence contributions to the result at generation step t are given by $R_t(source) = \sum_i x_i$, $R_t(target) = \sum_{j=1}^{t-1} y_j$. At every step t, Relevance follows the conservation principle: $R_t(source) + R_t(target) = 1$. At step 1, we have $R_1(source) = 1$, $R_1(target) = 0$. For every target token past the currently generated one, LRP is 0.

3.4 Word Order

Our aim is to examine differences in word order between translations and reference or source sentences. We evaluate two different reordering metrics, Fuzzy Reordering Score (FRS) (Talbot et al., 2011; Nakagawa, 2015)[3] and Translation Edit Rate (TER) (Snover et al., 2006). FRS ranges between 0 and 1, with larger values for highly monotonic alignments (higher structural similarity and closer word order). For a translation y' and reference y (or source sentence y): $FRS(y', y) = 1 - \frac{C-1}{M-1}$. C is the number of chunks of contiguously aligned translation words, intuitively perceived as the number of times a reader would need to jump in order to read the system's reordering of the sentence in the order proposed by the reference of length M. With *fast_align*[4] we calculate word alignments. TER is defined as $TER(y', y) = \frac{E}{L_y}$, where E is the number of edits needed to modify the produced translation y' to match the reference sentence y (or source sentence y), and L_y is the length of the reference (or the source sentence, respectively). TER ranges between 0 and 1, and low values indicate more monotonic alignments.

3.5 Model Robustness

Niu et al. (2020) set *TQ* (y', y) to be the model quality of model M with translation y', and reference y, to define the concepts of Consistency and Robustness. Consistency of a model M on input x and perturbation δ is given by $CONSIS(M|x, \delta) = Sim(y_\delta, y)$, where y is the reference translation, and y_δ is the translation of the perturbed sentence. *Sim* is the harmonic mean between model quality *TQ*(y', y_δ') and *TQ*(y_δ', y'), measured between translations y', y_δ', respectively. On the other hand, Robustness of a model M is defined as the ratio between quality of the model producing translations y_δ' and y': $ROBUST(M|x, \delta) = \frac{TQ(y'_\delta, y)}{TQ(y', y)}$. It takes values in [0,1]. We evaluate Consistency and Robustness on test sets perturbed with two different approaches: a. **misspelling** - each word is misspelled (random deletion, insertion or substitution of characters) with a probability of 0.1, b. **case-changing** - each sentence is modified (upper-casing, lower-casing or title-casing all letters) with a probability of 0.5.

[1] https://wortschatz.unileipzig.de/en/download/
[2] http://data.statmt.org/news-crawl/
[3] https://github.com/google/topdown-btg-preordering
[4] https://github.com/clab/fast_align

3.6 Semantic Similarity

In evaluating semantic similarity between human translations or source sentences and generated translations, we calculate the Ratio Margin-based Similarity Score (RMSS) between each reference or source sentence and its k-nearest neighbors among all translations (Artetxe and Schwenk, 2018). We follow Keung et al. (2021) to obtain and mean-pool the mBERT embedding vectors of all sentences. We set $cos(\cdot, \cdot)$ to be the cosine similarity and $\text{NN}_k^{src}(x)$ the k nearest neighbors of x in the reference or source sentence embedding space. RMSS is high when source and target pairs compared are closer than their respective nearest neighbors. For a hypothesis y:

$$RMSS(x, y) = \frac{cos(x, y)}{\sum_{z \in \text{NN}_k^{tgt}(x)} \frac{cos(x,z)}{2k} + \sum_{z \in \text{NN}_k^{src}(y)} \frac{cos(y,z)}{2k}}.$$

4 Results & Discussion

BLEU scores of converged models are seen in Table 1. In En–Gu, En–Kk, Kk–En, BT-MT improves BLEU; Models trained with parallel data show higher BLEU, and absence of parallel data or often introduction of low quality data (eg in Gu–En) through BT lowers output quality.

FRS

In most En–Fr, Fr–En experiments, there is a large fluctuation yet a small and gradual FRS increase between translations and references (Fig. 1), and then a small decrease. Higher FRS shows more monotonic alignments. Starting from non-monotonic alignments in the first stage, we get maximum FRS values in the second stage of training - highly aligned translations and references - which slightly decrease in the third stage until model convergence. With parallel data we get the most monotonic alignments, while we have the least identical reorderings between references and translations in BT-only cases, in both high- and low-resource setups. Similar patterns are observed in En–Kk and Kk–En, where we have the least monotonic alignments for few/no parallel data (MT-22k, BT-MT-22k, BT), and the most for experiments with parallel data (MT, BT-MT), with values slightly increasing and then remaining stable throughout training in most cases. BT En–Gu, Gu–En models show high and steady FRS values: between languages with a complicated and non-monotonic alignment, BT produces translations more aligned with the reference.

For En–Fr, Fr–En, FRS (Fig. 2) values are stable throughout training, and BT, BT-MT experiments' results imply highly monotonic alignments; with BT, translations are closer to source sentences in terms of word order. FRS is lower in MT only experiments, as source and translation alignments are less monotonic when models are trained with parallel data alone. Results are similar in En–Gu, Gu–En. BT, BT-MT results show an almost perfect alignment between source sentences and translations.

For En–Kk, Kk–En, it is interesting to observe that in the majority of experiments, source sentences are highly monotonic to translations, with steady FRS values throughout training.

We see that BT yields more stable and higher alignment scores compared to models trained only on parallel data, suggesting it offers a significant advantage for improving translation quality.

TER

Observing TER between translations and references (Fig. 3), in En–Fr and Fr–En, low TER for MT, BT-MT means more monotonic alignments, in contrast to higher TER in low-resource and BT-only experiments. TER gradually decreases for all models, as sentences generated at the end of training highly resemble human translations. Results are different for En–Gu and Gu–En; TER is low in BT- only and BT-MT models, but rather high, and increasing, in the MT-only model, with translations in the former case very close to references. BT produces more monotonic to the reference translations for a language diverse in terms of script, morphological complexity and word order from English.

Stable or slightly increasing TER values between source sentences and translations (Fig. 4) mean high structural resemblance. For En–Fr and Fr–En, BT and BT-MT show the lowest TER values, hence generated with BT sequences and source sentences have high monotonicity. Similarly, in En–Gu and Gu–En, BT, BT-MT models show lower TER and higher and more monotonic alignment of translations and source sentences.

For En–Kk, Kk–En, we see that we have higher monotonicity in higher/no supervision experiments, and lower in low-resource models- we can assume training with few parallel data

Method	en-fr	fr-en	en-gu	gu-en	en-kk	kk-en
Other methods						
	45.9[5]	-	0.1[6]	0.3[7]	2.5[8]	7.4[9]
Our method						
BT+AE	21.76	21.69	0.4	0.46	0.7	1.0
Parallel data: 22k						
MT	31.12	30.63	1.04	**2.65**	2.4	2.6
BT+AE+MT	34.54	34.02	**1.16**	2.19	2.8	2.9
132k						
MT	37.8	35.6	-	-	5.2	8.0
BT+AE+MT	38.6	38.4	-	-	**6.6**	**8.9**
1m						
MT	41.25	41.33	-	-	-	-
BT+AE+MT	40.37	40.4	-	-	-	-
2.5m						
MT	40.46	40.71	-	-	-	-
BT+AE+MT	39.88	39.62	-	-	-	-
5m						
MT	41.52	41.18	-	-	-	-
BT+AE+MT	40.89	40.8	-	-	-	-
23m						
MT	**41.75**	**41.41**	-	-	-	-
BT+AE+MT	41.29	40.99	-	-	-	-

Table 1: BLEU scores for En–Fr, Fr–En, En–Gu, Gu–En, En–Kk, Kk–En NMT. Test and validation sets are WMT19 for Gujarati and Kazakh, newstest2013-14 for French pairs. State-of-the art results given for the sake of consistency. *MT* stands for machine translation objective, *BT* stands for Back-Translation and *AE* for denoising auto-encoding.

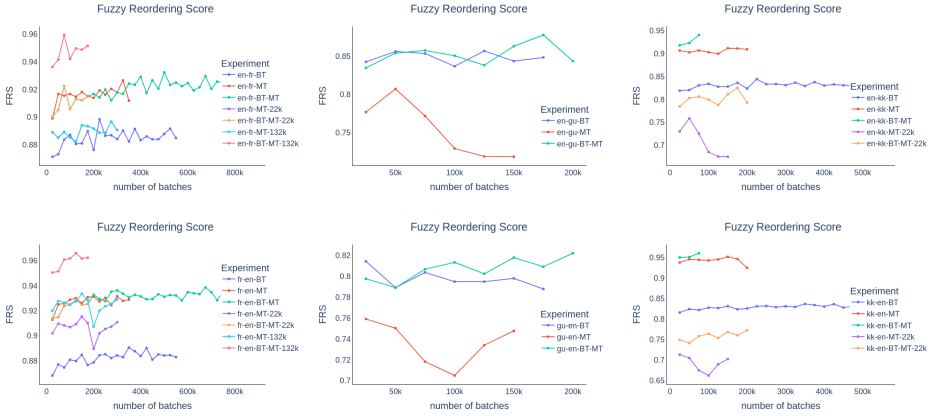


Figure 1: Fuzzy Reordering Scores (FRS) between references and generated translations, for En–Fr, Fr–En, En–Gu, Gu–En, En–Kk, Kk-En during training.

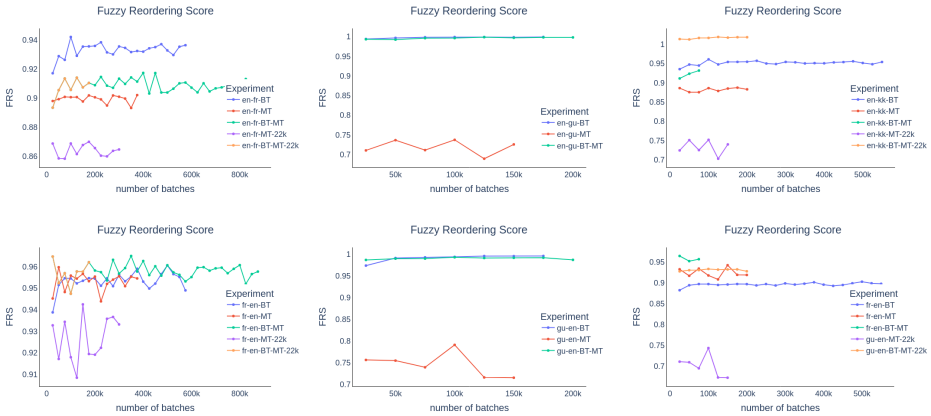


Figure 2: Fuzzy Reordering Scores (FRS) between source sentences and generated translations for En–Fr, Fr–En, En–Gu, Gu–En, En–Kk, Kk–En during training.

and for few epochs highly cannot help the model properly align produced translations and references or source sentences. Between translations and source sentences, when the model is sufficiently

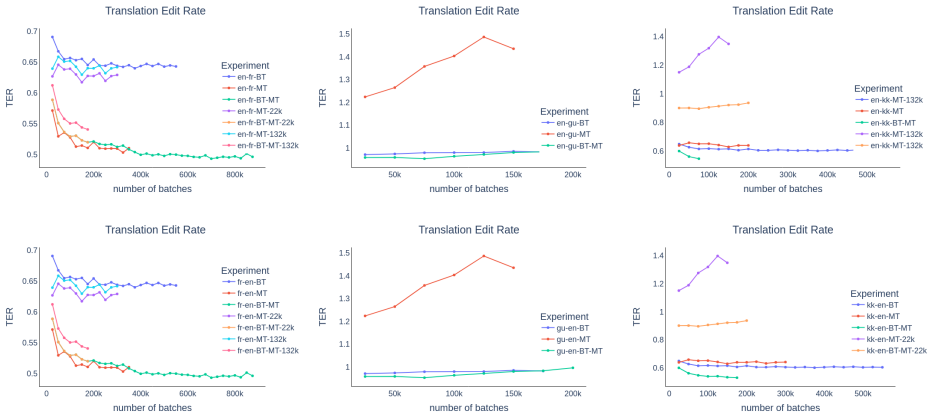


Figure 3: Translation Edit Rate (TER) between references and generated translations for En–Fr, Fr–En, En–Gu, Gu–En, En–Kk, Kk–En during training.

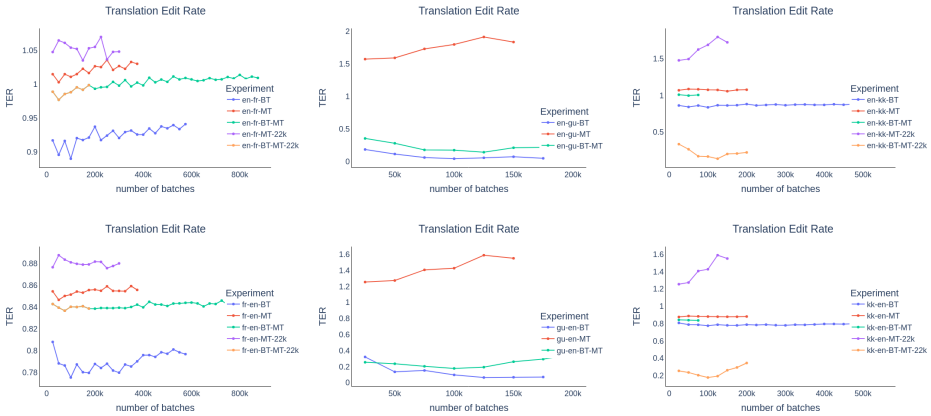


Figure 4: Translation Edit Rate (TER) between source sentences and generated translations for En–Fr, Fr–En, En–Gu, Gu–En, En–Kk, Kk–En during training.

trained, we surprisingly observe high monotonicity across experimental setups.

We can deduce that for En–Fr, Fr–En, translations have higher monotonicity to references in MT, BT-MT, lower in BT-only experiments, but higher monotonicity to source sentences in BT, BT-MT and lower in MT. Training supervision leads to better translation to hypothesis alignment, while BT induces better translation to source sentence alignment.

Hence, we see that the effectiveness of different training methods like MT and BT varies by language pair, with BT showing particular promise for languages are structurally diverse from English.

LRP analysis

Our observations from average source contribution, entropy of source contributions and entropy of target contributions during training confirm the findings of Voita et al. (2020, 2021). Changes in sentence contributions are not necessarily monotonic to the result, can help distinguish different training stages, and identify the balance between source and target sequences' relevance to the result (Fig. 5, 6, 7, 8).

For En–Fr and Fr–En, En–Kk and Kk–En NMT models (Fig. 5, 6) average source sentence contributions drop at the very beginning of training, while contributions are lowest in both directions in MT, and slightly higher in BT, BT-MT experiments; using only parallel (natural) data during training, average source contributions are lower (Voita et al., 2020) and the model relies more on the target prefix for sequence generation, while BT boosts the influence of source sentence to the result. Average contributions are mostly stable or slightly decrease as training progresses, and the source sentence becomes less important in sequence generation. For models trained with less data, contributions and relevance of the source sentence tokens to the generated sentence is high, due to the lack of substantial supervision.

Entropy of source contributions is high for MT-only experiments, contributions are more focused, and the model is more confident in choosing the important source tokens, while in BT-only and BT-MT experiments it requires broader source context for target sequence generation, and entropy of

contributions is high, for both evaluation directions. In MT-setups, training converges faster.

Studying the entropy of the target contributions, in both En–Fr and Fr–En directions, target entropy is more focused during the first part of training. We then notice either a small (BT, MT-22k, BT-MT-22k) or a larger (MT, BT-MT) increase, which gradually evens out as the model converges. Experiments with a small amount of training data, and/or BT have significantly lower entropy contributions than MT-only, with BT contributing to the model having higher confidence in choosing the target tokens generated. On the contrary, in Fr–En, combined experiments seem to have the highest, hence less focused target contributions;

Contributions' patterns are not similar for En–Gu and Gu–En models (Fig. 8, 9). Average source contributions in MT experiments are higher than those with BT, implying that using parallel data in training forces the model to rely on source tokens more heavily. Average source contributions are lowest in BT-only experiment and target sentence reliance for generation is highest.

Patterns in entropy of source contributions resemble those in En–Fr, Fr–En experiments. Entropy is low in MT-only; training with parallel data increases model confidence in selecting the important source tokens for target generation, while entropy in BT, BT-MT experiments is similarly high. We notice an increase in entropy of target contributions and high values in MT-only experiments in both directions, which validates our hypothesis that source contributions are more focused in these cases while the entropy in BT experiments is lower. Looking for differences between evaluation directions, En–Gu contributions in MT- and BT-MT are similar to those in the En–Fr low resource experiments, in contrast to training in the other direction.

We can conclude that back-translation (BT) boosts the influence of source sentences, particularly in low-resource settings, while also highlighting that sentence contributions are not necessarily monotonic and can indicate different training stages.

In Tables 3, 4 in the Supplementary material we show a few example sentences and their translations, at the beginning and end of training of each model. Examples of sentences and their perturbations are given in Table 5.

5 Conclusions

We conduct an extensive analysis of Supervised and/or Unsupervised NMT models' behavior for French, Gujarati and Kazakh NMT, to and from English, and examine the output in terms of quality, word order, semantic similarity and reliance on source and reference sentences. Our results highlight the importance of supervision for output quality, yet outline the superiority of UNMT in generating sentences highly aligned to references and in preserving models' robustness. We hope our work sets the ground for better understanding and improving UNMT and our findings can be utilized to improve real-world UNMT systems.

6 Limitations

It is a computationally hard task to train large Neural Machine Translation models from scratch and the complexity of the training process is high, calling for more efficient training solutions, in terms of memory distribution of the model and parallelization. It is strongly recommended to design a more systematic approach to addressing those factors and expand to more languages, in order to achieve further generalization of the method and overcome all current limitations. Moreover, results for low-resource NMT systems may often be poor, or marginally improving state-of-the-art, calling for improvement in NMT methods to boost performance.

7 Ethical Considerations

The authors of the paper are aware that when training large language models, several issues ought to be taken into account, related to quality, toxicity and bias related to their training process and output (Bender et al., 2021; Chowdhery et al., 2022; Brown et al., 2020).

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. *arXiv preprint arXiv:1902.01313*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

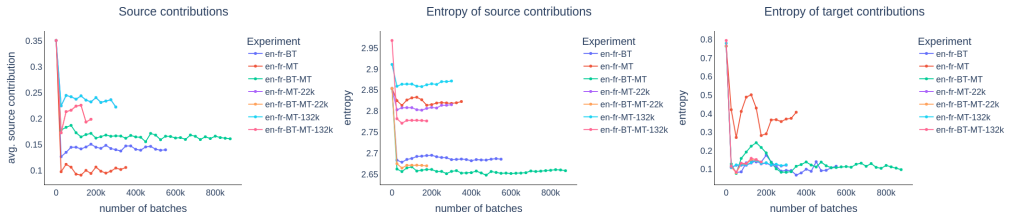


Figure 5: En–Fr Average Source Contribution, Entropy of Source Contributions and Entropy of Target Contributions during training.

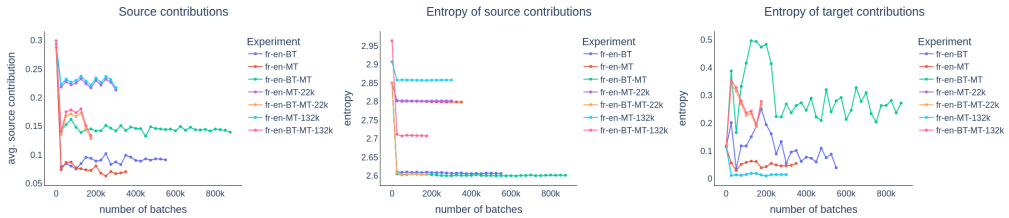


Figure 6: Fr–En Average Source contribution, Entropy of Source Contributions and Entropy of Target Contributions during training.

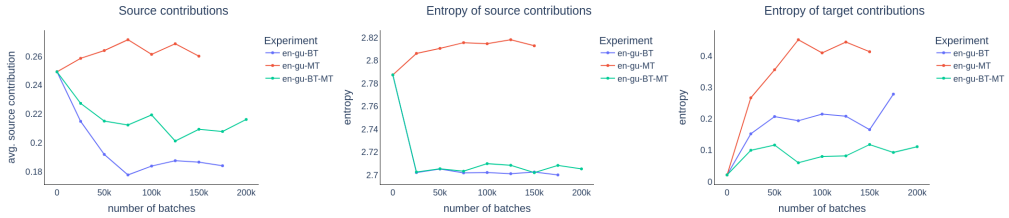
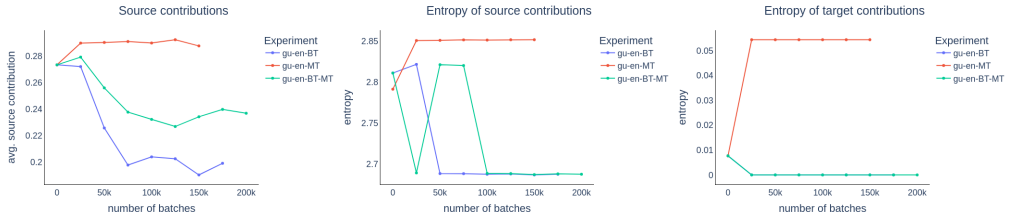


Figure 7: En–Gu Average Source Contribution, Entropy of Source Contributions and Entropy of Target Contributions during training.



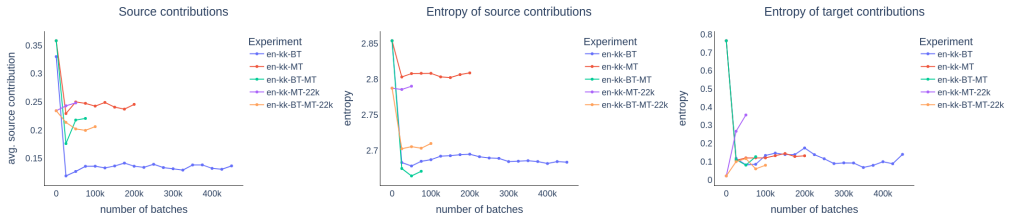


Figure 8: En–Kk Average Source Contribution, Entropy of Source Contributions and Entropy of Target Contributions during training.

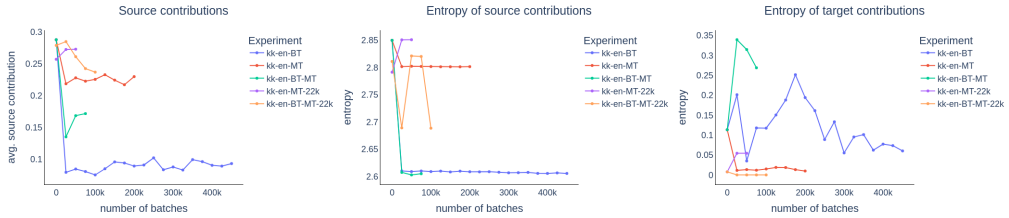


Figure 9: Kk–En Average Source Contribution, Entropy of Source Contributions and Entropy of Target Contributions during training.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On

the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur P Parikh. 2020a. A multilingual view of unsupervised machine translation. *arXiv preprint arXiv:2002.02955*.

Xavier Garcia, Aditya Siddhant, Orhan Firat, and Ankur P Parikh. 2020b. Harnessing multilinguality in unsupervised machine translation for rare languages. *arXiv preprint arXiv:2009.11201*.

Zhiwei He, Xing Wang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2022. Bridging the data gap between training and inference for unsupervised neural machine translation. *arXiv preprint arXiv:2203.08394*.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah A. Smith. 2021. Unsupervised Bitext Mining and Translation via Self-Trained Contextual Embeddings. *Transactions of the Association for Computational Linguistics*, 8:828–841.

Yunsu Kim, Miguel Graça, and Hermann Ney. 2020. When and why is unsupervised neural machine translation useless? *arXiv preprint arXiv:2004.10581*.

Emanuele La Malfa and Marta Kwiatkowska. 2022. The king is naked: on the notion of robustness for natural language processing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11047–11057.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Yihong Liu, Haris Jabbar, and Hinrich Schütze. 2022. Flow-adapter architecture for unsupervised machine translation. *arXiv preprint arXiv:2204.12225*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? *arXiv preprint arXiv:2004.05516*.

Kelly Marchisio, Markus Freitag, and David Grangier. 2022. On systematic style differences between unsupervised and supervised mt and an application for high-resource machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2214–2225.

Tetsuji Nakagawa. 2015. Efficient top-down btg parsing for machine translation preordering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 208–218.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. 2019. Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10482–10491.

David Talbot, Hideto Kazawa, and Hiroshi Ichikawa. 2011. A lightweight evaluation framework for machine translation reordering.

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*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Language modeling, lexical translation, reordering: The training process of nmt through the lens of classical smt. *arXiv preprint arXiv:2109.01396*.

Rui Wang and Hai Zhao. 2021. Advances and challenges in unsupervised neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 17–21.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2021. Measure and improve robustness in nlp models: A survey. *arXiv preprint arXiv:2112.08313*.

Zhengxuan Wu and Desmond C Ong. 2021. On explaining your explanations of bert: An empirical study with sequence classification. *arXiv preprint arXiv:2101.00196*.

Yu Yu, Abdul Rafae Khan, and Jia Xu. 2022. Measuring robustness for nlp. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3908–3916.

Robustness

En–Fr and Fr–En NMT models are highly robust in all MT, BT-MT setups (Table 2), especially when test sets are misspelled. On the contrary, En–Gu and Gu–En models are highly robust on BT, BT-MT experiments, on test sets perturbed by case-changing; for highly morphological complex languages, BT may help boost model robustness.

A similar behavior is observed for En–Kk, Kk–En. Models are highly robust on case-changing in all unsupervised, supervised and semi-supervised scenarios, and as the amount parallel sentences increases, we see an expected increased robustness to sentences' misspelling; the models become more robust to a high percentage of sentence perturbations with higher training supervision. In En–Fr and Fr–En NMT models, consistency patterns are similar to those found for Robustness: models are highly consistent in MT, BT-MT experiments, primarily when test sentences are misspelled, with their consistency increasing by the amount of parallel train data. BT-only training does not seem to help. Model consistency patterns for En–Gu, Gu–En follow those of Robustness, with BT outperforming other methods.

Semantic Similarity

MT-only and BT-MT experiments show high RMSS values in En–Fr, Fr–En between translations and references (Fig. 10), which have a higher semantic similarity than in BT-only or in reduced dataset experiments, On the contrary, in En–Gu and Gu–En, translations from MT models are less similar to references, and most similar in BT-only experiments, for which RMSS is highest. Source sentences show high semantic similarity to translations in MT-only experiments, followed by reduced-data model training results, outperforming BT-only or BT-MT models, in En–Fr and Fr–En; in the first direction, RMSS is very similar across models, while in the latter, behavior of the model in different setups is significantly more distinct. For En–Gu and Gu–En, BT-only experiments show highest semantic similarity between source sentences and translations (Fig. 11).

	En–Fr			Fr–En			En–Gu			Gu–En			En–Kk			Kk–En		
	BLEU	R	C	BLEU	R	C	BLEU	R	C	BLEU	R	C	BLEU	R	C	BLEU	R	C
BT																		
• original	**22.6**	-	-	**21.78**	-	-	0.31	-	-	0.36	-	-	0.7	-	-	1.0	-	-
• misspelling	14.77	0.65	**17.82**	16.86	**0.77**	**16.52**	**2.49**	0.03	**1.59**	**3.27**	0.08	**1.33**	**1.5**	0.14	**0.8**	0.7	0.7	**1.2**
• case-changing	14.87	**0.66**	13.34	14.56	0.66	11.3	1.22	**0.93**	0.45	0.91	**0.52**	0.65	1.2	**0.71**	0.62	**1.8**	**0.8**	0.84
22k																		
MT																		
• original	**31.12**	-	-	**30.63**	-	-	**2.51**	-	-	**0.77**	-	-	**2.4**	-	-	**2.6**	-	-
• misspelling	30.22	**0.97**	**26.03**	30.4	**0.99**	**31.25**	0.05	**0.01**	0	0.23	0.29	0.3	1.4	0.58	1.1	1.2	0.46	**1.3**
• case-changing	20.01	0.64	22.13	23.34	0.76	24.93	0	0	0	0.33	**0.42**	**0.47**	1.9	**0.79**	1.3	2.2	**0.84**	1.9
BT+AE+MT																		
• original	**34.42**	-	-	**33.87**	-	-	**1.08**	-	-	2.19	-	-	2.8	-	-	2.9	-	-
• misspelling	31.79	**0.92**	**28.18**	32.62	**0.96**	**33.44**	0.72	0.66	**0.79**	**3**	0.36	**2.37**	**3.2**	0.14	1.7	3.0	0.02	**3.1**
• case-changing	23.06	0.66	22.52	27	0.79	24.22	0.9	**0.83**	0.6	2.12	**0.96**	1.88	2.5	**0.89**	**1.83**	**3.5**	**0.2**	2.7
132k																		
MT																		
• original	**37.8**	-	-	**35.6**	-	-	-	-	-	-	-	-	5.2	-	-	8.0	-	-
• misspelling	35.2	**0.93**	**30.2**	33.8	**0.94**	**35.3**	-	-	-	-	-	-	4.9	0.94	14.6	7.8	**0.97**	7.5
• case-changing	24.01	0.63	23.1	25.2	0.7	25.4	-	-	-	-	-		5.0	**0.96**	4.7	7.7	0.96	**7.6**
BT+AE+MT																		
• original	**38.6**	-	-	**38.4**	-	-	-	-	-	-	-	-	5.8	-	-	8.9	-	-
• misspelling	36.9	0.95	32.5	36.2	**0.94**	**34.7**	-	-	-	-	-	-	**4.7**	**0.81**	5.4	8.2	**0.92**	8
• case-changing	25.1	0.68	24.5	26.0	0.67	26.1	-	-	-	-	-	-	4.2	0.72	4.0	8	0.89	**8.4**
23m																		
MT																		
• original	**41.84**	-	-	**41.41**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
• misspelling	40.16	**0.96**	**35.36**	40.63	**0.98**	**41.5**	-	-	-	-	-	-	-	-	-	-	-	-
• case-changing	27.26	0.65	29.68	30.08	0.72	**30.79**	-	-	-	-	-	-	-	-	-			
BT+AE+MT																		
• original	**42.63**	-	-	**42.37**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
• misspelling	39.71	**0.93**	**35.04**	40.72	**0.96**	**41.49**	-	-	-	-	-	-	-	-	-	-	-	-
• case-changing	28.12	0.65	27.31	33.58	0.79	30.33	-	-	-	-	-	-	-	-	-	-	-	-

Table 2: BLEU scores, Robustness (R) and Consistency (C) values for Unsupervised (BT), Supervised (MT), and Unsupervised + Supervised (BT+AE+MT) NMT experiments, for the converged model for En–Fr, Fr–En, En–Gu, Gu–En and En–Kk, Kk–En. Test and validation sets are from WMT19 for Gujarati and Kazakh, and newstest 2013-14 for French, and are perturbed following the method suggested in (Niu et al., 2020) for misspelling and case-changing.

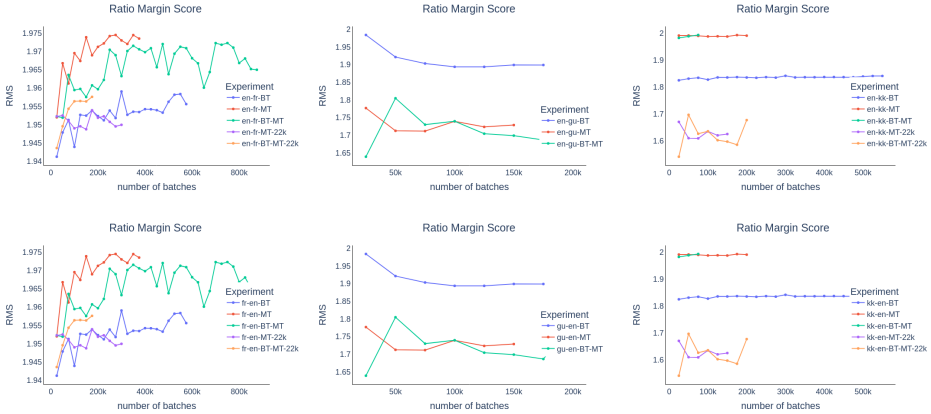


Figure 10: RMSS between references and generated translations for En–Fr, Fr–En, En–Gu, Gu–En, En–Kk, Kk–En during training.

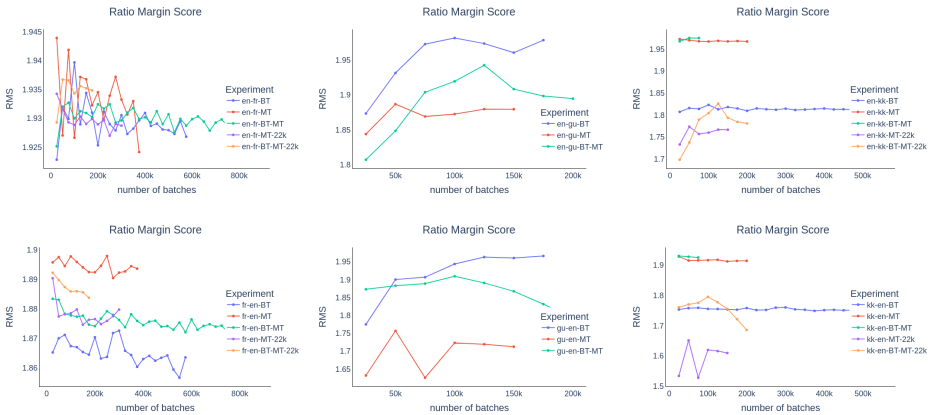


Figure 11: RMSS between source sentences and generated translations for En–Fr, Fr–En, En–Gu, Gu–En, En–Kk, Kk–En during training.

Original sentence	**fr**	la certification mondiale de la polio eradication ;
	en	global certification of polio eradication ;
MT-22k	**en** - FC	world world dication dication conference
	en - LC	world world prohibition of poliomyelitis ;
BT-MT-22K	**en** - FC	global eradication of geromyelite ;
	en - LC	global eradication of poliomyelitis ;
BT	**en** - FC	the global eradication of poliomyelite;
	en - LC	the global eradication of poliomyelite;
MT	**en** - FC	the global eradication of poliomyelitis;
	en - LC	global eradication of polio;
BT-MT	**en** - FC	global eradication of poliomyelitis ;
	en - LC	global eradication of poliomyelitis;

Original sentence	**fr**	dans notre region , la democratie est une valeur centrale .
	en	in our region , democracy is a fundamental value .
MT-22k	**en** - FC	in our region , democracy is a core value .
	en - LC	democracy in our region is a central value .
BT-MT-22K	**en** - FC	in our region , democracy is a central value .
	en - LC	in our region , democracy is a central value .
BT	**en** - FC	In our region , democratie is a central value .
	en - LC	In our region , democratie is a central value .
MT	**en** - FC	in our region , democracy is a central value .
	en - LC	democracy our region has a central value .
BT-MT	**en** - FC	in our region , democracy is a central value .
	en - LC	democracy our region is a central value .

Table 3: Two examples of a French sentence, their English ground-truth translation, and their English translations with each model's first and last checkpoint (**en** - FC, **en** - LC respectively)

Original sentence	**gu**	The આતંકી નેતાઓને સતત ધમકીઓ આપી રહ ◌ યા છે કે તેઓ ચૂંટણીમાં ભાગ ન લે .
	en	The militants are constantly intimidating the politician not to participate in the election .
BT	**en** - FC	The આતંકી નેતાઓને સતત ધમકીઓ આપી રહ ◌ યા છે કે તેઓ ચૂંટણીમાં ભાગ ન લે .
	en - LC	આતંકી નેતાઓને સતત ધમકીઓ આપી રહ ◌ યા છે કે તેઓ ચૂંટણીમાં ભાગ ન લે .
MT	**en** - FC	આતંકી નેતાઓને ધમકીઓ આપી રહ ◌ે યા છે કે તેઓ ચૂંટણીમાં ભાગ ન લે .
	en - LC	The આતંકી સતત નેતાઓને ધમકીઓ આપી રહ ◌ યા છે કે તેઓ ચૂંટણીમાં \.;
BT-MT	**en** - FC	આતંકી સતત નેતાઓને ધમકીઓ આપી રહ ◌ે યા છે કે તેઓ ચૂંટણીમાં ભાગ ન લે .
	en - LC	આતંકી સતત નેતાઓને ધમકીઓ આપી રહ ◌ યા છે કે તેઓ ચૂંટણીમાં ભાગ ન લે .

Original sentence	**gu**	જેમાં સાત નેપાળી ગાઇડોના મોત નિપજ ◌ે યાં હતા અને ઘણા ઘાયલ થયા હતા .
	en	In which seven Nepalese guides were found dead and many were injured .
BT	**en** - FC	In addition to નેપાળી ગાઇડોના મોત નિપજ ◌ે યાં હતા and ઘણા ઘાયલ થયા હતા .
	en - LC	જેમાં સાત નેપાળી ગાઇડોના મોત નિપજ ◌ યાં હતા અને ઘણા ઘાયલ થયા હતા .
MT	**en** - FC	સાત નેપાળી ગાઇડોના મોત નિપજ ◌ે યાં હતા અને ઘણા ઘાયલ થયા.
	en - LC	સાત નેપાળી ગાઇડોના મોત નિપજ ◌ે યાં હતા ઘાયલ થયા હતા .;
BT-MT	**en** - FC	In the past , the સાત નપાળા ગાઇડાના માત નંપજ ◌ે યા હતા and many ઘાયલ થયા હતા
	en - LC	જેમાં સાત નેપાળી ગાઇડોના મોત નિપજ ◌ે યાં હતા અને ઘણા ઘાયલ થયા હતા .

Table 4: Two examples of a Gujarati sentence, their English ground-truth translation, and their English translations with each model's first and last checkpoint (**en** - FC, **en** - LC respectively)

Original (En)	despite the measures taken by developing countries to enhance the dissemination of information and strengthen regulations and surveillance of financial markets , they remained extremely vulnerable to economic crises and international macroeconomic cycles .
Misspelling	despute the measures taken by developing countries to enhance the dissemination of information and strengthen regulations and surveillance of financial markets , they remained extremely vulnerable to economic crises and international macroeconomic cycles .
Case-changing	Despite The Measures Taken By Developing Countries To Enhance The Dissemination Of Information And Strengthen Regulations And Surveillance Of Financial Markets , They Remained Extremely Vulnerable To Economic Crises And International Macroeconomic Cycles .
Original (Fr)	il a ete souligne que , par souci de coherence , le libelle de cet alinea devrait etre aligne sur celui du paragraphe 4 concernant la constitution d' une garantie dans le contexte des mesures provisoires inter partes , si ce n' est que les mots " peut faire obligation " pourraient etre remplaces par les mots " fait obligation " .
Misspelling	l a ete souligne que , par souci de coherence , le libelle de cet alinea devrait etre aligne sur celui du paragraphe 4 concernant la constitution d ' une garantie dans le contexte des mesures provisoires inter partes , si ce n ' est que les mots " peut faire obligation " pourraient etre remplaces par les mots " fait obligation " .
Case-changing	IL A ETE SOULIGNE QUE , PAR SOUCI DE COHERENCE , LE LIBELLE DE CET ALINEA DEVRAIT ETRE ALIGNE SUR CELUI DU PARAGRAPHE 4 CONCERNANT LA CONSTITUTION D' UNE GARANTIE DANS LE CONTEXTE DES MESURES PROVISOIRES INTER PARTES , SI CE N' EST QUE LES MOTS " PEUT FAIRE OBLIGATION " POURRAIENT ETRE REMPLACES PAR LES MOTS " FAIT OBLIGATION " .
Original (Gu)	ભવ ય ગાધી ( ) એક ભારતીય ટલિવિઝન અભિનતા છ .
Misspelling	ભtવ ય ગા @ @ ધી ( ) એક ભારતીય ટ @ @ લિ @ @ વિઝન અભિન @ @ તા છ .
Case-changing	ભવ ય ગા@@ ધી ( ) એક ભારતીય ટ@@ લિ@@ વિઝન અભિન@@ તા છ .

Table 5: Examples of sentences in our test dataset, and their perturbed versions after misspelling and case-changing