Do noises break GRU Seq2Seq Translation Model?

Nguyen Quynh Anh
quynguyen@student.ethz.ch

Abstract

The project aims to assess the robustness of the Seq2Seq neural machine translation by testing trained model on data which had been added different types of noises. The research is thus focused on presenting these features in the setting of French-English translation. Two bi general categories of noises including character-level and word-level are added to the clean texts. Afterthat, built model is tested on the 'dirty' data. BLEU-4 is the metric used to evaluate the performance NMT model. The experiment reveals that word-level noises have less ability to break seq2seq model meanwhile randomizing character noises including Fully Random and Middle Random are two character-level noises whose the most negative affect to the model.

1 Introduction

Humans have surprisingly robust language processing systems that can easily overcome typos, misspellings, and the complete omission of letters when reading (Rawlinson, 2007). Moreover, this robustness extends to audio as well (Saberi and Perrott, 1999). The exact mechanisms and limitations of our understanding system are unknown. In stark contrast, neural machine translation (NMT) systems, despite their pervasive use, are immensely brittle compared to robustness of human's natural language processing ability (Belinkov and Bisk, 2018).

In this research, the robustness of seq2seq model is examined by training on noisy texts approach. The experiment shows that noisy texts in character-level effects more negatively to the robustness of seq2seq neural machine translation (NMT) systems compared to the influences of noisy text in word-level.

2 Related Works

Several related works were implemented to test the robustness of Machine Translation before. The source-target pairs of language might be different, natural noise and synthetic noise are usually used. Besides, researches were rarely designed to experiment on the noise at word-level.

Synthetic and Natural Noise Both Break Neural Machine Translation research (Belinkov and Bisk, 2018) was implemented on three different NMT systems with access to character information at different levels including *char2char*, *Nematus*, and *CharCNN*. The synthetic and natural noises are added are all in character level. They explored two approaches to increase model robustness: black box and white box. In this analysis, Seq2seq RNN (Sutskever et al., 2014) model is the unique model which is robustnedd tested. The black-box approach is applied so that the original model architecture is used without any modifying. Besides, natural noise are not considered as done by (Belinkov and Bisk, 2018) meanwhile synthetic noise in character-level and word-level are both added to the clean French dataset.

In the Robust Open-Vocabulary Translation from Visual Text Representations research (Salesky et al., 2021), natural noise from editing history of Reddit users are used meanwhile in this experiment, text are modified by adding seven type of synthetic noise in character-level and word-level. Besides, multiple language and visually similar characters for the languages such as Arabic or Korean are also considered meanwhile in this experiment, French-English corresponding to source-target language are exclusively experimented. Last but not least, testing Open-Vocabulary Translation robusness (Salesky et al., 2021) evaluates the quality of machine translation by using BLEU score as a main evaluation metric.

3 Dataset

The experiment is be conducted using a subset of approximately 30k sentences from *Europarl* dataset (Koehn, 2005), a parallel corpus of the European Parliament proceedings. By using the *Europarl* data, we would like not to only use a political data where precise translation is essential to its operations but also investigate how seq2seq model deals with long-range dependencies.

For the sake of measuring how good seq2seq machine translation model performs on corpus with different noises, the original text in French language will be modified using different approaches. Two main types of synthetic noise in character level and word level will be added to clean text to create new corpus.

Table 1 reveals the size of three dataset including training/validating/testing set. *Europarl* dataset is a parralel corpus so the size of French data had the same

Number of sentences		
Train	Validation	Test
29.824	996	987

Table 1: Statistics for the source-side of French to English parallel corpora.

size with the target data, English.

3.1 Character-level synthetic noise

Noise 1: Middle Random

Humans tend to easily process words where the first and last letters are in place, the internal letters being reordered (Keith Rayner, 2006). The new data with noise will be created by modifying all words in vanilla data which have at least 4 characters.

Noise 2: External Letters Swap

The second type of chosen noise is replacing clean words by swapping its external letters. In particular, the first and last letters of all words of more than four letters in our corpus are swapped.

Noise 3: Fully Random

Fully random noise is created by shuffling all letter of each words so that the order of all the letters in a word will be randomized.

Noise 4: Keyboard Typo

Keyboard typo is a common noise of digital users. Each letter in each word will be replaced with an adjacent key in the US-Keyboard.

3.2 Word-Level

Noise 5: Word Missing

We first implemented a word removal noise method. For each sentence, removed each word with a given frequency. We found that a 10% frequency of removal yielded a good trade-off.
We drop certain words in a sentence (e.g. The cat is the table).

Noise 6: Word Duplication

We also implemented a word duplication method. Each word in the sentence got duplicated with the same arbitrary chosen 10% frequency.
We double certain words in a sentence (e.g. The cat cat is on the table).

Noise 7: Phonetics Ambiguity

In many languages, and this is especially true in French, there are many words sharing the same or very similar pronunciation, homonyms. These are common candidates for human mistakes in writing and are interesting noise additions to study. To implement this, there is no other alternative than to manually create a dictionary for common homonyms in French. Using multiple resources, we thus created a list of around 200 homonyms. We then allowed for a maximum of 3 homonym replacement in each sentence.
We substitute words with similar phonetics ones. (e.g. there → their)

4 Methodology

Black-box training strategy, in which seq2seq model is trained without adjusting models' architectures, is applied. The original training set is replaced by different training data which had been added different type of noises. Exposing models to multiple types of noise during training benefits us in testing the robustness of the model.

4.1 Seq2seq Model

Seq2seq model, i.e. sequence to sequence models (Sutskever et al., 2014) is a special class of Recurrent Neural Network architectures which are used to solve complex Natural Language Processing tasks such as Machine Translation. In this research, a GRU Seq2Seq model is built for the French - English translation task. GRU has been chosen for the experiment in stead of LSTM because GRU can handle memory issues better and achieve similar performance compared to LSTM models. We use the GRU as a base model for both the encoder and the decoder. Seq2seq model includes 3 parts which are encoder, encoder-decoder bridge and decoder.

Encoder

Several GRU cells where each accepts a single element of the input sequence, collects information for that element and propagates it forward. The hidden states h_i are computed as the following:

$$h_t = f(W^{hh}h_{t-1} + W^{hx}x_t) \quad (1)$$

As shown in Figure 1, in order to calculate the final hidden state, weights are added to the previous hidden state h_{t-1} and the input vector x_t.

Encoder-Decoder Bridge

The hidden state resulted from the encoder part of the model capture the information for all input elements and acts as the initial hidden state of the decoder part of the model.

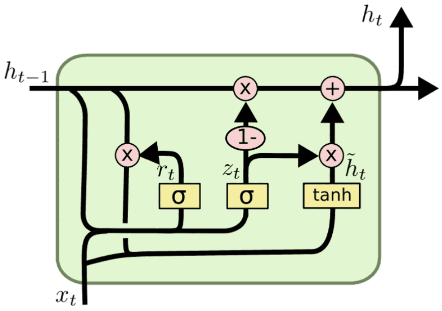


Figure 1: Encoder GRU

Decoder

Each GRU cell predicts an output y_t at a time step t. This recurrent unit will accept a hidden state from the previous unit and result an output y_t and an hidden state h_i. They are computed as following:

$$h_t = f(W^{hh}h_{t-1}) \quad (2)$$

$$y_t = softmax(W^S h_t) \quad (3)$$

The output y_t is calculated using the hidden state h_t and the corresponding weight W(S) and $Softmax$ function is applied to result a probability vector which will help to determine the final output.

4.2 Evaluation

In this research, we use BLEU score, i.e. Bilingual Evaluation Understudy score, to evaluate the translation quality of the model using data which is added different types of noises. In particular, the cumulative 4-gram BLEU score, also called BLEU-4, will be calculated for each case. A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0 (Koehn, 2009).

5 Results and Discussion

Table 2 shows BLEU-4 scores of models trained on Vanilla texts and tested on clean and noisy texts. In general, seq2seq model does not reach the good performance on translation tasks because the highest BLEU-4 scores is in the vicinity of 7. However, there are several insights that could be drawn from the experiment.

Level	Noise	BLEU-4 score
	Vanilla	6.702
Char Level	Middle Random	4.686
	Ext-Letters Swap	7.006
	Fully Random	4.168
	Keyboard Typo	4.943
Word Level	Word Missing	6.598
	Word Duplicate	7.045
	Phonetics Ambiguity	6.741

Table 2: The effect of synthetic noise on models trained on clean texts.

Seq2seq model is effected strongly by character-level types of noise. In particular, the model performs worst on the testing data with Middle Random and Fully Random noise. The BLEU-4 score of seq2seq model tested on them are respectively 4.6% and 4.1%. Being different from all other types of character level noise, External Letter Swap helps model reach a similar BLEU score as it is tested on Vanilla data.

In contrast, seq2seq translation model behaves robustly when it comes to data with word-level noises. In fact, the model uniformly achieve similar BLUE score on test set with added one of three word-level noises as shown in Table 2.

Besides, translation examples of the model using different testing data are also shown in Appendix A.

6 Conclusion

In conclusion, word-level noises are less likely to break seq2seq model robustness meanwhile randomizing character noises including Fully Random and Middle Random are two character-level noises whose the most negative affect to the model.

Future Works As shown in 2, seq2seq model was not well-performed in translation task even with vanilla data. For the future research, testing the robustness of other NMT models could be a good experiment. Also, we improve the research by adding noise with certain proportion in stead of fully-replaced noises as in this research. Besides, tuning better hyper-parameter and increasing the training data size could improve the model performance.

Data and Code

Github Repository | Data

Acknowledgements

This work benefited from discussions with exteammates.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation.

Rebecca L. Johnson Simon P. Liversedge Keith Rayner, Sarah J. White. 2006. Raeding wrods with jubmled lettres: There is a cost.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn. 2009. *Evaluation*, page 217–246. Cambridge University Press.

Graham Rawlinson. 2007. The significance of letter position in word recognition. *Aerospace and Electronic Systems Magazine, IEEE*, 22:26 – 27.

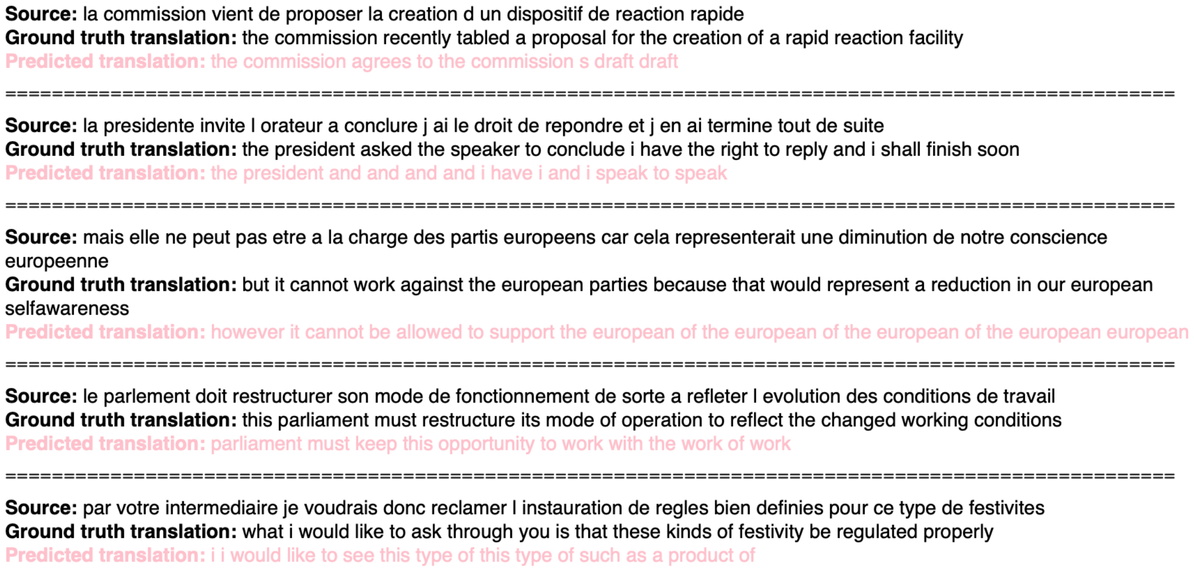
Kourosh Saberi and David R. Perrott. 1999. Cognitive restoration of reversed speech. *Nature*, 398:760–760.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

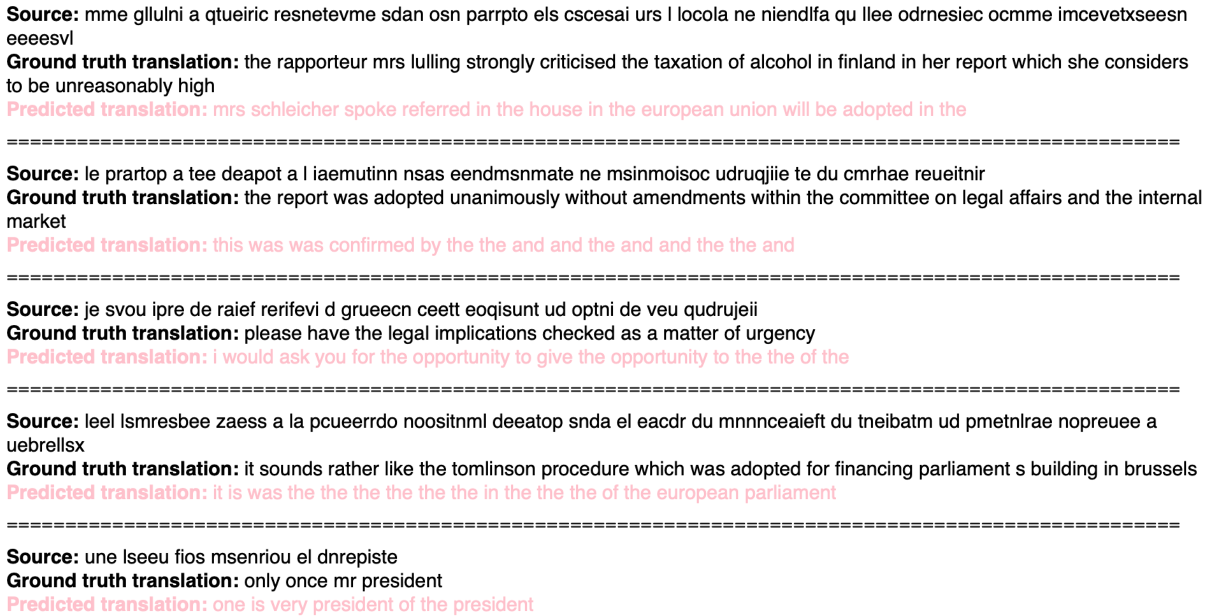
Appendix

Appendix A



```
Source: la commission vient de proposer la creation d un dispositif de reaction rapide
Ground truth translation: the commission recently tabled a proposal for the creation of a rapid reaction facility
Predicted translation: the commission agrees to the commission s draft draft
=====================================================================================================
Source: la presidente invite l orateur a conclure j ai le droit de repondre et j en ai termine tout de suite
Ground truth translation: the president asked the speaker to conclude i have the right to reply and i shall finish soon
Predicted translation: the president and and and and i have i and i speak to speak
=====================================================================================================
Source: mais elle ne peut pas etre a la charge des partis europeens car cela representerait une diminution de notre conscience
europeenne
Ground truth translation: but it cannot work against the european parties because that would represent a reduction in our european
selfawareness
Predicted translation: however it cannot be allowed to support the european of the european of the european of the european european
=====================================================================================================
Source: le parlement doit restructurer son mode de fonctionnement de sorte a refleter l evolution des conditions de travail
Ground truth translation: this parliament must restructure its mode of operation to reflect the changed working conditions
Predicted translation: parliament must keep this opportunity to work with the work of work
=====================================================================================================
Source: par votre intermediaire je voudrais donc reclamer l instauration de regles bien definies pour ce type de festivites
Ground truth translation: what i would like to ask through you is that these kinds of festivity be regulated properly
Predicted translation: i i would like to see this type of this type of such as a product of
```

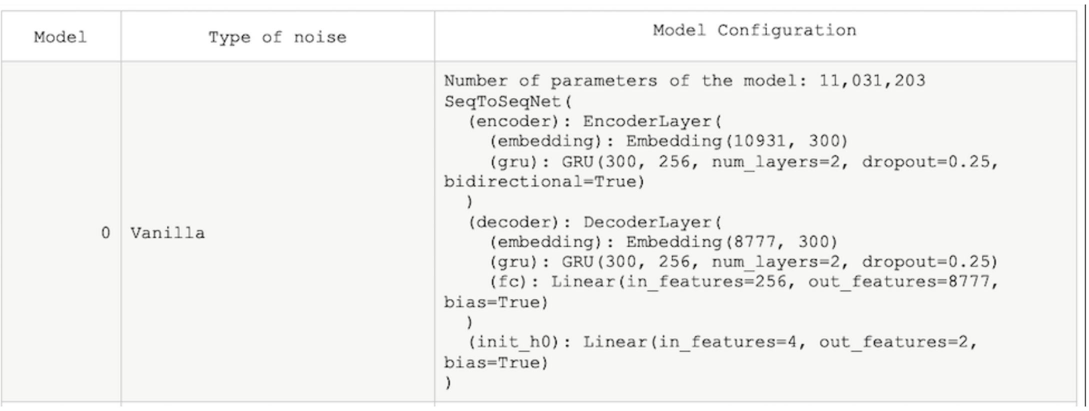
Figure 2: Translation example from model trained ans tested with vanilla data.



```
Source: mme gllulni a qtueiric resnetevme sdan osn parrpto els cscesai urs l locola ne niendlfa qu llee odrnesiec ocmme imcevetxseesn
eeeesvl
Ground truth translation: the rapporteur mrs lulling strongly criticised the taxation of alcohol in finland in her report which she considers
to be unreasonably high
Predicted translation: mrs schleicher spoke referred in the house in the european union will be adopted in the
=====================================================================================================
Source: le prartop a tee deapot a l iaemutinn nsas eendmsnmate ne msinmoisoc udruqjiie te du cmrhae reueitnir
Ground truth translation: the report was adopted unanimously without amendments within the committee on legal affairs and the internal
market
Predicted translation: this was was confirmed by the the and and the and and the the and
=====================================================================================================
Source: je svou ipre de raief rerifevi d grueecn ceett eoqisunt ud optni de veu qudrujeii
Ground truth translation: please have the legal implications checked as a matter of urgency
Predicted translation: i would ask you for the opportunity to give the opportunity to the the of the
=====================================================================================================
Source: leel lsmresbee zaess a la pcueerrdo noositnml deeatop snda el eacdr du mnnnceaieft du tneibatm ud pmetnlrae nopreuee a
uebrellsx
Ground truth translation: it sounds rather like the tomlinson procedure which was adopted for financing parliament s building in brussels
Predicted translation: it is was the the the the the the in the the the of the european parliament
=====================================================================================================
Source: une lseeu fios msenriou el dnrepiste
Ground truth translation: only once mr president
Predicted translation: one is very president of the president
```

Figure 3: Translation results by testing model on data whose characters are fully randomized.

Appendix B



Model	Type of noise	Model Configuration
0	Vanilla	Number of parameters of the model: 11,031,203 SeqToSeqNet((encoder): EncoderLayer((embedding): Embedding(10931, 300) (gru): GRU(300, 256, num_layers=2, dropout=0.25, bidirectional=True)) (decoder): DecoderLayer((embedding): Embedding(8777, 300) (gru): GRU(300, 256, num_layers=2, dropout=0.25) (fc): Linear(in_features=256, out_features=8777, bias=True)) (init_h0): Linear(in_features=4, out_features=2, bias=True))

Figure 4: Model’s Architecture

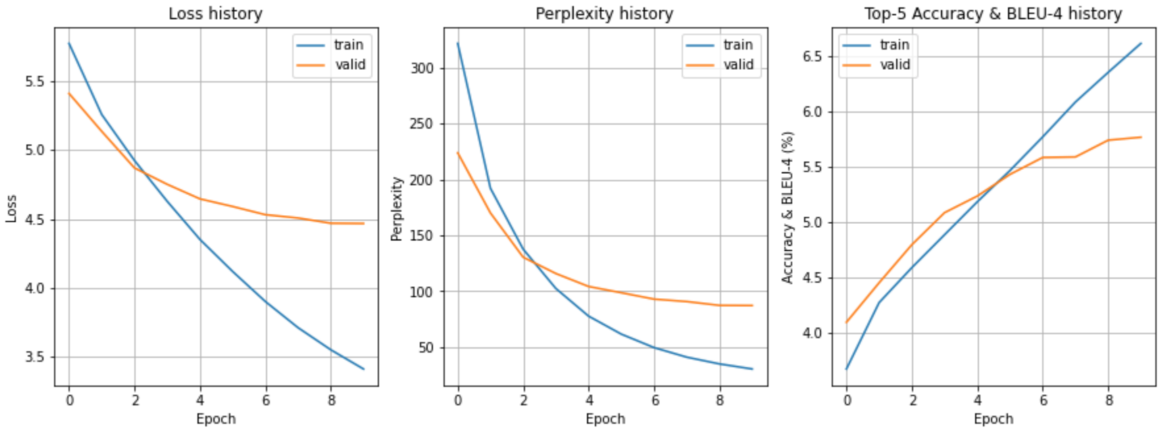


Figure 5: Loss - Perplexity - Blue-4 score results on vanilla data.