Adversarial Stability Training in Neural Machine Translation of Chinese-to-English Text

Mandy Lu
mlu355@stanford.edu

Kaylie Zhu
kayliez@stanford.edu

Abstract

A major challenge for neural machine translation (NMT) models is truly understanding semantics, creating common issues such as difficulty parsing long sentences and different performance on semantically similar inputs (8). This is especially relevant to Chinese-to-English translation because Chinese contains many synonyms and a relative lack of word order rules, resulting in a wide variety of ways to express sentiments. Traditional techniques to combat this problem incorporate specific domain knowledge or hand-crafted features and models, which can be time-consuming and difficult to generalize. We investigate the effectiveness of adversarial stability training in improving NMT robustness and decreasing the time manually analyzing linguistic features. This technique adds Gaussian noise to training data and incorporates a discriminator which encourages encoding outputs from noisy and regular data to be indistinguishable. We build a hybrid and more robust NMT model by implementing the adversarial stability training framework and combining it with the state-of-the-art Transformer model. To achieve this, we build upon Google's Transformer model and redesign the framework to incorporate a new CNN discriminator model that we implement, as well as a hybrid, joint loss function we create to facilitate similar behavior from encoder and decoder on an original input and its perturbed counterpart. Through a number of experiments and ablation studies conducted, we are thrilled to see that our model becomes more robust towards perturbations in data and model performance is significantly improved.

1 Introduction

One of the major problems in neural machine translation (NMT) tasks, and natural language processing in general, is a machine's inability to truly understand semantics. This leads to a variety of known issues, including difficulty parsing long sentences or under-performing on some phrases compared to other semantically similar ones due to lack of semantic understanding (8). In this project, our main goal is to explore new techniques to make neural machine translations from Chinese to English more robust. In particular, we seek to explore how the range of a NMT model's semantic mapping can be increased such that source texts with similar semantic meaning are mapped to similar, correct translations. This is an especially relevant problem to Chinese-to-English translation because Chinese contains a high amount of synonymous words and phrases, as a result of which many NMT models struggle with even the slightest rephrasing in Chinese inputs. In addition, the relative lack of spacial grammatical constructs in Chinese leads to a wide variety of possible ways to express the same sentiment, presenting additional difficulties for methods such as beam search encoding (1).

In this context, we introduce our adversarial stability training integrated, Transformer-based hybrid model, where we built upon the state of the art Transformer Model, on which we applied a adversarial stability training framework we implemented from scratch. The framework uses a hybrid loss function that encourages the encoder to produce similar embeddings for both the normal and perturbed inputs, and the decoder to translate to the same robust output when given these embeddings. It also makes use

Mentor: Abigail See

of a discriminator that tries to distinguish between the normal and perturbed inputs, whilst the rest of framework strives to fool the discriminator. We ran a number of experiments to test the performance of our model, using a specifically-chosen highly diverse, noisy and challenging dataset, and also conducted ablation study to test the efficacy of different components of our new hybrid model. We are elated to see that our new hybrid framework significantly improves translation performance, and the model exhibits more robustness and perturbation invariance in its deployment.

2 Related Work

Various methods have been implemented to combat current challenges in neural machine translation. Introduction of the encoder-decoder paradigm and neural models saw huge improvement over previous SMT models. Sutskever et al. and Cho et al. introduced seq2seq and the use of RNN and LSTM for NMT, which became the new gold standard (13) (3). In recent years, Sennrich and Haddow used linguistic features such as lemmas, part-of-speech tags, syntactic dependency labels and morphological features to enrich NMT input units (11). On the other hand, Garcia-Martinez et al proposed factored NMT by decomposing words with morphological and grammatical techniques in output units (4). In order to incorporate domain understanding into model architecture, Zhang et al. integrated topic knowledge into NMT for domain/topic adaptation (9). Most recently, Vaswani et al proposed a Transformer Model based on replacing RNN layers with multi-head attention modules which has outperformed most NMT models with respect to BLEU score (14). Convolutional sequence to sequence models, termed Fairseq, have seen similar success by using convolutional neural networks instead of recurrent neural networks in their encoder and decoder. (5).

Many of these techniques rely on incorporating specific domain knowledge or hand-crafted features or models into a traditional encoder-decoder architecture, which can be time-consuming and difficult to generalize. Additionally, many of these current methods struggle when given even the slightest perturbed inputs, and perform very poorly in such cases, unable to extract the similar semantic meanings behind the minor changes in wording inputs. These problems are addressed in and tackled by a variety of research works that inspired our work. For instance, Wu et al employs an adversarial NMT model that aims minimize the distinction between human translation and the translation given by an NMT model using an CNN adversary elaborately designed to differentiate the translation result generated by the NMT model from that by humans (15). Similarly, a conditional sequence generative adversarial net comprised of two adversarial sub models, a generator and a discriminator, is proposed by Yang et al to similar ends. Remarkably, Cheng et al proposes an adversarial framework to enhance the robustness of NMT model training, which aims to produce robust output that is invariant to small changes in inputs, by building into the hybrid model a discriminator as well as hybrid loss functions. (2). Inspired by these works, we investigate the application of adversarial stability training, by using Gaussian noise to develop robustness and allowing for less time manually analyzing linguistic features for each task. In particular, we were motivated to apply adversarial stability training to a modern architectures model, the self-attention Transformer network, which is consistently one of the best performing models for most NMT tasks (12).

3 Methodology

3.1 Task definition

We aim to produce a more robust, hybrid NMT model by incorporating adversarial stability training and implementing a modified version of the technique proposed in the research paper "Towards Robust Neural Machine Translation" by Cheng et al. (2) This proposed technique is model agnostic and applies to any NMT architecture with an encoder and decoder. In the paper, the authors applied their technique to a Gated RNN with 2 layers, showing that adversarial stability training (AST) with feature-level perturbations (see Figure 1) achieved the highest performances across all different datasets. This motivates us to apply AST to the state-of-the-art attention-based Transformer Model (14) in order to tackle deep-rooted challenges in Chinese-English translation, like its susceptibility to high variances and poorer performance when given slightly altered inputs. We sought to observe the application of the AST technique to a broad domain spanning many contexts in which it is crucial to be able to generalize and maintain robustness over noise, specifically unstructured web text and movie subtitles.

Source	zhongguo dianzi yinhang yewu guanli xingui jiangyu sanyue yiri qi shixing
Reference	china's new management rules for e-banking operations to take effect on march 1
MLE	china's electronic bank rules to be implemented on march 1
$AST_{lexical}$	new rules for business administration of china 's electronic banking industry will come into effect on march 1 .
$AST_{feature}$	new rules for business management of china 's electronic banking industry to come into effect on march 1
Perturbed Source	*zhong**fang*** dianzi yinhang yewu guanli xingui jiangyu sanyue yiri qi shixing
MLE	china to implement new regulations on business management
$AST_{lexical}$	the new regulations for the business administrations of the chinese electronics bank will come into effect on march 1 .
$AST_{feature}$	new rules for business management of china's electronic banking industry to come into effect on march 1

Figure 1: Example inputs and outputs. "zhongguo" is perturbed to "zhongfang", which acts as a synonym in this context. AST refers to adversarial stability training and MLE refers to a baseline model based on parameters trained by Maximum Likelihood Estimation

3.2 Dataset

We are using a dataset of 2 million parallel Chinese-English sentence pairs collected and shared by the China Workshop on Machine Translation (CWMT) community. Specifically, we are using their casict2015 corpus (ICT Web Chinese-English Parallel Corpus) provided by the Institute of Computing Technology, Chinese Academy of Sciences. The corpus contains about 2 million sentences pairs of sentences collected from from the web (60%), from, movie subtitles (20%), and from an English/Chinese thesaurus (20%). The sentence level alignment precision was reported to be higher than 99%. There are 22,802,353 total words and 435,010 distinct words (English). The casict2015 dataset is a subset of the CWMT (China Workshop on Machine Translation) Chinese-English parallel translations and was chosen due to its wide range of contexts, which cover a comprehensive overview of human communication, from biblical texts to colloquial speech. In addition, for data preprocessing we segmented the Chinese inputs by building a segmentation system base on the Jieba framework(6).

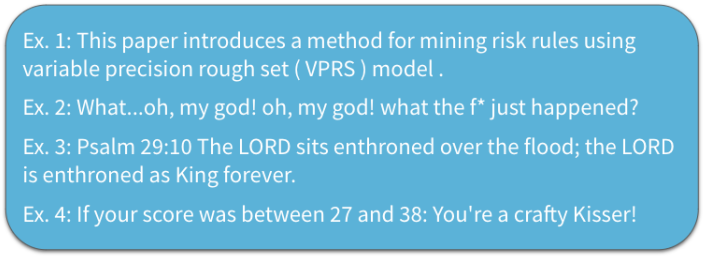


Figure 2: Examples of the diversity in the casict2015 dataset

3.3 Baseline

For our baseline, we implemented, trained and evaluated NMT performance on our dataset with the self-attention Transformer Model as is described in detail in the paper by Vaswani et al, preserving all listed parameters in Section 4.1 (14). We compared our performance to this Transformer baseline and results from the University of Cambridge submission WMT2018, which contained results on Chinese-English translation for several model architectures trained primarily on the CWMT parallel Chinese-English translations which include the casict2015 dataset (12).

3.4 Model and Techniques

We try to create a hybrid adversarial stability training model by combining the state-of-the-art Transformer model with an adversarial framework inspired by and developed based on Tencent AI's "Towards robust neural machine translation" (2). We draw inspirations from the paper written by researchers in the Tencent AI Lab which proposes a novel way to improve the robustness of NMT models using adversarial stability training. Essentially, by facilitating both the encoder and decoder in NMT models to behave similarly when given the original input and its perturbed counterpart, we aim to make the NMT model itself more robust against input perturbations. We are inspired by their work to apply a modified, hybrid version of adversarial stability training framework to new NMT

Link to dataset: http://nlp.nju.edu.cn/cwmt-wmt/

model architectures in an effort to make model performance on Chinese-English translation both better and more invariant towards input changes.

More specifically, we learn the perturbation-invariant NMT encoder and decoder in the methods proposed by Cheng et al. Given a minibatch of source sentence x, we construct a minibatch of perturbed sentences x' by obtaining the word embeddings for each word and adding random Gaussian noise to all word embeddings to simulate various types of feature-level perturbations.

$$E[x_i'] = E[x_i] + \varepsilon, \varepsilon \sim N(0, \sigma^2 I) \tag{1}$$

where ε is a vector sampled from Gaussian distribution with variance σ^2 (see equation 1), σ being a hyperparameter. As depicted in Figure 3, the architecture entails an input tuple of (x, x'), which correspondingly yields encoded representations H_x and $H_{x'}$. The aim is to make H_x and $H_{x'}$ as close as possible, while also enabling the decoder to generate the robust output y given $H_{x'}$.

We also implement from scratch an additional discriminator D, in an effort to distinguish the representations of perturbed input $H_{x'}$ from that of original input H_x. Whilst the encoder (G) aims to produce similar embeddings representing x and x' in order to fool the discriminator, the discriminator (D) attempts to correctly distinguish between the two. More specifically, D outputs a classification score when given an input representation, trying to maximize D(G(x)) to 1 and minimize D(G(x')) to 0. For this purpose, we build a binary CNN discriminator with a convolutional layer, max-pooling layer, followed by highway layer, dropout layer and a final sigmoid layer.

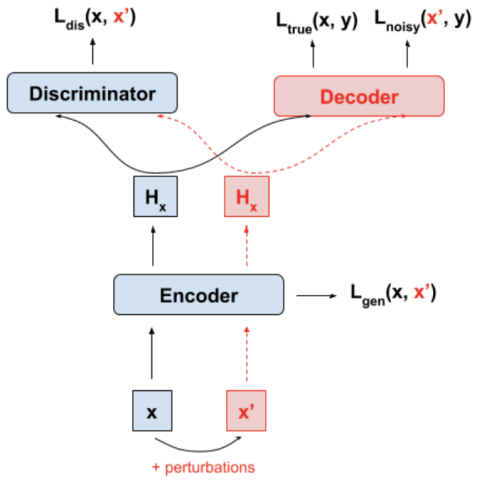


Figure 3: Architecture of NMT with adversarial stability training

3.5 Loss function

In order to achieve this series of objectives, we use the following losses, as defined by Cheng et al, all of which we will combine into creating a hybrid loss function. We aim to encourage the encoder to produce similar intermediate embeddings H_x and $H_{x'}$ when encoding the representations given x and x'. Formally, the perturbation-invariant encoder serves as a generator G, which defines the policy producing a sequence of hidden representations H_x given input sentence x. We also introduce a discriminator D as described in the previous section. Following the adversarial learning framework, the discriminator tries to maximize D(G(x)) to 1 and minimize D(G(x')) to 0, by minimizing the loss as described in the following with equation 2.

$$\mathscr{L}_{dis}(x, x'; \theta_{enc}, \theta_{dis}) = E_{x \sim S}[-logD(G(x))] + E_{x' \sim \mathscr{N}(x)}[-log(1-D)G)x'))))] \tag{2}$$

Now we further define $\mathscr{L}_{noisy}(x', y)$ to encourage the decoder to produce the robust output y when given the noisy input x', which is modeled as $-logP(y|x')$. In addition, we also keep the original training objective $\mathscr{L}(x, y)$, guaranteeing a good translation performance while targeting translation stability. These two objectives can be calculated using equation 3. (note that one may simply directly substitute x' for x when calculating $\mathscr{L}_{noisy}(x', y)$). For our model, we are using the label smoothing loss (see more details in section 4.3) as utilized in the Transformer paper (14) for both the true and

noisy losses.

$$\mathscr{L}(x,y;\theta) = \sum_{(x,y)\in S} -logP(y|x;\theta) \tag{3}$$

To incorporate the discriminator's classification output of $H_{x'}$, we define an additional loss

$$\mathscr{L}_{inv}(x,x';\theta_{enc},\theta_{dis}) = E_{x'\sim\mathscr{N}(x)}[-log(1-D)G)x'))] \tag{4}$$

We then integrate the three training losses $\mathscr{L}_{true}$, $\mathscr{L}_{inv}$, and $\mathscr{L}_{noisy}$into one overall loss function as the formal adversarial stability training objective as shown in equation 4.

$$\mathscr{J}(\theta) = \sum_{(x,y)\in\mathscr{S}}(\mathscr{L}_{true}(x,y;\theta_{enc},\theta_{dec}) - \alpha\mathscr{L}_{gen}(x,x';\theta_{enc},\theta_{dis}) + \beta\mathscr{L}_{noisy}(x',y;\theta_{enc},\theta_{dec}) \tag{5}$$

and we back propagate the discriminator loss $\mathscr{L}_{dis}$ separately and independently from this hybrid loss function. Note that $\{\theta_{enc},\theta_{dec},\theta_{dis}\}$ represent trainable parameters corresponding to the encoder, decoder and discriminator introduced previously. We aim to jointly minimize not only translation loss $\mathscr{L}_{true}$ but also the invariance losses $\mathscr{L}_{noisy}$ and $\mathscr{L}_{inv}$. For higher efficiency, Cheng updates both encoder and discriminator simultaneously at each iteration rather than the standard periodical training strategy in adversarial learning with a modified generator loss which is equal to negative L_{dis}. In order to evaluate the effect of this simplification on accuracy, we implemented both AST with traditional GAN loss and simplified generator loss, and performed ablation tests for comparison.

3.6 Training

We use mini-batch SGD to optimize the model. We adapt the framework we described to Google's Transformer model, a possibility enabled by the inherent versatility of the framework. In a typical forward pass, we use both a minibatch of x and y and a minibatch of perturbed inputs x' and y. We propagate both batches through the encoder, as shown in Figure 3, then pass both embeddings H_x and H'_x through both the discriminator and the decoder. We then calculate the three loss functions as defined in the previous section and jointly minimize them in an integrated manner. We collect all gradients in order to update all three sets of model parameters. Note that we back propagate the discriminator loss $\mathscr{L}_{dis}$ separately. Also, whilst other gradients are back-propagated as usual, the gradient of $\mathscr{L}_{inv}$ with respect to θ_{enc} is multiplied by -1.

4 Experiments

4.1 Experimental Details

We performed two ablation studies in order to measure the individual effectiveness of sub-parts of our training and model architecture. In the first study, we removed the GAN structure from the training framework by removing the discriminator from the model architecture and $Loss_{gen}$ from the objective function and loss function; thus we reduce to a data augmentation technique of simply training on both original and noisy data. In the second study, we implemented the original simplified GAN loss described by Cheng. Ablation studies were performed on smaller datasets of 250,000 sentences and we also trained a Transformer model with our modified AST on this dataset as a control for fair comparison.

For our main study, we first configured a baseline Transformer model following the guidelines detailed in Vaswani et al's "Attention is All you Need" (14) and configured our parameters to match those described in the paper. These parameters reproduced their WMT results as shown by OpenNMT researchers, although no specific metrics were provided, and we used the OpenNMT framework to develop our baseline model (7). We split our dataset into train, development, and test sets with a 20% of samples in development and 5% in train. The model configurations for the Transformer network include $N = 6$ identical layers for the encoder and 6 layers for the decoder, 2048 for the size of the hidden feed-forward in the Transformer, 8 heads, scaled-dot self attention type, and word embedding sizes of 512. The hyperparameters were set with an initial learning rate of 2, and a learning rate decay of 0.5 after 8000 epochs with noam decay. Adam optimization was used with beta1 of 0.9 and beta2 of 0.998. The model was run for 200,000 epochs, batch size was 4096 and dropout was set to 0.1. For our loss function, we used label-smoothing loss which minimizes the KL-divergence between the smoothed ground truth probability of a translation and the probability of the translation computed by the model, where the label smoothing value was $\varepsilon = 0.1$, and the probabilities of all non-true labels

are smoothed by $\varepsilon/((vocabsize)-1)$. Label smoothing loss was used to relax our confidence on the labels, to improve model performance where there may be noise in data input labels.
Four models were evaluated on the whole dataset, the default Transformer model described above, AST on Transformer with embeddings trained from scratch, AST on transformer with embeddings from the Chinese Wikipedia of dimension 300 and trained on 223 million tokens, and AST on Transformer with reduced batch size of 1024. (10). Due to resource and time constraints because each Transformer model takes up to 2 days to train, these four models were chosen to observe the effect of AST, batch size, and embeddings on the model.

4.2 Evaluation Methods

Both quantitative and qualitative methods were used to evaluate results. Quantitive metrics included BLEU (Bilingual Evaluation Understudy Score), perplexity and accuracy, defined as the percent of words in the NMT translation which match the source text. These metrics were used due to their effectiveness, simplicity, and ease of comparison to previous studies, including Cambridge WMT 2018. The model with lowest perplexity on the validation set was selected to produce translations to evaluate with BLEU.

However, since quantitative metrics alone are insufficient to evaluate our task performance, we performed qualitative analysis on the model. First, we manually created a set of perturbed source sentences by randomly selecting 10 source sentences and writing a semantically similar counterpart for each one. Then we produced translations on perturbed input for our baseline and final models and compared them to the translations of the original source sentences. Finally, we qualitatively compared NMT translations with the correct human translations to provide a more holistic understanding of our NMT performance.

5 Results

5.1 Quantitative Results and Analysis

5.1.1 Ablation Tests

Results for ablation studies are shown in (see Table 3).

Evaluation metric	No Discriminator	Simplified Loss	Modified AST
BLEU	4.46	4.64	4.67
Perplexity	3.14	3.06	2.43

Table 1: Ablation Studies

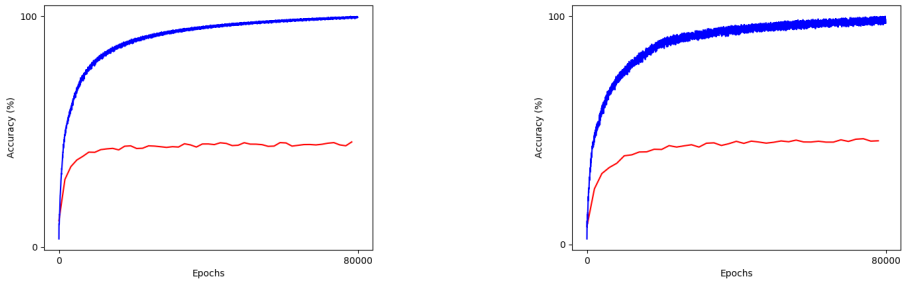
Modified AST had the highest BLEU score (4.67), followed closely by AST with simplified loss (4.64) and then the framework with no discriminator (4.46). Modified AST also had the lowest perplexity (2.43) which was significantly higher than the other two models: AST Simplified Loss (3.06) and No Discriminator (3.14).

5.1.2 Model Performance Comparison

We trained and evaluated four Transformer models trained on the whole dataset: baseline model, AST trained from scratch with random embeddings, AST Transformer with pre-trained embeddings and reduced batch size 1024, and AST Transformer with pre-trained embeddings with batch size 4096. A summary of our results are shown in Table 2. Out of the models we developed, AST Embed had the highest BLEU score (18.42) followed by AST Random Embed (18.23), AST Reduced Batch Embed (17.89) and the Transformer baseline (16.77). The order of lowest to highest perplexity is almost the same: AST Embed (1.23) is followed by AST Random Embed (1.44), the Transformer baseline (1.79) and then AST Reduced Batch Embed (2.86). The Cambridge WMT transformer model was our oracle and had higher BLEU at 25.6 with no reported perplexity.
Additionally, we plotted train and validation accuracy on our best model, AST Embed Transformer, and the Transformer baseline for comparison (Figure 4).

Metric	Tr Baseline	AST Reduced Batch Embed	AST Random Embed Tr	AST Embed Tr
BLEU	16.77	17.89	18.23	18.42
Text Perplexity	1.79	2.86	1.44	1.23

Table 2: Summary of Model Comparison. Transformer is (Tr) and AST Embed indicates AST with pretrained embeddings



Transformer Baseline AST Embed Transformer

Figure 4: Plots of train and val accuracy for Transformer Baseline (left) and AST Embed Transformer (right)

Metric	AST Embed Tr	Transformer Baseline
BLEU	30.7	27.9

Table 3: BLEU scores for perturbed outputs on AST Embed and Baseline

5.1.3 Perturbed Input Translations

BLEU scores were calculated for translations of the 10 perturbed source sentences by our best model, AST Embed Tr, and the Transformer Baseline. AST Embed had a much higher BLEU score at 30.7 and Transformer Baseline at 27.9.

5.2 Qualitative Analysis

Here we provide an example (see Figure 5) of running our hybrid model on pairs of original and perturbed inputs specifically targetted to confused standard NMT models, in an effort to demonstrate the robustness and perturbation-invariance of our model.

True text: 我吃了汤姆的三明治。
Perturbed Text: 我食用了汤姆的三明治。
Reference for true text: I ate Tom's sandwich.
Translations of perturbed text:
Transformer Baseline: I utilized Tom's sandwich.
Transformer + AST Embed: I ate Tom's sandwich.

Figure 5: Example of model performance on Chinese original and perturbed input pair

In this example, the perturbed text uses the less common word "食用" rather than the usual "吃" to express the verb "eat". Notably, the character "用" in "食用" also means "to use" or "utilise", hence as we see, the basic Transformer is confused by the perturbation and mistakenly translates it as "I utlised Tom's sandwich". Our hybrid model mechanism, conversely, is able to pick up the subtleties and semantics in the perturbed input, and correctedly translates to "I ate Tom's sandwich".

6 Discussion

In this study, we demonstrated the efficacy of the adversarial stability training technique proposed by Cheng et al applied to the Transformer model. To the best of our knowledge, this is the first study investigating the performance of AST on the Transformer network. We also report the effect of

various sub-components of the AST training framework which we measured through ablation testing. Finally, we performed qualitative analysis by examining translations manually and evaluating our best model and the baseline on a small set of perturbed outputs.

To measure individual effectiveness of sub-parts of our framework, we (1) removed the GAN structure from the training framework, reducing to a data augmentation problem of simply training on both original and noisy data and (2) implemented the original simplified generator loss L_{inv} as negative L_{dis}. Modified AST with a traditional GAN loss where discriminator and generator are propagated separately has the highest BLEU score (4.67) and lowest perplexity (2.43), which means that that the simplication of the GAN loss proposed by Cheng may result in a performance decrease (BLEU 4.64 Perplexity 3.06), but more experiments need to be done to verify this. The gap between Modified AST and the AST model with no discriminator was higher (BLEU 3.06, Perplexity 3.14), suggesting that the adversarial training objective is indeed making the NMT models more robust by encouraging intermediate embeddings of semantically similar outputs to be similar.
During more in-depth evaluation of various AST models on the whole dataset, we found that AST Embed was our best model, with a BLEU score of 18.42 followed and perplexity of 1.23. AST Embed improved over our Transformer baseline (BLEU 16.77 Perplexity 1.79) by 1.65 BLEU points and 0.56 perplexity, indicating that adversarial stability training is an effective technique to improve the performance and robustness of Transformer NMT models. Manual evaluation of the translation sentences corroborated these results, as the translation sentences were primarily comprehensible and well-formed even if the number of identical words to the reference text was low. This is likely due to the diversity and noisiness of our dataset, which includes primarily randomized web crawls and movie subtitles, leading to a very diverse range of possibilities for translations and a lower likelihood of matching reference words exactly. Chinese to English neural machine translation has also been one of the most challenging NMT tasks due to the different alphabet in Chinese and English as well as the high number of possible ways to segment Chinese text and the frequency of Chinese synonyms. Our best model's performance was lower than Cambridge's WMT 2018 submission who also trained on the casict2015 dataset, but they used a training dataset size of almost 50 million sentence pairings which included the larger CWMT dataset along with 16 GPUs and various ensembling methods, while we only used 2 million sentence pairings and a single GPU. The performance of the reduced batch embed (BLEU 18.23, Perplexity 2.86) confirms the common consensus that large batch size (in our case 4096) is crucial for Transformer performance. In addition, the random embeddings performed only slightly worse from BLEU (18.23) and perplexity (1.44), suggesting that the dataset is either large enough to accommodate learning embeddings from scratch or that our embeddings could perhaps be changed to be more effective. This is likely since the percent of our segmented token vocabulary with Chinese word embeddings was less than 20 percent.

Finally, the set of 10 perturbed source sentences evaluated by our best model, AST Embed Tr, and the Transformer Baseline (BLEU 30.7) showed that AST Embed had both a higher BLEU than the baseline (BLEU 27.9), and that both BLEU scores were significantly higher than the performance on the random test set. This is likely because the test set is so diverse and the sentences we selected were relatively short and straightforward with commmon words. Qualitatively, we also see a significant improvment in robustness visually as we evaluate the model's translations of the perturbed inputs targetted to confused most NMT models. Our hybrid model's ability to understand the subtleties and capture the semantic meanings behind the confusing word choice shows successful performance and manifests promising results that can be well extended to other research applications of adversarial learning in neural machine translation.

7 Conclusion and Future Work

After a careful analysis of our results, we conclude that AST can be an effective method to develop perturbation-robust Chinese-English NMT models. We showed that AST works performs well with Transformer network on diverse datasets. Furthermore, ablation studies demonstrate the efficacy of some components constituting our hybrid model: namely, the GAN component and our modified objective loss. Future work could include: (1) investigating techniques to handle rare words such as byte-pair encoding (2) using weak supervision to label semantically similar train text for AST input (3) more hyperparameter tuning to achieve better performance with more time and computational resources (4) evaluating which segmentation and embedding schema works the best.

References

[1] BRAZILL, S. Chinese to english translation: Identifying problems and providing solutions.

[2] CHENG, Y., TU, Z., MENG, F., ZHAI, J., AND LIU, Y. Towards robust neural machine translation. *CoRR abs/1805.06130* (2018).

[3] CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H., AND BENGIO, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[4] GARCÍA-MARTÍNEZ, M., BARRAULT, L., AND BOUGARES, F. Factored neural machine translation architectures. In *International Workshop on Spoken Language Translation (IWSLT'16)* (2016).

[5] GEHRING, J., AULI, M., GRANGIER, D., YARATS, D., AND DAUPHIN, Y. N. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122* (2017).

[6] JUNYI, S. "jieba" (chinese for "to stutter") : Chinese text segmentation: built to be the best python chinese word segmentation module. `https://https://github.com/fxsjy/jieba`, 2013.

[7] KLEIN, G., KIM, Y., DENG, Y., SENELLART, J., AND RUSH, A. M. Opennmt: Open-source toolkit for neural machine translation. *CoRR abs/1701.02810* (2017).

[8] KOEHN, P., AND KNOWLES, R. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation* (2017), Association for Computational Linguistics, pp. 28–39.

[9] LI, J., XIONG, D., TU, Z., ZHU, M., ZHANG, M., AND ZHOU, G. Modeling source syntax for neural machine translation. *arXiv preprint arXiv:1705.01020* (2017).

[10] LI, S., ZHAO, Z., HU, R., LI, W., LIU, T., AND DU, X. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2018), Association for Computational Linguistics, pp. 138–143.

[11] SENNRICH, R., AND HADDOW, B. Linguistic input features improve neural machine translation. *arXiv preprint arXiv:1606.02892* (2016).

[12] STAHLBERG, F., DE GISPERT, A., AND BYRNE, B. The university of cambridge's machine translation systems for WMT18. *CoRR abs/1808.09465* (2018).

[13] SUTSKEVER, I., VINYALS, O., AND LE, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (2014), pp. 3104–3112.

[14] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is all you need. In *Advances in Neural Information Processing Systems* (2017), pp. 5998–6008.

[15] WU, L., XIA, Y., ZHAO, L., TIAN, F., QIN, T., LAI, J., AND LIU, T. Adversarial neural machine translation. *CoRR abs/1704.06933* (2017).