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


**Mandy Lu** — mlu355@stanford.edu

**Kaylie Zhu** — kayliez@stanford.edu


## Motivation

- Major challenge for neural machine translation (NMT) models: **semantically similar input with severely dissimilar encodings** reducing translation performance

| | |
|---|---|
| zhongguo dianzi yinhang yewu guanli xingui jiangyu sanyue yiri qi shixing | |
| china's new management rules for e-banking operations to take effect on march 1 | |

**(Above)**: Source text and reference translation **(Below)**: Perturbed source

| | |
|---|---|
| *zhongfang* dianzi yinhang yewu guanli xingui jiangyu sanyue yiri qi shixing | |

- Use **adversarial stability training (AST)** framework to improve NMT robustness by integrating an adversarial objective to encourage noisy and true data encodings to be similar

## Task Definition

Develop a **perturbation-robust Chinese-English NMT model** with a modified AST in a Transformer network to **generalize better on difficult unstructured datasets** spanning multiple domains for which it is crucial to maintain robustness over noise

## Dataset: Casict2015

- 2 million parallel Chinese-English sentence pairs (22,802,353 words) from CWMT with 60% web crawl, 20% movie subtitles and 20% from English to Chinese thesaurus, segmented with Jieba
- Spans many contexts, from to technical writing to biblical texts to colloquial speech (e.g. below)

Ex. 1: This paper introduces a method for mining risk rules using variable precision rough set ( VPRS ) model .

Ex. 2: What...oh, my god! oh, my god! what the f* just happened?

Ex. 3: Psalm 29:10 The LORD sits enthroned over the flood; the LORD is enthroned as King forever.

Ex. 4: If your score was between 27 and 38: You're a crafty Kisser!

## Methodology and Model

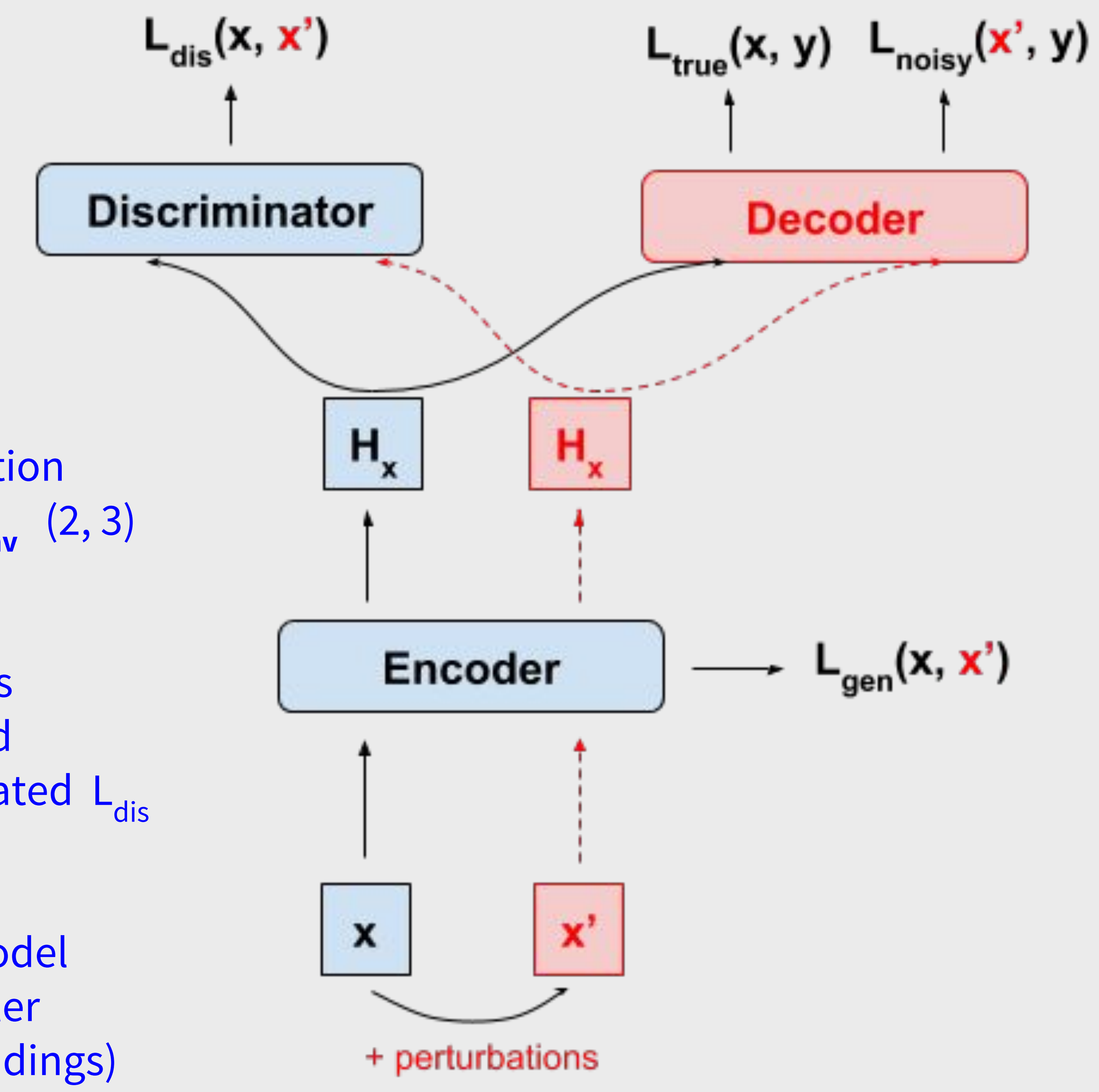
### I. AST

Given a minibatch of source sentences x, we construct a minibatch of perturbed sentences x' by adding random Gaussian noise to all word embeddings to simulate various types of feature-level perturbations.

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

1. Encoder acts as **generator (G)** to make embeddings $H_x$ and $H_{x'}$ as similar as possible to fool **discriminator (D)**
2. D tries to distinguish noisy from true embeddings by maximizing D(G(x)) to 1 and minimize D(G(x')) to 0
3. New **objective J** (1) is a hybrid loss function that incorporates $\mathbf{L_{true}}$ and $\mathbf{L_{noisy}}$ and $\mathbf{L_{inv}}$ (2, 3)
4. During training, a noisy batch is created for each true batch and gradient updates are halted until D finishes evaluation and $L_{true}$ and $L_{noisy}$ and $L_{inv}$ have been calculated $L_{dis}$ is backpropagated separately (3)

6. We evaluated a baseline Transformer model Against AST framework variations (smaller batch and random vs. pretrained embeddings)




### II. Hybrid loss functions

$$L(x,y;\theta) = \sum_{(x,y)\in S} -logP(y|x;\theta) \quad (1)$$

$$L_{inv}(x,x';\theta_{enc},\theta_{dis}) = E_{x'\sim\mathcal{N}(x)}[-log(1-D(G(x')))] \quad (2)$$

$$L_{dis}(x,x';\theta_{enc},\theta_{dis}) = E_{x\sim S}[-logD(G(x))] + E_{x'\sim\mathcal{N}(x)}[-log(1-D(G(x')))] \quad (3)$$

$$\mathcal{J}(\theta) = \sum_{(x,y)\in\mathcal{S}}(\mathcal{L}_{true}(x,y;\theta_{enc},\theta_{dec}) - \alpha\mathcal{L}_{inv}(x,x';\theta_{enc},\theta_{dis}) + \beta\mathcal{L}_{noisy}(x',y;\theta_{enc},\theta_{dec})) \quad (4)$$

### III. Ablation Studies

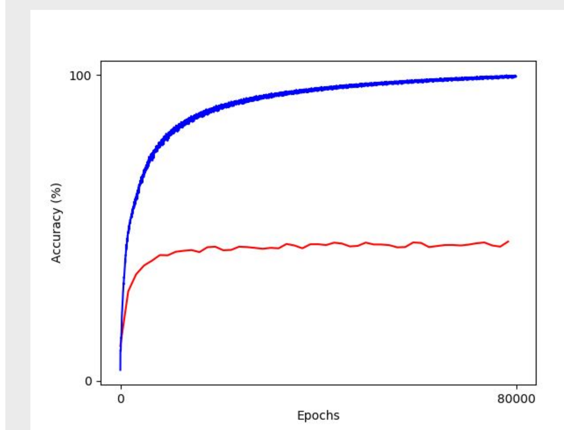
To measure individual effectiveness of sub-parts of our framework, we (1) removed the GAN structure from the training framework, reducing to a data augmentation problem of simply training on both original and noisy data and (2) implemented the original simplified generator loss $\mathbf{L_{inv}}$ as negative $\mathbf{L_{dis}}$

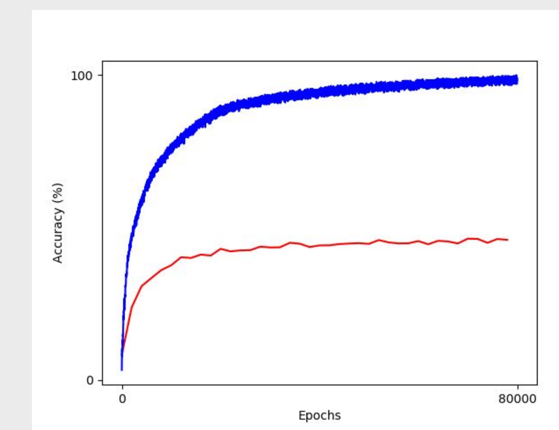
| Evaluation metric | No Discriminator | Simplified Loss | Modified AST |
|---|---|---|---|
| BLEU | 4.46 | 4.64 | 4.67 |
| Perplexity | 3.14 | 3.06 | 2.43 |

## Results

**Summary of Model Comparison**: Best performing model was AST Embed (pre-trained embeddings from Wikipedia) with 18.42 BLEU and 1.23 Perplexity

| Model Comparison | Tr baseline | AST embed Tr (small batch) | AST Random Embed Tr | AST Embed Tr |
|---|---|---|---|---|
| BLEU | 16.77 | 17.89 | 18.23 | **18.42** |
| Perplexity | 1.79 | 2.86 | 1.44 | **1.23** |




**(Middle)**: **Train and dev accuracy** (% words matching reference) for baseline **(left)** and AST Embed Tr **(right)**

**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.**

**(Bottom)**: Best Model (AST Embed Tr) evaluated against baseline on **semantically similar perturbed text**

## Conclusion and Future Work

- AST is an effective method to develop perturbation-robust Chinese-English NMT models and performs well with Transformer network on diverse datasets
- Ablation studies demonstrate the efficacy of some components constituting our hybrid model: the GAN component and our modified objective loss
- Future work could include: (1) techniques for rare words such as byte-pair encoding (2) using weak supervision to label semantically similar train text for AST input

## References


[1] Cheng, Yong, et al. "Towards robust neural machine translation." arXiv preprint arXiv:1805.06130 (2018).

[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in Neural Information Processing Systems. 2017.