Table 1: The non-robustness problem of neural machine translation. Replacing a Chinese word with its synonym (i.e., “不怕”→“不畏”) in example 1 or its homonym (i.e., “不得不”→“不的不”) in example 2 leads to significant erroneous changes in the English translation.

Small perturbations in the input can severely distort intermediate representations and thus impact translation quality of neural machine translation (NMT) models. Due to the introduction of NNN and attention, each contextual word can influence the model prediction in a global context. As shown in Table 1, although we only replace a source word with its synonym or its homonym, the generated translation has been completely distorted. In this paper, we propose to improve the robustness of NMT models with adversarial stability training (AST). The basic idea is to make both the encoder and decoder in NMT models robust against input perturbations by enabling them to behave similarly for the original input and its perturbed counterpart.

Constructing Perturbed Inputs

Our training framework can be easily extended to arbitrary noisy perturbations. Especially, we can design task-specific perturbation methods. In this paper, we propose two possible strategies to construct the perturbed inputs at different levels of representations.

- **At the lexical level:** We replace words at sampled positions with other words in the vocabulary according to the following distribution:

\[
P(x|x_i) = \sum_{\bar{E} \in \mathcal{V}_i} \exp(\cos(E[x_i], \bar{E}[x]))
\]

- **At the feature level:** We add the Gaussian noise to word embeddings to simulate possible types of perturbations. That is:

\[
E[x_i^\prime] = E[x_i] + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)
\]

Translation Experiments

Table 3: Case-insensitive BLEU scores on Chinese-English translation.

Table 4: Case-sensitive BLEU scores on WMT 14 English-German translation.

Ablation Study

Table 2: Ablation study of adversarial stability training AST with Chinese-English translation. “×” means the loss function is included in the training objective while “✓” means it is not.