Adversarially Improving NMT Robustness to ASR Errors with Confusion Sets

Shuaibo Wang[1], **Yufeng Chen**[1*], **Songming Zhang**[1]
Deyi Xiong[2], **Jinan Xu**[1]
[1] School of Computer and Information Technology, Beijing Jiaotong University,
Beijing 100044, China
[2] College of Intelligence and Computing, Tianjin University, Tianjin, China
{wangshuaibo,chenyf,zhangsongming,jaxu}@bjtu.edu.cn;
dyxiong@tju.edu.cn

Abstract

Neural machine translation (NMT) models are known to be fragile to noisy inputs from automatic speech recognition (ASR) systems. Existing methods are usually tailored for robustness against only homophone errors which account for a small portion of realistic ASR errors. In this paper, we propose an adversarial example generation method based on confusion sets that contain words easily confusable with a target word by ASR to conduct adversarial training for NMT models. Specifically, an adversarial example is generated from the perspective of acoustic relations instead of the traditional uniform or unigram sampling from the confusion sets. Experiments on different test sets with hand-crafted and real-world noise demonstrate the effectiveness of our method over previous methods. Moreover, our approach can achieve improvements on the clean test set.

1 Introduction

Neural machine translation (NMT) has been widely used and deployed as a “de facto standard” (Gehring et al., 2017; Vaswani et al., 2017). In many application scenarios, NMT models translate sentences generated by automatic speech recognition (ASR) systems. Although current ASR systems have made substantial progress, texts recognized by them still suffer from a variety of recognition errors, i.e., *deletion*, *insertion* or *substitution* of tokens, where substitution errors are the most common errors among them (Xue et al., 2020). These errors will result in severe degradation of translation quality due to the discrepancy between training and test data (Di Gangi et al., 2019; Cui et al., 2021).

In order to mitigate the negative impact of substitution errors on NMT models, many studies explore external phonetic information as extra representation or training objective. Liu et al. (2019) improve NMT robustness to homophone errors with joint textual and phonetic embeddings. Xue et al. (2020) utilize a gating mechanism to integrate phonetic information into the final output of the encoder to alleviate homophone errors. Qin et al. (2021) exploit a noise detector to convert homophone errors tokens into syllables and use a syllable-aware NMT model to translate the mixed sequences into target texts.

ASR-Ref	wǒ shēn biān haí yǒu gè lì zī 我 身 边 还 有 个 例 子
Trans-Base	There is another example around me.
ASR-Hyp	wǒ xiān biān haí yǒu gè lì zī 我 先 边 还 有 个 例 子
Trans-Base	I had another example before.
Trans-Pron	I have another example at the beginning.

Figure 1: An example in BSTC corpus.[1] The original character ‘身’ (‘body’) is recognized as a non-homophonous character ‘先’ (‘first’). Trans-Base and Trans-Pron represent the translation of the vanilla Transformer and the robust Transformer with external phonetic information, respectively.

These methods are usually designed for dealing with noisy tokens with same or similar pronunciation. However, realistic substitution noises in ASR-generated texts are not only limited to homophone errors due to complicated acoustics-linguistics relations, as shown in Figure 1. When the correct character ‘身 (shēn)’ is recognized as a non-homophonous character ‘先 (xiān)’ by an ASR system, previous methods fail to provide correct translation with the help of external phonetic information, indicating that employing phonetic information is not sufficient to handle realistic ASR errors.

To tackle this issue, we propose an adversarial example generation method based on confusion sets, where words in a confusion set for a target

[*] Corresponding author.

[1] A Chinese-English speech translation corpus introduced in Section 3.1.

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 221–227
November 20–23, 2022. ©2022 Association for Computational Linguistics

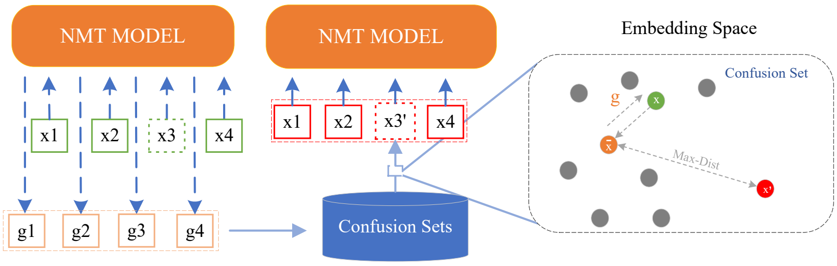

Figure 2: Illustration of the proposed method. The right most part denotes the embedding space for a confusion set. g is the gradient of input token x. Gradient descent is performed to change the original embedding x to $\overline{x}$. Then a token farthest from $\overline{x}$ is selected for substitution.

word are those that make ASR confusing with the target word semantically, lexically, or phonitically. Close to our work, Sperber et al. (2017) generate noisy training examples by uniformly selecting tokens from a sampled vocabulary. Martucci et al. (2021) propose a lexical noise model to emulate noisy transcripts by artificially corrupting clean transcripts. While they focused on heuristics for introducing noise to clean transcripts, without any explicit knowledge of acoustics or NMT models, which can not develop generalized and aggressive samples (Ebrahimi et al., 2018). In this paper, we propose to generate adversarial examples from the perspective of acoustic relations (Shivakumar and Georgiou, 2019). The acoustic relations reflect the acoustic similarity between words, and modeling the acoustic relations of confusing tokens is beneficial to mitigate the negative impact of ASR errors (Shivakumar et al., 2019).

Our key idea is to make the representations of confusing tokens close to those of corresponding golden tokens in the embedding space so as to model the acoustic relations of confusing tokens. To this end, we craft adversarial examples that have weak acoustic relations with original sentences to attack the NMT model according to both the gradient of the source token and the distance between token embeddings. With the generated adversarial examples, we conduct adversarial training to improve the robustness of NMT models against ASR errors.

To sum up, our contributions are as follows:

- We propose an adversarial example generation method from the perspective of acoustic relations based on confusion sets to handle realistic ASR errors.

- Experimental results show that our method can not only make NMT models resilient to ASR errors in both hand-crafted and real-world scenarios, but also outperform the baselines on the clean test sets.

2 Approach

We follow previous practice of using adversarial training to improve the robustness of NMT (Belinkov and Bisk, 2018; Cheng et al., 2020) by iteratively adding generated adversarial examples to the training set. In this section, we will introduce our approach (illustrated in Figure 2) in detail.

2.1 ASR Confusion Sets

Previous works (Xue et al., 2020; Cui et al., 2021) employ an external pronunciation dictionary to heuristically construct noisy candidates for each word. Some candidates generated in this way would not confuse ASR systems in real scenarios. Inspired by prior work (Wang et al., 2020), we construct confusion sets based on a corpus of ASR hypotheses and corresponding manual transcripts. Specifically, we first align each ASR hypothesis and its reference transcript at the word level by minimizing the Levenshtein distance between them. Then, we collect substitutions based on alignments.

2.2 Adversarial Example Generation

In order to improve the robustness of an NMT model against ASR errors, we generate adversarial examples with weak acoustic relations to the original source inputs to attack the victim NMT model, maintaining the acoustic rationality of generated sentences. In detail, we first randomly select a certain proportion of tokens to be replaced in source inputs and then choose candidate tokens for substitution from the corresponding confusion set constructed before. The chosen candidate tokens are farthest from the source input tokens in the embedding space.

Moreover, to make adversarial examples more generalized and aggressive, we take the gradients of the NMT model with respect to the source input tokens into account during adversarial example generation. Specifically, as shown in Figure 2, we first update token embeddings in the embedding space by gradient descent before choosing the replacement tokens, aiming to make the substitution based on the newly updated NMT model.

Formally, let $\mathbf{x} = (x_1, x_2, ..., x_N)$ and $\mathbf{y} = (y_1, y_2, ..., y_M)$ be the source input and target translation, respectively. The training loss of a single example is defined as:

$$\mathcal{L}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = -\frac{1}{M}\sum_{t=1}^{M} \log P(y_t|\mathbf{y}_{<t}, \mathbf{x}; \boldsymbol{\theta}) \quad (1)$$

where $\mathbf{y}_{<t} = (\langle s\rangle, y_1, y_2, ..., y_{t-1})$ is the partial target input and $\boldsymbol{\theta}$ denotes the parameters of the NMT model. With this the forward loss, we define $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = (g_1, g_2, ..., g_N)$ as the gradients of the input sentence $\mathbf{x}$ and $g_i = \nabla_{x_i}\mathcal{L}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ as the gradient for the ith token x_i.

We then induce an appropriate substitution token $x_i^{'}$ for token x_i from the corresponding confusion set C_{x_i}:

$$x_i^{'} = \underset{t_i \in C_{x_i}}{\arg\max}\, \text{Dist}^2(e_{t_i}, e_{x_i} - \lambda g_i) \quad (2)$$

where e represents token embeddings, Dist(.,.) denotes the euclidean distance between token embeddings, and λ is a hyperparameter.

For further analysis, we denote $e_{t_i} - e_{x_i}$ as d, and remove factors that have no effect on the choice of candidates. We then get:

$$\begin{aligned} \text{Dist}^2(e_{t_i}, e_{x_i} - \lambda g_i) &= \|e_{t_i} - e_{x_i} + \lambda g_i\|^2 \\ &= [d + \lambda g_i]^T [d + \lambda g_i] \\ &= d^T d + 2\lambda d^T g_i + \lambda^2 g_i^T g_i \\ &\propto \|d\|^2 + 2\lambda d^T g_i \quad (3) \end{aligned}$$

where we can see the substitution criterion is determined by two factors. The L2 norms of d represent the distance between token embeddings, and the second term is exactly the substitution strategy of Cheng et al. (2019). λ is a trade-off between the two factors. As demonstrated by our experiments (see Appendix A), small values of λ are preferred to improve the robustness of NMT models against ASR errors.

Dataset	Utterances	WER
Train	37,901	27.90%
Valid	956	15.21%
Top5-hyp.(asr)	188,317	19.09%†

Table 1: Statistics of the BSTC corpus. † denotes that the WER is calculated using the same tool reported in (Zhang et al., 2021) on the top-5 ASR hypotheses and corresponding manual transcripts provided by the BSTC corpus.

3 Experiments

3.1 Dataset

To be in line with previous work (Xue et al., 2020), we evaluated our approach on two Chinese-English datasets and constructed noisy test sets by randomly replacing tokens (more details in (Xue et al., 2020)).

Furthermore, to verify the effectiveness of our method in real-world scenarios, we used the public BSTC Chinese-English speech translation (ST) corpus[2] (Zhang et al., 2021) where the training set contains ASR results and corresponding manual transcripts and target sentences. Since the test set is not publicly available, we randomly excluded 1k pairs from the training data as our test set and used the public validation set to select the best checkpoint.

We constructed ASR confusion sets using all ASR hypothesis-reference pairs from the BSTC corpus. As shown in Table 1, to be consistent with the word error rate (WER) of real-world scenarios, we randomly selected 20% tokens of sentences for replacement to generate adversarial examples during training.

For all experiments, we segmented Chinese sentences into Chinese characters and employed Moses tokenizer for English tokenization. We learned byte pair encoding (BPE) (Sennrich et al., 2016) with 32K operations on the target side. We followed (Vaswani et al., 2017) to set the remaining configuration and implemented all NMT systems with Fairseq[3]. The NIST task was trained for 50K steps while the WMT17 task was trained for 150K steps due to larger training data. We report case-insensitive tokenized BLEU scores for NIST and WMT17 tasks and case-insensitive SacreBLEU (Post, 2018)[4] for BSTC.

[2] https://aistudio.baidu.com/aistudio/competition/detail/44

[3] https://github.com/pytorch/fairseq

[4] SacreBLEU hash: BLEU+case.mixed+lang.zh-

Method	NIST			WMT17		
	Clean	Noise	Δ	Clean	Noise	Δ
Vaswani et al. (2017)	45.05	39.40	-	23.27	20.35	-
Cheng et al. (2019)	45.32	43.72	+4.32	23.61	23.00	+2.65
Wang et al. (2020)	45.01	43.22	+3.82	23.52	22.20	+1.85
Martucci et al. (2021)	45.17	43.43	+4.03	23.52	22.88	+2.53
Ours	**45.65**	**44.24***	**+4.84**	**23.94**	**23.35***	**+3.00**

Table 2: Experiment results on the NIST (average BLEU scores on nist02,03,04,05,06,08) and WMT17 task. Results on noisy test sets are calculated by averaging BLEU scores on three artificial noisy test sets generated by randomly substituting one, two and three tokens in clean source sentences based on confusion sets. Δ represents BLEU improvements over Transformer on the noisy test sets. Results with mark $*$ are statistically (Koehn, 2004) better than (Cheng et al., 2019) with $p < 0.05$.

Method	Test-Ref	Test-Hyp
Vaswani et al. (2017)	20.48	15.51
Sperber et al. (2017)	20.46	16.11
Cheng et al. (2019)	20.92	15.75
Wang et al. (2020)	20.38	16.21
Martucci et al. (2021)	20.39	16.28
Ours	**21.17**	**16.66**

Table 3: Results of different methods on the BSTC ST corpus. **Hyp** and **Ref** represents ASR hypotheses and corresponding manual transcripts, respectively.

3.2 Main Results

We first compared against other noisy example generation methods proposed by Sperber et al. (2017) and Martucci et al. (2021). Besides, Cheng et al. (2019) present a gradient-based method to generate adversarial examples tightly guided by the training loss. Wang et al. (2020) simulate ASR hypotheses based on n-gram confusions where n can vary.

Results are shown in Table 2. Firstly, the vanilla Transformer suffers a great performance drop on the noisy test data, which is consistent with previous findings (Belinkov and Bisk, 2018). Secondly, among all methods trained with adversarial examples, our approach achieves the best performance on noisy test sets on the two corpora, i.e., 4.84 and 3.00 BLEU points over vanilla Transformer respectively, which suggests that adversarial examples generated by our strategy are more effective to make NMT models robust against ASR errors. Thirdly, our approach obtains higher BLEU scores on clean test sets than Cheng et al. (2019) that is the most related to our method, by 0.33 BLEU points on average, indicating that our adversarial examples can be used to improve translation quality as a regularization, whereas other methods only achieve small improvements or even drop.

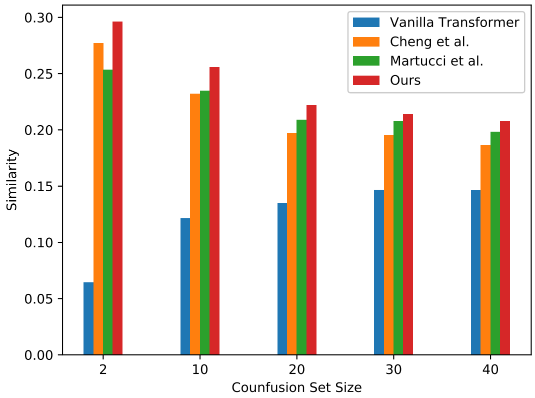

Figure 3: Average similarities between confusing tokens in the confusion set and corresponding ground-truth tokens. The confusion set size is the number of tokens in the confusion set.

Furthermore, we conducted experiments on the BSTC speech translation dataset to verify the effectiveness of our approach in real-world scenarios. We first trained the NMT model on the WMT17 Chinese-English corpus and then fine-tuned it on the BSTC training set. As shown in Table 3, we can see that most other methods improve the robustness of NMT, but slightly degrade the translation performance on the clean test set. Instead, the consistent improvements achieved by our approach on clean test sets and realistic ASR noise test set suggest that our method is also applicable and outstanding in real application scenarios with complex errors.

3.3 Acoustic Relations

To further analyse acoustic relations between words, we chose the checkpoint achieving the best

en+numrefs.1+smooth.exp+tok.13a+version.1.5.1

Method	Clean	HP Noise	ASR Noise	ADV Noise
Vaswani et al. (2017)	45.05	39.65 (5.40 ↓)	39.40 (5.65 ↓)	39.29 (5.76 ↓)
Li et al. (2018)	45.16	44.87 (**0.29** ↓)	41.42 (3.74 ↓)	40.00 (5.16 ↓)
Liu et al. (2019)	45.26	42.47 (2.79 ↓)	40.47 (4.79 ↓)	39.79 (5.47 ↓)
Xue et al. (2020)	45.07	44.74 (0.33 ↓)	41.22 (3.85 ↓)	39.96 (5.11 ↓)
Qin et al. (2021)	45.29	**44.99** (0.30 ↓)	41.37 (3.92 ↓)	40.37 (4.92 ↓)
Ours	**45.65**	44.79 (0.86 ↓)	**44.24 (1.41 ↓)**	**44.06 (1.59 ↓)**

Table 4: Results of different methods handling homophone errors on the NIST translation dataset. **HP Noise** and **ASR Noise** test sets are generated based on homophones and confusing tokens in the confusion sets, respectively. **ADV Noise** test set is generated by our substitution strategy. Note that the way of noisy test sets construction and the results calculation are consistent with those described in the main paper.

robustness on the NIST02 noise validation set. Following (Shivakumar and Georgiou, 2019), we employ the cosine similarity between confusing tokens to reflect the acoustic relations between words modeled by our method in the embedding space.

As shown in Figure 3, the worst results calculated by vanilla Transformer (Vaswani et al., 2017) show that the traditional approach can not capture the acoustic similarity between confusing tokens. Over all different size of confusing sets, our method achieves higher similarities than baselines, suggesting that our method can effectively model the acoustic relations for confusing tokens. This makes NMT models be able to alleviate the influence of real ASR errors by learning to adjust to similar representations of these erroneous tokens. Moreover, we can also see that the degree of similarity between confusing tokens is also consistent with the NMT model robustness in real-world scenarios shown in Table 2, which further validates our motivation of generating adversarial examples in the perspective of acoustic relations.

3.4 Homophone Errors vs. ASR Errors

We also examined the performance of our method in solving homophone errors. As shown in Table 4, we can see that these methods can greatly reduce the negative impact of homophone errors on NMT models but drop a lot when dealing with real-word errors, which indicates that ASR errors are not limited to homophone errors and the robustness of NMT models improved by exploiting external phonetic information fail to generalize over real errors. Additionally, previous methods achieve much worse performance than our method on the **ADV** noise test set and the performance gap from our method is enlarged to 3.88 BLEU, which suggests that adversarial examples generated by our method can attack NMT models more effectively. On the contrary, our method not only obtain higher performance on the clean test set and make NMT more robust to various real noises, but also can achieve competitive results on the **HP** noise test set compared with previous methods only tailored for homophone errors.

4 Conclusion

In this paper, we have presented an adversarial example generation method based on confusion sets to make NMT models robust against real ASR errors. The acoustic relations between confusing tokens modeled by our approach can make NMT models more resilient to ASR errors. Experimental results on two Chinese-English text translation tasks and one Chinese-English speech translation task prove that the effectiveness of our method. Moreover, our method does not require any changes to models. It could be therefore orthogonal and complementary to other methods to further improve the robustness of NMT model.

Acknowledgement

This research work is supported by the National Key R&D Program of China (2020AAA0108001), the National Nature Science Foundation of China (No. 61976016, 61976015 and 61876198) and Toshiba (China) Co.,Ltd. The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333, Florence, Italy. Association for Computational Linguistics.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. AdvAug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5961–5970, Online. Association for Computational Linguistics.

Tong Cui, Jinghui Xiao, Liangyou Li, Xin Jiang, and Qun Liu. 2021. An approach to improve robustness of NLP systems against ASR errors. *CoRR*, abs/2103.13610.

Mattia Antonino Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. 2019. Robust neural machine translation for clean and noisy speech transcripts. *arXiv preprint arXiv:1910.10238*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1243–1252. JMLR.org.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Xiang Li, Haiyang Xue, Wei Chen, Yang Liu, Yang Feng, and Qun Liu. 2018. Improving the robustness of speech translation. *arXiv preprint arXiv:1811.00728*.

Hairong Liu, Mingbo Ma, Liang Huang, Hao Xiong, and Zhongjun He. 2019. Robust neural machine translation with joint textual and phonetic embedding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3044–3049, Florence, Italy. Association for Computational Linguistics.

Giuseppe Martucci, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. Lexical Modeling of ASR Errors for Robust Speech Translation. In *Proc. Interspeech 2021*, pages 2282–2286.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Wenjie Qin, Xiang Li, Yuhui Sun, Deyi Xiong, Jianwei Cui, and Bin Wang. 2021. Modeling homophone noise for robust neural machine translation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7533–7537.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Prashanth Gurunath Shivakumar and Panayiotis Georgiou. 2019. Confusion2vec: towards enriching vector space word representations with representational ambiguities. *PeerJ Computer Science*, 5:e195.

Prashanth Gurunath Shivakumar, Mu Yang, and Panayiotis Georgiou. 2019. Spoken language intent detection using confusion2vec. *Interspeech 2019*.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *International Workshop on Spoken Language Translation (IWSLT)*, page 18.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Longshaokan Wang, Maryam Fazel-Zarandi, Aditya Tiwari, Spyros Matsoukas, and Lazaros Polymenakos. 2020. Data augmentation for training dialog models robust to speech recognition errors. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 63–70, Online. Association for Computational Linguistics.

Haiyang Xue, Yang Feng, Shuhao Gu, and Wei Chen. 2020. Robust neural machine translation with ASR errors. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 15–23, Seattle, Washington. Association for Computational Linguistics.

Ruiqing Zhang, Xiyang Wang, Chuanqiang Zhang, Zhongjun He, Hua Wu, Zhi Li, Haifeng Wang, Ying Chen, and Qinfei Li. 2021. BSTC: A large-scale Chinese-English speech translation dataset. In *Proceedings of the Second Workshop on Automatic Simultaneous Translation*, pages 28–35, Online. Association for Computational Linguistics.

A Effect of Hyperparameter λ

We evaluated the performance of our proposed method with different λs. As shown in Table 5, the robustness of NMT is improving as λ decreases, which implies that the distance between confusing

Test set	$\lambda =$				
	0.0	0.1	0.5	1.0	5.0
Clean	45.45	45.65	45.60	45.43	45.46
Noise	43.92	**44.24**	44.06	43.90	43.88

Table 5: Effect of λs on the NIST clean and noisy test sets.

token and ground-truth token embeddings is critical to handle ASR errors. Moreover, the poor result obtained when λ = 0.0 on the noisy test set indicates that gradient information of the victim model benefits the robustness of NMT to ASR noise. We conjecture the addition of NMT gradient information can help generate diversified adversarial examples.