Multimodal Robustness for Neural Machine Translation

Yuting Zhao*
Tokyo Metropolitan University
zhao-yuting@ed.tmu.ac.jp

Ioan Calapodescu
Naver Labs Europe
ioan.calapodescu@naverlabs.com

Abstract

In this paper, we look at the case of a Generic text-to-text NMT model that has to deal with data coming from various modalities, like speech, images, or noisy text extracted from the web. We propose a two-step method, based on composable adapters, to deal with this problem of Multimodal Robustness. In the first step, we separately learn domain adapters and modality specific adapters, to deal with noisy input coming from various sources: ASR, OCR, or noisy text (UGC). In a second step, we combine these components at runtime via dynamic routing or, when the source of noise is unknown, via two new transfer learning mechanisms (Fast Fusion and Multi Fusion). We show that our method provides a flexible, state-of-the-art, architecture able to deal with noisy multimodal inputs.

1 Introduction

Neural Machine Translation (NMT) has achieved great performances (Gehring et al., 2017; Vaswani et al., 2017) but still suffers from various robustness problems, as shown by many previous works (Koehn and Knowles, 2017; Belinkov and Bisk, 2017; Khayrallah and Koehn, 2018).

Specific datasets like Michel and Neubig (2018); Berard et al. (2019a); Li et al. (2019a); Specia et al. (2020a), consisting in noisy comments from Reddit, or from Wikipedia or from restaurant reviews, were proposed to showcase this robustness problem.

Based on these datasets, several solutions were proposed to deal with the problem. For example, using synthetic noise, with data augmentation Vaibhav et al. (2019); Karpukhin et al. (2019) or other adversarial training methods (Ebrahimi et al., 2018; Cheng et al., 2018, 2019, 2020). But these solutions, based on purely artificial noise do not guarantee the best results on real noise (Michel et al., 2019).

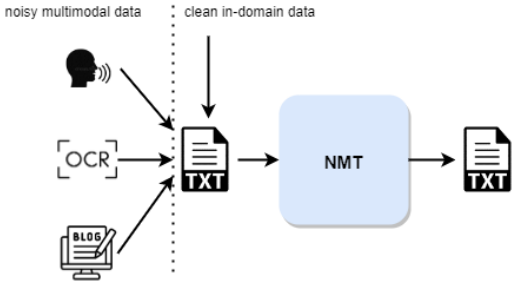

Figure 1: Setting for a text-to-text NMT model robust to clean and noisy multimodal data (coming from speech, images or a web page) on a specific domain

Another set of solutions make use of real noisy data to fine-tune or adapt generic models, so that they become more robust to realistic noise distributions (Michel and Neubig, 2018; Murakami et al., 2019; Helcl et al., 2019; Alam and Anastasopoulos, 2020; Berard et al., 2020). But, as shown in Specia et al. (2020a), these noise specific methods do not generalise well on domains or noise distributions not seen at training time.

In this work, we want to propose an extensible robustness solution for NMT able to overcome some of these limitations. To have a realistic setting, we propose to build a model able to translate several clean domains and their respective noisy versions, coming from various sources or modalities. As shown in Figure 1, we work with clean in-domain data and noisy versions of the same data coming from an automatic speech recognition (ASR) system, from an optical character recognition (OCR) application, or from social media with presence of keyboard typos or spelling errors in the user-generated content (UGC).

Our contributions are as follows:

*Work done during an internship at Naver Labs Europe.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8505–8516
December 7-11, 2022

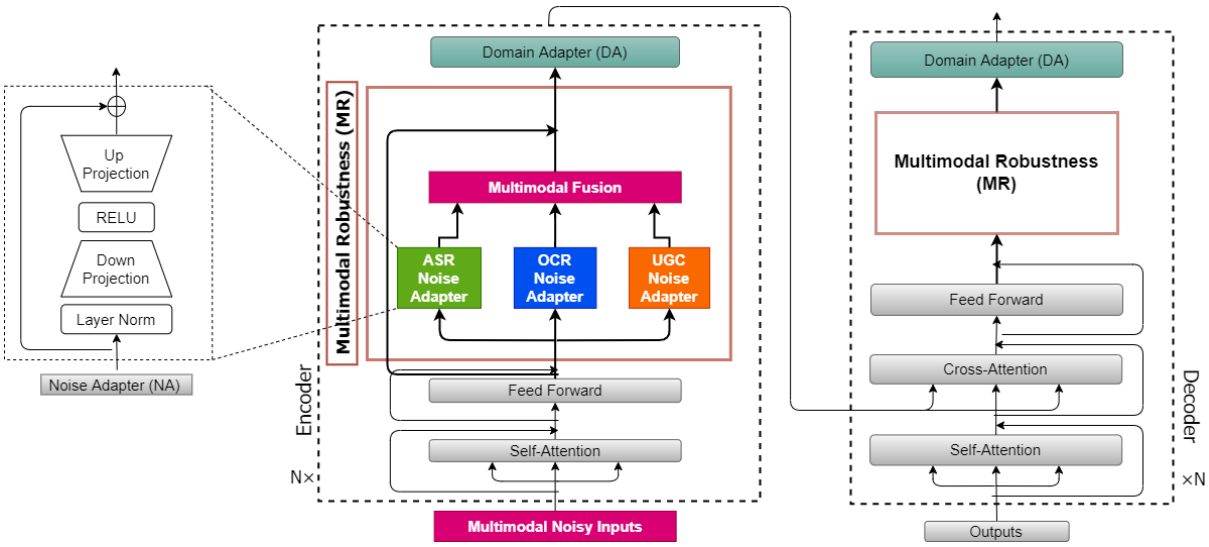

Figure 2: Robust Multimodal NMT model architecture. Domain Adaptation and Noise Adaptation are separated in their own Adapter Layers, and an optional Multimodal Fusion component is inserted between the Noise and Domain Adapters. This MR component is integrated both in the encoder and the decoder.

- We propose a setting to explore the robustness problem for NMT from the realistic viewpoint of a generic NMT model which has to deal with multimodal inputs, such as ASR input, OCR input, and UGC data.
- We implement a robust multimodal NMT model, which can handle different types of noise (but also clean data) simultaneously via composable Adapter Layers (for noise type and domains).
- We show that this model can easily be extended to new sources of noise and new domains.
- We also propose two new fusion mechanisms to deal transparently with the source of noise (a.k.a dealing with input when the source of noise is unknown).

2 Methodology

In Figure 2, we show the overall architecture of our proposition, which separates domain adaptation (DA) and a multimodal robustness (MR), with various noise adapters (NA) and an optional fusion mechanism, inside a conventional transformer encoder and decoder MT model.

2.1 Separate Domain Adaptation and Noise Adaptation Learning

The first step in our method is to separately learn Domain Adaptation (DA) and Noise Adaptation (NA).

For that, we start from a Generic NMT model, based on a transformer encoder-decoder architecture.

Given our training data separated into clean and noise specific data (as described in our experimental settings), we first inject Adapter Layers (Houlsby et al., 2019) to learn the Domain Adaptation task on clean in-domain data.

In a second step, these domain adapter layers are loaded but frozen alongside the other parameters of the model. We inject new Adapter Layers between the domain adapter layers and the feed-forward component of the transformer cell. We call them Noise Adapters (NA) and we create one NA for each type of noise. Each NA is trained only on his specific type of noise.

The DA and NA have the same structure, as shown in Figure 2, which is a down projection to a bottleneck dimension followed by an up projection to the initial embedding size.

Once, the DA and the various NA are separately learned, we can load them inside the same model and recompose them at runtime for decoding as explained in the next section.

2.2 Domain Adaptation and Noise Adaptation composition at Runtime

2.2.1 Dynamic Routing

The first way to recompose these various Adapter Layers, when we know the type of

noise, is to dynamically route their outputs at runtime. For example, to process OCR input, the data is first forwarded through the multi-head attention mechanisms (self-attention inside the encoder or self-attention and cross-attention inside the decoder), the feed-forward component, the OCR NA and finally the DA.

This solution works, when we actually know the source of noise. But, to be able to deal with an unknown source of noise, we need an additional component to combine what was learned by each NA.

For that, we propose two new fusion mechanisms, Fast Fusion and Multi Fusion, to combine automatically all NA layers outputs. We compare later these two solutions to the well known Adapter Fusion method proposed by (Pfeiffer et al., 2021), which learns a parameterized mixer of the outputs from trained NAs.

2.2.2 Fast Fusion (FF)

Fast Fusion (FF) is our first proposition to combine the knowledge from the various noise adapters. This simple solution consists in learning a linear projection W from the concatenation of the output of all the NAs (H) to the DA embedding size (d_{m}), followed by a residual connection x.

$$\mathrm{FF}(H) = \mathrm{W}(\mathrm{Concat}(H)) + x$$
$$H : \{\mathbf{h}_{\mathrm{asr}}, \mathbf{h}_{\mathrm{ocr}}, \mathbf{h}_{\mathrm{ugc}} \ldots \mathbf{h}_{\mathrm{noise_n}}\}$$
$$\mathrm{W} : \mathbb{R}^{|\mathrm{H}| \times \mathrm{d_m}} \rightarrow \mathbb{R}^{\mathrm{d_m}}$$

This module is learned on a mix of all types of noises. Everything, but the projection, is frozen inside the model.

2.2.3 Multi Fusion (MF)

Multi Fusion (MF) is our second proposal for merging the knowledge from all the NAs. Inspired by Adapter Fusion (AF), we implement an attention mechanism to learn how to combine various adapters. Contrarily to AF, we use a multi-head attention mechanism, like in the traditional transformer (Vaswani et al., 2017). Several attention heads are learned on a partition of the embedding space formed by the output of the NAs, and then followed by a residual x.

$$\mathrm{MF}(Q, K, V) = \mathrm{Concat}(\mathrm{head}^1, \ldots, \mathrm{head}^M) + x$$
$$\mathrm{head}^{i \in [1,M]} = \mathrm{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}})V$$
$$= \mathrm{softmax}(\frac{\mathbf{x} \cdot \mathrm{H}_{\mathrm{R}}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}})\mathrm{H}_{\mathrm{R}}$$

Where d_{k} is d_{m} divided by the number of attention heads M.

Like FF, MF is learned on a mix of all types of noise.

3 Experiments

3.1 Initial Corpus

To build our multimodal dataset, we start with the Multilingual TEDx (mTEDx) corpus (Salesky et al., 2021b), which is a multilingual corpus created from TEDx talks and suited for speech recognition and machine translation tasks. Table 1 shows the number of sentences available for translation in the mTEDx corpus.

This corpus is composed of audio recordings and their human provided transcriptions in 8 languages[1] and translations into up to 5 languages.[2]

These translations in 12 language pairs[3] can be obtained from OpenSLR.[4]

mTEDx	Train set	Valid set	Test set
Fr→En	30,171	1,036	1,059
Fr→Es	20,826	1,036	1,059
Fr→Pt	13,286	1,036	1,059
It→En	24,576	931	999
It→Es	2,261	931	999
Es→En	36,263	905	1,012
Es→Fr	3,663	905	1,012
Es→It	5,600	16	267
Es→Pt	21,107	905	1,012
Pt→En	30,855	1,013	1,020
Pt→Es	11,499	1,013	1,020
El→En	4,384	982	1,027

Table 1: Number of sentences available for translation in the mTEDx corpus.

3.2 Multi-modal version

From the initial mTEDx Corpus, we create four versions of the dataset to simulate clean data and noisy data coming from various sources

[1]Spanish (es), French (fr), Portuguese (pt), Italian (it), Russian (ru), Greek (el), Arabic (ar), German (de)

[2]English (en), Spanish, French, Portuguese, Italian

[3]Fr→En, Fr→Es, Fr→Pt, It→En, It→Es, Es→En, Es→Fr, Es→It, Es→Pt, Pt→En, Pt→Es, El→En.

[4]http://www.openslr.org/100

Clean in-domain	Pour trier, il faut trois secondes.
Noise ASR	pourtrier il faut trois secondes
Noise OCR	Pour trier, -, Tl faut trois secondes
Noise UGC	Poug trier, il faut trois secondes

Table 2: An example of noisy multimodal data.

(images, speech and web). An example of data from multi-modal versions is shown in Table 2.

3.3 Clean In-domain Data

We simply use the human transcripts and their translations as our clean, in-domain, dataset.

3.4 Noisy ASR data

To create the Noisy ASR version of the dataset, we use the audio files in the initial corpus and simply transcribe them using an off-the-shelf ASR system (SpeechBrain[5]).

We transcribe the mTEDx audio files for Fr and It, as it was the only available pre-trained models for Speechbrain.[6] So in total, this Noisy ASR dataset contains 5 language pairs: Fr→En, Fr→Es, Fr→Pt, It→En, It→Es.

In terms of noise, outside the usual ASR errors, we can note that the model only outputs lowercase text.

3.5 Noisy OCR data

To create this second noisy version of the mTEDx corpus, we simply print the human transcriptions to images.[7] We then use an OCR system, using CRAFT as a segmenter (Baek et al., 2019) and CRNN (Shi et al., 2015) as a recognizer, trained on Latin, Greek and Korean alphabets (case sensitive), to extract back the transcripts from the images.

3.6 Noisy UGC data

Finally, to simulate User Generated Content, as we can find on the web, we use NL-Augmenter (Dhole et al., 2021) to generate perturbations in the original mTEDx transcriptions. More specifically, we use the Butter Fingers perturbation to simulate typos based on keyboard layouts.[8]

[5]https://speechbrain.github.io/

[6]https://huggingface.co/speechbrain

[7]https://pillow.readthedocs.io/en/stable/

[8]https://github.com/GEM-benchmark/NL-Augmenter/tree/main/nlaugmenter/transformations/butter_fingers_perturbation

3.7 Evaluation

For evaluation, we use SacreBLEU (Post, 2018) on the test set to evaluate the translation quality and report BLEU (Papineni et al., 2002) and chrF (Popovic, 2015) scores.

3.8 Setup

Generic NMT model As a baseline, we trained a single multilingual NMT model trained on a huge out-of-domain dataset.

We use ParaCrawl v7.1 (Bañón et al., 2020) and select the 19 highest-resource languages paired with English. Then, like (Freitag and Firat, 2020), we build a multi-parallel corpus by aligning all pairs of languages through their English side.

We train a shared BPE model with 64k merge operations and inline casing (Berard et al., 2019b), by sampling from this data with temperature 5. We set the encoder/decoder to contain $N = 6$ layers. The embedding dimensions of all the input and output layers were set to $d_{\text{m}} = 1024$. The number of heads in all multi-head modules was set to $M = 8$. The label smoothing was set at 0.1, and the dropout was 0.1. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$. The learning rate was 0.0005, with a warm-up step of 8,000. We train the model for 120k steps, with joint BPE vocabularies of size 16k. The evaluation was performed every 20k steps and the best checkpoint was selected on the average of the validation loss.

Domain Adaptation Freezing the pre-trained multilingual NMT model, we fine-tuned the DA layers on the clean in-domain dataset to create a domain-adapted model for this setup. We kept the same parameters as the pre-trained model, and set DA to a size of 1024. We fine-tuned the DA for 3k steps with validation every 200 steps. The best checkpoint was saved according to the average of validation loss.

Noise Adaptation Keeping the DA setup fixed, we fine-tune the three types of NAs, with their respective noisy datasets: ASR NA trained on the Noisy ASR data, OCR NA trained on the Noisy OCR data and UGC NA trained on the Noisy UGC data. We keep the same parameters for the model and set NA layers to have a size of 1024.

Noise[ASR] Testsets	Multilingual NMT (out-of-domain)		DA (Clean data)		DA (Synthetic Noise)		DA (Real Noise)		DA + ASR NA		DA + AF		DA + FF		DA + MF	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
Fr→En	17.3	47.1	20.1	48.7	23.0	49.5	27.0	51.2	27.9	52.0	27.2	51.6	27.5	51.9	27.5	51.9
Fr→Es	19.3	50.7	23.0	52.8	25.2	53.4	29.6	55.3	30.1	55.8	29.1	55.0	29.7	55.3	29.7	55.5
Fr→Pt	15.9	49.0	23.1	53.1	25.2	53.7	27.7	55.1	28.0	55.3	28.9	55.9	28.7	55.7	28.7	56.0
It→En	13.1	41.8	14.8	42.8	16.8	43.4	19.5	44.9	20.0	45.2	19.5	45.0	19.2	44.7	19.9	45.1
It→Es	17.1	48.3	19.4	49.5	21.6	50.0	25.4	51.7	24.7	51.4	23.3	50.1	24.3	50.8	24.8	51.2
Avg	16.5	47.4	20.1	49.4	22.4	50.0	25.8	51.6	*26.1*	*51.9*	25.6	51.5	25.9	51.7	**26.1**	**51.9**

Table 3: BLEU and chrF scores on noisy ASR data. Multilingual NMT is our generic NMT model trained on out-of-domain Paracrawl data. DA (Clean data) is the domain adapted model on clean mTEDx data. DA (Synthetic Noise) is an adapter layer trained on synthetically generated noisy mTEDx data. DA (Real Noise) is the same but on the real type noisy mTEDx data. DA + X-NA is our proposition of decomposed DA and specific NA for noise X. DA-(AF|FF|MF) are decomposed DA and NA with a specific fusion mechanism for when we do not know the type of noise (the model has to combine all the NA present in the model). In italic are the best scores in an Oracle mode (when we know the noise source) and in bold are the best scores for the Blind mode (when the type of noise in input is not known).

Noise[OCR] Testsets	Multilingual NMT (out-of-domain)		DA (Clean data)		DA (Synthetic Noise)		DA (Real Noise)		DA + ASR NA		DA + AF		DA + FF		DA + MF	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
Fr→En	23.1	52.5	25.9	53.8	27.4	54.7	36.2	59.9	36.0	59.8	35.8	59.7	36.3	59.9	35.5	59.5
Fr→Es	22.1	55.7	26.9	57.4	27.2	57.8	38.0	63.0	37.9	63.2	37.7	62.7	37.6	62.9	37.6	63.0
Fr→Pt	18.0	53.3	24.9	56.8	25.6	57.3	37.4	63.8	35.4	62.9	36.9	63.8	36.7	63.6	36.9	63.8
It→En	17.0	49.3	18.9	50.1	20.1	50.7	29.3	55.6	28.7	55.2	28.5	55.2	28.4	55.3	28.2	54.9
It→Es	20.8	56.1	24.9	57.8	26.1	58.4	37.0	63.2	37.6	63.7	33.6	61.3	34.6	61.9	35.9	62.8
Es→En	19.6	48.3	21.9	49.9	22.5	50.4	31.1	55.6	30.9	55.6	31.2	55.6	31.6	55.9	30.9	55.6
Es→Fr	19.9	51.1	23.0	52.0	23.8	52.7	31.3	57.1	31.5	57.2	28.2	54.9	30.3	56.0	30.9	56.6
Es→It	18.9	53.4	22.2	53.8	22.0	54.4	32.2	59.2	31.8	59.6	29.4	57.9	30.2	58.7	30.4	59.2
Es→Pt	22.6	55.7	27.3	57.8	28.9	58.8	39.7	65.0	39.0	64.6	39.1	65.2	39.6	64.8	39.6	65.0
Pt→En	21.4	52.3	24.9	54.3	26.1	54.7	35.7	60.4	34.8	59.8	34.5	60.0	35.1	60.3	35.2	60.1
Pt→Es	24.7	58.9	29.9	61.3	30.2	61.7	41.8	67.1	41.5	67.0	41.2	67.3	41.4	67.2	41.3	67.2
El→En	4.6	33.8	7.1	35.1	7.4	35.9	17.6	44.6	17.3	44.4	16.4	43.6	17.6	44.2	17.2	44.1
Avg	19.4	51.7	23.2	53.3	23.9	54.0	33.9	59.5	*33.5*	*59.4*	32.7	58.9	**33.3**	59.2	**33.3**	**59.3**

Table 4: BLEU and chrF scores on noisy OCR data. See Table 3 for the legend.

Multimodal Fusion This setup integrates an additional fusion component, below the DA, to merge the three fine-tuned NAs. During training, we only fine-tune the multimodal fusion layer with a merge of the noisy multimodal datasets while keeping the rest frozen. In addition to our proposals (Multi Fusion and Fast Fusion), we also test Adapter Fusion (AF) (Pfeiffer et al., 2021) as a baseline.

Joint learning of domain and noise To compare our solution with previously proposed joint learning of noise and domain with Adapter Layers, we also train a single Adapter Layer tuned on all types of data, clean and noisy, as in (Berard et al., 2020).

Real vs synthetic noise To check the differences between realistic and synthetic noise, we also train an Adapter layer with basic random noise injection as in (Berard et al., 2020): such as common spelling errors, punctuation substitutions, letter swaps, spaces around punctuation, and accent removal.

Compositionality To test the ability to compose our DAs and NAs, we also train a DA on another domain Covost2 (Wang et al., 2020), which is a speech translation dataset created from Common Voice. The adapter layer is trained only on clean Covost data. We test this behavior when composed with NAs trained only on mTEDx noisy data. The test is performed on noisy Covost data created in the same way as the noisy mTEDx data. See Figure 3.

4 Results

Tables 3, 4, 5 show our evaluation results, BLEU and chrF, on each type of noise. In the following sections, we provide some qualitative analysis of these metrics.

4.1 Impact of multimodal noise

As seen in Table 6, on clean data, as expected, the DA model performs better than the Generic NMT model (+3.28 BLEU).

We can also easily see the impact of noisy data on these strong baselines: both suffer severely from noise with losses going from (-

Noise[UGC] Testsets	Multilingual NMT (out-of-domain)		DA (Clean data)		DA (Synthetic Noise)		DA (Real Noise)		DA + ASR NA		DA + AF		DA + FF		DA + MF	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
Fr→En	29.9	54.1	32.2	55.1	35.9	57.9	40.6	61.6	40.9	61.7	39.6	60.9	39.3	60.8	39.2	60.6
Fr→Es	29.9	56.6	34.9	59.1	37.7	61.6	42.2	65.0	42.5	64.9	41.6	64.4	42.3	64.9	41.8	64.8
Fr→Pt	25.3	54.1	32.4	58.3	35.3	60.8	41.8	65.7	41.1	65.0	41.5	65.7	41.3	65.3	41.2	65.3
It→En	22.9	49.6	25.0	51.1	27.9	53.7	31.7	57.0	31.9	57.0	31.2	56.5	30.9	56.5	31.3	56.7
It→Es	29.7	56.8	33.4	59.2	36.4	61.6	40.4	64.8	41.0	65.2	37.9	63.3	39.2	63.9	40.3	64.7
Es→En	25.3	49.7	27.5	51.4	30.0	54.1	34.1	57.6	34.0	57.3	34.4	57.8	33.8	57.4	33.7	57.5
Es→Fr	25.8	51.8	29.1	53.6	32.1	56.3	34.0	58.4	35.3	59.3	32.0	56.7	33.3	57.8	34.3	58.6
Es→It	26.0	54.1	28.1	55.3	31.0	58.1	34.5	60.9	34.6	61.1	33.2	59.8	33.0	60.1	33.7	60.1
Es→Pt	30.0	57.9	35.2	60.3	39.5	62.9	44.9	67.5	43.8	66.8	44.2	67.2	44.8	67.2	44.7	67.2
Pt→En	27.5	52.9	30.4	54.9	34.0	58.0	39.0	62.0	38.4	61.7	38.6	61.9	38.7	61.6	37.9	61.4
Pt→Es	32.2	60.6	36.4	62.8	40.6	65.6	45.7	69.5	45.8	69.4	45.7	69.5	45.8	69.5	45.6	69.4
El→En	31.5	55.3	33.1	56.4	33.4	56.5	32.3	55.7	32.9	56.3	28.9	52.9	30.8	54.4	31.7	55.0
Avg	28.0	54.5	31.5	56.5	34.5	58.9	38.4	62.1	*38.5*	*62.1*	37.4	61.4	37.8	61.6	**38.0**	**61.8**

Table 5: BLEU and chrF scores on noisy UGC data. See Table 3 for the legend.

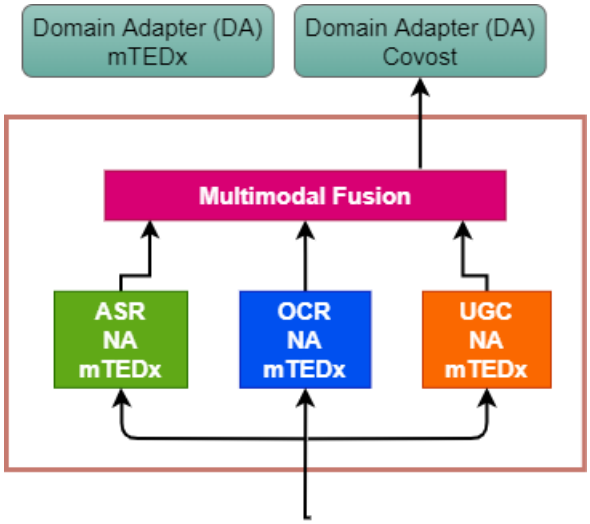

Figure 3: Transfer Learning (new domain). Noise adapters trained on a specific Noise type and Domain can be combined with other domain adapters. In this case, mTEDx NAs can be used with a Covost DA.

23.84) to (-12.09) BLEU points depending on the noise type. In our case, the ASR system seems to be the most noisy, followed by the OCR output and finally the UGC perturbations seem to have the less impact in terms of BLEU loss.

Finally, in the same table, we can also observe that the DA model is still in average better than the Generic model when faced with noisy multimodal data (+4.2 BLEU in average).

4.2 Noise Adaptation efficiency

As seen in 3, 4, 5, all the dedicated noise adaptation methods outperform the Generic NMT model and the DA model.

If we look into the details, for example in Table 3, for the ASR noise, we can see that the single DA trained with Synthetic Noise is outperformed by the DA trained with the Real

Data	Generic NMT BLEU($\Delta Clean$)	DA NMT BLEU($\Delta Clean$)	ΔDA
Clean	40.37 (+0.00)	43.65 (+0.00)	+3.28
ASR	16.53 (-23.84)	20.09 (-23.56)	+3.56
OCR	20.18 (-20.19)	25.20 (-18.45)	+5.02
UGC	27.53 (-12.84)	31.56 (-12.09)	+4.03

Table 6: Impact of noisy data, in terms of BLEU, on the Generic NMT model and on the Domain Adapted model who saw only clean data.

Noise (+3.4 BLEU). This confirms previous observations like the ones done by (Specia et al., 2020a). The same conclusion can be made for the other types of noise: OCR Table 4 and UGC Table 5.

If we look at our proposal of learning separately the DA and the NA, we can observe that our method is actually competitive with the previously proposed state-of-the-art methods, that jointly learn domain and noise. We have in average the same exact quality, but with the added benefice of easy extension (to new domains or types of noise) and the ability to handle inputs with unknown type of noise (see analysis in 4.3).

4.3 Multimodal Fusion mechanisms

Before looking at the Transfer Learning abilities of our architecture, to new domains and types of noise, let's look on how we deal with input containing an unknown type of noise.

If we look in Table 3, when we don't know the exact type of noise in input, we can see that Adapter Fusion, the current state-of-the-art on fusing Adapter Layers, loses (-0.5) BLEU when compared with the oracle system (where we choose the right Adapter for the type of input). Our solutions Fast Fusion and Multi

Fusion both obtain better results with a relative improvement of (+0.3 BLEU) and (+0.5) when compared with AF. The MF solution actually is as good as selecting the oracle Adapter Layer.

The same observation can be done on the other types of noises: FF and MF outperform the AF technique and bridge the gap with the Oracle systems (with an average of 0.16 BLEU points difference only).

4.4 Transfer Learning (new domain)

As seen in Table 7, when we train separately a new domain adapter, on clean data from Covost (Wang et al., 2020), like for the observations on mTEDx, the Domain Adapted model suffers from noisy input (-13.9 BLEU). But, when we combine this new DA with a previously trained NA (that was trained only on Noisy mTEDx data), we observe that we gain back most of the losses: to 46.6 BLEU on noisy covost data compared to 47.8 on clean covost data (only -1.2 BLEU points loss).

It shows that our method allows to easily extend the model to new domains while still being able to deal with specific, already known, types of noises. This type of extension being a lot more costly for all the methods doing a joint learning of domain and noise. For example, in case of joint learning, to handle 6 domains and 4 types of noises (UGC, OCR, UGC and Clean) one needs to train $6 \times 4 = 24$ adapter layers, while our solution only necessitates $6 + 4 = 10$ adapter layers to provide a some level of robustness to all domains.

Model	Clean covost	Noisy covost
DA(Covost)	47.8	33.9
DA(Covost) + NA(mTEDx)	47.8	46.6

Table 7: Transfer Learning (new domain). DA (Covost) Adapter Layer trained on clean covost data. NA (mTEDx) trained on noisy mTEDx data.

4.5 Transfer Learning (new noise)

Finally, to check the ability of our fusion mechanisms, FF and MF, to deal with an unknown type of noise, we evaluate them on a synthetic type of noise (different from UGC, OCR and ASR). This synthetic type of noise, similar to (Berard et al., 2020), consists of punctuation substitutions, letter swaps, spaces around punctuation, accent removal, etc.

As seen in Table 8, like before, the Generic NMT model and the DA model suffer from this new type of noise. When trying to deal with this new type of noise, without retraining any of our components, we observe again the FF and MF both outperform AF.

4.6 Convergence speed

As a last observation, we can see in Figure 4, that both FF and MF converge faster in terms of training steps than AF, giving us good results after only a few hundred steps of tuning.

5 Related Work

5.1 Robustness Task for NMT

Previous works have made several attempts to handle noise (Li et al., 2019b; Specia et al., 2020b). Data augmentation is used to generate more noisy training sentences, by injecting synthetic noise to emulate specific types of noise (Khayrallah and Koehn, 2018; Lui et al., 2019; Vaibhav et al., 2019; Karpukhin et al., 2019; Berard et al., 2020), back-translating data is also used to create artificial noise (Li and Specia, 2019; Zheng et al., 2019; Helcl et al., 2019; Post and Duh, 2019), and injecting made-up words breaks the text naturalness (Xu et al., 2021). In addition to data augmentation, Berard et al. (2019b); Murakami et al. (2019) apply data cleaning techniques in order to filter noisy data in a preprocessing setup to avoid catastrophic failures. Other works (Sperber et al., 2017; Cheng et al., 2018, 2019, 2020) propose adversarial methods to synthesize adversarial attacks in the training data. Michel and Neubig (2018); Murakami et al. (2019); Helcl et al. (2019); Alam and Anastasopoulos (2020); Berard et al. (2020) use various fine-tuning/adaptation techniques to help with specific types of noise.

5.2 Adapter Layers

Recent works has studied Adapter Layers (Houlsby et al., 2019) for various types of tasks. In computer vision, Rebuffi et al. (2017, 2018) introduce residual adapters for learning visually-diverse domains. In NLP, Stickland and Murray (2019); Pilault et al. (2020); Pfeiffer et al. (2021) mix adapters and multi-task learning for natural language understanding

Noise[Synthetic] Testsets	Multilingual NMT		DA		DA + AF		DA + FF		DA + MF	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
Fr→En	39.1	61.3	41.1	62.3	41.2	62.4	41.2	62.6	41.6	62.8
Fr→Es	39.3	64.0	44.0	66.2	44.1	66.0	44.6	66.5	43.9	66.5
Fr→Pt	33.0	61.3	42.1	66.0	43.5	67.2	43.1	66.9	42.8	66.9
It→En	32.0	58.3	34.1	59.3	34.1	59.2	34.1	59.2	34.2	59.3
It→Es	40.8	66.4	42.8	67.7	41.7	66.2	42.6	66.8	43.3	67.5
Es→En	32.2	56.6	34.5	58.2	34.8	58.4	35.1	58.7	35.1	58.9
Es→Fr	33.1	59.0	36.5	60.2	32.7	57.4	34.9	59.1	36.2	60.1
Es→It	32.9	61.0	35.0	60.9	34.3	60.4	34.2	60.7	34.9	61.4
Es→Pt	38.9	64.8	44.1	67.4	44.9	67.8	45.7	68.1	45.8	68.4
Pt→En	37.0	61.8	39.9	63.2	40.1	63.3	40.7	63.6	40.4	63.4
Pt→Es	42.5	68.4	46.8	70.5	47.0	70.8	48.1	71.1	47.9	71.2
El→En	29.4	53.6	30.8	54.5	27.6	51.2	29.5	52.9	30.0	53.5
Avg	35.9	61.4	39.3	63.0	38.8	62.5	39.5	63.0	**39.7**	**63.3**

Table 8: Evaluation of transfer learning abilities of the fusion mechanisms to an unknown type of noise (synthetic). AF/FF/MF are trained on UGC/OCR/ASR data and tested on unknown Synthetic noise.

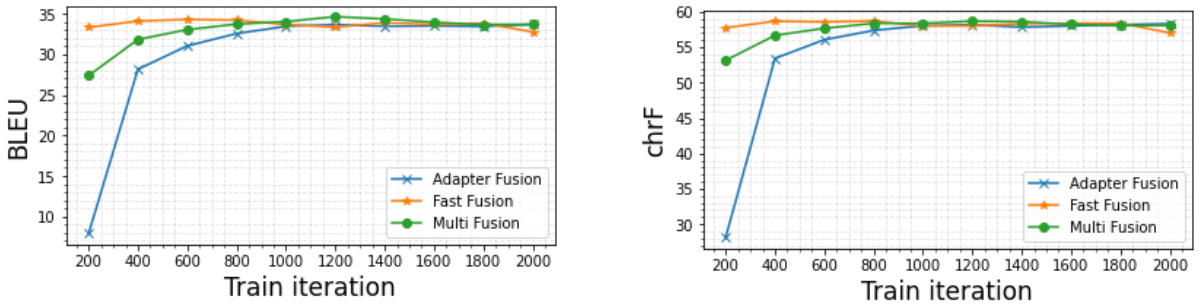

Figure 4: Validation performance vs train steps when adapting multimodal fusion layer with AF/FF/MF.

(NLU) tasks; Lin et al. (2020) exploit adapters to language generation tasks; Pfeiffer et al. (2020) propose an adapter-based framework for cross-lingual transfer; Ustun et al. (2020) apply adapter to dependency parsing. In speech processing, adapters are mostly used in ASR tasks (Kannan et al., 2019; Lee et al., 2021; Winata et al., 2021). Recently, they have also been explored for speech translation task (Escolano et al., 2021; Li et al., 2021a).

For NMT, Bapna et al. (2019) initially apply task specific adapter layers for multilingual NMT. Then, Philip et al. (2020); Stickland et al. (2021); Ustun et al. (2021) train adapters with different motivations: zero-shot NMT, cross-lingual transfer, and unsupervised NMT.

5.3 Multimodal Translation

Many NLP tasks benefit from multimodal integration, such as spoken language translation (Akiba et al., 2004), visual question answering (Agrawal et al., 2015), image captioning (Bernardi et al., 2016), multimodal sentiment analysis (Zadeh et al., 2016), image-guided translation (Zhao et al., 2020, 2022a,b). These works indicate that multimodal sensory integration is an important aspect of information processing and reasoning in NLP.

In contrast, multimodal robustness for text-to-text NMT remains relatively less explored. Recently, Salesky et al. (2021a) transform texts as images followed by OCR to cover some cases of noise for the robustness of open-vocabulary translation. Li et al. (2021b) combines an adversarial training on artificial noise with an image-guided machine translation model for translation robustness.

6 Conclusion

We propose an architecture to deal with the robustness problem in case of multimodal data for text-to-text NMT. Our solution is able to deal with realistic noise coming from a speech signal (via ASR processing), from an image (via OCR processing) or from noisy text as found in UGC on the web.

Our method proposes to first decompose the Domain Adaptation and Noise Adaptation learning tasks. In a second step, we show how we can dynamically recompose the specific DA and NA layers for handling, in the same model, various types of noise.

Finally, we also show how we can dynamically fuse the knowledge from these various adapters to provide robust translations, when the source of noise is unknown, when we have a new incoming domain or a new incoming source of noise.

Limitations

While we show some good capacities of this architecture to deal with unknown type of noise and new domains, via our new FF and MF mechanisms, we still lack a global mechanism to actually fully integrate both DA and NA. A question to be asked is: can we actually build a fusion mechanism on both levels Domain and Noise? Current attention mechanisms, like our multi-head attention in MF, do not support seamlessly this stack of Adapter Layers. So for now, we are limited to build a fusion mechanism for NAs and a separate one for DAs.

Also, on these new fusion mechanisms, while they perform better than Adapter Fusion, we should probably dig further to see if they generalize well to other NMT tasks (like domain adaptation for example) or to other NLP tasks (like the ones for which the Adapter Fusion was actually created).

Finally, while we believe our setting is a very realistic one, because most of the current multimodal NMT systems work in a pipeline way, it's not clear if our solutions will be of any use to fully multimodal systems working for example directly from the raw signal for speech.

Ethics Statement

We ensure that our work is conformant to the ACM Code of Ethics.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4–31.

Yasuhiro Akiba, MarcelloFederico, N. Kando, Hiromi Nakaiwa, and Junichi Tsujii. 2004. Overview of the iwslt evaluation campaign. In *IWSLT*.

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2020. Fine-tuning mt systems for robustness to second-language speaker variations. In *WNUT*.

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. *CoRR*, abs/1904.01941.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *ACL*.

Ankur Bapna, N. Arivazhagan, and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *EMNLP*.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. In *ICLR*.

Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. 2019a. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *WMGT*.

Alexandre Berard, Ioan Calapodescu, Vassilina Nikoulina, and Jerin Philip. 2020. Naver labs europe's participation in the robustness, chat, and biomedical tasks at wmt 2020. In *WMT*.

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019b. Naver labs europe's systems for the wmt19 machine translation robustness task. In *WMT*.

Raffaella Bernardi, Ruken Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.*, 55:409–442.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *ACL*.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. In *ACL*.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *ACL*.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Haoyue Shi, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari, Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, M. Yee, Jing Zhang, and Yue Zhang. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation.

J. Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *COLING*.

Carlos Escolano, Marta Ruiz Costa-jussà, José A. R. Fonollosa, and Carlos Segura. 2021. Enabling zero-shot multilingual spoken language translation with language-specific encoders and decoders. In *ASRU*.

Markus Freitag and Orhan Firat. 2020. Complete multilingual neural machine translation. In *WMT*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*.

Jindřich Helcl, Jindřich Libovický, and Martin Popel. 2019. Cuni system for the wmt19 robustness task. In *WMT*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *ICML*.

Anjuli Kannan, Arindrima Datta, Tara N. Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Z. Chen, and Seungjin Lee. 2019. Large-scale multilingual speech recognition with a streaming end-to-end model. In *INTERSPEECH*.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In *EMNLP*.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *NMT@ACL*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *NMT@ACL*.

Taewoo Lee, Min-Joong Lee, Tae Gyoon Kang, Seokyeong Jung, Minseok Kwon, Yeona Hong, Jungin Lee, Kyoung-Gu Woo, Ho-Gyeong Kim, Jiseung Jeong, Jihyun Lee, Hosik Lee, and Young Sang Choi. 2021. Adaptable multi-domain language model for transformer asr. *ICASSP*, pages 7358–7362.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019a. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation*.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Philipp Koehn, Philipp Koehn, Graham Neubig, Juan Miguel Pino, and Hassan Sajjad. 2019b. Findings of the first shared task on machine translation robustness. In *WMT*.

Xian Li, Changhan Wang, Yun Tang, C. Tran, Yuqing Tang, Juan Miguel Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021a. Multilingual speech translation from efficient finetuning of pretrained models. In *ACL*.

Zhenhao Li, Marek Rei, and Lucia Specia. 2021b. Visual cues and error correction for translation robustness. In *EMNLP*.

Zhenhao Li and Lucia Specia. 2019. Improving neural machine translation robustness via data augmentation: Beyond back-translation. In *EMNLP*.

Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. In *Findings of EMNLP*.

Alison Lui, Antonios Anastasopoulos, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. In *NAACL*.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *NAACL*.

Paul Michel and Graham Neubig. 2018. Mtnt: A testbed for machine translation of noisy text. In *EMNLP*.

Soichiro Murakami, Makoto Morishita, Tsutomu Hirao, and Masaaki Nagata. 2019. Ntt's machine translation systems for wmt19 robustness task. In *WMT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *EACL*.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *EMNLP*.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *EMNLP*.

Jonathan Pilault, Amine Elhattami, and Christopher Joseph Pal. 2020. Conditionally adaptive multi-task learning: Improving transfer learning in nlp using fewer parameters & less data. In *ICLR*.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *WMT*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *WMT*.

Matt Post and Kevin Duh. 2019. Jhu 2019 robustness task system description. In *WMT*.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *NIPS*.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2018. Efficient parametrization of multi-domain deep neural networks. In *CVPR*.

Elizabeth Salesky, David Etter, and Matt Post. 2021a. Robust open-vocabulary translation from visual text representations. In *EMNLP*.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021b. The multilingual tedx corpus for speech recognition and translation. In *INTERSPEECH*.

Baoguang Shi, Xiang Bai, and Cong Yao. 2015. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *CoRR*, abs/1507.05717.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020a. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*.

Lucia Specia, Zhenhao Li, Juan Miguel Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020b. Findings of the wmt 2020 shared task on machine translation robustness. In *WMT*.

Matthias Sperber, Jan Niehues, and Alexander H. Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *IWSLT*.

Asa Cooper Stickland, Alexandre Berard, and Vassilina Nikoulina. 2021. Multilingual domain adaptation for nmt: Decoupling language and domain information with adapters. In *WMT*.

Asa Cooper Stickland and Iain Murray. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *ICML*.

A. Ustun, Alexandre Berard, Laurent Besacier, and Matthias Gallé. 2021. Multilingual unsupervised neural machine translation with denoising adapters. In *EMNLP*.

A. Ustun, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. Udapter: Language adaptation for truly universal dependency parsing. In *EMNLP*.

Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. In *NAACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Changhan Wang, Anne Wu, and Juan Miguel Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus. *ArXiv*, abs/2007.10310.

Genta Indra Winata, Guangsen Wang, Caiming Xiong, and Steven C. H. Hoi. 2021. Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition. In *INTERSPEECH*.

Weiwen Xu, Ai Ti Aw, Yang Ding, Kui Wu, and Shafiq R. Joty. 2021. Addressing the vulnerability of nmt in input perturbations. In *NAACL*.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *IEEE Intelligent Systems*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Double attention-based multimodal neural machine translation with semantic image regions. In *EAMT*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022a. Region-attentive multimodal neural machine translation. *Neurocomputing*, 476:1–13.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022b. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:244–259.

Renjie Zheng, Hairong Liu, Mingbo Ma, Baigong Zheng, and Liang Huang. 2019. Robust machine translation with domain sensitive pseudo-sources: Baidu-osu wmt19 mt robustness shared task system report. In *WMT*.