NAIST's Machine Translation Systems for IWSLT 2020 Conversational Speech Translation Task

Ryo Fukuda[1], **Katsuhito Sudoh**[1], **Satoshi Nakamura**[1,2]
[1]Nara Institute of Science and Technology
[2]AIP Center, RIKEN, Japan
{fukuda.ryo.fo3, sudoh, s-nakamura}@is.naist.jp

Abstract

This paper describes NAIST's NMT system submitted to the IWSLT 2020 conversational speech translation task. We focus on the translation disfluent speech transcripts that include ASR errors and non-grammatical utterances. We tried a domain adaptation method by transferring the styles of out-of-domain data (United Nations Parallel Corpus) to be like in-domain data (Fisher transcripts). Our system results showed that the NMT model with domain adaptation outperformed a baseline. In addition, slight improvement by the style transfer was observed.

1 Introduction

Neural Machine Translation (NMT) has significantly improved the quality of Machine Translation (MT) (Bahdanau et al., 2014; Sutskever et al., 2014; Luong et al., 2015). However, domain-specific translation is still difficult in low-resource scenarios, although high performance can be achieved in resource-rich scenarios (Chu and Wang, 2018). Another major problem is the difficulty in translating noisy input sentences including filler, hesitation, etc. Belinkov and Bisk (2017) suggests the difficulty in learning to translate noisy sentences compared to clean ones. The translation of noisy sentences is very important for spoken language translation. In the IWSLT 2020 Conversational Speech Translation Task, we are going to tackle these two problems.

The task includes speech-to-text and text-to-text translation from disfluent Spanish speeches/transcripts to fluent English text. We chose the text-to-text subtask for our challenge task participation. The data for this task consists of about 130K bilingual pairs, would not be enough to learn a highly accurate NMT (Koehn and Knowles, 2017). In such a low-resource scenario, one promising way is domain adaptation using out-of-domain parallel corpora and in-domain monolingual corpora (Wang et al., 2016; Chu et al., 2017).

In domain adaptation, the "similarity" between in-domain and out-of-domain data affects the translation accuracy significantly (Koehn and Knowles, 2017). A domain can be defined by any property of the training data such as topic and style. We expect that the domain similarity comes from these properties.

Let us return to the task description. In the task, the inputs are conversational speech transcripts by Automatic Speech Recognition (ASR). They can include ASR errors as well as disfluent and non-grammatical utterances in spontaneous speech. In contrast, the outputs are fluent sequences. In other words, the purpose of this task is to translate disfluent transcripts into fluent sentences. As mentioned before, domain adaptation is a common practice in a low-resource scenario. However, it is difficult to prepare external parallel data in a disfluent source language and a fluent target language. Although fluent written parallel data are widely available, the effects of training with them are limited because the style of the input sentences differs from the in-domain data. We need a new strategy for training that can effectively use out-of-domain data with low similarity to in-domain data.

In this paper, we propose a novel domain adaptation method through style transfer of out-of-domain data using unsupervised machine translation. We increase the similarity between out-of-domain and in-domain data by transferring out-of-domain fluent input sentences into disfluent styles. This enables effective domain adaptive training and provides a robust NMT system for noisy input sentences.

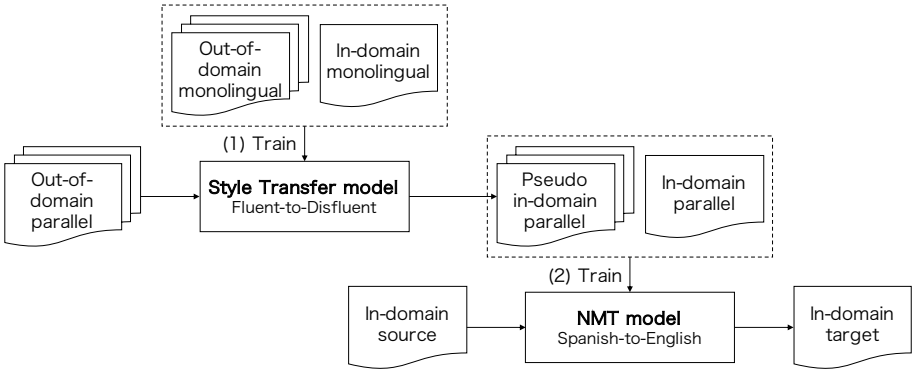

Figure 1: Overview of the proposed method.

2 System Details

Our method consists of two components: (1) Style Transfer model from fluent to disfluent Spanish. (2) Translation model from disfluent Spanish to fluent English, as illustrated in Figure 1. First, we transferred fluent Spanish in out-of-domain data into disfluent Spanish (Section 2.1). Then we trained the NMT model leveraging both out-of-domain parallel data as well as in-domain parallel data (Section 2.2).

2.1 Unsupervised Style Transfer

We employed an unsupervised learning method for the style transfer of Spanish of out-of-domain data. This is because there is no parallel corpus of fluent and disfluent Spanish and it is not possible to adapt supervised learning methods. Artetxe et al. (2018); Lample et al. (2018a,b) proposed Unsupervised Neural Machine Translation (UNMT) that learns the translation using monolingual corpora of two languages. In this system, we built a fluent-to-disfluent style transfer model based on UNMT with out-of-domain fluent data and in-domain disfluent data.

2.2 Domain Adaptation

For the challenge task, we apply fine-tuning, which is one of the conventional domain adaptation methods of MT (Sennrich et al., 2016a). The fine-tuning can result in significant improvements compared to both only in-domain training or only out-of-domain training (Dakwale and Monz, 2017). In this method, an NMT is pre-trained on a resource rich out-of-domain data until convergence, and then its parameters are fine-tuned on a low-resource in-domain data.

In this study, we pre-trained the NMT model on the pseudo in-domain data generated in 2.1, and then fine-tuned on true in-domain data.

Table 1: The number of sentence pairs of the data.

	# sentences
Fisher/Train	138,720
Dev	3,977
Test	3,641
UNCorpus/Train	1,000,000
Dev	4,000
Test	4,000

3 Results

3.1 Datasets

We used the LDC Fisher Spanish speech (disfluent) with new English translations (fluent) (Post et al., 2013; Salesky et al., 2018) as parallel in-domain data and the United Nations Parallel Corpus (UNCorpus) (Ziemski et al., 2016) as parallel out-of-domain data.

Fisher has the following multi-way parallel data distributed by the task organizer:

1. Spanish disfluent speech
2. Spanish disfluent transcripts (gold)
3. Spanish disfluent transcripts (ASR output)
4. English disfluent translations
5. English fluent translations

When training, we used (3) as input and (4) or (5) as output. UNCorpus consists of manually translated UN documents of the 25 years (1990 to 2014) for the six official UN languages, Arabic, Chinese, English, French, Russian, and Spanish. For our submission, one million Spanish-English bilingual sentence pairs were chosen randomly and used as out-of-domain data. Data statistics are shown in Table 1.

3.2 Spanish Style Transfer

3.2.1 Experimental Settings

Data We trained the style transfer from fluent to disfluent sentences using both Fisher and UNCorpus Spanish data. We preprocessed the data with Byte Pair Encoding (Sennrich et al., 2016b) to split sentences into subwords. The vocabulary size was set to 32,000 and sentences longer than 175 subwords were excluded from the training. We apply lowercasing and punctuation removal to UNCorpus same as Fisher corpus.

Model We used the implementation of UNMT[1] by Lample et al. (2018b). UNMT model was based on Transformer (Vaswani et al., 2017). Our models follow the suggested parameters from implementation of UNMT. We used three-layer shared encoder and shared decoder. We set the word embedding dimensions, hidden state dimensions, feed-forward dimensions to 512, 512, and 2048, respectively. We employed eight attention heads for both the encoder and the decoder. We chose Adam (Kingma and Ba, 2014) with a learning rate of 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$ as the optimizer. Each mini-batch contained 16 sentences.

In order to gain robustness to the content of the sentence, we first pre-trained the model using only UNCorpus/Train. During pre-training, early stopping was applied on the BLEU score between source sentences and back-translated sentences of the UNCorpus/Dev with a patience of 10 iterations, and the model with the highest score was stored. After that, additional training of 1 iteration using the Fisher/Train was performed.

Evaluation Axelrod et al. (2011) used a language model of in-domain data for out-of-domain data selection in domain adaptation. Following this study, we estimated the similarity between domains by measuring the perplexity (PPL) of the training set W of the out-of-domain data using a 3-gram language model M made from the in-domain data (Equation 1).

$$PPL = 10^{H(W|M)} \quad (1)$$

$H(W|M)$ is the entropy, defined as the average of the negative log-likelihood per token, as shown in the following equation:

$$H(W|M) = \frac{1}{|W|} \sum_{s \in W} -\log_{10} P(s|M) \quad (2)$$

[1]https://github.com/facebookresearch/UnsupervisedMT

Table 2: Perplexity and the number of unknown words (# UNK) for Fisher/train in the 3-gram language model.

Training data	perplexity	# UNK
Fisher	72.46	0
UNCorpus	589.81	5,173,539
Fisher-like UNCorpus	474.47	4,217,819

$P(s|M)$ is the probability of sentence s in the language model M. We used the SRI Language Modeling Toolkit to build the language model[2].

3.2.2 Results

Table 2 shows the perplexity of the language model for the Fisher/train. By transferring the fluent UNCorpus into the disfluent Fisher tone (Fisher-like UNCorpus) reduced the perplexity and number of unknown words.

3.3 NMT with Domain Adaptation

We trained the NMT models which translate from disfluent Spanish to fluent English.

3.3.1 Experimental Settings

Data For training data, we used Fisher/train as in-domain data and UNCorpus/Train and Fisher-like UNCorpus/Train as out-domain data. Fisher-like UNCorpus has the same number of sizes as UNCorpus. During training, we used Fisher/Dev as a validation set. Fisher/Test was used for evaluation. We preprocessed the data in the same way as in the previous experiment. However, for practical use, lowercasing and punctuation removal were applied only to the source language.

Model We used OpenNMT-py[3]. The NMT model was based on Transformer. The hyperparameters of the model almost follow the *transformer_base* settings (Vaswani et al., 2017). Note that in the Fisher-only experiment without domain adaptation, the batch size was halved to 2048 tokens. The model was trained for 20,000 iterations using out-of-domain data, and then fine-tuned for 1,000 iterations using in-domain data. The model parameters saved every 100 iterations.

Evaluation To evaluate the performance, we calculated the BLEU scores (Papineni et al., 2002) with sacreBLEU[4].

[2]http://www.speech.sri.com/projects/srilm/
[3]https://github.com/OpenNMT/OpenNMT-py
[4]https://github.com/mjpost/sacreBLEU

Table 3: BLEU scores of trained NMT models for Disfluent Spanish to Fluent English.

System	Fisher/Test
Fisher	14.8
UNCorpus	7.8
Fisher-like UNCorpus	6.7
UNCorpus + Fisher	18.3
Fisher-like UNCorpus + Fisher	**18.5**

Table 4: BLEU scores for Disfluent Spanish to Fluent English. NMT models used Fisher's disfluent references for training.

System	Fisher/Test
Fisher	11.6
UNCorpus + Fisher	15.2
Fisher-like UNCorpus + Fisher	**15.6**

3.3.2 Results

Tables 3 and 4 show the BLEU scores of the systems evaluated with single fluent references. In Table 3, "Fisher", "UNCorpus" and "Fisher-like UNCorpus" are models trained on a single training data. "UNCorpus + Fisher" and "Fisher-like UNCorpus + Fisher" are models that were pre-trained on UNCorpus and Fisher-like UNCorpus and then fine-tuned on Fisher/Train, respectively. The models in Table 4 did not use Fisher's fluent references when training but instead used disfluent references.

Both with and without Fisher's fluent references, domain adaptation training outperformed the baseline. Furthermore, when the pseudo-disfluent Spanish generated by the style transfer was used for training, the score was better than the use of the original UNCorpus without the style transfer. We submitted six systems in total: "Fisher", "UNCorpus + Fisher" and "Fisher-like UNCorpus + Fisher" in Table 3, and all of Table 4.

4 Discussion

Effect of Style Transfer In domain adaptation training, the accuracy was slightly improved by transferring the style of out-of-domain data to be like in-domain data. This shows that there is some significance in increasing the similarity between domains through style transfer.

However, when we did not perform domain adaptation and only trained with out-of-domain data, the accuracy for in-domain data was reduced by style transfer. The following is an example of style transferred sentence:

nueva york 1 a 12 de junio de 2015 (original)
nueva york oh a mi eh de de de de (generated)

As shown above, some generated sentences lost the meaning of the sentence due to missing phrases. As a result, the quality of the parallel data decreased and the final translation performance was also degraded. One of the causes of this problem is style transfer constraints are too strong. Thus, it may be mitigated by a model that could control the trade-off between style transfer and content preservation (Niu et al., 2017; Agrawal and Carpuat, 2019; Lample et al., 2019).

Further improvement can be expected by preventing changes in the meaning of sentences and converting only the style.

Fluent vs Disfluent references The model trained using Fisher's original disfluent data had a BLEU score of about three points lower than the model trained using the fluent data. In other words, in this task, we found that removing the disfluency of reference sentences improves the BLEU by about three points for all the learning strategies we tried. In domain adaptation, we expected this problem to be mitigated by training on large out-of-domain data with fluent reference sentences, but the desired results were not obtained.

5 Conclusion

In this paper, we presented NAIST's submission to the IWSLT2020 Conversational Speech Translation task. We experimentally show that domain adaptation can improve the translation accuracy of disfluent sentences. Moreover, the translation accuracy was improved by increasing the similarity between domains through style transfer, but the effect was limited due to the parallel data quality degradation.

Furthermore, The loss of accuracy caused by not using clean reference sentences of in-domain data could not be resolved by domain adaptation either.

In future work, we will pursue a style transfer system that does not reduce the quality of the parallel data and use it to improve the translation accuracy of NMT. High-quality style transfer may allow us to acquire robustness to the disfluency of input sentences and to learn fluent outputs by removing the disfluency of output sentences.

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564, Hong Kong, China. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate.

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

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Praveen Dakwale and Christof Monz. 2017. Finetuning for neural machine translation with limited degradation across in-and out-of-domain data. In *the 16th Machine Translation Summit*, pages 156–169.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome spanish–english speech translation corpus. In *International Workshop on Spoken Language Translation*.

Elizabeth Salesky, Susanne Burger, Jan Niehues, and Alex Waibel. 2018. Towards fluent translations from disfluent speech. *2018 IEEE Spoken Language Technology Workshop (SLT)*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Rui Wang, Hai Zhao, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2016. Connecting phrase based statistical machine translation adaptation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3135–3145, Osaka, Japan. The COLING 2016 Organizing Committee.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on*

Language Resources and Evaluation (LREC 2016), pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).