On Knowledge Distillation for Direct Speech Translation

Marco Gaido[1,2], Mattia Antonino Di Gangi[3*], Matteo Negri[1], Marco Turchi[1]
[1]Fondazione Bruno Kessler, Trento, Italy
[2]University of Trento, Trento, Italy
[3]AppTek, Aachen, Germany
{mgaido,digangi,negri,turchi}@fbk.eu

Abstract

English. Direct speech translation (ST) has shown to be a complex task requiring knowledge transfer from its sub-tasks: automatic speech recognition (ASR) and machine translation (MT). For MT, one of the most promising techniques to transfer knowledge is knowledge distillation. In this paper, we compare the different solutions to distill knowledge in a sequence-to-sequence task like ST. Moreover, we analyze eventual drawbacks of this approach and how to alleviate them maintaining the benefits in terms of translation quality.

Italiano. *È stato dimostrato che la speech translation (ST) diretta è un'operazione complessa che richiede l'adozione di tecniche di knowledge transfer sia da automatic speech recognition (ASR) che da machine translation (MT). Per quanto riguarda MT, una delle tecniche più promettenti è la knowledge distillation (KD). In questo lavoro, confrontiamo diverse possibili soluzioni di KD per addestrare modelli sequence-to-sequence come quelli di ST. Inoltre, analizziamo eventuali problemi causati da questa tecnica e come attenuarli mantenendo i benefici in termini di qualità della traduzione.*

1 Introduction

Speech translation (ST) refers to the process of translating utterances in one language into text in a different language. Direct ST is an emerging paradigm that consists in translating without intermediate representations (Bérard et al., 2016; Weiss et al., 2017). It is a newer and alternative approach to cascade solutions (Stentiford and Steer, 1988; Waibel et al., 1991), in which the input audio is first transcribed with an automatic speech recognition (ASR) model and then the transcript is translated into the target language with a machine translation (MT) model.

The rise of the direct ST paradigm is motivated by its theoretical and practical advantages, namely: *i)* during the translation phase it has access to information present in the audio that is lost in its transcripts (eg. prosody, characteristic of the speaker[1]), *ii)* there is no *error propagation* (in cascade systems the errors introduced by the ASR are propagated to the MT, which has no cues to recover them), *iii)* the latency is lower (as data flows through a single system instead of two), and *iv)* the management is easier (as there is a single model to maintain and no integration between separate modules is needed).

On the downside, direct ST suffers from the lack of large ST training corpora. This problem has been addressed by researchers through transfer learning from the high-resource sub-tasks (Bérard et al., 2018; Bansal et al., 2019; Liu et al., 2019), multi-task trainings (Weiss et al., 2017; Anastasopoulos and Chiang, 2018; Bahar et al., 2019a), and the proposal of data augmentation techniques (Jia et al., 2019; Bahar et al., 2019b; Nguyen et al., 2020). In this work, we focus on the transfer learning from MT. The classic approach consists in pretraining the decoder with that of an MT model. Its benefit, however, is controversial: indeed, (Bahar et al., 2019a) showed that it is effective only with

[*]Work done during the PhD at Fondazione Bruno Kessler and University of Trento.

[1]For instance, the pitch of the voice is a cue for the sex of the speaker. Although the gender is a social aspect and does not depend on physical attributes, in many cases sex and gender coincide, so systems relying on this are likely to have a better accuracy than those that do not have access to any information regarding the speaker (Bentivogli et al., 2020; Gaido et al., 2020b).

the addition of an *adapter* layer, but this has not been confirmed in (Gaido et al., 2020a), while in (Inaguma et al., 2020) it always brought improvements. Another, more promising possibility consists in distilling knowledge from an MT model.

Knowledge distillation (KD) is a knowledge transfer technique introduced for model compression (Hinton et al., 2015). A small *student* model is trained computing the KL-divergence (Kullback and Leibler, 1951) with the output probability distribution of a big *teacher* model. Although KD was introduced in the context of image processing, its effectiveness suggested its adoption in other fields. Specifically, Liu et al. (2019) showed that using an MT system as teacher brings significant improvements to direct ST models. However, they did not compare the different methods to distill knowledge in sequence-to-sequence models (Kim and Rush, 2016) and they did not analyze possible negative effects of adopting this technique.

In this paper, we analyze different sequence-to-sequence KD techniques (*word level*, *sequence-level*, *sequence interpolation*) and their combination in the context of direct ST. Then, we study the effect of the best technique on a strong system trained on a large amount of data to reach state-of-the-art results. We show that word-level KD is the best approach and that fine-tuning the resulting model without KD brings further improvements. Finally, we analyze the limitations and the problems present in models trained with KD, which are partly solved by the final finetuning.

2 Sequence-level Knowledge Distillation

We focus on distilling knowledge from an MT model to an ST model. This is helpful due to the better results achieved by MT, which is an easier task than ST, as it does not involve the recognition of the audio content, and it also benefits from the availability of large training corpora. Our student (ST) model is trained to produce the same output distribution of the teacher (MT) model when the latter is fed with the transcript of the utterances passed as input to the ST model. As KD was introduced in the context of *classification* tasks, while ST and MT are sequence-to-sequence *generation* tasks, an adaptation is required for its application. Kim and Rush (2016) introduced three methods to distill knowledge in sequence-to-sequence models: *i)* word-level KD, *ii)* sequence-level KD, and *iii)* sequence interpolation.

Word-level KD (`Word-KD`) refers to computing the KL-divergence between the distribution of the teacher and student models on each token to be predicted. As recomputing the teacher output at each iteration is computationally expensive (it needs a forward pass of the MT model), we explored the possibility to pre-compute and store the teacher outputs. To this aim, we experimented with truncating the output distribution to have a lower memory footprint, as proposed in MT (Tan et al., 2019).

Sequence-level KD (`Seq-KD`) consists in considering as target the output generated by the teacher model using the beam search.

Sequence interpolation (`Seq-Inter`) is similar to Seq-KD, but the target is the sentence with the highest BLEU score (Papineni et al., 2002) with respect to the ground truth among the n-best generated by the beam search with the teacher model.

As done in (Kim and Rush, 2016), we also combine these methods to analyze whether they are complementary or not. Finally, we experiment with fine-tuning the model trained with KD on the reference translations.

3 Experimental Settings

We performed preliminary experiments on a limited amount of data to compare the three KD methods. Then, we created a model exploiting all the available corpora with the best technique to analyze the KD behavior in a real scenario.

3.1 Data

We first experiment using only Librispeech (Kocabiyikoglu et al., 2018), an ST corpus with English audio, transcripts and French translations. We use the *(audio, transcript)* pairs for the ASR pre-training, the *(transcript, translation)* pairs to train the MT teacher, and the *(audio, translation)* pairs for the ST training.

Then, we built an English-Italian model. In addition to Librispeech, the ASR pre-training involves TED-LIUM 3 (Hernandez et al., 2018), Mozilla Common Voice,[2] How2 (Sanabria et al., 2018) and the en-it section of MuST-C (Di Gangi et al., 2019a). The MT teacher is trained on the OPUS datasets (Tiedemann, 2016), cleaned using the ModernMT framework (Bertoldi et al., 2017).[3]

[2] `https://voice.mozilla.org/`

[3] With the `CleaningPipelineMain` class.

For ST, we use the en-it section of MuST-C and Europarl-ST (Iranzo-Sánchez et al., 2020).

We pre-process the input audio extracting a 40-dimensional feature vector from a span of 25 ms every 10 ms using Mel filter bank. During this pre-processing performed with XNMT (Neubig et al., 2018), we also apply speaker normalization. The text is tokenized and the punctuation is normalized with Moses (Koehn et al., 2007). We create 8,000 shared BPE merge rules on the MT data of each experiment and apply them to divide the text into sub-word units. Samples lasting more than 20 seconds are discarded in order to avoid out of memory issues during training.

3.2 Models

For ST and ASR we use the S-Transformer architecture (Di Gangi et al., 2019b; Di Gangi et al., 2019c) with logarithmic distance penalty in the encoder. In particular, in the experiments on Librispeech we train a small model using the basic configuration by Di Gangi et al. (2019b), while in the experiment with all the data we follow the BIG configuration. In the second case, we also slightly modify the architecture to improve performance by removing the 2D attention layers and changing the number of Transformer Encoder layers and Transformer Decoder layers to be respectively 11 and 4 in ST and 8 and 6 in the ASR pre-training (Gaido et al., 2020a). The different number of layers between ASR and ST is motivated by the idea of having adaptation layers (Jia et al., 2019; Bahar et al., 2019a).

For MT we use a Transformer with 6 layers for both the encoder and the decoder. In the preliminary experiments, we use a small model with 512 hidden features in the attention layers, 2,048 hidden units in the feed-forward layers and 8 attention heads; in the experiment with more data we double all these parameters.

3.3 Training

We optimize our models with Adam (Kingma and Ba, 2015) using betas (0.9, 0.98). The learning rate increases linearly for 4,000 steps starting from 1e-7 to 5e-3. Then it decays according to the inverse square root policy. In fine-tunings, the learning rate is fixed at 1e-4. A 0.1 dropout is applied and the total batch size is 64. When we do not use KD, the loss is label smoothed cross entropy (Szegedy et al., 2016) with 0.1 smoothing factor.

In the final training with all the data, we apply SpecAugment (Park et al., 2019) with probability 0.5, 13 *frequency masking pars*, 20 *time masking pars*, 2 *frequency masking num*, and 2 *time masking num*. We also increase the overall batch size to 512. Moreover, the ASR pre-training is performed as a multi-task training in which we add a CTC loss (predicting the output transcripts) on the encoder output (Kim et al., 2017).

Our code is based on the Fairseq library (Ott et al., 2019), which relies on PyTorch (Paszke et al., 2019), and it is available open source at `https://github.com/mgaido91/FBK-fairseq-ST`. The models are trained on 8 GPU K80 with 11 GB of RAM.

4 Results

First, we experiment truncating the output distribution generated by the teacher model. Table 1 shows that truncating the output to few top tokens does not affect significantly the performance. On the contrary, the best result is obtained using the top 8 tokens. Hence, all our experiments with `Word-KD` use the top 8 tokens of the teacher.

Top K	BLEU
4	16.43
8	**16.50**
64	16.37
1024	16.34

Table 1: Results with different K values, where K is the number of tokens considered for `Word-KD`.

Then, we try different values for the temperature T parameter. The temperature is a parameter used to sharpen (if $T < 1$) or soften (if $T > 1$) the output distribution. In particular, by adding the temperature, the $softmax$ function that converts the *logits* z_i into probabilities p_i becomes:

$$p_i = \frac{e^{z_i/T}}{\sum(e^{z_i/T})} \quad (1)$$

A higher temperature has been claimed to help learning the so-called *dark knowledge* (Hinton et al., 2015), one of the possible reasons alluded to justify the success of KD. Indeed, with a high temperature, the cost function is similar to minimizing the squared distance between the logits produced by the student and teacher networks. So logits with very negative values – which are basically ignored with low temperature – become important to be learnt by the student network. For a demon-

stration, please refer to (Hinton et al., 2015). Table 2 reports the BLEU score for different values of T and indicates that the default $T = 1$ is the best value. This result suggests that, in ST, the networks do not have the capacity of MT models trained on the same data. So focusing on the mode of the probability distribution works best.

T	BLEU
1.0	**16.50**
4.0	16.11
8.0	14.27

Table 2: Results with different temperatures (T).

	BLEU
Baseline	9.4
`Word-KD`	16.5
`Seq-KD`	13.4
`Seq-Inter`	13.3
`Seq-KD` + `Word-KD`	15.7
`Word-KD` + FT `Seq-KD`	16.7
`Seq-KD` + FT `Word-KD`	**16.8**
`Word-KD` + FT w/o KD	**16.8**

Table 3: Results of the small model on Librispeech with different KD methods and combining them in a single training or in consecutive trainings through a fine-tuning (FT).

Then, we compare the different sequence-level KD techniques. We also combine them either in the same training or in consecutive trainings through a fine-tuning (FT). The results are presented in Table 3. We can notice that all the methods improve significantly over the baseline: KD makes the training easier and more effective. Among them, `Word-KD` achieves the best results by a large margin. Combining it with another method in the same training is harmful (`Seq-KD` + `Word-KD`), while a fine-tuning on a different KD method or without KD (i.e. using the ground-truth target and label smoothed cross entropy) improves results by up to 0.3 BLEU (`Seq-KD` + FT `Word-KD` and `Word-KD` + FT w/o KD). These results confirm the choice by (Liu et al., 2019), but differ from those of (Kim and Rush, 2016). So, we can conclude that the best sequence-to-sequence KD technique is task-dependent and that the best option to distill knowledge from MT to ST is the word-level KD.

To validate the effectiveness of KD in a real case, we create a model translating English utterances into Italian text leveraging all the available corpora for each task. Our ASR pre-trained model scores 10.21 WER on the MuST-C test set, while the teacher MT model scores 30.3 BLEU on the Italian reference for same test set. We train our ST model first on the ASR corpora for which we generated the target with the MT model (resulting in a `Seq-KD` + `Word-KD` training). Note that we could not use this data without `Seq-KD` or `Seq-Inter`, hence we opted for the best training including one of them (`Seq-KD` + `Word-KD`). Second, we fine-tune the model on the ST corpora with `Word-KD`. Third, we fine-tune without KD as in the case leading to the best result (Table 3). So, our training is: `Seq-KD` + `Word-KD` (on ASR data) + FT `Word-KD` + FT w/o KD. After the first two steps, our ST model scores 22.8 BLEU on the MuST-C test set, while after the final fine-tuning the result is it scores 27.7 BLEU. This highlights the importance of fine-tuning without KD.

5 Analysis

We analyze the outputs of the en-it model to assess whether, despite the benefits in terms of translation quality, KD introduces limitations or issues. Namely, we checked whether the lack of access of the MT teacher to information present in the audio and not in the text (such as the gender[4] of the speaker) hinders the ability of the final model to exploit such knowledge. Moreover, we compared the output generated by the model before fine-tuning without KD and after it to determine the reasons of the significant BLEU improvement.

Direct ST systems have been shown to be able to exploit the audio to determine the gender of the speaker and reflect it better in the translations into languages rich of gender marked words (Bentivogli et al., 2020). This is not possible for an MT system that has no clue regarding the speaker's gender. We tested the performance of our models on the category 1 of the MuST-SHE test set (Bentivogli et al., 2020) (which contains gender marked word related to the speaker) to check whether distilling knowledge from MT harms this advantage of ST systems or not. Table 4 shows that, indeed, systems trained with KD inherit the bias from the MT system and, although the final fine-tuning mitigates the issue, the final model has a higher gender bias than a base ST system without KD (regarding the words related to the speaker).

The better translation of speaker's gender marked words does not explain the big BLEU im-

[4]This is true if the gender identity coincides with the biological sex. This assumption holds true in nearly all our data.

		Female			Male			Bias
	BLEU	Corr.	Wrong	Diff.	Corr.	Wrong	Diff.	Diff. M - Diff. F
Base ST (Bentivogli et al., 2020)	21.5	**26.7**	27.2	**-0.5**	*46.3*	6.8	*39.5*	**40.0**
MT	**30.3**	10.8	55.5	-44.7	**54.4**	7.1	**47.3**	92.0
`Seq-KD` + `Word-KD` + FT `Word-KD`	22.8	12.3	46.5	-34.2	45.4	8.1	37.3	71.5
+ FT w/o KD	*27.7*	19.8	39.0	-19.2	43.2	10.5	32.7	51.9

Table 4: Accuracy on Category 1 of the MuST-SHE test set of a base direct ST model and models created using KD. A high *Diff.* means that the model is able to recognize the speaker's gender and the gap between the *Diff.* on the two genders indicates the bias towards one of them. The reported BLEU score refers to the MuST-C test set and shows the translation quality of the model.

provement obtained with fine-tuning. Hence, we performed a manual analysis of sentences with the highest TER (Snover et al., 2006) reduction. The analysis revealed three main types of enhancements, with the first being the most significant.

Samples with multiple sentences. Some utterances contain more than one sentence. In this case, the model trained with KD tends to generate the translation of only the first sentence, ignoring the others. This is likely caused by the fact that MT training data is mostly sentence-level. For this reason, the MT model tends to assign a high probability of the `EOS` symbol after the dot. The student ST model learns to mimic this harmful behavior and, as in ST training and test samples often include more than one sentence, to wrongly truncate the generation once the first sentence is completed. The fine-tuned model, instead, generates all the sentences.

Verbal tenses. The fine-tuned model tends to produce the correct verbal tense, while before the fine-tuning the verbal tense is often not precise, likely because the MT model favors more generic forms. For instance, *"That meant I was going to be on television"* should be translated as *"Significava che **sarei andata** in televisione"*. The model before fine-tuning produces *"Questo significava che **stavo andando** in tv"* while the fine-tuned model uses the correct verbal tense *"Questo significava che **sarei andata** in televisione"*. Despite relevant for the final score, it is debatable whether this is a real improvement of the fine-tuned model, as in some cases both verbal tenses are acceptable or their correctness depends on the context (e.g. in informal conversations, the usage of conjunctive forms is often replaced with indicative tenses).

Lexical choices. In some cases, the fine-tuned model chooses more appropriate words, probably thanks to the fine-tuning on in-domain data. For instance, the reference translation for *"She has taken a **course** in a business school, and she has become a veterinary doctor"* is *"Ha seguito un **corso** in una scuola di business, ed è diventata una veterinaria"*. The corresponding utterance was translated by the model before the fine-tuning into *"Ha frequentato una **lezione** di economia ed è diventata una dottoressa veterinaria"*, while after the fine-tuning the translation is *"Ha frequentato un **corso** in una business school, ed è diventata una dottoressa veterinaria"*.

We can conclude that KD provides a benefit in terms of overall translation quality, but the resulting ST system also learns negative behaviors (such as the masculine default for the speaker-related words that exacerbates the gender bias). These are partly solved by performing a fine-tuning without KD, which keeps (and even enhances) on the other side the translation capabilities.

6 Conclusions

We presented and analyzed the benefits and issues brought by distilling knowledge from an MT system for direct ST models. We compared the different KD techniques and our experiments indicated that the best training procedure consists in a pre-training with word-level KD and a fine-tuning without KD. Then, we showed that KD from MT models causes an increased gender bias, omission of sentences in multi-sentential utterances and more generic word/verbal-tense choices. Finally, we demonstrated that a fine-tuning helps resolving these issues, although the exacerbation of gender bias is not solved, but only alleviated.

Acknowledgments

This work is part of the "End-to-end Spoken Language Translation in Rich Data Conditions" project,[5] which is financially supported by an Amazon AWS ML Grant.

[5] `https://ict.fbk.eu/units-hlt-mt-e2eslt/`

References

Antonios Anastasopoulos and David Chiang. 2018. Tied Multitask Learning for Neural Speech Translation. In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 82–91, New Orleans, Louisiana.

Parnia Bahar, Tobias Bieschke, and Hermann Ney. 2019a. A Comparative Study on End-to-end Speech to Text Translation. In *Proc. of International Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 792–799, Sentosa, Singapore.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019b. On Using SpecAugment for End-to-End Speech Translation. In *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, China.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on High-resource Speech Recognition Improves Low-resource Speech-to-text Translation. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 58–68, Minneapolis, Minnesota.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 6923–6933, Virtual.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. 2018. End-to-End Automatic Speech Translation of Audiobooks. In *Proc. of ICASSP 2018*, pages 6224–6228, Calgary, Alberta, Canada.

Nicola Bertoldi, Roldano Cattoni, Mauro Cettolo, et al. 2017. MMT: New Open Source MT for the Translation Industry. In *Proc. of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 86–91, Prague, Czech Republic.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. 2016. Listen and Translate: A Proof of Concept for End-to-End Speech-to-Text Translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*, Barcelona, Spain.

Mattia Antonino Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. MuST-C: a Multilingual Speech Translation Corpus. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, page 2012–2017, Minneapolis, Minnesota.

Mattia Antonino Di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessi, and Marco Turchi. 2019b. Enhancing Transformer for End-to-end Speech-to-Text Translation. In *Proc. of Machine Translation Summit XVII*, pages 21–31, Dublin, Ireland.

Mattia Antonino Di Gangi, Matteo Negri, and Marco Turchi. 2019c. Adapting Transformer to End-to-end Spoken Language Translation. In *Proc. of INTERSPEECH*, Graz, Austria, September.

Marco Gaido, Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2020a. End-to-End Speech-Translation with Knowledge Distillation: FBK@IWSLT2020. In *Proc. of the 17th International Conference on Spoken Language Translation*, pages 80–88, Virtual.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020b. Breeding Gender-aware Direct Speech Translation Systems. In *Proc. of The 28th International Conference on Computational Linguistics (COLING 2020)*, Virtual.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation. In *Proc. of the Speech and Computer - 20th International Conference (SPECOM)*, pages 198–208, Leipzig, Germany.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. In *Proc. of NIPS Deep Learning and Representation Learning Workshop*, Montréal, Canada.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-One Speech Translation Toolkit. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Virtual.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Giménez. Adrià, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A Multilingual Corpus For Speech Translation Of Parliamentary Debates. In *Proc. of ICASSP 2020*, pages 8229–8233, Barcelona, Spain.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging Weakly Supervised Data to Improve End-to-End Speech-to-Text Translation. In *Proc. of ICASSP 2019*, pages 7180–7184, Brighton, UK.

Yoon Kim and Alexander M. Rush. 2016. Sequence-Level Knowledge Distillation. In *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning. In *Proc. of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4835–4839, New Orleans, Louisiana.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proc. of 3rd International Conference on Learning Representations (ICLR)*, San Diego, California.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting librispeech with French translations: A multimodal corpus for direct speech translation evaluation. In *Proc. of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Philipp Koehn, Hieu Hoang, Alexandra Birch, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

Solomon Kullback and Richard Arthur Leibler. 1951. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-End Speech Translation with Knowledge Distillation. In *Proc. of Interspeech 2019*, pages 1128–1132, Graz, Austria.

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Sachan, Philip Arthur, Pierre Godard, John Hewitt, Rachid Riad, and Liming Wang. 2018. XNMT: The eXtensible Neural Machine Translation Toolkit. In *Proc. of the 13th Conference of the Association for Machine Translation in the Americas*, pages 185–192, Boston, MA.

Thai-Son Nguyen, Sebastian Stueker, Jan Niehues, and Alex Waibel. 2020. Improving Sequence-to-sequence Speech Recognition Training with On-the-fly Data Augmentation. In *Proc. of the 2020 International Conference on Acoustics, Speech, and Signal Processing – IEEE-ICASSP-2020*, Barcelona, Spain.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. of Interspeech 2019*, pages 2613–2617, Graz, Austria.

Adam Paszke, Sam Gross, Francisco Massa, et al. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proc. of Advances in Neural Information Processing Systems 32 (NIPS)*, pages 8024–8035. Curran Associates, Inc.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset For Multimodal Language Understanding. In *Proc. of Visually Grounded Interaction and Language (ViGIL)*, Montréal, Canada.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of the 7th Conference of the Association for Machine Translation in the Americas,*, pages 223–231, Cambridge.

Fred W. M. Stentiford and Martin G. Steer. 1988. Machine translation of speech. *British Telecom Technology Journal*, 6(2):116–122.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, Nevada, United States.

Xu Tan, Yi Ren, Di He, Tao Qin, and Tie-Yan Liu. 2019. Multilingual Neural Machine Translation with Knowledge Distillation. In *Proc. of International Conference on Learning Representations (ICLR)*, New Orleans, Louisiana, United States.

Jörg Tiedemann. 2016. Opus – parallel corpora for everyone. *Baltic Journal of Modern Computing*, page 384. Special Issue: Proc. of the 19th Annual Conference of the European Association of Machine Translation (EAMT).

Alex Waibel, Ajay N. Jain, Arthur E. McNair, Hiroaki Saito, Alexander G. Hauptmann, and Joe Tebelskis. 1991. JANUS: A Speech-to-Speech Translation System Using Connectionist and Symbolic Processing Strategies. In *Proc. of the International Conference on Acoustics, Speech and Signal Processing, ICASSP 1991*, pages 793–796, Toronto, Canada.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. In *Proc. of Interspeech 2017*, pages 2625–2629, Stockholm, Sweden.