The IWSLT 2019 Evaluation Campaign

J. Niehues(1) *R. Cattoni*(2) *S. Stüker*(3) *M. Negri*(2) *M. Turchi*(2) *T. Ha*(3)
E. Salesky(4) *R. Sanabria*(5) *L. Barrault*(6) *L. Specia*(6) *M. Federico*(2)

(1) DKE, Maastricht University, Netherlands
(2) FBK - Via Sommarive 18, 38123 Trento, Italy
(3) KIT - Adenauerring 2, 76131 Karlsruhe, Germany
(4) Center for Language & Speech Processing, Johns Hopkins University, USA
(5) School of Informatics, University of Edinburgh, U.K.
(6) Department of Computing, Imperial College London, U.K.

Abstract

The IWSLT 2019 evaluation campaign featured three tasks: speech translation of (i) TED talks and (ii) How2 instructional videos from English into German and Portuguese, and (iii) text translation of TED talks from English into Czech. For the first two tasks we encouraged submissions of end-to-end speech-to-text systems, and for the second task participants could also use the video as additional input. We received submissions by 12 research teams. This overview provides detailed descriptions of the data and evaluation conditions of each task and reports results of the participating systems.

1. Introduction

This report informs on the evaluation campaign organized by the 16th International Workshop on Spoken Language Translation (IWSLT). Spoken language translation (SLT) is the problem of translating speech input from a source to a target language. In its most general variant, named speech-to-speech translation, the goal of SLT is to produce spoken output in the target language. However, since the begin, mainly due to the difficulty in developing and evaluating speech-to-speech translation systems, the scope of IWSLT has been limited to speech-to-text translation systems. Hence, IWSLT's focus is on the integration of two problems: automatic speech recognition (ASR), *i.e.* the conversion of a speech signal into a transcript, and machine translation (MT), *i.e.* the translation of a transcript into another language. SLT is more difficult than the simple concatenation of ASR and MT systems [1], for mainly two reasons: i) ASR systems are prone to make errors, ii) MT systems, which are typically trained on written language, do not perform well on noisy transcripts of spoken language [2].

On the other hand, SLT is a very important problem to solve, given the vast amount of human-to-human communication based on speech. Besides personal communication for business and traveling, important applications of SLT are for instance international teleconferencing and the subtitling of audiovisual content, just to mention two.

Since 2004 [3], IWSLT has put effort in organizing challenging SLT tasks, which were at the same time affordable for the currently available technology. Thus, IWSLT initially focused mainly on multilingual communication in the traveling domain [4, 5] a progressively moved to translation of speeches [6], university lectures [7], teleconference chats [8], dialogues in the health-care domain [9], etc. In parallel with the advance of technology, from statistical to neural models, also data collections prepared for the evaluations evolved, too. While for instance previous training data reflected the separation of ASR from MT, the recently collected MuST-C corpus [10] permits now to directly train end-to-end speech-to-text neural systems [11, 12, 13].

This year, two speech translation tasks and one text translation task of spoken language have been organized.[1]. The two speech translation tasks address the translation of TED talks and How2 instructional videos [14] from English to German and Portuguese. For the How2 task, SLT systems could also use the videos as additional input. The text translation task addresses the translation of TED talks from English into Czech.

This year, 12 groups participated in the evaluation (see Table 1). In the following, we describe each task in detail, provide a summary of the received submissions, and report tables with performance results in the appendix.

2. Speech Translation - TED

2.1. Definition

One of the speech-to-text translation tasks proposed this year required participants to translate English audio data extracted from TED talks[2] into German and Portuguese. This TED task accepted both cascade (i.e. ASR + MT pipelined architectures) and end-to-end system submissions, inviting partic-

[1] A fourth task addressing the translation of conversational Spanish speech into fluent English text was set up but could not be run due to unforeseen data licensing problems

[2] `http://www.ted.com`

Table 1: List of Participants

Team ID	Organization	Text	TED		How2	
		EN-CZ	EN-DE	EN-PT	EN-DE	EN-PT
BSLEE	Individual participant		✓			
CMU	Carnegie Mellon University, USA	✓				
CUNI	Charles University - Institute of Formal and Applied Linguistics, Czech Republic	✓				
FBK	Fondazione Bruno Kessler, Italy		✓			
IMPERIAL	Imperial College, UK					✓
KIT	Karlsruhe Institute of Technology, Germany	✓	✓	✓	✓	✓
LIG	Laboratoire d'Informatique de Grenoble, France	✓				
SRC-B	Samsung Research China - Beijing (SRC-B), China		✓	✓		
ON-TRAC	ON-TRAC Consortium (LIG, LIA, LIUM), France			✓		✓
OPPO	OPPO Beijing Research Institute	✓				
SRPOL	Samsung R&D Institute Poland, Poland		✓			
SRPOL-UEDIN	Samsung R&D Institute Poland and University of Edinburgh, Poland/UK	✓				

ipants to explicitly indicate which of the two architectural choices was made for their system.

In the cascade case, participants were provided with a baseline implementation of the traditional pipeline as a Docker container.[3] This implementation (comprising a neural ASR system, a sentence segmentation system and an attention-based MT system) was released for participants willing to focus on one component of the pipeline and exploit baseline components for the other parts.

In the end-to-end case, valid submissions had to be obtained by models that:

- Do not exploit intermediate discrete representations (e.g., source language transcription or hypotheses fusion in the target language);
- Rely on parameters that are all jointly trained on the end-to-end task

2.2. Data

In addition to the data also used last year (i.e. WIT^3 [6] and the Speech-Translation TED corpus downloadable from the task web page[4]), participants were provided with MuST-C, a recently released speech translation dataset [10]. MuST-C is a multilingual corpus aimed to facilitate the training of end-to-end systems for SLT from English into 8 languages (German, Spanish, French, Italian, Dutch, Portuguese, Romanian and Russian). For each target language, it comprises at least 385 hours of audio recordings from English TED Talks, which are automatically aligned at the sentence level with their manual transcriptions and translations. This was done by first aligning transcription and translations of each original English talk using the Gargantua sentence alignment tool [15], and then by aligning the English transcripts with the corresponding audio tracks using Gentle,[5] an off-the-shelf English forced-aligner built on the Kaldi ASR toolkit [16].

Language	#Talks	#src words
En-De (DEV)	22	21K
En-De (TEST)	23	41K
En-Pt (DEV)	14	20K
En-Pt (TEST)	26	41K

Table 2: Statistics of the development and test sets created for the IWSLT 2019 Speech Translation - TED task.

The same approach was applied to create new development and test data for IWSLT 2019 starting from talks that were not included yet in the current version of MuST-C. Some statistics about the newly-created En-De and En-Pt data are reported in Table2.

Additional allowed datasets include: the How2 corpus (only En-Pt, see Section 3), the TED LIUM corpus[6] [17], all the data provided by the WMT 2018 Conference on Machine Translation[7] and the OpenSubtitles corpus.[8]

2.3. Submissions

In total, we received 16 submissions (3 of them marked as "late" submissions) from 6 teams. Five teams (BSLEE, FBK, KIT, SRC-B, SRPOL) participated in the English-German sub-task (submitting 13 runs), while three teams (KIT, SRC-B, ON-TRAC) participated in the English-Portuguese sub-task (3 runs). The participating systems are briefly described

[3] https://github.com/isl-mt/SLT.KIT

[4] http://sites.google.com/view/iwslt-evaluation-2019/speech-translation

[5] github.com/lowerquality/gentle

[6] Except for a list of 35 "off-limit" talks included in the 2017 progressive and evaluation sets of the TED based MT tasks.

[7] http://www.statmt.org/wmt18/

[8] http://opus.nlpl.eu/OpenSubtitles2018.php

in Section 5. While the top-performing one in both the subtasks is based on the pipeline approach (ASR+MT), the others are fully end-to-end and exploit different audio segmentation techniques (different from the How2-task, the TED evaluation data are not supplied with pre-defined segmentation) as well as different data augmentation strategies and overall architectural choices.

2.4. Results

Case-sensitive BLEU [18] is the task's primary evaluation metric. In addition, for a more informative assessment automatic evaluation results were also computed in terms of case-insensitive BLEU, case-sensitive/insensitive TER [19], BEER [20], and CharacTER [21].

The results of the submitted primary runs are shown in Appendix A.1. The top results in the two sub-tasks suggest a higher difficulty for the En-De setting, for which the BLEU score is ~ 5.0 points lower (21.55 for En-De *Vs.* 26.53 for En-Pt). It's worth noting that, although in both the language settings the winning submission (a late submission in the En-Pt case) is still based on a pipeline architecture, the best direct systems are less than 2.0 BLEU points worse. This relatively small performance gap between traditional and fully end-to-end approaches is an interesting indicator of the progress made by end-to-end speech translation technology.

3. Speech Translation - How2

In this year's edition we added a new speech-to-text translation track which explores multimodality. In this track we offered two tasks. The first one was to translate English speech to Portuguese text by using explicit speech-text supervision and vision as supporting modality (*i.e.* grounding or adaptation). The second one was to translate English speech to German text without explicit German supervision (*i.e.* only by using speech-text Portuguese supervision and images as support).

3.1. Data

The How2 dataset is made of 79,114 instructional videos (2,000 hours), where each clip has an average duration of 90 seconds. The scripts for (re-)creating the dataset are made available at `https://github.com/srvk/how2-dataset`. The repository also contains information on obtaining the pre-computed features for validation or saving computation. Some baseline ASR systems, based on nmtpy [22], are also available.

We collected videos with ground-truth English subtitles from *YouTube* by using a keyword-based web spider as in [23]. We also downloaded metadata and video descriptions generated by the author of the video. To collect Portuguese and German translations, we first re-segmented the English subtitles into sentences. We then word-aligned these sentences to the audio speech by using an ASR system pretrained on Wall-Street Journal. Using the audio word-alignments and the segmented text sentences, we defined the segments of the How2 dataset. See Table 3 for more detailed statistics of the segment and word distribution.

Name	Language	Hours	# Hours	# Clips	Clip Stats
train	PT	13,168	298.2	184,949	5.8 s & 17 w
val	PT	150	3.2	2,305	5.8 s & 17 w
	DE	150	3.2	2,305	5.8 s & 15 w
test	PT	159	2.8	1,905	5.4 s & 17 w
	DE	175	3.9	2,497	5.7 s & 15 w

Table 3: Summary of the statistics of the datasets involved in the multimodal track in IWSLT 2019. In the "Clip Stats" column, we show the average time per segment and number of words.

3.1.1. Portuguese Annotation

We collected Portuguese translations of a 300 hours subset by using the *Figure Eight* crowdsourcing platform. To speed up the process, we gave automatic translation to the annotators and framed the task as a post-editing task. To do so, we first crowdsourced a quality score of the translations from three state-of-the-art commercial translation systems. After that, we selected the best system for each segment and used its outputs as a translation candidate for the post-edit task. We applied geographic restrictions so that the post-editing task could be performed only by people living in Portugal or Brazil. Post-editing had to also tke into consideration the video. To assure quality of the post-edited translations, we replaced a content word in every five automatic translations with a random content word, independent of the actual translation. After post-editing, if the replaced word was still present in the translation we excluded the worker's annotations.

Finally, we performed a verification experiment by comparing the results of the post-edited annotations with the ones generated by the state-of-the-art commercial translation system, which performs really well on this data and language pair. We observed that post-edits improve performance by 1 BLEU point confirming that the approach is justified.

3.1.2. German Annotation

The translations of the development and test sets of the evaluation were performed by two professional translators from scratch. Both translators were German native speakers and had obtained a Bachelor's degree in translation from English into German.

As the videos sometimes contained very specific content with unusual technical terms, the translators were also provided with the links to the videos and the English transcripts.

3.2. Submissions

We received 19 submissions from 3 teams. KITparticipated in the English-German sub-task, while three teams (KIT, IMPERIAL, ON-TRAC) participated in the English-Portuguese sub-task (18 runs). The participating systems are briefly described in Section 5.

While the top performing system for the English-to-Portuguese sub-task is a cascaded system, the second-best system is an end-to-end system. Only one team (IMPERIAL) used multi-modal information.

3.3. Results

The evaluation metrics are the same as for the TED speech translation task (see Section 2.4). The results of the submitted primary runs are shown in Appendix A.1. The top performing system in the English-to-Portuguese subtask performs about 3.5 BLEU points better than the end-to-end system in second place. Just as for the TED task, this comparatively small gap seems to be an indication of the progress in performance in end-to-end speech translation systems.

4. Text translation

4.1. Definition

The Text Translation Task this year addressed a new translation direction: from English to Czech. We invited participants to investigate MT into a moderate morphologically rich language and to overcome the difficulty of having less in-domain resources. Furthermore, the participants need to consider applying domain and genre adaptation methods, as we would test the translation system on spoken style TED talks.

The main data for the task is a compilation of TED talks from English and translated into Czech, collected by FBK as a special part of MuST-C[9], a recently released multilingual speech translation data set[10]. The participants have been asked to translate some test set in the similar genre and domain (TED talks).

Statistics of IWSLT-2019 from MuST-C for English-Czech text translation:

Dataset	#Talks	#src words
Training set	1257	2.4M
Development set	10	25K
Test set	43	47K

In addition, the participants in the text translation task have been provided a large portion of general data from the similar task presented in WMT's news translation campaign[10]. Those data come from various sources with mixed domains, e.g. Europarl, News Commentary, ParaCrawl, CommonCrawl and a large corpus CzEng provided from the Charles University (CUNI). The table below shows the statistics of those data sets.

Dataset	#sentence pairs	#src words
Europarl	641K	15.6M
CommonCrawl set	162K	3.3M
ParaCrawl	3M	48.8M
News Commentary	240k	5.1M
CzEng	57M	617M

Although these datasets allow the participants to use substantially large parallel data compared to the MuST-C, the fact that parts of those data are noisy requires the participants investigate suitable filtering and adaptation methods in order to get benefit from using those data sets.

4.2. Results

We received 30 submissions from 6 different participants (CUNI, KIT, SPROL-UEDIN, OPPO, LIG and CMU). The results on the tst2019 evaluation set for each participant's primary submission are shown in Appendix A.1, sorted by the BLEU metric.

5. Submissions

We received submissions from 12 research teams to the three tasks. In the following, the submissions will be briefly described.

5.1. CMU

The system by CMU used Block Multitask Learning (BMTL) to predict multiple targets of different granularities simultaneously. To do so, they incorporate a multitask learning approach to a traditional encoder-decoder machine translation model. More concretely, BMTL uses a single encoder that accepts subwords of one only granularity. The encoded representation is later used by multiple decoders, which have their individual attention mechanism and parameters, to generate translations on different subword granularities. At training time the losses of each subword granularity decoder are length-normalized, summed and averaged. BMTL forces the encoder to generate a more general representation independent from the output subword granularity. Finally, as a post-processing step, they combine the different output granularities of BMTL by using Multi-Engine Machine Translation (MEMT) into a single word-based translation hypothesis.

5.2. CUNI

CUNI participated in the the English to Czech translation task. All four CUNI systems are based on the Transformer model implemented in the Tensor2Tensor framework. Two of the systems serve as baselines, which are not adapted to

[9] https://ict.fbk.eu/must-c-release-v1-1/

[10] http://www.statmt.org/wmt19/translation-task.html

the TED talks domain: SentBase (contrastive3) is trained on single sentences, DocBase (contrastive1) on multi-sentence (document-level) sequences. The other two submitted systems are adapted to TED talks: SentFine (contrastive2) is fine-tuned on single sentences, DocFine (primary) is fine-tuned on multi-sentence sequences.

5.3. FBK

FBK's system is based on S-Transformer [24] with logarithmic distance penalty, an ST-oriented adaptation of Transformer. For training, the team focused on data augmentation techniques drawn from ST and ASR. The augmented data were exploited in three different ways at different stages of the process. First, by training an end-to-end ASR system and using the weights of its encoder to initialize the decoder of the ST model (transfer learning). Second, by using an English-German MT system trained on large data to translate the English side of the English-French MuST-C training set into German, and using the resulting data as additional training material. Third, by training the model with SpecAugment [25], an augmentation technique that randomly masks portions of the spectrograms in order to make them different at every training epoch.

5.4. IMPERIAL

All 15 systems submitted by IMPERIAL were trained on the How2 dataset only and they then conducted inference on the evaluation set (2497 examples) provided by the organizers. For all submissions, they use a cascaded speech translation system, transcribing an (English) How2 video segment first and then translating the transcript into Portuguese. The ASR model is unimodal and identical in all the 15 models, that is, it only utilizes the audio. Among the submissions, contrastives 1, 5, and 10 rely only on the transcripts to translate, whereas all the other systems are multimodal, i.e. they exploit visual features during translation. Among the multimodal systems, contrastives 2, 3, 6, 7, 11 and 12 use the 2048-D pooling features officially provided, whereas the other ones exploit visual features that were extracted from an action recognition network applied to the How2 videos.

5.5. KIT

KIT participated in all 5 conditions. For the SLT tasks, all submissions are cascaded systems of an speech recognition system, a punctuation prediction system and a machine translation system. They have conducted the speech recognition experiments with two different end-to-end architectures. The final model is the ensemble of those two architecture models, where it has achieved the best results on the development sets of SLT sub-tasks. For the machine translation part, a Transformer-based multilingual model has employed, thus, they are able to produce the translations of all the sub-tasks with a single model.

5.6. LIG

The MT system by LIG is based on a neural encoder-decoder Transformer architecture, with the ability to use a pre-trained language model (LM) in input. They trained the model on very few data: only 128k sentences of specialized data (TED talks) and 247k sentences of general data (news commentary), and in order to study the impact of the language model on the MT performance, they compared three configurations: First, a system without any LM (primary submission). Secondly, a model with an LM trained on the allowed data only, and finally one with an external LM trained on a large quantity of data. The results show a clear improvement with the external LM, whereas their system does not always benefit from the constrained LM compared to the no-LM scenario.

5.7. SRC-B

SRC-B proposes layer-wise tied self-attention for end-to-end speech translation. Their method takes advantage of sharing weights of speech encoder and text decoder. The representation of source speech and the representation of target text are coordinated layer by layer, so that the speech and text can learn a better alignment during the training procedure. They also adopt data augmentation to enhance the parallel speech-text corpus. The En-De experimental results show that their single end-to-end model achieves a BLEU score of 17.68 on tst2015. Their ASR achieves a WER of 6.6% on the TED-LIUM test set. The En-Pt model achieves a BLEU score of 11.83 on the MuST-C dev set.

5.8. ON-TRAC

ON-TRAC developed an end-to-end speech translation system to translate English speech into Portuguese text (EN-PT). A single end-to-end model was used for two primary submissions corresponding to two EN-PT evaluations sets: TED (MuST-C) and How2. The model is trained on the data from two training corpora provided by the IWSLT-2019 organizers: English-Portuguese part of the MuST-C corpus and How2. The total amount of training data corresponds to 674 hours of English audio and to about 390K segments. In this work, 80-dimensional Mel filter-bank features, concatenated with 3-dimensional pitch features, are used for training. Data augmentation, based on speed perturbation (with factors 0.9, 1.0, 1.1), is applied to speech data. They used an attention-based encoder-decoder architecture for the end-to-end speech translation model. The encoder has VGG-like CNN layers (two CNN blocks, where each block consists of 2-layer CNNs with max-pooling) followed by five stacked 1024-dimensional BLSTM layers. The decoder has two 1024-dimensional LSTM layers. The target tokens are characters. Experiments are conducted using the ESPnet – an end-to-end speech processing toolkit. The end-to-end translation model shows the following results in terms of case-sensitive BLEU score: 26.91 on the tst-COMMON (MuST-C's dev set), and 42.97 on the val (How2's dev set).

5.9. OPPO

OPPO's system is based on Transformer architecture which uses self-attention largely. As the training data may be crawled from the Web, which contains a lot of errors such as translation and spelling errors, mismatch sentences etc.., they cleaned nearly half of the data with the help of "fast align", resulting in considerable improvements in the final result. Besides, they also study the effect of model combination. On the devsets of IWSLT 2019, their system reaches a BLEU score of 19.94.

5.10. SRPOL

SRPOL's system is based on an altered Transformer ASR architecture. It was trained on augmented IWSLT, TED-LIUM and MuST-C corpora. Additionally it introduces a second ASR decoder to improve audio feature extraction.

5.11. SRPOL-UEDIN

SRPOL-UEDIN's system is a combination of four individually trained Transformer models. Two models were trained from scratch using a mixture of WMT19, MuST-C, and back-translated synthetic data and two were existing WMT19 models that were fine-tuned using MuST-C data. The final translations were produced by rescoring and reranking a combined n-best list using weighted scores from the four models.

6. Conclusions

The 2019 IWSLT evaluation campaign was about three tasks: speech-to-text translation of talks and instructional videos from English to German to Portuguese, and text-to-text translation of talks from English to Czech. Twelve teams took part in the evaluation. After describing the data and evaluation conditions of each task, we gave account of the results of all primary runs submitted by the participants. While this report focuses on the automatic evaluations only, an independent paper will report on a human evaluation carried out on a subset of the submitted runs.

7. References

[1] F. Casacuberta, M. Federico, H. Ney, and E. Vidal, "Recent efforts in spoken language processing," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 80–88, May 2008.

[2] N. Ruiz and M. Federico, "Complexity of spoken versus written language for machine translation," in *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT)*, Dubrovnik, Croatia, 2014, pp. 173–180.

[3] Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii, "Overview of the IWSLT04 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 1–12.

[4] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world." in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2002, pp. 147–152.

[5] M. Yang, H. Jiang, T. Zhao, and S. Li, "Construct Trilingual Parallel Corpus on Demand," in *Chinese Spoken Language Processing*, ser. Lecture Notes in Computer Science, Q. Huo, B. Ma, E.-S. Chng, and H. Li, Eds. Berlin, Heidelberg: Springer, 2006, pp. 760–767.

[6] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web Inventory of Transcribed and Translated Talks," in *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012. [Online]. Available: http://hltshare.fbk.eu/EAMT2012/html/Papers/59.pdf

[7] E. Cho, S. Fünfer, S. Stüker, and A. Waibel, "A corpus of spontaneous speech in lectures: The kit lecture corpus for spoken language processing and translation," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Reykjavik, Iceland: European Language Resources Association (ELRA), may 2014.

[8] C. Federmann and W. Lewis, "Microsoft Speech Language Translation (MSLT) Corpus: The IWSLT 2016 release for English, French and German," in *Proceedings of IWSLT 2016*, Dec. 2016. [Online]. Available: https://www.microsoft.com/en-us/research/publication/mslt-corpus-iwslt-2016-release/

[9] H. Tanaka, K. Yoshino, K. Sugiyama, S. Nakamura, and M. Kondo, "Multimodal interaction data between clinical psychologists and students for attentive listening modeling," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, Oct. 2016, pp. 95–98, iSSN: 2472-7695.

[10] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *Proc. NAACL*, 2019, pp. 2012–2017.

[11] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-Sequence Models Can Directly Translate Foreign Speech," in *Interspeech*

2017. ISCA, Aug. 2017, pp. 2625–2629. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0503.html

[12] L. Cross Vila, C. Escolano, J. A. R. Fonollosa, and M. R. Costa-Jussà, "End-to-End Speech Translation with the Transformer," in *IberSPEECH 2018*. ISCA, Nov. 2018, pp. 60–63. [Online]. Available: http://www.isca-speech.org/archive/IberSPEECH_2018/abstracts/IberS18_P1-9_Cross-Vila.html

[13] M. Sperber, G. Neubig, J. Niehues, and A. Waibel, "Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation," *arXiv:1904.07209 [cs]*, Apr. 2019, arXiv: 1904.07209. [Online]. Available: http://arxiv.org/abs/1904.07209

[14] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: a large-scale dataset for multimodal language understanding," in *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS, 2018. [Online]. Available: http://arxiv.org/abs/1811.00347

[15] F. Braune and A. Fraser, "Improved Unsupervised Sentence Alignment for Symmetrical and Asymmetrical Parallel Corpora," in *Proceedings of COLING 2010*, Beijing, China, 2010, pp. 81–89.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, K. N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The Kaldi Speech Recognition Toolkit," in *Proceedings of ASRU 2011*, Big Island, Hawaii, USA, 2011, pp. 1–4.

[17] A. Rousseau, P. Deléglise, and Y. Esteve, "Enhancing the ted-lium corpus with selected data for language modeling and more ted talks," in *LREC*, 2014. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1104_Paper.pdf

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, USA, 2002, pp. 311–318.

[19] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the The Seventh Conference of the Association for Machine Translation in the Americas (AMTA)*, Cambridge, USA, 2006, pp. 223–231.

[20] M. Stanojevic and K. Sima'an, "BEER: BEtter evaluation as ranking," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2014.

[21] W. Wang, J.-T. Peter, H. Rosendahl, and H. Ney, "CharacTer: Translation edit rate on character level," in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. Association for Computational Linguistics, 2016. [Online]. Available: https://doi.org/10.18653%2Fv1%2Fw16-2342

[22] O. Caglayan, M. García-Martínez, A. Bardet, W. Aransa, F. Bougares, and L. Barrault, "NMTPY: A flexible toolkit for advanced neural machine translation systems," *The Prague Bulletin of Mathematical Linguistics*, 2017.

[23] S.-I. Yu, L. Jiang, and A. Hauptmann, "Instructional videos for unsupervised harvesting and learning of action examples," in *Proceedings of the International Multimedia Conference (ACMM)*. ACM, 2014.

[24] M. A. Di Gangi, M. Negri, and M. Turchi, "Adapting Transformer to End-to-end Spoken Language Translation," in *Proceedings of INTERSPEECH*, Graz, Austria, September 2019.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

Appendix A. Automatic Evaluation

A.1. Official Testset (*tst2019*)

· All the sentence IDs in the IWSLT 2019 testset were used to calculate the automatic scores for each run submission.
· MT systems are ordered according to the *BLEU* metrics.
· *WER*, *BLEU* and *TER* scores are given as percent figures (%). · End-to-end systems are indicated by gray background.

Text Translation : English-Czech

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
CUNI	29.03	53.05	58.19	41.97	29.82	51.93
KIT	28.62	52.76	57.80	42.78	29.45	51.62
SRPOL-UEDIN	28.07	54.57	57.76	42.19	28.84	53.40
OPPO	26.67	55.65	56.64	43.79	27.59	54.39
LIG	22.72	58.51	53.94	48.27	23.47	57.41
CMU	16.93	64.65	50.06	53.60	17.6	63.69

Speech Translation : TED English-German

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
KIT	21.55	65.73	50.50	52.56	22.84	63.35
SRPOL	19.96	65.25	49.54	55.07	20.89	63.53
SRC-B	19.5	67.68	48.94	57.35	20.81	65.17
FBK	15.67	76.04	43.30	62.50	16.76	74.03
BSLEE	13.67	76.61	42.95	66.83	14.57	74.88

Speech Translation : TED English-Portuguese

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
ON-TRAC	24.57	67.92	49.16	52.33	25.87	65.41
SRC-B	9.95	106.28	30.85	76.62	10.6	103.22
KIT[11]	26.53	61.58	51.39	50.41	27.97	59.38

Speech Translation : How2 English-German

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
KIT	14.59	75.87	48.24	54.78	15.31	74.06

Speech Translation : How2 English-Portuguese

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
KIT	47.86	35.96	66.47	27.72	49.65	33.75
ON-TRAC	44.08	39.94	64.22	31.27	44.55	39.31
IMPERIAL	39.63	43.63	60.80	36.60	40.05	43.10

[11] late submission