FINDINGS OF THE IWSLT 2020 EVALUATION CAMPAIGN

Ebrahim Ansari
Charles U./IASBS

Amittai Axelrod
DiDi Labs

Nguyen Bach
Alibaba

Ondřej Bojar
Charles U.

Roldano Cattoni
FBK

Fahim Dalvi
QCRI

Nadir Durrani
QCRI

Marcello Federico
Amazon AI

Christian Federmann
Microsoft Research

Jiatao Gu
Facebook AI

Fei Huang
Alibaba

Kevin Knight
DiDi Labs

Xutai Ma
JHU/Facebook AI

Ajay Nagesh
DiDi Labs

Matteo Negri
FBK

Jan Niehues
Maastricht U.

Juan Pino
Facebook AI

Elizabeth Salesky
JHU

Xing Shi
DiDi Labs

Sebastian Stüker
KIT

Marco Turchi
FBK

Alex Waibel
CMU/KIT

Changhan Wang
Facebook AI

Abstract

The evaluation campaign of the International Conference on Spoken Language Translation (IWSLT 2020) featured this year six challenge tracks: (i) Simultaneous speech translation, (ii) Video speech translation, (iii) Offline speech translation, (iv) Conversational speech translation, (v) Open domain translation, and (vi) Non-native speech translation. A total of 30 teams participated in at least one of the tracks. This paper introduces each track's goal, data and evaluation metrics, and reports the results of the received submissions.

1 Introduction [Marcello]

The International Conference on Spoken Language Translation (IWSLT) is an annual scientific conference (Akiba et al., 2004; Eck and Hori, 2005; Paul, 2006; Fordyce, 2007; Paul, 2008, 2009; Paul et al., 2010; Federico et al., 2011, 2012; Cettolo et al., 2013, 2014, 2015, 2016, 2017; Niehues et al., 2018, 2019) for the study, development and evaluation of spoken language translation technology, including: speech-to-text, speech-to-speech translation, simultaneous and consecutive translation, speech dubbing, cross-lingual communication including all multimodal, emotional, para-linguistic, and stylistic aspects and their applications in the field. The goal of the conference is to organize evaluations and sessions around challenge areas, and to present scientific work and system descriptions. This paper reports on the evaluation campaign organized by IWSLT 2020, which features six challenge tracks:

- **Simultaneous speech translation**, addressing low latency translation of talks, from English to German, either from a speech file into text, or from a ground-truth transcript into text;
- **Video speech translation**, targeting multimodal speech translation of video clips into text, either from Chinese into English or from English into Russian
- **Offline speech translation**, proposing speech translation of talks from English into German, using either cascade architectures or end-to-end models, able to directly translate source speech into target text;
- **Conversational speech translation**, targeting the translation of highly disfluent conver-

Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT), pages 1–34

https://doi.org/10.18653/v1/P17

sations into fluent text, from Spanish to English, starting either from audio or from a verbatim transcript;

- **Open domain translation**, addressing Japanese-Chinese translation of unknown mixed-genre test data by leveraging heterogeneous and noisy web training data.
- **Non-native speech translation**, considering speech translation of English-to-Czech and English-to-German speech in a realistic setting of non-native spontaneous speech, in somewhat noisy conditions.

The challenge tracks were attended by 30 participants (see Table 1), including both academic and industrial teams. This correspond to a significant increment with respect to the last year's evaluation campaign, which saw the participation of 12 teams. The following sections report on each challenge track in detail, in particular: the goal and automatic metrics adopted for the challenge, the data used for training and testing data, the received submissions and the summary results. A detailed account of the results for each challenge is instead reported in a corresponding appendix.

2 Simultaneous Speech Translation

Simultaneous machine translation has become an increasingly popular topic in recent years. In particular, simultaneous speech translation enables interesting applications such as subtitle translations for a live event or real-time video-call translations. The goal of this challenge is to examine systems for translating text or audio in a source language into text in a target language from the perspective of both translation quality and latency.

2.1 Challenge

Participants were given two parallel tracks to enter and encouraged to enter both tracks:

- text-to-text: translating ground-truth transcripts in real-time.
- speech-to-text: translating speech into text in real-time.

For the speech-to-text track, participants were able to submit systems either based on cascaded or end-to-end approaches. Participants were required to implement a provided API to read the input and write the translation, and upload their system as a Docker image so that it could be evaluated by the organizers. We also provided an example implementation and a baseline system[1].

Systems were evaluated with respect to quality and latency. Quality was evaluated with the standard metrics BLEU (Papineni et al., 2002a), TER (Snover et al., 2006b) and METEOR (Lavie and Agarwal, 2007). Latency was evaluated with the recently developed metrics for simultaneous machine translation including average proportion (AP), average lagging (AL) and differentiable average lagging (DAL) (Cherry and Foster, 2019). These metrics measure latency from an algorithmic perspective and assume systems with infinite speed. For the first edition of this task, we report wall-clock times only for informational purposes. In the future, we will also take wall-clock time into account for the official latency metric.

Three regimes, low, medium and high, were evaluated. Each regime was determined by a maximum latency threshold. The thresholds were measured with AL, which represents the delay to a perfect real-time system (milliseconds for speech and number of words for text). The thresholds were set to 3, 6 and 15 for the text track and to 1000, 2000 and 4000 for the speech track, and were calibrated by the baseline system. Participants were asked to submit at least one system per latency regime and were encouraged to submit multiple systems for each regime in order to provide more data points for latency-quality trade-off analyses.

2.2 Data

Participants were allowed to use the same training and development data as in the Offline Speech Translation track. More details are available in §4.2.

2.3 Submissions

The simultaneous task received submissions from 4 teams: 3 teams entered both the text and the speech tracks while 1 team entered the text track only. Teams followed the suggestion to submit multiple systems per regime, which resulted in a total of 56 systems overall.

ON-TRAC (Elbayad et al., 2020) participated in both the speech and text tracks. The authors used a hybrid pipeline for simultaneous speech

[1] https://github.com/pytorch/fairseq/tree/simulastsharedtask/examples/simultaneous_translation

Team	Organization
AFRL	Air Force Research Laboratory, USA (Ore et al., 2020)
APPTEK/RWTH	AppTek and RWTH Aachen University, Germany (Bahar et al., 2020a)
BHANSS	Samsung Research, South Korea (Lakumarapu et al., 2020)
BUT	Brno University of Technology, Czech Republic (no system paper)
CASIA	Inst. of Automation, Chinese Academy of Sciences, China (Wang et al., 2020b)
CUNI	Charles University, Czech Republic (Polák et al., 2020)
DBS	Deep Bleu Sonics, China (Su and Ren, 2020)
DIDI LABS	DiDi Labs, USA (Arkhangorodsky et al., 2020)
ELITR	CUNI + KIT + UEDIN (Macháček et al., 2020)
FBK	Fondazione Bruno Kessler, Italy (Gaido et al., 2020)
HY	University of Helsinki, Finland (Vázquez et al., 2020)
HW-TSC	Huawei Co. Ltd, China (Wang et al., 2020a)
IITB	Indian Institute of Technology Bombay, India (Saini et al., 2020)
ISTIC	Inst. of Scientific and Technical Inf. of China (Wei et al., 2020)
KINGSOFT	Kingsoft, China. (no system paper)
KIT	Karlsruhe Institute of Technology, Germany (Pham et al., 2020)
KSAI	Kingsoft AI Lab, China (no system paper)
NAIST	Nara Institute of Science and Technology, Japan (Fukuda et al., 2020)
NICT	National Institute of Comm. Techn., Japan (no system paper)
OCTANOVE	Octanove Labs LLC, USA (Hagiwara, 2020)
ON-TRAC	ONTRAC Consortium, France (Elbayad et al., 2020)
OPPO	Beijing OPPO Telecommunications Co., Ltd., China (Zhang et al., 2020)
SJTU	Shanghai Jiao Tong University, China (no system paper)
SRC-B	Samsung Research, China (Zhuang et al., 2020)
SRPOL	Samsung Research , Poland (Potapczyk and Przybysz, 2020)
SRSK	Samsung Research, South Korea (Han et al., 2020)
TAMKANG	Tamkang University, Taiwan (no system paper)
TSUKUBA	University of Tsukuba, Japan (Cui et al., 2020)
UEDIN	University of Edinburgh, UK(Chen et al., 2020)
XIAOMI	Xiaomi AI Lab, China (Sun et al., 2020)

Table 1: List of Participants

translation track, with a Kaldi-based speech recognition cascaded with transformer-based machine translation with wait-k strategy (Ma et al., 2019). In order to save the cost of encoding every time an input word is streamed, a uni-directional encoder is used. Multiple wait-k paths are jointly optimized in the loss function. This approach was found to be competitive with the original wait-k approach without needing to retrain for a specific k.

SRSK (Han et al., 2020) participated in the speech and text tracks. This is the only submission to use an end-to-end approach for the speech track. The authors use transformer-based models combining the wait-k strategy (Ma et al., 2019) with a modality-agnostic meta learning approach (Indurthi et al., 2020) to address data sparsity. They also use the ST task along with ASR and MT as the source task, a minor variation explored compared to the original paper. In the text-to-text task, the authors also explored English-German and French-German as source tasks. This training setup is facilitated using a universal vocabulary. They analyzed models with different values in wait-k during training and inference and found the meta learning approach to be effective when the data is limited.

AppTek/RWTH (Bahar et al., 2020a) participated in the speech and text tracks. The authors proposed a novel method to simultaneous translation, by training an additional binary output to predict chunk boundaries in the streaming input. This module serves as an agent to decide when the contextual information is sufficient for the decoder

to write output. The training examples for chunk prediction are generated using word alignments. On the recognition side, they fixate the ASR system to the output hypothesis that does not change when further context is added. The model chooses chunk boundaries dynamically.

KIT (Pham et al., 2020) participated in the text track only. The authors used a novel read-write strategy called Adaptive Computation Time (ACT) (Graves, 2016). Instead of learning an agent, a probability distribution derived from encoder timesteps, along with the attention mechanism from (Arivazhagan et al., 2019b) is used for training. The ponder loss (Graves, 2016) was added to the cross-entropy loss in order to encourage the model towards shorter delays. Different latency can be achieved by adjusting the weight of the ponder loss.

2.4 Results

We discuss results for the text and speech tracks. More details are available in Appendix A.1.

2.4.1 Text Track

Results for the text track are summarized in the first table of Appendix A.1. Only the ON-TRAC system was able to provide a low latency model. The ranking of the systems is consistent throughout the latency regimes. The results for all systems are identical between the high latency regime and the unconstrained regime except for SRSK who submitted a system above the maximum latency threshold of 15.

In the table, only the models with the best BLEU score for a given latency regime are reported. In order to obtain a broader sense of latency-quality thresholds, we plot in Figure 1 all the systems submitted to the text track. The ON-TRAC models present competitive trade-offs across a wide latency range. The APPTEK/RWTH system obtains competitive performance for medium latency, but its characteristics in low and high latency regimes are unclear.

2.4.2 Speech Track

Results for the speech track are summarized in the second table of Appendix A.1. We also report latency-quality trade-off curves in Figure 2. The ON-TRAC system presents better trade-offs across a wide latency range. We also note that the APPTEK/RWTH systems are all above the highest latency threshold of 4000, which makes it difficult to compare its trade-offs to other systems.

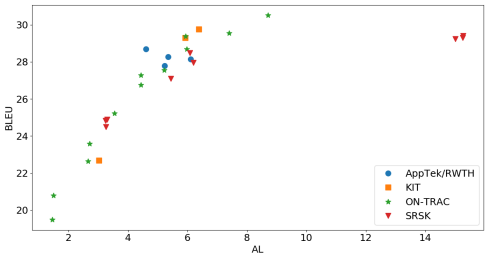

Figure 1: Latency-quality trade-off curves, measured by AL and BLEU, for the systems submitted to the text track.

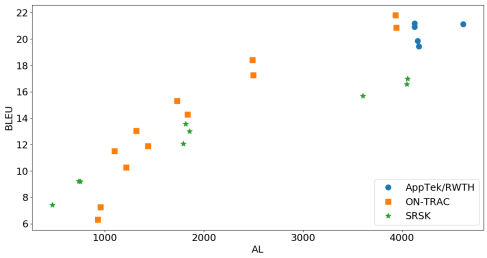

Figure 2: Latency-quality trade-off curves, measured by AL and BLEU, for the systems submitted to the speech track.

2.5 Future Editions

In future editions, we will include wall-clock time information as part of the official latency metric. This implies that the evaluation will be run in a more controlled environment, for example, the hardware will be defined in advance. We will also encourage participants to contrast cascade and end-to-end approaches for the simultaneous speech track.

3 Video Speech Translation

We are living the multiple modalities world in which we see objects, hear sounds, feel texture, smell odors, and so on. The purpose of this shared task is to ignite possibilities of multimodal machine translation. This shared task examines methods for combining video and audio sources as input of translation models.

3.1 Challenge

In this year's evaluation campaign, we added the video translation track to ignite possibilities of

multimodal machine translation. This track examines methods for combining video and audio sources as input of translation models. We offer two evaluation tasks. The first one is the constrained track in which systems are required to only use the datasets we provided in the data section. The second one was unconstrained systems in which additional datasets are allowed. Both tasks are available for Chinese-English and English-Russian language pairs.

3.2 Data

We are focusing on e-Commerce domain, particularly on the live video shows similar to the ones on e-Commerce websites such as AliExpress, Amazon, and Taobao. A typical live show has at least one seller in a wide range of recording environments. The live show contents cover product description, review, coupon information, chitchat between speakers, interactive chat with audiences, commercial ads, and breaks. We planned to collect videos from Taobao for Chinese-English, and videos from AliExpess for English-Russian.

We have experienced data collection and annotation challenges during these unprecedented times. Our English-Russian plan could not be carried out smoothly. Therefore, instead of collecting and annotating e-Commerce videos, we use the How2 dataset[2] and translate the dev and test sets from English to Russian.

For Chinese-English, we collected ten Taobao full live shows which last between fifteen minutes and four hours. After quality check, we keep seven live shows for annotation. For each live show we sampled video snippets ranging from 1 to 25 minutes relatively to the length of the original show. Audio files are extracted from video snippets. Each audio file is further split into smaller audios based on the silence and voice activities. We ask native Chinese speakers to provide human transcriptions. For human translation, we encourage annotators to watch video snippets before translating. There are 2 English translation references for a total of 104 minutes of Chinese live shows. All data is available on GitHub[3].

3.3 Submissions

We received 4 registrations, however, due to the pandemic we received only 1 submissions from team HW-TSC. We also used the cascaded speech translation cloud services from 2 providers which will be named as Online A and Online B.

Team HW-TSC participated in the Chinese-English unconstrained sub-task. HW-TSC submission is a cascaded system of a speech recognition system, a disfluency detection system, and a machine translation system. They simply extract the sound tracks from videos, then feed them to their proprietary ASR system and proceed transcripts to downstream modules. ASR outputs are piped into a BERT-based disfluency detection system which performs repeat spoken words removal, detect insertion and deletion noise. For the machine translation part, a transformer-big has been employed. They experimented multi-task learning with NMT decoding and domain classification, back translation and noise data augmentation. For the details of their approach, please refer to their paper (Table 1).

3.4 Results

We use vizseq[4] as our main scoring tool. We evaluate ASR systems in CER without punctuations. The final translation outputs are evaluated with lower-cased BLEU, METEOR, and chrF. We also break down the translation performances by the CER error buckets with sentence-level BLEU scores. HW-TSC has a better corpus-level performance than other online cloud services. All systems are sensitive to speech recognition errors.

4 Offline Speech Translation

In continuity with last year (Niehues et al., 2019), the offline speech translation task required participants to translate English audio data extracted from TED talks[5] into German. Participants could submit translations produced by either *cascade* architectures (built on a pipeline of ASR and MT components) or *end-to-end* models (neural solutions for the direct translation of the input audio), and were asked to specify, at submission time, which of the two architectural choices was made for their system.

Similar to last year, valid end-to-end submissions had to be obtained by models that:

- Do not exploit intermediate discrete representations (e.g., source language transcrip-

[2] https://srvk.github.io/how2-dataset/
[3] https://github.com/nguyenbh/iwslt2020_video_translation
[4] https://github.com/facebookresearch/vizseq
[5] http://www.ted.com

tion or hypotheses fusion in the target language);

- Rely on parameters that are all jointly trained on the end-to-end task

4.1 Challenge

While the cascade approach has been the dominant one for years, the end-to-end paradigm has recently attracted increasing attention as a way to overcome some of the pipeline systems' problems, such as higher architectural complexity and error propagation. In terms of performance, however, the results of the IWSLT 2019 ST task still showed a gap between the two approaches that, though gradually decreasing, was still of about 1.5 BLEU points. In light of this, the main question we wanted to answer this year is: *is the cascaded solution still the dominant technology in spoken language translation?* To take stock of the situation, besides being allowed to submit systems based on both the technologies, participants were asked to translate also the 2019 test set, which last year was kept undisclosed to enable future comparisons.

This year's evaluation also focused on a key issue in ST, which is *the importance of a proper segmentation of the input audio*. One of the findings of last year's campaign, which was carried out on unsegmented data, was indeed the key role of automatically segmenting the test data in way that is close to the sentence-level one present in the training corpora. To shed light on this aspect, the last novelty introduced this year is the possibility given to participants to process the same test data released in two versions, namely with and without pre-computed audio segmentation. The submission instructions included the request to specify, together with the type of architecture (cascade/end-to-end) and the data condition (constrained/unconstrained – see §4.2) also the chosen segmentation type (own/given).

Systems' performance is evaluated with respect to their capability to produce translations similar to the target-language references. To enable performance analyses from different perspectives, such similarity is measured in terms of multiple automatic metrics: case-sensitive/insensitive BLEU (Papineni et al., 2002b), case-sensitive/insensitive TER (Snover et al., 2006a), BEER (Stanojevic and Sima'an, 2014), and CharacTER (Wang et al., 2016). Similar to last year, the submitted runs are ranked based on the case-sensitive BLEU calculated on the test set by using automatic re-segmentation of the hypotheses based on the reference translations by mwerSegmenter.[6]

4.2 Data

Training and development data. Also this year, participants had the possibility to train their systems using several resources available for ST, ASR and MT. The training corpora allowed to satisfy the "constrained" data condition include:

- MuST-C (Di Gangi et al., 2019a)
- WIT[3] (Cettolo et al., 2012)
- Speech-Translation TED corpus[7]
- How2 (Sanabria et al., 2018)[8]
- LibriVoxDeEn (Beilharz and Sun, 2019)[9]
- Europarl-ST (Iranzo-Sánchez et al., 2020)
- TED LIUM v2 (Rousseau et al., 2014) and v3 (Hernandez et al., 2018)
- all the data provided by WMT 2019[10]
- OpenSubtitles 2018 (Lison et al., 2018)
- Augmented LibriSpeech (Kocabiyikoglu et al., 2018)[11]
- Mozilla Common Voice[12]
- LibriSpeech ASR corpus (Panayotov et al., 2015)

The list of allowed development data includes the dev set from IWSLT 2010, as well as the test sets used for the 2010, 2013, 2014, 2015 and 2018 IWSLT campaigns. Using other training/development resources was allowed but, in this case, participants were asked to mark their submission as an "unconstrained" one.

[6] https://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz

[7] http://i13pc106.ira.uka.de/~mmueller/iwslt-corpus.zip

[8] only English - Portuguese

[9] only German - English

[10] http://www.statmt.org/wmt19/

[11] only English - French

[12] https://voice.mozilla.org/en/datasets – English version en_1488h_2019-12-10

Test data. A new test set was released by processing, with the same pipeline used to build MuST-C (Di Gangi et al., 2019a), a new set of 22 talks that are not included yet in the public release of the corpus. To measure technology progress with respect to last year's round, participants were asked to process also the undisclosed 2019 test set. Both test corpora were released with and without sentence-like automatic segmentation. For the segmented versions, the resulting number of segments is 2,263 (corresponding to about 4.1 hours of translated speech from 22 talks) for the 2020 test set and 2,813 (about 5.1 hours from 25 talks) for the 2019 test set.

4.3 Submissions

We received submissions from 10 participants (twice as much compared to last year's number) coming from the industry, the academia and other research institutions. Eight teams submitted at least one run obtained with end-to-end technology, showing a steady increase of interest towards this emerging paradigm. In detail:

- 5 teams (DiDiLabs, FBK, ON-TRAC, BHANSS, SRPOL) participated only with end-to-end systems;
- 3 teams (AppTek/RWTH, KIT, HY) submitted runs obtained from both cascade and end-to-end systems;
- 2 teams (AFRL, BUT) participated only with cascade systems.

As far as input segmentation is concerned, participants are equally distributed between the two possible types, with half of the total submitting only runs obtained with the given segmentation and the other half submitting at least one run with in-house solutions. In detail:

- 5 teams (BHANSS, BUT, DiDiLabs, FBK, HY) participated only with the given segmentation of the test data;
- 2 teams (AFRL, ON-TRAC) participated only with their own segmentation;
- 3 teams (AppTek/RWTH, KIT, SRPOL) submitted runs for both segmentation types.

Finally, regarding the data usage possibilities, all teams opted for constrained submissions exploiting only the allowed training corpora listed in §4.2.

In the following, we provide a bird's-eye description of each participant's approach.

AFRL (Ore et al., 2020) participated with a cascade system that included the following steps: (1) speech activity detection using a neural network trained on TED-LIUM, (2) speech recognition using a Kaldi system (Povey et al., 2011) trained on TED-LIUM, (3) sentence segmentation using an automatic punctuator (a bidirectional RNN with attention trained on TED data using Ottokar Tilk[13]), and (4) machine translation using OpenNMT (Klein et al., 2017). The contrastive system differs from the primary one in two aspects: Step 3 was not applied, and the translation results were obtained using Marian (Junczys-Dowmunt et al., 2018) instead of openNMT.

AppTek/RWTH (Bahar et al., 2020b) participated with both cascade and end-to-end speech translation systems, paying attention to careful data selection (based on sentence embedding similarity) and weighting. In the cascaded approach, they combined: (1) high-quality hybrid automatic speech recognition (based on hybrid LSTM/HMM model and attention models trained on data augmented with a variant SpecAugment (Park et al., 2019), layer-wise pretraining and CTC loss (Graves et al., 2006) as additional loss), with (2) the Transformer-based neural machine translation. The end-to-end direct speech translation systems benefit from: (1) pre-training of adapted LSTM-based encoder and Transformer-based decoder components, (2) an adapter component in-between, and (3) synthetic data and fine-tuning. All these elements make the end-to-end models able to compete with the cascade ones in terms of MT quality.

BHANSS (Lakumarapu et al., 2020) built their end-to-end system adopting the Transformer architecture (Vaswani et al., 2017a) coupled with the meta-learning approach proposed in (Indurthi et al., 2020). Meta-learning is used mitigate the issue of over-fitting when the training data is limited, as in the ST case, and allows their system to take advantage of the available ASR and MT data. Along with meta-learning, the submitted system also exploits training on synthetic data created with different techniques. These include automatic English to German translation to generate artificial text data, and speech perturbation with

[13] https://pypi.org/project/punctuator/

the Sox audio manipulation tool[14] to generate artificial audio data similar to (Potapczyk et al., 2019).

BUT (unpublished report) participated with cascade systems based on (Vydana et al., 2020). They rely on ASR-MT Transformer models connected through neural hidden representations and jointly trained with ASR objective as an auxiliary loss. At inference time, both models are connected through n-best hypotheses and the hidden representation that correspond to the n-best hypotheses. The n-best hypothesis from the ASR model are processed in parallel by the MT model. The likelihoods of the final MT decoder are conditioned on the likelihoods of the ASR model. The discrete symbol token sequence, which is obtained as the intermediate representation in the joint model, is used as an input to an independent text-based MT model, whose outputs are ensembled with the joint model. Similarly, the ASR module of the joint model is ensembled from a separately trained ASR model.

DiDiLabs (Arkhangorodsky et al., 2020) participated with an end-to-end system based on the S-Transformer architecture proposed in (Di Gangi et al., 2019b,c). The base model trained on MuST-C was extended in several directions by: (1) encoder pre-training on English ASR data, (2) decoder-pre-training on German ASR data, (3) using wav2vec (Schneider et al., 2019) features as inputs (instead of Mel-Filterbank features), and (4) pre-training on English to German text translation with an MT system sharing the decoder with S-Transformer, so to improve the decoder's translation ability.

FBK (Gaido et al., 2020) participated with an end-to-end-system adapting the S-Transformer model (Di Gangi et al., 2019b,c). Its training is based on: *i)* transfer learning (via ASR pre-training and – word/sequence – knowledge distillation), *ii)* data augmentation (with SpecAugment (Park et al., 2019), time stretch (Nguyen et al., 2020a) and synthetically-created data), *iii)* combining synthetic and real data marked as different "domains" as in (Di Gangi et al., 2019d), and *iv)* multitask learning using the CTC loss (Graves et al., 2006). Once the training with word-level knowledge distillation is complete the model is fine-tuned using label smoothed cross entropy (Szegedy et al., 2016).

HY (Vázquez et al., 2020) participated with both cascade and end-to-end systems. For the end-to-end system, they used a multimodal approach (with audio and text as the two modalities treated as different languages) trained in a multi-task fashion, which maps the internal representations of different encoders into a shared space before decoding. To this aim, they incorporated the inner-attention based architecture proposed by (Vázquez et al., 2020) within Transformer-based encoders (inspired by (Tu et al., 2019; Di Gangi et al., 2019c)) and decoders. For the cascade approach, they used a pipeline of three stages: (1) ASR (trained with S-Transformer (Di Gangi et al., 2019c)), (2) re-punctuation and letter case restoration (based on Marian's implementation (Junczys-Dowmunt et al., 2018) of Transformer), and (3) MT (also based on Marian).

KIT (Pham et al., 2020) participated with both end-to-end and cascade systems. For the end-to-end system they applied a deep Transformer with stochastic layers (Pham et al., 2019b). Position encoding (Dai et al., 2019) is incorporated to mitigate issues due to processing long audio inputs, and SpecAugment (Park et al., 2019) is applied to the speech inputs for data augmentation. The cascade architecture has three components: (1) ASR (both LSTM (Nguyen et al., 2020b) and Transformer-based (Pham et al., 2019a)) (2) Segmentation (with a monolingual NMT system (Sperber et al., 2018) that adds sentence boundaries and case, also inserting proper punctuation), and (3) MT (a Transformer-based encoder-decoder model implementing Relative Attention following (Dai et al., 2019) adapted via fine-tuning on data incorporating artificially-injected noise). The WerRTCVAD toolkit[15] is used to process the unsegmented test set.

ON-TRAC (Elbayad et al., 2020) participated with end-to-end systems, focusing on speech segmentation, data augmentation and the ensembling of multiple models. They experimented with several attention-based encoder-decoder models sharing the general backbone architecture described in (Nguyen et al., 2019), which comprises an encoder with two VGG-like (Simonyan and Zisserman, 2015) CNN blocks followed by five stacked BLSTM layers. All the systems were developed using the ESPnet end-to-end speech processing toolkit (Watanabe et al., 2018). An ASR

[14] http://sox.sourceforge.net/

[15] https://github.com/wiseman/py-webrtcvad

model trained on Kaldi (Povey et al., 2011) was used to process the unsegmented test set, training the acoustic model on the TED-LIUM 3 corpus. Speech segments based on the recognized words with timecodes were obtained with rules, whose thresholds were optimised to get a segment duration distribution in the development and evaluation data that is similar to the one observed in the training data. Data augmentation was performed with SpecAugment (Park et al., 2019), speed perturbation, and by automatically translating into German the English transcription of MuST-C and How2. The two synthetic corpora were combined in different ways producing different models that were eventually used in isolation and ensembled at decoding time.

SRPOL (Potapczyk and Przybysz, 2020) participated with end-to-end systems based on the one (Potapczyk et al., 2019) submitted to the IWSLT 2019 ST task. The improvements over last year's submission include: (1) the use of additional training data (synthetically created, both by translating with a Transformer model as in (Jia et al., 2019) and via speed perturbation with the Sox audio manipulation tool); (2) training data filtration (applied to WIT^3 and TED LIUM v2); (3) the use of SpecAugment (Park et al., 2019); (4) the introduction of a second decoder for the ASR task, obtaining a multitask setup similar to (Anastasopoulos and Chiang, 2018); (5) the increase of the encoder layer depth; (6) the replacement of simpler convolutions with Resnet-like convolutional layers; and (7) the increase of the embedding size. To process the unsegmented test set, the same segmentation technique used last year was applied. It relies on iteratively joining, up to a maximal length of 15s, the fragments obtained by dividing the audio input with a silence detection tool.

4.4 Results

Detailed results for the offline ST task are provided in Appendix A.3. For each test set (i.e. this year's *tst2020* and last year's *tst2019*), the scores computed on unsegmented and segmented data (i.e. *own* vs *given* segmentation) are reported separately. Background colours are used to differentiate between cascade (white background) and end-to-end architectures (grey).

Cascade vs end-to-end. Looking at the results computed with case-sensitive BLEU (our primary evaluation metric), the first interesting thing to remark is that the highest score (25.3 BLEU) is achieved by an end-to-end system, which outperforms the best cascade result by 0.24 BLEU points. Although the performance difference between the two paradigms is small, it can be considered as an indicator of the steady progress done by end-to-end approaches to ST. Back to our initial question *"is the cascaded solution still the dominant technology in ST?"*, we can argue that, at least in this year's evaluation conditions, the two paradigms are now close (if not on par) in terms of final performance.

The importance of input segmentation. Another important aspect to consider is the key role played by a proper segmentation of the input speech. Indeed, the top five submitted runs are all obtained by systems operating under the "unsegmented" condition, that is with own segmentation strategies. This is not surprising considering the mismatch between the provided training material (often "clean" corpora split into sentence-like segments, as in the case of MuST-C) and the supplied test data, whose automatic segmentation can be far from being optimal (i.e. sentence-like) and, in turn, difficult to handle. The importance of a good segmentation becomes evident looking at the scores of those teams that participated with both segmentation types (i.e. AppTek/RWTH, KIT, SRPOL): in all cases, their best runs are obtained with own segmentations. Looking at these systems through the lens of our initial question about the distance between cascade and end-to-end approaches, it's interesting to observe that, although the two approaches are close when participants applied their own segmentation, the cascade is still better when results are computed on pre-segmented data.[16] Specifically, on unsegmented data, AppTek/RWTH's best cascade score (22.49 BLEU) is 2 points better than their best end-to-end score (20.5). For KIT's submissions the distance is slightly larger (22.06 - 19.82 = 2.24). In light of this consideration, as of today it is still difficult to draw conclusive evidence about the real distance between cascade and end-to-end ST since the effectiveness of the latter seems to highly depend a critical pre-processing step.

Progress wrt 2019. Comparing participants' results on *tst2020* and *tst2019*, the progress made by

[16] This is only possible for the submissions by AppTek/RWTH and KIT, since SRPOL participated only with their own segmentation.

the ST community is quite visible. Before considering the actual systems' scores, it's worth observing that the overall ranking is almost identical on the two test sets. This indicates that the top-ranked approaches on this year's evaluation set are consistently better on different new test data coming from the TED Talks domain. Three systems, two of which end-to-end, were able to outperform last year's top result (21.55 BLEU), which was obtained by a cascade system. Moreover, two out of the three systems that also took part in the IWSLT 2019 campaign (FBK, KIT and SRPOL) managed to improve their previous scores on the same dataset. In both cases, they did it with a large margin: from 3.85 BLEU points for FBK to 4.0 BLEU points for SRPOL. As the 2019 test set was kept undisclosed, this is another confirmation of the progress made in one year by ST technology in general, and by the end-to-end approach in particular.

5 Conversational Speech Translation

In conversational speech, there are many phenomena which aren't present in well-formed text, such as disfluencies. Disfluencies comprise e.g., filler words, repetitions, corrections, hesitations, or incomplete sentences. This differs strongly from typical machine translation training data. This mismatch needs to be accounted for when translating conversational speech both for domain mismatch as well as generating well-formed, fluent translations. While previously handled with intermediate processing steps, with the rise of end-to-end models, how and when to incorporate such a pre- or post-processing steps between speech processing and machine translation is an open question.

Disfluency removal typically requires token-level annotations for that language. However, most languages and translation corpora do not such annotations. Using recently collected fluent references (Salesky et al., 2018) for the common Fisher Spanish-English dataset, this task poses several potential questions: how should disfluency removal be incorporated into current conversational speech translation models where translation may not be done in a pipeline, and can this be accomplished without training on explicit annotations?

5.1 Challenge

The goal of this task is to provide fluent, English translations given disfluent Spanish speech or text. We provide three ways in which submissions may differ and would be scored separately:

- Systems which translate from speech, or from text-only
- Systems may be *unconstrained* (use additional data beyond what is provided) or *constrained* (use only the Fisher data provided)
- Systems which do and do not use the fluent references to train

Submissions were scored against the fluent English translation references for the challenge test sets, using the automatic metric BLEU (Papineni et al., 2002a) to assess fluent translations and METEOR (Lavie and Agarwal, 2007) to assess meaning preservation from the original disfluent data. By convention to compare with previous published work on the Fisher translation datasets (Post et al., 2013), we score using lowercased, detokenized output with all punctuation except apostrophes removed. At test time, submissions could only be provided with the evaluation data for their track. We compare submissions to the baseline models described in Salesky et al. (2019).

5.2 Data

This task uses the LDC Fisher Spanish speech (disfluent) (Graff et al.) with new target translations (fluent) Salesky et al. (2018). This dataset has 160 hours of speech (138k utterances): this is a smaller dataset than other tasks, designed to be approachable. We provide multi-way parallel data for training:

- disfluent Spanish speech
- disfluent Spanish transcripts (gold)
- disfluent Spanish transcripts (ASR output)
- disfluent English translations
- fluent English translations

Each of these are parallel at level of the training data, such that the disfluent and fluent translation references have the same number of utterances. Additional details for the fluent translations

can be found here: Salesky et al. (2018). We arranged an evaluation license agreement with the LDC where all participants could receive this data without cost for the purposes of this task.

The `cslt-test` set is originally Fisher dev2 (for which the fluent translations are released for this first time with this task). We provided participants with two conditions for each test set for the text-only track: gold Spanish transcriptions, and ASR output using the baseline's ASR model.

5.3 Submissions

We received two submissions, both for the text-only track, as described below.

Both teams described both constrained and unconstrained systems. While NAIST submitted multiple (6) systems, IIT Bombay submitted ultimately only their unconstrained system. Both teams submitted at least one model without fluent translations used in training – rising to the challenge goal of this task to generalize beyond available annotations.

NAIST (Fukuda et al., 2020) used a two-pronged approach: first, to leverage both a larger dataset which is out-of-domain (UN Corpus: i.e. both fluent and also out-of-domain for conversational speech) they utilize an unsupervised style transfer model, and second, to adapt between fluent and disfluent parallel corpora for NMT they pretrain on the original disfluent-disfluent translations and fine-tune to the target disfluent-fluent case. They find that their style transfer domain adaptation was necessary to make the most effective use of style-transfer, as without it, the domain mismatch was such that meaning was lost during disfluent-fluent translation.

IIT Bombay (Saini et al., 2020) submit both unconstrained and constrained systems, both without use of the parallel fluent translations. They use data augmentation through noise induction to create disfluent–fluent English references from English NewsCommentary. Their translation model uses multiple encoders and decoders with shared layers to balance shared modeling capabilities while separating domain-specific modeling of e.g. disfluencies within noised data.

5.4 Results

This task proved challenging but was met by very inventive and different solutions from each team. Results are shown in Appendix A.4.

In their respective description papers, the two teams scored their systems differently, leading to different trends between the two papers than may be observed in our evaluation.

The unconstrained submissions from each site utilized external data in very different ways, though with the same underlying motivation. Under the matched condition — unconstrained but no fluent references used during training — given gold source Spanish transcripts, The submissions from NAIST (Fukuda et al., 2020) were superior by up to 2.6 BLEU. We see that this is not the case, however, when ASR output is the source, where the IITB submission performs ≈3.4 better on BLEU; this submission, in fact, outperforms all submitted under any condition, though it has not been trained on the parallel fluent references. This may suggest perhaps that the multi-encoder and multi-decoder machine translation model from IITB transferred better to the noise seen in ASR output. Interestingly, we see a slight improvement in BLEU for both sites with ASR output as source under this matched conditions (e.g. for those models where the fluent data is not used).

Turning to our second metric, METEOR, where we assess meaning preservation with the original disfluent references, we see that the IITB submission from ASR output preserves much more of the content contained in the disfluent references, resulting in a much higher METEOR score than all other submissions. The utterances in these outputs are also 10% longer than those of NAIST-e. Qualitatively, these segments also appear to have more repetitions than the equivalents translated from gold transcripts. This suggests perhaps that NAIST's noised training using the additional unconstrained data may have transferred better to the noise seen in ASR output, causing less of a change given this challenge condition. This may not reflected by BLEU computed against fluent references, because in addition to removing disfluent content, other tokens have been changed. This reminds us this metric may not capture all aspects of producing fluent translations.

NAIST submitted 6 models, allowing us to see additional trends though there are no additional submissions with matched conditions. The unconstrained setting where they leveraged noising of UN Corpus data gave significant improvements of ≈ 5 BLEU. Surprisingly to us, their submissions

which do not leverage fluent references in training are not far behind those which do — the respective gap between otherwise matched submissions is typically ≈ 2 BLEU.

Overall, we are very encouraged to see submissions which did not use the fluent parallel data, and encourage further development in this area!

6 Open Domain Translation

The goals of this task were to further promote research on translation between Asian languages, the exploitation of noisy parallel web corpora for MT, and thoughtful handling of data provenance.

6.1 Challenge

The open domain translation task focused on machine translation between Chinese and Japanese, with one track in each direction. We encouraged participation in both tracks.

We provided two bilingual parallel Chinese-Japanese corpora, and two additional bilingual Zh-Ja corpora. The first was a large, noisy set of segment pairs assembled from web data. Section 6.2 describes the data, with further details in Appendix A.5. The second set was a compilation of existing Japanese-Chinese parallel corpora from public sources. These include both freely-downloadable resources and ones released as part of previous Chinese-Japanese MT efforts. We encouraged participants to use only these provided corpora. The use of other data was allowed, as long as it was disclosed.

The submitted systems were evaluated on a held-out, mixed-genre, test set curated to contain high-quality segment pairs. The official evaluation metric was 4-gram character BLEU (Papineni et al., 2002c). The scoring script[17] was shared with participants before the evaluation phase.

6.2 Parallel Training Data

We collected all the publicly available, parallel Chinese-Japanese corpora we could find, and made it available to participants as the `existing_parallel`. These include Global Voices, News Commentary, and Ubuntu corpora from OPUS Tiedemann (2012); OpenSubtitles (Lison and Tiedemann, 2016); TED talks (Dabre and Kurohashi, 2017); Wikipedia (Chu et al., 2014, 2015); Wiktionary.org; and WikiMatrix (Schwenk et al., 2019). We also collected parallel sentences from Tatoeba.org, released under a CC-BY License. Table 2 lists the size of each of these existing corpora. In total, we found fewer than 2 million publicly available Chinese-Japanese parallel segments.

Corpus	Segments	ZH Chars
Crawled (pipeline)	18,966,595	493,902,539
Ubuntu	92,250	1,549,964
Open Subtitles	914,355	10,932,722
TED	376,441	5,345,867
Global Voices	16,848	337,194
Wikipedia	228,565	5,067,489
Wiktionary	62,557	222,562
News Commentary	570	65,038
Tatoeba	4,243	50,846
WikiMatrix	267,409	9,950,657
Total	**20,929,833**	**527,424,878**

Table 2: Provided Chinese-Japanese parallel data.

We therefore built a data-harvesting pipeline to crawl the web for more parallel text. The data collection details can be found in Appendix A.5. The result was the `webcrawled_parallel_filtered` dataset, containing nearly 19M hopefully-parallel segment pairs (494M Zh chars) with provenance information. This crawled data combined with the existing corpora provide 20.9M parallel segments with 527M Chinese characters. We included provenance information for each segment pair.

6.3 Unaligned and Unfiltered Data

In addition to the aligned and filtered output of the pipeline, we released two other variations on the pipeline output. We hoped these larger yet noisier versions of the data would be of use for working on upstream data processing.

We provided a larger aligned, but unfiltered, version of the web-crawled data produced by the pipeline after Stage 5 (`webcrawled_parallel_unfiltered`). This corpus contains 161.5M segment pairs, and is very noisy (e.g. it includes languages other than Chinese and Japanese). Our expectation is that more sophisticated filtering of this noisy data will increase the quantity of good parallel data.

We also released the parallel document contents, with boundaries, from Step 4 in the

[17] https://github.com/didi/iwslt2020_open_domain_translation/blob/master/eval/bleu.py

pipeline shown in Appendix A.5. These documents are the contents of the webpages paired by URL (e.g. `gotokyo.org/jp/foo` and `gotokyo.org/zh/foo`), and processed with `BeautifulSoup`, but before using Hunalign (Varga et al., 2005) to extract parallel sentence pairs. We released 15.6M document pairs as `webcrawled_unaligned`. Sentence aligner improvements (and their downstream effects) could be explored using this provided data.

6.4 Dev and Test Sets

The provided development set consisted of 5304 basic expressions in Japanese and Chinese, from the Kurohashi-Kawahara Lab at Kyoto University.[18] The held-out test set was intended to cover a variety of topics not known to the participants in advance. We selected test data from high-quality (human translated) parallel web content, authored between January and March 2020. The test set curation process can be found in Appendix A.5.

This curation produced 1750 parallel segments, which we divided randomly in half: 875 lines for the Chinese-to-Japanese translation test set, and 875 lines for the other direction. The Japanese segments have an average length of 47 characters, and the Chinese ones have an average length of 35.

6.5 Submissions

Twelve teams submitted systems for both translation directions, and three more submitted only for Japanese-to-Chinese. Of the 15 participants, 6 were from academia and 9 were from industry.

We built a baseline system before the competition began, based on Tensor2Tensor (Vaswani et al., 2018), and provided participants with the baseline BLEU scores to benchmark against. We also provided the source code for training the baseline, as a potential starting point for experimentation and development. Our source code for the baseline system is now publicly available.[19]

The following summarizes some key points of the participating teams that submitted system descriptions; broad trends first, and then the individual systems in reverse-alphabetical order. Further details for these systems can be found in the relevant system description papers in the full workshop proceedings.

Architecture: All participants used either the Transformer architecture (Vaswani et al., 2017b) or a variant, such as dynamic linear combination of layers, or transformer-evolved with neural architecture search. Most participants submitted ensemble models, showing consistent improvement over the component models on the dev set.

Data Filtering: As anticipated, all teams invested significant effort in data cleaning, normalization and filtering of the provided noisy corpora. A non-exhaustive list of the techniques used includes length ratios, language id, converting traditional Chinese characters to simplified, sentence deduplication, punctuation normalization, and removing html markup.

XIAOMI (Sun et al., 2020) submitted a large ensemble, exploring the performance of a variety of Transformer-based architectures. They also incorporated domain adaptation, knowledge distillation, and reranking.

TSUKUBA (Cui et al., 2020) used the unfiltered data for backtranslation, augmented with synthetic noise. This was done in conjunction with n-best list reranking.

SRC-B (Samsung Beijing) (Zhuang et al., 2020) mined the provided unaligned corpus for parallel data and for backtranslation. They also implemented relative position representation for their Transformer.

OPPO (Zhang et al., 2020) used detailed rule-based preprocessing and multiple rounds of backtranslation. They also explored using both the unfiltered parallel dataset (after filtering) and the unaligned corpus (after alignment). Their contrastive system shows the effect of character widths on the BLEU score.

OCTANOVE (Hagiwara, 2020) augmented the dev set with high-quality pairs mined from the training set. This reduced the size of the webcrawled data by 90% before using. Each half of the discarded pairs was reused for backtranslation.

ISTIC (Wei et al., 2020) used the provided unfiltered webcrawl data after significant filtering. They also used adaptation, using elasticsearch to find sentence pairs similar to the test set, and optimizing the system on them.

DBS Deep Blue Sonics (Su and Ren, 2020) successfully added noise to generate augmented data for backtranslation. They also experimented with

[18] `http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JEC%20Basic%20Sentence%20Data`

[19] DiDi baseline source code available at: github.com/didi/iwslt2020_open_domain_translation

language model fusion techniques.

CASIA (Wang et al., 2020b) ensembled many models into their submission. They used the unfiltered data for backtranslation, used a domain classifier based on segment provenance, and also performed knowledge-distillation. They also used 13k parallel sentences from external data; see the "External data" note in Section 6.6.

6.6 Results and Discussion

Appendix A.5 contains the results of the Japanese-to-Chinese and Chinese-to-Japanese open-domain translation tasks. Some comments follow below.

Data filtering was unsurprisingly helpful. We released 4 corpora as part of the shared task. All participants used `existing_parallel` and `webcrawled_parallel_filtered`. Overall, participants filtered out 15%-90% of the data, and system performance increased by around 2-5 BLEU points. The `webcrawled_parallel_unfiltered` corpus was also used successfully, but required even more aggressive filtering. The `webcrawled_unaligned` data was even harder to use, and we were pleased to see some teams rise to the challenge. *Data augmentation* via backtranslation also consistently helped. However, there was interesting variation in how participants selected the data to be translated. *Provenance* information is not common in MT evaluations; we were curious how it would be used. Hagiwara (2020) tried filtering `web_crawled_parallel_filtered` using a provenance indicator, but found it was too aggressive. Wang et al. (2020b) instead trained a domain classifier, and used it at decoding time to reweight the domain-specific translation models in the ensemble.

External data was explicitly allowed, potentially allowing the sharing of external resources that were unknown to us. Hagiwara (2020) improved on their submitted system, in a separate experiment, by gathering 80k external parallel question-answer pairs from HiNative and incorporating them into the training set. Wang et al. (2020b) also improved their system by adding 13k external sentence pairs from hujiangjp. However, this inadvertently included data from one of the websites from which the task's blind test set was drawn, resulting in 383/875 and 421/875 exact matching segments on the Chinese side and Japanese side respectively.

Overall, we are heartened by the participation in this first edition of the open-domain Chinese-Japanese shared task, and encourage participation in the next one.

7 Non-Native Speech Translation

The non-native speech translation task has been added to IWSLT this year. The task focuses on the very frequent setting of non-native spontaneous speech in somewhat noisy conditions, one of the test files even contained speech transmitted through a remote conferencing platform. We were interested in submissions of both types: the standard two-stage pipeline (ASR+MT, denoted "Cascaded") as well as end-to-end ("E2E") solutions.

This first year, we had English as the only source language and Czech and German as the target languages. Participants were allowed to submit just one of the target languages.

The training data sets permitted for "constrained" submissions were agreed upon the training data with the Offline Translation Task (Section 4) so that task participants could reuse their systems in both tasks. Participants were however also allowed to use any other training data, rendering their submissions "unconstrained".

7.1 Challenge

The main evaluation measure is translation quality but we invited participants to report time-stamped outputs if possible, so that we could assess their systems also using metrics related to *simultaneous* speech translation.

In practice, the translation quality is severely limited by the speech recognition quality. Indeed, the nature of our test set recordings is extremely challenging, see below. For that reason, we also asked the participants with cascaded submissions to provide their intermediate ASR outputs (again with exact timing information, if possible) and score it against our golden transcripts.

A further critical complication is the lack of input sound segmentation to sentence-like units. The Offline Speech Translation Task (Section 4) this year allowed the participants to come up either with their own segmentation, or to rely upon the provided sound segments. In the Non-Native task, no sound segmentation was available. In some cases, this could have caused even a computational challenge, because our longest test document is

25:55 long, well beyond the common length of segments in the training corpora. The reference translations in our test set do come in segments and we acknowledge the risk of automatic scores being affected by the (mis-)match of candidate and reference segmentation, see below.

7.1.1 SLT Evaluation Measures

The SLT evaluation measures were calculated by SLTev,[20] a comprehensive tool for evaluation of (on-line) spoken language translation.

SLT Quality (BLEU$_1$ and BLEU$_{mw}$) As said, we primarily focus on *translation quality* and we approximate it with BLEU (Papineni et al., 2002a) for simplicity, despite all the known shortcomings of the metric, e.g. Bojar et al. (2010).

BLEU was designed for text translation with a clear correspondence between source and target segments (sentences) of the text. We have explored multiple ways of aligning the segments produced by the participating SLT systems with the reference segments. For systems reporting timestamps of individual source-language words, the segment-level alignment can be based on the exact timing. Unfortunately, only one system provided this detailed information, so we decided to report only two simpler variants of BLEU-based metrics:

BLEU$_1$ The whole text is concatenated and treated as *one* segment for BLEU. Note that this is rather inappropriate for longer recordings where many n-grams could be matched far from their correct location.

BLEU$_{mw}$ (mwerSegmenter + standard BLEU). For this, first we concatenate the whole document and segment it using the mwerSegmenter tool (Matusov et al., 2005). Then we calculate the BLEU score for each document in the test set and report the average.

Since the BLEU implementations differ in many details, we rely on a stable one, namely sacreBLEU (Post, 2018).[21]

SLT Simultaneity In online speech translation, one can trade translation quality for delay and vice versa. Waiting for more input generally allows the system to produce a better translation. A compromise is sought by systems that quickly produce first candidate outputs and *update* them later, at the cost of potentially increasing cognitive load for the user by showing output that will become irrelevant.

The key properties of this trade-off are captured by observing some form of *delay*, i.e. how long the user has to wait for the translation of the various pieces of the message compared to directly following the source, and *flicker*, i.e. how much "the output changes". We considered several possible definitions of delay and flicker, including or ignoring information on timing, segmentation, word reordering etc., and calculated each of them for each submission. For simplicity, report only the following ones:

Flicker is inspired by Arivazhagan et al. (2019a). We report a normalized revision score calculated by dividing the total number of words produced by the true output length, i.e. by the number of words in the completed sentences. We report the average score across all documents in the test set.

Delay$_{ts}$ relies on timing information provided by the participants for individual *segments*. Each produced word is assumed to have appeared at the time that corresponds proportionally to its (character) position in the segment. The same strategy is used for the reference words. Note that the candidate segmentation does not need to match the reference one, but in both cases, we get an estimated time span for each word.

Delay$_{mw}$ uses mwerSegmenter to first find correspondences between candidate and reference segments based on the actual words. Then the same strategy of estimating the timing of each word is used.

The Delay is summed over all words and divided by the total number of words considered in the calculation to show the average delay per word.

Note that we use a simple exact match of the candidate and reference word; a better strategy would be to use some form of monolingual word alignment which could handle e.g. synonyms. In our case, non-matched words are ignored and do not contribute to the calculation of the delay at all,

[20] https://github.com/ELITR/SLTev

[21] We use the default settings, i.e. the signature `BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.4.6`.

Domain	Files	Overall Duration	Segments	EN Words	CS Words	DE Words
Antrecorp	28	0h38m	427	5040	4071	4660
Khan Academy	5	0h18m	346	2886	2272	2660
SAO	6	1h39m	654	11928	9395	10613
Total	39	2h35m	1427	19854	15738	17933

Table 3: Non-Native Speech Translation Task test data composition. Words are estimated simply by splitting at whitespace without tokenization.

reducing the reliability of the estimate. To provide an indication of how reliable the reported Delays are, we list also the percentage of reference words matched, i.e. successfully found in the candidate translation. This percentage ranges from 20% to up to 90% across various submissions.

Note that only one team provided us with timing details. In order to examine the empirical relations between these conflicting measures, we focus on the several contrastive runs submitted by this them in Section 7.4.1.

7.1.2 ASR Evaluation Measures

The ASR-related scores were also calculated by SLTev, using the script ASRev which assumes that the "translation" is just an identity operation.

We decided to calculate WER using two different strategies:

WER$_1$ concatenating all segments into one long sequence of tokens, and

WER$_{mw}$ first concatenating all segments provided by task participants and then using mwerSegmenter to reconstruct the segmentation that best matches the reference.

In both cases, we pre-process both the candidate and reference by lower casing and removing punctuation.

7.2 Data

7.2.1 Training Data for Constrained Submissions

The training data was aligned with the Offline Speech Translation Task (Section 4) to allow cross-submission in English-to-German SLT. English-to-Czech was unique to the Non-Native Task.

The permitted data for constrained submissions were:

For English ASR:

- LibriSpeech ASR corpus (Panayotov et al., 2015),
- Mozilla Common Voice,[22]
- Speech-Translation TED corpus.[23]

For English→Czech Translation:

- MuST-C (Di Gangi et al., 2019a), release 1.1 contains English-Czech pair,
- CzEng 1.7 (Bojar et al., 2016).[24] Note that CzEng overlaps with English-German test data of the Offline Speech Translation Task so it was not allowed to use this English-Czech corpus to train English-German (multi-lingual) systems.

For English→Czech Translation:

- All the data for English-German track by WMT 2019[25] News Translation Task, i.e.:
  - English-German parallel data,
  - German monolingual data,
- MuST-C (Di Gangi et al., 2019a), release 1.0 contains English-German pair,
- Speech-Translation TED corpus,[26] the English-German texts,
- WIT^3 (Cettolo et al., 2012).

[22] https://voice.mozilla.org/en/datasets – English version en_1488h_2019-12-10

[23] http://i13pc106.ira.uka.de/~mmueller/iwslt-corpus.zip

[24] https://ufal.mff.cuni.cz/czeng/czeng17

[25] http://www.statmt.org/wmt19/

[26] http://i13pc106.ira.uka.de/~mmueller/iwslt-corpus.zip

7.2.2 Test Data

The test set was prepared by the EU project ELITR[27] which aims at automatic simultaneous translation of speech into subtitles in the particular domain of conference speeches on auditing.

The overall size of the test set is in Table 3. The details about the preparation of test set components are in Appendix A.6.

7.3 Submissions

Five teams from three institutions took part in the task. Each team provided one "primary" submission and some teams provided several further "contrastive" submissions. The primary submissions are briefly described in Table 4. Note that two teams (APPTEK/RWTH and BUT) took the opportunity to reuse their systems from Offline Translation Task (Section 4) also in our task.

For the purposes of comparison, we also included freely available ASR services and MT services by two companies and denote the cascaded run for each of them as PUBLIC-A and PUBLIC-B. The ASR was run at the task submission deadline, the MT was added only later, on May 25, 2020.

7.4 Results

Appendix A.6 presents the results of the Non-Native Speech Translation Task for English→German and English→Czech, resp.

Note that the primary choice of most teams does not agree with which of their runs received the best scores in our evaluation. This can be easily explained by the partial domain mismatch between the development set and the test set.

The scores in both German and Czech results indicate considerable differences among the systems both in ASR quality as well as in BLEU scores. Before drawing strong conclusions from these scores, one has to consider that the results are heavily affected by the lack of reliable segmentation. If MT systems receive sequences of words not well matching sentence boundaries, they tend to reconstruct the sentence structure, causing serious translation errors.

The lack of golden sound segmentation also affects the evaluation: mwerSegmenter used in preprocessing of WER_{mw} and BLEU_{mw} optimizes WER score but it operates on a slightly different tokenization and casing. While the instability will be small in WER evaluation, it could cause more problems in BLEU_{mw}. Our BLEU calculation comes from sacreBLEU it its default setting. Furthermore, it needs to be considered that this is the first instance of the Non-Native shared task and not all peculiarities of the used evaluation measures and tools are quite known.[28] A manual evaluation would be desirable but even that would be inevitably biased depending on the exact way of presenting system outputs to the annotators. A procedure for a reliable manual evaluation of spoken language translation without pre-defined segmentation is yet to be sought.

The ASR quality scores[29] WER_1 and WER_{mw} are consistent with each other (Pearson .99), ranging from 14 (best submission by APPTEK/RWTH) to 33 WER_1. WER_{mw} is always 1–3.5 points absolute higher.

Translation quality scores BLEU_1 and BLEU_{mw} show a similarly high correlation (Pearson .987) and reach up to 16. For English-to-German, the best translation was achieved by the secondary submissions of APPTEK/RWTH, followed by the primary ELITR-OFFLINE and one of the secondary submissions of CUNI-NN. The public services seem to score worse, PUBLIC-B follows very closely and PUBLIC-A seems to seriously underperform, but it is quite possible that our cascaded application of their APIs was suboptimal. The only on-line set of submissions (ELITR) score between the two public systems.

The situation for English-to-Czech is similar, except that APPTEK/RWTH did not take part in this, so ELITR-OFFLINE provided the best ASR as well as translations (one of their secondary submissions).

Often, there is a big variance of BLEU scores across all the submissions of one team. This indicates that the test set was hard to prepare for and that for a practical deployment, testing on the real input data is critical.

As expected, the ASR quality limits the trans-

[27] http://elitr.eu/

[28] In our analysis, we also used BLEU as implemented in NLTK (Bird et al., 2009), observing substantial score differences. For instance, BUT1 received NLTK-BLEU of 12.68 instead of 0.63 reported in Appendix A.6 BLEU_{mw}. For other submissions, NLTK-BLEU dropped to zero without a clear reason, possibly some unexpected character in the output. The explanation of why NLTK can inflate scores is still pending but it should be performed to be sure that sacreBLEU does not unduly penalize BUT submissions.

[29] Note that the same ASR system was often used as the basis for translation into both Czech and German so the same ASR scores appear on multiple lines in Tables in Appendix A.6.

Team	Paper	Training Data	Off/On-Line	Cascaded
APPTEK/RWTH	Bahar et al. (2020a)†	Unconstrained	Off-Line	Cascaded
BUT	(unpublished draft)	Unconstrained	Off-Line	Ensemble E2E+Cascaded
CUNI	Polák et al. (2020)	Unconstrained	Off-Line	Cascaded
ELITR	Macháček et al. (2020)	Unconstrained	On-Line	Cascaded
ELITR-OFFLINE	Macháček et al. (2020)	Unconstrained	Off-Line	Cascaded
PUBLIC-A	– (public service)	Unconstrained	Off-Line	Cascaded
PUBLIC-B	– (public service)	Unconstrained	Off-Line	Cascaded

† The paper describes the basis of the systems but does not explicitly refer to non-native translation task.

Table 4: Primary submissions to Non-Native Speech Translation Task. The public web-based services were added by task organizers for comparison, no details are known about the underlying systems.

lation quality. WER_1 and BLEU_1 correlate negatively (Pearson -.82 for translation to German and -.66 for translation to Czech). Same correlations were observed for WER_{mw} and BLEU_{mw}.

The test set as well as the system outputs will be made available at the task web page[30] for future deep inspection.

7.4.1 Trade-Offs in Simultaneous SLT

The trade-offs in simultaneity of the translation can be studied only on submissions of ELITR, see Appendix A.6. We see that the Delay ranges between 1 and up to 2.5 seconds, with Delay_{mw} giving sligthly lower scores on average, correlated reasonably well with Delay_{ts} (Pearson .989). Delay into German seems higher for this particular set of MT systems.

The best score observed for Flicker is 5.18 and the worst is 7.51. At the same time, Flicker is not really negatively correlated with the Delays, e.g. Delay_{ts} vs. Flicker have the Pearson correlation of -.20.

Unfortunately, our current scoring does not allow to study the relationship between the translation quality and simultaneity, because our BLEU scores are calculated only on the final segments. Any intermediate changes to the translation text are not reflected in the scores.

Note that the timing information on when each output was produced was provided by the participants themselves. A fully reliable evaluation would require participants installing their systems on our hardware to avoid effects of network traffic, which is clearly beyond the goals of this task.

8 Conclusions

The evaluation campaign of the IWSLT 2020 conference offered six challenge tracks which attracted a total of 30 teams, both from academy and industry. The increasing number of participants witnesses the growing interest towards research on spoken language translation by the NLP community, which we believe has been partly driven by the availability of suitable training resources as well as the versatility of neural network models, which now permit to directly tackle complex tasks, such as speech-to-text translation, which formerly required building very complex system. We hope that this trend will continue and invite researchers interested in proposing new challenges for the next edition to get in touch with us. Finally, results of the human evaluation, which was still ongoing at the time of writing the overview paper, will be reported at the conference and will be included in an updated version of this paper.

[30] http://iwslt.org/doku.php?id=non_native_speech_translation

9 Acknowledgements

The offline Speech Translation task has been partially supported by the "End-to-end Spoken Language Translation in Rich Data Conditions" Amazon AWS ML Grant. The Non-Native Speech Translation Task was supported by the grants 19-26934X (NEUREM3) of the Czech Science Foundation, and H2020-ICT-2018-2-825460 (ELITR) of the EU. We are also grateful to Mohammad Mahmoudi for the assistance in the task evaluation and to Jonáš Kratochvíl for processing the input with public web ASR and MT services by two well-known companies.

The Open Domain Translation Task acknowledges the contributions of Yiqi Huang, Boliang Zhang and Arkady Arkhangorodsky, colleagues at DiDi Labs, for their help with the organization and sincerely thank Anqi Huang, a bilingual speaker, for validating the quality of the collected evaluation dataset.

References

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT04 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

Antonios Anastasopoulos and David Chiang. 2018. Tied Multitask Learning for Neural Speech Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

Naveen Arivazhagan, Colin Cherry, Te I, Wolfgang Macherey, Pallavi Baljekar, and George Foster. 2019a. Re-translation strategies for long form, simultaneous, spoken language translation.

Naveen Arivazhagan, Colin Cherry, Wolfgang Macherey, Chung-Cheng Chiu, Semih Yavuz, Ruoming Pang, Wei Li, and Colin Raffel. 2019b. Monotonic infinite lookback attention for simultaneous machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1313–1323, Florence, Italy. Association for Computational Linguistics.

Arkady Arkhangorodsky, Yiqi Huang, and Amittai Axelrod. 2020. DiDi Labs' End-to-End System for the IWSLT 2020 Offline Speech Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian Herold. 2020a. Start-Before-End and End-to-End: Neural Speech Translation by AppTek and RWTH Aachen University. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian Herold. 2020b. Start-Before-End and End-to-End: Neural Speech Translation by AppTek and RWTH Aachen University. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Benjamin Beilharz and Xin Sun. 2019. LibriVoxDeEn - A Corpus for German-to-English Speech Translation and Speech Recognition.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, pages 231–238, Cham / Heidelberg / New York / Dordrecht / London. Masaryk University, Springer International Publishing.

Ondřej Bojar, Kamil Kos, and David Mareček. 2010. Tackling Sparse Data Issue in Machine Translation Evaluation. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 86–91, Uppsala, Sweden. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, K. Sudoh, K. Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT 2017)*, pages 2–14, Tokyo, Japan.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. 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.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2015. The IWSLT 2015 Evaluation Campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT Evaluation Campaign. In *Proceedings of the Tenth International Workshop on Spoken Language Translation (IWSLT 2013)*, Heidelberg, Germany.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *Proceedings of the Eleventh International Workshop on Spoken Language Translation (IWSLT 2014)*, Lake Tahoe, USA.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2016. The IWSLT 2016 Evaluation Campaign. In *Proceedings of the 13th International Workshop on Spoken Language Translation (IWSLT 2016)*, Seattle, USA.

Pinzhen Chen, Nikolay Bogoychev, and Ulrich Germann. 2020. Character Mapping and Ad-hoc Adaptation: Edinburgh's IWSLT 2020 Open Domain Translation System. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Colin Cherry and George Foster. 2019. Thinking slow about latency evaluation for simultaneous machine translation. *arXiv preprint arXiv:1906.00048*.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2014. Constructing a Chinese-Japanese Parallel Corpus from Wikipedia. *LREC (International Conference on Language Resources and Evaluation)*.

Chenhui Chu, Toshiaki Nakazawa, and Sadao Kurohashi. 2015. Integrated Parallel Sentence and Fragment Extraction from Comparable Corpora: A Case Study on Chinese–Japanese Wikipedia. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Hongyi Cui, Yizhen Wei, Shohei Iida, Masaaki Nagata, and Takehito Utsuro. 2020. University of Tsukuba's Machine Translation System for IWSLT20 Open Domain Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Raj Dabre and Sadao Kurohashi. 2017. MMCR4NLP: multilingual multiway corpora repository for natural language processing. *CoRR*, abs/1710.01025.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019a. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota.

Mattia A. Di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessì, and Marco Turchi. 2019b. Enhancing Transformer for End-to-end Speech-to-Text Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 21–31, Dublin, Ireland. European Association for Machine Translation.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019c. Adapting Transformer to End-to-End Spoken Language Translation. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1133–1137. ISCA.

Mattia A. Di Gangi, Matteo Negri, and Marco Turchi. 2019d. One-To-Many Multilingual End-to-end Speech Translation. In *Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 585–592, Sentosa, Singapore.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–22, Pittsburgh, PA.

Maha Elbayad, Ha Nguyen, Fethi Bougares, Natalia Tomashenko, Antoine Caubrière, Benjamin Lecouteux, Yannick Estève, and Laurent Besacier. 2020. ON-TRAC Consortium for End-to-End and Simultanouus Speech Translation Challenge Tasks at IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2011. Overview of the IWSLT 2011 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 11–27, San Francisco, USA.

Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2012. Overview of the IWSLT 2012 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 11–27, Hong Kong, HK.

Cameron Shaw Fordyce. 2007. Overview of the IWSLT 2007 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Trento, Italy.

Ryo Fukuda, Katsuhito Sudoh, and Satoshi Nakamura. 2020. NAIST's Machine Translation Systems for IWSLT 2020 Conversational Speech Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Marco Gaido, Mattia Antonio Di Gangi, Matteo Negri, and Marco Turchi. 2020. End-to-End Speech-Translation with Knowledge Distillation: FBK@IWSLT2020. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

David Graff, Shudong Huang, Ingrid Cartagena, Kevin Walker, and Christopher Cieri. Fisher spanish speech (LDC2010S01). Https://catalog.ldc.upenn.edu/ ldc2010s01.

Alex Graves. 2016. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*.

Alex Graves, Santiago Fernández, Faustino J. Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd international conference on Machine learning (ICML)*, pages 369–376, Pittsburgh, Pennsylvania.

Masato Hagiwara. 2020. Octanove Labs' Japanese-Chinese Open Domain Translation System. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Houjeung Han, Mohd Abbas Zaidi, Sathish Indurthi, Nikhil Kumar Lakumarapu, Beomseok Lee, and Sangha Kim. 2020. End-to-End Simultaneous Translation System for the IWSLT2020 using Modality Agnostic Meta-Learning. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. *CoRR*, abs/1805.04699.

S. Indurthi, H. Han, N. K. Lakumarapu, B. Lee, I. Chung, S. Kim, and C. Kim. 2020. End-end speech-to-text translation with modality agnostic meta-learning. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7904–7908.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *Proc. of 45th Intl. Conf. on Acoustics, Speech, and Signal Processing (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. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Ali Can Kocabiyikoglu, Laurent Besacier, and Olivier Kraif. 2018. Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation. In *Proceedings of LREC 2018*, Miyazaki, Japan.

Nikhil Kumar Lakumarapu, Beomseok Lee, Sathish Indurthi, Houjeung Han, Mohd Abbas Zaidi, and Sangha Kim. 2020. End-to-End Offline Speech Translation System for IWSLT 2020 using Modality Agnostic Meta-Learning. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT)*, pages 228–231, Prague, Czech Republic.

Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M Cohen, Huyen Nguyen, and Ravi Teja Gadde. 2019. Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. *LREC (International Conference on Language Resources and Evaluation)*.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mingbo Ma, Liang Huang, Hao Xiong, Renjie Zheng, Kaibo Liu, Baigong Zheng, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, Hua Wu, and Haifeng Wang. 2019. STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

Dominik Macháček, Jonáš Kratochvíl, Tereza Vojtěchová, and Ondřej Bojar. 2019. A speech test set of practice business presentations with additional relevant texts. In *Statistical Language and Speech Processing*, pages 151–161, Cham, Switzerland. Springer Nature Switzerland AG.

Dominik Macháček, Jonáš Kratochvíl, Sangeet Sagar, Matús Žilinec, Ondřej Bojar, Thai-Son Nguyen, Felix Schneider, Philip Williams, and Yuekun Yao. 2020. ELITR Non-Native Speech Translation at IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

E. Matusov, G. Leusch, O. Bender, , and H. Ney. 2005. Evaluating Machine Translation Output with Automatic Sentence Segmentation. In *Proceedings of the 2nd International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, USA.

I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. 2005. The ami meeting corpus. In *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J.*

Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen: Noldus Information Technology.

Ha Nguyen, Natalia, Marcely Zanon Boito, Antoine Caubriere, Fethi Bougares, Mickael Rouvier, Laurent Besacier, and Esteve Yannick. 2019. ON-TRAC consortium end-to-end speech translation systems for the IWSLT 2019 shared task. In *Proceedings of 16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

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

Thai-Son Nguyen, Sebastian Stuker, Jan Niehues, and Alex Waibel. 2020b. Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

J. Niehues, R. Cattoni, S. Stüker, M. Negri, M. Turchi, T. Ha, E. Salesky, R. Sanabria, L. Barrault, L. Specia, and M. Federico. 2019. The IWSLT 2019 Evaluation Campaign. In *Proceedings of the 16th International Workshop on Spoken Language Translation (IWSLT 2019)*, Hong Kong, China.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The IWSLT 2018 Evaluation Campaign. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, pages 2–6, Bruges, Belgium.

Brian Ore, Eric Hansen, Timothy Anderson, and Jeremy Gwinnup. 2020. The AFRL IWSLT 2020 Systems: Work-From-Home Edition. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-jing Zhu. 2002c. BLEU: A Method for Automatic Evaluation of Machine Translation. *ACL (Association for Computational Linguistics)*.

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. *Interspeech 2019*.

Michael Paul. 2006. Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–15, Kyoto, Japan.

Michael Paul. 2008. Overview of the IWSLT 2008 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–17, Waikiki, Hawaii.

Michael Paul. 2009. Overview of the IWSLT 2009 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–18, Tokyo, Japan.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the IWSLT 2010 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 3–27, Paris, France.

Ngoc-Quan Pham, Thai-Son Nguyen, Thanh-Le Ha, Juan Hussain, Felix Schneider, Jan Niehues, Sebastian Stuker, and Alexander Waibel. 2019a. The IWSLT 2019 KIT Speech Translation System. In *Proceedings of 16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, Sebastian Stüker, and Alexander Waibel. 2019b. Very deep self-attention networks for end-to-end speech recognition.

Ngoc-Quan Pham, Felix Schneider, Tuan-Nam Nguyen, Thanh-Le Ha, Thai-Son Nguyen, Maximilian Awiszus, Sebastian Stüker, and Alexander Waibel. 2020. KIT's IWSLT 2020 SLT Translation System. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Peter Polák, Sangeet Sagar, Dominik Macháček, and Ondřej Bojar. 2020. Neural ASR with Phoneme-Level Intermediate Step — A Non-Native Speech Translation Task Submission to IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

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

Tomasz Potapczyk and Pawel Przybysz. 2020. SR-POL's System for the IWSLT 2020 End-to-End Speech Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Tomasz Potapczyk, Pawel Przybysz, Marcin Chochowski, and Artur; Szumaczuk. 2019. Samsung's System for the IWSLT 2019 End-to-End Speech Translation Task. In *Proceedings of 16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2014. Enhancing the ted-lium corpus with selected data for language modeling and more ted talks. In *LREC*.

Nikhil Saini, Jyotsana Khatri, Preethi Jyothi, and Pushpak Bhattacharyya. 2020. Generating Fluent Translations from Disfluent Text Without Access To Fluent References. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Elizabeth Salesky, Susanne Burger, Jan Niehues, and Alex Waibel. 2018. Towards fluent translations from disfluent speech. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 921–926. IEEE.

Elizabeth Salesky, Matthias Sperber, and Alex Waibel. 2019. Fluent translations from disfluent speech in end-to-end speech translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2786–2792.

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 *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *INTERSPEECH*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*.

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006a. 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)*, pages 223–231, Cambridge, USA.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006b. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*.

Matthias Sperber, Ngoc Quan Pham, Thai Son Nguyen, Jan Niehues, Markus Muller, Thanh-Le Ha, Sebastian Stuker, and Alex Waibel. 2018. KIT's IWSLT 2018 SLT Translation System. In *15th International Workshop on Spoken Language Translation (IWSLT 2018)*, Bruges, Belgium.

Milos Stanojevic and Khalil Sima'an. 2014. BEER: BEtter evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Enmin Su and Yi Ren. 2020. Deep Blue Sonics' Submission to IWSLT 2020 Open Domain Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Yuhui Sun, Mengxue Guo, Xiang Li, Jianwei Cui, and Bin Wang. 2020. Xiaomi's Submissions for IWSLT 2020 Open Domain Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

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

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. *LREC (International Conference on Language Resources and Evaluation)*.

Mei Tu, Wei Liu, Lijie Wang, Xiao Chen, and Xue Wen. 2019. End-to-end speech translation system description of LIT for IWSLT 2019. In *Proceedings of 16th International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel Corpora for Medium Density Languages. *RANLP (Recent Advances in Natural Language Processing)*.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation. *arXiv preprint arXiv:1803.07416*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017a. Attention is All You Need. In *Proceedings of NIPS 2017*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017b. Attention Is All You Need. *NeurIPS (Neural Information Processing Systems)*.

Raúl Vázquez, Mikko Aulamo, Umut Sulubacak, and Jörg Tiedemann. 2020. The University of Helsinki submission to the IWSLT2020 Offline Speech Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Raúl Vázquez, Alessandro Raganato, Mathias Creutz, and Jorg Tiedemann. 2020. A systematic study of inner-attention-based sentence representations in multilingual neural machine translation. *Computational Linguistics*, 0((ja)):1–53.

Pavel Vondricka. 2014. Aligning Parallel Texts with InterText. *LREC (International Conference on Language Resources and Evaluation)*.

Hari Krishna Vydana, Martin Karafi'at, Katerina Zmolikova, Luk'as Burget, and Honza Cernocky. 2020. Jointly trained transformers models for spoken language translation.

Minghan Wang, Hao Yang, Yao Deng, Ying Qin, Lizhi Lei, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Ning Xie, and Xiaochun Li. 2020a. The HW-TSC Video Speech Translation system at IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Qian Wang, Yuchen Liu, Cong Ma, Yu Lu, Yining Wang, Long Zhou, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020b. CASIA's Submission for IWSLT 2020 Open Domain Translation. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. 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.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018-September:2207–2211.

Jiaze Wei, Wenbin Liu, Zhenfeng Wu, You Pan, and Yanqing He. 2020. ISTIC's Neural Machine Translation System for IWSLT'2020. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Qian Zhang, Tingxun Shi, Xiaopu Li, Dawei Dang, Di Ai, Zhengshan Xue, and Jie Hao. 2020. OPPO's Machine Translation System for the IWSLT 2020 Open Domain Translation Task. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

Yimeng Zhuang, Yuan Zhang, and Lijie Wang. 2020. LIT Team's System Description for Japanese-Chinese Machine Translation Task in IWSLT 2020. In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.

‘

Appendix A. Evaluation Results and Details

A.1. Simultaneous Speech Translation

· Summary of the results of the simultaneous speech translation **text track.**
· Results are reported on the blind test set and systems are grouped by latency regime.
· Tabulated raw data will also be provided on the task web site[31] and the repository[32].

Team	BLEU	AP	AL	DAL
Low Latency				
ON-TRAC	23.59	0.77	2.71	3.92
Medium Latency				
ON-TRAC	29.38	0.65	5.95	6.94
KIT	29.31	0.63	5.93	6.84
APPTEK/RWTH	28.69	0.65	4.61	7.26
SRSK	27.10	0.91	5.44	6.44
High Latency				
ON-TRAC	30.51	0.63	8.71	9.63
KIT	29.76	0.63	6.38	7.32
APPTEK/RWTH	28.69	0.65	4.61	7.26
SRSK	28.49	0.91	6.09	7.13
Unconstrained				
ON-TRAC	30.51	0.63	8.71	9.63
KIT	29.76	0.63	6.38	7.32
SRSK	29.41	0.90	15.28	15.68
APPTEK/RWTH	28.69	0.65	4.61	7.26

· Summary of the results of the simultaneous speech translation **speech track.**
· Results are reported on the blind test set and systems are grouped by latency regime.
· Tabulated raw data will also be provided on the task web site[33] and the repository[34].

Team	BLEU	AP	AL	DAL
Low Latency				
SRSK	9.25	1.17	738.75	1102.96
ON-TRAC	7.27	0.97	955.11	1833.27
Medium Latency				
ON-TRAC	15.31	0.86	1727.49	3280.03
SRSK	13.58	1.07	1815.93	2243.25
High Latency				
ON-TRAC	21.80	0.74	3932.21	5029.31
SRSK	15.70	1.07	3602.02	4677.22
Unconstrained				
ON-TRAC	21.80	0.74	3932.21	5029.31
APPTEK/RWTH	21.19	0.74	4123.67	4750.24
SRSK	16.99	1.03	4054.18	4799.37

A.2. Video Speech Translation

· Systems are ordered according to the *CER* metrics.
· *BLEU* and *METEOR* scores are given as percent figures (%).

Video Translation: Chinese-English

System	CER	BLEU	METEOR	chrF
HW_TSC	36.54	14.96	30.4	35.2
Online A	37.65	11.97	26.3	32.4
Online B	47.89	13.19	26.0	30.4

Chinese-English: Average sentence-level BLEU score within CER ranges

CER	HW_TSC	Online A	Online B
< 15.0	14.55	11.61	20.75
(15.0, 20.0]	15.72	14.78	17.17
(20.0, 25.0]	13.94	15.18	21.21
(25.0, 30.0]	13.10	7.84	16.38
(30.0, 35.0]	9.58	5.54	15.48
(35.0, 40.0]	5.85	5.77	15.82
> 40.0	7.65	3.32	4.71

A.3. Offline Speech Translation

· Systems are ordered according to the *BLEU* metrics.
· *BLEU* and *TER* scores are given as percent figures (%).
· End-to-end systems are indicated by gray background.

Speech Translation : TED English-German tst 2020 (own segmentation)

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
SRPOL	25.3	59.45	53.16	49.35	26.4	57.60
APPTEK/RWTH	25.06	61.43	53.51	48.24	26.29	59.20
AFRL	23.33	62.12	52.46	50.05	24.53	59.96
APPTEK/RWTH	23.29	64.77	52.31	49.12	24.67	62.42
KIT	22.56	65.56	50.04	53.15	23.71	63.42
ON-TRAC	22.12	63.87	51.20	51.46	23.25	61.85
KIT	21.81	66.50	50.99	51.30	24.21	63.06

Speech Translation : TED English-German tst 2020 (given segmentation)

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
APPTEK/RWTH	22.49	65.20	51.40	52.75	23.73	62.93
KIT	22.06	65.38	51.22	51.26	23.24	63.10
SRPOL	21.49	65.74	49.81	56.20	22.7	63.82
FBK	20.75	68.11	49.87	55.31	21.88	66.04
APPTEK/RWTH	20.5	70.08	49.65	54.85	21.84	67.95
KIT	19.82	70.51	48.62	56.91	22	67.36
BHANSS	18.09	71.78	47.09	60.96	19	70.06
HY	17.02	76.37	47.03	58.32	18.07	74.23
DIDI LABS	10.14	101.56	41.95	62.60	10.83	99.60
HY	6.77	86.31	36.81	76.30	7.26	84.91

Speech Translation : TED English-German tst 2019 (own segmentation)

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
SRPOL	23.96	60.79	51.45	51.16	24.94	59.12
APPTEK/RWTH	23.4	63.53	52.13	49.23	24.6	61.27
APPTEK/RWTH	21.58	66.15	50.87	50.54	22.85	63.79
AFRL	21.28	64.96	51.11	51.88	22.5	62.66
KIT	21.07	66.59	49.88	52.74	22.33	64.32
KIT	20.43	66.29	50.99	50.26	22.99	62.46
ON-TRAC	20.19	66.38	49.89	52.51	21.23	64.26

Speech Translation : TED English-German tst 2019 (given segmentation)

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
SRPOL	20.1	67.73	47.76	59.08	21.17	65.92
FBK	19.52	68.93	48.07	58.26	20.65	66.87
APPTEK/RWTH	19.23	71.22	47.94	57.96	20.53	68.97
KIT	18.83	70.08	47.83	57.88	21.2	66.66
BHANSS	17.85	70.32	46.63	61.01	18.85	68.55
HY	16.44	76.26	46.06	60.42	17.46	74.17
DIDI LABS	10.22	97.01	42.13	62.77	10.95	94.93
HY	7.64	83.85	37.48	75.74	8.25	82.47

A.4. Conversational Speech Translation

· MT systems are ordered according to the *BLEU* metric.
· BLEU scores utilize 2 **fluent** English references to assess fluent translation.
· METEOR scores utilize 4 **disfluent** English references to test meaning preservation from the original disfluent data.

* = *submitted with an off-by-one error on L2077; corrected by the organizers*

Text Translation : `test, gold transcript`

System	Constrained?	No Fluent Data?	BLEU	METEOR
NAIST-b			**25.6**	28.5
NAIST-c			25.4	28.1
NAIST-a	✓		**20.8**	25.7
NAIST-f		✓	**23.6**	33.8
NAIST-e		✓	23.1	34.1
IITB		✓	21.0	33.0
NAIST-d	✓	✓	**18.5**	30.8

Text Translation : `test, ASR output`

System	Constrained?	No Fluent Data?	BLEU	METEOR
NAIST-b			**23.9**	23.5
NAIST-c			22.0	22.0
NAIST-a	✓		**17.0**	21.6
IITB		✓	**28.1***	39.1
NAIST-e		✓	24.7	31.3
NAIST-f		✓	24.7	30.9
NAIST-d	✓	✓	**13.7**	22.3

A.5. Open Domain Translation

Shared translation task overall results for all participants, evaluated with 4-gram character BLEU.
* = *collected external parallel training data that inadvertently overlapped with the blind test set.*

JA → ZH		ZH → JA	
Baseline	22.0	Baseline	26.3
CASIA	55.8*	CASIA	43.0*
SRC-B	34.0	XIAOMI	34.3
OPPO	32.9	TSUKUBA	33.0
XIAOMI	32.5	OCTANOVE	31.7
TSUKUBA	32.3	DBS	31.2
UEDIN	30.9	OPPO	30.1
KSAI	29.4	UEDIN	29.9
ISTIC	28.2	SRC-B	28.4
DBS	26.9	ISTIC	27.7
OCTANOVE	26.2	NICT	26.3
KINGSOFT	25.3	KSAI	25.9
NICT	22.6	HW-TSC	7.1
HW-TSC	11.6		
TAMKANG	1.8		
SJTU	0.1		

Pipeline for crawling parallel Chinese-Japanese data

The pipeline's stages, diagrammed in Figure 3, are:

1. Deep-crawl the target URL list. We skipped this step in the first run, and instead started with 5 billion entries from CommonCrawl.[35]

2. Identify potentially-parallel Chinese-Japanese webpage pairs using URL structure. For example, `https://www.gotokyo.org/jp/` and `https://www.gotokyo.org/cn/` only differ by the country codes `jp` and `cn`.

3. Download the potentially parallel page pairs.

4. Strip HTML and markup metadata with the `BeautifulSoup` Python module. Split each page into sentence segments.

5. Align segments to be parallel, using Hunalign (Varga et al., 2005).

6. Filter pairs by language ID and length ratio.

The first pipeline run produced 227k URL pairs (1.4m segment pairs) of parallel data containing 28.7m characters on the Chinese side. We used the 227k URL pairs to trace which domains yielded the most parallel data. We then re-ran the pipeline on each of the 6000 most-promising domains, but now deep-crawling the domain using `scrapy` in Step 1 to produce the URL list examined in Step 2.

We concatenated the parallel output from all the runs, keeping track of the provenance URL of each segment. Finally, we applied a filter to remove objectionable content. The result was `webcrawled_parallel_filtered` dataset, containing nearly 19m hopefully-parallel segment pairs (494m Zh chars) with provenance information.

[35] https://commoncrawl.org/

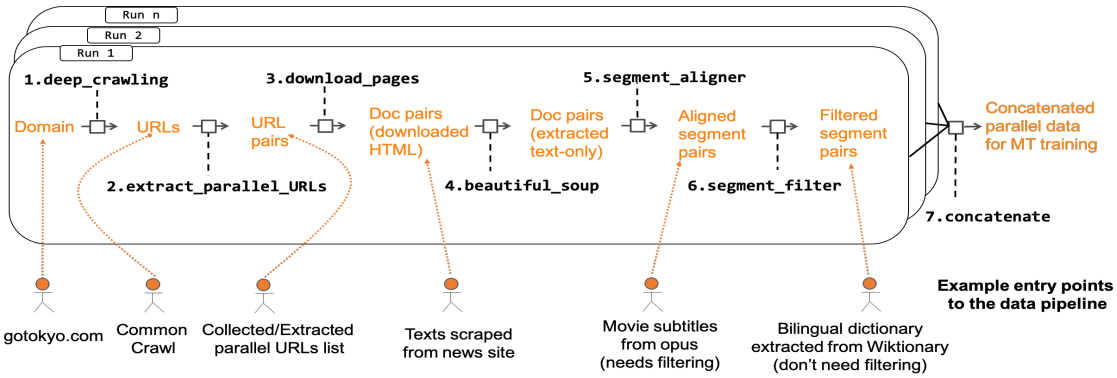

Figure 3: Pipeline to harvest parallel zh-jp text. The modules are numbered in black, with inputs/outputs in orange. The examples at the bottom show how the pipeline can be entered at intermediate stages.

Test Set Provenance

The held-out test set was intended to cover a variety of topics not known to the participants in advance. We selected test data from high-quality (human translated) parallel web content, authored between January and March 2020. Because of this timeframe, COVID19 is a frequent topic in the test set. We collected bilingual material from 104 webpages, detailed in the Appendix. Table 5.

Pages	Source
54	jp.hjenglish.com : Chinese website with Japanese learning material.
38	j.people.com.cn : the Japanese version of the People's Daily newspaper.
4	china-embassy.or.jp : the Embassy of China in Japan
4	people.com.cn : the People's Daily newspaper, in Chinese.
3	emb-japan.go.jp : the Embassy of Japan in China
1	kantei.go.jp : the Prime Minister of Japan's office

Table 5: Provenance of the Chinese-Japanese test set.

To build the test set, we first identified articles on these sites with translations, and copied their contents into separate files. All segments were then manually aligned by a native Chinese speaker with basic knowledge of Japanese, using the InterText tool (Vondricka, 2014). Lastly, a bilingual speaker filtered the aligned pairs, excluding pairs that were not parallel. This produced 1750 parallel segments, which we divided randomly in half: 875 lines for the Chinese-to-Japanese translation test set, and 875 lines for the other direction. The Japanese segments have an average length of 47 characters, and the Chinese ones have an average length of 35.

A.6. Non-Native Speech Translation

English→German

· Complete result for English-German SLT systems followed by public systems PUBLIC-A and PUBLIC-B for comparison.
· Primary submissions are indicated by gray background. Best results in bold.

	SLT					ASR	
	Quality		Simultaneity			Quality	
System	$BLEU_1$	$BLEU_{mw}$	Flicker	$Delay_{ts}$[Match%]	$Delay_{mw}$[Match%]	WER_1	WER_{mw}
APPTEK/RWTH1	14.70	13.28	-	-	-	**14.27**	**16.26**
APPTEK/RWTH2	**16.14**	**15.00**	-	-	-	**14.27**	**16.26**
APPTEK/RWTH3	15.92	14.50	-	-	-	**14.27**	**16.26**
BUT1	2.25	0.63	-	-	-	32.33	34.09
BUT2	2.25	0.67	-	-	-	32.91	34.46
BUT3	1.93	0.59	-	-	-	32.91	34.46
BUT4	2.29	0.72		-	-	32.91	34.46
CUNI-NN11	6.37	5.86	-	-	-	28.68	32.10
CUNI-NN12	14.08	12.38	-	-	-	17.39	20.46
CUNI-NN13	14.32	12.73	-	-	-	17.02	19.98
CUNI-NN14	6.65	6.20	-	-	-	28.75	32.23
CUNI-NN15	12.51	10.88	-	-	-	16.54	18.19
CUNI-NN16	13.15	11.50	-	-	-	16.33	17.95
ELITR31	9.72	7.22	6.71	**1.901 [50.91%]**	1.926 [30.01%]	23.77	25.15
ELITR32	9.18	7.32	7.48	1.926 [30.01%]	1.944 [30.42%]	22.91	24.26
ELITR33	9.18	7.32	7.48	1.972 [52.61%]	1.945 [30.43%]	22.91	24.26
ELITR34	9.18	7.32	7.43	1.951 [52.53%]	1.923 [30.41%]	22.91	24.26
ELITR35	9.18	7.32	6.48	2.038 [52.84%]	2.024 [30.76%]	22.91	24.26
ELITR36	9.18	7.32	5.97	2.034 [52.66%]	2.029 [30.79%]	22.91	24.26
ELITR37	9.39	7.05	6.33	2.471 [34.14%]	**1.828 [31.81%]**	23.81	25.25
ELITR38	9.40	7.06	6.35	2.461 [34.24%]	1.846 [31.85%]	23.81	25.25
ELITR39	9.40	7.06	6.33	2.380 [33.37%]	**1.810 [31.63%]**	23.81	25.25
ELITR40	9.39	7.05	5.66	2.544 [34.28%]	1.964 [32.28%]	23.81	25.25
ELITR41	9.39	7.06	**5.30**	2.391 [34.09%]	1.957 [32.28%]	23.81	25.25
ELITR-OFFLINE21	14.83	12.67	-	-	-	15.29	17.67
ELITR-OFFLINE22	13.31	11.35	-	-	-	15.29	17.67
ELITR-OFFLINE23	14.08	12.33	-	-	-	15.29	17.67
ELITR-OFFLINE24	13.03	10.76	-	-	-	15.29	17.67
ELITR-OFFLINE25	12.88	10.83	-	-	-	15.29	17.67
ELITR-OFFLINE26	10.45	8.32	-	-	-	15.29	17.67
ELITR-OFFLINE27	11.58	9.87	-	-	-	16.33	17.95
ELITR-OFFLINE28	11.76	9.83	-	-	-	16.33	17.95
ELITR-OFFLINE29	12.51	10.88	-	-	-	16.33	17.95
ELITR-OFFLINE30	11.34	9.42	-	-	-	16.33	17.95
ELITR-OFFLINE31	12.51	10.53	-	-	-	16.33	17.95
ELITR-OFFLINE32	7.89	5.72	-	-	-	16.33	17.95
CUNI-KALDI01	-	-	-	-	-	22.88	24.53
CUNI-KALDI02	-	-	-	-	-	30.42	31.17
CUNI-KALDI03	-	-	-	-	-	21.25	23.40
PUBLIC-A	4.29	3.02	-	-	-	30.10	31.09
PUBLIC-B	13.75	12.35	-	-	-	21.54	23.59

English→Czech

· Complete result for English-Czech SLT systems followed by public systems PUBLIC-A and PUBLIC-B for comparison.
· Primary submissions are indicated by gray background. Best results in bold.

	SLT					ASR	
	Quality		Simultaneity			Quality	
System	$BLEU_1$	$BLEU_{mw}$	Flicker	$Delay_{ts}$[Match%]	$Delay_{mw}$[Match%]	WER_1	WER_{mw}
CUNI-NN01	10.57	10.34	-	-	-	28.68	32.10
CUNI-NN02	10.89	11.50	-	-	-	17.39	20.46
CUNI-NN03	12.74	11.38	-	-	-	17.02	19.98
CUNI-NN04	10.24	10.21	-	-	-	28.75	32.23
CUNI-NN05	11.85	10.57	-	-	-	16.54	18.19
CUNI-NN06	12.27	11.00	-	-	-	16.33	17.95
ELITR01	7.87	6.22	7.00	1.530 [42.45%]	1.575 [23.93%]	23.77	25.15
ELITR02	7.56	5.95	6.46	1.696 [22.01%]	1.561 [25.25%]	23.81	25.25
ELITR03	7.56	5.95	6.38	1.744 [22.26%]	1.618 [25.34%]	23.81	25.25
ELITR04	7.54	5.93	6.38	1.725 [22.09%]	1.603 [25.32%]	23.81	25.25
ELITR05	8.93	7.67	7.51	1.605 [44.80%]	1.623 [92.49%]	23.81	25.25
ELITR06	8.79	7.54	7.00	**1.198 [52.55%]**	**1.082 [32.18%]**	23.81	25.25
ELITR07	8.93	7.67	6.97	1.596 [44.79%]	1.630 [24.86%]	23.81	25.25
ELITR08	8.93	7.67	6.54	1.586 [44.64%]	1.629 [24.91%]	23.81	25.25
ELITR09	8.93	7.65	7.38	1.520 [42.80%]	1.503 [23.23%]	23.81	25.25
ELITR10	8.93	7.67	7.41	1.630 [44.77%]	1.667 [24.96%]	23.81	25.25
ELITR11	6.50	4.94	6.00	1.677 [20.99%]	1.595 [24.58%]	23.81	25.25
ELITR12	6.50	4.94	6.26	1.610 [20.87%]	1.504 [24.35%]	23.81	25.25
ELITR13	6.50	4.94	6.26	1.495 [19.47%]	1.399 [23.30%]	23.81	25.25
ELITR14	6.52	4.95	5.69	1.650 [20.88%]	1.597 [24.63%]	23.81	25.25
ELITR15	6.50	4.94	**5.18**	1.541 [20.71%]	1.594 [24.59%]	23.81	25.25
ELITR16	7.40	5.74	6.64	1.633 [21.89%]	1.468 [24.43%]	23.81	25.25
ELITR17	8.45	7.32	6.56	1.597 [44.85%]	1.728 [25.35%]	22.91	24.26
ELITR18	8.36	7.17	6.00	1.514 [45.58%]	1.629 [26.54%]	22.91	24.26
ELITR19	8.56	7.45	5.31	1.600 [46.81%]	1.713 [27.94%]	22.91	24.26
ELITR20	8.55	7.41	6.31	1.557 [45.78%]	1.704 [26.51%]	22.91	24.26
ELITR-OFFLINE01	13.33	11.75	-	-	-	**15.29**	**17.67**
ELITR-OFFLINE02	13.44	11.64	-	-	-	**15.29**	**17.67**
ELITR-OFFLINE03	13.56	11.79	-	-	-	**15.29**	**17.67**
ELITR-OFFLINE04	**14.08**	**12.57**	-	-	-	**15.29**	**17.67**
ELITR-OFFLINE05	10.07	8.23	-	-	-	**15.29**	**17.67**
ELITR-OFFLINE06	8.42	6.99	-	-	-	**15.29**	**17.67**
ELITR-OFFLINE07	9.62	8.16	-	-	-	**15.29**	**17.67**
ELITR-OFFLINE08	11.88	10.26	-	-	-	16.33	17.95
ELITR-OFFLINE09	11.52	9.83	-	-	-	16.33	17.95
ELITR-OFFLINE10	11.43	9.99	-	-	-	16.33	17.95
ELITR-OFFLINE11	11.85	10.57	-	-	-	16.33	17.95
ELITR-OFFLINE12	9.29	7.76	-	-	-	16.33	17.95
ELITR-OFFLINE13	7.76	6.35	-	-	-	16.33	17.95
ELITR-OFFLINE14	7.37	6.54	-	-	-	16.33	17.95
PUBLIC-A	3.30	2.47	-	-	-	30.10	31.09
PUBLIC-B	10.79	9.85	-	-	-	21.54	23.59

Test Set Provenance

Only a limited amount of resources could have been invested in the preparations of the test set and the test set thus build upon some existing datasets. The components of the test sets are:

Antrecorp[36] (Macháček et al., 2019), a test set of up to 90-second mock business presentations given by high school students in very noisy conditions. None of the speakers is a native speaker of English (see the paper for the composition of nationalities) and their English contains many lexical, grammatical and pronunciation errors as well as disfluencies due to the spontaneous nature of the speech.

For the purposes of this task, we equipped Antrecorp with manual translations into Czech and German. No MT system was used to pre-translate the text to avoid bias in automatic evaluation.

[36] http://hdl.handle.net/11234/1-3023

Because the presentations are very informal and their translation can vary considerably, we created two independent translations into Czech. In the end, only the first one of them was used as the reference, to keep BLEU scores across test set parts somewhat comparable.

Khan Academy[37] is a large collection of educational videos. The speaker is not a native speaker of English but his accent is generally rather good. The difficulty in this part of the test lies in the domain and also the generally missing natural segmentation into sentences.

SAO is a test set created by ELITR particularly for this shared task, to satisfy the need of the Supreme Audit Office of the Czech Republic. The test sets consists of 6 presentations given in English by officers of several supreme audit institutions (SAI) in Europe and by the Europan Court of Auditors. The speakers nationality (Austrian, Belgian, Dutch, Polish, Romanian and Spanish) affects their accent. The Dutch file is a recording of a remote conference call with further distorted sound quality.

The development set contained 2 other files from Antrecorp, one other file from the SAO domain and it also included 4 files from the AMI corpus (Mccowan et al., 2005) to illustrate non-native accents. We did not include data from AMI corpus in the test set because we found out that some participants trained their (non-constrained) submissions on it.

For SAO and Antrecorp, our test set was created in the most straightforward way: starting with the original sound, manual transcription was obtained (with the help of ASR) as a line-oriented plaintext. The transcribers were instructed to preserve all words uttered[38] and break the sequence of words into sentences in as natural way as possible. Correct punctuation and casing was introduced at this stage, too. Finally, the documents were translated in Czech and German, preserving the segmentation into "sentences".

For the evaluation of SLT simultaneity, we force-aligned words from the transcript to the sound using a model trained with Jasper (Li et al., 2019) and resorted to fully manual identification of word boundaries in the few files where forced alignment failed.

Despite a careful curation of the dataset, we are aware of the following limitations. None of them are too frequent or too serious but they still deserve to be mentioned:

- Khan Academy subtitles never had proper segmentation into sentences and manual correction of punctuation and casing. The subtitles were supposedly manually refined but the focus was on their presentation in the running video lecture, not on style and typesetting.
- Khan Academy contains many numbers (written mostly as numbers). For small numbers, both digits and words are often equally suitable but automatic metrics treat this difference as a mistranslation and no straightforward reliable normalization is possible either, so we did not apply any.
- Minor translation errors into German were seen in Khan Academy videos and in the "Belgian" SAO file.

[37] http://www.khanacademy.org/

[38] This decision is possibly less common in the ASR community but it is motivated by the subsequent translation step which has the capacity to recover from disfluences as needed.