The IWSLT 2018 Evaluation Campaign

J. Niehues[(1)] R. Cattoni[(2)] S. Stüker[(1)] M. Cettolo[(2)] M. Turchi[(2)] M. Federico[(3)†]

[(1)] KIT - Adenauerring 2, 76131 Karlsruhe, Germany
[(2)] FBK - Via Sommarive 18, 38123 Trento, Italy
[(3)] Amazon AI - East Palo Alto, CA 94303, USA

Abstract

The International Workshop of Spoken Language Translation (IWSLT) 2018 Evaluation Campaign featured two tasks: the low-resourced machine translation task and the speech translation task. In the first task, manual transcribed speech needs to be translated from Basque to English. Since this translation direction is a under-resourced language pair, participants were encouraged to used additional parallel data from related languages. In the second task, the participants need to translate English audio into German text by building a full speech-translation system. In the baseline condition, participants were free to used any architecture, while they are restricted to use a single model for the end-to-end task.

This year, eight research groups took part in the Basque English translation task, and nine in the speech translation task

1. Introduction

We report here on the outcomes of the 2018 evaluation campaign organized by the International Workshop of Spoken Language Translation (IWSLT). The IWSLT workshop was started in 2004 [1] with the purpose of enabling the exchange of knowledge among researchers working on speech translation and creating an opportunity to develop and compare translation systems on a common test bed. The evaluation campaign built on one of the outcomes of the C-STAR (Consortium for Speech Translation Advanced Research) project, namely the BTEC (Basic Travel Expression Corpus) multilingual spoken language corpus [2], which initially served as a primary source of evaluation. Since its beginning, translation tasks of increasing difficulty were offered and new data sets covering a large number of language pairs were shared with the research community. In the fifteenth editions organized from 2004 to 2018, the campaign attracted around 70 different participating teams from all over the world.

Automatic spoken language translation is particularly challenging for a number of reasons. On one side, machine translation (MT) systems are required to deal with the specific features of spoken language. With respect to written language, speech is structurally less complex, formal and fluent. It is also characterized by shorter sentences with a lower amount of rephrasing but a higher pronoun density [3]. On the other side, speech translation [4] requires the integration of MT with automatic speech recognition, which brings with it the additional difficulty of translating content that may have been corrupted by speech recognition errors.

Along the years, three main evaluation tracks were progressively introduced, addressing all the core technologies involved in the spoken language translation task, namely:

- Automatic speech recognition (ASR), *i.e.* the conversion of a speech signal into a transcript
- Machine translation (MT), *i.e.* the translation of a polished transcript into another language
- Spoken language translation (SLT), *i.e.* the conversion and translation of a speech signal into a transcript in another language

In previous years, the ASR transcript was provided to the participants of the SLT task. Therefore, the SLT task main focus on investigating translation methods for automatic transcripts.

The recent development in deep learning lead to the usage of similar techniques in machine translation and automatic speech recognition. Furthermore, the success of sequence-to-sequence model allowed the development of end-to-end speech translation systems [5]. Therefore, in this years edition, we drooped the ASR task and included the transcription of the audio into the SLT task. So, for the first time, the participants need to develop the full speech translation pipeline.

The 2018 IWSLT evaluation focused on translating talks from two sources of data: translation of TED talks corpus [6] and, for the speech translation task, university lectures collected at KIT [7].

The TED translation task of IWSLT has become a seasoned task by now. Its introduction was motivated by its higher complexity with respect to the previous travel tasks, and by the availability of high quality data. In order to keep the tasks interesting and to follow current trends in research and industry, we expanded and developed the IWSLT tasks further. Motivated by last years success of the multi-lingual machine translation task, we created task on a low resources

† Work performed while the author was at FBK, Italy.

language pair. Furthermore, we developed the speech translation task further. Participants need to build a complete speech translation system and we encourage research on end-to-end models. Unlike in previous years, we also limited the scope of the evaluation to few languages. The main reason for this was to avoid dispersion of participants in too many tasks.

The translation directions considered this year for the SLT track were English to German. For the MT tracks, the participants need to translate from Basque to English.

For all tracks and tasks, permissible training data sets were specified and instructions for the submissions of test runs were given together with the detailed evaluation schedule.

All runs submitted by participants were evaluated with automatic metrics. In particular, for the MT tracks, an evaluation server was set up so that participants could autonomously score their runs on different dev and test sets. This year, 16 groups participated in the evaluation (see Table 1). In following, we provide a description of the tasks introduced this year followed by a detailed report of each track we organised which include a summary of the main results. The paper ends with an appendix reporting all the detailed results of this year's evaluation.

2. Low Resource Machine Translation

2.1. Definition

The Low Resource Translation Task addresses a conventional bilingual text translation task in the domain of the TED talks. Participants were required to translate TED talks from Basque to English. Given the difficulty of the proposed translation direction and the scarcity of available parallel data, additional parallel data from related languages were prepared.

Concerning Basque-English data, training set included 64 TED talks with 5.6K parallel sentences (81K Basque and 109K English tokens). Development set contained 10 talks with 1.1K parallel sentences (17K Basque and 23K English tokens). The valuation set *tst2018* consisted of 10 talks with 1.1K parallel sentences (15K Basque and 20K English tokens).

In-domain parallel training data included also talks from related languages: 73 talks for Basque-French, 74 for Basque-Spanish, 2595 for French-English, 2589 for Spanish-French and 2650 for Spanish-English. Moreover, an additional archive with the original xml files of all the TED talks available at April 2018 – excluding those in the tst2018 evaluation set – was provided. Finally, participants could download any data of the original TED talks from the TED website – excluding those in the tst2018 evaluation set.

Out-of-domain training data were restricted to parallel and monolingual corpora (including Basque data) provided by the OPUS[1] and WMT[2] organizations on their respective websites. Moreover, participants were allowed to utilize Basque-Spanish parallel and monolingual data from the Open Data Euskadi Repository kindly provided by the Vicomtech[3] research center.

In-domain training and development data were supplied through the website of the WIT3 ([6]), while out-of-domain training data were made available through the workshop's website.

2.2. Evaluation

Automatic translation of the test2018 tst2018 evaluation set were required to be in NIST XML format with case-sensitive, detokenized and punctuated texts. Translations quality was measured automatically by means of the three automatic standard metrics BLEU, NIST, and TER. Case sensitive scores were calculated with the software tools mteval-v13a.pl3 and tercom-0.7.254, by invocating:

- mteval-v13a.pl -c
- java -Dfile.encoding=UTF8 -jar tercom.7.25.jar -N -s

It is worth noticing here that the two scoring scripts apply their own internal tokenization.

In order to allow participants to evaluate their progresses automatically and under identical conditions, an evaluation server was developed. Participants could submit the translation of the development set to either a REST Webservice or through a GUI on the web, receiving as output BLEU, NIST and TER scores computed as described above. The core of the evaluation server is a shell script wrapping the mteval and tercom scorers. The REST service is implemented with a PHP script running over Apache HTTP Server, while the GUI on the web is written in HTML with AJAX code. The evaluation server was utilized also by the organizers for the automatic evaluation of the official submissions. After the evaluation period, the evaluation on the test2018 set was enabled to all participants as well.

2.3. Submissions

We received 15 submissions from 8 different participants (4 participants sent primary submissions only).

2.4. Results

The results on the tst2018 evaluation set for each participant are shown in Appendix A.1, sorted by the BLEU metric.

3. Speech Translation

3.1. Definition

In contrast to previous years, this year the participants needed to build the whole speech translation systems. The organizers did not provide any intermediate results as done in previous

[1] http://opus.nlpl.eu/

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

[3] http://www.vicomtech.org

Table 1: List of Participants

ALIBABA	Machine Intelligence Technology Lab, Alibaba Group
APPTEC	Applications Technology (AppTek), Aachen, Germany
AFRL	Air Force Research Laboratory, United States of America
ADAPT	ADAPT Centre, Ireland
CUNI	Charles University - Institute of Formal and Applied Linguistics, Czechia
FBK	Fondazione Bruno Kessler, Italy
HY	University of Helsinky, Finland
JHU	Johns Hopkins University, Baltimore, USA
KIT	Karlsruhe Institute of Technology, Germany
MEMAD	Department of Digital Humanities / HELDIG University of Helsinki, Finland
	Department of Signal Processing and Acoustics Aalto University, Finland
PROMPSIT	Prompsit Language Engineering, Spain
SGNLP	NLP Laboratory in Sogang University, South Korea
SRPOL-UEDIN	Samsung R&D Institute Poland and University of Edinburgh, Poland/UK
TIIC	Voice Interaction Technology Center, Sogou Inc., Beijing, China
	Tiangong Institute for Intelligent Computing, Tsinghua University, Beijing, China
USTC-NEL	University of Science and Technology of China and IFLYTEK Co. LTD.

years. Instead, a baseline system was provided [8]. Participants were free to used parts of this system or purely rely on the own models.

This year edition of the speech translation task contained two different conditions. In the first condition *Baseline*, the participants could use any architecture to generate the translations in the target language. The second condition *End-to-End* concentrated on end-to-end models. In this condition, participants need to train one large model to perform the whole process from source language audio to target language text.

In both tasks the same test data is used. The test data is English audio and needs to be translated into German. The test data consisted of two related types of data. One part of the training data are TED talks. These talks are well-prepared and address a broad audience. Therefore, they contain only very few disfluencies and contain only very few special terms. The second part of the test sets contain university talks and research presentations. Since the talks are targeted to a small target audience, the test sets contain more special terms.

For training the system, different data sources were provided to the participants. For training the ASR components, the TED LIUM corpus could be used [9]. For the training of the machine translation component, the data available form the WMT evaluation[4] was allowed. In addition, the organizers provide the WIT corpus []. Furthermore, for the first time, also a corpus to train the end-to-end corpus was provided. This corpus consists of English TED talks aligned with their German transcription[5].

[4]https://www.statmt.org/wmt18/

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

3.2. Evaluation

Since the audio was not segmented by a human into sentence-like units, the generated translation were segmented into different sentences than reference transcript and translation. Therefore, in a first step of the evaluation we need to realign the sentences of the reference and the automatic translation. This was done by minimizing the WER between the automatic translation and reference as described in [10]. Two segmentation were generated, one used case information for the case-sensitive metrics and one using no case information for the case-insensitive metrics.

Using the resegmented input, we used 4 different metrics to evaluate the results. For BLEU [11] and TER[12], we calculated case-sensitive and case-insensitive scores. In addition, we calculated the BEER score [13] and the characTER [14].

3.3. Submissions

In total we received 27 submissions from 9 partners. We received 7 primary submissions in the baseline condition and 4 primary submissions in the end-to-end submission. Two participants submitted output to both conditions. The results of all primary submissions are summarized in Appendix A.1.

3.4. Results

The detailed results of the automatic evaluation in terms of BLEU, TER, BEER and characTER can be found in Appendix A.1.

4. Conclusions

We reported results of the 2018 IWSLT Evaluation Campaign which featured two tasks: the translation of TED talks

from Basque to English and the speech translation task from English to German. In the second one, the test set contains TED talks as well as university lectures and research talks. In this task, two tracks were offered: a baseline condition and the end-to-end condition. In total, 14 international research groups joined the evaluation campaign. For the first time, traditional pipeline approaches for speech translation were compared to end-to-end translation models.

5. Acknowledgements

6. References

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

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

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

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

[5] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, "Sequence-to-sequence models can directly translate foreign speech," in *Interspeech 2017*, 08 2017, pp. 2625–2629.

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

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

[8] T. Zenkel, M. Sperber, J. Niehues, M. Müller, N.-Q. Pham, S. Stüker, and A. Waibel, "Open Source Toolkit for Speech to Text Translation," *The Prague Bulletin of Mathematical Linguistics*, vol. 111, pp. 125–135, October 2018. [Online]. Available: https://ufal.mff.cuni.cz/pbml/111/art-zenkel-et-al.pdf

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

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

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002. [Online]. Available: https://www.aclweb.org/anthology/P02-1040.pdf

[12] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of association for machine translation in the Americas*, 2006. [Online]. Available: https://www.cs.umd.edu/~snover/pub/amta06/ter_amta.pdf

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

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

Appendix A. Automatic Evaluation

A.1. Official Testset (*tst2018*)

· All the sentence IDs in the IWSLT 2018 testset were used to calculate the automatic scores for each run submission.
· MT systems are ordered according to the *BLEU* metrics.
· *WER*, *BLEU* and *TER* scores are given as percent figures (%).

Low Resource MT : Basque-English

System	BLEU	NIST	TER
SRPOL-UEDIN	26.21	6.51	59.49
HY	25.01	6.45	59.48
PROMPSIT	24.02	6.24	60.81
FBK	23.99	6.34	59.43
CUNI	22.86	6.10	60.31
ADAPT	13.89	4.46	69.98
AFRL	12.25	4.03	80.63
SGNLP	10.42	3.49	103.96

Speech Translation : English-German

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)	#Words
Baseline condition							
TIIC	28.09	55.74	54.73	84.72	29.44	53.73	39611
USTC-NEL	26.47	58.03	52.69	92.24	27.86	55.98	38372
ALIBABA	22.36	63.03	51.77	69.26	24.23	60.22	39751
APPTEC	21.45	64.12	51.56	63.47	22.72	61.69	41210
KIT	19.44	67.94	50.61	58.16	20.78	65.52	42128
AFRL	17.24	69.10	49.23	64.27	18.37	66.78	41155
MEMAD	15.8	74.51	47.01	82.56	17.13	72.00	41848
End-to-End condition							
USTC-NEL	19.4	68.20	48.77	87.30	20.77	65.73	41372
FBK	10.24	78.20	40.68	129.47	11.16	76.38	36627
KIT	8.4	88.54	41.48	80.38	9.22	86.55	44155
JHU	5.45	89.59	35.46	99.89	6.09	88.20	40932

Speech Translation TED Only : English-German

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
TIIC	28.18	57.31	52.74	61.06	29.36	55.65
USTC-NEL	26.79	59.89	51.28	92.50	27.89	58.23
ALIBABA	22.77	63.66	50.62	65.54	24.57	60.96
APPTEC	21.05	66.31	49.96	60.96	22.17	64.20
KIT	18.84	69.05	48.73	57.97	20.02	66.92
AFRL	15.46	72.23	47.26	61.02	16.51	70.06
MEMAD	15.57	74.83	45.35	87.54	16.8	72.56
End-to-End condition						
USTC-NEL	18.32	70.50	46.65	88.73	19.58	68.36
FBK	9.75	77.57	38.98	150.35	10.57	75.95
KIT	7.99	86.68	39.55	86.36	8.82	84.76
JHU	4.51	85.84	32.71	112.77	4.97	84.63

Speech Translation Lecture Only : English-German

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
TIIC	27.55	54.25	57.43	117.57	29.06	51.89
USTC-NEL	25.95	56.24	54.59	91.89	27.6	53.82
ALIBABA	21.77	62.42	53.30	74.42	23.68	59.52
APPTEC	21.84	62.03	53.73	66.96	23.28	59.28
KIT	20.01	66.88	53.16	58.43	21.5	64.18
AFRL	18.94	66.12	51.92	68.77	20.13	63.65
MEMAD	16.01	74.20	49.25	75.65	17.44	71.46
End-to-End condition						
USTC-NEL	20.41	66.00	51.67	85.31	21.87	63.22
FBK	10.7	78.81	42.99	100.49	11.7	76.79
KIT	8.76	90.32	44.11	72.07	9.58	88.26
JHU	5.84	93.18	39.24	82.03	6.58	91.60