# The IWSLT 2015 Evaluation Campaign

M. Cettolo[(1)] J. Niehues[(2)] S. Stüker[(2)] L. Bentivogli[(1)] R. Cattoni[(1)] M. Federico[(1)]

[(1)] FBK - Via Sommarive 18, 38123 Trento, Italy
[(2)] KIT - Adenauerring 2, 76131 Karlsruhe, Germany

## Abstract

The IWSLT 2015 Evaluation Campaign featured three tracks: automatic speech recognition (ASR), spoken language translation (SLT), and machine translation (MT). For ASR we offered two tasks, on English and German, while for SLT and MT a number of tasks were proposed, involving English, German, French, Chinese, Czech, Thai, and Vietnamese. All tracks involved the transcription or translation of TED talks, either made available by the official TED website or by other TEDx events. A notable change with respect to previous evaluations was the use of unsegmented speech in the SLT track in order to better fit a real application scenario. Thus, from one side participants were encouraged to develop advanced methods for sentence segmentation, from the other side organisers had to cope with the automatic evaluation of SLT outputs not matching the sentence-wise arrangement of the human references. A new evaluation server was also developed to allow participants to score their MT and SLT systems on selected dev and test sets. This year 16 teams participated in the evaluation, for a total of 63 primary submissions. All runs were evaluated with objective metrics, and submissions for two of the MT translation tracks were also evaluated with human post-editing.

## 1. Introduction

We present the results of the 2015 evaluation campaign organized by the International Workshop of Spoken Language Translation. The IWSLT evaluation has been running for twelve years and has been offering a variety of speech recognition, speech translation and text translation tasks [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11].

The 2015 IWSLT evaluation focused on the automatic transcription and translation of TED and TEDx talks, i.e. public speeches covering many different topics. The evaluation included three tracks:

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

As a major difference from the previous editions, not only participants in the ASR track but also those participating in the SLT track had to cope with *unsegmented speech* instead of pre-segmented speech. Thus, both ASR and SLT systems had to face the more realistic working condition of transcribing and translating a speech signal corresponding to an entire talk rather than a sequence of isolated speech segments, as in the past editions.

This year, the ASR track was on two languages, namely English and German. The SLT track included German to English and English to Chinese, Czech, French, German, Thai, and Vietnamese; the MT track offered the same tasks as SLT but in both directions.

For all tasks, all permissible training data sets were specified and instructions for the submissions of test runs were given together with the detailed evaluation schedule. This year, parallel data made available to the participants included an updated version of the WIT$^3$ [12] corpus of TED talks, data from the WMT 2015 shared tasks, the MULTIUN corpus, and Wikipedia translations kindly made available by PJAIT[13].

The test sets used for this year's evaluation (tst2015) include new TED or TEDx talks not previously released. Furthermore, for the ASR and MT tasks offered both in 2014 and 2015, progresses were assessed by asking participants to run their systems also on the test sets of edition 2014 (tst2014), which were specifically released again to this purpose.

All runs submitted by participants were evaluated with automatic metrics. In particular, for the SLT and MT tracks, an evaluation server was set up so that participants could autonomously score their runs on different dev and test sets. For two MT tasks, English-German and Vietnamese-English, systems were also evaluated by calculating HTER values on post-edits created by professional translators.

This year, 16 groups participated in the evaluation (see Table 1) submitting a total of 63 primary runs: 18 to the ASR track (9 for tst2015 and 9 for tst2014), 5 to the SLT track, and 40 to the MT track (26 for tst2015 and 14 for tst2014).

In the following, we overview each of the offered tracks (Sections 2, 3, 4), whose detailed results are provided in Appendix A. We also report on human evaluation (Section 5 and Appendix B), and finally draw some conclusions.

Table 1: List of Participants

| | |
|---|---|
| UNETI | University Of Economic And Technical Industries, Vietnam [14] |
| IOIT | Institute of Information Technology, Vietnam [15] |
| HLT-I2R | Institute for Infocomm Research, Singapore [16] |
| JAIST | Japan Advanced Inst. of Sc. and Technology; U. of Eng. and Technology; MITI [17] |
| PJAIT | Polish-Japanese Academy of Information Technology, Poland [13] |
| NAIST | Nara Institute of Science and Technology, Japan [18] |
| TUT | Toyohashi University of Technology, Japan [19] |
| RWTH | Rheinisch-Westfälische Technische Hochschule Aachen, Germany [20] |
| MITLL-AFRL | MIT Lincoln Laboratory and Air Force Research Laboratory, USA [21] |
| UEDIN | University of Edinburgh, United Kingdom [22] |
| MLLP | Machine Learning and Language Processing Research Group, Spain [23] |
| HDU | Dept. of Computational Linguistics, Heidelberg University, Germany [24] |
| LIUM | Laboratoire d'Informatique de l'Université du Maine, France [25] |
| UMD | University of Maryland, USA [26] |
| KIT | Karlsruhe Institute of Technology, Germany [27, 28] |
| SU | Stanford University, USA [29] |

# 2. ASR Track

## 2.1. Definition

The goal of the *Automatic Speech Recognition* (ASR) track for IWSLT 2015 was to transcribe English TED and German TEDx talks. The speech in TED lectures is in general planned, well articulated, and recorded in high quality. Actually TED talks are often rehearsed rigorously for several days with experts advising on and designing the presentation. Thus, to a certain degree, they almost resemble a stage performance. The main challenges for ASR in these talks are to cope with a large variability of topics, the presence of non-native speakers, and the rather informal speaking style that is often used in order to make talks entertaining. For the TEDx talks the recording conditions are often more difficult than for the English TED talks, as recording is usually done with a lower budget with worse equipment and less trained personnel. While the TEDx talks aim to mimic the TED talks, they are not as well prepared and well rehearsed as the TED lectures, thus portraying a more difficult to recognize speaking style and more adverse recording conditions for ASR.

The result of the recognition of the talks is used for two purposes. It is used to measure the performance of ASR systems on the talks and it is used as input to the spoken language translation evaluation (SLT), see Section 3.

## 2.2. Evaluation

Participants had to submit the results of the recognition of the tst2015 set in CTM format. The word error rate was measured case-insensitive. After the end of the evaluation a preliminary scoring was performed with the first set of references. This was followed by an adjudication phase in which participants could point out errors in the reference transcripts. The adjudication results were collected and combined into the final set of references with which the official scores were calculated.

For German, the transcriptions of the talks were generated manually by trained transcribers at KIT, while the initial English transcripts were derived from the subtitles available via TED by performing a forced alignment of the subtitles to the audio file. Then, a fast manual check was performed by listening to the talk and simultaneously scanning the aligned transcripts. In this way major deviations of the subtitles from the audio were detected. The subtitles were then either manually corrected or the affected portions of the audio were excluded from scoring. The more subtle differences between the subtitles and the actual spoken words were left for detection during the adjudication phase.

In order to measure the progress of the systems over the years, participants also had to provide results on the test set from 2014, i.e. *tst2014*.

## 2.3. Submissions

For this year's evaluation we received primary submissions from seven sites. For English we received six primary runs on tst2015 and six on tst2014, while for German we received 3+3 primary submissions. For English we further received a total of five contrastive submissions from three sites.

## 2.4. Results

The detailed results of the primary submissions of the evaluation in terms of word error rate (WER) can be found in Appendix A. The word error rates of the submitted systems on tst2015 are in the range of 6.6%–13.8% for English and 17.6%–43.3% for German.

In German, the fact that TEDx talks sometimes worse recording conditions than TED talks was reflected by the fact that one talk in the German tst2015 set had WERs above 45% and another above 30%, while for all other talks WERs were in the range from 10% to 23%.

Three participants of this year's English ASR track also participated last year. All of them showed significant progress on tst2014, absolute WER improvements ranging from 1.7–5.8 percentage points. This year the lowest WER on tst2014 was 7.1% as compared to 8.4% last year.

Only one participant from this year's German ASR evaluation also participated last year and did not show any progress on tst2014.

# 3. SLT Track

## 3.1. Definition

The SLT track required participants to translate the English and German talks of *tst2015* from the audio signal (see Section 2). The challenge of this translation task over the MT track is the necessity to deal with automatic, and in general error prone, transcriptions of the audio signal, instead of correct human transcriptions. Furthermore, in contrast to the previous years, this year no manual segmentation into sentences was provided. Therefore, participants needed to develop methods to automatically segment the text and insert punctuation marks.

For German as a source language, participants had to translate into English. For English as source language, participants could choose to translate into one or more languages between Chinese, Czech, French, German, Thai, Vietnamese.

## 3.2. Evaluation

For the evaluation, participants could choose to either use their own ASR technology, or to use ASR output provided by the conference organizers.

For English, the ASR output provided by the organizers was a single system output from one of the five submissions to the ASR track. For German we also used the a single best scored submissions from a different participant.

The results of the translation had to be submitted in NIST XML format, the same format used in the MT track (see Section 4).

Since the participants needed to segment the input into sentences, the segmentation of the reference and the automatic translation was different. In order to calculate the automatic evaluation metric, 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 [30].

## 3.3. Submissions

We received 5 primary and 9 contrastive submissions from nine participants, German to English receiving the most submissions.

## 3.4. Results

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

# 4. MT Track

## 4.1. Definition

The MT TED track basically corresponds to a subtitling translation task. The natural translation unit considered by the human translators volunteering for TED is indeed the single caption — as defined by the original transcript — which in general does not correspond to a sentence, but to fragments of it that fit the caption space. While translators can look at the context of the single captions, arranging the MT task in this way would make it particularly difficult, especially when word re-ordering across consecutive captions occurs. For this reason, we preprocessed all the parallel texts to re-build the original sentences, thus simplifying the MT task.

As already stated in the Introduction, for each translation direction, in-domain training and development data were supplied through the website of the WIT$^3$ [12], while out-of-domain training data were made available through the workshop's website. With respect to edition 2014 of the evaluation campaign, some of the talks added to the TED repository during the last year have been used to define the new evaluation sets (tst2015), while the remaining talks have been included in the training sets. For reliably assessing progress of MT systems over the years, the evaluation set of edition 2014 (tst2014) were distributed as progressive test set, when available. Development sets are either the same of past editions or have been built upon the same talks; tst2013 sets were included into the list of development sets.

With respect to all the other directions, the $DeEn$ MT task is an exception; in fact, its evaluation sets (tst2014 and tst2015) derive from those prepared for the ASR/SLT tracks, which consist of TEDx talks delivered in German language; therefore, no overlap exists with TED talks involved in other tasks. Both TEDx- and TED-based development sets have been released for this direction.

Table 2 provides statistics on in-domain texts supplied for training and evaluation purposes for each MT task. Texts are pre-processed (tokenization, Chinese and Thai segmentation) with the tools used for setting-up baseline systems (see below). Statistics on most development sets can be found in the overview paper of the 2014 edition [11].

MT baselines were trained from TED data only, i.e. no additional out-of-domain resources were used. The standard tokenization via the tokenizer script released with the Europarl corpus [31] was applied to all languages, with the exception of Chinese and Thai; the former was preprocessed by means of the Stanford Chinese Segmenter [32], while the Thai texts were segmented according to the guidelines[1] de-

[1] http://hltshare.fbk.eu/IWSLT2015/InterBEST2009Guidelines-2.pdf

Table 2: Bilingual training and evaluation corpora statistics.

| task | data set | sent | tokens $En$ | tokens foreign | talks |
|---|---|---|---|---|---|
| $En \leftrightarrow Zh$ | train | 210k | 4.27M | 4.02M | 1718 |
| | tst2014 | 1,068 | 20,3k | 20,0k | 12 |
| | tst2015 | 1,080 | 20,8k | 20,7k | 12 |
| $En \leftrightarrow Cs$ | train | 106k | 2.09M | 1.76M | 918 |
| | tst2015 | 1,080 | 20,8k | 17,9k | 12 |
| $En \leftrightarrow Fr$ | train | 208k | 4.23M | 4.51M | 1711 |
| | tst2014 | 1,305 | 24,8k | 27,5k | 15 |
| | tst2015 | 1,080 | 20,8k | 22,0k | 12 |
| $En \leftrightarrow De$ | train | 194k | 3.94M | 3.68M | 1597 |
| $\rightarrow$ | tst2014 | 1,305 | 24,8k | 23,8k | 15 |
| $\rightarrow$ | tst2015 | 1,080 | 20,8k | 19,7k | 12 |
| $\leftarrow$ | $\text{tst2014}_{\text{TEDx}}$ | 1,414 | 28,1k | 27,6k | 10 |
| $\leftarrow$ | $\text{tst2015}_{\text{TEDx}}$ | 2,809 | 41,0k | 38.8k | 14 |
| $En \leftrightarrow Th$ | train | 84k | 1.66M | 2.84M | 746 |
| | tst2015 | 756 | 15,1k | 25,7k | 9 |
| $En \leftrightarrow Vi$ | train | 131k | 2.63M | 3.32M | 1192 |
| | tst2015 | 1,080 | 20,8k | 24,6k | 12 |

fined at InterBEST 2009.[2]

Translation and lexicalized reordering models were trained on the parallel training data by means of the Moses toolkit; 5-gram LMs with improved Kneser-Ney smoothing were estimated on the target side of the training data with the IRSTLM toolkit [33]. The weights of the log-linear interpolation model were optimized on tst2010 with the MERT procedure provided with Moses.

Reference results from baseline MT systems on evaluation sets have been shared among participants after the Evaluation Period, in order to allow them to assess their scores.

### 4.2. Evaluation

The participants to the MT track had to provide the automatic translation of the test sets in NIST XML format. The output had to be case-sensitive, detokenized and had to contain punctuation.

The quality of the translations was measured both automatically, against the human translations created by the TED open translation project, and via human evaluation (Section 5).

Case sensitive scores were calculated for the three automatic standard metrics BLEU, NIST, and TER, as implemented in mteval-v13a.pl[3] and tercom-0.7.25[4], by calling:

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

[2] http://thailang.nectec.or.th/interbest/

[3] http://www.itl.nist.gov/iad/mig/tests/mt/2009/

[4] http://www.cs.umd.edu/ snover/tercom/

Detokenized texts were passed, since the two scorers apply an internal tokenizer. Before the evaluation, Chinese texts were segmented at char level, keeping non-Chinese strings as they are.

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

### 4.3. Submissions

We received submissions from 11 different sites. The total number of primary runs is 40: 26 on *tst2015* and 14 on *tst2014*; 16 primary runs regard the $EnDe$ pair in either one or the other direction, 10 $EnVi$, 6 $EnFr$, 6 $EnZh$ and 2 $EnCs$; in addition, we were asked to evaluate 33 contrastive runs. No submission were received for Thai.

### 4.4. Results

The results on the 2015 official test set for each participant are shown in Appendix A.1. Scores of baseline systems developed as described in Section 4.1 are reported as well.

For all language pairs but one, we show the case-sensitive BLEU, NIST and TER scores. The exception is the English to Chinese task, for which character-level scores are given.

On three language pairs out of five (*En-*{*Zh,Cs,Fr*}), too few submissions were received to make general comments; we can just observe that all systems setup by participants outperformed the baselines. The tasks involving German and Vietnamese attracted more attention. On German, which is a language notoriously difficult to process, the better systems largely beat the basic methods featured in the baselines (the BLEU scores of the best ranked runs are higher than baselines by about 50%); the SU MT English-German system deserves to be mentioned since its approach outclasses even the runner-up. On Vietnamese tasks, participant scores vary a lot as well; differently than on German, submitted runs hardly provided higher quality than baselines; in particular, on Vietnamese-to-English direction, none was able to improve the baseline translation: despite a deep analysis, we were unable to find a plausible explanation for this surprising outcome.

In Appendix A.2 the results on the progress test sets *test2014* are shown. For each task, the baseline performance is provided again, together with the score of the best tst2014 run submitted in 2014 edition of the Evaluation Cam-

paign. The latter scores can slightly differ from those officially disclosed last year because they have been recomputed by means of the Evaluation Server. Only tasks involving Chinese, French and German are considered here since Czech and Vietnamese languages were not proposed in edition 2014.

In comparing the 2015 results to the best 2014 submissions, different remarks can be done depending on the language. On Chinese tasks, no improvement is observed with respect to last year, when four participants sent primary runs: likely, the larger number of attendees increases the chance of measuring good scores. If on the English-to-French task the 2014 best system is definitely better than the unique participant to the 2015 edition, in the opposite direction both 2015 runs outperform the 2014 best run: therefore, we can softly state that some progress has been made on French, at least in translating it into English. On the contrary, no doubts that the German 2015 systems (in both directions) definitely improved over the 2014 edition, especially noting that the two best 2014 runs were a Rover combination of some of the other best runs. Therefore, the systems by SU, on English-German, and by RWTH and KIT, for German-English, resulted outstandingly effective.

## 5. Human Evaluation

Human evaluation was carried out on primary runs submitted by participants to two of the MT TED tasks, namely the MT English-German ($EnDe$) task and MT Vietnamese-English ($ViEn$) task. Following the methodology introduced in 2013, human evaluation was based on *Post-Editing* and systems were ranked according to the HTER (Human-mediated Translation Edit Rate) evaluation metric.

Post-Editing, *i.e.* the manual correction of machine translation output, has long been investigated by the translation industry as a form of machine assistance to reduce the costs of human translation. Nowadays, Computer-aided translation (CAT) tools incorporate post-editing functions, and a number of studies [34, 35] demonstrate the usefulness of MT to increase translators' productivity. The MT TED task offered in IWSLT can be seen as an interesting application scenario to test the utility of MT systems in a real subtitling task.

From the point of view of the evaluation campaign, our goal is to adopt a human evaluation framework able to maximize the benefit for the research community, both in terms of information about MT systems and data and resources to be reused. With respect to other types of human assessment, such as judgments of translation quality (i.e. adequacy/fluency and ranking tasks), the post-editing task has the double advantage of producing *(i)* a set of edits pointing to specific translation errors, and *(ii)* a set of additional reference translations. Both these byproducts are very useful for MT system development and evaluation.[5] Furthermore, the HTER metric [36] - which consists of measuring the minimum edit distance between the MT output and its manually post-edited version (*targeted* reference) - has been shown to correlate quite well with human judgments of MT quality.

[5]All the data produced for human evaluation are publicly available through the WIT$^3$ repository (wit3.fbk.eu).

Table 3: $EnDe$ task: Post-editing information for each Post-editor. PE effort is estimated with HTER. Scores are given in percentage (%).

| PEditor | PE Effort | *std-dev* | Sys TER | *std-dev* |
|---|---|---|---|---|
| PE 1 | 22.49 | 16.44 | 56.43 | 20.77 |
| PE 2 | 42.68 | 26.51 | 55.59 | 20.82 |
| PE 3 | 29.21 | 22.18 | 56.00 | 20.49 |
| PE 4 | 27.66 | 15.50 | 55.77 | 21.17 |
| PE 5 | 22.19 | 17.62 | 56.38 | 20.85 |

The human evaluation dataset and the collected post-edits are described in Section 5.1, whereas the results of the evaluation are presented in Section 5.2.

### 5.1. Evaluation Data

The human evaluation (HE) datasets contain around 10,000 words each and include subsets of the 12 TED Talks composing the IWSLT 2015 official test sets. We selected around the initial 56% of each talk for the $EnDe$ HE dataset, and around 45% for the $ViEn$ one.[6] This choice of selecting a consecutive block of sentences for each talk was determined by the need of realistically simulating a caption post-editing task on several TED talks. The resulting HE sets are composed of 600 segments for $EnDe$ and 500 segments for $ViEn$.

This year we received five primary submissions both for the $EnDe$ task and the $ViEn$ task. For each task, the output of the five systems on the HE set was assigned to five professional translators to be post-edited. To cope with translators' variability, an equal number of outputs from each MT system was assigned randomly to each translator (for all the details about data preparation and post-editing see [11] and Appendix B). The resulting evaluation data for each task consist of five new reference translations for each of the sentences in the HE set. Each one of these five references represents the targeted translation of the system output from which it was derived, and four additional translations are available as well for the evaluation of each MT system.

The main characteristics of the work carried out by post-editors are presented in Tables 3 and 4. In the tables, the post-editing effort for each translator is given. Post-editing effort is to be interpreted as the number of actual edit operations performed to produce the post-edited version and - consequently - it is calculated as the HTER of all the sentences post-edited by each single translator.

As we can see from the tables, PE effort is highly variable among post-editors, even though in different proportions depending on the task (from 22.19% to 42.68% for $EnDe$, and

[6]This different percentage is due to the fact that the number of words for each HE dataset was fixed to 10,000 but the Vietnamese source texts contain a higher number of words with respect to English.

Table 4: $ViEn$ task: Post-editing information for each Post-editor. PE effort is estimated with HTER. Scores are given in percentage (%).

| PEditor | PE Effort | *std-dev* | Sys TER | *std-dev* |
|---|---|---|---|---|
| PE 1 | 37.14 | 21.25 | 61.38 | 20.96 |
| PE 2 | 40.38 | 20.46 | 60.34 | 20.94 |
| PE 3 | 44.76 | 23.57 | 61.66 | 21.74 |
| PE 4 | 46.39 | 25.71 | 61.69 | 21.59 |
| PE 5 | 38.57 | 26.64 | 60.14 | 20.43 |

Table 5: $EnDe$ Task: human evaluation results. Scores are given in percentage (%). The system name next to the HTER score indicates the first system in the ranking with respect to which differences are statistically significant at $p < 0.01$.

| **System Ranking** | **HTER** ***HE Set all PErefs*** | **HTER** HE Set tgt PEref | **TER** HE Set ref | **TER** Test Set ref |
|---|---|---|---|---|
| SU | **16.16**$^{\text{UEDIN}}$ | 21.09 | 51.15 | 51.13 |
| UEDIN | **21.84**$^{\text{PJAIT}}$ | 27.99 | 56.39 | 56.05 |
| KIT | **22.67**$^{\text{PJAIT}}$ | 28.98 | 55.82 | 55.52 |
| HDU | **23.42**$^{\text{PJAIT}}$ | 29.93 | 57.32 | 56.94 |
| PJAIT | **28.18** | 35.68 | 59.51 | 59.03 |
| **Rank Corr.** | | 1.00 | 0.90 | 0.90 |

from 37.14% to 46.39% for $ViEn$). Data about weighted standard deviation confirm post-editor variability, showing that translators produced quite different post-editing effort distributions.

To further study post-editors' behaviour, we exploited the official reference translations available for the two MT tasks and we calculated the TER of the MT outputs assigned to each translator for post-editing (*Sys TER* Column in Tables 3 and 4), as well as the related standard deviation. As we can see from the tables, the documents presented to translators (composed of segments produced by different systems) are very homogeneous, as they show very similar TER scores and standard deviation figures. This also confirms that the procedure followed in data preparation was effective.

The variability observed in post-editing effort - despite the similarity of the input documents - is most probably due to translators' subjectivity in carrying out the post-editing task. These results are in line with those observed in IWSLT 2013 and 2014 for different datasets and language pairs.

### 5.2. Results

The outcomes of the two previous rounds of human evaluation through post-editing [10, 11] demonstrated that HTER computed against all the references produced by all post-editors allow a more reliable and consistent evaluation of MT systems with respect to HTER calculated against the targeted reference only. In light of these findings, also this year systems were officially ranked according to HTER calculated on all the collected post-edits.

Table 6: $ViEn$ Task: human evaluation results. Scores are given in percentage (%). The system name next to the HTER score indicates the first system in the ranking with respect to which differences are statistically significant at $p < 0.01$ (the asterisk indicates significance at $p < 0.05$).

| **System Ranking** | **HTER** ***HE Set all PErefs*** | **HTER** HE Set tgt PEref | **TER** HE Set ref | **TER** Test Set ref |
|---|---|---|---|---|
| JAIST | **32.24**$^{\text{TUT}}$ | 37.25 | 60.10 | 62.35 |
| UMD | **32.71**$^{\text{TUT}}$ | 37.99 | 58.92 | 59.19 |
| PJAIT | **34.27**$^{\text{TUT*}}$ | 40.50 | 59.48 | 62.20 |
| TUT | **38.50** | 43.42 | 62.49 | 62.69 |
| UNETI | **41.42** | 47.97 | 64.21 | 66.33 |
| **Rank Corr.** | | 1.00 | 0.70 | 0.70 |

Official results and rankings are presented in bold in Tables 5 and 6, which also present HTER scores calculated on the targeted reference only and TER results – both on the HE set and on the full test set – calculated against the official reference translation used for automatic evaluation (see Section 4.2 and Appendix A).[7]

To establish the reliability of system ranking, for all pairs of systems we calculated the statistical significance of the observed differences in performance. Statistical significance was assessed with the *approximate randomization* method [37], a statistical test well-established in the NLP community [38] and that, especially for the purpose of MT evaluation, has been shown [39] to be less prone to type-I errors than the bootstrap method [40]. In this study, the approximate randomization test was based on 10,000 iterations. For the $EnDe$ task, we can see in Table 5 that the top-ranked system (SU) is significantly better than all the other systems, while UEDIN, KIT, and HDU are not significantly different from each other but only with respect to PJAIT. For the $ViEn$ task, Table 6 shows that a winning system cannot be indicated, as there is no system that is significantly better than all other systems; the three top-ranking systems (JAIST, UMD, PJAIT) are significantly better than the two bottom-ranking systems (TUT, UNETI).

Some additional observations can be drawn by comparing HTER and TER results given in the tables, which largely confirm previous years' findings. First, we observe a considerable HTER reduction when using all collected post-edits (*all PErefs*) with respect to both the HTER obtained using the targeted post-edit (*tgt PEref*) and the TER obtained using the independent reference (*ref*). This reduction clearly confirms that exploiting all the available reference translations is a viable way to control and overcome post-editors' variability, giving an HTER which is more informative about the real performances of the systems. Moreover, the correlation between evaluation metrics is measured using *Spearman's rank*

[7] Note that since HTER and TER are edit-distance measures, lower numbers indicate better performances.

*correlation coefficient* $\rho \in$ [-1.0, 1.0]. We can see from the tables that TER rankings correlate well with the official HTER. Also, the observed shifts in the ranking occur only where the differences between systems are not statistically significant.

To conclude, the post-editing task introduced for manual evaluation brought benefit to the IWSLT community, and in general to the MT field. Indeed, producing post-edited versions of the participating systems' outputs allowed us to carry out a quite informative evaluation which minimizes the variability of post-editors, who naturally tend to diverge from the post-editing guidelines and personalize their translations. Furthermore, a number of additional reference translations are made available to the community for further development and evaluation of MT systems.

## 6. Conclusions

In this paper, we presented the organisation and outcomes of the 2015 IWSLT Evaluation Campaign. The IWSLT evaluation provides a venue where core technologies for spoken language translation can be evaluated on many different languages and compared not only across research teams but also over time. This year the evaluation was attended by 16 groups – i.e. 6 from Asia, 7 from Europe, and 3 from America. To honor the local organizer of this year, we added among the offered translation directions also English-Vietnamese, which finally attracted several participants. In order to simulate a real subtitling use case, the ASR and SLT tracks were run this year without providing any segmentation of the input speech. Then, in order to improve the automatic evaluation of the MT and SLT tracks, a new evaluation server was developed where participants could submit primary and contrastive runs at any time. Finally, for the two most popular MT runs, a manual evaluation was carried out with professional translators aiming at measuring MT quality in terms of post-editing effort required to fix the MT outputs. Concerning future plans, we are considering to extend the translation task, which now focus on TED talks only, to two other application scenarios: video conferences and lectures.

## 7. Acknowledgements

The human evaluation and part of the work by FBK's authors were carried out under the CRACKER project, which receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 645357.

## 8. 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] M. Eck and C. Hori, "Overview of the IWSLT 2005 evaluation campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Pittsburgh, PA, 2005, pp. 1–22.

[3] P. Michael, "Overview of the IWSLT 2006 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 1–15.

[4] C. S. Fordyce, "Overview of the IWSLT 2007 evaluation campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Trento, Italy, 2007, pp. 1–12.

[5] M. Paul, "Overview of the IWSLT 2008 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Waikiki, Hawaii, 2008, pp. 1–17.

[6] ——, "Overview of the IWSLT 2009 Evaluation Campaign," in *Proceedings of the sixth International Workshop on Spoken Language Translation*, Tokyo, Japan, 2009, pp. 1–18.

[7] M. Paul, M. Federico, and S. Stüker, "Overview of the IWSLT 2010 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Paris, France, 2010, pp. 3–27.

[8] M. Federico, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2011 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, San Francisco, USA, 2011, pp. 11–27.

[9] M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2012 Evaluation Campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Hong Kong, HK, 2012, pp. 11–27.

[10] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "Report on the 10th IWSLT Evaluation Campaign," in *Proceedings of the Tenth International Workshop on Spoken Language Translation (IWSLT 2013)*, Heidelberg, Germany, 2013.

[11] ——, "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, 2014.

[12] 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

[13] K. Wolk and K. Marasek, "PJAIT Systems for the IWSLT 2015 Evaluation Campaign Enhanced by Comparable Corpora," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[14] V. T. Hong, H. V. Thuong, V. N. Van, and T. L. Tien, "System description: IWSLT 2015 for Machine Translation," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[15] V. H. Nguyen, Q. B. Nguyen, T. T. Vu, and C. M. Luong, "The IOIT English ASR system for IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[16] H. D. Tran, J. Dennis, and W. Z. Ng, "The I2R ASR System for IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[17] H.-L. Trieu, T.-Q. Dang, P.-T. Nguyen, and L.-M. Nguyen, "The JAIST-UET-MITI Machine Translation Systems for IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[18] M. Heck, Q. T. Do, S. Sakti, G. Neubig, and S. Nakamura, "The NAIST English Speech Recognition System for IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[19] T. Nomura, H. Tsukada, and T. Akiba, "Improvement of Word Alignment Models for Vietnamese-to-English Translation," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[20] J.-T. Peter, F. Toutounchi, S. Peitz, P. Bahar, A. Guta, and H. Ney, "The RWTH Aachen Machine Translation Systems for IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[21] M. Kazi, B. Thompson, E. Salesky, T. Anderson, G. Erdmann, E. Hansen, B. Ore, K. Young, J. Gwinnup, M. Hutt, and C. May, "The MITLL-AFRL IWSLT 2015 MT System," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[22] M. Huck and A. Birch, "The Edinburgh Machine Translation Systems for IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[23] M. Á. D. A. Teba, A. A. M. Villaronga, S. P. Gozalbes, A. G. Pastor, J. A. S. Navarro, J. C. Saiz, and A. J. Císcar, "The MLLP ASR Systems for IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[24] L. Jehl, P. Simianer, J. Hitschler, and S. Riezler, "The Heidelberg University English-German translation system for IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[25] M. G. Martinez, L. Barrault, A. Rousseau, P. Deléglise, and Y. Estève, "The LIUM ASR and SLT Systems for IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[26] A. Axelrod and M. Carpuat, "The UMD Machine Translation Systems at IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[27] T.-L. Ha, J. Niehues, E. Cho, M. Mediani, and A. Waibel, "The KIT Translation Systems for IWSLT 2015," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[28] M. Müller, T. S. Nguyen, M. Sperber, K. Kilgour, S. Stüker, and A. Waibel, "The 2015 KIT IWSLT Speech-to-Text Systems for English and German," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[29] M.-T. Luong and C. D. Manning, "Stanford Neural Machine Translation Systems for Spoken Language Domains," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam, 2015.

[30] 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.

[31] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand, September 2005, pp. 79–86.

[32] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning, "A conditional random field word segmenter," in *Fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

[33] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," in *Proceedings of Interspeech*, Melbourne, Australia, 2008, pp. 1618–1621.

[34] M. Federico, A. Cattelan, and M. Trombetti, "Measuring user productivity in machine translation enhanced computer assisted translation," in *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*, 2012. [Online]. Available: http://www.mt-archive.info/AMTA-2012-Federico.pdf

[35] S. Green, J. Heer, and C. D. Manning, "The efficacy of human post-editing for language translation," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 439–448.

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

[37] E. W. Noreen, *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Wiley Interscience, 1989.

[38] N. Chinchor, L. Hirschman, and D. D. Lewis, "Evaluating message understanding systems: An analysis of the third message understanding conference (muc-3)," *Computational Linguistics*, vol. 19, no. 3, pp. 409–449, 1993.

[39] S. Riezler and J. T. Maxwell, "On some pitfalls in automatic evaluation and significance testing for MT," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 57–64. [Online]. Available: http://www.aclweb.org/anthology/W/W05/W05-0908

[40] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[41] M. Federico, N. Bertoldi, M. Cettolo, M. Negri, M. Turchi, M. Trombetti, A. Cattelan, A. Farina, D. Lupinetti, A. Martines, A. Massidda, H. Schwenk, L. Barrault, F. Blain, P. Koehn, C. Buck, and U. Germann, "The MateCat Tool," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, August 2014, pp. 129–132. [Online]. Available: http://www.aclweb.org/anthology/C14-2028

# Appendix A. Automatic Evaluation

## A.1. Official Testset (*tst2015*)

· All the sentence IDs in the IWSLT 2015 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 (%).

**TED : ASR English (ASR$_{EN}$)**

| System | WER | (# Errors) |
|---|---|---|
| HLT-I2R | 7.7 | (1,403) |
| IOIT | 13.8 | (2,523) |
| KIT | 9.2 | (1,689) |
| NAIST | 12.0 | (2,197) |
| MITLL-AFRL | **6.6** | **(1,201)** |
| MLLP | 13.3 | (2,421) |

**TED : ASR German (ASR$_{DE}$)**

| System | WER | (# Errors) |
|---|---|---|
| KIT | 20.3 | (6,931) |
| LIUM | **17.6** | **(6,010)** |
| MLLP | 43.3 | (14,787) |

**TED : SLT English-Chinese (SLT$_{EnZh}$)**

| System | *character-based* | |
|---|---|---|
| | BLEU | TER |
| MITLL-AFRL | **18.02** | **75.75** |

**TED : SLT English-French (MT$_{EnFr}$)**

| System | *case sensitive* | | *case insensitive* | |
|---|---|---|---|---|
| | BLEU | TER | BLEU | TER |
| LIUM | **18.51** | **79.06** | **20.02** | **76.41** |

**TED : SLT English-German (MT$_{EnDe}$)**

| System | *case sensitive* | | *case insensitive* | |
|---|---|---|---|---|
| | BLEU | TER | BLEU | TER |
| KIT | **0.1618** | **78.28** | **16.92** | **76.71** |

**TED : SLT German-English (MT$_{DeEn}$)**

| System | *case sensitive* | | *case insensitive* | |
|---|---|---|---|---|
| | BLEU | TER | BLEU | TER |
| KIT | **19.64** | **62.22** | **20.83** | **60.23** |
| RWTH | 18.79 | 65.18 | 20.23 | 62.62 |

**TED : MT English-Chinese (MT$_{EnZh}$)**

| System | *character-based* | | |
|---|---|---|---|
| | BLEU | NIST | TER |
| UEDIN | **25.39** | 6.3985 | 60.83 |
| MITLL-AFRL | 24.31 | **6.4136** | **59.00** |
| BASELINE | 21.86 | 5.8640 | 65.94 |

**TED : MT Chinese-English (MT$_{ZhEn}$)**

| System | *case sensitive* | | |
|---|---|---|---|
| | BLEU | NIST | TER |
| MITLL-AFRL | **16.86** | **5.2565** | **67.31** |
| BASELINE | 13.59 | 4.8918 | 68.01 |

**TED : MT English-Czech (MT$_{EnCs}$)**

| System | *case sensitive* | | |
|---|---|---|---|
| | BLEU | NIST | TER |
| PJAIT | **17.17** | **5.1056** | **63.00** |
| BASELINE | 14.74 | 4.7458 | 65.80 |

**TED : MT Czech-English (MT$_{CsEn}$)**

| System | *case sensitive* | | |
|---|---|---|---|
| | BLEU | NIST | TER |
| PJAIT | **25.07** | 6.4026 | 55.74 |
| BASELINE | 22.44 | 6.1186 | 57.99 |

**TED : MT English-French (MT$_{EnFr}$)**

| System | *case sensitive* | | |
|---|---|---|---|
| | BLEU | NIST | TER |
| PJAIT | **32.79** | **7.3222** | **49.15** |
| BASELINE | 30.54 | 6.9957 | 51.51 |

**TED : MT French-English (MT$_{FrEn}$)**

| System | *case sensitive* | | |
|---|---|---|---|
| | BLEU | NIST | TER |
| PJAIT | **32.75** | 7.2769 | 48.41 |
| UMD | 32.59 | **7.3708** | **47.12** |
| BASELINE | 31.94 | 7.3415 | 47.55 |

**TED : MT English-German (MT$_{EnDe}$)**

| System | *case sensitive* | | |
|---|---|---|---|
| | BLEU | NIST | TER |
| SU | **30.85** | **6.9898** | **51.13** |
| KIT | 26.18 | 6.4640 | 55.52 |
| UEDIN | 26.02 | 6.4518 | 56.05 |
| HDU | 24.96 | 6.3170 | 56.94 |
| PJAIT | 22.51 | 6.0412 | 59.03 |
| BASELINE | 20.08 | 5.7613 | 61.37 |

**TEDX : MT German-English (MT$_{DeEn}$)**

| System | *case sensitive* | | |
|---|---|---|---|
| | BLEU | NIST | TER |
| RWTH | **31.50** | **7.7932** | **47.11** |
| KIT | 31.08 | 7.7471 | 47.24 |
| PJAIT | 26.08 | 7.0350 | 52.34 |
| BASELINE | 21.78 | 6.4984 | 55.45 |

**TED : MT English-Vietnamese (MT$_{EnVi}$)**

| System | *case sensitive* | | |
|---|---|---|---|
| | BLEU | NIST | TER |
| PJAIT | **28.39** | 6.6650 | 56.01 |
| JAIST | 28.17 | **6.7092** | 55.84 |
| KIT | 26.60 | 6.4014 | 58.26 |
| SU | 26.41 | 6.5986 | **55.60** |
| UNETI | 22.93 | 6.0218 | 60.33 |
| BASELINE | 27.01 | 6.4716 | 58.42 |

**TED : MT Vietnamese-English (MT$_{ViEn}$)**

| System | *case sensitive* | | |
|---|---|---|---|
| | BLEU | NIST | TER |
| PJAIT | **23.46** | 5.7314 | 62.20 |
| UMD | 21.57 | **5.7831** | **59.19** |
| JAIST | 21.53 | 5.6413 | 62.35 |
| UNETI | 20.18 | 5.1443 | 66.33 |
| TUT | 19.78 | 5.4559 | 62.69 |
| BASELINE | 24.61 | 5.9259 | 59.32 |

# A.2. Progress Testset (*tst2014*)

· All the sentence IDs in the IWSLT 2014 testset were used to calculate the automatic scores for each run submission.
· MT systems are ordered according to the *BLEU* metric.
· *WER*, *BLEU* and *TER* scores are given as percent figures (%).

**TED : ASR English (ASR$_{EN}$)**

| System | WER | (# Errors) |
|---|---|---|
| HLT-I2R | 8.9 | (1,950) |
| IOIT | 13.9 | (3,036) |
| KIT | 9.7 | (1,689) |
| NAIST | 10.4 | (2,268) |
| MITLL-AFRL | **7.1** | **(1,549)** |
| MLLP | 19.5 | (4,258) |

**TED : ASR German (ASR$_{DE}$)**

| System | WER | (# Errors) |
|---|---|---|
| KIT | **(24.0)** | **(5,660)** |
| LIUM | 26.5 | (6,254) |
| MLLP | 49.4 | (11,657) |

**TED : MT English-Chinese (MT$_{EnZh}$)**

| System | *character-based* BLEU | NIST | TER |
|---|---|---|---|
| UEDIN | **19.63** | **5.5483** | 68.05 |
| MITLL-AFRL | 18.51 | 5.5294 | **66.73** |
| BASELINE | 17.74 | 5.2514 | 71.23 |
| BEST IWSLT2014 | 21.64 | 5.8732 | 65.66 |

**TED : MT Chinese-English (MT$_{ZhEn}$)**

| System | *case sensitive* BLEU | NIST | TER |
|---|---|---|---|
| MITLL-AFRL | **14.14** | **4.6736** | **72.55** |
| BASELINE | 11.43 | 4.3935 | 72.65 |
| BEST IWSLT2014 | 15.63 | 4.9138 | 69.67 |

**TED : MT English-French (MT$_{EnFr}$)**

| System | *case sensitive* BLEU | NIST | TER |
|---|---|---|---|
| PJAIT | **31.88** | **7.4901** | **47.92** |
| BASELINE | 30.31 | 7.2488 | 50.18 |
| BEST IWSLT2014 | 36.99 | 7.9127 | 45.20 |

**TED : MT French-English (MT$_{FrEn}$)**

| System | *case sensitive* BLEU | NIST | TER |
|---|---|---|---|
| UMD | **33.20** | **7.4807** | **46.32** |
| PJAIT | 32.92 | 7.3747 | 48.25 |
| BASELINE | 32.20 | 7.3677 | 47.60 |

**TED : MT English-German (MT$_{EnDe}$)**

| System | *case sensitive* BLEU | NIST | TER |
|---|---|---|---|
| SU | **27.58** | **6.8218** | **52.50** |
| UEDIN | 24.01 | 6.3821 | 57.04 |
| KIT | 23.31 | 6.4106 | 56.51 |
| HDU | 23.22 | 6.2500 | 57.81 |
| PJAIT | 20.68 | 5.9978 | 59.78 |
| BASELINE | 18.49 | 5.7409 | 61.66 |
| BEST IWSLT2014 | 23.25 | 6.3415 | 57.27 |

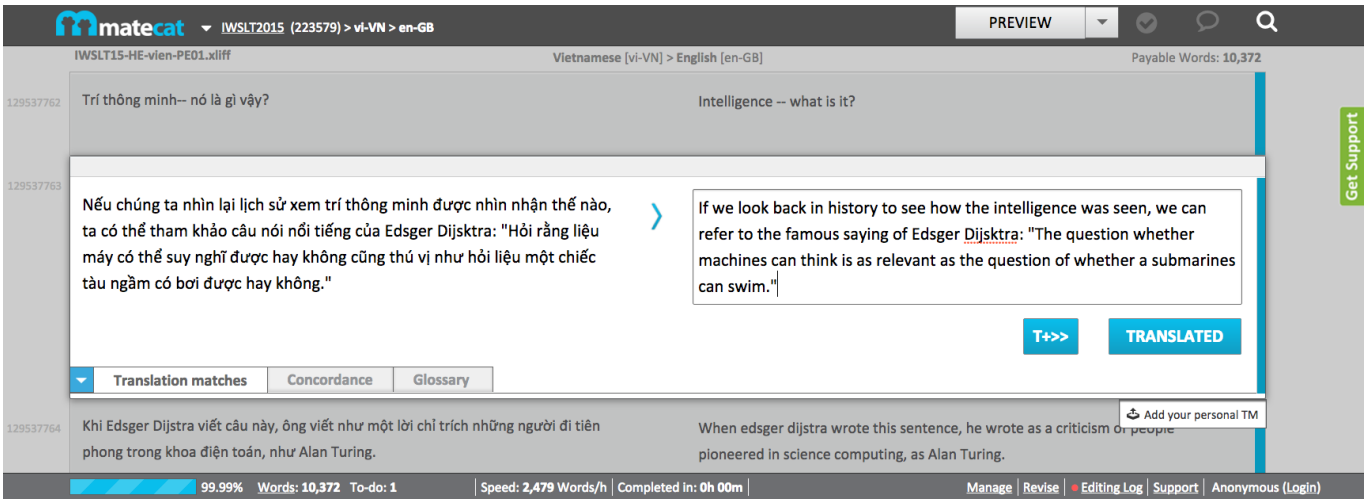
**TEDX : MT German-English (MT$_{DeEn}$)**

| System | *case sensitive* BLEU | NIST | TER |
|---|---|---|---|
| RWTH | **26.18** | **6.7160** | **55.15** |
| KIT | 25.18 | 6.5795 | 55.76 |
| PJAIT | 21.92 | 6.0407 | 60.59 |
| BASELINE | 17.99 | 5.5186 | 64.36 |
| BEST IWSLT2014 | 25.80 | 6.7011 | 55.07 |

# Appendix B. Human Evaluation

## Interface used for the bilingual post-editing task

Post-editing was carried out using MateCat[8] [41], which is a web-based open-source professional CAT tool developed within the EU funded project Matecat.




## Post-editing instructions given to professional translators

In this task you are presented with automatic translations of TED Talks captions.

You are asked to post-edit the given automatic translation by applying the minimal edits required to transform the system output into a fluent sentence with the same meaning as the source sentence.

While post-editing, remember that the post-edited sentence is to be intended as a transcription of spoken language. Also, depending on the style of the source language talk, you can use the corresponding style in the target language (*e.g.* if the talk uses a friendly/colloquial style you can use informal words too).

Note also that the focus is the correctness of the single sentence within the given context, NOT the consistency of a group of sentences. Hence, surrounding segments should be used to understand the context but NOT to enforce consistency on the use of terms. In particular, different but correct translations of terms across segments should not be corrected.

The document you have to post-edit is composed of around the first half of 12 different talks. Below you can find the name of the speaker and the title of each talk.

1. Alex Wissner-Gross: A new equation for intelligence.
2. Ash Beckham: We're all hiding something let's find the courage to open up.
3. Mary Lou Jepsen: Could future devices read images from our brains?
4. Ziauddin Yousafzai: My daughter Malala.
5. Geena Rocero: Why I must come out.
6. Kevin Briggs: The bridge between suicide and life.
7. Chris Kluwe: How augmented reality will change sports and build empathy.
8. Stella Young: I'm not your inspiration thank you very much.
9. Zak Ebrahim: I am the son of a terrorist here's how I chose peace.
10. David Chalmers: How do you explain consciousness.
11. Meaghan Ramsey: Why thinking you're ugly is bad for you.
12. Marc Kushner: Why the buildings of the future will be shaped by you.

[8] www.matecat.com