The IWSLT 2016 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 2016 Evaluation Campaign featured two tasks: the translation of talks and the translation of video conference conversations. While the first task extends previously offered tasks with talks from a different source, the second task is completely new. For both tasks, three tracks were organised: automatic speech recognition (ASR), spoken language translation (SLT), and machine translation (MT). Main translation directions that were offered are English to/from German and English to French. Additionally, the MT track included English to/from Arabic and Czech, as well as French to English. We received this year run submissions from 11 research labs. All runs were evaluated with objective metrics, while submissions for two of the MT talk tasks were also evaluated with human post-editing. Results of the human evaluation show improvements over the best submissions of last year.

1. Introduction

We report here on the outcomes of the 2016 evaluation campaign organized by the International Workshop of Spoken Language Translation. 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 thirteen editions organized from 2004 to 2016, 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

The 2016 IWSLT evaluation focused on two tasks: the Talk task, including translation of TED talks corpus [5] and lectures from the QED corpus [6], and the Microsoft Speech Language Translation (MSLT) task [7], that consists of translating conversations conducted via Skype.

The translation directions considered this year for the SLT track were English to German and French for the Talk task, and English to/from German and English to French for the MSLT task. The ASR track included task for English and German, while the MT track offered additional translation directions for the TED Talk task, namely: English to/from Czech and Arabic and French 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 SLT and MT tracks, an evaluation server was set up so that participants could autonomously score their runs on different dev and test sets. This year, 11 groups participated in the evaluation (see Table 3). 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. Then, we describe the protocol and outcomes of the human evaluation that we carried out on a subset of runs submitted to the MT track. The paper ends with an appendix reporting all the detailed results of this year's evaluation.

Table 1: Example of a sentence pair from the QED data.

Language	Transcript
English	So in this video I'm just going to do a ton of examples.
German	Daher werde ich in diesem Video viele Beispiele durchrechnen.

2. Tasks

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. We augment the Talk task by including more challenging lecture data. Further, we introduced a new task on video-conference conversations. Unlike in previous years, we also limited the scope of the evaluation to few languages: English, German, French, and one low resourced European language. The main reason for this was to avoid dispersion of participants in too many tasks.

2.1. Talk Task

TED talks are challenging due to their variety in topics, which can be considered unlimited for all practical purposes. With respect to the type of language, TED talks are, however, very well behaved. Before being delivered, TED talks are rehearsed rigorously. Therefore, the talks tend not to show spontaneous speech phenomena, but are rather well formed. However, the majority of talks held in the world are not that well formed and well rehearsed, but rather more spontaneous and of lower quality. A prominent example of such type of talk is given by academic lectures. In order to address more lifelike talks, we thus included data from from the QCRI Educational Domain (QED) Corpus[1] [6] into our talk task. This data is obtained from subtitles created on the Amara platform of videos from Khan Academy, Coursera, Udacity, etc. Table 1 gives an example of a transcription and translation from the corpus.

2.2. MSLT Task

MSLT stands for Microsoft Speech Language Translation and refers to data collected within a video conference scenario.[2] Translating video conference conversations is a challenging task due to the nature of the language used in conversations, which is often not planned, informal in nature, ungrammatical, using special idioms etc. Therefore, while maybe not as broad in domain as talks and lectures, this task represents a challenge that goes beyond the translation of talks. A detailed description of the data we have been used in the evaluation is provided in [7].

Table 2: Example of a sentence pair from the MSLT data.

Language	Transcript
German	ähm wir haben grade über Platten geredet, und über, über Musik, Musik Stream, was mich halt irgendwie nervt ist das bei so vielen Platten vorn so krass viel Werbung dazwischen geschaltet wird, und das find ich äh sehr störend, ja.
English	We just talked about albums and about streaming music, which just bugs me somehow, that for so many albums, so much advertising is placed before and in between them. And I find that very disruptive, yes.

The test data that has been made available from Microsoft Research consists of bilingual conversations, where each speaker was speaking in his own language but was able to understand the other dialog partner's language. In this way natural conversations could be recorded. Audio was then manually processed to produce transcripts, transformed transcripts (cleaned of disfluencies), and translations (in or out of English). Table 2 shows an example from such a dialogue in English and German. For proprietary issues, development and evaluation sets for the MSLT task were distributed only to participants who signed a data license agreement.

3. ASR Track

3.1. Definition

The *Automatic Speech Recognition* (ASR) track for IWSLT 2016 addressed both the Talk and the MSLT tasks described in Section 2.1 and 2.2, respectively.

The results of the recognition of the Talk task is used for two purposes. It is used to measure the performance of ASR systems on this task and it is used as input for the SLT track, see Section 4.

3.2. Evaluation

Participants had to submit the results of the recognition of the *tst2016* sets 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.

In order to measure the progress of the systems over the years, participants to the English Talk task also had to provide results on the test set from 2015, i.e. *tst2015*.

[1] http://alt.qcri.org/resources/qedcorpus/

[2] (http://research.microsoft.com/en-us/about/speech-to-speech-milestones.aspx

Table 3: List of Participants

RWTH	Rheinisch-Westfälische Technische Hochschule Aachen, Germany [8, 9]
MITLL-AFRL	MIT Lincoln Laboratory and Air Force Research Laboratory, USA [10]
UEDIN	University of Edinburgh, United Kingdom [11]
LIMSI	LIMSI, France [12]
UMD	University of Maryland, USA [13]
KIT	Karlsruhe Institute of Technology, Germany [14, 15]
FBK	Fondazione Bruno Kessler, Italy [16]
RACAI	Research Institute for AI of the Romanian Academy, Romania [17]
UFAL	Charles University, Czech Republic [18]
QCRI	Qatar Computing Research Institute, Qatar Foundation, Qatar [19]
IOIT	University of Information and Communication Technology, Thai Nguyen University, Vietnam [20]

3.3. Submissions

For this year's evaluation we received primary submissions from five sites.

For the English Talk task we received four primary runs on tst2016 and three on *tst2015*. We also received five contrastive submissions from two sites for *tst2016* and two contrastive submissions from one site for *tst2015*.

For the English MSLT task we received primary submissions from two sites, while for German we received two primary submissions. For German we further received a total of six contrastive submissions from two sites.

3.4. Results

The detailed results of the primary submissions of the evaluation in terms of word error rate (WER) can be found in Appendix A.

For the English Talk task the word error rates of the submitted systems on tst2016 are in the range of 7.2%–59.2%. On the TED only portion of that test set the best WER is 6.4% while for the QED portion the best WER is 10.4%. This shows that the QED data is significantly more difficult than the TED data.

For the English MSLT task WERs range from 22.3% to 29.5%, while for the German MSLT task WERs scored are between 19.7% and 25.5%.

Three participants of this year's English Talk task also participated last year. All of them showed significant progress on *tst2015*, absolute WER improvements ranging from 1.9–0.5 percentage points. This year the lowest WER on *tst2014* was 6.1% as compared to 6.6% last year.

4. SLT Track

4.1. Definition

The SLT track covered both the MSLT and Talk tasks. In particular, results of the two Talk sources were kept distinct, namely TED an QED. The participants should translate from the English and German audio signal (see Section 3). 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, for the lecture tasks no manual segmentation into sentences was provided. Therefore, participants needed to develop methods to automatically segment the automatic transcript and insert punctuation marks.

For the lecture tasks, participants could translate from English into German and French. For the MSLT task, the translation directions English to German and French as well as German to French were offered.

4.2. 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 both input languages, the ASR output provided by the organizers was a single system output from one of the submissions to the ASR track.

The results of the translation had to be submitted in NIST XML format, the same format used in the MT track (see Section 5).

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 [21].

4.3. Submissions

We received one primary submissions for every task. These submissions were created by two different participants.

4.4. Results

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

5. MT Track

5.1. Definition

Also, the MT track featured the Talk and the MSLT tasks. As for the other tracks, tests on the different Talk sources (TED and QED) were kept distinct.

Statistics of the distributed sets for the MSLT task are provided in Table 5.

In this edition, the QED exercise was considered as a dry-run and as such no specific training nor development sets were released; participants could exploit in any way the data from the QED corpus, with the exception of a specific list of QED talks.[3]

The TED exercise was in all respects the same as that proposed in the last editions of the evaluation campaign. Differently than for QED, in-domain training and development data were supplied through the website of the WIT[3] [5], while out-of-domain training data were made available through the workshop's website. With respect to edition 2015, some of the talks recently added to the TED repository have been used to define the new evaluation sets (*tst2016*), while the remaining new talks have been included in the training sets. For reliably assessing progress of MT systems over the years, the evaluation sets of edition 2015 (*tst2015*) were distributed as progressive test sets, when available. Development sets are either the same of past editions or have been built upon the same talks.

Table 4 provides statistics on in-domain texts supplied for training and evaluation purposes for each language pair of the TED and QED exercises. All texts were tokenized with the tokenizer script released with the Europarl corpus [22], but Arabic texts, which were processed by means of the QCRI Arabic Normalizer 3.0 [23].

Statistics on TED development sets can be found in the overview papers of 2014 and 2015 editions.

5.2. Evaluation

Participants of the track had to provide MT outputs of the test sets in NIST XML format. Outputs had to be case-sensitive, detokenized and punctuated.

The quality of the translations was measured both automatically, against human translations created by the TED open translation project, and via human evaluation (Section 6).

Case sensitive scores were calculated with the three automatic standard metrics BLEU, NIST, and TER, as implemented in mteval-v13a.pl[4] and tercom-0.7.25[5], by calling:

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

[3] available here: https://sites.google.com/site/iwsltevaluation2016/home/off-limit-ted-talks

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

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

Table 4: Talk task: statistics on bilingual train and test sets.

direction/source		data set	seg	tokens *En*	tokens foreign	talks
En↔ Ar	TED	train	240k	4.91M	3.91M	1,852
		tst2015	1,080	20,8k	16,2k	12
		tst2016	1,133	23,2k	18,1k	13
	QED	tst2016	549	5,2k	3,9k	3
En↔ Cs	TED	train	114k	2.26M	1.90M	999
		tst2015	1,080	20,8k	17,9k	12
		tst2016	1,133	23,2k	19,5k	13
	QED	tst2016	549	5,2k	3,8k	3
En↔ Fr	TED	train	220k	4.50M	4.79M	1,824
		tst2015	1,080	20,8k	22,0k	12
		tst2016	1,133	23,2k	23,9k	13
	QED	tst2016	549	5,2k	5,1k	3
En↔ De	TED	train	197k	3.96M	3.69M	1,611
		tst2015	1,080	20,8k	19,7k	12
		tst2016	1,133	23,2k	20,7k	13
	QED	tst2016	549	5,2k	4,6k	3

Table 5: MSLT task: statistics on bilingual dev and test sets.

direction	data set	seg	tokens source	tokens target
En→ Fr	dev2016	5,292	44,9k	49,6k
	tst2016	4,854	45,3k	49,3k
En→ De	dev2016	5,292	44,9k	44,6k
	tst2016	4,854	45,3k	45,2k
De→ En	dev2016	3,335	31,1k	29,2k
	tst2016	3,798	33,1k	31,2k

Detokenized texts were used, since the two scoring scripts apply their own internal tokenizers. Before the evaluation, Arabic texts were normalized with the QCRI Arabic Normalizer 3.0 [23].

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 any 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 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.

5.3. Submissions

We received submissions from 10 different sites. The total number of primary runs is 60: 40 on *tst2016* and 20 on the progressive *tst2015* set; the 40 primary on *tst2016* are dis-

tributed between the three MT exercises as follows: 9 on MSLT, 11 on QED and 20 on TED.

5.4. Results

The results on the 2016 official test set for each participant are shown in Appendix A.1. Appendix A.2 provides results on the progress test sets *test2015*, which only regard the TED exercise; for the language pairs proposed also in edition 2015, the score of the best TED *test2015* run submitted last year is given as well. For each test set, case-sensitive BLEU, NIST and TER scores are reported. (Notice tht QED runs were also scored in case insensitive mode for the reason discussed below.)

Assuming that for a given language pair the quality of the translation is related to the difficulty of the task, it results that MSLT seems easier than TED, while QED seems in general more difficult than TED, with few exceptions (e.g. QCRI in English-to-Arabic).

One issue of the QED test set came out lately. Both the transcriptions and the translations of the three lectures of the QED test set show inconsistent letter casing. For this reason, we scored the submissions also in case insensitive mode. In fact, the observed difference with the case sensitive scores is unusually high (over 5 absolute BLEU points), suggesting that this aspect should be handled better in the future.

By comparing the 2016 results on the progress test set to the best 2015 results, outstanding improvements can be observed on the TED French-English and English-French directions. The two participants which adopted a neural MT approach outperformed the best system of 2015 by 6-7 absolute BLEU points. It is also worth noticing that the outstanding scores reached in 2015 on the English-German TED task have been almost matched this year by three participants, also using NMT systems.

Finally, we want to highlight the "asymmetry" of the pairs involving the Arabic and Czech languages: in both cases, the translation from English into those two languages is much more difficult than the translation in the other way round.

6. Human Evaluation

Human evaluation was carried out for two MT TED tasks, namely English-German ($EnDe$) and English-French ($EnFr$). Following the methodology introduced in IWSLT 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 [24, 25] 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.[6] Furthermore, the HTER metric [26] - 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.

The human evaluation dataset and the collected post-edits are described in Section 6.1, whereas the results of the evaluation are presented in Section 6.2.

6.1. Evaluation Data

The human evaluation (HE) dataset created for each task was a subset of the 2015 test set *(tst2015)*. Both the $EnDe$ and $EnFr$ *tst2015* test sets are composed of the same 12 TED Talks, and around the initial 56% of each talk was included in the HE set. 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 $EnDe$ and $EnFr$ HE sets are identical and include 600 segments, corresponding to around 10,000 English words.

As regards the MT systems selected for human evaluation, different criteria were followed for the two tasks. For the $EnDe$ task, all four submitted primary runs were post-edited. For the $EnFr$ task, the top-two systems according to automatic evaluation (see Appendix A) were included in the evaluation. Since both top-ranking submissions were neural MT systems, for comparison purposes we additionally run and evaluated two state-of-the-art phrase-based systems, namely Google Translate and ModernMT.[7] Finally, to measure the progress with respect to last year's campaign, a system participating in IWSLT 2015 was also added to evaluation.

For each task, the output of the selected systems on the HE set was assigned to professional translators to be post-edited, namely four MT outputs for $EnDe$ and five for $EnFr$. 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 [27] and Appendix B).

The resulting evaluation data consists of multiple new

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

[7] www.modernmt.eu

Table 6: $EnDe$ TED Talk task (HE *tst2015*): 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	22.48	17.48	53.78	22.20
PE 2	23.22	18.92	54.20	22.82
PE 3	10.68	14.04	53.26	21.55
PE 4	42.22	24.25	53.43	22.24

Table 7: $EnFr$ TED Talk task (HE *tst2015*): 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	35.60	20.43	46.08	21.80
PE 2	21.89	15.64	46.32	20.89
PE 3	19.69	15.27	45.99	21.16
PE 4	13.90	12.70	46.40	20.51
PE 5	23.95	17.08	46.43	21.52

reference translations for each of the sentences in the HE set. Each one of these references represents the targeted translation of the system output from which it was derived, while the post-edits of the other systems are available for evaluation as additional references.

The main characteristics of the work carried out by post-editors are presented in Tables 6 and 7. In the tables, the post-editing (PE) effort for each translator is given. PE 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 10.68% to 42.22% for $EnDe$, and from 13.90% to 35.60% for $EnFr$). Data about weighted standard deviation confirm post-editor variability, showing that translators produced quite different PE 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 6 and 7), as well as the related weighted 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 PE 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 starting from IWSLT 2013 for different datasets and language pairs.

Table 8: $EnDe$ TED Talk task (HE *tst2015*): human evaluation results. Scores are given in percentage (%). The system name next to the mTER score indicates the first system in the ranking w.r.t. which differences are statistically significant at $p < 0.01$.

System Ranking	**mTER HE Set 5 PErefs**	**HTER** HE Set tgt PEref	**TER** HE Set ref	**TER** Test Set ref
UEDIN	**13.31**$^{\text{SU-15}}$	21.72	52.40	52.02
KIT	**14.12**$^{\text{FBK}}$	22.29	52.97	52.47
SU-15	**14.98**$^{\text{UFAL}}$	*21.09*	*51.15*	*51.13*
FBK	**15.95**$^{\text{UFAL}}$	25.42	*51.88*	*51.56*
UFAL	**21.89**	28.82	57.41	57.08
Rank Corr.		0.70	0.20	0.20

Table 9: $EnFr$ TED Talk task (HE *tst2015*): human evaluation results. Scores are given in percentage (%). The system name next to the mTER score indicates the first system in the ranking w.r.t. which differences are statistically significant at $p < 0.01$.

System Ranking	**mTER HE Set 5 PErefs**	**HTER** HE Set tgt PEref	**TER** HE Set ref	**TER** Test Set ref
UEDIN	**12.41**$^{\text{MMT}}$	17.89	43.46	44.46
FBK	**12.98**$^{\text{MMT}}$	18.51	*42.72*	*43.96*
MMT	**19.50**$^{\text{PJAIT-15}}$	25.18	48.15	49.46
GT	**19.98**$^{\text{PJAIT-15}}$	25.29	48.80	49.82
PJAIT-15	**21.90**	28.28	*48.09*	*49.15*
Rank Corr.		1.00	0.60	0.60

6.2. Results

The outcomes of the previous rounds of human evaluation through post-editing [28, 27, 29] demonstrated that multi-reference TER (mTER) – where TER is computed against all available post-edits – allows a more reliable and consistent evaluation of the real overall MT system performance with respect to HTER – where TER is calculated against the targeted reference only. In light of these findings, also this year systems were officially ranked according to mTER calculated on all the collected post-edits.

To allow a comparable overview of the results obtained for the two different language pairs, the evaluation framework of the two tasks was kept as similar as possible. To this purpose, since we collected five post-edits for $EnFr$ and only four for $EnDe$, we added to the evaluation of the $EnDe$ task the winning run (and corresponding post-edit) of last year's campaign, *i.e.* the neural MT system SU-15 [30].

Results and rankings are presented in bold in Tables 8 and 9, which also give 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

5.2 and Appendix A).[8]

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 [31], a statistical test well-established in the NLP community [32] and that, especially for the purpose of MT evaluation, has been shown [33] to be less prone to type-I errors than the bootstrap method [34]. In this study, the approximate randomization test was based on 10,000 iterations. The results of the test are also shown in Tables 8 and 9, where we report - next to the mTER score of each system - the name of the first system in the ranking with respect to which differences are statistically significant.

In the $EnDe$ task, a winning system cannot be indicated, since the top-ranking system (UEDIN) is not significantly different from the second one (KIT). In general, the ranking is not clearly defined, since the four top-ranking systems – which are all neural – are very close to each other, with UEDIN (first) significantly better than SU-15 (third), and KIT (second) significantly better than FBK (fourth) but not different from SU-15 (third). Moreover, all the neural MT systems are significantly better than the UFAL phrase-based system, overtooking it with a large margin (ranging from 6 to 9 mTER points). This outcome confirms last year's findings,[9] since the new neural systems perform very similarly to SU-15, UFAL compares well with last year's state of the art phrase-based systems, and the neural approach markedly outperfoms the phrase-based one.

The outcome of the $EnFr$ task is quite similar to that of the $EnDe$ task. There is not a single winning system, since the two top-ranking systems – which are both neural – are not significantly different from each other. Also for this language pair, neural systems are significantly better than all the three phrase-based systems, with an impressive improvement of at least 7 mTER points. Finally, the two external state of the art systems (MMT and GT) rank on par, while significantly outperforming last year's system PJAIT-15.

As a general comparison between $EnFr$ and $EnDe$ language pairs, mTER scores confirm that translating from English to German is more difficult than translating into French. However it is interesting to note that the differences revealed are not so marked as those given by a fully automatic metric such as TER computed on one independent reference. As an example, by taking the average performance of the two top-ranking systems in both tasks, we see that the relative difference between $EnDe$ and $EnFr$ in terms of mTER amounts to around 7%, while in terms of TER it amounts to around 16%. The evaluation carried out on multiple post-edits is more reliable and gives more accurate information about differences between language pairs.

Some additional observations can be drawn by comparing mTER and TER results given in the tables, which largely confirm previous years' findings. First, we observe a considerable TER reduction when using all collected post-edits (*5 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 overall score which is more informative about the real performances of the systems.

Moreover, the correlation between evaluation metrics is measured using *Spearman's rank correlation coefficient* $\rho \in$ [-1.0, 1.0]. We can see from the tables that TER rankings do not correlate well with the official mTER. A possible explanation is that – differently from mTER – when systems are very close to each other TER calculated against one independent reference does not allow to discriminate between systems. To verify this hypothesis, we calculated the statistical significance of the differences between systems according to TER. Indeed, for the $EnFr$ task, shifts in the ranking occur only where the differences between systems are not statistically significant (FBK *vs.* UEDIN and PJAIT-15 *vs.* MMT and GT). For the $EnDe$ task, the situation is more blurred, since SU-15 and FBK are not significantly different from UEDIN but are significantly better than KIT.

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.

7. Conclusions

We reported results of the 2016 IWSLT Evaluation Campaign which featured two tasks: the translation of video conference conversations, a brand new task, and the translation of talks from the TED talk collection and the QED corpus. For both tasks, automatic speech recognition, machine translation, and spoken language translations tracks were organised. In total, ten international research groups joined the evaluation campaign. Performance improvements observed last year on the translation, thanks to the application of deep neural networks, were confirmed and even enhanced this year.

8. Acknowledgements

The human evaluation and part of the work by FBK's authors were supported by the CRACKER and ModernMT projects, which receive funding from the EU's Horizon 2020 research and innovation programme (grants No. 645357 and 645487).

[8] Note that since TER is an edit-distance measure, lower numbers indicate better performances.

[9] For a detailed analysis of the outputs of the systems participating in the IWSLT 2015 MT $EnDe$ task, see [35]

9. 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] 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

[6] A. Abdelali, F. Guzman, H. Sajjad, and S. Vogel, "The amara corpus: Building parallel language resources for the educational domain," 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.

[7] C. Federmann and W. D. Lewis, "Microsoft Speech Language Translation (MSLT) Corpus: The IWSLT 2016 release for English, French and German," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, US-WA, 2016.

[8] W. Michel, Z. Tüske, M. A. B. Shaik, R. Schlüter, and H. Ney, "The RWTH Aachen LVCSR system for IWSLT-2016 German Skype conversation recognition task," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[9] J.-T. Peter, A. Guta, N. Rossenbach, M. Graca, and H. Ney, "The RWTH Aachen Machine Translation System for IWSLT 2016," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[10] M. Kazi, E. Salesky, B. Thompson, J. Taylor, J. Gwinnup, T. Anderson, G. Erdmann, E. Hansen, B. Ore, K. Young, and M. Hutt, "The MITLL-AFRL IWSLT 2016 Systems," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[11] M. Junczys-Dowmunt and A. Birch, "The University of Edinburgh's systems submission to the MT task at IWSLT," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[12] F. Burlot, M. Labeau, E. Knyazeva, T. Lavergne, A. Allauzen, and F. Yvon, "LIMSI@IWSLT'16: MT Track," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[13] X. Niu and M. Carpuat, "The UMD Machine Translation Systems at IWSLT 2016: English-to-French Translation of Speech Transcripts," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[14] E. Cho, J. Niehues, T.-L. Ha, M. Sperber, M. Mediani, and A. Waibel, "Adaptation and Combination of NMT Systems: The KIT Translation Systems for IWSLT 2016," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[15] T. S. Nguyen, M. Mueller, M. Sperber, T. Zenkel, K. Kilgour, S. Stueker, and A. Waibel, "The 2016 KIT IWSLT Speech-to-Text Systems for English and German," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[16] M. A. Farajian, R. Chatterjee, C. Conforti, S. Jalalvand, V. Balaraman, M. D. Gang, D. Ataman, M. Turchi, M. Negri, and M. Federico, "FBK's Neural Machine Translation Systems for IWSLT 2016," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[17] S. Pipa, A.-F. Vasile, I. Ionascu, S. D. Dumitrescu, and T. Boroş, "RACAI Entry for the IWSLT 2016 Shared Task," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[18] O. Bojar, R. Sudarikov, T. Kocmi, J. Helcl, and O. Cífka, "UFAL Submissions to the IWSLT 2016 MT Track," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[19] N. Durrani, F. Dalvi, H. Sajjad, and S. Vogel, "QCRI's Machine Translation Systems for IWSLT'16," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[20] T. T. V. Van Huy Nguyen, Trung-Nghia Phung and C. M. Luong, "The IOIT English ASR system for IWSLT 2016," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Seattle, WA, 2016.

[21] 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.

[22] 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.

[23] H. Sajjad, F. Guzmán, P. Nakov, A. Abdelali, K. Murray, F. A. Obaidli, and S. Vogel, "QCRI at IWSLT 2013: Experiments in Arabic-English and English-Arabic Spoken Language Translation," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[24] 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

[25] 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.

[26] 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.

[27] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico, "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.

[28] ——, "Report on the 10th IWSLT Evaluation Campaign," in *Proceedings of the Tenth International Workshop on Spoken Language Translation (IWSLT 2013)*, Heidelberg, Germany, 2013.

[29] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, R. Cattoni, and M. Federico, "The IWSLT 2015 Evaluation Campaign," in *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam, 2015.

[30] 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.

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

[32] 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.

[33] 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

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

[35] L. Bentivogli, A. Bisazza, M. Cettolo, and M. Federico, "Neural versus phrase-based machine translation quality: a case study," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, November 2016, pp. 257–267. [Online]. Available: https://aclweb.org/anthology/D16-1025

[36] 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 (*tst2016*)

· All the sentence IDs in the IWSLT 2016 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 (%).

ASR: Talk English (ASR$_{EN}$)

System	WER	# Errors
MITLL-AFRL	7.2%	1,796
KIT	8.5%	2,119
IOIT	16.0%	4,000
RACAI	59.2%	14,835

ASR: QED English (ASR$_{EN}$)

System	WER	# Errors
MITLL-AFRL	10.4%	491
KIT	11.6%	545
IOIT	16.6%	780
RACAI	113.6%	5,345

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

System	WER	# Errors
MITLL-AFRL	6.4%	1,305
KIT	7.7%	1,574
IOIT	15.8%	3,220
RACAI	46.6%	9,490

ASR : MSLT English (ASR$_{EN}$)

System	WER	# Errors
KIT	22.3%	9,807
IOIT	29.5%	12,970

ASR : MSLT German (ASR$_{DE}$)

System	WER	# Errors
RWTH	19.7%	5,899
KIT	25.5%	7,671

SLT : TED English-German

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	18.11	69.29	19.05	67.12

SLT : QED English-German

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	13.57	77.78	14.85	75.65

SLT : MSLT English-German

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	21.15	67.41	22.71	65.06

SLT : MSLT German-English

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	21.20	64.24	22.24	62.40

SLT : MSLT English-French

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
RACAI	4.30	79.53	4.62	78.61

MT : TED Arabic-English

System	*case sensitive*		
	BLEU	NIST	TER
QCRI	**31.78**	**7.1876**	**49.34**
MITLL-AFRL	28.68	6.7696	53.44

MT : QED Arabic-English

System	*case sensitive*			*case insensitive*		
	BLEU	NIST	TER	BLEU	NIST	TER
QCRI	**28.09**	**5.5085**	**58.88**	**33.47**	**6.2812**	**52.48**
MITLL-AFRL	14.26	3.9917	75.77	16.84	4.4232	71.82

MT : TED English-Arabic

System	*case sensitive*		
	BLEU	NIST	TER
QCRI	18.06	5.1625	69.04

MT : QED English-Arabic

System	*case sensitive*			*case insensitive*		
	BLEU	NIST	TER	BLEU	NIST	TER
QCRI	23.13	4.9507	65.67	23.14	4.9538	65.67

MT : TED Czech-English

System	*case sensitive*		
	BLEU	NIST	TER
UFAL	30.15	7.3063	49.09

MT : QED Czech-English

System	*case sensitive*			*case insensitive*		
	BLEU	NIST	TER	BLEU	NIST	TER
UFAL	19.44	4.6596	66.29	22.16	5.1922	61.36

MT : TED English-Czech

System	*case sensitive*		
	BLEU	NIST	TER
LIMSI	**16.24**	**5.0044**	**64.66**
UFAL	12.71	4.4875	69.49

MT : QED English-Czech

System	*case sensitive*			*case insensitive*		
	BLEU	NIST	TER	BLEU	NIST	TER
LIMSI	**15.89**	**3.9547**	**75.40**	**17.98**	**4.3363**	**71.24**
UFAL	14.18	3.5939	78.93	17.63	4.0832	73.86

MT : TED French-English

System	*case sensitive*		
	BLEU	NIST	TER
UEDIN	**37.56**	**8.2806**	**40.95**
FBK	37.19	8.2385	41.14

MT : TED English-French

System	*case sensitive*		
	BLEU	NIST	TER
UEDIN	**36.88**	7.7007	46.02
FBK	36.77	**7.7475**	**45.89**
RACAI	26.91	6.6369	54.91

MT : MSLT English-French

System	*case sensitive*		
	BLEU	NIST	TER
UMD	**43.47**	8.5433	**38.04**
FBK	42.98	**8.6440**	38.20

MT : TED German-English

System	*case sensitive*		
	BLEU	NIST	TER
RWTH	**33.68**	**7.7562**	45.80
KIT	33.61	7.7304	**45.40**
UEDIN	32.56	7.5873	46.15
UFAL	30.97	7.4057	47.54
FBK	30.30	7.2259	47.65

MT : QED German-English

System	*case sensitive*			*case insensitive*		
	BLEU	NIST	TER	BLEU	NIST	TER
RWTH	**29.65**	**5.8406**	**55.59**	**35.33**	**6.6282**	**49.27**
KIT	26.47	5.3082	60.03	30.74	5.9851	54.26
UFAL	23.19	5.1916	60.19	26.93	5.8378	54.68

MT : MSLT German-English

System	*case sensitive*		
	BLEU	NIST	TER
RWTH	**40.07**	**8.1521**	**39.36**
KIT	36.55	7.7232	40.21
FBK	35.06	7.7489	41.24
UFAL	32.84	7.4284	44.33

MT : TED English-German

System	*case sensitive*		
	BLEU	NIST	TER
UEDIN	**27.34**	**6.5588**	**55.26**
KIT	26.82	6.4517	56.27
FBK	26.56	6.5499	55.51
UFAL	23.14	5.9512	60.76

MT : QED English-German

System	*case sensitive*			*case insensitive*		
	BLEU	NIST	TER	BLEU	NIST	TER
UFAL	**18.11**	**4.2771**	**72.19**	**20.45**	**4.6769**	**67.95**
KIT	17.91	4.2513	73.56	20.24	4.6584	69.36

MT : MSLT English-German

System	*case sensitive*		
	BLEU	NIST	TER
KIT	**40.17**	**8.3286**	**39.26**
FBK	38.78	8.2610	39.52
UFAL	35.57	7.7262	42.56

A.2. Progress 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* metric.
· *WER*, *BLEU* and *TER* scores are given as percent figures (%).

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

System	WER	# Errors
MITLL-AFRL	6.1%	1,119
KIT	7.3%	1,334
IOIT	12.5%	2,277

SLT : TED English-German

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	17.67	79.53	18.55	77.51

MT : TED Arabic-English

System	*case sensitive*		
	BLEU	NIST	TER
QCRI	**34.09**	**7.3943**	**46.78**
MITLL-AFRL	30.53	6.9285	50.94

MT : TED English-Arabic

System	*case sensitive*		
	BLEU	NIST	TER
QCRI	19.50	5.3894	63.22

MT : TED Czech-English

System	*case sensitive*		
	BLEU	NIST	TER
UFAL	33.22	7.5290	46.64
BEST IWSLT2015	25.07	6.4026	55.74

MT : TED English-Czech

System	*case sensitive*		
	BLEU	NIST	TER
LIMSI	**19.18**	**5.3772**	**60.81**
UFAL	15.71	4.8609	65.76
BEST IWSLT2015	17.17	5.1056	63.00

MT : TED French-English

System	*case sensitive*		
	BLEU	NIST	TER
UEDIN	**39.69**	**8.2895**	**40.38**
FBK	38.44	8.2285	40.75
BEST IWSLT2015	32.75	7.2769	48.41

MT : TED English-French

System	*case sensitive*		
	BLEU	NIST	TER
FBK	**39.71**	**7.9694**	**43.96**
UEDIN	39.14	7.8646	44.46
RACAI	29.68	6.8445	52.66
BEST IWSLT2015	32.79	7.3222	49.15

MT : TED German-English

System	*case sensitive*		
	BLEU	NIST	TER
KIT	**34.13**	7.6087	**45.83**
RWTH	33.90	**7.6416**	46.39
UEDIN	33.83	7.6218	46.05
FBK	32.38	7.4886	46.78
UFAL	31.81	7.4467	47.16
BEST IWSLT2015	31.50	7.7932	47.11

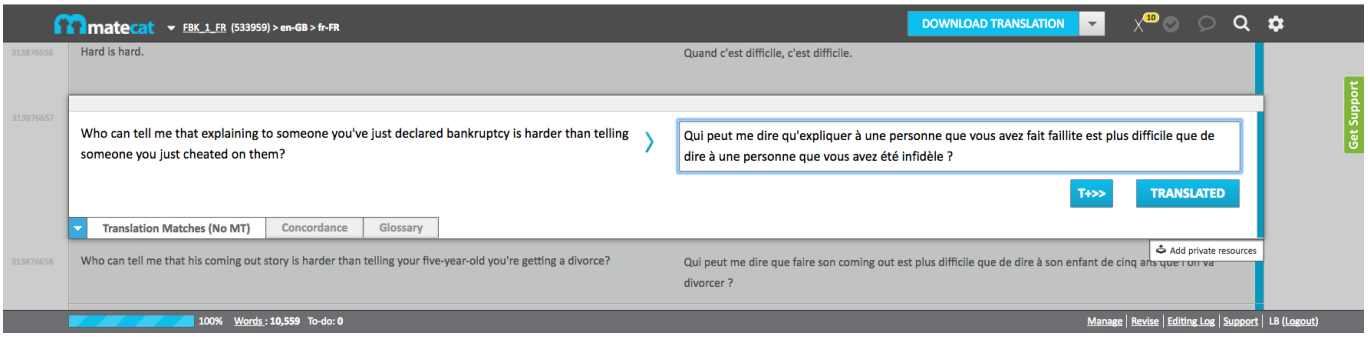
MT : TED English-German

System	*case sensitive*		
	BLEU	NIST	TER
UEDIN	**30.42**	6.8813	52.02
KIT	30.21	6.8135	52.47
FBK	30.05	**6.9381**	**51.58**
UFAL	25.63	6.2446	57.09
BEST IWSLT2015	30.85	6.9898	51.13

Appendix B. Human Evaluation

Interface used for the bilingual post-editing task

Post-editing was carried out using MateCat[10] [36], 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.

[10] www.matecat.com