Report on the 10th IWSLT Evaluation Campaign

Mauro Cettolo[(1)] Jan Niehues[(2)] Sebastian Stüker[(2)] Luisa Bentivogli[(1)] Marcello Federico[(1)]

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

Abstract

The paper overviews the tenth evaluation campaign organized by the IWSLT workshop. The 2013 evaluation offered multiple tracks on lecture transcription and translation based on the TED Talks corpus. In particular, this year IWSLT included two automatic speech recognition tracks, on English and German, three speech translation tracks, from English to French, English to German, and German to English, and three text translation track, also from English to French, English to German, and German to English. In addition to the official tracks, speech and text translation optional tracks were offered involving 12 other languages: Arabic, Spanish, Portuguese (B), Italian, Chinese, Polish, Persian, Slovenian, Turkish, Dutch, Romanian, Russian. Overall, 18 teams participated in the evaluation for a total of 217 primary runs submitted. All runs were evaluated with objective metrics on a current test set and two progress test sets, in order to compare the progresses against systems of the previous years. In addition, submissions of one of the official machine translation tracks were also evaluated with human post-editing.

1. Introduction

This paper overviews the results of the evaluation campaign organized by the International Workshop of Spoken Language Translation. The IWSLT evaluation has been now running for a decade and has offered along these years a variety of speech translation tasks [1, 2, 3, 4, 5, 6, 7, 8, 9]. The 2013 IWSLT evaluation continued along the line set in 2010, by focusing on the translation of TED Talks, a collection of public speeches covering many different topics. As in the previous two years, the evaluation included tracks for 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), that addressed the conversion and translation of a speech signal into a transcript in another language.

However, with respect to previous rounds, new languages have been added to each track. The ASR track included besides English also German, and the SLT and MT track offered English-French, English-German, and German-English translation directions. Besides the official evaluation tracks, many other optional translation directions were also offered. Optional SLT directions were from English to Spanish, Portuguese (B), Italian, Chinese, Polish, Slovenian, Arabic, and Persian. Optional MT translation directions were: English from/to Arabic, Spanish, Portuguese (B), Italian, Chinese, Polish, Persian, Slovenian, Turkish, Dutch, Romanian, and Russian. For each official and optional translation direction, training and development data were supplied by the organizers through the workshop's website. Major parallel collections made available to the participants were the WIT^3 [10] corpus of TED talks, all data from the WMT 2013 workshop (CITE), the MULTIUN corpus (CITE), and the SETimes parallel corpus (CITE). A list of monolingual resources was provided too, that includes both freely available corpora and corpora available from the LDC. Test data were released at the begin of each test period, requiring participants to return one primary run and optional contrastive runs within one week. The schedule of the evaluation was organized as follows: June 8, release of training data; Sept 2-8, ASR test of period; Sept 9-15, SLT test period; Oct 7-13, MT test period; Oct 7-20, test period of all optional directions.

All runs submitted by participants were evaluated with automatic metrics. In addition, MT runs of the English-French direction were evaluated manually. While in the past years SLT and MT outputs were evaluated through subjective rankings, this year another method was investigated. In particular, we tried to address the utility of MT output by measuring the post-editing effort needed by a professional translator to fix it.

This year, 18 participant sites registered (see Table 1) submitting a total of 217 primary runs: 28 to the ASR track, 10 to the SLT track, and 179 to the MT track (see Sections 3.3, 4.3, 5.3 for details).

In the rest of the paper we first outline the main goals of the IWSLT evaluation and then each single track in detail, in particular: its specifications, supplied language resources, evaluation methods, and results. The paper ends with some concluding remarks about the experience made in this evaluation exercise, followed by appendixes that complement the information given in the specific sections.

2. TED Talks

2.1. TED events

The translation of TED talks was introduced for the first time at IWSLT 2010. TED is a nonprofit organization that "invites the world's most fascinating thinkers and doers [...] to give the talk of their lives". Its website[1] makes the video recordings of the best TED talks available under the Creative Commons license. All talks have English captions, which have also been translated into many languages by volunteers worldwide. In addition to the official TED events held in North America, a series of independent TEDx events are regularly held around the world, which share the same format of the original TED talks but are hold in the language of the hosting country. Recently, an effort was made to set up a web repository [10] that distributes dumps of the available TED talks transcripts and translations under form of parallel texts, ready to use for training and evaluating MT systems. At this time, parallel data between English and 15 foreign languages are available in addition to evaluation sets results achieved by baseline MT systems trained for each translation direction.

Besides representing a popular benchmark for spoken language technology, the TED Talks task embeds interesting research challenges which are unique among the available speech recognition and machine translation benchmarks. TED Talks is a collection of rather short speeches (max 18 minutes each, roughly equivalent to 2,500 words) which cover a wide variety of topics. Each talk is delivered in a brilliant and original style by a very skilled speaker and, while addressing a wide audience, it pursues the goal of both entertaining and persuading the listeners on a specific idea. From the point of view of ASR, TED talks require copying with background noise – e.g. applauses and laughs by the public –, different accents including non native speakers, varying speaking rates, prosodic aspects, and, finally, narrow topics and personal language styles. From an application perspective, TED Talks transcription is the typical life captioning scenario, which requires producing polished subtitles in real-time.

From the point of view of machine translation, translating TED Talks implies dealing with spoken rather than written language, which is hence expected to be structurally less complex, formal and fluent. Moreover, as human translations of the talks are required to follow the structure and rythm of the English captions [2], a lower amount of rephrasing and reordering is expected than in ordinary translation of written documents.

From an application perspective, TED Talks suggest translation tasks ranging from off-line translation of written captions, up to on-line speech translation, requiring a tight integration of MT with ASR possibly handling stream-based processing.

[1] http://www.ted.com

[2] See recommendations to translators in http://translations.ted.org/wiki.

3. ASR Track

3.1. Definition

The goal of the *Automatic Speech Recognition* (ASR) track for IWSLT 2013 was to transcribe English TED talks and German TEDx talks. The speech in TED lectures is in general planned, well articulated, and recorded in high quality. 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. For the German TEDx talks the recording conditions are a little bit more difficult than for the English TED talks. While the TEDx talks aim to mimic the TED talks, they are not as well prepared and well rehearsed as the TED lectures, and recording is often done by amateurs resulting in often worse recording quality than the TED lectures.

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 for the spoken language translation evaluation (SLT), see Section 4.

3.2. Evaluation

Participants had to submit the results of the recognition of the tst2013 set in CTM format. The word error rate was measured case-insensitive. After the end of the evaluation a first 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 score were calculated.

In order to measure the progress of the systems over the years on English, participants also had to provide results on the test sets from 2011 and 2012, i.e. tst2011 and tst2012.

3.3. Submissions

For this year's evaluation we received primary submissions from eight sites: all of which participated in the English ASR task and four also in the German ASR task. For English we further received a total of nine contrastive submissions from six sites. For German we received eight contrastive submissions from three 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.1. The word error rate of the submitted systems in in the range of 13.5%-27.2% for English and 25.2%-37.8% for German.

In German, the fact that TEDx have sometimes worse recording conditions than TED talks was reflected by the fact that one talk in the German tst2013 had WERs above 80%, due to a bad recording set-up with high noise. All other WERs were mostly below 30% and 20%, for two talks even below 10%.

Table 1: List of Participants

NTT-NAIST	NTT Communication Science Labs, Japan & NAIST[11]
KIT	Karlsruhe Institute of Technology, Germany [12, 13]
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen, Germany [14, 15]
EU-BRIDGE	RWTH& UEDIN& KIT& FBK[16]
HDU	Dept. of Computational Linguistics, Heidelberg University, Germany [17]
UEDIN	University of Edinburgh, UK [18, 19, 20]
FBK	Fondazione Bruno Kessler, Italy [21, 22]
PRKE-IOIT	Inst. of Inform. and Techn., Vietnamese Academy of Science and Technology [23]
POSTECH	Pohang University of Science and Technology, Korea [24]
MITLL-AFRL	Mass. Institute of Technology/Air Force Research Lab., USA [25]
QCRI	Qatar Computing Research Institute, Qatar Foundation, Qatar [26]
MSR-FBK	Microsoft Corporation, USA, and FBK[27]
HKUST	Hong Kong University of Science and Technology, Hong Kong [28]
NICT	National Institute of Communications Technology, Japan [29, 30]
NAIST	Nara Institute of Science and Technology, Japan [31]
PJIIT	Polish-Japanese Institute of Information Technology, Poland [32]
CASIA	Institute of Automation, Chinese Academy of Sciences, China [33]
TUBITAK	TUBITAK - Center of Research for Advanced Technologies, Turkey

For English, it can be seen that all participants from IWSLT2011 and IWSLT2012 made significant progresses over the years, e.g., bringing down the WER from 13.5% to 7.9% on tst2011, a relative reduction by 41% over the course of three years.

4. SLT Track

4.1. Definition

The SLT track required participants to translate the English and German talks of tst2013 from the 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.

For German, participants had to translate into English. For English as source language, participants had to translate into French and German. In addition, participants could also optionally translate from English into one of the following languages: Arabic, Spanish, Farsi, Italian, Polish, Brazilian Portuguese, Slovenian, and Mandarin Chinese.

4.2. Evaluation

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

ence organizers. In order to facilitate scoring, participants had to to segment the audio according to the manual reference segmentation provided by the organizers of the evaluation.

For English, the ASR output provided by the organizers was a ROVER combination of the output from five submissions to the ASR track. The result of the ROVER had a WER of 12.4%. For German we used the output from KIT, as ROVER combination with other systems did not give any performance gains, and the German KIT ASR system scored best before the end of the adjudication.

The results of the translation had to be submitt ed in the same format as for the machine translation track (see Section 5).

4.3. Submissions

We received ten primary and nine contrastive submissions from five participants, English to French receiving the most submissions. In English to Arabic and English to Chinese only one participant each submitted results.

4.4. Results

The detailed results of the automatic evaluation in terms of BLEU and TER can be found in Appendix A.1. Appendix A.2 contains the results of the progress test set for English to French.

5. MT Track

5.1. 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

Table 2: Monolingual resources for official language pairs

data set	lang	sent	token	voc
train	De	146k	2.66M	107.4k
	En	159k	3.20M	58.3k
	Fr	158k	3.36M	70.7k

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 official and optional translation direction, in-domain training and development data were supplied through the website of the WIT[3] [10], while out-of-domain training data through the workshop's website. With respect to edition 2012 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 (tst2013), while the remaining talks have been included in the training sets. For reliably assessing progress of MT systems over the years, the evaluation sets of editions 2011 and 2012 were distributed together with tst2013 as progressive test sets, when available. Development sets (dev2010 and tst2010) are either the same of past editions or have been built upon the same talks.

With respect to all the other directions, the $DeEn$ MT task is an exception; in fact, its dev2012 and tst2013 - development and evaluation sets, respectively - derives from those prepared for the ASR/SLT tracks, which consist of TEDX talks delivered in German language; therefore, no overlap exists with any other TED talk involved in other tasks. Anyway, the standard dev2010 and tst2010 development sets have been released as well.

Tables 2 and 3 provides statistics on in-domain texts supplied for training, development and evaluation purposes for the official directions.

Reference results from baseline MT systems on the development set tst2010 are provided via the WIT3 repository. This helps participants and MT scientists to assess their experimental outcomes.

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 [34] was applied to all languages, with the exception of Chinese and Arabic languages, which were preprocessed by, respectively: the Stanford Chinese Segmenter [35]; either AMIRA [36], in the Arabic-to-English direction, or the QCRI-normalizer,[3] in the English-to-Arabic direction.

The baselines were developed with the Moses toolkit. Translation and lexicalized reordering models were trained on the parallel training data; 5-gram LMs with improved Kneser-Ney smoothing were estimated on the target side of the training parallel data with the IRSTLM toolkit. The weights of the log-linear interpolation model were optimized on dev2010 with the MERT procedure provided with Moses.

Table 3: Bilingual resources for official language pairs

task	data set	sent	tokens source	tokens target	talks
MT$_{EnFr}$	train	154k	3.06M	3.27M	1169
	dev2010	887	20,1k	20,2k	8
	tst2010	1,664	32,0k	33,9k	11
	tst2011	818	14,5k	15,6k	8
	tst2012	1,124	21,5k	23,5k	11
	tst2013	1,026	21,7k	23,3k	16
MT$_{DeEn}$	train	139k	2.59M	2.75M	1064
	dev2010	887	19,1k	20,1k	8
	tst2010	1,565	30,3k	32,0k	11
	dev2012	1,165	20,8k	21,6k	7
	tst2013	1,369	22,4k	22,8k	9
MT$_{EnDe}$	train	139k	2.75M	2.59M	1064
	dev2010	887	20,1k	19,1k	8
	tst2010	1,565	32,0k	30,3k	11
	tst2011	1,436	27,1k	26,4k	16
	tst2012	1,704	30,8k	29,3k	15
	tst2013	993	20,9k	19.7k	16

5.2. Evaluation

The participants to the MT track had to provide the results of the translation of the test sets in NIST XML format. The output had to be true-cased and had to contain punctuation.

The quality of the translations was measured automatically against the human translations created by the TED open translation project, and by human subjective evaluation (Section 5.5).

The evaluation specifications for the MT track were defined as case-sensitive with punctuation marks (case+punc). Tokenization scripts were applied automatically to all run submissions prior to evaluation.

Evaluation scores were calculated for the two automatic standard metrics BLEU and TER, as implemented in mteval-v13a.pl[4] and tercom-0.7.25[5], respectively.

5.3. Submissions

We received 68 submissions from 15 different sites, distributed as follows: 20 for the three official language pairs, 48 on optional directions.

The pairs that attracted the most interest are the official pairs – seven each for EnFr and DeEn, six for EnDe – and those involving Chinese (a total of nine in the two directions), Arabic (seven), Farsi (five) and Russian (five). Each pair received at least one submission.

The total number of primary runs, on evaluation set tst2013 and on progressive test sets tst2011 and tst2012, is

[3] Specifically developed for IWSLT 2013 by P. Nakov and F. Al-Obaidli at Qatar Computing Research Institute.

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

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

179; in addition, we were asked to evaluate also 156 contrastive runs.

5.4. Results

Table 4: BLEU and TER scores of baseline SMT systems on tst2013 for all language pairs. (⋆) Char-level scores.

pair		direction → BLEU	direction → TER	direction ← BLEU	direction ← TER
En	Fr	31.94	48.59	–	–
	De	19.58	59.81	19.07	65.94
	Ar	12.12	68.73	22.71	59.02
	Es	29.01	50.99	33.18	45.58
	Fa	8.94	72.74	12.17	88.88
	It	26.59	52.75	30.82	50.35
	Nl	22.82	57.66	28.00	54.49
	Pl	10.31	76.16	16.31	67.33
	Pt	29.65	46.85	35.80	42.93
	Ro	16.18	68.29	24.85	54.21
	Ru	13.69	71.30	18.57	64.99
	Sl	9.49	72.16	14.62	69.70
	Tr	6.62	79.96	12.24	75.90
	Zh	⋆18.15	⋆72.34	12.29	70.60

First of all, for reference purposes Table 4 shows BLEU and TER scores on the tst2013 evaluation sets of the baseline systems we developed as described in Section 5.1.

The results on the official test set for each participant are shown in Appendix A.1. For most languages, we show the case-sensitive and case-insensitive BLEU and TER scores. In contrast to the other language pairs, in the German to English translation task the source contained disfluencies. Therefore, the translation are evaluated once against translation containing disfluencies and once against reference containing no disfluencies. Furthermore, for English to Chinese we report character-level and word-level scores.

These results also show again the scores of the baseline system. Thereby, it is possible to see the improvements of the submitted systems on the different languages over the baseline system. The largest improvements could be gained on Slovenian-English by 9.44 BLEU points.

In Appendix A.2 the results on the progress test sets test2011 and test2012 are shown. When comparing the results to the submissions from last year, the performance could be improved in nearly all tasks.

5.5. Human Evaluation

Human evaluation was carried out on all primary runs submitted by participants to one of the official tracks of the TED task, namely the *official* MT English-French track.

This year's human evaluation saw the introduction of a major novelty. In fact, the traditional *Relative Ranking* task was substituted by a *Post-Editing* task and, accordingly, HTER (Human-mediated Translation Edit Rate) was adopted as the official evaluation metrics to rank the systems.

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 functionalities, and a number of studies [37, 38] demonstrate the usefulness of MT to increase professional 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 was to adopt a human evaluation framework able to maximize the benefit to the research community, both in terms of information about MT systems and data and resources to be reused. With respect to traditional 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. Furthermore, HTER[39] - which consists of measuring the minimum edit distance between the machine translation and its manually post-edited version - has been shown to correlate quite well with human judgments of MT quality.

The human evaluation setup and the collection of post-editing data are presented in Section 5.5.1, whereas the results of the evaluation are presented in Section 5.5.2.

5.5.1. Evaluation Setup and Data Collection

All 2013 systems participating in the English-French MT track were manually evaluated on a subset of the 2012 progress test set *(tst2012)*[6]. The Human Evaluation (HE) set represents around the initial 50% of each of the 11 *tst2012* talks, for a total of 580 segments and around 10,000 words. 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.

In order to evaluate the MT systems, the *bilingual* post-editing task was chosen, where professional translators are required to post-edit the MT output directly according to the source sentence. Bilingual post-editing is expected to give more accurate results than monolingual post-editing as post-editors do not depend on an given - and possibly imprecise - translation.

As far as evaluation metrics are concerned, HTER [39] is a semi-automatic metric derived from TER (Translation Edit Rate). TER measures the amount of editing that a human would have to perform to change a machine translation so that it exactly matches a given reference translation. HTER

[6]Since all the data produced for human evaluation will be made publicly available thorough the WIT^3 repository, we used the 2012 test set in order to keep the 2013 test set blind to be used as a progress test for next year's evaluation.

is a variant of TER where a new reference translation is generated by applying the minimum number of post-edits to the given MT output. This new *targeted* reference is then used as the only reference translation to calculate the MT output TER.

In the preparation of the data to be collected, some constraints were identified to ensure the soundness of the evaluation of the seven systems participating in the task: *(i)* each translator must post-edit all segments of the HE set, *(ii)* each translator must post-edit the segments of the HE set only once, and *(iii)* each MT system must be equally post-edited by all translators.

Given that we had seven systems to evaluate, in order to satisfy the above constraints we resorted to seven professional translators. Moreover, in order to cope with variability of post-editors (i.e. some translators could systematically post-edit more than others) we devised a scheme that dispatches MT outputs to translators both randomly and satisfying the uniform assignment constraints. Seven documents were hence prepared including all source segments of the HE set and, for each source segment, one MT output selected from one of the seven systems.

Documents were delivered to a language service provider together with instructions to be passed on to the translators, and the post-editing tasks were run using the tool developed under the MateCat project[7], an enterprise-level CAT tool. Both the post-editing interface and the guidelines given to translators are presented in Appendix B.

The resulting collected data consist of seven new reference translations for each of the 580 sentences of the HE set. Each one of these seven references represents the targeted translation of the system output from which it was derived. From the point of view of the system output, one targeted translation and other six untargeted translations are available.

Table 5 shows information about the characteristics of the work carried out by post-editors. First, 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 system sentences post-edited by each single translator. As we can see from the table, PE effort is highly variable among post-editors, ranging from 19.51% to 42.60%. Data about standard deviation confirm post-editor variability, showing that the seven translators produced quite different post-editing effort distributions.

To further study post-editor variability, we exploited the official reference translations available for this TED track and we calculated the TER of the outputs assigned to each translator for post-editing (*Sys TER* Column in Table 5), as well as the related standard deviation.

As we can see from the table, 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.

[7] www.matecat.com

Table 5: Post-editing information for each Post-editor

PEditor	PE Effort	*std-dev*	Sys TER	*std-dev*
PE 1	24.93	17.74	40.27	20.32
PE 2	34.03	19.86	39.48	19.89
PE 3	42.60	22.47	40.61	20.19
PE 4	32.78	21.07	39.98	20.97
PE 5	19.51	15.55	40.82	20.95
PE 6	30.64	19.48	40.42	20.70
PE 7	34.60	23.92	39.39	20.62

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. Thus, post-editor variability ai an issue to be addressed to ensure a sound evaluation of the systems.

5.5.2. Evaluation Results

As seen in the previous section, being able to reduce post-editors' variability would allow a more reliable and consistent evaluation of MT systems. To this purpose, the HTER for each system submission was calculated under two different settings, namely *(i)* using the targeted reference only (*Tgt PEref* setting), and *(ii)* using all the seven references produced by all the post-editors for each sentence (*All PErefs* setting).

The scores resulting from the application of the two HTER settings are shown in Table 6, which also presents a comparison of HTER scores and rankings with those obtained using the related automatic metrics TER[8].

Table 6: Official human evaluation results and comparisons with other metrics

System Ranking	**HTER** ***HE Set all PErefs***	**HTER** HE Set Tgt PEref	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**18.67**	29.83	38.71	38.72
KIT	**20.01**	*29.64*	39.20	39.22
UEDIN	**20.69**	31.61	39.81	39.83
RWTH	**21.06**	31.64	*39.70*	39.95
FBK	**21.41**	32.29	40.38	40.56
MITLL-AFRL	**22.24**	32.31	41.37	41.47
PRKE-IOIT	**22.26**	*32.01*	41.81	41.52
Rank Corr.		.857	.964	1.00

As shown in the table, the HTER reduction obtained in the *All PErefs* setting (Column 2) with respect to the *Tgt PEref* setting (Column 3) clearly shows that exploiting all the available reference translations is a viable way to control and overcome post-editors' variability, obtaining an HTER

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

which is more informative about the real performances of the systems. This is also confirmed by the range of standard deviations observed for the scores of the systems, which for *Tgt PEref* ranges from 20.57 to 23.18, while for *All PEref* ranges from 12.84 to 14.31.

For this reason, the scores and overall ranking of the systems as resulting in the *All PErefs* setting have been chosen as the official results of human evaluation.

In general, the very low HTER results obtained demonstrate that the overall quality of the systems is very high. Moreover, all systems are very close to each other. 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 [40], a statistical test well-established in the NLP community [41] and that, especially for the purpose of MT evaluation, has been shown [42] to be less prone to type-I errors than the bootstrap method [43]. According to the approximate randomization test based on 10,000 iterations, a winning system cannot be indicated, as there is no system that is significantly better than all other systems. Significant differences can be found only between the top-scoring system (EU-BRIDGE) and the three bottom-scoring ones. In particular, significance with respect to FBK is at $p \leq 0.1$, while significance with respect to MITLL-AFRL and PRKE-IOIT is at $p \leq 0.05$.

A number of additional observations can be drawn by comparing the official results with results obtained with other metrics (Columns 3,4,5 in Table 6).

In general, HTER reduces the edit rate with respect to TER. More specifically, we can see a reduction of around 25% for HTER calculated with only one targeted reference(*Tgt PEref* setting), and of around 50% for HTER calculated with all post-edited references (*All PErefs* setting).

Moreover, the correlation between evaluation metrics is measured using *Spearman's rank correlation coefficient* $\rho \in$ [-1.0, 1.0], with ρ = 1.0 if all systems are ranked in same order, ρ = -1.0 if all systems ranked in reverse order and ρ = 0.0 if no correlation exists. We can see from Table 6 that completely automatic metrics (TER) correlate well with the official HTER. In particular, TER calculated on the whole 2012 test set correlates perfectly, confirming that automatic metrics are more reliable when the quantity of evaluation data increases.

To conclude, the post-editing task introduced this year for manual evaluation brought benefit to the IWSLT community, and in general to the MT field. In fact, producing post-edited versions of all the participating systems' outputs allowed us to carry out a quite informative evaluation by minimizing the variability of post-editors, who naturally tend to diverge from the post-editing guidelines and personalize their translations. Moreover, a number of additional reference translations will be available for further development and evaluation of MT systems.

6. Conclusions

We have reported on the evaluation campaign organized for the tenth edition of the IWSLT workshop. The evaluation has addressed three tracks: automatic speech recognition of talks (in English and German), speech-to-text translation, and text-to-text translation, both from German to English, English to German, and English to French. Besides the official translation directions, many optional translation tasks were available, too, including 12 additional languages. For each task, systems had to submit runs on three different test sets: a newly created official test set, and two progress test sets created and used for the 2012 and 2011 evaluations, respectively. This year, 18 participants took part in the evaluation, submitting a total of 217 primary runs, which were all scored with automatic metrics. We also manually evaluated runs of the English-French text translation track. In particular, we asked professional translators to post-edit all system outputs on a subset of the 2012 progress test set, in order to produce *close references* for them. While we have observed a significant variability among translators, in terms of post-edit effort, we could obtain more reliable scores by using all the produced post-edits as reference translations. By using the HTER metric, the post-edit effort of the best performing system results remarkably low, namely less than 19%. Considering that this is still an upper bound of the ideal HTER score, this percentage of post-editing seems to be another strong argument supporting the utility of machine translation for human translators.

7. Acknowledgements

Research Group 3-01' received financial support by the '*Concept for the Future*' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative. The work leading to these results has received funding from the European Union under grant agreement no 287658 — Bridges Across the Language Divide (EU-BRIDGE).

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 evalu-

ation 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, 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

[11] K. Sudoh, G. Neubig, K. Duh, and H. Tsukada, "NTT-NAIST SMT Systems for IWSLT 2013," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[12] K. Kilgour, C. Mohr, M. Heck, Q. B. Nguyen, V. H. Nguyen, E. Shin, I. Tseyzer, J. Gehring, M. M̈ller, M. Sperber, S. Stüker, and A. Waibel, "The 2013 KIT IWSLT Speech-to-Text Systems for German and English," in *Proceedings of the tenth International Workshop on Spoken Language Trans lation (IWSLT)*, Heidelberg, Germany, 2013.

[13] T.-L. Ha, T. Herrmann, J. Niehues, M. Mediani, E. Cho, Y. Zhang, I. Slawik, and A. Waibel, "The KIT Translation Systems for IWSLT 2013," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[14] J. Wuebker, S. Peitz, T. Alkhouli, J.-T. Peter, M. Feng, M. Freitag, and H. Ney, "The RWTH Aachen Machine Translation Systems for IWSLT 2013," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[15] M. A. B. Shaik1, Z. Tüske, S. Wiesler, M. Nušbaum-Thom, S. Peitz, R. Schl̈er, and H. Ney, "The rwth aachen german and english lvcsr systems for iwslt-2013," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[16] M. Freitag, S. Peitz, J. Wuebker, H. Ney, N. Durrani, M. Huck, P. Koehn, T.-L. Ha, J. Niehues, M. Mediani, T. Herrmann, A. Waibel, N. Bertoldi, M. Cettolo, and M. Federico, "EU-BRIDGE MT: Text Translation of Talks in the EU-BRIDGE Project," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[17] P. Simianer, L. Jehl, and S. Riezler, "The Heidelberg University Machine Translation Systems for IWSLT2013," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[18] P. Bell, F. McInnes, S. R. Gangireddy, M. Sinclair, A. Birch, and S. Renals, "The UEDIN English ASR System for the IWSLT 2013 Evaluation," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[19] J. Driesen, P. Bell, M. Sinclair, and S. Renals, "Description of the uedin system for german asr," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[20] A. Birch, N. Durrani, and P. Koehn, "Edinburgh SLT and MT System Description for the IWSLT 2013 Evaluation," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[21] D. Falavigna, R. Gretter, F. Brugnara, D. Giuliani, and R. H. Serizel, "FBK @ IWSLT 2013 - ASR tracks," in *Proceedings of the tenth International Workshop on Spoken Language Trans lation (IWSLT)*, Heidelberg, Germany, 2013.

[22] N. Bertoldi, M. A. Farajian, P. Mathur, N. Ruiz, and M. Federico, "FBK's Machine Translation Systems for the IWSLT 2013 Evaluation Campaign," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[23] N.-Q. Pham, H.-S. Le, T.-T. Vu, and C.-M. Luong, "The Speech Recognition and Machine Translation System of IOIT for IWSLT 2013," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[24] H. Na and J.-H. Lee, "A Discriminative Reordering Parser for IWSLT 2013," in *Proceedings of the tenth International Workshop on Spoken Language Trans lation (IWSLT)*, Heidelberg, Germany, 2013.

[25] M. Kazi, M. Coury, E. Salesky, J. Ray, W. Shen, T. Gleason, T. Anderson, G. Erdmann, L. Schwartz, B. Ore, R. Slyh, J. Gwinnup, K. Young, and M. Hutt, "The MIT-LL/AFRL IWSLT-2013 MT Systems," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[26] 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.

[27] A. Aue, Q. Gao, H. Hassan, X. He, G. Li, N. Ruiz, and F. Seide, "MSR-FBK IWSLT 2013 SLT System Description," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[28] C. kiu Lo, M. Beloucif, and D. Wu, "Improving machine translation into Chinese by tuning against Chinese MEANT," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[29] C.-L. Huang, P. R. Dixon, S. Matsuda, Y. Wu, X. Lu, M. Saiko, and C. Hori, "The nict asr system for iwslt 2013," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[30] A. Finch, O. Htun, and E. Sumita, "The NICT Translation System for IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[31] S. Sakti, K. Kubo, G. Neubig, T. Toda, and S. Nakamura, "The naist english speech recognition system for iwslt 2013," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[32] K. Wolk and K. Marasek, "Polish - englishspeechstatistical machine translationsystems for the iwslt 2013," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[33] X. Peng, X. Fu, W. Wei, Z. Chen, W. Chen, and B. Xu, "The casia machine translation system for iwslt 2013," in *Proceedings of the tenth International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany, 2013.

[34] 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.

[35] 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.

[36] M. Diab, K. Hacioglu, and D. Jurafsky, "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," in *HLT-NAACL 2004: Short Papers*, D. M. Susan Dumais and S. Roukos, Eds. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2 - May 7 2004, pp. 149–152.

[37] 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

[38] 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.

[39] 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.

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

[41] 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.

[42] 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

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

Appendix A. Automatic Evaluation

"case+punc" evaluation : case-sensitive, with punctuations tokenized
"no_case+no_punc" evaluation : case-insensitive, with punctuations removed

A.1. Official Testset (*tst2013*)

· All the sentence IDs in the IWSLT 2012 testset were used to calculate the automatic scores for each run submission.
· ASR and MT systems are ordered according to the *WER* and *BLEU* metrics, respectively.
· All automatic evaluation metric scores are given as percent figures (%).

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

System	WER	(# Errors)
NICT	**13.5**	**(5,734)**
KIT	14.4	(6,115)
MITLL-AFRL	15.9	(6,788)
RWTH	16.0	(6,827)
NAIST	16.2	(6,897)
UEDIN	22.1	(9,413)
FBK	23.2	(9,899)
PRKE-IOIT	27.2	(11,578)

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

System	WER	(# Errors)
RWTH	**25.2**	**(4,845)**
KIT	25.7	(4,932)
FBK	37.5	(7,199)
UEDIN	37.8	(7,250)

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**26.81**	**55.08**	**27.53**	**54.06**
RWTH	25.62	57.21	26.41	56.09
UEDIN	22.45	61.34	23.30	60.06
MSR-FBK	22.42	63.69	23.72	62.20

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**18.05**	**64.46**	**18.66**	**63.22**
RWTH	17.27	66.33	17.88	65.09

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

System	*Ref. with disfluencies*				Ref. without disfluencies			
	case sensitive		case insensitive		*case sensitive*		case insensitive	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
KIT	**19.34**	**62.27**	**19.80**	**61.34**	**19.54**	**62.74**	**20.01**	**61.80**
UEDIN	14.92	68.12	15.39	67.28	15.03	68.70	15.52	67.86

TED : SLT English-Arabic (SLT$_{EnAr}$)

System	BLEU	TER
QCRI	**10.33**	**73.72**

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

System	*character-based*		*word-based*	
	BLEU	TER	BLEU	TER
KIT	**16.91**	**74.07**	**9.20**	**80.63**

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
EU-BRIDGE	**38.86**	**42.96**	**39.74**	**42.02**
KIT	38.63	43.20	39.60	42.11
UEDIN	38.45	43.96	39.39	42.91
FBK	37.69	44.13	38.46	43.23
RWTH	37.67	44.00	38.49	43.04
PRKE-IOIT	37.59	45.07	38.39	44.15
MITLL-AFRL	37.05	45.36	38.27	44.10
BASELINE	31.94	48.59	32.56	47.75

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
KIT	**25.71**	**54.46**	**26.47**	**53.34**
RWTH	24.74	55.52	25.41	54.42
NTT-NAIST	24.60	54.86	25.79	53.37
UEDIN	24.00	55.94	24.68	54.87
POSTECH	22.43	57.57	23.00	56.58
BASELINE	19.58	59.81	20.14	58.84

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

System	*Ref. with disfluencies* *case sensitive* BLEU	*Ref. with disfluencies* *case sensitive* TER	*Ref. with disfluencies* case insensitive BLEU	*Ref. with disfluencies* case insensitive TER	Ref. without disfluencies *case sensitive* BLEU	Ref. without disfluencies *case sensitive* TER	Ref. without disfluencies case insensitive BLEU	Ref. without disfluencies case insensitive TER
KIT	**26.48**	57.52	**27.11**	56.60	**26.57**	58.31	**27.16**	57.41
EU-BRIDGE	26.33	**56.70**	26.91	**55.78**	**26.57**	**57.29**	27.14	**56.38**
NTT-NAIST	25.69	60.96	26.29	60.06	25.83	60.75	26.45	59.82
UEDIN	25.54	59.99	26.12	59.07	25.35	60.98	25.87	60.08
RWTH	25.32	59.67	25.94	58.67	25.27	60.46	25.86	59.51
HDU	22.91	59.65	23.94	58.35	23.06	60.38	24.07	59.11
POSTECH	21.26	67.61	21.74	66.72	21.17	68.91	21.65	68.04
BASELINE	19.25	65.03	19.79	64.19	19.07	65.94	19.55	65.11

TED : MT English-Arabic (MT$_{EnAr}$)

System	BLEU	TER
QCRI	**15.78**	**65.43**
KIT	15.51	65.64
BASELINE	12.12	68.73
UEDIN	11.49	70.58

TED : MT Arabic-English (MT$_{ArEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
QCRI	**30.49**	51.37	**31.21**	50.37
RWTH	29.95	**50.61**	31.07	**49.44**
MITLL-AFRL	26.64	55.17	27.54	54.05
UEDIN	26.29	56.69	26.92	55.70
BASELINE	22.71	59.02	23.52	57.94

TED : MT English-Spanish (MT$_{EnEs}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**34.74**	**45.75**	**35.42**	**44.79**
BASELINE	29.01	50.99	29.57	50.08

TED : MT Spanish-English (MT$_{EsEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**39.12**	**41.36**	**39.74**	**40.59**
BASELINE	33.18	45.58	33.68	45.00

TED : MT English-Farsi (MT$_{EnFa}$)

System	BLEU	TER
FBK	**10.12**	**71.58**
UEDIN	9.49	72.92
BASELINE	8.94	72.74

TED : MT Farsi-English (MT$_{FaEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
MITLL-AFRL	**16.03**	**78.82**	**16.51**	**77.84**
UEDIN	15.10	88.06	15.42	87.20
FBK	14.47	85.84	14.86	84.87
BASELINE	12.17	88.88	12.56	87.84

TED : MT English-Italian (MT$_{EnIt}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**29.17**	**50.84**	**29.90**	**49.87**
BASELINE	26.59	52.75	27.16	51.88

TED : MT Italian-English (MT$_{ItEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**34.89**	**47.50**	**35.55**	**46.64**
BASELINE	30.82	50.35	31.30	49.63

TED : MT English-Dutch (MT$_{EnNl}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**25.52**	**55.92**	**26.49**	**54.31**
BASELINE	22.82	57.66	23.54	56.33

TED : MT Dutch-English (MT$_{NlEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**32.73**	**51.32**	**33.74**	**49.93**
BASELINE	28.00	54.49	28.94	53.08

TED : MT English-Polish (MT$_{EnPl}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
PJIIT	**14.29**	**73.54**	**15.04**	**72.06**
UEDIN	11.51	77.66	12.03	76.48
BASELINE	10.31	76.19	10.79	75.05

TED : MT Polish-English (MT$_{PlEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
PJIIT	**22.60**	**62.56**	**23.54**	**61.12**
UEDIN	20.91	64.32	21.59	63.11
BASELINE	16.31	67.33	16.85	66.26

TED : MT English-Portuguese (MT$_{EnPt}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	**33.18**	**44.92**	**33.92**	**43.90**
BASELINE	29.65	46.85	30.18	46.06

TED : MT English-Romanian (MT$_{EnRo}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	**17.57**	**66.96**	**18.10**	**65.83**
BASELINE	16.18	68.29	16.70	67.16

TED : MT English-Russian(MT$_{EnRu}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	**16.14**	**70.28**	**16.15**	**69.12**
HDU	15.87	69.00	15.95	67.63
BASELINE	13.69	71.30	13.69	70.22

TED : MT English-Slovenian(MT$_{EnSl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	**13.68**	**67.68**	**14.21**	**66.55**
RWTH	10.10	71.66	10.47	70.71
BASELINE	9.49	72.16	9.87	71.19

TED : MT English-Trukish(MT$_{EnTr}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
TUBITAK	**8.97**	**76.12**	**9.78**	**74.42**
UEDIN	6.76	82.32	7.24	81.09
BASELINE	6.62	79.96	6.94	78.80

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

System	*character-based*		*word-based*	
	BLEU	TER	BLEU	TER
CASIA	**20.55**	**65.12**	**12.45**	**72.21**
KIT	19.83	69.75	11.47	76.72
HKUST	18.66	70.36	10.85	78.12
UEDIN	18.57	69.71	10.56	77.90
BASELINE	18.15	72.34	10.01	81.77

TED : MT Portuguese-English (MT$_{PtEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	**37.33**	**42.91**	**37.80**	**42.31**
BASELINE	35.80	42.93	36.14	42.44

TED : MT Romanian-English (MT$_{RoEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	**29.82**	**50.53**	**30.58**	**49.55**
BASELINE	24.85	54.21	25.46	53.23

TED : MT Russian-English (MT$_{RuEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
HDU	**23.78**	**59.51**	**25.00**	**58.04**
UEDIN	22.67	61.99	23.37	60.93
MITLL-AFRL	21.65	60.71	22.59	59.38
BASELINE	18.57	64.99	19.12	63.90

TED : MT Slovenian-English (MT$_{SlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	**24.06**	**58.40**	**24.87**	**57.08**
RWTH	17.46	64.42	18.00	63.30
BASELINE	14.62	69.70	15.16	68.66

TED : MT Turkish-English (MT$_{TrEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
TUBITAK	**18.67**	**68.28**	**19.68**	**66.73**
UEDIN	14.87	74.19	15.63	72.85
BASELINE	12.24	75.90	12.89	74.79

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
RWTH	**16.17**	**65.37**	**17.00**	**64.17**
UEDIN	15.26	69.73	15.91	68.61
MITLL-AFRL	14.85	68.99	15.53	67.85
CASIA	14.55	69.08	15.52	67.37
BASELINE	12.29	70.60	12.85	69.56
HKUST	9.58	74.82	10.17	73.75

A.2. Progress Testset (*tst2011*) and (*tst2012*)

· All the sentence IDs in the IWSLT 2011 testset were used to calculate the automatic scores for each run submission.
· ASR and MT systems are ordered according to the *WER* and *BLEU* metrics, respectively.
· For each task, the best score of each metric is marked with **boldface**.
· All automatic evaluation metric scores are given as percent figures (%).

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

tst2011

System	IWSLT 2011 WER	(# Errors)	IWSLT 2012 WER	(# Errors)	IWSLT 2013 WER	(# Errors)
FBK	16.2	(2,091)	15.4	(1,991)	13.6	(1,754)
KIT	15.0	(1,938)	12.0	(1,552)	9.3	(1,196)
MITLL-AFRL	13.5	(1,741)	11.1	(1,432)	10.6	(1,360)
NAIST	—		12.0	(1,553)	9.1	(1,172)
NICT	25.6	(3,301)	10.9	(1,401)	7.9	(1,016)
PRKE-IOIT	—		—		14.6	(1,883)
RWTH	—		13.4	(1,731)	10.2	(1,319)
UEDIN	—		—		10.2	(1,318)

tst2012

System	IWSLT 2012 WER	(# Errors)	IWSLT 2013 WER	(# Errors)
FBK	16.8	(3,227)	16.2	(3,090)
KIT	12.7	(2,435)	9.6	(1,834)
MITLL-AFRL	13.3	(2,565)	11.3	(1,360)
NAIST	12.4	(2,392)	10.0	(1,913)
NICT	12.1	(2,318)	8.6	(1,636)
PRKE-IOIT	—		16.2	(3,101)
RWTH	13.6	(2,621)	11.3	(2,166)
UEDIN	14.4	(2775)	11.6	(2,212)

TED : SLT English-French test 2012(SLT$_{EnFr}$)

System	*case sensitive* BLEU	TER	case insensitive BLEU	TER
KIT	32.21	48.58	32.86	47.65
MSR-FBK	29.92	53.30	31.03	52.10

TED : SLT English-French test2011(SLT$_{EnFr}$)

System	*case sensitive* BLEU	TER	case insensitive BLEU	TER
KIT	31.06	50.70	31.93	49.61
MSR-FBK	27.21	56.22	28.32	54.82

TED : MT English-French test 2012(MT$_{EnFr}$)

System	*case sensitive* BLEU	TER	*case insensitive* BLEU	TER
EU-BRIDGE	42.13	38.72	42.99	37.83
UEDIN	41.21	39.83	42.02	38.94
KIT	41.02	39.22	41.96	38.34
RWTH	40.06	39.95	40.79	39.11
PRKE-IOIT	39.94	41.52	40.64	40.75
MITLL-AFRL	39.76	41.47	40.97	40.31
FBK	39.51	40.56	40.11	39.80

TED : MT English-French test 2011(MT$_{EnFr}$)

System	*case sensitive* BLEU	TER	*case insensitive* BLEU	TER
EU-BRIDGE	40.71	40.56	41.55	39.72
UEDIN	40.61	40.97	41.48	40.08
MITLL-AFRL	39.35	42.18	40.62	41.08
RWTH	39.25	41.24	40.16	40.29
KIT	39.11	41.74	40.33	40.63
PRKE-IOIT	38.80	42.86	39.54	42.12
FBK	38.41	42.02	39.09	41.25

TED : MT English-German test 2012 (MT$_{EnDe}$)

System	*case sensitive* BLEU	TER	*case insensitive* BLEU	TER
KIT	23.24	56.17	24.00	55.02
NTT-NAIST	22.86	56.12	24.10	54.57
UEDIN	22.53	57.43	23.26	56.27
RWTH	22.32	57.11	23.04	55.91
POSTECH	20.43	59.14	21.02	58.05

TED : MT English-German test2011 (MT$_{EnDe}$)

System	*case sensitive* BLEU	TER	*case insensitive* BLEU	TER
UEDIN	27.13	50.97	27.75	50.09
KIT	26.29	50.67	26.97	49.76
NTT-NAIST	26.04	50.13	27.27	48.82
RWTH	25.86	51.56	26.58	50.52
POSTECH	23.48	53.71	24.06	52.89

TED : MT English-Arabic test 2012(MT$_{EnAr}$)

System	BLEU	TER
QCRI	15.54	65.57
KIT	15.07	66.46
UEDIN	12.37	69.79

TED : MT English-Arabic test 2011(MT$_{EnAr}$)

System	BLEU	TER
QCRI	15.54	69.19
KIT	14.59	70.60
UEDIN	11.90	72.60

TED : MT English-Spanish test 2012 (MT$_{EnEs}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	26.84	55.86	27.78	54.42

TED : MT English-Spanish test 2011 (MT$_{EnEs}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	33.17	47.77	34.02	46.59

TED : MT English-Farsi test 2012 (MT$_{EnFa}$)

System	BLEU	TER
FBK	10.94	72.66
UEDIN	10.24	74.24

TED : MT English-Farsi test 2011 (MT$_{EnFa}$)

System	BLEU	TER
FBK	12.55	70.06
UEDIN	12.29	71.73

TED : MT English-Italian test 2012(MT$_{EnIt}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	25.28	56.67	26.09	55.55

TED : MT English-Italian test 2011(MT$_{EnIt}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	24.40	57.35	25.15	56.30

TED : MT English-Dutch test 2012(MT$_{EnNl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	26.66	53.21	27.74	51.62

TED : MT Arabic-English test 2012 (MT$_{ArEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
QCRI	30.26	49.55	31.13	48.51
RWTH	29.31	49.46	30.28	48.39
UEDIN	27.72	53.28	28.46	52.34
MITLL-AFRL	27.66	52.18	28.61	51.05

TED : MT Arabic-English test 2011 (MT$_{ArEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
QCRI	27.76	55.17	28.64	54.02
RWTH	27.34	54.41	28.52	53.05
MITLL-AFRL	25.66	57.60	26.58	56.32
UEDIN	25.58	58.91	26.25	57.89

TED : MT Spanish-English test 2011(MT$_{EsEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	30.78	48.65	31.67	47.48

TED : MT Spanish-English test 2012(MT$_{EsEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	37.09	43.45	38.08	42.21

TED : MT Farsi-English test 2012 (MT$_{FaEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	14.98	89.78	15.52	88.79
FBK	14.40	87.26	14.95	86.13

TED : MT Farsi-English test 2011 (MT$_{FaEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
MITLL-AFRL	20.04	62.76	20.90	61.55
UEDIN	19.15	67.64	19.80	66.49
FBK	18.85	66.38	19.48	65.20

TED : MT Italian-English test2012 (MT$_{ItEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	29.62	52.36	30.29	51.40

TED : MT Italian-English test2011 (MT$_{ItEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	30.24	51.81	31.04	50.81

TED : MT Dutch-English test2012 (MT$_{NlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	33.02	47.96	34.46	46.19

TED : MT English-Dutch test 2011(MT$_{EnNl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	30.33	47.48	31.54	45.92

TED : MT Dutch-English test2011 (MT$_{NlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	36.02	45.55	37.36	43.75

TED : MT English-Polish test2012 (MT$_{EnPl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
PJIIT	13.49	75.03	14.29	73.36
UEDIN	10.48	79.05	11.04	77.73

TED : MT Polish-English test2012 (MT$_{PlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
PJIIT	19.77	65.34	20.75	63.79
UEDIN	18.51	66.75	19.39	65.33

TED : MT English-Polish test2011 (MT$_{EnPl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
PJIIT	15.66	68.65	16.61	67.16
UEDIN	13.10	70.96	13.69	69.86

TED : MT Polish-English test2011 (MT$_{PlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
PJIIT	23.29	60.99	24.37	59.36
UEDIN	21.69	62.73	22.57	61.24

TED : MT English-Portuguese test 2012(MT$_{EnPt}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	34.88	43.66	35.84	42.50

TED : MT Portuguese-English test 2012 (MT$_{PtEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	40.56	39.64	41.18	38.95

TED : MT English-Portuguese test 2011(MT$_{EnPt}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	33.59	44.41	34.40	43.37

TED : MT Portuguese-English test 2011 (MT$_{PtEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	39.02	41.24	39.66	40.43

TED : MT English-Romanian test 2012 (MT$_{EnRo}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	19.21	63.03	19.74	62.08

TED : MT Romainan-English test2012 (MT$_{RoEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	31.84	49.19	32.52	48.28

TED : MT English-Romanian test 2012 (MT$_{EnRo}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	23.19	56.60	23.77	55.72

TED : MT Romanian-English test2012 (MT$_{RoEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	36.05	43.99	36.92	42.90

TED : MT English-Russian test 2012(MT$_{EnRu}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
HDU	13.76	73.13	13.83	71.13
UEDIN	13.53	74.66	13.54	72.87

TED : MT Russian-English test 2012 (MT$_{RuEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	20.71	62.78	21.58	61.50
MITLL-AFRL	19.61	62.46	20.53	61.14
HDU	18.20	63.40	19.37	61.74

TED : MT English-Russian test 2011(MT$_{EnRu}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	15.93	67.63	15.94	66.45
HDU	15.53	67.43	15.61	65.79

TED : MT Russian-English test 2011 (MT$_{RuEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	22.13	61.24	22.82	60.05
MITLL-AFRL	21.49	60.10	22.41	58.74
HDU	20.16	61.72	21.30	60.22

TED : MT English-Slovenian test 2012 (MT$_{EnSl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	12.35	70.12	12.88	69.05
RWTH	8.81	73.11	9.22	72.17

TED : MT Slovenian-English test2012 (MT$_{SlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	21.20	61.54	22.03	60.27
RWTH	16.41	65.22	17.00	64.19

TED : MT English-Trukish test 2012 (MT$_{EnTr}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
TUBITAK	9.29	75.46	10.00	73.85
UEDIN	7.41	81.67	7.84	80.20

TED : MT English-Trukish test 2011 (MT$_{EnTr}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
TUBITAK	9.16	75.89	10.19	73.90
UEDIN	7.36	81.30	8.14	79.57

TED : MT English-Chinese test2012 (MT$_{EnZh}$)

System	*character-based*		*word-based*	
	BLEU	TER	BLEU	TER
CASIA	21.88	65.57	13.41	72.64
UEDIN	18.07	71.31	10.80	79.72
KIT	17.93	73.04	10.04	80.39

TED : MT English-Chinese test2011 (MT$_{EnZh}$)

System	*character-based*		*word-based*	
	BLEU	TER	BLEU	TER
CASIA	24.04	62.90	14.94	70.60
KIT	20.41	69.37	11.76	77.88
UEDIN	19.75	68.51	11.54	78.20

TED : MT Turkish-English test 2012 (MT$_{TrEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
TUBITAK	18.93	67.03	19.84	65.49
UEDIN	15.00	72.58	15.77	71.38

TED : MT Turkish-English test 2011 (MT$_{TrEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
TUBITAK	18.63	67.60	19.61	65.99
UEDIN	15.02	73.90	15.89	72.53

TED : MT Chinese-English test 2012(MT$_{ZhEn}$)

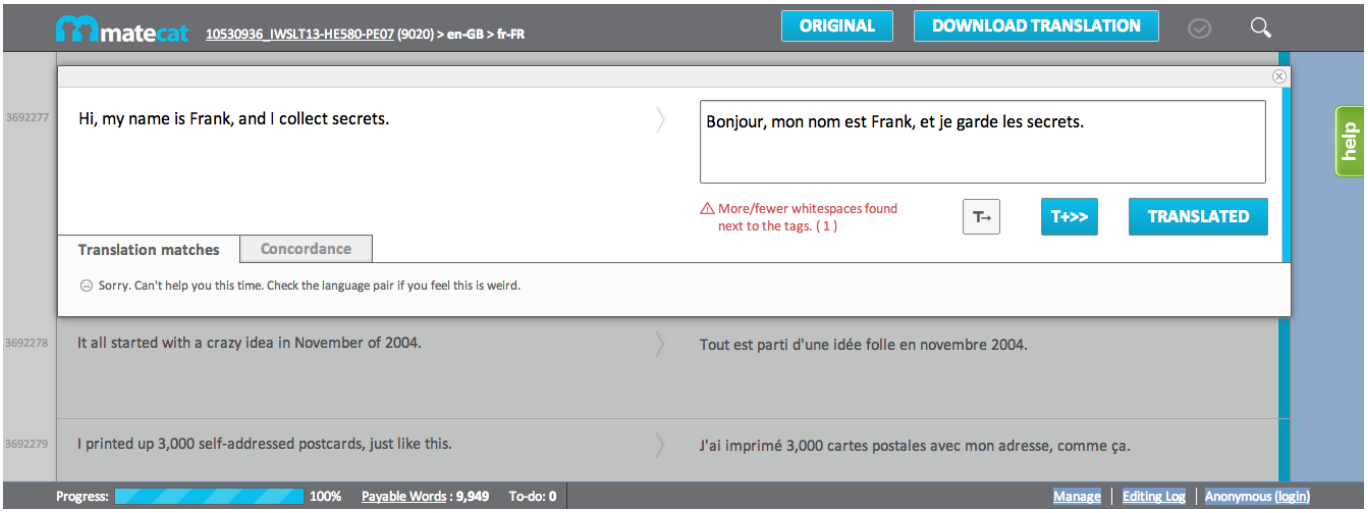
System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
RWTH	14.62	65.73	15.64	64.17
UEDIN	14.19	68.93	15.02	67.54
MITLL-AFRL	14.05	68.26	14.92	66.85
CASIA	12.36	68.76	13.52	66.98

TED : MT Chinese-English test 2011(MT$_{ZhEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
RWTH	16.61	63.37	17.57	61.96
UEDIN	16.10	65.45	16.82	64.18
MITLL-AFRL	15.92	65.68	16.82	64.40
CASIA	14.40	65.60	15.32	64.01

Appendix B. Human Evaluation

Interface used for the bilingual post-editing task



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. 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.

Examples:

Source: This next one takes a little explanation before I share it with you.
Automatic translation: ...avant que je partage avec vous.
Post-editing 1: ...avant de le partager avec vous.
Post-editing 2: ...avant que je le partage avec vous. (preferred - minimal editing and acceptable in spoken language)

Source: And the table form is important.
Automatic translation: Et la forme de la table est importante.
Post-editing 1: La forme de la table est également importante.
Post-editing 2: Et la forme de la table est importante. (preferred - no editing - slightly less fluent but better fitting the source speech transcription)

Source: Everyone who knew me before 9/11 believes...
Automatic translation: ...avant le 11/9...
Post-editing 1: ...avant le 11 septembre...
Post-editing 2: ...avant le 11/9... (preferred - no editing - better fitting the source)