Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014

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 11th evaluation campaign organized by the IWSLT workshop. The 2014 evaluation offered multiple tracks on lecture transcription and translation based on the TED Talks corpus. In particular, this year IWSLT included three automatic speech recognition tracks, on English, German and Italian, five speech translation tracks, from English to French, English to German, German to English, English to Italian, and Italian to English, and five text translation track, also from English to French, English to German, German to English, English to Italian, and Italian to English. In addition to the official tracks, speech and text translation optional tracks were offered, globally involving 12 other languages: Arabic, Spanish, Portuguese (B), Hebrew, Chinese, Polish, Persian, Slovenian, Turkish, Dutch, Romanian, Russian. Overall, 21 teams participated in the evaluation, for a total of 76 primary runs submitted. Participants were also asked to submit runs on the 2013 test set (progress test set), in order to measure the progress of systems with respect to the previous year. All runs were evaluated with objective metrics, and submissions for two of the official text translation tracks were also evaluated with human post-editing.

1. Introduction

This paper overviews the results of the 2014 evaluation campaign organized by the International Workshop of Spoken Language Translation. The IWSLT evaluation has been running now for over a decade and has offered along these years a variety of speech translation tasks [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. The 2014 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,
- Spoken language translation (SLT), that addressed the conversion and translation of a speech signal into a transcript in another language,
- Machine translation (MT), i.e. the translation of a polished transcript into another language.

However, with respect to previous rounds, new languages have been added to each track. The ASR track that previously included German and English, was extended by Italian. The SLT and MT track offered official English-French, English-German, German-English, English-Italian, and Italian-English translation directions. Besides the official evaluation tracks, many other optional translation directions were also offered. Optional SLT directions were English-Arabic and English-Chinese. Optional MT translation directions were: English from/to Arabic, Spanish, Portuguese (B), Hebrew, 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 [11] corpus of TED talks, all data from the WMT 2014 workshop [12], the MULTIUN corpus, and the SETimes parallel corpus. A list of monolingual resources was provided too, that includes both freely available corpora and corpora available from LDC. Test data were released at the beginning 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 2, release of training data; Sept 1–10, ASR test period; Sept 16–25, SLT test period (official directions); Sept 26–Oct 5, MT test period (official directions); Oct 6–17, MT and SLT test period of all optional directions.

All runs submitted by participants were evaluated with automatic metrics. In addition, manual evaluation was carried out for two MT tracks, namely the English-French and English-German tracks. Following the methodology introduced last year, systems were evaluated by calculating HTER values on post-edits created by professional translators. The rational behind this evaluation is to assess the utility of an MT output by measuring the post-editing effort needed by a professional translator to fix it.

This year, 21 sites participated (see Table 1) submitting a total of 76 primary runs: 15 to the ASR track, 16 to the SLT track, and 45 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 experiences gained in this eval-

Proceedings of the 11th International Workshop on Spoken Language Translation
Lake Tahoe, December 4th and 5th, 2014

uation 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 [11] 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.

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 2014 was to transcribe English TED talks, as well as German and Italian 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 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 poorer recording quality than for 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 tst2014 set in CTM format. The word error rate was measured case-insensitive. After the end of the evaluation a preliminary scoring was performed with the first set of references. This was followed by an adjudication phase in which participants could point out errors in the reference transcripts. The adjudication results were collected and combined into the final set of references with which the official scores were calculated.

In order to measure the progress of the systems over the years on English and German, participants also had to provide results on the test set from 2013, i.e. *tst2013*.

3.3. Submissions

For this year's evaluation we received primary submissions from eight sites as well as one combined submission by the EU-BRIDGE project. Seven sites participated in the English evaluation, three sites in the German evaluation and four sites in the Italian one. For English we further received a total of seven contrastive submissions from five sites. For German we received three contrastive submissions from one participant. For Italian we receieved five contrastive submissions from three sites. Also, for English we received a joint submission by the project EU-BRIDGE which was a ROVER combination of the partners' outputs and for which no separate system description was submitted.

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 8.4%–19.7% for English, 24.0%–38.8% for

Table 1: List of Participants

EU-BRIDGE	RWTH& UEDIN& KIT& FBK[13]
FBK	Fondazione Bruno Kessler, Italy [14, 15]
HKUST	Hong Kong University of Science and Technology, Hong Kong [16]
IOIT	Inst. of Inform. and Techn., Vietn. Acad. of Science and Techn. & Thai Nguyen University, Vietnam[17]
KIT	Karlsruhe Institute of Technology, Germany [18, 19]
KLE	Pohang University of Science and Technology, Republic of Korea
LIA	Laboratoire Informatique d'Avignon (LIA) University of Avignon, France [20]
LIMSI	LIMSI - LIMSI, France [21]
LIUM	LIUM, University of Le Mans, France [22]
MIRACL	MIRACL Laboratory Pôle Technologique, Tunisia & LORIA Nancy, France [23]
MITLL-AFRL	Mass. Institute of Technology/Air Force Research Lab., USA
NICT	National Institute of Communications Technology, Japan [24, 25]
NTT-NAIST	NTT Communication Science Labs, Japan & NAIST[26]
PJIIT	Polish-Japanese Institute of Information Technology, Poland [27]
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen, Germany [28]
SFAX	Sfax University, Tunisia
UEDIN	University of Edinburgh, United Kingdom [29, 30]
UMONTREAL	Université de Montréal, Canada
USFD	University of Sheffield, United Kingdom [31]
USTC	National Engineering Laboratory of Speech and Lang. Inform. Proc., Univ. of Science and Techn. of China [32]
VECSYS-LIUM	Vecsys Technologies, France & University of Le Mans, France [22]

German, and 21.9%–25.4% for Italian.

In German, the fact that TEDx have sometimes worse recording conditions than TED talks was reflected by the fact that two talks in the German tst2014 had WERs above 40%. WERs for all other talks were in the range from 9% to 32%.

For English, it can be seen that all participants from IWSLT 2013 made progress, many significant progress, e.g., bringing down the WER from 13.5% to 10.6% on *tst2013*, a relative reduction of 21% over the course of one year. For German, the best performing system only made minor progress, while one of the runner-ups made significant progress and one participant essentially stood the same.

4. SLT Track

4.1. Definition

The SLT track required participants to translate the English, German and Italian talks of *tst2014* 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 and Italian, participants had to translate into English. For English as source language, participants had to translate into French. In addition, participants could also optionally translate from English into one of the following languages: German, Italian, Arabic 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 conference organizers. In order to facilitate scoring, participants had 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 8.2%. For German and Italian we used the two single best scored submissions, as ROVER combination with other systems did not give any performance gains.

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

4.3. Submissions

We received 16 primary and 31 contrastive submissions from nine participants, English to French receiving the most submissions.

4.4. Results

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

Table 2: Monolingual resources for official language pairs

data set	lang	sent	token	voc
train	De	183k	3.36M	124.7k
	En	188k	3.81M	63.4k
	Fr	186k	4.00M	77.0k
	It	185k	3.49M	90.2k

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 reason, we preprocessed all the parallel texts to re-build the original sentences, thus simplifying the MT task.

For each official and optional translation direction, in-domain training and development data were supplied through the website of WIT[3] [11], while out-of-domain training data through the workshop's website. As usual, some of the talks added to the TED repository during the last year have been used to define the new evaluation sets (*tst2014*), 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 *tst2013* of edition 2013 were distributed together with *tst2014* as progressive test sets, when available. Development sets (*dev2010, tst2010, tst2011* and *tst2012*) are either the same of past editions or, in case of new language pairs, have been built upon the same talks.

Evaluation sets *tst2014* of $DeEn$ and $ItEn$ MT tasks derive from those prepared for ASR/SLT tracks, which consist of TEDx talks delivered in German and Italian language, respectively; therefore, no overlap exists with any other TED talk involved in other tasks. Since the $DeEn$ TEDx based MT task was proposed in 2013 as well, the *tst2013* has been released as progressive test set; on the contrary, it is the first time that Italian is involved in ASR/SLT tracks, therefore no evaluation set is available for assessing progress. A single TEDx based development set was released for each pair, together with standard TED based development sets *dev2010, tst2010, tst2011* and *tst2012* sets.

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

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 [33] was applied to all languages, with the exception of Chinese and Arabic languages, which were preprocessed by, respectively: the Stanford Chinese Segmenter [34] and the QCRI-normalizer.[3]

Table 3: Bilingual resources for official language pairs.

MT task	set	sent	tokens		talks
En Fr			En	Fr	
	train	179k	3.63M	3.88M	1415
	TED.dev2010	887	20,1k	20,2k	8
	TED.tst2010	1,664	32,0k	33,9k	11
	TED.tst2011	818	14,5k	15,6k	8
	TED.tst2012	1,124	21,5k	23,5k	11
	TED.tst2013	1,026	21,7k	23,3k	16
	TED.tst2014	1,305	24,8k	27,5k	15
En De			En	De	
	train	172k	3.46M	3.24M	1361
	TED.dev2010	887	20,1k	19,1k	8
	TED.tst2010	1,565	32,0k	30,3k	11
	TED.tst2011	1,433	26,9k	26,3k	16
	TED.tst2012	1,700	30,7k	29,2k	15
	TED.tst2013	993	20,9k	19,7k	16
	TED.tst2014	1,305	24,8k	23,8k	15
	TEDx.dev2012	1,165	21,6k	20,8k	7
	TEDx.tst2013	1,363	23,3k	22,4k	9
	TEDx.tst2014	1,414	28,1k	27,6k	10
En It			En	It	
	train	182k	3.68M	3.44M	1434
	TED.dev2010	887	20,1k	17,9k	8
	TED.tst2010	1,529	31,0k	28,7k	10
	TED.tst2011	1,433	26,9k	24,5k	16
	TED.tst2012	1,704	30,7k	28,2k	15
	TED.tst2013	1,402	30,1k	28,7k	21
	TED.tst2014	1,183	22,6k	21,2k	14
	TEDx.dev2014	1,056	28,9k	28,6k	13
	TEDx.tst2014	883	25,9k	26,5k	13

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 with the MERT procedure provided with Moses, mostly on the development sets *tst2010*; the exceptions are: TEDx tasks, where the TEDx based development sets were used; the two pairs involving Slovenian, where *dev2012* were employed.

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 case-sensitive and had to contain punctuation

[3]QCRI-normalizer was specifically developed for IWSLT Evaluation Campaigns by P. Nakov and F. Al-Obaidli at Qatar Computing Research Institute.

(case+punc).

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). 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 submissions from 14 different sites. On official pairs, the total number of primary runs is 39: 20 on *tst2014* and 19 on *tst2013*; 15 primary runs regard the $EnFr$ pair, 10 the $EnDe$ and 14 the $DeEn$; in addition, we were asked to evaluate also 64 contrastive runs.

Concerning the optional pairs, we received 48 primary runs (25 on *tst2014* and 23 on *tst2013*) and 20 contrastive submissions. The tasks that attracted the most interest are those involving Chinese: 8 primary runs were submitted for $EnZh$, 8 for $ZhEn$. The other submissions involve Arabic, Polish, Farsi, Hebrew, Turkish and Slovenian.

5.4. Results

Table 4: BLEU and TER scores of baseline SMT systems on all *tst2014* sets. (†) TEDx test set. ( ) Char-level scores.

pair		direction			
		BLEU	TER	BLEU	TER
En	Fr	32.07	48.62	–	–
	De	18.33	62.11	†17.89	†64.91
	It	27.15	53.19	†26.12	†55.30
	Ar	11.13	73.01	20.59	62.62
	Es	31.31	48.29	33.88	45.96
	Fa	11.31	71.20	16.74	72.02
	He	15.91	65.62	24.41	58.38
	Nl	22.77	58.38	27.82	52.98
	Pl	9.63	82.81	14.28	68.96
	Pt	31.25	47.25	36.44	42.80
	Ro	18.05	65.25	25.06	54.62
	Ru	11.74	71.99	15.91	69.73
	Sl	8.46	73.94	14.27	71.03
	Tr	7.75	78.69	12.88	77.15
	Zh	16.49	79.50	11.74	72.31

First of all, for reference purposes Table 4 shows BLEU and TER scores on the *tst2014* 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, for English to Chinese character-level scores are reported.

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.

In Appendix A.2 the results on the progress test sets *test2013* 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 primary runs submitted by participants to two of the official MT TED tracks, namely the MT English-German ($EnDe$) track and MT English-French ($EnFr$) track. Following the methodology introduced last year, human evaluation was based on *Post-Editing*, and HTER (Human-mediated Translation Edit Rate) was adopted as the official evaluation metric 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 [35, 36] 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 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. Furthermore, HTER[37] - 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

The human evaluation (HE) dataset created for each MT track was a subset of the corresponding 2013 progress test

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

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

set *(tst2013)*.[6] Both the $EnDe$ and $EnFr$ *tst2013* datasets are composed of 16 TED Talks, and we selected around the initial 60% of each talk. This choice of selecting a consecutive block of sentences for each talk was determined by the need of realistically simulating a caption post-editing task on several TED talks. The resulting HE sets are composed of 628 segments for $EnDe$ and 622 segments for $EnFr$, both corresponding to around 11,000 words.

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. Then, HTER scores were calculated on the created post-edits. HTER [37] 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 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 TER of the MT output.

An interesting outcome of last year's manual evaluation [10] was that the most informative and reliable HTER was not obtained by using only the targeted reference but by exploiting all the post-edits of the evaluated MT outputs. According to these results, also this year systems were officially ranked according to HTER calculated on multiple references.

As for the systems to be evaluated, this year we received five primary runs for the $EnDe$ track and seven for the $EnFr$ track. All the five $EnDe$ MT outputs were post-edited, whereas for the $EnFr$ track we decided to post-edit only five MT outputs out of the seven received. This reduction is not supposed to affect the official evaluation results - since all the participating systems are evaluated with HTER based on multiple post-edits - and it allowed us to respect the budget limitations while offering the community five additional reference translations for a high number of segments (around 60% of the test sets) and for two different language pairs. The five MT outputs selected for post-editing in the $EnFr$ task are the top-5 ranked systems according to automatic evaluation (see Appendix A).

In the preparation of the post-editing data to be collected, some constraints were identified to ensure the soundness of the evaluation: *(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. Furthermore, in order to cope with the 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. For each task, five 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 five 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 an enterprise-level CAT tool developed under the MateCat project[7]. Both the post-editing interface and the guidelines given to translators are presented in Appendix B.

For each task, the resulting collected data consist of five new reference translations for each of the sentences of the HE set. Each one of these five 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 four translations are available.

The main characteristics of the work carried out by post-editors are presented in Table 5 for the $EnDe$ task and in Table 6 for the $EnFr$ task, and largely confirm last year's findings. In the tables, the post-editing effort for each translator is given. Post-editing effort is to be interpreted as the number of actual edit operations performed to produce the post-edited version and - consequently - it is calculated as the HTER of all the system sentences post-edited by each single translator. It is interesting to see that the PE effort is similar for both language pairs, and also highly variable among post-editors, ranging from 19.69% to 40.91% for the $EnDe$ task, and from 17.47% to 39.65% for the $EnFr$ task. Data about weighted standard deviation confirm post-editor variability, showing that the five translators produced quite different post-editing effort distributions.

Table 5: En-De task: Post-editing information for each Post-editor

PEditor	PE Effort	*std-dev*	Sys TER	*std-dev*
PE 1	32.17	18.80	56.05	20.23
PE 2	19.69	13.56	56.32	20.34
PE 3	40.91	17.23	56.18	19.58
PE 4	27.56	14.71	55.93	20.02
PE 5	24.99	15.62	55.63	19.88

Table 6: En-Fr task: Post-editing information for each Post-editor

PEditor	PE Effort	*std-dev*	Sys TER	*std-dev*
PE 1	34.96	20.21	42.60	17.61
PE 2	17.47	14.76	42.81	17.98
PE 3	23.68	14.17	43.02	17.74
PE 4	39.65	20.47	42.27	17.78
PE 5	19.73	14.07	42.86	17.72

[6] Since all the data produced for human evaluation will be made publicly available thorough the WIT^3 repository, we used the 2013 test set in order to keep the 2014 test set blind to be used as a progress test for next year's evaluation.

[7] www.matecat.com

To further study post-editor variability, we exploited the official reference translations available for the two TED tracks and we calculated the TER of the MT outputs assigned to each translator for post-editing ("Sys TER" Column in Tables 5 and 6), as well as the related standard deviation.

As we can see from the tables, the documents presented to translators (composed of segments produced by different systems) are very homogeneous, as they show very similar TER scores and standard deviation figures. This also confirms that the procedure followed in data preparation was effective.

The variability observed in post-editing effort - despite the similarity of the input documents - is most probably due to translators' subjectivity in carrying out the post-editing task. Thus, post-editor variability is an issue to be addressed to ensure a sound evaluation of the systems.

5.5.2. *Evaluation Results*

As anticipated above, last year's human evaluation results demonstrated that HTER computed against all the references produced by all post-editors allowed a more reliable and consistent evaluation of MT systems with respect to HTER calculated against the targeted reference only. Indeed, the HTER reduction obtained using all post-edits clearly showed that exploiting all the available reference translations is a viable way to control and overcome post-editors' variability. For this reason, also this year systems were officially ranked according to HTER calculated on multiple references.

For the $EnDe$ task, HTER was calculated using all the five post-edits available, i.e. for each system the targeted translation and the additional four references were used. For the $EnFr$ task, since the post-edits for two MT outputs had not been created, in order to avoid biases only four post-edits out of five were used to calculate HTER, namely excluding from each system's evaluation its targeted translation.

The official results of human evaluation are given in Tables 7 and 8, which also present a comparison of HTER scores and rankings with TER results - on the HE set and on the full test set - calculated against the official reference translation used for automatic evaluation (see Section 5.2).[8] For the $EnFr$ task, the official HTER results presented in Table 8 for FBK and MIRACL (which do not have a corresponding post-edit) are those obtained on the combination of the four post-edits which gave the best results.

In general, the very low HTER results obtained in both tasks 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 [38], a statistical test well-established in the NLP community [39] and that, especially for the purpose of MT evaluation, has been shown [40] to be less prone to type-I errors than the bootstrap method [41]. The approximate randomization test was based on 10,000 iterations, and differences were considered statistically significant at $p < 0.01$. According to this test, for both tasks a winning system cannot be indicated, as there is no system that is significantly better than all other systems. In particular, for the $EnDe$ task only the bottom-ranking system (KLE) is significantly worse than all the other systems. For the $EnFr$ task, in Table 8 we report - next to the HTER score of each system - the name of the first system in the ranking with respect to which differences are statistically significant. We can see that only the two top-ranking systems are significantly better than the four bottom-ranking systems (from UEDIN to MIRACL), whereas all the other systems significantly differ only with respect to MIRACL.

Furthermore, for comparison purposes, Table 8 presents additional HTER results calculated on all the five post-edits available for the $EnFr$ task. First, it is interesting to note the further HTER reduction achieved, especially for the five top-scoring systems since their corresponding targeted reference was added. Also, comparing the two language pairs, we see that the HTER scores obtained for $EnFr$ with five reference translations are overall lower than those obtained for $EnDe$, indicating that systems translating into French perform better than systems translating into German.

A number of additional observations can be drawn by comparing the official HTER results with TER results. In general, for both tasks we can see that HTER reduces the edit rate of more than 50% with respect to TER. Moreover,

Table 7: En-De Task: Official human evaluation results

System Ranking	**HTER** ***HE Set 5 PErefs***	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**19.22**	54.55	53.62
UEDIN	**19.93**	56.32	55.12
KIT	**20.88**	54.88	53.83
NTT-NAIST	**21.32**	54.68	53.86
KLE	**28.75**	59.67	58.27
Rank Corr.		0.60	0.70

Table 8: En-Fr Task: Official human evaluation results

System Ranking	**HTER** ***HE Set 4 PErefs***	**HTER** HE Set 5 PErefs	**TER** HE Set ref	**TER** Test Set ref
EU-BRIDGE	**19.21**$^{\text{UEDIN}}$	16.48	42.64	43.27
RWTH	**19.27**$^{\text{UEDIN}}$	16.55	41.82	42.58
KIT	**20.89**$^{\text{MIRACL}}$	17.64	42.33	43.09
UEDIN	**21.52**$^{\text{MIRACL}}$	17.23	43.28	43.80
MITLL-AFRL	**22.64**$^{\text{MIRACL}}$	18.69	43.48	44.05
FBK	**22.90**$^{\text{MIRACL}}$	22.29	44.28	44.83
MIRACL	**33.61**	32.90	52.19	51.96
Rank Corr.		0.96	0.90	0.90

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

the correlation between evaluation metrics is measured using *Spearman's rank correlation coefficient* [-1.0, 1.0], with = 1.0 if all systems are ranked in same order, = -1.0 if all systems ranked in reverse order and = 0.0 if no correlation exists. We can see from the tables that TER rankings correlate well with the official HTER.

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 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 eleventh edition of the IWSLT workshop. The evaluation has addressed three tracks: automatic speech recognition of talks (in English, German, and Italian), 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 a progress test set created and used for the 2013 evaluation. This year, 21 participants took part in the evaluation, submitting a total of 76 primary runs, which were all scored with automatic metrics. We also manually evaluated runs of the English-German and English-French text translation tracks. In particular, we asked professional translators to post-edit system outputs on a subset of the 2013 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, for both tracks the post-edit effort of the best performing system results remarkably low, namely around 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 evaluation campaign," in *Proceedings of the International Workshop on Spoken Language Translation*, Trento, Italy, 2007, pp. 1–12.

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

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

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

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

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

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

[11] 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

[12] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna, "Findings of the 2014 Workshop on Statistical Machine Translation," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, 2014.

[13] M. Freitag, J. Wuebker, S. Peitz, H. Ney, M. Huck, A. Birch, N. Durrani, P. Koehn, M. Mediani, I. Slawik, J. Niehues, E. Cho, A. Waibel, N. Bertoldi, M. Cettolo, and M. Federico, "Combined Spoken Language Translation," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[14] B. Babaali, R. Serizel, S. Jalalvand, D. Falavigna, R. Gretter, and D. Giuliani, "FBK @ IWSLT 2014 - ASR track," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[15] N. Bertoldi, P. Mathur, N. Ruiz, and M. Federico, "FBK's Machine Translation and Speech Translation Systems for the IWSLT 2014 Evaluation Campaign," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[16] M. Beloucif, C.-K. Lo, and D. Wu, "Improving tuning against MEANT," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[17] Q. B. Nguyen, T. T. Vu, and C. M. Luong, "The Speech Recognition Systems of IOIT for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[18] K. Kilgour, M. Heck, M. Müller, M. Sperber, S. Stüker, and A. Waibel, "The 2014 KIT IWSLT Speech-to-Text Systems for English, German and Italian," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[19] I. Slawik, M. Mediani, J. Niehues, Y. Zhang, E. Cho, T. Herrmann, T.-L. Ha, and A. Waibel, "The KIT Translation Systems for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[20] M. Morchid, S. Huet, and R. Dufour, "A Topic-based Approach for Post-processing Correction of Automatic Translations," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[21] N. Segal, H. Bonneau-Maynard, Q. K. Do, A. Allauzen, J.-L. Gauvain, L. Lamel, and F. Yvon, "LIMSI English-French Speech Translation System," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[22] A. Rousseau, L. Barrault, P. Deléglise, Y. Estève, H. Schwenk, S. Bennacef, A. Muscariello, and S. Vanni, "The LIUM English-to-French Spoken Language Translation System and the Vecsys/LIUM Automatic Speech Recognition System for Italian Language for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[23] A. B. Romdhane, S. Jamoussi, A. B. Hamadou, and K. Smaili, "Phrase-based Language Modelling for Statistical Machine Translation," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[24] P. Shen, X. Lu, X. Hu, N. Kanda, M. Saiko, and C. Hori, "The NICT ASR System for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[25] X. Wang, A. Finch, M. Utiyama, T. Watanabe, and E. Sumita, "The NICT Translation System for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[26] K. Sudoh, G. Neubig, K. Duh, and K. Hayashi, "NTT-NAIST Syntax-based SMT Systems for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[27] K. Wolk and K. Marasek, "Polish - English Speech Statistical Machine Translation Systems for the IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[28] J. Wuebker, S. Peitz, A. Guta, and H. Ney, "The RWTH Aachen Machine Translation Systems for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[29] P. Bell, P. Swietojanski, J. Driesen, M. Sinclair, F. McInnes, and S. Renals, "The UEDIN ASR Systems for the IWSLT 2014 Evaluation," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[30] A. Birch, M. Huck, N. Durrani, N. Bogoychev, and P. Koehn, "Edinburgh SLT and MT System Description for the IWSLT 2014 Evaluation," in *Proceedings of*

the 11th International Workshop on Spoken Language Translation (IWSLT), Lake Tahoe, CA, 2014.

[31] R. W. M. Ng, M. Doulaty, R. Doddipatla, W. Aziz, K. Shah, O. Saz, M. Hasan, G. Alharbi, L. Specia, and T. Hain, "The USFD SLT system for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[32] S. Wang, Y. Wang, J. Li, Y. Cui, and L. Dai, "The USTC Machine Translation System for IWSLT 2014," in *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, Lake Tahoe, CA, 2014.

[33] 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.

[34] 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.

[35] 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

[36] 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.

[37] 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.

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

[39] 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.

[40] 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

[41] 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 (*tst2014*)

· All the sentence IDs in the IWSLT 2014 test set 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	**8.4**	**(1,831)**
EU-BRIDGE	9.8	(2,138)
MITLL-AFRL	9.9	(2,153)
KIT	11.4	(2,475)
FBK	11.4	(2,492)
LIUM	12.3	(2,689)
UEDIN	12.7	(2,763)
IOIT	19.7	(4,283)

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

System	WER	(# Errors)
KIT	**24.0**	**(5,660)**
UEDIN	35.7	(8,438)
FBK	38.8	(9,167)

TED : ASR Italian (ASR$_{IT}$)

System	WER	(# Errors)
VECSYS-LIUM	**21.9**	**(5,165)**
MITLL-AFRL	23.0	(5,440)
FBK	23.8	(5618)
KIT	25.4	(5,997)

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**27.45**	57.80	**28.16**	56.87
RWTH	26.94	57.29	27.74	**56.22**
LIUM	26.82	59.03	27.85	57.69
UEDIN	25.50	**57.23**	26.26	56.24
FBK	25.39	59.53	26.11	58.57
LIMSI	25.18	60.70	25.88	59.69
USFD	23.45	59.94	24.14	58.97

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
KIT	**17.05**	**68.01**	**17.58**	**66.97**
UEDIN	17.00	68.36	17.51	67.30
USFD	14.75	70.15	15.24	69.15
KLE	13.00	71.70	13.64	70.33

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

System	*case sensitive*		case insensitive	
	BLEU	TER	BLEU	TER
EU-BRIDGE	**19.09**	**63.80**	**19.59**	**62.94**
KIT	18.34	63.91	18.85	62.99
UEDIN	17.67	66.04	18.18	65.12
RWTH	17.24	65.04	17.78	64.07
KLE	9.95	74.05	10.36	72.97

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
EU-BRIDGE	**36.99**	45.20	**37.85**	**44.32**
KIT	36.22	**45.18**	36.97	44.37
UEDIN	35.91	45.78	36.64	45.04
RWTH	35.72	44.54	36.46	43.77
MITLL-AFRL	35.48	45.69	36.90	44.49
FBK	34.24	46.75	34.85	46.04
BASELINE	30.55	49.66	31.13	49.00
MIRACL	25.86	54.16	26.97	53.02
SFAX	16.09	62.89	17.33	61.48

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
EU-BRIDGE	**23.25**	**57.27**	**24.06**	**56.15**
KIT	22.66	57.70	23.35	56.66
UEDIN	22.61	58.95	23.14	57.92
NTT-NAIST	22.09	57.60	22.63	56.65
KLE	19.26	61.36	19.75	60.48
BASELINE	18.44	61.89	18.92	61.02

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
EU-BRIDGE	**25.77**	**54.61**	**26.36**	**53.76**
RWTH	25.04	55.49	25.61	54.65
KIT	24.62	55.62	25.16	54.77
NTT-NAIST	23.77	56.43	24.52	55.49
UEDIN	23.32	57.50	24.06	56.55
FBK	20.52	63.37	21.77	60.66
KLE	19.31	63.88	20.60	61.38
BASELINE	17.50	65.56	18.61	63.08

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

System	BLEU	TER
UEDIN	**13.24**	**69.16**
KIT	13.05	71.62
BASELINE	11.12	72.88

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
MITLL-AFRL	**27.52**	**54.54**	**28.41**	**53.44**
UEDIN	25.46	57.07	26.22	56.02
BASELINE	19.88	63.30	20.48	62.31

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**35.63**	**45.10**	**36.47**	**44.12**
BASELINE	31.26	48.43	31.95	47.48

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**37.29**	**43.73**	**38.07**	**42.85**
BASELINE	33.31	46.07	33.80	45.38

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

System	BLEU	TER
BASELINE	**6.48**	**81.14**

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
MITLL-AFRL	**18.37**	**66.02**	**19.03**	**65.03**
UEDIN	16.94	72.66	17.52	71.66
BASELINE	16.22	72.13	16.72	71.05

TED : MT English-Hebrew (MT$_{EnHe}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
BASELINE	**15.69**	**65.62**	**15.69**	**65.62**

TED : MT Hebrew-English (MT$_{HeEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**26.58**	**56.99**	**27.14**	**56.25**
BASELINE	23.66	58.66	24.20	57.83

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
PJIIT	**16.10**	**74.82**	**16.60**	**73.64**
BASELINE	9.75	82.60	10.16	81.44
LIA	7.79	86.89	10.12	82.31

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
PJIIT	**18.33**	**65.60**	**18.96**	**64.59**
BASELINE	13.94	68.75	14.49	67.63

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**32.41**	**45.85**	**33.12**	**44.87**
BASELINE	31.32	47.06	31.97	46.19

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

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
BASELINE	**35.78**	**43.13**	**36.16**	**42.61**
UEDIN	34.66	46.11	35.28	45.52

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**11.21**	**73.15**	**11.21**	**72.24**

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
LIA	**10.36**	**71.81**	**12.69**	**67.80**
BASELINE	8.53	73.75	8.87	72.76

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**6.97**	**79.93**	**7.36**	**78.65**
UMONTREAL	4.76	80.67	5.51	79.28

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

System	*character-based*	
	BLEU	TER
USTC	21.64	65.71
KIT	18.31	66.43
HKUST	16.41	74.35
BASELINE	15.56	80.48
UMONTREAL	7.40	81.89

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
MITLL-AFRL	**19.30**	**63.95**	**20.22**	**62.64**
BASELINE	15.48	69.93	15.95	68.91

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**13.69**	**70.79**	**14.07**	**69.83**

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**12.52**	**76.96**	**13.10**	**75.77**

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
USTC	**15.65**	**69.65**	**16.35**	**68.62**
NICT	14.05	71.68	14.88	70.42
MITLL-AFRL	12.83	74.74	13.51	73.58
BASELINE	11.22	72.43	11.79	71.37
HKUST	9.64	76.67	10.83	74.16

A.2. Progress Test Set (*tst2013*)

· All the sentence IDs in the IWSLT 2013 test set 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 tst2013

System	IWSLT 2013 WER	IWSLT 2013 (# Errors)	IWSLT 2014 WER	IWSLT 2014 (# Errors)
NICT	**13.5**	**(5,734)**	**10.6**	**(4,518)**
MITLL-AFRL	15.9	(6,788)	13.7	(5,856)
KIT	14.4	(6,115)	14.2	(6,044)
FBK	23.2	(9,899)	14.7	(6,247)
LIUM	—		16.0	(6,818)
UEDIN	22.1	(9,413)	16.3	(6,963)
IOIT	27.2	(11,578)	24.0	(10,206)

TED: ASR German tst2013

System	IWSLT 2013 WER	IWSLT 2013 (# Errors)	IWSLT 2014 WER	IWSLT 2014 (# Errors)
KIT	**25.7**	**(4,932)**	**25.4**	**(5,885)**
UEDIN	37.8	(7,250)	35.0	(6,720)
FBK	37.5	(7,199)	37.8	(7,261)

TED : MT English-French test 2013(MT$_{EnFr}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
EU-BRIDGE	**40.50**	43.27	**41.65**	42.06
KIT	40.12	43.09	41.11	42.04
RWTH	39.72	**42.58**	40.73	**41.52**
UEDIN	39.59	43.80	40.45	42.78
MITLL-AFRL	39.08	44.05	40.59	42.73
FBK	38.20	44.83	38.99	43.88
BASELINE	33.20	48.91	33.81	48.07
MIRACL	29.63	51.96	30.91	50.65

TED : MT English-German test 2013 (MT$_{EnDe}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
EU-BRIDGE	**26.22**	**53.62**	**27.30**	**52.34**
KIT	26.03	53.83	26.77	52.81
NTT-NAIST	25.80	53.86	26.55	52.75
UEDIN	25.33	55.12	26.13	53.93
KLE	21.69	58.27	22.25	57.32
BASELINE	20.96	58.48	21.52	57.58

TED : MT German-English test 2013 (MT$_{DeEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
EU-BRIDGE	**28.77**	**50.52**	**29.29**	**49.63**
KIT	27.98	50.92	28.55	50.04
NTT-NAIST	27.81	51.62	28.32	50.82
UEDIN	27.60	52.43	28.26	51.44
RWTH	27.59	51.33	28.08	50.41
FBK	25.45	55.80	26.07	54.88
KLE	23.59	57.38	24.18	56.47
BASELINE	20.26	60.33	20.89	59.48

TED : MT English-Arabic test 2013(MT$_{EnAr}$)

System	BLEU	TER
UEDIN	**14.20**	**65.97**
KIT	14.15	68.29
BASELINE	12.68	68.94

TED : MT Arabic-English test 2013 (MT$_{ArEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
MITLL-AFRL	**31.48**	**49.88**	**32.41**	**48.76**
UEDIN	29.06	53.02	29.74	52.03
BASELINE	21.63	60.32	22.46	59.12

TED : MT English-Spanish test 2013 (MT$_{EnEs}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**34.74**	**45.75**	**35.42**	**44.78**
BASELINE	30.63	49.39	31.14	48.57

TED : MT Spanish-English test 2013(MT$_{EsEn}$)

System	*case sensitive* BLEU	*case sensitive* TER	*case insensitive* BLEU	*case insensitive* TER
UEDIN	**39.13**	**41.37**	**39.75**	**40.60**
BASELINE	34.18	44.63	34.70	44.00

TED : MT English-Farsi test 2013 (MT$_{EnFa}$)

System	BLEU	TER
BASELINE	7.05	78.90

TED : MT Farsi-English test 2013 (MT$_{FaEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
MITLL-AFRL	**19.47**	**63.27**	**20.11**	**62.27**
UEDIN	16.51	82.50	16.87	81.58
BASELINE	14.04	83.01	14.44	82.09

TED : MT English-Hebrew test 2013(MT$_{EnHe}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**15.92**	**64.16**	**15.92**	**64.16**

TED : MT Hebrew-English test2013 (MT$_{HeEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
UEDIN	**29.70**	**52.40**	**30.51**	**51.35**
BASELINE	25.97	55.40	26.74	54.23

TED : MT English-Polish test2013 (MT$_{EnPl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
PJIIT	**25.92**	**61.04**	**26.62**	**59.94**
BASELINE	11.12	75.95	11.67	74.78

TED : MT Polish-English test2013 (MT$_{PlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
PJIIT	**27.99**	**58.01**	**28.61**	**57.10**
BASELINE	17.25	66.44	17.75	65.44

TED : MT English-Portuguese test 2013(MT$_{EnPt}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	31.38	46.42	31.89	45.66
UEDIN	33.20	44.90	33.93	43.90

TED : MT Portuguese-English test 2013 (MT$_{PtEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	37.85	40.87	38.26	40.35
UEDIN	37.34	42.91	37.80	42.30

TED : MT English-Russian test 2013(MT$_{EnRu}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**14.01**	**70.47**	**14.01**	**69.44**

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

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
MITLL-AFRL	**24.30**	**57.59**	**25.39**	**56.25**
BASELINE	19.82	63.56	20.40	62.46

TED : MT English-Slovenian test 2013 (MT$_{EnSl}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**9.63**	**73.32**	**9.97**	**72.34**

TED : MT Slovenian-English test2013 (MT$_{SlEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**14.64**	**68.68**	**15.19**	**67.63**

TED : MT English-Turkish test 2013 (MT$_{EnTr}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**6.85**	**80.40**	**7.21**	**79.08**
UMONTREAL	4.06	83.97	4.77	82.50

TED : MT Turkish-English test 2013 (MT$_{TrEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
BASELINE	**13.30**	**75.17**	**13.95**	**74.00**

TED : MT English-Chinese test2013 (MT$_{EnZh}$)

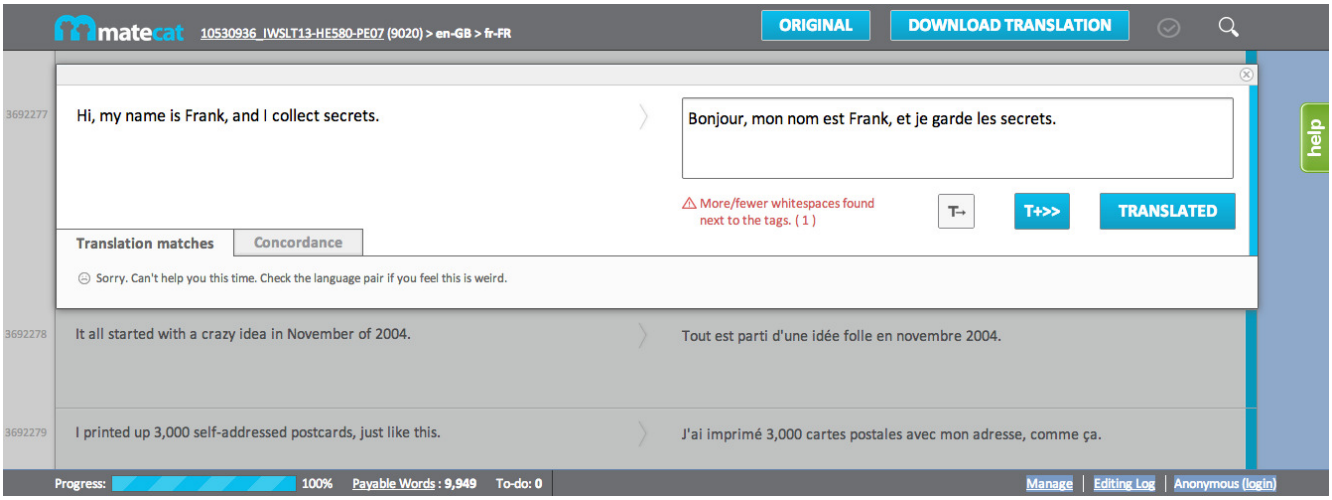
System	*character-based*	
	BLEU	TER
USTC	**22.49**	**63.74**
KIT	21.01	63.12
HKUST	18.81	70.94
BASELINE	18.23	76.15
UMONTREAL	7.93	80.47

TED : MT Chinese-English test 2013(MT$_{ZhEn}$)

System	*case sensitive*		*case insensitive*	
	BLEU	TER	BLEU	TER
USTC	**18.12**	**66.28**	**18.85**	**65.23**
NICT	16.57	67.96	17.36	66.76
MITLL-AFRL	15.59	70.89	16.32	69.68
BASELINE	13.40	68.85	14.00	67.90
HKUST	11.89	72.33	13.08	70.10

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)