# Overview of the IWSLT 2011 Evaluation Campaign

*Marcello Federico*
FBK
via Sommarive 18,
38123 Povo (Trento), Italy
federico@fbk.eu

*Luisa Bentivogli*
CELCT
Via alla Cascata 56/c,
38123 Povo (Trento), Italy
bentivo@fbk.eu

*Michael Paul*
NICT
Hikaridai 3-5,
619-0289 Kyoto, Japan
michael.paul@nict.go.jp

*Sebastian Stüker*
KIT
Adenauerring 2,
76131 Karlsruhe, Germany
sebastian.stueker@kit.edu

## Abstract

We report here on the eighth Evaluation Campaign organized by the IWSLT workshop. This year, the IWSLT evaluation focused on the automatic translation of public talks and included tracks for speech recognition, speech translation, text translation, and system combination. Unlike previous years, all data supplied for the evaluation has been publicly released on the workshop website, and is at the disposal of researchers interested in working on our benchmarks and in comparing their results with those published at the workshop. This paper provides an overview of the IWSLT 2011 Evaluation Campaign, which includes: descriptions of the supplied data and evaluation specifications of each track, the list of participants specifying their submitted runs, a detailed description of the subjective evaluation carried out, the main findings of each exercise drawn from the results and the system descriptions prepared by the participants, and, finally, several detailed tables reporting all the evaluation results.

## 1. Introduction

Over the last 8 years, the International Workshop on Spoken Language Translation (IWSLT) has been proposing challenging research tasks and an open experimental infrastructure for the scientific community working on the automatic translation of spoken and written language. The focus of the IWSLT Evaluation Campaign for this year was the translation of TED[1] talks, a collection of public talks covering a variety of topics. Four different tracks were offered to the participants: (i) ASR, the automatic transcription of talks from audio to text in English; (ii) SLT, addressing the automatic translation of talks from audio (or ASR output) to text, from English to French; (iii) MT, the automatic text translation of talks from English to French, Arabic to English, and Chinese to English; (iv) SC, the application of system-combination methods on ASR outputs in English and MT outputs in English and French.

As is traditionally done at IWSLT, evaluation specifications were prepared for each track, language resources for system training, development and evaluation were made freely available to the participants, and automatic and subjective evaluations were carried out on the outputs of the submitted systems.

The aim of these tracks is to provide an experiment framework for exploring research challenges in both speech recognition and machine translation such as domain, topic and style adaptation, rich transcription and translation from speech, spontaneous speech modeling, and translation between distant languages.

In order to support research on these themes beyond the participation in the IWSLT evaluation, all the prepared benchmarks will be available to the research community until updated with the next IWSLT edition.

In this paper we overview each track by describing its task, evaluation specifications, and language resources made available to the participants. Then, we provide information about the participants in the evaluation and systems that they developed for each track. Notice, that each participant had to submit at least one run for each of the tracks he registered for. Multiple run submissions were allowed, but participants had to explicitly indicate one *primary* run for each track. All other run submissions were treated as *contrastive* runs.

In the following section, we describe the human evaluation that was carried out to evaluate SLT and MT primary runs submitted by the participants. Finally, we provide the results and major findings of the evaluation, by surveying the system papers supplied by the participants. Three appendixes are also provided which report detailed tables of results for the participants' perusal.

## 2. Automatic Speech Recognition

### 2.1. Task Definition

The *Automatic Speech Recognition* (ASR) task for IWSLT 2011 was to recognize the recordings made available by TED on their website. The TED talks collection is a Web repository of recordings of public speeches, mostly held in English, covering a variety of topics, and for which high quality transcriptions and translations into several languages are available.

This task reflects the recent increase of interest in automatically transcribing lectures, in order to make them either searchable, or accessible across languages to also reach au-

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

diences that do not understand the language of the lecturer. Research in this area is especially driven by the fact that nowadays large repositories of lectures are available and distributed through the World Wide Web.

While the speech in TED lectures is in general planned, rather well articulated, and recorded in high quality, challenges arise from the large domain due to the many varying topics that can be the subject of a TED talk. Further challenges arise from the fact that talks in English are also often given by non-native speakers.

### 2.2. Language Resources

#### *2.2.1. Acoustic Model Training Data*

For acoustic model training, no specific data was provided by the evaluation campaign. Instead, participants were allowed to use any data available to them, but recorder before 31 December 2010.

#### *2.2.2. Language Modeling*

For language model training, we defined a closed set of publicly available English texts, so that no participant on the language model side could gain an advantage by including special in-domain data. The data consists of 2M words of TED transcripts, the English portion of the English-French training data from the *Sixth Workshop on Statistical Machine Translation* (WMT 2011) and Google Books N-Grams[2].

### 2.3. Evaluation Specifications

For the evaluation, participants had to provide automatic transcripts of several test talk recordings. The talks were accompanied by an UEM file that marked the portion of each talk that needed to be transcribed. Specifically excluded were the beginning portions of each talk containing a jingle and possibly introductory applause, and the applause and jingle at the end of each file after the speaker has concluded his talk.

In addition, the UEM file also provides a segmentation of each talk into sentence-like units. The segmentation was derived from the human captioning of each talk available for the TED talks on the Web. The use of these segmentations was compulsory for the participants to the evaluation. While giving human-defined segmentation makes the transcription task easier than it would be in real life, the use of it facilitates the speech translation evaluation since the segmentation of the input language perfectly matches the segmentation of the reference translation used in evaluating the translation task.

Participants were required to provide the results of the automatic transcription in CTM format. Participants were allowed multiple submissions for the task, but one submission had to be marked as the primary run.

The quality of the submissions was then scored in terms of word error rate (WER). The results were scored case-insensitive, but were allowed to be submitted case-sensitive. Numbers, dates, etc., had to be transcribed in words as they are spoken, not in digits. Common acronyms, such as NATO and EU, had to be written as one word, without any special markers between the letters. This applies no matter whether they are spoken as one word or spelled out as a letter sequence. All other letter spelling sequences had to be written as individual letters with spaces in between. Standard abbreviations, such as "etc." and "Mr." were accepted as specified by the GLM file in the scoring package that was provided to participants for development purposes. For words pronounced in their contracted form, it was permitted to use the orthography for the contracted form, as these cases were nevertheless normalized (after the GLM file) into their canonical form.

[2]http://ngrams.googlelabs.com/datasets

## 3. Spoken Language Translation

### 3.1. Task Definition

In the *Spoken Language Translation* (SLT) task, participants were required to translate the English TED talks into French, starting from the audio signal. The challenge of this translation task over the *Machine Translation* (MT) task, described in Section 4, is the necessity to deal with automatic, and in general error prone, transcriptions of the audio signal, instead of correct human transcriptions. In addition, the automatic transcripts supplied by the organizers do not contain true-casing nor punctuation information.

To get the most information out of the error prone output of the speech recognition system, the organizers also supplied word lattices in addition to the first best output of the ASR system.

### 3.2. Language Resources

For the SLT task the language resources available to participants are the union of those of the ASR track, described in Section 2.2, and of the English-to-French MT track, described in Section 4.2.

In addition, for development purposes, ASR outputs for the IWSLT 2010 development and test sets were also made available to participants.

### 3.3. Evaluation Specifications

The participants had to provide the result of the translation of the English audio in NIST XML format. The output had to be true-cased and had to contain punctuation. The participants could either use the audio files directly, or use the output—either first best hypotheses in CTM format or word lattices in SLF — of KIT, LIUM, and FBK from the ASR task.

The quality of the translations was measured automatically with BLEU [1] by scoring against the human translations created by the TED open translation project, and by human subjective evaluation (paired comparison, Section 7). Since the reference translations from the TED website match the segmentation of the reference transcriptions of the talks,

automatic evaluation scores for the MT outputs could be directly computed.

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

Moreover, automatic evaluation scores were also calculated for case-insensitive (lower-case only) MT outputs with punctuation marks removed (*no_case+no_punc*). Besides the BLEU metric scores, automatic evaluation using six additional standard metrics (METEOR [2], WER [3], PER [4], TER [5], GTM [6], and NIST [7]) were calculated offline and are listed in Appendix A.

In order to decide whether the translation output at the document-level of one MT engine is significantly better than another, we used the *bootStrap*[3] method that (1) performs a random sampling with replacement from the evaluation data set, (2) calculates the respective evaluation metric score of each engine for the sampled test sentences and the difference between the two MT system scores, (3) repeats the sampling/scoring step iteratively, and (4) applies the *Student's t-test* at a significance level of 95% confidence to test whether the score differences are significant [8]. In this year's evaluation, 2000 iterations were used for the analysis of the automatic evaluation results. Omitted lines between scores in the automatic evaluation result tables listed in Appendix A indicate non-significant differences in performance between the MT engines.

Correlations between different metrics were calculated using the *Spearman rank correlation coefficient* $\rho$, which is a non-parametric measure of correlation that assesses how well an arbitrary monotonic function can describe the relationship between two variables without making any assumptions about the frequency distribution of the variables. It is calculated as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2-1)},$$

where $d_i$ is the difference between the rank of the system $i$ and $n$ is the number of systems. The correlation results are listed in Appendix C.

# 4. Machine Translation

## 4.1. Task Definition

IWSLT 2011 features three different *Machine Translation* (MT) tasks for the text translation of English to French ($\text{MT}_{EF}$), Arabic to English ($\text{MT}_{AE}$), and Chinese to English ($\text{MT}_{CE}$).

The $\text{MT}_{EF}$ task was carried out using the same TED corpus as the Spoken Language Translation ($\text{SLT}_{EF}$) task described in Section 3. The order of $\text{SLT}_{EF}$ and $\text{MT}_{EF}$ talks is unchanged, so that the evaluation of both tasks is carried out using the same reference translation data sets. However, in contrast to the $\text{SLT}_{EF}$ task, the text input data of the $\text{MT}_{EF}$ task did not contain any speech recognition errors.

For the $\text{MT}_{AE}$ and $\text{MT}_{CE}$ tasks, TED talks not included in the $\text{MT}_{EF}$ data sets were selected from the TED website and used as the language resources for the respective task. Although the TED talk IDs of the $\text{MT}_{AE}$ and $\text{MT}_{CE}$ tasks are identical, the order of the talks were shuffled randomly resulting in different evaluation data sets for both tasks.

For the MT tasks, participants were requested to translate 8 ($\text{MT}_{EF}$ task) or 16 talks ($\text{MT}_{AE,CE}$), each comprising 90 $\sim$100 sentences on average. The text input file format was XML. The MT output results had to be submitted either in the original XML format or as plain text via e-mail.

## 4.2. Language Resources

The language resources provided to the participants of the MT tasks comprise monolingual and parallel corpora of TED talks ($train$) that are copyright of the TED Conference LLC[4] and distributed under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 license[5]. The development ($dev2010$) and evaluation ($tst2010$) data sets for last year's IWSLT 2010 TED translation task were also provided to the participants for system tuning and translation quality evaluation. Concerning the official evaluation data set for IWSLT 2011 ($tst2011$), only the source language texts were distributed to the participants. All text resources were case-sensitive and included punctuation marks. Details on the supplied monolingual and parallel resources are given in Table 1 and Table 2, respectively.

Moreover, several out-of-domain parallel corpora, including texts from the United Nations, European Parliament, and news commentaries, which were kindly provided by the organizers of the *6th Workshop on Statistical Machine Translation*[6] and the *EuroMatrixPlus project*[7], could also be used by the participants to train their systems. The *Google Books ngrams*[8] copyright of Google Inc. and distributed under a Creative Commons Attribution 3.0 license were also available to the participants. For details on the out-of-domain language resources, please refer to the IWSLT homepage[9].

## 4.3. Evaluation Specifications

The evaluation specifications of the MT tasks were identical to the ones of the SLT task described in Section 3.3. In addition to the MT outputs provided by the participants, the organizers used an online MT server to translate the testset data sets for the MT tasks. The online system (*online*) represents a state-of-the-art general-domain MT system that differs from the participating MT systems in two aspects: (1) its language resources are not limited to the supplied corpora and (2) its

[3] http://projectile.sv.cmu.edu/research/public/tools/bootStrap/tutorial.htm

[4] http://www.ted.com/talks

[5] http://creativecommons.org/licenses/by-nc-nd/3.0/

[6] http://www.statmt.org/wmt11/translation-task.html

[7] http://www.euromatrixplus.net/

[8] http://books.google.com/ngrams/datasets

[9] http://iwslt2011.org/doku.php?id=06_evaluation#download_of_training_data

Table 1: Monolingual Resources.

| Data | Lang | Sent | Token | Voc |
|---|---|---|---|---|
| train | en | 123,914 | 2.40M | 51.2K |
| | fr | 111,431 | 2.40M | 60.2K |

Table 2: Bilingual Resources.

| Task | Data | Lang | Sent | Token | Voc | Talks (Sen) |
|---|---|---|---|---|---|---|
| $MT_{EF}$ | train | en | 107,324 | 2.07M | 46.5K | 764 (140) |
| | | fr | | 2.21M | 58.1K | |
| | dev2010 | en | 934 | 20.1K | 3.4K | 8 (117) |
| | | fr | | 20.3K | 3.9K | |
| | tst2010 | en | 1,664 | 32.0K | 3.9K | 11 (151) |
| | | fr | | 33.8K | 4.8K | |
| | tst2011 | en | 818 | 14.5K | 2.5K | 8 (102) |
| | | fr | | 15.6K | 3.0K | |
| $MT_{AE}$ | train | ar | 90,590 | 1.62M | 71.1K | 672 (135) |
| | | en | | 1.74M | 42.4K | |
| | dev2010 | ar | 934 | 18.3K | 4.6K | 8 (117) |
| | | en | | 20.1K | 3.4K | |
| | tst2010 | ar | 1,664 | 29.2K | 6.0K | 11 (151) |
| | | en | | 32.0K | 3.9K | |
| | tst2011 | ar | 1,450 | 25.3K | 5.8K | 16 (91) |
| | | en | | 27.0K | 3.7K | |
| $MT_{CE}$ | train | zh | 107,097 | 1.95M | 56.8K | 755 (142) |
| | | en | | 2.07M | 46.8K | |
| | dev2010 | zh | 934 | 21.6K | 3.7 | 8 (117) |
| | | en | | 20.1K | 3.4K | |
| | tst2010 | zh | 1,664 | 33.3K | 4.4K | 11 (151) |
| | | en | | 32.0K | 3.9K | |
| | tst2011 | zh | 1,450 | 24.8K | 3.9K | 16 (91) |
| | | en | | 27.0K | 3.7K | |

parameters are not optimized using in-domain data. Its purpose is to investigate the applicability of a baseline system with unlimited language resources to the spoken language translation tasks investigated by the IWSLT evaluation campaign.

## 5. System Combination

*System Combination* (SC) is an approach for computing a consensus hypothesis from the outputs of multiple systems. The combination of different hypotheses can be based on confusion networks constructed by aligning the hypotheses with regard to word similarities and has been shown to be quite successful in automatic speech recognition [9] as well as machine translation [10].

Four System Combination tasks were carried out for IWSLT 2011, one ASR System Combination ($ASR^{SC}$) task and three MT System Combination ($MT^{SC}$) tasks, i.e., one for each of the MT tasks described in Section 4.

### 5.1. Language Resources

Table 3 summarizes the amount of ASR and MT output provided by the participants of the respective ASR and MT tasks that could be exploited by the participants of the respective System Combination tasks ($ASR^{SC}_{E}$, $MT^{SC}_{EF,AE,CE}$).

Table 3: System Combination Resources.

| Task | Lang | Runs | Systems |
|---|---|---|---|
| $ASR^{SC}_{E}$ | en | 5 | fbk, kit, lium, mit, nict |
| $MT^{SC}_{EF}$ | fr | 7 | dfki, fbk, kit, lig, limsi, mit, rwth |
| $MT^{SC}_{AE}$ | en | 4 | dcu, fbk, mit, rwth |
| $MT^{SC}_{CE}$ | en | 4 | dcu, msr, nict, rwth |

### 5.2. Task Definition

For the $ASR^{SC}$ and $MT^{SC}$ tasks, participants were requested to generate new recognition and translation hypotheses on the basis of the respective ASR and MT task translation results of the IWSLT 2011 testset. The system combination output results had to be submitted either in the original format of the input files or as plain text via e-mail.

### 5.3. Evaluation Specifications

The evaluation specifications of the ASR and SLT/MT tasks described in Section 2.3 and Section 3.3 were also used for the evaluation of the $ASR^{SC}$ and $MT^{SC}$ task run submissions. Both automatic ($ASR^{SC}$ and $MT^{SC}$ runs) and subjective ($MT^{SC}$ runs) evaluation metrics were applied and the scores of the system combination runs were directly compared to the $ASR_E$ and $MT_{EF,AE,CE}$ systems whose translation results were used to carry out the system combination. System ranking was performed on the combined set of ASR (MT) task and $ASR^{SC}$ ($MT^{SC}$) task run submission results, respectively.

## 6. Participants

A list of the participants of this year's evaluation is shown in Table 4. The number of primary and contrastive run submissions for each tasks are summarized in Table 5. In total, 30 primary runs and 51 contrastive runs were submitted by the participants.

Table 5: Run Submissions.

| Task | Primary (+Online) | Contrastive [Systems] |
|---|---|---|
| $ASR_E$ | 5 | 3 [FBK:2, MIT:1] |
| $SLT_{EF}$ | 5 | 6 [FBK:3, LIG:1, LIUM:1, RWTH:1] |
| $MT_{EF}$ | 7 (+1) | 13 [MIT:9, FBK:3, DFKI:1] |
| $MT_{AE}$ | 4 (+1) | 15 [MIT:6, DCU:4, RWTH:3, FBK:2] |
| $MT_{CE}$ | 4 (+1) | 6 [RWTH:3, MSR:2, DCU:1] |
| $ASR^{SC}$ | 2 | 3 [FBK:2, LIUM:1] |
| $MT^{SC}$ | 3 | 5 [DFKI:3, MSR:2] |

## 7. Human Evaluation

The subjective evaluation was carried out on all primary runs submitted by participants to the SLT, MT, and $MT^{SC}$ tracks. Regarding all MT tasks, individual systems were jointly eval-

Table 4: List of Participants.

| Short | Full names and system paper references | ASR | SLT | $MT_{FE}$ | $MT_{AE}$ | $MT_{CE}$ | $ASR^{SC}$ | $MT^{SC}$ |
|---|---|---|---|---|---|---|---|---|
| DCU | Centre For Next Generation Localization, Dublin City U., Ireland [11] | | | | X | X | | |
| DFKI | Deutsche Forschungszentrum für Künstliche Intelligenz, Germany [12] | | | X | | | | X |
| FBK | Fondazione Bruno Kessler, Italy [13] | X | X | X | X | | X | |
| KIT | Karlsruhe Institute of Technology, Germany [14] | X | X | X | | | | |
| LIG | Laboratory of Informatics of Grenoble, France [15] | | X | X | | | | |
| LIMSI | LIMSI, France [16] | | | X | | | | |
| LIUM | Laboratoire d'Informatique de l'Université du Maine, France [17] | X | X | | | | X | |
| MIT | Mass. Institute of Technology/Air Force Research Lab., USA [18] | X | | X | X | | | |
| MSR | Microsoft Research, USA [19] | | | | | X | | X |
| NICT | National Institute of Communications Technology, Japan [20, 21] | X | | | | X | | |
| RWTH | Rheinisch-Westfälische Technische Hochschule Aachen, Germany [22] | | X | X | X | X | | |
| | | 5 | 5 | 7 | 4 | 4 | 2 | 2 |

uated with the SC runs and the additional *online* system runs prepared by the organizers.

For each task, systems were evaluated by a subjective evaluation set composed of 400 sentences randomly taken from the test set used for automatic evaluation. Each evaluation set represents the various lengths of the sentences included in the corresponding test set, with the exception of sentences with less than 5 words, which were excluded from the subjective evaluation.

The IWSLT 2011 subjective evaluation focused solely on the *Ranking* task[10] and a number of novelties were introduced with respect to the traditional system ranking evaluation carried out in previous campaigns.

Firstly, this year's evaluation was not carried out by hired expert graders but by relying on crowdsourced data. The feasibility of using crowdsourcing methodologies as an effective way to reduce the costs of MT evaluation without sacrificing quality was investigated in a previous experiment [23], where the ranking evaluation of the IWSLT 2010 Arabic-English BTEC task was replicated by hiring non-experts through Amazon's Mechanical Turk. The analysis of the collected data showed that agreement rates for non-experts were comparable to those for experts, and that the crowd-based system ranking had a very strong correlation with expert-based ranking.

Secondly, the cost reduction obtained by using crowdsourcing allowed us to focus on modifying and extending the ranking methodology in different respects, with the aim of maximizing the overall evaluation reliability.

The goal of the *Ranking* evaluation is to produce a complete ordering of the systems participating in a given task. The ranking task requires human judges to decide whether one system output is better than another for a given source sentence. The judgments collected through these comparisons are used to obtain the ranking scores, which are calculated as the average number of times that a system was judged better than any other system.

Traditionally, in the ranking task, the judge was presented with the output of five submissions for a given source sentence and was asked to rank them from best to worst (ties were allowed) [24]. Each evaluation block contained the implicit pairwise comparisons (i.e. each system against the other systems presented in the same block) which constituted the basis of the ranking scores.

Although ranking a number of translated sentences relative to each other is quite intuitive, a 5-fold ranking task is less reliable than a direct comparison between only two translated sentences due to the higher cognitive load required to perform the task. Thus, a major innovation introduced this year to address annotation reliability was to abandon the traditional 5-fold ranking task and to directly collect pairwise comparisons.

The second innovation addresses system ranking reliability, and focuses on the number of human judgments collected for each single task. In order to achieve a complete ordering over the systems, full coverage of pairwise comparisons would be required. In previous campaigns, the traditional 5-fold ranking task data were created through a random selection of a (reasonably large) sample of all the possible system comparisons. In IWSLT 2011, we achieved full coverage by collecting pairwise comparisons following a round-robin tournament structure.

In a round-robin structure each system competes against every other system. We carried out multiple round-robins, where systems competed in 400 tournaments corresponding to the subjective evaluation sentences. On the one hand, the round-robin tournament is the the most complete way to determine system ranking. On the other hand, we wanted to investigate whether collecting a higher number of assessments for each task can highlight significant differences between systems that would otherwise be insignificant.

The system scores resulting from human evaluation are listed in Appendix B, which also provides detailed tables about pairwise head-to-head comparisons.

In the following sections we analyze the data that we collected by posting the ranking task on Amazon's Mechanical

[10] Last year human evaluation was also carried out for the *Fluency* and *Adequacy* metrics.

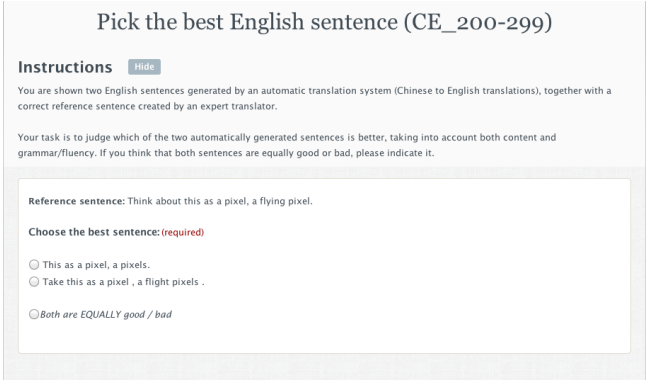


Figure 1: The ranking task based on pairwise comparisons as displayed on the MTurk interface.

Turk[11] (MTurk) through the CrowdFlower[12] (CFlower) interface.

### 7.1. Data Collection

For each task, we first prepared all the paired comparisons necessary for a complete round-robin over the 400 evaluation sentences. Details about the number of comparisons for which systems were evaluated are given in Table 6. As an example, for the $\text{SLT}_{EF}$ task we had five system submissions, and thus each system was evaluated against each of the other 4 systems 400 times (once for each evaluation sentence), leading to a total of 1,600 comparisons. Considering all systems, there are 10 pairwise comparisons for each evaluation sentence, corresponding to 4,000 comparisons for the whole evaluation set.

Table 6: Summary of the IWSLT 2011 *Ranking* task.

| **Task** | # systems | # comparisons per system | # comparison in total |
|---|---|---|---|
| $\text{SLT}_{EF}$ | 5 | 1,600 | 4,000 |
| $\text{MT}_{EF}$ | 9 | 3,200 | 14,400 |
| $\text{MT}_{AE}$ | 6 | 2,000 | 6,000 |
| $\text{MT}_{CE}$ | 6 | 2,000 | 6,000 |

All the pairwise comparisons to be evaluated were posted to MTurk through the CFlower interface[13]. Figure 1 shows the ranking task based on pairwise comparisons as presented to MTurk contributors.

For each pairwise comparison we requested three redundant judgments from different MTurk contributors. This means that for each task we collected three times the number of the necessary judgments (e.g. 12,000 for the $\text{SLT}_{EF}$ task).

Redundant judgment collection is a typical method to ensure the quality of crowdsourced data. In fact, instead of relying on a single judgment, label aggregation is computed by applying majority voting. Moreover, agreement information can be collected to find and manage the most controversial annotations[25].

In our ranking task (i.e. assessing two system outputs on the same source sentence), there are three possible assessments: *(i)* output A is better than output B, *(ii)* output A is worse than output B, or *(iii)* both output A and B are equally good or bad (tie). Given that we had three judgments and three possible values, we were faced with a number of comparisons for which it was not possible to assign a majority vote.

In order to calculate the ranking scores, *undecidable* comparisons can be managed in two ways, namely *(i)* interpreting the result of the comparison as a tie between the systems (neither of them won) and thus keeping all the collected data, or *(ii)* keeping only the data for which a majority assessment was actually obtained (without any further interpretation) and thus excluding undecidable comparisons from the evaluation.

We carried out the evaluation with both dataset configurations in order to better understand the impact of undecidable comparisons on system ranking. We found that undecidable comparisons represent a small percentage of all comparisons, ranging from 6.42% for the $\text{MT}_{EF}$ task to 13.33% for the $\text{SLT}_{EF}$ task. Given the large amount of data obtained with round-robin tournaments, a ranking evaluation carried out excluding undecidable comparisons can still be based on a high number of judgments. Moreover, undecidable comparisons are equally distributed among all comparisons (i.e. they do not affect specific head-to-head system comparisons) and system ranking does not change when they are excluded from the evaluation. Given that undecidable comparisons do not affect system ranking, we can conclude that they can be excluded from the evaluation, which can thus be based on the most consistent data only.

### 7.2. Inter-Annotator Agreement

In order to investigate the degree of consistency between human evaluators, we calculated inter-annotator agreement using *Fleiss' kappa coefficient* $\kappa$ [26, 27][14]. This coefficient measures the agreement between multiple raters (three or more) each of whom classifies $N$ items into $C$ mutually exclusive categories, taking into account the agreement occurring by chance. It is calculated as:

$$\kappa = \frac{P(a)-P(e)}{1-P(e)}$$

[11] http://www.mturk.com

[12] http://www.crowdflower.com

[13] A detailed description of the crowdsourcing methodology and of the quality control mechanisms used for data collection (i.e. locale qualifications and gold units) are described in detail in [23].

[14] This year, intra-annotator agreement was not calculated. On the one hand, it is inherently less significant for crowdsourced data, as a very high number of judges works at each task. On the other hand, gathering enough data to compute a meaningful intra-annotator agreement rate can be very difficult/expensive as CFlower does not allow requesters to ask that the same contributor completes the same work unit more than once (or to select a group of units to be completed by the same contributor).

Table 7: Inter-annotator agreement rates for the manual evaluation of the IWSLT 2011 tasks.

| Task | # judgments | P(*a*) | P(*e*) | $\kappa$ |
|---|---|---|---|---|
| SLT$_{EF}$ | 12,000 | 0.47 | 0.34 | 0.20 |
| MT$_{EF}$ | 43,200 | 0.61 | 0.37 | 0.39 |
| MT$_{AE}$ | 18,000 | 0.54 | 0.35 | 0.29 |
| MT$_{CE}$ | 18,000 | 0.51 | 0.37 | 0.22 |

where $P(a)$ is the observed pairwise agreement between the raters and $P(e)$ is the estimated agreement due to chance, calculated empirically on the basis of the cumulative distribution of judgments by all raters. If the raters are in complete agreement then $\kappa$ = 1. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$.

Table 7 shows inter-annotator agreement rates for the various tasks. It is worthwhile to note that, whereas in previous years inter-annotator agreement rates were calculated on a sample of repeated pairwise comparisons, this year the collection of redundant judgments allowed us to systematically calculate inter-annotator agreement for each pairwise comparison. The interpretation of the $\kappa$ values according to [28] is given in Table 8.

Table 8: Interpretation of the $\kappa$ coefficient.

| $\kappa$ | Interpretation |
|---|---|
| $< 0$ | No agreement |
| 0.0 – 0.20 | Slight agreement |
| 0.21 – 0.40 | Fair agreement |
| 0.41 – 0.60 | Moderate agreement |
| 0.61 – 0.80 | Substantial agreement |
| 0.81 – 1.00 | Almost perfect agreement |

According to this interpretation, the agreement over all the collected judgments results to be "slight" for the SLT$_{EF}$ task and "fair" for all the other tasks. If we consider only the comparisons with a real majority judgment (i.e. excluding undecidable cases), the agreement rates increase to 0.30 for the SLT$_{EF}$ task, 0.45 for the MT$_{EF}$ task, 0.38 for the MT$_{AE}$ task, and 0.31 for the MT$_{CE}$ task.

Comparing the inter-annotator agreement trend with data about automatic and manual evaluation, significant relations do not emerge. However, we notice that inter-annotator agreement rates seem to reflect the difficulty of the tasks, as higher agreement rates are recorded for those tasks where system performances are generally higher.

# 8. Main Findings

In this section, we try to point out methods and solutions that, according to the participants' descriptions, contributed most significantly to the performance of their systems. Our ideal goal is to provide some useful guidelines for setting up strong baselines for each track for the benefit of future participants or any interested researcher. The complete list of system description papers that we consulted is included in the references and can be tracked from Table 4.

## 8.1. ASR Track

Most ASR systems, including the best one by MIT, have acoustic models trained on TED talk recordings crawled from the Web. Supervised data for each talk was automatically selected after transcribing audio with a baseline ASR engine exploiting the captions available for each talk. This resulted in about 150 hours of speech. Notably, KIT achieved the second best performance by taking advantage of a different but significantly larger amount of supervised and unsupervised data (450 hour), including political speeches, news, and lectures.

Concerning speech pre-processing, most systems used rich sets of acoustic features, including up to third order MFCC, MVDR, and PLP coefficients, that were finally compressed with dimensionality reduction methods such as LDA and HLDA. In particular, LIUM used additional acoustic features in a re-scoring step that were generated by a multi-layer neural network. Acoustic models of the top three performing systems, MIT, KIT, LIUM, were trained discriminatively, using the MPE or MMIE criterions, while all other systems were trained with MLE.

All of the best systems employed 4-gram language models that were trained by interpolating TED data with other provided corpora.

All systems performed multi-pass decoding by applying speaker adaptation and by increasing resolution of models, e.g. from non-SAT to SAT acoustic models, and from 3-gram to 4-gram languages models. Some systems also included decoding steps employing acoustic models based on different acoustic features or lexicons, e.g. LIUM and FBK.

## 8.2. MT Track

The best submitted MT runs took advantage of data selection and domain adaptation techniques, both for translation and language modeling. Criteria for data selection include entropy or perplexity (DCU, LIUM, RWTH) and alignment-score based criteria (LIUM, RWTH), while adaptation methods considered the well known linear interpolation (see in particular the empirical weight estimation technique by MIT), log-linear interpolation (see in particular the additional scores introduced by KIT), and the more recent fill-up method by FBK.

Concerning language specific pre-processing of data, we point out the combined use of multiple word segmentation methods for Arabic and Chinese by RWTH and NICT. All participants computed word alignments with GIZA++, but MSR, which used an internal HMM-based tool, RWTH, which also used an internal EM-based phrase-extraction tool, and NICT, which employed in addition a Bayesian aligner based on a Pitman-Yor process model. All submitted runs were produced by either phrase-based (PB) or hierarchical phrase-

based (HPB) decoders. Best performance using single MT decoders were achieved for all tasks by using PB systems. (Notice that MSR submitted as primary run a combination of PB and HPB systems.) Among the sites that directly compared the two decoders (RWTH, MSR, DCU,DFKI), only half reported slightly better BLEU scores with HPB decoding. In particular, RWTH for $\text{MT}_{EF}$ and $\text{MT}_{CE}$, DCU for $\text{MT}_{AE}$ and $\text{MT}_{CE}$. This apparent inconsistency points out that it is still difficult to exactly compare the two approaches, but also that for same translation directions they perform rather similarly.

Finally, among the interesting features of the submitted MT systems, we point out the neural network language models employed by LIUM and LIMSI, the hybrid language model used to model the style of talks proposed by FBK, the syntax-based language model based on categorial grammars by DCU, the bilingual language model and re-ordering model used by KIT, the re-ordering constraints based on punctuation introduced by LIG and NICT, the topic-specific translation model and the discriminative training proposed by MSR, and the use of shallow rules for HPB decoding introduced by RWTH.

### 8.3. SLT Track

In this task, systems had to process ASR output, that was without punctuation and capitalization. RWTH performed comparative experiments showing that it is more convenient to recover punctuation before the translation step, rather than during or after it. Among the five participants, FBK and RWTH introduced punctuation before the translation step, KIT during translation, and LIUM and LIG after translation. In the later two cases, all training data were suitably preprocessed. All participants introduced capitalization on the MT output after translation. Concerning the MT engine, all participants used more or less the same set of features. The two best performing systems (LIUM and LIG) tuned their system on a dev set of ASR output, and investigated the use of multiple ASR hypotheses in input. LIUM compared SLT performance by using as input their own system 1-best output, confusion network output, and SC output. The latter input, showing the lowest WER, also resulted in the highest BLEU score. Finally, LIG showed significant improvements by instead separately translating different ASR outputs and then combining the translations with a Rover technique.

### 8.4. SC Track

System combinations of ASR runs were submitted by FBK and LIUM. Best results were obtained by FBK though the application of the Rover method to all systems but the least performing one. LIUM compared the bag of n-gram (BANG) and Rover methods on a different subset of runs, reporting a superiority of the Rover method. A further small improvements was reported after adding the outcome of the BANG method to the Rover configuration.

System combinations of MT runs were submitted by DFKIfor English-French and Arabic-English, and by MSR for Chinese-English. SC by MSR was performed with the Incremental Indirect HMM method, while DFKI implemented a sentence selection algorithm based on a log-linear model exploiting an heterogenous set features.

Unfortunately, both sites reported difficulties in tuning their SC methods, manly due to a significant mismatch in number and quality between the early and final MT runs provided to the participants. According to them, this mismatch was a main reason for the lack of meaningful improvements by SC over the single best runs. For this reason, we plan to improve this track in the future by postponing the submission of development data for SC to a date close to the submission of the MT evaluation sets.

## 9. Conclusions

The IWSLT 2012 Evaluation Campaign represents a break from previous editions in several aspects. First, we radically changed the application scenario from human-human dialogues in the travel domain to public talks on a variety of topics. Second, we publicly released all the supplied data and benchmarks used in the evaluation, to the advantage of any researcher interested in replicating or improving the results published at the workshop. Third, we added automatic speech recognition and system combination among the evaluation tracks. Finally, we carried out a subjective evaluation of the machine translation outputs by means of crowdsourcing and paired comparisons.

As expected, the increase in the complexity of the translation task has impacted the number of participants, which decreased with respect to the past editions. However, we hope that the high quality of results achieved by the participants for this year will attract more research labs in the future. In addition, by making all language resources and benchmarks freely available to the research community, we hope to significantly increase the interest around the translation of talks.

## 10. Acknowledgements

Special thanks go to Mauro Cettolo, Christian Girardi, Teresa Hermann, Giovanni Moretti, and Jan Niehues for contributing in the preparation of all TED data sets and in the analysis of results. Research Group 3-01'Multilingual Speech Recognition' received financial support from the 'Concept for the Future' of the Karlsruhe Institute of Technology within the framework of the German Excellence Initiative. The IWSLT organizers acknowledge support by the EuroMatrixPlus project (IST-231720) and the T4ME network of excellence (IST-249119), which are both funded by the European Commission under the Seventh Framework Programme for Research and Technological Development.

## 11. References


[1] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, USA, 2002, pp. 311–318.

[2] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation (WMT)*, Prague, Czech Republic, 2007, pp. 228–231.

[3] S. Niessen, F. J. Och, G. Leusch, and H. Ney, "An Evaluation Tool for Machine Translation: Fast Evaluation for Machine Translation Research," in *Proceedings of the Second International Conference on Language Resources & Evaluation (LREC)*, Athens, Greece, 2000, pp. 39–45.

[4] F. J. Och, "Minimum Error Rate Training in SMT," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, 2003, pp. 160–167.

[5] 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.

[6] J. P. Turian, L. Shen, and I. D. Melamed, "Evaluation of Machine Translation and its Evaluation," in *Proceedings of the MT Summmit IX*, New Orleans, USA, 2003, pp. 386–393.

[7] G. Doddington, "Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics," in *Proceedings of the Second International Conference on Human Language Technology (HLT)*, San Diego, USA, 2002, pp. 257–258.

[8] Y. Zhang, S. Vogel, and A. Waibel, "Interpreting Bleu/NIST Scores: How Much Improvement do We Need to Have a Better System?" in *Proceedings of the Second International Conference on Language Resources & Evaluation (LREC)*, 2004, pp. 2051–2054.

[9] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, Santa Barbara, USA, 1997, pp. 347–354.

[10] E. Matusov, G. Leusch, R. E. Banchs, N. Bertoldi, D. Dechelotte, M. Federico, M. Kolss, Y.-S. Lee, J. B. Marino, M. Paulik, S. Roukos, H. Schwenk, and H. Ney, "System Combination for Machine Translation of Spoken and Written Language," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14(5), pp. 1222–1237, 2008.

[11] P. Banerjee, H. Almaghout, S. Naskar, J. Jiang, A. Way, J. Roturier, and J. van Genabith, "The DCU Machine Translation Systems for IWSLT 2011," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[12] D. Vilar, E. Avramidis, M. Popović, and S. Hunsicker, "DFKI's SC and MT Submissions to IWSLT 2011," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[13] N. Ruiz, A. Bisazza, F. Brugnara, D. Falavigna, D. Giuliani, S. Jaber, R. Gretter, and M. Federico, "FBK @ IWSLT 2011," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[14] M. Mediani, E. Cho, J. Nieheus, T. Herrmann, and A. Waibel, "The KIT English-French Translation system for IWSLT 2011," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[15] B. Lecouteux, L. Besancier, and H. Blanchon, "LIG English-French Spoken Language Translation System for IWSLT 2011," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[16] T. Lavergne, A. Allauzen, H.-S. Le, and F. Yvon, "LIMSI's experiments in domain adaptation for IWSLT11," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[17] A. Rousseau, F. Bougares, P. Deléglise, H. Schwenk, and Y. Estève, "LIUM's system for the IWSLT 2011 Speech Translation Tasks," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[18] R. Aminzadeh, T. Anderson, R. Slyh, B. Ore, E. Hansen, W. Shen, J. Drexler, and T. Gleason, "The MIT-LL/AFRL IWSLT-2011 MT System," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[19] X. He, A. Axelrod, L. Deng, A. Acero, M.-Y. Hwang, A. Nguyen, A. Wang, and X. Huang, "The MSR System for IWSLT 2011 Evaluation," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[20] K. Abe, Y. Wu, C. lin Huang, P. R. Dixon, S. Matsuda, C. Hori, and H. Kashioka, "The NICT ASR System for IWSLT 2011," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[21] A. Finch, C.-L. Goh, G. Neubig, and E. Sumita, "The NICT Translation System for IWSLT 2011," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[22] J. Wuebker, M. Huck, S. Mansour, M. Freitag, M. Feng, S. Peitz, C. Schmidt, and H. Ney, "The RWTH Aachen Machine Translation System for IWSLT 2011," in *Proceedings of the eight International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA, 2011.

[23] L. Bentivogli, M. Federico, G. Moretti, and M. Paul, "Getting Expert Quality from the Crowd for Machine Translation Evaluation," in *Proceedings of the MT Summmit XIII*, Xiamen, China, 2011, pp. 521–528.

[24] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(Meta-) Evaluation of Machine Translation," in *Proceedings of the Second Workshop on Statistical Machine Translation (WMT)*, Prague, Czech Republic, 2007, pp. 136–158.

[25] C. Callison-Burch, "Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, August 2009, pp. 286–295. [Online]. Available: http://www.aclweb.org/anthology/D/D09/D09-1030

[26] S. Siegel and N. J. Castellan, *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1988.

[27] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76(5), 1971.

[28] J. Landis and G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33 (1), pp. 159–174, 1977.

# Appendix A. Automatic Evaluation

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

## A.1. Full Testset

· All the sentence IDs in the IWSLT 2011 testset were used to calculate the automatic scores for each run submission.
· The "$^{SC}$" system name suffixes indicate runs submitted to the system combination tasks.
· ASR and MT systems are ordered according to the *WER* and *BLEU* metrics, respectively.
· For each task, the best (worst) score of each metric is marked with **boldface** (`typewriter`).
· Besides the NIST metrics, all automatic evaluation metric scores are given as percent figures (%).

### A.1.1. Significance Test

· The mean scores of 2000 iterations were calculated for each MT output according to the *bootStrap* method [8].
· Omitted lines between scores indicate non-significant differences in performance between the MT engines.

**SLT English-French (SLT$_{EF}$)**

| *"case+punc" evaluation* | | | | | | | **System** | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **28.15** | **19.28** | 57.50 | **48.05** | 54.73 | **58.05** | **6.152** | LIUM | **29.38** | **20.23** | 57.29 | 48.50 | 55.69 | **57.35** | 6.446 |
| 26.77 | 19.23 | **56.66** | 49.03 | **53.70** | 56.17 | 6.119 | KIT | 28.25 | 19.75 | **56.08** | **48.47** | **54.80** | 56.95 | **6.461** |
| 26.74 | 18.74 | 57.58 | 48.51 | 54.88 | 57.21 | 6.150 | RWTH | 28.04 | 19.40 | 58.03 | 49.05 | 56.50 | 56.77 | 6.387 |
| 24.84 | `17.64` | 58.54 | 49.19 | 55.71 | 56.62 | 5.956 | LIG | `25.97` | `18.14` | `59.36` | `50.16` | `57.78` | 55.74 | `6.188` |
| `24.30` | 17.65 | `58.89` | `49.46` | `55.79` | `55.70` | `5.906` | FBK | 26.09 | 18.34 | 58.58 | 49.63 | 57.04 | `55.67` | 6.215 |

**MT English-French (MT$_{EF}$)**

| *"case+punc" evaluation* | | | | | | | **System** | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **37.64** | **27.14** | **43.84** | **37.49** | **41.70** | **66.92** | **7.489** | KIT | **36.50** | **25.82** | **46.31** | **39.79** | **45.03** | **64.88** | **7.579** |
| 36.49 | 26.07 | 46.06 | 39.12 | 43.66 | 65.49 | 7.173 | LIMSI | 35.25 | 24.76 | 48.73 | 41.67 | 47.36 | 63.25 | 7.255 |
| 36.12 | 26.04 | 45.80 | 39.09 | 43.74 | 65.61 | 7.266 | RWTH | 34.31 | 24.39 | 48.98 | 41.75 | 47.62 | 63.19 | 7.298 |
| 35.28 | 25.32 | 46.67 | 39.17 | 43.97 | 65.26 | 7.182 | MIT | 34.19 | 23.87 | 49.40 | 41.38 | 47.70 | 63.28 | 7.292 |
| 34.86 | 25.38 | 47.54 | 40.10 | 44.73 | `64.68` | 7.099 | FBK | 34.31 | 24.02 | 49.44 | 41.82 | 48.03 | 63.11 | 7.256 |
| 34.55 | 25.36 | 46.92 | 39.93 | 44.13 | 64.81 | 7.118 | LIG | 33.76 | 23.72 | 49.83 | 42.03 | 48.24 | 63.04 | 7.239 |
| `34.38` | `24.45` | `48.00` | `40.25` | `45.69` | 64.70 | `7.039` | DFKI | `32.60` | `22.93` | `51.14` | `43.02` | `49.56` | `62.15` | `7.062` |
| 40.71 | 29.47 | 43.37 | 37.16 | 41.41 | 67.99 | 7.488 | ONLINE | 39.48 | 28.08 | 46.11 | 39.49 | 44.84 | 66.03 | 7.554 |
| 37.53 | 26.91 | 44.42 | 37.63 | 42.07 | 66.82 | 7.416 | DFKI$^{SC}$ | 36.42 | 25.62 | 46.97 | 39.89 | 45.61 | 64.75 | 7.509 |

**MT Arabic-English (MT$_{AE}$)**

| *"case+punc" evaluation* | | | | | | | **System** | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **26.34** | **61.12** | **58.41** | **48.64** | **55.79** | **58.39** | **6.205** | RWTH | **25.37** | **58.79** | **60.71** | **50.27** | **59.15** | **57.06** | **6.304** |
| 24.32 | 59.13 | 61.84 | 51.09 | 59.05 | 56.90 | 5.908 | FBK | 22.98 | 56.62 | 64.70 | 53.29 | 62.86 | 55.23 | 5.924 |
| 19.80 | `54.67` | `68.60` | `55.96` | `64.64` | `52.40` | `5.245` | DCU | 19.31 | 52.41 | `68.22` | `56.21` | `66.16` | `51.92` | `5.418` |
| `19.56` | 55.36 | 64.64 | 53.06 | 61.19 | 54.31 | 5.352 | MIT | `19.18` | `52.09` | 67.47 | 55.16 | 65.50 | 52.40 | 5.482 |
| 24.60 | 59.21 | 62.40 | 51.86 | 59.34 | 55.71 | 5.954 | ONLINE | 23.61 | 57.66 | 63.45 | 52.65 | 61.64 | 55.40 | 6.061 |
| 23.66 | 58.85 | 62.20 | 51.53 | 59.32 | 56.49 | 5.760 | DFKI$^{SC}$ | 22.67 | 56.03 | 64.71 | 53.63 | 62.99 | 54.70 | 5.832 |

**MT Chinese-English (MT$_{CE}$)**

| *"case+punc" evaluation* | | | | | | | **System** | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **16.88** | **52.84** | **67.24** | **54.56** | **63.05** | **51.96** | **5.155** | MSR | **15.43** | **48.62** | 71.06 | **57.61** | 68.10 | **49.06** | **5.150** |
| 15.12 | 50.03 | 66.56 | 55.48 | 62.81 | 50.60 | 4.829 | RWTH | 13.61 | 45.87 | **70.41** | 58.73 | **67.93** | 47.58 | 4.689 |
| 12.12 | `45.91` | `75.82` | `60.70` | `70.67` | `45.62` | `4.382` | DCU | 11.31 | `42.95` | `76.60` | `62.04` | `73.27` | `44.18` | `4.433` |
| `11.90` | 48.46 | 71.77 | 57.46 | 67.15 | 48.44 | 4.582 | NICT | `11.05` | 45.01 | 75.24 | 59.78 | 71.85 | 46.19 | 4.596 |
| 15.18 | 51.17 | 69.95 | 56.08 | 65.55 | 49.79 | 5.157 | ONLINE | 14.12 | 49.12 | 73.14 | 57.79 | 69.49 | 48.51 | 5.211 |
| 17.02 | 53.18 | 68.25 | 54.59 | 63.80 | 52.26 | 5.157 | MSR$^{SC}$ | 15.64 | 49.05 | 72.05 | 57.42 | 68.71 | 49.42 | 5.191 |

### A.1.2. Evaluation Server

· The results are obtained using the IWSLT 2011 online evaluation tool at https://mastarpj.nict.go.jp/EVAL/IWSLT11/automatic/testset_IWSLT11

**ASR English (ASR$_{EF}$)**

| System | WER | (Count) |
|---|---|---|
| MIT | **15.30** | (1943) |
| KIT | 17.10 | (2172) |
| LIUM | 17.40 | (2208) |
| FBK | 18.20 | (2306) |
| NICT | 27.30 | (3466) |
| FBK$^{SC}$ | 13.60 | (1726) |
| LIUM$^{SC}$ | 13.90 | (1762) |

**SLT English-French (SLT$_{EF}$)**

| *"case+punc" evaluation* | | | | | | | System | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **28.23** | **19.30** | 57.49 | **48.06** | 54.73 | **58.06** | **6.411** | LIUM | **29.40** | **20.24** | 57.28 | 48.50 | 55.68 | **57.36** | 6.736 |
| 26.78 | 19.24 | **56.66** | 49.04 | **53.71** | 56.17 | 6.382 | KIT | 28.26 | 19.76 | **56.07** | **48.47** | **54.80** | 56.95 | **6.752** |
| 26.76 | 18.75 | 57.57 | 48.51 | 54.88 | 57.21 | 6.410 | RWTH | 28.06 | 19.41 | 58.02 | 49.05 | 56.49 | 56.78 | 6.671 |
| 24.85 | 17.64 | 58.54 | 49.20 | 55.72 | 56.61 | 6.197 | LIG | 25.98 | 18.14 | 59.36 | 50.16 | 57.78 | 55.74 | 6.453 |
| 24.31 | 17.66 | 58.90 | 49.47 | 55.80 | 55.70 | 6.146 | FBK | 26.11 | 18.34 | 58.58 | 49.64 | 57.05 | 55.67 | 6.484 |

**MT English-French (MT$_{EF}$)**

| *"case+punc" evaluation* | | | | | | | System | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **37.65** | **27.14** | **43.84** | **37.50** | **41.70** | **66.92** | **7.841** | KIT | **36.50** | **25.83** | **46.31** | **39.79** | **45.03** | **64.89** | **7.948** |
| 36.49 | 26.06 | 46.07 | 39.13 | 43.68 | 65.48 | 7.502 | LIMSI | 35.26 | 24.76 | 48.73 | 41.68 | 47.37 | 63.25 | 7.601 |
| 36.12 | 26.04 | 45.80 | 39.09 | 43.75 | 65.61 | 7.596 | RWTH | 34.31 | 24.40 | 48.98 | 41.74 | 47.62 | 63.20 | 7.640 |
| 35.28 | 25.32 | 46.68 | 39.18 | 43.99 | 65.25 | 7.505 | MIT | 34.19 | 23.87 | 49.41 | 41.39 | 47.72 | 63.27 | 7.635 |
| 34.87 | 25.39 | 47.54 | 40.10 | 44.74 | 64.68 | 7.417 | FBK | 34.31 | 24.03 | 49.44 | 41.82 | 48.03 | 63.11 | 7.596 |
| 34.54 | 25.35 | 46.94 | 39.95 | 44.15 | 64.80 | 7.433 | LIG | 33.74 | 23.71 | 49.84 | 42.04 | 48.26 | 63.03 | 7.574 |
| 34.39 | 24.46 | 48.01 | 40.26 | 45.69 | 64.70 | 7.351 | DFKI | 32.62 | 22.94 | 51.14 | 43.02 | 49.56 | 62.16 | 7.386 |
| 40.69 | 29.46 | 43.37 | 37.18 | 41.43 | 67.97 | 7.845 | ONLINE | 39.47 | 28.07 | 46.13 | 39.50 | 44.86 | 66.02 | 7.927 |
| 37.53 | 26.91 | 44.43 | 37.64 | 42.08 | 66.81 | 7.757 | DFKI$^{SC}$ | 36.42 | 25.62 | 46.97 | 39.89 | 45.62 | 64.75 | 7.869 |

**MT Arabic-English (MT$_{AE}$)**

| *"case+punc" evaluation* | | | | | | | System | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **26.32** | **61.10** | **58.44** | **48.66** | **55.81** | **58.37** | **6.417** | RWTH | **25.35** | **58.77** | **60.74** | **50.29** | **59.18** | **57.05** | **6.525** |
| 24.31 | 59.12 | 61.86 | 51.11 | 59.08 | 56.88 | 6.102 | FBK | 22.97 | 56.61 | 64.72 | 53.31 | 62.88 | 55.21 | 6.123 |
| 19.80 | 54.66 | 68.61 | 55.96 | 64.65 | 52.39 | 5.404 | DCU | 19.30 | 52.40 | 68.23 | 56.20 | 66.16 | 51.91 | 5.586 |
| 19.56 | 55.35 | 64.65 | 53.07 | 61.20 | 54.30 | 5.503 | MIT | 19.21 | 52.08 | 67.48 | 55.17 | 65.51 | 52.39 | 5.646 |
| 24.60 | 59.20 | 62.41 | 51.86 | 59.35 | 55.69 | 6.163 | ONLINE | 23.61 | 57.65 | 63.46 | 52.65 | 61.65 | 55.39 | 6.278 |
| 23.64 | 58.83 | 62.23 | 51.56 | 59.34 | 56.47 | 5.945 | DFKI$^{SC}$ | 22.66 | 56.02 | 64.73 | 53.65 | 63.01 | 54.68 | 6.024 |

**MT Chinese-English (MT$_{CE}$)**

| *"case+punc" evaluation* | | | | | | | System | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **16.89** | **52.84** | 67.23 | **54.54** | 63.04 | **51.96** | **5.300** | MSR | **15.45** | **48.63** | 71.05 | **57.59** | 68.09 | **49.07** | **5.301** |
| 15.13 | 50.04 | **66.56** | 55.48 | **62.80** | 50.60 | 4.964 | RWTH | 13.61 | 45.87 | **70.40** | 58.73 | **67.92** | 47.58 | 4.824 |
| 12.12 | 45.91 | 75.82 | 60.69 | 70.66 | 45.62 | 4.492 | DCU | 11.30 | 42.96 | 76.60 | 62.03 | 73.26 | 44.19 | 4.549 |
| 11.90 | 48.47 | 71.7 7 | 57.46 | 67.14 | 48.44 | 4.693 | NICT | 11.06 | 45.03 | 75.23 | 59.77 | 71.84 | 46.20 | 4.714 |
| 15.19 | 51.19 | 69.93 | 56.06 | 65.52 | 49.81 | 5.313 | ONLINE | 14.14 | 49.15 | 73.12 | 57.76 | 69.47 | 48.53 | 5.373 |
| 17.02 | 53.18 | 68.25 | 54.58 | 63.80 | 52.26 | 5.301 | MSR$^{SC}$ | 15.65 | 49.05 | 72.05 | 57.41 | 68.70 | 49.43 | 5.341 |

## A.2. Human Assessment Subset

· A 400 sentence ID subset of the subjective evaluation was used for calculating the automatic scores of each run submission.
· The "$SC$" system name suffixes indicate runs submitted to the system combination tasks.
· ASR and MT systems are ordered according to the *WER* and *BLEU* metrics, respectively.
· For each task, the best (worst) score of each metric is marked with **boldface** (`typewriter`).
· Besides the NIST metrics, all automatic evaluation metric scores are given as percent figures (%).

### A.2.1. Significance Test

· The mean scores of 2000 iterations were calculated for each MT output according to the *bootStrap* method [8].
· Omitted lines between scores indicate non-significant differences in performance between the MT engines.

**SLT English-French (SLT$_{EF}$)**

| *"case+punc" evaluation* | | | | | | | **System** | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **27.30** | **18.97** | 57.85 | **48.28** | 55.12 | **57.56** | 5.763 | LIUM | **27.93** | **19.62** | 57.98 | 48.98 | 56.32 | **56.68** | **5.990** |
| 25.69 | 18.48 | 57.96 | 48.75 | 55.21 | 56.65 | **5.777** | RWTH | 26.65 | 18.99 | 58.32 | 49.05 | 56.82 | 56.28 | 5.987 |
| 25.64 | 18.67 | **57.30** | 49.23 | **54.26** | 55.58 | 5.705 | KIT | 26.83 | 19.02 | **56.83** | **48.90** | **55.48** | 56.14 | 5.988 |
| 23.82 | 17.39 | 58.85 | 49.59 | 56.02 | 55.55 | 5.616 | FBK | 25.28 | 18.02 | 58.58 | 49.74 | 57.09 | 55.32 | 5.850 |
| 23.78 | 17.31 | 58.77 | 49.43 | 55.86 | 56.11 | 5.598 | LIG | 24.58 | 17.53 | 59.62 | 50.28 | 57.98 | 55.38 | 5.799 |

**MT English-French (MT$_{EF}$)**

| *"case+punc" evaluation* | | | | | | | **System** | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **36.43** | **26.65** | **44.44** | **37.86** | **42.31** | **66.32** | **7.013** | KIT | **35.15** | **25.28** | **46.88** | **40.05** | **45.50** | **64.30** | **7.060** |
| 36.30 | 26.21 | 45.94 | 39.23 | 43.66 | 65.18 | 6.809 | LIMSI | 34.84 | 24.79 | 48.57 | 41.72 | 47.36 | 62.94 | 6.850 |
| 35.49 | 25.79 | 46.01 | 39.25 | 44.01 | 65.30 | 6.860 | RWTH | 33.48 | 24.20 | 49.15 | 41.85 | 47.86 | 62.90 | 6.856 |
| 34.48 | 25.09 | 47.01 | 39.17 | 44.14 | 64.92 | 6.777 | MIT | 33.40 | 23.62 | 49.52 | 41.20 | 47.74 | 63.03 | 6.852 |
| 34.39 | 25.28 | 47.28 | 39.71 | 44.52 | 64.93 | 6.761 | FBK | 33.64 | 23.94 | 49.18 | 41.36 | 47.71 | 63.41 | 6.867 |
| 34.02 | 24.59 | 47.75 | 40.10 | 45.41 | 64.57 | 6.683 | DFKI | 32.05 | 22.99 | 50.87 | 42.86 | 49.24 | 62.00 | 6.664 |
| 33.99 | 25.39 | 46.51 | 39.69 | 43.65 | 64.75 | 6.758 | LIG | 33.17 | 23.72 | 49.27 | 41.44 | 47.62 | 63.26 | 6.848 |
| 40.13 | 29.08 | 43.87 | 37.59 | 41.93 | 67.44 | 7.044 | ONLINE | 38.72 | 27.65 | 46.45 | 39.75 | 45.21 | 65.58 | 7.074 |
| 37.45 | 26.96 | 43.89 | 37.07 | 41.68 | 67.10 | 7.075 | DFKI$^{SC}$ | 35.98 | 25.59 | 46.47 | 39.37 | 45.15 | 64.95 | 7.105 |

**MT Arabic-English (MT$_{AE}$)**

| *"case+punc" evaluation* | | | | | | | **System** | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **26.28** | **60.60** | **58.44** | **48.67** | **55.88** | **58.05** | **5.682** | RWTH | **24.95** | **58.26** | **60.95** | **50.38** | **59.42** | **56.63** | **5.731** |
| 24.15 | 58.16 | 62.97 | 52.01 | 60.33 | 56.08 | 5.376 | FBK | 22.49 | 55.52 | 65.99 | 54.65 | 64.26 | 54.11 | 5.330 |
| 19.65 | 55.55 | 64.42 | 52.84 | 60.88 | 54.16 | 4.995 | MIT | 19.27 | 52.17 | 67.34 | 55.02 | 65.27 | 52.33 | 5.095 |
| 19.31 | 54.30 | 68.68 | 56.31 | 64.79 | 51.99 | 4.813 | DCU | 18.81 | 51.85 | 68.48 | 56.67 | 66.48 | 51.42 | 4.996 |
| 23.72 | 58.42 | 62.98 | 52.09 | 59.74 | 55.32 | 5.442 | ONLINE | 22.47 | 56.81 | 64.17 | 52.91 | 62.16 | 54.93 | 5.513 |
| 23.41 | 57.92 | 62.87 | 51.93 | 59.96 | 55.81 | 5.269 | DFKI$^{SC}$ | 22.18 | 55.08 | 65.55 | 54.32 | 63.92 | 53.80 | 5.298 |

**MT Chinese-English (MT$_{CE}$)**

| *"case+punc" evaluation* | | | | | | | **System** | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **16.78** | **53.03** | 67.65 | **54.24** | 63.49 | **52.12** | **4.799** | MSR | **15.31** | **48.95** | 71.47 | **57.25** | 68.46 | **49.18** | **4.773** |
| 15.23 | 50.39 | **66.24** | 54.63 | **62.60** | 51.09 | 4.568 | RWTH | 13.44 | 46.38 | **69.89** | 58.08 | **67.54** | 47.75 | 4.403 |
| 12.52 | 46.65 | 75.39 | 59.90 | 70.29 | 46.24 | 4.167 | DCU | 11.55 | 43.66 | 76.42 | 61.25 | 73.03 | 44.81 | 4.192 |
| 12.16 | 48.70 | 71.71 | 57.09 | 67.15 | 48.80 | 4.346 | NICT | 10.96 | 45.25 | 75.01 | 59.58 | 71.68 | 46.39 | 4.310 |
| 16.38 | 51.90 | 69.54 | 55.19 | 65.34 | 50.55 | 4.898 | ONLINE | 15.20 | 49.76 | 72.56 | 57.03 | 69.11 | 49.14 | 4.916 |
| 17.13 | 53.49 | 68.43 | 54.30 | 64.14 | 52.54 | 4.835 | MSR$^{SC}$ | 15.72 | 49.75 | 71.94 | 57.03 | 68.60 | 49.75 | 4.849 |

### A.2.2. Evaluation Server

· A 400 sentence ID subset of the run submissions listed in Appendix A.1.2 was used for calculating the automatic scores of each run submission.

**ASR English (ASR$_{EF}$)**

| System | WER | (Count) |
|---|---|---|
| MIT | **15.40** | (1029) |
| KIT | 17.40 | (1160) |
| LIUM | 18.10 | (1204) |
| FBK | 18.20 | (1213) |
| NICT | 27.30 | (1819) |
| FBK$^{SC}$ | 13.80 | ( 919) |
| LIUM$^{SC}$ | 14.20 | ( 949) |

**SLT English-French (SLT$_{EF}$)**

| *"case+punc" evaluation* | | | | | | | System | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **27.34** | **18.96** | 57.87 | **48.30** | 55.14 | **57.55** | 6.032 | LIUM | **27.92** | **19.61** | 58.00 | 49.00 | 56.35 | **56.66** | **6.287** |
| 25.67 | 18.46 | 57.98 | 48.78 | 55.24 | 56.63 | **6.049** | RWTH | 26.62 | 18.97 | 58.35 | 49.09 | 56.85 | 56.25 | 6.282 |
| 25.61 | 18.65 | **57.33** | 49.27 | **54.29** | 55.55 | 5.975 | KIT | 26.79 | 18.99 | **56.87** | **48.94** | **55.52** | 56.11 | 6.285 |
| 23.82 | 17.38 | 58.86 | 49.60 | 56.04 | 55.53 | 5.874 | FBK | 25.26 | 18.01 | 58.61 | 49.77 | 57.11 | 55.30 | 6.135 |
| 23.77 | 17.30 | 58.79 | 49.45 | 55.89 | 56.09 | 5.850 | LIG | 24.57 | 17.51 | 59.65 | 50.30 | 58.01 | 55.36 | 6.075 |

**MT English-French (MT$_{EF}$)**

| *"case+punc" evaluation* | | | | | | | System | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **36.40** | **26.63** | **44.47** | **37.89** | **42.35** | **66.29** | **7.373** | KIT | **35.12** | **25.27** | **46.91** | **40.09** | **45.54** | **64.26** | **7.439** |
| 36.28 | 26.18 | 45.96 | 39.25 | 43.69 | 65.17 | 7.157 | LIMSI | 34.82 | 24.77 | 48.59 | 41.74 | 47.38 | 62.93 | 7.217 |
| 35.47 | 25.76 | 46.05 | 39.29 | 44.05 | 65.27 | 7.204 | RWTH | 33.46 | 24.18 | 49.19 | 41.89 | 47.90 | 62.87 | 7.213 |
| 34.47 | 25.08 | 47.03 | 39.19 | 44.17 | 64.91 | 7.116 | MIT | 33.39 | 23.61 | 49.54 | 41.22 | 47.75 | 63.01 | 7.212 |
| 34.38 | 25.27 | 47.30 | 39.73 | 44.54 | 64.92 | 7.097 | FBK | 33.62 | 23.93 | 49.20 | 41.38 | 47.73 | 63.39 | 7.224 |
| 34.00 | 24.57 | 47.79 | 40.13 | 45.45 | 64.56 | 7.012 | DFKI | 32.03 | 22.98 | 50.90 | 42.89 | 49.28 | 61.98 | 7.005 |
| 33.98 | 25.38 | 46.53 | 39.70 | 43.68 | 64.77 | 7.094 | LIG | 33.17 | 23.72 | 49.29 | 41.45 | 47.64 | 63.25 | 7.204 |
| 40.13 | 29.07 | 43.88 | 37.61 | 41.95 | 67.42 | 7.418 | ONLINE | 38.72 | 27.65 | 46.46 | 39.77 | 45.23 | 65.57 | 7.462 |
| 37.41 | 26.94 | 43.92 | 37.11 | 41.71 | 67.08 | 7.435 | DFKI$^{SC}$ | 35.95 | 25.57 | 46.49 | 39.41 | 45.18 | 64.92 | 7.483 |

**MT Arabic-English (MT$_{AE}$)**

| *"case+punc" evaluation* | | | | | | | System | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **26.40** | **60.60** | **58.43** | **48.67** | **55.87** | **58.06** | **5.937** | RWTH | **25.08** | **58.27** | **60.94** | **50.37** | **59.40** | **56.64** | **6.001** |
| 24.18 | 58.17 | 62.96 | 51.99 | 60.32 | 56.09 | 5.610 | FBK | 22.52 | 55.54 | 65.98 | 54.63 | 64.24 | 54.12 | 5.570 |
| 19.69 | 55.55 | 64.42 | 52.84 | 60.88 | 54.16 | 5.188 | MIT | 19.34 | 52.18 | 67.33 | 55.01 | 65.26 | 52.34 | 5.309 |
| 19.35 | 54.32 | 68.67 | 56.30 | 64.77 | 52.01 | 5.006 | DCU | 18.88 | 51.88 | 68.46 | 56.65 | 66.46 | 51.45 | 5.177 |
| 23.76 | 58.44 | 62.95 | 52.08 | 59.71 | 55.32 | 5.688 | ONLINE | 22.49 | 56.83 | 64.14 | 52.88 | 62.13 | 54.95 | 5.771 |
| 23.43 | 57.93 | 62.86 | 51.92 | 59.95 | 55.82 | 5.493 | DFKI$^{SC}$ | 22.21 | 55.09 | 65.54 | 54.31 | 63.91 | 53.81 | 5.536 |

**MT Chinese-English (MT$_{CE}$)**

| *"case+punc" evaluation* | | | | | | | System | *"no_case+no_punc" evaluation* | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BLEU | METEOR | WER | PER | TER | GTM | NIST | | BLEU | METEOR | WER | PER | TER | GTM | NIST |
| **16.78** | **53.03** | 67.63 | **54.21** | 63.47 | **52.14** | **4.985** | MSR | **15.30** | **48.95** | 71.46 | **57.23** | 68.45 | **49.19** | **4.969** |
| 15.22 | 50.38 | **66.24** | 54.63 | **62.60** | 51.08 | 4.741 | RWTH | 13.42 | 46.36 | **69.90** | 58.10 | **67.55** | 47.95 | 4.579 |
| 12.52 | 46.66 | 75.37 | 59.86 | 70.26 | 46.26 | 4.317 | DCU | 11.55 | 43.67 | 76.40 | 61.22 | 73.00 | 44.82 | 4.354 |
| 12.15 | 48.69 | 71.71 | 57.07 | 67.15 | 48.81 | 4.497 | NICT | 10.95 | 45.24 | 75.01 | 59.57 | 71.68 | 46.39 | 4.469 |
| 16.37 | 51.89 | 69.54 | 55.18 | 65.34 | 50.55 | 5.100 | ONLINE | 15.19 | 49.75 | 72.57 | 57.03 | 69.11 | 49.13 | 5.130 |
| 17.12 | 53.48 | 68.43 | 54.28 | 64.13 | 52.55 | 5.020 | MSR$^{SC}$ | 15.70 | 49.75 | 71.95 | 57.02 | 68.59 | 49.75 | 5.048 |

# Appendix B. Human Evaluation

## B.1. Ranking

· A subset of 400 test sentences was used to carry out the subjective ranking evaluation.
· The *"All systems"* scores indicate the average number of times that a system was judged better then ($>$) or better/equal to ($\geq$) any other system.
· The *"Head to head"* scores indicate the number of pairwise head-to-head comparisons won by a system.
· The "$^{SC}$" system name suffixes indicate runs submitted to the system combination tasks.
· The "online" system represents a state-of-the-art general-domain MT system.

**SLT English-French (SLT$_{EF}$)**

| System | ALL SYSTEMS $>$ others | ALL SYSTEMS $\geq$ others | System | HEAD-TO-HEAD # wins |
|---|---|---|---|---|
| LIUM | **0.3198** | **0.7678** | KIT | 4 / 4 |
| KIT | 0.3026 | 0.7563 | LIUM | 3 / 4 |
| LIG | 0.2738 | 0.7318 | LIG | 2 / 4 |
| RWTH | 0.2683 | 0.7317 | RWTH | 1 / 4 |
| FBK | 0.2119 | 0.6357 | FBK | 0 / 4 |

**MT English-French (MT$_{EF}$)**

| System | ALL SYSTEMS $>$ others | ALL SYSTEMS $\geq$ others | System | HEAD-TO-HEAD # wins |
|---|---|---|---|---|
| LIMSI | **0.4581** | **0.6225** | LIMSI | 7 / 8 |
| KIT | 0.4153 | 0.5947 | KIT | 6 / 8 |
| DFKI | 0.3859 | 0.5841 | MIT | 4 / 8 |
| MIT | 0.3859 | 0.5703 | DFKI | 3 / 8 |
| RWTH | 0.3775 | 0.5672 | RWTH | 2 / 8 |
| LIG | 0.3456 | 0.5534 | LIG | 1 / 8 |
| FBK | 0.3169 | 0.4975 | FBK | 0 / 8 |
| ONLINE | 0.5866 | 0.7369 | ONLINE | 8 / 8 |
| DFKI$^{SC}$ | 0.3250 | 0.6766 | DFKI$^{SC}$ | 5 / 8 |

**MT Arabic-English (MT$_{AE}$)**

| System | ALL SYSTEMS $>$ others | ALL SYSTEMS $\geq$ others | System | HEAD-TO-HEAD # wins |
|---|---|---|---|---|
| RWTH | **0.4012** | **0.8344** | RWTH | 4 / 5 |
| FBK | 0.3307 | 0.7670 | FBK | 3 / 5 |
| MIT | 0.1491 | 0.5841 | MIT | 1 / 5 |
| DCU | 0.1086 | 0.5065 | DCU | 0 / 5 |
| ONLINE | 0.5030 | 0.8369 | ONLINE | 5 / 5 |
| DFKI$^{SC}$ | 0.2296 | 0.7509 | DFKI$^{SC}$ | 2 / 5 |

**MT Chinese-English (MT$_{CE}$)**

| System | ALL SYSTEMS $>$ others | ALL SYSTEMS $\geq$ others | System | HEAD-TO-HEAD # wins |
|---|---|---|---|---|
| MSR | **0.4250** | **0.7360** | MSR | 3 / 5 |
| RWTH | 0.3330 | 0.6200 | RWTH | 2 / 5 |
| NICT | 0.2920 | 0.5545 | NICT | 1 / 5 |
| DCU | 0.1125 | 0.3605 | DCU | 0 / 5 |
| ONLINE | 0.5995 | 0.8035 | ONLINE | 5 / 5 |
| MSR$^{SC}$ | 0.4135 | 0.7500 | MSR$^{SC}$ | 4 / 5 |

## B.2. Pairwise System Comparisons

The following tables show pairwise comparisons between systems for each task. Wins read column by raw, i.e. the numbers in the table cells indicate the percentage of times that the system in that column was judged to be better than the system in that row. The winner of the two systems is indicated in bold. The difference between 100 and the sum of the complementary cells corresponds to the percentage of ties.

**SLT English-French (SLT$_{EF}$)**

| | FBK | KIT | LIG | LIUM | RWTH |
|---|---|---|---|---|---|
| FBK | - | **35.26** | **34.25** | **41.25** | **34.92** |
| KIT | 21.66 | - | 24.06 | 27.75 | 24.00 |
| LIG | 21.50 | **28.82** | - | **30.08** | 26.88 |
| LIUM | 20.50 | **29.25** | 21.55 | - | 21.55 |
| RWTH | 21.11 | **27.75** | **29.65** | **28.82** | - |
| $>$ | 21.19 | 30.26 | 27.38 | **31.98** | 26.83 |
| $\geq$ | 63.57 | 75.63 | 73.18 | **76.78** | 73.17 |

**MT English-French (MT$_{EF}$)**

| | DFKI | FBK | KIT | LIG | LIMSI | MIT | RWTH | ONLINE | DFKI$^{SC}$ |
|---|---|---|---|---|---|---|---|---|---|
| DFKI | - | 32.00 | **43.00** | 37.25 | **46.50** | **42.50** | 39.50 | **57.00** | **35.00** |
| FBK | **50.25** | - | **49.75** | **43.25** | **53.50** | **48.25** | **50.75** | **64.25** | **42.00** |
| KIT | 42.25 | 33.00 | - | 36.50 | **43.50** | 39.75 | 38.25 | **58.25** | 32.75 |
| LIG | **42.25** | 35.75 | **46.50** | - | **51.00** | **42.50** | **45.00** | **61.50** | **32.75** |
| LIMSI | 35.00 | 33.50 | 41.25 | 32.50 | - | 37.75 | 37.00 | **53.50** | 31.50 |
| MIT | 40.00 | 35.00 | **44.75** | 39.75 | **51.00** | - | 37.75 | **59.00** | **36.50** |
| RWTH | **42.25** | 34.50 | **45.25** | 37.25 | **50.50** | **47.25** | - | **61.25** | **28.00** |
| ONLINE | 28.50 | 25.75 | 26.25 | 25.00 | 30.50 | 27.00 | 26.00 | - | 21.50 |
| DFKI$^{SC}$ | 28.25 | 24.00 | **35.50** | 25.00 | **40.00** | 23.75 | 27.75 | **54.50** | - |
| $>$ | 38.59 | 31.69 | 41.53 | 34.56 | 45.81 | 38.59 | 37.75 | **58.66** | 32.50 |
| $\geq$ | 58.41 | 49.75 | 59.47 | 55.34 | 62.25 | 57.03 | 56.72 | **73.69** | 67.66 |

**MT Arabic-English (MT$_{AE}$)**

| | DCU | FBK | MIT | RWTH | ONLINE | DFKI$^{SC}$ |
|---|---|---|---|---|---|---|
| DCU | - | **55.00** | **29.75** | **57.50** | **62.81** | **41.75** |
| FBK | 08.75 | - | 17.00 | **33.25** | **43.18** | 14.50 |
| MIT | 21.00 | **44.25** | - | **53.75** | **58.04** | **31.00** |
| RWTH | 08.00 | 21.00 | 08.25 | - | **36.59** | 09.00 |
| ONLINE | 07.79 | 19.44 | 11.31 | 24.56 | - | 18.48 |
| DFKI$^{SC}$ | 08.75 | **25.50** | 08.25 | **31.50** | **50.89** | - |
| $>$ | 10.86 | 33.07 | 14.91 | 40.12 | **50.30** | 22.96 |
| $\geq$ | 50.65 | 76.70 | 58.41 | 83.44 | **83.69** | 75.09 |

**MT Chinese-English (MT$_{CE}$)**

| | DCU | MSR | NICT | RWTH | ONLINE | MSR$^{SC}$ |
|---|---|---|---|---|---|---|
| DCU | - | **69.75** | **55.25** | **53.75** | **75.00** | **66.00** |
| MSR | 10.00 | - | 24.00 | 27.75 | **48.75** | **21.50** |
| NICT | 15.25 | **46.50** | - | **42.75** | **67.75** | **50.50** |
| RWTH | 12.75 | **48.00** | 30.75 | - | **56.50** | **42.00** |
| ONLINE | 08.25 | 27.00 | 14.50 | 21.75 | - | 26.75 |
| MSR$^{SC}$ | 10.00 | 21.25 | 21.50 | 20.50 | **51.75** | - |
| $>$ | 11.25 | 42.50 | 29.20 | 33.30 | **59.95** | 41.35 |
| $\geq$ | 36.05 | 73.60 | 55.45 | 62.00 | **80.35** | 75.00 |

# Appendix C. Evaluation Metric Correlation

· The correlation between evaluation metrics is measured using *Spearman's rank correlation coefficient* $\rho \in$ [-1.0, 1.0] with $\rho$ = 1.0 if all systems ranked in same order, $\rho$ = -1.0 if all systems ranked in reverse order and $\rho$ = 0.0 if no correlation exists.

· The automatic evaluation metrics that correlate best with the respective human assessments are marked in boldface.

## C.1. Human Assessment and Automatic Evaluation

| $SLT_{EF}$ | BLEU | METEOR | WER | PER | TER | GTM | NIST |
|---|---|---|---|---|---|---|---|
| **Ranking (>)** | **0.9000** | 0.7000 | -0.8000 | -0.7000 | -0.8000 | 0.6000 | 0.7000 |
| **Ranking (≥)** | **0.9000** | 0.7000 | -0.8000 | -0.7000 | -0.8000 | 0.6000 | 0.7000 |
| **Head-to-Head** | 0.8000 | 0.6000 | **-0.9000** | -0.5000 | **-0.9000** | 0.3000 | 0.5000 |

| $MT_{EF}$ | BLEU | METEOR | WER | PER | TER | GTM | NIST |
|---|---|---|---|---|---|---|---|
| **Ranking (>)** | 0.5125 | 0.3958 | -0.4708 | -0.4708 | -0.5375 | **0.5542** | 0.4042 |
| **Ranking (≥)** | 0.7500 | 0.6667 | -0.7000 | -0.7000 | **-0.7667** | **0.7667** | 0.6000 |
| **Head-to-Head** | 0.8000 | 0.6833 | -0.7333 | -0.7333 | **-0.8167** | 0.7833 | 0.6500 |

| $MT_{AE}$ | BLEU | METEOR | WER | PER | TER | GTM | NIST |
|---|---|---|---|---|---|---|---|
| **Ranking (>)** | 0.8857 | **0.9429** | -0.6571 | -0.6571 | -0.6571 | 0.6571 | **0.9429** |
| **Ranking (≥)** | 0.8857 | **0.9429** | -0.6571 | -0.6571 | -0.6571 | 0.6571 | **0.9429** |
| **Head-to-Head** | 0.8857 | **0.9429** | -0.6571 | -0.6571 | -0.6571 | 0.6571 | **0.9429** |

| $MT_{CE}$ | BLEU | METEOR | WER | PER | TER | GTM | NIST |
|---|---|---|---|---|---|---|---|
| **Ranking (>)** | 0.7143 | 0.7714 | -0.4857 | -0.6571 | -0.4857 | 0.6000 | **0.9429** |
| **Ranking (≥)** | 0.7714 | 0.8286 | -0.4286 | -0.6000 | -0.4286 | 0.6571 | **1.0000** |
| **Head-to-Head** | 0.7714 | 0.8286 | -0.4286 | -0.6000 | -0.4286 | 0.6571 | **1.0000** |