Overview of the IWSLT 2010 Evaluation Campaign

Michael Paul
NICT
Hikaridai 2-2-2,
619-0288 Kyoto, Japan
michael.paul@nict.go.jp

Marcello Federico
FBK
via Sommarive 18,
38100 Trento, Italy
federico@fbk.eu

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

Abstract

This paper gives an overview of the evaluation campaign results of the 7*th* *International Workshop on Spoken Language Translation* (IWSLT 2010)[1]. This year, we focused on three spoken language tasks: (1) public speeches on a variety of topics (*TALK*) from English to French, (2) spoken dialog in travel situations (*DIALOG*) between Chinese and English, and (3) traveling expressions (*BTEC*) from Arabic, Turkish, and French to English. In total, 28 teams (including 7 first-time participants) took part in the shared tasks, submitting 60 primary and 112 contrastive runs. Automatic and subjective evaluations of the primary runs were carried out in order to investigate the impact of different communication modalities, spoken language styles and semantic context on automatic speech recognition (ASR) and machine translation (MT) system performances.

1. Introduction

The *International Workshop on Spoken Language Translation* (IWSLT) is a yearly, open evaluation campaign for spoken language translation. IWSLT's evaluations are not competition-oriented; their goal is to foster cooperative work and scientific exchange. In this respect, IWSLT proposes challenging research tasks and an open experimental infrastructure for the scientific community working on spoken and written language translation.

Previous IWSLT workshops focused on the establishment of evaluation metrics for multilingual speech-to-speech translation, innovative technologies for the translation of automatic speech recognition results from read-speech and spontaneous-speech input, and monolingual and bilingual dialog conversations [1].

This year, the standard BTEC task was provided for the translation of Arabic and Turkish spoken language text into English. For the first time, *French* was used as an input language for the BTEC task, attracting new groups to participate in this year's event.

As a continuation of last years efforts in translating spoken dialog, the DIALOG task focused on task-oriented cross-lingual human dialog in travel situations where the speech data was annotated with dialog and speaker information that could be exploited by the participant to incorporate contextual information into the translation process. For the DIALOG task, IWSLT participants had to translate both the Chinese and the English outputs of the automatic speech recognizers into English and Chinese, respectively.

The new challenge for this year's evaluation campaign was the translation of public speeches from English to French. The TALK task was based on a collection of recordings of public speeches covering a variety of topics, for which high quality transcriptions and translations into several languages are available. This task not only imposes new challenges on the development of MT systems, i.e., on how to deal with unlimited domains, but also on the applicability of standard evaluation protocols for the evaluation of translation results of automatic speech recognition outputs based on reference translations that are segmented differently.

All participants had to submit at least one run (*primary submission*) for each translation task they registered for. The evaluation of the primary runs was carried out using standard automatic evaluation metrics for all translation tasks. In addition to the single-metric scores, all automatic metric scores for the MT output were combined by normalizing each metric score distribution and calculating the average of all the normalized metric scores. Human assessments of translation quality ranking multiple MT systems were also applied for the DIALOG and BTEC tasks. Based on the evaluation results, the impact of different communication modalities (monologue vs. dialog), spoken language (planned vs. spontaneous) and semantic context (open vs. limited) was investigated.

The outline of the IWSLT 2010 evaluation campaign (its translation tasks and evaluation specifications) are descibed in detail in Section 2. The evaluation results are summarized and discussed in Section 3.

2. Outline of IWSLT 2010

This year's IWSLT campaign took place during the period of June-September 2010 and featured six different translation tasks that are summarized in Table 1.

In total, 28 research groups (including 7 first-time participants) from all over the world[2] participated in the event, producing a total of 60 MT engines. Information on the research groups, the utilized translation systems, and transla-

[1] http://iwslt2010.fbk.eu

[2] China: 2, France: 4, Germany: 2, Ireland: 1, Israel: 1, Italy: 1, Japan: 3, Korea: 1, Netherlands: 2, Portugal: 1, Qatar: 1, Singapore: 1, Spain: 3, Tunesia: 1, Turkey: 2, UK: 1, USA: 1

Table 1: Translation Tasks

Task	Translation Direction		Participants
TALK	English-French	TT_{EF}	9
DIALOG	English-Chinese	DT_{EC}	11
	Chinese-English	DT_{CE}	11
BTEC	Arabic-English	BT_{AE}	12
	French-English	BT_{FE}	9
	Turkish-English	BT_{TE}	8

tion task participation is summarized in Appendix A. Most participants used phrase-based and syntax-based statistical machine translation (SMT) systems. However, one example-based MT (EBMT) system and various hybrid approaches combining multiple SMT engines or SMT engines with rule-based (RBMT) systems were also exploited.

A detailed description of the translation tasks and the language resources (*supplied corpora*) which were provided to the participating research groups are given in Section 2.1 (TALK), Section 2.2 (DIALOG), and Section 2.3 (BTEC), respectively. The supplied resources were released to the participants three months ahead of the official run submissions period. The official run submission period was limited to two weeks for the BTEC and DIALOG tasks and five weeks for the TALK task. Run submission was carried out via email to the organizers with multiple runs permitted. However, the participant had to specify which runs should be treated as *primary* (evaluation using human assessments and automatic metrics) or *contrastive* (automatic evaluation only). The organizers set-up online evaluation servers for the TALK develop data sets[3] as well as the testdata sets[4] of the BTEC and DIALOG tasks that could be used by the participants to tune their systems (TALK) or carry out additional experiments after the official run submission period (DIALOG, BTEC). The schedule of the evaluation campaign is summarized in Table 2.

Table 2: Evaluation Campaign Schedule

Event	Date
Training/Develop Corpus Release	May 28, 2010
Evaluation Corpus Release	Aug 23, 2010
Translation Result (BTEC/DIALOG)	Sep 6, 2010
Automatic Evaluation Results	Sep 17, 2010
Translation Results (TALK)	Sep 30, 2010
Subjective Evaluation Results	Nov 12, 2010
Workshop	Dec 2-3, 2010

2.1. TALK Task

The new challenge of this year was the translation of public speeches from English to French. The so-called TALK task was based on the TED[5] talks collection, 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.

The proposed new challenge clearly departs from and completes the application scenarios proposed so far in the IWSLT evaluations. Macroscopic differences between the TALK task and the BTEC and DIALOG tasks are in the assumed communication modality, i.e. monologue vs. dialogue, spoken language style, i.e. planned vs. spontaneous, and semantic context, i.e. open vs. limited.

From a translation point of view, the TALK task is basically a subtitling translation task, in which the ideal translation unit is a single caption as defined by the original transcript. In fact, some word re-ordering across consecutive captions is also permitted in order to accommodate syntactic differences between source and target languages. The wide variety of topics covered by the TED talks has determined the type and volume of training data that has been prepared and released for this challenge. This in fact comprises a small (less than 1 million word) parallel corpus of TED talks and several out-of-domain large parallel corpora including texts from the United Nations, European Parliament, news commentaries, and the Web.

From a speech translation point of view, the problem of processing full transcripts rather than isolated utterances requires handling possible inconsistencies between the speech segmentation introduced by the ASR system and the text segmentation used in the reference transcripts and translations. In particular, this discrepancy impacts when word-graphs produced by the ASR system are used as MT input.

While the significantly larger amount of available training data clearly has an impact on the complexity of the MT systems being developed for the TALK task, the problem of aligning ASR and reference segments also required some important revision of the automatic evaluation method.

These major shifts with respect to the previous IWSLT evaluations are the reasons why we declared that this first evaluation is to be considered as an exercise for establishing reference baselines and appropriate evaluation protocols for future evaluations. Hence, although an evaluation server was set up to compute several translation accuracy metrics, no official ranking of the participants will be reported in this evaluation. As no human evaluation was planned for the new challenge, a different schedule from the other tasks was established in order to ease participation in all of the offered tasks. Before the submission deadline, we received primary submissions by nine teams in total. The majority of these teams also participated in other IWSLT tasks (see Appendix A).

2.1.1. Language Resources

The TALK task is about the translation of speeches taken from the TED website. TED LLC is a nonprofit organization with the declared goal of "disseminating ideas worth spreading". It regularly organizes two annual conferences in the US and one in UK, in which prominent experts from different fields are invited to give short talks about topics relevant

[3] http://isl.ira.uka.de/iwslt2010

[4] https://mastarpj.nict.go.jp/EVAL/IWSLT10/automatic/testset_IWSLT10

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

to the global society. Although TED stands for Technology Entertainment Design, over the years its scope has become much broader, indeed. TED is supported by industrial sponsors and a community of volunteers, which organize similar conferences in other countries of the world and help in creating content for the TED website, namely videos, transcripts and translations of talks. All content is copyrighted and made publicly available under a Creative Commons license. At this time, the TED website hosts around 800 English talks and the TED Open Translation Project has been managing translation of talks into 80 languages. Translations grow at a rapid pace thanks to more than 4,000 volunteers, which have contributed some 12,500 translations in total so far. To ensure quality in the process, all English source transcripts are prepared by professional transcribers and are revised by another translator prior to publication. Languages with the most translated talks at this time are Arabic, Bulgarian, Chinese, French, Italian, Portuguese, and Spanish.

For this IWSLT evaluation, a first parallel corpus of 345 English talks with their French translations was released, named `TED English-French ver.1.1`. The TED corpus and all additional parallel data allowed for this exercise have been made available at the workshop website[6]. Development and test sets, however, were released only to registered participants according to the evaluation schedule. The statistics[7] of the supplied TALK corpus are summarized in Table 3.

Table 3: Supplied Corpus (TALK)

TALK	data	lang	sent	avg.len	word token	word type
train	(text)	E	86,225	9.8	842,125	31,429
	(text)	F	86,225	10.0	867,963	42,599
dev	(speech)	E	1,368	9.5	12,962	2,687
	(text)	F	1,368	9.3	12,712	3,246
test10	(speech)	E	3,584	9.0	32,155	4,153
	(text)	F	3,584	9.2	33,010	5,571

2.1.2. Task Definition

For the TALK task, participants were requested to translate two input conditions: (1) the reference text that was extracted from the subtitles of the TED talks, and (2) the output from an automatic speech recognition (ASR) system run on the audio of the TED talks selected for the evaluation set. The reference texts were in true case and contained punctuation marks. The segmentation was given by the segmentation of the closed captioning of the TED data. The output from the ASR system was case-sensitive, but did not contain any punctuation marks. The segmentation was obtained automatically from the audio data and thus did not match the reference segmentation of the closed captions. The ASR output provided to the participants consisted of the single best output, a 20-best list, and the word lattices from the recognition system provided in standard lattice format (SLF).

[6]http://iwslt2010.fbk.eu/node/27

[7]For details on the additional language resources that were permitted for the TALK task, please refer to http://www.statmt.org/wmt10/translation-task.html.

The ASR system used for producing the automatic transcripts was the 2009 KIT English Quaero Evaluation system with a language model that was updated with the data from the TED training data. In order to measure the ASR performance we took the subtitles of the talks as provided by the TED website and re-annotated the time boundaries of the sentences to exactly match the speech. The ASR system achieved a case-insensitive word error rate (WER) of 26.4% on the TED development set, and 22.3% on the evaluation set, respectively.

The quality of the results of the automatic translation systems was measured with BLEU, NIST, and TER scores (see Table 9) using one reference translation. The reference translation was taken from the translations provided by the TED open translation project. All translations were supposed to be case-sensitive. Also, regardless of whether the reference transcription of the talks, which contain punctuation marks, or the automatic transcriptions, which do not, were used, the translation systems were supposed to produce punctuation marks which were considered in the automatic measures.

Since the reference translations from the TED website match the segmentation of the reference transcriptions of the talks, the scores for the automatic translation results could be directly computed. This was not the case for the translation of the ASR output, as the segmentation of the ASR output does not match that of the reference translation. We therefore used the method and scoring scripts from [2] which align the automatic translation and reference translation based on the Levenshtein distance first and then computes the automatic translation scores.

2.2. DIALOG Task

As a continuation of last year's efforts [1], the DIALOG task focused on the translation of task-oriented human dialog in travel situations. The speech data was recorded through human interpreters, where native speakers of different languages were asked to complete certain travel-related tasks like hotel reservations using their mother tongue. The translation of the freely-uttered conversation was carried out by human interpreters. The obtained speech data was annotated with dialog and speaker information. In total, 11 research groups participated in this year's DIALOG task (see Appendix A).

2.2.1. Language Resources

The DIALOG task was carried out using the Spoken Language Databases (SLDB) corpus, a collection of human-mediated cross-lingual dialogs in travel situations. Similar to last year, bilingual Chinese-English dialogs were provided to the participants for the training of the MT systems. In addition, the Chinese/English parts of the BTEC corpus (see Section 2.3), were provided to the participants of the DIALOG task and could be used as additional training bitext.

Linguistic tools like word segmentation tools, parsers, etc., were allowed to preprocess the supplied corpus, but par-

Table 4: Supplied Corpus (DIALOG)

BTEC	data	lang	sent	avg.len	word token	word type
train	(text)	C	19,972	7.4	148,224	8,408
	(text)	E	19,972	7.7	153,178	7,294
dev	(speech)	C	1,495	9.4	14,002	3,409
	(ref)	E	15,029	10.3	139,212	6,176
	(speech)	E	506	6.2	3,119	840
	(ref)	C	3,542	7.1	25,037	1,665
	(text)	C	1,741	5.5	9,666	2,920
	(ref)	E	20,762	6.8	141,262	6,306

SLDB	data	lang	sent	avg.len	word token	word type
train	(text)	C	10,061	8.9	89,110	3,734
	(text)	E	10,061	11.8	118,648	3,271
dev	(dialog)	C	200	9.3	1,859	377
	(ref)	E	800	9.8	7,829	418
	(dialog)	E	210	11.8	2,474	403
	(ref)	C	840	11.2	9,379	621
	(speech)	C	750	5.1	3,818	633
	(ref)	E	5,208	6.6	33,827	1,387
	(speech)	E	749	5.5	4,146	454
	(ref)	C	5,243	6.5	34,693	1,265
test09	(dialog)	C	405	11.3	4,562	653
	(ref)	E	1,620	13.7	22,253	886
	(dialog)	E	393	11.0	4,321	569
	(ref)	C	1,572	12.0	18,789	875
test10	(dialog)	C	532	8.2	4,361	900
	(ref)	E	2,128	13.3	28,384	1,636
	(dialog)	E	453	11.0	5,004	870
	(ref)	C	1,812	11.2	20,314	1,470

ticipants were asked to declare their usage in the system description paper and to measure the impact of these tools on the system performance. No additional parallel or monolingual corpora or word-lists were permitted to be used for the primary run. However, in order to motivate participants to explore the effects of additional language resources, the organizers also accepted contrastive runs based on additional resources.

Table 4 summarizes the characteristics of the Chinese (C) and English (E) training (*train*), development (*dev*) and evaluation (*test*) data sets. For evaluation purposes, two data sets, i.e., the testset of IWSLT 2009 (*test09*) and the new testset of IWSLT 2010 (*test10*) were used. The first two columns specify the given data set and its type. Besides the source language text ("text") and target language reference translation ("ref") resources, all data sets consist of the ASR output and manual transcriptions of the respective *dialog* or *speech* recordings of language *lang*. The number of sentences are given in the "*sent*" column, and the "*avg.len*" column shows the average number of words per training sentence, where the word segmentation for the source language was the one given by the output of the ASR engines without punctuation marks. "*word token*" refers to the number of words in the corpus and "*word type*" refers to the vocabulary size.

For the automatic evaluation of development data sets, 7 (16) reference translations for the SLDB (BTEC) were also included in the supplied corpus. For the DIALOG testset data sets, up to 4 reference translations were available.

2.2.2. Task Definition

For the DIALOG task, participants were asked to translate two input conditions: (1) the automatic speech recognition (ASR) outputs, i.e., word lattices (SLF), N-best lists (NBEST) and 1-best (1BEST) speech recognition results, and (2) the correct recognition results (CRR), i.e., text input without speech recognition errors.

For both input conditions, the input text contained neither case nor punctuation information. However, the reference translations were in true case and contained punctuation marks. Therefore, the participants had to recover case/punctuation information for the MT output run submissions. Instructions[8] on how to build a baseline tool for case/punctuation insertions using the *SRI Language Modeling Toolkit* [3] was also provided.

Participants of the DIALOG task had to translate both the English ASR outputs into Chinese and the Chinese ASR outputs into English, whereby they could choose the ASR output condition (SLF, NBEST, or 1BEST) that best suits their MT system. Translation of the CRR text input was mandatory for all participants.

The ASR systems used to create the ASR outputs were the Chinese and English ATRASR systems provided by NICT [4]. The recognition accuracies for the DIALOG testdata sets are summarized in Table 5. Besides the ASR output files (lattices, 20-BEST and 1-BEST lists), tools to extract larger NBEST lists were also provided to the participants.

Table 5: Speech Recognition Accuracy (DIALOG)

DIALOG	lang	word (%)		sentence (%)	
		lattice	1BEST	lattice	1BEST
test09	C	92.67	81.46	64.63	39.12
	E	89.58	82.20	50.13	37.15
test10	C	89.36	83.29	61.60	54.64
	E	89.06	81.11	44.30	34.81

2.3. BTEC Task

In order to (1) enable small groups and even newcomers to the field of machine translation to join the evaluation campaign and (2) provide a testbed for new ideas for spoken language translation techniques, a BTEC translation task focusing on frequently used utterances in the domain of travel conversations was provided for the translation of Arabic (A), French (F) and Turkish (T) spoken language text into English (E). In total, 20 research groups took part in at least one of the three BTEC translation tasks, submitting 12 primary runs for Arabic-English, 9 primary runs for French-English, and 8 primary runs for Turkish-English.

2.3.1. The BTEC Corpus

The BTEC task was carried out using the Basic Travel Expression Corpus (BTEC), a multilingual speech corpus con-

[8] http://mastarpj.nict.go.jp//IWSLT2009/downloads/case+punc_tool_using_SRILM.instructions.txt

Table 6: Supplied Corpus (BTEC)

BTEC	data	lang	sent	avg.len	word token	word type
train	(text)	A	19,972	8.0	158,926	18,154
	(text)	F	19,972	9.5	189,665	10,735
	(text)	T	19,972	7.0	139,514	20,106
	(text)	E	19,972	9.1	182,627	8,344
dev	(text)	A	2,508	6.3	15,797	3,875
	(ref)	E	35,238	8.1	284,612	5,609
	(text)	F	1,512	7.5	11,409	2,244
	(ref)	E	24,192	8.1	196,806	4,660
	(text)	T	1,006	5.7	5,766	2,083
	(ref)	E	16,096	8.1	130,518	3,712
test09	(text)	A	469	6.1	2,875	1,099
	(text)	F	469	7.8	3,642	976
	(text)	T	469	5.8	2,741	1,115
	(ref)	E	3,283	8.4	27,507	1,739
test10	(text)	A	464	6.4	2,953	1,180
	(text)	F	464	7.7	3,582	1,004
	(text)	T	464	5.8	2,710	1,149
	(ref)	E	3,248	8.4	27,183	1,580

taining tourism-related sentences similar to those that are usually found in phrasebooks for tourists going abroad.

The participants were supplied with a training corpus of 20K sentence pairs which covered the same sentence IDs for all translation directions. In addition, the testsets of previous IWSLT evaluation campaigns were also provided to the participants and could be used to improve the MT system performance for the respective translation tasks. In contrast to the DIALOG task, the supplied corpus of the BTEC task was in true case and contained punctuation marks. The corpus statistics are summarized in Table 6.

2.3.2. Task Definition

The translation input condition of all BTEC tasks consisted of correct recognition results, i.e., text input, for Arabic, Turkish, and French. The target language for all BTEC tasks was English. The monolingual and bilingual language resources that were allowed for training the translation engines for the primay runs were limited to the supplied corpus. All other BTEC language resources besides the ones for the given language pair were treated as additional language resources.

Similar to the DIALOG task, the evaluation specifications for the BTEC task were defined as case-sensitive with punctuation marks (*case+punc*). Tokenization scripts were applied automatically to all run submissions prior to evaluation. In addition, automatic evaluation scores were also calculated for case-insensitive (lower-case only) MT outputs with punctuation marks removed (*no_case+no_punc*).

2.4. Evaluation Specifications

In this section, we summarize the subjective and automatic evaluation metrics used to assess the translation quality of the primary run submissions.

2.4.1. Subjective Evaluation

Human assessments of translation quality were carried out using the *Ranking* metrics. For the *Ranking* evaluation, human graders were asked to *"rank each whole sentence translation from Best to Worst relative to the other choices (ties are allowed)"* [5]. The *Ranking* evaluation was carried out using a web-browser interface and graders had to order up to five system outputs by assigning a grade between 5 (*best*) and 1 (*worse*). This year's evaluations were carried out by paid evaluation experts, i.e., three graders for each of the target languages. The *Ranking* scores were obtained as the average number of times that a system was judged better than any other system. In addition, normalized ranks (*NormRank*) on a per-judge basis using the method of [6] were calculated for each run submission. The *Ranking* metric was applied to all submitted primary runs of all translation tasks.

Similar to last year's IWSLT edition [1], the difference of each MT system and the system that obtained the highest *Ranking* score (*BestRankDiff*) was calculated and used to define an alternative method to rank MT systems of a given translation task.

In addition, human assessments of the overall translation quality of a single MT system were carried out with respect to the *Fluency* and *Adequacy* of the translation. *Fluency* indicates how the evaluation segment sounds to a native speaker of the target language. For *Adequacy*, the evaluator was presented with the source language input as well as a "gold standard" translation and had to judge how much of the information from the original translation was expressed in the translation [7]. The *Fluency* and *Adequacy* judgments consisted of one of the grades listed in Table 7. The evaluation of both metrics, *Fluency* and *Adequacy*, was carried out separately using a web-browser tool. For each input sentence, the MT translation outputs of the respective systems were displayed on one screen and judgments were done by selecting one of the possible grades for each MT output.

In addition to the above standard metrics, a modified version of the *Adequacy* metrics (*Dialog*) that takes into account information beyond the current input sentence was applied to the translation results of the DIALOG task in order to judge a given MT output in the context of the respective dialog. For the *Dialog* assessment, the evaluators were presented with the history of previously uttered sentences, the input sentence and the "gold standard" translation. The evaluator had to read the dialog history first and then had to judge how much of the information from the reference translation is expressed in the translation in the context of the given dialog history by assigning one of the *Dialog* grades listed in Table 7. In cases where parts of the information were omitted in the current translation, but they could be understood in the context of the given dialog, such omission would not result in a lower *Dialog* score.

Due to high evaluation costs, the *Fluency*, *Adequacy*, and *Dialog* assessments were limited to the top-ranked MT system for each translation task according to the *Ranking* evaluation results. In addition, the translation results of each translation task were *pooled*, i.e., in cases of identical translations of the same source sentence by multiple engines, the pooled

Table 7: Human Assessment

Fluency	
5	Flawless Chinese/English
4	Good Chinese/English
3	Non-native Chinese/English
2	Disfluent Chinese/English
1	Incomprehensible

Adequacy / Dialog	
5	All Information
4	Most Information
3	Much Information
2	Little Information
1	None

translation was graded only once, and the respective rank was assigned to all MT engines with the same output.

For the final metric scores, each system score is calculated as the *median* of the assigned grades. All paid graders took part in a dry-run evaluation exercise prior to this year's evaluation period in order to get used to the evaluation metrics as well as the browser-based graphical user interfaces.

2.4.2. Grader Consistency

In order to investigate the degree of grading consistency between the human evaluators, we calculated *Fleiss' kappa coefficient* κ, which measures the agreement between two raters who each classify N items into C mutually exclusive categories taking into account the agreement occurring by chance. It is calculated as:

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

where $Pr(a)$ is the relative observed agreement among graders, and $Pr(e)$ is the hypothetical probability of chance agreement. 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 $\kappa \leq 0$. The interpretation of the κ values according to [8] is given in Table 8.

Table 8: Interpretation of κ Coefficient

κ	**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

2.4.3. Automatic Evaluation

The automatic evaluation of run submissions was carried out using the standard automatic evaluation metrics listed in Table 9. Besides the NIST metrics, all automatic evaluation metric scores listed in Appendix C are given as percent figure (%). For the DT_{EC} translation task, F1 scores calculated based on the *unigram precision* and *recall* system-level figures of each MT systems are used instead of the METEOR metric scores.

In addition to the single-metric scores of each MT output, the average of all automatic evaluation scores (**z-avg**) is calculated as follows. In the first step, all metric scores are normalized so that the score distribution of the respective metric has a zero mean and unit variance (*z-transform*). In the second step, the obtained z-scores of a given MT system are averaged to obtained the final *z-avg* system score [1].

Table 9: Automatic Evaluation Metrics

BLEU:	the geometric mean of n-gram precision by the system output with respect to reference translations. Scores range between 0 (worst) and 1 (best) [9] → 'mteval-v13a.pl'
NIST:	a variant of BLEU using the arithmetic mean of weighted n-gram precisions. Scores are positive with 0 being the worst possible [10] → 'mteval-v13a.pl'
METEOR:	calculates unigram overlaps between a translation and reference texts taking into account various levels of matches (*exact*, *stem*, *synonym*). Scores range between 0 (worst) and 1 (best) [11] → 'meteor-v1.0'
GTM:	measures the similarity between texts by using a unigram-based F-measure. Scores range between 0 (worst) and 1 (best) [12] → 'gtm-v1.4'
WER:	*Word Error Rate*: the edit distance between the system output and the closest reference translation. Scores are positive with 0 being the best possible [13]
PER:	Position independent word error rate: a variant of WER that disregards word ordering [14]
TER:	*Translation Edit Rate*: a variant of WER that allows phrasal shifts [15] → 'tercom-0.7.25'

2.4.4. Statistical Significance of Evaluation Results

In order to decide whether the translation output on the document-level of one MT engine is significantly better than another, we used the *bootStrap* 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 [16]. In this year's evaluation, 2000 iterations were used for the analysis of the automatic evaluation results.

2.4.5. Correlation between Evaluation Metrics

Correlations between different metrics were calculated using the *Spearman rank correlation coefficient* ρ 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.

3. Main Findings of IWSLT 2010

The subjective evaluation results of IWSLT 2010 are summarized in Appendix B. In addition to the MT outputs provided by the participants, the organizers used an online MT server to translate the testset data sets. 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 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. Section B.1 illustrates the overall performance of the best MT system for each translation task and the online system in terms of the *Fluency*, *Adequacy*, and *Dialog* metric scores. The *Ranking*, *NormRank*, and *BestRankDiff* metric results of all MT systems participating in the DIALOG and BTEC tasks are given in Section B.2 and Section B.3, respectively.

The automatic evaluation results of two testsets (*test*10 and *test*09) are given in Appendix C for two different subsets of the evaluation data: (1) the subset of testset sentence IDs used for human assessment where the scores are given as the mean score of the signficance test described in Section 2.4.4 (see Section C.1), and (2) the full testset translated by the participants where the scores were obtained by the online evaluation server[9] (see Section C.2). The MT systems are ordered according to the *z-avg* score of the metric combination for the *case+punc* evaluation specifications that achieved the highest rank correlation coefficients toward the subjective *Ranking* evaluation metric. If system performances *do not* differ significantly according to the *bootStrap* method, horizontal lines between two MT engines in the MT engine ranking tables are omitted. For each translation task, the highest (lowest) scores of the respective evaluation metric are highlighted in **boldface** (*italics*).

Finally, Appendix D summarizes the rank correlation coefficients of subjective and automatic evaluation results.

3.1. TALK Task

This section summarizes the main features of the systems that have been developed by the nine participants of the TALK task. This information is derived from the system descriptions provided by each team.

All participants approached the exercise with phrase-based statistical MT relying on linear combination of feature functions. In particular, seven teams employed the Moses decoder, *limsi* used an *n*-gram decoder, and *kit* used an in-house phrase-based decoder.

The best, median and worst BLEU scores of the primary submissions are given in Table 10 for two evaluation specifications: (1) case-sensitive with punctuations tokenized (*"case+punc"*) and (2) case-insensitive with punctuations removed (*"no_case+no_punc"*).

Table 10: Automatic Evaluation (TALK)

Eval Spec.	Input Cond.	BLEU (%) Best	Median	Worst
case+punc	Text	29.90	25.02	24.24
	ASR	16.34	15.68	12.13
no_case+no_punc	Text	29.98	26.42	23.32
	ASR	20.27	18.75	16.43

The important difference between the best scores achieved in the text and ASR conditions can be explained by the relatively high word-error-rate in the ASR transcripts, i.e., more than one word in every five was wrongly transcribed, and the impact of missing punctuation and letter case information in the ASR transcripts. The specific impact of speech recognition errors can be in part measured by comparing the scores computed on rich and on plain outputs; that is, BLEU scores considering letter case and punctuation (*case+punc*) versus BLEU scores disregarding such information (*no_case+no_punc*). While under the text input condition there is basically no difference between the rich and plain output evaluation, a more significant difference is observed under the ASR condition: the best system score changes from 16.34 to 20.27. As the best scores of the text-plain and ASR-plain conditions are from the same system, we can infer that the impact of speech recognition errors on BLEU scores is around 32%.

Concerning the systems that were developed for the evaluation, most of the participants focused on data filtering, data selection, and model adaptation. The reason for this in general was to find effective ways to make use of the large amount of out-of-domain parallel data that was provided. These approaches account indeed for most of the improvements claimed by the participants over their baselines.

Data filtering methods to extract reliable parallel data from the training data were reported by *fbk*, *kit*, *lium*, and *tubitak*. Data selection to extract parallel data relevant or close to the TALK task was applied by *fbk*, *iti-upv*, *lium*, and *ntt*. Model combination techniques were applied to the language and translation models to weight the contributions of different data sources. In particular, LM interpolation was applied by *fbk*, *kit*, *lig*, *limsi*, *lium*, *mit* and *ntt*. Interpolation of translation models was applied by *fbk* and *mit*. In contrast, *kit* applied a fixed combination scheme to merge two phrase tables. Parameter tuning of the scoring functions in the log-linear model was performed mostly with MERT, with the exception of *mit*, which also reports results with the MIRA algorithm, and *iti-upv*, which compares MERT with a new Bayesian adaptation method.

Concerning the introduction of novel feature functions, *kit* integrated a bilingual LM in its phrase-based decoder, and *limsi* introduced a re-ordering POS-based LM in its n-gram model. System combination was applied by only *mit*, while the use of additional resources (Wikipedia) for language modeling was only explored by *limsi*.

[9] *test*10: https://mastarpj.nict.go.jp/EVAL/IWSLT10/automatic/testset_IWSLT10
*test*09: https://mastarpj.nict.go.jp/EVAL/IWSLT10/automatic/testset_IWSLT09

Work to cope with issues related to the ASR input condition ranged from pre/post-processing methods that handle true casing and punctuation (all participants) up to training a specific MT system that processes ASR word-graphs (*lium*). From the reports of the participants, it seems that developing specific systems for each input condition definitely has rewards in terms of performance.

3.2. DIALOG Task

For the DIALOG task, eleven primary run submissions were submitted. Five participants (*inesc-id*, *postech*, *tubitak*, *uva-illc*, *uva-isca*) employed a single-engine phrase-based SMT approach based on the Moses decoder to translate the bilingual task-oriented human dialogs between Chinese and English. However, the majority of the participants (*dcu*, *i2r*, *ict*, *iti-upv*, *msra*, *nict*) made use of a hybrid MT system architecture combining two or more phrase-based SMT (*PBSMT*), hierarchical phrase-based SMT (*HPBSMT*), or syntax-based SMT (*SBSMT*) engines.

In particular, a standard phrase-based SMT system based on the Moses toolkit was combined with (a) an ITG-based SBSMT system by *iti-upv*, (b) an in-house PBSMT (*Lavender*) and SBSMT system (*Tranyu*) by *i2r*, (c) two HPBSMT systems (SAMT, CCG-based SMT) by *dcu*, (d) two in-house SBSMT systems (*SuperSilenus*, *TemBruin*) and a HPBSMT system (*John*) by *ict*, (e) in-house implementations of PBSMT and HPBSMT, two extended versions of the previous systems using a dependency tree language model, an SBSMT system, and a Treelet-based SMT system by *msra*. *Nict* combined in-house implementations of a PBSMT system (*CleopATRa*) and a HPBSMT system (*Linparse*).

Concerning system combination, median string computation (*iti-upv*). rescoring of combined n-best lists (*i2r*, *ict*, *nict*), and confusion network decoding techniques (*dcu*, *i2r*, *msra*, *nict*) were used. Moreover, new techniques investigated by this year's participants to improve system performance on the DIALOG task include: (1) the paraphrasing of the training data to adddress the data sparseness problem (*dcu*), (2) the handling of ASR errors using word-to-pinyin conversion (*ict*), confusion network decoding (*iti-upv*) and reranking of ASR output prior to decoding (*msra*), (3) the integration of multiple segmentation schemes for Chinese (*ict*, *nict*, *postech*), (4) the source language side re-odering via tree induction (*uva-illc*), (5) the combination of multiple word alignment methods (*i2r*, *insec-id*, *msra*), (6) the incorporation of syntactic constraints (*dcu*, *i2r*, *msra*), and (7) the exploitation of contextual information of the given dialog (*nict*, *uva-isca*). Experiments involving additional resources beyond the supplied corpus were conducted by *iti-upv*.

The human assessment results for the IWSLT 2010 DIALOG testset based on the system ranking evaluation are summarized in Appendix B.2 for all participating MT systems. The *NormRank* scores achieved for the CRR input condition are much higher than the ones obtained for the translation of the ASR output for both translation directions. Moreover, the translation quality of the English-Chinese (EC) MT systems is higher than the Chinese-English (CE) MT systems for the majority of the participating teams.

Comparing the *Ranking* and *NormRank* results, quite different MT system rankings are obtained for DT_{CE}, especially in the case of the ASR output translation task. In contrast, the DT_{EC} systems are ranked very similarly with minor differences for systems in the mid-range. Both metrics, however, agree at least on the top-ranked MT system for both translation directions.

The more stable rankings for the DT_{EC} vs. the DT_{CE} systems and the CRR vs. the ASR input condition indicate that the reliabilty of human assessment grading depends to some extend on the overall translation quality of the MT system outputs. For humans, it is more difficult to distinguish between MT systems with relatively lower translation quality, but it is easier to identify the best performing systems.

As an alternative ranking method, we investigated the gain that the best performing system achieved over the other systems on sentence-level. For each MT system, we calculated the ratio of translations that were ranked worse and those that were ranked better than the top-ranked system for a subset of around 300 translations where both systems were judged together. The results summarized in Appendix B.3 show that much smaller gains were achieved by the best system for the ASR output condition (DT_{CE}: 18%∼43%, DT_{EC}: 6%∼40%) compared to the CRR translation results (DT_{CE}: 35%∼68%, DT_{EC}: 37%∼71%). Moreover, the difference was much lower for DT_{EC} than for DT_{CE}.

The MT systems ranked most consistently for both translation directions are *ict*, *msra*, and *nict*. In addition, much higher ranks for the ASR vs. the CRR input condition were achieved by *i2r* and *postech* for DT_{CE} and *iti-upv* for DT_{EC}.

In order to get an idea of the absolute translation quality of this year's participating MT systems, *Fluency/Adequacy* (isolated sentences) and *Dialog* (within the context of the given dialog) assessments were carried out for the best ranked *ict* system and the *online* system outputs.

The results listed in Appendix B.1 confirm that the translation quality of the DT_{EC} systems is much higher than the DT_{CE} systems for both input conditions and all subjective evaluation metrics (Fluency: +0.45∼0.57, Adequacy: +0.40∼0.81, Dialog: +0.29∼0.62). However, the Fluency/Adequacy scores are relatively low for the ASR input condition (2.4/2.9 out of 5) for both translation directions and the DT_{CE} translations of the correct recognition results (2.9 out of 5). On the other hand, the best DT_{EC} system achieved moderate scores of 3.6/3.7. Moreover, the lower human assessment scores of the *online* system for both translation directions indicate that current state-of-the-art general-domain MT system have difficulties in handling ill-formed inputs like noisy speech (ASR errors) or spontaneous language styles (ungrammtical constructions).

Comparing the Adequacy and Dialog results, consistenly higher scores (DT_{CE}: +0.3, DT_{EC}: +0.2∼0.3) for both input

conditions were achieved when the context of the dialog was taken into consideration. This confirms the findings of last year's evaluation campaign on the same task (DT_{CE}: +0.3, DT_{EC}: +0.1) and indicates that much information necessary to understand a given translation is provided by the history of previously uttered sentences. Therefore, evaluation metrics for the translation of task-oriented dialogs should not be carried out on a sentence-by-sentence basis, but within the context of the given dialog.

The automatic evaluation results confirm the findings of the subjective evaluation, i.e., the scores for the noisy ASR input condition are lower than for the CRR inputs for both translation directions. However, the differences are much larger for the DT_{EC} task (BLEU: +7.1, TER: -6.8, GTM: +6.7) compared to the DT_{CE} task (BLEU: +2.1, TER: -1.9, GTM: +3.4), indicating a higher negative impact of ASR errors on the translation of English input sentences. Looking at the speech recognition results listed in Table 5, we can see that the word-level recognition accuracies of English and Chinese ASR engines are quite similar for both lattices and 1BEST recognition results. However, on sentence-level, the accuracy figures for English (lattice: 44.3%, 1best: 34.8%) are far worse than those for Chinese (lattice: 61.6%, 1best: 54.6%), which underlines the importance of handling ASR recognition errors in the context of the whole input sentence and the preceeding dialog. More gains are to be expected for dealing with n-best lists or even lattice input than single best recognition hypotheses.

In addition, we compared the automatic evaluation results obtained for the *test*09 testset for all participants that took part in both the 2009 (see [1], Appendix D.2) and the 2010 (see Appendix C.2.2) evaluation campaigns. The results showed that the majority of systems were able to improve their system performance for all automatic evaluation metrics based on last year's experiences, thus confirming the progress over time made by the best performing systems.

3.3. BTEC Task

A total of 29 MT engines were developed by the 20 participants of the BTEC tasks, with 12, 9, and 8 primary run submissions for the translation of Arabic, French and Turkish spoken language text into English, respectively.

The majority of the participants (14 teams) focused on phrase-based SMT approaches. In addition, an example-based approach was used by *tau* and an n-gram-based SMT approach was used by *dsic-upv*. *Tottori* combined a pattern-based MT approach with a standard phrase-based MT approach. Moreover, hybrid MT approaches combining phrase-based and hierarchical phrase-based SMT systems were investigated by *dcu*, *lig*, and *rwth*. Besides the Moses decoder, in-house phrase-based SMT decoders were employed by *apptek*, *kit*, and *nict*. An open-source hierarchical phrase-based SMT system (*Jane*) was used by *rwth*.

One of the main points of interest of this year's BTEC task was the identification of word segmentation that helps improve translation performance. Especially for Arabic, many segmentation schemes were explored, including BAMA (*qmul*, *tau*), MADA (*dcu*), ASVM (*lig*), AMIRA (*fbk*) and several in-house segmenters (*greyc*, *miracl*). In addition, the integration of multiple segmentation schemes into the translation process was investigated by *apptek* and *rwth*. For Turkish, the MORFESSOR segmentation toolkit was used by several participants including *qmul* and *tubitak*, and in-house segmenters were applied by *apptek*, *fbk*, and *limsi*. In contrast, the morphological analysis of the French input data was limited to simple tokenization preprocessing for most of the submitted primary runs. However, a stemming approach to reduce the data sparseness problem was applied by *kit* and a learning approach focusing on collocation segmentations was investigated by *upc*.

Other techniques exploited by the participants to improve system performance on the BTEC task include: (1) the combination of multiple word alignment methods (*apptek*, *mit*), (2) a phrase training method using forced alignment (*rwth*), (3) the incorporation of neural network language models (*dsic-upv*), (4) the application of new reordering models covering part-of-speech-based reordering (*apptek*, *kit*), short distance morpheme reordering (*limsi*) and dynamic distortion (*qmul*), (5) the incorporation of syntactic constraints (*rwth*), (6) the handling of unknown words (*apptek*, *limsi*, *qmul*), and (7) system combination techniques based on confusion network decoding (*lig*, *mit*, *rwth*).

The human assessment results for the BTEC task are summarized in Appendix B. The Ranking results listed in Appendix B.2 showed that the *online* system slightly outperformed the participating MT systems. This indicates the potential of using general-domain MT systems for the translation of spoken language input text that does not contain recognition errors. However, the gains were quite small, despite the fact that the MT systems of the IWSLT participants were trained on only 20k bitext.

Fluency/Adequacy grades were obtained for the best ranked MT systems of each translation task (*apptek* for BT_{AE}, *dsic-upv* for BT_{FE}, and *tubitak* for BT_{TE}) and the *online* system. The highest scores were achieved for BT_{FE} (4.0/4.3 out of 5), followed by BT_{TE} (3.7/4.0 out of 5), and BT_{AE} (3.3/3.6 out of 5).

The pair-wise comparison of each MT system with the *online* system listed in Appendix B.3 revealed that the difference in translation performance at sentence level for the BT_{FE} task is very small, i.e., less than 9% of the testset sentences were translated better by the *online* system compared to the majority of the participant's MT systems. In particular, around 30% of the testset sentences were ranked equally, 31.6%∼35.5% were ranked worse, and 35.3%∼40.0% were ranked better. The gains for the BT_{TE}task are slighly higher, i.e. up to 18% of the testset sentences. For the BT_{AE}, however, the *online* system is outperformed by the *tubitak* and the *mit* systems, gaining +6.6%/+2.3%, respectively.

Similar to the DIALOG task, the comparison of the auto-

matic evaluation scores obtained for the *test*09 evaluation data set that were submitted by participants who also took part in last year's shared task also confirmed that progress is being made over time for the BTEC task. For the BT_{AE} task, signficant gains (BLEU: +1.1, METEOR: +0.8, TER: -0.2, GTM: +1.0, NIST: +0.5) were achieved by the *mit* system against last year's best system, a joined submission of *mit* and *tubitak*. For the BT_{TE} task, last year's best performing system combination (*mit*+*tubitak*) would not be outperformed by this year's participants. However, the individual system performance of the *tubitak*, *fbk*, and *apptek* systems improved by +1.0∼8.7% BLEU, +4.2∼8.4% METEOR, -0.3∼2.1% TER, +0.2∼4.2% GTM, and +0.2∼0.6 NIST points.

3.4. Evaluation Metric Correlation

In order to get an idea of how closely the human assessment and automatic evaluation metrics are related, the *Spearman rank correlation coefficients* are summarized in Appendix D.

For each translation task, the MT system ranking obtained for the subjective *Ranking*, *NormRank*, *BestRankDiff* metrics and all investigated automatic evaluation metrics including the *z-avg* metric combination method are compared. For the DIALOG task, the correlation coefficients for ASR and CRR translation results are calculated separately.

The results show that the highest correlation to subjective evaluation metrics is obtained for the *z-avg* metric for the majority of the investigated translation tasks. In contrast to last year's evaluation campaign where the *z-avg* score was calculated as the average of all investigated automatic evaluation metrics, this year we calculated the *z-avg* score for all possible combinations and selected the metric subset that achieved the highest correlation for each translation task separately. The selected metric combinations are summarized in Appendix D.

However, the optimal subset and correlation coefficients largely depend on the translation task. For the DT_{CE} and DT_{EC} tasks, the highest correlation was achieved for the *BestRankDiff* (*Ranking*) metric when the *online* system is included in (excluded from) the MT system rankings. For the BTEC tasks, in general, the *z-avg* scores correlates best with the *NormRank* metric.

3.5. Grader Consistency

Each sentence was evaluated by three human judges. Due to different levels of experience and background of the evaluators, variations in judgments were to be expected. Besides the *inter-grader* consistency, we also calculated the *intra-grader* consistency using 100 randomly selected evaluation pages that had to be graded a second time. Concerning the *intra-grader* and *inter-grader* consistencies, the κ coefficients are given in Table 11.

The obtained overall *intra-grader* κ coefficients were high. Substantial agreement coefficients were obtained for the *Ranking* metrics for all translation tasks. Concerning the human assessment in terms of *Fluency*/*Adequacy*, substantial agreement was achieved for all BTEC tasks and moderate agreement for the DIALOG tasks.

Table 11: Grader Consistency

Metric	Intra-Grader κ			Inter-Grader κ		
	DT_{EC}	DT_{CE}	BT_{*E}	DT_{EC}	DT_{CE}	BT_{*E}
Ranking	0.69	0.66	0.78	0.51	0.43	0.59
Fluency	0.61	0.54	0.75	0.34	0.27	0.47
Adequacy	0.62	0.57	0.60	0.39	0.18	0.39
Dialog	0.60	0.52	–	0.38	0.26	–

Concerning the *inter-grader* consistency, the κ coefficients are much lower for the *Fluency*/*Adequacy*/*Dialog* metrics achieving only fair agreements for the DIALOG tasks and moderate agreement for the BTEC tasks. However, Moderate ro substantial agreements were achieved for the *Ranking* metrics resulting in a high reliability of this years human assessment results.

4. Conclusions

This year's workshop provided a testbed for verifying the quality of state-of-the-art speech-to-speech translation technologies for the translation of different communication modalities, spoken language styles and semantic context.

The standard BTEC task of IWSLT 2010 focused on the translation of frequently used utterances in the domain of travel conversations from Arabic, French, and Turkish into English. The analysis of the 29 MT system results submitted by 20 teams showed that even in a resource-limited setting, good translation performances can be achieved for the BTEC task, providing a valuable testbed to investigate new ideas for spoken language translation techniques. Due to the similarity of the source and target language, the French-English task proved to be the easiest task, achieving the highest subjective and automatic evaluation scores. For Turkish and Arabic, word segmentation issues seem to be crucial in order to deal with the significant amount of unknown words contained in this year's testset and to achieve high quality translation performance.

The DIALOG task was a repetition of last year's Challenge Task. The participants had to translate a collection of task-oriented dialogs in travel situations for both translation directions (Chinese-English and vice versa) using two input conditions, i.e., automatic speech recognition outputs containing recognition errors and text input without speech recognition errors. The automatic and subjective evaluation of the 11 primary run submissions of the IWSLT 2010 testset resulted in lower scores compared to last year's testset due to the higher translation complexity of the IWSLT 2010 testset. However, the comparison of the system outputs of last year's testset submitted by participants that took part in both the 2009 and 2010 evaluation campaigns showed an improvement in automatic evaluation scores, indicating the progress made over time on the DIALOG task. Many new techiques to improve translation quality were investigated in this year's shared task, including the paraphrasing of training data to

reduce the data sparseness problem, the integration of multiple word segmentation schemes to reduce the problem of unknown words, the handling of ASR errors to cope with noisy input, the introduction of syntactic constraints into hierarchical system to improve grammaticality of the MT output, and system combination techniques to overcome short comes of specific machine translation approaches. In addition, the application of a new evaluation metric taking into account information beyond the current input sentence to judge the quality of a translation in the context of a dialog resulted in new insights into the requirements of the translation and evaluation of human conversations that will help to advance the current state of the art in speech-to-speech translation.

This year for the first time, we ran an evaluation exercise on the translation of talks from English to French. The TALK task was organized around a collection of recordings, transcriptions, and translations of real public speeches covering a variety of topics. In addition to domain-specific parallel data, participants could try to exploit a fairly large amount of out-of-domain training data. The task required translating both manually and automatically generated transcripts. ASR transcripts were provided by the task organizers. Nine teams participated in this first exercise and our analysis of the results confirmed for us that the proposed task is definitely sound, original, interesting, and sufficiently complex. Future work will consider improving the ASR input condition by providing automatic transcripts of better quality and in larger quantity.

5. Acknowledgements

The authors would like to thank all the people involved in the preparation of this workshop and the subjective evaluation task. In particular, we would like to thank Christian Girardi for collecting the TED training data, and Jan Niehues and Teresa Hermann for their help in collecting the TED development and evaluation data and in setting up the evaluation server. Thanks also go to Kevin Kilgour for providing the language model for the ASR system for the TED task and Hermann Ney for providing us with the RWTH sentence segmentation tool for scoring the ASR output of the TALK task. In addition, we'd like to thank Shigeki Matsuda for preparing the speech data sets and generating the ASR outputs of the DIALOG task. Special thanks to the TUBITAK-UEDIN and CEA teams, for providing us with the Turkish and French data sets and to Chris Callison-Burch for letting us use the browser-interface scripts of the subjective *Ranking* metrics. In addition, we thank all the paid experts and volunteers who carried out the human assessment of the translation outputs. We also thank the program committee members for reviewing a large number of MT system descriptions and technical paper submissions. Last, but not least, we thank all the research groups for their active participation in the IWSLT evaluation campaign and for making the IWSLT workshop a success.

6. References

[1] M. Paul, "Overview of the IWSLT 2009 Evaluation Campaign," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 1–18.

[2] E. Matusov, G. Leusch, O. Bender, and H. Ney, "Evaluating Machine Translation Output with Automatic Sentence Segmentation," in *Proc. of IWSLT*, Pittsburgh, PA, 2005, pp. 148–154.

[3] A. Stolcke, "Srilm: an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, Denver (Colorado), 2002.

[4] S. Matsuda, T. Jitsuhiro, K. Markov, and S. Nakamura, "ATR Parallel Decoding Based Speech Recognition System Robust to Noise and Speaking Styles," *IEEE Transactions on Information and Systems*, vol. E89-D(3), pp. 989–997, 2006.

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

[6] J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing, "Confidence estimation for statistical machine translation," in *Final Report of the JHU Summer Workshop*, 2003.

[7] J. S. White, T. O'Connell, and F. O'Mara, "The ARPA MT evaluation methodologies: evolution, lessons, and future approaches," in *Proc of the AMTA*, 1994, pp. 193–205.

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

[9] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of the 40th ACL*, Philadelphia, USA, 2002, pp. 311–318.

[10] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proc. of the HLT 2002*, San Diego, USA, 2002, pp. 257–258.

[11] A. Lavie and A. Agarwal, "METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proc. of ACL Workshop on SMT*, Prague, Czech Republic, 2007, pp. 228–231.

[12] J. P. Turian, L. Shen, and I. D. Melamed, "Evaluation of machine translation and its evaluation," in *Proc. of the MT Summmit IX*, New Orleans, USA, 2003, pp. 386–393.

[13] S. Niessen, F. J. Och, G. Leusch, and H. Ney, "An evaluation tool for machine translation: Fast evaluation for machine translation research," in *Proc. of the 2nd LREC*, Athens, Greece, 2000, pp. 39–45.

[14] F. J. Och, "Minimum error rate training in smt," in *Proc. of the 41st ACL*, Sapporo, Japan, 2003, pp. 160–167.

[15] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proc. of the AMTA*, Cambridge and USA, 2006, pp. 223–231.

[16] Y. Zhang, S. Vogel, and A. Waibel, "Interpreting Bleu/NIST Scores: How Much Improvement do We Need to Have a Better System?" in *Proc of the LREC*, 2004, pp. 2051–2054.

[17] E. Matusov and S. Köprü, "AppTek's APT Machine Translation System for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 29–36.

[18] H. Almaghout, J. Jiang, and A. Way, "The DCU Machine Translation Systems for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 37–44.

[19] F. Zamora-Martinez, M. J. Castro-Bleda, and H. Schwenk, "N-gram-based Machine Translation enhanced with Neural Networks for the French-English BTEC-IWSLT'10 task," in *Proc. of IWSLT*, Paris, France, 2010, pp. 45–52.

[20] A. Bisazza, I. Klasinas, M. Cettolo, and M. Federico, "FBK @ IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 53–58.

[21] J. Gosme, W. Mekki, F. Debili, Y. Lepage, and N. Lucas, "The GREYC/LLACAN Machine Translation Systems for the IWSLT 2010 Campaign," in *Proc. of IWSLT*, Paris, France, 2010, pp. 59–65.

[22] X. Duan, R. E. Banchs, J. Lang, D. Xiong, A. Aw, M. Zhang, and H. Li, "I^2R Machine Translation System for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 67–72.

[23] H. Xiong, J. Xie, H. Yu, K. Liu, W. Luo, H. Mi, Y. Liu, Y. Lü, and Q. Liu, "The ICT Statistical Machine Translation Systems for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 73–79.

[24] W. Ling, T. Luís, J. Graça, L. Coheur, and I. Trancoso, "The INESC-ID Machine Translation System for the IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 81–84.

[25] G. Gascó, V. Alabau, J. Andrés-Ferrer, J. González-Rubio, M.-A. Rocha, G. Sanchis-Trilles, F. Casacuberta, J. González, and J.-A. Sánchez, "ITI-UPV system description for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 85–92.

[26] J. Niehues, M. Mediani, T. Herrmann, M. Heck, C. Herff, and A. Waibel, "The KIT Translation system for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 93–98.

[27] L. Besacier, H. Afli, T. N. D. Do, H. Blanchon, and M. Potet, "LIG Statistical Machine Translation Systems for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 99–104.

[28] A. Allauzen, J. M. Crego, I. D. El-Kahlout, H.-S. Le, G. Wisniewski, and F. Yvon, "LIMSI @ IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 105–112.

[29] A. Rousseau, L. Barrault, P. Deléglise, and Y. Estève, "LIUM's Statistical Machine Translation System for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 113–117.

[30] I. T. Khemakhem, S. Jamoussi, and A. B. Hamadou, "The MIRACL Arabic-English Statistical Machine Translation System for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 119–125.

[31] W. Shen, T. Anderson, R. Slyh, and A. R. Aminzadeh, "The MIT/LL-AFRL IWSLT-2010 MT System," in *Proc. of IWSLT*, Paris, France, 2010, pp. 127–134.

[32] C.-H. Li, N. Duan, Y. Zhao, S. Liu, L. Cui, M.-Y. Hwang, A. Axelrod, J. Gao, Y. Zhang, and L. Deng, "The MSRA Machine Translation System for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 135–138.

[33] C.-L. Goh, T. Watanabe, M. Paul, A. Finch, and E. Sumita, "The NICT Translation System for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 139–146.

[34] K. Sudoh, K. Duh, and H. Tsukada, "NTT Statistical Machine Translation System for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 147–152.

[35] H. Na and J.-H. Lee, "The POSTECH's Statistical Machine Translation System for the IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 153–156.

[36] S. Yahyaei and C. Monz, "The QMUL System Description for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 157–162.

[37] S. Mansour, S. Peitz, D. Vilar, J. Wuebker, and H. Ney, "The RWTH Aachen Machine Translation system for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 163–168.

[38] K. Bar and N. Dershowitz, "Tel Aviv University's System Description for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 169–174.

[39] J. Murakami, T. Nishimura, and M. Tokuhisa, "Statistical Pattern-Based Machine Translation with Statistical French-English Machine Translation," in *Proc. of IWSLT*, Paris, France, 2010, pp. 175–182.

[40] C. Mermer, H. Kaya, and M. U. Doğan, "The TÜBİTAK-UEKAE Statistical Machine Translation System for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 183–188.

[41] C. Henríquez, M. R. Costa-jussá, V. Daudaravicius, R. E. Banchs, and J. Marino, "UPC-BMIC-VDU system description for the IWSLT 2010: testing several collocation segmentations in a phrase-based SMT system," in *Proc. of IWSLT*, Paris, France, 2010, pp. 189–195.

[42] M. Khalilov and K. Sima'an, "The ILLC-UvA SMT System for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 197–203.

[43] S. Martzoukos and C. Monz, "The UvA System Description for IWSLT 2010," in *Proc. of IWSLT*, Paris, France, 2010, pp. 205–208.

Appendix A. MT System Overview

Research Group	MT System Description	Type	System	Tasks
Apptek, Inc. (Turkey)	AppTek's APT Machine Translation System for IWSLT 2010 [17]	PBSMT	apptek	BT_{AE},BT_{TE}
Carnegie Mellon University, Qatar Campus (Qatar)	Morphology-to-Syntax Alignment for Factored Phrase-based SMT	PBSMT	cmu_qatar	BT_{TE}
Dublin City University, School of Computing (Ireland)	The DCU Machine Translation Systems for IWSLT 2010 [18]	Hybrid	dcu	DT_{CE},BT_{AE}
Universidad CEU-Cardenal Herrera & Politecnica de Valencia (Spain)	N-gram-based Machine Translation enhanced with Neural Networks for the French-English BTEC-IWSLT'10 task [19]	NBSMT	dsic-upv	BT_{FE}
Fondazione Bruno Kessler, Ricerca Scientifica e Tecnologica (Italy)	FBK @ IWSLT 2010 [20]	PBSMT	fbk	TT_{EF},BT_{AE}, BT_{TE}
University of Caen Basse-Normandie, GREYC (France)	The GREYC/LLACAN Machine Translation Systems for the IWSLT 2010 Campaign [21]	PBSMT	greyc	BT_{AE}
Institute for Infocomm Research (Singapore)	I^2R Machine Translation System for IWSLT 2010 [22]	Hybrid	i2r	DT_{CE}
Chinese Academy of Sciences, Institute of Computing Technology (China)	The ICT Statistical Machine Translation Systems for the IWSLT 2010 [23]	Hybrid	ict	DT_{CE}
Instituto de Engenharia de Sistemas e Computadores Investigacao e Desenvolvimento (Portugal)	The INESC-ID Machine Translation System for the IWSLT 2010 [24]	PBSMT	inesc-id	DT_{CE},BT_{FE}
Universidad Politécnica de Valencia, Instituto Universitario Mixto de Tecnología Informática (Spain)	ITI-UPV system description for IWSLT 2010 [25]	Hybrid	iti-upv	TT_{EF},DT_{CE}
Karlsruhe Institute of Technology, interACT (Germany)	The KIT Translation system for IWSLT 2010 [26]	PBSMT	kit	TT_{EF},BT_{FE}
University J. Fourier, GETALP, LIG (France)	LIG Statistical Machine Translation Systems for IWSLT 2010 [27]	Hybrid	lig	TT_{EF},BT_{AE}
LIMSI-CNR (France)	LIMSI @ IWSLT 2010 [28]	PBSMT	limsi	TT_{EF},BT_{TE}
University of Le Mans, LIUM (France)	LIUM's Statistical Machine Translation System for IWSLT 2010 [29]	PBSMT	lium	TT_{EF}
MIRACL Laboratory (Tunesia)	The MIRACL Arabic-English Statistical Machine Translation System for IWSLT 2010 [30]	PBSMT	miracl	BT_{AE}
MIT Lincoln Laboratory (USA)	The MIT/LL-AFRL IWSLT-2010 MT System [31]	PBSMT	mit	TT_{EF},BT_{AE}, BT_{FE},BT_{TE}
Microsoft Research Asia, Natural Language Computing (China)	The MSRA Machine Translation System for IWSLT 2010 [32]	Hybrid	msra	DT_{CE}
National Institute of Information and Communications Technology (Japan)	The NICT Translation System for IWSLT 2010 [33]	Hybrid	nict	DT_{CE},BT_{FE}
NTT Comm. Science Labs (Japan)	NTT Statistical Machine Translation System for IWSLT 2010 [34]	PBSMT	ntt	TT_{EF}
Pohang University of Science and Technology (Korea)	The POSTECH's Statistical Machine Translation System for the IWSLT 2010 [35]	PBSMT	postech	DT_{CE}

EBMT : Example-based MT **NBSMT** : Ngram-based SMT **Hybrid** : Hybrid MT
PBSMT : Phrase-based SMT **HPSMT** : Hierachical Phrase-based SMT

: MT system description paper is not included in the proceedings.

Research Group	MT System Description	Type	System	Tasks
Queen Mary, University of London (United Kingdom)	The QMUL System Description for IWSLT 2010 [36]	PBSMT	qmul	BT_{AE},BT_{FE}, BT_{TE}
Rheinisch Westfälische Technische Hochschule (Germany)	The RWTH Aachen Machine Translation system for IWSLT 2010 [37]	Hybrid	rwth	BT_{AE}
Tel Aviv University (Israel)	Tel Aviv University's System Description for IWSLT 2010 [38]	EBMT	tau	BT_{AE}
Tottori University (Japan)	Statistical Pattern-Based Machine Translation with Statistical French-English Machine Translation [39]	Hybrid	tottori	BT_{FE}
TÜBİTAK-UEKAE (Turkey)	The TÜBİTAK-UEKAE Statistical Machine Translation System for IWSLT 2010 [40]	PBSMT	tubitak	TT_{EF},DT_{CE}, BT_{AE},BT_{TE}
Universitat Politécnica de Catalunya (Spain)	UPC-BMIC-VDU system description for the IWSLT 2010: testing several collocation segmentations in a phrase-based SMT system [41]	PBSMT	upc	BT_{FE}
University Amsterdam, Institute for Logic Language and Computation (Netherlands)	The ILLC-UvA SMT System for IWSLT 2010 [42]	PBSMT	uva-illc	DT_{CE}
University Amsterdam, Intelligence Systems Lab (Netherlands)	The UvA System Description for IWSLT 2010 [43]	PBSMT	uva-isca	DT_{CE},BT_{AE}, BT_{FE},BT_{TE}

EBMT : Example-based MT **NBSMT** : Ngram-based SMT **Hybrid** : Hybrid MT
PBSMT : Phrase-based SMT **HPSMT** : Hierachical Phrase-based SMT

: MT system description paper is not included in the proceedings.

Appendix B. Human Assessment

B.1. Fluency / Adequacy / Dialog

(best = 5.0, …, worst = 1.0)

· Only the top-ranked (*NormRank*) primary run submissions (cf. Appendix B.2.) were evaluated.
· *Fluency* indicates how the evaluation segment sounds to a native speaker of the target language.
· *Adequacy* indicates how much of the information from the reference translation was expressed in the MT output.
· *Dialog* is an adequacy assessment taking into account the context of the given dialog.

(testset_IWSLT10)

DIALOG	MT	Fluency	Adequacy	Dialog
DT_{CE}	ict.ASR	**2.41**	**2.42**	**2.72**
	online	1.75	1.84	2.07
	ict.CRR	**2.94**	**2.93**	**3.31**
	online	1.94	2.05	2.35
DT_{EC}	ict.ASR	**2.86**	**2.83**	**3.11**
	online	2.19	2.34	2.59
	ict.CRR	**3.61**	**3.74**	**3.93**
	online	2.41	2.62	2.88

BTEC	MT	Fluency	Adequacy
BT_{AE}	apptek	**3.43**	3.48
	online	3.28	**3.56**
BT_{FE}	dsic-upv	3.91	4.05
	online	**4.02**	**4.30**
BT_{TE}	tubitak	3.50	3.74
	online	**3.69**	**3.99**

B.2. Ranking

(**Ranking**: best = 1.0, …, worst = 0.0) (**NormRank**: best = 5.0, …, worst = 1.0)

· The *Ranking* scores are the average numbers of times that a system was judged better than any other system.
· The *NormRank* scores are normalized ranks on a per-judge basis using the method of [6].

DIALOG

DT$_{CE}$ (ASR)

MT	Ranking
ict	**0.5928**
nict	0.5197
i2r	0.4524
online	0.4442
msra	0.4392
iti-upv	0.3966
inesc-id	0.3850
uva-illc	0.3788
postech	0.3558
dcu	0.3439
tubitak	0.3420
uva-isca	*0.0736*

MT	NormRank
ict	**3.52**
nict	3.35
i2r	3.17
msra	3.12
inesc-id	3.05
online	3.05
iti-upv	3.00
uva-illc	2.96
tubitak	2.93
postech	2.93
dcu	2.83
uva-isca	*2.10*

DT$_{EC}$ (ASR)

MT	Ranking
ict	**0.5875**
i2r	0.5347
nict	0.5316
msra	0.4929
iti-upv	0.4730
postech	0.4670
inesc-id	0.4670
online	0.4467
tubitak	0.4296
dcu	0.3145
uva-illc	0.2819
uva-isca	*0.0307*

MT	NormRank
ict	**3.56**
nict	3.44
i2r	3.44
msra	3.24
iti-upv	3.22
postech	3.19
inesc-id	3.14
tubitak	3.10
online	3.02
dcu	2.63
uva-illc	2.52
uva-isca	*1.51*

DT$_{CE}$ (CRR)

MT	Ranking
ict	**0.7212**
nict	0.5720
i2r	0.5147
msra	0.5145
online	0.4746
dcu	0.4011
inesc-id	0.3911
iti-upv	0.3769
postech	0.3284
tubitak	0.3245
uva-illc	0.2483
uva-isca	*0.0766*

MT	NormRank
ict	**3.84**
nict	3.43
i2r	3.29
msra	3.26
online	3.10
inesc-id	3.00
dcu	2.91
iti-upv	2.89
tubitak	2.82
postech	2.80
uva-illc	2.62
uva-isca	*2.02*

DT$_{EC}$ (CRR)

MT	Ranking
ict	**0.7607**
i2r	0.5614
nict	0.5233
postech	0.4980
msra	0.4867
tubitak	0.4776
online	0.4577
inesc-id	0.4308
iti-upv	0.4086
dcu	0.3688
uva-illc	0.2986
uva-isca	*0.0311*

MT	NormRank
ict	**4.07**
i2r	3.50
nict	3.38
postech	3.24
tubitak	3.21
msra	3.19
inesc-id	2.99
online	2.95
iti-upv	2.90
dcu	2.75
uva-illc	2.48
uva-isca	*1.38*

BTEC

BT$_{AE}$

MT	Ranking
online	**0.4863**
apptek	0.4485
mit	0.4396
rwth	0.4020
qmul	0.3991
dcu	0.3889
fbk	0.3438
lig	0.3300
miracl	0.2967
uva-isca	0.2588
tau	0.2535
greyc	0.2529
tubitak	*0.2249*

MT	NormRank
apptek	**3.34**
mit	**3.34**
online	3.30
rwth	3.23
dcu	3.23
qmul	3.22
fbk	3.00
lig	2.91
miracl	2.87
uva-isca	2.78
tubitak	2.63
greyc	2.58
tau	*2.57*

BT$_{FE}$

MT	Ranking
online	**0.4114**
tottori	0.3482
kit	0.3256
dsic-upv	0.3248
mit	0.3135
inesc-id	0.3069
upc	0.3057
nict	0.3046
qmul	0.2794
uva-isca	*0.1437*

MT	NormRank
online	**3.24**
dsic-upv	3.13
kit	3.13
tottori	3.11
mit	3.09
inesc-id	3.08
upc	3.08
nict	3.03
qmul	2.94
uva-isca	*2.19*

BT$_{TE}$

MT	Ranking
online	**0.4437**
tubitak	0.3378
mit	0.3160
fbk	0.3137
apptek	0.3118
limsi	0.2923
qmul	0.2724
cmu_qatar	0.2697
uva-isca	*0.2432*

MT	NormRank
online	**3.22**
tubitak	3.13
mit	3.05
fbk	3.04
apptek	3.01
limsi	2.89
qmul	2.87
cmu_qatar	2.79
uva-isca	*2.72*

B.3 Difference To System With Best Ranking Score

(best = 0.0, …, worst = 1.0)

· The *BestRankDiff* scores are the ratio of translations that the system with the highest *Ranking* score (MT^{top}) gains to the respective system, i.e. $BestRankDiff = \frac{|translations\ ranked\ worse\ than\ MT^{top}| - |translations\ ranked\ better\ than\ MT^{top}|}{number\ of\ translations\ ranked\ together}$.

· The systems are ordered according to the *BestRankDiff* ratios.

DIALOG

DT$_{CE}$ (ASR)

ict	BestRankDiff	Better	Same	Worse
i2r	**0.1757**	0.2969	0.2303	0.4727
msra	0.1788	0.3196	0.1818	0.4985
nict	0.1953	0.3007	0.2031	0.4960
postech	0.3278	0.2450	0.1821	0.5728
online	0.3333	0.2371	0.1924	0.5704
inesc-id	0.3712	0.2006	0.2274	0.5719
tubitak	0.3880	0.2276	0.1567	0.6156
uva-illc	0.4169	0.2149	0.1530	0.6319
dcu	0.4308	0.1897	0.1897	0.6205
iti-upv	0.4361	0.2021	0.1595	0.6382
uva-isca	*0.7927*	0.0493	0.1085	0.8421

DT$_{EC}$ (ASR)

ict	BestRankDiff	Better	Same	Worse
nict	**0.0574**	0.3869	0.1685	0.4444
iti-upv	0.1363	0.3371	0.1893	0.4734
msra	0.1558	0.3290	0.1861	0.4848
tubitak	0.1966	0.3138	0.1757	0.5104
i2r	0.2034	0.3290	0.1385	0.5324
inesc-id	0.2321	0.3080	0.1517	0.5401
postech	0.2672	0.3017	0.1293	0.5689
online	0.3378	0.2702	0.1216	0.6081
dcu	0.3909	0.2510	0.1069	0.6419
uva-illc	0.6460	0.1150	0.1238	0.7610
uva-isca	*0.9033*	0.0210	0.0546	0.9243

DT$_{CE}$ (CRR)

ict	BestRankDiff	Better	Same	Worse
msra	**0.3464**	0.2105	0.2324	0.5570
nict	0.3843	0.2313	0.1529	0.6156
i2r	0.4334	0.1931	0.1802	0.6266
inesc-id	0.4771	0.2033	0.1161	0.6804
online	0.5296	0.1857	0.0988	0.7154
iti-upv	0.5530	0.1704	0.1060	0.7234
tubitak	0.5627	0.1578	0.1214	0.7206
dcu	0.6228	0.1315	0.1140	0.7543
postech	0.6837	0.1209	0.0744	0.8046
uva-illc	0.7456	0.1052	0.0438	0.8508
uva-isca	*0.8915*	0.0283	0.0518	0.9198

DT$_{EC}$ (CRR)

ict	BestRankDiff	Better	Same	Worse
i2r	**0.3733**	0.2360	0.1545	0.6094
nict	0.4553	0.2008	0.1428	0.6562
msra	0.5144	0.1893	0.1069	0.7037
postech	0.5590	0.1590	0.1227	0.7181
tubitak	0.5603	0.1767	0.0862	0.7370
online	0.5840	0.1681	0.0796	0.7522
iti-upv	0.6290	0.1169	0.1370	0.7459
inesc-id	0.6331	0.1310	0.1048	0.7641
dcu	0.7167	0.1115	0.0600	0.8283
uva-illc	0.8171	0.0622	0.0583	0.8793
uva-isca	*0.9723*	0.0000	0.0276	0.9723

BTEC

BT$_{AE}$

online	BestRankDiff	Better	Same	Worse
apptek	**-0.0659**	0.4258	0.2142	0.3598
mit	-0.0231	0.4035	0.2159	0.3804
qmul	0.0169	0.3785	0.2259	0.3954
dcu	0.0716	0.3253	0.2776	0.3970
rwth	0.0831	0.3490	0.2188	0.4321
miracl	0.1531	0.3213	0.2042	0.4744
lig	0.2159	0.3017	0.1804	0.5177
fbk	0.2682	0.2760	0.1796	0.5442
uva-isca	0.2988	0.2486	0.2039	0.5474
greyc	0.3577	0.2323	0.1775	0.5900
tubitak	0.3668	0.2100	0.2130	0.5769
tau	*0.4330*	0.1784	0.2099	0.6115

BT$_{FE}$

online	BestRankDiff	Better	Same	Worse
dsic-upv	**0.0082**	0.3476	0.2965	0.3558
mit	0.0246	0.3550	0.2653	0.3796
tottori	0.0270	0.3260	0.3209	0.3530
upc	0.0539	0.3252	0.2956	0.3791
kit	0.0630	0.3257	0.2854	0.3887
nict	0.0658	0.3309	0.2722	0.3967
qmul	0.0782	0.3435	0.2346	0.4217
inesc-id	0.0842	0.3157	0.2842	0.4000
uva-isca	*0.4595*	0.1742	0.1919	0.6338

BT$_{TE}$

online	BestRankDiff	Better	Same	Worse
tubitak	**0.0086**	0.3683	0.2547	0.3769
mit	0.0424	0.3663	0.2247	0.4088
fbk	0.0730	0.3483	0.2303	0.4213
limsi	0.1071	0.3464	0.2000	0.4535
apptek	0.1243	0.3204	0.2346	0.4448
cmu_qatar	0.1853	0.3239	0.1666	0.5093
qmul	0.1858	0.3097	0.1946	0.4955
uva-isca	*0.2890*	0.2630	0.1840	0.5520

Appendix C. Automatic Evaluation

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

C.1. Significance Test

· Only the subset of sentence IDs used for the human assessments were used for calculating the automatic scores of each MT output
· The mean score and the 95% confidence intervals were calculated for each MT output according to the *bootStrap* method [16].
· *Z-avg* is the average system score of *z*-transformed automatic evaluation metric scores achieving the highest rank correlation towards *Ranking*.
· MT systems are ordered according to the *z-avg* and the best (worst) score of each metric is marked with **boldface** (*italics*).
· Omitted lines between scores indicate non-significant differences in performance between the MT engines according to the *bootStrap* method [16].
· Besides the NIST metrics, all automatic evaluation metric scores are given as percent figure (%).

C.1.1 testset_IWSLT10

DIALOG Chinese-English (DT_{CE})

"case+punc" evaluation								**ASR**	*"no_case+no_punc" evaluation*							
BLEU	METEOR	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	METEOR	WER	PER	TER	GTM	NIST
18.75	**45.66**	72.53	59.61	63.44	54.27	4.392	1.377	ict	1.558	**19.86**	**43.79**	74.13	**59.50**	66.65	**54.56**	4.691
17.98	44.31	72.54	**59.59**	65.29	**55.44**	**4.618**	1.328	nict	1.245	17.99	41.52	75.63	60.63	69.17	53.95	**4.693**
18.23	43.04	72.59	60.60	65.13	54.30	4.270	1.185	i2r	1.107	18.78	40.39	75.44	62.02	68.97	52.23	4.362
19.04	42.20	**69.48**	62.00	**61.83**	53.10	3.002	1.087	msra	0.963	18.52	39.58	**71.89**	63.41	**64.97**	51.61	2.794
15.80	42.05	73.04	61.18	65.52	54.39	4.106	1.004	iti-upv	0.830	15.47	39.01	75.81	63.12	69.04	52.09	4.006
17.32	42.36	74.89	61.82	68.31	52.34	4.394	0.904	tubitak	0.642	16.53	39.11	78.44	64.14	72.80	49.50	4.402
15.62	40.76	73.97	63.13	66.30	50.56	3.897	0.721	inesc-id	0.600	15.50	38.04	76.92	64.73	69.87	48.78	3.953
15.83	38.80	74.85	64.08	67.72	49.64	3.683	0.515	postech	0.640	15.44	38.25	76.63	64.05	69.72	48.82	3.965
09.87	36.29	76.23	65.88	67.89	47.22	2.883	-0.065	uva-illc	0.250	12.31	36.19	77.75	65.06	70.30	48.27	3.352
10.53	36.04	77.31	66.61	68.82	46.54	2.980	-0.132	dcu	-0.163	10.38	33.98	78.95	67.25	71.20	45.32	2.973
07.15	30.15	83.50	73.45	74.95	37.61	*2.369*	-1.321	online	-0.567	08.85	33.11	82.24	69.56	75.54	42.84	3.265
09.07	*27.52*	*84.75*	*74.64*	*78.64*	*37.47*	2.676	-1.461	uva-isca	-1.963	*07.75*	*23.71*	*88.11*	*78.51*	*82.62*	*32.40*	*2.545*

DIALOG Chinese-English (DT_{CE})

"case+punc" evaluation								**CRR**	*"no_case+no_punc" evaluation*							
BLEU	METEOR	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	METEOR	WER	PER	TER	GTM	NIST
20.14	47.43	69.89	**56.63**	62.61	**58.78**	**5.028**	1.550	nict	1.523	20.44	44.86	72.42	**57.38**	66.02	**57.79**	**5.133**
21.08	**48.09**	69.34	58.22	60.98	56.96	4.710	1.515	ict	1.630	**21.92**	**46.11**	71.13	58.12	63.99	56.88	4.924
21.15	44.80	**67.17**	59.47	**59.84**	55.94	3.548	1.250	msra	1.079	20.56	42.30	**69.92**	61.07	**63.07**	54.25	3.329
18.87	44.93	70.89	59.59	62.41	56.00	3.859	1.045	i2r	1.001	19.79	42.55	73.07	60.20	65.49	54.44	3.798
18.63	44.08	72.74	60.30	65.53	55.11	4.558	0.956	tubitak	0.701	17.81	41.04	76.25	62.12	69.55	52.54	4.552
16.79	42.74	71.74	61.29	63.71	52.89	3.848	0.732	inesc-id	0.649	17.46	40.28	74.48	62.51	66.85	51.53	3.900
15.84	41.55	71.48	61.06	63.43	54.47	3.472	0.658	iti-upv	0.447	15.53	38.90	74.08	62.76	66.84	52.45	3.255
16.45	40.08	73.04	62.39	65.98	51.63	3.636	0.441	postech	0.572	16.25	39.72	74.39	62.30	67.61	51.09	3.926
11.24	38.01	76.62	65.10	67.92	48.33	3.189	-0.190	dcu	-0.178	11.37	36.40	78.00	65.10	70.14	47.94	3.265
10.82	37.46	75.48	65.03	66.87	47.69	3.051	-0.191	uva-illc	0.094	13.23	37.34	77.18	63.93	69.30	49.25	3.573
08.26	32.64	81.81	71.54	73.28	40.08	*2.581*	-1.244	online	-0.487	10.15	35.77	80.44	67.19	73.47	45.59	3.563
10.41	*29.52*	*82.89*	*72.70*	*76.05*	*39.68*	2.786	-1.377	uva-isca	-1.89	*09.00*	*25.72*	*86.11*	*76.29*	*79.43*	*35.48*	*2.647*

DIALOG English-Chinese (DT_{EC})

"case+punc" evaluation								**ASR**	*"no_case+no_punc" evaluation*							
BLEU	F1	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	F1	WER	PER	TER	GTM	NIST
29.76	**56.48**	**66.17**	**49.27**	**57.74**	**65.69**	**6.044**	1.558	ict	1.592	**29.70**	**55.41**	**68.17**	**50.78**	**59.03**	**64.03**	**6.065**
27.49	55.10	66.53	51.76	58.49	64.49	5.379	1.299	msra	1.199	26.95	52.76	68.62	54.05	59.90	61.27	5.134
28.62	54.58	69.32	52.05	62.41	65.16	5.879	1.271	i2r	1.190	27.55	52.83	71.81	54.21	63.92	62.06	5.863
27.51	54.36	70.30	51.81	62.49	64.68	5.805	1.207	nict	1.162	26.97	52.56	72.50	53.56	64.08	62.12	5.825
25.67	53.05	71.05	53.16	64.05	64.58	5.565	1.038	iti-upv	1.016	25.01	51.57	72.72	54.66	64.61	61.57	5.584
23.90	51.71	73.53	55.11	66.60	62.52	5.358	0.781	tubitak	0.703	23.13	49.72	75.39	56.70	67.31	58.58	5.373
24.05	50.09	77.13	58.40	71.83	62.57	5.216	0.512	inesc-id	0.552	23.72	49.54	77.70	58.97	71.78	59.86	5.327
22.64	49.41	76.28	58.25	69.99	60.01	5.067	0.455	postech	0.703	23.59	49.32	75.01	57.21	67.10	58.37	5.362
16.25	41.58	83.97	62.64	75.83	56.30	4.635	-0.286	uva-illc	-0.454	16.21	37.90	86.46	64.64	77.97	54.39	4.648
16.32	43.87	85.14	64.14	79.64	55.76	4.449	-0.398	dcu	-0.386	16.61	42.92	85.35	64.50	79.02	52.76	4.588
15.89	43.40	82.69	64.60	78.74	51.97	4.556	-0.409	online	-0.213	16.42	44.56	81.78	63.84	77.65	53.86	4.676
08.11	*27.65*	*88.80*	*77.99*	*80.50*	*34.44*	*2.293*	-1.885	uva-isca	-1.920	*08.11*	*27.29*	*89.14*	*78.34*	*80.04*	*34.61*	*2.313*

DIALOG English-Chinese (DT_{EC})

"case+punc" evaluation								**CRR**	*"no_case+no_punc" evaluation*							
BLEU	F1	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	F1	WER	PER	TER	GTM	NIST
36.78	**63.19**	**57.48**	**42.63**	**51.16**	**72.50**	**7.009**	1.784	ict	1.850	**36.79**	**61.52**	**59.51**	**44.25**	**52.65**	**70.50**	**7.040**
32.08	59.05	64.20	46.93	56.89	69.05	6.431	1.212	i2r	1.153	31.29	56.75	67.01	49.35	59.11	66.38	6.400
30.64	58.77	64.84	46.53	56.61	68.81	6.307	1.162	nict	1.145	30.53	56.86	67.34	48.35	58.45	66.64	6.334
29.38	57.97	64.38	48.13	56.20	67.73	5.899	1.052	msra	1.052	29.29	56.10	66.11	49.91	57.17	65.48	5.833
28.08	56.21	66.83	49.77	59.66	67.38	5.992	0.869	iti-upv	0.824	27.45	54.45	69.15	51.72	61.27	64.95	5.968
27.16	56.16	67.45	49.46	60.09	66.41	5.927	0.815	tubitak	0.733	26.56	53.99	70.00	51.63	61.60	62.77	5.906
27.62	54.52	69.13	51.33	61.92	65.93	5.859	0.688	postech	0.700	26.92	53.28	70.59	52.44	62.09	63.23	5.909
27.10	54.87	71.69	51.97	65.64	66.50	5.840	0.564	inesc-id	0.629	27.18	53.87	72.59	52.87	65.74	64.51	5.980
19.05	49.10	78.03	57.89	71.57	60.17	5.043	-0.189	dcu	-0.294	19.02	47.17	80.13	59.58	73.32	57.37	5.109
17.98	45.60	78.53	60.49	73.54	54.33	5.034	-0.441	online	-0.144	19.70	48.20	77.19	58.78	71.93	57.46	5.260
18.09	41.78	80.90	60.34	72.29	58.60	4.976	-0.461	uva-illc	-0.630	18.14	38.11	83.60	62.33	74.41	56.62	5.002
08.48	*29.84*	*84.52*	*75.66*	*75.48*	*35.46*	*1.918*	-1.910	uva-isca	-1.875	*09.44*	*30.78*	*84.72*	*74.62*	*75.63*	*36.84*	*2.286*

BTEC Arabic-English (BT_{AE})

"case+punc" evaluation								**CRR**	*"no_case+no_punc" evaluation*							
BLEU	METEOR	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	METEOR	WER	PER	TER	GTM	NIST
46.73	**73.22**	**37.51**	**32.72**	**32.30**	72.97	**7.345**	1.669	mit	1.618	**43.95**	**69.34**	**42.89**	**36.78**	**36.42**	69.42	7.228
43.76	71.48	39.95	35.22	34.93	**73.29**	7.248	1.365	rwth	1.322	41.25	67.19	45.62	39.12	39.59	**70.14**	7.177
42.96	72.88	40.72	35.46	35.10	71.41	7.285	1.312	apptek	1.333	40.77	69.02	45.83	39.06	39.58	68.34	**7.365**
43.47	71.69	40.31	35.73	34.68	71.73	7.123	1.285	dcu	1.240	40.64	67.78	45.78	39.57	39.13	68.10	7.036
41.55	70.84	42.27	36.76	36.31	70.25	7.042	1.053	qmul	0.999	39.33	66.64	48.19	41.09	40.82	66.36	7.023
40.57	69.23	42.39	36.71	36.25	70.12	6.734	0.934	fbk	0.864	38.12	65.23	48.29	41.09	40.69	66.23	6.574
35.92	65.95	46.64	40.33	40.32	66.75	6.482	0.314	tubitak	0.198	33.65	61.18	53.04	45.64	45.44	62.25	6.443
35.15	66.13	47.61	41.45	41.05	68.65	6.522	0.294	lig	0.470	34.69	63.18	52.04	44.68	44.50	65.70	6.651
33.62	68.37	49.26	41.69	41.71	68.16	6.586	0.262	miracl	0.468	33.58	64.80	53.45	44.72	46.38	66.36	6.907
33.85	67.75	48.91	42.47	43.54	68.53	6.666	0.211	online	0.143	30.81	63.24	55.48	46.75	49.21	64.98	6.772
29.04	64.14	50.70	45.17	43.13	63.23	5.857	-0.365	uva-isca	-0.219	29.24	59.02	55.15	48.17	48.56	61.23	6.107
27.04	*56.88*	53.79	48.13	45.75	59.84	*4.602*	-1.131	greyc	-1.375	24.02	*51.12*	61.23	53.79	51.20	54.65	*4.105*
20.11	57.60	*58.97*	*52.59*	*50.18*	*56.46*	4.978	-1.631	tau	-1.488	*21.42*	52.21	*64.78*	*56.00*	*56.70*	*54.36*	5.310

BTEC French-English (BT_{FE})

"case+punc" evaluation								**CRR**	*"no_case+no_punc" evaluation*							
BLEU	METEOR	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	METEOR	WER	PER	TER	GTM	NIST
52.69	**78.43**	**32.36**	**27.90**	**27.72**	**77.58**	**8.148**	1.174	mit	1.232	**51.27**	**75.36**	**36.69**	**31.11**	**31.16**	**74.20**	**8.201**
50.56	77.55	32.97	29.09	27.83	75.59	7.777	0.856	dsic-upv	0.842	48.76	74.52	37.74	32.34	31.35	72.42	7.784
50.52	77.31	33.41	28.85	28.24	75.79	7.844	0.852	nict	0.833	48.84	74.00	38.21	32.21	31.73	72.45	7.871
50.46	77.56	33.69	29.26	28.39	75.13	7.814	0.803	upc	0.784	48.28	74.47	38.29	32.42	31.70	71.78	7.831
49.30	77.45	35.08	30.24	30.09	76.00	7.963	0.721	tottori	0.728	47.09	74.23	39.81	33.33	33.82	72.72	8.053
48.59	77.33	35.33	30.25	30.36	76.01	7.896	0.666	kit	0.647	46.12	74.37	40.51	33.89	34.07	72.96	7.967
49.38	77.01	34.97	29.75	29.53	75.34	7.756	0.660	inesc-id	0.694	47.75	74.02	39.56	32.47	33.16	72.13	7.845
50.46	76.19	35.08	29.85	29.28	74.95	7.657	0.608	qmul	0.512	48.46	72.70	39.83	33.47	32.71	71.41	7.574
46.93	76.01	37.13	32.45	32.14	75.16	7.777	0.332	online	0.314	44.11	72.89	42.13	35.56	35.93	72.19	7.861
30.13	*67.28*	*50.80*	*45.02*	*42.52*	*64.28*	*6.118*	-2.387	uva-isca	-2.302	*31.00*	*62.97*	*54.55*	*46.71*	*48.29*	*63.28*	*6.621*

BTEC Turkish-English (BT_{TE})

"case+punc" evaluation								**CRR**	*"no_case+no_punc" evaluation*							
BLEU	METEOR	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	METEOR	WER	PER	TER	GTM	NIST
50.47	**77.80**	37.31	29.51	29.83	**75.00**	**7.871**	1.350	tubitak	1.412	49.07	**74.23**	42.14	**32.43**	33.68	**72.09**	**7.930**
49.87	76.55	36.27	**29.48**	30.26	75.13	7.744	1.275	fbk	1.202	47.56	72.89	41.58	33.11	34.22	71.61	7.630
51.27	76.20	**35.22**	29.99	**29.04**	74.53	7.300	1.232	mit	1.284	**50.11**	72.58	**39.35**	32.58	**32.39**	71.87	7.076
48.17	75.03	39.24	31.79	31.02	72.28	7.155	0.717	apptek	0.823	47.69	71.31	44.29	34.23	34.90	70.03	7.179
44.21	74.76	40.05	33.44	33.34	72.50	7.348	0.486	online	0.445	41.37	71.02	45.75	36.58	37.77	69.75	7.344
47.59	72.80	39.86	33.70	32.65	71.39	6.996	0.389	qmul	0.324	45.73	68.81	44.90	37.13	36.59	68.28	6.770
45.23	72.98	40.40	33.44	32.98	72.13	7.028	0.351	limsi	0.292	43.29	68.67	45.96	36.47	37.02	69.08	6.870
43.89	71.70	43.30	35.69	35.87	70.28	7.107	-0.072	cmu_qatar	-0.223	41.42	67.19	49.58	40.03	40.45	66.77	7.056
32.70	*66.01*	*50.86*	*43.86*	*40.03*	*63.09*	*5.746*	-1.872	uva-isca	-1.702	*33.36*	*60.35*	*55.71*	*45.49*	*45.26*	*62.34*	*5.926*

C.1.2 testset_IWSLT09

DIALOG Chinese-English (DT_{CE})

"case+punc" evaluation								**ASR**	*"no_case+no_punc" evaluation*							
BLEU	METEOR	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	METEOR	WER	PER	TER	GTM	NIST
34.06	61.01	**54.32**	45.37	**47.85**	67.52	5.841	1.603	msra	1.704	**34.06**	59.21	**55.28**	45.39	**50.33**	67.30	6.077
33.22	**63.08**	58.01	**44.70**	50.64	**67.95**	**6.188**	1.565	ict	1.673	33.36	**61.44**	59.83	**44.57**	53.97	**68.01**	**6.527**
31.06	60.90	59.21	46.77	52.53	66.89	5.809	1.234	i2r	1.295	32.58	58.59	60.92	47.16	56.09	65.92	6.204
27.35	57.17	60.24	48.08	54.52	65.55	5.428	0.846	iti-upv	0.664	26.16	54.27	62.68	49.68	58.01	64.22	5.570
28.60	58.05	60.62	49.16	54.49	62.91	5.540	0.816	inesc-id	0.734	28.73	55.69	62.95	50.11	58.23	61.91	5.805
28.16	57.97	62.99	49.83	57.92	63.87	5.442	0.648	tubitak	0.376	26.36	55.03	66.27	51.51	62.11	61.75	5.571
26.88	55.19	60.76	48.97	53.29	63.23	4.863	0.616	nict	0.410	26.99	52.56	63.07	50.80	56.59	61.02	4.921
28.29	56.01	63.80	51.20	57.97	61.87	5.540	0.517	postech	0.418	27.19	54.84	66.44	51.93	61.58	61.88	5.696
25.39	54.47	62.31	50.94	56.17	60.30	5.016	0.362	dcu	0.359	25.10	52.83	63.82	50.78	58.88	60.57	5.299
17.29	50.32	65.77	53.60	58.82	58.39	4.568	-0.279	uva-illc	0.253	23.11	52.80	65.61	51.12	59.95	61.02	5.386
19.70	45.91	71.64	59.97	*67.93*	53.27	4.072	-1.082	uva-isca	-1.693	17.99	*42.16*	*75.59*	*63.00*	*72.90*	*49.43*	*4.013*
13.54	*41.78*	*74.58*	*63.37*	67.91	*47.86*	*3.638*	-1.703	online	-1.050	*16.68*	46.52	72.35	58.39	67.58	52.62	4.608

DIALOG Chinese-English (DT_{CE})

"case+punc" evaluation								**CRR**	*"no_case+no_punc" evaluation*							
BLEU	METEOR	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	METEOR	WER	PER	TER	GTM	NIST
36.98	65.47	**51.10**	**41.21**	**44.77**	70.87	6.470	1.676	msra	1.672	**37.24**	64.05	**51.73**	41.01	**46.65**	71.00	6.746
34.97	**66.45**	52.21	41.23	45.88	**71.81**	**6.540**	1.639	ict	1.791	36.72	**66.02**	52.03	**39.45**	47.86	**71.48**	**7.072**
32.92	66.03	58.37	42.24	49.52	70.78	6.281	1.264	i2r	1.327	35.77	64.22	59.47	41.95	52.64	70.63	6.800
30.81	62.17	59.24	46.37	52.58	67.11	6.018	0.803	inesc-id	0.691	31.61	60.30	61.33	47.13	56.25	66.53	6.351
28.62	60.25	59.33	46.18	52.51	67.58	5.772	0.679	iti-upv	0.522	27.97	58.23	61.52	46.83	55.96	67.56	6.003
29.23	58.72	59.16	46.27	51.22	66.94	5.432	0.607	nict	0.447	29.88	56.75	61.06	47.09	54.23	65.86	5.597
29.87	61.96	62.15	47.49	56.60	67.43	5.781	0.559	tubitak	0.276	28.44	59.86	65.28	49.22	61.13	66.08	5.935
28.79	59.06	59.09	46.92	52.91	65.06	5.625	0.545	dcu	0.538	28.98	57.93	60.20	46.23	55.14	65.75	5.942
29.82	59.73	62.33	47.97	56.32	65.50	5.910	0.475	postech	0.376	29.29	59.27	64.41	48.59	59.70	66.27	6.070
19.52	53.95	65.64	51.04	58.42	61.67	5.194	-0.314	uva-illc	0.047	25.80	56.77	66.23	49.53	60.55	64.37	5.807
19.18	47.33	*72.28*	59.01	*68.02*	54.66	*4.180*	-1.388	uva-isca	-1.898	*17.59*	*44.44*	*76.44*	*62.07*	*73.83*	*52.84*	*4.106*
16.25	*47.01*	70.72	*59.37*	64.45	*52.72*	4.202	-1.401	online	-0.645	20.18	52.45	67.64	53.53	62.86	58.04	5.284

DIALOG English-Chinese (DT_{EC})

"case+punc" evaluation								**ASR**	*"no_case+no_punc" evaluation*							
BLEU	F1	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	F1	WER	PER	TER	GTM	NIST
37.92	**65.59**	54.24	**40.41**	**47.17**	71.89	6.287	1.490	ict	1.600	**38.51**	**64.38**	56.15	**41.70**	**48.50**	**70.96**	6.378
37.31	65.01	**53.33**	41.89	47.18	70.82	5.855	1.358	msra	1.381	37.96	63.50	**55.22**	43.62	48.63	69.06	5.770
36.32	63.56	57.27	41.92	50.73	70.65	6.352	1.247	i2r	1.186	36.02	61.53	59.91	44.20	53.10	68.39	6.361
36.78	63.93	59.68	41.98	53.05	**72.16**	**6.374**	1.210	nict	1.205	36.51	62.29	62.20	43.66	55.01	70.56	**6.412**
32.00	61.51	60.20	44.15	53.72	69.09	5.929	0.857	tubitak	0.762	31.80	59.66	62.49	46.42	55.44	66.17	5.921
34.22	61.12	62.70	44.91	56.36	70.43	6.108	0.842	inesc-id	0.827	34.20	59.64	64.80	46.61	57.82	68.51	6.206
32.85	60.23	60.19	45.68	54.22	68.82	5.912	0.800	iti-upv	0.730	32.64	58.35	62.32	47.86	55.74	66.79	5.922
32.07	59.86	62.57	45.84	55.82	67.36	5.833	0.656	postech	0.698	32.15	59.00	63.63	46.81	56.02	65.81	5.897
24.55	52.41	73.35	53.95	64.60	61.39	5.114	-0.423	uva-illc	-0.616	24.88	50.32	76.00	56.07	66.82	60.24	5.165
23.84	53.90	74.03	53.41	66.23	62.02	5.077	-0.433	dcu	-0.588	23.89	52.44	75.73	54.96	67.63	59.72	5.154
21.33	49.66	73.18	56.27	67.79	56.42	4.982	-0.775	online	-0.396	24.27	52.23	71.11	53.77	65.75	59.63	5.340
17.52	*42.54*	*76.16*	*63.88*	*69.30*	*48.79*	*3.235*	-1.686	uva-isca	-1.645	*18.80*	*43.63*	*76.09*	*62.73*	*68.79*	*50.20*	*3.544*

DIALOG English-Chinese (DT_{EC})

"case+punc" evaluation								**CRR**	*"no_case+no_punc" evaluation*							
BLEU	F1	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	F1	WER	PER	TER	GTM	NIST
49.26	**73.90**	**41.76**	**31.74**	**36.78**	**79.81**	**7.609**	1.870	ict	1.992	**49.55**	**72.55**	**43.82**	**33.29**	**38.48**	**78.72**	**7.696**
42.45	70.41	48.34	36.36	41.87	76.34	6.622	1.194	msra	1.238	43.02	68.98	49.77	37.89	43.09	75.07	6.610
41.33	69.50	49.86	35.50	43.60	75.99	6.999	1.163	i2r	1.096	41.00	67.44	52.91	37.74	46.00	74.15	6.995
42.63	69.30	52.99	36.01	45.87	77.64	7.139	1.133	nict	1.137	42.70	67.73	55.55	37.70	48.11	76.51	7.215
37.93	67.43	52.73	37.56	46.16	75.37	6.690	0.877	tubitak	0.776	37.72	65.59	55.45	39.83	48.10	72.69	6.668
40.17	67.09	54.90	38.14	48.00	76.06	6.885	0.872	inesc-id	0.901	40.22	65.99	56.56	39.45	49.11	74.59	7.019
36.15	65.42	55.16	40.41	48.32	73.68	6.410	0.604	iti-upv	0.569	36.59	64.09	57.19	42.07	49.66	72.28	6.458
37.19	64.78	56.20	40.65	49.72	72.70	6.495	0.558	postech	0.521	37.38	63.67	57.91	42.05	50.92	70.88	6.549
29.75	60.89	65.16	45.13	56.85	68.73	5.949	-0.157	dcu	-0.374	29.22	59.19	67.86	47.17	59.16	66.36	5.981
28.41	58.42	67.77	48.50	58.65	66.09	5.662	-0.456	uva-illc	-0.629	29.21	56.48	70.19	50.14	60.37	65.00	5.762
25.27	54.51	68.17	51.31	62.56	61.58	5.550	-0.816	online	-0.438	29.21	57.70	65.88	48.35	59.88	64.98	6.010
21.01	*47.18*	*70.74*	*59.41*	*63.70*	*52.95*	*3.474*	-1.700	uva-isca	-1.645	*22.82*	*48.77*	*70.43*	*57.50*	*63.20*	*54.92*	*4.017*

BTEC Arabic-English (BT_{AE})

"case+punc" evaluation								**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	METEOR	WER	PER	TER	GTM	NIST
58.60	**81.89**	**29.76**	**26.77**	**24.84**	**78.59**	**8.169**	1.844	mit	1.845	**56.58**	**78.89**	**34.87**	**30.60**	**28.44**	**76.17**	**8.258**
55.15	80.11	33.37	30.18	27.78	76.86	7.652	1.246	rwth	1.220	53.76	76.66	38.58	34.22	31.81	74.31	7.645
53.95	80.15	33.68	30.40	27.89	76.47	7.669	1.207	dcu	1.134	52.23	76.40	39.11	34.68	32.21	73.67	7.693
52.69	79.17	34.72	30.67	28.91	75.88	7.769	1.087	qmul	0.977	50.74	75.28	41.03	35.21	33.18	72.80	7.827
52.45	79.16	35.07	31.31	29.36	75.84	7.860	1.056	apptek	1.083	51.04	75.67	40.10	34.86	33.65	73.47	8.063
51.09	78.70	36.58	32.59	30.57	75.42	7.921	0.927	online	0.870	48.91	75.06	42.80	36.82	35.21	72.83	8.182
51.94	78.56	35.08	31.55	28.85	74.09	7.158	0.831	fbk	0.710	49.89	74.66	41.05	36.05	33.01	71.06	7.043
43.25	75.75	40.92	35.79	34.25	74.13	7.577	0.252	miracl	0.469	44.45	72.00	45.69	38.69	38.42	72.84	8.050
46.36	76.42	39.95	35.54	33.03	72.59	6.821	0.141	lig	0.441	46.41	74.17	43.38	37.61	35.57	71.33	7.049
44.48	73.70	38.95	34.76	32.54	70.09	6.845	0.056	tubitak	-0.126	42.83	68.64	45.56	40.18	37.42	66.40	6.923
39.26	73.00	42.87	38.76	34.45	67.65	6.549	-0.468	uva-isca	-0.335	40.78	68.15	47.43	41.83	39.65	66.36	6.995
39.17	71.67	43.78	40.15	36.92	67.97	5.606	-0.941	greyc	-1.303	35.47	66.51	51.05	46.16	42.51	63.84	*5.205*
26.72	*69.92*	*49.10*	*44.27*	*39.80*	*63.45*	*5.685*	-1.667	tau	-1.413	*29.18*	*65.18*	*53.91*	*46.77*	*45.25*	*62.97*	6.287

BTEC French-English (BT_{FE})

"case+punc" evaluation								**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	METEOR	WER	PER	TER	GTM	NIST
63.53	**82.95**	**25.90**	**23.54**	**21.91**	**82.09**	**9.003**	1.162	mit	1.190	61.82	**80.01**	**29.54**	**26.43**	**24.94**	**80.23**	**9.298**
63.52	81.95	26.69	24.19	22.16	80.40	8.710	0.921	dsic-upv	0.948	**62.23**	78.99	30.04	26.82	24.99	78.44	8.972
61.90	81.89	27.31	23.95	22.92	79.84	8.738	0.834	upc	0.775	59.48	79.10	31.66	27.47	25.97	77.64	9.013
61.41	81.39	27.73	24.70	23.08	80.28	8.741	0.781	nict	0.786	59.58	78.42	31.31	27.43	26.06	78.25	9.016
61.70	81.65	28.37	24.76	23.53	80.38	8.683	0.753	qmul	0.793	60.41	78.79	31.99	27.40	26.55	78.53	8.982
58.86	81.71	29.35	25.72	24.78	80.13	8.826	0.627	tottori	0.599	56.74	78.45	33.54	28.64	28.20	77.97	9.169
60.97	80.94	29.95	26.86	25.19	80.22	8.813	0.568	online	0.606	58.96	77.97	33.83	29.78	28.80	78.63	9.194
59.44	80.73	28.63	25.86	24.01	79.12	8.575	0.528	inesc-id	0.502	57.52	77.50	32.65	28.78	27.38	77.21	8.896
58.79	81.77	29.65	26.43	25.07	79.34	8.691	0.522	kit	0.440	56.12	78.71	34.46	29.92	28.67	77.31	9.011
38.51	*72.36*	*44.84*	*40.22*	*36.68*	*69.17*	*6.987*	-2.411	uva-isca	-2.354	*39.79*	*68.06*	*49.22*	*42.39*	*42.16*	*68.81*	*7.605*

BTEC Turkish-English (BT_{TE})

"case+punc" evaluation								**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST	**z-avg**		**z-avg**	BLEU	METEOR	WER	PER	TER	GTM	NIST
57.59	**81.63**	31.39	25.97	24.39	78.10	**8.342**	1.425	tubitak	1.524	55.74	**78.47**	36.09	**28.49**	27.98	**76.20**	**8.666**
57.66	80.19	31.61	**25.58**	25.01	**78.73**	8.218	1.352	fbk	1.374	55.51	77.12	36.67	28.93	28.51	76.16	8.379
60.20	79.44	**30.15**	25.92	**23.44**	77.86	7.419	1.279	mit	1.312	**59.16**	76.06	**34.58**	29.26	**26.58**	75.80	7.326
52.88	76.91	35.58	29.45	27.75	74.76	7.401	0.519	limsi	0.445	50.64	72.96	41.18	33.05	31.88	72.08	7.422
52.60	77.91	36.65	29.45	27.88	74.72	7.408	0.519	apptek	0.544	50.82	74.19	42.94	32.67	31.99	73.59	7.596
53.51	75.50	36.21	30.95	29.07	73.91	7.190	0.301	qmul	0.167	50.97	71.43	42.18	35.96	32.82	71.09	7.131
49.45	77.09	38.27	31.60	30.25	74.59	7.580	0.259	online	0.232	46.75	73.18	44.59	34.86	34.76	72.89	7.788
48.98	75.75	39.45	32.32	31.44	74.02	7.544	0.076	cmu_qatar	-0.062	46.38	71.14	45.92	37.04	36.24	71.33	7.710
36.01	*68.23*	*48.94*	*42.24*	*35.74*	*63.93*	*5.670*	-1.872	uva-isca	-1.678	*38.00*	*62.41*	*53.40*	*43.54*	*41.38*	*64.03*	*6.048*

C.2. Full Testset

· All sentence IDs of the IWSLT 2010 testset were used for calculating the automatic scores of each MT output.
· The order of the systems is the same as for the results reported in Appendix C.1.
· The results are obtained using the online evaluation tool: https://mastarpj.nict.go.jp/EVAL/IWSLT10/automatic/testset_IWSLT10

C.2.1 testset_IWSLT10

DIALOG Chinese-English (DT_{CE})

"case+punc" evaluation							**ASR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
21.40	**47.91**	69.66	57.88	60.70	55.63	4.681	ict	**22.20**	**45.49**	71.83	**58.37**	64.33	**55.35**	4.979
20.99	47.11	69.35	**57.44**	62.26	**57.18**	**4.968**	nict	20.73	43.81	72.97	59.04	66.40	55.12	**5.042**
20.77	45.90	69.55	58.43	61.89	56.04	4.604	i2r	21.36	42.73	72.77	60.25	65.97	53.68	4.721
22.68	45.54	**66.27**	59.34	**58.79**	55.33	3.562	msra	21.85	42.40	**69.07**	61.21	**62.34**	53.41	3.331
18.53	44.73	69.81	58.96	62.44	56.13	4.416	iti-upv	18.13	41.40	72.96	61.21	66.11	53.44	4.299
19.69	45.04	71.67	59.65	64.92	54.03	4.695	tubitak	18.43	41.28	75.75	62.48	69.72	50.74	4.698
18.10	43.69	71.02	61.02	63.08	52.20	4.152	inesc-id	18.03	40.47	74.25	62.91	67.03	50.24	4.231
18.41	41.46	72.13	62.11	65.01	51.26	3.970	postech	17.70	40.57	74.21	62.44	67.17	50.12	4.263
11.50	38.50	73.67	64.11	64.99	48.47	3.022	uva-illc	14.31	38.03	75.72	63.81	67.70	49.11	3.522
12.79	38.92	74.40	64.40	65.66	48.33	3.243	dcu	12.53	36.56	76.40	65.27	68.40	47.00	3.245
07.61	30.52	82.44	*73.36*	73.41	37.58	*2.297*	online	09.50	33.80	80.76	69.05	73.67	43.08	3.331
10.89	*29.85*	*82.55*	73.13	*76.09*	*39.07*	2.888	uva-isca	*09.34*	*25.63*	*86.31*	*77.29*	*80.49*	*33.82*	*2.770*

DIALOG Chinese-English (DT_{CE})

"case+punc" evaluation							**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
23.32	50.23	66.62	**54.48**	59.43	**60.38**	**5.394**	nict	23.34	47.23	69.57	**55.72**	63.04	**58.89**	**5.508**
23.47	**50.65**	66.35	56.21	57.93	58.55	5.037	ict	**24.58**	**48.42**	68.34	56.34	61.18	58.14	5.294
24.45	47.96	**64.16**	56.97	**56.81**	57.98	4.062	msra	23.54	44.93	**67.11**	58.89	**60.35**	55.98	3.823
22.07	48.00	67.53	57.19	59.10	57.78	4.272	i2r	22.80	45.22	70.04	58.21	62.44	55.94	4.220
21.05	46.77	69.48	58.06	62.13	56.75	4.861	tubitak	19.80	43.30	73.38	60.25	66.40	53.79	4.850
19.59	45.90	68.49	58.91	60.23	54.72	4.158	inesc-id	20.23	42.98	71.41	60.35	63.72	53.21	4.235
18.97	44.54	68.26	58.62	60.43	56.51	3.862	iti-upv	18.50	41.54	71.14	60.63	63.82	54.05	3.589
19.18	42.85	70.18	60.27	63.11	53.31	3.948	postech	18.65	42.15	71.86	60.56	64.91	52.44	4.243
13.58	40.90	73.61	62.88	64.67	50.13	3.459	dcu	13.58	38.97	75.31	63.12	67.20	49.57	3.540
12.56	39.70	72.76	63.02	63.85	49.19	3.190	uva-illc	15.23	39.24	74.85	62.46	66.45	50.24	3.731
08.71	33.13	*80.72*	*71.31*	71.66	*40.12*	*2.519*	online	10.84	36.63	78.89	66.52	71.50	45.88	3.643
12.28	*31.95*	80.47	71.06	*73.34*	41.31	3.007	uva-isca	*10.74*	*27.80*	*84.09*	*74.83*	*77.21*	*37.00*	*2.893*

DIALOG English-Chinese (DT_{EC})

"case+punc" evaluation							**ASR**	*"no_case+no_punc" evaluation*						
BLEU	F1	WER	PER	TER	GTM	NIST		BLEU	F1	WER	PER	TER	GTM	NIST
31.14	**57.67**	**64.40**	**48.13**	**56.11**	**66.40**	**6.296**	ict	**30.80**	**56.38**	**66.69**	**49.88**	**57.63**	**64.56**	**6.324**
28.73	56.25	64.97	50.77	56.99	65.07	5.626	msra	27.99	53.59	67.41	53.33	58.70	61.71	5.394
29.90	55.55	67.77	51.13	61.05	65.78	6.106	i2r	28.60	53.54	70.60	53.55	62.89	62.55	6.113
28.74	55.43	68.63	50.76	60.97	65.35	6.023	nict	28.04	53.49	71.12	52.74	62.80	62.63	6.062
26.80	53.74	69.73	52.45	62.80	65.03	5.762	iti-upv	25.97	52.06	71.68	54.18	63.58	61.89	5.801
24.98	52.61	72.12	54.33	65.20	63.18	5.536	tubitak	24.12	50.36	74.19	56.08	66.10	59.04	5.579
25.16	51.11	75.48	57.45	70.15	63.14	5.407	inesc-id	24.69	49.73	76.48	58.35	70.50	60.28	5.527
23.80	50.41	74.60	57.20	68.24	60.60	5.259	postech	24.58	50.20	73.62	56.35	65.61	58.75	5.572
17.56	42.64	82.15	61.39	73.79	57.11	4.827	uva-illc	17.27	38.90	84.99	63.66	76.21	54.98	4.844
17.32	44.86	83.70	63.37	78.06	56.33	4.597	dcu	17.39	43.60	84.28	64.00	77.81	53.12	4.747
16.23	43.72	82.09	64.46	77.96	51.96	4.679	online	16.85	44.87	81.00	63.60	76.80	53.98	4.812
08.50	*27.96*	*88.40*	*77.81*	*80.18*	*34.59*	*2.350*	uva-isca	*08.50*	*27.54*	*88.83*	*78.27*	*79.78*	*34.83*	*2.393*

DIALOG English-Chinese (DT_{EC})

"case+punc" evaluation							**CRR**	*"no_case+no_punc" evaluation*						
BLEU	F1	WER	PER	TER	GTM	NIST		BLEU	F1	WER	PER	TER	GTM	NIST
37.92	**64.16**	**56.01**	**41.72**	**49.76**	**72.95**	**7.272**	ict	**37.67**	**62.28**	**58.36**	**43.59**	**51.50**	**70.81**	**7.318**
33.03	59.89	62.87	46.19	55.72	69.40	6.649	i2r	32.14	57.36	65.98	48.88	58.21	66.60	6.646
31.72	59.69	63.30	45.68	55.30	69.31	6.529	nict	31.41	57.62	66.14	47.74	57.40	66.96	6.571
30.57	58.99	63.09	47.32	54.98	68.21	6.149	msra	30.24	56.76	65.14	49.43	56.23	65.79	6.099
29.15	57.12	65.42	48.92	58.31	67.86	6.206	iti-upv	28.38	55.16	68.07	51.15	60.17	65.25	6.198
28.13	57.02	66.12	48.74	58.86	66.93	6.116	tubitak	27.45	54.58	68.90	51.12	60.52	63.12	6.128
28.73	55.55	67.51	50.31	60.30	66.43	6.063	postech	27.82	54.13	69.29	51.68	60.74	63.54	6.125
28.17	55.96	70.03	51.07	64.00	66.97	6.042	inesc-id	28.13	54.50	71.36	52.28	64.46	64.82	6.199
20.13	50.16	76.38	56.93	69.86	60.74	5.217	dcu	19.81	47.93	78.89	58.97	71.98	57.75	5.292
18.26	45.91	78.01	60.41	72.90	54.37	5.171	online	20.11	48.48	76.53	58.61	71.23	57.52	5.422
19.33	41.93	79.37	59.55	70.81	59.13	5.147	uva-illc	19.13	37.80	82.48	61.86	73.27	56.92	5.179
09.02	*30.49*	*83.85*	*75.18*	*74.85*	*35.93*	*1.981*	uva-isca	*10.12*	*31.42*	*84.03*	*74.17*	*75.00*	*37.34*	*2.396*

BTEC Arabic-English (BT_{AE})

"case+punc" evaluation							**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
49.26	**75.25**	**35.43**	**30.97**	**30.28**	74.08	**7.944**	mit	**46.12**	**71.29**	**40.76**	**34.98**	**34.35**	70.50	7.893
46.60	73.69	37.71	33.31	32.70	**74.29**	7.842	rwth	43.72	69.41	43.35	37.22	37.27	**71.12**	7.845
45.70	74.95	38.38	33.55	32.86	72.60	7.878	apptek	42.79	71.06	43.50	37.23	37.24	69.36	**8.014**
46.15	73.82	38.01	33.72	32.47	72.88	7.710	dcu	42.89	69.78	43.49	37.60	36.85	69.17	7.675
44.25	73.03	39.91	34.82	34.07	71.49	7.627	qmul	41.43	68.74	45.83	39.13	38.54	67.55	7.675
43.07	71.22	40.27	35.01	34.12	71.07	7.254	fbk	40.24	67.09	46.17	39.37	38.51	67.09	7.123
38.64	68.26	44.18	38.38	37.88	68.03	6.998	tubitak	35.66	63.26	50.63	43.73	42.95	63.51	6.996
37.69	68.14	45.35	39.57	38.78	69.72	7.048	lig	36.51	65.02	49.79	42.84	42.28	66.67	7.218
35.86	70.44	46.82	39.66	39.22	69.24	7.074	miracl	35.23	66.66	50.98	42.68	43.82	67.31	7.483
36.62	70.14	46.03	40.13	40.56	69.80	7.210	online	33.26	65.57	52.62	44.42	46.09	66.14	7.381
31.73	66.76	48.00	42.95	40.20	64.68	6.340	uva-isca	31.66	61.53	52.25	45.79	45.48	62.74	6.688
29.55	*59.09*	51.59	46.35	43.57	61.07	*4.987*	greyc	25.82	*53.07*	59.32	52.32	49.07	*55.67*	*4.441*
22.05	59.92	*56.49*	*50.60*	*47.51*	*57.95*	5.370	tau	*23.21*	54.43	*62.20*	*53.90*	*53.93*	55.76	5.790

BTEC French-English (BT_{FE})

"case+punc" evaluation							**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
55.65	**80.36**	**29.97**	**25.97**	**25.52**	**78.48**	**8.827**	mit	**53.68**	**77.29**	**34.36**	**29.25**	**28.91**	**75.07**	**8.961**
53.56	79.42	30.75	27.20	25.83	76.53	8.444	dsic-upv	51.44	76.38	35.54	30.53	29.28	73.28	8.543
53.59	79.26	31.14	26.98	26.19	76.80	8.529	nict	51.33	76.03	35.90	30.29	29.58	73.42	8.644
53.27	79.46	31.41	27.43	26.32	76.13	8.460	upc	50.64	76.39	36.09	30.66	29.58	72.74	8.555
52.01	79.16	33.05	28.62	28.20	76.73	8.581	tottori	49.49	76.06	37.76	31.76	31.81	73.41	8.771
51.65	79.28	32.93	28.26	28.15	76.91	8.549	kit	48.74	76.34	38.13	31.99	31.78	73.79	8.701
52.33	79.04	32.58	27.87	27.30	76.35	8.418	inesc-id	50.29	76.08	37.14	30.65	30.84	73.08	8.598
53.62	78.40	32.44	27.67	26.96	76.03	8.347	qmul	51.13	74.95	37.23	31.34	30.33	72.42	8.339
49.91	77.85	34.72	30.45	29.82	76.00	8.380	online	46.82	74.84	39.66	33.63	33.47	72.97	8.548
32.34	*69.52*	*48.27*	*42.92*	*39.81*	*65.52*	*6.636*	uva-isca	*33.26*	*65.21*	*51.92*	*44.55*	*45.51*	*64.50*	*7.276*

BTEC Turkish-English (BT_{TE})

"case+punc" evaluation							**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
54.05	**79.94**	34.36	**27.30**	27.30	76.21	**8.579**	tubitak	51.98	**76.35**	39.28	**30.40**	31.06	**73.16**	**8.705**
53.29	78.71	33.68	27.50	27.86	**76.26**	8.443	fbk	50.55	75.01	38.95	31.09	31.72	72.72	8.408
54.55	78.27	**32.75**	28.00	**26.83**	75.72	8.031	mit	**53.05**	74.60	**37.00**	30.68	**30.15**	72.94	7.887
51.07	77.12	36.56	29.85	28.64	73.40	7.782	apptek	50.14	73.38	41.61	32.41	32.47	71.02	7.860
47.80	76.86	37.22	31.21	30.75	73.77	8.009	online	44.37	73.19	42.92	34.46	35.10	70.89	8.051
51.28	75.32	36.72	31.15	29.87	72.81	7.688	qmul	48.89	71.26	41.85	34.71	33.74	69.60	7.506
48.42	75.26	37.58	31.29	30.52	73.27	7.675	limsi	46.00	70.93	43.27	34.61	34.53	70.05	7.573
47.60	74.30	40.10	33.20	32.94	71.74	7.775	cmu_qatar	44.55	69.78	46.39	37.62	37.42	68.09	7.781
36.01	*69.08*	*47.52*	*41.13*	*36.85*	*64.90*	*6.324*	uva-isca	*36.46*	*63.46*	*52.22*	*42.78*	*41.98*	*64.03*	*6.600*

C.2.2 testset_IWSLT09

· All sentence IDs of the IWSLT 2009 testset were used for calculating the automatic scores of each MT output.
· The order of the systems is the same as for the results reported in Appendix C.1.2.
· The results are obtained using the online evaluation tool: https://mastarpj.nict.go.jp/EVAL/IWSLT10/automatic/testset_IWSLT09

DIALOG Chinese-English (DT_{CE})

"case+punc" evaluation							**ASR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
33.99	60.97	**54.34**	45.40	**47.89**	67.49	5.989	msra	34.01	59.19	**55.31**	45.42	**50.36**	67.28	6.242
33.19	**63.04**	58.06	**44.73**	50.69	**67.94**	**6.341**	ict	**33.34**	**61.40**	59.87	**44.60**	54.02	**67.99**	**6.701**
31.04	60.88	59.26	46.79	52.56	66.89	5.938	i2r	32.56	58.56	60.97	47.19	56.13	65.91	6.357
27.35	57.17	60.26	48.09	54.54	65.54	5.544	iti-upv	26.16	54.28	62.70	49.68	58.02	64.23	5.700
28.58	58.03	60.66	49.18	54.52	62.90	5.673	inesc-id	28.73	55.67	62.98	50.12	58.26	61.90	5.956
28.16	57.96	63.02	49.85	57.95	63.86	5.560	tubitak	26.38	55.02	66.29	51.53	62.13	61.74	5.699
26.87	55.16	60.79	49.00	53.33	63.22	4.969	nict	27.00	52.55	63.09	50.83	56.62	61.02	5.042
28.30	56.00	63.81	51.21	57.99	61.86	5.669	postech	27.22	54.84	66.44	51.94	61.59	61.87	5.835
25.37	54.46	62.32	50.95	56.19	60.30	5.120	dcu	25.10	52.82	63.84	50.80	58.90	60.56	5.414
17.29	50.31	65.79	53.63	58.85	58.37	4.658	uva-illc	23.14	52.79	65.61	51.13	59.98	61.02	5.509
19.70	45.89	71.67	59.99	*67.96*	53.27	4.135	uva-isca	18.02	*42.15*	*75.63*	*63.02*	*72.93*	*49.43*	*4.083*
13.54	*41.78*	74.60	*63.37*	67.95	*47.86*	*3.717*	online	*16.67*	46.51	72.36	58.40	67.61	52.62	4.716

DIALOG Chinese-English (DT_{CE})

"case+punc" evaluation							**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
36.94	65.45	**51.12**	**41.23**	**44.79**	70.86	6.645	msra	**37.21**	64.02	**51.76**	41.04	46.68	70.99	6.939
34.95	**66.43**	52.23	41.26	45.91	**71.80**	**6.712**	ict	36.70	**65.99**	52.06	**39.48**	**47.89**	**71.49**	**7.275**
32.89	66.02	58.43	42.26	49.56	70.79	6.427	i2r	35.73	64.18	59.53	41.99	52.68	70.60	6.973
30.79	62.15	59.27	46.38	52.61	67.11	6.167	inesc-id	31.59	60.30	61.37	47.14	56.28	66.53	6.524
28.62	60.24	59.36	46.19	52.54	67.58	5.900	iti-upv	27.96	58.21	61.56	46.85	56.00	67.55	6.141
29.24	58.72	59.17	46.29	51.25	66.94	5.564	nict	29.89	56.75	61.08	47.10	54.25	65.87	5.745
29.84	61.95	62.19	47.50	56.62	67.43	5.910	tubitak	28.43	59.85	65.32	49.24	61.16	66.07	6.072
28.77	59.03	59.12	46.93	52.92	65.06	5.745	dcu	28.97	57.90	60.21	46.26	55.14	65.73	6.077
29.81	59.72	62.36	47.98	56.35	65.49	6.051	postech	29.32	59.27	64.44	48.60	59.72	66.25	6.223
19.50	53.93	65.68	51.07	58.46	61.64	5.303	uva-illc	25.80	56.76	66.25	49.55	60.58	64.35	5.944
19.23	47.30	*72.32*	59.03	*68.05*	54.63	*4.244*	uva-isca	*17.60*	*44.42*	*76.46*	*62.07*	*73.85*	*52.82*	*4.169*
16.25	*47.00*	70.74	*59.39*	64.48	*52.71*	4.295	online	20.18	52.44	67.65	53.54	62.88	58.03	5.414

DIALOG English-Chinese (DT_{EC})

"case+punc" evaluation							**ASR**	*"no_case+no_punc" evaluation*						
BLEU	F1	WER	PER	TER	GTM	NIST		BLEU	F1	WER	PER	TER	GTM	NIST
37.97	**65.62**	54.21	**40.38**	**47.14**	71.91	6.431	ict	**38.57**	**64.41**	56.12	**41.67**	**48.47**	**70.98**	6.532
37.34	65.02	**53.33**	41.88	47.17	70.83	5.990	msra	38.00	63.51	**55.22**	43.61	48.62	69.07	5.911
36.34	63.56	57.27	41.92	50.73	70.66	6.486	i2r	36.04	61.54	59.91	44.20	53.09	68.40	6.501
36.79	63.93	59.69	41.99	53.06	**72.16**	**6.494**	nict	36.52	62.29	62.20	43.67	55.01	70.58	**6.541**
32.02	61.51	60.22	44.17	53.75	69.12	6.044	tubitak	31.83	59.66	62.50	46.43	55.46	66.18	6.044
34.24	61.11	62.72	44.93	56.40	70.44	6.229	inesc-id	34.28	59.64	64.81	46.62	57.85	68.52	6.337
32.87	60.23	60.18	45.68	54.22	68.83	6.034	iti-upv	32.68	58.36	62.31	47.85	55.74	66.80	6.052
32.09	59.86	62.56	45.84	55.82	67.39	5.950	postech	32.18	59.01	63.61	46.79	56.01	65.82	6.023
24.57	52.42	73.33	53.93	64.58	61.41	5.207	uva-illc	24.94	50.34	75.99	56.05	66.80	60.26	5.265
23.90	53.89	74.07	53.43	66.27	62.03	5.159	dcu	23.90	52.43	75.77	54.97	67.66	59.72	5.244
21.31	49.65	73.21	56.32	67.82	56.40	5.066	online	24.25	52.21	71.14	53.81	65.78	59.60	5.436
17.54	*42.53*	*76.16*	*63.89*	*69.31*	*48.79*	*3.281*	uva-isca	*18.81*	*43.62*	*76.11*	*62.74*	*68.80*	*50.20*	*3.595*

DIALOG English-Chinese (DT_{EC})

"case+punc" evaluation							**CRR**	*"no_case+no_punc" evaluation*						
BLEU	F1	WER	PER	TER	GTM	NIST		BLEU	F1	WER	PER	TER	GTM	NIST
49.32	**73.92**	**41.72**	**31.71**	**36.74**	**79.83**	**7.819**	ict	**49.61**	**72.58**	**43.76**	**33.26**	**38.45**	**78.74**	**7.921**
42.48	70.43	48.33	36.34	41.86	76.36	6.781	msra	43.06	69.02	49.76	37.86	43.07	75.10	6.775
41.37	69.52	49.84	35.49	43.59	76.01	7.161	i2r	41.04	67.47	52.89	37.72	45.97	74.17	7.164
42.63	69.29	53.01	36.02	45.88	77.65	7.291	nict	42.71	67.73	55.58	37.71	48.11	76.52	7.377
37.95	67.43	52.74	37.56	46.18	75.38	6.836	tubitak	37.76	65.60	55.47	39.84	48.11	72.71	6.822
40.23	67.09	54.90	38.15	48.01	76.06	7.037	inesc-id	40.27	65.99	56.55	39.44	49.12	74.61	7.184
36.18	65.43	55.15	40.39	48.30	73.68	6.545	iti-upv	36.62	64.10	57.16	42.06	49.64	72.28	6.605
37.22	64.78	56.20	40.65	49.73	72.71	6.638	postech	37.43	63.68	57.91	42.04	50.92	70.89	6.705
29.75	60.89	65.20	45.13	56.87	68.74	6.062	dcu	29.22	59.18	67.90	47.18	59.18	66.38	6.099
28.44	58.45	67.73	48.46	58.60	66.12	5.775	uva-illc	29.26	56.53	70.14	50.10	60.31	65.03	5.885
25.26	54.49	68.20	51.34	62.58	61.56	5.653	online	29.19	57.67	65.92	48.38	59.91	64.96	6.133
21.04	*47.18*	*70.73*	*59.42*	*63.70*	*52.95*	*3.530*	uva-isca	*22.86*	*48.77*	*70.42*	*57.51*	*63.20*	*54.92*	*4.085*

BTEC Arabic-English (BT_{AE})

"case+punc" evaluation							**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
58.69	**79.26**	**29.73**	**26.73**	**24.81**	**78.61**	**8.560**	mit	**56.68**	**76.04**	**34.83**	**30.55**	**28.39**	**76.20**	**8.693**
55.25	76.52	33.34	30.15	27.74	76.88	8.009	rwth	53.88	72.81	38.53	34.16	31.75	74.34	8.045
54.04	76.81	33.66	30.37	27.87	76.49	8.025	dcu	52.32	72.73	39.07	34.64	32.17	73.70	8.092
52.76	76.02	34.69	30.64	28.88	75.90	8.125	qmul	50.84	71.87	41.00	35.18	33.14	72.82	8.227
52.55	76.13	35.06	31.31	29.34	75.86	8.223	apptek	51.14	72.49	40.08	34.84	33.63	73.49	8.474
51.20	76.73	36.56	32.57	30.54	75.45	8.287	online	49.05	72.86	42.77	36.79	35.17	72.86	8.601
52.04	74.72	35.05	31.51	28.81	74.13	7.494	fbk	50.01	70.49	41.00	36.00	32.96	71.11	7.411
43.35	73.71	40.85	35.74	34.20	74.17	7.906	miracl	44.57	70.76	45.61	38.61	38.34	72.90	8.440
46.47	71.87	39.92	35.50	33.01	72.64	7.132	lig	46.54	69.64	43.33	37.56	35.53	71.38	7.410
44.61	70.54	38.89	34.69	32.47	70.14	7.136	tubitak	42.98	65.54	45.48	40.09	37.35	66.45	7.256
39.37	70.23	42.81	38.72	34.41	67.68	6.820	uva-isca	40.88	65.10	47.37	41.78	39.58	66.41	7.332
39.22	65.02	43.76	40.13	36.90	68.00	*5.835*	greyc	35.55	*59.18*	51.02	46.14	42.49	63.88	*5.442*
26.79	*64.69*	*49.07*	*44.23*	*39.78*	*63.48*	5.895	tau	*29.27*	60.63	*53.88*	*46.73*	*45.21*	*63.00*	6.551

BTEC French-English (BT_{FE})

"case+punc" evaluation							**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
63.62	**82.99**	**25.87**	**23.52**	**21.89**	**82.13**	**9.442**	mit	**61.78**	**79.99**	**29.57**	**26.46**	**24.98**	**80.23**	**9.789**
63.59	81.98	26.66	24.17	22.14	80.44	9.142	dsic-upv	62.21	78.98	30.07	26.84	25.01	78.43	9.456
62.00	81.92	27.26	23.91	22.88	79.89	9.170	upc	59.46	79.08	31.70	27.51	26.01	77.62	9.495
61.58	81.45	27.67	24.64	23.03	80.34	9.167	nict	59.65	78.40	31.37	27.47	26.10	78.24	9.488
61.80	81.70	28.33	24.73	23.49	80.42	9.109	qmul	60.38	78.78	32.03	27.42	26.58	78.54	9.461
58.94	81.73	29.32	25.70	24.76	80.15	9.255	tottori	56.70	78.44	33.60	28.67	28.24	77.95	9.647
61.07	80.95	29.97	26.87	25.19	80.22	9.243	online	58.97	77.95	33.89	29.82	28.85	78.60	9.681
59.54	80.76	28.59	25.82	23.97	79.15	8.994	inesc-id	57.50	77.50	32.69	28.79	27.40	77.20	9.367
58.90	81.80	29.61	26.40	25.04	79.38	9.113	kit	56.14	78.71	34.49	29.93	28.70	77.33	9.485
38.59	*72.40*	*44.80*	*40.19*	*36.65*	*69.21*	*7.289*	uva-isca	*39.78*	*68.07*	*49.25*	*42.41*	*42.18*	*68.81*	*7.972*

BTEC Turkish-English (BT_{TE})

"case+punc" evaluation							**CRR**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
57.63	**81.66**	31.37	25.95	24.35	78.12	**8.743**	tubitak	55.78	**78.50**	36.06	**28.46**	27.94	**76.23**	**9.124**
57.70	80.22	31.61	**25.57**	24.99	**78.76**	8.612	fbk	55.56	77.15	36.65	28.91	28.49	76.18	8.822
60.21	79.45	**30.16**	25.93	**23.44**	77.87	7.781	mit	**59.18**	76.07	**34.59**	29.26	**26.58**	75.81	7.728
52.97	76.95	35.55	29.43	27.71	74.80	7.750	limsi	50.75	73.02	41.14	33.01	31.84	72.13	7.811
52.64	77.92	36.66	29.45	27.87	74.73	7.748	apptek	50.87	74.20	42.92	32.64	31.96	73.60	7.979
53.54	75.50	36.20	30.93	29.06	73.93	7.532	qmul	50.99	71.44	42.17	35.95	32.81	71.11	7.507
49.42	77.09	38.29	31.62	30.26	74.60	7.922	online	46.71	73.17	44.59	34.88	34.78	72.89	8.178
49.06	75.78	39.42	32.30	31.40	74.03	7.880	cmu_qatar	46.48	71.17	45.88	37.02	36.20	71.35	8.090
36.07	*68.28*	*48.91*	*42.21*	*35.70*	*63.97*	*5.899*	uva-isca	*38.08*	*62.46*	*53.37*	*43.50*	*41.34*	*64.09*	*6.341*

Appendix D. Evaluation Metric Correlation

· The correlation between evaluation metrics are measured using the *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

· *Z-avg* is the average system score of the best *z*-transformed automatic evaluation metric subset obtained for the respective translation task. The z-avg scores are given for all MT systems including (*w/ online*) and excluding (*w/o online*) the online translation system).

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

(testset_IWSLT10)

DT_{CE} (ASR)	**z-avg**		single metrics						
	(w/ online)	(w/o online)	BLEU	METEOR	WER	PER	TER	GTM	NIST
Ranking	**0.7272**	(0.9090)	0.5174	0.6503	-0.7062	-0.6713	0.6433	-0.7202	0.3986
NormRank	**0.8216**	(0.9340)	0.6363	0.7587	-0.7937	-0.7412	0.7027	-0.7972	0.5104
BestRankDiff	**0.6909**	(0.7939)	**0.6909**	0.6181	-0.6727	-0.5181	0.4818	-0.6818	0.3727

DT_{CE} (CRR)	**z-avg**		single metrics						
	(w/ online)	(w/o online)	BLEU	METEOR	WER	PER	TER	GTM	NIST
Ranking	0.7342	(0.8454)	0.6433	**0.7412**	-0.6923	-0.6853	0.7272	-0.6853	0.5314
NormRank	0.7762	(0.9000)	0.6783	**0.7832**	-0.7272	-0.7272	0.7622	-0.7202	0.5804
BestRankDiff	**0.8741**	(0.9727)	0.7902	0.8391	-0.8531	-0.8111	0.8111	-0.8461	0.6153

DT_{EC} (ASR)	**z-avg**		single metrics						
	(w/ online)	(w/o online)	BLEU	F1	WER	PER	TER	GTM	NIST
Ranking	**0.9510**	(0.9636)	0.9300	0.9160	-0.9300	-0.8881	0.9160	-0.9090	0.9160
NormRank	**0.9702**	(0.9613)	0.9562	0.9423	-0.9458	-0.9248	0.9423	-0.9318	0.9423
BestRankDiff	**0.9090**	(0.9151)	0.8545	0.8909	-0.8818	-0.8909	0.8545	-0.8636	0.8818

DT_{EC} (CRR)	**z-avg**		single metrics						
	(w/ online)	(w/o online)	BLEU	F1	WER	PER	TER	GTM	NIST
Ranking	**0.8881**	(0.9272)	0.8601	0.8461	-0.8601	-0.8601	0.8041	-0.8041	0.8391
NormRank	**0.8951**	(0.9090)	0.8601	0.8531	-0.8531	-0.8741	0.8181	-0.7972	0.8601
BestRankDiff	**0.9230**	(0.9727)	0.8951	0.8881	-0.9090	-0.8951	0.8391	-0.8601	0.8741

BT_{AE} (CRR)	**z-avg** (w/ online)	(w/o online)	single metrics BLEU	METEOR	WER	PER	TER	GTM	NIST
Ranking	0.8241	(0.9370)	0.6483	0.7802	-0.6483	-0.6153	0.7582	-0.5329	**0.8351**
NormRank	0.9203	(0.9755)	0.8049	0.8928	-0.7994	-0.7774	0.8598	-0.7225	**0.9258**
BestRankDiff	0.8516	(0.9090)	0.6923	0.8461	-0.6923	-0.6648	0.7692	-0.6098	**0.8626**

BT_{FE} (CRR)	**z-avg** (w/ online)	(w/o online)	single metrics BLEU	METEOR	WER	PER	TER	GTM	NIST
Ranking	0.5393	(0.8333)	-0.1242	0.2727	0.0393	0.1272	**0.5636**	0.1151	0.5515
NormRank	0.5090	(0.7916)	-0.0242	0.3212	-0.0333	0.0606	**0.5272**	0.0242	0.4787
BestRankDiff	0.5151	(0.8833)	0.2818	0.4909	-0.3060	-0.2242	0.4303	-0.2969	**0.5272**

BT_{TE} (CRR)	**z-avg** (w/ online)	(w/o online)	single metrics BLEU	METEOR	WER	PER	TER	GTM	NIST
Ranking	**0.8333**	(1.0000)	0.4523	0.7380	-0.5238	-0.6250	0.7857	-0.4523	0.8095
NormRank	**0.8333**	(1.0000)	0.4523	0.7380	-0.5238	-0.6250	0.7857	-0.4523	0.8095
BestRankDiff	0.7857	(0.9285)	0.3333	0.6666	-0.3809	-0.5654	0.7380	-0.3333	**0.8095**

(Z-avg Metric Combinations)

Task		**Ranking**	**NormRank**	**BestRankDiff**
DT_{CE}	(ASR)	WER, TER, NIST	METEOR, TER	WER, TER
	(CRR)	BLEU, METEOR	BLEU, METEOR	METEOR. WER, PER, TER
DT_{EC}	(ASR)	F1, NIST	F1, NIST	F1, PER
	(CRR)	F1, NIST	F1, PER	BLEU, NIST
BT_{AE}	(CRR)	GTM, NIST	GTM, NIST	BLEU, METEOR, NIST
BT_{FE}	(CRR)	METEOR, GTM	METEOR, GTM	METEOR, GTM
BT_{TE}	(CRR)	METEOR, NIST	METEOR, NIST	METEOR, NIST