Overview of the IWSLT 2009 Evaluation Campaign

Michael Paul

National Institute of Information and Communications Technology
Hikaridai 2-2-2, Keihanna Science City, 619-0288 Kyoto, Japan
Michael.Paul@nict.go.jp

Abstract

This paper gives an overview of the evaluation campaign results of the *International Workshop on Spoken Language Translation* (IWSLT) 2009[1]. In this workshop, we focused on the translation of task-oriented human dialogs 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.

The translation directions were English into Chinese and vice versa for the *Challenge Task*, and Arabic, Chinese, and Turkish, which is a new edition, into English for the standard *BTEC Task*. In total, 18 research groups participated in this year's event. Automatic and subjective evaluations were carried out in order to investigate the impact of task-oriented human dialogs on automatic speech recognition (ASR) and machine translation (MT) system performance, as well as the robustness of state-of-the-art MT systems for speech-to-speech translation in a dialog scenario.

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, but oriented 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 and innovative technologies for the translation of automatic speech recognition results from read/spontaneous-speech input, and monolingual dialog conversations [1].

The focus of this year's evaluation campaign was the translation of task-oriented cross-lingual human dialogs 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. For the *Challenge Task*, IWSLT participants had to translate both the Chinese and the English outputs of the automatic speech recognizers (lattice, N/1BEST) into English and Chinese, respectively.

Like in previous IWSLT events, a standard *BTEC Task* was provided for the translation of Arabic and Chinese into English. For IWSLT 2009, however, the BTEC Task focused on text input only, i.e. no automatic speech recognition results had to be translated this time. Another innovative aspect of this year's edition was that Turkish was used as an input language for the first time, attracting new groups to participate in this year's event.

For the IWSLT evaluation campaigns of 2007 and 2008, participants were encouraged to list linguistic resources and tools that could be shared by the participants. However, it was difficult to distinguish whether system improvements were triggered by better suited (or simply more) language resources or by improvements in the underlying decoding algorithms and statistical models. In order to focus more on the research aspects, only the supplied resources listed in Appendix B were allowed for the training of the MT engines for the IWSLT 2009 official run submission.

All primary run submissions were judged and compared according to the *Ranking* metric where human graders were asked to rank whole sentence translations from best to worst relative to the other choices [2]. Then, human assessments of *Fluency* and *Adequacy* [3] were carried out for the top-ranked MT outputs of each translation task. In addition, a modified version of the *Adequacy* metrics that takes into account information beyond the current input sentence was applied to the translation results of the *Challenge Task* in order to judge the overall translation quality of a given MT output in the context of the respective dialog.

The translation quality of all primary and contrastive run submissions was also evaluated using various standard automatic evaluation metrics. In addition to the single-metric scores, all automatic metric scores for the MT output were combined by normalizing each metric score distribution and the final system scores were obtained by calculating the average of all normalized metric scores. Based on the evaluation results, the impact of task-oriented human dialogs on

[1] http://mastarpj.nict.go.jp/IWSLT2009

automatic speech recognition (ASR) and machine translation (MT) system performance as well as the robustness of state-of-the-art MT systems towards speech-to-speech translation in a dialog scenario was investigated.

2. IWSLT 2009 Evaluation Campaign

This year's IWSLT campaign took place during the period of April-July 2009 and featured five different translation tasks:

Table 1: Translation Tasks

Task	Translation Direction		Participants
Challenge	English-Chinese	CT_{EC}	7
	Chinese-English	CT_{CE}	7
BTEC	Arabic-English	BT_{AE}	9
	Chinese-English	BT_{CE}	12
	Turkish-English	BT_{TE}	7

In total, 18 research groups from all over the world[2] participated in the event, producing a total of 35 machine translation engines for the above five translation tasks. For the *Challenge Task*, one participant submitted its runs after the subjective evaluation period, so only 6 systems were assessed by humans. Information on the research groups, the utilized translation systems, and translation task participation is summarized in Appendix A. Most participants used 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.

For training purposes, the two spoken language corpora described in Section 2.1 were provided to all participating research groups. The supplied resources for IWSLT 2009 were released two months ahead of the official run submissions period. The official run submission period was limited to two weeks. 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). Each participant registered for the *Challenge Task* had to translate in both translation directions (English-Chinese AND Chinese-English). In total, 35 primary runs and 67 contrastive runs were submitted. After the official run submission period, the organizers set-up an online evaluation server[3] that could be used by the participants to carry out additional experiments on the evaluation testset. The schedule of the evaluation campaign is summarized in Table 2.

2.1. IWSLT 2009 Spoken Language Corpus

The IWSLT 2009 evaluation campaign was carried out using two multilingual spoken language corpora. (1) The *Basic Travel Expression Corpus* (BTEC*) contains tourism-related sentences similar to those which are usually found in phrase books for tourists traveling abroad [4]. Parts of this corpus were already used in previous IWSLT evaluation campaigns. Besides the sentence-aligned training corpus, the evaluation data sets of previous workshops including multiple reference translations were provided to the participants as a development corpus. (2) The *Spoken Language Databases* (SLDB) corpus is a collection of human-mediated cross-lingual dialogs in travel situations and parts of this corpus were provided to the participants of the *Challenge Task*.

Table 2: Evaluation Campaign Schedule

Event	Date
Training Corpus Release	Jun 19, 2009
Development Corpus Release	Jun 19, 2009
Evaluation Corpus Release	Aug 14, 2009
Result Submission Due	Aug 28, 2009

The monolingual and bilingual language resources that could be used to train the translation engines for the primary runs were limited to the supplied corpus for each translation task. All resources of the BT_{CE} translation task were also permitted for the *Challenge Task*.

2.1.1. Supplied Resources

Details of the IWSLT 2009 spoken language corpus are given in Appendix B.1. The first two columns specify the given data set and provide its type. Besides the "text" resources, all data sets consist of the ASR output (lattices, 1/NBEST lists) and manual transcriptions of the respective *read-speech* or *spontaneous-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. The English and Chinese target sentences were tokenized according to the evaluation specifications used for this year's evaluation campaign. "*Word token*" refers to the number of words in the corpus and "*word type*" refers to the vocabulary size. The number of reference translations used for the evaluation of the respective evaluation data sets is given in the "*ref.trans*" column. In addition, all translation tasks that permitted the usage of the respective resources are listed in the "*task*" column.

For this year's evaluation campaign, parts of the Arabic (A), Chinese (C), English (E), and Turkish[4] (T) subsets of the BTEC* corpus were used. The participants were supplied with a training corpus of 20K sentence pairs which covered the same sentence IDs for CT_{EC}, CT_{CE}, BT_{AE}, $\text{BT}_{CE}$and BT_{TE}. The amount of development corpus data sets differed between the translation tasks. The evaluation data sets of the *BTEC Task*, consisted of 469 randomly selected sentences from parts of the BTEC* corpus reserved for evaluation purposes. For automatic evaluation, up to 7 (16) reference trans-

[2]China: 2, France: 3, Ireland: 1, Italy: 1, Japan: 3, Singapore: 2, Spain: 2, USA: 2, Turkey: 2

[3]https://mastarpj.nict.go.jp/EVAL/IWSLT09/automatic/testset_IWSLT09

[4]The Turkish data sets were kindly provided by *The Scientific and Technological Research Council of Turkey* (TUBITAK-UEKAE).

lations were used for the evaluation (development) data sets for each of the BTEC⋆ translation tasks.

In addition, 394 SLDB training dialogs (10k sentence pairs) consisting of the transcripts of the uttered sentences and the simultaneous interpreter translations were provided to the participants of the *Challenge Task*. Speech data sets (ASR output) that could be used to adopt the MT systems to the new domain were provided for 10 dialogs (410 sentences). The evaluation data sets consisted of 27 dialogs covering 405 Chinese utterances and 393 English utterances and up to 4 reference translations used for each of the target languages.

ASR engines provided by the organizers were applied to the above speech data sets and produced word lattices from which NBEST/1BEST lists were extracted automatically using publicly available tools. Participants were free to choose the ASR output condition that best suited their machine translation technology for the input of the respective MT engine. In addition, the cleaned transcripts of the speech recordings, i.e., the *correct recognition results* (CRR), were also given to all participants for translation. Word segmentations according to the output of the ASR engines were also provided for all supplied resources.

Appendix B.2 summarizes the out-of-vocabulary (*OOV*) rates of the respective data sets, i.e., the percentage of words in the evaluation data that do not appear in the training data. The OOV rates are listed for all source languages and input conditions (CRR, 1BEST, NBEST). Concerning the *BTEC Tasks*, the amount of OOV is generally larger for Turkish (6%-7%) than for Arabic (5%-9%) or Chinese (3%-5%). Concerning the $devset$ and $testset$ of the *Challenge Task*, the OOV rates for the BTEC⋆ training corpus are smaller than the ones for the SLDB training corpus.

In order to get an idea of how difficult the IWSLT 2009 translation tasks were, we used the *SRI Language Modeling Toolkit*[5] to train standard 5-gram language models on the target language side of the supplied training corpora and evaluated the *entropy* and *total entropy*, i.e., the *entropy* multiplied by *word counts*, for each language in the respective evaluation data sets. The total entropy figures given in Appendix B.3 indicate that the *Challenge Task* can be expected to be more difficult to translate than the *BTEC Task* which was confirmed for the CRR inputs by the automatic evaluation results listed in Appendix D.

The recognition accuracies of the utilized ASR engines for the *Challenge Task* data sets are summarized in Appendix B.4. The *lattice accuracy* figures show the percentage of correct recognition results contained in the lattices, and the *1BEST accuracy* is the accuracy of the best path extracted from each lattice. The *word accuracies* of the utilized ASR engines ranged between 89%-94% (lattice) and 81%-85% (1BEST), where the percentages of correctly recognized sentences (*sentence accuracy*) ranged between 50%-74% (lattice) and 29%-48% (1BEST).

[5]http://www.speech.sri.com/projects/srilm

2.2. Evaluation Specifications

The *case+punc* evaluation specifications for IWSLT 2009 were defined as:

- case-sensitive
- with punctuation marks (. , ? ! ") tokenized

For the convenience of the participants, automatic evaluation scores were also calculated for the *no_case+no_punc* evaluation specifications:

- case-insensitive (lower-case only)
- no punctuation marks (remove . , ? ! ")

The focus of this year's evaluation campaign was the translation of speech data. Therefore, all input data files of the *Challenge Task* were case-insensitive and without punctuation information. Concerning the *BTEC Task* data sets, truecase and punctuation information were provided for all training and development data sets. This could be used together with the *Challenge Task* training data sets for recovering case/punctuation information according to the *case+punc* evaluation specifications for the *Challenge Task*. Instructions[6] on how to build a baseline tool for case/punctuation insertions using the *SRI Language Modeling Toolkit* was provided to all participants.

2.2.1. 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)"* [2]. 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*). 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 [5] were calculated for each run submission. The *Ranking* metric was applied to all primary runs submitted by the participants for each of the translation tasks.

Although the *Ranking* metric requires relatively low evaluation costs because multiple systems are judged simultaneously, the *Ranking* scores can only define a relative order of the judged MT systems. Its usage alone is not sufficient to provide information on the overall (absolute) translation quality of the respective MT systems. In the extreme case, all MT systems could be good or all MT systems could be bad. Moreover, the *Ranking* metric compares a single system against more than one other system simultaneously, but the points of reference, i.e., the subset of other systems ranked together, might differ for each system. Therefore, a direct comparison between two MT systems using *Ranking* and *NormRank* scores might be difficult.

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

In order to overcome the short-comes of the *Ranking* metrics, additional subjective evaluation metrics were applied. Similar to last year's IWSLT edition, a *paired-comparison* evaluation based on the obtained *Ranking* results was carried out in order to compare two MT systems directly, i.e., given two MT system outputs, the first system was compared against the second system on a sentence-by-sentence basis according to the *Ranking* grades where both systems were ranked together. The *gain* of the first system towards the second system was defined as the difference between the number of translations ranked better and the number of translations ranked worse divided by the total amount of gradings carried out together. Moreover, 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 [3]. The *Fluency* and *Adequacy* judgments consisted of one of the grades listed in Table 3. 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.

Table 3: Human Assessment

Fluency	
4	Flawless C/E
3	Good C/E
2	Non-native C/E
1	Disfluent C/E
0	Incomprehensible

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

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 *Challenge 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 *Adequacy* grades listed in Table 3. In case that parts of the information were omitted in the current translation, but they could be understood in the context of the given dialog, such omission should 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 *NormRank* evaluation results. In addition, the translation results of each translation task were *pooled*, i.e., in case of identical translations of the same source sentence by multiple engines, the pooled translation was graded once, and the respective rank was assigned to all MT engines with the same output.

The subjective evaluations were carried out by paid evaluation experts (4x English, 3x Chinese) and a small number of volunteers provided by this year's participants. For the final metric scores, we selected the judgements of the three most self-consistent graders of each translation task and 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.

2.2.2. 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 κ = 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 [6] is given in Table 4.

Table 4: Interpretation of κ Coefficient [6]

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

2.2.3. Automatic Evaluation

The automatic evaluation of run submissions was carried out using the seven standard automatic evaluation metrics listed in Table 5. Unfortunately, the METEOR metrics could not be applied to the evaluation of the CT_{EC} translation task, because the METEOR metrics requires language-specific parameters tuned on human judgments to calculate final scores which were not available for CT_{EC}. However, the utilized METEOR script provided additional system level information such as *unigram precision*, *recall*, and *f1 score* which

Table 5: 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) [7] → 'mteval-v13.pl'
NIST:	a variant of BLEU using the arithmetic mean of weighted n-gram precision values. Scores are positive with 0 being the worst possible [8] → 'mteval-v13.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) [9] → 'meteor-v0.8.3'
GTM:	measures the similarity between texts by using a unigram-based F-measure. Scores range between 0 (worst) and 1 (best) [10] → '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 [11]
PER:	Position independent WER: a variant of WER that disregards word ordering [12]
TER:	*Translation Edit Rate*: a variant of WER that allows phrasal shifts [13] → 'tercom-0.7.25'

are language independent. For the automatic evaluation results of the CT_{EC} task, we list the *f1* score instead of the METEOR score used for CT_{CE} and all *BTEC tasks*.

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. An alternative method to combine the results of multiple evaluation metrics used for IWSLT 2009 is to calculate the average system rank (**r-avg**) that a MT system achieved based on the system rankings of each automatic evaluation metrics.

2.2.4. 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 [14]. In this year's evaluation, 2000 iterations were used for the analysis of the automatic evaluation results.

2.2.5. 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. Evaluation Results

The evaluation results of the IWSLT 2009 workshop are summarized in Appendix C (*human assessment*) and Appendix D (*automatic evaluation*). The rank correlation coefficients of subjective and automatic evaluation results are given in Appendix E. For each evaluation metric, the best correlation coefficient of each translation task is marked in boldface.

3.1. Subjective Evaluation Results

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 of the evaluators selected to calculate the final scores are given in Table 6.

Table 6: Grader Consistency

Metric	**Intra-Grader** κ			**Inter-Grader** κ		
	CT_{EC}	CT_{CE}	BTEC	CT_{EC}	CT_{CE}	BTEC
Ranking	0.59	0.52	0.60	0.50	0.53	0.58
Fluency	0.75	0.68	0.75	0.39	0.50	0.50
Adequacy	0.71	0.64	0.74	0.39	0.44	0.47
Dialog	0.80	0.59	–	0.38	0.45	–

The obtained overall *intra-grader* κ coefficients were high and showed that all graders submitted very consistent evaluation results achieving *substantial agreement* levels for most of the translation tasks for the single-system overall translation quality metrics. Only moderate agreement was achieved for the *Ranking* assessments where evaluators had to grade multiple system outputs.

Concerning the *inter-grader* consistency, the κ coefficients are much lower. However, moderate agreement was achieved for all translation tasks having English as the target language. The lowest agreement was achieved for the Chinese translations of the *Challenge Task*. Concerning the type of evaluation metrics, the levels of *inter-grader* agreement were: *Ranking* > *Fluency* > *Adequacy* = *Dialog*.

3.1.1. Ranking Performance

The results of the IWSLT 2009 *Ranking* evaluation are summarized in Appendix C.2. For each translation task, the

MT system rankings are listed according to the *Ranking* and *NormRank* scores. For the *Challenge Task*, all MT systems are ranked in the same order for both metrics, but the MT systems of the *BTEC tasks* are ranked quite differently even though both metrics agree at least on the top-ranked MT system aside from the BT_{TE} task.

In order to get an idea of how different the performances of two given systems are, we performed a paired-comparison for all system combinations and calculated the gain of the first system against the second system as the ratio of improved translations, i.e., the difference of *better* translations and *worse* translations divided by the total number of translations judged together. The results are listed in Appendix C.3, where the order of the system for each translation task is defined by the positive gains of the *pairwise comparison*. For CT_{EC}, the obtained MT system ranks are identical to the *Ranking* and *NormRank*. However, the system orders differ largely for the CT_{CE} and all *BTEC tasks*. Although the *pairwise comparison* metric is an appropriate measure to directly judge differences in system performance on sentence-level for two given MT systems, it cannot be generalized to rank multiple MT system outputs. For example, the IWSLT 2009 results of the *BTEC task* show negative gains for the upv vs. $bmrc$ system and the $tottori$ vs. $tokyo$ system.

In order to avoid these inconsistencies, we calculated the *BestRankDiff* scores that rank all MT systems of each translation task according to the percentage of translations the top-scoring system gains to the respective system. The alternative MT system rankings based on the *BestRankDiff* scores are given in Appendix C.4.

3.1.2. Fluency/Adequacy/Dialog Performance

The results of the IWSLT 2009 *Fluency/Adequacy/Dialog* evaluation for the top-ranked MT submitted of each translation task are summarized in Appendix C.1. For the *Challenge Task*, both the ASR and CRR primary run submissions were graded together on the same evaluation page. The *BTEC tasks* were evaluated separately.

The highest *Fluency* and *Adequacy* scores were achieved for the BT_{TE} task followed by the BT_{CE} and BT_{AE} tasks. The lower *Challenge Task* results for the correct recognition results confirm that the task-oriented human dialogs in travel situations are more difficult to translate than the standard BTEC* data sets, as predicted by the total entropy figures in Appendix B.3.

Comparing the *Adequacy* and the *Dialog* results obtained for the *Challenge Task*, higher scores were achieved for the *Dialog* metrics consistently for all ASR and CRR translation results. This 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.

Moreover, a large drop in system performance for the ASR compared to the CRR results can be seen for both *Challenge Task* translation tasks. However, the CT_{EC} results are more affected by recognition errors than the CT_{CE} results, although the 1BEST word and sentence recognition accuracies for the Chinese *testset* utterances are far worse than the ones for the English utterances. The reason for this is that the the top-ranked *nlpr* system used the NBEST list to produce the translation outputs, thus benefiting from the higher recognition accuracy figures for the Chinese input lattices.

3.2. Automatic Evaluation Results

The automatic evaluation results of all MT engines using the *case+punc* evaluation specifications, i.e., *case-sensitive with punctuation marks tokenized*, as well as the *no_case+no_punc* evaluation specifications, i.e., *lowercase without punctuation marks*, are listed in Appendix D. The MT systems are ordered according to the *z-avg* score, i.e., the average of all normalized evaluation metric scores obtained for the respective MT output, for the *case+punc* evaluation specifications.

Appendix D.1 list the evaluation results based on the statistical significance test described in Section 2.2.4. For all automatic evaluation metrics, the MT system scores are calculated as the mean score of all metric scores obtained for 2000 iterations of the same random sampling with replacement from the evaluation data set. 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** (*italic*).

Besides the BT_{AE} task, the MT systems of all translation tasks that obtain the highest *z-avg* scores agree with the top-ranked systems according to the the human assessment results. However, the MT system rankings based on the automatic evaluation scores differ largely from those of the subjective evaluation scores.

In addition to the significance test results, the automatic evaluation scores obtained for the full testset are listed in Appendix D.2. Concerning the order of the MT system rankings, different rankings were obtained for the CRR translation results of the CT_{CE} task and the BT_{CE} task.

3.3. Evaluation Metric Correlations

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 E. For each translation task, the MT system ranking obtained for the subjective *Ranking*, *NormRank*, *BestRankDiff* metrics and all investigated automatic evaluation metrics including the two metric combination methods (*z-avg*, *r-avg*) are compared. For the *Challenge Task*, the correlation coefficients for ASR and CRR translation results are calculated separately as well as for the merged set of ASR and CRR MT outputs.

The results show that the highest correlation to automatic evaluation metrics is generally achieved for the *NormRank*

followed by the *Ranking* metric for all translation tasks where 9 or more MT systems are compared (CT_{EC}, CT_{CE}, BT_{CE}, BT_{AE}). The highest correlation coefficient for translation tasks with less MT systems involved ($CT_{EC}{}^{ASR}$, $CT_{EC}{}^{CRR}$, $CT_{CE}{}^{ASR}$, $CT_{CE}{}^{CRR}$, BT_{TE}), are achieved for the *BestRankDiff* metric.

Comparing the predictive power of the automatic evaluation metrics, we can see that the *METEOR* (*f1*) metric achieves a very high correlation with the *NormRank* and *Ranking* metrics for most of the translation tasks. Moreover, *TER*, followed by *BLEU*, performs best for all tasks having Chinese as the target language.

In addition to the combination of all investigated automatic evaluation metrics, we also calculated the rank correlation coefficient for all subsets of metric combinations. For the *Challenge Task*, however, the combination of all seven metrics performed best with the exception of the $CT_{CE}{}^{CRR}$ translation task where the combination of the *METEOR* and *TER* metrics achieved a perfect correlation. On the other hand, significantly higher correlations could be achieved for the *BTEC tasks* when the *TER* or *PER* metrics were combined with *METEOR* (BT_{AE}, BT_{CE}) or *BLEU* (BT_{TE}).

Concerning the combination of automatic evaluation metrics, better correlations were achieved for the average of normalized scores (*z-avg*) compared to the average MT system ranking (*r-avg*) for the *NormRank* and *Ranking* metrics. For the later metrics, the *z-avg* score based on the optimal subset combination outperformed even the best-performing single automatic evaluation metric for all *BTEC tasks* (BT_{CE}: ρ=0.5564, BT_{AE}: ρ=0.5000, BT_{TE}: ρ=0.8929). Concerning the *BestRankDiff* metric, the *r-avg* metric outperforms *z-avg* for most translation tasks.

4. Conclusion

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 task-oriented human dialogs in travel situations. Various innovative ideas were explored, most notably *advanced techniques for morphological preprocessing*, *improved statistical modeling techniques integrating syntactic and source language information*, *cross domain adaptation*, *new parameter optimization techniques*, *lattice decoding*, *system combinations*, and *semi-supervised reranking methods of NBEST lists*. 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.

5. Acknowledgments

I would like to thank all the people involved in the preparation of this workshop and the subjective evaluation task. In particular, I would like to thank Shigeki Matsuda from NICT for preparing the speech data sets and generating the ASR outputs. Special thanks to the TUBITAK-UEDIN team, for providing us with the Turkish data sets and to Chris Callison-Burch for letting us use the browser-interface scripts of the subjective *Ranking* metrics. In addition, I thank all the paid exports and volunteers (including members of *apptek*, *fbk*, *lig*, *uw*, and *nict*) who carried out the human assessment of the translation outputs. I also thank the program committee members for reviewing a large number of MT system descriptions and technical paper submissions. Last, but not least, I thank all research groups for their active participation in the IWSLT 2009 evaluation campaign and for making the IWSLT 2009 workshop a success.

6. References

[1] M. Paul, "Overview of the IWSLT 2008 evaluation campaign," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 1–17.

[2] C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder, "(Meta-) Evaluation of Machine Translation," in *Proceedings of the Second Workshop on SMT*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 136–158. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-0718

[3] 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.

[4] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14(5), pp. 1674–1682, 2006.

[5] 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.

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

[7] 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.

[8] 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.

[9] A. Lavie and A. Agarwal, "METEOR: An automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of Workshop on SMT at the 45th Annual Meeting of the Association*

of Computational Linguistics, Prague, Czech Republic, 2007, pp. 228–231.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] S. Köprü, "AppTek Turkish-English Machine Translation System Description for IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 19–23.

[16] M. R. Costa-Jussà and R. E. Banchs, "Barcelona Media SMT system description for the IWSLT 2009: introducing source context information," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 24–28.

[17] Y. Ma, T. Okita, O. Çetinoğlu, J. Du, and A. Way, "Low-Resource Machine Translation Using MaTrEx: The DCU Machine Translation System for IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 29–36.

[18] N. Bertoldi, A. Bisazza, M. Cettolo, G. Sanchis-Trilles, and M. Federico, "FBK @ IWSLT-2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 37–44.

[19] Y. Lepage, A. Lardilleux, and J. Gosme, "The GREYC Translation Memory for the IWSLT 2009 Evaluation Campaign: one step beyond translation memory," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 45–49.

[20] X. Duan, D. Xiong, H. Zhang, M. Zhang, and H. Li, "I^2R's Machine Translation System for IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 50–54.

[21] H. Mi, Y. Liu, T. Xia, X. Xiao, Y. Feng, J. Xie, H. Xiong, Z. Tu, D. Zheng, Y. Lu, and Q. Liu, "The ICT SMT Systems for the IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 55–59.

[22] F. Bougares, L. Besacier, and H. Blanchon, "LIG approach for IWSLT09 : Using Multiple Morphological Segmenters for Spoken Language Translation of Arabic," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 60–64.

[23] H. Schwenk, L. Barrault, Y. Estève, and P. Lambert, "LIUM's SMT Systems for IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 65–70.

[24] W. Shen, B. Delaney, A. Aminzadeh, T. Anderson, and R. Slyh, "The MIT-LL/AFRL IWSLT-2009 System," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 71–78.

[25] M. Utiyama, H. Yamamoto, and E. Sumita, "Two methods for stabilizing MERT: NICT at IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 79–82.

[26] M. Li, J. Zhang, Y. Zhou, and C. Zong, "The CASIA SMT System for IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 83–90.

[27] P. Nakov, C. Liu, W. Lu, and H. T. Ng, "The NUS SMT System for IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 91–98.

[28] X. Wu, T. Matsuzaki, N. Okazaki, Y. Miyao, and J. Tsujii, "The UOT System: Improve String-to-Tree Translation Using Head-Driven Phrasal Structure Grammar and Predicate-Argument Structures," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 99–106.

[29] J. Murakami, M. Tokuhisa, and S. Ikehara, "Statistical Machine Translation adding Pattern-based Machine translation in Chinese-English Translation," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 107–112.

[30] C. Mermer, H. Kaya, and M. U. Doğan, "The TÜBITAK-UEKAE SMT System for IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 113–117.

[31] G. Gascó and J. A. Sánchez, "UPV Translation System for IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 118–123.

[32] M. Yang, A. Axelrod, K. Duh, and K. Kirchhoff, "The University of Washington Machine Translation System for IWSLT 2009," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 124–128.

[33] A. Bisazza and M. Federico, "Morphological Pre-Processing for Turkish to English SMT," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 129–135.

[34] M. Cmejrek, B. Zhou, and B. Xiang, "Enriching SCFG Rules Directly From Efficient Bilingual Chart Parsing," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 136–143.

[35] K. Hayashi, T. Watanabe, H. Tsukada, and H. Isozaki, "Structural Support Vector Machines for Log-Linear Approach in SMT," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 144–151.

[36] H. Hoang, P. Koehn, and A. Lopez, "A Unified Framework for Phrase-Based, Hierarchical, and Syntax-Based SMT," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 152–159.

[37] G. Sanchis-Trilles, M. Cettolo, N. Bertoldi, and M. Federico, "Online Language Model Adaptation for Spoken Dialog Translation," in *Proc. of IWSLT*, Tokyo, Japan, 2009, pp. 160–167.

[38] C. Hori, S. Sakti, M. Paul, N. Kimura, Y. Ashikari, R. Isotani, E. Sumita, and S. Nakamura, "Network-based Speech-to-Speech Translation," in *Proc. of IWSLT*, Tokyo, Japan, 2009, p. 168.

Appendix A. MT System Overview

Research Group	MT System Description	Type	System	Submissions
Apptek, Inc. (Turkey)	AppTek Turkish-English Machine Translation System Description for IWSLT 2009 [15]	SMT	apptek	BT_{TE}
Barcelona Media (Spain)	Barcelona Media SMT system description for the IWSLT 2009: introducing source context information [16]	SMT	bmrc	BT_{AE}, BT_{CE}
Dublin City University, School of Computing (Ireland)	Low-Resource Machine Translation Using MaTrEx: The DCU Machine Translation System for IWSLT 2009 [17]	Hybrid SMT	dcu	BT_{CE}, BT_{TE}, CT_{CE}, CT_{EC}
Fondazione Bruno Kessler, Ricerca Scientifica e Tecnologica (Italy)	FBK @ IWSLT-2009 [18]	SMT	fbk	BT_{AE}, BT_{TE}, CT_{CE}, CT_{EC}
University of Caen Basse-Normandie, GREYC (France)	The GREYC Translation Memory for the IWSLT 2009 Evaluation Campaign: one step beyond translation memory [19]	EBMT	greyc	BT_{AE}, BT_{CE}, BT_{TE}
Institute for Infocomm Research (Singapore)	I^2R's Machine Translation System for IWSLT 2009 [20]	Hybrid SMT	i2r	BT_{CE}
Chinese Academy of Sciences, Institute of Computing Technology (China)	The ICT Statistical Machine Translation Systems for the IWSLT 2009 [21]	SMT	ict	BT_{CE}, CT_{CE}, CT_{EC}
University of Grenoble, LIG (France)	LIG approach for IWSLT09 : Using Multiple Morphological Segmenters for Spoken Language Translation of Arabic [22]	SMT	lig	BT_{AE}
University of Le Mans, LIUM (France)	LIUM's Statistical Machine Translation Systems for IWSLT 2009 [23]	SMT	lium	BT_{AE}, BT_{CE}
MIT Lincoln Laboratory (USA)	The MIT-LL/AFRL IWSLT-2009 System [24]	SMT	mit	BT_{AE}, BT_{TE}
National Institute of Information and Communications Technology (Japan)	Two methods for stabilizing MERT: NICT at IWSLT 2009 [25]	SMT	nict	CT_{CE}, CT_{EC}
Chinese Academy of Sciences, National Laboratory of Pattern Recognition (China)	The CASIA Statistical Machine Translation System for IWSLT 2009 [26]	Hybrid SMT	nlpr	BT_{CE}, CT_{CE}, CT_{EC}
National University of Singapore (Singapore)	The NUS Statistical Machine Translation System for IWSLT 2009 [27]	SMT	nus	BT_{CE} $CT_{CE}{}^{\ddagger}$, $CT_{EC}{}^{\ddagger}$
University of Tokyo (Japan)	The UOT System: Improve String-to-Tree Translation Using Head-Driven Phrasal Structure Grammar and Predicate-Argument Structures [28]	SMT	tokyo	BT_{CE}
University of Tottori (Japan)	Statistical Machine Translation adding Pattern-based Machine Translation in Chinese-English Translation [29]	Hybrid RBMT SMT	tottori	BT_{CE}, $CT_{CE}{}^{\dagger}$, $CT_{EC}{}^{\dagger}$
TÜBİTAK-UEKAE (Turkey)	The TÜBİTAK-UEKAE Statistical Machine Translation System for IWSLT 2009 [30]	SMT	tubitak	BT_{AE}, BT_{TE}
University Politècnica de València (Spain)	UPV Translation System for IWSLT 2009 [31]	Hybrid SMT	upv	BT_{CE}
University of Washington (USA)	The University of Washington Machine Translation System for IWSLT 2009 [32]	SMT	uw	BT_{AE}, BT_{CE}

† : runs were submitted after the official run submission period.
‡ : runs were submitted after the subjective evaluation period.

Appendix B. Language Resources

B.1. The IWSLT 2009 Spoken Language Corpus

data set	(data type)	lang	sent	avg.len	word token	word type	ref.trans	task
training	(text)	A	19,972	6.5	130,624	18,147	–	BT_{AE}
	(text)	C	10,061	8.9	89,109	3,734	–	CT_{CE}, CT_{EC}
	(text)		19,972	7.4	148,224	8,408	–	BT_{CE}, CT_{CE}, CT_{EC}
	(text)	E	10,061	11.8	118,648	3,271	–	CT_{CE}, CT_{EC}
	(text)		19,972	7.7	153,178	7,294	–	BT_{AE}, BT_{CE}, BT_{TE}, CT_{CE}, CT_{EC}
	(text)	T	19,972	5.6	112,364	17,612	–	BT_{TE}
devset1$_{CSTAR03}$	(text)	A	506	5.0	2,555	1,156	–	BT_{AE}
	(text)	C	506	5.5	2,808	877	–	BT_{CE}, CT_{CE}
	(text)	E	8,096	6.8	55,383	2,134	16	BT_{AE}, BT_{CE}, BT_{TE}, CT_{CE}
	(text)	T	506	4.5	2,274	1,103	–	BT_{TE}
devset2$_{IWSLT04}$	(text)	A	500	5.3	2,660	1,237	–	BT_{AE}
	(text)	C	500	5.8	2,906	917	–	BT_{CE}, CT_{CE}
	(text)	E	8,000	6.9	55,027	2,233	16	BT_{AE}, BT_{CE}, BT_{TE}, CT_{CE}
	(text)	T	500	4.5	2,262	1,168	–	BT_{TE}
devset3$_{IWSLT05}$	(text)	A	506	5.1	2,566	1,263	–	BT_{AE}
	(read-speech)	C	506	6.3	3,209	929	–	BT_{CE}, CT_{CE}
	(text)		3,542	7.1	25,037	1,665	7	CT_{EC}
	(text)	E	8,096	6.9	55,959	2,323	16	BT_{AE}, BT_{CE}, CT_{CE}
	(read-speech)		506	6.2	3,119	840	–	CT_{EC}
devset4$_{IWSLT06}$	(spontaneous)	C	489	10.7	5,226	1,142	–	CT_{CE}
	(text)	E	3,423	11.4	39,174	1,817	7	CT_{CE}
devset5$_{IWSLT06}$	(spontaneous)	C	500	11.1	5,566	1,338	–	CT_{CE}
	(text)	E	3,500	12.6	44,079	2,036	7	CT_{CE}
devset6$_{IWSLT07}$	(text)	A	489	4.9	2,383	1,164	–	BT_{AE}
	(text)	C	489	5.4	2,647	878	–	BT_{CE}, CT_{CE}
	(text)	E	2,934	6.4	18,776	1,362	7	BT_{AE}, BT_{CE}, CT_{CE}
devset7$_{IWSLT08}$	(text)	A	507	5.1	2,585	1,205	–	BT_{AE}
	(text)	C	246	5.3	1,305	248	–	BT_{CE}, CT_{CE}
	(text)	E	1,722	7.0	12,076	577	7	BT_{AE}, BT_{CE}, CT_{CE}
devset8$_{IWSLT08}$	(spontaneous)	C	246	5.3	1,305	248	–	CT_{CE}
	(text)	E	1,722	7.0	12,076	577	7	CT_{CE}
devset9$_{IWSLT08}$	(spontaneous)	C	504	5.0	2,513	385	–	CT_{CE}
	(text)	E	3,528	6.2	21,751	810	7	CT_{CE}
devset10$_{IWSLT08}$	(text)	C	1,757	6.2	10,971	555	7	CT_{EC}
	(spontaneous)	E	251	5.1	1,279	241	–	CT_{EC}
devset11$_{IWSLT08}$	(text)	C	3,486	6.8	23,722	710	7	CT_{EC}
	(spontaneous)	E	498	5.8	2,867	213	–	CT_{EC}
devset$_{IWSLT09}$	(spontaneous)	C	200	9.3	1,859	377	–	CT_{CE}
	(text)		840	11.2	9,379	621	4	CT_{EC}
	(text)	E	800	9.8	7,829	418	4	CT_{CE}
	(spontaneous)		210	11.8	2,474	403	–	CT_{EC}
testset$_{IWSLT09}$	(text)	A	469	4.8	2,333	1,095	–	BT_{AE}
	(spontaneous)	C	405	11.3	4,562	653	–	CT_{CE}
	(text)		1,572	10.5	16,558	872	4	CT_{EC}
	(text)		469	5.5	1,808	877	–	BT_{CE}
	(text)	E	3,283	7.1	23,149	1,526	7	BT_{AE}, BT_{CE}, BT_{TE}
	(text)		1,620	11.5	18,594	764	4	CT_{CE}
	(spontaneous)		393	11.0	4,329	570	–	CT_{EC}
	(text)	T	469	4.6	2,143	1,029	–	BT_{TE}

B.2. Out-Of-Vocabulary Rates

· concerning the IWSLT 2009 $devset$ and $testset$, the OOV rates for the BTEC (CHALLENGE) training corpus are listed.

data set	lang	OOV (%) CRR	1BEST	NBEST	task
devset1$_{CSTAR03}$	A	5.5	–	–	BT$_{AE}$
	C	5.0	–	–	BT$_{CE}$, CT$_{CE}$
	T	6.7	–	–	BT$_{TE}$
devset2$_{IWSLT04}$	A	5.7	–	–	BT$_{AE}$
	C	4.1	–	–	BT$_{CE}$, CT$_{CE}$
	T	7.7	–	–	BT$_{TE}$
devset3$_{IWSLT05}$	A	6.2	–	–	BT$_{AE}$
	C	3.3	3.0	4.0	BT$_{CE}$, CT$_{CE}$
	E	2.0	1.7	2.9	CT$_{EC}$
devset4$_{IWSLT06}$	C	3.5	4.2	4.5	CT$_{CE}$
devset5$_{IWSLT06}$	C	4.2	4.1	4.2	CT$_{CE}$
devset6$_{IWSLT07}$	A	4.9	–	–	BT$_{AE}$
	C	5.3	–	–	BT$_{CE}$, CT$_{CE}$
devset7$_{IWSLT08}$	A	5.1	–	–	BT$_{AE}$
	C	4.1	3.4	4.2	CT$_{CE}$
	E	3.0	2.1	3.1	CT$_{EC}$
devset8$_{IWSLT08}$	C	4.2	3.4	4.2	CT$_{CE}$
devset9$_{IWSLT08}$	C	2.6	2.5	3.8	CT$_{CE}$
devset10$_{IWSLT08}$	E	2.9	2.1	3.1	CT$_{EC}$
devset11$_{IWSLT08}$	E	3.0	2.6	3.4	CT$_{EC}$
devset$_{IWSLT09}$	C	3.3 (4.1)	3.5 (4.5)	3.7 (6.1)	CT$_{CE}$
	E	1.1 (0.6)	1.2 (1.9)	1.6(3.1)	CT$_{EC}$
testset$_{IWSLT09}$	A	9.3	–	–	BT$_{AE}$
	C	3.6 (7.1)	4.5(5.5)	4.8 (7.0)	CT$_{CE}$
		5.1	–	–	BT$_{CE}$
	E	1.5 (0.8)	1.9 (2.7)	2.2 (3.5)	CT$_{EC}$
	T	6.4	–	–	BT$_{TE}$

B.3. Language Model Perplexity

· standard 5gram language models trained on the CHALLENGE and BTEC training data sets were used to calculate the language perplexity of each target language for the CHALLENGE and BTEC tasks, respectively.

data set	lang	entropy	words	total entropy	task
devset$_{IWSLT09}$	C	6.07	2,342	14,214	CT$_{EC}$
	E	5.22	1,922	10,035	CT$_{CE}$
testset$_{IWSLT09}$	C	6.18	4,142	25,580	CT$_{EC}$
	E	5.43	4,501	24,446	CT$_{CE}$
		5.80	2,844	15,063	BT$_{AE}$, BT$_{CE}$, BT$_{TE}$

B.4. Speech Recognition Accuracy

data set (data type)	lang	word (%) lattice	1BEST	sentence (%) lattice	1BEST	task
devset$_{IWSLT09}$ (spontaneous)	C	94.41	81.46	74.63	39.12	CT$_{CE}$
	E	89.15	85.63	52.39	48.57	CT$_{EC}$
testset$_{IWSLT09}$ (spontaneous)	C	91.82	75.81	57.64	29.32	CT$_{CE}$
	E	89.58	82.20	50.13	37.15	CT$_{EC}$

Appendix C. Human Assessment

C.1. Fluency / Adequacy / Dialog

(best = 4.0, . . ., worst = 0.0)

· only the top-ranked (*NormRank*) primary run submissions (cf. Appendix C.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.

Task	MT	Fluency	Adequacy	Dialog
CT_{EC}	nlpr.ASR	2.35	2.45	2.53
	nlpr.CRR	2.60	2.81	2.90
CT_{CE}	nlpr.ASR	2.37	2.59	2.92
	nlpr.CRR	2.53	2.88	3.19
BT_{CE}	nlpr	2.78	2.99	
BT_{AE}	mit	2.70	2.76	
BT_{TE}	mit+tubitak	2.90	3.06	

C.2. Ranking

(**Ranking**: best = 1.0, . . ., worst = 0.0) (**NormRank**: best = 4.0, . . ., worst = 0.0)

· 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 [5].

CT_{EC}

MT	Ranking
nlpr.ASR	**0.4933**
nict.ASR	0.3348
dcu.ASR	0.2852
fbk.ASR	0.2785
ict.ASR	0.2399
tottori.ASR	0.1360

MT	NormRank
nlpr.ASR	**3.48**
nict.ASR	3.02
dcu.ASR	2.80
fbk.ASR	2.79
ict.ASR	2.63
tottori.ASR	2.18

CT_{EC}

MT	Ranking
nlpr.CRR	**0.5912**
ict.CRR	0.5429
nict.CRR	0.4496
fbk.CRR	0.4375
dcu.CRR	0.4235
tottori.CRR	0.2392

MT	NormRank
nlpr.CRR	**3.84**
ict.CRR	3.67
nict.CRR	3.42
fbk.CRR	3.32
dcu.CRR	3.31
tottori.CRR	2.58

CT_{CE}

MT	Ranking
nlpr.ASR	**0.5062**
ict.ASR	0.3095
dcu.ASR	0.3007
nict.ASR	0.2812
fbk.ASR	0.2676
tottori.ASR	0.2320

MT	NormRank
nlpr.ASR	**3.52**
ict.ASR	2.90
dcu.ASR	2.84
nict.ASR	2.80
fbk.ASR	2.75
tottori.ASR	2.60

CT_{CE}

MT	Ranking
nlpr.CRR	**0.5655**
ict.CRR	0.4527
dcu.CRR	0.4299
nict.CRR	0.4055
fbk.CRR	0.3833
tottori.CRR	0.3157

MT	NormRank
nlpr.CRR	**3.67**
ict.CRR	3.32
dcu.CRR	3.26
nict.CRR	3.20
fbk.CRR	3.11
tottori.CRR	2.83

BT_{AE}

MT	Ranking
mit	**0.3465**
mit+tubtak	0.3443
lium	0.3016
fbk	0.2939
lig	0.2784
bmrc	0.2765
tubitak	0.2660
uw	0.2625
greyc	0.1668

MT	NormRank
mit	**3.29**
mit+tubtak	3.28
fbk	3.03
bmrc	3.03
lium	3.01
uw	2.95
lig	2.87
tubitak	2.86
greyc	2.38

BT_{CE}

MT	Ranking
nlpr	**0.4985**
nus	0.3891
i2r	0.3781
ict	0.3737
uw	0.3219
tottori	0.3174
upv	0.3125
bmrc	0.3066
lium	0.2976
tokyo	0.2956
dcu	0.2900
greyc	0.2697

MT	NormRank
nlpr	**3.55**
nus	3.24
i2r	3.17
ict	3.12
uw	3.01
upv	2.99
bmrc	2.95
dcu	2.91
tokyo	2.87
tottori	2.84
lium	2.78
greyc	2.63

BT_{TE}

MT	Ranking
tubitak	**0.3594**
mit+tubtak	0.3335
fbk	0.3319
mit	0.3306
dcu	0.2655
apptek	0.2380
greyc	0.1568

MT	NormRank
mit+tubtak	**3.26**
tubitak	3.25
mit	3.23
fbk	3.13
dcu	2.92
apptek	2.74
greyc	2.39

C.3. Pairwise Comparison

(best = 1.0, . . ., worst = -1.0)

· the outputs of the first system are compared against a second system on a sentence-by-sentence basis according to the *Ranking* grades.
· the given scores are the ratio of improved translations, i.e. *gain* = $\frac{|better\ translations|-|worse\ translations|}{total\ translations}$.
· the order of the systems is defined by positive gains of the pairwise system comparison.

CT$_{EC}$

↓ 1^{st} . 2^{nd} →	**nict.ASR**	**dcu.ASR**	**fbk.ASR**	**ict.ASR**	**tottori.ASR**
nlpr.ASR	0.3071	0.3656	0.3566	0.4810	0.6092
	nict.ASR	0.0861	0.1557	0.1931	0.3804
		dcu.ASR	0.0132	0.1266	0.2556
			fbk.ASR	0.0867	0.3715
				ict.ASR	0.2670

↓ 1^{st} . 2^{nd} →	**ict.CRR**	**fbk.CRR**	**dcu.CRR**	**nict.CRR**	**tottori.CRR**
nlpr.CRR	0.1981	0.3824	0.3496	0.2197	0.5629
	ict.CRR	0.1763	0.1711	0.1607	0.6052
		fbk.CRR	0.0117	0.0275	0.3647
			dcu.CRR	0.0094	0.3494
				nict.CRR	0.3732

CT$_{CE}$

↓ 1^{st} . 2^{nd} →	**dcu.ASR**	**ict.ASR**	**fbk.ASR**	**nict.ASR**	**tottori.ASR**
nlpr.ASR	0.2633	0.3848	0.3854	0.3601	0.4752
	dcu.ASR	0.0103	0.0665	0.0689	0.1514
		ict.ASR	0.0611	0.0143	0.1898
			fbk.ASR	0.0490	0.1221
				nict.ASR	0.0811

↓ 1^{st} . 2^{nd} →	**ict.CRR**	**dcu.CRR**	**nict.CRR**	**fbk.CRR**	**tottori.CRR**
nlpr.CRR	0.2660	0.1939	0.2010	0.3267	0.5447
	ict.CRR	0.0190	0.0423	0.1778	0.2698
		dcu.CRR	0.0560	0.0789	0.2210
			nict.CRR	0.0208	0.2051
				fbk.CRR	0.1671

BT$_{CE}$

↓ 1^{st} . 2^{nd} →	**nus**	**ict**	**i2r**	**upv**	**uw**	**bmrc**	**tottori**	**dcu**	**lium**	**tokyo**	**greyc**
nlpr	0.1509	0.2554	0.1663	0.3729	0.3048	0.3480	0.4050	0.2847	0.3461	0.4015	0.5301
	nus	0.0516	0.0000	0.1843	0.1327	0.1811	0.2079	0.2160	0.2183	0.2236	0.3868
		ict	0.0183	0.1225	0.0684	0.0076	0.1926	0.0620	0.1763	0.1512	0.2437
			i2r	0.0704	0.1742	0.1905	0.1808	0.0698	0.1827	0.2418	0.3198
				upv	0.0182	**-0.0103**	0.1087	0.0955	0.0825	0.0364	0.1594
					uw	0.0696	0.1108	0.0897	0.1130	0.0823	0.2721
						bmrc	0.1303	0.0122	0.0683	0.0898	0.1868
							tottori	0.0155	0.1588	**-0.0072**	0.1094
								dcu	0.0584	0.0190	0.1401
									lium	0.0430	0.1731
										tokyo	0.1327

BT$_{AE}$

↓ 1^{st} . 2^{nd} →	**mit+tubitak**	**fbk**	**bmrc**	**lium**	**uw**	**tubitak**	**lig**	**greyc**
mit	0.0049	0.1526	0.1601	0.1124	0.2066	0.2048	0.2252	0.4405
	mit+tubitak	0.1054	0.1447	0.1804	0.1230	0.2350	0.2630	0.4602
		fbk	0.0520	0.0016	0.0393	0.0858	0.0909	0.3645
			bmrc	0.0518	0.0254	0.0949	0.0488	0.3498
				lium	0.0323	0.0698	0.1084	0.4041
					uw	0.0250	0.0469	0.2759
						tubitak	0.0224	0.2864
							lig	0.2818

BT$_{TE}$

↓ 1^{st} . 2^{nd} →	**tubitak**	**mit**	**fbk**	**dcu**	**apptek**	**greyc**
mit+tubitak	0.0589	0.0151	0.0886	0.1762	0.3294	0.4908
	tubitak	0.0223	0.0526	0.1787	0.3473	0.5000
		mit	0.0375	0.1913	0.3272	0.4669
			fbk	0.1240	0.2625	0.4288
				dcu	0.1153	0.3133
					apptek	0.2862

C.4. 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.

CT$_{EC}$

nlpr.ASR	BestRankDiff	Better	Same	Worse
nict.ASR	0.2985	0.5024	0.2937	0.2039
fbk.ASR	0.3672	0.5309	0.3054	0.1637
dcu.ASR	0.3696	0.5170	0.3356	0.1474
ict.ASR	0.4864	0.6126	0.2612	0.1262
tottori.ASR	0.6165	0.7042	0.2081	0.0877

nlpr.CRR	BestRankDiff	Better	Same	Worse
ict.CRR	0.2071	0.4432	0.3207	0.2361
nict.CRR	0.2146	0.4536	0.3074	0.2390
dcu.CRR	0.3534	0.5145	0.3244	0.1611
fbk.CRR	0.3837	0.5869	0.2099	0.2032
tottori.CRR	0.5524	0.6659	0.2206	0.1135

CT$_{CE}$

nlpr.ASR	BestRankDiff	Better	Same	Worse
nict.ASR	0.3601	0.5346	0.2909	0.1745
dcu.ASR	0.3621	0.5027	0.3567	0.1406
ict.ASR	0.3848	0.5112	0.3624	0.1264
fbk.ASR	0.3854	0.5156	0.3542	0.1302
tottori.ASR	0.4752	0.5940	0.2872	0.1188

nlpr.CRR	BestRankDiff	Better	Same	Worse
dcu.CRR	0.1939	0.4182	0.3575	0.2243
nict.CRR	0.2010	0.4334	0.3342	0.2324
ict.CRR	0.2595	0.4732	0.3131	0.2137
fbk.CRR	0.3267	0.5312	0.2643	0.2045
tottori.CRR	0.5447	0.6422	0.2603	0.0975

BT$_{CE}$

nlpr	BestRankDiff	Better	Same	Worse
nus	0.1509	0.3603	0.4303	0.2094
i2r	0.1662	0.3752	0.4158	0.2090
ict	0.2554	0.4385	0.3784	0.1831
dcu	0.2846	0.4501	0.3844	0.1655
uw	0.3047	0.4904	0.3239	0.1857
lium	0.3460	0.5322	0.2816	0.1862
bmrc	0.3480	0.4933	0.3614	0.1453
upv	0.3729	0.5059	0.3611	0.1330
tokyo	0.4014	0.5474	0.3066	0.1460
tottori	0.4050	0.5775	0.2500	0.1725
greyc	0.5300	0.6620	0.2060	0.1320

BT$_{AE}$

mit	BestRankDiff	Better	Same	Worse
mit+tubitak	0.0048	0.0326	0.9396	0.0278
lium	0.1123	0.3090	0.4943	0.1967
fbk	0.1526	0.3100	0.5326	0.1574
bmrc	0.1601	0.3267	0.5067	0.1666
tubitak	0.2047	0.3984	0.4079	0.1937
uw	0.2065	0.3508	0.5049	0.1443
lig	0.2252	0.4249	0.3754	0.1997
greyc	0.4405	0.5755	0.2895	0.1350

BT$_{TE}$

mit+tubitak	BestRankDiff	Better	Same	Worse
mit	0.0151	0.0571	0.9009	0.0420
tubitak	0.0589	0.2239	0.6111	0.1650
fbk	0.0886	0.2675	0.5536	0.1789
dcu	0.1761	0.3255	0.5251	0.1494
apptek	0.3293	0.4685	0.3923	0.1392
greyc	0.4908	0.5993	0.2922	0.1085

Appendix D. Automatic Evaluation

D.1. Significance Test

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

· the mean score and the 95% confidence intervals were calculated for each MT output according to the *bootStrap* method [14].
· *z-avg* is the average system score of all z-transformed automatic evaluation metric scores obtained by a single MT system.
· the MT systems are ordered according to the *z-avg* and the best (worst) score of each metric is marked with *boldface* (*italic*).
· omitted lines between scores indicate non-significant differences in performance between the MT engines according to the *bootStrap* method [14].

CHALLENGE English-Chinese (CT_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
35.70	**64.82**	**54.55**	**40.91**	**48.60**	**70.09**	**6.268**	**1.364**	nlpr	**1.572**	**37.71**	**64.59**	**55.32**	**41.48**	**48.83**	**70.25**	**6.410**
35.87	62.84	60.60	43.11	51.85	69.16	5.909	1.017	nict	0.828	35.48	61.00	63.10	45.31	53.49	66.79	5.783
33.38	61.28	61.20	44.83	54.38	69.57	5.999	0.881	fbk	0.749	33.34	59.63	63.21	46.46	55.70	67.07	6.019
32.85	60.11	59.39	45.85	52.86	66.99	5.738	0.659	dcu	0.674	33.17	59.43	61.03	46.90	54.04	65.91	5.785
29.03	58.06	63.52	47.71	56.15	64.74	5.553	0.194	ict	0.308	29.87	57.63	65.29	48.41	57.34	64.71	5.688
22.16	***45.16***	***85.20***	***64.50***	***80.84***	***63.99***	***4.441***	***-1.544***	tottori	***-1.559***	***22.60***	***46.34***	***84.42***	***63.56***	***79.35***	***60.95***	***4.590***

"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
40.80	**68.99**	**49.14**	**36.13**	**43.52**	74.80	**6.885**	**1.249**	nlpr	**1.411**	**43.09**	**69.09**	**49.79**	**36.70**	**43.57**	**75.06**	**7.055**
40.08	67.28	54.87	38.32	47.77	**76.01**	6.876	1.026	fbk	0.882	40.11	66.17	56.12	39.41	48.59	73.81	6.943
38.90	66.91	51.44	39.04	45.11	72.48	6.506	0.748	ict	0.889	40.03	67.04	52.42	39.32	45.56	72.64	6.639
38.49	67.08	55.03	39.41	46.50	72.99	6.257	0.613	nict	0.408	38.22	65.51	57.30	41.35	48.11	70.99	6.163
37.37	65.55	56.48	40.20	49.28	73.32	6.674	0.570	dcu	0.590	37.69	65.17	57.99	41.07	50.32	72.82	6.789
27.61	***54.97***	***74.23***	***53.85***	***68.71***	***69.12***	***5.295***	***-1.634***	tottori	***-1.607***	***27.56***	***54.71***	***74.96***	***54.18***	***68.87***	***66.79***	***5.408***

CHALLENGE Chinese-English (CT_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
35.49	**65.58**	**53.44**	**41.87**	**47.53**	**70.99**	**6.484**	**2.063**	nlpr	**2.132**	**37.12**	**64.07**	**54.64**	**42.39**	**49.87**	**69.80**	**6.801**
31.56	57.79	56.53	47.28	50.92	64.65	5.480	0.704	dcu	0.580	30.60	54.92	58.96	48.85	54.24	62.76	5.475
30.08	59.04	62.20	48.98	57.33	66.51	5.637	0.530	fbk	0.431	28.62	56.07	64.69	50.14	60.88	64.08	5.832
28.56	59.21	62.30	48.60	56.03	63.88	5.676	0.405	ict	0.537	28.50	57.61	64.37	49.22	59.74	63.65	5.890
26.65	58.31	***72.92***	***55.45***	***66.91***	66.54	5.181	-0.378	nict	-0.333	25.78	55.71	***74.88***	56.05	***70.21***	64.09	5.389
24.81	***54.88***	69.41	54.56	64.83	***61.33***	***4.970***	***-0.751***	tottori	***-0.775***	***23.23***	***52.69***	72.37	***57.06***	68.90	***59.46***	***5.035***

"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
36.42	**67.99**	**50.86**	**39.19**	**45.10**	**73.53**	**6.787**	**1.851**	nlpr	**1.873**	**38.06**	**66.93**	**52.27**	**39.17**	**47.16**	**72.49**	**7.137**
36.88	64.13	53.03	42.37	47.16	70.73	6.458	1.214	dcu	1.133	36.71	62.32	54.48	42.93	49.82	70.15	6.644
31.84	63.21	60.16	45.75	53.61	69.75	6.026	0.398	fbk	0.293	31.15	60.92	62.10	46.28	56.68	68.15	6.288
30.74	63.11	60.42	45.07	52.81	68.25	6.137	0.326	ict	0.623	31.76	62.60	61.61	44.12	55.89	69.60	6.500
29.70	63.11	***71.07***	***51.99***	***63.90***	70.08	5.605	-0.424	nict	-0.543	28.73	60.87	***73.75***	***52.58***	***67.56***	68.04	5.771
27.95	***59.69***	65.91	51.00	61.39	***65.90***	***5.412***	***-0.793***	tottori	***-0.809***	***27.16***	***57.90***	68.07	52.17	65.27	***65.06***	***5.563***

BTEC Chinese-English (BTEC_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
49.70	**72.67**	**41.02**	**35.54**	**33.65**	**72.53**	**7.363**	**2.178**	nlpr	**2.221**	**48.97**	**69.18**	**45.38**	**38.01**	**37.35**	**71.24**	**7.643**
44.77	68.09	44.03	38.96	35.85	69.66	6.494	1.344	nus	1.284	43.90	63.82	49.69	42.76	40.49	67.61	6.603
45.94	67.24	43.83	39.39	35.71	69.55	6.110	1.250	i2r	1.180	45.31	63.56	49.19	43.53	39.72	66.89	5.994
40.58	66.20	50.05	42.42	42.01	69.46	6.774	0.786	uw	0.857	39.68	62.18	54.89	45.41	45.69	67.45	6.967
42.38	64.48	45.66	41.73	36.25	66.83	***4.858***	0.545	dcu	0.518	41.98	59.79	50.99	45.28	40.47	64.79	***4.499***
39.53	64.18	48.45	42.81	39.38	66.87	5.853	0.489	bmrc	0.498	39.42	59.61	53.40	46.24	43.56	64.48	5.724
40.15	60.78	49.18	43.75	41.46	67.68	5.890	0.323	lium	0.109	38.22	55.76	55.35	49.21	45.98	63.78	5.621
35.33	62.70	51.93	44.82	41.81	65.96	5.821	0.068	upv	0.134	35.18	58.05	57.04	48.77	46.91	64.28	5.936
35.41	62.71	49.96	44.65	40.58	63.47	5.647	0.022	tokyo	0.139	35.45	58.05	55.12	47.81	45.71	61.91	5.843
35.65	62.27	50.78	45.06	41.57	64.60	5.610	-0.011	ict	0.013	34.77	58.21	57.11	49.14	46.57	62.57	5.682
31.49	61.71	55.88	47.60	48.06	64.79	6.154	-0.405	tottori	-0.465	29.33	56.82	62.50	52.09	54.00	61.95	6.360
27.92	***55.35***	***59.22***	***53.25***	***51.61***	***59.62***	5.457	***-1.444***	greyc	***-1.344***	***27.67***	***50.94***	***65.50***	***57.91***	***57.22***	***56.49***	5.686

BTEC Arabic-English (BTEC_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
57.53	**78.43**	**30.34**	**27.40**	**25.01**	**77.65**	7.715	**1.504**	mit+tubitak	**1.417**	**55.82**	**74.74**	**35.88**	**31.65**	**28.75**	**74.82**	7.628
57.17	78.23	30.64	27.91	25.19	77.20	7.651	1.432	mit	1.378	55.61	74.62	36.04	31.86	28.99	74.59	7.584
52.20	75.72	35.25	31.15	29.41	76.38	**7.862**	0.940	fbk	0.869	49.90	71.66	40.97	35.73	33.79	73.38	**7.975**
49.35	73.28	36.32	33.06	30.39	74.10	7.333	0.504	tubitak	0.431	47.13	68.69	42.33	37.52	34.84	71.05	7.328
50.88	73.15	36.67	32.94	30.31	74.60	6.888	0.465	lium	0.328	48.48	68.41	42.59	37.90	34.81	70.93	6.708
49.51	73.99	35.57	31.94	28.91	73.48	6.610	0.456	bmrc	0.412	48.34	69.52	40.94	36.13	32.98	70.58	6.350
46.60	73.68	39.90	34.58	33.14	74.87	7.534	0.325	lig	0.694	47.04	71.43	43.67	36.13	36.11	74.18	7.902
48.16	72.84	37.99	34.47	30.83	72.82	6.564	0.173	uw	0.298	48.01	68.95	42.69	38.11	34.80	70.64	6.609
32.91	***61.65***	***51.19***	***45.25***	***43.25***	***66.08***	***5.453***	***-1.941***	greyc	***-1.970***	***30.71***	***56.61***	***58.67***	***51.03***	***48.83***	***62.33***	***5.319***

BTEC Turkish-English (BTEC_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
60.75	**81.94**	**29.49**	**24.34**	**22.73**	**78.65**	8.127	**1.304**	mit+tubitak	**1.301**	**59.49**	**79.04**	**33.95**	**27.28**	**25.84**	76.83	8.276
60.11	81.30	30.10	24.89	23.10	78.43	7.992	1.216	mit	1.238	58.67	78.35	34.25	**27.28**	26.22	**76.86**	8.133
55.84	81.20	32.65	26.76	25.20	77.93	**8.215**	1.043	tubitak	1.059	53.86	77.64	37.18	29.31	29.01	76.49	**8.574**
56.82	79.08	32.82	27.04	25.81	77.26	7.948	0.912	fbk	0.819	54.76	75.71	38.88	30.82	29.55	74.73	8.062
56.07	76.03	33.57	29.71	26.48	75.70	6.948	0.502	dcu	0.335	54.17	71.54	39.01	34.07	30.43	72.84	6.657
43.28	74.73	***52.49***	35.62	34.70	70.51	6.824	-0.536	apptek	-0.226	43.74	70.95	55.18	36.15	38.93	72.22	7.328
35.48	***64.71***	50.96	***43.74***	***41.64***	***67.74***	***6.106***	***-1.441***	greyc	***-1.526***	***34.58***	***59.96***	***57.02***	***48.46***	***47.06***	***64.34***	***6.150***

D.2. Full Testset

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

· the systems were ranked according to the average system scores reported in Appendix D.1.
· the best (worst) score of each metric is marked with *boldface* (*italic*).

CHALLENGE English-Chinese (CT_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
35.66	**64.79**	**54.57**	**40.94**	**48.62**	**70.07**	**6.396**	**1.412**	nlpr	**1.627**	**37.66**	**64.56**	**55.34**	**41.51**	**48.84**	**70.22**	**6.553**
35.83	62.82	60.65	43.14	51.87	69.14	6.026	1.033	nict	0.837	35.44	60.98	63.15	45.34	53.52	66.76	5.902
33.37	61.27	61.21	44.84	54.39	69.57	6.117	0.920	fbk	0.788	33.33	59.63	63.21	46.46	55.71	67.07	6.145
32.82	60.10	59.41	45.86	52.87	66.97	5.853	0.677	dcu	0.694	33.14	59.42	61.05	46.90	54.05	65.89	5.907
29.01	58.05	63.54	47.72	56.14	64.72	5.657	0.179	ict	0.304	29.85	57.62	65.30	48.41	57.33	64.69	5.801
22.14	***45.16***	***85.18***	***64.47***	***80.82***	***63.99***	***4.509***	***-1.509***	tottori	***-1.592***	***22.56***	***46.35***	***84.40***	***63.53***	***79.33***	***60.95***	***4.663***

bleu	f1	wer	per	ter	gtm	nist	**z-avg**	**CRR**	**z-avg**	bleu	f1	wer	per	ter	gtm	nist
40.75	**68.97**	**49.17**	**36.14**	**43.54**	74.80	**7.035**	**1.305**	nlpr	**1.467**	**43.04**	**69.07**	**49.82**	**36.71**	**43.59**	**75.07**	**7.223**
40.05	67.27	54.88	38.32	47.78	**76.02**	7.028	1.101	fbk	0.947	40.07	66.16	56.13	39.40	48.60	73.81	7.104
38.86	66.90	51.46	39.04	45.12	72.47	6.660	0.750	ict	0.906	39.98	67.03	52.44	39.32	45.57	72.63	6.803
37.34	65.54	56.52	40.21	49.30	73.33	6.821	0.612	dcu	0.638	37.66	65.16	58.02	41.07	50.34	72.83	6.942
38.42	67.05	55.09	39.44	46.55	72.96	6.386	0.593	nict	0.380	38.15	65.49	57.36	41.38	48.15	70.96	6.295
27.59	***55.00***	***74.21***	***53.82***	***68.70***	***69.14***	***5.389***	***-1.592***	tottori	***-1.563***	***27.54***	***54.73***	***74.94***	***54.15***	***68.85***	***66.81***	***5.509***

CHALLENGE Chinese-English (CT_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
35.52	**65.59**	**53.44**	**41.86**	**47.51**	**71.01**	**6.659**	**2.116**	nlpr	**2.186**	**37.17**	**64.08**	**54.62**	**42.37**	**49.84**	**69.81**	**6.997**
31.61	57.83	56.51	47.25	50.91	64.69	5.614	0.749	dcu	0.621	30.64	54.95	58.94	48.82	54.21	62.80	5.616
30.13	59.07	62.18	48.95	57.31	66.54	5.763	0.573	fbk	0.472	28.66	56.10	64.68	50.10	60.86	64.12	5.972
28.59	59.21	62.30	48.59	56.03	63.89	5.806	0.444	ict	0.574	28.53	57.62	64.38	49.21	59.74	63.66	6.033
26.67	58.34	***72.90***	***55.45***	***66.89***	66.56	5.285	-0.344	nict	-0.300	25.80	55.74	***74.87***	56.05	***70.20***	64.11	5.506
24.82	***54.89***	69.43	54.56	64.84	***61.36***	***5.071***	***-0.721***	tottori	***-0.750***	***23.23***	***52.70***	72.39	***57.07***	68.90	***59.47***	***5.139***

bleu	meteor	wer	per	ter	gtm	nist	**z-avg**	**CRR**	**z-avg**	bleu	meteor	wer	per	ter	gtm	nist
36.44	**67.99**	**50.86**	**39.19**	**45.10**	**73.54**	**6.977**	**1.908**	nlpr	**1.930**	**38.08**	**66.92**	**52.27**	**39.17**	**47.16**	**72.50**	**7.350**
36.91	64.15	53.02	42.35	47.15	70.74	6.629	1.269	dcu	1.185	36.75	62.34	54.47	42.92	49.80	70.16	6.825
31.92	63.23	60.15	45.75	53.58	69.76	6.167	0.445	fbk	0.340	31.27	60.93	62.10	46.27	56.66	68.16	6.442
30.78	63.10	60.42	45.06	52.81	68.25	6.283	0.370	ict	0.668	31.85	62.59	61.62	44.13	55.89	69.59	6.664
29.70	63.09	***71.09***	***52.01***	***63.92***	70.08	5.726	-0.391	nict	-0.513	28.72	60.85	***73.78***	***52.61***	***67.59***	68.03	5.900
27.97	***59.71***	65.90	50.99	61.39	***65.92***	***5.531***	***-0.754***	tottori	***-0.775***	***27.16***	***57.91***	68.07	52.18	65.27	***65.06***	***5.689***

BTEC Chinese-English (BTEC_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
49.69	**72.66**	**41.04**	**35.55**	**33.67**	**72.52**	**7.696**	**2.239**	nlpr	**2.323**	**48.97**	**69.17**	**45.40**	**38.03**	**37.38**	**71.23**	**8.029**
44.81	68.08	44.04	38.97	35.86	69.66	6.780	1.462	nus	1.416	43.95	63.82	49.68	42.76	40.49	67.60	6.927
45.95	67.25	43.83	39.38	35.70	69.56	6.384	1.370	i2r	1.304	45.31	63.56	49.19	43.51	39.70	66.89	6.294
40.61	66.21	50.04	42.39	41.99	69.47	7.048	0.944	uw	1.018	39.72	62.20	54.87	45.38	45.66	67.45	7.285
42.37	64.47	45.68	41.75	36.26	66.83	***5.063***	0.704	dcu	0.642	41.97	59.78	51.02	45.30	40.49	64.77	***4.711***
39.55	64.19	48.46	42.80	39.37	66.86	6.096	0.667	bmrc	0.653	39.45	59.63	53.40	46.21	43.54	64.47	5.994
40.14	60.76	49.21	43.78	41.48	67.68	6.119	0.497	lium	0.272	38.22	55.74	55.37	49.23	45.99	63.78	5.867
35.38	62.69	49.97	44.66	40.59	63.44	5.862	0.237	tokyo	0.307	35.44	58.03	55.13	47.82	45.72	61.88	6.095
35.29	62.66	51.99	44.86	41.86	65.93	6.047	0.268	upv	0.301	35.15	58.01	57.10	48.82	46.96	64.25	6.197
35.63	62.26	50.80	45.07	41.58	64.59	5.841	0.204	ict	0.191	34.77	58.20	57.12	49.14	46.57	62.56	5.942
31.51	61.69	55.90	47.60	48.07	64.78	6.383	-0.160	tottori	-0.253	29.35	56.80	62.52	52.09	54.01	61.93	6.626
27.95	***55.37***	***59.23***	***53.24***	***51.61***	***59.64***	5.657	***-1.117***	greyc	***-1.095***	***27.73***	***50.98***	***65.50***	***57.88***	***57.21***	***56.53***	5.927

BTEC Arabic-English (BTEC_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
57.54	**78.45**	**30.34**	**27.40**	**25.02**	**77.65**	8.083	**1.612**	mit+tubitak	**1.501**	**55.82**	**74.75**	**35.89**	**31.66**	**28.76**	**74.83**	8.031
57.17	78.24	30.65	27.92	25.19	77.21	8.015	1.536	mit	1.459	55.61	74.62	36.06	31.87	29.00	74.59	7.984
52.23	75.73	35.23	31.14	29.39	76.40	**8.218**	1.021	fbk	0.926	49.94	71.68	40.95	35.71	33.78	73.42	**8.375**
49.33	73.27	36.34	33.08	30.41	74.10	7.651	0.555	tubitak	0.460	47.12	68.66	42.36	37.54	34.87	71.05	7.683
50.86	73.15	36.69	32.95	30.33	74.60	7.198	0.516	lium	0.358	48.46	68.40	42.61	37.90	34.84	70.94	7.052
49.51	73.99	35.57	31.94	28.91	73.49	6.904	0.507	bmrc	0.447	48.34	69.51	40.95	36.14	32.99	70.60	6.669
46.62	73.69	39.89	34.58	33.13	74.88	7.856	0.371	lig	0.738	47.08	71.44	43.65	36.13	36.11	74.20	8.279
48.12	72.83	38.01	34.49	30.84	72.80	6.845	0.205	uw	0.324	47.99	68.93	42.70	38.13	34.81	70.64	6.933
32.92	***61.66***	***51.21***	***45.25***	***43.26***	***66.07***	***5.654***	***-2.015***	greyc	***-2.054***	***30.72***	***56.62***	***58.69***	***51.03***	***48.84***	***62.34***	***5.536***

BTEC Turkish-English (BTEC_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
60.71	**81.92**	**29.53**	**24.37**	**22.78**	**78.64**	8.525	**1.468**	mit+tubitak	**1.466**	**59.44**	**79.03**	**34.01**	27.32	**25.88**	76.82	8.726
60.09	81.28	30.12	24.91	23.13	78.44	8.381	1.375	mit	1.399	58.63	78.34	34.29	**27.30**	26.25	**76.87**	8.571
55.82	81.20	32.67	26.76	25.22	77.92	**8.602**	1.194	tubitak	1.214	53.85	77.63	37.21	29.32	29.03	76.49	**9.023**
56.77	79.06	32.87	27.08	25.86	77.25	8.328	1.049	fbk	0.950	54.69	75.68	38.96	30.88	29.60	74.72	8.487
56.06	76.02	33.60	29.72	26.49	75.71	7.274	0.607	dcu	0.427	54.14	71.52	39.05	34.09	30.45	72.84	7.006
43.27	74.73	***52.50***	35.62	34.71	70.52	7.121	-0.497	apptek	-0.157	43.73	70.95	55.19	36.15	38.95	72.24	7.685
35.50	***64.75***	50.94	***43.72***	***41.63***	***67.78***	***6.347***	***-1.463***	greyc	***-1.544***	***34.60***	***60.00***	***57.00***	***48.44***	***47.05***	***64.39***	***6.425***

Appendix E. Evaluation Metric Correlation

· the correlation between evaluation metrics are measured using the *Spearman's rank correlation coefficient* $\rho \in$ [-1.0, 1.0]
with ρ = 1.0 if all systems ranked in same order, ρ = -1.0 if all systems ranked in reverse order and ρ = 0.0 if no correlation exists
· the number in parentheses behind each translation task label indicates the number of ranked MT systems
· *z-avg* is the average system score of all z-transformed automatic evaluation metric scores obtained by a single MT system.
· *r-avg* is the average system rank that a MT system achieved base on the system rankings of each automatic evaluation metrics.
· the automatic evaluation metrics that correlate best with the respective human assessments are marked in boldface

CT_{EC} (12)	**z-avg**	**r-avg**	bleu	f1	wer	per	ter	gtm	nist
Ranking	0.7413	0.2168	**0.8112**	0.7552	0.2448	0.2168	0.1329	-0.2308	0.3916
NormRank	0.7413	0.2168	**0.8112**	0.7552	0.2448	0.2168	0.1329	-0.2308	0.3916
BestRankDiff	0.0909	0.6364	0.0629	0.1189	0.7832	0.6364	**0.8951**	0.2308	0.0699

CT_{CE} (12)	**z-avg**	**r-avg**	bleu	meteor	wer	per	ter	gtm	nist
Ranking	0.2867	0.0909	0.0280	**0.8322**	0.3427	-0.2308	0.2517	0.0839	0.3706
NormRank	**0.6154**	0.1818	0.0420	0.3357	0.0140	0.0699	-0.0350	0.5804	0.0769
BestRankDiff	0.5804	0.3427	-0.0560	0.1538	-0.0769	0.5105	0.0559	**0.6294**	0.3147

$CT_{EC}{}^{ASR}$ (6)	**z-avg**	**r-avg**	bleu	f1	wer	per	ter	gtm	nist
Ranking	**0.9429**	0.8857	0.8857	**0.9429**	0.6571	0.8857	0.8285	0.6000	0.6000
NormRank	**0.9429**	0.8857	0.8857	**0.9429**	0.6571	0.8857	0.8285	0.6000	0.6000
BestRankDiff	0.8857	0.9429	0.7143	0.8857	0.4857	0.9429	**1.0000**	0.7714	0.7714

$CT_{EC}{}^{CRR}$ (6)	**z-avg**	**r-avg**	bleu	f1	wer	per	ter	gtm	nist
Ranking	0.8286	0.2168	0.8286	**0.9429**	0.7143	0.7714	0.4857	0.1429	0.4286
NormRank	0.8286	0.2168	0.8286	**0.9429**	0.7143	0.7714	0.4857	0.1429	0.4286
BestRankDiff	0.7143	0.6364	0.7143	0.5429	0.6000	0.6571	**0.9429**	0.2571	0.8857

$CT_{CE}{}^{ASR}$ (6)	**z-avg**	**r-avg**	bleu	meteor	wer	per	ter	gtm	nist
Ranking	0.7143	0.8286	0.7143	**0.9429**	0.4286	0.6000	0.6000	0.6571	0.6000
NormRank	0.7143	0.8286	0.7143	**0.9429**	0.4286	0.6000	0.6000	0.6571	0.6000
BestRankDiff	0.6000	0.8286	0.6000	**0.8857**	0.3143	0.3714	0.3714	0.4286	0.3147

$CT_{CE}{}^{CRR}$ (6)	**z-avg**	**r-avg**	bleu	meteor	wer	per	ter	gtm	nist
Ranking	**0.7143**	0.6571	0.2571	**0.7143**	0.4286	0.6000	0.6000	0.4857	0.5429
NormRank	**0.7143**	0.6571	0.2571	**0.7143**	0.4286	0.6000	0.6000	0.4857	0.5429
BestRankDiff	0.6000	**0.7714**	-0.3143	0.6000	0.3143	0.3714	0.3714	0.6571	0.5429

BT_{CE} (12)	**z-avg**	**r-avg**	bleu	meteor	wer	per	ter	gtm	nist
Ranking	**-0.5524**	-0.2867	0.1259	-0.2308	0.5385	-0.3846	0.2098	-0.4965	-0.2727
NormRank	-0.3846	0.0629	**0.6783**	-0.4336	0.3986	-0.0350	-0.1608	-0.0559	0.0839
BestRankDiff	0.2098	-0.2238	-0.6434	0.2378	0.1329	-0.0839	**0.7342**	-0.1189	-0.0489

BT_{AE} (9)	**z-avg**	**r-avg**	bleu	meteor	wer	per	ter	gtm	nist
Ranking	**0.4667**	0.0667	0.1333	0.4167	0.0000	-0.1000	-0.2167	0.0500	0.1500
NormRank	0.0333	-0.2833	0.3667	**0.9500**	-0.5000	-0.7000	-0.6167	0.8167	0.3667
BestRankDiff	0.1667	0.3500	-0.2333	-0.4667	0.4167	**0.7667**	0.7167	-0.3833	-0.0167

BT_{TE} (7)	**z-avg**	**r-avg**	bleu	meteor	wer	per	ter	gtm	nist
Ranking	0.6071	-0.5000	0.3571	0.6071	-0.6071	-0.5000	-0.5000	0.6071	**0.8571**
NormRank	0.8571	-0.6786	0.3928	0.8571	-0.5714	-0.6786	-0.6786	0.8571	**0.9643**
BestRankDiff	-0.6071	**1.0000**	0.2500	-0.6071	0.6786	**1.0000**	**1.0000**	-0.6071	-0.5714