Overview of the IWSLT 2008 Evaluation Campaign

Michael Paul

†National Institute of Information and Communications Technology
‡Advanced Telecommunications Research Laboratories
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) 2008[1]. In this workshop, we focused on the translation of spontaneous speech recorded in a real situation and the feasability of pivot-language-based translation approaches. The translation directions were English into Chinese and vice versa for the *Challenge Task*, Chinese into English and English into Spanish for the *Pivot Task*, and Arabic, Chinese, Spanish into English for the standard *BTEC Task*. In total, 19 research groups building 58 MT engines participated in this year's event. Automatic and subjective evaluations were carried out in order to investigate the impact of spontaneity aspects of field data experiments on 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 real environments.

1. Introduction

The *International Workshop on Spoken Language Translation* (IWSLT) is a yearly, open evaluation campaign for spoken language translation organized by the *Consortium for Speech Translation Advanced Research* (C-STAR)[2]. IWSLT's evaluations are not competition-oriented, but 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, the translation of automatic speech recognition results from read-speech and spontaneous-speech input, and dialog conversations [1, 2, 3, 4].

The focus of this year's evaluation campaign was the translation of spontaneous speech recorded in a real situation. Foreign travelers were provided with a state-of-the-art speech-to-speech translation hand-held device and were asked to carry-out specific tourism-related tasks (e.g., buying entrance tickets) using the device to communicate with local staff. Speech data was collected for 50 English and 50 Chinese travelers at 5 different locations, each carrying out 3-4 tasks. For the *Challenge Task*, IWSLT participants translated the Chinese/English output of the automatic speech recognizers (lattice, N/1BEST) into English/Chinese, respectively.

Another innovative aspect of this year's edition was the investigation of the feasability of pivot-language-based translation approaches. In the *Pivot Task*, participants were provided with read-speech recordings (lattice, N/1BEST) of Chinese utterances from the travel domain and had to apply Chinese-English and English-Spanish systems to produce the Spanish output. Like in previous IWSLT events, a standard *BTEC Task*, i.e. the translation of read-speech recordings (lattice, N/1BEST) and correct recognition results (text) of frequently used utterances in the travel domain, was also carried out for the translation of Chinese and Arabic into English as well as Chinese (directly) into Spanish.

Continuing the efforts started in 2007 to provide a list of linguistic resources and tools that can be shared by the participants[3], we asked that each participant send us information about non-proprietary resources used in the development of this year's submission so that other groups could also utilize these resources for the various tasks. It should be noted though, that participants did not had to provide resources directly. Nor were participants required to provide resources that they have acquired elsewhere and had then modified in some way (i.e. cleaned, corrected, or enhanced). In this latter example, a group provided a reference or link to the original provider or creator. Moreover, acceptable resources should be affordable by most research groups (publicly available monolingual or bilingual corpora, LDC data, etc.). In contrast, participants were not allowed to train or tune their systems on privately developed linguistic resources and/or corpora, NIST or LDC data which require participation in an evaluation campaign like GALE or NIST-MT, or publicly available linguistic resources which require high licensing fees. In addition, the use of resources supplied for other data tracks and previous IWSLT evaluation campaigns was also not allowed.

In total, 19 research groups participated in this year's evaluation campaign. A total of 58 MT engines were built

[1] http://www.slc.atr.jp/IWSLT2008
[2] http://www.c-star.org/
[3] http://www.slc.atr.jp/IWSLT2008/archives/2008/10/resources.html

to cover six different data tracks. The translation quality of all primary run submissions was evaluated using automatic evaluation metrics (*BLEU* [5], *METEOR* [6]) and a subjective evaluation metric that ranks each whole sentence translation from best to worst relative to the other choices [7]. In addition, human assessments of *fluency* and *adequacy* [8] were carried out for four selected MT system outputs for each of the data tracks. Based on the evaluation results, the impact of the spontaneity aspects of speech in real situations on the ASR and MT system performance as well as the robustness of state-of-the-art MT systems against speech recognition errors were investigated.

2. IWSLT 2008 Evaluation Campaign

This year's IWSLT evaluation campaign took place in the period of April-July 2008 and featured six different data track conditions:

Table 1: Data Tracks

Task	Translation Direction		Participants
Challenge	English-Chinese	CT_{EC}	7
	Chinese-English	CT_{CE}	11
BTEC	Arabic-English	BT_{AE}	10
	Chinese-English	BT_{CE}	14
	Chinese-Spanish	BT_{CS}	8
Pivot	Chinese-(English)-Spanish	PV_{CS}	8

In total, 19 research groups from all over the world[4] participated in the event, producing a total of 58 machine translaton engines for the above six data tracks. Information on the organisations, the utilized translation systems, and data track 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 SMT engines with EBMT systems or rule-based (RBMT) systems were also exploited.

For training purposes, a spoken language corpus described in Section 2.1 was provided to all participating research groups. In addition, the participants were free to use additional resources that could be shared[5] among the participants. The supplied resources of IWSLT 2008 were released one month ahead of the official run submissions period. The official run submission period was limited to five days. Run submission was carried out via email to the organizers whereby multiple runs were permitted. However, the participant had to specify which run should be treated as *primary* (evaluation using human assessments and automatic metrics) or *contrastive* (automatic evaluation only). In total, 58 primary runs and 101 contrastive runs were submitted. After the official run submission period, the organizers set-up an online evaluation server[6] that could be used by the particpants to carry out additional experiments on the evaluation testset.

The schedule of the evaluation campaign is summarized in Table 2.

Table 2: Evaluation Campaign Schedule

Event	Date
Training Corpus Release	June 2, 2008
Development Corpus Release	June 2, 2008
Evaluation Corpus Release	June 30, 2008
Result Submission Due	July 4, 2008

2.1. IWSLT 2008 Spoken Language Corpus

The IWSLT 2008 evaluation campaign was carried out using a multilingual spoken language corpus. The *Basic Travel Expression Corpus* (BTEC*) contains tourism-related sentences similar to those that are usually found in phrase books for tourists traveling abroad [9]. Parts of this corpus were already used in previous IWSLT evaluation campaigns [1, 2, 3, 4]. In addition to a sentence-aligned training corpus, the evaluation data sets of previous workshops including multiple reference translations were provided to the participants as a development corpus.

The evaluation data sets of IWSLT 2008 consisted of two different types of data. For the *Challenge Task*, machine-mediated conversational speech was recorded in a real situation. For the *Pivot Task* and the *BTEC Task*, read-speech recordings of randomly selected sentences from parts of the BTEC* corpus reserved for evaluation purposes were used.

ASR engines provided by the C-STAR partners 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.

2.1.1. Supplied Resources

For this year's evaluation campaign, parts of the Arabic (A), Chinese (C), English (E), and Spanish (S) 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}, BT_{CE}, and the English-Spanish part (PV_{ES}) of the PV_{CS} training corpus. In order to avoid a trilingual scenario for the *Pivot Task*, a separate set[7] of 20K sentence pairs were selected for the Chinese-English part (PV_{CE}) of PV_{CS}.

In order to optimize and evaluate their translation engines on in-domain data, the testsets of previous IWSLT evaluation

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

[5] Please refer to the MT system descriptions of each participant for details on what kind of additional resources were used.

[6] https://www.slc.atr.jp/EVAL/IWSLT08/automatic/testset_IWSLT08

[7] The union of both 20K sentence ID sets is identical to the 40K sentence pairs released for the Chinese-English data track of IWSLT 2006 and 2007.

Table 3: The IWSLT 2008 Spoken Language Corpus

data set	(data type)	lang	sent	avg.len	word token	word type	ref.trans	data track
training	(text)	A	19,972	6.5	130,624	18,147	–	BT_{AE}
	(text)	C	19,972	7.4	148,224	8,408	–	CT_{EC}, CT_{CE}, BT_{CE}, BT_{CS}
	(text)		20,000	6.8	135,518	9,185	–	PV_{CE}
	(text)	E	19,972	7.7	153,178	7,294	–	CT_{EC}, CT_{CE}, BT_{CE}, BT_{AE}, PV_{ES}
	(text)		20,000	9.1	182,793	8,286	–	PV_{CE}
	(text)	S	19,972	7.4	147,560	9,021	–	BT_{CS}, PV_{ES}
devset1$_{CSTAR03}$	(read-speech)	A	506	5.0	2,555	1,156	–	BT_{AE}
	(read-speech)	C	506	5.5	2,808	877	–	BT_{CE}
	(text)	E	8,096	6.8	55,383	2,134	16	BT_{CE}, BT_{AE}
devset2$_{IWSLT04}$	(read-speech)	A	500	5.3	2,660	1,237	–	BT_{AE}
	(read-speech)	C	500	5.8	2,906	917	–	BT_{CE}
	(text)	E	8,000	6.9	55,027	2,233	16	BT_{CE}, BT_{AE}
devset3$_{IWSLT05}$	(read-speech)	A	506	5.1	2,566	1,263	–	BT_{AE}
	(read-speech)	C	506	6.3	3,209	929	–	BT_{CE}, BT_{CS}, PV_{CE}
	(text)	E	506	6.2	3,119	840	–	PV_{ES}
	(text)		8,096	6.9	55,959	2,323	16	BT_{CE}, BT_{AE}, PV_{CE}
	(text)	S	8,096	6.3	50,420	2,616	16	BT_{CS}, PV_{ES}
devset4$_{IWSLT06}$	(read-speech)	A	489	8.6	4,185	1,618	–	BT_{AE}
	(spontaneous)	C	489	10.7	5,226	1,142	–	BT_{CE}
	(text)	E	3,423	11.4	39,174	1,817	7	BT_{CE}, BT_{AE}
devset5$_{IWSLT06}$	(read-speech)	A	500	9.3	4,652	1,950	–	BT_{AE}
	(spontaneous)	C	500	11.1	5,566	1,338	–	BT_{CE}
	(text)	E	3,500	12.6	44,079	2,036	7	BT_{CE}, BT_{AE}
devset6$_{IWSLT07}$	(read-speech)	A	489	4.9	2,383	1,164	–	BT_{AE}
	(text)	C	489	5.4	2,647	878	–	BT_{CE}
	(text)	E	2,934	6.4	18,776	1,362	7	BT_{CE}, BT_{AE}
devset$_{IWSLT08}$	(text)	C	1,757	5.2	9,136	553	7	CT_{EC}
	(spontaneous)		246	5.3	1,305	248	–	CT_{CE}
	(spontaneous)	E	251	5.1	1,283	239	–	CT_{EC}
	(text)		1,722	7.0	12,076	577	7	CT_{CE}
testset$_{IWSLT08}$	(text)	C	3,486	5.7	20,016	708	7	CT_{EC}
	(spontaneous)		504	5.0	2,513	385	–	CT_{CE}
	(spontaneous)	E	498	5.8	2,867	312	–	CT_{EC}
	(text)		3,528	6.2	21,751	810	7	CT_{CE}
	(read-speech)	A	507	5.1	2,585	1,205	–	BT_{AE}
	(read-speech)	C	507	5.5	2,808	885	–	BT_{CE}, BT_{CS}
	(text)	E	8,112	6.8	55,082	2,146	16	BT_{CE}, BT_{AE}
	(text)	S	8,112	6.2	50,169	2,569	16	BT_{CS}, PV_{CS}

campaigns as well as one third of the newly collected spontaneous data sets of the *Challenge Task* were provided to the participants together with up to 16 reference translations for each of the target languages.

Concerning the evaluation data sets of the *Challenge Task* of IWSLT 2008, machine-mediated conversational speech was recorded in a real situation. Foreign travelers were provided with a state-of-the-art speech-to-speech translation handheld device and were asked to carry-out specific tourism-related tasks (e.g., buying entrance tickets) by using the device to communicate with local staff. In total, speech data of 50 English and 50 Chinese travelers were released as the evaluation data set of CT_{EC} and CT_{CE}, respectively. Recordings were done at 5 different locations where each speaker carried out 3-4 tasks.

The evaluation data sets of the *BTEC Task* and the *Pivot Task* consisted of read-speech recordings whereby the source language texts were read aloud by 10 native speakers[8]. The reference translations for BT_{CE} and BT_{AE} as well as BT_{CS} and PV_{CS} were the same. They were produced by 5 human translators who created up to three paraphrases of the original corpus sentences each.

Details of the IWSLT 2008 spoken language corpus are given in Table 3. 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

[8] An exception was Arabic, with only two native speakers.

Table 4: Out-Of-Vocabulary Rates

data set	lang	OOV (%) CRR	1BEST	NBEST	data track
devset1$_{CSTAR03}$	A	5.5	–	–	BT$_{AE}$
	C	5.0	–	–	BT$_{CE}$, BT$_{CS}$
		2.3	–	–	PV$_{CE}$
devset2$_{IWSLT04}$	A	5.7	–	–	BT$_{AE}$
	C	4.1	–	–	BT$_{CE}$, BT$_{CS}$
		2.5	–	–	PV$_{CE}$
devset3$_{IWSLT05}$	A	6.2	–	–	BT$_{AE}$
	C	3.3	3.0	4.0	BT$_{CE}$, BT$_{CS}$
		4.8	3.4	4.7	PV$_{CE}$
	E	2.0	–	–	PV$_{ES}$
devset4$_{IWSLT06}$	A	16.1	17.4	18.8	BT$_{AE}$
	C	3.5	4.3	4.5	BT$_{CE}$, BT$_{CS}$
		3.8	4.4	4.7	PV$_{CE}$
devset5$_{IWSLT06}$	A	17.3	19.5	20.6	BT$_{AE}$
	C	4.2	3.9	4.2	BT$_{CE}$, BT$_{CS}$
		4.7	4.5	4.9	PV$_{CE}$
devset6$_{IWSLT07}$	A	18.0	16.0	17.4	BT$_{AE}$
	C	5.3	–	–	BT$_{CE}$, BT$_{CS}$
		2.7	–	–	PV$_{CE}$
devset$_{IWSLT08}$	C	4.1	3.4	4.2	CT$_{CE}$
	E	3.0	2.1	3.1	CT$_{EC}$
testset$_{IWSLT08}$	A	9.9	11.7	14.5	BT$_{AE}$
	C	2.9	2.5	3.8	CT$_{CE}$
		3.9	3.5	4.5	BT$_{CE}$, BT$_{CS}$
		2.2	2.5	3.7	PV$_{CE}$
	E	3.0	2.6	3.4	CT$_{EC}$

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 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 data tracks that permitted the usage of the respective resource are listed in the "*data track*" column. All resources of the BT$_{CE}$ were also permitted for the CT$_{CE}$ data track.

Table 4 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). In general, the OOV rates of CRR are higher than the OOV rates of the 1BEST data sets, because unknown words might either be ignored or mis-recognized as known words by the ASR engine. For NBEST lists, OOV rates are naturally higher than those of the 1BEST data sets. Very large OOV rates of up to 17% are obtained for the Arabic data sets which are mainly caused by word segmentation issues (*prefix/postfix* attachment) and spelling variations in Arabic. The lowest OOV rates for the CRR data are found for Chinese for the *Pivot Task*. The figures of the Chinese *BTEC Task* are twice as high as the ones for the *Pivot Task*. This indicates that the Chinese evaluation data sets are better covered by the PV$_{CE}$ corpus compared to the BT$_{CE}$ resources.

In order to get an idea of how difficult the IWSLT 2008 translation tasks were, we used the *SRI Language Modeling Toolkit*[9] 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*, of each language on the respective evaluation data sets. The entropy figures given in Table 5 indicate that CT$_{EC}$ can be expected to be the easiest task and that the *BTEC Task* can be expected to be more difficult to translate than the *Challenge Task*. This is confirmed for the CRR inputs by the automatic evaluation results listed in Appendix C. However, this is not the case for the *ASR output* translation results which indicates that recognition errors have a larger impact on the *Challenge Task* translation results compared to the ones of the *BTEC Task*.

The recognition accuracies of the utilized ASR engines for the IWSLT 2008 evaluation data sets are summarized in Table 6. The *lattice accuracy* figures show the percentage of correct recognition results contained in the lattices, where the *1BEST accuracy* is the accuracy of the best path extracted from each lattice.

Apart from Arabic, the *word accuracies* of the utilized ASR engines ranged between 87%-95% (lattice) and 79%-86% (1BEST), where the percentages of correctly recognized sentences (*sentence accuracy*) ranged between 65%-80% (lattice) and 53%-63% (1BEST). The 1BEST recognition results for the Arabic speech data were much lower (word: 73%, sentence: 36%). Unfortunately, the lattice accuracies for Arabic were not available.

Concerning different data types, similar lattice accuracies were obtained for CT$_{CE}$ and BT$_{CE}$. However, CT$_{CE}$'s 1BEST recognition results on sentence-level are 10% lower than the BT$_{CE}$ recognition results which seem to cause worse CT$_{CE}$ ASR output translation results than for the BT$_{CE}$ task (see Section 3).

2.2. Evaluation Specifications

The *official evaluation* specifications for IWSLT 2008 were identical to the ones used in the IWSLT 2006 and 2007 evaluation campaigns and were defined as:

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

For the convenience of the participants, automatic evaluation scores were also calculated for the following *additional 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 were case-insensitive and without punctuation information. However, true-case and punctuation information was provided

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

Table 5: Language Model Perplexity

data set	lang	entropy	words	total entropy	data track
devset$_{IWSLT08}$	C	9.71	1,710	16,604	CT$_{EC}$
	E	9.84	1,980	19,483	CT$_{CE}$
testset$_{IWSLT08}$	C	9.51	3,962	35,111	CT$_{EC}$
	E	10.10	3,662	36,986	CT$_{CE}$
		9.83	4,057	39,880	BT$_{CE}$, BT$_{AE}$
	S	10.25	3,885	39,821	BT$_{CS}$, PV$_{CS}$

Table 6: Recognition Accuracy

data set (data type)		lang	word (%)		sentence (%)		data track
			lattice	1BEST	lattice	1BEST	
devset$_{IWSLT08}$	(spontaneous)	C	95.33	86.90	78.46	58.54	CT$_{CE}$
		E	90.41	80.98	72.11	53.78	CT$_{EC}$
testset$_{IWSLT08}$	(spontaneous)	C	95.07	85.79	79.56	53.77	CT$_{CE}$
		E	87.27	79.77	65.06	53.01	CT$_{EC}$
	(read-speech)	A	–	72.80	–	36.10	BT$_{AE}$
		C	94.20	83.61	80.47	63.31	BT$_{CE}$, BT$_{CS}$, PV$_{CS}$

for all training data sets that could be used for recovering case/punctuation information according to the official evaluation specifications. Instructions[10] 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. Automatic Evaluation

The automatic evaluation of run submissions was carried out using a large number of standard automatic evaluation metrics whereby the automatic metric scores of all primary and contrastive runs were sent back to the participants one week after the run submission period.

For the official evaluation results[11] of the IWSLT 2008 workshop, we utilized the average score ("*(B+M)/2*") of the two automatic evaluation metrics listed in Table 7.

Table 7: 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) [5]
METEOR:	a metric that 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) [6]

2.2.2. Subjective Evaluation

Human assessments of translation quality 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. The *fluency* and *adequacy* judgments consist of one of the grades listed in Table 8[12]. 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 8: Human Assessment

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

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

Due to high evaluation costs, the *fluency* and *adequacy* assessments were limited to MT outputs of four systems per data track. The systems were selected based on the obtained "(B+M)/2" automatic evaluation scores as well as the amount of innovative ideas carried out for this year's event by the participants. Moreover, in order to get an idea of the range of translation quality, MT systems covering the top, middle, and lower performance levels of the respective data track were selected. In total, 24 run submissions were evaluated using the *fluency* and *adequacy* criteria. In order to reduce the costs further, the human assessment was limited to the translation outputs of 300 input sentences selected from the

[10] http://www.slc.atr.jp/IWSLT2008/downloads/case+punc_tool_using_SRILM.instructions.txt

[11] In addition to the official evaluation metrics used for IWSLT 2008, the *word error rate* (WER) [10], the *position-independent WER* (PER) [11], the translation error rate (TER) ([12]) and the *general text matcher* (GTM) [13] and NIST [14] scores were also calculated and provided to the participants for the analysis of their systems.

[12] Fluency grades are defined for the respective target language (C: Chinese, E: English, S: Spanish).

Table 9: Human Evaluators

lang	native	non-native
C	6	1
E	5	1
S	11	–
$\sum$	22	2

respective *testset* data sets. In addition, all translation results were *pooled*, i.e., in case of identical translations of the same source sentence by multiple engines, the translation was graded only once, and the respective rank was assigned to all MT engines with the same output. Each translation was evaluated by at least three judges where each system score is calculated as the *median* of the assigned grades. The evaluation was carried out by 22 native speakers and 2 non-native speakers with sufficient knowledge of the target language (see Table 9). All graders took part in a dry-run evaluation exercise in order to get used to the evaluation metrics as well as the browser-based graphical user interfaces.

In addition to the fluency/adequacy evaluation, an additional subjective evaluation metric that ranks MT system outputs according to their translation quality was applied to all primary runs submitted by the participants. 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)"* [7]. Similar to the *fluency/adequacy* assessments, 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 [15] were calculated for each run submission.

Moreover, 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 translations of the evaluation data set, the first system was compared towards the second system output 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. In addition, 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 data track.

2.2.3. 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 [16] is given in Table 10.

Table 10: Interpretation of κ Coefficient [16]

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

2.2.4. Statistical Significance of Evaluation Results

In order to decide whether the translation output on document-level of one MT engine is significantly better than another one, 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 [17]. In this year's evaluation, 2000 iterations were used for the analysis of the IWSLT 2008 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 could 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 2008 workshop are summarized in Appendix B (*human assessment*) and Appendix C (*automatic evaluation*). The correlation rank coefficients of subjective and automatic evaluation results are given in Appendix D. For each evaluation metric, the best score of each data track is marked in boldface.

Table 11: Intra-Grader Consistency

Metric	κ Coefficient					
	CT_{EC}	CT_{CE}	BT_{CE}	BT_{AE}	BT_{CS}	PV_{CS}
fluency	0.64	0.71	**0.75**	0.71	0.52	0.54
adequacy	0.74	0.68	**0.81**	0.61	0.67	0.70
ranking	**0.73**	0.56	0.69	0.56	0.52	0.56

Table 12: Inter-Grader Consistency

Metric	κ Coefficient					
	CT_{EC}	CT_{CE}	BT_{CE}	BT_{AE}	BT_{CS}	PV_{CS}
fluency	0.41	0.38	0.41	**0.44**	0.19	0.25
adequacy	0.40	0.41	0.46	**0.47**	0.26	0.30
ranking	**0.57**	0.52	0.56	0.54	0.47	0.51

3.1. Subjective Evaluation Results

Each sentence was evaluated by at least 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. The κ coefficients for *intra-grader* and *inter-grader* consistencies are given in Table 11 and Table 12. The highest κ coefficient for each subjective metric is marked in boldface.

The obtained overall *intra-grader* κ coefficients were high (*fluency*: 0.64, *adequacy*: 0.70, *ranking*: 0.59) and showed that all graders submitted very consistent evaluation grades. Substantial agreement levels were achieved for most evaluation tasks with Chinese and English as the target language. Only moderate agreement was achieved for the repeated *fluency* assessments of Spanish translation results. Concerning the evaluation types, the levels of *intra-grader* consistency were: *adequacy* > *fluency* > *ranking*.

However, the picture is reversed for the *inter-grader* consistency evaluation. The best level of agreement was achieved for the *ranking* metrics (overall: 0.50, moderate agreement), followed by *adequacy* (overall: 0.38) and *fluency* (overall: 0.35) achieving only a fair agreement between the graders of the respective data tracks. In the case of *fluency* only a slight agreement could be achieved for Spanish. These low κ coefficients for Spanish might be partly caused by (1) the lower translation quality of the Spanish MT outputs and (2) the lower level of experience of the first-time Spanish volunteers compared to the (partly professional) evaluators who had already taken part in the Chinese and English translation tasks of previous IWSLT evaluation campaigns.

The criteria for *fluency* and *adequacy* seems to be more precise ($\rightarrow$ higher *intra-grader* consistency), but allow for more variations in grading results due to a larger amount of choices and to different interpretations of the grades by each evaluator ($\rightarrow$ lower *inter-grader* consistency).

In order to minimize the impact of grader inconsistencies, only the grading results of the three most self-consistent graders of each data track were utilized and the *median* of the assigned grades was selected for the *fluency/adequacy* assessments as the final judgment for each sentence.

3.1.1. Fluency/Adequacy Performance

The results of the IWSLT 2008 *fluency/adequacy* evaluation for the primary ASR output runs are summarized in Appendix B.1. For each of the selected MT system outputs, the mean score and the 95% confidence intervals were calculated according to the *bootStrap* method [17]. The systems are ordered according to the average of the respective mean *fluency* and *adequacy* scores with the highest metric scores marked in boldface. The four systems were selected so that they cover the full range of translation quality (*high - middle - low*) for each data track. Therefore, it was to be expected that all differences in the metric scores were significantly different and that the system with highest average score was also ranked first in the single metrics. However, an exception was the *greyc* MT system of the PV_{CS} data track which obtained the highest *fluency* score, but the worst *adequacy* score.

Moreover, the *fluency/adequacy* scores differ largely between the data tracks. The highest average scores were obtained for BT_{CE} (*fluency*: 3.21, adequacy: 2.46), followed by CT_{CE} (*fluency*: 3.27, adequacy: 2.36), CT_{EC} (*fluency*: 2.75, adequacy: 2.39), and BT_{AE} (*fluency*: 3.15, adequacy: 2.18). The lowest translation quality was achieved for the Spanish translation tasks, i.e. BT_{CS} (*fluency*: 2.50, adequacy: 1.87) and PV_{CS} (*fluency*: 2.5, adequacy: 1.81).

3.1.2. Ranking Performance

The results of the IWSLT 2008 *ranking* evaluation are summarized in Appendix B.2. For each data track, all systems are ranked according to the *ranking* scores, i.e., the average number of times that a system was judged better than any other systems. Although the given rankings slightly differ compared to those rankings based on the *NormRank* scores for all data tracks besides BT_{CS}, both metrics agree on the top-ranked MT system, which is the *tch* system for all non-English data tracks (CT_{EC}, BT_{CS}, PV_{CS}), the *nlpr* system for the Chinese-English tasks (CT_{CE}, BT_{CE}) and the *mitll* system for BT_{AE}.

In order to get an idea about 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 towards the second system as the difference of the number of translations of the first system ranked better than the ones of the second system and the number of translations ranked worse, divided by the number of times both systems were ranked together. The results listed in Appendix B.3. indicate some inconsistencies of the *rankings* metric, because several MT system combinations result in negative gains when compared directly.

In order to avoid these inconsistencies, we calculated the *BestRankDiff* scores that rank all MT systems of each data track 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 B.4.

3.2. Automatic Evaluation Results

The automatic evaluation results of all MT engines using the official evaluation specifications, i.e., *case-sensitive with punctuation marks tokenized*, as well as the additional evaluation specifications, i.e., *lowercase without punctuation marks*, are listed in Appendix C. All primary run submission results for the ASR output translations are given on the lefthandside and the ones for the CRR input conditions are given on the righthandside of the tables. The MT systems are ordered according to the average of the BLEU and METEOR scores obtained for the primary run submission of the ASR output translation condition. 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 data track, the highest scores of the respective evaluation metric are highlighted in boldface.

Besides the CT_{CE} task where the two top-ranked systems were not significantly different, the MT systems of all data tracks that obtain the highest automatic evaluation 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.

3.3. Evaluation Metric Correlations

In order to get an idea of how closely the respective metrics are related, the *Spearman rank correlation coefficients* were calculated for all automatic evaluation metric combinations. Appendix D summarizes the comparison of the human assessment results and the official automatic evaluation metrics. Due to the limited number of graded systems, the obtained correlation coefficients for *fluency* and *adequacy* in Appendix D.1. might not be conclusive, but the results seem to confirm findings of previous IWSLT campaigns that *fluency* correlates well with BLEU and that *adequacy* correlates well with METEOR for the given travel tasks. An exception again is the PV_{CS} data track, where none of the automatic evaluation metrics agreed with the obtained *fluency* rankings.

A larger number of systems was used for the calculation of the correlation coefficients for the *ranking* metrics. The results in Appendix D.2. show that the *NormRank* metric correlates consistently better with automatic evaluation metrics than the *ranking* metric. However, the highest correlation coefficients were obtained for the *BestRankDiff* metric, especially when the systems are ranked according to the official automatic evaluation metrics "(B+M)/2".

4. Discussion

4.1. Challenge Task 2008

The novelty of this year's evaluation campaign was the usage of machine-mediated spontaneous speech data collected in field experiments using inexperienced users and state-of-the-art speech-to-speech technologies. In order to identify differences between such real-world data sets towards the more synthetic data collections of the BTEC* corpus, we compared the outcomes of the *Challenge Task* (CT_{CE}) and the *BTEC Task* using the Chinese-English translation results.

CT_{CE} sentences were on average shorter (see Table 3) and less complex (see Table 5) than the BTEC*sentences, but the translation quality of the system outputs for the ASR output condition were worse than those of the BT_{CE} translation results. Although quite similar ASR recognition performance was achieved in terms of lattice input (word: 95%, sentence: 80%) and *1BEST* word accuracy (84%), the recognition performance for the CT_{CE} data sets on sentence-level were much worse, i.e., a large drop of 10% in sentence-level recognition accuracy for 1BEST were obtained (see Table 6). Unfortunately, most participants used only the 1BEST input for both, the CT_{CE} and the BT_{CE}, translation tasks, thus resulting in lower ASR output scores for CT_{CE}.

4.2. Language Dependency

Although it is difficult to compare translation results across different languages and evaluation data sets, the overall translation quality of the primary run submissions of the CT_{CE} translation tasks seems to be higher than the results for the *Challenge Task* of 2006. The reasons are the lower complexity of this year's *Challenge Task*. However, the small amount of in-domain language resources might have prevented better translations due to the out-of-vocabulary and domain-mismatch problems of statistical models. Concerning the translation quality of this year's translations tasks, the data tracks can be ordered according to the *fluency* and *adequacy* assessment of the ASR output condition as:

$$\text{BT}_{CE} > \text{CT}_{CE} \approx \text{CT}_{EC} > \text{BT}_{AE} > \text{BT}_{AE} > \text{BT}_{CS} > \text{PV}_{CS}$$

4.3. Evaluation Metrics

It is well known that human assessments of *fluency* and *adequacy* judgments is quite expensive, even if cost reduction methods, like *pooling* and *evaluation data size limitation*, are applied. Therefore, not enough systems could be evaluated for *fluency* and *adequacy* during this year's subjective evaluation to make reliable comparisons to automatic evaluation metrics. The *ranking* metric requires lower evaluation costs, because multiple systems are judged simultaneously. However, the usage of this metric alone is not sufficient because *ranking* scores can only define a relative order, without providing 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 towards 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 incorrect, as shown by the negative gains obtained for several systems of the *paired-comparison* evaluation.

In contrast, the *BestRankDiff* metric scores are based on those ranking results where two MT systems were ranked together and the same point of reference is used. Thus, a direct relative comparison where the absolute translation quality is defined by the difference in performance with the best scoring system is possible. For IWSLT 2008, the *BestRankDiff* metric achieved the best correlation towards the official automatic evaluation metrics "*(B+M)/2"*. However, the agreement between the top-scoring MT systems according to the "*(F+A)/2*" and the *ranking* metrics for all IWSLT 2008 data tracks and the highest *inter-grader* consistency coefficients showed that the *ranking* method is a reliable method to identify the best performing system.

In order to minimize evaluation costs, to obtain consistent judgments of overall machine translation quality, and to be able to reliably define a relative ordering on MT systems, future IWSLT evaluations could be carried out as follows: for each data track, (1) a *ranking* evaluation for all primary system outputs will be performed to identify the best-performing MT system for each track, (2) a *fluency/adequacy* assessment will be carried out for best MT system only using the full data set and more than three human graders, and (3) systems will be ranked according to the *BestRankDiff* metric scores.

5. Conclusion

This year's workshop provided a testbed for verifying the quality of state-of-the-art speech-to-speech translation technologies for real-world applications using machine-mediated spontaneous speech data collected from inexperienced users in a real situation. Various innovative ideas were explored, most notably *increase of synthetic training resources by translating in-domain monolingual resources*, *advanced techniques for phrase extraction from NBEST alignments*, *improved statistical modeling techniques*, *system combinations*, and *rescoring/reranking methods of NBEST lists*. Although this year's evaluation data sets did not take into account the full context of the face-to-face conversations, new insights into the requirements of speech translation technologies for real world applications were obtained that will help to advance the current state-of-the-art in speech-to-speech translation.

6. Acknowledgments

I thank the C-STAR partners for their accomplishments during the preparation of this workshop and the subjective evaluation task. In particular, I would like to thank Gen Itoh, Shigeki Matsuda, and Mark Fuhs for their support in recording the speech data sets and generating the ASR outputs. Special thanks to Matthias Eck and Chris Callison-Burch for providing us with the software to run the automatic evaluation server and the subjective ranking evaluation, respectively. In addition, I thank all volunteers who carried out the human assessment of the translation outputs, including Jewel Faulkner, Cameron Fordyce, Ann Gethin, Yoshiko Kakitani, Maxim Khalilov and the UPC-TALP team, Lisa Mauti, Haifeng Wang and his colleagues at Toshiba. 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 2008 evaluation campaign and for making the IWSLT 2008 workshop a success.

7. References

[1] Y. Akiba, M. Federico, N. Kando, H. Nakaiwa, M. Paul, and J. Tsujii, "Overview of the IWSLT04 evaluation campaign," in *Proc. of IWSLT*, Kyoto, Japan, 2004, pp. 1–12.

[2] M. Eck and C. Hori, "Overview of the IWSLT 2005 evaluation campaign," in *Proc. of IWSLT*, Pittsburgh, USA, 2005, pp. 11–32.

[3] M. Paul, "Overview of the IWSLT 2006 evaluation campaign," in *Proc. of IWSLT*, Kyoto, Japan, 2006, pp. 1–15.

[4] C. S. Fordyce, "Overview of the IWSLT 2007 evaluation campaign," in *Proc. of IWSLT*, Trento, Italy, 2007, pp. 1–12.

[5] 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.

[6] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, 2005, pp. 65–72.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] A. Zollmann, A. Venugopal, and S. Vogel, "The CMU Syntax-Augmented Machine Translation System: SAMT on Hadoop with N-best Alignments," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 18–25.

[19] Y. Ma, J. Tinsley, H. Hassan, J. Du, and A. Way, "Exploiting Alignment Techniques in MaTrEx: the DCU Machine Translation System for IWSLT08," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 26–33.

[20] N. Bertoldi, R. Cattoni, M. Federico, and M. Barbaiani, "FBK @ IWSLT-2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 34–38.

[21] Y. Lepage, A. Lardilleux, J. Gosme, and J.-L. Manguin, "The GREYC Machine Translation System for the IWSLT 2008 Evaluation Campaign," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 39–45.

[22] B. Chen, D. Xiong, M. Zhang, A. Aw, and H. Li, "I^2R Multi-Pass Machine Translation System for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 46–51.

[23] Y. Liu, Z. He, H. Mi, Y. Huang, Y. Feng, W. Jiang, Y. Lu, and Q. Liu, "The ICT System Description for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 52–57.

[24] L. Besacier, A. B. Youssef, and H. Blanchon, "The LIG Arabic/English Speech Translation System at IWSLT08," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 58–62.

[25] H. Schwenk, Y. Estève, and S. A. Rauf, "The LIUM Arabic/English Statistical Machine Translation System for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 63–68.

[26] W. Shen, B. Delaney, T. Anderson, and R. Slyh, "The MIT-LL/AFRL IWSLT-2008 MT System," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 69–76.

[27] M. Utiyama, A. Finch, H. Okuma, M. Paul, H. Cao, H. Yamamoto, K. Yasuda, and E. Sumita, "The NICT/ATR Speech Translation System for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 77–84.

[28] Y. He, J. Zhang, M. Li, L. Fang, Y. Chen, Y. Zhou, and C. Zong, "The CASIA Statistical Machine Translation System for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 85–91.

[29] K. Sudoh, T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki, "NTT Statistical Machine Translation System for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 92–97.

[30] J. Lee and G. G. Lee, "POSTECH Machine Translation System for IWSLT 2008 Evaluation Campaign," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 98–103.

[31] S. Carter, C. Monz, and S. Yahyaei, "The QMUL System Description for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 104–107.

[32] D. Vilar, D. Stein, Y. Zhang, E. Matusov, A. Mauser, O. Bender, S. Mansour, and H. Ney, "The RWTH Machine Translation System for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 108–115.

[33] M. Khalilov, M. R. Costa-Jussà, C. A. Henríquez, J. A. R. Fonollosa, A. Hernández, J. B. Mariño, R. E. Banchs, C. Boxing, M. Zhang, A. Aw, and H. Li, "The TALP & I2R SMT Systems for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 116–123.

[34] H. Wang, H. Wu, X. Hu, Z. Liu, J. Li, D. Ren, and Z. Niu, "The TCH Machine Translation System for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 124–131.

[35] J. Murakami, M. Tokuhisa, and S. Ikehara, "Statistical Machine Translation without Long Parallel Sentences for Training Data," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 132–137.

[36] C. Mermer, H. Kaya, Ö. F. Güneş, and M. U. Doğan, "The TÜBİTAK-UEKAE Statistical Machine Translation System for IWSLT 2008," in *Proc. of IWSLT*, Hawaii, USA, 2008, pp. 138–142.

Appendix A. MT System Overview

Research Group	MT System Description	Type	System	Submissions
Carnegie Mellon University, InterACT Research Labs (USA)	The CMU Syntax-Augmented Machine Translation System: SAMT on Hadoop with N-best Alignments [18]	SMT	cmu	BT_{CE}
Dublin City University, School of Computing (Ireland)	Exploiting Alignment Techniques in MaTrEx: the DCU Machine Translation System for IWSLT08 [19]	SMT EBMT	dcu	CT_{EC}, CT_{CE}, BT_{CE}, BT_{AE}, BT_{CS}, PV_{CS}
Fondazione Bruno Kessler, Ricerca Scientifica e Tecnologica (Italy)	FBK @ IWSLT-2008 [20]	SMT	fbk	CT_{CE}, BT_{CE}, BT_{CS}, PV_{CS}
University of Caen Basse-Normandie, GREYC (France)	The GREYC Machine Translation System for the IWSLT 2008 Evaluation Campaign [21]	EBMT	greyc	BT_{CE}, BT_{AE}, BT_{CS}, PV_{CS}
Institute for Infocomm Research (Singapore)	I^2R Multi-Pass Machine Translation System for IWSLT 2008 [22]	SMT	i2r	CT_{CE}, BT_{CE}
Chinese Academy of Sciences, Insititute of Computing Technology (China)	The ICT System Description for IWSLT 2008 [23]	SMT	ict	CT_{EC}, CT_{CE}, BT_{CE}
University J. Fourier, GETALP, LIG (France)	The LIG Arabic/English Speech Translation System at IWSLT08 [24]	SMT	lig	BT_{AE}
University of Le Mans, LIUM (France)	The LIUM Arabic/English Statistical Machine Translation System for IWSLT 2008 [25]	SMT	lium	BT_{AE}
MIT Lincoln Laboratory (USA)	The MIT-LL/AFRL IWSLT-2008 MT System [26]	SMT	mitll	CT_{EC}, CT_{CE}, BT_{AE}
National Institute of Information and Communications Technology (Japan)	The NICT/ATR Speech Translation System for IWSLT 2008 [27]	SMT	nict	CT_{EC}, CT_{CE}, BT_{CE}, BT_{CS}, PV_{CS}
Chinese Academy of Sciences, National Laboratory of Pattern Recognition (China)	The CASIA Statistical Machine Translation System for IWSLT 2008 [28]	SMT	nlpr	CT_{EC}, CT_{CE}, BT_{CE}
NTT Communication Science Laboratories (Japan)	NTT Statistical Machine Translation System for IWSLT 2008 [29]	SMT	ntt	CT_{EC}
Pohang University of Science and Technology (Korea)	POSTECH Machine Translation System for IWSLT 2008 Evaluation Campaign [30]	SMT	postech	BT_{CE}, BT_{AE}, BT_{CS}
Queen Mary University of London (UK)	The QMUL System Description for IWSLT 2008 [31]	SMT	qmul	BT_{CE}, BT_{AE}, PV_{CS}
Rheinisch Westfälische Technische Hochschule (Germany)	The RWTH Machine Translation System for IWSLT 2008 [32]	SMT	rwth	CT_{CE}, BT_{CE}, BT_{AE}
UPC, TALP Research Center (Spain)	The TALP & I2R SMT Systems for IWSLT 2008 [33]	SMT	talp	BT_{AE}, BT_{CS}, PV_{CS}
Toshiba China R&D Center (China)	The TCH Machine Translation System for IWSLT 2008 [34]	SMT RBMT	tch	CT_{EC}, CT_{CE}, BT_{CE}, BT_{CS}, PV_{CS}
Tottori University (Japan)	Statistical Machine Translation without Long Parallel Sentences for Training Data [35]	SMT	tottori	CT_{EC}, CT_{CE}, BT_{CE}
TÜBİTAK-UEKAE (Turkey)	The TÜBİTAK-UEKAE Statistical Machine Translation System for IWSLT 2008 [36]	SMT	tubitak	BT_{CE}, BT_{AE}, BT_{CS}, PV_{CS}

Appendix B. Human Assessment

B.1. Fluency/Adequacy

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

· the mean score and the 95% confidence intervals were calculated for each MT output according to the *bootStrap* method [17].
· the MT systems are ordered according to the average of mean fluency and adequacy scores.

CT$_{EC}$

MT	Adequacy	Fluency
tch.ASR.1	**2.3894** [2.2071, 2.5716]	**2.7540** [2.6019, 2.9060]
ict.ASR.1	2.2442 [2.0600, 2.4284]	2.7363 [2.5819, 2.8907]
nlpr.ASR.5	2.2784 [2.1006, 2.4563]	2.5823 [2.4236, 2.7410]
tottori.ASR.1	1.6592 [1.4804, 1.8379]	2.0518 [1.8812, 2.2224]

CT$_{CE}$

MT	Adequacy	Fluency
tch.ASR.1	**2.3578** [2.1810, 2.5346]	**3.2748** [3.1563, 3.3934]
ict.ASR.1	2.0926 [1.9179, 2.2673]	3.0951 [2.9692, 3.2209]
rwth.ASR.1	1.9579 [1.7822, 2.1337]	2.8250 [2.6890, 2.9611]
fbk.ASR.1	1.7020 [1.5359, 1.8681]	2.5522 [2.3989, 2.7054]

BT$_{CS}$

MT	Adequacy	Fluency
tch.ASR.1	**1.8667** [1.7028, 2.0306]	**2.4969** [2.3501, 2.6437]
nict.ASR.1	1.3854 [1.2313, 1.5396]	2.0882 [1.9369, 2.2395]
fbk.ASR.1	1.4793 [1.3291, 1.6295]	1.8541 [1.7126, 1.9955]
postech.ASR.1	1.2177 [1.0777, 1.3576]	1.2256 [1.0959, 1.3554]

BT$_{CE}$

MT	Adequacy	Fluency
tch.ASR.1	**2.4619** [2.2860, 2.6378]	**3.2112** [3.0893, 3.3332]
i2r.ASR.1	2.2398 [2.0625, 2.4172]	3.1415 [3.0673, 3.3262]
cmu.ASR.1	2.1533 [1.9650, 2.3416]	3.1967 [3.0139, 3.2690]
tubitak.ASR.1	1.7577 [1.5820, 1.9334]	2.3530 [2.1952, 2.5107]

PV$_{CS}$

MT	Adequacy	Fluency
tch.ASR.1	**1.8082** [1.6546, 1.9617]	2.4762 [2.3263, 2.6260]
talp.ASR.1	1.4185 [1.2753, 1.5618]	2.1740 [2.0314, 2.3167]
greyc.ASR.1	0.7495 [0.6169, 0.8821]	**2.8143** [2.6835, 2.9451]
dcu.ASR.5	1.4131 [1.2609, 1.5653]	2.0312 [1.8748, 2.1875]

BT$_{AE}$

MT	Adequacy	Fluency
mitll.ASR.SLF	**2.1813** [1.9979, 2.3648]	**3.1476** [3.0174, 3.2779]
talp.ASR.1	1.9968 [1.8208, 2.1727]	2.6828 [2.5273, 2.8383]
dcu.ASR.1	1.9877 [1.8165, 2.1588]	2.3684 [2.2181, 2.5187]
qmul.ASR.1	1.4510 [1.2879, 1.6141]	1.8962 [1.7523, 2.0401]

B.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 [15].

CT$_{EC}$

MT	Ranking	NormRank
tch.ASR.1	**0.3966**	**2.38**
nlpr.ASR.5	0.3806	2.24
ict.ASR.1	0.3752	2.28
dcu.ASR.1	0.3043	2.10
nict.ASR.20	0.2095	1.71
mitll.ASR.SLF	0.1952	1.69
tottori.ASR.1	0.1838	1.61

BT$_{CS}$

MT	Ranking	NormRank
tch.ASR.1	**0.4773**	**2.45**
fbk.ASR.1	0.3342	2.11
nict.ASR.1	0.2979	2.04
tubitak.ASR.1	0.2899	2.00
dcu.ASR.1	0.2832	2.00
talp.ASR.1	0.2642	1.95
postech.ASR.1	0.2332	1.83
greyc.ASR.1	0.1994	1.62

PV$_{CS}$

MT	Ranking	NormRank
tch.ASR.1	**0.4932**	**2.47**
fbk.ASR.1	0.3990	2.27
talp.ASR.1	0.3545	2.10
tubitak.ASR.1	0.3416	2.12
dcu.ASR.1	0.3172	2.03
nict.ASR.1	0.3088	2.01
qmul.ASR.1	0.1772	1.56
greyc.ASR.1	0.1546	1.44

CT$_{CE}$

MT	Ranking	NormRank
nlpr.ASR.5	**0.5274**	**2.48**
tch.ASR.1	0.4657	2.37
ict.ASR.1	0.3869	2.13
i2r.ASR.1	0.3863	2.11
mitll.ASR.SLF	0.3686	2.00
rwth.ASR.1	0.3423	1.96
dcu.ASR.1	0.3331	1.95
ntt.ASR.1	0.3327	1.87
nict.ASR.1	0.3127	1.89
fbk.ASR.1	0.2585	1.71
tottori.ASR.1	0.2074	1.53

BT$_{CE}$

MT	Ranking	NormRank
nlpr.ASR.5	**0.5255**	**2.60**
tch.ASR.1	0.4900	2.54
ict.ASR.1	0.4668	2.44
i2r.ASR.1	0.4393	2.38
rwth.ASR.1	0.4060	2.20
cmu.ASR.1	0.4051	2.24
dcu.ASR.1	0.3302	2.02
fbk.ASR.1	0.3291	2.01
nict.ASR.1	0.2965	1.83
tubitak.ASR.1	0.2813	1.88
tottori.ASR.1	0.2342	1.66
postech.ASR.1	0.2138	1.58
greyc.ASR.1	0.1468	1.26
qmul.ASR.1	0.1603	1.35

BT$_{AE}$

MT	Ranking	NormRank
mitll.ASR.SLF	**0.4415**	**2.42**
talp.ASR.1	0.3901	2.27
rwth.ASR.1	0.3822	2.21
lig.ASR.SLF	0.3756	2.09
lium.ASR.1	0.3741	2.19
dcu.ASR.1	0.3634	2.18
tubitak.ASR.1	0.3574	2.20
qmul.ASR.1	0.2289	1.64
postech.ASR.1	0.1977	1.59
greyc.ASR.1	0.1498	1.21

B.3. Pairwise Comparison

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

· the outputs of the first system are compared against the 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 systems are ordered according to the *ranking* scores of Appendix B.2.

CT$_{EC}$

↓ 1^{st} . 2^{nd} →	**nlpr**	**ict**	**dcu**	**nict**	**mitll**	**tottori**
tch	0.0632	0.0665	0.1889	0.3147	0.3231	0.3555
	nlpr	**–0.0088**	0.0794	0.2915	0.2781	0.2911
		ict	0.1024	0.2800	0.3064	0.3587
			dcu	0.1883	0.1803	0.2342
				nict	0.0317	0.0417
					mitll	0.0108

CT$_{CE}$

↓ 1^{st} . 2^{nd} →	**tch**	**ict**	**i2r**	**mitll**	**rwth**	**dcu**	**ntt**	**nict**	**fbk**	**tottori**
nlpr	0.0878	0.2087	0.1639	0.2892	0.3587	0.3630	0.3547	0.4053	0.4084	0.6436
	tch	0.1526	0.1169	0.2022	0.2297	0.3030	0.3780	0.2779	0.3636	0.5300
		ict	0.0354	0.0036	0.0648	0.0604	0.1140	0.1847	0.2934	0.4560
			i2r	0.0402	0.1911	0.0871	0.1518	0.1721	0.2667	0.2911
				mitll	**–0.0396**	0.0901	0.0031	0.1042	0.2178	0.3406
					rwth	0.0059	0.0513	0.0031	0.2108	0.2623
						dcu	**–0.0066**	0.0440	0.1867	0.2529
							ntt	**–0.1269**	0.1377	0.1905
								nict	0.0831	0.3119
									fbk	0.1789

BT$_{CE}$

↓ 1^{st} . 2^{nd} →	**tch**	**ict**	**i2r**	**rwth**	**cmu**	**dcu**	**fbk**	**nict**	**tubitak**	**tottori**	**postech**	**greyc**	**qmul**
nlpr	0.0530	0.0912	0.1309	0.1910	0.1293	0.3660	0.3993	0.4276	0.4312	0.5135	0.5610	0.5993	0.7438
	tch	0.0140	0.1648	0.2148	0.1111	0.3086	0.3034	0.3977	0.3669	0.4779	0.5662	0.5979	0.6631
		ict	**–0.0352**	0.0977	0.1725	0.2538	0.2809	0.3623	0.2941	0.4846	0.4740	0.5789	0.5830
			i2r	0.0890	0.1623	0.2263	0.2292	0.2842	0.2030	0.5090	0.4160	0.5625	0.5520
				rwth	**–0.0351**	0.1200	0.1190	0.1748	0.2037	0.3459	0.3630	0.5393	0.5052
					cmu	0.1803	0.1034	0.2243	0.2358	0.3245	0.3571	0.4894	0.4948
						dcu	**–0.0278**	0.0660	0.0417	0.2772	0.2278	0.4409	0.4324
							fbk	0.0891	0.1018	0.1931	0.2727	0.4151	0.4100
								nict	0.0190	0.1434	0.2045	0.3922	0.3197
									tubitak	0.1223	0.1873	0.3267	0.2699
										tottori	0.0369	0.3210	0.2114
											postech	0.2117	0.1716
												greyc	**–0.2235**

BT$_{AE}$

↓ 1^{st} . 2^{nd} →	**talp**	**rwth**	**lig**	**lium**	**dcu**	**tubitak**	**qmul**	**postech**	**greyc**
mitll	0.0451	0.1273	0.2200	0.0855	0.1843	0.1222	0.3966	0.3823	0.5728
	talp	0.0483	0.0317	0.0628	0.0671	0.0879	0.3278	0.3537	0.5556
		rwth	0.0644	0.0558	0.0110	0.0742	0.3005	0.3962	0.5130
			lig	**–0.0799**	**–0.0309**	**–0.0620**	0.3529	0.2964	0.5187
				lium	0.0023	0.0112	0.2842	0.3296	0.5149
					dcu	**–0.0611**	0.3104	0.3487	0.4878
						tubitak	0.2946	0.3364	0.5090
							qmul	0.0122	0.2272
								postech	0.2181

BT$_{CS}$

↓ 1^{st} . 2^{nd} →	**fbk**	**nict**	**tubitak**	**dcu**	**talp**	**postech**	**greyc**
tch	0.2649	0.2712	0.3053	0.3658	0.3584	0.3727	0.3639
	fbk	0.0577	0.0666	0.0479	0.1047	0.2537	0.3049
		nict	0.0140	0.0680	0.0365	0.1371	0.2449
			tubitak	0.0023	0.0533	0.1136	0.2804
				dcu	0.0380	0.1448	0.2588
					talp	0.0867	0.1952
						postech	0.1919

PV$_{CS}$

↓ 1^{st} . 2^{nd} →	**fbk**	**talp**	**tubitak**	**dcu**	**nict**	**qmul**	**greyc**
tch	0.1638	0.2347	0.1902	0.2982	0.2885	0.5042	0.5892
	fbk	0.1232	0.0848	0.1882	0.1856	0.3727	0.4237
		talp	**–0.0073**	0.0117	0.0880	0.3530	0.3889
			tubitak	0.0985	0.0564	0.2950	0.3864
				dcu	0.0628	0.2797	0.3770
					nict	0.2947	0.3478
						qmul	0.0479

B.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* = $\frac{|translations\ ranked\ worse\ than\ MT^{top}| - |translations\ ranked\ better\ than\ MT^{top}|}{number\ of\ translations\ ranked\ together}$.

CT$_{EC}$

tch.ASR.1	BestRankDiff	Better	Same	Worse
nlpr.ASR.5	0.0632	0.2493	0.4380	0.3125
ict.ASR.1	0.0665	0.2405	0.4525	0.3070
dcu.ASR.1	0.1888	0.1774	0.4564	0.3662
nict.ASR.20	0.3147	0.1413	0.4027	0.4560
mitll.ASR.SLF	0.3231	0.1307	0.4155	0.4538
tottori.ASR.1	0.3555	0.1270	0.3905	0.4825

CT$_{CE}$

nlpr.ASR.5	BestRankDiff	Better	Same	Worse
tch.ASR.1	0.0906	0.2875	0.3344	0.3781
ict.ASR.1	0.1639	0.2459	0.3443	0.4098
i2r.ASR.1	0.2086	0.2466	0.2982	0.4552
mitll.ASR.SLF	0.2915	0.2177	0.2731	0.5092
ntt.ASR.1	0.3548	0.2262	0.1928	0.5810
rwth.ASR.1	0.3587	0.1762	0.2889	0.5349
dcu.ASR.1	0.3630	0.1782	0.2806	0.5412
nict.ASR.1	0.4071	0.1740	0.2449	0.5811
fbk.ASR.1	0.4083	0.1704	0.2509	0.5787
tottori.ASR.1	0.6435	0.1023	0.1519	0.7458

BT$_{AE}$

mitll.ASR.SLF	BestRankDiff	Better	Same	Worse
talp.ASR.1	0.0450	0.2575	0.4400	0.3025
lium.ASR.1	0.0855	0.2660	0.3825	0.3515
tubitak.ASR.1	0.1222	0.2219	0.4340	0.3441
rwth.ASR.1	0.1273	0.2090	0.4547	0.3363
dcu.ASR.1	0.1843	0.2260	0.3637	0.4103
lig.ASR.SLF	0.2200	0.2244	0.3312	0.4444
postech.ASR.1	0.3838	0.1534	0.3094	0.5372
qmul.ASR.1	0.3958	0.1595	0.2852	0.5553
greyc.ASR.1	0.5739	0.1034	0.2193	0.6773

BT$_{CE}$

nlpr.ASR.5	BestRankDiff	Better	Same	Worse
tch.ASR.1	0.0528	0.2781	0.3910	0.3309
ict.ASR.1	0.0912	0.2631	0.3826	0.3543
cmu.ASR.1	0.1292	0.2517	0.3674	0.3809
i2r.ASR.1	0.1304	0.2282	0.4132	0.3586
rwth.ASR.1	0.1910	0.2247	0.3596	0.4157
fbk.ASR.1	0.3660	0.1547	0.3246	0.5207
dcu.ASR.1	0.4014	0.1131	0.3724	0.5145
nict.ASR.1	0.4276	0.1448	0.2828	0.5724
tubitak.ASR.1	0.4312	0.1557	0.2574	0.5869
tottori.ASR.1	0.5154	0.1153	0.2540	0.6307
postech.ASR.1	0.5610	0.0975	0.2440	0.6585
greyc.ASR.1	0.5993	0.1198	0.1611	0.7191
qmul.ASR.1	0.7436	0.0512	0.1540	0.7948

BT$_{CS}$

tch.ASR.1	BestRankDiff	Better	Same	Worse
fbk.ASR.1	0.2649	0.1797	0.3757	0.4446
nict.ASR.1	0.2712	0.1373	0.4542	0.4085
tubitak.ASR.1	0.3053	0.1378	0.4191	0.4431
talp.ASR.1	0.3584	0.1394	0.3628	0.4978
greyc.ASR.1	0.3638	0.1610	0.3142	0.5248
dcu.ASR.1	0.3658	0.1305	0.3732	0.4963
postech.ASR.1	0.3727	0.1501	0.3271	0.5228

PV$_{CS}$

tch.ASR.1	BestRankDiff	Better	Same	Worse
fbk.ASR.1	0.1638	0.2267	0.3828	0.3905
tubitak.ASR.1	0.1901	0.2140	0.3819	0.4041
talp.ASR.1	0.2346	0.2048	0.3558	0.4394
nict.ASR.1	0.2885	0.2000	0.3115	0.4885
dcu.ASR.1	0.2987	0.1817	0.3379	0.4804
qmul.ASR.1	0.5042	0.0892	0.3174	0.5934
greyc.ASR.1	0.5892	0.0728	0.2652	0.6620

Appendix C. Automatic Evaluation

official evaluation : case-sensitive, with punctuations tokenized
additional evaluation : case-insensitive, with punctuations removed

· the systems were ranked according to the average score of the BLEU and METEOR metric results of the primary runs for the ASR Output data track.
· omitted lines between scores indicate non-significant differences in performance between the respective MT engines according to the *bootStrap* method [17].
· the best score of each metric is marked with *boldface*.

ASR Output							**Correct Recognition Result**					
official evaluation			*additional evaluation*			**CT**$_{EC}$	*official evaluation*			*additional evaluation*		
(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR		(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR
0.6173	**0.4890**	**0.7456**	**0.6082**	**0.4795**	**0.7369**	**tch**	**0.7039**	**0.5919**	**0.8159**	**0.6947**	**0.5814**	**0.8080**
0.5749	0.4329	0.7169	0.5680	0.4277	0.7083	**ict**	0.6400	0.5039	0.7760	0.6374	0.5042	0.7705
0.5708	0.4242	0.7173	0.5663	0.4228	0.7098	**nlpr**	0.6476	0.5070	0.7882	0.6374	0.5000	0.7747
0.5599	0.4102	0.7096	0.5536	0.4057	0.7014	**dcu**	0.6164	0.4683	0.7645	0.6140	0.4682	0.7597
0.5366	0.3783	0.6949	0.5387	0.3762	0.7011	**nict**	0.5931	0.4260	0.7602	0.5788	0.4187	0.7389
0.5166	0.3529	0.6803	0.5003	0.3339	0.6667	**tottori**	0.5691	0.4005	0.7377	0.5584	0.3869	0.7299
0.4930	0.3296	0.6564	0.5275	0.3564	0.6986	**mitll**	0.5379	0.3777	0.6980	0.5773	0.4092	0.7454

ASR Output							**Correct Recognition Result**					
official evaluation			*additional evaluation*			**CT**$_{CE}$	*official evaluation*			*additional evaluation*		
(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR		(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR
0.5061	**0.3982**	0.6140	**0.5084**	**0.4118**	**0.6050**	**nlpr**	**0.5774**	**0.4894**	**0.6653**	**0.5768**	**0.5010**	**0.6525**
0.5060	0.3969	**0.6151**	0.5011	0.4081	0.5940	**tch**	0.5694	0.4779	0.6609	0.5627	0.4834	0.6420
0.4882	0.3788	0.5976	0.4811	0.3917	0.5704	**i2r**	0.5457	0.4529	0.6384	0.5387	0.4614	0.6159
0.4681	0.3632	0.5729	0.4735	0.3869	0.5601	**ict**	0.5194	0.4254	0.6133	0.5260	0.4547	0.5972
0.4502	0.3388	0.5616	0.4636	0.3793	0.5479	**rwth**	0.4832	0.3883	0.5780	0.5017	0.4388	0.5646
0.4396	0.3247	0.5545	0.4600	0.3742	0.5458	**mitll**	0.4745	0.3663	0.5827	0.4958	0.4182	0.5733
0.4341	0.2963	0.5718	0.4303	0.3086	0.5520	**ntt**	0.4867	0.3617	0.6116	0.4874	0.3791	0.5957
0.4173	0.2934	0.5412	0.4134	0.3088	0.5180	**dcu**	0.4699	0.3576	0.5822	0.4679	0.3788	0.5570
0.3748	0.2386	0.5110	0.3822	0.2643	0.5000	**nict**	0.4302	0.3009	0.5594	0.4394	0.3318	0.5470
0.3580	0.2299	0.4861	0.3609	0.2482	0.4736	**fbk**	0.3961	0.2736	0.5185	0.4092	0.3081	0.5103
0.3406	0.2217	0.4594	0.3758	0.2526	0.4990	**tottori**	0.3747	0.2658	0.4836	0.4144	0.2986	0.5302

ASR Output							**Correct Recognition Result**					
official evaluation			*additional evaluation*			**BT**$_{CE}$	*official evaluation*			*additional evaluation*		
(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR		(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR
0.5347	**0.4494**	**0.6200**	**0.5888**	0.5155	**0.6622**	**tch**	**0.5361**	**0.4711**	0.**6012**	0.5880	0.5312	**0.6449**
0.5226	0.4459	0.5993	0.5665	0.4973	0.6357	**ict**	0.5188	0.4529	0.5848	0.5623	0.5028	0.6219
0.5219	0.4437	0.6001	0.5562	0.4829	0.6295	**cmu**	0.5227	0.4588	0.5867	0.5605	0.5020	0.6191
0.5148	0.4242	0.6054	0.5882	**0.5195**	0.6569	**nlpr**	0.5170	0.4430	0.5910	**0.5900**	**0.5381**	0.6419
0.5130	0.4338	0.5922	0.5624	0.4925	0.6324	**i2r**	0.5045	0.4444	0.5647	0.5577	0.5048	0.6107
0.4873	0.4016	0.5730	0.5355	0.4575	0.6135	**rwth**	0.4856	0.4190	0.5523	0.5313	0.4662	0.5965
0.4639	0.3612	0.5667	0.4997	0.4000	0.5994	**fbk**	0.4644	0.3712	0.5577	0.5037	0.4128	0.5946
0.4408	0.3397	0.5419	0.4879	0.3950	0.5809	**dcu**	0.4299	0.3380	0.5218	0.4751	0.3901	0.5601
0.4295	0.3146	0.5445	0.4646	0.3582	0.5711	**tubitak**	0.4350	0.3366	0.5335	0.4690	0.3782	0.5603
0.4180	0.3127	0.5233	0.4593	0.3518	0.5668	**nict**	0.4223	0.3356	0.5091	0.4637	0.3769	0.5505
0.3901	0.2941	0.4861	0.4240	0.3288	0.5192	**tottori**	0.4224	0.3230	0.5218	0.4578	0.3576	0.5580
0.3656	0.2551	0.4762	0.3880	0.2787	0.4974	**postech**	0.3932	0.2781	0.5083	0.4195	0.3051	0.5339
0.3160	0.2016	0.4305	0.3290	0.2124	0.4456	**greyc**	0.3226	0.2324	0.4128	0.3377	0.2464	0.4291
0.2638	0.1214	0.4062	0.3306	0.2086	0.4527	**qmul**	0.3175	0.1845	0.4506	0.3739	0.2588	0.4891

ASR Output						$\mathbf{BT}_{AE}$	Correct Recognition Result					
official evaluation			*additional evaluation*				*official evaluation*			*additional evaluation*		
(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR		(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR
0.4292	**0.3050**	**0.5534**	**0.5784**	**0.5217**	**0.6350**	**mitll**	**0.4739**	**0.3425**	**0.6053**	**0.6403**	**0.5866**	**0.6944**
0.4020	0.2745	0.5295	0.5343	0.4642	0.6043	**rwth**	0.4567	0.3354	0.5780	0.6119	0.5627	0.6610
0.3952	0.2562	0.5342	0.5041	0.3982	0.6100	**lium**	0.4479	0.3181	0.5776	0.5874	0.5147	0.6609
0.3915	0.2563	0.5267	0.5280	0.4505	0.6054	**talp**	0.4439	0.3131	0.5747	0.6015	0.5420	0.6617
0.3904	0.2519	0.5289	0.5115	0.4189	0.6040	**tubitak**	0.4227	0.2778	0.5675	0.5602	0.4723	0.6487
0.3881	0.2545	0.5216	0.5020	0.4080	0.5959	**lig**	0.4320	0.2838	0.5802	0.5686	0.4741	0.6639
0.3798	0.2403	0.5192	0.4993	0.4049	0.5937	**dcu**	0.4334	0.2923	0.5744	0.5710	0.4860	0.6567
0.3360	0.1935	0.4785	0.4312	0.3152	0.5472	**postech**	0.3950	0.2493	0.5407	0.5194	0.4198	0.6199
0.3126	0.1613	0.4639	0.4151	0.3001	0.5300	**qmul**	0.3379	0.1904	0.4853	0.4479	0.3408	0.5558
0.2467	0.1366	0.3568	0.3197	0.2304	0.4090	**greyc**	0.2698	0.1532	0.3863	0.3511	0.2601	0.4420

ASR Output						$\mathbf{BT}_{CS}$	Correct Recognition Result					
official evaluation			*additional evaluation*				*official evaluation*			*additional evaluation*		
(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR		(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR
0.3109	**0.3052**	**0.3165**	**0.2780**	**0.2699**	**0.2860**	**tch**	**0.3427**	**0.3457**	**0.3396**	**0.3142**	**0.3206**	**0.3077**
0.2619	0.2424	0.2813	0.2461	0.2256	0.2666	**fbk**	0.2966	0.2960	0.2972	0.2705	0.2643	0.2766
0.2594	0.2440	0.2747	0.2459	0.2332	0.2585	**tubitak**	0.2817	0.2662	0.2971	0.2661	0.2549	0.2773
0.2521	0.2389	0.2653	0.2303	0.2190	0.2416	**dcu**	0.2776	0.2710	0.2841	0.2518	0.2447	0.2588
0.2500	0.2331	0.2669	0.2397	0.2283	0.2511	**nict**	0.2763	0.2641	0.2884	0.2653	0.2566	0.2739
0.2359	0.2214	0.2504	0.2209	0.2102	0.2316	**talp**	0.2635	0.2565	0.2704	0.2476	0.2425	0.2526
0.2223	0.2010	0.2436	0.2142	0.1823	0.2460	**postech**	0.2569	0.2572	0.2566	0.2552	0.2518	0.2585
0.2021	0.1891	0.2150	0.2097	0.2039	0.2155	**greyc**	0.2103	0.1970	0.2235	0.2181	0.2127	0.2234

ASR Output						$\mathbf{PV}_{CS}$	Correct Recognition Result					
official evaluation			*additional evaluation*				*official evaluation*			*additional evaluation*		
(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR		(B+M)/2	BLEU	METEOR	(B+M)/2	BLEU	METEOR
0.3452	**0.3543**	**0.3360**	**0.3253**	**0.3448**	**0.3057**	**tch**	**0.3889**	**0.4042**	**0.3736**	**0.3756**	**0.4035**	**0.3477**
0.3212	0.3251	0.3172	0.2992	0.3079	0.2904	**fbk**	0.3758	0.3941	0.3574	0.3496	0.3652	0.3339
0.3144	0.3413	0.2875	0.2853	0.3088	0.2618	**talp**	0.3494	0.3809	0.3178	0.3225	0.3510	0.2939
0.3103	0.3281	0.2925	0.3005	0.3275	0.2735	**nict**	0.3440	0.3711	0.3168	0.3397	0.3782	0.3012
0.2863	0.2815	0.2910	0.2578	0.2581	0.2574	**tubitak**	0.3170	0.3188	0.3152	0.2886	0.2968	0.2804
0.2844	0.2847	0.2840	0.2685	0.2659	0.2710	**dcu**	0.3165	0.3242	0.3087	0.2979	0.3019	0.2938
0.1755	0.1505	0.2004	0.1799	0.1604	0.1993	**greyc**	0.1813	0.1580	0.2046	0.1825	0.1696	0.1954
0.1469	0.1159	0.1778	0.1666	0.1423	0.1909	**qmul**	0.0724	0.0287	0.1161	0.0921	0.0444	0.1397

Appendix D. Correlation between Evaluation Metrics

· 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 parantheses behind each data track label indicates the number of ranked MT systems
· the automatic evaluation metrics that correlate best with the respective human assessments are marked in boldface

D.1. Fluency/Adequacy vs. Automatic Evaluation

CT_{EC} (4)	(B+M)/2	BLEU	METEOR
(F+A)/2	**1.0000**	**1.0000**	0.8000
Fluency	**1.0000**	**1.0000**	0.8000
Adequacy	0.8000	0.8000	**1.0000**

CT_{CE} (4)	(B+M)/2	BLEU	METEOR
(F+A)/2	**1.0000**	**1.0000**	**1.0000**
Fluency	**1.0000**	**1.0000**	**1.0000**
Adequacy	**1.0000**	**1.0000**	**1.0000**

BT_{CS} (4)	(B+M)/2	BLEU	METEOR
(F+A)/2	0.8000	0.8000	8.0000
Fluency	0.8000	0.8000	8.0000
Adequacy	**1.0000**	**1.0000**	**1.0000**

BT_{CE} (4)	(B+M)/2	BLEU	METEOR
(F+A)/2	0.8000	0.8000	8.0000
Fluency	**1.0000**	**1.0000**	**1.0000**
Adequacy	0.8000	0.8000	8.0000

PV_{CS} (4)	(B+M)/2	BLEU	METEOR
(F+A)/2	0.8000	0.8000	8.0000
Fluency	0.0000	0.0000	0.0000
Adequacy	**1.0000**	**1.0000**	**1.0000**

BT_{AE} (4)	(B+M)/2	BLEU	METEOR
(F+A)/2	**1.0000**	**1.0000**	**1.0000**
Fluency	**1.0000**	**1.0000**	**1.0000**
Adequacy	**1.0000**	**1.0000**	**1.0000**

D.2. Ranking vs. Automatic Evaluation

CT_{EC} (7)	(B+M)/2	BLEU	METEOR
Ranking	0.1071	0.1071	**0.4286**
NormRank	**0.4286**	**0.4286**	0.1071
BestRankDiff	0.1071	0.1071	**0.4286**

CT_{CE} (11)	(B+M)/2	BLEU	METEOR
Ranking	0.4000	0.4000	**0.5727**
NormRank	0.4909	0.4909	**0.6636**
BestRankDiff	**0.8727**	**0.8727**	0.7909

BT_{CS} (8)	(B+M)/2	BLEU	METEOR
Ranking	0.1190	-0.1667	**0.6190**
NormRank	0.1190	-0.1667	**0.6190**
BestRankDiff	-0.1667	**-0.4524**	-0.0238

BT_{CE} (14)	(B+M)/2	BLEU	METEOR
Ranking	**0.3736**	0.3571	0.1758
NormRank	**0.6429**	0.6264	-0.0165
BestRankDiff	**0.5769**	0.5604	0.3132

PV_{CS} (8)	(B+M)/2	BLEU	METEOR
Ranking	0.0238	0.0476	**0.4524**
NormRank	0.1190	0.1429	**0.2619**
BestRankDiff	**0.6190**	-0.3571	**0.6190**

BT_{AE} (10)	(B+M)/2	BLEU	METEOR
Ranking	-0.0667	-0.0182	**-0.1273**
NormRank	0.6121	0.2364	**0.6364**
BestRankDiff	**0.9152**	-0.0424	0.6970