Overview of the IWSLT 2006 Evaluation Campaign

Michael Paul

ATR Spoken Language Communication Research Labs
Hikaridai 2-2-2, Keihanna Science City, 619-0288 Kyoto
Michael.Paul@atr.jp

Abstract

This paper gives an overview of the evaluation campaign results of the *International Workshop on Spoken Language Translation* (IWSLT) 2006[1]. In this workshop, we focused on the translation of spontaneous speech. The translation directions were Arabic, Chinese, Italian, or Japanese into English. In total, 21 translation systems from 19 research groups participated in this year's evaluation campaign. Both automatic and subjective evaluations were carried out in order to investigate the impact of spontaneity aspects 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 recognition errors.

1. Introduction

The *International Workshop on Spoken Language Translation* (IWSLT) is an evaluation campaign organized by the *Consortium for Speech Translation Advanced Research* (C-STAR)[2], that provides a common framework to compare and improve current state-of-the-art speech-to-speech translation technologies. Previous IWSLT workshops focused on the establishment of evaluation metrics for multilingual speech-to-speech translation [1] and the translation of automatic speech recognition results from read-speech input [2].

The focus of this year's IWSLT was the translation of spontaneous-speech input. The evaluation campaign was carried out using a multilingual spoken language corpus including Arabic, Chinese, Italian, Japanese, and English sentences from the travel domain. The input to the machine translation (MT) engines was either the output of an automatic speech recognition (*ASR*) system applied to spontaneous-speech and read-speech input or the correct recognition result (*CRR*). The translation was carried out from Arabic, Chinese, Italian, or Japanese into English.

Participants were supplied with in-domain resources, but were free to use additional resources as well. Depending on the amount of in-domain training data, two different data tracks (OPEN, CSTAR) were distinguished. In total, 21 MT systems from 19 research groups participated in this year's evaluation campaign. A total of 73 MT engines were built to cover different combinations of language pairs and data tracks. The translation quality of all official run submissions was evaluated using automatic evaluation metrics. In addition, human assessments were carried out for the most popular track, i.e., the translation of Chinese ASR output into English. Based on the evaluation results, the impact of the spontaneity aspects of speech on the ASR and MT systems performance as well as the robustness of state-of-the-art MT systems towards speech recognition errors were investigated.

[1] http://www.slc.atr.jp/IWSLT2006
[2] http://www.c-star.org/

2. IWSLT 2006 Evaluation Campaign

2.1. IWSLT 2006 Spoken Language Corpus

The IWSLT 2006 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 going abroad [3]. Parts of this corpus were already used in previous IWSLT evaluation campaigns [1, 2]. In addition to 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.

The evaluation data set of IWSLT 2006 consisted of spontaneous answers to questions in the tourism domain. This "*Challenge Task 2006*" differed greatly from the translation tasks of previous workshops. In addition to the spontaneous speech data, read-speech recordings of the cleaned transcripts were also used for evaluation purposes. ASR engines provided by the C-STAR partners were applied to the speech input and produced word lattices from which NBEST/1BEST lists were extracted automatically using publicly available tools. 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), Italian (I), Japanese (J), and English (E) subsets of the BTEC* corpus were used. The participants were supplied with a training corpus of 40K sentence pairs for CE/JE, and 20K sentence pairs for AE/IE and three development data sets (*dev1*, *dev2*, *dev3*; 500 sentences each) consisting of the evaluation data sets of previous IWSLT evaluation campaigns including up to 16 English reference translations for evaluation purposes.

Table 1: The IWSLT 2006 spoken language corpus

type	language	sentence count total	sentence count unique	avg. length	word tokens	word types
training	C/E	39,953	37,559 / 39,633	8.6 / 9.2	342,362 / 367,265	11,174 / 7,225
	J/E	39,953	37,173 / 39,633	10.0 / 9.2	398,498 / 367,265	11,407 / 7,225
	A/E	19,972	19,777 / 19,880	7.7 / 9.2	154,279 / 183,673	18,292 / 5,465
	I/E	19,972	19,641 / 19,880	8.6 / 9.2	171,764 / 183,673	10,085 / 5,465
development	C/E_{16}	1,512	1,458 / 20,585	7.0 / 8.2	10,570 / 198,872	1,882 / 2,882
dev1	J/E_{16}	1,512	1,462 / 20,585	8.2 / 8.2	12,416 / 198,872	1.686 / 2,882
dev2	A/E_{16}	1,512	1,455 / 20,585	6.3 / 8.2	9,466 / 198,872	2,698 / 2,882
dev3	I/E_{16}	1,512	1,450 / 20,585	6.8 / 8.2	10,318 / 198,872	2.014 / 2,882
development	C/E_7	489	487 / 3,379	11.7 / 13.4	5,702 / 45,956	1,138 / 1,392
dev4	J/E_7	489	488 / 3,379	14.0 / 13.4	6,836 / 45,956	1,084 / 1,392
	A/E_7	489	486 / 3,379	9.8 / 13.4	4,772 / 45,956	1,622 / 1,392
	I/E_7	489	485 / 3,379	11.8 / 13.4	5,770 / 45,956	1,236 / 1,392
evaluation	C/E_7	500	499 / 3,472	12.1 / 14.4	6,050 / 50,589	1,329 / 1,575
eval	J/E_7	500	499 / 3,472	14.8 / 14.4	7,420 / 50,589	1,238 / 1,575
	A/E_7	500	499 / 3,472	10.4 / 14.4	5,213 / 50,589	1,952 / 1,575
	I/E_7	500	499 / 3,472	13.4 / 14.4	6,699 / 50,589	1,471 / 1,575

E_N : 'N' English reference translations provided for evaluation purposes

Details of the IWSLT 2006 spoken language corpus are given in Table 1. The *total sentence counts* show the number of bilingual sentence pairs and the *unique sentence counts* refer to the number of unique monolingual sentences. The *average length* 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. 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.

2.1.2. Challenge Task 2006

In order to obtain speech input with a certain level of spontaneity, question/answer conversations between Chinese speakers were recorded by the C-STAR partners. In the preparation phase, around 1000 questions were extracted manually from the original BTEC* corpus, avoiding redundancy and an attempt was made to maximize the diversity of the topics addressed. In addition, *answer keys*, i.e. short phrases providing hints on the answer contents, were added to each question.

For recording, the questions were split into 20 subsets and pairs of native Chinese speakers[3] were asked to carry-out a "one-turn" role play. A brief *scene description* (outline of the role-play) was given to both speakers. Speaker *SQ* obtained a list of questions and asked one question after the other. Speaker *SA* obtained a list of answer keys and answered to each question using the following guidelines:

- answer in a natural way based on the answer keys
- avoid direct recitation of answer keys
- in case of Yes/No-questions, try to explain the reason

[3] 20 speakers, gender: 10x female/male each, age: 21 – 32 (avg: 25.7)

Table 2: Data preparation of Challenge Task 2006

scene:	[airplane] passenger asks flight attendance for help
question:	Okay. Where can I put my luggage? Is it here okay?
key:	(not here, overhead compartement)
answer:	"*sorry you'd better put it in the overhead compartement*"

scene:	[airport] asking directions
question:	Take me to this address. How long will it take?
key:	(depending on traffic condition, around 20 minutes)
answer:	"*it's hard to say it depends on the traffic conditions it should take only twenty minutes or so if there's no traffic jam*"

Examples of questions, answer keys, and recorded answers are given in Table 2. The obtained *Challenge Task 2006* data sets were split into two subsets: *dev4* (489 sentences, development corpus) and *eval* (500 sentences, evaluation corpus). The difficulty of this year's evaluation data set is illustrated in Table 3. It lists the target language perplexity of all translation tasks according to the supplied resources of IWSLT 2006. Compared to last year's evaluation data sets, the language perplexities of *dev4* and *eval* were three times higher.

Table 3: English language perplexity of IWSLT 2006 translation tasks

translation type	task	training data 40K (CE/JE)	training data 20K (AE/IE)
development	dev1	27.5	32.6
	dev2	31.4	36.7
	dev3	32.9	38.8
	dev4	85.6	98.3
evaluation	eval	105.9	113.9

In addition to the Chinese spontaneous-speech recordings, read-speech recordings of the *Challenge Task 2006* were produced for all source languages. The cleaned transcriptions of

the Chinese spontaneous-speech recordings were translated into English, Japanese, Arabic, and Italian by human translators. For English, two native speakers produced three alternative translations each resulting in a total of seven reference translations for the *dev4* and *eval* data set, respectively.

The source language texts were read aloud by 20 native speakers of the respective source language[4] and recognition results were obtained using ASR engines provided by the C-STAR partners.

Table 4 summarizes the out-of-vocabulary (*OOV*) rates of the obtained data sets. The OOV rates are listed for all source languages and input conditions (CRR, 1BEST, NBEST) and for the English reference translations using the 20K/40K training corpus. 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.

Table 4: OOV rates of IWSLT 2006 spoken language corpus

type	language	*OOV rates (%)*		
		CRR	1BEST	NBEST
dev4	C_s	2.0	2.0	2.3
	C_r	2.0	1.7	2.3
	J	1.7	1.3	1.3
	A	13.1	14.2	15.4
	I	3.6	2.1	2.2
	E_7	1.7 / 1.4		
eval	C_s	2.6	2.1	2.4
	C_r	2.6	2.4	2.5
	J	2.2	1.6	2.3
	A	14.3	16.0	17.1
	I	4.3	2.5	2.6
	E_7	2.7 / 1.9		

C_s: spontaneous speech, C_r: read speech

The lowest OOV rates for the CRR data are found for Japanese and Chinese (1.2-2.6%). The figures for Italian are twice as high. However, very large OOV rates of 13-17% are obtained for Arabic which are caused mainly by word segmentation issues (*prefix/postfix* attachment) and spelling variations in Arabic. The spontaneous speech data sets have slightly lower OOV rates than the read speech data.

The recognition accuracies of the utilized ASR engines for the *Challenge Task 2006* data sets are summarized in Table 5. 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[5] extracted from each lattice. Besides for Italian, the *word accuracies* of the read-speech recordings ranged between 82%-90% (lattice) and 74%-85% (1BEST), where the percentages of correctly recognized sentences (*sentence accuracy*) ranged between 30%-50% (lattice) and 20%-40% (1BEST). However, a large difference can be seen between the different source languages. The lattice accuracies of Chinese were 5%-8% lower than those obtained for Japanese and Arabic. For Chinese and Arabic, a large drop in recognition performance was seen when comparing lattice and 1BEST accuracies.

Table 5: Recognition accuracy of IWSLT 2006 spoken language corpus

type	language	*word* (%)		*sentence* (%)	
		lattice	1BEST	lattice	1BEST
dev4	C_s	76.95	67.38	22.49	18.00
	C_r	83.24	74.78	30.47	23.31
	J	88.95	84.35	50.31	40.08
	A	86.71	73.36	41.10	19.84
	I	76.02	74.10	15.34	13.91
eval	C_s	79.08	68.11	22.80	16.60
	C_r	82.07	73.64	28.40	22.80
	J	90.48	85.14	52.60	38.00
	A	88.20	73.88	41.60	16.60
	I	72.90	70.88	5.40	4.60

C_s: spontaneous speech, C_r: read speech

Concerning different speech types, a drop in recognition performance of 3%-6% in word accuracy and 5%-8% in sentence accuracy was seen for the spontaneous-speech data compared to the read-speech results.

2.2. Translation Input Conditions

In order to investigate the effects of recognition errors on the MT performance, the participants were asked to translate two types of input using the same MT engine:

1. *speech input* (wave forms) or *ASR output* (lattices, NBEST/1BEST lists)
2. *correct recognition results* (plain text)

The translation of the correct recognition results was mandatory for all participants. For the ASR output, most of the participants applied their MT engines to the 1BEST recognition results. Three research groups reported a gain in translation performance by translating NBEST lists and combining the obtained translation hypotheses. In addition, three groups exploited the ASR lattices directly to obtain its translation results. Concerning the *speech input*, the participants were allowed to use their own ASR engine, however none of the participants took this opportunity.

2.3. Data Track Conditions

For training purpose, the 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[6] as well.

The past IWSLT workshop results have shown that the amount of BTEC* sentence pairs used for training has a dominant effect on the performance of the MT systems on the given task. However, only C-STAR partners have access to

[4] An exception was Arabic with only one native speaker.

[5] We used the *lattice-toolkit* of the *SRI Language Modeling Toolkit* (http://www.speech.sri.com/projects/srilm) to automatically extract NBEST lists from ASR lattices.

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

the full BTEC* corpus[7] consisting of 172K sentence pairs. In order to allow a fair comparison between the systems, the following two data tracks were distinguished:

- *Open Data Track*:

  Any resources, except for the full BTEC* corpus and proprietary data, can be used as the training data for the MT engines. Concerning the BTEC* and proprietary data, only the *Supplied Resources* (see Section 2.1.1) were allowed to be used for training purposes.

- *C-STAR Data Track*:

  Any resources (including the full BTEC* corpus and proprietary data) can be used as the training data of MT engines.

2.4. Run Submissions

The supplied resources of IWSLT 2006 were released three months ahead of the official run submissions. The organizers also set-up an online evaluation server that could be used to evaluate system performance on the provided development data sets using automatic scoring metrics (see Section 2.5.1). The official run submission period was limited to three days during which the automatic scoring result feedback to the participant via email was made unavailable in order to avoid any system tuning towards the *eval* data. The schedule of the evaluation campaign is summarized in Table 6.

Table 6: IWSLT 2006 evaluation campaign schedule

Event	Date
Training Corpus Release	May 12, 2006
Development Corpus Release	June 30, 2006
Evaluation Corpus Release	August 7, 2006
Result Submission Due	August 9, 2006

In total, 19 research groups took part in this year's evaluation campaign and two groups registered multiple translation systems. Information on the organisations and the utilized translation systems is summarized in Appendix A. Most participants used statistical machine translation (SMT) systems. In addition, example-based MT (EBMT) systems, rule-based MT (RBMT) systems and hybrid approaches combining multiple MT engines were also exploited. Five of the MT systems were applied to all input conditions. Each participant submitted at least one run. In total, 73 official and 83 contrastive runs were submitted for the *eval*. The distribution of run submissions for the respective data track/input condition is summarized in Table 7.

After the official run submission period, the participants still had access to the evaluation server and in order to do additional experiments.

2.5. Evaluation Specifications

In order to deliver more usable translations, both for reading and for listening, and to make the IWSLT evaluation campaign results more comparable to outcomes of other MT evaluation workshops like those organized by NIST[8] or TC-STAR[9], the *official evaluation* specifications of this year's IWSLT were defined as:

Table 7: Distribution of run submissions

Translation Input Condition	*Open Data Track* Research Groups	*Open Data Track* Official (Contrastive)	*CSTAR Data Track* Research Groups	*CSTAR Data Track* Official (Contrastive)
CE spont	12	12 (11)	2	3 (3)
read	12	14 (17)	2	3 (3)
JE read	12	14 (14)	2	2 (3)
AE read	9	11 (14)	1	1 (1)
IE read	10	12 (14)	1	1 (3)
TOTAL	19	63 (70)	2	10 (13)

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

However, in order to be able to compare this year's IWSLT results to the outcomes of previous IWSLT workshops, the evaluation specifications of last year were also applied as an *additional evaluation*:

- case-insensitive (lower-case only)
- no punctuation marks (remove '.' ',' '?' '!' '"')
- no word compounds (replace hyphens '-' with space)

The focus of this year's evaluation campaign was the translation of speech data. Therefore, all input data files (see Section 2.2) were case-insensitive and without punctuation information. However, true-case and punctuation information was provided 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.

2.5.1. Automatic Evaluation

The automatic evaluation of run submissions was carried out using an online evaluation server. The participants had to upload two translation files (see Section 2.2). Text preprocessing was carried out automatically according to the evaluation specification described above. For the *official evaluation*, an English tokenizer tool, that was made available to all participants, was applied. For the *additional evaluation* all punctuation marks were removed and the text was converted to lower-case.

For development purposes, the participants had access to the online evaluation server of the *dev4* data set three weeks

[7] http://cstar.atr.jp/cstar-corpus

[8] http://www.nist.gov/speech/tests/{mt|gale}

[9] http://www.elda.org/en/proj/tcstar-wp4/index.htm

[10] http://www.slc.atr.jp/IWSLT2006/downloads/case+punc_tool_using_SRILM.instructions.txt

before the *eval* run submission period. For the official evaluation results[11] of the IWSLT 2006 workshop, we utilized the following three metrics:

Table 8: Automatic evaluation metrics

BLEU4:	the geometric mean of n-gram precision by the system output with respect to reference translations. Scores range between 0 (worst) and 1 (best) [6]
NIST:	a variant of BLEU4 using the arithmetic mean of weighted n-gram precision values. Scores are positive with 0 being the worst possible [7]
METEOR:	calculates unigram overlaps between a translations and reference texts using various levels of matches (*exact*, *stem*, *synonym*) are taken into account. Scores range between 0 (worst) and 1 (best) [8]

2.5.2. 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 English. For *adequacy*, the evaluator was presented with the source language input as well as a "gold standard" translation and has to judge how much of the information from the original translation is expressed in the translation. The *fluency* and *adequacy* judgments consist of one of the grades listed in Table 9.

Table 9: Human assessment

Fluency	
4	Flawless English
3	Good English
2	Non-native English
1	Disfluent English
0	Incomprehensible

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

The subjective evaluation was carried out only for the most popular track, i.e., the translation of Chinese ASR output into English. In order to compare different translation input conditions (*CE spont*, *CE read*, *CE CRR*), 7 MT systems that were applied to all input conditions were selected according to the automatic scoring results. In total, 21 run submissions were evaluated by humans.

The human assessment was limited to the translation outputs of 400 input sentences selected randomly from the *eval* data. In order to reduce the costs further, all translation results were *pooled*, i.e., in case of identical translations of the same source sentence by multiple MT engines, the translation was graded only once, and the respective rank was assigned to all MT engines with the same output.

Each translation of a single MT engine was evaluated by three judges where each system score is calculated as the *median* of the assigned grades. For fluency, only native speakers of English were used. The adequcay evaluation was carried out by native speakers and non-native speakers with sufficient knowledge of English. In total, 12 English native speakers and 11 non-native speakers were involved in this year's evaluation task. A total of 38,198 grading operations were performed.

[11] In addition to the official evaluation metrics used for IWSLT 2006, the *word error rate* (WER) [4] and *position-independent WER* (PER) [5] were also calculated by the evaluation server and provided to the participants for the analysis of their systems.

3. Evaluation Results

The evaluation results of the IWSLT 2006 workshop are summarized in Appendix B (*human assessment*) and Appendix C (*automatic evaluation*). For each translation condition/evaluation metric, the best score is marked in bold-face.

Based on the obtained evaluation results, the respective MT engines were ranked. In order to decide whether the translation output of one MT engine is significantly better than another one, we used the *bootStrap*[12] method that (1) performs a random sampling with replacement from the *eval* 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[13], and (4) applies the *Student's t-test* at a significance level of 95% confidence to test whether the score differences are significant [9]. In this paper, we omit a horizontal line between two MT engines in the MT engine ranking tables, if the system performances *do not* differ significantly according to the *bootStrap* method.

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. The grader consistency is investigated in Section 3.1.1.

The subjective evaluation results of the MT outputs for the CE translation tasks are summarized in Appendix B.1. where the MT engines are in descending order with respect to the *adequacy* score. Some general findings are given in Section 3.1.2.

3.1.1. Grader Consistency

In order to investigate the degree of grading consistency between the human evaluators, the *Kappa statistics* for the agreement of *fluency* and *adequacy* ratings were calculated. Only low agreement levels (*fluency*: 0.24, *adequacy*: 0.31) were obtained for both metrics. In addition, the average grading difference between two graders was 0.80 points for *fluency* and 0.68 points for *adequacy*.

In order to check the self-consistency of subjective evaluations, each grader had to evaluate a set of 100 sentences a second time. Based on these grades, the average difference between the first and second grade (*fluency*: 0.50, *adequacy*: 0.40) and the probability that the grader will assign a different grade (*fluency*: 0.44, *adequacy*: 0.39) were calculated.

The grader consistency figures are slightly worse than those obtained in the previous IWSLT workshops, which

[12] http://projectile.is.cs.cmu.edu/research/public/tools/bootStrap/tutorial.htm

[13] 2000 iterations were used for the analysis of the IWSLT 2006 results

might be partly caused by the lower translation quality of the MT outputs. In order to minimize the impact of grader inconsistencies, the *median* of the three assigned grades was selected as the final judgment for each sentence.

3.1.2. System Performance

The highest *fluency* and *adequacy* scores were obtained for the translation of the correct recognition results (1.67 for *adequacy*, 1.59 for *fluency*). Speech recognition errors for *read speech* input led to a drop in MT performance of 0.33-0.47 points for *adequacy* and 0.12-0.35 points for *fluency*. This indicates that recognition errors affected primarily the information content of the translation output. Moreover, only minor degradations in both metrics can be seen when comparing *read-speech* with *spontaneous speech* results.

However, the degree of degradation varies between MT engines. The smallest drop in performance can be seen for the JHU_WS06 system [16]. Although it does not achieve the best performance on the CRR task, it seems to be quite robust against recognition errors. One reason might be that it does not restrict its input to 1BEST ASR outputs, instead it uses information from the word lattice to overcome recognition problems. In contrast, the MIT-LL_AFRL system [18] achieved the highest *adequacy* score on the CRR task, but performance became worse on the *CE spont* task. Curiously, its *fluency* score for *spontaneous speech* is higher than its *read speech* score.

Such system specific phenomena lead to quite different MT engine rankings depending on which metric is used (see Appendix B.2.). In order to get an idea on the "overall" performance of each system, MT engine rankings of multiple metrics are combined by simply calculating the average rank for each MT engine. If no significant difference between two MT engine scores could be determined, the same rank was assigned to both MT engines. Table 10 summarizes the MT engine rankings when combining *fluency* and *adequacy* results. An omitted horizontal line between MT engines indicates the systems were not significantly different.

Table 10: Combination of Subjective Evaluation Rankings

CE spont	*CE read*	*CRR*
JHU_WS06	JHU_WS06	MIT-LL_AFRL
RWTH	MIT-LL_AFRL	RWTH
NTT	RWTH	NTT
MIT-LL_AFRL	NTT	JHU_WS06
UKACMU_SMT	NiCT-ATR	NiCT-ATR
NiCT-ATR	UKACMU_SMT	UKACMU_SMT

3.2. Automatic Evaluation Results

The automatic evaluation results of all MT engines using the official as well as the additional evaluation specifications are listed in Appendix C.1. The MT systems are ordered according to the BLEU4 metrics. The correct recognition results of all MT systems that were applied to the *CE spont* as well as the *CE read* translation task are identical and they are listed redundantly for the convenience of the reader.

The MT engine rankings are summarized in Appendix C.2. Similar to the subjective evaluation results, the rankings vary with respect to the utilized automatic evaluation metrics. In order to get an idea of how closely the respective metrics are related, the *Pearson correlation coefficients* were calculated for all automatic evaluation metric combinations. For each translation direction, we used the official run submissions for both (ASR, CRR) input conditions, i.e., 24 runs for *CE spont*, 28 runs for *CE read*, 28 for *JE*, 22 runs for *AE*, and 24 runs for *IE*, respectively. The correlation coefficients are given in Table 11. On the left hand side of the table, the BLEU4 metric is compared to the NIST and METEOR metric. The correlation between NIST and METEOR is given on the right hand side.

Table 11: Correlation between Automatic Evaluation Metrics

BLEU4	NIST	METEOR
CE spont	0.78	**0.86**
CE read	0.69	**0.73**
JE	**0.95**	0.88
AE	0.85	**0.98**
IE	**0.98**	0.95

NIST	METEOR
CE spont	**0.86**
CE read	0.72
JE	0.91
AE	0.90
IE	0.97

With the exception of the CE translation task, very high correlation coefficients were obtained, but large differences can be seen for each source language. BLEU4 seems to correlate better with NIST for JE, but better with METEOR for AE. These characteristics also affect the MT engine rankings (see Appendix B.2.). Analogous to the subjective evaluation, an "overall" MT engine ranking combining all automatic evaluation metrics for the translation of ASR output is given in Table 12. Again, an omitted horizontal line between MT engines indicates the systems were not significantly different.

3.3. Correlation between Subjective and Automatic Evaluation Results

The evaluation metrics listed in Table 8 were selected because the outcomes of last year's IWSLT workshop showed that these metrics were closely related to human judgement. Table 13 summarizes the *Pearson correlation coefficients* between BLEU4/NIST/METEOR and *adequacy*/*fluency* for this year's CE translation task.

The results confirm previous findings that *fluency* correlates best with BLEU4 and that *adequacy* correlates best with METEOR. The NIST metric has only moderate correlation to both subjective evaluation metrics.

Interestingly, the correlation coefficients are much higher for correct recognition results than for the translation of ASR outputs. This is especially so for the *spontaneous speech* translation task where only low correlations were obtained for *adequacy*. This indicates that standard evaluation metrics might not be appropriate for dealing with spontaneous speech translation tasks. Future investigations should focus

Table 12: Combination of Automatic Evaluation Rankings

CE spont	*CE read*	*JE read*	*AE read*	*IE read*
RWTH	RWTH	RWTH	IBM	Washington-U
JHU_WS06	MIT-LL_AFRL	NiCT-ATR	NiCT-ATR	NiCT-ATR
NiCT-ATR	NiCT-ATR	UKACMU_SMT	TALP_tuples	TALP_tuples
UKACMU_SMT	JHU_WS06	NTT	TALP_comb	MIT-LL_AFRL
HKUST	ITC-irst	MIT-LL_AFRL	NTT	TALP_comb
ITC-irst	TALP_tuples	ITC-irst	UKACMU_SMT	ITC-irst
MIT-LL_AFRL	TALP_phrases	SLE	TALP_phrases	TALP_phrases
NTT	UKACMU_SMT	HKUST	ITC-irst	NTT
Xiamen-U	HKUST	TALP_tuples	DCU	DCU
ATT	TALP_comb	NAIST	HKUST	UKACMU_SMT
NLPR	NTT	Kyoto-U	CLIPS	HKUST
CLIPS	Xiamen-U	TALP_comb		CLIPS
	NLPR	TALP_phrases		
	ATT	CLIPS		

Table 13: Correlation between Subjective and Automatic Evaluation Metrics

CE spon	BLEU4	NIST	METEOR
Fluency	**0.88**	0.55	0.72
Adequacy	0.34	**0.57**	0.54

CE read	BLEU4	NIST	METEOR
Fluency	**0.89**	0.63	0.66
Adequacy	0.83	0.64	**0.89**

CE CRR	BLEU4	NIST	METEOR
Fluency	**0.96**	0.84	0.93
Adequacy	0.95	0.82	**0.96**

on how to measure the impact of spontaneity aspects on the MT translation quality in order to improve the reliability of automatic evaluation metrics.

4. Discussion

4.1. Challenge Task 2006

As indicated by the English language perplexity figures listed in Table 3, the *Challenge Task 2006* of this year's evaluation campaign was much more difficult than the translation tasks of previous IWSLT workshops. The MT performance for all translation conditions on this year's evaluation set was much lower compared to the results of previous IWSLT evaluation campaigns.

One of the reasons is the discrepancy between the supplied resources and this year's evaluation data set. The supplied resources contain mainly short sentences, whereas the evaluation data sentences were much longer. In addition, the OOV rate is quite high for this year's IWSLT 2006 evaluation data. Hence, current state-of-the-art MT systems need to improve their capability to deal with input sentences having characteristics not covered by the training corpus or containing phrases never seen before. Further research on automatic text preprocessing techniques (*sentence splitting*, *word segmentation*, etc,), *model adaptation* and the translation of *unknown words* is necessary to fill the gap.

4.2. Additional Resources

Comparing the *Open Data Track* with the *CSTAR Data Track* results improvements of up to 4%-5% in BLEU as well as METEOR and 0.5-0.7 points in NIST were obtained when using additional in-domain training data for CE and JE.

In addition, some participants also investigated in the utilization of additional out-of-domain training resources [14, 29] and reported mixed success depending on the input condition and translation task.

4.3. Evaluation Specifications

When comparing the results of the *official* and *additional* evaluation specification, the utilized evaluation metrics showed quite different phenomena. The NIST scores are generally lower for the evaluation taking into account punctuation and case information.

Very similar scores were obtained for METEOR. However, the current version of this metric is not compatible with the *official* evaluation specifications. The script removes punctuation/case information and separates word compounds, differing from the *additional* evaluation specifications and therefore resulting in slightly different scores.

An unexpected effect, however, was seen for the BLEU metric. In contrast to NIST, many MT engines achieved higher BLEU scores for the *official* evaluation specifications, despite punctuation/case errors in the MT output. The extent of this phenomenon, however, differed between the language pairs (JE: 50%, AE: 30%, CE: 30% of the utilized MT engines). Interestingly, this phenomenon was not found for the translation of Italian where the BLEU scores of the *additional* evaluation specifications were always higher.

4.4. Language Dependency

For the IWSLT 2006 evaluation data, the same set of English reference translations were used for the evaluation of all translations outputs. Therefore, the translation results of MT engines using different source languages as the input can be directly compared.

Looking at the automatic evaluation results of the *Open Data Track*, the highest scores were obtained on the IE translation task for the CRR and the ASR output translation conditions. The latter was surprising given Italian had the worst recognition accuracies. One reason might be the close relationship between these two languages.

For Arabic, the high OOV rate caused problems for MT systems that relied on the supplied word segmentations. However, resegmenting the data set proved to be effective for increasing the vocabulary coverage and improving translation quality [14].

For Japanese, the highest recognition accuracy was obtained. However, due to large differences in syntactic structure and word order, the JE translation task seems to be one of the most difficult tasks and the best performing systems obtained lower scores compared to the AE and IE results. Interestingly, the JE task featured the largest number of non-SMT engines including a commercial system that achieved quite good performances (see [24, 17]).

For Chinese, the recognition accuracy for read speech is similar to the Arabic recognition results, but the automatic evaluation scores obtained for the top-scoring MT engines are much lower. The complexity of the CE translation task seems to be similar to JE. Altogether, the complexity[14] of the translation tasks of this year's IWSLT evaluation campaign can be summarized as:

$$\text{CE} \approx \text{JE} > \text{AE} \gg \text{IE}$$

4.5. Robustness towards ASR Output

When comparing the results of the *ASR Output* condition and the CRR data sets, all MT engines achieved lower scores for the translation of ASR output. The complexity of the translation input conditions can be summarized as:

$$\text{ASR spont} > \text{ASR read} \gg \text{CRR}$$

The impact of recognition errors, however, differs between the languages and is closely related to the recognition accuracy obtained for the respective speech input. A moderate degradation was seen for the JE/AE/CE translation tasks (0.5-3% for BLEU, 0.3-0.7 points for NIST, 3-7% for METEOR). However, the low recognition performance for Italian caused the largest difference (5-8% for BLEU, 0.9-1.2 points for NIST, 6-12% for METEOR) for IE.

In addition, MT engines that were only applied to the first-best recognition output showed a larger drop in performance than MT engines that directly used information from the word lattice. In order to make MT systems more robust against speech recognition errors and to tap the full potential of the ASR systems, more research on how to directly exploit word lattices efficiently is required. The results on using *confusion networks* reported by IWSLT 2006 participants [15, 16, 29] are promising and lead into the right direction.

[14] $\approx$: "*similar*", $>$: "*more difficult*", $\gg$: "*much more difficult*".

5. Conclusion

This year's workshop provided a testbed for applying novel ideas on how to deal with problems in the area of spontaneous speech translation. Various innovative ideas were explored, most notably the *usage of out-of-domain training data* [14, 29], new methods for *distortion modeling* [15, 26], *topic-dependent model adaptation* [20, 23], *efficient decoding of word lattices* [16], and *rescoring/reranking methods of NBEST list* [22, 23, 29]. Although not all ideas proved to be effective, new insights into the complexity of combining speech recognition and machine translation technologies were obtained that will help to advance the current state-of-the-art in 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 Roldano Cattoni, Roger Hsiao, Gen Itoh, Shigeki Matsuda, Jinsong Zhang, Shuwu Zhang for their support in recording the speech data sets and generating the ASR outputs. Special thanks to Matthias Eck for his extensive technical support in setting-up and maintaining the online evaluation servers. 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 2006 evaluation campaign and for making the IWSLT 2006 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 the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 1–12.

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

[3] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proc. of the EUROSPEECH03*, Geneve, Switzerland, 2003, pp. 381–384.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] S. Bangalore, S. Kanthak, and P. Haffner, "Finite-State Transducer-based Statistical Machine Translation using Joint Probabilities," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 16–22.

[11] C. Boitet, Y. Bey, M. Tomokiyo, W. Cao, and H. Blanchon, "IWSLT-06: experiments with commercial MT systems and lessons from subjective evaluations," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 23–30.

[12] N. Stroppa and A. Way, "MATREX: DCU Machine Translation System for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 31–36.

[13] M. Carpuat, Y. Shen, X. Yu, and D. Wu, "Toward Integrating Word Sense and Entity Disambiguation into Statistical Machine Translation," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 37–44.

[14] Y.-S. Lee, "IBM Arabic-to-English Translation for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 45–52.

[15] B. Chen, R. Cattoni, N. Bertoldi, M. Cetello, and M. Federico, "The ITC-irst SMT System for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 53–58.

[16] W. Shen, R. Zens, N. Bertoldi, and M. Federico, "The JHU Workshop 2006 IWSLT System," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 59–63.

[17] T. Nakazawa, K. Yu, D. Kawahara, and S. Kurohashi, "Example-based Machine Translation based on Deeper NLP," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 64–70.

[18] W. Shen, B. Delaney, and T. Anderson, "The MIT-LL/AFRL IWSLT-2006 MT System," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 71–76.

[19] M. Komachi, M. Nagata, and Y. Matsumoto, "Phrase Reordering for Statistical Machine Translation Based on Predicate-Argument Structure," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 77–82.

[20] R. Zhang, H. Yamamoto, M. Paul, H. Okuma, K. Yasuda, Y. Lepage, E. Denoual, D. Mochihashi, A. Finch, and E. Sumita, "The NiCT-ATR Statistical Machine Translation System for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 83–90.

[21] C. Chai, J. Du, W. Wei, P. Liu, K. Zhou, Y. He, and C. Zong, "NLPR Translation System for IWSLT 2006 Evaluation Campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 91–94.

[22] T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki, "NTT Statistical Machine Translation for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 95–102.

[23] A. Mauser, R. Zens, E. Matusov, S. Hasan, and H. Ney, "The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 103–110.

[24] P. Whitelock and V. Poznanski, "The SLE Example-Based Translation System," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 111–115.

[25] J. M. Crego, A. de Gispert, P. Lambert, M. Khalilov, M. R. Costa-jussà, J. B. Mariño, R. Banchs, and J. A. Fonollosa, "The TALP Ngram-based SMT System for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 116–122.

[26] M. R. Costa-jussà, J. M. Crego, A. de Gispert, P. Lambert, M. Khalilov, J. A. Fonollosa, J. B. Mariño, and R. Banchs, "TALP Phrase-Based System and TALP System Combination for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 123–129.

[27] M. Eck, I. Lane, N. Bach, S. Hewavitharana, M. Kolss, B. Zhao, A. S. Hildebrand, S. Vogel, and A. Waibel, "The UKA/CMU Statistical Machine Translation System for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 130–137.

[28] A. Zollmann, A. Venugopal, S. Vogel, and A. Waibel, "The CMU-UKA Syntax Augmented Machine Translation System for IWSLT-06," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 138–144.

[29] K. Kirchhoff, K. Duh, and C. Lim, "The University of Washington Machine Translation System for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 145–152.

[30] Y. Chen, X. Shi, and C. Zhou, "The XMU Phrase-Based Statistical Machine Translation System for IWSLT 2006," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2006, pp. 153–157.

Appendix A. MT System Overview

Research Group	MT System Description	Type	MT System	Input
AT&T Research	Finite-State Transducer-based Statistical Machine Translation using Joint Probabilities [10]	SMT	ATT	C_s, C_r
CLIPS-GETA	IWSLT-06: experiments with commercial MT systems and lessons from subjective evaluations [11]	RBMT	CLIPS	C_s^*,J*,A*,I*
Dublin City University, School of Computing	MATREX: DCU Machine Translation System for IWSLT 2006 [12]	SMT	DCU	A,I*
Hong Kong University of Science and Technology	Toward Integrating Word Sense and Entity Disambiguation into Statistical Machine Translation [13]	SMT	HKUST	C_s, C_r,J,A,I
IBM, Multilingual NLP Dept.	IBM Arabic-to-English Translation for IWSLT 2006 [14]	SMT	IBM	A
Instituto Trentino di Cultura, Center for Scientific and Technological Research	The ITC-irst SMT System for IWSLT 2006 [15]	SMT	ITC-irst	C_s, C_r,J,A,I
Johns Hopkins University, Summer Workshop 2006	The JHU Workshop 2006 IWSLT System [16]	SMT	JHU_WS06	C_s, C_r
Kyoto University, Kurohashi Lab	Example-based Machine Translation based on Deeper NLP [17]	EBMT	Kyoto-U	J
MIT Lincoln Laboratory / Air Force Research Laboratory	The MIT-LL/AFRL IWSLT-2006 MT System [18]	SMT	MIT-LL_AFRL	C_s, C_r,J,I
National Institute of Science and Technology	Phrase Reordering for Statistical Machine Translation Based on Predicate-Argument Structure [19]	SMT	NAIST	J
National Institute of Information and Communication Technology / ATR	The NiCT-ATR Statistical Machine Translation System for IWSLT 2006 [20]	SMT	NiCT-ATR	C_s, C_r,J,A,I
National Laboratory of Pattern Recognition, Chinese Academy of Science	NLPR Translation System for IWSLT 2006 Evaluation Campaign [21]	RBMT SMT	NLPR	C_s, C_r
NTT Communication Research Laboratories	NTT Statistical Machine Translation for IWSLT 2006 [22]	SMT	NTT	C_s, C_r,J,A,I
Rheinisch Westfaelische Technische Hochschule, Lehrstuhl für Informatik 6	The RWTH Statistical Machine Translation System for the IWSLT 2006 Evaluation [23]	SMT	RWTH	C_s, C_r,J
SHARP Laboratories of Europe	The SLE Example-Based Translation System [24]	EBMT	SLE	J
TALP-UPC Research Center	The TALP Ngram-based SMT System for IWSLT 2006 [25]	SMT	TALP_tuples	C_r,J,A,I
TALP-UPC Research Center	TALP Phrase-Based System and TALP System Combination for IWSLT 2006 [26]	SMT	TALP_phrases TALP_comb	C_r,J,A,I
InterACT Reserach Labs, Carnegie Mellon University / Univ. of Karlsruhe	The UKA/CMU Statistical Machine Translation System for IWSLT 2006 [27]	SMT	UKACMU_SMT	C_s, C_r,J,A,I
InterACT Reserach Labs, Carnegie Mellon University / Univ. of Karlsruhe	The CMU-UKA Syntax Augmented Machine Translation System for IWSLT-06 [28]	SMT	UKACMU_SAMT	C_s, C_r
University of Washington	The University of Washington Machine Translation System for IWSLT 2006 [29]	SMT	Washington-U	I
Xiamen University, Institute of Artifical Intelligence	The XMU Phrase-Based Statistical Machine Translation System for IWSLT 2006 [30]	SMT	Xiamen-U	C_s, C_r

'*' indicates late run submissions that were submitted after the official run submission period.

Appendix B. Human Evaluation

B.1. Evaluation Results

CE – ASR output

spontaneous speech

Adequacy	**MT Engine**	Fluency
Open Data Track		
0.9647	JHU_WS06	1.2175
0.9255	NTT	0.9258
0.9103	RWTH	1.2426
0.9096	UKACMU_SMT	0.8092
0.8852	NiCT-ATR	0.7832
0.7853	MIT-LL_AFRL	**1.3734**
CSTAR Data Track		
1.1933	NiCT-ATR	**1.3169**

read speech

Adequacy	**MT Engine**	Fluency
Open Data Track		
1.0297	JHU_WS06	1.2104
1.0261	MIT-LL_AFRL	1.1278
1.0136	RWTH	**1.2952**
0.9885	NTT	0.9973
0.9245	NiCT-ATR	0.8271
0.9112	UKACMU_SMT	0.8375
CSTAR Data Track		
1.2477	NiCT-ATR	**1.3156**

correct recognition

Adequacy	**MT Engine**	Fluency
Open Data Track		
1.4319	MIT-LL_AFRL	1.3472
1.4225	NTT	1.3008
1.3896	RWTH	**1.6498**
1.3503	JHU_WS06	1.3302
1.3418	NiCT-ATR	1.0300
1.3080	UKACMU_SMT	1.0818
CSTAR Data Track		
1.6711	NiCT-ATR	**1.5922**

B.2. MT Engine Rankings

(lines between MT engines indicate significant differences in performance between the respective MT engines according to the *bootStrap* method [9])

CE – ASR output

spontaneous speech

Adequacy	Fluency
Open Data Track	
JHU_WS06	MIT-LL_AFRL
NTT	RWTH
RWTH	JHU_WS06
UKACMU_SMT	NTT
NiCT-ATR	UKACMU_SMT
MIT-LL_AFRL	NiCT-ATR
CSTAR Data Track	
NiCT-ATR	NiCT-ATR

read speech

Adequacy	Fluency
Open Data Track	
JHU_WS06	RWTH
MIT-LL_AFRL	JHU_WS06
RWTH	MIT-LL_AFRL
NTT	NTT
NiCT-ATR	UKACMU_SMT
UKACMU_SMT	NiCT-ATR
CSTAR Data Track	
NiCT-ATR	NiCT-ATR

correct recognition

Adequacy	Fluency
Open Data Track	
MIT-LL_AFRL	RWTH
NTT	MIT-LL_AFRL
RWTH	JHU_WS06
JHU_WS06	NTT
NiCT-ATR	UKACMU_SMT
UKACMU_SMT	NiCT-ATR
CSTAR Data Track	
NiCT-ATR	NiCT-ATR

Appendix C. Automatic Evaluation

C.1. Evaluation Results

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

('*' indicates late run submissions that were submitted after the official submission period)

ASR Output						**MT**	**Correct Recognition Result**					
official evaluation			*additional evaluation*			**Engine**	*official evaluation*			*additional evaluation*		
BLEU4	NIST	METEOR	BLEU4	NIST	METEOR		BLEU4	NIST	METEOR	BLEU4	NIST	METEOR
CE – spontaneous speech												
Open Data Track												
0.1898	5.0523	0.4198	**0.1858**	5.2331	0.4197	RWTH	**0.2423**	**6.0961**	**0.5033**	**0.2446**	6.4609	**0.5031**
0.1807	**5.1513**	0.4138	0.1768	5.4270	0.4134	JHU_WS06	0.2140	6.0225	0.4802	0.2184	**6.4992**	0.4798
0.1657	4.2363	0.3800	0.1661	4.3968	0.3798	MIT-LL_AFRL	0.2157	6.0537	0.4895	0.2178	6.4985	0.4892
0.1630	4.9732	0.4043	0.1680	5.3175	0.4042	UKACMU_SMT	0.1996	5.7603	0.4729	0.2045	6.1850	0.4726
0.1591	4.9696	0.4117	0.1615	5.3592	0.4114	NiCT-ATR	0.2060	5.8613	0.4870	0.2123	6.3848	0.4862
0.1559	4.1801	0.3946	0.1584	4.5173	0.3945	NTT	0.2135	5.1271	0.4743	0.2166	5.5115	0.4736
0.1505	4.6813	0.3763	0.1623	4.9573	0.3768	Xiamen-U	0.1976	5.5640	0.4783	0.2162	5.9756	0.4791
0.1441	4.6365	**0.4238**	0.1653	5.3703	**0.4242**	HKUST	0.1804	5.3615	0.4915	0.2038	6.2078	0.4917
0.1422	4.9188	0.4119	0.1577	**5.4799**	0.4121	ITC-irst	0.1837	5.8267	0.4852	0.1992	6.4263	0.4851
0.1155	4.1762	0.3584	0.1229	4.5258	0.3583	ATT	0.1439	4.8954	0.4164	0.1511	5.2806	0.4165
0.1070	3.5755	0.3901	0.1005	3.6311	0.3899	NLPR	0.1284	4.0658	0.4601	0.1237	4.2242	0.4597
0.0585	3.7981	0.3143	0.0649	4.1666	0.3149	CLIPS*	0.0749	4.4256	0.3694	0.0804	4.8077	0.3701
CSTAR Data Track												
0.2008	**5.4009**	**0.4502**	**0.2039**	**5.8205**	**0.4492**	NiCT-ATR	**0.2645**	**6.5274**	**0.5425**	**0.2751**	**7.0860**	**0.5419**
0.1622	5.1865	0.4180	0.1605	5.5303	0.4177	UKACMU_SMT	0.2057	6.0548	0.4987	0.2103	6.5941	0.4983
0.1566	4.6606	0.3833	0.1481	4.7742	0.3828	UKACMU_SAMT	0.1954	5.7681	0.4642	0.1918	6.1137	0.4632
CE – read speech												
Open Data Track												
0.2111	5.4045	0.4432	**0.2032**	5.5644	0.4430	RWTH	**0.2423**	**6.0961**	**0.5033**	**0.2446**	6.4609	**0.5031**
0.1863	5.3290	0.4276	0.1894	5.6685	0.4275	JHU_WS06	0.2140	6.0225	0.4802	0.2184	**6.4992**	0.4798
0.1861	**5.4154**	0.4355	0.1877	**5.7698**	0.4354	MIT-LL_AFRL	0.2157	6.0537	0.4895	0.2178	6.4985	0.4892
0.1834	4.5306	0.4215	0.1876	4.9075	0.4212	NTT	0.2135	5.1271	0.4743	0.2166	5.5115	0.4736
0.1775	5.2286	0.4336	0.1772	5.6649	0.4323	NiCT-ATR	0.2060	5.8613	0.4870	0.2123	6.3848	0.4862
0.1710	5.0768	0.4227	0.1738	5.3809	0.4222	UKACMU_SMT	0.1996	5.7603	0.4729	0.2045	6.1850	0.4726
0.1650	4.8933	0.4266	0.1728	5.3577	0.4264	TALP_comb	0.1916	5.3980	0.4749	0.2021	5.9698	0.4749
0.1624	4.9779	0.4336	0.1748	5.5128	0.4333	TALP_tuples	0.1863	5.5714	0.4824	0.2034	6.2119	0.4825
0.1599	5.1255	0.4307	0.1676	5.6413	0.4307	TALP_phrases	0.1899	5.8030	0.4833	0.2008	6.4275	0.4832
0.1579	5.0115	0.4049	0.1718	5.3595	0.4052	Xiamen-U	0.1976	5.5640	0.4783	0.2162	5.9756	0.4791
0.1560	5.2207	0.4374	0.1698	5.7435	0.4370	ITC-irst	0.1837	5.8267	0.4852	0.1992	6.4263	0.4851
0.1545	4.7769	**0.4456**	0.1740	5.4961	**0.4457**	HKUST	0.1804	5.3615	0.4915	0.2038	6.2078	0.4917
0.1226	4.3813	0.3729	0.1297	4.7122	0.3733	ATT	0.1439	4.8954	0.4164	0.1511	5.2806	0.4165
0.1037	3.6384	0.4073	0.1022	3.7433	0.4076	NLPR	0.1284	4.0658	0.4601	0.1237	4.2242	0.4597
CSTAR Data Track												
0.2155	**5.6857**	**0.4787**	**0.2214**	**6.1453**	**0.4783**	NiCT-ATR	**0.2645**	**6.5274**	**0.5425**	**0.2751**	**7.0860**	**0.5419**
0.1685	5.0292	0.4111	0.1630	5.2834	0.4108	UKACMU_SAMT	0.1954	5.7681	0.4642	0.1918	6.1137	0.4632
0.1645	5.2372	0.4315	0.1647	5.6395	0.4308	UKACMU_SMT	0.2057	6.0548	0.4987	0.2103	6.5941	0.4983

ASR Output						MT	Correct Recognition Result					
official evaluation			*additional evaluation*			**Engine**	*official evaluation*			*additional evaluation*		
BLEU4	NIST	METEOR	BLEU4	NIST	METEOR		BLEU4	NIST	METEOR	BLEU4	NIST	METEOR
JE – read speech												
Open Data Track												
0.2142	**5.6502**	0.4570	**0.2107**	5.9364	**0.4571**	RWTH	**0.2368**	5.9183	0.4886	**0.2315**	6.2325	0.4890
0.1984	5.4843	0.4500	0.1906	5.7956	0.4498	NTT	0.2203	5.9077	0.4877	0.2147	6.3156	0.4873
0.1899	5.5915	**0.4574**	0.1832	**5.9428**	0.4569	NiCT-ATR	0.2122	5.9494	**0.4900**	0.2077	**6.3325**	**0.4893**
0.1891	5.5967	0.4421	0.1793	5.8474	0.4419	MIT-LL/AFRL	0.2099	**5.9866**	0.4757	0.1995	6.2570	0.4753
0.1868	5.6343	0.4505	0.1794	5.9031	0.4509	UKACMU_SMT	0.2030	5.9322	0.4820	0.1917	6.1468	0.4824
0.1604	5.4171	0.4397	0.1617	5.8325	0.4392	ITC-irst	0.1839	5.8538	0.4744	0.1882	6.3197	0.4742
0.1599	5.3393	0.4257	0.1467	5.6063	0.4257	SLE	0.1726	5.6497	0.4570	0.1601	5.9765	0.4568
0.1523	4.9022	0.4283	0.1647	5.4583	0.4283	HKUST	0.1560	4.8750	0.4579	0.1786	5.7071	0.4578
0.1418	4.8804	0.4057	0.1375	5.2416	0.4047	Kyoto-U	0.1655	5.4325	0.4497	0.1629	5.8843	0.4491
0.1390	4.7672	0.4105	0.1439	5.1993	0.4101	TALP_comb	0.1467	4.9743	0.4382	0.1566	5.5050	0.4383
0.1370	4.9437	0.4133	0.1397	5.3995	0.4134	TALP_tuples	0.1461	5.2717	0.4425	0.1495	5.8068	0.4426
0.1311	4.8372	0.4415	0.1360	5.3846	0.4410	NAIST	0.1431	5.2105	0.4664	0.1508	5.8442	0.4656
0.1280	4.7596	0.4066	0.1343	5.2119	0.4066	TALP_phrases	0.1370	5.0665	0.4331	0.1451	5.5877	0.4333
0.0755	3.7685	0.3325	0.0814	4.0450	0.3326	CLIPS*	0.0921	4.2723	0.3773	0.1006	4.6379	0.3773
CSTAR Data Track												
0.2487	**6.2778**	**0.5039**	**0.2468**	**6.7157**	**0.5032**	NiCT-ATR	**0.2861**	**6.8327**	**0.5536**	**0.2867**	**7.3021**	**0.5529**
0.1841	5.3980	0.4316	0.1760	5.5606	0.4319	UKACMU_SMT	0.2007	5.8584	0.4830	0.1956	6.3340	0.4831
AE – read speech												
Open Data Track												
0.2274	5.8466	0.4845	**0.2428**	**6.4867**	0.4842	IBM	**0.2549**	**6.3769**	**0.5316**	**0.2773**	**7.1681**	**0.5314**
0.2136	5.8213	0.4786	0.2146	6.2598	0.4783	TALP_tuples	0.2323	6.2380	0.5134	0.2383	6.7958	0.5133
0.2117	**5.9216**	**0.4867**	0.2164	6.3959	**0.4869**	NiCT-ATR	0.2365	6.3521	0.5224	0.2463	6.8893	0.5229
0.2101	5.5583	0.4747	0.2131	6.0012	0.4740	TALP_comb	0.2327	6.0337	0.5091	0.2395	6.5972	0.5087
0.2071	4.8403	0.4397	0.1967	4.7567	0.4384	NTT	0.2265	5.3316	0.4776	0.2216	5.3577	0.4758
0.1995	5.3359	0.4513	0.2086	5.6303	0.4511	UKACMU_SMT	0.2208	5.9059	0.4932	0.2349	6.3037	0.4929
0.1908	5.5448	0.4652	0.1989	6.0147	0.4646	TALP_phrases	0.2122	6.0177	0.5010	0.2220	6.5405	0.5004
0.1723	4.7352	0.4186	0.1780	5.1899	0.4182	ITC-irst	0.2005	5.1816	0.4581	0.2048	5.6040	0.4564
0.1477	3.3318	0.3920	0.1584	3.7237	0.3911	HKUST	0.1663	3.8863	0.4288	0.1800	4.4473	0.4273
0.1450	4.5307	0.4020	0.1391	4.7936	0.4000	DCU	0.1624	4.8902	0.4336	0.1589	5.2900	0.4320
0.0490	3.6202	0.2861	0.0455	3.7567	0.2864	CLIPS*	0.0601	3.9051	0.3110	0.0573	4.0840	0.3112
CSTAR Data Track												
0.2123	**5.8693**	**0.4875**	0.2234	6.3717	0.4873	UKACMU_SMT	**0.2420**	**6.4073**	**0.5275**	0.2584	6.9741	0.5276
IE – read speech												
Open Data Track												
0.2989	6.8985	0.5744	**0.3194**	7.4724	0.5739	NiCT-ATR	**0.3763**	8.1318	0.6630	0.4120	8.9027	0.6625
0.2837	6.7065	0.5660	0.3067	7.3139	0.5657	TALP_comb	0.3396	7.6405	0.6332	0.3774	8.4035	0.6328
0.2818	6.8723	0.5764	0.3067	7.5256	0.5761	TALP_tuples	0.3331	7.7474	0.6398	0.3738	8.5922	0.6394
0.2798	6.8593	0.5679	0.3007	7.5070	0.5678	MIT-LL/AFRL	0.3574	8.0089	0.6669	0.3920	8.8548	0.6669
0.2797	6.6217	0.5592	0.2969	7.2595	0.5588	ITC-irst	0.3497	7.8155	0.6468	0.3797	8.6186	0.6461
0.2787	**6.9318**	**0.5853**	0.3168	**7.6902**	**0.5853**	Washington-U	0.3543	**8.1890**	**0.7017**	**0.4206**	**9.2410**	**0.7019**
0.2769	6.6959	0.5607	0.2864	7.1949	0.5602	NTT	0.3449	7.8259	0.6431	0.3750	8.5266	0.6428
0.2684	6.6443	0.5634	0.2940	7.2944	0.5631	TALP_phrases	0.3200	7.5248	0.6256	0.3555	8.3201	0.6254
0.2598	6.5845	0.5497	0.2783	7.2281	0.5495	DCU*	0.3126	7.5462	0.6246	0.3467	8.3579	0.6245
0.2388	6.1999	0.5376	0.2577	6.8230	0.5371	UKACMU_SMT	0.3030	7.3011	0.6293	0.3419	8.1405	0.6286
0.2374	6.0956	0.5403	0.2778	7.0994	0.5398	HKUST	0.2964	7.1816	0.6239	0.3567	8.3486	0.6236
0.1368	5.1528	0.4322	0.1538	5.5129	0.4326	CLIPS*	0.1894	6.0183	0.5279	0.2210	6.4866	0.5283
CSTAR Data Track												
0.263	**6.6617**	**0.5638**	**0.2826**	**7.3188**	**0.5633**	UKACMU_SMT	**0.3312**	**7.7622**	**0.6587**	**0.3756**	**8.6779**	**0.6583**

C.2. MT Engine Rankings

(*official evaluation*: case-sensitive, with punctuations tokenized)

(lines between MT engines indicate significant differences in performance
between the respective MT engines according to the *bootStrap* method [9])

('*' indicates late run submissions that were submitted after the official submission period)

CE - spontaneous speech

ASR Output

BLEU4	NIST	METEOR
Open Data Track		
RWTH	JHU_WS06	HKUST
JHU_WS06	RWTH	RWTH
MIT-LL_AFRL	UKACMU_SMT	JHU_WS06
UKACMU_SMT	NiCT-ATR	ITC-irst
NiCT-ATR	ITC-irst	NiCT-ATR
NTT	Xiamen-U	UKACMU_SMT
Xiamen-U	HKUST	NTT
HKUST	MIT-LL_AFRL	NLPR
ITC-irst	NTT	MIT-LL_AFRL
ATT	ATT	Xiamen-U
NLPR	CLIPS*	ATT
CLIPS*	NLPR	CLIPS*
CSTAR Data Track		
NiCT-ATR	NiCT-ATR	NiCT-ATR
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT
UKACMU_SAMT	UKACMU_SAMT	UKACMU_SAMT

Correct Recognition Result

BLEU4	NIST	METEOR
Open Data Track		
RWTH	RWTH	RWTH
MIT-LL_AFRL	MIT-LL_AFRL	HKUST
JHU_WS06	JHU_WS06	MIT-LL_AFRL
NTT	NiCT-ATR	NiCT-ATR
NiCT-ATR	ITC-irst	ITC-irst
UKACMU_SMT	UKACMU_SMT	JHU_WS06
Xiamen-U	Xiamen-U	Xiamen-U
ITC-irst	HKUST	NTT
HKUST	NTT	UKACMU_SMT
ATT	ATT	NLPR
NLPR	CLIPS*	ATT
CLIPS*	NLPR	CLIPS*
CSTAR Data Track		
NiCT-ATR	NiCT-ATR	NiCT-ATR
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT
UKACMU_SAMT	UKACMU_SAMT	UKACMU_SAMT

CE - read speech

ASR Output

BLEU4	NIST	METEOR
Open Data Track		
RWTH	MIT-LL_AFRL	HKUST
JHU_WS06	RWTH	RWTH
MIT-LL_AFRL	JHU_WS06	ITC-irst
NTT	NiCT-ATR	MIT-LL_AFRL
NiCT-ATR	ITC-irst	NiCT-ATR
UKACMU_SMT	TALP_phrases	TALP_tuples
TALP_comb	UKACMU_SMT	TALP_phrases
TALP_tuples	Xiamen-U	JHU_WS06
TALP_phrases	TALP_tuples	TALP_comb
Xiamen-U	TALP_comb	UKACMU_SMT
ITC-irst	HKUST	NTT
HKUST	NTT	NLPR
ATT	ATT	Xiamen-U
NLPR	NLPR	ATT
CSTAR Data Track		
NiCT-ATR	NiCT-ATR	NiCT-ATR
UKACMU_SAMT	UKACMU_SMT	UKACMU_SMT
UKACMU_SMT	UKACMU_SAMT	UKACMU_SAMT

Correct Recognition Result

BLEU4	NIST	METEOR
Open Data Track		
RWTH	RWTH	RWTH
MIT-LL_AFRL	MIT-LL_AFRL	HKUST
JHU_WS06	JHU_WS06	MIT-LL_AFRL
NTT	NiCT-ATR	NiCT-ATR
NiCT-ATR	ITC-irst	ITC-irst
UKACMU_SMT	TALP_phrases	TALP_phrases
Xiamen-U	UKACMU_SMT	TALP_tuples
TALP_comb	TALP_tuples	JHU_WS06
TALP_phrases	Xiamen-U	Xiamen-U
TALP_tuples	TALP_comb	TALP_comb
ITC-irst	HKUST	NTT
HKUST	NTT	UKACMU_SMT
ATT	ATT	NLPR
NLPR	NLPR	ATT
CSTAR Data Track		
NiCT-ATR	NiCT-ATR	NiCT-ATR
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT
UKACMU_SAMT	UKACMU_SAMT	UKACMU_SAMT

JE – read speech

ASR Output

BLEU4	NIST	METEOR
Open Data Track		
RWTH	RWTH	NiCT-ATR
NTT	UKACMU_SMT	RWTH
NiCT-ATR	MIT-LL_AFRL	UKACMU_SMT
MIT-LL_AFRL	NiCT-ATR	NTT
UKACMU_SMT	NTT	MIT-LL_AFRL
ITC-irst	ITC-irst	NAIST
SLE	SLE	ITC-irst
HKUST	TALP_tuples	HKUST
Kyoto-U	HKUST	SLE
TALP_comb	Kyoto-U	TALP_tuples
TALP_tuples	NAIST	TALP_comb
NAIST	TALP_comb	TALP_phrases
TALP_phrases	TALP_phrases	Kyoto-U
CLIPS*	CLIPS*	CLIPS*
CSTAR Data Track		
NiCT-ATR	NiCT-ATR	NiCT-ATR
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT

Correct Recognition Result

BLEU4	NIST	METEOR
Open Data Track		
RWTH	MIT-LL_AFRL	NiCT-ATR
NTT	NiCT-ATR	RWTH
NiCT-ATR	UKACMU_SMT	NTT
MIT-LL_AFRL	RWTH	UKACMU_SMT
UKACMU_SMT	NTT	MIT-LL_AFRL
ITC-irst	ITC-irst	ITC-irst
SLE	SLE	NAIST
Kyoto-U	Kyoto-U	HKUST
HKUST	TALP_tuples	SLE
TALP_comb	NAIST	Kyoto-U
TALP_tuples	TALP_phrases	TALP_tuples
NAIST	TALP_comb	TALP_comb
TALP_phrases	HKUST	TALP_phrases
CLIPS*	CLIPS*	CLIPS*
CSTAR Data Track		
NiCT-ATR	NiCT-ATR	NiCT-ATR
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT

AE – read speech

ASR Output

BLEU4	NIST	METEOR
Open Data Track		
IBM	NiCT-ATR	NiCT-ATR
TALP_tuples	IBM	IBM
NiCT-ATR	TALP_tuples	TALP_tuples
TALP_comb	TALP_comb	TALP_comb
NTT	TALP_phrases	TALP_phrases
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT
TALP_phrases	NTT	NTT
ITC-irst	ITC-irst	ITC-irst
HKUST	DCU	DCU
DCU	CLIPS*	HKUST
CLIPS*	HKUST	CLIPS*
CSTAR Data Track		
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT

Correct Recognition Result

BLEU4	NIST	METEOR
Open Data Track		
IBM	IBM	IBM
NiCT-ATR	NiCT-ATR	NiCT-ATR
TALP_comb	TALP_tuples	TALP_tuples
TALP_tuples	TALP_comb	TALP_comb
NTT	TALP_phrases	TALP_phrases
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT
TALP_phrases	NTT	NTT
ITC-irst	ITC-irst	ITC-irst
HKUST	DCU	DCU
DCU	CLIPS*	HKUST
CLIPS*	HKUST	CLIPS*
CSTAR Data Track		
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT

IE – read speech

ASR Output

BLEU4	NIST	METEOR
Open Data Track		
NiCT-ATR	Washington-U	Washington-U
TALP_comb	NiCT-ATR	TALP_tuples
TALP_tuples	TALP_tuples	NiCT-ATR
MIT-LL_AFRL	MIT-LL_AFRL	MIT-LL_AFRL
ITC-irst	TALP_comb	TALP_comb
Washington-U	NTT	TALP_phrases
NTT	TALP_phrases	NTT
TALP_phrases	ITC-irst	ITC-irst
DCU*	DCU*	DCU*
UKACMU_SMT	UKACMU_SMT	HKUST
HKUST	HKUST	UKACMU_SMT
CLIPS*	CLIPS*	CLIPS*
CSTAR Data Track		
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT

Correct Recognition Result

BLEU4	NIST	METEOR
Open Data Track		
NiCT-ATR	Washington-U	Washington-U
MIT-LL_AFRL	NiCT-ATR	MIT-LL_AFRL
Washington-U	MIT-LL_AFRL	NiCT-ATR
ITC-irst	NTT	ITC-irst
NTT	ITC-irst	NTT
TALP_comb	TALP_tuples	TALP_tuples
TALP_tuples	TALP_comb	TALP_comb
TALP_phrases	DCU*	UKACMU_SMT
DCU*	TALP_phrases	TALP_phrases
UKACMU_SMT	UKACMU_SMT	DCU*
HKUST	HKUST	HKUST
CLIPS*	CLIPS*	CLIPS*
CSTAR Data Track		
UKACMU_SMT	UKACMU_SMT	UKACMU_SMT