Overview of the IWSLT04 Evaluation Campaign

Yasuhiro Akiba†1, *Marcello Federico*†2, *Noriko Kando*†3,
Hiromi Nakaiwa†1, *Michael Paul*†1, *Jun'ichi Tsujii*†4

†1ATR, †2ITC-irst, †3NII, †4University of Tokyo
yasuhiro.akiba@atr.jp, federico@itc.it, kando@nii.ac.jp
hiromi.nakaiwa@atr.jp, michael.paul@atr.jp, tsujii@is.s.u-tokyo.ac.jp

Abstract

This paper gives an overview of the evaluation campaign results of the IWSLT04[1] workshop, which is organized by the C-STAR[2] consortium to investigate novel speech translation technologies and their evaluation. The objectives of this workshop is to provide a framework for the applicability validation of existing machine translation evaluation methodologies to evaluate speech translation technologies. The workshop also strives to find new directions in how to improve current methods.

1. Introduction

The drastic increase in demands for the capability to assist trans-lingual conversations, triggered by IT technologies such as the Internet and the expansion of borderless communities such as the increased number of EU countries, has accelerated research activities on speech-to-speech translation technology. Many research projects have been designed to advance this technology, such as VERBMOBIL, C-STAR, NESPOLE!, and BABYLON. These projects, except for C-STAR, have mainly focused on the construction of a prototype system for several language pairs. On the contrary, one of C-STAR's ongoing projects is the joint development of a speech corpus that can handle a common task in multiple languages. As a first result of this activity, a Japanese-English speech corpus comprising tourism-related sentences, originally compiled by ATR, has been translated into the native languages of the C-STAR members. The corpus serves as a primary source for developing and evaluating broad-coverage speech translation technologies [1]. This corpus is used in the research and development of multi-lingual speech-to-speech translation systems on a "common use" basis.

For the effective and efficient research and development of speech-to-speech translation systems, the evaluation of current translation quality is very important. In particular, the system developments done by using a common corpus, like C-STAR project, require careful evaluation of the prominent translation techniques. Therefore, there is strong demand for the establishment of evaluation metrics for multilingual speech-to-speech translation systems.

For this purpose, the Evaluation Campaign 2004 was carried out using parts of the multilingual corpus (cf. Section 2.1). The task was to translate 500 Chinese or Japanese sentences into English. Depending on the amount of permitted training data, three different language resource conditions (*Small Data Track*, *Additional Data Track*, *Unrestricted Data Track*) were distinguished. The translation quality was measured using both human assessments (*subjective evaluation*) and automatic scoring techniques (*automatic evaluation*). The evaluation results of the submitted MT systems are summarized in Section 3.

The corpus supplied for this year's conference, the reference translations, the output of the participating MT systems, and the evaluation results will be made publicly available after the workshop. These resources can be used as a benchmark for future research on MT systems and MT evaluation methodologies.

We hope that IWSLT2004 will become the first step toward establishing standard metrics and a standard corpus for speech-to-speech multi-lingual translation technology.

2. Evaluation Campaign 2004

The Evaluation Campaign 2004 was carried out using parts of the multilingual corpus jointly developed by the C-STAR partners (cf. Section 2.1). The task was to translate 500 Chinese or Japanese sentences into English.

Depending on the amount of permitted training data, three different language resource conditions (*Small Data Track*, *Additional Data Track*, *Unrestricted Data Track*) were distinguished (cf. Section 2.2). Each participant was allowed to register only one MT system in each of the data tracks but could submit multiple translation results (*runs*) for the same track.

In total, 14 institutions took part in this year's workshop, submitting 20 MT systems for the Chinese-to-English (**CE**) and 8 MT systems for the Japanese-to-English (**JE**) translation tasks.

The translation quality was measured using both human assessments (*subjective evaluation*) and automatic scoring

[1] International Workshop on Spoken Language Translation, http://www.slt.atr.jp/IWSLT2004

[2] Consortium for Speech Translation Advanced Research, http://www.c-star.org/

techniques (*automatic evaluation*). The subjective evaluation was carried out by English native speakers. The translation quality was judged based on the *fluency* and *adequacy* of the translation (cf. Section 2.4). For the automatic evaluation, five different automatic scoring metrics (BLEU, NIST, WER, PER, GTM) were applied (cf. Section 2.5). All run submissions were evaluated using the automatic evaluation schemes. However, due to high evaluation costs, the subjective evaluation was limited to one run submission per track for each participant, which could be selected by the participants themselves. The results of the submitted MT systems are summarized in Section 3.

2.1. Multilingual Spoken Language Corpus

The *Basic Travel Expressions Corpus* (BTEC*)[3] is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country and cover utterances for every potential subject in travel situations [2].

For the Evaluation Campaign 2004, parts of the Chinese, Japanese, and English subsets of the BTEC* corpus were used. Details of the supplied IWSLT04 corpus are given in Table 1, where *word token* refers to the number of words in the corpus and *word type* refers to the vocabulary size. The participants were supplied with 20,000 sentence pairs for each translation direction randomly selected from the BTEC* corpus, and the training sets for CE and JE were disjunct.

Table 1: The IWSLT04 corpus

type	language	sentence count		avg.	word	word
		total	unique	length	tokens	types
training	Chinese	20,000	19,288	9.1	182,904	7,643
	English		19,949	9.4	188,935	8,191
	Japanese	20,000	19,046	10.5	209,012	9,277
	English		19,923	9.4	188,712	8,074
develop	Chinese	506	495	6.9	3,515	870
	Japanese	506	502	8.6	4,374	954
	English*	8,089	7,173	7.5	67,410	2,435
test	Chinese	500	492	7.6	3,794	893
	Japanese	500	491	8.7	4,370	979
	English*	8,000	6,907	8.4	66,994	2,496

* reference translations used for automatic evaluation

A development set of additional 506 sentences, including up to 16 reference translations, was provided to the participants to use for the tuning of their MT systems.

The test set consisted of 500 sentences randomly selected from parts of the BTEC* corpus reserved for evaluation purposes. The Chinese data set was created from the original Japanese-English test set, and a consistency check was carried out to guarantee that the Japanese and Chinese source sentences had the same meaning as the English reference translations. Therefore, the test set of the CE and JE translation tasks were identical except for having different sentence orders. Up to 16 reference translations were used for the automatic evaluation of the translation results; the distributions of the number of unique reference translations for each source sentence are summarized in Table 2.

Table 2: Distributions of unique reference translations

# of reference	16	15	14	13	12	11	10	9	8	7
test (500)	169	101	73	60	44	25	12	9	6	1
develop (506)	191	95	89	67	32	19	9	2	2	0

Word segmentations for the Chinese[4] and Japanese[5] subsets are provided when appropriate tools were not available for the participants. However, the participants were permitted to use their own language resources as long as they did not interfere with the language resource conditions of the respective data tracks (cf. Section 2.2).
Table 3 gives some examples of the English IWSLT04 corpus.

Table 3: English sample sentences

My name is Paul Smith .
Are the tickets on sale yet ?
Thank you for the nice meal .
I 'd like one of those , please .
Let me have ten thirty-five cent stamps .
I have to transfer to another flight in Hong Kong at three .
Okay , here you are . If you need anything else , let me know .

2.2. Data Track Conditions

Three different language resource conditions were distinguished. The training data of the *Small Data* track was limited to the supplied corpus only. The *Additional Data* track was set-up for CE only and limited the use of bilingual resources to the ones listed in Table 4. These resources are publicly available from the LDC[6].

Table 4: LDC resources

corpus ID	*corpus name*
LDC2000T46	Hong Kong News Parallel Text
LDC2000T47	Hong Kong Laws Parallel Text
LDC2000T50	Hong Kong Hansards Parallel Text
LDC2001T11	Chinese Treebank 2.0
LDC2001T57	TDT2 Multilanguage Text Version 4.0
LDC2001T58	TDT3 Multilanguage Text Version 2.0
LDC2002L27	Chinese English Translation Lexicon version 3.0
LDC2002T01	Multiple-Translation Chinese Corpus
LDC2003T16	SummBank 1.0
LDC2003T17	Multiple-Translation Chinese (MTC) Part 2
LDC2004T05	Chinese Treebank 4.0
LDC2004T09	ACE 2003 Multilingual Training Data

No restrictions on linguistic resources were imposed for the *Unrestricted Data* track.

[3] Up-to-date information on the BTEC* corpus can be found at http://cstar.atr.jp/cstar-corpus

[4] Semi-automatic segmentation using a tool provided by NLPR, a member of the C-STAR consortium, http://nlpr-web.ia.ac.cn/english/index.html

[5] Automatic segmentation using the CHASEN tool, http://chasen.naist.jp

[6] Linguistic Data Consortium, http://www.ldc.upenn.edu/

Table 5 gives an overview of the kinds of linguistic resources permitted (√) or not-permitted (×) for each data set condition.

Table 5: Permitted linguistic resources

Resources	Data Track		
	Small	Additional	Unrestricted
IWSLT04 corpus	√	√	√
LDC resources	×	√	√
tagger	×	√	√
chunker	×	√	√
parser	×	√	√
external bilingual dictionaries	×	×	√
other resources	×	×	√

2.3. Evaluation Specifications

In contrast to the TIDES[7] series of MT evaluations, the objective of the IWSLT workshops is the evaluation of speech translation technologies. In this framework, orthographic features are less relevant and therefore ignored in the evaluation of the MT output results. The evaluation parameters used for automatic and subjective evaluation are as follows:

- *case-insensitive*, i.e., lower-case only
- *no punctuation marks*, i.e., remove '.' ',' '?' '!' '"'
- *no word compounds*, i.e., replace '-' with blank space
- *spelling-out of numerals*

No text pre-processing was carried out, i.e., the participants were responsible for providing their translation output in agreement with the above mentioned evaluation specifications. In the case of a sentence count mismatch or the existence of non-ASCII characters (source language words that were not translated) in the English translation output, the run submission was rejected and no evaluation was carried out.

2.4. Subjective Evaluation

Previous competitive MT evaluations, like the series of DARPA MT evaluations in the mid 1990's [3], evaluated machine translation output with human reference translations on the basis of *fluency* and *adequacy* [4]. *Fluency* refers to the degree to which the translation is well-formed according to the grammar of the target language. *Adequacy* refers to the degree to which the translation communicates the information present in the reference output. The fluency and adequacy judgments consist of one of the grades listed in Table 6.

This evaluation methodology was adopted for the IWSLT04 workshop. The grading assignments for each grader were split into two parts. First, the MT output was displayed and the grader had to judge the *fluency* of the translation. In the second step, a reference translation was given and the grader had to evaluate the *adequacy* of the translation. In order to minimize grading inconsistencies between graders due to contextual misinterpretations of the translations, the situation in which the sentence is uttered (corpus annotations like "sightseeing" or "restaurant") was provided for the adequacy judgment.

Table 6: Human assessment

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

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

Table 7: Workload of graders

	1st grader	2nd grader	3rd grader	# input data
Team 1	G0	G2	G9	200
Team 2	G4	G5	G8	160
Team 3	G1	G3	G6	80
Team 4	G0	G3	G7	60

Each translation of a single MT system was evaluated by three judges. However, in order to minimize the costs of subjective evaluation, all translation results were *pooled*, i.e., in the 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. When the MT engines failed to output any translation for a given input, a score of 0 was assigned to the empty output.

In total, ten English native speakers were involved in the evaluation task, where each grader had to evaluate the output of all MT systems for a certain number of source sentences as summarized in Table 7.

In order to validate the reliability of each grader, two additional evaluation data sets were prepared. The first (*common*) data set was used to compare the grading differences between graders. It consisted of 100 sentences randomly selected from all MT outputs submitted for subjective evaluation. The common data set was evaluated by all human graders. The second (*grader-specific*) data set was used to validate the self-consistency of each grader, who had to evaluate 100 sentences randomly selected from the subset of MT outputs assigned to him or her a second time.

2.5. Automatic Evaluation

Various automatic scoring metrics have been proposed within the MT evaluation community. For the IWSLT04 workshop, we utilized the five metrics summarized in Table 8.

Excluding NIST, the scores of all automatic evaluation metrics are in the range of [0,1]. NIST is always positive, and its scoring range does not have a theoretical upper limit. In contrast to mWER and mPER, higher BLEU, NIST, and GTM scores indicate better translations. For the BLEU/NIST

[7] http://www.ldc.upenn.edu/Projects/TIDES/tidesmt.html

Table 8: Automatic evaluation metrics

mWER: *Multiple Word Error Rate*: the edit distance between the system output and the closest reference translation [5]

mPER: Position independent mWER: a variant of mWER that disregards word ordering [6]

BLEU: the geometric mean of n-gram precision by the system output with respect to reference translations [7]

NIST: a variant of BLEU using the arithmetic mean of weighted n-gram precision values [8]

GTM: measures the similarity between texts by using a unigram-based F-measure [9]

(v11a)[8] and GTM (v1.2)[9] scores, the software versions indicated in parentheses were used.

The translation output was compared with up to 16 reference translations that were pre-processed in order to conform with the format required by the evaluation specifications described in Section 2.3.

Before comparing the data sets, the MT system output and the reference translations were tagged by using a publicly available part-of-speech tagger[10].

All scoring scripts were applied, and the results were sent back automatically to the participants via email using the IWSLT04 evaluation server[11].

2.6. Evaluation Campaign Participants

In total, 14 institutions took part in the Evaluation Campaign 2004 using a large variety of translation methodologies (cf. Table 9). The most frequently entered types were *statistical machine translation* (SMT) engines (7), but *example-based* (EBMT) systems (3) and one *rule-based* (RBMT) translation system were also entered. Moreover, four institutions exploited hybrid MT approaches that combined corpus-based MT, translation memories, and interlingua approaches.

Table 9: MT system types

SMT	7	ATR-SMT, IBM, IRST, ISI, ISL-SMT, RWTH, TALP	
EBMT	3	HIT, ICT, UTokyo	
RBMT	1	CLIPS	
Hybrid	4	ATR-HYBRID	(SMT+EBMT)
		IAI	(SMT+TM)
		ISL-EDTRL	(SMT+IF)
		NLPR	(RBMT+TM)

For CE, 13 participants submitted 20 MT systems, and ten MT systems were submitted by six participants for the JE translation task (cf. Table 10).

In total, 11,134 translations (after pooling) from 28 MT systems had to be evaluated. To summarize the evaluation results, we assigned an ID to each MT system as listed in Table 11.

Table 10: MT system submissions

Data Track	**CE**	**JE**
Small	9	4
Additional	2	–
Unrestricted	9	4
Organization	13	6

Table 11: MT system ID

CE

Small		**Additional**		**Unrestricted**	
ATR-SMT (ATR-S)	: [10]	IRST	: [14]	CLIPS	: [19]
HIT	: [11]	ISI	: [15]	HIT	: [11]
IAI	: [12]			IBM	: [13]
IBM	: [13]			ICT	: [20]
IRST	: [14]			IRST	: [14]
ISI	: [15]			ISI	: [15]
ISL-SMT (ISL-S)	: [16]			ISL-EDTRL (ISL-E)	: [21]
RWTH	: [17]			ISL-SMT (ISL-S)	: [16]
TALP	: [18]			NLPR	: [22]

JE

Small		**Unrestricted**	
ATR-SMT (ATR-S)	: [10]	ATR-HYBRID (ATR-H)	: [10]
IBM	: [13]	CLIPS	: [19]
ISI	: [15]	RWTH	: [17]
RWTH	: [17]	UTokyo	: [23]

2.7. Evaluation Campaign Schedule

The schedule of the evaluation campaign is summarized in Table 12. The training corpus of the supplied IWSLT04 corpus was released three months in advance of the official test runs. The participants were able to validate their system performance one week ahead by submitting translation results of the development data set by using the automatic evaluation server.

Table 12: Evaluation campaign schedule

Event	Date
Evaluation Specifications	February 15, 2004
Application Submission	April 15, 2004
Notification of Acceptance	April 30, 2004
Sample Corpus Release	May 7, 2004
Training Corpus Release	May 21, 2004
Development Corpus Release	July 15, 2004
Evaluation Server Online	August 1, 2004
Test Corpus Release	August 9, 2004
Run Submission	August 12, 2004
Result Feedback to Participants	September 10, 2004
Camera-ready Paper Submission	September 17, 2004
Workshop	September 30 - October 1, 2004

The official test run period was limited to three days, during which the automatic scoring result feedback from the evaluation server to the participant via email was switched off in order to avoid any system tuning with the test set data.

After the official test run period, the participants still had access to the evaluation server in order to try out new ideas and compare their effectiveness toward their own official test run results (automatic scoring only). In addition, the par-

[8] http://www.nist.gov/speech/tests/mt/resources/scoring.htm

[9] http://nlp.cs.nyu.edu/GTM/

[10] http://www.cs.jhu.edu/∼brill/RBT1_14.tar.Z

[11] https://www.slt.atr.jp/EVAL/interface

Table 13: Expected difference between two assessments when each translation was evaluated twice by the same grader.

Grader ID	Fluency	Adequacy
G0	0.21	0.33
G1	0.37	0.39
G2	0.35	0.44
G3	0.49	0.38
G4	0.34	0.34
G5	0.22	0.44
G6	0.77	0.64
G7	0.29	0.44
G8	0.44	0.44
G9	0.46	0.55
Average	0.39	0.44

ticipants had access to an extended version[12] of the evaluation server that allowed them to select specific evaluation parameters and validate the robustness of their MT systems for different evaluation specifications. Please refer to the participants MT system descriptions in the IWSLT04 workshop proceedings for details on their findings.

3. Evaluation Results

3.1. Subjective Evaluation Results

This section reports results on subjective evaluations with regards to the following points:

- How consistently did each grader evaluate translations?
- How consistently did a group of three graders evaluate them?
- How were MT systems ranked according to subjective evaluation?

The variance among subjective evaluations was caused by the variance within graders (intra-grader variance) and the variance between graders (inter-grader variance). Thus, after analyzing these variances, the MT systems were ranked according to subjective evaluation with regard to the analysis.

3.1.1. Self-consistency of Graders

This section shows the results of checking the self-consistency of subjective evaluation for each grader. For this purpose, the graders evaluated 100 randomly selected translations[13] twice. This checking was scheduled so that the second assessments did not follow the first assessments. For each grader, the average difference between the first and second grades was calculated, and the rate of the first and second grades being different was also calculated. Table 13 shows the expected difference between the two assessments for each grader. Table 14 shows the error rate for each grader.

Table 14: Error rate of each grader in the same trial as Table 13

Grader ID	Fluency	Adequacy
G0	0.19	**0.23**
G1	0.33	0.34
G2	0.32	0.34
G3	0.47	0.33
G4	0.26	0.32
G5	**0.14**	0.40
G6	0.53	0.55
G7	0.28	0.39
G8	0.37	0.37
G9	0.33	0.37
Average	0.322	0.364

The expected differences of fluency and adequacy ranged from 0.21 to 0.77 and from 0.33 to 0.64, respectively, which were around 0.4 on average. This indicates that the quality of two MT systems whose difference in either fluency or adequacy is less than 0.8 cannot be distinguished. In other words, we cannot judge which MT system is better by this subjective evaluation results by regarding individual grader's error.

As shown in Table 14, the error rates were considerably larger than expected. Even in the smallest case, indicated by bold-face figures, they were around 20%.

Subjective evaluation is a classification task. If we merge two classes that are difficult to distinguish, we can reduce the error rate in practice. To reduce the error rate, the authors considered four binary classifications as follows:

- “5” versus “less than 5”,
- “larger than or equal to 4” versus “less than 4”,
- “larger than or equal to 3” versus “less than 3”, and
- “larger than or equal to 2” versus “less than 2”.

Tables 15 and 16 show the error rates of the above binary classifications in fluency and adequacy, respectively. The error rates of the binary classifications in fluency and adequacy (cf. Tables 15 and 16) were much smaller than those of the 5-grade classification (cf. Table 14). The minimal error rates of the binary classifications ranged from 0.01 to 0.07.

3.1.2. Consistency of Median Grades

This section shows the results of checking consistency among the median grades of three graders. For this check, 100 translations were randomly selected from all MT outputs submitted for subjective evaluation and evaluated by all human graders. For each pair of teams for three graders as shown in Table 7, the average of differences between median grades was calculated. Table 17 shows the expected difference between two median grades. The expected differences of fluency and adequacy ranged from 0.44 to 0.75 and from 0.34 to 0.61, respectively, which were around 0.55 on average.

This indicates that the performance of two MT systems whose difference in either fluency or adequacy is less than

[12] https://www.slt.atr.jp/EVAL

[13] Note that the 100 translations were different for each grader.

Table 15: Error rate of binary fluency classifications

Grader's ID	5 or less	≥ 4 or not	≥ 3 or not	≥ 2 or not
G0	**0.01**	0.07	**0.06**	0.07
G1	0.05	0.10	0.15	0.07
G2	0.08	**0.03**	0.13	0.11
G3	0.12	0.06	0.23	0.08
G4	0.07	0.07	0.09	0.11
G5	0.05	0.09	**0.06**	**0.02**
G6	0.11	0.17	0.30	0.19
G7	0.04	0.09	0.13	0.03
G8	0.09	0.06	0.19	0.10
G9	0.11	0.06	0.18	0.11
Average	0.073	0.08	0.152	0.089

Table 16: Error rate of binary adequacy classifications

Grader's ID	5 or less	≥ 4 or not	≥ 3 or not	≥ 2 or not
G0	0.08	0.10	**0.07**	0.08
G1	0.17	0.10	0.08	**0.04**
G2	0.07	0.13	0.13	0.11
G3	**0.05**	0.13	0.16	**0.04**
G4	0.09	**0.06**	0.08	0.11
G5	0.11	0.11	0.10	0.12
G6	0.07	0.22	0.27	0.08
G7	0.08	0.10	0.16	0.10
G8	0.11	0.13	0.15	0.05
G9	0.13	0.13	0.15	0.14
Average	0.096	0.121	0.135	0.087

1.1 cannot be distinguished. thus we cannot judge which MT system is better by these subjective evaluation results without giving consideration to individual grader's error.

3.1.3. Ranking MT Systems

From the discussions in Sections 3.1.1 and 3.1.2, the authors show two types of ranking lists according to the average grades and the ratios of 5-grade translations. The error rate of the first binary classification described in Section 3.1.1, "5" versus "less than 5", is the smallest on average among the four binary classifications. For more reliable ranking, a ranking based on the first binary classification was additionally calculated.

The first ranking lists are typically used in MT system comparisons, which are hereafter called **the regular ranking lists**. The scores in this ranking are in the range of [0,5]. Higher scores indicate that the corresponding MT systems are better, but they are not necessarily useful for our analysis because the self-consistency of each grader was low; here they are given for comparison purposes only.

Therefore, in addition to the regular ranking of MT systems, we conducted an alternative ranking according to ratios of 5-grade translations. In this ranking, we used assessments by the grader whose error rate was the smallest among three graders. Table 18 shows the error rate for this ranking. The values whose head is "Total" are the weighted average of the error rates for Teams 1 to 4, where the weights are the number of source sentences assigned to each team (cf. Table 18). The second ranking lists are more reliable, and these are hereafter called **the alternative ranking lists**. The scores in this ranking are in the range of [0,1]. Higher scores indicate that the corresponding MT systems are better.

Table 17: Expected difference between two medians of three graders by team

	Fluency			Adequacy		
	T2	T3	T4	T2	T3	T4
Team 1 (T1)	0.49	0.75	0.47	0.54	0.61	0.34
Team 2 (T2)	–	0.68	0.66	–	0.59	0.48
Team 3 (T3)	–	–	0.44	–	–	0.51
Ave.	0.58			0.51		

Table 18: Error rate of assessments by grader with the smallest error rate among three graders

	Fluency	Adequacy
Team 1	0.01	0.07
Team 2	0.05	0.09
Team 3	0.05	0.05
Team 4	0.01	0.05
Total	0.03	0.07

Note that the regular ranking lists are based on the medians, each of which is a median among the grades by three graders, while the alternative ranking lists are based on the grade assigned by the grader with the smallest error rate.

Table 19 shows the regular ranking lists and the alternative ranking lists. In some tracks, a line is found between two MT system IDs. In the regular ranking, this indicates that the difference between scores above the line is within twice the value of the corresponding expected difference on average, 0.58 for fluency or 0.51 for adequacy as shown in Table 17. In the alternative ranking, this indicates that the difference between scores above the line is within twice the value of the corresponding error rate in total, 0.03 for fluency or 0.07 for adequacy as shown in Table 18. For the CE supplied track in the regular ranking in Table 19 , for example, a line is found between the MT's ID in the first place and the MT's ID in the last place in the second and third columns. With regards to fluency, we cannot judge which MT system is better than the others among the top five MT systems. On the other hand, no line is found between the MT's ID in the first place and the MT's ID in the last place in the forth and fifth columns. With regards to adequacy, we cannot judge which MT system is better than the others among all the MT systems on this track. Moreover, in some tracks, we can find a asterisk mark (∗): The score of a marked MT system is not significantly less than that of the MT system placed in the first position, which was calculated according to a t-test based on 5-fold cross validation. Hereafter, all scores of either adequacy or fluency are compared with the highest scores of each track.

A summary of the alternative ranking for each track is as follows: in the CE supplied track, the second best MT system has a fluency score that is twice smaller than the expected difference of 0.03; and all MT systems except the last have adequacy scores within twice the value of the expected difference of 0.07. In the CE additional track, the second best MT system has a fluency score that is twice smaller than the expected difference; and all MT systems have adequacy scores within twice the value of the expected difference. In the CE unrestricted track, the two best MT systems have fluency and adequacy scores within twice the value of the expected difference. In the JE supplied track, the second best MT systems has a fluency score that is twice smaller than the expected difference; and all MT systems except the last have adequacy scores within twice the value of the expected difference. In the JE unrestricted track, the second best MT system has fluency score that is twice smaller than the expected difference; and the best two MT systems have adequacy scores within twice the value of the expected difference. A summary of the regular ranking for each track is omitted because it is easy for the readers to follow them.

Although the objective of this paper is not to discuss the superiority of MT systems, for convenience, the authors briefly summarize the common tendencies with regular and alternative rankings and some distinctive observations as follows: (1) SMT or Hybrid MT systems were ranked in the upper-half positions, that is, ATR-SMT, ISL-SMT, ISI, IRST, and RWTH for the CE supplied track; IRST for the CE additional track; IRST, and ISL-SMT for the CE unrestricted track; ATR-SMT, ISI, and RWTH for the JE supplied track; and ATR-Hybrid and RWTH for the JE unrestricted track. (2) Most MT system were better than only one RBMT. (3) For adequacy, ATR-SMT was ranked in a lower-half position in the regular ranking, but it was ranked in an upper-half position in the alternative ranking.

3.2. Automatic Evaluation Results

Table 20 shows the ranking lists according to the automatic evaluation metrics. Asterisk marks ($*$) in this table denote the same status as in Table 19 (insignificant difference between the marked MT system and the best MT system).

A summary of ranking lists on each track according to mWER is as follows: In the CE supplied track, only the second and third best MT systems in mWER have scores that are not significantly different from that of the best MT system. In the CE additional track, the second best or worse MT systems have mWER-scores that are significantly inferior to that of the best MT system. In the remaining three tracks, the results are the same as in the CE additional track. The summaries of ranking lists on each track according to the remaining automatic evaluation metrics are omitted because it is easy for the readers to follow them.

As with the subjective evaluation results, the authors briefly summarize the common tendencies in ranking according to the automatic evaluation metrics as follows: (1) SMT and Hybrid MT systems were ranked in the upper-half positions, that is, ATR-SMT, RWTH, IRST, ISI, and ISL-SMT for the CE supplied track; IRST for the CE additional track; IBM, IRST, ISL-SMT, and ISL-EDTRL for the CE unrestricted track; RWTH, and ISI for the JE supplied track; and ATR-Hybrid, RWTH for the CE unrestricted track. (2) Some SMT systems, including ISL-SMT and RWTH, were ranked in the best or second best position in the ranking lists corresponding to almost of all the automatic evaluation metrics, even if they might have been optimized with a particular automatic evaluation metric.

3.3. Correlation between Subjective and Automatic Evaluation Results

Table 21 shows the correlation co-efficients between subjective and automatic evaluation results. (A) shows the correlation co-efficients between average grades for either fluency or adequacy and automatic evaluation scores; (B) and (C) show the correlation co-efficients between ratios of 5-grade translations for either fluency or adequacy and automatic evaluation scores.

(A) and (B) are the results based on either all CE MT systems or all JE MT systems; (C) is the results based on either partial CE MT systems or partial JE MT systems, which were selected such that the differences in the their scores were twice larger than the error rates. A partial version of (A) was not be calculated because the number of the remaining MT system was two or three.

As shown in the upper result of (A), BLEU is the automatic evaluation metric most closely correlated to average grades of fluency for CE or JE MT systems. Thus, BLEU is the most promising automatic evaluation metric according to average grades of fluency.

As shown in the lower result of (A), NIST is the automatic evaluation metric most closely correlated to the average grades of adequacy for CE or JE MT systems. Thus, NIST is the most promising automatic evaluation metric according to average grades of adequacy.

As shown in the upper result of (B) and (C), BLEU is the automatic evaluation metric most closely correlated to ratios of 5-grade translations in fluency for CE or JE MT systems. Thus, BLEU is the most promising automatic evaluation metric according to ratios of 5-grade translations in fluency.

As shown in the lower result of (B), BLEU is the automatic evaluation metric most closely correlated to ratios of 5-grade translations in adequacy for CE MT systems but only the fourth most closely correlated to ratios of 5-grade translations in adequacy for JE MT systems. mWER is the automatic evaluation metric most closely correlated to ratios of 5-grade translations in adequacy for JE MT systems but as well as the third most closely correlated to ratios of 5-grade translations in adequacy for CE MT systems. On the other hand, as shown in the lower result of (C), mPER is the automatic evaluation metric most closely correlated to ratios of 5-grade translations in adequacy for CE or JE MT systems. Considering these observations, we could not say which is best, the lower result of (B) or (C). Therefore, from these results, we could not judge which automatic evaluation metric is the most promising in regard to ratios of 5-grade translations in adequacy.

4. Discussion

Various problems of subjective evaluation (fluency, adequacy) were found through this evaluation campaign. A summary of key findings is given as follows:

- The grade description is ambiguous. As a result, the interpretation of grades for fluency or adequacy depended on each grader.
- Self-consistency of a grader's subjective evaluation results was poor due to the way the translations to be evaluated were displayed or to the way the graders were selected.
- The variance of median grades by multiple graders was large due to the way the graders were selected. The number of graders is required to be as small as possible.
- Some single reference translations shown to graders for adequacy evaluation were ambiguous, which resulted in the interpretations of some reference translations being dependent to graders.
- (adequacy) Sometimes the translation has MORE information than the reference translation. If the information was of considerable importance or essential, the lowest rank was assigned.
- (adequacy) Sometimes the English is OK in part of the sentence but had incomprehensible parts as well. If the incomprehensible parts came at the beginning of the sentence, the graders often gave the adequacy a higher score because the sentence ended with something understandable. However, if the sentence ended with something incomprehensible, lower rankings were given. If it came in the middle (e.g., "do you have red coffee or tea") it could go either way.
- (adequacy) If numerical expressions were not translated exactly, frequently the lowest grade ("None of it") was assigned.
- (fluency) The rank of very short utterances such as "Okay" depends on the word and the situation context.
- (fluency) Sometimes the target sentence was flawless, but the meaning of the translation and the reference translations were completely different (highest rank for fluency and lowest for adequacy).

We need subjective evaluation results that are as error-free as possible to further promote automatic evaluation research. The authors will discuss the above problems in this workshop.

5. Acknowledgments

The research reported here was supported in part by the National Institute of Information and Communications Technology project entitled *"A study of speech dialogue translation technology based on a large corpus"* (ATR), the project *FU-PAT WebFaq* funded by the Province of Trento (IRST), and the Nature Science Foundation of China under grant number 60175012 and 6037 5018 (NLPR).

The IWSLT 2004 organizing committee would like to thank the organizing committees of ICSLP 2004 for their support and for kindly providing the template files. Moreover, we would like to thank the C-STAR partners for their accomplishments during the subjective evaluation task.

6. References

[1] M. Paul, H. Nakaiwa, and M. Federico, "Towards innovative evaluation methodologies for speech translation," in *Working Notes of the NTCIR-4 2004 Meeting, Supplement Volume 2*, Tokyo, Japan, 2004, pp. 17–21.

[2] 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.

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

[4] *Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Chinese-English Translations Revision 1.0*, Linguistic Data Consortium, 2002, http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess02.pdf.

[5] 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.

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

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

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

[9] 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.

[10] E. Sumita, Y. Akiba, T. Doi, A. Finch, K. Imamura, H. Okuma, M. Paul, M. Shimohata, and T. Watanabe, "EBMT, SMT, Hybrid and More: ATR spoken language translation system," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 13–20.

[11] M. Yang, T. Zhao, H. Liu, X. Shi, and H. Jiang, "Auto word alignment based Chinese-English EBMT," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 27–29.

[12] P. Langlais, M. Carl, and O. Streiter, “Experimenting with phrase-based statistical translation within the IWSLT 2004 Chinese-to-English Shared Translation Task,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 31–38.

[13] Y.-S. Lee and S. Roukos, “IBM spoken language translation system evaluation,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 39–46.

[14] N. Bertoldi, R. Cattoni, M. Cettolo, and M. Federico, “The ITC-irst statistical machine translation system for IWSLT-2004,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 51–58.

[15] E. Ettelaie, K. Knight, D. Marcu, D. S. Munteanu, F. J. Och, I. Thayer, and Q. Tipu, “The ISI/USC MT system,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, p. 59.

[16] S. Vogel, S. Hewavitharana, M. Kolss, and A. Waibel, “The ISL statistical translation system for spoken language translation,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 65–72.

[17] O. Bender, R. Zens, E. Matusov, and H. Ney, “Alignment templates: the RWTH SMT system”,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 79–84.

[18] A. D. GISPERT and J. B. MARINO, “TALP: Xgram-based spoken language translation system,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 85–90.

[19] H. Blanchon, C. Boitet, F. Brunet-Manquat, M. Tomokiyo, A. Hamon, V. T. Hung, and Y. Bey, “Towards fairer evaluations of commercial MT systems on Basic Travel Expressions Corpora,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 21–26.

[20] H. Hou, D. Deng, G. Zou, H. Yu, Y. Liu, D. Xiong, and Q. Liu, “An EBMT system based on word alignment,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 47–49.

[21] J. Reichert and A. Waibel, “The ISL EDTRL system,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 61–64.

[22] Y. Zuo, Y. Zhou, and C. Zong, “Multi-engine based Chinese-to-English translation system,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 73–77.

[23] E. Aramaki and S. Kurohashi, “Example-based machine translation using structual translation examples,” in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 91–94.

Table 19: Ranking lists according to subjective evaluation results

Chinese-to-English (CE)

Regular ranking

Track		Fluency		Adequacy	
		Score	MT_ID	Score	MT_ID
CE	U	3.776	^{s}IRST	3.662	^{s}ISL-S
		3.776	^{s}ISL-S*	3.526	^{s}IRST*
		3.400	^{h}NLPR	3.254	^{h}ISL-E
		3.036	^{s}IBM	3.188	^{e}HIT
		2.954	^{s}ISI	3.082	^{e}ICT
		2.934	^{h}ISL-E	2.996	^{s}IBM
		2.718	^{e}ICT	2.960	^{r}CLIPS
		2.648	^{e}HIT	2.800	^{h}NLPR
		2.570	^{r}CLIPS	2.784	^{s}ISI
	A	3.256	^{s}IRST	3.110	^{s}IRST
		2.846	^{s}ISI	2.724	^{s}ISI
	S	3.820	^{s}ATR-S	3.338	^{s}RWTH
		3.356	^{s}RWTH	3.088	^{s}IRST
		3.332	^{s}ISL-S	3.084	^{s}ISI
		3.120	^{s}IRST	3.056	^{e}HIT
		3.074	^{s}ISI	3.048	^{s}ISL-S
		2.948	^{s}IBM	3.022	^{s}TALP
		2.914	^{h}IAI	2.950	^{s}ATR-S
		2.792	^{s}TALP	2.938	^{h}IAI
		2.504	^{e}HIT	2.906	^{s}IBM

Alternative ranking

Track		Fluency		Adequacy	
		Score	MT_ID	Score	MT_ID
CE	U	0.558	^{s}IRST	0.446	^{s}ISL-S
		0.532	^{s}ISL-S*	0.394	^{s}IRST
		0.406	^{h}NLPR	0.294	^{h}ISL-E
		0.326	^{s}IBM	0.258	^{e}ICT
		0.296	^{h}ISL-E	0.250	^{s}IBM
		0.286	^{s}ISI	0.228	^{h}NLPR
		0.224	^{e}HIT	0.226	^{e}HIT
		0.222	^{e}ICT	0.178	^{s}ISI
		0.180	^{r}CLIPS	0.164	^{r}CLIPS
	A	0.410	^{s}IRST	0.316	^{s}IRST
		0.284	^{s}ISI	0.212	^{s}ISI
	S	0.582	^{s}ATR-S	0.338	^{s}RWTH
		0.420	^{s}ISL-S	0.296	^{s}ATR-S*
		0.390	^{s}RWTH	0.290	^{s}ISL-S*
		0.356	^{s}IRST	0.284	^{s}IRST*
		0.344	^{s}ISI	0.268	^{s}ISI
		0.314	^{s}IBM	0.258	^{h}IAI
		0.278	^{h}IAI	0.250	^{s}TALP
		0.246	^{s}TALP	0.232	^{s}IBM
		0.186	^{e}HIT	0.196	^{e}HIT

Japanese-to-English (JE)

Regular ranking

Track		Fluency		Adequacy	
		Score	MT_ID	Score	MT_ID
JE	U	4.308	^{h}ATR-H	4.208	^{h}ATR-H
		4.036	^{s}RWTH	4.066	^{s}RWTH
		3.650	^{e}UTokyo	3.316	^{e}UTokyo
		2.472	^{r}CLIPS	2.602	^{r}CLIPS
	S	3.484	^{s}ATR-S	3.412	^{s}RWTH
		3.480	^{s}RWTH*	3.086	^{s}ISI
		3.106	^{s}IBM	2.990	^{s}IBM
		3.102	^{s}ISI	1.942	^{s}ATR-S

Alternative ranking

Track		Fluency		Adequacy	
		Score	MT_ID	Score	MT_ID
JE	U	0.698	^{h}ATR-H	0.600	^{h}ATR-H*
		0.608	^{s}RWTH	0.564	^{s}RWTH
		0.506	^{e}UTokyo	0.360	^{e}UTokyo
		0.170	^{r}CLIPS	0.120	^{r}CLIPS
	S	0.520	^{s}ATR-S	0.358	^{s}RWTH
		0.440	^{s}RWTH	0.304	^{s}ISI
		0.368	^{s}ISI	0.262	^{s}IBM
		0.334	^{s}IBM	0.126	^{s}ATR-S

$s,\ e,\ r,\ or\ h$ appended to MT_ID indicates that the MT system is SMT, EBMT, RBMT, or Hybrid MT, respectively.

Table 20: Ranking lists according to automatic evaluation results

Chinese-to-English (CE)

Track		mWER		mPER		BLEU		NIST		GTM	
		Score	MT_ID	Score	MT_ID	Score	MT_ID	Score	MT_ID	Score	MT_ID
CE	U	0.379	^{s}ISL-S	0.319	^{s}ISL-S	0.524	^{s}ISL-S	9.56	^{s}ISL-S	0.748	^{s}ISL-S
		0.457	^{s}IRST	0.393	^{s}IRST	0.440	^{s}IRST	7.50	^{h}ISL-E	0.684	^{s}IBM
		0.525	^{s}IBM	0.427	^{h}ISL-E	0.350	^{s}IBM	7.36	^{s}IBM	0.671	^{s}IRST
		0.531	^{h}ISL-E	0.442	^{s}IBM	0.311	^{h}NLPR	7.24	^{s}IRST	0.666	^{h}ISL-E
		0.573	^{s}ISI	0.487	^{e}HIT	0.275	^{h}ISL-E	6.13	^{e}HIT	0.611	^{e}HIT
		0.578	^{h}NLPR	0.499	^{s}ISI	0.243	^{e}HIT	6.00	^{r}CLIPS	0.602	^{s}ISI
		0.594	^{e}HIT	0.531	^{h}NLPR	0.243	^{s}ISI	5.92	^{h}NLPR	0.584	^{r}CLIPS
		0.658	^{r}CLIPS	0.542	^{r}CLIPS	0.162	^{r}CLIPS	5.42	^{s}ISI	0.563	^{h}NLPR
		0.846	^{e}ICT	0.765	^{e}ICT	0.079	^{e}ICT	3.64	^{e}ICT	0.386	^{e}ICT
	A	0.496	^{s}IRST	0.420	^{s}IRST	0.351	^{s}IRST	7.39	^{s}IRST	0.655	^{s}IRST
		0.572	^{s}ISI	0.480	^{s}ISI	0.311	^{s}ISI	5.82	^{s}ISI	0.632	^{s}ISI
	S	0.455	^{s}RWTH	0.390	^{s}RWTH	0.454	^{s}ATR-S	8.55	^{s}RWTH	0.720	^{s}RWTH
		0.469	^{s}ATR-S*	0.404	^{s}ISL-S*	0.414	^{s}ISL-S	8.34	^{s}ISL-S*	0.694	^{s}ISL-S
		0.471	^{s}ISL-S*	0.420	^{s}ATR-S	0.408	^{s}RWTH	7.85	^{h}IAI	0.685	^{h}IAI
		0.488	^{s}ISI	0.425	^{s}ISI	0.374	^{s}ISI	7.74	^{s}ISI	0.672	^{s}ISI
		0.507	^{s}IRST	0.430	^{s}IRST	0.349	^{s}IRST	7.48	^{s}ATR-S	0.670	^{s}ATR-S
		0.532	^{h}IAI	0.451	^{h}IAI	0.346	^{s}IBM	7.12	^{s}IBM	0.665	^{s}IBM
		0.538	^{s}IBM	0.452	^{s}IBM	0.338	^{h}IAI	7.09	^{s}IRST	0.647	^{s}TALP
		0.556	^{s}TALP	0.465	^{s}TALP	0.278	^{s}TALP	6.77	^{s}TALP	0.644	^{s}IRST
		0.616	^{e}HIT	0.500	^{e}HIT	0.209	^{e}HIT	5.95	^{e}HIT	0.601	^{e}HIT

Japanese-to-English (JE)

Track		mWER		mPER		BLEU		NIST		GTM	
		Score	MT_ID	Score	MT_ID	Score	MT_ID	Score	MT_ID	Score	MT_ID
JE	U	0.263	^{h}ATR-H	0.233	^{h}ATR-H	0.630	^{h}ATR-H	11.25	^{s}RWTH	0.824	^{s}RWTH
		0.305	^{s}RWTH	0.249	^{s}RWTH*	0.619	^{s}RWTH*	10.72	^{h}ATR-H*	0.796	^{h}ATR-H
		0.485	^{e}UTokyo	0.420	^{e}UTokyo	0.397	^{e}UTokyo	7.88	^{e}UTokyo	0.672	^{e}UTokyo
		0.730	^{r}CLIPS	0.597	^{r}CLIPS	0.132	^{r}CLIPS	5.64	^{r}CLIPS	0.568	^{r}CLIPS
	S	0.418	^{s}RWTH	0.337	^{s}RWTH	0.453	^{s}RWTH	9.49	^{s}RWTH	0.764	^{s}RWTH
		0.484	^{s}ISI	0.379	^{s}ISI	0.400	^{s}ISI	8.46	^{s}ISI	0.732	^{s}ISI
		0.527	^{s}IBM	0.430	^{s}IBM	0.366	^{s}IBM	7.97	^{s}IBM	0.698	^{s}IBM
		0.614	^{s}ATR-S	0.570	^{s}ATR-S	0.364	^{s}ATR-S	3.41	^{s}ATR-S	0.539	^{s}ATR-S

$s,\ e,\ r,\ or\ h$ appended to MT_ID indicates that the MT system is SMT, EBMT, RBMT, or Hybrid MT, respectively.

Table 21: Correlation co-efficients between subjective and automatic evaluation results

(A) Regular ranking vs. automatic ranking (All)

Fluency

	mWER	mPER	**BLEU**	NIST	GTM
CE	-0.7124	-0.5830	**0.8505**	0.5995	0.5132
JE	-0.8867	-0.7836	**0.9404**	0.5995	0.6387

Adequacy

	mWER	mPER	BLEU	**NIST**	GTM
CE	-0.4324	-0.4404	0.4376	**0.5318**	0.3711
JE	-0.8978	-0.9376	0.7884	**0.9701**	0.9401

(B) Alternative ranking vs. automatic ranking (All)

Fluency

	mWER	mPER	**BLEU**	NIST	GTM
CE	-0.7214	-0.6010	**0.8600**	0.5950	0.5214
JE	-0.8252	-0.7032	**0.9070**	0.4871	0.5383

Adequacy

	mWER	mPER	**BLEU**	NIST	GTM
CE	-0.6427	-0.5779	**0.7407**	0.6820	0.5136
JE	**-0.9690**	-0.9641	0.9157	0.9176	0.9152

(C) Alternative ranking vs. automatic ranking (Partial)

Fluency

	mWER	mPER	**BLEU**	NIST	GTM
CE	-0.8734	-0.6743	**0.9548**	0.5736	0.5454
JE	-0.8376	-0.7223	**0.9288**	0.5089	0.5632

Adequacy

	mWER	**mPER**	BLEU	NIST	GTM
CE	-1.0000	**-1.0000**	1.0000	1.0000	1.0000
JE	-0.9894	**-0.9984**	0.9195	0.9907	0.9977