Overview of the IWSLT 2007 Evaluation Campaign

Cameron S. Fordyce

Center for the Evaluation of Language and Communication Technologies, Trento

fordyce@celct.it

Abstract

In this paper we give an overview of the 2007 evaluation campaign for the *International Workshop on Spoken Language Translation* (IWSLT)[1]. As with previous evaluation campaigns, the primary focus of the workshop was the translation of spoken language in the travel domain. This year there were four language pairs; the translation of Chinese, Italian, Arabic, and Japanese into English. The input data consisted of the output of ASR systems for read speech and clean text. The exceptions were the challenge task of the Italian English language pair which used spontaneous speech ASR outputs and transcriptions and the Chinese English task which used only clean text. A new characteristic of this year's evaluation campaign was an increased focus on the sharing of resources. Participants were requested to submit the data and supplementary resources used in building their systems so that the other participants might be able to take advantage of the same resources. A second new characteristic this year was the focus on the human evaluation of systems. Each primary run was judged in the human evaluation for every task using a straightforward ranking of systems. This year's workshop saw an increased participation over last year's workshop. This year 24 groups submitted runs to one or more of the tasks, compared to the 19 groups that submitted runs last year [1]. Automatic and human evaluation were carried out to measure MT performance under each condition, ASR system outputs for read speech, spontaneous travel dialogues, and clean text.

1. Introduction

IWSLT is an MT evaluation campaign organized by the Consortium for Speech Translation Advanced Research (C-Star)[2]. This consortium provides a common framework to compare and improve the state-of-the-art speech-to-speech translation (SST) technologies[1]. C-Star has organized annual workshops with progressively more challenging SST tasks with Japanese, Chinese, Arabic, Italian into English. The 2004 IWSLT workshop focused on evaluation metrics for SST[2]. The 2005 IWSLT focused on the translation of ASR outputs from read-speech inputs[3]. The 2006 IWSLT workshop focused on spontaneous translation of Chinese into English, and the translation of read Japanese, Arabic, and Italian into English[1].

The theme of this year's evaluation campaign remained the same as last year's, the translation of spontaneous-speech input. As with last year, the evaluation tasks were divided into two major groups, two "Challenge" tasks for spontaneous speech and two "Classical" tasks focusing on read-speech. The challenge tasks included the languages Chinese and Italian to English. The Chinese challenge task was structured to mirror last year's CE challenge task. Unfortunately, due to the unavailability of new CE test data at the last moment, clean text was substituted. The Italian to English challenge task marked a departure from the previous year in that the spontaneous speech came from a collection of transcribed dialogues from travel agent and client interactions via telephone.

The classical tasks included read speech for both Japanese to English and Arabic to English translation directions.

Participants were supplied with in-domain resources from several sources. The principal source for training, development, and evaluation data was the *Basic Travel Expression Corpus* (BTEC)[4]. Training and development data was made available from previous editions of the workshop. In addition, the SITAL[5][3] corpus of transcribed travel agent-client dialogues was made available to participants for the Italian to English language pair.

In the previous year's workshop, tasks were further divided in two data tracks, (OPEN, CSTAR)[1]. The primary difference between these two tracks was the possibility of the participants in the CSTAR track to use all the proprietary BTEC data rather than the BTEC data made available to all participants. In order to create a more level field for the comparison of systems, for this year's evaluation campaign it was decided to reduce the possible data conditions to one, the equivalent of an open track. Participants were allowed to use any publically available resource as long as it was affordable. Resources that were proprietary and unable to the general public were strongly discouraged. BTEC data from previous years, both training, development, and previous test sets were made available to this year's participants.

For the evaluation of system submissions, automatic evaluation and human evaluation were carried out. For the automatic metric, BLEU[6], with six references was used for the Japanese, Arabic, and Chinese tasks. For the Italian task, BLEU with four references was used. For the human

[1] http://iwslt07.itc.it

[2] http://www.c-star.org/

[3] The acronym SI-TAL or SITAL is used in during the evaluation campaign. This corpus is also referred to as the ADAM[5] corpus. SI-TAL (Integrated System for the Automatic treatment of Language) was a National Project for the creation of large linguistic resources and software tools for Italian written and spoken language processing.

evaluation, all primary submissions for all tasks were evaluated this year using a ranking system based on work done by Callison-Burch, et al. for the WMT07 shared task[7]. In addition to this approach, NIST adequacy/fluency metric was also applied for three submissions of each of the ASR tasks and for the CE clean task.

2. IWSLT 2007 Evaluation Campaign

2.1. IWSLT 2007 Spoken Language Corpus

This year's evaluation campaign relied on two distinct corpora in the travel domain, the BTEC and the SITAL corpora, a corpus of transcribed spoken Italian. Some additional linguistic resources such as Named-Entity lists were provided by the organizers. As part of the goal of this year's workshop, additional resources such as parallel corpora, linguistic tools, etc. were solicited from participants.

2.1.1. The BTEC Corpus

The BTEC corpus contains data for all the included languages of this year's evaluation campaign. BTEC contains sentences similar to those found in travelers phrase books[4]. The development, and training data has been released in previous campaigns[1, 2, 3]. The test set differed from last year's edition of IWSLT in that the recorded speech prompts came directly from the BTEC corpus rather than the transcripts of semi-spontaneous speech elicited for the Chinese to English challenge task[1]. There were 489 read sentences in this year's test set and each sentence had one canonical translation, with 5 additional translations created by paraphrasing the canonical translation.

2.1.2. The SITAL Corpus

The SITAL corpus consists of recorded simulated interactions between a travel agent and clients of a fictious travel agency in Italian[5]. The interactions consisted mainly of transactions concerning plane, railroad ticket purchases and hotel reservations. The corpus consists of human-human and human-machine interactions. Only recordings of the human-human interactions were used in this workshop. Participants were provided with data for development that included 996 transcribed utterances without case or punctuation information. The test set contained 724 sentences of complete dialogues. The utterances contained transcribed speech events such as repetitions, hesitations, and corrections which make translation very difficult. The utterances contained contiguous dialogues and participants were provided with dialogue boundaries for the development set.

For the development set one reference translation in English was provided. Both the test and development reference translations had punctuation and case information inserted manually. Translators were instructed to disregard some of the speech events, such as repetitions, but corrections were translated into English.

2.1.3. Additional Resources and participant supplied Resources

Some additional resources were provided by the organizers such as a named entity list for the IE challenge task, and scripts to tokenize the translation system output.

In addition, participants were requested to share the resources that were used in the building of their systems. This request reflected one of the main intentions of the workshop which was to foster cooperation in the creation of MT systems[4]. Further, systems were to be built with publically available and reasonably affordable data resources.

Participants did not have to provide resources directly. Nor were participants required to provide resources that they had acquired elsewhere and then modified in some way (i.e. cleaned, corrected, enhanced, etc.). In the latter case, participants were asked to provide a reference to the original provider or creator of the resource.

Acceptable Resources. Some examples of resources that could be used include:

- Publicly available aligned or monolingual corpora such as the EuroParl corpus or LDC data

- Publicly available annotated treebanks.

While the number of participants who contributed resources was not overwhelming, only 7 of 24 groups submitted resources, the list of publically available resources for all the tasks is quite long[5]. Submitted resources include monolingual and parallel corpora as well as treebanks, open source decoders, sentence aligners, and morphological analyzers.

2.2. Input Data Specifications

Two input types were provided this year. ASR system outputs in the form of 1-Best, N-Best lists, and lattices (HTK word lattice format) were provided to the participants for the ASR input task. For the clean data, transcriptions of the read speech was provided. Input data was case-insensitive and without punctuation information.

2.3. Evaluation Specifications

2.3.1.Data Specifications for Submissions

The evaluation specifications for IWSLT 2007 for system outputs follow closely the *official* evaluation specifications for IWSLT06[1], i.e. submitted sentences were to be case-sensitive and with punctuation marks tokenized. No other specifications were considered this year.

[4] See the call for participation, http://iwslt07.itc.it/menu/cfp.pdf.

[5] See http://iwslt07.itc.it/menu/resources.html.

2.3.2. Automatic Evaluation

Participants were asked to submit their runs via a web interface. The first run submitted was considered the “primary” run, or the run that each participant wanted considered for system comparison and for human evaluation. Additional runs could be submitted subsequently and were considered contrastive runs.

The BLEU[6] automatic metric was used to automatically rank systems for each task. For JE, AE classical tasks and for CE, six references were used. For Italian, four reference translations were prepared. The BLEU metric was chosen to measure system performance as it has been shown to correlate with human judgments[12, 13].

After submitting runs, participants were provided with the system rankings for all primary submissions and all other contrastive runs via email. See Tables 6-9 for the primary rankings according to BLEU scores. See Appendix B for a complete ranking of all submissions by BLEU score for each task.

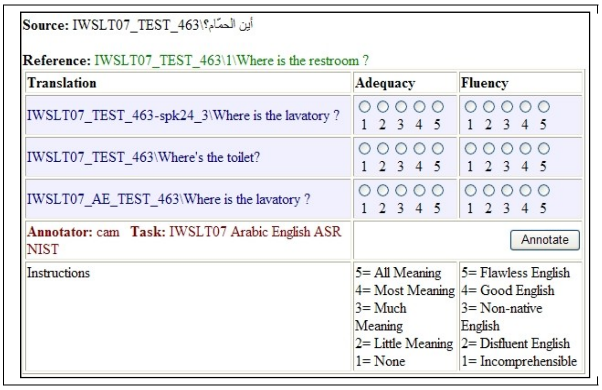
2.3.3. Human Evaluation

A recently introduced human evaluation metric, the ranking of sentences[7], was adopted for this year’s workshop and was applied to all the submitted runs. In addition, the NIST adequacy/fluency subjective evaluation metrics were applied to the top three systems as judged by the automatic metric. The ranking of sentences was used in conjunction with two other approaches during the recent WMT07 shared task[7][6].

Human evaluation of MT systems is typically a time-consuming and expensive endeavor. Many different approaches to the human evaluation of translation have been proposed from reading comprehension tests[9] to subjective scores of adequacy and fluency where adequacy refers generally to the preservation of information and fluency refers generally to the naturalness of the translation[8]. The latter method has been the most widely used for the evaluation of MT system outputs in such evaluation campaigns as the annual NIST Machine Translation Workshops[7].

Each metric has a five-point scale. For adequacy, the five point scale indicates how much of the information expressed in a reference translation is preserved in the system translation. 1 equals no information and 5 equals all information has been preserved. For fluency, a similar scale from 1 to 5 indicates how similar the submitted run is to natural English. Figure 1 presents an example of the adequacy/fluency metric.

These measures were conceived with the goal of obtaining independent measures. In many cases, however, these metrics appear to be highly correlated [7, 10].



Source: IWSLT07_TEST_463\أين الحمّام؟

Reference: IWSLT07_TEST_463\1\Where is the restroom ?

Translation	**Adequacy**	**Fluency**
IWSLT07_TEST_463-spk24_3\Where is the lavatory ?	○ ○ ○ ○ ○ 1 2 3 4 5	○ ○ ○ ○ ○ 1 2 3 4 5
IWSLT07_TEST_463\Where's the toilet?	○ ○ ○ ○ ○ 1 2 3 4 5	○ ○ ○ ○ ○ 1 2 3 4 5
IWSLT07_AE_TEST_463\Where is the lavatory ?	○ ○ ○ ○ ○ 1 2 3 4 5	○ ○ ○ ○ ○ 1 2 3 4 5
Annotator: cam **Task:** IWSLT07 Arabic English ASR NIST		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

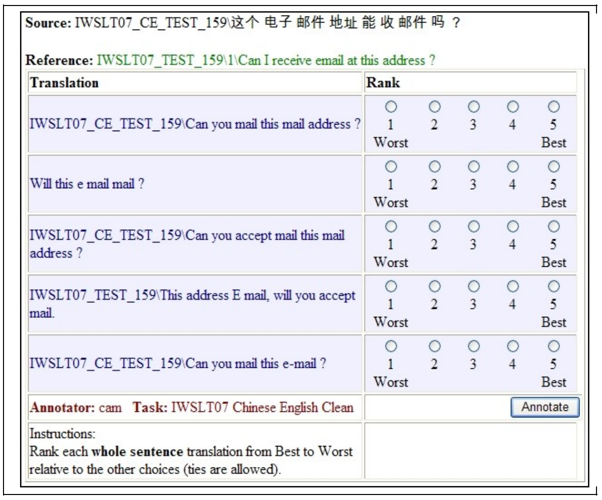
Figure 1 An example of the adequacy/fluency metric for the Arabic task.

2.3.4. Ranking Sentences

When evaluating multiple submitted sentences together using NIST adequacy/fluency, it has been observed that evaluators tend to assign fluency and adequacy scores relative to the other presented sentences[7, 10]. Further, evaluators using this metric often do so without training, which sometimes makes it difficult for them to regard the five-point scales as absolutes.

In the ranking metric, no more than five of the submitted sentences are presented to the evaluator with the source sentence and one reference translation. The evaluator must then rank the sentences from best to worst using a five point scale. Ties between systems are allowed. The system outputs were presented so that each system’s output was presented together with the outputs of all the other systems during the course of the evaluation.

Figure 2 shows the web-based interface for the ranking metric.



Source: IWSLT07_CE_TEST_159\这个 电子 邮件 地址 能 收 邮件 吗 ？

Reference: IWSLT07_TEST_159\1\Can I receive email at this address ?

Translation	**Rank**
IWSLT07_CE_TEST_159\Can you mail this mail address ?	○ ○ ○ ○ ○ 1 2 3 4 5 Worst Best
Will this e mail mail ?	○ ○ ○ ○ ○ 1 2 3 4 5 Worst Best
IWSLT07_CE_TEST_159\Can you accept mail this mail address ?	○ ○ ○ ○ ○ 1 2 3 4 5 Worst Best
IWSLT07_TEST_159\This address E mail, will you accept mail.	○ ○ ○ ○ ○ 1 2 3 4 5 Worst Best
IWSLT07_CE_TEST_159\Can you mail this e-mail ?	○ ○ ○ ○ ○ 1 2 3 4 5 Worst Best
Annotator: cam **Task:** IWSLT07 Chinese English Clean	Annotate
Instructions: Rank each **whole sentence** translation from Best to Worst relative to the other choices (ties are allowed).	

Figure 2 An example of the ranking metric for the Chinese Clean task.

All human judgments for both ranking and adequacy/fluency metrics were collected with a web-based interface. Unlike in [7], the different metrics were not alternated. For each task,

[6] See http://www.statmt.org/wmt07/shared-task.html

[7] See http://www.nist.gov/speech/tests/mt/ for more information on NIST MT evaluations.

systems were evaluated first with the ranking metric and then later with the adequacy/fluency metric. Also, evaluators were specifically assigned tasks to evaluate.

For the classical tasks, 300 sentences from the 489 sentences present in each of the JE, AE, and CE test sets were randomly selected and presented to at least 3 evaluators. Since the ranking metric requires that each submission be compared to the other system outputs, each sentence may be presented multiple times but in the company of different sets of systems.

For the challenge task, 300 sentences from the 724 sentences in the evaluation set were randomly selected after the 724 sentences were pruned of duplicates entries. This resulted in a set of 689 sentences from which the 300 sentences were chosen for the human evaluation.

3. Evaluation Results

3.1. Human Evaluation Results

In this section the results of the human evaluations are presented. For each task and input condition all submissions were evaluated by at least 3 human evaluators with the ranking metric described above. Evaluators included 2 volunteers with experience in evaluating machine translation and 6 paid evaluators who were provided with a brief training in machine translation evaluation.

In the ranking tables, the score is the average number of times that a system was judged to be better than any other system[7].

For the adequacy/fluency measures, only the top three systems for each ASR task and the Chinese English Clean task were evaluated. In order to account for variations in evaluator scoring for adequacy and fluency, the scores were normalized on a per-judge basis as suggested by Blatz et al[11].

3.1.1. System Results

Tables 1 through 4 show the results of the human evaluation using the ranking method. The best score is presented in bold.

IE ASR		**IE Clean**	
SYSTEM	% BETTER	SYSTEM	% BETTER
FBK	**48.5**	FBK	**52.5**
RWTH	42.4	RWTH	50.6
ATR	40.2	ATR	45.9
UEDIN	29.0	MIT	33.1
UW	27.8	NTT	32.5
MIT	24.6	INESCID	28.9
NTT	24.2	HKUST	23.3
RALI	24.2	ITI	19.6
INESCID	18.8	UW	4.0
HKUST	18.4		

Table 1 Human Rankings: IE, ASR and Clean.

JE ASR		**JE Clean**	
SYSTEM	% BETTER	SYSTEM	% BETTER
ATR	**27.3**	CMU	**32.7**
CMU-UKA	26.8	ATR	30.5
UEKAE	24.2	FBK	30.5
NTT	23.5	TOTTORI	28.0
FBK	23.3	UEKAE	27.4
DCU	19.2	NTT	27.3
HKUST	18.3	HKUST	21.9
		DCU	21.2
		GREYC	21.0

Table 2 Human Rankings: JE, ASR and Clean.

AE Clean		**AE ASR**	
SYSTEM	% BETTER	SYSTEM	% BETTER
DCU	**45.1**	UPC	**31.8**
UPC	42.9	MIT	31.4
UEKAE	36.4	DCU	28.1
UMD	36.0	UW	26.9
UW	35.4	NTT	25.5
MIT	35.1	CMU	25.5
CMU	33.9	UMD	25.0
LIG	33.9	LIG	24.2
NTT	25.3	UEKAE	19.8
GREYC	21.7	HKUST	11.2
HKUST	13.1		

Table 3 Human Rankings: AE, Clean and ASR..

CE Clean	
SYSTEM	% BETTER
CASIA	**37.6**
I2R	37.0
ICT	34.8
RWTH	32.4
FBK	30.6
CMU	30.6
UPC	28.3
XMU	28.1
HKUST	25.5
MIT	25.0
NTT	24.6
ATR	24.2
UMD	23.6
DCU	18.6
NUDT	16.1

Table 4 Human Rankings: CE Clean.

In order to compare tasks from this evaluation campaign with previous workshops, the top three systems for each ASR input condition for IE, JE, AE and the CE clean tasks were evaluated using the NIST fluency/adequacy metrics. The best scores are presented in bold.

NIST IE ASR		
SYSTEM	ADEQUACY	FLUENCY
ATR	0.529	0.446
FBK	**0.564**	0.479
RWTH	0.544	**0.484**
NIST JE ASR		
SYSTEM	ADEQUACY	FLUENCY
CMU-UKA	**0.501**	0.505
ATR	0.492	**0.540**
UEKAE	0.491	0.510
NIST CE Clean		
SYSTEM	ADEQUACY	FLUENCY
CMU	0.472	0.528
ICT	**0.511**	0.521
I2R	0.507	**0.547**
NIST AE ASR		
SYSTEM	ADEQUACY	FLUENCY
UW	0.430	0.404
MIT	0.447	**0.450**
UPC	**0.453**	0.431

Table 5 NIST adequacy and fluency scores normalized for all ASR input conditions and CE Clean. Top three systems to be evaluated for adequacy and fluency were chosen by BLEU rankings.

IE Clean	
System	BLEU
RWTH_IE_clean_primary_01	0.4531
FBK_IE_clean_primary_01	0.4432
ATR_IE_CLEAN_primary_01	0.3828
NTT_IE_clean_primary_01	0.3091
UEDIN_IE_clean_primary_01	0.2909
MIT-LL+AFRL_IE_clean_primary_01	0.2842
INESCID_IE_clean_primary_02	0.2657
UW_IE_clean_primary_01	0.2651
HKUST_IE_clean_01	0.1702
ITI_UPV_IE_clean_primary_01	0.1613
IE ASR	
FBK_IE_ASR_primary_01	0.4229
RWTH_IE_ASR_primary_01	0.4128
ATR_IE_ASR_primary_01	0.3550
NTT_IE_ASR_primary_01	0.2868
UEDIN_IE_ASR_primary_01	0.2662
UW_IE_ASR_primary_01	0.2540
MIT-LL+AFRL_IE_ASR_primary_01	0.2500
INESCID_IE_ASR_primary_02	0.2416
RALI_IE_ASR_primary_01	0.2106
HKUST_IE_ASR_01	0.1702

Table 6 Italian systems ranked by BLEU score.

JE Clean	
System	BLEU
TUBITAK-UEKAE_JE_clean_primary_01	0.4841
CMU-UKA_JE_clean_primary	0.4828
FBK_JE_clean_primary_01	0.4789
ATR_JE_CLEAN_primary_01	0.4745
NTT_JE_clean_primary_01	0.4365
TOTTORI_JE_clean_01	0.4321
HKUST_JE_CLEAN_01	0.4051
GREYC_JE_clean_primary_1	0.3964
DCU_JE_CLEAN_primary_01	0.3959
JE ASR	
System	BLEU
CMU-UKA_JE_ASR_primary	0.4386
TUBITAK-UEKAE_JE_ASR_primary_01	0.4269
ATR_JE_ASR_primary_01	0.4144
FBK_JE_ASR_primary_01	0.3946
NTT_JE_ASR_primary_01	0.3535
HKUST_JE_ASR_01	0.3249
DCU_JE_ASR_primary_01	0.3182

Table 7 Japanese systems ranked by BLEU score.

3.2.Automatic Evaluation Results

The following tables show the ranking of the primary submitted runs for all tasks according to BLEU score.

For both input conditions of the IE challenge task, the same three participants, RWTH, FBK and NiCT/ATR are clustered together at the head of the list.

AE Clean	
System	BLEU
TUBITAK-UEKAE_AE_clean_primary_01	0.4923
UMD_AE_clean_01	0.4858
UPC_AE_clean_primary_01	0.4804
DCU_AE_clean_primary_01	0.4709
MIT-LL+AFRL_AE_clean_primary_01	0.4553
CMU_AE_CLEAN_primary_02	0.4463
UW_AE_clean_primary_01	0.4162
LIG_AE_clean_primary_01	0.4135
NTT_AE_clean_primary_01	0.3403
GREYC_AE_clean_primary_1	0.3290
HKUST_AE_clean_01	0.1951
AE ASR	
UPC_AE_ASR_primary_01	0.4445
MIT-LL+AFRL_AE_ASR_primary_01	0.4429
UW_AE_ASR_primary_01	0.4092
DCU_AE_ASR_primary_01	0.3942
UMD_AE_ASR_primary_01	0.3908
LIG_AE_ASR_primary_01	0.3804
CMU_AE_ASR_primary_02	0.3756
TUBITAK-UEKAE_AE_ASR_primary_01	0.3679
NTT_AE_ASR_primary_01	0.3626
HKUST_AE_ASR_01	0.1420

Table 8 Arabic systems ranked by BLEU score.

CE Clean	
System	BLEU
I2R_CE_clean_primary_01	0.4077
ICT_CE_clean_Primary_01	0.3750
CMUsamt_CE_CLEAN_primary_01	0.3744
RWTH_CE_clean_primary_01	0.3708
CASIA_CE_clean_primary_01	0.3648
MIT-LL+AFRL_CE_clean_primary_01	0.3631
FBK_CE_clean_primary_01	0.3472
HKUST_CE_clean_01	0.3426
UMD_CE_clean_01	0.3211
ATR_CE_CLEAN_primary_01	0.3133
UPC_CE_clean_primary_01	0.2991
XMU_CE_clean_primary_01	0.2888
NTT_CE_clean_primary_00	0.2789
DCU_CE_CLEAN_primary_01	0.2737
NUDT_CE_clean_primary_01	0.1934

Table 9 Chinese systems ranked by BLEU score.

4. Discussion

4.1. Challenge and Classical Tasks for 2007

The challenge tasks planned for this year were intended to further the direction begun last year towards the translation of spontaneous. The Italian task presented a much more difficult type of input speech.

4.2. Participant Supplied Resources

While the number of participants that submitted resources by the deadline (approximately five weeks before test submission deadline) was somewhat limited, the number of resources collected was very encouraging. A problem with the request, however, was the definition of "publicly available" and of "affordable". It was clear that both terms are open to interpretation especially when resources require license agreements to be signed and when some resources may be with the allowable budget of some research groups but not others.

4.3. Human Evaluation

This year's evaluation campaign adopted a new human evaluation metric which simplified the evaluation process. This metric has been shown to be more efficient in terms of judgement times, more consistent in inter-annotator agreements[7]. Here, we used the kappa coefficient[14] to measure inter-annotator agreement using the same values as in [7] for P(E), i.e. 1/3. For all ranking tasks, the inter-annotator agreement was relatively good, with K = 0.608. According to Landis and Koch[15], the range of K 0.41 to 0.6 is moderate agreement. Individual rankings for certain tasks showed higher inter-annotator agreement.

With this metric, human evaluation of submitted runs was able to be offered to all runs of all tasks.

5. Conclusions

The 2007 IWSLT evaluation campaign saw increased number of groups submitting systems to one or more tasks continuing the growth of the IWSLT series of workshops.

A new human evaluation metric was adopted which proved to be efficient and allowed the evaluation of all tasks by human evaluators with this metric.

6. Acknowledgements

I would like to acknowledge the C-Star partners for their support in the preparation of this year's evaluation campaign. In particular, I would like to thank Mark Fuhs and Matthias Eck from CMU/UKA, and Nicola Bertoldi and Roldano Cattoni from FBK-Irst for their help in preparing the Arabic and Italian data. Very special thanks go to Michael Paul for his help both for the Japanese data, but especially for his help with the Chinese data. Thanks go to Philipp Koehn of the University of Edinburgh for permission to use the human evaluation interface used for the WMT07 shared task and to Josh Schroeder for explaining how it all worked. Thanks also to Chris Callison-Burch for his scripts and advice in scoring the human evaluations. Thanks to the all evaluators who took part in the human evaluation, especially, to John Henderson of MITRE. Finally, thanks go to all the participants of this year's workshop.

7. References

[1] M. Paul, "Overview of the IWSLT 2006 Evaluation Campaign," in *Proc. of the International Workshop of Spoken Language Translation*, Kyoto, Japan, 2006, pp.1-15 .

[2] 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.

[3] 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.

[4] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. "Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversation in the Real World" in *Proc. of LREC 2002*, Las Palmas, Spain, 2002.

[5] R. Cattoni, M. Danieli, V. Sandrini, C. Soria, "ADAM: the SI-TAL Corpus of Annotated Dialogues", in *Proc. of LREC 2002*, Las Palmas, Spain, 2002.

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

[8] LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Revision 1.5.

[9] D. Jones, W Shen, N. Granoien, M. Herzog, and C. Weinstein, “Measuring translation quality by testing English speakers with a new defense language proficiency test for Arabic,” in *Proc. of the 2005 International Conference on Intelligence Analysis*, McLean, VA, 2005.

[10] P. Koehn and C. Monz, “Manual and automatic evaluation of machine translation between European language,” in *Proc. of NAACL 2006 Workshop on Statistical Machine Translation*, New York, 2006.

[11] J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing, "Confidence estimation for machine translation," in *CLSP Summer Workshop Final Report WS2003*, Johns Hopkins University, 2003.

[12] D. Coughlin, “Correlating automated and human assessments of machine translation quality,” in *Proc. Of MT Summit IX*, New Orleans, 2003.

[13] G. Doddington, “Automatic evaluation of machine translation quality using n-grams co-occurence statistics,” in *Human Language Technology: Notebook Proceedings*, San Diego, 2002, pp. 128-132.

[14] J. Carletta, “Assessing Agreement on classification tasks: The kappa statistic,” in *Computational Linguistics*, 22(2):249-254, 1996.

[15] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, 33:159-174, 1977.

[16] A. Finch, E. Denoual, H. Okuma, M. Paul, H. Yamamoto, K. Yasuda, R. Zhang and E. Sumita, "The NICT/ATR Speech Translation System for IWSLT 2007"," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[17] Z. He, Haitao Mi, Y. Liu, W. Luo, Y. Huang, Z. Ren, Y. Lu and Q. Liu, "The ICT Statistical Machine Translation Systems for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[18] Y. Zhou, Y. He and C. Zong, "The CASIA Phrase-Based Statistical Machine Translation System for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[19] Y. Chen, X. Shi and C. Zhou, "The XMU SMT System for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[20] L. Besacier, M. Amar and L. Viet-Bac, "The LIG Arabic / English Speech Translation System at IWSLT 07," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[21] J. Murakami, M. Tokuhisa and S. Ikehara, "Statistic Machine Translation using Large J/E Parallel Corpus and Long Phrase Tables," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[22] A. Patry, P. Langlais and F. Béchet, "MISTRAL: A Lattice Translation System for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[23] Y. Yves and A. Lardilleux, "The GREYC Machine Translation System for the IWSLT 2007 Evaluation Campaign," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[24] B. Chen, J. Sun, H. Jiang, M. Zhang and A. Ti Aw, "I2R Chinese-English Translation System for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[25] N. Bertoldi, M. Cettolo, R. Cattoni and M. Federico, "FBK @ IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[26] W. Chao and Z. Li, "NUDT Machine Translation System for IWSLT2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[27] P. Lambert, M. Costa-jussà, J. Crego, M. Khalilov, J. Giménez, J. Mariño, R. Banchs, J. Fonollosa and H. Schwenk, "The TALP Ngram-based SMT System for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[28] J. Schroeder and P. Koehn, "The University of Edinburgh System Description for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[29] H. Hassan, Y. Ma and A. Way, "MaTrEx: the DCU Machine Translation System for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[30] A. Mauser, D. Vilar, G. Leusch, Y. Zhang and H. Ney, "The RWTH Machine Translation System for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[31] J. Graça, D. Caseiro and L. Coheur, "The INESC-ID IWSLT07 SMT System," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[32] W. Shen, B. Delaney, T. Anderson and R. Slyh, "The MIT-LL/AFRL IWSLT-2007 MT System," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[33] T. Watanabe, J. Suzuki, K. Sudoh, H. Tsukada and H. Isozaki, "Larger Feature Set Approach for Machine Translation in IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[34] K. Kirchhoff, "The University of Washington Machine Translation System for the IWSLT 2007 Competition," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[35] I. Lane, A. Zollmann, T. Linh Nguyen, N. Bach, A. Venugopal, S. Vogel, K. Rottmann, Y. Zhang and A. Waibel, "The CMU-UKA Statistical Machine Translation Systems for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[36] Y. Lepage and A. Lardilleux, "The GREYC Machine Translation System for the IWSLT 2007 Evaluation Campaign," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[37] Y. Shen, C. Lo, M. Carpuat and D. Wu, "HKUST Statistical Machine Translation Experiments for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

[38] V. Alabau, A. Sanchis and F. Casacuberta, "Using Word Posterior Probabilities in Lattice Translation," Proc. of the

International Workshop on Spoken Language Translation, Trento, 2007.

[39] C. Mermer, H. Kaya and M. Ugur Dogan, "The TUBITAK-UEKAE Statistical Machine Translation System for IWSLT 2007," Proc. of the International Workshop on Spoken Language Translation, Trento, 2007.

[40] C. Dyer, "The University of Maryland Translation System for IWSLT 2007," *Proc. of the International Workshop on Spoken Language Translation*, Trento, 2007.

8. Appendix: MT System Overview:

Research Group	MT System Description	Type	MT System
ATR Spoken Language Communication Research Lab	The NICT/ATR Speech Translation System for IWSLT 2007	Phrase-based SMT	NICT/ATR
Chinese Academy of Sciences, Inst. of Computing Technology, Key Laboratory of Intelligent Information Processing	The ICT Statistical Machine Translation Systems for IWSLT 2007	Syntax-based SMT	ICT
Chinese Academy of Sciences, Institute of Automation, National Laboratory of Pattern Recognition	The CASIA Phrase-Based Statistical Machine Translation System for IWSLT 2007	Phrase-based SMT	CASIA
Xiamen University, School of Information Sciences and Technologies, Dept. of Cognitive Science	The XMU SMT System for IWSLT 2007	Phrase-based SMT	XMU
Univ. J. Fourier (Grenoble), LIG Laboratory, GETALP Team	The LIG Arabic / English Speech Translation System at IWSLT 07	SMT	LIG
Tottori Univ., Faculty of Eng., Dept. of Information and Knowledge Engineering	Statistic Machine Translation using Large J/E Parallel Corpus and Long Phrase Tables	SMT	TOTTORI
Univ. de Montréal, Univ. of Avignon	MISTRAL: A Lattice Translation System for IWSLT 2007	Phrase-based SMT	MISTRAL
GREYC, Univ. of Caen Basse-Normandie	The GREYC Machine Translation System for the IWSLT2007 Evaluation Campaign	EBMT	GREYC
Institute for Infocomm Research (Singapore), Dept. of Human Language Technology	I2R Chinese-English Translation System for IWSLT 2007	SMT	I2R
FBK - Fondazione Bruno Kesler	FBK @ IWSLT 2007	SMT	FBK
National Univ. of Defence technology, School of Science, Beihang University, School of Computer Science	NUDT Machine Translation System for IWSLT2007	SMT	NUDT
Universitat Politècnica de Catalunya, TALP Research Center	The TALP Ngram-based SMT System for IWSLT 2007	SMT	TALP
U. of Edinburgh, School of Informatics	The University of Edinburgh System Description for IWSLT 2007	Phrase-based SMT	UEDIN
Dublin City Univ., School of Computing	MaTrEx: the DCU Machine Translation System for IWSLT 2007	EBMT	DCU
RWTH Aachen Univ., Computer Science Dept., Human Language Technology and Pattern Recognition	The RWTH Machine Translation System for IWSLT 2007	Phrase-based SMT	RWTH
INESC-ID, Spoken Language Lab (L2F)	The INESC-ID IWSLT07 SMT System	SMT	INESC-ID
MIT Lincoln Laboratory, Information Systems and Technology Group, Air Force Research Laboratory	The MIT-LL/AFRL IWSLT-2007 MT System	SMT	MIT-LL
NTT Communication Science Laboratories	Larger Feature Set Approach for Machine Translation in IWSLT 2007	Phrase-based SMT	NTT
Univ. of Washington, Dept. of Electrical Engineering	The University of Washington Machine Translation System for the IWSLT 2007 Competition	SMT	UW
InterACT Research Laboratories: Carnegie Mellon Univ. (Pittsburgh), Univ. of Karlsruhe, (Karlsruhe)	The CMU-UKA Statistical Machine Translation Systems for IWSLT 2007	Syntax-augmented SMT	CMU-UKA
Univ. of Science and Technology, Hong Kong, Dept. of Computer Science	HKUST Statistical Machine Translation Experiments for IWSLT 2007	Phrase-based SMT	HKUST
Institut Tecnològic d'Informàtica, Departament de Sistemes Informàtics i Computaciòn	Using Word Posterior Probabilities in Lattice Translation	SMT	ITI/UPV
National Research Institute of Electronics and Cryptology & The Scientific and Technological Research Council of Turkey	The TUBITAK-UEKAE Statistical Machine Translation System for IWSLT 2007	Phrase-based SMT	TUBITAK-UEKAE
Univ. of Maryland, Dept. of Linguistics	The University of Maryland Translation System for IWSLT 2007	Phrase-based SMT	UMD

9. Appendix B: Automatic Rankings by BLEU score for all submitted runs

IE Clean	
System	BLEU
RWTH_IE_clean_primary_01	0.4531
FBK_IE_clean_02	0.4444
FBK_IE_clean_primary_01	0.4432
RWTH_IE_clean_09	0.4415
FBK_IE_clean_04	0.4341
FBK_IE_clean_03	0.4341
RWTH_IE_clean_06	0.4287
RWTH_IE_clean_07	0.4284
RWTH_IE_clean_03	0.4246
RWTH_IE_clean_02	0.4201
RWTH_IE_clean_05	0.4166
RWTH_IE_clean_04	0.4162
ATR_IE_CLEAN_05	0.4037
ATR_IE_CLEAN_04	0.3958
ATR_IE_CLEAN_primary_01	0.3828
ATR_IE_CLEAN_02	0.3761
ATR_IE_CLEAN_03	0.3586
RWTH_IE_clean_08	0.3349
NTT_IE_clean_primary_01	0.3091
NTT_IE_clean_02	0.2983
NTT_IE_clean_04	0.2948
NTT_IE_clean_03	0.2947
NTT_IE_clean_05	0.2914
UEDIN_IE_clean_primary_01	0.2909
MIT-LL+AFRL_IE_clean_primary_01	0.2842
INESCID_IE_clean_primary_02	0.2657
UW_IE_clean_primary_01	0.2651
INESCID_IE_clean_01	0.2635
ITI_UPV_IE_clean_04	0.2100
ITI_UPV_IE_clean_03	0.2037
HKUST_IE_clean_01	0.1702
ITI_UPV_IE_clean_primary_01	0.1613

IE ASR	
System	BLEU
FBK_IE_ASR_primary_01	0.4229
FBK_IE_ASR_02	0.4206
FBK_IE_ASR_06	0.4165
FBK_IE_ASR_10	0.4155
FBK_IE_ASR_05	0.4151
FBK_IE_ASR_09	0.4146
RWTH_IE_ASR_primary_01	0.4128
FBK_IE_ASR_04	0.4100
FBK_IE_ASR_03	0.4099
FBK_IE_ASR_08	0.4075
FBK_IE_ASR_12	0.4074
FBK_IE_ASR_07	0.4074
FBK_IE_ASR_11	0.4045
ATR_IE_ASR_05	0.3717
ATR_IE_ASR_04	0.3665
ATR_IE_ASR_primary_01	0.3550
ATR_IE_ASR_02	0.3487
ATR_IE_ASR_03	0.3349
NTT_IE_ASR_primary_01	0.2868
UEDIN_IE_ASR_primary_01	0.2662
NTT_IE_ASR_02	0.2601
NTT_IE_ASR_03	0.2552
UW_IE_ASR_primary_01	0.2540
MIT-LL+AFRL_IE_ASR_primary_01	0.2500
INESCID_IE_ASR_01	0.2435
INESCID_IE_ASR_primary_02	0.2416
MIT-LL+AFRL_IE_ASR_02	0.2278
RALI_IE_ASR_primary_01	0.2106
RALI_IE_ASR_02	0.2055
RALI_IE_ASR_04	0.1850
ITI_UPV_IE_ASR_02	0.1822
HKUST_IE_ASR_01	0.1702
RALI_IE_ASR_03	0.0560

AE Clean	
System	BLEU
TUBITAK-UEKAE_AE_clean_primary_01	0.4923
UMD_AE_clean_01	0.4858
UPC_AE_clean_primary_01	0.4804
MIT-LL+AFRL_AE_clean_02	0.4741
DCU_AE_clean_primary_01	0.4709
MIT-LL+AFRL_AE_clean_primary_01	0.4553
CMU_AE_CLEAN_primary_02	0.4463
UW_AE_clean_primary_01	0.4162
LIG_AE_clean_primary_01	0.4135
NTT_AE_clean_02	0.3446
NTT_AE_clean_primary_01	0.3403
GREYC_AE_clean_primary_1	0.3290
NTT_AE_clean_03	0.3078
NTT_AE_clean_05	0.2947
NTT_AE_clean_04	0.2947
HKUST_AE_clean_01	0.1951

AE ASR	
System	BLEU
UPC_AE_ASR_primary_01	0.4445
MIT-LL+AFRL_AE_ASR_primary_01	0.4429
MIT-LL+AFRL_AE_ASR_02	0.4293
UW_AE_ASR_primary_01	0.4092
DCU_AE_ASR_primary_01	0.3942
UMD_AE_ASR_primary_01	0.3908
LIG_AE_ASR_primary_01	0.3804
CMU_AE_ASR_primary_02	0.3756
TUBITAK-UEKAE_AE_ASR_primary_01	0.3679
LIG_AE_ASR_secondary_01	0.3644
NTT_AE_ASR_primary_01	0.3626
NTT_AE_ASR_02	0.3037
NTT_AE_ASR_03	0.2813
HKUST_AE_ASR_01	0.1420

JE Clean	
System	BLEU
FBK_JE_clean_02	0.4893
TUBITAK-UEKAE_JE_clean_primary_01	0.4841
CMU-UKA_JE_clean_primary	0.4828
FBK_JE_clean_primary_01	0.4789
ATR_JE_CLEAN_primary_01	0.4745
ATR_JE_CLEAN_03	0.4630
ATR_JE_CLEAN_04	0.4559
ATR_JE_CLEAN_02	0.4512
NTT_JE_clean_02	0.4459
NTT_JE_clean_primary_01	0.4365
NTT_JE_clean_04	0.4337
TOTTORI_JE_clean_02	0.4321
TOTTORI_JE_clean_01	0.4321
NTT_JE_clean_03	0.4205
NTT_JE_clean_05	0.4192
TOTTORI_JE_clean_04	0.4184
TOTTORI_JE_clean_03	0.4184
HKUST_JE_CLEAN_01	0.4051
GREYC_JE_clean_primary_1	0.3964
DCU_JE_CLEAN_primary_01	0.3959
DCU_JE_CLEAN_04	0.3918
DCU_JE_CLEAN_03	0.3898

JE ASR	
System	BLEU
CMU-UKA_JE_ASR_primary	0.4386
TUBITAK-UEKAE_JE_ASR_primary_01	0.4269
ATR_JE_ASR_primary_01	0.4144
ATR_JE_ASR_02	0.4106
FBK_JE_ASR_04	0.3969
FBK_JE_ASR_primary_01	0.3946
ATR_JE_ASR_03	0.3931
FBK_JE_ASR_02	0.3897
FBK_JE_ASR_03	0.3848
ATR_JE_ASR_04	0.3665
NTT_JE_ASR_primary_01	0.3535
NTT_JE_ASR_02	0.3533
HKUST_JE_ASR_01	0.3249
DCU_JE_ASR_03	0.3248
DCU_JE_ASR_04	0.3231
DCU_JE_ASR_02	0.3215
DCU_JE_ASR_primary_01	0.3182
NTT_JE_ASR_03	0.2945

CE Clean	
System	BLEU
I2R_CE_clean_primary_01	0.4077
I2R_CE_clean_02	0.3942
RWTH_CE_clean_04	0.3849
RWTH_CE_clean_10	0.3791
RWTH_CE_clean_08	0.3785
ICT_CE_clean_Primary_01	0.3750
CMUsamt_CE_CLEAN_primary_01	0.3744
RWTH_CE_clean_09	0.3723
RWTH_CE_clean_05	0.3718
RWTH_CE_clean_primary_01	0.3708
RWTH_CE_clean_12	0.3674
RWTH_CE_clean_07	0.3655
CASIA_CE_clean_primary_01	0.3648
MIT-LL+AFRL_CE_clean_03	0.3634
MIT-LL+AFRL_CE_clean_primary_01	0.3631
MIT-LL+AFRL_CE_clean_02	0.3614
CMUsamt_CE_CLEAN_02	0.3597
ICT_CE_clean_02	0.3573
FBK_CE_clean_05	0.3508
RWTH_CE_clean_03	0.3473
FBK_CE_clean_primary_01	0.3472
HKUST_CE_clean_01	0.3426
FBK_CE_clean_04	0.3421
RWTH_CE_clean_02	0.3414
FBK_CE_clean_02	0.3410

CE Clean (cont.)	
System	BLEU
FBK_CE_clean_03	0.3394
RWTH_CE_clean_14	0.3364
RWTH_CE_clean_13	0.3298
UMD_CE_clean_01	0.3211
ATR_CE_CLEAN_02	0.3185
ATR_CE_CLEAN_primary_01	0.3133
ATR_CE_CLEAN_03	0.3124
ATR_CE_CLEAN_04	0.3117
RWTH_CE_clean_06	0.3081
UPC_CE_clean_primary_01	0.2991
ATR_CE_CLEAN_08	0.2937
UPC_CE_clean_03	0.2920
ATR_CE_CLEAN_07	0.2897
XMU_CE_clean_primary_01	0.2888
UPC_CE_clean_02	0.2885
XMU_CE_clean_03	0.2879
ATR_CE_CLEAN_05	0.2850
ATR_CE_CLEAN_06	0.2832
NTT_CE_clean_04	0.2807
ICT_CE_clean_03	0.2802
NTT_CE_clean_primary_00	0.2789
NTT_CE_clean_03	0.2780
XMU_CE_clean_02	0.2742
NTT_CE_clean_05	0.2737
DCU_CE_CLEAN_primary_01	0.2737
DCU_CE_CLEAN_03	0.2701
DCU_CE_CLEAN_02	0.2681
NTT_CE_clean_02	0.2627
NUDT_CE_clean_primary_01	0.1934
ICT_CE_clean_04	0.1777
NUDT_CE_clean_02	0.1758

10. Appendix C: Unnormalized NIST adequacy/fluency scores

The following tables show unnormalized adequacy and fluency scores. The best scores are shown in bold.

Arabic English ASR NIST		
System	ADEQUACY	FLUENCY
MIT	3.10	**3.24**
UW	3.01	2.97
UPC	**3.13**	3.13

Chinese English Clean NIST		
System	ADEQUACY	FLUENCY
CMU	3.26	3.69
ICT	**3.51**	3.67
I2R	3.48	**3.80**

Italian English ASR NIST		
System	ADEQUACY	FLUENCY
ATR	3.62	3.27
RWTH	3.69	**3.46**
FBK	**3.80**	**3.46**

Japanese English ASR NIST		
System	ADEQUACY	FLUENCY
CMU-UKA	**3.39**	3.54
ATR	3.35	**3.73**
UEKAE	3.34	3.56