Overview of the IWSLT 2012 Evaluation Campaign

M. Federico *M. Cettolo*
FBK
via Sommarive 18,
38123 Povo (Trento), Italy
{federico,cettolo}@fbk.eu

L. Bentivogli
CELCT
Via alla Cascata 56/c,
38123 Povo (Trento), Italy
bentivo@fbk.eu

M. Paul
NICT
Hikaridai 3-5,
619-0289 Kyoto, Japan
michael.paul@nict.go.jp

S. Stüker
KIT
Adenauerring 2,
76131 Karlsruhe, Germany
sebastian.stueker@kit.edu

Abstract

We report on the ninth evaluation campaign organized by the IWSLT workshop. This year, the evaluation offered multiple tracks on lecture translation based on the TED corpus, and one track on dialog translation from Chinese to English based on the Olympic trilingual corpus. In particular, the TED tracks included a speech transcription track in English, a speech translation track from English to French, and text translation tracks from English to French and from Arabic to English. In addition to the official tracks, ten unofficial MT tracks were offered that required translating TED talks into English from either Chinese, Dutch, German, Polish, Portuguese (Brazilian), Romanian, Russian, Slovak, Slovene, or Turkish. 16 teams participated in the evaluation and submitted a total of 48 primary runs. All runs were evaluated with objective metrics, while runs of the official translation tracks were also ranked by crowd-sourced judges. In particular, subjective ranking for the TED task was performed on a progress test which permitted direct comparison of the results from this year against the best results from the 2011 round of the evaluation campaign.

1. Introduction

The International Workshop on Spoken Language Translation (IWSLT) offers challenging research tasks and an open experimental infrastructure for the scientific community working on the automatic translation of spoken and written language. The focus of the 2012 IWSLT Evaluation Campaign was the translation of lectures and dialogs. The task of translating lectures was built around the TED[1] talks, a collection of public lectures covering a variety of topics. The TED Task offered three distinct tracks addressing automatic speech recognition (ASR) in English, spoken language translation (SLT) from English to French, and machine translation (MT) from English to French and from Arabic to English. In addition to the official MT language pairs, ten other unofficial translation directions were offered, with English as the target language and the source language being either Chinese, Dutch, German, Polish, Portuguese (Brazilian), Romanian, Russian, Slovak, Sloven, or Turkish.

[1] http://www.ted.com

This year, we also launched the so-called OLYMPICS Task, which addressed the MT of transcribed dialogs, in a limited domain, from Chinese to English.

For each track, a schedule and evaluation specifications, as well as language resources for system training, development and evaluation were made available through the IWSLT website. After the official evaluation deadline, automatic scores for all submitted runs we provided to the participants. In this edition, we received run submissions by 16 teams from 11 countries. For all the official SLT and MT tracks we also computed subjective rankings of all primary runs via crowd-sourcing. For the OLYMPICS Task, system ranking was based on a round-robin tournament structure, following the evaluation scheme adopted last year. For the TED task, as a novelty for this year, we introduced a double-elimination tournament, which previous experiments showed to provide rankings very similar to the more exhaustive but more costly round-robin scheme. Moreover, for the TED Task we run the subjective evaluation on a progress test—i.e., the evaluation set from 2011 that we never released to the participants. This permitted the measure of progress of SLT and MT against the best runs of the 2011 evaluation campaign.

In the rest of the paper, we introduce the TED and OLYMPICS tasks in more detail by describing for each track the evaluation specifications and the language resources supplied. For the TED MT track, we also provide details for the reference baseline systems that we developed for all available translation directions. Then, after listing the participants, we describe how the human evaluation was organized for the official SLT and MT tracks. Finally, we present the main findings of this year's campaign and give an outlook on the next edition of IWSLT. The paper concludes with two appendices, which present detailed results of the objective and subjective evaluations.

2. TED Task

2.1. Task Definition

The translation of TED talks was introduced for the first time at IWSLT 2010. TED is a nonprofit organization that "invites the world's most fascinating thinkers and doers [...] to give the talk of their lives". Its website makes the video

recordings of the best TED talks available under the Creative Commons Attribution-NonCommercial-NoDerivs 3.0 license[2]. All talks have English captions, which have also been translated into many other languages by volunteers worldwide.

This year we proposed three challenging tracks involving TED talks:

`ASR track`: automatic transcription of the talks' English audio;

`SLT track`: speech translation of talks from audio (or ASR output) to text, from English to French;

`MT track`: text translation of talks from:

`official`: English to French and Arabic to English

`unofficial`: German, Dutch, Polish, Portuguese-Brazil, Romanian, Russian, Slovak, Slovenian, Turkish and Chinese to English

In the following sections, we give an overview of the released language resources and provide more details about these three tracks.

2.2. Supplied Textual Data

Starting this year, TED data sets for the IWSLT evaluations are distributed through the WIT^3 web repository [1].[3] The aim of this repository is to make the collection of TED talks effectively usable by the NLP community. Besides offering ready-to-use parallel corpora, the WIT^3 repository also offers MT benchmarks and text-processing tools designed for the TED talks collection.

The language resources provided to the participants of IWSLT 2012 comprise monolingual and parallel training corpora of TED talks (`train`). Concerning the two official language pairs, the development and evaluation data sets (`dev2010` and `tst2010`), used in past editions, were provided for development and testing purposes. For evaluation purposes, two data sets were released: a new test set (`tst2012`) and the official test set of 2011 (`tst2011`) that was used as the `progress test` set to compare the results of this year against the best results achieved in 2011.

For the unofficial language pairs similar development/test set were prepared, most of them overlapping with the dev/test sets prepared for Arabic-English.

As usual, only the source part of the evaluation sets was released to the participants. All texts were UTF-8 encoded, case-sensitive, included punctuation marks, and were not tokenized. Parallel corpora were aligned at sentence level, even though the original subtitles were aligned at sub-sentence level. Details on the supplied monolingual and parallel data for the two official language pairs are given in Tables 1 and 2; the figures reported refer to tokenized texts.

[2]http://creativecommons.org/licenses/by-nc-nd/3.0/

[3]http://wit3.fbk.eu

Table 1: Monolingual resources for official language pairs

data set	lang	sent	token	voc
train	En	142k	2.82M	54.8k
	Fr	143k	3.01M	67.3k

Table 2: Bilingual resources for official language pairs

task	data set	lang	sent	token	voc	talks
MT$_{EnFr}$	train	En	141k	2.77M	54.3k	1029
		Fr		2.91M	66.9k	
	dev2010	En	934	20.1k	3.4k	8
		Fr		20.3k	3.9k	
	tst2010	En	1,664	32.0k	3.9k	11
		Fr		33.8k	4.8k	
	tst2011	En	818	14.5k	2.5k	8
		Fr		15.6k	3.0k	
	tst2012	En	1,124	21,5k	3.1k	11
		Fr		23,5k	3.7k	
MT$_{ArEn}$	train	Ar	138k	2.54M	89.7k	1015
		En		2.73M	53.9k	
	dev2010	Ar	934	18.3k	4.6k	8
		En		20.1k	3.4k	
	tst2010	Ar	1,664	29.3k	6.0k	11
		En		32.0k	3.9k	
	tst2011	Ar	1,450	25.6k	5.6k	16
		En		27.0k	3.7k	
	tst2012	Ar	1,704	27.8k	6.1k	15
		En		30.8k	4.1k	

Similar to last year, several out-of-domain parallel corpora, including texts from the United Nations, European Parliament, and news commentaries, were supplied to the participants. These corpora were kindly provided by the organizers of the *7th Workshop on Statistical Machine Translation*[4] and the *EuroMatrixPlus project* [5].

2.3. Speech Recognition

The goal of the *Automatic Speech Recognition* (ASR) track for IWSLT 2012 was to transcribe the English recordings of the `tst2011` and `tst2012` MT$_{EnFr}$ test sets (Table 2) for the TED task. This task reflects the recent increase of interest in automatic subtitling and audiovisual content indexing.

Speech in TED lectures is in general planned, well articulated, and recorded in high quality. The main challenges for ASR in these talks are to cope with a large variability of topics, the presence of non-native speakers, and the rather informal speaking style.

Table 3 provides statistics on the two sets; the counts of reference transcripts refer to lower-cased text without punctuation after the normalization described in detail in Section 2.6.

[4]http://www.statmt.org/wmt12/translation-task.html

[5]http://www.euromatrixplus.net/

Table 3: Statistics of ASR evaluation sets

task	data set	duration	sent	token	voc	talks
ASR_{En}	tst2011	1h07m28s	818	12.9k	2.3k	8
	tst2012	1h45m04s	1124	19.2k	2.8k	11

2.3.1. Language Resources

For acoustic model training, no specific data was provided by the evaluation campaign. Instead, just as last year, participants were allowed to use any data available to them, but recorded before December 31^{st}, 2010.

For language model training, the training data was restricted to the English monolingual texts and the English part of the provided parallel texts as described in Section 2.2.

2.4. Spoken Language Translation

The SLT track required participants to translate the English TED talks of `tst2011` and `tst2012` into French, starting from the audio signal (see Section 2.3). The challenge of this translation task over the MT track is the necessity to deal with automatic, and in general error prone, transcriptions of the audio signal, instead of correct human transcriptions.

Participants not using their own ASR system could resort to automatic transcriptions distributed by the organizers. These were the primary runs submitted by three participants to the ASR track:

Table 4: WER of ASR runs released for the SLT track

system num.	system name	tst2011	tst2012
1	NICT	10.9	12.1
2	MITLL	11.1	13.3
3	UEDIN	12.4	14.4

Table 4 shows their WERs. Participants could freely choose which set of transcriptions to translate; they were allowed even to create a new transcription, e.g., by means of system combination methods. Details on the specifications for this track are given in Section 2.6.

2.4.1. Language Resources

For the SLT task the language resources available to participants are the union of those of the ASR track, described in Section 2.3.1, and of the English-to-French MT track, described in Section 2.2.

2.5. Machine Translation

The MT TED track basically corresponds to a subtitling translation task. The natural translation unit considered by the human translators volunteering for TED is indeed the single caption—as defined by the original transcript—which in general does not correspond to a sentence, but to fragments of it that fit the caption space. While translators can look at the context of the single captions, arranging the MT task in this way would make it particularly difficult, especially when word re-ordering across consecutive captions occurs. For this reason, we preprocessed all the parallel texts to re-build the original sentences, thus simplifying the MT task.

Reference results from baseline MT systems on the official evaluation set (`tst2012`) are provided via the WIT^3 repository. This helps participants and MT scientists to assess their experimental outcomes, but also to set reference systems for the human evaluation experiments (Section 5).

MT baselines were trained from TED data only, i.e,. no additional out-of-domain resources were used. Preprocessing was applied as follows: Arabic and Chinese words were segmented by means of AMIRA [2] and the Stanford Chinese Segmenter [3], respectively; while for all the other languages the `tokenizer` script released with the Europarl corpus [4] was applied.

The baselines were developed with the Moses toolkit [5]. Translation and lexicalized reordering models were trained on the parallel training data; 5-gram LMs with improved Kneser-Ney smoothing [6] were estimated on the target side of the training parallel data with the IRSTLM toolkit [7]. The weights of the log-linear interpolation model were optimized on `dev2010` with the MERT procedure provided with Moses. Performance scores were computed with the `MultEval` script implemented by [8].

Table 5 collects the %BLEU, METEOR, and TER scores ("case sensitive+punctuation" mode) of all the baseline systems developed for all language pairs. In addition to the scores obtained on `dev2010` after the last iteration of the tuning algorithm, we also report the scores measured on the second development set (`tst2010`) and on the official test sets of the evaluation campaign (`tst2011`, `tst2012`). Note that the tokenizers and the scorer applied here are different from those used for official evaluation.

2.6. Evaluation Specifications

ASR—For the evaluation of ASR submissions, participants had to provide automatic transcripts of test talk recordings. The talks were accompanied by an UEM file that marked the portion of each talk that needed to be transcribed. Specifically excluded were the beginning portions of each talk containing a jingle and possibly introductory applause, and the applause and jingle at the end of each file after the speaker has concluded his talk. Also excluded were larger portions of the talks that did not contain the lecturer's speech.

In addition, the UEM file also provides a segmentation of each talk into sentence-like units. The segmentation was that at sentence-level used in the MT track (Section 2.2). While giving human-defined segmentation makes the transcription task easier than it would be in real life, the use of it facilitates the speech translation evaluation since the segmentation of the input language perfectly matches the segmentation of the reference translation used in evaluating the translation task.

Participants were required to provide the results of the au-

	%bleu	σ	mtr	σ	ter	σ
En-Fr						
dev2010	26.28	0.59	47.57	0.47	56.80	0.70
tst2010	28.74	0.47	49.63	0.37	51.30	0.47
tst2011	34.95	0.70	54.53	0.51	44.11	0.60
tst2012	34.89	0.61	54.68	0.44	43.35	0.50
Ar-En						
dev2010	24.70	0.54	48.66	0.39	55.41	0.59
tst2010	23.64	0.45	47.61	0.34	57.16	0.50
tst2011	22.66	0.49	46.37	0.37	60.27	0.59
tst2012	24.05	0.44	48.62	0.31	54.72	0.43
De-En						
dev2010	28.14	0.60	52.83	0.40	50.37	0.57
tst2010	26.18	0.48	50.86	0.34	52.59	0.50
tst2011	30.28	0.51	55.00	0.32	47.86	0.47
tst2012	26.55	0.48	50.99	0.32	52.42	0.46
Nl-En						
dev2010	23.79	0.62	47.04	0.49	57.14	0.64
tst2010	31.23	0.48	54.62	0.32	47.90	0.45
tst2011	33.45	0.55	56.31	0.36	45.11	0.49
tst2012	29.89	0.46	53.16	0.31	47.60	0.42
Pl-En						
dev2010	20.56	0.58	44.74	0.46	62.47	0.67
tst2010	15.27	0.36	40.03	0.31	69.95	0.47
tst2011	18.68	0.42	43.64	0.32	65.42	0.53
tst2012	15.89	0.39	39.11	0.32	68.56	0.48
Ptb-En						
dev2010	33.57	0.64	56.06	0.41	45.53	0.57
tst2010	35.27	0.47	58.85	0.31	43.01	0.43
tst2011	38.56	0.54	61.26	0.32	39.87	0.45
tst2012	40.74	0.50	62.09	0.29	37.96	0.40
Ro-En						
dev2010	29.30	0.57	53.26	0.40	49.54	0.56
tst2010	28.18	0.47	52.32	0.33	51.13	0.46
tst2011	32.46	0.52	55.92	0.34	45.99	0.48
tst2012	29.08	0.48	52.73	0.33	50.32	0.45
Ru-En						
dev2010	17.37	0.50	41.63	0.40	66.96	0.60
tst2010	16.82	0.37	41.93	0.29	66.28	0.47
tst2011	19.11	0.42	43.82	0.32	62.63	0.49
tst2012	17.44	0.39	41.73	0.31	63.94	0.43
Sk-En						
dev2012	19.23	0.42	42.65	0.32	62.03	0.46
tst2012	21.79	0.58	45.01	0.41	58.28	0.55
Sl-En						
dev2012	15.90	0.45	40.16	0.36	67.23	0.53
tst2012	14.33	0.39	39.42	0.33	69.20	0.50
Tr-En						
dev2010	11.13	0.40	36.29	0.37	78.25	0.54
tst2010	12.13	0.32	37.87	0.27	75.56	0.45
tst2011	13.23	0.37	39.21	0.30	74.00	0.49
tst2012	12.45	0.33	38.76	0.29	73.63	0.43
Zh-En						
dev2010	9.62	0.39	33.97	0.36	82.47	1.01
tst2010	11.39	0.32	36.80	0.28	75.99	0.76
tst2011	14.13	0.39	39.62	0.32	65.02	0.42
tst2012	12.33	0.33	37.67	0.30	67.80	0.39

Table 5: Performance of baselines in terms of %BLEU, METEOR (mtr) and TER scores, with standard deviations (σ). Values were computed in case-punctuation sensitive mode.

tomatic transcription in CTM format. Multiple submissions were allowed, but one submission had to be marked as the primary run.

The quality of the submissions was then scored in terms of word error rate (WER). The results were scored case-insensitive, but were allowed to be submitted case-sensitive. Numbers, dates, etc. had to be transcribed in words as they are spoken, not in digits. Common acronyms, such as NATO and EU, had to be written as one word, without any special markers between the letters. This applies no matter if they are spoken as one word or spelled out as a letter sequence. All other letter spelling sequences had to be written as individual letters with spaces in between. Standard abbreviations, such as "etc." and "Mr." were accepted as specified by the GLM file in the scoring package that was provided to participants for development purposes. For words pronounced in their contracted form, it was permitted to use the orthography for the contracted form, as these were normalized into their canonical form according to the GLM file.

SLT/MT—The participants to the SLT and MT tracks had to provide the results of the translation of the test sets in NIST XML format. The output had to be true-cased and had to contain punctuation. Participants to the SLT track could either use the audio files directly, or use automatic transcriptions selected from the ASR submissions (Table 4).

The quality of the translations was measured automatically with BLEU [9] by scoring against the human translations created by the TED open translation project, and by human subjective evaluation (Section 5).

The evaluation specifications for the SLT/MT tracks were defined as case-sensitive with punctuation marks (*case+punc*). Tokenization scripts were applied automatically to all run submissions prior to evaluation.

Moreover, automatic evaluation scores were also calculated for case-insensitive (lower-case only) translation outputs with punctuation marks removed (*no_case+no_punc*). Besides BLEU, six additional automatic standard metrics (METEOR [10], WER [11], PER [12], TER [13], GTM [14], and NIST [15]) were calculated offline.

3. OLYMPICS Task

As a continuation of previous spoken dialog translation tasks [16, 17], this year's IWSLT featured a translation task in the Olympics domain. The OLYMPICS task is a small-vocabulary task focusing on human dialogs in travel situations where the utterances were annotated with dialog and speaker information that could be exploited by the participant to incorporate contextual information into the translation process.

3.1. Task Definition

The translation input condition of the OLYMPICS task consisted of correct recognition results, i.e., text input. Participants of the OLYMPICS task had to translate the Chinese sentences into English.

The monolingual and bilingual language resources that could be used to train the translation engines for the primary

runs were limited to the supplied corpora described in Section 3.2. These include all supplied development sets, i.e., the participants were free to use these data sets as they wish for tuning model parameters or as training bitext, etc. All other language resources, such as any additional dictionaries, word lists, or bitext corpora were treated as "additional language resources".

3.2. Supplied Data

The OLYMPICS task was carried out using parts of the Olympic Trilingual Corpus (HIT), a multilingual corpus that covers 5 domains (traveling, dining, sports, traffic and business) closely related to the Beijing 2008 Olympic Games [18]. It includes dialogs, example sentences, articles from the Internet and language teaching materials.

Moreover, the Basic Travel Expression Corpus (BTEC) [19], a multilingual speech corpus containing tourism-related sentences, was provided as an additional training corpus. The BTEC corpus consists of 20k training sentences and the evaluation data of previous IWSLT evaluation campaigns [17].

Both corpora are aligned at sentence level. Table 6 summarizes the characteristics of the Chinese (*zh*) and English (*en*) training (*train*), development (*dev*) and evaluation (*eval*) data sets. The first two columns specify the given data set and its type. The source language text ("*text*") and target language reference translation ("*ref*") resources also include annotated sample dialogs ("*dialog*") and their translation into the respective language ("*lang*"). The number of sentences are given in the "*sent*" column, and the "*avg.len*" column shows the average number of characters/words per training sentence for Chinese/English, respectively. The reported figures refer to tokenized texts.

The BTEC development data sets include up to 16 English reference translations for 3k Chinese inputs sentences. For the HIT data sets, only single reference translations were available.

For each sentence of the HIT corpus, context information on the *type of text* (dialog, samples, explanation), *scene* (airplane, airport, restaurant, water/winter sports, etc.), *topic* (asking about traffic conditions, bargaining over a price, front desk customer service, etc.), and the *speaker* (customer, clerk, passenger, receptionist, travel agent, etc.) was provided to the participants.

The dialogs of the two development and the evaluation data sets were randomly extracted from the HIT corpus after disregarding dialogs containing too short (less than 5 words) or too long (more than 18 words) sentences. The evaluation and development data sets included a total of 123 and 157 dialogs consisting on average of 8 and 13 utterances, respectively.

The supplied resources were released to the participants three months ahead of the official run submission period. The official run submission period was limited to one week.

Table 6: Supplied Data (OLYMPICS)

BTEC	data	lang	sent	avg.len	token	voc
train	(text)	Zh	19,972	11.8	234,998	2,483
	(text)	En	19,972	9.1	182,627	8,344
dev	(text)	Zh	2,977	9.4	27,888	1,515
	(ref)	En	38,521	8.1	312,119	5,927

HIT	data	lang	sent	avg.len	token	voc
train	(text)	Zh	52,603	13.2	694,100	4,280
	(text)	En	52,603	9.5	515,882	18,964
dev1	(dialog)	Zh	1,050	12.8	13,416	1,296
	(ref)	En	1,050	9.6	10,125	1.992
dev2	(dialog)	Zh	1,007	13.3	13,394	1,281
	(ref)	En	1,007	10.0	10,083	1,900
eval	(dialog)	Zh	998	14.0	14,042	1,310
	(ref)	En	998	10.6	10,601	2,023

3.3. Run Submissions

Participant registered for the OLYMPICS translation task had to submit at least one run. Run submission was carried out via email to the organizers with multiple runs permitted. However, the participant had to specify which runs should be treated as *primary* (evaluation using human assessments and automatic metrics) or *contrastive* (automatic evaluation only). Re-submitting runs was allowed as far as they were submitted prior to the submission deadline.

In total, 4 research groups participated in the OLYMPICS task and 4 primary and 4 contrastive runs were submitted.

3.4. Evaluation Specifications

The evaluation specification for the OLYMPICS task was defined as case-sensitive with punctuation marks (*case+punc*). The same tokenization script was applied automatically to all run submissions and reference data sets prior to evaluation. In addition, automatic evaluation scores were also calculated for case-insensitive (lower-case only) MT outputs with punctuation marks removed (*no_case+no_punc*).

All primary and contrastive run submissions were evaluated using the standard automatic evaluation metrics described in Section 2.6 for both evaluation specifications (see Appendix A).

In addition, human assessments of the overall translation quality of a single MT system were carried out with respect to the *adequacy* of the translation with and without taking into account the context of the respective dialog. The differences in translation quality between MT systems were evaluated using a paired comparison method that adopts a round-robin tournament structure to determine a complete system ranking, as described in Section 5.

4. Participants

A list of the participants of this year's evaluation is shown in Table 7. In total, 14 research teams from 11 countries took part in the IWSLT 2012 evaluation campaign. The number of primary and contrastive run submissions for each tasks

Table 7: List of Participants

FBK	Fondazione Bruno Kessler, Italy [20, 21]
HIT	Harbin Institute of Technology, China [22]
KIT	Karlsruhe Institute of Technology, Germany [23]
KIT-NAIST	KIT& NAIST collaboration [24, 25]
KYOTO-U	Kyoto University, Kurohashi-Kawahara Lab, Japan [26]
LIG	Laboratory of Informatics of Grenoble, France [27]
MITLL	Mass. Institute of Technology/Air Force Research Lab., USA [28]
NAIST	Nara Institute of Science and Technology, Japan [29]
NAIST-NICT	NAIST& NICT collaboration [30]
NICT	National Institute of Communications Technology, Japan [31, 32]
PJIIT	Polish-Japanese Institute of Information Technology, Poland [33]
POSTECH	Pohang University of of Science and Technology, Korea [34]
RACAI	Research Institute for AI of the Romanian Academy, Romania [35]
RWTH	Rheinisch-Westfälische Technische Hochschule Aachen, Germany [36]
TUBITAK	TUBITAK - Center of Research for Advanced Technologies, Turkey [37]
UEDIN	University of Edinburgh, UK [38]

	TED													OLY
	ASR	SLT	MT											MT
	En	$EnFr$	$EnFr$	$ArEn$	$DeEn$	$NlEn$	$PlEn$	$PtbEn$	$RoEn$	$RuEn$	$SkEn$	$TrEn$	$ZhEn$	$ZhEn$
FBK	X		X	X	X	X					X	X		
HIT														X
KIT	X	X	X											
KYOTO-U														X
LIG			X											
MITLL	X	X	X	X										
NAIST	X		X	X	X	X	X	X	X	X	X	X	X	
NICT	X									X				X
PJIIT							X							
POSTECH														X
RACAI									X					
RWTH	X	X	X	X	X						X		X	
TUBITAK				X								X		
UEDIN	X	X	X		X									
	7	4	7	5	4	2	2	1	2	2	3	3	2	4

are summarized in Table 8. In total, 48 primary runs and 54 contrastive runs were submitted by the participants.

Table 8: Run Submissions

Task	Primary (Contrastive) [Systems]
TED ASR$_{En}$	7 (8) [FBK,KIT,KIT-NAIST,MITLL,NICT,RWTH,UEDIN]
TED SLT$_{EnFr}$	4 (8) [KIT,MITLL,RWTH,UEDIN]
TED MT$_{EnFr}$	7 (13) [FBK,KIT,LIG,MITLL,NAIST,RWTH,UEDIN]
TED MT$_{ArEn}$	5 (5) [FBK,MITLL,NAIST,RWTH,TUBITAK]
TED MT$_{DeEn}$	4 (5) [FBK,NAIST,RWTH,UEDIN]
TED MT$_{NlEn}$	2 (2) [FBK,NAIST]
TED MT$_{PlEn}$	2 (2) [NAIST,PJIIT]
TED MT$_{PtbEn}$	1 (0) [NAIST]
TED MT$_{RoEn}$	2 (4) [NAIST,RACAI]
TED MT$_{RuEn}$	2 (1) [NAIST,NICT]
TED MT$_{SkEn}$	3 (0) [FBK,NAIST,RWTH]
TED MT$_{TrEn}$	3 (1) [FBK,NAIST,TUBITAK]
TED MT$_{ZhEn}$	2 (1) [NAIST,RWTH]
OLY MT$_{ZhEn}$	4 (4) [HIT,KYOTO-U,NAIST-NICT,POSTECH]

5. Human Evaluation

Subjective evaluation was carried out on all primary runs submitted by participants to the official tracks of the TED task, namely the SLT track (English-French) and the MT *official* track (English-French and Arabic-English) and to the OLYMPICS task (Chinese-English).

For each task, systems were evaluated using a subjective evaluation set composed of 400 sentences randomly taken from the test set used for automatic evaluation. Each evaluation set represents the various lengths of the sentences included in the corresponding test set, with the exception of sentences with less than 5 words, which were excluded from the subjective evaluation.

Two metrics were used for the IWSLT 2012 subjective evaluation, i.e. *System Ranking* evaluation and, only for the OLYMPICS task, *Adequacy* evaluation.

The goal of the *Ranking* evaluation is to produce a complete ordering of the systems participating in a given task [39]. In the ranking task, human judges are given two MT outputs of the same input sentence as well as a reference translation and they have to decide which of the two translation hypotheses is better, taking into account both the content and fluency of the translation. Judges are also given the possibility to assign a tie in case both translations are equally good or bad. The judgments collected through these pairwise

comparisons are then used to produce the final ranking.

Following the practice consolidated in the previous campaign, the ranking evaluation in IWSLT 2012 was carried out by relying on crowd-sourced data. All the pairwise comparisons to be evaluated were posted to Amazon's Mechanical Turk[6] (MTurk) through the CrowdFlower[7] interface. Data control mechanisms including *locale qualifications* and *gold units* (items with known labels which enable distinguishing between trusted and untrusted contributors) implemented in CrowdFlower were applied to ensure the quality of the collected data [40].

For each pairwise comparison we requested three redundant judgments from different MTurk contributors. This means that for each task we collected three times the number of the necessary judgments. Redundant judgment collection is a typical method to ensure the quality of crowd-sourced data. In fact, instead of relying on a single judgment, label aggregation is computed by applying majority voting. Moreover, agreement information can be collected to find and manage the most controversial annotations.

In our ranking task, there are three possible assessments: *(i)* output A is better than output B, *(ii)* output A is worse than output B, or *(iii)* both output A and B are equally good or bad (tie). Having three judgements from different contributors and three possible values, it was not possible to assign a majority vote for a number of comparisons. These *undecidable comparisons* were interpreted as a tie between the systems (neither of them won) and were used in the evaluation.

In order to measure the significance of result differences for each pairwise comparison, we applied the Approximate Randomization Test[8]. The results for all the tasks are available in Appendix B.

Besides system ranking, an additional evaluation metrics was used in the OLYMPICS task, where the overall translation quality of a single run submission was also evaluated according to the translation *adequacy*, i.e., how much of the information from the source sentence was expressed in the translation with and without taking into account the context of the respective dialog. Details on the adequacy evaluation are given in Section 5.2.2.

Finally, in order to investigate the degree of consistency between human evaluators, we calculated inter-annotator agreement[9] using the *Fleiss' kappa coefficient* κ [42, 43]. This coefficient measures the agreement between multiple raters (three in our evaluation) each of whom classifies N items into C mutually exclusive categories, taking into account the agreement occurring by chance. It is calculated as:

$$\kappa = \frac{P(a)-P(e)}{1-P(e)}$$

where $P(a)$ is the observed pairwise agreement between the raters and $P(e)$ is the estimated agreement due to chance, calculated empirically on the basis of the cumulative distribution of judgments by all raters. If the raters are in complete agreement then κ = 1. If there is no agreement among the raters (other than what would be expected by chance) then $\kappa \leq 0$. The interpretation of the κ values according to [44] is given in Table 9.

Table 9: Interpretation of the κ coefficient.

κ	**Interpretation**
< 0	No agreement
0.0 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Within this common evaluation framework, different procedures were applied to the TED and OLYMPICS tasks.

5.1. TED Task

For the TED Task, subjective ranking was performed on the *Progress Test*, i.e. on the 2011 evaluation set[10], with the goal of measuring the progress of SLT and MT with respect to the top-ranked 2011 systems.

As a major novelty for this year, a change in the type of tournament used for the ranking evaluation was introduced. In IWSLT 2011, we adopted a round robin tournament, which is the most accurate way to determine system ranking due to its completeness (each system competes against every other system). The drawback of round robin is that completeness comes at a high cost, due to the large number of comparisons to be carried out. Thus, our goal for this year's evaluation was to adopt a tournament structure comparable with round robin in terms of reliability, but requiring less comparisons in favor of cost effectiveness.

Existing studies about the efficacy of sport tournament structures [45] demonstrated that knockout tournaments are comparable to round robin, if double elimination procedures are used and the allocation of players to the tournament structure is accurately assigned a-priori according to some criterion (seeding). The most promising structure, given its ability of ranking all players and the relatively few comparisons required, is the *Double Seeded Knockout with Consolation* (DSKOC) tournament.

In the DSKOC scheme proposed in [45], each player must loose twice before elimination from the tournament. The loss of one game does not therefore preclude that player from winning that tournament, provided that all future contests are won. Consolation play-offs are allowed at each stage of the tournament in order to place all players, and the a-priori seeding protocol is: P_1 - P_8, P_5 - P_4, P_3 - P_6, P_7 - P_2.

[6] http://www.mturk.com

[7] http://www.crowdflower.com

[8] To calculate Approximate Randomization we used the package available at: http://www.nlpado.de/∼sebastian/software/sigf.shtml [41]

[9] Agreement scores are presented in Section 5.1, Section 5.2, and in Appendix B.

[10] The reference translations for the 2011 test set were never released to the participants.

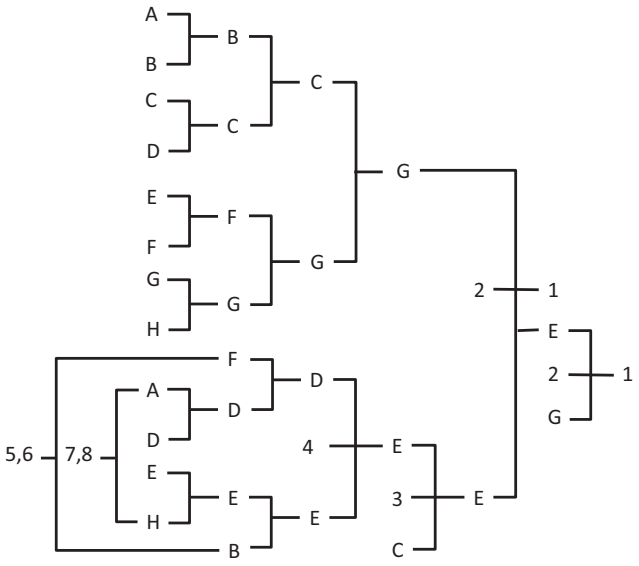


Figure 1: Example of double knockout tournament with consolation at each stage of the competition

Figure 1 shows an example of a DSKOC tournament structure.

The DSKOC scheme was tested on the IWSLT 2011 data. For all IWSLT 2011 tasks, the system ranking obtained with DSKOC was the same as the one obtained with the round robin scheme. Therefore, the DSKOC tournament was adopted with an 8-player scheme for the human evaluation of the IWSLT 2012 TED tasks. For the a-priori seeding, we used the *BLEU* scores obtained by the systems on the full 2011 test set.

Our evaluation scenario raises two issues that differentiate it from real sport tournaments, namely:

1. *Tied matches*. In case of tied outcome - i.e. equal number of evaluation sentences for which one system was judged better than the other - majority voting was not applied. Instead, we took into account all the judgments of each match and calculated which system obtained the highest number of "win" judgments[11].

2. *Systems competing more than once against each other*. The idea of giving a chance of recovery from an aberrant result, which is at the basis of the double elimination scheme in real sport tournaments, is not viable in our scenario were crowd-sourced judgments are collected only once. Thus, if two systems have to be compared twice, a second evaluation is not run and the results of the first comparison are used.

Our aim for IWSLT 2012 was not only to evaluate all the primary runs submitted for IWSLT 2012, but also to assess their progress with respect to the best 2011 systems. Given that an 8-player tournament was adopted, different system selection criteria were applied, depending on the participants in each track.

- **SLT$_{\mathbf{EnFr}}$**: all four 2012 runs were evaluated together with the four best runs of 2011[12].

- **MT$_{\mathbf{ArEn}}$**: all five 2012 runs were evaluated together with a baseline created by the organizers and the top two runs of 2011. In this track, only the top two 2011 systems were selected. This is due to the fact that the evaluation of last year showed a large gap between the two top-ranked systems and the last two systems, which obtained poor results both in terms of automatic metrics and subjective ranking.

- **MT$_{\mathbf{EnFr}}$**: as eight primary runs were submitted this year, two subsequent tournaments were carried out. In the first tournament, only the bottom four runs of 2012 were ranked. The top four runs of 2012 were ranked jointly with the top four 2011 runs in the second tournament.

A summary of the TED *Ranking* task is given in Table 10. Concerning the number of different matches, not all the comparisons required in a standard scenario[13] were crowdsourced in this evaluation, because *(i)* already evaluated matches were not repeated, and *(ii)* the results for the matches involving 2011 systems were taken from the IWSLT 2011 evaluation data. As far as inter-annotator agreement is concerned, all the three tracks are in the range of "Fair agreement". These results are comparable with those obtained last year and confirm the general trend among the tracks, where SLT$_{EnFr}$ shows the lowest agreement rate and MT$_{EnFr}$ the highest one.

Table 10: Summary of the TED *Ranking* task

Task	# ranked systems	# different HtH matches	I.A.A. (κ)
SLT$_{EnFr}$	8	15 (3 from 2011)	0.2263
MT$_{ArEn}$	8	14 (0 from 2011)	0.2496
MT$_{EnFr}$	12	21 (3 from 2011)	0.2861

For each TED track, Appendix B provides the system rankings according to the BLEU scores and the human evaluation results, as well as the complete structure of the tournaments and detailed tables about pairwise head-to-head comparisons.

5.2. OLYMPICS Task

The human evaluation of the new IWSLT 2012 translation task on the Olympics domain was performed with respect to system ranking and dialog adequacy. Both methodologies are described below.

[11] In other words, ties were resolved considering all the 1,200 judgments collected for the 400 evaluation sentences, instead of using the 400 labels resulting from majority voting.

[12] All the best 2011 systems where chosen according to the results of last year's human evaluation.

[13] In an 8-player DSKOC structure, 16 or 17 matches are necessary to complete the tournament.

5.2.1. System ranking

Following last year's evaluation methodology, system ranking for the OLYMPICS task was achieved through a paired comparison method that adopts a round-robin tournament structure. Round-robin is the most complete way to determine system ranking as it ensures a full coverage of paired comparisons between systems. We first prepared all the paired comparisons necessary for a complete round-robin over the selected 400 evaluation sentences (m=#sentences). Each system was evaluated against each of the other systems for each evaluation sentence. Considering all systems (n=#systems), there are $n(n-1)/2$ pairwise comparisons for each evaluation sentence, and thus $m * n(n-1)/2$ comparisons for the whole evaluation set. The complete ranking of the four system submissions to the task (n=4) using 400 evaluation sentences (m=400) required 2,400 comparisons.

Table 11: *Paired Comparison* Evaluation.

# systems	# comparisons per system	# comparison in total	# collected judgments	I.A.A. κ
4	1,200	2,400	7,200	0.3653

A summary of the OLY_{ChEn} paired comparison task is given in Table 11. As far as inter-annotator agreement is concerned, the results obtained compare well with the overall results obtained last year, falling in the class of "Fair agreement". The complete ranking of the systems and the results of all the pairwise comparisons are given in Appendix B.4.

5.2.2. Dialog Adequacy

In addition to the system ranking based on paired comparison, human assessments of the overall translation quality of a single MT system were carried out with respect to the *Adequacy* of the translation for all OLYMPICS task run submissions. For *Adequacy*, the evaluator was presented with the source language input as well as a reference translation and had to judge how much of the information from the original sentence was expressed in the translation [46]. The *Adequacy* judgments consisted of one of the grades listed in Table 12.

Table 12: Dialog Adequacy

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

In addition to the above standard metrics, a modified version of the *adequacy* metrics (*dialog*) that takes into account information beyond the current input sentence was applied to the translation results of the OLYMPICS task in order to judge a given MT output in the context of the respective dialog. For the *dialog* assessment, the evaluators were presented with the history of previously uttered sentences, the input sentence and the reference translation. The evaluator had to read the dialog history first and then had to judge how much of the information from the reference translation is expressed in the translation in the context of the given dialog history by assigning one of the *dialog* grades listed in Table 12. In cases where parts of the information were omitted in the system output, but they could be understood in the context of the given dialog, such omission would not result in a lower *dialog* score. For the final adequacy metric scores, each system score was calculated as the *median* of the assigned grades. The adequacy evaluation was carried out by an expert grader trained on the given tasks.

The *adequacy* evaluation results of all run submissions are summarized in Appendix B.4. The *dialog* assessment was carried out one week after the *adequacy* evaluation was finished. In order to reduce evaluation costs, only the best performing system (HIT) according to the *adequacy* metric was selected for the subjective evaluation using the *dialog* metric. We measured the *intra-grader* consistency[14] and obtained a κ coefficient of 0.51 (*moderate* agreement) and 0.74 (*substantial* agreement) for the *adequacy* and *dialog* assessment, respectively.

6. Main Findings

In this section, we point out the methods and solutions that, according to the participants' descriptions, contributed most to the performance of their systems. Our intent is to provide some useful suggestions for setting up strong baselines for each track for the benefit of future participants or any interested researcher. The complete list of the system description papers that we consulted is included in the references and can be found in Table 7.

6.1. TED Task

In the following, we briefly comment on the general outcomes of each track and point out relevant features of the systems that participated this year. Notice that our selection cannot be considered exhaustive nor objective.

6.1.1. ASR Track

Seven teams participated this year in the ASR track. A comparison of the 2011 and 2012 results on the progress test set is given in Appendix A.2. We indeed observe a significant drop in WER[15] between the two best performing systems, from 13.5% to 10.9%. Remarkable progress is observed for all teams that participated in both editions.

All the ASR system developed this year have complex architectures performing multiple adaptation and system combination steps. Some of their relevant aspects are briefly highlighted:

[14] The proportion of times that the same judge assigns the same grade when assessing the same system output twice.

[15] Notice that these figures differ from those reported in [47] as the references were afterwards manually improved.

Acoustic training data: The NICT system was trained only on TED recordings, roughly 170h of speech, which means much less data was used than for other systems.
Acoustic front-end: The best performing systems employed multiple acoustic front-ends, including MLPs (KIT, RWTH) and deep NN features (UEDIN), to lower feature dimensionality.
Acoustic models: The top performing systems employed AMs trained on different acoustic features and with different methods, combining SAT and discriminative methods.
Language models: The NICT and MITLL engines include a RNN LM for n-best re-scoring. All participants used n-gram LMs adapted via data selection and interpolation, both before and after decoding. FBK reports comparable results when adaptation is done after decoding.

6.1.2. SLT Track

Four teams participated in this track. Subjective rankings were carried out on the progress test by considering all 2012 primary SLT runs and the four best SLT runs of 2011. Detailed automatic scores and subjective ranking results are reported in Appendices A and B, respectively. The reported BLEU rankings on the current and progress tests result are consistent and statistically significant. According to the subjective ranking, the top three 2012 systems are better than the top 2011 run. Notice, however, that the subjective ranking of the 2012 runs differs from the corresponding BLEU ranking.

Participants in the SLT track used their own ASR system output which was post-processed in order to match the standard MT input conditions (punctuation and casing). MT was performed on the single best ASR output using the same engine as the French-English MT track, or after minor changes.

6.1.3. MT Track

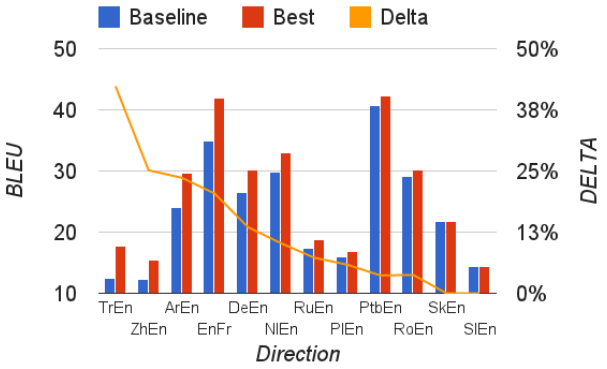
The official English-French and Arabic-English tracks had 7 participants, 5 respectively. For English-French, the BLEU rankings on the current and progress tests differ slightly. Subjective ranking on the progress test was carried out with two subsequent tournaments: one to select the top four runs of 2012, and another to determined their ranking jointly with the top four systems in the 2011 runs. The final outcome tells that the best two 2012 runs improved over the best 2011 run, and that the top three 2012 runs had identical BLEU ranks.

For Arabic-English, BLEU rankings on the current and progress tests are also slightly different. Subjective ranking was performed on the progress test by also including the best two 2011 runs. The best 2012 run ranks above the best 2011 run. The best two 2012 teams also improved their own 2011 runs. The subjective and BLEU rankings are again in perfect agreement.

A comparison between the baseline and the best performing systems is given in Figure 2.

The work carried out by the participants of TED MT tasks focused on the following aspects:

Figure 2: TED MT track: best runs vs. baselines



Data selection and adaptation: Basically all participants exploited data selection on the available out-of-domain resources to reduce the size and improve the accuracy of translation and language models. Out-domain models are combined by linear interpolation, log-linear interpolation, union (KIT,UEDIN), or fill-up (FBK). UEDIN also performed translation model adaptation with sparse lexical features.
Language model: Some well performing MT systems used class-based (KIT,RWTH) or hybrid (FBK) LMs to model the style of talks. KIT reports slight improvements with a continuous space LM (Restricted Bolzmannn Machine) applied during decoding.
Translation model: RWTH employed an improved phrase-extraction method that drastically reduces the size of the phrase-table. RWTH also reports gains with HPBT on Chinese-English, and FBK on Turkish-English. On the other side, UEDIN reports better results with PBT on German-English and French-English.
Reordering: For distant translation directions, RWTH and KIT applied POS based re-ordering rules, while FBK applied a hierarchical orientation model and early distortion cost estimates.
System combination: RWTH reports significant gains through confusion-network-based system combination.
To conclude, a few remarks concerning language specific issues. **Arabic and Chinese**: RWTH reports improvements by combining MT systems using multiple word segmentation models. For Chinese, RWTH also employs MT decoders processing text in reverted word order. **Turkish**: FBK reports relevant gains by using morphological segmentation and HPBT models. **Polish**: PJIIT reported negative results by applying morpho-syntactic factored models.

6.2. OLYMPICS Task

Four teams participated in the OLYMPICS task using quite different MT architectures including phrase-based SMT (HIT, NICT), syntax-based SMT (POSTECH), and syntax-based EBMT (KYOTO-U) approaches. The difficulty of this year's dialog translation tasks lay in the handling of out-of-vocabulary words and the sentence structure differences

(non-parallel sentence) of the supplied language resources, leading to lower evaluation scores for the structured-based MT approaches.

The work carried out by the participants of the OLYMPICS task focused on the following aspects:

Data preprocessing: The pre-processing of the Chinese language resources was carried out using the Stanford word segmenter [3] with the PKU model (HIT, NAIST-NICT) and in-house segmenters (KYOTO-U, POSTECH). For English, all participants only applied simple tokenization scripts. In addition, KYOTO-U applied sub-sentence splitting and non-parallel sentences filtering to improve the bilingual sentence alignment quality of the supplied corpus.

Additional language resources: KYOTO-U investigated the effects of using of external resources such as Wikipedia in order to reduce the out-of-vocabulary problem. Unfortunately, none of the participants used the dialog and speaker information annotated in the supplied corpus.

Translation model: HIT focused on model combination of phrase tables generated by *GIZA++* and *Pialign*.

Decoding: NICT extended the Minimum Bayes Risk decoding approach by considering maximum a-posteriori translation similarities and by taking advantage of the nearest neighbors of the source sentence. POSTECH focused on a forest-to-string machine translation approach based on binarized dependency forests. KYOTO-U carried out a tree-based decoding approach that uses an example-based MT (EBMT) system and integrates a Bayesian subtree alignment model based on dependency trees.

Clear and consistent rankings were obtained for human assessment using both paired comparison and adequacy metrics. Differences between all systems were statistically significant. Moreover, a comparison of the *adequacy* and *dialog* score differences of this year's and previous dialog translation tasks [16, 17] indicate that *dialog* metrics more closely reflect the reluctance of humans to accept machine translated output when taking into account the context of the conversation across different dialog types and domains.

7. Conclusions

We presented the organization and outcomes of the 2012 IWSLT Evaluation Campaign. This year the evaluation introduced several novelties: a small vocabulary translation tasks (OLYMPICS), unofficial TED talk MT tasks from 10 different languages into English, the use of a progress test set to compare this year's systems with the best runs of last year, and finally the adoption of new a tournament scheme to run the subjective evaluation on the official tracks. 16 teams participated in the evaluation, submitting a total of 48 primary runs. According to the automatic and subjective rankings of the official tracks on the progress test, performance was improved over the best results of last year. For the unofficial track, results by Moses baseline systems were made available for all 10 language pairs. For most of the tasks, participants were able to perform significantly better than the baseline. The plan for 2013 is to include additional unofficial language pairs and to adopt as progress test the 2012 test set, which for this reason will not be publicly released.

8. Acknowledgements

Research Group 3-01' received financial support by the '*Concept for the Future*' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative. The work leading to these results has received funding from the European Union under grant agreement no 287658 — Bridges Across the Language Divide (EU-BRIDGE). The subjective evaluation of the Evaluation Campaign was financed by a grant of the European Association for Machine Translation.

9. References

[1] M. Cettolo, C. Girardi, and M. Federico, "WIT^3: Web Inventory of Transcribed and Translated Talks," in *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012. [Online]. Available: http://hltshare.fbk.eu/EAMT2012/html/Papers/59.pdf

[2] M. Diab, K. Hacioglu, and D. Jurafsky, "Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks," in *HLT-NAACL 2004: Short Papers*, D. M. Susan Dumais and S. Roukos, Eds. Boston, Massachusetts, USA: Association for Computational Linguistics, May 2 - May 7 2004, pp. 149–152.

[3] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning, "A conditional random field word segmenter," in *Fourth SIGHAN Workshop on Chinese Language Processing*, 2005.

[4] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand, September 2005, pp. 79–86.

[5] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open Source Toolkit for Statistical Machine Translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, 2007, pp. 177–180. [Online]. Available: http://aclweb.org/anthology-new/P/P07/P07-2045.pdf

[6] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 4, no. 13, pp. 359–393, 1999.

[7] M. Federico, N. Bertoldi, and M. Cettolo, "IRSTLM: an Open Source Toolkit for Handling Large Scale Language Models," in *Proceedings of Interspeech*, Melbourne, Australia, 2008, pp. 1618–1621.

[8] J. Clark, C. Dyer, A. Lavie, and N. Smith, "Better hypothesis testing for statistical machine translation: Controlling for optimizer instability," in *Proceedings of the Association for Computational Lingustics*, ser. ACL 2011. Portland, Oregon, USA: Association for Computational Linguistics, 2011, available at http://www.cs.cmu.edu/ jhclark/pubs/significance.pdf.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, USA, 2002, pp. 311–318.

[10] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments," in *Proceedings of the Second Workshop on Statistical Machine Translation (WMT)*, Prague, Czech Republic, 2007, pp. 228–231.

[11] S. Niessen, F. J. Och, G. Leusch, and H. Ney, "An Evaluation Tool for Machine Translation: Fast Evaluation for Machine Translation Research," in *Proceedings of the Second International Conference on Language Resources & Evaluation (LREC)*, Athens, Greece, 2000, pp. 39–45.

[12] F. J. Och, "Minimum Error Rate Training in SMT," in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, 2003, pp. 160–167.

[13] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *Proceedings of the The Seventh Conference of the Association for Machine Translation in the Americas (AMTA)*, Cambridge, USA, 2006, pp. 223–231.

[14] J. P. Turian, L. Shen, and I. D. Melamed, "Evaluation of Machine Translation and its Evaluation," in *Proceedings of the MT Summmit IX*, New Orleans, USA, 2003, pp. 386–393.

[15] G. Doddington, "Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics," in *Proceedings of the Second International Conference on Human Language Technology (HLT)*, San Diego, USA, 2002, pp. 257–258.

[16] M. Paul, "Overview of the IWSLT 2009 Evaluation Campaign," in *Proceedings of the sixth International Workshop on Spoken Language Translation (IWSLT)*, Tokyo, Japan, 2010, pp. 1–18.

[17] M. Paul, M. Federico, and S. Stüker, "Overview of the IWSLT 2010 Evaluation Campaign," in *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, Paris, France, 2010, pp. 3–27.

[18] M. Yang, H. Jiang, T. Zhao, and S. Li, "Construct Trilingual Parallel Corpus on Demand," *Chinese Spoken Language Processing, Lecture Notes in Computer Science*, vol. 4274, pp. 760–767, 2006.

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

[20] D. Falavigna, R. Gretter, F. Brugnara, and D. Giuliani, "FBK @ IWSLT 2012 - ASR track," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[21] N. Ruiz, A. Bisazza, R. Cattoni, and M. Federico, "FBKs Machine Translation Systems for IWSLT 2012s TED Lectures," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[22] X. Zhu, Y. Cui, C. Zhu, T. Zhao, and H. Cao, "The HIT-LTRC Machine Translation System for IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[23] M. Mediani, Y. Zhang, T.-L. Ha, J. Niehues, E. Cho, T. Herrmann, R. Kärgely, and A. Waibel, "The KIT Translation systems for IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[24] C. Saam, C. Mohr, K. Kilgour, M. Heck, M. Sperber, K. Kubo, S. Stüker, S. Sakti, G. Neubig, T. Toda, S. Nakamura, and A. Waibel, "The 2012 KIT and KIT-NAIST English ASR Systems for the IWSLT Evaluation," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[25] M. Heck, K. Kubo, M. Sperber, S. Sakti, S. Stüker, C. Saam, K. Kilgour, C. Mohr, G. Neubig, T. Toda, S. Nakamura, and A. Waibel, "The KIT-NAIST (Contrastive) English ASR System for IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[26] C. Chu, T. Nakazawa, and S. Kurohashi, "EBMT System of Kyoto University in OLYMPICS Task at IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[27] L. Besancier, B. Lecouteux, M. Azouzi, and L. N. Quang, "The LIG English to French Machine Translation System for IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[28] J. Drexler, W. Shen, T. Gleason, T. Anderson, R. Slyh, B. Ore, and E. Hansen, "The MIT-LL/AFRL IWSLT-2012 MT System," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[29] G. Neubig, K. Duh, M. Ogushi, T. Kano, T. Kiso, S. Sakti, T. Toda, and S. Nakamura, "The NAIST Machine Translation System for IWSLT2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[30] H. Shimizu, M. Utiyama, E. Sumita, and S. Nakamura, "Minimum Bayes-Risk Decoding Extended with Two Methods: NAIST-NICT at IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[31] H. Yamamoto, Y. Wu, C.-L. Huang, X. Lu, P. Dixon, S. Matsuda, C. Hori, and H. Kashioka, "The NICT ASR System for IWSLT2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[32] A. Finch, O. Htun, and E. Sumita, "The NICT Translation System for IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[33] K. Marasek, "TED English-to-Polish translation system for the IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[34] H. Na and J.-H. Lee, "Forest-to-String Translation using Binarized Dependency Forest for IWSLT 2012 OLYMPICS Task," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[35] S. D. Dumitrescu, R. Ion, D. Stefanescu, T. Boros, and D. Tufis, "Romanian to English Automatic MT Experiments at IWSLT12," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[36] S. Peitz, S. Mansour, M. Freitag, M. Feng, M. Huck, J. Wuebker, M. Nuhn, M. Nußbaum-Thom, and H. Ney, "The RWTH Aachen Speech Recognition and Machine Translation System for IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[37] C. Mermer, "The TUBITAK Statistical Machine Translation System for IWSLT 2012," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[38] E. Hasler, P. Bell, A. Ghoshal, B. Haddow, P. Koehn, F. McInnes, S. Renals, and P. Swietojanski, "The UEDIN Systems for the IWSLT 2012 Evaluation," in *Proceedings of the ninth International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 2012.

[39] M. Federico, S. Stüker, L. Bentivogli, M. Paul, M. Cettolo, T. Herrmann, J. Niehues, and G. Moretti, "The IWSLT 2011 Evaluation Campaign on Automatic Talk Translation," in *Proceedings of the eighth International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012, pp. 3543–3550.

[40] L. Bentivogli, M. Federico, G. Moretti, and M. Paul, "Getting Expert Quality from the Crowd for Machine Translation Evaluation," in *Proceedings of the MT Summmit XIII*, Xiamen, China, 2011, pp. 521–528.

[41] S. Padó, *User's guide to* `sigf`*: Significance testing by approximate randomisation*, 2006.

[42] S. Siegel and N. J. Castellan, *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1988.

[43] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76(5), 1971.

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

[45] T. McGarry and R. Schutz, "Efficacy of traditional sport tournament structures," *The Journal of the Operational Research Society*, vol. 48(1), pp. 65–74, 1997.

[46] 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.

[47] M. Federico, L. Bentivogli, M. Paul, and S. Stüker, "Overview of the IWSLT 2011 Evaluation Campaign," in *Proceedings of the eighth International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, USA, 2011, pp. 11–27.

[48] Y. Zhang, S. Vogel, and A. Waibel, "Interpreting Bleu/NIST Scores: How Much Improvement do We Need to Have a Better System?" in *Proceedings of the Second International Conference on Language Resources & Evaluation (LREC)*, 2004, pp. 2051–2054.

Appendix A. Automatic Evaluation

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

A.1. Official Testset (*tst2012*)

· All the sentence IDs in the IWSLT 2012 testset were used to calculate the automatic scores for each run submission.
· ASR and MT systems are ordered according to the *WER* and *BLEU* metrics, respectively.
· For each task, the best score of each metric is marked with **boldface**.
· Besides the NIST metrics, all automatic evaluation metric scores are given as percent figures (%).
· Besides the ASR scores, the mean scores of 2000 iterations were calculated for each MT output according to the *bootStrap* method [48].
· Omitted lines between scores indicate non-significant differences in performance between the MT engines.

TED : ASR English (ASR$_{En}$)

System	WER	(Count)
NICT	**12.1**	**(2318)**
KIT-NAIST	12.4	(2392)
KIT	12.7	(2435)
MITLL	13.3	(2565)
RWTH	13.6	(2621)
UEDIN	14.4	(2775)
FBK	16.8	(3227)

TED : SLT English-French (SLT$_{EnFr}$)

"case+punc" evaluation							System	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
29.78	**59.35**	**53.56**	**44.94**	**50.89**	**60.17**	**6.730**	KIT	**31.09**	**58.35**	**53.40**	**45.15**	**51.86**	**59.73**	**7.031**
29.09	58.83	54.38	45.29	51.83	59.67	6.646	UEDIN	30.70	58.08	53.96	45.38	52.59	59.39	6.946
28.51	57.50	54.93	46.11	52.56	59.18	6.611	RWTH	29.96	56.95	54.37	46.13	53.07	58.90	6.901
24.67	55.59	61.05	50.93	58.44	55.86	5.908	MITLL	25.52	54.58	61.59	51.75	60.16	55.12	6.100

TED : MT English-French (MT$_{EnFr}$)

"case+punc" evaluation							System	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
40.65	**69.21**	42.02	34.91	39.96	68.95	7.969	UEDIN	39.22	**66.32**	44.73	37.09	43.32	67.02	8.031
40.44	68.74	**40.82**	**34.62**	**38.82**	**69.32**	**8.102**	KIT	**39.23**	65.94	**43.33**	**36.78**	**42.01**	**67.44**	**8.187**
39.45	68.01	42.49	35.82	40.60	68.30	7.916	NAIST	38.06	65.16	45.35	38.15	44.13	66.29	7.967
39.40	68.37	41.61	35.23	39.53	69.03	8.034	RWTH	38.16	65.46	44.22	37.57	42.98	67.04	8.099
37.58	67.23	43.00	35.96	41.00	68.04	7.856	LIG	36.04	64.27	45.72	38.31	44.44	65.98	7.892
37.27	66.76	44.15	36.91	42.27	67.16	7.712	FBK	35.73	63.78	47.05	39.40	45.77	64.93	7.740
32.93	64.34	50.09	41.49	47.77	64.02	6.980	MITLL	31.57	61.24	53.49	44.32	51.99	61.68	6.989

TED : MT Arabic-English (MT$_{ArEn}$)

"case+punc" evaluation							System	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
29.32	**65.71**	**50.86**	**41.79**	**48.18**	**63.23**	**7.046**	RWTH	**28.24**	**63.13**	**53.67**	**43.99**	**51.99**	**61.43**	**7.156**
27.87	63.85	54.45	44.57	51.63	61.03	6.656	FBK	26.40	61.03	57.94	47.28	55.98	58.79	6.686
25.33	61.14	56.57	46.70	54.01	59.06	6.356	NAIST	23.77	58.03	60.12	47.37	58.32	56.46	6.360
25.30	62.33	54.20	44.75	51.53	60.17	6.519	TUBITAK	23.90	59.38	57.53	48.64	55.77	57.89	6.568
19.32	61.59	61.29	51.85	53.61	53.37	5.390	MITLL	22.95	58.51	60.07	49.62	58.16	57.07	6.370

TED : MT German-English (MT$_{DeEn}$)

"case+punc" evaluation							System	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
29.84	**66.28**	**52.78**	**41.71**	**49.05**	**63.74**	**7.053**	RWTH	**28.85**	63.73	**54.90**	**43.25**	**52.10**	**62.20**	**7.269**
28.80	66.23	53.85	42.21	50.01	63.38	6.930	UEDIN	28.45	**64.00**	55.75	43.57	52.74	61.86	7.153
28.18	65.41	55.48	43.60	51.67	62.72	6.771	FBK	27.76	62.88	57.37	44.84	54.41	61.08	7.003
27.97	64.66	55.14	43.56	51.53	62.30	6.754	NAIST	26.95	62.00	57.54	45.31	54.66	60.36	6.934

TED : MT Dutch-English (MT$_{NlEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
32.69	**67.59**	**50.12**	**39.45**	**46.15**	**65.51**	**7.463**	FBK	**31.96**	**65.19**	**51.76**	**40.55**	**49.12**	**64.47**	**7.714**
30.97	66.14	51.80	40.94	47.68	64.09	7.238	NAIST	30.29	63.74	53.64	42.10	50.84	63.06	7.471

TED : MT Polish-English (MT$_{PlEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
16.66	**49.90**	**70.49**	**58.21**	**66.88**	**49.55**	**5.062**	NAIST	**15.33**	**46.27**	**73.38**	**60.60**	**71.04**	**47.08**	**5.151**
15.32	47.94	71.85	59.61	67.97	48.32	4.844	PJIIT	14.28	44.08	73.88	61.18	71.53	46.14	4.983

TED : MT Portuguese(Brazilian)-English (MT$_{PtbEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
41.67	**75.91**	**39.84**	**32.60**	**37.82**	**72.05**	**8.318**	NAIST	**40.01**	**73.45**	**42.77**	**34.89**	**41.29**	**70.02**	**8.399**

TED : MT Romanian-English (MT$_{RoEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
29.64	**65.19**	**52.41**	**43.06**	**49.90**	**62.91**	**6.931**	NAIST	**27.59**	**61.93**	**56.13**	**45.93**	**54.27**	**60.27**	**6.951**
27.00	64.46	56.30	46.20	51.09	60.03	6.514	RACAI	26.92	61.36	56.95	46.50	55.02	59.85	6.894

TED : MT Russian-English (MT$_{RuEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
18.31	**52.37**	**65.74**	**54.53**	**62.52**	**51.75**	**5.332**	NAIST	**16.97**	**48.67**	**68.59**	**57.06**	**66.57**	**49.22**	**5.385**
10.24	40.31	70.60	60.93	67.76	47.06	2.979	NICT	08.89	35.74	74.43	65.70	72.67	42.71	2.251

TED : MT Slovak-English (MT$_{SkEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
21.50	52.85	**62.26**	**54.34**	**59.38**	**54.11**	**5.545**	FBK	20.82	50.11	**64.41**	**56.29**	**62.48**	51.78	**5.686**
20.55	**53.91**	66.76	58.42	60.68	50.93	5.168	NAIST	**21.43**	**51.51**	65.89	56.89	63.85	**52.12**	5.685
16.24	53.63	68.31	61.41	59.84	47.42	4.691	RWTH	19.71	50.08	65.77	57.65	63.97	51.29	5.593

TED : MT Turkish-English (MT$_{TrEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
17.16	**53.51**	**74.32**	**52.32**	**66.65**	**54.61**	**5.551**	FBK	**16.06**	**50.37**	**77.81**	**54.53**	**70.86**	**52.43**	**5.691**
14.87	50.47	77.47	55.41	69.79	51.86	5.148	NAIST	13.66	47.16	81.37	57.78	74.37	49.44	5.256
12.86	47.36	80.04	58.58	72.78	48.90	4.745	TUBITAK	11.96	43.79	83.23	60.69	76.89	46.45	4.876

TED : MT Chinese-English (MT$_{ZhEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
15.08	**49.76**	**69.52**	**56.64**	**65.05**	**49.73**	**4.931**	RWTH	**13.95**	**45.97**	**73.08**	**59.58**	**69.84**	**47.18**	**4.904**
12.04	45.62	71.78	59.10	67.82	46.76	4.364	NAIST	10.91	41.47	75.59	62.49	72.91	43.74	4.222

OLYMPICS : MT Chinese-English (MT$_{ZhEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
19.17	**53.79**	**66.88**	**56.34**	**61.36**	**51.51**	**4.777**	HIT	**18.85**	**48.90**	**72.21**	**59.26**	**68.85**	**49.85**	**5.197**
16.95	50.21	69.82	59.18	65.42	49.79	4.531	NICT	16.37	45.55	75.85	63.28	72.65	46.55	4.749
12.79	46.34	75.46	63.92	71.10	45.94	3.994	KYOTO-U	12.38	41.44	82.83	68.54	79.74	43.06	4.177
12.16	38.90	84.14	71.98	79.68	43.67	3.631	POSTECH	10.89	32.38	92.71	78.64	89.66	39.22	3.650

A.2. Progress Testset (*tst2011*)

· All the sentence IDs in the IWSLT 2011 testset were used to calculate the automatic scores for each run submission.
· ASR and MT systems are ordered according to the *WER* and *BLEU* metrics, respectively.
· For each task, the best score of each metric is marked with **boldface**.
· Besides the NIST metrics, all automatic evaluation metric scores are given as percent figures (%).
· Besides the ASR scores, the mean scores of 2000 iterations were calculated for each MT output according to the *bootStrap* method [48].
· Omitted lines between scores indicate non-significant differences in performance between the MT engines.

TED : ASR English (ASR$_{En}$)

System	WER	(Count)
NICT	**10.9**	(1401)
MITLL	11.1	(1432)
KIT	12.0	(1552)
KIT-NAIST	12.0	(1553)
UEDIN	12.4	(1599)
RWTH	13.4	(1731)
FBK	15.4	(1991)

IWSLT 2011	WER	(Count)
MITLL	13.5	(1741)
KIT	15.0	(1938)
LIUM	15.4	(1992)
FBK	16.2	(2091)
NICT	25.6	(3301)

TED : SLT English-French (SLT$_{EnFr}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
28.85	**58.25**	**54.63**	**46.32**	**52.07**	**58.96**	**6.360**	KIT	**29.60**	**56.87**	**55.10**	**47.10**	**53.67**	**58.22**	**6.619**
27.83	56.37	55.87	47.43	53.38	58.15	6.298	RWTH	28.62	55.24	56.15	48.17	54.74	57.35	6.524
26.53	56.19	56.57	48.00	54.06	57.27	6.130	UEDIN	27.65	55.07	56.76	48.55	55.36	56.54	6.377
24.28	54.75	61.40	51.49	58.75	55.59	5.711	MITLL	24.86	53.71	62.31	52.55	60.69	54.69	5.873

TED : MT English-French (MT$_{EnFr}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
39.00	**67.73**	43.79	36.97	41.56	67.48	7.483	UEDIN	**37.86**	**64.64**	46.19	39.20	44.90	65.43	7.583
38.64	67.11	**42.98**	**36.75**	**40.88**	**67.69**	**7.607**	RWTH	37.37	63.90	**45.47**	39.11	44.38	65.59	7.681
38.49	67.12	43.08	36.86	41.00	67.59	7.587	KIT	37.35	64.09	45.53	**39.10**	**44.27**	**65.67**	**7.691**
37.90	66.62	43.90	37.58	41.79	66.88	7.442	NAIST	36.63	63.53	46.87	39.93	45.59	64.80	7.514
37.43	66.10	44.78	37.94	42.80	66.53	7.375	FBK	35.86	62.89	47.88	40.62	46.54	64.15	7.419
36.87	66.08	44.13	37.48	42.04	66.87	7.437	LIG	35.66	62.78	47.09	39.98	45.79	64.60	7.492
31.43	62.92	52.07	43.45	49.67	62.60	6.535	MITLL	30.09	59.49	55.78	46.42	54.19	60.14	6.568

TED : MT Arabic-English (MT$_{ArEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
27.29	**62.11**	**56.96**	**47.23**	**54.08**	**59.40**	**6.409**	RWTH	**26.25**	**59.76**	**59.00**	**48.76**	**57.33**	**58.16**	**6.519**
25.47	59.61	60.38	50.20	57.73	57.56	6.029	FBK	24.03	57.03	63.06	52.38	61.44	55.76	6.058
23.85	58.45	59.96	49.65	57.09	56.84	5.990	TUBITAK	22.43	55.68	62.96	52.14	61.10	54.78	6.006
23.66	58.52	61.79	51.39	58.85	56.46	5.826	NAIST	22.20	55.58	64.94	54.15	63.26	54.13	5.814
18.00	58.18	66.37	56.41	59.20	50.96	4.949	MITLL	21.38	55.14	65.05	53.25	63.04	54.44	5.830

TED : MT German-English (MT$_{DeEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
34.02	**70.46**	**48.05**	**37.99**	**44.50**	**67.03**	**7.426**	RWTH	**32.98**	**68.00**	**50.25**	**39.68**	**47.70**	**65.53**	**7.587**
32.42	70.32	49.91	38.28	45.77	66.99	7.311	UEDIN	31.68	67.94	52.17	39.99	48.94	65.42	7.450
32.38	69.87	50.30	39.06	46.56	66.68	7.243	FBK	31.77	67.56	52.28	40.53	49.32	65.14	7.421
31.53	69.21	50.87	39.34	46.83	66.10	7.193	NAIST	30.82	66.69	53.00	41.06	49.94	64.43	7.355

TED : MT Dutch-English (MT$_{NlEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
36.11	**71.40**	**47.94**	**37.51**	**43.91**	**67.81**	**7.623**	FBK	**35.30**	**69.30**	**49.70**	**38.56**	**47.06**	**66.95**	**7.842**
34.63	70.48	49.20	38.55	44.99	66.64	7.436	NAIST	33.82	68.21	51.24	39.72	48.49	65.77	7.632

TED : MT Polish-English (MT$_{PlEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
20.27	**55.81**	**66.07**	**53.92**	**62.49**	**54.13**	**5.484**	NAIST	**19.27**	**52.31**	**68.92**	**55.94**	**66.54**	**51.97**	**5.587**
18.65	53.61	68.11	55.42	64.19	53.10	5.279	PJIIT	18.00	50.30	69.91	56.86	67.45	51.12	5.469

TED : MT Portuguese(Brazilian)-English (MT$_{PtbEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
39.72	**75.06**	**41.67**	**34.11**	**39.45**	**71.04**	**7.990**	NAIST	**37.96**	**72.58**	**44.60**	**36.40**	**42.97**	**69.05**	**8.007**

TED : MT Romanian-English (MT$_{RoEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
33.62	**69.57**	**47.48**	**38.53**	**44.79**	**66.71**	**7.402**	NAIST	**31.84**	**66.62**	**50.62**	**40.92**	**48.79**	**64.58**	**7.447**
29.93	68.44	52.13	42.06	46.71	63.45	6.881	RACAI	30.10	65.57	52.58	42.05	50.53	63.61	7.266

TED : MT Russian-English (MT$_{RuEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
20.17	**55.09**	**64.35**	**52.91**	**61.14**	**53.76**	**5.436**	NAIST	**18.54**	**51.36**	**67.46**	**55.40**	**65.26**	**51.06**	**5.479**
11.52	42.37	68.93	58.62	66.03	49.02	3.473	NICT	09.97	38.04	72.56	63.22	70.83	44.80	2.791

TED : MT Turkish-English (MT$_{TrEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
17.23	**52.85**	**75.46**	**53.62**	**67.71**	**54.39**	**5.411**	FBK	**16.02**	**49.73**	**78.79**	**55.64**	**71.92**	**52.32**	**5.522**
15.04	50.02	79.38	57.42	71.74	51.55	4.965	NAIST	13.95	46.86	83.08	59.39	76.18	49.34	5.060
13.30	47.66	81.47	58.86	73.70	49.64	4.709	TUBITAK	12.34	44.19	84.41	60.48	77.63	47.59	4.847

TED : MT Chinese-English (MT$_{ZhEn}$)

"case+punc" evaluation							**System**	*"no_case+no_punc" evaluation*						
BLEU	METEOR	WER	PER	TER	GTM	NIST		BLEU	METEOR	WER	PER	TER	GTM	NIST
17.20	**52.21**	**67.25**	**54.70**	**62.86**	**51.92**	**5.189**	RWTH	**15.67**	**48.36**	**70.65**	**57.43**	**67.48**	**49.41**	**5.128**
13.74	48.01	69.51	57.22	65.77	49.17	4.628	NAIST	12.12	43.84	73.27	60.58	70.71	45.95	4.463

Appendix B. Human Evaluation

B.1. TED SLT English-French Task - Progress Testset (*tst2011*)

System Ranking

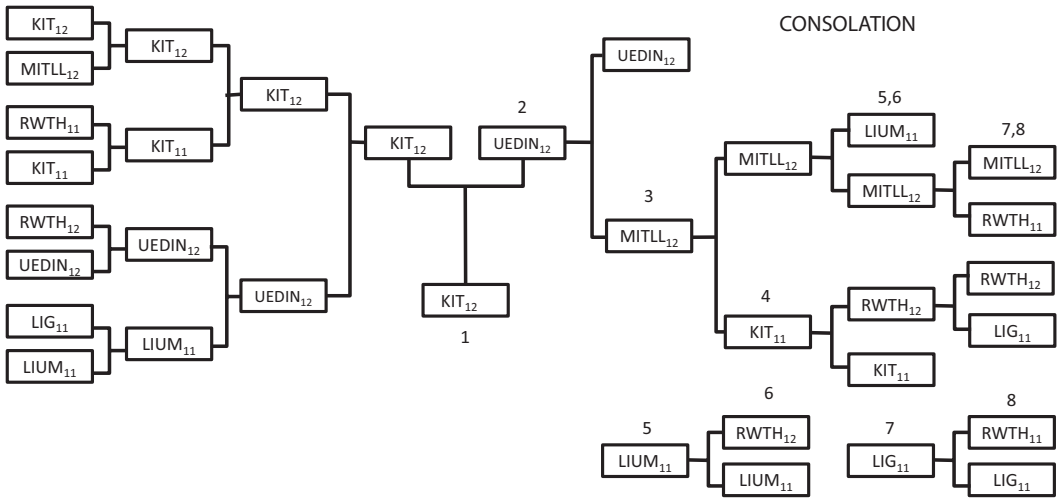
BLEU Ranking
(used for tournament seeding)

Ranking	System	BLEU score
1	KIT$_{12}$	28.86
2	LIUM$_{11}$	28.23
3	RWTH$_{12}$	27.85
4	KIT$_{11}$	26.78
5	RWTH$_{11}$	26.76
6	UEDIN$_{12}$	26.54
7	LIG$_{11}$	24.85
8	MITLL$_{12}$	24.27

Human Ranking
(resulting from tournament)

Ranking	System
1	KIT$_{12}$
2	UEDIN$_{12}$
3	MITLL$_{12}$
4	KIT$_{11}$
5	LIUM$_{11}$
6	RWTH$_{12}$
7	LIG$_{11}$
8	RWTH$_{11}$

Double Seeded Knockout with Consolation Tournament



Head to Head Matches Evaluation

· Head to Head matches: Wins indicate the percentage of times that one system was judged to be better than the other. The winner of the two systems is indicated in bold. The difference between 100 and the sum of the systems'wins corresponds to the percentage of ties.
· Statistical significance: † indicates statistical significance at $p \leq 0.10$, ‡ indicates statistical significance at $p \leq 0.05$, and ⋆ indicates statistical significance at $p \leq 0.01$, according to the Approximate Randomization Test based on 10,000 iterations.
· Inter Annotator Agreement: calculated using *Fleiss' kappa coefficient.*

HtH Matches	% Wins	I.A.A.
KIT$_{11}$- KIT$_{12}$	KIT$_{11}$: 23.75 KIT$_{12}$: **41.75***	0.1916
KIT$_{11}$- MITLL$_{12}$	KIT$_{11}$:28.50 MITLL$_{12}$: **33.50**	0.1716
LIG$_{11}$- RWTH$_{12}$	LIG$_{11}$: 31.25 RWTH$_{12}$: **31.75**	0.1993
LIUM$_{11}$- UEDIN$_{12}$	LIUM$_{11}$: 38.00 UEDIN$_{12}$: **38.00** *(a)*	0.1887
RWTH$_{11}$- MITLL$_{12}$	RWTH$_{11}$: 28.25 MITLL$_{12}$: **30.50**	0.1415

HtH Matches	% Wins	I.A.A.
MITLL$_{12}$- LIUM$_{11}$	MITLL$_{12}$: **39.75** LIUM$_{11}$: 37.50	0.2025
MITLL$_{12}$- KIT$_{12}$	MITLL$_{12}$: 18.00 KIT$_{12}$: **25.50**‡	0.3730
RWTH$_{12}$- KIT$_{11}$	RWTH$_{12}$: 37.50 KIT$_{11}$: **38.00**	0.2413
RWTH$_{12}$- LIUM$_{11}$	RWTH$_{12}$: 27.00 LIUM$_{11}$: **36.00**‡	0.2245
UEDIN$_{12}$- KIT$_{12}$	UEDIN$_{12}$: 37.25 KIT$_{12}$: **41.75**	0.2760

HtH Matches	% Wins	I.A.A.
UEDIN$_{12}$- MITLL$_{12}$	UEDIN$_{12}$: **40.75** MITLL$_{12}$: 34.50	0.2618
UEDIN$_{12}$- RWTH$_{12}$	UEDIN$_{12}$: **19.25** RWTH$_{12}$: 16.00	0.4009
RWTH$_{11}$- KIT$_{11}$	RWTH$_{11}$: 24.00 KIT$_{11}$: **27.75**	0.1784
LIG$_{11}$- LIUM$_{11}$	LIG$_{11}$: 21.55 LIUM$_{11}$: **30.08**‡	0.1743
RWTH$_{11}$- LIG$_{11}$	RWTH$_{11}$: 26.88 LIG$_{11}$: 29.65	0.1697

(a) Total number of wins considering all the judgments by the three annotators: UEDIN$_{12}$= 475; LIUM$_{11}$= 461.

B.2. TED MT Arabic-English Task - Progress Testset (*tst2011*)

System Ranking

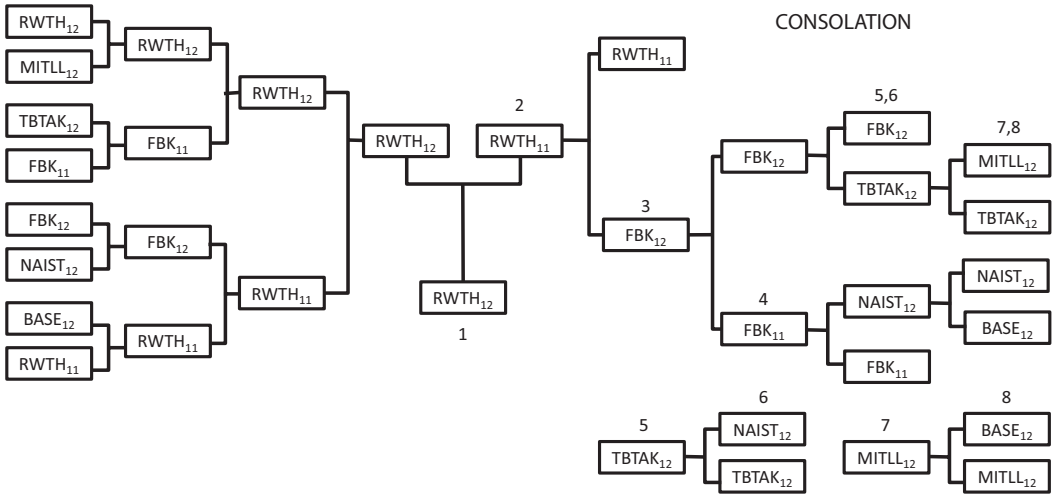
BLEU Ranking
(used for tournament seeding)

Ranking	System	BLEU score
1	RWTH$_{12}$	27.28
2	RWTH$_{11}$	26.32
3	FBK$_{12}$	25.46
4	FBK$_{11}$	24.31
5	TUBITAK$_{12}$	23.85
6	NAIST$_{12}$	23.65
7	BASELINE$_{12}$	22.08
8	MITLL$_{12}$	17.99

Human Ranking
(resulting from tournament)

Ranking	System
1	RWTH$_{12}$
2	RWTH$_{11}$
3	FBK$_{12}$
4	FBK$_{11}$
5	TUBITAK$_{12}$
6	NAIST$_{12}$
7	MITLL$_{12}$
8	BASELINE$_{12}$

Double Seeded Knockout with Consolation Tournament



Head to Head Matches Evaluation

· Head to Head matches: Wins indicate the percentage of times that one system was judged to be better than the other. The winner of the two systems is indicated in bold. The difference between 100 and the sum of the systems'wins corresponds to the percentage of ties.
· Statistical significance: † indicates statistical significance at $p \leq 0.10$, ‡ indicates statistical significance at $p \leq 0.05$, and ⋆ indicates statistical significance at $p \leq 0.01$, according to the Approximate Randomization Test based on 10,000 iterations.
· Inter Annotator Agreement: calculated using *Fleiss' kappa coefficient.*

HtH Matches	% Wins	I.A.A.
FBK$_{11}$- FBK$_{12}$	FBK$_{11}$: 23.75 FBK$_{12}$: **24.75**	0.2766
MITLL$_{12}$- RWTH$_{12}$	MITLL$_{12}$: 12.50 RWTH$_{12}$: **59.00***	0.2834
FBK$_{11}$- NAIST$_{12}$	FBK$_{11}$: **37.50**$\star$ NAIST$_{12}$: 18.25	0.2693
FBK$_{11}$- RWTH$_{12}$	FBK$_{11}$: 21.50 RWTH$_{12}$: **40.75***	0.2417
FBK$_{11}$- TUBITAK$_{12}$	FBK$_{11}$: **41.00*** TUBITAK$_{12}$: 26.50	0.1971

HtH Matches	% Wins	I.A.A.
NAIST$_{12}$- FBK$_{12}$	NAIST$_{12}$: 20.50 FBK$_{12}$: **47.25***	0.2352
NAIST$_{12}$- TUBITAK$_{12}$	NAIST$_{12}$: 24.00 TUBITAK$_{12}$: **24.00** [(a)]	0.2545
TUBITAK$_{12}$- FBK$_{12}$	TUBITAK$_{12}$: 18.25 FBK$_{12}$: **37.25***	0.2937
TUBITAK$_{12}$- MITLL$_{12}$	TUBITAK$_{12}$: **39.75*** MITLL$_{12}$: 19.75	0.2030
RWTH$_{11}$- FBK$_{12}$	RWTH$_{11}$: **38.50*** FBK$_{12}$: 25.25	0.2297

HtH Matches	% Wins	I.A.A.
RWTH$_{11}$- RWTH$_{12}$	RWTH$_{11}$: 20.25 RWTH$_{12}$: **27.25**†	0.3236
RWTH$_{11}$- BASELINE$_{12}$	RWTH$_{11}$: **58.75*** BASELINE$_{12}$: 10.25	0.2654
BASELINE$_{12}$- MITLL$_{12}$	BASELINE$_{12}$: 16.50 MITLL$_{12}$**25.25***	0.1933
BASELINE$_{12}$- NAIST$_{12}$	BASELINE$_{12}$:17.75 NAIST$_{12}$: **37.25***	0.2284

(a) Total number of wins considering all the judgments by the three annotators: TUBITAK$_{12}$= 358; NAIST$_{12}$= 327.

B.3.1 TED MT English-French - Progress Testset (*tst2011*)

· First tournament: all 2012 systems to determine the top four ones.

System Ranking

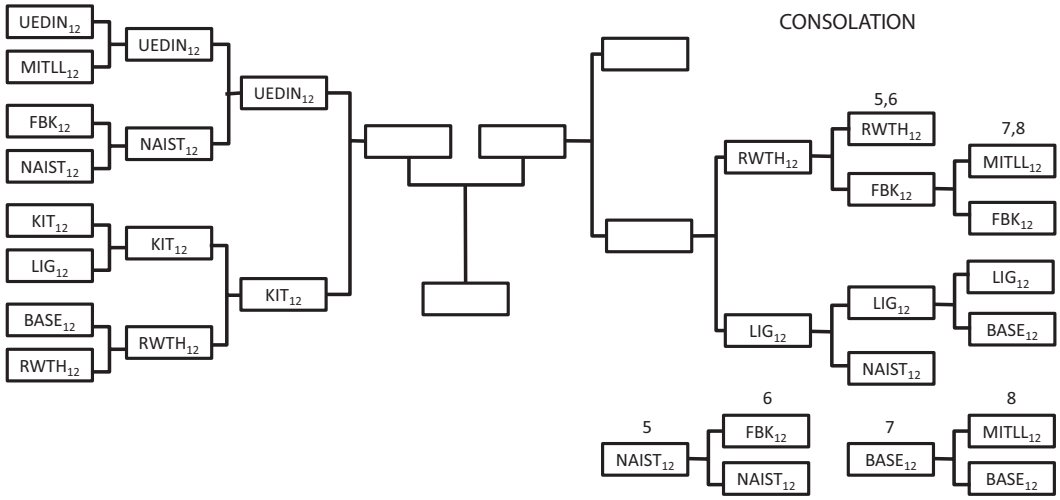
BLEU Ranking
(used for tournament seeding)

Ranking	System	BLEU score
1	UEDIN$_{12}$	39.01
2	RWTH$_{12}$	38.66
3	KIT$_{12}$	38.49
4	NAIST$_{12}$	37.90
5	FBK$_{12}$	37.43
6	LIG$_{12}$	36.88
7	BASELINE$_{12}$	33.90
8	MITLL$_{12}$	31.44

Human Ranking
(resulting from tournament)

Ranking	System
	KIT$_{12}$
	LIG$_{12}$
	RWTH$_{12}$
	UEDIN$_{12}$
5	NAIST$_{12}$
6	FBK$_{12}$
7	BASELINE$_{12}$
8	MITLL$_{12}$

Double Seeded Knockout with Consolation Tournament



Head to Head Matches Evaluation

· Head to Head matches: Wins indicate the percentage of times that one system was judged to be better than the other. The winner of the two systems is indicated in bold. The difference between 100 and the sum of the systems'wins corresponds to the percentage of ties.
· Statistical significance: † indicates statistical significance at $p \leq 0.10$, ‡ indicates statistical significance at $p \leq 0.05$, and ⋆ indicates statistical significance at $p \leq 0.01$, according to the Approximate Randomization Test based on 10,000 iterations.
· Inter Annotator Agreement: calculated using *Fleiss' kappa coefficient.*

HtH Matches	% Wins	I.A.A.
BASELINE$_{12}$- LIG$_{12}$	BASELINE$_{12}$: 24.75 LIG$_{12}$: **45.75***	0.1665
BASELINE$_{12}$- MITLL$_{12}$	BASELINE$_{12}$: **39.75** MITLL$_{12}$: 32.75	0.1963
FBK$_{12}$- MITLL$_{12}$	FBK$_{12}$: **43.50**‡ MITLL$_{12}$: 32.75	0.1508
FBK$_{12}$- RWTH$_{12}$	FBK$_{12}$: 27.25 RWTH$_{12}$: **36.75**‡	0.2500

HtH Matches	% Wins	I.A.A.
LIG$_{12}$- KIT$_{12}$	LIG$_{12}$: 26.00 KIT$_{12}$: **33.50**†	0.2921
MITLL$_{12}$- UEDIN$_{12}$	MITLL$_{12}$: 16.50 UEDIN$_{12}$: **47.50***	0.2367
NAIST$_{12}$- UEDIN$_{12}$	NAIST$_{12}$: 20.50 UEDIN$_{12}$: **33.00***	0.4014
NAIST$_{12}$- FBK$_{12}$	NAIST$_{12}$: **34.75**‡ FBK$_{12}$: 25.25	0.3085

HtH Matches	% Wins	I.A.A.
NAIST$_{12}$- LIG$_{12}$	NAIST$_{12}$: 32.00 LIG$_{12}$**34.50**	0.2622
RWTH$_{12}$- BASELINE$_{12}$	RWTH$_{12}$: **34.25*** BASELINE$_{12}$: 22.25	0.2298
RWTH$_{12}$- KIT$_{12}$	RWTH$_{12}$: 32.50 KIT$_{12}$: **33.50**	0.3218

B.3.2 TED MT English-French Progressive Task - Progress Testset (*tst2011*)

· Second tournament: the four top-ranked 2012 systems + the four top-ranked 2011 systems

System Ranking

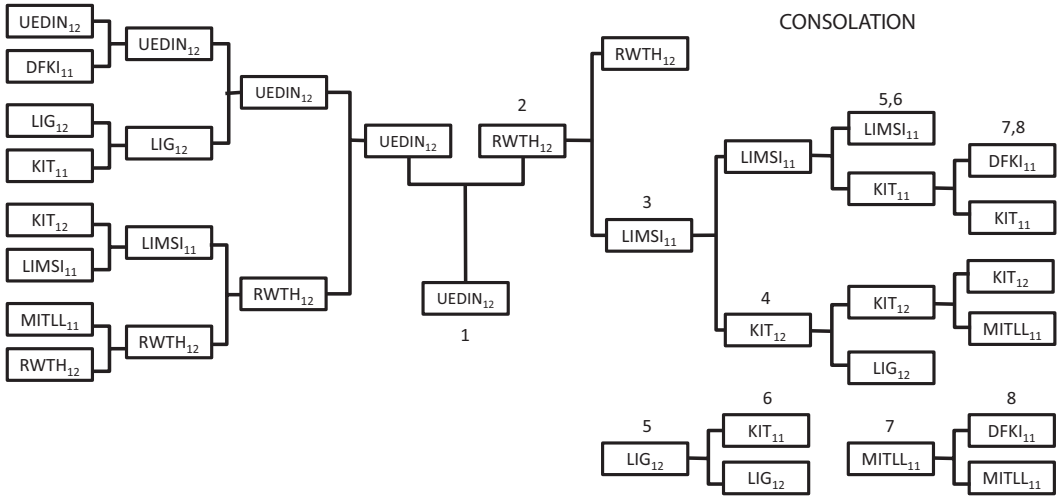
BLEU Ranking
(used for tournament seeding)

Ranking	System	BLEU score
1	UEDIN$_{12}$	39.01
2	RWTH$_{12}$	38.66
3	KIT$_{12}$	38.49
4	KIT$_{11}$	37.65
5	LIG$_{12}$	36.88
6	LIMSI$_{11}$	36.49
7	MITLL$_{11}$	35.28
8	DFKI$_{11}$	34.39

Human Ranking
(resulting from tournament)

Ranking	System
1	UEDIN$_{12}$
2	RWTH$_{12}$
3	LIMSI$_{11}$
4	KIT$_{12}$
5	LIG$_{12}$
6	KIT$_{11}$
7	MITLL$_{11}$
8	DFKI$_{11}$

Double Seeded Knockout with Consolation Tournament



Head to Head Matches Evaluation

· Head to Head matches: Wins indicate the percentage of times that one system was judged to be better than the other. The winner of the two systems is indicated in bold. The difference between 100 and the sum of the systems'wins corresponds to the percentage of ties.
· Statistical significance: † indicates statistical significance at $p \leq 0.10$, ‡ indicates statistical significance at $p \leq 0.05$, and ⋆ indicates statistical significance at $p \leq 0.01$, according to the Approximate Randomization Test based on 10,000 iterations.
· Inter Annotator Agreement: calculated using *Fleiss' kappa coefficient.*

HtH Matches	% Wins	I.A.A.
DFKI$_{11}$- UEDIN$_{12}$	DFKI$_{11}$: 22.75 UEDIN$_{12}$: **46.00***	0.2681
LIG$_{12}$- UEDIN$_{12}$	LIG$_{12}$: 23.00 UEDIN$_{12}$: **39.50***	0.2871
RWTH$_{12}$- LIMSI$_{11}$	RWTH$_{12}$: **35.25** LIMSI$_{11}$: 34.25	0.2625
RWTH$_{12}$- MITLL$_{11}$	RWTH$_{12}$: **39.25** MITLL$_{11}$: 33.75	0.2794

HtH Matches	% Wins	I.A.A.
KIT$_{11}$- LIG$_{12}$	KIT$_{11}$: 35.00 LIG$_{12}$: **37.75**	0.3218
LIMSI$_{11}$- KIT$_{12}$	LIMSI$_{11}$: **42.75** KIT$_{12}$: 38.50	0.2779
MITLL$_{11}$- KIT$_{12}$	MITLL$_{11}$: 28.75 KIT$_{12}$: **41.75***	0.2347
RWTH$_{12}$- UEDIN$_{12}$	RWTH$_{12}$: 23.50 UEDIN$_{12}$: **32.00**‡	0.3296

HtH Matches	% Wins	I.A.A.
DFKI$_{11}$- MITLL$_{11}$	DFKI$_{11}$: 40.00 MITLL$_{11}$: **42.50**	0.3777
KIT$_{11}$- LIMSI$_{11}$	KIT$_{11}$: 41.25 LIMSI$_{11}$: **43.50**	0.4154
DFKI$_{11}$- KIT$_{11}$	DFKI$_{11}$: 42.25 KIT$_{11}$: **43.00**	0.4235

B.4. OLYMPICS MT Chinese-English Task

System Ranking

· A subset of 400 test sentences was used to carry out the subjective ranking evaluation.
· The *"All systems"* scores indicate the average number of times that a system was judged better then ($>$) or better/equal to ($\geq$) any other system.
· The *"Head to head"* scores indicate the number of pairwise head-to-head comparisons won by a system.

System	ALL SYSTEMS $>$ others	ALL SYSTEMS $\geq$ others
HIT	**0.3808**	**0.8642**
NAIST-NICT	0.3025	0.8242
KYOTO-U	0.2150	0.7242
POSTECH	0.0850	0.6042

System	HEAD-TO-HEAD # wins
HIT	3 / 3
NAIST-NICT	2 / 3
KYOTO-U	1 / 3
POSTECH	0 / 3

Head to Head Matches Evaluation

· Head to Head matches: Wins indicate the percentage of times that one system was judged to be better than the other. The winner of the two systems is indicated in bold. The difference between 100 and the sum of the systems'wins corresponds to the percentage of ties.
· Statistical significance: † indicates statistical significance at $p \leq 0.10$, ‡ indicates statistical significance at $p \leq 0.05$, and $\star$ indicates statistical significance at $p \leq 0.01$, according to the Approximate Randomization Test based on 10,000 iterations.
· Inter Annotator Agreement: calculated using *Fleiss' kappa coefficient.*

HtH Matches	% Wins	I.A.A.
HIT- POSTECH	HIT: **47.75*** POSTECH: 6.25	0.3881
NAIST-NICT- KYOTO-U	NAIST-NICT: **32.50*** KYOTO-U: 17.25	0.3251
NAIST-NICT- HIT	NAIST-NICT: 17.75 HIT: **29.50***	0.3484

HtH Matches	% Wins	I.A.A.
KYOTO-U- HIT	KYOTO-U: 16.75 HIT: **37.00***	0.3819
KYOTO-U- POSTECH	KYOTO-U: **30.50*** POSTECH: 13.25	0.3722
NAIST-NICT- POSTECH	NAIST-NICT: **40.50*** POSTECH: 6.00	0.3616

Dialog Adequacy

(best = 5.0, . . ., worst = 1.0)

The following tables show how much of the information from the input sentence was expressed in the translation with (*adequacy*) and without (*dialog*) taking into account the context of the respective dialog.

OLYMPICS	**MT**	**Adequacy**	**Dialog**
MT_{ZhEn}	HIT	**3.17**	**3.42**
	NAIST-NICT	3.00	
	KYOTO-U	2.90	
	POSTECH	2.49	