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

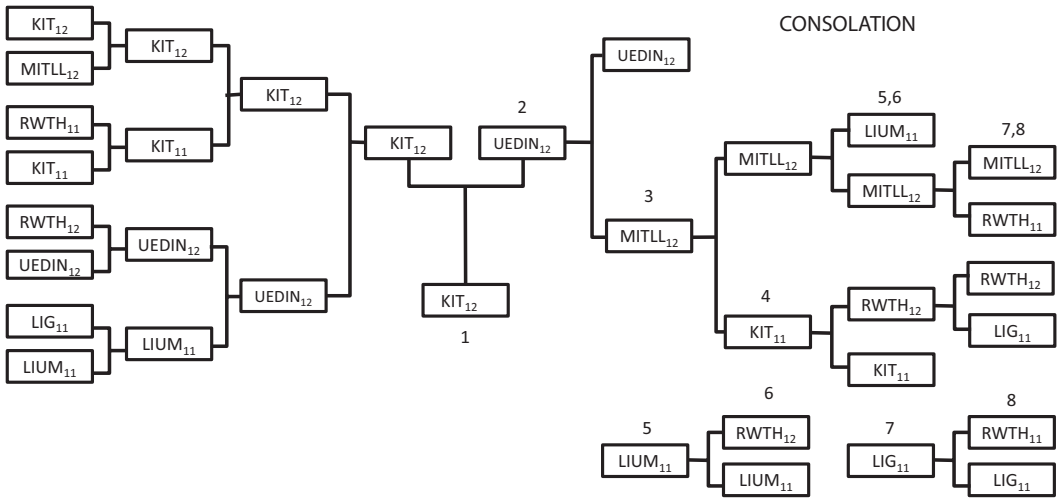
BLEU Ranking
(used for tournament seeding)

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

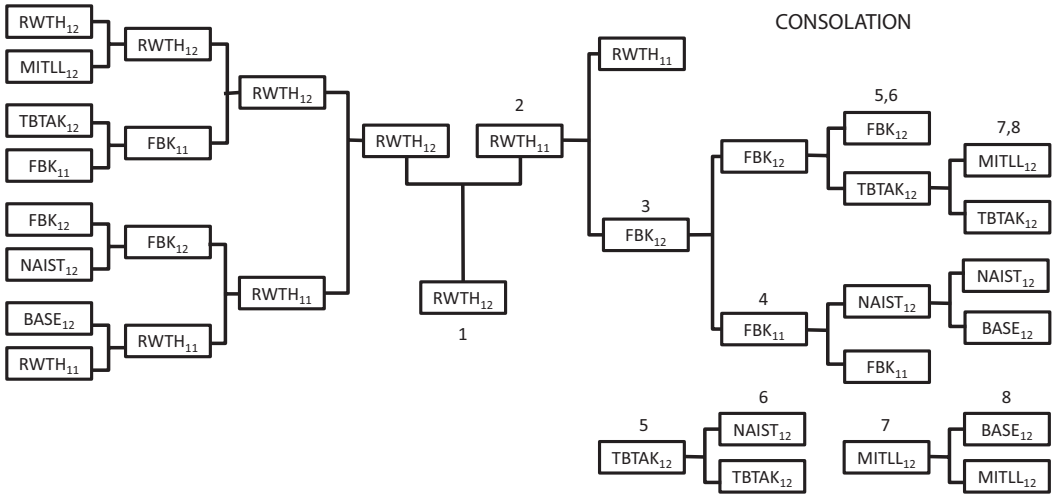
BLEU Ranking
(used for tournament seeding)

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

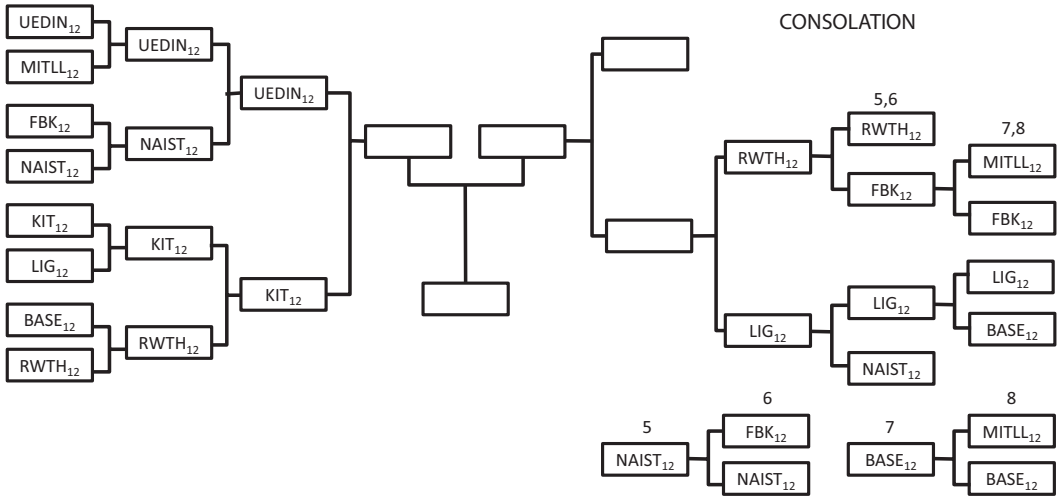
BLEU Ranking
(used for tournament seeding)

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

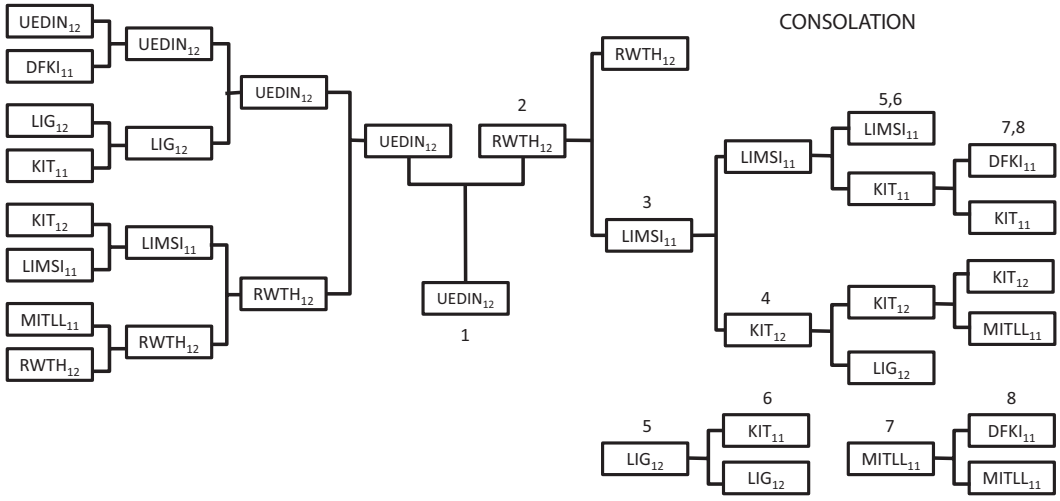
BLEU Ranking
(used for tournament seeding)

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	