

	de-en (21 systems)	fr-en (21 systems)	es-en (13 systems)	cz-en (5 systems)	hu-en (6 systems)	Average
ulc	<b>.78</b>	.92	.86	<b>1</b>	.6	<b>.83</b>
maxsim	.76	.91	<b>.98</b>	.7	.66	.8
rte (absolute)	.64	.91	.96	.6	<b>.83</b>	.79
meteor-rank	.64	<b>.93</b>	.96	.7	.54	.75
rte (pairwise)	.76	.59	.78	.8	<b>.83</b>	.75
terp	-.72	-.89	-.94	-.7	-.37	-.72
meteor-0.6	.56	<b>.93</b>	.87	.7	.54	.72
meteor-0.7	.55	<b>.93</b>	.86	.7	.26	.66
bleu-ter/2	.38	.88	.78	.9	-.03	.58
nist	.41	.87	.75	.9	-.14	.56
wpF	.42	.87	.82	<b>1</b>	-.31	.56
ter	-.43	-.83	-.84	-.6	-.01	-.54
nist (cased)	.42	.83	.75	<b>1</b>	-.31	.54
bleu	.41	.88	.79	.6	-.14	.51
bleusp	.39	.88	.78	.6	-.09	.51
bleusp4114	.39	.89	.78	.6	-.26	.48
bleu (cased)	.4	.86	.8	.6	-.31	.47
wpbleu	.43	.86	.8	.7	-.49	.46
wcd6p4er	-.41	-.89	-.76	-.6	.43	-.45

Table 7: The system-level correlation of the automatic evaluation metrics with the human judgments for translation into English.