

	fr-en (6268 pairs)	de-en (6382 pairs)	es-en (4106 pairs)	cz-en (2251 pairs)	hu-en (2193 pairs)	Overall (21200 pairs)
Oracle	.61	.63	.59	.61	.67	.62
rte (absolute)	.60	.61	<b>.59</b>	.57	<b>.65</b>	<b>.61</b>
ulc	<b>.61</b>	<b>.62</b>	.58	<b>.61</b>	.59	.60
maxsim	<b>.61</b>	<b>.62</b>	<b>.59</b>	.57	.61	.60
meteor-rank	<b>.61</b>	.61	<b>.59</b>	.57	.61	.60
meteor-0.6	<b>.61</b>	.61	.58	.57	.60	.60
rte (pairwise)	.56	.61	.57	.59	.64	.59
terp	.60	.61	<b>.59</b>	.57	.56	.59
meteor-0.7	<b>.61</b>	.61	.58	.57	.55	.59
ter	.60	.59	.57	.55	.51	.58
wpF	.60	.59	.57	<b>.61</b>	.46	.58
bleusp	<b>.61</b>	.59	.56	.55	.48	.57
bleusp4114	<b>.61</b>	.59	.56	.55	.46	.57
wcd6p4er	<b>.61</b>	.59	.57	.55	.44	.57
wpbleu	.60	.59	.57	.57	.43	.57

Table 12: Consistency of the automatic metrics when their system-level ranks are treated as sentence-level scores. Oracle shows the consistency of using the system-level human ranks that are given in Table 6.