Neural Machine Translation Quality and Post-Editing Performance

Vilém Zouhar, Aleš Tamchyna, Martin Popel, Ondřej Bojar

Model	P0→P1	P1→P2	P0→P2
Source	0.23	0.88	0.23
M01	0.65	0.94	0.63
M02	0.75	0.92	0.71
M03	0.72	0.90	0.69
M04	0.74	0.88	0.70
M05	0.74	0.94	0.73
M06	0.77	0.93	0.74
M07	0.80	0.93	0.78
M08	0.77	0.94	0.76
M09	0.77	0.93	0.76
M10	0.77	0.94	0.77
M11	0.80	0.95	0.80
M11*	-	-	0.92
Google	0.80	0.93	0.76
Microsoft	0.74	0.91	0.70
Reference	0.90	0.96	0.87
Reference*	-	-	0.87
Average	0.73	0.93	0.73
Lin. fit, all	0.011	0.001	0.015
Lin. fit, >36	0.004	0.000	0.027

Table 4: Average ChrF similarity per system between different stages of post-editing. Bottom two lines show linear fit coefficient on either all MT systems or on MT systems with BLEU > 36 (reference and source excluded). P0: system output, P1: post-editors' output, P2: reviewers' output.

Think & Total Time

$\hat{T} \approx T + \epsilon_T$	Measured think time
$\hat{A} \approx \hat{T} + \hat{W}$	Measured total time
$= T + W + \epsilon_T + \epsilon_W$	
$\overset{*}{W} := \hat{A} - \hat{T}$	Measured write time
$\approx W + \epsilon_W$	
$\overset{*}{T} := \min\{10s, \hat{T}\}$	Estimated think time
$\overset{*}{A} := \overset{*}{W} + \min\{10s, \hat{T}\}$	Estimated total time

8 documents translated from English to Czech by 13 MTs
Reference translation and source added →15 versions

P1: Post-edited by 15 professional post-editors
P2: Reviewed by 17 professionals

How strong is the relationship between MT quality, post-editing speed and post-edited quality?

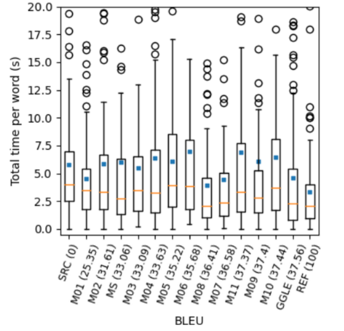
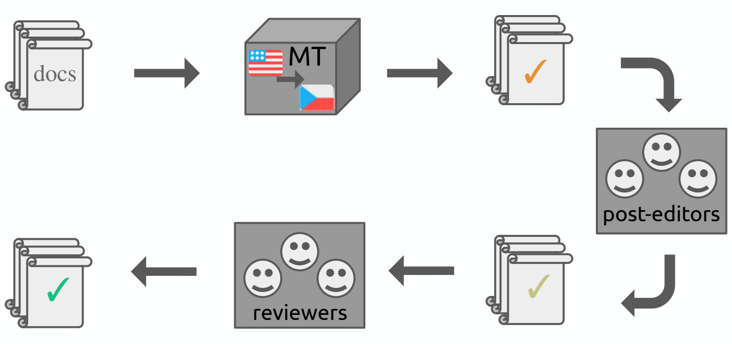


Figure 2: Total time per word in relation to MT system BLEU score. Every dot is a single post-edited sentence. Zoomed to [0, 20] on the y-axis. Orange bars represent medians and blue squares means. Upper whiskers are the 3rd quartile + 1.5× inter-quartile range.

Model	TER	BLEU	Steps [k]	ACh
M01	0.729	25.35	25.4	8
M02	0.678	31.61	29.0	8
M03	0.655	33.09	29.3	8
M04	0.648	33.63	33.0	8
M05	0.622	35.22	72.8	6
M06	0.624	35.68	997.1	0
M07	0.604	36.58	1015.2	5
M08	0.600	36.41	1022.4	6
M09	0.603	37.40	1055.0	8
M10	0.600	37.44	1058.6	6
M11	0.601	37.37	698.5	5
Google	0.623	37.56	–	–
Microsoft	0.632	33.06	–	–

Table 1: Overview of MT systems used. TER and BLEU were measured by SacreBLEU[7] (Post, 2018). Steps mark the number of training steps in thousands. ACh is the number of authentic-data-trained checkpoints in an average of 8 checkpoints.

Commercial Translation Pipeline



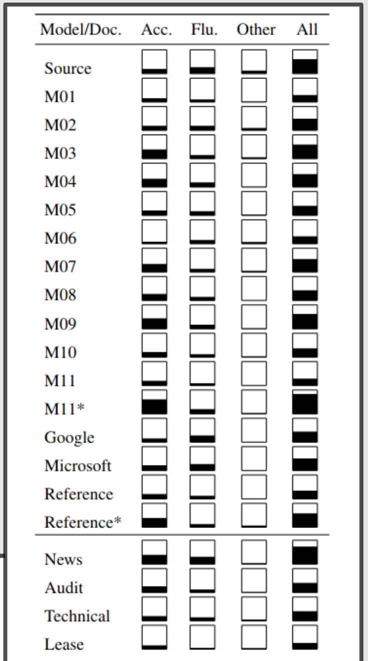


Table 7: Average LQA severity (reported from 0 to 3) of models and documents across three categories: Adequacy/accuracy, fluency and other. Their average is reported in the last column. Empty and full squares represent severities of 0 and 1, respectively.

Only top 8 systems: +1 BLEU → -0.51s / word

- *Trend not confirmed on larger sets of NMT systems*
- *Relationship weaker than for PBMT*
- *Do not expect small improvements in MT to lead to much {lower post-editing times, higher post-edited quality}*

Translating from scratch not that slower than post-editing

- *6.00s/word (src) | 5.66s/word (avg.) | 3.17s/word (ref)*

Diminishing results of additional phases

- *Much more edits in the first phase*
- *No noticeable relationship between MT quality and the second phase*

CHARLES UNIVERSITY
MEMSOURCE
We thank the company České překlady, NEUREM3 and LUSyD grants and the LINDAT/CLARIAHCZ project.