Neural Machine Translation Quality and Post-Editing Performance

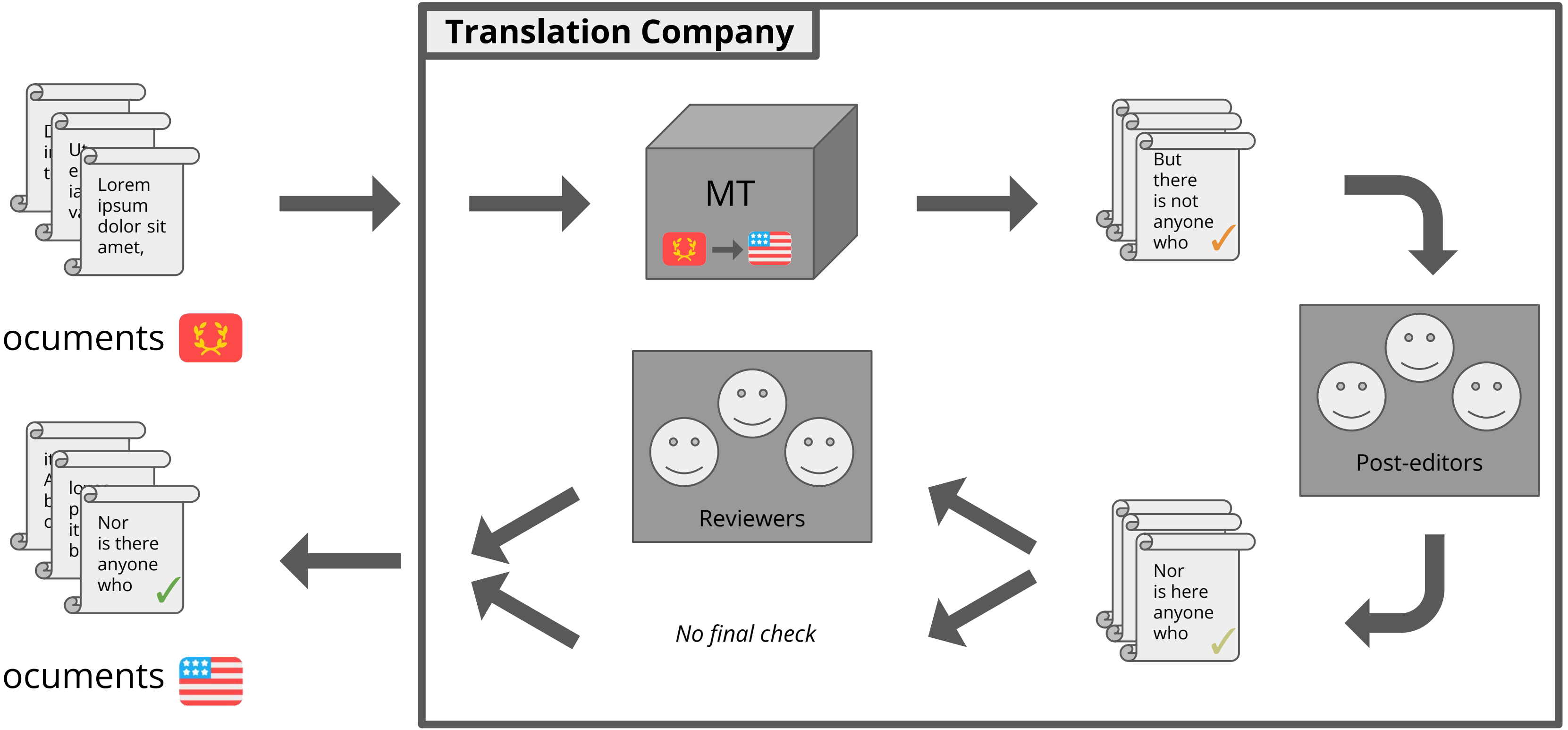
Vilém Zouhar, Aleš Tamchyna, Martin Popel, Ondřej Bojar

EMNLP 2021

CHARLES UNIVERSITY

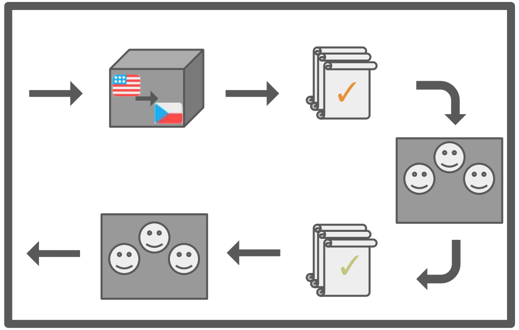
MEMSOURCE

Commercial Translation Workflow

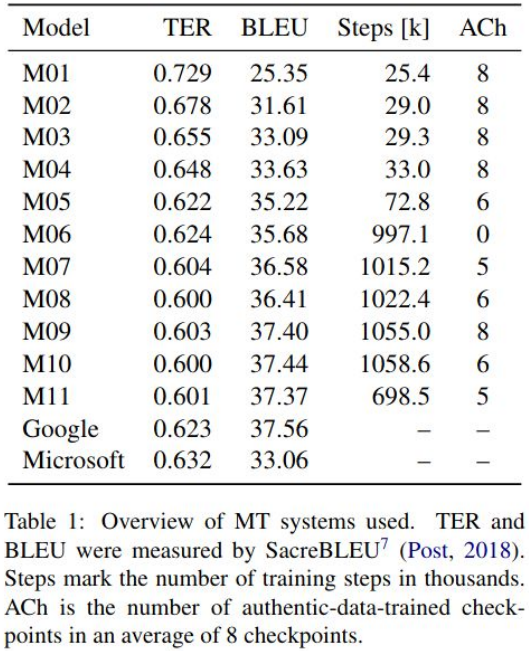


Research Overview

1. Translate 8 documents by 13 MTs (+SRC & REF)
2. Post-editing phase by 15 professionals
3. Revision phase by 15 (different) professionals



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



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.

Prior Work (Sanchez-Torron and Koehn, 2016)

	S-T & K 2016	**Ours**
MT	Phrase-based	Neural
Range	25-30 BLEU	25-38 BLEU
Count	9	13 (+2)
Distribution	Uniform	Non-uniform
Phases	One	Two

Conclusion:

...for each 1-point increase in BLEU, there is a PE time decrease of 0.16 seconds per word...

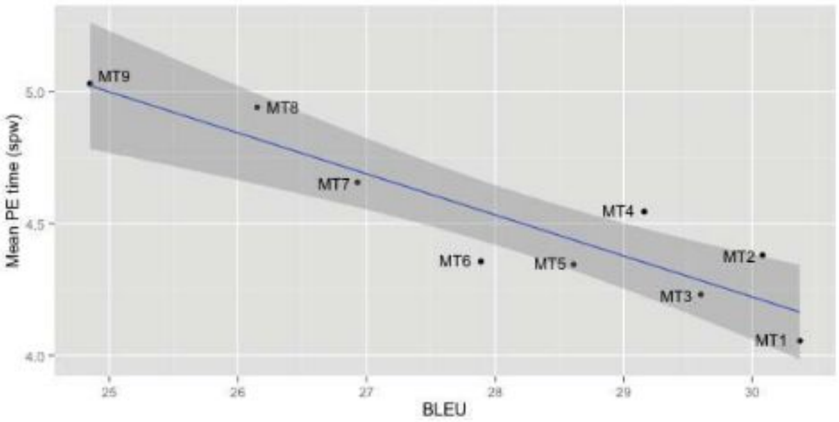


Figure 1: Scatter plot of systems' mean PE time against systems' BLEU and regression line with 95% confidence bounds

Edits in Phases

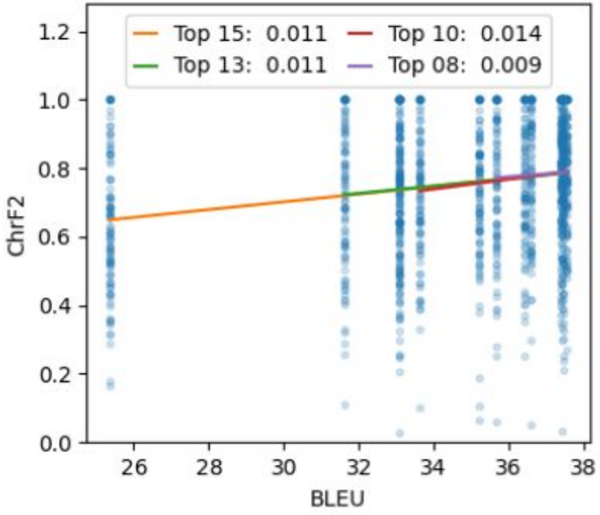
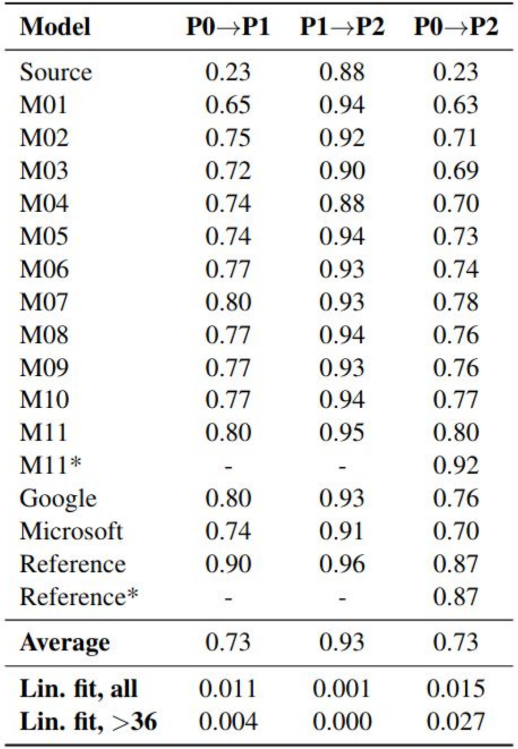


Figure 1: Sentence similarity measured by ChrF2 between the provided translation and first-phase (P0→P1). Every dot is a single sentence translated by a given MT. Source and Reference measurements are omitted for scale.

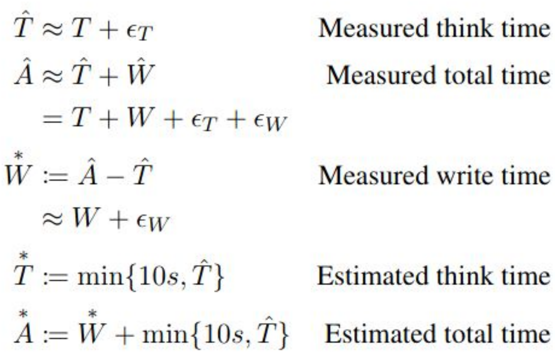


Model	P0→P1	P1→P2	P0→P2
Source	0.23	0.88	0.23
M01	0.65	0.94	0.63
M02	0.75	0.92	0.71
M03	0.72	0.90	0.69
M04	0.74	0.88	0.70
M05	0.74	0.94	0.73
M06	0.77	0.93	0.74
M07	0.80	0.93	0.78
M08	0.77	0.94	0.76
M09	0.77	0.93	0.76
M10	0.77	0.94	0.77
M11	0.80	0.95	0.80
M11*	-	-	0.92
Google	0.80	0.93	0.76
Microsoft	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.

Post-Editing Time

- Noisy data
  - Need for capping through heuristics
- Large differences between systems



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

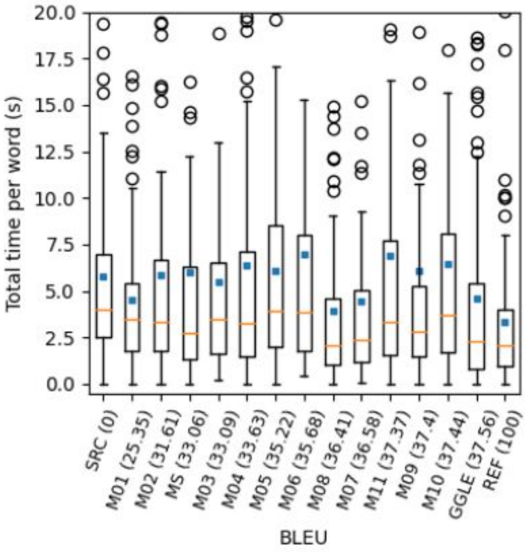
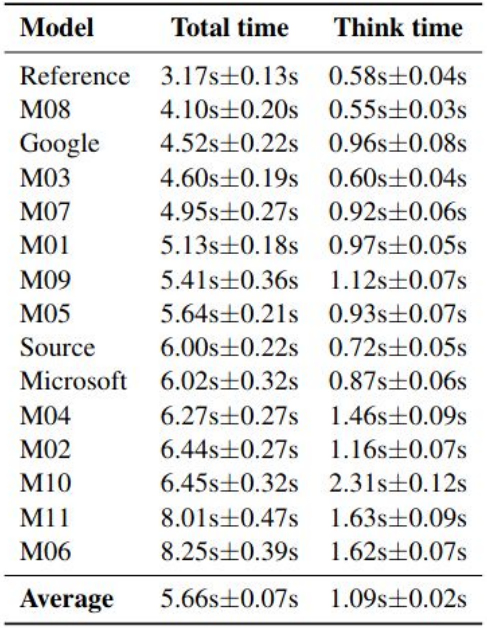


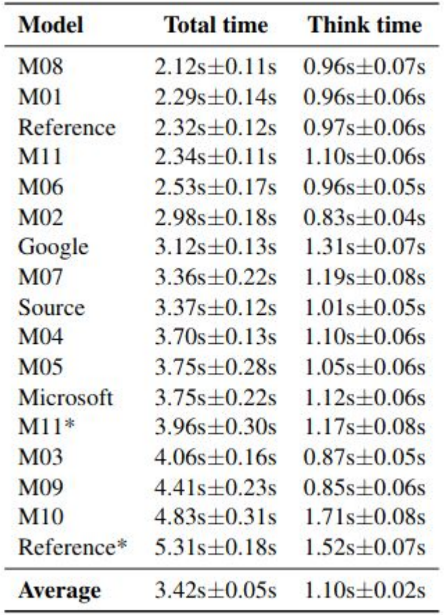
Figure 2: Total time per word in relation to MT system BLEU score. Every dot is a single post-edited sentence. Zoomed to [0, 20] on the y-axis. Orange bars represent medians and blue squares means. Upper whiskers are the 3rd quartile + 1.5× inter-quartile range.



Model	Total time	Think time
Reference	3.17s±0.13s	0.58s±0.04s
M08	4.10s±0.20s	0.55s±0.03s
Google	4.52s±0.22s	0.96s±0.08s
M03	4.60s±0.19s	0.60s±0.04s
M07	4.95s±0.27s	0.92s±0.06s
M01	5.13s±0.18s	0.97s±0.05s
M09	5.41s±0.36s	1.12s±0.07s
M05	5.64s±0.21s	0.93s±0.07s
Source	6.00s±0.22s	0.72s±0.05s
Microsoft	6.02s±0.32s	0.87s±0.06s
M04	6.27s±0.27s	1.46s±0.09s
M02	6.44s±0.27s	1.16s±0.07s
M10	6.45s±0.32s	2.31s±0.12s
M11	8.01s±0.47s	1.63s±0.09s
M06	8.25s±0.39s	1.62s±0.07s
Average	5.66s±0.07s	1.09s±0.02s

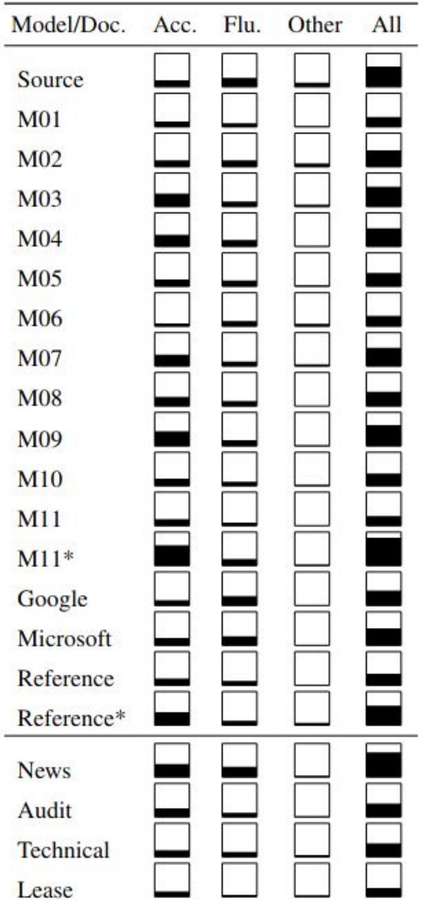
Table 5: Total and think time estimations for first phase of post-editing for all MT systems (+Source and Reference). Confidence intervals computed for 95%. Sorted by total time.

Second Phase & Errors



Model	Total time	Think time
M08	2.12s±0.11s	0.96s±0.07s
M01	2.29s±0.14s	0.96s±0.06s
Reference	2.32s±0.12s	0.97s±0.06s
M11	2.34s±0.11s	1.10s±0.06s
M06	2.53s±0.17s	0.96s±0.05s
M02	2.98s±0.18s	0.83s±0.04s
Google	3.12s±0.13s	1.31s±0.07s
M07	3.36s±0.22s	1.19s±0.08s
Source	3.37s±0.12s	1.01s±0.05s
M04	3.70s±0.13s	1.10s±0.06s
M05	3.75s±0.28s	1.05s±0.06s
Microsoft	3.75s±0.22s	1.12s±0.06s
M11*	3.96s±0.30s	1.17s±0.08s
M03	4.06s±0.16s	0.87s±0.05s
M09	4.41s±0.23s	0.85s±0.06s
M10	4.83s±0.31s	1.71s±0.08s
Reference*	5.31s±0.18s	1.52s±0.07s
Average	3.42s±0.05s	1.10s±0.02s

Table 6: Total and think time estimations for the review phase of post-editing for all MT systems (+Source and Reference). Confidence intervals computed for 95%. Sorted by total time.

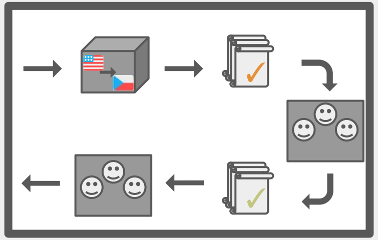


Model/Doc.	Acc.	Flu.	Other	All
Source	▂	▂	▁	▆
M01	▁	▁	□	▂
M02	▁	▁	▁	▄
M03	▃	▁	□	▅
M04	▃	▁	□	▅
M05	▁	▁	□	▃
M06	□	▁	□	▂
M07	▃	▁	□	▅
M08	▂	▁	□	▄
M09	▄	▁	□	▆
M10	▁	▁	□	▃
M11	▁	▁	□	▂
M11*	▅	▁	□	▇
Google	▁	▂	□	▄
Microsoft	▁	▂	□	▄
Reference	▁	▁	□	▃
Reference*	▃	▁	▁	▅
News	▃	▂	□	▇
Audit	▂	▁	□	▃
Technical	▁	▁	▁	▃
Lease	▁	□	□	▂

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.

Takeaways from NMT-PE effects

github.com/ufal/nmt-pe-effects-2021



- 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 second-phase