A Further Statistical Analysis

Table 1 shows detailed results, including those of individual raters, for all four experimental conditions. Raters choose between three labels for each item: MT is better than HUMAN (a), HUMAN is better than MT (b), or tie (t). Table 3 lists inter-rater agreement. Besides percent agreement (same label), we calculate Cohen's kappa coefficient

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}, \quad (1)$$

where $P(A)$ is the proportion of times that two raters agree, and $P(E)$ the likelihood of agreement by chance. We calculate Cohen's kappa, and specifically $P(E)$, as in WMT (Bojar et al., 2016, Section 3.3), on the basis of all pairwise ratings across all raters.

In pairwise rankings of machine translation outputs, κ coefficients typically centre around 0.3 (Bojar et al., 2016). We observe lower inter-rater agreement in three out of four conditions, and attribute this to two reasons. Firstly, the quality of the machine translations produced by Hassan et al. (2018) is high, making it difficult to discriminate from professional translation particularly at the sentence level. Secondly, we do not provide guidelines detailing error severity and thus assume that raters have differing interpretations of what constitutes a "better" or "worse" translation. Confusion matrices in Table 4 indicate that raters handle ties very differently: in document-level adequacy, for example, rater E assigns no ties at all, while rater F rates 15 out of 50 items as ties (Table 4g). The assignment of ties is more uniform in documents assessed for fluency (Tables 1, 4a–4f), leading to higher κ in this condition (Table 3).

Despite low inter-annotator agreement, the quality control we apply shows that raters assess items carefully: they only miss 1 out of 40 and 5 out of 128 spam items in the document- and sentence-level conditions overall, respectively, a very low number compared to crowdsourced work (Kittur et al., 2008). All of these misses are ties (i. e., not marking spam items as "better", but rather equally bad as their counterpart), and 5 out of 9 raters (A, B1, B2, D, F) do not miss a single spam item.

A common procedure in situations where inter-rater agreement is low is to aggregate ratings of different annotators (Graham et al., 2017). As shown in Table 2, majority voting leads to clearer discrimination between MT and HUMAN in all conditions, except for sentence-level adequacy.

	Document			Sentence		
Rater	MT	Tie	Human	MT	Tie	Human
Fluency						
A	13	8	29	30	32	42
B1				36	4	64
B2	8	18	24			
C	12	14	24	40	14	50
D	11	17	22	32	30	42
total	44	57	99	66	36	106
Adequacy						
E	26	0	24	59	3	42
F	10	15	25	44	16	44
G	18	4	28	38	23	43
H	20	3	27	38	11	55
total	74	22	104	103	19	86

Table 1: Ratings by rater and condition. Greyed-out fields indicate that raters had access to full documents for which we elicited sentence-level judgements; these are not considered for total results.

	Document			Sentence		
Aggregation	MT	Tie	Human	MT	Tie	Human
Fluency						
Average	22	29	50	32	17	51
Majority	24	10	66	26	23	51
Adequacy						
Average	37	11	52	50	9	41
Majority	32	18	50	38	32	31

Table 2: Aggregation of ratings (%).

	Document	Sentence
Fluency		
Same label	55 %	45 %
Cohen's κ	0.32	0.13
Adequacy		
Same label	49 %	50 %
Cohen's κ	0.13	0.14

Table 3: Inter-rater agreement.

A \ B2	a	t	b
a	5	4	4
t	1	5	2
b	2	9	18

(a) fluency, document, N=50

A \ C	a	t	b
a	7	2	4
t	2	4	2
b	3	8	18

(b) fluency, document, N=50

A \ D	a	t	b
a	6	3	4
t	2	6	0
b	3	8	18

(c) fluency, document, N=50

B2 \ C	a	t	b
a	5	1	2
t	4	5	9
b	3	8	13

(d) fluency, document, N=50

B2 \ D	a	t	b
a	6	1	1
t	3	7	8
b	2	9	13

(e) fluency, document, N=50

C \ D	a	t	b
a	7	3	2
t	1	7	6
b	3	7	14

(f) fluency, document, N=50

E \ F	a	t	b
a	4	9	13
t	0	0	0
b	6	6	12

(g) adequacy, document, N=50

E \ G	a	t	b
a	9	4	13
t	0	0	0
b	9	0	15

(h) adequacy, document, N=50

E \ H	a	t	b
a	11	1	14
t	0	0	0
b	9	2	13

(i) adequacy, document, N=50

F \ G	a	t	b
a	7	1	2
t	7	1	7
b	4	2	19

(j) adequacy, document, N=50

F \ H	a	t	b
a	6	1	3
t	8	0	7
b	6	2	17

(k) adequacy, document, N=50

G \ H	a	t	b
a	11	2	5
t	1	1	2
b	8	0	20

(l) adequacy, document, N=50

A \ B1	a	t	b
a	16	1	13
t	10	1	21
b	10	2	30

(m) fluency, sentence, N=104

E \ F	a	t	b
a	31	6	22
t	2	0	1
b	11	10	21

(n) adequacy, sentence, N=104

Table 4: Confusion matrices: MT is better than HUMAN (a), HUMAN is better than MT (b), or tie (t). Participant IDs (A–H) are the same as in Table 1.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation (WMT16). In *Proceedings of WMT*, pages 131–198, Berlin, Germany.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone? *Natural Language Engineering*, 23(1):3–30.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *Computing Research Repository*, arXiv:1803.05567.

Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of CHI*, pages 453–456, Florence, Italy.