Appendix

Corpus	Emoji	Upper	Spelling
4SQ-test	0.17	0.14	3.3
MTNT-test	0.02	0.18	3.8

Table 13: Noise comparison between 4SQ-test and MTNT-test (Michel and Neubig, 2018). Emojis, all-uppercase words and spelling + grammar mistakes (according to MS Word) per 100 tokens.

Model	4SQ-valid	MTNT-test
Berard et al. (2019)		
WMT (Inline case)	–	39.1
+ MTNT domain adaptation	–	44.3
+ Ensemble	–	**45.7**
Our models (single)		
WMT (Cased)	24.4	39.0
UGC (Cased)	30.3	41.5
UGC (Inline case)	29.4	41.6
UGC $\oplus$ BT + FT	33.7	44.5
UGC $\oplus$ BT $\oplus$ PE + tags	**33.9**	**44.9**
Nat noise $\oplus$ BT + FT	33.8	44.6
Nat noise $\oplus$ BT $\oplus$ PE + tags	33.5	**44.9**

Table 14: Comparison of our models against the winner of the WMT 2019 Robustness Task on the MTNT test set (similar robustness challenges but different domain). We also give cased BLEU of our models on the 4SQ-valid set. Results on 4SQ-test are shown in the paper.

Large-Scale monolingual evaluation We conducted a larger scale monolingual evaluation using Amazon Mechanical Turk (AMT), as reported in Table 15. We evaluated the translations of 1800 test sentences. To filter poor quality work, which occurs frequently in our experience, we also created gold questions by selecting 40 additional sentences for which we built 3 fake translations each, whose ranking was intentionally unambiguous and easy. We created HITs (Human Intelligence Tasks) of 10 sentences each, of which 3 sentences were gold questions. Workers were also required to have at least 98% task approval rate on AMT and 1000 tasks approved. We aimed for 6 submissions per HIT from 6 different workers. Compared to the in-house evaluation, the inter-judge agreement was low (Kappa of 0.15).

Pairs	Win	Tie	Loss
Tags + noise $\gg$ Tags	1939	7414	1667
Tags + noise $\gg$ Base	2718	6108	2178
Tags + noise $\gg$ GT	3008	5801	2173
Tags $\gg$ Baseline	2657	6110	2225
Tags $\gg$ GT	2950	5794	2234
Baseline $\gg$ GT	2205	6918	1889

Table 15: Large-scale Human Evaluation on Amazon Mechanical Turk ("$\gg$" means $p \leq 0.01$). The 4 models *Baseline*, *GT*, *Tags* and *Tags + noise* correspond respectively to rows 2 (UGC with inline case), 3 (Google Translate), 6 (Combination of BT, PE and tags) and 8 (Same as 6 with natural noise) in Table 8.

Both human evaluations agree and are consistent with the BLEU evaluation, except for the impact of natural noise, where the AMT evaluation found a significant improvement.

Evaluation	# Tasks	# Ties	% Ties	Kappa
In-house	12	3588	57%	0.47
AMT	1321	65988	58%	0.15

Table 16: Size of the human evaluations. AMT: Amazon Mechanical Turk. The AMT kappa (inter-judge agreement) is very low, while the in-house kappa is moderate.

SRC	On s'y sent comme a la maison ! <s> Équipe de serveurs très sympa! <s> Goutez au burger LE Retour d'Hervé, il est a tomber :-)
REF	It feels like home!! <s> Team of waiters very nice! <s> Taste the burger LE Retour d'Hervé, it's to die for :-)
Type	Bar, Bistro
Location	Paris, FR
Rating	8.29
SRC	Je conseille le crumble fraise/rhubarbe CHAUD. <s> C'est délicieux !!
REF	I recommend the strawberry/rhubard crumble HOT. <s> It's delicious!!
Type	Bakery, Breakfast Spot
Location	Brussels, BE
Rating	8.88
SRC	Très bons burgers, cheesecake à tomber par terre.... <s> Sans oublier <NAME>, <NAME> et <NAME> en un mot CHAR-MANTS!
REF	Very good burgers, cheesecake to die for... <s> Not to mention <NAME>, <NAME> and <NAME>: in a word CHAR-MING!
Type	American Restaurant
Location	Paris, FR
Rating	-
SRC	Friterie sympathique collée au Grand Boulevards. <s> On retrouve les incontournables frites belges. <s> Elle sont DELICIEUSESEMENT grosses comme on aiment :) a tester. <s> Ouverture tardive le we.
REF	Friendly chip shop stuck to Grand Boulevards. <s> We find the essential Belgian fries. <s> They are DELICIOUSLY big as we like them :) to test. <s> Late opening on the weekend.
Type	Belgian Restaurant, Fast Food Restaurant
Location	Paris, FR
Rating	7.91
SRC	Que de bon souvenir , fillet de boeuf au patte. <s> Merci pour l accueille Mr <NAME>
REF	Great memories, beef fillet with pasta. <s> Thank you for being so welcoming Mr <NAME>
Type	Café, Pizza Place
Location	Libreville, GA
Rating	8.21
SRC	La carte est souvent enrichie. <s> La gérance est top.
REF	The menu is often supplemented. <s> The management is top notch.
Type	Sushi Restaurant
Location	Sid'Bou Said, TN
Rating	7.70

Table 17: Examples of challenging examples from 4SQ-PE. We show the full reviews with sentence delimiters (<s>) and metadata. The words that contain typos or that could cause trouble for a regular NMT model are shown in red.

SRC	Le meilleur resto de Belleville, DE LOIN!
REF	The best restaurant in Belleville, BY FAR!
Cased	Best restaurant in Belleville, DE LOIN!
Inline case	The best restaurant in Belleville, BY FAR!
SRC	ESCALOPE DE VEAU MONTAGNARDE à tomber, et à ne plus pouvoir se lever de sa chaise
REF	ESCALOPE DE VEAU MONTAGNARDE is an absolute knock out and you'll have difficulty recovering
Cased	Falling down and not being able to get up from his chair
Inline case	ESCALOPE OF MOUNTAIN CALF to fall, and not be able to rise from his chair

Table 18: Examples of sentences from 4SQ-test with capitalized words, where default (cased) MT gets the translation wrong, and inline case helps.

SRC	**Bcp** de choix, peut-être Trop :-)
REF	**Plenty** of choice, maybe too much of it :-)
Inline case	**Bcp** of choice, maybe Too much :-)
Natural noise	**A lot** of choices, maybe Too much :-)
SRC	Service **loooooonnnng**.
REF	**Looooooong** wait.
Inline case	Service **loooooonnnng**.
Natural noise	**Long** service.

Table 19: Examples of sentences from 4SQ-test with noisy spelling (in red bold), where natural noise helps.

SRC	**Carte** attractive et pas excessive.
REF	Nice **menu** and not over the top.
Inline case	Attractive and not excessive **card**.
BT + FT	Attractive **menu** and not excessive.
SRC	**Cuisine** pas originale, service passable, mais l'endroit est joli !
REF	Not very original **food**, acceptable service, but the place itself is beautiful!
Inline case	Not an original **kitchen**, fair service, but the place is nice!
BT + FT	**Food** not original, service passable, but the place is nice!

Table 20: Examples of sentences from 4SQ-test with polysemous words (in red bold), where domain adaptation helps (with 4SQ-PE fine-tuning and back-translation).

SRC	Les frittes boff mais leurs burger, une tuerie!	Typo and slang (“bof”)
REF	The fries are meh, but the burgers, to die for!	
MT	The fries are great but their burgers are to die for!	
SRC	Le merveilleux du Merveilleux c'est merveilleux...	“merveilleux” is a pastry, “Merveilleux” is a pastry shop (named entity).
REF	The merveilleux at Merveilleux is marvelous...	
MT	The wonderful of the Wonderful it's wonderful...	
SRC	La souris d'agneau est délicieuse !	Dish name (translated literally)
REF	The lamb shank is delicious!	
MT	The lamb mouse is delicious!	
SRC	La quantité 5 raviolis qui se battent pour 12.70 euros.	Idiomatic expression (“qui se battent en duel”)
REF	Poor quantity, 5 raviolis or so for 12.70 Euros.	
MT	The quantity 5 dumplings that fight for 12.70 euros.	
SRC	Après le palais du facteur nous voici à la halte qui est un restaurant correct.	Named entities (“Palais Idéal du Facteur Cheval” and “La Halte du Facteur”)
REF	After the Palais du Facteur we stopped at La Halte, which is a reasonable restaurant.	
MT	After the mailman's palace here we are at the rest stop which is a decent restaurant.	

Table 21: Examples of bad translations by our best model (Noise $\oplus$ BT $\oplus$ PE + tags). All examples are from 4SQ-test, except for the last one, which is from SemEval.