Understanding the Impact of UGC Specificities on Translation Quality

José Carlos Rosales Núñez[1,2,3] **Djamé Seddah**[3] **Guillaume Wisniewski**[4]
[1] CNRS, LISN [2] Université Paris-Saclay
[3] INRIA Paris [4] Université de Paris, LLF, CNRS
jose.rosales@limsi.fr djame.seddah@inria.fr
guillaume.wisniewski@univ-paris-diderot.fr

Abstract

This work takes a critical look at the evaluation of user-generated content automatic translation, the well-known specificities of which raise many challenges for MT. Our analyses show that measuring the average-case performance using a standard metric on a UGC test set falls far short of giving a reliable image of the UGC translation quality. That is why we introduce a new data set for the evaluation of UGC translation in which UGC specificities have been manually annotated using a fine-grained typology. Using this data set, we conduct several experiments to measure the impact of different kinds of UGC specificities on translation quality, more precisely than previously possible.

1 Introduction

This work takes a critical look at the evaluation of user-generated content (UGC) automatic translation. The well-known specificities of UGC (high rate of OOVs, rare, grammatical constructs, ...) raise many challenges for Machine Translation and has been the topic of many recent works (Rosales Núñez et al., 2019; Specia et al., 2020).

Several UGC parallel corpora (Michel and Neubig, 2018a; Rosales Núñez et al., 2019) have been introduced to evaluate the robustness of MT, some of which, such as (Fujii et al., 2020), are specially annotated to identify UGC idiosyncrasies allowing to measure the impact of a given specificity. Our analyses (§2), indeed, show that measuring the average-case performance using a standard metric on a UGC test set falls far short of giving a reliable image of the UGC translation quality: explaining the observed performance gap requires a particular evaluation framework made of tailored metrics and specific test sets in which UGC idiosyncrasies have been precisely annotated.

That is why, following this line of works, we introduce `PMUMT`, a new parallel data set for the evaluation of UGC translation between French and English in which UGC specificities have been manually annotated using a fine-grained typology. `PMUMT` is larger, relies on a more refined error typology and, more importantly, its annotations are more detailed than existing noisy parallel corpora. Its annotation scheme enables us to generate automatically parallel corpora in which the kind and number of UGC specificities are precisely controlled. Contrary to many works studying the robustness of NMT systems by adding artificial noise to canonical corpora, `PMUMT` is made of attested UGC examples.

Using this framework, we conduct several experiments on three out-of-the-box NMT architectures in a zero-shot scenario, to measure more precisely than what was possible before the impact of the different kinds of UGC specificities on translation quality. Surprisingly enough, our experiments (§3) on natural data show that out-of-the-box models exhibit unexpected strong robustness against several kinds of noise, questioning several results reported in the literature (Michel and Neubig, 2018a; Belinkov and Bisk, 2018). We believe that this data set and its associated evaluation framework will pave the way for a better understanding of the interactions at play in neural machine translation of noisy user-generated content contexts.

2 Testing Out-of-the-Box NMT models on UGC

2.1 Experimental Setting

Training Data Because of the lack of a large parallel data set of noisy sentences, we train our systems on 'standard' parallel data sets: `WMT` (Bojar et al., 2016) and `OpenSubtitles` (Lison et al., 2018). The former contains canonical texts (2.2M sent.) and the latter (9.2M sent.) is made of informal dialogues found in popular sitcoms.

↓ Metric / Test set →	PFSMB[†]	PMUMT[†]	MTNT[†]	4SQ[†]	NewsTest	OpenSubsTest
3-gram KL-Div	1.563	1.442	0.471	0.500	0.406	0.006
%OOV	12.63	11.47	6.78	3,46	3.81	0.76
PPL	599.48	596.12	318.24	293.67	288.83	62.06

Table 1: Domain-related measure on the source side (FR), between used Test sets and other noisy UGC corpora using OpenSubtitles as training set. Dags indicate UGC corpora. *4SQ is the 4Square UGC data set introduced in (Berard et al., 2019). PPL: perplexity, KL-Div: Kullback-Leibler divergence.*

	WMT				OpenSubtitles			
	PFSMB[†]	MTNT[†]	News[◇]	OpenTest	PFSMB[†]	MTNT[†]	News	OpenTest[◇]
BPE-based models								
Seq2seq	9.9	21.8	27.5	14.7	17.1	27.2	19.6	28.2
+ <UNK> rep.	**17.1**	**24.0**	**29.1**	**16.4**	26.1	28.5	24.5	28.2
Transformer	15.4	21.2	27.4	**16.4**	**27.5**	**28.3**	**26.7**	**31.4**
Character-based models								
char2char	7.1	13.9	18.1	8.8	23.8	25.7	17.8	26.3

Table 2: BLEU scores for our models. The † symbol indicates the UGC test sets, and ◇ in-domain test sets.

UGC Test Sets To evaluate the different NMT models, we consider two data sets of manually translated UGC: MTNT (Michel and Neubig, 2018a) and the Parallel French Social Media Bank corpus (PFSMB) (Rosales Núñez et al., 2019)[1] which extends the French Social Media Bank (Seddah et al., 2012) with English translations. These two data sets raise many challenges for MT systems: they notably contain characters that have not been seen in the training data (e.g. emojis), rare character sequences (e.g. inconsistent casing or usernames) as well as many OOVs denoting URL, mentions, hashtags or more generally named entities (NE). Most of the time, OOVs are exactly the same in the source and target sentences.

NMT Models [2] In our experiments, we use three translation models. The first two models are standard NMT models that take as input BPE tokenized sentences: the model used in (Michel and Neubig, 2018a), a Seq2seq bi-LSTM architecture with global attention decoding as implemented in XNMT (Neubig et al., 2018) as well as a vanilla Transformer model as implemented in the OpenNMT toolkit (Klein et al., 2018).

We also consider a char-based model, namely the char2char of Lee et al. (2017). Using char-based models which are, by nature, open-vocabulary to translate UGC is intuitively appealing as these models are designed specifically to address the problem of translating OOVs and to deal with noisy input (Belinkov and Bisk, 2018).

As the Seq2seq model we consider in our experiments is not able to translate OOVs, we introduce, as part of our translation pipeline, a post-processing step in which the translation hypothesis is aligned with the source and <UNK> tokens are replaced by their aligned source token. In the case of our Transformer model, OpenNMT performs this automatically.

2.2 Results

Table 2 reports the BLEU scores (Papineni et al., 2002) [3] of the different models we consider both on canonical and non-canonical test sets. Contrary to the first results of Michel and Neubig (2018a), the quality of UGC translation does not appear to be so bad: the drop in performance observed on non-canonical corpora is of the same order of magnitude as the drop observed when translating out-of-domain data.

These results seem to indicate that, counterintuitively, translating UGC does not raise any specific challenges. We however believe that they are biased by the evaluation metric used: as UGC contains many mentions, URLs emoticons, or named entities that are the same in the source and in the target sentence, BLEU scores estimated on a canonical and on a non-canonical can not be directly compared: BLEU scores on non-canonical data are artificially high as systems are rewarded for simply coping source tokens, which is the most natural

[1] https://gitlab.inria.fr/seddah/parallel-french-social-mediabank

[2] Models parameters are detailed in the appendix.

[3] All BLEU scores are calculated by Post (2018)'s SacreBleu using the *intl* tokenization

solution to translate OOVs. For instance, the BLEU score between the *sources* and references of the PFSMB is 15.1 while it is only 2.7 on the WMT test set. That is why, we believe that the usual MT metrics overestimate the translation quality on UGC and we introduce, in the next section, a new corpus and a new way to measure the real impact of UGC specificities on translation quality.

3 Analyzing the Impact of UGC on Translation Quality

In order to understand the impact of UGC specificities on translation quality, we have annotated a new corpus in which UGC peculiarities are identified in each source sentence and 'normalized' to a canonical form.

3.1 A Corpus Annotated with UGC Specificities

The PMUMT corpus To understand the impact of UGC peculiarities, we manually annotated 400 source sentences sampled from the PFSMB: one of the authors, fluent in French and with good knowledge of UGC, identified spans in the sentence that differ from canonical French and characterized these specificities using the fine-grained typology of Sanguinetti et al. (2020) (see Table 4). Since the whole annotation process was done by a single person, no inter-annotator agreement can be calculated. Nevertheless, results of our pilot study for each individual UGC peculiarity (cf. Table 3 and Table A.3 for cross-metrics analysis), show that MT performance consistently performs better on our normalized corpus than on the original noisy set.

Each span containing an UGC specificity has been 'normalized' to a form closer to canonical French.[4] Table A.1 shows some examples of annotated (source) sentence. A normalized form of each target (i.e. English) sentence has also been produced to ensures that the target can be generated from the 'normalized' source.

At the end, the annotation of this corpus represents 200h of work, comprising an iterative improvement and debugging of the annotations to achieve the corpus' current version. [5]

The resulting corpus contains more than 1, 310 annotations. On average, each sentence contains 2.8 UGC peculiarities. Figure 1 describes the distribution of UGC peculiarities in the corpus.

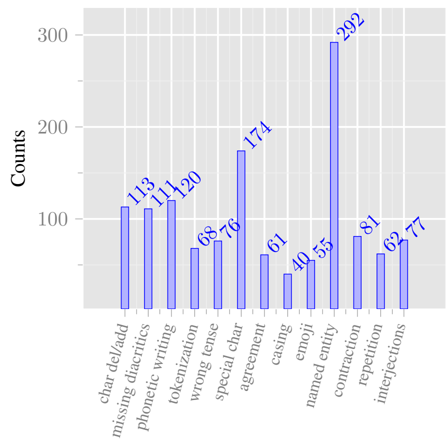

Figure 1: Distribution of UGC specificites of the FR UGC sources in PMUMT.

Controlling the Number of Specificities per Sentence Comparing the predictions of an NLP system taking either the normalized sentences or the original non-canonical sentences as input allows us to measure the impact of UGC on this system. However, it is impossible to perform a fine-grained analysis in which, for example, the impact of different types of specificities are compared, since UGC sentences generally contain several specificities of different types and the interactions between them cannot be easily neutralized.

That is why we have also constructed automatically a second version of our corpus to help us analyze the interactions between the UGC specificities in a sentence: by substituting only some of the span we have annotated, we can create corpora in which the number and the kind of specificities present in each sentence is tightly controlled. In this framework, each original sentence can be (partially) rewritten into as many sentences as there are UGC specificities in it.

This possibility of partial substitution greatly reduces the amount of data to be annotated for our analyses: instead of having to annotate a large amount of data to find enough sentences fulfilling the requested criteria, we are able to generate these

[4]To ensure that this normalization has actually made our corpus closer to a canonical corpus, we have computed the perplexity of the original sentences and of the normalized sentences estimated by a 5-gram Kneser-Ney language model trained on the OpenSubtitles corpus: the normalized version has a perplexity of 2,214 (and 11.60% of its token are OOVs) far lower that the original version (with a perplexity of 8,546 and an OOV ratio of 19.60%).

[5]The annotated corpus and code collection can be found in https://github.com/josecar25/PMUMT_annotated_UGC_corpus/

	char del/add	missing diacritics	phonetic writing	tokenization	wrong tense	special char	agreement	casing	emoji	named entity	contraction	repetition	interjections
s2s	0.80 (28.7)	0.95 (33.9)	0.93 (27.3)	0.96 (30.8)	0.94 (30.7)	0.88 (26.1)	0.95 (27.1)	0.75 (27.7)	0.91 (31.0)	0.86 (31.7)	0.95 (30.8)	0.90 (30.2)	0.93 (29.2)
c2c	0.99 (32.5)	0.99 (29.6)	0.86 (25.2)	1.00 (31.9)	0.97 (28.8)	0.81 (24.6)	0.96 (28.9)	0.86 (28.0)	0.83 (26.2)	0.94 (32.7)	0.91 (30.4)	0.95 (26.2)	0.91 (28.7)
TX	0.98 (35.3)	1.02 (34.0)	1.03 (33.2)	0.98 (32.9)	1.02 (33.7)	0.92 (29.2)	0.97 (33.8)	0.90 (26.9)	0.75 (28.3)	0.99 (35.4)	0.93 (31.1)	0.89 (36.8)	0.86 (30.2)

Table 3: BLEU score ratios between pairs of noisy and normalized sets of sentences, containing only one UGC specificity. BLEU scores on noisy sets are shown in parenthesis.

code	kind of specificities
1	Letter deletion/addition
2	Missing diacritics
3	Phonetic writing
4	Tokenisation error
5	Wrong verb tense
6	#; @, URL
7	Wrong gender/grammatical number
8	Inconsistent casing
9	Emoji
10	Named Entity
11	Contraction
12	Graphemic/punctuation stretching
13	Interjections

Table 4: Typology of UGC specificities used in our manual annotation.

sentences from our original annotation of 400 sentences. We believe that this approach could be of great interest to perform fine-grained error analysis for NLP systems dealing with UGC.

3.2 Impact of UGC Peculiarities on Translation Quality

We used the PMUMT corpus to evaluate the impact of UGC peculiarities on translation quality: we have reported in Table 5 the BLEU scores achieved by the considered systems on both the 400 original sentences and the 400 normalized sentences. As expected, translations of normalized sentences, that are more similar to the training data, are of better quality than translations of original (noisy) sentences: the BLEU scores achieved when translating normalized UGC content are close to those obtained on the in-domain test-set.

For all systems, considering the non-canonical original sentences results in a drop in translation quality of the same order of magnitude, which shows that, even if these models build sentence representations from completely different information, the presence of UGC peculiarities has a similar impact on all of them.

	original	normalized
`Seq2seq`	25.8	32.4
`char2char`	24.1	30.5
`Transformer`	28.6	33.6

Table 5: BLEU scores on the original and normalized source sentences of the PMUMT corpus.

Individual UGC Errors To get a more precise picture of the impact of UGC on translation quality, we have computed, for each kind of peculiarities, the BLEU scores achieved on the corpus built to contain only this peculiarity and the BLEU score computed on the 'normalized' version of the same sentences. Table 3 reports the ratio between these two scores (detailed results are reported in Table A.3 in the supplementary material).

The impact of a given kind of UGC specificity on translation quality is very different from one system to another: it appears that the source sentences representation that MT systems learn to construct are not sensitive to the same kind of noise or errors in the source sentence and even seem to be complementary. For instance, inconsistent casing strongly penalizes the `Seq2seq` model but has only a limited impact on the `char2char` model. On the contrary, the presence of characters specific to online conversation such as @ or # results in a substantial decrease of translation quality for `char2char`, but has less impact for `Seq2seq` or `Transformer`, suggesting that char-based models are not able to properly modeled characters that hardly appear in the training set.

Interestingly, the `Transformer` model appears to be very robust to a wide array of UGC peculiarities, even if it was not designed specifically to handle noisy input: in particular, the presence of named entities, spelling errors (i.e. substitution, deletion or insertion of letters), agreement error (of verb tense or in gender and number) as well as `tokenization` errors hardly hurt translation quality. Similarly, the `char2char` model succeed in translating correctly sentences with letter addition or suppression, showing that the model actually manage to learn sentence representations that are robust to spelling errors even if such errors are not present at training time. This results is at odd with the conclusion drawn by Belinkov and Bisk (2018) on artificial data.

Combination of Peculiarities To better understand the impact of combinations of UGC peculiarities on translation quality, Table 6 reports the ratio between the BLEU scores computed on the translation of a corpus in which there are exactly N different UGC peculiarities in a sentence and on the translation of the normalized version of these sentences. It appears that for all our systems translation quality decreases linearly with the number of specificities, suggesting that the impacts of the different specificities are independent of each other. Surprisingly enough, the gap between the `char2char` and `Transformer` is getting smaller with the number of specificities in each sentence.

	1	2	3	4+
# sents.	1,306	1,776	1,439	1,089
s2s	0.90 (30.1)	0.83 (27.0)	0.77 (24.2)	0.75 (23.2)
c2c	0.92 (29.5)	0.87 (26.6)	0.83 (24.3)	0.83 (23.2)
TX	0.96 (32.8)	0.89 (30.0)	0.86 (28.3)	0.84 (26.5)

Table 6: BLEU score ratio between pairs of normalized and noisy sentences containing N specificities. BLEU scores on noisy sentences are shown in parenthesis.

4 Conclusions

This work introduces `PMUMT` a new corpus of UGC translation in which UGC specificities are manually annotated using a fine-grained typology. Thanks to our detailed annotation process, we were able to build a new framework that allows us to automatically generate several parallel corpora in which the number and kind of UGC specificities is precisely controlled.

Our experiments show that, contrary to what was previously believed, out-of-the-box NMT models are robust to many different kind of UGC specificities and that the different architectures we tested are complementary, in the sense that they are not sensitive to the same specificities. In our future work, we plan to explore the intricacies of the robustness that seem to be linked to specific UGC idiosyncrasies. We make this data set and its associated evaluation framework public[6] as we believe it can pave the way for a better understanding of the interactions at play in neural machine translation of noisy user-generated content contexts.

Acknowledgments

We thank the reviewers for their valuable feedbacks. This work was funded by the French Research Agency via the ANR ParSiTi project (ANR-16-CE33-0021).

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. 2019. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019, Hong Kong, November 4, 2019*, pages 168–176. Association for Computational Linguistics.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin M. Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 131–198.

Ryo Fujii, Masato Mita, Kaori Abe, Kazuaki Hanawa, Makoto Morishita, Jun Suzuki, and Kentaro Inui. 2020. PheMT: A phenomenon-wise dataset for machine translation robustness on user-generated contents. In *Proceedings of the 28th International Con-*

[6] `https://github.com/josecar25/PMUMT_annotated_UGC_corpus/`

ference on Computational Linguistics, pages 5929–5943, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. Opennmt: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*, pages 177–184.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *TACL*, 5:365–378.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*

Paul Michel and Graham Neubig. 2018a. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018b. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 543–553.

Graham Neubig, Matthias Sperber, Xinyi Wang, Matthieu Felix, Austin Matthews, Sarguna Padmanabhan, Ye Qi, Devendra Singh Sachan, Philip Arthur, Pierre Godard, John Hewitt, Rachid Riad, and Liming Wang. 2018. XNMT: the extensible neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*, pages 185–192.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Comparison between NMT and PBSMT performance for translating noisy user-generated content. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 2–14, Turku, Finland. Linköping University Electronic Press.

Manuela Sanguinetti, Lauren Cassidy, Cristina Bosco, Özlem Çetinoglu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *CoRR*, abs/2011.02063.

Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The french social media bank: a treebank of noisy user generated content. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 2441–2458.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

Supplementary Materials

①	src	**JohnDoe389** (10) qui n'arrive pas **a** (2) **depasser** (2) 1 a (2) **FlappyBird** ... **ptddddr** (12,13)
	ref	JohnDoe389 who can't score more than 1 at FlappyBird ... lmaooooo
	N. src	Jean qui n'arrive pas à dépasser 1 à Jean ...
	N. ref	Jean who can't score more than 1 at Jean...
②	src	**#CaMeVénèreQuand** (6) le matin **a** (2) 7h on me parle alors que je suis pas encore **réveiller.** (5)
	ref	#ItAnnoysMeWhen in the moring at 7 am someone talks to me although I didn't wake up yet.
	N. src	le matin à 7h on me parle alors que je suis pas encore réveillé.
	N. ref	in the moring at 7 am someone talks to me although I didn't wake up yet.
③	src	vu sa **tete** (2) **c** (3) normal **kon** (3) **est** (3) **jms** (11) **parler** (5) d'elle !
	ref	in light of her face it's normal no one ever spoke about her!
	N. src	vu sa tête c'est normal qu'on a jamais parlé d'elle !
	N. ref	in light of her face it's normal no one ever spoke about her!
④	src	y a ma cousine qui **joue a** (2) flappy bird (10) **mdrrrrrrrrrrr** (12, 13) **elle et plus nuuul** (12,7) **que moi**
	ref	my cousin plays flappy bird looooooooooool she's more hopeless than me
	N. src	y a ma cousine qui joue à Jean Jean elle et plus nulle que moi
	N. ref	my cousin plays Jean Jean she's more hopeless than me

Table A.1: Examples from our annotated noisy UGC corpus. Source sentences have been annotated with UGC specificities of Table 4 (in blue) according to their numerical code. For each example, the original source and reference (*src* and *ref*) and their corresponding normalized version (*N. src* and *N. ref*) are shown.

		Letter deletion/addition/change
①	src	j'arrive pas à **boir normalemen**
	norm	j'arrive pas à boire normalement
	ref	I can't **drink normally**
	s2s	I can't **drink normal.**
	c2c	I can't **drink normally.**
	Tx	I can't **drink normal men.**
②	src	Je conseille à **toux** ceux qui ont l'esprit disons, un peu fermé de regarder sur les "Français d'origine contrôlée"
	norm	Je conseille à tous ceux qui ont l'esprit disons, un peu fermé de regarder sur les "Français d'origine contrôlée"
	ref	I advise **everyone** with a, let's say a little narrow mind to watch about the "Français d'origine contrôlée"
	s2s	I suggest **cough** those who have minds say, a little closed to look at the "frances of controlled origins"
	c2c	I counsel **those** who have the mind, a little close to looking at the French original controlled original controlled.
	Tx	I advise **anyone** with a mind, say, a little closed to look at the controlled French.
⑤	src	le côté suis **tro cool** au quotidien et je **relach** tout **quan** j'ai bu
	norm	les gens qui m'aiment me détestent quand j'ai bu
	ref	my side **very cool** in everyday life and **loosen** everything **when** I've been drinking
	s2s	I've been drinking all the time and I've been drinking everything **quan** I've been drinking.
	c2c	the side of the daily **cool** side and **relacing** everything **when** I've been drinking
	Tx	I'm the **cool** side. I'm the cool one.
		Tokenization
⑥	src	**J'sais pas vous**, mais de voir la joie des grands joueurs comme Zlatan, Motta, Verratti je trouve ça magnifique
	norm	Je sais pas vous, mais de voir la joie des grands joueurs comme Jean, Jean, Jean je trouve ça magnifique
	ref	**I don't know about you**, but seeing the joy of great players like Zlatan, Motta, Verratti I think it's wonderful
	s2s	**I don't know you**, but seeing the joy of the great players like Zlatan, Motta, Verratti, I think it's beautiful.
	c2c	**I don't know about you**, but to see the joy of great players like Zlatan, Motta, Varratt, I think it's beautiful.
	Tx	**I don't know about you**, but seeing the joy of big players like Zlatan, Motta, Verratti, I think it's beautiful.
⑦	src	pendant que vous me laissez en chien à l'atelier mon score de flappy bird fait que **d augmenter**
	norm	pendant que vous me laissez en chien à l'atelier mon score de Jean fait que d'augmenter
	ref	while you're bailing out on me at the workshop my flappy bird score **is just increasing**
	s2s	when you leave me as a dog when you leave me as a dog at the workshop.
	c2c	while you leave me dog at the workshop my flappy bird score **is that increasing**
	Tx	while you leave me as a dog at the workshop my flappy bird score **is just up.**
⑧	src	il **ma** dit que c'était normal aussi et que ça allait redescendre,
	norm	il m'a dit que c'était normal aussi et que ça allait redescendre,
	ref	**he told me** it was normal too and that it would come down,
	s2s	**He said it** was normal, too, and it was going to go down,
	c2c	**He told me it** was normal, too, and it was going back,
	Tx	**He told me it** was normal, too, and it was gonna come down,
		Inconsistent casing
⑨	src	Jean **DANS VOS YEUX**
	norm	Jean dans vos yeux
	ref	Jean **IN YOUR EYES**
	s2s	Jean **D in VOSY**
	c2c	Jean **in your eyes**
	Tx	Jean **in your eyes**
⑩	src	**JE VIENS DE VOIR** Jean **ET** Jean **JE PEUX PLUS**
	norm	Je viens de voir Jean je peux plus
	ref	I JUST WATCHED Jean AND Jean CAN'T TAKE IT
	s2s	I'm going to kill Jean and Jean I can't believe it.
	c2c	I'm here to see Jean And Jean I can no longer.
	Tx	I just saw Jean and Jean again.
		Domain-specific characters and emojis
❶	src	Avec mes magnifiques jumeaux Jean et Jean **@maxcarver @Charlie_Carver ♥**
	norm	Avec mes magnifiques jumeaux Jean et Jean
	ref	With my wonderful twins Jean and Jean **@maxcarver @Charlie_Carver ♥**
	s2s	with my gorgeous Jean and Jean **@maxarver @Carlie_Carver @Carlie_Carver #**
	c2c	with my beautiful Jean twins, Jean Jean and Jean **Charlier Charlier Carver.**
	Tx	with my beautiful twins Jean and Jean **imexcarver Charlie_Charver @Charver**

Table A.2: Examples from our noisy UGC corpus showing the `Transformer`, `char2char` and `Seq2seq` predictions. Present UGC specificities of Table 4 (in blue) are marked in bold.

As some of the data sets contain as few as 40 sentences, we have also computed the 95% confidence interval for all BLEU scores in Table A.3 using the bootstrapping method described in Koehn (2004). The width of all intervals is smaller than 0.30 for the BLEU scores (roughly 1% of the corresponding score) and than 0.006 for the ratios, which shows that we can trustfully compare their values.

Similarly, we have included results for the CHRF (Popović, 2015) and MULTI-BLEU-DETOK.PERL evaluation metrics since SACREBLEU showed a ratio between performances on noisy and clean text versions, indicating that the noisy version could be easier to translate than its normalization. In this regard, we can see in Table A.3 that at least one of the three reported metrics gives a ratio value equal or smaller than 1.0 within the 95% confidence interval error (CI Err.), suggesting that our normalization introduce limited artificial noise, comparable to the difference between correlated evaluation metrics.

	Metric	char del/add	missing diacritics	phonetic writing	tokenization	wrong tense	special char	agreement	casing	emoji	named entity	contraction	repetition	interjections
s2s	MB	0.78 (25.3)	0.94 (31.1)	0.92 (23.2)	0.95 (30.6)	0.98 (28.6)	0.83 (24.8)	0.95 (25.1)	0.77 (26.7)	0.87 (29.6)	0.87 (30.9)	0.91 (28.4)	0.83 (29.9)	0.90 (26.5)
	chrF	0.93 (46.7)	0.97 (53.1)	0.89 (43.1)	0.95 (50.9)	0.99 (50.6)	0.91 (44.3)	0.98 (49.3)	0.76 (40.2)	0.94 (51.1)	0.94 (51.7)	0.93 (47.3)	0.93 (47.3)	0.96 (48.0)
	SB	0.80 (28.7)	0.95 (33.9)	0.93 (27.3)	0.96 (30.8))	0.94 (30.7)	0.88 (26.1)	0.95 (27.1)	0.75 (27.7)	0.91 (31.0)	0.86 (31.7)	0.95 (30.8)	0.90 (30.2)	0.93 (29.2)
c2c	MB	1.00 (29.5)	1.00 (27.4)	0.85 (22.5)	0.99 (29.7)	0.97 (26.9)	0.80 (23.5)	0.97 (25.5)	0.91 (27.7)	0.83 (25.1)	0.95 (31.7)	0.88 (26.6)	0.93 (28.0)	0.91 (25.7)
	chrF	0.99 (48.5)	1.00 (50.6)	0.92 (44.8)	0.95 (50.1)	0.99 (49.1)	0.84 (44.0)	0.98 (49.9)	0.78 (40.6)	0.93 (49.5)	0.95 (51.6)	0.92 (48.8)	0.90 (47.8)	0.95 (49.7)
	SB	0.99 (32.5)	0.99 (29.6)	0.86 (25.2)	1.00 (31.9)	0.97 (28.8)	0.81 (24.6)	0.96 (28.9)	0.86 (28.0)	0.83 (26.2)	0.94 (32.7)	0.91 (30.4)	0.95 (26.2)	0.91 (28.7)
TX	MB	0.96 (30.3)	1.01 (33.0)	0.98 (33.2)	0.98 (31.5)	1.01 (31.6)	0.90 (28.4)	0.97 (31.4)	0.98 (25.8)	0.72 (26.7)	1.06 (35.7)	0.90 (28.4)	0.81 (25.9)	0.83 (27.0)
	chrF	0.95 (48.2)	1.00 (52.3)	0.98 (46.6)	0.99 (51.0)	1.01 (52.4)	0.93 (46.5)	0.97 (50.9)	0.80 (30.7)	0.88 (49.1)	1.00 (52.6)	0.93 (48.9)	0.87 (46.2)	0.92 (46.2)
	SB	0.98 (35.3)	1.02 (34.0)	1.03 (33.2)	0.98 (32.9)	1.02 (33.7)	0.92 (29.2)	0.97 (33.8)	0.90 (26.9)	0.75 (28.3)	0.99 (35.4)	0.93 (31.1)	0.89 (36.8)	0.86 (30.2)
CI Err.	(E-3)	4.5 (0.17)	1.5 (0.13)	2.7 (0.11)	2.6 (0.17)	2.4 (0.15)	1.8 (0.12)	1.7 (0.23)	5.7 (0.30)	3.0 (0.23)	2.2 (0.11)	2.2 (0.16)	2.5 (0.24)	3.1 (0.22)

Table A.3: BLEU score ratios between pairs of noisy and normalized sets of sentences, containing only one UGC specificity. BLEU scores on noisy sets are shown in parenthesis. *Three different metrics are shown for comparison:* `MultiBleu-detok.perl` *(MB)*, `chrF` *and* `SacreBleu` *(SB).* Error for 95% confidence intervals (CI Err).

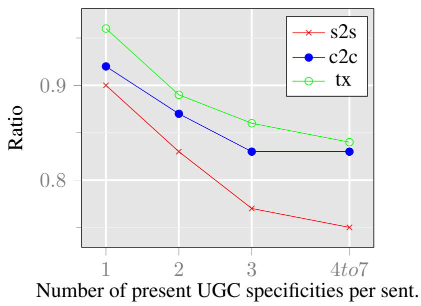

Figure A.1: Noisy/Clean BLEU scores' ratios for accumulated number of UGC specificities present per sentence for each model, corresponding to the results in Table 6. The *4to7* bin groups more than 4 types to provide a larger subcorpus, which weighted average is 4.34 UGC specificities per sentence.

A Reproducibility

Data All the UGC test sets and source code for our experiments are provided in the supplementary materials. For training data, we let the reader refer to each project's website for `WMT`[7] (consisting of `Europarlv7`, `Newcommentariesv10` and `Open Subtitles`[8], both accessed on November, 2019. Regarding clean test sets, we used `newstest15` from `WMT` and a subset of 11,000 unique phrases from `Open Subtitles`. We make the former test available in the link provided above for exact performance comparison.

Computation The NMT systems were trained using 1 Tesla V100, during an average of 72 hours to converge to the final solution for the `char2char` model and 56 hours for the BPE-based baselines.

A.1 NMT Models

Character-based models `char2char` models were trained as out-of-the box systems using the implementations provided by (Lee et al., 2017) .[9]

BPE-based models We consider, as baseline, two standard NMT models that take, as input, tokenized sentences. The first one is a `seq2seq` bi-LSTM architecture with global attention decoding. The `seq2seq` model was trained using the XNMT toolkit (Neubig et al., 2018).[10] It consists in a 2-layered Bi-LSTM layers encoder and 2-layered Bi-LSTM decoder. It considers, as input, word embeddings of 512 components and each LSTM units has 1,024 components.

We also study a vanilla Transformer model using the implementation proposed in the OpenNMT framework (Klein et al., 2018). It consists of 6 layers with word embeddings of 512 components, a feed-forward layers made of 2,048 units and 8 self-attention heads.

Hyper-parameters In Table A.4, we list the training variables set for our experiments. They match their corresponding default hyper-parameters.

Batch size	64
Optimizer	Adam
Learning rate	1e-4
Epochs	10 (best of)
Patience	2 epochs
Gradient clip	[-1.0, 1.0]

Table A.4: Hyper-parameters used for training the NMT systems.

Pre-processing For the BPE models, we used a 16K merging operations tokenization employing `sentencepiece`[11]. For word-level statistics we segmented the corpora using the `Moses` tokenizer [12].

[7] https://www.statmt.org/wmt15/translation-task.html

[8] http://opus.nlpl.eu/download.php?f=OpenSubtitles/v2018/moses/en-fr.txt.zip

[9] https://github.com/nyu-dl/dl4mt-c2c

[10] We decided to use XNMT, instead of OpenNMT in our experiments in order to compare our results to the ones of Michel and Neubig (2018b).

[11] https://github.com/google/sentencepiece

[12] https://github.com/moses-smt/mosesdecoder