RoCS-MT: Robustness Challenge Set for Machine Translation

Rachel Bawden
Inria, Paris, France
rachel.bawden@inria.fr

Benoît Sagot
Inria, Paris, France
benoit.sagot@inria.fr

Abstract

RoCS-MT, a Robust Challenge Set for Machine Translation (MT), is designed to test MT systems' ability to translate user-generated content (UGC) that displays non-standard characteristics, such as spelling errors, devowelling, acronymisation, etc. RoCS-MT is composed of English comments from Reddit, selected for their non-standard nature, which have been manually normalised and professionally translated into five languages: French, German, Czech, Ukrainian and Russian. In the context of the WMT23 test suite shared task, we analyse the models submitted to the general MT task for all from-English language pairs, offering some insights into the types of problems faced by state-of-the-art MT models when dealing with non-standard UGC texts. We compare automatic metrics for MT quality, including quality estimation to see if the same conclusions can be drawn without references. In terms of robustness, we find that many of the systems struggle with non-standard variants of words (e.g. due to phonetically inspired spellings, contraction, truncations, etc.), but that this depends on the system and the amount of training data, with the best overall systems performing better across all phenomena. GPT4 is the clear frontrunner. However we caution against drawing conclusions about generalisation capacity as it and other systems could be trained on the source side of RoCS and also on similar data.

1 Introduction

As the quality of state-of-the-art machine translation (MT) systems is becoming indistinguishable in certain scenarios and domains from that of human translators (Kocmi et al., 2022), the task of tackling the translation non-standard texts is becoming an increasingly realisable aim. A considerable proportion of texts produced today are done so online in informal, unedited settings, e.g. on forums such as Twitter and Reddit, and MT is frequently to make posts accessible to a global audience. However, it has been shown that MT still struggles with user-generated content (UGC) (Gupta et al., 2023), as the type of language can differ considerably from the edited texts that have traditionally been used to train and evaluate MT models.

The RoCS-MT challenge set (Robust Challenge Set for Machine Translation) is designed to provide a test bed for the automatic translation of non-standard UGC phenomena. It contains approximately 2k sentences from the online forum Reddit that have been manually normalised and professionally translated into five languages: French, German, Czech, Ukrainian and Russian. The sentences were selected specifically for the presence of non-standard phenomena, of which we provide manual annotations (e.g. spelling errors, devowelling, capitalisations, acronymisms, etc.). Inspired by other datasets such as the French Social Media Bank (Seddah et al., 2012) and its parallel component (Rosales Núñez et al., 2019), our aim is to provide an evaluation set that is more challenging than certain previous efforts, such as the commonly used MTNT dataset (Michel and Neubig, 2018). We also make different choices from most previous efforts concerning the guidelines for normalisation and translation of the source sentences. We choose to first normalise the source sentences before translation in order to optimise the quality of the translation and to reduce the arbitrariness that may be introduced when transferring non-standard variation to the target language (e.g. on which characters to apply spelling errors, how many characters to duplicate when elongating words). For normalisation, we aim to strike a balance between normalisation as much as possible while making sure that the normalised text remains natural.[1]

In this paper, we describe the creation of the challenge set, and in the context of the WMT23 test suite shared task, we analyse the models submitted

[1] E.g. We choose to not normalise the acronym *lol* 'laughing out loud', as it is rarely/never used in its expanded form.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 198–216
December 6–7, 2023.

to the general MT shared task for the from-English shared task language pairs: English→{Czech, German, Hebrew, Japanese, Russian, Ukrainian, Chinese} (en→{cs, de, he, ja, ru, uk, zh}). Through automatic and manual analysis of system outputs, we find that many of the phenomena remain challenging for most systems (in particular those that create potential out-of-vocabulary or rare words such as phonetically inspired spellings, contractions, devowelling and truncation). However, the difficulty varies depending on the phenomenon, the particular instance (notably how frequent the non-standard word is) and the system, especially with respect to the quantity of training data. The highest performing systems overall generally do better across the board on all phenomena, whereas the weaker systems struggle in particular with certain phenomena. GPT4 has a clear lead over other systems, correctly translating even some of the most challenging examples and sometimes (although inconsistently) reproducing non-standardness in its outputs. However conclusions are limited given that the training data is unknown (as is the case of other unconstrained systems).

We make the challenge set, system outputs, evaluation code and guidelines (for the normalisation, annotation and translation) openly available for research purposes.[2]

2 Related Work

Several parallel UGC datasets exist across different language pairs. While some are extracted automatically from crawled data (Ling et al., 2013; Vicente et al., 2016; Mubarak et al., 2020), a majority are based on monolingual sentences that are then translated into the target language (Sluyter-Gäthje et al., 2018; Michel and Neubig, 2018; Rosales Núñez et al., 2019; Fujii et al., 2020; McNamee and Duh, 2022). The closest to our RoCS-MT dataset are (Michel and Neubig, 2018) and (Rosales Núñez et al., 2019), which were designed to contain challenging non-standard phenomena, whereas many of the existing datasets do not apply any such filter. Like RoCS-MT, the MTNT dataset (Michel and Neubig, 2018) contains texts from Reddit. To target non-standard language, they select sentences that have a low probability using a language model trained on standard data. In practice, and as shown by Rosales Núñez et al. (2019), the amount of non-standard language remains limited with this method. Rosales Núñez et al. (2019) base their parallel dataset on the French Social Media Bank dataset (Seddah et al., 2012), which targets non-standard language by searching for specific non-standard keywords. They show that this leads to a higher level of non-standard language, although the method is by nature more biased towards the keywords and phenomena used for data selection. An error analysis of the dataset was conducted in (Rosales Núñez et al., 2021), showing MT quality (using BLEU) for different UGC phenomena.

Despite significant effort to describe and classify UGC phenomena (Michel and Neubig, 2018; Sanguinetti et al., 2020), there is no consensus as to how texts should be normalised (and indeed translated). One extreme is to normalise all phenomena to standard forms, as is often done in lexical normalisation tasks (Han and Baldwin, 2011; van der Goot et al., 2021), but which in several cases would lead to unnatural outputs (e.g. if *lol* and *lmao*, were systematically normalised to *laughing out loud* and *laughing my ass off*). This makes translation difficult too, as the translations would also be unnatural. At the other end of the spectrum is the choice to not normalise source texts and in addition to attempt to translate the phenomena into the target language, with the disadvantage that some phenomena are language-specific[3] and others would result in arbitrary decisions being made such as to which characters to apply spelling errors. The current datasets targeting particularly non-standard phenomena choose to at least in part transfer some phenomena to the target language, whereas we adopt a higher degree of normalisation (see Section 3.1.1 for more details), producing standard but natural-sounding translations.

3 Challenge Set Creation

3.1 Data Sourcing and Selection

The source sentences are taken from English posts on discussion platform Reddit[4] using the API.[5] We do not target a particular variety of English

[2] https://github.com/rbawden/RoCS-MT

[3] Two examples of this are French *verlan*, which consists in inverting syllables in words (e.g. *louche→chelou* 'bizarre') and English cockney rhyming slang (e.g. *loaf* meaning *head* thanks to its rhyme with the expression *loaf of bread*). However, even phenomena that do exist crosslingually do not necessarily apply to the same words (e.g. the use of digits to replace their homophones as in *2day* 'today', where the translation does not necessarily contain a homophone of a digit in the target language).

[4] www.reddit.com

[5] Using the free version of the API (December 2022).

(e.g. British, American, etc.) and even include some non-native English,[6] although we do not include code-switched texts. We get an initial pool of posts by searching for specific keywords from a manually drawn-up list as in (Sanguinetti et al., 2020), e.g. *ttyl*, *ppl*, *gr8*, *alot*. The full list is given in Appendix A. For each keyword, we crawled both Reddit-wide and 3 specific subreddits (CasualUK, MadeMeSmile and entertainment) to ensure a diversity of informal topics[7] at 6-month intervals between 2017 and 2022.

Once we had the initial pool of examples, in order to reduce the number of posts to manually review, we applied a very coarse-grained in-house 'non-standardness' classifier that we had trained on a small set of manually annotated tweets according to 4 labels (standard, mildly non-standard, moderately non-standard and very non-standard), and look at posts whose title or text was marked as anything other than 'standard'. From those posts, we manually select titles and passages from the text that contain interesting non-standard phenomena, including sentences not containing the initial keyword associated with the post. This means that although our initial search process is biased to our word list, the effect is diminished by taking additional non-standard phenomena. We automatically filter out any 18+ content (using the Reddit meta-information), and manually filter out any content that is sexually inappropriate, insulting or deals with sensitive (potentially triggering) topics such as suicide or drug addiction.

3.1.1 Sentence Splitting and Normalisation

We start by manually splitting the texts into sentences. In many cases, this corresponds to splitting on final punctuation (e.g. full stop, exclamation marks, etc.). However, the non-standard nature of the texts increases the number of cases where texts are split in places that are not marked by punctuation or where punctuation or newlines are added unexpectedly in the middle of what would ordinarily be considered a sentence.

For instance, the sequence *I went grocery shopping I'm down to my last dollars soon (...)* was split into the first sentence *I went grocery shopping* and the second sentence beginning with *I'm down to my last dollars soon*, despite the lack of a final punctuation between *shopping* and *I'm*.

The first author (a native English speaker) manually normalised each of the sentences produced by our manual sentence splitting, seeking help from people knowledgeable in the topics (e.g. video gaming) where necessary. The complete normalisation guidelines with examples can be found in the dedicated Github repository.[8] As with any guidelines for dealing with complex and evolving non-standard phenomena, the decisions made are certainly not bulletproof and are likely to evolve in future work. Our aim was to reach a compromise between (i) normalising as much as possible of the text while (ii) rendering the output natural and realistic and (iii) not over-normalising such as to remove the style of the original text. We therefore normalise words such that the normalised variant could be spontaneously and naturally used.

3.1.2 Translation

Translation of the English sentences was carried out by paid professional translators. They had access to the original posts and both the raw and normalised versions of each sentence. Translation was carried out at the sentence level (following the manual segmentation and using as the source the normalised translation), although the translators had access to surrounding linguistic context, as well as additional context and translation notes provided by the first author during the normalisation step. There were also several exchanges between the first author and the translators in order to provide additional context and to answer questions. In order to preserve author anonymity, translators did not have access to meta-information about the authors (e.g. their gender). A single translation was produced for each sentence (we left the choice of speaker gender to the translators) with the exception of Ukrainian, for which two translations were produced for sentences where the speaker gender has an impact.

The target languages were chosen to cover four of those in the WMT2023 general translation task (Czech, German, Ukrainian and Russian), as well as French, which is an important language for our own research, although we do not analyse the French portion of the data in this article.

Translation Guidelines Translators were provided with guidelines (see Appendix B). They were

[6] We do not have access to any personal information about the post authors, but we know this because some posters apologise for their level of English in the posts included.

[7] The subreddits were chosen to have topics that were informal and could have a reasonable number of posts, although in reality, the number of non-standard posts found from these specific subreddits was limited.

[8] https://github.com/rbawden/RoCS-MT

instructed to translate the normalised versions of each sentence into the target language, using standard language but best matching the intention, naturalness and familiarity level of the sentence, similar to the guidelines set out in (McNamee and Duh, 2022). The decision to use standard language was to avoid the arbitrariness associated with attempting to reproduce non-standard phenomena in translation, which would make comparisons, particularly automatic ones, more difficult (e.g. which characters to alter to reproduce a spelling error, how many characters to repeat in the case of expressive repetition, etc.). They were also instructed to respect the manual segmentation provided,[9] to respect punctuation choices made in the source where appropriate (e.g. conserving full stops) and to preserve English words in meta-linguistic discussions (i.e. where authors are writing specifically about English words). As in the normalisation guidelines, abbreviations, acronyms and simplifications were to be expanded unless the result would not make a natural sentence that could realistically be found. However, abbreviations linked to the names of places and institutions were to be kept as they were if used as such in the target language (e.g. French *OTAN* for English *NATO*). They were requested not to use MT systems to help them translate in order not to bias the translations produced.

3.2 Challenge Set Subsets

We create four subsets of the challenge set to test the impact of sentence segmentation (manual or automatic using spaCy) and of normalisation (manual or none, i.e. the original raw text):

- `manseg-raw`: Manual segmentation with original (raw) text
- `manseg-norm`: Manual segmentation with manual normalisation
- `spacyseg-raw`: spaCy segmentation with original (raw) text
- `spacyseg-norm`: spaCy segmentation with manual normalisation[10]

As shown in Section 3.3, the two different segmentation methods result in different numbers of individual sentences, and automatic segmentation with spaCy differs depending on whether the text has been normalised or not. In practice, in this article, we focus only on the `manseg-raw` and `manseg-norm` subsets, although we also release the system outputs for the `spacyset-` subset. We leave research on these other subsets (i.e. looking at the impact of sentence segmentation) to future work.

3.3 Dataset Characteristics

Some basic quantitative characteristics of the data are given Table 1.

Impact of sentence splitting While the number of sentences is fixed for the manual segmentation, spaCy segmentation is highly dependent on whether the text has been normalised or not, likely due to the tool being less well adapted to non-standard text; when applied to raw text, the resulting number of sentences is far lower than manual segmentation (1660 vs. 1922), whereas the resulting number of sentences is more similar to manual segmentation when applied to the normalised text.

Tokenisation Normalisation impacts the number of tokens in the texts, as well as the number of unique tokens. When comparing the two normalised subsets on the one hand and the two raw subsets on the other (i.e. differing only in the sentence splitting), the number of tokens differs due to the fact that automatic segmentation tends to oversplit sentences on punctuation that in the manual segmentation would remain part of a token in the preceding sentence. The number of unique tokens inevitably drops after normalising, due to the homogenisation of non-standard forms (7175 vs. 6612) for manual segmentation.

Normalisation Types We manually annotated the texts for non-standard phenomena (e.g. spelling errors, acronyms, devowelling, capitalisation, pronoun drop, etc.), with the possibility of there being several types for a single span of text. Our annotations are at the word-level, with some phenomena spanning several words (e.g. capitalisation). Table 2 provides some statistics for the annotations occurring in at least 10 sentences, and some examples are given in Examples 1-4.

(1) btw I wud prefer them rlly quick.
By the way, I would prefer them really quick.
acronym contraction devow.
capitalisation
punct_diff

[9] A segment's translation can contain several sentences but sentence boundaries cannot be overridden.

[10] The spaCy segmentation was obtained by concatenating all normalised sentences from a single text and then automatically splitting.

Subset	Seg.	Norm.	#sents.	#toks.	#toks. (unique)	Ave. sent. len.	#posts	#titles	#body
`manseg-raw`	Manual	×	1922	27971	7175	14.55	391	80	263
`manseg-norm`	Manual	✓	1922	28800	6612	14.98			
`spacyseg-raw`	spaCy	×	1660	28095	7297	16.92			
`spacyseg-norm`	spaCy	✓	1996	28881	6615	14.47			

Table 1: Basic statistics of the four subsets of the test suite. Tokens are defined as whitespace delimited character sequences. Sentences can either come from post titles or the body of the post.

Annotation	#toks	#diff toks	#sents
punct_diff	2500	136	1259
capitalisation	2122	802	1059
norm_punct	542	46	339
acronymisation	329	100	277
phonetic_distance	566	285	268
spelling_error	345	306	261
spacing	294	111	250
truncation	203	104	169
contraction	161	37	146
devowelling	137	33	122
elongation	139	96	117
pronoun_drop	114	1	110
word_drop	97	2	85
grammar	75	54	73
inflection	78	64	67
lex_choice	65	52	63
article_drop	69	1	63
scrambled	38	36	37
words_to_digits	45	18	37
word_to_symbol	26	12	22
dialectism	24	15	22
double_to_single_character	17	10	17
word_add	16	13	15
digits_to_words	16	13	14
interjection	13	8	10
surrounding_emphasis	12	11	10
word_order	11	11	10
emoticon	10	10	10

Table 2: For each annotation appearing in at least 10 sentences, the number of words, unique words (lowercased) and sentences for which it appears.

(2)
So any	idea s	on	wot	I shud	be
So any	ideas	on	what	I should	be ?
	spacing		phon._dst.	contraction	punct.

(3)
Dhat	kwik	beizh	fawks	jmmpd	
That	quick	beige	fox	jumped	
phon._dst.	phon._dst.	phon._dst.	phon._dst.	phon._dst.	

(4)
Em	HOW	DARE	YOU	SWEAR	IN
EM:	How	dare	you	swear	in
caps.	caps.	caps.	caps.	caps.	caps.
punct_diff					

FRONT	OF	MY	SUN	
front	of	my	son	?
caps.	caps.	caps.	spelling	punct_diff

4 Translation Systems

In this article, we evaluate the systems submitted to the general translation task at WMT2023. There are both constrained and unconstrained systems, the two settings presenting significant differences in training data that should be taken into account when comparing systems.

Constrained systems Constrained systems followed similar strategies, with many systems doing data filtering/cleaning and data augmentation, using either bilingual or multilingual models and reranking. The constrained systems submitted were AIRC (Rikters and Miwa, 2023), ANVITA, CUNI-Transformer and CUNI-DocTransformer (Popel, 2020) (we refer to these system as CUNI-Trans and CUNI-DocTrans to save space in the results tables), CUNI-GA (Jon et al., 2023), HW-TSC (Wu et al., 2023b), IOL_Research (Zhang, 2023), NAIST-NICT (Deguchi et al., 2023), Samsung_Research_Philippines (Cruz, 2023) (hereafter Samsung_RP), SKIM (Kudo et al., 2023) and UvA-LTL (Wu et al., 2023a).

Unconstrained systems As in previous years of the shared task, translations were produced from anonymised online systems, corresponding in this addition to ONLINE-{A,B,G,M,W,Y} submissions. This year, translations from GPT4 were also produced using 5 few-shot examples (GPT4-5shot).[11] Note that caution should be taken when comparing results from GPT4, given that it is very possible that source sentences from RoCS-MT are included in GPT4's training data. Two systems based on NLLB (Team et al., 2022) were also submitted in the context of the metrics shared task: NLLB_Greedy and NLLB_MBR_BLEU (hereafter NLLB_MBR), which both rely on the same model but differ by the decoding strategy, either standard (greedy) or based on the Minimum Bayes Risk strategy (Freitag et al., 2022). A number of unconstrained systems were also submitted by participants, namely Lan-BridgeMT (Wu and Hu, 2023), KYB , GTCOM (Zong, 2023), (Li et al., 2023), PROMT (Molchanov and Kovalenko, 2023), Yishu

[11]The prompt used is the sentence-level prompt from (Hendy et al., 2023), which is also shown in Appendix C.

(Min et al., 2023) and ZengHuiMT (Zeng, 2023).

5 Evaluation and Analysis

Evaluation of UGC translation is more challenging than standard text; a correct translation can either be standard or non-standard in the target language, and there may be multiple ways of being non-standard that may not all be covered by available references. In our case, we chose to produce standard reference translations (See Section 3.1.2). Any system that produces non-standard language may therefore be underestimated using reference-based metrics.

We test three different metrics (BLEU, COMET and COMET-QE) to evaluate the systems' translations of RoCS-MT, looking at how coherent they are between each other, and whether it is possible to use quality estimation to evaluate MT robustness in order to remove the need for reference translations (Section 5.1). We also look at the MT quality of each system per phenomenon by calculating COMET scores over subsets of the data. Finally, we perform a qualitative analysis, manually looking at how the different systems handle UGC phenomena, and confirming some of the trends using some simple automatic analyses (Section 5.2).

5.1 Automatic evaluation

BLEU (Papineni et al., 2002), as a surface-level metric, is intuitively not robust to variation. It is therefore likely to be particularly ill-adapted to MT robustness evaluation, since MT systems' outputs can display standard or non-standard characteristics. We choose nevertheless to test this here, calculating BLEU scores using the sacreBLEU toolkit (Post, 2018).[12] We compare BLEU to reference-based COMET (Rei et al., 2020)[13] for those language pairs for which we have a reference, and to COMET's reference-less (quality estimation) version, which we refer to as COMET-QE (Rei et al., 2022).[14] We notably aim to test whether it is possible to use COMET-QE for evaluation rather than reference-based COMET, which would remove the dependency on reference translations and make evaluation possible for a wider range of languages.

Ukrainian has two reference translations for sentences for which the speaker's gender results in different translations is ambiguous between male and female. While BLEU is designed to handle multiple references, this is not inbuilt into COMET. For these sentences, we choose to take the best COMET score of the two references. For COMET and COMET-QE, which also use the source sentence, we choose to evaluate system outputs against both the `manseg-norm` and `manseg-raw` source sentences, regardless of which set was translated by the system and take the highest score of all the source-reference combinations. This covers the case where non-standard (i.e. raw) sentences are normalised during the translation process.

We provide full results for COMET and COMET-QE in Table 3 and 4 respectively, and we include results for BLEU in Table 7 in Appendix D.

How coherent are the metrics? The trends of the three metrics are similar but not at all systematic (in terms of rankings) when evaluating translations of the normalised data (`manseg-norm`), with the same systems getting the highest scores across language pairs (amongst the best systems being ONLINE-W, ONLINE-B, GPT4). However, there are some clear inconsistencies between BLEU and the two COMET metrics when evaluating non-standard data (`manseg-raw`). For example GPT4 is ranked above other systems by COMET and COMET-QE, whereas the BLEU scores of other systems (and in particular ONLINE-W and sometimes ONLINE-B) are higher. This indicates that GPT4 outputs are more surfacically different from the reference translations, which could be a result of paraphrasing or non-standard translations rather than a reflection of MT quality, especially given the high scores by COMET.

This confirms that BLEU is poorly adapted to evaluating MT robustness and could even lead to misleading conclusions, confirming previous conclusions drawn by Rosales Núñez et al. (2021) about the inadequacy of BLEU for the evaluation of UGC MT. On the other hand, COMET-QE scores show more similar trends to COMET, suggesting that it could be possible to use it to evaluate without having to produce reference translations. We nevertheless add that COMET remains an automatic metric that does not produce perfect correlation with human judgments, more research would be necessary to stress-test the metric for MT robustness evaluation, particularly in terms of evaluating which of COMET and COMET-QE is better correlated with human judgments.

[12] `case:mixed|eff:no|tok:13a|smooth:exp|v:2.2.1`
[13] We use the default `wmt22-comet-da` model.
[14] We use the default `wmt22-cometkiwi-da`.

Which systems come out on top? The highest performing systems are the unconstrained online systems, with GPT4 getting significantly higher COMET and COMET-QE scores than other systems when translating non-standard (`raw`) text for all languages tested. Other systems that tend to produce high scores are ONLINE-W and to a lesser extent ONLINE-B. Apart from these online models, both NLLB models are the best-scoring ones, which might come from the fact that they are highly multilingual and therefore could be more robust to language variation. The constrained systems, whilst not the highest performing systems, appear to get comparable scores to at least some of the online systems.

Which systems are most robust? This question is linked to the previous question about MT quality on non-standard data. To take into account the base performance of the systems, we look at the difference in score between each system's translation of the non-standard sentences and their normalised versions (also in the previously mentioned Tables 3 and 4. While there is a general trend that the higher performing systems also also have a smaller difference in quality (i.e. they are also more robust), there are some stand-out systems. GPT4 is the system with the lowest quality difference between original and normalised sentences for all language pairs tested. The NLLB models also have a low delta between the two subset, lower than or comparable to some of the more robust online systems. Similarly to the previous question, constrained systems are not the most robust in terms of their score difference. Notably for en–cs and en–de, the score differences are amongst the highest. However, some of the systems do show peformance in the same ballpark as some of the online systems.

Automatic analysis by UGC phenomenon In order to analyse how systems handle different non-standard phenomena, we evaluate sentences by annotation types, by calculating COMET and COMET-QE scores for sentences containing at least one occurrence of a particular normalisation annotation. COMET results are given in Table 5 and we include a fuller analysis for COMET-QE results in Table 8 in Appendix E. Note that we only include annotation types that appear in at least 50 sentences, and that the 'all' column refers to the scores over all sentences and not just the ones annotated for UGC phenomena.

Scores are not directly comparable across annotation type. Performance by annotation type is consistent with previous conclusions, with GPT4 getting the highest scores across the board, and online systems and NLLB also doing well. It is striking that the systems that have higher scores in general tend to do better across the board on all annotation types, whereas the lower-scoring systems struggle with certain non-standard phenomena. They correspond in particular to phonetic distance, where a word is spelt differently according to how it is pronounced (e.g. *HEERE'Z A QWESHCHUN FER YA* 'Here's a question for you'), contractions (e.g. *wud* 'would'), devowelling (e.g. *nvr* 'never'), truncation (e.g. *intro* 'introductory') and spelling errors. These are notably phenomena that could well result in out-of-vocabulary words.

Are certain language pairs more difficult than others? It is tricky to compare across language pairs, since scores are not comparable. However, there are some indications that the en–cs set is more challenging, given the low scores across multiple annotation types for all systems other than GPT4. The fact that GPT4 has high scores for all annotation types listed shows that the lower scores of other models are not due to quality issues in the reference translations, and provides an upper bound against which other systems can be compared, thereby indicating that the systems struggled more.

5.2 Qualitative analysis

Non-standard variants of words Many of the non-standard phenomena that characterise the texts (e.g. acronyms, truncations, contractions, devowelling) represent a similar difficulty to unknown or rare tokens in MT. The treatment of these words differs according to the system used, and inevitably largely on the training data of the model. Many of the constrained systems struggle to translate such words, either copying the words into the translation or omitting them entirely. The degree to which the systems succeed in correctly translating these words appears to depend on how common it is. For example, *tho*, phonetically-inspired spelling of *though*, was translated successfully by multiple systems, although the devowelled word *tmro* 'tomorrow' proved more difficult.

Markers of expressivity It is common for UGC texts to have markers of expressivity such as capitalisation or repetition of letters. We removed these markers in our normalised versions and reference

Systems	en-cs			en-de			en-ru			en-uk		
	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ
Unconstrained												
GPT4-5shot	0.857	0.825	*0.031*	0.869	0.837	*0.032*	0.818	0.793	*0.025*	0.858	0.838	*0.021*
ONLINE-A	0.836	0.724	*0.112*	0.858	0.771	*0.087*	0.806	0.730	*0.076*	0.830	0.741	*0.090*
ONLINE-B	0.844	0.760	*0.084*	0.867	0.815	*0.052*	0.812	0.748	*0.063*	0.856	0.787	*0.069*
ONLINE-G	0.812	0.699	*0.113*	0.847	0.763	*0.084*	0.828	0.773	*0.055*	0.853	0.803	*0.050*
ONLINE-M	0.838	0.720	*0.118*	0.847	0.714	*0.133*	0.787	0.686	*0.102*	-	-	-
ONLINE-W	0.865	0.782	*0.082*	0.892	0.809	*0.083*	0.834	0.786	*0.048*	0.862	0.819	*0.043*
ONLINE-Y	0.819	0.725	*0.095*	0.862	0.795	*0.067*	0.814	0.756	*0.058*	0.823	0.750	*0.073*
NLLB_MBR	0.837	0.792	*0.045*	0.836	0.786	*0.049*	0.799	0.755	*0.045*	0.826	0.778	*0.049*
NLLB_Greedy	0.839	0.791	*0.049*	0.837	0.783	*0.054*	0.798	0.753	*0.046*	0.827	0.775	*0.052*
Lan-BridgeMT	0.820	0.723	*0.097*	0.830	0.737	*0.094*	0.784	0.699	*0.084*	0.795	0.705	*0.090*
GTCOM_Peter	0.822	0.725	*0.098*	-	-	-	-	-	-	0.807	0.714	*0.092*
PROMT	-	-	-	-	-	-	0.780	0.685	*0.095*	-	-	-
ZengHuiMT	0.811	0.717	*0.094*	0.833	0.760	*0.073*	0.772	0.706	*0.066*	0.786	0.709	*0.077*
Unconstrained												
AIRC	-	-	-	0.779	0.669	*0.110*	-	-	-	-	-	-
CUNI-Trans	0.831	0.719	*0.112*	-	-	-	-	-	-	-	-	-
CUNI-DocTrans	0.840	0.694	*0.146*	-	-	-	-	-	-	-	-	-
CUNI-GA	0.840	0.694	*0.146*	-	-	-	-	-	-	-	-	-

Table 3: COMET scores of systems on the `manseg-norm` and `manseg-raw` subsets.

Systems	en-cs			en-de			en-he			en-ja			en-ru			en-uk			en-zh		
	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ
Unconstrained																					
GPT4-5shot	0.817	0.800	*0.018*	0.822	0.805	*0.017*	0.806	0.793	*0.013*	0.846	0.838	*0.008*	0.806	0.789	*0.017*	0.809	0.797	*0.012*	0.797	0.786	*0.011*
ONLINE-A	0.807	0.724	*0.083*	0.816	0.765	*0.050*	0.807	0.737	*0.070*	0.824	0.772	*0.052*	0.807	0.750	*0.058*	0.791	0.726	*0.065*	0.786	0.725	*0.061*
ONLINE-B	0.814	0.756	*0.058*	0.821	0.793	*0.028*	0.812	0.767	*0.045*	0.848	0.822	*0.027*	0.808	0.761	*0.047*	0.805	0.759	*0.046*	0.805	0.766	*0.039*
ONLINE-G	0.791	0.705	*0.086*	0.812	0.766	*0.045*	0.786	0.720	*0.067*	0.782	0.700	*0.082*	0.821	0.784	*0.036*	0.809	0.775	*0.034*	0.765	0.704	*0.062*
ONLINE-M	0.807	0.710	*0.096*	0.810	0.724	*0.086*	-	-	-	0.798	0.711	*0.088*	0.790	0.702	*0.089*	-	-	-	0.762	0.692	*0.069*
ONLINE-W	0.822	0.765	*0.057*	0.822	0.780	*0.042*	-	-	-	0.822	0.790	*0.031*	0.819	0.786	*0.033*	0.812	0.782	*0.030*	0.802	0.767	*0.036*
ONLINE-Y	0.799	0.732	*0.067*	0.822	0.786	*0.036*	0.808	0.753	*0.056*	0.842	0.811	*0.031*	0.814	0.764	*0.050*	0.787	0.731	*0.056*	0.796	0.752	*0.044*
NLLB_MBR	0.802	0.762	*0.040*	0.801	0.763	*0.038*	0.796	0.756	*0.040*	0.721	0.682	*0.039*	0.794	0.754	*0.040*	0.784	0.744	*0.040*	0.617	0.596	*0.021*
NLLB_Greedy	0.806	0.765	*0.041*	0.802	0.761	*0.041*	0.795	0.756	*0.039*	0.749	0.711	*0.038*	0.795	0.754	*0.041*	0.786	0.745	*0.041*	0.664	0.645	*0.019*
Lan-BridgeMT	0.799	0.724	*0.075*	0.805	0.741	*0.064*	0.797	0.757	*0.040*	0.827	0.774	*0.053*	0.796	0.724	*0.071*	0.769	0.696	*0.072*	0.803	0.792	*0.011*
GTCOM_Peter	0.796	0.722	*0.074*	-	-	-	0.797	0.719	*0.077*	-	-	-	-	-	-	0.774	0.704	*0.070*	-	-	-
KYB	-	-	-	-	-	-	-	-	-	0.788	0.691	*0.097*	-	-	-	-	-	-	-	-	-
PROMT	-	-	-	-	-	-	-	-	-	-	-	-	0.789	0.710	*0.079*	-	-	-	-	-	-
Yishu	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.805	0.766	*0.039*
ZengHuiMT	0.781	0.713	*0.067*	0.792	0.748	*0.045*	0.790	0.734	*0.055*	0.828	0.791	*0.037*	0.772	0.724	*0.048*	0.748	0.696	*0.052*	0.772	0.711	*0.061*
Constrained																					
AIRC	-	-	-	0.763	0.684	*0.079*	-	-	-	0.779	0.701	*0.078*	-	-	-	-	-	-	-	-	-
ANVITA	-	-	-	-	-	-	-	-	-	0.797	0.716	*0.080*	-	-	-	-	-	-	0.630	0.536	*0.094*
CUNI-Trans	0.798	0.705	*0.093*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CUNI-DocTrans	0.803	0.677	*0.126*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CUNI-GA	0.803	0.677	*0.126*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HW-TSC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.793	0.740	*0.054*
IOL_Research	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.770	0.696	*0.074*
NAIST-NICT	-	-	-	-	-	-	-	-	-	0.830	0.764	*0.066*	-	-	-	-	-	-	-	-	-
Samsung_RP	-	-	-	-	-	-	0.797	0.732	*0.065*	-	-	-	-	-	-	-	-	-	-	-	-
SKIM	-	-	-	-	-	-	-	-	-	0.837	0.785	*0.052*	-	-	-	-	-	-	-	-	-
UvA-LTL	-	-	-	-	-	-	0.799	0.731	*0.068*	-	-	-	-	-	-	-	-	-	-	-	-

Table 4: COMET-QE scores of systems on the `manseg-norm` and `manseg-raw` subsets.

translations for consistency. However, there are no guidelines as to how the different systems should translate them: either preserving the markers in the translation or normalising as we do in the reference. What we observe is variable behaviour depending on both the system and on the specific textual instances. Table 6 shows one of the more extreme examples of character repetition. Most of the systems fail to translate the words with character repetition and instead copy them (the case of *moooorrreeee* 'more', *poollliitte* 'polite' and *Discouuurrrse* 'Discourse'). There is greater robustness for the word *neeeeed* 'need', which is translated correctly as *brauchen* by a majority of systems, but not by AIRC, ONLINE-M and ONLINE-W, which is interesting since ONLINE-W obtains very high BLEU scores for en–de. When systems do translate the words, it tends to be the standard form that is generated (i.e. without repetition). The only example here is GPT4, which translates *moooorrreeee* as *viiieeeel* 'viel', conserving the expressivity of the source sentence. What is interesting is that this behaviour is far from consistent for GPT4, with the other non-standard words in the same example not undergoing the same treatment. This could suggest that while the system succeeded in deciphering the

System	acronymisation	article_drop	capitalisation	contraction	devowelling	elongation	grammar	inflection	lex_choice	norm_punct	phonetic_distance	pronoun_drop	punct_diff	spacing	spelling_error	truncation	word_drop	all
en-cs																		
GPT4-5shot	0.785	0.823	0.811	0.815	0.848	0.810	0.816	0.817	0.802	0.809	0.798	0.833	0.819	0.821	0.824	0.795	0.810	0.825
ONLINE-A	0.606	0.663	0.696	0.591	0.599	0.665	0.655	0.721	0.649	0.715	0.567	0.712	0.702	0.698	0.619	0.624	0.676	0.724
ONLINE-B	0.664	0.726	0.733	0.665	0.667	0.717	0.719	0.761	0.695	0.751	0.629	0.756	0.746	0.750	0.698	0.697	0.728	0.760
ONLINE-G	0.591	0.644	0.682	0.572	0.581	0.658	0.636	0.657	0.625	0.680	0.529	0.693	0.677	0.653	0.588	0.608	0.653	0.699
ONLINE-M	0.601	0.661	0.689	0.616	0.586	0.659	0.676	0.693	0.633	0.720	0.574	0.700	0.693	0.695	0.636	0.627	0.671	0.720
ONLINE-W	0.682	0.730	0.754	0.700	0.708	0.748	0.704	0.771	0.702	0.772	0.663	0.763	0.766	0.754	0.717	0.726	0.727	0.782
ONLINE-Y	0.617	0.668	0.703	0.616	0.626	0.678	0.652	0.707	0.660	0.709	0.589	0.728	0.708	0.708	0.635	0.643	0.672	0.725
NLLB_MBR	0.711	0.786	0.778	0.757	0.771	0.787	0.745	0.785	0.754	0.778	0.724	0.794	0.786	0.773	0.757	0.738	0.774	0.792
NLLB_Greedy	0.718	0.770	0.775	0.750	0.781	0.759	0.750	0.789	0.745	0.773	0.721	0.793	0.785	0.766	0.760	0.747	0.765	0.791
Lan-BridgeMT	0.618	0.670	0.701	0.612	0.594	0.674	0.650	0.693	0.651	0.709	0.573	0.730	0.703	0.693	0.632	0.636	0.674	0.723
GTCOM_Peter	0.597	0.675	0.704	0.628	0.619	0.673	0.671	0.689	0.635	0.704	0.577	0.714	0.703	0.691	0.640	0.632	0.666	0.725
ZengHuiMT	0.610	0.674	0.695	0.605	0.593	0.649	0.658	0.723	0.645	0.693	0.576	0.707	0.702	0.692	0.635	0.653	0.690	0.717
CUNI-Trans	0.605	0.655	0.683	0.591	0.612	0.663	0.635	0.670	0.644	0.719	0.563	0.699	0.693	0.677	0.622	0.632	0.669	0.719
CUNI-DocTrans	0.583	0.617	0.660	0.553	0.562	0.660	0.609	0.653	0.610	0.692	0.521	0.669	0.667	0.654	0.589	0.610	0.644	0.694
CUNI-GA	0.583	0.617	0.660	0.553	0.562	0.660	0.609	0.653	0.610	0.692	0.521	0.669	0.667	0.654	0.589	0.610	0.644	0.694
en-de																		
GPT4-5shot	0.800	0.801	0.824	0.814	0.836	0.810	0.803	0.835	0.797	0.821	0.812	0.834	0.831	0.824	0.823	0.809	0.817	0.837
ONLINE-A	0.670	0.708	0.751	0.660	0.698	0.727	0.726	0.783	0.693	0.765	0.642	0.774	0.759	0.758	0.699	0.728	0.742	0.771
ONLINE-B	0.746	0.768	0.801	0.767	0.778	0.788	0.779	0.823	0.771	0.809	0.734	0.812	0.810	0.800	0.770	0.780	0.802	0.815
ONLINE-G	0.663	0.701	0.743	0.679	0.725	0.725	0.711	0.759	0.708	0.754	0.667	0.760	0.751	0.745	0.711	0.724	0.732	0.763
ONLINE-M	0.592	0.615	0.679	0.591	0.586	0.665	0.650	0.694	0.631	0.714	0.548	0.679	0.686	0.671	0.609	0.630	0.643	0.714
ONLINE-W	0.717	0.725	0.784	0.749	0.762	0.775	0.758	0.808	0.765	0.797	0.718	0.822	0.797	0.802	0.768	0.765	0.787	0.809
ONLINE-Y	0.710	0.746	0.777	0.745	0.762	0.760	0.753	0.797	0.746	0.789	0.703	0.794	0.787	0.784	0.761	0.756	0.750	0.795
NLLB_MBR	0.702	0.749	0.776	0.751	0.756	0.763	0.724	0.786	0.729	0.768	0.718	0.780	0.778	0.767	0.744	0.735	0.755	0.786
NLLB_Greedy	0.699	0.745	0.772	0.748	0.761	0.749	0.732	0.788	0.743	0.765	0.715	0.767	0.776	0.767	0.745	0.742	0.759	0.783
Lan-BridgeMT	0.624	0.652	0.722	0.627	0.629	0.701	0.678	0.726	0.656	0.725	0.602	0.716	0.720	0.705	0.637	0.673	0.678	0.737
ZengHuiMT	0.671	0.698	0.741	0.686	0.694	0.724	0.719	0.765	0.699	0.743	0.649	0.755	0.752	0.741	0.706	0.734	0.726	0.760
AIRC	0.556	0.588	0.646	0.533	0.557	0.625	0.597	0.666	0.595	0.646	0.522	0.631	0.647	0.636	0.557	0.595	0.609	0.669
en-ru																		
GPT4-5shot	0.751	0.794	0.780	0.788	0.811	0.795	0.766	0.797	0.773	0.781	0.760	0.780	0.787	0.755	0.781	0.758	0.771	0.793
ONLINE-A	0.633	0.687	0.716	0.641	0.689	0.668	0.712	0.751	0.688	0.719	0.609	0.713	0.721	0.709	0.653	0.680	0.703	0.730
ONLINE-B	0.664	0.724	0.732	0.684	0.730	0.709	0.721	0.770	0.712	0.744	0.652	0.740	0.740	0.747	0.705	0.706	0.722	0.748
ONLINE-G	0.689	0.762	0.759	0.730	0.754	0.757	0.749	0.791	0.747	0.767	0.700	0.766	0.765	0.765	0.736	0.743	0.768	0.773
ONLINE-M	0.566	0.636	0.657	0.584	0.583	0.634	0.653	0.691	0.636	0.687	0.555	0.668	0.661	0.657	0.608	0.602	0.640	0.686
ONLINE-W	0.724	0.769	0.770	0.761	0.773	0.759	0.769	0.797	0.753	0.777	0.721	0.784	0.780	0.781	0.763	0.762	0.781	0.786
ONLINE-Y	0.668	0.738	0.742	0.708	0.723	0.714	0.730	0.781	0.727	0.741	0.668	0.747	0.747	0.751	0.724	0.706	0.730	0.756
NLLB_MBR	0.679	0.748	0.743	0.731	0.749	0.763	0.721	0.763	0.718	0.738	0.690	0.737	0.746	0.732	0.717	0.703	0.733	0.755
NLLB_Greedy	0.675	0.738	0.740	0.720	0.745	0.734	0.718	0.758	0.710	0.729	0.695	0.734	0.744	0.733	0.720	0.703	0.723	0.753
Lan-BridgeMT	0.604	0.653	0.679	0.614	0.635	0.659	0.672	0.714	0.636	0.694	0.571	0.699	0.687	0.683	0.623	0.635	0.668	0.699
PROMT	0.574	0.636	0.664	0.590	0.645	0.646	0.635	0.686	0.637	0.677	0.553	0.678	0.667	0.656	0.601	0.620	0.644	0.685
ZengHuiMT	0.623	0.694	0.691	0.615	0.669	0.647	0.687	0.730	0.663	0.683	0.600	0.700	0.696	0.690	0.658	0.657	0.688	0.706
en-uk																		
GPT4-5shot	0.804	0.838	0.828	0.834	0.834	0.818	0.817	0.836	0.804	0.831	0.809	0.834	0.832	0.835	0.826	0.811	0.819	0.838
ONLINE-A	0.626	0.688	0.725	0.635	0.665	0.679	0.705	0.744	0.680	0.736	0.608	0.728	0.724	0.721	0.656	0.676	0.721	0.741
ONLINE-B	0.701	0.744	0.770	0.723	0.748	0.742	0.766	0.790	0.750	0.788	0.679	0.774	0.779	0.788	0.734	0.744	0.764	0.787
ONLINE-G	0.718	0.788	0.791	0.771	0.773	0.779	0.773	0.807	0.767	0.797	0.736	0.797	0.795	0.795	0.766	0.775	0.808	0.803
ONLINE-W	0.763	0.783	0.802	0.805	0.805	0.789	0.807	0.831	0.808	0.818	0.757	0.805	0.814	0.817	0.787	0.793	0.812	0.819
ONLINE-Y	0.647	0.712	0.737	0.703	0.684	0.706	0.706	0.754	0.685	0.742	0.656	0.749	0.738	0.734	0.710	0.691	0.713	0.750
NLLB_MBR	0.698	0.760	0.763	0.768	0.760	0.765	0.730	0.777	0.743	0.762	0.715	0.759	0.768	0.765	0.742	0.731	0.756	0.778
NLLB_Greedy	0.692	0.740	0.761	0.741	0.758	0.724	0.732	0.767	0.730	0.756	0.703	0.768	0.766	0.762	0.742	0.735	0.757	0.775
Lan-BridgeMT	0.591	0.655	0.685	0.606	0.603	0.657	0.660	0.672	0.632	0.679	0.572	0.689	0.684	0.675	0.620	0.628	0.660	0.705
GTCOM_Peter	0.591	0.670	0.702	0.613	0.645	0.671	0.677	0.706	0.671	0.707	0.580	0.693	0.695	0.685	0.637	0.636	0.658	0.714
ZengHuiMT	0.608	0.684	0.696	0.620	0.639	0.650	0.672	0.712	0.660	0.698	0.600	0.701	0.697	0.696	0.652	0.662	0.687	0.709

Table 5: COMET scores by normalisation annotation type and averaged over all RoCS-MT sentences ('all').

non-standard English texts, there is no systematic notion of generating non-standard translations; it is possible that an expressive version of *viel* was seen far more often in the training data, thus being a probable translation in this case.

Number of copied words Bearing in mind the fact that unknown source words are often either copied or removed, we apply a simple automatic analysis to judge how often systems copy source words to diagnose a potential problem with out-of-vocabulary words. We calculate how often a sentence's translation contains words that appear in both the source and system's translations (counting occurrences for each sentences and taking the average). We calculate the difference in the average number of copied words for the`manseg-raw` and `manseg-norm` subsets (i.e. how much more are words copied when systems are translating non-standard data). Results are shown in Table 9 of Appendix F. Note that a certain number of 'copied' words is normal, as certain words are expected to be identical in source and target texts (e.g. some named entities, numbers, etc.). There are similar trends across languages: the same systems tend to have relatively fewer/more copied words in raw than norm. GPT4 has the smallest difference in number of copied words between `norm` and `raw` translations. Unsurprisingly, most other systems have more copying in `raw` than in `norm`, indicating a problem handling unknown vocabulary and confirming our previous qualitative analysis.

On a similar note, we also analysed the source-translation length ratios for both `raw` and `norm` subsets. While this could be linked to copying or omit-

	norm/raw	Text
Source	norm	“politics used to be more polite. We need a return to the discourse”
Source	raw	"politics used to be mooooorrreeee poolllliitte we neeeeed a return to The Discouuurrrse"
Reference (de)	norm	„Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs.“
AIRC	norm	"Politik war früher höflicher. Wir brauchen eine Rückkehr zum Diskurs"
	raw	"Politiker waren früher mooooorrreeee poolllliitte wir neeeed eine Rückkehr zu The Discouuurrrse"
GPT4-5shot	norm	"die Politik früher höflicher war. Wir brauchen eine Rückkehr zur Diskussion"
	raw	"die Politik früher viiieeeel höflicher war, wir brauchen eine Rückkehr zur Diskussion"
Lan-BridgeMT	norm	"die Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs"
	raw	"die Politik früher mooooorrreeee poolllliitte war, dass wir eine Rückkehr zu The Discouuurrrse brauchten"
NLLB_MBR	norm	Politik früher höflicher war
	raw	"Politik war früher mooooorrreeeeeee poolllliitte, wir brauchen eine Rückkehr zu The Discouuurrrse"
ONLINE-A	norm	"Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs"
	raw	"Politik früher mooooorrreeee poolllliitte wir brauchten eine Rückkehr zu The Discouuurrrse"
ONLINE-B	norm	„die Politik früher höflicher war.“ Wir brauchen eine Rückkehr zum Diskurs“
	raw	„Politik früher mal mooooorrreeee poolllliitte war, wir brauchen eine Rückkehr zu The Discouuurrrse“
ONLINE-G	norm	“Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs
	raw	"Politik früher mooooorrreeee poolllliitte war, wir brauchen eine Rückkehr zur Discouurrrse"
ONLINE-M	norm	„Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs“
	raw	„Politik war früher mooooorrreeee poolllliitte wir neeeeed a return to The Discouurrrse“
ONLINE-W	norm	"die Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs"
	raw	"Politik früher mooooorrreeee poolllliitte we neeeeed a return to The Discouuurrrse"
ONLINE-Y	norm	„Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs“
	raw	„Politik früher mooooorrreeee poolllliitte war, wir brauchen eine Rückkehr zu The Discouuurrrse“
ZengHuiMT	norm	"Politik früher höflicher war", war früher höflicher. Wir brauchen eine Rückkehr zum Diskurs“
	raw	"Politik früher mooooorrreeee poolllliitte war, wir brauchten eine Rückkehr zu The Discouuurrrse“

Table 6: Example of character repetition linked to a mark of expressivity for en–de.

ting unknown words (as we have seen), it is more likely to be linked to overgeneration problems, linked to systems encountering text that is out-of-domain, which we occasionally observed in the system outputs. We observed that for all systems, the length ratio between `manseg-raw` translations and their source sentences was greater than those for `manseg-norm`. The effect was even greater for the texts when automatic sentence segmentation was applied (i.e. for `spacyseg-` subsets).

6 Conclusion

We have presented a new resource, RoCS-MT, a robustness challenge set for MT, designed to test MT systems on non-standard UGC. Our automatic and manual analysis show that non-standard texts are still a problem for many of the systems, including the unconstrained ones, and that certain phenomena such as phonetically inspired spellings pose a problem in particular. The comparison of COMET and COMET-QE metrics suggest that it may be possible to draw similar conclusions from automatic scoring without using references, although future work could go into more depth into analysing what is captured by the different metrics.

Limitations

The current test set is available for five from-English directions and it would be interesting to study other language directions, including those not involving English. The current version of the challenge set only contains variants for speaker gender for one of the language pairs, and we plan to add these for the other target languages in a future version.

Finally, a major limitation is one that is becoming widespread nowadays, which is that many of the systems trained and even used in research are trained on an unknown quantity of data for which the sources are unknown. Without being able to verify the fact, GPT4 and potentially some of the other systems are likely to be trained on some of the source sentences in the challenge set, and future models may even be trained on the reference translations we provide, despite it being indicated as a test set. This is a blocking factor for scientific comparison and one that goes beyond this particular resource.

Acknowledgements

This paper was funded by both authors chair positions in the PRAIRIE institute funded by the French national agency ANR as part of the “Investissements d’avenir” programme under the reference ANR-19-P3IA-0001. It was also funded by Rachel Bawden’s Emergence project, DadaNMT, funded by Sorbonne Université.

References

Jan Christian Blaise Cruz. 2023. Samsung R&D Institute Philippines at WMT 2023. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hiroyuki Deguchi, Kenji Imamura, Yuto Nishida, Yusuke Sakai, Justin Vasselli, and Taro Watanabe. 2023. NAIST-NICT WMT'23 General MT Task Submission. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

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

Ananya Gupta, Jae Takeuchi, and Bart Knijnenburg. 2023. On the real-world performance of machine translation: Exploring social media post-authors' perspectives. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 302–310, Toronto, Canada. Association for Computational Linguistics.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Josef Jon, Martin Popel, and Ondřej Bojar. 2023. CUNI-GA submission at WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Keito Kudo, Takumi Ito, Makoto Morishita, and Jun Suzuki. 2023. SKIM at WMT 2023 General Translation Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Ben Li, Yoko Matsuzaki, and Shivam Kalkar. 2023. KYB General Machine Translation Systems for WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, Sofia, Bulgaria. Association for Computational Linguistics.

Paul McNamee and Kevin Duh. 2022. The multilingual microblog translation corpus: Improving and evaluating translation of user-generated text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 910–918, Marseille, France. European Language Resources Association.

Paul Michel and Graham Neubig. 2018. 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.

Luo Min, yixin tan, and Qiulin Chen. 2023. Yishu: Yishu At WMT2023 Translation Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Alexander Molchanov and Vladislav Kovalenko. 2023. PROMT Systems for WMT23 Shared General Translation Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hamdy Mubarak, Sabit Hassan, and Ahmed Abdelali. 2020. Constructing a bilingual corpus of parallel tweets. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 14–21, Marseille, France. European Language Resources Association.

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*, pages 311–318, Philadelphia,

Pennsylvania, USA. Association for Computational Linguistics.

Martin Popel. 2020. CUNI English-Czech and English-Polish systems in WMT20: Robust document-level training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 269–273, Online. 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, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Matīss Rikters and Makoto Miwa. 2023. AIST AIRC Submissions to the WMT23 Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). 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.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2021. Understanding the impact of UGC specificities on translation quality. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 189–198, Online. Association for Computational Linguistics.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5240–5250, Marseille, France. European Language Resources Association.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a treebank of noisy user generated content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India. The COLING 2012 Organizing Committee.

Henny Sluyter-Gäthje, Pintu Lohar, Haithem Afli, and Andy Way. 2018. FooTweets: A bilingual parallel corpus of world cup tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Iñaki San Vicente, Iñaki Alegría, Cristina España-Bonet, Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martínez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. 2016. TweetMT: A parallel microblog corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2936–2941, Portorož, Slovenia. European Language Resources Association (ELRA).

Di Wu, Shaomu Tan, David Stap, Ali Araabi, and Christof Monz. 2023a. UvA-MT's Participation in the WMT 2023 General Translation Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe YU, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin GUO, Yuhao Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023b. Treating General MT Shared Task as a Multi-Domain Adaptation Problem: HW-TSC's Submission to the WMT23 General MT Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hui Zeng. 2023. Achieving State-of-the-Art Multilingual Translation Model with Minimal Data and Parameters. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Wenbo Zhang. 2023. IOL Research Machine Translation Systems for WMT23 General Machine Translation Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hao Zong. 2023. Gtcom neural machine translation systems for wmt23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

A Non-standard keywords for sourcing of posts

The full list of keywords we searched for using the Reddit API is as follows: *yyy*, *iii*, *eee*, *ppl*, *btw*, *imo*, *wtf*, *shes*, *hes*, *ima*, *shud*, *wud*, *cud*, *afaik*, *bcuz*, *hahaaa*, *dat*, *wen*, *wot*, *woz*, *bout*, *bro*, *gonna*, *lmao*, *ppl*, *smh*, *yall*, *omg*, *barley*, *fyi*, *beleive*, *seperate*, *lol*, *ttyl*, *muaha*, *mwah*, *air*, *afaik*, *fr*, *fyi*, *idk*, *ikr*, *irl*, *jk*, *nvm*, *plz*, *pls*, *cu*, *tbh*, *ur*, *wth*, *kk*, *2mo*, *2moro*, *tmrw*, *fwiw*, *nvm*, *thx*, *b4*, *ruok*, *m8*, *l8r*, *2nite*, *gr8*, *lk*, *wt*, *w/*, *peeps*, *sooo*, *verry*, *innit*, *wasnt*, *ain't*, *definately*, *yous*, *nae*, *awfull*, *freind*, *untill*, *wierd*, *aweful*, *wether* and *alot*.

The keywords were chosen as they illustrate well known non-standard phenomena, including:

- spelling errors (e.g. *wierd* 'weird', *wether* 'whether', *alot*)
- acronymisation (*nvm* 'never mind', *fyi* 'for your information)
- repetition of characters (e.g. *hahaaa*, *eee*, *sooo*)
- contractions (e.g. *cud*, *gonna*, *shud*))
- dialectisms (e.g. *ain't*, *yous*, *nae*, *innit*, *yall*)
- devowelling (e.g. *tmrw* 'tomorrow', *pls* 'please', *jk* 'joke')
- truncations, including abbrevations (e.g. *peeps* 'people', *w/* "with')
- digit phonetisation (e.g. *2nite* 'tonight', *b4* 'before', *l8r* 'later', *cu* 'see you', *ruok* 'are you ok')
- other phonetic spellings (e.g. *wot* 'what', *thx* 'thanks', *dat* 'that')
- missing whitespace (e.g. *cu* 'see you', *ruok* 'are you ok', both examples also corresponding to phonetic spellings)
- missing punctuation (e.g. *ur* (sometimes) 'you're', *wasnt* 'wasn't')
- etc.

Although the choice of keywords does create a certain bias in the types of language retrieved (especially given that several variants of some keywords are included), these keywords are used to identify posts that likely to contain other non-standard phenomena, so the final selected sentences are not restricted to those containing these keywords.

B Translation Guidelines

These guidelines are included because there are some specific constraints as to how the translations are to be carried out, and some particularities of the dataset to explain. The sentences to be translated are found in the excel spreadsheet in the column "Normalised segment". However, we also provide additional information that can help translation (see below for more information).

Origin of the text The texts to be translated are from the Reddit online forum (extracted using the API), taken from a range of different subreddits (so of different genres of text, e.g. relationship advice, advice about pets, video gaming strategy, etc.). They were selected due to their non-standard nature (spelling mistakes, abbreviations, lack of punctuation etc.).

Preprocessing of the text The texts have been manually pseudo-anonymised (usernames and names other than those representing celebrities and other well-known public figures are replaced with new names), split into "sentences" and normalised. It is the normalised versions of the sentences that are to be translated.

The sentences have been filtered to remove offensive or sensitive content (hate speech, taking drugs, suicide, etc.). However, profanities were kept as they were taken to be illustrative of the sociolect of online language. If however, you do not feel comfortable with translating something, please leave it blank and write a comment indicating that you have not translated it.

Additional context provided to help translation The text is split into short documents with one or several sentences per document. In the excel document, a sentence's document is indicated by the value in the column "Post number", and the cells are also coloured such that it is visually easier to see which sentences belong to the same document (alternating grey and white). A Reddit post is associated with a title and a text with the main content of the post. The documents can contain either the title or a subset of the text or even both. The type of text associated with each sentence is indicated in the column "Text type". Titles are marked in bold to make them visually easier to see. Although the normalised text may be sufficient to carry out the translation, we also give access to the additional information just in case:

- the title of the post
- the entire body of text associated with the post
- the raw version of the sentence (after pseudo-anonymisation and segmentation into sentences)
- some translation notes have been added to provide some context about the posts (e.g. to give an idea of what is the subject of conversation, the meaning of some expressions and abbreviations, etc. in order to make translation easier). Very occasionally there are indications about how to translate (for instance for meta-linguistic questions where people discuss particular words, it is best to keep the English words, e.g. *One word I simply can't say properly is water...* → water should be kept in English in the translation.

Constraints (important) The dataset will be used to evaluate machine translation systems on their ability to handle non-standard texts. This crucially means that: the sentence boundaries that have been defined must not be modified. It is possible to translate a sentence using several sentences if that is what is natural. However, it is not possible to merge several source sentences to produce a single translation of both (i.e. one translation per row). translators should not use machine translation systems or other computational systems to aid translation as this could bias the translations to look like translations produced by Google Translate, DeepL, ChatGPT, etc.

More specific guidelines There are multiple posts that use slang terms (e.g. gaming or general online slang such as *lol*) and it possible that the correct translation will be an English borrowing. It is fine to use an English borrowing in this case, if this is what is generally used online. The punctuation choices should be kept as much as possible, as appropriate for the target language of translation (e.g. conserving full stops, exclamation marks, quotes, etc.). As described above, there are some instances of people talking about English words, and in this case, the English words should be kept as is. Another example: *One says "Let's eat granny" making it seem like someone's going to eat their nan. However, the other example says "Lets eat, granny", implying a different meaning to the sentence.* The phrases *"Let's eat granny"* and *"Let's eat, granny"* should be kept in English. These are indicated in the translation notes.

Use of "non-standard" language:

- Any spelling mistakes that were in the raw sentence should not be reproduced in the translation (i.e. the normalised version should be used as the source sentence to translate).
- Formatting, including things like capitalisation, should (for the same reasons) follow the conventions of the normalised translation.
- Abbreviations, acronyms and simplifications (e.g. in English *wdym* = *what do you mean*, *bc* = *because*, *rly* = *really*, etc.) should be expanded, unless the result would not make a natural sentence that could realistically be found. An example of a non-natural expansion would be *lol* = *laughing out loud*, since this is not practically used.
- However, abbreviations linked to the names of places (e.g. *USA*, *UK*, *UCL* (=*University College London*) should be kept as they are if the acronym is also commonly used in the target language. In other cases, the most frequent equivalent translation should be used. (e.g. English *UN* = French *ONU*, English *NATO* = French *OTAN*).

The overall idea is that the translations should be natural and not contain the types of non-standard language that were normalised in the English versions, although they should match as best possible the style and familiarity.

Additional questions If you have any doubts or questions about the meaning of the sentences, please contact me at rachel.bawden@inria.fr to discuss things further.

C Prompt used for GPT4-5-shot

The prompt used for the GPT4-shot is the one from (Hendy et al., 2023), i.e. the following:

```
Translate this into 1. [target language]:
[shot n source]
1. [shot n reference]
Translate this into 1. [target language]:
[input]
1.
```

D BLEU scores

We provide BLEU scores for language pairs with reference translations in Table 7. The results are provided (as with the COMET scores in the main part of the paper) for the original raw subset (`manseg-raw`) and for its normalised version ((`manseg-norm`)) as well as the difference between the two scores (δ).

E COMET-QE scores by annotation type

We provide in Table 8 COMET-QE scores per annotation type for all from-English language pairs of the shared task.

F Copying analysis

Table 9 shows results for our automatic analysis of the number of source words that are found in the output translations. We calculate the number of such words, averaged over the number of sentences for each of the subsets `manseg-raw` and `manseg-norm` and we calculate the difference between the two. Positive numbers indicate that more copied words are found when systems translate the non-standard output and negative numbers indicate that more copied words are found when systems translated the normalised sentences.

Systems	en-cs			en-de			en-ru			en-uk		
	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ
Unconstrained												
GPT4-5shot	25.9	22.5	*3.4*	46.6	40.8	*5.8*	23.4	19.6	*3.9*	27.8	25.4	*2.4*
ONLINE-A	27.1	19.3	*7.8*	49.0	38.5	*10.5*	26.1	20.4	*5.8*	31.3	25.0	*6.3*
ONLINE-B	28.4	20.9	*7.5*	47.7	40.7	*7.1*	25.7	20.6	*5.1*	39.0	29.6	*9.4*
ONLINE-G	25.0	17.4	*7.6*	46.2	35.5	*10.7*	27.9	22.5	*5.4*	29.2	25.1	*4.0*
ONLINE-M	27.8	19.1	*8.7*	44.5	29.5	*15.0*	23.6	16.5	*7.1*	-	-	
ONLINE-W	30.0	22.9	*7.2*	66.0	47.1	*18.9*	29.3	23.7	*5.6*	31.5	27.4	*4.0*
ONLINE-Y	25.6	18.8	*6.7*	48.3	39.3	*9.0*	24.7	20.1	*4.5*	30.8	25.1	*5.7*
NLLB_MBR	25.5	20.8	*4.7*	41.5	34.1	*7.4*	22.3	18.2	*4.2*	26.9	22.3	*4.6*
NLLB_Greedy	25.4	20.8	*4.6*	42.0	34.0	*8.0*	22.1	18.4	*3.7*	26.2	22.0	*4.2*
Lan-BridgeMT	26.1	18.7	*7.4*	41.3	31.2	*10.1*	22.8	17.4	*5.4*	25.8	19.9	*5.9*
GTCOM_Peter	25.3	19.2	*6.2*	-	-	-	-	-	-	26.7	21.4	*5.3*
PROMT	-	-	-	-	22.6	16.4	*6.2*	-	-			
ZengHuiMT	26.1	20.1	*6.0*	46.7	39.2	*7.5*	23.5	19.6	*3.8*	27.9	23.3	*4.6*
Constrained												
AIRC	-	-	-	35.1	24.4	*10.6*	-	-	-	-		
CUNI-Trans	27.7	19.6	*8.1*	-	-	-	-	-	-			
CUNI-DocTrans	28.9	18.0	*10.9*	-	-	-	-	-	-			
CUNI-GA	28.9	18.0	*10.9*	-	-	-	-	-	-			

Table 7: BLEU scores of systems on the `manseg-norm` and `manseg-raw` subsets.

System	acronymisation	article_drop	capitalisation	contraction	devowelling	elongation	grammar	inflection	lex_choice	norm_punct	phonetic_distance	pronoun_drop	punct_diff	spacing	spelling_error	truncation	word_drop	all
en-cs																		
GPT4-5shot	0.773	0.783	0.793	0.804	0.817	0.806	0.780	0.801	0.782	0.794	0.776	0.802	0.798	0.802	0.785	0.782	0.777	0.800
ONLINE-A	0.654	0.671	0.705	0.637	0.625	0.696	0.671	0.722	0.654	0.727	0.589	0.699	0.710	0.717	0.625	0.640	0.689	0.724
ONLINE-B	0.704	0.740	0.741	0.702	0.684	0.744	0.720	0.761	0.710	0.758	0.649	0.747	0.750	0.754	0.687	0.709	0.729	0.756
ONLINE-G	0.638	0.658	0.692	0.619	0.605	0.693	0.648	0.676	0.635	0.696	0.557	0.694	0.690	0.672	0.597	0.631	0.656	0.705
ONLINE-M	0.626	0.661	0.688	0.638	0.595	0.684	0.674	0.695	0.628	0.721	0.582	0.698	0.691	0.693	0.621	0.627	0.688	0.710
ONLINE-W	0.703	0.715	0.749	0.715	0.696	0.761	0.679	0.759	0.706	0.764	0.668	0.753	0.756	0.748	0.695	0.725	0.711	0.765
ONLINE-Y	0.677	0.686	0.719	0.673	0.659	0.719	0.675	0.718	0.669	0.726	0.622	0.725	0.725	0.717	0.645	0.673	0.686	0.732
NLLB_MBR	0.706	0.750	0.753	0.749	0.744	0.760	0.720	0.759	0.719	0.761	0.692	0.768	0.761	0.750	0.706	0.721	0.745	0.762
NLLB_Greedy	0.717	0.749	0.755	0.751	0.745	0.750	0.724	0.758	0.722	0.760	0.692	0.764	0.765	0.750	0.718	0.730	0.736	0.765
Lan-BridgeMT	0.665	0.675	0.708	0.639	0.617	0.700	0.668	0.710	0.657	0.720	0.592	0.729	0.714	0.701	0.635	0.660	0.682	0.724
GTCOM_Peter	0.643	0.660	0.708	0.653	0.631	0.709	0.672	0.686	0.642	0.715	0.599	0.713	0.710	0.702	0.630	0.644	0.680	0.722
ZengHuiMT	0.650	0.663	0.699	0.638	0.612	0.686	0.656	0.716	0.660	0.700	0.590	0.695	0.706	0.702	0.625	0.665	0.682	0.713
CUNI-Trans	0.627	0.644	0.679	0.619	0.621	0.689	0.636	0.667	0.628	0.716	0.566	0.692	0.687	0.666	0.611	0.641	0.678	0.705
CUNI-DocTrans	0.602	0.604	0.647	0.571	0.571	0.683	0.604	0.653	0.599	0.692	0.529	0.648	0.656	0.655	0.583	0.619	0.633	0.677
CUNI-GA	0.602	0.604	0.647	0.571	0.571	0.683	0.604	0.653	0.599	0.692	0.529	0.648	0.656	0.655	0.583	0.619	0.633	0.677
en-de																		
GPT4-5shot	0.795	0.790	0.800	0.797	0.819	0.804	0.795	0.813	0.785	0.806	0.788	0.798	0.804	0.803	0.796	0.783	0.784	0.805
ONLINE-A	0.720	0.725	0.755	0.692	0.725	0.740	0.735	0.769	0.719	0.769	0.668	0.758	0.762	0.759	0.698	0.737	0.749	0.765
ONLINE-B	0.770	0.767	0.789	0.760	0.772	0.784	0.774	0.797	0.766	0.798	0.729	0.783	0.794	0.787	0.755	0.770	0.786	0.793
ONLINE-G	0.725	0.730	0.758	0.713	0.742	0.756	0.730	0.763	0.731	0.767	0.689	0.761	0.763	0.758	0.712	0.738	0.746	0.766
ONLINE-M	0.658	0.649	0.702	0.646	0.624	0.710	0.676	0.717	0.658	0.737	0.596	0.697	0.707	0.687	0.635	0.671	0.674	0.724
ONLINE-W	0.738	0.725	0.769	0.748	0.753	0.761	0.745	0.789	0.752	0.782	0.715	0.790	0.777	0.776	0.741	0.749	0.759	0.780
ONLINE-Y	0.747	0.752	0.780	0.761	0.759	0.777	0.758	0.796	0.762	0.791	0.720	0.784	0.786	0.779	0.748	0.760	0.766	0.786
NLLB_MBR	0.720	0.729	0.755	0.744	0.732	0.741	0.719	0.772	0.724	0.756	0.701	0.758	0.760	0.751	0.720	0.731	0.745	0.763
NLLB_Greedy	0.721	0.721	0.752	0.744	0.736	0.739	0.730	0.769	0.733	0.763	0.702	0.748	0.759	0.751	0.718	0.731	0.750	0.761
Lan-BridgeMT	0.688	0.680	0.730	0.676	0.673	0.709	0.700	0.746	0.691	0.746	0.636	0.736	0.733	0.722	0.657	0.709	0.714	0.741
ZengHuiMT	0.710	0.700	0.737	0.708	0.700	0.728	0.718	0.763	0.708	0.746	0.662	0.732	0.745	0.739	0.699	0.734	0.716	0.748
AIRC	0.627	0.610	0.664	0.582	0.615	0.645	0.633	0.690	0.622	0.681	0.560	0.662	0.672	0.665	0.582	0.631	0.654	0.684
en-he																		
GPT4-5shot	0.772	0.791	0.789	0.794	0.816	0.796	0.771	0.798	0.780	0.790	0.760	0.784	0.791	0.792	0.777	0.780	0.770	0.793
ONLINE-A	0.664	0.693	0.726	0.658	0.654	0.697	0.698	0.719	0.692	0.729	0.621	0.725	0.728	0.727	0.649	0.683	0.695	0.737
ONLINE-B	0.715	0.719	0.762	0.712	0.741	0.756	0.732	0.767	0.733	0.766	0.682	0.763	0.765	0.756	0.711	0.734	0.726	0.767
ONLINE-G	0.650	0.655	0.711	0.658	0.662	0.694	0.677	0.702	0.666	0.713	0.634	0.707	0.708	0.701	0.643	0.668	0.677	0.720
ONLINE-Y	0.698	0.701	0.745	0.702	0.695	0.733	0.710	0.754	0.701	0.750	0.660	0.750	0.748	0.738	0.683	0.711	0.725	0.753
NLLB_MBR	0.713	0.714	0.748	0.739	0.730	0.741	0.713	0.740	0.723	0.743	0.690	0.754	0.753	0.733	0.707	0.712	0.727	0.756
NLLB_Greedy	0.712	0.724	0.750	0.738	0.740	0.741	0.718	0.752	0.725	0.746	0.691	0.753	0.753	0.746	0.713	0.730	0.718	0.756
Lan-BridgeMT	0.709	0.721	0.749	0.741	0.733	0.735	0.718	0.750	0.718	0.750	0.686	0.758	0.752	0.733	0.710	0.725	0.716	0.757
GTCOM_Peter	0.651	0.677	0.705	0.639	0.643	0.673	0.676	0.720	0.669	0.719	0.601	0.702	0.708	0.701	0.627	0.657	0.670	0.719
ZengHuiMT	0.677	0.677	0.728	0.670	0.677	0.713	0.696	0.728	0.677	0.718	0.638	0.723	0.729	0.717	0.660	0.685	0.701	0.734
Samsung_RP	0.654	0.673	0.721	0.671	0.653	0.700	0.697	0.734	0.691	0.717	0.616	0.721	0.723	0.690	0.655	0.665	0.690	0.732
UvA-LTL	0.663	0.684	0.721	0.661	0.660	0.696	0.699	0.736	0.691	0.723	0.626	0.730	0.719	0.715	0.656	0.676	0.689	0.731
en-ja																		
GPT4-5shot	0.824	0.830	0.834	0.837	0.834	0.841	0.815	0.836	0.815	0.836	0.819	0.832	0.835	0.834	0.825	0.829	0.820	0.838
ONLINE-A	0.705	0.749	0.763	0.700	0.703	0.751	0.740	0.790	0.742	0.774	0.685	0.770	0.763	0.767	0.720	0.740	0.748	0.772
ONLINE-B	0.786	0.805	0.816	0.792	0.803	0.810	0.812	0.831	0.810	0.826	0.769	0.819	0.819	0.821	0.792	0.804	0.796	0.822
ONLINE-G	0.633	0.645	0.687	0.611	0.618	0.685	0.636	0.672	0.651	0.681	0.562	0.689	0.683	0.665	0.605	0.625	0.658	0.700
ONLINE-M	0.632	0.657	0.692	0.618	0.601	0.675	0.673	0.708	0.658	0.711	0.596	0.707	0.688	0.689	0.641	0.645	0.674	0.711
ONLINE-W	0.737	0.773	0.779	0.771	0.750	0.775	0.745	0.796	0.761	0.775	0.732	0.788	0.779	0.767	0.757	0.757	0.781	0.790
ONLINE-Y	0.769	0.794	0.806	0.782	0.777	0.793	0.788	0.828	0.799	0.815	0.749	0.804	0.806	0.808	0.788	0.789	0.790	0.811
NLLB_MBR	0.612	0.643	0.670	0.639	0.638	0.676	0.644	0.659	0.619	0.641	0.611	0.678	0.668	0.644	0.619	0.610	0.649	0.682
NLLB_Greedy	0.640	0.673	0.704	0.685	0.673	0.705	0.669	0.694	0.662	0.688	0.640	0.709	0.701	0.679	0.671	0.666	0.668	0.711
Lan-BridgeMT	0.721	0.730	0.766	0.712	0.705	0.747	0.742	0.763	0.737	0.777	0.669	0.779	0.765	0.760	0.710	0.746	0.738	0.774
KYB	0.613	0.634	0.666	0.616	0.593	0.651	0.621	0.710	0.659	0.679	0.572	0.702	0.673	0.670	0.629	0.659	0.670	0.691
ZengHuiMT	0.742	0.772	0.785	0.738	0.741	0.769	0.774	0.807	0.762	0.790	0.720	0.786	0.785	0.781	0.752	0.772	0.770	0.791
AIRC	0.629	0.644	0.677	0.613	0.610	0.663	0.639	0.724	0.664	0.681	0.583	0.702	0.684	0.675	0.631	0.661	0.677	0.701
ANVITA	0.647	0.681	0.707	0.648	0.623	0.664	0.667	0.736	0.671	0.706	0.605	0.734	0.701	0.699	0.644	0.673	0.696	0.716
NAIST-NICT	0.699	0.733	0.749	0.696	0.675	0.722	0.735	0.779	0.728	0.762	0.642	0.762	0.752	0.742	0.702	0.731	0.724	0.764
SKIM	0.719	0.753	0.777	0.734	0.720	0.752	0.757	0.797	0.755	0.782	0.683	0.786	0.778	0.785	0.734	0.761	0.753	0.785
en-ru																		
GPT4-5shot	0.758	0.786	0.778	0.789	0.801	0.790	0.756	0.785	0.759	0.787	0.765	0.777	0.784	0.760	0.772	0.758	0.767	0.789
ONLINE-A	0.690	0.702	0.740	0.690	0.718	0.706	0.732	0.764	0.708	0.747	0.650	0.735	0.744	0.740	0.683	0.713	0.727	0.750
ONLINE-B	0.717	0.741	0.750	0.725	0.746	0.734	0.724	0.766	0.731	0.764	0.679	0.745	0.758	0.766	0.715	0.736	0.728	0.761
ONLINE-G	0.737	0.772	0.776	0.766	0.767	0.764	0.760	0.796	0.763	0.784	0.723	0.775	0.782	0.780	0.748	0.766	0.774	0.784
ONLINE-M	0.616	0.640	0.676	0.613	0.604	0.660	0.660	0.692	0.645	0.711	0.587	0.674	0.681	0.685	0.625	0.636	0.659	0.702
ONLINE-W	0.758	0.764	0.774	0.778	0.765	0.754	0.762	0.790	0.758	0.784	0.732	0.782	0.785	0.788	0.752	0.769	0.777	0.786
ONLINE-Y	0.695	0.750	0.751	0.727	0.725	0.730	0.735	0.775	0.746	0.761	0.683	0.758	0.759	0.762	0.726	0.723	0.741	0.764
NLLB_MBR	0.702	0.736	0.743	0.734	0.727	0.758	0.713	0.745	0.716	0.745	0.687	0.745	0.751	0.742	0.706	0.716	0.734	0.754
NLLB_Greedy	0.695	0.720	0.744	0.724	0.727	0.736	0.719	0.752	0.710	0.741	0.692	0.741	0.752	0.742	0.707	0.716	0.731	0.754
Lan-BridgeMT	0.670	0.669	0.706	0.658	0.676	0.685	0.693	0.724	0.654	0.728	0.604	0.717	0.717	0.714	0.647	0.676	0.694	0.724
PROMT	0.638	0.658	0.693	0.652	0.670	0.689	0.663	0.705	0.653	0.711	0.594	0.690	0.697	0.689	0.625	0.658	0.671	0.710
ZengHuiMT	0.680	0.706	0.716	0.661	0.692	0.684	0.705	0.743	0.679	0.711	0.630	0.707	0.719	0.714	0.676	0.697	0.701	0.724
en-uk																		
GPT4-5shot	0.777	0.794	0.789	0.806	0.807	0.794	0.771	0.790	0.772	0.795	0.770	0.792	0.795	0.796	0.784	0.777	0.774	0.797
ONLINE-A	0.656	0.686	0.713	0.655	0.677	0.682	0.688	0.726	0.672	0.727	0.619	0.716	0.715	0.713	0.651	0.674	0.698	0.726
ONLINE-B	0.707	0.731	0.749	0.716	0.732	0.747	0.738	0.759	0.726	0.761	0.672	0.747	0.756	0.763	0.709	0.724	0.733	0.759
ONLINE-G	0.726	0.756	0.768	0.759	0.760	0.765	0.746	0.776	0.750	0.778	0.714	0.769	0.772	0.773	0.740	0.759	0.763	0.775
ONLINE-W	0.748	0.760	0.771	0.779	0.774	0.765	0.755	0.791	0.770	0.783	0.721	0.776	0.780	0.785	0.744	0.763	0.765	0.782
ONLINE-Y	0.666	0.705	0.720	0.695	0.671	0.699	0.695	0.737	0.686	0.729	0.646	0.730	0.724	0.722	0.685	0.689	0.691	0.731
NLLB_MBR	0.697	0.713	0.733	0.729	0.723	0.738	0.698	0.738	0.708	0.734	0.674	0.738	0.739	0.737	0.702	0.708	0.722	0.744
NLLB_Greedy	0.693	0.700	0.734	0.717	0.714	0.719	0.702	0.748	0.714	0.736	0.675	0.751	0.740	0.741	0.707	0.717	0.732	0.745
Lan-BridgeMT	0.632	0.644	0.677	0.618	0.625	0.665	0.660	0.675	0.634	0.683	0.577	0.686	0.682	0.675	0.621	0.638	0.658	0.696
GTCOM_Peter	0.626	0.665	0.691	0.633	0.654	0.676	0.670	0.694	0.657	0.706	0.596	0.685	0.690	0.686	0.637	0.647	0.664	0.704
ZengHuiMT	0.637	0.677	0.687	0.632	0.651	0.671	0.675	0.708	0.658	0.691	0.604	0.687	0.688	0.688	0.639	0.663	0.671	0.696
en-zh																		
GPT4-5shot	0.770	0.763	0.782	0.780	0.798	0.792	0.763	0.780	0.774	0.781	0.769	0.762	0.784	0.776	0.780	0.760	0.776	0.786
ONLINE-A	0.661	0.697	0.714	0.647	0.669	0.694	0.700	0.744	0.682	0.727	0.621	0.704	0.717	0.710	0.658	0.686	0.703	0.725
ONLINE-B	0.726	0.745	0.756	0.737	0.732	0.753	0.733	0.773	0.737	0.773	0.700	0.757	0.762	0.762	0.727	0.741	0.741	0.766
ONLINE-G	0.633	0.652	0.693	0.639	0.627	0.688	0.664	0.700	0.647	0.704	0.595	0.679	0.697	0.683	0.616	0.648	0.682	0.704
ONLINE-M	0.644	0.650	0.680	0.620	0.617	0.635	0.657	0.718	0.646	0.692	0.591	0.675	0.686	0.690	0.638	0.655	0.663	0.692
ONLINE-W	0.732	0.745	0.756	0.754	0.739	0.751	0.743	0.787	0.746	0.774	0.715	0.755	0.763	0.756	0.733	0.749	0.740	0.767
ONLINE-Y	0.702	0.722	0.745	0.722	0.709	0.742	0.713	0.761	0.724	0.751	0.675	0.734	0.747	0.751	0.708	0.719	0.719	0.752
NLLB_MBR	0.522	0.531	0.587	0.569	0.555	0.565	0.551	0.551	0.503	0.556	0.521	0.585	0.578	0.551	0.521	0.510	0.555	0.596
NLLB_Greedy	0.578	0.593	0.634	0.619	0.604	0.625	0.606	0.632	0.579	0.624	0.568	0.625	0.634	0.610	0.599	0.580	0.613	0.645
Lan-BridgeMT	0.779	0.784	0.785	0.789	0.796	0.802	0.765	0.789	0.765	0.791	0.780	0.779	0.790	0.786	0.781	0.771	0.774	0.792
Yishu	0.726	0.745	0.756	0.737	0.732	0.754	0.733	0.773	0.737	0.773	0.700	0.757	0.762	0.762	0.727	0.741	0.741	0.766
ZengHuiMT	0.645	0.671	0.697	0.664	0.641	0.689	0.674	0.725	0.669	0.715	0.625	0.687	0.706	0.703	0.648	0.677	0.686	0.711
ANVITA	0.509	0.505	0.510	0.519	0.488	0.536	0.544	0.554	0.498	0.531	0.470	0.515	0.524	0.526	0.488	0.500	0.486	0.536
HW-TSC	0.677	0.705	0.729	0.685	0.680	0.717	0.704	0.744	0.709	0.744	0.651	0.722	0.734	0.732	0.685	0.707	0.699	0.740
IOL_Research	0.627	0.655	0.679	0.619	0.613	0.654	0.677	0.692	0.618	0.697	0.573	0.665	0.685	0.667	0.609	0.638	0.657	0.696

Table 8: COMET-QE scores by normalisation annotation type and averaged over all RoCS-MT sentences (‘all’).

Lang. pair	en-cs	en-de	en-he	en-ja	en-ru	en-uk	en-zh
GPT4-5shot	0.14	-0.04	-0.04	-0.06	-0.07	-0.06	-0.04
NLLB_Greedy	0.08	-0.09	0.03	-0.01	-0.03	-0.01	-0.01
NLLB_MBR	0.06	-0.04	0.03	-0.00	-0.02	-0.01	-0.02
ONLINE-W	0.62	0.31	-	0.04	0.12	0.04	0.06
ONLINE-B	0.51	0.07	0.20	0.06	0.15	0.23	0.17
ONLINE-Y	0.54	0.10	0.27	0.01	0.07	0.11	0.25
Yishu	-	-	-	-	-	-	0.17
Lan-BridgeMT	0.69	0.46	0.02	0.44	0.28	0.42	-0.04
Samsung_Research_Philippines	-	-	0.32	-	-	-	-
HW-TSC	-	-	-	-	-	-	0.18
ONLINE-G	0.77	0.26	0.01	0.74	0.11	0.11	0.54
GTCOM_Peter	0.57	-	0.63	-	-	0.34	-
ONLINE-A	0.68	0.34	0.40	0.25	0.23	0.35	0.26
UvA-LTL	-	-	0.38	-	-	-	-
SKIM	-	-	-	0.33	-	-	-
ZengHuiMT	0.72	0.32	0.39	0.27	0.33	0.42	0.26
ONLINE-M	0.64	0.74	-	0.49	0.36	-	0.46
AIRC	-	0.71	-	0.56	-	-	-
PROMT	-	-	-	-	0.46	-	-
CUNI-Trans	0.88	-	-	-	-	-	-
NAIST-NICT	-	-	-	0.61	-	-	-
IOL_Research	-	-	-	-	-	-	0.54
CUNI-DocTrans	1.53	-	-	-	-	-	-
ANVITA	-	-	-	0.62	-	-	1.90
CUNI-GA	1.53	-	-	-	-	-	-
KYB	-	-	-	1.15	-	-	-

Table 9: The difference in the number of source words present in the MT output between the `manseg-raw` and `manseg-norm` subsets, averaged across all sentences for each system. This indicates how much more (or less) source words are copied in the raw (unnormalised) sentences with respect to their normalised versions.