A Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content

José Carlos Rosales Núñez[1,2,3] **Djamé Seddah**[3] **Guillaume Wisniewski**[1,2]
[1]Université Paris Sud, LIMSI
[2] Université Paris Saclay
[3] INRIA Paris
{jose.rosales,guillaume.wisniewski}@limsi.fr djame.seddah@inria.fr

Abstract

This work compares the performances achieved by Phrase-Based Statistical Machine Translation systems (PBSMT) and attention-based Neural Machine Translation systems (NMT) when translating User Generated Content (UGC), as encountered in social medias, from French to English. We show that, contrary to what could be expected, PBSMT outperforms NMT when translating non-canonical inputs. Our error analysis uncovers the specificities of UGC that are problematic for sequential NMT architectures and suggests new avenue for improving NMT models.

1 Introduction[1]

Neural Machine Translation (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014a; Cho et al., 2014) and, more specifically, attention-based models (Bahdanau et al., 2015; Jean et al., 2015; Luong et al., 2015; Mi et al., 2016) have recently become the method of choice for machine translation: many works have shown that Neural Machine Translation (NMT) outperforms classic Phrase-Based Statistical Machine Translation (PBSMT) approaches over a wide array of datasets (Bentivogli et al., 2016; Dowling et al., 2018; Koehn and Knowles, 2017). Indeed, NMT provides better generalization and accuracy capabilities (Bojar et al., 2016; Bentivogli et al., 2016; Castilho et al., 2017) even if it has well-identified limits such as over-translating and dropping translations (Mi et al., 2016; Koehn and Knowles, 2017; Le et al., 2017).

This work aims at studying how these interactions impact machine translation of noisy texts as generally found in social media and web forums and often denoted as User Generated Content (UGC). Given the increasing importance of social medias, this type of texts has been extensively studied over the years, *e.g.* (Foster, 2010; Seddah et al., 2012; Eisenstein, 2013).

In this work we focus on UGC in which no grammatical, orthographic or coherence rules are respected, other than those considered by the writer. Such rule-free environment promotes a plethora of vocabulary and grammar variations, which account for the large increase of out-of-vocabulary tokens (OOVs) in UGC corpora with respect to canonical parallel training data.

Translating UGC raises several challenges as it corresponds to both a low-resource scenario — producing parallel UGC corpora is very costly and often problematic due to inconsistencies between translators — and a domain adaptation scenario — only canonical parallel corpora are widely available to train MT systems and they must be adapted to the specificities of UGC. We therefore believe that translating UGC provides a challenging testbed to identify the limits of NMT approaches and to better understand how they are working.

Our contributions are fourfold:

- we compare the performance of PBSMT and NMT systems when translating either canonical or non-canonical corpora;
- we analyze both quantitatively and qualitatively several cases in which PBSMT translations outperform NMT on highly noisy UGC and we discuss the advantages, in terms of robustness, that PBSMT offers over NMT approaches;
- we explain how these findings highlight the limits of seq2seq (Sutskever et al., 2014b) and Transformer (Vaswani et al., 2017) NMT architectures, by studying cases in which, as opposed to the PBSMT system, the attention

[1] We thank our anonymous reviewers for their insightful comments. This work was funded by the ANR ParSiTi project (ANR-16-CE33-0021).

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 2–14
Turku, Finland, 30 September – 2 October, ©2019 Linköping University Electronic Press

mechanism fails to provide a correct translation;

- we introduce the `Cr#pbank` a new French-English parallel corpus made of UGC content built on the French Social Media Bank (Seddah et al., 2012). This corpus is much noisier than existing UGC corpora.

All our data sets are available at https://gitlab.inria.fr/seddah/parsiti.

2 Related Work

The comparison between NMT and PBSMT translation quality has been documented and revisited many times in the literature. Several works, such as (Bentivogli et al., 2016) and (Bojar et al., 2016), conclude that the former outperforms the latter as NMT translations require less post-editing to produce a correct translation. For instance, Castilho et al. (2017) present a detailed comparison of NMT and PBSMT and show that NMT outperforms PBSMT in terms of both fluency and translation accuracy, even if there is no improvement in terms of post-editing needs.

However, other case studies, such as Koehn and Knowles (2017), have defended the idea that NMT was still outperformed by PBSMT in cross-domain and low-resource scenarios. For instance, Negri et al. (2017) showed that, when translating English to French, PBSMT outperforms NMT by a great margin in multi-domain data realistic conditions (heterogeneous training sets with different sizes). Dowling et al. (2018) also demonstrated a significant gap of performance in favor of their PBSMT system's over an out-of-the-box NMT system in a low-resource setting (English-Irish). These conclusions have recently been questioned by Sennrich and Zhang (2019) who showed NMT could achieve good performance in low-resource scenario *when* all hyper-parameters (size of the byte-pair encoding (BPE) vocabulary, number of hidden units, batch size, ...) are correctly tuned and a proper NMT architecture is selected.

The situation for other NMT approaches, such as character-based NMT, is also confusing: Wu et al. (2016) have shown that character-based methods achieve state-of-the-art performance for different language pairs; Belinkov et al. (2017) and Durrani et al. (2019) have demonstrated their systems respective abilities to retrieve good amount of morphological information leveraging on subword level features. However, Belinkov and Bisk (2018) found that these approaches are not robust to noise (both synthetic and natural) when trained only with *clean* corpora. On the other hand, Durrani et al. (2019) concluded that character-based representations were more robust to synthetic and natural noise than word-based approaches. However, they did not find a substantial improvement over BPE tokenization, their BPE MT system even slightly outperforming the character-based one on 3 out of 4 of their test sets, including the one with the highest OOV rate.

Similarly to all these works, we also aim at comparing the performance of PBSMT and NMT approaches, hoping that the peculiarities of UGC will help us to better understand the pros and cons of these two methods. Our approach shares several similarity with the work of Anastasopoulos (2019) that described different experiments to determine how source-side errors can impact the translation quality of NMT models.

3 Experimental Setup

As the goal of this work is to compare the output of NMT and PBSMT when translating UGC corpora. Because of the lack of manually translated UGC, we consider a out-domain scenario in which our systems are trained on the canonical corpora generally used in MT evaluation campaigns and tested on UGC data. We will first describe the datasets used in this work (§3.1), then the different systems we have considered (§3.2) and finally the pre- and post-processing applied (§3.3).

3.1 Data Sets

Parallel corpora We train our models on two different corpora. We first consider the traditional corpus for training MT systems, namely the WMT data made of the `europarl` (v7) corpus[2] and the `newscommentaries` (v10) corpus[3]. We use the `newsdiscussdev2015` corpus as a development set. This is exactly the setup used to train the system described in (Michel and Neubig, 2018) which will be used as a baseline throughout this work.

We also consider, as a second training set, the French-English parallel portion of `OpenSubtitles'18` (Lison et al., 2018), a collection of crowd-sourced peer-reviewed subtitles for movies. We assume that, because it is made of informal dialogs, such as those found in popular *sitcoms*, sentences from `OpenSubtitles` will be much more similar to UGC data than `WMT` data,

[2] www.statmt.org/europarl/

[3] www.statmt.org/wmt15/training-parallel-nc-v10.tgz

in part because most of it originates from social media and consists in streams of conversation. It must however be noted that UGC differs significantly from subtitles in many aspects: emotion denoted with repetitions, typographical and spelling errors, emojis, etc.

To enable a fair comparison between systems trained on WMT and on OpenSubtitles, we consider a small version of the OpenSubtitles that has nearly the same number of tokens as the WMT training set and a large version that contains all OpenSubtitles parallel data.

To evaluate our system on in-domain data, we use the newstest'14 as a test set as well as 11,000 sentences extracted from OpenSubtitles.

Non-canonical UGC To evaluate our models, we consider two data sets of manually translated UGC.

The first one is a collection of French-English parallel sentences manually translated from an extension of the French Social Media Bank (Seddah et al., 2012) which contains texts collected on Facebook, Twitter, as well as from the forums of JeuxVideos.com and Doctissimo.fr.[4]

This corpus, called Cr#pbank, consists of 1,554 comments in French annotated with different kind of linguistic information: Part-of-Speech tags, surface syntactic representations, as well as a normalized form whenever necessary. Comments have been translated from French to English by a native French speaker and extremely fluent, near-native, English speaker. Typographic and grammatical error were corrected in the gold translations but the language register was kept. For instance, idiomatic expressions were mapped directly to the corresponding ones in English (e.g. "mdr" has been translated to "lol" and letter repetitions were also kept (e.g. "ouiii" has been translated to "yesss"). For our experiments, we have divided the Cr#pbank into a test set and a blind test set containing 777 comments each.

We also consider in our experiments, the MTNT corpus (Michel and Neubig, 2018), a dataset made of French sentences that were collected on Reddit and translated into English by professional translators. We used their designated test set and added a blind test set of 599 sentences we sampled from the MTNT validation set. The Cr#pbank and MTNT corpora both differ in the domain they consider, their collection date, and in the way sentences were collected to ensure they are noisy enough. We will see in Section 4 that the Cr#pbank contains much more variations and noise than the MTNT corpus.

Table 3 presents examples of UGC sentences and their translation found in these two corpora. As shown by these examples, UGC sentences contain many orthographic and grammatical errors and differ from canonical language both in their content (i.e. the topic they address and/or the vocabulary they are using) and their structure. Several statistics of these two corpora are reported in Table 1. As expected, our two UGC test sets have a substantially higher token to type ratio than the canonical test corpora, indicating a higher lexical diversity.

3.2 Machine Translation Systems

We experiment with three MT models: a traditional phrase-based approach and two neural models.

3.2.1 Phrase-based Machine Translation

We use the Moses (Koehn et al., 2007) toolkit as our phrase-based model, using the default features and parameters.

The language model is a 5-gram language model with Knesser-Ney smoothing on the target side of the parallel data. We decided to consider only the parallel data (and not any monolingual data) so that the PBSMT and NMT systems use exactly the same data.

3.2.2 seq2seq model

The first neural model we consider is a seq2seq bi-LSTM architecture with global attention decoding. The seq2seq model was trained using the XNMT toolkit (Neubig et al., 2018).[5] 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. A dropout probability of 0.3 was introduced (Srivastava et al., 2014). The model was trained using the ADAM optimizer (Kingma and Ba, 2015) with vanilla parameters ($\alpha = 0.02$, $\beta = 0.998$). Other more specific settings include keeping unchanged the learning rate (LR) for the first two epochs, a LR decay method based on the improvement of the performance on

[4] Popular French websites devoted respectively to videogames and health.

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

Corpus	#sentences	#tokens	ASL	TTR
train set				
`WMT`	2.2M	64.2M	29.7	0.20
Small	9.2M	57.7M	6.73	0.18
Large	34M	1.19B	6.86	0.25
test set				
`OpenSubTest`	11,000	66,148	6.01	0.23
`WMT`	3,003	68,155	22.70	0.23

Corpus	#sentences	#tokens	ASL	TTR
UGC test set				
`Cr#pbank`	777	13,680	17.60	0.32
`MTNT`	1,022	20,169	19.70	0.34
UGC blind test set				
`Cr#pbank`	777	12,808	16.48	0.37
`MTNT`	599	8,176	13.62	0.38

Table 1: Statistics on the French side of the corpora used in our experiments. *TTR stands for Type-to-Token Ratio, ASL for average sentence length.*

UGC Corpus		Example
MTNT	FR (src)	Je sais mais au final c'est moi que le client va supplier pour son offre et comme **Jsui** un gars cool, **jfai** au mieux.
	EN(ref)	I don't know but in the end I am the one who will have to deal with the customer begging for his offer and because **I'm** a cool guy, **I do** whatever I can to help him.
`Cr#pbank`	FR (src)	si vous me comprenez **vivé** la **mm** chose ou ***[vous]*** **avez passé le cap** je **pren tou** ce qui peu m'aider.
	EN (ref)	if you understand me **leave** the **same** thing or **have gotten over it** I **take everything** that can help me.

Table 3: Excerpts of the UGC corpora considered. Common UGC idiosyncrasies are highlighted: non-canonical contractions, spelling errors, missing elements, colloquialism, etc. See (Foster, 2010; Seddah et al., 2012; Eisenstein, 2013) for more complete linguistic descriptions.

the development set and a 0.1 label smoothing (Pereyra et al., 2017).

3.2.3 Transformer architecture

We consider a vanilla Transformer model (Vaswani et al., 2017) 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. It was trained using the ADAM optimizer with OpenNMT default parameters.

3.3 Data processing

3.3.1 Preprocessing

All of our datasets were tokenized with byte-pair encoding (BPE) (Sennrich et al., 2016) using `sentencepiece` (Kudo and Richardson, 2018). We use a BPE vocabulary size of 16K. As a point of comparison we also train a system on `Large OpenSubs` with 32K BPE operations. As usual, the training corpora were cleaned so each sentence has, at least, 1 token and, at most, 70 tokens.

We did not perform any other pre-processing. In particular, the original case of the sentences was left unchanged in order to help disambiguate subword BPE units (see example in Figure 1) especially for Named Entities that are vastly present in our two UGC corpora.

3.3.2 Post-processing : handling OOVs

Given the high number of OOVs in UGC, special care must be taken in choosing the strategy to handle them. The BPE pre-processing aims at encoding rare and unknown words as sequence of subword units reducing the number of tokens for which the model has no information. But, because of the many named-entities, contractions and unusual character repetitions, this strategy is not effective for UGC as it leads the input sentence to contain many unknown BPE tokens (that are all mapped to the special symbol `<UNK>` before translating).

The most common strategy for handling OOVs in machine translation systems is simply copying the unknown tokens from the source sentence to the translation hypothesis. This is done in the Moses toolkit (using the alignments produced during translation) and in OpenNMT (that uses the soft-alignments to copy the source token with the highest attention weight at every decoding step when necessary). At the time we conducted the MT experiments, the XNMT toolkit (Neubig et al., 2018) has no straightforward possibilities of replacing unknown tokens present in the test set.[6] For our seq2seq NMT predictions, we performed such replacement through aligning the translation hypothesis with the source sentences (both already tokenized with BPE) with fastalign (Dyer et al., 2013) and copying the source words aligned with the <UNK> token.

4 Measuring noise levels as corpus divergence

Several metrics have been proposed to quantify the domain drift between two corpora. In particular, the perplexity of a language model the KL-divergence between the character-level 3-gram distribution of the train and test sets were two useful measurements capable of estimating the noise-level of UGC corpora as shown respectively by Martínez Alonso et al. (2016) and Seddah et al. (2012).

We also propose a new metric to estimate the noise level tailored to the BPE tokenization. The *BPE stability*, BPEstab, is an indicator of how many BPE-compounded words tend to form throughout a test corpus. Formally BPEstab is defined as:

$$\frac{1}{N} \cdot \sum_{v \in \mathcal{V}} \text{freq}(v) \cdot \frac{\text{n\_unique\_neighbors}(v)}{\text{n\_neighbors}(v)} \quad (1)$$

where N is the number of tokens in the corpus, $\mathcal{V}$ the BPE vocabulary, freq(v) the frequency of the token v and n_unique_neighbors(v) the number of unique tokens that surrounds the token v. Neighbors are counted only within the original word limits. Low average BPE stability refers to a more variable BPE neighborhood, and thus, higher average vocabulary complexity.

Table 4 reports the noise-level of our test sets introduced in Section 3.1 with respect to our largest training set, Large OpenSubtitles. These measures all show how divergeent are our UGC corpora from our largest training set. As shown by its OOVs ratio and its KL-divergence score, our Cr#pbank corpus is much more noisier than the MTNT corpus, making it a more difficult target in our translation scenario.

5 Experimental Results

5.1 MT Performance

Table 5 reports the BLEU scores[7] achieved by the three systems we consider on the different combinations of train/test sets. These results show that, while NMT systems achieve the best scores on in-domain settings, their performance drops when the test set departs from the training data. On the contrary, the phrase-based system performs far better in out-domain setting than in-domain settings. It even appears that the quality of the translation of phrase-based system increases with the noise-level (as measured by the metrics introduced in §4): when trained on OpenSubtitles, its score for the Cr#pbank is surprisingly better than for in-domain data. This is not the case for neural models. In the next section we present a detailed error analysis to explain this observation.

Interestingly enough, we also notice that a MT system trained on the OpenSub corpora performed much better on UGC test sets than the system trained on the WMT collection. To further investigate whether this observation results from a badly chosen number of BPE operations, we have also trained using the Large OpenSubtitles corpus tokenized with a 32K operation BPE. We have selected these numbers of BPE operations (16K and 32K), beacause they are often used as maintream values, but this BPE parameter has been shown to have a significant impact on the MT system performance (Salesky et al., 2018; Ding et al., 2019). Thus, the number of merging BPE operations should be carefully optimized in order to garantee the best performance. However, this matter is out of the scope of our work.

Comparing both Large OpenSubtitles with BPE tokenization 16K and 32K, BLEU scores reveal that PBSMT has considerably lower performance as the vocabulary size doubles. Regarding the seq2seq NMT and, specially, PBSMT, we can notice these systems underperform for such vocabulary size, whereas the Transformer architecture shows slightly better performances. However, the Transformer still does not outperforms our best PBSMT benchmark on Cr#pbank. It is worth noting that performances of the in-domain test OpenSubsTest are kept almost invariable for PBSMT both and NMT models.As expected, these performance gaps between PBSMT and NMT models are

[6] Note that the models described in (Michel and Neubig, 2018) do not handle unknown words, its reported translation performance (Table 8 in the Appendix) would be thus underestimated if compared to our own results on the MTNT (Table 5).

[7] All BLEU scores evaluation are computed with SacreBLEU (Post, 2018).

↓ Metric / Test set →	Cr#pbank †	MTNT†	Newstest	OpenSubsTest
3-gram KL-Div	1.563	0.471	0.406	0.0060
%OOV	12.63	6.78	3.81	0.76
BPEstab	0.018	0.024	0.049	0.13
PPL	599.48	318.24	288.83	62.06

Table 4: Domain-related measure on the source side (FR), between Test sets and Large OpenSubtitles training set. Dags indicate UGC corpora.

	PBSMT				seq2seq				Transformer			
	Crap	MTNT	News	Open	Crap	MTNT	News	Open	Crap	MTNT	News	Open
WMT	20.5	21.2	22.5†	13.3	17.1	24.0	**29.1**†	16.4	15.4	21.2	27.4†	16.3
Small	28.9	27.3	20.4	26.1†	26.1	28.5	24.5	28.2†	27.5	28.3	26.7	31.4†
Large	**30.0**	**28.6**	22.3	27.4†	21.8	22.8	17.3	28.5†	26.9	28.3	26.6	31.5†
Large 32K	22.7	22.1	16.1	27.4†	25.3	27.2	21.9	28.4†	27.8	28.5	27.1	**31.9**†

Table 5: BLEU score results for our three models for the different train-test combinations. All the MT predictions have been treated to replace UNK tokens according to Section 3.3.2. The best result for each test set is marked in bold, best result for each system (row-wise) in blue color and score for in-domain test sets with a dag. 'Crap', 'MTNT', 'News' and 'Open' stand, respectively, for the Cr#pbank, MTNT, newstest'14 and OpenSubtitlesTest test sets.

substantial to out-of-domain test corpora, whereas scores on the in-domain test sets remain almost invariable regardless the chosen BPE vocabulary size.

5.2 Error Analysis

The goal of this section is to analyze both quantitatively and qualitatively the output of NMT systems to explain their poor performance in translating UGC. Several works have already identified two main limits of NMT systems: translation dropping and excessive token generation, also known as over-generation (Roturier and Bensadoun, 2011; Kaljahi et al., 2015; Kaljahi and Samad, 2015; Michel and Neubig, 2018). We will analyze in detail how these two problems impact our models in the following subsections.

It is also interesting to notice how performances lowered on the LargeOpenSubtitles system tokenized with 16K BPE operations for the seq2seq system. Specifically the newstest'14 translation results, for which we noticed a drop of 7.2 BLEU points with respect to the SmallOpenSubtitles configuration, despite having roughly 4 times more training data. This is due to a faulty behaviour of the fastalign method, directly caused by a considerable presence of UNK on the seq2seq output. Concisely, there were 829 UNK tokens on the newtest'14 prediction for the Small model and 3,717 of such tokens in the output of the Large setup. As soon as we double the number of operations on the further to train the Large 32K system, performances on all the out-of domain testsets substantially increase, having 862 UNK tokens on the newstest'14. This points to the fact that keeping the same size of BPE vocabulary while increasing the size of the trainig data several times causes to have too many UNK subword tokens on cross-domain corpora due to a small vocabulary given the size and the lexical variability of the training corpus. This is also suggested by the fact that the LargeOpenSubtitles 16K system results for the in-domain test set are the only ones with no performance loss. On the othe hand, it is important to note that the PBSMT and Transformer architecture did not showed a performance decrease for the Large model either.

Additionally, the PBSMT results for the Large 32K system are considerably lower than for any of the other 2 OpenSubtitles configurations. This shows that the PBSMT performs worse when we have 32K vocabulary size keeping the same data size, when compared to the Large system results. We hypothesize that this is caused by a loss of generalization capability due to the fact that phrase-tables are less factorized when having bigger vocabularies of whole words, rather than relatively few sub-word vocabulary elements.

5.2.1 Translation Dropping

By manually inspecting the systems outputs, we found that NMT models tend to produce shorter outputs than the translation hypotheses of our phrase-based system, often avoiding to translate the noisiest parts of the source sentence, such as in the example described in Figure 1. Sato et al. (2016) reports a similar observation.

Analyzing the attention matrices shows that this issue is often triggered by very unusual token sequences (e.g. letter repetitions that are quite frequent in UGC corpora), or when the BPE tokenization results in a subword token that can generate a translation that has a high probability according to a corpus of canonical texts. For instance, in Figure 1, a rare BPE token, part of the Named Entity "`teen wolf`" gets confused with the very common french token "`te`" *(you)*. As a consequence, the `seq2seq` model suddenly stops translating because the hypothesis "`I want to look at you`" is a very common English sentence with a much lower perplexity than the (correct) UGC translation. Similar pattern can be observed with the Transformer architecture in case of rare token sequences on the source side, such as in the third example of Table 9, causing the translation to stop abruptly.

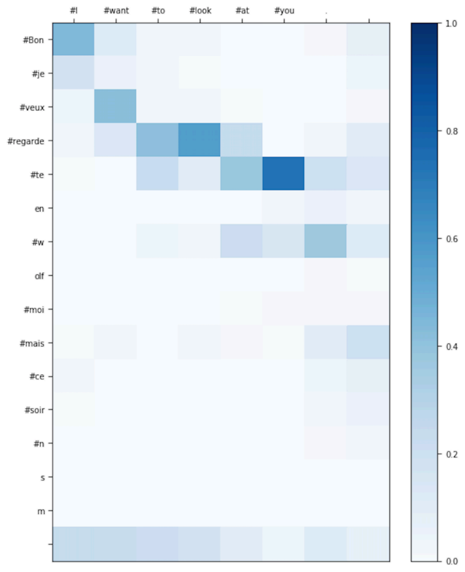

Figure 1: Attention matrix for the source sentence 'Bon je veux regardé teen wolf moi mais ce soir nsm*' predicted by a `seq2seq` model. **Ok, I do want to watch Teen Wolf tonight motherf..r*

Our phrase-based model does not suffer from this problem as there is no entry in the phrase table that matches the sequence of BPE tokens of the source sentence. This illustrates how hard alignment tables can be more efficient than soft-alignment produced by attention mechanisms for highly noisy cases, in particular when the BPE tokenization generates ambiguous tokens, which confuses the NMT model.

To quantify the translation dropping phenomenon, we show, in Figure 2, the distribution of the ratio between the reference (ground truth) translation sentence length and the one produced by PBSMT and NMT for `Cr#pbank`. This figure shows that both the NMT and Transformers models have a consistent tendency of producing shorter sentences than expected, while PBSMT does not. This is a strong evidence that NMT systems produce overall shorter translations, as has been noticed by several other authors. Moreover, there are a substantial percentage of the NMT predictions that are 60% shorter than the references, which demonstrates the presence of translations being dropped or shortened.

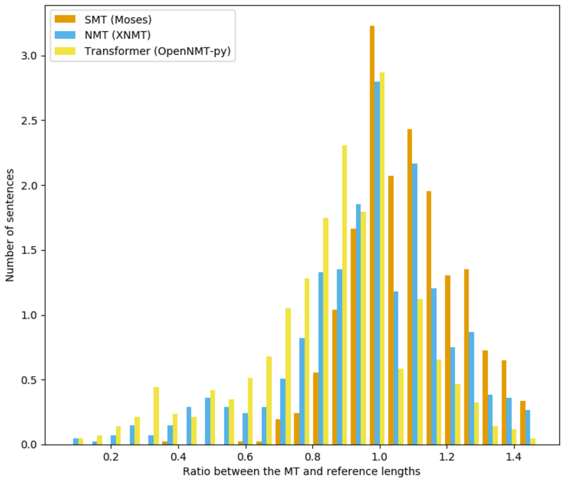

Figure 2: Distribution of `Cr#pbank` translations length ratio w.r.t ground truth translations.

5.2.2 Over-translation

A second well-known issue with NMT is that the model sometimes repeatedly outputs tokens lacking any coherence, thus adding considerable artificial noise to the output (Tu et al., 2016).

When manually inspecting the output, we noticed that this phenomenon occurred in UGC sentences that contain a rare, and often repetitive, sequence of tokens, such as those present in sentences like "`ne spoooooooilez pas teen` wolf non non non et non je dis non” *(don’t spoooooil Teen wolf no and no I say no)* in which the speaker emotion is expressed by repetitions of words or letters. The attention matrix obtained when translating such sentences with a seq2seq model often shows that the attention mechanism gets stalled due to the repetition of some BPE token (cf. the attention matrix in Figure 3 that corresponds to the example above). More generally, we noticed many cases in which the attention weights start focusing more and more on the end-of-sentence token until the translation is terminated while ignoring the source sentence tokens thereafter.

The transformer model exhibits similar problems (for instance it translates the previous example to “No no no no no no no no no no no no no no no no no no”). The PBSMT system does not suffer for this problem and arguably produces the best translation: “don't spoooooooozt Teen Wolf, no, no, no, no, I say no”.

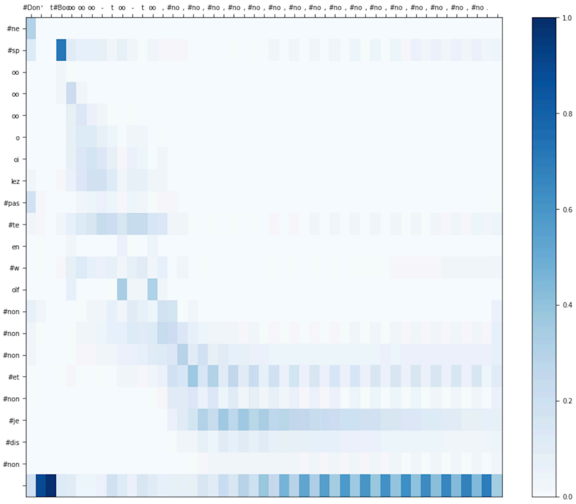

Figure 3: Attention matrix of a seq2seq model that exhibits the excessive token repetition problem. The sharp symbol (#) indicates spaces between words before the BPE tokenization.

To quantify the amount of noise artificially added by each of our models, we report, in Table 6 the Target-Source Noise Ratio (TSNR), recently introduced by Anastasopoulos (2019). A TSNR value higher than 1 indicates that the MT system adds more noise on top of the source-side noise, i.e. the rare and noisy tokens present in the source create even more noise on the output. This metric assumes that we have access to a corrected version of each source sentence. So in order to quantify this noise, we manually corrected 200 source sentences of the Cr#pbank corpus. In Table 6, we can observe that PBSMT has a better TSNR score, thus adding less artifacts (including dropped translations) to the output. We notice that the gap between PBSMT and NMT architectures (about 0.3) is much larger when training on WMT than when training in our OpenSubtitles (about 0.1).

	PBSMT	seq2seq	Transformer
WMT	4.62	5.00	4.92
Small	4.11	4.27	4.19
Large	3.99	4.27	4.09

Table 6: Noise added by the MT system estimated with the TSNR metric for the Cr#pbank corpus, the lower the better.

5.2.3 Qualitative analysis

In Table 9, in the Appendix for space reasons, we present some more MT outputs to qualitatively compare the PBSMT and NMT models. These predictions were produced using Large OpenSubtitles, trained with 16K fixed size vocabulary. From Example 9.1, we can see both NMT models exhibiting better grammatical coherence on the output. Specifically, the Transformer displays the most well-formatted and fluid translation. From Example 9.2, the seq2seq model produces several potential translations to unknown expressions (“*Vous m’avez tellement soulé*”) and translates “*soulé*” → “*soiled*”. Note that “*flappy*” is also often translated as “*happy*” throughout the Cr#pbank translations. The Transformer model produces arguably the worst results for this example because of this unknown expression (“*You’ve got me so flappy*”). Example 9.3 shows one symptomatic example of the transformer producing a shorter translation than the source and a common tendency to the seq2seq and Transformer models to basically “*crash*” when problematic cases are added (bad casing, rare word, incorrect syntax..). Finally, on Example 9.4, we can notice that neither of the NMT systems can correctly translate the upper-cased source token “*CE SOIR*” → “*TONIGHT*”, whereas PBSMT achieves to do so. It is interesting to note that the Transformer model generated a non-existent word (“*SOIRY*”) in its attempt to translate the OOV.

6 Discussion

The results presented in the previous two sections confirm the conclusions of Anastasopoulos (2019) that found a correlation between NMT performance and the level of noise in the source sentence. Note that, for computational reasons we have considered a single NMT architecture in all our experiments. However, Sennrich and Zhang (2019) have recently shown that hyper-parameters such as batch size, size of BPE vocabulary, model depth, etc., can have a large impact on translation performance especially in low-resource scenario, a conclusion that should be confirmed in cross-domain setting such as the one considered in this work.

As shown by the differential of performance in favor of the smaller training sets when used with the neural models, our results suggest that the specificities of UGC raise new challenges for NMT systems that cannot simply be solved by feeding ours models more data. Nevertheless, Koehn and Knowles (2017) highlighted 6 challenges faced by Neural Machine Translation, one of them being the lack of data for resource poordomain. This issue is strongly emphasized when it comes to UGC which does not constitute a domain on its own and which is subjected to a degree of variability only seen in the processing historical document over a large period of times (Bollmann, 2019) or in emerging dialects which can greatly varies over geographic or socio-demographic factors (transliterated Arabic dialects for example). This is why the availability of new UGC data sets is crucial and as such the release of the Cr#pbank is a welcome, small, stone in the edifice that will help evaluating machine translation architectures in near-real conditions such as blind testing.

In order to avoid common leaderboard pitfalls in such settings, we did not use the Cr#pbank's blind test set for any of our experiments, neither did we for the MTNT validation test. Nevertheless, evaluating models on unseen data is necessary, the more being the better. Therefore, in the absence of a MTNT blind test, we used a sample of its validation set, approximately matching the same average sentence length than its reference test set. In Table 7 are presented results of our best systems, based on their performance on our UGC test sets. They confirm the tendency exposed earlier: our PBSMT system is more robust to noise than our transformer-based NMT with respectively +4.4 and +11.4 BLEU points for the MTNT and Cr#pbank blind tests. For completeness, we run the seq2seq system of Michel and Neubig (2018), trained on their own data set (Europarl-v7, news-commentary-v10), without any domain-adaptation, on our blind tests. Results are on the same range than the same seq2seq model we trained on our edited data set (WMT). It would be interesting to see how their domain-adaptation technique, fine-tuning on the target domain data, which brought their system's performance to BLEU 30.29 on the MTNT test set, would fare on unseen data. As UGC domain is a constantly moving, almost protean, target, adding more data seems unsustainable on the long run. Exploring unsupervised adaptive normalization could provide a solid alternative.

System	Blind Test Sets MTNT	 Cr#pbank
Large 16K - PBSMT	**29.3**	**30.5**
Large 32K - Transformer	24.9	19.1
N&G18	19.3	13.3
N&G18 + our UNK	21.9	15.4

Table 7: BLEU score results comparison on the Cr#pbank and MTNT blind test sets. N&G18 stands for (Michel and Neubig, 2018)'s baseline system

7 Conclusions

This work evaluates the capacity of both phrase-based and NMT models to translate UGC. Our experiments show that phrase-base systems are more robust to noise than NMT systems and we provided several explanations about this*relatively* surprising fact, among which the discrepancy between BPE tokens as interpreted by the translation model at decoding time and the addition of lexical noise factors are among the most striking. We have also shown, by producing a new data set with more variability, that using more training data was not necessarily the solution for coping with UGC idiosyncrasies. The aim of this work is of course not to discourage the NMT system deployment for UGC, but to better understand what in PBSMT methods contribute to noise robustness.

In our future work, we plan to see whether theses conclusions still hold for other languages and even noisier corpora. We also plan to see whether it is possible to bypass the limitations of NMT systems we have identified by pre-processing and normalizing the input sentences.

References

Antonios Anastasopoulos. 2019. An analysis of source-side grammatical errors in NMT. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 213–223, Florence, Italy. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James R. Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 861–872.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 257–267.

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.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio, Antonio Valerio Miceli Barone, and Maria Gialama. 2017. A comparative quality evaluation of pbsmt and nmt using professional translators. In *Proceedings of MT Summit XVI, vol.1: Research Track, Nagoya, Japan, September 18-22, 2017*.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations. *CoRR*, abs/1905.10453.

Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. SMT versus NMT: preliminary comparisons for irish. In *Proceedings of the Workshop on Technologies for MT of Low Resource Languages, LoResMT@AMTA 2018, Boston, MA, USA, March 21, 2018*, pages 12–20.

Nadir Durrani, Fahim Dalvi, Hassan Sajjad, Yonatan Belinkov, and Preslav Nakov. 2019. One size does not fit all: Comparing NMT representations of different granularities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1504–1516.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369.

Jennifer Foster. 2010. "cba to check the spelling": Investigating parser performance on discussion forum posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384, Los Angeles, California. Association for Computational Linguistics.

Sébastien Jean, KyungHyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1–10.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1700–1709.

Rasoul Kaljahi, Jennifer Foster, Johann Roturier, Corentin Ribeyre, Teresa Lynn, and Joseph Le Roux. 2015. Foreebank: Syntactic analysis of customer support forums. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1341–1347.

Zadeh Kaljahi and Rasoul Samad. 2015. *The role of syntax and semantics in machine translation and quality estimation of machine-translated user-generated content*. Ph.D. thesis, Dublin City University.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71.

An Nguyen Le, Ander Martinez, Akifumi Yoshimoto, and Yuji Matsumoto. 2017. Improving sequence to sequence neural machine translation by utilizing syntactic dependency information. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 21–29.

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.*

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421.

Héctor Martínez Alonso, Djamé Seddah, and Benoît Sagot. 2016. From noisy questions to Minecraft texts: Annotation challenges in extreme syntax scenario. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 13–23, Osaka, Japan. The COLING 2016 Organizing Committee.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 955–960.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 543–553.

Matteo Negri, Marco Turchi, Marcello Federico, Nicola Bertoldi, and M. Amin Farajian. 2017. Neural vs. phrase-based machine translation in a multidomain scenario. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 280–284.

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.

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey E. Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon,* *France, April 24-26, 2017, Workshop Track Proceedings*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191.

Johann Roturier and Anthony Bensadoun. 2011. Evaluation of mt systems to translate user generated content. *Proceedings of Machine Translation Summit XIII, Xiamen, China*, pages 244–251.

Elizabeth Salesky, Andrew Runge, Alex Coda, Jan Niehues, and Graham Neubig. 2018. Optimizing segmentation granularity for neural machine translation. *CoRR*, abs/1810.08641.

Takayuki Sato, Jun Harashima, and Mamoru Komachi. 2016. Japanese-english machine translation of recipe texts. In *Proceedings of the 3rd Workshop on Asian Translation, WAT@COLING 2016, Osaka, Japan, December 2016*, pages 58–67.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 211–221.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014a. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014b. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Appendix

↓ System / Test set →	Newstest'14	Discusstest'15	MTNT†
out-of-domain set-up			
WMT-seq2seq N&G18	28.93	30.76	23.27
WMT-seq2seq (Ours)	28.70	30.00	23.00
domain adaptation set-up			
WMT-seq2seq N&G18+fine tuning	-	-	30.29

Table 8: BLEU score results comparison between our seq2seq system and thoses reported by Michel and Neubig (2018). None of the system outputs have been treated to replace UNK tokens. Dags indicate UGC corpora. N&G18 stands for (Michel and Neubig, 2018)'s system.

①	src	**Nen sans rire, j'ai bu hier soir** mais ca faisait deux semaines.
	ref	Yeah no kidding, I drank last night but it had been two weeks.
	PBSMT	**No, no, I've been drinking last night**, but it's been two weeks.
	seq2seq	**No laughing, I drank last night**, but it's been two weeks.
	Transformer	**No kidding, I drank last night**, but it's been two weeks.
②	src	**Vous m'avez tellement soulé avec votre** flappy bird j'sais pas quoi. Mais je vais le télécharger.
	ref	You annoyed me so much with your flappy bird whatever. But I'm going to download it.
	PBMST	**You're so drunk with your flappy bird** I don't know. But I'm going to download.
	seq2seq	**You have soiled me happy bird** I don't know what, but I'm going to download it.
	Transformer	**You've got me so flappy** I don't know what, but I'm gonna download it.
③	src	**Vos gueul ac vos** Zlatan
	ref	**Shut the fck up with your** Zlatan.
	PBMST	**Your scream in your** Zlatan
	seq2seq	**Your shrouds with your** Zlatan
	Transformer	Zlatan!
④	src	**CE SOIR Y A L'ÉPISODE DE** #TeenWolf OMFGGGG
	ref	**TONIGHT THERE'S THE** #TeenWolf **EPISODE** OMFGGGGG
	PBMST	**Tonight's It At The EPISODE OF** #Teen Wolf OMFGGGG
	seq2seq	Teenwolf OMFGGGGGGGGGG
	Transformer	**THIS SOIRY HAS THE** #TeenWOL OMFGGGGGGGGGG

Table 9: Examples from our noisy UGC corpus.

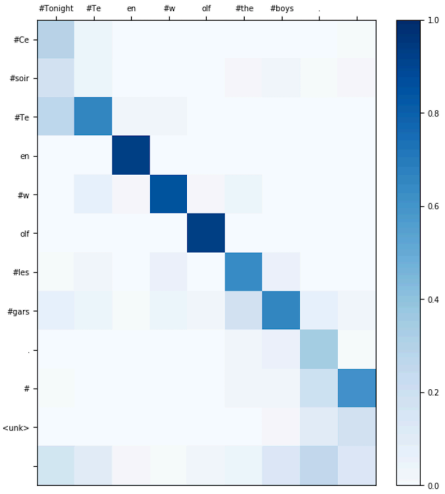

Figure 4: Attention matrix for the source sentence 'Ce soir Teen Wolf les gars.*' showing a proper translation thanks to correct casing of the named-entity BPE parts.**Tonight Teen Wolf guys.*