Machine Translation of Restaurant Reviews: New Corpus for Domain Adaptation and Robustness

Alexandre Bérard **Ioan Calapodescu** **Marc Dymetman**
Claude Roux **Jean-Luc Meunier** **Vassilina Nikoulina**
Naver Labs Europe
firstname.lastname@naverlabs.com

Abstract

We share a French-English parallel corpus of Foursquare restaurant reviews, and define a new task to encourage research on Neural Machine Translation robustness and domain adaptation, in a real-world scenario where better-quality MT would be greatly beneficial. We discuss the challenges of such user-generated content, and train good baseline models that build upon the latest techniques for MT robustness. We also perform an extensive evaluation (automatic and human) that shows significant improvements over existing online systems. Finally, we propose task-specific metrics based on sentiment analysis or translation accuracy of domain-specific polysemous words.

1 Introduction

Very detailed information about social venues such as restaurants is available from user-generated reviews in applications like Google Maps, TripAdvisor or Foursquare[1] (4SQ). Most of these reviews are written in the local language and are not directly exploitable by foreign visitors: an analysis of the 4SQ database shows that, in Paris, only 49% of the restaurants have at least one review in English, and the situation can be much worse for other cities and languages (e.g., only 1% of Seoul restaurants for a French-only speaker).

Machine Translation of such user-generated content can improve the situation and make the data available for direct display or for downstream NLP tasks (e.g., cross-lingual information retrieval, sentiment analysis, spam or fake review detection), provided its quality is sufficient.

We asked professionals to translate 11.5k French 4SQ reviews (18k sentences) to English. We believe that this resource[2] will be valuable to the community for training and evaluating MT systems addressing challenges posed by user-generated content, which we discuss in detail in this paper.

We conduct extensive experiments and combine techniques that address these challenges (e.g., factored case, noise generation, domain adaptation with tags) on top of a strong Transformer baseline. In addition to BLEU evaluation and human evaluation, we use targeted metrics that measure how well polysemous words are translated, or how well sentiments expressed in the original review can still be recovered from its translation.

2 Related work

Translating restaurant reviews written by casual customers presents several challenges for NMT, in particular robustness to non-standard language and adaptation to a specific style or domain (see Section 3.2 for details).

Concerning robustness to noisy user generated content, Michel and Neubig (2018) stress differences with traditional domain adaptation problems, and propose a typology of errors, many of which we also detected in the 4SQ data. They also released a dataset (MTNT), whose sources were selected from a social media (Reddit) on the basis of being especially noisy (see Appendix for a comparison with 4SQ). These sources were then translated by humans to produce a parallel corpus that can be used to engineer more robust NMT systems and to evaluate them. This corpus was the basis of the WMT 2019 MT Robustness Task (Li et al., 2019), in which Berard et al. (2019) ranked first. We use the same set of robustness and domain adaptation techniques, which we study more in depth and apply to our review translation task.

Sperber et al. (2017), Belinkov and Bisk (2018) and Karpukhin et al. (2019) propose to improve robustness by training models on data-augmented

[1] https://foursquare.com/

[2] https://europe.naverlabs.com/research/natural-language-processing/machine-translation-of-restaurant-reviews/

corpora, containing noisy sources obtained by random word or character deletions, insertions, substitutions or swaps. Recently, Vaibhav et al. (2019) proposed to use a similar technique along with noise generation through replacement of a clean source by one obtained by back-translation.

We employ several well-known domain adaptation techniques: back-translation of large monolingual corpora close to the domain (Sennrich et al., 2016b; Edunov et al., 2018), fine-tuning with in-domain parallel data (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Servan et al., 2016), domain tags for knowledge transfer between domains (Kobus et al., 2017; Berard et al., 2019).

Addressing the technical issues of robustness and adaptation of an NMT system is decisive for real-world deployment, but evaluation is also critical. This aspect is stressed by Levin et al. (2017) (NMT of curated hotel descriptions), who point out that automatic metrics like BLEU tend to neglect semantic differences that have a small textual footprint, but may be seriously misleading in practice, for instance by interpreting *available parking* as if it meant *free parking*. To mitigate this, we conduct additional evaluations of our models: human evaluation, translation accuracy of polysemous words, and indirect evaluation with sentiment analysis.

3 Task description

We present a new task of restaurant review translation, which combines domain adaptation and robustness challenges.

3.1 Corpus description

We sampled 11.5k French reviews from 4SQ, mostly in the *food* category,[3] split them into sentences (18k), and grouped them into train, valid and test sets (see Table 1). The French reviews contain on average 1.5 sentences and 17.9 words. Then, we hired eight professional translators to translate them to English. Two of them created the training set by post-editing (PE) the outputs of baseline NMT systems.[4] The other six translated the valid and test sets from scratch. They were asked to translate (or post-edit) the reviews sentence-by-sentence (to avoid any alignment problem), but they could see the full context. We manually filtered the test set to remove translations that were not satisfactory. The full reviews and additional metadata (e.g., location and type of the restaurant) are also available as part of this resource, to encourage research on contextual machine translation. 4SQ-HT was translated from scratch by the same translators who post-edited 4SQ-PE. While we did not use it in this work, it can be used as extra training or development data. We also release a human translation of the French-language test set (668 sentences) of the Aspect-Based Sentiment Analysis task at SemEval 2016 (Pontiki et al., 2016).

Corpus	Sentences	Reviews	Words (FR)
4SQ-PE	12 080	8 004	141 958
4SQ-HT	2 784	1 625	29 075
4SQ-valid	1 243	765	13 976
4SQ-test	1 838	1 157	21 525

Table 1: 4SQ corpora. 4SQ-PE is the training set. 4SQ-HT is not used in this work.

3.2 Challenges

Translating restaurant reviews presents two main challenges compared to common tasks in MT. First, the reviews are written in a casual style, close to spoken language. Some liberty is taken w.r.t. spelling, grammar, and punctuation. Slang is also very frequent. MT should be robust to these variations. Second, they generally are reactions, by clients of a restaurant, about its food quality, service or atmosphere, with specific words relating to these aspects or sentiments. These require some degree of domain adaptation. The following table illustrates these issues, with outputs from an online MT system. Examples of full reviews from 4SQ-PE along with metadata are shown in Appendix.

(1)	é qd g vu sa ... and when I saw that ... é qd g seen his ...	(source) (reference) (online MT)
(2)	c'est trooop bon ! it's toooo good! it's good trooop!	
(3)	le cadre est nul the setting is lousy the frame is null	
(4)	le garçon a pété un cable the waiter went crazy the boy farted a cable	
(5)	pizza nickel, tres bonnes pattes great pizza, very good pasta nickel pizza, very good legs	

[3] https://developer.foursquare.com/docs/resources/categories

[4] ConvS2S or Transformer Big trained on the "UGC" corpus described in Section 6, without domain adaptation or robustness tricks.

Examples 1 and 2 fall into the robustness category: 1 is an extreme form of SMS-like, quasi-phonetic, language (*et quand j'ai vu ça*); 2 is a literal transcription of a long-vowel phonetic stress (*trop* → *trooop*). Example 3 falls into the domain category: in a restaurant context, *cadre* typically refers to the *setting*. Examples 4 and 5 involve both robustness and domain adaptation: *pété un cable* is a non-compositional slang expression and *garçon* is not a *boy* in this domain; *nickel* is slang for *great*, *très* is missing an accent, and *pâtes* is misspelled as *pattes*, which is another French word.

Regarding robustness, we found many of the same errors listed by Michel and Neubig (2018) as noise in social media text: SMS language (*é qd g vu sa*), typos and phonetic spelling (*pattes*), repeated letters (*trooop, merciiii*), slang (*nickel, bof, mdr*), missing or wrong accents (*tres*), emoticons (':-)') and emojis (☺), missing punctuation, wrong or non-standard capitalization (lowercase proper names, capitalized words for emphasis). Regarding domain aspects, there are polysemous words with typical specific meaning *carte* → *map*, *menu*; *cadre* → *frame*, *executive*, *setting*), idiomatic expressions (*à tomber par terre* → *to die for*), and venue-related named entities (*La Boîte à Sardines*).

4 Robustness to noise

We propose solutions for dealing with non-standard case, emoticons, emojis and other issues.

4.1 Rare character placeholder

We segment our training data into subwords with BPE (Sennrich et al., 2016c), implemented in SentencePiece (Kudo and Richardson, 2018). BPE can deal with rare or unseen words by splitting them into more frequent subwords but cannot deal with unseen characters.[5] While this is not a problem in most tasks, 4SQ contains a lot of emojis, and sometimes symbols in other scripts (e.g., Arabic). Unicode now defines around 3k emojis, most of which are likely to be out-of-vocabulary.

We replace rare characters on both sides of the training corpus by a placeholder (`<x>`); a model trained on this data is typically able to copy the placeholder at the correct position. Then, at inference time, we replace the output tokens `<x>` by the rare source-side characters, in the same order. This approach is similar to that of Jean et al. (2015), who used the attention mechanism to replace output `UNK` symbols with the aligned word in the source. Berard et al. (2019) used the same technique to deal with emojis in the WMT robustness task.

[5] Unless actually doing BPE at the *byte* level, as suggested by Radford et al. (2019).

	Uppercase	Lowercase
Input	`UNE HONTE !`	`une honte !`
Pre-proc	`UN E _H ON TE _!`	`une _honte _!`
MT output	`A _H ON E Y !`	`A _dis gra ce !`
Post-proc	`A HONEY!`	`A disgrace!`

Table 2: Capital letters break NMT. BPE segmentation and translation of capitalized or lowercase input.

4.2 Capital letters

As shown in Table 2, capital letters are another source of confusion. *HONTE* and *honte* are considered as two different words. The former is out-of-vocabulary and is split very aggressively by BPE. This causes the MT model to hallucinate.

Lowercasing A solution is to lowercase the input, both at training and at test time. However, when doing so, some information may be lost (e.g., named entities, acronyms, emphasis) which may result in lower translation quality.

Factored translation Levin et al. (2017) do factored machine translation (Sennrich and Haddow, 2016; Garcia-Martinez et al., 2016) where a word and its case are split in two different features. For instance, *HONTE* becomes *honte* + *upper*.

We implement this with two embedding matrices, one for words and one for case, and represent a token as the sum of the embeddings of its factors. For the target side, we follow Garcia-Martinez et al. (2016) and have two softmax operations. We first predict the word in its lowercase form and then predict its case.[6] The embeddings of the case and word are then summed and used as input for the next decoder step.

Inline casing Berard et al. (2019) propose another approach, *inline casing*, which does not require any change in the model. We insert the case as a regular token into the sequence right after the word. Special tokens `<U>`, `<L>` and `<T>` (upper, lower and title) are used for this purpose and appended to the vocabulary. Contrary to the previous

[6] Like the "dependency model" of Garcia-Martinez et al. (2016), we use the current state of the decoder and the embedding of the output word to predict its case.

solution, there is only one embedding matrix and one softmax.

In practice, words are assumed to be lowercase by default and the <L> tokens are dropped to keep the factored sequences as short as possible. *"Best fries EVER"* becomes *"best <T> _f ries _ever <U>"*. Like Berard et al. (2019), we force SentencePiece to split mixed-case words like *MacDonalds* into single-case subwords (*Mac* and *Donalds*).

Synthetic case noise Another solution that we experiment with (see Section 6) is to inject noise on the source side of the training data by changing random source words to upper (5% chance), title (10%) or lower case (20%).

4.3 Natural noise

One way to make an NMT system more robust is to train it with some of the most common errors that can be found in the in-domain data. Like Berard et al. (2019), we detect the errors that occur naturally in the in-domain data and then apply them to our training corpus, while respecting their natural distribution. We call this "natural noise generation" in opposition to what is done in (Sperber et al., 2017; Belinkov and Bisk, 2018; Vaibhav et al., 2019) or in Section 4.2, where the noise is more synthetic.

Detecting errors We compile a general-purpose French lexicon as a transducer,[7] implemented to be traversed with extended edit distance flags, similar to Mihov and Schulz (2004). Whenever a word is not found in the lexicon (which means that it is a potential spelling mistake), we look for a French word in the lexicon within a maximum edit distance of 2, with the following set of edit operations:

(1)	deletion (e.g., *apelle* instead of *appelle*)
(2)	insertion (e.g., *appercevoir* instead of *apercevoir*)
(3)	constrained substitution on diacritics (e.g., *mangè* instead of *mangé*)
(4)	swap counted as one operation: (e.g., *mnager* instead of *manger*)
(5)	substitution (e.g., *menger* instead of *manger*)
(6)	repetitions (e.g., *Merciiii* with a threshold of max 10 repetitions)

We apply the transducer to the French monolingual Foursquare data (close to 1M sentences) to detect and count noisy variants of known French words. This step produces a dictionary mapping the correct spelling to the list of observed errors and their respective frequencies.

In addition to automatically extracted spelling errors, we extract a set of common abbreviations from (Seddah et al., 2012) and we manually identify a list of common errors in French:

(7)	Wrong verb endings (e.g., *il a manger* instead of *il a mangé*)
(8)	Wrong spacing around punctuation symbols (e.g., *Les.plats ...* instead of *Les plats...*)
(9)	Upper case/mixed case words (e.g., *manQue de place* instead of *manque de place*)
(10)	SMS language (e.g., *bcp* instead of *beaucoup*)
(11)	Phonetic spelling (e.g., *sa* instead of *ça*)

Generating errors With this dictionary, describing the real error distribution in 4SQ text, we take our large out-of-domain training corpus, and randomly replace source-side words with one of their variants (rules 1 to 6), while respecting the frequency of this variant in the real data. We also manually define regular expressions to randomly apply rules 7 to 11 (e.g., `"er "→"é "`).

We obtain a noisy parallel corpus (which we use instead of the "clean" training data), where about 30% of all source sentences have been modified, as shown below:

Error type	Examples of sentences with injected noise
(1) (6) (9)	L'Union **eU**ropéen**ne** esp**ere** que la réunion de suiv**iii** entre le Président [...]
(2) (3) (10)	Le Comité no**tte** avec **bcp** d'int**eret k** les projets d'articles [...]
(4) (7) (8)	Réun**oin sur.la** comptabilit**er** nationale [...]

5 Domain Adaptation

To adapt our models to the restaurant review domain we apply the following types of techniques: back-translation of in-domain English data, fine-tuning with small amounts of in-domain parallel data, and domain tags.

5.1 Back-translation

Back-translation (BT) is a popular technique for domain adaptation when large amounts of in-domain monolingual data are available (Sennrich et al., 2016b; Edunov et al., 2018). While our in-domain parallel corpus is small (12k pairs), Foursquare contains millions of English-language reviews. Thus, we train an NMT model[8] in the reverse direction (EN→FR) and translate all the 4SQ

[7] In Tamgu: https://github.com/naver/tamgu

[8] Like the "UGC" model with rare character handling and inline case described in Section 6.3.

English reviews to French.[9] This gives a large synthetic parallel corpus.

This *in-domain* data is concatenated to the out-of-domain parallel data and used for training.

Edunov et al. (2018) show that doing back-translation with sampling instead of beam search brings large improvements due to increased diversity. Following this work, we test several settings:

Name	Description
BT-B	Back-translation with beam search.
BT-S	Back-translation with sampling.
BT-S × 3	Three different FR samplings for each EN sentence. This brings the size of the back-translated 4SQ closer to the out-of-domain corpus.
BT	No oversampling, but we sample a new version of the corpus for each training epoch.

We use a temperature[10] of $T = \frac{1}{0.9}$ to avoid the extremely noisy output obtained with $T = 1$ and strike a balance between quality and diversity.

5.2 Fine-tuning

When small amounts of in-domain parallel data are available, fine-tuning (FT) is often the preferred solution for domain adaptation (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016). It consists in training a model on out-of-domain data, and then continuing its training for a few epochs on the in-domain data only.

5.3 Corpus tags

Kobus et al. (2017) propose a technique for multi-domain NMT, which consists in inserting a token in each source sequence specifying its domain. The system can learn the particularities of multiple domains (e.g., polysemous words that have a different meaning depending on the domain), which we can control at test time by manually setting the tag. Sennrich et al. (2016a) also use tags to control politeness in the model's output.

As our corpus (see Section 6.1) is not clearly divided into domains, we apply the same technique as Kobus et al. (2017) but use *corpus* tags (each sub-corpus has its own tag: `TED`, `Paracrawl`, etc.) which we add to each source sequence. Like in (Berard et al., 2019), the 4SQ post-edited and back-translated data also get their own tags (`PE` and `BT`).

[9] This represents ≈15M sentences. This corpus is not available publicly, but the Yelp dataset (`https://www.yelp.com/dataset`) could be used instead.

[10] with $p(w_i) = \frac{exp(z_i/T)}{\sum_{k=1}^{|V|} exp(z_k/T)}$

Corpus tag	SRC: La **carte** est trop petite.
`TED`	The map is too small.
`Multi-UN`	The card is too small.
`PE`	The **menu** is too small.

Figure 1: Example of ambiguous source sentence, where using corpus tags help the model pick a more adequate translation.

Corpus	Lines	Words (FR)	Words (EN)
WMT	29.47M	1 003M	883.5M
UGC	51.39M	1 125M	1 041M

Table 3: Size of the WMT and UGC training corpora (after filtering).

Figure 1 gives an example where using the `PE` corpus tag at test time helps the model pick a more adequate translation.

6 Experiments

6.1 Training data

After some initial work with the WMT 2014 data, we built a new training corpus named UGC (User Generated Content), closer to our domain, by combining: Multi UN, OpenSubtitles, Wikipedia, Books, Tatoeba, TED talks, ParaCrawl[11] and Gourmet[12] (See Table 3). Notably, UGC does not include Common Crawl (which contains a lot of misaligned sentences and caused hallucinations), but it includes OpenSubtitles (Lison and Tiedemann, 2016) (spoken-language, possibly closer to 4SQ). We observed an improvement of more than 1 BLEU on news-test 2014 when switching to UGC, and almost 6 BLEU on 4SQ-valid.

6.2 Pre-processing

We use `langid.py` (Lui and Baldwin, 2012) to filter sentence pairs from UGC. We also remove duplicate sentence pairs, and lines longer than 175 words or with a length ratio greater than 1.5 (see Table 3). Then we apply SentencePiece and our rare character handling strategy (Section 4.1). We use a joined BPE model of size 32k, trained on the concatenation of both sides of the corpus, and set SentencePiece's vocabulary threshold to 100. Finally, unless stated otherwise, we always use the *inline casing* approach (see Section 4.2).

[11] Corpora available at `http://opus.nlpl.eu/`

[12] 3k translations of dishes and other food terminology `http://www.gourmetpedia.eu/`

6.3 Model and settings

For all experiments, we use the Transformer Big (Vaswani et al., 2017) as implemented in Fairseq, with the hyperparameters of Ott et al. (2018). Training is done on 8 GPUs, with accumulated gradients over 10 batches (Ott et al., 2018), and a max batch size of 3500 tokens (per GPU). We train for 20 epochs, while saving a checkpoint every 2500 updates ($\approx \frac{2}{5}$ epoch on UGC) and average the 5 best checkpoints according to their perplexity on a validation set (a held-out subset of UGC).

For fine-tuning, we use a fixed learning rate, and a total batch size of 3500 tokens (training on a single GPU without delayed updates). To avoid overfitting on 4SQ-PE, we do early stopping according to perplexity on 4SQ-valid.[13] For each fine-tuned model we test all 16 combinations of dropout in $\{0.1, 0.2, 0.3, 0.4\}$ and learning rate in $\{1, 2, 5, 10\} \times 10^{-5}$. We keep the model with the best perplexity on 4SQ-valid.[14]

6.4 Evaluation methodology

During our work, we used BLEU (Papineni et al., 2002) on *news-valid* (concatenation of news-test 2012 and 2013) to ensure that our models stayed good on a more general domain, and on *4SQ-valid* to measure performance on the 4SQ domain.

For sake of brevity, we only give the final BLEU scores on *news-test 2014* and *4SQ-test*. Scores on 4SQ-valid, and MTNT-test (for comparison with Michel and Neubig, 2018; Berard et al., 2019) are given in Appendix. We evaluate "detokenized" MT outputs[15] against raw (non-tokenized) references using SacreBLEU (Post, 2018).[16]

In addition to BLEU, we do an indirect evaluation on an Aspect-Based Sentiment Analysis (ABSA) task, a human evaluation, and a task-related evaluation based on polysemous words.

6.5 BLEU evaluation

Capital letters Table 4 compares the case handling techniques presented in Section 4.2. To better evaluate the robustness of our models to changes of case, we built 3 synthetic test sets from 4SQ-test, with the same target, but all source words in upper, lower or title case.

[13]The best perplexity was achieved after 1 to 3 epochs.

[14]The best dropout rate was always 0.1, and the best learning rate was either 2×10^{-5} or 5×10^{-5}.

[15]Outputs of our models are provided with the 4SQ corpus.

[16]SacreBLEU signature: BLEU+case.mixed+numrefs.1 +smooth.exp+tok.13a+version.1.2.10

Model	BLEU	Case insensitive BLEU		
	4SQ	Upper	Lower	Title
Cased	31.78	16.02	32.42	26.67
LC to cased	30.91	**33.09**	33.09	33.09
Factored case	31.62	32.31	32.96	29.86
Inline case	31.55	31.08	32.63	29.61
Noised case	**31.99**	32.64	**33.73**	**33.63**

Table 4: Robustness to capital letters (see Section 4.2). 4SQ's source side has been set to upper, lower or title case. The first column is case sensitive BLEU. "LC to cased" always gets the same scores because it is invariant to source case.

Model	news	noised news	4SQ
UGC (Inline case)	**40.68**	35.59	31.55
+ natural noise	40.43	**40.35**	**31.69**

Table 5: Baseline model with or without natural noise (see Section 4.3). *Noised news* is the same type of noise, artificially applied to news-test.

Inline and factored case perform equally well, significantly better than the default (cased) model, especially on all-uppercase inputs. Lowercasing the source is a good option, but gives a slightly lower score on regular 4SQ-test.[17] Finally, synthetic case noise added to the source gives surprisingly good results. It could also be combined with factored or inline case.

Natural noise Table 5 compares the baseline "inline case" model with the same model augmented with natural noise (Section 4.3). Performance is the same on 4SQ-test, but significantly better on news-test artificially augmented with 4SQ-like noise.

Domain adaptation Table 6 shows the results of the back-translation (BT) techniques. Surprisingly, BT with beam search (BT-B) deteriorates BLEU scores on 4SQ-test, while BT with sampling gives a consistent improvement. BLEU scores on news-test are not significantly impacted, suggesting that BT can be used for domain adaptation without hurting quality on other domains.

Table 7 compares the domain adaptation techniques presented in Section 5. We observe that:

1. Concatenating the small 4SQ-PE corpus to the 50M general domain corpus does not help much, unless using tags.

[17]The "LC to cased" and "Noised case" models are not able to preserve capital letters for emphasis (as in Table 2), and the "Cased" model often breaks on such examples.

Model	news	4SQ
UGC (Inline case)	40.68	31.55
UGC ⊕ BT-B	40.56	30.17
UGC ⊕ BT-S	40.64	32.64
UGC ⊕ BT	**40.84**	32.69
UGC ⊕ BT-S × 3	40.63	**32.84**

Table 6: Comparison of different back-translation schemes (see Section 5.1). ⊕ denotes the concatenation of several training corpora.

Model	Tag	news	4SQ
UGC (Inline case)	–	40.68	31.55
UGC ⊕ 4SQ-PE	–	40.80	32.05
UGC + FT	–	39.78	**35.02**
UGC ⊕ 4SQ-PE + tags	–	40.71	32.12
	PE	38.97	34.36
UGC ⊕ BT + tags	–	40.67	33.47
	BT	39.02	33.00

Table 7: Domain adaptation with 4SQ-PE fine-tuning (FT) or corpus tags. The "tag" column represents the corpus tag used at test time (if any).

2. *4SQ-PE + tags* is not as good as fine-tuning with 4SQ-PE. However, fine-tuned models get slightly worse results on news.
3. Back-translation combined with tags gives a large boost.[18] The BT tag should not be used at test time, as it degrades results.
4. Surprisingly, using no tag at test time works fine, even though all training sentences had tags.[19]

As shown in Table 8, these techniques can be combined to achieve the best results. The natural noise does not have a significant effect on BLEU scores. Back-translation combined with fine-tuning gives the best performance on 4SQ (+4.5 BLEU vs UGC). However, using tags instead of fine-tuning strikes a better balance between general domain and in-domain performance.

6.6 Targeted evaluation

In this section we propose two metrics that target specific aspects of translation adequacy: translation accuracy of domain-specific polysemous words and Aspect-Based Sentiment Analysis performance on MT outputs.

[18] Caswell et al. (2019); Berard et al. (2019) observed the same thing.

[19] We tried keeping a small percentage of UGC with no tag, or with an ANY tag, but this made no difference.

Model	news	4SQ
WMT	39.37	26.26
UGC (Inline case)	40.68	31.55
Google Translate (Feb 2019)	36.31	29.63
DeepL (Feb 2019)	?	32.82
UGC ⊕ BT + FT	39.55	35.95
UGC ⊕ BT ⊕ PE + tags	**40.99**	35.72
Nat noise ⊕ BT + FT	39.91	**36.35**
Nat noise ⊕ BT ⊕ PE + tags	40.72	35.60

Table 8: Combination of several robustness or domain adaptation techniques. At test time, we don't use any tag on news, and use the PE tag on 4SQ (when applicable). BT: back-translation. PE: 4SQ-PE. FT: fine-tuning with 4SQ-PE. ⊕: concatenation.

French word	Meanings
Cadre	setting, frame, executive
Cuisine	food, kitchen
Carte	menu, card, map

Table 9: French polysemous words found in 4SQ, and translation candidates in English. The most frequent meanings in 4SQ are underlined.

Translation of polysemous words We propose to count polysemous words specific to our domain, similarly to (Lala and Specia, 2018), to measure the degree of domain adaptation. TER between the translation hypotheses and the post-edited references in 4SQ-PE reveals the most common substitutions (e.g., "card" is often replaced with "menu", suggesting that "card" is a common mistranslation of the polysemous word "carte"). We filter this list manually to only keep words that are polysemous and that have a high frequency in the test set. Table 9 gives the 3 most frequent ones.[20]

Table 10 shows the accuracy of our models when translating these words. We see that the domain-adapted model is better at translating domain-specific polysemous words.

Indirect evaluation with sentiment analysis We also measure adequacy by how well the translation preserves the polarity of the sentence regarding various aspects. To evaluate this, we perform an indirect evaluation on the SemEval 2016 Aspect-Based Sentiment Analysis (ABSA) task (Pontiki et al., 2016). We use our internal ABSA systems trained on English or French SemEval

[20] Rarer ones are: *adresse* (place, address), *café* (coffee, café), *entrée* (starter, entrance), *formule* (menu, formula), *long* (slow, long), *moyen* (average, medium), *correct* (decent, right), *brasserie* (brasserie, brewery) and *coin* (local, corner).

Model	cadre	cuisine	carte	Total
Total (source)	23	32	29	100%
WMT	13	17	14	52%
UGC (Inline case)	22	27	18	80%
UGC $\oplus$ PE + tags	**23**	31	**29**	99%

Table 10: Number of correct translations for difficult polysemous words in 4SQ-test by different models. The first row is the number of source sentences that contain this word. Other domain-adapted models (e.g., "UGC + FT" or "UGC $\oplus$ BT") also get $\approx$ 99% accuracy.

ABSA Model	Aspect	Polarity
ABSA French	**64.7**	**83.2**
ABSA English	59.5	72.1
ABSA English on MT outputs		
WMT	54.5	66.1
UGC (Inline case)	58.1	70.7
UGC $\oplus$ BT $\oplus$ PE + tags	60.2	72.0
Nat noise $\oplus$ BT $\oplus$ PE + tags	60.8	73.3

Table 11: Indirect evaluation with Aspect-Based Sentiment Analysis (accuracy in %). *ABSA French*: ABSA model trained on French data and applied to the SemEval 2016 French test set; *ABSA English*: trained on English data and applied to human translations of the test set; *ABSA English* on MT outputs: applied to MT outputs instead of human translations.

2016 data. The evaluation is done on the SemEval 2016 French test set: either the original version (ABSA French), or its translation (ABSA English). As shown in Table 11, translations obtained with domain-adapted models lead to significantly better scores on the ABSA task than the generic models.

6.7 Human Evaluation

We conduct a human evaluation to confirm the observations with BLEU and to overcome some of the limitations of this metric.

We select 4 MT models for evaluation (see Table 12) and show their 4 outputs at once, sentence-by-sentence, to human judges, who are asked to rank them given the French source sentence in context (with the full review). For each pair of models, we count the number of wins, ties and losses, and apply the Wilcoxon signed-rank test.

We took the first 300 test sentences to create 6 tasks of 50 sentences each. Then we asked bilingual colleagues to rank the output of 4 models by their translation quality. They were asked to do one or more of these tasks. The judge did not know about the list of models, nor the model that produced any given translation. We got 12 answers. The inter-judge Kappa coefficient ranged from 0.29 to 0.63, with an average of 0.47, which is a good value given the difficulty of the task. Table 12 gives the results of the evaluation, which confirm our observations with BLEU.

Pairs	Win	Tie	Loss
Tags $\approx$ Tags + noise	82	453	63
Tags $\gg$ Baseline	187	337	74
Tags $\gg$ GT	226	302	70
Tags + noise $\gg$ Baseline	178	232	97
Tags + noise $\gg$ GT	218	315	65
Baseline $\gg$ GT	173	302	123

Table 12: In-house human evaluation ("$\gg$" means better with $p \leq 0.05$). The 4 models *Baseline*, *GT*, *Tags* and *Tags + noise* correspond respectively to rows 2 (UGC with inline case), 3 (Google Translate), 6 (Combination of BT, PE and tags) and 8 (Same as 6 with natural noise) in Table 8.

We also did a larger-scale monolingual evaluation using Amazon Mechanical Turk (see Appendix), which lead to similar conclusions.

7 Conclusion

We presented a new parallel corpus of user reviews of restaurants, which we think will be valuable to the community. We proposed combinations of multiple techniques for robustness and domain adaptation, which address particular challenges of this new task. We also performed an extensive evaluation to measure the improvements brought by these techniques.

According to BLEU, the best single technique for domain adaptation is fine-tuning. Corpus tags also achieve good results, without degrading performance on a general domain. Back-translation helps, but only with sampling or tags. The robustness techniques (natural noise, factored case, rare character placeholder) do not improve BLEU.

While our models are promising, they still show serious errors when applied to user-generated content: missing negations, hallucinations, unrecognized named entities, insensitivity to context.[21] This suggests that this task is far from solved.

We hope that this corpus, our natural noise dictionary, model outputs and human rankings will help better understand and address these problems. We also plan to investigate these problems on lower resource languages, where we expect the task to be even harder.

[21] See additional examples in Appendix.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation. In *ICLR*.

Alexandre Berard, Calapodescu Ioan, and Claude Roux. 2019. NAVER LABS Europe's Systems for the WMT19 Machine Translation Robustness Task. In *WMT*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged Back-Translation. In *WMT*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *EMNLP*.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast Domain Adaptation for Neural Machine Translation. *arXiv*.

Mercedes Garcia-Martinez, Loic Barrault, and Fethi Bougares. 2016. Factored Neural Machine Translation. *arXiv*.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On Using Very Large Target Vocabulary for Neural Machine Translation. *NAACL-HLT*.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on Synthetic Noise Improves Robustness to Natural Noise in Machine Translation. *arXiv*.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain Control for Neural Machine Translation. In *RANLP*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *EMNLP*.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *LREC*.

Pavel Levin, Nishikant Dhanuka, Talaat Khalil, Fedor Kovalev, and Maxim Khalilov. 2017. Toward a full-scale neural machine translation in production: the Booking.com use case. In *MT Summit XVI*.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir K. Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan M. Pino, and Hassan Sajjad. 2019. Findings of the First Shared Task on Machine Translation Robustness. In *WMT*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *LREC*.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*, ACL.

Minh-Thang Luong and Christopher D. Manning. 2015. Stanford Neural Machine Translation Systems for Spoken Language Domain. In *IWSLT*.

Paul Michel and Graham Neubig. 2018. MTNT: A Testbed for Machine Translation of Noisy Text. In *EMNLP*.

Stoyan Mihov and Klaus U. Schulz. 2004. Fast Approximate Search in Large Dictionaries. *Computational Linguistics*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling Neural Machine Translation. In *WMT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wj Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud Maria Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *WMT*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. Building a treebank of noisy user-generated content: The French Social Media Bank. In *The 11th International Workshop on Treebanks and Linguistic Theories (TLT)*.

Rico Sennrich and Barry Haddow. 2016. Linguistic Input Features Improve Neural Machine Translation. In *WMT*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling Politeness in Neural Machine Translation via Side Constraints. In *NAACL-HLT*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving Neural Machine Translation Models with Monolingual Data. In *ACL*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.

Christophe Servan, Josep Crego, and Jean Senellart. 2016. Domain specialization: a post-training domain adaptation for neural machine translation. *arXiv*.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward Robust Neural Machine Translation for Noisy Input Sequences. In *IWSLT*.

Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving Robustness of Machine Translation with Synthetic Noise. In *NAACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *NIPS*.

Appendix

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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