MTNT: A Testbed for Machine Translation of Noisy Text

Paul Michel and **Graham Neubig**
Language Technologies Institute
Carnegie Mellon University
{pmichel1,gneubig}@cs.cmu.edu

Abstract

Noisy or non-standard input text can cause disastrous mistranslations in most modern Machine Translation (MT) systems, and there has been growing research interest in creating noise-robust MT systems. However, as of yet there are no publicly available parallel corpora of with naturally occurring noisy inputs and translations, and thus previous work has resorted to evaluating on synthetically created datasets. In this paper, we propose a benchmark dataset for Machine Translation of Noisy Text (MTNT), consisting of noisy comments on Reddit[1] and professionally sourced translations. We commissioned translations of English comments into French and Japanese, as well as French and Japanese comments into English, on the order of 7k-37k sentences per language pair. We qualitatively and quantitatively examine the types of noise included in this dataset, then demonstrate that existing MT models fail badly on a number of noise-related phenomena, even after performing adaptation on a small training set of in-domain data. This indicates that this dataset can provide an attractive testbed for methods tailored to handling noisy text in MT.[2]

1 Introduction

#nlproc is actualy f*ing hARD tbh 😂

This handcrafted sentence showcases several types of noise that are commonly seen on social media: abbreviations ("#nlproc"), typographical errors ("actualy"), obfuscated profanities ("f*ing"), inconsistent capitalization ("hARD"), Internet slang ("tbh" for "to be honest") and emojis (😂). Although machine translation has achieved significant quality improvements over the past few years due to the advent of Neural Machine Translation (NMT) (Kalchbrenner and Blunsom; Sutskever et al., 2014; Bahdanau et al., 2014; Wu et al., 2016), systems are still not robust to noisy input like this (Belinkov and Bisk, 2018; Khayrallah and Koehn). For example, Google Translate[3] translates the above example into French as:

#nlproc est en train de f * ing dur hb

which translates back into English as "#nlproc is in the process of [f * ing] hard hb". This shows that noisy input can lead to erroneous translations that can be misinterpreted or even offensive.

Noise in social media text is a known issue that has been investigated in a variety of previous work (Eisenstein; Baldwin et al.). Most recently, Belinkov and Bisk (2018) have focused on the difficulties that character based NMT models have translating text with character level noise within individual words (from scrambling to simulated human errors such as typos or spelling/conjugation errors). This is a good first step towards noise-robust NMT systems, but as we demonstrate in §2, word-by-word replacement or scrambling of characters doesn't cover all the idiosyncrasies of language on the Internet.

At this point, despite the obvious utility of creating noise-robust MT systems, and the scientific challenges contained therein, there is currently a bottleneck in that there is no standard open benchmark for researchers and developers of MT systems to test the robustness of their models to these and other phenomena found in noisy text on the Internet. In this work, we introduce MTNT, a new, realistic dataset aimed at testing robustness of MT systems to these phenomena. The dataset contains naturally created noisy source

[1] www.reddit.com

[2] The data is publicly available at http://www.cs.cmu.edu/~pmichel1/mtnt/.

[3] translate.google.com as of May 2018

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 543–553
Brussels, Belgium, October 31 - November 4, 2018.

sentences with professionally sourced translations both in a pair of typologically close languages (English and French) and distant languages (English and Japanese). We collect noisy comments from the Reddit online discussion website (§3) in English, French and Japanese, and ask professional translators to translate to and from English, resulting in approximately 1000 test samples and from 6k to 36k training samples in four language pairs (English-French (`en-fr`), French-English (`fr-en`), English-Japanese (`en-ja`) and Japanese-English (`ja-en`)). In addition, we release additional small monolingual corpora in those 3 languages to both provide data for semi-supervised adaptation approaches as well as noisy Language Modeling (LM) experiments. We test standard translation models (§5) and language models (§6) on our data to understand their failure cases and to provide baselines for future work.

2 Noise and Input Variations in Language on the Internet

2.1 Examples from Social Media Text

The term "noise" can encompass a variety of phenomena in natural language, with variations across languages (*e.g.* what is a typo in logographic writing systems?) and type of content (Baldwin et al.). To give the reader an idea of the challenges posed to MT and Natural Language Processing (NLP) systems operating on this kind of text, we provide a non-exhaustive list of types of noise and more generally input variations that deviate from standard MT training data we've encountered in Reddit comments:

- **Spelling/typographical errors**: "across" → "accross", "receive" → "recieve", "could have" → "could of", "temps" → "tant", "除く" → "覗く"
- **Word omission/insertion/repetition**: "je n'aime pas" → "j'aime pas","je pense" → "moi je pense"
- **Grammatical errors**: "a ton of" → "a tons of", "There are fewer people" → "There are less people"
- **Spoken language**: "want to" → "wanna", "I am" → "I'm", "je ne sais pas" → "chais pas", "何を笑っているの" → "何わろてんねん",
- **Internet slang**: "to be honest" → "tbh", "shaking my head" → "smh", "mort de rire" → "mdr", "笑" → "w"/"草"
- **Proper nouns** (with or without correct capitalization): "Reddit"→ "reddit"
- **Dialects**: African American Vernacular English, Scottish, Provençal, Québécois, Kansai, Tohoku...
- **Code switching**: "This is so cute" → "This is so kawaii", "C'est trop conventionel" → "C'est trop mainstream", "現在捏造中..." → "Now 捏造ing..."
- **Jargon**: on Reddit: "upvote", "downvote", "sub", "gild"
- **Emojis and other unicode characters**: ❤️,😀,😂,😍,😡, 😘, 😭
- **Profanities/slurs** (sometimes masked) "f*ck", "m*rde" ...

2.2 Is Translating Noisy Text just another Adaptation Problem?

To a certain extent, translating noisy text is a type of *adaptation*, which has been studied extensively in the context of both Statistical Machine Translation (SMT) and NMT (Axelrod et al.; Li et al.; Luong and Manning, 2015; Chu et al.; Miceli Barone et al.; Wang et al.; Michel and Neubig, 2018). However, it presents many differences with previous domain adaptation problems, where the main goal is to adapt from a particular topic or style. In the case of noisy text, it will not only be the case that a particular word will be translated in a different way than it is in the general domain (e.g. as in the case of "sub"), but also that there will be increased lexical variation (e.g. due to spelling or typographical errors), and also inconsistency in grammar (e.g. due to omissions of critical words or mis-usage). The sum of these differences warrants that noisy MT be treated as a separate instance than domain adaptation, and our experimental analysis in 5.4 demonstrates that even after performing adaptation, MT systems still make a large number of noise-related errors.

3 Collection Procedure

We first collect noisy sentences in our three languages of interest, English, French and Japanese.

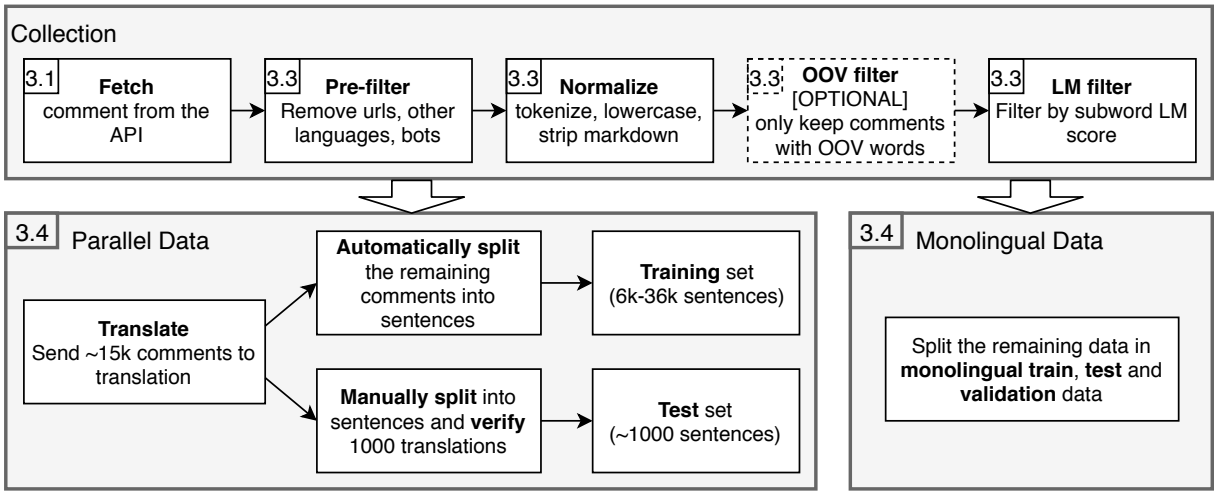

Figure 1: Summary of our collection process and the respective sections addressing them. We apply the same procedure for each language.

We refer to Figure 1 for an overview of the data collection and translation process.

We choose Reddit as a source of data because (1) its content is likely to exhibit noise, (2) some of its sub-communities are entirely run in different languages, in particular, English, French and Japanese, and (3) Reddit is a popular source of data in curated and publicly distributed NLP datasets (Tan et al.). We collect data using the public Reddit API. [4]

Note that the data collection and translation is performed at the comment level. We split the parallel data into sentences as a last step.

3.1 Data Sources

For each language, we select a set of communities ("subreddits") that we know contain many comments in that language:

English: Since an overwhelming majority of the discussions on Reddit are conducted in English, we don't restrict our collection to any community in particular.

French: `/r/france`, `/r/quebec` and `/r/rance`. The first two are among the biggest French speaking communities on Reddit. The third is a humor/sarcasm based offspring of `/r/france`.

Japanese: `/r/newsokur`, `/r/bakanewsjp`, `/r/newsokuvip`, `/r/lowlevelaware` and `/r/steamr`. Those are the biggest Japanese speaking communities, with over 2,000 subscribers.

We collect comments made during the 03/27/2018-03/29/3018 time period for English, 09/2018-03/2018 for French and 11/2017-03/2018 for Japanese. The large difference in collection time is due to the variance in comment throughput and relative amount of noise between the languages.

3.2 Contrast Corpora

Not all comments found on Reddit exhibit noise as described in Section 2. Because we would like to focus our data collection on noisy comments, we devise criteria that allow us to distinguish potentially noisy comments from clean ones. Specifically, we compile a *contrast* corpus composed of clean text that we can compare to, and find potentially noisy text that differs greatly from the contrast corpus. Given that our final goal is MT robust to noise, we prefer that these contrast corpora consist of the same type of data that is often used to train NMT models. We select different datasets for each language:

English: The English side of the preprocessed parallel training data provided for the German-English WMT 2017 News translation task,[5] as provided on the website. This amounts to ≈ 5.85 million sentences.

[4] In particular, we use this implementation: `praw.readthedocs.io/en/latest`, and our complete code is available at `http://www.cs.cmu.edu/~pmichel1/mtnt/`.

[5] `http://www.statmt.org/wmt17/translation-task.html`

French: The entirety of the French side of the parallel training data provided for the English-French WMT 2015 translation task.[6] This amounts to ≈ 40.86 million sentences.

Japanese: We aggregate three small/medium sized MT datasets: KFTT (Neubig, 2011), JESC (Pryzant et al.) and TED talks (Cettolo et al., 2012), amounting to ≈ 4.19 million sentences.

3.3 Identifying Noisy Comments

We now describe the procedure used to identify comments containing noise.

Pre-filtering First, we perform three pre-processing to discard comments that do not represent natural noisy text in the language of interest:

1. Comments containing a URL, as detected by a regular expression.
2. Comments where the author's username contains "bot" or "AutoModerator". This mostly removes automated comments from bots.
3. Comments in another language: we run `langid.py`[7] (Lui and Baldwin) and discard comments where $p(\text{lang} \mid \text{comment}) > 0.5$ for any language other than the one we are interested in.

This removes cases that are less interesting, i.e. those that could be solved by rule-based pattern matching or are not natural text created by regular users in the target language. Our third criterion in particular discards comments that are blatantly in another language while still allowing comments that exhibit code-switching or that contain proper nouns or typos that might skew the language identification. In preliminary experiments, we noticed that these criteria 14.47, 6.53 and 7.09 % of the collected comments satisfied the above criteria respectively.

Normalization After this first pass of filtering, we pre-process the comments before running them through our noise detection procedure. We first strip Markdown[8] syntax from the comments. For English and French, we normalize the punctuation, lowercase and tokenize the comments using the Moses tokenizer. For Japanese, we simply lowercase the alphabetical characters in the comments. Note that this normalization is done for the purpose of noise detection only. The collected comments are released without any kind of pre-processing. We apply the same normalization procedure to the contrast corpora.

Unknown words In the case of French and English, a clear indication of noise is the presence of out-of-vocabulary words (OOV): we record all lowercased words encountered in our reference corpus described in Section 3.2 and only keep comments that contain at least one OOV. Since we did not use word segmentation for the Japanese reference corpus, we found this method not to be very effective to select Japanese comments and therefore skipped this step.

Language model scores The final step of our noise detection procedure consists of selecting those comments with a low probability under a language model trained on the reference monolingual corpus. This approach mirrors the one used in Moore and Lewis and Axelrod et al. to select data similar to a specific domain using language model perplexity as a metric. We search for comments that have a low probability under a sub-word language model for more flexibility in the face of OOV words. We segment the contrast corpora with Byte-Pair Encoding (BPE) using the sentencepiece[9] implementation. We set the vocabulary sizes to $1,000$, $1,000$ and $4,000$ for English, French and Japanese respectively. We then use a 5-gram Kneser-Ney smoothed language model trained using `kenLM`[10] (Heafield et al.) to calculate the log probability, normalized by the number of tokens for every sentence in the reference corpus. Given a reddit comment, we compute the normalized log probability of each of its lines under our subword language model. If for any line this score is below the 1st percentile of scores in the reference corpus, the comment is labeled as noisy and saved.

3.4 Creating the Parallel Corpora

Once enough data has been collected, we isolate $15,000$ comments in each language by the follow-

[6] http://www.statmt.org/wmt15/translation-task.html
[7] https://github.com/saffsd/langid.py
[8] https://daringfireball.net/projects/markdown
[9] https://github.com/google/sentencepiece
[10] https://kheafield.com/code/kenlm/

	#samples	#src tokens	#trg tokens
en-fr	1,020	15,919	18,445
fr-en	1,022	16,662	16,038
en-ja	1,002	11,040	20,008
ja-en	1,020	23,997	33,429

Table 1: Test set numbers.

	#samples	#src tokens	#trg tokens
en-fr	36,058	841k	965k
fr-en	19,161	661k	634k
en-ja	5,775	281k	506k
ja-en	6,506	172k	128k

Table 2: Training sets numbers.

	#samples	#src tokens	#trg tokens
en-fr	852	16,957	18,948
fr-en	886	41,578	46,886
en-ja	852	40,124	46,886
ja-en	965	25,010	23,289

Table 3: Validation sets numbers.

ing procedure:

- Remove all duplicates. In particular, this handles comments that might have been scraped twice or automatic comments from bots.
- To further weed out outliers (comments that are too noisy, *e.g.* ASCII art, wrong language…or not noisy enough), we discard comments that are on either end of the distribution of normalized LM scores within the set of collected comments. We only keep comments whose normalized score is within the 5-70 percentile for English (resp. 5-60 for French and 10-70 for Japanese). These numbers are chosen by manually inspecting the data.
- Choose $15,000$ samples at random.

We then concatenate the title of the thread where the comment was found to the text and send everything to an external vendor for manual translations. Upon reception of the translations, we noticed a certain amount of variation in the quality of translations, likely because translating social media text, with all its nuances, is difficult even for humans. In order to ensure the highest quality in the translations, we manually filter the data to segment the comments into sentences and weed out poor translations for our test data. We thereby retain around $1,000$ sentence pairs in each direction for the final test set.

We gather the samples that weren't selected for the test sets to be used for training or fine-tuning models on noisy data. We automatically split comments into sentences with a regular expression detecting sentence delimiters, and then align the source and target sentences. Should this alignment fail (*i.e.* the source comment contains a different number of sentences than the target comment after automatic splitting), we revert back to providing the whole comment without splitting. For the training data, we do not verify the correctness of translations as closely as for the test data. Finally, we isolate $\approx$ 900 samples in each direction to serve as validation data.

Information about the size of the data can be found in Table 1, 2 and 3 for the test, training and validation sets respectively. We tokenize the English and French data with the Moses (Koehn et al.) tokenizer and the Japanese data with Kytea (Neubig et al., 2011) before counting the number of tokens in each dataset.

3.5 Monolingual Corpora

After the creation of the parallel train and test sets, a large number of unused comments remain in each language, which we provide as monolingual corpora. This additional data has two purposes: first, it serves as a resource for in-domain training using semi-supervised methods relying on monolingual data (e.g. Cheng et al.; Zhang and Zong). Second, it provides a language modeling dataset for noisy text in three languages.

We select $3,000$ comments at random in each dataset to form a validation set to be used to tune hyper-parameters, and provide the rest as training data. The data is provided with one comment per line. Newlines within individual comments are replaced with spaces. Table 4 contains information

		#samples	#tok	#char
en	train	81,631	3,99M	18,9M
	dev	3,000	146k	698k
fr	train	26,485	1,52M	7,49M
	dev	3,000	176k	867k
ja	train	32,042	943k	3.9M
	dev	3,000	84k	351k

Table 4: Monolingual data numbers.

		Spelling	Grammar	Emojis	Profanities
en	newstest2014	0.210	0.189	0.000	0.030
	newsdiscusstest2015	0.621	0.410	0.021	0.076
	MTNT (en-fr)	**2.180**	**0.559**	**0.289**	**0.239**
fr	newstest2014	2.776	0.091	0.000	0.245
	newsdiscusstest2015	1.686	0.457	0.024	0.354
	MTNT	**4.597**	**1.464**	**0.252**	**0.690**
ja	TED	0.011	0.266	0.000	0.000
	KFTT	0.021	0.228	0.000	0.000
	JESC	0.096	0.929	0.090	**0.058**
	MTNT	**0.269**	**1.527**	**0.156**	0.036

Table 5: Numbers, per 100 tokens, of quantifiable noise occurrences. For each language and category, the dataset with the highest amount of noise is highlighted.

on the size of the datasets. As with the parallel MT data, we provide the number of tokens after tokenization with the Moses tokenizer for English and French and Kytea for Japanese.

4 Dataset Analysis

In this section, we investigate the proposed data to understand how different categories of noise are represented and to show that our test sets contain more noise overall than established MT benchmarks.

4.1 Quantifying Noisy Phenomena

We run a series of tests to count the number of occurrences of some of the types of noise described in Section 2. Specifically we pass our data through spell checkers to count spelling and grammar errors. Due to some of these tests being impractical to run on a large scale, we limit our analysis to the test sets of MTNT.

We use slightly different procedures depending on the tools available for each language. We test for spelling and grammar errors in English data using Grammarly[11], an online resource for English spell-checking. Due to the unavailability of an equivalent of Grammarly in French and Japanese, we test for spelling and grammar error using the integrated spell-checker in Microsoft Word 2013[12]. Note that Word seems to count proper nouns as spelling errors, giving higher numbers of spelling errors across the board in French as compared to English.

For all languages, we also count the number of profanities and emojis using custom-made lists and regular expressions[13]. In order to compare results across datasets of different sizes, we report all counts per 100 words.

The results are recorded in the last row of each section in Table 5. In particular, for the languages with a segmental writing system, English and French, spelling errors are the dominant type of noise, followed by grammar error. Unsurprisingly, the former are much less present in Japanese.

4.2 Comparison to Existing MT Test Sets

Table 5 also provide a comparison with the relevant side of established MT test sets. For English and French, we compare our data to newstest2014[14] and newsdiscusstest2015[15] test sets. For Japanese, we compare with the test sets of the datasets described in Section 3.2.

Overall, MTNT contains more noise in all metrics but one (there are more profanities in JESC, a Japanese subtitle corpus). This confirms that MTNT indeed provides a more appropriate benchmark for translation of noisy or non-standard text.

Compared to synthetically created noisy test sets (Belinkov and Bisk, 2018) MTNT contains less systematic spelling errors and more varied types of noise (*e.g.* emojis and profanities) and is thereby more representative of naturally occurring noise.

[11] https://www.grammarly.com/

[12] https://products.office.com/en-us/microsoft-word-2013

[13] available with our code at https://github.com/pmichel31415/mtnt

[14] http://www.statmt.org/wmt15/dev-v2.tgz

[15] http://www.statmt.org/wmt15/test.tgz

5 Machine Translation Experiments

We evaluate standard NMT models on our proposed dataset to assess its difficulty. Our goal is not to train state-of-the art models but rather to test standard off-the-shelf NMT systems on our data, and elucidate what features of the data make it difficult.

5.1 Model Description

All our models are implemented in DyNet (Neubig et al., 2017) with the XNMT toolkit (**?**). We use approximately the same setting for all language pairs: the encoder is a bidirectional LSTM with 2 layers, the attention mechanism is a multi layered perceptron and the decoder is a 2 layered LSTM. The embedding dimension is 512, all other dimensions are 1024. We tie the target word embeddings and the output projection weights (Press and Wolf). We train with Adam (Kingma and Ba, 2014) with XNMT's default hyper-parameters, as well as dropout (with probability 0.3). We used BPE subwords to handle OOV words. Full configuration details as well as code to reproduce the baselines is available at `https://github.com/pmichel31415/mtnt`.

5.2 Training Data

We train our models on standard MT datasets:

- en ↔ fr: Our training data consists in the europarl-v7[16] and news-commentary-v10[17] corpora, totaling $2,164,140$ samples, $54,611,105$ French tokens and $51,745,611$ English tokens (non-tokenized). We use the newsdiscussdev2015[14] dev set from WMT15 as validation data and evaluate the model on the newsdiscusstest2015[15] and newstest2014[14] test sets.

- en ↔ ja: We concatenate the respective train, validation and test sets of the three corpora mentioned in 3.2. In particular we detokenize the Japanese part of each dataset to make sure that any tokenization we perform will be uniform (in practice we remove ASCII spaces). This amounts to $3,900,772$ training samples ($34,989,346$ English tokens without tokenization). We concatenate the dev sets associated with these corpora to serve as validation data and evaluate on each respective test set separately.

[16] `http://www.statmt.org/europarl/`

[17] `http://www.statmt.org/wmt15/training-parallel-nc-v10.tgz`

	en-fr	fr-en
newstest2014	33.52	28.93
newsdiscusstest2015	33.03	30.76
MTNT	21.77	23.27
MTNT (+tuning)	29.73	30.29
	en-ja	**ja-en**
TED	14.51	13.25
KFTT	20.82	20.77
JESC	15.77	18.00
MTNT	9.02	6.65
MTNT (+tuning)	12.45	9.82

Table 6: BLEU scores of NMT models on the various datasets.

5.3 Results

We use `sacreBLEU`[18], a standardized BLEU score evaluation script proposed by Post (2018), for BLEU evaluation of our benchmark dataset. It takes in detokenized references and hypotheses and performs its own tokenization before computing BLEU score. We specify the `intl` tokenization option. In the case of Japanese text, we run both hypothesis and reference through KyTea before computing BLEU score. We strongly encourage that evaluation be performed in the same manner in subsequent work, and will provide both scripts and an evaluation web site in order to facilitate reproducibility.

Table 6 lists the BLEU scores for our models on the relevant test sets in the two language pairs, including the results on MTNT.

5.4 Analysis

To better understand the types of errors made by our model, we count the n-grams that are over- and under- generated with respect to the reference translation. Specifically, we compare the count ratios of all 1- to 3-grams in the output and in the reference and look for the ones with the highest (over-generated) and lowest (under-generated) ratio.

We find that in English, the model undergenerates the contracted form of the negative ("do not"/"don't") or of auxiliaries ("That is"/"I'm").

[18] `https://github.com/mjpost/sacreBLEU`

Source	Moi faire la gueule dans le métro me manque, c'est grave ?
Target	I miss sulking in the underground, is that bad?
Our model	I do not know what is going on in the metro, that is a serious matter.
+ fine-tuning	I do not want to be in the metro, it's serious?
Source	:o 'tain je me disais bien que je passais à côté d'un truc vu les upvotes.
Target	:o damn I had the feeling that I was missing something considering the upvotes.
Our model	o, I was telling myself that I was passing over a nucleus in view of the Yupvoots.
+ fine-tuning	o, I was telling myself that I was going next to a nucleus in view of the <unk>upvotes.
Source	* C'est noël / pâques / pentecôte / toussaint : Pick One, je suis pas catho
Target	Christmas / Easter / Pentecost / All Saints: Pick One, I'm not Catholic!
Our model	<unk> It is a pale/poward, a palec<unk>te d'<unk>tat: Pick One, I am not a catho!
+ fine-tuning	<unk> It's no<unk>l / pesc<unk>e /pentecate /mainly: Pick One, I'm not catho!

Table 7: Comparison of our model's output before and after fine-tuning in fr-en.

Similarly, in French, our model over generates "de votre" (where "votre" is the formal 2nd person plural for "your") and "n'ai pas" which showcases the "ne [...] pas" negation, often dropped in spoken language. Conversely, the informal second person "tu" is under-generated, as is the informal and spoken contraction of "cela", "ça". In Japanese, the model under-generates, among others, the informal personal pronoun 俺 ("ore") or the casual form だ ("da") of the verb です ("desu", to be). In ja-en the results are difficult to interpret as the model seems to produce incoherent outputs (*e.g.* "no, no, no...") when the NMT system encounters sentences it has not seen before. The full list of n-grams with the top 5 and bottom 5 count ratios in each language pair is displayed in Table 8.

fr-en	en-fr	ja-en	en-ja
Over generated			
<unk>	<unk>	no, no,	※
it is not	qu'ils	i	が
I do not	de votre	no, no, no,	か ?
That is	s'il	so on and	て
not have	n'ai pas	on and so	す か ?
Under generated			
it's	tu	!	?
I'm	ça	Is	よ 。
I don't	que tu	>	って
>	!	""The	俺
doesn't	as	those	だ 。

Table 8: Over and under generated n-grams in our model's output for en-fr

5.5 Fine-Tuning

Finally, we test a simple domain adaptation method by fine-tuning our models on the training data described in Section 3.4. We perform one epoch of training with vanilla SGD with a learning rate of 0.1 and a batch size of 32. We do not use the validation data at all. As evidenced by the results in the last row of Table 6, this drives BLEU score up by 3.17 to 7.96 points depending on the language pair. However large this increase might be, our model still breaks on very noisy sentences. Table 7 shows three examples in fr-en. Although our model somewhat improves after fine-tuning, the translations remain inadequate in all cases. In the third case, our model downright fails to produce a coherent output. This shows that despite improving BLEU score, naive domain adaptation by fine-tuning doesn't solve the problem of translating noisy text.

6 Language Modeling Experiments

In addition to our MT experiments, we report character-level language modeling results on the monolingual part of our dataset. We use the data described in Section 3.5 as training and validation sets. We evaluate the trained model on the source side of our en-fr, fr-en and ja-en test sets for English, French and Japanese respectively.

We report results for two models: a Kneser-Ney smoothed 6-gram model (implemented with KenLM) and an implementation of the AWD-LSTM proposed in (Merity et al., 2018)[19]. We report the Bit-Per-Character (bpc) counts in table 9.

[19] https://github.com/salesforce/awd-lstm-lm

	6-gram		AWD LSTM	
	dev	test	dev	test
English	2.081	2.179	1.706	1.810
French	1.906	2.090	1.449	1.705
Japanese	5.003	5.497	4.801	5.225

Table 9: Language modeling scores

We intend these results to serve as a baseline for future work in language modeling of noisy text in either of those three languages.

7 Related work

Handling noisy text has received growing attention among various language processing tasks due to the abundance of user generated content on popular social media platforms (Crystal, 2001; Herring, 2003; Danet and Herring, 2007). These contents are considered as noisy when compared to news corpora which have been the main data source for language tasks (Baldwin et al.; Eisenstein). They pose several unique challenges because they contain a larger variety of linguistic phenomena that are absent in the news domain and that lead to degraded quality when applying an model to out-of-domain data (Ritter et al.; Luong and Manning, 2015). Additionally, they are live examples of the Cmabrigde Uinervtisy (Cambridge University) effect, where state-of-the-art models become brittle while human's language processing capability is more robust (Sakaguchi et al., 2017; Belinkov and Bisk, 2018).

Efforts to address these challenges have been focused on creating in-domain datasets and annotations (Owoputi et al.; Kong et al.; Blodgett et al., 2017), and domain adaptation training (Luong and Manning, 2015). In MT, improvements were obtained for SMT (Formiga and Fonollosa). However, the specific challenges for neural machine translation have not been studied until recently (Belinkov and Bisk, 2018; Sperber et al.; Cheng et al., 2018). The first provides empirical evidence of non-trivial quality degradation when source sentences contain natural noise or synthetic noise within words, and the last two explore data augmentation and adversarial approaches of adding noise efficiently to training data to improve robustness.

Our work also contributes to recent advances in evaluating neural machine translation quality with regard to specific linguistic phenomena, such as manually annotated test sentences for English to French translation, in order to identify errors due to specific linguistic divergences between the two languages (Isabelle et al.), or automatically generated test sets to evaluate typical errors in English to German translation (Sennrich). Our contribution distinguishes itself from this previous work and other similar initiatives (Peterson, 2011) by providing an open test set consisting of naturally occurring text exhibiting a wide range of phenomena related to noisy input text from contemporaneous social media.

8 Conclusion

We proposed a new dataset to test MT models for robustness to the types of noise encountered in natural language on the Internet. We contribute parallel training and test data in both directions for two language pairs, English $\leftrightarrow$ French and English $\leftrightarrow$ Japanese, as well as monolingual data in those three languages. We show that this dataset contains more noise than existing MT test sets and poses a challenge to models trained on standard MT corpora. We further demonstrate that these challenges cannot be overcome by a simple domain adaptation approach alone. We intend this contribution to provide a standard benchmark for robustness to noise in MT and foster research on models, dataset and evaluation metrics tailored for this specific problem.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. *International Conference on Learning Representations*.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2017. A dataset and classifier for recognizing social

media english. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 56–61.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391.

David Crystal. 2001. *Language and the Internet*. Cambridge University Press.

Brenda Danet and Susan Herring. 2007. *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press., New York.

Jacob Eisenstein. 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.

Lluís Formiga and José A. R. Fonollosa. Dealing with input noise in statistical machine translation. In *Proceedings of the Conference on Computational Linguistics 2012: Posters*, pages 319–328.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 690–696.

Susan Herring, editor. 2003. *Media and Language Change*. Special issue of Journal of Historical Pragmatics 4:1.

Pierre Isabelle, Colin Cherry, and George Foster. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Huda Khayrallah and Philipp Koehn. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions*, pages 177–180.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012.

Mu Li, Yinggong Zhao, Dongdong Zhang, and Ming Zhou. Adaptive development data selection for log-linear model in statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 662–670.

Marco Lui and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the Association for Computational Linguistics 2012 System Demonstrations*, pages 25–30.

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and optimizing lstm language models. *International Conference on Learning Representations*.

Antonio Valerio Miceli Barone, Barry Haddow, Ulrich Germann, and Rico Sennrich. Regularization techniques for fine-tuning in neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1490–1495.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation.

Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the Association for Computational Linguistics 2010 Conference Short Papers*, pages 220–224.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, et al. 2017. Dynet: The dynamic neural network toolkit. *arXiv preprint arXiv:1701.03980*.

Graham Neubig, Yosuke Nakata, and Shinsuke Mori. 2011. Pointwise prediction for robust, adaptable japanese morphological analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 529–533.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics.

Kay Peterson. 2011. Openmt12 evaluation.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163.

R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz. Jesc: Japanese-english subtitle corpus. *ArXiv e-prints*.

Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1524–1534. Association for Computational Linguistics.

Keisuke Sakaguchi, Kevin Duh, Matt Post, and Benjamin Van Durme. 2017. Robsut wrod reocginiton via semi-character recurrent neural network. In *AAAI*, pages 3281–3287.

Rico Sennrich. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382.

Matthias Sperber, Jan Niehues, and Alex Waibel. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 90–96.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1483–1489.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Jiajun Zhang and Chengqing Zong. Exploiting source-side monolingual data in neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545.