Findings of the First Shared Task on Machine Translation Robustness

Xian Li[1], Paul Michel[2], Antonios Anastasopoulos[2], Yonatan Belinkov[3], Nadir Durrani[4],
Orhan Firat[5], Philipp Koehn[6], Graham Neubig[2], Juan Pino[1], Hassan Sajjad[4]
[1]Facebook AI, [2]Carnegie Mellon University, [3]Harvard University and MIT,
[4]Qatar Computing Research Institute, [5]Google AI, [6]Johns Hopkins University

Abstract

We share the findings of the first shared task on improving robustness of Machine Translation (MT). The task provides a testbed representing challenges facing MT models deployed in the real world, and facilitates new approaches to improve models' robustness to noisy input and domain mismatch. We focus on two language pairs (English-French and English-Japanese), and the submitted systems are evaluated on a blind test set consisting of noisy comments on Reddit[1] and professionally sourced translations. As a new task, we received 23 submissions by 11 participating teams from universities, companies, national labs, etc. All submitted systems achieved large improvements over baselines, with the best improvement having +22.33 BLEU. We evaluated submissions by both human judgment and automatic evaluation (BLEU), which shows high correlations (Pearson's $r = 0.94$ and 0.95). Furthermore, we conducted a qualitative analysis of the submitted systems using `compare-mt`[2], which revealed their salient differences in handling challenges in this task. Such analysis provides additional insights when there is occasional disagreement between human judgment and BLEU, e.g. systems better at producing colloquial expressions received higher score from human judgment.

1   Introduction

In recent years, Machine Translation (MT) systems have seen great progress, with neural models becoming the *de-facto* methods and even approaching human quality in news domain (Hassan et al., 2018). However, like other deep learning models, neural machine translation (NMT) models are found to be sensitive to synthetic and natural noise in input, distributional shift, and adversarial

examples (Koehn and Knowles, 2017; Belinkov and Bisk, 2018; Durrani et al., 2019; Anastasopoulos et al., 2019; Michel et al., 2019). From an application perspective, MT systems need to deal with non-standard, noisy text of the kind which is ubiquitous on social media and the internet, yet has different distributional signatures from corpora in common benchmark datasets.

The goal of this shared task is to provide a testbed for improving MT models' robustness to orthographic variations, grammatical errors, and other linguistic phenomena common in user-generated content, via better modelling, training, adaptation techniques, or leveraging monolingual training data. Specifically, the shared task aims to bring improvements on the following challenges:

- To improve NMT's robustness to orthographic variations, grammatical errors, informal language, and other linguistic phenomena or noise common on social media.

- To explore effective approaches to leverage abundant out-of-domain parallel data.

- To explore novel approaches to leverage abundant monolingual data on the Web (e.g., tweets, Reddit comments, commoncrawl, etc.).

- To thoroughly investigate and understand the overall challenges in translating social media text and identify major themes of efforts which needs more research from the community.

In this first iteration, the shared-task used the MTNT dataset (Michel and Neubig, 2018) that contains noisy social media texts and their translations between English (Eng) and French (Fra) and English and Japanese (Jpn), in four translation directions: Eng→Fra, Fra→Eng, Eng→Jpn, and

---

[1]www.reddit.com
[2]https://github.com/neulab/compare-mt

Jpn→Eng. We describe the dataset and the task setup in Section 3. The shared-task attracted a total of 23 submissions from 11 teams. The teams employed a variety of methods to improve robustness. A specific challenge was the small size of the in-domain noisy parallel dataset. We summarize the participating systems in Section 4 and the notable methods in Section 5. The contributions were evaluated both automatically and via a human evaluation. The results demonstrate a significant progress of the state-of-the-art in MT robustness, with multiple teams surpassing the shared-task baseline by a large margin. These results are discussed in Section 6.

We hope that this task leads to more efforts from the community in building robust MT models.

2 Related Work

The fragility of neural networks (Szegedy et al., 2013) has been shown to extend to neural machine translation models (Belinkov and Bisk, 2018; Heigold et al., 2017) and recent work focused on various aspects of the problem. From the identification of the causes of this brittleness, to the induction of (adversarial) inputs that trigger the unwanted behavior (attacks) and making such models robust against various types of noisy inputs (defenses); improving robustness has been receiving increasing attention in NMT.

While Koehn and Knowles (2017) mentioned domain mismatch as a challenge for neural machine translation, Khayrallah and Koehn (2018) addressed noisy training data and focus on the types of noise occurring in web-crawled corpora. Michel and Neubig (2018) proposed a new dataset (MTNT) to test MT models for robustness to the types of noise encountered in the Internet and demonstrated that these challenges cannot be overcome by simple domain adaptation techniques alone.

Belinkov and Bisk (2018) and Heigold et al. (2017) showed that NMT systems are very sensitive to slightly perturbed input forms, and hinted at the importance of injecting noisy examples during training, also known as adversarial examples. Further research proposed several methods of generating and using noisy examples as NMT input to advance the understanding and improve the translation quality. Following machine vision, two major branches being explored when generating noisy examples, *i)* white box methods, where adversarial

examples are generated with access to the model parameters (Ebrahimi et al., 2018; Cheng et al., 2018a,b, 2019) and *ii)* black-box attacks, where examples are generated without accessing model internals (Zhao et al., 2018; Lee et al., 2018; **?**; Anastasopoulos et al., 2019; Vaibhav et al., 2019); see Belinkov and Glass (2019) for a categorization of such work. In particular, some have focused on specific variations of naturally-occurring noise, such as grammatical errors produced by non-native speakers (Anastasopoulos et al., 2019) or errors extracted from Wikipedia edits (Belinkov and Bisk, 2018). It has also been shown that adding synthetic noise does not trivially increase robustness to natural noise (Belinkov and Bisk, 2018) and may require specific recipes (Karpukhin et al., 2019).

Michel et al. (2019) recently emphasized the importance of meaning-preserving perturbations and along with Cheng et al. (2019) demonstrated the utility of adversarial training without significantly impairing performance on clean data and domain. Durrani et al. (2019) showed that character-based representations are more robust towards noise compared to such learned using BPE-based sub-word units in the task of machine translation.

3 Task

This is the first year we introduce the robustness task. The goal of the task setup is to examine MT systems' performance on non-standard, noisy, user-generated text, which often resemble mixed challenges around orthographic variations, grammar errors, domain shift and stylistic lexical choice, etc. We use the MTNT dataset (Michel and Neubig, 2018) as a testbed for the above-mentioned robustness challenges. To give readers an idea of the natural "noise" present in the MTNT dataset, and the challenges for MT systems to robustly understand and translate them, we provide some examples of input variations:

- **Spelling/typographical errors:** accross (across), recieve (receive), tant (temps)

- **Grammatical errors:** a tons of, there are less people

- **Spoken language and internet slang:** wanna, chais pas, tbh, smh, mdr

- **Code switching:** This is so kawaii, C'est trop mainstream

- **Profanity/slurs:** f*ck, m*rde

Readers are encouraged to refer to Michel and Neubig (2018) for more details. This year's task probes MT robustness for two language pairs, French to/from English and Japanese to/from English.

3.1 Task Setup

The task includes two tracks, *constrained* and *unconstrained* depending on whether the system is trained on a predefined training datasets or not. The two tracks are evaluated by the same automatic and human evaluation protocol, however, they are compared separately.

For the constrained system track, the task specifies two types of training data in addition to MTNT train set:

- **"Out-of-domain" parallel data:** This facilitates MT model's capability to perform supervised learning from examples with different distribution such as lexical choice, language style, genre etc. For example, parallel corpora from WMT news translation task, subtitles and TED talks are specified.

- **Monolingual data:** We encourage participants to develop novel solutions to learn from unlabelled data, improve existing semi-supervised approach such as backtranslation. We provide both in-domain (MTNT) and out-of-domain (News Commentary, News Crawl, etc) monolingual data.

3.2 Training Data

In the constrained setting, participants were allowed to use the WMT15 training data[3] for Eng↔Fra and any of the KFTT (Neubig, 2011), JESC (Pryzant et al.) and TED talks (Cettolo et al., 2012) corpora for Jpn↔Eng. Additionally, the use of the MTNT corpus (Michel and Neubig, 2018) was allowed in order to adapt models on limited in-domain data.

3.3 Test Data

The test sets were collected following the same protocol as the MTNT dataset, *i.e.* collected from

Reddit, filtered out for noisy comments using a sub-word language modeling criterion and translated by professional translators. The statistics of the test sets are reported in Table 1.

3.4 Evaluation protocol

The system outputs were evaluated by professional translators. The translators were presented the original source sentence, the reference and the system output side by side. The order between the reference and the system output was randomized by the user interface. The translators rated both the reference and the translation on a scale from 1 to 100. For both the original source sentence and the reference, the original text was presented except for Eng-Jpn where the Japanese reference tokenized with KyTea was presented in order to be consistent with the systems' outputs. The user interface for annotation is illustrated in Figure 1.

We also evaluated BLEU (Papineni et al., 2002) for each system using SacreBLEU (Post, 2018). For all language pairs except Eng-Jpn, we used the original reference and SacreBLEU with the default options. In the case of Eng-Jpn, we used the reference tokenized with KyTea and the option `--tokenize none`.

4 Participants and System Descriptions

We received 23 submissions from 11 teams. Except two submissions on the Eng-Fra language pair, all systems used the *constrained* setup. Below we briefly describe the systems from the 8 teams which submitted corresponding system description papers:

Baidu & Oregon State University's submission (Zheng et al., 2019): Their system is based on the Transformer implementation in OpenNMT-py (Klein et al., 2017). The main methods applied in their submission are: domain-sensitive data mixing and data augmentation with back-translation. For data mixing, they used a special symbol on the source side to indicate the data domain. For data augmentation, they back-translate from a target language to its noisy source. The intuition, also observed by Michel and Neubig (2018), is that the source sentences are noisier than their target translations. They include out-of-domain clean data during this step and differentiate data types with a special symbol on the target side. In addition, they also run a model ensemble.

Please use the below guidelines to give your rate:

- **1-10 Completely incorrect**
  The translation is completely incorrect and inaccurate. It bears no resemblance to the source string and there are major grammar mistakes and/or typos, or the meaning is impossible to understand.
- **11-30 Overall meaning is different / lost**
  The translation has a few correct keywords from the source string, but the overall meaning is different, or lost, or impossible to understand.
- **31-50 Major grammar mistakes / major typos**
  The translation represents parts of the source string, but does not cover the overall meaning, or there are major grammar mistakes and/or major typos.
- **51-70 Understandable, with few translation errors / grammar mistakes**
  The translation is understandable and represents the overall meaning of source string. There are a few translation errors, grammar mistakes, and/or major typos.
- **71-90 Very good translation, only a couple of minor mistakes**
  The translation renders the source string in the target language very closely. Only a couple of minor grammar mistakes and/or minor typos are present.
- **91-100 Accurate translation, no mistakes**
  The translation renders the source string in the target language accurately. There are no mistakes.

ORIGINAL	TRANSLATIONS	COMMENTS	
			#1
I Mean I hit max this month.	Je veux dire, je frappe max ce mois-ci. Rate (A) **75** **Very good translation, only a couple of minor mistakes** Je veux dire que j'ai atteint max ce mois-ci. Rate (B) **75** **Very good translation, only a couple of minor mistakes**	✓ A B ✗	I do not know this language / There are no words to translate
			#2
I think the worst clapping I experienced was at the hands of a Gem Knights player who got out multiple beastly fusions on turn 1 before Silent Magician could even negate the fusion.	Je pense que le pire moment que j'ai vécu a été provoqué par un joueur de Chevalier-gemme qui a sorti plusieurs fusions bestiales au premier tout avant que le Magicien Silencieux ne puisse même annuler la fusion. Rate (A) **61** **Understandable, with few translation errors / grammar mistakes** Je pense que la pire claque que j'ai vécue était aux mains d'un joueur de Gem Knights qui a fait de multiples fusions au tour 1 avant que Silent Magian puisse même annuler la fusion. Rate (B) **86** **Very good translation, only a couple of minor mistakes**	✓ A B ✗	I do not know this language / There are no words to translate

Figure 1: Annotation interface for human evaluations.

	Eng-Fra	Fra-Eng	Eng-Jpn	Jpn-Eng
# samples	1,401	1,233	1,392	1,111
# source tokens	20.0k	19.8k	20.0k	18.7k
# target tokens	22.8k	19.2k	33.6k	13.4k

Table 1: Statistics of the test sets.

The team experimented with the Fra→Eng and Eng→Fra translation directions, obtaining 43.6 and 36.4 BLEU-cased, respectively (3rd place in both). Their ablations show significant benefit from domain-sensitive training (+3 BLEU), with additional improvements from back-translation and ensembling.

CMU's submission (Zhou et al., 2019): This submission only participated in the Fra→Eng direction. They proposed the use of tied multitask learning, where the noisy source sentences are first decoded by a same-language denoising decoder, and both information is passed on to the translation decoder. This approach requires data triples of noisy source, clean source, translation, which they created by data augmentation over the provided data, using tag-informed translation systems trained on either noisy (MTNT) or clean (Europarl) data. As the participants point out though, their performance improvements seems to be attributed to data augmentation and not to the intermediate denoising decoder.

CUNI's submission (Helcl et al., 2019): They participated in Eng→Fra and Fra→Eng directions, following a classical two stage approach, i) training of a base model using a mix of parallel (WMT15 Eng-Fra News Translation) and back-translated monolingual data (from News Crawl and Europarl - excluding News Discussions), ii) fine-tuning of the base model using the training portion of the MTNT dataset. All models follow the Transformer-Big architecture, with the hyperparameters and optimization recipe from the 2018 WMT News Translation shared task submission of CUNI, without ensembles. For both Eng-Fra and Fra-Eng directions, fine-tuning brought about 2+ BLEU points on top of the base models with the Transformer-Big architecture, whereas improvements were substantially larger when the base models were RNN-Based MTNT baselines, about 8+ BLEU points. Participants emphasized the importance of their strong Transformer-Big base

model which was already 10+ BLEU points better than the MTNT baseline provided by the shared task. The effect of individual partitions of the base model training set (parallel and backtranslated-mono) on final system quality is not experimented. Finally, participants point out one peculiarity they've noticed in the train/validation partitioning of the original MTNT dataset; validation source sentences being started with the letter "Y" followed by alphabetically sorted sentences (test partition not effected).

FOKUS' submission (Grozea, 2019): This team participated in three directions: Eng→Fra, Fra→Eng and Jpn→Eng. For the Eng→Fra and Fra→Eng language pairs, the submissions are *unconstrained* systems, where the model was trained on the medical domain corpus provided by the WMT biomedical shared task [4]. Despite the training data being out-of-domain, removing "low-quality" parallel data such as "Subtitles" as the author hypothesized helped to bring 2 to 4 BLEU points improvement over the baseline models. Their Jpn→Eng submission is a *constrained* system, using the same model architecture as the Eng→Fra language pair. To improve robustness, they introduced synthetic noise (omitting and duplicating letters) in the training data to both source and target sentences.

JHU's submission (Post and Duh, 2019): This submission participated in the Fra→Eng and Jpn↔Eng tasks. The participants used data dual cross-entropy filtering for reducing the monolingual data, then back-translate these, and train their Transformer models (Vaswani et al., 2017). They compared Moses tokenization+Byte Pair Encoding (BPE) (Sennrich et al., 2016), and sentence-piece (Kudo and Richardson, 2018) (without any pre-processing) and found the two comparable, and that using larger sentence-piece models improved over smaller ones. For Jpn↔Eng (both di-

---

rections) they first used both in-domain (MTNT) and out-of-domain data (other constrained), and then continued training (fine-tune) using MTNT only. They also reported many results from their hyper-parameter search (albeit without a clear recommendation). The final submission is an ensemble of 4 models.

NaverLabsEurope(NLE)' submission (Bérard et al., 2019): The participants carried substantial effort to clean the CommonCrawl data, applying length filtering (length ratio threshold), language identification-based filtering, and attention based filtering. They used the Transformer-Big architecture for Fra→Eng and Jpn→Eng, and Transformer-Base for the Eng→Jpn direction.

The participants incorporated several methods to encourage robustness (detailed ablations on the effect of each method were not provided). They lowercase all data. However in order to preserve casing information in the input, they propose a technique called *inline casing* which adds additional casing tags (one per non-lowercased subword) in the sequence. Emojis were replaced with a special symbol. Natural noise based on manually defined noise rules was added on the source side of the training data. Lastly, MTNT monolingual data was back-translated to be used during training of the final system. They trained their system on all available data with special tags for each domain and for each data type e.g. real, back-translated, or noisy data. They found that adding tags is as good as fine-tuning the system, allowing for more flexibility at test time. Their final submission with an ensemble of 6 systems for Eng→Jpn and ensembles of 4 systems for the other language directions performed the best in the evaluation campaign.

NICT's submission (Dabre and Sumita, 2019): The authors used Transformer models to train their systems and employed two strategies namely: i) mixed fine-tuning and ii) multilingual models for making the systems robust. The former helps as the in-domain data is available in a very small quantity. Using a mix of in-domain and out-domain data for fine-tuning helps overcome the problem of adjusting learning rate, applying better regularization and other complicated strategies. It is not clear how these two methods contributed towards making the models more robust. According to the authors, mixed fine-tuning and multilingual training (bidirectional) helped. In the error analy-

sis, they found that their system performs poorly in translating emojis. The segmentation errors generated by KyTea resulted in further errors in the translation.

NTT's submission (Murakami et al., 2019): The participants submitted systems for the Eng→Jpn and Jpn→Eng directions in the constrained setting. Their techniques include the placeholder mechanism for copying non-standard tokens (emojis, emoticons, etc), back-translation, fine-tuning on in-domain corpus, and ensemble. Especially, the placeholder mechanism provides +1.4 BLEU and +0.7 BLEU points for Jpn→Eng and Eng→Jpn respectively. Finetuning provides a larger improvement for Eng→Jpn (+1.2 BLEU) than Jpn→Eng (-0.3 BLEU). Their model is Transformer-Base configuration, where they demonstrated its capacity to noise-robustness can be further improved by the above-mentioned techniques.

5 Summary of Methods

In this section, we give a common theme and summary of methods applied by the various participants.

Data Cleaning Data cleaning played an important part in training successful MT systems in this campaign. Unlike other participants, the winning team Naver Labs Bérard et al. (2019) and NTT (Murakami et al., 2019) applied data cleaning techniques in order to filter noisy parallel sentences. They filtered i) identical sentences on source and target side, ii) sentences that belonged to a language other than the source and target language, iii) sentences with length mismatch, and iv) also applied attention-based filtering. Data cleaning gave an improvement of more than 5 BLEU points with substantial reduction in the hallucination of the model for the winning team.

Placeholders Training and test data contained tokens (such as emoticons) which do not require translation. Murakami et al. (2019) and Bérard et al. (2019) preserved these in a preprocessing step using special placeholders and copied them in the translation output. Murakami et al. (2019) reported a gain of up to 1.4 BLEU points by using placeholders.

Data Augmentation Other than handling noisy data, one of the challenges related to this task was

data sparsity. All the participants back-translated in-domain monolingual data and used synthetic data as part of their training pipeline. In addition, Bérard et al. (2019) created a noisy version of all the available in-domain and out-of-domain data by randomly replacing words with their noisy variants. For training, they appended source sentences with a tag <noisy> to distinguish them from the original data. Zhou et al. (2019) used translation systems using placeholders in order to create both clean versions of the noisy in-domain datasets, as well as noisy versions of the clean out-of-domain dataset. To get additional data, other than back-translation, the JHU team (Post and Duh, 2019) used cross-entropy based filtering to select top 1 million sentences from Gigaword, CommonCrawl and the UN corpus. Adding large filtered data gave then an improvement of +5.8 BLEU points.

Domain-aware Training In order to differentiate different data, real from synthetic, in-domain from out-domain, several participants used additional tags. Zheng et al. (2019); Bérard et al. (2019) used domain tags during training to indicate data domain. Bérard et al. (2019) additionally included data type tags (real or back-translated) for further categorization of the training data. Compared to fine-tuning, adding tags provides them additional flexibility, resulting in a generalized system, robust towards a variety of input data.

Fine-tuning Along with the noisy in-domain MTNT data, general domain data typically made available for WMT campaign was also allowed for this task. Most participants (Murakami et al., 2019; Dabre and Sumita, 2019; Helcl et al., 2019) trained on general domain data and fine-tuned the models towards the task. Murakami et al. (2019) did not see a consistent improvement with fine-tuning. Due to the small size of the in-domain data, Dabre and Sumita (2019) fine-tuned on a mix of in-domain and a subset of the out-of-domain data.

Ensembles To benefit from the different trained models and to make the performance more stable, many participants performed ensemble over their models. Murakami et al. (2019), Bérard et al. (2019), Zheng et al. (2019), and Post and Duh (2019) ensembled between 4 and 6 checkpoints of their model for the final submission. They observed a consistent performance improvement over using a single model.

6 Results

In this section we describe quantitative results, and also perform a qualitative analysis of the results.

6.1 Quantitative Results

The quantitative analysis of the submitted systems yields fairly consistent results. On automatic evaluation (BLEU) the best system across all translation directions is the NaverLabsEurope(NLE) one. The same system received also the highest human judgment scores, with the exception of the Eng→Jpn task, where the NTT system was ranked higher. Overall, the correlation between human judgments and BLEU is very high. For Eng→Fra, the Pearson's correlation coefficient is 0.94, while for the other three tasks it is over 0.97.

Human Evaluation The results of human evaluation following the evaluation protocol described in Section 3.4 are outlined in Table 2.

Automatic Evaluation The automatic evaluation (BLEU) results of the Shared Task are summarized in Table 3.

6.2 Qualitative Analysis

In order to discover salient differences between the methods, we performed analysis using compare-mt (Neubig et al., 2019), and present a few of the salient findings below.

Stronger Submissions were Stronger at Everything: The submissions to the track achieved a wide range of BLEU and human evaluation scores. In our analysis we found that the systems at the higher end of the spectrum with regards to BLEU also tended to be the best by most other measures (human evaluation, word F-measure by various frequency buckets, sentence-level scores, etc.). Because of this, we limit our remaining analysis to the top three systems in the Fra→Eng and Eng→Fra tracks, and the top two systems in the Eng→Jpn and Jpn→Eng tracks.

Generalization to Words not in Adaptation Data is Essential: The MTNT corpus provides a small amount of training data that can be used to adapt systems to the task of translating social media. One large distinguishing factor between the best-performing system by Naver Labs Europe (NLE) and the second- or third-place systems was

System	Human judgment scores (RANK)			
	Eng→Fra	Fra→Eng	Eng→Jpn	Jpn→Eng
Constrained				
Baidu+OSU	71.5 (2)	80.6 (3)	–	–
CMU	–	58.2 (6)	–	–
CUNI	66.3 (3)	82.0 (2)	–	–
FOKUS	–	–	–	48.5 (5)
JHU	–	76.3 (4)	58.5 (3)	65.4 (3)
NaverLabs	**75.5** (1)	**85.3** (1)	63.9 (2)	**74.1** (1)
NTT	–	–	**66.5** (1)	71.3 (2)
NICT	–	–	44.7 (4)	49.1 (4)
Unconstrained				
FOKUS	52.5 (4)	62.6 (5)	–	–

Table 2: Average human judgments over all submitted systems (the higher the better). The systems' rank for each translation direction is shown in parentheses. The best system is **highlighted**.
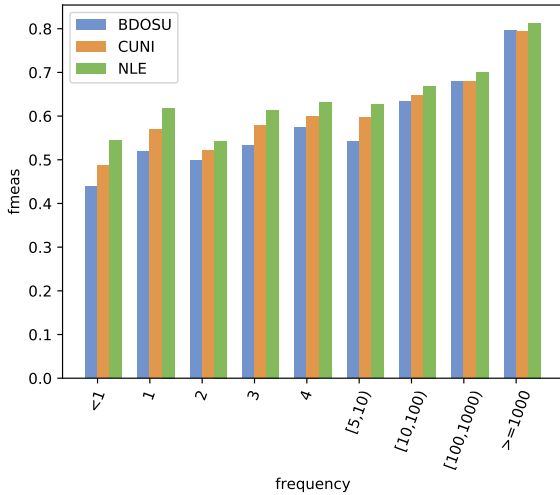

Figure 2: Word F-measure by frequency in the MTNT training data for Fra-Eng.

Figure 3: Word F-measure by casing of the words in the target: all lower-case, title case, all upper-case, or other.

performance on words that were *not* included in this training data that nonetheless appeared in the test set. We show the example of word-level F-measure bucketed by frequency of the words in the MTNT test set for Fra→Eng in Figure 2. From this figure we can see that the NLE system does a bit better in all frequency categories, but the difference is particularly stark for words that appear only once or not at all in the MTNT training set.

Proper Handling of Casing is Important: One other innovation performed by the NLE team was lowercasing of words and separate prediction of casing information. This modeling decision apparently resulted in significantly better results partic-
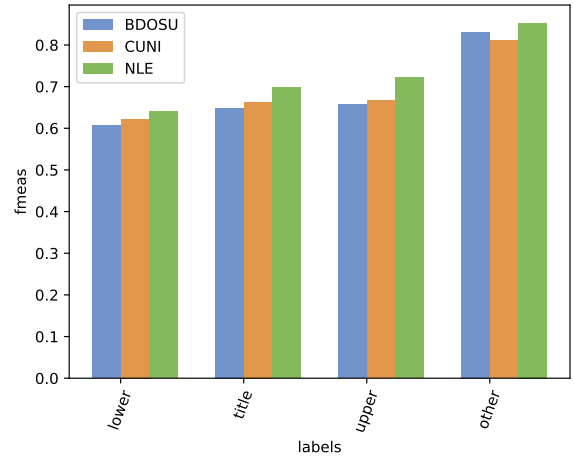
ularly on words that were written in all upper-case, as demonstrated in the results of word F-measure by casing in the target language, demonstrated for Eng→Fra in Figure 3. In addition, we show an example for Fra→Eng in Table 4, where the NLE system translates upper-case characters perfectly, but the CUNI system struggles.

Special Handling of Special Characters is Beneficial: Special characters such as Emojis or symbols were difficult for some systems. Interestingly, even among the top systems, some systems were better at handling different varieties of these characters than others. As an example, in Jpn→Eng, the NTT system performed better on Japanese-style smileys written with standard char-

System	BLEU (RANK)			
	Eng→Fra	Fra→Eng	Eng→Jpn	Jpn→Eng
Baseline	22.1	25.6	8.4	5.8
Constrained				
Baidu+OSU	36.39 (3)	43.59 (3)	–	–
CMU	–	32.25 (5)	–	–
CUNI	38.49 (2)	44.83 (2)	–	–
FOKUS	–	–	–	6.42 (5)
JHU	–	40.24 (4)	14.67 (3)	12.01 (3)
NaverLabs	**41.39** (1)	**47.93** (1)	**17.73** (1)	**16.41** (1)
NTT	–	–	16.86 (2)	14.82 (2)
NICT	–	–	11.09 (4)	7.56 (4)
Unconstrained				
FOKUS	24.22 (4)	29.94 (6)	–	–

Table 3: Automatic evaluation (BLEU, cased) over all submitted systems, with the system's rank in parentheses. The best system is **highlighted**.

	Output	BLEU+1
Ref	From Sri Lanka , to Russia , to the United States , to Japan I mean it 's a market THAT GOES EVERYWHERE .	
CUNI	from sri lanka , to russia , to the united states , to japon I mean it 's a market QUI VA PARTOUT .	33.0
NLE	From Sri Lanka , to Russia , to the United States , to Japan I mean it 's a market THAT GOES EVERYWHERE .	100

Table 4: An example of handling of casing in two Fra→Eng systems

	Output	BLEU+1
Ref	Kawaii ♪ (* ・ ω ・ 人)	
NTT	Cute (* ・ ω ・ 人)	76.0
NLE	It 's cute .	0.0
Ref	👉💥👩♂	
NTT		0.0
NLE	👉💥👩♂	100

Table 5: Examples of translation results on special characters.

acters, while the NLE system performed better on Unicode-standard Emojis, as shown in Table 5.

Non-standard Sentence Structure can be Difficult: Some systems also found sentences with unusual structures, including brackets or other types of punctuation interspersed with actual text, particularly difficult. For example, Table 6 shows an example of Jpn→Eng sentences where the

NTT system had trouble generating the appropriate number of symbols in the appropriate places, while the NLE system was more robust in this regard.

Colloquial Expressions are Key: There was also a marked difference among the top systems in their ability to produce the more informal register reflected in the MTNT test data. We show an example in Table 7 of n-grams that the NTT system was better at producing than the NLE system. All of these are relatively colloquial ways of expressing common function word phrases (1. "is not doing", 2. "but", 3. "lots", 4. "right?", 5. "but,") that can also be expressed with more formal expressions. Clearly the NTT system is producing a slightly less formal register than the NLE system, although a manual examination of the outputs found that even the NTT system was still commonly producing register that was more formal than is commonly found on social media. This may be attributed to the fact that the NTT system

	Output	BLEU+1
Ref	* *] (# mm-e9) [* * Because there 's now protection * *] (# mm-e4)	
NTT	* * * *) (# m-e9) [* * * * * * * * * * * * * * * * * -e4 because there is more protection .)	14.3
NLE	* * * (# mm-e9) [* * Because there is already protection * *] (# mm-e4)	72.0

Table 6: An example of translation results on as sentence with an unusual number of special symbols.

	n-gram	NTT	NLE
1.	ていない	5	0
2.	だけど	4	0
3.	多くの	4	0
4.	ね	3	0
5.	だけど、	3	0

Table 7: Examples of n-grams where one the NTT Eng→Jpn system was more accurate than the NLE system

performed fine-tuning on the MTNT data, moving it towards a more appropriately colloquial register.

7 Conclusions

As a new WMT shared task, this year we focused on building MT systems which are robust to input variations commonly observed in informal language, social media text etc.

From a methodological perspective, the "constrained" setup of the task encouraged participants to leverage both out-of-domain parallel data and in-domain monolingual data to improve performance. Some techniques were utilized by multiple participants and proved their effectiveness in boosting MT models' robustness to noisy input and domain mismatch, including data cleaning, domain-aware training, data augmentation (including backtranslation and copying place-holder tags), finetuning, etc.

In terms of evaluation, we found an automatic metric (BLEU) to be roughly consistent with human judgment. Qualitative analysis found that strong baseline systems were important, but on top of this additional methods specifically aimed at trying to handle various types of noise found in social media text were effective and necessary to further improve within the upper echelons of systems submitted to the shared task.

There are several directions to be explored in the future editions of the task. First, it can exhibit a separate track for "probing" models' robustness so as to understand current models' weaknesses. Second, it could further disentangle improvements for different challenges, e.g., due to noise in training data or due to distribution shift at test time. Controlling the kind of noise introduced, e.g. natural vs. artificial, may be useful in this regard.

Acknowledgements

References

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. In *Proc. NAACL HLT*.

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

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics (TACL)*, 7:49–72.

Alexandre Bérard, Ioan Calapodescu, and Claude Roux. 2019. Naver Labs Europe's Systems for the WMT19 Machine Translation Robustness Task. In *Proceedings of the 2019 Shared task on Machine Translation Robustness, Conference on Machine Translation (WMT)*.

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.

Minhao Cheng, Jinfeng Yi, Huan Zhang, Pin-Yu Chen, and Cho-Jui Hsieh. 2018a. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. *CoRR*, abs/1803.01128.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *ACL*. Association for Computational Linguistics.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018b. Towards robust neural machine translation. *CoRR*, abs/1805.06130.

Raj Dabre and Eiichiro Sumita. 2019. NICT's Supervised MT Systems for the Translation Robustness Task in WMT19. In *Proceedings of the 2019 Shared task on Machine Translation Robustness, Conference on Machine Translation (WMT)*.

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, Volume 1 (Long and Short Papers)*, pages 1504–1516, Minneapolis, Minnesota. Association for Computational Linguistics.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Cristian Grozea. 2019. The submission of FOKUS to the WMT 19 robustness task. In *Proceedings of the 2019 Shared task on Machine Translation Robustness, Conference on Machine Translation (WMT)*.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Georg Heigold, Günter Neumann, and Josef van Genabith. 2017. How robust are character-based word embeddings in tagging and mt against wrod scramling or randdm nouse? *arXiv preprint arXiv:1704.04441*.

Jindřich Helcl, Jindřich Libovický, and Martin Popel. 2019. CUNI System for the WMT19 Robustness Task. In *Proceedings of the 2019 Shared task on Machine Translation Robustness, Conference on Machine Translation (WMT)*.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. *CoRR*, abs/1902.01509.

Huda Khayrallah and Philipp Koehn. 2018. 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, Melbourne, Australia. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

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: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. In *Interpretability and Robustness in Audio, Speech, and Language Workshop Conference on Neural Information Processing Systems*.

Paul Michel, Xian Li, Graham Neubig, and Juan Miguel Pino. 2019. On evaluation of adversarial perturbations for sequence-to-sequence models. In *Proc. NAACL HLT*.

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 (EMNLP)*.

Soichiro Murakami, Makoto Morishita, Tsutomu Hirao, and Masaaki Nagata. 2019. NTT's Machine Translation Systems for WMT19 Robustness Task. In *Proceedings of the 2019 Shared task on Machine Translation Robustness, Conference on Machine Translation (WMT)*.

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

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. 2019. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 35–41, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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

Matt Post and Kevin Duh. 2019. JHU 2019 Robustness Task System Description. In *Proceedings of the 2019 Shared task on Machine Translation Robustness, Conference on Machine Translation (WMT)*.

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

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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. 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 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

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 *Advances in neural information processing systems*, pages 5998–6008.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations*.

Renjie Zheng, Hairong Liu, Mingbo Ma, Baigong Zheng, and Liang Huang. 2019. Robust Machine Translation with Domain Sensitive Pseudo-Sources: Baidu-OSU WMT19 MT Robustness Shared Task System Report. In *Proceedings of the 2019 Shared task on Machine Translation Robustness, Conference on Machine Translation (WMT)*.

Shuyan Zhou, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. 2019. Improving Robustness of Neural Machine Translation with Multi-task Learning. In *Proceedings of the 2019 Shared task on Machine Translation Robustness, Conference on Machine Translation (WMT)*.