Findings of the WMT 2020 Shared Task on Machine Translation Robustness

Lucia Specia[1], Zhenhao Li[1], Juan Pino[2], Vishrav Chaudhary[2], Francisco Guzmán[2]
Paul Michel[3], Graham Neubig[3], Hassan Sajjad[4], Nadir Durrani[4], Yonatan Belinkov[5],
Philipp Koehn [2,6], Xian Li[2]
[1]Imperial College London, [2]Facebook AI, [3]Carnegie Mellon University,
[4]Qatar Computing Research Institute, [5]Harvard University and MIT, [6]Johns Hopkins University

Abstract

We report the findings of the second edition of the shared task on improving robustness in Machine Translation (MT). The task aims to test current machine translation systems in their ability to handle challenges facing MT models to be deployed in the real world, including domain diversity and non-standard texts common in user generated content, especially in social media. We cover two language pairs – English-German and English-Japanese and provide test sets in zero-shot and few-shot variants. Participating systems are evaluated both automatically and manually, with an additional human evaluation for “catastrophic errors”. We received 59 submissions by 11 participating teams from a variety of types of institutions.

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.

Following the first shared task on Machine Translation (MT) Robustness, we now propose a second edition, which aims at testing MT systems' robustness towards domain diversity. Specifically, this year's task aims to evaluate a general MT system's performance in the following two scenarios:

- Zero-shot: the goal is to evaluate a general MT system's performance in unseen domains at test time, which are likely to be different from a training domain (e.g. News, Wikipedia). For that, no domain-specific data or information on the test sets is given to participants.

- Few-shot: the goal is to test an MT system's performance if a few in-domain training examples are provided for the target domain. The question we ask is: can the general MT system leverage those training examples to improve performance on this domain while not dropping its performance on other domains?

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 some trends on approaches used by various systems in Section 4.1. The contributions were evaluated both automatically and via a human evaluation and the results discussed in Section 5.

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

2 Related Work

Domain mismatch is a key challenge in machine translation (Koehn and Knowles, 2017). Most approaches for improving robustness of MT systems to domain shift assume the existence of significant amounts of parallel data in both the source and target domain. In this scenario, a common approach is to first train an MT system on a (generic) source domain and then to fine-tune it on a (specific) target domain (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016; Servan et al., 2016; Chu et al., 2017), to continuously fine-tune on datasets increasingly similar to the target domain (Sajjad et al., 2017), or to dynamically change the balance of data towards the target domain (van der Wees et al., 2017). Another approach trains a system on multiple domains at the same time, while adding domain-specific tags to the input examples (Kobus et al., 2016). Both these approaches were employed by participants of the first shared task on MT robustness (Li et al., 2019).

Other methods for domain adaptation of MT systems include instance weighting (Joty et al., 2015; Wang et al., 2017b), incorporating a domain classifier (Chen et al., 2017; Britz et al., 2017), and data selection (Durrani et al., 2015; Wang et al., 2017a). Some make use of monolingual data available either in the target domain—for example by training the decoder on such data (Domhan and Hieber, 2017) or by backtranslating it (Sennrich et al., 2016)—or in the source domain, via similar techniques (Zhang and Zong, 2016).

Chu and Wang (2018) provide a broad survey of domain adaptation for neural MT, which demonstrates that the predominant setup assumes knowledge of the target domain and availability of target domain data at training time. In light of this prior work, the shared task proposed a relatively underexplored scenario, where examples in the target domain are either unavailable or relatively few.

Other aspects of robustness are robustness to adversarial examples or noisy inputs. The fragility of neural MT models has been previously demonstrated in various settings (Belinkov and Bisk, 2018; Heigold et al., 2017; Durrani et al., 2019; Anastasopoulos et al., 2019; Lee et al., 2018). Michel and Neubig (2018) proposed a new dataset (MTNT) to test MT models for robustness to the types of noise encountered in the Internet. The previous iteration of the shared task focused on robustness of MT systems to such noise (Li et al., 2019). We refer to that report for more details.

3 Task

To facilitate comparability with the News translation task and also to reduce the participation cost, we suggest the same training data as the WMT20 News task.[1] The focus of the Robustness Task is to both evaluate models built on this type of data on more challenging test sets, as well as to encourage participants to explore novel training and modeling approaches so that models have more robust performance at test time on multiple domains, including unseen and diversified domains. We offer two language pairs: English-German (En→De) and English-Japanese (En→Ja), with different test sets focusing on one or both these language pairs, or one particular language direction.

3.1 Phases

The test cycle is divided into two phases. In the first phase – **zero-shot phase**, we release blind test sets from a mixture of domain(s), and participants submit their system's output without any information on these blind domains or training/development data for them. In the second phase – **few shot phase**, we release a small amount of training data (10K sentence pairs) from one of the test domains and participants submit their system's output after utilizing these training examples.

3.2 Training Data

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.

- **Constrained:** Participants can only use the training data made available for this year's News translation task for training. They can use both the parallel data and monolingual data provided in this year's task. Multilingual systems trained with data provided by WMT20 News task are also allowed (and participants should indicate whether this is the case).

- **Unconstrained:** Participants can develop novel solutions to learn from unlabelled data, especially additional monolingual data from domains such as biomedical and/or Reddit. The online systems that we evaluated also fall in this category.

- **Few-shot:** Participants are provided a few in-domain training examples. The data provided consist of the German-English train and valid portions of the CoVoST dataset (deduplicated by source German sentences) and the Japanese-English and English-Japanese

[1] http://www.statmt.org/wmt20/translation-task.html

train and valid portions of the MTNT dataset (Michel and Neubig, 2018).

3.3 Development Data

The task specified the following data to help participants evaluate their system's performance on unseen and multiple domains.

- English-German: participants can use the development data from the News translation task, development data from QED (Abdelali et al., 2014) corpus, development data from WMT19 Medical translation task, and development data from the WMT16 IT translation task.
- English-Japanese: participants can use the development data from the News translation task, and development data from the MTNT dataset, which contains noisy social media texts and their clean translations.

3.4 Test Data

We have three test sets which were created using different sources and approaches. The general statistics are reported in Table 1.

Wikipedia Comments Test Set (set1): This data was collected by Imperial College London and Facebook. We created this to be a particularly challenging test set where the source segments contain various types of linguistic constructs that could lead to what we call *catastrophic errors* in the MT output. For that, we chose user-generated content, namely comments on Wikipedia edits by Wikipedia editors. More specifically, we took English Wikipedia comments from an existing dataset from the Toxic Comment Classification Challenge.[2] The Challenge made available 160,000 comments on Wikipedia edits tagged with multi-grade toxicity labels (toxic, severe toxic, obscene, threat, insult, or identity hate). We believe that the presence of toxic content can be very challenging for MT systems.

After filtering out non-English segments and segments that were too long (>50 words or >1000 characters) or too short (<5 words), we kept all the remaining comments with any toxic label (approx. 7K) and randomly selected 10K non-toxic samples.

Based on this initial selection of 17K English comments, we defined heuristics to further sample from the selection and diversify the potential sources of catastrophic errors. To that end, we first machine translated all comments using an in-house transformer-based model into Japanese and German. The goal of that was to be able to examine potential differences in source and (one example of) translation segments.[3] We then pre-processed and automatically annotated all 17K segments with the following soft labels for catastrophic errors:

1. Introduction of toxicity: we checked both source and machine translation for toxic words (using in-house lists) and labelled as positive (i.e. potentially containing errors) cases where the source does not contain such words, but the translation does (at least one).
2. Mistranslation of named entities: we annotated person, organisation and location named entities in the source and translation (using an in-house named entity recognition model) and labelled as positive cases where (a) the translation has fewer named entities than the source and the translation has at least one toxic word, (b) the translation has at least 2 fewer named entities than the source, and (c) the list of named entity types (e.g. person vs location) in source and translation differ and translation has at least one toxic word.
3. Inversion of sentiment: we applied the Google Cloud Sentiment Analysis tool[4] to annotate each source and machine translation and labelled as positive cases with very different sentiments, i.e. the source is very positive (>0.5) and the translation is very negative (<-0.5) or vice-versa (scores range from -1 (negative) to 1 (positive).
4. Difference in emojis: we detected emojis in the source and machine translation[5] and labelled as positive cases where source and translation have a different number of emojis.

[2] www.kaggle.com/c/jigsaw-toxic-comment-classification-\challenge

[3] We are aware that using one particular translation model can bias the selection to cases that are challenging for this particular model. In future work following this methodology, we recommend that multiple MT models be used.

[4] https://cloud.google.com/natural-language/docs/analyzing-sentiment

[5] https://github.com/carpedm20/emoji/)

5. Presence of idioms: we checked if the source contains idiomatic expressions, using an in-house list of idioms built from various sources, and labelled those cases as positive.

We note that the automatic labelling using our various pre-processing techniques may have introduced errors, but we believe that basing the selection on such heuristics will still lead to higher chances of selecting very challenging source segments than arbitrarily sampling the data.

We divided the original data (toxic and non-toxic 17K) into 5 sets, one for each of these soft labels (allowing for duplicates samples across sets). Finally, we uniformly selected a test set per language pair, containing 1,098 unique segments for English→German and 1,100 unique segments for English→Japanese. We provided the test sets for experiments in both directions, but we will only report results on the original source→target direction. For each of these test sets, we discarded the machine translation and collected reference translations from scratch using professional translators.

Reddit Test Set (set2): This data was collected by Carnegie Mellon University following the same procedure as last year's test set (described in Michel and Neubig (2018)): comments from the social media website `reddit.com` were scraped, filtered for noisy comments and translated by professional translators. This year, data was collected for two translation directions: English→Japanese and Japanese→English. For English, comments were collected from the `/r/all` feed, which encompasses all communities, and filtered for English. Since Japanese is a minority language on Reddit, comments were scraped from a selection of japanese-speaking communities, detailed in Michel and Neubig (2018).

Common Voices Test Set (set3): This data was obtained from from the CoVoST corpus (Wang et al., 2020). CoVoST is derived from Common Voice (Ardila et al., 2020), a crowdsourced speech recognition corpus with an open CC0 license. Transcripts were sent to professional translators and the quality of translations was controlled by automatic and manual checks (Guzmán et al., 2019). For this task, we used the German→English test set with source German sentences deduplicated.

3.5 Evaluation protocol

Automatic evaluation: We first computed BLEU (Papineni et al., 2002) for each system using SacreBLEU (Post, 2018). For all language pairs except En→Ja, we used the original reference and SacreBLEU with the default options. In the case of En→Ja, we used the reference tokenized with KyTea and the option `--tokenize none`.

Human evaluation: 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, as well as the different MT systems, was randomized and not disclosed to the translator. The translators rated both the reference and the translation. We believe that the reference translation in this evaluation setup to serves the purpose of calibration by offering the human annotators one (hopefully) good example of translation. We also report metrics for these reference translations as an upperbound for the data.

We sampled 400 translations from each MT system in each of the test sets and language pairs (28 combinations), resulting in 11,200 segments and their references to be annotated (22,400 segments in total). Each translation/reference segment was annotated by three raters. Quality control was manged by the company providing the ratings, where the main check was that the three ratings could not disagree by more than one category (in which case additional raters are enlisted until agreement is reached).

The rating of translations was done using a different metric from last year's task. Instead of direct assessment (DA), we chose a discrete *likert* rating ranging from 1 to 5, which we found to lead to higher agreement between raters in other annotation projects (Diab et al., 2020). A summary of the guidelines provided for this *likert* rating is as follows:

1 Bad: translation errors are so severe that they cause the target text to be incomprehensible. This may be mainly due to major grammatical, typographical, or lexical errors, or omission of critical or important salient information.

2 Poor: the target text contains translation errors, but these errors do not hinder overall comprehension and do not mistranslate overall intent.

	En→De	De→En	En→Ja	Ja→En
Wikipedia Comments (set1)	1,098 / 26,549	-	1,100 / 29,419	-
Reddit (set2)	-	-	1,376 / 20,011	997 / 20,842
Common Voice (set3)	-	5,609 / 43,119	-	-

Table 1: Number of sentences/words per test set (Japanese words are counted after tokenization with KyTea).

The errors may be mainly due to partial differences in intent, grammatical or typographical errors, or omission of important salient information.

3 Acceptable: the target text is fully comprehensible and fully translated (i.e. no information is omitted), even if it contains minor errors. These errors may be mainly due to partial lack of fluency, or a few grammatical or typographical errors.

4 Very Good: the target text is fully comprehensible, fully correct, and does not miss any information. Style matters may not be transferred faithfully, such as level of formality, or the translation of idioms does not need to be perfect but their meaning needs to be correctly conveyed.

5 Excellent: the target text is fully comprehensible, fully correct, and does not miss any information. Additionally, source style is reflected in the translation and if present, idioms are perfectly handled.

Catastrophic error annotation: As an additional form of human annotation, alongside the *likert* ratings described above, we instructed the annotators to indicate, for translations rated below **3 - poor or bad**, whether they contained any catastrophic errors, and to categorise the type of error. This is a new type of evaluation and we provided detailed guidelines, which we summarise below.

Annotators were asked to provide a **YES/NO flag** to indicate whether the translation contains any error (one or more words) that changes the meaning of the source segment in a critical way. Critical errors are those that lead to misleading translations which may carry religious, health, safety, legal or financial implications, or introduce toxicity. The set of critical errors used for the guidelines (which also included examples of these errors) includes – but is not limited to – the cases below:

- Introduction of toxicity (profanity, violence, hate or abuse) (TOX).
- Introduction of health/safety risks (SAF).
- Mistranslation of named entities (NAM).
- Reverse negation (NEG).
- Reverse of sentiment/polarity (SEN).
- Change in units/time/date/numbers (NUM).
- Other (OTH).

If the answer is YES, annotators were asked select one of the categories indicating the type of critical error. They were asked to choose the category that compromises the meaning of the sentence the most if more than one error was found in the same segment. Three raters flagged and categorized errors.

4 Participants and System Descriptions

We received submissions from 8 teams participating across different tasks, test sets and languages we provided this year. Below we briefly present the systems we were able to get a system description paper for:

Naver Labs (NLE): They participated in Chat and Biomedical tasks along with the Robustness task. They trained a general big-transformer model using *FairSeq* toolkit (Ott et al., 2019) and adapted it towards different tasks using lightweight adapter layers for each task (Bapna and Firat, 2019). They compared results against the more traditional fine-tuning method (Luong and Manning, 2015) to show that the former provides a viable alternative, while significantly reducing the amount of parameters per task. They also explored using embedding from pre-trained language models in their NMT system of which they tried two MLM variants: i) using NMT encoder's setting, using Roberta (Liu et al., 2019). The latter was found more robust to novel domains and noise. The authors found that initializing with first 8 layers instead of the entire model to be optimal. Another notable finding included the use of single bidirectional model instead of monodirectional models to give similar performance. For the robustness task specifically they added source side synthetic noise and used BPE drop-out. While this was found to be useful in handling noisy data, no gains towards domain robustness were observed.

LIMSI: LIMSI participated in Biomedical and Robustness tasks. For the robustness challenge their main exploration was using adapter layers (Bapna and Firat, 2019) applied on 8 domains (parallel data released in the News task). The architecture adds an additional, domain-specific layer on top of every layer of the encoder and the decoder. This allows the test sets from known domains to use adapter layers and for novel domains to use the generic system. They created a noisy domain by adding synthetic noise to source data. The idea is that residual adapter layer learned from such data learns how to deal with noisy domain and is also able to preserve performance on the cleaner domains. However this did not work as well. The residual adapter fine-tuned using the ParaCrawl corpus gave better performance.

e-Translation: Their effort was mainly directed towards the News translation task, however they submitted two systems to the Robustness task. Their general systems were built using big-transformer configuration trained using Marian (Luong and Manning, 2015) after up-sampling original training data. The system was then fine-tuned for another round with an LM scored subset of original data. Finally ensembling four checkpoints produced their final systems. The authors reported an interesting finding that their models performed better on the noisy test sets released for this task than on the standard news test set, suggesting that systems trained on the diverse domains were already robust enough.

UEDIN: Team UEDIN mainly trained their system towards News translation task, but added Gumbel noise to the output layer of the systems submitted to the Robustness task. They followed standard NMT training pipeline and boasted their systems with additional data filtered from the paracrawl corpus. The data was carefully selected using dual cross-entropy (Junczys-Dowmunt, 2018) and length-normalized cross-entropy.

OPPO: Team OPPO also trained their systems for the language pairs released for the News translation task and did not carry any specific exploration towards the task of Robustness. Their systems followed standard training regime of training transformer models with Marian toolkit, with back-translation to generate synthetic data and ensembles of models. As additional module, they added to their system a reranker trained on six forward and backward models, the scores of which are used as features in training the reranker.

PROMPT: Team PROMPT also participated mainly in the News translation task. Their systems were trained using OpenNMT (Klein et al., 2017) toolkit. They applied several stages of data preprocessing including length-based filtering, removing duplications, and using in-house classifier based on `Hunalign` aligner to identify and discard non-parallel sentences. They used two types of synthetic data to improve their systems: i) randomly selecting subset of Wikipedia equal to the size of news data and generating parallel corpus through back-translation, ii) creating synthetic data with unknown words using the procedure described in (Pinnis et al., 2017). Systems were trained with tags to differentiate between original data and synthetic data from each other. Named entities were handled through a post-processing module with redecoding whenever a named entity was not translated or translated incorrectly.

Online systems: We also evaluated three top performing online MT systems, which are also commonly used in the WMT News translation task: online-A, online-B, and online-G. While we do not have access to details of the architectures of these models, to the best of our knowledge they are are all neural MT models with one case including a selection between translations from statistical and neural models.

4.1 Common Trends

Participating systems were trained following a standard recipe, i) using big-transformer models, ii) boasting performance with tagged back-translation, iii) continued training with filtered data and in-domain data (where available), iv) ensembling different models to obtain further improvements. Only two teams, namely Naver Labs and LIMSI made specific efforts towards the task of Robustness. Both of them used lightweight domain adaptors proposed by Bapna and Firat (2019). Both teams also explored making the systems robust by adding noisy synthetic data. While they found using adaptor layers instead of fine-tuning the entire model to be a viable alternative, no success was observed adding noise to the training process.

5 Results

In this section we describe the results of both automatic and manual evaluation of general translation quality (Section 5.1), as well as an analysis of catastrophic errors (Section 5.2).

5.1 General Quality

Overall, the correlation between human judgments and BLEU is not strong. For En→De (set1), the Pearson's correlation coefficient is 0.97, while for the other four tasks the coefficients are lower, with 0.78, 0.65, 0.52, 0.79 for En→De (set1), Ja→En (set2), En→Ja (set2), and De→En (set3) respectively.

Automatic Evaluation The automatic evaluation (BLEU) results of the Shared Task are summarized in Table 2, where we also include the three online translation systems. We performed significance test using `compare-mt` (Neubig et al., 2019) where systems are considered as significantly different at p <0.05. The result of significance test is used for the automatic evaluation ranking.

Overall, the **unconstrained** online-B system provides strong results and outperforms most systems in the five language pairs, except the De→En (set3) and En→Ja (set1).

Among the participating teams, the best **zero-shot** systems were OPPO, which outperforms other zero-shot systems in En→De (set1), Ja→En (set2), and En→Ja (set2) tasks, and NLE, which outperforms other systems in the other two tasks.

Only Naver Labs participated in the **few-shot** stage (NLE-few) and submitted their systems in four language directions except the En→De (set1) subtask. Their few-shot systems ranked the first in all the four directions they participated, tying online-B system in three language pairs.

Human Evaluation The results of human evaluation following the evaluation protocol described in Section 3.5 are outlined in Table 3. The *likert* score is calculated by averaging ratings from the three human annotators over the 400 sampled translations for each MT system, and we performed significance test using the `testSignificance.py` script[6] (Dror et al., 2018) with p <0.05. The result of significance test in *likert* score is used for the human judgement ranking. Interestingly, the correlation in the system rankings between human judgments and BLEU is not strong. In other words, the best performing systems in BLEU do not rank high according human judgement, sometimes even rank the lowest. For example, in Ja→En (set2), the online-B system ranks first in BLEU but last in *likert* score. OPPO outperforms all systems in both directions on set2, and is overall the best system among the **constrained, zero-shot** submissions.

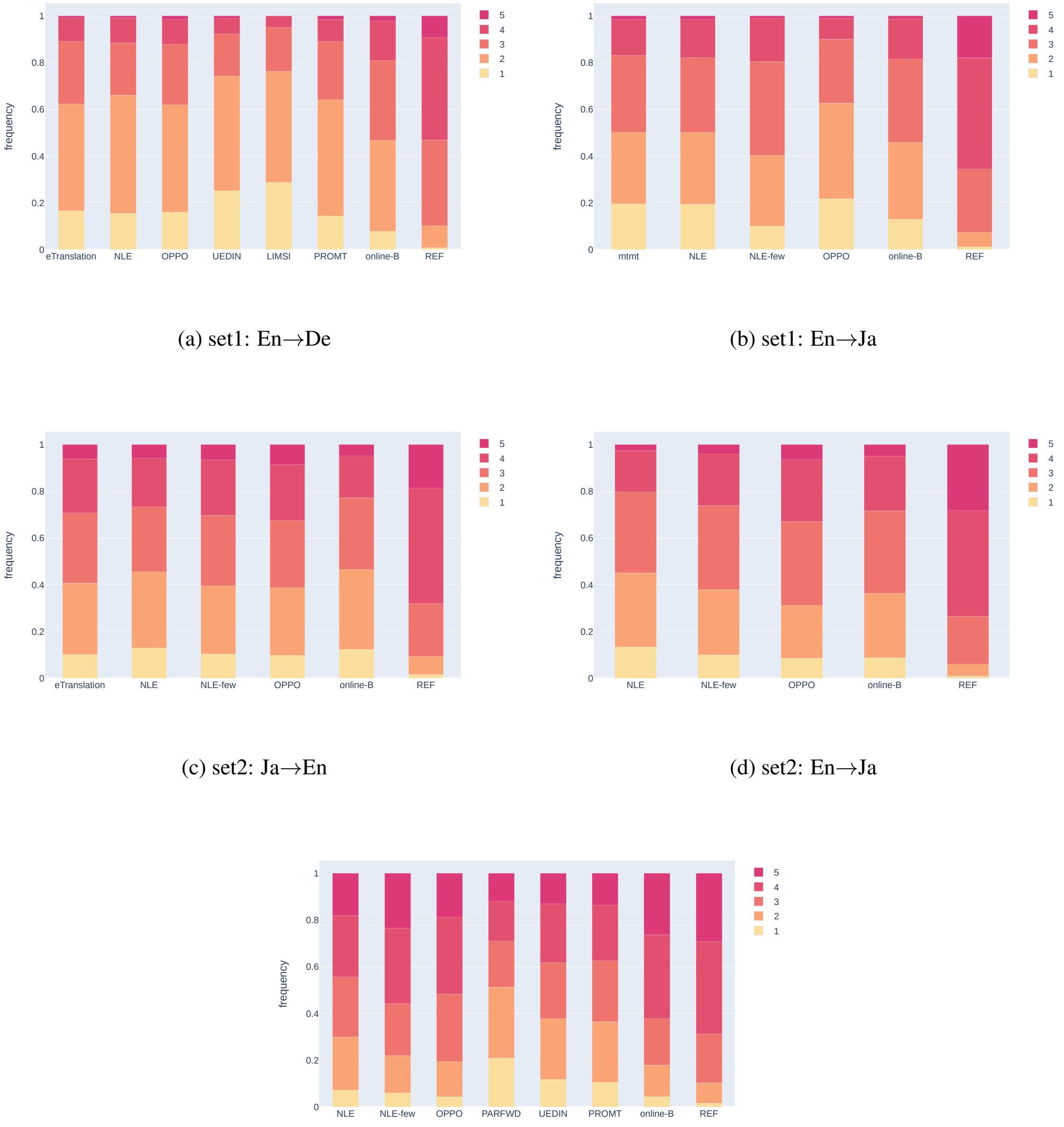
To get insight on the proportion of sentences with each of the categories of human score, Figure 1 displays the distribution of *likert* ratings for all systems. The most frequent ratings for the participating systems are 2 and 3 while for the human-translated references it is 4. Comparing the few-shot and zero-shot systems, the NLE-few outperforms most systems because the frequency of lower ratings (1 or 2) is lower, but the frequency of high ratings (5) is similar to the zero-shot systems.

5.2 Evaluation on Catastrophic Failures

Here we turn our attention to the extra level of annotation where human raters flag and categorise catastrophic errors in sentences. We note that we had three raters for each translation, and that in some cases different categories of errors were flagged. This naturally happened since the raters were asked to choose the category with the biggest negative impact, which is a subjective decision. For example, in En→De (set1), each system has 28 sentences in average flagged with multiple errors. We report this average multi-error counts in Figure 3. In addition, we note that there may also be cases of disagreement, where only a subset of raters flag errors (we will perform agreement analysis later).

Error rate of systems Table 3 shows the proportion of sentences containing as least one error in (which we will refer as "error rate"). The error rates vary among different test sets. Regarding set1, which is sourced from Wikipedia comments, oversampling for more challenging content, the error rate for different systems is high, ranging from 51% to 76%. It is interesting that annotators indicate that the human-translated references contain catastrophic errors as well, with an error rate of 23% for both language pairs in set1. The error rate in

[6] `https://github.com/rtmdrr/testSignificanceNLP`



(a) set1: En→De

(b) set1: En→Ja

(c) set2: Ja→En

(d) set2: En→Ja

(e) set3: De→En

Figure 1: Distribution of *likert* ratings for all submitted systems (the darker the color, the higher the ratings - higher quality).

System	BLEU (RANK)				
	set1		set2		set3
	En→De	En→Ja	Ja→En	En→Ja	De→En
Constrained					
eTranslation	41.9 (3)	–	13.9 (2)	–	–
mtmt	–	18.2 (5)	–	–	–
NLE	42.2 (4)	22.5 (3)	13.3 (2)	16.2 (3)	44.7 (2)
NLE(FEW)	–	**25.4** (1)	**15.3** (1)	18.4 (1)	**45.4** (1)
OPPO	42.9 (2)	19.1 (5)	15.2 (1)	17.3 (2)	43.3 (3)
PARFWD	–	–	–	–	30.8 (5)
UEDIN	35.1 (7)	–	–	–	43.8 (3)
LIMSI	30.2 (8)	–	–	–	–
Unconstrained					
PROMT	41.4 (5)	–	–	–	41.4 (4)
online-A	38.6 (6)	23.1 (2)	13.6 (2)	17.8 (2)	43.2 (3)
online-B	**48.0** (1)	**25.4** (1)	14.3 (1)	**18.8** (1)	44.3 (2)
online-G	37.9 (7)	20.4 (4)	9.4 (3)	14.8 (3)	43.4 (3)

Table 2: Automatic evaluation (corpus-level BLEU, cased) over all submitted systems, with the system's rank in parentheses ($p < 0.05$). Bold highlights the system with highest BLEU score.

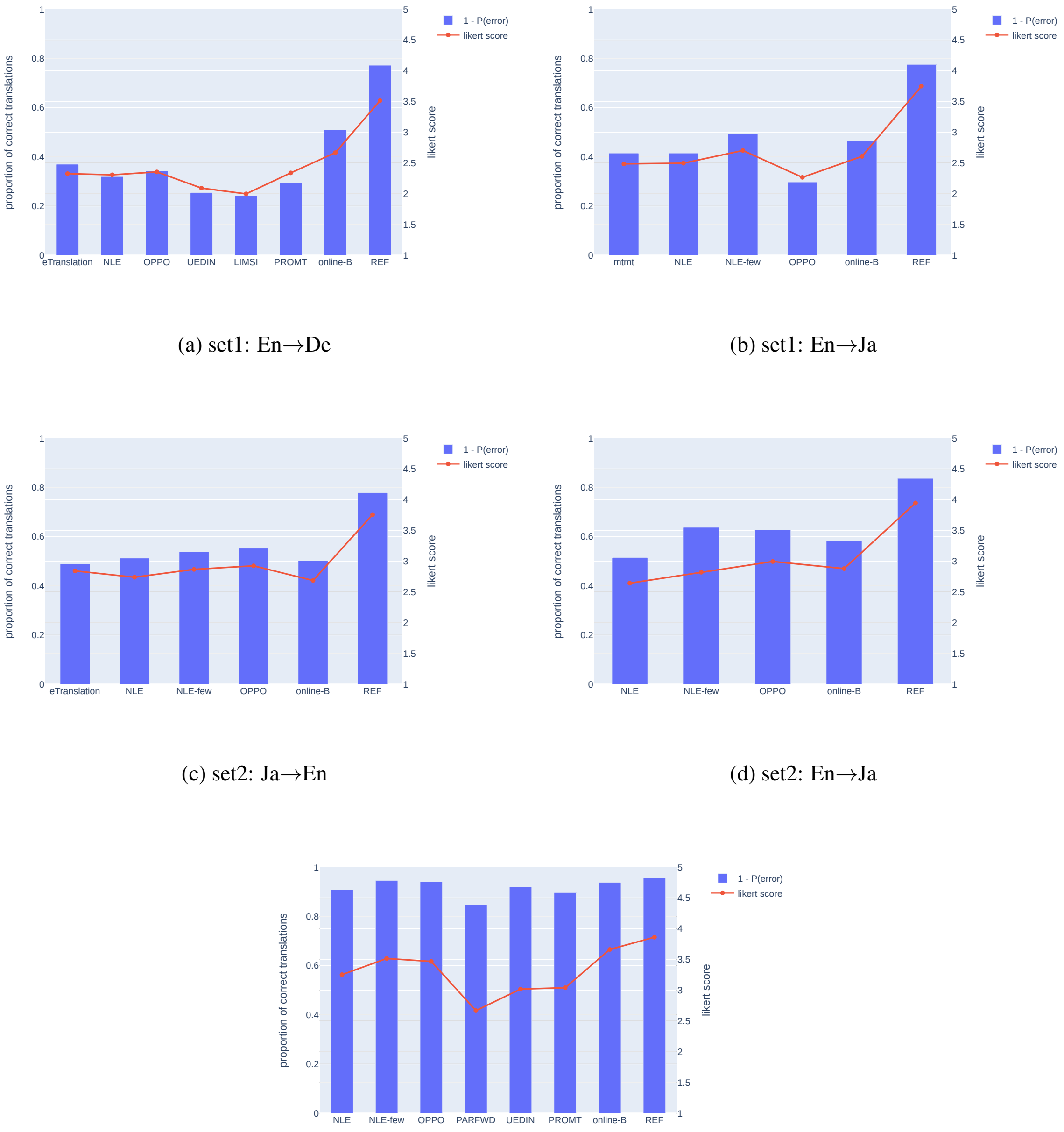
set2, sourced from Reddit, is lower, which is within 36%-51% for participating systems and 16%-18% for the references. In set3, which is sourced from Common Voice data, the error rate is the lowest. All systems except one achieve less than 10% error rate. The issue of catastrophic errors in the reference translations needs further investigation. We speculate that this could be due to misinterpretation of the guidelines, as we discuss below.

The error rate is highly correlated with the *likert* score reported in Section 5.1. We show in Figure 2 the relation of the proportion of translations without catastrophic errors (blue bars) and the *likert* scores (red lines). As expected, systems with more translations without errors get higher *likert* scores. The Pearson's correlation coefficient for De→En (set3) is 92%, while for the other four language pairs, the coefficients are over 96%.

Distribution of error types In Figure 3 we show the absolute counts and proportion of different types of catastrophic errors per system. We note that some sentences may have been annotated with more than one error type (by different human annotators), and therefore the counts may seem inflated. To provide a better idea of the distribution of errors, for each system the error proportion is calculated as the number of translations with certain error divided by the number of sampled translations, i.e. 400. In all five language pairs, the OTH error is the main source of catastrophic errors, however, this OTH error is not clearly defined and might indicate different translation errors, e.g. some translations simply copy the source sentence and are therefore labelled as OTH error. This requires further analysis.

Excluding the OTH error (Figure 4), the catastrophic error distribution varies in different subtasks. Named entities (NAM) account for a large proportion of errors in all subtasks except En→De (set3). In En→De (set1), Ja→En (set2), and De→En (set3) subtasks, sentiment (SEN) errors are very frequent, similar to NAM errors. The TOX error is predominant only in En→Ja subtask. Other types of catastrophic errors occupy much smaller proportion.

This figure also highlights the different catastrophic error types flagged for reference translations. While this needs further inspection and investigation, we suspect that annotators might have misinterpreted the guidelines. For example, in the Wikipedia comments En→Ja, there is a large proportion of sentences with catastrophic errors of the type "toxic" (TOX): almost 10% of the reference translations contain such error type. Translations (human or machine) containing toxic content might have been tagged as containing errors, even though the source segments also contained such toxic content and the translation is simply transferring it.



(a) set1: En→De

(b) set1: En→Ja

(c) set2: Ja→En

(d) set2: En→Ja

(e) set3: De→En

Figure 2: Proportion of translations without any error (bars) and *likert* over all submitted systems (red points/line).

System	*likert* score / error rate (RANK)				
	set1		set2		set3
	En→De	En→Ja	Ja→En	En→Ja	De→En
Constrained					
eTranslation	2.33 / 63% (2)	–	2.84 / 51% (1)	–	–
mtmt	–	2.49 / 59% (3)	–	–	–
NLE	2.31 / 69% (2)	2.50 / 59% (3)	2.74 / 49% (2)	2.64 / 49% (3)	3.25 / 9% (3)
NLE(FEW)	–	**2.70** / 51% (1)	2.87 / 46% (1)	2.82 / 36% (2)	3.51 / 6% (2)
OPPO	2.36 / 66% (2)	2.27 / 70% (4)	**2.93** / 45% (1)	**3.00** / 37% (1)	3.47 / 6% (2)
PARFWD	–	–	–	–	2.67 / 15% (5)
UEDIN	2.09 / 75% (3)	–	–	–	3.02 / 8% (4)
LIMSI	2.00 / 76% (4)	–	–	–	–
Unconstrained					
PROMT	2.34 / 71% (2)	–	–	–	3.04 / 10% (4)
online-B	**2.67** / 49% (1)	2.61 / 54% (2)	2.69 / 50% (2)	2.88 / 42% (2)	**3.66** / 6% (1)
Reference	3.51 / 23 %	3.75 / 23%	3.76 / 18%	3.95 / 16%	3.86 / 4%

Table 3: Average human judgments and catastrophic error translation rates over all submitted systems and the reference translations (p <0.05). The systems' rank for each translation direction is shown in parentheses. The best system is **highlighted**.

However, this would not explain other error types, which are defined in terms of mistranslation or mismatches between source and target content, such as incorrect named entity translation (NAM). We will analyse the data for that, as well as make it available.

6 Conclusions

The second edition of this WMT shared task focused on testing MT systems in more challenging conditions than last year, in two ways: (i) by making this in a zero-shot setting, where no training set and no in-domain development set were provided, (ii) by biasing the selection of the test sets to make them even harder to translate, for example, by oversampling segments with toxic content. We hoped to encourage participants in the other WMT translation tasks to submit to this task.

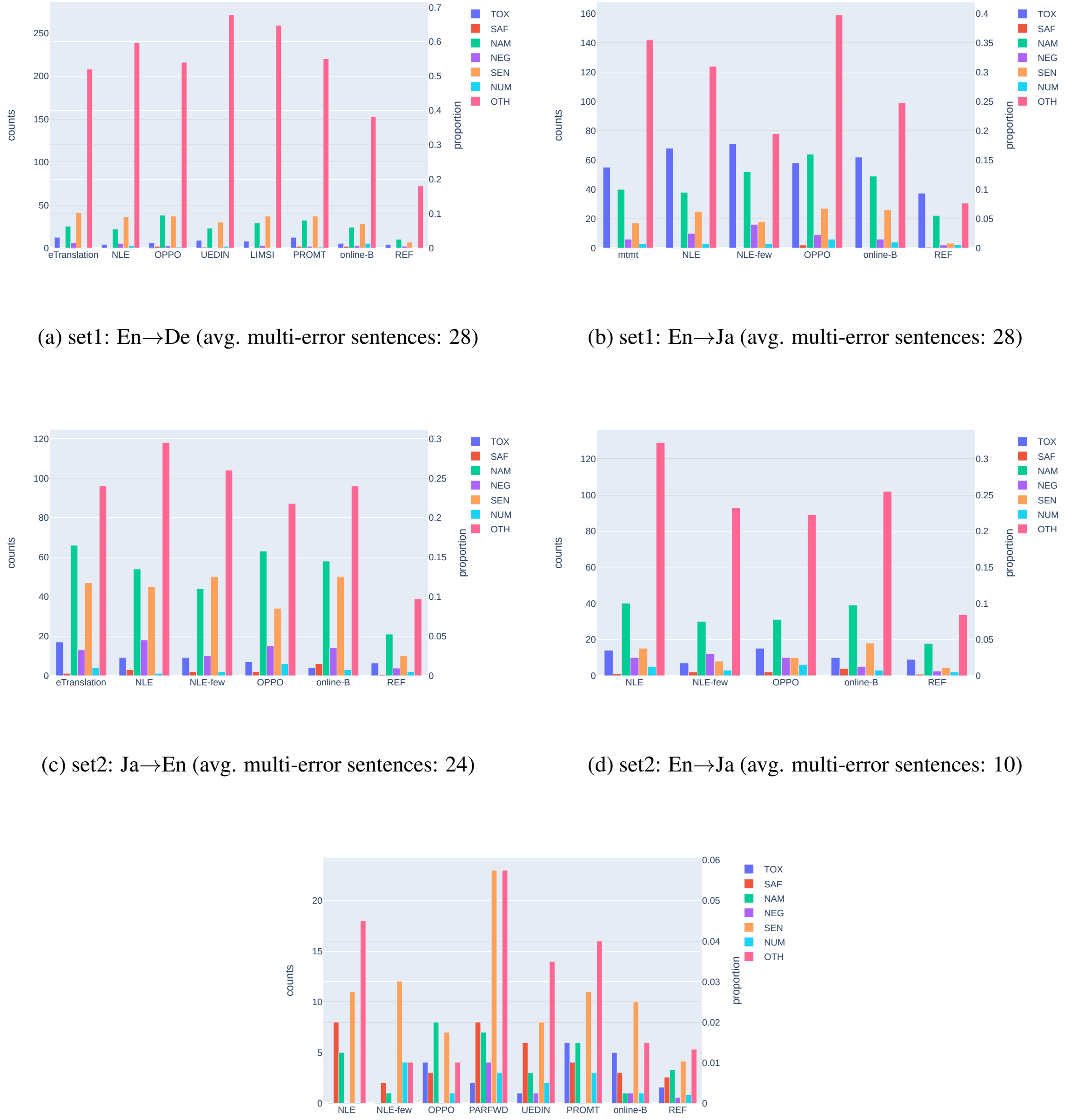
Indeed, most participating teams submitted standard NMT models trained on other types of data and other WMT tasks. Very few teams introduced specific techniques for robustness, such as augmenting training data with synthetic noise. Perhaps not entirely surprisingly, strong online systems, which are trained on a large variety of text types and domains, performed well according to both automatic and human evaluation. The only few-shot submission, however, managed to outperform online systems in most test sets, even in those from a different domain from the small training set provided. This is an interesting outcome and shows that few-shot settings are promising.

A new protocol was used for human evaluation: for general quality, direct assessment was replaced by *likert* scores with more detailed guidelines. The ranking of systems according to this human evaluation does not always agree with that given by BLEU, which is not surprising. According to human evaluation, systems were ranked together more often.

In addition to general quality, we also introduced a flag for catastrophic errors, which is a novel way to evaluate translations. The proportion of sentences containing such errors seems a lot higher than expected. This could be an artefact of the perception of human annotators on what constitutes a catastrophic error. This would explain why even the reference translations are found to contain such errors, albeit on a much smaller scale. In future work we will carry out in depth analysis on the annotation to investigate this high number of catastrophic errors in human and machine translations.

Acknowledgements

Lucia Specia was supported by funding from the Bergamot project (EU H2020 Grant No. 825303). We thank Facebook for funding the human evaluation. We thank Khetam Al Sharou for her help with



(a) set1: En→De (avg. multi-error sentences: 28)

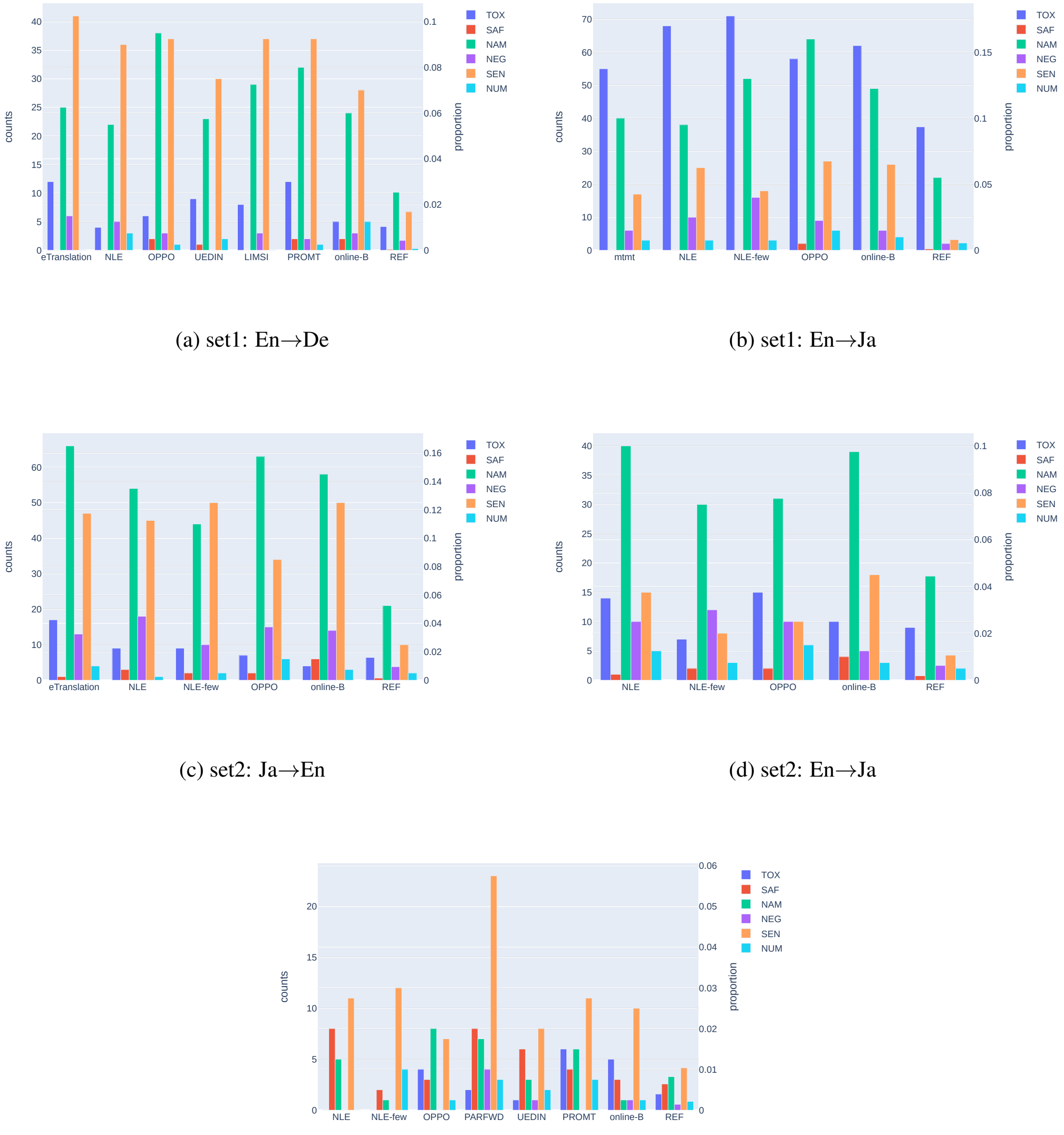
(b) set1: En→Ja (avg. multi-error sentences: 28)

(c) set2: Ja→En (avg. multi-error sentences: 24)

(d) set2: En→Ja (avg. multi-error sentences: 10)

(e) set3: De→En (avg. multi-error sentences: 4)

Figure 3: Distribution of different types of catastrophic errors for all systems: Absolute count or each error type per system, as well as proportion of sentences in each system that contain that error. The average number of sentences labelled with multiple errors per system is reported in parentheses.



(a) set1: En→De

(b) set1: En→Ja

(c) set2: Ja→En

(d) set2: En→Ja

(e) set3: De→En

Figure 4: Distribution of different types of catastrophic errors for all systems **excluding OTH**: Absolute count or each error type per system, as well as proportion of sentences in each system that contain that error.

guidelines for catastrophic errors.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland.

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

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-Âmultilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

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

Denny Britz, Quoc Le, and Reid Pryzant. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126, Copenhagen, Denmark. Association for Computational Linguistics.

Boxing Chen, Colin Cherry, George Foster, and Samuel Larkin. 2017. Cost weighting for neural machine translation domain adaptation. In *Proceedings of the First Workshop on Neural Machine Translation*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of simple domain adaptation methods for neural machine translation. *CoRR*, abs/1701.03214.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mona Diab, Denise Diaz, Ahmed Kishky, Anh Ngo, Ashley Chen, Paco Guzman, and Cynthia Gao. 2020. Rethinking direct assessment machine translation evaluation protocols for user-generated data: A comparative study. In *preparation*.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1500–1505, Copenhagen, Denmark. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

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.

Nadir Durrani, Hassan Sajjad, Shafiq Joty, Ahmed Abdelali, and Stephan Vogel. 2015. Using joint models for domain adaptation in statistical machine translation. In *Proceedings of the Fifteenth Machine Translation Summit (MT Summit XV)*, Florida, USA. AMTA.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *CoRR*, abs/1612.06897.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

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 scramlbing or randdm nouse? *arXiv preprint arXiv:1704.04441*.

Shafiq Joty, Hassan Sajjad, Nadir Durrani, Kamla Al-Mannai, Ahmed Abdelali, and Stephan Vogel. 2015. How to avoid unwanted pregnancies: Domain adaptation using neural network models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1259–1270, Lisbon, Portugal. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Catherine Kobus, Josep Maria Crego, and Jean Senellart. 2016. Domain control for neural machine translation. *CoRR*, abs/1612.06140.

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

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*.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the first shared task on machine translation robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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*, Da Nang, Vietnam.

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)*.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, 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.

Marcis Pinnis, Rihards Krislauks, Daiga Deksne, and Toms Miks. 2017. Neural machine translation for morphologically rich languages with improved subword units and synthetic data. In *Text, Speech, and Dialogue - 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings*, volume 10415 of *Lecture Notes in Computer Science*, pages 237–245. Springer.

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.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Yonatan Belinkov, and Stephan Vogel. 2017. Neural machine translation training in a multi-domain scenario. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

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.

Christophe Servan, Josep Maria Crego, and Jean Senellart. 2016. Domain specialization: a post-training domain adaptation for neural machine translation. *CoRR*, abs/1612.06141.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. 2020. Covost: A diverse multilingual speech-to-text translation corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. European Language Resources Association.

Rui Wang, Andrew Finch, Masao Utiyama, and Eiichiro Sumita. 2017a. Sentence embedding for neural machine translation domain adaptation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*.

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

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the the Conference on Empirical Methods in Natural Language Processing*.

Jiajun Zhang and Chengqing Zong. 2016. 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, Austin, Texas. Association for Computational Linguistics.