NRC Systems for the 2020 Inuktitut–English News Translation Task

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and **Patrick Littell**
National Research Council Canada
{Rebecca.Knowles, Darlene.Stewart, Samuel.Larkin, Patrick.Littell}@nrc-cnrc.gc.ca

Abstract

We describe the National Research Council of Canada (NRC) submissions for the 2020 Inuktitut–English shared task on news translation at the Fifth Conference on Machine Translation (WMT20). Our submissions consist of ensembled domain-specific finetuned transformer models, trained using the Nunavut Hansard and news data and, in the case of Inuktitut–English, backtranslated news and parliamentary data. In this work we explore challenges related to the relatively small amount of parallel data, morphological complexity, and domain shifts.

1 Introduction

We present the National Research Council of Canada (NRC) Inuktitut–English[1] machine translation (MT) systems in both translation directions for the 2020 WMT shared task on news translation.

Inuktitut is part of the dialect continuum of Inuit languages, the languages spoken by Inuit, an Indigenous people whose homeland stretches across the Arctic. Included in this continuum are Indigenous languages spoken in northern Canada, including but not limited to the Territory of Nunavut. The term *Inuktut* is used by the Government of Nunavut (2020) to describe Inuit languages spoken in Nunavut, such as Inuktitut and Inuinnaqtun. The majority of the Inuit language text provided for the shared task comes from ᓄᓇᕗᒻᒥ ᒪᓕᒐᓕᐅᖅᑕᐊ (Nunavut Maligaliurvia; Legislative Assembly of Nunavut) through the Nunavut Hansard, the published proceedings of the Legislative Assembly of Nunavut. The Nunavut Hansard is released publicly by the Government of Nunavut in Inuktitut and English (also an official language of Nunavut), and with their generous assistance was recently processed and released for use in building MT systems (Joanis et al., 2020).[2]

In this work, we examined topics related to morphological complexity and writing systems, data size, and domain shifts. Our submitted systems are ensembled domain-specific finetuned transformer models, trained using Nunavut Hansard and news data and, in the case of Inuktitut–English, backtranslated news and parliamentary data. We measured translation performance with BLEU (Papineni et al., 2002),[3] metrics specific to the production of Roman text in Inuktitut, and human evaluation (to be reported). We hope that human evaluation will provide insight as to whether the current state of the art is sufficient to start building computer aided translation tools of interest and use to Inuit language translators, or whether more work is needed to make the systems usable.

2 Related Work and Motivation

Initial experiments on building neural machine translation (NMT) systems for Inuktitut–English using the most recent Nunavut Hansard corpus are reported in Joanis et al. (2020). Earlier work includes Micher (2018) and Schwartz et al. (2020), and, predating the recent wave of NMT, Martin et al. (2003), Martin et al. (2005), and Langlais et al. (2005). There has also been work on morphological analysis of Inuktitut, including Farley

[1] Abbreviated `iu` and `en` using ISO 639-2 codes.

[2] Though we note that Inuktitut, Inuinnaqtun, English, and French may all be spoken in the House, we use the term Inuktitut in describing our MT systems for two main reasons: 1) the official website describes the Nunavut Hansard as being published "in both Inuktitut and English" (Legislative Assembly of Nunavut, 2020) and 2) because we wish to make clear the limitations of our work; there is no reason to expect that the systems built using the data provided for WMT will perform well across various Inuit languages and dialects (or even across a wider range of domains).

[3] Computed using `sacreBLEU` version 1.3.6 (Post, 2018) with mteval-v13a tokenization: `BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a`.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 156–170
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

(2009) and Micher (2017). In this work, we focus mainly on approaches that are not language-specific, but that are motivated by specific challenges of translating relatively low-resource, morphologically complex languages; thus they are also not entirely language-agnostic.

2.1 Language Typology and Writing Systems

Inuit languages are highly morphologically-complex; many Inuktitut words consist of a large number of morphemes, and can translate to entire phrases or clauses in English (Mallon, 2000; Micher, 2017; Joanis et al., 2020).[4]

Moreover, these morphemes are not easily segmented from one another, as they exhibit phonological changes at morpheme boundaries. That is to say, a given morpheme may be spelled in a number of different ways (or may even appear to merge with a neighbouring morpheme) depending on the morphemes adjacent to it. This means that automatic segmentation approaches may not be optimal. Nevertheless we try using them, and see if we can mitigate some of those challenges via experiments on joint vs. disjoint vocabularies and inserting noise into the segmentation process.

English is written in the Roman script (ISO 15924: Latn), while the Inuktitut data used for this task is primarily written in syllabics (ISO 15924: Cans).[5] There is some Roman text in the Inuktitut side of the data and some syllabics text in the English side of the data, though the former is much more common than the latter.

2.2 Domains and Recency Effects

The Inuktitut–English training corpus released for WMT 2020 consists of parliamentary transcriptions and translations from the Legislative Assembly of Nunavut (Joanis et al., 2020), while the development and test sets are a mix of parliamentary text and news text, the latter drawn from Nunatsiaq News.[6] These two domains are quite different from one another, and in our initial baseline experiments (training only on parliamentary data), we observed very low BLEU scores when translating news data. As we wished to build a constrained system, our only source of Inuktitut news was the data in the development set. In order to retain the ability to use news data in the development and test sets, we utilized an approach of dividing the news development data into thirds, including a third in the training set, using a third as part of the validation set, and holding the remaining third out as test.

The Nunavut Hansard is known to exhibit recency effects, i.e., when testing on a recent subset of the corpus, training on a recent subset is better than training on an early subset (Joanis et al., 2020). Although we have not fully examined the reasons behind this, it could be due to topic shift, a shift in the named entities in the corpus, changes in transcription and translation practices, or any combination of these and more.

We consider tagging domain as one approach to this. Sennrich et al. (2016a) use side constraints (tags) in order to control English–German MT output formality. Yamagishi et al. (2016) add information about voice (e.g. active/passive) to Japanese–English translation via source-side tags. Johnson et al. (2017) also use tags at the start of source sentences, in their case to indicate what language the multilingual translation system should translate into.[7] One might consider domain to fall somewhere between these use cases; Kobus et al. (2017) use domain tags to influence translation in a multi-domain setting. Caswell et al. (2019) use tags to indicate when data has been backtranslated.

3 Data

While the parallel text size for this language pair is quite small compared to high-resource language pairs in the news translation task, Inuktitut is one of the few Indigenous languages in Canada (or possibly the only) for which there exists enough parallel text (with any other language) to train robust statistical or NMT systems outside of the strictly low-resource paradigm (Littell et al., 2018). Thus we expect that it should be helpful to incorporate available monolingual data.

We trained our baseline models using the full 1.3 million line Nunavut Hansard 3.0 (NH) parallel corpus. For IU-EN, we also used a random subselection of 1.3M sentences of English Europarl v10 (Koehn, 2005; Tiedemann, 2012) and 1.3M sentences of English 2019 News data[8] backtranslated into Inuktitut (Section 5.4). We did not use

[4] The Inuktut Tusaalanga website provides an overview of grammar: https://tusaalanga.ca/node/1099

[5] A number of different writing systems, including both syllabics and Roman orthography, are used to write Inuit languages. Inuit Tapiriit Kanatami (ITK) is in the process of creating a unified writing system (Inuit Tapiriit Kanatami, 2020).

[6] https://nunatsiaq.com/

[7] Wang et al. (2018) add a target-side tag.

[8] From the WMT 2020 task page: https://www.statmt.org/wmt20/translation-task.html

Wiki Titles or Common Crawl Inuktitut data.[9] We incorporated the news portion of the development data in training our models to alleviate the domain mismatch issue (Section 5.1).

4 Preprocessing and Postprocessing

We first applied an internal script to convert control characters to spaces as well as normalizing spaces and hyphens; this was effectively a no-op for the Nunavut Hansard parallel corpus, but removed some problematic characters in the monolingual training data. Parallel training corpora were then cleaned with the Moses `clean-corpus-n.perl` script (Koehn et al., 2007), using a sentence length ratio of 15:1 and minimum and maximum lengths of 1 and 200, respectively. For monolingual training data, the second cleaning step consisted of removing empty lines. For Inuktitut, we used the `normalize-iu-spelling.pl` script provided by the organizers.

We then performed punctuation normalization. This included specific `en` and `iu` normalization scripts, to more accurately capture and retain information about directional quotation marks, different types of dashes, and apostrophes, normalizing to the most common form. For Inuktitut, this included treating word-internal apostrophes as `U+02BC` MODIFIER LETTER APOSTROPHE.[10] Appendix C provides a detailed description. After this preliminary normalization, we applied the Moses `normalize-punctuation.perl` script, with the language set to `en` (or backing off to `en`, as there are currently no Inuktitut-specific rules implemented).

Having noted that some of the lines in the training data contained more than one sentence (which results in unintended tokenization behavior), we next performed sentence splitting using the Portage sentence splitter (Larkin et al., 2010) on each side of the training data before feeding it to the Moses tokenizer (using aggressive hyphen splitting). Sentences that had been split were then re-merged following tokenization.

We trained joint byte-pair encoding (BPE; Sennrich et al., 2016c) models on the full Nunavut Hansard parallel training data using `subword-nmt`, then extracted English and Inuktitut vocabularies separately.[11] Using a joint BPE model improves performance on Roman text in Inuktitut output (Section 5.2 and Appendix B).

As postprocessing, we de-BPE the data, run the Moses detokenizer, and then convert the placeholder tokens from our normalization scripts to their corresponding symbols (dashes, apostrophes, quotation marks, etc.).[12]

5 Experiments

All models were typical transformers (Vaswani et al., 2017) with 6 layers, 8 attention heads, network size of 512 units, and feedforward size of 2048 units, built using Sockeye (Hieber et al., 2018) version 1.18.115. We have changed the default gradient clipping type to absolute, used the whole validation set during validation, an initial learning rate of 0.0001, batches of ∼8192 tokens, and maximum sentence length of 200 tokens. We have optimized for BLEU. Custom checkpoint intervals have been used during training, with final systems using between 2 and 11 checkpoints per epoch, consistent within sets of experiments (e.g., vocabulary size sweeping). For finetuning, the checkpoint interval is set to 9, resulting in about 2 checkpoints per epoch for news and 13 for Hansard. For finetuning, we used an initial learning rate of 0.00015 (decreasing by a factor of 0.7 if there was no improvement after 8 checkpoints). Decoding was done with beam size 5.

In the following sections, we describe the experiments that led to our submitted systems. Our final systems were trained on a mix of news and Hansard data (Section 5.1), using joint BPE (Section 5.2), BPE-dropout (for EN-IU; Section 5.3), tagged backtranslation (for IU-EN; Section 5.4), finetuning (Section 5.5), ensembling, and the use of domain-specific models (Section 5.6).

[9] Appendix E provides additional detail about noise and other concerns with the Common Crawl data.

[10] The apostrophe sometimes represents a glottal stop, so when it appeared between syllabic characters, we treated it as a letter that should not be tokenized.

[11] When extracting the BPE vocabulary (which we then used consistently for all experiments) and when applying the BPE model, we used a glossary containing the special tokens produced in preprocessing, Moses special tokens, and special tags (Section 5.4), to ensure they would not be split.

[12] During the task test period, we noted that the test data contained spurious quotation marks, wrapping some entire sentences. After notifying the organizers and confirming that those were produced in error, we handled them as followed: removed the straight quotes that surrounded complete lines, preprocessed, translated, and postprocessed the text that had been contained inside of them, and then reapplied the quotes to the output. There is not an exact match between the source and target for these spurious quotes, so this approach is effective but *not* an oracle.

5.1 Training and Development Splits

In baseline experiments, training only on the Nunavut Hansard training data provided, we noted a major difference in BLEU scores between the Hansard and news portions of the development set. While BLEU scores should not be compared directly across different test sets, the magnitude of this difference (in the EN-IU direction, BLEU scores in the mid-20s on Hansard and in the mid-single digits on news) and the knowledge of differences between parliamentary speech and the news domain suggested that there was a real disparity, likely driven by train/test domain mismatch.

To test this we divided the news portion of the development set in half, maintaining the first half as development data, and adding the second half to the training corpus. Adding half the news nearly doubled the BLEU score on the held out half of the news data, if we duplicated it between 5 and 50 times (to account for how much more Hansard data was available).[13] Initial experiments on vocabulary types and sizes were performed in this setting (Section 5.2).

For the remainder of our experiments, we switched to a setting where we divided the news data into three approximately equally sized thirds; to maintain most documents separate across splits, we split the data into consecutive chunks. Most experiments were run with the first third added to training data, the second third as part of the development set alongside the Hansard development set, and the final third as a held-out test set. This permitted additional experiments on finetuning (Section 5.5) with a genuinely held-out test set.[14] For our final systems, we ensembled systems that had been trained on each of the thirds of the news development data.

5.2 BPE

Ding et al. (2019) highlight the importance of sweeping the number of subword merges (effectively, vocabulary size) parameter, particularly in lower-resource settings. We swept a range of disjoint BPE size pairs (see Appendix A for details of vocabulary size and sweep), and found that disjoint 1k vocabularies performed well for IU-EN, while the combination of disjoint 5k (EN) and 1k (IU) vocabularies performed well for EN-IU (on the basis of averaged Hansard development and news development BLEU score).

As noted in Section 2.1, the Inuktitut data is written in syllabics. However, it contains some text in Roman script, in particular, organization names and other proper nouns. Over 93% of the Roman tokens that appear in the Inuktitut development data also appear in the corresponding English sentence. The ideal behavior would be for a system to copy such text from source to target. When the BPE vocabulary model is learned jointly the system can learn a mapping between identical source and target tokens, and then learn to copy. When the vocabulary is disjoint, there may not be identical segmentations for the system to copy, posing more of a challenge. Appendix B provides details of our experiments on joint vocabulary for successfully producing Roman text in Inuktitut output.

Due to the similarity in BLEU scores, and for simplicity and consistency, the remainder of our experiments *in both directions* were performed with jointly learned (and separately extracted) BPE vocabularies. We experimented with joint BPE vocabulary sizes of 1k, 2k, 5k, 10k and 15k.

5.3 BPE-Dropout

Knowing that the morphology of Inuktitut may make BPE suboptimal, we chose to apply BPE-dropout (Provilkov et al., 2020) as implemented in `subword-nmt` in an attempt to improve performance. BPE-dropout takes an existing BPE model, and when determining the segmentation of each token in the data randomly drops some merges at each merge step. The result is that the same word may appear in the data with multiple different segmentations, hopefully resulting in more robust subword representations. Rather than modifying the NMT system itself to reapply BPE-dropout during training, we treated BPE-dropout as a preprocessing step. We ran BPE-dropout with a rate of 0.1 over both the source and target training data 5 times using the same BPE merge operations, vocabularies and glossaries as before, concatenating these to form a new extended training set.[15]

In our initial baseline experiments (without

[13] Adding all of the data would not have allowed us to evaluate the outcome on news data, and not including any news data in the development set also hurt performance.

[14] An alternative approach would be to select pseudo in-domain data from the Hansard, by finding the Hansard data that is most similar to the news data (Axelrod et al., 2011; van der Wees et al., 2017). While this may be worth exploring, we felt the extreme discrepencies between news and Hansard merited examination with gold in-domain data.

[15] We also experimented with 11 and 21 duplicates of the training data, and dropout rates of 0.2; we did not observe major differences between the settings.

news data in training), we found that BPE-dropout was more helpful in the IU-EN direction (+∼0.4 BLEU) than in the reverse (+∼0.2 BLEU). After incorporating a third of the news data in training, we found the reverse: a small increase for IU-EN (+∼0.1) and a slightly larger increase for EN-IU (+∼0.3).

5.4 Tagging and Backtranslation

By incorporating news data into our training set (Section 5.1), we improve performance on news data. However, the approach is sensitive to the number of copies of news data added, which can decrease performance on both Hansard and news data if not carefully managed. Both English news data and monolingual English parliamentary data (from Europarl) are plentiful in WMT datasets, so we incorporated them into our models via backtranslation (Sennrich et al., 2016b).

We apply approaches from Kobus et al. (2017) and Caswell et al. (2019): tagging source data domain (<NH> or <NEWS>) and (for IU-EN) tagging backtranslated source data (<BT>). Tagging domain appears to be particularly important for translating into Inuktitut, with between 1.4 and 2.4 BLEU points improvement on a subset of the news development test and minimal effect on the Hansard development data scores.

For backtranslation, we chose random samples of Europarl and WMT 2019 news data, experimenting with 325k, 650k, and 1.3M lines each, with 1.3 million performing best.[16] Ablation experiments with just news or just Europarl data showed less promise than the two combined. We did not perform backtranslation of Inuktitut (see Appendix E).

We performed two rounds of backtranslating the randomly sampled 1.3M lines each of Europarl and WMT 2019 news data. The first round (BT1) used our strongest single 5k joint BPE (with dropout) EN-IU system at the time. The second round (BT2) used a stronger three-way ensemble, with improved performance on both Hansard and news.

We experimented with combinations of tags for the backtranslated data (other parallel corpora have source domain tags unless otherwise stated):

- tagging all backtranslated data with <BT>;
- tagging backtranslated data with both <BT> and domain tags, where the domain tag matches the closest parallel corpus domain, i.e., <BT> <NH> or <BT> <NEWS>.[17]
- tagging backtranslated data with just a domain tag matching the closest parallel corpus domain, i.e. <NH> or <NEWS>.
- tagging all backtranslated data with <BT>, but not domain tagging the parallel corpora.
- tagging nothing.

As Table 1 shows for IU-EN translation, using backtranslated Europarl and news text clearly helped translating news text (as much as 8.0 BLEU) while only slightly impacting the translation of Hansard text. Without any backtranslated text, using domain tags (<NH> for Hansard and <NEWS>) appears to have a small positive effect on Hansard translation, and none on news (contrary to what we observed in the EN-IU direction).

The main observation from these experiments was that it was most important to distinguish backtranslation from true bitext (an observation similar to those noted in Marie et al. (2020)). Our best results were observed with no tags for the bitext and the <BT> tag for the backtranslated data. These experiments finished after the deadline, so our final submission uses the the next best combination: domain tags for bitext and <BT> tags for backtranslation.[18]

5.5 Finetuning

After building models with domain tags and backtranslation (in the case of IU-EN), we turned to finetuning to see if there was room to improve.

For systems that had been trained on Hansard data concatenated with the first third of the news development data, we experimented with finetuning on just that same first third of news data (using the second third for early stopping and the final third for evaluation), as well as both the first and second thirds (using the final third for both early stopping and evaluation). These approaches improved translation performance in terms of BLEU on the remaining third, with the use of more news data being more effective.[19]

[16] This was the largest size tested; it remains possible that increasing it even more could lead to even better performance.

[17] We also experimented with using novel tags for the domains of the backtranslated data (<PARL> and <EN-NEWS>) with and without additional <BT> tags, but found this had approximately the same effect as combining the backtranslation and domain tags, so we omit it from Table 1.

[18] Additional details of the backtranslation systems and these experiments are in Appendix D.

[19] We expect that training on more of the news data from the start (i.e., two thirds) might improve performance even more, but for our initial experiments we chose to use one third in

Backtranslation Data	Bitext Source Tag	Backtranslation Tag	NH	News.03	Avg
—	—	—	41.0	18.0	29.5
—	<NH|NEWS>	—	**41.3**	18.0	29.7
BT1	—	—	41.0	22.9	32.0
BT1	<NH|NEWS>	<NH|NEWS>	41.0	21.6	31.3
BT1	<NH|NEWS>	<BT> <NH|NEWS>	40.9	23.5	32.2
BT1	<NH|NEWS>	<BT>	40.7	23.8	*32.3*
BT2	—	—	40.8	23.6	32.2
[FINAL] BT2	<NH|NEWS>	<BT>	41.0	25.1	33.1
BT2	—	<BT>	40.9	**26.3**	**33.6**

Table 1: Backtranslation tag experiments on: IU-EN 15k Joint BPE, NH + News.01 (duplicated 15 times), 1.3M EuroParl, 1.3M News. Cased word BLEU scores measured on Hansard (NH) and last third of news (News.03; final 718 lines) portions of newsdev2020-iuen.

We also found that we were able to improve translation of Hansard data by finetuning on recent data. Joanis et al. (2020) observed recency effects when building models with subsets of the data. Here we take that observation a step further and find that finetuning with recent data (Hansard training data from 2017, which was already observed in training) produces BLEU score improvements on Hansard development data on the order of 0.5 BLEU into English, and on the order of 0.7 BLEU into Inuktitut (Tables 2 and 3).[20]

Despite the use of domain tags, finetuning on one domain has negative results for the other (see Tables 2 and 3).

System	NH Dev.	ND 3	NH Test	News Test	Full Test
Base: NH+ND.1	24.7	11.7	16.7	11.6	14.1
Base: NH+ND.2	24.7	11.3	16.7	11.2	13.9
Base: NH+ND.3	24.7	–	16.9	12.2	14.5
Ensemble	25.0	–	17.1	13.3	15.1
F.t. ND.{1,2}	21.5	**13.5**	15.0	12.2	13.8
F.t. ND.{2,3}	21.9	–	15.0	13.2	14.4
F.t. ND.{3,1}	20.9	–	13.8	13.1	13.7
Ens.: F.t. ND	21.7	–	14.9	**14.1**	14.8
F.t. NH (from 1)	25.4	11.9	16.9	11.3	14.0
F.t. NH (from 2)	25.4	11.0	16.8	11.0	13.9
F.t. NH (from 3)	25.3	–	16.8	11.3	14.0
Ens.: F.t. NH	**25.7**	–	**17.5**	12.9	15.1
Final hybrid	**25.7**	–	**17.5**	**14.1**	**15.8**

Table 2: BLEU scores of 10k joint BPE EN-IU systems. The best performer is in **bold**. ND=News dev., indexed by thirds. F.t.=Finetuning. Dashes mean a score should not be computed due to test/training data overlap.

5.6 Ensembling and Hybrid Model

Our hope was to build a single system to translate both news and Hansard but, in the end, we found that our attempts at finetuning for the combination of news and Hansard were outperformed by systems finetuned to one specific domain. Maintaining a held-out third of news data allowed us to measure performance of ensembled models on news data, so long as we only ensembled systems that had not trained on that held-out data. In order to create our final submissions, we chose finetuned systems based on the held-out third, and then ensembled them with the assumption that the strong ensemble with access to the full news development data would outperform the individual systems or pairs of systems trained on subsets. In general, we found that ensembling several systems (using Sockeye's built-in ensembling settings) improved performance. However, this had some limits: for EN-IU if we combined a strong news system whose performance on Hansard had degraded too much with a strong Hansard system whose performance on news had degraded, the final result would be poor performance on both domains.

Our solution to this was simple: decode news data with an ensemble of models finetuned on news, and decode Hansard data with an ensemble of models finetuned on Hansard. Our final submissions are hybrids of domain-specific systems.[21]

6 Submitted Systems

6.1 English–Inuktitut

Our primary submission for EN-IU is a hybrid of two joint BPE 10k ensembled systems with

order to enable us to measure improvements on a held-out set; see Section 5.6 for our efforts to use ensembling to balance the usefulness of training on more data with the ability to measure progress during preliminary experiments.

[20]Note that there is a fine distinction between the two settings here: when finetuning on recent Hansard data, the system is training on data it has already seen. When finetuning on news data, we expose the system to some data it has already seen (one third of the news data) and some data that it has *not* trained on (another third of the news data).

[21]This leaves questions open, e.g., if a Hansard system trained without any news data would perform as well or better on Hansard test data than one trained with news data.

System	NH Dev.	ND 3	NH Test	News Test	Full Test
BT1:NH+ND.1	40.7	23.8	29.0	21.6	25.6
BT2:NH+ND.1	41.0	25.1	29.3	22.1	25.9
BT2:NH+ND.2	41.1	25.1	28.9	22.9	26.1
BT2:NH+ND.3	41.1	–	28.7	22.6	25.9
Ensemble	41.7	–	29.6	24.8	27.4
F.t. ND.{1,2}	39.9	**26.7**	28.5	23.9	26.4
F.t. ND.{2,3}	39.6	–	28.2	23.8	26.1
F.t. ND.{3,1}	40.1	–	28.4	23.7	26.2
Ens.: F.t. ND	40.9	–	29.1	**25.8**	27.6
F.t. NH (from 1)	41.6	23.6	29.0	21.0	25.3
F.t. NH (from 2)	41.5	24.6	28.9	22.8	26.1
F.t. NH (from 3)	41.5	–	28.8	21.6	25.5
Ens.: F.t. NH	**42.4**	–	**29.9**	24.3	27.3
Final hybrid	**42.4**	–	**29.9**	**25.8**	**28.0**

Table 3: BLEU scores of IU-EN systems. The best performer is in **bold** font. ND=News dev., indexed by thirds. F.t.=Finetuning. Dashes mean a score should not be computed due to test/training data overlap.

domain tags. To translate the Nunavut Hansard data, we used an ensemble of three systems, all finetuned on 2017 Hansard data using only the Hansard development data for validation during finetuning. The three base systems used for finetuning were trained on the full Hansard along with the first, second, or third news third (duplicated 15 times), respectively, with BPE-dropout on both the source and target sides.

To translate the news data, we again used an ensemble of three base systems trained with BPE-dropout on both the source and target sides: a base system trained on all Hansard data with the first third of news data (duplicated 15 times) finetuned on the first and second third of news data, another such base system trained instead with the second third of news data (duplicated 15 times) and finetuned on the second and third third of news data, and a final base system trained with the third third of news data (duplicated 15 times) and finetuned on the first and third thirds. The hybrid system had a BLEU score of 15.8 on the test data (Table 2).

6.2 Inuktitut–English

Our primary submission for IU-EN is a hybrid of two joint BPE 15k ensembled systems with domain tags (for news and Hansard bitext) and backtranslation tags (for the backtranslated data). Due to time constraints, we did not run BPE-dropout. Like the EN-IU direction, we built three baseline systems. All baseline systems were trained on the full Hansard training data, along with 1.3 million lines of backtranslated Europarl data and 1.3 million lines of backtranslated news 2019 data. The three baseline systems differed in which third of news was used for training, as described for EN-IU. Backtranslation was performed using an ensemble of the three baseline systems used for the EN-IU task (joint BPE 10k, BPE-dropout).

We performed finetuning on news and recent Hansard in the same manner as for EN-IU. The news test data was translated with the ensemble of news-finetuned systems, while the Hansard test data was translated with the ensemble of the Hansard-finetuned systems. The final system had a BLEU score of 28.0 on the test data (Table 3).

7 Conclusions and Future Work

We have presented the results of our IU-EN and EN-IU systems, showing that a combination of BPE-dropout (for EN-IU), backtranslation (for IU-EN), domain-specific finetuning, ensembling, and hybrid systems produced competitive results. We performed automatic evaluations of the submitted systems in terms of BLEU, chrF (Popović, 2015), and YiSi-1 (Lo, 2020). Our EN-IU system performed best out of the constrained systems in terms of BLEU (15.8, +2.5 above the next-best system), chrF (37.9, +1.7 above the next-best), and YiSi (82.4, +0.5 above the next-best). Our IU-EN system performed third-best out of all systems in terms of BLEU (28.0, -1.9 below the best system), third-best in terms of chrF (48.9, -2.0 below the best system), and third-best in terms of YiSi-1 (92.3, -0.6 behind the best system).[22]

There remains a wide range of future work to be done to improve translation for this language pair. There is still space to improve Roman text output in Inuktitut, perhaps even as simply as an automatic postediting approach. Different subword segmentations (or ones complementary to BPE-dropout like He et al. (2020)), particularly ones that capture morphological and phonological aspects of Inuktitut may also be promising.

In terms of adding monolingual data, we expect that improved data selection for backtranslated data (i.e., to increase topic relevance) may be useful, as would additional Inuktitut monolingual data. Due to time constraints, we were unable to complete BPE-dropout for IU-EN systems; we expect this would have resulted in improved performance.

[22] We do not have information about whether any of these systems were unconstrained. It is also worth noting that the highest-ranked systems differed depending on the metric used, so we await human evaluation.

Domain finetuning remains a challenge given the very small amount of parallel news data available. We did experiment with mixing Hansard and news data for finetuning, but, contrary to Chu et al. (2017), were unable to outperform news-only systems on news. It may be worth trying approaches designed to prevent catastrophic forgetting in domain adaptation (Thompson et al., 2019).

The real test, of course, will be human evaluation; are the systems producing output that might be usable, whether for computer aided translation (via postediting or interactive translation prediction) or for use in other downstream applications?

Acknowledgments

We thank the reviewers for their comments and suggestions. We thank Eddie Santos, Gabriel Bernier-Colborne, Eric Joanis, Delaney Lothian, and Caroline Running Wolf for their comments and feedback on the paper. We thank Chi-kiu Lo for providing automatic evaluation scores of submitted systems. We thank the language experts at Pirurvik Centre for their work on the forthcoming human annotations, and the Government of Nunavut and Nunatsiaq News for providing and allowing the use and processing of their data in this shared task.

References

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

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

Benoît Farley. 2009. Uqailaut. `www.inuktitutcomputing.ca/Uqailaut/info.php`.

Government of Nunavut. 2020. We speak Inuktut. `https://www.gov.nu.ca/culture-and-heritage/information/we-speak-inuktut`. Accessed August 11, 2020.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2020. Dynamic programming encoding for subword segmentation in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3042–3051, Online. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

Inuit Tapiriit Kanatami. 2018. National Inuit Strategy on Research.

Inuit Tapiriit Kanatami. 2020. Unification of the Inuit language writing system.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Rebecca Knowles and Philipp Koehn. 2018. Context and copying in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3034–3041, Brussels, Belgium. Association for Computational Linguistics.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of the 10th Machine Translation Summit (MT Summit)*, pages 79–86.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Philippe Langlais, Fabrizio Gotti, and Guihong Cao. 2005. NUKTI: English-Inuktitut word alignment system description. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 75–78, Ann Arbor, Michigan. Association for Computational Linguistics.

Samuel Larkin, Boxing Chen, George Foster, Ulrich Germann, Eric Joanis, Howard Johnson, and Roland Kuhn. 2010. Lessons from NRC's Portage system at WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics-MATR*, WMT '10, pages 127–132, Stroudsburg, PA, USA. Association for Computational Linguistics.

Legislative Assembly of Nunavut. 2020. FAQS: What is Hansard? `https://assembly.nu.ca/faq#n125`. Accessed August 11, 2020.

Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kūpono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaoleohaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Maroussia Levesque, Keoni Mahelona, Caleb Moses, Isaac ('Ika'aka) Nahuewai, Kari Noe, Danielle Olson, 'Ōiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hēmi Whaanga. 2020. Indigenous protocol and artificial intelligence position paper. Project Report 10.11573/spectrum.library.concordia.ca.00986506, Aboriginal Territories in Cyberspace, Honolulu, HI. Edited by Jason Edward Lewis.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chi-kiu Lo. 2020. Extended study of using pretrained language models and YiSi-1 on machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation: Shared Task Papers*.

Mick Mallon. 2000. Inuktitut linguistics for technocrats. Ittukuluuk Language Programs. `https://www.inuktitutcomputing.ca/Technocrats/ILFT.php`.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Joel Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and using an English-Inuktitut parallel corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 115–118.

Joel Martin, Rada Mihalcea, and Ted Pedersen. 2005. Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 65–74, Ann Arbor, Michigan. Association for Computational Linguistics.

Jeffrey Micher. 2017. Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106, Honolulu. Association for Computational Linguistics.

Jeffrey Micher. 2018. Using the Nunavut hansard data for experiments in morphological analysis and machine translation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 65–72, Santa Fe, New Mexico, USA. 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 the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Carla Parra Escartín, Wessel Reijers, Teresa Lynn, Joss Moorkens, Andy Way, and Chao-Hong Liu. 2017. Ethical considerations in NLP shared tasks. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 66–73, Valencia, Spain. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Lane Schwartz, Francis Tyers, Lori Levin, Christo Kirov, Patrick Littell, Chi kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, Jeffrey Micher, Lonny Strunk, Han Liu, Coleman Haley, Katherine J. Zhang, Robbie Jimmerson, Vasilisa Andriyanets, Aldrian Obaja Muis, Naoki Otani, Jong Hyuk Park, and Zhisong Zhang. 2020. Neural polysynthetic language modelling.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. 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.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. 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 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul, Turkey. European Languages Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2955–2960, Brussels, Belgium. Association for Computational Linguistics.

Marlies van der Wees, Arianna Bisazza, and Christof Monz. 2017. Dynamic data selection for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1410, Copenhagen, Denmark. Association for Computational Linguistics.

Hayahide Yamagishi, Shin Kanouchi, Takayuki Sato, and Mamoru Komachi. 2016. Controlling the voice of a sentence in Japanese-to-English neural machine translation. In *Proceedings of the 3rd Workshop on Asian Translation (WAT2016)*, pages 203–210, Osaka, Japan. The COLING 2016 Organizing Committee.

A Data and BPE Sizes

For reproducibility, we show the data sizes after cleaning in Table 4. Exact sizes for BPE vocabularies (including codes, and extracted vocabulary sizes) are shown in Table 5.

For IU-EN, we tried disjoint (Inuktitut, English) models with BPE size pairs: (500,500), **(1k,1k)**, (5k,5k), (10k,10k), (500,5k), (1k,5k), and (5k,10k). For EN-IU, we tried disjoint (English, Inuktitut) models with BPE size pairs: (500,500), (1k,1k), (5k,5k), (10k,10k), **(5k,1k)**, (10k,1k), (10k,5k). The typological differences between English and Inuktitut motivated these experiments on vocabulary sizes, testing whether unbalanced sizes may be better than balanced sizes when using disjoint vocabularies. Our intuition was that smaller vocabulary sizes for Inuktitut and larger vocabulary sizes for English might lead to mappings between vocabulary items that were closer to one-to-one.

We also tried using SentencePiece Unigram models (Kudo, 2018) trained for the same size pairs, but they did not yield improvements in BLEU score, so we used `subword-nmt` for the remainder of our experiments.

B Roman Text in Inuktitut Output

When Roman text appears in the Inuktitut side of the Hansard, that same text almost always appears in the English side of the data. In the test data, this is the case for all but 9 of the 151 Roman tokens on the Inuktitut side of the data (over 94%).[23] Thus we expect a good NMT system should learn to produce Roman output that matches some portion of the Roman text in the English input (or at least, that should be true for *most* Roman text it produces).

When the BPE vocabulary model is learned jointly, this task has the potential to be trivial: the system simply needs to learn which source vocabulary tokens are the same as which target vocabulary tokens and then copy them, and RNN-based NMT systems are known to be able to use context in order to do this (Knowles and Koehn, 2018). When the BPE vocabularies are learned independently, however, such words may be broken into different sequences of subwords on the source and target sides, a more challenging task for the system to handle. Intuitively, the system must learn to spell to do this successfully. This led us to experiment with joint vocabulary models. We measured precision, recall, and F1 averaged across Hansard and news development data for Roman words (whole words, after de-BPE, but prior to detokenization) for systems trained using disjoint vocabulary models and those with vocabularies trained jointly (but extracted separately). We found comparable BLEU scores between the two settings, but found that average F1 was higher in the joint setting.

In Table 6, we see three systems from an early set of experiments with identical BLEU scores (the best out of their respective vocabulary size and data balancing sweeps; the disjoint system had the news data repeated 30 times, while the 2k joint system had it repeated 10 times and the 10k joint system had it repeated 30 times). The joint systems had higher F1 scores, particularly driven by improvements in precision (on Hansard, an increase from 30.9% precision to 38.7% and 35.7% and on news an increase from 28.1% to 49.2% and 43.2%).

Nevertheless, as evidenced by the relatively low F1 scores, there best systems still make some errors, as shown in the example below:

Src: Mr. Speaker, the Indspire Awards represent the highest honour the indigenous community presents to its people.

Ref: ᐅᖃᖅᑏ, ᑖᒃᑯᐊ ᐱᑖᖑᖃᑦᑕᖅᑐᑦ ᑭᒡᒐᖅᑐᐃᓂᖅ ᐳᖅᑐᓂᖅᐸᒥᒃ, ᐅᐱᓐᓇᖅᑐᒥᐅᑕᑯᓗᐃᑦ ᓄᓇᖃᖅᑳᖅᑐᕕᓂᕐᓄᑦ ᑕᕝᕗᖓ ᐃᓄᕐᑯᑎᖕᒋᓐᓄᑦ.

MT: ᐅᖃᖅᑏ, ᑖᒃᑯᐊ Inspiration Award-ᖓᓂᕋᖅᑕᐅᔪᑦ ᕐᑯᑦᑎᓂᖅᐸᒥᒃ ᓄᓇᖃᖅᑳᖅᑐᒥᓂᕐᓄᑦ ᑐᓂᓯᖃᑦᑕᕐᒪᑕ ᐃᓄᕐᑯᑎᒥᓄᑦ.

There are in fact two errors in this MT output: first, the system generates Roman text that it perhaps ought not to have generated, and second it does not successfully copy *Indspire*, instead producing *Inspiration*. This suggests that, although using joint BPE has improved Roman text output in Inuktitut, there is still room for additional improvement. Our final submission had an F1 score of 27.4 (41.3% precision and 20.5% recall)

[23] Of the 9 exceptions, 6 were cases where one side used an abbreviation and the other expanded it, 1 was a plural/singular distinction, 1 was a capitalization difference, and 1 was a spelling difference.

Data set	Sentences	IU words	EN words
Nunavut Hansard 3.0	1299349	7992376	17164079
Nunavut Hansard 3.0 (2017 only)	40951	275248	582480
News (from newsdev2019-eniu)	2156	24980	44507
EN Europarl v10 (full)	2295044		56029587
EN Europarl v10 (subselect)	1300000		31750842
EN News.2019 (full)	33600797		836569124
EN News.2019 (subselect)	1300000		32380145

Table 4: Dataset sizes (post cleaning) of data used in our experiments. Of the monolingual data, only the subselection was used, not the full dataset.

BPE model	Codes	IU Vocab	EN Vocab
IU 1k	573	1006	
EN 1k	794		995
IU 2k	1573	2000	
EN 2k	1794		1978
IU 5k	4573	4991	
EN 5k	4794		4895
IU 10k	9573	9989	
EN 10k	9794		9749
JNT 1k	557	977	519
JNT 2k	1557	1919	1086
JNT 5k	4557	4480	2754
JNT 10k	9557	8597	5071
JNT 15k	12590	12590	7038

Table 5: BPE codes and extracted vocabulary sizes using `subword-nmt` with the `--total-symbols` flag. Single language BPE models are indicated by ISO code and joint models by *JNT*.

System	BLEU	Ave. F1
Disjoint BPE: IU 1k, EN 5k	24.7	24.2
Joint BPE 2k	24.7	25.9
Joint BPE 10k	24.7	27.6

Table 6: Comparison of best disjoint and joint BPE systems trained using Nunavut Hansard and half of the news data as training, scored with BLEU and with Roman text F1 averaged over the Hansard development data and the other half of the news development data. These were early systems trained without tags or back-translation.

C Preprocessing and Postprocessing

In this appendix, we provide detail about our additional language-specific preprocessing and postprocessing.

C.1 Preprocessing

Our additional preprocessing focuses on quotation marks, apostrophes, and some other punctuation. We first describe English-specific preprocessing.

We normalize double quotation marks to three distinct special tokens, `-LDQ-`, `-RDQ-`, and `-UDQ-` (left, right, and unknown double quote, respectively), separated from any surrounding characters by a space. For directional quotation marks ('LEFT DOUBLE QUOTATION MARK' (U+201C) and 'RIGHT DOUBLE QUOTATION MARK' (U+201D)), this is a simple substitution. For straight quotations ('QUOTATION MARK' (U+0022)), we apply the following heuristics: those followed by a space are right, those preceded by a space are left, those followed by punctuation (period, comma, question mark, semicolon) are right, those at the beginning of a line are left, those at the end of a line are right. All that remain are considered unknown.

For single quotes or apostrophes ('LEFT SINGLE QUOTATION MARK' (U+2018) and 'RIGHT SINGLE QUOTATION MARK' (U+2019)), we do as follows. We first convert any instances of 'GRAVE ACCENT' (U+0060) to the right single quote (this is rare but manual examination of the training data suggests that they are used as apostrophes). We then convert any instances of left and right single quotation marks to special (space-separated) tokens `-LSA-` and `-RSA-`, respectively. We next consider 'APOSTROPHE' (U+0027). That token followed by a space is mapped to `-RSA-`, while any instances preceded by a space are mapped to `-LSA-`. Any that are sandwiched between alphanumeric characters (a-z, A-Z, 0-9) are treated as a word internal apostrophe, `-RSI-`. Remaining ones preceded by alphanumeric characters are mapped to `-RSA-`, while those followed by alphanumeric characters are mapped to `-LSA-`. Any remaining at this point are mapped to `-ASO-` (other).

We also map 'EN DASH' (U+2013) to -NDA- and 'EM DASH' (U+2014) to -MDA- (as ever, keeping these special tokens space-separated from remaining text).

For Inuktitut, we use similar substitutions, noting the differences below. This is run after the spelling normalization script provided. For quotation marks, any instances of 'LEFT SINGLE QUOTATION MARK' (U+2018) followed immediately by 'RIGHT SINGLE QUOTATION MARK' (U+2019) are treated as -LDQ-, while any instances of two 'RIGHT SINGLE QUOTATION MARK' (U+2019) in a row are treated as -RDQ-. Double apostrophe is first mapped to 'QUOTATION MARK' (U+0022). Those straight double quotes preceded *or* followed by punctuation (period, comma, question mark, semicolon) are treated as -RDQ-. We expand the earlier alphanumeric matching to include the unicode character range 1400-167F, which contains all syllabics present in the data.

There are five observed types of single quotes or apostrophes in the data. The most common is 'RIGHT SINGLE QUOTATION MARK' (U+2019), appearing more than 9000 times, followed by 'APOSTROPHE' (U+0027), appearing more than 1300 times, followed by 'GRAVE ACCENT' (U+0060), over 600 times, 'LEFT SINGLE QUOTATION MARK' (U+2018), which appears fewer than 200 times, and 'ACUTE ACCENT' (U+00B4), which appears very rarely. We first map the grave accent to 'RIGHT SINGLE QUOTATION MARK' (U+2019). Then, for the remaining four types, if they appear within syllabics (range U+1400 to U+167F), we map them to 'MODIFIER LETTER APOSTROPHE' (U+02BC). This is important because this is then treated as a *non-breaking* character for the purposes of Moses tokenization. It often represents a glottal stop, which *should* be treated as part of the word, not necessarily as something to split on. When one of the four types appears at the end of a word, it is treated as a -RSA- if a left single apostrophe was observed before it in the sentence. Any remaining at the ends of syllabic words are treated as modifier letter apostrophe. Any of the four that appear between non-syllabic alphanumeric characters are mapped to -RSI-. Remaining left single quotation marks are mapped to -LSA-, while remaining right single quotations and acute accents are mapped to -RSA-. Apostrophes are then mapped in the same manner as English, with the addition of the syllabic range to the alphanumeric range.

C.2 Postprocessing

The postprocessing is done to revert the placeholder tokens to appropriate characters and is done after de-BPE-ing and Moses detokenization.

For English, we do as follows. The placeholder -LDQ- and any spaces to the right of it are replaced with 'LEFT DOUBLE QUOTATION MARK' (U+201C), while -RDQ- and any spaces to the left of it are replaced with 'RIGHT DOUBLE QUOTATION MARK' (U+201D), and -UDQ- is replaced with 'QUOTATION MARK' (U+0022) with no modification to spaces.

The -RSI- token and any surrounding spaces are replaced with 'RIGHT SINGLE QUOTATION MARK' (U+2019), -RSA- and any spaces preceding it are replaced with 'RIGHT SINGLE QUOTATION MARK' (U+2019), -LSA- and any spaces following it are replaced with 'LEFT SINGLE QUOTATION MARK' (U+2018), and -ASO- is replaced with 'APOSTROPHE' (U+0027).

The em-dash placeholder is replaced with an em-dash without surrounding spaces, while the en-dash placeholder is replaced with an en-dash *with* surrounding spaces. We also perform some other small modifications to match the most common forms in the original text: spaces around dashes and forward slashes are removed, times are reformatted (spaces removed between numbers with colons and other numbers), space between greater than signs is removed, space is removed before asterisks, spaces are removed following a closing parenthesis that follows a number, three periods in a row are replaced with 'HORIZONTAL ELLIPSIS' (U+2026), space is removed after asterisk that begins a line, space is removed after the pound sign, and space is removed between a right apostrophe and a lowercase s.

For Inuktitut, the postprocessing is similar, with the following changes/additions: the modifier letter apostrophe is replaced with the 'RIGHT SINGLE QUOTATION MARK' (U+2019), no spaces are placed around the en-dash, and spaces are removed between a close parenthesis followed by an open parenthesis.

D Backtranslation Details

Here we describe details of our backtranslation experiments. The first pass (BT1) employed our strongest English–Inuktitut system at the time, trained on the Nunavut Hansard bitext plus the first third of the news bitext (from newsdev2020-eniu) using 5k joint BPE with BPE-dropout on both source and target. Later, we backtranslated the data a second time (BT2) using a stronger three-way ensemble of systems, each of which was trained on the NH corpus and a different third of the news bitext from newsdev2020-eniu using 10k joint BPE with BPE-dropout on both source and target. This ensemble improved the BLEU score on the NH portion of newsdev2020-eniu by 0.5 BLEU (from 24.5 to 25.0); while we could not measure the improvement of the 3-way ensemble on news data, an ensemble of two of these systems (trained using one of the first two thirds of news) yielded a 1.5 boost in BLEU measured on the final news third (from 12.1 to 13.6) over the system used for the first round. Thus the ensembled system used for this second round of backtranslation was stronger at translating both parliamentary and news data.

With BT1 backtranslated data, positive effects came from ensuring that backtranslated data and true bitext are tagged differently. Tagging the backtranslated source with the exact same domain tags as the parallel data leads to a decrease in performance of 1.7–2.2 BLEU for translating news; it is even worse (by 1.3 BLEU on news) than using no tags at all.

While most round one (BT1) backtranslation tagging methods yielded news data BLEU increases between 0.4 and 0.8 (over not tagging), a larger improvement of $\geq$1.5 BLEU occurred when using our second round of backtranslated data (BT2); notably, the worst system trained using BT2 scores only 0.1 BLEU (average) below the best BT1 system. Our best performance was achieved using BT2 backtranslations with <BT> tags but no domain tagging (for either the parallel or backtranslated source). It outperformed the next best system by 1.2 BLEU on news; unfortunately those experiments did not complete before the deadline. Thus our submitted system used the best available systems at the time for additional finetuning: domain tags on parallel data and <BT> tags on the backtranslated data.

Each of the individual systems that contributed to our final Inuktitut–English system combination used 1.3 million lines of Europarl (tagged as <BT>), 1.3 million lines of news (tagged as <BT>), approximately 1.3 million lines of Nunavut Hansard (tagged as <NH>), and 719 or 718 lines of news (tagged as <NEWS> and duplicated 15 times).

E Inuktitut Common Crawl and Additional Data

We did not use any data from Inuktitut Common Crawl in our submissions. In our initial experiments, we found it generally harmed translation quality. Nevertheless, we provide here some observations from our analysis, in the hopes that they are useful to other researchers. First, the Common Crawl data provided contains fairly large amounts of non-Inuktitut data. This includes noise, such as long sequences of characters (like lists of characters) as well as text art (such as English words spelled using visually similar syllabics and other characters, e.g., ᕼᕮᒪᒪᑐ). There is also text in several other languages and dialects, including, but likely not limited to: ᓇᓴᒃᐱ (Naskapi),[24] ᖃᐧᐃᔭᐁᐧᐃᐧᐣ (Plains Cree), and ᓄᓇᕗᒻᒥᐅᑎᑐᑦ (Nunavimmiutitut, an Inuit language spoken in Nunavik).[25] Of particular note is the latter, which – while it is the only one of the three within the Inuit dialect continuum that includes Inuktitut – is an Inuit language (sometimes called the Nunavik dialect of Inuktitut) that makes use of one additional column in the syllabary (∇, ∨, ∪, ..., or *ai*, *pai*, *tai*, ...). Those characters do not appear in the Hansard, thus rendering it impossible for our systems to translate them exactly without some form of modification, even if they might otherwise share similarities with words that appear in the Hansard. Removing characters that were not observed in the Hansard data (which helps filter out some non-Inuktitut language data) and filtering out potential text art results in a much smaller Common Crawl data set, less than half the size of the original.

While additional monolingual or bilingual data would likely benefit English to Inuktitut translation, we encourage non-Inuit researchers who plan to perform data collection to do so in collaboration with Inuit communities and language speakers.

[24] Text appears to be scraped from the Naskapi Community Web Site, http://www.naskapi.ca/.

[25] https://www.kativik.qc.ca/our-schools/resources/

The efforts of Inuit language experts at Pirurvik Centre were vital to the analysis of the data used for this task (Joanis et al., 2020), collected through communications with Nunatsiaq News and the Government of Nunavut with the goal of selecting data usable for this translation task, both in terms of public availability and language. Aside from the machine learning related risks of accidentally collecting data from other languages and labeling it as Inuktitut (as we observed in the Common Crawl data), there are also ethical concerns. While it does not focus specifically on language data, the National Inuit Strategy on Research (NISR, Inuit Tapiriit Kanatami, 2018) highlights as a priority "Ensuring Inuit access, ownership and control over data and information" and focuses on partnership with Inuit organizations, transparency, and data sharing to end exploitative research practices and build research relationships that respect Inuit self-determination.[26] The NISR contains a discussion of potential harms of research done without relationships to the communities impacted by it, with both Inuit-specific concerns and concerns from a broader history of colonialism. Lewis et al. (2020) provide a discussion of guidelines for Indigenous-centred AI from a variety of Indigenous perspectives (though not specifically from Inuit perspectives), including topics of ethics, data sovereignty, and responsibility and relationships in AI. Building and maintaining community relationships and collaborations can help ensure that data is handled and shared in ways that respect cultural values and Indigenous intellectual property,[27] which outsiders may not be familiar with. A full discussion of these topics is beyond the scope of this paper, but we raise the discussion here as part of the process of working towards best practices in building respectful research relationships that centre community goals at all steps of the research process.

F Statement on Avoiding Conflicts of Interest

In their work on ethical considerations in shared tasks, Parra Escartín et al. (2017) raise the issue of actual or perceived conflicts of interest between task organizers and participants. We provide the following information in the interest of transparency.

The data for the shared task on Inuktitut-English was collected by researchers at the National Research Council of Canada (NRC) in collaboration with the Pirurvik Centre.[28] The team of researchers at NRC was divided into two groups: those working on task organization and those participating in the shared task (the latter group are the authors of this paper). In order to prevent unfair advantages to the task participants, the organizers did not discuss the web source or details of the evaluation set with the participants at any time before the submission of the systems.

We did communicate with the organizers to receive clarification regarding the spurious quotes in the test data; the response to this was distributed to the full WMT mailing list.

[26] https://www.itk.ca/wp-content/uploads/2018/04/ITK_NISR-Report_English_low_res.pdf

[27] United Nations Declaration on the Rights of Indigenous Peoples, Article 31.

[28] https://www.pirurvik.ca/