Machine Translation for English–Inuktitut with Segmentation, Data Acquisition and Pre-Training

Christian Roest† Lukas Edman‡ Gosse Minnema‡
Kevin Kelly† Jennifer Spenader† Antonio Toral‡
†Institute for Artificial Intelligence ‡Center for Language and Cognition,
University of Groningen
The Netherlands

c.roest@student.rug.nl, j.l.edman@rug.nl, g.f.minnema@rug.nl
kevin.kelly@live.se, j.spenader@ai.rug.nl, a.toral.ruiz@rug.nl

Abstract

Translating to and from low-resource polysynthetic languages present numerous challenges for NMT. We present the results of our systems for the English–Inuktitut language pair for the WMT 2020 translation tasks. We investigated the importance of correct morphological segmentation, whether or not adding data from a related language (Greenlandic) helps, and whether using contextual word embeddings improves translation. While each method showed some promise, the results are mixed.

1 Introduction

This paper presents the neural machine translation (NMT) systems submitted by the University of Groningen to the WMT 2020 translation task[1] between Inuktitut and English in both directions (EN↔IU), describing both constrained and unconstrained systems where we investigated the following research questions:

- RQ1. Does morphological segmentation benefit translation with polysynthetic languages? Existing NMT research showed that morphological segmentation outperforms byte-pair encoding (BPE) (Sennrich et al., 2016) for some agglutinative languages. For example, rule-based morphological segmentation improved English-to-Finnish translation (Sánchez-Cartagena and Toral, 2016). and unsupervised morphological segmentation improved Turkish-to-English translation (Ataman et al., 2017). We investigate if morphological segmentation also improves translation performance for polysynthetic languages, and if effects differ depending on translation direction.
- RQ2. Does the use of additional data from a related language, Greenlandic (KL), improve the outcome? Due to the scarcity of EN–IU parallel data, we investigate if adding Greenlandic data to the Inuktitut data to train a multilingual NMT system (Johnson et al., 2017), improves the performance of the NMT systems on the unconstrained task (Zoph et al., 2016).
- RQ3. Does the translation benefit from using contextual word embeddings? The use of such embeddings has proven beneficial for many tasks in natural language processing (Devlin et al., 2019), including MT (Zhu et al., 2020), so we deem it sensible to test this for a polysynthetic language, which we will do by means of masked language modelling pre-training.

In section 2 we present the main data and evaluation measures used. In section 3 we present experiments with morphological segmentation methods. Section 4 presents the results of our translation systems, and in section 5 we present our conclusions.

2 Corpora and Evaluation

The preprocessing followed the procedure of Joanis et al. (2020), carrying out the following steps in order: spelling normalisation and romanisation (only for IU), punctuation normalisation, tokenisation, and truecasing (only for EN). Parallel data is additionally filtered (ratio 15, minimum and maximum length 1 and 200, respectively). As monolingual data we use the Common Crawl (CC) corpus for Inuktitut, and the 2019 version of Newscrawl for English. For CC we also filter out duplicate lines, lines of which more than 10% of the characters are neither alphanumerical nor standard punctuation, and lines that contain more than 200 words. These

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

Proceedings of the 5th Conference on Machine Translation (WMT), pages 274–281
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

steps reduce the amount of data considerably, from 164,766 to 28,391 lines. Line deduplication is also applied to Hansards.[2]

Since the parallel training data contains only Hansards, we used part of the news from the dev set as additional training data by splitting the news part of the dev set: the first 1859 lines are used for training and the last 567 for development. We refer to these subsets as `newsdevtrain` and `newsdevdev`, respectively.

Tables 1 and 2 show the parallel and monolingual datasets, respectively, used for training after preprocessing.

Corpus	Sentences	Words EN	Words IU
Hansards	769810	17303903	8236210
Newsdevtrain	1859	40154	24121

Table 1: Preprocessed EN–IU parallel training data.

Lang.	Corpus	Sentences	Words
IU	Common Crawl	28391	381805
EN	Newscrawl	5000000	143776337

Table 2: Preprocessed monolingual training data.

During development, we evaluated our systems on the news and Hansards portions of the development set, separately. We used two automatic evaluation metrics: BLEU (Papineni et al., 2002) and CHRF (Popović, 2015). CHRF is our primary evaluation metric for EN→IU, due to the fact that this metric has been shown to correlate better than BLEU with human evaluation when the target language is agglutinative (Bojar et al., 2016). BLEU is our primary evaluation metric for IU→EN systems, as the correlations with human evaluation of BLEU and CHRF are roughly on par for EN as the target language. Prior to evaluation the MT output is detruecased (only EN) and detokenized with Moses' scripts.

3 Segmentation with intrinsic evaluation

Like many polysynthetic languages, Inuktitut has a high degree of inflection and agglutination, leading to very long words with a very high morpheme-to-word ratio (Mager et al., 2018). By our estimation, Inuktitut has an average of around 4.39 morphemes per word.

This means on average there are more potential boundaries, as well as more actual segmentation boundaries to locate per word, making segmentation particularly challenging.

Inconsistent segmentation harms an NMT model's ability to extract knowledge, because it reduces the frequency and activation of all vocabulary items during training, such that for each individual element in the vocabulary is found in fewer contexts. At inference, inconsistent segmentation can result in morphs that are out-of-vocabulary, resulting in information loss.

We hypothesize that linguistically correct segmentation may be particularly beneficial for translation with polysynthetic languages because it could provide more consistent isolation of concepts into subwords.

We evaluated a broad pool of segmenters to determine how close various methods can achieve linguistically correct segmentation, comparing results to reference segmentations obtained from the Inuktitut Computing GitHub repository[3]. This repository contains 1096 Inuktitut words, manually segmented at the National Research Council of Canada (NRC).

Our experiments include: Rule-based with Uqailaut[4]; Morfessor Baseline (semi-supervised) (Creutz and Lagus, 2002); Morfessor FlatCat (semi-supervised) (Grönroos et al., 2014); LMVR (unsupervised) (Ataman et al., 2017); and Neural Transformer segmentation (supervised).

We used Uqailaut's rule-based segmenter to create additional annotated segmentations used to train the supervised and semi-supervised systems. In total 600,000 segmentations of unique words from the Hansard training dataset were created. All semi-supervised and unsupervised systems were trained with the Hansard training corpus. For training semi-supervised methods, we use 60,000 of the collected segmentations with Uqailaut as annotated training data, and another 3,000 as validation data. For LMVR we set the maximum lexicon size to 20,000.

Related to our work, a previous study (Kann et al., 2018) compared segmentation methods based on their ability to generate linguistically correct segmentations for several low-resource Mexican polysynthetic languages. Their proposed RNN-

[2] We used Hansards for training with and without deduplication and the former led to better results.

[3] https://github.com/LowResourceLanguages/InuktitutComputing

[4] http://www.inuktitutcomputing.ca/Uqailaut/info.php

based neural approach outperformed baselines of other common approaches, so we also tested a neural segmentation method, but instead of an RNN we use a Transformer architecture. We implement this neural segmenter using Marian[5]. On the source side, the unsegmented words are used as input data. The corresponding segmented words are used as target data. On the target side we denote the segmentation boundary by adding a boundary token (@), like in the following example:

Source: a k i r a q t u q t u t
Target: a k i r a q @ t u q @ t u t

We trained three neural segmentation models: one on all 600,000 annotated segmentations, plus two with 45,000 annotated segmentations, one with only unambiguous annotations[6] and one with a random selection from the pool of 600,000.

Table 3 shows the intrinsic evaluation results. Similar to Kann et al. (2018), the neural segmentation model improves over existing segmentation methods by a considerable margin. The neural model trained on the 45,000 unambiguous data outperformed the model trained on all the 600,000 segmentations, suggesting that the consistency of the data is more important than the quantity. The other segmenters clearly struggled with the long words, often splitting words into a combination of very long root, and very short morphs. FlatCat scored the highest of the existing methods on both F1 and accuracy.

Unfortunately, both the neural and rule-based models sometimes fail to segment the input word. This makes them unfit to use in a translation system; since some words are left unsegmented, and this leads to a very large vocabulary size which hurts the translation performance. Micher (2017) previously explored improving the coverage of the Uqailaut morphological analyser with the use of an RNN based approach. In Micher (2018), an SRNN extension to the Uqailaut morphological analyzer is used in an SMT system, and yields a statistically significant improvement for IU→EN translation compared to the unextended rule-based analysis. Similar to their approach, we combined the best performing models of the intrinsic evaluation, to construct a custom 3-step segmenter to improve the coverage. This method initially applies the rule-based segmenter. If the rule-based segmenter fails, it falls back on the Transformer (unambiguous 45K) model. For non-alphabetic tokens we apply the BPE 5K model, because the Transformer fails for these tokens.

Method	F1	Acc.	Fail (%)
M. Baseline	0.317	0.222	-
M. FlatCat	0.397	0.328	-
LMVR	0.296	0.240	-
Trf. (45K rand.)	0.378	0.297	-
Trf. (45K single)	**0.680**	**0.539**	0.09
Trf. (all 600K)	0.625	0.433	0.55
3-Step	0.741	0.696	-
3-Step + LMVR	0.292	0.258	-
Rule-based	0.716	0.681	11.50

Table 3: Results of the intrinsic evaluation for each segmentation approach. The F1 score is calculated on segmentation boundaries, while the accuracy is calculated on the full segmentation. The *fail* statistic signifies the percentage of words that the approach failed to reconstruct for the methods for which that can occur.

Preliminary experiments with this approach still resulted in a very large vocabulary size. To reduce the vocabulary size further and combine all steps into a single model, afterwards we perform vocabulary reduction using LMVR. We specify a lexicon size of 20,000, which results in an actual vocabulary size of 41,024. The vocabulary reduction applied to the 3-step model leads to a drop in F1 and accuracy. This could be either because the vocabulary reduction leads to fewer segmentation boundaries per word, or because LMVR changes the model too much.

4 Translation experiments

Unless mentioned otherwise, the translation models are trained using Marian (Junczys-Dowmunt et al., 2018) v1.9.0 on an Nvidia V100. The translation models use the `transformer` model type with default settings. We use the `ce-mean-words` cost function. We perform a validation run every 5,000 update steps and apply early stopping after the validation cost stalls 5 times in a row. The model with the best translation score on the validation set (Section 2) is stored for each experiment.

4.1 Constrained Systems

Our constrained systems can be divided into four groups according to the techniques used: tags, backtranslation and domain-specific data (section 4.1.1), morphological segmentation (4.1.2),

[5] https://marian-nmt.github.io/

[6] Out of the 600,000 words, Uqailaut produces unambiguous segmentations for 45,000 words

contextual word embeddings (4.1.3) and ensembling and fine tuning (4.1.4).

4.1.1 Initial Systems

In these systems, following Joanis et al. (2020), we segment the training data with BPE (Sennrich et al., 2016) separately on each language. 5,000 and 2,000 merges are performed on both languages for MT systems into EN and IU, respectively.

Table 4 shows our initial constrained systems and their results on the development set.

	IU→EN		EN→IU	
System	**News**	**Hansards**	**News**	**Hansards**
1	14.73	29.62	40.29	52.97
2	17.96	29.7	47.47	**54.20**
3	17.24	28.88	**51.31**	53.86
4	**22.24**	**30.05**	NA	NA

Table 4: Results of the initial constrained systems for both translation directions and both dev sets. The scores are BLEU (IU→EN) and CHRF (EN→IU). Best result shown in bold.

Initial Systems System 1 is trained on Hansards. System 2 adds newsdevtrain, oversampled (5 times) given its small size compared to the other corpus used for training, i.e. Hansards (see Table 1). This results in a notable improvement for news (over 3 points into EN and over 7 into IU) and, as expected, a minor difference for Hansards.

Tags System 3 differs from system 2 in that each source sentence is preprended with a tag (<H> for Hansards and <N> for news); this degrades results into EN, but improves results into IU considerably for news (almost 4 points), with minimal change to Hansards.

Backtranslation In system 4 different amounts of newscrawl 2019 were backtranslated and concatenated to the training data of previous systems 3 and 2, with (<B>) and without a tag, respectively. This system is used only for IU→EN and its best results were obtained with 1 million sentences without tags; compared to system 2, adding backtranslation results in over 3 points improvement for news (22.2 vs 18) and a smaller increase for Hansards (30 vs 29.7).

We also explored the use of backtranslation for EN→IU. CC (backtranslated into EN) was concatenated to the training data of the previous systems 3 and 2, with and without a tag, respectively. Results were very similar. We conjecture this was due to its limited size and noisy nature, since it is web crawled.

	IU→EN		EN→IU	
Model	**News**	**Hans.**	**News**	**Hans.**
BPE 5K	14.77	**28.31**	32.52	39.81
Morfessor	13.39	26.82	28.75	38.20
FlatCat	12.86	26.49	23.25	29.88
LMVR	14.98	27.50	**34.84**	**41.25**
Trf. (single)	11.31	24.56	31.34	39.33
3-St.+LMVR	**15.25**	28.06	34.51	40.54

Table 5: Results of the extrinsic evaluation for the selected segmentation methods. Scores for IU→EN are in BLEU, and for EN→IU are in CHRF. Best results for each dataset and metric are in bold. All models are trained only on the Hansard training data.

Topic-specific News Because the texts in both dev sets concern (mostly) events in Nunavut, we hypothesised that Nunavut-related news *only* from our backtranslated news might be beneficial. We selected only documents from the document-delimited version of newscrawl that contain any word from a topic list.[7] Topic words were picked due to being frequent in newsdevtrain and unambiguosly related to Nunavut. 2,845 newsstories were extracted, after preprocessing 150,472 sentences and 3,220,925 words. We trained systems with this topic-specific backtranslated news as well as a similar amount of news randomly selected. Contrary to our hypothesis, the random news outperformed topic-specific news: 18.92 vs 20.2 BLEU on the news part of the dev set.

4.1.2 Morphological segmentation

We train translation models for the segmentation methods described in Section 3. For these experiments, the English data was segmented using BPE with 5,000 merges. Results are reported in Table 5. Both models that use LMVR for vocabulary reduction perform well for translation into IU, outperforming BPE on both Hansard and News data. There seems to be no benefit from the use of a more morphologically correct segmenter, as the highest scoring segmenters on the intrinsic evaluation (Table 3) generally performed worse on the extrinsic evaluation.

Based on the results of this extrinsic evaluation, we decide to use the BPE, LMVR, and 3-Step seg-

[7] Baffinland, Inuit, Inuits, inuits, Inuktut, Inuktitut, Iqaluit, Kivalliq, Nunatsiaq, Nunavik, Nunavut and Savikataaq.

mentations in our best systems so far (system 3 into IU and 4 into EN, see Table 4). Different amounts of BPE merges were tried for EN. The best results were obtained with 32,000 into IU and 20,000 into EN, whose results are reported in Table 6. The LMVR segmenter improved the translation into IU for the Hansard data, but not for news. For translation into EN there was no improvement from using a different segmenter.

	IU→EN		EN→IU	
System	**News**	**Hans.**	**News**	**Hans.**
Sys. 4 & 3 resp.	**22.24**	**30.05**	**51.31**	53.86
LMVR	21.89	29.20	50.36	**54.45**
3-Step + LMVR	21.79	29.66	50.19	52.18

Table 6: Results of the constrained systems that use morphological segmentation for both translation directions and both dev sets. The scores are BLEU (IU→EN) and CHRF (EN→IU). Best results shown in bold. The IU→EN models are based on system 4, while the EN→IU models are based on system 3 (Section 4.1.1).

4.1.3 Contextual Word Embeddings

With the recent success of pretrained contextual embeddings in MT (Lample and Conneau, 2019; Zhu et al., 2020), we try using this technique for a polysynthetic language. Specifically, we use the XLM model (Lample and Conneau, 2019), not only as a means of having contextual embeddings, but also to leverage available monolingual data for the task. For our XLM experiments, pretraining uses both masked language modeling (MLM) and translation language modeling (TLM). For the NMT training step, we include both denoising and back-translation for the monolingual data, as well as the standard MT training with the parallel data. Both the pretraining step and the NMT step use the monolingual data and the parallel data.

Pretraining	**IU→EN**	**EN→IU**
No	**19.32**	48.36
Yes	18.58	**49.10**

Table 7: Comparison of pretrained and non-pretrained XLM systems on the News dev set. The scores are BLEU (IU→EN) and CHRF (EN→IU).

To observe the effect of language model pretraining, we train a model using the same data used in system 4 (see Table 4), with 10,000 BPE joins applied jointly to both languages.[8] See results in Table 7. Interestingly, the performance decreases for IU→EN but increases for EN→IU when pretraining is added. A possible explanation for this is that Inuktitut stands to benefit more from pretraining as it uses more of the total joint vocabulary (around 90% of the tokens compared to 70%).

To use the existing monolingual data (Section 2), we train XLM models with the News Crawl data for English and Common Crawl data for Inuktitut, as specified in Table 2. We also use Hansards and `Newsdevtrain` oversampled 5 times for parallel data. We try both tagging the data (with the Common Crawl data receiving its own tag, <C>) and leaving it untagged. We report the results in Table 8. The results indicate an improvement with tagged data in the EN→IU direction. This is consistent with our observations with Marian-run models (systems 2 and 3 in Table 4). The XLM model results show that despite removing back-translated parallel data, results are similar. This is almost certainly due to the on-the-fly back-translation present in the training scheme. The results for EN→IU are improved, which is likely due to even a small amount of Inuktitut Common Crawl data being indeed useful for training.

Tagged	**IU→EN**	**EN→IU**
No	**18.96**	48.9
Yes	16.76	**49.97**

Table 8: Results of the XLM models using monolingual data on the News dev set. Scores are BLEU (IU→EN) and CHRF (EN→IU).

The best result with XLM (19.32 BLEU for IU→EN) is almost 3 points behind the result of the system trained with Marian on the same data (22.24, system 5 in Table 4). A difference between these two systems is that XLM uses joint BPE (since the encoder is shared by both languages), while with Marian we used separate BPE models for each language, following Joanis et al. (2020). To have a fairer comparison, we train the same Marian model with joint BPE, which leads to a score of 21.43, still 2 points ahead of the XLM model.

This difference in performance can be attributed, we hypothesise, to two reasons: (i) the XLM models use a joint encoder and decoder for both languages so the model must learn to translate in both

[8] We apply BPE jointly as it follows the methods of Lample and Conneau (2019).

directions and (ii) differences in implementation of the Transformer model in both toolkits.

4.1.4 Ensembles

For our final submissions, we depart from the best system so far (3 into IU and 4 into EN) and experiment with the use of ensembling and fine-tuning techniques. While some systems that used morphological segmentation performed similarly to those with BPE, their ensembles lagged behind. We therefore focused on BPE-based systems. In the following experiments we varied the value of the decoder's penalty length based on results on the dev set (until now we had used the value 1.0): for IU→EN we use 0.8 for news and 1.4 for Hansards while for EN→IU 1.2 was used for both dev sets. The results are shown in Table 9.

	IU→EN		**EN→IU**	
System	**News**	**Hans.**	**News**	**Hans.**
best single system	22.38	38.41	51.83	54.35
ens normal	23.72	39.07	52.92	55.05
ens FT	24.01	**39.72**	53.19	55.31
ens normal + FT	**24.25**	39.67	**53.46**	**55.39**

Table 9: Results of the constrained systems that use ensembling (referred to as ens) and fine tuning (FT) for both translation directions and both dev sets. The scores are BLEU (IU→EN) and CHRF (EN→IU). Best results shown in bold.

Ensembles are built by training the same system with different seeds (4 into EN and 3 into IU) and picking the model from each training seed with the highest score. These bring consistent improvements for both directions and dev sets: from 0.66 points for IU→EN Hansards to 1.34 for news in the same direction (row "ens normal" in Table 9).

We fine tune on `newsdevtrain` on its own and together with backtranslated news (only into EN) for the news dev set and on Hansards for the Hansards dev set. The ensembles of fine-tuned models bring consistent improvements compared to ensembles of non fine-tuned systems (row "ens FT" versus "ens normal" in Table 9). Finally, ensembling both fine-tuned and no fine-tuned systems (row "ens normal + FT" in Table 9) pushes the scores further (except for Hansards IU→EN) though rather slightly.

4.2 Unconstrained Systems

4.2.1 Data Acquisition

We use three additional parallel corpora that we acquired. First, we use data from the Inuktitut magazine[9], which contains parallel articles about Inuit culture and society in Inuktitut (IU), English (EN), and French; we manually extracted the text (IU syllabics, romanized IU, and EN) from several recent issues. Second, we use data from a Kalaallisut (KL) magazine[10] containing parallel news articles in Danish (DA) and KL. These texts were also manually extracted. Thirdly, parallel data from 21 multilingual websites containing DA and KL texts, was crawled using bitextor[11].

4.2.2 MT with Unconstrained Data

These datasets are pre-processed just like the ones from the constrained setup. In addition, we select a subset using their sentence alignment confidence score.[12] The KL crawl is paired with Danish. We performed language classification on the Danish data using LangID[13], removing any sentence pairs not classified as Danish. Danish was translated into English with a pretrained DA→EN system[14] from OPUS-MT (Tiedemann and Thottingal, 2020). Dataset details are presented in Table 10.

		Words	
Corpus	**Sentences**	**EN**	**IU/KL**
IU Magazine	1134	29312	18152
KL Magazine	657	13009	7491
KL crawl	14778	277159	163468

Table 10: Preprocessed unconstrained parallel training data.

We added these corpora atop the best constrained systems (3 into IU and 4 into EN) one at a time and evaluated on the news part of the dev set. Table 11 shows the results. Into EN, adding IU magazine (for which we tried different oversampling values) did not improve results. Due to this and time limitations we did not add the remaining unconstrained

[9] *Inuktitut Magazine*, https://www.itk.ca/category/inuktitut-magazine/.

[10] *Atuagagdliutit*, https://timarit.is

[11] https://github.com/bitextor/bitextor

[12] The datasets were aligned with Hunalign, which provides a confidence score. We experimented with different thresholds and based on results on the dev set and used 0.4 for IU and KL magazines and 0.5 for KL crawl (Varga et al., 2007).

[13] https://github.com/saffsd/langid.py

[14] https://object.pouta.csc.fi/OPUS-MT-models/da-en/opus-2019-12-04.zip

data. Into IU, adding IU magazine (with a tag and oversampled 5 times) resulted in a slight improvement (51.9 vs 51.3). Adding to this KL magazine (also oversampled 5 times) degraded results, as did adding KL crawl (although to a lesser extent).

System	**IU→EN**	**EN→IU**
Best constrained (5, 3 resp.)	**22.24**	51.31
+ IU magazine	22.22	**51.88**
+ IU mag + KL mag		50.57
+ IU mag + KL crawl		51.27

Table 11: Results of the unconstrained systems for both translation directions and both dev sets. The scores are BLEU (IU→EN) and CHRF (EN→IU). Best results shown in bold.

5 Conclusions

This paper has reported on the systems submitted by the University of Groningen to the English↔Inuktitut translation directions of the news shared task at WMT 2020.[15] Our best results were obtained using well-established techniques, including oversampling domain-specific training data, backtranslation, tags, fine-tuning and ensembling.

The use of morphological segmentation (RQ1) led to results that were on par with those obtained by BPE in terms of automatic evaluation metrics. One problem is that existing morphological segmenters for low-resourced languages like Inuktitut suffer from poor coverage, which impedes making a complete comparison with more automatic methods. The extrinsic comparisons between segmenters showed that a more accurate morphological segmentation does not lead to improved translation performance. We further found that existing language agnostic segmenters struggle to produce correct segmentations on Inuktitut, and that neural methods appear to be more suitable for polysynthetic languages (cf. (Kann et al., 2018)) . Note also the importance of limiting the vocabulary size of morphological segmentation for MT, which could be explored further.

The use of additional data from Inuktitut did improve the results slightly, but not the addition of data from a related language, Greenlandic (RQ2). The fact that its usefulness was limited could be due to the fact that half of the test set was from a specific domain for which considerable amounts of data were already available to train (Hansards).

Finally, the use of contextual embeddings (RQ3), led to mixed results since it resulted in an improvement for one direction but a degradation for the other.

Acknowledgments

We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster. Thanks also to Ben Shaffrey, Barbera de Mol and Adna Bliek for help preparing the Inuktitut magazine data.

References

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *The Prague Bulletin of Mathematical Linguistics*, 108(1):331–342.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185.

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*

[15] We will provide links to the additional datasets we used in the camera-ready version.

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.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Manuel Mager, Elisabeth Mager, Alfonso Medina-Urrea, Ivan Vladimir Meza Ruiz, and Katharina Kann. 2018. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 73–83, Santa Fe, New Mexico, USA. 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.

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.

Víctor M. Sánchez-Cartagena and Antonio Toral. 2016. Abu-MaTran at WMT 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 362–370, Berlin, Germany. Association for Computational Linguistics.

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.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.