The University of Helsinki and Aalto University submissions to the WMT 2020 news and low-resource translation tasks

Yves Scherrer
University of Helsinki
yves.scherrer@helsinki.fi

Stig-Arne Grönroos
Aalto University
stig-arne.gronroos@aalto.fi

Sami Virpioja
University of Helsinki
sami.virpioja@helsinki.fi

Abstract

This paper describes the joint participation of University of Helsinki and Aalto University to two shared tasks of WMT 2020: the news translation between Inuktitut and English and the low-resource translation between German and Upper Sorbian. For both tasks, our efforts concentrate on efficient use of monolingual and related bilingual corpora with scheduled multi-task learning as well as an optimized subword segmentation with sampling.

Our submission obtained the highest score for Upper Sorbian → German and was ranked second for German → Upper Sorbian according to BLEU scores. For English–Inuktitut, we reached ranks 8 and 10 out of 11 according to BLEU scores.

1 Introduction

Our work is motivated by Grönroos et al. (2020), who provide a detailed study of different transfer learning and regularization approaches for low-resource machine translation. They focus on an asymmetric-resource scenario in which the target language is underresourced, but related to a higher-resource language that can be used in a multilingual setting. For example, in the English-to-Estonian task, Estonian is assumed to be a low-resource language (LRL) which is complemented by a second higher-resource target language (HRL), Finnish. Among the WMT 2020 shared tasks, the **German → Upper Sorbian** low-resource translation task exactly corresponds to this setup, with Czech being a high-resource language closely related to Upper Sorbian. We adapt the approach proposed by Grönroos et al. (2020) also to three slightly different scenarios: in the **Upper Sorbian → German** task, the low-resource language is on the source side, but can be complemented with Czech in the same way; for the **English → Inuktitut** task, no related high-resource language is available; and for **Inuktitut → English**, the low-resource language is on the source side and no high-resource language is available.

Grönroos et al. (2020) recommend the combination of the following techniques to reach optimal translation performance in their examined setup:

Scheduled multi-task learning The learning process is split in two phases. The first phase only sees data from the source and the HRL, whereas LRL data is only added in the second phase.

Backtranslation The addition of synthetic data has become a staple of neural machine translation. They recommend marking synthetic data and controlling its weight in the task scheduler.

Subword regularization Following Kudo (2018), each time a word is used during training, a new segmentation into subwords is sampled from the probabilistic segmentation model.

Monolingual tasks In order to benefit from more easily available monolingual data and to make the model more robust to noise, they propose to include denoising sequence autoencoder tasks. A first variant applies small changes to the input side of the corpus (e.g. word deletions, substitutions and reorderings). A second variant, called taboo sampling, relies on the subword regularization idea and generates two maximally different segmentations of the source and target text. For English–Inuktitut,[1] we extend this idea to a transliteration task between romanized and syllabic Inuktitut.

Subword regularization and taboo sampling require the subword segmentation to be based on

[1] We use dashes to refer to language pairs independently of translation direction.

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1129–1138
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

	Parallel		Monolingual		
Corpus	EN→IU	IU→EN	EN	IU	IU Translit.
NH train	771 382	771 382			771 382
Wikititles	455	455			455
NH unaligned (EN)		*319 045*			
NH unaligned (IU)	*356 005*			356 005	
NewsCommentary		*557 628*			
NewsCrawl 2019	*2 000 000*		1 000 000		
NewsDiscuss 2019	*2 000 000*		1 000 000		
CommonCrawl	*80 244*			80 244	
Total	1 208 086	5 648 510	2 000 000	436 249	771 837

Table 1: Training corpora sizes (number of lines) for the English–Inuktitut systems. Numbers in italics designate synthetic datasets whose source side is produced by backtranslation.

a probabilistic model. While subword regularization has been introduced in conjunction with SentencePiece, Grönroos et al. (2020) show that the EM+Prune variant of Morfessor (Grönroos et al., 2020) outperforms SentencePiece.

The paper is structured as follows. In Section 2, we present the datasets, their sizes and their usage in our submission. Section 3 reports additional experiments with different approaches to word segmentation. Section 4 provides more details about our multi-task approach and the underlying NMT architecture. Section 5 summarizes the results.

2 Data

Both the Inuktitut–English and Upper Sorbian–German tasks can be qualified as low-resource settings, with less than 800K (deduplicated) parallel training instances for the former and 60K for the latter. For both tasks, we follow the constrained setting, which limits the allowed data to those made available on the WMT website. In this section, we present the parallel and monolingual resources that we used for our systems.

2.1 Inuktitut–English

Training data The training resources for the Inuktitut–English tasks are summarized in Table 1. Two allowed parallel resources are provided, the training part of the Nunavut Hansard (NH) corpus (Joanis et al., 2020) and the small WikiTitles corpus. Since the NH training corpus contained a significant proportion of duplicates and preliminary experiments suggested a slight adverse effect of duplicates, we removed them with the OpusFilter tools (Aulamo et al., 2020). We also cleaned the WikiTitles corpus, removing Inuktitut entries not in syllabic script and identical entries. The Inuktitut side of both training corpora was also used to create a parallel corpus for the romanized ↔ syllabic transliteration task. The romanized version was converted from the syllabic one using the *uniconv* + *iconv* pipeline proposed by the corpus providers.

The NH corpus contains a large amount of unaligned data, which we used as additional monolingual corpora. We removed all sentences that were already covered by one of the parallel NH datasets. The English and Inuktitut parts were processed separately. Both parts were backtranslated to the other language using baseline models trained on the parallel corpora, and filters were applied to both sides of the parallel datasets (see below). The Inuktitut unaligned data was used both as a monolingual dataset and as a synthetic parallel dataset for the EN→IU task, whereas the English unaligned data was only used as a synthetic parallel dataset for the IU→EN task (see Table 1).

Among the wealth of monolingual English data provided by WMT, we selected the NewsCommentary corpus and the 2019 sections of NewsCrawl and NewsDiscuss. We produced Inuktitut backtranslations for NewsCommentary and for 2M sentences each (after filtering) of the NewsCrawl and NewsDiscuss corpora. Of the latter two corpora, we held out distinct sets of 1M sentences each for monolingual tasks.

In terms of monolingual Inuktitut data, besides the unaligned NH data, the organizers only provided a CommonCrawl dump. This corpus was again backtranslated to English and filtered. The resulting corpus was used both as a monolingual

	Parallel				Monolingual		
Corpus	DE→HSB	HSB→DE	DE↔CS	HSB→CS	DE	HSB	CS
Training	60 000	60 000					
Europarl		*560 608*	567 422	*568 573*			
JW300		*1 114 024*	1 140 474	*1 161 656*			
NewsComm.			184 341	*185 132*			
Tatoeba		*4 425*	4 431	*4 448*			
Sorb. Inst.	*334 643*					334 643	
Sorb. Web	*94 980*					94 980	
Witaj	*218 249*					218 249	
NewsComm. (mono)		*389 199*			389 199		184 341
NewsCrawl 2018		*11 529 295*			11 529 295		6 723 691
NewsCrawl 2019		*9 041 245*			9 041 245		9 508 788
Total	707 872	22 698 796	1 896 668	1 919 809	20 959 739	647 872	16 416 820

Table 2: Training corpora sizes (number of lines) for the German–Sorbian systems. Numbers in italics designate synthetic datasets whose source side is produced by backtranslation.

dataset and as a synthetic parallel dataset for the EN→IU task.

Validation data We used the NH *dev* partition as primary validation set, and the *devtest*, *test* and *NewsDev2020* as secondary validation sets.

Preprocessing All datasets were processed with a translation-direction-specific pipeline. Inuktitut spelling and apostrophe normalization scripts were applied both on source and target sides. The Moses punctuation normalization script was applied only to the English target sides of the parallel corpora. No further preprocessing or tokenization was applied.

Filtering The monolingual and backtranslated parallel corpora were filtered with OpusFilter (Aulamo et al., 2020). The main purpose of this step was to remove too short (i.e., less than 1 word or less than 5 characters on either side) and too long sentences (i.e., more than 300 words or 3000 characters on either side). Furthermore, since crawled input data could be noisy and backtranslation could produce suboptimal results for certain sentences, we applied an additional language model filter based on 5-gram language models trained on the NH training part. Sentences with an average character cross-entropy higher than 30 on either side were removed.

2.2 Upper Sorbian–German

Training data The training data for the Upper Sorbian–German tasks are summarized in Table 2. The organizers provide a parallel German–Sorbian corpus of 60k sentence pairs that we use without further filtering or processing. Moreover, we use four sources of parallel German–Czech data for both directions: the Europarl and JW300 corpora provided on OPUS, as suggested by the organizers, and additionally the Tatoeba and NewsCommentary corpora, which are also available through OPUS (Tiedemann, 2012). The German side of three datasets[2] is backtranslated to Upper Sorbian using a baseline system. The Czech side of the four datasets is backtranslated to Upper Sorbian using an unsupervised character-level translation system (see below). Length filters are applied to all data from external resources (see below).

The organizers provide three monolingual Sorbian corpora: *Sorbian Institute*, a Sorbian *Web Crawl*, and *Witaj*. All corpora are backtranslated to German using a baseline system and filtered.

As monolingual German and Czech resources, we selected the NewsCommentary corpus and the 2018 and 2019 sections of NewsCrawl. These datasets were again filtered. The German datasets were backtranslated to Sorbian.

Validation data We use the *dev* partition as primary validation data and the *devtest* partition as secondary validation data (2000 sentence pairs each).[3]

[2] The full (i.e., unaligned) German version of NewsCommentary is also backtranslated, see below.

[3] The validation and test data for Sorbian consist of fairly short and syntactically simple sentences, which explains why even baseline systems such as those reported in Table 4 obtain BLEU scores around 50.

		EN→IU BLEU			IU→EN BLEU		
	Segmentation model and parameters	Dev	Devtest	Test	Dev	Devtest	Test
0	BPE, raw data, 2k+2k/5k+5k separate, no sampling	24.2	17.9	19.3	41.4	31.4	35.0
1	SentencePiece, raw data, 20k+20k separate, no sampling	23.1	16.9	18.4	36.8	27.2	30.9
2	SentencePiece, raw data, 5k+5k separate, no sampling	24.3	18.0	19.3	40.7	30.9	34.3
3	SentencePiece, raw data, 10k joint, no sampling	24.1	18.0	19.5	40.8	30.8	34.3
4	SentencePiece, dedup data, 10k joint, no sampling	24.2	17.7	19.0	40.7	30.8	34.4
5	SentencePiece, dedup data, 10k joint, with sampling	24.0	17.8	19.2	40.6	30.7	34.4
6	Morfessor, dedup data, 10k joint, no sampling	24.1	17.6	19.0	40.5	30.2	33.9
7	Morfessor, dedup data, 10k joint, with sampling	24.4	18.1	19.3	40.5	30.5	34.2

Table 3: Segmentation model experiments for English–Inuktitut. The baseline model (0) was trained using a Sockeye Transformer with default settings, whereas models 1–7 were trained using OpenNMT-py Transformers with default settings. The segmentation models were trained on the raw or deduplicated versions of the NH training corpus.

For the training phases using exclusively German and Czech data, we use the aligned WMT-News corpus (20 549 sentence pairs), made available on OPUS, as validation set.

Filtering A simple length filter was applied to all corpora sourced from OPUS: sentence pairs where at least one side is empty or longer than 300 words were removed. The same filter was also applied to parallel corpora obtained by backtranslation, which explains the slightly diverging numbers for identical corpora in Table 2.

The Sorbian web crawl was filtered by a 5-gram language model trained on the remaining original Sorbian data. Sentences with a cross-entropy higher than 50 were removed.

All corpus filtering tasks were implemented with OpusFilter (Aulamo et al., 2020). No other preprocessing or tokenization was applied.

Czech–Sorbian backtranslation The task organizers do not provide any Czech–Sorbian parallel corpora that could be used to train a baseline system for producing backtranslations. We therefore resort to unsupervised machine translation. Since Czech and Sorbian are closely related, we extract word n-grams from monolingual corpora and match them using string similarity and frequency criteria.[4] This results in a list of 620k distinct bigram pairs and 230k distinct trigram pairs. They are weighted by frequency to constitute a training corpus for a character-level Czech-to-Sorbian translation system. The translation system is based on bi-directional RNNs with two encoder and two decoder layers. In order to produce backtranslations, the Czech input sentences are chunked into overlapping trigram sequences, translated to Sorbian and merged back again.

3 Segmentation models

NMT models should ideally be able to represent the entire vocabulary of their source and target languages. The simplest solution however, in which word forms are represented as atomic vocabulary items, leads to sparse statistics, issues with out-of-vocabulary words, and heavy computational costs due to large vocabularies. Moreover, such word-level modeling does not allow the productive recombination of morphemes and is thus unsuitable for morphologically rich languages such as Inuktitut or Sorbian. In recent years, a consensus has emerged that NMT vocabularies should consist of subwords of variable size. Various unsupervised word segmentation algorithms have been proposed, among which byte-pair encoding (BPE) (Sennrich et al., 2016), SentencePiece (Kudo and Richardson, 2018), and several variants of Morfessor (Ataman et al., 2017; Banerjee and Bhattacharyya, 2018; Grönroos et al., 2018, 2020).

Besides the actual word segmentation algorithm, various parameters influence the quality of the resulting translation system:

- Separate word segmentation models for each language or one joint vocabulary for all languages. The joint approach scales better to multilingual models, and enables consistent segmentation of named entities and cognate

[4] We use Europarl, NewsCommentary, Taoeba and WMT-News as Czech monolingual corpora, and Training, Sorbian Institute and Witaj as Sorbian monolingual corpora.

	Algorithm	Segmentation model Training data (tokens)	Translation model Training data (lines)	DE→HSB BLEU Dev	 Devtest	HSB→DE BLEU Dev	 Devtest
1	SentencePiece	0.6M HSB + 0.7M DE	60k	56.93	49.76	57.11	48.74
2	Morfessor	0.6M HSB + 0.7M DE	60k	53.42	46.93	53.93	45.79
3	SentencePiece	8.4M HSB + 8.9M DE	60k	57.39	51.00	57.69	49.91
4	Morfessor	8.4M HSB + 8.9M DE	60k	55.34	48.99	55.51	47.61
5	SentencePiece	8.4M HSB + 8.9M DE + 8.4M CS	60k	57.82	51.30	58.45	49.86
6	Morfessor	8.4M HSB + 8.9M DE + 8.4M CS	60k	56.27	49.76	56.68	48.81
7	SentencePiece	8.4M HSB + 8.9M DE + 8.4M CS	708k / 1931k	61.90	55.06	62.41	53.78
8	Morfessor	8.4M HSB + 8.9M DE + 8.4M CS	708k / 1931k	61.56	55.04	62.16	53.83

Table 4: Segmentation model experiments for German–Upper Sorbian. All segmentation models are joint models with 20 000 units, but trained on variable amounts of data. All translation models are OpenNMT-py Transformers with default settings with active subword sampling, trained either without (1–6) or with (7–8) additional backtranslations.

words across languages, assuming they are written in the same script.

- The chosen vocabulary size and the amount of training data from which the segmentation model is learned. Denkowski and Neubig (2017) recommend a vocabulary size of 32k units, trained jointly on all languages, for normal-sized datasets. In contrast, Ding et al. (2019) obtain the best results with small vocabularies of only 500 units in low-resource scenarios. Optimal vocabulary size varies thus depending on the size of the parallel and monolingual data.

- If the segmentation algorithm is based on a probabilistic model (such as SentencePiece or Morfessor), it can be used to sample different segmentations for any given word. This technique is known as subword regularization (Kudo, 2018) and has been shown to improve the robustness of translation models.

Grönroos et al. (2020) tested various segmentation model configurations on a multilingual translation task and obtained best results with Morfessor EM+Prune, followed by SentencePiece and BPE. Furthermore, when trained on the same amount of data and using subword regularization, the vocabulary size (tested between 5K and 20K entries) turned out to be irrelevant for both SentencePiece and Morfessor EM+Prune.

We carried out some additional experiments with the English–Inuktitut task, which differs from their setup in the sense that the languages use different scripts and there is no third language involved. Table 3 compares different parameter settings with the baseline results provided by the organizers (Joanis et al., 2020). A first set of experiments shows that the vocabulary size does matter when not using subword sampling (1 vs 2), but that separate and joint segmentation models perform equivalently (2 vs 3). SentencePiece does not perform better than BPE (2 vs 0), although different preprocessing choices may be responsible for the generally lower results obtained in the IU→EN direction. The second set of experiments shows that Morfessor EM+Prune lags slightly behind SentencePiece when not using sampling (6 vs 4), but that sampling has a more beneficial effect to Morfessor EM+Prune than to SentencePiece (7 vs 6, 5 vs 4).

For German–Upper Sorbian, the setup differs from Grönroos et al. (2020) with respect to the amount of available training data. We therefore ran additional experiments to measure the impact of both the training data used for the segmentation model and the training data used for the translation model. Table 4 summarizes our findings. All experiments are based on joint word segmentation models with a total of 20K vocabulary items.

When training both the segmentation model and the translation model on the provided parallel data (experiments 1 and 2), SentencePiece performs much better than Morfessor EM+Prune. The addition of monolingual training data for the segmentation model (experiments 3 and 4) helps both segmentation algorithms about equally well (+ 1–2 BLEU). In contrast, the further addition of Czech data for the segmentation model (experiments 5 and 6) benefits Morfessor more than SentencePiece on average.[5] Finally, augmenting the trans-

[5] The additional monolingual Sorbian data comes from the

				EN→IU BLEU		IU→EN BLEU	
	Training data	Weighting	Monoling. tasks	NH Dev	Newsdev	NH Dev	Newsdev
1	EN↔IU + BT	—	—	24.13	15.72	41.07	32.86
2	EN↔IU + BT	✓	Noise + Translit.	*25.15	*15.95	**41.89**	33.47
3	EN↔IU + BT	✓	Noise + Taboo	**25.28**	**16.15**	*41.67	***33.49**

Table 5: Inuktitut translation experiments. Systems marked with * were used for the final primary submissions.

lation model training data with backtranslations obviously increases the overall translation scores, but also brings Morfessor EM+Prune on par with SentencePiece.

We were thus not able to reproduce the substantial gains in translation quality with Morfessor EM+Prune observed by Grönroos et al. (2020). Rather, we found that SentencePiece was generally more robust to different data conditions and setups. Nevertheless, Morfessor EM+Prune remains competitive with its default parameters if subword sampling is enabled and the training data are carefully chosen. For the final Inuktitut models, we decided to used configuration 7 from Table 3, since it allowed us to use monolingual tasks relying on subword sampling. For the final Sorbian models, we used configuration 8 from Table 4.

4 Translation models

All our models are based on the Transformer architecture and use, by and large, the same hyperparameters as Grönroos et al. (2020). The Transformer contains 8 encoder and 8 decoder layers with 16 attention heads each. The hidden layer size is 1024, the filter size 4096. The minibatch varies between 7200 and 9200 tokens, depending on the task, and gradients are accumulated over 4 minibatches. All models were trained for 200 000 steps, which corresponded to 5–7 days training time on a single V100 GPU. The best savepoint was selected on the basis of development set accuracy; this measure turned out to be more stable than development set BLEU score.

We use the *dynamicdata* branch of the OpenNMT-py toolkit (Klein et al., 2017) for our experiments.[6] This branch provides the necessary adaptations for the techniques introduced by Grönroos et al. (2020): scheduled multi-task learning requires the ability to adjust the task mix during training, whereas subword regularization and the denoising sentence autoencoder task require sampling fresh noise for each minibatch.

The experiments presented in Tables 3 and 4 already confirmed the positive impact of subword regularization and backtranslation. Row 1 of Tables 5 and 6 provide baseline results with these two techniques. Backtranslated training instances are marked with a special token.

Scheduled multi-task learning As row 2 in Table 6 shows, the mere inclusion of a German↔Czech task with language labels but without any task scheduling already increases BLEU scores by 1.5 points. However, simple transfer learning setups such as this are prone to catastrophic forgetting, especially in low-resource settings such as ours.

Kiperwasser and Ballesteros (2018) propose a general strategy called scheduled multi-task learning, in which different tasks are mixed according to a task-mix schedule. Grönroos et al. (2020) propose a partwise constant task-mix schedule with an arbitrary number of steps, any of which can be mixing multiple tasks. This flexibility is useful when training with a large number of heterogeneous tasks: multiple language pairs with different amounts of data, data from different domains (oversampling the in-domain data), natural vs synthetic (e.g. back-translated) data, and auxiliary tasks (e.g. autoencoder).

A training schedule with two phases (row 3 in Table 6) further increases scores slightly. Details of the schedule and the task weights are given in Table 7.

Witaj and Sorbian Institute corpora. We added an equivalent amount of German data from NewsCommentary and WMT-News. The Czech data also stems from NewsCommentary and WMT-News and is complemented by a subset of Czech Europarl.

[6] https://github.com/Waino/OpenNMT-py The functionality of the *dynamicdata* branch is included by default in the upcoming release v2.0 of OpenNMT-py, albeit in a different implementation.

	Training data	Weight./Schedul.	Monoling. tasks	DE→HSB BLEU		HSB→DE BLEU	
				Dev	Devtest	Dev	Devtest
1	DE↔HSB + BT	—	—	61.56	55.04	62.16	53.83
2	DE↔CS + DE↔HSB + BT	—	—	63.15	56.71		
3	DE↔CS + DE↔HSB + BT	✓	—	***63.93**	*56.82	64.48	56.27
4	DE↔CS + DE↔HSB + BT	✓	Noise	63.84	56.45	***64.88**	*56.76
5	DE↔CS + DE↔HSB + BT	✓	Taboo	63.61	**57.11**	64.72	**56.96**

Table 6: Sorbian translation experiments. Systems marked with * were used for the final primary submissions.

4.1 Monolingual tasks

Denoising sequence autoencoder task. In the denoising autoencoder (Vincent et al., 2008; Hill et al., 2016) clean text is corrupted by sampling from a noise model, and fed in as a pseudo-source. The target is a reconstruction of the clean input. The goal of the autoencoder tasks is to use monolingual data to strengthen target language modeling in the decoder and source language understanding in the encoder. In addition, the autoencoder task acts as regularization. Noise has been used as a regularizer in many NLP techniques, including dropout (Srivastava et al., 2014), label smoothing (Szegedy et al., 2016), SwitchOut (Wang et al., 2018), and subword regularization (Kudo, 2018). Sampling fresh noise for each minibatch is important, especially in low-resource conditions where the small data set is reused for many epochs. The denoising sequence autoencoder has previously been applied to language model pretraining in BART (Lewis et al., 2019).

Typical noise models for denoising sequence autoencoder apply small changes to the input side of the corpus: local reordering (Lample et al., 2018), deletions (Iyyer et al., 2015), insertions (Vaibhav et al., 2019), substitutions (Wang et al., 2018), and masking (Devlin et al., 2019). Of these, our method applies local reordering and token deletion.

Taboo sampling segmentation task. Grönroos et al. (2020) propose taboo sampling, a noise model extending the subword regularization idea specifically for monolingual data. It takes in monolingual text and generates two maximally different segmentations, e.g. *dys + functional* on the source side and *dysfunction + al* on the target side. During taboo sampling, all multi-character subwords used in the first segmentation have their probability temporarily set to zero, to ensure that they are not used in the second segmentation.

Transliteration task. As an alternative to taboo sampling, we take advantage of the fact that Inuktitut can be written in two different scripts, romanized and syllabic. Since the segmentation model is trained only on syllabic Inuktitut (and the occasional romanized proper name occurring on the English side of the NH corpus), we assume that the same word will be segmented very differently in the two scripts, leading to a similar effect as taboo sampling. We include a romanized→syllabic transliteration task in the EN→IU model, and a syllabic→romanized task in the IU→EN model.

Experiments 2–3 in Table 5 as well as experiments 4–5 in Table 6 explore different combinations of monolingual tasks. For Inuktitut, the addition of monolingual tasks increases BLEU scores markedly, but there is no clear winner between the transliteration and taboo tasks. For Sorbian, the monolingual tasks only help when translating towards German, but not when translating towards Sorbian. One reason for this somewhat surprising finding could be that the Sorbian monolingual data is identical with the Sorbian target of the backtranslations, so that no additional data is added with the monolingual tasks.

5 Submissions and results

For the best-performing configurations, we trained two models each, one (“basic”) with the hyperparameters listed above, and an alternative one with relative position distance clipping at 4 (see Shaw et al., 2018). However, this setting did not yield any consistent accuracy gains or losses.

For the **Inuktitut** task, we submitted single systems of settings 2 and 3 for both directions. For EN→IU, the alternative model of setting 2 obtained the best scores on the test set (10.1 BLEU / 0.301 chrF), whereas for IU→EN, the basic model of setting 3 obtained the best scores on the test set (23.0 BLEU / 0.455 chrF). Among the 11 primary submissions in both translation directions, our sub-

	EN→IU	IU→EN		DE→HSB		HSB→DE	
Training steps	0–200k	0–200k	Training steps	0–60k	60–200k	0–60k	60–200k
Bilingual	45%	45%	Bilingual DE↔CS	90%	50%	85%	25%
Backtranslation	40%	40%	Bilingual DE↔HSB		20%		30%
Noise EN	5%	5%	Backtr. DE↔HSB		20%		30%
Noise IU	5%	5%	Backtr. HSB→CS			5%	5%
Taboo IU / Translit. rom.→syll.	5%		Noise / Taboo DE	5%		5%	5%
Taboo EN / Translit. syll.→rom.		5%	Noise / Taboo CS	5%		5%	
			Noise / Taboo HSB		10%		5%

Table 7: Task schedules for the Inuktitut (left) and Sorbian (right) experiments.

mission obtained rank 8 for EN→IU and rank 10 for IU→EN in terms of (inofficial) BLEU scores. It will be instructive to examine the manual evaluation results and the other system descriptions to identify the reasons behind these rather disappointing results.

For the **Sorbian** task, we submitted ensembles of the basic and alternative models. Setting 3 turned out to be the best choice for DE→HSB (57.9 BLEU / second rank), and setting 4 for HSB→DE (59.6 BLEU / first rank). For both directions, ensembling has raised the BLEU scores by 0.6. Our submissions would obtain first rank in both directions if only single systems were considered.

6 Conclusion

In this work, we tested various methods for low-resource machine translation proposed by Grönroos et al. (2020) on the English–Inuktitut and German–Upper Sorbian tasks in WMT 2020. In particular, we investigated several subword segmentation approaches and the inclusion of monolingual tasks.

In terms of **subword segmentation**, we were not able to reproduce the reported gains for the Morfessor EM+Prune method over SentencePiece. We obtained comparable results with both methods though. We also found that increasing the size of segmentation model training data was useful, and that Morfessor EM+Prune was more sensitive to training data size than SentencePiece. Furthermore, we obtained slight improvements from subword sampling, confirming earlier results.

During the development phase, we also found curious interactions between the subword vocabulary size and different NMT toolkits. We were able to reproduce the organizer-provided Inuktitut baselines with both small and large vocabularies using the Sockeye toolkit, but obtained significantly lower scores with OpenNMT-py and large vocabularies, even after harmonizing the training hyperparameters between toolkits. With small subword vocabularies, OpenNMT-py became competitive again.

The inclusion of **monolingual tasks** yielded clear improvements for the Inuktitut experiments. The noise model had the most positive effect, whereas the transliteration and taboo sampling tasks showed minor effects. In contrast, the effect of the monolingual tasks on the Sorbian experiments was more subtle. The **two-phase training schedule** introduced by Grönroos et al. (2020) proved useful in the Sorbian experiments.

Acknowledgments

This work is part of the FoTran project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 771113).

The authors gratefully acknowledge the support of the CSC – IT Center for Science, Finland, for computational resources. We also acknowledge NVIDIA for their GPU grant.

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.

Mikko Aulamo, Umut Sulubacak, Sami Virpioja, and Jörg Tiedemann. 2020. OpusTools and parallel corpus diagnostics. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3782–3789, Marseille, France. European Language Resources Association.

Tamali Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the Second Workshop on Subword/Character Level Models*, pages 55–60. Association for Computational Linguistics.

Michael Denkowski and Graham Neubig. 2017. Stronger baselines for trustable results in neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 18–27, Vancouver. 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.

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.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2018. Cognate-aware morphological segmentation for multilingual neural translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 386–393, Belgium, Brussels. Association for Computational Linguistics.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2020. Morfessor EM+Prune: Improved subword segmentation with expectation maximization and pruning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3944–3953, Marseille, France. European Language Resources Association.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2020. Transfer learning and subword sampling for asymmetric-resource one-to-many neural translation. ArXiv:2004.04002 [cs.CL].

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP) (Volume 1: Long Papers)*, pages 1681–1691.

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.

Eliyahu Kiperwasser and Miguel Ballesteros. 2018. Scheduled multi-task learning: From syntax to translation. *Transactions of the Association for Computational Linguistics*, 6:225–240.

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

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. ArXiv:1910.13461 [cs.CL].

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.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1916–1920.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning (ICML)*, pages 1096–1103.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. Switchout: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 856–861.