Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4933–4943
Marseille, 20-25 June 2022

Evaluating Pre-training Objectives for Low-Resource Translation into Morphologically Rich Languages

Prajit Dhar, Arianna Bisazza, Gertjan van Noord
Center for Language and Cognition Groningen (CLCG)
University of Groningen
{p.dhar, a.bisazza, g.j.m.van.noord}@rug.nl

Abstract

The scarcity of parallel data is a major limitation for Neural Machine Translation (NMT) systems, in particular for translation into morphologically rich languages (MRLs). An important way to overcome the lack of parallel data is to leverage target monolingual data, which is typically more abundant and easier to collect. We evaluate a number of techniques to achieve this, ranging from back-translation to random token masking, on the challenging task of translating English into four typologically diverse MRLs, under low-resource settings. Additionally, we introduce Inflection Pre-Training (or PT-Inflect), a novel pre-training objective whereby the NMT system is pre-trained on the task of re-inflecting lemmatized target sentences before being trained on standard source-to-target language translation. We conduct our evaluation on four typologically diverse target MRLs, and find that PT-Inflect surpasses NMT systems trained only on parallel data. While PT-Inflect is outperformed by back-translation overall, combining the two techniques leads to gains in some of the evaluated language pairs.

Keywords: low resource nmt, morphology, inflection

1. Introduction

Machine translation has improved significantly in the last decade, mostly due to Transformer based Neural Machine Translation (NMT) models. Since their proposal by Vaswani et al. (2017), transformers have dominated the machine translation field and have established state-of-the-art results for various language pairs, even achieving human parity in some language pairs like Chinese→English (Hassan et al., 2018) and English↔Czech (Popel et al., 2020)[1].

The improvement in the quality of machine translation has not just been limited to high-resource language pairs. Sennrich and Zhang (2019) and Araabi and Monz (2020) show that the quality of recurrent and Transformer-based NMT, respectively, can be considerably improved by a careful selection of the hyperparameters. Multilingual NMT systems such as Aharoni et al. (2019) and Zhang et al. (2020) have also led to improvements for low-resource languages. However, such gains are very uneven across target languages and strongly dependent on the presence of a closely related language in the training data, (e.g. Spanish for Galician and Russian for Belarussian). Moreover, training massively multilingual systems is computationally very expensive.

An important obstacle to improving the quality of NMT *into* low-resource languages is the rich morphology of many of these languages. In this paper, we focus on the task of translating from a morphologically poor language, English, into various Morphologically Rich Languages (MRLs) with very different morphological systems (Estonian, Lithuanian, Tamil, Turkish and German). We assume a common scenario where (i) only little parallel data is available for training but (ii) a sizeable amount of target monolingual data is available, and (iii) there are no closely related high-resource language pairs.

We consider various well-established ways to circumvent the lack of large parallel corpora by leveraging target monolingual data, and evaluate their effectiveness in the presence of very complex target morphologies. These techniques include: *back-translation* (Sennrich et al., 2016a), a very popular but also expensive data augmentation technique; its light-weight alternative *stupid back-translation* (Burlot and Yvon, 2018); and *random token masking* (Raffel et al., 2020) which has proved its effectiveness for a wide range of downstream tasks (Cooper Stickland et al., 2021).

Alongside this evaluation, we also aim to study if the addition of a linguistically motivated pre-training objective aids low-resource translation in our scenario. We hypothesize that learning to correctly inflect target words in MRLs is a major challenge for NMT models trained on limited data. To address this, we propose to exploit existing morphological analyzers and lemmatizers, which exist for a wide range of languages (Straka, 2018; Kirov et al., 2018) and deliver reasonable quality even for languages where only little parallel data is available. The new technique, which we call PT-Inflect, consists of pre-training the NMT model on the task of *inflecting* the target language, or in other words, transforming a sequence of lemmatized target words into the corresponding sequence of surface forms. PT-Inflect offers a cost-effective way to gather synthetic data, as it does not involve the training of an additional NMT system, and is expected to provide complementary benefits.

Based on our experiments in two simulated low-resource settings, we find that PT-Inflect outperforms

[1] There are however doubts raised on these claims (Toral et al., 2018)

the baseline in all our language pairs. We also observe that adding more monolingual data up to 1M tokens yields further improvements. When comparing different pre-training objectives, the most computationally costly technique, back-translation, remains the most effective. And finally, combining back-translation with PT-Inflect in the same system leads to further gains in English→Lithuanian (as well as English→German, very low-resource setup).

2. Previous Work

Unlike pre-neural statistical translation systems, NMT systems require large amounts of parallel data to generate satisfactory translations. While parallel corpora might not be abundant for several languages, there may be monolingual data that could be leveraged. Several strategies have been proposed to exploit these resources. We discuss a few of them below. For a more detailed overview on the current state of low-resource translation see Wang et al. (2021).

Denoising using random token masking Massive language models pre-trained on monolingual data using the masked language modelling (MLM) objective (Devlin et al., 2018; Lample and Conneau, 2019; Shoeybi et al., 2019) have become the absolute state of the art for a wide range of classification tasks within NLP.

More relevant to Machine Translation, MASS (Song et al., 2019), BART (Lewis et al., 2020) and its multilingual version mBART (Liu et al., 2020) extend the idea of pre-training to a fully fledged encoder-decoder model, which can then be fine-tuned for various downstream sequence-to-sequence tasks including translation. Various forms of denoising and masking objectives can be used to pre-train such models. Several such objectives are given in Raffel et al. (2020). A notable objective is Random Token Masking, where the source side tokens are randomly replaced by a masking token before being fed to the NMT as input. We experiment with Random Token Masking as a pre-training objective in our experiments.

Cooper Stickland et al. (2021) fine-tune mBART models for low-resource translation with several configurations, such as freezing and increasing the encoder side weights. They report improvements in difficult language pairs like Nepali- and Simhala-to-English. For translations to English, Liu et al. (2021) adapted the mBART model and report improvements over mBART scores for English-to-Bengali and -Tamil.

Language modelling as a pre-training task Baziotis et al. (2020) have used pre-trained language models to aid low-resource NMT. They trained their language models on the target-side monolingual data. The target side information is fed to the translation model by employing a posterior regularization objective. They reported a +2.9 BLEU improvement over the baseline system with their best configuration for English-Turkish.

Target Language	Token Similarity (%)	Accuracy (%)
Estonian	60.9	57.2
Lithuanian	66.8	60.8
Tamil	56.3	55.7
Turkish	54.3	55.0
German	67.2	67.4

Table 1: Statistics from the inflection pre-training task. Token similarity: percentage of target tokens that remain unmodified by lemmatization, in different target languages. Lower percentages suggest a more complex morphology. Accuracy denotes the total percentage of tokens correctly re-inflected by our PT-Inflect models. Languages are sorted by increasing token similarity.

Data augmentation using monolingual data Back-translation (BT) as a technique to generate synthetic data was introduced by Sennrich et al. (2016a). Using a NMT system that was initially trained on a parallel corpus, target side monolingual sentences are translated to the source-side language. The resulting pairs of (back-translated, original) sentences are then used to augment the existing parallel data and train a better system.

Since then, subsequent papers have suggested improvements to the traditional BT algorithm: Caswell et al. (2019) find that prepending the back-translated sentences with a dedicated tag aids the NMT system. Edunov et al. (2018) report that introducing noise to the source sentences only is beneficial in high-resource settings.

Some cheaper alternatives to BT have also been proposed: Currey et al. (2017) find that simply copying the target sentences to the source side outperforms a BT system for English→Turkish. Burlot and Yvon (2018) refine this technique by copying the target sentences to the source side and appending each token on the source side with a language tag. They call this technique Stupid Back-Translation (StupidBT), and show it can significantly increase the BLEU score, compared to NMT systems only trained on parallel data.

Linguistically motivated auxiliary tasks With the aim to improve machine translation, several pre-training tasks have been envisioned. Syntactic information such as part-of-speech or lemmas are plugged into a component of the NMT (Shavarani and Sarkar, 2021; Eriguchi et al., 2017; Chakrabarty et al., 2020). Others have proposed inserting linguistic information via a task, either in a multi-task setting (Sánchez-Cartagena et al., 2021; Armengol-Estapé and Costa-jussa, 2021; Shearing et al., 2018) or prior to the translation phase (Zhou et al., 2019; Li et al., 2020). In particular, Shearing et al. (2018) find that including linguistic information such as glosses to the training data greatly improve low-resource machine translation for the pairs Russian-English and Spanish-English.

In this work, we choose to experiment with the *de facto*

	Lemmatized sentence T$_{lemmas}$ (with EN glosses)		Inflected sentence T$_{original}$ (with EN translation)
ET	Farish ei olema esitama üks süüdistus	→	Farishile ei olnud esitatud ühtegi süüdistust
	Farish no to be brought single charge		*No charges had been brought against Mr Farish*
LT	kelias m8 po policija incidentas vėl atidaryti	→	Kelias M8 po policijos incidento vėl atidarytas
	way m8 after police incident again open		*M8 fully reopens after police incident*
TR	bu ol	→	Bu olmayacak
	this to be		*That's not going to happen*
TA	kaṭṭāyam avar nam iḻivupaṭuttu *viṭakkūṭu(m)	→	kaṭṭāyam avarkaḷ nammai iḻivupaṭutta viṭakkūṭātu
	force he us humiliate may		*We must not let them humiliate us*
DE	dass man er|es|sie selbst treu bleiben müssen	→	Dass man sich selbst treu bleiben muss
	that one it self true to remain must		*That you have to be true to who you are*

Table 2: Examples of PT-Inflect in our five target languages: pairs of (T$_{lemmas}$ →T$_{original}$) sentences are provided to the NMT model during pre-training. Tamil is transliterated with the ISO 15919 standard. Incorrectly lemmatized words are marked with an asterisk and the corrections are given within parenthesis.

standard technique for data augmentation in MT, back-translation; its light-weight alternative, stupid back-translation; and random token masking as another light-weight, non-MT specific pre-training technique that has become very popular since the advent of large pre-trained language models. As the linguistically motivated pre-training task, we introduce a novel objective that is specifically intended to improve translation into MRLs.

3. Inflection Pre-Training (PT-Inflect)

Based on the observation that morphological analysers such as UD-Pipe (Straka and Straková, 2017) and Unimorph (Kirov et al., 2018) exist for many low resource languages, we propose a new pre-training technique, called Inflection Pre-Training (or PT-Inflect), which leverages such analyzers to generate synthetic training data. The UD-Pipe tool is trained on the official Universal Dependency treebanks and is capable of performing various linguistic tasks such as lemmatization, part-of-speech tagging, dependency parsing, etc. While several other lemmatizers exist, we opted for UD-Pipe as it is available for all the languages in our study.

Firstly, the UD-Pipe models are run on the target monolingual data (T$_{original}$). We then take the resulting lemmatized sentences (T$_{lemmas}$) as the source side text, while the original sentences (T$_{original}$) constitute the target side of the synthetic data. Examples of PT-Inflect sentence pairs can be seen in Table 2. This auxiliary inflection task differs from a translation task in at least two ways, namely: (i) there is no reordering involved, and (ii) the input and output sequences have exactly the same number of tokens. Furthermore, many lemmas are identical to the surface form, and only need to be copied over during inflection. The percentage of tokens that remain unmodified by lemmatization is provided in Table 1, as calculated on the monolingual training data described in Section 5.1. Given the number of inflected forms a word can possess (Table 3), we expect the NMT will learn about the morphology and syntax of the target language during pre-training. Additionally the accuracy is calculated. The accuracy is akin to the Word Edit Rate without re-ordering, i.e. the total number of tokens correctly translated without considering the position of the tokens.

Training scheme We consider two ways to use the dataset of lemmatized-original sentence pairs: In **pre-training**, we initially train the NMT models on the artificial data (T$_{lemmas}$ → T$_{original}$). Once the models have converged, they are then trained on the parallel data. In **joint training** we directly train the models on a mix of original parallel data and PT-Inflect data.

4. Target Languages

We choose five target languages that differ widely by morphological typology and morphological complexity. The first four languages display very high morphological complexity: Estonian is a Finnic language with mixed agglutinative-fusional morphology, Lithuanian is a higly inflecting-fusional language belonging to the Baltic family, Tamil and Turkish are extensively agglutinative languages belonging to the Dravidian and Turkic family, respectively. To put results into perspective, we also include German, a moderately inflecting-fusional language that belongs to the same family as the source language, English.

For each language, Table 3 presents statistics on the number of tokens and English/Target-Language token ratio, as calculated on the parallel training data (Section 5.1). We also report Morphology Counting Complexity (MCC) values (Sagot, 2013), which correspond to the number of unique morphological categories found in each language. In addition to the MCC values computed on UD and reported in Cotterell et al. (2019), we also include MCC values calculated on the UniMorph datasets.

The following observations can be made: (i) Tamil has the biggest source/target token ratio, with around

Language	Tokens(k)	EN/Trg Token Ratio	MCC		Type/Token Ratio	UDPipe Acc.
			UniMorph	UD		
Estonian	72	1.39	108	110	0.340	90.5
Lithuanian	81	1.23	139	123	0.383	85.3
Tamil	26	3.81	360	201	0.422	84.1
Turkish	83	1.21	883	140	0.307	90.0
German	95	1.05	37	38	0.266	95.4
English	100	-	5	6	0.174	94.9

Table 3: Statistics involving the source and target languages. The score for MCC (UniMorph) is unavailable for Tamil as Tamil is not found in the current version of UniMorph. Instead we have included the number of possible noun and verb forms for Tamil as mentioned in Sarveswaran et al. (2019).

four English tokens for every Tamil one. (ii) All languages, except English and German, have extremely high MCC values (over 100 unique morphological categories). (iii) All the target languages have a higher type/token ratio when compared to English, with Tamil having the most skewed ratio amongst them.

UD-Pipe lemmatization accuracy for each target language, as reported here, is also shown in Table 3. Accuracy varies considerably across languages, ranging from 96.4% in German to 84.1% in Tamil.

We decided to conduct experiments on these languages under simulated low-resource settings rather than on truly low-resource languages for various reasons: First, the selected languages with their varying morphology types and language families, are a testbed to see how the pre-training objectives fare in very different contexts. Secondly, from the perspective of datasets, our chosen languages are all present in WMT evaluations (more specifically, news translation evaluations), which increases replicability and comparability of the results across target languages. Finally, abundant monolingual data is available for all of the selected target languages. In the future, we would also like to experiment with truly low-resource and endangered languages for which lemmatizers already exist, such as Pashto and Occitan represented in WMT 2020 (Koehn et al., 2020) and 2021 (Akhbardeh et al., 2021) respectively, or Telugu represented at WAT20 (Nakazawa et al., 2020).[2]

5. Experimental Setup

5.1. Datasets

We use the data provided from the past three editions of WMT (Bojar et al., 2018; Barrault et al., 2019; Barrault et al., 2020). The corpus for Estonian is taken from WMT 2018, Lithuanian and Turkish from WMT 2019 and the Tamil and German corpora come from WMT 2020. Table 4 shows parallel data size and composition for each language pair.

For the monolingual data, we considered the resources listed by the last three editions of WMT. Given the relative copiousness of monolingual data, we resorted to using Common Crawl[3] for all target languages.

Low resource setting for parallel data To simulate a low-resource setting for all language pairs we choose two settings: 100k and 2M tokens. These token numbers are calculated on the source side (English). Then for each target language, the equivalent target sentences are extracted.

To control for variations that could arise due to training data sub-sampling (Liu and Prud'hommeaux, 2022), we run each experiment on five disjoint subsets of the parallel training data and report the average results.

Pre-processing Empty sentences and duplicates are removed, as well as sentence pairs with more than 150 tokens or having a target/source length ratio above 0.7. Sentences where the language identification score computed by fastText (Joulin et al., 2017) is below 0.4 are also discarded. The datasets are then shuffled and the 2M tokens (around 110k sentences) are randomly selected. A smaller subset of 100k tokens is then extracted from this, and constitutes our very-low resource setting. Similar pre-processing is performed on the monolingual data.

Subword segmentation Prior studies such as Ataman and Federico (2018) have presented evidence that linguistically-motivated vocabulary reduction (LMVR) techniques can outperform purely frequency-based techniques such as BPE (Sennrich et al., 2016b) for Turkish and other MRLs. However, more recent work (Dhar et al., 2020; Dhar et al., 2021; Saleva and Lignos, 2021) reports the opposite result for languages like Tamil and Kazakh. Given the mixed results, we opt for the widely used byte-pair-encoding (BPE) in our experiments. We use the implementation of BPE as provided by SentencePiece (Kudo and Richardson, 2018). For all our experiments the BPE models are trained together on the source and target sentences (i.e. there is one joint BPE model for each experiment). We opt for a sub-token dictionary size of 10k. This choice of dictionary size is based on the optimal settings for 10k dataset size from Araabi and Monz (2020).

[2] Pashto, Occitan, and Telugu are just three of the 142 languages currently covered by Unimorph analyzers (Kirov et al., 2018), see `https://unimorph.github.io`

[3] https://commoncrawl.org/

Trg. Language	#Sent.	Sources
Estonian	1M	Europarl (59%), Paracrawl (21%) and Tilde (20%)
Lithuanian	4.9M	Paracrawl (83%), Europarl (12%), Tilde (4%) and Wikititles (<1%)
Tamil	1.3M	JW (46%), UFAL (13%), Wikimatrix (11%), PIB (9%), Tanzil (7%), Wikititles (5%), others (9%)
Turkish	0.2M	SET Times (100%)
German	130M	Wikimatrix (74%), Europarl (23%) and Wikititles (4%)

Table 4: Number of parallel sentences available in each language, and relative data sources.

Effect of monolingual data size We also investigate the effect of monolingual data size on translation quality. Specifically, we consider two settings: when the parallel and the artificial data are of the same size (100k tokens) or when the artificial data is 10 times the size of the parallel corpora (1M). In order to overcome the size difference bias encountered during joint training, we over-sample the parallel data to match the artificial data size.

Evaluation and test sets We use the development sets and testsets provided by WMT, specifically: Estonian (newsdev-2018, newstest-2018), Lithuanian (newsdev-2019, newstest-2019), Tamil (newsdev-2020, newstest-2020), Turkish (newsdev-2016, newstest-2018) and German (newsdev-2020, newstest-2020). All development/testsets sets have around 2k sentences.

5.2. NMT Baseline

All the NMT models are based on the Transformer architecture and implemented using Fairseq (Ott et al., 2019). In particular, the configuration of our models is based on the optimized settings from Araabi and Monz (2020), who investigated optimal hyperparameters to train Transformer-based NMT models under low-resource settings. The encoder and decoder are set to 5 layers with embedding dimension of 512. The attention heads are reduced from the default 8 to 2 and the embeddings of the feed-forward neural network is 512 dimensions. Layer normalization is performed for both encoder and decoder layers. The dropout, attention dropout as well as the activation dropout are all set to 0.3. The batch size during training is set to 4096 tokens and the loss function is cross-entropy with label smoothing of 0.6. Given the extreme small size of the training data, all models are trained for 200 epochs with the early stopping criterion set to 5.[4]
The NMT systems trained only on the parallel corpora are considered as our BASELINE models.

5.3. Pre-training objectives

In all our experiments, pre-training follows the setup of the BASELINE models, with the only changes being to the training time and early stopping criteria. We found that some of the pre-training models took longer to converge and hence we set the maximum epochs to 300 epochs and an early stopping criteria 10. The dictionary size for all the models is 10K, as mentioned in Section 5.1.

Lang	Model	Training	Mono	BLEU	CHRF
ET	Base	-	-	3.4	22.3
	PT-Inf	Joint-	100k	3.5	23.1
	PT-Inf	Pre-	100k	3.8	26.4
	PT-Inf	Pre-	1M	**4.4**	**27.1**
LT	Base	-	-	3.0	27.8
	PT-Inf	Joint-	100k	3.1	28.1
	PT-Inf	Pre-	100k	3.1	28.3
	PT-Inf	Pre-	1M	**3.7**	**29.9**
TA	Base	-	-	1.7	20.2
	PT-Inf	Joint-	100k	2.0	24.4
	PT-Inf	Pre-	100k	2.4	25.7
	PT-Inf	Pre-	1M	**3.3**	**26.8**
TR	Base	-	-	2.4	19.9
	PT-Inf	Joint	100k	2.4	19.9
	PT-Inf	Pre-	100k	2.5	19.9
	PT-Inf	Pre-	1M	**3.3**	**21.0**
DE	Base	-	-	7.1	34.5
	PT-Inf	Joint	100k	7.6	39.4
	PT-Inf	Pre-	100k	7.5	39.4
	PT-Inf	Pre-	1M	**8.1**	**43.0**

Table 5: Comparison of PT-Inflect (PT-Inf) with baseline (Base) in the very low-resource setting (100k-token parallel training dataset). BLEU and CHRF scores are averaged over five disjoint training subsets. Monolingual data size (Mono) is given in number of tokens. The best system, for each language pair and score, is highlighted in boldface. Lang refers to the target language.

PT-Inflect As previously noted, the PT-Inflect data is generated with the UD-Pipe tool (Straka and Straková, 2017). The same configurations and hyper-parameters as the BASELINE models are used in both joint training and pre-training settings. Note that, for PT-Inflect, the BPE models are trained on the combination of regular parallel sentences and PT-Inflect data. This results in the source dictionary comprising English sub-tokens as well as lemmatized target sub-tokens. We also experimented with adding a special tag to denote syn-

[4] In preliminary experiments, the aforementioned hyper-parameter settings did not provide satisfactory results for the English-Tamil pair. We hence modified some of the hyper-parameter for this language pair based on Dhar et al. (2020), specifically: setting the activation dropout to 0.3, sharing the embeddings and training for 200 epochs.

thetic data as in Tagged-BT, however that did not improve performance.

Random Token Masking We follow (Raffel et al., 2020) to implement random token masking. We experiment with two masking rates that appeared to be beneficial in their translation experiments, namely: 15%, the value used in BERT (Devlin et al., 2018), and 50% to discern if more denoising leads to improvements in low-resource translation.

Back-translation We develop our BT systems using the TaggedBT approach (Caswell et al., 2019). For each language pair, a NMT system is trained in the reverse direction (i.e. English→ Target). These NMT systems are then used to generate the synthetic English sentences, given the target monolingual sentences. As with PT-Inflect, the BT systems are trained with the same configuration as the BASELINE models.[5]

Stupid Back-translation For Stupid-BT we implement the *copy-marked* technique from Burlot and Yvon (2018). This involves copying the target sentences to the source side and prepending each token with a unique tag to prevent the system from simply learning to copy the sentences token by token.

Regarding the computational cost of these techniques, BT is by far the most expensive one as it involves the training of an additional NMT model and the generation of a large number of back-translated sentences. PT-Inflect is considerably less expensive than BT, but requires a lemmatizer to be available in the target language. Finally, stupid BT and random token masking are the cheapest and fastest running techniques.

6. Results

In this section we present a number of experimental results: first, we investigate the effect of PT-Inflect on very low-resource translation (100k training tokens) and identify its optimal setup. Secondly, we run a comparative evaluation of various pre-training objectives, including PT-Inflect, in both very low- (100K) and low- (2M) resource settings. Lastly, we combine the best pre-training objectives into a single system to find out if they provide complementary benefits.

We use SacreBLEU (Post, 2018) and CHRF++ (Popović, 2015) for calculating the BLEU and CHRF scores, respectively. We consider CHRF our main metric because it has been shown to be more insightful and to correlate better with human evaluation than BLEU, for translation into MRLs (Popović, 2015; Bojar et al., 2016). In fact, being based on full word matches, BLEU is hardly suitable to evaluate highly agglutinative languages like Tamil. On the other hand, CHRF can capture partial word matches in the form of character n-grams.

	Model Name	100K parallel tok		2M parallel tok	
		BLEU	CHRF	BLEU	CHRF
ET	BASELINE	3.4	22.3	8.8	46.1
	RandMask15	3.6	24.0	8.9	46.4
	RandMask50	3.8	24.8	8.8	45.8
	StupidBT	4.1	25.2	8.9	46.2
	BT	**5.4**	**27.2**	**9.1**	46.6
	PT-Inflect	4.4	27.1	**9.1**	**47.5**
LT	BASELINE	3.0	27.8	11.2	47.6
	RandMask15	3.2	27.7	11.5	47.7
	RandMask50	3.0	27.1	11.7	48.0
	StupidBT	3.4	28.7	12.0	47.9
	BT	**4.3**	**32.1**	**12.6**	**49.2**
	PT-Inflect	3.7	29.9	12.3	48.8
TA	BASELINE	1.7	20.2	5.3	41.8
	RandMask15	2.6	24.8	5.9	43.1
	RandMask50	2.2	19.9	5.6	43.2
	StupidBT	2.5	24.3	5.9	43.1
	BT	**3.7**	**27.1**	**6.1**	**43.3**
	PT-Inflect	3.3	26.8	6.0	43.2
TR	BASELINE	2.4	19.9	10.9	44.9
	RandMask15	2.5	20.2	11.2	45.3
	RandMask50	2.6	20.7	11.2	45.4
	StupidBT	2.9	20.5	11.1	44.8
	BT	**3.8**	**26.6**	**12.6**	**49.3**
	PT-Inflect	3.4	21.0	11.5	46.6
DE	BASELINE	7.1	34.5	26.2	55.4
	RandMask15	7.7	38.1	26.1	55.0
	RandMask50	7.5	37.4	26.1	54.9
	StupidBT	7.9	40.2	26.3	55.7
	BT	**8.9**	42.9	**28.6**	**58.4**
	PT-Inflect	8.1	**43.0**	26.3	55.6

Table 6: Comparison of different pre-training objectives in the very low- (100K) and low- (2M) resource settings. All techniques are applied to a 1M-token monolingual dataset. The best performing technique, for each language pair and score, is highlighted in boldface.

A first, general observation on the BASELINE results shown in Table 5 is that BLEU scores are very low for this low-resource settings. This is in spite of implementing our NMT systems with the hyper-parameters suggested by Araabi and Monz (2020). In their paper, they were able to obtain a BLEU score of 11.3 for English→German, while our BASELINE scores 7.1 BLEU. We note, however, that these scores are not directly comparable as they evaluated their models on the IWSLT task (speech transcripts) and we on the WMT news task. When comparing performance across languages, we find that the BLEU and CHRF scores are best for English-German. This should not come as a surprise, as German is related to English and has the least complex morphology among our target languages.

[5] We always use back-translation in a pre-training regime, that is, our systems are first trained on the synthetic BT data, and then on the gold parallel data. In experiments not shown here, we also trained models on a mixture of synthetic BT and gold sentence pairs (i.e., joint training), but that worked similarly or worse than BT pre-training.

6.1. PT-Inflect

As shown in Table 5, our proposed inflection pre-training technique achieves consistent gains over the baseline in all languages and according to both metrics. Somewhat surprisingly, the largest CHRF gains are seen on German (+8.5), followed by Tamil (+6.4) and Esthonian (+4.8), whereas the smallest gains are found in Lithuanian (+2.1) and Turkish (+1.1). These results clearly indicate that morphological complexity is not the only factor at play. The accuracy of the off-the-self lemmatizer (cf. Table 3, last column) or that of the trained re-inflection model (cf. Table 1) may also play a role, as well as the similarity between monolingual data and test set.

Finally, we assess the optimal settings for PT-Inflect: (i) Regarding the type of training, we find that pre-training performs better than jointly training (rows 2 and 3 for each language). For all languages, the CHRF scores of the pre-training models are either on par or better than their pre-training counterparts. (ii) Regarding the effect of monolingual data size, we report gains in all five language pairs when we increase this from 100k to 1M tokens (rows 3 vs. 4).

Following these observations, we use pre-training on 1M monolingual tokens for all the remaining experiments.

6.2. Different pre-training objectives

The experiments so far show that PT-Inflect is an effective way to improve very low-resource translation into different target MRLs. We now compare PT-Inflect to the other pre-training objectives. Additionally we include the results on the larger 2M-token parallel dataset. The results are presented in Table 6.

Overall, back-translation (BT) remains the most effective way to use target monolingual data: considerable gains over the baseline are observed in all language pairs, in both the very low- and low-resource setups. The largest gain by BT over PT-Inflect is seen for Turkish, with a gain of +5.6 CHRF score in the very low-resource setup.

PT-Inflect appears as the second best technique, leading to results that are competitive with BT in several language pairs.

Next, we find that all lightweight pre-training techniques (different rates of random token masking and stupid BT) clearly underperform both BT and PT-Inflect in the large majority of settings. Under computational or time constraints, we recommend choosing StupidBT as a lightweight technique, as this performs on par with, or better than Random Masking in most cases. We find no clear winner between the two Random Masking rates, moreover the differences between the two are nearly offset in the low resource setting.

Finally, we note that moving from the very low- (100K parallel tokens) to the low-resource setting (2M) leads to much stronger baselines and smaller gains by all PT techniques, which was to be expected. Differences among PT techniques also become smaller, however the relative trends are similar as in the very low-resource setup.

Lang	Model	100K parallel tok		2M parallel tok	
		BLEU	CHRF	BLEU	CHRF
ET	PT-Inf	4.4	27.1	9.1	47.5
	BT	**5.4**	**27.2**	9.1	46.6
	Comb	5.3	27.0	**9.3**	**47.8**
LT	PT-Inf	3.7	29.9	12.3	48.8
	BT	4.3	32.1	12.6	49.2
	Comb	**4.5**	**33.2**	**13.0**	**49.9**
TA	PT-Inf	3.3	26.8	6.0	43.2
	BT	**3.7**	**27.1**	**6.1**	**43.3**
	Comb	3.5	26.3	5.9	43.1
TR	PT-Inf	3.4	21.0	11.5	46.6
	BT	**3.8**	**26.6**	**12.6**	**49.3**
	Comb	3.6	24.8	11.9	47.1
DE	PT-Inf	8.1	43.0	26.3	55.6
	BT	8.9	42.9	**28.6**	**58.4**
	Comb	**9.3**	**44.4**	27.9	57.1

Table 7: Comparison of COMBINE (Comb) with PT-Inflect (PT-Inf) and Back-Translation (BT). The best performing pre-training objective, for each language (Lang) pair and score, is highlighted in boldface.

6.3. Combining multiple objectives

We have seen that Back-Translation is the best pre-training objective for low-resource translations, with PT-Inflect being competitive in a few settings. We hence investigate whether a combination of the two techniques can lead to further improvements, as follows: A comparable amount of sentences (equivalent to 1M English tokens for back-translation and 1M inflected tokens for PT-Inflect) from both techniques are taken as the artificial data. A NMT model is first trained on the combined artificial data using the settings mentioned in Section 5.3. Once this model has converged, it is trained on the gold parallel data. The results of this technique, called COMBINE, are given in Table 7.

We find no clear winner between BT-only and COMBINE. More specifically, in the 100k-parallel setting, COMBINE is the best technique for Lithuanian and German, however it underperforms BT-only in Tamil and Turkish. Results in Estonian are almost the same. As for the 2M-parallel setting, differences are again less pronounced, and results rather inconclusive: BT wins in Tamil, Turkish and German, but COMBINE slightly outperforms it in Estonian and Lithuanian.

In summary, PT-Inflect appears to bring small complementary benefits over BT in some languages and settings, but it cannot be concluded that COMBINE is the best technique overall.

7. Conclusions

Building high-quality NMT systems for MRLs is challenging, especially in low-resource conditions (Sennrich and Zhang, 2019; Araabi and Monz, 2020). In this work, we have introduced a new pre-training technique, PT-Inflect, as a solution to overcome the scarcity of parallel data for MRLs. Rather than training a reverse-direction NMT system as needed for back-translation, PT-Inflect exploits linguistic tools, which are readily available in many languages, to generate pairs of lemmatized-original target sentences. This data is then used to pre-train the NMT model with the goal of improving its ability to generate complex target inflected forms when the source language is morphologically poor.

Through the course of our experiments, we conclusively find that PT-Inflect outperforms NMT systems trained only on parallel corpora, in both a very low- (100K parallel training tokens) and low-resource (2M) setting. Additionally, we found that pre-training is better than joint training, both in terms of performance and usability, and that increasing the monolingual data used by PT-Inflect leads to better NMT quality.

Back-translation still proved to be a better pre-training technique than PT-Inflect across the board. However, the combination of the two techniques brought further benefits in some of the evaluated languages, suggesting they affect complementary aspects of the translation model. In the future, we would like to experiment with truly low-resource and endangered languages for which lemmatizers already exist, such as Pashto, Occitan, or Telugu.

Acknowledgements

Arianna Bisazza was funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.021.646. We also would like to thank the Center for Information Technology of the University of Groningen for providing access to the Peregrine high performance computing cluster.

8. Bibliographical References

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual 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 3874–3884, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Araabi, A. and Monz, C. (2020). Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Armengol-Estapé, J. and Costa-jussa, M. (2021). Semantic and syntactic information for neural machine translation: Injecting features to the transformer. *Machine Translation*, 35, 04.

Ataman, D. and Federico, M. (2018). An evaluation of two vocabulary reduction methods for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–110, Boston, MA, March. Association for Machine Translation in the Americas.

Baziotis, C., Haddow, B., and Birch, A. (2020). Language model prior for low-resource neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online, November. Association for Computational Linguistics.

Bojar, O., Graham, Y., Kamran, A., and Stanojević, M. (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, August. Association for Computational Linguistics.

Burlot, F. and Yvon, F. (2018). Using monolingual data in neural machine translation: a systematic study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium, October. Association for Computational Linguistics.

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

Chakrabarty, A., Dabre, R., Ding, C., Utiyama, M., and Sumita, E. (2020). Improving low-resource NMT through relevance based linguistic features incorporation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4263–4274, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Cooper Stickland, A., Li, X., and Ghazvininejad, M. (2021). Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online, April. Association for Computational Linguistics.

Cotterell, R., Kirov, C., Hulden, M., and Eisner, J. (2019). On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342, March.

Currey, A., Miceli Barone, A. V., and Heafield, K. (2017). Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Copenhagen, Denmark, September.

Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dhar, P., Bisazza, A., and van Noord, G. (2020). Linguistically motivated subwords for English-Tamil translation: University of Groningen's submission to WMT-2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 126–133, Online, November. Association for Computational Linguistics.

Dhar, P., Bisazza, A., and van Noord, G. (2021). Optimal word segmentation for neural machine translation into Dravidian languages. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 181–190, Online, August. Association for Computational Linguistics.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, October-November. Association for Computational Linguistics.

Eriguchi, A., Tsuruoka, Y., and Cho, K. (2017). Learning to parse and translate improves neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 72–78, Vancouver, Canada, July. Association for Computational Linguistics.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April.

Kudo, T. and Richardson, J. (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, November. Association for Computational Linguistics.

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

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Li, Y., Li, X., Yang, Y., and Dong, R. (2020). A diverse data augmentation strategy for low-resource neural machine translation. *Information*, 11(5).

Liu, Z. and Prud'hommeaux, E. (2022). Data-driven model generalizability in crosslinguistic low-resource morphological segmentation.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Liu, Z., Winata, G. I., and Fung, P. (2021). Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online, August. Association for Computational Linguistics.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Popel, M., Tomková, M., Tomek, J., Łukasz Kaiser, Uszkoreit, J., Bojar, O., and Žabokrtský, Z. (2020). Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11.

Popović, M. (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, September. Association for Computational Linguistics.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Sagot, B. (2013). Comparing complexity measures. *Computational approaches to morphological complexity*, 02.

Saleva, J. and Lignos, C. (2021). The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 164–174, Online, April. Association for Computational Linguistics.

Sánchez-Cartagena, V. M., Esplà-Gomis, M., Pérez-Ortiz, J. A., and Sánchez-Martínez, F. (2021). Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8502–8516, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Sarveswaran, K., Dias, G., and Butt, M. (2019). Using meta-morph rules to develop morphological analysers: A case study concerning Tamil. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 76–86, Dresden, Germany, September. Association for Computational Linguistics.

Sennrich, R. and Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy, July. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). 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, August. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016b). 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, August. Association for Computational Linguistics.

Shavarani, H. S. and Sarkar, A. (2021). Better neural machine translation by extracting linguistic information from BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2772–2783, Online, April. Association for Computational Linguistics.

Shearing, S., Kirov, C., Khayrallah, H., and Yarowsky, D. (2018). Improving low resource machine translation using morphological glosses (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 132–139, Boston, MA, March. Association for Machine Translation in the Americas.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. *CoRR*, abs/1909.08053.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. (2019). Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Straka, M. (2018). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium, October. Association for Computational Linguistics.

Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, R., Tan, X., Luo, R., Qin, T., and Liu, T.-Y. (2021). A survey on low-resource neural machine translation. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4636–4643. International Joint Conferences on Artificial Intelligence Organization, 8. Survey Track.

Zhang, B., Williams, P., Titov, I., and Sennrich, R. (2020). Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online, July. Association for Computational Linguistics.

Zhou, Z., Levin, L. S., Mortensen, D. R., and Waibel, A. H. (2019). Using interlinear glosses as pivot in low-resource multilingual machine translation. *arXiv: Computation and Language*.

9. Language Resource References

Akhbardeh, Farhad and Arkhangorodsky, Arkady and Biesialska, Magdalena and Bojar, Ondřej and Chatterjee, Rajen and Chaudhary, Vishrav and Costa-jussa, Marta R. and España-Bonet, Cristina and Fan, Angela and Federmann, Christian and Freitag, Markus and Graham, Yvette and Grundkiewicz, Roman and Haddow, Barry and Harter, Leonie and Heafield, Kenneth and Homan, Christopher and Huck, Matthias and Amponsah-Kaakyire, Kwabena and Kasai, Jungo and Khashabi, Daniel and Knight, Kevin and Kocmi, Tom and Koehn, Philipp and Lourie, Nicholas and Monz, Christof and Morishita, Makoto and Nagata, Masaaki and Nagesh, Ajay and Nakazawa, Toshiaki and Negri, Matteo and Pal, Santanu and Tapo, Allahsera Auguste and Turchi, Marco and Vydrin, Valentin and Zampieri, Marcos. (2021). *Findings of the 2021 Conference on Ma-*

chine Translation (WMT21). Association for Computational Linguistics.

Barrault, Loïc and Bojar, Ondřej and Costa-jussà, Marta R. and Federmann, Christian and Fishel, Mark and Graham, Yvette and Haddow, Barry and Huck, Matthias and Koehn, Philipp and Malmasi, Shervin and Monz, Christof and Müller, Mathias and Pal, Santanu and Post, Matt and Zampieri, Marcos. (2019). *Findings of the 2019 Conference on Machine Translation (WMT19)*. Association for Computational Linguistics.

Barrault, Loïc and Biesialska, Magdalena and Bojar, Ondřej and Costa-jussà, Marta R. and Federmann, Christian and Graham, Yvette and Grundkiewicz, Roman and Haddow, Barry and Huck, Matthias and Joanis, Eric and Kocmi, Tom and Koehn, Philipp and Lo, Chi-kiu and Ljubešić, Nikola and Monz, Christof and Morishita, Makoto and Nagata, Masaaki and Nakazawa, Toshiaki and Pal, Santanu and Post, Matt and Zampieri, Marcos. (2020). *Findings of the 2020 Conference on Machine Translation (WMT20)*. Association for Computational Linguistics.

Bojar, Ondřej and Federmann, Christian and Fishel, Mark and Graham, Yvette and Haddow, Barry and Koehn, Philipp and Monz, Christof. (2018). *Findings of the 2018 Conference on Machine Translation (WMT18)*. Association for Computational Linguistics.

Kirov, Christo and Cotterell, Ryan and Sylak-Glassman, John and Walther, Géraldine and Vylomova, Ekaterina and Xia, Patrick and Faruqui, Manaal and Mielke, Sabrina J. and McCarthy, Arya and Kübler, Sandra and Yarowsky, David and Eisner, Jason and Hulden, Mans. (2018). *UniMorph 2.0: Universal Morphology*. European Language Resources Association (ELRA).

Koehn, Philipp and Chaudhary, Vishrav and El-Kishky, Ahmed and Goyal, Naman and Chen, Peng-Jen and Guzmán, Francisco. (2020). *Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment*. Association for Computational Linguistics.

Nakazawa, Toshiaki and Nakayama, Hideki and Ding, Chenchen and Dabre, Raj and Higashiyama, Shohei and Mino, Hideya and Goto, Isao and Pa Pa, Win and Kunchukuttan, Anoop and Parida, Shantipriya and Bojar, Ondřej and Kurohashi, Sadao. (2020). *Overview of the 7th Workshop on Asian Translation*. Association for Computational Linguistics.

Straka, Milan and Straková, Jana. (2017). *Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe*. Association for Computational Linguistics.