Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3650–3656
Marseille, 11–16 May 2020

Being Generous with Sub-Words towards Small NMT Children

Arne Defauw, Tom Vanallemeersch, Koen Van Winckel, Sara Szoc, Joachim Van den Bogaert
CrossLang
Kerkstraat 106, 9050 Gentbrugge, Belgium
{firstname.lastname}@crosslang.com

Abstract
In the context of under-resourced neural machine translation (NMT), transfer learning from an NMT model trained on a high resource language pair, or from a multilingual NMT (M-NMT) model, has been shown to boost performance to a large extent. In this paper, we focus on so-called *cold start* transfer learning from an M-NMT model, which means that the *parent* model is not trained on any of the *child* data. Such a set-up enables quick adaptation of M-NMT models to new languages. We investigate the effectiveness of cold start transfer learning from a *many-to-many* M-NMT model to an under-resourced child. We show that sufficiently large sub-word vocabularies should be used for transfer learning to be effective in such a scenario. When adopting relatively large sub-word vocabularies we observe increases in performance thanks to transfer learning from a parent M-NMT model, both when translating to and from the under-resourced language. Our proposed approach involving dynamic vocabularies is both practical and effective. We report results on two under-resourced language pairs, i.e. Icelandic-English and Irish-English.

Keywords: Neural Machine Translation, Multilingual Models, Transfer Learning

1. Introduction

In recent years, the performance of machine translation systems has been improving significantly thanks to the shift from statistical to neural machine translation (Bahdanau et al., 2014; Sutskever et al., 2014; Vaswani et al., 2017). However, for under-resourced language pairs, the performance of MT systems can still be disappointing, as pointed out for instance by Koehn et al. (2017).

To improve MT quality for under-resourced languages, a potential strategy consists of involving other language pairs during MT training than the one under scrutiny. In such a multilingual scenario, different language pairs, often belonging to closely related languages, are combined into a single translation model (Ha et al., 2016; Johnson et al., 2017). Recently, first efforts towards training massively multilingual (or 'universal') models have been undertaken (Aharoni et al., 2019; Arivazhagan et al., 2019), by extending systems towards dozens or even hundreds of language directions, which obviously introduces various bottlenecks.

An alternative, although related, strategy to improve NMT performance in an under-resourced scenario is transfer learning (Zoph et al., 2016). It is a concept that has been applied to various subfields of AI, such as computer vision and NLP. In NMT, transfer learning consists of training a (multilingual) *parent* MT model on one or more language pairs, in order to initialize a *child* MT model that continues training with data from a (different) set of (under-resourced) language pairs.

Transfer learning to an under-resourced language pair, either from a parent model trained on a high resource language pair (Zoph et al. 2016; Kocmi and Bojar, 2018; Lakew et al. 2018) or from a large multilingual parent model (Neubig and Hu, 2018), has been shown to increase translation performance.

While previous work on transfer learning has focused on a scenario in which the parent and child pairs share their target language (English), this paper investigates the effectiveness of transfer learning from a *many-to-many* multilingual parent model to a bilingual child model able to translate from an under-resourced language (Icelandic or Irish) into English and vice versa.

The focus of this work is on *cold start* transfer learning. This means that the parent model is not trained on any of the child data, which avoids the need of training a large multilingual model including all the language pairs at once. Such a set-up opens up various challenges. We show that, in order for transfer learning to be effective in such a scenario, sufficiently large sub-word vocabularies should be used.

The effect of sub-word vocabulary size in the context of (multilingual) transfer learning has only been scarcely investigated. Most related to our work are the papers by Nguyen and Chiang (2017) and Lakew et al. (2018). Nguyen and Chiang (2017) show that increasing sub-word vocabulary size increases translation performance when transferring between two under-resourced language pairs. From Lakew et al. (2018) we borrow the concept of *dynamic vocabulary* which means that the vocabulary of the parent model will be updated with the child vocabulary (see section 2.3.3).

Finally, we note that in related work (Nguyen and Chiang, 2017; Neubig and Hu, 2018; Lakew et al. 2018), under-resourced language pairs contain around 50-100k sentence pairs or less. In such an extreme low-resource scenario, improvements in translation quality are easier to obtain via transfer learning than in a mid-resource setting (Zoph et al. 2016; Kocmi and Bojar, 2018). In this paper, we focus on language pairs that can be considered low to mid-resourced (~1M sentence pairs) and present a simple, yet effective strategy for *cold start* multilingual transfer learning.

2. Data and Methods

2.1 Data

Our experiments involve 5 Germanic languages, i.e. 3 West-Germanic languages (English, German, Dutch) and two North-Germanic languages (Swedish and Icelandic).[1] We will refer to these five languages using the ISO-693-1 language codes EN, DE, NL, SV and IS. Another experiment involves Irish (Gaeilge, GA) instead of Icelandic; in this case, none of the other four languages is closely related to the under-resourced language, as Irish is a Celtic language.

[1] Another potential candidate for inclusion in the set of North-Germanic languages is Norwegian.

In Table 1 we give an overview of the parallel data used for training the NMT systems. We only made use of EN-XX parallel data, although for training our multilingual systems (see 2.3) we could have added data for other language pairs (e.g. DE-SV, NL-SV,...).

Language pair	Corpus[2]	#unique sent. pairs	#tokens (EN)
DE-EN	DGT	1,683k	40,215k
	DCEP	1,957k	49,511k
	Tatoeba	130k	961k
	Total	**3,725k**	**89,756k**
NL-EN	DGT	1,698k	40,201k
	DCEP	2,017k	50,310k
	Tatoeba	16k	115k
	Total	**3,684k**	**89,664k**
SV-EN	DGT	1,697k	40,503k
	DCEP	1,941k	46,955k
	Tatoeba	8k	51k
	Total	**3,601k**	**86,569k**
IS-EN	EUbookshop	8k	169k
	Tatoeba	7k	53k
	JW300	451k	8,940k
	TildeModel	363k	6,138k
	ELRC	53k	977k
	Total	**878k**	**16,237k**
GA-EN	DGT	39k	948k
	DCEP	7k	158k
	EUbookshop	96k	2,183k
	Irish legislation	172k	4,286k
	EU constitution	7k	140k
	Crawled data	131k	3,475k
	ParaCrawl data	785k	17,646k
	Total	**1,195k**	**27,861k**

Table 1: Overview of the parallel data used for training our multilingual and bilingual NMT systems.

For more details concerning the GA-EN parallel data, especially the web-crawled data and the processing of the GA-EN portion of the ParaCrawl Corpus, we refer to Defauw et al. (2019a,b).

We will report NMT performance on 3 language pairs in both translation directions: SV↔EN, IS↔EN and GA↔EN. For the SV-EN language pair we held out a test set of 3k unique sentence pairs from the training data. For IS-EN we held out 2k unique sentence pairs from the *EUbookshop, Tatoeba* and *ELRC* corpus. For GA-EN a test set of 3k unique sentence pairs was held out from the *DGT* and *DCEP* corpus.

[2] *DGT:* http://opus.nlpl.eu/DGT-v2019.php; *DCEP:* https://wt-public.emm4u.eu/Resources/DCEP-2013/DCEP-Download-Page.html; *Tatoeba:* http://opus.nlpl.eu/Tatoeba-v20190709.php; *EUbookshop*: http://opus.nlpl.eu/Eubookshop-v2.php; *JW300*: http://opus.nlpl.eu/JW300-v1.php; *TildeModel:* http://opus.nlpl.eu/TildeMODEL-v2018.php; *ELRC*: financial-economic corpora from www.elrc-share.eu; Irish legislation: www.gaois.ie/en; *EU constitution*: http://opus.nlpl.eu/Euconst.php; *Crawled data*: www.education.ie and www.courts.ie; *ParaCrawl data*: https://paracrawl.eu.

All occurrences of test sentences were removed from the training data: e.g. if an English sentence in the IS-EN test set was also found in the DE-EN training data, it was removed from the latter. This was done to prevent a bias in favor of the systems involving a multilingual parent engine.

2.2 MT Architecture

MT engines were trained with OpenNMT-tensorflow[3] using the Transformer architecture and default training settings. This configuration is the same as the 'base model' in the original paper on the Transformer architecture by Vaswani et al. (2017).

In terms of preprocessing, we tokenize the data and train a shared byte pair encoding (BPE) model (Sennrich et al. 2016) on the concatenation of the source and target data. The maximal sub-word vocabulary size of our NMT models is equal to the number of BPE merge operations. For more details we refer to section 2.3.3.

We report BLEU, TER and METEOR scores (the latter only for the XX→EN translation direction). Scores are obtained using the multeval library (Clark et al. 2011).

2.3 Experimental set-up

In this section we provide more details with regard to the followed methodology. As mentioned in the introduction, this paper investigates the impact of sub-word vocabulary size on the effectiveness of transfer learning from a multilingual NMT model in an under-resourced scenario. We first introduce the concept of multilingual NMT, followed by an overview of our multilingual transfer learning experiments and of our dealing with the sub-word vocabularies of our models.

2.3.1 Multilingual NMT

For building multilingual NMT models (M-NMT), we follow the strategy proposed by Johnson et al. (2017). This strategy consists of adding language codes to the source side of the training corpus, and has shown to be very effective, especially in the context of low-resource NMT. We prepend special tokens to the source sentence of a pair: for instance when a Swedish source is paired with an English target we prepend the tokens <*_src_sv*> and <*_tgt_en*> to the Swedish source sentence.

In this paper, we focus on so-called many-to-many M-NMT models that can translate to and from multiple languages. NMT models that can only translate to and from two languages (e.g. IS↔EN) will be referred to as bilingual NMT models (B-NMT).

For training our M-NMT systems we concatenated the XX→EN data with the EN→XX training data for the various languages. Data for language pairs not involving English were not included. This means that our M-NMT systems are able to translate to and from English, and, via zero shot translation, between other pairs of languages (e.g. SV↔DE).

2.3.2 Transfer Learning

Transfer learning across two NMT models involves a parent model, typically trained on a large amount of data, the encoder-decoder components of which are then transferred to initialize the parameters of a low-resourced child model. In this paper, we use a large M-NMT model

[3] https://github.com/OpenNMT/OpenNMT-tf

as a parent and an B-NMT system involving an under-resourced language combination as a child. This approach is similar to the one followed by Neubig and Hu (2018), although they only investigated M-NMT transfer learning in the context of *many-to-one* M-NMT models, i.e. multi-source models with a single target (EN).

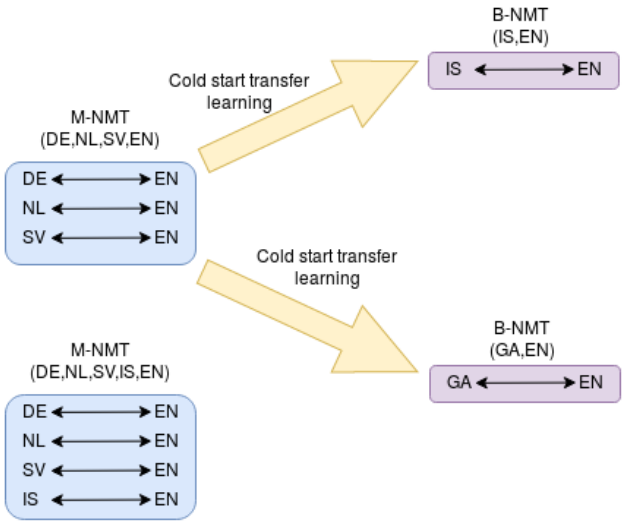

Figure 1: Overview of two M-NMT systems and two cold start transfer learning scenarios: M-NMT (DE,NL,SV,EN) ⟹ B-NMT (IS,EN) and M-NMT (DE,NL,SV,EN) ⟹ B-NMT (GA,EN).

As we focus on cold start transfer learning, the parent model of our transfer learning systems is not trained on the child data, in contrast to *warm start* transfer learning, in which the data of the under-resourced language pair of the child system is available during training of the parent system. Although it adds another layer of complexity, we chose to focus on a cold start scenario: from a practical point of view, it is more interesting to avoid the need of training a large multilingual model including all language pairs at once.

In Figure 1 we give an overview of the various transfer learning experiments we present in this paper. Our parent models will always be trained on DE, NL, SV and EN, and will be referred to as M-NMT(DE,NL,SV,EN). Our child systems will either be trained on IS and EN or on GA and EN, and will be referred to as M-NMT (DE,NL,SV,EN) ⟹ B-NMT(IS,EN) and M-NMT (DE,NL,SV,EN) ⟹ B-NMT(GA,EN). M-NMT systems trained on DE, NL, SV, IS and EN will be used as a benchmark for the systems involving IS.

2.3.3 Dynamic Vocabulary

For our transfer learning experiments, we used a dynamic vocabulary, in line with Lakew et al. (2018). This approach avoids the need to have all training data, including the training set that will be used for continued training, available when creating the parent model, and is thus suited for a cold start scenario.

In order to create a dynamic vocabulary, the sub-word (BPE) vocabulary of the parent NMT system (V_p) is updated with the sub-word vocabulary of the child system (V_c). This leads the entries V_c that are also present in V_p ($V_p \cap V_c$) to inherit the embedding space of the parent NMT model, while the entries in V_c that were not present in V_p ($V_c \backslash V_p$) will be initialized randomly in the child NMT model. Entries in V_p absent in V_c ($V_p \backslash V_c$) are discarded. This corresponds to the progAdapt methodology proposed by Lakew et al. (2018).

As mentioned in 2.2, our BPE models are trained on the concatenation of both source and target data. In the context of M-NMT and B-NMT (see 2.3.1), this implies that we trained our BPE models on the concatenation of all training data: e.g. for our M-NMT model involving DE, NL, SV and EN, we trained the BPE model on the concatenation of the DE-EN, NL-EN and SV-EN training data. For each NMT model (either parent or child) a new BPE model is trained, and for each NMT model one sub-word (BPE) vocabulary is extracted from the training data, with a maximal size equal to the number of BPE merge operations. We refer to Figure 2, where we schematically explain the relation between the BPE models and extracted vocabularies for the cold start transfer learning scenario M-NMT (DE,NL,SV,EN) ⟹ B-NMT (IS,EN).

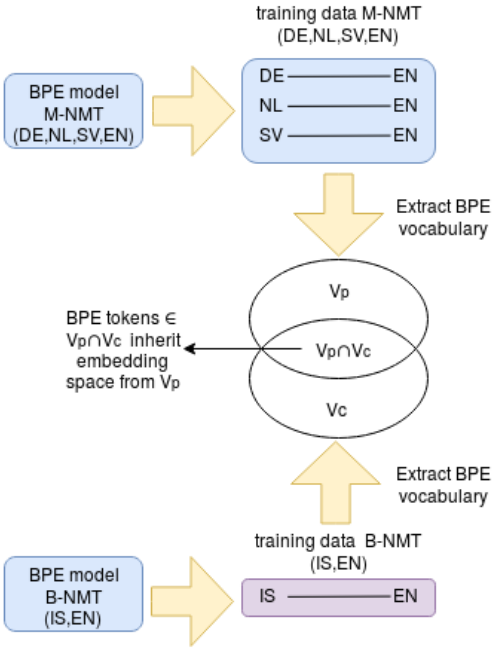

Figure 2: Overview of the relation between the BPE models and vocabularies for the transfer learning scenario M-NMT (DE,NL,SV,EN) ⟹ B-NMT (IS,EN). Maximal size of the sub-word vocabularies is equal to the number of BPE merge operations of the underlying BPE model.

We note that it is also possible to train a BPE model or SentencePiece model (Kudo and Richardson 2018) on the different languages separately, as in Neubig and Hu (2018). However, we chose to use a shared vocabulary created through BPE as this was shown to improve the alignment of embedding spaces across languages that share the same alphabet (Smith et al. 2017; Lample et al. 2018).

In our experiments we varied the sizes of the sub-word vocabularies, both of the parent M-NMT and the child B-NMT models. This is achieved by varying the number of merge operations of the BPE model used for extracting the sub-word vocabulary. The maximal sub-word

vocabulary size is taken equal to the number of BPE merge operations. Numbers of BPE merge operations considered are 10k, 32k and 64k. In the following we will, for convenience, refer to sub-word vocabulary sizes as 10k, 32k and 64k, while in reality they can be slightly smaller than these numbers.

Previous publications on transfer learning applied vocabulary sizes of 8k (Lakew et al. 2018; Neubig and Hu, 2018)[4] and 15k-32k (Zoph et al. 2016; Kocmi and Bojar, 2018).

3. Results

As mentioned in section 2.3.2, we trained parent M-NMT models on the four Germanic languages DE, NL, SV and EN. In a first set of experiments, we varied the sub-word vocabulary sizes of the M-NMT(DE,NL,SV,EN) models, and evaluated translation performance EN↔SV. We observe an increase in performance when using a larger sub-word vocabulary (Figure 3). This is in line with previous work on this subject (Arivazhagan et al. 2019), where it was shown that M-NMT models with smaller vocabulary perform noticeably worse on high-resource languages.

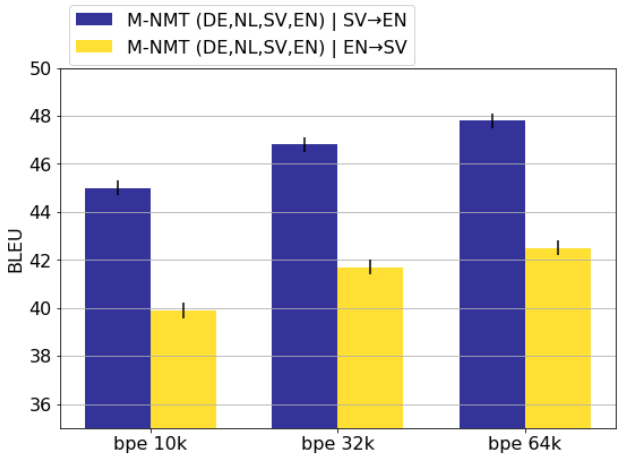

Figure 3: BLEU scores EN↔SV for the M-NMT(DE,NL,SV,EN) models when varying vocabulary sizes.

However, it is unclear how this increase in translation performance of a parent M-NMT model thanks to a larger vocabulary can be transferred to an under-resourced child language pair. For instance, it has been argued that using a smaller sub-word vocabulary (and thus shorter sub-word BPE tokens) leads to better performance due to improved generalization for under-resourced languages (Cherry et al. 2018; Kreutzer and Sokolov, 2018).

To answer the above question we performed a variety of experiments on two under-resourced language pairs: IS-EN and GA-EN. We trained B-NMT models for IS-EN and GA-EN with a sub-word vocabulary size of 10k, 32k and 64k. Next, we performed transfer learning from the parent model M-NMT(DE,NL,SV,EN) trained with varying vocabulary size (10k, 32k, 64k) to B-NMT(IS,EN) and B-NMT(GA,EN) child models with the same vocabulary size using a dynamic vocabulary (see 2.3.3 for more details). We did not use larger vocabulary sizes than 64k because these may lead to a long training and decoding time.

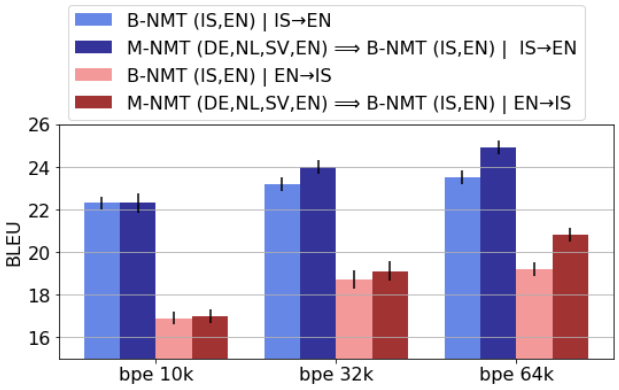

Figure 4: BLEU scores IS↔ EN. Results for the IS→EN translation direction are shown in blue and those for EN→IS in red.

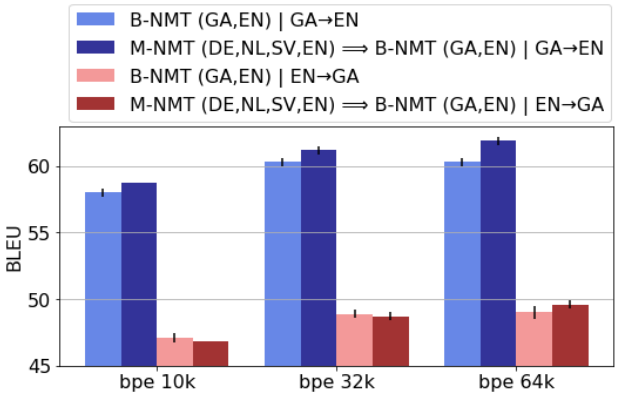

Figure 5: BLEU scores GA↔EN.

We refer to Figure 4 for the results on the IS-EN language pair. We observe that when no transfer learning is used, the performance of our B-NMT model improves when increasing the sub-word vocabulary sizes (light blue and light red colors). However, more interestingly, the BLEU scores of the B-NMT models obtained after transfer learning (i.e. the M-NMT (DE,NL,SV,EN) ⟹ B-NMT (IS,EN) models, dark blue and red colors) largely improve when increasing the sub-word vocabulary sizes of the M-NMT and B-NMT models. In order for multilingual transfer learning to be effective, a parent and child M-NMT model with a sufficient large vocabulary are needed. Especially in the EN→IS translation direction the difference in BLEU/METEOR score between the transfer learning systems using a large and small vocabulary is clear (see Table 2). We also compare our results to the performance of an M-NMT system including IS, i.e. M-NMT (DE,NL,SV,IS,EN), (see last two rows of Table 2). We observe that our transfer learning systems perform better than this system, although increasing vocabulary sizes in the latter also leads to an increase in BLEU and METEOR score, in line with the results on EN↔SV (Figure 3).

[4] Neubig and Hu (2018) trained SentencePiece models on each language separately and used a sub-word vocabulary of 8k for each language.

Model type	Voc size	Metric	IS→EN	EN→IS
B-NMT	10k	BLEU	22.3 (0.5/0.1)	16.9 (0.5/0.1)
		METEOR	24.3 (0.3/0.0)	/
		TER	65.5 (0.8/0.3)	70.6 (0.6/0.2)
B-NMT	32k	BLEU	23.2 (0.6/0.1)	18.7 (0.5/0.2)
		METEOR	24.7 (0.3/0.1)	/
		TER	65.1 (0.8/0.4)	69.6 (0.7/0.1)
B-NMT	64k	BLEU	23.5 (0.6/0.1)	19.2 (0.6/0.1)
		METEOR	24.6 (0.3/0.0)	/
		TER	64.7 (0.7/0.1)	69.6 (0.7/0.2)
M-NMT ⟹ B-NMT	10k ⟹ 10k	BLEU	22.3 (0.5/0.2)	17.0 (0.5/0.1)
		METEOR	24.4 (0.3/0.1)	/
		TER	64.9 (0.6/0.2)	70.5 (0.6/0.3)
M-NMT ⟹ B-NMT	32k ⟹ 32k	BLEU	24.0 (0.6/0.1)	19.1 (0.5/0.2)
		METEOR	25.4 (0.3/0.0)	/
		TER	63.2 (0.7/0.1)	68.8 (0.7/0.2)
M-NMT ⟹ B-NMT	64k ⟹ 64k	BLEU	**24.9 (0.6/0.1)**	**20.8 (0.6/0.1)**
		METEOR	**25.8 (0.3/0.1)**	/
		TER	**63.0 (0.7/0.1)**	**67.7 (0.7/0.3)**
M-NMT +IS	10k	BLEU	21.1 (0.5/0.1)	14.7 (0.5/0.1)
		METEOR	23.7 (0.3/0.1)	/
		TER	66.6 (0.6/0.2)	72.9 (0.7/0.1)
M-NMT +IS	32k	BLEU	22.7 (0.6/0.1)	16.8 (0.5/0.1)
		METEOR	24.7 (0.3/0.1)	/
		TER	65.1 (0.7/0.1)	70.7 (0.6/0.2)
M-NMT +IS	64k	BLEU	**24.9 (0.6/0.2)**	20.3 (0.6/0.2)
		METEOR	25.1 (0.3/0.1)	/
		TER	64.1 (0.7/0.3)	68.4 (0.7/0.2)

Table 2: Results IS↔EN. Model types: baseline (B-NMT), transfer (M-NMT ⟹ B-NMT), multilingual including IS (M-NMT+IS). In brackets: variance due to test set selection and optimizer instability (last 5 epochs).

We performed a similar set of experiments on the GA-EN language pair, see Figure 5 and Table 3. We observed similar results as for IS-EN: increasing the sub-word vocabulary sizes of both parent and child results in increases in translation quality that are not observed on the baseline B-NMT models. Although results on EN→GA are similar, the effect of transfer learning is smaller than for the other translation directions reported. This may be caused by the larger linguistic distance between GA and the languages of the parent M-NMT model, which is trained on SV, DE, NL and EN.

The results reported above demonstrate that increased performance of a parent M-NMT model caused by increased sub-word vocabulary size (Figure 3) can be successfully transferred to a child B-NMT system. To obtain more insight into this mechanism we calculated the overlap in sub-word vocabulary between parent and child models for varying vocabulary sizes. When the vocabulary overlap is high, we expect a lot of information reuse between the parent and the child (see Figure 6). Both for IS-EN and GA-EN we see that the absolute number of overlapping sub-word tokens increases and the relative amount decreases as vocabulary size grows: for a size of 10k more than one third of the tokens overlap between parent and child, for 64k only a quarter.

Model type	Voc size	Metric	GA→EN	EN→GA
B-NMT	10k	BLEU	58.0 (0.5/0.1)	47.1 (0.5/0.1)
		METEOR	44.3 (0.3/0.0)	/
		TER	31.7 (0.5/0.1)	39.8 (0.5/0.2)
B-NMT	32k	BLEU	60.3 (0.5/0.1)	48.9 (0.5/0.1)
		METEOR	45.5 (0.3/0.1)	/
		TER	29.6 (0.5/0.1)	38.4 (0.5/0.0)
B-NMT	64k	BLEU	60.3 (0.5/0.1)	49.0 (0.5/0.2)
		METEOR	45.5 (0.3/0.0)	/
		TER	29.6 (0.5/0.0)	38.6 (0.5/0.2)
M-NMT ⟹ B-NMT	10k ⟹ 10k	BLEU	58.7 (0.5/0.0)	46.8 (0.5/0.0)
		METEOR	44.7 (0.3/0.0)	/
		TER	31.1 (0.5/0.0)	40.0 (0.5/0.0)
M-NMT ⟹ B-NMT	32k ⟹ 32k	BLEU	61.2 (0.5/0.1)	48.7 (0.5/0.1/)
		METEOR	45.9 (0.3/0.0)	/
		TER	29.1 (0.5/0.1)	38.5 (0.5/0.1)
M-NMT ⟹ B-NMT	64k ⟹ 64k	BLEU	**61.9 (0.5/0.1)**	**49.6 (0.5/0.1)**
		METEOR	**46.3 (0.3/0.0)**	/
		TER	**28.3 (0.5/0.1)**	**37.9 (0.5/0.1)**

Table 3: Results GA↔EN.

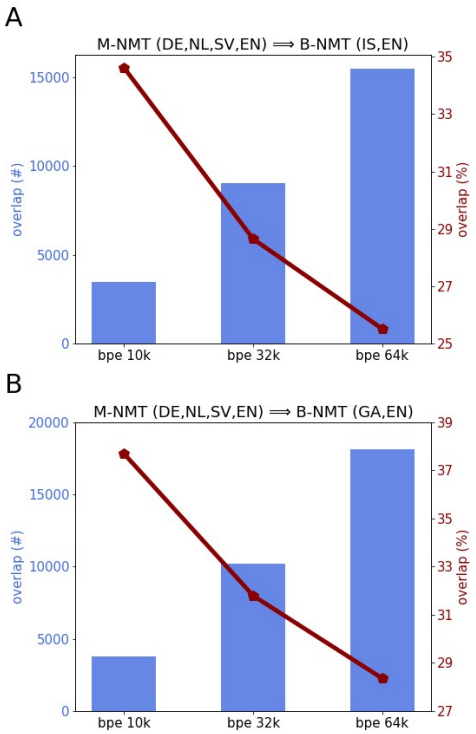

Figure 6: Vocabulary overlap between parent and child NMT models for varying vocabulary sizes. Absolute numbers are shown in blue (left *y*-axis), and relative numbers in red (right *y*-axis).

In terms of translation performance, the relative decrease of overlapping sub-word tokens between parent and child is compensated by the absolute increase of overlapping tokens and the increased performance of the parent model on a related language pair (SV↔EN). It would be interesting to further investigate if a minimum overlap percentage is needed for multilingual transfer learning to

be effective, and how this is related to the size of the parent and child's sub-word vocabulary.

4. Conclusion and Future Work

In this paper, we have investigated the effect of sub-word vocabulary size in the context of cold start transfer learning from a many-to-many M-NMT model to an under-resourced language pair. We showed that cold start transfer learning from a parent M-NMT to an under-resourced child model only results in increased translation performance of the child when a sufficiently large sub-word vocabulary is used.

Our proposed multilingual cold start transfer learning approach using dynamic vocabularies is both practical, as it only requires training one 'large' M-NMT model, and effective, resulting in increased performance on under-resourced language pairs.

In future work we want to investigate whether it is possible to follow the same approach when transferring from a many-to-many M-NMT model trained on a larger amount of language pairs. For instance, when more languages belonging to different language families are involved, re-grouping languages before training BPE or SentencePiece models may have a positive effect. Also, the effect of sampling temperature (Arivazhagan et al. 2019) to prevent BPE or SentencePiece models to be overwhelmed by certain languages (e.g. English) should be investigated. Regrouping languages before generating sub-word vocabularies and tuning the sampling temperature to its optimal value may both have a positive effect on vocabulary overlap between parent and child, resulting in more information reuse.

Finally, it should be investigated whether increasing the number of parameters of the parent M-NMT model (apart from vocabulary size) leads to increased performance on the downstream task in a cold start transfer learning scenario, in a similar way as observed in 'universal' M-NMT (Aharoni et al. 2019; Arivazhagan et al. 2019).

Acknowledgements

This work was performed in the framework of the SMART 2015/1091 project ("Tools and resources for CEF automated translation"), funded by the CEF Telecom programme (Connecting Europe Facility).

5. Bibliographical References

Aharoni, R., Johnson, M., and Firat O. (2019). Massively Multilingual Neural Machine Translation. *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)* (NAACL'19), pages 3874–3884, Minneapolis, Minnesota, June. Association for Computational Linguistics (ACL).

Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M.X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., and Wu., Y (2019). Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. *arXiv preprint arXiv:1907.05019*.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR, abs/1409.0473*.

Cherry, C., Foster, G., Bapna, A., Firat, O., and Macherey, W. (2018). Revisiting Character-based Neural Machine Translation with Capacity and Compression. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (EMNLP'18), pages 4295–4305, Brussels, Belgium, November. Association for Computational Linguistics (ACL).

Clark, J., Dyer, C., Lavie, A., and Smith, N. (2011). Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. *Proceedings of the Association for Computational Linguistics* (ACL), pages 176–181.

Defauw, A., Vanallemeersch, T., Szoc, S., Everaert, F., Van Winckel, K., Scholte, K., Brabers, J., and Van den Bogaert, J. (2019). Collecting Domain Specific Data for MT: an Evaluation of the ParaCrawl Pipeline. *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 186–195, Dublin, Ireland.

Defauw, A., Szoc, S., Vanallemeersch, T., Bardadym, A., Brabers, J., Everaert, F., Scholte, K., Van Winckel, K., and Van den Bogaert, J. (2019). Developing a Neural Machine Translation System for Irish. *Proceedings of the 2nd Workshop on Technologies for MT for Low Resource Languages*, pages 32–38, Dublin, Ireland.

Ha, T., Niehues, J., and Waibel A. (2016). Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. *CoRR, abs/1611.04798*.

Johnson, M., Schuster, M., V Le, Q., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenber, M., Corrado, G., Hughes, M., and Dean J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kocmi, T. and Bojar, O. (2018). Trivial Transfer Learning for Low-Resource Neural Machine Translation. *Proceedings of the Third Conference on Machine Translation, Volume 1: Research papers* (WMT'18), pages 224–252, Brussels, Belgium, November. Association for Computational Linguistics (ACL).

Koehn, P. and Knowles, R. (2017). Six Challenges for Neural Machine Translation. *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada, August. Association for Computational Linguistics (ACL).

Kreutzer, J. and Sokolov, A. (2018). Learning to Segment Inputs for NMT Favors Character-level Processing. *CoRR, abs/1810.01480.*

Kudo, T. and Richardson, J. (2018). SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations)* (EMNLP'18), pages 66–71, Brussels, Belgium, November. Association for Computational Linguistics (ACL).

Lakew, S.M., Erofeeva, A., Negri, M., Federico, M., and Turchi, M. (2018). Transfer Learning in Multilingual Neural Machine Translation with Dynamic Vocabulary. *Proceedings of the International Workshop on Spoken Language Translation* (IWSLT'18), pages 54–61.

Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Unsupervised Machine Translation Using

Monolingual Corpora Only. *Proceedings of the International Conference on Learning Representations* (ICLR'18).

Neubig, G. and Hu, J. (2018). Rapid Adaptation of Neural Machine Translation to New Languages. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (EMNLP'18), pages 875–880, Brussels, Belgium, November. Association for Computational Linguistics (ACL).

Nguyen, T.Q. and Chiang, D. (2017). Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 296–301, Taipei, Taiwan, November. (AFNLP).

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. *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 (ACL).

Smith, S.L., Turban, D.H.P., Hamblin, S., and Hammerla, N.Y. (2017). Offline Bilingual Word Vectors, Orthogonal Transformations and the Inverted Softmax. *International Conference on Learning Representations.*

Sutskever, I., Vinyals, O., and V Le, Q. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in neural information processing systems*, 27:3104–3112.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, L.J., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* 30:6000–6010.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer Learning for Low-Resource Neural Machine Translation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (EMNLP'16), pages 1568–1575, Austin, Texas, November. Association for Computational Linguistics (ACL).