Subword Segmentation and a Single Bridge Language Affect Zero-Shot Neural Machine Translation

Link:
[Link to publication record in Edinburgh Research Explorer](#)

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proceedings of the Fifth Conference on Machine Translation

OPEN ACCESS

Subword Segmentation and a Single Bridge Language Affect Zero-Shot Neural Machine Translation

Annette Rios[1] and **Mathias Müller**[1] and **Rico Sennrich**[1,2]
[1]Department of Computational Linguistics, University of Zurich
[2]School of Informatics, University of Edinburgh

Abstract

Zero-shot neural machine translation is an attractive goal because of the high cost of obtaining data and building translation systems for new translation directions. However, previous papers have reported mixed success in zero-shot translation. It is hard to predict in which settings it will be effective, and what limits performance compared to a fully supervised system. In this paper, we investigate zero-shot performance of a multilingual EN↔{FR,CS,DE,FI} system trained on WMT data. We find that zero-shot performance is highly unstable and can vary by more than 6 BLEU between training runs, making it difficult to reliably track improvements. We observe a bias towards copying the source in zero-shot translation, and investigate how the choice of subword segmentation affects this bias. We find that language-specific subword segmentation results in less subword copying at training time, and leads to better zero-shot performance compared to jointly trained segmentation. A recent trend in multilingual models is to not train on parallel data between all language pairs, but have a single *bridge* language, e.g. English. We find that this negatively affects zero-shot translation and leads to a failure mode where the model ignores the language tag and instead produces English output in zero-shot directions. We show that this bias towards English can be effectively reduced with even a small amount of parallel data in some of the non-English pairs.

1 Introduction

Zero-shot translation has first been introduced by Firat et al. (2016) and refers to the ability of a multilingual NMT model to translate between all its source and target languages, even those pairs for which no parallel data was seen in training. In the simplest setting, all parameters in the network are shared between the different languages and the translation is guided only by special tags to indicate the desired output language (Johnson et al., 2017; Ha et al., 2016). While this capability is attractive because it is an alternative to building N^2 dedicated translation systems to serve N languages, performance on zero-shot pairs tends to lag behind pivot translation. Recent papers, such as Arivazhagan et al. (2019), Gu et al. (2019) and Zhang et al. (2020), have suggested training techniques to improve the generalization to unseen language pairs, but performance varies considerably across settings.

In this paper, we examine in detail the behavior of the multilingual model proposed by Johnson et al. (2017) on zero-shot translation directions. Our experiments show the following:

- Translation quality for zero-shot language pairs is highly unstable between different training runs, and between training checkpoints, which calls for more rigour to avoid false positive results.

- The incorrect copying of source text into the output is affected by the extent of subword copying at training time, and can be reduced by performing language-specific subword segmentation.

- English-centric models have a tendency to produce English text for non-English input. Multi-bridge models that include data from non-English pairs mitigate this problem.

Overall, we observe improvements of 8.1 BLEU (15.9→24.0) on 6 zero-shot directions with simple changes to the multilingual training setup.

2 Related Work

Our experiments are based on the multilingual model proposed by (Johnson et al., 2017; Ha et al.,

528

2016): A single model is trained on multiple language pairs with a standard encoder-decoder architecture, all parameters in the network are shared for all languages, including the vocabulary. An artificial target language token determines the output language. We prefix this special token to the source sentence as in Johnson et al. (2017). The major advantage of this model lies in its simplicity, since it does not require changing the architecture or training objective.

Several recent studies have explored approaches to improve generalization to zero-shot language pairs, for example through semi-supervised training (Gu et al., 2019; Currey and Heafield, 2019; Zhang et al., 2020) or alignment of encoder representations (Arivazhagan et al., 2019).

Our study is concerned with data conditions that enable zero-shot generalization for multilingual NMT, specifically preprocessing and data settings. While initial work used separate encoders and decoders for different languages (Firat et al., 2016), sharing of encoder and decoder parameters was established by Johnson et al. (2017); Ha et al. (2016) and has since been widely adopted. Johnson et al. (2017) use a shared subword segmentation model across languages, and this strategy is followed by later work (e.g. Aharoni et al., 2019; Zhang et al., 2020). Ha et al. (2016) do not share embeddings across languages, but use language-specific codes. We will show that both strategies cause errors.

In terms of data settings, the number of languages involved in multilingual models has increased from 3–4 (Firat et al., 2016; Johnson et al., 2017) to over 100 (Aharoni et al., 2019). The most popular setup are English-centric datasets, where the model is trained on translations between English and a number of other languages. A multi-way parallel corpus between 5 languages has been provided for the IWSLT17 multilingual task (Cettolo et al., 2012). Results on this dataset show strong zero-shot generalization, close or even exceeding the supervised condition (Lakew et al., 2017), but multi-way parallel corpora are only available in small amounts and specific domains, so we investigate alternatives to English-centric models that do not rely on multi-way parallelism.

3 Data and Models

Following Aharoni et al. (2019), our baseline setup is English-centric. For training, we use 5 million parallel sentences per language pair for
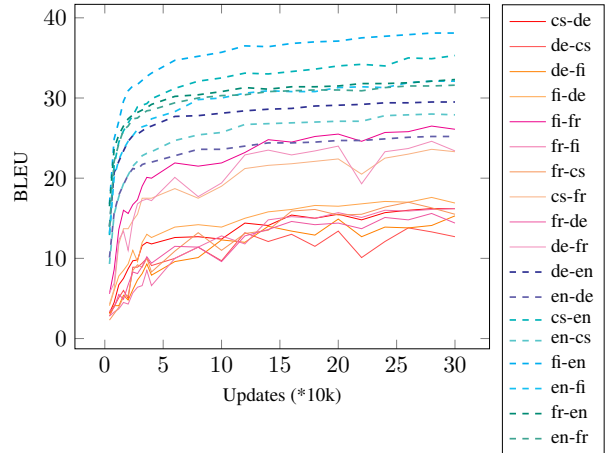

Figure 1: Baseline BLEU scores on test set as a function of training time. Dashed lines: trained pairs; solid lines: zero-shot pairs.

English↔{French,Czech,German,Finnish} from WMT (Barrault et al., 2019). For all zero-shot language pairs, we sample test sets from OPUS (Tiedemann, 2012), see Table 1 for details.

To indicate the target language, we prefix a language tag on the source side (e.g. <2en>). Following Johnson et al. (2017), we segment all data with a byte-pair encoding model trained jointly on the training data in all five languages (Sennrich et al., 2016), with a threshold of 32k BPE operations. All our systems are base Transformers (Vaswani et al., 2017) implemented in Sockeye (Hieber et al., 2018), trained with early stopping based on BLEU on a development set that consists in equal parts of parallel sentences from all trained translation directions. See Appendix A and B for training details.

4 Baseline Experiments

BLEU[1] on zero-shot pairs is relatively unstable, see Fig. 1: while BLEU on the trained pairs increases steadily during training (dashed lines), performance on unseen language pairs fluctuates considerably, as also observed by Aharoni et al. (2019). Furthermore, multiple training runs result in relatively large differences in BLEU on the zero-shot directions. Across three training runs, average BLEU varies up to 0.24 points on trained language pairs (standard deviation: 0.12), but up to 6.28 BLEU on zero-shot pairs (standard deviation: 3.14) – see Table 2 for full results. We suspect that this fluctuation is due to the fact that the model is not op-

---

[1]SacreBLEU (Post, 2018): `BLEU+c.mixed+#.1 +s.exp+t.13a+v.1.2.21.`

corpora		training	dev	test
Language Pairs with English:				
de↔en	Commoncrawl, Europarl-v9, Wikititles-v1	5M	250	2000
cs↔en	Europarl-v9, CzEng1.7	5M	250	2000
fr↔en	Commoncrawl, Europarl-v7	5M	250	2000
fi↔en	Europarl-v9, Wikititles-v1, Paracrawl-v3	4.35M*	250	2000
Multi-Bridge Pairs:				
fr↔fi	Rapid2016	350k	200	2000
cs↔de	Rapid2016, NewsCommentary, GlobalVoices	343k**	200	2000
Zero-shot test sets:				
de↔fi	Rapid2016			2000
de↔fr	Rapid2016, NewsCommentary, GlobalVoices			2000
cs↔fr	Rapid2016, NewsCommentary, GlobalVoices			2000

* oversampled to 5M
** oversampled to 350k

Table 1: Parallel corpora for training and testing. Sampled development sets are combined for training to a total of 2000 sentences (baselines) or 2800 sentences (training with cs↔de and fi↔fr). Europarl-v7, NewsCommentary and GlobalVoices retrieved from OPUS (Tiedemann, 2012), all other corpora are part of the WMT19 translation shared task (Barrault et al., 2019)

timized on zero-shot directions: models converge to different local minima that may be similarly good for trained pairs, but with no mechanism that stabilizes generalization to zero-shot pairs. If not stated otherwise, we will report the mean and standard deviation of three training runs with different seeds throughout the paper.

As an alternative to zero-shot translation, we report results obtained via pivot translation through English (e.g. German-English-Czech). On our data set, this approach works better than zero-shot translation. Pivot translation is stable across training runs, with a standard deviation of 0.19.

5 Copy Bias and Language-Specific Subword Segmentation

One failure mode we observe in zero-shot translation is over-copying of the input.[2] We suspect that for the translation of zero-shot directions, the model relies heavily on (sub-) words in the vocabulary that are shared between languages. To test this hypothesis, we train two models with language-specific subword segmentation:

a) a model with language-specific subword seg-

___
[2]See also (Ha et al., 2017; Arivazhagan et al., 2019; Zhang et al., 2020), who make similar observations in different settings.

Trained Directions			
	sampled test	official wmt test sets	
de-en	29.6 ±0.12	2019	31.9 ±0.35
cs-en	35.0 ±0.30	2018	25.7 ±0.20
fi-en	38.2 ±0.06	2019	25.2 ±0.20
fr-en	32.1 ±0.21	2015	33.9 ±0.56
en-de	25.2 ±0.21	2019	30.7 ±0.25
en-cs	28.4 ±0.45	2019	18.0 ±0.17
en-fi	32.1 ±0.20	2018	12.7 ±0.26
en-fr	31.7 ±0.21	2015	32.5 ±0.40
average	31.6 ±0.12		

Zero-Shot Directions		
	direct	pivot
cs-de	14.7 ±1.39	20.3 ±0.32
de-cs	8.9 ±5.14	20.1 ±0.44
cs-fr	22.0 ±2.71	28.3 ±0.31
fr-cs	11.5 ±5.89	22.1 ±0.31
de-fr	23.3 ±2.48	29.0 ±0.15
fr-de	12.0 ±3.01	21.6 ±0.06
fi-fr	23.5 ±4.12	30.4 ±0.06
fr-fi	12.2 ±4.61	20.7 ±0.26
fi-de	15.1 ±1.69	21.3 ±0.15
de-fi	11.2 ±4.53	20.0 ±0.38
average	15.4 ±3.14	23.4 ±0.19

Table 2: Baseline (BLEU). Average and standard deviation of 3 training runs reported. For zero-shot directions, we compare direct zero-shot translation and pivot translation via English.

530

mentation and no vocabulary overlap. We
limit BPE operations to 10k per language.

b) similar to model a), with the exact same sub-
word segmentation, but with vocabulary over-
lap.

For model a), we remove any potential vocabu-
lary overlap by adding a language identifier to each
subword. For instance, consider the preposition *in*
in German and English: instead of one token *in*,
the network vocabulary has an entry for *in#de#* and
an additional entry for *in#en#*. This corresponds
to the language-specific coding introduced by Ha
et al. (2016).

For model b), we split words with the same
language-specific BPE models as for a), but we
allow vocabulary overlap, i.e. homographic forms
in different languages are represented by a single
entry in the network's vocabulary. This results in
a vocabulary size of ~50k for model a), whereas
for model b), the vocabulary amounts to a total of
~36k subwords.

Table 3 shows that removing vocabulary overlap
does not affect the trained language pairs greatly,
however, the effect on the zero-shot directions is
quite harsh: For the first evaluation of model a),
we remove only the correct target language tag
(i.e. homographic forms with wrong language tag
count as wrong), while for the second evaluation,
we remove all language tags from the translations
(i.e. homographic forms in other languages count
as correct). In the first case, the model averages at
only 4.7 BLEU on zero-shot directions, however,
the more lenient second evaluation results in better
scores (12.7 BLEU). This difference is due to the
fact that the no-overlap model tends to produce a lot
of English subwords (marked by #en#), especially
for proper names and numbers.[3]

The second evaluation improves BLEU because
the no-overlap model will often output the correct
form, e.g. for proper names, if the word in the target
language has the same spelling as in English.

Model b), with language-specific BPE and over-
lapping vocabularies, represents a compromise
between a fully shared representation and fully
language-specific coding. We hypothesize that al-
lowing some vocabulary overlap helps aligning the
representation between sentences with the same

---

[3]This essentially means that the strict evaluation gives us a
more realistic estimate of the translation quality we can expect
if the source and target language do not happen to share word
forms, e.g. languages in different scripts.

	trained	zero-shot
jointly trained BPE	31.6 ±0.12	15.4 ±3.14
language-specific BPE:		
a) no overlap, strict*	30.9 ±0.58	4.7 ±1.90
a) no overlap, lenient**	31.3 ±0.59	12.7 ±2.52
b) vocabulary overlap	31.2 ±0.60	20.5 ±0.43

* homographic words in other languages=wrong
** homographic words in other languages=correct

Table 3: Average BLEU for models with language spe-
cific BPE, with and without vocabulary overlap.

	BPE	subwords	words
training set	jointly trained	9.70%	*5.70%
	lang.-specific	7.96%	*5.70%
translations	jointly trained	24.82%	20.58%
	lang.-specific	6.97%	4.70%

* identical

Table 4: Average word and subword overlap between
source and target in training set, and in zero-shot trans-
lation output with jointly trained and language-specific
BPE.

meaning in different languages, which is also sup-
ported by the effectiveness of cross-lingual pre-
training with shared vocabularies for unsupervised
MT and cross-lingual transfer (Conneau and Lam-
ple, 2019). We observe that models with jointly
trained BPE develop a strong bias towards copying
the input in zero-shot conditions. However, using
language-specific BPE reduces the subword over-
lap between source and target sentences at training
time, and consequently reduces this copying be-
havior at test time (see Table 4). Model b) indeed
performs better (+5.1 BLEU) on the zero-shot di-
rections than the original baseline with shared BPE
(see Table 3).

6 Multi-Bridge Models

A common issue in zero-shot translation is out-
put in the wrong language. Previous work has
addressed this with semi-supervised training (Gu
et al., 2019; Arivazhagan et al., 2019; Zhang et al.,
2020). We explore whether the recent trend to train
English-centric models is to blame for this behav-
ior. In most cases, the model will wrongly produce
English output in zero-shot directions, since for all
non-English languages, English was the only target
language seen in training.

We suspect that adding even a small amount of
parallel data in pairs without English will improve

	Single-Bridge	Multi-Bridge
	trained	
de-en	29.3 ±0.31	29.3 ±0.56
cs-en	35.1 ±0.81	34.9 ±0.65
fi-en	37.5 ±0.90	37.7 ±0.75
fr-en	31.5 ±0.42	31.6 ±0.30
en-de	24.9 ±0.38	24.9 ±0.40
en-cs	28.1 ±0.81	28.0 ±0.70
en-fi	31.6 ±0.60	31.6 ±0.70
en-fr	31.3 ±0.67	31.5 ±0.42
average	31.2 ±0.60	31.2 ±0.56
	zero-shot	*trained*
cs-de	17.6 ±0.30	21.7 ±0.60
de-cs	18.3 ±0.42	21.7 ±0.78
fi-fr	26.3 ±0.61	33.7 ±1.01
fr-fi	17.8 ±0.49	23.1 ±0.51
average	20.0 ±0.44	25.1 ±0.72
	zero-shot	
cs-fr	24.6 ±0.36	28.2 ±0.71
fr-cs	20.0 ±0.53	22.2 ±0.66
de-fr	26.6 ±0.30	29.5 ±0.31
fr-de	19.0 ±0.46	21.5 ±0.66
fi-de	18.3 ±0.35	21.6 ±0.60
de-fi	16.9 ±0.61	20.8 ±0.83
average	20.9 ±0.43	24.0 ±0.62

Table 5: BLEU for single-bridge baseline with language-specific BPE (see Table 3), and model trained with 350k pairs in de↔cs and fi↔fr (multi-bridge). Both models use language-specific BPE segmentation.

	single-bridge			multi-bridge		
	tgt	en	src	tgt	en	src
cs-fr	95.92	1.33	0.03	97.28	0.55	0
fr-cs	95.33	0.38	0.50	95.57	0.32	0.22
de-fr	94.00	1.72	1.40	95.43	0.97	0.82
fr-de	92.47	2.75	1.23	95.43	1.17	0.43
fi-de	91.65	2.40	0.60	94.38	0.88	0.33
de-fi	91.93	1.57	1.52	93.58	0.77	0.85
average	93.55	1.69	0.89	95.30	0.78	0.44

Table 6: Percentage of output produced in the correct target language (tgt), English, and the source language (src) in zero-shot translation according to automatic language identification. Models from Table 5.

version of langdetect[5] to estimate the number of translations in the correct language. Even though the amount of parallel data in de↔cs and fi↔fr was small compared to the directions with English (350k vs. 5 million sentence pairs), the new model is less likely to produce output in the wrong target language, as shown in Table 6.

7 Comparison to Back-Translation and Encoder Alignment

Previous work on the zero-shot generalization of multilingual NMT systems has proposed back-translation or changes to the training objective to improve translation in unsupervised directions. While we consider our proposed solutions on the data side to be complementary, and easier to adopt widely, we still want to discuss the question how our solutions compare to previous work, and whether they can be combined.

7.1 Back-Translation

Previous work has used fine-tuning with synthetic, back-translated data for translation directions that were unseen at training time (Gu et al., 2019; Currey and Heafield, 2019; Zhang et al., 2020). While this can mitigate the problem of producing output in the wrong language, this approach is sensitive to the zero-shot translation quality of back-translation.[6] We perform experiments following Gu et al. (2019) where we create synthetic corpora for all zero-resource directions via back-translations (250k sentences per translation direction), and fine-tune our models on the concatena-

generalization, make models more sensitive to the language tag, and reduce the amount of English translations in the zero-shot directions. To test this hypothesis, we collect a small amount of parallel data in German-Czech and Finnish-French[4] and train our model with the additional language pairs. This new model has seen all non-English languages paired with exactly one other non-English language, but it still has zero-shot directions in de↔fr, fr↔cs and de↔fi. We use language-specific BPE segmentation and thus use the model with the best zero-shot performance from Table 3 as baseline.

The results in Table 5 show that even a small amount of parallel data in non-English language pairs increases generalization to unseen translation directions. The increase in BLEU scores for the newly added pairs de↔cs and fi↔fr are expected, but the new model also performs better on cs↔fr, de↔fr and fi↔de (+3.1 BLEU on average).

Following Zhang et al. (2020), we use the Python

---

[4] See Table 1 for details.

[6] Unless back-translation is done via a pivot language, but note that Gu et al. (2019) report slightly better results for direct zero-shot back-translation.

	single-bridge	+align	+bt	multi-bridge	+align	+bt
en↔* (avg)	31.2 ±0.60	31.4 ±0.20	30.5 ±0.33	31.2 ±0.56	31.4 ±0.28	30.3 ±0.23
cs-fr	24.6 ±0.36	25.8 ±0.46	25.8 ±0.40	28.2 ±0.71	28.8 ±0.49	29.0 ±0.06
fr-cs	20.0 ±0.53	20.5 ±0.12	20.4 ±0.32	22.2 ±0.66	22.6 ±0.21	23.8 ±0.06
de-fr	26.6 ±0.30	27.4 ±0.26	26.1 ±0.38	29.5 ±0.31	29.7 ±0.47	30.8 ±0.17
fr-de	19.0 ±0.46	19.6 ±0.15	19.8 ±0.21	21.5 ±0.66	21.6 ±0.31	22.0 ±0.00
de-fi	16.9 ±0.61	17.9 ±0.26	17.4 ±0.29	20.8 ±0.83	21.2 ±0.50	21.9 ±0.29
fi-de[1]	18.3 ±0.35	18.8 ±0.12	*20.1 ±0.40*	21.6 ±0.60	22.0 ±0.23	*23.9 ±0.00*
average	20.9 ±0.43	21.7 ±0.19	21.6 ±0.30	24.0 ±0.62	24.3 ±0.36	25.2 ±0.03

[1] used as development set for early stopping for +bt

Table 7: Zero-resource translation performance (BLEU) with single-bridge and multi-bridge multilingual models, fine-tuned with a cosine loss to reward encoder representation alignment (+align), and back-translation for zero-resource translation directions (+bt).

tion of this data, plus 250k sentence pairs per supervised translation direction. As base system for both back-translation and fine-tuning, we consider both our single-bridge and our multi-bridge system.

As stopping criterion during fine-tuning, we use BLEU on the Finnish↔German test set, one of the zero-resource language pairs. This leaves us with 5 translation directions that are still purely zero-resource.

7.2 Encoder Alignment

Arivazhagan et al. (2019) propose to use cosine distance as an additional loss term for multilingual models. The cosine distance loss encourages the model to produce encoder representations for sentences in the source language that are similar to the encoder representation of the same sentence in the target language. This, directly and indirectly, rewards similarity of encoder representations across all languages. We implement cosine loss in Sockeye, but instead of normalising sequence lengths by max pooling like Arivazhagan et al. (2019), we average encoder states, as proposed by Gouws et al. (2015). We introduce a new hyperparameter λ that scales cosine distance (CD) loss w.r.t. the standard cross-entropy (CE):

$$\mathcal{L} = (1 - \lambda) * CE + \lambda * CD \qquad (1)$$

We train models with $\lambda = 0.5$. As in our experiments with back-translation, we do not train from scratch, but fine-tune each of the single-bridge and multi-bridge models with a patience of 10.[7]

7.3 Results

Results are shown in Table 7. The gains from using more than one bridge language and back-translation are cumulative: Both the single- and the multi-bridge baseline improve with encoder alignment and back-translation, but the multi-bridge performs better overall in zero-resource directions.

Aligning encoder representations leads to an increase of 0.8 BLEU for the zero-shot directions for the single bridge data. In the multi-bridge scenario however, the effect of the additional loss is smaller (+0.3 BLEU on average). Table 7 contains only results for models with language-specific subword segmentation; but preliminary experiments show that aligning encoder representations of one of the baselines from Table 2 with jointly trained BPE gives a similar result: Encoder alignment alone does not fix the underlying issue caused by vocabulary overlap and English-centric models, even though we observe an increase of ~ 1.5 BLEU points in zero-shot directions over the baseline.

Back-translation leads to an average improvement of 0.7 BLEU with single-bridge data, and 1.2 BLEU with multi-bridge data. On the fully supervised pairs English↔{Czech,German,Finnish,French}, we observe a performance drop by 0.7–0.9 BLEU with back-translation. Again, back-translation alone does not seem to solve the issues of single-bridge setups, and the model benefits from additional supervised translation directions.

On the 6 remaining zero-shot translation direc-

---

[7] In a new training run with random initialization, the encoder produces highly similar representations for all languages from the start. Arivazhagan et al. (2019) report that fine-tuning yields better results.

tions, our pivoting baseline (Table 2) achieves an average BLEU of 23.7. Our best system with multi-bridge data and back-translation achieves 25.2, and thus outperforms our pivoting baseline by 1.5 BLEU.

8 Conclusions

We analyze the importance of shared subwords in multilingual models and find that language-specific BPE segmentation helps to reduce the amount of untranslated segments in zero-shot directions. Furthermore, we explore whether the tendency to produce the wrong output language can be attributed to using English as the only bridge language, and show that even with a small amount of additional training data in non-English language pairs, generalization to unseen translation directions improves as the model is less likely to produce output in the wrong language.

Compared to previous work, the methods we propose are easier to implement, since they only concern data collection and pre-processing, and result in higher gains for zero-shot directions. They are also compatible in principle with approaches that introduce new training objectives or model modifications, and we report best results when fine-tuning a multi-bridge model with back-translation for zero-resource translation directions.

For future work, we are interested in testing the effects of subword regularization (Kudo, 2018; Provilkov et al., 2020) on zero-shot translation performance, and scaling multi-bridge setups to massively multilingual settings.

Acknowledgements

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 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. Association for Computational Linguistics.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019. The Missing Ingredient in Zero-Shot Neural Machine Translation. *CoRR*, abs/1903.07091.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 Conference on Machine Translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Anna Currey and Kenneth Heafield. 2019. Zero-resource neural machine translation with monolingual pivot data. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 99–107, Hong Kong. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-Resource Translation with Multi-Lingual Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. Bilbowa: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille, France. PMLR.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved Zero-shot Neural Machine Translation via Ignoring Spurious Correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward Multilingual Neural Machine Translation with Universal Encoder and Decoder. In *International Workshop on Spoken Language Translation 2016*, Seattle, USA.

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2017. Effective strategies in zero-shot neural machine translation. In *International Workshop on Spoken Language Translation 2017*, Tokyo, Japan.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The Sockeye Neural Machine Translation Toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (AMTA)*, volume 1: Research Papers, pages 200–207.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

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.

Surafel Melaku Lakew, Quintino Lotito, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. FBK's Multilingual Neural Machine Translation System for IWSLT 2017. In *International Workshop on Spoken Language Translation 2017*, pages 113–119, Tokyo, Japan.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-Dropout: Simple and Effective Subword Regularization. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.

A   Model Size and Training

All models are trained with the Sockeye toolkit (Hieber et al., 2018)[8] on 5 Tesla-V100 (16GB) GPUs for 4-5 days.

model type	number of parameters
1 joint bpe baseline	60,516,602
2 language specific bpe, no vocabulary overlap	69,728,543
3 language specific bpe, vocabulary overlap	62,470,619
4 language specific bpe, vocabulary overlap, multi-bridge	62,490,626

Table 8: Number of Parameters per Model Type. Numbers vary between models due to different vocabulary sizes. Vocabulary is built automatically based on training data, therefore, 4 has a slightly larger vocabulary than 3. Cosine-loss models have the same number of parameters as 3 and 4.

model type	best checkpoint and BLEU					
	seed=1		seed=2		seed=3	
1 joint bpe baseline	106	30.7	95	30.9	94	31.3
2 language specific bpe, no vocabulary overlap	66	29.3	115	31.0	60	30.0
3 language specific bpe, vocabulary overlap	90	30.3	120	31.0	55	29.4
4 language specific bpe, vocabulary overlap, multi-bridge	117	29.3	80	28.6	59	28.2

Table 9: Best checkpoint according to BLEU on development set (patience=10). Sentence pairs in the development sets are identical for each model, however the dev set for model 4 contains additional samples in cs↔de and fi↔fr.

---

[8] https://github.com/awslabs/sockeye

B Hyperparameters

Training Hyperparameters for all Models	
training settings:	
batch type	word
batch size	16384
max-seq-len	100:100
word-min-count	1:1
seed	1/2/3
model settings:	
encoder	transformer
decoder	transformer
num-layers	6:6
transformer-model-size	512
transformer-attention-heads	8
transformer-feed-forward-num-hidden	2048
transformer-preprocess	n
transformer-postprocess	dr
transformer-positional-embedding-type	fixed
num-embed	512:512
weight-tying-type	src_trg_softmax
optimization settings:	
optimizer	adam
optimized-metric	bleu
checkpoint interval	4000
max-num-checkpoint-not-improved	10
min-num-epochs	0
max-updates	1001000
label-smoothing	0.1
gradient-clipping-threshold	-1
initial-learning-rate	0.0001
learning-rate-reduce-num-not-improved	8
learning-rate-reduce-factor	0.7
learning-rate-scheduler-type	plateau-reduce
learning-rate-warmup	0
initialization settings:	
weight-init	xavier
weight-init-scale	3.0
weight-init-xavier-factor-type	avg
dropout settings:	
embed-dropout	0:0
transformer-dropout-attention	0.1
transformer-dropout-act	0.1
transformer-dropout-prepost	0.1

Table 10: Sockeye hyperparameters for all models (values with ':' = encoder:decoder)