Surprise Language Challenge: Developing a Neural Machine Translation System between Pashto and English in Two Months

Surprise Language Challenge: Developing a Neural Machine Translation System between Pashto and English in Two Months

Alexandra Birch,[1] **Barry Haddow,**[1] **Antonio Valerio Miceli Barone,**[1]
Jindřich Helcl,[1] **Jonas Waldendorf,**[1] **Felipe Sánchez-Martínez,**[2]
Mikel L. Forcada,[2] **Miquel Esplà-Gomis,**[2] **Víctor M. Sánchez-Cartagena,**[2]
Juan Antonio Pérez-Ortiz,[2] **Wilker Aziz,**[3] **Lina Murady,**[3] **Sevi Sariisik,**[4]
Peggy van der Kreeft,[5] **Kay MacQuarrie**[5]

[1]University of Edinburgh, [2]Universitat d'Alacant, [3]Universiteit van Amsterdam,
[4]BBC , [5]Deutsche Welle

Abstract

In the media industry, the focus of global reporting can shift overnight. There is a compelling need to be able to develop new machine translation systems in a short period of time, in order to more efficiently cover quickly developing stories. As part of the low-resource machine translation project GoURMET, we selected a surprise language for which a system had to be built and evaluated in two months (February and March 2021). The language selected was Pashto, an Indo-Iranian language spoken in Afghanistan, Pakistan and India. In this period we completed the full pipeline of development of a neural machine translation system: data crawling, cleaning, aligning, creating test sets, developing and testing models, and delivering them to the user partners. In this paper we describe the rapid data creation process, and experiments with transfer learning and pretraining for Pashto-English. We find that starting from an existing large model pre-trained on 50 languages leads to far better BLEU scores than pretraining on one high-resource language pair with a smaller model. We also present human evaluation of our systems, which indicates that the resulting systems perform better than a freely available commercial system when translating from English into Pashto direction, and similarly when translating from Pashto into English.

1 Introduction

The Horizon 2020 European-Union-funded project GoURMET[1] (Global Under-Resourced MEdia Translation) aims to improve neural machine translation for under-resourced language pairs with a special emphasis on the news domain. The two media partners in the GoURMET project, the BBC in the UK and Deutsche Welle (DW) in Germany, publish news content in 40 and 30 different languages, respectively, and gather news in over 100 languages. In such a global information scenario, machine translation technologies become an important element in the everyday workflow of these media organisations.

[1] https://GoURMET-project.eu/

Surprise language exercises (Oard et al., 2019) started in March 2003, when the US Defense Advanced Research Projects Agency (DARPA) designated Cebuano, the second most widely spoken indigenous language in the Philippines, as the focus of an exercise. Teams were given only ten days to assemble language resources and to create whatever human language technology they could in that time. These events have been running annually ever since.

The GoURMET project undertook its surprise language evaluation as an exercise to bring together the whole consortium to focus on a language pair of particular interest to the BBC and DW for a short period of time. Given the impact of the COVID-19 pandemic, a two-month period was considered realistic. On 1 February 2021, BBC and DW revealed the chosen language to be Pashto. By completing and documenting how this challenge was addressed, we prove we are able to bootstrap a new high quality neural machine translation task within a very limited window of time.

There has also been a considerable amount of recent interest in using pretrained language models for improving performance on downstream natural language processing tasks, especially in a low resource setting (Liu et al., 2020; Brown et al., 2020; Qiu et al., 2020), but how best to do this is still an open question. A key question in this work is how best to use training data which is not English (`en`) to Pashto (`ps`) translations. We experimented, on the one hand, with pretraining models on a high-resource language pair (German–English, one of the most studied high-resource language pairs) and, on the other hand, with fine-tuning an existing large pretrained translation model (mBART50) trained on parallel data involving English and 49 languages including Pashto (Tang et al., 2020). We show that both approaches perform comparably or better than commercial machine translation systems especially when Pashto is the output language, with the large multilingual model achieving the highest translation quality between our two approaches.

The paper is organised as follows. Section 2 motivates the choice of Pashto and presents a brief analysis of the social and technical context of the language. Section 3 describes the efforts behind the crawling of additional monolingual and parallel data in addition to the linguistic resources already available for English–Pashto. Section 4 introduces the twofold approach we followed in order to build our neural machine translation systems: on the one hand, we developed a system from scratch by combining mBART-like pretraining, German–English translation pretraining and fine-tuning; on the other hand, we also explored fine-tuning on the existing pretrained multilingual model mBART50. We present automatic results and preliminary human evaluation of the systems in Section 5.

2 The Case for Pashto

The primary goal, when selecting which low-resource language pair to work on, was to provide a tool that would be useful to both the BBC and Deutsche Welle. It had to be an under-resourced language with high news value and audience growth potential, and one that could pose a satisfactory research challenge to complement the wider goals of the project. Pashto ticked all of these boxes.

Pashto is one of the two official languages of Afghanistan along with Dari. Almost half of the country's 37.5 million people, up to 10 percent of the population in neighbouring Pakistan, and smaller communities in India and Tajikistan speak Pashto, bringing estimates of Pashto speakers worldwide around 45–50 million (Brown, 2005). Europe hosts a growing number of Pashto speakers, too. As of the end of 2020, there were over 270,000 Afghans living in Ger-

many[2] and 79,000 in the UK[3]. Projecting from Afghanistan's national linguistic breakdown,[4] up to half of these could be Pashto speakers.

Pashto (also spelled *Pukhto* and *Pakhto* is an Iranian language of the Indo-European family and is grouped with other Iranian languages such as Persian, Dari, Tajiki, in spite of major linguistic diferences among them. Pashto is written with a unique enriched Perso-Arabic script with 45 letters and four diacritics.

Translating between English and Pashto poses interesting challenges. Pashto has a richer morphology than that of English; the induced data sparseness may partly be remedied with segmentation in subword units tokenization models such as SentencePiece (Kudo and Richardson, 2018), as used in mBART50. There are Pashto categories in Pashto that do not overtly exist in English (such as verb aspect or the oblique case in general nouns) and categories in English that do not overtly exist in Pashto (such as definite and indefinite articles), which may pose a certain challenge when having to generate correct text in machine translation output.

Due to the chronic political and social instability and conflict that Afghanistan has experienced in its recent history, the country features prominently in global news coverage. Closely following the developments there remains a key priority for international policy makers, multilateral institutions, observers, researchers and the media, alongside the wider array of individual news consumers. Pashto features in BBC Monitoring's language portfolio. Enhancing the means to better follow and understand Pashto resources first hand through machine translation offers a valuable contribution.

The interest of commercial providers of machine translation solutions in Pashto is recent and there is room for improvement for existing solutions. Google Translate integrated Pashto in 2016, ten years after its launch.[5] Amazon followed suit in November 2019 and Microsoft Translator added Pashto into its portfolio in August 2020.[6] Nevertheless, Pashto has been of interest to the GoURMET Media Partners long before that. Deutsche Welle started its Pashto broadcasts in 1970 and BBC World Service in 1981. Both partners are currently producing multimedia content (digital, TV, radio) in Pashto. BBC Pashto reaches 10.4 million people per week, with significant further growth potential.

3 Data Creation

The process of data collection and curation is divided into two clearly different processes to obtain: (a) training data, and (b) development and test data. This section describes these two proceses. Note that our neural systems were trained with additional training data which will be described in Section 4.

3.1 Training Data

Traning data consists of English–Pashto parallel data and Pashto monolingual data, and was obtained by two means: directly crawling websites likely to contain parallel data, and crawling the top-level domain (TLD) of Afganistan (`.af`), where Pashto is an official language.

Direct crawling was run using the tool Bitextor (Espla-Gomis and Forcada, 2010) on a collection of web domains that were identified as likely to contain English–Pashto parallel data.

[2]German Federal Statistical Office, `https://bit.ly/3fg5LGr`

[3]ONS statistics, `https://bit.ly/3oh92cS`

[4]World Factbook, `https://www.cia.gov/the-world-factbook/field/languages/`

[5]`https://blog.google/products/translate/google-translate-now-speaks-pashto`

[6]`https://bit.ly/3w4WMPi`

This list was complemented by adding the web domains used to build the data sets released for the parallel corpus filtering shared task at WMT2020 (Koehn et al., 2020). A total of 427 websites were partially crawled during three weeks following this strategy, from which only 50 provided any English–Pashto parallel data.

Crawling the Afganistan TLD was carried out by using the tool LinguaCrawl.[7] An initial set of 30 web domains was manually identified, mostly belonging to national authorities, universities and news sites. Starting from this collection, a total of 150 new websites were discovered containing documents in Pashto. After document and sentence alignment (using the tool Bitextor), 138 of them were identified to contain any English–Pashto parallel data.

3.2 Test and Development Data

The development and test sets were extracted from a large collection of news articles in Pashto and English, both from the BBC and the DW websites. In both cases, documents in English and documents in Pashto were aligned using the URIs of the images included in each of them, as, in both cases, these elements are language-independent. Given the collection of image URLs in a document in English (I_{en}) and that collection in a document in Pashto (I_{ps}), the similarity score between these two documents was computed as:

$$\mathrm{score}(I_{\mathrm{en}}, I_{\mathrm{ps}}) = \frac{1}{|I_{\mathrm{en}} \cup I_{\mathrm{ps}}|} \sum_{i \in I_{\mathrm{en}} \cap I_{\mathrm{ps}}} \mathrm{IDF}(i)$$

where $\mathrm{IDF}(i)$ is the inverse document frequency (Robertson, 2004) of a given image. English–Pashto pairs of documents were ranked using this score, and document pairs with a score under 0.1 were discarded.

After document alignment, documents were split into sentences and all the Pashto segments were translated into English using Google Translate.[8] English segments and machine-translated Pashto segments in each pair of documents were compared using the metric `chrF++` (Popović, 2017), and the best 4,000 segment pairs were taken as candidate segments for human validation.

Finally, a team of validators from BBC and DW manually checked the candidate segments. Through human validation, 2,000 valid segment pairs were obtained from the BBC dataset, and 815 for the DW dataset. The BBC dataset whas then divided into two sub-sets: 1,350 segment pairs for testing and 1,000 segment pairs for development; for the DW data, the whole set of 815 segment pairs was used as a test set.

3.3 Final Data Set Statistics

Table 1 shows the number of segment pairs, the number of tokens both in Pashto and English, and the average number of tokens per segment for the corpus obtained.

4 Training of Neural Machine Translation Systems

We developed two different neural models: a *from-scratch* system, and a larger and slower system based on an existing pretrained model. The development of the former starts with a mediun-size randomly-initialized transformer (Vaswani et al., 2017), whereas the latter is obtained by fine-tuning the larger downloadable mBART50 pretrained system (Tang et al., 2020).

[7] `https://github.com/transducens/linguacrawl`
[8] `https://translate.google.com`

		Pastho		English	
Corpus name	# segm. pairs	# tokens	tokens/segm.	# tokens	tokens/segm.
Crawled	59,512	759,352	12.8	709,630	11.9
BBC Test	1,350	25,453	18.8	30,417	22.5
BBC Dev	1,000	18,793	18.8	22,438	22.4
DW Test	813	14,956	18.3	20,797	25.5

Table 1: Crawled and in-house parallel corpora statistics.

Remarkably, mBART50 has been pretrained with some Pashto (and English) data which makes it a very convenient model to explore.

The size of pretrained models make them poor candidates for production environments, especially where they are required to run on CPU-only servers as it is the case in the GoURMET project, yet translations have to be available at a fast rate. In those scenarios, the from-scratch system may be considered a more efficient alternative. Our mBART50 systems can still be useful in those scenarios to generate synthetic data with which to train smaller models.

4.1 From-scratch Model

This has been trained "from scratch" in the sense that it does not exploit third-party pretrained models. It was built by using a combination of mBART-like pretraining (Liu et al., 2020), German–English translation pretraining and fine-tuning. We used the Marian toolkit (Junczys-Dowmunt et al., 2018) to implement this model.

Data preparation. We use different version of training data in different rounds, starting from a small and relatively high-quality dataset and adding more data as it becomes available in parallel to our model training efforts.

Initial data. For our initial English-Pashto parallel training corpus we use the WMT 2020 data excluding ParaCrawl. This dataset consists mostly of OPUS data.[9] We did not use existing data from the ParaCrawl project[10] at this point because it requires filtering to be used effectively and we first wanted to build initial models on relatively clean data. For our initial monolingual corpus we use all the released Pashto NewsCrawl[11] and the 2019 version of the English NewsCrawl[12]. Finally, we also use the Pashto-English corpus that was submitted by the Bytedance team to the WMT 2020 cleaning shared task (Koehn et al., 2020).

For pretraining the German–English model we use exisiting WMT data (Europarl, Common Crawl and News Commentary). We use WMT dev and test sets[13] for early stopping and evaluation, and the BBC development and test sets (see Section 3) for additional evaluation. We process these data with standard Moses cleaning and punctuation normalization scripts[14]. For Pashto we also filter the training data with a language detector based on Fasttext word embeddings to remove the sentences in incorrect languages, and we apply an external character

[9] http://opus.nlpl.eu
[10] https://paracrawl.eu/
[11] http://data.statmt.org/news-crawl/ps/
[12] http://data.statmt.org/news-crawl/en/news.2019.en.shuffled.deduped.gz
[13] http://www.statmt.org/wmt20/translation-task.html
[14] https://github.com/marian-nmt/moses-scripts

normalization script[15].

We generate a shared SentencePiece vocabulary (BPE mode) on a merged corpus obtained by concatenating the German–English training data, the first 6,000,000 English monolingual sentences, and all the Pashto monolingual and Pashto–English parallel data each upsampled to approximately match the size of the English monolingual data. We reserve a small number of special tokens for language id and mBART masking. The total vocabulary size is $32,000$ token types.

mBART-like pretraining. We pretrain a standard Marian transformer-based model (Junczys-Dowmunt et al., 2018) with a reproduction of the mBART (Liu et al., 2020) pretraining objective with our English and Pashto monolingual data. We use only the masking distortion, but not the consecutive sentences shuffling distortion, as our monolingual data is already shuffled and therefore the original ordering of the sentences is not available. We also did not use online backtranslation as it is not available in Marian. We upsample the Pashto data so that each batch contains an equal amount of English and Pashto sentences. The output language is specified by a language identification token at the beginning of the source sentence. We perform early stopping on cross-entropy evaluated on a monolingual validation set obtained in the same way as the training data.

Exploitation of German–English data. We pretrain a bidirectional German–English model with the same architecture as the mBART-like model defined above (see Section 4.1 above). As in the mBART model, we use a language id token prepended to the source sentence to specify the output language. We use WMT data (see Section 4.1) for training and early stopping.

Training of the from-scratch system. Training consists of fine-tuning a pretrained model with Pashto–English parallel data, using it to generate initial backtranslations which are combined with the parallel data and used to train another round of the model, starting again from a pretrained model. At this point, we include the first 220,000 sentence pairs of "Bytedance" filtered parallel data, sorted by filtering rank.

Following similar work with English–Tamil (Bawden et al., 2020), we start with our mBART-like model and we fine-tune it in the Pashto→English direction with our parallel data. Then we use this model to backtranslate the Pashto monolingual data, generating a pseudo-parallel corpus which we combine with our true parallel corpus and use to train a English→Pashto model again starting from mBART. We use this model to backtranslate the first 5,000,000 monolingual English sentences (we also experimented with the full corpus, but found minimal difference), and we train another round of Pashto→English followed by another round of English→Pashto, both initialized from mBART pretraining.

After this phase we switch to German–English pretraining. Due to the limited available time, we did not experiment on the optimal switching point between the two pretraining schemes; we based this decision instead on our previous experience with English–Tamil (Bawden et al., 2020). We perform four rounds (counting each translation direction as one round) of iterative backtranslation with initialization from German–English pretraining.

On the last round we evaluate multiple variants of training data as more data became available. We found that including additional targeted crawls on news websites (see Section 3) improved translation quality. Adding synthetic English paraphrases or distillation data from the large mBART50 model however did not provide improvements.

[15] https://github.com/rnd2110/SCRIPTS_Normalization

4.2 mBART50-Based Model

The experiments in this section try to show how far we can get by building our English–Pashto NMT systems starting from the recently released (January 2021) pretrained multilingual model mBART50 (Tang et al., 2020).[16] mBART50 is an extension of mBART (Liu et al., 2020) additionally trained on collections of parallel data with a focus on English as source (*one-to-many* system or mBART50 1–to–n for short) or target (*many-to-one* system). As of March 12th 2021 the n–to–1 system is not available for download; therefore, we used the *many-to-many* (mBART50 n–to–n for short) version as a replacement. As regards mBART50 1–to–n, our preliminary experiments showed that the bare model without further fine-tuning gave in the English→Pashto direction results similar to mBART50 n–to–n. We also confirmed that mBART50 1–to–n gives very bad results on Pashto→English as the system has not been exposed to English during pretraining. Consequently, our experiments focus on mBART50 n–to–n for both translation directions; being a multilingual model, this will also reduce the number of experiments to consider as the same system is trained at the same time in both directions. As already mentioned, mBART50 was pretrained with Pashto and English data which makes it a very convenient model to start with.

Experimental set-up. Although these models have already processed English and Pashto texts (not necessarily mutual translations) during pretraining, fine-tuning them on English–Pashto parallel data may improve the results. Therefore, apart from evaluating the plain non-fine-tuned mBART50 n–to–n system, we *incrementally* fine-tuned it in three consecutive steps:

1. First, we fine-tuned the model with a very small parallel corpus of 1,400 sentences made of the TED Talks and Wikimedia files in the clean parallel data set provided for the WMT 2020 shared task on parallel corpus filtering and allignment for low-resource conditions.[17] Validation-based early stopping was used and training stopped after 20 epochs (this took around 20 minutes on one NVIDIA A100 GPU). This scenario may be considered as a few-shot adaptation of the pretrained model.
2. Then, we further fine-tuned the model obtained in the first step with a much larger parallel corpus of 343,198 sentences made of the complete WMT 2020 clean dataset and the first 220,000 sentences in the corpus resulting from the system submitted by Bytedance to the same shared task (Koehn et al., 2020). Training stopped after 7 epochs (around 2 hours and 20 minutes on one A100 GPU).
3. Finally, we additionally fine-tuned the model previously obtained with a synthetic English–Pashto parallel corpus built by translating 674,839 Pashto sentences[18] into English with the model resulting from the second step. The Pashto→English model in the second step gave a BLEU score of 25.27 with the BBC test set, allowing us to assume that the synthetic English generated has reasonable quality. Note that we carried out a multilingual fine-tuning process and therefore the synthetic corpus is used to fine-tune the system in both directions, which yields giving a system which will be probably worse than the initial one in the Pashto→English direction. Training stopped after 7 epochs (around 4 hours on one A100 GPU). Only sentences in the original Pashto monolingual corpus with lengths between 40 and 400 characters were included the synthetic corpus.

Fine-tuning configuration. Validation-based early stopping was applied with a patience value of 10 epochs. The development set evaluated by the stopping criterion was the in-house

[16] `https://github.com/pytorch/fairseq/blob/master/examples/multilingual`

[17] `http://www.statmt.org/wmt20/parallel-corpus-filtering.html`

[18] Concatenation of all files available at `http://data.statmt.org/news-crawl/ps` on March 2021 except for `news.2020.Q1.ps.shuffled.deduped.gz`.

	BBC test	DW test	FLORES devtest
Google	12.84	10.19	9.16
from-scratch	15.00	10.41	9.73
mBART50	2.47	1.53	7.56
+ small	9.93	7.67	8.24
+ small, large	11.85	10.31	**10.82**
+ small, large, synthetic	**18.55**	**12.54**	8.61

Table 2: BLEU scores of the English→Pashto systems. Each column represents a different test set used to compute the score. The first row contains the results for a commercial general-purpose system. The second row contains the scores for the model trained from scratch. The results for mBART50 correspond, from top to bottom, to a non-fine-tuned mBART50 n–to–n system, and this system incrementally fine-tuned with a small parallel corpus of 1,400 sentences, a larger parallel corpus of 343,198 sentences, and a synthetic corpus of 674,839 sentences obtained from Pashto monolingual text.

	BBC test	DW test	FLORES devtest
Google	0.413	**0.374**	**0.345**
from-scratch	0.411	0.336	0.331
mBART50	0.170	0.147	0.284
+ small	0.351	0.301	0.314
+ small, large	0.389	0.341	0.343
+ small, large, synthetic	**0.463**	**0.374**	0.330

Table 3: chrF2 scores of the English→Pashto systems. See table 2 for details.

validation set made of 1,000 sentences curated by the BBC presented in Section 3. Note that no hyper-parameter tuning was performed and, therefore, better results could be attained after a careful grid search hyper-parameter optimization.

Embedding table filtering. As already shown, these models may have strong memory requirements. As a way to verifying whether these requirements could be relaxed, we ran a script to reduce the embedding tables by removing those entries corresponding to tokens not found in a collection of English and Pashto texts. The original vocabulary size of 250,054 tokens of the mBART50 model was reduced to 16,576. This resulted in a relatively small decrease in memory consumption: for example, the GPU memory requirements of mBART n–to–n at inference time (setting the maximum number of tokens per mini-batch to 100) moved from around 4 GB to around 3 GB.

5 Results and Discussion

Tables 2 and 3 show BLEU and chrF2 scores, respectively, for the English to Pashto systems with different test sets. The evaluation metrics for the Google MT system are also included for reference purposes. Similarly, tables 4 and 5 show BLEU and chrF2 scores, respectively, for the Pashto to English systems. All the scores were computed with `sacrebleu` (Post, 2018).

The test sets considered are the two in-house parallel sets created by BBC and DW (see Section 3) and the devtest set provided in the FLORES[19] benchmark (2,698 sentences).

[19] `https://github.com/facebookresearch/flores`

	BBC test	DW test	FLORES devtest
Google	**35.03**	**24.65**	**21.54**
from-scratch	20.00	15.06	14.90
mBART50	19.42	15.30	14.59
+ small	22.55	17.50	14.77
+ small, large	25.27	19.13	17.71
+ small, large, synthetic	25.38	17.88	17.08

Table 4: BLEU scores of the Pashto→English systems. See table 2 for details.

	BBC test	DW test	FLORES devtest
Google	**0.628**	**0.532**	**0.506**
from-scratch	0.482	0.445	0.411
mBART50	0.456	0.431	0.423
+ small	0.512	0.471	0.420
+ small, large	0.527	0.481	0.451
+ small, large, synthetic	0.535	0.477	0.448

Table 5: chrF2 scores of the Pashto→English systems. See table 2 for details.

As can be seen, the from-scratch system provides worse results than the mBART50-based model obtained after the three-step fine-tuning procedure, which may be easily explained by the smaller number of parameters and the lack of initial knowledge.

Regarding the mBART50-based models, for the English→Pashto direction, the scores obtained with the non-fine-tuned models for the FLORES test set are considerably higher than those corresponding to the BBC and DW test sets, which suggests that either they belong to different domains, or they contain very different grammatical or lexical structures, or the FLORES corpus was used to pretrain mBART50. This indicates that fine-tuning could provide a twofold benefit: on the one hand, it may allow the model to focus on our two languages of interest, partially forgetting what it learned for other languages; on the other hand, it may allow the model to perform domain adaptation. In the English→Pashto direction each successive fine-tuning step improves the scores, except when the last model is evaluated against the FLORES devtest set, which makes sense as the development set belongs to the domain of the BBC and DW test sets. Notably, the system resulting from the three-step fine-tuning process improves Google's scores as of April 2021. In the Pashto→English direction, the same trend can be observed, although in this case the best mBART50-based system is noticeably behind the scores of Google's system, yet it still provides scores higher than those for the other translation direction.

5.1 Human Evaluation

Four senior editors from BBC Pashto were asked to score translations in a blind exercise from 1 to 100, with 100 indicating top quality. The evaluators were provided with four outputs for both English→Pashto and Pashto→English samples; these outputs were obtained from the mBART50-based models with beam widths of 1 and 5, from the from-scratch system and from Google Translate. Table 6 demonstrates the average scores by human evaluators for 20 selected sentences. This small sample means that the scores are indicative of the model performance, but together with the BLEU scores gives the user partners confidence in the translation quality.

	Pashto→English	English→Pashto
Google	83.80	68.50
from-scratch	63.50	67.65
mBART50 (beam width 1)	**85.15**	83.60
mBART50 (beam width 5)	83.15	**92.30**

Table 6: Average human scores for 20 translations generated by 3 of our models and a commercial general-purpose system.

Both mBART50-based models performed very strongly, with outcomes significantly better than Google or from-scratch models into Pashto. The model will be made available for further utilization for monitoring and content creation purposes of the media partners as well as the API's public-facing site.[20] The confidence derived from the human evaluation has encouraged the BBC and DW to adopt Pashto↔English machine translation solutions.

6 Conclusion

We present a description of our rapid Pashto↔English machine translation system building exercise. We performed extensive crawling and data cleaning and alignment, combined with pretraining experiments to deliver a strong translation system for a low-resource language. We test different transfer learning approaches and show that large, multilingual models perform better than smaller models from a high-resource language pair. The data [21], models [22] and tools [23] are shared publically.

Acknowledgments

Work funded by the European Union's Horizon 2020 research and innovation programme under grant agreement number 825299, project Global Under-Resourced Media Translation (GoURMET). Some of the computational resources used in the experiments have been funded by the European Regional Development Fund (ERDF) through project IDIFEDER/2020/003.

References

Bawden, R., Birch, A., Dobreva, R., Oncevay, A., Miceli Barone, A. V., and Williams, P. (2020). The university of edinburgh's english-tamil and english-inuktitut submissions to the wmt20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 92–99, Online. Association for Computational Linguistics.

Brown, K. (2005). *Encyclopedia of language and linguistics*, volume 1. Elsevier.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are Few-Shot learners.

[20] https://translate.GoURMET.newslabs.co/
[21] http://data.statmt.org/gourmet/models/en-ps
[22] http://data.statmt.org/gourmet/models/en-ps
[23] https://gourmet-project.eu/data-model-releases/

Espla-Gomis, M. and Forcada, M. (2010). Combining content-based and url-based heuristics to harvest aligned bitexts from multilingual sites with bitextor. *The Prague Bulletin of Mathematical Linguistics*, 93(2010):77–86.

Junczys-Dowmunt, M., Heafield, K., Hoang, H., Grundkiewicz, R., and Aue, A. (2018). Marian: Cost-effective high-quality neural machine translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 129–135.

Koehn, P., Chaudhary, V., El-Kishky, A., Goyal, N., Chen, P.-J., and Guzmán, F. (2020). Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742. Association for Computational Linguistics.

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.

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.

Oard, D. W., Carpuat, M., Galuscakova, P., Barrow, J., Nair, S., Niu, X., Shing, H.-C., Xu, W., Zotkina, E., McKeown, K., Muresan, S., Kayi, E. S., Eskander, R., Kedzie, C., Virin, Y., Radev, D. R., Zhang, R., Gales, M. J. F., Ragni, A., and Heafield, K. (2019). Surprise languages: Rapid-response cross-language IR. In *Proceedings of the Ninth International Workshop on Evaluating Information Access, EVIA 2019*.

Popović, M. (2017). chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618.

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, Belgium, Brussels. Association for Computational Linguistics.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained models for natural language processing: A survey.

Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60.

Tang, Y., Tran, C., Li, X., Chen, P.-J., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint*.

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 Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.