Global Under-Resourced MEedia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D1.4 – Final report on data gathering and augmentation

Nature	Report	**Work Package**	WP1
Due Date	31/03/2022	**Submission Date**	31/03/2022
Main authors	Felipe Sanchez-Martinez (UA)		
Co-authors	Miquel Esplà-Gomis (UA), Víctor M. Sánchez-Cartagena (UA), Aarón Galiano (UA), Alexandra Birch (UEDIN), Barry Haddow (UEDIN)		
Reviewers	Wilker Aziz (UVA)		
Keywords	data gathering, data augmentation, corpus crawling, machine translation		
Version Control			
v0.8	**Status**	First draft	06/06/2022
v1.0	**Status**	Final	31/06/2022

Contents

Abstract

This deliverable reports the work conducted within work package WP1 on data gathering and data augmentation during the second half of the project. It focuses on three main tasks: crawling of monolingual and bilingual corpora from the web, processing the data dumps provided by the user partners to obtain in-domain corpora for development and testing, and the application and development of data augmentation techniques for generating synthetic training corpora.

1 Introduction

Work package WP1 focuses on the identification and collection on linguistic resources for the languages of interest to GoURMET (task T1.1), the identification, collection and evaluation of monolingual and bilingual corpora (task T1.2), and the generation of synthetic data and lexical augmentation (task T1.3). This deliverable is the continuation of deliverable *D1.2 Initial progress report on data gathering and augmentation* and reports the work conducted within WP1 on data gathering for the languages we have worked with during the second half of the project, and on data augmentation. Deliverable *D1.5 Final release of project data* provides pointers for downloading the corpora we have crawled and made available from the project webpage.

The crawling of parallel corpora for the languages we report in this deliverable —Burmese, Hausa, Igbo, Macedonian, Pashto, Tigrinya and Yoruba— has been challenging because the amount of existing resources was extremely scarce, which made it difficult to identify reliable parallel segments.

The rest of this report is organized as follows. The next section reports our work on data crawling from the web, the problems encountered and the way they have been addressed. Section 3 describes the processing of the BBC and DW data dumps to obtain parallel corpora for development and testing. Section 4 describes the work we have conducted on data augmentation by generating synthetic training corpora. The report ends with a section summarizing the work conducted within WP1 during the second half of the project and a list of publications and software we have released or contributed to.

2 Parallel and monolingual data crawling

This section describes the process followed to acquire parallel and monolingual data from the Internet as a resource to train neural MT (NMT) systems. Deliverable *D1.5 Final release of project data* provides pointer for downloading them.

The process followed to acquire the corpora is very similar to that reported in deliverable *D1.2 Initial progress report on data gathering and augmentation* and can be split into three different steps: crawling, document and segment alignment, and cleaning. In what follows we describe what changes with respect to the process described in *D1.2 Initial progress report on data gathering and augmentation*, for a detailed description of the process we refer the reader to that deliverable.

2.1 Crawling

The process has not changed with respect to what was reported in Section 2.1 of deliverable *D1.2 Initial progress report on data gathering and augmentation*. We followed two complementary approaches: downloading as many documents as possible from a known collection of websites, and exploring a specific top-level domain to find websites from which to crawl data. The list of websites to crawl was automatically obtained by leveraging automatic-language-identification metadata from the CommonCrawl corpus[1] and identifying URLs in Wikimedia dumps[2] pointing to external websites. We also crawled websites manually identified after inspecting websites containing text in the targeted languages.

[1] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages
[2] https://dumps.wikimedia.org/backup-index.html

2.2 Document and segment alignment

In this step, parallel data is identified and obtained from the monolingual data crawled in the two languages of interest. As in the first half of the project, this stage is carried out using Bitextor (Espla-Gomis and Forcada, 2010),[3] a free/open-source tool to which we have contributed, although with some changes as regards the way document alignment is performed for one of the languages. In general, candidate parallel documents are identified using a method based on bag-of-word-overlapping metrics that rely on bilingual lexicons automatically obtained from the available parallel data (as described in deliverable *D1.2 Initial progress report on data gathering and augmentation*, Section 2.2). However, for Burmese we used a NMT system trained on the available parallel corpora. In this way we avoided the problems caused by the fact that Burmese, in general, does not segment sentence into tokens or words, and when it does, the tokenization is arbitrary because it is not mandatory and depends solely on the writing style of the author.

After document alignment, every pair of parallel documents is then aligned at the sentence level. For all languages, with the exception of Burmese, we used Hunalign (Varga et al., 2007) for this task. The same bilingual dictionary used for document alignment was provided to Hunalign to improve the accuracy of the alignment. In the case of Burmese we used BLEUAlign[4](Sennrich and Volk, 2011) for sentence alignment, a method that uses MT; the system used was the very same used for document alignment.

2.3 Cleaning

Cleaning corpora implies removing noisy sentence pairs; pairs that are either incorrectly aligned or do not contain valid text in the expected language. After experimenting with Bicleaner[5] (Sánchez-Cartagena et al., 2018) and LASER (Schwenk, 2018) during the first half of the project, for the language in this second half, with the exception of Burmese, we used an improved version of Bicleaner (Esplà-Gomis et al., 2020) using extremely randomized trees (Geurts et al., 2006) and additional features using the frequency of the words in monolingual corpora to better exploit the discriminatory power of low-frequency words. For Burmese we used Bicleaner AI (Zaragoza-Bernabeu et al., 2022),[6] a fork of Bicleaner using neural networks.

Bicleaner models are language-pair specific and use probabilistic bilingual dictionaries in both translation directions. These dictionaries were obtained as a by-product of the process of producing the bilingual lexicon for document alignment (see Section 2.2). The classifier used to score each sentence pair was trained on additional parallel data. In addition to the score provided by the classifier, a character-level language model was also used to provide monolingual confidence.

Bicleaner AI provides two types of models, light models for fast scoring and full models for high performance. Full models use fine-tuned XLMRoberta (Conneau et al., 2020) trained on bilingual corpora. The experiments we have performed with Burmese show that Biclenaer AI with a full model performs better than Bicleaner when both classifiers are trained on the same parallel corpora.

[3] https://github.com/bitextor/bitextor/branches
[4] https://github.com/rsennrich/Bleualign
[5] https://github.com/bitextor/bicleaner/
[6] https://github.com/bitextor/bicleaner-ai

Language pair	# sentences	# left tokens	# right tokens
English–Hausa	19,904	417,458	413,871
English–Igbo	29,969	550,567	565,343
English–Pashto	59,512	709,630	759,352
English–Yoruba	113	896	1091
English–Burmese	1033	16,343	23,300
English–Macedonian	71,506	2,259,609	1,814,624
English–Tigrinya	—	—	—

Table 1: Amount of parallel sentences and amount of left and right tokens in the corpora crawled from the Internet.

2.4 Crawled corpora

This section provides the specific details of the different corpora we have crawled during the second half of the project. The usefulness of these parallel corpora for the project has been indirectly evaluated by measuring the quality of the NMT systems trained on them. Table 1 provides, for each corpus, the number of parallel sentences and tokens in each language. Deliverable *D1.5 Final release of project data* provides pointers for downloading these corpora.

2.4.1 English–Hausa

We identified 72,776 English–Hausa parallel sentence pairs from the data available in the Internet Archive. However, the original crawled dataset contained many duplicate pairs. After deduplication, the released dataset contains 19,904 sentence pairs.

Complementary to the parallel data, we also used Internet Archive to obtain monolingual data in Hausa. In total, the noisy dataset consisted of 6,480,327 sentences. After cleaning and deduplication, we collected a dataset of 1,163,513 sentences.

In addition to the GoURMET releases, we prepared extra training and test data for the WMT21 Hausa↔English news translation task. Our contribution to the training data consisted of 5937 parallel sentences crawled and aligned from the President of Iran's website. For the test data we prepared both a development and a test set consisted of newly translated data. For each of these we selected 1000 sentences of Hausa and English news, and arranged for professional translation into the other language.

2.4.2 English–Igbo

The parallel data for English–Igbo was also obtained from the Internet Archive. After filtering and deduplication, the corpus consists of 29,969 sentences pairs coming from a total of 174 websites.

We also crawled an Igbo monolingual corpus from the Internet Archive. After cleaning a deduplication we obtained a monolingual corpus with 350,062 sentences.

2.4.3 English–Pashto

English–Pashto parallel data and Pashto monolingual data were obtained by two means: directly crawling websites likely to contain parallel data, and crawling the top-level domain of Afganistan (`.af`), where Pashto is an official language. As regards the first method, a total of 427 websites were partially crawled during three weeks following this strategy, from which only 50 provided any English–Pashto parallel data. As regards crawling the Afghanistan top-level domain, 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, 138 of them were identified to contain any English–Pashto parallel data.

The bilingual lexicon for document alignment was built on the concatenation of the following parallel corpora: GNOME, KDE4,[7] Tatoeba, TED2020,[8] Wikimedia,[9] and Ubuntu.

After document and segment alignment, 219,061 unique segment pairs were extracted. For the Bicleaner model, both the regressor and the language models were trained on the same datasets as used for building the bilingual lexicon mentioned in the previous paragraph. The threshold for the score provided by the regressor was set to 0.325. In addition, CLD3 language identifier was applied to both the right and the left sides of the parallel corpus to confirm the languages in every pair of segments.

As a by-product of the crawling, a Pashto monolingual corpus containing 381,951 sentences was obtained. This corpus is the result of cleaning a larger one. Language detection was performed using CLD3.

2.4.4 English–Yoruba

The first attempt to obtain parallel corpora for English–Yoruba was crawling a collection of 606 websites extracts from Wikidumps and the top level domain from the three countries in which Yoruba is an official language. Unfortunately, the Yoruba content was nearly non-existent. For this reason a manual search was done. After discarding websites with translation plugins, 14 websites with content in Yoruba and English were selected. The process of crawling and alignment was carried out with the tool Bitextor. Every website was crawled for a maximum of 72 hours.

Document alignment was performed using a bilingual dictionary automatically built by Bitextor from the following corpora available at OPUS: wikimedia,[10] YW300,[11] CCAligned,[12] GNOME,[13] Mozilla-I10n,[14] XLEnt,[15] GlobalVoices,[16] Tatoeba[17] and Ubuntu.[18]

[7] http://opus.nlpl.eu/KDE4-v2.php
[8] https://opus.nlpl.eu/TED2020-v1.php
[9] https://opus.nlpl.eu/wikimedia-v20190628.php
[10] https://opus.nlpl.eu/wikimedia-v20210402.php
[11] https://opus.nlpl.eu/JW300-v1c.php
[12] https://opus.nlpl.eu/CCAligned-v1.php
[13] https://opus.nlpl.eu/GNOME-v1.php
[14] https://opus.nlpl.eu/Mozilla-I10n-v1.php
[15] https://opus.nlpl.eu/XLEnt-v1.1.php
[16] https://opus.nlpl.eu/GlobalVoices-v2018q4.php
[17] https://opus.nlpl.eu/Tatoeba-v2021-07-22.php
[18] https://opus.nlpl.eu/Ubuntu-v14.10.php

Bicleaner was used to filter the raw corpus obtained by Bitextor (569 pairs of segments); bicleaner model was trained by using all the corpora available at OPUS listed before as a source of correctly-aligned segments. The result for the parallel corpus was only 113 sentences pairs.

The monolingual corpus contains 50 988 sentences. This corpus is the result of cleaning a larger one. Language detection was performed using CLD2. Sentences containing only URLs were removed.

2.4.5 English–Burmese

We crawled a total of 901 websites extracted from Wikidumps, Langstats and the top level domain. At the time of crawling, some government and university websites were down due to the political situation in Myanmar, and most of the remaining websites had content only in English or offered Burmese in headers, but not in content.

As the use of a bilingual lexicon for document and segments alignment led to wrong alignments due to the fact that Burmese does not consistently use blank spaces to delimit the words, we used machine translation for alignment. In particular, we used the systems described in *D5.5 – GoURMET Final progress report on integration* for English–Burmese. With machine translation, 43 946 segment pairs were obtained.

Parallel corpus cleaning was performed with Bicleaner-AI, which was trained on the ALT corpus.[19] Using a threshold of 0.2 and removing duplicates, the corpus finally obtained contains 1 033 parallel sentences.

In addition to the parallel corpora, a Burmese monolingual corpus with 601 048 sentences was obtained. This corpus is the result of removing duplicates and applying language detection with CLD3.

2.4.6 English–Macedonian

We crawled 330 websites from Common Crawl, from which 16 were not available at the time of crawling. 43 of these websites contained English–Macedonian parallel data, while 60 of them contained ony monolingual data in Macedonian. From the top-level domain, we crawled 365 additional websites, from which 53 were identified to contain parallel data, and 318 contained monolingual data in Macedonian. In total, 96 websites contained English–Macedonian parallel data and 378 websites contained monolingual data in Macedonian.

The bilingual lexicon for document alignment was built on the concatenation of the following parallel corpora: EUbookshop,[20] GNOME,[21] JW300,[22] KDE4,[23] OpenSubtitles,[24] SETIMES,[25] and Ubuntu.[26]

[19] https://www2.nict.go.jp/astrec-att/member/mutiyama/ALT/
[20] http://opus.nlpl.eu/EUbookshop-v2.php
[21] http://opus.nlpl.eu/GNOME.php
[22] http://opus.nlpl.eu/JW300-v1.php
[23] http://opus.nlpl.eu/KDE4-v2.php
[24] http://opus.nlpl.eu/OpenSubtitles-v2018.php
[25] http://opus.nlpl.eu/SETIMES-v2.php
[26] http://opus.nlpl.eu/Ubuntu.php

During document alignment, a total of 180,520 pairs of documents were obtained, from which 2,051,678 segment pairs were extracted. For the Bicleaner model, the regressor was trained on the parallel corpus GlobalVoices2015, available at OPUS. The threshold for the score provided by the regressor was set to 0.68. Bicleaner's character-level language model was trained on the same corpora used to build the bilingual lexicons and the threshold was set to 0.5. The resulting parallel corpus consisted of 71,506 segment pairs.

As a by-product of the crawling, a Macedonian monolingual corpus containing 1,153,565 sentences was obtained. This corpus is the result of cleaning a larger one. Language detection was performed using CLD3. Additional cleaning was performed using the same language model trained for cleaning parallel data.

2.4.7 English–Tigrinya

Tigrinya is a language with especially scarce resources. For this reason, crawling English–Tigrinya data required some additional manual work when compared to the process followed for other language pairs. A small list with about 10 news sites containing articles in Tigrinya was manually created and crawled using Bitextor. After crawling, only 5 of them were found to be productive in texts in Tigrinya. Intensive crawling was applied for two weeks. Unfortunately, no parallel data was spotted in the data crawled. However, about 197,000 unique segments in Tigrinya were obtained from crawling. After additional language identification using CLD3, 152,554 segments in Tigrinya were finally obtained. It is worth noting that this monolingual corpus can be considered domain specific, as all the text used to build it were downloaded from news websites.

2.5 Monolingual News Crawl

The monolingual news crawl contains text gathered from online news sources from around the world using their RSS feeds, and we release an update each year. The corpus is used in the WMT[27] shared tasks as the main source of monolingual data. The 2021 update (released in January 2022) contains 265 million lines, and 5.4 billion running words. The total size of the corpus is 1.7 billion lines, in 59 languages. In the GoURMET project we have focused on adding sources from India and Africa to the news crawl.

3 Processing of BBC and DW data dumps

Following the approach already described in *D1.2 Initial progress report on data gathering and augmentation*, we processed data dumps provided by the user partners for the languages addressed during the second half of the project. Data dumps are compilations of documents containing, each of them, a piece of news that has been published by BBC or DW in their respective websites. They have been processed to obtain test corpora and development sets. It is worth noting that, in order to be able to freely distribute our NMT models, BBC data cannot be used for training, since BBC has restrictions on derivative work. In the case of BBC, they provided the data dumps already compiled in JSONL format. For DW, we crawled their website by following the `sitemap.xml` file and built a similar JSONL document. Table 2 reports the number of documents in these data dumps.

[27] http://www.statmt.org/wmt22

Language	# documents BBC	# documents DW
Burmese	11,899	—
English	270,107	186,501
Hausa	18,049	30,731
Igbo	4,767	—
Macedonian	—	55,380
Pashto	16,398	29,649
Tigrinya	6,308	—
Turkish	30,201	—
Urdu	29,547	—
Yoruba	8,268	—

Table 2: Number of documents in the monolingual data dumps provided by BBC and DW for the second round of languages.

The process carried out on the data dumps aimed at identifying pairs of segments that are parallel. Given the fact that they would be used for testing and development, the quality of the result of this process was critical. The approach followed consisted in:

1. *Identifying parallel documents*: We exploited the fact that the images included in the pieces of news were independent of the language in which they are written and aligned documents containing a very similar set of images.

2. *Extracting promising pairs of segments*: Pairs of documents were segmented and, for each pair, the set of all possible segment pairs were produced.[28] From the collection of aligned segments, all the segments written in the language other than English were machine-translated with Google Translate.[29] The English side of each segment was then compared to the machine-translated counterpart, and segment pairs were ranked using the chrF2++ (Popović, 2017) translation quality metric.

3. *Human validation* The 4,000 segment pairs in the highest part of the ranking were conveniently presented to human validators using a form in which each validator could mark if a pair of segments was parallel or not.

The sizes of the resulting collections of validated parallel segments is shown in Table 3; they were later split into test and development sets, except for Igbo and Tigrinya.

4 Data augmentation

As we did for training the MT systems during the first half of the project, we have applied some data augmentation techniques to generate synthetic parallel corpora to be used for training. The

[28] Segments shorter than 5 words were discarded in this process.

[29] Google Translate was chosen because it covers all except for Tigrinya. For this specific language, we used the Amharic–English translation direction instead, with the hope that, given the similarity between Amharic and Tigrinya, the resulting translations could be of enough quality for our purpose.

Language	# validated segment pairs
Burmese	2,000
Hausa	2,000
Igbo	260
Macedonian	2,000
Pashto	3,165
Tigrinya	1,200
Turkish	2,636
Urdu	2,636
Yoruba	1112

Table 3: Number human-validated segment pairs extracted from the alignments between data dumps.

most common approaches are back-translation (Sennrich et al., 2016b), which leverages target monolingual corpora, and iterative back-translation (Hoang et al., 2018), which simultaneously leverages monolingual data in the two languages involved in the translation. In addition to these approaches, we have applied other data augmentation approaches developed within the GoURMET project, which are described in this section, to make the most of the available resources. We refer the reader to Appendix A of deliverable *D5.5 Final report on integration* for the specific details as regards the data augmentation techniques used for building each translation model.

4.1 Multi-task learning for data augmentation

This work has been published in the 2021 Conference on Empirical Methods in Natural Language Processing (Sánchez-Cartagena et al., 2021).

Data augmentation (Feng et al., 2021) is formalized by many authors as a solution to a data distribution mismatch problem (Wang et al., 2018; Wei et al., 2020): the data distribution of the sentence pairs observed in the training corpus differs from the true data distribution, and hence the system should be trained on an augmented training set which, hopefully, will be more representative of the true data distribution. In this way, the trained system is less likely to face totally out-of-distribution data when translating.

However, the method described next follows a completely different approach for data augmentation. Inspired by one-to-many multilingual NMT systems (Johnson et al., 2017), which are able to translate from a single language into many target languages, and that learn richer encoder representations (Dong et al., 2015), we devised a data-augmentation strategy that is applied within a multi-task learning framework. In this new approach, we can apply naïve transformations to the training samples (such as reversing the order of the target words) and generate additional parallel sentences that, despite being completely unlikely under the data distribution, are able to improve the quality of the resulting NMT system because, during training, they expose the network to new situations where the target-language context is not sufficient to achieve a low loss error.

The following two sections summarizes the approach and the main results obtained, for further details we refer the reader to the paper cited above.

Task	Lang.	Synthetic training sample
original training sample	source	Es gibt andere Möglichkeiten , die Pyramide zu durchbrechen .
	target	There 's other ways of breaking the pyramid .
swap	target	There . other ways of breaking pyramid 's the
token	target	There 's other UNK of UNK UNK UNK .
source	target	Es gibt andere Möglichkeiten , die Pyramide zu durchbrechen .
reverse	target	. pyramid the breaking of ways other 's There
mono	target	's There other ways the pyramid of breaking .
replace	source	Es gibt aufzurüsten kalt , Schach Spezialwissen zu durchbrechen .
	target	There 's arming cold of breaking chess specialties .

Table 4: A German–English, word-aligned training sample (first row) and the result of applying the transformations described in Sec. 4.1 using $\alpha = 0.5$ for those transformations controlled by this hyperparameter. Words modified by each transformation are coloured; for *swap* and *replace*, a different colour identifies each pair of words that are either swapped or replaced together, respectively.

Approach. The approach consists of using a vanilla NMT system —in the experiments reported in our paper (Sánchez-Cartagena et al., 2021) and summarised below, it is a transformer system as defined by Vaswani et al. (2017)— where all (main and auxiliary) tasks share the encoder and the decoder. In order to avoid harmful interferences by the out-of-distribution target data generated for the auxiliary tasks, we add a task-specific artificial token to the source sentence to constrain the kind of output to be produced (Sennrich et al., 2016a; Johnson et al., 2017), much like in multilingual NMT. For each auxiliary task, we append a synthetic corpus of the same size to the original training data, which is obtained by applying a transformation to each original pair of sentences. In almost all the tasks, the source sentence is left unchanged while the target sentence is substantially modified.

The transformations to be applied are designed to force the systems towards learning a richer encoder representation and trust less the target prefix when generating a new word. What follows is a brief summary of them (see Table 4), where α is a hyperparameter that determines the proportion of target words affected by the transformation:

swap: Pairs of random target words are swapped until only $(1-\alpha)\cdot t$ words remain in their original position.

token: $\alpha \cdot t$ random target words are replaced by a special (UNK) token (Xie et al., 2017).

source: The target sentence becomes a copy of the source sentence.

reverse: The order of the words in the target sentence is reversed.

mono: Target words are reordered so as to make the alignment between source and target words monotonic.

replace: $\alpha \cdot t$ source–target aligned words are selected at random and replaced by random entries in a bilingual lexicon obtained from the training corpus.

Summary of results. We conducted experiments for the translation from English to German, Hebrew and Vietnamese, and for the translation in the reverse direction, using corpora commonly used for evaluating data augmentation techniques in low-resource scenarios. We evaluated the effect of using each auxiliary task in isolation, as well as the combination of the best performing ones. We also evaluated two strong data augmentation methods that aim at extending the support of the empirical data distribution by replacing some words by random samples from the vocabulary.

The results show that our approach consistently outperforms the baseline system in all language pairs and translation directions. In general, the auxiliary tasks *reverse* (translation into the target language but in the reverse order) and *replace* (random replacement of target words and the source words they are aligned with) are the best performing ones. *swap* (random swapping of words) and *source* (copying the source sentence) often perform worse than the former tasks, which suggests that a non-systematic word order or a completely different vocabulary in the target could negatively influence the main task.

Interestingly, using the three best auxiliary tasks together further improves the performance, achieving the best results in all translation tasks with BLEU scores between 1.1 and 1.9 points over the baseline. This suggests that different auxiliary tasks affect the encoder in different ways and are somehow complementary. Additional experiments combining our data augmentation approach with back-translation, show that both approaches are complementary and bring improvements as compared to the baseline systems.

Finally, we performed a study of the performance of the NMT systems trained with our data augmentation approached under domain shift. Theses experiments we carried out using texts in the IT, medical and law domains, and show that the systems trained with our approach are more robust under domain shift and tend to hallucinate —produce completely inadequate translations not related to the source sentence to be translated— less.

4.2 Improving the translation of out-of-vocabulary words using bilingual lexicon induction

The work described in this section is in progress with a PhD student at the University of Edinburgh, Jonas Waldendorf.

An often overlooked difficulty of low-resource neural-machine translation (NMT) is the ability of models to correctly predict translations of words that are out of vocabulary (OOV). English–Pashto is one such low-resource language pair for which the translation of OOV words is of particular interest. One reason is that Pashto is a morphologically rich language and hence the probability of observing specific surface forms of a given word is lower. Another key issue that is also common among other low-resource languages is the different distribution of content between the training and test data. This is due to the parallel data containing a significant amount of content from specifics domains, such as IT and religious texts, which is not reflected in common downstream tasks for NMT systems such as the translation of news articles. Finally, the low amount of availability of parallel data inherent in the task means that naturally a smaller vocabulary is covered. Due to these factors, it is important for English–Pashto, as well as for low-resource language pairs in general,

to improve the translation of OOV words; improving the translation of these words is integral to providing good translations of the sentences seen at inference time.

Incorporating monolingual target side data by generating synthetic source-side data has been shown to improve the overall performance of low-resource NMT systems in terms of automatic evaluation metrics such as BLEU or chrF. However, an important benefit of incorporating monolingual data is the significant increase in the amount of vocabulary items that are observed during training, the effects of which can only be seen when evaluating NMT predictions at the single OOV word level. Back-translation (BT) (Sennrich et al., 2016c) is the most frequently used data augmentation technique for generating synthetic source-side sentences. However, work in the field of domain adaptation has shown that word-for-word back-translation using bilingual dictionaries extracted from bilingual word embeddings is a suitable alternative to BT when specifically targeting improved translation of OOV words (Hu et al., 2019a). Unlike the existing dictionary-based techniques in domain adaptation (Hu et al., 2019a; Huck et al., 2019), we apply this technique to the realistic low resource scenario of training English–Pashto NMT systems with a less distinct shift in content.

Søgaard et al. (2018) and Ormazabal et al. (2019) demonstrate that joint training leads to more isomorphic bilingual word embedding spaces for linguistically distinct languages. Based on this we propose a new approach of using a joint training methodology (Luong et al., 2015) to anchor an embedding space whilst also training on monolingual data, in addition to incorporating sub-word information. We contrast this to a frequently used approach of independently training two sets of word embeddings before mapping them into a shared space using a seed dictionary (Conneau et al., 2017). The main contributions of this work are as follows: adapting the word-for-word BT methodology from high-resource domain to English–Pashto, a realistic low-resource translation scenario, and proposing an extension to the joint training methodology which incorporates monolingual data.

4.3 Exploring diversity in back translation

This work has been published in the 3rd Workshop on Deep Learning for Low-Resource NLP (Burchell et al., 2022).

The data augmentation technique of back translation (BT) is used in nearly every current NMT system to reach improved performance. It involves creating a pseudo-parallel dataset by translating target-side monolingual data into the source language using a secondary NMT system (Sennrich et al., 2016c). In this way, it enables the incorporation of monolingual data into the NMT system. Whilst adding data in this way helps nearly all language pairs, it is particularly important for low-resource NMT where parallel data is scarce by definition.

Because of its ubiquity, there has been extensive research into how to improve BT (Burlot and Yvon, 2018; Hoang et al., 2018; Fadaee and Monz, 2018; Caswell et al., 2019), especially in ways which increase the 'diversity' of the back-translated data (Edunov et al., 2018; Soto et al., 2020). Previous work (Ott et al., 2018; Vanmassenhove et al., 2019) has found that machine translations lack the diversity of human productions. This is because most translation systems use some form of mode-seeking decoding (a.k.a. maximum a posteriori decoding, or MAP decoding), meaning that they will always favour the most probable output. Edunov et al. (2018) and Soto et al. (2020) argue that this makes standard BT data worse training data since it lacks 'richness' or diversity.

Despite the focus on increasing diversity in BT, what 'diversity' actually is in the context of NMT

training data is ill-defined. In fact, Tevet and Berant (2021) point out that there is no standard metric for measuring diversity. Most previous work uses the BLEU score between candidate sentences or another n-gram based metric to estimate similarity (Zhu et al., 2018; Hu et al., 2019b; Shu et al., 2019). However, such metrics mostly measure changes in the vocabulary or spelling. Because of this, they are likely to be less sensitive to other kinds of variety such as changes in structure.

We argue that quantifying 'diversity' using n-gram based metrics alone is insufficient. Instead, we split diversity into two aspects: variety in the word choice and spelling, and variety in structure. We call these aspects *lexical diversity* and *syntactic diversity*, respectively. Here, we follow recent work in the natural language generation (e.g. Iyyer et al., 2018; Huang and Chang, 2021; Hosking and Lapata, 2021) which explicitly models the meaning and form of the input separately. Of course, there are likely more kinds of diversity than this, but this definition provides a common-sense framework to extend our understanding of the concept. To our knowledge, no other previous work has attempted to isolate and automatically measure syntactic and lexical diversity.

Building from our definition, we introduce novel metrics aimed at measuring lexical and syntactic diversity separately. We then carry out an empirical study into what effect training data with these two kinds of diversity has on final NMT performance in the context of low-resource machine translation. We do this by creating BT datasets using different generation methods and measuring their diversity. We then evaluate what impact different aspects of diversity have on final model performance. We find that a high level of diversity is beneficial for final NMT performance, though lexical diversity seems more important than syntactic diversity. Importantly though there are limits to both; the data should not be so 'diverse' that it affects the accuracy of the parallel data.

We summarize our contributions as follows:

- We put forward a more nuanced definition of 'diversity' in NMT training data, splitting it into *lexical diversity* and *syntactic diversity*. We present two novel metrics for measuring these different aspects of diversity.
- We carry out empirical analysis into the effect of these types of diversity on final NMT model performance for low-resource English↔Turkish and mid-resource English↔Icelandic.
- We find that nucleus sampling is the highest-performing method of generating BT, and it combines both lexical and syntactic diversity.

5 Conclusion

This deliverable has reported the work conducted within WP1 on data gathering and data augmentation during the second half of the project. We crawled data from the web to obtained training resources; we faced some difficulties due to the scarceness of parallel texts for some languages, as well as the lack of consistent word segmentation for one of them (Burmese). We have also processed data dumps provided by the user partners to obtain high-quality in-domain development and test corpora. Finally, we have conducted research on exploring diversity in back-translation, on the translation of out-of-vocabulary words using bilingual lexicon induction, and on a multi-task learning framework for integrating synthetic training samples generated by applying simple transformation to real ones.

A summary of the research outcomes of WP1 follows:

Corpora. We have released all the corpora we have crawled from the web. Deliverable *D1.5 Final release of project data* provides the URLs from which they can be downloaded.

Publications. The following research papers resulted from the work described in this deliverable:

- *Rethinking data augmentation for low-resource neural machine translation: a multi-task learning approach*, Sánchez-Cartagena et al. (2021).
- *Findings of the 2021 Conference on Machine Translation (WMT21)*, Akhbardeh et al. (2021).
- *Exploring diversity in back translation for low resource machine translation*, Burchell et al. (2022).

Software. The following is a list of free/open-source software we have released or contributed to:

- MTL-DA (https://github.com/vitaka/mtl-da). Scripts for training machine translation systems using different data augmentation techniques in the target language.
- Diversity in back-translation. Code publicly available at github.com/laurieburchell/exploring-diversity-bt.
- Bitextor (https://github.com/bitextor/bitextor). Contributed to. It is the most widely-used tool to automatically harvest bilingual corpora from multilingual websites.
- Bicleaner (https://github.com/bitextor/bicleaner). Contributed to. It is a tool for the detection of noisy sentence pairs in a parallel corpus.

References

Akhbardeh, F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., Costa-jussa, M. R., España-Bonet, C., Fan, A., Federmann, C., Freitag, M., Graham, Y., Grundkiewicz, R., Haddow, B., Harter, L., Heafield, K., Homan, C., Huck, M., Amponsah-Kaakyire, K., Kasai, J., Khashabi, D., Knight, K., Kocmi, T., Koehn, P., Lourie, N., Monz, C., Morishita, M., Nagata, M., Nagesh, A., Nakazawa, T., Negri, M., Pal, S., Tapo, A. A., Turchi, M., Vydrin, V., and Zampieri, M. (2021). Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Burchell, L., Birch, A., and Heafield, K. (2022). Exploring diversity in back translation for low-resource machine translation. In *Deeplo*.

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

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

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Dong, D., Wu, H., He, W., Yu, D., and Wang, H. (2015). Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China.

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

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(1):77–86.

Esplà-Gomis, M., Sánchez-Cartagena, V. M., Zaragoza-Bernabeu, J., and Sánchez-Martínez, F. (2020). Bicleaner at wmt 2020: Universitat d'alacant-prompsit's submission to the parallel corpus filtering shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 950–956, Online. Association for Computational Linguistics.

Fadaee, M. and Monz, C. (2018). Back-translation sampling by targeting difficult words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 436–446, Brussels, Belgium. Association for Computational Linguistics.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. (2021). A survey of data augmentation approaches for NLP.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1):3–42.

Hoang, V. C. D., Koehn, P., Haffari, G., and Cohn, T. (2018). Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Hosking, T. and Lapata, M. (2021). Factorising meaning and form for intent-preserving paraphrasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1418, Online. Association for Computational Linguistics.

Hu, J., Xia, M., Neubig, G., and Carbonell, J. (2019a). Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Hu, J. E., Rudinger, R., Post, M., and Van Durme, B. (2019b). Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6521–6528.

Huang, K.-H. and Chang, K.-W. (2021). Generating syntactically controlled paraphrases without using annotated parallel pairs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1022–1033, Online. Association for Computational Linguistics.

Huck, M., Hangya, V., and Fraser, A. (2019). Better OOV Translation with Bilingual Terminology Mining. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5809–5815, Florence, Italy. Association for Computational Linguistics.

Iyyer, M., Wieting, J., Gimpel, K., and Zettlemoyer, L. (2018). Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, 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.

Luong, T., Pham, H., and Manning, C. D. (2015). Bilingual Word Representations with Monolingual Quality in Mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, Colorado. Association for Computational Linguistics.

Ormazabal, A., Artetxe, M., Labaka, G., Soroa, A., and Agirre, E. (2019). Analyzing the Limitations of Cross-lingual Word Embedding Mappings. *arXiv:1906.05407 [cs]*. arXiv: 1906.05407.

Ott, M., Auli, M., Grangier, D., and Ranzato, M. (2018). Analyzing uncertainty in neural machine translation. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965. PMLR.

Popović, M. (2017). chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Sánchez-Cartagena, V. M., Bañón, M., Ortiz-Rojas, S., and Ramírez-Sánchez, G. (2018). Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium.

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

Schwenk, H. (2018). Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne, Australia. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany.

Sennrich, R., Haddow, B., and Birch, A. (2016c). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Sennrich, R. and Volk, M. (2011). Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Shu, R., Nakayama, H., and Cho, K. (2019). Generating diverse translations with sentence codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1827, Florence, Italy. Association for Computational Linguistics.

Soto, X., Shterionov, D., Poncelas, A., and Way, A. (2020). Selecting backtranslated data from multiple sources for improved neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3898–3908, Online. Association for Computational Linguistics.

Søgaard, A., Ruder, S., and Vulić, I. (2018). On the Limitations of Unsupervised Bilingual Dictionary Induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Tevet, G. and Berant, J. (2021). Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Vanmassenhove, E., Shterionov, D., and Way, A. (2019). Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA, USA.

Wang, X., Pham, H., Dai, Z., and Neubig, G. (2018). SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium.

Wei, X., Yu, H., Hu, Y., Weng, R., Xing, L., and Luo, W. (2020). Uncertainty-aware semantic augmentation for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2724–2735, Online.

Xie, Z., Wang, S. I., Li, J., Lévy, D., Nie, A., Jurafsky, D., and Ng, A. Y. (2017). Data noising as smoothing in neural network language models. In *Proceedings of the 5th International Conference on Learning Representations, ICLR 2017*, Toulon, France.

Zaragoza-Bernabeu, J., Bañón, M., Ramírez-Sánchez, G., and Ortiz-Rojas, S. (2022). Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., and Yu, Y. (2018). Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D1.4 Final report on data gathering and augmentation