Survey of Low-Resource Machine Translation

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, Alexandra Birch

▶ **To cite this version:**

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, Alexandra Birch. Survey of Low-Resource Machine Translation. 2021. hal-03479757v1

HAL Id: hal-03479757

https://inria.hal.science/hal-03479757v1

Preprint submitted on 14 Dec 2021 (v1), last revised 1 Nov 2022 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Survey of Low-Resource Machine Translation

Barry Haddow
University of Edinburgh

Rachel Bawden
Inria, France

Antonio Valerio Miceli Barone
University of Edinburgh

Jindřich Helcl
University of Edinburgh

Alexandra Birch*
University of Edinburgh

We present a survey covering the state of the art in low-resource machine translation. There are currently around 7000 languages spoken in the world and almost all language pairs lack significant resources for training machine translation models. There has been increasing interest in research addressing the challenge of producing useful translation models when very little translated training data is available. We present a high level summary of this topical field and provide an overview of best practices.

* 10 Crichton Street, EH89AB, UK, a.birch@ed.ac.uk

© 2016 Association for Computational Linguistics

1. Introduction

Current machine translation (MT) systems have reached the stage where researchers are now debating whether or not they can rival human translators in performance (Hassan et al. 2018; Läubli, Sennrich, and Volk 2018; Toral et al. 2018; Popel et al. 2020). However these MT systems are typically trained on data sets consisting of tens or even hundreds of millions of parallel sentences. Data sets of this magnitude are only available for a small number of highly resourced language pairs (typically English paired with some European languages, Arabic and Chinese). The reality is that for the vast majority of language pairs in the world the amount of data available is extremely limited, or simply non-existent.

Resource Level	Language Pair	Speakers	Parallel Sentences
High	English–French	267M	280M
Medium	English–Myanmar	30M	0.7M
Low	English–Fon	2M	0.035M

Table 1: Examples of language pairs with different levels of resources. The number of speakers is obtained from Ethnologue.

In Table 1 we illustrate the differences between what can be considered to be high, medium and low resource language pairs, based on the number of first language speakers[1] and the number of parallel sentences in Opus (Tiedemann 2012), the largest collection of publicly available translated data. Although there is a correlation between speaker numbers and size of parallel data, there are many exceptions where either widely spoken languages have little parallel data, or languages with very small speaker numbers are richly resourced (mainly European languages). For one of the world's most spoken languages, French, there are nearly 280 million parallel sentences of English–French in OPUS. However when we search for English–Myanmar, we find only around 700,000 parallel sentences, despite Myanmar having a large number of speakers. If we consider the Nigerian language Fon, which still has over 2 million speakers, then we find far fewer parallel sentences, only 35,000. Developing MT systems for these three language pairs will require very different techniques.

The lack of parallel data for most language pairs is only one part of the problem. Existing data is often noisy or from a very specialised domain. Looking at the resources that are available for English–Fon, we see that the only corpus available is extracted from Jehovah's Witness publications (Agić and Vulić 2019, JW300). For many language pairs, the only corpora available are those derived from religious sources (e.g. Bible, Koran) or from IT localisation data (e.g. from open-source projects such as GNOME and Kubuntu). Not only is such data likely to be in a very different domain from the text that we would like to translate, but such large-scale multilingual automatically extracted corpora are often of poor quality (Caswell et al. 2021) and this problem is worse for low-resource language pairs. This means that low-resource language pairs suffer from multiple compounding problems: lack of data, out-of-domain data and noisy data. Apart from data issues, they are also often understudied languages where it

1 from Ethnologue: https://www.ethnologue.com/

is difficult to access language speakers and experts, and even basic tools like language identification do not exist or are not reliable.

Perhaps in part due to all these challenges, there has been an increasing interest in the research community in exploring more diverse languages, and language pairs that do not include English. This survey paper presents a high-level summary of approaches to low-resource neural MT (NMT), which should be useful for those interested in this broad and rapidly evolving field. There are currently a number of other survey papers in related areas, for example a survey of monolingual data in low-resource NMT (Gibadullin et al. 2019) and a survey of multilingual NMT (Dabre, Chu, and Kunchukuttan 2020). There have also been two very recent surveys of low-resource MT, which have been written concurrently with this survey (Ranathunga et al. 2021; Wang et al. 2021). Our survey aims to provide the broadest coverage of existing research in the field and we also contribute an extensive overview of the tools and techniques validated across 18 low-resource shared tasks that ran between 2018 and 2021.

One of the challenges of surveying the literature on low-resource MT is how to define what a low-resource language pair is. This is hard, because "resourced-ness" is on a continuum and any criterion must be arbitrary. We also note that the definition of low-resource can change over time. We could crawl more parallel data, or we could find better ways of using related language data or monolingual data which means that some language pairs are no longer so resource-poor. We maintain that for research to be considered to be on "low-resource MT", there should be some way in which the research should either aim to understand the implications of the lack of data, or propose methods for overcoming the lack of data. We do not take a strict view of what to include in this survey though; if the authors consider that they are studying low-resource MT, then that is sufficient. We do feel however that it is important to distinguish between simulated low-resource settings (where a limited amount of data from otherwise high-resource language pairs is used) and genuinely low-resource languages (where additional difficulties apply). We also discuss some papers that do not explicitly consider low-resource MT but which present important techniques and we mention methods that we think have the potential to improve low-resource MT.

In Figure 1 we show how we structure the diverse research addressing low-resource MT, and this paper follows this structure. We start the survey by looking at work that aims to increase the amount and quality of parallel and monolingual data available for low-resource MT (Section 2). We then look at work that uses other types of data: monolingual data (Section 3), parallel data from other language pairs (Section 4), and other types of linguistic data (Section 5). Another avenue of important research is how to make better use of existing, limited resources through better training or modelling (Section 6). We conclude by looking at efforts in the community to build research capacity through shared tasks and language-specific collectives (Section 7), providing a practical summary of commonly used approaches and other techniques often used by top-performing shared task systems.

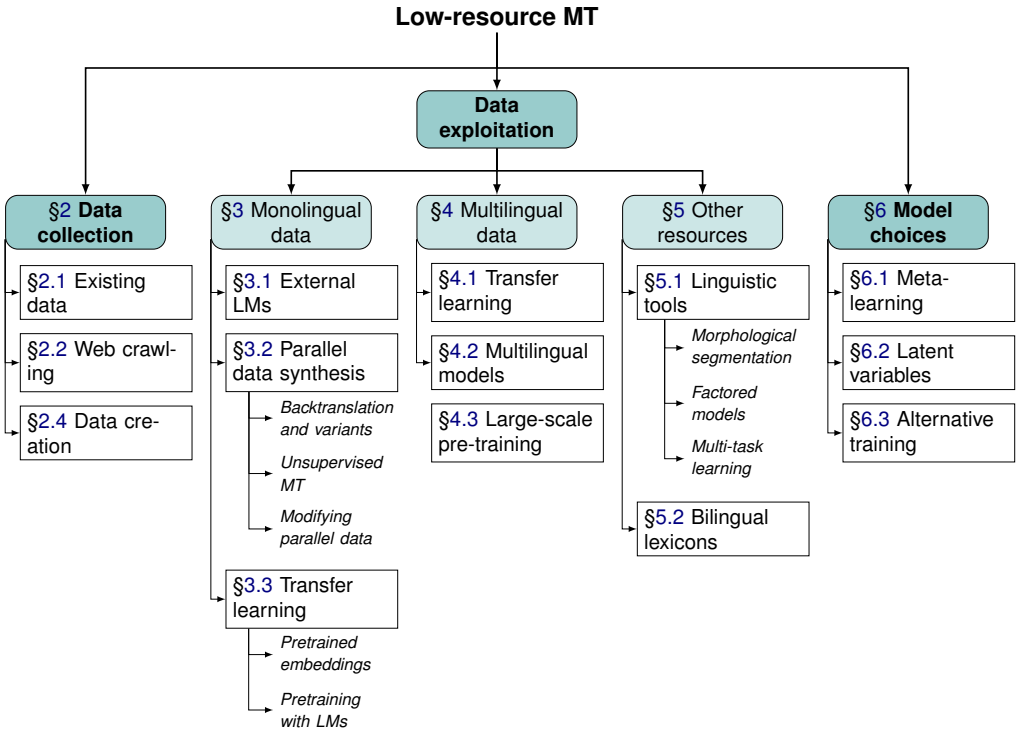

Figure 1: Overview of research covered in this survey.

2. Data Sources

The first consideration when applying data-driven MT to a new language pair is to determine what data resources are already available. In this section we discuss commonly used data sets and how to extract more data, especially parallel data, for low-resource languages. A recent case-study (Hasan et al. 2020) has shown how carefully targeted data gathering can lead to clear MT improvements in a low-resource language pair (in this case, Bengali–English). Data is arguably the most important factor in our success at modelling translation and we encourage readers to consider data set creation and curation as important areas for future work.

2.1 Searching Existing Data Sources

The largest range of freely available parallel data is found on Opus[2] (Tiedemann 2012), which hosts parallel corpora covering more than 500 different languages and variants. Opus collects contributions of parallel corpora from many sources in one convenient website, and provides tools for downloading corpora and metadata.

Monolingual corpora are also useful, although not nearly as valuable as parallel data. There have been a few efforts to extract text from the CommonCrawl[3] collection of websites, and these generally have broad language coverage. The first extraction effort was (Buck, Heafield, and van Ooyen 2014), although more recent efforts such as Oscar[4] (Ortiz Suárez, Sagot, and Romary 2019) and CC100 (Conneau et al. 2020) have focused on cleaning the data, and making it easier to access. A much smaller (but generally less noisy) corpus of monolingual news[5] is updated annually for the WMT shared tasks (Barrault et al. 2020), and currently covers 59 languages.

2.2 Web-crawling for Parallel Data

Once freely available sources of parallel data have been exhausted, one avenue for improving low-resource NMT is to obtain more parallel data by web-crawling. There is a large amount of translated text available on the web, ranging from small-scale tourism websites to large repositories of government documents. Identifying, extracting and sentence-aligning such texts is not straightforward, and researchers have considered many techniques for producing parallel corpora from web data. The links between source texts and their translations are rarely recorded in a consistent way, so techniques ranging from simple heuristics to neural sentence embedding methods are used to extract parallel documents and sentences.

Paracrawl,[6] a recent large-scale open-source crawling effort (Bañón et al. 2020) has mainly targeted European languages, only a few of which can be considered as low-resource, but it has created some releases for non-European low-resource language pairs, and the crawling pipeline is freely available. Related to this are other broad parallel data extraction efforts, where recently developed sentence-embedding based alignment methods (Artetxe and Schwenk 2019b,a) were used to extract large parallel corpora in many language pairs from Wikipedia (Schwenk et al. 2021) and Common-

2 http://opus.nlpl.eu/
3 https://commoncrawl.org/
4 https://oscar-corpus.com/
5 http://data.statmt.org/news-crawl/
6 https://www.paracrawl.eu/

Crawl (Schwenk et al. 2019). Similar techniques were used to create the largest parallel corpora of Indian languages, Samanantar (Ramesh et al. 2021).

Broad crawling and extraction efforts are good for gathering data from the "long tail" of small websites with parallel data, but they tend to be much more effective for high-resource language pairs, because there is more text in these languages, and routine translation of websites is more common for some languages. Focused crawling efforts, where the focus is on particular language pairs, or particular sources, can be more effective for boosting the available data for low-resource language pairs. Religious texts are often translated into many different languages, with the texts released under permissive licences to encourage dissemination, so these have been the focus of some recent crawling efforts, such as corpora from the Bible (Mayer and Cysouw 2014; Christodouloupoulos and Steedman 2015) or the publications of Jehovah's witnesses (Agić and Vulić 2019).[7] In India, government documents are often translated into many of the country's official languages, most of which would be considered low-resource for MT. Recent efforts (Haddow and Kirefu 2020; Siripragada et al. 2020; Philip et al. 2021) have been made to gather and align this data, including producing parallel corpora *between* different Indian languages, rather than the typical English-centric parallel corpora. The last of these works uses an iterative procedure, starting with an existing NMT system for alignment (Sennrich and Volk 2011), and proceeding through rounds of crawling, alignment and retraining, to produce parallel corpora for languages of India. Further language-specific data releases for low-resource languages, including not only crawled data from MT but annotated also data for other NLP tasks, were recently provided by the Lorelei project (Tracey et al. 2019), although this data is distributed under restrictive and sometimes costly LDC licences.

2.3 Low-resource Languages and Web-crawling

Large-scale crawling faces several problems when targeting low-resource language pairs. Typical crawling pipelines require several stages and make use of resources such as text preprocessors, bilingual dictionaries, sentence-embedding tools and translation systems, all of which may be unavailable or of poor quality for low-resource languages. Also, parallel sentences are so rare for low-resource languages that even a small false-positive rate will result in a crawled corpus that is mostly noise (e.g. sentences that are badly aligned, sentences in the wrong language, or fragments of html/javascript).

One of the first stages of any crawling effort is language identification, perhaps thought to be a solved problem with wide-coverage open-source toolkits such as CLD3.[8] However it has been noted (Caswell et al. 2020) that language identification can perform quite poorly on web-crawled corpora, especially on low-resource languages, where it is affected by class imbalance, similar languages, encoding problems and domain mismatches. Further down the crawling pipeline, common techniques for document alignment rely on the existence of translation systems (Uszkoreit et al. 2010; Sennrich and Volk 2011) or sentence embeddings (Artetxe and Schwenk 2019a), which again may not be of sufficient quality in low-resource languages, and so we often have to fall back on older, heuristic alignment techniques (Varga et al. 2005). The consequence is that the

7 A cleaned-up version of JW300 has recently been added to opus (https://opus.nlpl.eu/JW300.php)
8 https://github.com/google/cld3

resulting corpora are extremely noisy and require extensive filtering before they can be useful for NMT training.

Filtering of noisy corpora is itself an active area of research, and has been explored in recent shared tasks, which particularly emphasised low-resource settings (Koehn et al. 2019, 2020). In an earlier version of the task (Koehn et al. 2018), dual conditional cross-entropy (Junczys-Dowmunt 2018) was found to be very effective for English–German. However this method was less effective for language pairs that were low-resource and more distant. In the 2019 and 2020 editions of the task, all participants apply some heuristic filtering (e.g. based on language identification and length) and then strong submissions typically used a combination of embedding-based methods (such as LASER (Artetxe and Schwenk 2019b), GPT-2 (Radford et al. 2019) and YiSi (Lo 2019)) with feature-based systems such as zipporah (Xu and Koehn 2017) or bicleaner (Sánchez-Cartagena et al. 2018). Whilst the feature-based methods are much faster than sentence-embedding based methods, both types of methods require significant effort in transferring to a new language pair, especially if no pre-trained sentence embeddings or other models are available.

Barry: is there a citation for this?

The conclusion is that all crawled data sources should be treated with care, especially in low-resource settings as they will inevitably contain errors. A large-scale quality analysis (Caswell et al. 2021) of crawled data has highlighted that many contain incorrect language identification, non-parallel sentences, low quality texts, as well as offensive language, and these problems can be more acute in low-resource languages.

2.4 Other Data Sources

In addition to mining existing sources of translations, researchers have turned their attention to ways of creating new parallel data. One idea for doing this in a cost-effective manner is crowd-sourcing of translations. Post, Callison-Burch, and Osborne (2012) showed this method is effective in collecting a parallel corpus covering several languages of India. Pavlick et al. (2014) used crowd-sourcing to collect bilingual dictionaries covering a large selection of languages. Whilst not specifically aimed at MT, the Tatoeba[9] collection of crowd-sourced translations provides a useful resource with broad language coverage. An MT challenge set covering over 100 language pairs has been derived from Tatoeba (Tiedemann 2020).

Obtaining more training data is important, but we should not forget the role of standardised and reliable test sets in improving performance on low-resource translation. Important contributions have come from shared tasks, such as those organised by the WMT Conference on Machine Translation (Bojar et al. 2018; Barrault et al. 2019, 2020; Fraser 2020), the Workshop on Asian Translation (Nakazawa et al. 2018, 2019, 2020) and the Workshop on Technologies for MT of Low Resource Languages (LoResMT) (Karakanta et al. 2019; Ojha et al. 2020). In addition to the shared task test sets, the FLORES-101 benchmark (Goyal et al. 2021) covers a large number of low-resource languages with multi-parallel test sets, vastly expanding on the original FLORES release (Guzmán et al. 2019).

We summarise the data sets described in this section in Table 2.

9 https://tatoeba.org/

Corpus name	URL	Description
CC100	http://data.statmt.org/cc-100/	Monolingual data from CommonCrawl (100 languages).
Oscar	https://oscar-corpus.com/	Monolingual data from CommonCrawl (166 languages).
news-crawl	http://data.statmt.org/news-crawl/	Monolingual news text in 59 languages.
Opus	https://opus.nlpl.eu/	Collection of parallel corpora in 500+ languages, gathered from many sources.
WikiMatrix	https://bit.ly/3DrTjPo	Parallel corpora mined from wikipedia
CCMatrix	https://bit.ly/3Bin6rQ	Parallel corpora mined from Common-Crawl
Samanantar	https://indicnlp.ai4bharat.org/samanantar/	A parallel corpus of 11 languages of India paired with English
Bible	https://github.com/christos-c/bible-corpus	A parallel corpus of 100 languages extracted from the Bible
JW300	https://opus.nlpl.eu/JW300.php	A parallel corpus of 300 languages extracted from Jehovah's witness publications.
Tatoeba Challenge	https://bit.ly/3Drp36U	Test sets in over 100 language pairs.
WMT corpora	http://www.statmt.org/wmt21/	Test (and training) sets for many shared tasks.
WAT corpora	https://lotus.kuee.kyoto-u.ac.jp/WAT/	Test (and training) sets for many shared tasks.
FLORES-101	https://github.com/facebookresearch/flores	Test sets for 100 languages, paired with English
Pavlick dictionaries	https://bit.ly/3DgI0cu	Crowd-sourced bilingual dictionaries in many languages

Table 2: Summary of useful sources of monolingual, parallel and benchmarking data discussed in this section

3. Use of monolingual data

For low-resource language pairs, parallel text is, by definition, in short supply, and even applying the data collection strategies of Section 2 may not yield sufficient text to train high-quality MT models. However, monolingual text will almost always be more plentiful than parallel, and leveraging monolingual data has therefore been one of the most important and successful areas of research in low-resource MT.

In this section, we provide an overview of the main approaches used to exploit monolingual data in low-resource scenarios. We start by reviewing work on integrating external language models into NMT (Section 3.1), work largely inspired from the use of language models in statistical MT (SMT). We then discuss research on synthesising parallel data (Section 3.2), looking at the dominant approach of backtranslations and its variants, unsupervised MT and the modification of existing parallel data using language models. Finally, we turn to transfer learning (Section 3.3), whereby a model trained on monolingual data is used to initialise part or all of the NMT system, either through using pre-trained embeddings or through the initialising of model parameters from pre-trained language models. The use of monolingual data for low-resource has previously been surveyed by Gibadullin et al. (2019), who choose to categorise methods according to whether they are architecture-independent or architecture-dependent. This categorisation approximately aligns with our split into (i) approaches based on synthetic data and the integration of external LMs (architecture-independent), and (ii) those based on transfer learning (architecture-dependent).[10]

3.1 Integration of external language models

For SMT, monolingual data was normally incorporated into the system using a language model, in a formulation that can be traced back to the noisy channel (Brown et al. 1993). In early work on NMT, researchers drew inspiration from SMT, and several works have focused on integrating external language models into NMT models.

The first approach, proposed by Gülçehre et al. (2015), was to modify the scoring function of the MT model by either interpolating the probabilities from a language model with the translation probabilities (they call this *shallow fusion*) or integrating the hidden states from the language model within the decoder (they call this *deep fusion*). Importantly, they see improved scores for a range of scenarios, including a (simulated) low-resource language direction (Turkish→English), with best results achieved using deep fusion. Building on this, Stahlberg, Cross, and Stoyanov (2018) proposed *simple fusion* as an alternative method for including a pre-trained LM. In this case, the NMT model is trained from scratch with the fixed LM, offering it a chance to learn to be complementary to the LM. The result is improved translation performance as well as training efficiency, with experiments again on low-resource Turkish–English, as well as on larger data sets for Xhosa→English and Estonian→English.

The addition of a language model to the scoring function as in the works described above has the disadvantage of increasing the time necessary for decoding (as well as training). An alternative approach was proposed by Baziotis, Haddow, and Birch (2020), who aim to overcome this by using the language model as a regulariser during training – pushing the NMT probabilities to be more probable under the language model prior.

10 With the difference that unsupervised MT is architecture-dependent and we choose to discuss it in Section 3.2 on synthesising parallel data.

They see considerable gains in very low-resource settings (albeit simulated), using small data sets for Turkish–English and German–English.

3.2 Synthesising Parallel Data using Monolingual Data

One direction in which the use of monolingual data has been highly successful is in the production of synthetic parallel data. This is particularly important in low-resource settings when genuine parallel data is scarce. It has been the focus of a large body of research and has become the dominant approach to exploiting monolingual data due to the improvements it brings (particularly in the case of backtranslation as we shall see) and the potential for progress in the case of unsupervised MT. Both backtranslation and unsupervised MT belong to a category of approaches involving self-learning, which we discuss in Section 3.2.1. We then discuss an alternative method of synthesising parallel data in Section 3.1, which involves modifying existing parallel data using a language model learnt on monolingual data.

3.2.1 Self-learning: backtranslation and its variants. One of the most successful strategies for leveraging monolingual data has been the creation of synthetic parallel data through translating monolingual texts either using a heuristic strategy or an intermediately trained MT model. This results in parallel data where one side is human generated, and the other is automatically produced. We focus here on backtranslation and its variants, before exploring in the next section how unsupervised MT can be seen as an extension of this idea.

Backtranslation. Backtranslation corresponds to the scenario where target-side monolingual data is translated using an MT system to give corresponding synthetic source sentences, the idea being that it is particularly beneficial for the MT decoder to see well-formed sentences. Backtranslation was first introduced in SMT (Bertoldi and Federico 2009; Bojar and Tamchyna 2011), but since monolingual data could already be incorporated easily into SMT systems using language models, and because inference in SMT was quite slow, backtranslation was not widely used. For NMT however, it was discovered that backtranslation was a remarkably effective way of exploiting monolingual data (Sennrich, Haddow, and Birch 2016a), and it remains an important technique, for both low-resource MT and MT in general.

There has since been considerable interest in understanding and improving backtranslation. For example, Edunov et al. (2018a) showed that backtranslation improved performance even at a very large scale, but also that it provided improvements in (simulated) low-resource settings, where it was important to use beam-search rather than sampling to create backtranslations (the opposite situation to high-resource pairs).

Variants of backtranslation. Forward translation, where monolingual source data is translated into the target language, also exists but has received considerably less interest than backtranslation for low-resource MT, presumably because of the noise it introduces for the decoder. A related, and even simpler, technique, that of copying from target to source to create synthetic data, was introduced by Currey, Miceli Barone, and Heafield (2017). They showed that this was particularly useful in low-resource settings (tested on Turkish–English and Romanian–English) and hypothesise that it helped particularly with the translation of named entities.

Iterative backtranslation. For low-resource NMT, back-translation can be a particularly effective way of improving quality (Guzmán et al. 2019). However one possible issue is that the initial model used for translation (trained on available parallel data) is often of poor quality when parallel data is scarce, which inevitable leads to poor quality backtranslations. The logical way to address this is to perform *iterative backtranslation*, whereby intermediate models of increasing quality in both language directions are successively used to create synthetic parallel data for the next step. This has been successfully applied to low-resource settings by several authors (Hoang et al. 2018; Dandapat and Federmann. 2018; Bawden et al. 2019; Sánchez-Martínez et al. 2020), although successive iterations offer diminishing returns, and often two iterations are sufficient, as has been shown experimentally (Chen et al. 2020).

Other authors have sought to improve on iterative backtranslation by introducing a round-trip (i.e. autoencoder) objective for monolingual data, in other words performing backtranslation and forward translation implicitly during training. This was simultaneously proposed by Cheng et al. (2016) and He et al. (2016) and also by Zhang and Zong (2016) who also added forward translation. However, none of these authors applied their techniques to low-resource settings. In contrast, Niu, Xu, and Carpuat (2019) developed a method using Gumbel softmax to enable back-propagation through backtranslation: they tested in low-resource settings but achieved limited success.

Despite the large body of literature on applying back-translation and related techniques, and evidence that it works in low-resource NMT, there are few systematic experimental study of back-translation specifically for low-resource NMT, apart from (Xu et al. 2019), which appears to confirm the findings of (Edunov et al. 2018a) that sampling is best when there are reasonable amounts of data and beam search is better when data is very scarce.

3.2.2 Unsupervised MT. The goal of *unsupervised MT* is to learn a translation model without any parallel data, so this can be considered an extreme form of low-resource MT. The first unsupervised NMT models (Lample et al. 2018a; Artetxe et al. 2018) were typically trained in a two-phase process: a rough translation system is first created by aligning word embeddings across the two languages (e.g. using bilingual seed lexicons), and then several rounds of iterative backtranslation and denoising autoencoding are used to further train the system. Whilst this approach has been successfully applied to high-resource language pairs (by ignoring available parallel data) it has been shown to perform poorly on genuine low-resource language pairs (Guzmán et al. 2019; Marchisio, Duh, and Koehn 2020; Kim, Graça, and Ney 2020), mainly because the initial quality of the word embeddings and their cross-lingual alignments is poor (Edman, Toral, and van Noord 2020). The situation is somewhat improved using transfer learning from models trained on large amounts of monolingual data (Section 3.3), and some further gains can be achieved by adding a supervised training step with the limited parallel data (i.e. semi-supervised rather than unsupervised) (Bawden et al. 2019). However the performance remains limited, especially compared to high-resource language pairs.

These negative results have focused researchers' attention on making unsupervised MT work better for low-resource languages. Chronopoulou, Stojanovski, and Fraser (2021) improved the cross-lingual alignment of word embeddings in order to get better results on unsupervised Macedonian–English and Albanian–English. A separate line of work is concerned with using corpora from other languages to improve unsupervised NMT (see Section 4.2.2)

3.2.3 Modification of existing parallel data. Another way in which language models have been used to generate synthetic parallel data is to synthesise parallel examples from new ones by replacing certain words.[11] In translation, it is important to maintain the relation of translation between the two sentences in the parallel pair when modification of the pair occurs. There are to our knowledge few works so far in this area. Fadaee, Bisazza, and Monz (2017) explore data augmentation for MT for a simulated low-resource setting (using English–German). They rely on bi-LSTM language models to predict plausible but rare equivalents of words in sentences. They then substitute in the rare words and replace the aligned word in the corresponding parallel sentence with its translation (obtained through a look-up in an SMT phrase table). They see improved BLEU scores and find that it is a complementary technique to backtranslation. More recently, Arthaud, Bawden, and Birch (2021) apply a similar technique to improve the adaptation of a model to new vocabulary for the low-resource translation direction Gujarati→English. They use a BERT language model to select training sentences that provide the appropriate context to substitute new and unseen words in order to create new synthetic parallel training sentences. While their work explores the trade-off between specialising to the new vocabulary and maintaining overall translation quality, they show that the approach can improve the translation of new words more effectively following data augmentation.

3.3 Introducing Monolingual Data Using Transfer Learning

The third category of approaches we explore looks at transfer learning, by which we refer to techniques where a model trained using the monolingual data is used to initialise some or all of the NMT model. A related, but different idea, multilingual models, where the low-resource NMT model may be trained with the help of other (high-resource) languages will be considered in Section 4.

Pre-trained embeddings. When neural network methods were introduced to NLP, transfer learning meant using pre-trained word embeddings, such as word2vec (Mikolov et al. 2013) or GloVe (Pennington, Socher, and Manning 2014), to introduce knowledge from large unlabelled monolingual corpora into the model. The later introduction of the multilingual fastText embeddings (Bojanowski et al. 2017) meant that pre-trained embeddings could be tested with NMT (Di Gangi and Federico 2017; Neishi et al. 2017; Qi et al. 2018). Pre-trained word embeddings were also used in the first phase of unsupervised NMT training (Section 3.2.2). Of most interest for low-resource NMT was the study by Qi et al. (2018), who showed that pre-trained embeddings could be extremely effective in some low-resource settings.

Pre-trained language models. Another early method for transfer learning was to pre-train a language model, and then to use it to initialise either the encoder or the decoder or both (Ramachandran, Liu, and Le 2017). Although not MT *per se*, Junczys-Dowmunt et al. (2018b) applied this method to improve grammatical error correction, which they modelled as a low-resource MT task.

11 These techniques are inspired by data augmentation in computer vision, where it is much simpler to manipulate examples to create new ones (for example by flipping and cropping images) whilst preserving the example label. The difficulty in constructing synthetic examples in NLP in general is the fact that the modifying any of the discrete units of the sentence is likely to change the meaning or grammaticality of the sentence.

The pre-trained language model approach has been extended with new objective functions based on predicting masked words, trained on large amounts of monolingual data. Models such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019) have been shown to be very beneficial to natural language understanding tasks, and researchers have sought to apply related ideas to NMT. One of the blocking factors identified by Yang et al. (2020) in using models such as BERT for pre-training is the problem of catastrophic forgetting (Goodfellow et al. 2014). They propose a modification to the learning procedure involving a knowledge distillation strategy designed to retain the model's capacity to perform language modelling during translation. They achieve increased translation performance according to BLEU, although they do not test on low-resource languages.

Despite the success of ELMo and BERT in NLP, large-scale pre-training in NMT did not become popular until the success of the XLM (Conneau and Lample 2019), MASS (Song et al. 2019b) and mBART (Liu et al. 2020) models. These models allow transfer learning for NMT by initial training on large quantities of monolingual data in several languages, before fine-tuning on the languages of interest. Parallel data can also be incorporated into these pre-training approaches. Since they use data from several languages, they will be discussed in the context of multilingual models in Section 4.3.

4. Use of multilingual data

In the previous section we considered approaches that exploit monolingual corpora to compensate for the limited amount of parallel data available for low-resource language pairs. In this section we consider a different but related set of methods, which use additional data from different languages (i.e. in languages other than the language pair that we consider for translation). These multilingual approaches can be roughly divided into two categories: (i) transfer learning and (ii) multilingual models.

Transfer learning (Section 4.1) was introduced in Section 3.3 in the context of pre-trained language models. These methods involve using some or all of the parameters of a "parent" model to initialise the parameters of the "child" model. The idea of multilingual modelling (Section 4.2) is to train a system that is capable of translating between several different language pairs. This is relevant to low-resource MT, because low-resource language pairs included in a multilingual model may benefit from other languages used to train the model. Finally (Section 4.3), we consider more recent approaches to transfer learning, based on learning large pre-trained models from multilingual collections of monolingual and parallel data.

4.1 Transfer Learning

In the earliest form of multilingual transfer learning for NMT, a parent model is trained on one language pair, and then the trained parameters are used to initialise a child model, which is then trained on the desired low-resource language pair.

This idea was first explored by Zoph et al. (2016), who considered a French–English parent model, and child models translating from 4 low-resource languages (Hausa, Turkish, Uzbek and Urdu) into English. They showed that transfer learning could indeed improve over random initialisation, and the best performance for this scenario was obtained when the values of the target embeddings were fixed after training the parent, but the training continued for all the other parameters. Zoph et al. (2016) suggest that the choice of the parent language could be important, but did not explore this further for their low-resource languages.

Whilst Zoph et al. (2016) treat the parent and child vocabularies as independent, Nguyen and Chiang (2017) showed that when transferring between related languages (in this case, within the Turkic family), it is beneficial to share the vocabularies between the parent and child models. To boost this effect, subword segmentation such as BPE (Sennrich, Haddow, and Birch 2016b) can help to further increase the vocabulary overlap. In cases where there is little vocabulary overlap (for example, because the languages are distantly related), mapping the bilingual embeddings between parent and child can help (Kim, Gao, and Ney 2019). In some cases, even though the languages are related, they use different scripts. This issue can be solved by transliteration (Goyal, Kumar, and Sharma 2020).

The question of how to choose the parent language for transfer learning, as posed by Zoph et al. (2016), has been taken up by later authors. One study suggests that language relatedness is important (Dabre, Nakagawa, and Kazawa 2017). However in (Kocmi and Bojar 2018) the authors showed that the main consideration in transfer learning is to have a strong parent model, and it can work well for unrelated language pairs. Still, if languages are unrelated and the scripts are different, for example transferring from an Arabic–Russian parent to Estonian–English, transfer learning is less useful. Lin et al. (2019) perform an extensive study on choosing the parent language for transfer learning, showing that data-related features of the parent models and lexical overlap

are often more important than language similarity. Further insight into transfer learning for low-resource settings was provided by (Aji et al. 2020), who analysed the training dynamics and concluded that the parent language is not important. The effectiveness from transfer learning with strong (but linguistically unrelated) parent models has been confirmed in shared task submissions such as (Bawden et al. 2020) – see Section 7.2.4.

Multi-stage transfer learning methods have also been explored. Dabre, Fujita, and Chu (2019) propose a two-step transfer with English on the source side for both parent and child models. First, a one-to-one parent model is used to initialise weights in a multilingual one-to-many model, using a multi-way parallel corpora that includes the child target language. Second, the intermediate multilingual model is fine-tuned on parallel data between English and the child target language. Kim et al. (2019) use a two-parent model and a pivot language. One parent model is between the child source language and the pivot language (e.g. German–English), and the other translates between the pivot and the child target language (e.g. English–Czech). Then, the encoder parameters from the first model and the decoder parameters of the second models are used to initialise the parameters of the child model (e.g. German–Czech).

4.2 Multilingual Models

The goal of multilingual MT is to have a universal model capable of translation between any two languages. Including low-resource language pairs in multilingual models can be seen as means of exploiting additional data from other, possibly related, languages. Having more languages in the training data helps developing an universal representation space, which in turn allows for some level of parameter sharing among the language-specific model components.

The degree to which parameters are shared across multiple language directions varies considerably in the literature, with early models showing little sharing across languages (Dong et al. 2015) and some later models explore the sharing of most or all parameters (Johnson et al. 2017). The amount of parameter sharing can be seen as a trade-off between ensuring that each language is sufficiently represented (has enough parameters allocated) and that low-resource languages can benefit from the joint training of parameters with other (higher-resource) language pairs (which also importantly reduces the complexity of the model by reducing the number of parameters required).

Dong et al. (2015) present one of the earliest studies in multilingual NMT, focused on translation from a single language into multiple languages simultaneously. The central idea of this approach is to have a shared encoder and many language-specific decoders, including language-specific weights in the attention modules. By training on multiple target languages (presenting as a multi-task setup), the motivation is that the representation of the source language will not only be trained on more data (thanks to the multiple language pairs), but the representation may be more universal, since it is being used to decode several languages. They find that the multi-decoder setup provides systematic gains over the bilingual counterparts, although the model was only tested in simulated low-resource settings.

As an extension to this method, Firat, Cho, and Bengio (2016) experiment with multilingual models in the many-to-many scenario. They too use separate encoders and decoders for each language, but the attention mechanism is shared across all directions, which means adding languages increases the number of model parameters linearly (as opposed to quadratic increase when attention is language-direction-specific). In all

cases, the multi-task model performed better than the bilingual models according to BLEU scores, although it was again only tested in simulated low-resource scenarios.

More recent work has looked into the benefits of sharing only certain parts of multilingual models, ensuring language-dependent components. For example, Platanios et al. (2018) present a contextual parameter generator component, which allows finer control of the parameter sharing across different languages. Fan et al. (2020) also include language-specific components by sharing certain parameters across pre-defined language groups in order to efficiently and effectively upscale the number of languages included (see Section 4.2.1).

In a bid to both simplify the model (also reducing the number of parameters) and to maximally encourage sharing between languages, Ha, Niehues, and Waibel (2016) and Johnson et al. (2017) proposed to use a single encoder and decoder to train all language directions (known as the universal encoder-decoder). Whereas Ha, Niehues, and Waibel (2016) propose language-specific embeddings, Johnson et al. (2017) use a joint vocabulary over all languages included, which has the advantage of allowing shared lexical representations (and ultimately this second strategy is the one that has been retained by the community). The control over the target language was ensured in both cases by including pseudo-tokens indicating the target language in the source sentence. Although not trained or evaluated on low-resource language pairs, the model by Johnson et al. (2017) showed promise in terms of the ability to model multilingual translation with a universal model, and zero-shot translation (between language directions for which no parallel training data was provided) was also shown to be possible. We shall see in the next section (Section 4.2.1) how scaling up the number of languages used for training can be beneficial in the low-resource setting.

Combining multilingual models, with the transfer learning approaches of the previous section, Neubig and Hu (2018) present a number of approaches for adaptation of multilingual models to new languages. The authors consider cold- and warm-start scenarios, depending on whether the training data for the new language was available for training the original multilingual model. They find that multilingual models fine-tuned with the low-resource language training data mixed in with data from a similar high-resource language (i.e. similar-language regularisation) give the best translation performance.

4.2.1 Massively multilingual models. In the last couple of years, efforts have been put into scaling up the number of languages included in multilingual training, particularly for the universal multilingual model (Johnson et al. 2017). The motivation is that increasing the number of languages should improve the performance for all language directions, thanks to the addition of extra data and to increased transfer between languages, particularly for low-resource language pairs. For example, Neubig and Hu (2018) trained a many-to-English model with 57 possible source languages, and more recent models have sought to include even more languages; Aharoni, Johnson, and Firat (2019) train an MT model for 102 languages to and from English as well as a many-to-many MT model between 59 languages, and Fan et al. (2020), Zhang et al. (2020a) and Arivazhagan et al. (2019) train many-to-many models for over 100 languages.

While an impressive feat, the results show that it is non-trivial to maintain high translation performance across all languages as the number of language pairs is increased (Mueller et al. 2020; Aharoni, Johnson, and Firat 2019; Arivazhagan et al. 2019). There is a trade-off between *transfer* (how much benefit is gained from the addition of other languages) and *interference* (how much performance is degraded due to having to also learn to translate other languages) (Arivazhagan et al. 2019). It is generally

bad news for high-resource language pairs, for which the performance of multilingual models is usually below that of language-direction-specific bilingual models. However, low-resource languages often do benefit from multilinguality. It has also been shown that for zero-shot translation, the more languages included in the training, the better the results are (Aharoni, Johnson, and Firat 2019; Arivazhagan et al. 2019).

There is often a huge imbalance in the amount of training data available across language pairs, and for low-resource language pairs, it is beneficial to upsample the amount of data. However, upsampling low-resource pairs has the unfortunate effect of harming performance on high-resource pairs (Arivazhagan et al. 2019). A solution to this problem is the commonly used strategy of temperature-based sampling, which involves adjusting how much we sample from the true data distribution (Devlin et al. 2019; Fan et al. 2020), providing a certain compromise between making sure low-resource languages are sufficiently represented but reducing the deterioration in performance seen in high-resource language pairs. Temperature-based sampling can also be used when training the subword segmentation to create a joint vocabulary across all languages so that the low-resource languages are sufficiently represented in the joint vocabulary despite there being little data.

Several works have suggested that the limiting factor is the capacity of the model (i.e. the number of parameters). Whilst multilingual training with shared parameters can increase transfer, increasing the number of languages decreases the per-task capacity of the model. Arivazhagan et al. (2019) suggest that model capacity may be the most important factor in the transfer-interference trade-off; they show that larger models (deeper or wider) show better translation performance across the board, deeper models being particularly successful for low-resource languages, whereas wider models appeared more prone to overfitting. Zhang et al. (2020a) show that online backtranslation combined with a deeper Transformer architecture and a special language-aware layer normalisation and linear transformations between the encoder and the decoder improve the translation in many-to-many setup.

4.2.2 Multilingual Unsupervised Models. As noted in Section 3.2.2, unsupervised MT performs quite poorly in low-resource language pairs, and one of the ways in which researchers have tried to improve its performance is by exploiting data from other languages. Sen et al. (2019b) demonstrate that a multilingual unsupervised NMT model can perform better than bilingual models in each language pair, but they only experiment with high-resource language pairs. Later works (Garcia et al. 2021; Ko et al. 2021) directly address the problem of unsupervised NMT for a low-resource language pair in the case where there is parallel data in a related language. More specifically, they use data from a third language (Z) to improve unsupervised MT between a low-resource language (X) and a high-resource language (Y). In both works, they assume that X is closely related to Z, and that there is parallel data between Y and Z. As in the original unsupervised NMT models (Lample et al. 2018a; Artetxe et al. 2018), the training process uses denoising autoencoders and iterative backtranslation.

4.3 Large-scale Multilingual pre-training

The success of large-scale pre-trained language models such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019) has inspired researchers to apply related techniques to MT. Cross-lingual language models (XLM; Conneau and Lample 2019) are a direct application of the BERT masked language model (MLM) objective to learn from parallel data. The training data consists of concatenated sentence pairs, so that the model

learns to predict the identity of the masked words from the context in both languages simultaneously. XLM was not applied to low-resource MT in the original paper, but was shown to improve unsupervised MT, as well as language modelling and natural language inference in low-resource languages.

The first really successful large-scale pre-trained models for MT were mBART (Liu et al. 2020) and MASS (Song et al. 2019b), which demonstrated improvements to NMT in supervised, unsupervised and semi-supervised (i.e. with back-translation) conditions, including low-resource language pairs. The idea of these models is to train a noisy autoencoder using large quantities of monolingual data in 2 or more languages. The autoencoder is a transformer-based encoder-decoder, and the noise is introduced by randomly masking portions of the input sentence. Once the autoencoder has been trained to convergence, its parameters can be used to initialise the MT model, which is trained as normal. Using mBART, Liu et al. (2020) were able to demonstrate unsupervised NMT working on the distant low-resource language pairs Nepali–English and Sinhala–English, as well as showing improvements in supervised NMT in low-resource language pairs such as Gujarati–English.

The original mBART was trained on 25 different languages and its inclusion in HuggingFace (Wolf et al. 2020) makes it straightforward to use for pre-training. It has since been extended to mBART50 (Tang et al. 2020), which is trained on a mixture of parallel and monolingual data, and includes 50 different languages (as the name suggests); mBART50 is also available on HuggingFace. A recent case study (Birch et al. 2021) has demonstrated that mBART50 can be combined with focused data gathering techniques to quickly develop a domain-specific, state-of-the-art MT system for a low-resource language pair (in this case, Pashto–English).

A recent multilingual pre-trained method called mRASP (Lin et al. 2020) has shown strong performance across a range of MT tasks: medium, low and very low-resource. mRASP uses unsupervised word alignments generated by MUSE (Conneau et al. 2018) to perform random substitutions of words with their translations in another language, with the aim of bringing words with similar meanings across multiple languages closer in the representation space. They show gains of up to 30 BLEU points for some very low-resource language pairs such as Belarusian–English. mRASP2 (Pan et al. 2021) extends this work by incorporating monolingual data into the training.

Of course, pre-trained models are useful if the languages you are interested in are included in the pre-trained model, and you have the resources to train and deploy these very large models. On the former point, Muller et al. (2021) have considered the problem of extending multilingual BERT (mBERT) to new languages for natural language understanding tasks. They find greater difficulties for languages which are more distant from those in mBERT and/or have different scripts – but the latter problem can be mitigated with careful transliteration.

5. Use of external resources and linguistic information

For some languages, alternative sources of linguistic information, for example (i) linguistics tools (Section 5.1) and (ii) bilingual lexicons (Section 5.2), can be exploited. They can provide richer information about the source or target languages (in the case of tagging and syntactic analysis) and additional vocabulary that may not be present in parallel data (in the case of bilingual lexicons and terminologies). Although there has been a large body of work in this area in MT in general, only some have been applied to true low-resource settings. We therefore review work looking at exploiting these two sources of additional information, for languages where such resources are available and put a particular focus on those that have been applied to low-resource languages.

5.1 Linguistic tools and resources

Additional linguistic analysis such as part-of-speech tagging, lemmatisation and parsing can help to reduce sparsity by providing abstraction from surface forms, as long as the linguistic tools and resources are available. A number of different approaches have developed for the integration of linguistic information in NMT. These include morphological segmentation, factored representations (Section 5.1.2), multi-task learning (Section 5.1.3), interleaving of annotations (Section 5.1.4) and syntactic reordering (Section 5.1.5).

5.1.1 Morphological segmentation. A crucial part of training NMT system is the choice of subword segmentation, a pre-processing technique providing the ability to represent an infinite vocabulary with a fixed number of units and to better generalise over shorter units. For low-resource languages, it is even more important because there is a greater chance of coming across words that were not seen at training time. The most commonly used strategies are statistics-based, such as BPE (Sennrich, Haddow, and Birch 2016b) and sentencepiece (Kudo and Richardson 2018). Not only might these strategies not be optimal from a point of view of linguistic generalisation, but for low-resource languages especially they have also been shown to give highly variable results, depending on what degree of segmentation is selected; this degree is a parameter which therefore must be chosen wisely (Ding, Renduchintala, and Duh 2019; Sennrich and Zhang 2019).

Several works have tested the use morphological analysers to assist the segmentation of texts into more meaningful units for low-resource languages. In their submission to the WMT19 shared task for English→Kazakh, Sánchez-Cartagena, Pérez-Ortiz, and Sánchez-Martínez (2019) use the morphological analyser from Apertium (Forcada and Tyers 2016) to segment Kazakh words into stem (often corresponding to the lemma in Kazakh) and the remainder of the word. They then learnt BPE over the morphological segmented data. Saleva and Lignos (2021) also test morphologically aware subword segmentation for three low-resource language pairs: Nepali, Sinhala and Kazakh to and from English. They test segmentations using the LMVR (Ataman et al. 2017) and MORSEL (Lignos 2010) analysers, but found no gain over BPE and no consistent pattern in the results. These results go against previous results from Grönroos et al. (2014) that showed that an LMVR segmentation can outperform BPE when handling low-resource Turkish, but they are in accordance with more recent ones for Kazakh–English (Toral et al. 2019) and Tamil–English (Dhar, Bisazza, and van Noord 2020), where it does not seem to improve over BPE.

5.1.2 Factored models. Factored source and target representations (Garcia-Martinez, Barrault, and Bougares 2016; Sennrich and Haddow 2016; Burlot et al. 2017) were designed as a way of decomposing word units into component parts, which can help to provide some level of composite abstraction from the original wordform. For example, a wordform may be represented by its lemma and its part-of-speech, which together can be used to recover the original surface form. This type of modelling can be particularly useful for morphologically rich languages (many of which are already low-resource), for which the large number of surface forms can result in greater data sparsity and normally necessitate greater quantities of data.

Factored models originated in SMT (Koehn and Hoang 2007), but were notably not easy to scale. The advantage of factored representations in NMT is that the factors are represented in continuous space and therefore may be combined more easily, without resulting in an explosion in the number of calculations necessary. Garcia-Martinez, Barrault, and Bougares (2016), Sennrich and Haddow (2016) and Burlot et al. (2017) evaluate on language pairs involving at least one morphologically rich languages and show that improvements in translation quality can be seen, but this is dependent on the language pair and the type of linguistic information included in the factors. Nădejde et al. (2017) use factors to integrate source-side syntactic information in the form of CCG tags, which they combine with an interleaving approach on the target side (see Section 5.1.4) to significantly improve MT performance, for high-resource (German→English) and mid-low-resource (Romanian→English) language directions.

5.1.3 Multi-task learning. Multi-task learning can be seen more as a way of forcing the model to learn better internal representations of wordforms by training the model to produce a secondary output (in this case linguistic analyses) as well as a translation.

Initial work in multi-task learning for MT did not concentrate on the low-resource scenario. Luong et al. (2016) explore different multi-task setups for translation (testing on English–German), among which a setup where they use parsing as an auxiliary task to translation, which appears to help translation performance as long as the model is not overly trained on the parsing task. The question of how to optimally train such multi-task models has inevitably since been explored, inspired in part by concurrent work in multi-encoder and multi-decoder multilingual NMT (See Section 4), since it appears that sharing all components across all tasks is not the optimal setting. Niehues and Cho (2017) experiment with part-of-speech (PoS) tagging and named entity recognition as auxiliary tasks to MT and test different degrees of sharing. They find that sharing the encoder only (i.e. separate attention mechanisms and decoders) works best and that using both auxiliary tasks enhances translation performance in a simulated low-resource DE→EN scenario.

Since then, there have been some applications of multi-task learning to lower-resource scenarios, with slight gains in translation performance. Nădejde et al. (2017) also share encoders in their multi-task setting for the integration of target-side syntactic information in the form of CCG supertags (for DE→EN and mid-low-resource RO→EN). Similarly, Zaremoodi, Buntine, and Haffari (2018) develop a strategy to avoid task interference in a multi-task MT setup (with named entity recognition, semantic parsing and syntactic parsing as auxiliary tasks). They do so by extending the recurrent components of the model with multiple blocks and soft routing between them to act like experts. They test in real low-resource scenarios (Farsi–English and Vietnamese–English) and get gains of approximately 1 BLEU point by using the additional linguistic information in the dynamic sharing setup they propose.

5.1.4 Interleaving of linguistic information in the input. As well as comparing factored representations and multi-task decoding, Nădejde et al. (2017) also introduce a new way of integrating target-side syntactic information, which they call *interleaving*. The idea is to annotate the target side of the training data with token-level information (CCG supertags in their case) by adding a separate token before each token containing the information pertaining to it, so that the model learns to produce the annotations along with the translation. They found this to work better than multi-task for the integration of target-side annotations and was also complementary with the use of source factors. Inspired by these results, Tamchyna, Weller-Di Marco, and Fraser (2017) also followed the interleaving approach (for English→Czech and English→German, so not low-resource scenarios), but with the prediction of interleaved morphological tags and lemmas, followed by a deterministic wordform generator. Whereas Nădejde et al. (2017) seek to create representations that are better syntactically informed, the aim of (Tamchyna, Weller-Di Marco, and Fraser 2017) is different: they aim to create a better generalisation capacity for translation into a morphologically rich language by decomposing wordforms into their corresponding tags and lemmas. They see significantly improved results with the two-step approach, but find that simply interleaving morphological tags (similar to (Nădejde et al. 2017)) does not lead to improvements. They hypothesise that the morphological tags are less informative than CCG supertags and therefore the potential gain in information is counterbalanced by the added difficulty of having to translate longer target sequences.

In a systematic comparison with both RNN and transformer architectures and for 8 language directions (and in particular for low-resource languages), Sánchez-Cartagena, Pérez-Ortiz, and Sánchez-Martínez (2020) find that interleaving (with part-of-speech information and morphological tags) is beneficial, in line with the conclusions from (Nădejde et al. 2017). Interestingly, they find that (i) interleaving linguistic information in the source sentence can help, and morphological information is better than PoS tags and (ii) interleaving in the target sentence can also help, but PoS tagging is more effective than morphological information, despite the translations being more grammatical with added morphological information.

5.1.5 Syntactic Reordering. Other than being used as an additional form of input, syntactic information can also be used *a priori* to facilitate the translation task by reordering words within sentences to better match a desired syntactic order. Murthy, Kunchukuttan, and Bhattacharyya (2019) found this to be particularly effective for very low-resource languages in a transfer learning setup, when transferring from a high-resource language pair to a low-resource pair (see Section 4.1). Testing on translation into Hindi from Bengali, Gujarati, Marathi, Malayalam and Tamil, having transferred from the parent language direction English→Hindi, they apply syntactic reordering rules on the source-side to match the syntactic order of the child source language, resulting in significant gains in translation quality.

5.2 Bilingual lexicons

Bilingual lexicons are lists of terms (words or phrases) in one language associated with their translations in a second language. The advantage of bilingual lexicons is that they may well provide specialist or infrequent terms that do not appear in available parallel data, with the downside that they do not give information about the translation of terms in context, notably when there are several possible translations of a same term. However, they may be important resources to exploit, since they provide complemen-

tary information to parallel data and may be more readily available and cheaper to produce.[12]

The approaches developed so far to exploit bilingual lexicons in MT can be summarised as follows: (i) as seed lexicons to initialise unsupervised MT (Lample et al. 2018b; Duan et al. 2020) (as described in Section 3.2.2), (ii) as an additional scoring component, particularly to provide coverage for rare of otherwise unseen vocabulary (Arthur, Neubig, and Nakamura 2016; Feng et al. 2017) and (iii) as annotations in the source sentence by adding translations from lexicons just after their corresponding source words (Dinu et al. 2019)[13] or by replacing them in a code-switching-style setup (Song et al. 2019a).

The most recent work on using lexicons in pretrained multilingual models (Lin et al. 2020, mRASP) shows the most promise. Here translations of words are substituted into the source sentence in pretraining, with the goal of bringing words with similar meanings across multiple languages closer in the representation space. Please see Section 4.3 for more details.

12 Note that many of the works designed to incorporate bilingual lexicons actually work on automatically produced correspondences in the form of phrase tables (produced using SMT methods). Although these may be extracted from the same parallel data as used to train the NMT model, it may be possible to give more weight to infrequent words than may be the case when the pairs are learnt using NMT.

13 The origin of the source units (i.e. original or inserted translation) is distinguished by using factored representations. A similar technique was used to insert dictionary definitions rather than translations by Zhong and Chiang (2020).

6. Model-centric Techniques

In the previous two sections we have looked at using monolingual data, and data from other langauage pairs to improve translation. In this section we explore work that aims to make better use of the data we already have by investigating better modelling, training and inference techniques.

In recent years, MT systems have converged towards a fairly standardised architecture: a sequence-to-sequence neural network model with an encoder and an autoregressive decoder, typically implemented as a Transformer (Vaswani et al. 2017), although recurrent models (Bahdanau, Cho, and Bengio 2015) are still used. Training is performed on a parallel corpus by minimising the cross-entropy of the target translations conditional on the source sentences. Monolingual examples, if available, are typically converted to parallel sentences, as discussed in Section 3. Once the model is trained, translations are usually generated by beam search with heuristic length control, which approximates, though not quite accurately, maximum a posteriori (MAP) inference on the learned conditional probability distribution.

This approach has been very successful for MT on high-resource language pairs where there is enough high-quality parallel and monolingual text covering a wide variety of domains to wash out most of the misaligned inductive bias that the model might have. However, for low-resource language pairs the inductive bias of the model becomes more prominent, especially when the model operates out of the training distribution, as it frequently does when the training data has sparse coverage of the language. Therefore it can be beneficial to design the neural network architecture and training and inference procedures to be more robust to low-resource conditions, for instance by explicitly modelling the aleatoric uncertainty that is intrinsic to the translation task due to its nature of being a many-to-many mapping.

In this section, we will review recent machine learning techniques that can improve low-resource MT, such as meta-learning for data-efficient domain adaptation and multilingual learning (Section 6.1), Bayesian and latent variable models for explicit quantification of uncertainty (Section 6.2) and alternatives to cross-entropy training (Section 6.3) and beam search inference (Section 6.4).

6.1 Meta Learning

In Section 4 we discussed using multilingual training to improve low-resource MT by combining training sets for different language pairs in joint-learning or transfer learning schemes. A more extreme form of this approach involves the application of *meta learning*: rather than training a system to directly perform well on a single task or fixed set of tasks (language pairs in our case), a system can be trained to quickly adapt to a novel task using only a small number of training examples, as long as this task is sufficiently similar to tasks seen during (meta-)training.

One of the most successful meta-learning approaches is Model-Agnostic Meta-Learning (*MAML*) (Finn, Abbeel, and Levine 2017), which was applied to multilingual MT by Gu et al. (2018). In MAML, we train task-agnostic model parameters $\bar{\theta}$ so that they can serve as a good initial value that can be further optimised towards a task-specific parameter vector $\overset{*}{\theta}_m$ based on a task-specific training set D_m. This is accomplished by repeatedly simulating the fine-tuning procedure, evaluating each fine-tuned model on its task-specific evaluation set and then updating the task-agnostic parameters in the direction that improves this score.

Once training is completed, the fine-tuning procedure can be directly applied to any novel task. Gu et al. (2018) apply MAML by meta-training on synthetic low-resource tasks obtained by randomly subsampling parallel corpora of high-resource language pairs and then fine-tune on true low-resource language pairs, obtaining substantial improvements.

An alternative approach to meta-learning involves training *memory-augmented networks* that receive the task-specific training examples at execution time and maintain a representation of them which they use to adapt themselves on the fly (Vinyals et al. 2016; Santoro et al. 2016), an approach related to the concept of "fast weights" computed at execution time as opposed to "slow weights" (the model parameters) computed at training time (Schmidhuber 1992). Lake (2019) applied memory-augmented networks to synthetic sequence-to-sequence tasks in order to evaluate out-of-distribution generalisation under a variety of conditions. Curiously, very large language models such as GPT-2 (Radford et al. 2019) and in particular GPT-3 (Brown et al. 2020) also exhibit this meta-learning capability even without any modification of the network architecture or training procedure, suggesting that meta-learning itself can be learned from a sufficient large amount of data. In fact, GPT-3 achieves near-SOTA quality when translating into English even with a single translation example, for multiple source languages including Romanian, a medium-low resource language.

6.2 Latent variable models

Auto-regressive NMT models can in principle represent arbitrary probability distributions given enough model capacity and training data. However in low-resource conditions, the inductive biases of the models might be insufficient for a good generalisation.

Various approaches have attempted to tackle these issues by introducing *latent variables*, random variables that are neither observed as source sentences nor as target sentences, and are rather inferred internally by the model. This can be done with a *source-conditional* parametrisation, which applies latent-variable modelling only on the target sentence or with a *joint* parametrisation, which applies it to both the source and the target sentences.

Latent variable models enable a higher model expressivity and more freedom in the engineering of the inductive biases, at the cost of more complicated and computationally expensive training and inference. For this reason, approximation techniques such as Monte Carlo sampling or MAP inference over the latent variable are used, typically based on the *Variational Autoencoder* framework (VAE) (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014).

In the earliest Variational NMT approach by Zhang et al. (2016), a source-conditional parametrisation is used and the latent variable is a fixed-dimension continuous variable that is intended to capture global information about the target sentence. Training is performed by maximising a lower bound, known as the *Evidence Lower Bound* (ELBO) of the conditional cross-entropy of the training examples, which is computed using an auxiliary model component known as *inference network* that approximates the posterior of the latent variable as a diagonal Gaussian conditional on both the source and the target sentence. During inference the latent variable is either sampled from the prior or more commonly approximated as its mode (which is also its mean). This basic approach, similar to image VAEs and the Variational Language Model of Bowman et al. (2016), adds limited expressivity to autoregressive NMT because a fixed-dimensional unimodal distribution is not especially well suited to represent the variability of a sentence, but it can be extended in various ways: Su et al. (2018) and Schulz, Aziz, and

Cohn (2018) use a sequence of latent variables, one for each target token, parametrised with temporal dependencies between each other.

Setiawan et al. (2020) parametrise the latent posterior using *normalising flows* (Rezende and Mohamed 2015), which can represent arbitrary and possibly multi-modal distributions as a sequence of transformation layers applied to a simple base distribution.

Eikema and Aziz (2019) use a joint parametrisation as they claim that explicitly modelling the source sentence together with the target sentence provides additional information to the model. Inference is complicated by the need to post-condition the joint probability distribution on the source sentence, hence a series of approximations is used in order to ensure efficiency.

The latent variable models described so far have been evaluated on high-resource language pairs, although most of them have been evaluated on the IWSLT dataset, which represents a low-resource domain. However, latent-variable MT has also been applied to fully low-resource language pairs, using models where the latent variables have been designed to have linguistically-motivated inductive bias. Ataman, Aziz, and Birch (2019) introduce a NMT model with latent word morphology in a hierarchical model, allowing for both word level representations and character level generation to be modelled. This is beneficial for morphologically rich languages, which include many Turkic and African low-resource languages. These languages use their complex morphologies to express syntactic and semantic nuance, which might not be captured by the purely unsupervised and greedy BPE preprocessing, especially when the BPE vocabulary is trained on a small corpus. The proposed model uses for each word one multivariate Gaussian latent variable representing a lemma embedding and a sequence of quasi-discrete latent variables representing morphology. Training is performed in a variational setting using a relaxation based on the Kumaraswamy distribution (Kumaraswamy 1980; Louizos, Welling, and Kingma 2018; Bastings, Aziz, and Titov 2019), and inference is performed by taking the modes of the latent distributions, as the model is source-conditional. This approach has been evaluated on morphologically rich languages including Turkish, yielding improvements both in in-domain and out-of-domain settings.

6.3 Alternative training objectives

When an autoregressive model is trained to optimise the cross-entropy loss, it is only exposed to ground-truth examples during training. When this model is then used to generate a sequence with ancestral sampling, beam search or another inference method, it has to incrementally extend a prefix that it has generated itself. Since the model in general cannot learn exactly the "true" probability distribution of the target text, the target prefix that it receives as input will be out-of-distribution, which can cause the estimate of the next token probability to become even less accurate. This issue, named *exposure bias* by Ranzato et al. (2016), can compound with each additional token and might result in the generated text to eventually become completely nonsensical. Exposure bias theoretically occurs regardless of the task, but while its impact has been argued to be small in high-resource settings (Wu et al. 2018), in low-resource MT it has been shown to be connected to the phenomenon of *hallucination*, where the system generate translations that are partially fluent but contain spurious information not present in the source sentence (Wang and Sennrich 2020).

A number of alternatives to cross-entropy training have been proposed in order to avoid exposure bias, which all involve exposing the model during training to complete

or partial target sequences generated by itself. Ranzato et al. (2016) explore multiple training strategies and propose a method called *MIXER* which is a variation of the *REINFORCE* algorithm (Williams 1992; Zaremba and Sutskever 2015). In practice REINFORCE suffers from high variance, and they therefore apply it only after the model has already been pre-trained with cross-entropy, a technique also used by all other the training methods described in this section. They further extend the algorithm by combining cross-entropy training and REINFORCE within a each sentence according to a training schedule which interpolates from full cross-entropy to full REINFORCE. They do not evaluate on a true low-resource language pair, but they do report improvements on German→English translation on the relatively small IWSLT 2014 dataset (Cettolo et al. 2014).

Contrastive Minimum Risk Training (CMRT or just MRT) (Och 2003; Shen et al. 2016; Edunov et al. 2018b) is a similar training technique that can be considered a biased variant REINFORCE that focuses on high-probability translations generated by decoding from the model itself. Wang and Sennrich (2020) apply CMRT to low-resource translation (German→Romansh) as well as German→English IWSLT 2014, reporting improvements in the out-of-domain test case, as well as a reduction of hallucinations.

Both REINFORCE and CMRT use a reward function that measures the similarity between a sampled and reference translations. However, the exact mechanism that makes such approaches work is not completely clear, Choshen et al. (2020) show that REINFORCE and CMRT also work when the reward function is a trivial constant function rather than a sentence similarity metric, suggesting that their primary effect is to regularise the model pretrained with cross-entropy by exposing it to its own translations, hence reducing exposure bias.

An alternative training technique involving beam search decoding in the training loop has been proposed by Wiseman and Rush (2016), based on the *LaSO* (Daumé and Marcu 2005) structured learning algorithm. This approach also exposes the model to its own generations during training, and it has the benefit that training closely matches the inference process, reducing any mismatch. The authors report improvements on German→English IWSLT 2014. However they do not evaluate on a true low-resource language pair.

An even simpler technique that exposes the model's own generations during training is *scheduled sampling* (Bengio et al. 2015), which also starts with cross-entropy training and progressively replaces part of the ground truth target prefixes observed by the model with its own samples. Plain scheduled sampling is theoretically unsound (Huszar 2015), but it can be made more consistent by backpropagating gradients through a continuous relaxation of the sampling operation, as shown by Xu, Niu, and Carpuat (2019) who report improvements on low-resource (Vietnamese→English) MT.

Regularisation techniques have also been applied to low-resource MT. Sennrich and Zhang (2019) evaluated different hyperparameter settings, in particular batch size and dropout regularisation, for German→English with varying amounts of training data and low-resource Korean→English. Müller, Rios, and Sennrich (2020) experimented with various training and inference techniques for out-of-distribution MT both for a high-resource (German→English) and low-resource (German→Romansh) pair. For the low-resource pair they report improvements by using sub-word regularisation (Kudo 2018), defensive distillation and source reconstruction. An alternate form of subword regularisation, known as BPE dropout has been proposed by Provilkov, Emelianenko, and Voita (2019), reporting improvements on various high-resource and low-resource language pairs. He, Haffari, and Norouzi (2020) apply a dynamic programming approach to BPE subword tokenisation, evaluating during training all possible ways of

tokenising each target word into subwords, and computing an optimal tokenisation at inference time. Since their method is quite slow however, they only use it to tokenise the training set and then train a regular Transformer model on it, combining it with BPE dropout on source words, reporting improvements on high-resource and medium-resource language pairs.

6.4 Alternative inference algorithms

In NMT, inference is typically performed using a type of beam search algorithm with heuristic length normalisation[14] (Jean et al. 2015; Koehn and Knowles 2017). Ostensibly, beam search seeks to approximate maximum a posteriori (MAP) inference. However it has been noticed that increasing the beam size, which improves the accuracy of the approximation, often degrades translation quality after a certain point (Koehn and Knowles 2017). It is actually feasible to exactly solve the maximum a posteriori inference problem, and the resulting mode is often an abnormal sentence; in fact, it is often the empty sentence (Stahlberg and Byrne 2019). It is arguably dismaying that NMT relies on unprincipled inference errors in order to generate accurate translations. Various authors have attributed this "beam search paradox" to modelling errors caused by exposure bias or other training issues and they have proposed alternative training schemes such as these discussed in section 6.3. Even a perfect probabilistic model, however, could well exhibit this behaviour due to a counter-intuitive property of many high-dimensional random variables that causes the mode of the distribution to be very different from *typical* samples, which have a log-probability close to the entropy of the distribution. See Cover and Thomas (2006) for a detailed discussion of *typicality* from an information theory perspective. Eikema and Aziz (2020) recognise this issue in the context of NMT and tackle it by applying *Minimum Bayes Risk* (MBR) inference (Goel and Byrne 2000).

Minimum Bayes Risk seeks to generate a translation which is maximally similar, according to a metric such as BLEU or METEOR (Denkowski and Lavie 2011), to other translations sampled from the model itself, each weighted according to its probability. The intuition is that the generated translation will belong to a high-probability cluster of similar candidate translations; highly abnormal translations such as the empty sentence will be excluded. Eikema and Aziz (2019) report improvements over beam search on the low-resource language pairs of the FLoRes dataset (Nepali–English and Sinhala–English) (Guzmán et al. 2019) while they lose some accuracy on English–German. They also evaluate inference though ancestral sampling, the simplest and theoretically least biased inference technique, but they found that it performs worse than both beam search and MBR.

Energy-based models (EBMs) (LeCun et al. 2006) are alternative representations of a probability distribution which can be used for inference. An ERM of a random variable (a whole sentence, in our case) is a scalar-valued *energy function*, implemented as a neural network, which represents an unnormalised log-probability. This lack of normalisation means that only probability ratios between two sentences can be computed efficiently, for this reason training and sampling from EBMs requires a proposal distribution to generate reasonably good initial samples to be re-ranked by the model, in the context of MT this proposal distribution is a conventional autoregressive NMT

14 The implementations of beam search differ between MT toolkits in details that can have significant impact over translation quality and are unfortunately often not well documented in the accompanying papers.

model. Naskar et al. (2020) define source-conditional or joint EBMs trained on ancestral samples from an autoregressive NMT model using a reference-based metric (e.g. BLEU). During inference they apply the EBM to re-rank a list of N ancestral samples from the autoregressive NMT model. This approximates MAP inference on a probability distribution that tracks the reference-based metric, which would not give high weight to abnormal translations such as the empty sentence. They report improvements on multiple language pairs, in particular for medium-resource and low-resource language pairs such as Romanian, Nepali, and Sinhala to English.

Reranking has been also applied under the generalised noisy channel model approach initially developed for SMT (Koehn, Och, and Marcu 2003), where translations are scored not just under the probability of the target conditional on the source (direct model) but also under the probability of the source conditional on the target (channel model) and the unconditional probability of the target (language model prior), combined by a weighted sum of their logarithms. This reranking can be applied at sentence level on a set of candidate translations generated by the direct model by conventional beam search (Chen et al. 2020) or at token level interleaved with beam search (Bhosale et al. 2020), resulting in improvements in multiple language pairs including low-resource ones.

7. Shared Tasks

MT is a big field and many interesting papers are published all the time. Because of the variety of language pairs, toolkits and settings, it can be difficult to determine what research will have a impact beyond the published experiments. Shared tasks provide an opportunity to reproduce and combine research, while keeping the training and testing data constant.

System description papers can offer valuable insights into how research ideas can transfer to real gains when aiming to produce the best possible system with the data available. Whilst standard research papers often focus on showing that the technique or model proposed in the paper is effective, the incentives for system descriptions are different; authors are concerned with selecting the techniques (or most commonly the combination of techniques) that work best for the task. System descriptions therefore contain a good reflection of the techniques that researchers believe will work (together with their comparison) in standardised conditions. The difficulty with system descriptions is that the submitted systems are often potpourris of many different techniques, organised in pipelines with multiple steps, a situation that rarely occurs in research papers presenting individual approaches. In light of this, it is not always easy to pinpoint exactly which techniques lead to strongly performing systems, and it is often the case that similar techniques are used by both leading systems and those that perform less well. Moreover, the papers do not tend to provide an exhaustive and systematic comparison of different techniques due to differences in implementation, data processing and hyperparameters. We also note that the evaluation of shared tasks normally focuses on quality alone, although a multi-dimensional analysis may be more appropriate (Ethayarajh and Jurafsky 2020), and that even if the task has manual evaluation, there is still debate about the best way to do this (Freitag et al. 2021).

Apart from the system descriptions, an important output of shared tasks is the publication of standard training sets and test sets (Section 2). These can be used in later research, and help to raise the profile of the language pair for MT research.

In this section we survey the shared tasks that have included low-resource language pairs, and we draw common themes from the corresponding sets of system description papers, putting into perspective the methods previously described in this survey. Rather than attempting to quantify the use of different techniques *à la* (Libovický 2021), we aim to describe how the most commonly used techniques are exploited, particularly for high-performing systems, providing some practical advice to training systems for low-resource language pairs. We begin with a brief description of shared tasks featuring low-resource pairs (Section 7.1), before surveying techniques commonly used (Section 7.2).

7.1 Low-resource MT in Shared Tasks

There are many shared tasks that focus on MT, going all the way back to the earliest WMT shared task (Koehn and Monz 2006). However they have tended to focus on well-resourced European languages and Chinese. Tasks specifically for low-resource MT are fairly new, coinciding with the recent interest in expanding MT to a larger range of language pairs.

We choose to focus particularly on shared tasks run by WMT (WMT Conference on Machine Translation), IWSLT (International Conference on Spoken Language Translation), WAT (Workshop on Asian Translation) and LowResMT (Low Resource Machine Translation). In Table 3, we list the shared MT task that have focused on low-resource

pairs. In addition to the translation tasks, we should mention that the corpus filtering task at WMT has specifically addressed low-resource MT (Koehn et al. 2019, 2020).

Year	Task name and reference	Language pairs
2018	IWSLT (Niehues et al. 2018)	Basque–English
2018	WAT Mixed domain (Nakazawa et al. 2018)	Myanmar–English
2019	WAT Mixed domain (Nakazawa et al. 2019)	Myanmar–English and Khmer–English
2019	WAT Indic (Nakazawa et al. 2019)	Tamil–English
2019	WMT news (Barrault et al. 2019)	Kazakh–English and Gujarati–English
2019	LowResMT (Ojha et al. 2020)	{Bhojpuri, Latvian, Magahi and Sindhi} – English
2020	WMT news (Barrault et al. 2020)	{Tamil, Inuktitut, Pashto and Khmer} – English
2020	WMT Unsupervised and very low resource (Fraser 2020)	Upper Sorbian–German
2020	WMT Similar language (Barrault et al. 2020)	Hindi–Marathi
2020	WAT Mixed domain (Nakazawa et al. 2020)	Myanmar–English and Khmer–English
2020	WAT Indic (Nakazawa et al. 2020)	Odia–English
2021	AmericasNLP (Mager et al. 2021)	Ten indigenous languages of Latin America, to/from Spanish
2021	WAT News Comm (Nakazawa et al. 2021)	Japanese–Russian
2021	WAT Indic (Nakazawa et al. 2021)	Ten Indian languages, to/from English
2021	LowResMT	Taiwanese Sign Language–Trad. Chinese, Irish–English and Marathi–English
2021	WMT News	Hausa–English and Bengali–Hindi
2021	WMT Unsupervised and very low resource	Chuvash–Russian and Upper Sorbian–German
2021	WMT Similar language	Tamil–Telugu, Bambara–French and Maninka–French

Table 3: Shared tasks that have included low-resource language pairs

7.2 Commonly used Techniques

In this section, we review the choices made by participants to shared tasks for low-resource MT, focusing on those techniques that are particularly widespread, those that work particularly well and the choices that are specific to particular languages or language families. We describe these choices in an approximately step-by-step fashion: starting with data preparation (Section 7.2.1) and data processing (Section 7.2.2), then proceeding to model architecture choices (Section 7.2.3), exploiting additional data, including backtranslation, pretraining and multilinguality (Section 7.2.4) and finally looking at model transformation and finalisation, including ensembling, knowledge distillation and fine-tuning (Section 7.2.5).

7.2.1 Data preparation. An important initial step to training an NMT model is to identify available data (See Section 2) and to potentially filter it depending on the noisiness of the dataset and how out-of-domain it is or to use an alternative strategy to indicate domain or data quality (i.e. tagging). So what choices do participants tend to make

in terms of using (or excluding) data sources, filtering and cleaning of data and using meta-information such as domain tags?

Choice of data. We focus on constrained submissions only (i.e. where participants can only use the data provided by the organisers), so most participants use all available data. Hesitancy can sometimes be seen with regards to web-crawled data (other than WMT newscrawl, which is generally more homogeneous and therefore of better quality), some choosing to omit the data (Singh 2020) and others to filter it for quality (Chen et al. 2020; Li et al. 2020). It is very unusual to see teams do their own crawling (Hernandez and Nguyen (2020) is a counter-example); teams doing so run the risk of crawling data that overlaps with the development set or one side of the test set.

Data cleaning and filtering. Although not exhaustively reported, many of the submissions apply some degree of data cleaning and filtering to the parallel and monolingual data. In its simplest form, this means excluding sentences based on their length (if too long) and the ratio between the lengths of parallel sentences (if too different). Some teams also remove duplicates (e.g. Li et al. (2019a)). More rigorous cleaning includes eliminating sentences containing fewer than a specified percentage of alpha-numeric characters in sentences (depending on the language's script), those identified as belonging to another language (e.g. using language identification) or those less likely to belong to the same distribution as the training data (e.g. using filtering techniques such as Moore-Lewis (Moore and Lewis 2010)). Data filtering is also a commonly used technique for back-translation data (see the paragraph on data augmentation below), often using similar filtering techniques such as dual conditional cross-entropy filtering (Junczys-Dowmunt 2018) to retain only the cleanest and most relevant synthetic parallel sentences. Unfortunately the effect of data filtering is rarely evaluated, probably because it would involve expensive re-training.

Data tagging. Some teams choose to include meta-information in their models through the addition of pseudo-tokens. For example, Dutta et al. (2020) choose to tag sentences according to their quality for the Upper Sorbian–German task, this information being provided by the organisers. Domain tagging (i.e. indicating the type of data), which can be useful to indicate whether data is in-domain or out-of-domain was used by Chen et al. (2020), one of the top-scoring systems for Tamil–English. For the Basque–English task, Scherrer (2018) find that using domain tags gives systematic improvements over not using them, and Knowles et al. (2020a) come to the same conclusion when translating into Inuktitut.

7.2.2 Data pre-processing. There is some variation in which data pre-processing steps are used. For example, it has been shown that for high-resource language pairs such as Czech–English, it is not always necessary to applying tokenisation and truecasing steps (Bawden et al. 2019) before apply subword segmentation. We do not observe a clear pattern, with many systems applying all steps, and some excluding tokenisation (Wu et al. 2020 for Tamil) and truecasing. Among the different possible pre-processing steps, we review participants choices concerning tokenisation, subword segmentation and transliteration/alphabet mapping (relevant when translating between languages that use different scripts).

Tokenisation. If a tokeniser is used before subword segmentation, it is common for it to be language-specific, particularly for the low-resource language in question. For

example IndicNLP[15] (Kunchukuttan 2020) is widely used for Indian languages (e.g. for the shared tasks involving Gujarati and Tamil), and many of the Khmer–English submissions also used Khmer-specific tokenisers. For European languages, the Moses tokeniser (Koehn et al. 2007) remains the most commonly used option.

Subword segmentation. All participants perform some sort of subword segmentation, with most participants using either sentencepiece (Kudo and Richardson 2018)[16] or subword_nmt toolkits (Sennrich, Haddow, and Birch 2016b).[17] Even though the BPE toolkit is not compatible with the Abugida scripts used for Gujarati, Tamil and Khmer (in these scripts, two unicode codepoints can be used to represent one glyph), we only found one group who modified BPE to take this into account (Shi et al. 2020). BPE-dropout (Provilkov, Emelianenko, and Voita 2020), a regularisation method, was found to be useful by a number of teams (Knowles et al. 2020b; Libovický et al. 2020; Chronopoulou et al. 2020).

The size of the subword vocabulary is often a tuned parameter, although the range of different values tested is not always reported. Surprisingly, there is significant variation in the subword vocabulary sizes used, and there is not always a clear pattern. Despite the low-resource settings, many of the systems use quite large subword vocabularies (30k-60k merge operations). There are exceptions: a large number of the systems for Tamil–English use small vocabularies (6k-30k merge operations), which may be attributed to the morphologically rich nature of Tamil coupled with the scarcity of data.

Joint subword segmentation is fairly common. Its use is particularly well motivated when the source and target languages are similar (e.g. for Upper Sorbian-German) and when 'helper languages' are used to compensate for the low-resource scenario (e.g. addition of Czech and English data). However, it is also used in some cases even where there is little lexical overlap, for example for Tamil–English, where the languages do not share the same script, including by some of the top-scoring systems (Shi et al. 2020; Wu et al. 2018). Although few systematic studies are reported, one hypothesis could be that even if different scripts are used there is no disadvantage to sharing segmentation; it could help with named entities and therefore reducing the overall vocabulary size of the model (Ding, Renduchintala, and Duh 2019).

A few works explore alternative morphology-driven segmentation schemes, but without seeing any clear advantage: Scherrer, Grönroos, and Virpioja (2020) find that, for Upper-Sorbian–German, Morfessor can equal the performance of BPE when tuned correctly (but without surpassing it), whereas Sánchez-Cartagena (2018) find gains for Morfessor over BPE. Dhar, Bisazza, and van Noord (2020) have mixed results for Tamil–English when comparing linguistically motivated subword units compared to the use of statistics-based sentencepiece (Kudo and Richardson 2018).

Transliteration and alphabet mapping. Transliteration and alphabet mapping has been principally used in the context of exploiting data from related languages that are written in different scripts. This was particularly present for translation involving Indian languages, which often have their own script. For the Gujarati–English task, many of the top systems used Hindi–English data (see below the paragraph on using other language data) and performed alphabet mapping into the Gujarati script (Li et al. 2019b;

15 https://github.com/anoopkunchukuttan/indic_nlp_library
16 https://github.com/google/sentencepiece
17 https://github.com/rsennrich/subword-nmt

Bawden et al. 2019; Dabre et al. 2019). For Tamil–English, Goyal et al. (2020) found that when using Hindi in a multilingual setup, it helped for Hindi to be mapped into the Tamil script for the Tamil→English direction, but did not bring improvements for English→Tamil. Transliteration was also used in the Kazakh–English task, particularly with the addition of Turkish as higher-resourced language. Toral et al. (2019), a top-scoring system, chose to cyrillise Turkish to increase overlap with Kazakh, whereas Briakou and Carpuat (2019) chose to romanise Kazakh to increase the overlap with Turkish, but only for the Kazakh→English direction.

7.2.3 Model architectures and training. The community has largely converged on a common architecture (the transformer (Vaswani et al. 2017)), although differences can be observed in terms of the number of parameters in the model and certain training parameters. It is particularly tricky to make generalisations about model and training parameters given the dependency on other techniques used (which can affect how much data is available). However a few generalisations can be seen, which we review here.

SMT versus NMT. There is little doubt that NMT has overtaken SMT, even in the low-resource tasks. The majority of submissions use neural MT models and more specifically transformers (rather than recurrent models). Some teams compare SMT and NMT (Dutta et al. 2020; Sen et al. 2019a) with the conclusion that NMT is better when sufficient data is available, including synthetic data. Some teams use SMT only for back-translation, on the basis that SMT can work better (or at least be less susceptible to hallucinating) on the initial training using a very small amount of parallel data. For SMT systems, the most commonly used toolkit is Moses (Koehn et al. 2007), whereas there is a little more variation for NMT toolkits; the most commonly used being Fairseq (Ott et al. 2019), Marian (Junczys-Dowmunt et al. 2018a), OpenNMT (Klein et al. 2017) and Sockeye (Hieber et al. 2020).

Model size. Although systematic comparisons are not always given, some participants did indicate that architecture size was a tuned parameter (Chen et al. 2020), although this can be computationally expensive and therefore not a possibility for all teams. The model sizes chosen for submissions varies, and there is not a clear and direct link between size and model performance. However there are some general patterns worth commenting on. While many of the baseline models are small (corresponding to transformer-base or models with fewer layers), a number of high-scoring teams found that it was possible to train larger models (e.g. deeper or wider) as long as additional techniques were used, such as monolingual pretraining (Wu et al. 2020) or additional data from other languages in a multilingual setup or after synthetic data creation through pivoting (Li et al. 2019b) through a higher-resource language or backtranslation (Chen et al. 2020; Li et al. 2019b). For example the Facebook AI team (Chen et al. 2020), who fine-tuned for model architecture, started with a smaller transformer (3 layers and 8 attention heads) for their supervised English→Tamil baseline, but were able to increase this once backtranslated data was introduced (to 10 layers and 16 attention heads). Although some systems perform well with a transformer-base model (Bawden et al. 2019 for Tamil–English), many of the best systems use larger models, such as the transformer-big (Hernandez and Nguyen 2020; Kocmi 2020; Bei et al. 2019; Wei et al. 2020; Chen et al. 2020).

Alternative neural architectures. Other than variations on the basic transformer model, there are few alternative architectures tested. Wu et al. (2020) tested the addition of

dynamic convolutions to the transformer model following (Wu et al. 2019), which they used along with other wider and deep transformers in model ensembles. However they did not compare the different models. Another alternative form of modelling tested by several teams was factored representations (see Section 5.1.2). Dutta et al. (2020) explored the addition of lemmas and part-of-speech tags for Upper-Sorbian–German but without seeing gains, since the morphological tool used was not adapted to Upper-Sorbian. For Basque–English, Williams et al. (2018) find that source factors indicating the language of the subword can help to improve the baseline system.

Training parameters. Exact training parameters are often not provided, making comparison difficult. Many of the participants do not seem to choose training parameters that are markedly different from the higher-resource settings (Zhang et al. 2020b; Wu et al. 2020).

7.2.4 Using additional data. Much of this survey has been dedicated to approaches for the exploitation of additional resources to compensate for the lack of data for low-resource language pairs: monolingual data (Section 3), multilingual data (Section 4) or other linguistic resources (Section 5). In shared tasks, the following approaches have been shown to be highly effective to boosting performance in low-resource scenarios.

Backtranslation. The majority of high-performing systems carry out some sort of data augmentation, the most common being backtranslation, often used iteratively, although forward translation is also used (Shi et al. 2020; Chen et al. 2020; Zhang et al. 2020b; Wei et al. 2020). For particularly challenging language pairs (e.g. for very low-resource between languages that are not very close), it is important for the initial model that is used to produce the backtranslations to be of sufficiently high quality. For example, some of the top Gujarati–English systems employed pretraining before backtranslation to boost the quality of the initial model (Bawden et al. 2019; Bei et al. 2019). Participants do not always report the number of iterations of backtranslations performed, however those that do often cite the fact that few improvements are seen beyond two iterations (Chen et al. 2020). Tagged backtranslation, whereby a pseudo-token is added to sentences that are backtranslated to distinguish then from genuine parallel data have previously shown to provide improvements (Caswell, Chelba, and Grangier 2019). Several participants report gains thanks to the addition of backtranslation tags (Wu et al. 2020; Chen et al. 2020; Knowles et al. 2020a), although Goyal et al. (2020) find that tagged backtranslation under-performs normal backtranslation in a multilingual setup for Tamil–English.

Synthetic data from other languages. A number of top-performing systems successfully exploit parallel corpora from related languages. The two top-performing systems for Gujarati–English use an Hindi–English parallel corpus to create synthetic Gujarati–English data (Li et al. 2019b; Bawden et al. 2019). Both exploit the fact that there is a high degree of lexical overlap between Hindi and Gujarati once Hindi has been transliterated into Gujarati script. Li et al. (2019b) choosing to transliterate the Hindi side and then to select the best sentences using cross-entropy filtering, and Bawden et al. (2019) choosing to train an Hindi→Gujarati model, which they use to translate the Hindi side of the corpus. Pivoting through a higher-resource related language was also found to be useful for other language pairs: for Kazakh–English, Russian was the language of choice (Li et al. 2019b; Toral et al. 2019; Dabre et al. 2019; Budiwati et al. 2019), for Basque–English, Spanish was used as a pivot (Scherrer 2018; Sánchez-Cartagena 2018), which was found

to be more effective than backtranslation by Scherrer (2018), and was found to benefit from additional filtering by Sánchez-Cartagena (2018).

Transfer-learning using language modelling objectives. The top choices of language modelling objectives are *mBART* (Liu et al. 2020) (used by Chen et al. (2020) and Bawden et al. (2020) for Tamil–English), *XLM* (Conneau and Lample 2019) (used by Bawden et al. (2019) for Gujarati–English, by Laskar et al. (2020) for Hindi–Marathi, and by Kvapilíková, Kocmi, and Bojar (2020) and Dutta et al. (2020) for Upper-Sorbian–German), and *MASS* (Song et al. 2019b) (used by Li et al. (2020) and Singh, Singh, and Bandyopadhyay (2020) for Upper Sorbian–German). Some of the top systems used these language modelling objectives, but their use was not across the board, and pretraining using translation objectives was arguably more common. Given the success of pretrained models in NLP, this could be surprising. A possible explanation for these techniques not being used systematically is that they can be computationally expensive to train from scratch and the constrained nature of the shared tasks means that the participants are discouraged from using pretrained language models.

Transfer learning from other MT systems. Another commonly used technique used by participants was transfer learning involving other language pairs. Many of the teams exploited a high-resource related language pair. For example, for Kazakh–English, pretraining was done using Turkish–English (Briakou and Carpuat 2019) and Russian–English (Kocmi and Bojar 2019), Dabre et al. (2019) pretrained for Gujarati–English using Hindi–English, and Czech–German was used to pretrain for Upper-Sorbian–German (Knowles et al. 2020b).

An alternative but successful approach was to use a high-resource but not necessarily related language pair. For example, the CUNI systems use Czech–English to pretrain Inuktitut (Kocmi 2020) and Gujarati (Kocmi and Bojar 2019), and Bawden et al. (2020) found pretraining on English–German to be as effective as mBART training for Tamil–English. Finally, a number of teams opted for multilingual pretraining, involving the language pair in question and a higher-resource language or several higher-resource languages. Wu et al. (2020) use the mRASP approach: a universal multilingual model involving language data for English to and from Pashto, Khmer, Tamil, Inuktitut, German and Polish, which is then fine-tuned to the individual low-resource language pairs.

Multilingual models. Other than the pretraining strategies mentioned just above, multilingual models feature heavily in shared task submissions. The overwhelmingly most common framework used was the universal encoder-decoder models as proposed by Johnson et al. (2017). Some participants chose to include select (related) languages. Williams et al. (2018) use Spanish to boost Basque–English translation and find that the addition of French data degrades results. Goyal and Sharma (2019) add Hindi as an additional encoder language fro Gujarati–English and for Tamil–English, they test adding Hindi to either the source or target side depending on whether Tamil is the source or target language (Goyal et al. 2020). Other participants choose to use a larger number of languages. Zhang et al. (2020b) train a multilingual system on six Indian languages for Tamil–English and Hokamp, Glover, and Gholipour Ghalandari (2019) choose to train a multilingual model on all WMT languages for Gujarati–English (coming middle in the results table). Upsampling the lower-resourced languages in the multilingual systems is an important factor, whether the multilingual system is used as the main model or for pretraining (Zhang et al. 2020b; Wu et al. 2020).

7.2.5 Model transformation and finalisation. Additional techniques, not specific to low-resource MT, are often applied in the final stages of model construction, and they can provide significant gains to an already trained model. We regroup here knowledge distillation (which we consider as a sort of model transformation) and both model combination and fine-tuning (which can be considered model finalisation techniques).

Knowledge distillation. Knowledge distillation is also a frequently used technique and seems to give minor gains, although is not as frequently used as backtranslation or ensembling. Knowledge distillation (Kim and Rush 2016) leverages a large teacher model to train a student model. The teacher model is used to translate the training data, resulting in synthetic data on the target side. A number of teams apply this iteratively, in combination with backtranslation (Xia et al. 2019) or fine-tuning (Li et al. 2019b). Bei et al. (2019) mix knowledge distillation data with genuine and synthetic parallel data to train a new model to achieve gains in BLEU.

Model combination. Ensembling is the combination of several independently trained models and is used by a large number of participants to get gains over single systems. Several teams seek to create ensembles of diverse models, including deep and wide ones. For example Wu et al. (2020) experiment with ensembling for larger models (larger feed forward dimension and then deeper models), including using different sampling strategies to increase the number of different models. Ensembling generally leads to better results but not always. Wu et al. (2020) found that a 9-model ensemble was best for Khmer and Pashto into English, but they found that for English into Khmer and Pashto, a single model was best. A second way of combining several models is to use an additional model to rerank n-best hypothesis of an initial model. Libovický et al. (2020) attempted right-to-left rescoring (against the normally produced left-to-right hypothesis), but without seeing any gains for Upper-Sorbian–German. Chen et al. (2020) test noisy channel reranking for Tamil–English, but without seeing gains either, although some gains are seen for Inuktitut→English, presumably because of the high-quality monolingual news data available to train a good English language model.

Fine-tuning. Mentioned previously in the context of pretraining, fine-tuning was used in several contexts by a large number of teams. It is inevitably used after pretraining on language model objectives or on other language pairs (see above) to adapt the model to the language direction in question. It is also frequently used on models trained on backtranslated data, by fine-tuning on genuine parallel data (Sánchez-Cartagena, Pérez-Ortiz, and Sánchez-Martínez 2019). A final boost used by a number of top-systems is achieved through fine-tuning to the development set (Shi et al. 2020; Chen et al. 2020; Zhang et al. 2020b; Wei et al. 2020). This was choice not made by all teams, some of which chose to keep it as a held-out set, notably to avoid the risk of overfitting.

8. Conclusion

In the previous section (Section 7), we saw that even if shared tasks rarely offer definitive answers, they do give a good indication of what combination of techniques can deliver state-of-the-art systems for particular language pairs. This look at what methods are commonly used in practise, hopefully providing some perspective on the research we have covered in this survey.

In this survey we have looked at the entire spectrum of scientific effort in the field: from data sourcing and creation (Section 2), leveraging all types of available data, monolingual data (Section 3), multilingual data (Section 4) and other linguistic resources (Section 5), to improving model robustness, training and inference (Section 6).

Thanks to large-scale and also more focused efforts to identify, scrape and filter parallel data from the web, some language pairs have moved quickly from being considered low-resource to now being considered more medium-resourced (e.g. Romanian–English, Turkish–English and Hindi–English). The ability of deep learning models to learn from monolingual data (Sennrich, Haddow, and Birch 2016a; Cheng et al. 2016; He et al. 2016) and related languages (Liu et al. 2020; Conneau and Lample 2019; Pan et al. 2021) has had a big impact on this field. However, our models still have a voracious appetite for data, far hungrier than any human learner, and recent results on very low-resource and distant languages (Barrault et al. 2020) show that we still have a long way to go. Looking forward, we now discuss a number of key areas for future work.

Collaboration with Language Communities. Recent efforts by language communities (Nekoto et al. 2020; Onome Orife et al. 2020) to highlight their work and their lack of access to opportunities to develop expertise and progress language technology has resulted in bringing valuable talent and linguistic knowledge into our field. It is very clear that we need to work together with speakers of low-resource and endangered languages to understand their challenges, their needs and to create knowledge and resources which can help them benefit from progress in our field, and also help their communities overcome language barriers.

Massive Multilingual Models. The striking success of multilingual pre-trained models such as mBART (Liu et al. 2020) and mRASP (Pan et al. 2021) still needs further investigation. We should be able to answer questions such as whether the gains are more from the size of models, or from the number of languages the models are trained on, or is it from the sheer amount of data used. There are also questions about how to handle languages that are not included in the pretrained model.

Incorporating external knowledge. We will never have enough parallel data, and for many language pairs the situation is harder due to a lack of high-resourced related languages and a lack of monolingual data. We know that parallel data is not an efficient way to learn to translate. We have not fully explored questions such as what is a more efficient way of encoding translation knowledge – bilingual lexicons, grammars or ontologies – or indeed what type of knowledge is most helpful in creating MT systems and how to gather it. Further work looking at how we can best incorporate these resources is also needed: should they be incorporated directly into the model or should we use them to create synthetic parallel data?

Robustness. Modern MT systems, being large neural networks, tend to incur substantial quality degradation as the distribution of data encountered by the production system

becomes more and more different than the distribution of the training data (Lapuschkin et al. 2019; Hupkes et al. 2019; Geirhos et al. 2020). This commonly occurs in translation applications where the language domains, topics, and registers can be extremely varied and quickly change over time. Especially in low-resource settings, we are often limited to old training corpora from a limited sets of domains. Therefore it is of great importance to find ways to make the systems robust to distribution shifts. This is a big research direction in general machine learning, but it has a specific angle in MT due to the possibility of producing *hallucinations* (Martindale et al. 2019; Raunak, Menezes, and Junczys-Dowmunt 2021) that might mislead the user. We need to find ways to make the systems detect out-of-distribution conditions and ideally avoid producing hallucinations or at least warn the user that the output might be misleading.

References

Agić, Željko and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Association for Computational Linguistics, Florence, Italy.

Aharoni, Roee, 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.

Aji, Alham Fikri, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Association for Computational Linguistics, Online.

Arivazhagan, Naveen, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. 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.

Artetxe, Mikel, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation. In *Proceeedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada.

Artetxe, Mikel and Holger Schwenk. 2019a. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Association for Computational Linguistics, Florence, Italy.

Artetxe, Mikel and Holger Schwenk. 2019b. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Arthaud, Farid, Rachel Bawden, and Alexandra Birch. 2021. Few-shot learning through contextual data augmentation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Online.

Arthur, Philip, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Association for Computational Linguistics, Austin, Texas.

Ataman, Duygu, Wilker Aziz, and Alexandra Birch. 2019. A latent morphology model for open-vocabulary neural machine translation. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, LA, USA.

Ataman, Duygu, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically Motivated Vocabulary Reduction for Neural Machine Translation from Turkish to English. *The Prague Bulletin of Mathematical Linguistics*, 108:331–342.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.

Bañón, Marta, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Association for Computational Linguistics, Online.

Barrault, Loic, Ondřej Bojar, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Maller, 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, Association for Computational

Linguistics, Florence, Italy.
Barrault, Loïc, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Association for Computational Linguistics, Online.
Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Association for Computational Linguistics, Florence, Italy.
Bawden, Rachel, Alexandra Birch, Radina Dobreva, Arturo Oncevay, Antonio Valerio Miceli Barone, and Philip Williams. 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, Association for Computational Linguistics, Online.
Bawden, Rachel, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019. The University of Edinburgh's submissions to the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 103–115, Association for Computational Linguistics, Florence, Italy.
Baziotis, Christos, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Association for Computational Linguistics, Online.
Bei, Chao, Hao Zong, Conghu Yuan, Qingming Liu, and Baoyong Fan. 2019. GTCOM neural machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 116–121, Association for Computational Linguistics, Florence, Italy.
Bengio, Samy, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, Curran Associates, Inc.
Bertoldi, Nicola and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Association for Computational Linguistics, Athens, Greece.
Bhosale, Shruti, Kyra Yee, Sergey Edunov, and Michael Auli. 2020. Language models not just for pre-training: Fast online neural noisy channel modeling. In *Proceedings of the Fifth Conference on Machine Translation*, pages 584–593, Association for Computational Linguistics, Online.
Birch, Alexandra, Barry Haddow, Antonio Valerio Miceli Barone, Jindrich Helcl, Jonas Waldendorf, Felipe Sánchez Martínez, Mikel Forcada, Víctor Sánchez Cartagena, Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Wilker Aziz, Lina Murady, Sevi Sariisik, Peggy van der Kreeft, and Kay Macquarrie. 2021. Surprise language challenge: Developing a neural machine translation system between Pashto and English in two months. In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)*, pages 92–102, Association for Machine Translation in the Americas, Virtual.
Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.
Bojar, Ondřej and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Association for Computational Linguistics, Edinburgh, Scotland.
Bojar, Ondřej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Association for Computational Linguistics, Belgium, Brussels.

Bowman, Samuel R., Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Association for Computational Linguistics, Berlin, Germany.

Briakou, Eleftheria and Marine Carpuat. 2019. The University of Maryland's Kazakh-English Neural Machine Translation System at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 134–140, Association for Computational Linguistics, Florence, Italy.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, Curran Associates, Inc.

Buck, Christian, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the Common Crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3579–3584, European Language Resources Association (ELRA), Reykjavik, Iceland.

Budiwati, Sari Dewi, Al Hafiz Akbar Maulana Siagian, Tirana Noor Fatyanosa, and Masayoshi Aritsugi. 2019. DBMS-KU Interpolation for WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 141–146, Association for Computational Linguistics, Florence, Italy.

Burlot, Franck, Mercedes García-Martínez, Loïc Barrault, Fethi Bougares, and François Yvon. 2017. Word Representations in Factored Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, pages 20–31, Association for Computational Linguistics, Copenhagen, Denmark.

Caswell, Isaac, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, International Committee on Computational Linguistics, Barcelona, Spain (Online).

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

Caswell, Isaac, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *CoRR*, abs/2103.12028.

Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *Proceedings of the 11th International Workshop on Spoken Language Translation*, Lake Tahoe, CA, USA.

Chen, Peng-Jen, Ann Lee, Changhan Wang, Naman Goyal, Angela Fan, Mary

Williamson, and Jiatao Gu. 2020. Facebook AI's WMT20 news translation task submission. In *Proceedings of the Fifth Conference on Machine Translation*, pages 113–125, Association for Computational Linguistics, Online.

Cheng, Yong, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Association for Computational Linguistics, Berlin, Germany.

Choshen, Leshem, Lior Fox, Zohar Aizenbud, and Omri Abend. 2020. On the Weaknesses of Reinforcement Learning for Neural Machine Translation. In *Proceedings of the 8th International Conference on Learning Representations*, Online.

Christodouloupoulos, Christos and Mark Steedman. 2015. A Massively Parallel Corpus: The Bible in 100 Languages. *Language Resources and Evaluation*, 49(2):375–395.

Chronopoulou, Alexandra, Dario Stojanovski, and Alexander Fraser. 2021. Improving the Lexical Ability of Pretrained Language Models for Unsupervised Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 173–180, Association for Computational Linguistics, Online.

Chronopoulou, Alexandra, Dario Stojanovski, Viktor Hangya, and Alexander Fraser. 2020. The LMU Munich System for the WMT 2020 Unsupervised Machine Translation Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1084–1091, Association for Computational Linguistics, Online.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Association for Computational Linguistics, Online.

Conneau, Alexis and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32, Curran Associates, Inc.

Conneau, Alexis, Guillaume Lample, Marc'aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada.

Cover, Thomas M. and Joy A. Thomas. 2006. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA.

Currey, Anna, Antonio Valerio Miceli Barone, and Kenneth Heafield. 2017. Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156, Association for Computational Linguistics, Copenhagen, Denmark.

Dabre, Raj, Kehai Chen, Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. NICT's Supervised Neural Machine Translation Systems for the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 168–174, Association for Computational Linguistics, Florence, Italy.

Dabre, Raj, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Comprehensive Survey of Multilingual Neural Machine Translation. *CoRR*, abs/2001.01115.

Dabre, Raj, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage Fine-Tuning for Low-Resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416, Hong Kong, China.

Dabre, Raj, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286, The National University (Phillippines).

Dandapat, Sandipan and Christian Federmann. 2018. Iterative Data Augmentation for Neural Machine Translation: a Low Resource Case Study for English-Telugu. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages

287–292, Alacant, Spain.

Daumé, Hal and Daniel Marcu. 2005. Learning as Search Optimization: Approximate Large Margin Methods for Structured Prediction. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 169–176, Association for Computing Machinery, New York, NY, USA.

Denkowski, Michael and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Association for Computational Linguistics, Edinburgh, Scotland.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota.

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

Di Gangi, Mattia Antonino and Marcello Federico. 2017. Monolingual Embeddings for Low Resourced Neural Machine Translation. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 97–104, Tokyo, Japan.

Ding, Shuoyang, Adithya Renduchintala, and Kevin Duh. 2019. A Call for Prudent Choice of Subword Merge Operations in Neural Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 204–213, European Association for Machine Translation, Dublin, Ireland.

Dinu, Georgiana, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training Neural Machine Translation to Apply Terminology Constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Association for Computational Linguistics, Florence, Italy.

Dong, Daxiang, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 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, Association for Computational Linguistics, Beijing, China.

Duan, Xiangyu, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. 2020. Bilingual Dictionary Based Neural Machine Translation without Using Parallel Sentences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579, Association for Computational Linguistics, Online.

Dutta, Sourav, Jesujoba Alabi, Saptarashmi Bandyopadhyay, Dana Ruiter, and Josef van Genabith. 2020. UdS-DFKI@WMT20: Unsupervised MT and Very Low Resource Supervised MT for German-Upper Sorbian. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1092–1098, Association for Computational Linguistics, Online.

Edman, Lukas, Antonio Toral, and Gertjan van Noord. 2020. Low-resource unsupervised NMT: Diagnosing the problem and providing a linguistically motivated solution. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 81–90, European Association for Machine Translation, Lisboa, Portugal.

Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier. 2018a. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Association for Computational Linguistics, Brussels, Belgium.

Edunov, Sergey, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018b. Classical Structured Prediction Losses for Sequence to Sequence Learning. 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 355–364, Association for Computational Linguistics, New Orleans, Louisiana.

Eikema, Bryan and Wilker Aziz. 2019. Auto-Encoding Variational Neural Machine Translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 124–141,

Association for Computational Linguistics, Florence, Italy.
Eikema, Bryan and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, International Committee on Computational Linguistics, Barcelona, Spain (Online).
Ethayarajh, Kawin and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. *CoRR*. ArXiv: 2009.13888v1.
Fadaee, Marzieh, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Association for Computational Linguistics, Vancouver, BC, Canada.
Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-Centric Multilingual Machine Translation. *CoRR*, abs/2010.11125.
Feng, Yang, Shiyue Zhang, Andi Zhang, Dong Wang, and Andrew Abel. 2017. Memory-augmented Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1390–1399, Association for Computational Linguistics, Copenhagen, Denmark.
Finn, Chelsea, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, PMLR, International Convention Centre, Sydney, Australia.
Firat, Orhan, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, Association for Computational Linguistics, San Diego, California.
Forcada, Mikel L. and Francis M. Tyers. 2016. Apertium: a free/open source platform for machine translation and basic language technology. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation: Projects/Products*, pages 127–144, Baltic Journal of Modern Computing, Riga, Latvia.
Fraser, Alexander. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Association for Computational Linguistics, Online.
Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *CoRR*. ArXiv: 2104.14478v1.
Garcia, Xavier, Aditya Siddhant, Orhan Firat, and Ankur Parikh. 2021. Harnessing multilinguality in unsupervised machine translation for rare languages. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1126–1137, Association for Computational Linguistics, Online.
Garcia-Martinez, Mercedes, Loïc Barrault, and Fethi Bougares. 2016. Factored Neural Machine Translation Architectures. In *Proceedings of the 13th International Workshop on Spoken Language Translation*, Seattle, United States.
Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*.
Gibadullin, Ilshat, Aidar Valeev, Albina Khusainova, and Adil Khan. 2019. A survey of methods to leverage monolingual data in low-resource neural machine translation. *CoRR*, abs/1910.00373.
Goel, Vaibhava and William J Byrne. 2000. Minimum Bayes-Risk Automatic Speech Recognition. *Computer Speech and Language*, 14(2):115–135.
Goodfellow, Ian J, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2014. An Empirical Investigation of Catastrophic Forgetting in Gradient-Based Neural Networks. In *Proceedings of the*

2014 International Conference on Learning Representations, Banff, AB, Canada.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

Goyal, Vikrant, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient Neural Machine Translation for Low-Resource Languages via Exploiting Related Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Association for Computational Linguistics, Online.

Goyal, Vikrant, Anoop Kunchukuttan, Rahul Kejriwal, Siddharth Jain, and Amit Bhagwat. 2020. Contact Relatedness can help improve multilingual NMT: Microsoft STCI-MT @ WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 202–206, Association for Computational Linguistics, Online.

Goyal, Vikrant and Dipti Misra Sharma. 2019. The IIIT-H Gujarati-English Machine Translation System for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 191–195, Association for Computational Linguistics, Florence, Italy.

Grönroos, Stig-Arne, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin City University and Association for Computational Linguistics, Dublin, Ireland.

Gu, Jiatao, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Association for Computational Linguistics, Brussels, Belgium.

Gülçehre, Çaglar, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Guzmán, Francisco, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Association for Computational Linguistics, Hong Kong, China.

Ha, Thanh-Le, Jan Niehues, and Alexander Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, US.

Haddow, Barry and Faheem Kirefu. 2020. PMIndia - A Collection of Parallel Corpora of Languages of India. *CoRR*, abs/2001.09907.

Hasan, Tahmid, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Association for Computational Linguistics, Online.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *CoRR*, abs/1803.05567.

He, Di, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual Learning for Machine Translation. In *Advances in Neural Information Processing Systems*, volume 29, Curran Associates, Inc.

He, Xuanli, Gholamreza Haffari, and Mohammad Norouzi. 2020. Dynamic Programming Encoding for Subword Segmentation in Neural Machine

Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3042–3051, Association for Computational Linguistics, Online.

Hernandez, François and Vincent Nguyen. 2020. The Ubiqus English-Inuktitut System for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 213–217, Association for Computational Linguistics, Online.

Hieber, Felix, Tobias Domhan, Michael Denkowski, and David Vilar. 2020. Sockeye 2: A Toolkit for Neural Machine Translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 457–458, European Association for Machine Translation, Lisboa, Portugal.

Hoang, Vu Cong Duy, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 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.

Hokamp, Chris, John Glover, and Demian Gholipour Ghalandari. 2019. Evaluating the supervised and zero-shot performance of multi-lingual translation models. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 209–217, Association for Computational Linguistics, Florence, Italy.

Hupkes, Dieuwke, Verna Dankers, Mathijs Mul, and Elia Bruni. 2019. The compositionality of neural networks: integrating symbolism and connectionism. *CoRR*, abs/1908.08351.

Huszar, Ferenc. 2015. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *CoRR*, abs/1511.05101.

Jean, Sébastien, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal Neural Machine Translation Systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Association for Computational Linguistics, Lisbon, Portugal.

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

Junczys-Dowmunt, Marcin. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Association for Computational Linguistics, Belgium, Brussels.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018a. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Association for Computational Linguistics, Melbourne, Australia.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Shubha Guha, and Kenneth Heafield. 2018b. Approaching Neural Grammatical Error Correction as a Low-Resource Machine Translation Task. 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 595–606, Association for Computational Linguistics, New Orleans, Louisiana.

Karakanta, Alina, Atul Kr. Ojha, Chao-Hong Liu, Jonathan Washington, Nathaniel Oco, Surafel Melaku Lakew, Valentin Malykh, and Xiaobing Zhao, editors. 2019. *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*. European Association for Machine Translation, Dublin, Ireland.

Kim, Yoon and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Association for Computational Linguistics, Austin, Texas.

Kim, Yunsu, Yingbo Gao, and Hermann Ney. 2019. Effective cross-lingual transfer of neural machine translation models without shared vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257.

Kim, Yunsu, Miguel Graça, and Hermann Ney. 2020. When and Why is Unsupervised Neural Machine Translation Useless? In *Proceedings of the 22nd Annual Conference of the European Association for*

Machine Translation, pages 35–44, European Association for Machine Translation, Lisboa, Portugal.

Kim, Yunsu, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between Non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China.

Kingma, Diederik P. and Max Welling. 2014. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, Banff, AB, Canada.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Association for Computational Linguistics, Vancouver, BC, Canada.

Knowles, Rebecca, Samuel Larkin, Darlene Stewart, and Patrick Littell. 2020a. NRC Systems for Low Resource German-Upper Sorbian Machine Translation 2020: Transfer Learning with Lexical Modifications. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1112–1122, Association for Computational Linguistics, Online.

Knowles, Rebecca, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2020b. NRC Systems for the 2020 Inuktitut-English News Translation Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 156–170, Association for Computational Linguistics, Online.

Ko, Wei-Jen, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource NMT models to translate low-resource related languages without parallel data. 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 802–812, Association for Computational Linguistics, Online.

Kocmi, Tom. 2020. CUNI Submission for the Inuktitut Language in WMT News 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 171–174, Association for Computational Linguistics, Online.

Kocmi, Tom and Ondřej Bojar. 2018. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Association for Computational Linguistics, Brussels, Belgium.

Kocmi, Tom and Ondřej Bojar. 2019. CUNI Submission for Low-Resource Languages in WMT News 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 234–240, Association for Computational Linguistics, Florence, Italy.

Koehn, Philipp, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 724–740, Association for Computational Linguistics, Online.

Koehn, Philipp, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Association for Computational Linguistics, Florence, Italy.

Koehn, Philipp and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Association for Computational Linguistics, Prague, Czech Republic.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Association for Computational Linguistics, Prague, Czech Republic.

Koehn, Philipp, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of*

the Third Conference on Machine Translation: Shared Task Papers, pages 726–739, Association for Computational Linguistics, Belgium, Brussels.

Koehn, Philipp and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Association for Computational Linguistics, Vancouver, BC, Canada.

Koehn, Philipp and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, Association for Computational Linguistics, New York City.

Koehn, Philipp, Franz J. Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133, Edmonton Canada.

Kudo, Taku. 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, Association for Computational Linguistics, Melbourne, Australia.

Kudo, Taku and John Richardson. 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, Association for Computational Linguistics, Brussels, Belgium.

Kumaraswamy, P. 1980. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1):79–88.

Kunchukuttan, Anoop. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Kvapilíková, Ivana, Tom Kocmi, and Ondřej Bojar. 2020. CUNI Systems for the Unsupervised and Very Low Resource Translation Task in WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1123–1128, Association for Computational Linguistics, Online.

Lake, Brenden M. 2019. Compositional generalization through meta sequence-to-sequence learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 9791–9801, Curran Associates, Inc.

Lample, Guillaume, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised Machine Translation Using Monolingual Corpora Only. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada.

Lample, Guillaume, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Association for Computational Linguistics, Brussels, Belgium.

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *CoRR*, abs/1902.10178.

Laskar, Sahinur Rahman, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. Hindi-Marathi Cross Lingual Model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 396–401, Association for Computational Linguistics, Online.

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Association for Computational Linguistics, Brussels, Belgium.

LeCun, Yann, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Li, Bei, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao, and Jingbo Zhu. 2019a. The niutrans machine translation systems for wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266, Association for Computational Linguistics, Florence, Italy.

Li, Bei, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao, and Jingbo Zhu. 2019b. The NiuTrans machine translation systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 257–266, Association for Computational Linguistics, Florence, Italy.

Li, Zuchao, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2020. SJTU-NICT's Supervised and Unsupervised Neural Machine Translation Systems for the WMT20 News Translation Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 218–229, Association for Computational Linguistics, Online.

Libovický, Jindřich. 2021. Jindřich's Blog – Machine Translation Weekly 86: The Wisdom of the WMT Crowd. Online, Accessed: 24.07. 2021.

Libovický, Jindřich, Viktor Hangya, Helmut Schmid, and Alexander Fraser. 2020. The LMU Munich System for the WMT20 Very Low Resource Supervised MT Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1104–1111, Association for Computational Linguistics, Online.

Lignos, Constantine. 2010. Learning from unseen data. In *Proceedings of the Morpho Challenge 2010 Workshop*, pages 35–38, Aalto University, Espoo, Finland.

Lin, Yu-Hsiang, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Association for Computational Linguistics, Florence, Italy.

Lin, Zehui, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. *arXiv preprint arXiv:2010.03142*.

Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lo, Chi-kiu. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Association for Computational Linguistics, Florence, Italy.

Louizos, Christos, Max Welling, and Diederik P. Kingma. 2018. Learning sparse neural networks through l_0 regularization. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada.

Luong, Minh-Thang, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task Sequence to Sequence Learning. In *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico.

Mager, Manuel, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Association for Computational Linguistics, Online.

Marchisio, Kelly, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 569–581, Association for Computational Linguistics, Online.

Martindale, Marianna, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying fluently inadequate output in neural and statistical machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 233–243, European Association for Machine Translation, Dublin, Ireland.

Mayer, Thomas and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, European Language Resources Association (ELRA), Reykjavik, Iceland.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781:arXiv:1301.3781.

Moore, Robert C. and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Association for Computational Linguistics, Uppsala, Sweden.

Mueller, Aaron, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An Analysis of Massively Multilingual Neural Machine Translation for Low-Resource Languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3710–3718, European Language Resources Association, Marseille, France.

Muller, Benjamin, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Association for Computational Linguistics, Online.

Müller, Mathias, Annette Rios, and Rico Sennrich. 2020. Domain Robustness in Neural Machine Translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Association for Machine Translation in the Americas, Virtual.

Murthy, Rudra, Anoop Kunchukuttan, and Pushpak Bhattacharyya. 2019. Addressing word-order divergence in multilingual neural machine translation for extremely low resource languages. 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 3868–3873, Minneapolis, Minnesota.

Nădejde, Maria, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. Predicting target language CCG supertags improves neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 68–79, Association for Computational Linguistics, Copenhagen, Denmark.

Nakazawa, Toshiaki, Nobushige Doi, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Yusuke Oda, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2019. Overview of the 6th workshop on Asian translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 1–35, Association for Computational Linguistics, Hong Kong, China.

Nakazawa, Toshiaki, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. 2021. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 1–45, Association for Computational Linguistics, Online.

Nakazawa, Toshiaki, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, and Sadao Kurohashi. 2020. Overview of the 7th workshop on Asian translation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 1–44, Association for Computational Linguistics, Suzhou, China.

Nakazawa, Toshiaki, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on Asian translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Association for Computational Linguistics, Hong Kong.

Naskar, Subhajit, Amirmohammad Rooshenas, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2020. Energy-based reranking: Improving neural machine translation using energy-based models. *CoRR*, abs/2009.13267.

Neishi, Masato, Jin Sakuma, Satoshi Tohda, Shonosuke Ishiwatari, Naoki Yoshinaga, and Masashi Toyoda. 2017. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In *Proceedings of the 4th Workshop on Asian Translation*

(WAT2017), pages 99–109, Asian Federation of Natural Language Processing, Taipei, Taiwan.

Nekoto, Wilhelmina, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Association for Computational Linguistics, Online.

Neubig, Graham and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Association for Computational Linguistics, Brussels, Belgium.

Nguyen, Toan Q. and David Chiang. 2017. Transfer Learning across Low-Resource, Related Languages for Neural Machine Translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Asian Federation of Natural Language Processing, Taipei, Taiwan.

Niehues, Jan, Ronaldo Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcelo Federico. 2018. The IWSLT 2018 Evaluation Campaign. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 2–5, Bruges, Belgium.

Niehues, Jan and Eunah Cho. 2017. Exploiting linguistic resources for neural machine translation using multi-task learning. In *Proceedings of the Second Conference on Machine Translation*, pages 80–89, Association for Computational Linguistics, Copenhagen, Denmark.

Niu, Xing, Weijia Xu, and Marine Carpuat. 2019. Bi-directional differentiable input reconstruction for low-resource 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 442–448, Association for Computational Linguistics, Minneapolis, Minnesota.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Association for Computational Linguistics, Sapporo, Japan.

Ojha, Atul Kr., Valentin Malykh, Alina Karakanta, and Chao-Hong Liu. 2020. Findings of the LoResMT 2020 shared task on zero-shot for low-resource languages. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 33–37, Association for Computational Linguistics, Suzhou, China.

Onome Orife, Iroro Fred, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Ktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane–Machine Translation For Africa. In *AfricaNLP Workshop*, International Conference on Learning Representations (ICLR).

Ortiz Suárez, Pedro Javier, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom.

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), pages 48–53, Association for Computational Linguistics, Minneapolis, Minnesota.

Pan, Xiao, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

Pavlick, Ellie, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2(Feb):79–92.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Association for Computational Linguistics, Doha, Qatar.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. 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 2227–2237, Association for Computational Linguistics, New Orleans, Louisiana.

Philip, Jerin, Shashank Siripragada, Vinay P Namboodiri, and CV Jawahar. 2021. Revisiting Low Resource Status of Indian Languages in Machine Translation. In *Proceedings of the 3rd ACM India Joint International Conference on Data Science & Management of Data*, pages 178–187, Online.

Platanios, Emmanouil Antonios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435, Association for Computational Linguistics, Brussels, Belgium.

Popel, Martin, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11.

Post, Matt, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Association for Computational Linguistics, Montréal, Canada.

Provilkov, Ivan, Dmitrii Emelianenko, and Elena Voita. 2019. Bpe-dropout: Simple and effective subword regularization. *CoRR*, abs/1910.13267.

Provilkov, Ivan, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-Dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online.

Qi, Ye, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, Association for Computational Linguistics, New Orleans, Louisiana.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Ramachandran, Prajit, Peter Liu, and Quoc Le. 2017. Unsupervised pretraining for sequence to sequence learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 383–391, Association for Computational Linguistics, Copenhagen, Denmark.

Ramesh, Gowtham, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2021. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *CoRR*, abs/2104.05596.

Ranathunga, Surangika, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for Low-Resource languages: A survey. *CoRR*,

abs/2106.15115.
Ranzato, Marc'Aurelio, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence Level Training with Recurrent Neural Networks.
Raunak, Vikas, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. *CoRR*, abs/2104.06683.
Rezende, Danilo and Shakir Mohamed. 2015. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, PMLR, Lille, France.
Rezende, Danilo Jimenez, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, PMLR, Bejing, China.
Saleva, Jonne and Constantine Lignos. 2021. The Effectiveness of Morphology-aware Segmentation in Low-Resource Neural Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 164–174, Association for Computational Linguistics, Online.
Sánchez-Cartagena, Víctor M., Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 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*, Association for Computational Linguistics, Brussels, Belgium.
Sánchez-Cartagena, Víctor M., Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2019. The Universitat d'Alacant Submissions to the English-to-Kazakh News Translation Task at WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 356–363, Association for Computational Linguistics, Florence, Italy.
Sánchez-Cartagena, Víctor M., Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2020. Understanding the effects of word-level linguistic annotations in under-resourced neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3938–3950, International Committee on Computational Linguistics, Barcelona, Spain (Online).
Sánchez-Cartagena, Víctor M. 2018. Prompsit's Submission to the IWSLT 2018 Low Resource Machine Translation Task. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 95–103, Bruges, Belgium.
Sánchez-Martínez, Felipe, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, Mikel L. Forcada, Miquel Esplà-Gomis, Andrew Secker, Susie Coleman, and Julie Wall. 2020. An English-Swahili parallel corpus and its use for neural machine translation in the news domain. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 299–308, European Association for Machine Translation, Lisboa, Portugal.
Santoro, Adam, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-Learning with Memory-Augmented Neural Networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pages 1842–1850, New York, NY, USA.
Scherrer, Yves. 2018. The University of Helsinki submissions to the IWSLT 2018 low-resource translation task. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 83–88, Bruges, Belgium.
Scherrer, Yves, Stig-Arne Grönroos, and Sami Virpioja. 2020. The University of Helsinki and Aalto University submissions to the WMT 2020 news and low-resource translation tasks. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1129–1138, Association for Computational Linguistics, Online.
Schmidhuber, J. 1992. Learning to control fast-weight memories: An alternative to dynamic recurrent networks. *Neural Computation*, 4(1):131–139.
Schulz, Philip, Wilker Aziz, and Trevor Cohn. 2018. A stochastic decoder for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1243–1252, Association for Computational Linguistics, Melbourne, Australia.
Schwenk, Holger, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs

from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Association for Computational Linguistics, Online.

Schwenk, Holger, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the WEB. *CoRR*, abs/1911.04944.

Sen, Sukanta, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019a. IITP-MT System for Gujarati-English News Translation Task at WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 407–411, Association for Computational Linguistics, Florence, Italy.

Sen, Sukanta, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019b. Multilingual Unsupervised NMT using Shared Encoder and Language-Specific Decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Association for Computational Linguistics, Florence, Italy.

Sennrich, Rico and Barry Haddow. 2016. Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Association for Computational Linguistics, Berlin, Germany.

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

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Association for Computational Linguistics, Berlin, Germany.

Sennrich, Rico and Martin Volk. 2011. Iterative, MT-based Sentence Alignment of Parallel Texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Northern European Association for Language Technology (NEALT).

Sennrich, Rico and Biao Zhang. 2019. Revisiting Low-Resource Neural Machine Translation: A Case Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Association for Computational Linguistics, Florence, Italy.

Setiawan, Hendra, Matthias Sperber, Udhyakumar Nallasamy, and Matthias Paulik. 2020. Variational Neural Machine Translation with Normalizing Flows. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7771–7777, Association for Computational Linguistics, Online.

Shen, Shiqi, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Association for Computational Linguistics, Berlin, Germany.

Shi, Tingxun, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, and Jie Hao. 2020. OPPO's machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 282–292, Association for Computational Linguistics, Online.

Singh, Keshaw. 2020. Adobe AMPS's submission for very low resource supervised translation task at WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1144–1149, Association for Computational Linguistics, Online.

Singh, Salam Michael, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2020. The NITS-CNLP System for the Unsupervised MT Task at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1139–1143, Association for Computational Linguistics, Online.

Siripragada, Shashank, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3743–3751, European Language Resources Association, Marseille, France.

Song, Kai, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019a. Code-Switching for Enhancing NMT with

pre-Specified 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 449–459, Minneapolis, Minnesota.

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019b. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, California.

Stahlberg, Felix and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362, Association for Computational Linguistics, Hong Kong, China.

Stahlberg, Felix, James Cross, and Veselin Stoyanov. 2018. Simple Fusion: Return of the Language Model. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 204–211, Association for Computational Linguistics, Brussels, Belgium.

Su, Jinsong, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. 2018. Variational recurrent neural machine translation. *CoRR*, abs/1801.05119.

Tamchyna, Aleš, Marion Weller-Di Marco, and Alexander Fraser. 2017. Modeling Target-Side Inflection in Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, pages 32–42, Association for Computational Linguistics, Copenhagen, Denmark.

Tang, Yuqing, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, European Language Resources Association (ELRA), Istanbul, Turkey.

Tiedemann, Jörg. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Association for Computational Linguistics, Online.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Association for Computational Linguistics, Belgium, Brussels.

Toral, Antonio, Lukas Edman, Galiya Yeshmagambetova, and Jennifer Spenader. 2019. Neural Machine Translation for English–Kazakh with Morphological Segmentation and Synthetic Data. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 386–392, Association for Computational Linguistics, Florence, Italy.

Tracey, Jennifer, Stephanie Strassel, Ann Bies, Zhiyi Song, Michael Arrigo, Kira Griffitt, Dana Delgado, Dave Graff, Seth Kulick, Justin Mott, and Neil Kuster. 2019. Corpus building for low resource languages in the DARPA LORELEI program. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 48–55, European Association for Machine Translation, Dublin, Ireland.

Uszkoreit, Jakob, Jay Ponte, Ashok Popat, and Moshe Dubiner. 2010. Large Scale Parallel Document Mining for Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1101–1109, Beijing, China.

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 590–596, Borovets, Bulgaria.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *31st Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach, CA, USA.

Vinyals, Oriol, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, volume 29, Curran Associates, Inc.

Wang, Chaojun and Rico Sennrich. 2020. On exposure bias, wallucination and domain shift in neural machine translation. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3544–3552, Association for Computational Linguistics, Online.

Wang, Rui, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation.

Wei, Daimeng, Hengchao Shang, Zhanglin Wu, Zhengzhe Yu, Liangyou Li, Jiaxin Guo, Minghan Wang, Hao Yang, Lizhi Lei, Ying Qin, and Shiliang Sun. 2020. HW-TSC's Participation in the WMT 2020 News Translation Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 293–299, Association for Computational Linguistics, Online.

Williams, Philip, Marcin Chochowski, Pawel Przybysz, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2018. Samsung and University of Edinburgh's System for the IWSLT 2018 Low Resource MT Task. In *Proceedings of the 15th International Workshop on Spoken Language Translation*, pages 118–123, Bruges, Belgium.

Williams, Ronald J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256.

Wiseman, Sam and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Association for Computational Linguistics, Austin, Texas.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Association for Computational Linguistics, Online.

Wu, Felix, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *Proceedings of the 7th International Conference on Learning Representations*, New Orleans, USA.

Wu, Lijun, Xu Tan, Di He, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018. Beyond error propagation in neural machine translation: Characteristics of language also matter. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3602–3611, Association for Computational Linguistics, Brussels, Belgium.

Wu, Liwei, Xiao Pan, Zehui Lin, Yaoming ZHU, Mingxuan Wang, and Lei Li. 2020. The Volctrans Machine Translation System for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 305–312, Association for Computational Linguistics, Online.

Xia, Yingce, Xu Tan, Fei Tian, Fei Gao, Di He, Weicong Chen, Yang Fan, Linyuan Gong, Yichong Leng, Renqian Luo, Yiren Wang, Lijun Wu, Jinhua Zhu, Tao Qin, and Tie-Yan Liu. 2019. Microsoft Research Asia's Systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 424–433, Association for Computational Linguistics, Florence, Italy.

Xu, Hainan and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Association for Computational Linguistics, Copenhagen, Denmark.

Xu, Nuo, Yinqiao Li, Chen Xu, Yanyang Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2019. Analysis of Back-Translation methods for Low-Resource neural machine translation. In *Natural Language Processing and Chinese Computing*, pages 466–475, Springer International Publishing.

Xu, Weijia, Xing Niu, and Marine Carpuat. 2019. Differentiable sampling with flexible reference word order for 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 2047–2053, Association for Computational Linguistics, Minneapolis, Minnesota.

Yang, Jiacheng, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Yong Yu, Weinan Zhang, and Lei Li. 2020. Towards Making the Most of BERT in Neural Machine Translation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 9378–9385, New York, NY, USA.

Zaremba, Wojciech and Ilya Sutskever. 2015. Reinforcement learning neural turing machines. *CoRR*, abs/1505.00521.

Zaremoodi, Poorya, Wray Buntine, and Gholamreza Haffari. 2018. Adaptive knowledge sharing in multi-task learning: Improving low-resource neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 656–661, Association for Computational Linguistics, Melbourne, Australia.

Zhang, Biao, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Association for Computational Linguistics, Online.

Zhang, Biao, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Association for Computational Linguistics, Austin, Texas.

Zhang, Jiajun and Chengqing Zong. 2016. Exploiting Source-side Monolingual Data in Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Association for Computational Linguistics, Austin, Texas.

Zhang, Yuhao, Ziyang Wang, Runzhe Cao, Binghao Wei, Weiqiao Shan, Shuhan Zhou, Abudurexiti Reheman, Tao Zhou, Xin Zeng, Laohu Wang, Yongyu Mu, Jingnan Zhang, Xiaoqian Liu, Xuanjun Zhou, Yinqiao Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2020b. The NiuTrans Machine Translation Systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 338–345, Association for Computational Linguistics, Online.

Zhong, Xing Jie and David Chiang. 2020. Look it up: Bilingual and monolingual dictionaries improve neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 538–549, Online.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Association for Computational Linguistics, Austin, Texas.

Author biographies

Barry Haddow. is a senior researcher in the School of Informatics at the University of Edinburgh. He has worked in machine translation for more than 10 years, and his current interests include low-resource MT, spoken language translation and evaluation of MT. Barry coordinates the annual WMT conference on machine translation and associated shared tasks.

Rachel Bawden. is a researcher in Machine Translation (MT) at Inria, Paris, particularly interested in low-resource and robust MT. She obtained her PhD in contextual MT in France and has previously worked on the GoURMET low-resource MT project as a research associate at the University of Edinburgh.

Antonio Valerio Miceli Barone. is a researcher in Machine Translation at the University of Edinburgh. He has worked on deep learning methods for multi-lingual natural processing and machine translation, specifically cross-lingual embedding inductionn, deep recurrent neural networks and low-resource techniques. He is currently working on the GoURMET project.

Jindřich Helcl. is a PhD student at Charles University in Prague and a research associate in MT at the University of Edinburgh, working on the GoURMET project. His other topics of interest include non-autoregressive MT and multimodal MT.

Alexandra Birch. is a Reader at the University of Edinburgh. She co-ordinates the H2020 project called GoURMET (2019-2022) and is the PI on the EPSRC fellowship called MTStretch. Both projects focus on low-resource MT for the media industry.