Neural Machine Translation for the Indigenous Languages of the Americas: An Introduction

Manuel Mager[♡*] **Rajat Bhatnagar**[♠] **Graham Neubig**[♯]
Ngoc Thang Vu[♢] **Katharina Kann**[♠]
[♡]AWS AI Labs [♯]Carnegie Mellon University
[♠]University of Colorado Boulder [♢]University of Stuttgart

Abstract

Neural models have drastically advanced state of the art for machine translation (MT) between high-resource languages. Traditionally, these models rely on large amounts of training data, but many language pairs lack these resources. However, an important part of the languages in the world do not have this amount of data. Most languages from the Americas are among them, having a limited amount of parallel and monolingual data, if any. Here, we present an introduction to the interested reader to the basic challenges, concepts, and techniques that involve the creation of MT systems for these languages. Finally, we discuss the recent advances and findings and open questions, product of an increased interest of the NLP community in these languages.

1 Introduction

More than 7 billion people on Earth communicate in nearly 7000 different languages (Pereltsvaig, 2020). Of these, approximately 900 languages are native of the American continent (Campbell, 2000). Most of these indigenous languages of the Americas (ILA) are endangered at some degree (Thomason, 2015). This huge variety in languages is simultaneously a rich treasure for humanity and also a barrier to communication among people from different backgrounds. Human translators have been important in overcoming language barriers. However, trained translators are not accessible to everyone on Earth and even scarcer for endangered and minority languages. The need for translations is even written in the constitutions of several countries like Mexico, Peru, Paraguay, Venezuela, and Bolivia (Zajícová, 2017) to allow native speakers to have equal language rights regarding law.

This is why developing MT is crucial: it helps humanity overcome language barriers while simultaneously allowing people to continue using their native tongue. However, the challenges to achieving these problems are not trivial. It is not only the amount of available data (a common thesis among the NLP community) but also a set of challenging issues (dialectical and orthographic variations, noisy texts, complex morphology, etc.) that must be addressed.

MT has always been an important task within the larger area of natural language processing (NLP). In 1954, the Georgetown–IBM experiment (Hutchins, 2004) was the first that showed at least some effectiveness of MT. Further research resulted in rule-based systems and statistical models. In 2023, neural models define state of the art for MT if training data is plentiful – i.e., for so-called high-resource languages (HRLs) – and have also achieved impressive results for low-resource languages (LRLs). MT is also the most studied NLP task for the ILA (Mager et al., 2018b; Littell et al., 2018). The common issue among these languages is the extreme low-resource conditions they are confronted with. The research interest for these languages has increased in the last years, including the recent AmericasNLP 2021 shared task (Mager et al., 2021) on 10 indigenous languages to Spanish, and the WMT (Conference on Machine Translation) shared task for Inuktitut–English (Barrault et al., 2020).

In this work we aim to provide a comprehensive introduction to the challenges that involve creating MT systems for ILA, and the current status of the existing work. We organize this work as follows: We start by introducing state-of-the-art NMT models (§2). Then, we discuss the current challenges for these languages (§3); and we introduce the key concepts related to low-resource NMT and the implications for endangered languages of the Americas(§3). This is followed by a discussion of available data (§4). Afterwards, we introduce the important concepts for LRL and endangered languages (§5); then we introduce the main strategies

[*]Work done while at the University of Stuttgart.

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 109–133
July 14, 2023

aimed at improving NMT with limited training data (§6); and finally we give an overview of the work done for ILA on MT (§7). In doing so, we provide insights into the following questions: Which systems define the state of the art on low-resource NMT applied to the ILA? What is the route that ahead to improve the translations of the ILA?

2 Background and Definitions

Formally, the task of MT consists of converting text X in a source language L_x into text Y in a target language L_y that conveys the same meaning.[1] Translating text $X \in L_x$ into $Y \in L_y$ can be described as a function (Neubig, 2017):

$$Y = \texttt{MT}(X). \tag{1}$$

X and Y can be of variable length, such as phrases, sentences, or even documents.

If other languages are used during the translation process, e.g., as pivots, we denote them as $L_1, \ldots, L_n$. We refer to a corpus of monolingual sentences in language L_i as $M^{L_i} = S_1, ..., S_n$.

Probabilistic Modeling and Data When using probabilistic MT models, the goal is to find $Y \in L_y$ with the highest conditional probability, given $X \in L_x$. Under the supervised machine learning paradigm, a parallel corpus $C_{parallel} = (X_1, Y_1), ..., (X_n, Y_n)$ is used to learn a set of parameters θ, which define a probability distribution over possible translations. Given $C_{parallel}$, the training objective of an NMT model is generally to maximize the log-likelihood $\mathcal{L}$ with respect to θ:

$$\mathcal{L}_\theta = \sum_{(X_i,Y_i) \in C_{parallel}} \log p(Y_i|X_i;\theta). \tag{2}$$

Within this overall framework, there are a number of design decisions one has to make, such as which model architecture to use, how to generate translations, and how to evaluate.

Decoding Decoding refers to the generation of output $\hat{Y}$, given the parameters θ and an input X. Often, decoding is done by approximately solving the following maximization problem:

$$\text{argmax}_{\hat{Y}} p(\hat{Y}|X;\theta) \tag{3}$$

[1]This is an approximation, since it is in general not possible to map the meaning of text exactly into another language (Nida, 1945; Sechrest et al., 1972; Baker, 2018).

Most NMT systems factorize the probability of $\hat{Y} = \hat{y}_1, ..., \hat{y}_T$ in a left-to-right fashion:

$$p(\hat{Y}) = \prod_{t=1}^{T} p(\hat{y}_t|\hat{y}_{<t}, X, \theta) \tag{4}$$

Thus, the probability of token $\hat{y}_t$ at time step t is computed using the previously generated tokens $\hat{y}_{<t}$, the source sentence X and the model parameters θ. Common algorithms for finding a high-probability translation are greedy decoding, i.e., picking the token with the highest probability at each time step, and beam search (Lowerre, 1976).

2.1 Input Representations

The texts X and Y are input into an NMT system as sequences of continuous vectors. However, defining which units should be represented as such vectors is non-trivial. The classic way is to represent each *word* within X and Y as a vector (or embedding). However, in a low-resource setting, often not all vocabulary items appear in the training data (Jean et al., 2015; Luong et al., 2015). This issue especially effects languages with a rich inflectional morphology (Sennrich et al., 2016c): as many word forms can represent the same lemma, the vocabulary coverage decreases drastically. Furthermore, for many LRLs, boundaries between words or morphemes are not easy to obtain or not well defined in the case of languages without a standard orthography. Alternative input units have been explored, such as characters (Ling et al., 2015), byte pair encoding (BPE; Sennrich et al., 2016a), morphological representations (Vania and Lopez, 2017; Ataman and Federico, 2018), syllables (Zhang et al., 2019), or, recently, a visual representation of rendered text (Salesky et al., 2021). No clear advantage has been discovered for using morphological segmentations over BPEs when testing them on LRLs (Saleva and Lignos, 2021).

Input representations can be pretrained. The two most common options are: i) word embeddings, where each type is represented by a vector, e.g., Word2Vec (Mikolov et al., 2013), Glove (Pennington et al., 2014), or Fasttext (Bojanowski et al., 2017)) embeddings, and ii) contextualized word representations, where entire sentences are being encoded at a time, e.g., ELMo (Peters et al., 2018) or BERT (Devlin et al., 2019). However, training of these methods requires large monolingual training corpora, which may not be readily available for LRLs. As most ILA have rich morphology, this topic has gathered special interest. The discussion about the usage of morpholigical segmented input for NMT models is recurrent. (Mager et al., 2022) show that the unsupervised morphologically inspired models outperform BPE pre-processing (experimented on 4 language pares). Similar experiments done on Quechua–Spanish and Inuktitut–Enlgish (Schwartz et al., 2020), comparing BPEs against Morfessor (Smit et al., 2014). Also (Ortega et al., 2020a) improves the SOTA (state-of-the-art) for Quechua–Spanish MT using a morphological guided BPE algorithm.

2.2 Architectures

NMT models typically are sequence-to-sequence models. They encode a variable-length sequence into a vector or matrix representation, which they then decode back into a variable-length sequence (Cho et al., 2014). The two most frequent architectures are: i) recurrent neural networks (RNN), such as LSTMs (Hochreiter and Schmidhuber, 1997) or GRUs (Cho et al., 2014), and ii) transformers (Vaswani et al., 2017), which define the current state of the art in the high-resource setting.

As for most neural network models, training an NMT system on a limited number of instances is challenging (Fernández-Delgado et al., 2014). There are common problems that arise from limited data in the training set. One major advantage of neural models is their ability to learn representations from raw data, in contrast to manually engineered features (Barron, 1993). However, problems arise when not enough data is provided to enable effective learning of features. Another strength of neural networks is their generalization capacity (Kawaguchi et al., 2017). However, training a neural network on a small dataset easily leads to overfitting (Rolnick et al., 2017). Recent studies, however, show empirically that this does not necessarily happen if the network is tuned correctly (Olson et al., 2018).

2.3 Evaluation

Accurately judging translation quality is difficult and, thus, often still done manually: bilingual speakers assign scores according to provided criteria such as fluency and adequacy (*Does the output have the same meaning as the input?*). However, manual evaluation is expensive and slow. Moreover, in the case of endangered languages, bilingual speakers can be hard or impossible to find.

Automatic metrics provide an alternative.[2] These metrics assign a score to system output, given one or more ground truth reference translations. The most widely used metric is BLEU (Papineni et al., 2002), which relies on token-level n-gram matches between the translation to be rated and one or more gold-standard translations. For morphologically rich languages, character-level metrics, such as chrF (Popović, 2017), are often more suitable, as they allow for more flexibility. In the AmericasNLP ST (Mager et al., 2021) this metric was used over BLEU, as it fits better to the rich morphology of many ILA.

To have a concrete example, lets have the following Wixarika phrase with an English translation:

yu-huta-me	ne-p+-we-'iwa
an-two-ns	1sg:s-asi-2pl:o-brother

I have two brothers

As discussed in (Mager et al., 2018c) it is difficult to translate back from Spanish (or other Fusional language) the morpheme *p+* as it has not equivalent in these languages. So if we would ignore these morpheme at all, BLEU would penalize the entire word *nep+we'iwa*. In contrast, chrF would give credit to the translation, even if the *p+* is missing.

One shortcoming of these evaluation metrics is that the evaluation is very dependent on the surface forms and not on the ultimate goal of semantic similarity and fluency. Recent work uses pretrained models to evaluate semantic similarity between translations and the gold standard (Zhang et al., 2020d), but these methods are limited to languages for which such models are available. This is not possible for the ILA, as the amount of monolingual data is not enough to train a reliable pretrained language model[3].

3 Challenges and open questions

In an overview of the datasets and recent studies of MT for the ILA, we found the following main issues to be handled.

Extreme low-resource parallel datasets Even with the recent advances, the resources available to train MT systems are extremely scarce, having

[2] For a detailed overview of automatic metrics for MT we refer the interested reader to specialized reviews (Han, 2016; Celikyilmaz et al., 2020; Chatzikoumi, 2020).

[3] One exception to this is Quechua, that has a large enough monolingual dataset to train a BERT like model (Zevallos et al., 2022)

training set between 4k and 20k sentences (see §4), with notable exceptions for Inuktitut, Guarani and Quechua (Joanis et al., 2020; Ortega et al., 2020a).

Lack of monolingual data Most of these languages are mostly used in spoken form. In recent years, with the advancement and democratization of mobile technologies, indigenous languages have seen a slight increase in massaging systems and private spheres (Rosales et al.). However, the usage of these languages on the internet is rather limited. Even Wikipedia has a limited amount of these languages (Mager et al., 2018b).

Low domain diversity . As most parallel datasets are scarce, they are restricted to a small number of domains, making it challenging to adapt it, or try to aim for general translation models. This has been recognized as a major problem during the AmericasNLP ST (Mager et al., 2021).

Rich morphology An important number of these languages are morphological highly rich. In many cases, we find polysynthetic, with or highly agglutinative languages (Kann et al., 2018) or even fusional phenomenon (Mager et al., 2020).

Distant paired language The most common languages that we find that ILA is translated into are Spanish, English, and Portuguese. However, these languages are distantly related to the ILA, and have completely different linguistically phenomenons (Campbell, 2000; Romero et al., 2016).

Noisy text environments Monolingual texts, if exist, are found in social media that often use a non-canonical witting (Rosales et al.).

Code-Swithing This phenomenon is strongly present in ILA, as all of these languages are minority languages in their own countries. The bilingualism among their communities is strong (and CS is a common phenomenon in this setup (Çetinoğlu, 2017)). The final result of this phenomenon is the inclusion of code-switching on a common base (Mager et al., 2019) in their language.

Lack of orthographic normalization The usage of ILA faces the problem of having a unified orthographic standard. This is not always possible, as the suggestions of linguists and official entities do not always match the day-by-day writing of the speakers. Moreover, in some cases, special symbols present in the orthographic standards are not accessible in English or Spanish keyboard and need to be replaced with other symbols. The winner of the AmericasNLP ST got important improvements using orthographic normalizers developed specifically for each American language (Vázquez et al., 2021).

Dialectal variety The indigenous languages have a strong dialectal variety, making it hard for native speakers to understand even speakers from neighboring villages. The linguistic richness of entire regions is so diverse that even a single state like the Mexican Oaxaca could correspond to the diversity in the whole Europe (McQuown, 1955).

4 Available MT datasets for ILA

The parallel datasets available for MT have been increasing during the last years. At this moment, we can show in two folds the development of these resources: as shown in table 2 work on specific language has emerged; but also broader datasets have started to cover the ILA (see table 1).

Language-specific corpus collection work has been done for many languages, where parallel corpus has been the main component. In recent time we have seen Cherokee–English (OPUS) (Zhang et al., 2020c), Wixarika–Spanish (Mager et al., 2018a), Shipio–Konibo (Feldman and Coto-Solano, 2020), and others (see table 2). The most prominent of these datasets has been the Inuktitut–English parallel data. The last version of this dataset corpora (Joanis et al., 2020) is has medium size with 1,450,094 sentences. Previous versions of this corpus are (Martin et al., 2003). This data set was used for the WMT 2020 Shared Task on Unsupervised, and Low Resourced MT (Barrault et al., 2020).

For wide-spoken languages like Guarani, it is even possible to collect a web crawled dataset, including news articles and social media parallel aligned data (Chiruzzo et al., 2020; Góngora et al., 2021) This dataset also includes monolingual data. This is possible as Guaraní is one of the most spoken indigenous languages of the continent.

In contrast to the language-specific datasets, we find broader approaches (see table 1). The broadest multilingual dataset, which contains the Bible's New Testament, includes about 1600 languages (Mayer and Cysouw, 2014; McCarthy et al., 2020) of the 2,508 that have been collected by the Summer Institute of Linguistic (SIL) (Anderson and Anderson, 2012). Another remarkable effort to obtain broad language coverage is the PanLex project (Kamholz et al., 2014), which has gathered lexical

Dataset	**Paired-languages**	Authors
AmericasNLI	Aymara, Asháninka, Bribri, Guaraní, Nahuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo, Wixarika	(Ebrahimi et al., 2022)
CPML	Ch'ol, Maya, Mazatec, Mixtec, Nahuatl and Otomi	(Sierra Martínez et al., 2020)
OPUS	*	(Tiedemann, 2016)
New testament Bible	*	(McCarthy et al., 2020)

Table 1: Parallel dataset collections that contain one or more indigenous languages of the Americas

Language	**Paried-language**	**ISO**	**Family**	**Sentences**	Domain	Authors
Asháninka	Spanish	cni	Arawak	3883		(Ortega et al., 2020b)
Bribri	Spanish	bzd	Chibchan	5923		(Feldman and Coto-Solano, 2020)
Guarani	Spanish	gn	Tupi-Guarani		News, Blogs	(Abdelali et al., 2006)
Guarani	Spanish	gn	Tupi-Guarani	14,531	News, Blogs	(Chiruzzo et al., 2020)
Guarani	Spanish	gn	Tupi-Guarani	14,792	News, Social Media	(Góngora et al., 2021)
Guarani	Spanish	gn	Tupi-Guarani	30855	8 Domains	(Chiruzzo et al., 2022)
Nahuatl	Spanish	nah	Uto-Aztecan	16145	Diverse Books	(Gutierrez-Vasques et al., 2016)
Otomí	Spanish	oto	Oto-Manguean	4889	Diverse Books	`https://tsunkua.elotl.mx`
Rarámuri	Spanish	tar	Uto-Aztecan	14721	Dictionary Examples	(Mager et al., 2022)
Shipibo-Konibo	Spanish	shp	Panoan	14592	Educational, Religious	(Galarreta et al., 2017)
Wixarika	Spanish	hch	Uto-Aztecan	8966	Literature	(Mager et al., 2018a)
Cherokee	English	chr	Uto-Aztecan		OPUS	(Zhang et al., 2020c)
Inuktitut	English	iku	Eskimo–Aleut	1,450,094	Legislative	(Joanis et al., 2020)
Ayuuk	Spanish	mir	Mixe–Zoque	7553	Diverse	(Zacarías Márquez and Meza Ruiz, 2021)
Mazatec	Spanish	Many	Oto-Manguean	9799	Diverse	(Tonja et al., 2023)
Mixtec	Spanish	Many	Oto-Manguean	13235	Diverse	(Tonja et al., 2023)

Table 2: Parallel datasets that have been released focusing on one indigenous language

translation dictionaries for over 5,700 languages. However, for most languages, PanLex contains only a few dozen words. Duan et al. (2020) show that such dictionaries can be used to create an NMT system, making bilingual dictionaries relevant for further studies.

Recently community-driven research groups have started the creation of own parallel datasets, such as Masakhane (Orife et al., 2020; Nekoto et al., 2020) for African languages, and AmericasNLP for indigenous languages of the Americas (Ebrahimi et al., 2021; Mager et al., 2021). The AmericasNLI dataset is an important effort to have a common evaluation benchmark for the 10 indigenous languages of the Americas for the MT and NLI tasks.

Given the constitutional rights of indigenous languages in many countries of the Americas, it is possible to access this data. Vázquez et al. (2021) made available this resource during their shared task system development.

Finally, it is important to mention that many of the languages spoken in the Americas have Wikipedia's set of articles available[4].

Collection of New Data A common way to create parallel data with the help of bilingual speakers is via elicitation (translating the foreign text into another language). It has the disadvantage of biasing the created text to forms and topics, culture, and even grammatical forms towards the source language (Lörscher, 2005). A method that avoids this problem is language documentation, which consists of storing and annotating commonly used speech or text (Himmelmann, 2008). However, it is

[4]The available languages in wikipedia can be consulted at: `https://es.wikipedia.org/wiki/Portal:Lenguas_indígenas_de_América`. Until the publication of this article, there were only entries in Nahuatl, Navajo, Guarani, Aymara, Klaalisut, Esquimal, Inukitut, Cherokee, and Cree.

costly and requires specialists. In this process, involving the community members that are bilingual speakers is important (Bird, 2020).

5 Low-resource MT

For the purpose of this paper we define LRLs as languages for which standard techniques are unable to create well performing systems, which makes it necessary to resort to other techniques (cf. Figure 1) such as transfer learning. For MT, the amount of available resources differs widely across language pairs: some have less than 10k parallel sentences, while other have more than 500k, with some exceptions in the orders of several million.

Emulating a low-resource scenario by downsampling available data for high-resource languages is common and helps understanding a model's performance across different settings. However, further evaluating methods on a diverse set of low-resource languages is crucial, since many languages exhibit particular linguistic phenomena (Mager et al., 2020), that perturb the final results, especially since most large datasets are from the Indo-European language family, to which only 6.16% of the world's languages belong (Lewis, 2009).

Importantly, there is no strong correlation between the number of resources available per language and the number of speakers: Javanese with 95 million speakers and Kannada with 44 million are considered LRLs, while French, with only 64 million native speakers, is among the most widely studied languages. Improving models to handle LRLs will extend access to information online as well as human language technology to all monolingual speakers of those languages. In the case of ILA, most languages are endangered at some degree, but most of them have the same issue: they are low resourced for parallel and monolingual data.

Endangered Languages Krauss (1992) estimates that 50% of all languages are doomed or dying, and that in this century we will see either the death or the doom of 90% of all human languages. The current proportion of languages that are already extinct or moribund ranges from 31% down to 8% depending on the region, with the most severe cases in the Americas and Australia (Simons and Lewis, 2013). To determine how endangered a language is, Lewis and Simons (2010) proposes a classification scale called EGIDS with 13 levels. The higher the number on this scale, the greater the level of disruption of the language's inter-generational transmission.[5] MT for endangered LRLs has the potential to help with documentation, promotion and revitalization efforts (Galla, 2016; Mager et al., 2018b). However, as these languages are commonly spoken by small communities, or indigenous people, researchers should aim for a direct involvement of those communities (Bird, 2020).

What is polysynthesis? A polysynthetic language is defined by the following linguistic features: the verb in a polysynthetic language must have an agreement with the subject, objects and indirect objects (Baker, 1996); nouns can be incorporated into the complex verb morphology (Mithun, 1986); and, therefore, polysynthetic languages have agreement morphemes, pronominal affixes and incorporated roots in the verb (Baker, 1996), and also encode their relations and characterizations into that verb. The most common word orders present in these languages are SOV, VSO, SVO and free order. It is important to notice that a polysynthtic language can have a aggutinative [6] or can have also fusional characteristics, like Totonaco or Tepehua (Mager et al., 2020).

6 Low-resource MT paradigms

Most languages of the Americas do not have high amount of data for MT. Therefore, we introduce the most important paradigms to improve low-resourced machine translation. Figure 1 shows a general overview of the methods and options to improve LRL MT. For a more detailed understanding of this techniques we refer the reader to specialized low-resource MT surveys (Haddow et al., 2022; Wang et al., 2021; Ranathunga et al., 2021).

6.1 Multilingual Supervised Training

With a multilingual set of parallel data $D_{parallel}$ between different language pairs $\{(L_1, L_2), \ldots, (L_m, L_n)\}$ we can train a model that is able to map a sentence from any source language L_x into any target language L_y that is contained in $D_{parallel}$ (see 2). These multilingual NMT models have seen a growth in popularity and efficiency in recent years. We will now cover the different training algorithms for these models: 1) many source languages and one target

[5] The complete EGIDS scale can be found at https://www.ethnologue.com/about/language-status

[6] Agglutination refers to a concatenation of morphemes, with minimal changes to the surface form.

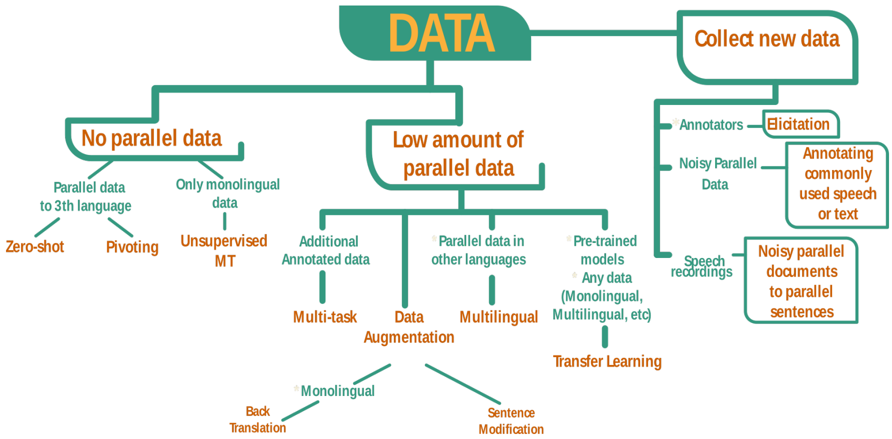

Figure 1: What to do when we have low o no data to train our machine translation models? This diagram shows basic scenarios, solutions, and common requirements for each method, with the section describing the method.

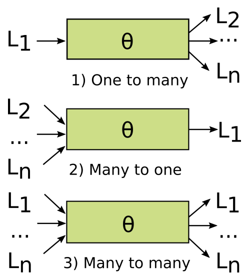

Figure 2: An overview of different multilingual setups.

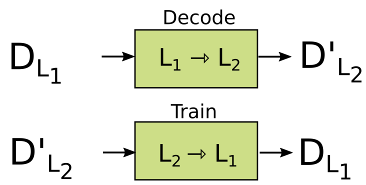

Figure 3: Backtranslation

language (*many-to-one*), 2) one source and many target languages (*one-to-many*), and 3) many source languages and many target languages (*many-to-many*). For a general overview of multilingual MT, we refer the reader to surveys dedicated to this topic (Tan et al., 2019; Dabre et al., 2019). Johnson et al. (2017) are the first to introduce a multilingual NMT model, trained on translating from a large number of languages to English as well as in the opposite direction. The authors show that these models improve over single-language pair models for LRLs.

6.2 Multi-task Training

Multi-task training (Caruana, 1997) aims to improve the performance of the main task – MT in our case – by adding one or more auxiliary tasks to the training. The easiest way is to share all parameters of the network, using the ideas already explored in multilingual NMT (§6.1). This can be done with a special flag in the input that specifies the current task. It is also possible to share only the encoder and have two separate decoders for each task.

Multilingual Modeling In order to handle multilinguality it is also possible to adapt modify the NMT models. The main proposals to do so has been: sharing all parameter except the attention mechanism of a RNN NMT model (Blackwood et al., 2018); parameter sharing in the transformer architecture Sachan and Neubig (2018);

6.3 Data Augmentation

Back-Translation A straightforward way to leverage monolingual data for low-resource MT is to generate a meaningful signal with the help of an already initialized MT model (see Figure 3). This method is called back-translation (BT; Sennrich et al., 2016b): With monolingual data M^{L_x} in source language L_x and a trained model that is able to translate from L_x into a target language L_y we can generate a translation M'^{L_y}. This pseudo parallel data (M^{L_x}, M'^{L_y}) is then used to train a new model in the opposite direction. This process can be applied iteratively to improve the translation (Hoang et al., 2018).

Sentence Modification Other methods to generate more parallel sentences are based on lexical substitution. Fadaee et al. (2017) explores replacing frequent words with low-frequency ones in both source and target to improve the translation of rare words. This is done using language models (LMs) and automatic alignment.

Pivoting If no parallel corpus between languages L_x and L_y is available, but both of them have parallel corpora with a third language L_p, pivoting is an option. The basic idea is to train two MT systems: one that translates $L_x \rightarrow L_p$ and another for $L_p \rightarrow L_y$. Pivoting has first been introduced for SMT (Wu and Wang, 2007; Cohn and Lapata, 2007; Utiyama and Isahara, 2007).

6.4 Semi-supervised and Unsupervised MT

Transfer Learning via Pretraining Transfer learning refers to using knowledge learned from one task to improve performance on a related task (Weiss et al., 2016). In recent years this approach has gained popularity with big multilingual models such as Conneau and Lample (2019) that proposes training the encoder and the decoder separately in order to get cross-language representations (XLM). This idea has further been extended by Song et al. (2019, MASS) to masking a *sequence* of tokens from the input (multilingual MASS (Siddhant et al., 2020)). Another approach is to train the entire transformer model as a denoising autoencoder (BART; Lewis et al., 2019) (multilingual BART (mBART) (Liu et al., 2020)). It is also possible to pretrain a transformer in a multi-task, text-to-text fashion, where one of the tasks is MT (T5; Raffel et al., 2020) (multilingual version (Xue et al., 2021)).

Unsupervised MT UMT covers approaches that do *not* require any parallel text, relying only on monolingual data. This differs from zero-shot translation, which uses parallel data for other language pairs. Early approaches tackled the problem with an auto-encoder with adversarial training (Lample et al., 2017) or with auto-encoders with a shared encoding space as well as separate decoders for each target language (Artetxe et al., 2018). The main problem for these approches is the need of a big monolingual dataset, that is not available for most ILA.

7 Advances in MT for the indigenous languages of the Americas

In recent years the interest in MT for indigenous languages of the Americas has increased. The task is not easy. The first usage of NMT systems has not been successful (Mager and Meza, 2021). However, with the use of LRL MT methods, we have witnessed great improvements.

The Cherokee–English (Zhang et al., 2020c) language pair has been explored using a pre-trained BERT (Devlin et al., 2019) for the English side. A system demonstration of this approach is also accessible (Zhang et al., 2021). The back translation strategy for Bribri–Spanish NMT transformers has also been explored (Feldman and Coto-Solano, 2020) and by (Oncevay, 2021) (for four Peruvian languages to Spanish) with good results. The scarce indigenous language monolingual text can be replaced to some extent with Spanish text or extracted from PDFs, and other sources (Bustamante et al., 2020).

One of the main challenges for the complex morphological languages in the area has been the prepossessing step. Schwartz et al. (2020) show that even if morphological segmentation has less perplexity a the language modeling time, it is still under-performing or equivalent against BPEs for MT (for Inuktitut-–English, Yupik—English Data, Guaraní—Spanish Data). A more comprehensive (on the segmentation modeling side) was done by (Mager et al., 2022) exploring a wide array of segmentation models.The latter study showed that supervised morphological segmentation underperform unsupervised. However, unsupervised morphological segmentation like LMVR (Ataman et al., 2017) and FlatCat (Grönroos et al., 2014) perform better than BPEs. (Ngoc Le and Sadat, 2020) studied how better to perform word segmentation for the Inuktitut–English pair. They found that for this language pair, a morphological segmentation, or a combination of BPEs and morphological segmentation, works better than just applying vanilla BPEs. Also, training word embeddings for Guarani–Spanish translation is an excellent opportunity to increase the MT performance of these languages (Góngora et al., 2022).

The usage of transfer learning from multilingual systems has been tried, with limited results (Nagoudi et al., 2021) (training an own T5 model for indigenous languages) and (Zheng et al., 2021). However, pertaining a Spanish–English model together with ILA, and then fine-tuning it (together with a careful prepossessing and filtering step) has been the most successful strategy (Vázquez et al., 2021).

The quality of MT systems of ILA has been a constant debate. However, Ebrahimi et al. (2021) shows that the quality of MT for these languages is enough to improve other tasks like natural language inference (NLI).

Inuktitut–Enlgish ST The WMT 2020 news translation task included Inuktitut–English translation (Barrault et al., 2020). The participating systems explored the difficulties of working with a polysynthetic language in a medium resource scenario. Participating teams in this competition were: (Kocmi, 2020; Hernandez and Nguyen, 2020; Scherrer et al., 2020; Roest et al., 2020; Lo, 2020; Knowles et al., 2020; Zhang et al., 2020e; Krubiński et al., 2020).

AmericasNLP 2021 and 2023 ST In 2021, the AmericasNLP community organized a workshop on Machine Translation for 10 indigenous languages of the Americas in 2021 (Mager et al., 2021) and 2023 (Ebrahimi et al., 2023) with an additional indigenous language (Chatino). The AmericasNLP shared task winner was (Vázquez et al., 2021) in 2021, and a more mixed result in 2023[7]. Other participants in this shared task are (Nagoudi et al., 2021; Bollmann et al., 2021; Zheng et al., 2021; Knowles et al., 2021; Parida et al., 2021; Nagoudi et al., 2021). It is important to point at the importance of clean datata. For Quechua, (Moreno, 2021) got the best results generating an additional amount of clean data.

AmericasNLP 2022 Competition is a competition on Speech-to-Text translation is organized and is targeting the following language pairs: Bribri–Spanish, Guaraní–Spanish, Kotiria–Portuguese, Wa'ikhana–Portuguese, and Quechua–Spanish (Ebrahim et al., 2023)[8].

8 Ethical aspects

When working with ILAs are also interacting with communities and nations that speak these languages. In most cases, these speakers have been exposed to a colonial past, or to a local oppression, by the majority language and culture. It is important to point to best practices and recommendations when performing our research. Bird (2020) and Liu et al. (2022) advocate to include community members as co-authors (Liu et al., 2022) as well as considering data and technology sovereignty. This is also aligned with the community building aimed at by Zhang et al. (2022). Mager et al. (2023) summarizes the main aspects that should be considered as follows: i) *Consultation, Negotiation and Mutual Understanding*. It is important to inform the community about the planned research, negotiating a possible outcome, and reaching a mutual agreement on the directions and details of the project should happen in all cases. ii) *Respect of the local culture and involvement*. As each community has its own culture and view of the world, researchers should be familiar with the history and traditions of the community. Also, it should be recommended that local researchers, speakers, or internal governments should be involved in the project. iii) *Sharing and distribution of data and research*. The product of the research should be available for use by the community, so they can take advantage of the generated materials, like papers, books, or data.

9 Conclusion

Machine translation for ILA has gained interest in the NLP community over the last few years. Here, we provide an exhaustive overview of the basic MT concepts and the particular challenges for MT for ILA (in the context of low-resource scenarios and its relation to endangered languages). We additionally survey the current advances of MT for these languages.

Limitations

This paper's aim is to give an introduction to researchers, students, of interested community indigenous community members to the topic of Machine Translation for Indigenous languages of the Americas. Therefore, this paper is not an in-depth survey of the literature on indigenous languages nor a more technical survey of low-resource machine translation. We would point the reader to more specific surveys on these aspects (Haddow et al., 2022; Mager et al., 2018b).

Ethical statement

We could not find any specific Ethical issue for this paper or potential danger. Nevertheless, we want to point to the reader that working with indigenous languages (in this case, MT) implies a set of ethical questions that are important to handle. For a deeper understanding of the matter, we suggest specialized literature to the reader (Mager et al., 2023; Bird, 2020; Schwartz, 2022).

[7] Up to this moment, no official desciption papers for the 2023 are published.

[8] http://turing.iimas.unam.mx/americasnlp/st.html

References

Ahmed Abdelali, James Cowie, Steve Helmreich, Wanying Jin, Maria Pilar Milagros, Bill Ogden, Mansouri Rad, and Ron Zacharski. 2006. Guarani: A case study in resource development for quick ramp-up MT. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 1–9, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Idris Abdulmumin, Bashir Shehu Galadanci, and Aliyu Garba. 2019. Tag-less back-translation. *arXiv preprint arXiv:1912.10514*.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884.

Benyamin Ahmadnia and Bonnie J Dorr. 2019. Augmenting neural machine translation through round-trip training approach. *Open Computer Science*, 9(1):268–278.

Antonios Anastasopoulos and David Chiang. 2018. Tied multitask learning for neural speech translation. 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 82–91, New Orleans, Louisiana. Association for Computational Linguistics.

Stephen R Anderson and Stephen Anderson. 2012. *Languages: A very short introduction*, volume 320. Oxford University Press.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *arXiv preprint arXiv:1903.07091*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Duygu Ataman and Marcello Federico. 2018. Compositional representation of morphologically-rich input for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *stat*, 1050:21.

Mark C Baker. 1996. *The polysynthesis parameter*. Oxford University Press.

Mona Baker. 2018. *In other words: A coursebook on translation*. Routledge.

Loïc Barrault, 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, Online. Association for Computational Linguistics.

Andrew R Barron. 1993. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. *arXiv preprint arXiv:2004.14928*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3112–3122.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Marcel Bollmann, Rahul Aralikatte, Héctor Murrieta Bello, Daniel Hershcovich, Miryam de Lhoneux, and Anders Søgaard. 2021. Moses and the character-based random babbling baseline: CoAStaL at AmericasNLP 2021 shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 248–254, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Lyle Campbell. 2000. *American Indian languages: the historical linguistics of Native America*, volume 4. Oxford University Press on Demand.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Isaac Caswell, 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.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Özlem Çetinoğlu. 2017. A code-switching corpus of Turkish-German conversations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 34–40, Valencia, Spain. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubhanker Banerjee, Richard Saldanha, John P. McCrae, Anand Kumar M, Parameswari Krishnamurthy, and Melvin Johnson. 2021. Findings of the shared task on machine translation in Dravidian languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 119–125, Kyiv. Association for Computational Linguistics.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2021. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 15–26, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yong Cheng. 2019. Joint training for pivot-based neural machine translation. In *Joint Training for Neural Machine Translation*, pages 41–54. Springer.

Yong Cheng, Lu Jiang, and Wolfgang Macherey. 2019. Robust neural machine translation with doubly adversarial inputs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4324–4333.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. AdvAug: Robust adversarial augmentation for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5961–5970, Online. Association for Computational Linguistics.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022. Jojajovai: A parallel guarani-spanish corpus for mt benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2098–2107.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2019. A survey of multilingual neural machine translation. *arXiv preprint arXiv:1905.05395*.

Jacob Devlin, 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.

Xiangyu Duan, 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.

Abteen Ebrahim, Manuel Mager, Pavel Oncevay Arturo Danni Liu Koneru Sai Ugan Enes Yavuz Wiemerslage, Adam Denisov, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, Bashar Talafha, Inciarte Alcides Alcoba, Muhammad Abdul-Mageed, Luis Chiruzzo, Rolando Coto-Solano, Hilaria Cruz, Sofía Flores-Solórzano, Aldo Andrés Alvarez López, Ivan Meza-Ruiz, John E. Ortega, Alexis Palmer, Rodolfo Joel Zevallos Salazar, Kristine, Thang Vu Stenzel, and Katharina Kann. 2023. Findings of the second americasnlp competition on speech-to-text translation. *preprint*.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Enora Rice, Cynthia Montaño, John Ortega, Shruti Rijhwani, Alexis Palmer, Rolando Coto-Solano, Hilaria Cruz, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Third Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. 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 4052–4059.

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

Marzieh Fadaee, 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.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181.

Alexander Fraser. 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, Online. Association for Computational Linguistics.

Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. Corpus creation and initial SMT experiments between Spanish and Shipibo-konibo. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244, Varna, Bulgaria. INCOMA Ltd.

Candace Kaleimamoowahinekapu Galla. 2016. Indigenous language revitalization, promotion, and education: Function of digital technology. *Computer Assisted Language Learning*, 29(7):1137–1151.

Xavier Garcia, Pierre Foret, Thibault Sellam, and Ankur P Parikh. 2020. A multilingual view of unsupervised machine translation. *arXiv preprint arXiv:2002.02955*.

Mozhdeh Gheini, Xiang Ren, and Jonathan May. 2021. Cross-attention is all you need: Adapting pretrained Transformers for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1765, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158, Online. Association for Computational Linguistics.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2022. Can we use word embeddings for enhancing Guarani-Spanish machine translation? In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 127–132, Dublin, Ireland. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Stig-Arne Grönroos, 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.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, pages 1–67.

Lifeng Han. 2016. Machine translation evaluation resources and methods: A survey. *arXiv preprint arXiv:1605.04515*.

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

Nikolaus P Himmelmann. 2008. Language documentation: What is it and what is it good for? In *Essentials of language documentation*, pages 1–30. De Gruyter Mouton.

Vu Cong Duy Hoang, 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

J Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850.

W John Hutchins. 2004. The georgetown-ibm experiment demonstrated in january 1954. In *Conference of the Association for Machine Translation in the Americas*, pages 102–114. Springer.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. On using very large target vocabulary for neural machine 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 1–10, Beijing, China. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Melvin Johnson, Mike Schuster, Quoc V 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.

David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. Panlex: Building a resource for panlingual lexical translation. In *LREC*, pages 3145–3150.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. 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 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. 2017. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. *arXiv preprint arXiv:2004.14524*.

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

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2021. NRC-CNRC machine translation systems for the 2021 AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 224–233, Online. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. 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 452–457.

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

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Mateusz Krubiński, Marcin Chochowski, Bartłomiej Boczek, Mikołaj Koszowski, Adam Dobrowolski, Marcin Szymański, and Paweł Przybysz. 2020. Samsung R&D institute Poland submission to WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 181–190, Online. Association for Computational Linguistics.

Surafel M Lakew, Quintino F Lotito, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Improving zero-shot translation of low-resource languages. In *Proceedings of the 14h IWSLT*, pages 113–119.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. Zero-shot neural machine translation: Russian-Hindi @LoResMT 2020. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 38–42, Suzhou, China. Association for Computational Linguistics.

Yichong Leng, Xu Tan, Tao Qin, Xiang-Yang Li, and Tie-Yan Liu. 2019. Unsupervised pivot translation for distant languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 175–183.

M Paul Lewis. 2009. *Ethnologue: Languages of the world*. SIL international.

M Paul Lewis and Gary F Simons. 2010. Assessing endangerment: expanding fishman's gids. *Revue roumaine de linguistique*, 55(2):103–120.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Zuchao Li, Rui Wang, Kehai Chen, Masso Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020. Data-dependent gaussian prior objective for language generation. In *International Conference on Learning Representations*.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages

2649–2663, Online. Association for Computational Linguistics.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Zihan Liu, Yan Xu, Genta Indra Winata, and Pascale Fung. 2019. Incorporating word and subword units in unsupervised machine translation using language model rescoring. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 275–282.

Zoey Liu, Crystal Richardson, Richard Hatcher Jr, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. *arXiv preprint arXiv:2204.05541*.

Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 895–902, Online. Association for Computational Linguistics.

Wolfgang Lörscher. 2005. The translation process: Methods and problems of its investigation. *Meta: journal des traducteurs/Meta: Translators' Journal*, 50(2):597–608.

Bruce T Lowerre. 1976. The harpy speech recognition system. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 84–92.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine 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 11–19, Beijing, China. Association for Computational Linguistics.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018a. Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2019. Subword-level language identification for intra-word code-switching. 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 2005–2011, Minneapolis, Minnesota. Association for Computational Linguistics.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the low-resource challenge for canonical segmentation. *arXiv preprint arXiv:2010.02804*.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018b. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. *arXiv preprint arXiv:2305.19474*.

Manuel Mager, Elisabeth Mager, Alfonso Medina-Urrea, Ivan Vladimir Meza Ruiz, and Katharina Kann. 2018c. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 73–83, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager and Ivan Meza. 2021. Retos en construcción de traductores automáticos para lenguas indígenas de México. *Digital Scholarship in the Humanities*, 36.

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

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Thang Vu. 2022. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5990–5997, Online. Association for Computational Linguistics.

Joel Martin, Howard Johnson, Benoit Farley, and Anna Maclachlan. 2003. Aligning and using an english-inuktitut parallel corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond-Volume 3*, pages 115–118. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. *Oceania*, 135(273):40.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The johns hopkins university bible corpus: 1600+ tongues for typological exploration. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Norman A McQuown. 1955. The indigenous languages of latin america. *American Anthropologist*, 57(3):501–570.

Antonio Valerio Miceli-Barone, Jindřich Helcl, Rico Sennrich, Barry Haddow, and Alexandra Birch. 2017. Deep architectures for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 99–107.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Marianne Mithun. 1986. On the nature of noun incorporation. *Language*, 62(1):32–37.

Oscar Moreno. 2021. The REPU CS' Spanish–Quechua submission to the AmericasNLP 2021 shared task on open machine translation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 241–247, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusoglu. 2021. IndT5: A text-to-text transformer for 10 indigenous languages. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 265–271, Online. Association for Computational Linguistics.

Wilhelmina Nekoto, 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, Online. Association for Computational Linguistics.

Graham Neubig. 2017. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*.

Graham Neubig 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.

Tan Ngoc Le and Fatiha Sadat. 2020. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Dat Quoc Nguyen, Kairit Sirts, and Mark Johnson. 2015. Improving topic coherence with latent feature word representations in MAP estimation for topic modeling. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 116–121, Parramatta, Australia.

Toan Q Nguyen 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.

Eugene Nida. 1945. Linguistics and ethnology in translation-problems. *Word*, 1(2):194–208.

Jan Niehues 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, Copenhagen, Denmark. Association for Computational Linguistics.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1828–1836, Osaka, Japan. The COLING 2016 Organizing Committee.

Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. 2020. Zero-shot cross-lingual transfer with meta learning. *arXiv preprint arXiv:2003.02739*.

Atul Kr. Ojha, 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, Suzhou, China. Association for Computational Linguistics.

Matthew Olson, Abraham Wyner, and Richard Berk. 2018. Modern neural networks generalize on small data sets. In *Advances in Neural Information Processing Systems*, pages 3619–3628.

Arturo Oncevay. 2021. Peru is multilingual, its machine translation should be too? In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 194–201, Online. Association for Computational Linguistics.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020a. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

John E Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020b. Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. Association for Computational Linguistics.

Yirong Pan, Xiao Li, Yating Yang, and Rui Dong. 2020. Multi-task neural model for agglutinative language translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 103–110, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shantipriya Parida, Subhadarshi Panda, Amulya Dash, Esau Villatoro-Tello, A. Seza Doğruöz, Rosa M. Ortega-Mendoza, Amadeo Hernández, Yashvardhan Sharma, and Petr Motlicek. 2021. Open machine translation for low resource South American languages (AmericasNLP 2021 shared task contribution). In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 218–223, Online. Association for Computational Linguistics.

Jeffrey Pennington, 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, Doha, Qatar. Association for Computational Linguistics.

Asya Pereltsvaig. 2020. *Languages of the World*. Cambridge University Press.

Matthew Peters, 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.

Alberto Poncelas, Maja Popović, Dimitar Shterionov, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Combining pbsmt and nmt back-translated data for efficient nmt. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 922–931.

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

Nima Pourdamghani and Kevin Knight. 2019. Neighbors helping the poor: improving low-resource machine translation using related languages. *Machine Translation*, 33(3):239–258.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Alessandro Raganato, Raúl Vázquez, Mathias Creutz, and Jörg Tiedemann. 2021. An empirical investigation of word alignment supervision for zero-shot multilingual neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8449–8456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2021. Neural machine translation for low-resource languages: A survey. *arXiv preprint arXiv:2106.15115*.

Shuo Ren, Yu Wu, Shujie Liu, Ming Zhou, and Shuai Ma. 2020. A retrieve-and-rewrite initialization method for unsupervised machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3498–3504, Online. Association for Computational Linguistics.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Christian Roest, Lukas Edman, Gosse Minnema, Kevin Kelly, Jennifer Spenader, and Antonio Toral. 2020. Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 274–281, Online. Association for Computational Linguistics.

David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. 2017. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Carlos Barron Romero, Jesús Manuel Mager Hois, and Fernando Reyes Avilés. 2016. Richard feynman, los alfabetos y los lenguajes. *Relingüística aplicada*, (19):2.

Mónica Jasso Rosales, Manuel Mager, and Ivan Vladimir Meza Ruız. Towards a twitter corpus of the indigenous languages of the americas.

Devendra Singh Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. *arXiv preprint arXiv:1809.00252*.

Elizabeth Salesky, David Etter, and Matt Post. 2021. Robust open-vocabulary translation from visual text representations. *arXiv preprint arXiv:2104.08211*.

Jonne Saleva 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, Online. Association for Computational Linguistics.

Motoki Sano, Jun Suzuki, and Shun Kiyono. 2019. Effective adversarial regularization for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 204–210.

Yves Scherrer, 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, Online. Association for Computational Linguistics.

Lane Schwartz. 2022. Primum non nocere: Before working with indigenous data, the acl must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731.

Lane Schwartz, Francis Tyers, Lori Levin, Christo Kirov, Patrick Littell, Chi-kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, et al. 2020. Neural polysynthetic language modelling. *arXiv preprint arXiv:2005.05477*.

Lee Sechrest, Todd L Fay, and SM Hafeez Zaidi. 1972. Problems of translation in cross-cultural research. *Journal of cross-cultural psychology*, 3(1):41–56.

Rico Sennrich, 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, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016c. 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, Berlin, Germany. Association for Computational Linguistics.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. In *International*

Conference on Machine Learning, pages 4548–4557. PMLR.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Gerardo Sierra Martínez, Cynthia Montaño, Gemma Bel-Enguix, Diego Córdova, and Margarita Mota Montoya. 2020. CPLM, a parallel corpus for Mexican languages: Development and interface. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2947–2952, Marseille, France. European Language Resources Association.

Gary F Simons and M Paul Lewis. 2013. The world's languages in crisis. *Responses to language endangerment: In honor of Mickey Noonan. New directions in language documentation and language revitalization*, 3:20.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, Mikko Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*. Aalto University.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788.

Haiyue Song, Raj Dabre, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, and Eiichiro Sumita. 2020. Pre-training via leveraging assisting languages and data selection for neural machine translation. *arXiv preprint arXiv:2001.08353*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

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

Tejas Srinivasan, Ramon Sanabria, and Florian Metze. 2019. Multitask learning for different subword segmentations in neural machine translation. *arXiv preprint arXiv:1910.12368*.

Dario Stojanovski, Viktor Hangya, Matthias Huck, and Alexander Fraser. 2019. The lmu munich unsupervised machine translation system for wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 393–399.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Robust unsupervised neural machine translation with adversarial training. *arXiv preprint arXiv:2002.12549*.

Xu Tan, Yichong Leng, Jiale Chen, Yi Ren, Tao Qin, and Tie-Yan Liu. 2019. A study of multilingual neural machine translation. *arXiv preprint arXiv:1912.11625*.

Sarah G Thomason. 2015. *Endangered languages*. Cambridge University Press.

Jörg Tiedemann. 2016. Opus–parallel corpora for everyone. *Baltic Journal of Modern Computing*, page 384.

Jörg Tiedemann. 2018. Emerging language spaces learned from massively multilingual corpora. *arXiv preprint arXiv:1802.00273*.

Atnafu Lambebo Tonja, Christian Maldonado-Sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023. Parallel corpus for indigenous language translation: Spanish-mazatec and spanish-mixtec. *arXiv preprint arXiv:2305.17404*.

Antonio Toral, 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.

Hai-Long Trieu, Duc-Vu Tran, Ashwin Ittoo, and Le-Minh Nguyen. 2019. Leveraging additional resources for improving statistical machine translation on asian low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(3).

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? 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 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? *arXiv preprint arXiv:2004.04070*.

Liang Wang, Wei Zhao, Ruoyu Jia, Sujian Li, and Jingming Liu. 2019a. Denoising based sequence-to-sequence pre-training for text generation. 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 3994–4006.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019b. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4636–4643. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019c. Multilingual neural machine translation with soft decoupled encoding. In *International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big Data*, 3(1):9.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond bleu: Training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

Haoran Xu, Benjamin Van Durme, and Kenton Murray. 2021. BERT, mBERT, or BiBERT? a study on contextualized embeddings for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6663–6675, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Delfino Zacarías Márquez and Ivan Vladimir Meza Ruiz. 2021. Ayuuk-Spanish neural machine translator. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 168–172, Online. Association for Computational Linguistics.

Lenka Zajícová. 2017. Lenguas indígenas en la legislación de los países hispanoamericanos. *Onomázein*, (NE III):171–203.

Poorya Zaremoodi, 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.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Aradiel, and Hilario Nelsi Melgarejo. 2022. Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13, Hybrid. Association for Computational Linguistics.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020b. 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, Online. Association for Computational Linguistics.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can nlp help revitalize endangered languages? a case study and roadmap for the cherokee language. *arXiv preprint arXiv:2204.11909*.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020c. ChrEn: Cherokee-English machine translation for endangered language revitalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 577–595, Online. Association for Computational Linguistics.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2021. ChrEnTranslate: Cherokee-English machine translation demo with quality estimation and corrective feedback. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 272–279, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020d. Bertscore: Evaluating text generation with bert. In *ICLR*.

Yuhao Zhang, 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. 2020e. The NiuTrans machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 338–345, Online. Association for Computational Linguistics.

Zhuosheng Zhang, Yafang Huang, and Hai Zhao. 2019. Open vocabulary learning for neural chinese pinyin ime. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1584–1594.

Francis Zheng, Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Low-resource machine translation using cross-lingual language model pretraining. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 234–240, Online. Association for Computational Linguistics.

Hao Zheng, Yong Cheng, and Yang Liu. 2017. Maximum expected likelihood estimation for zero-resource neural machine translation. In *IJCAI*, pages 4251–4257.

Shuyan Zhou, Xiangkai Zeng, Yingqi Zhou, Antonios Anastasopoulos, and Graham Neubig. 2019. Improving robustness of neural machine translation with multi-task learning. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 565–571, Florence, Italy. Association for Computational Linguistics.

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020a. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655, Online. Association for Computational Linguistics.

Jinhua Zhu, Fei Gao, Lijun Wu, Yingce Xia, Tao Qin, Wengang Zhou, Xueqi Cheng, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. *arXiv preprint arXiv:1905.10523*.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020b. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

A Appendix

In this appendix we expand the information regarding current work on MT for LRL.

A.1 Expanded LR work on Multilingual supervised training

Arivazhagan et al. (2019a) introduce a representational invariance training objective across languages that achieves comparable results with pivoting methods. Promising results of multilingual models have encouraged experiments with models trained on a massive amount of language pairs, resulting in large multilingual models: Aharoni et al. (2019) train a single model on 102 languages to and from English in contrast to the 58 languages used by Neubig and Hu (2018).

The negative aspect of this approach is the size of the network. Arivazhagan et al. (2019b) perform an extensive study on 102 language pairs to explore different settings and training setups and achieve good results for LRLs, while maintaining good performance for high-resource languages. Related massively multilingual NMT systems have been trained for analytic proposes (Tiedemann, 2018; Malaviya et al., 2017) and general zero-shot transfer learning (Artetxe and Schwenk, 2019). mRASP (Lin et al., 2020) use for pretraining of the multilingual model and add a randomly aligned substitution loss that aims to bring words and phrases closer in the cross-lingual space.

Zhang et al. (2020a) explores the main problems that arise for such models: multilingual NMT usually underperforms bilingual models (Arivazhagan et al., 2019b), the larger the number of languages gets the more the performance drops (Aharoni et al., 2019), languages in datasets used for multilingual training are unbalanced in size, and poor zero-shot performance compared to pivot models (cf. §6.3). Zhang et al. (2020a) addresses these problems with a language-aware input layer, a deep transformer architecture (Wang et al., 2019b), and an online back-translation approach. These modifications boost zero-shot translation performance for multilingual models.

To improve the problem of imbalanced and linguistically diverse training data, mostly heuristic methods have been proposed: Arivazhagan et al. (2019b) samples training data from different languages based on a data size scaled by temperature term. These heuristics have an impact on performance, and ignore other factors that are not size. Oversampling of data is used by Johnson et al. (2017); Neubig and Hu (2018); Conneau and Lample (2019). Wang et al. (2020) proposes a differentiable data selection method that automatically learns to weight training data, optimizing translation on all languages.

Multilingual modeling Sharing all parameters except for the attention mechanism shows improvements compared with sharing everything in an RNN NMT model (Blackwood et al., 2018). Sachan and Neubig (2018) explores parameter sharing in the transformer architecture for the decoder in the one-to-many translation setting and shows that transformers are more suitable than RNNs for this task. Also, parameter sharing in the decoder and embedding layer further improves performance. Lu et al. (2018) proposes a shared layer intended to capture the interlingua knowledge and an extension to the typical RNN network with multiple blocks along with a trainable routing network. The routing network enables adaptive collaboration by dynamic sharing of blocks conditioned on the task at hand, input, and model state (Zaremoodi et al., 2018). Zhang et al. (2020a) proposes a language-aware layer to improve such architectures further. With a similar idea, Zhu et al. (2020a) incorporates two special language embeddings into the self-attention mechanism. The first encodes the unique characteristics of each language, while the second captures common semantics across languages.

One problem in multilingual NMT systems is the translation into the wrong language. To address this problem, Zhang et al. (2020b) add a language-aware layer normalization and a linear transformation that is inserted between the encoder and the decoder to induce a language-specific translation. Raganato et al. (2021) explore to weight the target language label with jointly training one cross attention head with word alignments.

Other modifications of NMT model architectures to improve their performance on low-resource languages include: deep RNNs (Miceli-Barone et al., 2017), normalization layers (Ba et al., 2016), direct lexical connections (Nguyen et al., 2015), word embedding layers conducive to lexical sharing (Wang et al., 2019c).

A.2 Extended Multi-task training

Zhou et al. (2019) uses this approach, but extends it with a cascade architecture: the first decoder reads the encoder, and the second decoder reads the encoder and the first decoder (Niehues et al., 2016; Anastasopoulos and Chiang, 2018). The auxiliary task (first decoder) is a denoising decoder. With RNN NMT architectures, one can further decide if the attention mechanism should be shared among tasks (Niehues and Cho, 2017). The authors compare all architectures and find that they perform similarly, with only sharing the encoder being slightly better.

Using linguistic information as an auxiliary task has not yet been explored exhaustively. Niehues and Cho (2017) studies the usage of part-of-speech (POS) and named entity (NE) tags, finding that training on named entity recognition (NER), POS tagging and MT together improves performance the most. For agglutinative languages, morphological auxiliary tasks can be beneficial: Pan et al. (2020) uses stemming with fully shared parameters.

As an alternative to linguistically informed auxiliary tasks Srinivasan et al. (2019) uses multiple BPE vocabulary sizes to generate different segmentations. Each segmentation is treated as an individual task.

A.3 Data augmentation

Back-translation Caswell et al. (2019) shows that adding a special tag to the synthetic data improves performance. A technique that exploits this idea is training an initial translation model with synthetic data generated via BT and then finetune it with gold data (Abdulmumin et al., 2019). This simple yet effective training algorithm improves NMT for LRLs; however, it can also degrade performance on HRLs if trained without a tagging strategy (Marie et al., 2020).

Multiple improvements of BT have been proposed. Edunov et al. (2018) shows that sampling or noisy beam search can generate more effective pseudo-parallel data. However, for LRLs an optimal beam search and greedy decoding are better. A factor that influences BT's effectiveness is the quality of the initial MT systems (Hoang et al., 2018). Using back-translated data from multiple sources (Poncelas et al., 2019) or optimizing the ranking of back-translated data yields further gains (Soto et al., 2020).

BT results in gains when the parallel corpora are naturally occurring text and not translationese, as the latter would only improve automatic n metrics (Toral et al., 2018; Graham et al., 2020). **?** shows that BT produces more fluent text and is preferred by humans. Additionally, translationese and original data can be modeled as separate languages in a multilingual model (Riley et al., 2020). BT is also a central part of unsupervised MT (UMT; cf. §6.4) and zero-shot MT (Gu et al., 2019).

Sentence modification Zhu et al. (2019) proposes to replace a randomly chosen word in a sentence with a *soft-word*. That means that, instead of sampling a word from the lexical distribution of a LM like Kobayashi (2018), the authors use the hidden state vector of the LM directly. Wu et al. (2019) substitutes the RNN LMs from previous work and use BERT (Devlin et al., 2019) – a transformer trained with a masked language modeling objective – instead. The authors finetune BERT with a conditional masked language modeling objective that tries to avoid the prediction of words that do not correspond to the original sentence meaning.

Another way to augmented MT data is by paraphrasing. If a good paraphrase system exists, this can increase the number of training instances (Hu et al., 2019). Paraphrasing can also be used at training time by sampling paraphrases of the reference sentence from a paraphraser and training the MT model to predict the distribution of the paraphraser (Khayrallah et al., 2020). This helps the model to generalize. Wieting et al. (2019) propose a similar approach, using minimum risk training to optimize BLEU. To avoid BLEU's constraints to a specific reference, they use paraphrasing to diversify the given reference.

Finally, existing data can be augmented by adding noise. This noise can be continuous or discrete. In the case of applying continuous noise, noise vectors are added to the word embeddings (Cheng et al., 2018; Sano et al., 2019). Discrete noise is realized by inserting, deleting, or replacing words, BPE tokens, or characters to expand the training set in an adversarial fashion (Belinkov and Bisk, 2018; Ebrahimi et al., 2018; **?**; Cheng et al., 2019, 2020).

Pivoting While it is simple to implement and effective, pivot-based approaches suffer from error propagation. To overcome that for NMT, joint training Zheng et al. (2017); Cheng (2019) and round-trip training (Ahmadnia and Dorr, 2019) have been proposed.

Pivoting with NMT systems has been used for translating Japanese, Indonesian, and Malay into Vietnamese (Trieu et al., 2019), translation of related languages (Pourdamghani and Knight, 2019), multilingual zero-shot MT (Lakew et al., 2018), and UMT (cf. §6.4) between distant language pairs (Leng et al., 2019).

A.4 Recent low-resource Shared Tasks

First, the LoResMT 2020 shared task (Ojha et al., 2020) explores the case of language pairs which have no parallel data between them (Hindi–Bhojpuri, Hindi–Magahi, and Russian–Hindi). The winning system (Laskar et al., 2020) uses a MASS model in a zero-shot fashion with additional monolingual data (see §6.4). Second, the WMT 2020 shared tasks on UMT and very low-resource supervised MT (Fraser, 2020) provide text and 60k aligned phrases for German–Upper Sorbian., The most important technique in all tracks is transfer learning, achieving surprisingly good results. For the AmericasNLP 2021 shared task on open MT (Mager et al., 2021), 10 indigenous language languages were paired with Spanish, resulting in an extreme low-resource setting (4k to 125k paired sentences), with challenges out as domain, dialectical, and orthographic mismatches between splits and datasets. The best systems shows that data cleaning and collection (§**??**) as well as multilingual approaches (§6.1) result in the best performance in this conditions. Finally the shared task on MT in Dravidian languages (Chakravarthi et al., 2021) features 3 languages paired with English as well as Tamil–Telugu. Again, the winning system uses a multilingual approach. The best performing systems use BT (§6.3) and BPE word segmentation (§2.1).

The results from these challenges indicate that the optimal selection and combination of methods differs between cases (i.e., amount of monolingual, parallel data, cleanness of data, domain mismatch, linguistic closeness of languages). This implies that data analysis and linguistic knowledge are needed to improve a final system's performance.

A.5 Transfer learning

This helps low-resource tasks as a lower amount of data can be used for training. One application of transfer learning to MT is the usage of a pretrained RNN LM (Gulcehre et al., 2015) as the decoder in an NMT system. Zoph et al. (2016) is the first work that uses pretrained models to improve NMT systems. The authors perform two experiments with an RNN encoder–decoder architecture with an attention mechanism: the model is first pretrained on a high-resource language pair This works even better if related languages are used during pretraining (Nguyen and Chiang, 2017). Using pretrained LMs at decoding time and as priors at training time also improves vanilla models (Baziotis et al., 2020).

To avoid overfitting, models can be finetuned on both a HRLs pair and a LRLs pair in a multi-task fashion (Neubig and Hu, 2018).

However, how can we represent best the vocabulary? Zoph et al. (2016) use separate embeddings for the source and the target language. However, using tied embeddings has been shown to yield better results (Press and Wolf, 2017). Edunov et al. (2019) employs ELMO (Peters et al., 2018) representations as pretrained features in the encoder of a transformer model. Song et al. (2020) shows that it is possible to improve performance by combining monolingual texts from linguistically related languages, performing a script mapping. It is also possible to extract features from a BERT model in the source language and combining these with an NMT system (Zhu et al., 2020b), but using a BERT model pretrained with a mixed sentences from source and target languages lead to even better results (Xu et al., 2021).

Encoder-decoder pretrained models have gained popularity in the last years for low-resource MT. Conneau and Lample (2019) proposes training the encoder and the decoder separately in order to get cross-language representations (XLM). This idea has further been extended by Song et al. (2019, MASS) to masking a *sequence* of tokens from the input. Training MASS in a multilingual fashion and using monolingual data for pretraining helps to improve NMT for low-resource languages and zero-shot translation (Siddhant et al., 2020). Another approach is to train the entire transformer model as a denoising autoencoder (BART; Lewis et al., 2019). The multilingual version of BART (mBART) is more suitable for NMT tasks and yields important gains (Liu et al., 2020). It is also possible to pretrain a transformer in a multi-task, text-to-text fashion, where one of the tasks is MT (T5; Raffel et al., 2020). All four models can be finetuned for MT or used in an unsupervised fashion. Improvements to BART can be obtained by augmenting the maximum likelihood objective with an additional objective, which is a data-dependent Gaussian prior distribution (Li et al., 2020). Huge LMs can improve zero-shot and few-shot learning even further (Brown et al., 2020), but at a high computational cost. Pursuing another direction, Wang et al. (2019a) develops a hybrid architecture between a transformer and a pointer-generator network. At training time, the authors jointly train the encoder and the decoder in a denoising auto-encoding fashion.

One crucial problem for transfer-learning is minimizing catastrophic forgetting (Serra et al., 2018). Chen et al. (2021) show that it is possible to combine a pre-trained multilingual model, with fine-tuining it with one single language pair, to improve zero-shot machine translation. Another way to handle this problem is reducing the number of parameter to be updated. Gheini et al. (2021) propose to only update the cross attention parameters.

A.6 Unsupervised MT

The addition of other components such as masked LMs and denoising auto-encoding has also been tried (Stojanovski et al., 2019). Unsupervised methods are vulnerable to adversarial attacks of word substitution and order change in the input. Adversarial training can improve performance in such situations (Sun et al., 2020). Since the initialization step is crucial for UMT, Ren et al. (2020) aligns semantically similar sentences from two monolingual corpora with the help of cross-lingual embeddings. With these, an SMT system is trained to warm up an NMT system. However, UMT still has to overcome a set of challenges. Søgaard et al. (2018) shows that performance decays dramatically for languages with different typological features, since, in such situations, bilingual word embeddings (Conneau et al., 2017) are far from isomorphic. Vulić et al. (2020) finds that isomorphism is also less likely if small amounts of monolingual data are used for training bilingual word embeddings. Nooralahzadeh et al. (2020) discovers that performance quickly deteriorates for a mismatch of source and target domain and that the initialization of word embeddings can affect MT performance. All of this makes UMT for LRLs or endangered languages challenging.

Some of the described issues have been addressed: Liu et al. (2019) proposes to combine word-level and subword-level embeddings to account for morphological complexity. For the problem of distant language pairs, Leng et al. (2019) proposes pivoting (cf. §6.3). Isomorphism of bilingual word-embeddings can be improved with semi-supervised methods (Vulić et al., 2019).

Garcia et al. (2020) introduces multilingual UMT systems. The main idea consists of generalizing UMT by using a multi-way back-translation objective. Recently, pretrained multilingual transformer networks are used to improve UMT even further (cf. §6.4).

B Ethical Considerations

Ethical concerns when working on MT for endangered languages include a lack of community involvement during language documentation, data creation, and development and setup of MT systems. For more information, we refer interested readers to Bird (2020). Finally, we want to mention that publicly employing low-quality MT systems for LRLs bears a risk of translating incorrectly or in biased (e.g., sexist or racist) ways.