Massively Multilingual Text Translation for Low-Resource Languages

Zhong Zhou

CMU-LTI-23-014

Language Technology Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:

Alexander Waibel[†‡] (Chair)
Alon Lavie[†]
Graham Neubig[†]
Jan Niehues[‡]

[†] Carnegie Mellon University
[‡] Karlsruhe Institute of Technology

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Language and Information Technologies.

November 29, 2023
DRAFT

Keywords: multilingual machine translation, severely low-resource translation, endangered languages, active learning, large pretrained models, human machine translation, deep learning, neural networks, interlingual transfer, paraphrases, linguistic distance, information dissemination.

To God.

This job has been given to me to do.
Therefore, it is a gift.
Therefore, it is a privilege.
Therefore, it is an offering I may make to God.
Therefore, it is to be done gladly, if it is done for Him.
Here, not somewhere else, I may learn God's way.
In this job, not in some other, God looks for faithfulness.

Elisabeth Elliot

Abstract

Translation into severely low-resource languages has both the cultural goal of saving and reviving those languages and the humanitarian goal of assisting the everyday needs of local communities that are accelerated by the recent COVID-19 pandemic. In many humanitarian efforts, translation into severely low-resource languages often does not require a universal translation engine, but a dedicated *text-specific* translation engine. For example, healthcare records, hygienic procedures, government communication, emergency procedures and religious texts are all limited texts. While generic translation engines for all languages do not exist, translation of multilingually known limited texts into new, low-resource languages may be possible and reduce human translation effort. We attempt to leverage translation resources from rich-resource languages to efficiently produce best possible translation quality for well *known texts*, which are available in multiple languages, in a new, low-resource language.

To achieve this efficiency, we translate a closed text that is known in advance and available in multiple source languages into a new and low-resource language. Despite the challenges of little data and few human experts, we build methods to promote cross-lingual transfer, leverage paraphrase diversity, address the variable-binding problem, measure language similarity, build efficient active learning algorithms for learning seed sentences, activate knowledge in large pretrained models and produce quality translation with as small as a few hundreds lines of low-resource data. Working with extremely small data, we demonstrate that it is possible to produce useful translations for machines to work alongside human translators to expedite the translation process, which is exactly the goal of this thesis.

To reach this goal, we argue that in translating a closed text into low-resource languages, generalization to out-of-domain texts is not necessary, but generalization to new languages is. Performance gain comes from massive source parallelism by careful choice of close-by language families, style-consistent corpus-level paraphrases within the same language and strategic adaptation of existing large pretrained multilingual models to the domain first and then to the language. Such performance gain makes it possible for machine translation systems to collaborate with human translators to expedite the translation process into new, low-resource languages.

Acknowledgments

Many thanks to my advisor and all my committee members and mentors who has helped me grow over the years. This thesis is completed through much help and support from our scientific community that I am deeply grateful to.

I would like to thank my advisor Alex Waibel. Alex is an insightful researcher, an experienced entrepreneur and a proficient writer. I am grateful that he shares my research goal. Alex has generously funded the entire work, has helped me with academic writing and has purchased multiple machines for our research over the years. His support is pivotal in this thesis.

I am thankful to my committee members and my collaborators, Alon Lavie, Graham Neubig, Jan Niehues, Matthias Sperber, and Mark Bean. I am deeply grateful to Alon's sharing of wisdom on growing scientific career on Machine Translation evaluation, Graham's help with learning time management and prioritization skills, Jan's insights in prioritizing on key results in paper writing, Matthias' consistent support and Mark's expertise in Quechuan languages.

I am deeply grateful to the support of Carolyn Rosé and Jamie Callan. Both of them share wisdom and strengthen me to grow and mature in research. I will pass on their kindness and wisdom to others.

We stand on the shoulder of giants, and I would like to thank all brilliant minds in the fields who help me grow over the years. I want to thank Ramayya Krishnan, Tom Mitchell, Alan Black, Rita Singh, Lori Levin, Uri Alon, David Mortensen, Rema Padman, Rahul Telang, Brian K. Kovak, David Choi, Steven Shreve, Larry Wassermann and many others for their advice and mentorship.

I am so thankful to Angela Lusk, Jonny Cagwin and Suzie Laurich-McIntyre. They helped me learn and grow over the years. They are exemplary in their kindness, wisdom, character, strength and their dedication to student well-being. I want to thank Kevin Haworth, Keely Austin, Michael Laudenbach, Rose Chang, Laura DeLuca, Elizabeth Dietrich, Christian Hallstein and Jessica Hsu for improving my academic writing skills.

Thanks to family and friends especially Christian and Shirley Hallstein, Paul and Sharon Johnston, Krissy Geffel, Ong Ai Boon, Jimmy and Valerie Williams, Cammie Dunaway, Linda and Fred Griffin, Shannon Libengood, Kristen Emrick, Jessica Hsu, Abigail Holizna, Janice Turner, Bri Saleone, Michael and Denise Danko, Jeff Bergeson, Brenda Miller, Liam Ain Levi, Jamie Ian Joel, Ryan Len Reese and the Schöpfle family, and many who help me grow.

Publications

Parts of this thesis have previously appeared in the following publications:

1. Zhong Zhou, Matthias Sperber, and Alex Waibel. Massively parallel cross-lingual learning in low-resource target language translation. In *Proceedings of the 3rd conference on Machine Translation. Association for Computational Linguistics*, 2018.
2. Zhong Zhou, Matthias Sperber, and Alex Waibel. Paraphrases as foreign languages in multilingual neural machine translation. In *Proceedings of the Student Research Workshop at the 56th Annual Meeting of the Association for Computational Linguistics*, 2019.
3. Zhong Zhou and Alex Waibel. Family of origin and family of choice: Massively parallel lexiconized iterative pretraining for severely low resource text-based translation. In *Proceedings of the 3rd Workshop on Research in Computational Typology and Multilingual NLP in the 20th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, 2021.
4. Zhong Zhou and Alex Waibel. Active learning for massively parallel translation of constrained text into low resource languages. In *Proceedings of the 4th Workshop on Technologies for Machine Translation of Low Resource Languages in the 18th Biennial Machine Translation Summit*, 2021.
5. Zhong Zhou, Jan Niehues, and Alex Waibel. Train global, tailor local: Minimalist multilingual translation into endangered languages. In *Proceedings of the 6th Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT) of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023.

Contents

All figures, graphs and visuals in this thesis except photographs are created by the author. All photographs except those in Figure 1.5, Figure 2.3, and Figure 7.1 are taken by Mark Bean in Peru, reproduced with permission. Figure 1.5, Figure 2.3, and Figure 7.1 are provided by Mark Bean, reproduced with permission. Photographs with people are included with the permission of the Quechuan language communities involved. In all photograph captions, "Panao" refers to Panao Quechua, "Sihuas" refers to Sihuas Quechua and "Margos" refers to Margos-Yarowilca-Lauricocha Quechua. For permissions, please contact the author.

List of Figures

List of Tables

Chapter 1

Introduction

> "To have another language is to possess a second soul."
>
> *Charlemagne*

Translation into severely low resource languages has both the cultural goal of saving low-resource languages and the humanitarian goal of assisting the everyday needs of local communities that are accelerated by the recent COVID-19 pandemic. In many humanitarian efforts, translation into severely low resource languages often does not require a universal translation engine, but a dedicated *text-specific* translation engine. For example, healthcare records, hygienic procedures, government communication, emergency procedures and religious texts are all limited texts. Translation of limited texts have many real-world applications. One such application is the translation of water, sanitation and hygiene (WASH) guidelines to protect Indian tribal children against waterborne diseases and more recently COVID-19 infections, introducing earthquake preparedness techniques to Indonesian tribal groups living near volcanoes and delivering information to the disabled or the elderly in low-resource language communities in Uganda [11, 201, 221, 239]. These are useful examples of translating a closed text known in advance to the severely low-resource language. We show "wash your hands" in many languages in Figure 1.1.

While generic translation engines for all languages do not exist, translation of multilingually known limited texts into new, low-resource languages may be possible and reduce human translation effort. We attempt to leverage translation resources from rich-resource languages to efficiently produce best possible translation quality for well *known texts*, which are available in multiple languages, for a new, severely low-resource language.

To achieve better efficiency, we translate a closed text that is known in advance into a new and severely low-resource language by leveraging massive source parallelism. In other words, we make use of available translations in many known source languages to produce

Figure 1.1: "Washing your hands" in world languages [81].

a good translation in severely low-resource language. The source languages may be rich-resource, but may include other low-resource languages that have slightly more data or human expertise.

In this problem setup, given a text that is multilingually available, we are interested in translating it into a new, low-resource language. The problem has three unique aspects that are different from traditional Machine Translation (MT) problems:

1. Our text is closed, not arbitrary as in traditional MT problems.
2. Our text has multiple source languages with complete text translations while traditional MT is typically single-source.
3. Our text has little to no translation in the target low-resource language, while traditional MT assumes abundant data.

1.1 Thesis Statement

Machine Translation focuses on building a practical solution for human communication across cultures [141, 194, 229, 296]. Translation into severely low-resource languages is ultimately an extremely small data problem. The methods that usually work in big data settings may not work for severely low-resource scenarios. Many large and evolved models may not perform as well as small and simple models in cases with little to no data. To overcome this challenge of extremely small data, our focus is to build a practical text-based Machine Translation engine that joins forces with human translators and uses minimal resources to expedite the translation process into severely low-resource languages.

THESIS STATEMENT *In translating a multilingually known limited texts into a new, low-resource language, we argue that generalization to out-of-domain texts is not necessary, but generalization to new languages is necessary. Performance gain comes from massive source parallelism through the following: 1) close-by language families, 2) style-consistent corpus-level paraphrases within the same language, 3) carefully-constructed linguistic closeness, 4) selective choice of active learning methods, and 5) strategic adaptation of existing large pretrained multilingual models to the domain first and then to the language. Such performance gain makes it possible for machine translation systems to collaborate with human translators to expedite the translation process into new, low-resource languages.*

While the industry trend is to move towards bigger models with bigger data, our approach uses fewer languages, smaller data and minimal expert efforts. Given that expert efforts are mostly compensated or measured by the number of translates or edits, this saves computation power and resources. Therefore, this saves time and money, while improving translation performance. Saving time and money helps low-resource language communities to thrive on limited resources. Furthermore, with adequate scheduling and suitable model training on the whole text of multiple source languages, it is possible to build a sufficiently good translation model based on little data in a new and severely low-resource language. This does not mean that we can build a universal interlingua using such small data as the good translation results produced by learning about the text may not generalize to other texts. Nevertheless, generalization to other texts is desirable but not necessary in our goal of producing practical and high quality translation of the given closed text in our research problem. In this thesis, instead of focusing on translating any text into any language, we focus on the practical goal of translating a given text into a new, low-resource language.

1.2 Thesis Overview

To realize the goal of expediting the translation process of a multi-source text into new, low-resource languages, we show an overview of the work done as part of this thesis and how they are contributing to our main goal in Figure 1.3. Following this introduction, we examine related existing research work in Chapter 2. After literature review, we then present this thesis in two main parts: massively multilingual translation (Chapter 3, Chapter 4 and Chapter 5), and human machine translation (Chapter 6, Chapter 7 and Chapter 8). We focus on a case study in Quechuan language family for applying the methods built in this thesis to the real-world translation in Chapter 8. All these chapters contributes to the main goal of this thesis: translation of multilingually known limited text into new, low-resource languages. To conclude, Chapter 9 summarizes our main contributions and explores future research directions and opportunities in this research space.

Figure 1.2: Quechuan language community in Peru. Photograph by Mark Bean.

We focus on two parts of this thesis: massively multilingual translation (Chapter 3, Chapter 4 and Chapter 5), and human machine translation (Chapter 6, Chapter 7 and Chapter 8) and give an overview of these two parts below.

- **Massively multilingual translation** (Part I): We examine how source parallelism benefit translation of a given text into new, low-resource languages through multilingual training. In Chapter 3, we build cross-lingual transfer both within a given language family and also across different language families. We find that in practice, training on two closely related language families or equivalently around ten closely related languages is often enough for translating to a given low-resource language. We also propose an order-preserving named entity translation mechanism to resolve the variable binding problem and produce high quality lexiconized translations under severely low-resource scenarios. When data of similar languages is not available, we may leverage different translations of the same text in the same language as described in Chapter 4. We treat paraphrases as foreign languages, and train on corpus-level paraphrases to improve translation performance. We find that our multi-paraphrase translation models improve performance better than multilingual models and improves the sparsity issue of rare word translation as well as diversity in lexical choice.

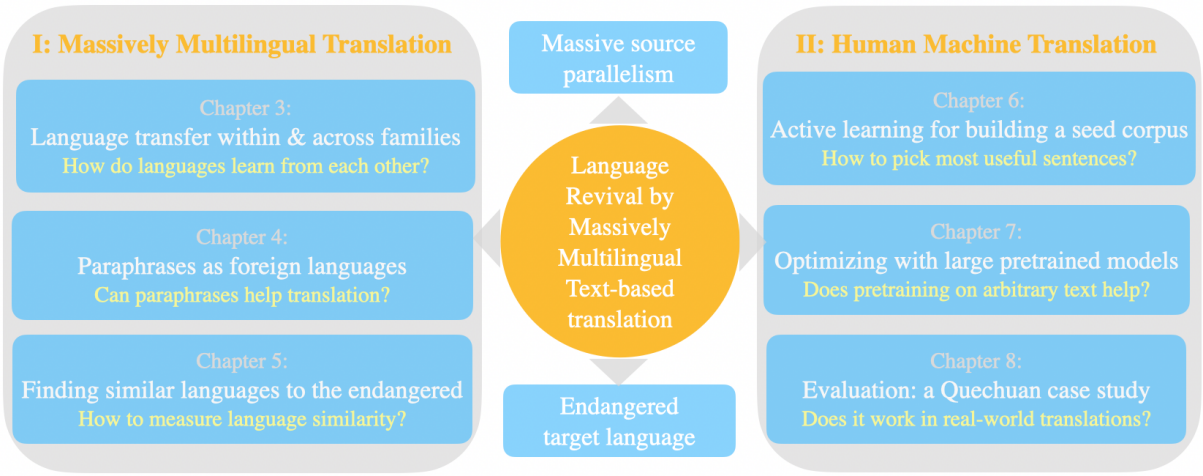


Figure 1.3: Overview of the work done as part of this thesis.

When paraphrase information is also not available, we may build our own linguistic distance metric based on translation distortion, fertility and performance. In Chapter 5, we propose a method, *Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer*, to train on low-resource language data. We push the limit by using only ∼1,000 lines (∼3.5% of the entire text) to translate the whole text and achieved good translation performance.

- **Human Machine translation** (Part II): Having examined source parallelism, we build a human machine translation workflow algorithm for machine translation systems to collaborate with human translators to expedite the process. Our proposed human machine translation is not to replace the human translators with machine translation systems, but instead, to get the best of both worlds as shown in Figure 1.4. In our translation process, human translators are informed by machine sentence ranking through active learning to produce a seed corpus. Machine systems then use this seed corpus to produce a full translation draft. Human translators post-edit the draft, and feed new data to machines each time they finish post-editing a portion of the text. In each iteration, machines produce better and better drafts with new data, and human translators find it easier and faster to post-edit. Together they complete the translation of the whole text into a severely low-resource language. We first develop various active learning methods on known languages and transfer ranking to the new, low-resource language in Chapter 6. Secondly, we activate the knowledge of large multilingual models by proposing multilingual and multi-stage adaptations through different training schedules in Chapter 7; we find that adapting pretrained models to the domain and then to the low-resource language works best. Thirdly, we aggregate scores from 115 languages to provide a universal ranking and increase

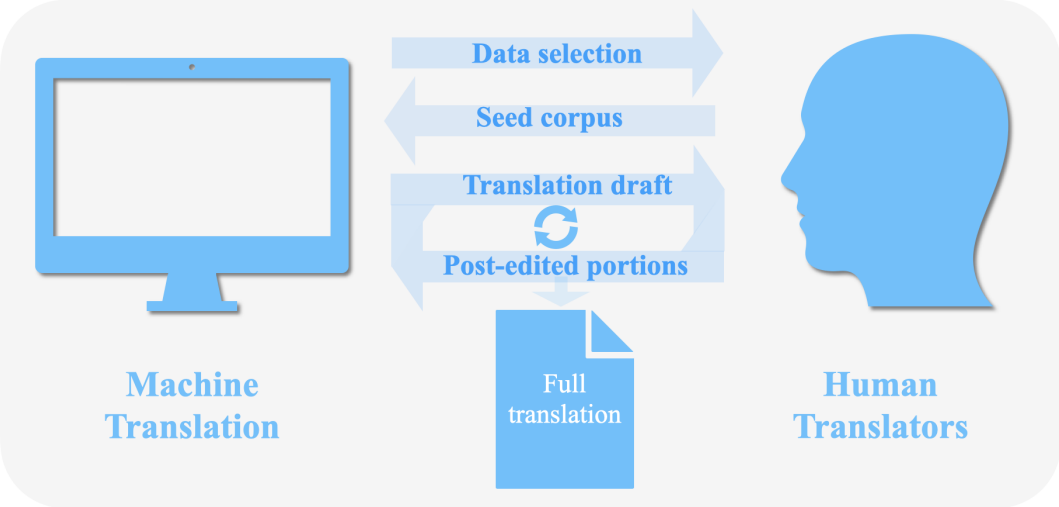


Figure 1.4: Human machine translation process.

robustness by *relaxed memoization* method. In Chapter 8, having examined both source parallelism and human machine translation workflow, we evaluate our work in all previous chapters by translating it into practical use in a case study in Quechuan language family in extensive collaboration with Mark Bean and his translation group with in-depth knowledge of various Quechuan languages. We find that machine translation performance is significantly positively correlated with language similarity. The more connected a language is, the higher the performance. Moreover, we find that decluttering poorly-connected languages improves translation score. Based on this finding, we show effectiveness of our models through good results in translation into a new, low-resource language called Sihuas Quechua.

Finally, we summarize our contributions and discuss future research opportunities in translating into new, low-resource languages in Chapter 9.

1.3 Practical Goal Setting

Having understood the overview of this thesis, we would like to clarify the goal of this work and how to make it practical and attainable given the real-life constraints of low-resource languages.

In essence, our goal is to minimize human translation and post-editing efforts required to generate a full publishable-standard translation of a given text. We aim to minimize human translators' efforts in both the translation process of the seed corpus sentences and the post-editing process of the subsequent iterations. In other words, we want to make translation and post-editing work easier and smaller for human translators through automation.

To realize this goal, ideally we want to hire a large number of human translators, measure and compare the resources (time and money) used to translate the same text into a target low-resource language that does not have any translations of the text. To show that this thesis has achieved the goal of reducing human translation efforts, optimally we need to measure and compare time taken and money spent in two translation scenarios:

1. Baseline: total time taken and total amount of money paid when human translators translate the text to meet publishable standards.
2. Total time taken and total amount of money paid when human translators work with machine translation systems to translate and post-edit the text to meet publishable standards through the human machine translation algorithm introduced in this thesis.

However, this ideal solution is unrealistic especially in large translation projects. Large translation projects in real-life usually takes decades, if not centuries of work. For example, in the case of Bible Translation, the average project cost of a complete written Bible is $937,446 and the average time of completing a Bible translation with sufficient resources is 15.8 years [142]. This is clearly not realistic and beyond the scope of this thesis.

To set more practical goals, we treat the translation process of the seed corpus and the post-editing process of the subsequent iterations separately, and use the following two sub-goals as the proxy sub-goals for minimizing time and money used for the translation project:

1. To minimize human translation effort at the translation process of the seed corpus, we optimize and minimize the amount of sentences to be used to construct seed corpus as a proxy.
2. To minimize human post-editing process of the subsequent iterations, we maximize the quality and utility of MT-generated translation of the full text and optimize translation efficiency.

To understand why we choose these two sub-goals as proxies, we first need to understand that most of the human translators are paid by the number of words they translate or post-edit [33, 82, 288]. The rate they translate a word is different from the rate they post-edit a word [82]. Therefore, the total human translation cost is entirely tied to the number of translated words and post-edits human translators need to perform to meet publishable standards. Firstly, minimizing the amount of sentences used to construct seed corpus directly reduces the number of words human translators need to translate. Therefore, it serves as a good proxy for minimizing human translation efforts of the seed corpus. Secondly, maximizing translation performance of the MT-system minimizes the number of edits required for generating a full publishable-standard translation of the given text [25, 26, 265, 292]. Since time and money saved is directly linked with the number of edits

saved, maximizing translation performance is an appropriate proxy sub-goal for minimizing human post-editing processes.

Using the proxy sub-goals, we transform our goal of minimizing human translation efforts required to generate a full translation of the given text into two practical proxy sub-goals as the following:

1. Optimizing and minimizing the amount of sentences used to construct seed corpus.
2. Maximizing the quality and utility of MT-generated translations of the full text and optimizing translation efficiency.

The first sub-goal of minimizing the seed corpus serves as a proxy in Chapter 6 to minimize the human translation efforts in the creation of the seed corpus, while the second sub-goal of maximizing translation performance serves as a proxy in Chapter 7 to minimize human translation efforts in the post-editing process during the subsequent iterations.

To measure translation performance, our primary automatic metric in this thesis is chrF [230]. We choose chrF for accuracy, fluency and expressive power in morphologically-rich languages [217]. We use the metric chrF for the beginning and the conclusion of this thesis to motivate and summarize our main contributions of this paper. In addition to our main evaluation metric chrF, in order to have a comprehensive understanding of each chapter, we give a wide range of evaluation metrics to supplement our understanding of translation performance through BLEU , characTER, COMET score, and BERTscore [231, 241, 275, 305, 317]. Moreover, we explore human evaluation methods in addition to automatic evaluation methods. For detailed analysis, we prioritize BLEU in Massively multilingual translation (Part I) as we mainly work with European languages while we prioritize chrF in Human Machine translation (Part II) as we mainly work with morphologically-rich low-resource languages. Overall, we will use our metric chrF to conclude this thesis.

1.4 Thesis in Practice

As we have discussed in the Thesis Overview, we integrate this thesis into real-world scenarios by examining a case study in Quechuan language family in Chapter 8. This provides practical insights in understanding how to use this thesis for practical situations. We will discuss the case study in Quechuan language family in detail at the end of our thesis. However, we would like to give a glimpse to this case study before we dive into the details of the different chapters of this thesis, so that we have a more intuitive understanding of the problem we are solving.

The Quechuan language family, also called Runasimi ("people's language") by local communities, is a varied group of languages covering a wide region of Peruvian Andes in South America, extending from Colombia, Ecuador, Peru, Bolivia to northwest Argentina [137, 184]. Its history traces back to the Inca Empire and there is broad spectrum of

Figure 1.5: A practical application of this thesis in Peru. Photograph provided by Mark Bean.

sociolinguistic diversity among Quechuan languages throughout the Spanish colonial history of the area [77]. Many are low-resource languages.

We make practical use of findings of this thesis to translate into a few low-resource languages in Quechua in Chapter 8. We provide extensive analysis to answer the following questions:

1. Under what conditions does our method work?
2. Under what conditions does our method fail to work?
3. When it works, does it provide real help to human translators working in the field?
4. When it does not work, is there ways we can improve user experience for human translators?

Through the extensive analysis to answer the above questions, we find that machine translation performance is significantly positively correlated with language similarity. The more connected a language is, the better it is to translate into this language. Using this result, we employ our thesis in practice and achieve good performance for translation into a new, low-resource language called Sihuas Quechua. We give a glimpse of this thesis in practice before we dive into the details of each chapter, and we will look closely at this Quechuan case study after the detailed discussion of this thesis in Chapter 8. Through this case study, we show a few opportunities for further research that this thesis present.

1.5 How to Read This Thesis

Given the overview, we aim to make this thesis accessible and inclusive of readers coming from diverse research backgrounds for the many potential collaborative opportunities in this research space.

Following this introduction, we give an overview of the research landscape by examining related existing research done in the space of translation into low-resource languages in Chapter 2. For Chapter 3 to Chapter 8, each chapter is relatively self-contained and readers can read them not in sequence. However, we suggest to read Chapter 3 and 4 first before Chapter 5 to gain an understanding what could be done with language closeness and why it is important to measure language closeness when such information is not available. We also recommend to read Chapter 6 and Chapter 7 together as most of the work in these two chapters have previously appeared in publications together. Lastly, we recommend reading Chapter 8 after having completed all the previous chapters as this chapter leads us from the comfort zone of academic research to enter the real-world translation process, which is exactly the place our work is most impactful.

Chapter 2

Literature Review

> "Every act of communication is a miracle of translation."
>
> *Ken Liu*

To understand the existing literature in the space of translation into low-resource languages, we focus on examining different aspects of our translation task through related research works. We only show related works that are relevant to all parts of this thesis in this chapter. For related works that are only relevant to a particular chapter, we will introduce them in the specific chapter directly.

In this chapter, we show related works that are closely connected to all parts of this thesis in three broad categories: low-resource languages, machine translation, and translation in practice. These three categories are intricately linked to each other. Firstly, related works in "low-resource languages" show the broader impact of empowering low-resource language communities and facilitating information dissemination in these communities. Secondly, related works in "Machine Translation" demonstrate the state-of-the-art tools researchers have built that could be applied in the severely low-resource translation, including massively multilingual translation and large pretrained multilingual models. Finally, related works in "translation in practice" bridge the two worlds by putting MT tools into practical use of real-world translation. More specifically, we show the related works in the areas of human machine translation, active learning and post-editing.

Having examined related works in low-resource languages, machine translation, and translation in practice, we conclude this chapter by introducing our research framework that is built based on existing literature, including the toolkits we use, our core datasets and evaluation metrics we use.

2.1 Low-Resource Languages

2.1.1 Information Dissemination

Interactive Natural Language Processing (NLP) systems are classified into information assimilation, dissemination, and dialogue [30, 238, 302, 303, 304]. Assimilation and dissemination have very broad definition and our definition below assumes the reference point of rich-resource communities and discusses them only in the context of rich and low-resource languages. *Information assimilation* involves information flow from low-resource to rich-resource language communities while *information dissemination* involves information flow from rich-resource to low-resource language communities. It helps low-resource language communities to make better-informed and autonomous decisions. Taken together, they allow *dialogue* and interaction of different groups at eye level. Most research is on information assimilation with examples span from urgent earthquake detection to infectious disease surveillance [27, 40, 79]. Few work on dissemination is on information dissemination with examples ranging from introducing disaster prevention techniques to delivering information to the disabled and the elderly in low-resource language communities [1, 12, 19, 213, 328]. Information dissemination is challenging because there is little low-resource language data, much less parallel corpus, little funding, and few human experts in training and evaluation [12, 75]. Some low-resource languages have no formal writing systems [1, 19]. Note that in a broader definition of dissemination, the field is largely under-researched even if the focus is not low-resource. However, our discussion focuses largely on dissemination in the severely low-resource scenarios.

2.1.2 Low-Resource Languages

A language is alive when many people speaks it and dies when no one speaks it. There are ∼7,139 languages in the world, unequally distributed across the world, with drastic differences in their number of speakers and vitality [111, 200, 219]. More than half of these languages will die in the next 80 years [15, 81]. Though there are cases where a language dies through war, genocide, natural disasters or infectious diseases, most languages dies while the speakers do not; the speakers either voluntarily or are forced to speak another mainstream language as part of the endangerment process that is deeply rooted in political, historical, social and economical reasons [187, 283].

A language needs attention when it is spoken by enough people that it could survive under favorable conditions but few or no children are learning it [62, 154, 313]. Such languages may survive and thrive if they gain prestige, power and visibility [62]. Language preservation is therefore an intricate and complex matter that invites many different views [14, 71, 122, 166, 245].

Figure 2.1: A native man in Peru reading translated text. Photograph by Mark Bean.

2.2 Machine Translation

2.2.1 Massively Multilingual Translation

Machine polyglotism, training machines to be proficient in many languages, is a new paradigm of multilingual NMT [4, 6, 69, 94, 103, 117, 147, 159, 290, 323, 327]. The objective is to translate from any of the input languages to any of the output languages [94]. Many multilingual NMT systems involve multiple encoders and decoders [117], and it is hard to combine attention for quadratic language pairs bypassing quadratic attention mechanisms [94]. In multi-source scenarios, multiple encoders share a combined attention mechanism [327]. In multi-target scenarios, every decoder handles its own attention with parameter sharing [69]. Attention combination schemes include simple combination and hierarchical combination [175].

Attentional Neural Machine Translation (NMT) is trained directly in an end-to-end system and has flourished recently [178, 260, 312]. The state-of-the-art of multilingual NMT is adding source and target language labels in training a universal model with a single attention scheme, and Byte-Pair Encoding (BPE) is used at preprocessing stage [117]. This method is elegant in its simplicity and its advancement in low-resource language translation as well as zero-shot translation using pivot-based translation scheme [147]. However, these works have training data that increases quadratically with the number of languages [69, 103], rendering training on massively parallel corpora difficult.

2.2.2 Large Pretrained Multilingual Models

As discussed before, the state-of-the-art multilingual machine translation systems translates from many source languages to many target languages [94, 117, 147, 323, 327]. The bottleneck in building such systems lies in the limitations of computing power, data, and machines. The training data quadratically increase with the number of languages, rendering training on many languages difficult. Companies equipped with vast multilingual datasets and substantial computational capabilities have built and released some large pretrained multilingual models for researchers to work on [181, 278]. These are very helpful for researchers to build on this pretrained model to further work in multilingual translation. M2M100, released by Facebook, is trained in 100 languages [85, 91, 254] and covers a few low-resource languages. DeltaLM, released by Microsoft, is pre-trained in 6 TB of multilingual data from CC100, CC-Net, and Wikipedia in 100 languages [185] and allows for flexible input languages. NLLB (No Language Left Behind), released by Meta AI, is trained on 200 languages [61]. With more advances in the future, there will be models covering more languages.

2.2.3 Low-Resource Machine Translation

Recent advances have succeeded in building multilingual methods to translate from multiple rich-resource languages to a new, low-resource language [94, 117, 147, 322, 323]. Many have demonstrated good transfer learning to low-resource languages as few as ∼4,000 lines [177, 234] and ∼1,000 lines [321] of data. Transliteration can replace parallel corpora in translation [150]. Many use simulated low-resource languages, some use real low-resource languages [32, 148, 306].

In addition to researching and training on extremely small data, some researchers work on zero-shot learning [47, 48, 150, 206, 224, 225, 316]. Zero-shot translation in severely low-resource settings exploits the massive multilingualism, cross-lingual transfer, pretraining, iterative back-translation and freezing sub-networks [22, 52, 72, 170, 174, 177, 209, 223, 284, 308, 310]. However, zero-shot learning is volatile and unstable, and is not suitable for tasks that requires high level of accuracy [244], which is crucial in our translation goal. Therefore, instead of zero-shot learning, we choose to use extremely small data instead.

2.3 Translation in Practice

2.3.1 Human and Machine Translation

Machine translation began about the same time as the first computer [133, 229]. Over the years, human translators have different reactions to machine translation advances, mixed

with doubt or fear [141]. Some researchers study human translation taxonomy for machine to better assist human translation and post-editing efforts [44, 66]. Human translators benefit from machine assistance as human individual bias and translation capacity limitations are compensated for by large-scale machine translation [37, 38, 160, 162, 173, 251]. On the other hand, machine translation benefits from professional human translators' context-relevant and culturally-appropriate translation and post-editing efforts [141]. Severely low-resource translation is a fitting ground for close human machine collaboration [44, 194, 326].

2.3.2 Active Learning

Active learning has long been used in machine translation [8, 83, 100, 107, 118, 138, 197, 262]. Random sampling and data selection through active learning has been surprisingly powerful [55, 114, 127, 135, 152, 158, 259]. Some researchers choose to train on simpler data with shorter sentences through curriculum learning and active learning [83, 228]. The mathematician Donald Knuth uses the population of Menlo Park to illustrate the value of random sampling [158]. The population of Menlo Park is approximately the same as the number of Bible verses (∼31,000). Knuth claims that random sampling across the city is likely to produce a better set of unbiased, reliable and independent statistical samples than choosing people working in the same building.

There is extensive research to beat random sampling by methods based on entropy [164], coverage and uncertainty [220, 318], clustering [100, 119], consensus [118], syntactic parsing [197], mixture based on density and diversity [9, 72, 164], and learning to learn active learning strategies [180].

2.3.3 Post-editing

Human post-editing helps neural MT systems to produce publishable materials [66]. Large MT systems often have issues with under-prediction, over-prediction, repetition, mislabelled data, and hallucination [36, 191]. The state-of-the-art MateCat Tool uses translation memories (TMs) as a reference point to search for matches, exact or approximate, of the current TM segment for translation [93]. It successfully integrates suggestions found TMs and helps human translators and provides a simple and useful interface for human translators to finish the post-editing process [93]. Other tools include PET, CATaLog Online, and iterative interfaces [17, 215, 272]. Many evaluate these tools for productivity and quality gain [115, 116, 210, 299].

Some researchers also work on automatic post-editing to improve MT performance through online learning approach and neural programmer-interpreter method [96, 301]. The post-editing feedback is often used to build adaptive MT systems in real time [67].

In addition, there is also research down for confidence measure on the translations on how much post-editing is required. One such example uses the human-targeted Translation Edit Rate (HTER), on both the word, and the character level [270, 305].

2.4 Research Framework

After reviewing relevant literature on language revival, machine translation and translation in practice, We introduce a few tool kits that we use in this thesis, the core datasets that we work with, and the evaluation metrics we use.

2.4.1 Tools

We use several natural language processing and translation tools for our translation systems. They are listed below.

1. *Fairseq*: a neural sequence learning framework on Pytorch for machine translation, language modelling and other generation jobs [214]. Fairseq allows us to work with a few large pretrained multilingual models and finetune those models to customize to our translation tasks. Fairseq is licensed under the MIT License.
2. *OpenNMT*: an open-source Neural Machine Translation (NMT) system [157]. It is very flexible with many severely low-resource languages. OpenNMT-py is the PyTorch version of the OpenNMT project. OpenNMT/OpenNMT-py is licensed under the MIT License.
3. *Sockeye*: an open-source neural machine translation framework on Pytorch [132]. Sockeye is licensed under the Apache License 2.0.
4. *Googletrans*: a free and simple python library for Google Translate API. This serves as a way to partially evaluate languages that our team do not speak or cannot find native speakers in. When we evaluate the Google's English translation of our system translations into low-resource languages, we are aware that the final outputs introduce another layer of potential error from Google's system. This is therefore used with caution. This does not compete with automatic evaluation methods, and is not comparable with real human qualitative evaluation, but it serves as an imperfect way to understand the overall performance of the system. Googletrans is licensed under the MIT License.
5. *COMET*: a neural framework for MT evaluation [240]. COMET offers a neural way of training multilingual MT evaluation models that attains levels of correlation with human evaluation for some of the rich-resource languages. COMET is licensed under the Apache License 2.0.

6. *MT Telescope*: a toolkit for comparing translations from different MT systems [241]. We use MT telescope to compare across multiple translation systems using a myriad of metrics including COMET above. MT-Telescope is licensed under the Apache License 2.0.
7. *Compare-MT*: a toolkit for comparing outputs from multiple systems for language generation tasks including MT, summarization, dialog generation and others [207]. Compare-mt is licensed under the BSD 3-Clause "New" or "Revised" License.
8. *Fast_align*: a light-weight, quick, and unsupervised word aligner [78]. Fast_align is licensed under the Apache License 2.0.
9. FastText: an open-source, free, simple framework that helps with learning text representations and text classifiers [35, 149]. FastText is MIT-licensed.
10. *SentencePiece*: a light-weight and unsupervised text tokenizer and detokenizer for neural generation systems with a fixed vocabulary size []. SentencePiece is licensed under the Apache License 2.0.
11. *Subword-nmt*: a simple word segmenter [260]. This is uses in our system as an alternative tool to SentencePiece. Subword-nmt is licensed under the MIT License.
12. *KenLM*: a Language Model (LM) toolkit [128, 129]. KenLM is released mostly under the GNU LGPL license, and is distributed under the GNU Lesser General Public License 2.1.
13. *NLTK*: a natural language processing platform that works seamlessly with python codes [31]. We use the LM from NLTK alongside KenLM above. We also use the automatic evaluation metrics from NLTK for MT evaluation. NLTK is licensed under the Apache License 2.0.
14. *Moses*: an open-source toolkit for Statistical Machine Translation (SMT) [163]. Since we are working the severely low-resource scenario, our low-resource data is extremely small. Statistical models which are often based in frequency of words is very helpful in text processing. One such feature we use often in our research is truecaser, where the first word of every sentence is normalized to its most frequent form. This prevents a named entity at the beginning of the sentence to be lowercased. Moses is licensed under the GNU LGPL.
15. *Lang2vec*: a lightweight library for querying the URIEL typological database, a carefully structured resource on language typology and universals [179, 188]. Though not every low-resource language we are working on is represented by Lang2vec, it does cover a broad set of languages. It focuses on the typology and various linguistic features of each language. Lang2vec is licensed under Creative Commons Attribution Share Alike 4.0 International.

2.4.2 Data

Our main text is the Bible in multiple languages [196]. Existing research classifies world languages into Resource 0 to 5, with 0 having the lowest resource and 5 having the highest [148]. We choose a few target languages as severely low-resource languages, both real and hypothetical, ranging from Resource 0 to 5. Each severely low-resource language seed corpus contains a extremely small portion (as low as ∼3%) the text, while all other languages have full text. Our goal is to translate the rest of the text into the severely low-resource language.

2.4.3 Baseline Systems

In our setup we have the new, low-resource language as the target language, and we have a few neighboring languages as the source languages that are either in the same linguistic language family or geographically close to facilitate linguistic transfer. In effect, we have a few source languages with full translations of the text and a new and low-resource language that has an extremely small seed corpus. We use the state-of-the-art multilingual machine translation systems including transformers prepending both source and target language labels to each source sentence with BPE [117, 147, 165, 260]. For precise translation for all named entities, we use an existing method of *order-preserving named entity translation* by masking each named entity with ordered `__NE`s using a parallel multilingual lexicon table in all available languages [311, 321]. We use BPE in our models [165, 260].

2.4.4 Automatic Evaluation

We use chrF as our main metric for introduction and conclusion of the key summary of the contribution of this thesis [230]. We choose chrF for accuracy, fluency and expressive power in morphologically-rich languages [217]. Additionally, we use a complete set of automatic evaluation metrics to supplement our understanding of translation performance: chrF, characTER, BLEU, COMET score, and BERTscore [230, 231, 241, 275, 305, 317]. For detailed analysis, we prioritize BLEU in Massively multilingual translation (Part 1) as we mainly work with European languages while we prioritize chrF in Human Machine translation (Part 2) as we mainly work with morphologically-rich low-resource languages. And we will use our overall metric chrF to conclude this thesis.

Part I

Massively Multilingual Translation

In the first part of this thesis, we explore how source parallelism benefit translation of a given text into new, low-resource languages through multilingual training. In Chapter 3, we build cross-lingual transfer both within a given language family and also across different language families. We also propose an order-preserving lexiconized machine translation model to resolve the variable binding problem, producing high quality lexiconized translations under severely low-resource scenarios. In Chapter 4, we treat paraphrases as foreign languages, propose a multi-paraphrase translation model which trains on corpus-level paraphrases to improve translation performance. We find that our multi-paraphrase translation models improve performance better than multilingual models and improve the sparsity issue of rare word translation as well as diversity in lexical choice. In Chapter 5, we build our own linguistic distance metric based on translation distortion, fertility and performance. We propose a method, *Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer* (IPML), to train on the low-resource language data. We use only ∼1,000 lines (∼3.5% of the entire text) to translate the whole text and achieved good translation performance.

Chapter 3

Language Transfer within and across Families

> "Learning, learned people knew, was a multilingual enterprise."
>
> *Michael D. Gordin*

Having examined closely the existing literature in space of translation into low-resource languages, we investigate intra-family and inter-family transfer in translating a multi-source, closed text into a new, low-resource language.

3.1 Introduction

We work on translation from a rich-resource language to a low-resource language. There is usually little low-resource language data, much less parallel data available [12, 75]; Despite the challenges of little data and few human experts, it has many useful applications. Applications include translating water, sanitation and hygiene (WASH) guidelines to protect Indian tribal children against waterborne diseases, introducing earthquake preparedness techniques to Indonesian tribal groups living near volcanoes and delivering information to the disabled or the elderly in low-resource language communities [11, 20, 221, 239]. We show part of the lyrics of a hand-washing song in a few low-resource languages in Table 3.2 [281]. These are useful examples of translating a closed text known in advance to the low-resource language.

There are three main challenges. Firstly, most of previous works research on individual languages instead of collective families. Cross-lingual impacts and similarities are very helpful when there is little data in low-resource language [45, 63, 130, 211, 250, 266, 271, 289]. Secondly, most of the multilingual translation works assume the same amount of training

Figure 3.1: A low-resource language community in Peru gathering to celebrate together. Photograph by Mark Bean.

data for all languages. In the low-resource case, it is important to exploit low or partial data in low-resource language to produce high quality translation. The third issue is the variable-binding problem that is common in neural systems, where "John calls Mary" is treated the same way as "Mary calls John" [95, 109]. It is more challenging when both "Mary" and "John" are rare words. Solving the binding problem is crucial because the mistakes in named entities change the meaning of the translation. It is especially challenging in the low-resource case because many words are rare words.

Our contribution[1] in addressing these issues is three-fold, extending from multi-source multi-target attentional models. Firstly, to examine intra-family and inter-family influences, we add source and target language family labels in training. Training on multiple families improves BLEU score significantly; moreover, we find training on two neighboring families closest to the low-resource language gives reasonably good BLEU scores, and we define neighboring families closely in Section 3.3.2. Secondly, we conduct an ablation study to explore how generalization changes with different amounts of data and find that we only need a small amount of low-resource language data to produce reasonably good BLEU scores. We use full data except for the ablation study. Finally, to address the variable-binding problem, we build a parallel lexicon table across twenty-three European languages and devise a novel method of order-preserving named entity translation method. Our method

[1]The material in this chapter was originally published as "Massively Parallel Cross-Lingual Learning in Low-Resource Target Language Translation" in WMT, 2018 [322].

Figure 3.2: An example of a hand-washing song that is translated into a few languages [281].

works in translation of any text with a fixed set of named entities known in advance. Our methods works best when we require high accuracy when there are many long sentences and out-of-vocabulary words, including many words that only occur once (hapax legomenon) [145, 192]. Our goal is to minimize manual labor, but not to fully automate to ensure the correct translation of named entities and their ordering.

3.2 Related Work

3.2.1 Sub-word Level Machine Translation

Many MT systems lack robustness with out-of-vocabulary words (*OOV*s) [312]. Most *OOV*s are treated as unknowns (*$UNK*s) uniformly, even though they are semantically important and different [178, 260]. To tackle the *OOV* problem, researchers work on byte-level [103] and character-level models [54, 178]. Many character-level models do not work as well as word-level models, and do not produce optimal alignments [285]. As a result, many researchers shift to sub-word level modeling between character-level and word-level. Many uses BPE which iteratively learns subword units and balances sequence length and expressiveness with robustness [260].

3.2.2 Lexiconized Machine Translation

Many work with lexicons and named entities in MT [13, 208, 307]. Some create a separate character-level named entity model and mark all named entities as *$TERM*s to train [307]. This method learns people's names well but does not improve BLEU scores [307]. It is time-consuming and adds to the system complexity. Others build lexicon translation seamlessly

Families	Languages
Germanic	German (de) Danish (dn) Dutch (dt) Norwegian (no) Swedish (sw) English (en)
Slavic	Croatian (cr) Czech (cz) Polish (po) Russian (ru) Ukrainian (ur) Bulgarian (bg)
Romance	Spanish (es) French (fr) Italian (it) Portuguese (po) Romanian (ro)
Albanian	Albanian (ab)
Hellenic	Greek (gk)
Italic	Latin (ln) [descendants: Romance languages]
Uralic	Finnish (fn) Hungarian (hg)
Celtic	Welsh (ws)

Table 3.1: Language families. Language codes are in parentheses.

with attentional MT by using an affine transformation of attentional weights [13, 208]. Some embed crosslingual lexicons into the same vector space for transfer [76].

3.3 Translation System

3.3.1 Baseline Translation System

Our baseline is multi-source multi-target attentional model within one language family through adding source and target language labels with a single unified attentional scheme, with BPE used at the preprocessing stage. The source and target vocabulary are not shared.

3.3.2 Proposed Extensions

We present our methods in solving three issues relevant to translation into low-resource language as our proposed extensions.

Language Families and Cross-lingual Learning

Cross-lingual and cross-cultural influences and similarities are important in linguistics [45, 63, 130, 172, 211, 250, 266, 271, 289]. The English word, "Beleaguer" originates from the Dutch word "belegeren"; "fidget" originates from the Nordic word "fikja". English and Dutch belong to the same family and their proximity has effect on each other [124, 246]. Furthermore, languages that do not belong to the same family affect each other too [10, 250, 289]. "Somatic" stems from the Greek word "soma"; "広告" (Japanese), "광고"(Korean), "Quảng cáo"(Vietnamese) are closely related to the Traditional Chinese word "廣告". Indeed, many cross-lingual similarities are present.

Language	German	Danish	Dutch	English	Norwegian	Swedish
German	N.A.	37.5	43.4	45.1	41.1	35.8
Danish	39.0	N.A.	37.1	41.1	42.6	37.4
Dutch	43.5	36.3	N.A.	45.1	39.0	34.3
English	40.4	34.5	41.1	N.A.	37.1	34.0
Norwegian	40.5	42.7	40.4	42.8	N.A.	40.6
Swedish	39.4	38.9	37.5	40.4	43.0	N.A.

Table 3.2: (Baseline model) Germanic family multi-source multi-target translation. Each row represents source, each column represents target.

In this work, we use the language phylogenetic tree as the measure of closeness of languages and language families [222]. The distance measure of language families is the collective of all of the component languages. Language families that are next to each other in the language phylogenetic tree are treated as neighboring families in our work, like Germanic family and Romance family. In our discussion, we will often refer to closely related families in the language phylogenetic tree as neighboring families.

We prepend the source and target family labels, in addition to the source and target language labels to the source sentence to increase translation performance. For example, all French-to-English translation pairs are prepended with four labels, the source and target family labels (`__opt_family_src_romance __opt_family_tgt_germanic`) and the source and target languages labels (`__opt_src_fr __opt_tgt_en`). In Section 3.4, we examine intra-family and inter-family effects more closely.

Ablation Study on Target Training data

To achieve high information transfer from rich-resource language to low-resource target language, we would like to find out how much target training data is needed to produce reasonably good performance. We vary the amount of low-resource training data to examine how to achieve reasonably good BLEU score using limited low-resource data. In the era of Statistical Machine Translation (SMT), researchers have worked on data sampling and sorting measures [16, 83].

To rigorously determine how much low-resource target language is needed for reasonably good results, we do a range of control experiments by drawing samples from the low-resource language data randomly with replacement and duplicate them if necessary to ensure all experiments carry the same number of training sentences. We keep the amount of training data in rich-resource languages the same, and vary the amount of training data in low-resource language to conduct rigorous control experiments. Our data selection process is different from prior research in that only the low-resource training data is reduced, simu-

lating the real world scenario of having little data in low-resource language. By comparing results from control experiments, we determine how much low-resource data is needed.

Order-preserving Lexiconized Model

The variable-binding problem is an inherent issue in connectionist architectures [95, 109]. "John calls Mary" is not equivalent to "Mary calls John", but neural networks cannot distinguish the two easily [95, 109]. The failure of traditional neural models to distinguish the subject and the object of a sentence is detrimental. For example, in the narration "John told his son Ryan to help David, the brother of Mary", it is a serious mistake if we reverse John and Ryan's father-son relationships or confuse Ryan's and David's relationships with Mary.

While many machine translation systems can successfully solve the Variable Binding Problem when there is abundant training data and there are a few named entities in a sentence, the Variable Binding Problem for severely low-resource scenarios remains very relevant even today due to the following reasons. Firstly, when there are many occurrences of hapax legomenon in a text, meaning when there are many words that only appear in the text once, it is almost impossible to translate an unseen named entity that never appears in the training data but appears in the test data. In the Bible, for instance, there are ∼1,480 words that only occur once. These words are very hard to translate. Additionally, if a text contains extremely long sentences with many rare named entities, it is hard for MT systems to translate all names correctly even if they appear in the training data. For example, the Bible dataset contain many sentences that are extremely long and has more than 10 named entities with an order that is very meaningful and cannot be mistranslated. Therefore, the Variable Binding Problem for severely low-resource languages is still relevant today. A successful solution to the Variable Binding Problem for low-resource languages is very useful when we aim for high accuracy when we translate long sentences with many unseen words, among which many only occur in the text once and never appear in the training data.

Variable Binding Problem is still relevant for hapax legomena, which are words that only appear in the text once. This means that a name may never appear in the training data, and only appear once in test data. the Bible has more than 2000 named entities that are hapaxes.

In our work, we focus mainly on text with a fixed set of named entities known in advance. We assume that experts help to translate a given list of named entities into low-resource language first before attempting to translate any text. Under this assumption, we propose an order-preserving named entity translation mechanism. Our solution is to first create a parallel lexicon table for all twenty-three European languages using a seed English lexicon table and fast-aligning it with the rest [78]. Instead of using *$UNK*s to replace the named

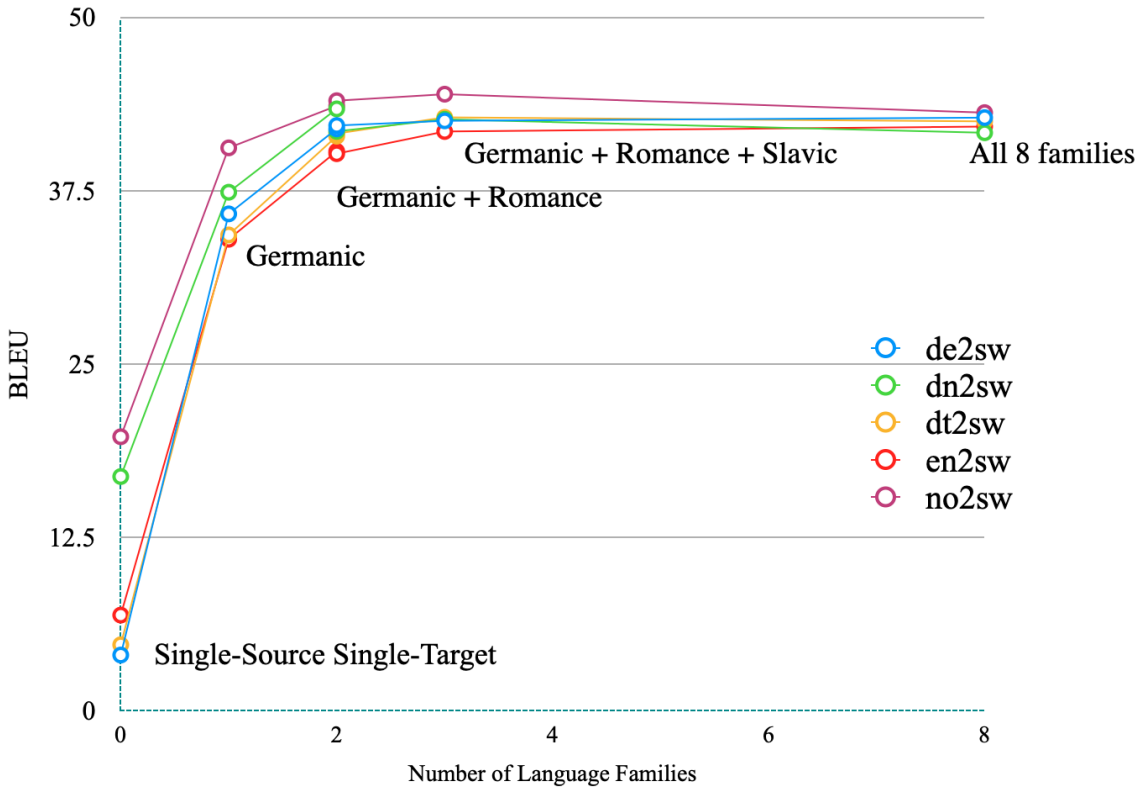


Figure 3.3: Intra-family and inter-family effects on BLEU scores with respect to increasing addition of language families.

entities, we use *$NE*s to distinguish them from the other unknowns. We also sequentially tag named entities in a sentence as *$NE1*, *$NE2*, . . . , to preserve their ordering. For every sentence pair in the multilingual training, we build a target named entity decoding dictionary by using all target lexicons from our lexicon table that matches with those appeared in the source sentence. During the evaluation stage, we replace all the numbered *$NE*s using the target named entity decoding dictionary to present our final translation. This method improves translation accuracy greatly and preserves the order.

As a result of our contribution, the experts only need to translate a few lexicons and a small amount of low-resource text before passing the task to our system to obtain good results. Post-editing and minor changes may be required to achieve 100% accuracy before the releasing the translation to the low-resource language communities.

3.4 Experiments

Having examined our baseline translation system and the proposed extensions, we are ready to discuss our experimental setups and premises.

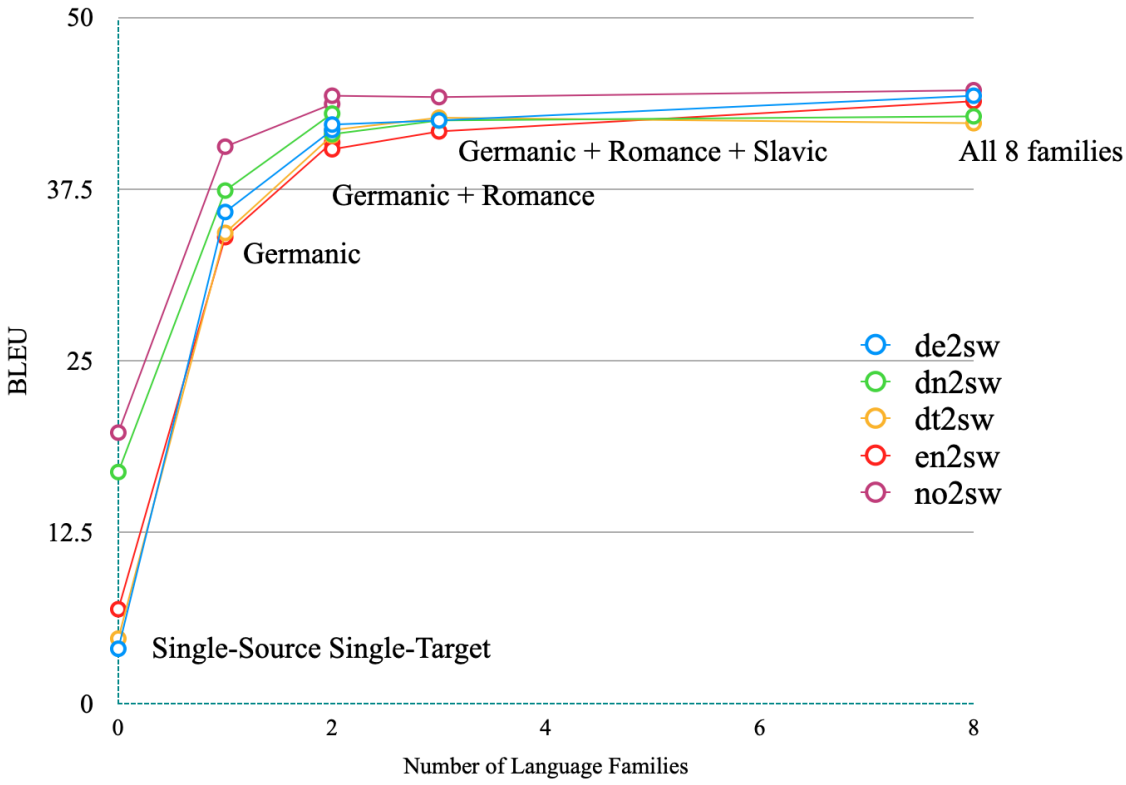


Figure 3.4: Effects of adding family labels on BLEU scores with respect to increasing addition of language families.

Our main goal is translating a multilingually known limited text into a new, low-resource language. There are three premises in our research that differ from traditional MT work:

1. Our text is closed, not arbitrary as in traditional MT problems.
2. Our text has multiple source languages with complete text translations while traditional MT is typically single-source.
3. Our text has little to no translation in the target low-resource language, while traditional MT assumes abundant data.

With these three premises, we build a multilingual model[2] from scratch[3], exclusively on limited massively multilingual data, in translation into a new, low-resource language. Training across all language pairs is done jointly to maximize interlingual transfer.

[2]In Chapter 5-9, all multilingual models are multilingual transformer-based. In Chapter 3-4, all models are multilingual LSTM-based model. This is because our research in Chapter 3-4 is done before transformer comes about.

[3]We train these models from scratch because we want to assume severely low-resource situation where there is no additional resource except for a little data from the low-resource language. Since all models are trained in one stage, there is no need to finetune. Results from this chapter serve as a baseline for Chapter 6-9 where we use multi-stage adaptation using large pretrained models.

Thus, we explore interlingual transfer strategies within and across language families in extremely low-training-data setups. Our ablation study aims to find out how to minimize the low-resource training data to achieve sufficiently good performance; our interlingual transfer studies the transfer within and across language families; and our ordered named entities translation mechanism aim to produce highly accurate named entity translations even when such words do not appear in the training data.

3.4.1 Data

We choose the Bible corpus as a test ground for our proposed extensions because the Bible is the most translated text that exists and is freely accessible. Though it has limitations, it has fewer copyright issues like most of literary works that are translated into many languages do. The Bible provides massively parallel high-quality translations into many severely low-resource languages and is rapidly growing in language coverage. There are many research works done using the Bible [18, 23, 46, 51, 74, 196, 203, 242, 252]. Unlike many past research works where only New Testament is used [74], we use both Old Testament and New Testament in our Bible corpus. We align all Bible verses across different languages.

We train our proposed model on twenty-three European languages across eight families on a parallel Bible corpus. For our purpose, we treat Swedish as our hypothetical low-resource target language, English as our rich-resource language in the single-source single-target case and all other Germanic languages as our rich-resource languages in the multi-source multi-target case.

Firstly, we present our data and training parameters. Secondly, we add family tags in different configurations of language families showing intra-family and inter-family effects. Thirdly, we conduct an ablation study and plot the generalization curves by varying the amount of training data in Swedish, and we show that training on one fifth of the data give reasonably good BLEU scores. Lastly, we devise an order-preserving lexicon translation method by building a parallel lexicon table across twenty-three European languages and tagging named entities in order.

3.4.2 Training Parameters

We clean and align the Bible in twenty-three European languages in Table 3.1. We randomly sample the training, validation and test sets according to the 0.75, 0.15, 0.10 ratio. Our training set contains 23K verses, but is massively parallel. In our control experiments, we also use the experiment training on the WMT'14 French-English dataset together with French and English Bibles to compare with our results. Note that our WMT baseline contains French and English Bibles in addition to the WMT'14 data, and is used to contrast our results with the effect of increasing data.

Experiment	S	G	GS	GR	3F	8F
de2sw	4.0	35.8	42.0	42.2	42.5	42.8
dn2sw	16.9	37.4	43.4	41.8	42.7	41.7
dt2sw	4.8	34.3	41.4	41.6	42.8	42.5
en2sw	6.9	34.0	40.3	40.2	41.8	42.1
no2sw	16.8	40.6	43.6	44.0	44.5	43.1

Table 3.3: Inter-family and intra-family effects on BLEU scores with respect to increasing addition of language families.
S: single-source single-target model.
G: training on Germanic family.
GS: training on Germanic, Slavic family.
GR: training on Germanic, Romance family.
3F: training on Germanic, Slavic, Romance family.
8F: training on all 8 European families together.

In all our experiments, we use a minibatch size of 64, dropout rate of 0.3, 4 RNN layers of size 1000, a word vector size of 600, learning rate of 0.8 across all LSTM-based multilingual experiments. For single-source single-target translation, we use 2 RNN layers of size 500, a word vector size of 500, and learning rate of 1.0. All learning rates are decaying at the rate of 0.7 if the validation score is not improving or it is past epoch 9. We use SGD as our learning algorithm. We build our code based on OpenNMT [157]. For the ablation study, we train on BLEU scores directly until the *Generalization Loss* (*GL*) exceeds a threshold of $\alpha = 0.1$ [233]. *GL* at epoch t is defined as $GL(t) = 100(1 - \frac{E^t_{val}}{E^t_{opt}})$, modified by us to suit our objective using BLEU scores [233]. E^t_{val} is the validation score at epoch t and E^t_{opt} is the optimal score up to epoch t. We evaluate our models using both BLEU scores [217] and qualitative evaluation.

3.5 Results

Having understood our data, training parameters and experimental setup, we investigate our results from three perspectives: interlingual transfer within and across families, ablation study and order-preserving lexiconized translation.

3.5.1 Interlingual Transfer Within and Across Families

We first investigate intra-family and inter-family influences and the effects of adding family labels. We use full training data in this subsection.

Experiment	S	G	GSl	GRl	3Fl	8Fl
de2sw	4.0	35.8	41.8	42.2	42.5	44.3
dn2sw	16.9	37.4	43.0	41.5	42.5	42.8
dt2sw	4.8	34.3	41.4	41.8	42.7	42.3
en2sw	6.9	34.0	40.9	40.4	41.7	43.9
no2sw	16.8	40.6	43.7	44.3	44.2	44.7

Table 3.4: Effects of adding family labels on BLEU scores with respect to increasing addition of language families.
S and G: same as in Table 3.3.
GSl: Germanic, Slavic family with family labels.
GRl: Germanic, Romance family with family labels.
3Fl: Germanic, Slavic, Romance family with family labels.
8Fl: all 8 European families together with family labels

Languages have varying closeness to each other: Single-source single-target translations of different languages in Germanic family to Swedish show huge differences in BLEU scores as shown in Table 3.3. These differences are well aligned with the multi-source multi-target results. Norwegian-Swedish and Danish-Swedish translations have much higher BLEU scores than the rest. This hints that Norwegian and Danish are closer to Swedish than the rest in the neural representation.

Multi-source multi-target translation improves greatly from single-source single-target translation: English-Swedish single-source single-target translation gives a low BLEU score of 6.9 as shown in Table 3.3, which is understandable as our dataset is very small. BLEU score for English-Swedish translation improves greatly to 34.0 in the multi-source multi-target model training on Germanic family as shown in Table 3.2. In this work, we treat the Germanic multi-source multi-target model as our baseline model. We present only relevant columns important for cross-lingual learning and translation into low-resource language here.

Adding languages from other families into training improves translation quality within each family greatly: English-Swedish translation's BLEU score improves significantly from 34.0 to 40.3 training on Germanic and Slavic families, and 40.2 training on Germanic and Romance families as shown in Table 3.3. After we add all three families in training, BLEU score for English-Swedish translation increases further to 41.8 in Table 3.3. Finally, after we add all eight families, BLEU score for English-Swedish translation increases to 42.1 in Table 3.3.

A Plateau is observed after adding more than one neighboring family: A plateau is observed when we plot Table 3.3 in Figure 3.3. The increase in BLEU scores

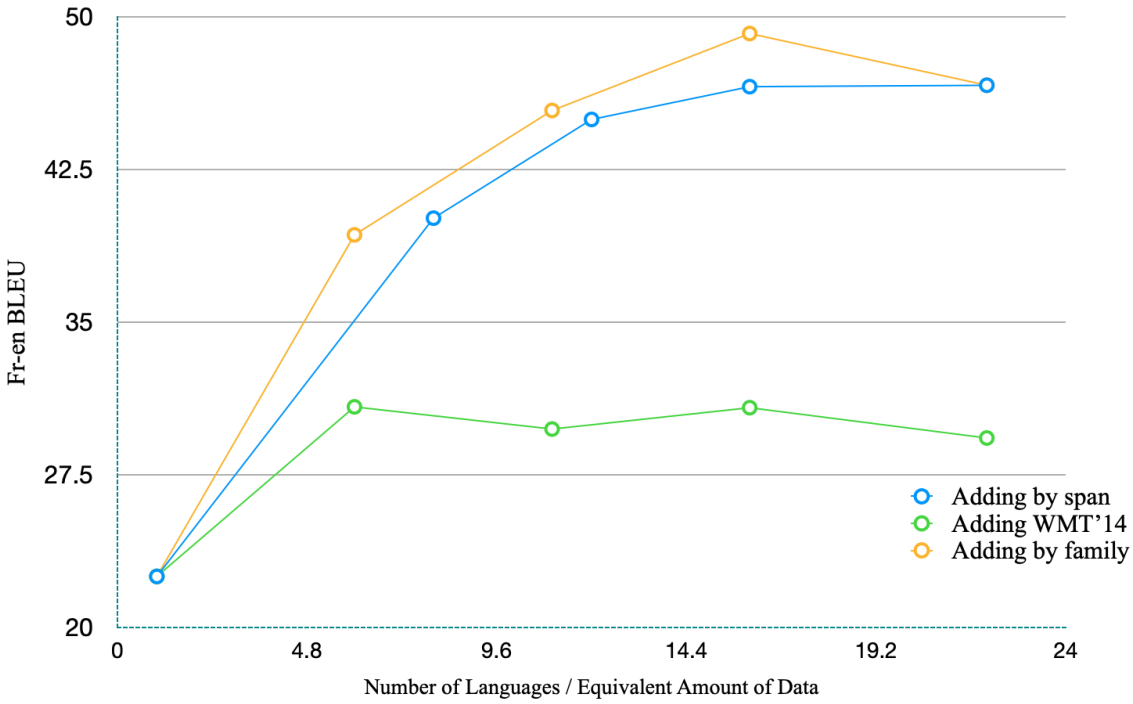


Figure 3.5: Comparison of different ways of increasing training Data in French-English translation.
Family: Adding data from other languages based on the family unit
WMT'14: Adding WMT'14 data as control experiment
Sparse: Adding data from other languages that spans the eight European families

after adding two families is much milder than that of the first addition of a neighboring family. This hints that using unlimited number of languages to train may not be necessary.

Adding family labels increases translation performance: We observe in Table 3.4 that BLEU scores for most language pairs improve with the addition of family labels. Training on eight language families, we achieve a BLEU score of 43.9 for English-Swedish translation, +9.9 above the Germanic baseline. Indeed, the more families we have, the more helpful it is to distinguish them.

Training on two neighboring families nearest to the low-resource language gives better result than training on languages that are further apart: Our observation of the plateau hints that training on two neighboring families nearest to the low-resource language is good enough as shown in Table 3.3. Before jumping to conclusion, we compare results of adding languages by family with that of adding languages by random samples that span all eight families, defined as the following.

Definition 3.5.1 (Language Spanning)**.** A set of languages spans a set of families when it contains at least one language from each family.

In Figure 3.5, we conduct a few experiments on French-English translation using different ways of adding training data. Let *family addition* describe the addition of training data

Data	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
#w	53589	107262	161332	214185	268228	322116	375439	429470	483440	538030
log#w	4.73	5.03	5.21	5.33	5.43	5.51	5.57	5.63	5.68	5.73
en2sw	25.2	30.6	32.9	32.7	34.2	34.2	33.8	33.6	34.3	34.9
de2sw	26.5	33.4	34.8	35.7	36.7	36.5	37.1	37.1	36.4	37.5
dn2sw	27.2	34.8	35.8	37.1	37.6	37.1	38.5	38.0	37.4	38.4
dt2sw	26.1	32.5	34.2	34.9	36.0	35.8	36.0	35.7	35.8	36.6
no2sw	27.7	36.9	37.9	39.5	39.4	39.2	41.3	40.8	39.2	40.5

Table 3.5: Ablation Study on Germanic Family. #w is the word count of unique sentences in Swedish data.

through adding close-by language families based on the unit of family; let *sparse addition* describe the addition of training data through adding language sets that spans eight language families. In sparse addition, languages are further apart as each may represent a different family. We find that family addition gives better generalization than that of sparse addition. It strengthens our earlier results that training on two families closest to our low-resource language is a reliable way to reach good generalization.

Generalization is not merely an effect of increasing amount of data: In Figure 3.5, we compare all methods of adding languages against a WMT'14 curve by using equivalent amount of WMT'14 French-English data in each experiment. The WMT'14 curve serve as our benchmark of observing the effect of increasing data, we observe that our addition of other languages improve BLEU score much sharply than the increase in the benchmark, showing that our generalization is not merely an effect of increasing data. We also observe that though increase WMT'14 data initially increases BLEU score, it reaches a plateau and adding more WMT'14 data does not increase performance from very early point.

3.5.2 Ablation Study on Target Training Data

We use full training data from all rich-resource languages, and we vary the amount of training data in Swedish, our low-resource language, spanning from one tenth to full length uniformly. We duplicate the subset to ensure all training sets, though having a different number of unique sentences, have the same number of total sentences.

Power-law relationship is observed between the performance and the amount of training data in low-resource language: Figure 3.7 shows how BLEU scores vary logarithmically with the number of unique sentences in the low-resource training data. It follows a linear pattern for single-source single-target translation from English to Swedish as shown in Figure 3.6. We also observe a linear pattern for the multi-source multi-target case, though more uneven in Figure 3.7. The linear pattern with BLEU scores against the

English	German	Czech	Spanish	Finnish	Swedish
Joseph	Joseph	Jozef	José	Joseph	Josef
Peter	Petrus	Petr	Pedro	Pietari	Petrus
Zion	Zion	Sion	Sion	Zionin	Sion
John	Johannes	Jan	Juan	Johannes	Johannes
Egypt	Ägypten	Egyptské	Egipto	Egyptin	Egyptens
Noah	Noah	Noé	Noé	Noa	Noa

Table 3.6: A few examples from the parallel lexicon table.

logarithmic data shows the power-law relationship between the performance in translation and the amount of low-resource training data. Similar power-law relationships are also found in past research and contemporary literature [131, 291].

We find one fifth of data is enough for sufficiently good translation performance: For the multi-source multi-target case, we find that using one fifth of the low-resource training data is sufficient for good translation performance. It gives reasonably good BLEU scores as shown in Figure 3.7. This is helpful when we have little low-resource data. For translation into the low-resource language, experts only need to translate a small amount of seed data before passing it to our system[4].

3.5.3 Order-preserving Lexiconized Model

We devise a mechanism to build a parallel lexicon table across twenty-three European languages using very little data and zero manual work. A few lexicon examples are shown in Table 3.6. We first extract named entities from the English Bible [190] and combine them with English biblically named entities from multiple sources [80, 134, 205, 243, 269]. Secondly, we carefully automate the filtering process to obtain a clean English lexicon list. Using this list as the seed, we build a parallel lexicon table across all twenty-three languages through fast-aligning [78]. The final parallel lexicon table has 2916 named entities. In the translation task into low-resource language, we assume that the experts first translate these lexicon entries, and then translate approximately one fifth random sentences before we train our model. If necessary, the experts evaluate and correct translations before releasing the final translations to the low-resource language community. We aim to reduce human effort in post-editing and increase machine accuracy. After labeling named entities in each sentence pair in order, we train and obtain good translation results.

We observe 60.6% accuracy in human evaluation where our translations are parallel to human translations: In Table 3.8, we show some examples of machine

[4]Note that using nine tenth of random samples yields higher performance than using full data, but it may not be generalized to other datasets.

Experiment	G	OG	OG1	OGM
de2sw	35.8	36.6	36.6	36.9
dn2sw	37.4	37.0	37.2	36.9
dt2sw	34.3	35.8	35.6	35.9
en2sw	34.0	33.6	33.9	33.4
no2sw	40.6	41.2	41.0	41.4

Table 3.7: Summary of order-preserving lexicon translation.
G: training on Germanic family without using order-preserving method.
OG: order-preserving lexicon translation.
OG1: OG translation using lexicons with frequency 1.
OGM: OG translation using lexicons with manual selection.

translated text, we also show the expected correct translations for comparison. Not only the named entities are correctly mapped, but also the ordering of the subject and the object is preserved. In a subset of our test set, we conduct human evaluation on 320 English-Swedish results to rate the translations into three categories: accurate (parallel to human translation), almost accurate (needing minor corrections) and inaccurate. More precisely, each sentence is evaluated using three criteria: correct set of named entities, correct positioning of named entities, and accurate meaning of overall translation. If a sentence achieves all three, then it is termed as accurate; if either a name entity is missing or its position is wrong, then it is termed as almost accurate (needing minor correction); if the meaning of the sentence is entirely wrong, then it is inaccurate. Our results are 60.6% accurate, 33.8% needing minor corrections, and 5.6% inaccurate. Though human evaluation carries bias and the sample is small, it does give us perspective on the performance of our model.

Order-preservation performs well especially when the named entities are rare words: In Table 3.8, the model without order-preservation lexiconized treatment performs well when named entities are common words, but fails to predict the correct set of named entities and their ordering when named entities are rare words. The last column shows the number of occurrences of each named entity. For the last example, there are many named entities that only occur in data once, which means that they never appear in training and only appear in the test set. The model without order-preservation lexiconized treatment predicts the wrong set of named entities with the wrong ordering. Our lexiconized order-preserving MT model, on the contrary, performs well at both the head and tail of the distribution, predicts the right set of named entities with the right ordering.

Prediction with longer sentences and many named entities are handled well: In Table 3.8, we see that the model without order-preservation lexiconized treatment performs well with short sentences and few named entities in a sentence. But as the number of the name entities per sentence increases, especially when the name entities are rare unknowns

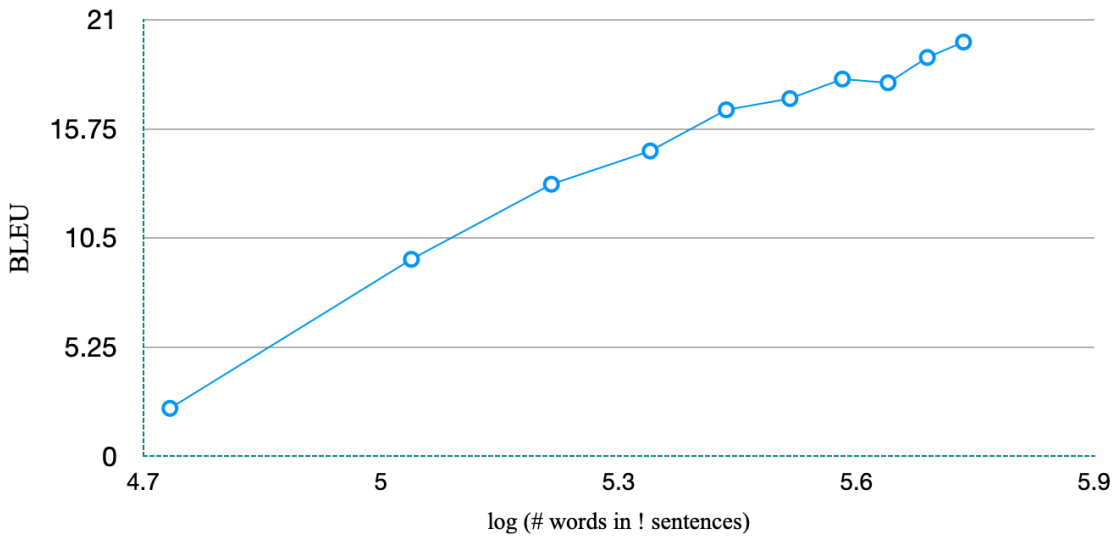


Figure 3.6: Single-source single-target English-Swedish BLEU plots against increasing amount of Swedish data.

as discussed before, normal model cannot make correct prediction of the right set of name entities with the correct ordering in Table 3.8. Our lexiconized order-preserving model, on the contrary, gives very high accuracy when there are many named entities in the sentence and maintains their correct ordering.

Trimming the lexicon list that keeps the tail helps to increase BLEU scores: Different from most of the previous lexiconized model works where BLEU scores never increase [307], our BLEU scores show minor improvements. BLEU score for German-Swedish translation increases from 35.8 to 36.6 in Table 3.7. As an attempt to increase our BLEU scores even further, we conduct two more experiments. In one setting, we keep only the tail of the lexicon table that occur in the Bible once. In another setting, we keep only a manual selection of lexicons. Note that this is the only place where manual work is involved and is not essential. There are minor improvements in BLEU scores in both cases.

33.8% of the translations require minor corrections: The sentence length for these translations that require minor corrections is often longer. We notice that some have repetitions that do not affect meaning, but need to be trimmed. Some have the under-prediction problem where certain named entities in the source sentence never appear; in this case, missing named entities need to be added. Some have minor issues with plurality and tense. Typically, sentences with longer sentence length and more complicated named entity relationships require minor corrections to achieve high translation quality.

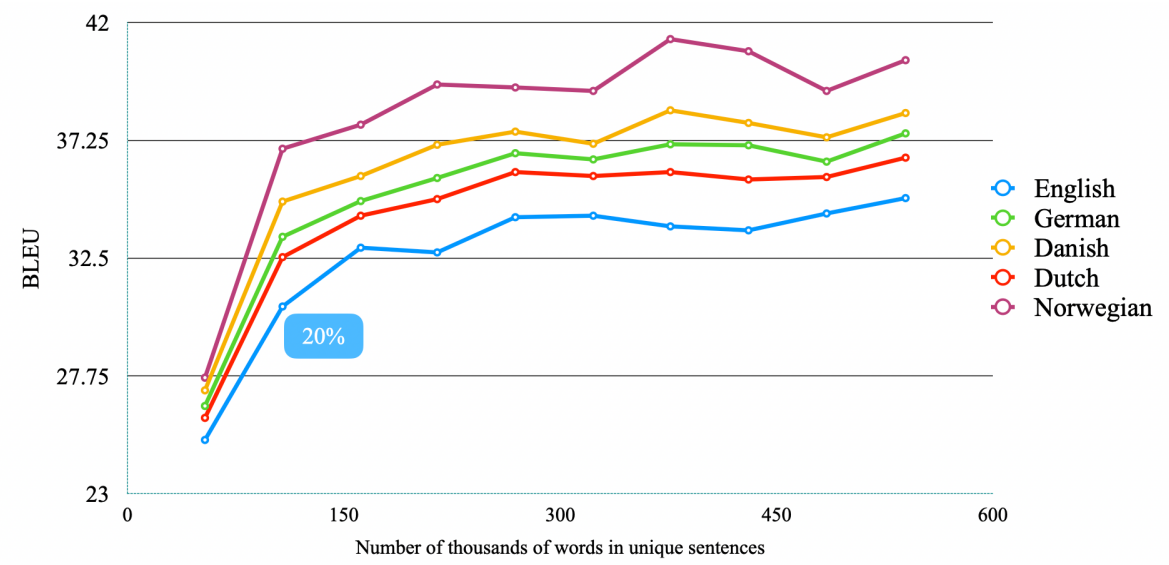


Figure 3.7: Multi-source multi-target Germanic-family-trained BLEU plots against increasing amount of Swedish data.

3.6 Conclusion and Future Directions

We present our order-preserving translation system for cross-lingual learning in European languages. We examine three issues that are important to translation into low-resource language: the lack of low-resource data, effective cross-lingual transfer, and the variable-binding problem.

Firstly, we add the source and the target family labels in training and examined intra-family and inter-family effects. We find that training on multiple families, more specifically, training on two neighboring families nearest to the low-resource language improves BLEU scores to a reasonably good level. Secondly, we devise a rigorous ablation study and show that we only need one fifth of the low-resource target data to produce sufficiently good translation performance. Thirdly, to address the variable-binding problem, we build a parallel lexicon table across twenty-three European languages and design a novel order-preserving named entity translation method by tagging named entities in each sentence in order. We achieve reasonably good quantitative and qualitative improvements in a preliminary study.

The order-preserving named entity translation labels named entities in order. Since there are relatively less number of long sentences with many named entities than short sentences with few named entities, underprediction of named entities in long sentences may occur. To seek solution to the underprediction problem, we are looking at randomized labeling of the named entities. Moreover, our order-preserving named entity translation method works well with a fixed pool of named entities in any static document known in advance.

This is due to our unique use cases for applications like translating water, sanitation and hygiene (WASH) guidelines written in the introduction. We devise our method to ensure high accuracy targeting translating named entities in static document known in advance. However, researchers may need to translate dynamic document to low-resource language in real-time. We are actively researching into the dynamic timely named entity discovery with high accuracy.

We are actively extending our work to cover more world languages, more diverse domains, and more varied sets of datasets to show our methods are generalizable. Since our experiments shown in this work are using European languages, we are also interested on non-European languages like Arabic, Indian, Chinese, Indonesian and many others to show that our model is widely generalizable. We also expect to discover interesting research ideas exploring a wider universe of linguistically dissimilar languages.

Our work is helpful for translation into low-resource language, where human translators only need to translate a few lexicons and a partial set of data before passing it to our system. Human translators may also be needed during post-editing before a fully accurate translation is released. Our future goal is to minimize the human correction efforts and to present high quality translation timely.

We would also like to work on real world low-resource tribal languages where there is no or little training data. Translation using limited resources and data in these tribal groups that fits with the culture-specific rules will be very important [172]. Real world low-resource languages call for cultural-aware translation. Careful research with real world low-resource languages will create great value and long impact for the low-resource language communities.

One situation when we experience real-world limits is that we may have little information of nearby languages, not to say translation of the multi-source text that is available in nearby languages. In those situations, we still aim to make the best of all the resources that we can find about the low-resource languages. In real-world situations, sometimes we do have multiple translators working on the same text, producing different paraphrases of the text in the same language, rather than in other languages. In the next chapter, we will explore ways to leverage paraphrases within the same language for translation of a multi-source, closed text into a new, low-resource language.

Source Sentence	Model Translation without Order Preservation (Before)	Model Translation with Order Preservation (After)	Reference	Frequency of Named Entities
And *Noah* fathered three sons, *Shem*, *Ham*, and *Japheth*.	Och *Noa* födde tre söner, *Sem*, *Ham* och *Jafet*.	Och *Noa* födde tre söner, *Sem* *Ham* och *Jafet*	Och *Noa* födde tre söner: *Sem*, *Ham* och *Jafet*.	*Noah*: 58, *Shem*: 18, *Ham*: 17, *Japheth*: 11
And *Saul* spoke to his son *Jonathan*, and to all his servants, to kill *David*.	Och *Saul* sade till *Jonatan*, hans son, och alla hans tjänare, så att de skulle döda *David*.	Och *Saul* talade till sin son *Jonatan* och alla hans tjänare för att döda *David*	Och *Saul* talade med sin son *Jonatan* och med alla sina tjänare om att döda *David*	*Saul*: 424, *Jonathan*: 121, *David*: 1134
And they killed *Parshandatha*, and *Dalphon*, and *Aspatha*, and *Poratha*, and *Adalia*, and *Aridatha*, and *Parmashta*, and *Arisai*, and *Aridai*, and *Vajezatha*,	Och de dräpte *Kedak*, *Ir-Fittim*, *Aquila*, *dörrvaktarna*, *Amarja*, *Bered*, *vidare Bet-Hadt*, *Berota*, *Gat-Rimmon*,	Och de dräpte *Parsandata* *Dalefon* och *Aspata* *Porata* *Adalja* *Aridata* *Parmasta* *Arisai* *Aridai* *Vajsata*	Och *Parsandata*, *Dalefon*, *Aspata*, *Porata*, *Adalja*, *Aridata*, *Parmasta*, *Arisai*, *Aridai* och *Vajsata*,	*Parshandatha*: 1, *Dalphon*: 1, *Aspatha*: 1, *Poratha*: 1, *Adalia*: 1, *Aridatha*: 1, *Parmashta*: 1, *Arisai*: 1, *Aridai*: 1, *Vajezatha*: 1

Table 3.8: Examples of order-preserving lexicon-aware translation for English to Swedish. The frequency of the named entities are the number of occurrences each named entity appears in the whole dataset; for example, all named entities in the last sentence do not appear in the training data.

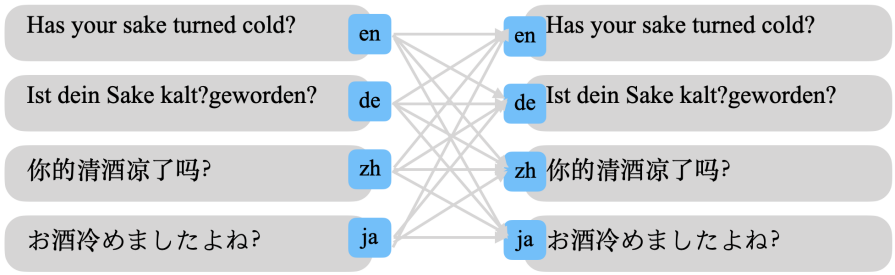
Chapter 4

Paraphrases as Foreign Languages

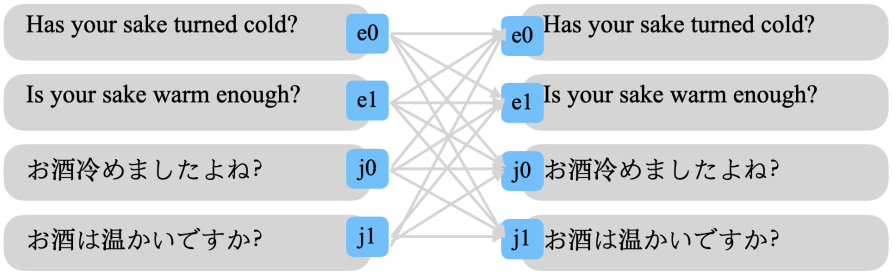
> "Learning another language is not only learning different words for the same things, but learning another way to think about things."
>
> *Flora Lewis*

Having demonstrated successful language transfer within and across families in the translation of a multi-source text into a new, low-resource languages, we would like to make full use of all the information that we can find in the low-resource language.

However, we face many translation cases where there is little information or incomplete information of nearby languages, not to mention translation of the given text in those nearby languages. In those situations, we might have translators translating the same text into the low-resource languages in different translating and writing styles due to quality control reasons. Therefore, paraphrases within the same language could be a by-product of the human translation process, and is another resource we could use for translation into low-resource languages. In other words, we are interested in researching into situations where there is more information in paraphrases within the same language, rather than across languages. We would like to use this as a new angle to make full use of all existing information that we can find in the target low-resource language. By using all information of paraphrases within the same language, we contribute to the main thesis of translation into low-resource languages by maximizing our use of the limited resources. In this chapter, we



(a) Multilingual model



(b) Multi-paraphrase model

Figure 4.1: Translation Paths in (a) multilingual model (b) multi-paraphrase model. Both form almost a complete bipartite graph.

seek to leverage paraphrases for translation of a multi-source text into a new, low-resource language.

4.1 Introduction

Paraphrases, rewordings of texts with preserved semantics, are often used to improve generalization and the sparsity issue in translation [42, 90, 101, 204, 255]. Unlike previous works that use paraphrases at the word/phrase level, we research on different translations of the whole corpus that are consistent in structure, content, coherence and style as paraphrases at the corpus level; we refer to paraphrases as the different translation versions of the same corpus. We train paraphrases in the style of multilingual translation systems [6, 69, 94, 103, 117, 147, 290, 327]. Implicit parameter sharing enables multilingual translation systems to learn across languages and achieve better generalization [147]. Training on closely related languages are shown to improve translation [322]. We view paraphrases as an extreme case of closely related languages and view multilingual data as paraphrases in different languages. Paraphrases can differ randomly or systematically as each carries the translator's unique style.

Source Sentence	Translation 1	Translation 2	Translation 3
My little horse must think it queer, to stop without a farmhouse near, between the woods and frozen lake, the darkest evening of the year. "Stopping by woods on a snowy evening", *Robert Frost.*	我的小马该满腹狐疑：附近没农舍，为什么歇息, 在树林和冰湖之间的地方，在这一年中最暗的昏里. Translated by *An tu.*	我的小马一定颇惊讶：四望不见有什么农家，偏是一年最暗的黄昏，寒林和冰湖之间停下. Translated by *Guangzhong Yu.*	我的小马肯定觉得奇怪,附近没有房子却停下来,在林子和结冰的湖水间,一年中最为黑暗的夜晚. Translated by *Tiejun Yang.*
Two roads diverged in a wood, and I– I took the one less traveled by, and that has made all the difference. "The road not taken", *Robert Frost*	林子里有两条路，朝着两个方向，而我– 我走上一条更少人迹的路，于是带来完全不同的一番景象. Translated by *Ping Fang.*	一片树林里分出两条路，而我选了人迹更少的一条，从此决定了我一生的道路. Translated by *Zixing Gu.*	林中两路, 岔道傍徨,我选一路, 人迹稀疏, 人生道路, 全不一样. Translated by *Ming Lei.*

Table 4.1: Examples of parallel paraphrasing data in English-Chinese poetry translation.

We treat paraphrases as foreign languages, and train a unified model[1] on paraphrase-labeled data with a shared attention in the style of multilingual translation systems. Similar to multilingual translation systems' objective of translating from any of the N input languages to any of the M output languages [94], multi-paraphrase translation systems aim to translate from any of the N input paraphrases to any of the M output paraphrases in Figure 4.1. In Figure 4.1, we see different expressions of a host showing courtesy to a guest to ask whether sake (a type of alcohol drink that is normally served warm in Asia) needs to be warmed. In Table 4.8, we show a few examples of parallel paraphrasing data in the Bible corpus. Different translators' styles give rise to rich parallel paraphrasing data, covering wide range of domains. In Table 4.7, we also show some paraphrasing examples from the modern poetry dataset, which we are considering for future research.

Indeed, we go beyond the traditional MT learning of one-to-one mapping between the source and the target text; instead, we exploit the many-to-many mappings between the source and target text through training on paraphrases that are consistent to each other at the corpus level. Our method achieves high translation performance and gives interesting findings. The differences between our work and the prior works are mainly the following.

Unlike previous works that use paraphrases at the word or phrase level, we use paraphrases at the entire corpus level to improve translation performance. We use different translations

[1]The material in this chapter was originally published in SRW at ACL, 2019 [323].

of the whole training data consistent in structure as paraphrases of the full training data. Unlike most of the multilingual MT works that uses data from multiple languages, we use paraphrases as foreign languages in a single-source single-target MT system training only on data from the source and the target languages.

Our work is meaningful because in translation into severely low-resource languages, we may have little information/data or incomplete information/data of nearby languages, not to mention translation of the text in those nearby languages. In those situations, for quality control reasons, we might have translators translating the same text into the low-resource languages in different translating and writing styles. This process provides us with paraphrases within the same language as a by-product of the human translation process. Indeed, this is another resource we could use for translation into low-resource languages. In other words, we are interested in researching into situations where there is more information in paraphrases within the same language, rather than across languages. We would like to use this as a new angle to make full use of all existing information that we can find in the target low-resource language. By using all information of paraphrases within the same language, we contribute to the main thesis of translation into low-resource languages by maximizing our use of the limited resources. When information is scarce, paraphrases within a given language becomes a powerful resource for increasing lexical diversity, improving sparsity and improving translation performance. Therefore, it is meaningful to leverage paraphrases for translation of a multi-source text into a new, low-resource language.

Additionally, our work is also meaningful in contributing findings in harnessing paraphrases in MT as the following:

1. Our multi-paraphrase MT[2] results show significant improvements in BLEU scores over all baselines.
2. Our paraphrase-exploiting MT[3] uses only two languages, the source and the target languages, and achieves higher BLEUs than the multi-source and multi-target MT that incorporates more languages.
3. We find that adding the source paraphrases helps better than adding the target paraphrases.
4. We find that adding paraphrases at both the source and the target sides is better than adding at either side.

[2]In Chapter 5-9, all multilingual models are multilingual transformer-based. In Chapter 3-4, all models are multilingual LSTM-based model. This is because our research in Chapter 3-4 is done before transformer comes about.

[3]We train these models from scratch because we want to assume severely low-resource situation where there is no additional resource except for a little data from the low-resource language. Though this chapter's main focus is on paraphrases, we make design decisions based on our goal in this thesis.

5. We also find that adding paraphrases with additional multilingual data yields mixed performance; its performance is better than training on language families alone, but is worse than training on both the source and target paraphrases without language families.
6. Adding paraphrases improves the sparsity issue of rare word translation and diversity in lexical choice.

4.2 Related Work

4.2.1 Paraphrasing

Many works generate and harness paraphrases [21, 39, 41, 101, 126, 186, 189, 216, 235, 276]. Some are on question and answer [70, 90], evaluation of translation [319] and translation [204, 255]. Past research includes paraphrasing unknown words/phrases/sub-sentences [42, 89, 204, 255]. These approaches are similar in transforming the difficult sparsity problem of rare words prediction and long sentence translation into a simpler problem with known words and short sentence translation. It is worthwhile to contrast paraphrasing that diversifies data, with knowledge distillation that benefits from making data more consistent [112].

Our work is different in that we exploit paraphrases at the corpus level, rather than at the word or phrase level.

4.2.2 Multilingual Attentional Translation Models

As discussed before, machine polyglotism which trains machines to translate any of the N input languages to any of the M output languages is a new paradigm in multilingual translation models [6, 69, 94, 103, 290, 327]. An interesting work is training a universal model with a shared attention mechanism with the source and target language labels and Byte-Pair Encoding (BPE) [117, 147]. Many multilingual translation systems involve multiple encoders and decoders [117], and it is hard to combine attention for quadratic language pairs bypassing quadratic attention mechanisms [94]. An interesting work is training a universal model with a shared attention mechanism with the source and target language labels and Byte-Pair Encoding (BPE) [117, 147]. This method is elegant in its simplicity and its advancement in low-resource language translation and zero-shot translation using pivot-based translation mechanism [94, 147].

Unlike previous works, our parallelism is across paraphrases, not across languages. In other words, we achieve higher translation performance in the single-source single-target paraphrase-exploiting MT systems than that of the multilingual MT systems.

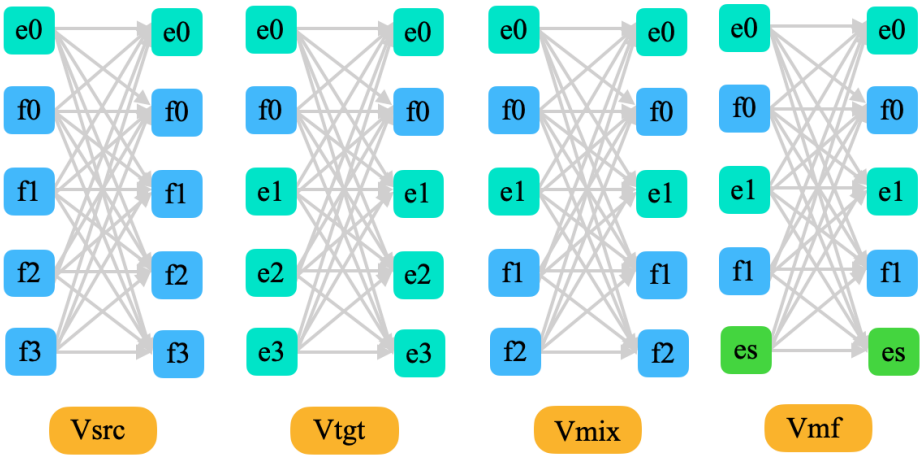


Figure 4.2: Examples of different ways of adding 5 paraphrases. `e[?n]` and `f[?n]` refers to different English and French paraphrases, `es` refers to the Spanish (an example member of Romance family) data. We always evaluate the translation path from `f0` to `e0`.

4.3 Models

We have four baseline models. Two are single-source single-target attentional models, the other two are multilingual MT models with a shared attention [117, 147]. In Figure 4.1, we show an example of multilingual attentional MT. Translating from all 4 languages to each other, we have 12 translation paths. For each translation path, we label the source sentence with the source and target language tags. Translating from "你的清酒凉了吗?" to "Has your sake turned cold?", we label the source sentence with `__opt_src_zh __opt_tgt_en`. More details are in Section 4.4.

In multi-paraphrase model, all source sentences are labeled with the paraphrase tags. For example, in French-to-English translation, a source sentence may be tagged with `__opt_src_f1 __opt_tgt_e0`, denoting that it is translating from version "f1" of French data to version "e0" of English data. In Figure 4.1, we show 2 Japanese and 2 English paraphrases. Translating from all 4 paraphrases to each other ($N = M = 4$), we have 12 translation paths as $N \times (N - 1) = 12$. For each translation path, we label the source sentence with the source and target paraphrase tags. For the translation path from "お酒冷めましたよね?" to "Has your sake turned cold?", we label the source sentence with `__opt_src_j1 __opt_tgt_e0` in Figure 4.1. Paraphrases of the same translation path carry the same labels. Our paraphrasing data is at the corpus level, and we train a unified MT model with a shared attention. Unlike the paraphrasing sentences in Figure 4.1, We show this example with only one sentence, it is similar when the training data contains many sentences. All sentences in the same paraphrase path share the same labels.

data	1	6	11	16	22	24
WMT	22.5	30.8	29.8	30.8	29.3	-
Family	22.5	39.3	45.4	49.2	46.6	-
Vmix	22.5	44.8	50.8	53.3	55.4	57.2
Vmf	-	-	49.3	-	-	-

Table 4.2: Comparison of adding a mix of the source paraphrases and the target paraphrases against the baselines. All acronyms including data are explained in Section 4.4.3.

Data	1	6	11	13
Vsrc	22.5	41.4	48.9	48.8
Vtgt	22.5	40.5	47.0	47.4

Table 4.3: Comparison of adding source paraphrases and adding target paraphrases. All acronyms including data are explained in Section 4.4.3.

4.4 Experiments

4.4.1 Data

Our main data is the French-to-English Bible corpus [196], containing 12 versions of the English Bible and 12 versions of the French Bible [4]. We translate from French to English. Since these 24 translation versions are consistent in structure, we refer to them as paraphrases at corpus level. In our work, each paraphrase refers to each translation version of whole Bible corpus. To understand our setup, if we use all 12 French paraphrases and all 12 English paraphrases so there are 24 paraphrases in total, i.e., $N = M = 24$, we have 552 translation paths because $N \times (N - 1) = 552$.

To prepare data for this experimental setup, we clean and align 24 paraphrases of the Bible corpus because the original corpus contains missing or extra verses for different paraphrases. After cleaning and aligning all 24 paraphrases of the Bible corpus, we randomly sample the training, validation and test sets according to the 0.75, 0.15, 0.10 ratio. Our training set contains only 23K verses, but is massively parallel across paraphrases.

For all experiments, we choose a specific English corpus as `e0` and a specific French corpus as `f0` which we evaluate across all experiments to ensure consistency in comparison, and we evaluate all translation performance from `f0` to `e0`. The reason we choose this particular

[4]We considered the open subtitles with different scripts of the same movie in the same language; they covers many topics, but they are noisy and only differ in interjections. We also considered the poetry dataset where a poem like "If" by Rudyard Kipling is translated many times, by various people into the same language, but the data is small.

data	6	11	16	22	24
Entropy	5.6569	5.6973	5.6980	5.7341	5.7130
Bootstrap 95% CI	(5.6564, 5.6574)	(5.6967, 5.6979)	(5.6975, 5.6986)	(5.7336, 5.7346)	(5.7125, 5.7135)
WMT	-	5.7412	5.5746	5.6351	-

Table 4.4: Entropy increases with the number of paraphrase corpora in *Vmix*. The 95% confidence interval is calculated via bootstrap resampling with replacement.

data	6	11	16	22	24
F1(freq1)	0.43	0.54	0.57	0.58	0.62
WMT	-	0.00	0.01	0.01	-

Table 4.5: F1 score of frequency 1 bucket increases with the number of paraphrase corpora in *Vmix*, showing training on paraphrases improves the sparsity at tail and the rare word problem.

language pair is because this language pair from `f0` to `e0` is the language pair that is part of the multilingual baseline that we are comparing our results with. We want to have the same language pair to ensure consistency when we compare all experiments against the baselines. This way of evaluation is necessary for our comparison with the baselines but it is also a limitation of our research which we will discuss at the end of this chapter.

4.4.2 Training Parameters

In all our experiments, we use a minibatch size of 64, dropout rate of 0.3, 4 RNN layers of size 1000, a word vector size of 600, number of epochs of 13, a learning rate of 0.8 that decays at the rate of 0.7 if the validation score is not improving or it is past epoch 9 across all LSTM-based experiments. Byte-Pair Encoding (BPE) is used at preprocessing stage [117]. Our code is built on OpenNMT [157] and we evaluate our models using BLEU scores [217], entropy [264], F-measure and qualitative evaluation.

4.4.3 Baselines

We introduce a few acronyms for our four baselines to describe the experiments in Table 4.3, Table 4.2 and Figure 4.3. Firstly, we have two single-source single-target attentional MT models, *Single* and *WMT*. *Single* trains on `f0` and `e0` and gives a BLEU of 22.5, the starting point for all curves in Figure 4.3. *WMT* adds the out-domain WMT'14 French-to-English data on top of `f0` and `e0`; it serves as a weak baseline that helps us to evaluate all experiments' performance discounting the effect of increasing data.

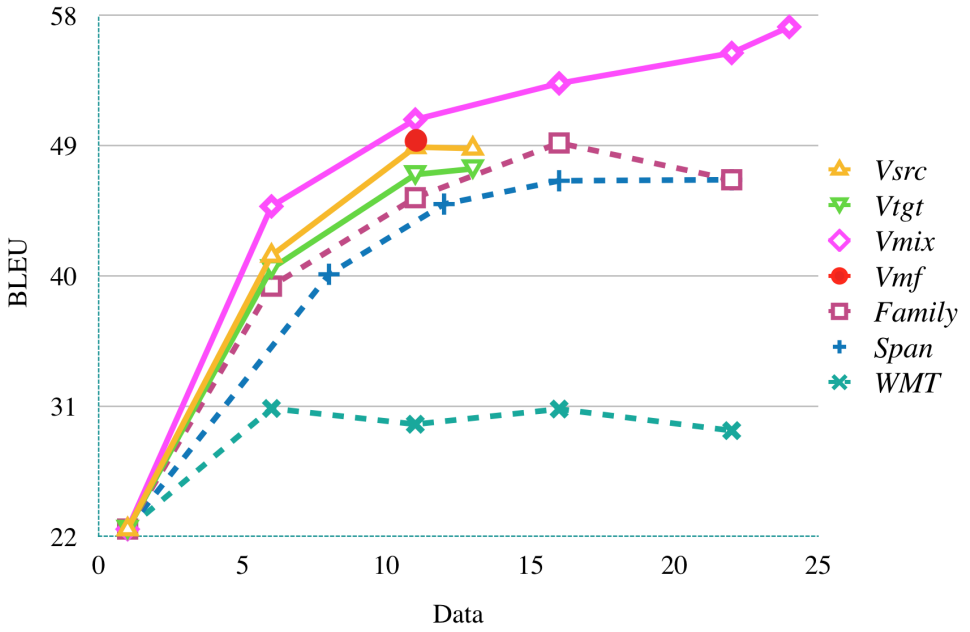


Figure 4.3: BLEU plots showing the effects of different ways of adding training data in French-to-English Translation. All acronyms including data are explained in Section 4.4.3.

Moreover, we have two multilingual baselines[5] built on multilingual attentional MT, ***Family*** and ***Span*** [322]. ***Family*** refers to the multilingual baseline by adding one language family at a time, where on top of the French corpus `f0` and the English corpus `e0`, we add up to 20 other European languages. ***Span*** refers to the multilingual baseline by adding one *span* at a time, where a span is a set of languages that contains at least one language from all the families in the data; in other words, span is a sparse representation of all the families. Both ***Family*** and ***Span*** trains on the Bible in 22 Europeans languages trained using multilingual MT. Since ***Span*** is always sub-optimal to ***Family*** in our results, we only show numerical results for ***Family*** in Table 4.3 and Table 4.2, and we plot both ***Family*** and ***Span*** in Figure 4.3. The two multilingual baselines are strong baselines while the ***WMT*** baseline is a weak baseline that helps us to evaluate all experiments' performance discounting the effect of increasing data. All baseline results are taken from a research work which uses the grid of (1, 6, 11, 16, 22) for the number of languages or equivalent number of unique sentences and we follow the same in Figure 4.3 [322]. All experiments for each grid point carry the same number of unique sentences.

Furthermore, ***Vsrc*** refers to adding more source (English) paraphrases, and ***Vtgt*** refers to adding more target (French) paraphrases. ***Vmix*** refers to adding both the source and the target paraphrases. ***Vmf*** refers to combining ***Vmix*** with additional multilingual data;

[5]For multilingual baselines, we use the additional Bible corpus in 22 European languages that are cleaned and aligned to each other.

note that only *Vmf*, *Family* and *Span* use languages other than French and English, all other experiments use only English and French. For the x-axis, data refers to the number of paraphrase corpora for *Vsrc*, *Vtgt*, *Vmix*; data refers to the number of languages for *Family*; data refers to and the equivalent number of unique training sentences compared to other training curves for *WMT* and *Vmf*.

4.5 Results

Training on paraphrases gives better performance than all baselines: The translation performance of training on 22 paraphrases, i.e., 11 English paraphrases and 11 French paraphrases, achieves a BLEU score of 55.4, which is +32.9 above the *Single* baseline, +8.8 above the *Family* baseline, and +26.1 above the *WMT* baseline. Note that the *Family* baseline uses the grid of (1, 6, 11, 16, 22) for number of languages, we continue to use this grid for our results on number of paraphrases, which explains why we pick 22 as an example here. The highest BLEU 57.2 is achieved when we train on 24 paraphrases, i.e., 12 English paraphrases and 12 French paraphrases.

Adding the source paraphrases boosts translation performance more than adding the target paraphrases: The translation performance of adding the source paraphrases is higher than that of adding the target paraphrases. Adding the source paraphrases diversifies the data, exposes the model to more rare words, and enables better generalization. Take the experiments training on 13 paraphrases for example, training on the source (i.e., 12 French paraphrases and the English paraphrase `e0`) gives a BLEU score of 48.8, which has a gain of +1.4 over 47.4, the BLEU score of training on the target (i.e., 12 English paraphrases and the French paraphrase `f0`). This suggests that adding the source paraphrases is more effective than adding the target paraphrases.

Adding paraphrases from both sides is better than adding paraphrases from either side: The curve of adding paraphrases from both the source and the target sides is higher than both the curve of adding the target paraphrases and the curve of adding the source paraphrases. Training on 11 paraphrases from both sides, i.e., a total of 22 paraphrases achieves a BLEU score of 50.8, which is +3.8 higher than that of training on the target side only and +1.9 higher than that of training on the source side only. The advantage of combining both sides is that we can combine paraphrases from both the source and the target to reach 24 paraphrases in total to achieve a BLEU score of 57.2.

Adding both paraphrases and language families yields mixed performance: We conduct one more experiment combining the source and target paraphrases together with additional multilingual data. This is the only experiment on paraphrases where we use multilingual data other than only French and English data. The BLEU score is 49.3, higher than training on families alone, in fact, it is higher than training on eight European

families altogether. However, it is lower than training on English and French paraphrases alone. Indeed, adding paraphrases as foreign languages is effective, however, when there is a lack of data, mixing the paraphrases with multilingual data is helpful.

Adding paraphrases increases entropy and diversity in lexical choice, and improves the sparsity issue of rare words: We use bootstrap resampling and construct 95% confidence intervals for entropies [264] of all models of *Vmix*, i.e., models adding paraphrases at both the source and the target sides. We find that the more paraphrases, the higher the entropy, the more diversity in lexical choice as shown in Table 4.4. From the word F-measure shown in Table 4.5, we find that the more paraphrases, the better the model handles the sparsity of rare words issue. Adding paraphrases not only achieves much higher BLEU score than the *WMT* baseline, but also handles the sparsity issue much better than the *WMT* baseline.

Adding paraphrases helps rhetoric translation and increases expressiveness: Qualitative evaluation shows many cases where rhetoric translation is improved by training on diverse sets of paraphrases. In Table 4.6, Paraphrases help the model to use a more contemporary synonym of "silver", "money", which is more direct and easier to understand. Paraphrases simplifies the rhetorical or subtle expressions, for example, our model uses "rejoice" to replace "break out into song", a personification device of mountains to describe joy, which captures the essence of the meaning being conveyed. However, we also observe that the model wrongly translates "clap the palm" to "strike". We find the quality of rhetorical translation ties closely with the diversity of parallel paraphrases data. Indeed, the use of paraphrases to improve rhetoric translation is a good future research question. Please refer to the Table 4.6 for more qualitative examples.

4.6 Conclusion

We train on paraphrases as foreign languages in the style of multilingual translation systems. Adding paraphrases improves translation quality, the rare word issue, and diversity in lexical choice. Adding the source paraphrases helps more than adding the target ones, while combining both boosts performance further. Adding multilingual data to paraphrases yields mixed performance.

Having understood our main contributions in this chapter, we need to understand a limitation of this research in our evaluation setup. Since we would like to measure how much improvement we have achieved above the multilingual baseline where there is a particular language pair (paraphrase pair) present, we can only choose this particular language pair (paraphrase pair) for evaluation across all experiments. Moreover, we want to explore using all of the existing paraphrases as alternative references in a multi-reference evaluation manner. However, we did not choose this setup, instead we chose to use a

particular language pair (paraphrase pair) that is part of the multilingual baseline to ensure consistency in comparison. This is a necessary design decision, but is also a limitation to our research. In the future, we are interested in multi-reference evaluation on top of the existing evaluation mechanism.

A natural question one may ask is that if we have known target paraphrases available, is there still a need to translate it in the first place? Our answer is yes. Each person's writing and translation style carries a unique signature, which is exactly what the machine learns. For example, if we have a text translated for adults, and we need to paraphrase for adolescents, the machine needs to learn the tone and style of adolescents. In this setting, our finding is very useful for training the translation of text with a specific style, and translation of text that is requires a specific level of formality. One may also ask, how does our method compare with monolingual adaptation in the target language? Why do we do multi-paraphrase training if we can just adapt on monolingual data? This is because the nature of the low-resource scenarios in our use case. We have little to no data in the target low-resource language. As we move to the following chapters, we will use as low as ∼600 sentences in the target language. The intention of this chapter is to contribute to the main thesis, which is to how to address machine translation and style adaptation under severely low-resource scenarios. Indeed, multi-paraphrase training provides better generalization and avoids overfitting than using monolingual adaptation on severely low-resource languages.

We would like to explore the common structure and terminology consistency across different paraphrases. Since structure and terminology are shared across paraphrases, we are interested in building an explicit representation of the paraphrases and extending our work for better translation, or translation with more explicit and more explainable hidden states, which is very important in all neural systems. All of these have immeasurable value for translation into low-resource languages. In low-resource situations, we might have translators translating the same text into the low-resource languages in different translating and writing styles. Paraphrases within the same language could be therefore a by-product of human translation process, and is another resource we could use for translation into low-resource languages. By using all information of paraphrases within the same language, we contribute to the main thesis of translation into low-resource languages by maximizing the use of our limited resources.

We are interested in broadening our dataset in our future experiments. In addition to the parallel paraphrasing data as shown in the Bible corpus in Table 4.8, we also show some paraphrasing examples from the modern poetry dataset in Table 4.7. There are very few poems that are translated multiple times into the same language, we therefore need to train on extremely small dataset. Rhetoric in paraphrasing is important in poetry dataset, which again depends on the training paraphrases. The limited data issue is also relevant to the low-resource setting.

We would like to effectively train on extremely small low-resource paraphrasing data. As discussed above about the potential research poetry dataset, dataset with multiple paraphrases is typically small and yet valuable. If we can train using extremely small amount of data, especially in the low-resource scenario, we would exploit the power of multi-paraphrase translation further.

Cultural-aware paraphrasing and subtle expressions are vital [169, 172]. Rhetoric in paraphrasing is a very important too. In Figure 4.1, "is your sake warm enough?" in Asian culture is an implicit way of saying "would you like me to warm the sake for you?". We would like to model the culture-specific subtlety through multi-paraphrase training.

In this chapter and the previous chapter, we have explored ways to leverage information in paraphrases in the same language and treat them as foreign languages, as well as nearby languages within the same or nearby language families. However, in real-world scenarios, information of which language is close may not be available, information of paraphrases within the same language might also not be available. In the next chapter, we focus on these information-scarce situations, and we research on how to make best use of the available multi-source text, and build multiple metric spaces of language closeness statistically to facilitate training.

Source Sentence	Machine Translation	Reference
Comme de l'eau fraîche pour une personne fatigué, Ainsi est une bonne nouvelle venant d'une terre lointaine.	As cold waters to a thirsty soul, so is good news from a distant land.	Like cold waters to a weary soul, so is a good report from a far country.
Lorsque tu seras invité par quelqu'un à des noces, ne te mets pas à la première place, de peur qu'il n'y ait parmi les invités une personne plus considérable que toi,	When you are invited to one to the wedding, do not be to the first place, lest any one be called greater than you.	When you are invited by anyone to wedding feasts, do not recline at the chief seat lest one more honorable than you be invited by him,
Car chaque arbre se connaît à son fruit. On ne cueille pas des figues sur des épines, et l'on ne vendange pas des raisins sur des ronces.	For each tree is known by its own fruit. For from thorns they do not gather figs, nor do they gather grapes from a bramble bush.	For each tree is known from its own fruit. For they do not gather figs from thorns, nor do they gather grapes from a bramble bush.
Nous la côtoyâmes avec peine, et nous arrivâmes à un lieu nommé Beaux Ports, près duquel était la ville de Lasée.	And coasting along it with difficulty, we came to a certain place called Fair Havens, near to which was the city of Lasea.	And coasting along it with difficulty, we came to a certain place named Fair Havens, near to which was a city , Lasea.
Vous tous qui avez soif, venez aux eaux, Même celui qui n'a pas d'argent! Venez, achetez et mangez, Venez, achetez du vin et du lait, sans argent, sans rien payer!	Come, all you thirsty ones, come to the waters; come, buy and eat. Come, buy for wine, and for nothing, for without money.	Ho, everyone who thirsts, come to the water; and he who has no silver, come buy grain and eat. Yes, come buy grain, wine and milk without silver and with no price.
Oui , vous sortirez avec joie , Et vous serez conduits en paix ; Les montagnes et les collines éclateront d'allégresse devant vous , Et tous les arbres de la campagne battront des mains .	When you go out with joy , you shall go in peace ; the mountains shall rejoice before you , and the trees of the field shall strike all the trees of the field .	For you shall go out with joy and be led out with peace . The mountains and the hills shall break out into song before you , and all the trees of the field shall clap the palm .

Table 4.6: Examples of French-to-English translation trained using 12 French paraphrases and 12 English paraphrases.

English	If you can fill the unforgiving minute with sixty seconds' worth of distance run, yours is the Earth and everything that's in it, and—which is more—you'll be a Man, my son! "if", *Rudyard Kipling.*
German	Wenn du in unverzeihlicher Minute Sechzig Minuten lang verzeihen kannst: Dein ist die Welt—und alles was darin ist— Und was noch mehr ist—dann bist du ein Mensch! Translation by *Anja Hauptmann.*
	Wenn du erfüllst die herzlose Minute Mit tiefstem Sinn, empfange deinen Lohn: Dein ist die Welt mit jedem Attribute, Und mehr noch: dann bist du ein Mensch, mein Sohn! Translation by *Izzy Cartwell.*
	Füllst jede unerbittliche Minute Mit sechzig sinnvollen Sekunden an; Dein ist die Erde dann mit allem Gute, Und was noch mehr, mein Sohn: Du bist ein Mann! Translation by *Lothar Sauer.*
Chinese	若胸有激雷, 而能面如平湖, 则山川丘壑, 天地万物皆与尔共, 吾儿终成人也！ Translation by *Anonymous.*
	如果你能惜时如金利用每一分钟不可追回的光阴；那么，你的修为就会如天地般博大，并拥有了属于自己的世界，更重要的是：孩子，你成为了真正顶天立地之人！ Translation by *Anonymous.*
	假如你能把每一分宝贵的光阴, 化作六十秒的奋斗—-你就拥有了整个世界, 最重要的是——你就成了一个真正的人，我的孩子！ Translation by *Shan Li.*
Portuguese	Se você puder preencher o valor do inclemente minuto perdido com os sessenta segundos ganhos numa longa corrida, sua será a Terra, junto com tudo que nela existe, e—mais importante—você será um Homem, meu filho! Translation by *Dascomb Barddal.*
	Pairando numa esfera acima deste plano, Sem receares jamais que os erros te retomem, Quando já nada houver em ti que seja humano, Alegra-te, meu filho, então serás um homem!... Translation by *Féliz Bermudes.*
	Se és capaz de dar, segundo por segundo, ao minuto fatal todo valor e brilho. Tua é a Terra com tudo o que existe no mundo, e—o que ainda é muito mais—és um Homem, meu filho! Translation by *Guilherme de Almeida.*

Table 4.7: Examples of parallel paraphrasing data with German, Chinese, and Portuguese paraphrases of the English poem "If" by Rudyard Kipling.

English	Consider the lilies, how they grow: they neither toil nor spin, yet I tell you, even Solomon in all his glory was not arrayed like one of these. *English Standard Version.*
	Look how the wild flowers grow! They don't work hard to make their clothes. But I tell you Solomon with all his wealth wasn't as well clothed as one of these flowers. *Contemporary English Version.*
	Consider how the wild flowers grow. They do not labor or spin. Yet I tell you, not even Solomon in all his splendor was dressed like one of these. *New International Version.*
French	Considérez les lis! Ils poussent sans se fatiguer à tisser des vêtements. Et pourtant, je vous l'assure, le roi Salomon lui-même, dans toute sa gloire, n'a jamais été aussi bien vêtu que l'un d'eux! *La Bible du Semeur.*
	Considérez comment croissent les lis: ils ne travaillent ni ne filent; cependant je vous dis que Salomon même, dans toute sa gloire, n'a pas été vêtu comme l'un d'eux. *Louis Segond.*
	Observez comment poussent les plus belles fleurs: elles ne travaillent pas et ne tissent pas; cependant je vous dis que Salomon lui-même, dans toute sa gloire, n'a pas eu d'aussi belles tenues que l'une d'elles. *Segond 21.*
Tagalog	Wariin ninyo ang mga lirio, kung paano silang nagsisilaki: hindi nangagpapagal, o nangagsusulid man; gayon ma'y sinasabi ko sa inyo, Kahit si Salomon man, sa buong kaluwalhatian niya, ay hindi nakapaggayak na gaya ng isa sa mga ito. *Ang Biblia 1978.*
	Isipin ninyo ang mga liryo kung papaano sila lumalaki. Hindi sila nagpapagal o nag-iikid. Gayunman, sinasabi ko sa inyo: Maging si Solomon, sa kaniyang buong kaluwalhatian ay hindi nadamitan ng tulad sa isa sa mga ito. *Ang Salita ng Diyos.*
	Tingnan ninyo ang mga bulaklak sa parang kung paano sila lumalago. Hindi sila nagtatrabaho ni humahabi man. Ngunit sinasabi ko sa inyo, kahit si Solomon sa kanyang karangyaan ay hindi nakapagdamit ng singganda ng isa sa mga bulaklak na ito. *Magandang Balita Biblia.*
Spanish	Considerad los lirios, cómo crecen; no trabajan ni hilan; pero os digo que ni Salomón en toda su gloria se vistió como uno de éstos. *La Biblia de las Américas.*
	Fíjense cómo crecen los lirios. No trabajan ni hilan; sin embargo, les digo que ni siquiera Salomón, con todo su esplendor, se vestía como uno de ellos. *Nueva Biblia al Día.*
	Aprendan de las flores del campo: no trabajan para hacerse sus vestidos y, sin embargo, les aseguro que ni el rey Salomón, con todas sus riquezas, se vistió tan bien como ellas. *Traducción en lenguaje actual.*

Table 4.8: Examples of parallel paraphrasing data with English, French, Tagalog and Spanish paraphrases in Bible translation.

Chapter 5

Building Language Family with Incomplete Information

> "The art of translation lies less in knowing the other language than in knowing your own."
>
> *Ned Rorem*

In the previous two chapters, we have carefully examined ways to make the best use of information on nearby languages/ language families as well as different paraphrases within the same language for our task of translation of a multi-source text into a new, low-resource language. However, in real-world situations, such information may not be complete. We may not have access to multiple paraphrases within the same language and we may have limited information about which languages are close by and which language family the target language belongs to. In such situations, we can best make use of the existing multi-source text by statistically building metric spaces of language closeness to predict which languages are close to the target language. We can then use the synthetic language families that we built for multilingual training.

5.1 Introduction

2020 is the year that we started the life-saving hand washing practice globally. Applications like translating water, sanitation, and hygiene (WASH) guidelines into severely low-resource languages are very impactful in tribes like those in Papua New Guinea with 839 living languages [108, 267]. Translating humanitarian texts like WASH guidelines with scarce data and expert help is key [30].

Figure 5.1: A Quechua-speaking community gathering in Peru. Photograph by Mark Bean.

Inspired by such applications, we translate multilingually known texts into a severely low-resource language. In the translation process, we focus on five challenges that are not addressed previously[1]. Most multilingual transformer works that translate into low-resource language limit their training data to available data in the same or close-by language families or the researchers' intuitive discretion; and are mostly limited to less than 30 languages [113, 322, 325]. Instead, we examine ways to pick useful source languages from 124 source languages in a principled fashion. Secondly, most works require at least 4,000 lines of low-resource data [177, 234, 322]; we use only ∼1,000 lines of low-resource data to simulate real-life situation of having extremely small seed target translation. Thirdly, many works use rich-resource languages as hypothetical low-resource languages. Moreover, most works do not treat named entities separately; we add an order-preserving lexiconized component for more accurate translation of named entities. Finally, many multilingual works present final results as sets of translations from all source languages; we build a novel method to combine all translations into one.

[1]The material in this chapter was originally published in SIGTYP at NAACL, 2021 [321].

Eastern Pokomchi		English	
FAMD	*FAMP*	*FAMD*	*FAMP*
Chuj*	Dadibi	Danish*	Dutch*
Cakchiquel*	Thai	Norwegian*	Afrikaans*
Guajajara*	Gumatj	Italian	Norwegian*
Toba	Navajo	Afrikaans*	German*
Myanmar	Cakchiquel*	Dutch*	Danish*
Slovenský	Kanjobal	Portuguese	Spanish
Latin	Guajajara*	French	Frisian*
Ilokano	Mam*	German*	Italian
Norwegian	Kim	Marshallese	French
Russian	Chuj*	Frisian*	Portuguese

Table 5.1: Top ten languages closest to Eastern Pokomchi (left) and English (right) in ranking 124 source languages. *FAMD* and *FAMP* are two constructions of Family of Choice (*FAMC*) by distortion and performance metrics respectively. All are trained on ∼1,000 lines. We star those in Family of Origin.

We have five contributions. Firstly, we rank the 124 source languages to determine their closeness to the low-resource language and choose the top few. We call the linguistic definition of language family *Family of Origin* (FAMO), and we call the empirical definition of higher-ranked languages using our metrics *Family of Choice* (FAMC). They often overlap, but may not coincide.

Secondly, we build an *Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer* (IPML) training on ∼1,000 lines of low-resource data. Using iterative pretraining, we get a +23.9 BLEU increase over a multilingual order-preserving lexiconized transformer baseline (MLc) using English as a hypothetical low-resource language, and a +10.3 BLEU increase over our asymmetric baseline. Training with the low-resource language on both the source and target sides boosts translation into the target side. Training on 1,093 lines from the book of Luke, we reach a 23.7 BLEU score testing on 30,022 lines of Bible. We have a 42.8 BLEU score for Portuguese-English translation on the medical EMEA dataset.

Thirdly, we use a real-life severely low-resource Mayan language, Eastern Pokomchi, a Class 0 language [148] as one of our experiment setups. In addition, we also use English as a hypothetical low-resource language for easy evaluation.

We also add an order-preserving lexiconized component to translate named entities well. To solve the variable-binding problem to distinguish "Ian calls Yi" from "Yi calls Ian" [95, 109, 322], we build a lexicon table for 2,939 Bible named entities in 124 source languages including more than 66 severely low-resource languages.

Finally, we combine translations from all source languages by using a novel method. For every sentence, we find the translation that is closest to the translation cluster center. The expected BLEU score of our combined translation is higher than translation from any of the individual source.

5.2 Related Works

5.2.1 Multilingual Pretraining

Recent research on machine polyglotism involves training machines to be adept in many languages by adding language labels in the training data with a single attention [94, 103, 117, 147, 323]. Some explores data symmetry [29, 97, 176]. Zero-shot translation in severely low-resource settings exploits the massive multilinguality, cross-lingual transfer, pretraining, iterative back-translation and freezing subnetworks [22, 52, 72, 170, 174, 177, 183, 209, 223, 284, 308, 310].

5.2.2 Linguistic Distance

To construct linguistic distances [53, 121, 212], researchers explore typological distance [50, 60, 123, 226, 237, 277], on World Atlas of Language Structures [60], lexical distance on the Swadesh list [139], normalized Levenshtein distance and Jaccard distance [2, 136, 261], sonority distance [218], trade cost [144], structural distance on immigrants' fluency [49] and spectral distance [73].

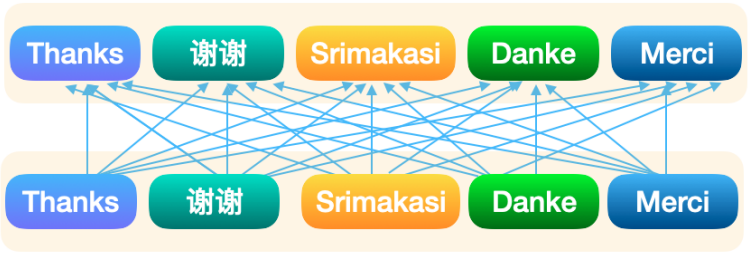
5.3 Methodology

5.3.1 Multilingual Order-preserving Lexiconized Transformer

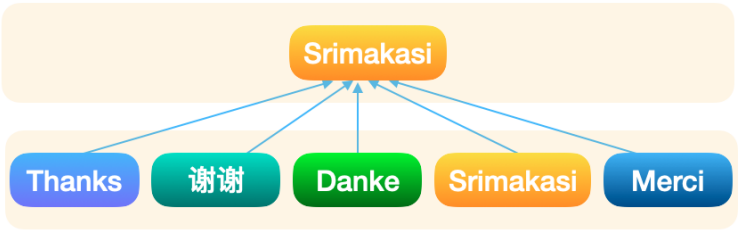
Multilingual Transformer

In training, each sentence is labeled with the source and target language label. For example, if we translate from Chuj ("ca") to Cakchiquel ("ck"), each source sentence is tagged with `__opt_src_ca __opt_tgt_ck`. A sample source sentence is "`__opt_src_ca __opt_tgt_ck` Tec'b'ejec e b'a mach ex tzeyac'och Jehová yipoc e c'ool".

We train on Geforce RTX 2080 Ti using ∼100 million parameters, a 6-layer encoder and a 6-layer decoder that are powered by 512 hidden states, 8 attention heads, 512 word vector size, a dropout of 0.1, an attention dropout of 0.1, 2,048 hidden transformer feed-forward units, a batch size of 6,000, "adam" optimizer, "noam" decay method, and a label smoothing



(a) Complete graph configuration



(b) Star graph configuration

Figure 5.2: (a) Complete graph configuration of translation paths (Many-to-many) in an example of multilingual translation. (b) Star configuration of translation paths (Many-to-one) using Indonesian as the low-resource example.

of 0.1 and a learning rate of 2.5 on OpenNMT [157, 298]. After 190,000 steps, we validate based on BLEU score with early stopping patience of 5.

Star Versus Complete Configuration

We show two configurations of translation paths in Figure 5.2: *star* graph (multi-source single-target) configuration and *complete* graph (multi-source multi-target) configuration. The complete configuration data increases quadratically with the number of languages while the star configuration data increases linearly.

Order-preserving Lexiconized transformer

The variable binding problem issue is difficult in severely low-resource scenario; most neural models cannot distinguish the subject and the object of a simple sentence like "Fatma asks her sister Wati to call Yi, the brother of Andika", especially when all named entities appear once or never appear in training [95, 109]. Recently, researchers use order-preserving lexiconized Neural Machine Translation models where named entities are sequentially tagged in a sentence as `__NE`s [322]. The previous example becomes "`__NE0` asks her sister `__NE1` to call `__NE2`, the brother of `__NE3`".

This method works under the assumption of translating a closed text known in advance. Its success relies on good coverage of named entities. To cover many named entities, we

Source Sentence	IPML Translation	Reference
En terwyl Hy langs die see van Galiléa loop, sien Hy Simon en Andréas, sy broer, besig om 'n net in die see uit te gooi; want hulle was vissers.	And as He drew near to the lake of Galilee, He Simon saw Andrew, and his brother, lying in the lake, for they were fishermen.	And walking along beside the Sea of Galilee, He saw Simon and his brother Andrew casting a small net in the sea; for they were fishers.
En toe Hy daarvandaan 'n bietjie verder gaan, sien Hy Jakobus, die seun van Sebedéüs, en Johannes, sy broer, wat besig was om die nette in die skuit heel te maak.	And being in a distance, He saw James, the son of Zebedee, and John, his brother. who kept the nets in the boat.	And going forward from there a little, He saw James the son of Zebedee, and his brother John. And they were in the boat mending the nets.
En verder Jakobus, die seun van Sebedéüs, en Johannes, die broer van Jakobus- aan hulle het Hy die bynaam Boanérges gegee, dit is, seuns van die donder-	And James the son of Zebedee, and John the brother of James; and He gave to them the name, which is called Boanerges, being of the voice.	And on James the son of Zebedee, and John the brother of James, He put on them the names Boanerges, which is, Sons of Thunder.

Table 5.2: Examples of Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer (IPML) translation from Afrikaans to English using *FAMP*. We train on only 1,093 lines of English data.

build on existing research literature [311, 322] to construct a massively parallel lexicon table that covers 2,939 named entities across 124 languages in our Bible database. Our lexicon table is an expansion of the existing literature that covers 1,129 named entities [311]. We add in 1,810 named entities that are in the extreme end of the tail occurring only once. We also include 66 more real-life severely low-resource languages.

For every sentence pair, we build a target named entity decoding dictionary by using all target lexicons from the lexicon table that match with those in the source sentence. In severely low-resource setting, our sequence tagging is larged based on dictionary look-up; we also include lexicons that are not in the dictionary but have small edit distances with the source lexicons. In evaluation, we replace all the ordered `__NE`s using the target decoding dictionary to obtain our final translation.

Let us translate "Fatma asks her sister Wati to call Yi, the brother of Andika" to Chinese and German. Our tagged source sentence that translates to Chinese is "`__opt_src_en __opt_tgt_zh __NE0` asks her sister `__NE1` to call `__NE2`, the brother of `__NE3`"; and we use `__opt_tgt_de` for German. The source dictionary is "`__NE0`: Fatma, `__NE1`: Wati, `__NE2`: Yi, `__NE3`: Andika" and we create the target dictionaries. The Chinese output is "`__NE0`叫她的姐妹`__NE1`去打电话给`__NE3`的兄弟`__NE2`" and the German output is "`__NE0` bittet ihre Schwester `__NE1` darum, `__NE2`, den Bruder `__NE3`, anzurufen". We decode the named entities to get final translations.

5.3.2 Ranking Source Languages

Existing works on translation from multiple source languages into a single low-resource language usually have at most 30 source languages [113, 322, 325]. They are limited within the same or close-by language families, or those with available data, or those chosen based on the researchers' intuitive discretion. However, the set of related languages chosen as source languages under such scenarios could be incomplete. Instead, we examine ways to pick useful source languages in a principled fashion motivated by cross-lingual impacts and similarities [45, 63, 130, 211, 250, 266, 271, 289]. We find that using many languages that are distant to the target low-resource language may produce marginal improvements, if not negative impact. Indeed, existing literature on zero-shot translation also suffers from the limitation of linguistic distance between the source languages and the target language [170, 177, 223]. We therefore rank and select the top few source languages that are closer to the target low-resource language using the two metrics below.

We rank source languages according to their closeness to the low-resource language. We construct the Family of Choice (FAMC) by comparing different ways of ranking linguistic distances empirically based on the small low-resource data.

To rank linguistic distances empirically based on the small low-resource data, we propose two different ways to construct FAMCs: one by distortion, the other by fertility. Distortion represents the amount of rearrangement work that needs to be done if we were to have a word-by-word translation of the source sentence. Fertility of a source word represents the number of target words produced by that source word. Ideally, if two languages are very close, we expect to see a one-on-one mapping between source and target words. Moreover, we also expect to do minimal rearrangement work if two languages are close. Therefore, at a higher level of understanding, we aim to rank languages by how high its probability of distortion equalling zero and fertility equalling one is. The higher the probability a source language has, the closer this source language is to the target language. With this understanding, let us explain the mathematical formulation of our distortion-based and fertility-based similarity measures in the following.

Let S_s and S_t be the source and target sentences, let L_s be the source length, let $P(S_t = s_t|s_s, l_s)$ be the alignment probability, let F_s be the fertility of how many target words a source word is aligned to, let D_t be the distortion based on the fixed distance-based reordering model [161].

We first construct a word-replacement model based on aligning the small amount of target low-resource data with that of each source language using `fast_align` [78]. We replace every source word with the most probable target word according to the product of the alignment probability and the probability of fertility equalling one and distortion equalling zero $P(F_s = 1, D_t = 0|s_t, s_s, l_s)$. We choose a simple word-replacement model

because we aim to work with around 1,000 lines of low-resource data. For fast and efficient ranking on such small data, a word-replacement model suits our purpose.

We use two alternatives to create our FAMCs. Our distortion measure is the probability of distortion equalling zero, $P(D_t = 0|s_t, s_s, l_s)$, aggregated over all words in a source language. We use the distortion measure to rank the source languages and obtain the distortion-based FAMC (*FAMD*); we use the translation BLEU scores of the word-replacement model as another alternative to build the performance-based FAMC (*FAMP*). In Table 5.1, we list the top ten languages in FAMD and FAMP for Eastern Pokomchi and English. We use both alternatives to build FAMCs.

To prepare for transformer training, we choose the top ten languages neighboring our target low-resource language in FAMD and FAMP. We choose ten because existing literature shows that training with ten languages from two neighboring language families is sufficient in producing quality translation through cross-lingual transfer [322]. Since for some low-resource languages, there may not be ten languages in FAMO in our database, we add languages from neighboring families to make an expanded list denoted by *FAMO*$^+$.

5.3.3 Iterative Pretraining

Having ranked similar languages to the low-resource language, we choose a suitable set of source languages for training to translate into the low-resource language. For training purposes, we propose two stages of pretraining using multilingual order-preserving lexiconized transformers on the complete and the star configuration. We design iterative pretraining on symmetric data to address catastrophic forgetting that is common in training [98, 156]. The word "iterative" refers to the multi-stage pretraining that is focused on the complete configuration on all source languages first followed by adding the target low-resource language. Each stage of pretraining has symmetric data for optimization. We propose iterative pretraining to promote data symmetry, address catastrophic forgetting, and increase robustness.

Stage 1: Pretraining on Neighbors

Firstly, we pretrain on the complete graph configuration of translation paths using the top ten languages neighboring our target low-resource language in FAMD, FAMP, and FAMO$^+$ respectively. Low-resource data is excluded in training.

We use the multilingual order-preserving lexiconized transformer. Our vocabulary is the combination of the vocabulary for the top ten languages together with the low-resource vocabulary built from the ∼1,000 lines. The final model can translate from any of the ten languages to each other.

Source Sentence	IPML Translation	Reference
Ket idi limmabas iti dinna ti baybay ti Galilea, nakitana ni Simon ken ni Andres a cabsatna, nga iwaywayatda ti iket iti baybay; ta dumadaclisda idi.	Eh noq ojik i rub'an i Jesús juntar i k'isa palaw i Galilea, xrilow reje i Simón ruch'ihil i Andres, re' i rutuut i k'isa palaw, ruum jinaj i k'isa palaw barco.	Noq k'ahchi' rik'iik i Jesús chi chii' i k'isa palaw ar Galilea, xrilow reje wach i Simón ruch'ihil i ruchaaq', Andres rub'ihnaal. Re' keh aj karineel taqe, k'ahchi' kikutum qohoq i kiya'l pan palaw.
Ket idi nagna pay bassit nakitana ni Santiago nga anac ni Zebedeo ken ni Juan a cabsatna, nga addada idi iti barangayda, a tartarimaanenda dagiti iketda.	Eh noq ojik i rub'an i Jesús, xrilow i Jacobo, re' i Jacobo rak'uun i Zebedeo, re' Juan rub'ihnaal, ruch'ihil taqe i raj tahqaneel. eh xkikoj wo' wach chinaah i k'isa palaw.	Eh junk'aam-oq chik i xb'ehik reje i Jesús, xrilow kiwach i ki'ib' chi winaq kichaaq' kiib', re' Jacobo, re' Juan, rak'uun taqe i Zebedeo. Eh wilkeeb' chupaam jinaj i barco, k'ahchi' kik'ojem wach i kiya'l b'amb'al kar.
Ket immasideg ni Jesus ket iniggamanna iti imana ket pinatacderna; ket pinanawan ti gorigor , ket nagservi cadacuada.	Eh re' Jesús xujil i koq riib', xutz'a'j i koq chinaah i q'ab'. eh re' i kaq tz'a' chi riij. eh jumehq'iil xwuktik johtoq, re' chik i reh xutoq'aa' cho yejanik kiwa'.	Eh re' i Jesús xujil i koq riib' ruuk' i yowaab', xuchop chi q'ab', xruksaj johtoq, eh jumehq'iil xik'ik i tz'a' chi riij. Eh re' chik i reh xutoq'aa' cho yej-anik kiwa'.

Table 5.3: Examples of Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer (IPML) translation from Ilokano to Eastern Pokomchi using *FAMD*. We train on only 1,086 lines of Eastern Pokomchi data.

Stage 2: Adding Low-Resource Data

We include the low-resource data in the second stage of training. Since the low-resource data covers $\sim 3.5\%$ of the text while all the source languages cover the whole text, the data is highly asymmetric. To create symmetric data, we align the low-resource data with the subset of data from all source languages. As a result, all source languages in the second stage of training have $\sim 3.5\%$ of the text that is aligned with the low-resource data. We therefore create a complete graph configuration of training paths using all the eleven languages.

Using the pretrained model from the previous stage, we train on the complete graph configuration of translation paths from all eleven languages including our low-resource language. The vocabulary used is the same as before. We employ the multilingual order-preserving lexiconized transformer for pretraining. The final model can translate from any of the eleven languages to each other.

Source Sentence	IPML Translation	Reference
Caso detecte efeitos graves ou outros efeitos não mencionados neste folheto, informe o médico veterinário.	If you notice any side effects or other side effects not mentioned in this leaflet, please inform the vétérinaire.	If you notice any serious effects or other effects not mentioned in this leaflet, please inform your veterinarian.
No tratamento de Bovinos com mais de 250 Kg de peso vivo, dividir a dose de forma a não administrar mais de 10 ml por local de injecção.	In the treatment of infants with more than 250 kg in vivo body weight, a the dose to not exceed 10 ml per injection.	For treatment of cattle over 250 kg body weight, divide the dose so that no more than 10 ml are injected at one site.
No entanto, uma vez que é possível a ocorrência de efeitos secundários, qualquer tratamento que exceda as 1-2 semanas deve ser administrado sob supervisão veterinária regular.	However, because any of side effects is possible, any treatment that 1-5 weeks should be administered under regular supravegghere.	However, since side effects might occur, any treatment exceeding 1–2 weeks should be under regular veterinary supervision.

Table 5.4: Examples of IPML translation on medical EMEA dataset from Portuguese to English using $FAMO^{+}$.

5.3.4 Final Training

Finally, we focus on translating into the low-resource language. We use the symmetric data built from the second stage of pretraining. However, instead of using the complete configuration, we use the star configuration of translation paths from the all source languages to the low-resource language. All languages have $\sim 3.5\%$ of the text.

Using the pretrained model from the second stage, we employ the multilingual order-preserving lexiconized transformer on the star graph configuration. We use the same vocabulary as before. The final trained model can translate from any of the ten source languages to the low-resource language. Using the lexicon dictionaries, we decode the named entities and obtain our final translations.

5.3.5 Combination of Translations

We have multiple translations, one per each source language. Combining all translations is useful for both potential post-editing works and systematic comparison of different experiments especially when the sets of the source languages differ.

Our combination method assumes that we have the same text in all source languages. For each sentence, we form a cluster of translations from all source languages into the low-resource language. Our goal is to find the translation that is closest to the center of the cluster. We rank all translations according to how centered this translation is with

respect to other sentences by summing all its similarities to the rest. The highest score is the closest to the cluster center. We take the most centered translation for each sentence and output our combined result. The expected BLEU score of our combined translation is higher than translation from any of the individual source language.

5.4 Data

We use the Bible dataset and the medical EMEA dataset [196, 286]. EMEA dataset is from the European Medicines Agency and contains a lot of medical information that may be beneficial to the low-resource communities. Our method can be applied to other datasets like WASH guidelines.

For the Bible dataset, we use 124 source languages with 31,103 lines of data and a target low-resource language with ∼1,000 lines (∼3.5%) of data. We have two setups for the target low-resource language. One uses Eastern Pokomchi, a Mayan language; the other uses English as a hypothetical low-resource language [2]. We train on only ∼1,000 lines of low-resource data from the book of Luke and test on the 678 lines from the book of Mark. Mark is topically similar to Luke, but is written by a different author. For the first stage of pretraining, we use 80%, 10%, 10% split for training, validation and testing. For the second stage onwards, we use 95%, 5% split of Luke for training and validation, and 100% of Mark for testing.

Eastern Pokomchi is Mayan, and English is Germanic. Since our database does not have ten members of each family, we use FAMO$^+$, the expanded version of FAMO. For English, we include five Germanic languages and five Romance languages in FAMO$^+$; for Eastern Pokomchi, we include five Mayan languages and five Amerindian languages in FAMO$^+$. The Amerindian family is broadly believed to be close to the Mayan family by the linguistic community.

We construct FAMCs by comparing different ways of ranking linguistic distances empirically based on ∼1,000 lines of training data. In Table 5.1, we list the top ten languages for Eastern Pokomchi and English in FAMD and FAMP respectively.

To imitate the real-life situation of having small seed target translation data, we choose to use ∼1,000 lines (∼3.5%) of low-resource data. We also include Eastern Pokomchi in addition to using English as a hypothetical low-resource language. Though data size can be constrained to mimic severely low-resource scenarios, much implicit information is still used for the hypothetical low-resource language that is actually rich-resource. For example, implicit information like English is Germanic is often used. For real low-resource scenarios,

[2] In Table 5.1 and Table 5.7, Kanjobal is Eastern Kanjobal, Mam is Northern Mam, Cuzco is Cuzco Quechua, Ayacucho is Ayacucho Quechua, Bolivian is South Bolivian Quechua, and Huallaga is Huallaga Quechua.

Experiments	IPML	MLc	MLs	PMLc	PMLs	AML
Pretrained	✓			✓	✓	
Iterative	✓					
Lexiconized	✓	✓	✓	✓	✓	✓
Symmetrical	✓	✓	✓	✓	✓	
Star	✓		✓		✓	
Complete	✓	✓		✓		✓
Combined	37.3	13.4	14.7	34.7	35.7	27.0
German	35.0	11.6	12.3	33.3	34.5	25.4
Danish	36.0	12.5	12.4	33.3	34.2	26.2
Dutch	35.6	11.5	11.1	32.3	33.7	25.0
Norwegian	35.7	12.3	12.0	33.2	34.1	25.8
Swedish	34.5	11.8	12.4	32.3	33.4	24.9
Spanish	36.4	11.7	11.8	34.1	35.0	26.2
French	35.3	10.8	10.8	33.1	34.0	25.8
Italian	35.9	11.7	11.7	34.3	34.5	26.1
Portuguese	31.5	9.6	10.1	30.0	30.4	23.1
Romanian	34.6	11.3	12.1	32.3	33.2	25.0

Table 5.5: Comparing our iteratively pretrained multilingual order-preserving lexiconized transformer (IPML) with the baselines training on 1,093 lines of English data in *FAMO*$^+$. We checkmark the key components used in each experiments and explain all the baselines in details in Section 5.5.

the family information may have yet to be determined; the neighboring languages may be unknown or incomplete, and if they are known, they are highly likely to be low-resource too. We thus use Eastern Pokomchi as our real-life severely low resource language.

To understand Eastern Pokomchi, it is a Mayan language that is morphological rich, ergative and agglutinative [5, 56]. It has been isolated from other languages historically and is seen by many as non-transparent, complex and obscure [86].

In addition to the Bible dataset, we work with the medical EMEA dataset [286]. Using English as a hypothetical language, we train on randomly sampled 1,093 lines, and test on 678 lines of data. Since there are only 9 languages in Germanic and Romance families in EMEA dataset, we include a slavic language Polish in our FAMO$^+$ for experiments.

The EMEA dataset is less than ideal comparing with the Bible dataset. The Bible dataset contains the same text for all source languages; however, the EMEA dataset does not contain the same text. It is built from similar documents but has different parallel data for each language pair. Therefore, during test time, we do not combine the translations from various source languages in the EMEA dataset.

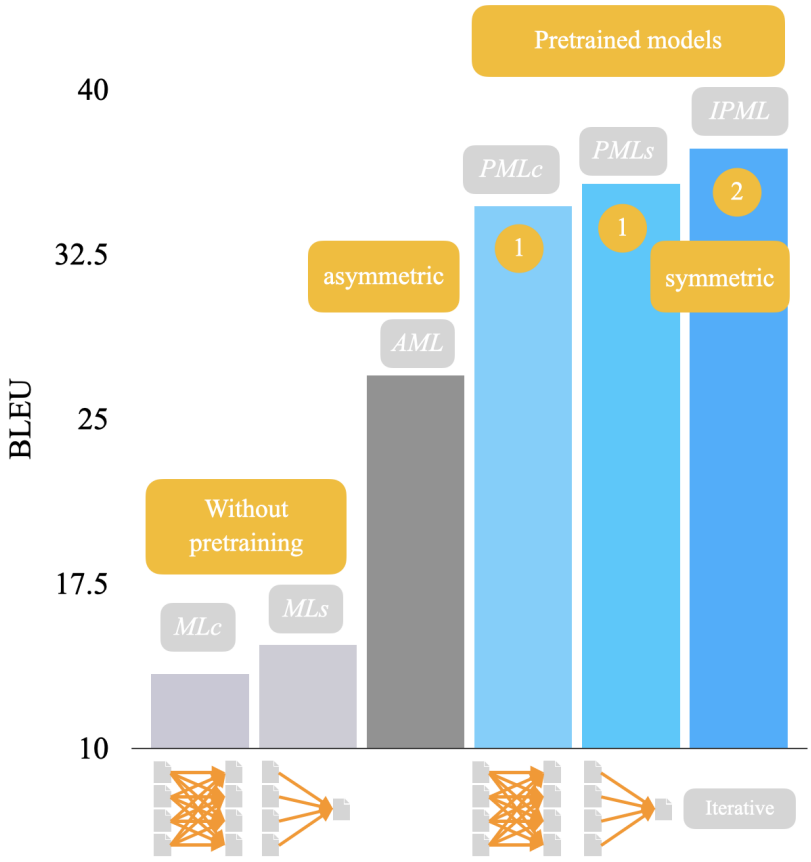


Figure 5.3: Comparing our method with different baselines for translation into English as a hypothetical low-resource language using ∼1,000 lines of data.

5.5 Results

We summarize our results in Figure 5.3 and we show our result is generalizable to medical domains by looking at EMEA datasets which we will focus on at the end of the results section.

In Figure 5.3, we compare our iteratively pretrained multilingual order-preserving lexiconized transformer (IPML) with five baselines in Table 5.5. *MLc* is a baseline model of multilingual order-preserving lexiconized transformer training on complete configuration; in other words, we skip the first stage of pretraining and train on the second stage in Chapter 5.3.3 only. *MLs* is a baseline model of multilingual order-preserving lexiconized transformer training on star configuration; in other words, we skip both steps of pretraining and train on the final stage in Chapter 5.3.4 only. *PMLc* is a baseline model of pretrained multilingual order-preserving lexiconized transformer training on complete configuration; in other words, we skip the final stage of training after completing both stages of pretraining. *PMLs* is a baseline model of pretrained multilingual order-preserving lexiconized transformer training

Input Language Family					
By Linguistics		By Distortion		By Performance	
FAMO$^+$		*FAMD*		*FAMP*	
Source	BLEU	Source	BLEU	Source	BLEU
Combined	37.3	Combined	38.3	Combined	39.4
German	35.0	German	36.7	German	37.6
Danish	36.0	Danish	37.1	Danish	37.5
Dutch	35.6	Dutch	35.6	Dutch	36.7
Norwegian	35.7	Norwegian	36.9	Norwegian	37.1
Swedish	34.5	Afrikaans	38.3	Afrikaans	39.3
Spanish	36.4	Marshallese	34.7	Spanish	38.4
French	35.3	French	36.0	French	36.6
Italian	35.9	Italian	36.9	Italian	37.7
Portuguese	31.5	Portuguese	32.9	Portuguese	33.1
Romanian	34.6	Frisian	36.1	Frisian	36.9

Table 5.6: Performance of Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer (IPML) training for English on *FAMO*$^+$, *FAMD* and *FAMP*. We train on only 1,093 lines of English data.

on star configuration; in other words, after the first stage of pretraining, we skip the second stage of pretraining and proceed to the final training directly. Finally, *AML* is a baseline model of multilingual order-preserving lexiconized transformer on asymmetric data. We replicate the ∼1,000 lines of the low-resource data till it matches the training size of other source languages; we train on the complete graph configuration using eleven languages. Though the number of low-resource training lines is the same as others, information is highly asymmetric.

Pretraining is key as IPML beats the two baselines that skip pretraining in Table 5.5. Using English as a hypothetical low-resource language training on FAMO$^+$, combined translation improves from 13.4 (MLc) and 14.7 (MLs) to 37.3 (IPML) with iterative pretraining. Training with the low-resource language on both the source and the target sides boosts translation into the target side. Star configuration has a slight advantage over complete configuration as it gives priority to translation into the low-resource language. Iterative pretraining with BLEU score 37.3 has an edge over one stage of pretraining with scores 34.7 (PMLc) and 35.7 (PMLs).

All three pretrained models on symmetric data, IPML, PMLc and PMLs, beat asymmetric baseline AML. In Table 5.5, IPML has a +10.3 BLEU increase over our asymmetric baseline on combined translation using English as a hypothetical low-resource language training on

Input Language Family					
By Linguistics		By Distortion		By Performance	
FAMO$^+$		*FAMD*		*FAMP*	
Source	BLEU	Source	BLEU	Source	BLEU
Combined	23.0	Combined	23.1	Combined	22.2
Chuj	21.8	Chuj	21.9	Chuj	21.6
Cakchiquel	22.2	Cakchiquel	22.1	Cakchiquel	21.3
Guajajara	19.7	Guajajara	19.1	Guajajara	18.8
Mam	22.2	Russian	22.2	Mam	21.7
Kanjobal	21.9	Toba	21.9	Kanjobal	21.4
Cuzco	22.3	Myanmar	19.1	Thai	21.8
Ayacucho	21.6	Slovenský	22.1	Dadibi	19.8
Bolivian	22.2	Latin	21.9	Gumatj	19.1
Huallaga	22.2	Ilokano	22.5	Navajo	21.3
Aymara	21.5	Norwegian	22.6	Kim	21.5

Table 5.7: Performance of Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer (IPML) training for Eastern Pokomchi on *FAMO*$^+$, *FAMD* and *FAMP*. We train on only 1,086 lines of Eastern Pokomchi data.

FAMO$^+$. All four use the same amount of data, but differ in training strategies and data configuration. In severely low-resource scenarios, effective training strategies on symmetric data improve translation greatly.

We show IPML results training on different sets of source languages in FAMO$^+$, FAMD, and FAMP, for English and Eastern Pokomchi in Table 5.6 and 5.7. FAMP performs the best for translation into English while both FAMP and FAMD outperforms FAMO$^+$ as shown in Table 5.6. FAMD performs best for translation into Eastern Pokomchi as shown in Table 5.7. Afrikaans has the highest score for English's FAMD and FAMP, outperforming Dutch, German or French. A reason may be that Afrikaans is the youngest language in the Germanic family with many lexical and syntactic borrowings from English and multiple close neighbors of English [108].

Note that we are not focusing on comparing FAMO$^+$ with FAMCs, our focus is what we could do to boost translation performance when there is incomplete information of which languages are close to our given low-resource language. When language family information is limited or incomplete, i.e., when FAMO$^+$ is limited or incomplete, constructing FAMC to determine neighbors is very useful in translation.

Comparing Eastern Pokomchi results with English results, we see that translation into real-life severely low-resource languages is more difficult than translation into hypothetical

Source	BLEU
Combined	N.A.
German	34.8
Danish	37.7
Dutch	39.7
Swedish	37.7
Spanish	42.8
French	41.6
Italian	39.2
Portuguese	42.8
Romanian	40.0
Polish	34.1

Table 5.8: IPML Performance on the EMEA dataset trained on only 1,093 lines of English data.

Source	BLEU
Combined	23.7
German	21.6
Danish	22.9
Dutch	21.2
Norwegian	21.3
Swedish	19.9
Spanish	22.9
French	22.3
Italian	21.8
Portuguese	20.7
Romanian	16.3

Table 5.9: IPML Performance on the entire Bible excluding ∼1k lines of training and validation data.

ones. The combined score is 38.3 for English in Table 5.6 and 23.1 for Eastern Pokomchi on FAMD in Table 5.7. Eastern Pokomchi has ejective consonants which makes tokenization process difficult. It is agglutinative, morphologically rich and ergative just like Basque [5, 56]. It is complex, unique and nontransparent to the outsider [86]. Indeed, translation into real severely low-resource languages is difficult.

We are curious of how our model trained on ∼1,000 lines of data performs on the rest of the Bible. In other words, we would like to know how IPML performs if we train on ∼3.5% of the Bible and test on ∼96.5% of the Bible. In Table 5.9, training on 1,093 lines from the book of Luke, we achieve a BLEU score of 23.7 for IPML using FAMP in English [3].

We show qualitative examples in Table 5.2 and 5.3. The source content is translated well overall and there are a few places for improvement in Table 5.2. The words "fishermen" and "fishers" are paraphrases of the same concept. IPML predicts the correct concept though it is penalized by BLEU.

Infusing the order-preserving lexiconized component to our training greatly improves qualitative evaluation. But it does not affect BLEU much as BLEU has its limitations in severely low-resource scenarios. This is why all experiments include the lexiconized component in training. The BLEU comparison in our paper also applies to the comparison of all experiments without the order-preserving lexiconized component. This is important in real-life situations when a low-resource lexicon list is not available, or has to be invented. For

[3]A previous version of this work shows higher BLEU scores with random sampling. Since active learning is not the focus of this section, we show all results training on the book of Luke in this paper. For further results in active learning, please refer to our follow-up work in the next section [320].

example, a person growing up in a local village in Papua New Guinea may have met many people named "Bosai" or "Kaura", but may have never met a person named "Matthew", and we may need to create a lexicon word in the low-resource language for "Matthew" possibly through phonetics.

We also see good results with the medical EMEA dataset. Treating English as a hypothetical low-resource language, we train on only 1,093 lines of English data. For Portuguese-English translation, we obtain a BLEU score of 42.8 while the rest of languages all obtain BLEU scores above 34 in Table 5.8 and Table 5.4. In Table 5.4, we see that our translation is very good, though a few words are carried from the source language including "vétérinaire". This is mainly because our ∼1,000 lines contain very small vocabulary; however, by carrying the source word over, key information is preserved.

5.6 Conclusion

We use ∼1,000 lines of low-resource data to translate a closed text that is known in advance to a severely low-resource language by leveraging massive source parallelism. We present two metrics to rank the 124 source languages and construct FAMCs. We build an iteratively pretrained multilingual order-preserving lexiconized transformer and combine translations from all source languages into one by using our centric measure. Moreover, we add a multilingual order-preserving lexiconized component to translate the named entities accurately. We build a massively parallel lexicon table for 2,939 Bible named entities in 124 source languages, covering more than 66 severely low-resource languages. Our good result for the medical EMEA dataset shows that our method is useful for other datasets and applications.

Our final result can also serve as a ranking measure for linguistic distances though it is much more expensive in terms of time and resources. In the future, we would like to explore more metrics that are fast and efficient in ranking linguistic distances to the severely low-resource language.

Having examined ways to build linguistic measure and build metric spaces of language closeness that is suitable for multilingual training, when language family information is not complete, we have established a proven way to train multilingual models for the translation task of a multi-source text into a new, low-resource language. And we have shown success translating the whole text using about 1,000 lines of training data as our seed corpus. And our next question is, what is the optimal 1,000 lines to produce as seed corpus to optimize machine translation performance? In other words, is there a way to use active learning to determine the most effective way to build seed corpus to optimize for machine translation? This is what we would like to address in the next chapter.

Part II

Human Machine Translation

Having examined source parallelism in the first part of the thesis (Chapter 3, Chapter 4 and Chapter 5), we build a human machine translation workflow algorithm for machine translation systems to collaborate with human translators to expedite the translation process of a multi-source text into new, low-resource languages (Chapter 6, Chapter 7 and Chapter 8). Our proposed human machine translation is not to replace the human translators with machine translation systems, but instead, to get the best of both worlds and to expedite the translation process. In our translation process, human translators are informed by machine sentence ranking through active learning to produce a seed corpus. Machine systems then use this seed corpus to produce a full translation draft. Human translators post-edit the draft, and feed new data to machines each time they finish post-editing a portion of the text. In each iteration, machines produce better and better drafts with new data, and human translators find it easier and faster to post-edit. Together they complete the translation of the whole text into an severely low-resource language.

In Chapter 6, we first develop various active learning methods on known languages and transfer ranking to the new, low-resource language. In Chapter 7, we activate the knowledge of large multilingual models by proposing multilingual and multi-stage adaptations through different training schedules; we find that adapting pretrained models to the domain and then to the low-resource language works best. Thirdly, we aggregate scores from 115 languages to provide a universal ranking and increase robustness by *relaxed memoization* method. Having examined both source parallelism and human machine translation workflow, we evaluate our work in all previous chapters by translating academic progress to the real-world translation process in a case study in Quechuan language family in Chapter 8. We collaborate extensively with a translation group with in-depth knowledge of various Quechuan languages and focus on evaluation. We find that machine translation performance is significantly positively correlated with language similarity. The more connected a language is, the easier it is to translate into it. In addition, we find that decluttering poorly-connected languages improves translation score. Using this finding, we achieve good results in translating into a new, low-resource language called Sihuas Quechua.

Chapter 6

Active Learning for Building a Seed Corpus

"Tell me and I forget.
Teach me and I remember.
Involve me and I learn."

Benjamin Franklin

Having examined interlingual transfer both within and across different language families, paraphrases within the same language, and the thorough way of building linguistic measure and building metric spaces of language closeness, we have established a proven way to use a small seed corpus (around 1,000 lines of low-resource language data) to complete our translation task of translating a multi-source text into a new, low-resource language. In this chapter, we examine the optimal way to build seed corpus to best help with machine translation. We incorporate active learning to learn sentence ranking in the low-resource language and inform human translators which set of sentences to translate first, in order to produce the most effective and useful seed corpus.

6.1 Introduction

Machine translation has flourished ever since the first computer was made [133, 229]. Over the years, human translation is assisted by machine translation to remove human bias and translation capacity limitations [37, 38, 160, 162, 173, 251]. By learning human translation taxonomy and post-editing styles, machine translation borrows many ideas from human translation to improve performance through active learning [44, 66, 262]. As discussed in

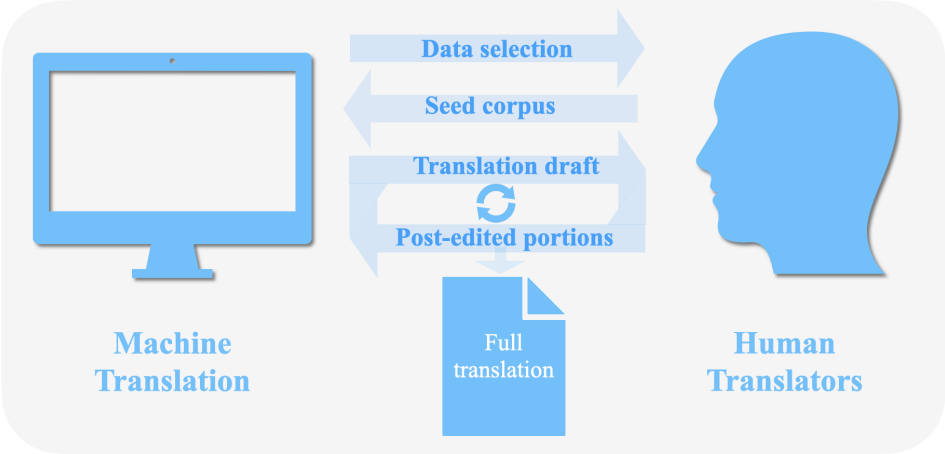


Figure 6.1: Translation workflow for severely low-resource languages.

Chapter 1, we propose a workflow[1] to bring human translation and machine translation to work together seamlessly in translation of a closed text into a severely low-resource language as shown in Figure 6.1 and Figure 6.2 and Algorithm 1. With our human machine translation framework, we are interested in translating a given text that has many existing translations in different languages into a severely low-resource language well.

In this translation framework, human translators are informed by machine data selection process through active learning to produce a seed corpus. Machine systems then use this seed corpus to produce a full translation draft. Human translators post-edit the draft, and feed new data to machines each time the newly post-edited text iteratively. In each iteration, machines produce better and better drafts with new data, and human translators find it easier and faster to post-edit. Together they complete the translation of the whole text into a severely low-resource language.

As discussed in Chapter 1, our overall goal in this thesis is to minimize human translation and post-editing efforts required to generate a full publishable-standard translation of the given text. Ideally, we want to hire a large number of human translators to measure and compare the resources (time and money) used to translate the same text into a target low-resource language that does not have any translations of the text with and without our help. However, this ideal solution is unrealistic especially in large translation projects. This is why we transform our goal of minimizing human translation efforts required to generate a full translation of the given text into two practical proxy sub-goals as the following:

1. Optimizing and minimizing the amount of sentences to be used to construct seed corpus.

[1]The material in this chapter was originally published in LoresMT at MT Summit, 2021 [320] and LoResMT at ACL, 2023 [324].

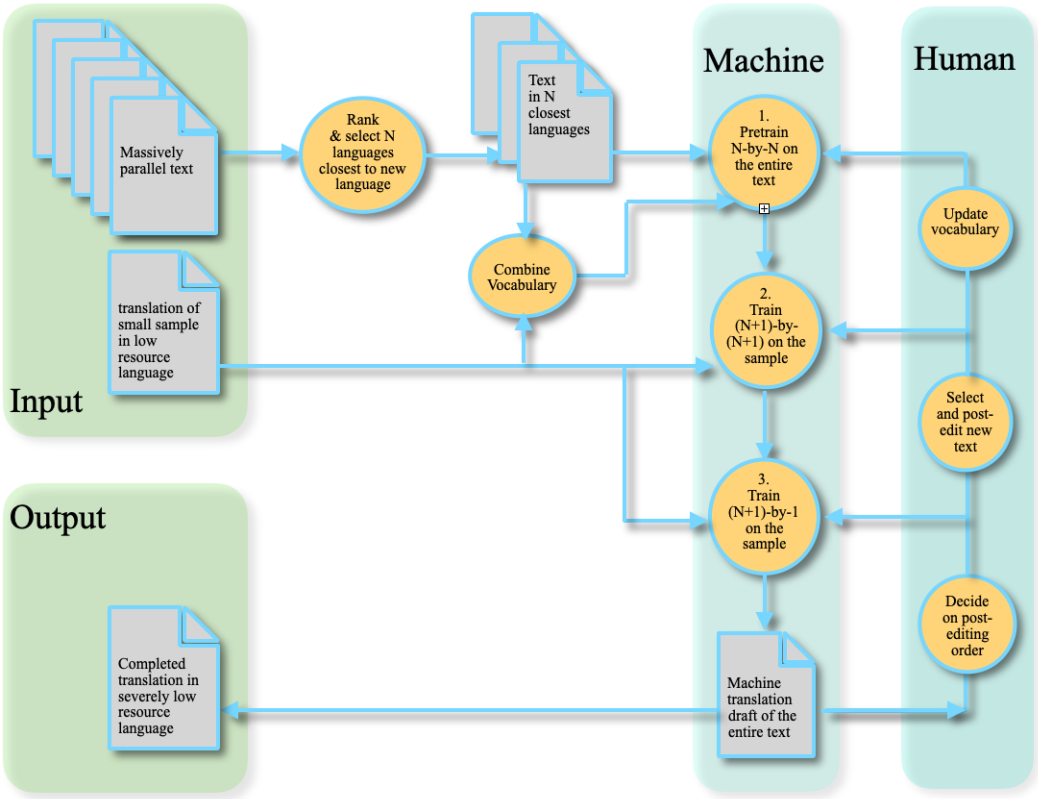


Figure 6.2: Proposed joint human machine translation sequence for a given closed text.

2. Maximizing the quality and utility of MT-generated translation of the full text and optimizing translation efficiency.

The first sub-goal of minimizing the seed corpus serves as a proxy in this chapter to minimize the human translation efforts in the creation of the seed corpus, while the second sub-goal of maximizing translation performance serves as a proxy in the next chapter to minimize human translation efforts in the post-editing process during the subsequent iterations.

With our work in the first part of the thesis, we minimize our seed corpus size to ∼3% of the text. In other words, we use ∼3% of the text to translate ∼97% of the text [320, 321, 322]. In this chapter, we are interested in finding out which ∼3% of the text is to be translated first to build a seed corpus to best improve the translation performance of the first translation draft of the machine translation system. This will help to minimize the post-editing efforts through the iterations afterwards.

To find which ∼3% of the text is to be translated first to optimize performance, we examine how this process of construction methods of such seed corpora is completed in past. Historically, this is mostly determined by field linguists' experiential and intuitive discretion. Many human translators employ a portion-based strategy when translating large texts. For

Algorithm 1: Proposed joint human machine translation sequence for a given closed text.

Input: A text of N lines consisting multiple books/portions, parallel in L source languages

Output: A full translation in the target low-resource language, l'

0. Initialize translation size, $n = 0$, vocabulary size, $v = 0$, vocabulary update size, $\triangle v = 0$;

1. Using Active Learning to choose S (∼1,000) sentences with vocabulary size v_S for human translators to produce the seed corpus, update $n = S$, $v = v_S$;

2. Rank and pick a family of close-by languages by linguistic, distortion or performance metric ;

while $n < N$ **do**

  if $\triangle v > 0$ **then**

    3. Pretrain on the full texts of neighboring languages ;

  4. Train on the n sentences of all languages in multi-source multi-target configuration ;

  5. Train on the n sentences of all languages in multi-source single-target configuration ;

  6. Combine translations from all source languages using the centeredness measure ;

  7. Review all books/portions of the translation draft ;

  8. Pick a book/portion with n' lines and v' more vocabulary ;

  9. Complete human post-editing of the portion chosen, $v = v + v'$, $n = n + n'$, $\triangle v = v'$;

return full translation co-produced by human (Step 1, 7-9) and machine (Step 0, 2-6) translation ;

example, translation of the book "The Little Prince" may be divided into smaller tasks of translating 27 chapters, or even smaller translation units like a few consecutive pages. Each translation unit contains consecutive sentences. Consequently, machine translation often uses seed corpora that are chosen based on human translators' preferences, but may not be optimal for machine translation. For optimal machine translation, researchers have yet to examine various Active Learning (AL) methods to improve accuracy and effectiveness in building better optimized seed corpora so as to minimize the initial human effort.

To solve this problem, we propose explainable and robust active learning methods that perform as well as or better than random sampling; we transfer methods learned on data of known languages to the new, severely low-resource language. Our contribution is two-fold: 1. in addition to random sampling, we develop 14 active learning methods on known languages and transfer ranking to the new, severely low-resource language; 2. we also aggregate scores

from 115 languages to provide a universal ranking and increase robustness by *relaxed memoization* method.

In this work, we aim to answer this question: when field linguists have limited time and resources, which lines would be given priority? Given a closed text, we propose that it would beneficial if field linguists translate ∼1,000 lines chosen based on active learning ranking first, getting the first machine translated draft of the whole text, and then post-edit to obtain final translation of each portion iteratively as shown in Algorithm 1. We recognize that the portion-based translation is very helpful in producing quality translation with formality, cohesion and contextual relevance. Thus, our proposed way is not to replace the portion-based approach, but instead, to get the best of both worlds and to expedite the translation process as shown in Figure 6.2.

6.1.1 Translation Workflow

In our translation workflow, human translators are informed by machine sentence ranking through active learning to produce a seed corpus. Machine systems then use this seed corpus to produce a full translation draft. Human translators post-edit the draft, and feed new data to machines each time they finish post-editing a portion of the text. In each iteration, machines produce better and better drafts with new data, and human translators find it easier and faster to post-edit. Together they complete the translation of the whole text into an severely low-resource language (Figure 6.1).

To produce sentence ranking, traditional active learning approaches assume abundant data, but we have little to no data in the target severely low-resource language. We question this assumption and build seed corpora by ranking all sentences in existing translations from other languages to generalize to a new, severely low-resource language. This ranking is target-independent as we do not require any severely low-resource language data. To produce such a ranking, we explore active learning methods including random sampling, unigram, n-gram, entropy and aggregation methods (Table 6.2). For each reference language, we build unigram, n-gram and entropy models (Figure 6.3). To prevent any language from overpowering the ranking, we aggregate sentence scores across multiple languages and rank the final aggregation. To select the pool of languages for aggregation, we build methods on different voting mechanisms.

To curate a seed corpus in the new, severely low-resource language where we have no data initially, we pass the sentence ranking learned from known languages to human translators. Human translators take this ranking, and translate the top few (∼1,000 or less) sentences, curating the seed corpus. This seed corpus is then used for training MT systems.

Once the MT systems produces the translation draft, human translators post-edit the draft. Once a stage of post-editing is finished, it is being fed back to the MT systems for its training to produce better drafts. Over many iterations of MT drafting and human

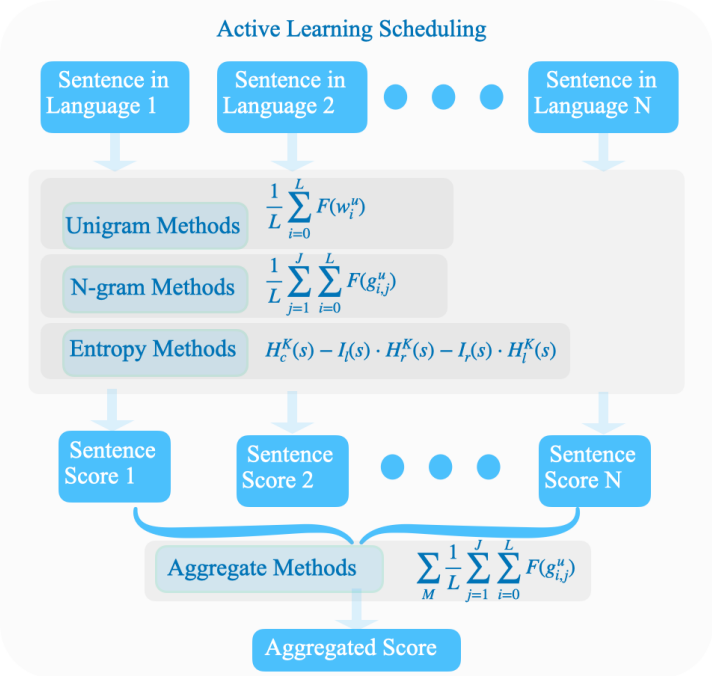


Figure 6.3: Visualizing different active learning methods. We score and rank each sentence in a text corpus.

post-editing, the final translation of the entire text into the severely low-resource language is completed.

6.1.2 Different Active Learning Approaches

Having understood our translation workflow, we are ready to examine different active learning methods. One may ask, how does our translation task into low-resource languages warrant special study given there are many existing research on active learning. To understand the reason why our translation needs warrant a special study, it is important that we understand the concept of the Knapsack Problem.

What is the Knapsack Problem? How is it relevant to translation? And how does it make active learning in our translation scenario different from existing active learning research? To answer these questions, let us define the Knapsack Problem first.

Definition 6.1.1 (Knapsack Problem)**.** Optimize the selection of items from a given set, considering their weights and values, with the aim of keeping the total weight within a given limit while maximizing the total value [151, 227].

The Knapsack Problem finds practical applications in a myriad of decision-making scenarios, spanning a wide array of fields and industries, including Machine Translation [82]. If we compare sentences in a text with items in a set, and use translation costs as weights, and

Book	Author	Books	Chapters	Pages	Languages
The Bible	Multiple	66	1,189	1,281	689
The Little Prince	Antoine de Saint Exupéry	1	27	96	382
Dao De Jing	Laozi	1	81	∼10	>250
COVID-19 Wiki Page	Multiple	1	1	∼50	155
The Alchemist	Paulo Coelho	1	2	163	70
Harry Potter	J. K. Rowling	7	199	3,407	60
The Lord of the Rings	J. R. R. Tolkien	6	62	1,037	57
Frozen Movie Script	Jennifer Lee	1	112	∼40	41
The Hand Washing Song	Multiple	1	1	1	28
Dream of the Red Chamber	Xueqin Cao	2	120	2500	23
Les Misérables	Victor Hugo	68	365	1,462	21

Table 6.1: Examples of different texts with the number of languages translated to date [58, 64, 99, 140, 168, 171, 196, 247, 281, 287, 295, 314].

use sentence scores as values, then we can define the Knapsack Problem in Machine Translation. The Knapsack Problem is NP-hard, similarly, the Knapsack Problem in Machine Translation is also NP-hard.

Definition 6.1.2 (Knapsack Problem in Machine Translation)**.** Optimize the selection of sentences from a given text, considering their translation costs and scores, with the aim of keeping the total cost within a given limit while maximizing the combined score [82].

With these definitions, we are ready to compare our active learning situations with existing active learning works in machine translation. Most of the existing research using active learning for machine translation can be framed as the Knapsack Problem. However, the difference lies mostly in what the sentence scores represent. Most of the existing work uses a coverage-based score with the goal of covering each item in vocabulary once [82, 83, 119]. This is very different when we are using active learning in translation into low-resource languages. Unlike existing work that covers the entire vocabulary, we use active learning to build a seed corpus that is as low as ∼600 sentences which only cover a small portion of the vocabulary. Our goal is to determine, with severely limited resources, which set of sentences, covering only a small subset of the vocabulary, is able to give the best translation draft in our translation efforts.

Indeed, many researchers count the number of unknown n-grams as score functions to solve the Knapsack Problem, covering all vocabulary [82, 83, 119]. Instead of solving the Knapsack Problem, we choose sentences to partially cover the vocabulary and build an extremely small seed corpus. To cover the vocabulary strategically, we sum the frequency counts of the unknown n-grams to increase density. These frequency counts promote frequent words for learning to be meaningful under extremely low-resource scenarios. In Table 6.2

we denote frequency function by $F(\cdot)$, denote sequence length by L and denote the highest n-gram order by J.

We first examine random sampling, before we introduce 14 more different active learning methods. Random sampling approach is different from the traditional portion-based approach of translation by field-linguists. The main difference of the two approaches is that the portion-based approach focuses on preserving coherence of the text locally, while the random-sampling approach focuses on increasing coverage of the text globally. Our results show that the random sampling approach performs better. When training on a seed corpus of ∼1,000 lines from the Bible and testing on the rest of the Bible (∼30,000 lines), random sampling beats the portion-based approach by +8.5 BLEU using English as a simulated low-resource language training on a family of languages built on the distortion measure, and by +1.9 using a Mayan language, Eastern Pokomchi, training on a family of languages based on the linguistic definition. Using random sampling, machine translation is able to produce a high-quality draft of the whole text that expedites the subsequent iterations of translation efforts.

In addition to random sampling, we explore 14 more different active learning methods. To improve accuracy and effectiveness in building better optimized seed corpora than randomly sampled data so as to minimize the initial human effort, we propose explainable and robust active learning methods that perform as well as or better than random sampling; we transfer methods learned on data of known languages to the new, severely low-resource language. We also examine different training schedules and we find a strategic way of growing large multilingual models in a multilingual and multi-stage fashion with extremely small severely low-resource seed corpora. Our work aims to bridge the gap between human translators and machine translation systems for them to work together better.

Moreover, we compare three different ways of incorporating incremental post-edited data during the translation process. We find that self-supervision using the whole translation draft affects performance adversely, and is best to be avoided. We also show that adding the newly post-edited text to training with vocabulary update performs the best.

6.2 Methodology

Our main goal is to translate a given text that is available in many languages to a new, severely low-resource language. In our translation workflow, we first develop active learning methods to transfer sentence ranking from known languages to a new, severely low-resource language. We then pass this ranking to human translators for them to translate the top few (∼1,000 or less) sentences into the severely low-resource language, curating the seed corpus. We finally train on the seed corpus, either from scratch or from a pretrained model.

To develop active learning methods to transfer sentence ranking from known languages to a new, severely low-resource language, we start with random sampling and comparing it with the traditional portion-based approach. In addition to random sampling, we propose and compare 14 active learning methods for machine translation into a new, severely low-resource language. To compare all active learning algorithms fairly, we use the same translation system unit as a control for all experiments, varying only the seed corpora built by different methods. We select the same number of words in all seed corpora as most translators are paid by the number of words [33, 82, 288].

6.2.1 Training Schedule

In our setup we have the new, severely low-resource language as the target language, and we have a few neighboring languages as the source languages that are either in the same linguistic language family or geographically close to facilitate linguistic transfer. In effect, we have N source languages with full translations of the text and a new and severely low-resource language that has an extremely small seed corpus.

We train our models using a state-of-the-art multilingual transformer by adding language labels to each source sentence [117, 147, 322, 323]. We borrow the order-preserving named entity translation method by replacing each named entity with `__NE`s [323] using a multilingual lexicon table that covers 124 source languages and 2,939 named entities [321]. For example, the sentence "Somchai calls Juan" is transformed to "`__opt_src_en __opt_tgt_ca __NE0` calls `__NE1`" to translate to Chuj. We use families of close-by languages constructed by ranking 124 source languages by distortion measure (*FAMD*), performance measure (*FAMP*) and linguistic family (*FAMO*$^+$); the distortion measure ranks languages by decreasing probability of zero distortion, while the performance measure incorporates an additional probability of fertility equalling one [321]. Using families constructed, we pretrain our model first on the whole text of nearby languages, then we train on the ∼1,000 lines of low-resource data and the corresponding lines in other languages in a multi-source multi-target fashion. We finally train on the ∼1,000 lines in a multi-source single-target fashion [321].

We combine translations of all source languages into one. Let all N translations be $t_i, i = 1, \ldots, N$ and let similarity between translations t_i and t_j be S_{ij}. We rank all translations according to how centered it is with respect to other sentences by summing all its similarities to the rest through $\sum_j S_{ij}$ for $i = 1, \ldots, N$. We take the most centered translation for every sentence, $\max_i \sum_j S_{ij}$, to build the combined translation output. The expectation of the combined score is higher than that of any of the source languages [321].

Name	Description	Score Function
S	Frequency sum of unknown words	$\sum_{i=0}^{L} F(w_i^u)$
SN	Normalized S by L	$\frac{1}{L}\sum_{i=0}^{L} F(w_i^u)$
SNG_J	Normalized Frequency sum of n-grams up to J	$\frac{1}{L}\sum_{j=1}^{J}\sum_{i=0}^{L} F(g_{i,j}^u)$
AGG_J^M	Aggregation of n-gram scores up to J with set M	$\sum_{M}\frac{1}{L}\sum_{j=1}^{J}\sum_{i=0}^{L} F(g_{i,j}^u)$
ENT^K	Entropy methods, K is KenLM or not	$H_c^K(s) - I_l(s)\cdot H_r^K(s) - I_r(s)\cdot H_l^K(s)$

Table 6.2: Summary of score functions.

6.2.2 Active Learning Strategies

Random Sampling

Our initial approach is random sampling to increase coverage of the text. To show the effect of random sampling, we compare it with the baseline of the linguistic baseline of the excerpt-based approach, *Luke*. The excerpt-based approach, which selects a portion of the text with consecutive sentences, preserves the text's formality, cohesion and context but lacks global coverage. Random sampling increases global coverage but sacrifices local coherence.

N-gram Approach

In addition to random sampling, we devise different active learning methods based on n-gram methods, entropy methods, and aggregation methods. To show the effectiveness of these active learning methods, we compare them with two baselines: linguistic baseline of the excerpt-based approach, *Luke*, and the statistical baseline of random sampling as our second baseline. Therefore, for the next section of active learning, we have 2 baselines: the linguistic baseline of the excerpt-based approach, *Luke*, and the statistical baseline of random sampling, *Rand*.

N-gram methods have often been used in research. Many researchers count the number of unknown n-grams as score functions to solve the Knapsack Problem, covering all vocabulary [82, 83, 119]. Instead of solving the Knapsack Problem, we choose sentences to partially cover the vocabulary and build an extremely small seed corpus. To cover the vocabulary strategically, we sum the frequency counts of the unknown n-grams to increase density. These frequency counts promote frequent words for learning to be meaningful in the extremely low-resource scenario. In Table 6.2 we denote frequency function by $F(\cdot)$, denote sequence length by L and denote the highest n-gram order by J.

Entropy Approach

Many have worked on entropy methods in modelling density and diversity [9, 82, 119, 315]. Our problem defers from previous work in that our data is much smaller, and we rely on other languages to generalize to the severely low-resource language. Therefore, we combine the existing research in combining density and diversity metrics into score functions in ranking sentences together with frequency counts[2]. Frequency counts here helps us to promote frequent words in our use case of extremely small data and partial vocabulary. We would like to cover as many words as possible, but we also would like to give priority to more frequent words for learning to be meaningful in the extremely low-resource scenario. We use traditional Language Models (LMs) instead of neural language models, as our data size is extremely small. For implementations of LMs, we use KenLM and NLTK's LM because of their simplicity and speed, especially KenLM [31, 84, 102, 120, 128]. In Table 6.2 we let $H(\cdot)$ be the cross entropy function, with the choice of KenLM (K) or NLTK (N). To separate training from testing in using language models, we divide the data into three portions, the sentences that we have chosen (c), and the remaining that are split equally into two parts, left (l) and right (r). Let $I_l(\cdot)$ and $I_r(\cdot)$ be indicator functions to show whether a sentence belongs to the left or the right. We aim to maximize the diversity H_c and optimize density by adjusting H_l and H_r [164].

Aggregation Approach

To prevent any language from overpowering the ranking, we aggregate sentence scores across different languages (Figure 6.3). We investigate the use of a customized set of languages for each severely low-resource language, versus the use of a universal set of languages representing world languages. The former requires some understanding of the neighboring languages, the latter requires careful choices of the representative set [32]. To build a universal ranking, we either aggregate over all existing languages, or create a representative pool for existing languages.

We have 4 aggregation methods: *one-vote-per-language* (L), where we aggregate over all languages, *one-vote-per-family* (F), where we aggregate over languages representing the top few families, *one-vote-per-person* (P), where we aggregate over the top few most spoken languages, and *one-vote-per-neighbor* (N), where we aggregate over a customized set of neighboring languages. For the world language distribution, L covers all, F samples across it, P covers the head, while N creates a niche area around the severely low-resource language.

[2]In the entropy score function in Table 6.2, we use highest n-gram order of 2 for NLTK's LM, we use highest n-gram order of 2 for the two halves (H_l^K and H_r^K) and order of 5 for the sampled data (H_c^K) for KenLM. Since KenLM needs at least a few words to start with, we use MLE as a warm start to select up to 5 sentences before launching KenLM.

Aggregation decreases variance and increases accuracy. Typical aggregation involve taking the sum or the average. Since they have the same effect on sentence ranking, we take the sum for simplicity.

To implement *one-vote-per-language*, we train on all available languages. that have full translations of the text , which calls for memory and time efficient algorithms. We use parallel algorithms, use hash maps for all data structures, skip disk-caching to save time and space, and reuse important data structures [7]. To save space and time, we devise *relaxed memoization*. We observe that only a few words are added to the vocabulary during each step, and therefore only a few sentences containing these new words have updated scores. At every step, we compute sentence score for each language, producing a score matrix of languages versus sentences. We update entries that are affected by the selected sentence, cache and reuse other entries. Further parallelism reduce memory consumption and results in >360 times speedup, from ∼6.5 months to ∼13 hours. Our code is efficient in memory and time.

6.2.3 Joint Human Machine Translation

Our work differs from the past research in that we put low-resource translation into the broad collaborative scheme of human machine translation. We compare the portion-based approach with the active learning approach in building seed corpora. We also compare three methods of updating models with increasing amount of human post-edited data. We add the newly post-edited data to training in three ways: with vocabulary update, without vocabulary update, or incorporating the whole translation draft in a self-supervised fashion additionally. For best performance, we build the seed corpus by active learning, update vocabulary iteratively, and add newly post-edited data to training without self-supervision. We also have a larger test set, we test on ∼30,000 lines rather than ∼678 lines from existing research [3].

We propose a joint human machine translation workflow in Algorithm 1. After pretraining on neighboring languages in Step 3, we iteratively train on the randomly sampled seed corpus of low-resource data in Step 4 and 5. The reason we include both Step 4 and 5 in our algorithm is because training both steps iteratively performs better than training either one [321]. Our model produces a translation draft of the whole text. Since the portion-based approach has the advantage with formality, cohesion and contextual relevance, human translators may pick and post-edit portion-by-portion iteratively. The newly post-edited data with updated vocabulary is added to the machine translation models without self-supervision. In this way, machine translation systems rely on quality parallel corpora that are incrementally produced by human translators. Human translators lean on machine translation for quality translation draft to expedite translation. This creates a synergistic collaboration between human and machine.

6.2.4 Evaluation Metrics

Existing multilingual systems produce multiple outputs from all source languages, rendering comparison messy. To simplify, we combine translations from all source languages into one by an existing *centeredness method* [321]. Using this method, we score each translated sentence by the sum of its similarity scores to all others, and we take the highest ranked sentence by score as our final translation.

To compare effectively, we control all test sets to be the same. Since different active learning strategies produce different seed corpora to be used as training and validation sets, the training and validation sets vary. Their complement, the test sets therefore also vary, rendering comparison difficult. To build the same test set, we devise an *intersection method*. We take the whole text and carve out all seed corpora, that is, all training and validation sets from all experiments. The remaining is the final test set, which is the intersection of all test sets.

Using our intersection method, we are able to build a common test set for all our experiments. One may ask, why not set aside a fixed amount of sentences for testing? The reason that we do not hold out a test set is because of our experiment setup. In each experiment, we use a different active learning algorithm to choose ∼3% of the text to build the seed corpus, then use this ∼3% of the text to translate ∼97% of the text. We have more than 200 experiments in this chapter, and each has a different seed corpus. In each experiment, we compare all the sentences of the text to select sentences to build a seed corpus. If we do hold out a test set for all experiments, this bars those held-out test sentences from being selected into the seed corpus for each experiment. Consequently, all seed corpora will be changed as a result of holding out a test set. This is not what we intend in our experiment design. This is why we propose the intersection method so that every experiment has access to all sentences to select into its seed corpus and we can compare them most effectively.

Our metrics are: chrF, characTER, BLEU, COMET score, and BERTscore [230, 231, 241, 275, 305, 317]. For random sampling, our metric used is mainly BLEU as we only show our result in two languages. However, for experiments on extended active learning strategies, we test on many more languages and we prioritize chrF over BLEU for better accuracy, fluency and expressive power in morphologically-rich languages [217].

6.3 Data

We show two sets of experiments: one on random sampling, the other on more extended active learning methods including n-gram methods, entropy methods, and aggregation methods.

Target	L	Family	Source Languages
Frisian	0	Germanic	English*, German, Dutch, Norwegian, Afrikaans, Swedish, French, Italian, Portuguese, Romanian
Hmong	0	Hmong–Mien	Komrem*, Vietnamese, Thai, Chinese, Myanmar, Haka, Tangsa, Zokam, Siyin, Falam
Pokomchi	0	Mayan	Chuj*, Cakchiquel, Mam, Kanjobal, Cuzco, Ayacucho, Bolivian, Huallaga, Aymara, Guajajara
Turkmen	1	Turkic	Kyrgyz*, Tuvan, Uzbek, Karakalpak, Kazakh, Azerbaijani, Japanese, Korean, Finnish, Hungarian
Sesotho	1	Niger–Congo	Yoruba*, Gikuyu, Xhosa, Kuanyama, Kpelle, Fon, Bulu, Swati, Venda, Lenje
Welsh	1	Celtic	English*, German, Danish, Dutch, Norwegian, Swedish, French, Italian, Portuguese, Romanian
Xhosa	2	Nguni	Swati*, Gikuyu, Sesotho, Yoruba, Lenje, Gbaya, Afrikaans, Wolaitta, Kuanyama, Bulu
Indonesian	3	Austronesian	Javanese*, Malagsy, Tagalog, Ilokano, Cebuano, Fijian, Sunda, Zokam, Wa, Maori
Hungarian	4	Uralic	Finnish*, French, English, German, Latin, Romanian, Swedish, Spanish, Italian, Portuguese
Spanish	5	Romance	English*, German, Danish, Dutch, Norwegian, Swedish, French, Italian, Portuguese, Romanian

Table 6.3: Summary of different target languages used [43, 59]. L, resource level, is from a scale of 0 to 5 [148]. Reference languages used for active learning methods except aggregate methods are starred.

6.3.1 Random Sampling

To first test random sampling, we work on the Bible in 124 source languages [196], and have experiments for English, a simulated language, and Eastern Pokomchi, a Mayan language. We train on ∼1,000 lines of low-resource data and on full texts for all the other languages. We aim to translate the rest of the text (∼30,000 lines) into the low-resource language. In pretraining, we use 80%, 10%, 10% split for training, validation and testing. In training, we use 3.3%, 0.2%, 96.5% split for training, validation and testing. Our test size is >29 times of the training size [3]. We use the book "Luke" for the portion-based approach as suggested by many human translators.

Training on ∼100 million parameters with Geforce RTX 2080 Ti, we employ a 6-layer encoder and a 6-layer decoder with 512 hidden states, 8 attention heads, 512 word vector size, 2,048 hidden units, 6,000 batch size, 0.1 label smoothing, 2.5 learning rate, 0.1 dropout and attention dropout, an early stopping patience of 5 after 190,000 steps, “BLEU” validation

metric, "adam" optimizer and "noam" decay method [157, 217]. We increase patience to 25 for larger data in the second stage of training in Figure 6.4a and 6.4b.

6.3.2 N-gram, Entropy and Aggregation Methods

In addition to random sampling, we are interested in comparing various n-gram methods, entropy methods and aggregation methods. To test these methods, we choose a more extensive set of target languages. When we choose target languages, we look at existing research, which classifies world languages into Resource 0 to 5, with 0 having the lowest resource and 5 having the highest [148]. We choose 10 target languages ranging from Resource 0 to 5 (Table 6.3). For each target language we choose ten neighboring languages as source languages (Table 6.3). These languages are Eastern Pokomchi, Hmong, and Frisian (Resource 0), Turkmen, Welsh and Sesotho (Resource 1), Xhosa (Resource 2), Indonesian (Resource 3), Hungarian (Resource 4), Chinese and Spanish (Resource 5) in Table 6.3. We prioritize Resource 0 to 2 languages as real low-resource languages, and we use Resource 3 to 5 languages as hypothetical ones. It is surprising to us that a lot of the Resource 0 languages are not too far away from the rich-resource languages. Frisian, for example, are spoken near the Northern Sea near Netherlands and Germany, and is in close proximity with a few rich-resource European languages [193]. However, because of the close proximity with rich-resource languages, low-resource languages like Frisian often suffer from lack of prestige and has a bigger threat to extinction as many younger people choose to speak the rich-resource languages nearby. This also suggests interesting research direction on low-resource languages and dialects that are in close proximity with rich-resource language communities.

To translate into these languages, our text is the Bible in 125 languages [196]. Each low-resource seed corpus contains ∼3% of the text, while all other languages have full text. Our goal is to translate the rest of the text into the low-resource language. In pretraining, we use a 80/10/10 split for training, validation and testing, respectively. In training, we use approximately a 3.0/0.2/96.8 split for training, validation and testing, respectively. Our training data for each experiment is ∼1,000 lines. We use BPE with size of ∼3,000 for the low-resource language and ∼9,000 for the combined [260].

Training on ∼100 million parameters with Geforce RTX 2080 Ti and RTX 3090, we use a 6-layer encoder and a 6-layer decoder with 512 hidden states, 8 attention heads, 512 word vector size, 2,048 hidden units, 6,000 batch size, 0.1 label smoothing, 2.5 learning learning rate and 1.0 finetuning learning rate, 0.1 dropout and attention dropout, a patience of 5 after 190,000 steps in $[N]^2$ with an update interval of 1000, a patience of 5 for $[N+1]^2$ with an update interval of 200, and a patience of 25 for $[N+1]$ and $[1]^2$ with an update interval of 50, "adam" optimizer and "noam" decay method [157, 217].

Input Language Family								
By Linguistics			By Distortion			By Performance		
FAMO$^+$			*FAMD*			*FAMP*		
Training	*Luke*	*Rand*	Training	*Luke*	*Rand*	Training	*Luke*	*Rand*
Testing	*Best All*	*Best All*	Testing	*Best All*	*Best All*	Testing	*Best All*	*Best All*
Combined	37.9 21.9	42.8 28.6	Combined	38.6 22.9	44.8 31.4	Combined	40.2 23.7	44.6 30.6
German	35.6 20.0	40.8 26.5	German	37.0 20.8	42.7 28.8	German	38.0 21.3	41.6 28.2
Danish	36.7 19.0	38.2 25.9	Danish	37.3 19.6	39.5 28.0	Danish	38.4 19.9	39.2 27.5
Dutch	36.4 20.4	39.7 27.2	Dutch	36.4 21.1	41.9 29.6	Dutch	37.5 21.6	41.6 28.9
Norwegian	36.5 20.2	40.0 26.9	Norwegian	37.2 20.8	41.4 29.1	Norwegian	37.5 21.1	41.0 28.4
Swedish	34.9 19.7	39.9 26.2	Afrikaans	38.3 22.2	42.8 30.5	Afrikaans	39.5 22.9	42.3 29.8
Spanish	36.8 21.5	39.8 27.6	Marshallese	35.1 21.6	41.4 28.8	Spanish	38.7 22.9	41.9 29.0
French	36.0 19.7	39.6 26.1	French	36.2 20.3	41.1 28.3	French	37.3 20.7	40.5 27.5
Italian	36.7 20.6	38.4 26.9	Italian	37.3 21.0	40.6 29.1	Italian	38.6 21.8	39.9 28.5
Portuguese	32.4 15.8	30.1 21.3	Portuguese	33.2 16.5	33.6 24.0	Portuguese	33.7 16.3	33.1 22.9
Romanian	34.9 19.3	37.1 26.0	Frisian	36.4 21.6	43.0 29.8	Frisian	37.8 22.3	42.2 29.1

Table 6.4: Performance training on 1,093 lines of English data on *FAMO*$^+$, *FAMD* and *FAMP*. We train using the portion-based approach in *Luke*, and using random sampling in *Rand*. During testing, *Best* is the book with highest BLEU score, and *All* is the performance on ∼29,000 lines of test data [3].

6.4 Results

We first look at our results on random sampling, and then we compare it with n-gram, entropy and aggregation methods through extensive experiments in multiple severely low-resource languages.

6.4.1 Random Sampling

We observe that random sampling performs better than the portion-based approach. Training on ∼3% of the text, and testing on ∼97% of the text, we see a sharp performance gain by random sampling. In Table 6.4 and 6.5, random sampling gives a performance gain of +8.5 for English on FAMD and +1.9 for Eastern Pokomchi on FAMO$^+$ [3]. The performance gain for Eastern Pokomchi may be lower because Mayan languages are morphologically

[3] In the previous chapter, we test on ∼30,000 lines excluding the ∼1,000 lines of training and validation data. In this chapter, we test on the intersection of different test sets. In Table 6.4 and 6.5, we test on ∼29,000 lines of data of the Bible excluding both the book of Luke and the randomly sampled ∼1,000 lines. In Table 6.6, we evaluate on ∼29,000 lines of data of the Bible excluding both the randomly sampled ∼1,000 lines and the book of 1 Chronicles.

Input Language Family								
By Linguistics			By Distortion			By Performance		
FAMO$^+$			*FAMD*			*FAMP*		
Training	*Luke*	*Rand*	Training	*Luke*	*Rand*	Training	*Luke*	*Rand*
Testing	*Best All*	*Best All*	Testing	*Best All*	*Best All*	Testing	*Best All*	*Best All*
Combined	23.1 8.6	19.7 10.5	Combined	23.3 8.5	17.7 9.5	Combined	22.4 7.2	15.8 7.8
Chuj	21.8 7.9	16.5 9.8	Chuj	22.0 7.9	15.4 8.9	Chuj	21.8 7.0	13.2 7.3
Cakchiquel	22.3 7.9	18.2 9.9	Cakchiquel	22.4 7.9	17.3 9.1	Cakchiquel	21.2 6.9	14.8 7.4
Guajajara	19.9 7.1	14.7 8.9	Guajajara	19.2 6.9	14.2 8.2	Guajajara	18.9 5.9	10.6 6.6
Mam	22.2 8.6	19.7 10.6	Russian	22.2 7.3	13.7 8.5	Mam	21.9 7.5	17.1 8.0
Kanjobal	21.8 8.1	17.5 10.0	Toba	22.0 8.3	16.8 9.4	Kanjobal	21.6 7.1	13.8 7.6
Cuzco	22.4 7.8	17.7 9.8	Myanmar	19.2 5.3	10.7 6.5	Thai	21.9 6.3	10.5 7.0
Ayacucho	21.6 7.6	18.5 9.7	Slovenský	22.2 7.5	13.5 8.7	Dadibi	19.9 6.2	15.3 6.9
Bolivian	22.3 7.8	17.4 9.8	Latin	22.0 7.8	14.8 9.0	Gumatj	19.2 3.8	8.9 3.3
Huallaga	22.2 7.7	18.0 9.7	Ilokano	22.6 8.4	17.8 9.4	Navajo	21.4 6.5	13.5 7.3
Aymara	21.5 7.5	18.6 9.6	Norwegian	22.6 8.3	16.7 9.4	Kim	21.6 7.0	13.9 7.5

Table 6.5: Performance training on 1,086 lines of Eastern Pokomchi data on *FAMO*$^+$, *FAMD* and *FAMP*. We train using the portion-based approach in *Luke*, and using random sampling in *Rand*. During testing, *Best* is the book with highest BLEU score, and *All* is the performance on ∼29,000 lines of test data [3].

rich, complex, isolated and opaque [5, 56, 86]. English is closely related to many languages due to colonization and globalization even though it is artificially constrained in size [30]. This may explain why Eastern Pokomchi benefits less.

In addition to evaluating all of the remaining text (∼97% of the text), we also evaluate the best book of the remaining text. The Bible has 66 books. The best performing book in our draft recommends human translators the easiest book to post-edit first.

To simulate human translation efforts in Step 7 and 8 in Algorithm 1, we rank 66 books of the Bible by BLEU scores on English's FAMD and Eastern Pokomchi's FAMO$^+$. We assume that BLEU ranking is available to us to simulate human judgment. In reality, this step is realized by human translators skimming through the translation draft and comparing performances of different books by intuition and experience. In Section 6.5, we will discuss the limitation of this assumption. Performance ranking of the simulated low-resource language may differ from that of the actual low-resource language. But the top few may coincide because of the nature of the text, independent of the language. In our results, we observe that narrative books performs better than philosophical or poetic books. The book of 1 Chronicles performs best for both English and Eastern Pokomchi with random sampling. A possible explanation is that the book of 1 Chronicles is mainly

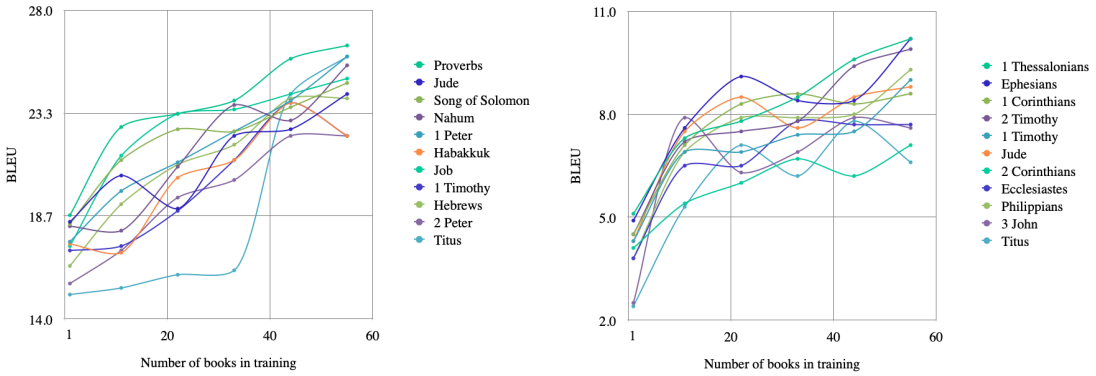
Source	*Seed*	*Self-Supervised*	*Old-Vocab*	*Updated-Vocab*
Combined	30.8	24.4 (-6.4)	32.1 (+1.3)	32.4 (+1.6)
Danish	27.7	21.6 (-6.1)	28.8 (+1.1)	29.2 (+1.5)
Norwegian	28.6	22.5 (-6.1)	29.8 (+1.2)	30.2 (+1.6)
Italian	28.7	22.3 (-6.4)	29.8 (+1.1)	30.2 (+1.5)
Afrikaans	30.1	23.8 (-6.3)	31.4 (+1.3)	31.6 (+1.5)
Dutch	29.2	22.9 (-6.3)	30.3 (+1.1)	30.6 (+1.4)
Portuguese	23.8	18.3 (-5.5)	24.6 (+0.8)	25.0 (+1.2)
French	27.8	21.7 (-6.1)	28.9 (+1.1)	29.4 (+1.6)
German	28.4	22.4 (-6.0)	29.5 (+1.1)	29.9 (+1.5)
Marshallese	28.4	22.4 (-6.0)	29.5 (+1.1)	29.9 (+1.5)
Frisian	29.3	23.2 (-6.1)	30.4 (+1.1)	30.8 (+1.5)

Table 6.6: Comparing three ways of adding the newly post-edited book of 1 Chronicles [3]. *Seed* is the baseline of training on the seed corpus alone, *Old-Vocab* skips the vocabulary update while *Updated-Vocab* has vocabulary update. *Self-Supervised* adds the complete translation draft in addition to the new book.

narrative, and contains many named entities that are translated well by the order-preserving lexiconized model. We included the BLEU scores of the best-performing book in Table 6.4 and 6.5. Note that only scores of "All" are comparable across experiments trained on the book of Luke with those trained by random sampling as they evaluate on the same set [3]. For the best-performing book, it is the book of 1 Chronicles for random sampling, and the book of Mark or the book of Matthew for experiments trained on the book of Luke. Thus, we cannot compare BLEU scores for the best-performing books across experiments. We include them in the tables to show the quality of the translation draft human translators will work on if they proceed to translate the best-performing book.

In Table 6.6, we compare three different ways of updating the machine translation models by adding a newly post-edited book that human translators produced. We call the baseline without addition of the new book *Seed*. *Updated-Vocab* adds the new book to training with updated vocabulary while *Old-Vocab* skips the vocabulary update. *Self-Supervised* adds the whole translation draft of ∼30,000 lines to pretraining in addition to the new book. Self-supervision refers to using the small seed corpus to translate the rest of the text which is subsequently used to train the model. We observe that the *Self-Supervised* performs the worst among the three. Indeed, *Self-Supervised* performs even worse than the baseline *Seed*. This shows that quality is much more important than quantity in severely low-resource translation. It is better for us not to add the whole translation draft to the pretraining as it affects performance adversely.

On the other hand, we see that both *Updated-Vocab* and *Old-Vocab* performs better than *Seed* and *Self-Supervised*. *Updated-Vocab*'s performance is better than *Old-Vocab*.



(a) English (b) Eastern Pokomchi

Figure 6.4: Performance of the most difficult 11 books with increasing number of training books.

An explanation could be that *Updated-Vocab* has more expressive power with updated vocabulary. Therefore, in our proposed algorithm, we prefers vocabulary update in each iteration. If the vocabulary has not increased, we may skip pretraining to expedite the process.

We show how the algorithm is put into practice for English and Eastern Pokomchi in Figure 6.4a and 6.4b. After producing the first draft of the text by training a MT system on the seed corpus, we hold out the most difficult 11 books (the worst-performing 11 books) and set them aside as the test set for evaluating the entire iterative post-editing process. Taking the most difficult 11 books as the held-out test set, we divide the other 55 books of the Bible into 5 portions to simulate 5 iterations of post-editing process. Using this setup, we translate the text by using the randomly sampled ∼1,000 lines of seed corpus first, and then proceed with human machine translation in Algorithm 1 in 5 iterations with increasing number of post-edited portions. Each portion contains 11 books, serving as post-edited portion for each iteration. In each iteration, we simulate human post-editing process by feeding the actual translation of the given text portion to the MT system. MT system produces better and better drafts and we show the improvement using the most difficult 11 books.

For English, we observe that philosophical books like "Proverbs" and poetry books like "Song of Solomon" perform very badly in the beginning, but begin to achieve above 20 BLEU scores after adding 11 books of training data. This reinforces our earlier result that ∼20% of the text is sufficient for achieving high-quality translation [322]. However, some books like "Titus" remains difficult to translate even after adding 33 books of training data. This shows that adding data may benefit some books more than the others. A possible explanation is that there are multiple authors of the Bible, and books differ from each other in style and content. Some books are closely related to each other, and may benefit from translations of other books. But some may be very different and benefit much less.

↑chrF	Frisian	Hmong	Pokomchi	Turkmen	Sesotho	Welsh	Xhosa	Indonesian	Hungarian	Spanish	Average
Baselines:											
+ *Luke*	47.5	41.6	39.4	34.9	41.2	41.2	32.0	43.3	34.4	46.7	40.2
+ *Rand*	50.5	43.9	42.8	38.9	43.2	46.0	34.9	47.2	37.4	50.1	43.5
Our Models:											
+ S	49.2	38.5	40.4	35.2	39.0	41.9	32.5	43.5	35.1	48.0	40.3
+ SN	50.9	43.9	43.2	38.3	41.6	43.2	36.1	46.9	36.7	50.3	43.1
+ SNG_2	53.2	**46.1**	43.3	39.5	44.4	45.8	36.6	48.4	37.8	51.8	44.7
+ SNG_3	52.7	46.0	**44.5**	39.6	**45.5**	47.5	**36.8**	48.9	**39.2**	52.3	45.3
+ SNG_4	**53.6**	45.7	44.4	**40.3**	44.9	47.7	36.8	**49.1**	39.0	**52.7**	**45.4**
+ SNG_5	53.0	45.6	43.9	39.7	45.4	46.7	36.8	49.1	38.4	52.5	45.1
+ ENT^N	50.9	43.7	38.1	37.2	42.5	44.5	34.7	46.7	36.0	49.9	42.4
+ ENT^K	52.7	45.7	43.5	40.2	44.6	45.2	36.4	49.0	39.1	51.8	44.8
+ AGG_5^L	47.1	41.5	39.8	34.0	39.9	42.1	31.4	43.5	33.7	45.2	39.8
+ AGG_5^F	45.0	38.4	38.5	32.4	38.8	47.1	30.4	41.2	33.3	44.2	38.9
+ AGG_5^P	45.5	38.8	38.0	32.0	38.8	**48.2**	30.5	41.0	33.2	44.0	39.0
+ AGG_5^N	45.4	39.1	38.3	32.4	38.8	48.0	30.7	41.2	33.2	44.3	39.1

Table 6.7: 140 experiments comparing 14 active learning methods translating into 10 different languages with Schedule *B*.

For Eastern Pokomchi, though the performance of the most difficult 11 books never reach BLEU score of 20s like that of English experiments, all books have BLEU scores that are steadily increasing. Challenges remain for Eastern Pokomchi, a Resource 0 language [148]. We hope to work with native Mayan speakers to see ways we may improve the results.

6.4.2 N-gram, Entropy, and Aggregation Methods

In addition to active learning, we compare 14 active learning methods across 10 different target languages (Table 6.7).

Normalizing by sequence length improves density: Without normalization, the model chooses longer sentences with many rare words. Normalization improves density. For Sesotho, the chrF score is 39.0 without normalization and 41.6 with it.

Marginal benefit of increasing n-gram order wanes: Existing research shows bigrams suffice [82]. As the n-gram order increases, the data gets sparser and the marginal benefit subsides. Hmong has the best score (46.1) using bigrams.

Tipping points vary with language: The optimal highest n-gram order may differ from language to language. 4-grams work best for Frisian while bigrams work best for Hmong. Hmong is an isolating language while Frisian is a fusional language. A possible explanation is that higher n-grams may have more impact on fusional languages.

Entropy and n-gram methods both beat baselines and higher n-gram models perform best: KenLM is much faster and performs better than NLTK. The entropy method using KenLM beats both baselines. Frisian has a chrF score of 52.7 with the entropy method using KenLM. This is much higher than the baselines: *Luke* (47.5) and *Rand* (50.5).

The 4-gram model (53.6) is higher because building LMs from a few lines of data may not be accurate. Simpler n-gram models work better than more evolved entropy models with small data.

Aggregation over all languages serves as a universal ranking: The first 10 active learning methods are based on learning from one reference language and generalizing to the low-resource language, while the last 4 focus on aggregation over multiple languages (Table 6.7). For Welsh, aggregation over multiple languages (48.2 with most spoken languages) performs better than those that rely on one reference language; but for other languages aggregation performs worse. Aggregation over all languages performs better than other aggregation methods for all languages except Welsh. This hinges on the reference language. For Frisian, choosing English (a Germanic language) as a reference language, performs better than aggregation. For Welsh (a Celtic language), choosing a reference language that is not as close, performs worse. But we often do not have such information for the low-resource languages. In such cases, universal ranking by aggregating over all languages is useful.

We recommend n-gram methods as the most efficient in computational cost: Among all of our active learning methods, random sampling does not require much computation, and therefore costs the least in terms of computational time. However, random sampling does not perform as well as n-gram models. Comparing n-gram models with entropy models and aggregation models, n-gram models is the most effective, costing the least in terms of computational time, and performs the best. For example, a 4-gram active learning model takes around 38 minutes to train, while an entropy model typically takes 14 hours to train. Since aggregation methods by design take multiple reference languages, they take longer time than than any models based on a single reference language. This is the reason we highly recommend n-gram methods as the most efficient and the least expensive.

Our active learning methods mimic curriculum learning: Our models pick short and simple sentences first, emulating curriculum learning and helping human translators [24, 110, 146].

All active learning methods cover different genres: Our methods pick a mix of sentences from different genres, sentence lengths and complexity levels. Moreover, our methods pick narrative sentences first, which is helpful for human translators.

Our model captures some language subtleties: Apart from the metrics, we showed our translation to native speakers (Table 6.8). We translate "He sees that it is good" to "lug ca rua huv nwg lu sab" ("He puts it in the liver") in Hmong, which uses liver to express joy. This increases lexical choice.

6.5 Conclusion

To serve our main goal of minimizing human translation and post-editing efforts, our contribution in this chapter is that we minimize the seed corpus to be ∼3% of the text. Indeed, we use ∼3% of the text to translate the ∼97% of the text in the low-resource language we want to translate to. Having minimized the training data to be ∼3% of the text in the given low-resource language, we optimize this process with active learning on which ∼3% of the text to translate first to produce the seed corpus. To determine which ∼3% of the text, we show results from random sampling, as well as n-gram, entropy, and aggregation methods in addition to the portion-based approach to build seed corpus without any low-resource language data. Of all these methods, we find that the n-gram method (in particular, the 4-gram method) is sufficient for producing high translation performance when we are given complete information of languages close to the given low-resource language. However, when we are not given complete information of the close-by languages, we recommend the aggregation method proposed in this chapter as a universal ranking to use. This minimizes human translation efforts in the production of the seed corpus. Additionally, we also compare three different ways of updating the machine translation models by adding newly post-edited data iteratively. We find that vocabulary update is necessary, but self-supervision by pretraining with whole translation draft is best to be avoided.

However, we still face challenges with the lack of local coherence and context. The excerpt-based approach enjoys advantages with formality, cohesion and contextual relevance. Active learning methods, on the contrary, do not have consecutive sentences and therefore lose local coherence and pose challenges to human translators [66, 195, 202, 248, 273, 309, 320]. Human translators may not be receptive to active learning methods like random sampling. Additionally, it may take longer for human translators to translate a set of non-consecutive sentences, because each sentence may be from a different chapter with a different context and it is harder for human translators to recycle translations. Indeed, the lack of local coherence and global context is a real challenge when human translators work with machine translation systems. Moreover, improving local coherence is important and is also an active research area.

One limitation of our work is that in real life scenarios, we do not have the reference text in low-resource languages to produce the BLEU scores to decide the post-editing order. Consequently, field linguists need to skim through and decide the post-editing order based on intuition. However, computational models can still help. One potential way to tackle it is that we can train on ∼1,000 lines from another language with available text and test on the 66 books. Since our results show that the literary genre plays important role in the performance ranking, it would be reasonable to determine the order using a "held-out language" and then using that to determine order in the target low-resource language.

In the future, we would like to work with human translators who understand and speak low-resource languages.

Another concern human translators may have is the creation of randomly sampled seed corpora. To gauge the amount of interest or inertia, we have interviewed some human translators and many are interested. However, it is unclear whether human translation quality of randomly sampled data differs from that of the traditional portion-based approach. We hope to work with human translators closely to determine whether the translation quality difference is manageable.

We are also curious how our model will perform with large literary works like “Lord of the Rings” and "Les Misérables". We would like to see whether it will translate well with philosophical depth and literary complexity. However, these books often have copyright issues and are not as easily available as the Bible data. We are interested in collaboration with teams who have multilingual data for large texts, especially multilingual COVID-19 data.

Having examined various active learning methods in building seed corpora that optimize machine translation, we have established the first step of the human machine translation workflow as show in Figure 6.1. In the next chapter, we will focus on the next few steps in the workflow by focusing on how to use large multilingual models to most effectively train on such a small seed corpus in the low-resource language.

Target	System Translation	Reference
Frisian	mar Ruth sei: Ik scil dy net forlitte, en ik scil fen dy net weromkomme; hwent hwer "tstû hinnegeane, den scil ik hinnegean, en dêr scil ik dy fornachtsje. dyn folk is myn folk, en dyn God is myn God.	mar Ruth sei: Sit net tsjin my oan, dat ik jo forlitte en weromtsjen scil; hwent hwer "t jo hinne geane, dêr scil ik hinne gean, en hwer "t jo fornachtsje, dêr scil ik fornachtsje; jins folk is myn folk en jins God is myn God;
Hmong	Lauj has rua nwg tas, "Tsw xob ua le ntawd, kuv yuav moog rua koj lub chaw kws koj moog, hab kuv yuav nyob huv koj haiv tuabneeg. koj yog kuv tug Vaajtswv."	tassws Luv has tas, "Tsw xob has kuas kuv tso koj tseg ncaim koj rov qaab moog. koj moog hovtwg los kuv yuav moog hab, koj nyob hovtwg los kuv yuav nyob hov ntawd hab, koj haiv tuabneeg los yog kuv haiv tuabneeg hab, koj tug Vaajtswv los yog kuv tug Vaajtswv.
Pokomchi	eh je' wili i xq'orarik reh i Rut: Maacanaa' chih taj i hin. re' hin naa nub'anam aweh chupaam i ye'aab' naa nuk'achariik ayu'. re' hin naa nuk'achariik awuuk', eh re' hin naa nukahniik chi nuDios, inki.	re' Rut je' wili i chaq'wik xub'an: Maa pahqaaj aakuyariik weh re' hin ma' jaruuj nee tinukanaa' kahnoq, xa aha' pa' nee tiooj i hat, nee wo' kinooj chawiij, xa aha' pa' nee ti k'achariik i hat ar nee kink'acharik i hin. eh re' aatinamiit re' wo' re' nutinamiit i hin, eh re' aaDios re' wo' re' nuDios i hin.
Turkmen	Rut: oňa: "Sen nirä gitseň, men hem seniň ýanyňa gitmerin. Sen nirä gitseň, men hem seniň halkym bolaryn. Men seniň Hudaýym bolaryn.	emma Rut: "Seni terk edip ýanyňdan gitmegi menden haýyş etme. sen Nirä gitseň, Menem şol ýere gitjek. sen nirede bolsaň, Menem şol ýerde boljak. seniň halkyň - meniň halkym, seniň Hudaýyň meniň Hudaýym bolar.
Sesotho	yaba Ruthe o re ho yena: "O se ke wa tloha ho wena, hobane ke tla ya le wena, ke tla ya le wena, mme ke tla ya hona moo. setjhaba sa ka, le Modimo wa hao."	empa Ruthe a re: "O se ke wa nqobella hore ke kgaohane le wena, kapa hore ke se ke ka tsamaya le wena, hobane" moo o yang teng ke tla ya teng, moo o phelang teng ke tla phela teng; tjhaba sa heno e be tjhaba sa heso, Modimo wa hao e be Modimo wa ka.
Welsh	a Ruth a ddywedodd, Nuw gael arnaf fi, atolwg, atolwg, oddi wrthyt: canys lle yr wyt yn myned, ac yno yr wyt yn myned, y byddaf fy hun. dy bobl yw fy bobl, a'th Dduw yw fy Duw.	a Ruth a ddywedodd, Nac erfyn arnaf fi ymado â thi, i gilio oddi ar dy ôl di: canys pa le bynnag yr elych di, yr af finnau; ac ym mha le bynnag y lletyech di, y lletyaf finnau: dy bobl di fydd fy mhobl i, a'th Dduw di fy Nuw innau:
Xhosa	URute waphendula wathi: "Undiyekeli ukuba ndixhamle, kuba ndiza kuhlala apho uthanda khona. mna ndiza kuba ngabantu bam, abe nguThixo wam."	Waphendula uRute wathi: "Sukundinyanzela usithi mandikushiye. apho uya khona, nam ndiya kuya, ndiye kuhlala nalapho uhlala khona, amawenu abe ngamawethu, noThixo wakho abe nguThixo wam.
Indonesian	tetapi Rut: menjawab: "Janganlah engkau meninggalkan aku dan pulang ke tempat kediamanmu, sebab aku akan pergi dan berdiam di mana engkau diam, sebab orang-orangmu akan menjadi umat-Ku dan Allahmu."	tetapi kata Rut: "Janganlah desak aku meninggalkan engkau dan pulang dengan tidak mengikuti engkau; sebab ke mana engkau pergi, ke situ jugalah aku pergi, dan di mana engkau bermalam, di situ jugalah aku bermalam: bangsamulah bangsaku dan Allahmulah Allahku;
Hungarian	Ruth így felelt: Nem kérlek téged, hogy gondolj meg téged, mert csak hozzád megyek, és én otthagytam, hogy legyenek hozzád. a te népem az én, és az én Istenem az én.	de Ruth azt felelte: Ne unszolj engem, hogy elhagyjalak és visszatérjek tőled. mert ahová te mégy, odamegyek, ahol te megszállsz, ott szállok meg. Néped az én népem, és Istened az én Istenem.
Spanish	y Rut: dijo a David: No me permite de ti, y me quitaré de ti; porque donde vayas, yo iré a donde vayas, y habitaré; y tu pueblo es mi pueblo, y tu Dios es mi Dios.	respondió Rut: No me ruegues que te deje, y me aparte de ti; porque a dondequiera que tú fueres, iré yo, y dondequiera que vivieres, viviré. tu pueblo será mi pueblo, y tu Dios mi Dios.

Table 6.8: Qualitative evaluation using SNG_5 to translate into each target language.

Chapter 7

Optimizing with Large Pretrained Models

> "The deepest connection you have with someone and their culture, is through learning their language."
>
> *Marisa J Taylor*

Having examined various active learning methods in building a seed corpus that optimizes for machine translation, we are interested to increase effectiveness of how to best train on such a small seed corpus. To serve our main goal of minimizing human translation and post-editing efforts, having minimized the seed corpus to ∼3% of the text, and optimized which ∼3% of the text is to be translated first, we are interested in how to best train on such a small sample of the text and produce the best translation performance to minimize post-editing efforts. In this chapter, we will show how to most effectively train using large pretrained models to minimize human post-editing efforts. We find it most useful to adapt the large pretrained models to the domain first, and then to the target low-resource language.

7.1 Introduction

A language dies when no one speaks it. A language needs attention when it is spoken by enough people that it could survive under favorable conditions but few or no children are learning it [62, 154, 313]. More than half of the 7,139 languages will die in the next 80 years [15, 81]. Many of these languages are low-resource. These languages may survive and thrive if they gain prestige, power and visibility [62]. Frisian, for example, struggles

Figure 7.1: A community in Peru that speaks Panao Quechua. Photograph provided by Mark Bean.

to gain prestige in Germany, and is low-resource even though it has a large number of speakers. Hebrew, conversely, has been revived as a spoken language because it is critical to the development and identity of the Jewish community. We empower these language communities by exercising a language[1]. This can be achieved by translating important texts to their language so that these communities can gain information, knowledge, power and visibility in their own language.

The problem in these scenarios, therefore, is not to build a high accuracy translation engine for *any texts* using huge data corpora, but rather to build a good translation for a *known* text (for which translations in many other languages exist), but in a new language with only extremely little seed data (a few hundred sentences). We assume there is little to no low-resource language data and few human translators. To produce high quality translation, existing methods rely on a seed corpus produced by human translators. Previous work has shown progress in using extremely small seed corpora with as small as ∼1,000 lines of data and has found that active learning performs better than choosing a fixed portion of the text to build a seed corpus [177, 234, 321]. However, researchers have yet to completely solve the problem of using large multilingual models for representational learning to train

[1]The material in this chapter was originally published in LoResMT at ACL, 2023 [324].

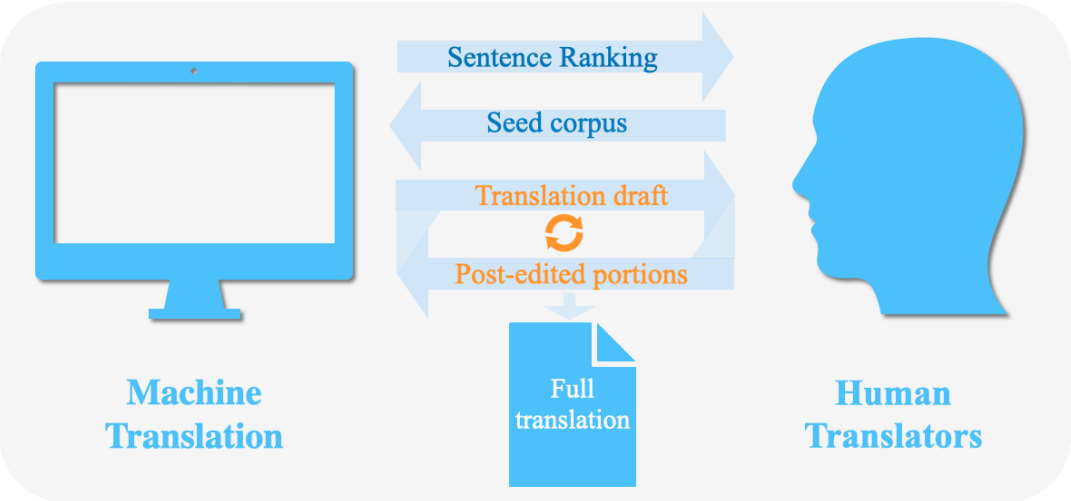


Figure 7.2: Translation workflow for low-resource languages, focusing on training on the seed corpus followed by iterations of post-editing and updated training.

(or adapt) them to a new, low-resource language by training on an extremely small seed corpus.

To solve this problem, we examine different training schedules and we find a strategic way of growing large multilingual models in a multilingual and multi-stage fashion with extremely small low-resource seed corpora.

In our translation workflow, human translators are informed by machine sentence ranking to produce a seed corpus. To curate a seed corpus in the new, low-resource language where we have no data initially, we pass the sentence ranking learned from known languages to human translators. Human translators take this ranking, and translate the top few (∼1,000 or less) sentences, curating the seed corpus.

Using the seed corpus created by active learning, machine systems then train on the seed corpus to produce a full translation draft. Human translators post-edit the draft, and feed new data to machines each time they finish post-editing a portion of the text. In each iteration, machines produce better and better drafts with new data, and human translators find it easier and faster to post-edit. Together they complete the translation of the whole text into an low-resource language (Figure 7.2).

To train on such small seed corpus, we find pretraining to be key. For the pretrained model, we either create our own pretrained model by training on known languages, or use an existing pretrained model. We explore both paths in our work, with and without activating the knowledge in existing large pretrained models. We observe an average increase of 28.8 in chrF score over the baselines.

Our contribution is therefore: we activate the knowledge of large multilingual models by proposing multilingual and multi-stage adaptations through 24 different training schedules;

Figure 7.3: 24 different training schedules.
[N]: multilingual model on N neighboring languages
$[N+1]^2$: multi-target model with low-resource language
[N+1]: single-target model with low-resource language
$[1]^2$: autoencoder in low-resource language.

we find that adapting pretrained models to the domain and then to the low-resource language works best. Our model can stand on its own and can also be boosted by large multilingual models. Our model works on many languages spanning different resource levels.

7.2 Methods

We translate a fixed text that is available in many languages to a new, low-resource language. In our translation workflow, we first develop active learning methods to transfer sentence ranking from known languages to a new, low-resource language. We then pass this ranking to human translators for them to translate the top few (∼1,000 or less) sentences into the low-resource language, curating the seed corpus. We finally train on the seed corpus, either from scratch or from a pretrained model.

We build training schedules on an extremely small seed corpus. We propose and compare 24 training schedules for machine translation into a new, low-resource language. To compare all experiments fairly, we use the same translation system unit as a control for all experiments, varying only the seed corpora built by different methods. We select the same number of words in all seed corpora as most translators are paid by the number of words they translate [33, 82, 288].

7.2.1 Training Schedules

In our setup we have the new, low-resource language as the target language, and we have a few neighboring languages as the source languages that are either in the same linguistic language family or geographically close to facilitate linguistic transfer. In effect, we have N source languages with full translations of the text and a new and low-resource language that has an extremely small seed corpus, which could be a few hundred lines of data.

We use the state-of-the-art multilingual transformer prepending both source and target language labels to each source sentence [117, 147]. For precise translation for all named entities, we use an existing method of *order-preserving named entity translation* by masking each named entity with ordered `__NE`s using a parallel multilingual lexicon table in 125 languages [311, 321].

Using this multilingual transformer architecture as a base, we build 5 training units on the small seed corpus of the new, low-resource language and the existing translations of known languages. We let $[N]^2$ denote the training of all source languages in a N-by-N multilingual transformer. We let $[N+1]^2$ denote the training of all languages including the low-resource language in a (N+1)-by-(N+1) multilingual transformer. We let [N+1] denote the (N+1)-by-1 multilingual transformer that focuses on translating into the low-resource language. We let $[1]^2$ be the autoencoder on the low-resource language.

Our translation system is built on these 5 training units: an optional [M2M100] [91], $[N]^2$, $[N+1]^2$, [N+1] and $[1]^2$. These 5 stages increase in specificity while they decrease in data size. Building on them, we show 24 different training schedules, among which 8 are pretrained with in-domain data and 16 are pretrained with out-of-domain large multilingual models (Figure 7.3). We only consider models with pretraining and therefore do not exhaust all 32 training schedules.

7.2.2 Active Learning Strategies

We have two baselines: the linguistic baseline of the excerpt-based approach, *Luke*, and the statistical baseline of random sampling, *Rand*. The excerpt-based approach, which selects a portion of the text with consecutive sentences, preserves the text's formality, cohesion and context but lacks global coverage. Random sampling increases global coverage but sacrifices local coherence.

In addition to random sampling, we also explore n-gram, entropy and aggregation methods as introduced in the previous chapter.

7.2.3 Evaluation Method and Metrics

Existing multilingual systems produce multiple outputs from all source languages, rendering comparison messy. To simplify, we combine translations from all source languages into one by an existing *centeredness method* [321]. Using this method, we score each translated sentence by the sum of its similarity scores to all others. We rank these scores and take the highest score as our combined score. The expected value of the combined score is higher than that of each source.

To compare effectively, we control all test sets to be the same. Since different experiments use different seed corpora as training and validation sets, the training and validation sets vary. Their complement, the test sets therefore also vary, rendering comparison difficult. To build the same test set, we devise an *intersection method*. We take the whole text and carve out all seed corpora, that is, all training and validation sets from all experiments. The remaining is the final test set, which is the intersection of all test sets across all experiments.

Target	L	Family	Source Languages
Frisian	0	Germanic	English*, German, Dutch, Norwegian, Afrikaans, Swedish, French, Italian, Portuguese, Romanian
Hmong	0	Hmong–Mien	Komrem*, Vietnamese, Thai, Chinese, Myanmar, Haka, Tangsa, Zokam, Siyin, Falam
Pokomchi	0	Mayan	Chuj*, Cakchiquel, Mam, Kanjobal, Cuzco, Ayacucho, Bolivian, Huallaga, Aymara, Guajajara
Turkmen	1	Turkic	Kyrgyz*, Tuvan, Uzbek, Karakalpak, Kazakh, Azerbaijani, Japanese, Korean, Finnish, Hungarian
Sesotho	1	Niger–Congo	Yoruba*, Gikuyu, Xhosa, Kuanyama, Kpelle, Fon, Bulu, Swati, Venda, Lenje
Welsh	1	Celtic	English*, German, Danish, Dutch, Norwegian, Swedish, French, Italian, Portuguese, Romanian
Xhosa	2	Nguni	Swati*, Gikuyu, Sesotho, Yoruba, Lenje, Gbaya, Afrikaans, Wolaitta, Kuanyama, Bulu
Indonesian	3	Austronesian	Javanese*, Malagsy, Tagalog, Ilokano, Cebuano, Fijian, Sunda, Zokam, Wa, Maori
Hungarian	4	Uralic	Finnish*, French, English, German, Latin, Romanian, Swedish, Spanish, Italian, Portuguese
Spanish	5	Romance	English*, German, Danish, Dutch, Norwegian, Swedish, French, Italian, Portuguese, Romanian

Table 7.1: Summary of different target languages used [43, 59]. L, resource level, is from a scale of 0 to 5 [148]. Reference languages used for active learning methods except aggregate methods are starred.

Our metrics are: chrF, characTER, BLEU, COMET score, and BERTscore [230, 231, 241, 275, 305, 317]. We prioritize chrF over BLEU for better accuracy, fluency and expressive power in morphologically-rich languages [217].

7.3 Data

Existing research classifies world languages into Resource 0 to 5, with 0 having the lowest resource and 5 having the highest [148]. We choose 10 target languages ranging from Resource 0 to 5 (Table 7.1). For each target language we choose ten neighboring languages as source languages (Table 7.1). These languages[2] are Eastern Pokomchi, Hmong, and Frisian (Resource 0), Turkmen, Welsh and Sesotho (Resource 1), Xhosa (Resource 2), Indonesian

[2]For simplicity, in Table 7.1 Pokomchi is Eastern Pokomchi, Hmong is Hmong Hoa, Kanjobal is Eastern Kanjobal, Mam is Northern Mam, Cuzco is Cuzco Quechua, Ayacucho is Ayacucho Quechua, Bolivian is

↑chrF	Frisian	Hmong	Pokomchi	Turkmen	Sesotho	Welsh	Xhosa	Indonesian	Hungarian	Spanish	Average
Baselines:											
+ Bilingual	23.1	25.0	28.7	18.9	25.2	22.2	21.4	27.2	20.1	22.1	23.4
+ Multilingual	28.0	28.1	31.9	22.6	28.3	26.5	23.9	29.7	22.3	26.8	26.8
Our Models:											
+ Schedule *B*	50.5	43.9	42.8	38.9	43.2	46.0	34.9	47.2	37.4	50.1	43.5
+ Active (AL)	53.6	45.7	44.4	40.3	44.9	47.7	36.8	49.1	39.0	52.7	45.4

Table 7.2: Results for translation into 10 languages that are new and severely low-resource to the system, independent of M2M100.

↑chrF	Frisian	Welsh	Hungarian	Spanish	Average
Baselines:					
+ Bilingual	23.1	22.2	20.1	22.1	21.9
+ Multilingual	28.0	26.5	22.3	26.8	25.9
+ M2M100	26.0	9.9	38.8	47.5	24.9
Our Models:					
+ Schedule *I*	53.5	49.5	42.2	53.2	49.6
+ Active (AL)	54.9	49.8	43.2	54.9	50.7

Table 7.3: Results for translation into 4 languages that are new and severely low-resource to the system, activating knowledge in M2M100 and leveraging active learning.

(Resource 3), Hungarian (Resource 4), Chinese and Spanish (Resource 5) in Table 7.1. We prioritize Resource 0 to 2 languages as real low-resource languages, and we use Resource 3 to 5 languages as hypothetical ones to show the spectrum. It is surprising to us that a lot of the Resource 0 languages are not too far away from the rich-resource languages. Frisian, for example, are spoken near the Northern Sea near Netherlands and Germany, and is in close proximity with a few rich-resource European languages [193]. However, because of the close proximity with rich-resource languages, low-resource languages like Frisian often suffer from lack of prestige and has a bigger threat to extinction as many younger people choose to speak the rich-resource languages nearby. This also suggests interesting research direction on low-resource languages and dialects that are in close proximity with rich-resource language communities.

South Bolivian Quechua, and Huallaga is Huallaga Quechua, Chinese is Traditional Chinese, Haka is Haka Chin, Siyin is Siyin Chin, Falam is Falam Chin, Kpelle is Kpelle Guinea.

Network	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
[M2M100]	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓
$[N]^2$	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓								
$[N+1]^2$	⇓	⇓	⇓	⇓					⇓	⇓	⇓	⇓				
[N+1]	⇓	⇓			⇓	⇓			⇓	⇓			⇓	⇓		
$[1]^2$	⇓		⇓		⇓		⇓		⇓		⇓		⇓		⇓	
↑chrF	**52.9**	**51.8**	49.5	**52.8**	**52.7**	**51.9**	27.4	16.9	49.6	48.5	39.6	48.7	48.5	45.7	27.8	26.3
↓cTER	0.492	0.508	0.482	0.488	0.493	0.502	0.654	0.800	0.530	0.546	0.553	0.539	0.538	0.579	0.650	0.667
↑BLEU	28.8	27.9	24.2	28.9	28.8	28.2	3.0	0.6	24.8	24.2	13.9	24.3	24.5	22.0	3.4	3.3
↑COMET	-0.56	-0.59	-0.63	-0.53	-0.56	-0.57	-1.28	-1.75	-0.67	-0.70	-0.89	-0.68	-0.69	-0.80	-1.21	-1.30
↑BERTS	0.891	0.889	0.886	0.892	0.891	0.890	0.813	0.775	0.883	0.881	0.861	0.882	0.880	0.873	0.823	0.819

Table 7.4: Comparing 16 training schedules with M2M100. BERTS is BERTScore, cTER is characTER and LRatio is length ratio.

Among these ten languages, Frisian, Welsh, Hungarian and Spanish are on the list of languages trained in M2M100 [91]. We apply a multi-stage adaptation of large pretrained models to these four languages. During training, we use the large sentence-piece vocabulary from M2M100 which contains 128k unique tokens for optimized performance.

To translate into these languages, our text is the Bible in 125 languages [196]. Each low-resource seed corpus contains ∼3% of the text, while all other languages have full text. Our goal is to translate the rest of the text into the low-resource language. In pretraining, we use a 80/10/10 split for training, validation and testing, respectively. In training, we use approximately a 3.0/0.2/96.8 split for training, validation and testing, respectively. Our training data for each experiment is ∼1,000 lines. We use BPE with size of ∼3,000 for the low-resource language and ∼9,000 for the combined [260].

Training on ∼100 million parameters with Geforce RTX 2080 Ti and RTX 3090, we use a 6-layer encoder and a 6-layer decoder with 512 hidden states, 8 attention heads, 512 word vector size, 2,048 hidden units, 6,000 batch size, 0.1 label smoothing, 2.5 learning learning rate and 1.0 finetuning learning rate, 0.1 dropout and attention dropout, a patience of 5 after 190,000 steps in $[N]^2$ with an update interval of 1000, a patience of 5 for $[N+1]^2$ with an update interval of 200, and a patience of 25 for [N+1] and $[1]^2$ with an update interval of 50, "adam" optimizer and "noam" decay method [157, 217].

For finetuning from a M2M100 Model, training on ∼418 million parameters with Geforce RTX 3090, we use a 12-layer encoder and a 12-layer decoder with 1024 hidden states, 16 attention heads, 1024 word vector size, 4,096 hidden units, 0.2 label smoothing, 0.0002 training learning rate and finetuning 0.00005 learning rate, 0.1 dropout and attention dropout, a patience of 10, "BLEU" validation metric, "adam" optimizer and "noam" decay method [85, 91, 254].

Network	A	B	C	D	E	F	G	H
$[N]^2$	⇓	⇓	⇓	⇓	⇓	⇓	⇓	⇓
$[N+1]^2$	⇓	⇓	⇓	⇓				
[N+1]	⇓	⇓			⇓	⇓		
$[1]^2$	⇓		⇓		⇓		⇓	
↑chrF	38.7	**51.1**	35.6	50.8	43.4	**51.2**	25.6	24.1
↓cTER	0.555	0.517	0.572	0.515	0.523	0.507	0.650	0.682
↑BLEU	12.5	24.9	9.2	24.5	17.5	26.2	2.5	2.1
↑COMET	-0.87	-0.66	-0.91	-0.65	-0.81	-0.63	-0.99	-1.02
↑BERTS	0.850	0.882	0.839	0.884	0.865	0.885	0.801	0.794

Table 7.5: Comparing 8 training schedules without M2M100.
$[N]^2$: multilingual model on N neighboring languages
$[N+1]^2$: multi-target model with low-resource language
[N+1]: single-target model with low-resource language
$[1]^2$: autoencoder in low-resource language.

7.4 Results

For simplicity, we use the centeredness method to combine translations from all source languages and have one score per metric. To compare across different methods, all experiments have the same test set (3,461 lines), the intersection of all test sets.

Our models improve over the baselines: With Schedule *I*, we observe an average improvement of 24.7 in chrF score over the M2M100 baseline (Table 7.3). By active learning with 4-gram model, we observe an increase of 28.8 in chrF score over the bilingual baseline.

Our strategic training schedule improves the translation further by activating the knowledge of M2M100 : With Schedule *B* and the 4-gram model, we observe an average improvement of 18.6 in chrF score over the multilingual baseline (Table 7.2). For Schedule *I*, the increase is 24.8 over the multilingual baseline (Table 7.3). Indeed, the increase with the activation of M2M100 is greater. This shows that our strategic schedules[3] improve translation performance by activating the knowledge of M2M100.

7.4.1 Training Schedules

We compare 24 training schedules using a randomly sampled seed corpus (∼1,000 lines) to translate into Frisian (Table 7.4 and 7.5).

[3]In Table 7.2, our model with training scheduling uses Schedule *B*, our model with active learning uses SNG_4. In Table 7.3, our model with training scheduling uses Schedule *I*, our model with active learning uses SNG_4.

↑chrF	Frisian	Hmong	Pokomchi	Turkmen	Sesotho	Welsh	Xhosa	Indonesian	Hungarian	Spanish	Average
Baselines:											
+ *Luke*	47.5	41.6	39.4	34.9	41.2	41.2	32.0	43.3	34.4	46.7	40.2
+ *Rand*	50.5	43.9	42.8	38.9	43.2	46.0	34.9	47.2	37.4	50.1	43.5
Our Models:											
+ S	49.2	38.5	40.4	35.2	39.0	41.9	32.5	43.5	35.1	48.0	40.3
+ SN	50.9	43.9	43.2	38.3	41.6	43.2	36.1	46.9	36.7	50.3	43.1
+ SNG_2	53.2	**46.1**	43.3	39.5	44.4	45.8	36.6	48.4	37.8	51.8	44.7
+ SNG_3	52.7	46.0	**44.5**	39.6	**45.5**	47.5	**36.8**	48.9	**39.2**	52.3	45.3
+ SNG_4	**53.6**	45.7	44.4	**40.3**	44.9	47.7	36.8	**49.1**	39.0	**52.7**	**45.4**
+ SNG_5	53.0	45.6	43.9	39.7	45.4	46.7	36.8	49.1	38.4	52.5	45.1
+ ENT^N	50.9	43.7	38.1	37.2	42.5	44.5	34.7	46.7	36.0	49.9	42.4
+ ENT^K	52.7	45.7	43.5	40.2	44.6	45.2	36.4	49.0	39.1	51.8	44.8
+ AGG_5^L	47.1	41.5	39.8	34.0	39.9	42.1	31.4	43.5	33.7	45.2	39.8
+ AGG_5^F	45.0	38.4	38.5	32.4	38.8	47.1	30.4	41.2	33.3	44.2	38.9
+ AGG_5^P	45.5	38.8	38.0	32.0	38.8	**48.2**	30.5	41.0	33.2	44.0	39.0
+ AGG_5^N	45.4	39.1	38.3	32.4	38.8	48.0	30.7	41.2	33.2	44.3	39.1

Table 7.6: 140 experiments comparing 14 active learning methods translating into 10 different languages with Schedule *B*.

Pretraining with [N]^2 works well without M2M100: We compare 8 training schedules without M2M100 (Table 7.5). We find that Schedule *B* (pretraining on [N]^2 and training on [N+1]^2 and [N+1]) and Schedule *F* (pretraining on [N]^2 and training on [N+1]) work well without M2M100. Schedule *B* gives a chrF score of 51.1 and Schedule *F* gives a chrF score of 51.2.

M2M100 is useful when a target language and its corresponding source languages are in the M2M100 list and the test set does not overlap with the M2M100 training set. However, we strongly advise discretion, as training data for large pretrained models is usually not clearly specified and most are not trained with low-resource languages in mind. M2M100 training data may very likely contain the Bible data, so it only serves as a comparison and provides an alternative view to show that our model is robust with large models. When M2M100 does not apply, our models pretrained with [N]^2 suffice.

Full stage training increases robustness: For models without M2M100 we can use Schedule *B* (Table 7.6) or F (Table 7.9). Though the results for Frisian are similar, B is much better than F for morphologically rich languages like Pokomchi, Turkmen and Xhosa. Indeed, B with full training is more robust than F, which skips [N+1]^2. Similarly, for models with M2M100, we can use Schedule *I* (Table 7.7) or *L* (Table 7.8). Again, Schedule *I* with full training stages perform better than Schedule *L*.

Applying M2M100 alone gives poor results: Schedule *X* produces poor results (Table 7.4). Problems include catastrophic forgetting, bias towards rich-resource languages, and unclean data. Existing research shows some released models mislabel their English data as Welsh [236].

↑chrF	Frisian	Welsh	Hungarian	Spanish	Average
Baselines:					
+ *Luke*	49.3	44.3	38.8	48.4	45.2
+ *Rand*	53.5	49.5	42.2	53.2	49.6
Our Models:					
+ S	51.9	45.9	40.4	51.1	47.3
+ SN	54.8	47.4	42.3	53.2	49.4
+ SNG_2	54.5	49.5	43.5	54.2	50.4
+ SNG_3	54.4	50.4	**43.9**	54.5	**50.8**
+ SNG_4	**54.9**	49.8	43.2	**54.9**	50.7
+ SNG_5	54.5	50.1	43.5	54.1	50.6
+ ENT^N	52.7	47.2	40.9	52.9	48.4
+ ENT^K	54.6	49.4	43.5	53.8	50.3
+ AGG_5^A	49.4	44.2	37.3	48.2	44.8
+ AGG_5^S	46.5	49.8	36.4	46.4	44.8
+ AGG_5^M	48.6	50.4	36.5	46.9	45.6
+ AGG_5^T	48.8	**50.8**	36.4	46.9	45.7

Table 7.7: 56 experiments activating the knowledge in M2M100 with Schedule *I*.

Mixed models with M2M100 perform well: A few training schedules beat those pretrained with $[N]^2$ (Table 7.5). Schedule *I* (training on 5 stages) gives a chrF score of 52.9, L (training 3 stages skipping [N+1] and $[1]^2$) gives 52.8, M (training 4 stages skipping $[N+1]^2$) gives 52.7, J (training 4 stages skipping $[1]^2$) gives 51.8, and N (training 3 stages skipping $[N+1]^2$ and $[1]^2$) gives 51.9. All are higher than those without M2M100.

Adapting M2M100 to the domain and then to the low-resource language works best: Schedule *I* (training on 5 stages) with score 52.9 performs best. These models first adapt M2M100 to the domain by doing another pretraining on N^2. After adapting M2M100 to the domain, we adapt the model to the low-resource language by training on $[N+1]^2$. The final two stages [N+1] and $[1]^2$ are optional.

7.4.2 Qualitative Evaluation

We examine the selected sentences from different active learning algorithms and gain insight into how sentences are chosen.

Our models and mixed models perform better than M2M100 alone: M2M100 often produces extremely short sentences or repetition. Our models do not have those issues.

7.5 Conclusion and Future Work

Our key contributions is that we compare 24 schedules with large pretrained models in translation to low-resource languages. Our model is robust with large multilingual models. We find that adapting large pretrained models first to the domain by training on all text translations in existing source languages (N^2), followed by adapting it to the low-resource language by training on all translations including the low-resource data ($[N+1]^2$) works best. These two stages are the most essential while the rest is optional, and we recommend Schedule *I* that trains on all 5 stages introduced in this chapter. This helps to minimize human post-editing efforts during the subsequent iterations after translation of the seed corpus.

While the industry trend is to move towards bigger models with bigger data, our minimalist approach not only uses fewer languages, but we also aggregate over fewer languages. Our vocabulary size is ∼3000 for low-resource languages; this is in sharp comparison with large multilingual models like M2M100 with vocabulary size 128,108 for 100 languages. This saves computation power and resources, and therefore time and money, while improving translation performance.

Evaluation is still a challenge. It is difficult to find native speakers and establish long-term collaborations. There is also much variety among low-resource languages. Some are more accessible than others and these might provide earlier, realistic evaluation of our method. Hmong and Eastern Pokomchi are harder to assess while Frisian and Welsh, and many Eastern dialects in southern China and Indonesia, offer easier access and evaluation. Once we widen the set of low-resource languages by including more accessible ones, there are more possibilities to evaluate less accessible ones. Empowering low-resource languages is not just a technology problem. It requires much efforts in communication and collaboration with local communities. We welcome collaboration with native speakers to broaden our research perspective and to deepen mutual understanding of its diversity and complexity. Through our technologies, we would like to work with local communities to revive low-resource languages by bringing more young people to speak and use those languages. This will empower local communities and bring them to flourish.

In the next chapter, we will focus on working with human translators in the field of low-resource languages, and deploy the human machine translation workflow to evaluate its effectiveness in the real-world.

↑chrF	Frisian	Welsh	Hungarian	Spanish	Average
Baselines:					
Luke	49.1	41.7	38.3	48.7	44.5
Rand	52.8	46.8	41.9	52.9	48.6
Our Models:					
S	51.6	44.8	40.7	52.0	47.3
SN	53.2	45.8	42.2	52.9	48.5
SNG_2	54.2	47.6	42.5	53.8	49.5
SNG_3	53.7	47.9	**43.3**	**54.5**	49.9
SNG_4	**54.3**	48.5	43.2	54.4	**50.1**
SNG_5	53.9	48.6	43.2	54.5	50.1
ENT^N	52.1	44.8	40.7	52.4	47.5
ENT^K	53.7	46.7	43.1	53.7	49.3
AGG_5^A	48.4	43.2	37.1	48.4	44.3
AGG_5^S	47.3	48.1	36.1	47.1	44.7
AGG_5^M	46.9	47.8	36.3	47.2	44.6
AGG_5^T	47.1	**48.8**	36.1	46.8	44.7

Table 7.8: 56 experiments integrated with M2M100 on Schedule L.

↑chrF	Frisian	Hmong	Pokomchi	Turkmen	Sesotho	Welsh	Xhosa	Indonesian	Hungarian	Spanish	Average
Baselines:											
Luke	47.5	38.2	37.4	33.8	38.5	38.5	29.2	41.7	31.5	46.3	38.3
Rand	51.3	38.9	41.5	36.4	39.0	43.1	32.1	45.3	34.8	50.2	41.3
Our Models:											
S	48.7	35.8	39.8	27.6	36.1	38.1	29.4	41.5	32.5	47.5	37.7
SN	50.9	38.4	41.5	36.9	38.7	41.1	32.5	44.8	33.1	49.2	40.7
SNG_2	52.9	40.9	42.4	37.3	41.0	44.3	33.4	45.8	35.8	51.2	42.5
SNG_3	53.1	41.8	**43.2**	**38.4**	41.9	45.6	**34.0**	47.0	36.4	52.2	**43.4**
SNG_4	**53.6**	41.8	42.2	38.1	41.7	44.5	33.5	**47.5**	36.7	**52.5**	43.2
SNG_5	53.0	41.5	42.0	38.1	**42.3**	45.1	33.5	47.3	36.4	52.2	43.1
ENT^N	50.7	39.5	34.0	34.8	39.4	42.5	32.4	44.4	33.9	48.6	40.0
ENT^K	52.5	**42.4**	42.3	38.5	41.6	43.4	33.6	47.1	**37.1**	51.7	43.0
AGG_5^L	47.4	38.8	38.9	33.2	37.3	40.1	28.9	41.6	31.7	45.7	38.4
AGG_5^F	44.6	36.0	37.1	30.9	35.8	44.3	27.8	39.2	30.7	43.9	37.0
AGG_5^P	45.2	36.6	36.9	30.8	35.6	44.9	27.9	39.0	30.5	43.8	37.1
AGG_5^N	45.4	36.8	37.1	31.3	35.7	**46.0**	28.0	39.2	30.2	43.8	37.4

Table 7.9: 140 experiments comparing 14 active learning methods translating into 10 different languages on Schedule F.

Seed Size	Frisian	Hmong	Pokomchi	Turkmen	Sesotho	Welsh	Xhosa	Indonesian	Hungarian	Spanish	Average
Word count	25695	31249	36763	17354	25642	25786	15017	22318	18619	22831	24127
Line count for each experiment											
Baselines:											
Luke	1151	1151	1151	1151	1151	1151	1151	1151	1151	1151	1151
Rand	1022	1001	1101	1045	976	1117	988	1065	1066	1023	1040
Our Models:											
S	692	654	832	689	657	771	598	634	644	682	685
SN	1522	1399	1522	1524	1434	1595	1501	1601	1545	1488	1513
SNG_2	1484	1350	1490	1454	1369	1557	1418	1513	1468	1463	1457
SNG_3	1385	1319	1468	1416	1317	1439	1368	1451	1415	1365	1394
SNG_4	1327	1295	1419	1367	1279	1409	1309	1426	1374	1310	1352
SNG_5	1289	1289	1397	1311	1280	1381	1256	1359	1334	1273	1317
ENT^N	1796	1721	1769	1840	1761	1914	1839	1967	1884	1805	1830
ENT^K	1340	1287	1507	1266	1132	1405	1128	1358	1264	1327	1301
AGG_5^A	984	1025	1060	998	967	1031	1016	1018	993	958	1005
AGG_5^S	1049	1084	1152	1043	1025	1182	1147	1093	1076	1019	1087
AGG_5^M	1058	1097	1159	1109	1025	1232	1159	1101	1087	1018	1105
AGG_5^T	1048	1094	1153	1101	1020	1274	1141	1101	1087	1014	1103

Table 7.10: Seed Corpus Size for different target languages. The seed corpus gives rise to both training data and validation data, therefore the training size is smaller than the above. Note that all experiments for a given target language share the same number of words, although they have different number of lines. Since each language use different number of words to express the same meaning of a given text, we choose the number of words in the given book "Luke" as the standard reference for each target language. For example, "Luke" in Xhosa contains 15,017 words while "Luke" in Frisian contains 25,695 words.

Chapter 8

A Quechuan Case Study

> "The difference between the right word and the almost right word is really a large matter – it's the difference between lightning and a lightning bug."
>
> *Mark Twain*

Having examined various active learning methods to build a seed corpus in the low-resource language, and having explored multiple scheduling methods to use large pretrained multilingual models to train on such a small seed corpus, we dive into the real-life applications of translation into low-resource languages by working closely with field linguists and human translators in Peru. In this chapter, we evaluate our progress from all previous chapters through a case study in Quechuan language family by working with field linguist Mark Bean and his translation team. We show effectiveness of our model and our proposed human machine translation workflow for translation of a multi-source text into a new, low-resource language. We find that translation performance is significantly positively correlated with language similarity. The more connected a language is, the higher the translation performance. Furthermore, we find decluttering poorly-connected languages improves performance. Using this finding, we show our results in translating into a new, low-resource language called Sihuas Quechua.

8.1 Introduction

Machine Translation is not only a core scientific problem for Computer Science, but also a cost-cutting business tool. The market for language services is estimated to be US $72.2 billions in 2023 [68, 143]. Many governments, businesses, international organizations like

the European Union, and numerous global missions will accelerate in their spending on translation services by US $20.83 billions during 2022-2026 in a recent forecast [182]. And the total market for language services is forecast to climb to US $98.1 billions in 2028 [143].

Although the current market for translation into low-resource languages are relatively small compared with the world's spending on translation services, its potential market is substantial. Indeed, translation into low-resource languages carries immense value in serving the cultural goal of saving and reviving low-resource languages and the humanitarian goal of assisting the everyday needs of local communities.

To serve the goal of cultivating and expanding potential translation markets in low-resource languages, machine translation cannot stand alone. Large MT system outputs have problems with repetition, mislabelled data, hallucination, inaccuracy, underproduction, over-fitting and inconsistencies [36, 191, 280]. Neural MT systems often requires human post-editing to produce readily publishable materials, independent of how evolved the models are [66]. Systems adapted from large MT systems may have some improvements in translation, but still have problems and need post-editing to produce publishable translations. Indeed, post-editing and human checking is crucial in producing publishable materials [280].

Given the importance of post-editing and human checking, symbiosis is necessary between Machine Translation and human translation to accelerate and benefit the translation industry. In our translation framework, our goal is to translate a text that is available in multiple source languages to a new, severely low-resource language. To accomplish this goal, we rely on the close collaboration between MT systems and human translators to finish a high quality translation of the given text. We begin by performing active learning methods to rank sentences in the given text on available languages. Once we finish ranking, we inform human translators. After human translators obtain the ranking, they prepare the translation of the top few ranked sentences (∼1,000 lines) to build our seed corpus. Once human translators finish building the seed corpus, they pass it to MT systems. Using this seed corpus, MT systems train and produce the first translation draft of the entire text in the severely low-resource language. Working on this draft, human translators post-edit and complete translation for a few portions of the text. These newly translated text is then fed back to the MT systems to improve training and produce improved version of the draft. Iterations of post-editing not only produce new data to improve translation quality by MT systems, but also serve as a positive feedback loop and expedite the entire process. In this way, human translators and MT systems work together to finish a high quality translation of the text.

Having understood this translation workflow between MT systems and human translators, we would like to know how it works in real life and evaluate the progress from all previous chapters. In Table 8.1 and Table 8.2, we show the summary of results from the previous chapters. We see that multilingual training by carefully selecting similar languages to train with (Chapter 3-5), active learning (Chapter 6) and staged finetuning with large

↑chrF	Frisian	Hmong	Pokomchi	Turkmen	Sesotho	Welsh	Xhosa	Indonesian	Hungarian	Spanish	Average
Baseline	23.1	25.0	28.7	18.9	25.2	22.2	21.4	27.2	20.1	22.1	23.4
Our Models:											
+ Chapter 3-5	28.0	28.1	31.9	22.6	28.3	26.5	23.9	29.7	22.3	26.8	26.8
+ Chapter 7	50.5	43.9	42.8	38.9	43.2	46.0	34.9	47.2	37.4	50.1	43.5
+ Chapter 6	53.6	45.7	44.4	40.3	44.9	47.7	36.8	49.1	39.0	52.7	45.4

Table 8.1: Result summary for translation into 10 languages that are new and severely low-resource to the system, independent of M2M100.

↑chrF	Frisian	Welsh	Hungarian	Spanish	Average
Baseline	23.1	22.2	20.1	22.1	21.9
Our Models:					
+ Chapter 3-5	28.0	26.5	22.3	26.8	25.9
+ Chapter 7	53.5	49.5	42.2	53.2	49.6
+ Chapter 6	54.9	49.8	43.2	54.9	50.7

Table 8.2: Result summary for translation into 4 languages that are new and severely low-resource to the system, leveraging knowledge in M2M100 and using active learning.

pretrained models (Chapter 7) improve translation performance in our experiments. To evaluate improvements made from Chapter 3-7, and to translate our progress in academic settings into the real-world applications, we work with field linguist Mark Bean and his team in Peru on a case study of the Quechuan language family.

Working on this Quechuan case study, we show in this chapter that for well-connected languages, our finding of multilingual training by carefully selecting similar languages to train with (Chapter 3-5), active learning (Chapter 6) and staged finetuning with large pretrained models (Chapter 7) improve translation performance. For poorly-connected languages, due to the low language similarity available for cross-lingual transfer, the impact of our method could not be tested.

8.2 A Case Study on Quechuan Languages

To show effectiveness of our methods, we collaborate with field linguist, Mark Bean, and his team who work closely with the low-resource Quechuan language communities in South America, especially in the region of Peru. Mark and his team have been spending their lives working closely with the Quechuan language community to translate the Bible text into each of the target Quechuan languages.

Our goal in this case study on the Quechuan language family is three-fold. Firstly, we aim to find out whether our method works in real-world translation efforts. Furthermore, we want to determine under which conditions it is most favorable to apply our method to

Figure 8.1: Sisters who speak Margos Quechua in Peru. Photograph by Mark Bean.

translate into the low-resource languages successfully. Lastly, in the instances where those conditions are not met, we suggest a few future research directions to improve translation quality.

8.2.1 History and Geography

The Quechuan language family is a varied group of languages covering a wide Andean region of South America, stretching from Colombia, Ecuador, Peru, Bolivia to northwest Argentina [137, 184]. There is broad spectrum of sociolinguistic diversity among Quechuan languages throughout the Spanish colonial history of the area [77].

Quechua is widely spoken in a wide range of Peruvian Andes before the expansion of the Inca Empire [57]. The Inca empire enforced Quechua as the official language during its rule where many diverse dialects are developed, influenced by local languages [155]. For example, Quechua is influenced by Aymara in the area of Cuzco, which is the old Inca capital [294]. In addition to the support of the Inca Empire, the Spanish rulers also helped Quechuan languages to grow and flourish [87]. The Catholic church encouraged and facilitated the first written form of Quechua [65].

Ethnologue has 45 Quechuan languages which are then divided into two groups: central (Quechua I) and peripheral (Quechua II) [3, 81]. Within the categories, they are part of the

ISO Code	Language	Total	Books	OT	NT
quz	Quechua, Cuzco	31099	66	23142	7957
quy	Quechua, Ayacucho	31099	66	23142	7957
quh	Quechua, South Bolivian	31099	66	23142	7957
qub	Quechua, Huallaga	31099	66	23142	7957
qxo (qxoc)	Quechua, Southern Conchucos	31099	66	23142	7957
qxo (qxoh)	Quechua, Huacaybamba	31099	66	23142	7957
qve	Quechua, Eastern Apurímac	31099	66	23142	7957
qvh	Quechua, Huamalíes-Dos de Mayo Huánuco	31099	66	23142	7957
qvm	Quechua, Margos-Yarowilca-Lauricocha	31099	66	23142	7957
qvw	Quechua, Huaylla Wanca	31099	66	23142	7957
qwh	Quechua, Huaylas Ancash	31099	66	23142	7957
qxn	Quechua, Northern Conchucos Ancash	31099	66	23142	7957
inb	Quechua, Inga	7957	27	0	7957
quf	Quechua, Lambayeque	7957	27	0	7957
qul	Quechua, North Bolivian	7957	27	0	7957
qup	Quechua, Southern Pastaza	7957	27	0	7957
qvc	Quechua, Cajamarca	7957	27	0	7957
qvn	Quechua, North Junín	7957	27	0	7957
qvs	Quechua, San Martín	8308	28	351	7957
qvz	Quechua, Northern Pastaza	7957	27	0	7957
qxh	Quechua, Panao	7957	27	0	7957
qws	Quechua, Sihuas	7373	23	1494	5879

Table 8.3: Quechuan Family. "Total" is the total number of lines in the text, "OT" is the number of lines in Old Testament while "NT" is that in New Testament, and "Books" is the number of books translated. To differentiate Southern Conchucos from Huacaybamba, we use "c" and "h".

dialect continuum as they mostly contain dialects spoken across the Peruvian Andean area where they could be mutually intelligible if they are close enough [34]. However, languages are not mutually intelligible across categories [167].

8.2.2 Key Languages in Analysis

We show the Quechuan family of languages[1] we are working on in Table 8.3. We have Quechuan languages that differ in resource levels: we have 12 languages that have complete Bible text, and 9 languages that have at least New testament text, and 1 language

[1]Note that Huacaybamba Quechua and Southern Conchucos Quechua both have the ISO code "qxo", in other tables of this thesis, we will differentiate the two as "qxoh"(Huacaybamba) and "qxoc"(Southern Conchucos).

that only has a few stories translated. There are twelve Quechuan languages that has complete translation of the Bible text: Cuzco Quechua, Ayacucho Quechua, South Bolivian Quechua, Huallaga Quechua, Southern Conchucos Quechua, Huacaybamba Quechua, Eastern Apurímac Quechua, Huamalíes-Dos de Mayo Huánuco Quechua, Margos-Yarowilca-Lauricocha Quechua, Huaylla Wanca Quechua, Northern Conchucos Ancash Quechua, Huaylas Ancash Quechua. These languages have complete Bibles in both Old Testament and New Testament. We also have nine other languages that have only partial translations of the Bible, having at least the entire translation of the New Testament. They are: Inga Quechua, Lambayeque Quechua, North Bolivan Quechua, Southern Pastaza Quechua, Cajamarca Quechua, North Junín Quechua, San Martín Quechua, Northern Pastaza Quechua, and Panao Quechua. The language that does not have the entire translation of the New Testament is Sihuas Quechua. Sihuas has partial translations of the New Testament and a few chapters from the Old Testament.

8.3 Data

We use both the partial and complete translations across 22 languages in Quechuan family. We have the complete translations of the Bible data in 12 languages. These data include both the Old Testament and the New Testament data. We also use the 9 languages that has New Testament data. We have a language, Sihuas, that has only partial New Testament and partial Old Testament data. There are three languages that we are focusing on in our detailed qualitative evaluation, and they are: Margos-Yarowilca-Lauricocha Quechua (*Margos*), Panao Quechua (*Panao*) and Sihuas Quechua (*Sihuas*) [77, 88, 268, 297]. There is an existing translation of the complete Bible in Margos, however, there is no existing translation of the Old Testament in Panao. Finally, no translation of the Old Testament or the New Testament exists in Sihuas.

8.4 Results Analysis

We evaluate the effectiveness of our model by investigating the following questions: 1.) Does our method presented in this thesis work in real-world translation? 2.) Under which conditions does our method work, and under which conditions does it not work? 3.) When our method works, how well does it work? 4.) When our method does not work, which future research directions could we pursue to improve translation performance?

To answer these questions, we would like to examine the core focus of this thesis by looking at the relationship between translation performance and language similarity. In this section, we first look at language similarity and determine its correlation with MT

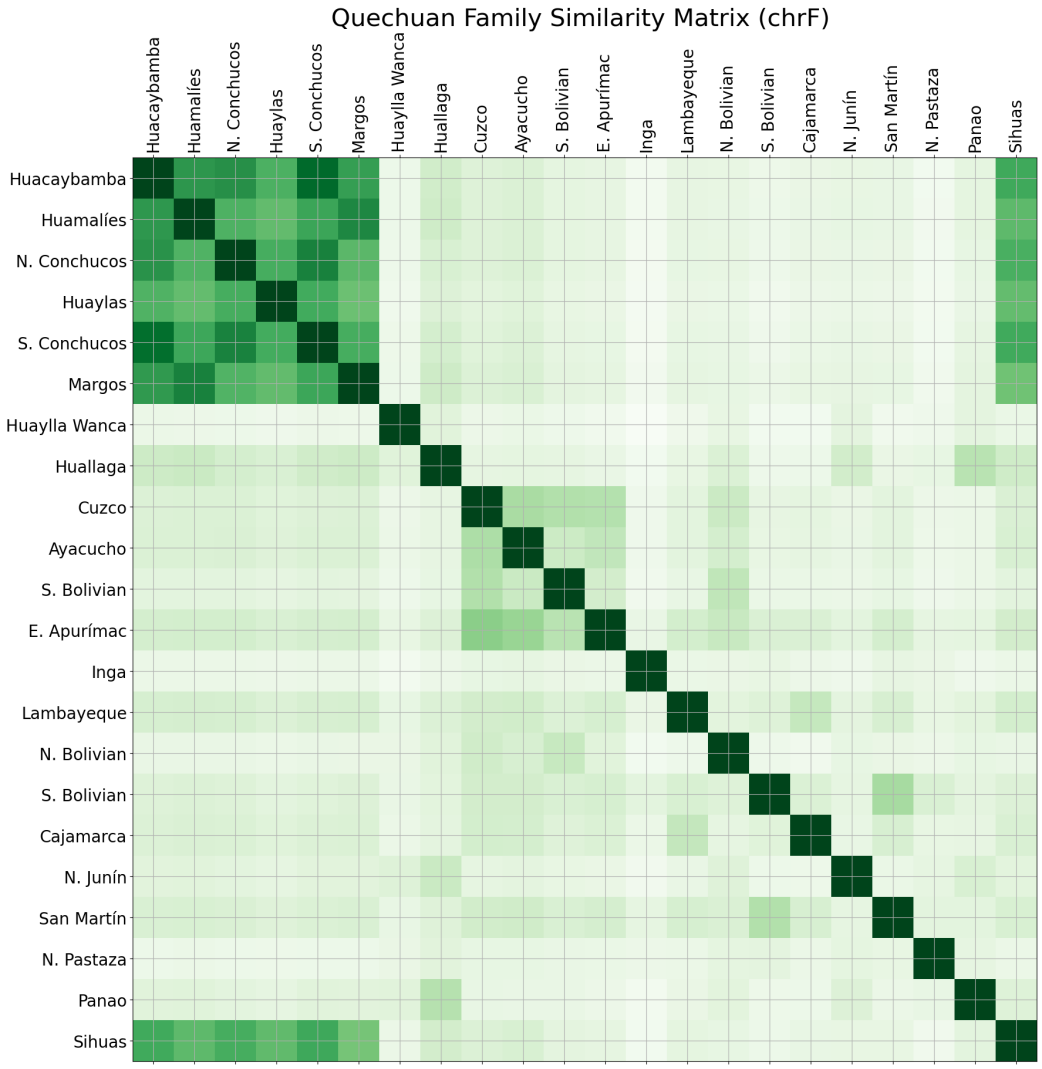
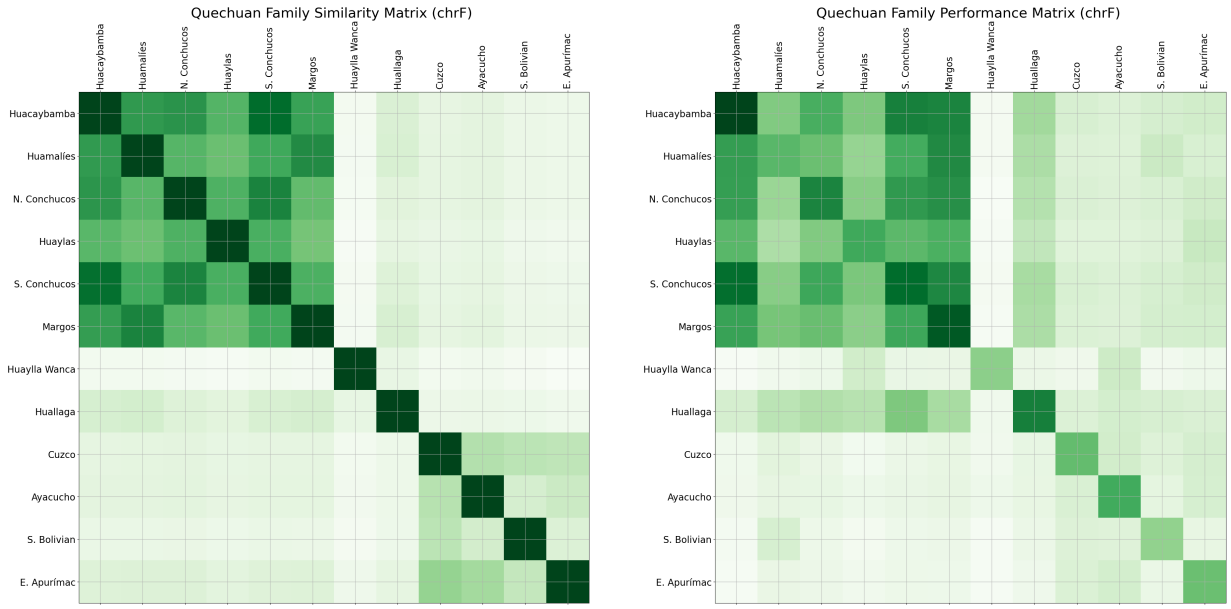


Figure 8.2: Similarity matrix on Quechuan family (chrF).

performance in three different perspectives. After determining their correlation, we conclude this analysis in translation into Sihuas Quechua, a new, low-resource language.

8.4.1 Similarity Analysis

We begin our analysis by examining language similarity. There are many methods to measure language [53, 121, 212]. Most methods require information on typology [50, 60, 123, 226, 237, 277], World Atlas of Language Structures [60], and the Swadesh list [139]. Given severely low-resource scenarios in our use case, we do not make assumptions on external resources, and we aim to measure language similarity in a light-weight manner by only using the text we are translating.



(a) Similarity matrix on 12 Quechuan languages.

(b) Performance matrix on 12 Quechuan languages.

Figure 8.3: Comparison of similarity and performance matrices (chrF).

Using the given text, we measure language similarity between any two given languages by looking at the text translations in them. We analyse sentence overlap (percentage of matching sentences in both languages) and word overlap (percentage of matching words in both languages). We also use metrics including characTER, chrf and 1-gram and 4-gram BLEU [217, 230, 305] on translations in both languages.

Using these pair-wise similarity measures, we show a 22-by-22 similarity matrix for all Quechuan languages in our case study using chrF metric in Figure 8.2 and a 12-by-12 close-up version for the 12 Quechuan languages with full text translation in Figure 8.10. Margos and Sihuas are in the dark green area (well-connected), Panao is in the almost white area (poorly-connected). We clearly see that there is a strong language cluster among these 6 Quechuan languages: Huacaybamba Quechua, Huamalíes-Dos de Mayo Huánuco Quechua, Northern Conchucos Ancash Quechua, Huaylas Ancash Quechua, Southern Conchucos Quechua, and Margos-Yarowilca-Lauricocha Quechua. There is a less obvious language cluster among 4 other Quechuan languages: Cuzco Quechua, Ayacucho Quechua, South Bolivian Quechua, and Eastern Apurimac Quechua. Apart from these two language clusters, the similarity matrix is mostly light green in other areas except for the language Sihuas Quechua.

Sihuas Quechua, represented as last row and column in the similarity matrix in Figure 8.2, stands out as deep green. Even though Sihuas is the least resourced among the 22 languages, it is well connected to the main cluster we identified earlier and is similar to each of the 6 languages in said cluster.

Target Language	chrf	characTER	1-gram BLEU	4-gram BLEU
Quechua Huacaybamba	81.4	0.858	75.0	49.4
Quechua Huamalíes-Dos de Mayo Huánuco	60.0	0.518	56.2	23.0
Quechua Northern Conchucos Ancash	70.3	0.666	66.0	30.9
Quechua Huaylas Ancash	61.9	0.555	59.6	22.7
Quechua Southern Conchucos	80.4	0.793	77.4	46.8
Quechua Margos-Yarowilca-Lauricocha	83.2	0.835	80.1	54.7
Quechua Huaylla Wanca	27.6	0.231	32.1	3.9
Quechua del Huallaga Huánuco	52.5	0.417	45.5	12.8
Quechua Cuzco	39.9	0.328	33.7	4.7
Quechua Ayacucho	39.4	0.320	29.8	2.9
Quechua South Bolivian	41.4	0.317	37.2	6.2
Quechua Eastern Apurímac	43.2	0.310	33.9	5.7

Table 8.4: Key result summary of the round robin experiments.

In addition to chrF metric, we show the full set of language similarity matrices using chrF, characTER, 1-gram BLEU, 4-gram BLEU, sentence overlap and word overlap in Figure 8.9. We see that all 6 matrices are very similar, though having different levels of granularity. The sentence overlap metric, for example, is the coarsest measure as every pair is mostly light green. This is very intuitive because it is indeed rare that the same sentence is expressed in the exactly the same way in two languages. For example, Margos has 5% sentence match and 88% word match. The low percentage of sentence match could be caused by alignment issues or close-to-zero edit distances that are not captured by this metric. Indeed, other metrics like chrF, characTER, 1-gram BLEU, 4-gram BLEU offer more fine-grained analysis portfolios for each language. The most fine-grained analysis is done through character-level metrics like chrF and characTER.

8.4.2 How Similarity Affect Performance

Having explored various similarity measures on 22 Quechuan languages, we would like to examine how similarity influences translation performance. We show three approaches in this section: the round robin experiment, and the effects of adding and removing similar languages. We examine each approach closely in the remaining section.

Round Robin Experiment

Firstly, we conduct a set of round robin experiments. To simplify our visualization and assure data symmetry, we choose 12 languages with complete text translations for this set of experiments: Cuzco Quechua, Ayacucho Quechua, South Bolivian Quechua, Huallaga Quechua,

Spearman	chrf	characTER	1-gram BLEU	4-gram BLEU
Correlation	0.771	0.729	0.703	0.741
P-value	4.25×10^{-25}	1.31×10^{-29}	2.51×10^{-26}	9.65×10^{-23}

Table 8.5: Key result summary of correlation between performance and similarity.

Southern Conchucos Quechua, Huacaybamba Quechua, Eastern Apurímac Quechua, Huamalíes-Dos de Mayo Huánuco Quechua, Margos-Yarowilca-Lauricocha Quechua, Huaylla Wanca Quechua, Northern Conchucos Ancash Quechua, Huaylas Ancash Quechua.

For each round of the round robin experiments, we take one of these 12 languages as the hypothetical target low-resource language, and assume that we only have New Testament data for this language. Since all the other 11 languages have complete translations of the whole Bible including both the Old Testament and the New Testament, we train on all the 11 complete Bibles and test on the Old Testament of the chosen target language. We use DeltaLM as our large pretrained model for multi-stage adaptations [185].

For DeltaLM, we use the sentence-piece vocabulary that comes directly with it. This vocabulary contains 250k unique tokens. We have tried to build our own vocabulary that is morpheme-based; however, training with existing vocabulary works much better than morpheme-based self-constructed vocabulary. Therefore, we use the default vocabulary with 250 unique tokens.

For training schedules, we train using both Schedule B (without using large pretrained models) and Schedule J (using large pretrained models). For example, if we choose Margos as the target low-resource language, we reach a BLEU score of 53.0 with Schedule B and we have a BLEU score of 54.7 with Schedule J. Indeed, Schedule J is also preferred by the field linguists through qualitative evaluation. For simplicity, we show results of Schedule J in this section.

We show key results from the round robin experiments in Table 8.4. Each row represents the combined score from all 11 source languages into the given target language specified in the first column. We see that there is a wide spectrum of machine translation performance. The experiment translating into Margos, for example, reaches a BLEU score of 54.7 and a chrF score of 83.2. However, the experiment translating into Cuzco reaches a BLEU score of 4.7 and a chrF score of 39.9. To understand this large difference in performance, we want to establish the correlation of performance and similarity to the target language.

Correlation between Performance and Similarity

To understand the correlation between translation performance and language similarity, we see that they are positively correlated in Figure 8.3. To visualize how translation performance is driven by language similarity, we show similarity matrices and performance

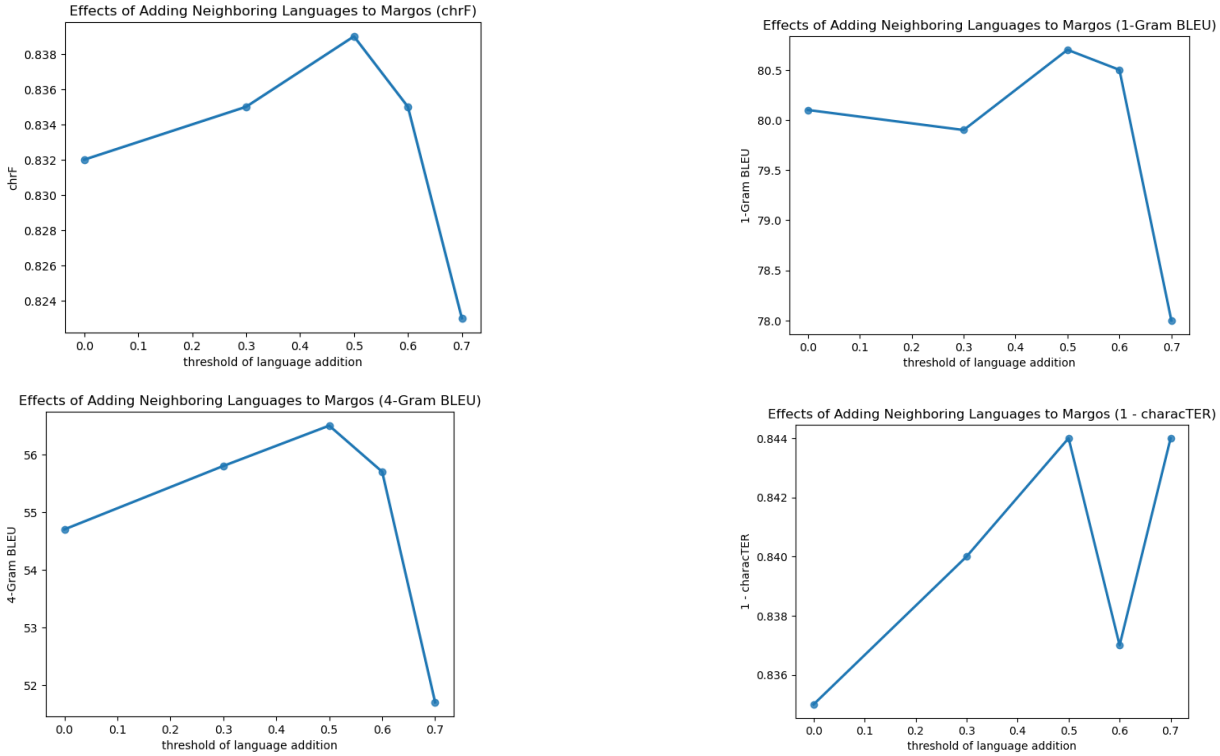


Figure 8.4: Effects of adding similar languages in translation into Margos

matrices for the 12 Quechuan languages we have shown earlier in Figure 8.11 and Figure 8.3. Comparing them side-by-side, we see their resemblance, especially at the 6-language cluster that is dark green (well-connected area).

To determine the statistical significance, we take the column-wise Spearman correlation between the performance matrix and the similarity matrix and show key results in Table 8.5. For the chrF metric, we have a correlation of 0.771 between the performance matrix and the similarity matrix. The associated p-value is 4.25×10^{-25}, which shows high statistical significance. A detailed analysis of Spearman correlation in all 4 metrics (chrF, characTER, 1-gram BLEU, 4-gram BLEU) shows extremely significant p-values and high correlation scores in Table 8.5. Indeed, this shows that the more well-connected the target language is, the higher the translation performance.

Having established the strong positive correlation between performance and similarity, we show a more fine-grained comparison in Figure 8.11. The last column of this figure shows the fine-grained correlation and p-value by each source language. We see that there is strong positive correlation for languages close to the 6-language cluster while there is little correlation for languages that are already far away from the other languages. This means the more connected a language is, the more language similarity correlates with translation performance. Moreover, the more connected a language is, the better the performance.

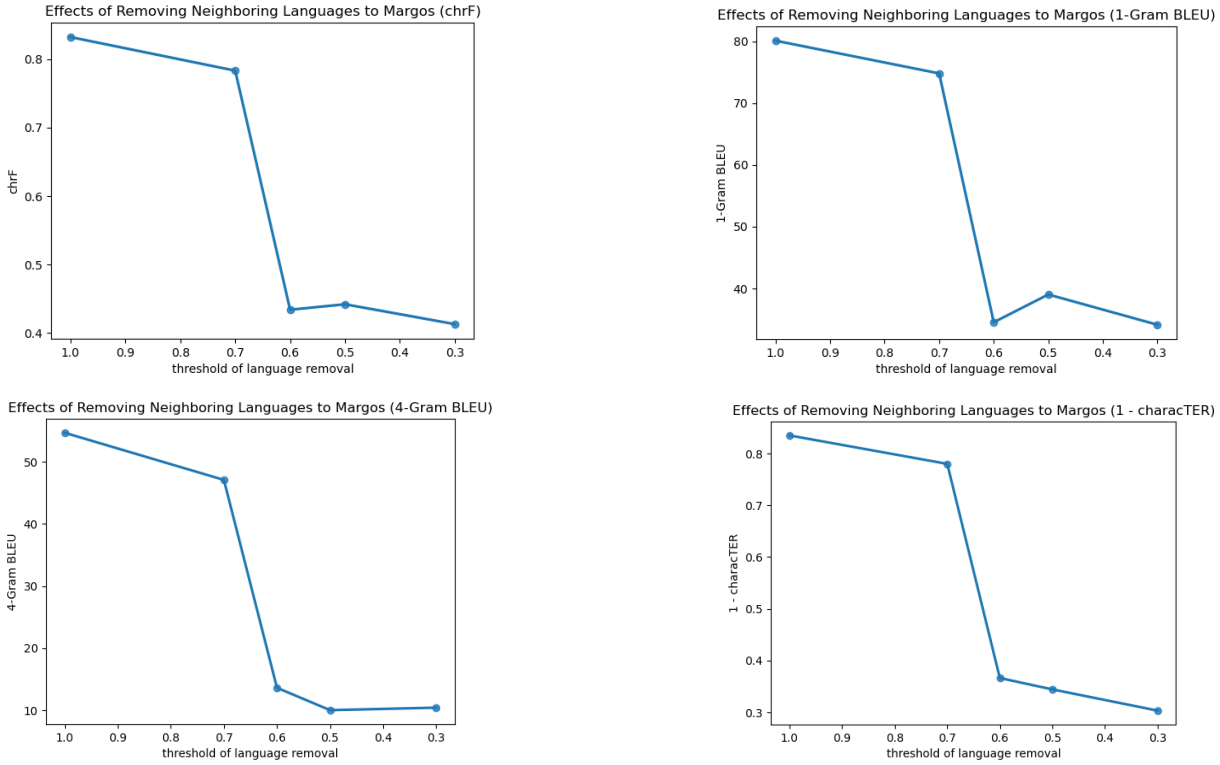


Figure 8.5: Effects of removing similar languages in translation into Margos

Effects of Adding and Removing Similar Languages

Having established that the more connected a language is, the better the translation performance, we now examine the relationship between translation performance and language similarity from another perspective. In Figure 8.3a, we show the degree of connection between languages in varying shades of green. The darker green represents high similarity and the lighter green represents low similarity. Using the varying similarity scores, we visualize how similarity correlates with performance by measuring performance while adding or removing similar languages.

We choose Margos as the target language to serve as the control, and rank other languages according to their similarity to Margos. As we drop or add similar languages, we measure the performance of our method with a varying number of source languages. We show the effects of adding languages in Figure 8.4 and that of removing languages in Figure 8.5. The x-axis is the chrF similarity threshold. For example, a threshold of 0.5 in Figure 8.4 means that the corresponding experiment only trains on source languages that achieves at least a chrF score of 0.5. On the contrary, a threshold of 0.5 in Figure 8.5 means that the corresponding experiment only trains on source languages that achieve a chrF score less than 0.5.

The optimal threshold for translating into Margos by adding languages is noticeably 0.5 chrF as shown in Figure 8.4, meaning that it is optimal to train on source languages that are closer to the target low-resource language. Moreover, adding more distant languages hurts translation performance.

Additionally, we find similar results by removing languages in Figure 8.5. If we were to train on source languages that achieve a chrF score less than 0.7, we still maintain good performance. However, once we move to 0.6, the performance plunges. This shows how crucial source languages that are closer to the low-resource language are. Therefore, it is important to keep these languages instead of removing them.

Both sets of experiments point to the same conclusion that choosing languages that are very similar to the target language is key in achieving high translation performance. Furthermore, decluttering poorly-connected languages helps to improve translation.

Understanding Poorly-Connected Languages

Having established that the more connected a language is, the higher the translation performance, it is easy to understand why our multi-stage adaptation method works very well with highly similar Quechuan languages as shown in Table 8.4. For example, the experiment translating into Margos reaches a BLEU score of 54.7 and a chrF score of 83.2. However, in the same round robin experiments in Table 8.4, we also observe that our multi-stage adaptation method does not work as well with a few poorly-connected Quechuan languages that are very different from the rest. For instance, the experiment translating into Cuzco reaches a BLEU score of 4.7 and a chrF score of 39.9.

To understand this low performance with poorly-connected languages, let us focus on a more detailed comparison between the performance matrix and similarity matrix in Figure 8.11. The last column of this figure shows the fine-grained correlation and p-value by each source language. 8.11. The p-value is close to 0 for strongly-connected languages while the p-value is very high for poorly-connected languages. In other words, there is a significant positive correlation for strongly-connected languages while there is little correlation for poorly-connected languages. This means the more connected a language is, the more language similarity correlates with translation performance. Consequently, for poorly-connected languages, we cannot deduce correlation between performance and similarity, and their performance is therefore uncertain.

One potential reason for such poor performance is that the advantage of multilinguality does not lend any leverage to the poorly-connected languages because generalization and cross-lingual learning is hard. Another reason is very related to the practice of the human translation team. In the past, translations are done through years. Each translation team may translate a set of languages together using the same conventions and notations. When a new translation team is formed, the new team may choose to use different notations and

System Translation	Human Post-edits	Edits
Tayta Diostam päyakö qamkuna kaqćhö kaykar Crisputa y Gayullata bautizanqäpita.	Tayta Diostam**i** päyakö qamkuna kaqćhö kaykar Crisputa**wan** y Gayullata bautizanqäpita.	2
Tsaymi mayqëkipis niyankimanku noqapa shutëćhö bautizakuyanqëkita.	**Ch**aymi mayqëkipis niyankimanku noqapa shutëćhö bautizakuyanqëkita.	1
Tsaynöllami Estéfanas castallata bautizash**q**ä. Tsaypita mastaqa manami pitapis bautizanqäta yarpäku.	**Ch**aynöllami Estéfanas castallata bautizash**k**ä. Tsaypita mastaqa manami pitapis bautizanqäta yarpäku.	2

Table 8.6: Qualitative evaluation in translation into Sihuas.

conventions. This may result in a lot changes as it may affect the language documentation process and could potentially add more difficulty in translating into languages that are already poorly-connected.

One may ask, what could help translation into such poorly-connected languages? Does adding more data help? In our experiments with one of the poorly-connected languages, Panao Quechua, we initially have poor results. In order to help us have more data for training, human translators translate more sentences and provide us with much more data for training. With more data, our translation result remains poor. This could be because when languages are far apart, multilinguality may not be realistically useful due to the limited cross-lingual transfer learning.

Beyond the Quechuan Family

We have considered languages in the Quechuan family so far. However, we have not considered languages outside of the Quechuan family. Are there languages outside the Quechuan family that are similar to the target language that we may use for training? To visualize this, we measure similarity of 142 source languages to a given target language. We show language similarity rankings for Margos (Figure 8.7a), Panao (Figure 8.7b) and Sihuas (Figure 8.7c).

We find that Aymara [125], which is a South American native language spoken in a nearby region (on the Altiplano) close to Peru, has relatively visible similarity with Margos. However, the closest languages are still from the Quechuan language family. This means that even though some Quechuan languages are very different from one another, they are still closer to each other than those outside the Quechuan language family. Therefore, we still recommend finding similar languages within the Quechuan family for translating into any of the Quechuan low-resource languages.

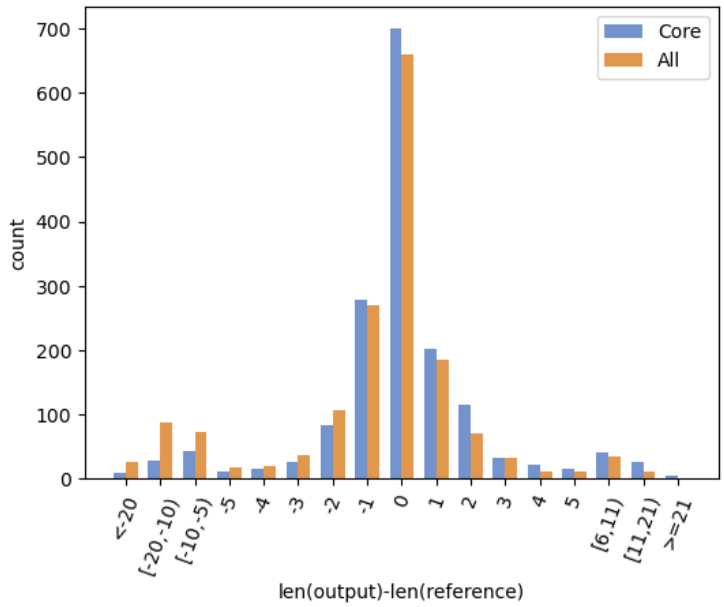


Figure 8.6: Output and reference length difference for two systems translating to Sihuas using compare-MT [207]. The blue system translates using languages that are at least 0.6 chrF with Sihuas, while the orange system trains on all.

8.4.3 Translation into Sihuas

Having understood that the more well-connected a language is, the higher the translation performance, we identified Sihuas. Sihuas, though having the least resources, is very close to the other Quechuan languages, especially to the 6 language cluster shown in Figure 8.3a. We therefore conduct experiments to translate into Sihuas.

Translation Results

When we train on the incomplete translations of the New Testament and test on partial translations of the Old Testament, we reach a BLEU score of 43.4 and a chrF score of 75.8 and a characTER score of 0.279. This indeed shows that the well-connected language is easier to translate into.

Furthermore, applying what we learned about the threshold of closeness by adding and removing similar languages from training, we conduct an experiment by training only on languages that have similarity scores of at least 0.5 chrF with Sihuas. We improve our translation score to 44.7 of BLEU, 75.2 of chrF and 0.248 of characTER. However, this is not the most optimal. This shows that we improve translation performance by decluttering poorly-connected source languages.

Finally, after working further with field linguists who speak Sihuas, we decide to train only on languages that have similarity scores of at least 0.6 chrF with Sihuas, and we improve

our translation score further to 46.3 of BLEU, 76.3 of chrF and 0.242 of characTER. This is our final translation system for Sihuas. In Figure 8.6 and 8.14, we show that the system that trains only on close languages performs better than the system that trains on all available languages.

Comparing with Post-Editing Existing Source Texts

Having shown the high performance of our MT system that translates to Sihuas Quechua, we compare our method with post-editing on existing source texts directly. Post-editing existing source texts is what human translators could do without help from any MT systems. Comparing these two is very meaningful to understand our contribution.

Our MT system is more effective than just post-editing existing books. If we post-edit from the translation of the text in Huacaybamba, the closest language to Sihuas, we have a chrF score of 68.3 and a BLEU score of 25.7 between Huacaybamba and Sihuas. Our machine translation has a chrF score of 76.3 and a BLEU score of 46.3. Even though there is a ∼8-point difference in chrF, there is a ∼20-point difference in terms of BLEU. Given the ∼20-point BLEU difference, there could be a lot of words in Huacaybamba that is similar to Sihuas, but not exactly the same. Our translation system is able to learn the sub-word level morphology changes, and therefore make it easier for the human translators.

In addition to the ∼20-point BLEU difference, human translators benefit more from a better post-editable draft of the text from our MT system, rather than post-editing from source texts directly. Our MT system lessens the chance of errors and reduces the amount of deletion, insertions and changes that human translators need to do to post-edit well.

8.5 Limitations and Future Work

From this Quechuan case study, we demonstrate that for well-connected languages, our finding of multilingual training by carefully selecting similar languages to train with (Chapter 3-5), active learning (Chapter 6) and staged finetuning with large pretrained models (Chapter 7) improve translation performance. For poorly-connected languages, due to the low language similarity available for cross-lingual transfer, the impact of our method could not be tested.

Our work is limited by the text and the translations of this text in different languages in Quechuan family that is available to us in this chapter. However, the relationship between similarity and performance is generalizable to other languages and language families. In Chapter 5, we have examined the translation performance using our method for the European medical EMEA dataset, have shown good performance. In Chapter 6 and 7, we

have shown our method is generalizable to different target languages. However, all of our results in this thesis is limited by the text and the amount of data that is available to us.

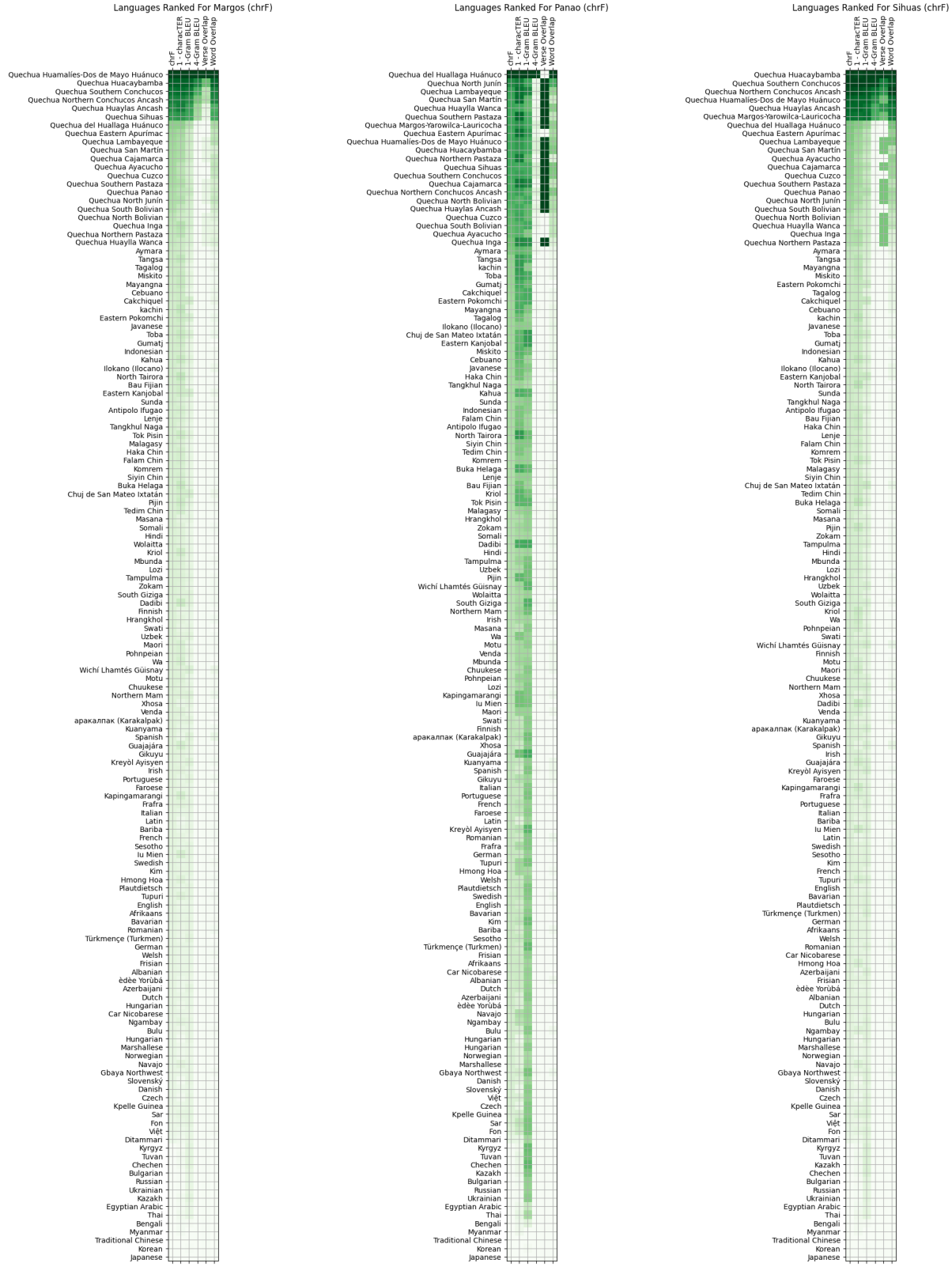
This limitation points us to many opportunities to extend this work. We are keen to explore more varied target low-resource languages and language families. We are also open to collaborate with more human translators working in other parts of the world to extend our work to other domains including legal, literary and educational documents.

In addition to wider domains and more languages, another area of interest is to find languages that may not be close to each other in the textual domain but in other domains that we could use to improve translation. One potential method is to transform our data from the textual space to the phonological space or other non-textual domains. In Figure 8.12, Figure 8.13 and Figure 8.7, we show typological features based on genetic, featural, geographic, inventory, phonological and syntactic similarities. Comparing these features with our similarity graphs in Figure 8.9, we see much richer connections among the Quechuan languages. If we can find a way to transform our data into the phonological space and find related languages to improve performance score, that will be very promising.

Furthermore, it is very important to explore what we can do when there are very few similar languages. In the case for translating into Panao Quechua, translation performance suffers as Panao is more distant from other languages in the textual form. We have tried to build a Quechuan-specific morpheme vocabulary. However, once we change the vocabulary from the large pretrained model to incorporate this Quechuan-specific morpheme vocabulary, training diverges. It is difficult to train large pretrained models using the new vocabulary in academic settings, but this gives room for creative solutions in future research.

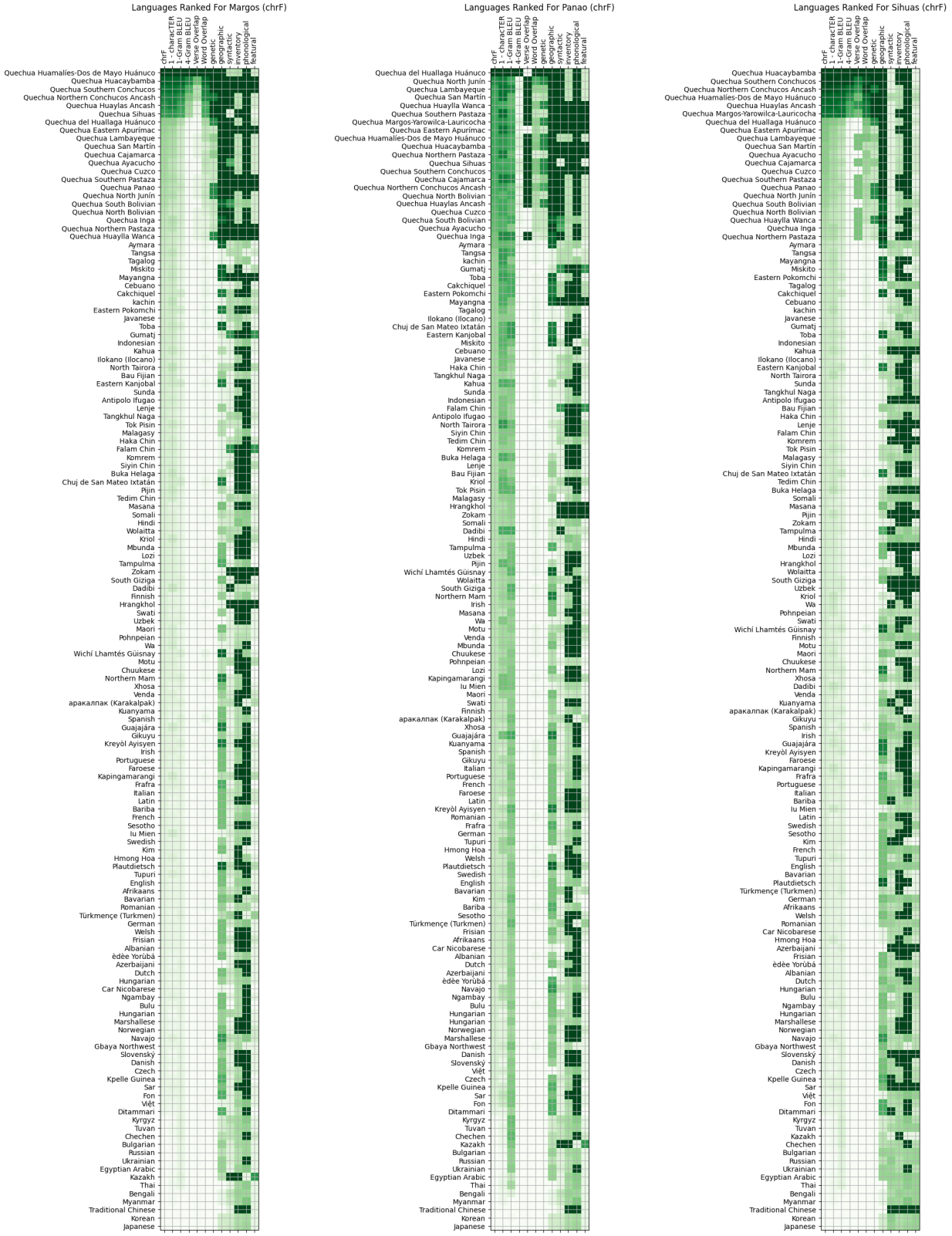
Finally, there are continued conversations with the field linguists and human translators working in the field. Understanding and working with their needs are key in building long-term collaborative relationships. This process is a continued dialogue. It is through this continued dialogue that we respect the dignity of both the indigenous low-resource language communities as well as the field linguists working with the natives. It is also through the same dialogue that the low-resource language community and field linguists learn to trust machine translation systems. We would like to continue this process of mutual trust and mutual understanding through continued conversations.

Having closely examined our case study in Quechuan language family, we conclude this thesis in the next chapter.



(a) Margos. (b) Panao. (c) Sihuas.

Figure 8.7: Language rankings by similarity



(a) Margos. (b) Panao. (c) Sihuas.

Figure 8.8: Language rankings by similarity with typological features.

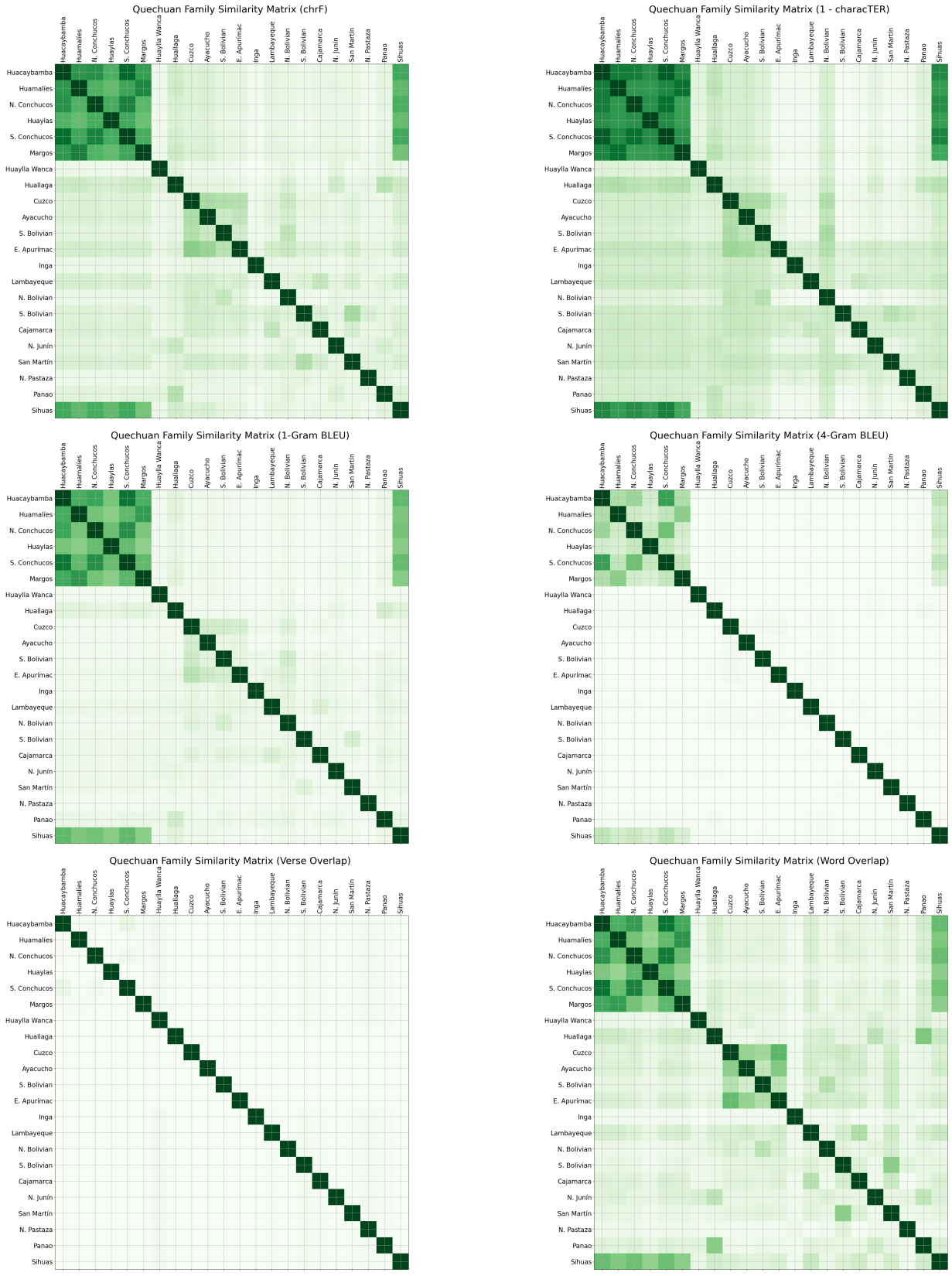


Figure 8.9: Similarity matrix based on chrF, characTER, 1-gram BLEU, 4-gram BLEU, sentence overlap and word overlap.

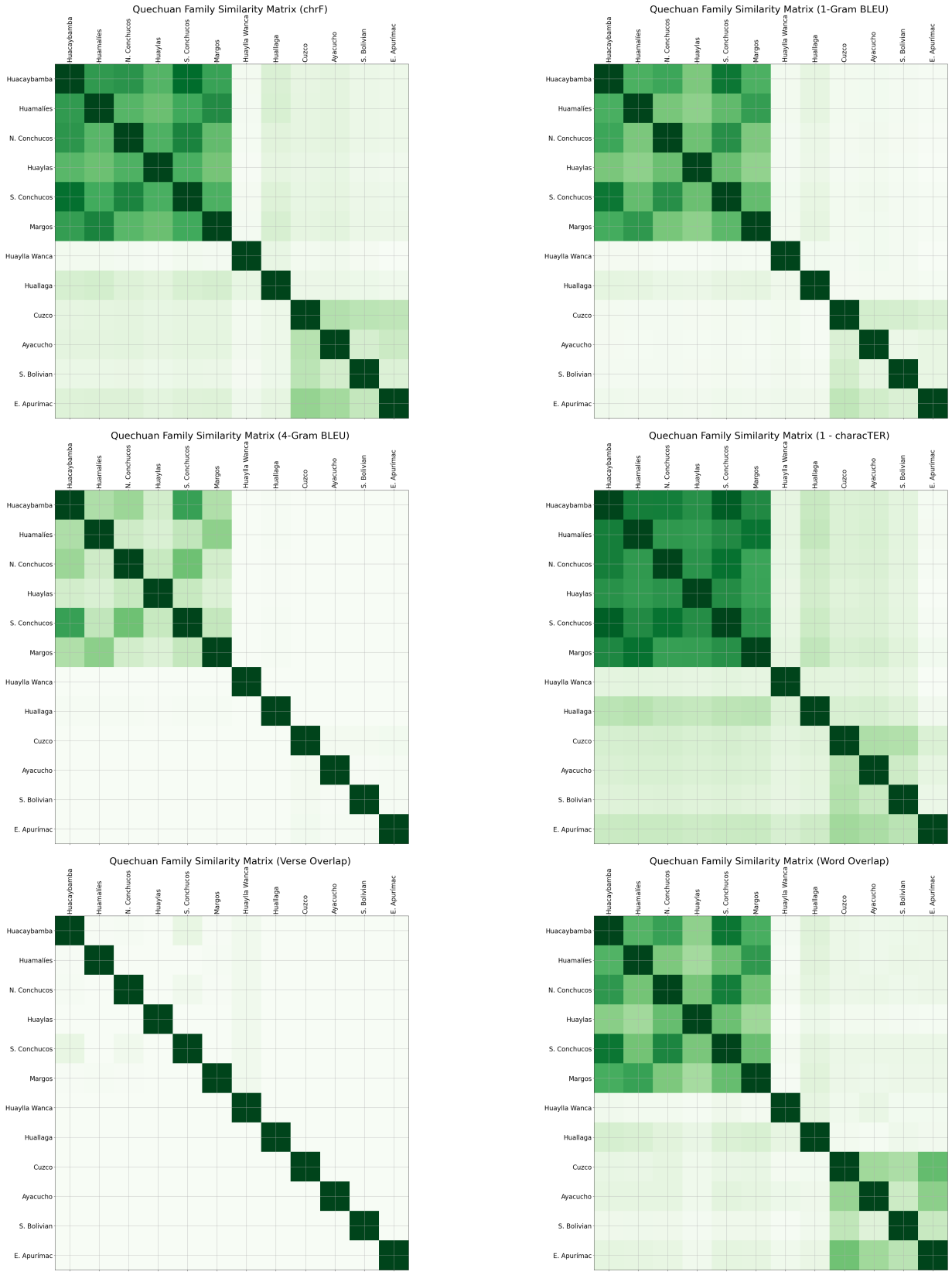


Figure 8.10: Similarity Matrix for 12 Quechuan languages (zooming into 12 main Quechuan languages in Figure 8.9).

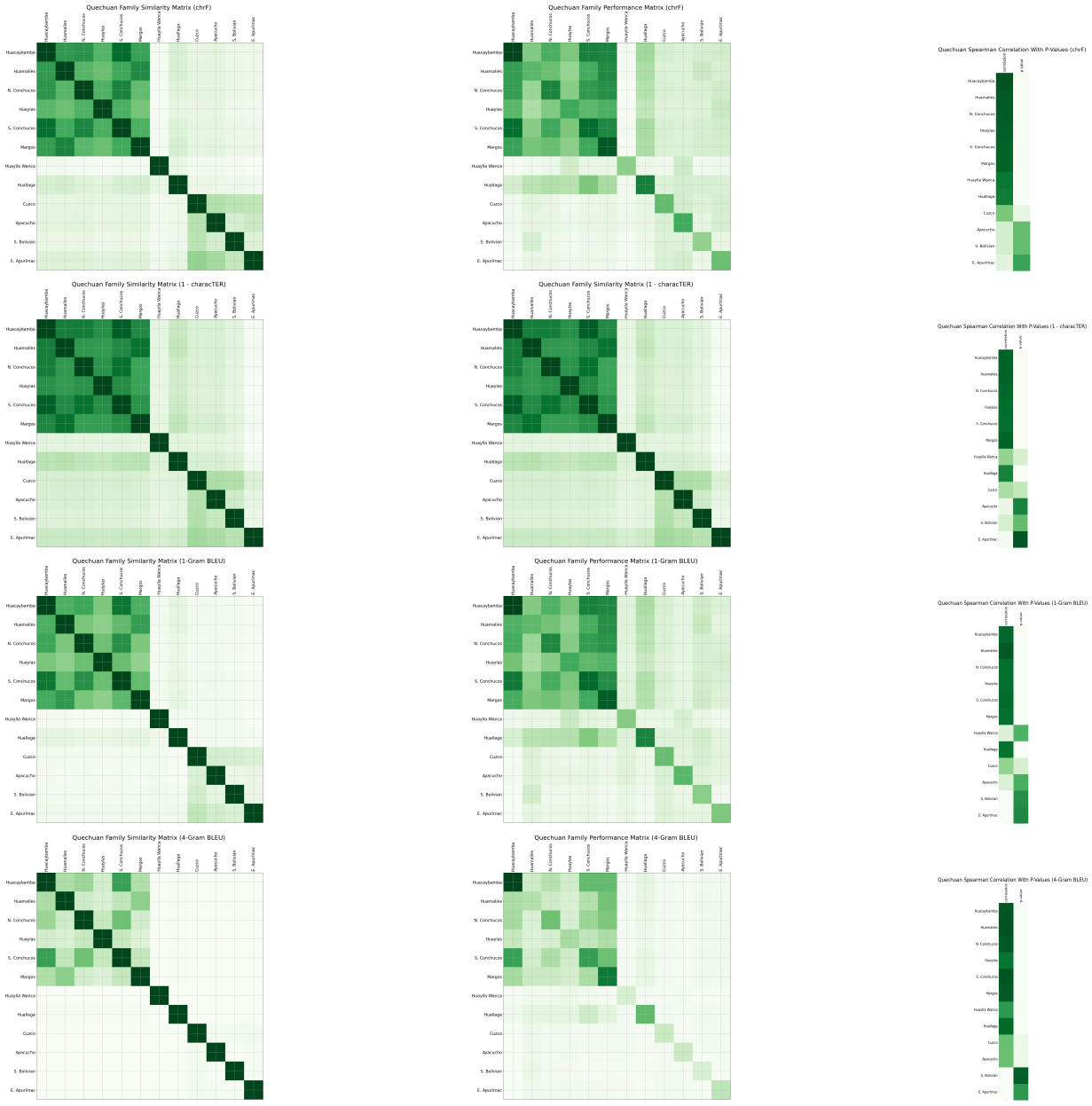


Figure 8.11: Complete comparison similarity and performance matrices using chrF, characTER, 1-gram BLEU and 4-gram BLEU. The last column shows fine-grained correlation and p-value for each source language.

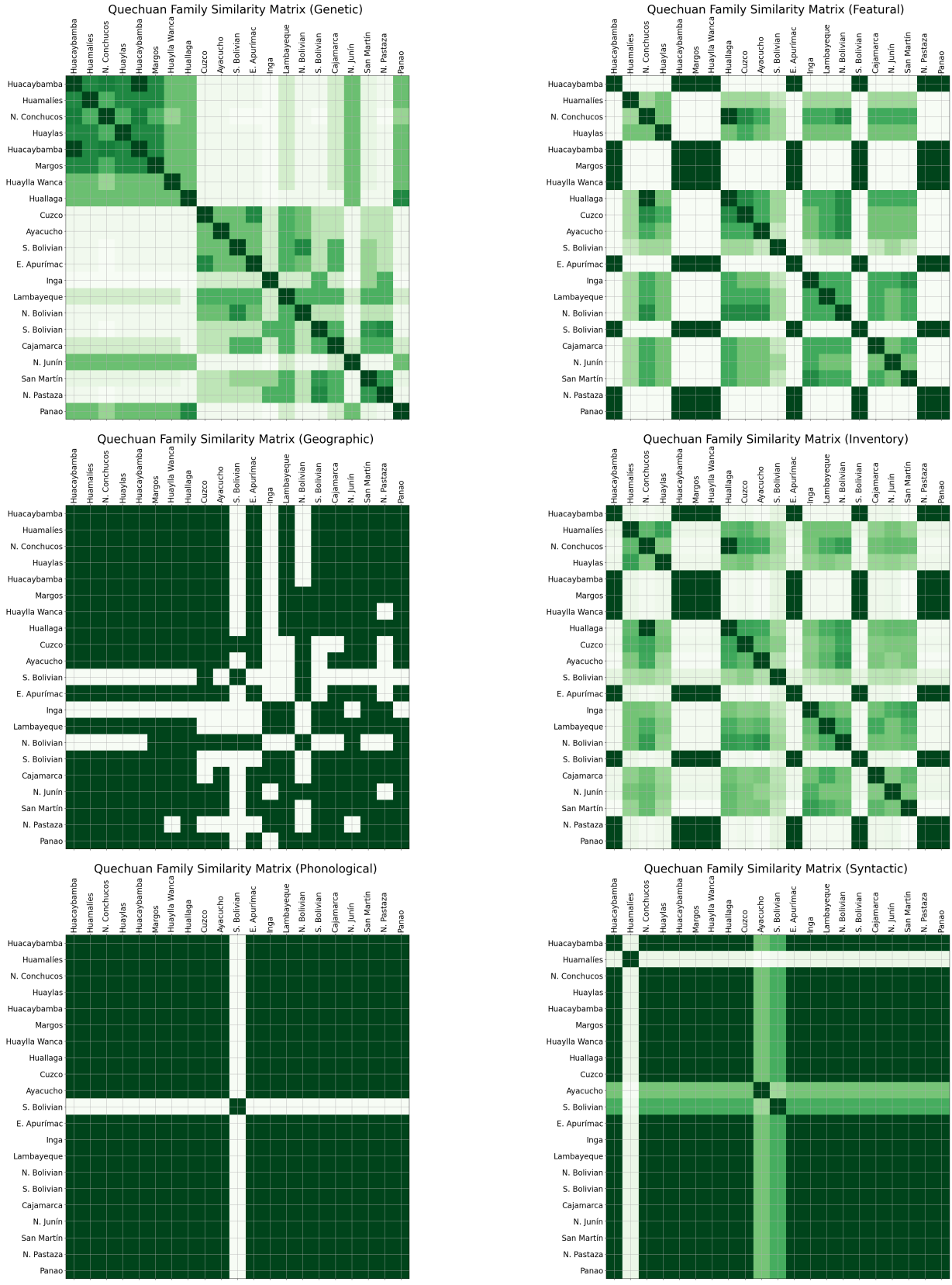


Figure 8.12: Similarity matrix based on genetic, featural, geographic, inventory, phonological and syntactic similarities [179, 188].

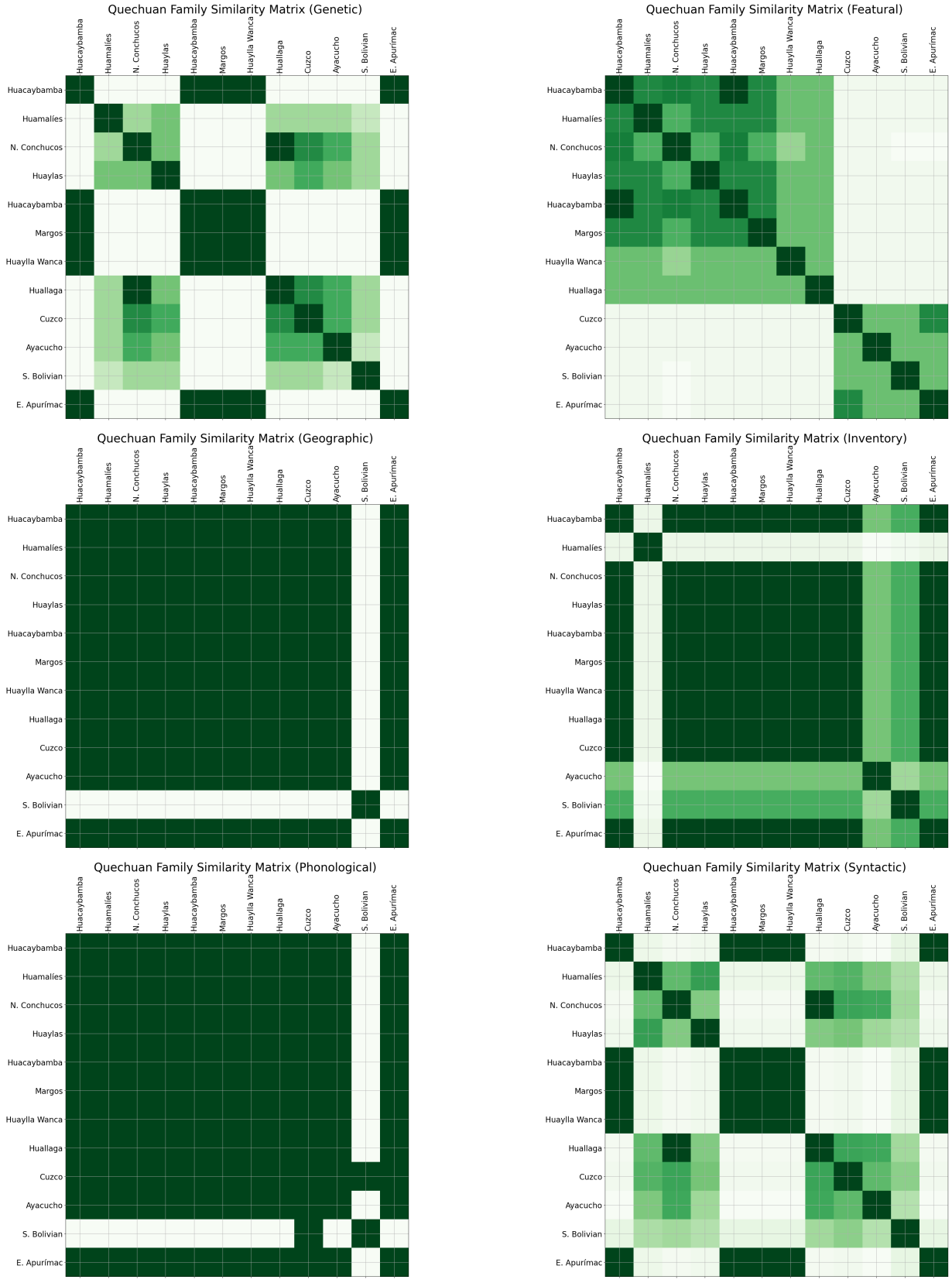


Figure 8.13: Similarity matrix based on genetic, featural, geographic, inventory, phonological and syntactic similarities for 12 Quechuan languages [179, 188].

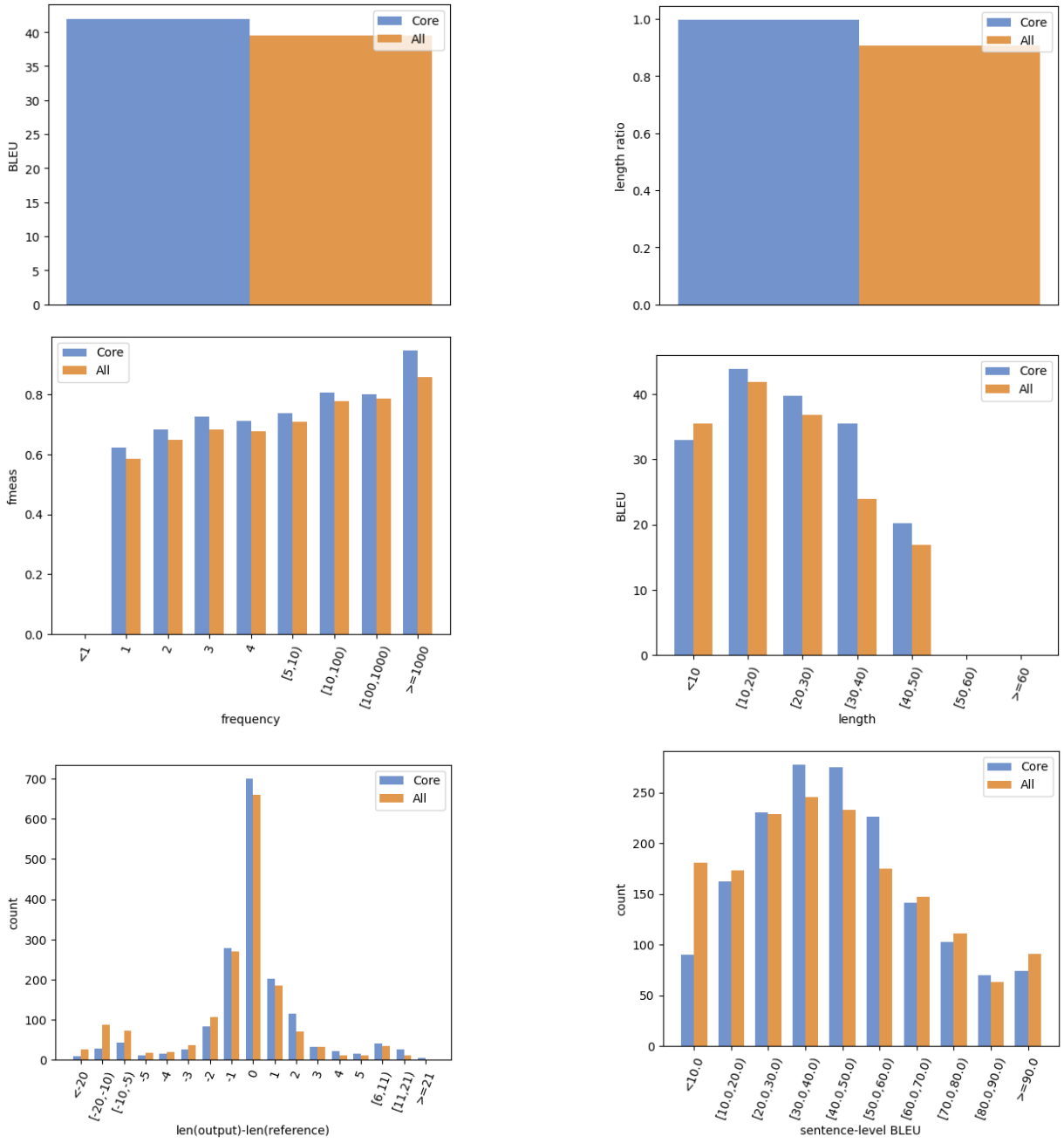


Figure 8.14: Detailed comparison between the system trained on only close languages that have similarity scores above 0.6 chrF (blue) versus the system trained on all (orange) using compare-MT [207].

Chapter 9

Conclusion

"Words travel worlds.
Translators do the driving."

Anna Rusconi

Earlier, we defined our thesis statement as the following:

THESIS STATEMENT *In translating a closed text that is known in advance and available in multiple source languages into a new and severely low-resource language, we argue that generalization to out-of-domain texts is not necessary, but generalization to new languages is necessary. Translation performance gain comes from massive source parallelism by careful choice of close-by language families, style-consistent corpus-level paraphrases within the same language and strategic adaptation of existing large pretrained multilingual models to the domain first and then to the language. Such performance gain makes it possible for machine translation systems to collaborate with human translators to expedite the translation process into new, low-resource languages.*

To conclude, we summarize our contributions and describe how they support our thesis statement. Moreover, we show the limitations of our research and propose different ways to further this work in the future. We also discuss the broader impact of our work from academic research to the real-world machine translation field.

9.1 Summary of Contributions

While we examine all of the following topics in the rest of the thesis, we summarize our contributions to the research community through two main parts: 1.) massively multilingual translation, and 2.) human machine translation as shown in Figure 9.1.

In severely low-resource scenarios, we explore ways to effectively learn from massive source parallelism in Part I. In Part II, we build a human machine translation workflow for

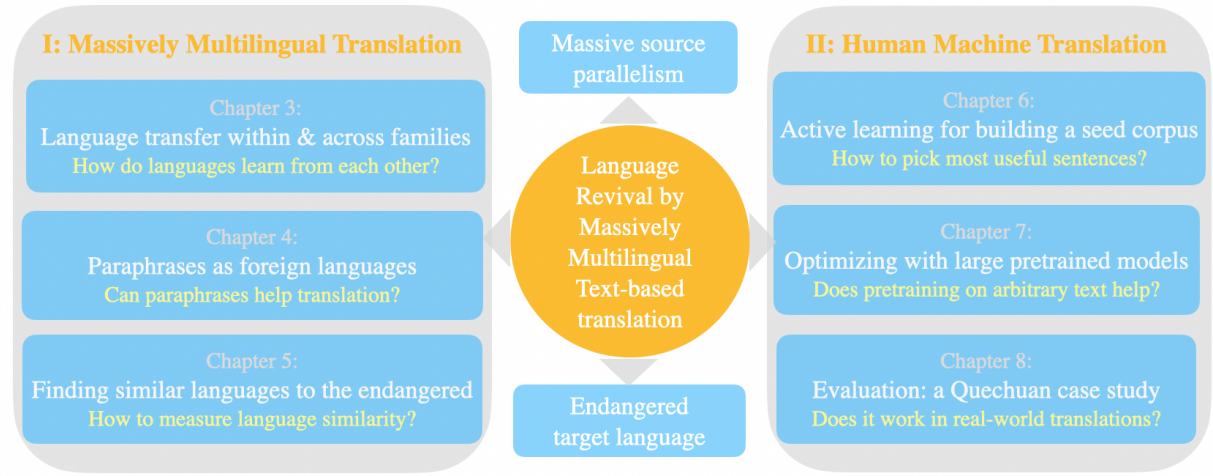


Figure 9.1: Recap of the work done as part of this thesis.

machine translation systems and human translators to work together seamlessly through active learning and large pretrained models. We show proof of concept that it is possible to produce a quality translation draft of the whole text through as little as a few hundred lines (∼3% of the text) of the low-resource data. On top of demonstrating that it is possible to translate using little resource, we build multiple mechanisms to improve effectiveness and accuracy.

9.1.1 Key contributions

As discussed in Chapter 1, our goal is to minimize human translation and post-editing efforts required to generate a full publishable-standard translation of the given text. Ideally we want to hire a large number of human translators, measure and compare the resources (time and money) used to translate the same text into a target low-resource language that does not have any translations of the text under two scenarios: with and without the using this thesis. However, this ideal solution is unrealistic especially in large translation projects. Large translation projects in real-life usually takes decades, if not centuries of work, which is beyond the scope of this thesis. This is why we transform our goal of minimizing human translation efforts required to generate a full translation of the given text into two practical proxy sub-goals as the following:

1. Optimizing and minimizing the amount of sentences to be used to construct seed corpus.
2. Maximizing the quality and utility of MT-generated translation of the full text and optimizing translation efficiency.

The first sub-goal of minimizing the seed corpus serves as a proxy in Chapter 6 to minimize the human translation efforts in the creation of the seed corpus, while the second

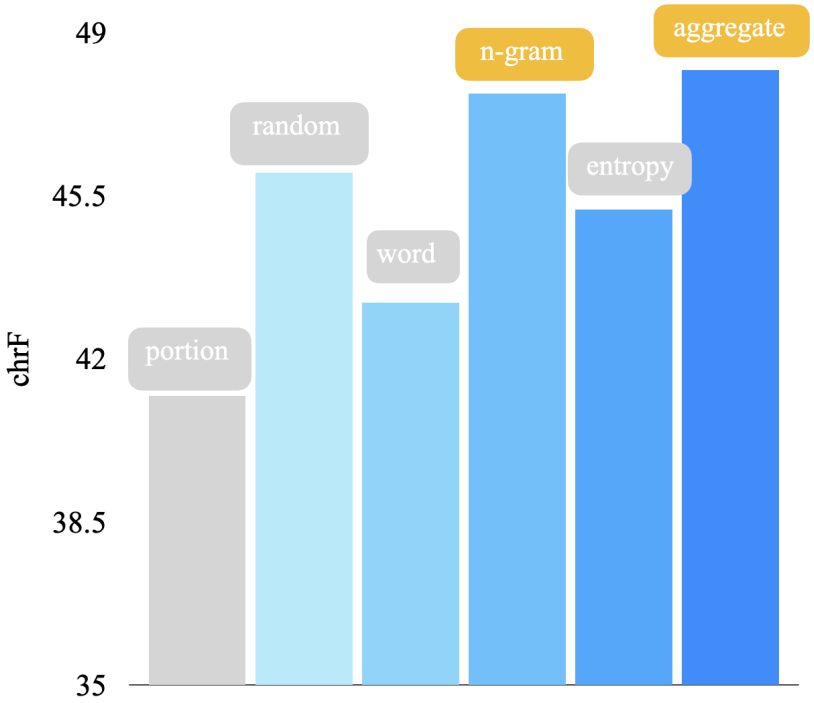


Figure 9.2: Key result of minimizing the amount of sentences to be used to construct seed corpus for translation into Welsh.

sub-goal of maximizing translation performance serves as a proxy in Chapter 7 to minimize human translation efforts in the post-editing process during the subsequent iterations.

To measure translation performance, our primary automatic metric in this thesis is chrF [230]. We choose chrF for accuracy, fluency and expressive power in morphologically-rich languages [217]. We use the metric chrF in Chapter 1 and this chapter of this thesis to motivate and summarize our main contributions of this paper.

Using chrF, we summarize the key contributions based on these two sub-goals in Figure 9.2 and Figure 9.3. Figure 9.2 shows our key results for the first sub-goal of minimizing the seed corpus serves as a proxy in Chapter 6 to minimize the human translation efforts in the creation of the seed corpus. And Figure 9.3 shows the second sub-goal of maximizing translation performance serves as a proxy in Chapter 7 to minimize human translation efforts in the post-editing process during the subsequent iterations.

In Figure 9.2, our main contribution is that we minimize the seed corpus to be ∼3% of the text, and we use ∼3% of the text to translate the ∼97% of the text in the low-resource language we want to translate to. Having minimized the training data to be ∼3% of the text in the given low-resource language, we optimize with active learning on which ∼3% of the text to translate first to produce the seed corpus. To determine which ∼3% of the text, we find that n-gram method (in particular, 4-gram method) is sufficient for producing high translation performance when we do not have access to complete information about

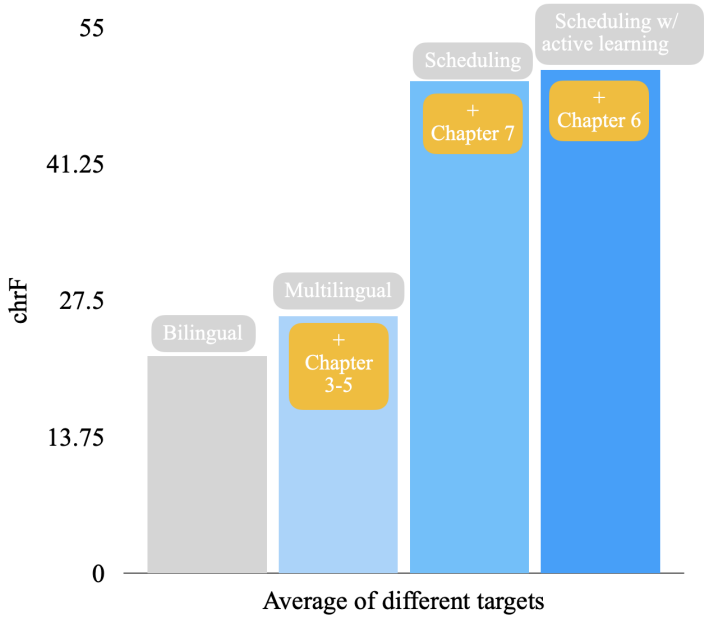


Figure 9.3: Key result of maximizing the quality and utility of MT-generated translation of the full text.

languages close to the low-resource language we want to translate to. However, when we do not have access to such information, we recommend aggregation method proposed in Chapter 6 as a universal ranking to use. This minimizes human translation efforts in the production of the seed corpus. These are the key contributions for our first sub-goal.

Moreover, in Figure 9.3, we further show our key contributions for our second sub-goal. We show the scheduling with large pretrained models using 4-gram method as our active learning method gives the best translation performance. This helps to minimize human post-editing efforts during the subsequent iterations after translation of the seed corpus.

The detailed contributions towards these two sub-goals by different chapters are described in the following two sections. We also show the limitations of this work followed by potential future research to accomplish the main goal beyond this thesis.

9.1.2 Massively Multilingual Translation

In Part I, we explore source parallelism in translation of a given text into new, low-resource languages through massively multilingual training. In Chapter 3, we build cross-lingual transfer both within a given language family and also across different language families. In Chapter 4, we treat paraphrases within the same language as foreign languages, and train on corpus-level paraphrases to improve translation performance. In Chapter 5, we build our own linguistic distance metric based on translation distortion, fertility and performance.

Massively Parallel Intra-family and Inter-family Learning: Our contribution in building cross-lingual transfer, resolving the variable binding problem and producing high quality translations under severely low-resource data scenario is three-fold, extending from multi-source multi-target attentional Neural MT (NMT).

Firstly, to examine intra-family and inter-family influences, we add source and target language family labels in training. Training on multiple families improves BLEU score significantly; moreover, we find training on two neighboring families closest to the low-resource language gives reasonably good BLEU scores.

Secondly, we conduct an ablation study to explore how generalization changes with different amounts of data and find that we only need a small amount of low-resource language data to produce reasonably good BLEU scores. We use full data except for the ablation study.

Finally, to address the variable-binding problem, we build a parallel lexicon table across twenty-three European languages and devise a novel method of order-preserving named entity translation method. Our method works in translation of any text with a fixed set of named entities known in advance. Our goal is to minimize manual labor, but not to fully automate to ensure the correct translation of named entities and their ordering.

Paraphrases as Foreign Languages: We treat paraphrases, rewordings of texts with preserved semantics, as foreign languages, and train a unified NMT model on paraphrase-labeled data with a shared attention in the style of multilingual NMT. Our main findings in harnessing paraphrases in NMT are the following.

Our multi-paraphrase NMT results show significant improvements in BLEU scores over all baselines. In addition, our paraphrase-exploiting NMT uses only two languages, the source and the target languages, and achieves higher BLEUs than the multi-source and multi-target NMT that incorporates more languages.

Furthermore, we find that adding the source paraphrases helps better than adding the target paraphrases, and find that adding paraphrases at both the source and the target sides is better than adding at either side. We also find that adding paraphrases with additional multilingual data yields mixed performance; its performance is better than training on language families alone, but is worse than training on both the source and target paraphrases without language families.

Moreover, adding paraphrases improves the sparsity issue of rare word translation and diversity in lexical choice.

Family of Origin and Family of Choice: Massively Parallel Lexiconized Iterative Pretraining for Severely low-resource Machine Translation: We have five contributions in building customized set of languages that are close to the severely low-resource language.

Firstly, we rank the 124 source languages to determine their closeness to the low-resource language and choose the top few. We call the linguistic definition of language family *Family*

of Origin (FAMO), and we call the empirical definition of higher-ranked languages using our metrics *Family of Choice* (FAMC). They often overlap, but may not coincide.

Secondly, we build an *Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer* (IPML) training on ∼1,000 lines of low-resource data. Using iterative pretraining, we get a +23.9 BLEU increase over a multilingual order-preserving lexiconized transformer baseline (MLc) using English as a hypothetical low-resource language, and a +10.3 BLEU increase over our asymmetric baseline. Training with the low-resource language on both the source and target sides boosts translation into the target side. We have a 42.8 BLEU score for Portuguese-English translation on the medical EMEA dataset.

Thirdly, we use a real-life severely low-resource Mayan language, Eastern Pokomchi, a Class 0 language [148] as one of our experiment setups. In addition, we also use English as a hypothetical low-resource language for easy evaluation.

Fourthly, we also add an order-preserving lexiconized component to translate named entities well. To solve the variable-binding problem to distinguish "Ian calls Yi" from "Yi calls Ian" [95, 109, 322], we build a lexicon table for 2,939 Bible named entities in 124 source languages including more than 66 severely low-resource languages.

Finally, we combine translations from all source languages by using a novel method. For every sentence, we find the translation that is closest to the translation cluster center. The expectation of BLEU score for our combined translation is higher than translation from any of the individuals.

9.1.3 Human Machine Translation

In Part II, having examined source parallelism, we build a human machine translation workflow algorithm for machine translation systems to collaborate with human translators to expedite the process. In Chapter 6, we develop various active learning methods on known languages and transfer ranking to the new, low-resource language. In Chapter 7, we activate the knowledge of large multilingual models by proposing multilingual and multi-stage adaptations through different training schedules; we find that adapting pretrained models to the domain and then to the low-resource language works best. In Chapter 8, we evaluate our work by translating academic progress to the real-world translation process in a case study in Quechuan language family.

Active Learning for Human Machine Translation: We use a random sampling approach to build seed corpora when resources are extremely limited. We recognize that the portion-based translation is very helpful in producing quality translation with formality, cohesion and contextual relevance. Thus, our proposed way is not to replace the portion-based approach, but instead, to get the best of both worlds and to expedite the translation process. The two approaches differ in that the portion-based approach focuses on preserving

coherence of the text locally, while the random-sampling approach focuses on increasing coverage of the text globally.

Moreover, we compare three different ways of incorporating incremental post-edited data during the translation process. We find that self-supervision using the whole translation draft affects performance adversely, and is best to be avoided.

Furthermore, we also show that adding the newly post-edited text to training with vocabulary update performs the best.

Train Global, Tailor Local: Minimalist Multilingual Translation into Low-Resource Languages: We push the limits of random sampling and explore more active learning methods. Our contribution is three-fold.

Firstly, we develop 14 active learning methods on known languages and transfer ranking to the new, low-resource language.

Secondly, we activate the knowledge of large multilingual models by proposing multilingual and multi-stage adaptations through 24 different training schedules; we find that adapting pretrained models to the domain and then to the low-resource language works best.

Thirdly, we aggregate scores from 115 languages to provide a universal ranking and increase robustness by *relaxed memoization* method.

Evaluation: We focus on the case study in Quechuan language family. We find that machine translation performance is significantly positively correlated with language similarity. The more connected a language is, the better it is to translate into this language. Furthermore, we find that decluttering poorly-connected languages improves translation score. Using this result, we demonstrate our methods in translating into a new, low-resource language called Sihuas and achieve high quality translation performance.

9.2 Limitations

Having shown our contributions to the scientific community, we discuss the limitations of our work based on system-level constraints, data-level constraints, task-level constraints, evaluation-level constraints and machine-level constraints.

9.2.1 System-Level Constraints

Incompleteness

From Gödel's incompleteness theorems [105, 106], we understand that there is no all-encompassing "Theory of Everything" that unifies all that is provable and true. Gödel argues that there exists mathematical statements that are undecidable within a formal system, and such systems cannot prove its own consistency. This theory not only shakes

the field of mathematics, physics, philosophy, Computer Science theory, and linguistics, it also offers profound insights into the limitation of Machine Translation.

Applying Gödel's incompleteness theorems, Machine Translation, a formal system, regardless of its depth and complexity, may not be able to solve every translation needs. However, we may still produce models that produce good enough translations that are useful by humans. Indeed, 25 years after Gödel presented incompleteness theorem, Turing asked the question "Can Machines Think?" and paved the road for Machine Translation [293]. Turing didn't propose that machines could think in the same way humans do, but he believed they could simulate intelligent behavior to a point where it could be hard to distinguish human and machines.

Following Gödel's and Turing's thoughts, we could build machine translation systems that produce good enough translations that are useful for communication. However, such machine translation systems, like any formal system, are always incomplete and limited.

End-to-End Learning

In addition to Gödel's incompleteness theorems, we also face limitations particular to end-to-end systems. There are a few known weaknesses of end-to-end systems. For example, most of the end-to-end networks depends on stochastic gradient descent where slow convergence or getting stuck at local optima could be a real problem [104]. Moreover, end-to-end systems suffer from interpretability [300], racial/gender bias [232], composability [274], resource-demanding nature [198] and many other shortcomings [104].

Indeed, our work is limited by limitations inherent to all Machine Translation systems, especially end-to-end learning systems.

Large Pretrained Models

Moreover, our work is also limited to the large pretrained models that are available to us at the time of research for each chapter. In Chapter 7, we use large pretrained models like M2M100. And in Chapter 8, we further use DeltaLM. The method of adapting large pretrained models to the domain and then to the low-resource language is generalizable to other large pretrained models. However, the effect of the translation performance gain we present in this thesis is limited to the large pretrained models.

9.2.2 Data-Level Constraints

On top of limitations at the system level, we are also limited by the following data-level constraints including data representation, data accessibility, and data design decisions.

Data Representation

Our work is limited to the representation of our data in text form. From the extended language similarity measures that we have shown in the previous chapter, even though many Quechuan languages are phonologically close, they may be very far apart in written form. Many local languages did not have a written form until a team of field linguists started the documentation and translation process. The initial team of field of linguists who started the process is therefore pivotal in determining the written form of the given low-resource language. Different teams may have different conventions, methodologies, and beliefs in how to best document and translate a language. Indeed, there are many factors in determining the differences between written forms of two languages even though they might be relatively close in the phonological space.

Without a multi-faceted data representation in many non-textual spaces including the phonological space, our work is limited to the similarities and features machines can learn from the written form. Indeed, our work is limited by the type of data. However, there is a wider range of data beyond text that is helpful for us, that includes audio, video, pictures and different varieties of multi-modal data.

Indeed, our research will benefit from more variety of data that includes audio, and video data that covers wide range of domains.

Data Accessibility

In addition to the limitation of data representation, our work is limited to data accessibility and copyrights. This work can be applied to any text, including large literary text, instruction menu, infectious disease prevention brochures, immigrant welcome booklets. However, most of these texts are copyrighted and not easily accessible. And for those that are more accessible, its translations in multiple source languages are usually non-existent, scattered and not normalized. Training and testing on the Bible dataset is therefore a viable way for testing our methods, as the Bible dataset is relatively accessible and consolidated.

However, even with the Bible dataset, different translations of the Bible carry different copyrights. There are hundreds of completed Bibles, but we only have $\sim$125 that are completed and available to us. A large number of completed Bibles are copyrighted and not available to us. This limits the number of source languages we can train on. It also limits the number target languages that we can test on as the chance of finding similar languages for extremely low-resource and isolated languages in a relatively small dataset is low. Both of the these limit the number of experiments we can do in this space.

Indeed, our work will thrive on more accessible data from a wider range of domains.

Data Design Decisions

In addition to data representation and data accessibility constraints, we also face constraints in our problem setup and data design decisions. In our problem setup, given a text that is multilingually available in many languages, we are interested in translating it into a new, low-resource language. The problem has three unique aspects that are different from traditional MT problems: 1.) our text is closed, not arbitrary 2.) our text has complete translations in all source languages 3.) our text has little to no translation in the target low-resource language.

In this unique setup, we make a few design decisions that suit the purpose of automatic evaluation needs. In designing experiments from Chapter 3-8, for each target low-resource language that already has full translations of the text, we train on ∼3% of the data, and test on ∼97% of the data. Since we are training on all source languages that have complete translations of the text, we indirectly have access to the ∼97% of the data in source languages as well. This is intended in our problem setup because we are translating a closed text. Since machine translation systems and human translators have seen the full text in many existing translations, we intend to encode the meaning of all of the sentences in the text and decode it into the target low-resource language. This design decision is limited by our unique problem setup and is suitable for this task. However, the training and testing setup will be different if we work on open-domain and open text in the future.

9.2.3 Task-Level Constraints

In addition to limitations at the system level and data level, we also face limitations at the task level.

As we have discussed in Chapter 1, ideally we want to measure time taken and money spent for the entire text translation project. However, this is beyond the scope of this thesis. In order to set practical goals for the scope of this thesis, we use two practical proxy sub-goals to transform our goal of minimizing human translation efforts required to generate a full translation of the given text into tangible forms. These two sub-goals are:

1. Optimizing and minimizing the amount of sentences to be used to construct seed corpus.
2. Maximizing the quality and utility of MT-generated translation of the full text and optimizing translation efficiency.

As a result, we are not directly measuring the exact time and money used for completion of the text translation with and without our methods, and we lack realistic meta-data of the entire text translation process to complete the whole text. The translation process with the ideas in this thesis includes seven stages:

1. **Language discovery stage:** This is the hardest stage that researchers do not usually focus on. There are many low-resource language communities that are very isolated from the rest of the world. Some of these low-resource communities are extremely suspicious and hostile towards outsiders [92, 249, 253]. Therefore, it is very important to build a friendly first contact with the low-resource community, and discern whether a language is a new language or a dialect. This is the language discovery process, which may take a long time.
2. **Language learning stage:** This stage is also important as human translators learn and formalize the orthography of the language.
3. **Language literacy stage:** After human translators have learned the language and formalized the orthography, human translators need to teach local communities these language tools if human translators want to involve native speakers in the subsequent translation stages.
4. **Seed corpus translation stage:** This is what we focused on in detail in Chapter 6 on active learning. Machine systems provide different data selection methods for human translators to produce the seed corpus.
5. **Iterative post-editing stage:** This is what we focused on in detail in Chapter 7 on large pretrained models. Machine systems optimize translation drafts for human translators to post-edit on, while human translators provide more post-edited data for machine translation systems to train on. Together, they complete a translation of the whole text.
6. **Quality control stage:** This stage may involve multiple levels of focus groups, including human translators, quality control teams and native low-resource language speakers. The translated text needs to be checked to pass multiple standards before it is released.
7. **Document production stage:** For any translation of a text, the final presentation of the document may include chapter headers, footnotes, paragraph numbers, section headers, titles and many various forms of formatting components. The document production process is therefore much more than the content translation and is paramount in the final publication of the translated text in the low-resource language.

Among the above seven stages of the text translation process, Stage 4 and 5 can be accelerated by models and algorithms introduced in this thesis. However, since the entire text translation process contains five other stages that are beyond the scope of this thesis, we face constraints of time and money estimation of the entire process in the real-world.

9.2.4 Evaluation-Level Constraints

In addition to limitations from the system level, data level and task level, we also face limitations during the evaluation process.

Qualitative Evaluation

On top of automatic evaluation metrics like chrF, characTER that we may use for low-resource languages that are often morphological-rich, our work is limited by human evaluation by native speakers. Most members of the research community do not speak the low-resource languages, and it is difficult to find native speakers and establish long-term collaborations.

There is also a lot of diversity across all low-resource languages. Some are more accessible than others. Hmong and Eastern Pokomchi are harder to assess while Frisian and Welsh, and many Eastern dialects in southern China and Indonesia, are easier to access. These languages could potentially provide easier access and evaluation opportunities.

Empowering and reviving these languages is not just a scientific problem, it requires communication, building trust with field linguists and native speakers and building long-term connections with low-resource language communities. In this thesis, we collaborated with a group of field linguists working on Quechuan language families. And in the future, we are looking to broaden the collaboration to include more human translators, field linguists, and native speakers.

Cultural-Aware Evaluation

Broadening our collaboration with native speakers and human translators is also important for understanding the target low-resource language community culture. This work is limited by the extent of our understanding of the native community culture. Cultural-aware translations and evaluation are key, especially in expressing subtleties in rhetoric [169, 172]. This is especially relevant in non-Western communities where expressions are implicit rather than explicit, and true meanings in communication might hinge on what is not said rather than what is said. If we return to one of our earlier examples, when a host is asking "is your sake cold", what the host is actually saying is that "would you like me to warm it for you". If we translate the sentence without understanding the culture behind this language, we may do a good job translating the sentence verbatim but completely miss the main message that the speaker wants to get across. Therefore, we would like to model culture-specific subtleties in the future work.

9.2.5 Machine-Level Constraints

In addition to the limitations at the system level, data level, task level and the evaluation level, we are also limited to the computing level. We use machines in academic settings. All our research is done with 2 cards of Geforce RTX 1080 Ti, 2 cards of Geforce RTX 2080 Ti and 1 card of RTX 3090. Our limitation is not just on the type of graphic cards (especially the memory), but also on the number of graphic cards.

As we have mentioned in the previous few chapters, in situations when computing power is a constraint, we usually devise parallel methods like *relaxed memoization* to compensate for our limitation. Though clever algorithm and parallel computing could help us to achieve good performance with limited resources, it could only help to a limited extent. In the future, if we have computing power to train a text-specific pretrained large model using our dedicated vocabulary on our own, we could achieve higher performance. Indeed, our research would benefit with more accessible and stronger large scale computing power.

9.3 Future Directions

Given the limitations we face at the system level, data level, evaluation level and computing level, we are interested in pushing our research in the following future directions. These future directions include: overcoming data-level constraints, broadening applications and tasks, improving post-editing user experience, moving beyond limitations on large pretrained models, and overcoming evaluation-level constraints.

9.3.1 Overcoming Data-Level Constraints

Firstly, to push the limitations of data representation, we are interested in venturing into multi-modal learning, and translation into sign languages.

Multi-Modal Learning

Our work will benefit from a multi-faceted data representation in many non-textual spaces including the phonological space. By venturing into speech and visual domain, our work will no longer be limited to similarities and features machines can learn from the written form. When two languages are not close in the written form, they might be close in the phonological space or other non-textual spaces. By representing our data in multiple spaces beyond text is helpful for us to identify and learn from other languages to better translate into a new, low-resource language.

Looking beyond text, there is a wider range of data representations that is helpful for us. It includes audio, video, pictures and different varieties of multi-modal data. Indeed, our research will benefit from more variety of data that covers wider range of domains.

Sign Language

Many researchers have worked on sign languages. Translation into sign languages in the low-resource language communities is a very meaningful and important work. For example, researchers find that a substantial portion of children in a native Nicaraguan community are genetically inclined to be born deaf; and these deaf children learn and create sign languages [199, 232, 256, 257, 258]. Translating into these low-resource languages where many children are deaf is therefore centered on translating into sign languages. Many researchers have ventured into sign language [153, 282], this is indeed a promising research direction.

Real-life Data Encoding

Working with real-life severely low resource languages presents a series of challenges that we may not expect. For example, among the languages that have complete Bible translations in our dataset, we have more than 66 real-life severely low-resource languages. Many of such languages have "latin-1" encoding rather than "utf-8" encoding. And many of them do not even have "latin-1" encoding. However, most transformer platforms work with only "utf-8" encoding. Preprocessing takes a lot of effort. In the future, when we expand to other datasets or other source languages, we will face more situations similar to this and we will overcome them.

9.3.2 Broadening Applications and Tasks

Wider Range of Applications

On top of exploring more diverse data representation, we also would like to explore a wider range of applications and datasets. In addition to the datasets that are used in this thesis, we are interested in collaborating with more diverse teams to translate medical and literary texts, including infectious disease prevention brochures, immigrant welcoming booklet, instruction menus, large literary texts, movie scripts, and song lyrics. These texts are helpful for low-resource language communities to understand and improve their own welfare and healthcare and communication with the outside world.

Our method can be applied to the medical and healthcare domains which is immensely valuable to the low-resource communities. In Chapter 5, we have applied our method to the European medical dataset EMEA and achieved high translation performance [321]. EMEA dataset is built by the European Medicines Agency and has a lot of medical information that may be beneficial to the low-resource communities. However, existing dataset in EMEA is limited to the European languages. Unlike European languages, a lot of severely low-resource languages do not receive a lot of attention and have extremely limited budgets in curating such useful datasets. In the future, if we can work with human translators to curate multilingual dataset that includes many low-resource languages in the medical

domain, this will be very helpful. One specific use case is the translation of COVID-19 guidelines. In an event of a global pandemic, many low-resource communities are affected and there is immense value in creating a multilingual healthcare and medical dataset as in the COVID-19 case. With such dataset, we could help to translate into many low-resource communities. Indeed, use cases of our model in the healthcare and medical domains has significant value.

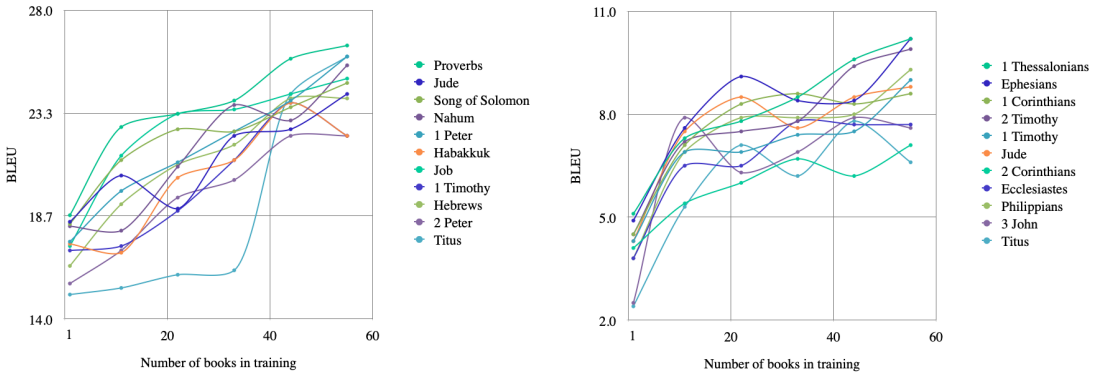
In addition to the medical and healthcare domains, we are also keen in exploring the education domain in all subject areas and levels that are relevant to the low-resource communities. There are many textbooks and course materials that are in English, and many are not in low-resource languages. There is tremendous value in translating and providing access for such texts to help the low-resource communities to learn and flourish. This is especially relevant in language preservation and revival for the young generations. A low-resource language will have a better chance at flourishing when young children are learning and speaking the low-resource languages rather than English. Indeed, there is immense value in translating textbooks and educational materials into low-resource languages in al subject areas and levels that are relevant.

Moreover, another example of future work that is very close to our research community is improving communication between mainstream rich-resource language communities and immigrants/refugees. Pittsburgh, like numerous European and North American cities, is increasingly welcoming a growing number of refugees and immigrants from various countries as the world becomes more globalized. While a lot of channels have been established to help these refugees and immigrants, most of these channels are in English. It would be much more helpful if such channels are in their own languages. For example, translating driving manuals into their languages would immensely help them to transition into cities like Pittsburgh. This not only helps the elderly population among these refugees and immigrants who may not speak or understand English, but also helps those that do to feel home. This will increase inclusivity and diversity.

Comprehensive Text Translation Tasks

As we have discussed in Section 9.2.3, there are at least seven stages in the text translation process, among which 5 that we have not focused on in this thesis.

1. **Language discovery stage:** Building the friendly first contact, language discovery, and the discernment process to determine whether a language is a new language or a dialect, is very important.
2. **Language learning stage:** Human translators learns and formalizes the orthography of the language.
3. **Language literacy stage:** Human translators teach local communities to involve native speakers in the subsequent translation stages.



(a) English (b) Eastern Pokomchi

Figure 9.4: Performance of the most difficult 11 books with increasing number of training books.

4. **Seed corpus translation stage:** This is what we focused in detail in Chapter 6 on active learning.
5. **Iterative post-editing stage:** This is what we focused in detail in Chapter 7 on large pretrained models.
6. **Quality control stage:** This stage may involve multiple levels of focus groups, including human translators, quality control teams and native low-resource language speakers.
7. **Document production stage:** For any translation of a text, the final presentation of the document may include chapter headers, footnotes, paragraph numbers, section headers, titles and many various forms of formatting components, and is important.

This thesis accelerates Stage 4 and 5. In the future, we are interested in building the entire seven stages to complete full text translation process. This requires many number of years and resources, and is a very important process.

9.3.3 Improving Post-Editing User Experience

In addition to exploring wider range of applications and datasets, we could also work with human translators in the post-editing process to improve translation. There are a number of ways that the machine translation system could optimize post-edits and use post-edits to expedite translation.

Iterative Post-Editing

As discussed in Section 6.4.1, we show simulation of the iterative post-editing process for translation into English and Eastern Pokomchi in Figure 9.4a and 9.4b. After producing the first draft of the text by training a MT system on the seed corpus, we hold out the most

difficult 11 books (the worst-performing 11 books) and set them aside as the test set for evaluating the entire iterative post-editing process. Taking the most difficult 11 books as the held-out test set, we divide the other 55 books of the Bible into 5 portions to simulate 5 iterations of post-editing process. Using this setup, we use Schedule B in 5 iterations with increasing number of post-edited portions. Each portion contains 11 books, serving as post-edited portion for each iteration. In each iteration, we simulate human post-editing process by adding the actual translation of the given text portion to the MT system. MT system produces better and better drafts and we show the improvement using the most difficult 11 books.

This simulation gives an understanding of the iterative post-editing, however, it is a simulation. Going beyond simulation, we are interested in measuring and recording the real post-editing process by working with real-life text translation process. In real-life text translation process, there is coordination between different teams, translation stages. These coordination may influence the translation improvement over the entire iterative process, and cannot be simulated. In the future, we look forward to see more work between teams in completing the iterative post-editing process in real-life.

Iterative Learning From Post-Edits

The key of our research on Human Machine Translation is the close collaboration between human translators and machine translation. In our work, once human translators receive the machine translated draft, they post-edits a portion of the text. The machine then take this newly post-edited portions and add to the training data, and therefore produce a better draft. Iterations of post-editing and training improve and expedite translation.

In addition to using post-edited data as additional training data, we could also learn from post-edits so that human translators do not need to repeat the same edits. For example, in the case of Sihuas Quechua, the correct spelling is "chaymi" instead of the machine translated "tsaymi". The edits involves changing "ts" to "ch". We could automate edits like this to save time for the human translators so that they do not have to repeat their edits.

Comprehensive Document Production

In addition to learning from post-edits from human translators, machine systems could produce a comprehensive document translation rather than content translation. For any translation of a text, the final presentation of the document may include chapter headers, footnotes, paragraph numbers, section headers, titles and many various forms of formatting components. The document production process is therefore much more than the content translation. In this thesis, all of the input data are stripped of the formatting information including titles, headers, and footnotes. Consequently, training with such inputs produces outputs that also does not contain such information. In real-world document translation

productions, these formatting issues are important. Therefore, we could aim to produce a comprehensive document translation in the future.

To produce a comprehensive document translation, we could train on the completely labeled document with all titles, headers and footnotes as inputs and aim to produce outputs that contain such information. However, while some title and headers are well aligned, footnotes may not be. Indeed, including such information may render data alignment and structure incomplete. Incomplete data alignment and structure may affect the performance of our multilingual training. Additionally, neural network beyond structured data is still an active area of research [28, 263, 279]. Therefore, directly adding formatting information to training may not be the best solution.

Instead of directly adding formatting information to training, we could add a pre-processing stage and a post-processing stage where we create an external data structure for the given document, and align titles and headers. For these aligned titles and headers, they could follow the same translation mechanism as the main content of the document. For footnotes, we can create a partial aligned data for training for the best possible translation. This method has a strong potential of maintaining accurate structured prediction.

Furthermore, there are a lot of success in industry where companies have successfully produced comprehensive document translations. In the future, we are open to collaborate with industry leaders to learn and to better our system so as to create a comprehensive document production process for human translators.

Adapting to Different Writing Styles

In addition to comprehensive document production, our work could benefit from adapting to more varied writing styles. We notice the writing style of the test data is not uniform. For example, the text in the book of Mark usually starts with conjunctives. “en”, a Dutch conjunctive word, is carried over to the translation output as almost every test sentence begins with conjunctives. This is a unique writing style used by the author Mark, but is not shared with other test data which is written by different authors.

When the writing styles of the test data differs from the training data, machine translation systems encounter challenges. We would like to improve the translation performance of testing on entirely different text with different writing styles. This is another way to minimize the post-editing efforts and improve user experience for human translators.

9.3.4 Moving Beyond Pretrained Models Limits

As we have discussed in our limitations, the performance gain in this thesis is limited by the large pretrained models we use. However, our finding of adapting large pretrained models to the domain and then to the language is universal and generalizable to other tasks, domains and large pretrained models.

Current and Future Large Language Models

Based on this thesis, in the future we want to focus on adapting or fine-tuning a current or future state-of-the-art multilingual Large Language Model to the task of translating a known text into a new, and low-resource target language. This task is largely three-fold:

1. Learning vocabulary from the target low-resource language.
2. Learning to generate grammatical and coherent text in the target low-resource language.
3. Learning to translate appropriate meaningful content into the target low-resource language.

For the first sub-task, we could benefit from collaboration with human translators and working with any available monolingual data in the given new, low-resource language. In the case of Sihuas Quechua that we studied in Chapter 8, our human translators provides us a list of morphemes in Sihuas Quechua that they learned during the language discovery process before we work on text translation. This information is very valuable. In addition, if there is any existing monolingual data in the given new, low-resource language, it will contribute greatly to discovery of vocabulary and morphemes. There are large number of low-resource languages that are morphologically rich, having monolingual data is not just helpful for learning word vocabulary but also helpful for learning sub-word level morphemes. Indeed, morpheme-related research in machine translation is difficult but important. From our research, we find that if we replace the vocabulary from the BPE-based Large Pretrained Models (which all the Large Pretrained Models present currently are) with morpheme-based vocabulary, the results are very poor. In the future, this problem may be solved in multiple ways including training our own domain-specific pretrained models with morpheme-based vocabulary in the target new, low-resource language, which we will discuss in the next section. Morpheme-level research is a very important future direction.

Moreover, for the second and third sub-task, they could be tackled jointly. There is much we could do based on the result of this thesis. Our method of adapting to the domain and then to the language can still be used in today's swiftly-evolving and fast-growing field of Large Language Models. In the future, we are interested in working with current and future large pretrained models and perform multi-stage adaptation to the domain first and then to the new, low-resource language.

Resolving these three sub-tasks in the future is very important. Additionally, we also would like to discuss training our own pretrained models suited to our task.

Training Domain-Specific Pretrained Models

In addition, we will also benefit from training our own pretrained model in the future. If we have computing power to train a domain-specific pretrained large model on all languages

Book	chrF	characTER	4-gram BLEU	1-gram BLEU	Number of Lines
Esther	0.868	0.142	57.4	83.2	167
1 Chronicles	0.864	0.133	60.9	84.2	958
Haggai	0.862	0.111	59.9	83.2	38
2 Chronicles	0.859	0.14	57.1	81.7	824
Joshua	0.85	0.154	56.2	81.6	658
2 Kings	0.85	0.157	57.0	82.0	719
Ezra	0.843	0.143	58.9	82.6	280
Jeremiah	0.84	0.141	54.3	79.2	1365
1 Kings	0.835	0.165	58.0	81.7	831
Habakkuk	0.834	0.149	48.7	75.8	56

Table 9.1: Top 10 ranked Old Testament books translating into Quechua Margos.

in our dataset (∼145 languages) using our dedicated vocabulary, we could achieve higher performance. If our dataset grows and we could train on more languages like a few key companies, we could contribute to the community by translating the text into any given language much more easily. When we train more future experiments, we want to work with more massively parallel systems. Indeed, we will benefit from more computing power and training our own domain-specific pretrained models.

9.3.5 Overcoming Evaluation-Level Constraints

More Fine-Grained Evaluation

In addition to improving post-editing user experience for human translators, we are interested in more fine-grained evaluation. In this work, we have used a few different automatic evaluation metrics. In addition to the varying evaluation metrics, there are many different ways to create a more fine-grained evaluation mechanism.

To build a more fine-grained evaluation system, we first need to understand that our test data is not uniform as we have discussed before. For example, the Bible has 66 books, covering different topics, genres and writing styles. The book of "Psalms" is poetry-based while Paul's letters like "Romans" are very philosophical. So far, we have looked at the the entire text as a whole. We would like to work with more fine-grained translation evaluation on all 66 books of the Bible. In Table 9.1, we show the top 20 performing Old Testament books when we test on the Bible based on a small portion of Margos data. Choosing books that are best helped by Machine Translation could inform human translators the sequence of post-editing steps. We have explored this briefly in Chapter 5. In the future, we would like to work with human translators iteration-by-iteration as rankings produced by each iteration differs from each another.

Continued Dialogue with Native Speakers

In translations into low-resource languages, oftentimes we as researchers do not speak these languages. This makes qualitative evaluation really hard. We may be able to understand the named entities. For example, in the case of translation into Eastern Pokomchi, we can read some of the named entities "Jesús", "Galilea", "Simón" and "Andres" in the machine translated text "Eh noq ojik i rub'an i Jesús juntar i k'isa palaw i Galilea, xrilow reje i Simón ruch'ihil i Andres, re' i rutuut i k'isa palaw, ruum jinaj i k'isa palaw barco". However, we cannot read and speak Eastern Pokomchi which is a Mayan language and we do not know anybody who speaks it. If we can work with native speakers of Eastern Pokomchi, it would greatly help with qualitative evaluation.

There are continued conversations with field linguists and human translators working in the field. Understanding their needs is pivotal in building a long-term collaborative relationship that benefits both sides. This process is a continued dialogue. It is through this continue dialogue that we respect the dignity of both the indigenous low-resource language communities as well as the field linguists working in the field, while they learn to trust machine translation systems through time. We would like to continue this process of mutual trust and mutual understanding of both sides through continued conversations.

9.4 Broader Impact

Our work is done with the intention of reviving and empowering low-resource languages, and more importantly, reviving the communities who speak low-resource languages. We want to lift up people and communities who are otherwise not in the spotlight of the world's attention. Lifting up low-resource language communities with this work, we aim to bring diversity, inclusivity, and equity to different language communities. And we want to provide NLP solutions that are inclusive and accessible to people across the world.

This could bring voices to low-resource language communities, understanding needs of the elderly population among the low-resource language speakers, dissemination of infectious disease prevention information, welcoming immigrants and refugees from low-resource language communities with books in their own language, educating deaf children in the low-resource language communities, communicating and bridging the low-resource communities with the world through translating large literary texts, movies and songs into the low-resource language.

9.5 Key takeaways

We have a few key takeaways from this thesis. Given data scarcity in the low-resource scenarios, we explore ways to effectively learn from massive source parallelism in Part I.

We then build a human machine translation workflow for machine translation systems and human translators to work together seamlessly through active learning and large pretrained models in Part II. We show proof of concept that it is possible to produce quality translation draft of the whole text through as little as a few hundred lines (∼3% of the text) of the low-resource data. In addition to demonstrate that it is possible to translate using little resource, we show various ways to improve effectiveness and accuracy.

Firstly, in Part I, we examine how source parallelism benefit translation of a given text into new, low-resource languages through massively multilingual training. In Chapter 3, we build cross-lingual transfer both within a given language family and also across different language families; we showed that training with two close-by families typically builds sufficiently good cross-lingual transfer in multilingual training. We also propose an order-preserving lexiconized machine translation model to resolve the variable binding problem and producing high quality lexiconized translations under severely low-resource scenarios. In Chapter 4, we treat paraphrases within the same language as foreign languages, and train on corpus-level paraphrases to improve translation performance. We find that our multi-paraphrase translation models improve performance better than multilingual models and improves the sparsity issue of rare word translation as well as diversity in lexical choice. In Chapter 5, we build our own linguistic distance metric based on translation distortion, fertility and performance. We propose a method, *Iteratively Pretrained Multilingual Order-preserving Lexiconized Transformer* (IPML), to train on low-resource language data. We push the limit by using only ∼1,000 lines (∼3.5% of the entire text) to translate the whole text and achieved good translation performance using IPML.

In Part II, having examined source parallelism, we build a human machine translation workflow algorithm for machine translation systems to collaborate with human translators to expedite the process. In Chapter 6, we first develop various active learning methods on known languages and transfer ranking to the new, low-resource language. Secondly, we activate the knowledge of large multilingual models by proposing multilingual and multi-stage adaptations through different training schedules in Chapter 7; we find that adapting pretrained models to the domain and then to the low-resource language works best. Thirdly, we aggregate scores from 115 languages to provide a universal ranking and increase robustness by *relaxed memoization* method. In Chapter 8, having examined both source parallelism and human machine translation workflow, we evaluate our work by translating academic progress to the real-world translation process in a case study in Quechuan language family. We collaborate extensively with a translation group with in-depth knowledge of various Quechuan languages and focus on evaluation. We find that machine translation performance is significantly positively correlated with language similarity. The more connected a language is, the better it is to translate into this language. Furthermore, decluttering poorly-connected languages improves translation performance. Using this finding, we show our results in translating into a new, low-resource language called Sihuas Quechua.

Bibliography

[1] Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 937–947, 2017. 2.1.1

[2] Ife Adebara, El Moatez Billah Nagoudi, and Muhammad Abdul Mageed. Translating similar languages: Role of mutual intelligibility in multilingual transformers. *Proceedings of the 5th Conference on Machine Translation*, 2020. 5.2.2

[3] Willem FH Adelaar. Quechua i y quechua ii: En defensa de una distinción establecida. *Revista Brasileira de Linguística Antropológica*, 5(1):45–65, 2013. 8.2.1

[4] Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *Proceedings of the 18th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies9*, 2019. 2.2.1

[5] Judith Aissen, Nora C England, and Roberto Zavala Maldonado. *The Mayan languages*. Taylor & Francis, 2017. 5.4, 5.5, 6.4.1

[6] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the 17th Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. 2.2.1, 4.1, 4.2.2

[7] Uri Alon, Frank F Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. Neuro-symbolic language modeling with automaton-augmented retrieval. *Proceedings of the 39th International Conference on Machine Learning*, 2022. 6.2.2

[8] Vamshi Ambati. *Active learning and crowdsourcing for machine translation in low resource scenarios*. PhD thesis, Carnegie Mellon University, 2012. 2.3.2

[9] Vamshi Ambati, Stephan Vogel, and Jaime G Carbonell. Multi-strategy approaches to active learning for statistical machine translation. In *Proceedings of the 13th Biennial Machine Translation Summit*, 2011. 2.3.2, 6.2.2

[10] Ulrich Ammon. *The dominance of English as a language of science: Effects on other languages and language communities*, volume 84. Walter de Gruyter, 2001. 3.3.2

[11] Dimitra Anastasiou and Reinhard Schäler. Translating vital information: Localisation, internationalisation, and globalisation. *Syn-thèses Journal*, 3:11–25, 2010. 1, 3.1

[12] Antonios Anastasopoulos, Sameer Bansal, David Chiang, Sharon Goldwater, and Adam Lopez. Spoken term discovery for language documentation using translations. In *Proceedings of the Workshop on Speech-Centric Natural Language Processing*, pages 53–58, 2017. 2.1.1, 3.1

[13] Philip Arthur, Graham Neubig, and Satoshi Nakamura. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 21st Conference on Empirical Methods in Natural Language Processing*, 2016. 3.2.2

[14] Peter Austin and Andrew Simpson. *Endangered languages*, volume 14. Buske Verlag, 2007. 2.1.2

[15] Peter K Austin and Julia Sallabank. *The Cambridge handbook of endangered languages*. Cambridge University Press, 2011. 2.1.2, 7.1

[16] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 16th Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, 2011. 3.3.2

[17] Wilker Aziz, Sheila Castilho, and Lucia Specia. PET: a tool for post-editing and assessing machine translation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3982–3987, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2012/pdf/985_Paper.pdf`. 2.3.3

[18] Rafael E Banchs and Marta R Costa-Jussà. A semantic feature for statistical machine translation. In *Proceedings of the 5th workshop on syntax, semantics and structure in statistical translation*, pages 126–134. Association for Computational Linguistics, 2011. 3.4.1

[19] Sameer Bansal, Herman Kamper, Sharon Goldwater, and Adam Lopez. Weakly supervised spoken term discovery using cross-lingual side information. In *Acoustics, Speech and Signal Processing*, pages 5760–5764. IEEE, 2017. 2.1.1

[20] Julia Barrett. Support and information needs of older and disabled older people in the uk. *Applied ergonomics*, 36(2):177–183, 2005. 3.1

[21] Regina Barzilay and Kathleen R McKeown. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 50–57. Association for Computational Linguistics, 2001. 4.2.1

[22] Christos Baziotis, Barry Haddow, and Alexandra Birch. Language model prior for low-resource neural machine translation. *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 2.2.3, 5.2.1

[23] Stephen Beale, Sergei Nirenburg, Marjorie McShane, and Tod Allman. Document authoring the bible for minority language translation. *Proceedings of 10th Biennial Machine Translation Summit*, 2005. 3.4.1

[24] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 6.4.2

[25] Luisa Bentivogli, Nicola Bertoldi, Mauro Cettolo, Marcello Federico, Matteo Negri, and Marco Turchi. On the evaluation of adaptive machine translation for human post-editing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(2):388–399, 2015. 1.3

[26] Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Federmann Christian. Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Workshop on Spoken Language Translation (IWSLT 2018)*, pages 62–69, 2018. 1.3

[27] Alexandre Bérard, Zae Myung Kim, Vassilina Nikoulina, Eunjeong Lucy Park, and Matthias Gallé. A multilingual neural machine translation model for biomedical data. In *Proceedings of the 1st Workshop on NLP for COVID-19 at the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 2.1.1

[28] Monica Bianchini, Giovanna Maria Dimitri, Marco Maggini, and Franco Scarselli. Deep neural networks for structured data. *Computational intelligence for pattern recognition*, pages 29–51, 2018. 9.3.3

[29] Alexandra Birch, Miles Osborne, and Philipp Koehn. Predicting success in machine translation. In *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*, pages 745–754, 2008. 5.2.1

[30] Steven Bird. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, 2020. 2.1.1, 5.1, 6.4.1

[31] Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/P04-3031. 13, 6.2.2

[32] Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. 2.2.3, 6.2.2

[33] Michael Bloodgood and Chris Callison-Burch. Bucking the trend: Large-scale cost-focused active learning for statistical machine translation. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010. 1.3, 6.2, 7.2

[34] Frederic Blum, Carlos Barrientos, Adriano Ingunza, and Zoe Poirier. A phylolinguistic classification of the quechua language family. 2023. 8.2.1

[35] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. ISSN 2307-387X. 9

[36] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2.3.3, 8.1

[37] Lynne Bowker. *Computer-aided translation technology: A practical introduction*. University of Ottawa Press, 2002. 2.3.1, 6.1

[38] Lynne Bowker and Des Fisher. Computer-aided translation. *Handbook of translation studies*, 1:60–65, 2010. 2.3.1, 6.1

[39] Florin Brad and Traian Rebedea. Neural paraphrase generation using transfer learning. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 257–261, 2017. 4.2.1

[40] John S Brownstein, Clark C Freifeld, Ben Y Reis, and Kenneth D Mandl. Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the healthmap project. *PLoS medicine*, 5(7):e151, 2008. 2.1.1

[41] Chris Callison-Burch, Colin Bannard, and Josh Schroeder. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 255–262. Association for Computational Linguistics, 2005. 4.2.1

[42] Chris Callison-Burch, Philipp Koehn, and Miles Osborne. Improved statistical machine translation using paraphrases. In *Proceedings of the 6th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 17–24. Association for Computational Linguistics, 2006. 4.1, 4.2.1

[43] Lyle Campbell and Anna Belew. *Cataloguing the world's endangered languages*, volume 711. Routledge New York, USA, 2018. (document), 6.3, 7.1

[44] Michael Carl, Barbara Dragsted, and Arnt Lykke Jakobsen. A taxonomy of human translation styles. *Translation journal*, 16(2):155–168, 2011. 2.3.1, 6.1

[45] Jasone Cenoz. The effect of linguistic distance, l2 status and age on cross-linguistic influence in third language acquisition. *Cross-linguistic influence in 2nd language acquisition: Psycholinguistic perspectives*, 111(45):8–20, 2001. 3.1, 3.3.2, 5.3.2

[46] Sin-wai Chan and David E Pollard. *An Encyclopaedia of Translation: Chinese-English, English-Chinese*. Chinese University Press, 2001. 3.4.1

[47] Guanhua Chen, Shuming Ma, Yun Chen, Li Dong, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. Zero-shot cross-lingual transfer of neural machine translation with multilingual pretrained encoders. *Proceedings of the 26th Conference on Empirical Methods in Natural Language Processing*, 2021. 2.2.3

[48] Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 142–157, 2022. 2.2.3

[49] Barry R Chiswick and Paul W Miller. Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of Multilingual and Multicultural Development*, 26(1):1–11, 2005. 5.2.2

[50] Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. Understanding translationese in multi-view embedding spaces. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6056–6062, 2020. 5.2.2, 8.4.1

[51] Christos Christodouloupoulos and Mark Steedman. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395, 2015. 3.4.1

[52] Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. Reusing a pretrained language model on languages with limited corpora for unsupervised nmt. *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 2.2.3, 5.2.1

[53] Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. Improving multilingual models with language-clustered vocabularies. *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 5.2.2, 8.4.1

[54] Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1693–

1703, 2016. 3.2.1

[55] Kenneth L Clarkson and Peter W Shor. Applications of random sampling in computational geometry, ii. *Discrete & Computational Geometry*, 4(5):387–421, 1989. 2.3.2

[56] Lauren Eby Clemens, Jessica Coon, Pedro Mateo Pedro, Adam Milton Morgan, Maria Polinsky, Gabrielle Tandet, and Matthew Wagers. Ergativity and the complexity of extraction: A view from mayan. *Natural Language & Linguistic Theory*, 33(2):417–467, 2015. 5.4, 5.5, 6.4.1

[57] Father Bernabe Cobo. *History of the Inca Empire: an account of the Indians' customs and their origin, together with a treatise on Inca legends, history, and social institutions*. University of Texas Press, 2010. 8.2.1

[58] Paulo Coelho. *The alchemist*. HarperOne; 25th edition, 2015. (document), 6.1

[59] Richard Oliver Collin. Ethnologue. *Ethnopolitics*, 9(3-4):425–432, 2010. (document), 6.3, 7.1

[60] Bernard Comrie. *The world atlas of language structures*. Oxford University Press, 2005. 5.2.2, 8.4.1

[61] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022. 2.2.2

[62] David Crystal. *Language death*. Cambridge University Press, 2002. 2.1.2, 7.1

[63] Boele De Raad, Marco Perugini, and Zsófia Szirmák. In pursuit of a cross-lingual reference structure of personality traits: Comparisons among five languages. *European Journal of Personality*, 11(3):167–185, 1997. 3.1, 3.3.2, 5.3.2

[64] Antoine de Saint-Exupéry. *El Principito: The Little Prince*. Editorial Verbum, 2019. (document), 6.1

[65] Domingo de Santo Tomás. *Lexicón o vocabulario de la lengua general del Perú*. por Francisco Fernandez de Cordoua, 1951. 8.2.1

[66] Michael Denkowski. Machine translation for human translators. *Unpublished doctoral dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania*, 2015. 2.3.1, 2.3.3, 6.1, 6.5, 8.1

[67] Michael Denkowski, Alon Lavie, Isabel Lacruz, and Chris Dyer. Real time adaptive machine translation for post-editing with cdec and transcenter. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 72–77, 2014. 2.3.3

[68] Donald A DePalma. Language demand and supply. In *The Routledge Handbook of Translation and Globalization*, pages 363–374. Routledge, 2020. 8.1

[69] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, pages 1723–1732, 2015. 2.2.1, 4.1, 4.2.2

[70] Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. Learning to paraphrase for question answering. In *Proceedings of the 22nd Conference on Empirical Methods in Natural Language Processing*, pages 875–886, 2017. 4.2.1

[71] Nancy C Dorian. A response to ladefoged's other view of endangered languages. *Language*, 69(3):575–579, 1993. 2.1.2

[72] Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. Dynamic data selection and weighting for iterative back-translation. *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 2.2.3, 2.3.2, 5.2.1

[73] Haim Dubossarsky, Ivan Vulić, Roi Reichart, and Anna Korhonen. The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, pages 2377–2390, 2020. 5.2.2

[74] Philipp Dufter and Hinrich Schütze. A universal semantic space. *arXiv preprint arXiv:1801.06807*, 2018. 3.4.1

[75] Long Duong, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. An attentional model for speech translation without transcription. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 949–959, 2016. 2.1.1, 3.1

[76] Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, and Trevor Cohn. Multilingual training of crosslingual word embeddings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 894–904, 2017. 3.2.2

[77] Alan Durston. *Pastoral Quechua: the history of Christian translation in colonial Peru, 1550-1654*. University of Notre Dame Pess, 2007. 1.4, 8.2.1, 8.3

[78] Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 12th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 644–648, 2013. 8, 3.3.2, 3.5.3, 5.3.2

[79] Paul S Earle, Daniel C Bowden, and Michelle Guy. Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6), 2012. 2.1.1

[80] Matthew George Easton. *Eastons Bible Dictionary: A Dictionary of Bible Terms*. Thomas Nelson, 1897. 3.5.3

[81] David M Eberhard, Gary F Simons, and Charles D Fennig. *Ethnologue*. SIL International, Global Publishing, 2021. (document), 1.1, 2.1.2, 7.1, 8.2.1

[82] Matthias Eck. *Developing deployable spoken language translation systems given limited resources*. PhD thesis, Karlsruhe Institute of Techonology, 2008. 1.3, 6.1.2, 6.1.2, 6.2, 6.2.2, 6.4.2, 7.2

[83] Matthias Eck, Stephan Vogel, and Alex Waibel. Low cost portability for statistical machine translation based on n-gram frequency and tf-idf. In *International Workshop on Spoken Language Translation*, 2005. 2.3.2, 3.3.2, 6.1.2, 6.2.2

[84] FY Edgeworth. Addendum on" probable errors of frequency-constants". *Journal of the Royal Statistical Society*, 72(1):81–90, 1909. 6.2.2

[85] Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, pages 5960–5969, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.480. URL https://aclanthology.org/2020.emnlp-main.480. 2.2.2, 7.3

[86] Nora C England. *A grammar of Mam, a Mayan language*. University of Texas Press, 2011. 5.4, 5.5, 6.4.1

[87] Anna María Escobar. Spanish in contact with quechua. *The handbook of Hispanic sociolinguistics*, pages 321–352, 2011. 8.2.1

[88] Alexandra Espichán-Linares and Arturo Oncevay-Marcos. Language identification with scarce data: A case study from peru. In *Information Management and Big Data: 4th Annual International Symposium, SIMBig 2017, Lima, Peru, September 4-6, 2017, Revised Selected Papers 4*, pages 90–105. Springer, 2018. 8.3

[89] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*, 2017. 4.2.1

[90] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1608–1618, 2013. 4.1, 4.2.1

[91] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22 (107):1–48, 2021. 2.2.2, 7.2.1, 7.3, 7.3

[92] Jerry L Faught et al. John davis and joseph islands: indigenous missionaries among the creeks in indian territory. *Baptist History and Heritage*, 43(2):32–44, 2008. 1

[93] Marcello Federico. Matecat. In *Proceedings of the 17th Annual conference of the European Association for Machine Translation*, 2014. 2.3.3

[94] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 866–875, 2016. 2.2.1, 2.2.2, 2.2.3, 4.1, 4.2.2, 5.2.1

[95] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71, 1988. 3.1, 3.3.2, 5.1, 5.3.1, 9.1.2

[96] Ariadna Font-Llitjós and Jaime G Carbonell. Automating post-editing to improve mt systems. 2006. 2.3.3

[97] Markus Freitag and Orhan Firat. Complete multilingual neural machine translation. *Proceedings of the 5th Conference on Machine Translation*, 2020. 5.2.1

[98] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 5.3.3

[99] Tze Long Fung et al. COVID-19. `https://en.wikipedia.org/wiki/COVID-19`, 2020. [Online; accessed 24-May-2021]. (document), 6.1

[100] Rashmi Gangadharaiah, Ralf D Brown, and Jaime G Carbonell. Active learning in example-based machine translation. In *Proceedings of the 17th Nordic Conference of Computational Linguistics*, pages 227–230, 2009. 2.3.2

[101] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. Ppdb: The paraphrase database. In *Proceedings of the 12th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 758–764, 2013. 4.1, 4.2.1

[102] Carl Friedrich Gauss. Bestimmung der genauigkeit der beobachtungen. *Abhandlungen zur Methode der kleinsten Quadrate, 1887*, 1816. 6.2.2

[103] Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. Multilingual language processing from bytes. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 1296–1306, 2016. 2.2.1, 3.2.1, 4.1, 4.2.2, 5.2.1

[104] Tobias Glasmachers. Limits of end-to-end learning. In *Asian conference on machine learning*, pages 17–32. PMLR, 2017. 9.2.1

[105] Kurt Gödel. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für mathematik und physik*, 38:173–198, 1931.

9.2.1

[106] Kurt Gödel. *Kurt Gödel: collected works: volume I: publications 1929-1936*, volume 1. Oxford University Press, USA, 1986. 9.2.1

[107] Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254. Association for Computational Linguistics, 2012. 2.3.2

[108] Raymond G Gordon Jr. Ethnologue, languages of the world. *http://www. ethnologue. com/*, 2005. 5.1, 5.5

[109] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014. 3.1, 3.3.2, 5.1, 5.3.1, 9.1.2

[110] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1311–1320. PMLR, 2017. 6.4.2

[111] Lenore A Grenoble and Lindsay J Whaley. *Endangered languages: Language loss and community response*. Cambridge University Press, 1998. 2.1.2

[112] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017. 4.2.1

[113] Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. Universal neural machine translation for extremely low resource languages. *Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, 2018. 5.1, 5.3.2

[114] Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor OK Li. Meta-learning for low-resource neural machine translation. *Proceedings of the 23rd Conference on Empirical Methods in Natural Language Processing*, 2018. 2.3.2

[115] Ana Guerberof. Productivity and quality in mt post-editing. In *Beyond Translation Memories: New Tools for Translators Workshop*, 2009. 2.3.3

[116] Ana Guerberof Arenas. What do professional translators think about post-editing. *JoSTrans The journal of specialised translation*, 19:75–95, 2013. 2.3.3

[117] Thanh-Le Ha, Jan Niehues, and Alexander Waibel. Toward multilingual neural machine translation with universal encoder and decoder. *International Workshop on Spoken Language Translation*, 2016. 2.2.1, 2.2.2, 2.2.3, 2.4.3, 4.1, 4.2.2, 4.3, 4.4.2, 5.2.1, 6.2.1, 7.2.1

[118] Gholamreza Haffari and Anoop Sarkar. Active learning for multilingual statistical machine translation. In *Proceedings of the Joint Conference of the 47th Annual*

Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 181–189, 2009. 2.3.2

[119] Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. Active learning for statistical phrase-based machine translation. In *Proceedings of the 8th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 415–423, 2009. 2.3.2, 6.1.2, 6.2.2

[120] G. Hagen. Grundzüge der wahrscheinlichkeits-rechnung. *Dümmler,Berlin*, 1816. 6.2.2

[121] Jan Hajič. Machine translation of very close languages. In *Proceedings of the 6th Applied Natural Language Processing Conference*, 2000. 5.2.2, 8.4.1

[122] Ken Hale. Endangered languages: On endangered languages and the safeguarding of diversity. *language*, 68(1):1–42, 1992. 2.1.2

[123] Lynne Hansen, Karri Lam, Livia Orikasa, Paul Rama, Geraldine Schwaller, and Ronald Mellado Miller. In the beginning was the word. *Second Language Acquisition Abroad: The LDS Missionary Experience*, 45:89, 2012. 5.2.2, 8.4.1

[124] Rosalind M Harding and Robert R Sokal. Classification of the european language families by genetic distance. *Proceedings of the National Academy of Sciences*, 85(23): 9370–9372, 1988. 3.3.2

[125] Martha J Hardman. Aymara and quechua: languages in contact. In *South American Indian languages: retrospect and prospect*, pages 617–643. University of Texas Press, 1985. 8.4.2

[126] Sadid A Hasan, Bo Liu, Joey Liu, Ashequl Qadir, Kathy Lee, Vivek Datla, Aaditya Prakash, and Oladimeji Farri. Neural clinical paraphrase generation with attention. In *Proceedings of the Clinical Natural Language Processing Workshop*, pages 42–53, 2016. 4.2.1

[127] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. *Advances in neural information processing systems*, 29:820–828, 2016. 2.3.2

[128] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the 6th workshop on Statistical Machine Translation*, pages 187–197, 2011. 12, 6.2.2

[129] Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H Clark, and Philipp Koehn. Scalable modified kneser-ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, 2013. 12

[130] Theo Hermans. Cross-cultural translation studies as thick translation. *Bulletin of the School of Oriental and African Studies*, 66(3):380–389, 2003. 3.1, 3.3.2, 5.3.2

[131] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017. 3.5.2

[132] Felix Hieber, Tobias Domhan, Michael Denkowski, and David Vilar. Sockeye 2: A toolkit for neural machine translation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 457–458, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.50. 3

[133] Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015. 2.3.1, 6.1

[134] RD Hitchcock. Hitchcock's bible names dictionary, art. *AJ Johnson Publishers, New York*, 1874. 3.5.3

[135] Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, 2018. 2.3.2

[136] Eric W Holman, Søren Wichmann, Cecil H Brown, Viveka Velupillai, André Müller, Dik Bakker, et al. Advances in automated language classification. *Quantitative investigations in theoretical linguistics*, pages 40–43, 2008. 5.2.2

[137] Rosaleen Howard. The quechua language in the andes today: Between statistics, the state, and daily life. In *History and language in the Andes*, pages 189–213. Springer, 2011. 1.4, 8.2.1

[138] Junjie Hu and Graham Neubig. Phrase-level active learning for neural machine translation. *arXiv preprint arXiv:2106.11375*, 2021. 2.3.2

[139] Chu-Ren Huang, Laurent Prévot, I-Li Su, and Jia-Fei Hong. Towards a conceptual core for multicultural processing: A multilingual ontology based on the swadesh list. In *International Workshop on Intercultural Collaboration*, pages 17–30. Springer, 2007. 5.2.2, 8.4.1

[140] Victor Hugo. *Les Miserables*. C. Lassalle, 1863. (document), 6.1

[141] John Hutchins. Machine translation and human translation: in competition or in complementation. *International Journal of Translation*, 13(1-2):5–20, 2001. 1.1, 2.3.1

[142] Sagamore Institute. A study of cost and use of funds in bible translation. https://ministrywatch.com/wp-content/uploads/2022/09/Sagamore-Institute-Study-copy-9.12.22.pdf, 2022. [Online; accessed 20-Nov-2023]. 1.3

[143] Mordor Intelligence. Language services market: Global industry trends, share, size, growth, opportunity and forecast 2023-2028. 2023. 8.1

[144] Ingo Eduard Isphording and Sebastian Otten. The costs of b abylon—linguistic distance in applied economics. *Review of International Economics*, 21(2):354–369, 2013. 5.2.2

[145] Karen Ann Jehn. Hapax legomenon ii: Theory, a thesaurus, and word frequency. *CAM*, 5(1):8–10, 1993. 3.1

[146] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-paced curriculum learning. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015. 6.4.2

[147] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017. 2.2.1, 2.2.2, 2.2.3, 2.4.3, 4.1, 4.2.2, 4.3, 5.2.1, 6.2.1, 7.2.1

[148] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. (document), 2.2.3, 2.4.2, 5.1, 6.3, 6.3.2, 6.4.1, 7.1, 7.3, 9.1.2

[149] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017. 9

[150] Alina Karakanta, Jon Dehdari, and Josef van Genabith. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1): 167–189, 2018. 2.2.3

[151] Hans Kellerer, Ulrich Pferschy, David Pisinger, Hans Kellerer, Ulrich Pferschy, and David Pisinger. *Multidimensional knapsack problems*. Springer, 2004. 6.1.1

[152] Maurice G Kendall and B Babington Smith. Randomness and random sampling numbers. *Journal of the royal Statistical Society*, 101(1):147–166, 1938. 2.3.2

[153] Lee Kezar, Riley Carlin, Tejas Srinivasan, Zed Sehyr, Naomi Caselli, and Jesse Thomason. Exploring strategies for modeling sign language phonology. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023. 9.3.1

[154] D. M. Kincade. *The decline of Native Language in Canada*. Stanford University Press, 1991. 2.1.2, 7.1

[155] Kendall A King and Nancy H Hornberger. Quechua as a lingua franca. *Annual Review of Applied Linguistics*, 26:177–196, 2006. 8.2.1

[156] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 5.3.3

[157] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. Opennmt: Open-source toolkit for neural machine translation. *Proceedings of the 55th annual meeting of the Association for Computational Linguistics, System Demonstrations*, pages 67–72, 2017. 2, 3.4.2, 4.4.2, 5.3.1, 6.3.1, 6.3.2, 7.3

[158] Donald E Knuth. *3: 16 Bible texts illuminated*. AR Editions, Inc., 1991. 2.3.2

[159] Tom Kocmi. Exploring benefits of transfer learning in neural machine translation. *arXiv preprint arXiv:2001.01622*, 2020. 2.2.1

[160] Philipp Koehn. A process study of computer-aided translation. *Machine Translation*, 23(4):241–263, 2009. 2.3.1, 6.1

[161] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009. 5.3.2

[162] Philipp Koehn and Barry Haddow. Interactive assistance to human translators using statistical machine translation methods. *Proceedings of the 12th Biennial Machine Translation Summit*, 2009. 2.3.1, 6.1

[163] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180, 2007. 14

[164] Sai Koneru, Danni Liu, and Jan Niehues. Cost-effective training in low-resource neural machine translation. *arXiv preprint arXiv:2201.05700*, 2022. 2.3.2, 6.2.2

[165] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 23rd Conference on Empirical Methods in Natural Language Processing*, page 66–71, 2018. 2.4.3

[166] Peter Ladefoged. Another view of endangered languages. *Language*, 68(4):809–811, 1992. 2.1.2

[167] Peter Nelson Landerman. *Quechua dialects and their classification*. University of California, Los Angeles, 1991. 8.2.1

[168] Laozi. *Dao de jing*. University of California Press, 2019. (document), 6.1

[169] Mildred L Larson. *Meaning-based translation: A guide to cross-language equivalence*. University press of America Lanham, 1984. 4.6, 9.2.4

[170] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, pages 4483–4499, 2020. 2.2.3, 5.2.1, 5.3.2

[171] Jennifer Lee. Frozen, 2013. (document), 6.1

[172] Lori Levin, Donna Gates, Alon Lavie, and Alex Waibel. An interlingua based on domain actions for machine translation of task-oriented dialogues. In *Proceedings of the 5th International Conference on Spoken Language Processing*, 1998. 3.3.2, 3.6, 4.6, 9.2.4

[173] Haiying Li, Arthur C Graesser, and Zhiqiang Cai. Comparison of google translation with human translation. In *The 27th International Flairs Conference*, 2014. 2.3.1, 6.1

[174] Zheng Li, Mukul Kumar, William Headden, Bing Yin, Ying Wei, Yu Zhang, and Qiang Yang. Learn to cross-lingual transfer with meta graph learning across heterogeneous languages. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, pages 2290–2301, 2020. 2.2.3, 5.2.1

[175] Jindřich Libovickỳ and Jindřich Helcl. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 196–202, 2017. 2.2.1

[176] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. Choosing transfer languages for cross-lingual learning. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 5.2.1

[177] Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. Pre-training multilingual neural machine translation by leveraging alignment information. *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 2.2.3, 5.1, 5.2.1, 5.3.2, 7.1

[178] Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. Character-based neural machine translation. *Proceedings of the 4th International Conference on Learning Representations*, 2016. 2.2.1, 3.2.1

[179] Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, 2017. (document), 15, 8.12, 8.13

[180] Ming Liu, Wray Buntine, and Gholamreza Haffari. Learning to actively learn neural machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 334–344, 2018. 2.3.2

[181] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*, 2020. 2.2.2

[182] TechNavio (Infiniti Research Ltd.). Global language services market 2022-2026. 2022. 8.1

[183] Minh-Thang Luong, Ilya Sutskever, Quoc V Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, 2015. 5.2.1

[184] Aurolyn Luykx, Fernando García Rivera, and Félix Julca Guerrero. Communicative strategies across quechua languages. *International Journal of the Sociology of language*, 2016(240):159–191, 2016. 1.4, 8.2.1

[185] Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. DeltaLM: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. 2021. 2.2.2, 8.4.2

[186] Nitin Madnani, Joel Tetreault, and Martin Chodorow. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 11th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 182–190. Association for Computational Linguistics, 2012. 4.2.1

[187] Luisa Maffi. Endangered languages, endangered knowledge. *International Social Science Journal*, 54(173):385–393, 2002. 2.1.2

[188] Chaitanya Malaviya, Graham Neubig, and Patrick Littell. Learning language representations for typology prediction. *Proceedings of the 22nd Conference on Empirical Methods in Natural Language Processing*, 2017. (document), 15, 8.12, 8.13

[189] Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 881–893, 2017. 4.2.1

[190] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics*,

pages 55–60, 2014. 3.5.3

[191] Gary Marcus. Deep learning is hitting a wall. *Nautilus, Accessed*, pages 03–11, 2022. 2.3.3, 8.1

[192] Hellen Mardaga. Hapax legomena: a neglected field in biblical studies. *Currents in Biblical Research*, 10(2):264–274, 2012. 3.1

[193] Thomas L Markey. *Frisian*, volume 13. Walter de Gruyter, 2011. 6.3.2, 7.3

[194] Lorena Guerra Martínez. *Human Translation Versuv Machine Translation and Full Post-editing of Raw Machine Translation Output*. Citeseer, 2003. 1.1, 2.3.1

[195] Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level machine translation: Methods and evaluation. *ACM Computing Surveys*, 2019. 6.5

[196] Thomas Mayer and Michael Cysouw. Creating a massively parallel bible corpus. *Oceania*, 135(273):40, 2014. (document), 2.4.2, 3.4.1, 4.4.1, 5.4, 6.1, 6.3.1, 6.3.2, 7.3

[197] Akiva Miura, Graham Neubig, Michael Paul, and Satoshi Nakamura. Selecting syntactic, non-redundant segments in active learning for machine translation. In *Proceedings of the the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 20–29, 2016. 2.3.2

[198] Joss Moorkens. Ethics and machine translation. *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 18:121, 2022. 9.2.1

[199] Gary Morgan and Judy Kegl. Nicaraguan sign language and theory of mind: The issue of critical periods and abilities. *Journal of Child Psychology and Psychiatry*, 47 (8):811–819, 2006. 9.3.1

[200] Christopher Moseley. *Encyclopedia of the world's endangered languages*. Routledge, 2008. 2.1.2

[201] Moses Mulumba, Juliana Nantaba, Claire E Brolan, Ana Lorena Ruano, Katie Brooker, and Rachel Hammonds. Perceptions and experiences of access to public healthcare by people with disabilities and older people in uganda. *International journal for equity in health*, 13(1):1–9, 2014. 1

[202] Víctor Muntés Mulero, Patricia Paladini Adell, Cristina España Bonet, and Lluís Màrquez Villodre. Context-aware machine translation for software localization. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation: EAMT 2012: Trento, Italy, May 28th-30th 2012*, pages 77–80, 2012. 6.5

[203] Martijn Naaijer and Dirk Roorda. Parallel texts in the hebrew bible, new methods and visualizations. *Young*, 140:157, 1993. 3.4.1

[204] Shashi Narayan, Claire Gardent, Shay Cohen, and Anastasia Shimorina. Split and rephrase. In *Proceedings of the 22nd Conference on Empirical Methods in Natural*

Language Processing, pages 617–627, 2017. 4.1, 4.2.1

[205] Orville James Nave. *Nave's Topical Bible: A Digest of the Holy Scriptures*. Topical Bible Publishing Company, 1903. 3.5.3

[206] Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. *Proceedings of the 23rd Conference on Empirical Methods in Natural Language Processing*, 2018. 2.2.3

[207] Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, and Xinyi Wang. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 18th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 35–41, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-4007. URL https://aclanthology.org/N19-4007. (document), 7, 8.6, 8.14

[208] Toan Q Nguyen and David Chiang. Improving lexical choice in neural machine translation. *arXiv preprint arXiv:1710.01329*, 2017. 3.2.2

[209] Farhad Nooralahzadeh, Giannis Bekoulis, Johannes Bjerva, and Isabelle Augenstein. Zero-shot cross-lingual transfer with meta learning. *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 2.2.3, 5.2.1

[210] Sharon O'Brien and Joss Moorkens. Towards intelligent post-editing interfaces. 2014. 2.3.3

[211] Terence Odlin. *Language transfer: Cross-linguistic influence in language learning*. Cambridge University Press, 1989. 3.1, 3.3.2, 5.3.2

[212] Arturo Oncevay, Barry Haddow, and Alexandra Birch. Bridging linguistic typology and multilingual machine translation with multi-view language representations. *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 5.2.2, 8.4.1

[213] Robert Östling and Jörg Tiedemann. Neural machine translation for low-resource languages. *arXiv preprint arXiv:1708.05729*, 2017. 2.1.1

[214] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 20th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, 2019. 1

[215] Santanu Pal, Marcos Zampieri, Sudip Kumar Naskar, Tapas Nayak, Mihaela Vela, and Josef van Genabith. Catalog online: Porting a post-editing tool to the web. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 599–604, 2016. 2.3.3

[216] Bo Pang, Kevin Knight, and Daniel Marcu. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the 3rd Conference of North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, 2003. 4.2.1

[217] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 1.3, 2.4.4, 3.4.2, 4.4.2, 6.2.4, 6.3.1, 6.3.2, 7.2.3, 7.3, 8.4.1, 9.1.1

[218] Steve Parker. Sonority distance vs. sonority dispersion–a typological survey. *The sonority controversy*, 18:101–165, 2012. 5.2.2

[219] Doris L Payne. Endangered languages: Current issues and future prospects, 1999. 2.1.2

[220] Alvaro Peris and Francisco Casacuberta. Active learning for interactive neural machine translation of data streams. *Proceedings of the 23rd Conference on Computational Natural Language Learning*, 2018. 2.3.2

[221] Robyn Perry and Steven Bird. Treasure language storytelling: Cross-cultural language recognition and wellbeing. *Proceedings of the 5th International Conference on Language Documentation and Conservation*, 2017. 1, 3.1

[222] Filippo Petroni and Maurizio Serva. Language distance and tree reconstruction. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(08):P08012, 2008. 3.3.2

[223] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 2.2.3, 5.2.1, 5.3.2

[224] Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alex Waibel. Improving zero-shot translation with language-independent constraints. *Proceedings of the 4th conference on Machine Translation*, 2019. 2.2.3

[225] Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, pages 4465–4470, 2020. 2.2.3

[226] Manfred Pienemann, Bruno Di Biase, Satomi Kawaguchi, and Gisela Håkansson. Processability, typological distance and l1 transfer. *Cross-linguistic aspects of Processability Theory*, pages 85–116, 2005. 5.2.2, 8.4.1

[227] David Pisinger. A minimal algorithm for the 0-1 knapsack problem. *Operations Research*, 45(5):758–767, 1997. 6.1.1

[228] Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M Mitchell. Competence-based curriculum learning for neural machine translation. *arXiv preprint arXiv:1903.09848*, 2019. 2.3.2

[229] Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15, 2020. 1.1, 2.3.1, 6.1

[230] Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, 2015. 1.3, 2.4.4, 6.2.4, 7.2.3, 8.4.1, 9.1.1

[231] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the 3rd Conference on Machine Translation*, pages 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319. 1.3, 2.4.4, 6.2.4, 7.2.3

[232] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381, 2020. 9.2.1, 9.3.1

[233] Lutz Prechelt. Early stopping-but when? *Neural Networks: Tricks of the trade*, pages 553–553, 1998. 3.4.2

[234] Ye Qi, Devendra Singh Sachan, Matthieu Felix, Sarguna Janani Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? *Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, 2018. 2.2.3, 5.1, 7.1

[235] Chris Quirk, Chris Brockett, and William Dolan. Monolingual machine translation for paraphrase generation. In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing*, 2004. 4.2.1

[236] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 7.4.1

[237] Taraka Rama and Prasanth Kolachina. How good are typological distances for determining genealogical relationships among languages? In *Proceedings of the 30th International Conference on Computational Linguistics*, pages 975–984, 2012. 5.2.2, 8.4.1

[238] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *Proceedings of the 4th International Conference on Learning Representations*, 2016. 2.1.1

[239] B Reddy, Yadlapalli S Kusuma, Chandrakant S Pandav, Anil Kumar Goswami, Anand Krishnan, et al. Water and sanitation hygiene practices for under-five children among households of sugali tribe of chittoor district, andhra pradesh, india. *Journal of environmental and public health*, 2017. 1, 3.1

[240] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213. 5

[241] Ricardo Rei, Ana C Farinha, Craig Stewart, Luisa Coheur, and Alon Lavie. Mt-telescope: An interactive platform for contrastive evaluation of mt systems. 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 73–80, 2021. 1.3, 6, 2.4.4, 6.2.4, 7.2.3

[242] Philip Resnik, Mari Broman Olsen, and Mona Diab. The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1-2): 129–153, 1999. 3.4.1

[243] Edwin W. Rice. *People's Dictionary of the Bible*. Forgotten Books, 2015. 3.5.3

[244] Annette Rios, Mathias Müller, and Rico Sennrich. Subword segmentation and a single bridge language affect zero-shot neural machine translation. *Proceedings of the 5th conference on Machine Translation*, 2020. 2.2.3

[245] Suzanne Romaine. Preserving endangered languages. *Language and Linguistics Compass*, 1(1-2):115–132, 2007. 2.1.2

[246] Malcolm Ross et al. Language families and linguistic diversity. In *Encyclopedia of Language and Linguistics*. Elsevier, 2 edition, 2006. 3.3.2

[247] JK Rowling. Harry potter. *The 100 Greatest Literary Characters*, page 183, 2019. (document), 6.1

[248] Pramod Salunkhe, Aniket D Kadam, Shashank Joshi, Shuhas Patil, Devendrasingh Thakore, and Shrikant Jadhav. Hybrid machine translation for english to marathi: A research evaluation in machine translation:(hybrid translator). In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 924–931. IEEE, 2016. 6.5

[249] Jane Samson. Translation teams: missionaries, islanders, and the reduction of language in the pacific. In *Critical Readings in the History of Christian Mission*, pages 704–721. Brill, 2021. 1

[250] Edward Sapir. How languages influence each other. *Language: an Introduction to the Study of Speech*, 1921. 3.1, 3.3.2, 5.3.2

[251] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874, 2021. 2.3.1, 6.1

[252] Kevin P Scannell. Machine translation for closely related language pairs. In *Proceedings of the Workshop Strategies for developing machine translation for minority languages*, pages 103–109. Citeseer, 2006. 3.4.1

[253] Michael Schönhuth. Dead missionaries, wild sentinelese: An anthropological review of a global media event. *Anthropology Today*, 35(4):3–6, 2019. 1

[254] Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.507. URL https://aclanthology.org/2021.acl-long.507. 2.2.2, 7.3

[255] Yuuki Sekizawa, Tomoyuki Kajiwara, and Mamoru Komachi. Improving japanese-to-english neural machine translation by paraphrasing the target language. In *Proceedings of the 4th Workshop on Asian Translation*, pages 64–69, 2017. 4.1, 4.2.1

[256] Ann Senghas. *Children's contribution to the birth of Nicaraguan Sign Language*. PhD thesis, Massachusetts Institute of Technology, 1995. 9.3.1

[257] Ann Senghas and Marie Coppola. Children creating language: How nicaraguan sign language acquired a spatial grammar. *Psychological science*, 12(4):323–328, 2001. 9.3.1

[258] Richard J Senghas, Ann Senghas, and Jennie E Pyers. The emergence of nicaraguan sign language: Questions of development, acquisition, and evolution. *Biology and knowledge revisited: From neurogenesis to psychogenesis*, pages 287–306, 2005. 9.3.1

[259] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009. 2.3.2

[260] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, 2016. 2.2.1, 11, 2.4.3, 3.2.1, 6.3.2, 7.3

[261] Maurizio Serva and Filippo Petroni. Indo-european languages tree by levenshtein distance. *EPL (Europhysics Letters)*, 81(6):68005, 2008. 5.2.2

[262] Burr Settles. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012. 2.3.2, 6.1

[263] Lei Sha, Lili Mou, Tianyu Liu, Pascal Poupart, Sujian Li, Baobao Chang, and Zhifang Sui. Order-planning neural text generation from structured data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 9.3.3

[264] Claude E Shannon. Prediction and entropy of printed english. *Bell Labs Technical Journal*, 30(1):50–64, 1951. 4.4.2, 4.5

[265] Chung-ling Shih. Re-looking into machine translation errors and post-editing strategies in a changing high-tech context. *Compilation & Translation Review*, 14(2), 2021. 1.3

[266] Philippa Shoemark, Sharon Goldwater, James Kirby, and Rik Sarkar. Towards robust cross-linguistic comparisons of phonological networks. In *Proceedings of the 14th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 110–120, 2016. 3.1, 3.3.2, 5.3.2

[267] Gary F Simons and Charles D Fennig. *Ethnologue: languages of Asia*. sil International Dallas, 2017. 5.1

[268] Karla Smith and Terry Smith. Panao runakuna pintashun. 2009. 8.3

[269] William Smith, Francis Nathan Peloubet, and Mary Abby Thaxter Peloubet. *Smith's Bible Dictionary*. Pyramid Books, 1967. 3.5.3

[270] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, 2006. 2.3.3

[271] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the 1st Conference on Machine Translation*, volume 2, pages 543–553, 2016. 3.1, 3.3.2, 5.3.2

[272] Matthias Sperber, Graham Neubig, Satoshi Nakamura, and Alex Waibel. Optimizing computer-assisted transcription quality with iterative user interfaces. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*

(LREC'16), pages 1986–1992, 2016. 2.3.3

[273] Matthias Sperber, Graham Neubig, Jan Niehues, Satoshi Nakamura, and Alex Waibel. Transcribing against time. *Speech communication*, 93:20–30, 2017. 6.5

[274] Biplav Srivastava and Francesca Rossi. Towards composable bias rating of ai services. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 284–289, 2018. 9.2.1

[275] Craig Stewart, Ricardo Rei, Catarina Farinha, and Alon Lavie. COMET - deploying a new state-of-the-art MT evaluation metric in production. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 78–109, Virtual, October 2020. Association for Machine Translation in the Americas. URL https://aclanthology.org/2020.amta-user.4. 1.3, 2.4.4, 6.2.4, 7.2.3

[276] Yui Suzuki, Tomoyuki Kajiwara, and Mamoru Komachi. Building a non-trivial paraphrase corpus using multiple machine translation systems. In *Proceedings of ACL 2017, Student Research Workshop*, pages 36–42, 2017. 4.2.1

[277] Agneta M-L Svalberg and Hjh Fatimah Bte Hj Awg Chuchu. Are english and malay worlds apart? typological distance and the learning of tense and aspect concepts. *International Journal of Applied Linguistics*, 8(1):27–60, 1998. 5.2.2, 8.4.1

[278] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*, 2020. 2.2.2

[279] Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Budrionis, and Fred Godtliebsen. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(6):e1549, 2021. 9.3.3

[280] Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. Exploring document-level literary machine translation with parallel paragraphs from world literature. *Proceedings of the 27th Conference on Empirical Methods in Natural Language Processing*, 2022. 8.1

[281] Nisha Thampi, Yves Longtin, Alexandra Peters, Didier Pittet, and Katie Overy. It's in our hands: a rapid, international initiative to translate a hand hygiene song during the covid-19 pandemic. *Journal of Hospital Infection*, 105(3):574–576, 2020. (document), 3.1, 3.2, 6.1

[282] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020. 9.3.1

[283] Sarah G Thomason. *Endangered languages*. Cambridge University Press, 2015. 2.1.2

[284] Brian Thompson, Huda Khayrallah, Antonios Anastasopoulos, Arya D McCarthy, Kevin Duh, Rebecca Marvin, Paul McNamee, Jeremy Gwinnup, Tim Anderson, and Philipp Koehn. Freezing subnetworks to analyze domain adaptation in neural machine translation. *Proceedings of the 3rd Conference on Machine Translation*, 2018. 2.2.3, 5.2.1

[285] Jörg Tiedemann. Character-based pivot translation for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 141–151. Association for Computational Linguistics, 2012. 3.2.1

[286] Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218, 2012. 5.4

[287] John Ronald Reuel Tolkien. *The Lord of the Rings: One Volume*. Houghton Mifflin Harcourt, 2012. (document), 6.1

[288] Katrin Tomanek and Udo Hahn. Semi-supervised active learning for sequence labeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1039–1047, 2009. 1.3, 6.2, 7.2

[289] Antonio Toral and Andy Way. *What Level of Quality can Neural Machine Translation Attain on Literary Text?*, chapter Translation Quality Assessment: From Principles to Practice. Springer, 2018. 3.1, 3.3.2, 5.3.2

[290] Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. Polyglot neural language models: A case study in cross-lingual phonetic representation learning. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 1357–1366, 2016. 2.2.1, 4.1, 4.2.2

[291] Marco Turchi, Tijl De Bie, and Nello Cristianini. Learning performance of a machine translation system: a statistical and computational analysis. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pages 35–43, 2008. 3.5.2

[292] Marco Turchi, Matteo Negri, M Farajian, and Marcello Federico. Continuous learning from human post-edits for neural machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):233–244, 2017. 1.3

[293] Alan M Turing. Can a machine think. *The world of mathematics*, 4:2099–2123, 1956. 9.2.1

[294] Thomas Turino. Quechua and aymara. In *The Garland Encyclopedia of World Music*, pages 205–224. Routledge, 2017. 8.2.1

[295] Index Translationum UNESCO. World bibliography of translation, 1932. (document), 6.1

[296] Jaap Van der Meer. At last translation automation becomes a reality: an anthology of the translation market. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, resources and tools for building MT*, 2003. 1.1

[297] Rik van Gijn and PC Muysken. Highland-lowland language relations. 2020. 8.3

[298] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5.3.1

[299] Lucas Nunes Vieira and Lucia Specia. A review of translation tools from a post-editing perspective. In *Proceedings of the Third Joint EM+/CNGL Workshop Bringing MT to the User: Research Meets Translators (JEC 2011)*, pages 33–42, 2011. 2.3.3

[300] Warren J von Eschenbach. Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, 34(4):1607–1622, 2021. 9.2.1

[301] Thuy Vu and Gholamreza Haffari. Automatic post-editing of machine translation: A neural programmer-interpreter approach. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3048–3053, 2018. 2.3.3

[302] Alex Waibel. Communicating in a multilingual world. `http://www.speech.kth.se/higgins/references/waibel.shtml`, 2016. [Online; accessed Dec 16, 2017]. 2.1.1

[303] Alex Waibel and Christian Fugen. Spoken language translation. *IEEE Signal Processing Magazine*, 25(3):70–79, 2008. 2.1.1

[304] Alex Waibel, Hidefumi Sawai, and Kiyohiro Shikano. Modularity and scaling in large phonemic neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(12):1888–1898, 1989. 2.1.1

[305] Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. Character: Translation edit rate on character level. In *Proceedings of the 1st Conference on Machine Translation*, pages 505–510, 2016. 1.3, 2.3.3, 2.4.4, 6.2.4, 7.2.3, 8.4.1

[306] Xinyi Wang, Sebastian Ruder, and Graham Neubig. Expanding pretrained models to thousands more languages via lexicon-based adaptation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. 2.2.3

[307] Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. Sogou neural machine translation systems for wmt17. In *Proceedings of the 2nd Conference on Machine Translation*, pages 410–415,

2017. 3.2.2, 3.5.3

[308] Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. On negative interference in multilingual models: Findings and a meta-learning treatment. *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 2.2.3, 5.2.1

[309] Rebecca Webster, Margot Fonteyne, Arda Tezcan, Lieve Macken, and Joke Daems. Gutenberg goes neural: Comparing features of dutch human translations with raw neural machine translation outputs in a corpus of english literary classics. In *Informatics*, volume 7, page 32. MDPI, 2020. 6.5

[310] Hao-Ran Wei, Zhirui Zhang, Boxing Chen, and Weihua Luo. Iterative domain-repaired back-translation. *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, 2020. 2.2.3, 5.2.1

[311] Winston Wu, Nidhi Vyas, and David Yarowsky. Creating a translation matrix of the bible's names across 591 languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, 2018. 2.4.3, 5.3.1, 7.2.1

[312] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 2.2.1, 3.2.1

[313] Stephen A Wurm. *Atlas of the World's Languages in Danger of Disappearing*. Unesco, 2001. 2.1.2, 7.1

[314] Cao Xueqin. *Dream of the Red Chamber*. Editorial Axioma, 2016. (document), 6.1

[315] Xiangkai Zeng, Sarthak Garg, Rajen Chatterjee, Udhyakumar Nallasamy, and Matthias Paulik. Empirical evaluation of active learning techniques for neural mt. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 84–93, 2019. 6.2.2

[316] Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 2.2.3

[317] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *Proceedings of the 9th International Conference on Learning Representations*, 2019. 1.3, 2.4.4, 6.2.4, 7.2.3

[318] Yuekai Zhao, Haoran Zhang, Shuchang Zhou, and Zhihua Zhang. Active learning approaches to enhancing neural machine translation. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing*, pages 1796–1806, 2020. 2.3.2

[319] Liang Zhou, Chin-Yew Lin, and Eduard Hovy. Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 11th Conference on Empirical Methods in Natural Language Processing*, pages 77–84. Association for Computational Linguistics, 2006. 4.2.1

[320] Zhong Zhou and Alex Waibel. Active learning for massively parallel translation of constrained text into low resource languages. *Proceedings of the 4th Workshop on Technologies for Machine Translation of Low Resource Languages in the 18th Biennial Machine Translation Summit*, 2021. 3, 1, 6.1, 6.5

[321] Zhong Zhou and Alex Waibel. Family of origin and family of choice: Massively parallel lexiconized iterative pretraining for severely low resource text-based translation. *Proceedings of the 3rd Workshop on Research in Computational Typology and Multilingual NLP in the 20th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, 2021. 2.2.3, 2.4.3, 1, 6.1, 6.2.1, 6.2.3, 6.2.4, 7.1, 7.2.1, 7.2.3, 9.3.2

[322] Zhong Zhou, Matthias Sperber, and Alex Waibel. Massively parallel cross-lingual learning in low-resource target language translation. In *Proceedings of the 3rd conference on Machine Translation*. Association for Computational Linguistics, 2018. 2.2.3, 1, 4.1, 4.4.3, 5.1, 5.3.1, 5.3.2, 6.1, 6.2.1, 6.4.1, 9.1.2

[323] Zhong Zhou, Matthias Sperber, and Alex Waibel. Paraphrases as foreign languages in multilingual neural machine translation. *Proceedings of the Student Research Workshop at the 56th Annual Meeting of the Association for Computational Linguistics*, 2019. 2.2.1, 2.2.2, 2.2.3, 1, 5.2.1, 6.2.1

[324] Zhong Zhou, Jan Niehues, and Alex Waibel. Train global, tailor local: Minimalist multilingual translation into endangered languages. *Proceedings of the 6th Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT) of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023. 1, 1

[325] Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655, 2020. 5.1, 5.3.2

[326] Zhaorong Zong. Research on the relations between machine translation and human translation. *Journal of Physics: Conference Series*, 1087(6):062046, 2018. 2.3.1

[327] Barret Zoph and Kevin Knight. Multi-source neural translation. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*, pages 30–34, 2016. 2.2.1, 2.2.2, 4.1, 4.2.2

[328] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *Proceedings of the 21st Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, 2016. 2.1.1