UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Unsupervised Neural Machine Translation between the Portuguese language and the Chinese and Korean languages

Catarina Francisca Nunes da Cruz Ferreira

Mestrado em Informática

Dissertação orientada por:
Prof. Doutor António Manuel Horta Branco
e coorientada por
Doutor João Ricardo Martins Ferreira da Silva

2022

Acknowledgments

First, I would like to express my gratitude to my supervisors, Professor Doctor António Branco, and Doctor João Silva. Without their guidance and expertise, this dissertation would not have been possible.

Next, I would like to thank the NLX group for granting me the resources to execute my project and all of the group members who helped me out and offered me insights into the area of NLP.

To conclude, I cannot forget to thank all of my family for their unconditional support during this intensive academic year.

Resumo

O propósito desta dissertação é apresentar um estudo comparativo e de reprodução sobre técnicas de Tradução Automática Neuronal Não-Supervisionada (Unsupervised Neural Machine Translation) para o par de línguas Português (PT) → Chinês (ZH) e Português (PT) → Coreano (KR) tirando partido de ferramentas e recursos online.

A escolha destes pares de línguas prende-se com duas grandes razões. A primeira refere-se à importância no panorama global das línguas asiáticas, nomeadamente do chinês, e também pela influência que a língua portuguesa desempenha no mundo especialmente no hemisfério sul. A segunda razão é puramente académica. Como há escassez de estudos na área de Processamento Natural de Linguagem (NLP) com línguas não-germânicas (devido à hegemonia da língua inglesa), procurou-se desenvolver um trabalho que estude a influência das técnicas de tradução não supervisionada em par de línguas poucos estudadas, a fim de testar a sua robustez.

Falada por um quarto da população mundial, a língua chinesa é o "Ás" no baralho de cartas da China. De acordo com o International Chinese Language Education Week, em 2020 estimava-se que 200 milhões pessoas não-nativas já tinham aprendido chinês e que no ano corrente se encontravam mais de 25 milhões a estudá-la. Com a influência que a língua chinesa desempenha, torna-se imperativo desenvolver ferramentas que preencham as falhas de comunicação. Assim, nesta conjuntura global surge a tradução automática como ponte de comunicação entre várias culturas e a China.

A Coreia do Sul, também conhecida como um dos quatro tigres asiáticos, concretizou um feito extraordinário ao levantar-se da pobreza extrema para ser um dos países mais desenvolvidos do mundo em duas gerações. Apesar de não possuir a hegemonia económica da China, a Coreia do Sul exerce bastante influência devido ao seu soft power na área de entretenimento, designado por hallyu. Esta "onda" de cultura pop coreana atraí multidões para a aprendizagem da cultura. De forma a desvanecer a barreira comunicativa entre os amantes da cultura coreana e os nativos, a tradução automática é um forte aliado porque permite a interação entre pessoas instantaneamente sem a necessidade de aprender uma língua nova.

Apesar de Portugal não ter ligações culturais com a Coreia, há uma forte ligação com a região administrativa especial de Macau (RAEM) onde o português é uma das línguas oficiais, sendo que a Tradução Automática entre ambas as línguas oficiais é uma das áreas estratégicas do governo local tendo sido estabelecido um laboratório de Tradução Automática no Instituto Politécnico de Macau que visa construir um sistema que possa ser usado na função pública de auxílio aos tradutores.

Neste trabalho foram realizadas duas abordagens: (i) Tradução Automática Neuronal Não Supervisionada (Unsupervised Neural Machine Translation) e; (ii) abordagem pivô (pivot approach). Como o foco da dissertação é em técnicas não-supervisionadas, nenhuma das arquiteturas fez uso de dados paralelos entre os pares de línguas em questão. Nomeadamente, na primeira abordagem usou-se dados monolingues. Na segunda introduziu-se uma terceira língua pivô que é utilizada para estabelecer a ponte entre a língua de partida e a de chegada.

Esta abordagem à tradução automática surgiu com a necessidade de criar sistemas de tradução para pares de línguas onde existem poucos ou nenhuns dados paralelos. Como demonstrado por Koehn and Knowles [2017a], a tradução automática neuronal precisa de grandes quantidades de dados a fim de ter um desempenho melhor que a Tradução Automática Estatística (SMT). No entanto, em pares de línguas com poucos recursos linguísticos isso não é exequível. Para tal, a arquitetura de tradução automática não supervisionada somente requer dados monolingues. A implementação escolhida foi a de Artetxe et al. [2018d] que é constituída por uma arquitetura encoder-decoder. Como contém um double-encoder, para esta abordagem foram consideradas ambas direções: Português $\leftrightarrow$ Chinês e Português $\leftrightarrow$ Coreano. Para além da reprodução para línguas dissimilares com poucos recursos, também foi elaborado um estudo de replicação do artigo original usando os dados de um dos pares de línguas estudados pelos autores: Inglês $\leftrightarrow$ Francês.

Outra alternativa para a falta de corpora paralelos é a abordagem pivô. Nesta abordagem, o sistema faz uso de uma terceira língua, designada por pivô, que liga a língua de partida à de chegada. Esta opção é tida em conta quando há existência de dados paralelos em abundância entre as duas línguas. A motivação deste método é fazer jus ao desempenho que as redes neuronais têm quando são alimentadas com grandes volumes de dados. Com a existência de grandes quantidades de corpora paralelos entre todas as línguas em questão e a pivô, o desempenho das redes compensa a propagação de erro introduzida pela língua intermediária. No nosso caso, a língua pivô escolhida foi o inglês pela forte presença de dados paralelos entre o pivô e as restantes três línguas. O sistema começa por traduzir de português para inglês e depois traduz a pivô para coreano ou chinês. Ao contrário da primeira abordagem, só foi considerada uma direção de Português $\rightarrow$ Chinês e Português $\rightarrow$ Coreano. Para implementar esta abordagem foi considerada a framework OpenNMT desenvolvida

por [Klein et al., 2017].

Os resultados foram avaliados usando a métrica BLEU [Papineni et al., 2002b]. Com esta métrica foi possível comparar o desempenho entre as duas arquiteturas e aferir qual é o método mais eficaz para pares de línguas dissimilares com poucos recursos.

Na direção Português → Chinês e Português → Coreano a abordagem pivô foi superior tendo obtido um BLEU de 13,37 pontos para a direção Português → Chinês e um BLEU de 17,28 pontos na direção Português → Coreano. Já com a abordagem de tradução automática neural não supervisionada o valor mais alto obtido na direção Português → Coreano foi de um BLEU de 0,69, enquanto na direção de Português → Chinês foi de 0,32 BLEU (num total de 100).

Os valores da tradução não supervisionada vão estão alinhados com os obtidos por [Guzmán et al., 2019], [Kim et al., 2020]. A explicação dada para estes valores baixos prende-se com a qualidade dos cross-lingual embeddings. O desempenho dos cross-lingual embeddings tende a degradar-se quando mapeia pares de línguas distantes e, sendo que modelo de tradução automática não supervisionado é inicializado com os cross-lingual embeddings, caso estes sejam de baixa qualidade, o modelo não converge para um ótimo local, resultando nos valores obtidos na dissertação.

Dos dois métodos testados, verifica-se que a abordagem pivô é a que tem melhor performance. Tal como foi possível averiguar pela literatura corrente e também pelos resultados obtidos nesta dissertação, o método neuronal não-supervisionado proposto por Artetxe et al. [2018d] não é suficientemente robusto para inicializar um sistema de tradução suportado por textos monolingues em línguas distantes. Porém é uma abordagem promissora porque permitiria colmatar uma das grandes lacunas na área de Tradução Automática que se cinge à falta de dados paralelos de boa qualidade. No entanto seria necessário dar mais atenção ao problema dos cross-lingual embeddings em mapear línguas distantes.

Este trabalho fornece uma visão sobre o estudo de técnicas não supervisionadas para pares de línguas distantes e providencia uma solução para a construção de sistemas de tradução automática para os pares de língua português-chinês e português-coreano usando dados monolingues.

Palavras-chave: Processamento de Linguagem Natural, Tradução automática não supervisionada, Coreano, Português, Chinês

Abstract

This dissertation presents a comparative and reproduction study on Unsupervised Neural Machine Translation techniques in the pair of languages Portuguese (PT) → Chinese (ZH) and Portuguese (PT) → Korean(KR).

We chose these language-pairs for two main reasons. The first one refers to the importance that Asian languages play in the global panorama and the influence that Portuguese has in the southern hemisphere. The second reason is purely academic. Since there is a lack of studies in the area of Natural Language Processing (NLP) regarding non-Germanic languages, we focused on studying the influence of non-supervised techniques in under-studied languages.

In this dissertation, we worked on two approaches: (i) Unsupervised Neural Machine Translation; (ii) the Pivot approach. The first approach uses only monolingual corpora. As for the second, it uses parallel corpora between the pivot and the non-pivot languages.

The unsupervised approach was devised to mitigate the problem of low-resource languages where training traditional Neural Machine Translations was unfeasible due to requiring large amounts of data to achieve promising results. As such, the unsupervised machine translation only requires monolingual corpora. In this dissertation we chose the implementation of Artetxe et al. [2018d] to develop our work.

Another alternative to the lack of parallel corpora is the pivot approach. In this approach, the system uses a third language (called pivot) that connects the source language to the target language. The reasoning behind this is to take advantage of the performance of the neural networks when being fed with large amounts of data, making it enough to counterbalance the error propagation which is introduced when adding a third language.

The results were evaluated using the BLEU metric and showed that for both language pairs Portuguese → Chinese and Portuguese → Korean, the pivot approach had a better performance making it a more suitable choice for these dissimilar low resource language pairs.

Keywords: Natural Language Processing, Unsupervised Neural Machine Translation, Neural Networks, Chinese, Korean, Portuguese

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In a more globalized and connected world where physical barriers are no longer a limitation to exploring new cultures, language differences become the only obstacle to communicating with other civilizations. Despite English's position as the *lingua franca*, the truth is that many people cannot hold a basic conversation in Shakespeare's language as, according to Eberhard et al. [2022], only around 1 billion people can speak English.

Although the act of translating is one of the oldest practices in the world, going back to the Mesopotamian era where Sumerian poems were translated into Asian languages, nowadays it has been revamped with the aid of technology. We went from focusing on translating religious texts to incorporating translation in every aspect of our society ranging from phone applications that do real-time translation to social networks adding features that allow us to read comments from other users written in their native languages. In times like these, technology can work as a bandage that patches the unintelligibility between cultures.

As we become more dependent on the cyber-world, so do our needs increase to communicate with every corner of the world. The problem is that the technology we use nowadays requires large amounts of parallel data. That is, data composed by texts and their translations, where each sentence and its corresponding translation are explicitly marked. This means that implementing state-of-the-art Neural Machine Translation (NMT) techniques is only feasible for languages with a solid online presence because that allows us to quickly obtain the needed data to train these data-hungry models. English, being one of the languages with the largest number of online users, is always on the front-line regarding advances in Natural Language Processing (NLP). The consequences of this are that many languages are left out, which hinders the ability of their native speakers to engage in meaningful cross-cultural interactions. These native speakers usually do not grasp English well,

making them isolated from the rest of the world.

One of the solutions that NLP proposes is using unsupervised techniques that leverage the abundance of existing monolingual data to train neural machine translation systems. As people's online presence increases, tremendous amounts of texts are generated, which can be used to create and improve machine translation systems. As such, unsupervised techniques are a promising solution to the dependence on parallel data, as this type of data are expensive and time-consuming to obtain since they require a body of expertise to create.

Many of the studies behind these unsupervised techniques have been conducted in English or other Indo-European languages. There is little scholarship that tests the feasibility of the unsupervised approaches in dissimilar languages, making it hard to assess their applicability when dealing with the creation of an MT system for dissimilar language pairs with little parallel corpora available.

Thus, the motivation behind this dissertation is to contribute to the area of Machine Translation by providing an in-depth study of the feasibility of unsupervised techniques for dissimilar language pairs. We chose the following language pairs: Portuguese → Chinese and Portuguese → Korean. These language pairs were chosen for two reasons. The first is the lack of good quality parallel corpora, making them perfect candidates as they are low-resource language pairs. The second is to attest to the robustness of unsupervised approaches when dealing with distant languages, as most of the literature is focused on using English paired with other Indo-European languages.

1.2 Research Context and Goals

This work was developed over 9 months at the NLX—Natural Language and Speech Group, a Natural Language Processing research group from the University of Lisbon, Faculty of Sciences.

The goal of this dissertation was to compare and reproduce two different machine translation systems in the pair of languages Portuguese, Chinese, and Korean. This dissertation aims to conduct a comparative study of different MT architectures to better understand how unsupervised machine translation behaves when dealing with dissimilar languages.

1.3 The Portuguese Language

With around 250 million native speakers and 24 million L2 (second language) speakers, Portuguese is the 6th most spoken language in the world [Eberhard et al., 2022]. It is the official language of 9 countries (Angola, Brazil, Cape Verde, East

Timor, Equatorial Guinea, Guinea-Bissau, Mozambique, Portugal, and São Tomé and Princípe) and the most widely used language in the southern hemisphere.

The Portuguese language traces its roots back to the Ibero-Romance group that evolved from the vulgar Latin [Posner, 1996].

Its writing system is based on the Latin script containing 26 letters, and there are 9 oral vowels, 2 semivowels and 21 consonants in European Portuguese. Due to being the official language of many nations spread out over the entire world, Portuguese boasts a variety of dialects, with European Portuguese and Brazilian Portuguese being the two most well-known ones. One of the major differences between these two variants is in the prosody. European Portuguese is considered a stress-timed dialect whereas Brazilian Portuguese is a syllable-timed one. Like all Romance languages, Portuguese follows SVO (subject, verb, object) as the canonical sentence order and preserves from Classical Latin the verb inflection, a feature common to all of the Romance languages, adding up to over a dozen of conjugational endings. Moreover, it boasts some grammar idiosyncrasies that are not commonly found in other languages. Some such features are the usage of the future subjunctive mood and the personal infinite inflection according to its subject in person and number (e.g., "é melhor voltarmos", Eng. *is best we-return*).

1.4 The Chinese Language

The Chinese language is better envisaged as a cluster of language families that belong to the Sino-Tibetan group. According to linguists, there are between 8 to 13 main dialects with several sub-dialects [Norman, 2002]. They can differ in various ways, such as in pronunciation, grammar, or vocabulary. An example of this difference is between Mandarin and Shanghainese. Compared to Mandarin, Shanghainese has more vowels and consonants. Another difference is in the vocabulary. For example, "I" in Mandarin is written and spelled as 我 (wo), whereas in Shanghainese it is pronounced differently 吾 (ngu). For Mandarin speakers, the word 吾 (ngu) would sound completely foreign as the consonant *ng* does not exist in Mandarin.

With over a billion speakers, Mandarin is the widest spread dialect from the Sino-Tibetan family and has been the official language of China since the 1930s.

The Chinese language is a tonal language. In the case of Mandarin, it uses 4 tones to distinguish words, while other dialects use different tones. Cantonese, for instance, makes use of 9 tones. Tonality is an important feature in the Chinese language phonology as it is essential for intelligibility due to the vast number of words that only differ in their tones. In Figure 1.1 there is a description of the tones in Mandarin Chinese. Failing to correctly pronounce a tone could lead to situations where, for instance, wrongly pronouncing "mother" could be mixed up with "horse"

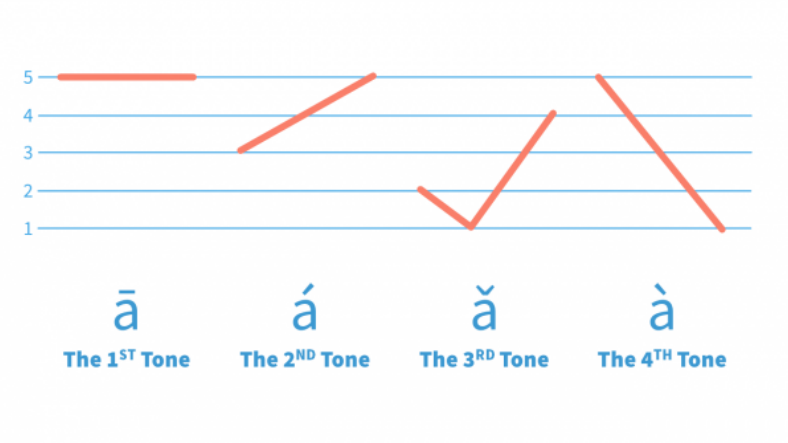

Figure 1.1: Illustrative example of how Chinese tones are pronounced.

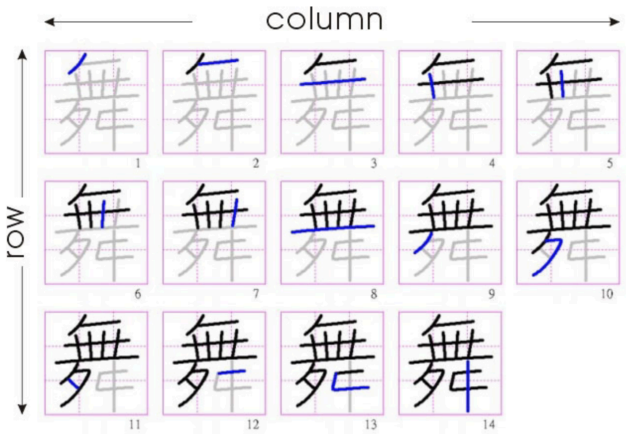

Figure 1.2: Stroke order in Chinese writing. The blue line depicts the order in which the strokes of the character must be written.

since both words only differ by one tone.

The Chinese written system differs from alphabetic systems in the sense that words are composed by characters rather than letters. For example, while in Romance languages each letter corresponds to a phoneme, in Chinese each logogram corresponds to a syllable with a semantic meaning, and each logogram can be monosyllabic or part of a polysyllabic word.

Chinese characters are composed by a set of strokes and radical components which all together make up a character. Figure 1.2 represents a structure of the character 舞 (Wǔ) where each line highlighted in blue corresponds to the strokes which must be written in a defined sequence, from the left to the right and from the top to the bottom. In total it is estimated that there exist around 50,000 characters, however the average college-educated student masters around 4,000, which are enough to be literate in the Chinese language.

Records of Chinese primordial writing system were reported from the late Shang

Dynasty circa 1250–1050 BC. The early Chinese script is called Oracle Bone due to the writings being carved on bronze vessels and oracle bones[1] Throughout Chinese history, characters went through a succession of alterations, and it was during the Qin Dynasty (221–206 BC) that they were standardized. By the 20th century, classical Chinese was replaced by the written vernacular Chinese (i.e., up to this point the written Chinese and spoken Chinese were mutually unintelligible) and during the Cultural Revolution (circa 1966–1976), in an effort to increase alphabetization, the Chinese government simplified a variety of characters gaining the form that is known nowadays.

Like Portuguese and English, Chinese follows SVO (subject-verb-object) as the canonical sentence order but, unlike Portuguese, Chinese lacks inflection. Instead, Chinese uses particles to express verbal aspects (e.g., the particle - 了 (le) is used to denote an action that happened in the past)

1.5 The Korean Language

Korean is the 13th most spoken language in the world, with a total number of 77 million speakers. There are two types of dialect: the Seoul dialect, which is spoken in South Korea, and the Pyongyang dialect, which is the official version used in North Korea.

The origins of Korean language remain as a source of mystery up to present times. It is a source of a heated debate amongst Korean scholars [Song, 2005, Campbell and Mixco, 2007] whether Korean should be considered an Altaic language, which is a group of Asian languages such as Japanese, Mongolian and Turkish, or remain as a "language isolate".

The Korean writing system is derived from Hangul which is an alphabet created by King Sejong from the Yi Dynasty in 1443. Before the creation of Hangul, Korean writing system used Chinese characters. However, due to the significant grammar differences between the two languages it was a particular challenge writing in Korean. As such, to make writing more accessible to the masses Hangul was invented.

Considered as one of the most scientific alphabets ever devised, Hangul consists of 24 letters: 14 consonants and 10 vowels. The letters are formed into building blocks that represent syllables which are constructed under a set of rules. For example, if the syllable (i.e., block) starts with a vowel ㅏ (read as *a*) then it must be preceded with a silent consonant ㅇ that will act as a placeholder. Another rule is that a syllable can start with a primary consonant but never a complex one (i.e., corresponds to diphthongs). The last rule is that if a syllable starts with a vowel, it can not occur with consecutive vowels, or if it starts with a consonant, it can not

[1] Oracle bones were animal bones used for divination, hence their name.

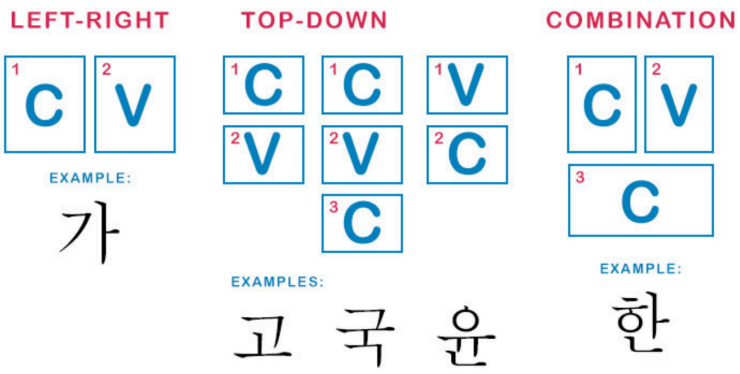

Figure 1.3: Illustrative example of the rules of the Korean alphabet. The first character -가 (ga) is written by the consonant -ㄱ (g) and followed by the vowel -ㅏ (a). On the middle picture there are three blocks and as it can be attested in the third character 윤 (yoon) it begins with a vowel -유 (yoo) and is followed by a consonant -ㄴ (n)

occur. consecutive consonants Figure 1.3 exemplifies these writing rules of Hangul.

The Korean language still preserves reminiscences of Confucianism in its structure. This is particularly noticeably in the case of honorifics which are a form of speech that reflect the hierarchical social status between the speaker and listener. It breaks down into three values for the dimensions of formality, politeness and honorifics.

The formal dimension is used when speaking with people that are close to the speaker or are of lower age. The formal speech can be identified by the verbs ending in the particle -다 (da) and by the usage of informal honorific pronouns and markers such as -님 (nim) or -야 (ya).

The second dimension is called politeness and is widely used in social situations where the speaker is unfamiliar with the listener and wants to show respect.

The final dimension is used in situations with a defined hierarchical structure like in a workplace context or in school. To show respect, it is expected that students approach their professor in this tone and the same is applied to workers when dealing with their superior. On the other hand, it is not expected for those in higher social position to use the honorifics or politeness form of speech.

Despite the vast influence of Chinese language on Korean, namely in vocabulary, Korean is considered an agglutinate language with a SOV (subject-object-verb) canonical sentence order. Such characteristics are also commonly shared with other Asian languages like Japanese and Mongolian.

1.6 Document Structure

This document has seven chapters and is structured in the following manner:

- Chapter 2 refers to the planning and goals of the dissertation
- Chapter 3 gives an overview of the field of NLP with special focus on the various types of Machine Translation architectures.
- Chapter 4 relates previous works done in the topic of the dissertation and introduces the concept of low-resource language pairs.
- Chapter 5 describes the work performed, the frameworks and tools used.
- Chapter 6 provides the evaluation results and a discussion.
- Chapter 7 gives final remarks and pointers for future work.

Chapter 2

Planning and Goals

In this chapter the objectives, planning and development of my work will be addressed. The methodology followed and the comparison between the planned and the actual work are also detailed below.

2.1 Objectives and Planning

The major objective of this dissertation is to develop a comparative study and reproduction of two different types of machine translation architectures and evaluate which method is the most suitable for low resource language pairs that are dissimilar.

To carry out this study three languages were chosen that fulfill the two following criteria: Having abundant monolingual corpora with little to none parallel data and being from distant language families. Given these criteria, Portuguese, Chinese and Korean were selected.

The first goal of this dissertation was to get acquainted with the field of Natural Language Processing, particularly Neural Machine Translation. As such, the first two months were dedicated to a review of the literature where it was needed to understand the differences between Supervised Neural Machine Translation, Unsupervised Neural Machine Translation, and Statistical Machine Translation. Plus, it was also required to gain familiarization with the state-of-the-art in Unsupervised Neural Machine Translation.

Given that we will be experimenting with data-driven models for which little data is already available, data gathering and curation are central to this work. The second goal was thus to gather and curate all the data necessary for the planned experiments. This data should be of good quality and in abundance as noisy or poor quality data could affect the performance of the MT system.

The third goal was to successfully develop the two proposed Machine Translation systems: the Unsupervised Neural Machine Translation and the Pivot-based Approach. To develop the pivot-based approach it was needed to grasp the workings

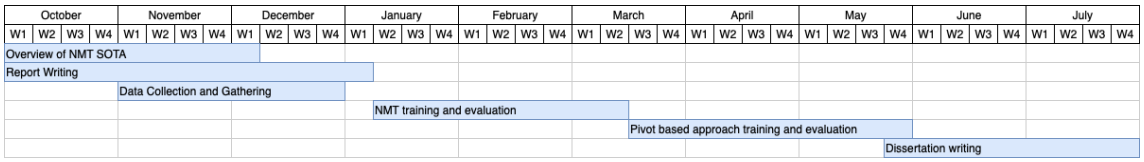

Figure 2.1: Dissertation execution timeline

of the OpenNMT framework. To train the UNMT model, we had to understand [Artetxe et al., 2018d] code's repository and do several experiments to test if it was working correctly. In addition, it was also vital to create our own cross-lingual embeddings as, without them, it would not have been possible to train the UNMT model. As such, we had to be familiarized with the VecMap framework developed by [Artetxe et al., 2018b].

In summary, the following steps were set for the development of my work:

- Acquire knowledge regarding the field of Natural Language Processing and Neural Machine Translation.
- Gain familiarization with the state of the art in Unsupervised Neural Machine Translation;
- Gather monolingual data in the three languages addressed by this work;
- Gather parallel data for the study of the pivot NMT model;
- Develop a MT system in the two proposed architectures;
- Develop the cross-lingual embeddings that are needed to train the UNMT system;
- Understand and test the code's repository to train the UNMT model;
- Gain familiarization with the OpenNMT framework and use it to build the pivot-based approach;

2.2 Planning

In order to achieve the aforementioned goals, a set of guidelines carried out during this dissertation were defined. The plan proposed is as follows:

A) Be familiarized with the field of Natural Language Processing and with Neural Machine Translation - 2 Months

B) Writing the report for the "Estudo Orientado a Informática" course - 2 Months (overlapping with A)

C) Gathering and collecting data - 1 Month (overlapping with A, B)

D) Training and evaluation of the unsupervised NMT system - 2 Months

E) Training and evaluation of the pivot MT system - 2 months

F) Writing of the dissertation - 2 months

2.2.1 Execution of the plan

In this section, we detail the plan's execution and point out any deviations from the initial plan.

As we can see in Figure 2.1, the first item to execute was the literature review (item A). Many information sources were consulted to obtain knowledge in the area of the NLP, such as the Annual Meetings of the Association for Computational Linguistics (ACL) and the Conference on Machine Translation (WMT). Item A was executed according to the plan, taking two months to accomplish. Nonetheless, given this dissertation's nature, different research papers were often conferred.

While consulting different research papers to build knowledge in the area of NLP, the report was being written (item B). The initial plan had an execution time of two months for this task, from October to December. Nonetheless, since the delivery time was in January, the execution time was further delayed until January because the report needed further revision.

During this time, around December, we began to gather the data (item C) needed to train the MT systems. The previous readings in the literature review phase proved helpful since the data used to train the UNMT system was referenced from [Artetxe et al., 2018d] paper: WMT monolingual News Crawl [1] contains millions of sentences in the three languages with good quality.

The initial plan to train and evaluate the unsupervised NMT (item D) was scheduled for two months, as shown in Figure 2.1. Experiments were done to assess whether the model was converging. As a result, we also conducted a reproduction study in the EN-FR language pair. We ended up deviating over two weeks from the original plan due to an undetected error in the early stages of training. Once we successfully trained the models, they were evaluated - a vital step to ensure our goal was achieved.

As soon as we concluded the evaluation of the UNMT, we advanced to the training and evaluation of the pivot-based approach (item E). Since we added the pivot language, which implies training an additional model, and the training time for each model has two-folded, we set the execution time to three months. That

[1] http://data.statmt.org/news-crawl/

means we started in March and concluded at the end of May with the training and evaluation of the pivot-based approach.

Ultimately, we set the writing of the dissertation (item F) to be for two months, starting from the end of May to the end of July. The execution time is to dedicate the month of June to the writing and July for the revision.

Chapter 3

Literature Review

This chapter provides an overview of the field of Machine Translation by introducing some key concepts on word embeddings, neural networks, and the current state of the art in the unsupervised neural machine translation area.

3.1 Word Embeddings

Word embeddings are vectorial representations of word semantics. Since word meaning is projected into a vector space in a way that ensures that related words are placed closer to each other than unrelated words, the semantic relatedness of words can be assessed based on metrics such as their distance within this vector space or the cosine of their vectors.[Turian et al., 2010].

Word2Vec

There are multiple methods for obtaining word embeddings. Here we take Word2Vec as an illustrative example.

Word2Vec is a neural network for constructing word embeddings developed by Mikolov et al. [2013].

A vector represents each word, and the cosine similarity computes the semantic closeness between each word. The model takes considerable input from corpus text, and each word in the corpus is assigned a vector represented in the space. Thus semantically similar words are close to each other in the vector space.

Word2Vec uses two model architectures: Continuous Bag of Words (CBOW) and Skip-Gram. Skip-gram gives better word representations when the monolingual data is small. CBOW is faster and more suitable for larger datasets [Mikolov et al., 2013].

Skip-Gram

In terms of architecture, Skip-gram is a simple neural network with only one hidden layer. Given a center word Skip-gram iteratively looks at words within a range set by the window size. In other words, it looks for n words to the right of the center word and n words to the left of the center word. The goal is to calculate the probability of each context word occurring given a center word.

At the input layer, words are encoded as a one-hot vector of dimensions $V \times 1$ where V is the vocabulary size. Then, the input is multiplied by a lookup table to generate a matrix of $V \times N$ dimensions where N corresponds to the word index in the one-hot encoding vector. The matrix is projected onto a projection layer P that acts as a hidden layer and will pass through an activation function as shown in Equation 3.1 that will output a probability distribution of the context words given the center word.

$$P(w_o|w_c) = \frac{\exp(v_0^T v_c)}{\sum_{i \in v} \exp(v_i^T v_c)} \tag{3.1}$$

where w_o is the context word, w_c is the center word and v is vector of word w.

Continuous Bag of Words (CBOW)

Like skip-gram, the architecture of CBOW is a one-layer neural network. The main difference between these two models is that on CBOW, we predict the target word based on the context words set up by the sliding window parameter, whereas in skip-gram we try to predict the context words given the center word.

3.1.1 Cross-Lingual Embeddings

Cross-lingual word embeddings are word embeddings for more than one language, where words from different languages are mapped into the same vector space, maintaining the property that words that are semantically related, even across languages, will be closer together than words that are not related. As such, cross-lingual word embeddings match the lexicon between the source and the target language in a common vector space, allowing to measure the semantic distance between words of different languages [Mikolov et al., 2013].

The typical approach that is used to obtain cross-lingual embeddings consists of learning separate word embeddings, one for each language, from monolingual corpora using normal word embedding methods such as Word2Vec (i.e., Skip-gram or CBOW), and then learning a linear transformation that maps the separate embeddings into a shared space, or that transforms the space of one of the embeddings into the space of the other. Next we cover the most relevant works and approaches.

Mikolov et al. [2013] was the first to popularize the linear projection with a transformation matrix after observing that words have a similar geometric arrangement

in the same shared vector space. To achieve this, the authors first created a small bilingual dictionary by translating the 5000 most frequent words from the source to the target language. Then they trained the transformation matrix W by minimizing the error between the translation output of W and the bilingual dictionary entries.

Lazaridou et al. [2015] suggested the margin-based (max-margin) to solve the *hubness* (i.e., some words appear as the nearest neighbors of other words), which is an observed phenomenon in high-dimensional spaces problem caused by the projection matrix.

Xing et al. [2015] proposed vector normalization after finding an inconsistency on the paper of Mikolov et al. [2013] that uses the inner product to compute the distance measure and then cosine distance to estimate word similarities. In [Mikolov et al., 2013] the Euclidean distance is used to calculate the loss function used to train the transformation matrix W. However, when word vectors are applied to estimate word similarities, the algorithm uses cosine distance. To solve the inconsistency, Xing et al. [2015] proposed to use cosine distance in the objective function and replace the inner product as shown in the following equation:

$$\max_{W} \sum_{i} (Wx_i)^T z_i \tag{3.2}$$

To achieve this, they normalize the word embeddings as unit vectors and constrain the linear transformation W into an orthogonal matrix.

Faruqui and Dyer [2014] propose a technique based on canonical correlation analysis (CCA) [Hotelling, 1936] to formulate cross-lingual embeddings. Their technique first constructs independent monolingual word embeddings and then projects them onto a common vector space. To obtain bilingual embeddings, they use CCA, which measures the linear relationship between two multidimensional variables. Unlike a linear projection, CCA learns a transformation matrix for each language.

Artetxe et al. [2016] proposed a new framework based on previous techniques from Xing et al. [2015] and Faruqui and Dyer [2014]. Like Xing et al. [2015], Artetxe et al. [2016] proposes the linear transformation to be orthogonal. However, they show that by doing so, it optimizes the objective function, whereas in [Xing et al., 2015] orthogonality was used to enforce the word vectors to be of unit length. While the model of Faruqui and Dyer [2014] changes the monolingual embeddings by applying restrictions, Artetxe et al. [2016] claim that forcing monolingual invariance in their model improves the learning of the bilingual mapping.

Seeking to map language pairs without dictionaries or word alignments, Barone [2016] proposed a new unsupervised method using an adversarial auto-encoder that does not need parallel data. The method combines an encoder that is used to transform the source embedding into the target embedding, a discriminator that discriminates between the actual target embedding and the mapped embedding,

and a decoder that reconstructs from the mapped embedding the source embedding. Despite the novelty, the author concluded that his approach was not competitive enough.

Following the same line of work in fully unsupervised cross-lingual embeddings, Artetxe et al. [2018b] proposes constructing a dictionary in a fully unsupervised manner by introducing two new methods: a fully unsupervised initialization scheme as an initial solution and robust self-learning.

Fully unsupervised scheme

The fully unsupervised initialization scheme is based on intra-lingual similarity distribution. The idea is to take a word in some language (e.g., "dog") and compute the similarity between that word and the rest of the words in that language using the monolingual embeddings. This comparison between one word and the rest of the words in the monolingual embeddings will give the similarity distribution for that word. Having done that, the model moves to another language (e.g., Italian) and repeats the same process for all the words in the vocabulary.

The rationale is that one would expect equivalent words in different languages to have similar distributions. If we plot two words in different languages, like "dog" in English and its corresponding word "cane" in Italian, we would expect them to have a similar distribution, that is have a similar plot.

Robust self-learning

One of the modifications in the cross-lingual mapping proposed in [Artetxe et al., 2017] is the introduction of the stochastic dictionary induction. This is done by randomly discarding some entries in the cross-lingual similarity matrix. Doing so makes the solution change more iteratively, and enables a broader exploration of the search space, thus escaping from poor local optima. In addition, Artetxe et al. [2018b] also incorporate a frequency-based vocabulary cutoff. Thanks to this, the authors produce the search space and simplify the optimization problem. Also, they incorporate Cross-domain Similarity Local Scaling retrieval (CSLS) [1] [Conneau et al., 2017b] which is shown to be beneficial in mitigating the hubness problem. Moreover, they make the dictionary induction process bidirectional. At last, after the final iteration, they apply symmetric re-weighting [Artetxe et al., 2018a].

Despite the numerous methods to train cross-lingual embeddings, in this dissertation we will be focusing on the Vecmap framework forwarded by Artetxe et al.

[1] CSLS is used to evaluate the quality of cross-lingual embeddings by computing the cosine similarity between the aligned words. It can be used to reduce the hubness problem by weighting the cosine similarity scores helping the k-nearest neighbor algorithm to find the correct data points in an over-represented vector space.

[2018b]. We chose this framework not only because it has achieved state-of-the-art results in this area but also due to it being easy to implement.

3.1.2 Sub-word techniques

Neural machine translation is the state of the art in MT [Vaswani et al., 2017]. However, it faces problems when dealing with out-of-vocabulary (OOV) words [Sennrich et al., 2016b] (i.e., words that never appeared in the training corpus) and rare words (i.e., words that appeared only a few times in the corpus), and also when faced with the open vocabulary issue which is concerned with the limited vocabulary entries that NMT systems can handle.

To address these issues, some algorithms [Sennrich et al., 2016b, Provilkov et al., 2020] have been proposed that segment words into smaller units which are always known to the model, known as *sub-words*. Any word can be formed by combining a different number of these sub-words, therefore avoiding the OOV issue.

Sennrich et al. [2016b] proposed a sub-word technique using the Byte Pair Encoding (BPE) compression algorithm [Gage, 1994]. This algorithm merges the most frequent pairs of characters in a given corpus and replaces the characters with the newly merged pair. The model starts with an initial vocabulary containing all the unique characters in the corpus (e.g., a, b, c, l, o, w). It then counts the most frequent n-gram occurring in the text. For example, the word "lowest" occurred 9 times and "low" 5 times. Thus, "lo" and "ow" are frequent n-gram pairs and give rise to new vocabulary entries (viz., "lo" and "ow"). This process is repeated until the desired size of the vocabulary is reached.

Since BPE splits words in a deterministic fashion, the model is prone to segmentation errors and flawed at accounting for the compositionality of words. This means that for each word, the model only learns to segment words in one way, failing to explore the morphology of words at its fullest potential. Provilkov et al. [2020] proposed BPE-dropout incorporate randomness into the approach. In contrast to the deterministic nature of BPE, which segments words similarly, BPE-dropout allows for multiple segmentation of the same word. It achieves this by randomly dropping merges during the training phase. In the training phase, the parameter p that ranges from 0 to 1 controls the drop-out level (i.e., how much segmentation can be applied to the word). When $p = 0$, the level of segmentation is equal to that of BPE, and when $p = 1$, every single character in the word is segmented. The authors proposed setting $p = 0.1$ for several languages as it achieved the best results during the experimentation [Provilkov et al., 2020].

3.2 Statistical Machine Translation

Statistical machine translation (SMT) [Koehn et al., 2003] is based on the noisy channel model [Shannon, 1948]. A SMT model has two components, namely $p(e)$, which is the language model, and $p(f|e)$, which is the translation model.

The language model calculates the probability of a given sentence e belonging to the target language. That is, the language model aids in identifying which sentences are fluent in the target language.

The translation model calculates the probability of a source sentence f being translated to a target sentence e. That is, the translation model is concerned with finding the probability that a given target sentence is a translation of the source sentence.

Having these two models, we can derive $p(e|f)$ using the Bayes rule, obtaining:

$$\arg\max_e \ p(e|f) = \arg\max_e \ p(f|e) \times p(e) \tag{3.3}$$

where f is the source sentence, e is the target sentence, $p(e)$ is the language model probability for the target sentence, $p(f|e)$ is the translation model probability for the source sentence given the target sentence.

To translate from the source language to the target language, we search for the target sentence e that maximizes the product of the language model probability $p(e)$ and the translation model probability $p(f|e)$, given the source sentence$f$. This is equivalent to finding the target sentence that maximizes the posterior probability $p(e|f)$ using Bayes rule.

3.2.1 Phrase based models

Figure 3.1 shows a representation of how a phrase based model works. Whereas in word based models the translations are done word by word, in phrase based models they are done by phrases. The mathematical model is shown below.

$$\arg\max_e \ p(e|f) = \arg\max_e \ p(f|e) \times p_{LM}(e) \times p_D(e,f) \times \omega^{\text{length}(e)} \tag{3.4}$$

where $p(f|e)$ is the translation model, $p_{LM}(e)$ the language model, $p_D(e,f)$ the reordering model and $\omega^{\text{length}(e)}$ is a word penalty. These components are further explained below.

Phrase translation model

The phrase translation model consists of a table with all the possible translations that are consistent with the word alignment of the sentence pair, where each phrase has a score associated. To build a phrase model SMT computes word alignments

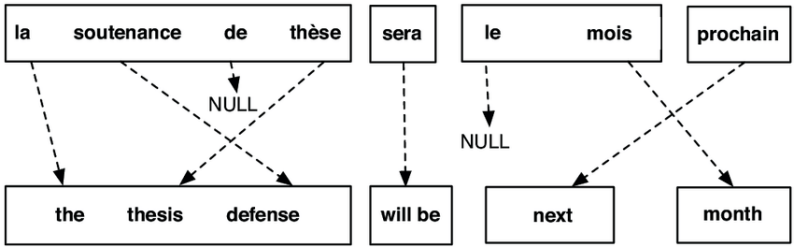

Figure 3.1: Example of how a SMT model works (from [Shah, 2012]).

using IBM models [Knight, 1999, 1997, Brown et al., 1990, 1988] or an EM algorithm [Dempster et al., 1977]. It then extracts translation phrase pairs and assigns a score to each of them using the maximum likelihood estimation.

Reordering model

The reordering model is used to tackle a common problem in translation: not all words in a given sentence can be consecutively translated due to word order differences between the source and target language.

To do this, the reordering model scores the word order correctness in a sentence, and is falls under two types, namely distance-based distortion models and a lexical reordering models.

The former uses a scoring factor that penalizes any deviation from the monotonic order, while the latter conditions reordering in phrases using three types of movement: (a) monotonic; (b) swap with the previous phrase; and (c) discontinuous.

Language model

Traditional SMT systems use n-grams as language models where the target word is assigned a probability within a word sequence using a monolingual corpus. The language model calculates the joint probability of $P(w)$ given a prefix of previous words from w_1 to w_n where w_n is set by the number of n-grams we want to look into (e.g., unigram where the probability of each word is independent from previous words, bigram where $P(w)$ is conditioned on the single previous word, etc). The model can be extended to trigrams, 4-grams and so on, but this usually leads to data sparseness issues.

3.3 Neural Machine Translation

With the considerable increase, over the past few years, in computational power and readily accessible data, Neural Machine Translation (NMT) has become the mainstream paradigm in machine translation [Sutskever et al., 2014, Cho et al., 2014b]. It uses an end-to-end approach that, in a single model, directly maps an

input sequence (e.g., English) to an output sequence (e.g., French). They solve some of the problems recurring with SMT, such as mitigating the sparsity problem and overcoming the locality problem by using unconstrained contexts [Artetxe et al., 2018c].

The most popular architectures in NMT are Sequence-to-Sequence (Seq2Seq) based on Recurrent Neural Networks (RNN) [Sutskever et al., 2014, Cho et al., 2014b], and the current state-of-the-art Transformer [Vaswani et al., 2017].

3.3.1 Sequence-to-Sequence (Seq2Seq)

Sutskever et al. [2014] were the first to use Recurrent Neural Networks (RNN) in Machine Translation to directly map input sequences into output sequences, in an approach called sequence-to-sequence (Seq2Seq).

To address the inability of Deep Neural Networks (DNN) to deal with sequence data of variable length, since these networks require the input and output to be vectors with fixed dimensions, Sutskever et al. [2014] applied the Long Short-Term Memory (LSTM) architecture [Hochreiter and Schmidhuber, 1997], a type of RNN, to solve the sequence-to-sequence task.

The main idea of Sutskever et al. [2014] is to use two LSTMs: the *encoder*, which encodes the source language sequence into a single vector, and the *decoder*, which takes that vector and decodes the target language sequence.

The input sequence is processed one word at a time by the encoder. Similarly, the decoder generates the output sequence one word at a time. Like all RNNs, information about the state of the process is managed by having the output and the hidden state of the network at any given time step be used as part of its input in the next time step.

3.3.2 Sequence-to-Sequence with Attention

The Seq2Seq encoder from the previous section compresses all the information of the source sentence into a fixed-sized vector. This results in performance problems due to information loss when dealing with long sentences [Cho et al., 2014a]. As a solution to this problem, Bahdanau et al. [2015] proposed the mechanism of attention.

The crucial innovation proposed by Bahdanau et al. [2015] consists of having the encoder expose all intermediate encodings it produces while processing the input, instead of only exposing the final one. The decoder, in turn, instead of having to work based solely on a single encoding, computes a weighted sum of all encoder states, this being the so-called *attention* mechanism.

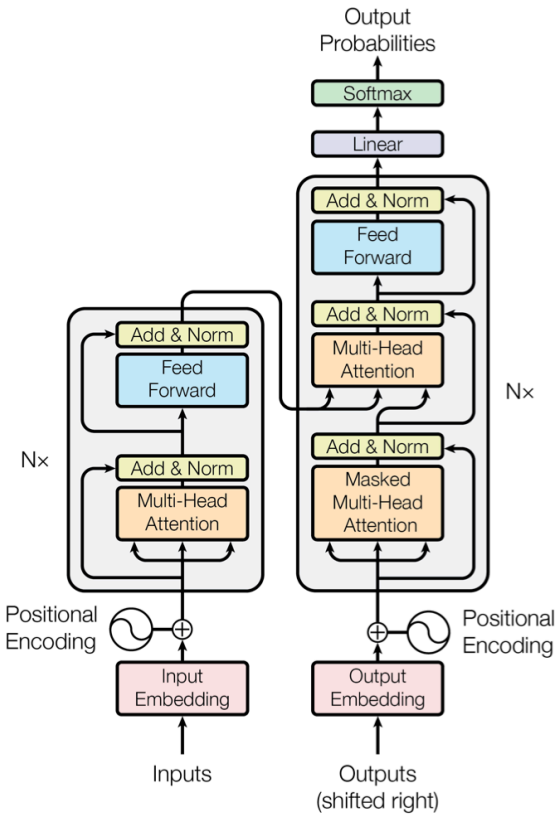

Figure 3.2: Transformer architecture (from [Vaswani et al., 2017])

3.4 Transformers

As stated in Section 3.3.1, RNNs generate sequences of hidden states based on the previous hidden state and current input. Since, at each time step, RNN models needs information from the previous time step, any attempt at input data parallelization is hindered. To reduce sequential computation costs caused by the RNN models, Vaswani et al. [2017] proposed a new architecture called Transformer that relies on self-attention, ditching RNNs entirely.

The Transformer also follows an encoder-decoder architecture (see Figure 3.2), the difference being that all input tokens are processed in parallel. To accomplish this parallelization, Vaswani et al. [2017] introduced several novel features to the encoder-decoder architecture.

Positional Encoding

A first novel feature is the introduction of positional encodings. Since the model does not use recurrence, which inherently maintains some information about word position, the authors propose a new vector that is added to the input embeddings to inform about the relative position of a word in a sentence. In this particular case,

the authors used *sinusoidal* positional embeddings, based on the sine and cosine functions, as shown in Equation 3.5.

$$\begin{aligned} \mathrm{PE}_{pos,2i} &= \sin(pos/10000^{2i/d}) \\ \mathrm{PE}_{pos,2i+1} &= \cos(pos/10000^{2i/d}) \end{aligned} \tag{3.5}$$

where pos is the position of the word in the sentence, d is the embedding dimension, and i is the index of the word in the embedding vector (even indices use sin while odd indices use cos).

Self-Attention

In RNNs, we needed the previous hidden state of the model to learn the context of the current input. However, as mentioned before, this caused long-term dependencies issues and computational costs due to the input being processed sequentially. To address this issue, Transformers use a mechanism called self-attention inspired by the attention that models the relationship between all the words in the sequence regardless of their position via an attention score.

The attention function maps a query to a set of key-value pairs where the output of the function is a weighted sum of the values where the weight of each value is computed by the dot product between the scaled query and keys.

Different functions can be used to calculate the weights for the attention weighted sum. Vaswani et al. [2017] use the scaled dot product. It takes as input a query **Q**, keys **K** and values **V**. Mathematically is represented as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.6}$$

where QK^T is the dot product between the query and keys, $\sqrt{d_k}$ is a normalization operation to scale down the dot product and lead to more stable gradients, and V is the value vector that is multiplied with the softmax to obtain the weighted sum.

Multi-Head Attention

Vaswani et al. [2017] also introduces the concept of Multi-Head Attention. It is an extension of attention that allows the model to focus on different aspects of the sequence by computing multiple attentions, each with their own set of learned weights. It is computed as follows:

$$\mathrm{MultiHead}(Q, K, V) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h)W^O \tag{3.7}$$

where

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{3.8}$$

Figure 3.3: Example of how pivot machine translation works.

Conceptually each attention head is calculated individually via the Scaled Dot Product. Then they are combined into a final matrix W by concatenating the result of each head.

3.5 Pivot Machine Translation

When dealing with low resource language pairs for which there is little to no parallel data, an alternative to building MT systems is using a pivot language. This approach assumes the existence of a third language, the *pivot* language, where parallel corpora between the source-pivot and pivot-target languages are abundant. Conceptually, the pivot language connects the two models by acting as an intermediary, through two separate models, one from source to pivot and another from pivot to target.

Figure 3.3 shows an illustrative example showing how pivot translation works.

3.5.1 Joint-Training

To reduce the error propagation caused by the additional training of a third language, Cheng [2019] introduced a new joint-training approach. The motivation behind this is to connect the source-to-pivot model and the pivot-to-target model during the training, instead of training them separately as is commonly done in other pivot-based approaches. To accomplish this, the authors proposed a training objective composed of three parts, as shown in the following equation.

$$\mathcal{J}(\Theta_{x\to z}, \Theta_{z\to y}) = \mathcal{L}(\Theta_{x\to z}) + \mathcal{L}(\Theta_{z\to y}) + \lambda\mathcal{R}(\Theta_{x\to z}, \Theta_{z\to y}) \tag{3.9}$$

Where $\Theta_{x\to z}$ is the likelihood of the source-to-pivot, $\Theta_{z\to y}$ is the likelihood of the pivot-to-target and $\lambda\mathcal{R}(\Theta_{x\to z}, \Theta_{z\to y})$ is the connection term where λ is the hyper-parameter used to tune between the likelihoods and connection term.

The authors used pivot word embeddings to make the connection between the models as it is naturally present in both the source and target model parameters. The approach introduces three connection terms used in the training objective to associate the source to the target model.

1. The first connection term encourages both models to generate the same vector representations of the pivot words.

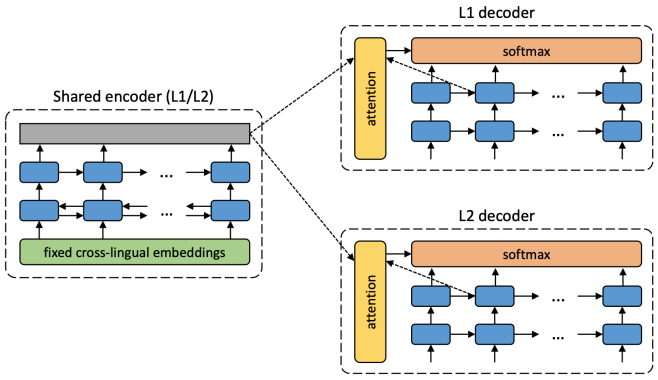

Figure 3.4: Unsupervised NMT architecture (from [Artetxe et al., 2018d])

2. In the second term, the authors introduced a penalization of the Euclidean distance between the two vectors to soften the constraint in the first term because vectors of different languages are not identical.

3. The final connection term assumes that there is a small bridging corpus between the source and the target parallel corpus.

3.6 Unsupervised Neural Machine Translation

The field of Machine Translation has reached state-of-the-art performances [Vaswani et al., 2017, Bahdanau et al., 2015, Sutskever et al., 2014] using Neural Networks. However, as Neural Machine Translation (NMT) systems are data-hungry models that require a considerable amount of parallel corpora to yield good results, low resource languages tend to perform poorly [Koehn and Knowles, 2017b]. To address this problem Artetxe et al. [2018d] introduced a novel method using only monolingual corpora to train NMT systems, achieving BLEU scores of [Papineni et al., 2002a] scores of the tests of 15.56 points (French to English) and 10.21 (German to English).

The Unsupervised NMT uses the standard encoder-decoder architecture [Bahdanau et al., 2015, Sutskever et al., 2014]. In Figure 3.4 there is an example depicting the UNMT architecture. It differs from the standard NMT models in five distinctive aspects:

Dual structure The Unsupervised NMT system translates to both directions rather than being unidirectional (e.g., Chinese to Portuguese or Portuguese to Chinese).

Shared Encoder It uses only one encoder to be shared by both languages. The

shared encoder mechanism aims to reproduce an independent language representation that a decoder will decode when generating the translated text.

Cross-lingual Embeddings Another fundamental change from standard NMT is using fixed embeddings via cross-lingual embeddings that remain fixed during training. The authors proposed these pre-trained fixed embeddings to make the model learn how to generate independent word-level representations. The training model relies on monolingual corpora, so the procedures used to train supervised models are impossible. However, the authors propose two strategies to solve this constraint.

Denoising autoencoding The UNMT model requires some constraints to acquire knowledge. Otherwise, it would be a mere copying task [Artetxe et al., 2019a][2]. To achieve that, the model uses the denoising auto-encoder [Vincent et al., 2010] which works the following way: It injects random noise in the input sentence (i.e., swapping words and performing local substitutions) to force the system to reconstruct the original sentence given the corrupted version. By reconstructing the noisy sentence, the system learns the ins and outs of the language structure.

On-the-fly backtranslation It is an adaptation of back-translation proposed by Sennrich et al. [2016a]. Given a sentence of a language, it translates to another language in inference mode. Then it uses the newly generated synthetic parallel data to train the model to predict the original sentence. The novelty of this proposed system is that it improves the quality of the synthetic parallel data on each iteration. Given its dual structure architecture, during the training phase of each iteration, it performs mini-batch denoising in the two languages and one mini-batch of on-the-fly back-translation of one language to the other vice-versa. This allows the model to upgrade on each iteration.

3.7 Unsupervised Statistical Machine Translation

Lample et al. [2018b] and Artetxe et al. [2018c] made some advances in unsupervised statistical machine translation obtaining 7–10 BLEU points [Artetxe et al., 2018c] more than UNMT. The proposed method takes two monolingual corpora, learns from their n-gram embeddings, and then maps them into a cross-lingual mapping in a fully unsupervised manner using self-learning [Artetxe et al., 2018c] or adversarial learning [Lample et al., 2018b].

[2]Without adding a constraint, the auto-encoder would simply learn to copy the input words. It would not be able to understand the language structure. If the input was a sentence full of random words, the model would simply copy those words in the same order indicating that it did not learn how the language works.

These newly formed cross-lingual phrase embeddings are utilized to induce a phrase-table model. To do so, the method looks up each source phrase and its 100 nearest neighbors in the L2 language pair using cosine similarity to calculate the distance between the source phrase and the possible translation candidates. Then, it applies a softmax function over the cosine similarity to obtain the phrase translation probabilities which is calculated as follows:

$$\phi(\bar{f}|\bar{e}) = \frac{\cos(\bar{e}, \bar{f})/\tau}{\sum_{\bar{f}'} \cos(\bar{e}, \bar{f}')/\tau} \tag{3.10}$$

where $\bar{e}$ is the source language phrase $\bar{f}$ is the translation candidate, and τ a temperature parameter is used to regulate the confidence in the predictions.

Statistical Machine Translation is a log-linear combination of several statistical models where a tuning process is used to optimize the weights. A popular choice to maximize the evaluation metric (e.g., BLEU) in the validation corpus is the Minimum Error Rate Training (MERT)[Och, 2003].

In standard SMT, a parallel corpus would be used to tune and update the weights that maximize the model. However, since using parallel data would violate the constraint of the system being fully unsupervised, the authors proposed two methods to optimize the model using monolingual data.

Unsupervised Tuning

Having trained all these different models, a tuning process is applied to optimize their weights in the resulting log-linear model, which typically maximizes some evaluation metric in a separate validation parallel data. The authors propose back-translation to create a synthetic parallel corpus. A small monolingual corpus (e.g., 10,000 sentences) in the source language is used to create a synthetic parallel corpus by translating it to the target language. This synthetic parallel corpus is then utilized to tune the model, which is subsequently used to translate in the opposite direction (i.e., from target to source). It performs the same in the reverse direction by making small batches of a monolingual corpus, uses to tune the model, and iterating until convergence.

Joint refinement

Refinement is applied to solve performance issues related to the phrase table initially induced with the cross-lingual embeddings. The joint refinement begins by building two synthetic corpora in opposite directions using the initial SMT system. After the synthetic data is created, it extracts phrase pairs from each parallel corpora and induces a new phrase table by taking their intersection. Doing so not only guarantees that the probability estimates of the phrase-table are meaningful but

also discards the ungrammatical phrases initially introduced by the cross-lingual fused phrase-table.

3.8 Evaluation Metrics

Machine Translation evaluation is a hot topic within NLP. The main advantage of using an automatic metric is to evaluate the translation output quickly and inexpensively. However, one of its shortcomings is the lack of correlation with human judgment [Callison-Burch et al., 2006].

There are many evaluation metrics such as METEOR, hLepor, TER, chrF, COMET, BertScore meteor, and BLEU [Papineni et al., 2002b]. Due to its wide acceptance within the scientific field, BLEU was used to evaluate the MT systems in this dissertation.

The core idea of BLEU (Bilingual Evaluation Understudy) is to assign a single numerical score to a translation to tell how good the generated text is compared to one or more reference translations. BLEU works by computing the modified precision of the n-grams. This is done by summing up the count clips, which are the number of times the n-gram appears in the reference text divided by the total n-grams of the generated translation. Equations 3.11, 3.12 and 3.13 present this more formally:

$$\text{Score}(y, \hat{y}) = \exp\left(\frac{1}{N}\sum_{n=1}^{N} P_n(y, \hat{y}) \times \text{BP}(y, \hat{y})\right) \tag{3.11}$$

where y is the reference translation, $\hat{y}$ is the predicted translation, P_n is the modified precision function, and BP is a brevity penalty function.

$$P_n(y, \hat{y}) = \frac{\sum_{ngrams} \text{CountClip}(ngram)}{\sum_{ngrams} \text{Count}(ngram)} \tag{3.12}$$

with CountClip being the minimum between the n-gram count in the predicted sentence $\hat{y}$ and the n-gram count in the reference sentence y, and Count the number of n-grams in the predicted sentence $\hat{y}$.

The brevity penalty is used to penalize translations that are too short using the following calculation.

$$\text{BP}(y, \hat{y}) = \begin{cases} 1 & \text{if length}(\hat{y}) > \text{length}(y) \\ \exp\left(1 - \dfrac{\text{length}(\hat{y})}{\text{length}(y)}\right) & \text{otherwise} \end{cases} \tag{3.13}$$

Chapter 4

Related Work and Low-Resource Language Pairs

In this chapter we address two major topics. The first section covers the related work on the machine translation systems studied in this dissertation. The second section explains the motivation behind the choice of the low-resource language pairs addressed in this work.

4.1 Unsupervised Neural Machine Translation

Unsupervised Neural Machine Translation (UNMT) began to gain traction with the publication of two papers in 2018, one by Artetxe et al. [2018d] and another by Lample et al. [2018a], where the authors proposed a new approach[1] using monolingual corpora only.

When applied to high-resource language pairs, these methods achieved some degree of success [Artetxe et al., 2018d, Lample et al., 2018a, Artetxe et al., 2019a, Sen et al., 2019], however when applied to low-resource language pairs they performed poorly.

Guzmán et al. [2019] obtained BLEU scores of 0.1 and 0.5 in a dataset called Flores on Nepali-English and Sinhala-English. The authors attributed the poor performance to the initialization of the cross-lingual word embeddings.

Marchisio et al. [2020] made an empirical evaluation in dissimilar languages. They found that translation performance deteriorates when the source and target corpora are from different language families. For instance, they observed that the performance for the English-Russian pair was worse than for the English-French pair (i.e., FR-EN loses 2.9 BLEU scores versus the 5.9 loss for RU-EN, when compared to supervised) due to having different scripts.

[1] The only difference between Lample et al. [2018a] and Artetxe et al. [2018d] is that the former incorporate adversarial training.

Kim et al. [2020] studies corroborate the findings of Marchisio et al. [2020] that state-of-the-art unsupervised neural machine translation fares poorly when dealing with language dissimilarity and domain mismatch between source and target language.

Given that phrase-based statistical machine translation (PBSMT) [Koehn et al., 2003] models perform better than UNMT when dealing with scarce labeled data [Lample et al., 2018b], Artetxe et al. [2019a] and Lample et al. [2018b] adapted UNMT to train unsupervised statistical machine translation, ending up to overpass previous state-of-the-art results in UNMT. Both approaches use cross-lingual embeddings from monolingual corpora, which are then used to train the initial phrase model combined with a distortion and n-gram language model.

However, as noted by Artetxe et al. [2019a], unsupervised statistical machine translation has some deficiencies and a new line of research looked into combining unsupervised NMT with unsupervised statistical machine translation into a hybrid approach[Lample et al., 2018b, Artetxe et al., 2019a, Marie and Fujita, 2018, Ren et al., 2019].

Lample et al. [2018b] conducted a study on five language pairs, including low resource (e.g., English-Urdu) and unrelated script (English-Russian). In all of the language pairs examined, the hybrid approach achieved state-of-the-art results compared to the previous approach that used the UNMT architecture.

Marie and Fujita [2018] proposed a new method using supervised NMT framework and synthetic data generated using the unsupervised statistical machine translation, achieving state-of-the-art results on WMT'16 English-German dataset. The authors claim that their approach could outperform previous works in dissimilar languages [Kim et al., 2020] due to assuming relatedness between source and target pairs.

The work of Artetxe et al. [2019a] began as an improvement of their previous unsupervised SMT, where they added subword information and developed a joint refinement procedure. Their proposed system achieved the best results compared to previous studies using the hybrid system USTM/UNMT, improving the BLEU scores by 5.5 points in English-to-German WMT'14 and 7.4 points more in English-German WMT'16 and achieving 0.5 points more than the winning supervised system in 2014.

Unlike the other methods [Artetxe et al., 2019c, 2018c, Lample et al., 2018b, Marie and Fujita, 2018] previously mentioned where SMT was used to initialize the system, Ren et al. [2019] employ SMT during the training phase. Looking to solve the introduction of noise when using iterative back-translation in UNMT, the authors proposed improving the quality of translation by combining SMT with UNMT with a training EM algorithm where SMT boosts UNMT performance by

dealing with the denoising while NMT is concerned with the fluency.

Marie et al. [2019b] proceeded to follow the same framework as Ren et al. [2019] by making some adjustments using both backward and forward translation in the induction of the phrase-model during the training of the USMT whereas previous works only used backward [Artetxe et al., 2019a] or forward [Ren et al., 2019] translation.

To curb the lack of supervision signals in UNMT, Li et al. [2020] proposed the reference language-based UNMT. It uses a pivot language that shares a parallel corpus with the source language and is supposed to aid the translation task through a proposed reference agreement mechanism.

4.2 Pivot Translation

Pivot translation has been widely studied in SMT [Cohn and Lapata, 2007, Wu and Wang, 2007, Utiyama and Isahara, 2007, Bertoldi et al., 2008, Zahabi et al., 2013]. It was adapted to NMT by Johnson et al. [2017]. As mentioned in Section 3.5, the main idea of pivot translation is to bridge the source and language target via an intermediate language that is highly resourceful.

Liu et al. [2018] did a comparative study between the direct and pivot-based approaches. In their experiments, they reported a better performance in both directions for the direct approach, with BLEU scores of 25.11 for ZH-PT and a BLEU of 18.68 for PT-ZH. Meanwhile, with the pivot method, they achieved a BLEU score of 14.60 in the ZH→EN→PT direction and a BLEU score of 11.29 in the opposite direction.

Liu et al. [2019] performed a comparative study using Chinese and English as the pivot languages to translate among 3 languages which were Russian, French and Spanish. They found that using a pivot that is linguistically close to both the source and target languages leads to a performance increase. In their experiments, they noticed that using English as a pivot rather than Chinese improved the model by 12 BLEU points on average.

Santos et al. [2019] achieved state-of-the-art results in the Portuguese-Chinese direction, also using English as pivot, with a score of 17.48 BLEU, surpassing the Google baseline.

We did not find previous works for the Portuguese-Korean direction. However, somewhat extensive literature exists for pivot-based approaches to Korean paired up with other languages. In a similar study by Liu et al. [2019] they compared English and Chinese as the pivot language for Indo-European languages, Paul et al. [2009] conducted comparative experiments using Asian languages. In the total of the eight Asian languages explored, they found that using Asian languages as pivot had a

better performance than using English as the pivot. For example, when the authors tested Korean as the source language, it achieved the highest BLEU points when Japanese was the pivot language and Chinese was the target. In comparison, the BLEU scored slightly lower when they trained Korean-to-Chinese direction using English as the pivot.

Kim et al. [2015] paper focuses on building a direct MT system from Korean to Spanish using a pivot approach for Bilingual Lexicon Extraction. Choi et al. [2018] tackled the low-resource language pair problem by devising a corpus extension on low-resource language pairs and training a multi-source neural machine translation system. The baseline model created was Korean-Arabic. The training data to improve the baseline model was Korean-Arabic, English-Arabic, Japanese-Arabic, and Chinese-Arabic corpora. With their experiments, the authors concluded that using the two approaches in synergy (i.e., augmenting the corpora synthetically and training a multi-source MT) improves the performance of the NMT in a low-resource setting as attested with the Korean-Arabic direction. As shown in the paper, the BLEU of the baseline model was 21.92 points, and with the two approaches together, it jumped over 6 points, achieving a BLEU of 27.07.

4.3 Low-Resource Language Pairs

The conventional NMT systems that achieve a translation output similar in quality to that of human translators require large amounts of good-quality parallel data. However, datasets of this magnitude exist only for very few language pairs. The consequence is that, due to data scarcity, using traditional NMT techniques to train in other language pairs becomes impractical. As noted before, for NMT to achieve comparable results to that of SMT it needs heaps of parallel data [Koehn and Knowles, 2017a]. As such, it is essential to study methods that alleviate the low-resource language (LRLs) problem as it hinders the capability of building robust NMT systems for all existing languages.

There are many indicators to classify a language pair as being low-resource. However, for this dissertation, we followed the criteria from [Lakew et al., 2019, Platanios et al., 2018, Qi et al., 2018] that classifies a language pair to be low-resource if the parallel corpora are below 0.5 million lines.

In our case, we opted for choosing the language pairs Korean-Portuguese and Chinese-Portuguese as it exists an abundance of monolingual corpora for all of them but the available parallel corpora is scarce and of poor quality.

Low-resource languages can be trained in several NMT architectures such as semi-supervised [Edunov et al., 2018, Stahlberg et al., 2018], using data augmentation methods [Platanios et al., 2018, Peng et al., 2020, Sennrich et al., 2016a],

apply transfer-learning techniques [Neubig and Hu, 2018, Johnson et al., 2017, Cooper Stickland et al., 2021] or with zero-shot approaches [Kim et al., 2019, Cheng et al., 2016]. In our dissertation, we employed the unsupervised NMT approach, considered the most extreme case of LRLs where no parallel data is available. Despite already existing small parallel corpora between the studied language pairs, we wanted to test the feasibility of building NMT solutions for extreme settings of dissimilar low-resource language pairs. As such, we constrained only to use monolingual data.

Chapter 5

Implementation

This chapter will discuss the implementation used to build the trilingual MT system.

Section 5.1 discusses the preprocessing, and the corpora we chose to train the systems.

Section 5.2 will present and explain the two approaches studied in this dissertation: (i) Unsupervised Neural Machine Translation and (ii) Pivot-Machine Translation.

Section 5.3 describes in detail how we implemented the two systems, and section 5.4 concludes this chapter with a summary of the work done in constructing these approaches.

5.1 Data and Preprocessing

In this section, we describe the data we used to train both NMT systems and the preprocessing that was needed.

5.1.1 UNMT Corpora

We extracted the datasets used in our translation task from WMT monolingual News Crawl.[1] To train the UNMT system, we extracted 140 million words from the Portuguese dataset and 180 million words from the Korean dataset. As for the Chinese dataset we extracted 30 million words. Table 5.1 reports the size of the datasets. To evaluate the results (i.e., cross-lingual embeddings), we obtained three WordSim-353 datasets, one in Portuguese[2], one in Korean[3] [Park et al., 2018] and one in Chinese[4] [Chen and Ma, 2018]. All datasets are a result of the translation of WordSim-353 and classified according to Agirre et al. [2009].

[1] http://data.statmt.org/news-crawl/

[2] https://portulanclarin.net/repository/browse/lx-wordsim-353/c4e08b72e6dd11e6a2aa782bcb074135a5ac38ba70a14fb3adbd5782b21dacb0/

[3] https://github.com/SungjoonPark/KoreanWordVectors

[4] http://ckipsvr.iis.sinica.edu.tw/cembeval/reg.php

Dataset	Source	Words
Portuguese	NewsCrawl	140M
Korean	NewsCrawl	180M
Chinese	NewsCrawl	30M

Table 5.1: Number of words from the datasets and source from where they were extracted

PT-EN	Domain	Sentences
Scielo	Literature	3M
Europarl	Legal	1.3M
Wikipedia	NetCrawl	38M
TED	Subtitles	300k
Paracrawl	NetCrawl	2.8M

Table 5.2: PT-EN: parallel corpora distribution

5.1.2 Pivot Corpora

For the pivot approach, it was needed to find parallel corpora for Portuguese, Chinese, Korean, and the pivot language, which in this case was English, as it was the only language available among all the studied languages.

Portuguese → English Corpora

The corpora shown in Table 5.2 was taken from OPUS repository[5] [Tiedemann, 2012]. The first dataset has around 3 million sentences and corresponds to translations, from Portuguese to English, of literary works. The second dataset, with around 1.3 million sentences, refers to legal translations retrieved from the proceedings of the European Parliament. The Wikipedia dataset is a compilation of articles written in both Portuguese and English. Followed by that is the TED dataset with around 300 thousand sentences which refers to a transcription of videos from Portuguese to English, and finally, is the Paracrawl dataset, which is a compilation of data crawled from the web, with 2.8 million parallel sentences. In total, 7.8 million sentences were used.

English → Chinese Corpora

We used three datasets to train the NMT model from English to Chinese, with around 2.24 million parallel sentences. The first one was Ted2020, which is comprised of video transcripts. The second is Tanzil, a compilation of Quran texts, and the

[5] https://opus.nlpl.eu

EN-ZH	Domain	Sentences
TED2020	Subtitles	122k
Tanzil	Religious	2M
WMT-News	News	120k

Table 5.3: EN-ZH: parallel corpora distribution

EN-KR	Domain	Sentences
Paracrawl	NetCrawl	4.0M
Tanzil	Religious	93.6k
TED2020	Subtitles	400k
Wikimatrix	NetCrawl	1.3M

Table 5.4: EN-KR: parallel corpora distribution

third is WMT-News which corresponds to data crawled from news articles. In Table 5.3 there is a summarized version of the dataset used to train this model.

English → Korean Corpora

The data shown in Table 5.4 was also obtained from the OPUS repository. In this model, 4 datasets were obtained comprising 5 million parallel sentences. As mentioned above, Tanzil is of religious domain with a total of 1 million sentences. Paracrawl is a collection of data crawled from the internet. Wikimatrix corresponds to articles and texts retrieved from the internet and Wikipedia. At last, Ted2020 is a collection of subtitles from the Ted Talkshow gathered from the year 2020.

As we can see from the Tables 5.2, 5.3 and 5.4, the datasets are of the same domain as we tried to have datasets that were as close to each other as possible. This is because domain mismatch is one of the causes of the error-propagation that often plagues the performance of pivot NMT systems [Cheng, 2019].

5.1.3 Preprocessing

As mentioned in Section 3.1.2, subwording is a process that divides words into smaller units. This preprocessing step is helpful in NMT because it helps to alleviate the out-of-vocabulary (OOV) words. This is due to conventional NMT being inadequate to translate rare words. What subwording helps to form a dictionary based on the partition of words into smaller units that will be used to translate OOV words rather than replace them with a <unk> symbol.

There are many subword implementations like Byte-Pair-Encoding (BPE) [Sennrich et al., 2016b], WordPiece [Schuster and Nakajima, 2012] and Sentence-Piece [Kudo and Richardson, 2018]. In our case we opted for sentence-piece for two rea-

sons: It does not require the raw text to be tokenized in order to implement the algorithm, unlike BPE which assumes a preprocessing to the text has been done before applying it; and it is already incorporated in the OpenNMT ecosystem and, to standardize the same preprocessing steps in both NMT architectures, we chose it so it could be implemented in the unsupervised NMT input.

Despite using the same subwording algorithm, the implementation pipeline differs. For the UNMT, the training was done separately on the monolingual corpus of each language with a vocabulary of 50,000 entries. As for the pivot approach, it is already embedded in the code from the OpenNMT framework.[6] The only essential step was configuring it to choose the vocabulary size and the input data path.

5.2 Unsupervised Neural Machine Translation

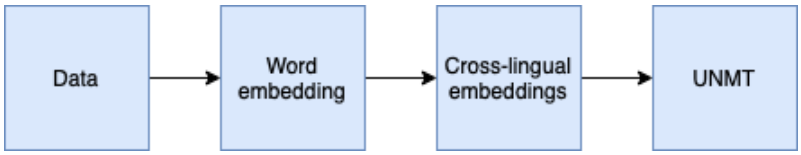

Figure 5.1: UNMT system pipeline

Figure 5.1 provides an overview of the steps needed to train the unsupervised neural machine translation. It starts by gathering large amounts of monolingual data converted to word embeddings. Afterward, the cross-lingual embeddings are generated and fed as inputs to the UNMT system. The quality of the cross-lingual embeddings is vital for the training of the UNMT. This is because cross-lingual embeddings are used to initialize the system, and if so happens to be of poor quality, then the entire model will be corrupted and converge to a poor local optima [Artetxe et al., 2018c].

5.2.1 Monolingual Embeddings

The monolingual embeddings were obtained using **Fasttext**[7]. We followed the same parameters used in the original training, the monolingual embeddings with a cbow model of 10 negative samples, a context window of five words, and embeddings of three hundred dimensions. We opted to use fasttext because it differs from word2vec since the vectors are constructed as an averaged sum of the character n-grams. This has the advantage of dealing with rare and OOV (out of vocabulary) words.

[6] It is the framework chosen to train the NMT models for the pivot approach
[7] https://fasttext.cc/

5.2.2 Cross-lingual Embeddings

After training the monolingual embeddings for each language used in this study, we will use them to initialize the cross-lingual embedding algorithm. The most widely used algorithms are MUSE [Lample et al., 2018a] and VecMap [Artetxe et al., 2018c]. We chose VecMap as it is the current state-of-the-art method. In an empirical study conducted by Glavaš et al. [2019] where several unsupervised cross-lingual embedding algorithms were tested, they found that VecMap had the best performance.

As Section 3.1.1 refers, VecMap is a cross-lingual embedding algorithm that requires only monolingual corpora to project both monolingual embeddings into a shared vector space. Previous methods required a bilingual training dictionary of a few thousand words which could be a strenuous task for low-resource language pairs such as the ones used in this work. To achieve this, the authors proposed a fully unsupervised bilingual dictionary induction based on the assumption of isometry (i.e., corresponding words in different languages have a similar geometric arrangement) and a robust self-learning algorithm. They formulate that similar words in different languages have the same values and use this principle to build an initial dictionary that will improve iteratively until convergence using the self-learning algorithm.

To achieve this, we used the publicly available authors' repository[8] and ran the method for both pairs of languages on the command line interface: Portuguese-Korean and Portuguese-Chinese. We ran it on GPU, taking approximately 10 minutes for each training session to converge.

5.2.3 Unsupervised Neural Machine Translation

Given that the unsupervised Neural Machine translation system uses fixed cross-lingual embedding in the encoder, we had to train both languages using VecMap as described in Section 3. Having done that, we are ready to train the system. We will be reproducing the unsupervised machine translation system of Artetxe et al. [2017], available to test on GitHub.[9] We followed the parameters used in the original paper to train the system with a batch size of 50 sentences and a vocabulary cutoff of up to 50,000 words. As suggested by the authors, this cutoff is needed to deal with out-of-memory issues due to hardware constraints. As the original paper reported, we used Adam as the optimizer and a learning rate of 0.0002, a dropout regularization with a drop probability 0.3 during training, and performed 300,000 iterations to train each language variant.

[8] https://github.com/artetxem/vecmap
[9] https://github.com/artetxem/undreamt

5.3 Pivot Neural Machine Translation

Pivot-based NMT is a two-step process where two systems are trained using source-to-pivot and pivot-to-target data. This approach is an alternative to the direct NMT, where concatenating the two trained systems allows for a better performance than a direct approach with low resources. Since we are concerned with the impact of low-resource languages on dissimilar languages, we trained two pivot systems in the direction Portuguese→Korean and Portuguese→Chinese. One of the requirements to build a pivot NMT is choosing a high-resource language to bridge the two models. We opted out for English as the pivot since it exists large parallel corpora between English and the remaining languages. In total, we trained three different systems, and they followed the general pipeline:

- Portuguese-English-Korean: Since Portuguese is the source language, we start by translating a model from Portuguese→English and then train an independent model from English→Korean. In the end, we concatenate the two models to translate from the source to the target.

- Portuguese-English-Chinese: As the Portuguese→English model has already been trained, we train an independent model from English→Chinese and follow the same procedure mentioned above.

5.3.1 Training Options

The hyperparameters used to train the transformer model are the same as in [Vaswani et al., 2017] with 6 encoder and 8 decoder layers, 8 attention heads, and a word vector size of 512. For a correct initialization of the parameters we set *param_init_glorot* equal to true and *param_init* equal to 0. The batch type used was tokens and we set it to 1000 tokens[10] The optimizer used was Adam, with a learning rate of 2. We performed 500,000 iterations to train each model and performed validation at every 10,000 steps. The entire configuration used to train the pivot-based approach can be consulted in the appendix A.

5.3.2 Open-NMT Framework

Nowadays it exists many frameworks that aid in the implementation of NMT systems. The most common ones are Marian Framework[11] from the University of Ed-

[10] In openNMT configuration guideline, they set it to 4095. However, we faced memory issues and had to reduce the batch size until training converged.

[11] https://marian-nmt.github.io/

inburgh, Fairseq[12] from Meta, tensor2tensor[13] from Google, and OpenNMT[14] from Harvard.

To aid in implementing the pivot model, we adopted the OpenNMT framework [Klein et al., 2017]. This framework was initially developed by the Harvard NLP labs and is currently maintained by SYSTRAN and Ubiqus.

It is implemented in Python and can be used in the two most popular deep learning frameworks: Pytorch and Tensorflow. We chose OpenNMT because it is fast, frequently updated, easy to use, and boasts an extensive array of features that allow the developer to customize its models.

5.4 Summary

The goal of this chapter was to provide an overview of the work done to implement the two NMT systems: (i) Unsupervised Neural Machine Translation and (ii) Pivot approach.

The chapter began by describing the type of parallel corpora chosen (viz., data source, data size, and domain) and the preprocessing steps.

Then, it introduced the implementation of the first model, explaining the steps needed, such as obtaining the monolingual word embeddings and cross-lingual embeddings and then training the UNMT system.

Finally, it presented the latter NMT model, which mentions the training options, the pivot approach pipeline, and the framework used.

The code described here can be found at https://github.com/nlx-group/UNMT-between-PT-and-ZH-KR.git.

[12] https://ai.facebook.com/tools/fairseq
[13] https://github.com/tensorflow/tensor2tensor
[14] http://opennmt.net/

Chapter 6

Evaluation

In this chapter, we present the results obtained in our dissertation.

Section 6.1 shows the results obtained in the UNMT approach, where each subsection corresponds to an evaluation and analysis of our proposed system. Subsection 6.1.1 concerns an ablation study from the original paper of Artetxe et al. [2018b] and a detailed evaluation of our reproduced values. Subsection 6.1.2 contains a replication of [Artetxe et al., 2018d] in the pair of language EN$\Longleftrightarrow$FR and our results in the pair of languages PT$\Longleftrightarrow$ZH and PT$\Longleftrightarrow$KR.

Section 6.2 refers to the results obtained in the pivot-based approach and a qualitative analysis where we extracted three sentences from each language pair to assess the model's translation output better.

The data for replicating the evaluation reported here can be found at https://github.com/nlx-group/UNMT-between-PT-and-ZH-KR/.

6.1 Unsupervised Neural Machine Translation

This section is divided into two subsections. The first subsection refers to the cross-lingual embeddings evaluation. The second subsection provides an in-detail evaluation of the UNMT translation output.

6.1.1 Cross-lingual evaluation datasets

Ablation Study

Following the original paper [Artetxe et al., 2017], we also performed an ablation study. These results are reported in Tables 6.1 and 6.2.[1] A more comprehensive version of the latter table may be seen in Table B.1, in Appendix B.

We conducted five runs per ablation test, and our accuracy results follow the original paper differing only with the unsupervised initialization and CSLS.

[1]The results for English-Finnish were left out because we could not replicate the experiment due to the absence of the dataset for this language pair.

	Pearson	Spearman	s	t
EN-DE	73.88	75.01	1.0	6.0
EN-ES	76.12	76.64	1.0	7.8
EN-FI	—	—	1.0	14.0
EN-IT	60.37	62.64	1.0	6.8

Table 6.1: Ablation study results evaluated using the WordSim 353 dataset.

	EN-DE best	**EN-ES** best	**EN-FI** best	**EN-IT** best
Full System	48.5	37.6	33.5	32.6
Reproduced	48.33	36.98	33.50	47.87
- Stochastic	48.1	37.8	0.28	48.2
Reproduced	48.47	37.87	0.35	48.33
- Cutoff	48.3	35.5	31.9	46.9
Reproduced	31.11	37.07	33.50	48.13
- CSLS	0.0	0.0	0.0	0.0
Reproduced	48.33	37.0	33.30	36.93
Bidirectional	48.3	36.2	31.4	46.0
Reproduced	48.53	47.87	33.50	36.93
- Re-weighting	48.1	36.0	32.9	46.1
Reproduced	47.13	36.33	33.50	47.73

Table 6.2: Ablation study results of [Artetxe et al., 2017] and our reproduction, evaluated using bilingual lexicon induction. Full table in Table B.1, in Appendix B.

The accuracy reported for CSLS in the original paper can be explained due to a bug found in the authors' code. The code available by the authors does not contemplate the method to run the unsupervised initialization. Thus, it was unclear how to reproduce it and opted not to include it in the ablation study.

As for the vocabulary cutoff, we also faced some constraints due to computational limitations. In the original paper, the authors conducted the ablation study by setting $K = 100$. However, when we tried reproducing it, we ran against GPU memory issues. As a solution, we were faced with two possibilities: either run it in the CPU or reduce the K parameter. We opted for the second option due to the expensive computation cost of running in CPU, which would take more than 3 days to achieve convergence as tested by Garneau et al. [2020].

Cross-lingual embeddings

To evaluate the results of our experiment, we needed to have a gold standard dataset to assess the quality of the embeddings. In the original paper, the authors used bilingual dictionaries to obtain the accuracy of the cross-lingual embeddings. Instead, we evaluated our task using cross-lingual word similarity datasets. The reason is that we wanted to maintain the method fully unsupervised (i.e., recreate the method without any bilingual signal), and to the best our knowledge there are no bilingual dictionaries in the pairs of languages used in this experiment. To create the cross-lingual word similarity datasets, we performed the following steps:

1. **Obtain WordSim 353**: This is a gold standard dataset developed by Finkelstein et al. [2002] and used to evaluate word similarity and relatedness in English. It contains 353-word pairs associated with an average of 13 to 16 human judgments. The dataset is divided into two subsets: The first subset has 153-word pairs evaluated by 13 human evaluators, and the second subset has 200-word pairs evaluated by 16 subjects. Each word pair receives a score from the annotators on a scale from 0 to 10, where 0 means unrelated and 10 refers to the same word.

2. **Using the Camacho-Collados et al. [2015] framework**: This framework is a generalization of the method of Kennedy and Hirst [2012], which created a manual cross-lingual word similarity data-set from two monolingual word similarity dataset in French-English. Camacho-Collados et al. [2015] expanded this to any pair of languages done automatically. This is done by pairing two monolingual datasets that were previously aligned. Then the algorithm seeks to pair them into a new paired dataset *a'-b'* using a mapping function that attributes a new similarity value (e.g., 0–4 in RG-65). It discards new pairs *a'-b'* if the difference between the scores is a quarter more significant than the

PT	KR		PT	ZH	Score
computador	키보드	7.62	advogado	法律	8.38
manteiga	빵	6.19	valores	市场	8.08
jaguar	고양이	7.42	telemóvel	股票	1.62
minoria	평화	3.69	reserva	电话	1.62
atraso	뉴스	3.31	CD	股票	1.31
centro	학교	3.44	estrela	电影	7.38
atividade	여행	5.0	utensílio	工具	6.46
roupeiro	옷	8.0	ave	鹤	7.38
lucro	미디어	2.88	acorde	微笑	0.54
software	컴퓨터	8.5	mágico	玻璃	2.08

Table 6.3: Sample pairs from the PT-KR and PT-ZH cross-lingual word similarity datasets (PT: Portuguese, KR: Korean, ZH: Chinese).

similarity scale size (e.g., 1.0 in RG-65). However, if smaller, a new aligned pair is merged with a score equal to the average of the original similarity scores.

Following this procedure, we created two cross-lingual word similarity datasets based on the WordSim-353 in Portuguese-Korean and Portuguese-Chinese. To create them we extracted from the internet three monolingual WordSim-353 datasets in Portuguese[2], Chinese[3] [Chen and Ma, 2018] and Korean[4] [Park et al., 2018]. Since they are all translated from the original English WordSim-353 [Finkelstein et al., 2002] data set they are aligned and thus can be used in the framework of Camacho-Collados et al. [2015] to generate cross-lingual word similarity data-sets. Table 6.3 shows the sample pairs with their corresponding similarity scores from two of the cross-lingual data sets.

Cross-lingual Embeddings on PT-KR and PT-ZH

The results of our experiments are reported in Tables 6.4 and 6.5. The first thing that stands out is the discrepancy in the values between Portuguese-Korean and Portuguese-Chinese. Although the scores of both languages are not statistically significant, the gap between them is moderately broad. Though we do not have a specific reason for this, we can speculate that one of the contributing factors is the vocabulary size of the Chinese word embeddings.

Despite everything, the values obtained seem to be following the current literature as noted in [Vulić et al., 2019], where the authors noticed that unsupervised

[2] LX-WordSim-353 (https://hdl.handle.net/21.11129/0000-000B-D38E-7)
[3] http://ckipsvr.iis.sinica.edu.tw/cembeval/reg.php
[4] https://github.com/SungjoonPark/KoreanWordVectors

PT-ZH						
	Pearson (best)	Spearman (best)	Pearson (avg)	Spearman (avg)	s	t
VecMap	8.66	9.69	8.0	9.0	1.0	6.8
Stochastic	–	–	–	–	–	–
Cutoff	12.97	12.72	11.36	10.73	1.0	25
CSLS	15.64	11.84	16.55	13.60	1.0	14
Bidirectional	8.4	9.60	9.69	7.2	1.0	8.2
Re-weighting	9.60	9.69	7.2	8.2	1.0	8.4

Table 6.4: Ablation study conducted in the language pair Portuguese-Chinese.

PT-KR						
	Pearson (best)	Spearman (best)	Pearson (avg)	Spearman (avg)	s	t
VecMap	42.46	40.79	40	41	1.0	6
Stochastic	44.00	43.51	41.81	41.35	1.0	14
Cutoff	44.00	43.51	41.81	41.35	1.0	14
CSLS	40.69	38.54	40.32	38.22	1.0	14
Bidirectional	40.52	38.27	34.78	32.52	1.0	13.31
Re-weighting	40.52	38.27	34.78	32.52	1.0	19.6

Table 6.5: Ablation study conducted in the language pair Portuguese-Korean.

word translation methods fail when language pairs are distant, and in [Garneau et al., 2020], where they tested the method using four languages that are distant from English, concluding that the method failed when dealing with Latvian and Vietnamese languages.

We also conducted an ablation study as in the original paper to test the robustness of the hyperparameters in dissimilar languages. Unlike Artetxe et al. [2018b], who reported that the method did not converge without stochastic procedure in dissimilar languages, we did not find a drop in performance in our experiments. Nor did the system struggle to converge when CSLS was turned off, unlike what was noticed by Garneau et al. [2020].

We want to highlight that our method was evaluated with a cross-lingual similarity dataset instead of bilingual lexicon induction[5]. To compare the results of our

[5]Cross-lingual embeddings evaluated using a bilingual lexicon induction are focused on measuring the accuracy of word translations and use a dictionary to induce the evaluation. Cross-lingual similarity data-sets evaluate the semantic similarity between the alignments and do not require a bilingual dictionary, thus making it more suitable for our task.

cross-lingual embeddings from the original paper, refer to Table 6.1. At first glance, these results show that the model performance is robust enough to changes done in the hyperparameters.

Hyperparameters analysis. Following the recommendation of Garneau et al. [2020], we also decided to test the VecMap framework robustness by conducting experiments in the hyperparameters where we altered a key hyperparameter and let the remaining ones stay fixed. The critical parameters chosen were the following: (1) Number of neighbors in CSLS, (2) frequency-based vocabulary cutoff, and (3) stochastic initialization p value.

Below, we explore in detail the impact on the performance of this assessment in all of the tested languages (e.g., original paper and our experiments) and the average runtime cost.

CSLS. We varied the k parameter within a range between 1 and 10. The results are reported in Tables 6.2, 6.4 and 6.5. We noticed that the values did not show a significant variation for all the tested language pairs. However, as expected, the performance tends to drop as k increases, as well as the runtime and number of iterations.

Frequency based vocabulary cutoff. To conduct the experiments under this parameter, we tested several k values ranging from 10,000 to 50,000 with an increment of 10,000 per run. As mentioned above, we had some initial difficulties implementing the cutoff when k parameter passed a threshold above 50,000.

As expected, when the method was initialized with the lowest k value, the accuracy dropped, and the runtime was faster. The only exception occurred when running EN-IT, where the highest value obtained was when k is 10,000. However, it remained remarkably stable throughout all the runs with variations around 1%. In other language pairs, the model's accuracy grew as k increased.

Our findings corroborate the results of Artetxe et al. [2018b], who noticed an improved model performance as k was set to a higher value minus in the language pair en-es where it dropped when vocabulary cutoff increased.

Stochastic dictionary induction. To conduct this experiment, we opted for values between 0.05 and 0.3 for the initial keep probability (p_0). As for p's growth factor p_{factor} we opted for a linear space of values between 1.5 and 3. Compared to the entire system reproduction, the stochastic induction showed a slight performance increase across all language pairs minus in EN-IT. Nonetheless, the variation was minimal (less than 1%).

Languages	BLEU
EN-FR	8.47
FR-EN	7.45

Table 6.6: BLEU scores in newstest2014 in language-pair EN-FR

Throughout all the runs conducted, we verified that the model performs best when the initial factor probability (p_0) is set to 0.1, which is the default parameter in the original paper.

6.1.2 Unsupervised Neural Machine Translation

This section will present the results of our study conducted in the pair of languages PT-ZH and PT-KR. Moreover, a subsection is dedicated to a reproduction study from the original paper in the language pair EN-FR.

Reproduction study

For the English-French dataset, the authors used the News Crawl[6] corpora articles from 2007 to 2013. Table 6.6 shows the results obtained in our reproduction.

The original paper obtained a BLEU of 9.98 in the direction FR-EN and a BLEU of 6.25 in the direction EN-FR using the baseline model. To conduct these experiments, we gathered data from the same source as the authors. However, we circumscribed the data to the news articles from the years 2007 to 2009 due to memory limitations.

To evaluate the unsupervised system, we used the test dataset NewsCommentary v16[7] which contained around 2,000 sentences. We obtained a BLEU of 7.48 in the direction FR-EN and a BLEU of 8.47 in the direction EN-FR, as presented in the table 6.6. Judging from our results, we were able to reproduce with success the UNMT system.

Low-Resource Language Pairs Evaluation

Tables 6.8 and 6.9 present the scores obtained from the unsupervised neural machine translation. Due to the dual encoder nature of the UNMT architecture, we could evaluate this approach's performance in both directions. Based on these results, it is possible to conclude that UNMT is not a viable approach for low-resource language pairs. Our results seem to be following the current literature regarding this approach.

[6] https://data.statmt.org/news-crawl/

[7] https://opus.nlpl.eu/News-Commentary.php

Source	Reference	System
L'or à 10.000 dollars l'once ?	10,000 Gold?	to 10,000 dollars
Mais devinez ce qui s'est passé ?	Wouldn't you know it?	But what would come ?
Une réponse est bien sûr un effondrement complet du dollar.	One answer, of course, is a complete collapse of the US dollar.	An answer is like a full collapse of the dollar.

Table 6.7: Translation output sample from FR→EN

Languages	BLEU
PT-ZH	0.32
PT-KR	0.69

Table 6.8: UNMT BLEU scores for PT→ZH and PT→KR

Kim et al. [2020] researched the feasibility of the unsupervised NMT. In their research, they tested 5 languages and found that on dissimilar language pairs, they obtained a BLEU of less than 3 points. As previously mentioned in Section 4.1, Guzmán et al. [2019] reported similar BLEU to ours in distant language pairs. In the Nepali-English direction, they obtained a BLEU of 0.5 and a BLEU of 0.1 in the direction English-Nepali and English-Sinhala.

In an empirical study conducted by Marchisio et al. [2020], where the authors stress-tested UNMT under several situations, including the behavior of UNMT under dissimilar language pairs with different scripts, like Guzmán et al. [2019] they also tested UNMT on the English-Nepali and English-Sinhala on a dataset provided by Facebook and obtained BLEU scores as low as 0.2.

A plausible explanation for these low scores is the poor cross-lingual embedding initialization. Since UNMT is initialized with the cross-lingual embeddings, which are used to provide the bilingual signal needed to initiate the iterative back-translation, the quality of the UNMT depends on the cross-lingual embeddings. If they have a poor degrading performance, then UNMT will not yield favorable results.

Cross-lingual embeddings have a poor performance when dealing with distant

Languages	BLEU
ZH-PT	0.07
KR-PT	0.59

Table 6.9: UNMT BLEU scores for ZH→PT and KR→PT

language pairs [Vulić et al., 2019, Bojanowski et al., 2016, Glavaš et al., 2019, Hoshen and Wolf, 2018]. The reason is that modern cross-lingual algorithms are built upon the hypothesis that languages are isomorphic, that is the assumption that words in different languages have a similar geometric distribution.

Despite this holding true in several languages, it fails for distant language pairs. For instance, Bojanowski et al. [2016] noticed different structures for the word *girl* in English and Japanese; Hoshen and Wolf [2018] did a similar study where they mapped orthogonal mappings using fastText embeddings and noticed an 81% accuracy for English-Spanish and a low 2% accuracy for English-Japanese; and Vulić et al. [2019] noticed the worst performers were Korean, Thai, and Basque, whose morphology is most distinct, while the best performers were found with similar languages and that none of the best performers in the fully unsupervised approach surpassed the weakly supervised method.

Our hypothesis that the failure of UNMT is due to the weak cross-lingual embeddings performance is backed up by [Marchisio et al., 2020], who in their study explain the poor performance of EN-SI and EN-NE as being due to "weak isomorphism" between dissimilar languages.

As a comparison, the BLEU of the EN→DE was around 6.89 points,[8] and the cross-lingual embeddings were around 73.88% in a Pearson's scale as shown in Table 6.1. Whereas in the PT→ZH direction the cross-lingual evaluation was around 9% and the BLEU of the UNMT was less than 1.

Qualitative Analysis

To better assess the quality of the UNMT translation output, we selected around 30 parallel sentences and manually analyzed them in the PT⟺ZH and PT⟺KR directions. Tables 6.10 and 6.11 present a small sample of the translation output in both directions for the language pairs studied in this dissertation.

At first glance, our analysis shows that the model can decode the target language but fails at conveying its meaning. For instance, in Table 6.10 in the first example, it translates *Uma Europa Alemã* into 在哪里 which roughly means *where?*. On the next example, it fails to translate *Líderes para um Mundo sem Liderança* into 实际上, 这也并非易事 which means *it is not easy in reality* when the correct translation should have been 领导缺乏领袖的世界 (*Leaders of a World without Leadership*).

On the same note, we noticed that the model seems to be able to capture the topic of the translating sentences but acts poorly at decoding the correct meaning. This example can be verified in Table 6.11 where in the first line, the reference text mentions a golf tournament and the proposed system comments on American football.

[8] This value is from the original paper [Artetxe et al., 2018d].

Source	Reference	System
Uma Europa Alemã?	德国的欧洲?	在哪里?
Líderes para um Mundo sem Liderança	领导缺乏领袖的世界	实际上, 这也并非易事.
Por exemplo, o Presidente Francês Nicolas Sarkozy impediu uma posição Europeia comum relativamente à proposta de soberania Palestiniana às Nações Unidas em Setembro. Por outras palavras, mesmo quando a França lidera, não o faz sempre de uma maneira construtiva.	但法国的单边主义做法往往使其与欧洲伙伴敌对。例如, 9 月份, 法国总统萨尔科齐在联合国巴勒斯坦建国投票中抢占了欧洲共同立场。换句话说, 即使由法国来领导, 它也并不总是用这样建设性的方式.	王沪宁在当天上午在全国生态保护工作会议上表示, “在长江中下游面临的“太严重”. 问题”最近, 长江经济带的发展, 也将成为 “一带一路” 的重要组成部分 。.

Table 6.10: UNMT translation output sample from PT→ZH

These results are a stark contrast with the translation output of the EN→FR system, as shown in Table 6.7, where the model was able to output a translation that is faithful to the meaning of the original sentence, despite the ungrammatical phrase constructions.

6.2 Pivot Neural Machine Translation

In this section, we present the results from the pivot-based approach in the language pairs PT→ZH and PT→KR. The results are shown in Table 6.12.

The pivot approach obtains 17.28 BLEU points in the PT→KR translation direction. To evaluate it, we used 2,000 parallel sentences from the Tanzil dataset[9] [Tiedemann, 2012] test set. For the language pair PT→ZH we used the News Commentary v11 test set available at Opus NLP[10]. As highlighted by Santos et al. [2019], to evaluate the PT→ZH approach, we had to use the jieba[11] tokenizer on the test set and in our proposed translation. Failing to do so would cause a decrease in the BLEU score due to the algorithm being based on a white-space token overlap.

[9] https://opus.nlpl.eu/Tanzil-v1.php
[10] https://opus.nlpl.eu/News-Commentary-v11.php
[11] https://github.com/fxsjy/jieba

Source	Reference	System
일반적으로메이저대회로알려져있고간단히메이저라고도불리는남자메이저골프대회는프로골프에서가장권위있는네개의연간토너먼트이다.	Os principais torneios masculinos do golfe, geralmente conhecidos como Major Championships, e muitas vezes referidos simplesmente como majors, são os quatro prestigiados torneios anuais de golfe profissional.	O percurso da seleção feminina de ouro, que chegou a um ano após a 2000, durante a medalha de prata da equipa da pela equipa inglesa de futebol americano .
그래서이들을먼저시험한다.	É por isto que estou contando a vocês primeiro.	Foi e se dedica à instalação.
그두지역사이에는초등학교하나가위치한다.	Estudo primário entre as duas cidades .	Também há uma participação na faculdade.

Table 6.11: UNMT translation output sample from KR→PT

Languages	BLEU
PT-KR	17.28
PT-ZH	13.28

Table 6.12: Pivot based approach BLEU scores for PT→KR and PT→ZH

Qualitative Analysis

Similar to the analysis done in Section 6.1.2 where we analyzed the output of the UNMT, we decided to follow the same evaluation on the pivot-based approach. Since the pivot-based approach was trained on the direction PT→KR and PT→ZH, we will only present an evaluation of this direction.

Table 6.13 shows an excerpt of the translation output of the pivot model from PT→ZH.

In the first example, the model can accurately translate the Portuguese phrase *O que falhou em 2018?* into Chinese. The only catch is that our proposed system added an unrelated word 共和党 which translates to *Republican Party*. As we can verify in this example, the Portuguese phrase which means *What happened in 2008?* makes no reference to political parties. Despite translating this phrase accurately, it added an unnecessary detail.

As for the second example *Um consenso de Berlim?* which in English means *A consensus from Berlin?* the model was able to translate perfectly into Chinese, even translating the city name Berlin into its official Chinese transliteration.

We also included an example of a lengthy sentence to test how the model would behave. At first look, the proposed system behaved well in dealing with a long sentence. Some examples are noticeable with the correct translation of European Union ("União Europeia" in Portuguese), which our system translated to 欧盟. Another example is accurate translation into Chinese of this complex sentence *Com efeito, as regras de origem (...), revelaram-se problemáticas em alguns dos anteriores acordos de reconhecimento da UE.* which refers to the excessive bureaucracy of EU in the origin of products.

We also detected some deficiencies. For instance, in the original sentence *Embora uma laranja de origem brasileira, cuja venda é permitida em Portugal* there is a reference to Brazilian oranges; however, our system failed to mention Brazil when referring to the oranges in the passage 尽管可以向葡萄牙出售的子橙.

The main difference between the reference text and our proposed system is that our system is a literal translation of the Portuguese text, whereas the reference strays away from the source sentence, writing it in a more target audience-focused way. The goal of translation is to produce a text that sounds fluent in the target language so that the audience does not realize that the text they are reading is a product of a translation. Despite our proposed system capturing the meaning of the source text, it still struggles in translation fluency and readers might find some passages awkward.

We also incorporated a detailed analysis of the PT→KR, which can be consulted in Table 6.14. In the first example, the model did poorly translating the source sentence into Korean. Compared with the reference, our proposed system fails to

capture the meaning of the Portuguese sentence. In the reference sentence 뒤에남아있던늙은한여인과 which roughly translates to *an older woman who was left behind* our proposed system outputs two words that are not part of the Korean dictionary.

Regarding the second example, there are some mismatches between the reference system "하나님께 권능이있으시니라" and our proposed system "사람들중에는그들은아니거늘." The reference text says *Do not think God will break his promise to his seekers. Truly, God has the power of retribution.*. The proposed translation means *There are many among people that only pretend to believe in God and the afterlife; however, they are genuinely not.*

The final example contains some mismatches between the reference and our proposal system. Interestingly, the original text contains an error regarding the misspelling of Ezra as 아자르 (reads as ajare) when in fact, it should have been written as 에즈라 (ejera). The reference text can be translated into *Does Abraham worship the idol of his father Ezra as a god? You and your people are wrong* whereas our proposed system translated into *He created you, shaped you, and gave you balance.* From this, we can observe that our proposed system understands that God is mentioned in the original text. However, it fails to do an accurate translation. The original text questions whether Abraham worships his father, Ezra, as a god, whereas our proposed system translated into God having created the subject and shaped him.

Source	Reference	System
O que falhou em 2008?	2008 年败在何处?	共和党,2008 年什么失败?
Um consenso de Berlim ?	柏林共识?	柏林达成共识?
Com efeito , as regras de origem excessivamente restritivas, como por exemplo as que regem as normas dos serviços profissionais, revelaram-se problemáticas em alguns dos anteriores acordos de reconhecimento da UE. Embora uma laranja de origem brasileira , cuja venda é permitida em Portugal, possa ser vendida em toda a UE, um engenheiro ou contabilista de nacionalidade brasileira habilitados para exercer em Portugal, têm de cumprir exigências distintas em matéria de autorização se quiserem exercer a sua profissão noutro país da UE, dificultando a tão necessária mobilidade laboral, obrigando os trabalhadores não-europeus a enfrentar procedimentos burocráticos dispendiosos e ineficazes .	事实上, 欧盟此前的不少承认协议 (比如在专业服务标准方面) 表明, 过度约束的来源地规则很有问题大有问题。尽管获准在葡萄牙出售的巴西橙可以在整个欧盟出售, 但获得葡萄牙执照的巴西工程师和会计师仍必须满足欧盟其他地区各自的执业要求, 这迫使非欧洲工人忍受费时费力的官僚主义流程, 不利于急需的劳动力流动。	事实上, 过度限制性来源的规则, 如治理专业服务规则的人, 在此前的欧盟识别协议中, 被证明是有问题的。尽管可以向葡萄牙出售的子橙, 可以在整个欧盟中, 但大多是工程师和会计师让葡萄牙得以在葡萄牙锻炼, 如果它们想在另一个欧盟国家执业时满足不同的要求, 从而迫使非欧洲劳工流动。

Table 6.13: Pivot-based approach output sample for PT→ZH

Source	Reference	System
Exceto uma a anciã, que foi deixada para trás.	뒤에남아있던늙은한여인과	오직와! 만도음!
Nunca penseis que Deus falte à promessa feita aos Seus mensageiros, porque Deus é Punidor, Poderosíssimo.	하나님께그분의선지자들에게약속을어길것이라생각치말라실로하나님은응보의권능이있으시니라	사람들중에는하나님과내세를믿는척말하는무리가있으나실로그들은아니거늘
Quando Abraão disse a Ezra, seu pai: Tomas os ídolos por deuses ? Eis que te vejo a ti e a teu povo em evidente erro.	일러가로되아브라함이그의아버지아자르깨우상을신으로모시나이까당신과그리고당신백성은분명히잘못하고있습니다.	그분께서너희를창조하고형상을만든후균형을주시었고.

Table 6.14: Pivot-based approach output sample for PT→KR

Chapter 7

Conclusion

The work conducted in this dissertation allowed me to grasp the basics of Machine Translation and understand a range of approaches and techniques which can be used when dealing with low-resource languages. As covered in this dissertation, there is still much work to be done to reach a situation where low-resource languages can be as competitive as their high-resource language counterparts. Nonetheless, we will mention some future research guidelines that can help mitigate the performance gap between low and high-resource language pairs.

This Chapter concludes our dissertation. It summarizes the main results in Section 7.1. Then, Section 7.2 lists the contributions of this dissertation. The final Section 7.3 closes this work by providing some guidelines for future research.

7.1 Summary

This dissertation aimed to study the feasibility of devising a machine translation system using only monolingual corpora. For this purpose, we tested two architectures: (i) Unsupervised Neural Machine Translation and (ii) Pivot-based approach. As for the languages chosen, we wanted to recreate a realistic setting. Thus we opted for low-resource language pairs that are from distant language families. By choosing PT→KR and PT→ZH, we get an insight into how unsupervised techniques behave in this setting, and a reference on how to proceed should be needed in the future to build an MT system for distant low-resource language pairs.

With a BLEU of 17.28 points for the PT→KR and a BLEU of 13.37 for PT→ZH, it is possible to gauge that the pivot-based approach is the most suitable choice for both language pairs. The results of the UNMT, as shown in Table 6.9 are aligned with the current literature [Marchisio et al., 2020, Guzmán et al., 2019, Kim et al., 2020] where they also tested an unsupervised based approach on low-resource languages and obtained low BLEU scores.

The code and data for the experiments reported here can be found at the footnote

below. [1]

7.2 Contributions

The major contributions made in this dissertation are as follows:

- **An exploratory study in dissimilar low-resource language pairs**
  To study the feasibility and robustness of building NMT systems using only monolingual corpora, two NMT systems were devised: (i) Unsupervised Neural Machine Translation; (ii) Pivot approach. As mentioned before, we concluded that UNMT is unfeasible for distant languages due to the cross-lingual embedding initialization that leads to poor performance of the model.

- **A comparative study of cross-lingual embeddings**
  Cross-lingual embeddings play a major part in the success of the UNMT approach. This is noticeable in our results that highlight the differences between the cross-lingual scores of Indo-European languages and their BLEU scores on the UNMT system versus our experiments. An extensive body of literature [Doval et al., 2020, Søgaard et al., 2018, Glavaš et al., 2019] has been conducted in studying the impact of distant languages in cross-lingual embedding algorithms. The consensus is that cross-lingual embeddings provide unsatisfactory solutions to distant language pairs due to a lack of isomorphism between the languages in the pair. Our results (cf. Tables 6.8 and 6.9) align with those found in the current literature.

- **The creation of competitive PT→ZH and PT→KR NMT systems**
  Using the pivot-based approach on the direction PT→KR and PT→ZH, we achieved competitive scores on a methodology that did not require parallel corpora between the tested languages. It achieves 17.28 BLEU on PT→KR and 13.37 BLEU on PT→ZH..

7.3 Future Work

Despite the pivot-based approach having obtained good scores, there is still work to improve the accuracy of MT models for low-resource language pairs. With that in mind, we present two research topics that can be further explored to improve the performance of unsupervised machine translation.

[1] https://github.com/nlx-group/UNMT-between-PT-and-ZH-KR.git - Mestrado de Informática - NLX-Group

7.3.1 Hybrid model

As mentioned in Section 3.7, the hybrid model is a viable architecture that connects statistical machine translation with neural machine translation. The primary purpose of using SMT is to provide a better initialization which was the major degrading factor in UNMT.

This approach follows the idea of using a phrase table to initialize the model and then feeding the phrase table to a standard NMT system that is improved iteratively. The idea of using a phrase-table comes from research conducted by Artetxe et al. [2018c] and Lample et al. [2018b], where it was noticed that UNMT systems were superseded by SMT approaches. This phrase-table is induced through cross-lingual embeddings in combination with an n-gram language model and is improved iteratively via back-translation.

Recent work [Artetxe et al., 2019c, Lample et al., 2018b] into hybrid models has yielded better results than UNMT approaches. Artetxe et al. [2019c] achieved 5.5 points more than the previous state-of-art unsupervised approach, obtaining a total score of 22.5 BLEU points in English-to-German WMT 2014.

This technique could prove its usefulness for PT $\leftrightarrow$ ZH and PT $\leftrightarrow$ KR language pairs since it only requires monolingual data and has achieved far better scores than the previous state-of-art UNMT system.

7.3.2 Cross-lingual embeddings

The assumption of approximate isomorphism between languages is the a critical factor the method of Artetxe et al. [2018b]. As mentioned before, this assumption assumes that words in different languages share a similar geometric composition. However, this isomorphism can only be attested for similar languages and domains, and for distant language pairs it leads to a degradation in performance.

One of the lines of research to mitigate this degrading factor is to explore new methods that can increase the isomorphism of monolingual spaces, as presented by the work of Zhang et al. [2019]. Many current cross-lingual embeddings methods [Conneau et al., 2017a, Artetxe et al., 2018b] introduced an orthogonal constraint into their algorithms. It is argued that by adding this constraint, the original structure of the monolingual embeddings is maintained and enriched by connecting with the other languages. Nonetheless, the success of using orthogonal methods is still tied to isomorphism. Concerned with that, Zhang et al. [2019] proposed an iterative approach that normalizes the length of the monolingual embeddings to make them more similar, which in return improves the alignment. The method is called Iterative Normalization and is backed up by two properties: *length-invariance* and *center-invariance*.

Length-Invariance: This property refers that all vectors should be of equal length. This constraint is added to solve an inconsistency in the cross-lingual embedding algorithms where the training objective is to maximize the dot product and cosine similarity by minimizing the Euclidean distance. In addition, length invariance satisfies another prerequisite of bilingual orthogonal mapping that refers to translation pairs should be of equal length.

Center-Invariance: It mentions that the mean vector of different languages should have the same length. This proposal has already been exploited by Artetxe et al. [2016], who reported that zero-mean centering (i.e., ensuring that the mean vector is zero) improved dictionary induction. The only difference between the approach of Artetxe et al. [2016] and that of Zhang et al. [2019] is the motivation behind zero-mean centering. Artetxe et al. [2016] assume that two randomly picked words should not be similar or dissimilar. However, as Zhang et al. [2019] argue, there is not an even semantic distribution of words as there are words prone to having more synonyms than other words. Instead, zero-mean should be implemented because it satisfies the center-invariance constraint.

Zhang et al. [2019] tested their approach using MUSE [Conneau et al., 2017a] dictionaries and reported a performance increase when Iterative Normalization was applied. The most exciting finding was a 40% accuracy improvement when their method was applied in the language pair English-Japanese.

Since this method provides a solution for orthogonal mapping for language pairs where isomorphism is not found and presented by the results in the dissimilar language pair English-Japanese, it could be advantageous to map the cross-lingual embeddings on this approach and then train a UNMT system. As the performance of UNMT is bound to the quality of the cross-lingual embeddings, it can be assumed that this would lead to a better translation output for dissimilar languages.

Appendix A

Hyper-Parameters for Training

The following hyperparameters were used to train the Transformer model on the openNMT framework:

- Train steps: valid steps: 10000 train steps: 500000
- Batching: queue size: 10000 bucket size: 32768 world size: 4 gpu ranks: [0] batch type: tokens batch size: 4096 valid batch size: 8 max generator batches: 2 accum-count: [4] accum-steps: [0]
- Optimization: model-dtype: fp32 optim: adam learning rate: 2 warmup-steps: 8000 decay-method: noam adam-beta2: 0.998 max-grad-norm: 0 label-smoothing: 0.1 param-init: 0 param-init-glorot: true normalization: tokens
- Model: encoder type: transformer decoder type: transformer position encoding: true encoder layers: 6 decoder layers: 6 heads: 8 rnn size: 512 word vector size: 512 transformer-ff: 2048 dropout-steps: [0] dropout: [0.1] attention dropout: [0.1]

Appendix B

Ablation Study on Cross-lingual Embeddings

	EN-DE				EN-ES				EN-FI				EN-IT			
	best	avg	s	t	best	avg	s	t	best	avg	s	t	best	avg	s	t
Full System	48.5	48.2	1.0	7.3	37.6	37.3	1.0	9.1	33.5	32.6	1.0	12.9	48.5	48.1	1.0	8.9
Reproduced	48.33	48.33	1.0	6	36.98	36.98	1.0	6.8	33.50	33.50	10	14	47.87	47.87	1.0	7.8
- Stochastic	48.1	48.35	1.0	2.5	37.8	37.8	1.0	2.6	0.28	0.28	0	4.3	48.2	48.2	1.0	2.7
Reproduced	48.47	48.33	1.0	5.2	37.87	37.29	1.0	4.46	0.35	0.35	0	5.2	48.33	48.18	1.0	4.33
- Cutoff	48.3	48.1	1.0	105.3	35.5	34.9	1.0	185.2	31.9	30.8	1.0	162.5	46.9	46.5	1.0	114.5
Reproduced	31.11	29.16	1.0	28	37.07	36.6	1.0	7.7	33.50	33.50	1.0	14	48.13	47.97	1.0	35
- CSLS	0.0	0.0	0	13.6	0.0	0.0	0.0	14.1	0.0	0.0	0.0	13.1	0.0	0.0	0.0	15.0
Reproduced	48.33	48.33	10	6	37.0	36.71	1.0	9.4	33.30	32.2	1.0	15	36.93	36.98	10	7.67
Bidirectional	48.3	48.0	1.0	5.5	36.2	35.8	1.0	7.3	31.4	24.9	0.8	5.6	46.0	45.4	1.0	5.6
Reproduced	48.53	48.13	1.0	3.8	47.87	47.87	1.0	6.8	33.50	33.50	10	14	36.98	36.93	1.0	8.1
- Re-weighting	48.1	47.4	1.0	7.0	36.0	35.5	1.0	9.1	32.9	31.8	1.0	11.2	46.1	45.6	1.0	8.4
Reproduced	47.13	47.13	1.0	7.5	36.33	36.33	10	6.8	33.50	33.50	1.0	5.50	47.73	47.20	1.0	6.8

Table B.1: Full table results on the ablation study conducted on cross-lingual embeddings.

Bibliography

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/N09-1003.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. Massively multilingual word embeddings. *CoRR*, abs/1602.01925, 2016. URL http://arxiv.org/abs/1602.01925.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of EMNLP*, 2016.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1042. URL https://aclanthology.org/P17-1042.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *AAAI*, 2018a.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, July 2018b. Association for Computational Linguistics. doi: 10.18653/v1/P18-1073. URL https://aclanthology.org/P18-1073.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels, Belgium, November 2018c. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D18-1399.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. *CoRR*, abs/1710.11041, 2018d. URL http://arxiv.org/abs/1710.11041.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. An effective approach to unsupervised machine translation. *CoRR*, abs/1902.01313, 2019a. URL http://arxiv.org/abs/1902.01313.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Bilingual lexicon induction through unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy, July 2019b. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1494.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July 2019c. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

Antonio Valerio Miceli Barone. Towards cross-lingual distributed representations without parallel text trained with adversarial autoencoders. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 121–126, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1614. URL https://aclanthology.org/W16-1614.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. volume 3, pages 932–938, 01 2000. doi: 10.1162/153244303322533223.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. Phrase-based statistical machine translation with pivot languages. In *Proceedings*

of the 5th International Workshop on Spoken Language Translation: Papers, pages 143–149, Waikiki, Hawaii, October 20-21 2008. URL https://aclanthology.org/2008.iwslt-papers.1.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 07 2016. doi: 10.1162/tacl_a_00051.

Peter F. Brown, John Cocke, Stephen A. Della-Pietra, Vincent J. Della-Pietra, Frederick Jelinek, Robert L. Mercer, and Paul Rossin. A statistical approach to language translation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 1988.

Peter F. Brown, John Cocke, Stephen A. Della-Pietra, Vincent J. Della-Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul Rossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):76–85, 1990.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy, April 2006. Association for Computational Linguistics. URL https://aclanthology.org/E06-1032.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1–7, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2001. URL https://aclanthology.org/P15-2001.

Lyle Campbell and Mauricio J. Mixco. *A Glossary of Historical Linguistics*. Edinburgh University Press, 2007. URL http://tscheer.free.fr/scan/Campbell%20&%20Mixco%2007%20-%20A%20Glossary%20of%20Historical%20Linguistics.pdf.

Chi-Yen Chen and Wei-Yun Ma. Word embedding evaluation datasets and wikipedia title embedding for chinese. 05 2018.

Yun Chen, Yang Liu, Yong Cheng, and Victor Li. A teacher-student framework for zero-resource neural machine translation. pages 1925–1935, 01 2017. doi: 10.18653/v1/P17-1176.

Yong Cheng. *Joint Training for Pivot-Based Neural Machine Translation*, pages 41–54. 08 2019. ISBN 978-981-32-9747-0. doi: 10.1007/978-981-32-9748-7_4.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1185. URL https://aclanthology.org/P16-1185.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259, 2014a. URL http://arxiv.org/abs/1409.1259.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR*, abs/1406.1078, 2014b. URL http://arxiv.org/abs/1406.1078.

Gyu-Hyeon Choi, Jong-Hun Shin, and Young-Kil Kim. Improving a multi-source neural machine translation model with corpus extension for low-resource languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1144.

Trevor Cohn and Mirella Lapata. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-1092.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. 10 2017a.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data, 2017b.

Asa Cooper Stickland, Xian Li, and Marjan Ghazvininejad. Recipes for adapting pre-trained monolingual and multilingual models to machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3440–3453, Online, April 2021.

Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.301. URL https://aclanthology.org/2021.eacl-main.301.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39 (1):1–38, 1977.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa Anke, and Steven Schockaert. On the robustness of unsupervised and semi-supervised cross-lingual word embedding learning. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4013–4023, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.495.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. Ethnologue: languages of the world - 22nd edition. internet. https://www.ethnologue.com/guides/ethnologue200, 2022. Accessed: 2022-07-14.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL https://aclanthology.org/D18-1045.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1049. URL https://aclanthology.org/E14-1049.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, 2002.

Philip Gage. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38, 1994.

Nicolas Garneau, Mathieu Godbout, David Beauchemin, Audrey Durand, and Luc Lamontagne. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings: Making the method robustly reproducible as well.

In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5546–5554, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.681.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 710–721, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1070. URL https://aclanthology.org/P19-1070.

Jiang Guo, Wanxiang Che, David Yarowsky, Haifeng Wang, and Ting Liu. Cross-lingual dependency parsing based on distributed representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1234–1244, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1119. URL https://aclanthology.org/P15-1119.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1632. URL https://aclanthology.org/D19-1632.

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

Yedid Hoshen and Lior Wolf. Non-adversarial unsupervised word translation. pages 469–478, 01 2018. doi: 10.18653/v1/D18-1043.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, 12 1936. ISSN 0006-3444. doi: 10.1093/biomet/28.3-4.321. URL https://doi.org/10.1093/biomet/28.3-4.321.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for*

Computational Linguistics, 5:339–351, 2017. doi: 10.1162/tacl_a_00065. URL https://aclanthology.org/Q17-1024.

Alistair Kennedy and Graeme Hirst. Measuring semantic relatedness across languages. In *Proceedings of xLiTe: Cross-Lingual Technologies Workshop at the Neural Information Processing Systems Conference*, volume 1130, 2012.

Jae-Hoon Kim, Hong-Seok Kwon, and Hyeongwon Seo. Evaluating a pivot-based approach for bilingual lexicon extraction. *Computational Intelligence and Neuroscience*, 2015:1–13, 04 2015. doi: 10.1155/2015/434153.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1080. URL https://aclanthology.org/D19-1080.

Yunsu Kim, Miguel Graça, and Hermann Ney. When and why is unsupervised neural machine translation useless? In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 35–44, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.5.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://aclanthology.org/P17-4012.

Kevin Knight. Automating knowledge acquisition for machine translation. *AI Magazine*, 18(4), 1997. URL http://nlp.cs.swarthmore.edu/~richardw/papers/pdf/knight1997-automating.pdf.

Kevin Knight. A statistical MT tutorial workbook. available at http://www.isi.edu/∼knight/, 1999. URL http://www.snlp.de/prescher/teaching/2007/StatisticalNLP/bib/1999jhu.knight.pdf.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017a. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL https://aclanthology.org/W17-3204.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation, 2017b.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, page 48—54, USA, 2003. Association for Computational Linguistics. doi: 10.3115/1073445.1073462. URL https://doi.org/10.3115/1073445.1073462.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Surafel M. Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. Adapting multilingual neural machine translation to unseen languages. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong, November 2-3 2019. Association for Computational Linguistics. URL https://aclanthology.org/2019.iwslt-1.16.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2018a.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, October-November 2018b. Association for Computational Linguistics. doi: 10.18653/v1/D18-1549. URL https://aclanthology.org/D18-1549.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1027. URL https://aclanthology.org/P15-1027.

Zuchao Li, Hai Zhao, Rui Wang, Masao Utiyama, and Eiichiro Sumita. Reference language based unsupervised neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4151–4162, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.371. URL https://aclanthology.org/2020.findings-emnlp.371.

Chao-Hong Liu, Catarina Silva, Longyue Wang, and Andy Way. *Pivot Machine Translation Using Chinese as Pivot Language: 14th China Workshop, CWMT 2018, Wuyishan, China, October 25-26, 2018, Proceedings*, pages 74–85. 01 2019. ISBN 978-981-13-3082-7. doi: 10.1007/978-981-13-3083-4_7.

Siyou Liu, Longyue Wang, and Chao-Hong Liu. Chinese-Portuguese machine translation: A study on building parallel corpora from comparable texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1236.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. When does unsupervised machine translation work? *CoRR*, abs/2004.05516, 2020. URL https://arxiv.org/abs/2004.05516.

Benjamin Marie and Atsushi Fujita. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *CoRR*, abs/1810.12703, 2018. URL http://arxiv.org/abs/1810.12703.

Benjamin Marie, Hour Kaing, Aye Myat Mon, Chenchen Ding, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. Supervised and unsupervised machine translation for Myanmar-English and Khmer-English. In *Proceedings of the 6th Workshop on Asian Translation*, pages 68–75, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-5206. URL https://aclanthology.org/D19-5206.

Benjamin Marie, Haipeng Sun, Rui Wang, Kehai Chen, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. NICT's unsupervised neural and statistical machine translation systems for the WMT19 news translation task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 294–301, Florence, Italy, August 2019b. Association for Computational Linguistics. doi: 10.18653/v1/W19-5330. URL https://aclanthology.org/W19-5330.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1103. URL https://aclanthology.org/D18-1103.

Jerry Norman. *The Chinese dialects: phonology.* Routledge, 2002. ISBN 9780203221051.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075117. URL https://aclanthology.org/P03-1021.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002a.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002b. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Sungjoon Park, Jeongmin Byun, Sion Baek, Yongseok Cho, and Alice Oh. Subword-level word vector representations for Korean. In *Proceedings of the 56th Annual Meeting of the ACL*, pages 2429–2438, 2018.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. On the importance of pivot language selection for statistical machine translation. pages 221–224, 01 2009. doi: 10.3115/1620853.1620914.

Wei Peng, Chongxuan Huang, Tianhao Li, Yun Chen, and Qun Liu. Dictionary-based data augmentation for cross-domain neural machine translation. *ArXiv*, abs/2004.02577, 2020.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1039. URL https://aclanthology.org/D18-1039.

Rebecca Posner. *The Romance languages*. Cambridge University Press, 1996.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.170. URL https://aclanthology.org/2020.acl-main.170.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2084. URL https://aclanthology.org/N18-2084.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. Unsupervised neural machine translation with SMT as posterior regularization. *CoRR*, abs/1901.04112, 2019. URL http://arxiv.org/abs/1901.04112.

Rodrigo Santos, João Silva, António Branco, and Deyi Xiong. *The Direct Path May Not Be The Best: Portuguese-Chinese Neural Machine Translation*, pages 757–768. 08 2019. ISBN 978-3-030-30243-6. doi: 10.1007/978-3-030-30244-3_62.

Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. pages 5149–5152, 03 2012. ISBN 978-1-4673-0045-2. doi: 10.1109/ICASSP.2012.6289079.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. Multilingual unsupervised NMT using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1297. URL https://aclanthology.org/P19-1297.

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 (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

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 (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany, August 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

Kashif Shah. *Model adaptation techniques in machine translation*. PhD thesis, 06 2012.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1072. URL https://aclanthology.org/P18-1072.

J.J. Song. *The Korean Language: Structure, Use and Context*. Routledge, 2005. ISBN 9780415328029. URL https://books.google.pt/books?id=rIk52cJ1vDEC.

Felix Stahlberg, James Cross, and Veselin Stoyanov. Simple fusion: Return of the language model. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 204–211, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6321. URL https://aclanthology.org/W18-6321.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. *CoRR*, abs/1409.3215, 2014. URL http://arxiv.org/abs/1409.3215.

Wilson L. Taylor. "cloze procedure" : A new tool for measuring readability. *Journalism & Mass Communication Quarterly*, 30:415 – 433, 1953.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Joseph Turian, Lev-Arie Ratinov, and Yoshua Bengio. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the*

48th Annual Meeting of the Association for Computational Linguistics, pages 384–394, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/P10-1040.

Masao Utiyama and Hitoshi Isahara. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York, April 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-1061.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(110):3371–3408, 2010. URL http://jmlr.org/papers/v11/vincent10a.html.

Ivan Vulić and Anna Korhonen. On the role of seed lexicons in learning bilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 247–257, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1024. URL https://aclanthology.org/P16-1024.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. Do we really need fully unsupervised cross-lingual embeddings?, 2019.

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

Hua Wu and Haifeng Wang. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/P07-1108.

Min Xiao and Yuhong Guo. Distributed word representation learning for cross-lingual dependency parsing. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 119–129, Ann Arbor, Michigan, June

2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1613. URL https://aclanthology.org/W14-1613.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado, 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1104. URL https://aclanthology.org/N15-1104.

Samira Zahabi, Somayeh Bakhshaei, and Shahram Khadivi. Using context vectors in improving a machine translation system with bridge language. volume 2, pages 318–322, 08 2013.

Mozhi Zhang, Keyulu Xu, Ken-ichi Kawarabayashi, Stefanie Jegelka, and Jordan Boyd-Graber. Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3180–3189, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1307. URL https://aclanthology.org/P19-1307.