UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Portuguese-Chinese Neural Machine Translation

Rodrigo Soares dos Santos

MESTRADO EM ENGENHARIA INFORMÁTICA
Especialização em Interação e Conhecimento

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

2019

Agradecimentos

O trabalho apresentado nesta dissertação foi possível graças à ajuda de várias pessoas e instituições, às quais gostaria de agradecer.

Aos meus orientadores, Professor Doutor António Branco e ao Doutor João Silva, agradeço a sua orientação e total disponibilidade.

Agradeço ao grupo NLX, que me forneceu um espaço de trabalho e ferramentas para a execução deste estudo, assim como a todos os membros do grupo que sempre estiveram disponíveis para me ajudar.

À FCT (Fundação para a Ciência e a Tecnologia), agradeço pelo financiamento no âmbito do projeto ASSET (ANI/3279/2016), e ao INCD (Infraestrutura Nacional de Computação Distribuída), pelo poder computacional necessário para o treino dos vários modelos.

Finalmente, agradeço ao meus pais por todo o apoio e paciência em todos estes anos de escolaridade.

A special thanks to Professor Deyi Xiong for his initial guidance, and to Yingying Peng for her help in the human evaluation process.

Lisboa, Julho de 2019
Rodrigo Soares dos Santos

Resumo

Esta dissertação apresenta um estudo sobre Tradução Automática Neuronal (*Neural Machine Translation*) para o par de línguas Português (PT) ↔ Chinês (ZH) culminando na criação de um sistema de tradução automática com desempenho ao nível do estado da arte, que tira partido apenas de recursos e ferramentas livremente disponíveis.

Este par de línguas foi escolhido devido ao seu impacto a nível global. O Português é a sexta língua mais falada no mundo, com presença em todos os continentes (sendo em particular a língua mais falada no hemisfério sul) e a língua Chinesa, que tem como país de origem a China, é a língua mais falada em todo o mundo.

Como super potência emergente, a China tem cada vez mais ligações aos países ocidentais e, como tal, a necessidade de instrumentos de comunicação adequados que possam atravessar as barreiras linguísticas é cada vez mais premente. A tradução automática surge assim como um apoio para o acesso rápido a grandes quantidades de informação.

Portugal e a língua portuguesa têm várias ligações à China. Uma destas ligações é Macau, uma região administrativa especial da República Popular da China onde o Português e o Chinês são ambas línguas oficiais e, assim sendo, onde o interesse num sistema que traduza entre as duas é muito grande. Porém, o problema da Tradução Automática entre estas duas línguas ainda não tem sido alvo de suficiente atenção pela comunidade científica.

Neste trabalho ambas as direções de tradução são consideradas, isto é, são criados sistemas de tradução para a direção de tradução Português → Chinês e para a direção Chinês → Português. A dificuldade na criação de tais sistemas passa pela aquisição de corpora de qualidade e em quantidade suficiente nas duas línguas, o que para o par de línguas escolhido é um grande desafio; e passa também pela escolha da arquitetura que melhor se adapta a esse corpora.

Para a criação destes sistemas de tradução, exploro três abordagens, que são referidas neste documento como: (i) abordagem direta (*direct approach*), que faz uso apenas de corpora paralelo entre Português e Chinês; (ii) abordagem pivô (*pivot approach*), que usa uma terceira língua como intermediário para a tradução; e (iii) abordagem muitos-para-muitos (*many-to-many approach*), que tira partido de toda a informação usada nas outras duas abordagens.

As várias abordagens são implementadas com recurso a redes neuronais, mais propriamente à arquitetura Transformer (Vaswani et al., 2017), e obtêm desempenho assinalável, com uma das abordagens a alcançar resultados superiores aos do Google Tradutor para o par de línguas escolhido em ambas as direções.

Para efeitos de teste e comparação entre as várias abordagens e as traduções do Google Tradutor, o mesmo corpus de teste é usado para avaliar todos os sistemas. Esse corpus de teste é constituído pelas primeiras 1000 frases do News Commentary v11 corpus (Tiedemann, 2012), sendo composto por textos jornalísticos bem curados e com grande qualidade gramatical.

A abordagem direta é a solução mais comum usada para a criação de um sistema de tradução automática. No caso deste estudo, um corpus paralelo entre Português e Chinês é usado para a criação de dois modelos, um para cada direção de tradução, isto é um para PT → ZH e outro para ZH → PT.

Apesar das dificuldades em encontrar corpora paralelo entre Português e Chinês, foi possível encontrar um corpus com cerca de 1 milhão de frases, o qual é usado para o treino desta abordagem. O artigo que apresenta este corpus (Chao et al., 2018) foi publicado poucos meses antes do início desta dissertação e tanto quanto sei não existem outros trabalhos que usem este corpus além de (Chao et al., 2018).

Usando a métrica BLEU (Papineni et al., 2002), a abordagem direta consegue um melhor desempenho que a base dada pelo Google Tradutor para a direção ZH → PT, não conseguindo, contudo, ultrapassar esta base para a direção de tradução PT → ZH.

A falta de qualidade e quantidade de corpora paralelos entre Português e Chinês motiva a experimentação com uma abordagem pivô. Numa abordagem pivô, o sistema faz uso de uma língua intermediária escolhida de forma a que haja grande quantidade e qualidade de corpora paralelos entre esta e as outras duas línguas. O sistema começa por traduzir de Português ou Chinês para a língua pivô e de seguida traduz da língua pivô para Chinês ou Português. A ideia por detrás desta abordagem é que as redes neuronais tendem a ter melhor performance quanto maior for o número de exemplos usados para treino da rede, e que esta melhoria será capaz de compensar a degradação da tradução introduzida pela passagem por uma língua intermédia.

Usando a métrica BLEU, esta abordagem obtém resultados superiores à base e à abordagem direta em ambas as direções de tradução.

Finalmente, a abordagem muitos-para-muitos segue as propostas de Johnson et al. (2017), Lakew et al. (2017) e Aharoni et al. (2019), que permitem o uso dos vários corpora paralelos usados para treino das outras duas abordagens.

Usando a métrica BLEU, os resultados deste sistema ficam entre os da abordagem direta e os da abordagem pivô, não conseguindo ultrapassar a base para a direção de tradução PT → ZH.

De entre os vários sistemas criados, a abordagem com melhores resultados é a abor-

dagem pivô, que por sua vez foi a única abordagem que não viu qualquer tipo de dados paralelos entre as línguas Portuguesa e Chinesa. Porém, a abordagem muitos-para-muitos é a que demonstra maior potencial de desenvolvimento pois tem a capacidade de facilmente incorporar mais dados e assim melhorar a qualidade de tradução.

O trabalho final, para além de uma panorâmica sobre o estado da arte da tradução automática, fornece uma solução prática com boa qualidade para a tradução entre Português e Chinês usando apenas recursos e ferramentas livremente disponíveis.

Foi também criado um serviço online de tradução entre Português e Chinês disponível gratuitamente em https://portulanclarin.net/workbench/lx/translator/, resultante do trabalho descrito neste documento.

Cabe notar que parte do trabalho apresentado nesta dissertação já foi alvo de revisão por pares (*peer review*) e aceite para publicação (Santos et al., to appear).

Palavras-chave: Processamento de Linguagem Natural, Tradução Automática, Redes Neuronais Artificiais, Tradução Automática Neuronal, Português, Chinês.

Abstract

This dissertation reports on a study addressing Neural Machine Translation for the language pair Portuguese ↔ Chinese and also on the development of a state of the art Machine Translation system for this pair using only freely available resources.

The choice of this particular language pair was due to the fact that China is regarded as an emerging super power whose ties are steadily increasing with western countries, and as such the need for appropriate communication tools that can cross linguistic barriers is becoming a more pressing issue. The use of Machine Translation supports fast access to big quantities of data in another language.

Portugal and its language have several ties with China. With Macau being a special administrative region of the People's Republic of China where the two languages are official languages, a Machine Translation system for this pair is of high importance.

In this work, both translation directions are considered. That is, there are systems for the translation direction Chinese → Portuguese, and systems for the direction Portuguese → Chinese. The key issue underlying the creation of such systems is twofold: (i) the gathering of corpora with good enough quality and quantity, which for this pair is a challenge; and (ii) the choice of a suitable architecture to accommodate such corpora.

Three approaches are followed to address the problem, with all the implemented systems making use of neural networks, namely the Transformer architecture, and with the performance of one approach surpassing that of the baseline Google Translate for the chosen language pairs in both translation directions.

An online translation service was also developed, showcasing one of the three approaches studied in this document for the two translation directions, and is freely available at https://portulanclarin.net/workbench/lx/translator/.

Note that part of the work presented in this dissertation already passed peer review, and was accepted for publication (Santos et al., to appear).

Keywords: Natural Language Processing, Machine Translation, Artificial Neural Networks, Neural Machine Translation, Portuguese, Chinese.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the last few years, Artificial Intelligence (AI) has been an area of major interest and advancements are published everyday on several topics. As a subfield of AI, Natural Language Processing (NLP) is no exception and Machine Translation (MT) has seen new architectures systematically improve the state of the art.

This chapter starts by addressing the motivation behind this dissertation. This is followed by a brief description of the two languages that the focus of this work and a presentation of the research context of the dissertation alongside with my contributions to the relevant project. Finally, an overview and structure of the present document is provided.

1.1 Motivation

Language is the prime vehicle for human communication and, since the early days of AI, it has been studied in the subdomain of NLP in order to try to understand and derive meaning from it in an useful way that can be handled by a machine.

Being a practical application of NLP, MT seeks to foster the ability of a machine of diluting the barriers imposed by the various communication systems used for human communication, which in an increasingly globalized world are obstacles for mutual understanding.

Literature about MT often revolves around English. Yet every language has particular problems that have to be faced. For instance, languages like Finnish and Czech are morphologically very rich and it can be hard to translate into them; while languages, such as Japanese or Chinese, which do not have word boundaries, are hard to represent in MT systems that expect words to be explicitly separated. While under-represented in the NLP literature, many of these languages are spoken in some of the most powerful countries in the world.

China is regarded as an emerging super power due to its large population and increasing political and economic influence. Industries and companies all over the world are highly dependent on what happens in China. Therefore, it is of high importance to be aware of what is happening there. However, there is a huge language barrier between China and the rest of the world, primarily because of its writing system based on unique logo-grams.

Like China and Chinese, Portugal and its language have strong positions world wide with Portuguese being one of the official languages of several international organizations, including Mercosur, the Organization of Ibero-American States, the Union of South American Nations, the Organization of American States, the African Union, the Economic Community of West African States, the Southern African Development Community and the European Union.

Other than in Portugal, Portuguese is spoken all around the globe, within growing economies like Brazil and Angola, and in big emigrant communities in many other countries.

Portuguese speaking countries have growing ties with China, and the connection between them demands a way of efficient communication. The use of educated personnel that has knowledge of both languages is one way of solving the problem, yet the high demand for such people makes the solution expensive and inefficient, due to the high effort required to educate translators and the slow pace of the human translation process. Machine Translation appears as a useful solution to this problem, providing great translation speed and affordable costs. Its downside is the low translation quality, which is a direct result of the quantity and nature of the available resources and of the choice of architecture used to work upon such resources. Normally, the output of these systems still requires revision by bilingual speakers when used for high quality document production.

The motivation of the present work is thus twofold. On the one hand, to undertake research on PT $\leftrightarrow$ ZH Machine Translation, contributing to the literature for this understudied language pair and advancing the state of the art; on the other hand, to study viable approaches for the creation of MT solutions for a pair of languages with few freely available resources, despite being languages with large number of speakers.

1.2 The Portuguese Language

The Portuguese language is the sixth language with the largest number of native speakers on the planet (Simons, 2019), and the most spoken language in the southern hemisphere with around 280 million speakers (4% of the world population). Figure 1.1 shows the distribution of Portuguese speaking people around the world.

Portuguese is the official language of 9 countries (Angola, Brazil, Cape Verde, East Timor, Equatorial Guinea, Guinea-Bissau, Mozambique, Portugal, and São Tomé and Príncipe). While population in Portugal is decreasing, populations of these other countries are growing and it is estimated that by 2050 around 400 million people will speak Portuguese (Reto et al., 2016).

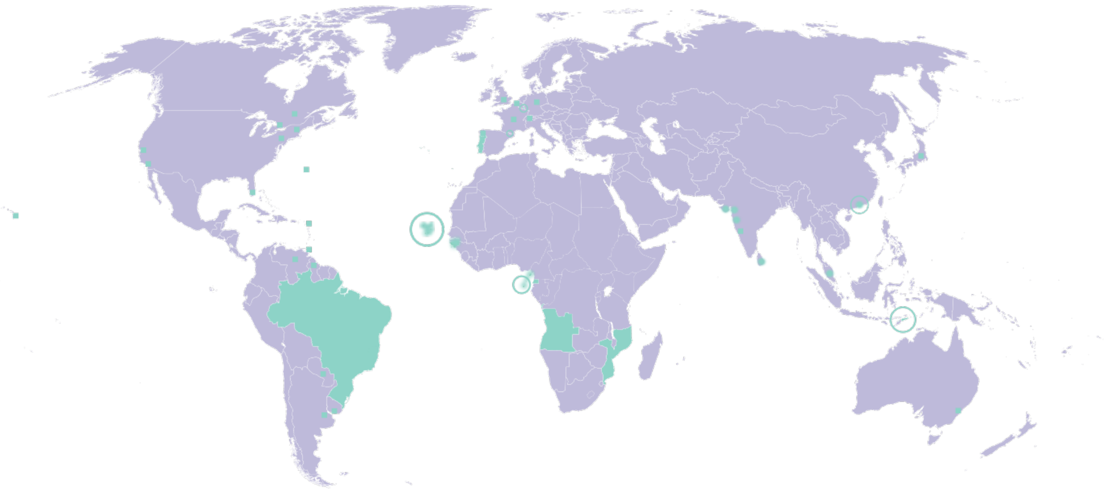


Figure 1.1: Geographic distribution of Portuguese speaking people world wide - Adapted from: Wikimedia Commons (2018)

The Portuguese language has Latin as its base because of the Roman occupation of the Iberian Peninsula. It has also acquired vocabulary from all over the world as the result of the Portuguese expansion during the Age of Discoveries from the XV century to the XVII century.

Grammatically, the canonical order in Portuguese sentences is SVO (subject-verb-object). Nouns, adjectives, pronouns, and articles are moderately inflected (gender and number), while verbs are highly inflected, with 11 conjugational paradigms.

The Portuguese writing system is based on the Latin script (with 26 letters), and has well marked word separation through the use of blank spaces.

1.3 The Chinese Language

The Chinese language is a group of dialects whose speakers make up around 19% of the world population (Simons, 2019). The written system is common to all dialects, making it a vehicle for mutual understanding between literate people. The pronunciation, however, varies between dialects to such an extent that it can lead to a lack of mutual understanding (Norman, 2003). Figure 1.2 shows the geographic distribution of several Chinese dialects in China.

The written system is composed of logo-grams that represent morphemes. These logo-grams can be based on representations of physical objects, abstract notions or pronunciation.

A college student knows between 4,000 to 5,000 logo-grams (DeFrancis et al., 1969), making the learning curve of the written system a real challenge. In order to ease learning of the Chinese written system and improve literacy across China, in 1950 efforts were

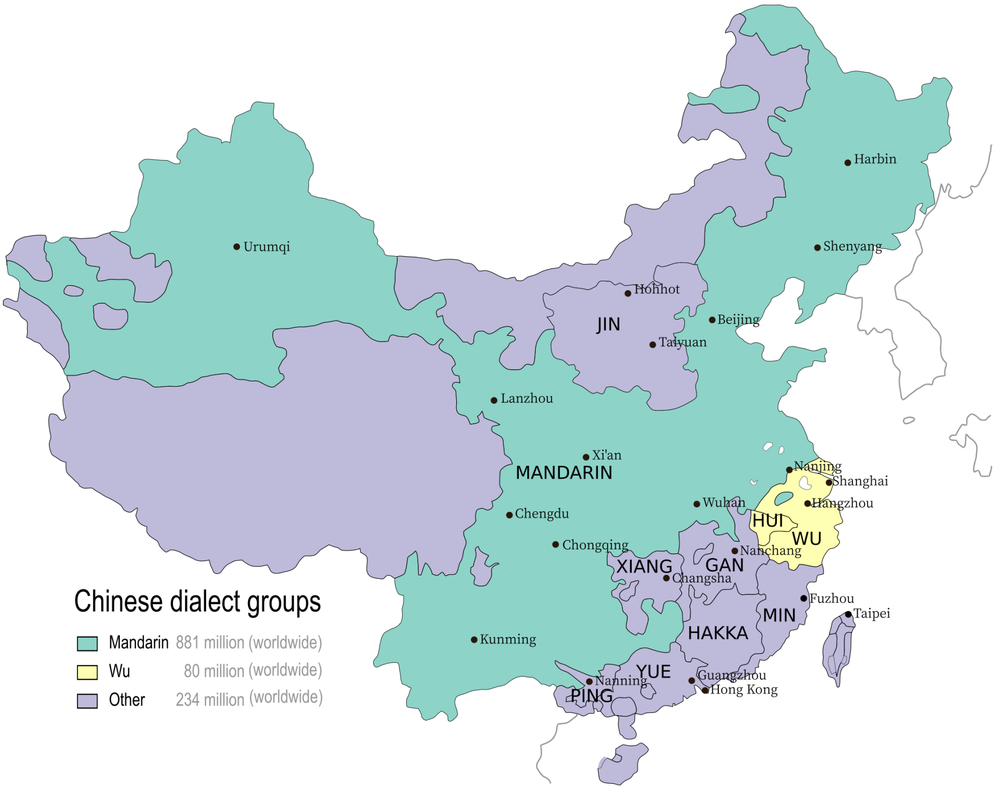


Figure 1.2: Geographic distribution of the Chinese dialects - Adapted from: Wikimedia Commons (2019)

made to simplify this system. Two separate initiatives were conducted, one attempting to simplify the existing characters and another attempting to adopt the Latin script.

While the simplification of existing characters was fairly straightforward, the implementation of the Latin script faced a big problem, namely the differences in pronunciation between the several Chinese dialects. As the Latin script is based on phonetics, the creation of a writing system based on the Latin script demanded the adoption of one of the Chinese dialects. The Mandarin dialect was chosen to be the base of this new Latin script writing system, the Pinyin, mainly due to it being the dialect with the highest number of speakers.

Pinyin was never truly adopted, probably because of this requirement to adapt the script to one form of Chinese. Instead, the simplified version of the written system was officialized (DeFrancis et al., 1969). Figure 1.3 shows a comparison between these writing systems. Throughout this dissertation only simplified Chinese is used.

Like Portuguese, the canonical order in Chinese sentences is SVO (subject-verb-object), although it has almost no word inflection.

Traditional	你叫什麼名字?
Simplified	你叫什么名字?
Pinyin	Nǐ jiào shénme míngzi?

Figure 1.3: Traditional vs Simplified vs Pinyin Chinese writing for the sentence “What is your name?”

1.4 Research Context and Contributions

The work whose results are presented in this document was undertaken during my stay at NLX—Natural Language and Speech Group,[1] a research group for Natural Language Processing from the Faculty of Sciences of the University of Lisbon.

NLX has several project on MT. This dissertation was performed within the scope of two projects: the ASSET (Intelligent Assistance for Everyone Everywhere) project, which aims to improve automatic assistance quality on various languages for the Information Technology domain; and the CNPTDeepMT-Chinese (Portuguese Deep Machine Translation in eCommerce Domain), which focuses on the improvement of automatic translation between the Portuguese and the Chinese languages.

With the help of the group I was able to carry out all the work presented in this document, excluding some frameworks, tools and corpora that I resorted to, which are properly credited to their authors when mentioned. I was given total autonomy to perform this study.

The major contributions described in this dissertation are: an overview of the current state of the art for MT, and of the tools and data available for the PT ↔ ZH language pair; an exploratory research on MT for PT ↔ ZH; and a translation system for PT ↔ ZH with state of the art performance.

The NLX group also made possible for a translation service resulting from the work described in this document to be freely available online.[2]

Part of the work presented in this dissertation already passed peer review, and was accepted for publication (Santos et al., to appear).

1.5 Overview and Document Structure

This document presents various MT systems for PT ↔ ZH where both translation directions are considered.

Three approaches for translation between these two languages are presented: (i) the direct approach, which only makes use of parallel corpora between Portuguese and Chinese; (ii) the pivot approach, which uses a third language as broker for the translation;

[1] http://nlxgroup.di.fc.ul.pt/

[2] https://portulanclarin.net/workbench/lx/translator/

and (iii) the many-to-many approach, which takes advantage from all the data given to the previous two approaches. Pros and cons of every approach are presented along with implementation details.

The remainder of this document is structured as follows. Chapter 2 refers to the objectives and planing of the dissertation.

Chapter 3 provides a description of the related work done on Machine Translation with a focus on neural approaches and on work for the PT $\leftrightarrow$ ZH pair.

Chapter 4 describes the work performed, and the frameworks and tools used, as well as provides the evaluation results and their discussion.

Chapter 5 gives final remarks and pointers for future work. Finally, the Appendices have various additional information.

Chapter 2

Objectives and Planning

This chapter addresses the objectives of my study, followed by the planning for the development work leading to the dissertation.

The methodology followed and the comparison between the planed and the actual work are also detailed below.

2.1 Objectives

The major objective of this work is to address the challenge of determining how far one is presently able to go when developing Neural Machine Translation (NMT) solutions for both directions of the Portuguese ↔ Chinese (PT ↔ ZH) language pair making use only of freely available resources, and ultimately to develop a state of the art NMT system that is able to translate from Portuguese to Chinese, and from Chinese to Portuguese.

Firstly, it is of high importance to get acquainted with the field and to be up to date with its state of the art. Therefore, the study of Natural Language Processing and Neural Machine Translation in particular was one of the goals of this dissertation.

For the implementation of a NMT system, there is a need to gather parallel corpora. That is, data comprised of sentences from one language and their respective translations in another language. Said corpora must be of the languages involved and it should be of good quality, i.e. the translation between the two languages has to be correct. It ought also to be as large and diverse as possible, in order to allow the system to observe and learn from as many phenomena of the languages involved as possible.

Together with the collection of a data set, the appropriate NMT architecture to accommodate information for both languages and create a mapping between them has to be chosen. For this we need to have in mind the corpora that we have available as well as the languages involved. Tuning of the hyper-parameters of the model may be required to better make use of the available data.

Finally, the pre-processing and post-processing of data is vital to improve the system performance, so choosing the best tools that fit our problem is important. These could range from relatively simple tokenizers, for languages that use the white space for word separation (like Portuguese), to more complex sentence segmenters, in the case of languages without word boundaries (like Chinese).

Taking this into account, the following potential objectives were set for my work leading to this dissertation:

- Acquire knowledge about Natural Language Processing, in general, and Neural Machine Translation, in particular;
- Learn about Machine Learning techniques for Neural Machine Translation, namely Deep Learning and Neural Networks;
- Familiarize myself with the state of the art of NMT;
- Collect information on NMT frameworks available for the development of translation systems;
- Collect parallel corpora for the study of NMT for the pair PT $\leftrightarrow$ ZH;
- Collect information on the tools available for the pre-processing and post-processing of the acquired corpora;
- Study the various approaches to train an NMT system for PT $\leftrightarrow$ ZH;
- Develop a translation system for PT $\leftrightarrow$ ZH.

2.2 Planning

To achieve the objectives mentioned above, a set of guidelines was set previous to starting the work leading to this dissertation. The plan initially proposed was:

A) Acquire knowledge about the foundations of Natural Language Processing and the state of the art on Neural Machine Translation.

- 2 Months

B) Collect information on NMT frameworks and experiment with MT systems, more specifically with Neural Machine Translation systems that are currently the state of the art.

- 4 Months (Overlapping with A)

C) Collection of data (corpora) and tools for the development of a PT $\leftrightarrow$ ZH translation system.

- 1 Month (Overlapping with B)

D) Development of a Neural Machine Translation system for both directions of the language pair Portuguese and Chinese.

- 4 Months

E) Evaluation of the PT ↔ ZH Neural Machine Translation system created.

- 1 Month (Overlapping with D)

F) Finally, writing of the dissertation.

- 3 Months - (Overlapping with E)

2.3 Plan Execution

The execution of the plan above is detailed below together with the steps taken to accomplish every item of the planning.

In order to keep up with the state of the art (item A), which is rapidly evolving in the field, I resorted to information sources such as the proceedings of conferences like the Annual Meetings of the Association for Computational Linguistics (ACL)[1] and the Conference on Machine Translation (WMT),[2] paper repositories like *arXiv*[3] and paper aggregators like *Google Scholar*.[4] This item of the planning was followed throughout all the work leading to this dissertation and not only at the beginning, given that the field is evolving rapidly and what was the state of the art at the beginning has been in many cases surpassed or improved upon during the duration of my work.

Several experiments (item B) were conducted before moving on to further tasks. I began by changing and adjusting some existing systems in order to study the impact that my changes had on those systems. These systems were normally small, simple and only toy data was used, and the frameworks they were implemented with were diverse. This way I got a broad understanding of what frameworks were better to use.

Some of the best corpora available were not on the desired pair so, to move forward with the initial experimentation, a first system was produced on the Spanish ↔ English language pair, for which corpora is more abundant than for the targeted languages of this study, as well as it being a pair easily understandable by me. Results were promising and confidence was built to start tackling the target language pair.

Data acquisition (item C) was one of the hardest tasks, considering that little research is done with the pair PT ↔ ZH. Nonetheless, due to the approach taken to the first point of the planning, eventually a suitable corpus for PT ↔ ZH translation was found in the literature.

[1] https://www.aclweb.org/
[2] http://statmt.org/
[3] https://arxiv.org/
[4] https://scholar.google.com/

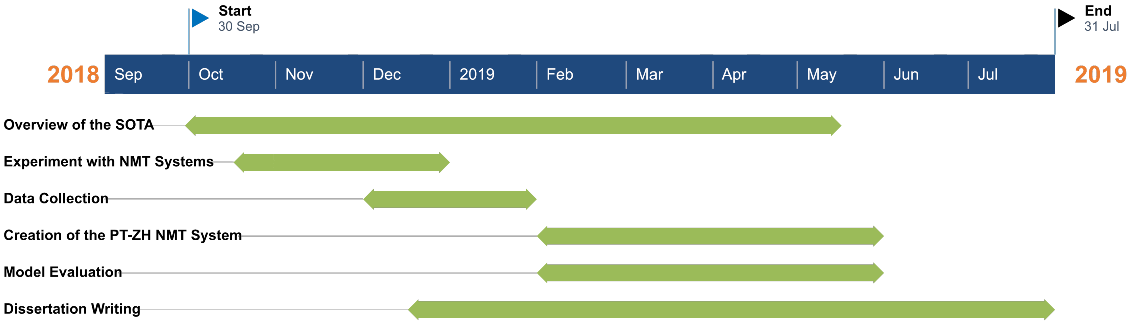


Figure 2.1: Development timeline

Regarding the creation of the Neural Machine Translation system (item D), the same explanation can be given. That is, constant following of the state of the art was necessary to eventually settle upon the best tools and NMT architectures to develop such systems.

Evaluation (item E) is a natural step on the creation of a Neural Network system as no system is finished without knowing its capabilities or if it even fulfills its objective. Accordingly, evaluation was run immediately as soon as training of interim systems had finished.

The writing of the dissertation (item F) was something done all along the way, with the biggest portion of it being done on the final months of this work.

Figure 2.1 provides an overview of the time taken to accomplish the various entries of the planning. Although the time taken for some tasks changed in relation to the initial plan, the plan was completed without any omissions.

Chapter 3

Related Work

Machine Translation has gone through significant advancements in the last few years with Neural Machine Translation (NMT) surpassing previous Phrase Based Statistical architectures. Though Neural Machine Learning is the now sought after approach, Machine Translation has a decades long history of research.

This chapter will mention some milestones of Machine Translation history, the current state of the art and, finally, some of the work done for Portuguese ↔ Chinese (PT ↔ ZH) Machine Translation specifically.

3.1 Non-Neural Machine Translation

The first mention of MT in the literature is from the XVII century with the idea of mechanical dictionaries, yet the first concrete proposal of such systems only came in 1933, by the hand of George Artsrouni and Petr Smirnov-Troyanskii (Hutchins, 1995). The former designed a storage device on a paper tape that could be used to store a word and its equivalent in another language. The latter had a three stage translation system where an editor knowing only the source language did a logic analysis and annotated the words into their base form and syntactic functions, a machine transformed the source language annotations into target language annotations, and finally another editor knowing the target language finished the translation by generating the sentence in the target language from the annotations.

With advancements in code breaking during World War II and the demand for translation systems during the Cold War, quite a lot of research was done in the field, with the English ↔ Russian language pair being the focus. These systems used rules to fix the word order after direct, word-to-word translation with the help of dictionaries.

In 1966 the ALPAC report written by a committee of seven US scientists deemed Machine Translation as more expensive, slower and less accurate than human translation, determining most of funding to research to be closed. Improvements in the field slowed down but they were not abandoned, so in the 1980s there were various systems on Machine Translation focused on inter-lingua and rule-based approaches with statistical architectures beginning to show promising results in the late 1980s.

Statistical Machine Translation (SMT) makes use of the probability distribution that a string in the target language is the translation of a string in the source language. IBM (International Business Machines Corporation) was one of the pioneers of SMT with word based SMT. Their first statistical model (Model 1) worked by splitting the sentence into words and attributing word translation probabilities by the frequency observed in a parallel corpus, that is a collection of texts where a sentence in a language is aligned with its corresponding translation in another language.

However, the most used SMT platform was eventually the Moses[1] system that makes use of phrase based SMT (PBSMT) (Koehn et al., 2003). PBSMT splits the sentences not only into words but also into phrases (or, more precisely, into n-grams, which are contiguous sequences of n words). SMT, mostly in the form of PBSMT, was the mainstream MT approach for more than 20 years.

Nowadays, Neural Machine Translation is regarded as the mainstream approach and there are several NMT architectures, each with its strengths and weaknesses.

3.2 Neural Machine Translation

With the advent of the age of Big Data and the evolution of computer hardware in terms of processing speed, Neural Networks have risen in popularity. The ability to learn almost any data pattern makes these networks the go-to architecture as long as there are lots of training data (even if quality does not necessarily have to be perfect), and computational power to put it all together.

In this section I give an overview of the main NMT topics that are relevant to this work, starting with the introduction of the encoder-decoder architecture and the *attention* mechanism present in most recent models, followed by an introduction of the Transformer model eventually used in this work.

3.2.1 Sequence-to-Sequence Encoder-Decoder

Sutskever et al. (2014) was the first to use Deep Neural Networks to map sequences to sequences (Seq2Seq), which is the core of a translation system where a source sequence is mapped to a target sequence.

The idea is that these systems learn to map a source sentence to a target sentence directly, in an end-to-end fashion, given enough training on a large parallel corpus.

In order to obtain a representation from a sentence, we need a method that can process sequences of words of variable size. An idea that was successful with speech recognition

[1] http://www.statmt.org/moses/

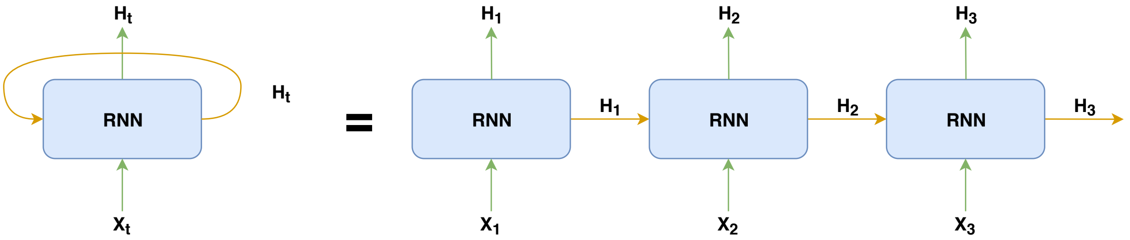


Figure 3.1: Unrolling of an RNN

was the use of Recurrent Neural Networks (RNN). When applied to MT these networks have recurrent units that process one word at each time step in a recurrent way, keeping an internal state between time steps.

Figure 3.1 shows one recurrent unit (left) and how it looks after it is unrolled in time (right), with its internal state (hidden state H_t) being kept between time steps.

There are various types of recurrent units, with the most used being the Long Short Term Memory (LSTM) unit (Hochreiter and Schmidhuber, 1997), and the Gated Recurrent Unit (GRU) (Cho et al., 2014).

By using LSTM units, Sutskever et al. (2014) devised a Seq2Seq method to encode the input sequence, regardless of its length, into a vector of fixed dimensionality, and then decode the target sequence from that vector, hence the name encoder-decoder.

In the Seq2Seq architecture, shown in Figure 3.2, the first RNN unit encodes the input sequence one word at a time in a recursive way. When the encoder gets to the end of the sequence, which is marked by a well defined token (e.g. <EOS>, for End Of Sentence), it stops and gives its last hidden state (H_n) to the decoder.

On the decoder side, the hidden state of its RNN unit is initialized with the last encoder state and its input is a start sequence token (e.g. <BOS>, for Begin Of Sentence). The resulting RNN unit vector state is passed to a linear layer that resizes the vector to the same size of the target vocabulary, to be used as input to a softmax layer that creates a probability distribution over the vocabulary. The word with the highest probability in the distribution is the word that is predicted by the decoder.

The predicted word is fed as input to the decoder on the next time step, and this process is repeated until the <EOS> token is predicted or a hard cut of length is reached, marking the end of the target sentence.

During training, there is an additional step where the output probability distribution vector given by the softmax is compared with the one-hot-encoded vector that represents the target word,[2] and the error (loss) between the objective vector and the predicted vector is backpropagated (Rumelhart et al., 1986) through the network adjusting its weights.

Most NMT architectures today follow these ideas of the Seq2Seq encoder-decoder,

[2] In the one-hot-encoded vector of a word, all positions have a value of zero, except for the position corresponding to that word, which has a value of one.

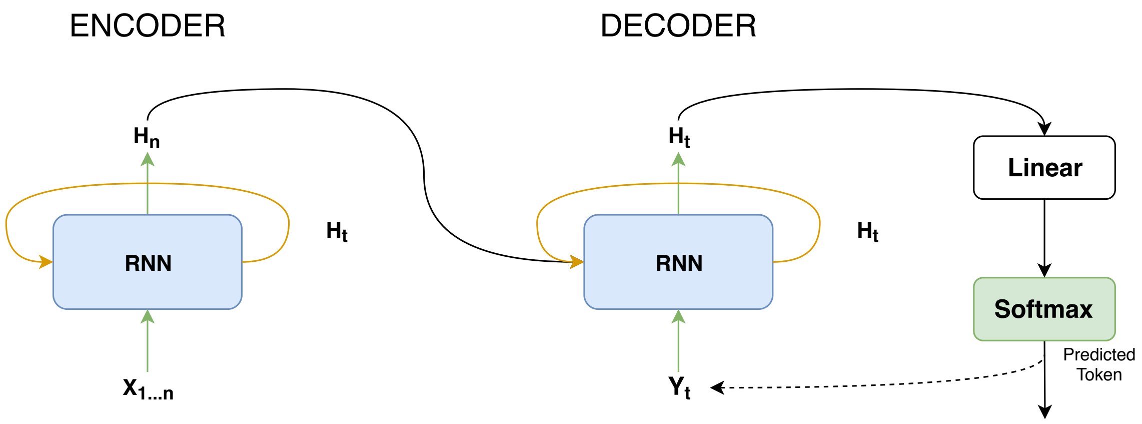


Figure 3.2: Seq2Seq encoder-decoder architecture

like Google's Neural Machine Translation (Wu et al., 2016) and the convolutional model ConvS2S of Gehring et al. (2017b).

Google's system achieved one of the best results on RNN Machine Translation. At the time, the system achieved the best BLEU scores (cf. Section 3.3) for English → French and English → German, with 38.95 BLEU and 24.67 BLEU respectively (Wu et al., 2016), on the newstest2014 test set which is a standard evaluation corpus introduced in the WMT 2014 workshop.

ConvS2S (Gehring et al., 2017b) spotlighted a different paradigm to NMT, bringing ideas from the image recognition field, namely the use of Convolutional Neural Networks (CNN) Krizhevsky et al. (2012). Inspired by the visual cortex, these networks apply various convolutions over the matrices that compose an image. This way, they make use of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns. In NMT these networks work in similar a way, yet instead of using pictures as input, the concatenation of the numerical representation of the words (see below for an explanation of embeddings) is used, forming a matrix with the words of the sentence. ConvS2S is the state of the art system on Seq2Seq NMT with Convolutional Neural Networks (CNN), where the authors improved upon their previous work (Gehring et al., 2017a). These improvements saw them surpass Google's model with 40.51 for English → French and 25.16 for English → German.

One exception to the Seq2Seq architecture is the CNN network of Elbayad et al. (2018) where they use a CNN that both encodes and decodes the sequences. This architecture allows them to reduce the number of parameters needed to train an NMT system, yet they could not surpass the performance of the state of the art Seq2Seq encoder-decoder architectures.

Embeddings. Note that the input of any neural model has to be represented numerically, typically as a large vector of real numbers. As such, the first layer of NMT models is what

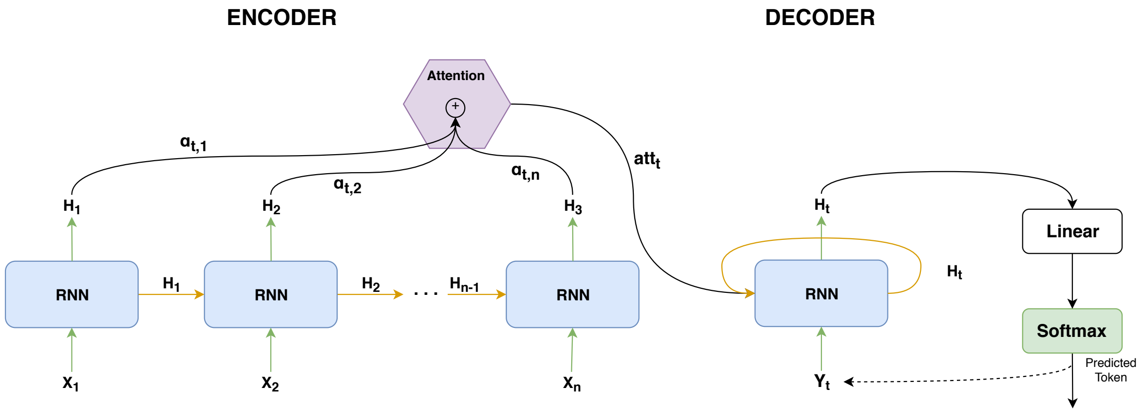


Figure 3.3: Seq2Seq encoder-decoder architecture with attention

is called an embedding layer, which takes the symbols in the input, such as words, and learns a mapping for each symbol into a vectorial space.

3.2.2 Attention

In the Seq2Seq encoder-decoder architecture described above, the encoder has to pack the representation of the whole input sequence into a single vector that is passed on to the decoder, which places a great burden on the model. The mechanism of attention, introduced in the seminal paper of Bahdanau et al. (2015), eases this burden by, instead of passing a single vector from the encoder to the decoder, allowing the decoder to access *all encoder states*, each contributing in a different amount for the final vector representation of the input.

Figure 3.3 illustrates the incorporation of the attention mechanism in the Seq2Seq encoder-decoder architecture.

The vector representation of the input sequence at time step t, which I refer to as the context vector att_t, is a weighted sum of encoder states, calculated as shown in Equation 3.1:

$$att_t = \sum_{j=1}^{T_n} \alpha_{tj} h_j \tag{3.1}$$

where h_j is an encoder state and α_{tj} is the weight assigned to that state at time step t.

In turn, the weight α_{tj} of each hidden state h_j is computed by the softmax shown in Equation 3.2:

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{T_n} exp(e_{tk})} \tag{3.2}$$

where e_{tj} is a score calculated for each encoder state. Recall that the softmax transforms its input into a probability distribution, that is a set of values between 0 and 1 that add up to 1. As such, α_{tj} work as weights for the weighted sum in Equation 3.1.

In (Bahdanau et al., 2015) this score is learned by a feed forward network (FF), as shown in Eq 3.3:

$$e_{tj} = FF(d_{t-1}, h_j) \tag{3.3}$$

where d_{t-1} is the previous decoder state.

There are alternative formulations of attention that vary the scoring function. Luong et al. (2015a) experimented with different scoring functions and concluded that, instead of a learned function, a simple dot product, as shown in Eq 3.4, both improves performance and reduces computation time and memory, since there are no parameters that need to be learned.

$$e_{tj} = d_{t-1} \cdot h_j \tag{3.4}$$

The attention mechanism brought large improvements to all encoder-decoder architectures and has since become a staple of all NMT systems.

From this point onward, I will adopt the terminology used in (Vaswani et al., 2017) for describing the attention mechanism, which describes it as being computed with the use of *Queries*, *Keys* and *Values*. The decoder state, d_{t-1} in Equations 3.3 and 3.4, is referred to as the *Query*; while the hidden states of the encoder, h_j, are referred to as *Keys* when used in the scoring function (Equations 3.3 and 3.4), and as *Values* when used in the weighted sum (Equation 3.1).

3.2.3 Transformer

Taking the attention mechanism to the extreme, Vaswani et al. (2017) drop the recurrent mechanisms of previous architectures and rely solely on attention. This not only results in a simpler model which is also more efficient than recurrent models due to its lack of temporal dependencies, it also achieves better results than other approaches Vaswani et al. (2017). The BLEU scores for the before mentioned tests are 41.8 (En $\rightarrow$ Fr) and 28.4 (En $\rightarrow$ De). Given its currently undisputed claim as the best NMT architecture, I chose it for the current study. The Transformer will be described in detail in Chapter 4.

3.2.4 Learning a Vocabulary

An NMT model cannot contain all the known words in a language. Not only in principle, as new words are constantly being formed, but also in practice, as the embedding layer (see Section 3.2.1) would grow too large to feasibly handle the model (the English language has more than 450 thousand words).[3] Therefore, a subset of words, usually the most frequent, is selected to form what is called the vocabulary.

[3] Webster's Third New International Dictionary, Unabridged, together with its 1993 Addenda Section, includes some 470,000 entries. The Oxford English Dictionary, Second Edition, reports that it includes a similar number.

Starting corpus	low lower newer wider
Initialize the vocabulary	l; o; w; e; r; n; i; d;
First merge	e + r = er
Vocabulary	l; o; w; e; s; t; r; i; d; er;
Second merge	l + o = lo
Vocabulary	l; o; w; e; s; t; r; i; d; er; lo;
Third merge	lo + w = low
Vocabulary	l; o; w; e; s; t; r; i; d; er; lo; low;
Apply vocabulary to corpus	low low@@ er n@@ e@@ w@@ er w@@ i@@ d@@ er

Figure 3.4: Word piece algorithm for a maximum of three merges

I like sing@@ ing in the rain with Fit@@ z@@ ge@@ rald .

Figure 3.5: A possible segmentation into word pieces for the sentence “I like singing in the rain with Fitzgerald.”. The sequence ‘@@’ denotes the continuation of a word.

A problem is faced when using word vocabularies, which is the problem of Out Of Vocabulary words (OOV). OOVs are words that are not contained in the vocabulary yet appear in a sentence that the models sees (either during training, testing or after deployment). In order for the model to keep working, these words are replaced with a special reserved symbol (eg. <OOV>) that is included in the vocabulary. However, performance takes a large hit whenever there is an occurrence of an OOV word.

There are various studies (Luong et al., 2015b; Jean et al., 2015) that try to mitigate this problem, a common solution being copying the OOV source word to the output, or using character vocabularies.

Sennrich et al. (2016b) devised yet another method to present data to a model, which could be described as a hybrid of word based and character based vocabularies. In this method, the vocabulary is made of *word pieces*, which can be whole words, parts of words or individual characters.

This method starts by building a vocabulary with all the individual characters, as this ensures that there is no word that cannot be represented. Then, it merges entries in the vocabulary by the frequency this aggregation appears in the training corpus. It stops when a predefined number of merges is made or the vocabulary reaches a certain size. Figure 3.4 illustrates the process of creating a word piece vocabulary, and Figure 3.5 shows an example of a sentence segmented into word pieces.

Most systems nowadays use this method for vocabulary creation, as are the cases of Google’s system, ConvS2S and Transformer.

3.3 Evaluation Metrics

While preferable for its quality, human evaluation is slow and expensive. Therefore, the use of automatic evaluation metrics is the main form of performance assessment for MT. These automatic evaluation metrics can range from simple ones like the F1-score or perplexity, to more complex ones like NIST (Doddington, 2002), ROUGE (Lin and Hovy, 2003), and BLEU (Papineni et al., 2002).

Bilingual Evaluation Understudy (BLEU) is the most used automatic metric for translation quality measure. It is given by an n-gram modified precision defined as

$$BLEUScore(y, \hat{y}) = exp\left(\frac{1}{N}\sum_{n=1}^{N} P_n(y, \hat{y}) * BP(y, \hat{y})\right) \tag{3.5}$$

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. These functions are defined by the following Equations.

$$P_n(y, \hat{y}) = \frac{\sum_{ngrams \in \hat{y}} CountClip(ngram)}{\sum_{ngrams \in \hat{y}} Count(ngram)} \tag{3.6}$$

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 $\hat{y}$.

$$BP(y, \hat{y}) = \begin{cases} 1, & \text{if } length(\hat{y}) > length(y) \\ exp\left(1 - \frac{length(\hat{y})}{length(y)}\right), & \text{otherwise} \end{cases} \tag{3.7}$$

Every evaluation metric has its strengths and weaknesses, and the shortcomings of BLEU are widely known in the scientific community (Callison-Burch et al., 2006), such as its lack of correlation with human judgment. However, probably because of BLEU being well studied and these weaknesses being well defined, as well as it allowing comparison with previous works, it still is the most used automatic metric.

In the present dissertation the 4-gram BLEU ($n = 4$) is used for performance assessment of the various MT systems studied.

3.4 Portuguese $\leftrightarrow$ Chinese Machine Translation

Work specifically done on the chosen languages is rare, with Macau driving most of the research effort for the pair. Macau has two official languages, Chinese and Portuguese, the latter due to it having been under Portuguese rule from the XVI to the XX century.

Wong (2001) and Wong and Chao (2010) published papers on the topic, offering the reader various options either on Machine Translation or tools to help translators and teachers of both languages. These tools range from bilingual dictionaries to rule based models.

Direction	**BLEU**
PT → ZH	33.42
ZH → PT	35.69

(a) (Chao et al., 2018)

Direction	**BLEU**
PT → ZH	18.68
ZH → PT	25.11

(b) (Liu et al., 2018)

Table 3.1: BLEU scores from related work

Suitable corpora for Neural Machine Translation between Chinese and Portuguese were not available until recently, when Chao et al. (2018) created a 6 million sentences parallel corpus for PT ↔ ZH by scraping governmental sites of Macau. From this corpus, they made 1 million sentences available for public use.

In their paper they use their full corpus (with 6M sentences) and train a Recurrent Neural Network system with two layers on the encoder and decoder sides. Results obtained seem high in absolute terms (see Table 3.1a), yet they make use of a test set taken from the same distribution as their training corpus. This makes comparison between approaches that do not use the same training corpus unfair. Due to this, as will be explained in Section 4.2.4, a different test set corpus will be adopted in this dissertation, in order to allow fair comparison between the various approaches studied here.

Like Chao et al. (2018), Liu et al. (2018) also created a PT ↔ ZH corpus extracted from the governmental sites of Macau, but only with 0.84 million parallel sentences. However, none of it has been publicly released. Reported results (see Table 3.1b) are once again on their own test set, therefore no reproduction or comparison is possible.

Chapter 4

Implementation

In order to pursue the goal of creating a system that translates between Portuguese and Chinese using only freely available data and tools, a few choices had to be made. In this Chapter I discuss these decisions.

Firstly, in Section 4.1, the three approaches followed for training the system are described. These are the (i) *direct approach*, the (ii) *pivot approach*, and the (iii) *many-to-many approach*.

Section 4.2 describes the corpora chosen for the training of each approach, as well as the corpus used as test set.

The Transformer, which is the current state of the art architecture for Neural Machine Translation[1] and my choice of architecture for this work, is presented in Section 4.3. In that same Section I present the training options used for each approach.

Finally, the training times for every approach are reported.

4.1 Approaches to Training

A core issue in Machine Translation is how to make the best use of the available parallel data. Hence, in the present work I experiment with three different approaches to training an NMT system, which are described below.

4.1.1 Using a Different Model for Each Direction (Direct)

A straightforward option to create an MT system for a pair of languages is to use a parallel corpus of these languages.

For the language pair under study, a single Portuguese-Chinese parallel corpus will allow to create two models, one for each translation direction, that is a PT → ZH model and a ZH → PT model.

[1] The superiority of this model is confirmed with the WMT 2018 conference (Bojar et al., 2018), where 29 out of the 38 systems presented there used the Transformer.

One might expect this approach to yield the best performance given I am training separate models, each specific to a language pair and direction. This is the way most of the literature tackles the problem of translation, and the approach that normally sets the state of the art performance for most language pairs.

As neural network models need large amounts of data, underperformance with this approach is encountered for languages for which there is little parallel corpora available.

I refer to this solution as the *direct approach* throughout this document.

4.1.2 Using a Pivot Language (Pivot)

For some pairs of languages, there are few parallel corpora available. The pair Portuguese $\leftrightarrow$ Chinese is one such case (Chao et al., 2018). In this circumstance, it might be more advantageous for the translation to go through an intermediate third language, the pivot language, in a two-step process, as there might be more data available for the source-pivot and pivot-target pairs than there is for the source-target pair. This may permit to train two systems where concatenation delivers better performance than a direct approach with fewer data, in spite of the accumulated losses in the two steps.

The first system starts by translating from Portuguese or Chinese to the pivot language and then, the second system, translates from the pivot language to Chinese or Portuguese, respectively. So, all in all, four models are needed in order to accomplish the translation in both directions. The data used are parallel corpora for Portuguese $\leftrightarrow$ pivot and Chinese $\leftrightarrow$ pivot. Note that there is no direct translation between Portuguese and Chinese in this approach.

A subtle problem in this approach is when there is an idiomatic expression specific to the pivot language. For example, the sentence “Ele pontapeou o balde” (He kicked the bucket) only has the literal meaning of kicking a bucket in Portuguese, yet if the expression is translated to English and then to another language, the probability of it being translated as if the subject of the sentence has died is high, as this non-literal reading is present in the pivot language.

This approach is referred to in this work as the *pivot approach*.

4.1.3 Using a Single Model for All Pairs (Many-to-Many)

Another approach that can be resorted to is to gather all available parallel data into a single corpus.

Following the ideas from Johnson et al. (2017), Lakew et al. (2017), and Aharoni et al. (2019), the so called zero-shot machine translation seems to be a useful approach for Neural Machine Translation between low resourced languages. This consists in giving more language pairs to a model for training than those available under the direct approach in order to improve translation quality of less resourced pairs, and even translate between

Source	**Target**
<pt> What is your name?	Qual é o teu nome?
<zh> What is your name?	你叫什么名字?
<en> 你叫什么名字?	What is your name?
<pt> 你叫什么名字?	Qual é o teu nome?
<zh> Qual é o teu nome?	你叫什么名字?
<en> Qual é o teu nome?	What is your name?

Figure 4.1: Tagging the source sentence with the target language in the corpus for the many-to-many approach (pt corresponds to Portuguese, zh to Chinese and en to English).

pairs that are not seen in training.

To this extent, a system consisting of a single model was created from a corpus composed by all the data used for the direct and pivot approaches. In order to know to which language the system should translate to, a special token is appended to the beginning of the source sentence denoting the language of the target sentence. In the present work, the ISO 639-1 code of the language is used as this special token, as exemplified in Figure 4.1.

An advantage of the many-to-many system is there being more data available than for either of the two previous approaches, as it is capable of using all their corpora for training, and in this way provide the model with more data. On the flip side, the model has to contend with a much more difficult task that may decrease its potential performance.

I refer to this solution as the *many-to-many approach*.

4.2 Corpora

The three approaches indicated above require or benefit from different types of parallel corpora, which are discussed in this section.

4.2.1 Direct Approach Corpora

Parallel data for the PT ↔ ZH pair is scarce (Chao et al., 2018). Existing corpora are normally of low quality and/or low quantity, which leads to training sub-optimal neural networks for Machine Translation. This happens because the PT ↔ ZH language pair has not been the focus of much research and, as such, there are few suitable corpora for training a Neural Machine Translation model.

The Open Subtitles 2016 PT ↔ ZH (Tiedemann, 2012) corpus is a corpus of subtitles of movies and TV shows. While its size is considerable (5 million sentence pairs), its quality is very low. Many sentences do not align with each other, the grammar is poor, and there are too many short sentences, with the average sentence length being around 14 tokens, making it not suitable for training.

		Chinese			Portuguese		
Domain	**Sentences**	**Average Length**	**Tokens**	**Vocabulary**	**Average Length**	**Tokens**	**Vocabulary**
News	146,095	28.40	4,148,669	69,691	36.00	5,259,712	65,462
Legal	173,420	18.92	3,280,904	77,081	21.22	3,680,346	77,701
Subtitle	250,000	9.16	2,289,436	48,842	10.79	2,698,296	70,461
Tech.	250,000	22.06	5,514,523	53,717	24.41	6,102,664	64,262
General	250,000	21.54	5,385,459	87,707	26.37	6,592,183	121,074
Total	**1,069,515**	**19.28**	**20,618,991**	**200,163**	**22.75**	**24,333,201**	**224,481**

Table 4.1: UMPCorpus training set distribution

Portuguese	Havia 28.000 pessoas na conferência de neurociências este ano.
Chinese	今年的神经系统科学研讨会我们有28000 个专家参与。

Figure 4.2: UMPCorpus example from the "general" category

The News Commentary V11 corpus for PT ↔ ZH (Tiedemann, 2012) is of good quality. Yet, its size is small, with only around 10 thousand sentence pairs. This makes the corpus not suitable for Neural Machine Translation training. Nevertheless, it will be used as test set because of its quality. This will be better described in Section 4.2.4.

Tanzil (Tiedemann, 2012) is a collection of religious Quran texts. Like the News Commentary corpus, is a good quality corpus but with a small size (12,000 sentence pairs). It would be useful for testing purposes but will not be used because it is on the very specific domain of religion.

After an extensive search in conferences proceedings, blogs, and paper archives to find sufficiently good corpora that would permit to proceed with my work, I found the UMPCorpus (Chao et al., 2018), developed in the department of Computer and Information Science of the University of Macau, China. This corpus was released on May 2018, a few months before I started the work leading to my dissertation. To the best of my knowledge, there is currently no other research done with this corpus beyond the one reported in the original Chao et al. (2018) paper and the one in the present dissertation.

Despite indicating that the corpus has around 6 million sentences, Chao et al. (2018) only make available for public use a subcorpus with 1 million PT ↔ ZH parallel sentences, together with an additional 5,000 sentences for testing purposes.

UMPCorpus includes texts from five domains, namely law, subtitles, tech, news and general. Tables 4.1 and 4.2 show the distribution of both test and train corpora as presented in the original paper. Note that the authors do not provide the number of tokens and the vocabulary size for the test set and, due to different possibilities for Chinese segmentation (see Section 4.3.3), these values could vary depending on the segmentation algorithm used.

The corpus is evenly distributed between domains, with a little less emphasis on the legal and news domains. This corpus has a large variety of texts ranging from small sentences with big lexical diversity (Text-Type Ratio) as in the case of the subtitle domain,

Test Set	Sentences	Average Length	
		Chinese	Portuguese
News	1,000	27.63	34.09
Legal	1,000	28.56	31.78
Subtitle	1,000	8.71	9.92
Tech	1,000	22.47	24.86
General	1,000	22.13	26.02
Total	**5,000**	**21.90**	**25.33**

Table 4.2: UMPCorpus test set distribution

to larger sentences from the news domain, and sentences with smaller lexical diversity as this is the case of the technology domain.

This corpus is the one that will be used for training in the direct approach. An example from the corpus is shown in Figure 4.2.

As mentioned previously, Chao et al. (2018) separate 5,000 parallel sentences to be used as test set. I found that using this corpus as test set could bias negatively the evaluation of the other two approaches because this test set is very similar to the corpus used for training of the direct approach. Hence, I decided to use these 5,000 sentences of the UMPCorpus as development set.

The corpus used for testing will be described in Section 4.2.4.

4.2.2 Pivot Approach Corpora

For the pivot approach, there was the need to find parallel data involving both Portuguese, Chinese and a pivot language. The pivot language chosen was English (EN) given the availability of parallel language data between English and both Portuguese and Chinese, and given the quality and quantity of those data.

Portuguese ↔ English Corpora

The corpus used for the pair Portuguese ↔ English resulted from the concatenation of four corpora.

These four corpora were taken from the OPUS repository (Tiedemann, 2012). The corpus with fewer sentences is Tanzil, with 0.1 million sentences that are translations from the Quran, followed by JRC-ACQUIS and Europarl (version 7), which have respectively 1.6 and 2 million sentences, with law texts of the European Union and the translations of the sessions of the European Union Parliament. Finally, Paracrawl,[2] which consists of data crawled from the web, is the largest with around 3.3 million parallel sentences.

[2] Opus only gives download option for the first version of Paracrawl (Paracrawl V1). I used a third version of this corpus found here: https://paracrawl.eu.

Corpus (Domain)	Sent.
Tanzil (Religious)	0.12M
JRC-ACQUIS (EU Law)	1.63M
Europarl (EU Parliament)	1.96M
Paracrawl (Web Crawl)	3.25M
Total	**6.96M**

(a) PT ↔ EN pair

Corpus (Domain)	Sent.
News Commentary v11 (News)	0.07M
Tanzil (Religious)	0.19M
UMCorpus (Various)	2.22M
MultiUN (United Nations)	9.56M
Total	**12.04M**

(b) ZH ↔ EN pair

Table 4.3: Pivot corpus distribution

The Paracrawl data set is a structured file that contains translations that were filtered by the cleaning tool Bicleaner. For each translation pair, the document has meta-data with properties such as sentence length or markers for the occurrence of special characters. Therefore, with the use of this meta-data, I further cleaned this corpus by removing:

- pairs where either sentence was marked as very short (shorter than 3 tokens);
- pairs of sentences with mismatching Arabic numerals;
- pairs of sentences that were identical in both translation sides;
- pairs where either sentence had no letters (only numbers or symbols);
- and pairs where the length ratio between the sentences was larger than 3:2.

The final corpus, summarized in Table 4.3a, has close to 7 million parallel sentences.

For development purposes I use the first 5,000 sentences from the News Commentary V11 PT ↔ EN corpus.[3]

Chinese ↔ English Corpora

For the ZH ↔ EN directions, four corpora were gathered, with around 12 million sentences in total.

These four corpora are, from smallest to largest: NewsCommentaryV11,[4] composed of texts from news articles, with 0.07 million sentences; Tanzil, with 0.19 million sentences from the religious domain; UMCorpus (Tian et al., 2014) with 2.2 million sentences, from the same research group as the UMPCorpus PT ↔ ZH paper; and finally

[3] There is no overlap between the Portuguese sentences in this development set and the Portuguese sentences in the PT ↔ ZH corpus used as test set.

[4] NewsCommentaryV11 has around 5% of noise. Several sentences are in Hindu and others in Indonesian. This makes evident that even corpora that is highly regarded as having good quality can have serious problems.

the MultiUN Corpus, with 9.5 million sentences of documents from the United Nations. An overview is presented in Table 4.3b.

Similar to what I did for the direct approach, I used the 5,000 sentences that are provided in addition with the ZH ↔ EN UMCorpus as the development set.

The test corpus for both directions of the pivot approach is the same that is used for the direct approach. It will be described in Section 4.2.4.

4.2.3 Many-to-many Corpora

The many-to-many approach benefits from being supported by more corpora than the other two approaches. It benefits from all kinds of parallel corpora where one of the languages of interest occurs, in our case Portuguese or Chinese.

The final corpus consisted of all the data used by the previous two approaches, i.e. the 1 million sentence pairs from the direct approach, the 7 million sentence pairs used in the pivot approach for the PT ↔ EN directions, and the 12 million pairs also used in the pivot approach for the ZH ↔ EN directions.

All the data was duplicated, and by means of appropriate prefixation (as described in Section 4.1.3), every sentence pair was given to the model in both directions, resulting in a corpus with 40 million sentence pairs.

For development purposes, I used the same development set as for the direct approach, that is the 5,000 sentences provided in addition to the UMPCorpus. They were duplicated for both directions PT ↔ ZH by prefixation.

The test corpus is the same that is used for the other two approaches. It will be described in Section 4.2.4.

4.2.4 Evaluation Corpora

A test set that is different from the training data and the development data is needed to assess the performance of a given model for input data not seen during training. This includes assessing its performance on data from distributions that are different from the training distribution, thus assessing its ability to generalize to new data.

As already mentioned, PT ↔ ZH corpora are scarce, yet there are some high quality corpora that even though small, are suitable for evaluation. One of these corpora is the News Commentary v11 Tiedemann (2012) that contains translations on the news domain.

Despite my objective not being a Neural Machine Translation system specifically on the news domain, this corpus turns out to be a good choice as a test set because its content is diverse (as news usually cover a variety of topics), its sentences are not trivial and, since news are well curated, it has high quality translations.

The first 1,000 sentences of the News Commentary v11 for PT ↔ ZH are used as evaluation test set for all approaches. This test set length was chosen as 1,000 is within

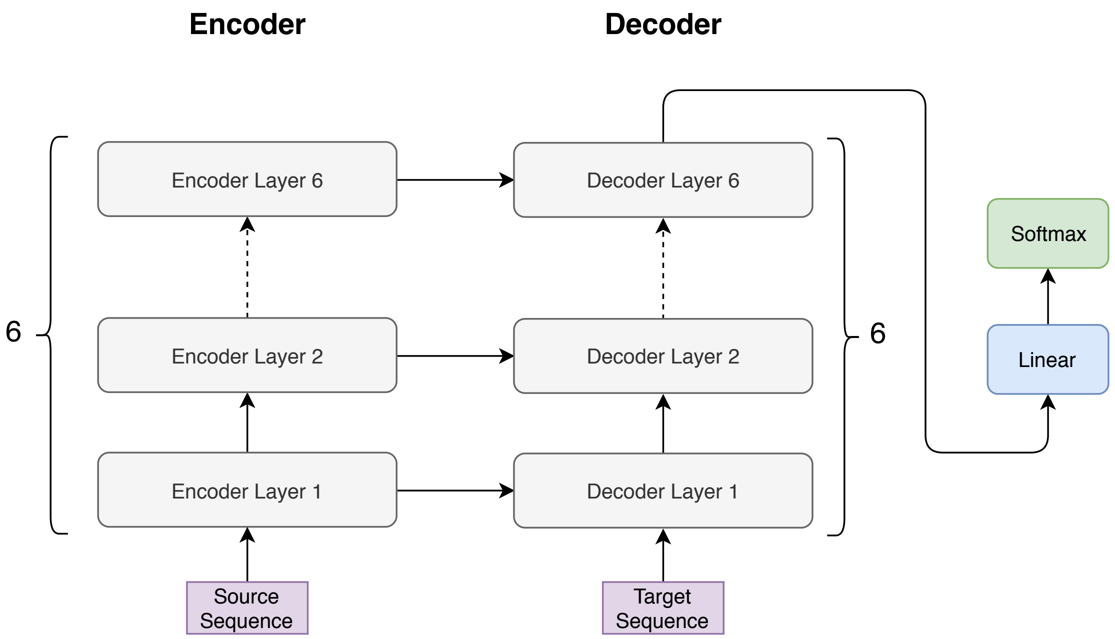


Figure 4.3: The Transformer - Overview.

the length range of test sets used in the literature. The first sentences were chosen in order to make it easy to reproduce my experiment, or compare with alternatives done by third parties, with this same test set.

4.3 The NMT System

As discussed in Section 3.2.1, there are different architectures available for Neural Machine Translation. Among them, the Transformer is widely regarded as being the state of the art, a claim supported by the last Conference on Machine Translation (WMT 2018, (Bojar et al., 2018)), where the best performing systems for every task were Transformer based. Accordingly, this is the architecture that I adopted for this work.

4.3.1 Transformer

The Transformer (Vaswani et al., 2017) is a rather recent architecture, but it has quickly established itself as the state of the art for NMT. It follows the standard deep encoder-decoder architecture to learn a mapping between a source sequence and a target sequence.

Similarly to other recent NMT architectures, the Transformer makes use of stacked layers (a deep network), as can be seen in Figure 4.3, which enables it to represent information at various levels of abstraction. Deep networks achieve great performance by learning to represent concepts hierarchically, with each concept being built upon simpler concepts from the previous layers. For example, given a sentence, one of the lower layers may learn mostly morphology, with the next layer being able to increasingly cope with

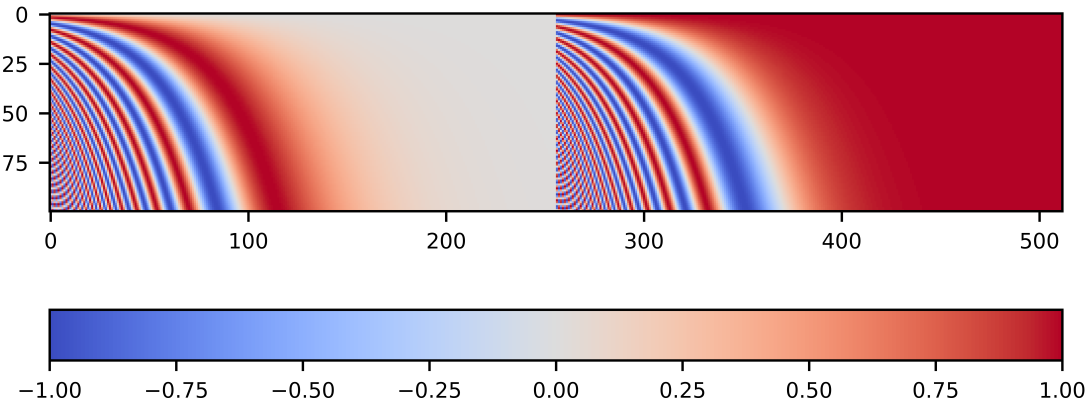


Figure 4.4: Sinusoidal positional embeddings. For word position up to 100 (vertical axis) and embedding space of size 512 (horizontal axis).

syntax, while another deeper layer representing, to a large extent, semantics content of the sentence.

The main innovations of the Transformer model are in (i) how it relies solely on attention, dispensing with any of the recurrent modules of previous architectures; and (ii) how it resorts to multiple heads of attention and self-attention. Next I describe each of these innovations in further detail.

No recurrent modules. As usual in neural approaches to language processing, the words in the source and target sequences are represented as vectors in an embedding space (see Section 3.2.1). Recurrent and convolutional architectures are intrinsically able to keep track of word positions in their input sequences. However, since the Transformer does not use a recurrent mechanism, information about the position of the words in the sequences needs to be explicitly added to the input source and target sequences. In (Vaswani et al., 2017), this is done through sinusoidal positional embeddings, which are provided by the following equation:

$$PE(pos, i) = \begin{cases} sin(pos/10000^{i/d}), & \text{if } i < d/2 \\ cos(pos/10000^{i/d}), & \text{if } i \geq d/2 \end{cases} \tag{4.1}$$

where pos is the word position in the sentence, d is the model embedding dimension and i is the index of a position in the embeddings vector.

A pictorial representation of positional embeddings, for vector size d=512 and word positions up to 100, is shown in Figure 4.4. Each horizontal line corresponds to a word position in the sentence (first word is at position 0) and represents the positional embedding vector that is to be added to the usual embedding of that word. The first half of the values of the positional embedding vector are given by a sin function while the second half are given by a cos function.

In their paper, Vaswani et al. (2017) also experimented with learned positional embeddings, which achieved similar performance to sinusoidal positional embeddings. Nevertheless, the authors argue that learned embeddings cannot extrapolate to sequence lengths longer than the ones encountered during training, and their learning is yet another computational burden to the model. As such, in this work I use sinusoidal positional embeddings.

Note that, as an additional benefit, not having recurrent modules allows to greatly accelerate training of the model since its layers are almost only feed forward layers and do not have temporal dependencies between them.

Multi-head attention. Another innovation of Vaswani et al. (2017) is the introduction of multi-head attention and self-attention.

As mentioned in Section 3.2.2, the attention mechanism of Bahdanau et al. (2015) receives as input a decoder state (the *Query*) and the set of all encoder states (which are simultaneously the *Keys* and the *Values*).

In multi-head attention, instead of a single attention calculation, there are multiple ones, each taking as input Queries, Keys and Values that have been transformed by different learned linear transformations, as shown in Equation 4.2.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{4.2}$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$.

Multi-head attention for a Query Q, set of Keys K and corresponding Values V is the concatenation of several separate attention calculations ($head_i$) with a linear transformation W^O. Each attention calculation $head_i$ works as the regular attention of (Bahdanau et al., 2015), but its inputs undergo different, learned linear transformations. Namely, for head i, W_i^Q transforms the Query, W_i^K transforms the Keys, and W_i^V transforms the Values. The rationale behind this technique is to allow each attention head to focus on some different aspect of its input.

Self-attention. In the normal attention mechanism, the Query corresponds to a decoder state while the Keys and Values originate from the encoder. In self-attention, both the encoder and the decoder have their own attention mechanisms which refer to their own stack of layers. That is, for the encoder self-attention, the Query, Keys and Values all originate from the encoder layer below (for the first layer this corresponds to the embedding layer), and similarly for the decoder.

While Figure 4.3, shown before, provided a high-level overview of the stacked layers, Figure 4.5 zooms in on a layer (the first one) to better show the multi-head and the self-attention mechanisms. Note that all subsequent layers have the same components yet their inputs are provided by the layer right before instead of the word embeddings of the

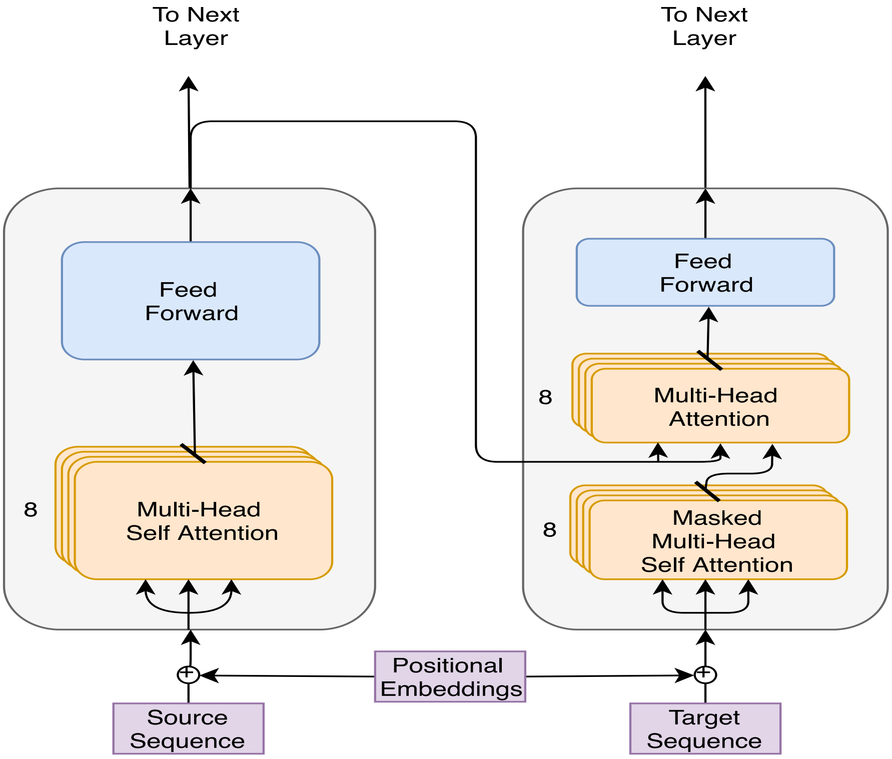


Figure 4.5: The Transformer - first layer.

source/target sequence. The number of deep layers (6) and the number of attention heads (8) depicted in the Figures 4.3 and 4.5 are the ones used by Vaswani et al. (2017) in their Transformer base model. I opted for using the same layout in the architecture of my system, as it has been empirically chosen by Vaswani et al. (2017) and empirically proven to lead to good results.

In training, the source and target sequences are fed to the stack of encoder and decoder blocks. These sequences are composed of the embeddings of the words that form the sentence. Each block begins by applying multi-head self-attention to all the word embeddings from the layer below.

Note that, for the decoder blocks, self-attention is masked in order to hide the word being predicted and the words that follow it. This prevents the leakage of future information as, while training, the model cannot be allowed to access the words that it has not predicted yet.

For the encoder blocks, the outputs of all heads are concatenated and the result is run through a feed forward layer that outputs a sequence of vectors, such that the sequence and the vectors have the same dimension, respectively, of the length of the sequence and the size of the word vectors in the source/target sequence input.

For the decoder blocks, the output of self-attention, before being run through a feed forward layer, goes through an additional multi-head attention component, this one being the attention mechanism that allows the decoder to attend over encoder states. This is the entry point of encoder information into the decoder, where the Query comes from the decoder self-attention layer and the Keys and Values come from the encoder block.

Finally, and coming back to the Transformer overview in Figure 4.3, the output of the final decoder block on the stack is fed to a linear layer that projects it into a larger vector, with the size of the target vocabulary, called the logits vector. This logits vector is then fed to a softmax layer. This softmax layer creates a probability distribution over the target vocabulary, from which the word with the highest probability is chosen to be the word predicted by the decoder.

During training, the output probability distribution given by the softmax is compared with the one-hot-encoded vector that represents the target word,[5] and the error (loss) between the objective vector and the predicted vector is backpropagated (Rumelhart et al., 1986) through the network, adjusting its weights.

During translation with an already trained model, at each time step, the word that is predicted by the decoder is appended to the target input sequence. This process is repeated until the end of sentence symbol is generated or a pre-defined maximum sentence length is reached.

4.3.2 Training Options

In the present study, the same hyper-parameters as the Transformer Base[6] from (Vaswani et al., 2017) are used, with 6 encoder and decoder layers, 8 attention heads and an embedding size of 512. The full configuration of the model is given in detail in the Appendices (cf. Appendix B).

Differently from the original Transformer, where shared embeddings are used, in this study the embeddings for the source language are separate from those for the target language. Shared embeddings is when both source and target languages share the same embedding mapping between their tokens and the embedding space. This allows the model to adjust the same embeddings two times during the encoder and decoder back propagation. Though it could happen that words that are written in the same way but not share the same meaning could be given the same embedding, the benefits of this approach normally outperform this drawback.

One of the first papers to present shared embeddings (Wu et al., 2016) mentions that in translation it often makes sense to copy rare entity names or numbers directly from the source to the target (i.e. leave them untranslated). To facilitate this type of direct copying, the use of shared embeddings between the source language and target language guarantees that the same token in source and target sentence will be represented by the same embedding, making it easier for the system to learn to copy these tokens. This

[5] Recall that, in the one-hot-encoded vector of a word, all positions have a value of zero, except for the position corresponding to that word, which has a value of one.

[6] ”Base” is the name of one version of the models presented in (Vaswani et al., 2017).

method is adopted by the most recent NMT architectures.

However, given the nature of the two languages in this work, whose writing system is different, using shared embeddings would severely reduce the vocabulary size available for each language, since there are no overlapping words between the two languages.

This was confirmed with two experiments, where one of them used shared embeddings and the other did not. These experiments are better reported in Section 5.

4.3.3 Pre-processing

Before the training of the model, data should be pre-processed in a way that it will help the model to achieve better performance, and this takes an even more important role when a language like Chinese is involved.

Chinese has little to no word separation. This is illustrated in Figure 4.2 (page 24), where what looks like one big word (before the number "28,000") are in fact several words concatenated. Separation in a sentence happens only by punctuation.

Therefore, in order to ensure better performance, the pre-processing steps described below were undertaken.

Segmentation

Neural Machine Translation models are based on sequences of symbols (words or word-segments), therefore word separation is necessary before training.

For Chinese, segmentation is a non-trivial NLP task. There are several ways of doing segmentation of sentences and differences in segmentation can lead to sentences not being well formed or even alter their meaning. As such, segmentation is an important step that heavily influences the quality of the model.

There are several alternatives, from rule based to neural based. The one used in this work is the Jieba Segmenter.[7] It was recommended by a Chinese native speaker,[8] who is a researcher in the area of Natural Language Processing and Neural Machine Translation, for its quality and ease of use.

Tokenization

All texts for Portuguese and English were pre-processed with the help of the Moses Tokenizer from the Sacremoses Package,[9]. This is a simple rule-based tokenizer that mainly separates punctuation and converts certain symbols to a different representation (for example the quote symbol ' is converted to &apos;). This makes it so that there are no words with punctuation symbols attached (for example, commas normally appear together with

[7] https://github.com/fxsjy/jieba

[8] My thanks to Prof. Deyi Xiong for his help in this matter.

[9] https://github.com/alvations/sacremoses

	Portuguese
Original	Faça o meu rapaz ver isso, que ao lado do eterno “porquê”, existe um “sim”. *Do the my boy see that, that at-the side of-the eternal “why”, exists a “yes”.*
Tokenized	Faça o meu rapaz ver isso , que ao lado do eterno &quot; porquê &quot; , existe um &quot; sim &quot; .
Final	Fa@@ ça o meu ra@@ paz ver isso , que ao lado do e@@ terno &quot; por@@ quê &quot; , existe um &quot; sim &quot; .
	Chinese
Original	在无穷的“为什么” 的边上，一定存在”是的” 这种肯定的一面！
Segmented	在/ 无穷的/ / “/ 为什么/ ”/ / 的/ 边上/ ，/ 一定/ 存在/ / “/ 是/ 的/ ”/ / 这种/ 肯定/ 的/ 一面/ / ! *In/ endless/ / “/ why/ ”/ / of/ on-the-side/ ,/ for sure/ exist/ / “/ yes/ of/ ”/ / this-kind/ sure/ of/ one-side/ / !*
Final	在/ 无@@ 穷@@ 的/ / “/ 为什么/ ”/ / 的/ 边@@ 上/ ，/ 一定/ 存在/ / “/ 是/ 的/ ”/ / 这种/ 肯定/ 的/ 一@@ 面/ / !

Figure 4.6: Example of the pre-processing steps

a word), so that the word and the punctuation are fed separately during the training of the model, which can help the model understand different uses of the same symbol. For instance, in English the quote may mark possession or it may be a quotation symbol.

Learning Vocabulary

As mentioned in Section 3.2.4, there are several ways to feed NMT models with text sequences, from character based to word based input.

Input divided into sub-word units (Sennrich et al., 2016b), that is the division of words into smaller units, is the one used in the present work. Sub-word units are used because this way one can have a limited dictionary yet be able to avoid out of vocabulary (OOV) words since any word can be represented as a sequence of sub-words units, the most extreme case being representing a word as a sequence of individual characters.

For all approaches described in this document, a vocabulary with 32,000 entries was learned for each language,[10] except for the many-to-many approach where a joined vocabulary with Portuguese, Chinese and the pivot language (English) is learned.

Figure 4.6 shows the various steps of pre-processing for the sentence “Make my boy see that, that beside the eternal ‘why’, there exists a ‘yes’. ” in Portuguese and Chinese. The slash symbol followed by a space (“/ ”) is used by the Chinese segmentation tool to

[10] The script used to create the vocabularies can be found at https://github.com/rsennrich/subword-nmt.

mark a segmentation boundary. The "@@" mark indicates a word that was split into subwords. The token that appears after the mark is part of the token that contains the mark (e.g. the word "rapaz" (boy) is represented as two sub-word units, "ra@@" and "paz").

4.3.4 Marian Framework

There are several NMT frameworks, like tensor2tensor from Google,[11] Fair2Seq from Facebook,[12] OpenNMT from MIT,[13] among others, that implement many of the most popular architectures.

To help my implementation of the desired models, I adopted the Marian Framework (Junczys-Dowmunt et al., 2018), which is being developed at the Adam Mickiewicz University in Poznań (AMU) and at the University of Edinburgh.

The Marian framework offers a C++ implementation,[14] that tends to be faster than Python or LuaJIT (used in some of the other frameworks), has an easy API, and good documentation. It offers also the option for GPU or CPU training and decoding, minimal software dependencies and a permissive MIT license.

4.3.5 Training Times

All training was performed on a NVIDIA Tesla K40m GPU[15] and on a NVIDIA Titan RTX GPU.

The direct approach used the NVIDIA Tesla K40m GPU, and training took about 4 days for each direction, more precisely 3 days and 21 hours for the ZH → PT direction and 4 days and 8 hours for the PT → ZH direction.

The training of the pivot approach consisted on four different models, with the two models used for PT→EN→ZH translation taking around 18 days to converge (2 days for PT → EN; 16 days for EN→ZH), and the models for ZH→EN→PT also taking around 18 days (13.5 days for the ZH → EN direction and 4.5 days for EN→PT). The full training totals around 37 GPU days on a NVIDIA Tesla K40m.

Finally, the many-to-many approach, due to its size, had to be trained on the NVIDIA Titan RTX GPU. This GPU is faster than the other one, so the model only required 5 days to converge. Note, however, that these times are not comparable with the ones from the previous approaches, as a different GPU was used.

Nevertheless, an informative comparison between the many-to-many approach and the other two approaches can be made because of an initial test run where only half of the corpus (i.e. 20 million sentence pairs) was used to train the many-to-many approach, which

[11] https://github.com/tensorflow/tensor2tensor

[12] https://ai.facebook.com/tools/fairseq

[13] http://opennmt.net/

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

[15] My thanks to INCD (Infraestrutura Nacional de Computação Distribuída) for providing the computational resources that supported these experiments.

made it possible to be run on the same, lower-spec GPU as the other two approaches. This version, with the half corpus, required 34 days to converge.

4.4 Summary

This Chapter presented the work performed in this dissertation. It starts by describing the various training approaches explored, namely the (i) direct approach, which only uses parallel corpora between Portuguese and Chinese; the (ii) pivot approach, which relies on a third language as a broker for translation; and the (iii) many-to-many approach, which benefits from all training data of the other two approaches.

The Chapter also introduced the corpora used for the three approaches together with the pre-processing steps used for every corpus, as well as the adopted architecture for NMT training, the Transformer model and the framework that was used.

Chapter 5

Evaluation

This Chapter is concerned with the evaluation results. It begins by introducing the baseline, this is followed by the results for each of the approaches. Finally, some manual evaluation for the best approach is presented.

All evaluation is performed on the News Commentary test set (see Section 4.2.4).

5.1 Google Translate Baseline

In order to have a baseline against which the performance of the various systems I developed in the present study can be compared, I resort to the online service Google Translate.[1]

To obtain the relevant baseline score, I evaluated this service on the News Commentary test set. This established a very strong baseline to be challenged by my systems.

With Google Translate being one of the most used translation services around the world, any score near this baseline would be praiseworthy, taking into account the dimension of the company and the resources available to its MT team, in terms of qualified expert human resources, data and computational power.

Evaluation against such an industry giant was possible because Google Translate allows document translation, even though with a limit of 5,000 characters at a time. To circumvent this constraint, the corpus was divided into several blocks of up to 5,000 characters, which are translated one block at a time and finally concatenated to be scored.

When scoring test data with the BLEU metric (cf. Section 3.3), an issue had to be addressed. When translating to Chinese, Google Translate outputs sentences without word separation, as this is what is expected by the human users. Since BLEU is based on white-space separated token overlap, the BLEU scores on these sentences would be either 1 or 0,[2] depending on whether the automatic translation and the reference translation are, respectively, equal or different.

[1] https://translate.google.com/

[2] BLEU score are normally multiplied by 100 for presentation.

	Reference	**Hypothesis**	**BLEU**
With spaces	Alice likes Bob!	Alice likes Bob!	100.00
No spaces	AlicelikesBob!	AlicelikesBob!	100.00
With spaces	Alice likes Bob!	Alice likes Tom!	35.36
No spaces	AlicelikesBob!	AlicelikesTom!	0.00

Table 5.1: Example in English of how word spacing affects scores.

Original	你叫什麼名字?
Jeiba	你/ 叫/ 什/ 麼/ 名字/ ？
Stanford	你/ 叫/ 什麼/ 名字/ ？
Spaces	你/ 叫/ 什/ 麼/ 名/ 字/ ？

Figure 5.1: The Chinese sentence equivalent to “What is your name” in various forms of segmentation.

Table 5.1 shows an example of this where a sentence, in English for the sake of readability, is measured against a reference with and without spaces. As we can see, altering the name “Bob” to “Tom”, while lowering the scores in the sentence with spaces, still allows for a high value. However, when we remove all spacing, BLEU treats the whole sentence as a single “word” and a single different character is enough to bring the BLEU score down to zero.

To circumvent this problem there was a need to perform sentence segmentation of the output of Google Translate.

The Jieba segmentation tool was used on the training corpora. Using this same tool to segment the output of Google Translate could raise claims of favoritism, as the models could be tuned for outputing text with a segmentation similar to that produced by Jieba. Accordingly, to preemptively address such claims, I evaluate the PT → ZH output of Google Translate and the output of the three approaches under three different segmentations.[3] Namely, (i) Jeiba segmentation, (ii) segmentation using the Stanford segmenter,[4] and (iii) segmentation by inserting a space between every character. Examples of these variants can be found in Figure 5.1.

Results from Tables 5.2a and 5.2b show scores far from the values of state of the art for languages pairs like English ↔ French/German (cf. Section 3). This is to be expected, as the latter are pairs equipped with more and better corpora, and are the pairs that most research is conducted upon.

For the News Commentary v11 test set, Google Translate has a score of 12.23 BLEU for the ZH → PT translation direction. In the case of PT → ZH translation, scores are a little higher, with scores for sentences segmented with the Stanford segmentation tool

[3] Segmentation for Chinese can be ambiguous. Different segmentation tools may produce different results.

[4] https://nlp.stanford.edu/software/segmenter.shtml

System	BLEU
Baseline	**12.23**

(a) ZH → PT direction

System	BLEU
Baseline	
Stanford	**14.29**
Jieba	13.69
Space	24.92

(b) PT → ZH direction

Table 5.2: BLEU Scores for Google Translate baseline

	Reference	Hypothesis	BLEU
No character spacing	Alice likes Bob!	Alice likes Bob!	100.00
Character spacing	A l i c e l i k e s B o b !	A l i c e l i k e s B o b !	100.00
No character spacing	Alice likes Bob!	Alice likes Tom!	35.36
Character spacing	A l i c e l i k e s B o b !	A l i c e l i k e s T o m !	70.83

Table 5.3: Example (in English) of how character spacing inflates BLEU scores.

reaching 14.29 BLEU.

Comparing between the BLEU scores obtained with the various Chinese segmentation techniques, we see that the test set segmented with Jeiba stays behind the test set segmented with the Stanford segmentation tool, and the test set segmented with spaces leaping more than 10 BLUE points in relation to the other two segmentation techniques.

All these segmentation techniques for scoring Chinese texts with BLEU are not optimal, because word segmentation is not perfect in either case, and space evaluation allows for a bigger n-gram match between hypotheses and reference sentences, as only characters have to match. As can be seen in Table 5.3 different words can have matching characters (Bob and Tom share the letter “o”) and these matching characters inflate BLUE scores.

The BLEU scores obtained after segmentation of Chinese with the Stanford segmentation tool are the main comparison measures used between the various approaches and the baseline throughout the rest of the dissertation. This is done, as already mentioned, as to not give favoritism to any system.[5]

5.2 Direct Approach

For the direct approach, before the model was trained, the corpus was pre-processed with the Moses tokenizer, for Portuguese, and with the Jieba segmentation tool, for Chinese. A vocabulary with a maximum of 32,000 entries was created for each language with the most frequent sub-word units.

[5] The space segmentation could also be used for this purpose, yet its output, where each character is individually separated, is not a valid word segmentation.

System	BLEU
Baseline	12.23
Direct Approach	**13.38**

(a) ZH → PT direction

System	BLEU
Baseline	**14.29**
Direct Approach	
Stanford	11.05
As-is	10.20
Jieba	10.72
Space	20.03

(b) PT → ZH direction

Table 5.4: Direct approach BLEU scores

The pre-processed corpora are then fed to Marian (running the Transformer architecture). The training proceeds until perplexity or cross-entropy on the development set does not decrease for 10 evaluation steps in a row. This stopping criteria was the default in the Marian framework.

The BLEU scores on the News Commentary v11 test set for the direct approach can be visualized in Table 5.4. The line "As-is" reports the BLEU scores on the segmentation output by the model without re-tokenization.

Table 5.4a reports the performance of the system for the ZH → PT direction. The direct approach achieves 13.38 BLEU points, which is an improvement of 1 BLEU point over the Google Translate baseline (reported in the first line of the table).

For the PT → ZH direction the direct approach achieves 11.05 BLEU points, falling behind the Google Translate baseline by 3 BLEU points.

One possible explanation for being this far from the baseline in one direction and surpassing it in the other can be found in (Johnson et al., 2017). The authors refer that Google has started to shift their translation models to more compact ones that make use of zero-shot translation, where a single model can handle several language pairs. Hence, its performance is better for those languages that it has seen the most. It is likely that in their training data, there are more Chinese sentences (paired with languages other than Portuguese) than there are Portuguese sentences.

An experiment using this approach and shared embeddings was conducted. For this, a shared vocabulary with 32,000 sub-word units was learned, instead of two separate vocabularies as reported in this approach. The remaining parameters were kept the same as in the direct approach. The system trained with shared embeddings achieved 2 BLEU points less, for the two translation directions, than the direct approach.

The performance of the direct approach is very satisfactory. Although it has not been able to surpass the baseline for the PT → ZH direction, it is capable of breaking the very strong baseline given by Google Translate for the ZH → PT direction.

Corpus	**BLEU**
PT → EN devset	37.78
ZH → EN devset	23.10
EN → PT devset	38.82
EN → ZH devset	17.20

Table 5.5: Blue scores for the pivot system with development sets

System	**BLEU**
Baseline	12.23
Direct approach	13.38
Pivot approach	**17.79**

(a) ZH → PT direction

System	**BLEU**
Baseline	14.29
Direct approach	11.05
Pivot approach	
Stanford	**15.25**
As-is	14.83
Jieba	14.84
Space	25.37

(b) PT → ZH direction

Table 5.6: Pivot approach BLEU scores

5.3 Pivot Approach

Like for the direct approach, for the pivot approach the training data set is pre-processed with the Moses tokenizer for Portuguese and English, and with the Jeiba Segmentation tool for Chinese. Vocabularies for the 32,000 most frequent sub-word units are created for each direction.

The corpora used is a combination of several corpora available for the PT ↔ EN and ZH ↔ EN directions, totaling around 7 million parallel sentences for PT ↔ EN, and around 12 million for ZH ↔ EN.

The performance results are displayed in Table 5.6. The pivot approach obtains 17.79 BLEU points for the ZH → PT translation direction, which is an improvement of more than 4 BLEU points over the direct approach, that already surpassed the Google Translate baseline by 1 BLEU point.

While the chosen baseline had already been surpassed for the ZH → PT direction by the direct approach, the PT → ZH direction was still out of reach by more than 3 BLEU points, which is a big gap to fill. In spite of this, the pivot approach is able to outperform the chosen baseline for the PT → ZH direction by 1 BLEU point, achieving 15.25 BLEU points, against the 14.83 of the Google Translate baseline.

These results differ from the ones in (Liu et al., 2018), where the direct approach fares better than the pivot approach. To understand this difference, it is worth noticing that the

System	BLEU
Baseline	12.23
Direct appraoch	13.38
Pivot appraoch	**17.79**
Many-to-many approach	16.22

(a) ZH → PT direction

System	BLEU
Baseline	14.29
Direct approach	11.05
Pivot appraoch	**15.25**
Many-to-many approach	
Stanford	13.98
As-is	13.28
Jieba	13.51
Space	23.48

(b) PT → ZH direction

Table 5.7: Many-to-many BLEU scores

pivot approach benefits if both intermediate steps of translation are a lot stronger than the single step of the direct approach. Liu et al. (2018) use only 2 million sentences for the PT ↔ EN translation direction, making it not worth going for the pivot approach.

The scores obtained here show that the benefits from having additional data for the pivot approach outweigh the drawback of not going direct, in line with the widely held opinion that these systems are highly dependent on the quantity of data available.

5.4 Many-to-Many Approach

The third approach studied is the many-to-many one. With all the data from the previous approaches, it had the potential to surpass them.

Differently from the two other approaches, a shared vocabulary of 32,000 sub-word units was created, i.e. with all the English, Chinese and Portuguese sentences together. This was done this way because both the encoder and the decoder see sentences from the three languages.

The amount of shared vocabulary between Chinese and the other two languages (Portuguese and English) is practically non existent.[6] It would be beneficial to have a bigger vocabulary, yet the same vocabulary size was kept in order to change as few training variables as possible between the approaches.

While this approach was the one trained with the most data, its performance was not above the performance of all other approaches, falling behind the pivot one in both directions, as presented in Table 5.7.

For the ZH → PT translation direction this approach achieved 16.22 points BLEU, outperforming both the Google Translate baseline, by 4 BLEU points, and the direct approach, by 3 BLEU points.

[6] There could be some words written in Latin script, such as names of people or companies.

Like the direct approach, the many-to-many approach could not beat the chosen baseline for the PT → ZH translation direction, yet performing near the baseline with 13.98 BLEU points, that is only 0.31 behind it. Comparison with the other two approaches sees the many-to-many approach surpassing the direct approach by 4 BLEU points, and falling behind the pivot one by more than 1 point BLEU.

Despite not being the top performing approach, it shows big promise because it has the ability of incorporating more data as any parallel corpus where a desired language occurs as either source or target can be used. However, by giving more data, the task the model has to face may be increasingly more difficult, which may make the model not to take full advantage of all the additional data.

5.5 Manual Evaluation

It is known that BLEU scores do not necessarily correlate with human judgment of translation quality (Callison-Burch et al., 2006). As such, it is desirable to complement automatic evaluation based on BLEU with human evaluation. However, the latter is costly, with a human evaluator taking several hours to evaluate a few hundred sentences. In contrast, an automatic metric can evaluate a corpus with thousands of sentences in a few seconds. Furthermore, the language pair used in this study has a small number of bilingual speakers, making it hard to find suitable human evaluators.

Despite these adversities, I was fortunate to have the help of a volunteer native Chinese speaker.[7]

The evaluation task consisted in, given a Portuguese sentence and two Chinese translations of it provided by two MT systems, indicate the best translation. Only the PT → ZH direction was evaluated because the human evaluator was a native speaker of Chinese, and it would be significantly harder for her to assess the difference in quality between two Portuguese translations given a Chinese source sentence, than to assess the quality difference between two Chinese translations given a Portuguese source sentence.

Given that manual evaluation is costly, it was restricted to two systems, namely the Google Translate baseline and the pivot approach, which is the best performing approach under the BLEU metric.

The human evaluator was asked to evaluate 50 instances where a Portuguese source sentence was shown together with the translation from the Google Translate baseline and the translation from the pivot approach. These two translations were randomly ordered among all instances as not to bias the evaluation.

Among the 50 instances, I randomly introduced 10 filler instances to assess the correctness of the human evaluator. These filler instances were composed by the golden reference translation and the translation from the direct approach, which is the approach

[7] My thanks to YingYing Peng on her help in this matter.

Under translation	**The sentence in red is not translated.** . . . 2003-2008, e do subsequente tsunami de execuções hipotecárias. Mas as turbulências no mercado imobiliário estão a dissipar-se. 由于2003-2008年房地产市场扩张期间施工过度以及随后发生的抵押贷款执行海啸，住宅投资在国内生产总值中所占百分比仍然处于历史上的低位。
Wrong translation	**The acronym (ELS) is mistranslated.** Os ataques aéreos apenas piorariam as coisas, ao reduzir a legitimidade de base do ELS e ajudar as forças Islamitas. 空袭只能通过削弱ERS的基本合法性和帮助伊斯兰力量来恶化局势。
Repetition	**The phrase "que o" is repeated.** . . . e não emprestar mais do que o que o Congresso autorizou".

Figure 5.2: Error analyses of the pivot system.

that had the worst BLEU scores. This was done under the assumption that, in these cases, the gold reference translation is always better than the translation given by the direct approach, and as such should always be indicated by the human evaluator.

The test set used for the manual evaluation was extracted from the News Commentary v11 corpus, the one used for the automatic evaluation. The selected sentences were the first 50 sentences where the remainder of the line number divided by 10 equals 3 (i.e. lines 3, 13, 23, . . .).

Manual evaluation resulted in a close match between the pivot approach and the Google Translate baseline, with the pivot system being chosen 17 times over the Google Translate system, and the Google Translate system being chosen 23 times over the pivot system.

As for the filler instances, the human evaluator chose the gold reference 7 out of 10 times, a value that indicates a good evaluation reliability.

These close values in manual evaluation were already expected as this translation direction is the one that is closer in terms of BLEU scores between these two systems, with a difference of less than a BLEU point in favor of the pivot approach.[8]

An additional fact that has to be noticed is that this manual evaluation process does not assess how much better one translation is in relation to the other, but rather how many times one system performs better than the other (even if by a small margin). This is in contrast to the automatic evaluation metric used (BLEU), which quantitatively evaluates how much better the translations of one system are over the other.

Further analysis of the pivot system (in this case both directions are considered) shows

[8] The BLEU scores for these 40 sentences are 15.31 for the pivot approach, and 14.09 for the Google Translate baseline.

the expected issues with Neural Machine Translation, such as under-translation, over-translation, word repetition and translating a word with the wrong meaning.

Figure 5.2 illustrates some of these problems. Note that, for the sake of readability part of the sentences are cut (see Appendix C for the full examples). In the first example there is a case of under-translation, where the last sentence (in red) does not appear in the Chinese translation. The next example illustrates the wrong translation of the ELS acronym (Free Syrian Army - FSA, in English) to ERS. Finally, the last example shows the repetition of phrases where “que o” (that the, in English) appears two times in a row.

5.6 Extending the Corpus with Back-Translation

The Conference on Machine Translation (WMT) is the main MT evaluation event, where the state of the art is pushed forward. One of the most adopted techniques to enhance NMT performance used in the various systems presented in the conference is back-translation (in WMT’18 (Bojar et al., 2018), 22 of the 38 participating systems used this technique).

Back-translation is a simple method for extending corpora where a previously trained MT system on a chosen pair of languages is used to translate monolingual corpora in one of these languages to the other, hence creating additional ”synthetic” parallel data for those two languages.

I performed an experiment on back-translation where all the 450,000 Chinese sentences from the news domain of the UMCorpus were collected (Tian et al., 2014) and given to the model of the many-to-many approach for translation into Portuguese. The chosen sentences are all from the news domain in order to be the sentences that most favor the test set, since the News Commentary v11 test set is also on the news domain.

A bigger monolingual Chinese corpus could be collected, however studies made by Poncelas et al. (2018) point out that continuing to increase the amount of monolingual data gives diminishing returns and can even be harmful. Following similar experiments in the literature, I opted for 450,000 sentences.

The many-to-many approach model was used because it is as fast as the direct approach and faster than the pivot approach, since it translates without passing through an intermediary (pivot) language that would double translation time,[9] and has better performance than the direct approach.

A system, using the direct approach, was trained on the concatenation of the back-translated “synthetic” corpus and the UMPCorpus (Chao et al., 2018).

As expected the resulting system outperforms the direct approach (cf. Table 5.8) in both directions, with 15.29 BLEU for the ZH → PT direction and 13.17 for the PT → ZH translation direction, an improvement of 2 BLEU points in each direction.

[9] Translation of these 450 thousand sentences took almost 3 days.

Corpus	BLEU
Google Translate baseline	12.23
Direct approach	13.38
Pivot approach	**17.79**
Many-to-Many approach	16.22

(a) ZH → PT direction

Corpus	BLEU
Google Translate baseline	14.29
Direct approach	11.05
Pivot approach	**15.25**
Many-to-Many approach	13.98

(b) PT → ZH direction

Table 5.8: Summary of BLEU scores

This system trained with data extended with the back-translated corpora still fares worse than both the pivot and the many-to-many approaches, as well as not being capable of surpassing the Google baseline for the PT → ZH direction.

In this experiment only Chinese monolingual data was used to create a back-translated corpus. A potential improvement could consist of also using back-translated Portuguese monolingual data for the extension of the training corpus.

This experiment shows big potential to further improve both many-to-many and direct approaches, which are the ones that make use of PT ↔ ZH parallel corpora.

5.7 Summary

This Chapter presented the results obtained in this dissertation, with the best performing approach, the pivot approach, achieving 17.79 BLEU points for the ZH → PT translation direction, and 15.25 BLEU for the PT → ZH direction, outperforming the very strong baseline given by the Google Translate.

A summary of the results obtained on this dissertation can be seen in Table 5.8 (the results for each approach all use the Stanford re-segmentation), together with the BLEU scores obtained by the chosen baseline.

Chapter 6

Conclusion

This dissertation led me from the basics of Machine Translation all the way to being able to understand and apply the state of the art in this field for the development of a top performing MT system.

This Chapter concludes my dissertation. It presents a summary of the main results (Section 6.1) followed by a listing of the major contributions of this dissertation (Section 6.2), and closes with some pointers for future work (Section 6.3).

6.1 Summary

The main objective of this work was to address the challenge of determining how far one is presently able to go when developing Machine Translation solutions for both directions of the Portuguese ↔ Chinese (PT ↔ ZH) language pair making use only of freely available resources, and ultimately to develop a state of the art MT system that is able to translate from Portuguese to Chinese, and from Chinese to Portuguese. This objective has been successfully completed.

The fist part of the objective, that is determining how far one is presently able to go when developing MT solutions for both directions of the PT ↔ ZH language pair making use only of freely available resources, was accomplished by studying three approaches that make use of the available parallel corpora. These approaches were, the (i) direct approach, which only uses parallel corpora between Portuguese and Chinese; the (ii) pivot approach, which relies on a third language as a broker for translation; and the (iii) many-to-many approach, which benefits from all training data of the other two approaches.

The second part of the objective, concerned with developing a state of the art MT system, led me to apply all three approaches in the creation of various MT systems. These systems were compared with a very strong baseline, the Google Translate service, which has been created by a tech giant with access to a very large supply of expert human resources, of parallel data and of computational resources. Table 6.1 summarizes the BLEU scores obtained by the baseline and the three approaches I developed.

Corpus	BLEU
Google Translate baseline	12.23
Direct approach	13.38
Pivot approach	**17.79**
Many-to-Many approach	16.22

(a) ZH → PT direction

Corpus	BLEU
Google Translate baseline	14.29
Direct approach	11.05
Pivot approach	**15.25**
Many-to-Many approach	13.98

(b) PT → ZH direction

Table 6.1: Summary of BLEU scores

Considering the three studied approaches, the direct approach was the one with the lowest scores. However, it was still able to surpass the Google Translate baseline for the ZH → PT translation direction by more than 1 BLEU point. For the ZH → PT translation direction, the direct approach achieved 13.38 BLEU points against the 12.23 points from the baseline. This is a very satisfactory result, considering that this approach is the most straightforward approach studied here and the one that used the least amount of training data.

The second approach to train a MT system studied in this dissertation was the pivot approach. This approach achieved the best results, outperforming the baseline for both translation directions. When translating from Portuguese to Chinese, the pivot approach achieves 15.35 points BLEU, an improvement of around 1 BLEU points over the baseline (14.29). For the other translation direction (ZH → PT) the pivot approach achieves 17.79 BLEU points, an impressive improvement of over 5 BLEU points on the score of the baseline.

Finally, the many-to-many approach fared between the other two approaches. For the PT → ZH direction, it falls slightly behind the baseline, with 13.98 BLEU points against 14.29. For the other translation direction (ZH → PT), this approach achieves 16.22 BLEU points, another impressive improvement, with 4 points above the baseline.

6.2 Contributions

The major contributions resulting of the work performed in this dissertation are:

- **An exploratory study of PT ↔ ZH machine translation.** Three approaches to train an NMT system for PT ↔ ZH were studied, namely the (i) direct approach, which only uses parallel corpora between Portuguese and Chinese; the (ii) pivot approach, which relies on a third language as a broker for translation; and the (iii) many-to-many approach, which benefits from all training data of the other two approaches. With the training data available, the pivot approach outperforms the other two.

- **A translation system for PT ↔ ZH with state of the art performance.** The translation system based on the pivot approach outperforms the very strong baseline given by Google Translate for the two translation directions. It achieves 15.35 points BLEU, an improvement of around 1 BLEU points over the baseline (14.29) when translating into Chinese, and outperforms the baseline by over 5 BLEU points when translating into Portuguese, achieving 17.79 BLEU points (against the 12.23 of the baseline).

- **Contribution to scientific projects.** The results obtained in this dissertation were also incorporated within the results of two projects: (i) the ASSET (Intelligent Assistance for Everyone Everywhere) project, which aims to improve automatic assistance quality on various languages for the Information Technology domain; and (ii) the CNPTDeepMT-Chinese (Portuguese Deep Machine Translation in eCommerce Domain), which focuses on the improvement of automatic translation between the Portuguese and the Chinese languages.

- **A PT ↔ ZH translation service.** The translation service, for both translation directions, supported by the best MT system I developed is freely available for use online at https://portulanclarin.net/workbench/lx/translator/.

- **A research paper accepted for publication.** Part of the work presented in this dissertation already passed peer review, and was accepted for publication (Santos et al., to appear).

6.3 Future Work

The work presented in this dissertation led to the development of state of the art MT systems for the translation pair Portuguese ↔ Chinese. Nevertheless, research for this pair is far from over, with performance still far from what is achieved for other language pairs.

While acquiring better and larger parallel corpora translation quality is sure to improve this performance, creating new corpora is not in the scope of this study. Instead, this work is concerned with the adoption of techniques that improve said quality with the use of existing data. Accordingly, this Section addresses future research work, that could plausibly further improve the results obtained in this dissertation.

6.3.1 Back-translation

As shown in Section 5.6, back-translation is a useful tool for increasing the training corpus size. Back-translation is a simple method for extending corpora where a previously trained MT system on a chosen pair of languages is used to translate monolingual corpora in one of these languages to the other, hence creating a “synthetic” parallel corpus for those two languages.

Much research has already been dedicated towards improving this method (Sennrich et al., 2016a; Edunov et al., 2018; Imamura et al., 2018), and most authors reached the consensus that a big problem with back-translation is that vocabulary tends to be reduced as these networks usually choose the same output, which diminishes the diversity and richness of the generated translations.

The use of noise during decoding as well as sampling from sub-optimal token predictions have shown to be better options than always choosing the token with highest probability.

The work presented in this dissertation appears as a valid alternative to these methods as the several approaches studied here serve as means for doing translation sampling since every approach outputs different translations arising from their diverse training methods.

6.3.2 Unsupervised Neural Machine Translation

While parallel corpora is difficult to create and to find, and for some language pairs even nonexistent, monolingual data is abundant, ranging from books, magazines, newspapers, thesis and research papers, to easily accessible web scrapings. Accordingly, there is active research on how to use monolingual texts to build MT systems.

Unsupervised Machine Translation tries to answer these questions. While not being able to perform as well as its supervised counterpart, recent work (Artetxe et al., 2019; Lample and Conneau, 2019) has significantly closed the gap. One of these research lines follows the idea of using cross-lingual pre-trained embeddings as anchors for translation, by allowing connections between the words of the two languages.

The use of such techniques could be advantageous for the PT ↔ ZH language pair, since parallel data for the pair is scarce yet monolingual texts are abundant for both languages.

Other than unsupervised Machine Translation, the use of monolingual data in order to improve supervised MT performance is also a sought after topic. Currey et al. (2017) tackle this problem by simply giving monolingual sentences as both source and target together with parallel corpora, and obtain better performance than without the monolingual texts.

6.3.3 Multilingual NMT

As mentioned during the description of the many-to-many approach (cf. Section 4.1.3), the use of several language pairs can improve translation performance for resource poor pairs. This broader idea of using many languages that exchange information that help to improve Machine Translation is referred to as Multilingual MT. Several studies have been conducted on this topic (Dabre et al., 2019), with interlingua representation being a popular objective.

Interlingua representation is the creation of a representation agnostic to any language, yet holding information essential to all languages. This way, by translating to the interlingua representation, one could translate to any desired language.

The idea of a interlingua is not new, being already a research topic in the mid-twentieth century (Richens, 1958). However, the use of rule based approaches lacked the necessary level of robustness.

The use of Neural Networks comes as a great help to the creation of such language, as they have the ability of acquiring knowledge from multiple language sources, and a great generalization capability. One possible research line is the creation of new neural architectures that are capable of producing an interlingua representation, through acquiring knowledge from several languages into a converging component.

By fixing/sharing one component of the neural network, which acts as an intermediary between the encoder and decoder, and alternate between several encoder and decoder components, each for a different language, one could make the network to converge to an interlingua representation on the fixed/shared component. The quality of this interlingua representation would improve with the number of different languages it sees as encoder/source and decoder/target.

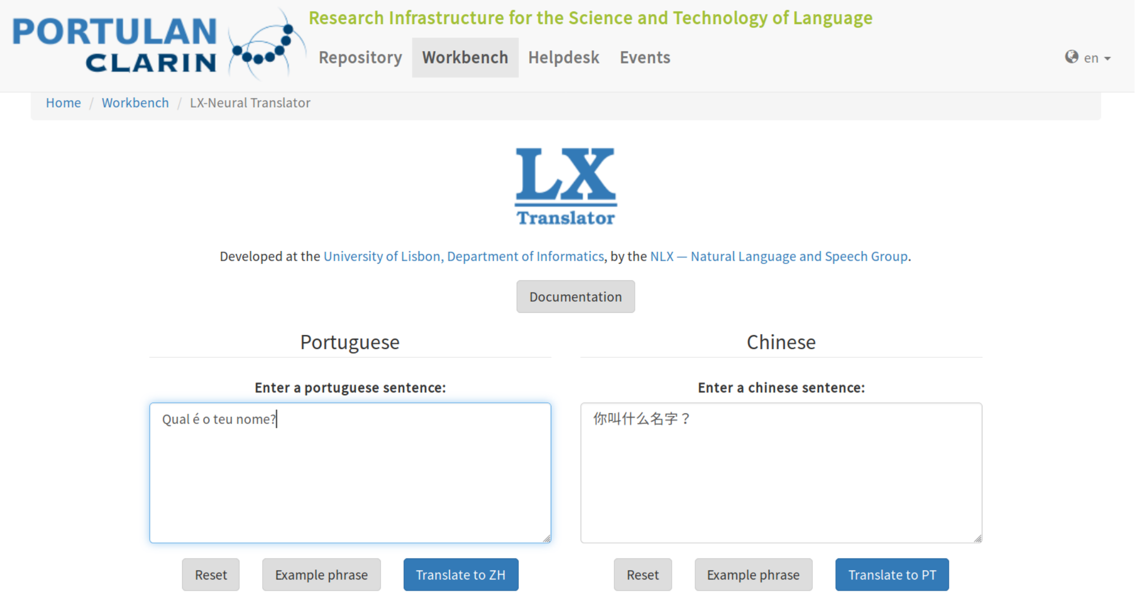
Ideally, by freezing this shared component, it would be possible to train new language pairs not previously seen, as the fixed/shared component would force the new encoder to work towards that interlingua representation and the decoder to generate from it. It may even be possible to train new encoders and decoders only with monolingual corpora.

For our chosen pair this method could improve translation quality, as one major problem faced in this dissertation was the lack of quantity/quality of the parallel corpora available for Chinese and Portuguese.

Appendix A

Online Translation Service

A screenshot of the online translation service freely available online that is supported by the results presented in this dissertation is displayed below.



PORTULAN CLARIN
Research Infrastructure for the Science and Technology of Language
Repository
Workbench
Helpdesk
Events
en
Home / Workbench / LX-Neural Translator
LX Translator

Developed at the University of Lisbon, Department of Informatics, by the NLX — Natural Language and Speech Group.

Enter a portuguese sentence: Qual é o teu nome?

Enter a chinese sentence: 你叫什么名字？

The online service was developed with the help of some members of the NLX Group[1] (Natural Language and Speech Group), and is meant as a simple demonstration of the system.

The user can input Portuguese or Chinese text into the corresponding text box. Upon pressing the button to translate, the text, translated into the other language appears in the other text box.

The LX-Translator can be found at: https://portulanclarin.net/workbench/lx/translator/

[1] My thanks to Luis Gomes, Ruben Branco and Frederico Apolónia for their help in this matter.

Appendix B

Hyper-Parameters for Training

The following hyper-parameters were passed to the Marian framework for training the various approaches:

```
train-sets:
- source.train
- source.train
model: model_pt-zh.npz
type: transformer
max-length: 100
maxi-batch: 1000
early-stopping: 10
valid-freq: 5000
save-freq: 5000
disp-freq: 500
valid-sets:
- source.dev
- target.dev
valid-metrics:
- cross-entropy
- perplexity
valid-mini-batch: 64
mini-batch-fit: true
beam-size: 6
normalize: 0.6
enc-depth: 6
dec-depth: 6
transformer-heads: 8
transformer-postprocess-emb: d
transformer-postprocess: dan
transformer-dropout: 0.1
label-smoothing: 0.1
learn-rate: 0.0003
lr-warmup: 16000
lr-decay-inv-sqrt: 16000
lr-report: true
exponential-smoothing: 0.0001v optimizer-params:
- 0.9
- 0.98
- 1e-09
clip-norm: 5
```

Appendix C

Pivot Approach Error Cases

Under translation - The sentence in red is not translated.

O investimento residencial ainda se situa num nível historicamente baixo como percentagem do PIB, em resultado do excesso de construção durante o período de expansão do mercado imobiliário, 2003-2008, e do subsequente tsunami de execuções hipotecárias. Mas as turbulências no mercado imobiliário estão a dissipar-se.

由于2003-2008年房地产市场扩张期间施工过度以及随后发生的抵押贷款执行海啸，住宅投资在国内生产总值中所占百分比仍然处于历史上的低位。

Wrong translation - The acronym (ELS) is mistranslated.

Na Síria, onde partes consideráveis de território estão já sob controlo Islamita e onde a Frente Al Nusra, pró-Al-Qaeda, ofusca o Exército Livre Sírio (ELS), apoiado pelos EUA, a administração Obama encara a amarga colheita das suas anteriores escolhas políticas. Os ataques aéreos apenas piorariam as coisas, ao reduzir a legitimidade de base do ELS e ajudar as forças Islamitas.

在叙利亚，相当一部分领土已经处于伊斯兰控制之下，亲基地组织的努斯拉阵线在美国的支持下破坏了叙利亚自由军（ERS），奥巴马政府看到了以前政治选择的痛苦收获。空袭只能通过削弱ERS的基本合法性和帮助伊斯兰力量来恶化局势。

Repetition - The phrase "que o" is repeated.

Na verdade, à medida que o aumento do teto da dívida se aproxima, Henry Arão, um pesquisador sênior distinto da Instituição de Brookings, aponta que a Constituição dos EUA exige que o governo dos EUA "gaste o dinheiro que o Congresso lhe deu, para cobrar os impostos que o Congresso lhe autorizou a cobrar, e não emprestar mais do que o que o Congresso autorizou". Se o Congresso se recusar a aumentar o teto da dívida, não será possível cumprir todas as três obrigações legais – mas pode ser o mais difícil.

Bibliography

Aharoni, Roee, Melvin Johnson, and Orhan Firat (2019). Massively multilingual neural machine translation. Available as arXiv preprint arXiv:1903.00089.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2019). An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Bojar, Ondřej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz (2018). Findings of the 2018 Conference on Machine Translation (WMT18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn (2006). Re-evaluation the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Chao, Lidia S., Derek F. Wong, Chi Hong Ao, and Ana Luísa Leal (2018). UM-PCorpus: A large Portuguese-Chinese parallel corpus. In *Proceedings of the LREC 2018 Workshop "Belt & Road: Language Resources and Evaluation"*, pages 38–43.

Cho, Kyunghyun, Bart van Merriënboer Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares Holger Schwenk, and Yoshua Bengio (2014). Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *The 2014 Conference on Empirical MethodsIn Natural Language Processing*.

Currey, Anna, Antonio Valerio Miceli Barone, and Kenneth Heafield (2017). Copied monolingual data improves low-resource neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 148–156.

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

DeFrancis, John, Chia-yee Teng, and Chih-sheng Yung (1969). *Advanced Chinese Reader (Yale Language S)*. Yale University Press.

Doddington, George (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145.

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

Elbayad, Maha, Laurent Besacier, and Jakob Verbeek (2018). Pervasive attention: 2d convolutional neural networks for sequence-to-sequence prediction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 97–107.

Gehring, Jonas, Michael Auli, David Grangier, and Yann Dauphin (2017a). A convolutional encoder model for neural machine translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135.

Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin (2017b). Convolutional sequence to sequence learning. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, pages 1243–1252.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). Long short-term memory. *Neural Computation*, pages 1735–1780.

Hutchins, W John (1995). Machine translation: A brief history. In *Concise history of the language sciences*, pages 431–445.

Imamura, Kenji, Atsushi Fujita, and Eiichiro Sumita (2018). Enhancement of encoder and attention using target monolingual corpora in neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 55–63.

Jean, Sébastien, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio (2015). On using very large target vocabulary for neural machine translation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10.

Johnson, Melvin, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean (2017). Google's multilingual neural machine translation system:

Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, pages 339–351.

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

Koehn, Philipp, Franz Josef Och, and Daniel Marcu (2003). 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*, pages 48–54.

Krizhevsky, Alex, Ilya Sutskever, and G Hinton (2012). Imagenet classification with deep convolutional networks. In *Proceedings of the Conference Neural Information Processing Systems (NIPS)*, pages 1097–1105.

Lakew, Surafel M, Quintino F Lotito, Matteo Negri, Marco Turchi, and Marcello Federico (2017). Improving zero-shot translation of low-resource languages. *14th International Workshop on Spoken Language Translation.*

Lample, Guillaume and Alexis Conneau (2019). Cross-lingual language model pretraining. Available as arXiv preprint arXiv:1901.07291.

Lin, Chin-Yew and Eduard Hovy (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Liu, Siyou, Longyue Wang, and Chao-Hong Liu (2018). 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)*, pages 1485–1492.

Luong, Thang, Hieu Pham, and Christopher D Manning (2015a). Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Luong, Thang, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba (2015b). Addressing the rare word problem in neural machine translation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19.

Norman, Jerry (2003). *The Sino-Tibetan Languages (Routledge Language Family Series)*.

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

Poncelas, Alberto, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban (2018). Investigating backtranslation in neural machine translation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation: 28-30 May 2018, Universitat d'Alacant, Alacant, Spain*, pages 249–258.

Reto, Luís, Fernando Machado, and Jose Esperanca (2016). *Novo Atlas da Língua Portuguesa/New Atlas of the Portuguese Language*.

Richens, R. H. (1958). Interlingual Machine Translation. *The Computer Journal*, pages 144–147.

Rumelhart, David E, Geoffrey E Hinton, and Ronald J Williams (1986). Learning representations by back-propagating errors. *Nature*, pages 533–536.

Santos, Rodrigo, João Silva, António Branco, and Deyi Xiong (to appear). The Direct Path May Not Be The Best: Portuguese — Chinese Neural Machine Translation. In *EPIA 2019 Conference on Artificial Intelligence*.

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

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

Simons, Gary F., editor (2019). Ethnologue: Languages of the world - 22nd edition. Internet, 24/07/2019 - http://www.ethnologue.com.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Tian, Liang, Derek F. Wong, Lidia S. Chao, Paulo Quaresma, Francisco Oliveira, and Lu Yi (2014). Um-corpus: A large English-Chinese parallel corpus for statistical machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1837–1842.

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

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wikimedia Commons (2018). Map of the portuguese language in the world.svg. [Online; accessed 21-May-2019].

Wikimedia Commons (2019). Map of sinitic languages full-en.svg. [Online; accessed 21-May-2019].

Wong, Fai Aliás Wong Hway (2001). Automatic translation: Overcome the barriers between european and chinese languages.

Wong, Fai and Sam Chao (2010). PCT: Portuguese-Chinese machine translation systems. *Journal of translation studies*, pages 181–196.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. Available as arXiv preprint arXiv:1609.08144.