UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

FELIPE SOARES

Machine Translation for the biomedical domain, corpora acquisition and translation experiments

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Prof. Dr. Karin Becker

Porto Alegre
July 2019

CIP — CATALOGING-IN-PUBLICATION

Soares, Felipe

Machine Translation for the biomedical domain, corpora acquisition and translation experiments / Felipe Soares. – Porto Alegre: PPGC da UFRGS, 2019.

65 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2019. Advisor: Karin Becker.

1. Scientific texts. 2. Biomedical domain. 3. Corpora acquisition. 4. Statistical Machine Translation. 5. Neural Machine Translation. I. Becker, Karin. II. Título.

“We should take astrology seriously.
No, I don’t mean we should believe in it.
I am talking about fighting it seriously
instead of humouring it as a piece of harmless fun.”
— Richard Dawkinson

AGRADECIMENTOS

Agradeço ao CNPq pelo financiamento recebido na forma de bolsa de mestrado pela duração de 1 ano.

ABSTRACT

Availability of biomedical documents in more than one language (e.g. not just in English) can broader the access to information and help patients and practitioners to keep up to date with the recent advances in biomedicine. In this work, we are interested in using machine translation to translate Spanish and Portuguese biomedical scientific texts to English, and vice-versa. We also present the development of three parallel corpora for scientific texts in the biomedical domain in English, Portuguese and Spanish. Our developed corpora are larger than the already available ones for this domain and languages. Regarding translation experiments, to create our training data, we concatenated several parallel corpora, both from in-domain and out-of-domain sources, as well as terminological resources from UMLS. We validated our approaches by participating in the biomedical translation track of the shared task at WMT conference. Our systems are based on statistical machine translation and neural machine translation, using the Moses and OpenNMT toolkits, respectively. We carried out experiments in four translation directions for the English/Spanish and English/Portuguese language pairs. Our systems achieved the best BLEU scores according to the official shared task evaluation.

Keywords: Scientific texts. Biomedical domain. Corpora acquisition. Statistical Machine Translation. Neural Machine Translation.

Tradução automática para o domínio biomédico: aquisição de corpora e experimentos de tradução

RESUMO

A disponibilidade de documentos biomédicos em mais de um idioma (por exemplo, não apenas em inglês) pode ampliar o acesso à informação e ajudar os pacientes e profissionais a se manterem atualizados sobre os recentes avanços na biomedicina. Neste trabalho, estamos interessados em usar a tradução automática para traduzir textos científicos biomédicos em espanhol e português para o inglês, e vice-versa. Também apresentamos o desenvolvimento de três corpora paralelos para textos científicos no domínio biomédico em inglês, português e espanhol. Nossos corpora desenvolvidos são maiores que os já disponíveis para este domínio e idiomas. Com relação aos experimentos de tradução, para criar nossos dados de treinamento, concatenamos vários corpora paralelos, tanto de fontes de domínio, quanto fora do domínio, bem como recursos terminológicos do UMLS. Nós validamos nossas abordagens participando da shared task de tradução biomédica da conferência WMT. Nossos sistemas são baseados em tradução automática estatística e tradução automática neural, e foram desenvolvidos usando os toolkits Moses e OpenNMT, respectivamente. Participamos de quatro direções de tradução para os pares de idiomas inglês/espanhol e inglês/português. Nossos sistemas alcançaram as melhores pontuações BLEU de acordo com a avaliação oficial da shared task.

Palavras-chave: Textos científicos, domínio biomédico, aquisição de corpora, tradução automática estatística, tradução automática neural.

LIST OF ABBREVIATIONS AND ACRONYMS

SMT	Statistical Machine Translation
NMT	Neural Machine Translation
TDC	Theses and Dissertations Catalog
NLP	Natural Language Processing
LM	Language Model
WMT	Conference on Machine Translation
MT	Machine Translation
BVS	*Biblioteca Virtual em Saúde*
CAPES	*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*
BIREME	*Biblioteca Regional de Medicina*
CSV	Comma Separated Value
OPAS	Pan American Health Organization
NLM	National Library of Medicine
UMLS	Unified Medical Language System
RBMT	Rule-based machine translation
PBSMT	Phrase-based statistical machine translation
CNN	Convolutional Neural Network
RNN	Recursive Neural Network
LSTM	Long Short-Term Memory
UPC	Technical University of Catalunya
UHH	University of Hamburg

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

1.1 Motivation

The availability of biomedical documents in more than one language (e.g. not just in English) can broader the access to information and help patients and practitioners to keep up to date with the recent advances in biomedicine. As stated by Lazarev and Nazarovets (2018) in an article published in Nature, one cannot dismiss papers written in languages other than English. In another correspondence in Nature, Prieto (2018) stresses the need to have multilingual research-paper databases: "It is absurd to put effort and public resources into research that has already been published. This will continue to be a risk as long as papers in non-English journals are not routinely indexed in the international databases".

Prieto (2018) suggests that the scientific community needs to develop a comprehensive multi-language translation tool to enable international researchers to access regional databases not compiled in English. This would allow them to find all the details about the study that may not be available in abstracts, such as detailed experimental design and results. Lazarev and Nazarovets (2018) also point out that in some countries, works can be accessed through automated translation services. In light of that, we see that the development of an automatic translation tool for biomedical articles is an important contribution to the scientific community.

Machine Translation (MT) is concerned on, provided an input sentence, translate it to one or more target languages, usually without the need of human input or action (HUTCHINS, 1995). MT systems can be used in a plethora of applications and with multiple goals. Currently, several providers have generic-domain tools aiming at translation, such as Google Translate, Microsoft Bing Translate, Amazon Translate. However, these services are free up to a certain point. After a determined number of characters, one has to pay to be able to translate the texts. In addition, they are not tailored for specific domains, which would also require payment by the user. In this work, we are interested in using MT to translate Spanish and Portuguese biomedical scientific texts to English, and vice-versa. This is very important to the biomedical field due to the abundance of state-of-the-art articles in English, which can be a barrier to native speakers of Portuguese and Spanish that do not have the necessary skills to read them in English. Past competitions in biomedical machine translation (e.g. WMT Biomedical track) have tried to fill this gap,

but there is still room for improvement.

Historically, MT systems have been built based on two approaches: following a pre-defined set of rules defined by a linguist, which is called rule-based MT, or taking advantage of texts written in the source and target languages (i.e. parallel corpora) to learn translation patterns, called corpus-based MT (TERUMASA, 2007). The main disadvantage of the former is that a high number of rules are required to produce quality translations, thus more human effort needed. The later has been the standard approach in the last years, since no hard-coded and human made rules are needed. To ensure high quality corpus-based MT systems, large amounts of parallel sentences are required (NAIR; PETER, 2012). However, acquiring parallel corpora that is large enough to train MT systems is not a trivial task (RESNIK; SMITH, 2003), specially when one is interested in specific domains, which is our case.

Systems trained on generic texts (e.g. speeches from the European Parliament or movies subtitles) may fail to provide good translations for texts from other sources, such as scientific texts. Since corpus-based MT systems learn translation patterns from data, it is essential for in-domain models to be trained on as much domain-related data as possible, aiming at capturing information regarding its domain-specific terms (KOEHN; SCHROEDER, 2007).

In addition to collect adequate and good quality parallel corpora for in-domain corpus-based MT, one has also to decide which MT paradigm will be used to train such system. Currently, there are two main approaches for MT modelling: Statistical Machine Translation (SMT), and Neural Machine Translation (NMT) (DOWLING et al., 2018). The former is already established and dates back to the 1990's (BROWN et al., 1990), while the latter has been increasingly explored in the past 5 years (VASWANI et al., 2018), providing several advantages over the traditional SMT, such as improved fluency and adequacy, and greater performance regarding the number of translated words per second. In a few words, SMT relies on linear estimations and tables of translation probabilities, while NTM models usually encode words in a vector space and then decode them in the target language.

In earlier works addressing automatic translation of biomedical texts, such as those from Neves, Yepes and Névéol (2016) and Wu et al. (2011), they have tried to compile in-domain parallel biomedical data and train SMT systems to perform automatic translation of articles. In the work of Yepes, Prieur-Gaston and Névéol (2013), they compared the use of general-domain (Google News) data and in-domain data (Pubmed) for the same

task. However, none of them have tried to combine specific general-domain data, or data from other domains, with biomedical data to enhance translation quality.

1.2 Hypotheses

The first hypothesis we investigate is that the concatenation of partially in-domain data (i.e. scientific texts and medicines documents), with in-domain (i.e. biomedical scientific texts), and out-of-domain data with similar textual structure (i.e. books, legislative texts), and terminological resources (i.e. biomedical controlled vocabulary) can provide enough information for the accurate translation of sentences from biomedical scientific texts. In a brief, our training set would be comprised of: partially in-domain data + in-domain data + out-of-domain data + terminological resources. This is an attempt to perform domain adaptation directly at the first training step. We consider the following four translation directions: English→Portuguese, Portuguese→English, English→Spanish, Spanish→English.

Given that the acquisition of large parallel corpora and their concatenation with other similar corpora will increase information availability, NMT systems should benefit from such resources, as stated in Belinkov and Glass (2016), NMT tends to present higher accuracy when the training corpus ir large. Thus, the second hypothesis we investigate is that NMT systems can outperform SMT systems for the the aforementioned translation directions and biomedical domain.

To verify our hypotheses, we have built SMT and NMT systems on the aforementioned data and tested their translation quality regarding the usual BLEU metric. Aiming at providing unbiased results, we submitted our models to the Third Conference on Machine Translation (WMT18) shared task on biomedical translation[1], and used the official evaluation as results for our study. Other participants in the same shared task provided systems trained only on in-domain data, which are used as our baseline of comparison for the combination of both data sources.

[1] <http://www.statmt.org/wmt18/biomedical-translation-task.html>

1.3 Objectives

With this work, we have the following primary objective: Verify how the combination of in-domain and general domain texts, in a large corpus, can lead to an state-of-the-art performance in machine translation for the following languages::

- Portuguese to English,
- Spanish to English,
- English to Portuguese,
- English to Spanish.

To accomplish such primary objective, we also need to accomplish the secondary objectives:

- Develop larger parallel corpora for the mentioned languages;
- Develop models to compare Statistic Machine Translation to Neural Machine Translation for this in-domain task.

1.4 Contributions

The main contributions of this work are:

- Development of a parallel corpus of full-text scientific articles from SciELO[2] database in English, Portuguese, and Spanish.
- Development of a parallel corpus of abstracts from the CAPES[3] repository of theses and dissertations in English and Portuguese.
- Development of a parallel corpus of abstracts from BVSalud[4] database in English, Portuguese, and Spanish.
- Training of translation models using SMT and NMT approaches for the aforementioned data and terminological resources, making the NMT final models available for public use. These models outperformed state-of-the-arth models, as well as the SMT ones.

  All contributions can be accessed at: <www.felipesoares.net>.

[2] <http://www.scielo.org>
[3] <https://catalogodeteses.capes.gov.br>
[4] <http://bvsalud.org>

1.4.1 Publications

During the development of the present work, three main articles were produced:

- SOARES, F.; MOREIRA, V; BECKER, K. A Large Parallel Corpus of Full-Text Scientific Articles. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan, 2018. CAPES Qualis: A1;
- SOARES, F.; BECKER, K. UFRGS Participation on the WMT Biomedical Translation Shared Task. In: **Proceedings of the Third Conference on Machine Translation: Shared Task Papers**. Bruxelas, Belgica, 2018. CAPES Qualis: Não Classificado;
- SOARES, F.; YAMASHITA, G.H.; ANZANELLO, M.J. A Parallel Corpus of Theses and Dissertations Abstracts. In: VILLAVICENCIO, Aline *et al*. **Computational Processing of the Portuguese Language - PROPOR 2018. Lecture Notes in Computer Science**. Cambridge: Springer,2018. CAPES Qualis: B3.

The following article is not directly related to this research, but was produced during the masters course:

- SOARES, F.; BECKER, K.; ANZANELLO, M.J. A hierarchical classifier based on human blood plasma fluorescence for non-invasive colorectal cancer screening. **Artificial Intelligence in Medicine**, v. 82, p. 1-10, 2017. CAPES Qualis: A2.

1.5 Organization

This work is organized as follows. In Chapter 2, we draw the theoretical foundations of our methods and experiments, focusing on multiligual corpora, machine translation and databases. In Chapter 3, we present the related work already done in the field, with focus on development of parallel corpora of scientific texts and biomedical machine translation. In Chapter 4, we detail the creation of our parallel corpora and their evaluation. In Chapter 5, we detail our experiments and results regarding Machine Translation, with focus on SMT and NMT and the pre-processing steps. Finally, in Chapter 6, we derive our conclusions and point out directions of possible future works.

2 THEORETICAL FOUNDATIONS

In this chapter, we present the basic theoretical foundations for the comprehension of the work presented in this dissertation. The following sections introduce concepts of natural language processing and machine translation.

2.1 Multilingual Corpora

Corpora in more than one language, usually called multilingual corpora, can be essentially of two types: parallel and comparable (MCENERY; ZHONGHUA, 2007). A comparable corpus has texts segments in two or more languages that are not a translation of each other, generally referring to the same domain (e.g. two versions of the same article in Wikipedia, one in English and the other in Spanish). A parallel corpus has the original segments in one language and their translations in one or more languages, which are aligned (e.g. the original Harry Potter in English and its translation to Portuguese and Greek, aligned according sentence or paragraph).

Parallel corpora have been used for a wide range of applications, such as: machine translation (NEVES et al., 2018; YEPES; PRIEUR-GASTON; NÉVÉOL, 2013), extraction of bilingual-terminologies (GUINOVART; SIMOES, 2009), creation of cross-language word embeddings (AMMAR et al., 2016), cross-language plagiarism detection (POTTHAST et al., 2011), cross-language information retrieval (PECINA et al., 2014). In addition to the mentioned computational uses, parallel corpora can be used in teaching translation for students (WANG; QIN; WANG, 2007), or to find patterns in the human translation process (MAURANEN; KUJAMÄKI, 2004).

In the case of corpus-based MT, parallel corpora play an essential role. As studied previously (CALLISON-BURCH; KOEHN; OSBORNE, 2006), the quality of translation improves with the size of the corpora used for training, thus, the larger the corpus, usually the better the translation is.

The acquisition of parallel corpora is not a trivial task, as it may demand considerable use of expert human curating. After collecting the texts in the desired languages, one has to first segment the text in small chunks of sentences or individual sentences, and then align them such that the sentence in one language has its counterpart in another language. For that instance, many segmenters and aligners have been proposed in the

literature. Some of the segmenters are language specific, such as Segtok[1], which focuses on Indo-European languages (i.e. Spanish, English, and German), and the one proposed by Sugisaki (2018) for German, while others are language-independent, such as PUNKT (KISS; STRUNK, 2006).

Many parallel corpora are already available, some with bilingual alignment, while others are multilingually aligned, with 3 or more languages, such as Europarl (KOEHN, 2005), from proceedings of the European Parliament, JRC-Acquis (STEINBERGER et al., 2006), from the European Commission; OpenSubtitles (ZHANG; LING; DYER, 2014), from movies subtitles.

For the creation of parallel corpus, one of the most important steps is the sentence alignment, which we will now discuss and present the system we used for sentence alignment in our studies.

2.1.1 Sentence alignment

Sentence alignment is the process of, given a corpus in two languages, finding the counterpart of one source sentence in the target corpus. One of the most used algorithms for sentence alignment is known as Gale-Church, since it was proposed by Gale and Church (1993). It is based on the principle that equivalent sentences in two or more languages should roughly have the same length. A probabilistic score is assigned to each proposed correspondence of sentences, and then a dynamic programming framework is used to find the maximum likelihood alignment of sentences.

A more recent algorithm, Hunalign (VARGA et al., 2007), uses Gale-Church sentence-length information to first automatically build a parallel dictionary based on this alignment. Once the dictionary is built, the algorithm realigns the input text in a second iteration, this time combining sentence-length information with the dictionary. When a dictionary is supplied to the algorithm, the first step is skipped. A drawback of Hunalign is that it is not designed to handle large corpora (above 10 thousand sentences), since it consumes large amounts of memory. In these cases, the algorithm cuts the large corpus in smaller manageable chunks, which may affect dictionary building.

In this work, we use Hunalign since it has proven to be a robust algorithm in several related works (STEINBERGER et al., 2006; YU; MAX; YVON, 2012; ABDUL-RAUF et al., 2010). In addition, we used the score provided by Hunalign, which is the

[1] <https://github.com/fnl/segtok>

combination of sentence-length score and dictionary score, to filter out very discrepant alignments. For such, we set as threshold an alignment score of 75%. Sentence pairs with scores below that point were discarded.

2.2 Machine Translation

Machine Translation is the automation of some or all the processes of translating sentences from one language (source) to another (target) (KOEHN, 2009). Translation itself is not an easy task, even for human translators (e.g. human A can translate a sentence differently from human B, without altering its meaning). Some of the difficulties are highlighted by Rebechi and Andreetto (2015), which compares the different translations of the book *Trauer und Melancholie*, from Freud, to Portuguese by the corpus linguistics perspective. Most of the difficulties arise from the differences in the source and target language structures, or the vocabulary used in an specific domain. In addition, polysemy (i.e. the existence of more than one possible meaning for a given word) presents an additional challenge for automatic translation.

Bird, Klein and Loper (2009) pointed out that MT methods can be divided in two main types: rationalists and empiricists. The most relevant example of rationalism method is the rule-based MT (RBMT), which uses information derived by linguists about the source and target language, using several manual developed rules to translate the given text. On the other hand, SMT methods are included in the empiricist methods, since they rely on corpora and past examples rather than human tailored rules. SMT methods are more robust regarding small training corpora and can provide fluent translations (XUAN; LI; TANG, 2012). NMT methods are also included in the empiricist category, and have been of great interest in the past few years, achieving remarkable results in translation quality (BRITZ et al., 2017a).

In the remaining of this section we detail the SMT and NMT approaches used in this work.

2.2.1 Statistical Machine Translation

Many different models of SMT have been proposed in the past decades (LOPEZ, 2008), but in this dissertation we will focus on phrase-based statistical machine translation

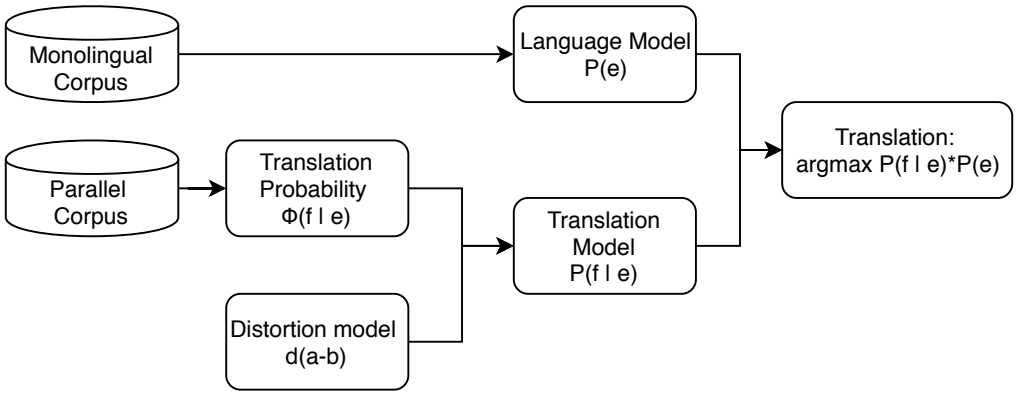
(PB-SMT). Zens, Och and Ney (2002) give us a concise definition of PB-SMT.

Given that the goal of MT is to transfer the meaning of a source language sentence f_1 segmented in J phrases ${f_1}^{\mathrm{J}} = f_1, ..., f_j, ..., f_J$, into a target language sentence ${e_1}^{\mathrm{I}} = e_1, ..., e_i, ..., f_I$, the conditional probability $Pr({e_1}^{\mathrm{I}}|{f_1}^{\mathrm{J}})$ is used to describe the correspondence between the two sentences. The translation problem can be formulated as a maximization problem after the use of Bayes rule, where $p(e)$ denotes the language model, and $p(f|e)$ denotes the translation model:

$$argmax_e p(e|f) = argmax_e p(f|e)p(e) \tag{2.1}$$

The language model (LM) describes the correctness of the target language sentence, aiming at avoid syntactically incorrect sentences. The translation model (TM) gives the likelihood that the source sentence and the candidate translation are equivalent. One could see the TM as the adequacy of a choice of translation, while the LM as the fluency of that given translation (HEARNE; WAY, 2011). In Figure 2.1, we show a graphical representation of a traditional SMT model and its source of information, such as monolingual an parallel corpora.

Figure 2.1: Schematic representation of a traditional PB-SMT model



Source: Adapted from (ZENS; OCH; NEY, 2002)

The LM $p(e)$ gives the likelihood that a string e is a valid and coherent sentence in the target language. The language model provides two main components: (i) a model of the monolingual training corpus, and (ii) a method for computing the probability of a previously seen, or unseen, string e using that model (HEARNE; WAY, 2011).

From the statistical point of view, an LM is a joint probability distribution of a sequence of words. One of the most traditional ways of constructing an LM is to use count-based probabilities on n-grams. Considering the Markov assumption (HUNNI-

CUTT; CARLBERGER, 2001), the process of predicting a word sequence is broken down into predicting one word at a time. Then, the probability of $P(w_1, w_2, w_3)$ is a product of word probabilities considering the preceding words, limited to m words (GOODMAN, 2001):

$$P(w_n|w_1, w_2, ..., w_{n-1}) \approx P(w_n|w_{n-m}, ..., w_{n-2}, w_{n-1}) \tag{2.2}$$

Considering a tri-gram model, the probability of a word w_i given w_{i-2} and w_{i-1} can be approximated by the number of occurrences:

$$P(w_i|w_{i-2}, w_{i-1}) \approx \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})} \tag{2.3}$$

During decoding (i.e. translation), the input sentence is segmented into a sequence of smaller phrases, which are individually translated. Each foreign phrase f is translated into an e phrase, with translation being modeled by a probability distribution $\phi(f|e)$. Since output phrases e may be reordered, this phenomenon is modeled by a relative distortion probability distribution $d(a_1 - b_{i-1})$, where a_i is the start position of the input phrase and b_i is the end position of the foreign input phrase translated to the target language (KOEHN; OCH; MARCU, 2003). Then, the translation probability $p(f|e)$ can be decomposed as:

$$p(f|e) = \prod \phi(f|e)d(a_1 - b_{i-1}) \tag{2.4}$$

To calculate the $\phi(f|e)$ probability, a word-based alignment is usually performed from parallel corpora.

From this brief review, we can see that SMT models are highly dependant of good LMs and based on linear assumptions for model building. Thus, SMT is somewhat "hard-coded" to follow a specific model and translation probability tables.

In this work we will use both SMT and NMT systems to build biomedical MT systems, which we will present in the following section.

2.2.2 Neural Machine Translation

Neural Machine Translation is a relatively new approach for automatic translation, and has led to significant improvements over SMT, especially regarding human evaluation

(Klein et al., 2017), when coherence and fluency is better evaluated, not only the percentage of correct words and sentence length. One of the first appearances of NMT was in the work of Kalchbrenner and Blunsom (2013), where they describe a new encoder-decoder architecture for machine translation. This model encodes, through an encoder, a source text into a continuous vector space and uses a decoder to transform the vectors into the target language.

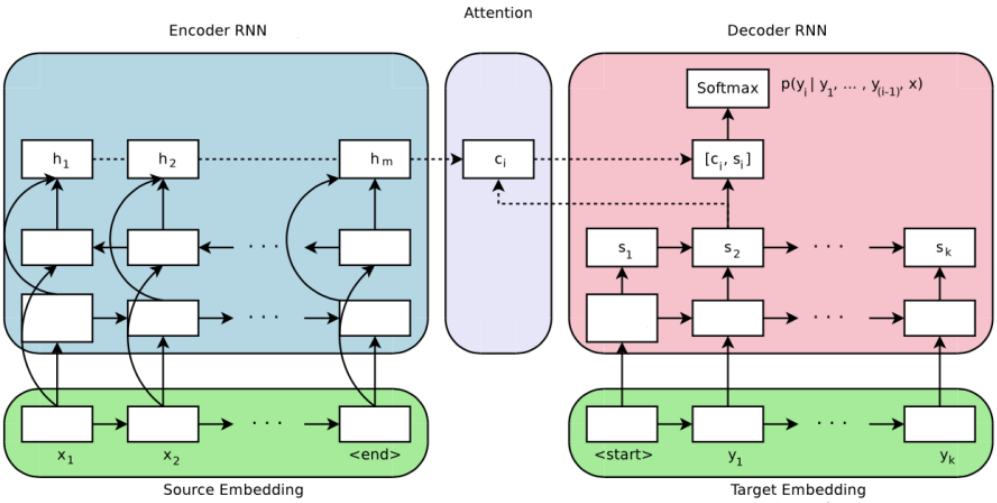
By introducing a gated Long Short-Term Memory (LSTM) (CHO et al., 2014), Sutskever, Vinyals and Le (2014) presented a new approach for neural translation, known as sequence to sequence (seq2seq) learning. LSTMs are used both in the encoder and decoder to learn the vector representation of sentences and perform their decoding back to sentences. LSTMs can be seen as neural cells able to "forget" information that is not necessary to the task in hand, while keeping the relevant information. This is achieved by the combination of an input gate, an output gate, and a forget gate. The input gate controls the extent to which a new value flows into the cell, while the output controls the amount of the value in the cell that is passed to the output activation function. Inside the cell, the forget gate controls the amount of information which will remain (GERS; SCHMIDHUBER; CUMMINS, 1999). At that point, most approaches of deep learning were only used to re-score SMT outputs, not doing the complete translation.

In Figure 2.2, we show the representation of a sequence to sequence network with attention, which will be explain later. We can see that the decoder is analogous to the language model in SMT, being responsible for the correctness of the sentence in the target language, while also accounting for the alignment.

The attention enables the network to give importance only to relevant parts of the input during prediction. Thus, when the decoder is producing a word to form the target sentence, only a specific part of the input sentence is relevant and taken into account. This causes the target sentence to be predicted based on context vectors, instead of a fixed-length vector, which leads to better translations. With the introduction of the attention mechanism (BAHDANAU; CHO; BENGIO, 2014), the use of NMT for the whole process of MT started to grow.

In Figure 2.3, we show the representation of a bidirectional LSTM encoder. One can see that each word in the source language ($m_i, ..., m_j$) is fed to two units. One unit is a forward LSTM, while the other one is backwards, thus giving the name of bidirectional LSTM. The reason for superior results in translation is that when using a forward encoder, the only information given to predict the next word is about the past. Similarly, in a

Figure 2.2: Representation of a seq2seq neural network with attention



Source: adapted from Britz et al. (2017b)

backward encoder, the only information is about the future. In a bidirectional LSTM, where the sentence is read forwards and backwards, the information is about both the future and the past, leading to more information and context about the input and greater prediction quality.

We can see that NMT models are more flexible than SMT ones regarding structure, allowing non-linear relationships. In addition, the fact that they do not rely on translation probability tables, and the inclusion of attention mechanisms, may make them more suitable for the translation of more complicated sentences, such as the ones in scientific writing.

In light of that, in this work we make use of a bidirectional LSTM network with attention to develop the translation models and to compare with the traditional SMT ones.

2.3 Automatic Evaluation Metrics

We now present the most usual automatic evaluation metrics for MT

2.3.1 BLEU

BLEU is a metric for the purpose of evaluating machine translation systems with greater economy, speed and independence of languages than manual evaluations. It is

based on the proximity between machine translation and translation performed by a qualified person in professional translations, called a reference translation (PAPINENI et al., 2002).

The BLEU metric evaluates automatic translations through the precision of n-letter or word sequences, called n-grams. The precision of n-grams indicates the number of n-grams compatible between the sentence to be evaluated, called the candidate sentence, and the equivalent reference sentence, dividing this number by the total words of the candidate sentence. This compatibility must be accounted for only once.

When the candidate sentence is greater than the corresponding reference, a penalty for brevity is calculated, as shown in equation 2.5, since the candidate sentence must be similar to the reference in size, choice and word order. The penalty for brevity is calculated on the whole corpus and not on the sentences. In equation 2.5, r is the number of words in the reference text and c is the number of words in the candidate text (MELO; MATOS; DIAS, 2015).

$$BP = \begin{cases} 1 & c > r \\ exp(1 - \frac{r}{c}) & c = r \end{cases} \tag{2.5}$$

Finally, to calculate the BLEU metric, the geometric mean of the modified precision is calculated by multiplying it by the penalty factor for brevity, as shown in equation 2.6.

$$BLEU = BP * exp(\sum_{n=1}^{N} w_n log(p_n)) \tag{2.6}$$

2.3.2 NIST

The NIST metric, as well as the BLEU, is also based on n-grams precision (varying in this case from 1 to 5), but it uses the arithmetic mean of n-grams rather than the geometric mean as it does in BLEU. Another difference between these two metrics is that in NIST, n-grams are weighted by weights according to the information they provide rather than simply being counted as in BLEU (CASELI, 2004).

NIST represents the average information, per word, given by the n-grams in the candidate that match an n-gram of one of the references in the set of references. NIST's brevity penalty (BP), in relation to BP of BLEU, penalizes more seriously the very small

candidates and less the candidates closest to the references, in size (CASELI, 2004; WOŁK; KORŽINEK, 2016).

2.3.3 METEOR

METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric for assessing the quality of machine translation . The metric is based on the use of n-grams and is focused on the use of statistical and accurate evaluation of the source text. Unlike the BLEU metric, this metric uses synonym matching functions along with exact word matching. The metric was developed to solve the problems that were found in the more popular BLEU metric, as well as to create a good correlation with the experts' assessment at the level of phrases or sentences (LAVIE; DENKOWSKI, 2009).

As a result of running the metrics at the level of phrases, the correlation with the human solution was 0.964, while the BLEU metric was 0.817 on the same input data set. At the proposal level, the maximum correlation with expert evaluation was 0.403 (BANERJEE; LAVIE, 2005).

As in the BLEU metric, the basic unit for evaluation is the sentence, the algorithm first performs text alignment between the two sentences, the reference translation line and the input text line for evaluation (see Figures a and b). This metric uses several steps to establish the correspondence between the words machine translation and reference translation for matching two strings(LAVIE; DENKOWSKI, 2009):

1. Exact matchmaking — strings that are identical in reference and machine translation are determined;
2. Establishing the correspondence of the basics - the stemming is carried out (highlighting the stem of the word), and words with the same root in the reference and machine translation are determined;
3. Matching synonyms - words that are synonymous in accordance with WordNet are defined.

Alignment is the set of correspondences between n-grams. The stages of comparison with reference translations are performed sequentially, and at each of them only those n-grams that were not matched at the previous stages are added to the set of matches. As soon as the last stage is completed, the final n-gram P is calculated using the following

equation:

$$P = \frac{m}{w_t} \tag{2.7}$$

Where m the number of n-grams in machine translation, which were also found in the reference translation, and w_t the number of n-grams in machine translation. N-gram R (total n-gram for reference translations) is calculated by the following formula:

$$R = \frac{m}{w_r} \tag{2.8}$$

Where w_r the number of n-grams in the reference translation. The following formula is used to determine the harmonic mean of the translation:

$$F_{mean} = \frac{10PR}{R + 9P} \tag{2.9}$$

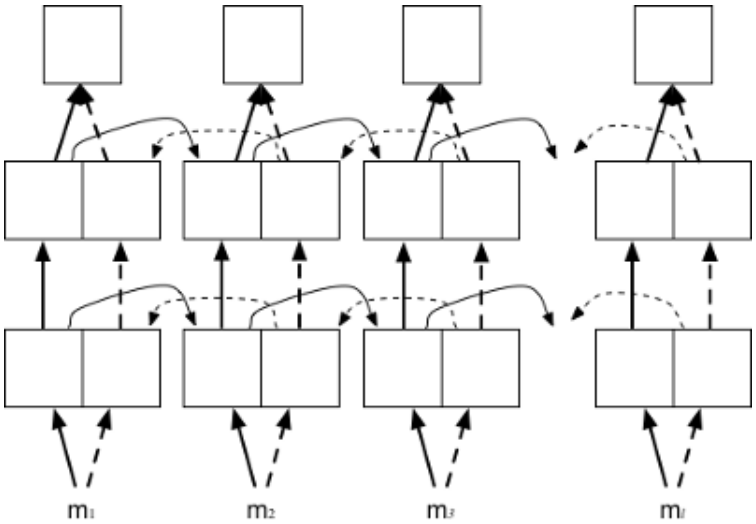
This formula is used only for comparing single words that are matched in the reference and machine translation. In order to take into account also phrases that match, the so-called penalty p is used. For this, n-gram is combined into several possible groups. Fine grained p is calculated by the following formula:

$$p = 0.5 \left(\frac{c}{u_m} \right)^3 \tag{2.10}$$

Where c is the number of n-gram groups, and u_m the number of n-grams, which are combined into groups. Then the final quality indicator is calculated by the following formula:

$$M = F_{mean}(1 - p) \tag{2.11}$$

Figure 2.3: Representation of a bidirectional LSTM network



Source: <http://opennmt.net/OpenNMT/training/models/>

3 RELATED WORK

3.1 Parallel corpora of scientific texts

The development of parallel corpora from scientific texts has been researched by several authors, aiming at translation of biomedical articles (WU et al., 2011; NEVES; YEPES; NéVéOL, 2016), or named entity recognition of biomedical concepts (KORS et al., 2015). Regarding Portuguese/English and English/Spanish, the FAPESP corpus (AZIZ; SPECIA, 2011), from the Brazilian magazine *revista pesquisa FAPESP*[1], contains more than 150,000 aligned sentences per language pair, constituting an important language resource.

Wu et al. (2011) constructed a parallel corpus of biomedical article titles from PubMed in English and other six languages (i.e. French, Spanish, German, Hungarian, Turkish, and Polish). They downloaded PubMed XMLs and used regular expressions to find the <ArticleTitle>, and <VernacularTitle> tags, which correspond to the title in English and the other language, respectively. In Table 3.1 we show the statics of the aforementioned corpus.

Table 3.1: Wu et al. (2011) corpus statistics

Language	**Sentences**
French	555,058
Hungarian	32,937
Polish	152,327
Spanish	233,881
German	706,258
Turkish	6,665

The work of Neves, Yepes and Névéol (2016), which is one of the main data sources for biomedical translation, uses SciELO as the source of parallel data. They developed a crawler to retrieve articles in SciELO listed in the "Biological Sciences" and "Health Sciences" subjects. After crawling the website, the titles and abstracts were stored and indexed in SAP HANA database, which was also used for language detection and sentence splitting. Later, sentences were automatically aligned using the Geometric Mapping and Alignment (GMA)[2] tool, where titles and abstracts were supplied separately. In addition, they also carried out manual evaluation of the sentence alignment quality.

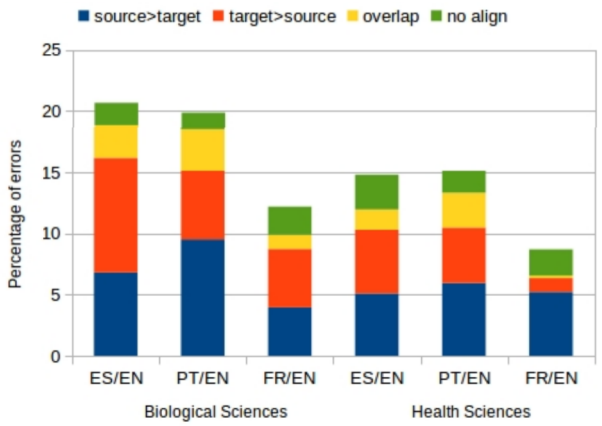
[1] <http://revistapesquisa.fapesp.br/>
[2] <https://nlp.cs.nyu.edu/GMA/>

Table 3.2: Neves, Yepes and Névéol (2016) corpus statistics

Language Pair	Language	Sentences
EN/ES	EN	767,039
	ES	735,125
EN/PT	EN	669,629
	PT	651,438
EN/FR	EN	9,393
	FR	9,501

They defined five alignment quality categories: (a) "OK", when correctly aligned, (b) "Source>Target" when correctly aligned, but with source containing more information, (c) "Targe>Source" when correctly aligned, but with target containing more information, (d) "Overlap", when there is an overlap in the information content of both sentences but they cannot be considered aligned, and (e) "No alignment", when sentences are unrelated. In Table 3.2 we show the corpus statistics, and in Figure 3.1 we show the alignment quality reported by the authors.

Figure 3.1: Percentage of the errors in the quality assessment of the SciELO corpus



Source: Neves, Yepes and Névéol (2016)

While for other translation tasks the availability of parallel data is relatively high, such as legal texts (e.g Europarl with 2 million sentences), and subtitles (with around 35 million sentences[3], the biomedical scientific writing is under-resourced. As shown in the related works, none of them can even achieve 1 million sentences.

[3]<http://opus.nlpl.eu>

3.2 Biomedical machine translation

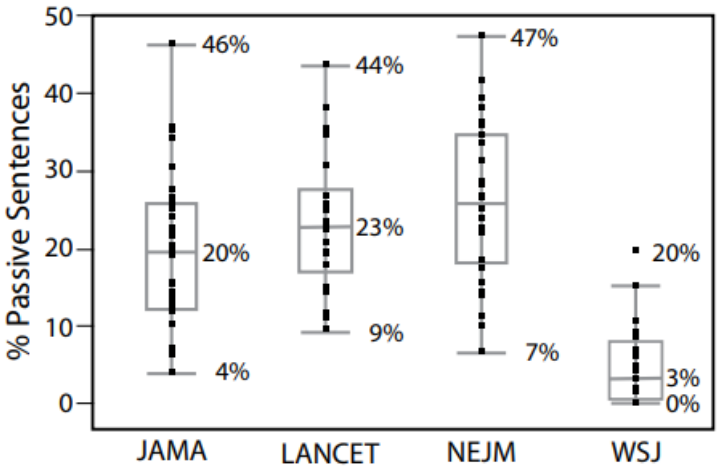
Translation of biomedical texts is a special case of in-domain machine translation and imposes some additional challenges to the case of general domain systems, being the following the most important ones:

- Long tail effect: The biomedical domain has a very long tail regarding specific terms and vocabulary, which means that a lot of words appear very few times, making the recognition of translation patterns more difficult (MOEN; ANANIADOU, 2013).
- Use of passive voice in medical texts: This is an interesting phenomena that can happen for certain languages and specific domains. A study by Amdur, Kirwan and Morris (2010) showed that the use of passive voice in medical journals is much higher than in general domain texts. They studied four medical journals and compared the use of passive voice The Wall Street Journal articles. They found a median of 20% to 26% of use of passive voice in medical articles, while only 3% in The Wall Street Journal. In Figure 3.2, we show the boxplot extracted from their study showing the differences.
- Availability of parallel corpora: In the Open Source Parallel Corpus (OPUS) database[4], as of February 2019, the EMEA corpus (from the European Medicines Agency) is ranked as the 10th for the Portuguese/English language pair, and 12th for English/Spanish, regarding number of sentences. Since OPUS is the largest available collection of parallel data, this position shows that domain-specific corpora is not largely available. Since most translation techniques rely on such type of corpora, this poses as a weakness that has to be accounted for when training MT systems.

Considering the aforementioned challenges, some related works have tried to produce MT systems to the biomedical domain and to expand the availability of parallel corpora. Regarding the latter, the work of Yepes, Prieur-Gaston and Névéol (2013) shows the use of Medline database[5] to retrieve information about articles not written in English to produce a parallel corpus. The authors' approach consists in: (i) retrieving citations from Medline that are in a foreign language, storing their DOIs (Document Object Identification); (ii) access publisher pages using DOI and gather their contents; (iii) based on regular expressions, extract the abstracts in the native language. Once the parallel

[4] <http://opus.nlpl.eu/>
[5] <https://www.nlm.nih.gov/bsd/medline.html>

Figure 3.2: Use of passive voice in medical texts



Box plot of the distribution of the percentages of passive voice frequency for the 4 publications that we analyzed in this study. The standard deviation was 10% for The Journal of the American Medical Association (JAMA) and The New England Journal of Medicine (NEJM), 9% for The Lancet, and 5% for The Wall Street Journal (WSJ). The horizontal line in each box marks the median percentage. The T-bars that extend from the lower and upper borders are defined by the interquartile range; their length is 1.5 times the distance from the 25th to the 75th percentile, which is the length of the box. The length of the upper and lower T-bars may differ because the end of the T-bar must be anchored to observed data points Source: Amdur, Kirwan and Morris (2010)

abstracts were available, they used Hunalign[6] to produce aligned sentences. As a proof-of-concept, they retrieved data for English/Spanish and English/French. An SMT system using Moses was trained with such data and its performance compared to a general domain SMT trained on the newstest2011 corpus[7]. The authors show that the in-domain SMT (i.e. trained wit the data collected by them) system can outperform a system trained with a general domain corpus. In Table 3.3 we show the performance of their system for the aforementioned data.

Table 3.3: BLEU scores for in-domain and general domain SMT systems

Training Set	**EN ->FR**	**FR ->EN**	**EN ->ES**	**ES ->EN**
Newstest2011	24.25	25.78	29.98	30.40
Medline (only titles)	19.29	21.12	25.00	25.59
Medline (titles and abstract)	24.25	25.78	29.98	30.40

Source: Adapted from Yepes, Prieur-Gaston and Névéol (2013)

In a comprehensive review, Wu et al. (2011) studied if automatic machine translation systems were already accurate enough to transled Pubmed titles for patients. They

[6]<https://github.com/danielvarga/hunalign>
[7]<http://www.statmt.org/moses/RELEASE-4.0/models/cs-en/evaluation/newstest2011.filtered.1/>

developed an MT system based on Moses (KOEHN et al., 2007) for bidirectional translation between English and French, Hungarian, Polish, Spanish, German, and Turkish. As a source of comparison, they used the Google Translate system available at the time, which was also SMT based. As shown in Table 3.4, the in-domain MT system outperforms Google for most of the language pairs, thus proving that tailored MT systems provide better accuracy in the biomedical area. In addition, they also developed a parallel corpus using Medline titles available in more than one language using simple querying and regular expressions for parsing of the XML results.

Table 3.4: BLEU scores for in-domain SMT and Google Translate

Language Pair	**Google**	**BioMT**
FR ->EN	37.74	45.46
EN ->FR	34.95	46.54
HU ->EN	19.08	17.35
EN ->HU	08.08	10.88
PT ->EN	29.98	36.04
EN ->PT	17.54	31.70
ES ->EN	45.65	47.64
EN ->ES	44.14	49.32
DE ->EN	36.39	39.63
EN ->DE	23.20	34.48
TR ->EN	26.52	17.33
EN ->TR	13.63	15.40

Source: Adapted from Wu et al. (2011)

Pecina et al. (2014) studied the use of SMT to translate medical texts and medical search queries from Czech, German and French to English. They also used Moses (KOEHN et al., 2007) as MT system, while using already available parallel corpora for training. To compare the developed system, they performed experiments in a test set with sentences from medical texts and search queries in a number of different settings for tuning the SMT system. Their results show that the in-domain trained models performed better in translating medical texts and search queries than the general-domain ones.

In a more recent work, Wołk and Marasek (2015) used the EMEA corpus in English/Polish to train MT systems for this specific domain. They explored the use of SMT with Moses (KOEHN et al., 2007) and an NMT with encoder-decode architecture. Their results show that SMT based translation performed better than the NMT one, by more than 10 BLEU points for both directions (i.e English to Polish and Polish to English). We must notice that their corpus is relative small for NMT purposes, consisting of approximately 1 million sentences, thus SMT performance tends to be better.

3.2.1 Shared task in biomedical translation

The Conference on Machine Translation (WMT) is one of the main venues related to translation technology, dating back to 2006, as a workshop of statistical machine translation. More recently, it was renamed to Conference on Machine Translation to better express the new advances in MT, especially related to NMT. Besides usual article submission and invited talks, this yearly event hosts several shared tasks related to machine translation, such as evaluation, translation and post-editing. In the last edition of WMT (2018)[8], three translation shared tasks took place: translation of news, biomedical documents and multimodal. In the biomedical shared task, which started in 2014, the aim is to develop translation systems to translate article titles and abstracts usually extracted from PubMed.

In the year of 2016, the bioemedical translation task included the following language pairs: English/French, English/Spanish, and English/Portuguese. That year, five teams participated in the task, submiting a total of 40 runs. Evaluation was carried out based on BLEU scores and manual validation (BOJAR et al., 2016). Later in 2017, the shared task included, in addition to the languages in the 2016 edition, the following language directions: English-Czech, English-German, English-Hungarian, English-Polish, English-Romanian, English-Swedish. A total of seven teams submitted their systems for evaluation, which was also carried by means of automatic BLEU scores and manual evaluation (YEPES et al., 2017).

In the year of 2018, when we submitted our systems, two other teams submitted runs for the competition, TALP-UPC and University of Hamburg (UHH). This competition allows one to utilize as many different corpora as wanted, in contrast to usual MT shared tasks, where the competitors can only use a pre-defined set of corpora. This allowed us to employ the parallel corpora generated as part of this thesis, and test them in a competitive scenario. The only restriction to participants was regarding to the use of Pubmed data, since part of it would be used as test set. With this in mind, we tried to remove them from our dataset, as further explained in Section 5.2. Regarding evaluation, this track of WMT uses BLEU scores as automatic evaluation tool, as well as human judgment to evaluate the performance between the systems.

The system from TALP UPC (TUBAY; COSTA-JUSSÃ, 2018) used the Transformer architecture (KAISER et al., 2017; VASWANI et al., 2017), a novel paradigm for

[8]<http://www.statmt.org/wmt18/>

training NMT systems. As RNNs takes a sequence of tokens as input and processes them word by word, parallelization is difficult. In addition, as explained before, when sentences are too long, the model tends to forget long dependencies, which is somewhat alleviated by attention mechanisms. In Transformer, however, there is no recurrence, and order information is given by positional encodings. It also includes Self-Attention layers both in encoder and decoder. The team used a multi-source language training approach, which consists of providing parallel sentences from multiple languages and one single language as output. In this case, they used romance languages as input (Spanish, Portuguese and French) and English as output. The datasets were from MEDLINE and SciELO abstracts. As for pre-processing, tokenization, truecasing and Byte Pair Encoding (BPE) segmentation was performed.

University of Hamburg used Moses to build a SMT system using the EMEA, Wikipedia, Commoncrawl, Paracrawl and the SciELO abstracts corpora. This shows that they also performed in-domain and out-of-domain corpora concatenation, similar to our work. However, for the language model, they build one for in-domain, and other for out-of-domain, which were later interpolated. In addition, they developed a scoring system to filter out-of-domain data based on the in-domain one. The purpose was to select from general corpora the sentences which have the same profile as the biomedical ones. In Section 5.3 we present the results achieved by both systems.

3.3 Summary

In Table 3.5, we summarize the related works regarding their use of parallel corpora in the biomedical domain, in terms of number of sentences, as well as the architecture of machine translation system (i.e. SMT or NMT). From the previous reviewed works, we can see that a very few of them dedicated their efforts to create new distributable parallel corpora for this scientific domain, as well as to explore the use of the more recent NMT systems. Given this finding, we identify two gaps in the literature that we aim to address: the development and distribution of large parallel corpora for biomedical translation and the study of modern NMT architectures for such domain.

We highlight that all mentioned works used only in-domain data for training, without exploring the possibility of concatenating out-of-domain data. This was only proposed by our current work and the UHH team which also participated in the 2018 edition of the

WMT shared task.

Table 3.5: Overview of the related works regarding use of SMT, NMT, corpora development and average number of sentences for training

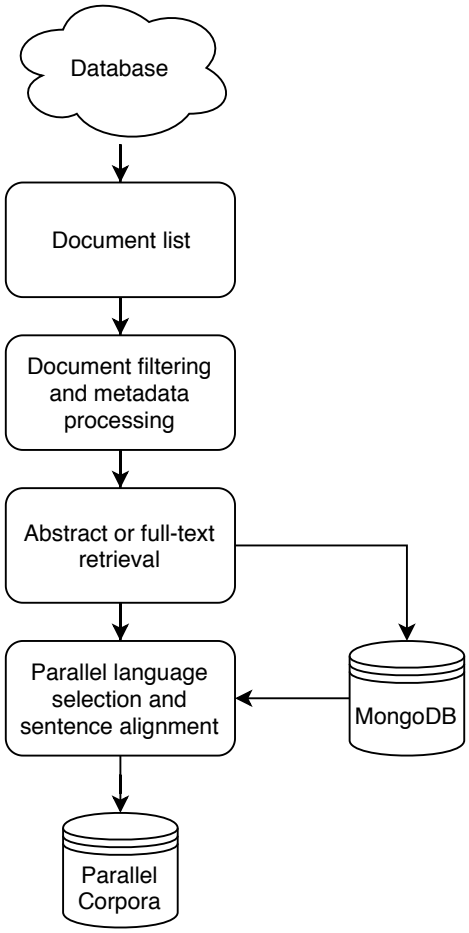
Reference	**SMT**	**NMT**	**Corpus Creation**	**Number of sentences**
Wu et al. (2011)	X		X	0.3M
Yepes, Prieur-Gaston and Névéol (2013)	X		X	0.1M
Pecina et al. (2014)	X			3,000
Wołk and Marasek (2015)	X	X		1.04M
Neves, Yepes and Névéol (2016)	X		X	0.6M
Tubay and Costa-jussÃ (2018)		X		0.8M
This work	X	X	X	2.8M

4 CORPORA DEVELOPMENT

As pointed out in Section 3.1, there is still room to improve the availability of parallel corpora in the biomedical domain, specially for the English to Spanish and Portuguese languages.

This chapter presents the databases we used in our work, as well as the process employed to acquire, pre-process and align the parallel corpora we developed for the biomedical domain. In Figure 4.1, we show an overview of the process we followed for corpora acquisition and processing. Details will be provide in the following sections.

Figure 4.1: Overall process employed for corpora creation.



4.1 Databases

In this Chapter, we aim at providing the reader information about the corpora used (CAPES Theses and Dissertation Catalog, SciELO, and BVSalud) and the processes for

data crawling, processing and storage.

We focus on the three mentioned databases for these main reasons:

- Data availability: all databases are easily accessible online, without the need of authentication of application for licences;
- In-domain data: the database BVSalud is completely oriented to the biomedical domain, while SciELO and CAPES encompass various scientific areas, but with a large number of biomedical texts;
- Open Access: Data from CAPES is completely Open Access, while data from BVSalud and SciELO contain information about licensing for each article.

4.1.1 Scientific databases

Academic (or scientific) databases are electronic collections of documents, such as articles, dissertations, clinical cases, which can be searched by the user. The complete content can be directly accessed or linked to external repositories. The access of scientific databases can be controlled by a subscription, that is, one has to pay a fee to be able to query the database, or rely on institutional access, while others are open access, meaning that the user can query it free of charge.

Some databases are tailored to domain-specific, such as computer science and engineering (e.g. Association for Computing Machinery[1] or IEEE Xplore[2]) and biomedical (e.g. Embase[3] and PubMed[4]). These databases can provide the full-text contents of the indexed articles, or only metadata that can be used to access the documents in its original repository, such as PDF links, titles and abstracts. In this work, we explore three main databases as source of parallel text: SciELO[5], CAPES Theses and Dissertations Catalog (TDC)[6], and Virtual Health Library (BVSalud)[7], which are further described in Section 4.1.

ScIELO and CAPES are databases of general-domain scientific texts, while BVSalud is completely oriented to the biomedical fields. However, as we will show in the Section 4.1, even the general-domain ones have a predominance of biomedical texts.

[1] <http://acm.org/>
[2] <https://ieeexplore.ieee.org/>
[3] <https://www.embase.com/>
[4] <https://www.ncbi.nlm.nih.gov/pubmed/>
[5] <www.scielo.org>
[6] <https://catalogodeteses.capes.gov.br>
[7] <www.bvsalud.org>

4.1.2 Terminological Resources

One can use parallel terminologies from the Unified Medical Language System[8] (UMLS) to train MT systems. The UMLS is a collection of knowledge sources maintained by the NLM (U.S. National Library of Medicine) that facilitates the development and interoperability of systems dealing with health and biomedical data. The Metathesaurus is the base of UMLS, providing a unified access to over 1 million biomedical concepts from several controlled vocabularies and terminologies. Essentially, it links synonyms and alternative views of the same concepts, and also identifies relationships among them (NATIONAL LIBRARY OF MEDICINE, 2009).

To extract parallel concepts, the MetamorphoSys application provided by NLM can be used to subset the language resources for the desired language pairs. This approach is similar to the one proposed by Naspre and Labaka (2016). Once the database is downloaded and imported to MetamorphoSys, the MRCONSO RRF file, which contains all concepts in several language, can be imported to a relational database to split the data in a parallel format in the language pairs. Table 4.1 shows the number of parallel concepts available for each pair considered in this work.

Table 4.1: Number of concepts from UMLS for each language pair

Language Pair	**Concepts**
EN/ES	14,399
EN/PT	26,194

4.1.3 SciELO

The SciELO database is a Latin American and Caribbean initiative developed to meet the needs of developing countries regarding scientific communications, increasing the visibility and access to scientific literature (PACKER, 2000). SciELO comprises a set of methodologies for electronic publication, access and preservation of science, technology, and medicine full-text journals, using the web. SciELO is one of the most important services provided by BIREME (*Biblioteca Regional de Medicina*) (MARCONDES et al., 2003), currently hosting more than 700,000 documents, as detailed in Table 4.2.

Another interesting aspect of SciELO is that several journals publish full-text ar-

[8]<https://www.nlm.nih.gov/research/umls/>

ticles of scientific articles in more than one language, a feature commonly limited to the abstracts in other scientific databases. Therefore, the SciELO database can be a valuable source for parallel corpora for various scientific domains.

Table 4.2: Distribution of SciELO documents according to thematic areas

SciELO Thematic Areas	**Documents**
Health Sciences	352,443
Human Sciences	145,521
Agricultural Sciences	87,866
Biological Sciences	86,216
Applied Social Sciences	79,065
Exact and Earth Sciences	40,881
Engineering	37,848
Multidisciplinary	24,753
Linguistics, Literature and Arts	16,374

Source: SciELO website (September/2018)

4.1.4 CAPES TDC

CAPES (*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) is the Brazilian governmental body responsible for overseeing post-graduate programs across the country. Among its roles, CAPES tracks every enrolled student and scientific production. In addition, CAPES maintains a freely accessible database of theses and dissertations produced by these graduate students (i.e. Theses and Dissertations Catalog - TDC) since 1998. However, abstracts are only available since 2013. Under recent governmental efforts in data sharing, CAPES made TDC available in CSV format, containing both abstracts and metadata, making it easily accessible for data mining tasks. In table 4.3 we show the number of documents available in TDC for each year.

Recent data files, from 2013 to 2016, contain valuable information not restricted to NLP purposes, such as abstracts in Portuguese and English, scientific categories, and keywords. Thus, TDC can be an important source of parallel Portuguese/English scientific abstracts. Details about the distribution of documents according to the area of knowledge can be found in Section 4.8.1.1.

Table 4.3: Distribution of documents according to year in CAPES TDC

Year	Documents
2013	68,103
2014	71,074
2015	76,296
2016	80,592

Source: CAPES website (September/2018)

4.1.5 BVSalud

In Latin America and Carib, the Pan American Health Organization (OPAS), in agreement with BIREME, maintains the BVS database, which is an important source of biomedical texts in three main languages: English, Spanish, and Portuguese. Founded in Brazil in 1967, under the name of Regional Medicine Library (from which the acronym BIREME comes from), it keeps pace with the growing demand for up-to-date scientific literature from the Brazilian health systems and the communities of healthcare researchers, professionals and students. Then, in 1982, its name changed to Latin-American and Caribbean Center on Health Sciences Information so as to better express its dedication to the strengthening and expansion of the flow of scientific and technical health information across the region (Brazil) (2011).

Currently, BVS has more than 1 million texts indexed, and provides integrated search capabilities with PUBMED. In Table 4.4, we show the different databases of interest in this study that are part of BVS, and their number of documents, which many of them are available in more than one language.

Table 4.4: Distribution of documents among the main databases in BVS

Database	Documents
LILACS	827,769
IBECS	171,323
BINACIS	141,014
CUMED	62,755
Index Psychology	55,441
BBO - Dentistry	46,294
BDENF - Nursing	38,476

Source: BVS website (September/2018)

4.2 Licensing

Most articles in the Scielo database are licensed under the Creative Commons copyright, with different types of licenses. In order to be able to distribute the contents of the gathered articles, we filtered only those licensed under terms that allow derivatives, since ND (No Derivatives) licenses require the content to be distributed without any modification. As we removed some parts from the articles (e.g. images, tables, references), we would be infringing such copyright rules. All articles distributed in our dataset contain the corresponding license, authorship, and unique identifiers of original sources.

For the CAPES dataset, no special requirements are needed related to licensing, since all data are open source according to the government open data program.

As for the BVSalud database, we included in our datasets only open access documents. To retrieve license information, we crawled the BVS website containing information about the indexed journals [9] as well as the Directory of Open Access Journals [10].

4.3 Document retrieval

Scielo's website provides unified access to a series of regional databases (such as from Argentina, Brazil, South Africa) with full-text articles, offering simple and advanced search capabilities. We iteratively queried the database to retrieve all lists of results, which were then parsed and all relevant contents, such as URLs for all available languages of each article, authorship, licensing, title, and abstract were stored in a database. The availability of full-text articles opens the possibility of paragraph and section alignment, which provides a more complete pre-treatement. Thus, we adopted the MongoDB database system, as it is document-oriented, and allows for the easy querying and storage of this type of data. Scripts for document retrieval can be found in <https://github.com/soares-f/scrapper-bvs>

Then we queried the results in MongoDB to filter only the articles meeting the following constraints: a) articles with full-text available in at least two of three languages of interest (i.e. English, Portuguese, and Spanish); and b) type of licensing is non ND terms. The full-text of all articles meeting these two criteria were downloaded from the

[9] <http://portal.revistas.bvs.br/>
[10] <https://doaj.org/>

Scielo database in HTML format.

For the BVSalud database, we performed a document retrieval similarly to SciELO, however, we just focused on the abstracts.

The TDC datasets are available in the CAPES open data website[11] divided by years, from 2013 to 2016 in CSV and XLSX formats. We downloaded all CSV files from the respective website and loaded them into a relational database for better manipulation. This differs from the case of Scielo, since only abstracts are available, thus there is no hierarchical relation in the extracted texts, not requiring a more flexible database such as MongoDB. The database was then filtered to remove documents without both Portuguese and English abstracts, and additional metadata collected (e.g. authorship, year of publication).

4.4 Document parsing

Since BVSalud and CAPES TDC datasets are only from abstracts, their document parsing process only requires database query. This process of identifying and selecting the relevant information is required in SciELO, since we were mining full-text data, which has an hierarchical structure (e.g. sections and subsections), as well as other items that are of no interest for the purpose of parallel corpus creation, such as figures and tables. We now describe the steps performed for document parsing in the SciELO database.

The HTML contents of all articles were parsed using an in-house Python script tailored to the Scielo format. First, all non-textual elements, such as images, tables, references, citations, and footnotes were removed. Our algorithm was designed to preserve the hierarchical and paragraph structure of the article across the different languages in order to produce results aligned at paragraph and section levels. This could help achieving good sentence level alignment.

The main challenges in parsing Scielo HTML contents are heterogeneity issues concerning HTML structure and formatting over different years. More recent articles are well-formated and contain specific tags for paragraphs, sections, subsections, and titles. We concentrated efforts in developing rules to tackle all ill-formated HTML issues identified, so as to cover as much content as possible, but to reduce the risk of misalignment, we discarded all documents that presented very different structures across the languages.

Each parsed full-text translation was stored in MongoDB aiming at preserving the

[11] <https://dadosabertos.capes.gov.br/dataset/catalogo-de-teses-e-dissertacoes-de-2013-a-2016>

structure of the articles. When our parsing algorithm failed at identifying the document structure, its content was stored as a unstructured list of paragraphs, as we assume that if two translations of the same article present the same number of parsed paragraphs, it is likely they can be simply aligned according to their order.

4.5 Pre-processing

In the pre-processing of all databases, we performed a language checking, using the Langdetect[12] Python package, to make sure that there was no misplacing of English abstracts in the Portuguese field, for instance, or the other way around, removing the documents that presented such inconsistency in the CAPES and BVSalud datasets. For the SciELO dataset, we performed such checking at the paragraph level, since segments are from full-text.

We removed newline/carriage return characters (i.e \n and \r), as they would interfere with the sentence alignment tool. In addition, to the CAPES dataset, we performed a case folding to lower case letters, since the TDC datasets present all fields with uppercase letters.

4.6 Sentence Alignment

In this work, we focus in sentence align corpora. Thys, we used the LF aligner tool[13], a wrapper around the Hunalign algorithm (VARGA et al., 2007).

In SciELO, for articles with the same structure across the languages, pairs of parallel paragraphs were input to the sentence aligner at a time, aiming at reducing the risk of misalignment. For the other cases, all paragraphs were passed to the aligner together. For BVSalud and CAPES, abstracts were directly supplied to Hunalign. Aligned sentences were stored as text files for post-processing.

After sentence alignment, the following post-processing steps were performed, since abstracts or full-text segments could contain texts in more than one language: (i) removal of all non-aligned sentences; (ii) removal of all sentences with fewer than three characters, since they are likely to be noise from ill-formatted HTML or abstract formatting; (iii) removal of all sentences written in the same language using the language

[12]<https://github.com/Mimino666/langdetect>
[13]<https://sourceforge.net/projects/aligner/>

detector.

4.7 Manual Evaluation

Although the Hunalign algorithm usually presents a good alignment between sentences, we also conducted a manual validation to evaluate the quality of the aligned sentences.

For the SciELO corpus, we randomly selected 300 pairs of sentences, 100 for each language pair, and 100 trilingual sentences. If the pair was correctly aligned, we marked it as "correct", otherwise, as "no alignment". For the BVS corpus, we randomly selected 300 sentences, 100 for the triligual subset, and 100 for each subset of EN/PT and EN/ES. If the pair was fully aligned, we marked it as "correct"; if the pair was incompletely aligned, due to segmentation errors, for instance, we marked it as "partial"; otherwise, when the pair was incorrectly aligned, we marked it as "no alignment". For CAPES BTD, we selected 400 pairs of sentences, and the strategy followed the same of BVS. We included a larger number of pairs since this database is prone to have more errors, as we had to perform casefolding, which impacts segmentation and alignment.

The reason why ScIELO presented different evaluation strategy is that since we were dealing with full-text data and well behaved paragraphs, segmentation errors were almost non-existent.

4.8 Results for Parallel Corpora Development

In this section, we show the results we achieved when building the corpora, regarding corpora statistics and the manual evaluation of alignment quality.

4.8.1 Corpora Statistics

4.8.1.1 CAPES TDC

Table 4.5 shows the number of documents and sentences for the aligned corpus according to the 9 main knowledge areas defined by CAPES. The dataset is available[14]

[14]<https://figshare.com/s/6f760a4f3610a83c2e3f>

in TMX format (RAWAT; CHANDAK; CHAUHAN, 2016), since it is the standard format for translation memories. We also made available the aligned corpus as an SQLite database in order to facilitate future stratification according to knowledge area, for instance. In this database, we included the following metadata information: year, university, title in Portuguese, type of document (i.e. theses or dissertation), keywords in both languages, knowledge areas and subareas according to CAPES, and URL for the full-text PDF in Portuguese. An excerpt of the corpus is shown in Table 4.6.

Table 4.5: CAPES TDC corpus statistics according to knowledge area.

Knowledge Area	**Docs**	**Sents**	**Tokens EN**	**Tokens PT**
Health Sciences	38,221	224,773	5.46M	5.51M
Humanities	38,493	189,648	5.63M	5.54M
Applied Social Sciences	32,176	160,131	4.66M	4.60M
Agricultural Sciences	26,740	154,710	3.92M	3.92M
Engineering	27,074	149,888	3.87M	3.92M
Multidisciplinary	26,502	140,849	3.84M	3.81M
Exact and Earth Sciences	19,630	106,098	2.64M	2.66M
Biological Sciences	16,465	98,994	2.33M	2.34M
Linguistic and Arts	13,717	64,281	1.99M	1.96M
Total	239,018	1,289,372	34.35M	34.28M

4.8.1.2 SciELO

Table 4.7 shows the overall corpus statistics in terms of documents, sentences and tokens for all language pairs and for the set of trilingual aligned documents. One may notice that EN-PT documents are predominant over other language pairs. This may be explained by the fact that almost all Brazilian journals are published through Scielo, thus favoring Portuguese-English translations.

The datasets are also available[15] in TMX format. Besides the aligned sentences, we included the following metadata for each document: aligned title, authors, copyright license, DOI (if available), journal name, Scielo's unique identifier, and subject area. This information was included either to fully comply with Creative Commons requisites, or to provide additional information for other possible applications, such as text classification or clustering. The metadata is also available in an SQLite database.

An example of trilingual sentence is shown in Table 4.8.

[15]<https://figshare.com/s/091fcaf8ad66a3304e90>

Table 4.6: Excerpt of the CAPES TDC corpus with document ID.

ID	Portuguese	English
127454	nessa tese apresentamos duas linhas de pesquisa distintas, a saber, na primeira, referente aos capítulos 1 e 3 aplicamos técnicas estatísticas à análise de imagens do satélite de abertura sintética (sar) e, na segunda, referente ao capítulo 2, examinamos problemas relativos à estimação de parâmetros por máxima verossimilhança na distribuição exponencial-poisson.	in this thesis we present two distinct research lines, namely, the first, referring to chapters 2 and 3, apply statistical techniques to the analysis of synthetic aperture radar (sar) images, and the second, referring to chapter 4, we examined problems concerning parameter estimation by maximum likelihood in exponential-poisson distribution.
1419264	para determinação dessa flora utilizamos os recursos de observação, coleta e identificação.	we use the resources of investigation, collection and identification to determine this flora.
439358	estimaram-se os benefícios ambientais da reciclagem de veículos com mais de 10 anos de uso, considerando os poluentes na fabricação de um veículo novo.	we estimated the environmental benefits of recycling vehicles in use more than 10 years, taking into consideration pollution engendered in the manufacture of a new vehicle.
675023	a coleta de dados se deu por meio de entrevista semiestruturada com 12 familiares cuidadores de crianças atendidas em pronto-socorro pediátrico de um hospital de ensino.	data collection was through semi-structured interviews with 12 family caregivers of children seen in a pediatric emergency department of a teaching hospital.
675023	os dados foram submetidos à análise de conteúdo temático conforme bardin (2011).	the data were subjected to thematic content analysis according to bardin (2011).
1173306	o planejamento e programação do projeto de construção naval têm dois objetivos por base: diminuir o tempo de fabricação e os custos.	shipbuilding project planning and scheduling possess two major objectives: manufacturing time and cost reduction.

4.8.1.3 BVSalud

Table 4.9 shows the statistics (i.e. number of sentences) for the aligned corpus according to the 2 language pairs and the trilingual subset. The dataset is also available in TMX format. In this database, we included the following metadata information: year, keywords in the available languages, database of origin, country, authorship, and URL for the full-text when available. Similarly, we made available an SQLite database for better sub-setting.

Table 4.7: SciELO corpus statistics for all language pairs and the trilingual set. Number of tokens are in the same order of the languages column.

Languages	Docs	Sents	Tokens
EN-ES	2,029	177,781	5.2M
			5.7M
PT-ES	76	4,987	140,434
			151,148
EN-PT	29,609	2.9M	76.0M
			77.3M
EN-PT-ES	3,142	255,914	7.0M
			7.8M
			7.2M

Table 4.8: Example of trilingual aligned sentences in the SciELO dataset.

Language	Text
English	Among its objectives, it aims to defend the interests of society and Nursing in the context of Public Policies and the Unified Health System with emphasis on Mental Health
Spanish	Entre sus objetivos está defender los intereses de la sociedad y de la Enfermería en el contexto de las Políticas Públicas y del Sistema Único de Salud con énfasis en el área de la Salud Mental.
Portuguese	Entre seus objetivos, visa defender os interesses da sociedade e da Enfermagem no contexto das Políticas Públicas e do Sistema Único de Saúde com ênfase na área de Saúde Mental.

4.8.2 Alignment Evaluation

All sentences were randomly sample for the developed corpora, as explained in Section 4.7.

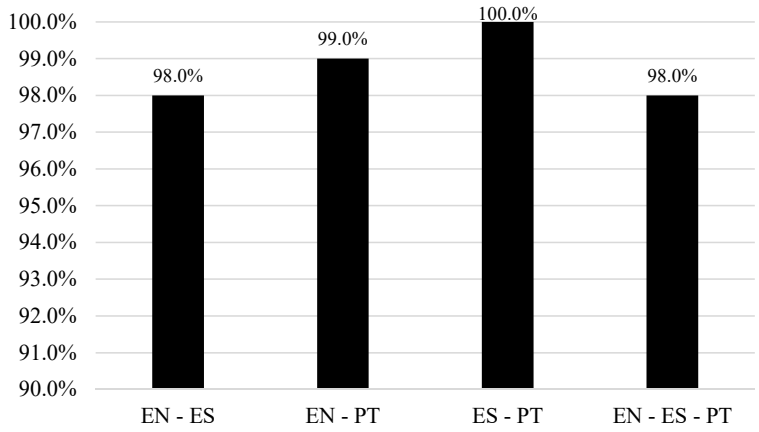
For SciELO, Figure 4.2 depicts the rate of correct alignments for each subset of parallel languages. All language combinations presented at least 98% of correct alignments, with the language pair ES - PT achieving 100%.

For CAPES TDC, 82.30% were correctly aligned, while 13.33% were partially aligned, and 4.35% presented no alignment. Regarding the partial alignment, most of the problems are due to segmentation issues previous to the alignment, which wrongly split the sentences. Since all words were case folded to lowercase letters, the segmenter lost an important source of information for the correct segmentation, generating malformed sentences. Some examples of partial alignment errors are shown in Table 4.10, where most sentences were truncated in the wrong part. This does not happen in other corpora, which did not require any case folding, thus preserving the original lowercase or uppercase letters, which provides important information for segmentation.

Table 4.9: Corpus statistics according to language pair in BVSalud. Number of tokens are in the same order of the languages column.

Language Pairs	Docs	Sents	Tokens
EN/PT	144,576	711,475	17.30M
			17.49M
EN/ES	184,434	789,547	19.23M
			19.41M
EN/ES/PT	50,463	203,719	4.94M
			4.98M
			4.99M

Figure 4.2: SciELO alignment accuracy for the four language subsets.



Regarding BVSalud, from all the evaluated sentences, average 96% were correctly aligned, while average 2% were partially aligned. The trilingual subset was the one with the best alignment, achieving 97% correct alignment. Figure 4.3 shows the alignment accuracy for all language subsets.

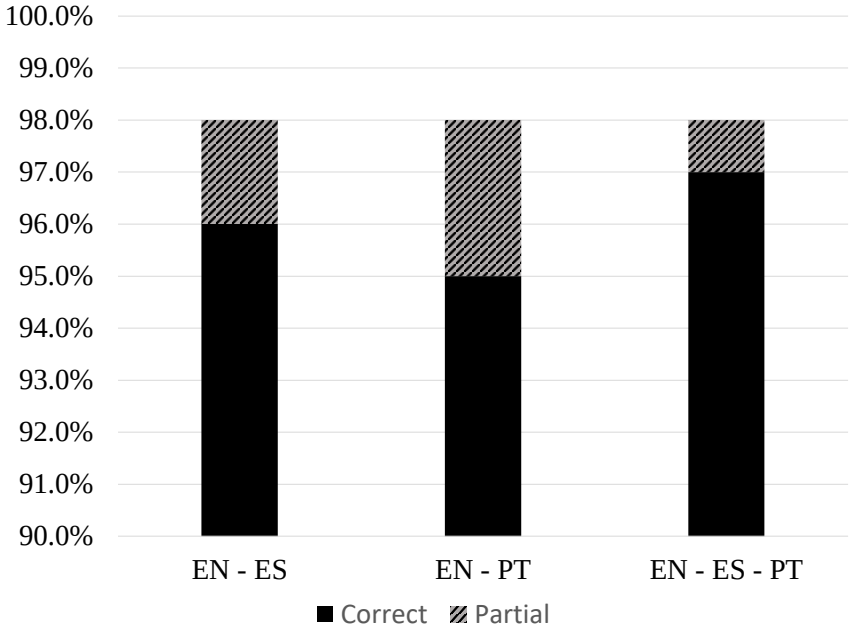
Different factors may have contributed to this high alignment quality. The use of Hunalign (VARGA et al., 2007) with a dictionary is perhaps the most probable reason, as it combines a dictionary with sentence-length information to boost alignments. For SciELO, the input of articles segmented by parallel paragraphs also contributed to quality enhancement, since this can reduce the probability of misalignment.

The development of the three biomedical parallel corpora is a significant contribution to the area of MT, since our presented corpora surpasses by a large margin the size of the already existing ones, as shown in Table 3.5. In addition to the size of corpus, in terms of sentences, we also highlight the high quality of the alignment, which reached figures of approximately 99% of correct alignment. All these features make these corpora an invaluable resource for the development of accurate MT systems for the biomedical

Table 4.10: Examples of partial alignment errors in CAPES TDC dataset

Portuguese	English
os dados foram comparados entre os grupos por anova de medidas repetida	data were compared by repeated measures anova. results: we found a significa
o estudo utilizará um software comercial para simular a peça	the study will use commercial software to simulate the piece with a number of different crack sizes and the
buscamos subsídios teóricos em autores que veem na reflexão e na pesquisa um grande potencial para o desenvolvimento d	we seek theoretical support in authors who see in reflection and research a great potential for

Figure 4.3: BVS alignment accuracy for the three language subsets.



domain.

As for publications, the SciELO corpus was presented in the LREC 2018, while the CAPES one in the PROPOR 2018. Currently, the BVS corpus is being considered for publication in the Corpora conference.

5 MACHINE TRANSLATION

Now we detail the resources and translation models used in our study, as well as the results achieved by our methods.

5.1 Language Resources

In this section, we describe the language resources used to train both models, which are from two main types: corpora and terminological resources. Corpora can provide evidences for MT systems regarding vocabulary and language structure, while terminological resources can cover in-domain terms that may not appear in the corpora.

5.1.1 Corpora

We used both in-domain and general domain corpora to train our systems. We expect that the general domain corpora can improve fluency, since they follow a similar grammar construction than scientific articles (e.g. formal language, use of passive voice in romance languages, no contractions in English), while the in-domain corpora may provide larger vocabulary coverage and fine-tuned grammar properties. For general domain data, we used the books corpus (TIEDEMANN, 2012), which is available for several languages, included the ones we explored in our systems, and the JRC-Acquis (TIEDEMANN, 2012). As for in-domain data, we included several different corpora:

- The corpus of theses and dissertations abstracts (TDC), as detailed in Section 4.1.4.
- The corpus of full-text scientific articles from Scielo, which acquisition was detailed in Section 4.1.3.
- The corpus from Virtual Health Library[1] (BVSalud), also detailed in Section 4.1.5
- A subset of the UFAL medical corpus[2], containing the Medical Web Crawl data for the English/Spanish language pair.
- The EMEA corpus (TIEDEMANN, 2012), consisting of documents from the European Medicines Agency.

Table 5.1 depicts the original number of parallel segments according to each cor-

[1] <http://bvsalud.org/>
[2] <https://ufal.mff.cuni.cz/ufal_medical_corpus>

pora source. In Section 5.2.1, we detail the pre-processing steps performed on the data to comply with the task evaluation.

Table 5.1: Original size of individual corpora used in our experiments

Corpus	Sentences	
	EN/ES	EN/PT
Books	93,471	-
UFAL	286,779	-
Full-text Scielo	425,631	2.86M
JRC-Acquis	805,757	1.64M
EMEA	-	1.08M
CAPES-BDTD	-	950,252
BVS	737,818	631,946
Total	2.37M	7.19M

5.1.2 Terminological Resources

Regarding terminological resources, we extracted parallel terminologies from the Unified Medical Language System[3] (UMLS). For that matter, we used the MetamorphoSys application provided by U.S. National Library of Medicine (NLM) to subset the language resources for our desired language pairs. Our approach is similar to what was proposed by Naspre and Labaka (2016), as described in Section 4.1.2. We expect that terminological resources can provide better translation for specific biomedical terms, since they are specific to this domain.

Once the resource was available, we imported the MRCONSO RRF file to a relational database to split the data in a parallel format in the two language pairs. Table 5.2 shows the number of parallel concepts for each pair.

Table 5.2: Parallel UMLS concepts for each language pair

Language Pair	Concepts
EN/ES	14,399
EN/PT	26,194

[3]<https://www.nlm.nih.gov/research/umls/>

5.2 Experimental Settings

In this section, we detail the pre-processing steps employed as well as the architecture of the SMT and NMT systems. All our settings and procedures regarding MT training were carried out considering the guidelines of the bioemdical translation shared task[4] in the Third Conference on Machine Translation (WMT18). For this shared task, we participated with the following language directions: English→Portuguese, Portuguese→English, English→Spanish, Spanish→English.

5.2.1 Pre-processing

As detailed in the description of the biomedical translation task, the evaluation is based on texts extracted from Medline. Since one of our corpora, the one comprised of full-text articles from Scielo, may contain a considerable overlap with Medline data, we decided to employ a filtering step in order to avoid including such data.

The first step in our filter was to download metadata from Pubmed articles in Spanish and Portuguese. For that matter, we used the Ebot utility[5] provided by NLM using the queries *POR[la]* and *ESP[la]*, retrieving all results available. Once downloaded, we imported them to a relational database which already contained the corpora metadata. To perform the filtering, we used the *pii* field from Pubmed to match the Scielo unique identifiers or the title of the papers, which would match documents not from Scielo.

Once the documents were matched, we removed them from our database and included the terminological resources. This step was done by concatenating the terminological parallel pairs as simple sentences in the training corpus. To try to alleviate the imbalance between sentences from parallel corpora and terminology, the terms were oversampled 10 times. Later, data were partitioned in training and validation sets. Please notice that the validation set does not contain any terminological pair. Table 5.3 contains the final number of sentences for each language pair and partition.

Table 5.3: Final corpora size for each language pair

Language	Train	Dev
EN/ES	2.35M	22,670
EN/PT	7.17M	24,206

[4]<http://www.statmt.org/wmt18/biomedical-translation-task.html>
[5]<https://www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot/ebot.cgi>

5.2.2 SMT System

We used the popular Moses toolkit (KOEHN et al., 2007) to train our SMT system for the two language pairs. As training parameters, we followed the Moses baseline steps[6] to train four MT systems (i.e. one for each translation direction). The baseline training is usually used as comparison method for evaluating corpora quality, which is our case, since the system is standardized. The following parameters are used:

- Language Model:KenLM 3-gram model
- Phrase-Table: Up to 3 tokens
- Beam search size: 5
- Reordering: msd-bidirectional-fe
- Optimization algorithm: MERT
- Epochs: Until convergence

Regarding training, we used the Amazon AWS spot virtual machines with 24 cores and 60GB of RAM, and used parallelization as much as possible to reduce training time and the associated cost. One drawback of Moses is the required amount of RAM needed for large corpora, which is our case. During word alignment and phrase-table construction, all hash tables are loaded in memory, requiring such type of virtual machine.

5.2.3 NMT System

As for the NMT system, we employed the OpenNMT toolkit (Klein et al., 2017) to train four MT systems, one for each translation direction. Tokenization was performed by the supplied OpenNMT algorithm. Regarding network parametrization, the following settings were used, while all other parameters were set as default:

- Encoder type: bidirectional recurrent neural network
- Decoder type: Seq2Seq with attention (default)
- Word vector size: 600
- Layers (encoder and decoder): 4
- RNN size: 800
- Batch size: 64

[6] <http://www.statmt.org/moses/?n=moses.baseline>

- Vocabulary size: 50000

The choice of bidirectional recurrent neural network was made due to its past records of high accuracy and performance through several tasks. At the time we performed such experiments, the Transformer model was not completely matured nor easily available in the OpenNMT toolkit.

To train our system, we used the Azure virtual machines with a single NVIDIA Tesla V100 GPU. The models with the best perplexity value were chosen as final models. During translation, OOV (Out-of-Vocabulary) words were replace by their original word in the source language, all other OpenNMT options for translation were kept as default. By replacing the OOV words with their original version, we expect that acronyms which are used in English for all languages are preserved (e.g. PCR - Polymerase Chain Reaction), thus achieving better BLEU scores. In addition, even if the word is not an acronym, from the user point of view, it is desirable that the OOV words are not just marked as unknown, but presented in an more understandable way, even though in another language. Moreover, there is a predominance of Latin terms in biomedical literature, which makes them understandable even when not translated.

5.3 Experimental Results and Discussion

5.3.1 Automatic Evaluation

We now detail the results achieved by our SMT and NMT systems on the official test data used in the shared task. Table 5.4 shows the BLEU scores (PAPINENI et al., 2002) for both systems and for the submissions made by other teams.

Table 5.4: Official BLEU scores for the English/Spanish and English/Portuguese language pairs in both translation directions. Bold numbers indicate the best result for each direction.

Team, Runs	EN/ES	EN/PT	ES/EN	PT/EN
UFRGS run1 (NMT)	39.62	**39.43**	43.31	**42.58**
UFRGS run2 (SMT)	**39.77**	**39.43**	**43.41**	**42.58**
TGF TALP UPC run1	-	-	40.49	39.49
TGF TALP UPC run2	-	-	39.06	38.54
UHH-DS run1	31.32	34.92	36.16	41.84
UHH-DS run2	31.05	34.19	35.17	41.80
UHH-DS run3	31.33	34.49	36.05	41.79

The organizers of the shared task also provided results for the subset of well aligned sentences extracted from Pubmed, which is a more realistic assessment of the performance, since it does not include errors due to malformed input data. These results are shown in Table 5.5.

Table 5.5: Official BLEU scores for the English/Spanish and English/Portuguese language pairs in both translation directions using only well aligned sentences. Bold numbers indicate the best result for each direction.

Team, Runs	EN/ES	EN/PT	ES/EN	PT/EN
UFRGS run1 (NMT)	**44.50**	**43.14**	**46.92**	**46.01**
UFRGS run2 (SMT)	**44.50**	**43.14**	**46.92**	**46.01**
TGF TALP UPC run1	-	-	42.91	42.55
TGF TALP UPC run2	-	-	41.26	41.56
UHH-DS run1	34.77	37.27	38.45	44.28
UHH-DS run2	34.70	36.76	37.17	44.32
UHH-DS run3	35.08	36.91	38.18	44.27

Our submissions achieved the best results for all translation directions we submitted, with remarkable BLEU scores for the ES/EN and PT/EN pairs. When compared to the other teams, our results presented similar behavior, with higher scores when English was the target language, which may be explained by the poor English morphosyntactic structure. For the English/Spanish pair, the SMT system presented slightly better results than the NMT one, probably due to the vocabulary size used in the NMT, which leads to more OOV words.

Regarding the superior results achieved, we assume that the large parallel corpora used in our experiments played an essential role. Although we did not use the provided Scielo abstracts corpus (NEVES; YEPES; NéVéOL, 2016), we used a newer parallel corpus also from Scielo, but comprised of full-text articles, which overlaps with the abstracts, but contains more data.

In addition to the biomedical and health corpora, we employed two out-of-domain corpora that we assumed to have a similar structure to scientific texts: the books and the JRC-Acquis (TIEDEMANN, 2012). Another option would be to use the large Europarl corpus (KOEHN, 2005), but we disregarded it since it is comprised of speeches transcripts, which do not follow the usual structure of scientific texts.

As for the better SMT performance, we assume that this happened because of the size of the vocabulary used in the NMT model (50,000 words). Thus, out-of-vocabulary words are more frequent. We were not able to use a larger vocabulary due to the computational resources available. We also point out that SMT models are less prone to out-of-

vocabulary words, since the phrase-table can be produced with very low occurences of a specific word.

5.3.2 Manual Evaluation

In addition to the automatic evaluation based on BLEU scores, the organizers of the shared task provided manual evaluations for all participant teams. They used the submission stated as primary by the participants. As described in (NEVES et al., 2018), they used a 3-way ranking method of evaluation. For each pairwise comparison, they checked a total of 100 randomly selected sentence pairs. The annotator should read the two sentences (A and B), i.e. translations from two systems or from reference, and choosing one of the following options:

- A<B: when the quality of translation B was higher than A.
- A=B: when both translations had similar quality.
- A>B: when the quality of the translation A was higher than B.
- Flag error: when the translation did not seem to be derived from the same input sentence. This is usually related to errors in corpus alignment.

In Table 5.6 we show the official results from the manual evaluations of the submissions for all language pairs we participated. We can see that for the language pair EN/ES, our submission presented better results than the reference (original sentences). Similarly, for the language EN/PT, we achieved similar or better performance than reference, since in 6 sentences our translations were better and in 43 sentences our performance was judged similar to the reference. Thus, in 49 sentences we were better or equal than reference, and in 42 worse.

When comparing Table 5.5 (automatic evaluation) with Table 5.6 (manual evaluation), one can notice that there is a shift in the ranking for the languages ES/EN and PT/EN. For ES/EN, the team TGF TALP UPC achieved best results than our submission, while we remained better than UHH-DS. Regarding PT/EN, TGF TALP UPC moved from being the last ranked in automatic evaluation to the first position, while we remained better than UHH-DS. For EN/ES and EN/PT, we remained better than the other teams.

As stated by the organizers of the shared task, our submission might have achieved better results regarding BLEU scores by correctly translating particular medical concepts, while TGF TALP UPC may have achieved better sentence coherence and fluency by the

Table 5.6: Official results for the manual evaluations. It is important to notice that the values are absolute counts, not percentages.

Languages	Runs (A vs B)	Total	A>B	A=B	A<B
EN/ES	UFRGS vs reference	86	37	23	26
	UFRGS vs UHH-DS	88	29	37	22
	reference vs UHH-DS	92	30	33	29
EN/PT	UFRGS vs reference	86	6	43	42
	UFRGS vs UHH-DS	100	32	53	15
	reference vc UHH-DS	81	46	28	7
ES/EN	TGF TALP UPC vs reference	72	26	12	34
	TGF TALP UPC vs UFRGS	100	51	38	11
	TGF TALP UPC vs UHH-DS	98	79	12	7
	reference vs UFRGS	77	50	15	12
	reference vs UHH-DS	77	54	10	13
	UFRGS vs UHH-DS	700	45	24	31
PT/EN	TGF TALP UPC vs reference	89	25	26	38
	TGF TALP UPC vs UFRGS	100	55	24	21
	TGF TALP UPC vs UHH-DS	100	58	24	18
	reference vs UFRGS	87	42	22	23
	reference vs UHH-DS	87	52	28	7
	UFRGS vs UHH-DS	100	48	27	25

use of the Transformer architecture. We stress that our effort to build quality large parallel biomedical corpora enhanced the in-domain vocabulary availability, improving the translation of medical concepts. In addition, the organizers (NEVES et al., 2018) point out that the TGF TALP UPC team also trained on the Scielo database, however, they did not mention any attempt to remove potential overlapping sentences between Medline (test set) and Scielo, which was performed in our experiments.

5.3.3 Discussion

From the results presented above and our initial hypotheses drawn in Section 1.2 we can point some important discussion insights.

The first hypothesis was that the concatenation of in-domain and out-of-domain similar corpora could lead to a greater accuracy in translation. When comparing the results from our NMT model with the TALP-UPC one, we can see that regarding automatic evaluation, our system presented higher score. In this sense, it is also important to highlight that the Transformer model is currently seen as the state-of-the-art in MT, as well as the use of multilingual translation has proved to increase translation performance (JOHN-

SON et al., 2017). Thus, the TALP-UPC included two important performance boosters in their model, but was not able to achieve the same evaluation as our NMT regarding BLEU points. This can show that the corpus size and the concatenation played an important role.

Regarding the second hypothesis, that the NMT model could outperform the SMT one, we do not have strong evidences to support that with the experiments we carried out. In fact, in some cases the SMT system outperformed the NMT. As explained before, we see the issue regarding the vocabulary size in the NMT as the leading factor for such. While TALP-UPC team used BPE to overcome vocabulary limitation, we used a fixed size. This can lead to more unknown translations, but is able to preserve frequent words and their collocations.

An additional important point to stress is regarding the inclusion of terminological resources. In our experiment, we tried to oversample the terms when concatenating them into the training corpus, but when setting the vocabulary size limit, such terms may disappear. Thus, we think it is important to try alternative ways of enforcing terminologies, such as during beam-search in decoding.

6 CONCLUSIONS

In this work, we aimed at developing automatic machine translation systems adapted for the biomedical domain, with focus on scientific texts. We explored the following languages: English to Spanish, Spanish to English, English to Portuguese, and Portuguese to English.

We identified a gap in the literature regarding the number of parallel corpora available for such domain and languages. We also found a scarce number of studies covering Neural Machine Translation for biomedical domain, as well as comparisons to established Statistical Machine Translation models. Given that, we decided that we would aim at filling those gaps by: developing and validating new and large parallel corpora for biomedical translation in the languages already mentioned, and performing experiments with both neural and statistical machine translation models.

Regarding corpora development, we explored the open access contents of CAPES Thesis and Dissertations Catalogue, ScIELO and BVSalud, creating parallel corpora from these databases. The corpora were manually validated regarding alignment accuracy and statistics about corpora size were provided. We achieved high quality alignment and large corpora, with more sentences than the ones already publicly available.

As for the experiments, we employed Moses for SMT and OpenNMT for NMT, with a sequence-to-sequence model. In general, SMT models presented better performance than NMT ones, which we owe due to the size of the vocabulary set for the NMT. In addition, to externally validate our experiments, we participated on the Third Conference on Machine Translation (WMT18) shared task on biomedical translation, achieving the highest BLEU scores for all language pairs we participated.

Regarding future work, to solve the out-of-vocabulary issue in NMT, we recommend and will research the use of subword methods, such as byte-pair-encoding (BPE). Recent researches also point out that training systems with multiple languages, such as (Portuguese, Spanish) to English might provide a further increase in performance, since the model can benefit from similarities and correlations in the source languages.

REFERENCES

ABDUL-RAUF, S. et al. **Evaluation of sentence alignment systems**. [S.l.], 2010.

AMDUR, R. J.; KIRWAN, J.; MORRIS, C. G. Use of the passive voice in medical journal articles. **AMWA Journal: American Medical Writers Association Journal**, v. 25, n. 3, 2010.

AMMAR, W. et al. Massively multilingual word embeddings. **arXiv preprint arXiv:1602.01925**, 2016.

AZIZ, W.; SPECIA, L. Fully automatic compilation of a Portuguese-English parallel corpus for statistical machine translation. In: **STIL 2011**. Cuiabá, MT: [s.n.], 2011.

BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. **arXiv preprint arXiv:1409.0473**, 2014.

BANERJEE, S.; LAVIE, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: **Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization**. [S.l.: s.n.], 2005. p. 65–72.

BELINKOV, Y.; GLASS, J. Large-scale machine translation between arabic and hebrew: Available corpora and initial results. **arXiv preprint arXiv:1609.07701**, 2016.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python: analyzing text with the natural language toolkit**. [S.l.]: " O'Reilly Media, Inc.", 2009.

BOJAR, O. et al. Findings of the 2016 conference on machine translation. In: **Proceedings of the First Conference on Machine Translation**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 131–198. Available from Internet: <http://www.aclweb.org/anthology/W/W16/W16-2301>.

(BRAZIL), B. . P. . W. **VHL Guide 2011**. VHL Guide 2011, 2011. Available from Internet: <http://modelo.bvsalud.org/wp-content/uploads/Guia_da_BVS_2011_pt.pdf>.

BRITZ, D. et al. Massive exploration of neural machine translation architectures. **CoRR**, abs/1703.03906, 2017. Available from Internet: <http://arxiv.org/abs/1703.03906>.

BRITZ, D. et al. Massive exploration of neural machine translation architectures. In: **Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2017. p. 1442–1451.

BROWN, P. F. et al. A statistical approach to machine translation. **Computational linguistics**, v. 16, n. 2, 1990.

CALLISON-BURCH, C.; KOEHN, P.; OSBORNE, M. Improved statistical machine translation using paraphrases. In: **Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics**. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006. (HLT-NAACL '06), p. 17–24. Available from Internet: <https://doi.org/10.3115/1220835.1220838>.

CASELI, H. d. M. Regras de tradução automáticas induzidas de textos paralelos envolvendo o português do brasil. **Projeto de Qualificação Doutorado) Instituto de Ciências matemáticas e Computação, USP**, 2004.

CHO, K. et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. **arXiv preprint arXiv:1406.1078**, 2014.

DOWLING, M. et al. Smt versus nmt: Preliminary comparisons for irish. In: **Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)**. [S.l.: s.n.], 2018. p. 12–20.

GALE, W. A.; CHURCH, K. W. A program for aligning sentences in bilingual corpora. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 19, n. 1, p. 75–102, mar. 1993. ISSN 0891-2017. Available from Internet: <http://dl.acm.org/citation.cfm?id=972450.972455>.

GERS, F. A.; SCHMIDHUBER, J.; CUMMINS, F. Learning to forget: Continual prediction with lstm. IET, 1999.

GOODMAN, J. A bit of progress in language modeling. **CoRR**, cs.CL/0108005, 2001. Available from Internet: <http://arxiv.org/abs/cs.CL/0108005>.

GUINOVART, X. G.; SIMOES, A. Parallel corpus-based bilingual terminology extraction. In: **CEUR Workshop Proceedings**. [S.l.: s.n.], 2009. v. 578.

HEARNE, M.; WAY, A. Statistical machine translation: A guide for linguists and translators. **Language and Linguistics Compass**, v. 5, n. 5, p. 205–226, 2011.

HUNNICUTT, S.; CARLBERGER, J. Improving word prediction using markov models and heuristic methods. **Augmentative and Alternative Communication**, Taylor & Francis, v. 17, n. 4, p. 255–264, 2001.

HUTCHINS, W. J. Machine Translation: A Brief History. **Concise history of the language sciences: From the Sumerians to the cognitivists**, p. 431 – 445, 1995. ISSN 0022-5061.

JOHNSON, M. et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. **Transactions of the Association for Computational Linguistics**, MIT Press, v. 5, p. 339–351, 2017.

KAISER, L. et al. One model to learn them all. **arXiv preprint arXiv:1706.05137**, 2017.

KALCHBRENNER, N.; BLUNSOM, P. Recurrent continuous translation models. In: **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2013. p. 1700–1709.

KISS, T.; STRUNK, J. Unsupervised multilingual sentence boundary detection. **Computational Linguistics**, MIT Press, v. 32, n. 4, p. 485–525, 2006.

Klein, G. et al. OpenNMT: Open-Source Toolkit for Neural Machine Translation. **ArXiv e-prints**, 2017.

KOEHN, P. Europarl: A parallel corpus for statistical machine translation. In: **MT summit**. [S.l.: s.n.], 2005. v. 5, p. 79–86.

KOEHN, P. **Statistical machine translation**. [S.l.]: Cambridge University Press, 2009.

KOEHN, P. et al. Moses: Open source toolkit for statistical machine translation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions**. [S.l.], 2007. p. 177–180.

KOEHN, P.; OCH, F. J.; MARCU, D. Statistical phrase-based translation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1**. [S.l.], 2003. p. 48–54.

KOEHN, P.; SCHROEDER, J. Experiments in domain adaptation for statistical machine translation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the second workshop on statistical machine translation**. [S.l.], 2007. p. 224–227.

KORS, J. A. et al. A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. **Journal of the American Medical Informatics Association**, v. 22, n. 5, p. 948–956, 2015.

LAVIE, A.; DENKOWSKI, M. J. The meteor metric for automatic evaluation of machine translation. **Machine Translation**, v. 23, n. 2, p. 105–115, Sep 2009. ISSN 1573-0573. Available from Internet: <https://doi.org/10.1007/s10590-009-9059-4>.

LAZAREV, V. S.; NAZAROVETS, S. A. Don't dismiss citations to journals not published in english. Springer, 2018.

LOPEZ, A. Statistical machine translation. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 40, n. 3, p. 8:1–8:49, aug. 2008. ISSN 0360-0300. Available from Internet: <http://doi.acm.org/10.1145/1380584.1380586>.

MARCONDES, C. H. et al. The scielo brazilian scientific journal gateway and open archives; usability of hypermedia educational e-books; building upon the mylibrary concept to better meet the information needs of college students; open archives and uk institutions; the utah digital newspapers project; examples of practical digital libraries. **D-Lib Magazine**, ERIC, v. 9, n. 3, p. n3, 2003.

MAURANEN, A.; KUJAMÄKI, P. **Translation universals: do they exist?** [S.l.]: John Benjamins Publishing, 2004.

MCENERY, T.; ZHONGHUA, X. Parallel and comparable corpora. **Corpus-Based Perspectives in Linguistics**, v. 6, p. 131, 2007.

MELO, F. R. de; MATOS, H. C. de O.; DIAS, E. R. B. Aplicação da métrica bleu para avaliação comparativa dos tradutores automáticos bing tradutor e google tradutor. **Revista e-scrita: Revista do Curso de Letras da UNIABEU**, v. 5, n. 3, p. 33–45, 2015.

MOEN, S.; ANANIADOU, T. S. S. Distributional semantics resources for biomedical text processing. **Proceedings of LBM**, p. 39–44, 2013.

NAIR, L. R.; PETER, S. D. Machine translation systems for indian languages. **International Journal of Computer Applications (0975–8887)**, Citeseer, v. 39, n. 1, 2012.

NASPRE, O. Perez-de V.; LABAKA, G. Ixa biomedical translation system at wmt16 biomedical translation task. In: **Proceedings of the First Conference on Machine Translation**. Berlin, Germany: Association for Computational Linguistics, 2016. p. 477–482. Available from Internet: <http://www.aclweb.org/anthology/W/W16/W16-2338>.

NATIONAL LIBRARY OF MEDICINE. **UMLS® Reference Manual [Internet]**. Bethesda, MD, 2009. Available from Internet: <https://www.ncbi.nlm.nih.gov/books/NBK9676/>. Accessed in: 25 sep. 2018.

NEVES, M. et al. Findings of the wmt 2018 biomedical translation shared task: Evaluation on medline test sets. In: **Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers**. Belgium, Brussels: Association for Computational Linguistics, 2018. p. 328–343. Available from Internet: <http://www.aclweb.org/anthology/W18-6403>.

NEVES, M.; YEPES, A. J.; NéVéOL, A. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In: CHAIR), N. C. C. et al. (Ed.). **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)**. Paris, France: European Language Resources Association (ELRA), 2016. ISBN 978-2-9517408-9-1.

PACKER, A. L. Scielo-a model for cooperative electronic publishing in developing countries. **D-Lib Magazine**, v. 6, n. 10, 2000.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the 40th annual meeting on association for computational linguistics**. [S.l.], 2002. p. 311–318.

PECINA, P. et al. Adaptation of machine translation for multilingual information retrieval in the medical domain. **Artificial intelligence in medicine**, Elsevier, v. 61, n. 3, p. 165–185, 2014.

POTTHAST, M. et al. Cross-language plagiarism detection. **Language Resources and Evaluation**, Springer, v. 45, n. 1, p. 45–62, 2011.

PRIETO, D. Make research-paper databases multilingual. **Nature**, v. 560, n. 7716, p. 29, 2018.

RAWAT, S.; CHANDAK, M.; CHAUHAN, N. An approach for efficient machine translation using translation memory. In: SPRINGER. **International Conference on Smart Trends for Information Technology and Computer Communications**. [S.l.], 2016. p. 285–291.

REBECHI, R. R.; ANDREETTO, M. D. As retraduções de trauer und melancholie para o português: o léxico freudiano sob o olhar da linguística de corpus. **Pandaemonium Germanicum**, v. 18, n. 26, p. 126–157, 2015.

RESNIK, P.; SMITH, N. A. The web as a parallel corpus. **Computational Linguistics**, MIT Press, v. 29, n. 3, p. 349–380, 2003.

STEINBERGER, R. et al. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. **arXiv preprint cs/0609058**, 2006.

SUGISAKI, K. Word and sentence segmentation in german: Overcoming idiosyncrasies in the use of punctuation in private communication. In: REHM, G.; DECLERCK, T. (Ed.). **Language Technologies for the Challenges of the Digital Age**. Cham: Springer International Publishing, 2018. p. 62–71. ISBN 978-3-319-73706-5.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2014. p. 3104–3112.

TERUMASA, E. Rule based machine translation combined with statistical post editor for japanese to english patent translation. In: **Proceedings of the MT Summit XI Workshop on Patent Translation**. [S.l.: s.n.], 2007. v. 11, p. 13–18.

TIEDEMANN, J. Parallel data, tools and interfaces in opus. In: CHAIR), N. C. C. et al. (Ed.). **Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)**. Istanbul, Turkey: European Language Resources Association (ELRA), 2012. ISBN 978-2-9517408-7-7.

TUBAY, B.; COSTA-JUSSÃ, M. R. Neural machine translation with the transformer and multi-source romance languages for the biomedical wmt 2018 task. In: **Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers**. Belgium, Brussels: Association for Computational Linguistics, 2018. p. 678–681. Available from Internet: <http://www.aclweb.org/anthology/W18-6450>.

VARGA, D. et al. Parallel corpora for medium density languages. **AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE**, p. 247, 2007.

VASWANI, A. et al. Tensor2tensor for neural machine translation. **arXiv preprint arXiv:1803.07416**, 2018.

VASWANI, A. et al. Attention is all you need. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2017. p. 5998–6008.

WANG, K.-f.; QIN, H.-w.; WANG, H.-x. Using parallel corpus in translation teaching [j]. **Computer-Assisted Foreign Language Education**, v. 6, 2007.

WOŁK, K.; KORŽINEK, D. Comparison and adaptation of automatic evaluation metrics for quality assessment of re-speaking. **arXiv preprint arXiv:1601.02789**, 2016.

WOŁK, K.; MARASEK, K. Neural-based machine translation for medical text domain. based on european medicines agency leaflet texts. **Procedia Computer Science**, Elsevier, v. 64, p. 2–9, 2015.

WU, C. et al. Statistical machine translation for biomedical text: are we there yet? In: AMERICAN MEDICAL INFORMATICS ASSOCIATION. **AMIA Annual Symposium Proceedings**. [S.l.], 2011. v. 2011, p. 1290.

XUAN, H.; LI, W.; TANG, G. An advanced review of hybrid machine translation (hmt). **Procedia Engineering**, Elsevier, v. 29, p. 3017–3022, 2012.

YEPES, A. J. et al. Findings of the wmt 2017 biomedical translation shared task. In: **Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 234–247. Available from Internet: <http://www.aclweb.org/anthology/W17-4719>.

YEPES, A. J.; PRIEUR-GASTON, E.; NÉVÉOL, A. Combining medline and publisher data to create parallel corpora for the automatic translation of biomedical text. **BMC bioinformatics**, BioMed Central, v. 14, n. 1, p. 146, 2013.

YU, Q.; MAX, A.; YVON, F. Aligning bilingual literary works: a pilot study. In: **Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature**. [S.l.: s.n.], 2012. p. 36–44.

ZENS, R.; OCH, F. J.; NEY, H. Phrase-based statistical machine translation. In: SPRINGER. **Annual Conference on Artificial Intelligence**. [S.l.], 2002. p. 18–32.

ZHANG, S.; LING, W.; DYER, C. Dual subtitles as parallel corpora. European Language Resources Association, 2014.