university of groningen

faculty of science and engineering

Getting the Null Subject Right with Neural Machine Translation

Author:
Federico Ferlito
Artificial Intelligence MSc

Internal supervisor:
dr J. K. Spenader
Faculty of Science and Engineering

External supervisor:
Christian Roest
UMCG

ABSTRACT

Null subjects are non overtly expressed subject pronouns found in pro-drop languages, such as Italian, Greek and Spanish. In the past, translating null subjects into a non-pro drop language, where the subject must be explicit, had shown to be problematic for older MT systems. The current state-of-the-art of MT offers many benefits compared to the previous methods, however there is limited research that investigates their quality during null-subject translation.

In this project, we quantify and compare the occurrence of the null-subject for several languages in the Europarl corpus. Next, we evaluate null subjects' translation into English, a "non pro-drop" language. We do so by training various NMT methods which are compared on their ability to generate the correct subjects during the null-subject translation, and their ability to produce quality translations. With the results, we determine the improvement compared to the previous research on the topic, explaining which mechanism allowed the models to overcome the difficulties in this task. Finally, we measure the bias of generated subjects with regard to gender, and we propose a novel method to alter the training data with the aim of reducing the bias.

CONTENTS

1 INTRODUCTION

Speakers of different languages must attend to and encode different aspects of the world in different ways in order to use their language correctly (Sapir, 1921; Slobin, 1996). On the one hand, some languages encode similar features in similar ways, and they look and sound similar to each other. Other languages share almost nothing between each other, as everything is encoded differently. As a consequence, the task of translating a language into another one becomes easier as the languages share more features. For example, all the Latin-based languages, like Italian, Spanish and Portuguese, share many properties, which makes the translation between them relatively straightforward. On the other hand, translating from a European language to an Asiatic one is a more complex task, as the languages have few commonalities.

One aspect which is encoded differently in many languages is related to the way the subject is expressed: *null-subject languages* (NSLs) are those languages that can omit the subject of a sentence (Haegeman, 1994). Translating a text from an NSL into a non-NSL, where pronouns are regularly retained, poses a challenge: pronouns in the target language have to be generated. The source language dictates how this task is solved: the majority of NSL, like Italian, allows the inference of these pronouns from the verb inflection. Other types of null-subject languages, like Chinese, differ from this as they do not have any verbal inflection: the subject must be retrieved from the discourse, and not from the grammar (d'Alessandro, 2015).

The field of *Machine Translation* (MT) has been growing recently, as many state-of-the-art advancements are being developed and many breakthroughs are being made. As the quality of such automated translations increases, also its demand is rising, since it can be used as a viable option for making translation less expensive and faster. Given this rise in popularity, such systems must be able to deliver grammatically correct translations. Therefore, when translating from a NSL into a non-NSL, for example from Italian to English, they must infer correctly the missing pronoun to generate a correct sentence.

Past studies have shown that translating from NSLs to non-NSLs can be difficult to resolve for *statistical machine translation systems* and *rule based systems* (Russo et al., 2012; Chung & Gildea, 2010). Both types of system fail to infer to right pronoun almost half of the time, resulting in grammatically wrong translations. However, there is little research that directly addresses this phenomenon using more recent methods based on Neural Machine Translation (NMT), like with the *transformer model* (Vaswani et al., 2017).

A reason for the failures of older systems during the null-subject translation relies on the fact that these models do not make use of the right morphological information in the input sentence to infer the missing pronoun. A recent study indicates how modern NMT encoders learn morphological features of the source words when these are transferable to the target language (Bisazza & Tump, 2018). In their research, Bisazza & Tump perform a fine-grained analysis of how morphological features are encoded in different parts of NMT systems, showing that some semantic features are learnt by the systems. One example is the verb conjugation. This is especially true when these features are good predictors for the target translation. For example, grammatical gender based on agreement of nouns and adjectives is learnt only if present in both source and target languages (i.e. French and Italian). If one of the target languages lack the grammatical gender, then this information is not learnt from the source language. This finding suggests that NMT systems should be able to use morphological features like verb conjugation to infer the right pronoun when going from a NSL to a non-NSL, as this is the best predictor for such a task.

In some cases, however, a sentence only encodes partial information about the subject. For example, the verb conjugations may indicate who is doing the action in some languages, but without specifying its gender. If the context does not contain other cues about this information, NMT models are forced to 'guess' the gender of the missing subject. In this case, models show the bias learnt from the data and exploiting statistical dependencies on the sentence level learned from large amounts of parallel data. For example, it's been observed that when inferring the null subject from Czech, there is a preference toward the masculine pronoun because the training data contains this bias, and the gender is not encoded in the verb (Popel, 2018). This phenomenon of *under-representation* decreases the visibility of certain social groups: in many MT systems feminine entities in a text are misrepresented as male in the translation (Frank et al., 2004; Schiebinger, 2014; Savoldi et al., 2021).

This work aims to investigate the quality of the translation of the null subjects with modern NMT systems, comparing the results of modern architectures with the previous research on the topic. The methods are compared based on the completeness and incorrectness of the translation, with regard to the null subject. We investigate several languages which differ in their grammar and in the amount of parallel data available, translating Greek, Italian, Spanish and Finnish (the NSLs) into English (a non-NSL). Finally, we investigate the influence of the bias in the training data with regard to the gender of the null subject translation, and we try to address this problem using a novel method that aims at reducing the imbalance of classes during the training of the system.

1.1 RESEARCH QUESTIONS

This study focus on the quality of the translation of NMT systems of the null-subject for different language pairs. We focus on the completeness of the translation, and its correctness. Measuring the correctness will allow us to see if there is any gender bias given by the training data, and try different approaches to reduce or remove it. With the obtained results, this work aims to find answers to the following research questions (RQs):

RQ1) Do languages show a similar frequency of null-subjects cross-linguistically?

RQ2) How well does the LSTM encoder-decoder model infer the null-subject when translating from NSLs into English?

RQ3) Does the attention mechanism improve the quality of the null-subject translation in the LSTM model?

RQ4) How well does the state-of-the-art transformer model translate the null-subjects from these languages into English compared to the LSTM architecture?

RQ5) Can we reduce the gender bias shown by the null-subject translation by balancing the training corpus of the NMT system?

To answer RQ1, we perform a quantitative analysis on the dependency parse in each sentence for different pro-drop languages using state-of-the-art dependency parsing methods. By answering Q1, we aim to measure the distribution of dropped subjects in different NSLs.

For RQ2, we train an LSTM translation model for each language pair. The performance of each system is measured using the BLEU score (Papineni et al., 2002), the STM score (Brown et al., 1993), and measuring the percentage of missing subjects, as was done in previous studies (Russo et al., 2012). Across the different null-subject languages, there is a variable amount of data, and the information about the subject that can be derived from the verb inflection is variable.

Similarly, for RQ3 we train the same LSTM models on the same language pairs, with the addition of the attention mechanism (Vaswani et al., 2017), in order to analyze its influence on the null-subject translation.

For RQ4, we train a transformer model for each language pair, gathering the same metric as RQ2 and RQ3. The parameters and hyper-parameters will be kept identical across the different language pairs and the different models of RQ2-3-4 to allow for easy comparison. Answering these questions allows us to learn to what extent NMT models can translate the dropped subject for different types of languages, and also which mechanism allows their successful translation.

For RQ5, we access the performance of the translation models using special test data-sets, in which we can measure the incorrectness of the gender of the null-subject translation. We then compare these results to the real gender distribution in order to measure the influence of the bias in the training data. We then check the effect of balancing the pronouns in the corpora, with regard to gender, and compare the performance obtained after re-training new transformer models.

Training different NMT systems for different NSLs allows us to investigate their strengths and the weaknesses, identifying what properties across languages are harder to model and to what extent the data has an influence during the training phase of such models. This can help to understand the current problems of MT and to give a direction for future research.

2 THE NULL SUBJECT

Across languages, some concepts are expressed in different ways. Therefore, when translating it's important to acknowledge these variations, otherwise the meaning could be lost or changed in the translation process. One such variation is how the subject is expressed. Before talking about this difference, let's have a quick review on the different types of subjects. We can distinguish two types of subjects in a sentence. The first type are the nominal subjects, which are expressed with a noun, and sometimes can be accompanied by modifiers. The second type are the pronominal subjects: instead of being a noun, the subject is one or more morphemes that encode the semantic features of the subject (i.e. number, gender, person). An example of a sentence with a nominal subject is shown in (1-a), while (1-b) presents one with a pronominal subject.

(1) a. John is eating.
b. I feel safe.

In the English example (1-b), the subject pronoun *'I'* is the morpheme indicating the subject. In other languages, it can be found as an affix on the verb. In the example shown in (2) the suffix *-amo* indicates the first plural person in Italian.

(2) Mangi-amo. [Italian]
We eat. [English translation]

In languages like Italian, the features of the subjects are encoded in the verb as conjugations, allowing the verb to distinguish all the person-number combinations. This property is called *rich agreement* (Taraldsen, 1980). A consequence of this property is that the use of subject pronouns is redundant. For example, both the Italian sentences in (3-a) and (3-b) are grammatically correct.

(3) a. Noi mangiamo. [Italian]
b. Mangiamo. [Italian with dropped subject]
c. We eat. [English translation]

We call a language *pro-drop* when the grammar allows pronoun dropping in all finite clauses. They are also called *Null-Subject Languages*

(NSLs). Figure 1 shows the geographical distribution of NSLs. From the map, it's clear that the majority of languages allows dropping the subject. The conditions in which this omission is allowed vary from language to language, and some attempts have been made to define its properties, like by Rizzi (1986). In the following sections, we give an overview of the different categories of NSLs.

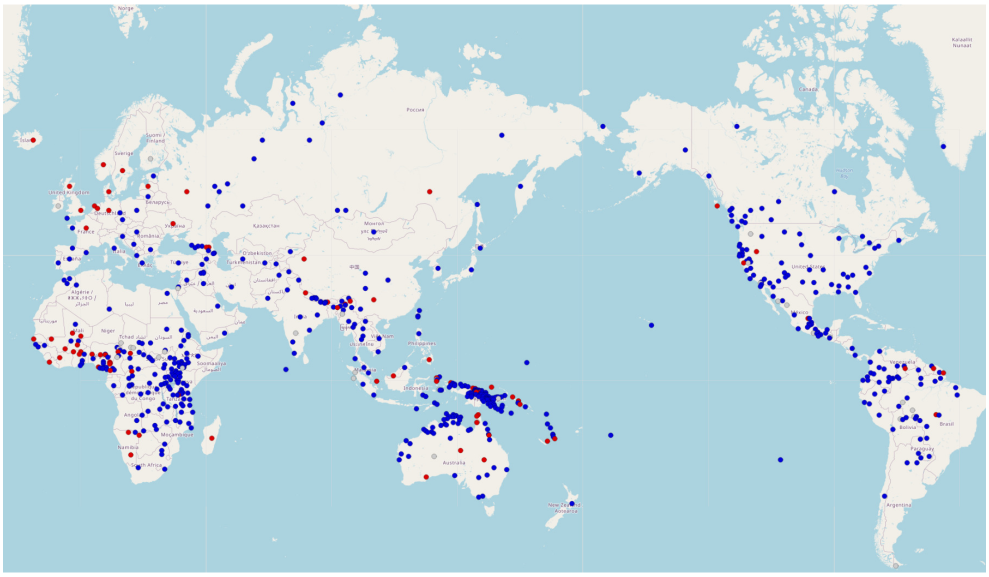

Figure 1: The map by Dryer (2013) shows a number of different types of languages based on the method they use for expressing pronominal subjects: different colours indicate NSLs (blue), non-NSLs (red), and languages that allow the null-subject only in certain cases (grey).

2.1 TYPES OF NULL-SUBJECT LANGUAGES

It's possible to classify languages into different types of NSL types: canonical, radical or partial. The division depends on the structural, lexical and morphological properties of the language. In the literature, there has been some disagreement concerning the terminology. In the next sections, we will provide all the possible nomenclatures, keeping an agnostic point of view.

2.1.1 Canonical NSLs

Also known as full or consistent NSLs [1], canonical NSLS are pro-drop languages for which a referential subject can be left unexpressed. The

[1] We will use the terms full NSL, canonical NSL, and consistent NSL more or less interchangeably.

Subject	Verb conjugation	Translation
Io	Ved-o	I see
Tu	Ved-i	You see
Lui / lei / esso	Ved-e	He / she / it see
Noi	Ved-iamo	We see
Voi	Ved-ete	You see
Loro	Ved-ono	They see

Table 1: Verb conjugation of the Italian verb "vedere" (to see). Each of the six verb form specify the number and person of the subject.

verb conjugation, which contains information about the subject, helps to understand who is the antecedent of the null-subject.

Examples of canonical NSLs are Italian, Spanish and Greek, which will be investigated in this thesis. Other languages belonging to this group are all Romance languages (except for French), Arabic, Turkish, Tamil, Berber, Hausa and Basque.

The extent of the information contained in the verb declination differs in each language. Frequently, the conjugation only specifies the number (i.e. singular or plural) and the person (i.e. 1st, 2nd or 3rd) of the subject, like in Italian, Spanish or Greek. Below in Table 1 it's possible to see the different verb forms for an Italian verb. However, other languages can include other information: Tamil, for example, specifies in the inflection also the gender and whether the subject is humanoid or not.

In this work we will focus mainly on this category of NSLs, investigating the null-subject translation from Italian and Spanish, as it was done in previous research by Russo et al. (2012). On top of that, we will investigate an additional canonical NSL, Greek.

2.1.2 Partial NSLs

In some languages, the null-subject is restricted to specific cases or syntactic structures. One such language is Finnish: the 3rd person subject pro-drop is restricted to contexts where it is a generic reference, like in (4), Otherwise, it must be specified, like in (5).

(4) Jos haluaa voittaa, täytyy harjoitella paljon [Finnish]
If one wants to win, one has to practice a lot [English]

(5) Jos hän haluaa voittaa, hänen täytyy harjoitella paljon [Finnish]
If he/she wants to win, he/she has to practice a lot [English]

With regard with the null-subject translation, the task is comparable to the canonical NSL translation, as the only difference is that the subject omission is only restricted to few cases. In practice, this restriction makes the translation easier: the subject needs to be inferred in fewer cases. The task could be more challenging when translating from a non-NSL to a partial NSL. However, this will not be covered in this thesis.

In general, partial NSLs allows omission of the subject depending on syntactic conditions. Finnish, which will be investigated in this research, is a partial NSL. Other languages that will not be covered are Russian, Icelandic Hebrew, Marathi, Assamese, and Brazilian Portuguese (Biberauer, 2008).

2.1.3 Radical NSLs

These languages, also referred to as discourse pro-drop or radical pro-drop languages, can leave both the object and the subject unexpressed, even though they lack verb inflection. Many Asiatic languages belong to this category: some of the most studied ones are Japanese, Korean, and Chinese. In this group of languages, the null subject is similar to the ellipsis, as the subject can only be inferred from previous sentences in the discourse. As example in Chinese is given in (6).

(6) 我洗过猫了。 ∅又变干净了。 [Chinese]
wo xi guo mao liao. you biangan jing liao
I washed my cat. turned clean again.
I washed my cat. It turned clean again. [English]

The translation of radical NSLs into non-NSLs poses many challenges, as the task of inferring the subject is more complex than in the other types of NSLs. Past research that aims at solving the task improved the translation providing more context, for example using *discourse-level information*, or using *extra-labelling* of the null-subjects (L. Wang et al., 2019; L. Wang et al., 2017; Chung & Gildea, 2010). As there is already extensive amount of research on this topic, this thesis will not be covering radical NSLs.

2.1.4 Non NSLs

Some languages do not allow the omission of the subject, except in few special cases. Among these, we find English, French, Swedish and Sindhi. In this category, both the pronominal and expletive pronouns have to be explicit. An example in French, with its English translation, is given below in (7):

(7) J'habite à Paris. [French]
I live in Paris. [English]

In this thesis, we will be focusing on the translation of the null-subject from canonical and partial NSL into English, a non-NSL. This latter language does not allow for subject dropping, except in few cases. These exceptions belong to special genres, for example, in diaries and spoken dialogues. Other cases where the subject can be dropped are when we want to avoid repetition, like in the example in (8):

(8) We don't believe it, but we will think about it .
We don't believe it, but will think about it .

2.2 CHALLENGES OF THE NULL-SUBJECT TRANSLATION

The translation between language becomes easier when the source and the target language share grammatical and syntactical properties, like whether or not they allow for the subject dropping. Past research on Machine Translation had shown that translation systems would fail to infer the covert subject in the majority of the cases, and also that a correct resolution can be impossible without enough context (Russo et al., 2012; Popel, 2018; L. Wang et al., 2019; L. Wang et al., 2017; Chung & Gildea, 2010). Depending on the type of NSL (chapter 2.1) of the source and the target languages, a translator may have to perform some extra steps to make the translation sound. Consider translating from a canonical NSL to a non-NSL, like shown in (9): a translator must infer and add the missing pronoun to create a grammatically correct sentence when going from Italian to English.

(9) Insieme (noi) potremo governare la galassia. [Italian]
Together we could rule the galaxy. [English]

The contextual information is essential to infer the missing pronoun. In rich morphological languages, like Italian, the verb inflection contains information about the subject. Empirically, canonical NSLs have been claimed to correlate with the rich agreement inflection of finite verbs (Rizzi, 1986). Therefore, in this group of languages, the verb form allows for the inferring of the subject. However, in radical-NSLs the null category must be inferred via the context and the discourse, as the verbs lack agreement inflection, as shown in (6). We can see how translating such implicit information poses different difficulties depending on the source language.

On one side, we have the radical NSLs, in which only the antecedent of a null-subject must be retrieved from the past, in previous sentences. When translating these to a non-NSL, it's necessary to have the complete context available to have a correct translation. On the other side, we have canonical NSL, in which the morphology of the language and the verb inflection contains information about the subject.

The issue of inferring the right pronoun does not only concern human translators. Machine Translation (MT) systems need to be provided with the right input to generate the right subject. For example, as we saw in section 2.1.3, radical NSLs must infer this information from the dialogue. Therefore, a translation system must be provided with all the sentences containing the relevant context, or it could miss some crucial details from previous sentences required to generate the right pronoun, like in the example given in (10).

(10) 我 前一会精神上太紧张。pro 现在比较 平静了。 [Chinese]
wo qian yihui jingshenshang tai jinzhang. xianzai bijiao pingjing liao
I was too nervous a while ago. be now calmer.
I was too nervous a while ago. I am now calmer. [English]

In the example, the pronoun 'I' must be inferred looking at the subject of the previous sentence. A system that splits its input based on single sentences, splitting them using the full stops, would fail to infer the right pronoun here. In contrast, when dealing with canonical NSLs the translation system must be able to focus on the right part of the verb in the same sentence to generate the right pronoun. On top of that, the system may be required to look into the past or in the future to find the verb.

Previous research in MT has been limited addressing the process of inferring and translating the null subject only from radical NSLs (L. Wang et al., 2019; L. Wang et al., 2017; Chung & Gildea, 2010). In this group of NSL, finding the right pronoun proved to be challenging even for the most recent MT models, like the *transformer model*. By contrast, the translation of the subjects in canonical NSLs have been studied only with older translation systems (Russo et al., 2012), or only indirectly when using the state-of-the-art in MT (Popel, 2018).

In this thesis, we are interested in the quality of the null-subject translation with modern MT systems based on Neural Machine Translation (NMT). We will investigate canonical NSLs that were already investigated in the past, like Italian and Spanish, but also another one, Greek. We expect that the use of modern MT methods will improve the null-subject translation for these three languages. On top of that, we will explore also a partial NSL, Finnish. We expect that the null-subject translation should be easier for this language, as the dropping of the subject is restricted to fewer cases. In the following chapters, we will give a review of the research on MT system and the translation of the null-subject in NSLs.

3 MACHINE TRANSLATION

When a person translates a piece of text into another language, many challenges arise: concepts and ideas are expressed in different ways across the world, and translators need in-depth knowledge about many different grammar books and cultures. Such tasks can be compared to the one of deciphering an encrypted message: many rules are needed to map a sentence from one form to another one, preserving the original meaning. During World War II, computers were mostly used to decode encrypted information, as they allowed to apply the many rules needed in a fast and reliable way. It was immediately clear that the same techniques could be applied to the translation of foreign languages as well. Warren Weaver, a mathematician and a pioneer in the field of automated translation, wrote a letter on the topic in 1947 (published later in 1949/1955):

> ”When I look at an article in Russian, I say: this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.” (Weaver, 1947)

This was the beginning of the field of Machine Translation (MT). A system that generates automated translation offers many benefits, as it's cost-efficient, durable and allows for a fast translation of large volumes of text. In the last decades, there has been a great progress in the field of MT, and some of the principles that were established in the early days are still valid today. Many different models have been developed, and many researchers studied in depth their benefits and their problems. Each developed system was improving on its predecessors, but all of them were characterized by the same problem, which is the translation of the null-subject.

In the following sections, we will review the history of MT systems, and for each one, we will give an overview of the research addressing the null-subject translation. Section 3.1 describes the rule-based machine translation systems. Following, section 3.2 outlines the statistical models for MT. In section 3.3, Neural Translation Models are introduced, and an overview of their benefits compared to the previous methods is given. In section 3.4 we explain segmentation methods, which allows data-driven model to handle large vocabularies and rare words. Section 3.5 describes evaluations metrics that provide rapid assessments of translation models. Here we also introduce two ways of measuring the correctness of the null-subject translation. Finally, section 3.6 introduces a common problem of data-driven methods, which is the bias in the training data. Specifically, we give an overview of gender bias, which is a problem affecting the null-subject translation.

3.1 RULE BASED MACHINE TRANSLATION

While most modern MT systems are systems based on neural language models, the field was dominated by *rule-based machine translation* (RBMTs) systems in the early years (Hutchins, 2007). These methods are based on linguistic information about the source and the target language, and they rely on extensive sets of rules and large bilingual dictionaries.

Direct machine translation was the main focus in the first years of the field. They were built on bilingual dictionary entries and some simple grammatical rules, aimed to fix simple issues like word ordering or morphology. Since this method is based on the word-level it had several problems. First, building bilingual dictionaries was expensive work. Second, these systems could not deal well with idiomatic expressions and ambiguities in the text. Lastly, this method is not scalable: adapting the model to new domains required new rules and lexicon.

Transfer-based machine translation was improving on its predecessor, adding a new level of abstraction to the translation process. It involved three steps. First, the source sentence is converted into an abstract representation, called *intermediate representation*, using its morphological and syntactical structure. Second, this representation is converted into the equivalent of the target language. Finally, the translation is generated using the first step in reverse.

Interlingua Machine translation was a popular research trend between the 1980s and 1990s, which replaced the intermediate representations of the transfer-based method with an *interlingua*. This new approach attempted to represent the meaning of sentences in an abstract way, independently from the source and target language. It differed from transfer-based methods, which required a unique representation for each language.

All the rule-based methods mentioned above consist of a set of rules which operates either directly at the word-level, or they operate via an abstract representation. Figure 2 illustrate the different levels at which the different systems work.

When two languages are very close to each other, the translation between them requires fewer rules, as their grammars are comparable and the mapping is very straightforward. When designing a rule-based

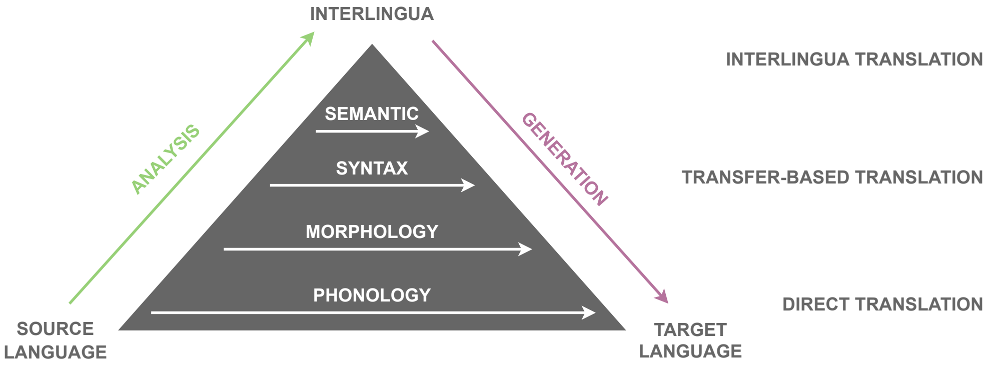

Figure 2: Bernard Vauquois' pyramid showing different depths of representation: interlingual machine translation is at the top, followed by transfer-based, then direct translation.

system, it's crucial to take into account the many differences between languages. A clear example is the allowance of the null-subject: when translating from a NSL to a non-NSL, these systems require some extra rule to generate the missing subject (Russo et al., 2012).

In a recent paper, Russo et al. investigated the quality of the translation of the null-subject using a transfer-based MT system, called *ITS-2* (2012). They tested the system on a special corpus, where every sentence dropped the subject in the source language. They manually extracted this test set from the training corpora. Their experiments show that the rate of missing pronouns in the translations, shown in Table 2, is considerable for such a system: when translating from Italian to French, 43.61 % of the subjects pronouns were missing, and an additional 9.6 % were incorrect. Similarly, when translating from Spanish to French, 48.78 % were missing, and another 6.93 % were wrongly translated.

Language pair	Correct	Incorrect	Missing
IT ->FR	46.78 %	9.6%	43.61%
ES ->FR	44.28%	6.93%	48.78%

Table 2: Results of the rule-based MT system, ITS-2, tested by Russo et al. (2012)

The next section introduces to Statistical Machine Translation, which, compared to the methods described so far, offers many benefits for the resolution of the null-subject translation.

3.2 STATISTICAL MACHINE TRANSLATION

The biggest limit in the early years of machine translation was the formalization of the many rules required to develop rule-based system. This limitation motivated the need for *data-driven methods*, which could translate from past examples.

In the late 1980s, *statistical methods* were applied for the first time to speech recognition. Following their success, IBM Research applied the same mathematical concepts for modelling the translation task with a system named *Candide* (Berger et al., 1994). The system was groundbreaking, as it only relied on probability models learnt from bilingual text data. However, these systems had important limitations, as the computers were not powerful enough, and they required a large amount of bilingual text data, which was hard to create or find. It was not until the 2000ss that these mathematical concepts became dominant in the field of machine translation. Many factors contributed to the increase in the popularity of Statistical Machine Translation (SMT) systems. *Parallel corpora*, which are datasets of texts with their aligned translations, became publicly available. One example is the Europarl corpus (Koehn, 2005). Softwares and libraries were made open source. Among them, the *Moses* system (Koehn et al., 2007), became the most popular toolkit for MT research.

A sentence in a source language s, can be translated in many ways in a target language t. In SMT, the main assumption is that every target sentence t can be a possible translation of s. We give to each pair of sentences (s, t) a number $P(t|s)$, which we can interpret as the probability of producing a translation t when presented with s. We can then use Bayes' theorem to write:

$$P(t|s) = \frac{P(t)P(s|t)}{P(t)} \tag{1}$$

Since the denominator $P(t)$ is independent from the product $P(t)P(s|t)$, finding the right translation involves making the product $P(t)P(s|t)$ as large as possible. This leads us to the equation:

$$\hat{t} = \underset{t}{argmax} P(t)P(s|t) \tag{2}$$

The equation above summaries the three main components of SMT: the language model probability $P(t)$, the translational model probability $P(s|t)$, and designing a simple and efficient way to maximize their product (Brown et al., 1993).

Many variants of SMT systems have been developed, depending upon how translation is modelled. Initially, SMT systems like the one of Berger et al. (1994) worked by splitting sentences into words, hence the name *word-based* SMT. One of the main limitations of such a system is that the way in which each word was mapped into the target language. Usually, a translated sentence has a different number of words compared to the original language. For example, the Italian word 'Mangiamo' means 'We eat' in English. While word-based SMT systems would manage to translate the word from Italian into English, the opposite was not possible: these systems were not able to produce many-to-one and many-to-many mappings. *Phrase-based* systems solved many of the problems of its predecessor (Koehn et al., 2003). Instead of using single words, groups of consecutive words, called n-grams, were considered as single units. This simple solution improved the quality of the translations noticeably when using units of three consecutive words.

Another less popular approach in SMT research had focused on the use of linguistically motivated models, which include syntactical and structural information in the model (Yamada & Knight, 2001; Imamura, 2002). In *syntax-based* systems, groups of consecutive words are considered as a single unit only if they were *constituents*, and they belonged to the same sub-tree in a syntax tree. While these systems could better handle better languages with very different syntax and word-ordering, this restriction proved to be harmful to the quality of the translations.

As the SMT systems only rely on parallel corpora to model the translation process, syntactical and structural differences between the source and the target languages do not need to have pre-programmed rules to be translated anymore. This has shown to be beneficial when the languages differ in the ways they allow the null-subject. Compared to RBMT systems, statistical methods have a higher rate of correctly translated personal pronouns when going from a NSL to a non-NSL. Research on this topic compared the performance of *Moses* with a rule-based system: the results, shown in Table 3, shows that the former made fewer errors when translating sentences with null-subjects (Russo et al., 2012). When the SMT model was translating from Spanish to French, the translation would lack 33 % of the subjects in the target language, and an additional 2.21 % of generated subjects were incorrect. Similarly, when translating from Italian to French the translations would show 33.81 % of missing subjects and 5.18 % of wrong ones. This experiment showed that SMT systems improved the translation when inferring the dropped subjects in 14.22 % for Italian sentences and 16.72 % for Spanish ones. While this is a significant improvement, the error rates are still considerable.

Language	Correct	Incorrect	Missing
IT ->FR	61.00 %	5.18 %	33.81%
ES ->FR	64.78 %	2.21 %	33.00 %

Table 3: Results of the Statistical MT system, Moses, tested by Russo et al. (2012)

The next section introduces to Neural Machine Translation, which, compared to the methods described so far, offers many benefits for the resolution of the null-subject translation.

3.3 NEURAL MACHINE TRANSLATION

Inspired by the networks of neurons in the animal brain, models of artificial neural networks (ANN) have been studied since the 1950s (Rosenblatt, 1957). During the early years of research in this field, the first neural network models were applied also for machine translation (Allen, 1987; Neco & Forcada, 1997; Waibel et al., 1991). However, the research in this area was abandoned for several decades, as the computational complexity of such models far exceeded the resources and computers available at the time. Even if their performance was far from remarkable, they have a striking similarity to the current MT approaches. Starting from 2006, a new wave of neural network research allowed these models to gain a lot of attention, and since then they have been successful in a lot of areas, like image recognition and speech recognition. In these years, the first neural networks were used together with SMT systems as pre or post-processing steps, for example for providing translation tables (Schwenk, 2012; R. Wang et al., 2014) or for reordering the words (Kanouchi et al., 2016; Li et al., 2014).

Only in recent years, have advancements that processing powers allowed for neural machine translation to completely replace SMT. The first models employing pure neural machine translation made use of convolutional neural networks (Kalchbrenner & Blunsom, 2013) and sequence-to-sequence models (Cho et al., 2014; Sutskever et al., 2014). While these models achieved good translation quality with short sentences, they performed poorly with longer sequences. The addition of some refinements, like the attention mechanism, byte-pair-encoding, and back-translation, allowed neural machine translation to become state-of-the-art.

The following sections give an overview of the different NMT methods developed in recent years. First, section 3.3.1 describe the first neural architecture proposed for MT, the encode-decoder model. Section 3.3.2 describes a more sophisticated model which is more capable of learning the long-term dependencies in long texts, the Long-Short-Term-Memory network. Section 3.3.3 introduces the attention mechanism, which improved NMT models allowing them to focus on the right part of the input sentence to generate each output word. Finally, section 3.3.4 describes the state-of-the-art in MT, the transformer architecture, which in the last year was the most popular choice in the field.

3.3.1 Encoder-decoder model

Neural networks for mapping a variable-length sequence to another variable-length sequence are called sequence-to-sequence models, or encoder-decoder. Figure 3 illustrates the system.

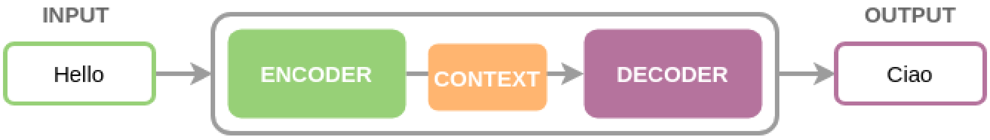

Figure 3: The architecture of an encoder-decoder (or sequence-to-sequence) model

This architecture was first proposed by Cho et al. (2014), and shortly after by Sutskever et al. (2014). The idea is very simple: first, an encoder (or reader) processes the input sequence using an *recurrent neural network* (RNN). The encoder emits a vector called *context* as a function of its final hidden state. Then, the decoder (or writer) is conditioned on that fixed-length vector to generate the output sequence using a vector-to-sequence RNN. The computations of this model are illustrated in Figure 4.

When translating a sentence, we want to output a prediction that depends on the whole input sequence. For example, the correct interpretation of a single word may depend on the word following it. If there is more than one possible translation for a word, we may have to look far in the past or in the future to disambiguate them. To address this need, the encoder implements a *bidirectional RNN* (Schuster & Paliwal, 1997). As the name suggests, bidirectional RNNs combine an RNN that move forward in the input sequence, with another RNN that move in the opposite direction, starting from the end of the sequence. This allows computing a hidden representation that depends on both past and future states.

As the context is captured in a fixed-size vector, the decoder is simply a *vector-to-sequence* RNN. The vector is given as input to the RNN nodes at each time state, for which it generates an output token. This is passed together with the context in the next time step.

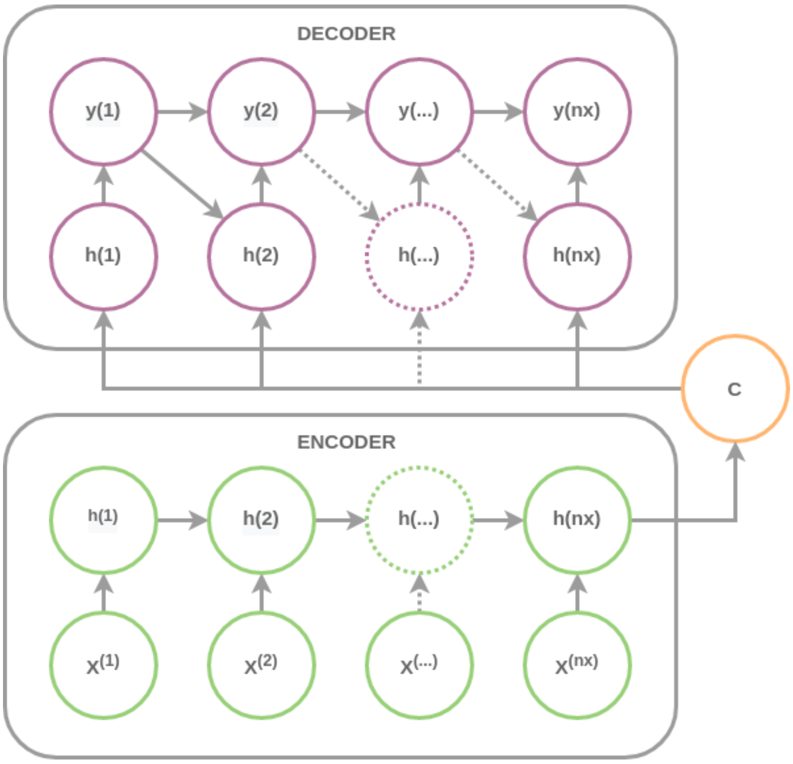

Figure 4: Example of encoder-decoder RNN architecture for learning to generate an input sequence $(y^1, y^2, ..., y^{n_y})$ given an input sequence $(x^1, x^2, ..., x^{n_x})$. It is composed of an encoder RNN that reads the input sequence and of a decoder that generate the output sequence. The final state of the hidden node of the encoder RNN is used to compute the context vector C, which represents a semantic summary of the input sequence

3.3.2 Long short term memory network

When we predict the next word in a sentence the context plays a crucial role, and usually, we need to look at previous words or sentences to make a confident prediction. While in most cases the cues to infer a word are in the recent past, in some cases the gap between the relevant cues and the place where they are needed can be very big, for example, in several previous sentences. Although the RNN layers of the encoder-decoder model can theoretically use the information from the distant past inputs to generate the next words, in practice they don't (Hochreiter et al., 2001). The problem is related to the way the model is optimized: most networks are trained using optimization algorithms based on *gradient descent*, which tunes the model's parameters computing a gradient on the errors produced. However, when training deeper models, the gradient is computed through many layers, and it becomes unstable. When the derivatives are too large and the gradient increases exponentially, we refer to it as the *exploding gradient problem*. When the derivatives are too small and the gradient decrease exponentially, we call it *vanishing gradient problem*. In RNN models, these two problems become even worse since the gradient must propagate also through different time steps.

Long-short-term-memory (LSTM) networks are a special type of RNN, which are more capable of learning long-term dependencies. They were first proposed by Hochreiter & Schmidhuber (1997). The recurrent module in this type of network is called the *cell state*. In the cell, the information from the previous states is modulated with different gates. Each gate has a different function: the *forget gate* decides which information from the previous state must be removed in the current state; the *input gate* decides what information of the new input must be retained; the *output gate* generates the output of the cell, and decide which information are sent to the next cell state. The whole system is illustrated in Figure 5.

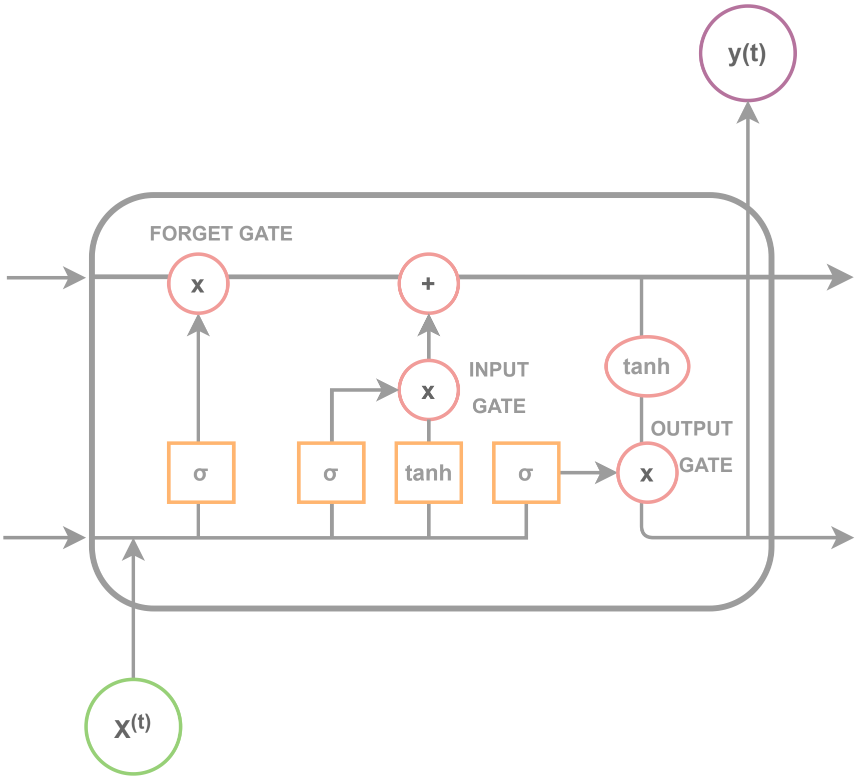

Figure 5: Example of cell state of a LSTM network. The upper arrow going through the diagram contains the state information from the previous cell, and it's propagated to the next cell. The circles (x) are point-wise operations that represent the three gates of the cell. The rectangles are neural networks layers.

Encoder-decoder systems implementing LSTM networks have been shown to improve the quality of the translation compared to the same systems implementing simple RNNs. The specialized gates aimed at forgetting old information and remembering new relevant features make them a natural choice for MT. The system was further improved using an *attention mechanism*, which learns which elements of the context vector to associate to elements of the output sequence. The next section describes it in detail.

3.3.3 Attention mechanism

By training a sufficiently large RNN (or LSTM) model for long enough, encoder-decoder models can capture the semantic details of long sentences into a fixed-sized representation, as demonstrated by Cho et al. (2014) and Sutskever et al. (2014). When the input sentences become longer, however, capturing this information becomes less trivial. This phenomenon is called the *bottleneck problem*. A more efficient approach is to produce translated words one at a time, each time focusing on a different part of the input sentence to gather the required details. This is exactly the idea proposed by Bahdanau et al. (2016), implemented in what is called an *attention mechanism*, which is illustrated in Figure 6.

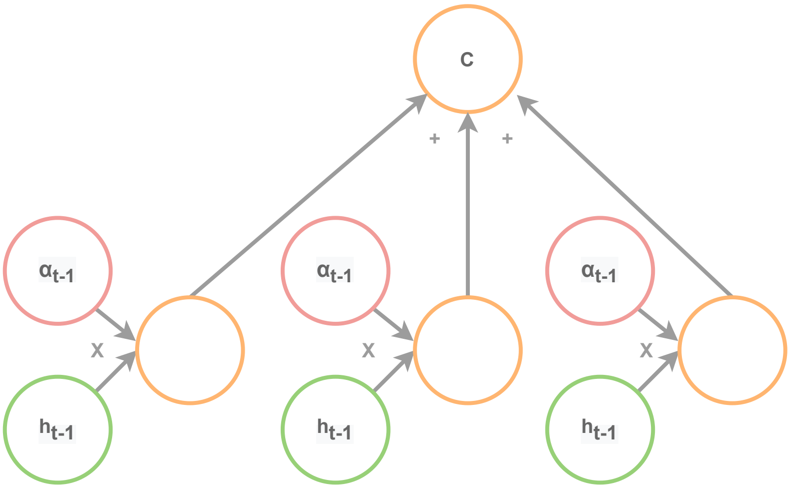

Figure 6: Illustration of the attention mechanism. The context vector C if formed by taking a weighted average of features h_t with weights α_t. The weights range in the interval [0,1], and are intended to concentrate the sum around a single h_t for each time step.

The *attention mechanism* creates a context vector C concatenating all the hidden states produced by the encoder RNN. To further improve this representation, each hidden state is weighted to give higher importance to the words that are most useful in predicting the next output word. The *attention weights* are computed using the last hidden state of the decoder: this state encapsulates all the words decoded so far, and it contains the information required to predict the next translated token. Following this intuition, the weight for each word is computed as a similarity score between the encoded representation of that word and the last state of the decoder. First, for each encoder hidden state $h(i)$, we compute a similarity score e_{ij} matching this vector to the last decoder state $d(j-1)$:

$$e_{ij} = a(h(i), d(j-1)) \tag{3}$$

The function can be any arbitrary function. In the original setup, a single layer feed-forward neural network was used (Bahdanau et al., 2016). Second, we use the SoftMax function to normalize the computed scores e, highlighting the largest ones and reducing the smallest ones, for each of the element of the input of length T_x:

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{4}$$

Finally, we can use these computed weights α on the hidden representations of the encoder to get out desired context vector:

$$C = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{5}$$

The resulting context vector C is used, together with the previous state of the decoder, to generate the next output word. Equation 5 is illustrated in Figure 6.

One of the main advantages of the attention mechanism is that the model learns by itself the alignments of the words between the source language and the target language. It also makes it possible to inspect the attention weights and the corresponding associations between the input and the generated output, as shown in Figure 7.

The matrix shows the attention weight for the following translation:

(11) Esta es mi situacion economica en este momento. [Spanish]
This is my economic situation in the time. [Generated]
This is my economic situation at the moment. [Reference]

From the matrix, it's possible to see how all the words are aligned in a monotonic way, except for the 2 words 'economic situation', which are mapped in a non-monotonic way and the order of the words is reversed from Spanish.

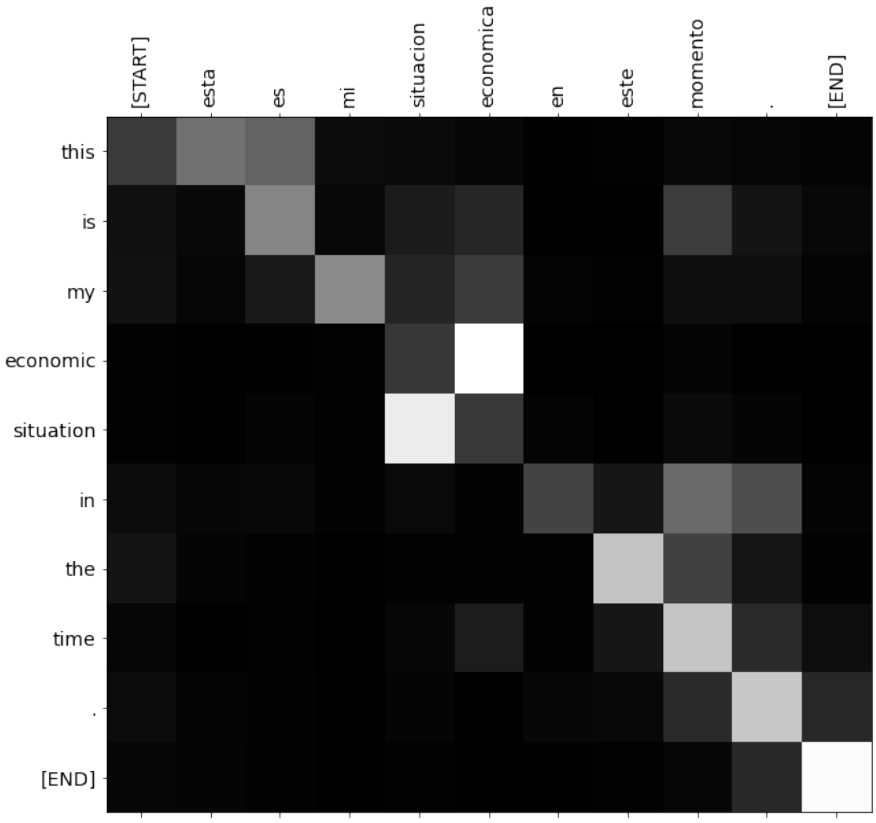

Figure 7: The attention weights. The x and y axis correspond to the words in the source sentence (Spanish) and the generated translation (English). Each pixel shows the weight α_{ij} between the j-th source word and the i-th translated word (see Equation 4). The whiter pixels indicate input tokens which are more influential for predicting some output words.

3.3.4 Transformer architecture

Encoder-decoder models quickly became the state-of-the-art of every NLP application, completely substituting SMT system. However, while this architecture was able to produce good and fluent translations, it had some limitations. One of the biggest problems of such a system is its sequential nature: each output word is generated linearly using the previous hidden states. This creates a bottleneck in the computational time, as it was not possible to parallelize the model. The second main issue with these systems is the way they handle long-distance dependencies: while the use of LSTM cells improved the handling of the memory using specialized gates, remembering things for a long period of time was still challenging.

The introduction of the *transformer* model had a huge impact in the MT field, as this new neural-network was able to produce better translations without using recurrent connections (Vaswani et al., 2017). A new mechanism, called *self-attention*, allowed for the model to be more parallelisable and to be trained in a faster way. An illustration of the model is shown in Figure 8

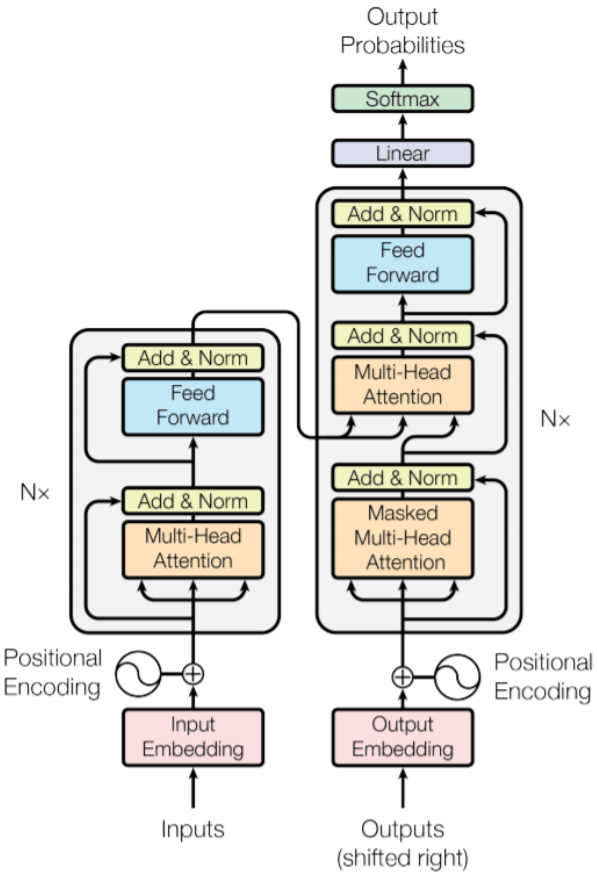

Figure 8: The transformer model architecture, as illustraded by Vaswani et al. (2017)

The architecture is made of two parts: a stack of encoders, and a stack of decoders. Each encoder can be broken down into two parts: first, the *self-attention layer* receives the input. This encodes specific words looking at other words in the input: it computes three vectors for each input word, called *query*, *key* and *value*, using three matrices created during the training. Then, we use the query vector of a word and we match it with all the key vectors of the words in the sentence, using the dot product. This operation returns us a vector of scores for that word, which we can then use to scale the value vectors. This helps to identify how important each input is in relation to the generated words. We do this for all the value vectors, which are then summed and passed to the second layer of the encoder.

The output is then sent to the second layer, a *feed forward neural network*, which is used independently for each encoding. This layer halps to normalize the values during the forward pass in the model. To take into account the order of the input sequence, the model adds to input an extra vector, the *positional encoding*. This vector represents the relative position of each word, following a pattern that the model learns. This vector also allows the model to scale up to inputs bigger than the ones present in the training data.

The decoders have the same two layers of the encoder, but between them, there is an attention layer that helps the decoder to focus on the attention matrix which contains the contextual relevant information for generating the current target output word.

The transformer model solved some of the biggest issues of the previous NMT model, achieving a spectacular quality in the translation task. However, all the NMT methods aforementioned have to deal with the same problem: the computational time scales up linearly with the size of the language vocabulary. The next section describes some methods to deal with this issue, which involves the segmentation of words into smaller units.

3.4 SEGMENTATION FOR MT

One of the biggest obstacles for data-driven MT methods is *data sparcity*. In corpora, the distribution of words is largely skewed. For example, in the corpus of the parliamentary proceedings of the European Parliament, the most frequent word, "*the*", accounts for 6.5 % of the 30-million words. On the other, there is are a plethora of words that do not occur frequently: 33,447 words only occur once. A mathematical law, *Zips' law*, describes this phenomenon, stating that frequencies of words depend on their ranking in the frequency table. This means that the most frequent word occurs twice as many times as the second one, three times as often as the third one, and so on until the last frequent word. Therefore, the majority of the words in a corpus will be rare words. Many of these words can be new words (i.e. *retweeting*, *e-bike*) or names (i.e. *Facebook*, *Covid-19*).

Today, the most common approach to handle rare words and new words is to break all the input words into smaller parts, called *subword units*. It's a similar approach to the methods used in SMT to handle compound words (i.e. *website* → *web+site*) and morphology (*unfollows* → *un* + *follow* + *s*). Section 3.4.1 describes the most popular algorithm for such task, *byte-pair-encoding*. Following this explanation, a more linguistically-inspired method, *morfessor*, is presented in section 3.4.2.

3.4.1 Byte-pair encoding (BPE)

A popular method for creating an inventory of subword units is *byte-pair encoding* (Sennrich et al., 2016). This method creates a vocabulary using a parallel corpus. First, the words are split into individual characters. Where the splitting occurs, a special character is placed. Then, we merge the most frequent pair of characters. We repeat this step a fixed number of times. Each of these steps increases the vocabulary size by one, expanding the initial inventory of characters.

Consider the following toy corpus, where we mirror the behaviour of the algorithm:

(12) t h e ␣ f a t ␣ c a t ␣ s e e s ␣ t h e ␣ t h i n ␣ t h i r s t y ␣ b a t

The most frequent pair of characters here is *th*, occurring 4 times. So we merge these into one single token:

(13) th e ␣ f a t ␣ c a t ␣ s e e s ␣ th e ␣ th i n ␣ b a t ␣ th i r s t y

Following, the most frequent pair is *at*, so we create this token next:

(14) th e ␣ f at ␣ c at ␣ s e e s ␣ th e ␣ th i n ␣ b at ␣ th i r s t y

Next, the most frequent token pair is *the*, so we merge this tokens into a full word:

(15) the ␣ f at ␣ c at ␣ s e e s ␣ the ␣ th i n ␣ b at ␣ th i r s t y

The algorithm starts grouping single characters combinations, and then it joins frequent words. At the end of the process, the most occurring words in the corpus will be present in the BPE vocabulary, and the rare words will be split into sub-words. A common practice is to run the algorithm on a concatenation of the source and target corpus: this helps with the transliteration of names (Sennrich et al., 2016).

In our experiments on the null-subject translation, we preprocess the different corpora using BPE to restrict the vocabulary of the studied languages. Reducing the vocabulary to a smaller number of subwords allows us to decrease the computational time of the models.

While BPE generates subwords based on the frequencies of morphemes, some languages with rich-morphology may benefit from using more linguistically-inspired segmentation methods. In the next section, we illustrate one such algorithm, morfessor, that allows us to improve the subword creation for certain languages.

3.4.2 Morfessor

Some segmentation methods put more emphasis on producing linguistically correct sub-words units. Morphologically inspired segmentation algorithms can give better results than straightforward methods like BPE, especially when analyzing highly-inflecting languages, like Finnish, Turkish, or Estonian. (Banerjee & Bhattacharyya, 2018). One such popular algorithm is called *morfessor* which focuses on finding morphemes using an unsupervised algorithm (Creutz et al., 2005).

Morfessor builds a probabilistic model of a language M, which consists of a morph vocabulary and a grammar. The goal of the model is to generate a concise segmentation of the corpus. The training involves finding the *Maximum a posteriori* (MAP) estimate for the parameters:

$$\underset{M}{argmax} P(M|corpus) = \underset{M}{argmax} P(corpus|M) \cdot P(M) \tag{6}$$

where the probability of the model of language $P(M)$ is:

$$P(M) = P(vocabulary, grammar) \tag{7}$$

The MAP estimate consists of two parts: the model of the language $P(M)$, and the maximum likelihood estimate $P(corpus|M)$, which is conditioned on the given model of language.

This probabilistic approach does not directly search for morphemes in the words: they are naturally induced from the unsupervised training process. This makes Morfessor a general tool that can be used in many different languages. During each training epoch, every possible split is considered for each word. The best splitting is selected depending on the associated cost. The training continues until the cost gain is lower than a certain threshold. However, the stopping criterion can also be a maximum number of epochs or an approximate number of split operations.

In our experiments, we will use morfessor as a preprocessing step on the Finnish corpus, as the generation of subword results improved for such highly-inflecting language. The use of an appropriate segmentation method for each language allows getting improved translation quality, without sacrificing the vocabulary size. This in turn permits us to focus on addressing the performance of NMT models in the null-subject translation.

The evaluation of MT models requires a fast and efficient way to measure the quality of the output. In general, this is accomplished by comparing a generated translation to a human-generated one. However, to investigate a specific aspect of the MT model, for example, the evaluation of the null-subject translation, alternative scoring metrics are more suited. The next section describes the evaluations metrics that are used in our experiment.

3.5 EVALUATION METRICS FOR MT

When we evaluate MT systems, we put more trust in the judgement of human evaluators, who look at the generated translations of several models and determine an overall score for each one, analyzing sentence by sentence. This method has a major disadvantage: it's time consuming, and very often evaluators need to be trained.

In MT research, evaluations need to be done frequently: models contain a lot of hyper-parameters to be tuned, and many variations of the same model must be trained in a fast fashion to find the best possible model. This motivates the implementation of automatic methods for accessing the quality of MT outputs. Usually, this is done by comparing a human-generated reference translation to the output generated from a model.

The next sections illustrate the automatic evaluation scores that are used in our experiments. Section 3.5.1 describes BLEU, one of the most popular metrics in MT research. Section 3.5.2 describes a more linguistically motivated metric, STM, which aims at checking the syntactical quality of the translations. Finally, section 3.5.3 describes a metric designed to measure directly the quality of the null-subject translation.

3.5.1 BLEU

BLEU (bilingual evaluation understudy) is one of the most common automatic evaluation metrics. It is based on a simple assumption: ”The closer a machine translation is to a professional human translation, the better it is” (Papineni et al., 2002). This metric has shown to correlate well with human judgements in ranking MT systems (Doddington, 2002; Papineni et al., 2002).

For every generated translation, BLEU is computed using the number of words that overlaps with the reference translation, and also the overlap with a higher order of *n*-grams matches. It includes in the score a brevity penalty, which is based on the ratio between the number of words in the generated translation and the reference one.

While BLEU score had shown to correlate well with human judgment over large test sets, it does not do so well at the sentence level

(Blatz et al., 2004). One of the biggest reasons is that this metric considers synonyms as wrong words. Accurate translations which use different words from their reference sentences would therefore score very low. Moreover, it's not linguistically motivated: generated sentences with incorrect syntax will have an inflated score if they happen to contain the correct words.

Despite its drawbacks, BLEU it's also the most standard MT evaluation metric. We use BLEU as a form of efficient bench-marking, as we can easily access the quality of the translation on a large test set.

3.5.2 STM

Most of the automatic methods for accessing the quality of MT systems are based on the same type of information, the *n*-gram sub-sequences of the hypothesis translation. While in practice this has shown to work well, this type of feature does not capture the grammatically of the sentence. So it's possible that translations that contain roughly the correct words get high scores, even though they do not form a coherent sentence. The *subtree metric* (STM) addresses this problem, comparing differences between the syntactic relations between words, generated as trees, like the ones shown in Figure (16). The syntax trees of the reference and the output translation are compared, giving a penalty to outputs with different syntactic structures. This allows quality evaluation beyond the word-level, as it relies on both lexical and syntactical information. Consider the following examples:

(16) Reference: I eat at home
Hypothesis 1: I snack at home
Hypothesis 2: Eat at home

If we use BLEU to evaluate the two generated sentences, the first hypothesis scores 0.707, while the second hypothesis is 0.72. The latter one scores higher than the first one, as there are more bi-grams in common with the reference translation. However, the evaluation is incorrect, as the sentence is lacking the subject. We can solve this problem by taking into account the syntactic similarity of the sentences. Figure 9 shows syntactic trees of the examples shown in (16).

It's clear from the syntactic trees that the first hypothesis has the same structure as the reference sentence, while the other hypothesis has a very different one. If we compute the STM metric on these syntax trees, we get a score of 1 for the first hypothesis and 0.39 for the second one.

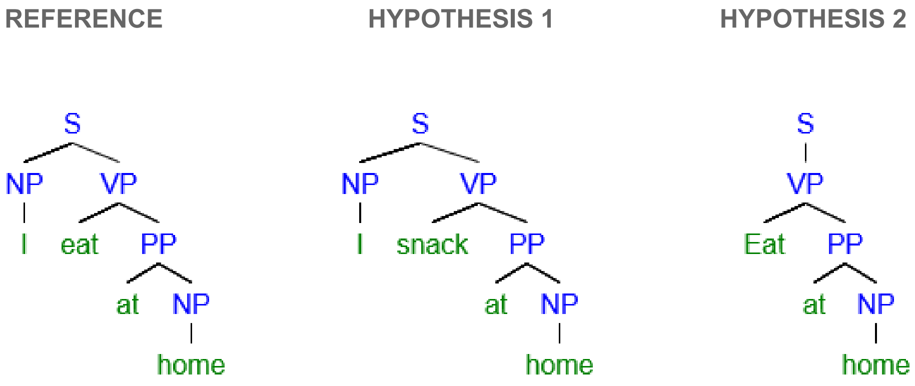

Figure 9: Syntax trees of some example sentence. The tree on the left shows the tree of the reference translation; the one in the middle shows the tree of a correct translation; the right one shows the tree of a wrong translation, which does not match with the reference.

The advantages of the STM make it a good candidate for accessing the quality of null-subject translations: when translating a sentence with a null-subject into a non-NSL, this metric reflects the integrity of the generated sentence, which indicates how well the pronominal subjects are inferred. If the translation does not contain the subject, the syntactic tree of the sentence will be very different from the reference one, resulting in a lower score. We will use STM for comparing the quality of the translation of the null subject across different MT systems, as it offers more insights than the BLEU score alone on the null-subject translation.

3.5.3 Percentage of subjects translated

When we evaluate the translation of null-subject sentences in cases where the target language is a non-NSL, common metrics like BLEU or STM do not directly tell us about the quality of the subject translated, as they compute scores over a text as a whole. For this reason, we include as an additional metric: the percentage of subjects correctly translated in the target language. This score allows us to evaluate the completeness of the translation, with regard to the syntactical structure. Additionally, we measure the percentage of pronouns with the correct gender, so that we have a score to measure the correctness of the null-subject translation. These metrics have been working well in previous research (Russo et al., 2012). To compute this metric automatically, we use a *dependency parsing* (DP) algorithm, which allows us to analyze the grammatical structure of sentences based on the relations of their words. The result of such analysis is a tree structure $T = (V, A)$, where V is a set of nodes, representing each word (including the punctuation), and A is a set of directed arcs, representing the dependencies and the grammar relationship between the elements of V. An example of such parsing is shown in 10.

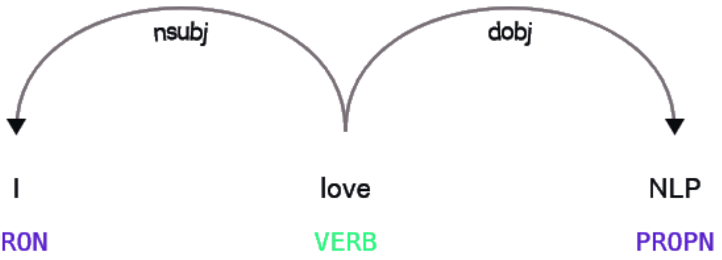

Figure 10: Example of dependency parse tree of the sentence "I love NLP". The tree is generated using the Dependency Vizualizer developed by Spacy

In the illustrated tree, the verb 'love' is at the root, and the arrows coming from it represents the dependencies from the root to the other words. For example, the word 'I' is the nominal subject (*nsubj*) of the verb, while the word 'NLP' is the direct object (*dobj*) of it. We use such a parsing tree to measure the percentage of verbs which are accompanied by a subject, searching in the tree the relevant arc connections between verbs and subjects [1]. For example, in (17) the subject-verb pairs found in a sentence are shown. On the other hand, when a verb is lacking the subject, the analysis returns only the verb, as shown in the example in (18).

(17) I do not believe you, but I will vote for you anyway
(I (do believe)) (I (will vote))

(18) I do not believe you, but will vote for you anyway
(I (do believe)) ((will vote))

We use such analysis to measure the performance of the different models in the null-subject translation task. We first run this analysis on the source language corpus to extract a subset of sentences where there are verbs with a null-subject. Then, we use the trained translation model to translate these extracted sentences, and we run another analysis on the generated output. We compute our special metric, the percentage of verbs with the missing subject, as:

$$\%\ \textit{missing subject} = \frac{\#\ \textit{tuples without subject}}{\#\ \textit{tuples with subject} + \#\ \textit{tuples without subject}}$$

[1] We use the library grammaregex to search in the parsed tree https://github.com/krzysiekfonal/grammaregex

Moreover, we can compute the percentage of subjects which are translated differently comparing the analysis on the translated sentence with the one of the reference one, as:

$$\%\ wrong\ subject = \frac{\#\ different\ subjects}{\#\ same\ subjects + \#\ different\ subjects}$$

The aforementioned evaluation metrics allows us to compare how different NMT models perform when they have to infer the null-subject in the target language. Another aspect to take into account when evaluating a data-driven model is the quality of the training corpus. If such data contains some form of bias, this will be reflected in the generated output. For example, it has been observed that when the NMT models were translating from non-gendered to gendered language, certain biases related to the gender of the subject would emerge. In the next section, we give an overview of the algorithmic bias of MT models, and how they can cause errors in the null-subject translation.

3.6 BIAS IN MT

Interest in understanding and mitigating biases in MT systems is steadily growing in the latest year. Recent studies showed how gender disparities are affecting these technologies. The problem was first highlighted by Schiebinger (2014), who criticized the phenomenon of masculine default after translating several interviews with MT systems. In her article, despite several feminine mentions being present in the text, she was repeatedly referred to with masculine pronouns. As such systems often translate texts sentence by sentence, they show this gender bias when the correct pronouns cannot be inferred from the local context.

Bias in NMT models reflects disparities in the data. Asymmetries in the use of the pronouns in the training corpus are learnt by the MT systems and rewarded during their optimization. This motivates the need for a careful data curation before training these models (Bender et al., 2021; Hutchinson et al., 2021). This problem becomes evident when the source sentence contains a null-subject, as the system must infer the right pronoun. If the context does not contain any gender specification, the translation is going to always give a masculine subject (i.e. *'he'*), as observed in previous research (Popel, 2018). An example of such a case is shown in (19).

(19) Ha lavorato molto. [Italian]
He/She worked a lot. [English]

In this Italian sentence, the verb conjugation helps to discriminate the person and the number of the subject pronoun, but not the gender. In cases like this one, an MT system would return the masculine pronoun 'he' as the default answer when resolving the null-subject. This problem does not occur when the context contains gender information in other parts of the sentence, for example in gendered adjectives or proper nouns, like in the example in (20).

(20) Maria è stanca. Ha lavorato molto. [Italian]
Maria is tired. She worked a lot. [English]

In this Italian example, the first sentence contains information about the gender of the subject: the feminine proper name "Maria". When an MT system must infer the null-subject of the second sentence, it will have enough context information to generate the right pronoun only when provided with both sentences together. Some methods that address the null-subject inference in radical NSL, like Chinese, improve the quality of these translations providing discourse-level information (Chung & Gildea, 2010) .

As the imbalance of different classes has been known to cause undesired biases and severe degradation of the performance (Johnson & Khoshgoftaar, 2019), for example when the wrong pronouns are generated, we measure the effect of balancing the use of pronouns expressing gender in the training corpus, with the aim of learning a distribution of pronouns that is less skewed during the translation of null-subjects, thus increasing the quality of the translation and the generalization of the model. A popular method for balancing the classes of the dataset implies down-sampling the larger ones, or up-sampling the smaller one until a more balanced distribution is reached (He & Ma, 2013). The down-sampling technique is suitable for the dataset where the underrepresented class is large enough for the model to learn the distribution of features. On the other hand, up-sampling the smaller class can lead to generalization errors, as the duplicated sentences can cause the model to over-fitting. Therefore, only the former technique will be used in our experiments.

4 CORPUS DATA

4.1 THE EUROPARL CORPUS

In this section, we will discuss the data that we will use in our experiments. As mentioned in Section 2, we will run experiments on canonical NSLs, namely Italian, Spanish and Greek, and one partial NSL, Finnish. We will train our MT system to translate to the same target language, which is English, a non-NSL. Therefore, we carry out a set of experiments on 5 language pairs: Italian-English, Spanish-English, Finnish-English and Greek-English. For each language pair, we used the Europarl corpus, a collection of the proceedings of the European Parliament (Koehn, 2005). Using a unique corpus allows for an easier interpretation and comparison of the results across the languages, as they all share the same topic domain. Table 4 presents the size of the various corpora.

Language pair	sentences
IT-EN	1,909,115
SP-EN	1,965,734
FI-EN	1,924,942
EL-EN	1,235,976

Table 4: Sentences of the parallel corpus for each language pair (Release v.7)

Below in Table 5, we report the mean length of the source and target corpus for each language pair. All the languages have a similar mean length between the source and target side, except for Finnish and Greek, where the mean length is slightly lower on the source side. Italian and Spanish have a similar length on the source side, which was expected as they have similar grammar and similar syntax.

Language pair	Mean length (source)	Mean length (target)
IT-EN	26	24
SP-EN	26	24
FI-EN	19	24
EL-EN	21	25

Table 5: Mean length of sentences of the parallel corpus for each language pair (Release v.7)

We first ran some exploratory analysis on the available data with regard to the subject distribution to get a better understanding of the data. First, we analyze the null-subject with regard to the verb conjugation in section 4.1.1. Then, we analyze the distribution of the gendered pronouns in section 4.1.2.

4.1.1 Analysis of the verbs

Figure 11 shows the distribution of the verb conjugation for Greek, Italian and Spanish, and Finnish, and whether the subject was present or not. The results are computed running an analysis on the dependency parsing trees, as explained in section 3.5.3, using the *Spacy* library (Honnibal & Montani, 2017). We ran this analysis on a subset of 50.000 randomly sampled sentences for each language.

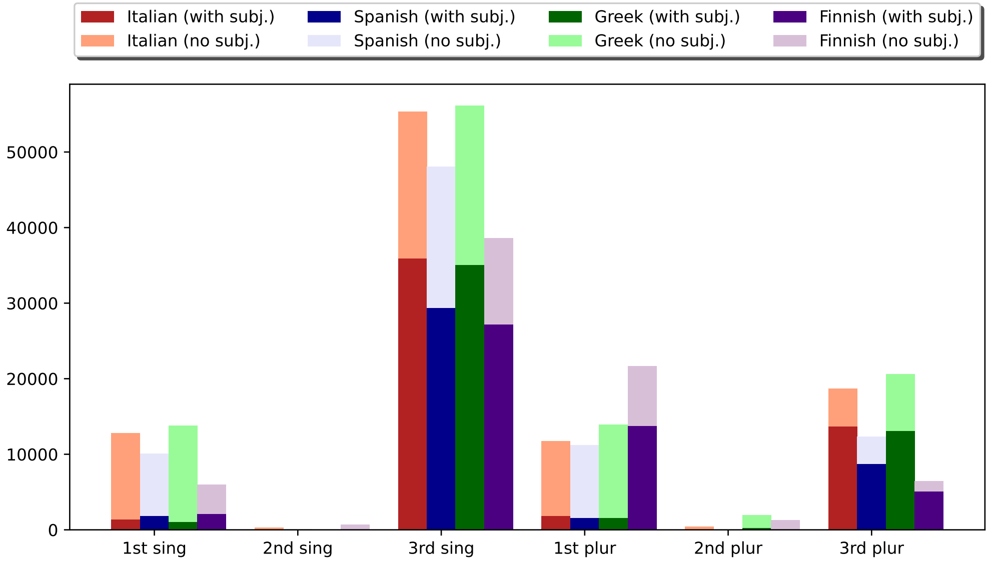

Figure 11: Null-subject presence across the conjugations of verbs in the Europarl corpus, on a subset of 50K sentences. The sentences are different for each language. Note that some sentences may contain more than one verb

Across the different languages, the speakers rarely refer to each other directly: the second conjugation, both singular and plural, is rarely used. The only exception is for Greek, in which the second plural conjugation is commonly used when speaking formally. Regarding the null subject distribution, the bar-plot shows that the subject is missing in the majority of the verbs conjugated in the first person, both plural and singular. In the third singular conjugation, around one-third of the subjects are null-subjects. For the plural form, this is less marked.

4.1.2 Analysis of the gender

We conducted a more focused analysis on the English side of the corpora with respect to the gender of the pronouns to obtain a better understanding of the data. We analyzed the frequencies of the gendered pronouns *he* and *she* in all the sentences of the Europarl corpus, as shown in Figure 6. The masculine pronoun is the subject in the majority of the sentences, which indicates a potential source of bias in the training data (as discussed in section 3.6).

	Sentences with subjects 'he'	Sentences with subject 'she'
IT-EN	31,278	8,618
EL-EN	19,245	5,662
ES-EN	30,908	8,470
FI-EN	25,404	7,107

Table 6: Gender of the English pronominal subjects found in the target side of the Europarl corpus.

This section gave an exploratory analysis of the data that we will use for the null-subject evaluation. The omission of the subject is very common across each NSL. Given the large percentage of sentences with missing subjects, we expect that our data-driven systems will be able to generalize the knowledge and learn to use the right contextual information for inferring the missing subjects. On top of that, we also expect that there may be a bias when inferring the gendered pronouns, as the data seems to contain an imbalanced use of pronouns. This should be reflected in a lower BLEU score, a lower STM, and a higher rate of wrong inferred subjects. In the next section, we will explain in detail our experimental setup aimed at solving our research questions.

5 THE TRANSLATION OF THE NULL SUBJECT WITH NMT

This chapter gives a quantitative analysis of the null-subject translation of NMT models for different language pairs. The source code necessary to replicate the experiments can be found on Github [1]. Section 5.1 describes the various pre-processing steps applied to the data, which include tokenization and segmentation. Section 5.2 outlines the parameters of the different NMT models, the evaluation metrics, and the procedure to train and evaluate the different models. Finally, in Section 5.3 the results of the experiments and their interpretation are reported.

5.1 PREPROCESSING

5.1.1 Tokenization and data cleaning

For the first steps of the pre-processing, we use the scripts from the *Moses* MT toolkit[2]. We first tokenize each corpus with the Moses tokenizer. After that, we remove the empty sentences and the redundant space characters. Finally, the data is cleaned with a minimum sentence length of 1 token, a maximum length of 200 tokens.

5.1.2 Segmentation

For the data available in Italian, Spanish, Greek and English, we build a lexicon of morphs using *BPE*. For each language pair, we apply BPE with 32K split operations (Sennrich et al., 2016) on the concatenation of the source and target corpus (i.e. Italian + English). For Finnish, which has a more rich morphology compared to other languages, we segment the corpus using *morfessor* (Creutz et al., 2016), the probabilistic machine learning algorithm. The corpus weight of the morfessor model is tuned so that there will be approximate 32K morph types in the lexicon, using the original word-list of Creutz et al. (2005). Using

[1] https://github.com/fferlito/Null_Subject_Analysis

[2] https://github.com/moses-smt/mosesdecoder

this number of morphs allows a better comparison between the models, as they will have a similar vocabulary size.

5.2 EXPERIMENTAL SETUP

We carry out experiments on 5 languages pairs: IT-EN, ES-EN, FI-EN, EL-EN. For each pair, we train different NMT models, from the pro-drop language (Italian, Spanish, Finnish and Greek) to the non-pro drop one (English). We expect that the three canonical NSLs, namely Greek, Italian and Spanish, will perform similarly in the null-subject translation task, as they have similar grammar and syntax. We expect that the language models will perform even better in the null-subject translation for the remaining language, Finnish: as this language is a partial NSL, there are fewer cases where the subject is dropped. This makes the task easier, as there are fewer cases to learn for each Finnish NMT model.

We trained all our models using *fairseq* (Ott et al., 2019), a sequence modeling toolkit written in PyTorch (version 0.1.2). Each model is trained with an Nvidia V100 GPU card, using CUDA 10.1. The hardware is provided by the *Peregrine High Performance Computing cluster* of the University of Groningen[3].

We train three different NMT architectures to investigate the difference in quality during the null-subject translation. We first train an LSTM network (described in section 3.3.2): this will be the simplest model, which we use as a baseline result. We will compare the results of this model with the ones obtained in previous research by the STM system and rule-based system. We expect that the LSTM model will translate a considerable higher percentage of null-subjects compared to the STM. Second, we train a LSTM with attention mechanism (introduced in section 3.3.3). We expect that the model will benefit from the usage of the attention mechanism for the null-subject translation, as the mapping between verb conjugation and subject is non-monotonic (as explained in section3.3.3). Therefore, the model with attention mechanism should translate a higher percentage of subjects compared to the base model without it. Finally, we train a model based on the transformer architecture with the attention mechanism (outlined in section 3.3.4). We expect that the improved architecture and the better attention mechanism will make this model outperform the previous two, both in terms of quality of translation and null-subject translation.

[3] https://wiki.hpc.rug.nl/peregrine/

Both the LSTM models, with and without attention mechanism, have the same hyper-parameters. The encoder and the decoder are composed of 2 layers each, with an embedded size of 256 units. The models are optimized with the Adam optimizer and a dropout of 0.2. Similarly, all the transformer models are based on the same hyper parameters. To maintain a manageable network size, we use a feed-forward network size of 1024, which gives reasonable translation quality. Both the encoder and the decoder has 6 layers each, with an embedded dimension of 512 and 4 attention heads. All the models are optimized using Adam, with a dropout of 0.3 and a beam size of 5. During the training, all the models are fed with batches of size 64, until they stop training with early stopping after 25 epochs without improvement on the validation set of size 1000.

We evaluate the quality of the translation of each model by computing the BLEU score on our special null-subject-test set, containing 1000 sentences for each language pair extracted using the dependency parsing tree analysis (explained in section 3.5.3). These test datasets are also manually checked to account for potential errors of the parsing analysis, so that in each sentence there is at least a verb without subject. The size of these sets is large enough to estimate the general quality of the translations, and it's used in many studies that apply this same metric. The sentences are extracted randomly before training the model, and they are different for each model. We use the same test set to compute the STM, which indicates if the models generate translations with the same syntactic structure of the reference translation (as described in section 3.5.2). To measure the completeness of the null-subject translation, we make use of a special test dataset where all the source sentences lack the subject. We then measure the percentage of subjects verbs with subjects in the translated sentence, as described in section 3.5.3. We use the Spacy dependency parser (Honnibal & Montani, 2017) to generate these special datasets for Greek, Italian, Spanish and Finnish, which are then manually checked to remove the wrong sentences. The accuracy of the dependency parsers is around 90% across all language models. We make sure that all these special tests contain the same sentences, which allows for an easier comparison of the results. To measure the percentage of null-subject correctly translated in the target language, we use again the dependency parser, which allows us to see directly in which sentence the subject is missing. An example of a sentence of this special dataset is given in Table 7.

Language	Sentence
English	We have obtained a new Europe today.
Italian	Oggi abbiamo un'Europa nuova.
Spanish	Hoy hemos conseguido una nueva Europa.
Finnish	Olemme nyt saaneet uuden Euroopan.
Greek	Έχουμε μια καινούργια Ευρώπη σήμερα. (Ékhoume mia kainoúryia Evrópi símera)

Table 7: Example of test sentence, where the target sentence, always in English, requires an overt subject. Notice that in none of the source languages is there an overt subject.

5.3 RESULTS

We report BLEU scores and the STM on each test set, as well as the percentages of verbs with an explicit subject on the English side. All metrics are high for every language pair, as shown in Table 8.

	BLEU	STM	correct	wrong	missing
			IT ->EN		
LSTM	23.05	0.472	91%	2%	7%
LSTM (Att.)	35.67	0.542	89%	1%	10%
Transformer	37.30	0.548	87%	1%	12%
			ES ->EN		
LSTM	29.85	0.533	88%	3%	9%
LSTM (Att.)	42.44	0.585	86%	3%	11%
Transformer	44.01	0.534	88%	3%	9%
			EL ->EN		
LSTM	26.38	0.511	89%	3%	8%
LSTM (Att.)	41.11	0.586	86%	3%	10%
Transformer	43.63	0.601	86%	3%	11%
			FI ->EN		
LSTM	18.89	0.345	77%	7%	16%
LSTM (Att.)	32.76	0.358	80%	6%	14%
Transformer	34.44	0.410	75%	15%	10%

Table 8: Results for the different models on the null-subject test set. The percentage of subjects translated is computed checking if each verb in English had a subject. Correct translations are considered when the null pronoun is translated by an overt pronoun with the correct gender, person and number features in English; otherwise, we considered it incorrect. Missing translation refers to cases where the null pronoun is not generated at all in the target language.

The BLEU scores obtained by the different models reflect our initial expectations: the simplest model, the encoder-decoder LSTN model, has the lowest score, followed by the counterpart with the attention mechanism. Finally, the transformer model has the highest score, confirming its superior capabilities for the translation task. The scores are similar to the ones obtained in the literature (Bugliarello et al., 2020). Surprisingly, the Italian models perform generally worse than the Spanish ones, even though the two languages share a lot of properties and they are very similar to each other. However, this trend is similar to the scores of other researchers. For example Bugliarello et al. obtained a BLEU score of 40.8 for the Italian model and 50.6 for the Spanish one, which seems to point to the fact that Italian has a more complex grammar that makes it harder to translate for the models. Finnish has the lowest scores for every MT architecture. We assume that its complex morphology and the different segmentation technique makes the mapping into English more complex to model.

If we compare the STM to the BLEU score, we observe a similar trend: the simple LSTM model has the lowest score, followed by the LSTM with attention and the transformer. This seems to indicate that, as we improve the model architecture, the generated translations have a syntactical structure more similar to their reference sentences.

If we compare the different percentages of the translated subject, we can see that all the models have a considerable amount of correct translations. If we check some generated translations, like the ones shown in Table 9, we can see that the model correctly translates the null subjects in English.

If we compare the performance of the simple LSTM model to the performance of the older MT systems (discussed in section 3.1 and 3.2), we can see that when the source language is Italian, there are 44.32 % more subjects correctly translated compared to the rule-based system and 30 % compared to the SMT. Similarly, when translating from Spanish, there is an improvement on 43.72 % of the translated subjects compared to the rule-based system and 24.32 % when comparing it to the SMT system.

Figure 12 shows the bar-plots of the correct, wrong and missing null-subject translations results of Table 8. From the plot is clear that all the different models perform similarly, as the rate of the null subject translation is around 88-90% for each language pair. The only exception is for Finnish, as the correct translations are between 75 and 77%.

The percentages for Italian, Spanish and Greek reflect our initial expectations: they all have similar scores across the different models, which indicates that the task difficulty is similar for these languages, as they all have similar grammar and syntax. Surprisingly, Finnish has the lowest values. This is against our expectations: even if this language is a partial NSL, and it contains fewer cases where the subject is dropped, the model struggles more to infer the right subject, as indicated by the higher percentage of wrong subjects. We have three possible expla-

Language	Text
Reference	I do not believe you , but I am going to vote for you anyway.
Italian	Non ti credo, ma ti voterò comunque.
Italian ->English	I do not believe you, but I will vote for you anyway.
Spanish	No le creo, pero voy a votar por usted de todas formas.
Spanish ->English	I do not believe you, but I am going to vote for you anyway.
Finnish	En usko sinuun , mutta äänestän sinua siitä huolimatta.
Fininsh ->English	I do not believe in you, but I will vote for you nonetheless.
Greek	Δεν σε πιστεύω , αλλά θα σε ψηφίσω έτσι και αλλιώς. (Den se pistévo , allá tha se psiphíso étsi kai alliós.)
Greek ->English	I do not believe you, but I will vote for you anyway.

Table 9: Example of translations when in the source sentence the subject is missing. For every language-pair, the model successfully inferred the right pronoun.

nations for this: (i) there may be too few cases where the subject is dropped, and the models may not have enough data to learn the rules for mapping; (ii) Finnish is a gender-neutral language, therefore the model has less context to correctly infer the gendered pronouns; (iii) Finnish has a more complex morphology, which makes the MT models themselves perform poorly in the translation task.

When we look at the percentage of translated subjects in Table 8, we can see that in around 10 % of the cases the subject is missing, even though English is a non-NSL and we would expect the rate to be zero percent in the most optimal scenario. Having an inspection of the cases where the subject is not retained in the target language gives a hint to the explanation. In few English constructions, pronouns are regularly omitted. Therefore, in many cases where we we consider the subject in the translation to be missing, the translation may be grammatically correct. In the prescriptive grammar, we can find three types of English null subjects. One of them is the deletion of the same subjects in conjoined sentences as shown in the generated translation (21).

(21) È un approccio a favore del mercato e dovrebbe funzionare. [Italian]
It is a pro-market approach and it should work. [Target

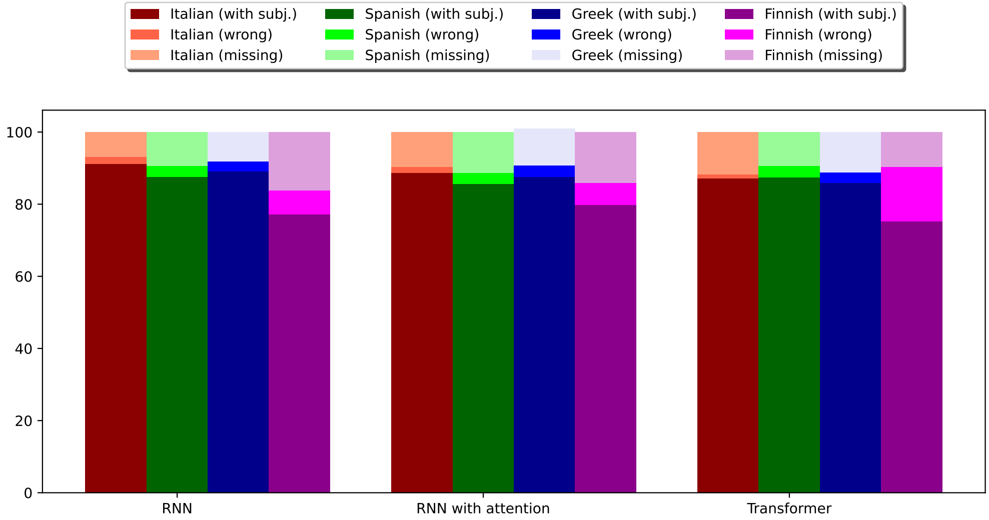

Figure 12: Null-subject translation across the conjugations of verbs in the Europarl corpus. For each language pair we show the result of three different models. The x-axis indicate the corresponding models trained.

translation]
It is also a market approach and should work. [Generated translation]

It's interesting to see that, even if the target sentence in (21) contained the subject in the conjoined sentence, this is not the case in the generated translation. This reflects how the model learnt this English rule from the training data.

The second type is the imperative null subject, which we find in imperative clauses, like in the generated translation (22)

(22) Ricordate che avete un minuto ciascuno. [Italian]
Remember that you each have one minute. [Target translation]
Remember that you have one minute each. [Generated translation]

The last type happens when there is a process of *trunctation*: the removal of one or more words at the beginning of the sentence. In (23) it's possible to see the truncation of the first words of a spoken sentence:

(23) He went up to one shelf, and scanned it. 'Hmm, it seems to be a section missing,' he said." [With over subject]

He went up to one shelf, and scanned it. 'Hmm, seems to be a section missing,' he said." [With truncated subject]

This type of null-subject is not present in the generated translations: this English construction is very informal, therefore we expect the corpus does not contain any of these cases, as the tone of the dialogues is very formal.

Having a closer look at the generated translation, it appeared that in few cases the reference translation was wrong, as shown in (24)

(24) Vi porgiamo il nostro benvenuto [Original]
I welcome you most warmly [Reference translation]
We welcome you most warmly [Generated translation]

In the example above, the verb in the Italian sentence, *porgiamo*, is conjugated in the first plural form. However, the reference translation wrongly contains the singular form of the pronouns. These few cases in the parallel data inflate the error rate during the evaluation of the null-subject translation. Other cases in which the subject is considered wrong is when the model does not use the subject of the reference sentence, but a synonym, like in (25).

(25) (pro) è il modo più efficace [Original]
It is the most effective way [Reference translation]
This is the most effective way [Generated translation]

In this example, the model uses the word 'this' instead of 'it'. Both of them are acceptable translations.

The use of a dependency parser for the metric computation has the consequence that potential errors during the parsing are propagated in the final score. In some cases, we noticed that the subjects of some sentences are wrongly classified by the parser, leading to a slightly inflated error rate. The cases just mentioned covers the majority of the sentences where the subject is considered wrong. Therefore, we expect the wrong null-subjects translation rate is close to 0 %.

The results show that for each language pair, all the model successfully manages to translate the subject correctly. On top of that, the spoken nature of the data influenced the model to learn some prescriptive rules of the target language. The next chapter will describe the second experiment, in which the correctness of the translation is checked, with regard to the gender bias.

6 THE GENDER BIAS AND THE NULL SUBJECT

This chapter gives an analysis of the gender bias that occurs during the null-subject translation for different language pairs, observing the variation of such bias after altering the training data of the model. Section 6.1 outlines the NMT model parameter, evaluation metrics, and the procedure to train and evaluate the different models, and our method to reduce potential gender bias in the translations. In Section 6.2 the results of the experiments and their interpretation are reported.

6.1 EXPERIMENTAL SETUP

We carried out experiments on the same 4 languages pairs of the previous experiment: IT-EN, ES-EN, FI-EN, EL-EN. For each one, we train a model using the same preprocessing steps and model configuration of the previous experiment, described in Sections 5.1 and 5.2.

In the first part of this experiment, we measure the incorrectness and the gender bias in the translations of the models trained in the Europarl corpus. In the target language, English, the only gender specification is in the third person singular, in which the pronoun can have a feminine, masculine and neutral form (i.e. *he, she, it*). The analysis will focus on the sentences where the target translation contains these pronouns. In the source languages Greek, Italian and Spanish, words like adjectives, determiners, and some nouns change their form depending on the word to which they refer. This allows inferring the gender of the subject, even when this is dropped from a sentence. For instance, in the examples shown in (26) it's possible to see how the word '*bravo*' (meaning 'good', 'well-behaved') has two forms, depending on whether it's referring to male or a female.

(26) a. È bravo [Italian]
He is good [English]

b. È brava [Italian]
She is good [English]

In other cases, the sentence does not contain any gender specification, and there is not enough context to infer the correct pronoun, like in the example(27) :

(27) Ha preso il treno [Italina]
He/She took the train [English]

In the example above, both pronouns (he and she) are possible translations. To disambiguate between them, more context is necessary, as a previous sentence or information about the speakers. MT has shown that in cases like this, the MT systems will always return the masculine pronoun.

Finnish is the only language that differs in this regard. Compared to the other languages under investigation, it does not have gender agreement in nouns, adjectives and articles. This makes inferring the right pronoun harder for Finnish, as there are fewer cues. On top of that, the 3rd person singular pronoun is always without gender. Consider the following example (28):

(28) Hän on hyvä [Finnish]
He/She is good [English]

In this example, the personal pronoun "hän" can be mapped to both he and she. The only way to disambiguate between the two is with additional context information, for example with a proper noun. This means that in Finnish inferring the right pronoun in the target language is more challenging even when there is an overt subject in the source sentence. Therefore we expect that the potential bias present in the training data will stronger for this language.

To measure the gender bias in the trained models, we manually filtered two special datasets from the Europarl corpus: one in which the subject in the target language is always the masculine pronoun 'he', and the other set which always has the feminine one, 'she'. We automatically filtered them checking in the English side which sentences contained the pronouns aforementioned. We then removed manually the ones where gender information about the subject is present in other parts of the sentence (i.e. proper nouns, gendered articles and adjectives). In total, we extracted 62 sentences where the target subject is feminine ('she'), and 100 sentences where this is masculine ('he'). An example of a test sentence for the validation is shown in Table 10. In every target translation, both the gendered pronoun "he" and "she" are potential correct translations: this allows us to easily see if the models have a preferred translation for these cases, which reflects a bias in the training data. To access the correctness of the translation in this special dataset, we compare the gender of the translated pronouns, compare to the original ones. We expect that the baseline model will use the masculine pronoun in the translation in the majority of the cases.

Language	Sentence
English	He took a shower.
Italian	Ha fatto la doccia.
Spanish	Darse una ducha.
Finnish	Hän kävi suihkussa.
Greek	Έκανε ντους. (Ékane dous.)

Table 10: Example of test sentence, in which it's not possible to infer the gender of the subject in the source language. If there isn't additional context, both the pronouns "he" and "she" are correct translations.

In the second part of the experiment, we balance the use of the gendered pronouns in the English side of the corpus. To do so, we under-sample the sentence pairs where there is the most frequent pronoun. We tried different ratios of masculine and feminine pronouns, and we report the gender of the generated translation for each case. Table 11 shows the number of sentences with masculine and feminine pronouns.

	Sentences with subjects 'he'	Sentences with subject 'she'
IT-EN	31.278	8618
EL-EN	19.245	5662
ES-EN	30.908	8470
FI-EN	25.404	7107

Table 11: Gender of the English pronominal subjects found in the Europarl corpus.

We use a novel method for under-sampling the sentences with the masculine pronouns in the target side of the corpus, generating two different scenarios: first, allowing both masculine and feminine pronouns usage to be equally balanced across the corpora; second, we allow the feminine pronouns to be the most frequent pronoun, reversing the initial distribution so that we have 2 feminine pronouns for every masculine one. We expect that the trained model will reflect the distribution of gendered pronouns in the generated translations, using more feminine pronouns as these become more prevalent in the training corpus.

6.2 RESULTS

We report the gender of the translated subject for each of the models. We do this by showing the gender of the target translation, using the Europarl sentences as a gold standard. The effect of balancing the use of gendered pronouns is the generated pronouns is shown in Figure 13. The bar-plot indicates that, when we do not alter the training data, the model for each language pair has a strong preference for the masculine pronoun 'he' when it must infer the pronoun. This bias is more prominent when translating from Finnish. If we look at the distribution of each pronoun in the training data (section 4.1.2), we can see that this bias reflects the imbalance in the use of each case. This result is in line with our initial expectations.

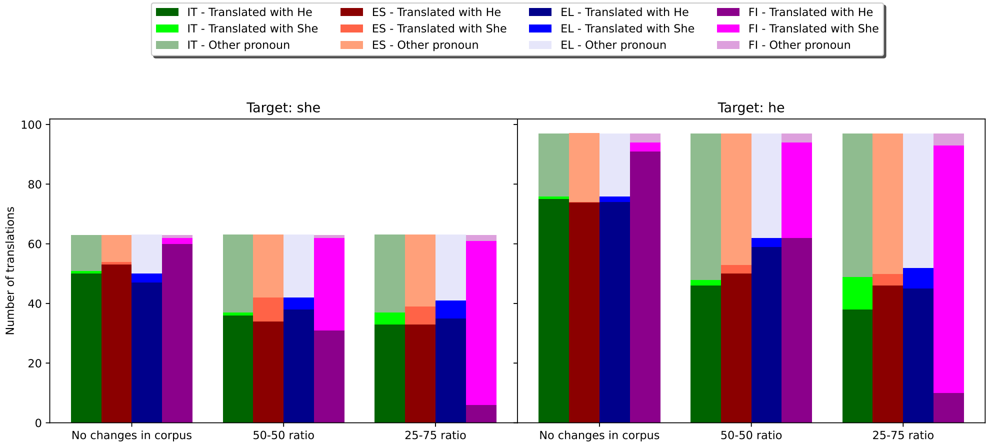

Figure 13: Gender of the pronominal subjects found in the English translation for different training datasets. The left plot shows the frequencies of the translations where the target translation has the feminine pronoun ('she'), while the right plot a masculine one ('he'). Each dataset has a different ratio of masculine-feminine subject pronouns.

The plots indicate that under-sampling the sentences where the masculine pronoun 'he' is the subject alters the bias in the translation: the generated translations use the masculine pronouns as default ones less frequently. However, the plot also shows that as the masculine pronouns decreases in the training corpus, the use of the feminine pronouns does not increase accordingly, except for Finnish. This is against our initial expectations: the models trained with the altered dataset seem to change the translation so that neither the masculine or feminine pronouns are used.

If we have a look at the generated translations, we see that in all language pairs, except for Finnish, the model would generate translations with the neutral form of the pronoun, "it", as shown in the example in (29):

(29) Ha fatto un lavoro notevole. [Source text]
He has done a remarkable job. [Default corpus]
It has done a great deal of work. [50-50 corpus]
It has done a remarkable job. [75-25 corpus]

As we did not measure the presence of the neutral gender, a possible explanation for these generated translations is that the neutral 3rd singular pronoun is the most common pronoun in the balanced corpus. As this makes fewer assumptions about the gender of the subject, we think that it's beneficial for the model. The Finnish model, on the other hand, generated more feminine pronouns as the training corpus has less masculine ones. While in the other languages there are a lot of cases where the gender agreement of the subject of nouns, adjectives and articles helps to disambiguate the gender, as shown in the example (26) above, in Finnish this is never the case, as the language is gender neutral. We think that this difference causes the Finnish model to be more dependant on the distribution of pronoun usage, as there are less lexical dependencies between words and gender.

Another interesting pattern that can be observed is in the Italian model. As the masculine pronouns are less frequent in the dataset, we noticed that many 3rd singular person pronouns are changed into a 2nd singular person form, like in the example shown in (30):

(30) Ha dimenticato di parlare del bilancio. [Source text]
She forgot to talk about the budget. [Default corpus]
You forgot to talk about the budget. [50-50 corpus]
You forgot to talk about the budget. [75-25 corpus]

In Italian, when a person addresses directly another person in a formal way, the 3rd singular person is used instead of the 2nd singular person. The model seems to generate more translation where the speaker is addressing another person after altering the training data. While this seems to generate more correct translations, as indicated by the higher BLEU score shown in Table 12, there is not enough context to understand whether the speaker is talking to someone else or not. Therefore, many of these translations could be wrong.

The Finnish-English model is the only one that reflects the balancing of the pronouns in the inference of the pronouns. We assume that the reason for this is the morphology of the Finnish language: compared to the other NSL, it is gender-neutral and there is almost no reference to the gender of the subjects in adjectives, pronouns and nouns. This makes the Finnish-English model have fewer word variations indicating the gender of the subjects, which makes the model learn fewer associations between words and genders. The other language pairs that we studied make rich use of the grammatical gender. As the majority of the subjects in the Europarl corpus are masculine, many words will be in the masculine form to agree with the gender of the subject. This leads the transformer model to associate many of these words to the masculine subject, as it's still the predominant gender form of the adjectives, articles, and verbs in these languages. Past research had shown that MT models would associate certain words to specific genders when going from and to gender-rich languages (Mikolov et al., 2013). For example, it has shown that the word "doctor" would be associated with the masculine gender, and the word "nurse" with the feminine one, as in the gendered language there would be a higher frequency with these corresponding genders. We assume that in our experiment the model would learn similar associations: even if the pronoun usage is balanced, there could be more masculine word variations, which make the model show a bias for these words when inferring the gender of the pronoun. This does not happen for Finnish, which does not allow to generate these word-dependencies related to gender.

We also report the BLEU score for the different models in Table 12. This allows seeing if changing the usage of the pronouns in the corpus has an effect on the quality of the translations.

Language	feminine-masculine pronoun	BLEU
Italian	unbalanced	37.30
	50-50 ratio	38.32
	25-75 ratio	37.96
Greek	unbalanced	44.01
	50-50 ratio	44.21
	25-75 ratio	43.43
Spanish	unbalanced	44.01
	50-50 ratio	44.6
	25-75 ratio	45.33
Finnish	unbalanced	34.44
	50-50 ratio	35.04
	25-75 ratio	34.04

Table 12: BLEU score for the models for each language pair, with and without balancing the use of gendered-pronouns in the training corpora.

As the table indicates, balancing the pronouns' usage in the training data has a beneficial effect: in every language pair, the model with the best BLEU is the one where the masculine and feminine pronouns have the same frequency.

7 CONCLUSION

Using the findings brought out in this study, we can now investigate and answer the research questions. We asked:

RQ1) Do languages show a similar frequency of null-subjects across different languages?

RQ2) How well does the LSTM encoder-decoder model infer the null-subject when translating from NSLs into English?

RQ3) Does the attention mechanism improve the quality of the null-subject translation in the LSTM model?

RQ4) How well does the state-of-the-art transformer model translate the null-subjects from these languages into English compared to the LSTM architecture?

RQ5) Can we reduce the gender bias showed by the null-subject translation by balancing the Europarl corpus?

From the analysis on the parallel corpora is clear that different NSLs shows a similar frequency of null-subjects (RQ1). The results show that not only the ratio of verbs with and without subjects are similar across languages, but also that the verb conjugations show the same distribution. This is expected, as the text within the corpus is translated by professional translators and the nature and the domain of the data remains unaltered.

Our LSTM models outperformed the previous MT methods in the null-subject translation task with a considerable margin (RQ1). The usage of a neural approach is beneficial: the ability to learn close and distant dependencies in the text allows that the model learns the relation between the missing subject and the verb conjugation. While this is clear inspecting the correct inferences of the pronouns, a disadvantage of the method is that it acts as a black-box: it's not possible to investigate directly the causal dependencies between the input and the output.

The effect of the addition of the attention mechanism in the LSTM model does not appear to improve directly the null-subject translation (RQ2). A closer inspection on translations lacking the subjects shows that the attention mechanism does not generate them in specific structures where it is not required, improving the fluency of the sentence removing, for example, repetitions.

Similarly, the transformer models that we trained did not outperform the LSTM models in the null-subject translation task (RQ3). The percentage of the translated subject is the lowest for this model. Similarly to the LSTM with attention mechanism, this model removes a lot of the subjects that are repeated in the same sentence, thus reducing the percentage of subjects present in the generated outputs.

We observed a consistent trade-off between the quality of the model, expressed in terms of BLEU and STM scores, and the percentage of subjects covertly expressed in the target language. In general, the inference of the null-subject benefits from the use of neural machine translation architectures. While the LSTM models infer most of the subjects correctly, the meaning and the structure of the translations are of poorer quality, as indicated by the relatively low BLEU and STM metrics. Conversely, the addition of the attention mechanism and the use of the transformer architecture with attention allow for increased translation quality. This is reflected in the learning of the structures in the target language in which the omission of the subject is allowed, as indicated by the higher STM score for these two models.

The effect of balancing the use of gendered pronouns in the training data shows a reduced bias in the inference of the null-subject (RQ5). The impact of balancing the pronouns varies depending on the type of source language. The influence is greater for gender-neutral languages, like Finnish. The lower effect on the languages with grammatical gender may be caused by the fact that the model learns to associate certain words with a specific gender. Across all languages, the model has a tendency to alter the sentences to avoid the use of gendered pronouns after the balancing of the pronouns, for example, using passive forms and using neutral pronouns, like ”it”.

Further experiments, using a broader set of gender-neutral languages, could shed more light on the gender bias in the null-subject translation, as the only factor influencing the bias in these languages is the vocabulary. The use of languages with rich morphology, that allows the expression of the gender thought adjectives and other structures, introduces new dependencies which make the investigation of the bias more complex.

Further research could also be conducted to determine the effectiveness of annotating the null-category in the source language: as in some cases there is not information to infer the right gendered pronoun, a possible approach can annotate these cases with some special tokens. These can then be used in a post-processing step to infer the right pronoun when more context is available, or by an external user.

Lastly, while our experiments showed that the NMT systems can solve the null-subject translation in canonical NSLs, the methods did not focus on whether or not the type of segmentation had any effect. This can be further investigated in future research. If only certain types of segmentation allow for the correct inferring of the null-subject, the finding would have a large impact for future practice.

BIBLIOGRAPHY

Allen, R. (1987, 01). Several studies on natural language and back propagation. , 2.

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural machine translation by jointly learning to align and translate.*

Banerjee, T., & Bhattacharyya, P. (2018, June). Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the second workshop on subword/character LEvel models* (pp. 55–60). New Orleans: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W18-1207 doi: 10.18653/v1/W18-1207

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (p. 610–623). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3442188.3445922 doi: 10.1145/3442188.3445922

Berger, A. L., Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Gillett, J. R., Lafferty, J. D., ... Ures, L. (1994). The Candide system for machine translation. In *Human Language Technology: Proceedings of a workshop held at Plainsboro, New Jersey, March 8-11, 1994.* Retrieved from https://www.aclweb.org/anthology/H94-1028

Biberauer, T. (2008). Semi null-subject languages, expletives and expletive pro reconsidered..

Bisazza, A., & Tump, C. (2018, 01). The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. In (p. 2871-2876). doi: 10.18653/v1/D18-1313

Blatz, J., Fitzgerald, E., Foster, G., G, S., Goutte, C., Kulesza, A., ... Ueng, N. (2004, 08). Confidence estimation for machine translation. *Proceedings of COLING 2004*. doi: 10.3115/1220355.1220401

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993, June). The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, *19*(2), 263–311.

Bugliarello, E., Mielke, S., Anastasopoulos, A., Cotterell, R., & Okazaki, N. (2020, 01). It's easier to translate out of english than into it: Measuring neural translation difficulty by cross-mutual information. In (p. 1640-1649). doi: 10.18653/v1/2020.acl-main.149

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014, October). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation* (pp. 103–111). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W14-4012 doi: 10.3115/v1/W14-4012

Chung, T., & Gildea, D. (2010, October). Effects of empty categories on machine translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 636–645). Cambridge, MA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D10-1062

Creutz, M., Lagus, K., Linden, K., & Virpioja, S. (2005). Morfessor and hutmegs: unsupervised morpheme segmentation for highly-inflecting and compounding languages. Retrieved from http://eprints.pascal-network.org/archive/00001841/

Doddington, G. (2002, 01). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. , 138-145. doi: 10.3115/1289189.1289273

Dryer, M. S. (2013). Expression of pronominal subjects. In M. S. Dryer & M. Haspelmath (Eds.), *The world atlas of language structures online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from https://wals.info/chapter/101

d'Alessandro, R. (2015). Null subject. *Contemporary linguistic parameters*, 201–226.

Frank, A., Hoffmann, C., Strobel, M., et al. (2004). Gender issues in machine translation. *Univ. Bremen*.

Haegeman, L. (1994). *Introduction to government and binding theory* (2nd ed. ed.). Oxford: Blackwell.

He, H., & Ma, Y. (2013). *Imbalanced learning: Foundations, algorithms, and applications* (1st ed.). Wiley-IEEE Press.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.*

Hochreiter, S., & Schmidhuber, J. (1997, 12). Long short-term memory. *Neural computation*, *9*, 1735-80. doi: 10.1162/neco.1997.9.8.1735

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.* (To appear)

Hutchins, J. (2007). Machine translation: A concise history. *Computer aided translation: Theory and practice*, *13*(29-70), 11.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (p. 560–575). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3442188.3445918 doi: 10.1145/3442188.3445918

Imamura, K. (2002). Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based mt..

Johnson, J., & Khoshgoftaar, T. (2019, 03). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*, 27. doi: 10.1186/s40537-019-0192-5

Kalchbrenner, N., & Blunsom, P. (2013, October). Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1700–1709). Seattle, Washington, USA: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D13-1176

Kanouchi, S., Sudoh, K., & Komachi, M. (2016, December). Neural reordering model considering phrase translation and word alignment for phrase-based translation. In *Proceedings of the 3rd workshop on Asian translation (WAT2016)* (pp. 94–103). Osaka, Japan: The COLING 2016 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/W16-4607

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit* (pp. 79–86). Phuket, Thailand: AAMT. Retrieved from http://mt-archive.info/MTS-2005-Koehn.pdf

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ... Herbst, E. (2007, June). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/P07-2045

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 human language technology conference of the north American chapter of the association for computational linguistics* (pp. 127–133). Retrieved from https://www.aclweb.org/anthology/N03-1017

Li, P., Liu, Y., Sun, M., Izuha, T., & Zhang, D. (2014, August). A neural reordering model for phrase-based translation. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 1897–1907). Dublin, Ireland: Dublin City University and Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/C14-1179

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.*

Neco, R., & Forcada, M. (1997). Asynchronous translations with recurrent neural nets. In *Proceedings of international conference on neural networks (icnn'97)* (Vol. 4, p. 2535-2540 vol.4). doi: 10.1109/ICNN.1997.614693

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., ... Auli, M. (2019). *fairseq: A fast, extensible toolkit for sequence modeling.*

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Popel, M. (2018, October). CUNI transformer neural MT system for WMT18. In *Proceedings of the third conference on machine translation: Shared task papers* (pp. 482–487). Belgium, Brussels: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W18-6424 doi: 10.18653/v1/W18-6424

Rizzi, L. (1986). Null objects in italian and the theory of pro. *Linguistic Inquiry*, *17*(3), 501–557. Retrieved from http://www.jstor.org/stable/4178501

Rosenblatt, F. (1957). *The perceptron, a perceiving and recognizing automaton project para.* Cornell Aeronautical Laboratory. Retrieved from https://books.google.nl/books?id=P_XGPgAACAAJ

Russo, L., Loáiciga, S., & Gulati, A. (2012). Improving machine translation of null subjects in italian and spanish. In *Proceedings of the student research workshop at the 13th conference of the european chapter of the association for computational linguistics* (p. 81–89). USA: Association for Computational Linguistics.

Sapir, E. (1921). *Language : an introduction to the study of speech / edward sapir*. New York: Hartcourt, Brace and Company.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. *arXiv preprint arXiv:2104.06001*.

Schiebinger, L. (2014, 03). Scientific research must take gender into account. *Nature*, *507*, 9. doi: 10.1038/507009a

Schuster, M., & Paliwal, K. (1997, 12). Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, *45*, 2673 - 2681. doi: 10.1109/78.650093

Schwenk, H. (2012, December). Continuous space translation models for phrase-based statistical machine translation. In *Proceedings of COLING 2012: Posters* (pp. 1071–1080). Mumbai, India: The COLING 2012 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C12-2104

Sennrich, R., Haddow, B., & Birch, A. (2016). *Neural machine translation of rare words with subword units.*

Slobin, D. I. (1996). From "thought and language"to "thinking for speaking". In J. Gumperz & S. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge University Press.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks.*

Taraldsen, K. (1980). *On the nominative island condition, vacuous application and the that-trace filter*. Indiana University Linguistics Club. Retrieved from https://books.google.nl/books?id=Z7lTNAAACAAJ

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). *Attention is all you need.*

Waibel, A., Jain, A., McNair, A., Saito, H., Hauptmann, A., & Tebelskis, J. (1991). Janus: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *[proceedings] icassp 91: 1991 international conference on acoustics, speech, and signal processing* (p. 793-796 vol.2). doi: 10.1109/ICASSP.1991.150456

Wang, L., Tu, Z., Wang, X., & Shi, S. (2019). *One model to learn both: Zero pronoun prediction and translation.*

Wang, L., Tu, Z., Zhang, X., Liu, S., Li, H., Way, A., & Liu, Q. (2017, 06). A novel and robust approach for pro-drop language translation. *Machine Translation*, *31*. doi: 10.1007/s10590-016-9184-9

Wang, R., Zhao, H., Lu, B.-L., Utiyama, M., & Sumita, E. (2014, October). Neural network based bilingual language model growing for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 189–195). Doha, Qatar: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D14-1023 doi: 10.3115/v1/D14-1023

Weaver, W. (1949/1955). Translation. In W. N. Locke & A. D. Boothe (Eds.), *Machine translation of languages* (pp. 15–23). Cambridge, MA: MIT Press. (Reprinted from a memorandum written by Weaver in 1949.)

Yamada, K., & Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of the 39th annual meeting on association for computational linguistics* (p. 523–530). USA: Association for Computational Linguistics. Retrieved from https://doi.org/10.3115/1073012.1073079 doi: 10.3115/1073012.1073079