Null Subjects in Spanish as a Machine Translation Problem

José Diego Suárez, Luis Chiruzzo
Instituto de Computación, Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay
jose.diego.suarez@fing.edu.uy, luischir@fing.edu.uy

Abstract

In this study we approach the detection of null subjects and impersonal constructions in Spanish using a machine translation methodology. We repurpose the Spanish AnCora corpus, converting it to a parallel set that transforms Spanish sentences into a format that allows us to detect and classify verbs, and train LSTM-based neural machine translation systems to perform this task. Various models differing on output format and hyperparameters were evaluated. Experimental results proved this approach to be highly resource-effective, obtaining results comparable to or surpassing the state of the art found in existing literature, while employing modest computational resources. Additionally, an improved dataset for training and evaluating Spanish null-subject detection tools was elaborated for this project, that could aid in the creation and serve as a benchmark for further developments in the area.

Keywords: Null-subject, Machine Translation, OpenNMT, Spanish

1. Introduction

A subject is a syntactic element, typically expressed as a noun phrase, which acts as a specifier for a verb phrase. The subject often, though not always, corresponds to a verbal argument in an agent-like semantic role. For instance, consider the following sentences in English:

- 1a) ***The man*** *drinks coffee.*
- 2a) *Are* ***they*** *here?*
- 3a) ***It*** *rains.*
- 4a) *Close the door!*

A subject (signaled by bold text) can be identified in sentences (1a) through (3a) whereas example (4a), a command, shows no explicit element that can be identified as its subject although context allows us to assert the existence of an elided subject ('you') who is expected to 'close the door'. It might also be noted that while the subjects for sentences (1a) and (2a) correspond to a real world entity (what is known as a *referent*), no such actor can be defined for sentence (3a): no referent can be identified for the *'it'* in *'it rains'*, this is an impersonal construction which requires a referent-less 'dummy pronoun' in order to fill what would otherwise be a syntactic gap in the subject position.

In standard English, overt or explicit subjects are always required except in imperative sentences; a subject-less sentence such as *"*drinks coffee"* results ungrammatical to most speakers.

This is not the case for other languages. In particular, there is a great number of languages dubbed 'null-subject languages' where any kind of sentence may be constructed without marking the subject explicitly. In such languages, the referent corresponding to the subject role (if any) is either hinted at by other syntactical or morphological features, such as verb agreement, or simply left to context. This is a special case of the more general 'pro-drop languages', which allow the omission of pronouns in different parts of the sentence, not only in the subject. One example of a null-subject language is Spanish, where the sample sentences presented before might be expressed as follows:

- 1b) ***El hombre*** *bebe café.*
- 2b) *¿Están aquí?*
- 3b) *Llueve.*
- 4b) *¡Cierre la puerta!*

While example (1b) contains an explicit subject (*el hombre*, 'the man'), no subjects are present in examples (2b) through (4b). Such sentences are considered to have 'null subjects'. It should be noted that sentence (2b) could also be expressed using an overt subject (*¿Están* ***ellos*** *aquí?*), more in line with the corresponding English example (2a), but this is entirely optional and bears little importance for interpretation as the existence of a third person plural referent in subject position can already be inferred from the verb due to Spanish subject-verb agreement. On a similar vein, sentence (1b) could be expressed with a null subject (*Bebe café.*) leaving the identity of the person who drinks coffee up to context.

A different situation can be identified for the third example sentence. Similarly to English, rainfall and other weather verbs are expressed in Spanish through impersonal verbs (like *llover*, 'to rain') for which no referent can be found. However, Spanish grammar departs from that of English in that a null subject construction is required in this case,

LREC-COLING 2024, pages 12313–12322

obviating the need for 'dummy pronouns' such as the 'it' in (3a); translating 'it rains' as *"*él llueve"* results ungrammatical. As a consequence, all impersonal constructions are necessarily expressed as null subject clauses in Spanish.

It should be noted that Spanish grammar allows many different impersonal constructions besides weather verbs, including existential clauses (*"hay"*, *"hubo"*, corresponding to English 'there is', 'there were') and sentences referring to a generic, unspecified actor such as *"se trabaja mejor en equipo"* ('it is better to work on teams'). The latter category involves verbs which could also appear in personal usages, making this a contextual distinction, so this phenomenon cannot be fully identified by simply checking against a list of intrinsically impersonal verbs.

The identification of null-subject and impersonal sentences is a necessary step for the syntactic parsing of a null-subject language (for example see Chiruzzo and Wonsever (2018) and Chiruzzo (2020)), bears significant relevance for semantic language processing tasks (as subjects often correspond to the primary role of a verb) and has implications for machine translation into a non-null-subject language such as English, where referents must be resolved and dummy pronouns introduced to conform to the syntax of the target language, or even for machine translation between pro-drop languages (see for example Ferlito (2021) for a discussion). Due to these reasons, the automatic detection of these phenomena constitutes a relevant Natural Language Processing (NLP) task.

2. Related Work

The problem of null-subject detection in Spanish has received moderate academic attention. Early takes on the subject include Ferrández and Peral (2000) which proposed a rule-based approach which achieved up to 88% accurate results, although evaluated against a very limited corpus, and without the distinction between null-subject as a result of omissions and impersonal usages where null-subject is mandatory.

This distinction was incorporated in later works such as Rello et al. (2012) and González and Martínez (2018). In those studies, a number of features including a verb's lemma and immediate context were processed through clustering algorithms such as K* (Rello et al., 2012) and support vector machines (González and Martínez, 2018). Results were evaluated on larger, more representative corpora such as EZSIC in Rello et al. (2012) and González and Martínez (2018) and the AnCora corpus (Delor, Mariona and Martí, Antonia and Recasens, M., 2022) for González and Martínez (2018).

A neural network approach to this problem was presented in Gerez et al. (2019), where they built several classifiers based on different features and a variety of neural networks such as feed-forward neural networks (FFNN), convolutional neural networks (CNN) and Long Short-Term Memory networks (LSTM).

To the extent of our knowledge, no published study has yet approached the problem as a sequence-to-sequence transformation as proposed in this work. This concept was motivated by the successful application of machine translation-like sequence to sequence techniques to syntactic parsing as shown in Vinyals et al. (2015).

3. A Machine Translation Approach to the Problem

The term Machine Translation (MT) refers to a number of techniques intended to translate text from a natural language (such as English, Spanish or Mandarin) into another through an automated method. In the last decade, MT research has been focused on Neural Machine Translation (NMT), where a neural network is purposed to build a sequence-to-sequence model. In such a model, the input is treated as a sequence of tokens (typically corresponding to words and punctuation in the source language) which are successively processed to generate an output sequence whose tokens map to the desired output. These models often make use of recursive neural network implementations such as LSTM (*Long Short-Term Memory* networks) which, with the addition of *attention mechanisms* to weight the relevance of past inputs, are able to process long sequences to a great degree of effectiveness, leading to continued improvements to the state of the art for MT tasks.

Many of the concepts and methods introduced for NMT are applicable to a wider range of NLP tasks which can be modelled as sequence to sequence transformations. For instance, the extraction of a certain feature or the addition of markers such as parsing tags or missing punctuation could be modelled as a translation problem between a certain kind of input and output sequences. By crafting a custom parallel corpus matching the expected input and output pairs for the problem, an NMT framework could be trained to map inputs to the problem to the expected output.

This approach was pioneered by Vinyals et al. (2015), where NMT techniques were adapted to build a highly effective syntax parser for English by mapping English plaintext to a linearized representation of the target syntax representation, effectively treating such a representation "as a foreign language".

This methodology has been adapted in further

works such as Stanovsky and Dagan (2018) where a machine translation model was repurposed to build a semantic parser for English.

In this study, we propose a machine translation model to "translate" text in Spanish into an intermediary representation that will allow us to classify verb phrases as having an explicit subject, a null subject or corresponding to an impersonal usage. Experiments were run on PyTorch LSTM-based NMT models trained using the OpenNMT framework (Klein et al., 2017).

In order to develop such an application, it is necessary to define the input and output formats that will define the source and target "languages" for the translation model, assemble a sufficiently large parallel corpus to successfully train the neural network, and explore the combination of external model parameters (hyperparameters) that will result in the best achievable performance. Considerations about these topics will be explored in the following sections.

4. Corpora

Neural machine translations models are commonly trained and evaluated on aligned parallel corpora, consisting of a collection of paired inputs and outputs in the source and target languages. Ideally, the parallel corpus should be vast and comprehensive enough to be sufficiently representative of any input the application may encounter during its operation.

In order to identify null-subjects and impersonal constructions in Spanish, this study required an annotated corpus where Spanish sentences were marked for the desired features, preferably one based on a diverse array of source materials so as to guarantee an adequate level of comprehensiveness. An existing language resource, the Spanish version of the **AnCora** corpus (Delor, Mariona and Martí, Antonia and Recasens, M., 2022) was identified as starting point for the definition of such a corpus. The AnCora corpus is a multi-level annotated corpus containing over 16,000 sentences and 500,000 words of Spanish text sampled from news text covering a diversity of topics and tagged with multiple annotations for syntactic structure, semantic roles and morphological derivations, among others. In particular, AnCora annotations include markers for most non-impersonal null-subject constructions as well as tags that allow to identify a certain subset of impersonal sentences (reflexive impersonal sentences). AnCora also indicates the part of speech (POS) of each lexical item, a feature that was incorporated into the output format for this project. Additionally, this corpus had been used in the evaluation of previous studies about null-subject recognition in Spanish, thus providing a good point of comparison between this and earlier results.

As part of this study, the AnCora corpus was processed in order to extract the elements that were relevant to the problem at hand, namely marks related to null-subject and impersonal clauses. During this process, we found that a fraction of the phrases in the corpus (amounting to 1324 verbs, 6% of the total) did not bear enough information to determine whether they had an explicit subject, a null-subject or whether they were impersonal. We classified these verbs manually in order to make a version of AnCora that is suitable for our task. The information extracted from AnCora as well as the hand-annotated additions constitute an improved corpus for null-subject detection, which we named *NullSubjAncoraCorpus*, published as part of this study[1]. This corpus also incorporates a default partition into training, development and test splits based on a *de facto* standard partition for AnCora as proposed by the CoNLL 2017 shared task (Zeman, 2017). Stats for the *NullSubjAncoraCorpus* and its proposed partitions are given in table 1.

	Training	Development	Test	Total
Sentences	13,489	1,567	1,623	**16,679**
Tokens	414,334	48 768	49,018	**512,120**
Finite verbs	34,925	4,018	4,162	**43,105**
Explicit subj.	25,047	2,850	2,979	**30,876**
Null subject	8,835	1,062	1,062	**10,959**
Impersonal	1,043	106	121	**1,270**

Table 1: Composition of the *NullSubjAncoraCorpus* and its proposed partitions.

In turn, the extended corpus *NullSubjAncoraCorpus* was transformed into a parallel corpus for training by extracting input sentences (plaintext) and output sentences following a format intended to facilitate the training of the NMT model and the posterior extraction of the identified features. A characterization of the output format will be given in the following section.

5. Development

In the proposed classifiers, an input sentence is fed into an NMT model which outputs a 'translation', an intermediary sequence in a certain output format which will encode the expected categories for each finite verb on the input. This intermediary sequence is subsequently post-processed to obtain the classification for each finite verb.

Non-finite verbforms (infinitives, participles and gerunds) are not classified as they are unable to head a verb phrase in Spanish and, as such, cannot be associated to a subject.

[1] https://github.com/jotadiego/NullSubjAncoraCorpus

5.1. Approaches

Initially, two NMT-based approaches were put forward for null-subject recognition. These approaches differ on how null-subjects are represented in the translation output and correspondingly define different output formats for the model. The first approach (Approach I) seeks to identify the phenomenon by incorporating an "elliptical subject" mark as a token in the translation output. This method closely follows the way null-subject clauses are represented in the original AnCora corpus, which incorporates an empty noun phrase element tagged as "elliptical" in a position where an explicit subject would otherwise be expected, usually preceding the corresponding verb. Under this approach, a verb phrase is classified as having null or explicit subjects depending on whether their output contains the "elliptical subject" token or not; the distinction between personal and impersonal null-subject sentences is not considered. A second approach (Approach II) consists in modelling null-subject and impersonal verb detection as a classification problem where each finite verb is to be classified into the following three categories:

- 1) Verbs with an **explicit subject**.
- 2) Verbs with a **null-subject** (excluding those resulting from impersonal constructions).
- 3) Verbs in an **impersonal** usage.

The expected output for Approach II models represents verbs using different labels depending on their respective category. Note that a binary distinction between explicit and null subjects (as the one implemented by Approach I) is also possible by merging the second and third classes. In our initial experiments, we used this binary format to compare between both approaches, and found that Approach II significantly outperforms Approach I (see section 6). Because of this, later experiments are done only with Approach II, and the *NullSubjAncoraCorpus* format corresponds to this second approach.

5.2. Translation Input and Output Format

The expected input and output formats for a translation model are determined by the aligned parallel corpus used to train the model.

In this study, the input only has minimal preprocessing consisting on transforming the text to lowercase in order to prevent purely typographic variations in capitalization from affecting the result. Although different output formats are required depending on the approach (section 5.1), there are considerations that apply to both, such as the need to consider input sentences that might contain multiple finite verbs and which, accordingly, will require multiple classification values. For the purpose of mapping values in the output to verbs in the input, we decided to use a format where each token in the input (verb or not) corresponded to a token in the output. In order to reduce the output space for the model (which results in smaller networks which might be trained using fewer computational resources), each word in the input was mapped to a tag corresponding to their part of speech similarly to a standard POS-tagging task, as shown on table 2, with additional tags being used depending on the approach.

Class	Tag
Adjective	**a**
Adverb	**r**
Conjunction	**c**
Date	**k**
Determiner	**d**
Interjection	**j**
Noun	**n**
Numeral	**z**
Preposition	**s**
Pronoun	**p**
Punctuation	**f**
Verb (finite)	**v**
Verb (non-finite)	**i**

Table 2: Part of speech (POS) tags used in the output formats; additional tags are used depending on the specific approach for null-subject tagging. Codes are based on *EAGLES* POS-tagging conventions as followed by the AnCora corpus. Modifications were added to distinguish between finite and non-finite verbforms as only the former are relevant for the task.

For Approach I, a special null-subject mark '**0**' is added to verb phrases lacking an explicit subject. Placement of the **0** tag follows the format given in the AnCora corpus, where a null-subject commonly precedes the verb (a canonical position for subjects in Spanish) but may sometimes be found in other positions. As the criteria for null-subject mark placement within the corpus seems to be inconsistent, only the presence or absence of a **0** mark in the output for a verb phrase is considered when evaluationg Approach I models, disregarding the relative positions.

By the Approach I scheme, sample sentences (1b) through (4b) are expected to generate the following outputs;

- 1b-I) ***El hombre*** *bebe café.* → **d n v n f**
- 2b-I) *¿_ Están aquí?* → **f 0 v r f**
- 3b-I) *_ Llueve.* → **0 v f**

Class	Category	Tag
Finite verb	Explicit subject	**v**
Finite verb	Null subject	**w**
Finite verb	Impersonal	**y**
Non-finite verb	-	**i**

Table 3: Tags used in the output format, corresponding to the three classification categories for finite verbs and part of speech (POS) tags for other lexical items. Codes are based on *EAGLES* POS-tagging conventions as followed by the AnCora corpus, adjusted with the inclusion of additional tags for null-subject mark.

- 4b-I) *¡Cierre _ la puerta!* → **f v 0 d n f**

Under Approach II, the discrimination is done by using different output tags for verbs depending on their category, as shown in table 3.
By the Approach II scheme, sample sentences (1b) through (4b) are expected to generate the following outputs;

- 1b-II) ***El hombre*** *bebe café.* → **d n v n f**
- 2b-II) *¿Están aquí?* → **f w r f**
- 3b-II) *Llueve.* → **y f**
- 4b-II) *¡Cierre la puerta!* → **f w d n f**

Ideally this would result in input and output sentences being the same length, with the classification value for the verb in the i-th position within the input sequence being recorded in the i-th position of the output. As NMT models are not perfect and might occasionally generate extraneous tokens or fail to generate expected tokens, the length of an actual output might still differ from that of the input. In such cases, relative positions within each sequence are used to correlate finite verbs in the input and classification values in the output.
Imperfections in the translation process might also result in a mismatch between the number of finite verbs in the input and the verb classification values in the output. While spurious verbs could be simply dismissed (after matching correct verbs given their relative positions), a missing label for a given finite verb is to be considered as an error. In practice this means that, although there are three possible values for a finite verb in the input (explicit subject, null-subject, impersonal), the classifier might output four different cases: the three valid categories, and an invalid fourth category corresponding to 'not detected'.

5.3. Hyperparameters

In Machine Learning a distinction is made between two classes of model parameters: ordinary or internal parameters whose values might be adjusted during the learning process and hyperparameters, which refer to characteristics that are fixed throughout the learning process. In the case of a neural network such as the OpenNMT models proposed in this study, parameters correspond to the weights of the LSTM neural networks and attention mechanisms, whose values are successively refined during the training of the model, whereas hyperparameters include such factors as the architecture of the neural network, how data is represented and interpreted in the model and the evaluation criteria, among others.
In this work, we experimented with three such hyperparameters: the usage of pre-trained word-embeddings, the number of hidden layers and the amount of units per hidden layer.

5.3.1. Word Embeddings

As neural networks operate on numeric values, it is necessary for an NMT application to transform a text input into a series of discrete tokens (typically corresponding to words or other lexical items such as punctuation marks) with a numerical representation such as a real-valued vector of a given dimensionality. Such representations are known as word embeddings, as they are able to embed a word (a token) into an n-dimensional vector space. The collection of tokens represented by a given schema is known as the vocabulary of the model, possible inputs not contemplated within the vocabulary are deemed out-of-vocabulary (OOV) items and require a fallback strategy such as being conflated into a single 'unknown' vector.
Although in principle the relationship between a token and its embedding could be arbitrary, multiple NLP applications report performance improvements when using collections of pre-trained word embeddings constructed in such a way that vector similarity correlates to semantic similarity. Pre-trained word embeddings may be generated from sufficiently large corpus (typically in excess of ten million words) through unsupervised learning algorithms such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) where semantic similarity is approximated by statistical correlations on the context of individual words.
In this work, we experimented with non pre-trained random embeddings generated by the OpenNMT framework, and with three Spanish word embedding sets: two collections pre-trained on the Spanish Billion Word Corpus (SBWC) using the GloVe and Word2Vec algorithms[2], and the 'emb39' collection trained with Word2Vec over a 6 billion word corpus (Azzinari and Martínez, 2016). In all cases, a vector size of 300 was used.

[2] https://github.com/dccuchile/spanish-word-embeddings

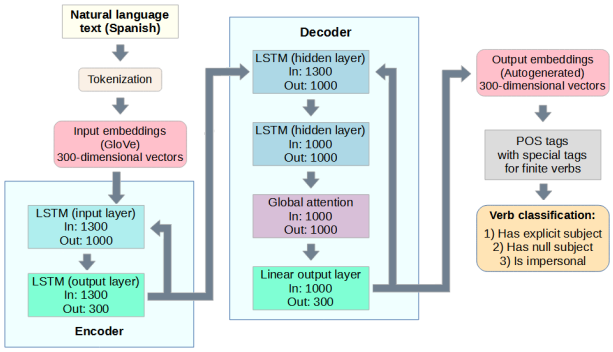

Figure 1: Diagram of the proposed architecture for a NMT model with 2 hidden layers of 1000 units.

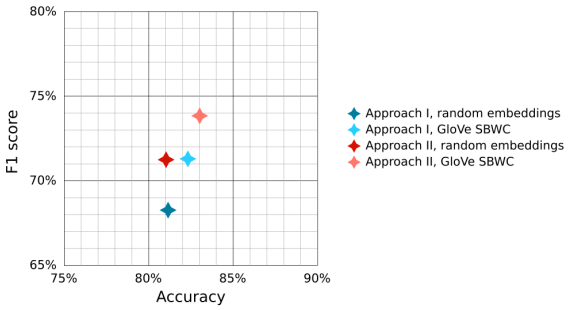

Figure 2: Scatterplot of F1 score and accuracy metrics of 2-layer models as described in table 4.

5.3.2. Neural Network Dimensions

Neural network models considered in this study follow an encoder-decoder LSTM architecture with an attention mechanism, with a layered structure comprised of an input layer, a series of hidden layers and an output layer. During this project only models with equal-sized hidden layers (i.e., composed of the same amount of units) were tested. As a result, the dimensions of a given model in this study are given by two hyperparameters: the number of hidden layers m and their size n.
The resulting networks are structured as follows:

- An embedding mechanism translating a predefined vocabulary (compiled from the training corpus) to the 300-dimensional word-embedding representation (using a fallback 'unk' vector for out-of-vocabulary items).
- The input LSTM layer composed of n units which receive the 300 values of the word embedding.
- An initial hidden layer of size n which receives the n outputs of the input layer as well as a 300-dimensional embedding of the last output of the model.
- $m-1$ size n hidden layers processing the n outputs of the preceding hidden layer.
- The output LSTM, which also contains m layers, combined with a global attention mechanism, operating on the n-dimensional output of the final hidden LSTM layer.
- A linear output layer which converts the n-dimensional into a softmax distribution for the value of the next token to be output.

The resulting architecture is depicted in Figure 1.

6. Experiments

Experimentation was carried on in a staged manner, with a first series of experiments intended as a proof of concept followed by progressive refinements obtained by adjusting model hyperparameters.

6.1. Binary Classifiers

Experiments were initially run on a binary classification scenario, contrasting explicit and null subjects; distinction between personal and impersonal null-subject clauses was incorporated at a later stage.
In the first stage, models with 1 and 2 hidden layers comprised of 500 units were tested as well as variations on the proposed output format (approaches I and II). Additionally, a comparison was made between training these models with random and pre-trained word embeddings (using GloVe SWBC vectors for the latter). Results for this stage are reported on table 4.

Appr.	Layers	Embeddings	Accuracy	Precision	Recall	F1 Score
I	1×500	Random	77,38%	74,04%	51,53%	60,77%
I	1×500	Glove SBWC	73,65%	68,05%	47,04%	55,62%
I	2×500	Random	81,16%	72,13%	64,82%	68,28%
I	2×500	Glove SBWC	82,33%	**74,87%**	68,07%	71,31%
II	1×500	Random	67.73%	70.06%	31.55%	43.51%
II	1×500	Glove SBWC	45.01%	48.97%	9.08%	15.32%
II	2×500	Random	81.04%	69.06%	73.61%	71.26%
II	2×500	Glove SBWC	**83.03%**	73.06%	**74.67%**	**73.85%**

Table 4: Results obtained over the development set for binary classification (explicit vs null subject) in the first stage of experimentation.

While models built off a single hidden layer displayed lackluster performance metrics, models with 2 hidden layers were able to obtain promising results, serving as a proof of concept for the application of an NMT technique to the problem. Results for 2-layer models as depicted in figure 2 show that models based on Approach II (using **v** and **w** tags for verbs with explicit or null subject, respectively) achieved better results than models based on Approach I (using a **0** mark as a null-subject stand-in). This led us to discard Approach I models in favor of Approach II for all later experiments.
Furthermore, results hinted at an improvement of

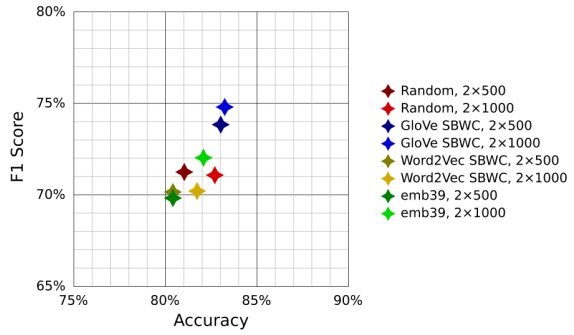

Figure 3: Scatterplot of F1 score and accuracy metrics of the models differing on hidden layer size and word embedding type, as described in table 5.

performance when using pre-trained word embeddings, with the model which incorporated Glove SBWC embeddings consistently outperforming its random embeddings counterpart. These observations motivated a second stage of experimentation where further word-embedding collections were tested, as well as the effect of rising the size of hidden layers from 500 to 1000 units per layer. Results for this second series of experiments are reported on table 5 and shown on figure 3.

Embeddings	Layers	Accuracy	Precision	Recall	F1 Score
Random	2×500	81.04%	69.06%	73.61%	71.26%
Random	2×1000	82.61%	74.27%	**78.16%**	71.09%
Glove SBWC	2×500	83.03%	73.06%	74.67%	73.85%
Glove SBWC	2×1000	**83.25%**	**77.87%**	71.99%	**74.81%**
W2V SBWC	2×500	80.42%	75.19%	65.77%	70.17%
W2V SBWC	2×1000	81.74%	76.61%	64.82%	70.22%
emb39	2×500	80.42%	73.37%	66.63%	69.84%
emb39	2×1000	82.09%	77.74%	67.11%	72.04%

Table 5: Results obtained over the development set for binary classification (explicit vs null subject) in the second stage of experimentation.

Results show that GloVe-based word embeddings allowed for a significantly better performance, even when compared to models using Word2Vec embeddings pre-trained over the same corpus. An increase from 500 to 1000 units per hidden layer also appeared to contribute to better performance although to a lesser degree than the election of a word embedding collection.

A third round of experimentation was carried on, focused on evaluating the performance of larger models with the top-performing Glove SWBC embeddings. Models with up to 4 hidden layers and up to 1500 units per hidden layer were tested. Results, as depicted on table 6, show the limits of increasing the model's dimensionality. In particular, a sharp decline in performance is observed when increasing the number of hidden layers to 4 whereas models with 1000 units per layer outperformed their counterparts with 1500 units as per the F1 score metric.

Layers	Accuracy	Precision	Recall	F1 Score
2×500	83.03%	73.06%	**74.67%**	73.85%
2×1000	83.25%	77.87%	71.99%	74.81%
2×1500	83.55%	**82.12%**	68.07%	74.44%
3×500	82.66%	76.51%	72.56%	74.48%
3×1000	**84.15%**	78.31%	73.52%	**75.84%**
3×1500	83.55%	81.39%	67.30%	73.68%
4×500	66.91%	56.42%	43.69%	49.25%
4×1000	72.51%	60.60%	46.46%	52.60%

Table 6: Results obtained over the development set for binary classification (explicit vs null subject) in the third stage of experimentation.

The best results were observed for models with 3 hidden layers of size 1000. It should be noticed, however, that the increase in performance is moderate when compared to smaller 2 hidden layer models whose training is less computationally expensive. Table 7 compares training and execution times for the three best performing models, evidencing a trade-off between better performing larger models and faster to train smaller models, suggesting that for particular resource-sensitive applications it might be preferable to opt for a less resource-intensive variant, even at the cost of lesser accuracy.

Layers	Training steps	Training time	Processing time	F1 score
2×500	6000	1 hours	10.2 ms	73.85%
2×1000	12000	2.5 hours	16.2 ms	74.81%
3×1000	90000	10.5 hours	19.2 ms	75.84%

Table 7: Training and execution stats for selected models. Training steps measure the number of iterations over the training set; processing time measures the average time for executing the model on a sentence from the development set. Training times as measured using an entry-level GPU (Nvidia GeForce GTX 1650).

6.2. Ternary Classifiers

The three most promising binary classification models, as identified on the third experimentation stage (table 6) were subsequently extended to the ternary classification case, incorporating the distinction between null-subject phrases arising from an optionally omitted subject and null-subject resulting from a verb in an impersonal usage where no subject can be admitted. As shown in table 8, models with 2 hidden layers outperformed the model with 3 hidden layers in this scenario, and were also much faster to train.

6.3. Results

Final results for the NMT-based classification model are presented in table 9. There appears to be a correlation between the F1-score for different verb categories and their support within the corpus, with explicit subjects (which account

Layers	Training steps	Training time	Macro-F1 score
2×500	10 000	1 hours	69.82%
2×1000	12 000	2.5 hours	**70.90%**
3×1000	110 000	16.5 hours	56.44%

Table 8: Training stats for ternary classification (explicit subject vs null-subject vs impersonal verb). Macro-F1 score computed over the development set.

for 71.7% of finite verbs within the training set) achieving the best results whereas the impersonal verb category (corresponding to only 2.9% of finite verbs in the training set) obtains less accurate results.

The best performance (as per macro F1 score) was achieved by a model with two hidden layers and 500 units per layer. A confusion matrix for this classifier is given in table 10.

Layers	Training steps	Explicit subj. F1	Null-subj. F1	Impersonal F1
2×500	10 000 steps	90.12%	**76.85%**	**50.47%**
2×1000	12 000 steps	**90.69%**	76.29%	46.43%
3×1000	110 000 steps	86.05%	63.31%	15.05%
Layers	Macro F1	Weighted F1	Accuracy	TER
2×500	**72.48%**	86.19%	82.96%	4.93%
2×1000	71.14%	**86.41%**	**84.14%**	**4.92%**
3×1000	54.80%	79.27%	76.84%	16.31%

Table 9: Performance for the ternary classification models (explicit subject, null-subject or impersonal) over the test set. Weighted F1 is the F1 score weighted by each category support. TER measures the distance between the output of the NMT model and the expected output sequence.

Actual category	Detected as Explicit	Detected as Null-subject	Detected as Impersonal	Not detected
Explicit	**2636 (86.6%)**	187 (6.2%)	15 (0.5%)	205 (6.7%)
Null-subject	158 (14.9%)	**790 (74.4%)**	8 (0.8%)	106 (9.9%)
Impersonal	13 (22.8%)	17 (29.8%)	**27 (47.4%)**	0 (0%)
(Spurious)	7	6	0	-

Table 10: Confusion matrix for the best performing model (2 hidden layers with 500 units each). Category values are retrieved from the output of a translation model which might occasionally fail to generate the expected number of results (see section 5.2). As a consequence, the classifier may fail to produce a valid classification value for a verb (not detected) or produce supernumerary values (spurious). Results evaluated over the test set.

6.4. Comparison to Previous Work

Table 11 compares the final results for the proposed classifiers and those reported in the existing literature. It can be concluded that null-subject classifiers based on a machine translation approach are able to achieve a performance comparable to those of more traditional methods. In particular, it must be noted that one of the 2x500 model obtained the highest F1 score for Spanish null subject recognition, 0.769, constituting an improvement in the state of the art over the score reported by Rello et al. (2012) (on a different corpus), and the scores reported for AnCora as well. The Macro F1 in general is on par with the one reported by (González and Martínez, 2018), which uses the unmodified version of AnCora.

Model	Corpus	Explicit S F1	Null-S F1	Impersonal F1	Macro F1
2×500	NSAnCora	90.1%	**76.9%**	50.5%	72.5%
2×1000	NSAnCora	90.7%	76.3%	46.4%	71.1%
3×1000	NSAnCora	86.1%	63.3%	15.1%	54.8%
González (2018)	AnCora	**90.9%**	71.8%	**55.5%**	**72.7%**
Gerez (2019)	AnCora	89%	67%	34%	63%
Rello (2012)	ESZIC	91.2	75.5%	72.7	79.8
González (2018)	ESZIC	89.1%	66.3%	62.2%	73.9%

Table 11: Comparison between results obtained in this study and previous works on different corpora. Top results over the AnCora corpus variants are highlighted.

While our classifier proved capable of distinguishing between explicit and null subject phrases with a reliability comparable or even surpassing previous proposals, it was not possible to achieve a similar level of performance in the identification of impersonal verbs. This might be a consequence of the limited number of examples for this category within the corpus; one way to improve this could be training on a custom corpus with an artificially increased ratio of impersonal verb phrases, as in Gerez et al. (2019). Note that the results for impersonal verbs in González and Martínez (2018) are not directly comparable, because they considered only the reflexive impersonal sentences marked in AnCora, while we considered other cases as well.

As evidenced by the confusion matrix given in table 10, failure to detect verbs during the translation process constitutes a major source of error. Further analysis of the results indicated that failed detections were often a consequence of verbs not being found within the vocabulary compiled from the training set and, as a result, being represented by the generic 'unk' token, irrespective of their POS. As nouns outnumber verbs in the corpus, models tend to classify OOV tokens as nouns. This problem is compounded by the fact that Spanish morphology allows for verbs to take a great number of conjugations depending on their person, tense, aspect and mood; a verb appearing in a form not covered in the training corpus will be treated as an OOV token even if other inflections of the same lemma are represented. Possible improvements include the usage of a larger vocabulary, pre-processing the input to transform OOV verb forms into a surrogate verbal value (e.g. Chiruzzo (2020) uses the embedding of the most frequent word given POS and morphological features), or the addition of lemma information to account for inflectional forms.

7. Conclusions

In this work, we presented a series of experiments adapting an NMT approach to the task of detecting explicit subjects, null subjects, or impersonal verbs in Spanish. Experimental results validated the approach, achieving results comparable to or surpassing those reported in prior literature, with a Macro F1 score of 0.725, in a similar range to the 0.727 score obtained in previous works on a similar corpus (González and Martínez, 2018), and an improvement to the state of the art in the recognition of null-subjects in particular, with an F1 score of 0.769 for this category.

This approach also proved to be efficient in the usage of resources both during the training of the models and its execution. The top-performing model was trained within 2.5 hours on a laptop with an entry-level GPU (Nvidia GeForce GTX 1650, 930-1395 MHz), processing input at a rate of 10.2 ms per sentence. It should also be noted that an NMT translation model has a time complexity of order $O(n^2)$ for an entry of length n (due to its usage of an attention mechanism), in contrast to $O(n^3)$ traditional syntactic parsers that might be applied to the same task.

In order to generate a custom aligned parallel corpus for the NMT model, a new Spanish-language corpus focused on null-subject and impersonal verb recognition was created by extracting information out of the existing AnCora corpus (Delor, Mariona and Martí, Antonia and Recasens, M., 2022) and manually annotating instances of verbs whose classification could not be deduced from its information. This resulted in a new resource dubbed 'NullSubjAncoraCorpus' which has been made available to the public in order to facilitate the creation and evaluation of further developments in the area.

8. Bibliographical References

Agustin Azzinari and Alejandro Martínez. 2016. Representación de palabras en espacios de vectores. Proyecto de grado. Universidad de la República. Uruguay.

Luis Chiruzzo. 2020. *Statistical Deep Parsing for Spanish*. Ph.D. thesis, Universidad de la República, Uruguay.

Federico Ferlito. 2021. *Getting the Null Subject Right with Neural Machine Translation*. Ph.D. thesis, University of Groningen.

Antonio Ferrández and Jesús Peral. 2000. A computational approach to zero-pronouns in spanish. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 166–172.

Ernesto Gerez, Juan Pablo Irazusta, and Mauricio Irace. 2019. Detección y resolución de sujetos nulos en español. Proyecto de grado. Universidad de la República. Uruguay.

Lucía González and Verónica Martínez. 2018. Detección de sujetos omitidos en el español. Proyecto de grado. Universidad de la República. Uruguay.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Luz Rello, Ricardo Baeza-Yates, and Ruslan Mitkov. 2012. Elliphant: Improved automatic detection of zero subjects and impersonal constructions in Spanish. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 706–715, Avignon, France. Association for Computational Linguistics.

Gabriel Stanovsky and Ido Dagan. 2018. Semantics as a foreign language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2412–2421.

Oriol Vinyals, Łukasz Kaiser, Terry Koo, Slav Petrov, Ilya Sutskever, and Geoffrey Hinton. 2015. Grammar as a foreign language. *Advances in neural information processing systems*, 28.

Daniel et al Zeman. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

9. Language Resource References

Luis Chiruzzo and Dina Wonsever. 2018. Spanish hpsg treebank based on the ancora corpus. In

Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Delor, Mariona and Martí, Antonia and Recasens, M. 2022. *AnCora Spanish 2.0.0*. distributed via ELRA: ELRA-Id ELRA-W0326, 2.0, ISLRN 252-495-813-736-1.