Cross-lingual Embeddings for Sentiment Analysis

Iñigo Alonso

Abstract

The main contributions to the field of Natural Language Processing (NLP) in the few last years have come thanks to the emergence of word embeddings. One of the challenges of this ground-breaking technology where researchers have set their sights on, is expanding it to low resource languages where there is not enough data to accurately train monolingual embeddings. With this purpose, Facebook developers trained and made available to the public a set of cross-lingual embeddings named MUSE. These embeddings were trained on English (high resource language) and can be used for different NLP tasks on different languages. This project proves the validity of their usage for sentiment analysis. Specifically, this case study focuses on movie reviews in Spanish.

1. Definition

1.1 Problem Statement

Since the representation of words on a vector space of a number came out first in 2013 (Mikolov et al), it has become the most ubiquitous technique in the field of Natural Language Processing (NLP). There are a variety of problems for which its use has been studied, text classification (Peng Jin et al , Qian Liu et al), document clustering (Inzamam Rahaman et al), part of speech tagging(Yu Dong et al), named entity recognition (Cıcero Nogueira dos Santos et al) sentiment analysis (Ruggero Petrolito et al). In particular, the project will focus on the latest of these tasks.

As it happened in the field of image recognition, where the weights from the most successful models in the ImageNet challenge where used from transfer learning, the embeddings that have performed the best until this day have been made available for the public. Being the most popular of this resources the GloVe embedding dataset(Jeffrey Pennington et al).

The current project utilizes a similar resource, the MUSE embeddings (Alexis Conneau et al). These MUSE embeddings are a cross-lingual set of pre-trained embeddings, which means that, they were created with the goal of fixing the approach that was in place for cross-lingual document classification. This old approach involved translating from a language to English and using an English classifier to classify the text originally written in another language. With this procedure, the errors in the translation get propagated and increase the error rate in the classification.

The resulting MUSE embeddings are language independent, that is, each language gets its own set of embeddings. With the constraint that all these embeddings are represented in the same vector space.

Hence, these state-of-the-art embeddings allow to take advantage of the data in high resource languages to perform NLP tasks over other low resource languages. Providing state-of-the-art technology to every language, which could be considered a milestone on the efforts to achieve globalization in the machine learning field.

In this paper, I examine the performance of cross-lingual embeddings. This evaluation is performed for Sentiment Classification of movie reviews, a task more complex than the traditional topic classification problem. The multilingual embeddings whose validity is tested obtained from English (high resource language) and proved on Spanish (lower resource language).

1.1.1 The Cross-lingual Embeddings Problem

While the accuracy of the embeddings claimed by the Facebook Developers Group is remarkable, there are some concerns about their application to sentiment analysis.

First, the projection of the words to a common space fails to explain how words that suffer from polysemy and synonymy in their original language are translated into this common space.

Other issue that will have to be solved in the preprocessing steps, is the absence of some of the words from the movie dataset in the embeddings and vice versa. This happens because some of the movie characters and titles are words that are not found in the resources used to train the multilingual embeddings. In the same way, the corpus on which the MUSE embddings are trained is way larger than the 50,000 movie reviews in English.

The nature of this corpus is not composed of reviews for the most part, which makes it less sensitive to capture sentiments. A case of this could be the words sucks, used for reviews in a very different way depending on its contextual meaning.

1.1.2 The Sentiment Analysis Problem

The distinction between positive and negative reviews is a fairly easy task to perform for humans, especially in comparison to that of the standard text categorization problem, where topics can be closely related. Nonetheless, machine learning algorithms have thrived to a major extent with the latter type of problems.

As pointed Shivakumar Vaithyanathan et al, the complexity of the problem is not appreciated by human intuition, as we fail to determine that a sufficiently long list of positive and negative sentiment adjectives such as *excellent* and *terrible* would suffice to achieve considerable performance on the task. Nonetheless, this approach fails to acknowledge the importance of comparative structures like *still* or as a *matter of fact* that having proved be as paramount of indicators as the previous adjectives. As a result, the strategy mentioned beforehand only yielded 69% accuracy.

Additionally, some preprocessing techniques that identify negation should be considered to speed up calculations. An example of this could be the sentence *I didn't like the scene where…,* the vector representation of *like* should not be the same as the one of *not_like*, for this reason the sentence could be rewrited as *I not_like the scene where…* for better performance. Although the standard LSTM architecture should be able to get feedback from previous words, the previous technique could be helpful to improve accuracy.

1.2 Project Overview

1.2.1 Cross-lingual embedding models

Many have been the researchers who have tried to take advantage of the fact that cross-lingual embeddings do not need parallel data to perform well in NLP tasks over low resource languages.

As a result, a first wave of cross-lingual models was published, which showed good performance in tasks like name entity recognition (Ying Lin et al, Shyam Upadhyay et al), Dependency parsing (Shyam Upadhyay et al, Jiang Guo et al) or Document Classification (Tianze Shi et al). These models achieved the first millestones for cross-lingual word embeddings, and allowed for their application to more complex problems such as the one studied in this project, sentiment analysis.

1.2.2 Sentiment Analysis

The complexity of sentiment analysis lies on particular traits that no other NLP task shares, and that are even more accentuated when working with online reviews. Some of these traits are subjectivity, tone, context, Irony, Comparisons, Emojis or the fact that it hard to define objectively a line between what is positive, negative or neutral. To make all this elements clear let's point out some answers the question Was the movie good? that make explicit the previous points. 1.Yeah. Sure.(Sarcastic tone). 2. It was second to none. 3. ¯_(ツ)_/¯. For these reasons, sentiment analysis task seem to be between 70 and 85% accuracy.

For the present problem (the application of this type of embeddings to sentiment analysis), other pre-trained non-cross-lingual embeddings, different to the ones provided by Facebook (MUSE embeddings), have already been proven to work successfully (Sebastian Ruder et al, Jan Deriu et al). The best accuracy measure (82%) of these 2 works, was obtained for the case where the deep learning architecture is trained on English corpus and utilized for sentiment analysis in other languages.

It has to be noted that both of these models used a Convolutional Neural Network (CNN) architecture. In the current problem, a Long Short Term Memory (LSTM) architecture is tested. While this approach could result costlier, the higher number of parameters that are used comparing to the CNN the following for 3 main reasons, made us believe that similar performance could be reached: 1. Since the

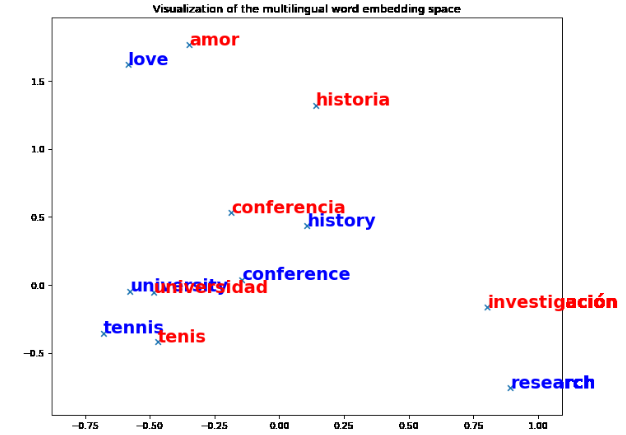


Figure 1. 2 dimensional representation of the closest neighbors in Spanish and English.

LSTM concept was introduced in 1997 by Sepp Hochreiter et al. it has consistently been proved its validity to model sequential data. 2. The MUSE embeddings have been trained on fastText and a combination of data from Facebook and Wikipedia, which constitutes a huge data resource. 3. The developers at Facebook have claimed one their blog an accuracy close to 95% on a language-specific classifier, when operating on languages not originally seen in training. Which is well over the results reached in previous works.

1.2.3 MUSE Embeddings

The particularity of the cross lingual word embeddings is the following: if you represent the embedding of a word in English and the one of the same word in a different language, using t-SNE or a similar technique, you will be able to see both words closely together in the vector space. This is just a way that shows the similarity across a chosen dimensionality space between the vectors representing 2 words that have the same meaning, but belong to different languages.

This magic, does not occur automatically when training each embedding on its own language, it requires to project the language independent embedding on a common space.

To make the projection happen a projection matrix W, is learnt. This matrix W, is the result of the optimization over a cost function with the goal of minimizing the loss between the word x_i in English and its translation in the objective language, y_i.

$$M = argmin_W \sum_i \|x_i - Wy_i\|^2$$

When one analyzes the process to obtain the MUSE embeddings presented in the work by (Alexis Conneau et al), 3 main steps are identified:

1. Get monolingual embeddings: trained separate embeddings for each language using fastText and a combination of data from Facebook and Wikipedia.

2. Alignment of monolingual embeddings: The projection of the embeddings into a common space can be done through 2 approaches.

a) The supervised way: This method uses parallel data, using a bilingual dictionary, it learns a mapping from the source to the target space by solving the orthogonal Procrustes problem.
b) The unsupervised way: Does not require parallel data. In this case it uses, adversarial training to do the mapping, and instead of, by minimizing the loss function the mapping is performed by maximizing the similarity measure CSLS.

3. Evaluation of Cross-lingual embeddings for several tasks.

1.3 Metrics

Being a supervised learning task and a classification problem accuracy metric will be paramount.

Nonetheless, for this particular task the F-β Score can end up being more interesting. The goal of a movie classifier needs to be considered to understand this. The goal of the algorithm is that the user views all the good movies available, it is not that concerning if the user watches a bad movie from time to time. As a movie fan one can live with that.

For this reason, besides calculating the precision and the recall, a F-β Score, where β=2,

On this case, for the alignment of the cross-lingual embeddings, I missed the document in the MUSE github where the alignment of the words was provided, in this case the index matching was done following the nearest neighbor approach over both embedding sets. That is, for every English embedding its nearest neighbor in Spanish was calculated, and associate the same index.

2. Analysis

2.1 Data Exploration and Visualization

2.1.1 High Resource Language Data

The training data consists of 50,000 reviews in English from IMDB, allowing no more than 30 reviews per movie. The constructed dataset contains an even number of positive and negative reviews, so a naive classifier would yield 50% accuracy.

For better results, only highly polarized reviews have only been considered.

In particular, the review is labelled as a negative review if it has a score ≤ 4 out of 10, and positive, if has a score ≥ 7out of 10. Neutral reviews are not included in the dataset.

This dataset was made available to the public after (Andrew L. et al.) provided a benchmark for future work in this area.

The training and test sets are divided in 40,000 and 10,000 splits.

Both sets have 50/50 ratio of positive and negative reviews. Even though the preprocessed dataset is available through keras, the decision to do my own preprocessing was taken. The reason being, not inducing any type of imbalance or bias in the preprocessing of the Spanish and English datasets. The number of total words is 35,741. The average review length is 218 words and 75% of the reviews have less than 265 words.

2.1.2 Low Resource Language Data

The test data is provided by Dr. Fermín L. Cruz Mata from the University of Seville (http://www.lsi.us.es/~fermin/index.php/Main_Page) containing 4380 movie reviews in Spanish, extracted from the site www.muchocine.net. This movie reviews are ranked from 1 to 5. Which, so as to follow the polarized approach in the training set, I considered as positive if the rate is 4 or 5 and negative if it is 1 or 2. Hence, once the reviews of rating equal to 3 are discarded, the final dataset size is 2,275 movies.

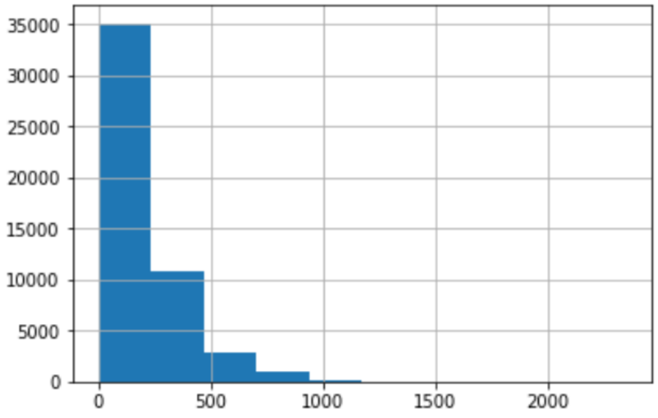


Figure 2. Histogram of the word distribution in the English Dataset

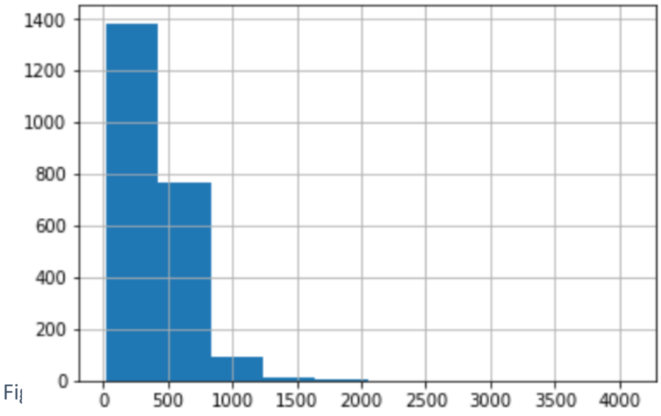


Figure 4. Histogram of the word distribution in the Spanish Dataset

Table 1. Summary of the Word count statistics in both datasets IMBD (English) and Muchocine (Spanish)

STATISTICS OF THE DATASETS								
	count	mean	std	min	25%	50%	75%	max
IMBD	50,000	218.28	161.55	2	120	164	264	2342
MUCHOCINE	2,275	437.81	258.01	21	284	376	522.5	4,082

Being 1,145 of those positive and 1,130 negative. Resulting in 20,385 unique words. As a remark, it has to be stated that reviews are longer than the ones in English, with an average review of 438 words and 75% of the reviews having more than 284 words
On both datasets the reviews are randomly shuffle to make sure no bias, such author repetition is included.

2.2 Algorithms and Techniques

The architecture tried for now is a many-to-one LSTM setting with a softmax layer in the output layer.

This model was trained with the training set of 40,000 documents in english over just 5 epochs due to limited computing resources.

In particular, this is one of the limitations of the chosen methodology compared to other ones used in NLP. Other architecture like RNTN models using Sentiment trees, Convolutional networks, Doc2vec or SVM are more computationally inexpensive.

Nonetheless, in the current project, as computational efficiency is not within the goals, the Long Short-Term Memory Recurrent Neural Network model is picked. It has a proven track record in modelling sequential data, as it is the algorithm fueling the state-of-the-art technologies in the speech recognition and translation field. It also has the memory and power to capture the short-term relationships that some of the traits described in 2.1 involve.

Some of the strengths and weaknesses around this choice will be presented in the Error Analysis section.

As per the functions of the gradient descent optimizer used during the training of the neural network, Adam optimizer was used (Diederik P. Kingma et al.):

$$\alpha_t = \alpha \,.\, \sqrt{1 - \frac{\beta_2^t}{1 - \beta_1^t}}$$

$$\theta_t \leftarrow \theta_{t-1} - m_t\,/(\sqrt{v_t} + \hat{\theta}$$

where the picked values are a learning rate of 0.001, a β1 value of 0.9, a β2 of 0.999 and a epsilon value of 0. This optimization algorithm implements the momentum and the adaptive learning rate concepts.

For the loss function, the binary cross-entropy function used was:

$$CE = -\sum_{i=1}^{C'=2} t_i \log\big(f(s_i)\big) = -t_1^i \log\big(f(s_1)\big) - (1 - t_1)\log(1 - f(s_1))$$

2.3 Benchmark

As described above, the study of the performance of cross-lingual embedding is pretty new so there are not any papers having studied the performance of cross-lingual embeddings for sentiment classification. The closest that was found, is the paper on Document Classification by Tianze Shi et al, that studied the problem over English and

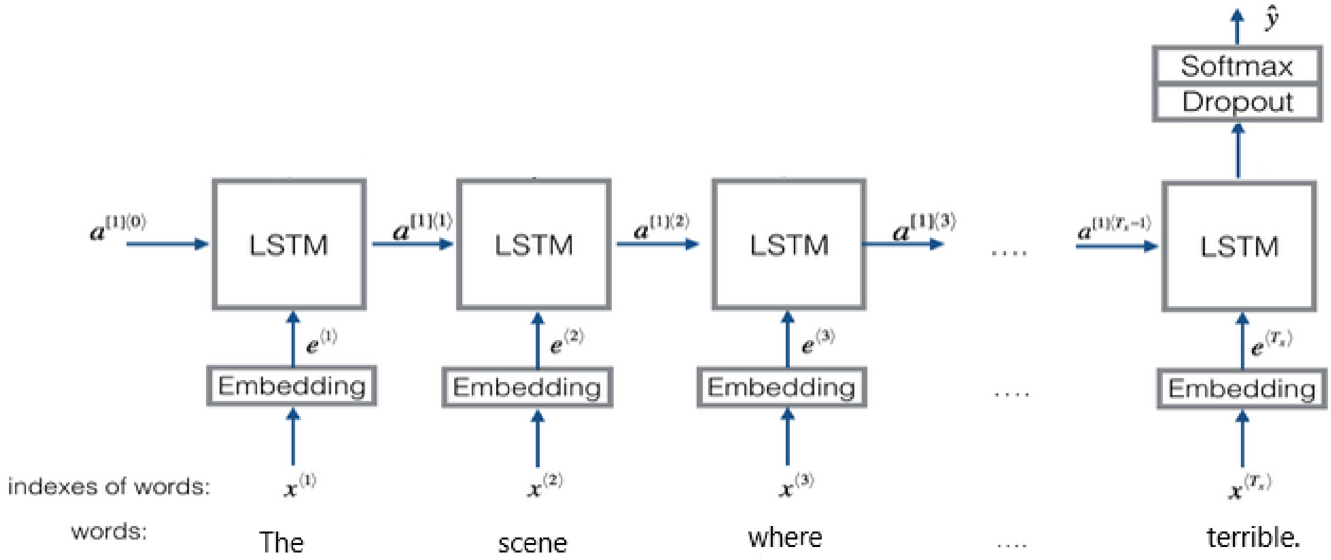


Figure 5. Representation of the Network architecture.

German. The accuracies in a task way easier than sentiment classification are shown in the table 2. In the language where the model is not trained it is registered an accuracy ranging from 75% to 80% depending on the model used.

For this lack of references, I have created my own benchmark for the performace of the embeddings over the English corpus. I trained my own embeddings on the English reviews as a benchmark.

For the performance of the model over the Spanish data, a drop in the accuracy from 12 to 15% with respect to the obtained in the English data is expected, just like it happened on the Tianze Shi et al paper.

3. Methodology

3.1 Data preprocessing and Implementation

I used documents from the movie-review corpus described in Section 3. To create a dataset with uniform class distribution (studying the effect of skewed class distributions was out of the scope of this study), the training set randomly selected 20,000 positive-sentiment and 20,000 negative-sentiment documents in English.

To prepare the documents in English the steps followed are the following:

1.Preprocessing: Text transformations such as making it lower case, getting rid of punctuation signs and stopwords; and tokenizing (nltk tokenizer) are performed.

2.Padding to equal length: all the documents were padded to the length of 300, taking the first 300 words, in both the training and the test sets.

3. Add <PAD>, <START> and <UNK> labels to the corpus: Indicating the padding, beginning of the text and unknown words, respectively.

4. Create the embedding matrix: The word vectors from the MUSE pre-trained embeddings were considered only if the word appeared in the documents, this way a null vector is assigned to null words.

The Spanish document preparation began with the same first 3 steps. After that, the closest English Neighbor was calculated for each Spanish word, so as to match indexes (pointed out in section 2.2).

In this step unfortunate events took place, due to text encoding issues in the source files, the Spanish words including accents and inexistent letters in the English alphabet such as ñ or ü were substituted by error symbols. Some of this symbols could be restored and thus, some vocabulary saved. But, in the majority of the cases, the error symbols where irreplaceable and the words had to be deleted,

resulting in a major loss of information due to an encoding error from the original text.

Still, the results obtained are optimistic at 72.97% accuracy of classification using MUSE embeddings on unseen Spanish corpus (15% below the performance on English movie reviews) with a model that has not been trained in Spanish corpus.

In this way, dictionaries matching the indexes to the words in both, English and Spanish where obtained and allowed for the implementation in Keras, as well as the posterior error analysis.

3.2 Refinement

The refinement of the model was highly limited by the computational resources. The computations of a model of a single layer with 100 neurons and of another one having 2 layers of 50 neurons each were cancelled after the ending of the calculations was estimated.

Also, the number of epochs did not exceed 5 due to the same reason.

The only tuning that was effectively performed was the increase of the cut off length of the words to the first 500. This resulted in an accuracy increase of a 0.5%.

4. Results

4.1 Model Evaluation and Evaluation

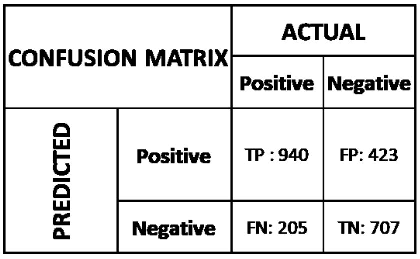
Different experiments are analyzed so as to obtain the real validity of the MUSE embedding for cross-lingual sentiment analysis purposes. Each of scenarios, has set difference references to determine the validity of MUSE embeddings. In the following lines, the experiment results on English corpus are going to be presented first and after that the Spanish ones.

English corpus In order to evaluate the accuracy of MUSE Embeddings for sentiment classification, a reference to be met should be set. While in section 2, the up-to-date milestones on the task were displayed, I decided to test the MUSE Embeddings against my own embeddings. These embeddings where trained in the 40,000 samples of the IMBD dataset, the same training set that is also used to train the weights of the model where the MUSE embeddings are tested.

The trained embeddings set a standard of 86.94 % accuracy, that was outperformed by the MUSE embeddings with 87.56%. This is a clear representation of the potential of MUSE embeddings, which are able to surpass the accuracy by embeddings which have been trained specifically for sentiment classification of movie reviews. Besides, this analysis lied some initial thoughts to the performance that could be expected on Spanish corpora.

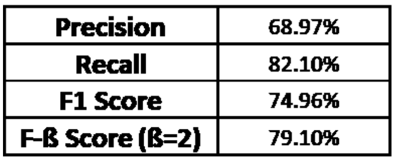
Spanish corpus The initial inspection of the Spanish corpus casted some doubts on whether the dataset was appropriate. Too much of informality in both the wording and the tone was detected, which added up to the skepticism rose by the lost vocabulary due to the text format. For this reason, it was decided to translate the Spanish reviews to English using one of the most cutting-edge translators out there, the Google Translator itself. This apparently tedious task, was helpful to determine that the dataset could effectively be used for this task, this was proven by the 87.26% accuracy of the MUSE embeddings (pretty much same performance as in original English movie reviews).

Table 2. Confusion Matrix of the classification results over the Spanish



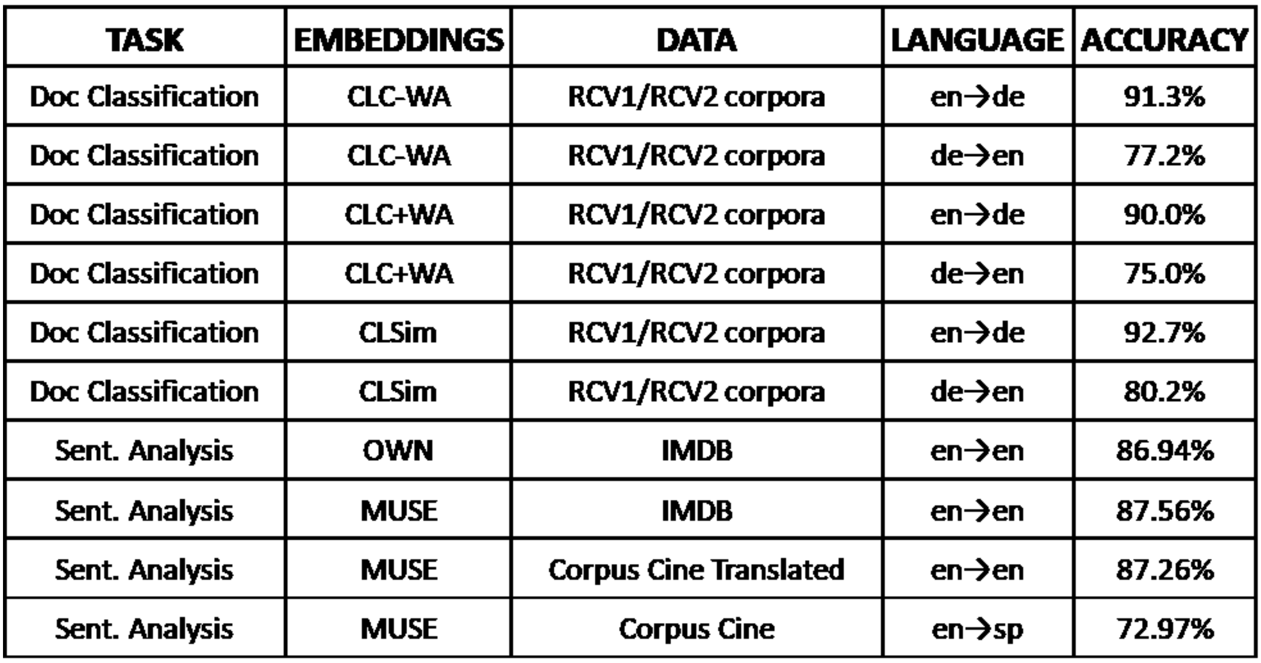
CONFUSION MATRIX		ACTUAL	
		Positive	Negative
PREDICTED	Positive	TP : 940	FP: 423
	Negative	FN: 205	TN: 707

Table 4. Metrics of the sentiment classification in the Spanish dataset.



Precision	68.97%
Recall	82.10%
F1 Score	74.96%
F-ß Score (ß=2)	79.10%

Table 3. Summary of task accuracies. The first 6 entries correspond to the only paper (Tianze Shi et al). that, until this date, has analyzed cross-lingual embeddings on classification tasks (document classification, easier on paper than sentiment classification). The las 4 entries correspond to the experiments run in this project. Note: In the language column trained language → predicted



TASK	EMBEDDINGS	DATA	LANGUAGE	ACCURACY
Doc Classification	CLC-WA	RCV1/RCV2 corpora	en→de	91.3%
Doc Classification	CLC-WA	RCV1/RCV2 corpora	de→en	77.2%
Doc Classification	CLC+WA	RCV1/RCV2 corpora	en→de	90.0%
Doc Classification	CLC+WA	RCV1/RCV2 corpora	de→en	75.0%
Doc Classification	CLSim	RCV1/RCV2 corpora	en→de	92.7%
Doc Classification	CLSim	RCV1/RCV2 corpora	de→en	80.2%
Sent. Analysis	OWN	IMDB	en→en	86.94%
Sent. Analysis	MUSE	IMDB	en→en	87.56%
Sent. Analysis	MUSE	Corpus Cine Translated	en→en	87.26%
Sent. Analysis	MUSE	Corpus Cine	en→sp	72.97%

Once the Spanish data was validated, the real challenge was encountered. The MUSE cross-lingual embeddings gave a 72.97% accuracy on original Spanish data.

This result is positive and shows the potential of cross lingual embeddings for cross-lingual sentiment analysis. Specially, considering the computational limitations of the model and preprocessing, where reviews were cut to 300 characters. Other limitations where that an average of 50 words were lost per Spanish review (from the 300 selected words) and the LSTM model had a single layer of 50 units.

4.2 Justification

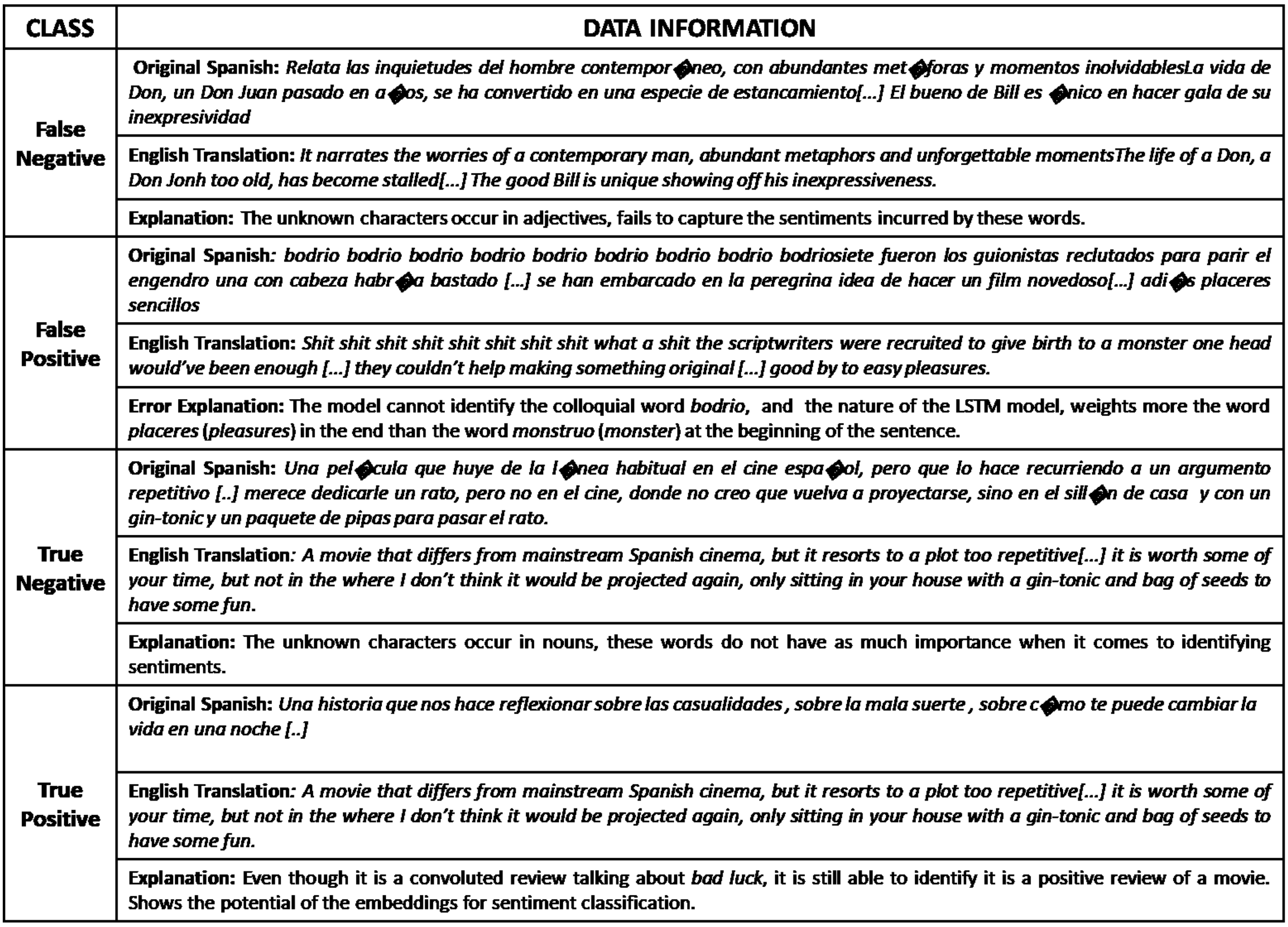
Even if the accuracy recorded in the metrics is pretty good there are some points that need to be pointed out to understand the real upside of cross-lingual embeddings.

First, the Spanish dataset was the main source of problems. The set had some particularities that have hampered further performance of the algorithm.

Let’s start with the length of the reviews. The reviews are on average way longer than the ones in English, having some of them has 2,000+ words. Apparently, while this may seem like a major issue it was not initially backed up by the numbers as the classification accuracy obtained on the truncated Spanish reviews was superior to the average one (73.40%). Nonetheless, a second analysis was performed to on an attempt to prove that longer reviews would result on better performance, this time, extending the length to 500 words per review. This time, the results backed up my assumption as the overall performance of the model rose to 73.56%.

On the other hand, the tone and wording was very informal on some cases and included to many curse words. As a consequence, it means that the reviews

Table 5. Understanding the model. Examples of reviews and how the classifer perfomed on each case.



CLASS	DATA INFORMATION
False Negative	**Original Spanish:** *Relata las inquietudes del hombre contempor�neo, con abundantes met�foras y momentos inolvidablesLa vida de Don, un Don Juan pasado en a�os, se ha convertido en una especie de estancamiento[...] El bueno de Bill es �nico en hacer gala de su inexpresividad*
	English Translation: *It narrates the worries of a contemporary man, abundant metaphors and unforgettable momentsThe life of a Don, a Don Jonh too old, has become stalled[...] The good Bill is unique showing off his inexpressiveness.*
	Explanation: The unknown characters occur in adjectives, fails to capture the sentiments incurred by these words.
False Positive	**Original Spanish***: bodrio bodrio bodrio bodrio bodrio bodrio bodrio bodrio bodriosiete fueron los guionistas reclutados para parir el engendro una con cabeza habr�a bastado [...] se han embarcado en la peregrina idea de hacer un film novedoso[...] adi�s placeres sencillos*
	English Translation: *Shit shit shit shit shit shit shit shit what a shit the scriptwriters were recruited to give birth to a monster one head would’ve been enough [...] they couldn’t help making something original [...] good by to easy pleasures.*
	Error Explanation: The model cannot identify the colloquial word *bodrio*, and the nature of the LSTM model, weights more the word *placeres* (*pleasures*) in the end than the word *monstruo* (*monster*) at the beginning of the sentence.
True Negative	**Original Spanish:** *Una pel�cula que huye de la l�nea habitual en el cine espa�ol, pero que lo hace recurriendo a un argumento repetitivo [..] merece dedicarle un rato, pero no en el cine, donde no creo que vuelva a proyectarse, sino en el sill�n de casa y con un gin-tonic y un paquete de pipas para pasar el rato.*
	English Translation*: A movie that differs from mainstream Spanish cinema, but it resorts to a plot too repetitive[...] it is worth some of your time, but not in the where I don’t think it would be projected again, only sitting in your house with a gin-tonic and bag of seeds to have some fun.*
	Explanation: The unknown characters occur in nouns, these words do not have as much importance when it comes to identifying sentiments.
True Positive	**Original Spanish:** *Una historia que nos hace reflexionar sobre las casualidades , sobre la mala suerte , sobre c�mo te puede cambiar la vida en una noche [..]*
	English Translation*: A movie that differs from mainstream Spanish cinema, but it resorts to a plot too repetitive[...] it is worth some of your time, but not in the where I don’t think it would be projected again, only sitting in your house with a gin-tonic and bag of seeds to have some fun.*
	Explanation: Even though it is a convoluted review talking about *bad luck*, it is still able to identify it is a positive review of a movie. Shows the potential of the embeddings for sentiment classification.

were riddled with some of the traits that make sentiment analysis a task so hard to succeed in, such as irony or language-specific expressions. In Table 4, there is an example of a false positive review that shows the poor text wording (*bodrio* is a colloquial word for shit that does not appear in the embeddings) and the limitations of the LSTM model that weight more the last words that the earlier ones, in order to classify a sentiment.

Lastly, to top it all off, some of the characters had unreadable encoding. This meant the loss of the majority words with an accented letter or a special character such as ñ or ü.

As a consequence, there was a loss of words and cutting of the reviews. The numbers that represent this phenomenon are that only 40% of the Spanish embeddings were used. As stated, the number of lost words per review was 50. Related to this, it is remarkable that out the average of words lost is the same in the rightly classified words and the misclassified ones.

Although this may suggest that the missing words did not have such an effect on the prediction outcome, this is highly dependent on the type of missing words. This is clearly shown on the false negative and true negative examples on the Table 4. The impact if the missing words are adjectives is way higher than if the missing words are nouns.

The point that more computationally expensive approaches would have been favorable to prove the validity of MUSE embeddings, was proved by the analysis performed where the reviews were truncated or padded to 500 words instead of 300. As a result, the accuracy of the model rose 0.5 per cent.

In the same way, a better model could have been built by adding a second layer of LSTMs. But at this point, the added complexity and computational time would have not been worth it, and it would be more interesting to study other models like convolutional networks or SVM, that perform just as good as LSTMs (or even better) for sentiment analysis, while being computationally less expensive.

Lastly, there is a really important point not to be overlooked. The model in reality is more effective than what it accuracy suggests, due to its task-specificity.. In other words, a movie sentiment classifier needs to emphasize recall over precision and that is exactly what it does, reaching a F-ß Score of 79.10% when I consider a weight of ß =2 to make the F1 score task specific (Table 5). The F-ß Score allows us to pick which metric is more important for the case study; the recall in this case. So, the bigger the value of ß, the higher the emphasis on recall, whereas a value closer to 0 means more importance given to the precision metric. The formula that defines the F-ß Score is:

$$F_{\beta} = (1+\beta)^2 \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision + Recall)}$$

4. Conclusion

4.1 Free- Form Visualization

Free form Visualization is provided on the Table 4.

4.2 Reflection

The results provided by the MUSE embeddings are quite good and have met my initial expectations. Comparing to the performance on the English dataset, the drop was around 15%, which considering the limitations of the Spanish dataset that were found in the preprocessing, it is remarkable.

As stated, only 40% of the Spanish embeddings were used, the reason being that it the reviews were truncated or padded up to 300 words (only around 25% of the reviews have less than 300 words). Additionally, I faced the loss of mostly all of the words including unknown characters (words with accents or including letters like ñ or ü), which as a Spanish speaker believed it would throw down completely the accuracy of the model (the amount of accented words in Spanish is huge). Many words had to be deleted due to formatting issues (50 per review of length 300, on average).

In this sense, I believe that if I had had the resources to increase the length considered for each review and to could have kept the lost words, due to the text encoding issues, it could have rose the performance to the high-70s or low-80s

Considering all this setbacks, MUSE embeddings have still been able to live up to the expectations, and it can be underlined finally a powerful method for cross-lingual NLP has been achieved. A fact that transcends the machine learning community. It is paramount to ensure that countries lacking enough AI infrastructure do not lag behind and have access to the same resources. Hence, it could be stated it is an achievement for the globalization.

When it comes to the application of the cross-lingual embeddings to sentiment analysis applied to movie reviews, a couple of particularities are remarked.

First, the difficulties inherent to the sentiment analysis problem (explained in section 2), make the accuracy reached by the cross-lingual embeddings used for this task lower than the one they obtained in other tasks, such as document classification (Table 2), the only classification task where the use of MUSE embeddings has been researched.

Second, the model performs very well for this specific task, sentiment analysis of movie reviews. Even though no tweaks were made on the loss function to guide learning on that direction, the model has better recall metric than precision.

Lastly, a proof of the power of the MUSE embeddings is that they outperform the state-of-the-art technology for sentiment classification mentioned on section 2, by 5%. Which, although the tasks are not the same (the one at Sebastian Ruder et al, Jan Deriu et al, has more categories and is more complex overall), is a good indicator of the quality of the embeddings.

4.3 Improvement

Along the document there have been pointed out ways that might result on better performance. The main alternatives for improvement are solving the encoding problem on the low resource data, building a more computationally expensive LSTM model or utilizing other machine learning approaches

4.3.1 Low resource data

Hours were spent trying to solve the encoding of the unknown characters. When it was explored the hexadecimal code of the text it was observed that the unknown characters had the same hexadecimal encoding, no matter what the missing character was (The missing characters ranged from accented words á, é,… to special characters such as ¿, ¡, ñ…).

So there was no way to recover them, using a "clean" data would certainly help. As shown in the example table, if the missing words are adjectives the impact is great, so in most of the case having all the words readable would certainly impact the model for good.

4.3.2 More complex LSTM model

Due to the dimensionality of the embeddings to be dealt with, as well as the length of the reviews the model takes a long time to train. Thus, only one layer of LSTMs with 50 neurons has been trained.

Adding a layer would increase the complexity of the model and therefore, capture some relationships that are not straightforward.

4.3.3 Other approaches

Deep learning is not the most popular approach for sentiment classification. The main reason being that other less expensive models have shown a similar accuracy (or even better) on the task. Support Vector Machines, Dependency Trees and Naïve Bayes are some of these model types.

Inside the field of deep learning, convolutional networks can be an alternative too. Cícero Nogueira dos Santos and Maíra Gatti outperformed state-of-the-art models of RNNs, NB and SVMs on *Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts.*

Out of the models mentioned in this section, the one using convolutional networks seems the better suited to solve the problem.

References

Tomas Mikolov, Ilya Sutshever, Kai Chen, Greg Corrado and Jeffrey Dean. *Distributed Representations of Words and Phrases and their Compositionality,* 2013.

Peng Jin, Yue Zhang, Xingyuan Chen, Yunqing Xia. *Bag-of-Embeddings for Text Classification*, 2016.

Qian Liu, Heyan Huang, Yang Gao, Xiaochi Wei, Yuxin Tian, Luyang Liu. *Task-oriented Word Embedding for Text Classification*, 2018

Inzamam Rahaman, Patrick Hosein. *Exploiting Gaussian Word Embeddings for Document Clustering*, 2017

Yu, Dong Jin Wang, Wei. *Part-Of-Speech Tag Embedding for Modeling Sentences and Documents* 2016

Cicero Nogueira dos Santos, Victor Guimaraes. *Boosting Named Entity Recognition with Neural Character Embeddings, 2015.*

Ruggero Petrolito, Felice Dell'Orletta. *Word Embeddings in Sentiment Analysis*, 2017

Jeffrey Pennington, Richard Socher, Christopher D. Manning. *GloVe: Global Vectors for Word Representation*, 2014.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Herve Jegou. *Word Translation Without Parallel Data*, 2017.

Ying Lin, Shengqi Yang, Veselin Stoyanov, Heng Ji. *A Multi-lingual Multi-task Architecture for Low-resource Sequence Labeling,* 2018.

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, Dan Roth. *Cross-lingual Models of Word Embeddings: An Empirical Comparison,* 2016.

Jiang Guo, Wanxiang Che1, David Yarowsky, Haifeng Wang, Ting Liu. *Cross-lingual Dependency Parsing Based on Distributed Representations,* 2016.

Tianze Shi Zhiyuan Liu Yang Liu Maosong Sun. *Learning Cross-lingual Word Embeddings via Matrix Co-factorization,* 2015.

Sebastian Ruder, Parsa Ghaffari, John G. Breslin. *Deep Learning for Multilingual, Aspect-based Sentiment Analysis*, 2016.

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, Martin Jaggi. *Leveraging Large Amounts of Weakly Supervised Data for Multi-Language Sentiment Classification,* 2017.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzatoy , Ludovic Denoyerx , Herve Jegouy. *Word Translation Without Parallel Data,* 2017.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng and Christopher Potts. *Learning Word Vectors for Sentiment Analysis,* 2011.

Shivakumar Vaithyanathan, Bo Pang and Lillian Lee. *Thumbs up? Sentiment Classification using Machine Learning Techniques,* 2002.

Diederik P. Kingma and Jimmy Lei Ba. *Adam: A Method For Stochastic Optimization*

https://github.com/facebookresearch/MUSE