Master in Cognitive
Science and Language

Master's Thesis

September 2018

Machine-translation inspired reordering as preprocessing for cross-lingual sentiment analysis

---

Alejandro Ramírez Atrio

Supervisor:

Toni Badia

Abstract

In this thesis we study the effect of word reordering as preprocessing for Cross-Lingual Sentiment Analysis. We try different reorderings in two target languages (Spanish and Catalan) so that their word order more closely resembles the one from our source language (English). Our original expectation was that a Long Short Term Memory classifier trained on English data with bilingual word embeddings would internalize English word order, resulting in poor performance when tested on a target language with different word order. We hypothesized that the more the word order of any of our target languages resembles the one of our source language, the better the overall performance of our sentiment classifier would be when analyzing the target language. We tested five sets of transformation rules for our Part of Speech reorderings of Spanish and Catalan, extracted mainly from two sources: two papers by Crego and Mariño (2006a and 2006b) and our own empirical analysis of two corpora: CoStEP and Tatoeba. The results suggest that the bilingual word embeddings that we are training our Long Short Term Memory model with do not improve any English word order learning by part of the model when used cross-lingually. There is no improvement when reordering the Spanish and Catalan texts so that their word order more closely resembles English, and no significant drop in result score even when applying a random reordering to them making them almost unintelligible, neither when classifying between 2 options (positive-negative) nor between 4 (strongly positive, positive, negative, strongly negative). We also replicated this with two different classifiers: a Convolutional Neural Network and a Support Vector Machine. The Convolutional Neural Network should primarily learn only short-range word order, while the Long Short Term Memory network should be expected to learn as well more long-range orderings. The Support Vector Machine does not take into account word order. Subsequently, we analyzed the prediction biases of these models to see how they affect the reordering results. Based on this analysis, we conclude that the lacking results of the Long Short Term Memory classifier when fed a reordered text do not respond to a problem of prediction bias. In the process of training our models, we use two bilingual lexicons (English-Spanish and English-Catalan) (Hu and Liu 2004) that contain words that typically are key for analyzing the sentiment of a sentence that we use to project our bilingual word embeddings between each language pair. Due to the results we got in the reordering experiments, we conjectured that what determines how our models are classifying the sentiment of the target languages is whether these lexicon words appear or not in the input sentence. Finally, because of this, we test different alterations on the target languages corpora to determine whether this conjecture is strengthened or not. The results seem to go in favor of it. Our main conclusion, therefore, is that Part of Speech-based word reordering of a target language to make its word order more similar to a source language does not improve the results on sentiment classification of our Long Short Term Memory classifier trained on source language data, regardless of the granularity of the sentiment, based on our bilingual word embeddings.

Contents

List of Tables

1 Introduction

1.1 Motivation

When facing a relatively simple text, such as a restaurant or hotel review, a tweet, or any product review, a general sentiment can typically be extracted. This sentiment can be considered as fundamentally positive or negative, and different levels of granularity can be determined (how many divisions should we have in a gradient between 'positive' an 'negative'). The automatic process in which a computer recognizes the overall sentiment of such texts is called Sentiment Analysis. These automatic processes can allow us to analyze large quantities of texts extracted from the Internet, for example. Depending on which topics these texts are about, we can get a view on how a section of the population might feel about a particular product, a touristic destination, or a public figure, which can be of interest for companies marketing their products, public institutions, or social studies, among many other areas or organizations. Nowadays, the sentiment analysis of languages with more computational resources (such as English) tends to produce better results than languages with fewer resources (such as Spanish or Catalan). Confronted with this problem, we have two main choices: the first option is to try to create resources for low-resource languages so that we have a comparable amount of tools and data with high-resource languages such as English, mainly, and therefore be able to obtain a similar level of results to Sentiment Analysis tasks performed in high-resource languages. The second choice is to try to use the data and tools already available and currently in use for high-resource languages to perform Sentiment Analysis tasks on low-resource languages. This, of course, requires some intermediate work so that resources on a language such as English can be used in a language such as Spanish. There is a variety of possibilities to do this. Cross-Lingual Sentiment Analysis (CLSA) is the study of the different ways in which we can use the data and tools of high-resource languages to improve the performance measure of Sentiment Analysis tasks in low-resource languages.

One of these possibilities is the following: when we want to analyze the sentiment of a hotel review in Catalan (we will call this the target language), for instance, we can simply translate it to the languages whose resources we are going to use. For example, we may have trained a model on English data that performs quite well when classifying the sentiment of English hotel reviews (we will call this the source language). We would translate the Catalan review to English, and simply use the English-trained sentiment classifier to analyze the translation to our original tweet. We can think of this approach as using Machine Translation (MT) as preprocessing for CLSA. Specifically, in this thesis we will use a different approach: instead of MT, we will use bilingual word embeddings between the source and target languages. We are interested in how much does the different word order between the target and source

languages affect the sentiment classification. Therefore, we want to study how the word reordering of the target languages that we will be working with, Spanish and Catalan, so that they more closely resemble the word order of our source language, English, affects the sentiment classification of an English-trained Model. While MT approaches typically require large amounts of parallel data, which may be difficult to acquire for low-resource languages, bilingual embeddings have been proved to be competitive in Sentiment Analysis while requiring fewer data (Abdalla and Hirst 2017, Barnes et al 2018b, Chen et al 2016). As we previously said, we will not use MT, but bilingual embeddings. A bilingual embedding consists of two monolingual embeddings that represent translation pairs: one in the source language, and one in the target language. For instance, ideally, if we were to compare a monolingual distribution of the triad 'food'-'dinner'-'orange' with 'comida'-'almuerzo'-'naranja', we would find strong similarities. What we do, then, is optimize between the two representations (the six embeddings) so that we remain each of the embeddings of these pairs being very similar to the other embedding, with only three pairs (one for each translation pair) that represent the respective distribution for both languages. For this we require a bilingual dictionary of key words between the source and target languages. We can rely on the larger amount of annotated data for our source language to analyze our target language. The main advantage of this strategy, then, is that we do not need as much parallel annotated data.

This work is strongly dependent on an ongoing PhD thesis by Jeremy Barnes (2017, 2018a, 2018b) as well as his constant input during its development. In the followings subsections 1.2 and 1.3 we explain the aims of this thesis and the approach we will follow, respectively. In section 2 we present the current state of the art of CLSA and word reordering. In section 3 we introduce the models, data, and tools we will be using throughout the thesis. In section 4 we extract and test the different sets of reordering rules on three different classifiers. In section 5 we perform an analysis of the different models used. Finally, in section 6 we present a summary of the obtained results.

1.2 Aims

We hypothesize that the Sentiment Analysis model we are using, a Long Short Term Memory (LSTM) network is, among other things, learning something about the word order of its source language. Because of this, we expect the following: if we compare the performance measure of our LSTM classifier when analyzing texts in Spanish or Catalan to the performance measure of this same sentiment classifier when analyzing texts in Spanish or Catalan whose word order has been reordered, the closer this new word order is to the English word order, the higher the performance measure of the analysis of the reordered Spanish and Catalan will be. Therefore we expect that any reordering that makes the word order of our target languages more similar to the

word order of our source language will improve performance, and that the closer this resemblance is, the higher the performance will be.

After the experiments reported in this thesis, we can say that our LSTM model does not take into account word order of the target language for sentiment analysis. Whether this is caused for the particular classifier we are using, or the bilingual embeddings is not entirely clear. Any type of reordering, even a random reordering of the target language did not seem to have much relevance (if any at all) in the sentiment analysis of the reordered test. For this reason we can say that our word embeddings and an LSTM network do not result in any learning on word order relevance for the sentiment analysis of the target languages. Further work should rely on learning whether this is a problem of the classifiers tested, or the embeddings used. The reordering itself does not seem to be a factor.

1.3 Approach

We decided to base our target language reordering on Part of Speech tags reordering. The main reason for this is the generalizing power POS tags. Intuitively, we can think of common reorderings like {NOUN, ADJ} and {ADJ, NOUN}. We used two main approaches for this. Firstly, we used a list of POS reorderings between Spanish and English developed by Crego and Mariño. Secondly, we analyzed two corpora: a parallel (English-Spanish) segment of CoStEP, a cleaned version of the Europarl Corpus, and also a parallel English-Spanish segment of Tatoeba, which contains a much more informal language than the Europarl corpus. Based on the comparison between Part of Speech (POS) tags combinations (bigram, trigram, tetragram, and pentagram) frequencies of this empirical analysis of English and Spanish parallel corpora, and using different criteria (section 4.3) we determine different sets of POS tags transformation rules. These sets of rules together with the list of rules extracted from Crego and Mariño should reorder our target languages so that they more closely resemble the word order of our source language.

As mentioned in section 1.2, we expected that the closer we could change the word order of our target languages to the word order of our source language, the better the sentiment classification would be on a source-language-trained LSTM network. Because of this we compare different sets of transformation rules, including two control sets, expecting that some of them will result in much better classification than others. We test a random reordering, a single transformation rule, the list of rules extracted from Crego and Mariño, a set of rules extracted from our own empirical analysis of the CoStEP corpus, and finally a set of rules extracted from our own empirical analysis of the Tatoeba corpus.

2 State of the art

2.1 Cross-lingual Sentiment Analysis

When we read a restaurant review, for instance, typically we can identify an author, an entity, be it a politician, a city, a monument, a football team, a restaurant, a product, etc. and a general attitude that the author has or wants to express towards this object. A short message such as 'This restaurant offers the best service I have ever experienced' has a clearly positive attitude, or sentiment, while if 'best' was instead 'worst' we would say it has a clear negative sentiment. In this case, we are classifying the sentiment of a text in a binary way, but we can also think of a 4-scale classification (strongly positive, positive, negative, strongly negative). Some complex or specialized texts such as novels or highly technical reviews of products may not necessarily have a clear overall attitude, but some short messages such as consumer-level product reviews, film reviews, tweets and so on, have a general positive or negative sentiment, or can be dissected in multiple parts that do (which can be classified according to different levels of granularity). Some texts might not have a discernible sentiment: in these cases we would label them as neutral.

Sentiment Analysis is the study and development of automatic classification procedures of such texts based on a previously decided attitude scale. One of the most important resources for this is labeled data. While unsupervised Sentiment Analysis exists (Turney 2002, Lin and He 2009), most state-of-the-art methods use large datasets of labeled corpora, and training classification models with this data (Al-Shabi et al 2017, Demirtas 2013). While there is no shortage of annotated corpora in English, other languages such as Catalan have very few of them in comparison. This means that low-resource languages cannot use the same classification methods as high-resource language and expect similar quality of results. For this reason, different methods of making use of the labeled data of high-resource languages for low-resource languages are used. This is called Cross-Lingual Sentiment Analysis. One of the most used methods for this is to perform Machine Translation of the target language text to the source language, and then feed the translated text to our classifier, that has been trained with a large quantity of labeled data from our source language. There are two main problems with this approach: firstly, it relies on the quality of the translation between the source and target languages, which, depending on the target language, might not be good enough, and therefore may harm our sentiment analysis of it. Secondly, and more importantly, it requires a much larger quantity of parallel annotated data than bilingual word embeddings. With bilingual word embeddings we can use machine translation of only some key words (and depending on the number of these key words, even manual translation is possible), and look at how well different monolingual embeddings of the source language capture the word semantics of the

target language, and select the embedding pairs that best transfer the semantic knowledge we have of our high-resource source language to our low-resource target language.

2.2 Word Reordering in Machine Translation (as Preprocessing)

Word Reordering is a complex area in MT as well as Contrastive linguistics (Pinedo 1997). Because of some of the problems we can find when changing the word order of the source language at the same time as making the rest of the translation process, sometimes it is considered as a best approach to firstly rearrange the word order of the language we are working with monolingually and with access to the full sentence before carrying out any other translation work on it. There are three main preprocessing strategies: deterministic, non-deterministic, and hybrid techniques. Deterministic preprocessing consists of finding the best possible reordering of the input sentence based on hard transformation rules. Non-deterministic strategies encode various alternative reorderings and rely on an n-gram based decoder to choose the best options depending on its training. Hybrid strategies consist on using hard rules exclusively on particularly difficult (typically long-range) word reorderings, and rely on non-deterministic reorderings on the less difficult reorderings (typically short-range) (Bisazza and Federico 2016). In this thesis, we will use a deterministic strategy.

We can also consider two main approaches to how the rules are determined: data-driven or manually (based on linguistic knowledge). We are not aware of any work in the scientific literature that studies both long and short-range word order differences between Spanish and English. In this thesis we will work with two classes of transformation rules: one of them is extracted from Crego and Mariño, and the other class, which consists of multiple sets of transformation rules, is our own work based on the empirical analysis of the two corpora we are working with. Both of these are POS-based, data-driven extracted from a previous statistical analysis of corpora. A purely data-based approach relies exclusively on a brute-force approach to reordering, which is computationally expensive. On large sentences, some constraints have to be introduced to control very long-range reordering. The main problem of this approach is that it does not use any of the linguistic data we may know regarding common reorderings between the source and target language. Crego and Mariño use a knowledge-informed data-driven approach to extract their transformation rules. Our own method is further explained in section 4.2. We are not aware of any work in the scientific literature studying the effects of word order in CLSA when using MT or bilingual word embeddings as preprocessing.

3 Models, Data, and Tools

3.1 Models

In this thesis we will be mainly working with an LSTM classifier. In order to compare the impact of word reordering as preprocessing for the analysis of the sentiment of the reordered text, we will also work with a Support Vector Machine and a Convolutional Neural Network. We describe these classifiers next.

Long Short Term Memory Network

A Recurrent Neural Network (RNN) is a type of Neural Network that specially takes into account sequential information. For example, when an RNN looks at a particular word of an input sentence, or any sequence of tokens, it will also take into consideration information gathered from previous words in the sentence. While theoretically an RNN does not have a strict limit on how long its input information can be for it to "remember" all the relevant information of the elements of the sequence, there is a practical limit of a handful of previous steps that an RNN can realistically take into account. An LSTM is a type of RNN particularly designed to improve the long-range consideration of previous elements of the information sequence it is treating that currently lead to state-of-the-art results in Sentiment Analysis (Barnes et al 2017, Howard and Ruder 2018, Tai et al 2015). An application for this can be, for instance, trying to predict the following word given an incomplete sentence: an LSTM will take into account the previous words of the sentence for it; which we would consider a reasonable approach.

Support Vector Machine

We will use a Support Vector Machine (SVM) to compare the results of our LSTM model in Section 5. We will use a bag-of-embeddings representation of our data, and because of this the SVM does not take into account any word order, so no reordering of a text will affect its sentiment classification. An SVM produces a strong baseline for Sentiment Analysis (Kiritchenko et al 2014).

Convolutional Neural Network

We will use a Convolutional Neural Network (CNN) to compare the results of our LSTM model in Section 5. A CNN can learn word orders, but it is much more restricted to local word order than an LSTM and it has a much more limited expressibility (Barnes et al 2017, Dos Santos and Gatti 2014, Severyn and Moschitti 2015).

Corpus	Sentences	Avg. Sent length	Avg. Word length
CoStEP (es)	4000	28.01	5.69
CoStEP (en)	4000	26.34	5.50
Tatoeba (es)	4000	9.71	5.08
Tatoeba (en)	4000	10.47	4.59
Multibooked (ca)	1149	14.25	5.02
OpeNER (es)	1472	16.39	5.33
OpeNER (en)	1731	13.85	5.12

Table 1: Corpora statistics

3.2 Corpora and Lexicons

In this thesis we will be using two sets of corpora for two different tasks. Firstly, for the empirical analysis we will use two parallel English-Spanish corpora: a section of CoStEP (a cleaned version of Europarl), and a section of Tatoeba (a web page in which native users can propose their preferred translations to given words). Secondly, for the training of the models and the sentiment analysis evaluation we will use a section of the MultiBooked corpus (Catalan) and the OpeNER Spanish and English corpora, all of them consisting on hotel reviews. We will also be using two bilingual lexicons English-Spanish and English-Catalan to train our classifiers. Table 1 shows the length of the sections of the corpora we will be using as well as relevant data to infer the complexity of the languages present in each corpus.

CoStEP - Europarl

The CoStEP corpus is a cleaned version of the Europarl corpus, organized and aligned based on speaker turns (Graën et al 2014).[1] Because of the lengthy POS annotation process, we will be working with a parallel section of 4000 sentences in English and Spanish. We will use it to determine the frequencies of the most common word reorderings between English and Spanish. After word and sentence tokenization, we use the Stanford tagger to annotate it for POS tags.

Tatoeba

Tatoeba is a web page in which translations of sentences are proposed by native speaker users, and typically presents a more informal language.[2] Because of the lengthy POS annotation process, we will be working too with a parallel section of 4000 sentences in English and Spanish. We will use it to determine the frequencies of the most common word reorderings between English and Spanish, and compare the differences between these and the ones extracted from the CoStEP corpus. After sentence and word tokenization, we use again the Stanford tagger to annotate it for POS tags.

[1]The corpus is available in https://pub.cl.uzh.ch/wiki/public/costep/start
[2]The corpus was scraped from https://tatoeba.org/eng/

Corpora	Train	Test	Dev.
OpeNER (es)	1029	296	147
OpeNER (en)	1210	347	174
MultiBooked (ca)	803	232	114

Table 2: Train, testing, and development sets

OpeNER

We will use a subset of the English and Spanish OpeNER corpora (Agerri et al 2013) that deal exclusively with hotel reviews. The corpora are annotated for POS tags and sentiment (strongly positive, positive, negative, strongly negative). We will use the English only for training and testing our model, and part of the Spanish to train and test our models, and part of it as one of our target language corpus to test how reordering affects the sentiment classification (Table 2).

MultiBooked

MultiBooked is an annotated corpus of Basque and Catalan Hotel reviews. We will be using the Catalan as one of our target language corpus to test how reordering affects the sentiment classification. (Barnes et al 2018a). The corpus is annotated for POS tags and sentiment (strongly positive, positive, negative, strongly negative) (Table 2).

Bilingual Lexicons

A projection lexicon can help us decide the sentiment of a sentence by looking at specific words that appear in it. Words such as 'richly' or 'amazing' appearing in a sentence can inform us with relatively high confidence about the overall sentiment of the sentence. We use a lexicon for each language pair (English-Spanish and English-Catalan) to determine common embeddings for these words based on the monolingual embeddings extracted from the data of the source and target languages. We will be using two bilingual English-Spanish and English-Catalan lexicons of 5700 words to determine how much is our LSTM, CNN, and SVM classifiers relying on these key words for sentiment classification (section 5.2) that perform better than general bilingual lexicons. We use GoogleNews vector for the English, and for Spanish and Catalan we use skip-gram embeddings using Word2Vec with 300 dimensions based on Wikipedia corpora (Barnes et al 2018b).

3.3 Tools

Stanford Tagger

We use the Stanford Tagger with Stanford-Postagger, and we use the English-left3words-distsim model for English tagging, and the Spanish-ud model for

Spanish tagging.[3]

NLTK

We use the NLTK library for sentence and word tokenizing and calculating the frequency distribution of corpora.[4]

4 Impact of word reorderings on an LSTM, CNN, and SVM

4.1 Motivation and Goals

The objective of this experiment is to find a set of transformation rules based on the most common POS reorderings between Spanish and English so that, when applied to our target languages (Spanish and Catalan), we would be able to reorder the text in a way that would make its word order more similar to English. We do not have parallel corpora between English and Catalan, but we use the data from the Spanish corpus to reorder the Catalan, under the assumption that the differences are not be very significant.

When evaluating the performance of our LSTM model (and in section 4.2 the other models as well) we focus on its F1 measure. The model is trained on the English data from the OpeNER corpus optimized by looking at the training Spanish data from the same corpus, and we reorder the complete Catalan and Spanish corpora from OpeNER and MultiBooked, and compare the sentiment analysis of this reordered texts to the analysis of the same non-reordered texts. We also compare the results of an LSTM with two different models, a CNN and an SVM.

We supply our models with: i) the original non-reordered target language and ii) various reorderings of it. We expect that the performance of this model improves when supplying it with ii). This is assuming that the reordering of the target language makes its word order more similar to the one of the source language. In the same way, if the reordering makes the word order of the target language even more different to the one of the source language than it is already, we expect the performance to drop.

4.2 Determining the Transformation Rules

We use two classes of transformation rules: the first class is extracted from Crego and Mariño (2006a, 2006b) and consists of only one set of transformation rules. The second class is extracted from our own empirical analysis of the CoStEP and Tatoeba corpora, and consists of multiple sets of transformation rules, with the initial objective

[3]Available here https://nlp.stanford.edu/software/tagger.shtml
[4]Available here https://www.nltk.org/

of testing each of them and ultimately retaining the one that produces the best scores. We add as Appendix 1 all the tagsets present in the annotated corpora.

Crego and Mariño

From Crego and Mariño (2006a, 2006b) we extract a list of 15 POS-based transformation rules. We will refer to this list as CREGO. As we mentioned earlier, their method for determining the rules is a knowledge-informed data-driven strategy, and the corpus used is Europarl. For this reason, some of the rules represent a highly formal and complex language not common in the hotel reviews corpora we are using to test the word reordering (OpeNER and MultiBooked). Appendix 2 shows the respective frequency of each Spanish POS combination in its n-gram list. Finally, we convert the POS tags of the set extracted to the tagset we work with. The tagsets are presented in Appendix 1.

Empirical Analysis

We decide to carry our own empirical analysis because of the low number (15) of rules present in the list by Crego and Mariño. Another reason is that their list is extracted from Europarl, which features a significantly more formal language than the sentiment-annotated corpora we will be testing (see Table 1 for the complexity difference between the corpora). If we look at Appendix 2 we can see that some of the Spanish POS tag combinations are not even present in the Tatoeba corpus, which bears a much more similar language to our testing corpora. Our empirical analysis will also show whether the frequencies of a formal and informal corpora result in different transformation rules. Firstly, we extract from the complete CoStEP corpus the Spanish and English parallel data. We perform tokenization at the sentence and word level of a segment of 4000 sentences of each, and annotate it. We convert the English tags to the Spanish tags. We then notice that the English text has a significant higher number of 'VERB' tags than the Spanish, which could affect negatively short-range POS combinations comparisons between the two. For this, we decide to simplify all consecutive sequences of 'VERB' tags to only one occurrence. Tables 3 and 4 show the frequencies and accumulated frequencies (Acc. Freq.) before and after the POS tag fusion.

We work as well with the Tatoeba corpus, whose language has a level of complexity and formality much more similar to the one from our hotel reviews corpora. We also work with a Spanish and English parallel corpus.

The objective, in both cases, is to use the statistical data that we get from these empirical analysis to find common POS tag combinations (from bigram to pentagram) in our target language (Spanish) whose order is changed in the source language (English). A predictable example of this would be the Spanish bigram

POS combinations	English		Spanish	
	Freq.	Acc. Freq	Freq.	Acc.Freq
'VERB'	17.003 %	17.003 %	9.012 %	9.012 %
('VERB', 'VERB')	4.861 %	21.864 %	0.475 %	9.487 %
('VERB', 'VERB', 'VERB')	1.278 %	23.142 %	0.001 %	9.488 %
('VERB', 'VERB', 'VERB', 'VERB')	0.239 %	23.381 %	0.0 %	9.488 %
('VERB', 'VERB', 'VERB', 'VERB', 'VERB')	0.041 %	23.422 %	0.0 %	9.488 %

Table 3: Frequencies of consecutive 'VERB' tag in CoStEP

POS Combination	English	Spanish
'VERB'	12.601 %	8.562 %

Table 4: 'VERB' tag frequencies after fusing the combinations

{NOUN, ADJ}, which in English would likely be {ADJ, NOUN}. In section 4.3 we explain in more detail the different ways in which we measure this commonality.

4.3 Sets of transformation rules

Having the data of POS frequencies between the Spanish and English texts of bigram, trigram, tetragram, and pentagram combinations, we decide on 3 criteria to extract the possible transformation rules from this data to add to the set of transformation rules extracted from Crego and Mariño.

Our intention is the following: each of the three criteria will automatically create transformation rules from the POS combinations from the corpora. Each of the criteria is more strict than the previous criterion, and it applies itself to the set of rules created by the previous criterion. Our objective is to keep applying increasingly stricter limits to our criterion and see when this criterion starts to improve results. The original expectation was that the first generous criterion would significantly reduce performance, since it makes the text, once the transformation rules are applied, almost impossible to understand. With stricter criteria, score was expected to gradually improve, until eventually the shrinking number of transformation rules would make the reordered text too close to the original, and score would drop again. This is clearly not the case, as we will see in section 4.5. A secondary expectation was that the rules produced by the empirical analysis of the Tatoeba corpus would result in a higher overall score, since the formality of the corpus and its linguistic complexity is more similar to our hotel reviews corpora than CoStEP. We cannot see any data that supports this hypothesis in the results either.

Criterion 1 (no criterion): We look through all the combinations of tags, from bigrams to pentagrams, ordered from more common to least common in our target

Spanish		English	
Bigram POS combs	Freqs	Bigram POS combs	Freqs
(‘DET’, ‘NOUN’)	10.10%	(‘DET’, ‘NOUN’)	6.22%
(‘NOUN’, ‘ADJ’)	2.96%	(‘ADJ’, ‘NOUN’)	4.73%
(‘VERB’, ‘DET’)	2.28%	(‘VERB’, ‘DET’)	3.33%

Table 5: Simplified comparison of frequencies for Criterion 1

language. We take each of them, and look at the set of all possible permutations (including the combination itself with no change). Then we look at the same list of combinations in the source language's respective list of POS combination frequencies, ordered from most common to least common, and we create the transformation rule in such a way that the first element of the rule is the Spanish tag combination, and the second element of it the English combination. If both are the same, we discard the rule. It is worth noting that in any n-gram combination where $n > 2$ and at least one of the tags is repeated, ambiguity occurs. Consider the possible rule:

$$\{\text{‘NOUN’, ‘NOUN’, ‘ADJ’}\} \rightarrow \{\text{‘ADJ’, ‘NOUN’, ‘NOUN’}\}.$$

Our original strategy was to assume a left to right reorder, and later a right to left and compare the best results. Since the initial results show clearly that reordering does not affect sentiment analysis, we do not carry through to deal with this ambiguity.

In order to illustrate more clearly the process stated by criterion 1, we present a simplified comparison of the highest frequencies of bigram POS combinations between English and Spanish in the CoStEP corpus in Table 5. Note that this is an idealization of the results to better explain the algorithm. The complete results (of the 10 most frequent combinations) are presented fully in Appendix 3.

Looking at Table 5 we see that first POS combination we would look up would be (‘DET’, ‘NOUN’). We would then look at its set of possible permutations: {(‘NOUN’, ‘DET’), (‘DET’, ‘NOUN’)}. We now would look at the combination that ranked higher in the English frequencies. In our example, this would be (‘DET’, ‘NOUN’). Therefore we would finally create the rule

$$\{\text{‘DET’, ‘NOUN’}\} \rightarrow \{\text{‘DET’, ‘NOUN’}\}.$$

In this case, the rule would be scraped (since it would not result in any change), and we would move on to the next Spanish POS combination. In this case it would be (‘NOUN’, ‘ADJ’). Once more, we would look at its set of possible permutations: {(‘NOUN’, ‘ADJ’), (‘ADJ’, ‘NOUN’)}. We now would look at the combination that ranked higher in the English frequencies. In this case it would be (‘ADJ’, ‘NOUN’), which would create the rule

$$\{\text{‘NOUN’, ‘ADJ’}\} \rightarrow \{\text{‘ADJ’, ‘NOUN’}\}.$$

This rule we would keep. We can also see that the following step would result in the rule

$$\{\text{'VERB'}, \text{'DET'}\} \rightarrow \{\text{'VERB'}, \text{'DET'}\}.$$

In this case, the rule would be once more scraped, and we would move on to the next combination.

Criterion 2: Building from each set of transformation rules we get from applying criterion 1 to the CoStEP corpus and the Tatoeba corpus, we delete from each set the rules that look at POS tag combinations that are too uncommon in their own frequencies. We take as a baseline an analysis of Crego and Mariño's list: the Spanish tag combination that is least common occurs at the top 48 percentile of its n-gram distribution. The higher combination of this is {NOUN, ADJ, CONJ, ADJ}, which is in the 1.35% of its n-gram frequency. The lowest is {ADJ, ADV}, in the 47.41%. Based on this, we eliminate all the transformation rules obtained following criterion 1 such that their target language combination is not found in the 47% of its respective n-gram list. Again, the original strategy consisted of repeatedly applying this criterion with an increasingly stricter limit (beginning as 48%, and tending towards $\sim$ 1%). Once more, the results showed futile continuing this approach.

Criterion 3: Building on the previous set of rules (the result of applying criterion 2 to the set of rules resulted of applying criterion 1 to each of the statistical data from the analysis of the CoStEP and Tatoeba corpora), we now eliminate the rules that connect a Spanish combination and an English combination that are too far apart in their respective frequencies. Once more, we take as a baseline an analysis of Crego and Mariño's list. In this case, we calculate in which top percentile both the Spanish and English combination of a rule are, and look at their difference (respectively to their n-gram). We find a minimum of 0.3 % and a maximum of 68.54 % (when compared to CoStEP). Therefore, as with the previous criterion, we take as a baseline maximum a difference of 68 %, and delete from our previous set all the rules that exceed this maximum. More intuitively, we want to make sure that our transformation rule is indeed capturing the same combination in each language: if we find a combination $\{\alpha,\ \beta\}$ in the 5% in Spanish and a combination $\{\beta,\ \alpha\}$ in the 65% in English, we should probably assume that it is just a coincidence, and not consider such a rule.

To sum up, we begin by creating a set of transformation rules in each of the corpora we are analyzing (CoStEP and Tatoeba) simply by comparing all the Spanish tag combinations of their n-gram to the English permutation with the

Criteria	2-gram	3-gram	4-gram	5-gram	Total
Cr. One	91	1312	7153	18845	27401
Cr. Two	42	648	3616	9598	13904
Cr. Three	15	246	1369	3565	5195

Table 6: Application of the rules criteria in CoStEP

Criteria	2-gram	3-gram	4-gram	5-gram	Total
Cr. One	77	727	2182	3621	6607
Cr. Two	34	366	1088	1724	3212
Cr. Three	15	148	436	655	1254

Table 7: Application of the rules criteria in Tatoeba

highest frequency, and delete from this baseline set the rules that i) rely on very infrequent tag combinations, and ii) show a significant difference between the respective frequencies of each combination, which would show that the connection between both combinations would probably be a coincidence. Tables 6 and 7 show the number of transformation rules created by every criterion on both corpora.

We also introduce two sets of transformation rules for comparison: a random reordering, and a set consisting of exclusively one transformation rule:

$$\{\text{NOUN, ADJ}\} \rightarrow \{\text{ADJ, NOUN}\}.$$

Therefore, we will compare the following sets of transformation rules (see Table 9 for their number of rules):

- **CREGO**: a set extracted from the articles by Crego and Mariño.
- **ONE**: a set consisting of one transformation rule.
- **EUR**: a set of rules resulted from the application of criterion 3 to the data from the CoStEP corpus.
- **TAT**: idem as 3) with the Tatoeba corpus.

Reordering	Example sentence
Original	“La relacio qualitat preu no es correspon .”
CREGO	“La relacio qualitat no preu es correspon .”
ONE	“La relacio qualitat preu no es correspon .”
EUR	“La preu relacio qualitat es no correspon .”
TAT	“La relacio qualitat preu es correspon no .”
RANDOM	“qualitat preu correspon es La relacio no .”

Table 8: Application of reorderings to an example sentence

Reordering	Rules
EUR	5195
TAT	1254
CREGO	15
ONE	1
RANDOM	-

Table 9: Final sets of rules

- **RANDOM**: a random reordering.

We also present in Table 8 an example Catalan sentence from Multibooked with all the different reorderings being applied.

4.4 Reordering

We will now apply CREGO, ONE, EUR, TAT and RANDOM to the Spanish OpeNER corpus and the Catalan MultiBooked corpus and obtain 5 different reordered corpora. We will feed this as input to our LSTM model, and compare the results of its analysis of them to the original non-reordered corpus. The way we apply the rules of each set is the following: we apply them in descending order, from higher n-gram to lower (5 to 2), and we only make one pass for each sentence. This means that it is possible that the application of one or several rules in the first pass on a particular sentence may have resulted in a combination of tags that would be again transformed if we were to apply our set of rules a second time. The original strategy was to apply indefinitely all rules of a set until it could not be applied anymore to the input sentence, but since our results showed that the reordering proved irrelevant, this was not pursued. We will obviate EUR for the reordering of the CNN and SVM because, as we see, its high number of rules makes it basically equivalent to RANDOM.

	LSTM Binary			LSTM Non-Binary		
Reordering	ES	CA	EN	ES	CA	EN
Original	0.695	0.622	0.775	0.366	0.316	0.713
EUR	0.668	0.641	-	0.369	0.305	-
TAT	0.688	0.623	-	0.368	0.310	-
CREGO	0.690	0.620	-	0.372	0.313	-
ONE	0.692	0.623	-	0.372	0.315	-
RANDOM	0.679	0.636	-	0.361	0.326	-

Table 10: LSTM results with different reorderings

	CNN Binary			CNN Non-Binary		
Reordering	ES	CA	EN	ES	CA	EN
Original	0.588	0.446	0.823	0.344	0.288	0.600
TAT	0.605	0.436	-	0.358	0.288	-
CREGO	0.593	0.444	-	0.353	0.283	-
ONE	0.592	0.447	-	0.362	0.280	-
RANDOM	0.606	0.479	-	0.351	0.304	-

Table 11: CNN results with different reorderings

	SVM Binary			SVM Non-Binary		
Reordering	ES	CA	EN	ES	CA	EN
Original	0.646	0.528	0.763	0.367	0.286	0.497

Table 12: SVM results

4.5 Results and Discussion

Tables 10 to 12 show the results from the reorderings with our three classifiers. We can see the results point to the following conclusion: word order does not seem to have a real effect on the sentiment analysis of our model. The difference between the different sets of transformation rules are non-existent in some cases, and in others the slight difference can be seen as the result of the small sample size of our corpora. We can conclude that our bilingual embeddings do not seem to allow our LSTM network to learn word order insofar as sentiment analysis for the target language. We see that even with a random reordering, the LSTM scores significantly higher than an SVM or any ordering of the CNN. The fact that RANDOM scores higher than any other ordering with a CNN is interesting, but we see no other explanation for this other than variation because of the small sample size of our corpora.

5 Model Analysis

5.1 Error Analysis

Based on the results the experiment described in section 4.4, we consider interesting to examine how biased the three models are towards positive sentiment predictions

	POS	NEG
Gold	83%	17%
LSTM (preds)	83%	17%
SVM (preds)	70%	30%
CNN (preds)	58%	42%

Table 13: Comparison of prediction biases of classifiers

(Table 13 shows the overall biases of the three classifiers as compared with the gold). For convenience, we will take 70% of our Spanish OpeNER corpus (1029 sentences): 851 sentences are positive (the sum of strong positive and positive), and 178 negative sentences (strong negative and negative): 83% of the corpus is positive, and 17% negative.

If we look at the predictions of our LSTM model on the latter corpus, we see the following: the model makes 849 positive predictions to 180 negative predictions of the Spanish segment, making 83% of its predictions positive and the remainder 17% negative. The SVM produces 716 positive predictions and 313 negative predictions, making the positive predictions amount to 70% of the total, and 30% of the predictions being negative. Finally, the CNN makes 597 positive predictions and 432 negative predictions, making the 58% of the predictions positive and 42% negative.

Taking into consideration the bias towards positive predictions of all of our models, we can see that the LSTM learns well the proportion of positives and negatives sentiments of the corpora. Because of this, we can say that low scoring of the LSTM with the different reorderings is not due to any bias issue. The SVM, however, learns worse this proportion than the LSTM. Finally, we can see that the CNN performs rather poorly on this regard, with the lowest score of the three classifiers.

5.2 Impact of Lexicon Words

Considering our results from the experiment described in section 4.4, we hypothesize that our LSTM model is not learning anything related to word order. We consider instead that it is mainly deciding the sentiment of a sentence predominantly based on the occurrence of certain words (or lack thereof) present in two bilingual lexicons (English-Spanish and English-Catalan) that we are using for our bilingual embeddings. Our expectation is that clear changes to our corpus based on targeting the words that appear in the lexicon should substantially change the overall score.

For convenience, for this additional experiment we only work with a section of our overall corpus: we work with 70 % of our Spanish corpus (1029 sentences). We take this segment and evaluate its analysis according to the following criteria:

	LSTM		CNN		SVM	
	Bi	Non-Bi	Bi	Non-Bi	Bi	Non-Bi
Original	0.628	0.313	0.588	0.359	0.591	0.367
RANDOM	0.626	0.299	0.583	0.338	0.591	0.367
ONLYLEX	0.624	0.324	0.393	0.217	0.424	0.308
NOLEX	0.597	0.292	0.550	0.340	0.573	0.326

Table 14: Classifiers results on a lexically-modified corpus

Lexical Modification	Example sentence
Original	"Muy amable el personal y perfecto el desayuno buffet con mucha variedad ."
ONLYLEX	"amable perfecto variedad ."
NOLEX	"Muy el personal y el desayuno buffet con mucha ."
RANDOM	"con perfecto buffet mucha el variedad personal amable Muy . el desayuno y"

Table 15: Application of lexical modifications to an example sentence

- **Original**: The original sentence.
- **ONLYLEX**: The segment with all the words that are not present in the lexicon deleted.
- **NOLEX**: The segment with all words present in the lexicon deleted.
- **RANDOM**: The segment with a random reordering of all words.

We also present in Table 15 an example Spanish sentence from OPeNER with all the criteria being applied.

Once more, we consider the F-1 measure of the evaluation, and compare the binary and non-binary classes. The results shown in Table 14 seem to reinforce our conjecture: the drop in score in the LSTM between the original corpus and the corpus with all words not present in the lexicon deleted is almost non-existent, despite the fact that we are reducing most input sentences to one or two words (the ones in the lexicon that appear in the given sentence). Likewise, according to our hypothesis, we see a significant drop when deleting the lexicon words from the text, and almost no difference when maintaining all the original words in an input sentence but reordering them randomly. As far as the CNN, we can see an interesting result: the main drop in score seems to be when we delete most of the words of the corpus, and only leave the key words that appear in the lexicon. This could mean that a CNN might be more susceptible to word order than an LSTM. But we see a similar result with an SVM, which does not take into account word order at all, so more data would be required for strengthening this possibility. We can also realize that the LSTM is by far the most stable of the three.

6 Conclusions

6.1 Summary of Results

Our main conclusion is that no change in reordering seems to affect in any significant way the sentiment analysis of our LSTM model, even when introducing a random reordering which renders the input sentences almost unintelligible. From this we can conclude that our LSTM is not learning much related to word order. It remains to be studied whether other classifiers would respond better to our bilingual embeddings, or other embeddings would make an LSTM classifier learn more about the source language word order. With the results of section 5.2 we can reasonably add to our main conclusion that our LSTM classifier is deciding the sentiment of a sentence based almost exclusively on either the appearance or non-appearance of some key words, present in a 5700 bilingual lexicon used to train it, while ignoring other words in the sentence and their order. We can also see that a CNN seems to be taking into account features in an input sentence other than lexicon words, but if we look at Table 10, we can see that a random reordering improves score as compared to the original, so it is doubtful that word order is being taken into account by a CNN. Finally, based on the comparison with other models, we can say that the LSTM's learning of the bias towards positive sentiment predictions is not affecting in any major way the overall results.

References

[1] Abdalla, M.; Hirst, G. 2017. Cross-Lingual Sentiment AnalysisWithout (Good) Translation. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*: 506-515.

[2] Al-Shabi, A.;Adel, A.; Omar, N.; Al-Moslmi, T. 2017. Cross-Lingual Sentiment Classification from English to Arabic using Machine Translation. In *International Journal of Advanced Computer Science and Applications (IJACSA)* 8 (12): 434-440.

[3] Agerri, R.; Cuadros, M.; Gaines, S.; Rigau, G. 2013. OpeNER: Open polarity enhanced named entity recognition. In *Sociedad Española para el Procesamiento del Lenguaje Natural* 51: 215–218.

[4] Barnes, J.; Klinger, R.; Schulte im Walde, S. 2017. Assessing state-of-the-art sentiment classifiers on state-of-the-art sentiment datasets. In *Proceedings of 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2017)*.

[5] Barnes, J.; Lambert, P.; Badia, T. 2018a. Multibooked: A corpus of basque and catalan hotel reviews annotated for aspect-level sentiment classification. In *Proceedings of 11th Language Resources and Evaluation Conference (LREC 2018)*.

[6] Barnes, J.; Klinger, R.; Schulte im Walde, S. 2018b. Bilingual sentiment embeddings: Joint projection of sentiment across languages. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2018)*: 2483-2493.

[7] Bisazza, A; Federico, M. 2016. A Survey of Word Reordering in Statistical Machine Translation: Computational Models and Language Phenomena. In *Computational Linguistics* 42(2): 163-205.

[8] Chen, X.; Sun, Y.; Athiwaratkun, B.; Cardie, C.; Weinberger, K. 2016. Adversarial Deep Averaging Networks for Cross-Lingual Sentiment Classification. In *CORR*.

[9] Crego, J.M.; Mariño, J.B. 2006a. Improving Statistical MT by coupling reordering and decoding. In *Mach. Translat.* 20: 199-215.

[10] Crego, J.M.; Mariño, J.B. 2006b. Integration of POStag-based source reordering into SMT decoding by an extended search graph. In *Proceedings*

of the 7th conference of the Association for Machine Translation in the America, Visions for the future of machine translation (AMTA): 29-36.

[11] Demirtas, E. 2013. Cross-Lingual Sentiment Analysis with Machine Translation. Eindhoven University of Technology. Masters Thesis.

[12] Dos Santos, C.N.; Gatti, M. 2014. Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*: 69-78.

[13] Graën, J.; Batinic, D.; Volk, M. 2014. Cleaning the Europarl corpus for linguistic applications. In *Konvens 2014. Stiftung Universität Hildesheim.*

[14] Howard, J.; Ruder, S. 2018. Universal language model fine-tunning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*: 328-339.

[15] Hu, M.; Liu, B. 2004. Mining opinion features in customer reviews. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*: 168–177.

[16] Kiritchenko, S.; Zhu, X.; Cherry, C.; Mohammad, S.M. 2014. Canada-2014: Detecting Aspects and Sentiment in Customer Reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*: 437-442.

[17] Lin, C.; He, Y. 2009. Joint Sentiment/Topic Model for Sentiment Analysis. In *CIKM 2009*: 375-384.

[18] Pinedo, A. 1997. Translating Spanish Verb-Subject Order into English: Strategies for maximising discourse-pragmatic equivalence. Lancaster University. PhD Thesis.

[19] Severyn, A.; Moschitti, A. 2015. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).*

[20] Tai, S.K.; Socher, R.; Manning, C. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*: 1556-1566.

[21] Turney, P.D. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*: 417-424.

Appendix

POS Tagsets

English Tagset	Spanish Tagset	CREGO Tagset
" ’ ’ "	’ADJ’	’AQ’
’ , ’	’ADP’	’CC’
’ . ’	’ADV’	’NC’
’ : ’	’AUX’	’PP’
’CC’	’CONJ’	’RG’
’CD’	’DET’	’RN’
’LS’	’INTJ’	’VA’
’DT’	’NOUN’	’VM’
’PDT’	’NUM’	’VS’
’PRP$’	’PART’	
’EX’	’PRON’	
’MD’	’PROPN’	
’FW’	’PUNCT’	
’IN’	’SCONJ’	
’JJ’	’VERB’	
’JJR’	’X’	
’JJS’		
’NN’		
’NNS’		
’NNP’		
’NNPS’		
’POS’		
’PRP’		
’RB’		
’RBR’		
’RBS’		
’RP’		
’TO’		
’VB’		
’VBD’		
’VBG’		
’VBN’		
’VBP’		
’VBZ’		
’UH’		
’WDT’		
’WP’		
’WRB’		

Table 16: Tagsets used for annotating English (38), Spanish (16), and used by Crego and Mariño (9)

CREGO Frequencies on CoStEP and Tatoeba

POS Combination	Freq.	Accum. Freq.
('NOUN', 'ADJ')	2.96 %	2.96 %
('NOUN', 'ADV')	0.534 %	3.494 %
('ADV', 'VERB')	0.522 %	4.016 %
('VERB', 'PRON')	0.209 %	4.225 %
('NOUN', 'ADV', 'ADJ')	0.199 %	4.424 %
('ADJ', 'ADJ')	0.178 %	4.602 %
('NOUN', 'ADJ', 'ADJ')	0.138 %	4.74 %
('ADJ', 'ADV')	0.107 %	4.847 %
('NOUN', 'ADJ', 'CONJ', 'ADJ')	0.094 %	4.941 %
('NOUN', 'CONJ', 'NOUN', 'ADJ')	0.031 %	4.972 %
('NOUN', 'ADV', 'ADJ', 'CONJ')	0.021 %	4.993 %
('NOUN', 'ADV', 'ADV')	0.021 %	5.014 %
('ADJ', 'ADV', 'ADJ')	0.013 %	5.027 %
('NOUN', 'ADJ', 'ADV', 'ADJ')	0.01 %	5.037 %
('NOUN', 'ADV', 'ADJ', 'CONJ', 'ADJ')	0.006 %	5.043 %

Table 17: CREGO Spanish POS combinations frequencies in CoStEP

POS Combinations	Freq.	Accum. Freq.
('ADV', 'VERB')	1.628 %	1.628 %
('NOUN', 'ADJ')	1.19 %	2.818 %
('VERB', 'PRON')	0.732 %	3.55 %
('NOUN', 'ADV')	0.695 %	4.245 %
('ADJ', 'ADV')	0.163 %	4.408 %
('NOUN', 'ADV', 'ADJ')	0.141 %	4.549 %
('ADJ', 'ADV', 'ADJ')	0.036 %	4.585 %
('ADJ', 'ADJ')	0.028 %	4.613 %
('NOUN', 'ADJ', 'ADV', 'ADJ')	0.02 %	4.633 %
('NOUN', 'ADJ', 'ADJ')	0.017 %	4.65 %
('NOUN', 'ADV', 'ADV')	0.015 %	4.665 %
('NOUN', 'ADJ', 'CONJ', 'ADJ')	0.014 %	4.679 %
('NOUN', 'CONJ', 'NOUN', 'ADJ')	0.006 %	4.685 %
('NOUN', 'ADV', 'ADJ', 'CONJ')	0.0 %	4.685 %
('NOUN', 'ADV', 'ADJ', 'CONJ', 'ADJ')	0.0 %	4.685 %

Table 18: CREGO Spanish POS combinations frequencies in Tatoeba

Extract of POS frequencies CoStEP

Spanish			English		
2-gram	Freq	Accum Freq	2-gram	Freq	Accum Freq
(DET, NOUN)	10.10%	10.10%	(DET, NOUN)	6.22%	6.22%
(ADP, DET)	7.26%	17.36%	(ADP, DET)	6.02%	12.24%
(NOUN, ADP)	6.47%	23.83%	(NOUN, ADP)	5.94%	18.18%
(NOUN, PUNCT)	3.79%	27.62%	(ADJ, NOUN)	4.73%	22.91%
(ADP, NOUN)	3.72%	31.35%	(NOUN, PUNCT)	4.18%	27.09%
(PUNCT, $)	3.44%	34.79%	(PUNCT, $)	3.65%	30.73%
(NOUN, ADJ)	2.96%	37.75%	(VERB, DET)	3.33%	34.06%
(VERB, DET)	2.28%	40.03%	(NOUN, VERB)	3.01%	37.07%
(VERB, ADP)	2.24%	42.27%	(DET, ADJ)	2.93%	40.00%
(PROPN, PUNCT)	2.24%	44.50%	(PROPN, PROPN)	2.53%	42.53%

Table 19: Ten most frequent 2-gram POS combinations in Spanish and English (CoStEP)