Multiple Evaluations of a Bigram Model

Daniel J Casas

Introduction

I am going to present multiple methods for accuracy evaluation of a bigram model that outputs actual words in Palenquero creole (Colombia).

A bigram model was trained using a corpus of 139000 words in Palenquero, which is about 70% of the total corpus of 198834 words. The remaining 59834 words were used for evaluation.

To evaluate the performance of this model, I propose the following three methods: cross-entropy loss, entity matching, and novelty detection. These three approaches offer a well-rounded mix between inferential and descriptive statistics to measure accuracy.

I will also discuss their advantages and disadvantages, as well as discussing the model's overall quality at predicting words in Palenquero.

Cross-Entropy Loss

Cross-entropy loss is a useful method to evaluate whether the model's predictions are closer to the ground truth labels. Lower values are preferred, since they indicate better performance. Looking at table 1 we can see that the values for training, validation, and test are very similar, almost identical. This is an indication that the model is not over-fitting, thus suggesting that the model is able to generalise to unseen data.

Training Loss	1.5836
Validation Loss	1.6060
Test Loss	1.6038

Table 1: Cross-Entropy Loss of the Palenquero Bigram Model

However, these values over one indicate that there is still room for improvement and that the loss might be smaller if the correct fine-tuning is applied to the hyperparametres of the model, such as batches and layers.

Eventhough cross-entropy loss is good at generalising "correctly for continuous probability densities" (Shore and Johnson, 1980), it might be sensitive to outliers and perform worse when the training set is not too large.

Entity Matching

"Entity Matching is the task of deciding if two entity descriptions refer to the same real-world entity" (Peeters and Bizer, 2023). With this evaluation method, I wanted to show how many words in the output where actual words in Palenquero. In contrast to evaluating the model with an inferential method, like cross-entropy loss, entity matching is a good close second alternative to human evaluation and it allows me to offer descriptive insights. This is an effective way to evaluate the quality of the model's output.

I asked the model to output 10, 100, 1000, 10000, and 100000 words—almost as many as the model was trained on—and I compared those results to a set of words from a test set. The model had not seen this corpus before, so I assumed it would be a good fit for the task. The goal, then, was seeing if the number of matches was dependent on the amount of output.

As Figure 1 shows, accuracy goes down as the output increases. This might suggest issues with the model struggling with the increased complexity of outputting more words, thus performing the same task over and over again. It could also could be an issue with scalability—simply the training data set does not provide much to work further. That can also be combined with data sparsity and noise coming from irrelevant matches.

	output	matches	accuracy_%
0	10	7	70.00
1	100	66	66.00
2	1000	651	65.10
3	10000	6319	63.19
4	100000	63140	63.14

Figure 1: Cross-Entropy Loss

Novelty Detection

"Novelty detection is the identification of new or unknown data or signals that a machine learning system is not aware of during training" (Miljković, 2010). With this method, my aim was to identify which words in the output where words that the model had not seen during training. This method is particularly good when new data has patterns that are rare (Miljković, 2010) and it gave me the chance to see how far the model was able to go in terms of being able to create "novel" vocabulary in Palenquero.

Similarly to what I did in entity matching, I asked the model to output 10, 100, 1000, 10000, and 100000 words. Then, I checked if there were actual words in Palenquero by matching to the test set and then I checked which of those words where already present in the training set. My goal was to find real words absent in training.

It is remarkable that the amount of matches for unseen words in training is far less than for those found in the previous method. There is a middle point in which the amount of matches and the accuracy seem to stabilise, between the 100 and 1000 words of output, but in general it follows the same downward tendency I saw in the results of entity matching.

This is an indication of the high reliance the model has on the training set and its difficulty to depart from it. This could be corrected by fine-tuning the model, thus obtaining even a lower cross-entropy loss, thus increasing the entity matches and, by extension, the novelty detection would be higher.

	output	matches	accuracy_%
0	10	0	0.000
1	100	1	1.000
2	1000	10	1.000
3	10000	86	0.860
4	100000	837	0.837

Figure 2: Novelty Detection Summary

Conclusion

As seen in the section about cross-entropy loss, the model has still has room to reach lower values, i.e., higher accuracy. This fine-tuning will in turn have a trickle down effect, improving the output, making it more Palenquero-like, and that can be tested by the two other methods I used.

As seen in each section, both inferential—cross-entropy loss—and descriptive approaches—entity matching and novelty detection—offer a complementary view on the models performance and it allows for a much more insightful look at the results. They account for the mathematical and statistical work the model is operating in the background and for the end result users will see. This is particularly effective at compensating each others drawbacks.

It is also very encouraging to see a language model trained in a class-room setting perform relatively well. These small numbers mean a lot to a community that is fighting to preserve and revitalise its ancestral language. It is definitely a motivation to gather more data and improve the models.

It would also be interesting exploring a human evaluation of the output that is not an "actual" word in the language and seeing if it could be used as part of the revitalisation efforts. They could be raw materials for a new lexicon in Palenquero.

Links to code

https://github.com/DanielCasas21/palenque/tree/main/modelo_bigram_pln

References

Miljković, D. (2010). Review of novelty detection methods, 593–598.

Peeters, R., & Bizer, C. (2023). Using chatgpt for entity matching [Accepted and to be published in Proceedings of ADBIS 2023 as short paper]. *arXiv preprint arXiv:2305.03423*, Article arXiv:2305.03423. https://doi.org/10.48550/arXiv.2305.03423

Shore, J. E., & Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, 26–37.

Additional Charts

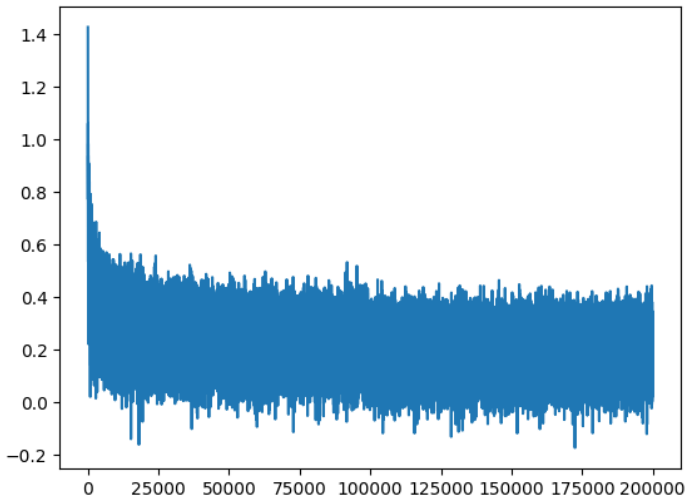

Figure 3: Model's Cross-Entropy Loss

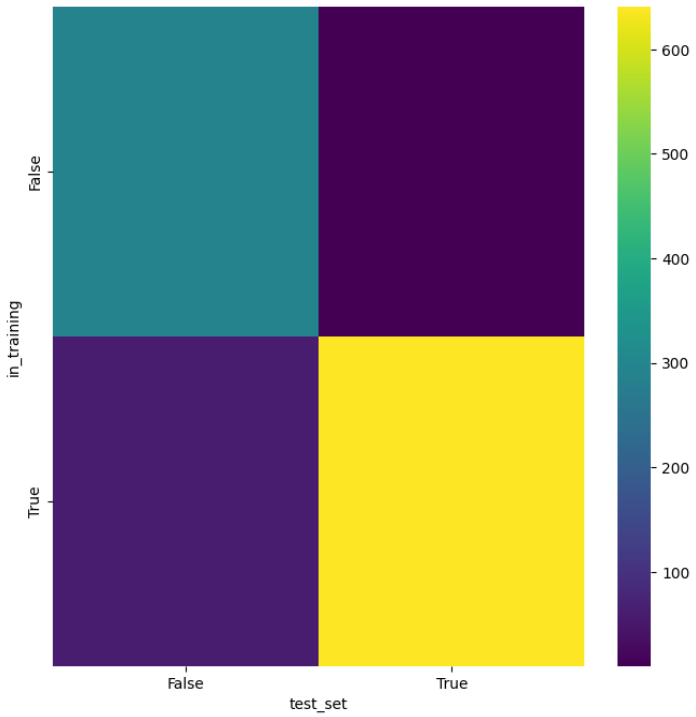

Figure 4: Heat map contrasting output matches to word in the training and test sets