A Study on Brazilian Portuguese Verbal Irregularities via Artificial Neural Networks

Beatriz Albiero
University of Sao Paulo
beatriz.albiero@usp.br

Abstract

This project is an attempt at reproducing Rumelhart and McClelland's [7] connectionist experiment described in the book Parallel Distributed Processing, chapter "On learning the past tense of English verbs", with Brazilian Portuguese as a target language. In this book, Rumelhart and McClelland describe a new theory of cognition called connectionism that is challenging the idea of symbolic computation that has traditionally been at the center of the debate in theoretical discussions about the mind. In the chapter "On learning the past tense of English verbs", Rumelhart and McClelland describe an experiment in which a feedforward neural network was developed in order to find patterns among phonological features between present and past tense forms of English verbs. In this research, an identical network has been built to predict Brazilian Portuguese irregularities. We will discuss our model's results and the difficulties in adapting this model in order to predict irregularities in Brazilian Portuguese.

Introduction

The process of verbal inflection from present to past tense in the English language is certainly one of the most controversial topics of debate among the main theoretical currents in linguistic study [3, 5, 7]. At the heart of the debate is the exact characterization of the mechanisms that enable a speaker to relate a verb in present tense to its past tense form.

The past tense of English is made up of a set of inflection families, which are split into two macro families: regular and irregular. Each macro family features subsets, grouped together by inflective similarity, with a noteworthy variety of families within the irregular group. This peculiar quality has motivated researchers D. Rumelhart and J. McClelland to build an artificial neural network model capable of predicting said inflective processes. While it is also possible to observe inflective regularities within irregular verb families in Brazilian Portuguese, the question remains: is it possible to adapt Rumelhart and McClelland's model to such language?

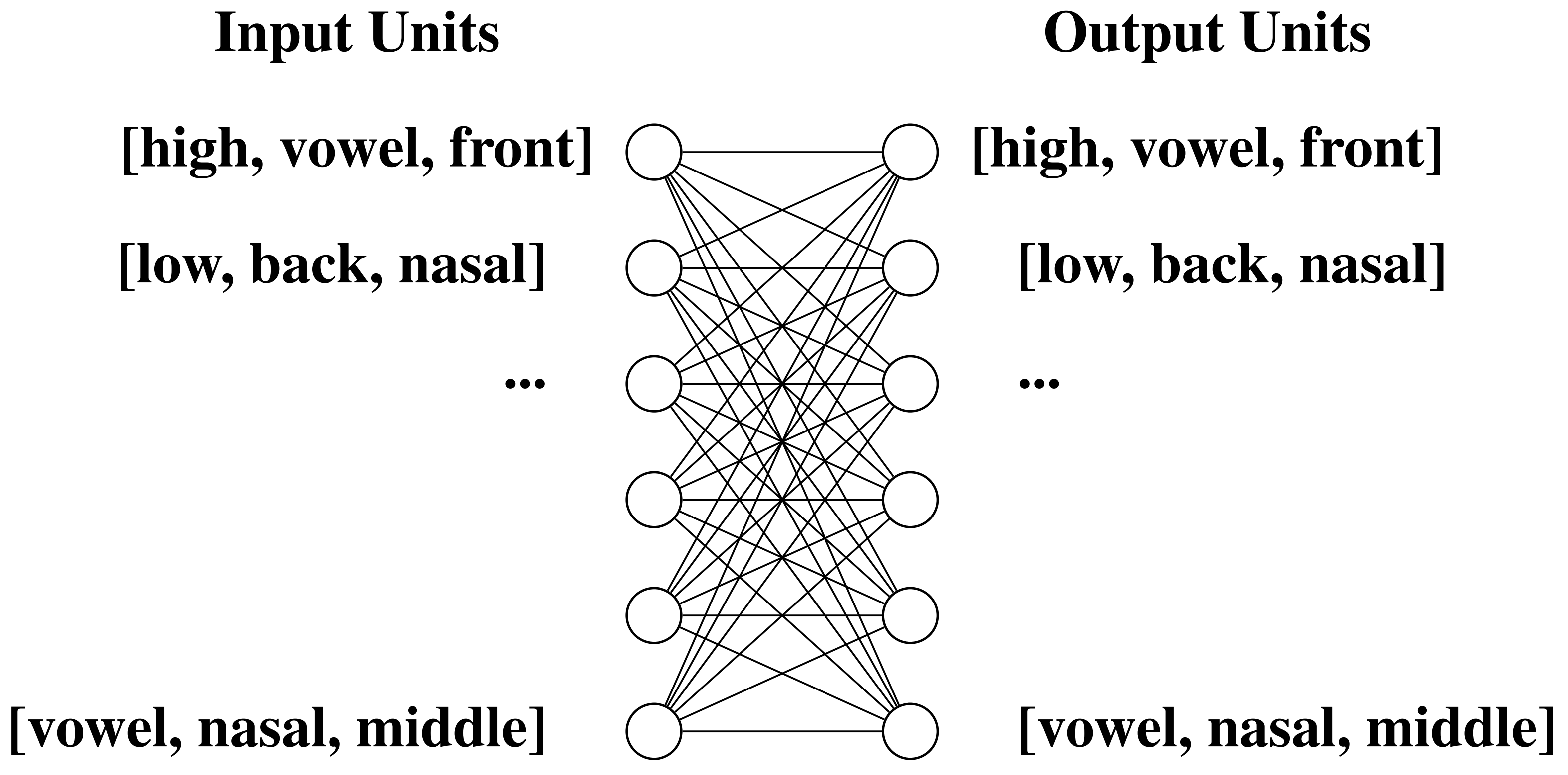
Main Objectives

- The development of predictive computational models (artificial neural networks) capable of detecting inflective phonological patterns within irregular verb families in Brazilian Portuguese.
- Analysis of the model's results.

Methodology

A corpus of four hundred and three verbs, both in the infinitive and in the first person singular of the indicative, was built via web-scraping techniques. The material was then translated into a pseudo-phonological writing system, which was further codified into Wickelfeatures. A Feed Forward Neural Network was then built with the help of Keras library [2]. The network, similar to the one presented by the authors, is basically made up of two simple layers. The first layer is responsible for receiving the input data, which is a codified representation of the phonological features, the sounds, of a verb in the infinitive form (**Wickelfeatures**)[7]. The second layer outputs identically codified phonological features, however, conjugated in the first person singular of the present tense. The first person singular was chosen by virtue of its moderate range of inflective variations. The results were then arranged into tables for further analysis.

Architecture



Input Units

Output Units

[high, vowel, front]

[low, back, nasal]

...

[vowel, nasal, middle]

[high, vowel, front]

[low, back, nasal]

...

[vowel, nasal, middle]

Figure 1: Neural Artificial Network Scheme

Corpus

403 verbs, divided into two parts: test (∼20%) and training (∼80%)

	Regulars	Irregulars	Total	%
Test	42	36	78	19.35
Train	172	153	325	80.65
Total	214	189	403	100

Table 1: Regular-irregular ratios in datasets

Results

A small set of 21 verbs was used to test the accuracy of the model. Since all verbs in the infinitive form end in the phoneme r, the deletion of this phoneme reduced redundancies and facilitated the decoding of wickelfeatures. In order to test the relevance of the irregular verb ratio for the models accuracy, five datasets, each with a different regular-irregular verb ratio, were tested, with the best infinitive to first person singular prediction accuracy being 47.62% for two sets: one made up of 55% irregular verbs and other made up of 95% irregular verbs.

Verbs exceeding four phonemes did not yield satisfactory Wickelfeature to Pseudo-Phonological-Writing decoding results.

Ratio	Epochs	Batch Size	Accuracy
55%	400	464	47.62%
65%	400	565	38.10%
75%	400	388	38.10%
85%	400	390	42.86%
95%	400	395	47.62%

Table 2: Accuracy results for different ratios of irregular verbs.

Verb	Input	Expected	Output
pegar	#pega#	#pEgu#	#pEku#
cegar	#sega#	#sEgu#	#sigu#
secar	#seka#	#sEku#	#siku#
levar	#leva#	#lEvu#	#levu#
orar	#ora#	#Oru#	#earu#
morar	#mora#	#mOru#	#mEru#
postar	#posta#	#pOstu#	#pOtu#*
mentir	#menti#	#mintu#	#mitu#*
tossir	#tosi#	#tusu#	#tutu#
fazer	#faze#	#fasu#	#fasu#
matar	#mata#	#matu#	#matu#
pagar	#paga#	#pagu#	#pagu#
#sair#	#sai#	#saiu#	#saiu#

Table 3: Some outputs of the model

Conclusions

- Changing the ratio of irregular verbs in the training set apparently did not cause significant changes in accuracy, as shown in Table 2. It happened probably because the training dataset was made up of a small variety of verbs and enlarging its irregular verb classes simply resulted in the strengthening of connections for the Wickelfeatures that it already knew, culminating in a poor generalization capacity.
- The process of decoding Wickelfeatures into the pseudo-phonological writing system failed to yield satisfactory outcomes due to excessive binding redundancy within the Feed Forward Neural Network. The inadequacy of such model called for the implementation of a new strategy, which led to further research into Recurrent Neural Networks, particularly, the Encoder-Decoder technique.

Forthcoming Research

- Encoder-Decoder Model implementation with the purpose of better decoding Wickelfeatures.
- Once the decoding of Wickelfeatures is properly contrived, a psycho-linguistic test will take place, in order to measure native speaker intuition against the Neural Network output.

References

[1] J. L. Bybee and C. L. Moder. Morphological classes as natural categories. language. *Language*, 59(2):251–270, June 1983.

[2] François Chollet et al. Keras. `https://keras.io`, 2015.

[3] N. Chomsky and M. Halle. *The sound pattern of English*. MIT Press, 1968/1991.

[4] et al. Marcus, G. *Overregularization in language acquisition. Monographs of the Society for Research in Child Development*. 1992.

[5] S. Pinker and A. Prince. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(2):73–193, March 1988.

[6] T. Rashid. *Make Your Own Neural Network: A Gentle Journey Through the Mathematics of Neural Networks, and Making Your Own Using the Python Computer Language*. 2016.

[7] D. E. Rumelhart and J. L McClelland. *On learning the past tenses of English verbs*. Bradford Books/MIT Press, 1986.

Acknowledgements

I would like to express my very great appreciation to Marcelo Barra, Livy Real, Romulo Ferreira, the DataLab team, NAACL Emerging Regions Fund and the Linguistics Graduation Department (FFLCH-USP).