A Study on Brazilian Portuguese Verbal Irregularities via Artificial Neural Networks

Beatriz Albiero[1]

[1] University of Sao Paulo, Sao Paulo, BR
beatriz.albiero@usp.br

Abstract. This project is an attempt at reproducing Rumelhart and McClelland's (1986) connectionist experiment described in the book Paralell 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 debate in theoretical discussions about the mind. The authors' theory assumes the mind is composed of a great number of elementary units connected in a neural network. Mental processes are interactions between these units which excite and inhibit each other in parallel rather than sequential operations. In this context, knowledge can no longer be thought of as stored in localized structures; instead, it consists of the connections between pairs of units that are distributed throughout the network. 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.

Keywords: COMPUTATIONAL LINGUISTICS, LANGUAGE ACQUISITION, NEURAL NETWORKS.

1 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 (Rumelhart & McClelland 1986, Pinker & Prince 1988, Pinker 1999). 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 composed of a variety of families, occurring not only the distinction between regular and irregular verbs but also the formation of groups within the set of irregular, composed by verbs that share the same process of flexion:

blow – blew, grow – grew, know – knew, throw – threw

bear – bore, swear – swore, tear – tore, wear – wore
drink – drank, shrink – shrank, sink – sank, stink – stank

It is possible to think that the learning of the pertinence of a verb to one or the other family would come from a case-by-case memorization. However, experiments showed that when test subjects were presented made-up verbs the tendency for allocating verbs in certain classes was observed, for example, for the artificial verb *splining*, most people opted for the splang or splung form (Bybee & Moder, 1983). This example contradicts the idea that speakers could only be reproducing memorized forms and suggests that they are actively identifying patterns, plus: they have a natural intuition about the appropriateness of the allocation of the verb in one class or another. Rumelhart and McClelland (1986) presented a computational model of empiricist character that was fundamental for the emergence of a new school within the cognitive sciences: connectionism. The model has been developed by analogy to the structure in which the neurons are related in the brain, and therefore, it has received the name of *artificial neural network*. It is basically composed of an artificial network of nodes connected in parallel (**Fig.1**). The first layer of nodes is responsible for receiving the input data, which is the data referring to the phonological features that characterize the sounds of a verb in the present tense. The second layer is a response layer (output) that ought to try return data referring to the traits that characterize the sounds of the same verb provided in the input, but in the past tense. After the completion of this step, the output data obtained should then be compared to the correct form of the verb in the past tense, a kind of feedback. The function of the connections between the layers is to strengthen (or weaken) the relationships between the input and output layers according to the comparisons made between the output layer and the template. It is important to mention that, a priori, the network does not have any type of information for its operation, learning will take place over multiple iterations.

Rumelhart and McClelland's model presented excellent results in the task of predicting the verbal forms expected for the past simple, being able to identify associations correctly among all the 420 verbs in which it was trained. In addition, it performed satisfactorily when submitted to 86 new verbs that were not part of the training, obtaining a 92% success rate for irregular verbs and 84% for irregular verbs (91% accuracy for all verbs in total). The model served, therefore, to corroborate the argument

that it is possible to accomplish this task efficiently without the use of explicit rules.

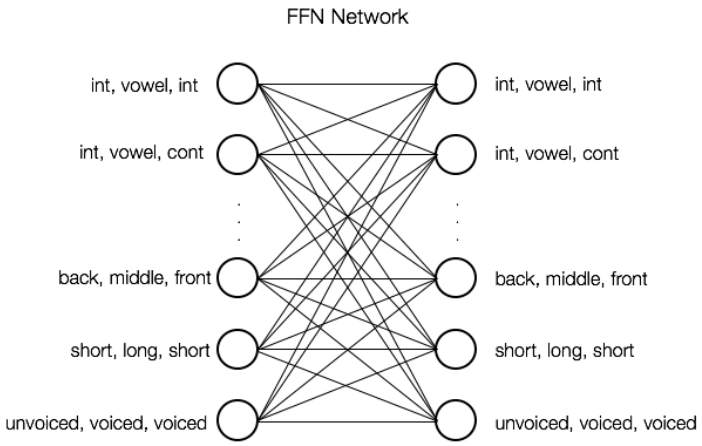

Fig. 1. Neural Artificial Network Scheme

2 Wickelfeatures

As described in **Section 1.1**, the input and output units correspond to the phonological features that characterize the sounds of a verb in the present tense. Rumelhart and McClelland's have proposed the characterization of each phoneme as a combination of features in 4 simple dimensions. The phoneme 'd', for example, can be characterized by a set of 4 features (considering 4 different dimensions in a simplified phonetic table described by the authors): Interrupted, Stop, Middle, Voiced.

Table 1. was based on Rumelhart and McClelland's coding scheme and it has been adapted to account for Brazilian Portuguese phonemes. The first dimension of the table divides the phonemes into three major types: interrupted consonants, continuous consonants and vowels. The second dimension further subdivides these major classes into: stop and nasals, fricatives and liquids, high and low. The third dimension classifies the phonemes into three rough places of articulation: front, middle, and back. The fourth subcategorizes the consonants into voiced vs. voiceless categories and subcategorizes the vowels into long and short. The authors also needed a representation scheme that allowed them to keep track of the sequence of the verbs sounds. For this reason, each activation unit should somehow still preserve some notion of sequence.

Therefore, their approach was to stablish each unit as a trigram of features, that means that they first divided each verb into trigrams of phonemes and then, for each phoneme, combined their phonological features maintaining the order.

Table 1. Categorization of phonemes on four simple dimensions.
(Adapted to the Portuguese language)

		front		**middle**		**back**	
		v/ Long	**u/ Short**	**v/ Long**	**u/ Short**	**v/ Long**	**u/ Short**
Interrupted	**Stop**	b	p	d	t	g	k
	Nasal	m		n			
Continuous	**Fricative**	v	f	z	s	j	x
	Liquid	l		r			h
Vowel	**high**	e	i			o	u
	low		E		a		O

*u = unvoiced, v = voiced.

They chose to name this particular type of input units as Wickelfeatures. This representation activates multiple units for each trigram and also allows the model to find patterns among features of sounds. One last feature had to be added to the coding scheme in order to assure that units would be translated back into trigrams of phonemes (and finally back into a verb) after the training: a special symbol (#) representing the beginning and the ending of a word.

3 Modeling in Neural Networks for Brazilian Portuguese Verbs

The same network scheme and flowchart presented by the authors were used for the construction of this project. However, the input verbs are verbs in the infinitive form and the expected output verbs are verbs conjugated to the first person singular (in the present tense of the indicative).

3.1 The Corpus

The Corpus is composed of a total of 403 verbs. It was then divided up into two parts: the test (containing approximately 20% of the corpus) and training (containing approximately 80% of the corpus) datasets.[1] **Table**

[1] More information about the corpus can be found here: https://github.com/beatrizalbiero/MsResearch/tree/master/WickelfeaturesProject/Corpus

2. features the number and ratio of regular-irregular classes in the datasets.

Table 2. Presence of Regular and Irregular verbs in the Corpus

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

3.2 First Results

A small set of 21 verbs was used to test the accuracy of the model. Since all verbs in the infinitive form end with the phoneme "r", the deletion of this phoneme reduced redundancies and facilitated the decoding of the wickelfeatures. In order to test the relevance of the irregular verbs ratio in the model's power of prediction, different training datasets were used to train the model, with each dataset containing different irregular-regular classes ratios.

Table 3. Accuracy results for different ratios of irregular verbs.

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%

Changing the ratio of irregular verbs in the training set apparently did not cause significant changes in accuracy. **Table 8.** features some predictions of the model.

Table 8. Some predictions of the model.

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#

4 Conclusion

Although the accuracy rates shown in **Table 7.** are far from the rates obtained by Rumelhart and McClelland on the same task, **Table 8.** Features some interesting results. Whereas the model failed to predict most irregular conjugations, it correctly predicted all regular verbs present in this test set and mistakenly regularized one ("levar"). It also correctly predicted the irregular families of the verbs "postar" and "mentir", despite the fact that it missed the central phoneme*.

References

1. Bybee, J. L., & Moder, C. L. (1983). Morphological classes as natural categories. Language, 59, 251–270.
2. Chomsky, N., & Halle, M. (1968/1991). The sound pattern of English. Cambridge, MA: MIT Press.
3. Marcus, G., et al. (1992). Overregularization in language acquisition. Monographs of the Society for Research in Child Development , 57, i+iii+v+vi+1-178.
4. Pinker, S., & Prince, A. (1988). On language and connectionism. Cognition 28 (1-2):73-193.
5. Rashid, T. (2016). Make Your Own Neural Network: A Gentle Journey Through the Mathematics of Neural Networks, and Making Your Own Using the Python Computer Language. CreateSpace Independent Publishing Platform.
6. Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group, Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 2. Psychological and biological models. Cambridge, MA: Bradford Books/MIT Press.