Revisiting Rumelhart and McClelland (1986) for Romanian Verb Conjugation

Victor Albu
albuvictor2016@gmail.com

Ciprian Cristescu
ciprian.cristescu@s.unibuc.ro

Abstract

This project reproduces the classic experiment of Rumelhart and McClelland (1986) on learning past tense verbs, adapted for the Romanian language. Unlike the original study, we employ a modern Bidirectional LSTM architecture with character-level embeddings. We demonstrate that the model can generalize to unseen verbs (including invented ones) while correctly retaining irregular forms, effectively solving the "dual-route" mechanism debate using a single neural network.

1 Introduction

The core problem addressed in this project is the acquisition of grammatical rules by neural networks. In 1986, Rumelhart and McClelland (1) demonstrated that a neural network could learn English past tense rules (adding "-ed") without explicit programming. We aim to investigate if this holds true for Romanian, a Romance language with complex conjugation patterns involving suffixes (e.g., *-at*, *-it*) and significant irregularities (e.g., *a fi* $\rightarrow$ *fost*).

Our main contributions are:

- Adaptation of the 1986 experiment for Romanian morphology.
- Implementation of a **Synthetic Data Generator** to overcome data scarcity.
- Comparison of generalization capabilities using a modern Bidirectional LSTM architecture.

We chose this project to understand the transition from rote memorization to rule generalization in deep learning models.

2 Approach

We implemented a character-level Sequence-to-Sequence model using Python, Keras, and TensorFlow. The code and dataset are available in our repository: https://github.com/Victor1e/Project_Rumelhart_Romanian

2.1 Data Generation

Since a large corpus of Romanian verb pairs is not readily available for this specific task, we created a **Synthetic Data Generator**. This algorithm creates pseudo-random verb roots adhering to Romanian phonotactics and generates their regular past tense forms (e.g., *boganiza* $\rightarrow$ *boganizat*). We combined 800 synthetic verbs with real irregular verbs to create a balanced training corpus.

2.2 Model Architecture

Instead of a simple MLP, we used a Recurrent Neural Network (RNN) specifically designed for sequence processing:

1. **Embedding Layer:** Maps input characters to a 64-dimensional dense vector space.
2. **Bidirectional LSTM:** Two LSTM layers (128 units) process the sequence in both directions to capture morphological context.
3. **Output:** A TimeDistributed Dense layer with Softmax activation predicts the character probability distribution.

Training took approximately 5 minutes on Google Colab GPU, running for 100 epochs with Early Stopping.

3 Preliminary Results

The LSTM model achieved $> 99\%$ accuracy on the validation set. Crucially, it correctly conjugated the invented test verb "boganiza" to "boganizat" (demonstrating rule acquisition) and "fi" to "fost" (demonstrating memory of exceptions). This confirms the "U-shaped learning" hypothesis where the model eventually masters both rules and exceptions.

4 Limitations

The primary limitation is the use of synthetic data. While effective for learning abstract rules, the generated verbs do not reflect the semantic distribution of real-world language. Additionally, the model operates strictly at the character level and does not account for sentence-level context.

5 Conclusions and Future Work

We successfully reproduced the connectionist experiment for Romanian. The Bidirectional LSTM proved superior to traditional feed-forward networks in handling variable-length sequences. Future work could involve training a Transformer-based model (like a small GPT) to compare performance on morphological tasks.

References

[1] David E. Rumelhart and James L. McClelland. 1986. *On learning the past tenses of English verbs*. MIT Press.