Orthographic Normalization of Historical Texts

Pryce Bevan, Luke Gessler, and Michael Kranzlein

Georgetown University
Department of Computer Science

{pwb8, lg876, mmk119}@georgetown.edu

Abstract

We explore several model architectures for token-based historical text normalization. At the token level, we observe that a naïve baseline imposes an upper bound on model performance on tokens seen in training data. Accordingly, we implement an LSTM model and a transformer model to predict unseen tokens and incorporate the baseline's predictions for seen tokens. We test these hybrid models on historical Spanish texts and achieve results that are on par with state-of-the-art models. We also note that TODO outperforms TODO.

1 Introduction

In many cases, researchers require or prefer to work with standardized data. In this work, we examine historical text normalization—"the problem of translating historical documents written in the absence of modern spelling conventions and making them amenable to search by today's scholars, processable by natural language processing models, and readable to laypeople" (Bollmann et al., 2018). This is a standardization problem that...

1.1 Motivating the Normalization Problem

2 Related Work

Historical text normalization has received renewed attention as neural approaches have gained traction in the NLP community. Over the past ten years, the state of the art has seen an evolution from Hidden Markov Models and rule-based models to simple neural networks to more context-aware recurrent neural networks. Most recently, there is a shift toward attention and transformers.

2.1 HMM and Rule-Based Methods

2.2 Neural Networks

2.3 Recurrent Neural Networks

2.4 Attention and Transformers

3 Model

3.1 Dataset

For this research, we make use of the Post Scriptum corpus, a resource built in 2014 to promote work in the digital humanities (Vaamonde et al., 2014). The corpus contains epistolary texts in Spanish and Portuguese, ranging from the 16th century to the 19th century. These documents are unpublished and represent a diverse set of social backgrounds. For our experiments, we use only the Spanish texts, totaling 2368 documents. Each original document is accompanied by a corresponding manually modernized document. Our models ingest an original token as input and predict the best modern equivalent.

3.1.1 Example

por me hazer md me ebye el bonete q conpre aqui

Figure 1: Test

Por me hacer merced, me envíe el bonete que compré aquí.

Figure 2: Modernized

3.2 Naïve Baseline

3.3 LSTM Model

3.4 Transformer Model

4 Results

Char	Char	Char	Word	Word	Word
1	1	1	1	1	1

5 Analysis

6 Conclusion

Acknowledgments

References

Marcel Bollmann, Natalia Korchagina, and Anders Søgaard. 2019. Few-Shot and Zero-Shot Learning for Historical Text Normalization. page 10.

Marcel Bollmann, Anders Søgaard, and Joachim Bingel. 2018. Multi-task learning for historical text normalization: Size matters. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 19–24, Melbourne, Australia. Association for Computational Linguistics.

Gael Vaamonde, Luisa Costa, Rita Marquilhas, Clara Pinto, and Fernanda Pratas. 2014. Post Scriptum: Archivo Digital de Escritura Cotidiana. *Janus: Estudos sobre el Siglo de Oro*, pages 473–482.