Integrating Vision Transformers and Large Language Models for Enhanced Handwriting Text Recognition and Named Entity Identification in Documentary Manuscripts

This study presents an innovative approach for Handwriting Text Recognition (HTR) and Named Entity Recognition (NER) by integrating Vision Transformer (ViT) and Large Language Models (LLMs), focusing on documentary manuscripts from the medieval and early modern periods, spanning from the late 11th to the late 16th centuries. Our aim is to construct a comprehensive end-to-end system specifically designed for the accurate transcription and understanding of historical texts, such as registers, charters, and manuscript series. This system employs an encoder-decoder architecture that synergizes Image and Language Transformer models in an autoregressive sequence, a significant leap from traditional methodologies.

The foundation of our research is an extensive annotated corpus, featuring nearly 3 million tokens and 250,000 graphical text lines derived from 52 diverse manuscripts in Latin, French, and Spanish. Notably, a substantial segment of this corpus —around 1 million tokens— is enriched with named entity annotations essential for NER tasks.

Transitioning from the earlier Convolutional Neural Network Recurrent (CNNR) models to the more sophisticated transformer architectures, we harness the advanced capabilities of pre-trained LLMs and Image Transformers. This shift, incorporating the Attention mechanism, represents a pivotal change, enhancing model development efficiency. Our findings demonstrate that the transformer-based encoder-decoder strategy not only surpasses CNNR models in HTR performance on CER, WER and BERT-score metrics but also adeptly manages NER and post-correction simultaneously. This dual-task capacity significantly diminishes the reliance on extensive ground-truth data for fine-tuning tasks, streamlining the processing of historical manuscripts.

Moreover, we explore the challenges faced during training and validation, particularly the creation of domain-specific language models, the implications of a causal word and subword prediction approach, and the biases introduced by adopting an Attention mechanism. We also address the challenge of transcribing texts with considerable graphical and orthographical variability using a fixed vocabulary and generative methods prone to yielding unauthentic transcriptions.

We show that the application of a hybrid training strategy, leveraging both supervised learning on annotated data and unsupervised learning on larger, unannotated datasets, further enhances the model's ability to generalize across different document types and time periods. Besides, this approach allows for the effective handling of both continuous script and discrete entity recognition within the same framework, significantly enhancing the model's versatility and applicability to a broad range of historical documents.