NLP Project Research Papers, their methodology, problems faced, dataset, results and evaluation metrics

1. **Ancient Script Image Recognition and Processing: A Review** [Link]
   Brief summary: -
   The authors categorize these writing systems into **logographic scripts** (Focus on semantic encoding through intricate symbols representing), which use complex symbols to represent meaning, and **phonographic scripts** (Represent phonemes or sounds through simpler, often connected characters (e.g., Ancient Greek, Sanskrit, Old Latin)**,** which use simpler characters to represent sounds. Digital analysis of these artifacts is often hindered by **extreme data scarcity** and significant **physical degradation**, such as erosion, cracks, and fading. To overcome these obstacles, the study examines advanced **deep learning techniques**, including generative adversarial networks for **data augmentation** and structural decomposition for **zero-shot recognition**.

   Datasets:-
   • **OBC306:** A large-scale dataset for Oracle Bone Inscriptions with 309,551 samples.
   • **HeiCuBeDa:** High-resolution 3D models and images of Mesopotamian Cuneiform.
   • **CATMuS Medieval:** A multilingual large-scale dataset for Latin scripts.

• **HKR:** A database for handwritten Slavic (Kazakh and Russian) scripts.

Results:- Specific models have achieved impressive results, such as **92.3%** for Oracle Bone Inscription recognition, **93.3%** for Bronze inscriptions, and up to **98.6%** for printed Sanskrit.

Evaluation Metrics

The effectiveness of the reviewed methods is measured using several metrics:

• **Recognition Performance:** Accuracy, Precision, Recall, and F-score.

• **Transcription Accuracy: Character Error Rate (CER)** and **Word Error Rate (WER)**, particularly for phonographic scripts.

• **Image Quality: Peak Signal-to-Noise Ratio (PSNR)** and **Structural Similarity Index Measure (SSIM)** for denoising and restoration tasks.

• **Retrieval/Detection:** mean Average Precision (mAP).

2. **Deep Learning Meets Egyptology: a Hieroglyphic Transformer for Translating Ancient Egyptian** [Link]

   This paper presents a novel approach to the automatic translation of ancient Egyptian hieroglyphs into modern languages. This paper introduced something like "**Hieroglyphic Transformer**"

   Methodology:- They have used **transfer learning** by fine-tuning a pre-trained multilingual model on a meticulously curated dataset. To address the scarcity of English data, they used **backtranslation**, where the model translated German entries into

English to augment the training set (I am unsure why they did it but it gave the best accuracy was for English and German). The process involved extensive **data cleaning** and the organization of data into specific source-target pairs, such as Hieroglyphs-to-German or Transliteration-to-POS tags.

Problems Faced

• **Data Scarcity and Accessibility:** Most ancient Egyptian data is **non-machine-readable**, existing as images or physical scans that require significant digitization.

• **Research Gap:** Previous AI efforts in Egyptology focused primarily on **Optical Character Recognition (OCR)** rather than linguistic translation.

• **Low-Resource Constraints:** There is a significant imbalance in available translations, with far more data in German than in English.

Dataset

The dataset was constructed from a snapshot of the Thesaurus Linguae Aegyptiae (TLA).

• Content: It focuses on Middle and Old Egyptian. After filtering and cleaning, it contained 61,605 data points.

• Structure: Each data point includes Gardiner codes (alphanumeric identifiers for signs), transliterations, German or English translations, Lemma IDs, and Part-of-Speech (POS) tags

Models Used

The primary model is the **Hieroglyphic Transformer**, built upon the **M2M-100** multilingual framework, which was originally designed to translate between 100 modern languages.

Evaluation Metrics

• **Automatic Metrics:** The researchers used **SacreBLEU** and **RougeL** to measure translation quality.

• **Statistical Validation:** They performed **10-fold cross-validation** to ensure the model's reliability and to mitigate selection bias.

• **Human Evaluation:** A qualitative assessment of 16 examples was conducted to check for **morphological accuracy, grammatical correctness, and semantic coherence**.

3. **(Research Article4.)Describing archival photographs using multimodal LLMs: a case study on evaluating vision-language model performance for creating descriptive metadata** [Link]

Purpose

This paper aims to examine the efforts of librarians in Northeastern University Library's Digital Production Services department to employ Gemini, a pre-trained multimodal large language model, for generating descriptive metadata for an archival photographic collection.

Design/methodology/approach

The project comprised three phases: 1) researching and selecting a multimodal LLM that was both accurate and cost-effective for generating descriptive metadata for photographs; 2) developing an application to batch-submit photographs to the chosen model's API and convert the results into a human-readable spreadsheet and 3) evaluating the output for completeness, accuracy, consistency and potential bias.

4. **Vision-Based Large Language Models for Vietnamese Handwriting Recognition** [Link]
   In short, this paper provides a comparative analysis between OCR-specific models and Vision-LLMs. The results indicate that while OCR-specialized models excel in accuracy and resource efficiency within narrow domains, Vision-LLMs provide greater adaptability across tasks - albeit with higher resource demands and sensitivity to script details.

5. **Few shots are all you need: A progressive learning approach for low resource handwritten text recognition** [Link]
   This paper proposes a method to transcribe manuscripts with rare or unknown alphabets while significantly reducing the need for manual human annotation.

   Methodology
   The authors treat handwritten text recognition as a **symbol detection and matching task** rather than a traditional classification problem. The methodology follows these steps:
   • **Pre-training:** The model is first trained on synthetic line images created from the **Omniglot dataset** to learn generic matching capabilities across different alphabets.
   • **Progressive Pseudo-Labeling:** To adapt to a specific target script, the user provides a "few shots" (ideally **5 samples per symbol type**). These shots are used to:
      1. **Generate synthetic lines** by randomly concatenating the shots with added noise and artifacts to mimic real handwriting.
      2. **Iteratively label** real unlabeled text lines. The model identifies symbols with high confidence scores, assigns them "pseudo-labels," and incorporates them into the training set for the next iteration.

• **Matching & Decoding:** A shared backbone extracts features from a text line (query) and a symbol (support). These are passed through an **Attention Region Proposal Network (RPN)** to generate similarity scores, which are then organized into a **similarity matrix** and decoded into the final text sequence using a modified **CTC-based algorithm**.

Problems Faced

• **Data Scarcity:** Ancient or enciphered manuscripts often have no existing labeled datasets or dictionaries, making deep learning difficult.
• **Manual Effort:** Traditional supervised HTR requires annotating thousands of characters with bounding boxes, which is extremely time-consuming.
• **Technical Challenges:** Cursive scripts and touching characters make **segmentation** (cutting lines into individual characters) prone to errors.
• **Domain Gap:** There is a significant performance drop when moving from synthetic training data to real, weathered historical manuscripts.

Datasets

The model was evaluated on three low-resource manuscripts:
• **Borg Cipher:** A 17th-century manuscript with 34 symbols; challenging due to **heavily connected and touching characters**.
• **Copiale Cipher:** An 18th-century manuscript with 100 symbols (Latin, Greek, and graphic signs); challenging due to the **large alphabet size**.
• **Codex Runicus:** A historical Nordic law manuscript featuring **runes**; the symbols are relatively well-segmented.

Models Used

• **Shared Backbone: VGG 16** for feature extraction from query and support images.
• **Attention Mechanism:** Performs **depth-wise cross-correlation** between the support and query feature maps within the RPN.

Evaluation Metrics
• **Symbol Error Rate (SER):** The primary metric, calculated as (substitutions, deletions, and insertions over total ground-truth length).
• **Intersection over Union (IoU):** Used to verify the accuracy of the predicted bounding boxes for pseudo-labels (a detection was considered correct if **IoU 0.7**).
• **Annotation Time:** Measured in minutes to quantify the reduction in human labor.

6. **Deep Learning in Archiving Indus Script and Motif Information**

[Link]

This paper focuses on applying modern computational techniques to one of the world's most famous undeciphered scripts, aiming to automate the recognition of its complex signs.

Dataset

The study primarily utilizes digitized versions of the major Indus corpora:

• Mahadevan Corpus (1977): A foundational digital concordance of the Indus script.

• Parpola Corpus (1994): Another major collection used to cross-reference sign forms.

Models Used

• **Convolutional Neural Networks (CNNs):** These are the core models used for image recognition.

Evaluation Metrics

• **Accuracy:** The primary metric for sign classification performance.

• **Top-k Accuracy:** Used to account for the visual similarity between different signs (checking if the correct sign is in the model's top-3 or top-5 predictions) .

• **Confusion Matrix:** Utilized to visualize and analyze the model's performance across different sign classes .

7. **A Hybrid Capsule Network-based Deep Learning Framework for Deciphering Ancient Scripts with Scarce Annotations: A Case Study on Phoenician Epigraphy**
[Link]
This paper introduces a novel framework designed to automate the recognition of ancient scripts, specifically focusing on the 22 letters of the Phoenician alphabet. The research aims to save epigraphists' time by providing a reliable tool for digitizing and deciphering historical inscriptions.

Methodology
The authors proposed a systematic multi-stage pipeline:
1. Preprocessing & Denoising: Initial noise filtering and binarization are followed by a convolutional autoencoder that reconstructs character images to remove noise.

2. Data Augmentation: To overcome data scarcity, a Variational Autoencoder (VAE) is used to generate new, human-like handwritten character samples by learning a latent probability distribution.
3. Classification: The augmented data is fed into a custom capsule network to classify each character into one of the 22 Phoenician alphabets.

Problems Faced

• **Data Scarcity:** Deep learning typically requires massive datasets, but ancient scripts suffer from a lack of annotated samples.
• **Complex Characteristics:** The Phoenician script uses *Scriptio continua* (no spaces or marks between words) and lacks vowels, making manual and automated deciphering difficult.
• **Inherent Variability:** Scripts varied significantly over time (12th century B.C. to 5th century A.D.) and across different geographic regions and materials.

Dataset

• The researchers created a unique **corpus of labeled Phoenician alphabets**.
• The dataset covers **all 22 characters** and includes a wide range of writing styles and historical stages.
• It includes diverse samples found across the Mediterranean region to ensure the model can handle various epigraphic styles.

Models Used

• **Hybrid Autoencoder Architecture:** Combines a **Convolutional Autoencoder** for denoising and a **Variational Autoencoder (VAE)** for generating augmented, human-like data.
• **Capsule Network:** Unlike standard CNNs, this network preserves **visuospatial features** and hierarchical relationships (spatial information) using **dynamic routing**. This model was chosen because it requires less training data and can handle overlapping characters common in inscriptions.

Evaluation Metrics

• **Accuracy:** Used to measure the overall recognition performance.
• **Loss:** Calculated using the **Euclidean distance loss function** within the capsule network's decoder to determine similarity between reconstructed and actual features.