Enhancing Information Extraction from Scanned Sleep Study Reports

A Comprehensive Evaluation of Workflow Components

By Kayvan Shah

Introduction

This study systematically evaluates the workflow for extracting information from scanned sleep study reports.

Context:

- Scanned documents in Electronic Health Records (EHR) pose unique challenges.
- Importance of accurate information extraction for clinical decision-making and analysis.

Significance:

- Fills a crucial gap in EHR-scanned document processing.
- Addresses challenges in handling diverse formats, images, and handwritten elements in sleep study reports.

Workflow Components

Image Preprocessing:

- Dilation/Erosion: Morphological operations to enhance or diminish regions of interest, improving text recognition.
- Contrast Enhancement: Techniques to increase the visual difference between text and background, aiding OCR accuracy.

OCR (Optical Character Recognition):

- Tool: Utilized Tesseract, a widely-used OCR engine, in conjunction with preprocessing methods.
- Grayscale: Conversion to grayscale, a common practice to simplify image data while retaining essential features.

NLP (Natural Language Processing):

- Bag-of-Words Models: Evaluated 7 machine learning-based models relying on the frequency of words in isolation, including Logistic Regression, SVM, kNN, NaiveBayes, and Random Forest.
- Deep Learning-Based Sequence Models: Explored 3 Transformer-based models - BiLSTM, BERT, and ClinicalBERT, capable of capturing contextual information for sequence data.

The workflow incorporates seamlessly connected processes, where enhanced images undergo OCR, and NLP models operate on the recognized text for information extraction.

Key Findings

Optimized Extraction:

- Dilation/Erosion and Contrast Enhancement: These image preprocessing techniques were crucial in optimizing OCR accuracy, enhancing the extraction of relevant information from scanned documents.
- Word-Layout Information: Incorporating word positions and layout details significantly boosted model performance, indicating the importance of document structure in information extraction.
- Deep Learning Impact: Transformer-based NLP models (BERT, ClinicalBERT) outperformed traditional bag-of-words models, showcasing the efficacy of advanced deep learning techniques in scanned document processing.

Novel Contributions:

- Comprehensive Evaluation: This study pioneers a systematic evaluation of key components in the information extraction workflow, including image preprocessing, OCR, and NLP model selection.
- Document Layout Utilization: Unprecedented exploration of leveraging document layout features as structured inputs, revealing a new direction for optimizing NLP model performance in scanned documents.

Improvements:

- Increased Accuracy: Adoption of advanced techniques, particularly deep learning-based NLP models, demonstrated notable improvements in accuracy, validating their efficacy in handling complex information extraction tasks from scanned documents.

Implications & Future Directions

Broader Relevance:

- EHR Applications: Insights from the study are not limited to sleep study reports, offering valuable applications in processing diverse scanned documents within Electronic Health Records (EHR).

Potential:

- Enhanced Efficiency: Optimizing workflows using advanced techniques can significantly improve information extraction, potentially revolutionizing data utilization in healthcare settings.

Future Research:

- Table Processing: Addressing challenges in natural language processing within tables, a common feature in medical reports, is crucial for advancing the capabilities of information extraction.
- Diverse Documents: Expanding the study to include various document types will provide a more holistic understanding, contributing to the development of versatile information extraction systems.

Interdisciplinary Collaboration:

- Collaborative Efforts: Encouraging collaboration between informaticians, healthcare professionals, and data scientists will facilitate the development of robust, real-world applications for scanned document information extraction in healthcare.

What is the purpose of image preprocessing in the proposed workflow?

A. Enhance document layout
B. Improve image quality for OCR
C. Remove handwritten notes
D. Extract information from tables

B

Improve image quality for OCR

THANK YOU