A robust section identification method for scanned electronic health records

Anand Subramanian
Praveen Kumar Suresh
Sudarsun Santhiappan

BUDDI AI

07-01-2023

Outline

An Introduction to EHR Processing

- **Electronic Health Records (EHRs):** Narrate and document all the clinical aspects of a patient's visit to a doctor or hospital.
- Usually created and stored in different formats such as free-text files, scanned images, and audio dictations.
- Many downstream clinical activities process EHRs as text. Thus, we use **Optical Character Recognition (OCR)** to extract text from scanned images.

Problems with OCR Processing

- However, OCR engines are sensitive to the quality of documents and images.
- Noise artifacts such as **poor brightness, motion blurs**, and other image issues are present in scanned EHRs.

Problems with OCR Processing

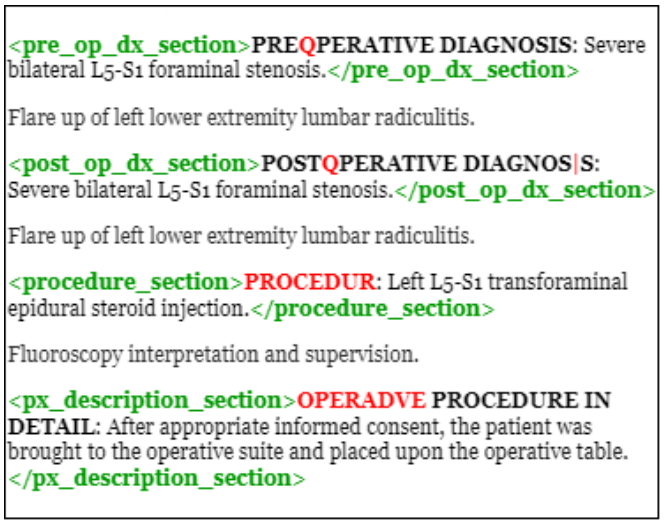

Figure 1: Sample OCR corrupted sections

What is Section Identification?

- EHRs document information such as diagnosed conditions and procedures performed under specific sections.
- The start of such sections is often clearly indicated by specific headers in the text.
- **Section Identification:** Identification of lines that indicate the start of a section.

Why is Section Identification important?

- Section Identification is crucial for downstream clinical activities such as medical coding, diagnosis identification, and medication extraction.
- **Named Entity Recognition (NER)** and **Relation Extraction (RE)** are some essential downstream services for enriching the information in an EHR.

Problems with Section Identification under OCR errors

- OCR errors are characterized by **character** and **word-level errors** in text.
- Traditionally, we use systems that employ lookup-based term matching logic for section identification.

Problems with Section Identification under OCR errors

- Updating term-based lookups variations with OCR corrupted section headers is infeasible due to the vast space of variations possible.
- Adding all possible variants leads to an inflated and non-generic lookup. Regularly updating systems deployed in production involve significant system downtime.

Considerations

- We hypothesize that sequence models with a character-level component may be robust in tackling such issues.
- Training deep-learning models require large amounts of annotated data, which involves significant human labor and cost.
- In contrast, lookup-based systems are more straightforward to update.

Contributions

- A simple **weakly-supervised method** of labeling data using a lookup-based module.
- Development of **character-level word sequence models** for section identification that are robust to OCR errors.
- Application of a **character-augmentation scheme** inspired by techniques from existing literature to improve the robustness of the character-level word sequence models on OCR errors.

Weakly Supervised Labeling Process

- Annotation of documents in the clinical domain can be prohibitively expensive regarding human resources and capital.
- Annotation workflows involve deidentifying documents, assessing the quality of annotations, and annotating significant amounts of data.

Weakly Supervised Labeling Process

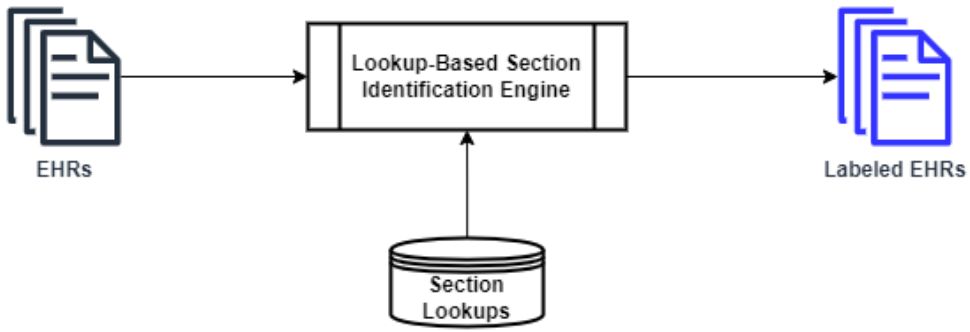

Figure 2: Representation of the Weakly Supervised Labeling Process.

Weakly Supervised Labeling Process

- We exploit the fact that specific headers denote the start of sections, as a source of supervision.
- We leverage a lookup-based system that relies on external lookups curated by subject matter experts (SMEs) and uses a regex-based approach to label sections.

Weakly Supervised Labeling Process

- We label **10000 clean documents** from the Ambulatory Surgery specialty from which the Personal Health Information (PHI) is de-identified.
- We partition this corpus into a train set comprising 8000 documents and a validation set containing 2000 documents and utilize this corpus for training and tuning our models.

Model Training

- Formally, we define the task of section identification as a multi-class text classification problem.
- We implement two architectures, where the first block is a character-level module that learns a contextualized representation of a word from its characters.
- The second block uses the representations from the first block, to learn a contextualized representation of the sentences from its words.

Model Overview

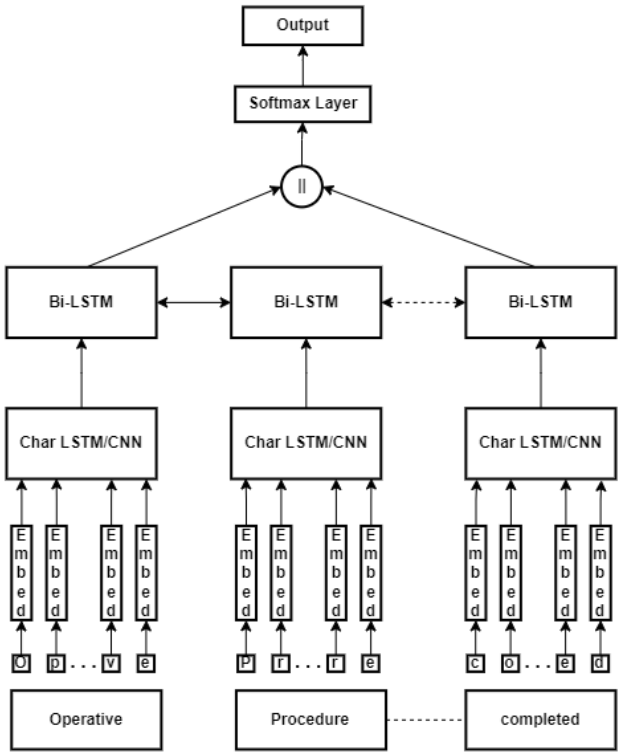

Figure 3: High level model overview

Model Architectures

- The first architecture utilizes a Bidirectional-LSTM [1, 2] at the character level followed by a Bidirectional-LSTM at the word level.
- The second architecture utilizes a 1D-CNN, followed by a MaxPooling layer at the character level followed by a Bidirectional-LSTM at the word level.

Character Augmentation

- Since OCR errors comprise errors at the character level, we introduce character augmentation to improve the model's robustness to OCR issues.
- Character augmentation schemes in previous literature have been shown to improve the performance of NLP systems in the presence of character and word errors [3].

Character Augmentation

- We define an augmentation scheme that applies one of the following transformations:
  1. **Character Substitution**
  2. **Character Insertion**
  3. **Character Deletion**
  4. **No Augmentation**

Evaluation

- We benchmark the performance on a corpus of 1000 documents from the Ambulatory Surgery specialty containing OCR errors with the Personal Health Information de-identified from the charts.

- We compare our models against two baselines:
  1. Lookup-based Module
  2. Word-Level Bidirectional LSTM
- We examine the **macro-average f1 score** to assess how well the systems fare across all sections.

Results

Method	Macro Average F1
Lookup System	0.716
Word LSTM	0.735
Char LSTM	**0.778**
Char CNN	**0.758**
Char LSTM (Aug)	**0.781**
Char CNN (Aug)	**0.780**

Table 1: Results of the models on the test set

Findings

- We observe that the char-LSTM and char-CNN based models outperform both baselines.
- Outperforming the lookup-based method intuitively makes sense as the system would not fare well in the presence of OCR errors.

Findings

- We observe that the character models, even without augmentation, perform strongly in the presence of OCR error compared to the word level model.
- Training the character-level models with augmentation results in an improved performance on the OCR corrupted test set.

Conclusion and Future Work

- We introduce character-level word models with character augmentation for section identification that are robust to OCR errors.
- We empirically demonstrate that these models significantly outperform baseline approaches on documents with character errors.
- Future extensions to this work could explore different model architectures or document-level modeling approaches.

References


Sepp Hochreiter and Jürgen Schmidhuber.
Long short-term memory.
Neural computation, 9(8):1735–1780, 1997.


M. Schuster and K.K. Paliwal.
Bidirectional recurrent neural networks.
IEEE Transactions on Signal Processing, 45(11):2673–2681, 1997.


Yonatan Belinkov and Yonatan Bisk.
Synthetic and natural noise both break neural machine translation.
arXiv preprint arXiv:1711.02173, 2017.