# Automated Digitization of Unstructured Medical Prescriptions

**Megha Sharma**[*] Amazon
**Tushar Vatsal**[*] Amazon
**Srujana Merugu** Amazon
**Aruna Rajan**[†] Google

## Abstract

Automated digitization of prescription images is a critical prerequisite to scale digital healthcare services such as online pharmacies. This is challenging in emerging markets since prescriptions are not digitized at source and patients lack the medical expertise to interpret prescriptions to place orders. In this paper, we present prescription digitization system for online medicine ordering built with minimal supervision. Our system uses a modular pipeline comprising a mix of ML and rule-based components for (a) image to text extraction, (b) segmentation into blocks and medication items, (c) medication attribute extraction, (d) matching against medicine catalog, and (e) shopping cart building. Our approach efficiently utilizes multiple signals like layout, medical ontologies, and semantic embeddings via LayoutLMv2 model to yield substantial improvement relative to strong baselines on medication attribute extraction. Our pipeline achieves +5.9% gain in precision@3 and +5.6% in recall@3 over catalog-based fuzzy matching baseline for shopping cart building for printed prescriptions.

## 1 Introduction

In recent years, prompted by the COVID pandemic, there has been a rise in the adoption of online pharmaceutical services leading to improved access to medications and health outcomes. However, in emerging markets such as India, online pharmacy ordering continues to be challenging since prescriptions tend to be paper-based, unstructured and often, handwritten, which makes digitization a vital prerequisite. For in-store purchases, customers follow a simple process of presenting a prescription to the store pharmacist who interprets it and fulfills the order. Current e-commerce purchase process, however, imposes a significant cognitive load on customers since they have to explicitly specify the medicines. This process is onerous for the customers due to their (a) unfamiliarity with the ordering process, (b) difficulty in understanding prescriptions, and (c) lack of expertise to interpret medical acronyms and identify substitute medicines. Further, most online pharmacies have a post-cart creation workflow where customers upload the prescription to be verified by a remote pharmacist. Lack of pharmacist capacity often leads to long wait time making the process unscalable. Therefore, an automated system that converts prescription images to a digitized form to facilitate search-less shopping is essential for the success of online pharmacies. In particular, we need to extract the medical advice section which contains a list of medication items, each of which is a record of multiple fields such as BRAND-NAME.

**Challenges.** Addressing this problem is non-trivial due to multiple reasons shown in Figure 3a: (a) variability in prescription image quality, background, and orientation, (b) diversity of layouts and doctor styles, (c) high prevalence of typos that create confusion between similar items (e.g., Fibrodone and Firodone), (d) specialized vocabulary of regional prescriptions, and (e) the need for converting dosage-specific instructions to a precise product order. Additionally, there are limited labeled prescriptions due to the high manual effort it entails.

**Related work.** While there have been significant advances in document AI [6, 15, 19] and information extraction [29, 21, 13] techniques, most of these methods are effective only on images of well-formatted documents such as invoices. Besides, these generic methods require significant supervision and are not sufficiently modular to support a phased automation of the prescription processing workflow. Recent work on digitizing medical prescriptions [27] is focused on using named entity recognition (NER) methods for medication attribute detection, but these models perform poorly on non-US prescriptions due to vocabulary gaps

[*]Equal contribution
[†]Work done while at Amazon



*Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*
*Volume 5: Industry Track*, pages 794–805

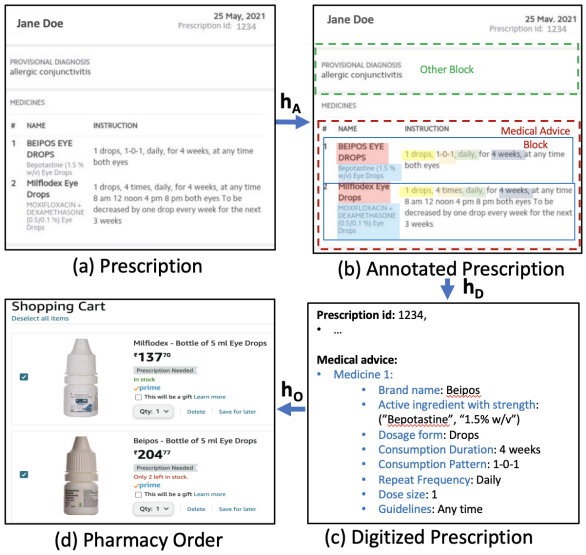



Figure 1: Prescription processing stages & entities. Only fictional prescriptions are shown in the paper for privacy reasons.

and do not utilize the layout or catalog information. We present additional related work in Appendix B.

**Contributions.** In this paper, we present a study on automated digitisation of prescription images with printed content which covers design, data, modeling and evaluation aspects. We discuss experimental results for our emerging modular pipeline which comprises a mix of ML and rule-based components for (a) image to text extraction and normalization (b) segmentation into blocks, medication items and extraction of medication attributes, (c) matching against pharmacy catalog, and (d) shopping cart building. We detail how our approach efficiently combines layout signal, medical ontologies, and semantic embeddings via LayoutLMv2 model to yield substantial improvement relative to strong baselines on medication attribute extraction, and results in +5.9% and +5.6% gain in precision@3 and recall@3 over catalog-based fuzzy matching baseline for printed prescriptions. We discuss key learnings relevant for low data regime document AI systems in addition to presenting component-wise efficacy of our pipeline and results from ongoing experiments (Appendix A.4.2) to highlight future directions. We present safety aspects in Section 9.

## 2 Prescription Digitization Problem

Given a prescription image, a natural choice for digitization is in terms of conversion to a structured prescription object as per a global standard such as the Fast Healthcare Interoperability Resources (FHIR) framework [3]. Since our objective is to create a shopping cart for automated medicine ordering we focus on populating the relevant fields only in the FHIR prescription schema (Table 5). To accommodate the nuances of regional medications, we define each *medicine* in the pharmacy catalog as a unique tuple of BRAND-NAME or GENERIC-NAME[1], FORM, INGREDIENT and STRENGTH. A unique pair of a medicine and package details corresponds to a stock keeping unit (SKU).

Figure 1 depicts the various stages of processing a prescription (denoted by $h_A$, $h_D$, $h_O$) that results in the successive creation of following entities: *(a) Annotated Prescription* is a visually rich document (VRD) comprising labeled rectangular bounding boxes (BBs). Each BB is associated with text and a list of annotations, which include the start-end offsets and labels corresponding to medication attributes, item boundaries, and block type, *(b) Digitized Prescription* is a structured object with canonical entries following the FHIR-based prescription schema, *(c) Pharmacy Order* is a list of SKUs from the prescription along with the recommended quantities. The conversion to a pharmacy order ($h_O$) can be enabled via a deterministic lookup using a medicine-SKU map if the medication codes in the digitized prescription are from the catalog. Hence, we focus on the non-trivial transformations $h_A$ and $h_D$ that entail a data-driven approach.

Let $\mathcal{P} = \{p_i\}_{i=1}^{N}$ denote the set of the available prescription images for training. For the $i^{th}$ prescription $p_i$, let $(a_i, d_i, o_i)$ denote the human annotated prescription, digitized prescription, and pharmacy order obtained from expert pharmacists, i.e., $a_i = h_A(p_i)$, $d_i = h_D(a_i)$, $o_i = h_O(d_i)$. Typically, pharmacists directly create or validate pharmacy orders from a prescription image without any record of the intermediate annotation and digitization. Since these prescription-order pairs are inadequate for an end-to-end neural model, we explicitly gather supervision for the intermediate stages for a subset of the prescriptions to enable a pipelined approach. Let $z_i^A, z_i^D, z_i^O$ denote binary indicators of the availability of $a_i, d_i, o_i$ respectively. Further $L^A(\cdot,\cdot), L^D(\cdot,\cdot), L^O(\cdot,\cdot)$ be suitable loss functions for comparison of pairs of candidate annotated versions, digitized versions, and orders corresponding to a prescription such as the accuracy of annotations, matching with canonical entities, and the constructed order respectively as shown in Table 2. Then, the training objective

[1] Generic names are globally approved, e.g., paracetamol, while brand names are manufacturer given e.g., Calpol.

is to learn mappings $\hat{h}_A$ and $\hat{h}_D$ that produce high fidelity reconstructions of the processed versions of the prescription and can be viewed as a loss minimization:

$$\min_{\hat{h}_A, \hat{h}_D} \Big[ \sum_{i|z_i^A=1} L_A(a_i, \hat{h}_A(p_i)) + \sum_{i|z_i^D=1} L_D(d_i, (\hat{h}_A \circ \hat{h}_D)(p_i)) + \sum_{i|z_i^O=1} L_O(o_i, (\hat{h}_A \circ \hat{h}_D \circ h_O)(p_i)) \Big]$$

Given a new prescription, the learned mappings $(\hat{h}_A, \hat{h}_D)$ along with $h_O$ yield a pharmacy order.

## 3 Solution Design

### 3.1 Design Choices

We discuss the key tenets and design choices of our prescription digitization approach (Figure 3b).

**Modularity.** Supporting phased automation of user-driven cart building and pharmacist-driven validation workflows entailed a modular pipeline.

**Solution choice dependent on input signals.** Limited labeled prescription data coupled with access to medical ontologies made it prudent to choose a hybrid combination of rule-based and ML modules instead of an end-to-end deep neural model.

**Interoperability.** The need to interface with other healthcare systems led us to choose a data representation based on global FHIR standards.

**Extension over reinvention.** Fast and scalable implementation required use of existing solutions for sub-problems wherever acceptable and focusing on exploration of the harder sub-problems.

### 3.2 Components

We describe components of Figure 3b below.

**Text extraction & VRD Normalization.** First, we identify OCR bounding boxes (BBs) and extract the text from these BBs. Then we perform rotation and background cropping, using the position coordinates of BBs, to create normalized VRDs with more homogeneous layouts as shown in Figure 2.

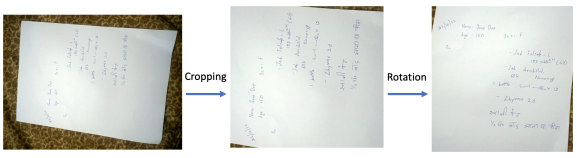



Figure 2: Steps in VRD normalization pipeline.

**Entity Annotation.** Annotating BBs comprises three tasks corresponding to stage (b) of Figure 1: (a) detecting block(s) containing medical advice of doctor, (b) chunking of words, within medical advice block, related to medication(s) into item(s), and (c) extracting medication attributes such as brand name, duration of consumption from an item. Though a joint model that optimizes $\sum_{i|z_i^A=1} L_A(a_i, \hat{h}_A(p_i))$ to simultaneously detect blocks, medication items, and attributes seems like a natural choice, it is prohibitive due to the constraints on amount of supervision, computational effort, and limits on context size of NLP models (usually 512 tokens). We simplify this problem by solving the sub-tasks in the order $(a \rightarrow c \rightarrow b)$. Advice block detection reduces sequence length (as shown in figure 4) permitting transformer-based encodings and increasing precision for later tasks. In this task, we construct latent representation of the BBs based on position, semantics, and membership in medical ontologies and learn a supervised classification model to predict whether a BB contains medical advice or not. We perform step (c) and (b) only on BBs predicted as advice blocks. For medication attributes, we label each token in advice BB using our NER model into one of 7 classes (DURATION, FORM, FREQUENCY, INGREDIENT, ITEM-MARKER, BRAND-NAME, and STRENGTH). Sequence of NER predictions are fed into our heuristic algorithm for medication item chunking that leverages relative positions of BRAND-NAME, STRENGTH and FORM tokens. **Matching and Canonicalization.** The next step ($\hat{h}_D$) is to map each annotated medication item in the prescription (e.g., T.[FORM] Crocin[BRAND-NAME], 5 ml[STRENGTH]) to a medicine ID in the pharmacy catalog using extracted attributes. For this we use our Pharmacy product catalog as a reference. This catalog contains all medicine products listed on our website and each product is described by a set of attributes such as BRAND-NAME and FORM. We adopt a two-stage approach comprising: (a) identifying candidates by fuzzy matching predicted BRAND-NAME with that in catalog, (b) computing a match score based on FORM and STRENGTH to identify the best matching medicine ID using either a rule-based or an ML classifier.

**Cart Building.** The final step ($h_O$) is to construct the pharmacy order, i.e., list of required SKUs and their quantities. To enable this, the standard dosage amount of SKU is computed during catalog creation, e.g., 3 packs of 30 ml bottle maps to 90 ml. From the digitized prescription, total recommended dosage amount can be computed from dosage duration, daily dosage pattern and units to be consumed at a time. Appropriate SKU and its quantity can be

derived to minimally exceed this amount.

## 4 Data Collection

Due to the sensitive nature of prescriptions and recent emergence of our medicine ordering application, a public dataset of unstructured prescription images does not exist to the best of our knowledge. External benchmarks such as [5] only contain clean text without any layout information. Hence, we use a proprietary dataset of 1359 Indian prescriptions paired with (fully or partial) digitized orders, delinked from customer IDs. These are mostly or fully printed prescriptions and have been validated by our in-house experts. The prescription images are modified as follows prior to modeling. AWS Textract, which is security certified for critical data, is used to extract text from images. The obtained text is then run through AWS Comprehend to detect personally identifiable information such as patient/doctor names, phone numbers and then the corresponding OCR bounding boxes are grayed out. For a subset of prescriptions, we procured in-house human annotations for supervised training of all components. Ground-truth text, BBs for medical advice blocks, labeled text spans for medication attributes, as well labels for pairs of candidate and ground truth SKUs for medication matching were annotated in the prescription image by the annotators. More details on the annotation tasks are given in Table 6. Table 1 lists details of the training and evaluating splits for various components. Given the expensive labeling effort, this data size is realistic for early-stage specialized document AI systems.

## 5 Experimental Results

We present our evaluation method and results on the efficacy of the full system and various components with focus on medication attribute extraction.

### 5.1 Evaluation Methodology

Practical systems need to be evaluated during development (offline metrics) and post deployment (online metrics). Table 1 lists our offline evaluation metrics. Most of these are self-explanatory except Brand match which is the percentage of medicine brand names ordered by the customers in the extracted text and indicates the medical text extraction efficacy. The online metrics of our system (not reported for proprietary reasons) depend on whether the digitization is integrated into the pharmacist processing flow or the customer-facing UI. These include rate of correction of automated cart suggestions, reduction in cart-building time, reduction in order rejections during verification stage as well as business metrics on the order volume.

### 5.2 Component-wise Efficacy

Table 2 lists the metrics of various components of our digitization pipeline, which we discuss below.

**Text Extraction.** Due to limited supervision, we use pretrained off-the-shelf solutions. AWS Textract is our preferred choice as it provides a higher brand match (+7%) than AWS Rekognition as the latter has limit on the number of extractable words.

**Advice Block detection.** To reduce complexity for downstream tasks, we first detect medical advice blocks. We employ a two-stage solution (see Figure 4) of (a) clustering BBs using K-means on their positional coordinates, and (b) classifying each cluster as advice block or not using XGBoost [8] classifier trained on cluster position, and fractions of medical and printed words. Lastly, adjacent advice blocks are merged. This method can be extended to other block types (e.g., header, footer) using a multi-class classifier and block-type indicators. Our solution results in an operational point with 94.8% recall, 88.1% precision, reduction in block size (Figure 5) and mostly homogeneous clusters (homogeneity score: 0.857). Common errors occur due to: (a) sparse text that cannot use local semantic context well leading to false positives, and (b) long lines that are ideally a single cluster but split because of the high divergence in the horizontal dimension.

**Medication Item Chunking.** We exploit the observation that attributes of a medication are contiguous with BRAND NAME preceding STRENGTH and the ordering relative to FORM being flexible. Let $t_k$ be the $k^{th}$ detected BRAND NAME token. For each $t_k$, we construct up to two candidate medications with brand based on $t_k$, STRENGTH based on the closest STRENGTH token to $t_k$ in the span $(t_k, t_{k+1})$, and FORM derived from the FORM tokens closest to $t_k$ on either side in the spans $(t_{k-1}, t_k)$ and $(t_{k-1}, t_k)$. Our approach yields high accuracy (97.2%) obviating the need for an ML system.

**Matching and Canonicalization.** As discussed in Section 3.2, we employ a two-stage approach of filtering and ranking using match score. For match score computation, we consider two methods using the same attribute fuzzy scores as inputs: 1) rule-

[2] https://en.wikipedia.org/wiki/Word_error_rate

[3] Strict matching metrics as per SemEval-13 [7].

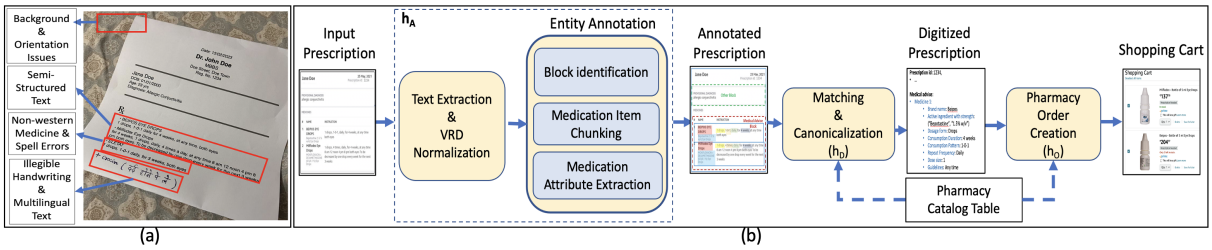



Figure 3: (a) Illustration of the challenges with prescription digitization in emerging markets. The image presents a representative Indian prescription, (b) Flow of automated order creation from prescription images

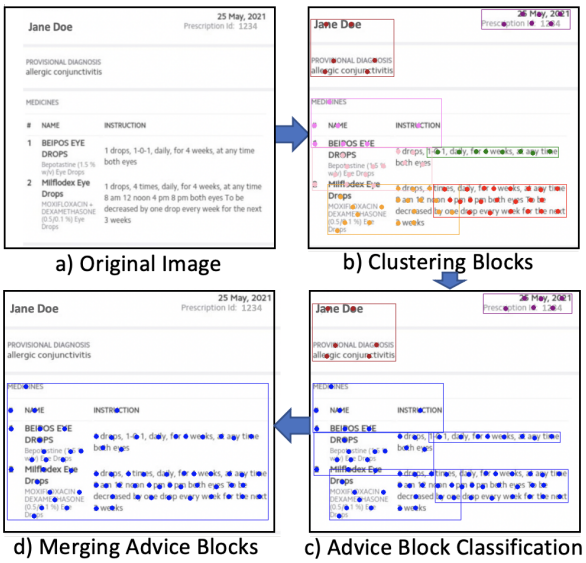



Figure 4: Flow chart of medical advice block detection module. Highlighted blue boxes are predicted advice blocks.

based one with heuristics for handling missing values and 2) XGBoost classifier that predicts whether a chunked item matches a candidate SKU using the same features as rule-based one except heuristics. Empirically, we see that the former approach yields +18% gain over the latter in precision@1 of the top matching item. Note that precision@1 is the same as recall@1 due to stand-alone evaluation.

### 5.3 Medication Attribute Extraction

Table 3 shows performance of various approaches based on multiple input signals: **(a) Catalog Features:** These include membership scores of tokens with respect to dictionaries of BRAND-NAME, INGREDIENT, STRENGTH (e.g., mg) and FORM (e.g., tablet) created from the catalog, **(b) Semantic features:** These include contextual text embeddings derived from transformer models such as BERT [10] and MedBERT [25] pretrained on Wikipedia and PubMed respectively, **(c) Layout features:** Since the layout provides extra information, e.g., text in the middle is usually medical advice, we use LayoutLM [32] and LayoutLMv2 [31] models, which have multi-modal Transformer architecture as backbone and utilize layout, visual, and textual features to learn cross-modal interactions, and **(d) Collective labeling:** We use the linear conditional random field (CRF) loss to exploit relationships amongst labels, e.g. BRAND-NAME often lies between FORM and STRENGTH.

Note than in Table 3, the token level metrics are weighted with token length so that errors on small tokens are less penalized and OTHER tokens are excluded as these are not critical for the application. From the results, we observe that XGBoost trained with catalog features performs comparable to custom Comprehend fine-tuned on our data illustrating the importance of catalog signal. While BERT-based models using semantic features further improve the performance, the best accuracy is seen when we incorporate layout features (LayoutLMV2 variants) as well. Note that models such as Comprehend Medical and MedBERT are not suitable for our problem as these are not trained on the Indian medicine vocabulary.

**Ablation Studies.** To evaluate the efficacy of the various signals as well as modeling sequential dependencies via CRF, we conducted ablation studies. Since LayoutLMv2 already uses semantic and layout features, we added collective labeling (LayoutLMv2 + CRF) and catalog features (LayoutLMv2 + CF) separately and in a combined setting (LayoutLMv2 + CF + CRF). We note from Table 3 that performance only changes marginally. Similar behavior is observed when using BERT-variants indicating that catalog and collective labeling are subsumed by the semantic and layout encoding. Good performance of XGBoost + catalog features variant points to presence of non-linear interactions and value of catalog signal. Further studies (Appendix A.4.2) indicate that the performance depends on the quality and diversity of supervision more than than the quantity pointing to

| Task | Data split | Metrics | Approaches |
|---|---|---|---|
| Text extraction | (-, -, 10) | Brand match % | Textract, Rekognition |
| Block detection | (44, 5, 13) | Precision (P), Recall (R) | K-means + XGBoost |
| Medication attribute extraction | (977, 190, 192) | Token & entity level P, R, F1 [3] | Refer Table 3 |
| Medication Item Chunking | (977, 190, 192) | Accuracy | Rule based |
| Matching & canonicalization | (272, 46, 58) | Precision@k (P@k) | Rule based, XGBoost |

Table 1: Details of dataset (train,val,test splits), evaluation metrics, and approaches for sub-tasks.

| Method | Brand match% |
|---|---|
| Textract | 56.17 |
| Rekognition | 49.38 |

(a) Text Extraction

| Method | Precision | Recall |
|---|---|---|
| K-means + XGBoost | 0.881 | 0.948 |

(b) Block Detection

| Method | P@1 |
|---|---|
| Rule based | 0.945 |
| XGBoost | 0.765 |

(c) Matching

| Method | Accuracy |
|---|---|
| Rule based | 0.972 |

(d) Medication Chunking

Table 2: Efficacy of various stages of pipeline excluding medication attribute extraction.

| Type | Model | Token Precision | Token Recall | Token F1 | Entity Precision | Entity Recall | Entity F1 |
|---|---|---|---|---|---|---|---|
| AWS Solutions | Custom Comprehend | 0.955 | 0.882 | 0.915 | 0.774 | 0.790 | 0.782 |
| Catalog Features (CF) | CF + XGBoost | 0.973 | 0.870 | 0.917 | 0.766 | 0.780 | 0.773 |
| Semantic Features | BERT | 0.975 | 0.913 | 0.942 | 0.802 | 0.829 | 0.815 |
| | MedBERT | 0.974 | 0.893 | 0.931 | 0.802 | 0.811 | 0.806 |
| Layout Features | LayoutLM (LLM) | 0.981 | 0.918 | 0.948 | 0.826 | 0.834 | 0.830 |
| | LLMv2 | **0.983** | **0.926** | **0.953** | 0.829 | **0.842** | 0.835 |
| | LLMv2 + CF | 0.981 | 0.915 | 0.946 | **0.835** | 0.837 | 0.836 |
| Collective Labeling | BERT + CRF | 0.974 | 0.903 | 0.936 | 0.800 | 0.816 | 0.808 |
| | BERT + CF + CRF | 0.974 | 0.896 | 0.932 | 0.808 | 0.814 | 0.811 |
| | LLMv2 + CRF | 0.982 | 0.921 | 0.950 | **0.835** | 0.840 | **0.838** |
| | LLMv2 + CF + CRF | 0.983 | 0.921 | 0.950 | 0.830 | 0.835 | 0.832 |

Table 3: Performance of various NER methods on medication attribute extraction.

the benefits of using active learning approaches.

**Error Diagnosis.** Table 4 presents an error diagnosis of our best model (LayoutLMv2) and areas of improvement such as deducing labels from context (e.g. "tablet once a day" → Frequency). Figure 9 presents the confusion matrix of different medication attribute classes.

### 5.4 Overall Cart Building Efficacy

We evaluate the overall pipeline on a test set of 179 orders (71% are partially digitized) consisting of 200 digitized medication items. We predict top K (K=3) SKUs for each medicine identified in the prescription image for customer safety and evaluate our approaches on precision@3 (i.e., fraction of predicted being in the ground truth orders) and recall@3 (i.e., fraction of actual ordered medications being detected). Since precision estimate is based on partial orders, it is pessimistic. The baseline method performs fuzzy matching of attributes (e.g., BRAND-NAME, FORM) of catalog items with n-grams from complete prescription text and selects top K SKUs for each prescribed medicine. Our proposed approach combines the best version of each component from Section 3.2 and gets **+5.9%** in precision@3 and **+5.6%** in recall@3 over the baseline.

**Error diagnosis.** Primary gaps in our approach include: (a) Text extraction errors, e.g., capsule extracted as "apsule" resulting in misclassification as form; (b) Limited semantic understanding of the model, e.g., "once a month" denotes Frequency but was predicted as Duration; (c) Token not exclusively associated with a label, e.g., “syrup” is usually Form, but "corn syrup” is an Ingredient, and (d) Minor variations in medication attributes (e.g., "LosarH" vs "LosarCH") which can be handled by including INGREDIENT during matching.

## 6 Learnings

Below are our key learnings on building document AI systems for low data regime:

**Annotation design is critical.** Annotation tasks (drawing BBs, text chunking) should be well-specified with low cognitive load and include all the

| Actual \ Predicted | ITEM-MARKER | INGREDIENT | FORM | DURATION | BRAND-NAME | STRENGTH | FREQUENCY | OTHER |
|---|---|---|---|---|---|---|---|---|
| ITEM-MARKER | | | | | | | | 6 . TAB |
| INGREDIENT | Motia 3 | | Corn Syrup | | Pregabalin | | | Para Sucphate |
| FORM | ctab duone-mer | apsule | | | T-FIL | | ointment at bedtime | tab after breakfast |
| DURATION | | | | | | | | tues / thurs / sat |
| BRAND-NAME | | 2 Clon-azepam | Threptin Diskettes | | | VOGS GM | | Rient OD |
| STRENGTH | | 100 billion spores | | | SPF 50 | | | 10 gm |
| FREQUENCY | 1 unit | | tablet once a day | once a for month | | 2 TSF | | bed Ativan time |
| OTHER | 1 . D-Rise | Glargin composi-tion Insulin | 1 tab oral | Continue | Bioderma Sebium | vertin tab 16mg | 1-0-1 single dose | |

Table 4: **Error diagnosis matrix**: Words colored in red belong to the row attribute and are confused for the column attribute. For example, "Corn Syrup" is labeled as INGREDIENT but Syrup is wrongly predicted as FORM. There are few primary reasons for the errors: (a) token being used with multiple labels, e.g., “syrup” is a common term in FORM, but "Corn Syrup” is a special case where it is INGREDIENT. (b) Text extraction errors, e.g., Capsule detected as "apsule" resulting in it being labeled as INGREDIENT instead of FORM. (c) Limited semantic understanding of the model (e.g., once a month is an expression for FREQUENCY), and (d) High fraction of the OTHER class resulting in biased decisions.

relevant input (e.g., raw images) to avoid cascading errors. This is especially true for annotations on VRD output from OCR which could itself be erroneous. Building an annotation UI that leverages existing models but allows for manual corrections as part of a semi-automated workflow is an ideal strategy for progressive improvement.

**Divide and conquer.** Despite the ubiquity of end-end neural models, it is vital to choose a solution approach based on application constraints, e.g., data limitations, the need for modularity to support phased development and audibility. We adopted a divide-and-conquer approach by partitioning our problem into sub-tasks which could be solved separately using domain knowledge where possible. Our multi-stage solution is extensible and reusable across different workflows and data segments.

**Model and problem complexity should match.** Ideal performance is obtained when complexity of approach matches that of the problem conditioned on available data and domain knowledge. We noticed in our case that richer ML models were comparable or under-performed simpler ML models and domain heuristic-based approaches in medicine chunking and matching tasks due to less data.

## 7 Concluding Remarks

Prescription digitization is a critical enabler of online pharmacy services. We present a holistic, modular approach to address this problem in a low data regime using hybrid ML and rule-based components. Our approach uses layout signals, medical ontologies, sequential dependencies, and semantic embeddings to yield significant improvement over baselines and good performance on unstructured printed prescriptions. Ongoing directions include using active learning to judiciously label data (section A.4.2), pseudo labeling of partially digitized orders and digitizing handwritten prescriptions.

## 8 Limitations

Our prescription digitization approach has a few limitations but is still effective for a broad enough application domain and permits future enhancements that address these limitations. First, our system uses an off-the-shelf text extraction tool (AWS Textract) that provides accurate extractions on printed prescriptions but has variable performance on hand written data depending on the legibility of the handwriting. In future, we plan to build a specialized extraction model trained to recognise medical practitioner’s handwriting to replace AWS TextExtract. Further, multiple components in our approach (e.g., attribute extraction) have been trained on primarily English transcriptions. Extension to other language prescriptions requires access to medical vocabulary and training data in those languages. Note that AWS Textract supports multiple languages and can be readily paired with an automated translator to convert the content to English. We did not consider this option since multilingual prescriptions in India tend to have mixed content with medications written in English itself. Lastly, the performance of multiple tasks such as advice block detection, medication attribute extraction and matching-canonicalization depends on the coverage of the available medical catalog.

## 9 Safety and Ethics Statement

Our motivation is to improve access and affordability to online pharmaceutical services in emerging markets such as India through accurate and easy digitization of medical prescriptions. Given the sensitive nature of medical prescriptions and the associated health impact, it was critical to pay attention to multiple aspects that we discuss below:

**Secure and Privacy-safe Data Collection:** Privacy of customer data is paramount to us. Hence, prior to modeling, we remove customer, facility and practitioner information by obscuring the regions containing personally identifiable fields such as names, phone numbers, and addresses, which are identified using security-certified AWS services (AWS Comprehend, AWS Textract).

**Model Bias:** A key limitation of the existing medical NER models is their poor performance on non-US and EU prescriptions due to bias in the training data, which is almost exclusively based on US-EU centric medical content and vocabulary. In our approach, we have deliberately chosen to have explicit dependence on aspects that vary across geographical regions (e.g., medical catalog), which enhances the applicability of our approach. To further limit the model bias and minimize distributional differences between training and production settings, we have trained our models on prescription images that are randomly sampled from customer uploads. These often include low resolution and improperly positioned images. In future, as the scope of deployment changes, we plan to periodically retrain the model with training images by sampling from the production data.

**Health Safety:** One of the primary concerns in prescription digitisation is the impact of errors on patient health and adherence to health regulations. To alleviate adverse outcomes, we have multiple guardrails. First, we present the top three suggestions along with scores for each medication for two-fold review by customer and pharmacist. Second, to avoid prescription abuse (e.g., manipulation of quantities, prescription reuse) and comply with regulations, there are additional checks based on the prescription date, patient purchase history, and recommended limits on medication quantities.

**Usage for a Limited Scope:** Our proprietary system has been trained for a specific-use case, i.e., prescription digitization with acceptable performance on primarily English printed prescriptions for India region. We plan to use the model within this limited scope and expand usage only after adequate benchmarking. To limit the risks of misuse, we do not plan to release this system externally.

## 10 Acknowledgements

We would like to thank our product team (Sourabh Jeswani, Jayaramakrishnan Balasubramanian) and engineering team (Varnit Agnihotri, Ranjith Sompalli, Saurabh Gupta, Amir Kamal, Vinay Vaidya) for supporting us during problem formulation, data gathering and providing continuous feedback.

## References

[1] Amazon Rekognition. https://aws.amazon.com/rekognition/.

[2] Amazon Textract. https://aws.amazon.com/textract/.

[3] Fast Healthcare Interoperability Resources. https://www.hl7.org/fhir/.

[4] Google Cloud Vision. https://cloud.google.com/vision/.

[5] NLP Research Data Sets. https://www.i2b2.org/NLP/DataSets/.

[6] Joe Barrow, Rajiv Jain, Vlad Morariu, Varun Manjunatha, Douglas Oard, and Philip Resnik. A joint model for document segmentation and segment labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 313–322, 2020.

[7] David S. Batista. Named-entity evaluation metrics based on entity-level. https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/, 2018.

[8] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 785–794, 2016.

[9] Arthur Flor de Sousa Neto, Byron Leite Dantas Bezerra, Alejandro Héctor Toselli, and Estanislau Baptista Lima. HTR-Flor: A deep learning system for offline handwritten text recognition. In *33rd SIBGRAPI Conference on Graphics, Patterns and Images*, pages 54–61, 2020.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[11] Son Doan, Lisa Bastarache, Sergio Klimkowski, Joshua C Denny, and Hua Xu. Integrating existing natural language processing tools for medication extraction from discharge summaries. *Journal of the American Medical Informatics Association*, 17(5):528–531, 2010.

[12] Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

[13] Benedict Guzman, Isabel Kayu Metzger, Yindalon Aphinyanagphongs, and Himanshu Grover. Assessment of amazon comprehend medical: Medication information extraction. *ArXiv*, abs/2002.00481, 2020.

[14] Cheryl Clark Meredith Keybl David Tresner-Kirsch John Aberdeen, Samuel Bayer. An annotation and modeling schema for prescription regimens. *J Biomed Semantics 2019; 10(1): 10.*, 2019.

[15] Anoop R Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. Chargrid: Towards understanding 2d documents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4459–4469, Brussels, Belgium, 2018.

[16] Ali Can Kocabiyikoglu, Jean-Marc Babouchkine, François Portet, and Raheel Qader. Neural medication extraction: A comparison of recent models in supervised and semi-supervised learning settings. In *IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 148–152, Aug 2021.

[17] A. L. Koerich, R. Sabourin, and C. Y. Suen. Large vocabulary off-line handwriting recognition: A survey. Pattern Analysis & Applications, 2003.

[18] Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 32–39, Minneapolis, Minnesota, 2019.

[19] Bodhisattwa Prasad Majumder, Navneet Potti, Sandeep Tata, James Bradley Wendt, Qi Zhao, and Marc Najork. Representation learning for information extraction from form-like documents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6504, 2020.

[20] Eric Medvet, Alberto Bartoli, and Giorgio Davanzo. A probabilistic approach to printed document understanding. International Journal on Document Analysis and Recognition (IJDAR), 2011.

[21] Rasmus Berg Palm, Dirk Hovy, Florian Laws, and Ole Winther. End-to-end information extraction without token-level supervision. In *Proceedings of the Workshop on Speech-centric Natural Language Processing*, pages 48–52, Copenhagen, Denmark, 2017. Association for Computational Linguistics.

[22] Rasmus Berg Palm, Ole Winther, and Florian Laws. Assessment of amazon comprehend medical: Medication information extraction. *ArXiv preprint*, abs/1708.07403, 2017.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 8024–8035, 2019.

[24] Jon Patrick and Min Li. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *Journal of the American Medical Informatics Association*, 17(5):524–527, 2010.

[25] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction. *ArXiv preprint*, abs/2005.12833, 2020.

[26] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, 2008.

[27] Carson Tao, Michele Filannino, and Zlem Uzuner. Prescription extraction using CRFs and word embeddings. *J. of Biomedical Informatics*, 72(C):60–66, 2017.

[28] Ozlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.

[29] Shaolei Wang, Yue Zhang, Wanxiang Che, and Ting Liu. Joint extraction of entities and relations based on a novel graph scheme. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4461–4467, 2018.

[30] Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. ChemNER: Fine-grained chemistry named entity recognition with ontology-guided distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5227–5240, 2021.

[31] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, 2021.

[32] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. LayoutLM: Pre-training of text and layout for document image understanding. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event*, pages 1192–1200, 2020.

[33] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. Every document owns its structure: Inductive text classification via graph neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 334–339, 2020.

[34] Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu. Document modeling with graph attention networks for multi-grained machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6708–6718, 2020.

# A Additional Solution Details

## A.1 Prescription Schema

Refer to Table 5 for prescription schema.

| Field | Type | Description |
|---|---|---|
| Medicine ID | Identifier | Unique code in catalog |
| Duration | Numeric | #Days to consume the medicine |
| Consumption Pattern | Enum | Consumption pattern of the doses, e.g., 1-0-0 |
| Repeat Frequency | Numeric | For medications consumed with gaps across days |
| As Needed Indicator | Boolean | Set to true if medicine is to be taken SOS |
| Dosage Size | Numeric | Size of the dose |
| Dosage Units | Enum | Units for quantifying dose (e.g., 1 ml, 1 tablet) |
| Additional Instruction | String | Guidelines on consuming the medicine |

Table 5: Schema for digitized prescription which is compliant with FHIR standard.

## A.2 Annotation Tasks

Refer to Table 6 for details on annotation tasks.

| Annotation Task | Labels |
|---|---|
| Block Identification | Medical advice, Other |
| Medication Item Chunking | B, I, O label for medication item segments |
| Medication Attribute Extraction | B, I, O labels based on entities below<br>a) DURATION<br>b) FORM<br>c) FREQUENCY<br>d) INGREDIENT<br>e) ITEM-MARKER<br>f) BRAND-NAME<br>g) STRENGTH |

Table 6: Annotation of VRD is done in three ways - (a) forming BB around relevant block such as medical advice, (b) identification of series of tokens which form one medication item, (c) extraction of attributes required for identifying medicinal items, e.g., BRAND-NAME, STRENGTH.

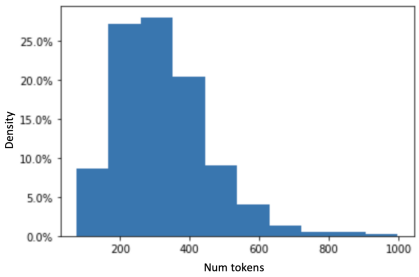
## A.3 Advice Block Detection

Figure 5 shows the reduction in length of prescriptions with advice block detection.

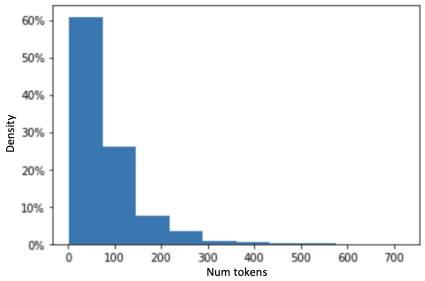
## A.4 Medication attribute extraction

### A.4.1 Training setup and details

For training LayoutLMV2 model (our best performing model), PyTorch [23] is used and the pretrained model is taken from open-source Huggingface library. Batch size of 2 and dropout of 0.1 is used



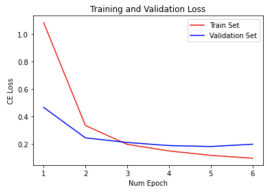

(a) Token count density distribution for prescriptions



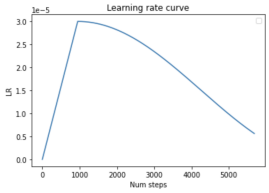

(b) Token count density distribution for Advice blocks

Figure 5: Histogram for token sequence length for the entire prescriptions and advice blocks. Note that advice blocks tend to be much smaller than the 512 tokens required for a transformer.

for model training. Learning rate schedule and loss convergence curves are shown in Figure 6. Model architecture for LayoutLMv2 is shown in Figure 7. Adam optimizer is used with exponential decay rates for first moment and second moment estimated as 0.9 and 0.99 respectively.




(a) Loss curve epoch wise




(b) Learning rate step wise

Figure 6: Details of the training set up for the LayoutLMv2 Model.

### A.4.2 Efficient Use of Unlabeled Data.

Training the model with an increasing number of randomly chosen prescriptions indicated that there is improvement in performance, but at a relatively slow rate. Since labeling effort is much more expensive than acquisition of unlabeled prescriptions, we explored using common active learning methods [26] to prioritize the selection of prescriptions for labeling. Figure 8 shows the learning curves using increasing training data size with selection based on random sampling, entropy of class posteriors, and product of entropy as well as normalized occurrence frequency in the unlabeled data. The results point to potential benefits of judicious prioritization but more exploration is required to optimally combine the entropy and frequency signals.

## B Related Work

Our work is primarily related to four areas of research that we briefly review below.

**Document AI** is a multi-disciplinary area centered

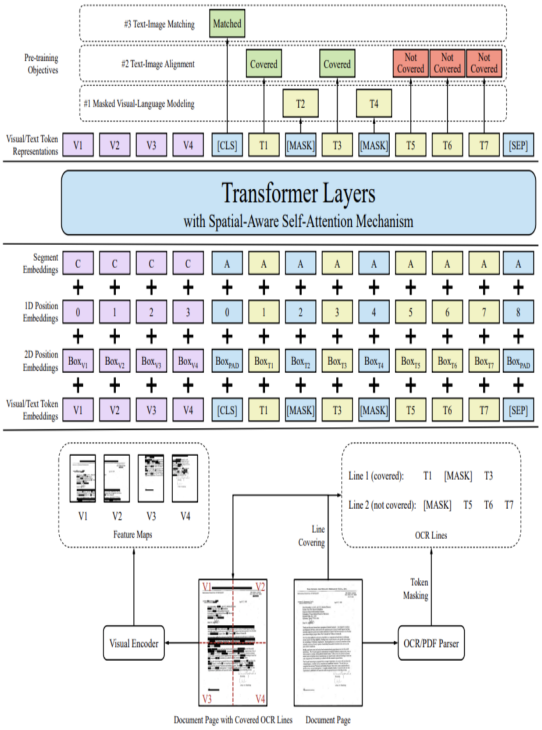



Figure 7: LayoutLMv2 Model Architecture from [31]

on understanding visually rich documents (VRDs) using techniques [22, 20] spanning computer vision, layout understanding, natural language understanding, and information retrieval. Document AI techniques that combine Optical Character Recognition (OCR) [4, 2, 1] with graph neural networks [33, 34, 18, 29, 19] have proven to be effective at extracting structured information from documents images, especially for well-formatted printed documents with tables and headers such as invoices. However, these methods perform poorly on documents with uneven layout and handwritten content, such as medical prescriptions. Recent models such as LayoutLM [32, 31] that jointly

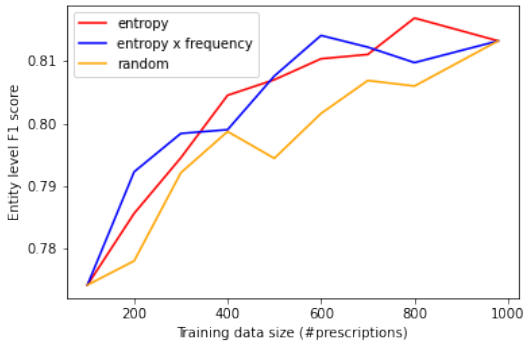



Figure 8: Plot of test entity level F1 score with model trained using different data selection strategy and data volume. Curves represent data selection strategies based on a) class posteriors entropy, b) product of entropy and normalized frequency and c) random sampling

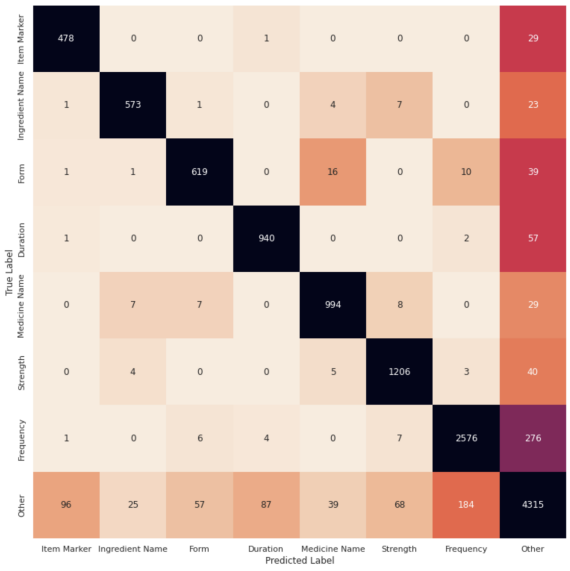



Figure 9: Confusion matrix showing detailed error reports.

learn the layout, visual, and text signal from a large corpus of document images improve performance with uneven layout. Handwritten text recognition (HTR) remains an open challenge despite advances in multi-dimensional RNNs and transformer models [9, 12, 17] due to the variability in author style and limited supervision. Incorporating domain-specific language models is, thus, critical for domain-specific HTR. We combine ideas from OCR, LayoutLM, and domain catalog-based matching to create a tailored solution for our application.

**Information extraction** techniques [29, 21, 13] that deal with conversion of unstructured text to structured form, especially forming blocks of interest comprising lists of multi-attribute records are directly relevant to our application. These methods primarily use coupled models for segmentation and attribute detection (i.e. entity recognition (ER)), based on conditional random fields in combination with semantic embeddings derived from seq2seq models such as BERT [10], Bi-LSTMs and require extensive labeled data. Since such supervision is limited in our scenario, we decouple segmentation and attribute extraction tasks, using simpler approaches for the former and exploring the SOTA ER techniques while incorporating ideas on exploiting ontologies [30].

**Prescription Digitization** has seen rising interest in recent years with standardization of health data resources [3, 14]. Most techniques [28, 24, 11, 16], however, fixate on the ER aspects assuming the input is an unstructured text sequence and present results on benchmark datasets [27, 13, 5] of printed clinical documents from Western marketplaces. These models are inadequate for unstructured prescriptions since these do not account for the extraction errors, layout signals, and the gaps in the vocabulary. Therefore, we focus on developing a holistic approach with raw noisy prescriptions as input.