ITCS 6156 - Project Proposal

Keywords Identification from Handwritten Doctor's Prescription

Team Members:

Name	Login	800#
Akanksha Bhalla	abhalla	801077998
Dhananjay Arora	darora2	801077164
Naman Manocha	nmanocha	801077765
Shubham Gupta	sgupta37	801081963

Introduction

Problem

As we all know that doctors have illegible handwriting and it is difficult for individuals from non-medical background to understand the disease and the medicines mentioned in the report. Most of the times, even pharmacists find it difficult to decipher the handwriting written in medical reports. This is the problem we have been observing from decades now and many have suffered due to this problem.

In the modern era, few doctors have started to provide digitized prescriptions to maintain records, but the majority of doctors still provide traditionally handwritten prescriptions on their printed letterhead. This is one of the main reasons we want to work on this problem.

There are many existing models which convert handwritten images to digital text, but our aim is to go an extra mile and identify the keywords or medical terms from the prescriptions such as the disease, medicines prescribed, medical tests suggested, etc.

Motivation

Many people face a problem in identifying or understanding keywords written by doctors in their prescriptions and end up purchasing wrong medicines. One of us also faced a similar problem. Naman got a prescription from the doctor when he was suffering from pain in his abdominal area. Even though he got the medication he still didn't know what it was for. On his next visit to the doctor, he questioned why he was still having pain. The doctor said that he had mentioned in his prescription that he was suffering from a kidney stone and that might take more time for him to recover. Similarly, many people face such problems, and by building this model we are trying to overcome such issues.

Dataset

We have got the IAM Handwriting Dataset for the initial training of our model. This dataset is very large, and it consists of images of handwritten text segregated by words and sentences which we will be using. We are looking up for the specific dataset related to medical reporting. One of the literature that we surveyed during our initial analysis talked about research in this area. We have reached out to the author of the research paper to provide the dataset and we should be getting it shortly.

Literature Survey:

Paper: Handwritten Text Recognition using Deep Learning

This paper classifies handwritten word and converts them to digital text. They try to accomplish the task using 2 approaches, one with classifying words directly, for which Convolutional Neural Network (CNN) was used and another approach based on character segmentation for which Long Short-Term memory was used to construct boundary boxes for each character.

They use Hidden Markov Model for the task of OCR which results in high accuracy for the OCR. A CNN with two convolution layers and two average pooling layers and the fully connected layer were used to classify. The architecture followed in this paper assumes that input images will be of the same size. To achieve this task, they pad the image with whitespace to the maximum width and height of the dataset.

Another important aspect mentioned by them was that some images of the words in the dataset was tilted. They decided to rotate the image to the right slightly with random probability before training. They then tried for the character level classification and further run various experiments for both approaches and with a different combination of hyperparameters.

Finally, they conclude that character level classification was more successful and mentioned that due to lack of enough data for their problem and their model suffers from segmenting cursive characters.

Paper: Offline Handwritten English Character Recognition based on Convolutional Neural Network

This paper applied LeNet-5 CNN model with special settings for several neurons in each layer on UNIPEN dataset. The author of this paper has described his self-developed algorithm for training based on reinforcement of error samples, which they called Error-Samples-Reinforcement-Learning (ESRL) algorithm.

The CNN Model used in this paper is modified LeNet-5 with various modifications to tradeoff time-cost and recognition performance. CNN with 32 X 32 and 28 X 28 normalization lose information more quickly than 20 X 20 normalization. For preprocessing the UNIPEN Dataset, they used the DDA method and anti-aliasing. LeNet-5 with ESRL algorithm had the ability to reject illegal samples for printed character recognition. Hamming Distance was used to compare with rejecting distance.

Paper: Support Vector Machine for English handwritten character recognition

This paper proposed the Freeman Chain Code as the representation technique for image character recognition. This code gives the boundary of a character image and the location of the next pixel. After that, the feature vector is built. Finally, the Support Vector Machine (SVM) is used to classify the text. Dataset was taken from NIST (National Institute of Standards and Technology).

The proposed model starts with pre-processing. It involves process like thinning which produces a skeleton of a character. This character is fed to the next stage which is feature extraction. FCC is generated from characters which are used as features for classification. It traverses through each pixel of the character. The heuristic function is used to generate FCC to represent characters correctly and in the final step, SVM classifiers help to build the recognition model. Test results showed high accuracy at that time of research.

Paper: Keyword spotting in doctor's handwriting on medical prescriptions

There are many significant impacts in this piece of proposed work, for example, this work can be used in developing expert diagnostic systems, in extracting information from patient history, in detecting wrong medication and in making different statistical analyses

of the medicines prescribed by the doctors. In order to extract the information from such document images, domain-specific knowledge of doctors was firstly extracted by identifying department names from the printed text that appears in the letterhead part. Then the specialty/expertise of doctors was understood by this letterhead text which helped to search only relevant prescription documents for word spotting in the handwritten portion.

The word spotting in letterhead part as well as in handwritten part has been performed using the Hidden Markov Model in this paper. The objective of this paper was aimed at the automatic transcription of the handwritten prescription with the help of keyword spotting.

Firstly, this paper proposes an approach by incorporating domain knowledge into the text retrieval process. From the printed letterhead part of the prescription, attributes are retrieved, regarding the doctor specifications, with the help of word spotting in printed text part. The doctor's domain thus extracted is indexed with the search for disease/medicine keywords.

It utilizes a handwritten word spotting methodology to search the queried keywords in the written text portion of the relevant prescription. A Tandem-HMM system is proposed for word spotting. A PHOG feature was used for an efficient word recognition system. The proposed system in this paper attempts to eradicate the problems that a patient faces while reading a prescription because of poor handwriting of doctors.

Paper: A New Approach to Information Retrieval based on Keyword Spotting from Handwritten Medical Prescriptions

This paper mainly focuses on spotting words in the handwritten portion of the prescription using the Hidden Markov Model (HMM). They performed experiments on 500 different types of prescriptions. This paper also proposed medical knowledge for text recognition.

Their first part was to get details about doctors like department and designation from the printed part of doctor's letterhead with the help of OCR. Thus, extracted details from the printed part mapped with disease and medicine prescribed by the doctor. And for the offline word spotting in the prescription is done with the help of the Hidden Markov Model. HMM is used to model sequential dependencies. Also, the separate Gaussian Mixture Model (GMM) is defined for each state of the model. This paper has used word level classification using the HMM approach.

Method

Our idea is to come with a model which can scan through a handwritten medical report or prescription and convert it into digital text and be able to identify the medical keywords related to patient's disease or medicines prescribed.

After the careful review of the research done in this area, we plan to implement this model using Convolutional Neural Networks (CNN), Long short-term memory (LSTM), Hidden Markov model (HMM), Optical Character Recognition (OCR). These are the methods that we found in the above-mentioned literature. We will be trying various model combinations by using them and will try to come up with something new out of these according to the results we achieve after running experiments. We will get more clarity on these topics once we learn them in depth during class and thus figure out what will be best for our model.

We will approach this problem towards its solution in the following path:

1. Data Preprocessing & Visualization of data for better understanding
2. Training the model on IAM Dataset for Handwriting Training
3. Training the model with medical reports or handwritten prescriptions
4. Trying various models to find the best fit for this problem
5. Enhancements and Improvements based on experimentation.

Timeline

Milestones/Tasks	**Estimated Bandwidth**
Literature Survey	02/01 - 02/25
Project Proposal	02/26 - 03/02
Dataset Collection	03/01 - 03/06
Preprocessing and Visualization	03/11 - 03/15
Training & Testing the model with IAM dataset	03/15 - 03/22
Training the medical data	03/23 - 03/29

Mid Term Report	03/30 - 04/01
Building Various Models using TensorFlow	04/02 - 04/12
Testing the model performance and interpret results	04/12 - 04/15
Find ways to improve the model	04/16 - 04/25
Final Written Report and Poster preparation	04/26 - 05/03

Note: Everyone will contribute towards every task.

Difference

Most of the literature we read discussed the handwritten data to digital text conversion using various Deep Learning methods. There is not much research done in the medical domain, especially with handwritten medical reports. The research work that we found in this domain was mainly done using Hidden Markov model (HMM) and Optical Character Recognition (OCR) but none of them mentioned or used Deep Learning to a considerable extent. We aim to solve this problem by using existing methods and involve deep learning to achieve more accuracy. The proposed system attempts to eradicate the problems that a patient faces while reading a prescription because of poor handwriting of doctors.

References

[1] A. Mukhejee, A. Halder, S. Nath, S. Sarkar. "A New Approach to Information Retrieval based on Keyword Spotting from Handwritten Medical Prescriptions", Advances in Industrial Engineering and Management, American Scientific Publishers, Vol. 6, No. 2 (2017),90-96,
https://pdfs.semanticscholar.org/0af8/ec347b87569e3f68da04891fafce0c3f1fc8.pdf

[2] P. Roya, A. Bhunia, A. Das, P. Dhar, U. Pal "Expert Systems With Applications", Elsevier, Volume 76 (2017), Pages 113-128, https://doi.org/10.1016/j.eswa.2017.01.027

[3] D. Nasien, H. Haron, S. Yuhaniz "Support Vector Machine (SVM) for English Handwritten Character Recognition", 2010 Second International Conference on Computer Engineering and Applications, IEEE COMPUTER SOC (2010), https://ieeexplore.ieee.org/document/5445830

[4] B. Balci, D. Saadati, D. Shiferaw "Handwritten Text Recognition using Deep Learning" (2017), http://cs231n.stanford.edu/reports/2017/pdfs/810.pdf

[5] A. Yuan, G. Bai, L. Jiao, Y. Liu "Offline handwritten English character recognition based on the convolutional neural network", IEEE COMPUTER SOC (2012), 10th IAPR International Workshop on Document Analysis Systems, https://ieeexplore.ieee.org/document/6195348