MediReport

San Jose State University FA19 CMPE-272-01 Group 14 Project

Vini Aswal
Computer Engineering Department
San Jose State University
vijendersingh.aswal@sjsu.edu

Sakshi Mahendru
Computer Engineering Department
San Jose State University
sakshi.mahendru@sjsu.edu

Anjana Menon Cherubala
Computer Engineering Department
San Jose State University
anjana.menoncherubala@sjsu.edu

Animesh Swain
Computer Engineering Department
San Jose State University
animesh.swain@sjsu.edu

***Abstract*—An application that enables end-users who have do not possess adequate medical literacy, to interpret the meanings of various medical/physiological parameters that are present in their medical reports or prescriptions, in an easy to understand and informative format which can be presented in multiple languages.**

***Index Terms*—Aspose,React,Java,RDF database,NLP/NLG libraries,token parsing,spaCy,DBpedia,SPARQL.**

I. Introduction

Everyone cannot read or understand the various medical jargon and parameters mentioned in their medical reports (lipid profile, CBC report, LFT, RFT, etc.). They usually have to take a day off from their schedules and visit their doctor/medical practitioner to get a sense of their annual health reports or even results of important tests. Our solution will save people time and money for a doctor's visit, by making an informed decision about their health using their medical reports. This will be useful, particularly in the case where patients are located in remote areas and where medical facilities are sparse. We have built a React-based UI, where the users can upload their medical reports. React Native has been used to implement the solution, which enables the application to be ported to mobile platforms as well. Tesseract OCR and Aspose APIs have been used to extract medical report data while scanning the report. We will store RDF medical data from DBpedia and other data sets in MarkLogic. We will also be using various NLP/NLG libraries such as Apache OpenNLP, SimpleNLP, etc., to generate a human-like response for the user. Our Java-based system will interface with the RDF database and NLP/NLG libraries to send back repose to the UI.

Multi-language support has been included for multiple space-delimited languages such as English, Spanish, and all Indian languages. Languages such as Chinese, Japanese, Korean, Thai, Khmer cannot be supported since their writing systems do not use spaces and OpenNLP performs space-level token parsing.

II. Business Value

Limited literacy skills are one of the strongest predictors of poor health outcomes for patients. Studies have shown that when patients have low reading fluency, they know less about their chronic diseases and they are worse at managing their care. Overall, studies of patient-accessible medical records suggest modest improvements in doctor-patient communication, adherence, patient empowerment, and patient education.Present health care system relies on the assumption that patients can understand complex written information presented to them in their reports.If people do not understand health information, they are less likely to take necessary actions for their health or make appropriate health decisions.With MediReport, the aim is to bring down the heavy reliance of people on their health care practitioners for common health problems leading to reduced medical bills. Reduction in the time involved in taking appropriate health decisions is also one value add of the solution proposed.

III. Use-case scenarios

A. Patient trying to understand a medical report :

Patients often have difficulties understanding the clinical data presented in portals. In response, increasingly, patients either ignore their reports or go online to make sense of this data. The medical information provided online in forums and discussion groups can lead to patient anxiety and such information may not always be applicable to that particular individual. MediReport will give the user a one-stop solution for patients to understand their medical reports, the meaning and impact of each term (for ex., Bilirubin, Creatinine, etc.) , the urgency of their medical problem and also suggest ways to manage the symptoms, if applicable.

B. Patients trying to understand a doctor's prescription :

40-80 percent of the medical information provided by healthcare practitioners is forgotten immediately. The greater the amount of information presented, the lower the proportion correctly recalled; Furthermore, almost half of the information that is remembered is incorrect. To help patients in recalling and understanding each prescribed medicine, MediReport will augment the prescription with explanation, categorization (antibiotic, antibacterial, etc.) side-effects, medical usage, mode of action, etc.

C. Physicians trying to make a diagnosis :

A study by Meyer and Payne suggests that the association between physicians' diagnostic accuracy and their confidence in that accuracy may be poor and that physicians may not request the required additional resources (ie, additional tests, second opinions, curbside consultations, referrals, and reference materials) to facilitate diagnosis when they most need it. These mismatched associations might prevent physicians from reexamining difficult cases when their diagnosis is incorrect. Improving these associations and the use of potential resources in handling difficult cases could potentially reduce diagnostic error. MediReport will help reduce diagnostic error by providing the physician the additional resources such as Signs and symptoms, Virology, Pathophysiology, Diagnosis, Prevention, Treatment, Management, Prognosis, etc.

D. Insurance company validating a claim :

The insurers know a lot about you, based on claims. They aggregate data, such as imaging, medications, referrals, admissions, and emergency department visits, as well as quality metrics around severity-adjusted episodes of care for specific diagnoses. MediReport will help insurance companies in making decisions about pre-existing conditions, valid claims, reporting malpractices, etc. by providing them with a clear understanding of the thousands of medical terms and jargon that can be difficult to remember. It will save insurance companies a lot of money by reducing work hours in understanding medical cases and well as reduce dependence on consultants that need to be paid high salaries.

IV. Implementation

A. Front-end

React Native has been used to create a single page cross-platform app that can be run on iOS, Android and Web with a Material UI. React Native is a multi-platform solution developed by Facebook that lets you build mobile apps using JavaScript. These mobile apps are considered multi-platform because they're written once and deployed across many platforms, like Android, iOS and the web.

Passport.js, which is authentication middle-ware for Node.js which generates JSON Web Tokens has been implemented for user authentication.

node.bcrypt.js library has been used to hash user passwords and store in the Mongo DB.

B. Node.js back-end

Node.js is an open-source, cross-platform, JavaScript [Library] that executes JavaScript code outside of a browser. The Node.js based back-end which contains Express based APIs such as login, sign-up, get PDF results, etc. It acts as a producer and routes its requests to the appropriate Kafka topic. The back-end communicates with Kafka to achieve scalability by queuing API requests. The Kafka based back-end, which contains code to perform actual operations such as using Passport/JWT based authentication,PDF parsing,parsed PDF storage etc. PDF is stored via Multer in back-end and is parsed using the node library pdf2json.

C. Python back-end

The Python based backend has a Klien server which accepts the parsed PDF. After using spaCy and ScispaCy's models for processing biomedical, scientific or clinical text to perform NER, we get medical entities. The named entities are searched in DBpedia using dynamically generated SPARQL and the result is returned to the UI.

D. Cloud deployment

The application has been divided into micro services namely front-End, back-End, Python back-end and Kafka back-end. Each of these functions has been containerized and is running within separate containers. All the containers are encapsulated within a single AWS EC2 instance. Since all of the containers belong to the same Virtual Private Cloud and by extension, the same subnet, inter-container communication is permitted amongst all the containers. Since the MediReport app is an CA model, the two axes of the AKF scaling cube are covered here, namely the X and Y axes. The Scale Cube is a model for defining micro services and scaling technology products. AKF Partners invented the Scale Cube in 2007, originally publishing it online in 2007.

X axis scaling: Cloning/Replication - In order to ensure that the application is able to handle heavy traffic loads and reliability, a cluster of 3 Docker instances has been deployed. A network load balancer performs the task of routing the incoming requests to the application to the appropriate primary

docker instance. Also the docker instances run in the private subnet space whereas the load balancer is in the public subnet space. This ensures all that any external traffic has only one single entry point via the load balancer into the underlying system. This is a vital feature in the case of preservation of patient medical reports and records.

Y axis scaling: Split dissimilar things – Scaling along the Y axes composes of functional decomposition. This is inherent by the design of this application and future iterations would allow each of these micro services to be scaled out and improved upon at different rates.

E. Security vulnerabilities testing

The OWASP Zap (Open Web Application Security Project Zed Attack Proxy) security tool has been used to check for security vulnerabilities of the web application.

It allows web application security testers to perform fuzzing, scripting, spidering, and proxying in order to attack web apps.

V. Named Entity Recognition (NER)

Named-entity recognition (NER) (also known as entity identification, entity chunking and entity extraction) is a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.

Notable NER platforms include:

- GATE supports NER across many languages and domains out of the box, usable via a graphical interface and a Java API.
- OpenNLP includes rule-based and statistical named-entity recognition.
- spaCy features fast statistical NER as well as an open-source named-entity visualizer.

For our application we have chosen to use the spaCy NER platform because it has been trained on the OntoNotes 5 corpus and it has an advantage over other platforms as we only need to apply NLP once, the entire background pipeline will return the objects/tokens.For processing biomedical, scientific or clinical text, we have used the ScispaCy NER models.

A. spaCy NER :

spaCy is an open-source software library for advanced natural language processing, written in the programming languages Python and Cython. The library is published under the MIT license and currently offers statistical neural network models for English, German, Spanish, Portuguese, French, Italian, Dutch and multi-language NER, as well as tokenization for various other languages.

B. Sci spaCy NER model :

This repository contains custom pipes and models related to using spaCy for scientific documents.In particular, there is a custom tokenizer that adds tokenization rules on top of spaCy's rule-based tokenizer, a POS tagger and syntactic parser trained on biomedical data and an entity span detection model. Separately, there are also NER models for more specific tasks.

VI. Personalisation

A. Multiple language support :

MediReport supports personalised results in the following languages:

- English
- Arabic
- Spanish/Español
- French/Le français
- German/Deutsch
- Chinese
- Italian/Italiano
- Japanese
- Dutch/Nederlands
- Polish/Polski
- Portuguese/Português
- Russian

B. Age and gender based recommendations :

Data from Institute for Health Metrics and Evaluation (IHME), an independent population health research centre at University of Washington, Medicine has been used to provide age and gender-based disease recommendations. It also provides data on food items that have an impact on the medical condition.

C. NLG personalised message for user :

GPT-2, a large transformer-based language model trained on a dataset of 8 million web pages has been used to provide personalised messages to the user.

ACKNOWLEDGMENT

The project team collectively wants to thank our mentor and instructor, Prof. Rakesh Ranjan, Director of Emerging Technologies at IBM Data AI, for directing our efforts in this project for selection of technology stack, market research, design thinking process, development and deployment of MediReport. His guidance, motivation and industry knowledge has been pivotal in making this project a success. We would also like to express gratitude to everyone who directly or indirectly contributed and the Computer Engineering Department at San Jose State University for allowing us the opportunity to work on this project.

REFERENCES

[1] Graham, S., Brookey, J. (2008). Do patients understand?. The Permanente journal, 12(3), 67–69. doi:10.7812/tpp/07-144.

[2] Ross, S. E., Lin, C. T. (2003). The effects of promoting patient access to medical records: a review. Journal of the American Medical Informatics Association : JAMIA, 10(2), 129–138. doi:10.1197/jamia.m1147.

[3] Reynolds, T. L., Ali, N., McGregor, E., O'Brien, T., Longhurst, C., Rosenberg, A. L., ... Zheng, K. (2018). Understanding Patient Questions about their Medical Records in an Online Health Forum: Opportunity for Patient Portal Design. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2017, 1468–1477.

[4] McGuire LC. Remembering what the doctor said: organization and older adults' memory for medical information. Exp Aging Res 1996;22: 403-28.

[5] Anderson JL, Dodman S, Kopelman M, Fleming A. Patient information recall in a rheumatology clinic. Rheumatol Rehabil 1979;18: 245-55.

[6] Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' Diagnostic Accuracy, Confidence, and Resource Requests: A Vignette Study. JAMA Intern Med. 2013;173(21):1952–1958. doi:10.1001/jamainternmed.2013.10081.

[7] Kaufman J. M. (2015). How to work with insurance companies. Neurology. Clinical practice, 5(5), 448–453. doi:10.1212/CPJ.0000000000000179.

[8] Facts Figures https://spacy.io/usage/facts-figures.

[9] Bizer, Christian; Heath, Tom; Berners-Lee, Tim (2009). "Linked Data – The Story So Far" (PDF). International Journal on Semantic Web and Information Systems. 5 (3). doi:10.4018/jswis.2009081901.

[10] Create Documents with Aspose.Pdf for .NET https://visualstudiomagazine.com/articles/2010/09/01/create-documents-with-asposepdf-for-net.aspx .

[11] Getting Started With MarkLogic Server https://docs.marklogic.com/guide/getting-started/intro .

[12] Why MarkLogic Will Lead the Next-Generation of Database Technology https://www.marklogic.com/blog/marklogic-will-lead-next-generation-multi-model-database/ .

[13] https://github.com/allenai/scispacy .