MACHINE LEARNING BASED MEDICAL INFORMATION RETRIEVAL SYSTEMS

by

Akhil Gudivada

Febuary, 2019

Director of Thesis: Dr. Nasseh Tabrizi

Major Department: Computer Science

As many fields progress with the assistance of cognitive computing, the field of health care is also adapting, providing many benefits to all users. However, advancements in this area are hindered by several challenges such as the void between user queries and the knowledge base, query mismatches, and range of domain knowledge in users. In this research, we explore existing methodologies as well as look into existing real-life applications that are used in the medical field today. We also look into specific challenges and techniques that can be used to overcome these barriers, specifically related to cognitive computing in the medical domain. Future information retrieval (IR) models that can be tailored specifically for medically intensive applications which can handle large amounts of data are explored as well. The purpose of this work is to give the reader an in-depth understanding of artificial intelligence being used in the medical field today, as well as future possibilities in the domain. The models and techniques designed and discussed in this research can help provide a framework, or starting point for those interested in effectively developing, maintaining, and using these models to help improve the quality of health-care. Furthermore, we explore the development process of such a model and discuss the steps including data collection, processing, model creation, and also improvement.

MACHINE LEARNING BASED MEDICAL INFORMATION RETRIEVAL SYSTEMS

A Thesis

presented to the faculty of the department of Computer Science

East Carolina University

In fulfillment of the requirements for the degree:

Master of Science in Computer Science

by

Akhil Gudivada

Febuary, 2019

©Akhil Gudivada, 2019

MACHINE LEARNING BASED MEDICAL INFORMATION RETRIEVAL SYSTEMS

by

Akhil Gudivada

APPROVED BY:

DIRECTOR OF THESIS: _____

Dr. Nasseh Tabrizi

COMMITTEE MEMBER: _____

Dr. Qin Ding

COMMITTEE MEMBER: _____

Dr. Rui Wu

CHAIR OF THE DEPARTMENT

OF COMPUTER SCIENCE: _____ Venkat Gudivada, PhD

DEAN OF THE

GRADUATE SCHOOL: _____ Paul J. Gemperline, PhD

Acknowledgements

I would like to thank Dr. Tabrizi, my advisor for all his hard work and dedication in pushing this thesis forward and encouraging me throughout the process as well as being a reliable and experienced source of professional advice. I would also like to thank IBM for allowing us to use their Watson Discovery services which served as a crucial tool in the development of this work.

Table of Contents

Chapter 1: Introduction

Through the advancement of the technology age, many aspects of life have become intertwined with machines. Likewise, medical-related information is also making this transition, making properly retrieving medical data a priority as well as a challenge. This thesis serves as an entrance into exploring the methodologies and tools which serve as the buffer which brings intelligent medical information to the user. While the concept may sound rather simple, many challenges such as ambiguous terminology, diversity of information, and specialized domain knowledge exist [1].

Today, cognitive computing is used more than ever in the field of health care. The partnership between artificial intelligence and human beings with the goal of transforming health care on a global scale is no longer a fantasy. Medical data is available in surplus, and much of it is even underutilized such as data from fitness trackers and mobile apps. The potential for new applications to develop by using the untapped, existent data presents an era to make unprecedented leaps in the medical domain.

In addition, medical data is not always available to the patients. Dealing with this data not only makes the topic a cognitive computing issue, but the vast amount of data that needs to be processed also makes it a big data issue. By the year 2020, the amount of medical data that is available will double every 73 days, and up to 80% of this data will remain unstructured [2]. Even if medical professionals have access to this data, there is simply no way for all of it to be analyzed and processed in a meaningful and beneficial way towards patients without the help of artificial intelligence. With the advance of system hardware, the efficiency of data processing and the reemergence of machine learning, cognitive computing

is affecting nearly every field. For example, automobiles are able to navigate independently, businesses learn about their customers better to more efficiently target them, and investors are employing it to make predictions and improve their research analysis. However, due to the reasons mentioned, no field is likely to be as disrupted in the near future by cognitive computing as health care [1].

Many applications and platforms to process large amounts of information exist, however very few are being used in the field of cognitive computing based medicine. Tools such as Mongo DB, Hadoop, and other new age databases are equipped to handle the needs in this field. Platforms such as Microsoft Azure, IBM Watson, and Google DeepMind exist and are able to process large amounts of information while running machine learning based algorithms to pick up key points of the data, sentiments, and even summaries [1].

With so much potential in this area, this research serves to both explore the current state of cognitive computing technologies in the field of medicine, while also building a potential model using some of the existing technologies to be used for practical purposes. First we explore previous and current systems, the applications which use them including their purposes. Then we dive into building a model which can be used practically.

Chapter 2: Literature Survey

2.1 Current IR Systems

2.1.1 Overview

As access to data has become more and more prevalent due to the reduced cost of storing and hosting data, the need for effective IR techniques has never been higher. Moreover, specific systems may have different requirements for each application. A medical application, as discussed further later on, has significantly different demands than a quantitative system for banking or trading stocks due to the domain containing its own language [3]. A document based IR system contains three sub-units: queries, document representations, and the algorithms which match the two. Although each IR system's architecture may vary, and some may even be unknown to the public for security or financial reasons, the general structure is displayed in Figure 2.1.

A document is simply data stored in a file that contain topics of interest. A corpus is the total collection of these documents. These documents are transformed into a document representation either automatically or manually to ensure matching these with queries becomes easier. Each document representation should reflect the author's intention, which is where natural language processing (NLP) and context-based models become important. Once these are properly indexed, the user's need (in the form of a query), is matched with the document representations and returned to the user in a meaningful way, such as ranked based off relevance. These models can then be tested to have their performance analyzed by calculating *precision* and *accuracy*. precision is simply the percentage of retrieved documents which were deemed relevant whereas accuracy is the percent of relevant documents retrieved

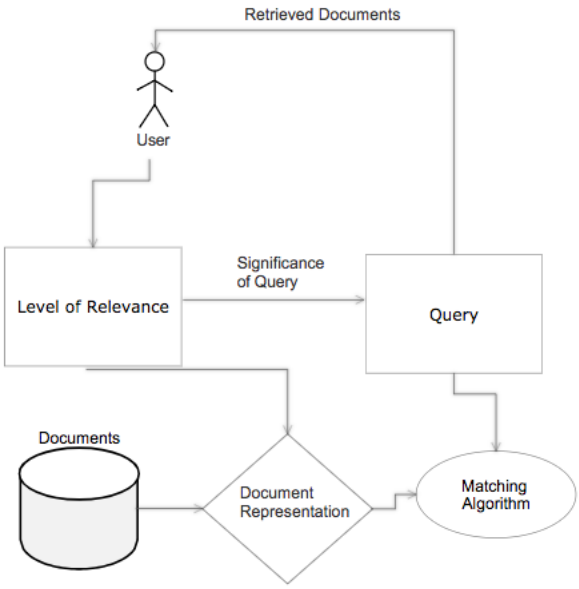

Figure 2.1: Outline of basic IR model

by the system. Although these two measurements typically have an adverse effect on each other, both are often used to analyze IR systems [4]. IR models can be broken down into three categories:

- Probabilistic models
- Knowledge-based models
- Learning systems based models

Probabilistic IR models are based on estimating a value, or probability of relevance of a document to the user based on the user-provided query. Generally, the relevance feedback from the user is utilized to determine the probability of similar documents being relevant or not relevant to the user. Different learning strategies are used in this model. Estimation of probabilities for relevance is performed on a set of sample documents or queries and extended to all the queries and documents. Inference networks use a network which captures probabilistic dependencies among nodes in the network.

Knowledge based system focus on modeling in two areas. First, the system tries to generate a model which takes into account an expert's domain knowledge for that specific

field. Therefore, these systems need to be adapted to a specific domain, and end up being better suited to handle the nuances of any specific domain. An example of this model is the Unified Medical Language System (UMLS) which will be discussed in detail in the *Applications* section.

Learning systems based approaches use algorithmic extraction of data or identifying underlying patterns in data. These include approaches such as symbolic learning, neural networks, and evolution based algorithms. In symbolic learning, discovery is attained through inductive learning while creating a hierarchical structure and producing If-Then rules and concepts [4]. The NeuroDiary application discussed later uses this approach.

2.1.2 Domain Specific Query Expansion

In order to address the challenges presented for retrieving medical data, an enhanced IR model needs to be used which takes into account the challenges and offers solutions for them. Various versions of knowledge-based query expansion techniques can be used to mitigate these issues [1]. This custom system needs to expand the original query by incorporating new terms from the initially retrieved documents and add context to the query. After this process is done, the system needs to capture relevant terms and semantic relations between terms by extracting them from knowledge bases. Not only does this system provide linear relationships between terms, but adds a tertiary level relationship by expansion. This proposed version of the model can help fix the incompleteness of medical information that is stored. The process itself, called tensor factorization, is widely used for problems related to link prediction [5]. In addition, this model, shown in Figure 2.2, also increases the ability of this system to work well under sparse settings such as recommender systems. With an expanded set of terms, the probability to find and retrieve documents which are relevant increases, even if the query directly doesn't have the appropriate terms to make a match match. Moreover, this proposed system is able to be properly integrated with existing IR systems.

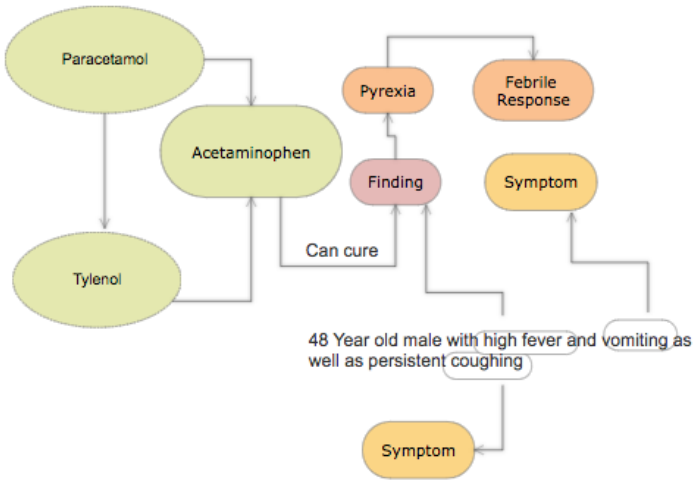

Figure 2.2: An example of a medical query being segmented

2.1.3 Cognitive IR Technologies

Due to the large amount of data that needs to be properly processed, as discussed in the introduction section, existing technologies such as ones from IBM exist which can help solve the issues this research focuses on. As many people may be aware, IBM solved the long-standing problem of answering multi-domain questions in the show *Jeopardy!*. IBM created an incomprehensibly massive parallel, probabilistic, evidence based model which is well known as *Watson* which launched in 2010 [2]. *Watson* overcame a large challenge in computer science by illustrating the closing void of natural language content with the integration of NLP, IR, machine learning, deep learning, knowledge representation and reasoning, as well as massively parallel computation. This project related all those ideas and merged it with open-domain, automatic, question-answer technology which not only performs well, but outperforms top human minds by a significant margin. Gaining notoriety from its success on *Jeopardy!*, IBM customized and delivered this technology through the cloud, to the public. With the algorithms already in place, IBM's cognitive machines gain evidence-based knowledge over time and also improves performance with each interaction [2]. Although specific implementation details may be protected by IBM, these technologies can still be incorpo-

rated with many other projects fairly easily to handle medical data. For example, in 2015 IBM health cloud was launched which allows information fed into the system to be identified, shared, and combined with a dynamically growing view of data which can be presented to the health care provider instantly. Currently, these technologies are being used to match patients to relevant clinical trials and even DNA insights are being examined for personal treatment. The goal in this project is to incorporate these existing technologies for natural language processing as well as training the model which can intake a large amount of data and properly train, process, test, and make accurate predictions based on this data.

2.1.4 Databases

Although medical applications may need to be specified to have access to their own personalized databases, some medical libraries exist and are used with high regularity for both computer scientists as well as medical professionals. Common public repositories for medical documents include [6]:

- Pubmed - literature
- GEO - genomics, proteomics and clinical data
- ArrayExpress- genomics, proteomics and clinical data
- miRGen - database of miRNA targets and relationships
- miRBase - database for miRNA sequences and annotations

Through these databases, the data can be broken down into levels of data based on what they pertain to. The data stored either falls in to the following categories: organism, organ layer, tissue layer, cellular level, and sub-cellular level. In the organism layer, data pertains to the organism as a whole along with a group of organs. This data is typically related to clinical data which comes from the hospital database manager. The organ level data comes from a pathologist and deals with issues that pertain to an entire organ such as lung cancer, hemochromatosis, hepatitis, and myocardial infarctions. Tissue level data is similar and generally comes from a pathologist as well. Cellular data arises from lab technicians who

work with data pertaining to individual cells or a group of cells such as cell count, cellular lysis, and cancer related issues. Sub-cellular data is composed by the structures of the cells and are usually provided by a data analyst who performs ontology analyses [6].

Two popular databases which contain documents in the medical field: Medline and Pubmed, offer over 23 million journals and receives over 2.7 billion queries per year [7]. According to a study from conducted by Susannah Fox, it is found that 80% of internet users in the united states searched for health related information online [8]. However, finding relevant medical information is difficult for both common people as well as specialists. Issues like vocabulary mismatches create a void in the IR system. For example, when a user has a term in their query such as *high blood pressure*, a relevant document may not come back as a match simply because it was written in a technical manner which contained the term *hypertension*. Although these issues exist in regular IR models, it becomes more significant in medical IR simply because many users may not be familiar with technical medical terms [7].

2.2 Challenges

Extracting medical data itself is a challenge, however the complexity increases along with the specificity and ambiguity of medical languages. This difference causes vernacular mismatches between queries and documents, making traditional IR methods and machine learning algorithms often ineffective. These issues also affect classification, hierarchies, and term dependencies. For example, if a user enters *how to treat a headache*, a document that contains the term *acetaminophen* may not be flagged as relevant for the machine, but obviously is from the standpoint of a physician or even a regular human. The issue in this example is that the model has to be trained with the proper domain knowledge to associate the terms *acetaminophen*, *may treat*, and *headaches*. Moreover such systems need to be able to cater to both people with and without the medical terminology background. The users could be physicians or patients, therefore technical terms need to mesh well with common, everyday

terms used. It is proven by several studies conducted that the proper incorporation of syntax and semantic information within the knowledge domain being put to use is critical to the performance of the IR system [9]. Some effective methods currently used in systems handling medical data including knowledge-based query expansion will be discussed later on.

In today's interconnected world, many different types of users could be using a medical IR system at different locations and need to collaborate. Most articles published on PubMed will prove this in the affiliation section itself. Most interdisciplinary domains that are medically related contain users with different backgrounds such as data-mining, mathematics, computer science, biology, statistics, IT, etc. Due to the heterogeneity of users, the IR system needs to strike a balance in order to be able to satiate the needs of these users. The system can neither be too generic nor too specific. Due to the diverse backgrounds of the users, some users may want predefined workflows while others might want to configure their own workflows via existing components or web-pages. In addition to heterogeneous users, the data they analyze is also heterogeneous. Often times, data is collected without knowledge of how exactly it may be analyzed and used, therefore, the data can be extremely unstructured, an issue that often arises when working with medical data. Moreover, the amount of data which is collected is extremely large due to the advances in computing power. With data coming from many locations such as hospitals, analysis offices, and even patients themselves, it is important to keep this in mind while designing systems which work with and display this data. [6]. Moreover, it is important to keep viewing options customizable for users as they may have different needs and may be from different backgrounds, some of which may be extremely technical, while others may not have any technical experience at all. Although these issues may exist to a certain extent in other fields, bio-informatics and any fields which combine computer science and medicine, experience the greatest level of diversity.

Multiple issues that need to be addressed still remain. Most databases containing medical knowledge remain unstructured and incomplete. This causes samples to be considered *noisy* and the inaccuracy of concept mapping sets higher requirements for the model to perform

as needed [10]. One study even showed that the precision of concept mapping through MetaMap was less than 72% [11]. In addition, observed data is sparse due to incompleteness of the knowledge bases and limited annotated samples. These challenges cause the need for a systematic method to address the issues present. This method needs to be handcrafted specifically to handle medical data.

Chapter 3: Current Cognitive Computing Applications

3.1 Applications

A collection of publications related to using cognitive computing technology for practical purposes in the medical domain were analyzed for the past 10 years. As visible, this is an up and coming field with artificial intelligence paving the way for medical advancements. With machine learning resurfacing in many domains, the combination of powerful hardware and efficient algorithms make it possible to integrate machines with medicine. While reviewing papers for this survey, only current applications that were either in use, or being developed for real-world applications in the medical field were explored. Hypothetical applications for future work were not considered. This survey extended both the computer science domain as well as the medical domain.

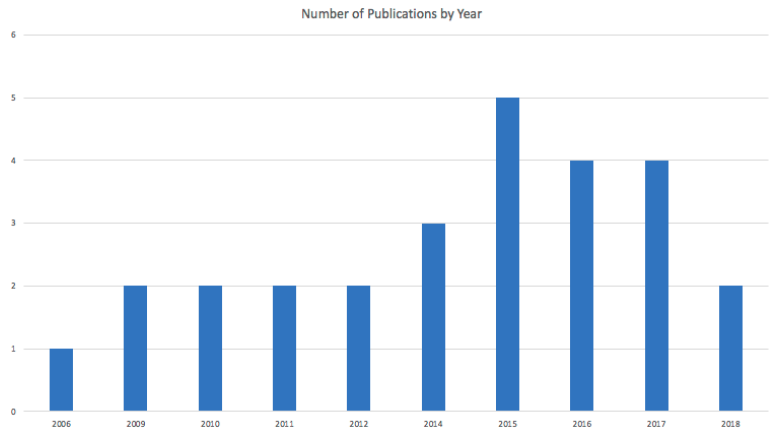

Figure 3.1: IEEE Published Papers on Cognitive Computing Application in the Medical Domain by Year

With more and more specific applications successfully integrating artificial intelligence into their tools, cognitive computing branches off deep into the medical domain making it

even more important for IR systems to properly function. As shown in 3.1, the amount of specific cognitive computing applications in the medical domain are shown by publication numbers by year. There is a clear upward trend with cognitive computing paving its way into the medical domain (as well as all others). While there is progress being made, the medical domain still holds limitless possibilities for practical uses of cognitive computing and machine learning.

3.1.1 EHR Conversion

Electronic Health Records (EHR) conversion from traditional methods of storing records is rapidly increasing. In addition, U.S. legislators passed the Health Information Technology for Economic and Clinical Health (HITECH) act which encourages the use of EHR by clinicians. Any application that health care providers use needs to be easy for them to utilize as well as be compliant with the specific clinical best practices as shown in Figure 3.2 [12].

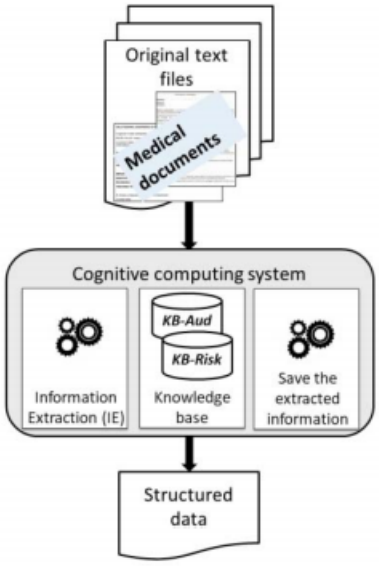

Figure 3.2: Structuring medical data

A proposed model for hearing impaired patients is proposed by Tognola et al. Structured from three modules, this model provides a real-life application for cognitive computing in the medical domain. The first step is to extract meaningful health information from medical

documents. Next, a knowledge base module is used to classify the extracted information. The last module simply saves this information into a secure location which can be revisited with ease. The information extraction stage of this application applies two techniques: regular expressions to retrieve information in regular text patterns, and MetaMap (MM) to extract any textual information that was not associated to regular expression patterns. In addition, MM extracts and codes medical text from medical literature into Unified Medical Language System (UMLS) which is currently used for analyzing clinical notes. For example, if this system receives the input of a sentence "The patient presents an inflammation in both ears", the term "inflammation in ear" would be coded to the system as the UMLS term with a unique internal token such as "C0029877" with a description: "Ear Inflammation". After this stage, the knowledge-base system takes the input and links it to a relation. For example, if the UMLS term returns "hearing disorder", then all pieces of information related to the input in the hierarchical structure will be returned to the clinician as illustrated in Figure 3.3 [12].

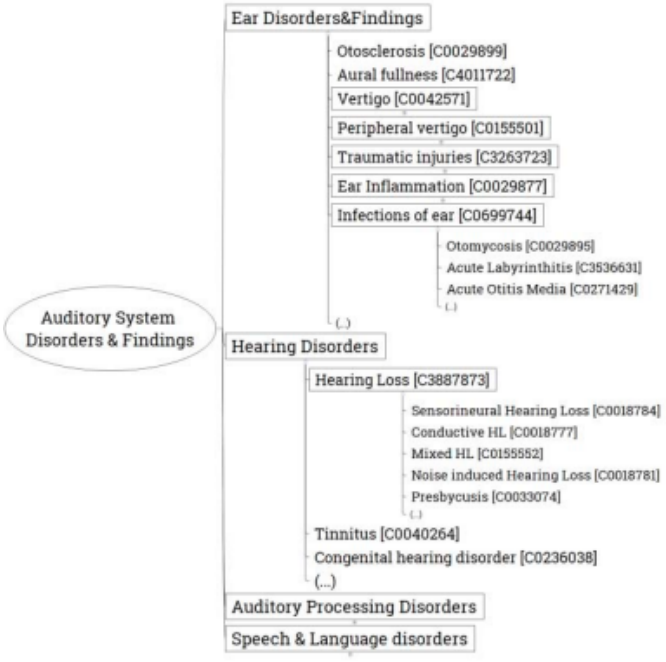

Figure 3.3: Hierarchical structure of the knowledge based portion

3.1.2 Neuro Diary

Very few cognitive computing medical applications are in use today, however it is a growing area. Another application called NeuroDiary exists in the prototype stage. This application which was built at Liverpool University, UK, helps perform 2 functions: collects data from patients diagnosed with hydrocephalus (buildup of cerebrospinal fluid in the head) as well as presents a visualization for the severity of the pain data across time to give the HCP a better understanding to help treat the condition. As displayed in Figure 3.4, Patients can log-in to the system and record instances of pain, headaches, and other symptoms related to the illness [13].

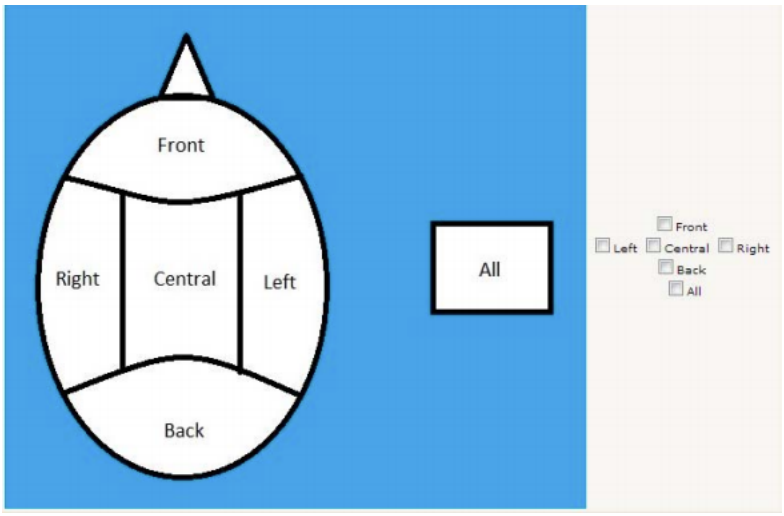

Figure 3.4: User input entry for NeuroDiary

In addition to inputting the location of the pain, the user can also note if the intensity on a scale of 1-10 as well as determine if it is a gradual or sudden change. The user also can input additional comments which will be tracked for the HCP. Currently, this system plots the pain intensity vs time as well as highlights any points of interest which show a significant change to the HCP. The proposed improvements to the existing system include: making larger data samples so that machine learning algorithms can help classify individual cases, improve the UI in order to display more precise location points which can even be turned into coordinates, and extending the application to mobile devices for extended use along with

two way communication in place between the user and the HCP [13]. Although individual services and applications may contain their own implementations, one common requirement for any application in the medical domain as stated in the introduction, is availability to a knowledge base data-set which contains specific information in the domain.

3.1.3 Descision Tree Model

Another model that is proposed by Papageorgiou states that the idea of a technique combining a decision tree using a Decision Tree Algorithm (DTA) such as ID3, with a Fuzzy Cognitive Map model (FCM). The FCM model is trained with an unsupervised learning algorithm in order to achieve improved accuracy. The Hebbian learning algorithm is used to train the FCM structure. In this model, if a large amount of input data is given, then the quantitative data is used to induce a decision tree and to construct the FCM. This allows flexibility to be maximized by fuzzification of strict decision tests which are simply derived if-then statements as displayed in Figure 3.5 [14]. Using this model, 92 cases of bladder

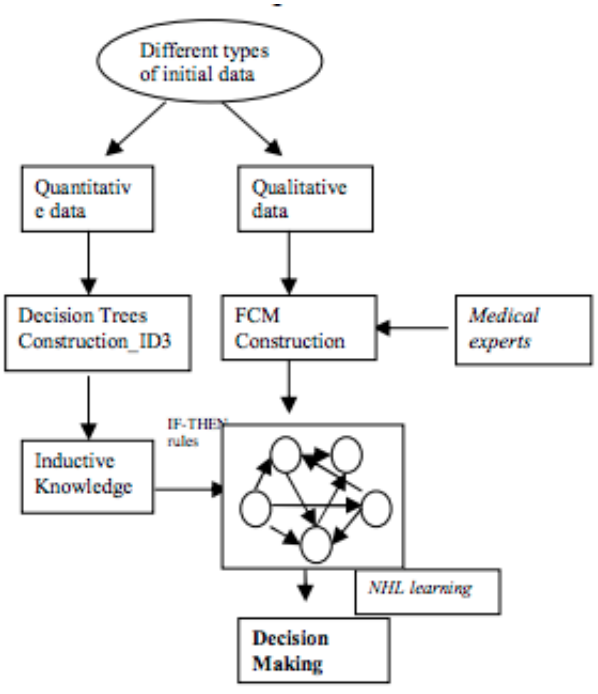

Figure 3.5: Decision tree model

cancer were collected from a hospital in Patras, Greece. Using this system, 63 cases were tagged as low-grade, and 29 as high-grade. Following this diagnosis, each tissue sample was evaluated retrospectively (with the diagnosis already predicted by the model). The FCM grading tool was implemented to provide distinct quantitative values for both the high and low-grade cases by using a simple Bayesian classifier for the output data. Sensitivity and specificity were calculated for this model in order to evaluate the performance. For the 92 samples, the results came out to 79% sensitivity and 87.5% specificity for the FCM grading tool [14].

3.1.4 Predicting Outcomes in Perinatal Medicine

As computer scientists and physicians continue to work together, the analysis and treatment of medical care will improve. In this application, artificial neural networks (ANN) allow the development of prediction models to make a diagnosis and develop a method for therapy. Perinatal care is the field of medicine that focuses on the period of time before a birth to ensure the proper growth of a fetus until one week after delivery (usually 28th week of gestation until 1 week post birth). The data collected during this period is both diverse due to the medical aspect as well as critical. The data collected for this application can be categorized into three groups: pre-birth, delivery type, and Apgar score [15]. pre-birth data is especially critical due to its direct correlation with infant mortality and increased risk of chronic health issues. Delivery type predicts whether the baby will be delivered through the traditional method or by C-section. This can help mothers prepare for care. The third type of data comes from the APgar score. Developed by Dr. Virginia Apgar, this test includes five attributes: appearance, pulse, reflex, muscle tone, and respiratory rate. This data is stored in a central location through the Parinatal Partnership Program of Eastern Ontario (PPPESO) who ensures the data is properly structured and accurate for all entries. The model is set up with a back-propagation feed-forward neural network with weight elimination cost function and the hyperbolic tangent transfer function. The training is done on two-thirds of the data

selected at random while the remaining portion is tested. A three later network is used with the cost function to prune the network which is also optimized for both sensitivity and specificity. The model was also manually tweaked for all the ANN parameters in order to obtain optimal results. This model, which is currently running around the clock, saves thousands of manual hours by making these predictions for the HCPS [15].

3.1.5 Medical Decision Support System for Detection of Downs syndrome

Using deep learning techniques, image features can be extracted using a range of computer models and data processing algorithms. Using these methods with existing IT infrastructure present in hospitals, Wojtowicz et al. developed a system to predict the presence of Down's syndrome in infants. By studying dermatoglyphics, the study of fingerprints, lines, mounts and shapes of hands, medical professionals are able to determine the occurrence of genetic disorders such as Down's in infants. Until now, the classification of these patterns are carried out by a professional anthropologist. However, now with the service of automatic pattern recognition, a machine can accurately predict genetic disorders. These results can be provided to clinics and hospitals where anthropologists may be limited or not staffed at all [16].

The process begins with the collection of data in a non-invasive manner using touch scanners or cameras. These images are sent to a database connected to a system which runs a tele-medical system through the internet. The data is then subject to detailed analysis and classification is carried out. The analysis of data containing specialized domain knowledge in the form of natural language, arithmetic expressions and logic relations formulate a set of conditions in order to determine the conclusion. In addition, the partial information from the images leads to creation of feature vectors which represent the characteristics of image patterns. Using these features, the probability of a genetic disorder being present in the individual can be determined. Since these results can be obtained remotely, this method overcomes the limitations that arise from the shortage of employees who are employed across

smaller medical centers, providing significant value [16].

Three sets of data need to be collected in order to run this analysis. First a classifier must determine whether the fingerprint scans belong to one of five classes: left loop, right loop, whorl, plain arch, and tented arch. These classifications are determined using Henny's classification method, and the classification can be viewed in Figure 3.6 [16].

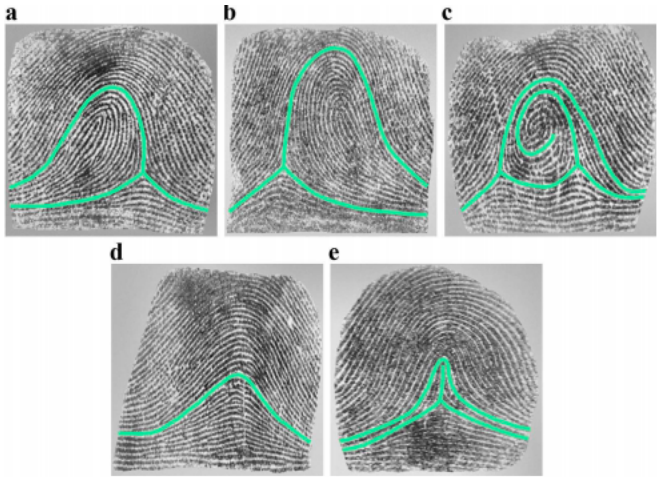

Figure 3.6: Classification of: (a) left loop, (b) right loop, (c) whorl, (d) plain arch, and (e) tented arch

The second set of data collected determines if the scan contain the soles of the hand belong to one of the five classes: small distal loop, large distal loop, tibial arch, whorl, and tibial loop by using a novel STFT method for fingerprint enhancement [17]. The data collected for this test can be viewed in Figure 3.7 [16].

The third pattern recognition tool is used to determine the angle of the right palm print. The value is calculated by the angle between the digital and axial triradius. The algorithm used for this calculation uses two independent local image descriptors calculated in two separate ways for accuracy. Although a full implementation of this system is still pending, a partial implementation written in the C language, using PostgresSQL as a database to store the images, classification results were deemed to be accurate. However, the main limitation in this study is the lack of a significant amount of data to properly train and test models. Once the system is implemented, it is up to the HCPS and others with proper data to

improve the system and improve results [16].

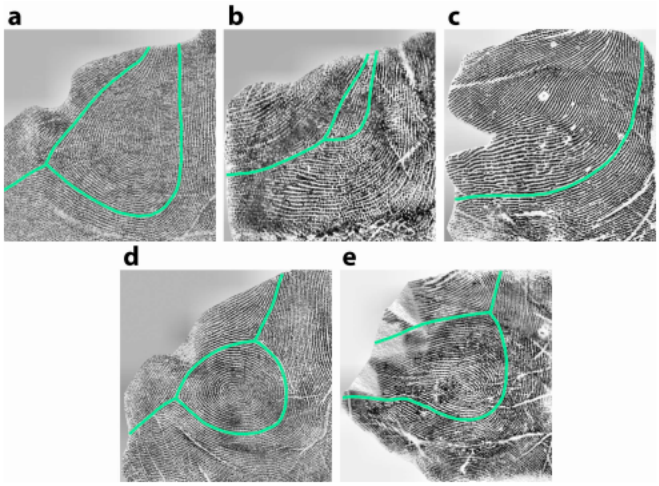

Figure 3.7: Classification of: (a) large distal loop, (b) small distal loop, (c) tibial arch, (d) whorl, and (e) tibial loop.

3.2 Limitations and Extensions

Although several applications and methods exist in order to make cognitive computing assist health care professionals with daily tasks, this field offers limitless possibilities for future work. IR models can be trained with better resources and data samples making them more effective for both the service providers and the patients in the health care field. One major issue that needs to be resolved is that the knowledge domain today is weak for this field in particular. With an increasing corpus which collects and stores relevant medical information for a domain and even a specific sub-domain, advances can be made and models can be trained more efficiently. While methods for specific specialties in medicine such as for ear, nose, and throat specialists exist, no model properly trained to encompass a full range of medical data exists partly due to the challenge of encompassing such a large variety of information which often comes from unstructured sources such as medical records which may also be protected by medical regulation laws such as HIPAA [2]. An IR system that extends a knowledge-based system tailored to handle user queries which reside in the medical

domain will also be explored for future work. However regardless of how well a system is properly implemented, large amounts of high quality data are still required to properly train any kind of medical IR model as proven in the Wojtowicz study [16]. Often times access to this data is restricted due to privacy information for medical records which are protected by HIPAA. While AI advancements are being made, proper data still remains a major barrier and will need to be addressed on a large scale sooner than later. Regardless of the challenges present, artificial intelligence in the medical domain has limitless capabilities and remains largely untapped as shown by the lack of implemented applications shown in the literature review section of this research.

This literature survey was conducted on the current state of cognitive computing, and specifically IR systems in the medical domain. The current applications using cognitive computing, the advantages, and the future possibilities were discussed. In addition, the challenges and requirements while working specifically in the medical domain are highlighted. Potential future work along with barriers which need to be addressed in detail are presented.

Chapter 4: Related Work

4.1 Revolution in Medicine

Though the human brain is an incredibly complex system, it has its own limitations. However, through the aide of cognitive computing, new heights which at once were unfathomable can be reached. Just this year, in 2018, the first reports emerged of artificial intelligence performing better than humans on a medical clinical examination. On June 28th, Dr. Mobasher Butt stood on stage in Londons Royal College of Physicians, where he announced that his company's trained AI received a score of 82%, beating out the average by medical students of 72% [18]. Dr. Ali Parsa, the founder of Babylon Health, states that on the planet, over 5 billion people lack the access to basic surgery. He claims that the U.S. has shifted its focus from health care to the economic benefits of it, and that there are large gaps in the health-care system. Dr. Parsa predicts that the U.S. will be the largest consumer of artificial intelligence in health-care in the near future [18]. With many other domains already using artificial intelligence to improve the quality of life, medicine has yet to make the breakthrough for various reasons, but the era of change is fast approaching. Starting in 2016, the U.S Department of Veterans Affairs (VA) hospitals in Durham, North Carolina have been using IBM Watson to help diagnose cancer patients by collecting DNA from tumors and analyzing the genetic material to determine possible causes as well as effective treatments. The V.A. treats nearly 4% of U.S. cancer patients, allowing IBM Watson to have a large sample size [19]. Dr. Kyu Rhee claims that "it is incredibly challenging to read, understand, and stay up-to-date with the breadth and depth of medical literature and link them to relevant mutations for personalized cancer treatments" which is a sentiment shared by many medical

professionals, justifying the need for effective usage of artificial intelligence in such a crucial domain [19]. In this research, we take a close look into some of the existing technologies and how to develop models to optimize them specifically for practical use in medical environments. While limited technologies currently exist for every-day clinical usage,the field remains wide open and a large market exists for new technologies to come into play [20]. In an unprecedented era where data is abundant and largely unused as illustrated in Figure 4.1. [21], the time to make advancements is now.

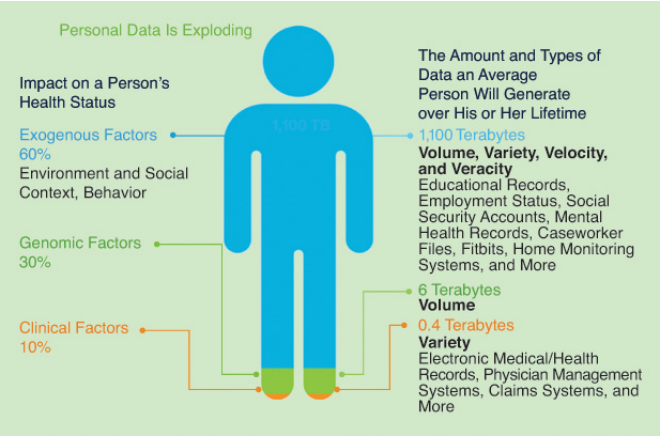

Figure 4.1: Large amounts of health-related data is being generated at an unprecedented rate

4.2 Related Research

Cognitive computing technologies have been incrementally changing the field of medicine itself. A study shows that artificial intelligence is being used in a database which holds records of diabetes for the Pima people, a group of Native Americans residing in central Arizona. The dataset holds personal records for these individuals and the data was acquired through the US National Institute of Diabetes and Digestive and Kidney Disease [22]. Notably, this dataset contains very clean data with no missing values taken from 768 females who are 21 years and older who may show signs of diabetes. This dataset contains eight attributes and 2 classification factors (diabetic or non diabetic). When using machine learning algorithms

in this test case, it is reported that there was a 77% accuracy in the classification of the test data [23].

Another recent study took into account various parameters related to medical grafting, a surgical procedure to transfer tissue from one location to another on the body, without bringing its own blood supply along with it. The benefits of this procedure are obvious and an example would be a kidney transplant where a healthy kidney replaces a defective one and can prevent kidney failure. A successful procedure enhances the quality of life for a patient as well as reduces overall cost of medical care. Although all the factors which determine successful graft procedures are still unknown, the following factors are known to have a correlation: the compatibility of the blood types between the two individuals involved in the procedure, the number of number of leukocyte antigen mismatches, and the results from cross-match testing which determines if the recipient's cells will accept or reject the new tissue. These factors were taken into account to be tested and processed by algorithms in the system because these were the factors also considered when making a professional medical determination by physicians. In this specific model, the problem was formalized by defining it as: selecting the correct kidney from the available pool of organs for a particular individual thereby maximizing the chances of a successful transplant. The factors that were taken into account for this study included: age, mismatches of related donor types, mismatches of related recipient types, recipient state, referring hospital, donor hospital, donor sex, and initial kidney preservation [23]. This study also illustrates the key point of domain specific information described above as well as domain experts in the field of medicine as well as the expansion of cognitive computing through other domains [24] as illustrated in Figure 4.2.

Once the model was built, a series of neural networks (n = 500) were independently trained to predict the outcome of a given transplant with unlabeled data. All networks consisted of 16 input neurons for the attributes and 2 output neurons to predict if the transplant was a success or not. The data was split into 3 groups: training data, test data, and validation data which showed that the model held accurate for slightly over 70% of the

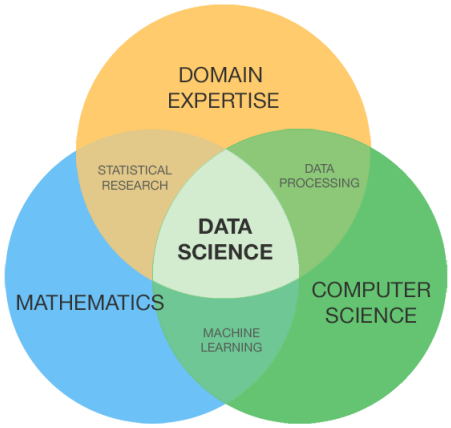

Figure 4.2: Interdisciplinary problems require interdisciplinary solutions

data [22].

Over 1.6 million Americans are diagnosed with cancer every year, and the ability of each individual to find the correct information to battle the disease can be a tough challenge. In today's world there is often an abundance of information, and filtering and sorting through this can be difficult. Moreover, the ability to extract insightful information from the literature available presents another challenge in a crucial period. Using the abundance of cancer information and data available to the public, researchers using IBM Watson developed a method using data from American Cancer Society (ACS) to create a virtual advisor. This advisor will mine data from ACS's website, cancer.org which contains over 14,000 pages of detailed information with over 70 cancer topics. The system takes into account the type of cancer as well as the stage in which the individual is in to connect the patient with various groups, activities, and education. This dynamic system will anticipate the needs of the patient and evolve over time as more input is processed through it. In the future, additional implementations such as natural language processing for speech can be implemented using IBM Watson and can help link the patients to phone calls with their physicians or other support groups [2].

Another application has been created using IBM Watson for personalized medicine. The idea behind this system is that the best cancer treatment is to detect, prevent, and treat

it before it reaches advanced stages. However, no two people or cancers are alike. The current process for trial matching is conducted through clinical coordinators who sort through thousands of patient records and match the patient with a given protocol. However, each one of these protocols has 46 requirements on average and range from containing a genetic marker to age, tumor stage, growth, and treatment history. No matter how much of an expert any individual is, this becomes a huge task to conduct without advanced computing capabilities as shown in Figure 4.3 [2]. This is why the system using Watson was created. A clinician is able to submit a patient's health data against the data in the clinical trial database and offers feedback to the physician regarding the matching relations to a specific clinical study. The need for this approach will only grow as the bounds of knowledge expands rapidly and as personalized medicine becomes common practice. Any procedure that is personalized will require targeting very small and specific instances or groups which may have no natural affiliation. Systems such as the ones described will support a higher level of personalization by enabling individual health data records to be securely connected with the clinical trials databases. These techniques can also help bring new methods of treatments on an individual basis which otherwise would logistically be possible. Whatever the case may be, bringing cognitive computing into the medical domain is no longer a luxury due to the ever expanding factual data becoming available in today's world, it is now a necessity.

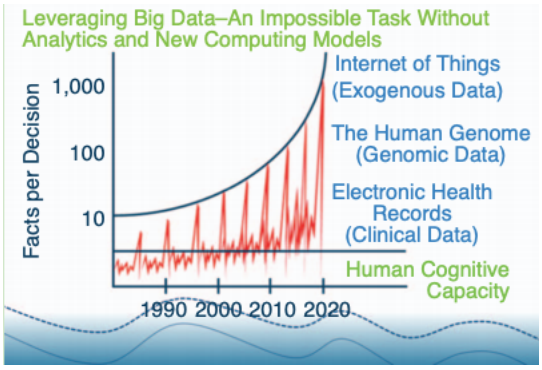

Figure 4.3: An increase in factual information eventually exceeds human cognitive abilities

Although many tools are being constantly developed in the field, few of these tools have developed far and made an impact on lives as the health-care industry still remains to be

minimally assisted by artificial intelligence. However, with powerful tools at the disposal of even common people thanks to the development of cloud architecture, the possibilities of developing an impactful product to enhance the quality of life for countless individuals remains open, and a breakthrough seems imminent.

Chapter 5: Tools for Data Analysis

5.1 Data Processing Tools

For processing medical data, certain properties can be assumed in order to optimize algorithms and enhancements to get proper results. First, data can be expected to be large in nature as is the case with most of the data being used across various domains today simply due to the data available through the advancement of technology [4]. Secondly, data can be expected to be incomplete with several attributes missing due to the source data itself not containing this information, as well as due to a lack of transparency in patient data from the health care providers due to legal limitations [9]. In addition, it can be assumed that the natural language of the data is relatively specialized and contains terms which may not be well known to the general public. A knowledge-base can be incorporated into the model in order to produce accurate results. For example, data may contain the term "pyrexia" which is more commonly known as a "fever." Part of the challenge in designing a model includes being aware of these subtleties [20].

5.1.1 Babylon Health

Founded in 2013, Babylon health's goal was to expand medical care to those who might otherwise not be able to have access to health-care services. Babylon uses artificial intelligence to receive a number of inputs from patients, uses undisclosed machine learning algorithms, and generates meaningful output which it returns to its clients. Recently, Babylon has merged with *We Chat*, a popular Chinese messaging network with over 1 billion users, to provide its artificial medical intelligence services to its users [25]. In addition, this platform allows

users to chat with a doctor in real time. The company's founder states that the goal of this artificial intelligence based system was to reach more users through the use of technology who otherwise would be unable to have access to proper health services [18]. Although this service, as shown in Figure 5.1, [26], this service is being used by many users today, the configurations of the model is fixed as well as undisclosed to the public, and the model can't be altered by anyone outside of the company, making it practical, but not adaptable.

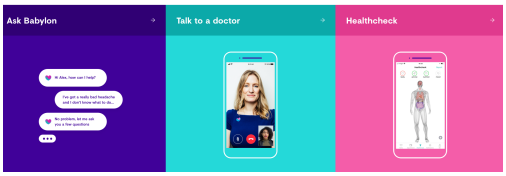

Figure 5.1: Implementation of Babylon Health being used

5.1.2 Apache UIMA

Although Babylon Health's application is promising and is already in use, its inflexible nature due to it being a private application meant for clients, makes it a poor candidate to build and customize a medical model on. The next application we examine, is Apache IUMA, a free, open-source tool for natural language processing. Unstructured Information Management Applications (UIMA) are software systems that analyze large volumes of unstructured information in order to discover knowledge that is relevant to an end user. An example model could take written input and detect entities, concepts, and keywords from the information provided. Many frameworks and languages including Java, C++, and XML for meta-data can be used. The Apache license allows any developer to make use of the frameworks and a full diagram can be seen in Figure 5.2 [27]. Through this frameworks, a model to analyze large amounts of medical data can be built, however the process itself could take many years and will be expensive for an individual or even small group. However, not all hope is lost as the next tool we discuss implements this framework within itself, allowing the aforementioned barriers to be transcended.

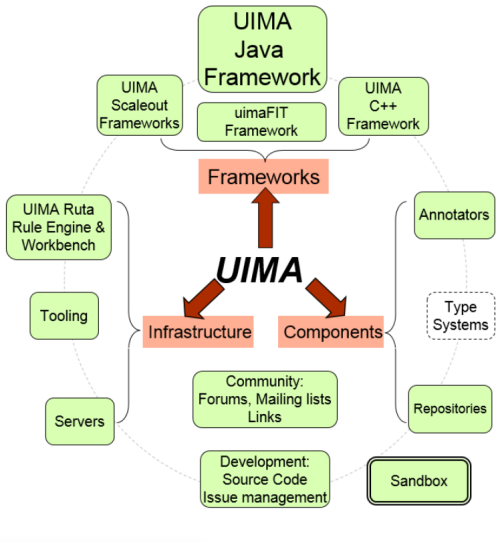

Figure 5.2: UIMA frameworks

5.1.3 IBM Watson

As many may already be familiar with, IBM Watson came to the public's attention after appearing on the popular game show, *Jeopardy* in February 2011 where it not only performed well, but decisively beat several human champions in the game. IBM Watson has since came forward as one of the leading tools in cognitive computing and is currently used across many domains today ranging from predicting disease outbreaks before they occur, to predicting who may score the next touchdown in a football game [2]. More importantly, in the context of this study of building a proper model to process large amounts of health-care related data, it implements the Apache UIMA framework within Watson Discovery [28]. Through the use of IBM Watson as a cognitive computing tool to store and process data, we will specifically be looking into enriching the customized model so that medical data can be processed as intended.

Chapter 6: Model Creation and Performance Analysis

6.1 Building a Custom Model

Although the need for various models may be highly dependant on the field, and even within the medical domain, the needs for every individual organization may be significantly different. In this study, we develop a personalized model which is adapted and trained with data regarding both type 1 and type 2 diabetes. This model is then improved within the IBM Watson environment using enrichment techniques discussed later.

6.1.1 Data Acquisition

The first step in building a valuable model includes collecting, sorting, and cleaning data to be used in the model. According to many data scientists, up to 70% of a project's life-cycle in the field may be devoted to solely collecting, cleaning up, and processing data. As previously discussed, data within the medical domain is especially susceptible to missing information, values, and even entire columns which is why it is important to have proper procedures and tools in place to process the data as needed [20]. For our model, first we build a knowledge base due to data being highly specialized. As stated, since the goal of the model is to process data regarding diabetes, a knowledge-based system needs to be built. For this model, various documents were considered and the most relevant documents were used in this scenario. Although the model should be fluid and adaptive depending on the results, a book about type 1 diabetes was selected which was authored by several domain knowledge experts (M.D.'s) [29]. This document was fed to IBM Watson Discovery to be ingested along with enrichments which will be discussed later. IBM Watson allows documents in

PDF, JSON, Word, and JSON formats up to 50 MB which can be upgraded as needed based on project requirements.

With a proper vocabulary provided with the documents mentioned, the next step involves searching for and finding relevant data which can provide meaningful output. As previously mentioned, various databases exist in the medical domain depending on the data being searched for, ranging from RNA genomes to data on bone fractures [20]. In this study, publicly available data from data.gov was selected. As discussed, when dealing with any medical information it is important to keep in mind that this data is potentially private information and only publicly available information without patient identity should be used.

Figure 6.1: Various domains of data available with public licensing

As shown in Figure 6.1, public data is available through the government for many different kinds of domains. The data specifically can be acquired through the desired formats such as JSON, CSV, PDF, etc. Although this method for data collection works well in the medical domain, it can be used across the fields of agriculture, economics, meteorology, education, safety, engineering, manufacturing and much more. Specifically withing the medical data archive, many datasets can be found with public data coming from local, state, and even the federal government. Depending on the study being conducted, data can specifically be obtained from a specific location. This can be beneficial when a study is being done on a population in a given city or state which can help researchers get the necessary data to

conduct their work. As an example, Figure 6.2 shows the result of the general query "cancer" being entered. As shown, a wide range of custom filters can be applied to find precise data for any given medical domain project or application. The filters can also be applied on the format of the data depending on which tools are being used in the application.

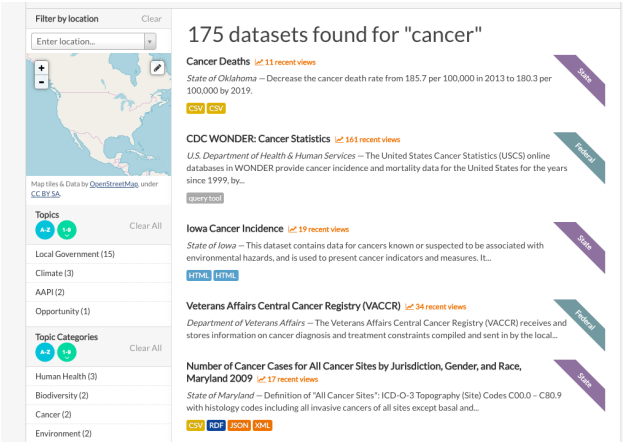

Figure 6.2: Result of a basic medical query

For this study, along with the documents used for domain knowledge on type 1 diabetes and prevalent genes, data from specific cases of diabetes reported by the government was also used with attributes such as age, gender, location, tagged genes, health condition, and much more. The purpose of adding this information is to sharpen the model to be more versatile and simulate a real-life use case of the model. The idea is for a health care professional not to be replaced, but to have the ability to be assisted with a cognitive computing model. Although a health care professional may have some level of knowledge about diabetes, it is impossible to process the amount of knowledge that a system like this can intake and provide valuable insight. However, the system still needs to be tweaked to perform well in test scenarios.

6.1.2 Model Enrichments

Once a model is loaded with data and all the files have been indexed, the time comes to adjust the system and provide it with the knowledge base discussed earlier. This step also

serves as a transfer of human knowledge into artificial intelligence. These adjustments or tweaks are referred to as enrichments within the IBM Watson environment as they help give the model being built more value. In this experiment, we take the existing model that we constructed for the medical information discussed, and enhance it using enrichments. All of the documentation for working with enrichments including entity, keyword, and concepts can be found here [30]. The first type of enrichment is entity enrichment. In entity enrichment, the model will sometimes return items such as persons, places, organizations, and references that are present in the input data when relevant. The entity extraction feature adds semantic information to content to help understand the subject and context of the text being analyzed. The next type of enrichment is concept enrichment which deals with the underlying concepts with relation to each other. Natural language processing features are used to identify underlying patterns, similar to the learning based models covered in [20]. Concept enrichments can be especially valuable in a domain such as medicine. For example, the epigenetic factors that deal with diabetes can be explored and underlying relationships that are often not obvious to the human eye can be identified through the help of artificial intelligence. The last form of enrichments include keyword enrichments. As the name suggests, critical words related to the context of the text or data are given a specific value based on their importance. Similar to how humans process information with a large amount of text, based on the context, the more significant keywords are given more value and significance when making an analysis. An example of these enrichments are shown in Figure 6.3 with an example related to the medical data being processed that we use in this work.

These enrichments can be performed within a model once the proper datasets and knowledge based system has been properly input-ed into Watson. The enrichments are coded using a JSON structure which can be easily edited and read. The attributes potentially calculated from these extractions include:

- Sentiment: boolean - optional - When true, sentiment analysis is performed on the

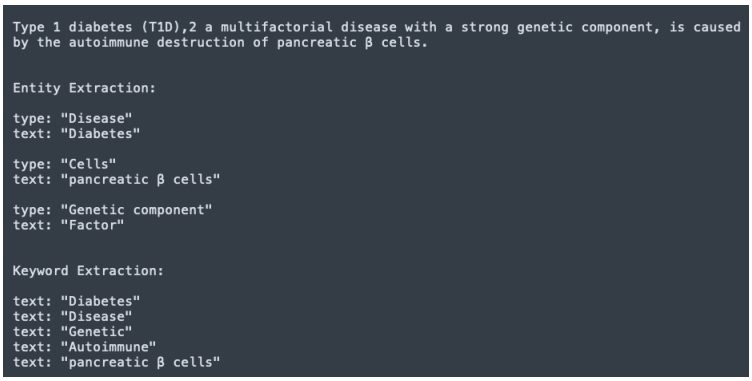

Figure 6.3: An example of enrichment extraction performed

extracted entity in the context of the surrounding content.

- Emotion: boolean - optional - When true, emotional tone analysis is performed on the extracted entity in the context of the surrounding content.
- Limit: int - optional - The maximum number of entities to extract from the ingested document. The default is 50.
- Mentions: boolean - optional - When true, the number of times that this entity is mentioned is recorded. The default is false.
- Mention Types: boolean - optional - When true, the mention type for each mention of this entity is stored. The default is false.
- Sentence Location: boolean - optional - When true, the sentence location of each entity mention is stored. The default is false.
- Model: string - optional - When specified, the custom model is used to extract entities instead of the public model.

6.1.3 Expanding Knowledge Base

Although we have already constructed a knowledge base within our IBM Watson model, that is a finite system which is also computationally expensive to maintain, and it is not advised to overload the capacity. Instead, an API which links information from DBpedia, can be used in concept enrichment. For example, we want to define a concept for a specific gene that is correlated with type 1 diabetes. It may be quite easy to create a concept for disease, diabetes, and even HLA-DQA1, a gene prominent in the onset of type 1 diabetes. However, if your model's needs are for thousands of diseases, genes, or other concepts, it becomes nearly

impossible to store this information in one system. Thus, a simple API call to DBpedia can be made. DBpedia is a crowd-sourced and open source effort to extract structured content from information from different projects. The structured information resembles an open knowledge system which is publicly available. The information is stored in a machine readable database which is architecturally set up to allow the information to be harvested, organized, shared, searched, and indexed. This allows a large amount of information to be available to the public, and allows for concepts to easily be tagged in our model. Once the API call is made, it can be easily referenced in the JSON format of data indexing used in our system. The model will now return relevance scores based on all the factors discussed and now take into account these concepts, for future queries, thus incrementally improving the model. An example is shown in Figure 6.4.

Figure 6.4: DBpedia library being linked to the model for concept enrichment

6.2 Results

After the model was trained with 5 instances of entities, 5 instances of concepts, and 5 instances of keywords, the goal was to see how this model would perform on its own when tagging these relationships, specifically after the enrichments discussed were made. The results for the query which results in displaying the top entities given the input of the diabetic dataset provided is shown in Figure 6.5

6.5 As shown, 46 entities were retrieved with a negative sentiment score. The results also showed that the entities were correctly matched for keywords such as "diabetes" which became properly tagged as a health condition. An obvious flaw arises when an uncommon phrase was thrown at the model. The phrase "t1d" which is the abbreviated form of "type

Figure 6.5: Results after entity enrichments

Genes Involved in Type 1 Diabetes

Sentiment	negative
Entities	celiac **disease**,**diabetes**,International **Diabetes**, **Diabetes** Genetics Consortium,Belgian **Diabetes** Registry
Categories	/**health** and **fitness**/**disease**/**diabetes**, /**health** and **fitness**/**disease**
Concepts	**Diabetes** mellitus,**Diabetes** mellitus type 1, **Diabetes** mellitus type 2
Text	"...Symposium on ''**Diabetes** and **health**''...." "...Shared and distinct genetic variants in type 1 **diabetes** and celiac **disease**...." "...Prediction and interaction in complex **disease** genetics: experi- ence in type 1 **diabetes**...." "...The Next-Generation Sequencing (NGS) technology has opened new avenues to elucidate the role of coding and noncoding RNAs in **health** and **disease** and would speed up the identification of causative gene variants in T1D...." "...Analysis of 17 autoimmune **disease**-associated variants in type 1 **diabetes** identifies 6q23/TNFAIP3 as a suscepti- bility locus...."

Figure 6.6: Results for the query "genes involved in type 1 diabetes"

1 diabetes" came up, the model incorrectly tagged this as a location as opposed to a health condition. Although the statistics such as relevance score and percent of entities correctly identified do provide some form of quantified analysis for this model, the best measure for the system is to be able to use it in a real world situation. The model returned 0 results when the knowledge base was tied in with the data processed into the system. The model was unable to differentiate the diabetic dataset with the data which was processed to build up

the knowledge base. By far the best analysis that was able to be detected was for sentiment analysis. Overall, the system had its strength and weaknesses, but had several shortcomings. A system such as this can be used to quickly sort various entities, keywords, and concepts, as well as summarize a large amount of text. However, the connection between underlying patterns can not be processed by such a model which is what is needed for an advancement in cognitive computing in medicine.

Chapter 7: Closing Thoughts

7.1 Model Limitations

One of the biggest challenges that physicians face today is that they do not routinely connect with their patients after a procedure, especially a surgical one. They have no way of knowing how well their patient is recovering or even the average case scenario for their patients to recover [31]. Though services and models like Watson can help bridge these gaps and do more, there still exist some limitations that will take time to overcome. In 2017, a collaboration project with MD Anderson Center in Houston Texas, and with IBM Watson for $39 million came to a standstill as it was terminated due to overambitious goals not being met by both sides. The medical center, which is overseen by University of Texas, had a plan to read data about patients' symptoms, gemomic sequences, and pathology reports. This data, alongside physician notes were to be combined and help produce a possible diagnosis and treatment for the patient. The project was simply too complicated for the technology in place today. The primary point of failure came with not having enough complete data to train the model with [31]. Although large amounts of data publicly exists, much of it remains unstructured.

Cognitive models learn by continually tweaking its internal processes in order to produce the highest percentage of correct answers based on some training set. An example would be to classify a set of radiological images as cancerous or not. The correct classifications in this case are well known and it is fairly easy to train a model with data. However, the true challenge comes when solving problems where critical thinking is involved and also with problems that go beyond human knowledge such as detecting relationships between specific gene variances and a particular disease. Similar to a chicken and egg problem, the correct

relationships need to be established first. In a cognitive computing area like self-driving vehicles, it is fairly easy for a common person with no domain knowledge to identify what a street is, what a stop sign is, where the road ends, etc. However in the medical domain, it takes experts to identify labels for data. Simply put, any system requires a large sample size of structured data [20]. In the medical domain, this is hard to come by, and even if the data exists, it needs to be labeled by domain experts. Perhaps the time is a bit too soon to see the next big advancement in medical cognitive computing, but it is the time to be proactive in taking measures to solve the solution. In this case, it is the time to start streamlining processes so that structured and potentially useful data can be collected in a way that can be used to train such cognitive models. As portrayed in Figure 7.1 [32], this needs to be a community wide approach. Institutions, corporations, researchers, and hospitals all need to be on the same page and work together to establish the necessary procedures in place for a smoother operation so that the limits of these systems can be expanded as the potential reward could be greater than imaginable in the next paradigm of technology.

Figure 7.1: Community wide approach to solving the problem

7.2 Future Work

Although the model had limitations, the biggest strengths of it included the identification of sentiment as well as the identification of entities. Such a model could particularly be used for an application for tracking patients and their moods throughout a period of time.

An application like Neuro-Diary [20] could greatly benefit from a model similar to the one created in this project. Such applications could help health-care providers better assess and monitor patient health, even when there are more patients than they can handle. A model like this would greatly reduce the amount of time spent analyzing as the process becomes automated. Similar models could also have a huge impact on social fields such as psychology and sociology to analyze sentiments, especially with large sets of data. Using the knowledge-based approach we used in this study can greatly benefit many various systems across many domains. Through the recent advancements in cognitive computing, nearly every domain has space to push forward, however, the necessity is pressing in the medical domain. As explored in this study, the biggest hurdle left to overcome in the medical domain still remains having access to large amounts of unbroken, useful, and structured data. The next breakthrough in this area will require more emphasis on a structured system for acquiring data which is structured as needed.

7.3 Conclusions

All the various procedures discussed were performed including:

- data acquisition - relevant data and a small knowledge base was built from scratch and used to generate a medical cognitive model. In this project, we used a type 1 diabetes dataset from data.gov where several genes were identified based on their correlation with the disease in individuals.
- data ingestion - The data which was collected was placed into a knowledge-based system within IBM Watson and stored, indexed, and processed.
- model enhancements - enhancements were made using keyword, entity, and content enrichment techniques, and the DBpedia API was linked and used to store these concepts to help improve the performance of the model.
- shortcomings and limitations- the limitations of both the model built as well as cognitive medical models in general and the complexities of building such a model and the lack of proper resources, especially clean data were discussed.
- future implementations and improvements - were mentioned and discussed.

BIBLIOGRAPHY

[1] H. Wang, Q. Zhang, and J. Yuan, "Semantically enhanced medical information retrieval system: A tensor factorization based approach," *IEEE Access*, vol. 5, pp. 7584–7593, 2017.

[2] M. N. Ahmed, A. S. Toor, K. O'Neil, and D. Friedland, "Cognitive computing and the future of health care cognitive computing and the future of healthcare: The cognitive power of ibm watson has the potential to transform global personalized medicine," *IEEE Pulse*, vol. 8, no. 3, pp. 4–9, May 2017.

[3] P. Pathak, M. Gordon, and W. Fan, "Effective information retrieval using genetic algorithms based matching functions adaptation," in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, Jan 2000, pp. 8 pp. vol.1–.

[4] J. Callan, "Distributed information retrieval," in *Advances in information retrieval*, W. B. Croft, Ed. Kluwer, 2000, pp. 127–150.

[5] B. Ermiş, E. Acar, and A. T. Cemgil, "Link prediction in heterogeneous data via generalized coupled tensor factorization," *Data Mining and Knowledge Discovery*, vol. 29, no. 1, pp. 203–236, Jan 2015. [Online]. Available: https://doi.org/10.1007/s10618-013-0341-y

[6] D. Wegener, S. Rossi, F. Buffa, M. Delorenzi, and S. Riiping, "Towards an environment for data mining based analysis processes in bioinformatics amp; personalized medicine," in *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, Nov 2011, pp. 570–577.

[7] J. Y. Nie and W. Shen, "Flexible concept matching for medical information retrieval," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, Oct 2015, pp. 1901–1906.

[8] S. Fox, "Health topics," 2011.

[9] G. Zuccon, B. Koopman, and P. Bruza, "Exploiting inference from semantic annotations for information retrieval: Reflections from medical ir," in *Proceedings of the 7th International Workshop on Exploiting Semantic Annotations in Information Retrieval*, ser. ESAIR '14. New York, NY, USA: ACM, 2014, pp. 43–45. [Online]. Available: http://doi.acm.org/10.1145/2663712.2666197

[10] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 926–934. [Online]. Available: http://papers.nips.cc/paper/5028-reasoning-with-neural-tensor-networks-for-knowledge-base-completion.pdf

[11] W. Shen and J.-Y. Nie, "Is concept mapping useful for biomedical information retrieval?" in *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. Jones, E. San Juan, L. Capellato, and N. Ferro, Eds. Cham: Springer International Publishing, 2015, pp. 281–286.

[12] G. Tognola, A. Murri, and D. Cuda, "Cognitive computing for the automated extraction and meaningful use of health data in narrative medical notes: An application to the clinical management of hearing impaired aged patients," in *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, March 2018, pp. 299–302.

[13] A. Farrugia, D. Al-Jumeily, M. Al-Jumaily, A. Hussain, and D. Lamb, "Medical diagnosis: Are artificial intelligence systems able to diagnose the underlying causes of specific headaches?" in *2013 Sixth International Conference on Developments in eSystems Engineering*, Dec 2013, pp. 376–382.

[14] E. Papageorgiou, C. Stylios, and P. Groumpos, "A combined fuzzy cognitive map and decision trees model for medical decision making," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2006, pp. 6117–6120.

[15] M. Frize, D. Ibrahim, C. Catley, and R. C. Walker, "Using artificial intelligence to estimate outcomes in perinatal medicine," in *2006 Canadian Conference on Electrical and Computer Engineering*, May 2006, pp. 730–733.

[16] H. Wojtowicz, J. Wojtowicz, W. Koziol, and W. Wajs, "Medical decision support system architecture for diagnosis of down's syndrome," in *2013 Federated Conference on Computer Science and Information Systems*, Sept 2013, pp. 179–182.

[17] S. Chikkerur, V. Govindaraju, and A. N. Cartwright, "Fingerprint image enhancement using stft analysis," in *Pattern Recognition and Image Analysis*, S. Singh, M. Singh, C. Apte, and P. Perner, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 20–29.

[18] P. Olson, "This ai just beat human doctors on a clinical exam," Jun 2018. [Online]. Available: https://www.forbes.com/sites/parmyolson/2018/06/28/ai-doctors-exam-babylon-health/#2b20579b12c0

[19] A. Moscaritolo, "Va reenlists ibm's watson in fight against cancer," Jul 2018. [Online]. Available: https://www.pcmag.com/news/362590/va-reenlists-ibms-watson-in-fight-against-cancer

[20] A. Gudivada and N. Tabrizi, *IEEE 2018 Symposium Series on Computational Intelligence*.

[21] "Cognitive computing and the future of health care." [Online]. Available: https://pulse.embs.org/may-2017/cognitive-computing-and-the-future-of-health-care/

[22] F. Shadabi and D. Sharma, "Artificial intelligence and data mining techniques in medicine amp;150; success stories," in *2008 International Conference on BioMedical Engineering and Informatics*, vol. 1, May 2008, pp. 235–239.

[23] W. Duch, R. Adamczak, and K. Grabczewski, "A new methodology of extraction, optimization and application of crisp and fuzzy logical rules," *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 277–306, March 2001.

[24] Shellypalmerdigitalliving, "Data science advisory." [Online]. Available: https://www.shellypalmer.com/data-science/

[25] H. Crouch, "Babylon expands its ai technology to mainland china," Apr 2018. [Online]. Available: https://www.digitalhealth.net/2018/04/babylon-ai-technology-china-tencent/

[26] "babylon dr app." [Online]. Available: https://www.babylonhealth.com/product

[27] "Welcome to the apache uima project." [Online]. Available: http://uima.apache.org/

[28] [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SS5RWK_3.5.0/com.ibm.discovery.es.nav.doc/iiysaovcapipe.htm

[29] A. P. Escher and A. Li, *Type 1 diabetes*. InTech, 2013.

[30] "Cognitive computing and the future of health care." [Online]. Available: https://pulse.embs.org/may-2017/cognitive-computing-and-the-future-of-health-care/

[31] D. H. Freedman, "What will it take for ibm's watson technology to stop being a dud in health care?" Jul 2017. [Online]. Available: https://www.technologyreview.com/s/607965/a-reality-check-for-ibms-ai-ambitions/

[32] L. Shapiro, "Ibm watson for healthcare startups," Jun 2015. [Online]. Available: https://www.slideshare.net/levshapiro/mhealth-israelibm-watson-for-healthcare-startups

[33] P. Domingos, "A unified bias-variance decomposition for zero-one and squared loss," in *Proc. National Conference on Artificial Intelligence and Proc. Conference Innovative Applications of Artificial Intelligence*. AAAI Press / The MIT Press, 2000, pp. 564–569.

[34] H. Turtle and J. Flood, "Query evaluation: strategies and optimizations," *IP&M*, vol. 31, no. 6, pp. 831–850, 1995.

[35] K. Aberer, "P-Grid: A self-organizing access structure for P2P information systems," in *Proc. International Conference on Cooperative Information Systems*. London, UK: Springer, 2001, pp. 179–194.

[36] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in *Proc. CIKM*, 2004, pp. 42–49.

[37] R. H. Creecy, B. M. Masand, S. J. Smith, and D. L. Waltz, "Trading MIPS and memory for knowledge engineering," *CACM*, vol. 35, no. 8, pp. 48–64, 1992.

[38] O. Sornil, "Parallel inverted index for large-scale, dynamic digital libraries," Ph.D. dissertation, Virginia Tech, 2001. [Online]. Available: scholar.lib.vt.edu/theses/available/etd-02062001-114915/

[39] D. Harman, R. Baeza-Yates, E. Fox, and W. Lee, "Inverted files," 1992, pp. 28–43.

[40] H. Garcia-Molina, J. Widom, and J. D. Ullman, *Database System Implementation*. Upper Saddle River, NJ, USA: Prentice Hall, 1999.

[41] I. S. Altingövde, R. Ozcan, H. C. Ocalan, F. Can, and Ö. Ulusoy, "Large-scale cluster-based retrieval experiments on Turkish texts," in *Proc. SIGIR*. ACM Press, 2007, pp. 891–892.

[42] D. Gauding, "Two sentara hospitals among ibm watson health 100 top hospitals." [Online]. Available: https://www.sentara.com/woodbridge-virginia/aboutus/news/news-articles/two-sentara-hospitals-among-ibm-watson-health-100-top-hospitals.aspx