International Journal of Software Science and Computational Intelligence
Volume 12 • Issue 3 • July-September 2020

Developing Concept Enriched Models for Big Data Processing Within the Medical Domain

Akhil Gudivada, East Carolina University, USA
James Philips, East Carolina University, USA
Nasseh Tabrizi, East Carolina University, USA
https://orcid.org/0000-0002-3949-0065

ABSTRACT

Within the past few years, the medical domain has endeavored to incorporate artificial intelligence, including cognitive computing tools, to develop enriched models for processing and synthesizing knowledge from Big Data. Due to the rapid growth in published medical research, the ability of medical practitioners to keep up with research developments has become a persistent challenge. Despite this challenge, using data-driven artificial intelligence to process large amounts of data can overcome this difficulty. This research summarizes cognitive computing methodologies and applications utilized in the medical domain. Likewise, this research describes the development process for a novel, concept-enriched model using the IBM Watson service and a publicly available diabetes dataset and knowledge-base. Finally, reflection is offered on the strengths and limitations of the model and enhancements for future experiments. This work thus provides an initial framework for those interested in effectively developing, maintaining, and using cognitive models to enhance the quality of healthcare.

KEYWORDS

Artificial Intelligence, Cognitive Computing, Big Data, Medical, Information Retrieval

INTRODUCTION

Though the human brain is an incredibly complex system, it has its own limitations on the amount of information it can synthesize and recall. However, through the aid of cognitive computing, human cognition can be supplemented with computer systems that implement semantic and neural models of human thought (Wang et al., 2018). Through the application of cognitive computing to a plethora of diverse domains, new advances once unfathomable can hopefully be attained. For example, in 2018 the first reports emerged of artificial intelligence performing better than humans on a medical clinical examination. On June 28th, 2018, Dr. Mobasher Butt stood on stage in London's 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% (Olson 2018). 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 United States has shifted its focus from health care to its economic benefits, and that there are large gaps in the health-care system. To fill these gaps in its healthcare infrastructure and services, Parsa

DOI: 10.4018/IJSSCI.2020070105

predicts that the United States will be the largest consumer of artificial intelligence in healthcare in the near future (Olson, 2018).

While many other domains already use artificial intelligence to enhance the quality of life for their users, medicine has yet to make the breakthrough for various reasons. Cognitive Computing for the medical field heralds an era of change rapidly approaching. Starting in 2016, the U.S Department of Veterans Affairs (VA) hospitals in Durham, North Carolina, have used IBM Watson to help with diagnosing cancer patients by collecting DNA from tumors and analyzing the genetic material to determine possible causes as well as effective treatments. The VA treats nearly 4% of U.S. cancer patients, allowing IBM Watson to have a large sample size (Moscaritolo, 2018). 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”; this sentiment is shared by many medical professionals, justifying the need for effective usage of artificial intelligence in the crucial domain of medicine (Moscaritolo, 2018).

In this paper, we examine existing cognitive computing technologies and the process for developing 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, untapped market exists for new technologies to emerge (Gudivada and Tabrizi, 2018). In an unprecedented era where data is abundant yet largely under-utilized, the time to make advancements has arrived (Ahmed, Toor, O’Neil, & Friedland, 2017).

The rest of the paper is organized as follows:

- Motivation
- Related Work
- Data Processing Tools
- Building a Custom Concept-Enriched Model
- Model Enrichments
- Results
- Future Work
- Conclusion

This paper reflects a partially revised and updated version of the authors’ research originally published at the 18th IEEE International Conference on Cognitive Informatics & Cognitive Computing (2019).

MOTIVATION AND CONTRIBUTION

As many fields progress with the assistance of cognitive computing, the field of health care is also adapting, providing many benefits to medical practitioners and patients. However, advancements in this area are hindered by several challenges, including discrepancies between user queries and the knowledge base, query mismatches that retrieve sub-optimal results, and a varying spectrum of domain knowledge in users. In this research, we explore existing methodologies as well as look into existing real-life applications of Cognitive Computing that are currently in use in the medical domain. These technologies include Babylon Health, Apache UIMA, and IBM Watson. Using IBM Watson, we also investigate the creation of concept-enriched models to overcome the challenges related to cognitive computing in the medical domain. These concept-enriched models that can be tailored specifically for medically intensive applications capable of handling large amounts of data.

The purpose of this work is to impart to the reader an understanding of artificial intelligence tools and methodologies being used in the medical field today, as well as future possibilities in the domain. The concept-enriched model 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, and model creation, and reflect on model improvement for future experiments.

Therefore, in this research our contributions include:

- A survey of existing technologies, methodologies, and challenges in medical cognitive computing present in the literature.
- Introduction of a novel concept-enriched model created using IBM Watson and a publicly available diabetes dataset and knowledge-base.
- Reflections on the model's strengths in entity recognition, text summarization, and sentiment analysis as well as directions for future work to overcome model limitations and enhance the application of Cognitive Computing to the medical domain.

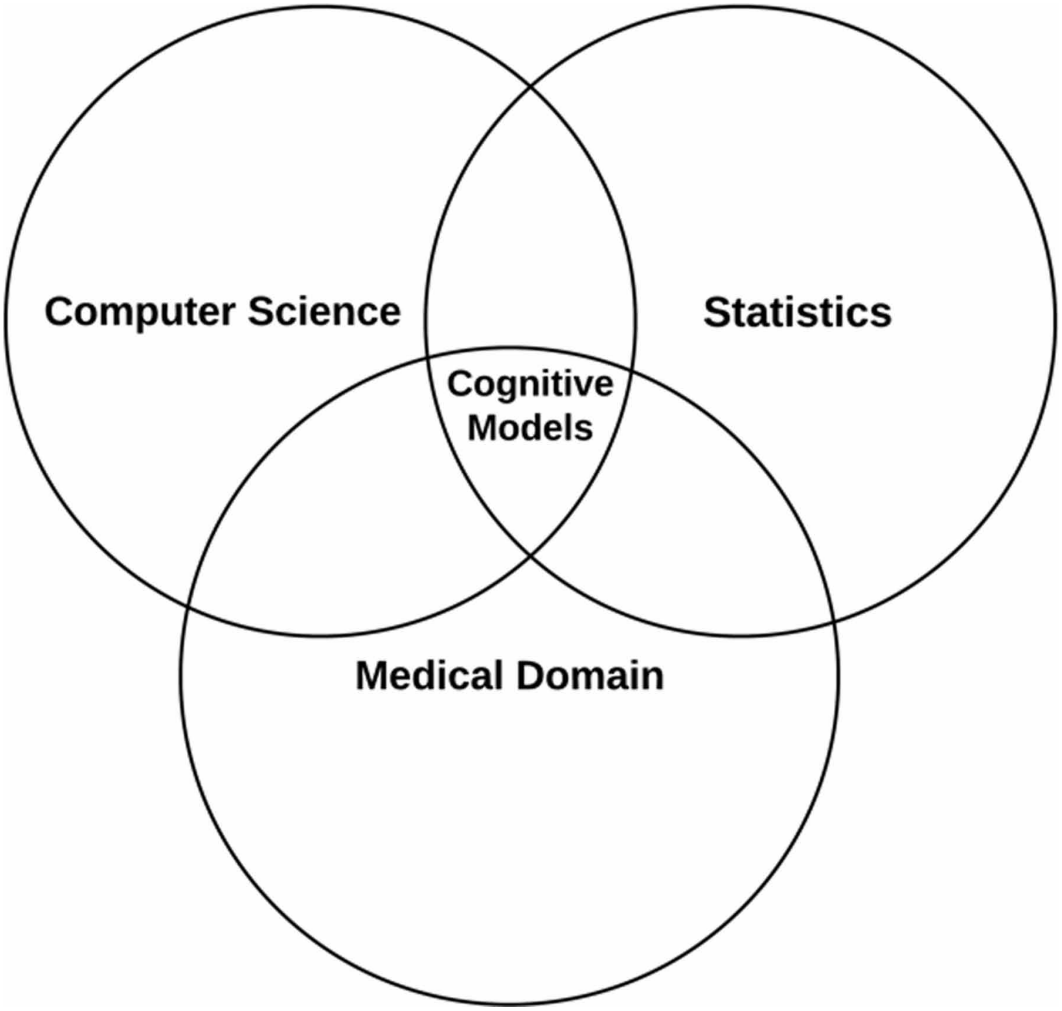
RELATED WORK

Cognitive computing technologies have been incrementally changing the field of medicine itself. One 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 (Shadbabi and Sharma, 2008). 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 (Duch, Adamczak, & Grabczewski, 2001).

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; an example would be a kidney transplant where a healthy kidney replaces a defective one, preventing kidney failure. 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 same factors considered by physicians prior to transplant. 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 (Duch, Adamczak, & Grabczewski, 2001). 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 as illustrated in Figure 1 (Palmer, 2019). 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 data (Shadbabi and Sharma, 2008).

Several recent studies have highlighted the growing challenges and importance of Big Data methodologies and cloud infrastructure for data aggregation, knowledge management and consumption from heterogeneous media sources beyond the medical domain. Originating in applications for media streaming of ultra-high definition video, security for and feature extraction from Internet-of-Things (IoT) networks in cloud computing contexts, and the application of soft computing to Big Data,

Figure 1. Interdisciplinary problems require interdisciplinary solutions



these studies have relevance as well for developing robust cognitive computing architectures for the medical domain. As Psannis, Stergiou, and Gupta observe, "intelligent clouds" can be helpful for compressing, storing, and processing streaming media on a Big Data scale (2019). Moreover, another Big Data challenge consists of the vast quantities of data generated by networked IoT devices. These collections of networked sensors present a challenge for developing efficient, sustainable and secure cloud computing infrastructure for Big Data processing, including monitoring patient health metrics (Stergiou, Psannis, Bupta, & Ishibashi, 2018; Din, Paul, Ahmad, Gupta, & Rho, 2018; Murugan, 2019). Furthermore, due to the heterogeneity of Big Data soft computing techniques will be essential (Gupta, Agrawal, Yamaguchi, Sheng 2018). Cognitive-based information retrieval systems will need to leverage text mining for bibliometric analysis to extract geographic, author collaborations, and topical summarization from published medical research (Hao, Chen, Li, & Yan, 2018; Al-akashi, 2019). Additionally, cognitive systems will need to interface with electronic health record management systems for patients (Ziebell, Albors-Garrigos, Schoeneberg, & Marin, 2019). Thus, cognitive computing systems in the medical domain will need to use efficient, sustainable cloud-native architectures and be capable of aggregating and synthesizing structured and unstructured textual,

sensor-based, and audio-visual data from patient health records, textual research and patient record knowledge-bases, and distributed networks of IoT-based medical sensors.

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 (Ahmed, Toor, O'Neil, & Friedland, 2017).

This is why the system using Watson was created. A clinician is able to submit a patient's health data and compare it against the data in the clinical trial database. Watson 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 expand rapidly and as personalized medicine becomes a pervasive 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 treatment on an individual basis which otherwise would be logistically impossible. 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.

Furthermore, as Chen, Argentinis, and Weber note the proliferation of the Big Data challenge necessitates collection, integration, synthesis of data in a plethora of formats on a vast scale (Chen, Argentinis, & Weber, 2016). Through predictive modeling enhanced by machine learning, cognitive systems such as Watson can assimilate domain-specific, technical content. This makes cognitive systems ideal tools for the modern medical domain that creates an incessant stream of clinical research much faster than any single doctor or medical researcher can comprehend it. As they note, analysis of medical Big Data due to its disparate data types (structured and unstructured), scale, and complexity including proteomics, metabolomics, and genomics thwarts traditional methods (Chen, Argentinis, & Weber, 2016), but cognitive tools enable synthesis and interpretation by combining and connecting these heterogeneous data at the terabyte scale yielding new insights into disease origins and treatments.

Although many tools are being developed in the field, few of these tools have developed far and made a significant impact on lives of patients since the health-care industry remains has yet to adopt artificial intelligence extensively. 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.

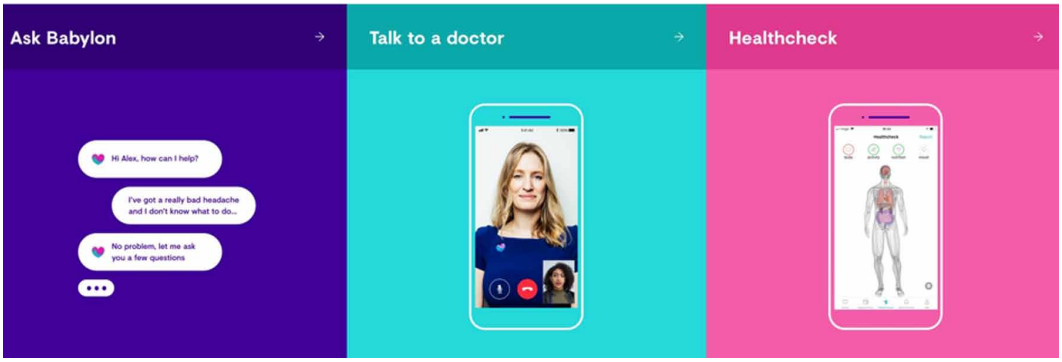
DATA PROCESSING TOOLS

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 (Callan 2000). 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 (Zuccon, Koopman, & Bruza, 2014). 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 (Moscaritolo, 2018).

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 (Crouch, 2018). 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 (Olson, 2018). Although this service, as shown in Figure 2 (Babylon Health Services, 2020), 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 cannot be altered by anyone outside of the company, making it practical, but not adaptable.

Figure 2. Implementation of Babylon Health being used



Apache UIMA

Although Babylon Health's application is promising and is already in use, its inflexible nature as a private application intended for clients makes it a poor candidate to build and customize a medical model. The next application we examined is Apache UIMA. This is 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 along with the UIMA tool. The Apache license allows any developer to make use of the frameworks (Ferrucci and Lally, 2004). Through these 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 incorporates this framework, allowing the aforementioned barriers to be transcended.

IBM Watson

IBM Watson came to the public's attention after appearing on the popular US 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 emerged 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 (Ahmed, Toor, O'Neil, & Friedland, 2017). 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 (IBM Watson Content Analytics, 2020). 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.

BUILDING A CUSTOM CONCEPT-ENRICHED MODEL

Although the need for various models may be highly dependent 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.

Data Acquisition

The first step in building a valuable model includes collecting, sorting, and cleaning data to be used in the model.

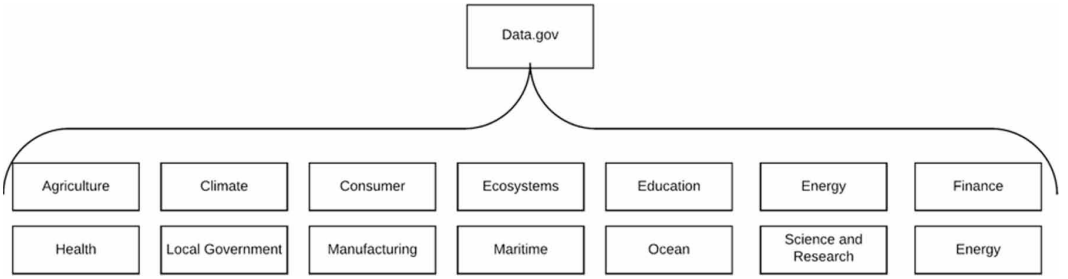
According to many data scientists, up to 70% of a project's life-cycle in the field may be devoted to solely collecting, cleaning up, and processing data. As previously discussed, data within the medical domain is especially susceptible to missing information, values, and even entire columns which is why it is important to have proper procedures and tools in place to process the data as needed (Moscaritolo, 2018). 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, an open-access book about type 1 diabetes was selected which was authored by several domain knowledge experts (M.D.s) (Escher, 2013). For the data acquisition and preprocessing phase, this document was loaded into *IBM Watson Discovery* to be ingested along with enrichments which will be discussed later. IBM Watson allows documents in PDF, Microsoft Word, and JSON formats up to 50 MB. This size constraint 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 (Moscaritolo, 2018). 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.

As shown in Figure 3, 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, within 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 4 shows the result of entering the general query "cancer". As shown, a wide range of custom filters can be applied to find precise data for any given medical

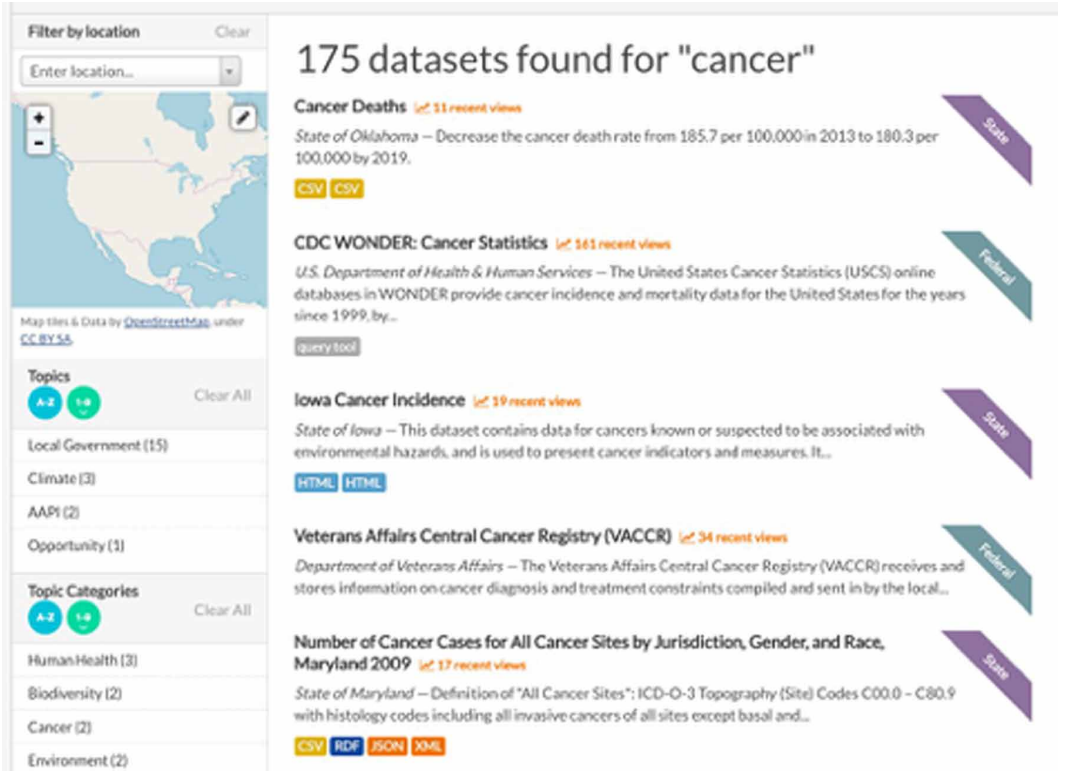
Figure 3. Various domains of data available with public licensing from Data.gov



domain project or application. The filters can also be applied to restrict the format of the data to tailor the dataset to the researcher's toolchain and workflow.

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 purpose of creating these enriched-models is not to replace health care professionals but to assist their domain knowledge through the use of these cognitive computing tools. Although a health care professional may have some level of knowledge about diabetes, it is impossible for a

Figure 4. Result of a basic medical query in Data.gov public data repository



single human to process the amount of data that a system like this can take in and provide valuable insight. However, the system still needs to be tuned to perform well in test scenarios.

Model Enrichments

Once a model has been 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 tunings are referred to as enrichments within the IBM Watson environment as they enrich the value of the model being constructed. 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 in the IBM Cloud documentation (IBM Cloud Docs). 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 increase understanding of the subject and context of the text being analyzed. The second type of enrichment is concept enrichment. This deals with the relation of underlying concepts to each other. Natural language processing features are used to identify underlying patterns, similar to the learning-based models covered in (Moscaritolo, 2018). Concept enrichments can be especially valuable in a domain such as medicine. For example, the epigenetic factors in 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 final enrichment type is keyword enrichment. 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 weight when making an analysis. An example of these enrichments is shown in Figure 5 with an example related to the diabetic data being processed that we utilized in this paper.

These enrichments can be performed within a model once the proper datasets and knowledge-based system has been properly given as input into Watson. The enrichments are coded using a JSON (Javascript Object Notation) structure conducive to human reading and editing. Attributes potentially calculated from these extractions include those shown in Table 1.

Figure 5. An example of enrichment extraction performed

```
Type 1 diabetes (T1D),2 a multifactorial disease with a strong genetic component, is caused
by the autoimmune destruction of pancreatic β cells.


Entity Extraction:

type: "Disease"
text: "Diabetes"

type: "Cells"
text: "pancreatic β cells"

type: "Genetic component"
text: "Factor"


Keyword Extraction:

text: "Diabetes"
text: "Disease"
text: "Genetic"
text: "Autoimmune"
text: "pancreatic β cells"
```

Table 1. Description of potential attributes calculated from extracted entities

Attribute name	Data type	Necessity	Description
Sentiment	Boolean	Optional	Performs sentiment analysis on the extracted entity in its surrounding context
Emotion	Boolean	Optional	Performs emotional tone analysis on the extracted entity in its surrounding context
Limit	Integer	Optional	Sets a maximum number of entities to extract from the document. Defaults to 50.
Mentions	Boolean	Optional	Number of occurrences of this entity in the document. Defaults to False.
Mention types	Boolean	Optional	Stores the mention type for each mention of the entity in the document. Defaults to False
Model	String	Optional	Specifies the use of the custom model to perform entity extraction instead of the public model
Sentence location	Boolean	Optional	Stores sentence location for each entity. Defaults to False.

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.

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

Figure 6. DBpedia library being linked to the model for concept enrichment


```
{
  "text": "Type 1 diabetes (T1D),2 a multifactorial disease with a strong genetic component,
  is caused by the autoimmune destruction of pancreatic β cells. .",
    "enriched_text": {
      "concepts": [
        {
          "text": "Diabetes",
          "relevance": 0.91136,
          "dbpedia_resource": "http://dbpedia.org/resource/Diabetes"
        },
        {
          "text": "Disease",
          "relevance": 0.886784,
          "dbpedia_resource": "http://dbpedia.org/resource/Disease"
        }
      ]
```


relationships, specifically after the enrichments discussed above 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 7.

As exemplified in Figure 8, the system retrieved 46 entities with a negative sentiment score. The results also indicated that the entities were correctly matched with keywords such as "diabetes" which were properly tagged as a health condition. An obvious flaw arises when an uncommon phrase was given to the model. When phrase "t1d" (the abbreviated form of "type 1 diabetes") was encountered, the model incorrectly tagged this as a location as opposed to a health condition. Although performance metrics, such as relevance score and percentage of entities correctly identified, do provide some form of quantified analysis for this model, the best performance measure for the system is its use it in a real-world situation. The model returned no results when the knowledge base was integrated 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 sentiment analysis. Overall, the system had its strengths in entity-recognition, keyword extraction, and textual summarization. However, the connection between underlying patterns cannot yet be processed by such a model. Ultimately, this is what is needed for an advancement in medical cognitive computing.

LIMITATIONS

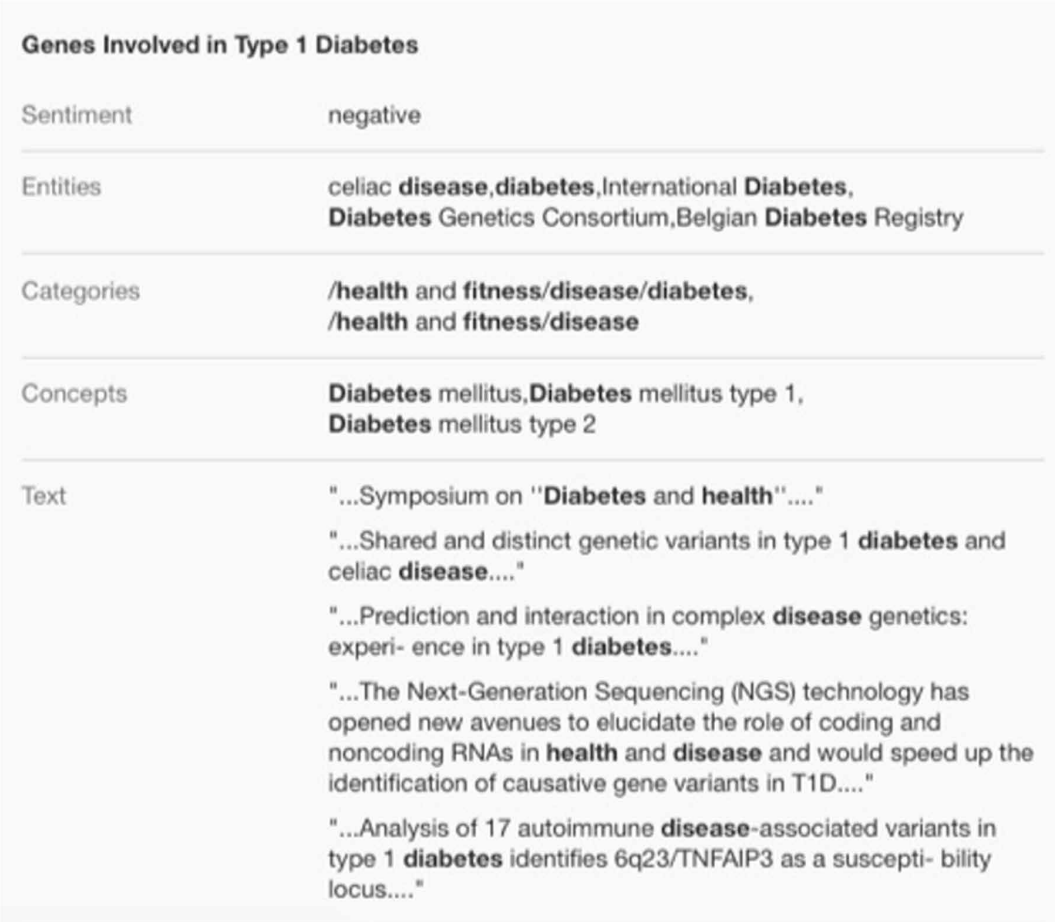
One of the biggest challenges that physicians face today is that they do not routinely connect with their patients after a procedure, especially a surgical one. They have no way of knowing how well their patient is recovering or even the average case scenario for their patients to recover (Freedman, 2017). 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 $39 million collaboration project between MD Anderson Center in Houston Texas, and IBM Watson came to a standstill; it was terminated due to overambitious goals not being met by both sides. The medical center, overseen by University of Texas, had a plan to use Watson technology to read data about patients' symptoms, genomic sequences, and pathology reports. These data, alongside physician notes, were to be combined and help produce a possible diagnosis and treatment for the patient. Despite its ambitious vision, the project was simply too complicated for the Cognitive Computing technology in place today. The primary point of failure came with not having enough complete data with which to train the model (Freedman, 2017). Although large amounts of data publicly exist, much of it remains unstructured.

Cognitive models learn by continually tweaking their internal processes in order to produce the highest percentage of correct answers based on some training set. An example would be to classify

Figure 7. Results after entity enrichments in JSON format

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

Figure 8. Results for the query "genes" involved in "type 1 diabetes" in our system



knowledge such as detecting relationships between specific gene variances and a particular disease. Similar to the chicken and egg problem, the correct relationships need to be established first. For example, in a Cognitive Computing area such as self-driving vehicles, it is fairly easy for a 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 (Moscaritolo, 2018). 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. Nevertheless, it is the time to be proactive in taking measures to reach 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 Shapiro notes (Shapiro 2015), 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.

FUTURE WORK

Although the model constructed in this research from the public diabetes dataset had limitations, its biggest strengths included the identification of sentiment as well as the correct 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 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 challenge 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. Additional future work would include comparison of this model with models created by other cognitive systems and strengthening the empirical evaluation of our model's performance.

CONCLUSION

In this work the authors present a novel model with customized enrichments for processing Big Data in the medical field. As part of the data acquisition step, the authors collected relevant data from the US Data.gov site on Type 1 diabetes and integrated this with a test knowledge-base. This was used to generate a medical cognitive model in IBM Watson. The dataset consisted of genes identified based on correlation with the diabetes disease in the individuals.

Next, as part of the data ingestion step, the collected data was entered into a knowledge-based system within the IBM Watson platform. The system stored it, indexed, and processed it. Once the data had been ingested into the system, enhancements based on keyword, entity, and content enrichment techniques were applied to the model. Moreover, the DBpedia API was linked with the enrichments and used to store concepts for improved model performance.

After discussing the results of the experiment, the authors assessed the shortcomings and limitations of the model and limitations of cognitive medical models generally. Likewise, the authors discussed complexities of building these models, including the lack of proper resources, especially clean data.

Due to the highly-technical domain knowledge in the medical field, this complexity presents challenges for natural language extraction, even for the Cognitive Computing systems discussed in this paper. The inherent ambiguity of human language coupled with the technicality of medical terminology can contribute to failures to retrieve relevant documents and even construct appropriate hierarchies and classifications of terminologies in documents. The authors' experience building the custom model discussed in this paper confirms the difficulties discussed above in the Limitations section. Despite the promise of Cognitive Computing and the model enrichments techniques discussed above, additional research effort will be essential in the future to overcome the dauting challenge of medical Big Data and leverage it as a tool for human welfare.

ACKNOWLEDGMENT

The authors wish to thank IBM CAS group for their assistance with this research. The authors also thank 18th IEEE International Conference on Cognitive Informatics & Cognitive Computing (2019) co-chairs for invitation to submit a revised version of this paper to IJSSCI. Finally, the authors acknowledge the helpful suggestions of the IJSSCI reviewers that have strengthened this work.

REFERENCES

Al-akashi, F. H. A. (2019, July 1). Abstract retrieval over Wikipedia articles using neural network. *International Journal of Software Science and Computational Intelligence*. doi:10.4018/IJSSCI.2019070102

Babylon Health Services. (2020). Retrieved from https://www.babylonhealth.com/product

Callan, J. (2000). Distributed information retrieval. In W. B. Croft (Ed.), *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval* (pp. 127–150). Springer. doi:10.1007/0-306-47019-5_5

Chen, Y., Elenee Argentinis, J., & Weber, G. (2016). IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research. *Clinical Therapeutics*, *38*(4), 688–701. doi:10.1016/j.clinthera.2015.12.001 PMID:27130797

Crouch, H. (2018, April 10). *Babylon expands its AI technology to mainland China*. Digital Health. Retrieved from https://www.digitalhealth.net/2018/04/babylon-ai-technology-china-tencent/

Din, S., Paul, A., Ahmad, A., Gupta, B. B., & Rho, S. (2018). Service orchestration of optimizing continuous features in industrial surveillance using big data based fog-enabled internet of things. *IEEE Access*, *6*, 21582–21591. doi:10.1109/ACCESS.2018.2800758

Duch, W., Adamczak, R., & Grabczewski, K. (2001). A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transactions on Neural Networks*, *12*(2), 277–306. doi:10.1109/72.914524 PMID:18244384

Escher, A. (Ed.). (2013). *Type 1 Diabetes*. InTech. doi:10.5772/45927

Ferrucci, D., & Lally, A. (2004). UIMA: An architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, *10*(3–4), 327–348. doi:10.1017/S1351324904003523

Freedman, D. H. (2017, June 27). *What will it take for IBM's Watson technology to stop being a dud in health care?* MIT Technology Review. Retrieved from https://www.technologyreview.com/s/607965/a-reality-check-for-ibms-ai-ambitions/

Gudivada, A., & Tabrizi, N. (2018). A literature review on machine learning based medical information retrieval systems. In *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 250–257). IEEE Press. doi:10.1109/SSCI.2018.8628846

Gupta, B. B., Agrawal, D. P., Yamaguchi, S., & Sheng, M. (2018). Advances in applying soft computing techniques for big data and cloud computing. *Soft Computing*, *22*(23), 7679–7683. doi:10.1007/s00500-018-3575-1

Hao, T., Chen, X., Li, G., & Yan, J. (2018). A bibliometric analysis of text mining in medical research. *Soft Computing*, *22*(23), 7875–7892. doi:10.1007/s00500-018-3511-4

IBM Cloud Docs. (n.d.). Retrieved from https://cloud.ibm.com/docs/services/discovery-data?topic=discovery-data-create-enrichments

IBM Watson Content Analytics. (2020). IBM Knowledge Center. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS5RWK

Levi Shapiro. (2015, June 30). *Mhealth israel_ibm watson for healthcare startups* [Technology]. Retrieved from https://www.slideshare.net/levshapiro/mhealth-israelibm-watson-for-healthcare-startups

Moscaritolo, A. (2018, July 19). *Va reenlists IBM's Watson in fight against cancer*. PCMAG. Retrieved from https://www.pcmag.com/news/va-reenlists-ibms-watson-in-fight-against-cancer

Murugan, R. (2019). A cloud-based patient health monitoring system using the internet of things. In Handbook of Research on Cloud Computing and Big Data Applications in IoT. Academic Press. doi:10.4018/978-1-5225-8407-0.ch010

Olson, P. (2018, June 28). *This ai just beat human doctors on a clinical exam*. Forbes. Retrieved from https://www.forbes.com/sites/parmyolson/2018/06/28/ai-doctors-exam-babylon-health/

Psannis, K. E., Stergiou, C., & Gupta, B. B. (2019). Advanced media-based smart big data on intelligent cloud systems. *IEEE Transactions on Sustainable Computing*, *4*(1), 77–87. doi:10.1109/TSUSC.2018.2817043

Shadabi, F., & Sharma, D. (2008). Artificial intelligence and data mining techniques in medicine – success stories. In *Proceedings of the 2008 International Conference on BioMedical Engineering and Informatics* (Vol. 1, pp. 235–239). Academic Press. doi:10.1109/BMEI.2008.170

Shelly Palmer Digital Living. (2015, August 1). *Data science advisory*. Retrieved from https://www.shellypalmer.com/data-science/

Stergiou, C., Psannis, K. E., Gupta, B. B., & Ishibashi, Y. (2018). Security, privacy & efficiency of sustainable cloud computing for big data & IoT. *Sustainable Computing: Informatics and Systems*, *19*, 174–184. doi:10.1016/j.suscom.2018.06.003

Wang, Y., Raskin, V., Rayz, J., Baciu, G., Ayesh, A., Mizoguchi, F., & Howard, N. et al. (2018, January 1). Cognitive computing: Methodologies for neural computing and semantic computing in brain-inspired systems. *International Journal of Software Science and Computational Intelligence*, *10*(1), 1–14. doi:10.4018/IJSSCI.2018010101

Ziebell, R.-C., Albors-Garrigos, J., Schoeneberg, K.-P., & Marin, M. R. P. (2019, April 1). Adoption and success of e-hrm in a cloud computing environment: A field study. *International Journal of Cloud Applications and Computing*, *9*(2), 1–27. doi:10.4018/IJCAC.2019040101

Zuccon, G., Koopman, B., & Bruza, P. (2014). 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 - ESAIR '14* (pp. 43–45). Academic Press. doi:10.1145/2663712.2666197

Akhil Gudivada is a former student at East Carolina University with research focused on big data and cognitive computing in the medical domain. Works published include research with using IBM Watson as a leveraging tool to better understand large datasets in the medical domain.

James Philips is a graduate student researcher at East Carolina with research interests in text mining, information retrieval, and software analytics. His prior research has focused on recommender systems.

Nasseh Tabrizi received his B.S. degree in Computer Science from Manchester University, UK. He then completed his M.S. and Ph.D. from Automatic Control and Systems Engineering Department, Sheffield University, UK. Tabrizi worked in Manchester University for two years prior to his appointment at East Carolina University in 1984. He was the Graduate Program Director of Computer Science until 12/15/2019, and founder and director of Software Engineering graduate program at East Carolina University. His research interests are in the areas of machine learning, Big Data analytics, computer vision, software engineering, and computer science education.