Proc. of the Sixth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2018

ISBN: 978-1-63248-149-8 doi: 10.15224/978-1-63248-149-8-19

Natural Language Processing for designing voice assistant for healthcare systems

[Dr.Meghana Nagori, Ashwin Kulkarni, Sumedh Sharangdhar, Kirti Keskar, Sameer Bhale, Dr.Vivek Kshirsagar]

***Abstract*— In the world of having everything in our pocket, we have developed a technology (product) which leverages the power of networking, artificial intelligence, data mining in a single package to revolutionize the traditional way of handling health data. Using the most convenient way of interacting with a device which is voice recognition we are trying to eliminate a range of documents, which leads to minimum human interference in the whole process of healthcare and we are collecting unstructured health data of a consumer (patient). Hence, by applying techniques of Natural Language Processing we are analyzing and extracting meaningful data from the narrated input. Our algorithm will enable clinicians to retrieve the information of a patient in the most precise format. Our product facilitates one another part of recognizing consumers' health-related need and suggest suitable, trusted and predicted (easily available) solution for it using well organized pharmaceutical data set. For a massive data set (history) this algorithm can be further implemented on block-chain technique.**

***Keywords*—Networking, Artificial Intelligence, Data Mining, Voice Recognition, Healthcare, Natural Language Processing.**

I. Introduction

The traditional and most general methodology to collect all the data from the patient is by interacting with him/her and to carry this data wherever needed is via documentation which includes paperwork for reports, test results, discharge papers, etc. It is observed that the interaction session, every time with a clinician consumes around 10 minutes for the old patient and 20 minutes for the new patients. Additionally, it is a time-consuming process when a clinician is changed. Carrying all the documents whenever needed is not feasible and sometimes due to lack of time, clinicians may skip some important data from vast dataset related to a patient. This way of data arrangement is not handy. In order to overcome those complications, there is a need for some simple yet specific and appropriate approach towards this. Out of all, the best way we could find is the ***unstructured textual data*.** Which can be captured/written by using the conversation between patient and doctor. Stated that it should be in electronic/digital form. Since for the clinicians, textual data is the simplest understandable format possible.

[Dr.Meghana Nagori, Ashwin Kulkarni, Sumedh Sharangdhar, Kirti Keskar, Sameer Bhale, Dr.Vivek Kshirsagar]

Government College of Engineering, Aurangabad, Maharashtra,INDIA.

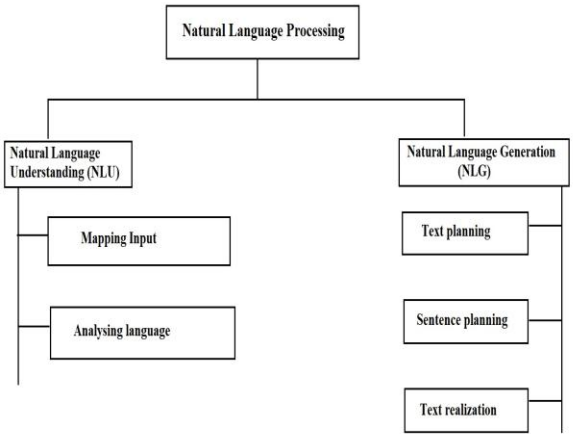

Figure 1. Components of *NLP*

They do not necessarily have thorough knowledge regarding data management software or any kind of technicality. Simply by reading, they are good to go.

The main problem is to find how likely computer systems are capable of generating such unstructured data. The best approach *is a **N**atural **L**anguage **P**rocessing **(NLP)***. ***NLP*** is a method of ***A**rtificial **I**ntelligence* which is used for communicating with computer systems by using natural language such as English. ***NLP*** helps computers to analyze, understand and derive meaning from human language in a useful and a smart way [1]

NLP has mainly two components as shown in figure:

1) ***N**atural **L**anguage **U**nderstanding **(NLU)***
2) ***N**atural **L**anguage **G**eneration **(NLG)***

N**atural **L**anguage **U**nderstanding:*** It mainly deals with analyzing the input text and machine reading comprehensions. It does the disassembly and parsing of the input. It is considered as an AI-hard problem. Since there is a difficulty in solving such problems and which needs the system to be as intelligent as people.

N**atural **L**anguage **G**eneration:* It deals with the production of meaningful phrases and sentences in the form of natural language [2]. ***NLG involves,

1) *Text Planning*: which deals with retrieving the relevant contents from the base of knowledge.
2) *Sentence Planning*: which includes choosing the required words and forming meaningful phrases.

3) *Text Realization*: it includes the mapping of the sentence plan into sentence structure.

Proc. of the Sixth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2018

ISBN: 978-1-63248-149-8 doi: 10.15224/978-1-63248-149-8-19

A. *History of NLP*:

In the early 1980s, the computational grammar became a popular area of research. The *NLP* was growing rapidly. The increase in computer power is what allowed this transformation to take place. There were so many new and advanced systems coming out and which helped in the growing of *NLP*. The popular projects such as *SHRDLU* [3] which dealt with rearranging blocks, pyramids by user input. *SHRDLU* would understand sentences like: 'put the red cube on top of the blue cube' and carry out that action in the real world. Then in 1982, the concept of a chatbot was created and the project Jabberwacky began. The project was intended to create a user-friendly and entertaining chat interface. Later In the early 1990s, the *NLP* was growing faster than ever. It was the era when the Internet was flooding. There was lots of a research carried out regarding information extraction and automatic summarizing. Canada started doing research regarding machine learning at this time. Since there were so many advances in this field *US* government started encouraging such AI-based systems which had a less reliance on database systems.

B. *Steps in NLP:*

NLP involves general five steps. *Lexical Analysis* involves identifying and analyzing the structure of words. *Syntactic Analysis* involves the analysis of sentence against grammar and arrangement of all the words in such a way that it shows the relationship between those words. *Semantic Analysis* draws the dictionary meaning from the text. The text is checked for meaningfulness [4]. *Disclosure Integration* brings the meaning of immediately succeeding sentence. *Pragmatic Analysis* involves re-interpretation of what was said into what it actually meant.

Moreover, to develop the *NLP* systems there are wide ranges of libraries available, out of which we are using *Stanford NLP*. Coming back to the problem of healthcare, this paper represents the system for successful and meaningful extraction of data using *NLP*. But, to use *NLP* tools we need the data In unstructured form. To capture the data from the conversation of patient and doctor we have used *Text-to-speech* feature.

II. Related work

The grip of *NLP* on computing is increasing day by day. In 2011, everybody was working on creating generalized structure but now we are in a phase of utility building and increasing efficiency of the application in *NLP*. Work such as interpreting genetic test result faster with greater accuracy by *IBM WATSON HEALTH DIVISION*, rationalizing neural predictions, teaching machine to describe image via natural language feedback and much more has already done. In this paper, we are focusing on healthcare domain which didn't get the spotlight as of now. *NLP* is being proved as an effective tool for creating utilities which will help human to move forward with greater capabilities and innovation, people are finding answers to the question related to the neural network using *NLP* and also they are successfully able to teach a machine to describe the image via natural language feedback. Additionally, they are able to make a machine communicate with the creator using labels and attributes (by providing certain dataset). We are using *Stanford NLP* to solve the neural problem in the simplest way. One in all we taught a machine to learn by itself and making it smarter to solve existing problems efficiently. They can be more data efficient than other models. Attempts like "supervisor trying to teach machine visual concepts" have been made and they get expected results to some extent by teaching machine vocabulary of attributes related to the certain domain. All those models are useful for solving many *NLP* problems improving both visualization, and interpretability along with greater accuracy. ("There's this huge amount of data in the healthcare space, and we need to find the best ways to extract what's relevant.") Finally, the goal is to make a machine which represents unstructured text in the simplest way so that it will be beneficial to reduce the time consumption and help both doctor and patient to systematically maintain the health record. To improve clinical data integrity and retrieve it in the quickest way possible along with the highest accuracy. Existing *NLP* based products like Amazon's Alexa and many automated customer service application are using Text-to-Speech feature but it has not been implemented in healthcare domain as of now. We are trying to use such features in our product so that it becomes more convenient to use.

III. Proposed Methodology

The real idea behind this system is for healthcare domain. The Consumer should enter his symptoms in an unstructured form. The information given by consumer should be precise and relevant. This data is processed using Stanford *coreNLP*. The ability to analyze and extract meaning from narrative text or other unstructured data sources is a major part of recognizing patient's need and giving results according to that. The result is displayed after internal analysis by *coreNLP*. As stated earlier the *Stanford CoreNLP* [5] has maximum simplicity while tackling the situation of our research problem. We are using quantitative as well as qualitative approach in this project. This approach fits the whole situation because we need dynamic and unstructured health data of a consumer which can be sometimes irrelevant as well, but our system will extract quantitative data using semantic actions and will make it structured. For more user convenience we are leveraging voice recognition to capture user data, as typing in data is not possible in every situation. For the first input, we collect data from the user through voice recognition when he/she is narrating his/her medical particulars to the clinicians. This data is completely unstructured, hence this voice is converted to text and there is need of text mining to extract meaning from those sentences. The next task is done using natural language processing where all the tasks of *NLP* are carried out which is done in two levels as low-level and high-level *NLP* tasks.

Proc. of the Sixth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2018

ISBN: 978-1-63248-149-8 doi: 10.15224/978-1-63248-149-8-19

A. *Low-level NLP tasks*:

1. Sentence Segmentation: Segmentation means detecting a boundary of every sentence which is determined by the location of a full stop(.) in it. But,this task becomes exhausting when comes to a situation like abbreviation as (Dr.), (i.e.), etc.
2. Tokenization: It can also be called as lexical analysis as every lexeme is extracted within the sentence here. Tokens often obtained in healthcare domain uses the hyphen, slashes in prescription as a dose of 10mg/day.
3. Part-of-speech tagging: This task labels tokens with their part-of-speech tags as to map semantics of that sentence.
4. Morphological decomposition: This task is specifically used for breaking compound words. This mainly includes lemmatization i.e. reaching to the root meaning of the word since there are many complex and multi-meaning terms in medical.

B. *High-level NLP tasks:*

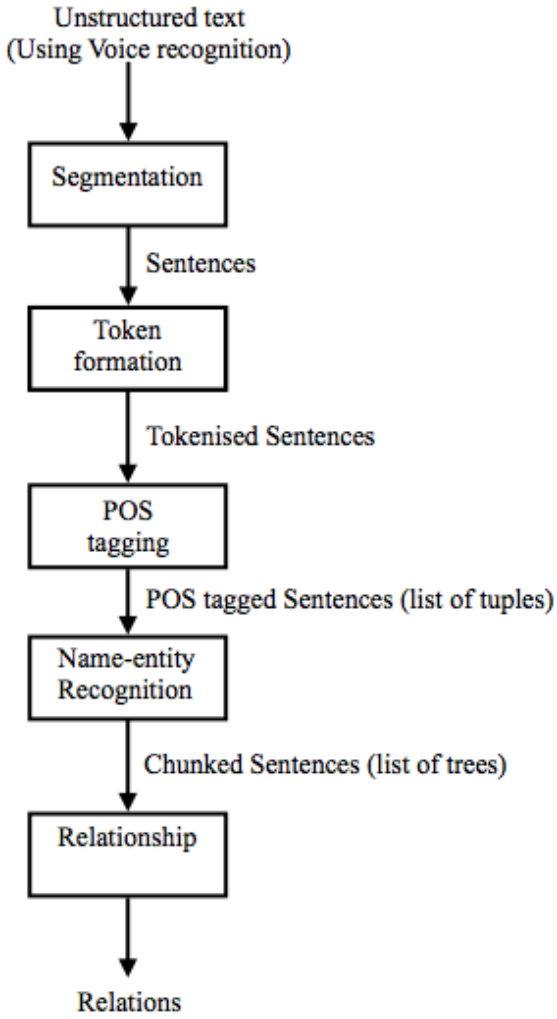

Figure 2. *NLP* Tasks

Name Entity Recognition(NER): This is a very arduous job in NLP. Identifying specific word as an entity and categorizing them as a person, institution, medicine, disease, location, etc. the task is to map these entities with the meanings in vocabulary as a noun, adjective, etc. Thus, we get a name entity relation in a sentence for further processing. This step is called as preprocessing to get semantic of a line which will be helpful in extracting exact results to the queries related to the patient. We faced some issues while performing low-level and high-level NLP tasks, such as:

1. Word order variation: changing order of words can change the meaning of a whole sentence, we need to check interrelations of words as well.
2. Relationship extractions: In this type of relationships between different words of a sentence as well as the relationship between different sentences of a paragraph is observed and then extracting appropriate information from it, is executed. After performing the steps stated above on the user provided unstructured input data, we need to store it in the form of structured data. We've used *MySQL* database to store the structured data. The data is stored in the form of *hash-maps* and can be used in O(1) time. The query given by a clinician to know anything about the patient is in the unstructured form and given using voice recognition. Thereafter, for processing that query we've used *longest matching subsequence* method to give the exact answer to that query. The whole algorithm is explained in the '*Algorithm*' section.

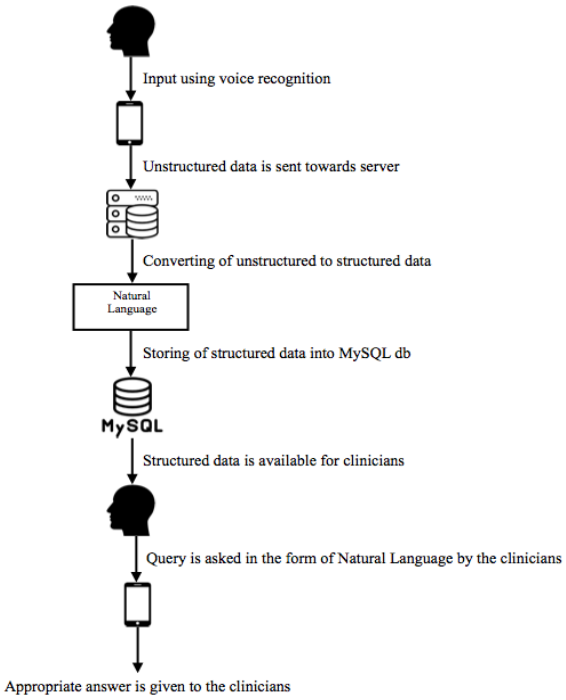

Figure 3. Proposed System

Proc. of the Sixth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2018

ISBN: 978-1-63248-149-8 doi: 10.15224/978-1-63248-149-8-19

spoken languages into text by computers. *Hidden Markov Models (HMM)* are used in speech recognition techniques which used to give exact text translation of voice input given by a consumer. The unstructured data then sent to a server for processing. Because we are using Stanford *CoreNLP* libraries which are quite big for any in-application data, hence we are performing our whole algorithm on the *MySQL* server. The whole unstructured data is processed using the *NLP* steps mentioned above and then structured data is saved into *MySQL* database using hash keys, the hashing is the most relevant technique we found during the whole process hence we finalized that thing. After *NLP* processing for converting unstructured to structured health data, it is available for the clinicians to perform any query on it (Which will also be taken through voice recognition in the unstructured format). By using our algorithm which is a mixture of *Hidden Markov Model (HMM)* and *Longest Common Subsequence (LCS*)[6] (As explained further in algorithm section) we are giving appropriate results to the clinicians.

IV. Algorithm

The foundation stone on which a computational software system works are the algorithms involved. By considering space and time complexities we have prepared almost efficient algorithm for this system. There exist two parts of the process: (1) Taking unstructured input from a consumer and then processing it using *NLP* and make it structured. (2) By using that structured information retrieving appropriate information for the clinicians.

$$x_i = y_j \Rightarrow C[i,j] = C[i-1,j-1] + 1 \quad \ldots.(1)$$

$$x_i \neq y_j \Rightarrow C[i,j] = max(C[i-1,j], C[i,j-1]) \quad \ldots.(2)$$

The first part consists voice recognition techniques (different computing mobile devices might have different sensors) to take an unstructured input from a consumer and then processing of it using Stanford *CoreNLP* library.

A. *Pipelining :*

Any NLP tasks should perform the whole algorithm using diving the whole data into subparts, and we know that pipelining handles concept of subparts effectively. Because of very big health document of a consumer, dividing it and then mixing is the best way to perform the algorithm. The *POS* tagger is used to divide all the sentences using part of speech rules of an English language. The intention behind pipelined architecture is that it follows *XML* based technology. *Unstructured Information Management Architecture (UIMA*)[7] was also on our list but its scope is now beyond *NLP*. Since our preferred language of programming is *JAVA*, we have used Stanford *CoreNLP* instead of Natural Language

Figure 4. Name Entity Recognition

Toolkit. The non*NLP* task has perfect accuracy, but if there exists an error in one of the sentences of a pipelined architecture when the process moves to next step and so on, accuracy degrades at each level. With using multiple branching this issue can be resolved. We are using multiple branching techniques at some stages where input data is more complex.

B. *Name Entity Recognition*

The process of name entity recognition is the task of dividing the whole data into particular entities[8]. e.g., in the Fig. 4, the algorithm correctly finds the person name, integral/ordinal part of the sentence and the location. There are many types of entities such as date, time, integer, person, location, etc. After processing takes place using the above tasks, the algorithm finds basic dependencies from the sentence. Like which is a noun, which is adjective, etc. After the strong data in the structured format on the database, we then take inputs from the clinicians using voice recognition technique in the unstructured form as well. Then, by processing the similar tasks as explained above, we extract the exact meaning of a query given by the clinician [9].

Here the use of *Longest Common Subsequence* comes into the picture. By checking the sentences of the consumers' input and by checking the sentences given by clinician where longest match among the words happens, we give the most appropriate results irrespective of sentence jumbling, capital, and small alphabets, etc. The complexity of an algorithm is O (mn) where m is the length of one sentence and n is the length of one query of a clinician.

The algorithm is as follows: If there's a match, it increases the count by 1. The difference between typical *LCS* and our algorithm is that it saves the occurrence. This makes the answer formation easy. Finally, the algorithm works efficiently up to 28,000 to 38,000 words given by a consumer. Simultaneously, we are providing a simplified predicted solution for it using well organized pharmaceutical data set such as "data.gov" [10]. For a massive data set (history) this algorithm can be further implemented on block-chain technique.

Conclusion and Future scope

We propose to develop an accurate model for predicting personalized healthcare data with a voice assistant and its components e.g. speech recognition etc. Given the growing complexity of medical decision making our system may be useful for facilitating effective decision making in critical

Proc. of the Sixth Intl. Conf. Advances in Computing, Communication and Information Technology- CCIT 2018

ISBN: 978-1-63248-149-8 doi: 10.15224/978-1-63248-149-8-19

cases due to ease in availability of data. In future, block-chain technology can be adapted for storing the patient's entire history since birth and for accessing such massive datasets optimized computations can be achieved in the distributed environment.

Acknowledgement

We extend our thanks to our principal Dr.P.B.Murnal for his encouraging words and provision of resources for conducting the research.

References

[1] Fred Popowich, "Using Text Mining and Natural Language Processing for Health Care Claims Processing", SIGKDD Explorations,Volume 7, Issue 1,pp 59-66

[2] José M. Alonso, Alberto Bugarín, "Natural Language Generation with Computational Intelligence", IEEE Computational Intelligence Magazine ,Volume: 12, Issue: 3, Aug. 2017,pp 8-9

[3] Understanding Natural Language, Academic Press, 1972

[4] Alfred V. Aho and Jeffrey D. Ullman,1972, "The theory of parsing, translation, and compiling",Prentice-Hall, Inc.

[5] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, David McClosky, "The Stanford CoreNLP Natural Language Processing Toolkit"

[6] Kang Ning1 , Hoong Kee Ng2 and Hon Wai Leong2, "Analysis of the Relationships among Longest Common Subsequences, Shortest Common Supersequences and Patterns and its application on Pattern Discovery in Biological Sequences"

[7] David Ferrucci, Adam Lally, "UIMA: An architectural approach to unstructured information processing in the corporate research environment", Natural Language Engineering, Vol(10),pp 327–348.

[8] Anne Kao & Steve Poteet, "Text Mining and Natural Language Processing – Introduction for the Special Issue", SIGKDD Explorations. Volume 7, Issue 1, 1-2

[9] José M. Alonso, Alberto Bugarín, "Natural Language Generation with Computational Intelligence", IEEE Computational Intelligence Magazine ,Volume: 12, Issue: 3, Aug. 2017,pp 8-9

[10] data.gov.uk/dataset/general_pharmaceutical_services

About Authors

Dr. Meghana Nagori works as an assistant professor in computer science and engineering department at Government Engineering college, Aurangabad, Maharashtra in India since 2002. She has supervised over thirty postgraduate dissertation titles. Her research interests lie in the domain of big data, algorithms and optimization, medical informatics, recommender systems, natural language processing.

Ashwin Kulkarni is pursuing his bachelor's of computer science and engineering from Government Engineering College, Aurangabad, Maharashtra in India and is currently a final year student in the department.

Sumedh Sharangdhar is pursuing his bachelor's of computer science and engineering from Government Engineering College, Aurangabad, Maharashtra in India and is currently a final year student in the department.

Kirti Keskar is pursuing her bachelor's of computer science and engineering from Government Engineering College, Aurangabad, Maharashtra in India and is currently a final year student in the department.

Sameer Bhale is pursuing his bachelor's of computer science and engineering from Government Engineering College, Aurangabad, Maharashtra in India and is currently a final year student in the department.

Dr.Vivek Kshirsagar is associate professor and head of computer science and engineering department at Government Engineering College, Aurangabad, Maharashtra in India since 2010. He has supervised over forty postgraduate dissertation titles. His research interests lie in the field of Networking and Security and big data analytics.