International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Development of Structured Data from Unstructured Homeopathic Case Sheet Documents Using Natural Language Processing Techniques

Savitha Shetty[1], Sarika Hegde[2], Saritha Shetty[3]*

Submitted:12/03/2024 **Revised**: 27/04/2024 **Accepted**: 04/05/2024

Abstract: Homeopathic physicians generally maintain the records of patients in the form of text documents, which contain valuable clinical information but lacks the standardized structure. Generation of structured data from unstructured homeopathic case sheet documents is a challenging process. The goal of the study is to identify and extract the key-value pairs from such unstructured documents to generate structured data that can be utilized for analysis and decision making in homeopathic practice. The paper outlines the methodology for collecting homeopathic case sheet documents, extracting relevant information and structuring it into a format suitable for further analysis. The methodology described in this study has been successfully implemented on a dataset comprising of 300 homeopathic case sheet documents. For each of these documents, our approach successfully extracted key-value pairs, representing various clinical parameters such as details of presenting complaints, family history, final diagnosis etc. The key-value pairs were structured and organized within excel spreadsheet, facilitating further analysis and interpretation of extracted clinical data. Evaluation shows that out of 300 documents three documents could not be properly converted to the structured form due to spacing constraints hence accuracy is 99%.

Keywords: *homeopathic case sheets, information extraction, key value pairs, structured data*

1. Introduction

Homeopathic medicine relies heavily on individualized treatments based on detailed patient histories and symptom descriptions [1]. The documentation of these records are called as homeopathic case sheet documents. These case sheet documents contain lot of information about the patients in unstructured form [2] and cannot be processed as it is. The case sheet document contains unstructured narratives, making it challenging to extract actionable insights from the data. The paper aims to bridge this gap by proposing a method to convert unstructured homeopathic case sheet documents into structured data, facilitating more efficient analysis and treatment planning.

The transformation of unstructured data into structured formats has become increasingly critical in healthcare research. In homeopathy, case sheet documents often contain rich clinical information presented in an unstructured manner, posing challenges for data analysis and research. This paper addresses the development of methodologies to convert unstructured homeopathic case sheet documents into structured data, enhancing accessibility and usability. By employing natural language processing (NLP) [3] techniques, the proposed approach aims to systematically extract and organize key clinical details. The structured data can significantly raise the effectiveness of homeopathic practice, facilitating better patient management and evidence-based research. Ultimately, this work contributes to the broader goal of integrating traditional medical practices with modern data analytics.

2. Literature Survey

The unstructured clinical narratives in EHRs contain critical patient information such as medical problems, treatments, and diagnostic tests. The NLP techniques are utilized to convert this unstructured data into structured formats, making it usable for various clinical applications [4]. The authors utilized a private database of 10,000 anonymised health records including 234,673 clinical notes. These notes were broken into 1,183,345 distinct sentences for retraining BioBERTptRT. Furthermore, a hand labeling of 100,021 sentences was conducted in order to refine the models. This extensive data collection and preprocessing enabled the effective training and evaluation of the proposed model [5]. Authors collected diverse health data formats, including unstructured, semi-structured, and structured data. Utilizing an ontology-based approach, they structured the collected data for efficient processing and retrieval in MongoDB, highlighting the importance of schema design for NoSQL databases [6].

The authors addresses the challenge of data scarcity in the medical field by leveraging generative model like GANs to produce synthetic data. The classification accuracy is more stable while using the smaller datasets. These synthetic datasets are then integrated with real data to train binary classifiers, aiming to enhance classification accuracy, because here only a small amount of original information is

[1] *NMAMIT (Nitte Deemed to be University), Nitte-574110, INDIA*
ORCID ID: 0000-0003-1295-2562
[2] *NMAMIT (Nitte Deemed to be University), Nitte-574110, INDIA*
ORCID ID: 0000-0002-8781-9451
[3] *NMAMIT (Nitte Deemed to be University), Nitte-574110, INDIA*
ORCID ID:0000-0001-5575-5134
* *Corresponding Author Email: shettysaritha1@nitte.edu.in*

accessible [7]. The author addresses about the organization of information from documents written in natural languages like English for various informational tasks. Author focuses on the methods for analyzing and formatting textual data to create structured databases. This process aims to enhance data retrieval and management, particularly in science information and medical records [8]. Medical institutes use EMR (Electronic Medical Records) to record patient conditions, but these records often have issues like diversity, incompleteness, and redundancy. Preprocessing is necessary to improve data quality for effective data mining, involving techniques like data cleansing and integration for structured data. For unstructured medical texts, complex methods are required [9].

The authors analyzed 20 electronic medical-record systems, categorizing them into 'classical' and 'experimental' systems. They identified three main challenges: enhanced data presentation, clear data understanding, and direct data entering. Dynamic data-input forms and free-text input with natural language interpretation are two promising techniques,highlighting the need for further research on optimal search structures for medical narratives. The measured performance of the pipeline achieved F1 score from 0.690 to 0.995 for various elements [10].

3. Problem Statement

The primary challenge addressed in this research is the identification and extraction of key-value pairs from unstructured homeopathic case sheet documents. These documents typically contain patient medical histories, description of symptoms, details of complaints, treatment histories, family histories etc but lacks a standardized format for data extraction. The simple block diagram of the extraction process is as shown in the Fig 1. The input to the work is homeopathic case sheet documents and the output is the structured data in the excel spreadsheets.

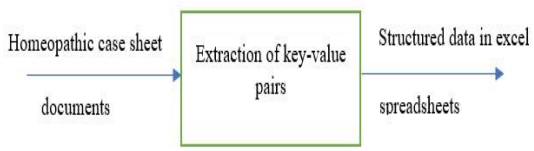


Fig. 1. simple block diagram of the process

4. Objectives

- Collection of homeopathic case sheet documents from the hospital.
- Development of algorithm for the extraction of key-value pairs from unstructured clinical documents.
- Evaluation and generation of structured data from the extracted information.

4.1. Data collection and preprocessing

The homeopathic case sheet documents are collected from the hospital. The sequence of steps followed during the data collection are shown as a flowchart below in Fig. 2. The steps included are as below:

Step 1: The homeopathic case sheet documents are collected from a homeopathic hospital.

Step 2: Physicians in the hospital are consulted during the document collection for scanning the case sheet documents from the hospital.

Step 3: Each of the case sheet document is scanned and stored in the PDF(Portable Document Format) format. The case sheet documents are written using pen by the physician in is own handwriting on the paper.

Step 4: Each of the PDF document is converted to docx format for further processing.

Step 5: Each of the docx file is the descriptions of each of the patient, list of docx files are stored in a directory for further processing.

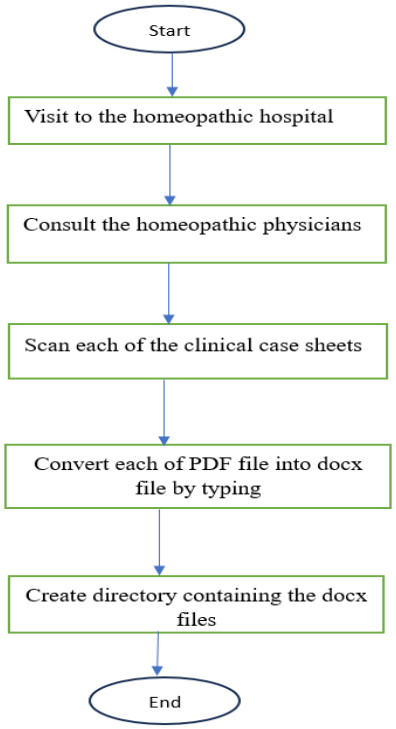


Fig. 2. Flow diagram of the data gathering process

4.2. Information extraction and data structuring

The collected case sheet documents are present in the hospital template format as shown below. Fig. 3 indicates the sample input collected from the hospital in the original handwritten text form and the Fig. 4 indicates the typed docx form.

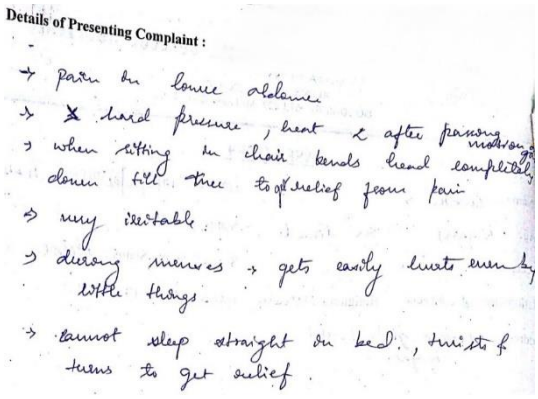


Fig. 3. Sample input in original handwritten text form

Details of Presenting Complaint:

pain in lower abdomen.
Hard pressure, heat less after passing motion.
When sitting in chair bends head completely down till knee to get relief from pain.
Very Irritable.
During menses easily gets hurts even by little things.
Cannot sleep straight in bed, twists and turns to get relief.

Fig. 4. Sample input in typed docx form

The Fig. 3 indicates the handwritten free text notes of the case sheet document. The "Details of presenting complaints" is one of the key and value is few sentences. The Fig. 4 indicates the typed text from the handwritten text. The data present in the homeopathic case sheet document is identified as key: value pair, which should be extracted to convert the unstructured data into the structured format. The algorithm is written for the conversion of directory containing docx files into structured excel spreadsheet format comprising of rows and columns.

Algorithm:

Initialize empty list as 'details' which will contain the list of directories, where each dictionary represents extracted information from word document.

for each 'docx' files in specified directory do

Read content of each word document

Set maximum number of words allowed for key as five.

Initialize 'last_key' variable to store last encountered key in the document

Initialize empty dictionary 'info' to store extracted information from the document

for each paragraph of document do

Extract text from each paragraph

if paragraph is empty or contains only whitespace characters then

Skip processing of this paragraph

Move to the next paragraph

end if

Split the text of paragraph into chunks using ':' as delimiter

if len(chunks)>1 and len(chunks[0].split())<=max_word_for_key then

Extracts the key and converts it to lowercase

Update the last_key with the current key

Join remaining chunks and store them in 'info' dictionary with the key extracted

else

Paragraph continues, append the paragraph text to the existing value for last_key in the 'info' dictionary

end if

end for

if the document contains any tables, then

Select first table in the document

Initialize the observations list

Initialize the prescriptions list

for each row of the table do

Extracts the text content of each cell in the current row

if content of row is non-empty then

Appends first cell contents to observations list

Appends second cell contents to prescriptions list

Concatenate observations and prescriptions list

Store the concatenated value in the 'info' dictionary under key 'table'

end if

end for

Append contents from 'info' dictionary to 'details' list

end if

end for

Create data frame from the list of dictionaries

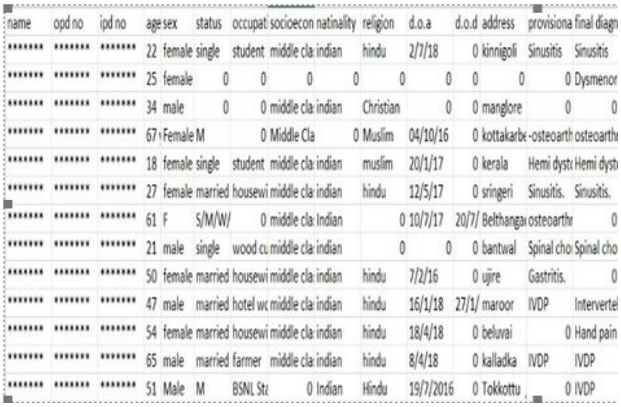
Write data frame into an excel file

The algorithm processes multiple Word documents in a directory to extract structured information. It starts with an empty list called "details" to store the extracted data from each document. For each document, it reads the content and uses up to five words to identify keys. It tracks the last key

found and stores the extracted data in a dictionary. If a document has tables, the algorithm extracts text from the first table's rows and appends this data to observations and prescriptions lists. Finally, the algorithm converts the details list into a DataFrame and writes it to an Excel file, organizing the extracted information systematically.

5. Results

After applying the algorithm on the collected data the structured excel spreadsheet document is generated which is shown in the Fig. 5. below.



name	opd no	ipd no	age	sex	status	occupati	socioecon	natinality	religion	d.o.a	d.o.d	address	provisiona	final diagn
*******	*******	*******	22	female	single	student	middle cla	indian	hindu	2/7/18	0	kinnigoli	Sinusitis	Sinusitis
*******	*******	*******	25	female	0	0	0	0	0	0	0	0	0	Dysmenor
*******	*******	*******	34	male	0	0	middle cla	indian	Christian		0	0	manglore	0	0
*******	*******	*******	67	Female	M		0	Middle Cla	0	Muslim	04/10/16	0	kottakarb	osteoarth	osteoarth
*******	*******	*******	18	female	single	student	middle cla	indian	muslim	20/1/17	0	kerala	Hemi dyst	Hemi dyst
*******	*******	*******	27	female	married	housewi	middle cla	indian	hindu	12/5/17	0	sringeri	Sinusitis.	Sinusitis.
*******	*******	*******	61	F	S/M/W/		0	middle cla	Indian		0	10/7/17	20/7/	Belthanga	osteoarth	0
*******	*******	*******	21	male	single	wood cu	middle cla	indian		0	0	0	bantwal	Spinal cho	Spinal cho
*******	*******	*******	50	female	married	housewi	middle cla	indian	hindu	7/2/16	0	ujire	Gastritis.	0
*******	*******	*******	47	male	married	hotel wc	middle cla	indian	hindu	16/1/18	27/1/	maroor	IVDP	Intervertel
*******	*******	*******	54	female	married	housewi	middle cla	indian	hindu	18/4/18	0	beluvai		0 Hand pain
*******	*******	*******	65	male	married	farmer	middle cla	indian	hindu	8/4/18	0	kalladka	IVDP	IVDP
*******	*******	*******	51	Male	M	BSNL Str	0	Indian	Hindu	19/7/2016	0	Tokkottu		0 IVDP

Fig. 5. structured excel file displaying the docx file contents in each row

Fig. 5. shows the excel file in which each case sheet document is shown as a separate row of the excel spreadsheet document. Here the column label indicates the "key" of the "key: value" pairs extracted from the case sheet document. Each row cells indicate the contents of the "key: value" pairs. The column labels "name", "ipd no", "opd no" are deidentified by mentioning it as "*******" to protect patient privacy while ensuring the data remains useful for research and public health purposes. Few of the fields are written are zero because the entry is missing in the case sheet document. Part of excel file is shown in Fig. 5.

The results demonstrate the effectiveness of proposed method in accurately extracting "key: value" pairs from unstructured homeopathic case sheet documents. The structured data generated enables the systematic analysis and visualization of patient records, facilitation the evidence-based decision making in homeopathic documents.

6. Evaluation

The evaluation of the work is done manually by checking whether the content of each key:value pair of each case sheet document is properly entered in the generated excel spreadsheet. The Table 1 indicates the evaluation results for 300 documents along with the accuracy. It is observed that for 3 of the documents out of 300 documents few entries are not done properly due to more spacing issues in the original document.

Table 1. Evaluation results for 300 documents

Total number of clinical documents considered	Total number of documents (few key:value pairs of the document) not converted to structured form due to spacing constraints	Accuracy
300	03	0.99

7. Conclusion

This research paper presents an approach to develop the structured data from unstructured homeopathic case sheet documents. By utilizing the algorithm, the valuable clinical information is extracted and organized into structured format, enhancing the usability and accessibility of 300 homeopathic patient records for the healthcare professionals. The suggested method's efficacy is assessed manually by checking the each key:value pairs of each patient record with the generated structured data in excel spreadsheet. It is observed that the accuracy is 0.99 because few of the key:value pairs in the three of the documents are not converted to structured form properly due to the spacing issues in the original document.

Acknowledgements

We sincerely thank our college for supplying the required materials.

Conflicts of interest

Conflicts of interest are not disclosed by the writers.

References

[1] Wilhelm, M., Hermann, C., Rief, W., Schedlowski, M., Bingel, U., & Winkler, A. (2024). Working with patients' treatment expectations–what we can learn from homeopathy. Frontiers in Psychology, 15, 1398865.

[2] Gupta, S. K., Basu, A., Nievas, M., Thomas, J., Wolfrath, N., Ramamurthi, A., ... & Singh, H. (2024). PRISM: Patient Records Interpretation for Semantic Clinical Trial Matching using Large Language Models. arXiv preprint arXiv:2404.15549.

[3] Hsu, J. C., Wu, M., Kim, C., Vora, B., Lien, Y. T., Jindal, A., ... & Wu, B. (2024). Applications of advanced natural language processing for clinical pharmacology. Clinical Pharmacology & Therapeutics, 115(4), 786-794.

[4] Gao, Y., Mahajan, D., Uzuner, Ö., & Yetisgen, M. (2024). Clinical natural language processing for secondary uses. Journal of Biomedical Informatics, 150, 104596.

[5] de Oliveira, J. M., Antunes, R. S., & da Costa, C. A. (2024). SOAP classifier for free-text clinical notes with domain-specific pre-trained language models. Expert Systems with Applications, 245, 123046.
[6] Sen, P. S., & Mukherjee, N. (2024). An ontology-based approach to designing a NoSQL database for semi-structured and unstructured health data. Cluster Computing, 27(1), 959-976.
[7] Abedi, M., Hempel, L., Sadeghi, S., & Kirsten, T. (2022). GAN-based approaches for generating structured data in the medical domain. Applied Sciences, 12(14), 7075.
[8] Sager, N. (1978). Natural language information formatting: the automatic conversion of texts to a structured data base. In Advances in computers (Vol. 17, pp. 89-162). Elsevier.
[9] Sun, W., Cai, Z., Li, Y., Liu, F., Fang, S., & Wang, G. (2018). Data processing and text mining technologies on electronic medical records: a review. Journal of healthcare engineering, 2018.
[10] Tange, H. J., Hasman, A., de Vries Robbé, P. F., & Schouten, H. C. (1997). Medical narratives in electronic medical records. International journal of medical informatics, 46(1), 7-29.