Chapter 73
Extractive Summarization of EHR Notes

Ajay Chaudhary, Merlin George, and Anu Mary Chacko

1 Introduction

EHRs are records of the patient's health information that contains both structured and the unstructured information. Clinical notes (unstructured information) within EHRs are a rich source of data where detailed information about the patient's medical history and clinical care process is documented. The increased adoption of electronic health records (EHRs) has led to an unprecedented amount of patient health information stored in electronic format. Physicians at the point of care find it difficult to review all the information due to the abundance of notes within the patient's EHR and the time constraint inherent in the clinical setting. The use of templates and copy-paste has introduced unnecessary or redundant data into clinical notes, worsening the problem of information overload and note bloat. Intelligent clinical decision support automatically processing and producing person specific inference for providing personalized smart care are becoming pressing need. The first step in this direction will be automatic summary generation of EHR [1].

Automated summary generation algorithms can help in organizing and synthesizing patient's medical history. This paper reviews the different approach for automatic summarization and explores the notion of using automatic summarization of discharge summary notes to improve clinical experience.

A. Chaudhary · M. George (✉) · A. M. Chacko
National Institute of Technology, Calicut, India
e-mail: merlin_p170080cs@nitc.ac.in

A. Chaudhary
e-mail: chaudharyajay_m180077cs@nitc.ac.in

A. M. Chacko
e-mail: anu.chacko@nitc.ac.in

M. Dave et al. (eds.), *Proceedings of the International Conference on Paradigms of Computing, Communication and Data Sciences*, Algorithms for Intelligent Systems,
https://doi.org/10.1007/978-981-15-7533-4_73

1.1 Types of Summarization

Pivovarov et al., 2015 [2] classifies summarization into two main types:

1. Extractive summarization—created by borrowing phrases or sentences from the original input text.
2. Abstractive summarization—new text generated by synthesizing the original text.

1.2 Main Challenges in Summarization

Some of the unsolved challenges in summarization are listed below.

1. Identification and aggregation of similar information.
2. Reduction information to the most salient
3. Usage of existing domain knowledge
4. Evaluation of summarization techniques

In this work, a tool was developed to do the extractive summarization of an EHR note. The tool highlights the important content in the note by scoring each sentence in the note. The rest of the paper is structured as follows. Section 2 discusses related work. Section 3 explains our proposed work and explains the prototype designed as a proof of concept. Limitations of current work and scope for future work are explored in Sect. 4. Section 5 concludes the paper.

2 Related Work

Dabek et al., 2017 [3] use structured data for building a timeline-based framework to effectively aggregate and summarize the clinical data of a patient enclosed within an EHR. The timeline is formed by summary nodes where each summary node corresponds to an encounter or a date range. Each summary node contains different data elements of the patient's EHR like laboratory test results, vitals, medication, clinical note, radiology note, etc.

Alsentzer et al., 2018 [4] provide an upper bound on extractive summarization of discharge notes but the low recall rate of 0.431 suggests that to recreate a discharge summary with extractive summarization alone is difficult. They also developed an LSTM model to sequentially label topics of history of present illness notes, a narrative section in the discharge summary that describes the patient's prior history and current symptoms, and achieved an F1 score of 0.876, which indicates that this model can be employed to create a dataset for evaluation of extractive summarization methods. They also claim that there is no literature on extractive or abstractive EHR summarization using neural networks.

Liang et al., 2019 [5] proposed an automated system for disease-specific extractive summarization on a single clinical note. It describes a clinical note processing pipeline that includes a basic NLP processing layer as well as additional HER-specific components such as note section classification, and disease context identification. Their model was limited to only two specific diseases hypertension and diabetes mellitus for which they achieved F-scores of 0.657 and 0.679, respectively. The F-scores were obtained by intrinsic evaluation where they compare the generated summary against the ground truth created by physicians. There future plan includes evaluation of their system by quantitative extrinsic measures to evaluate the usefulness of the system generated summaries for practicing physicians at the point of care.

Si et al., 2018 [6] proposed a frame-based natural language processing (NLP) method that extracts cancer-related information from clinical narratives. In frame semantics, the words or phrases that trigger the frame are known as lexical units (LU), and the frame elements are described as the participants or roles of a frame that defines the characteristics or attributes associated with the lexical units. They focus on three frames: cancer diagnosis, cancer therapeutic procedure, and tumor description. An expert physician devised a list of lexical units (trigger phrases) and corresponding elements (attributes) for each frame. They have utilized a deep learning-based approach, bidirectional long short-term memory (LSTM), conditional random field (CRF), which uses both character and word embeddings. The system consists of two constituent sequence classifiers: a frame identification (lexical unit) classifier and a frame element classifier. The classifier achieves an F1 score of 93.70 for cancer diagnosis, 96.33 for therapeutic procedure, and 87.18 for tumor description.

Clinical notes in EHR are intermittently updated and are sometimes missing information. Gong et al., 2018 [7] addressed this problem by learning to generate topics that should be in summaries of EHR. They developed a model that can be used to generate topics from structured health record data that should be in a patient's clinical note. These topics can be used as a checklist for clinicians while they are writing the note.

Plaza et al., 2012 [8] describe a summarization system for the biomedical domain that represents documents as graphs formed from concepts and relations in the UMLS Metathesaurus. The graph is constructed using the UMLS concepts identified within the sentence, extracting the complete hierarchy of hypernyms for each concept and finally merging the hierarchies. The two upper levels of these hierarchies are removed since they represent concepts with excessively broad meanings and may introduce noise to later processing. They also describe a variety of strategies that make use of MetaMap and Word Sense Disambiguation (WSD) to accurately map biomedical documents onto UMLS Metathesaurus concepts.

Rossiello et al., 2017 [9] proposed centroid-based text summarization for general domain using word embeddings. Embedding for each sentence is constructed using embeddings of words within it and a centroid embedding is constructed using vectors of words having high TF*IDF weights. All the sentences having high cosine similarity with the centroid vector are added to the summary. Also, before adding a sentence to the summary, cosine similarity with all the summary sentences are calculated to avoid similar sentences in the summary. Their future work includes using topic modeling techniques such as Latent Dirichlet Allocation (LDA) for construction of centroid embedding.

Chengzhang et al., 2018 [10] used Word2vec for summarization in general domain. Words in an article were represented as vectors trained by Word2vec, the weight of each word, the sentence vector and the weight of each sentence were calculated by combining word-sentence relationship with graph-based ranking model. Finally, the summary was generated on the basis of the final sentence vector and the final weight of the sentence. They obtained better results on real datasets compared to TF-IDF and TextRank (Table 1).

Arora et al., 2017 [11] proposed a simple but tough to beat baseline for sentence embeddings. They calculated a sentence embedding by weighted average of word vectors, and then modified them a bit using PCA/SVD. This weighting improved performance by about 10–30% in textual similarity tasks, and beats the sophisticated supervised methods including RNN's and LSTM's.

Hirao et al., 2002 [12] extracted important sentences using support vector machines in general domain. They also discussed various input features for text summarization. Some of the features are position of sentence (position in document, position in paragraph), length of sentence, weight of sentence (frequency-based), named entities (person, location, etc.), semantical depth of nouns, etc.

Beam et al., 2018 [13] present a new set of embeddings for medical concepts learned using an extremely large collection of medical data which includes insurance claims database of 60 million members, a collection of 20 million clinical notes, and 1.7 million full text biomedical journal articles. They have generated the largest set of embeddings for 108,477 medical concepts called cui2vec. They have also provided a downloadable set of pre-trained embeddings of the cui2vec embeddings.

Chen et al., 2019 [14] introduce BioSentVec: the first open set of sentence embeddings trained with over 30 million documents from both scholarly articles in PubMed and clinical notes in the MIMIC-III Clinical Database. The embeddings are also publicly available.

Much of the work of text summarization has been done in general domain. There has been a couple of work in the area of EHR summarization but it has been more in topic generation. Extractive summarization requires identifying and highlighting concepts that are important. In this work, an attempt is made to highlight the important sentences in an EHR note by evaluating and scoring sentences in the EHR notes based on the concepts discussed in the sentence.

3 Proposed Extractive Summarization Technique

The challenge in automatic summarization is the identification of key events/concepts that are there in EHR Note. The general approach used in identifying key concepts or topics in a document is TF/IDF approach. In our approach, we are augmenting extra steps like concept identification to deal with the limitations of TF/IDF approach.

Table 1 Work organized by publication date

Work	Key points and Limitations
Hirao et al., 2002 [12]	• Summarization using support vector machines • Done for documents in general domain
Bodenreider et al., 2004 [15] (UMLS)	• The Unified Medical Language System is a repository of biomedical vocabularies which integrates over 2 million names for some 900,000 concepts, as well as 12 million relations among these concepts
Aronson et al., 2010 [16] (Metamap)	• MetaMap is a highly configurable program to map biomedical text to the UMLS Metathesaurus or to discover Metathesaurus concepts referred to in text
Plaza et al., 2012 [8]	• A summarization system that represents documents as graphs formed from concepts and relations in the UMLS Metathesaurus
Rossiello et al., 2017 [9]	• A centroid-based text summarization for documents using Word Embeddings • Done for documents in general domain
Dabek et al., 2017 [3]	• A timeline-based framework to summarize the clinical data of a patient enclosed within an EHR • Used only structured data
Arora et al., 2017 [11]	• An improved technique for sentence embedding for textual similarity tasks
Beam et al., 2018 [13] (cui2vec)	• A new set of embeddings for 108,477 medical concepts called cui2vec
Alsentzer et al., 2018 [4]	• Provides an upper bound on extractive summarization of discharge notes of MIMIC-III dataset
Gong et al., 2018 [7]	• A model that can be used to generate topics that should be in patient clinical note • Used only structured data
Yuqi Si et al., 2018 [6]	• A frame-based NLP method that extracts information from clinical narratives • Done only for cancer clinical notes
Chengzhang et al., 2018 [10]	• Summarization with graph-based ranking model using Word2vec • Done for documents in general domain
Liang et al., 2019 [5]	• Summarization for only for two specific diseases, hypertension or diabetes mellitus
Chen et al., 2019 [14] (BioSentVec)	• The first open set of sentence embeddings trained with over 30 million documents from both scholarly articles in PubMed and clinical notes in the MIMIC-III Clinical Database

The first step will be to pre-process the EHR Note to remove all irrelevant words. Stop words specific to medical domain like hospital, treatment, etc., are also removed in this process. Medical domain-specific dictionary is useful for identifying such stop words. Once the data is pre-processed by removal of stop words, next process is to identify the medical concepts discussed in the notes. Redundant concepts need to be identified and removed and each of the concept identified needs to be scored relatively. The sentences representing key concepts can be selected based on score obtained.

Figure 1 shows the design for a simple extractive summarization pipeline. A proof of concept was designed to validate the above approach. The details of dataset and implementation is explained below.

3.1 Dataset

MIMIC-III is a freely available, deidentified database containing electronic health records of patients admitted to an Intensive Care Unit (ICU) at Beth Israel Deaconess Medical Center between 2001 and 2012, released on September 2, 2016. The current version of the database is v1.4. The database contains all of the notes associated with each patient's time spent in the ICU as well as 55,177 discharge reports and 4475 discharge addendums for 41,127 distinct patients [17].

3.2 Implementation

We developed a basic tool that can extract important medical concepts and score sentences based on TF*IDF weights and MetaMap for summarization. The following tasks were performed in our extractive summarizer.

- **Section segmentation**: Discharge summary notes in MIMIC-III dataset contains sections like 'History of Present Illness', 'Past Medical History', 'Family History', 'Brief Hospital Course', etc. Using regular expressions, the document was separated into these sections.
- **Sentence segmentation**: For each section, sentence segmentation was performed using spaCy. spaCy is an open-source software library for advanced natural language processing, written in the programming languages Python and Cython. We also added some additional document-specific sentence segmentation rules to spaCy. After sentence segmentation by spaCy, each sentence was assigned a unique id.
- **Stop words removal**: Stop words are commonly used words that do not add any importance to the sentence and hence are removed. We generated a list of stop words from spaCy and MetaMap. spaCy contains more than 300 stop words in

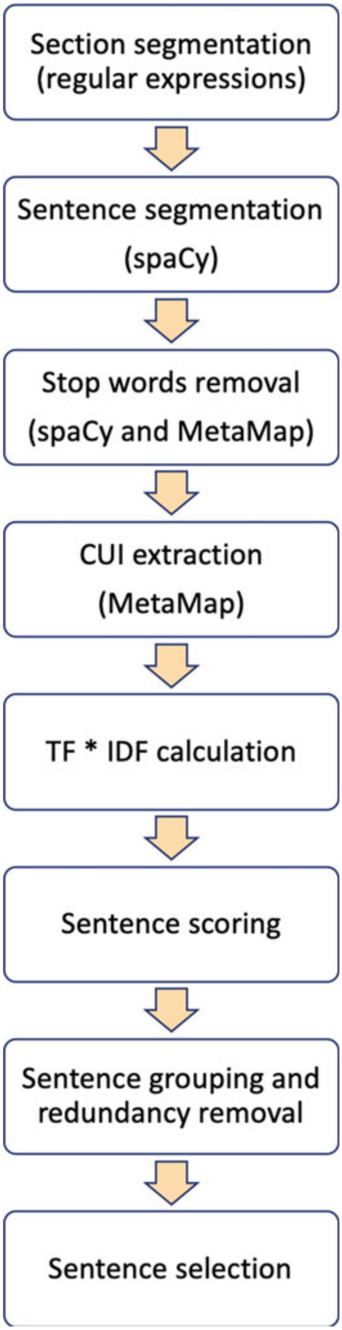

Fig. 1 Extractive summarization pipeline

general domain (such as ”the”, ”in”, ”a”, ”for”). MetaMap contains more than 10,000 stop words in medical domain (such as “age”, ”hospital”, ”date”).

- **Medical concepts extraction**: MetaMap is a highly configurable program developed by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM) to map biomedical text to the UMLS Metathesaurus or, equivalently, to discover Metathesaurus concepts referred to in text. We used MetaMap [16] for extraction of medical concepts from the note. Semantic types like ”*Therapeutic or Preventive Procedure*”, ”*Diagnostic Procedure*”, ”*Disease or Syndrome*”, etc. were included. And semantic types like ”Temporal Concept”, ”Quantitative Concept”, ”Idea or Concept”, etc. were excluded.
- **Sentence scoring**: The score of a sentence is sum of tf-idf score of each non-stop word in the sentence divided by total number of non-stop words in the sentence. The sentences were then sorted by descending order of their weights.
- **Sentence grouping and redundancy removal**: Sentences were grouped based on Concept Unique Identifier (CUI) that uniquely identifies a concept. Sentences whose all CUIs were covered by other sentences were removed. Sentences with no CUIs were also removed.
- **Sentence selection**: Finally, the top 10% sentences were selected and highlighted in the given input note.

Figure 2 displays output of simple extractive summarizer for a “History of Present Illness” section of the note. The red-colored words are the extracted medical concepts from MetaMap. The underlined sentences form the summary.

4 Results and Future Work

TF-IDF works based on the number of times a word appears in the document. So, it cannot identify similar sentences or identify negation in sentences or detect if a sentence has any medical terms or not. Hence, we used MetaMap to handle these problems. It was seen that summarization using only TF-IDF gives good results but it also had many limitations. In this section, we discuss these problems and how we overcame them.

Since TF-IDF uses only word count for scoring sentence, we might find sentences that have high similarity, repeated in the final summary. For example, sentences like “Mitral regurgitation is seen” and “Moderate mitral regurgitation persists” were repeated in the summary when we used only TF-IDF for summarization. Such limitations were eliminated by using concept of Concept Unique Identifiers (CUIs). CUIs from each sentence were extracted using Metamap. Sentences with same CUI implies sentences discussing the same concept. We grouped sentences that contained same CUIs and discarded repeated sentences. Also, the sentences that contained more CUIs and covered all the CUIs of other sentences were given more importance.

History of Present Illness :
Mr. [* * Known lastname 10722 * *] is an 84-year - old gentleman with a history of CAD s / p LAD stenting in [* * 2105 * *] , who has known diastolic heart failure with a normal LVEF (70 %) . He also has a history of tachy - brady syndrome and underwent a permanent pacemaker placement . In [* * Month (only) 1096 * *] , was admitted to [* * Hospital3 * *] with dyspnea and treated for a presumed CHF exacerbation . Trop at OSH was 0.07 (indeterminate) and had increased O2 requirement . He was transferred to [* * Hospital1 18 * *] for cardiac catheterization which revealed nonobstructive CAD . He was evaluated for aortic valve treatment options . After informed consent including full explanation of risks vs. benefits , he was enrolled into the high risk arm of the [* * Hospital1 10723**]/TAVI study . He met all inclusion criteria , and did not meet any exclusion criteria , was screened and accepted to be randomized to either [* * Hospital1 10723**]/TAVI or surgical AVR as per study protocol .
Patient was admitted from [* * 2112 - 2 - 7 * *] - [* * 2112 - 2 - 12 * *] for optimization of cardiac status prior to randomization in [* * Month / Day / Year 10723 * *] study .
Patient was randomized into [* * Month / Day / Year * *] arm , which he will receive during this hospitalization . During te last hospitalization , he was started on Pradaxa in preparation for procedure . Patient given clear instructions to stop taking Pradaxa 3 days prior to procedure (on [* * 2112 - 2 - 20 * *]) . He was also diuresed with lasix and achieved euvolemia , maintained on 2L oxygen , which he is on at home .

Since his last discharge , patient has been feeling well . He had one episode of severe epistaxis requiring packing by an ENT at [* * Hospital 882 * *] Hospital . Per the patient , Dr. [* * Last Name (STitle) * *] was aware of this and oversaw his care , and his pradaxa dose was decreased .
Patient stopped taking pradaxa the day prior to admission in anticipation of [* * Last Name (STitle) * *] procedure . Patient has not had any more episodes of epistaxis since then .

Additionally , while receiving 2 units of pRBC at [* * Location (un) 620 * *] on [* * 2 - 19 * *] , he developed abdominal pain which he thought was reflux , but was not relieved by tums . He was admitted for cardiac enzyme monitoring , out of concern abd pain could be cardiac in etiology . Troponins were trended and were 0.074 , then 0.102 , then 0.121 . CK - MB flat . Subsequently , he developed diarrhea , [* * First Name8 (NamePattern2) * *] [* * Location (un) 620 * *] report 10 episodes in one day . Patient reports diarrhea has improved and he now has 1 - 2 episodes of loose stool . He endorses a decreased appetite , but denies nausea / vomiting . C. diff was negative .

Fig. 2 Output of simple extractive summarizer for a "History of Present Illness" section of the note. The red-colored words are the extracted medical concepts from MetaMap. The underlined sentences form the summary

Using only TF-IDF resulted in high score for some sentences that had no medical terms. A sentence like "Please make an appointment to see your primary care physician" does not contain any medical term and is less important. Such sentences get high TF-IDF scores as they contain words that have high frequency in clinical notes. This is not relevant for our summary and hence scores for sentences that have no CUIs were reduced to remove such sentences from summary.

Clinical notes contains sentences like "No aortic regurgitation is seen" and "Left ventricular wall thickness, cavity size, and systolic function are normal". These sentences are less important as they indicate absence of disease. Negation of CUIs can also be detected using MetaMap and spaCy. Scores of such sentences that indicated absences of CUIs were reduced during sentences scoring.

The top 10% of the scored sentences was considered to be part of the summary. The qualitative evaluation by a domain expert can be used to test the correctness of the result.

This work can be further extended by the use of graph-based technique to identify depth of CUIs so that sentences with more specific CUIs can be selected and sentences with general CUIs like procedure, pain, etc., can be discarded. Use of word embeddings or sentences embeddings to know similarity in sentences is also another direction to explore.

5 Conclusion

The collection of data in EHR is growing day by day. It becomes difficult and time consuming for the clinicians to go through all the EHR notes of a patient. So, there is need of automatic summarization of EHR notes. In this paper, we propose a technique for extractive summarization of clinical notes using TF-IDF and Metamap. We also discuss various limitations of using TF-IDF for extractive summarization of EHR notes and discuss future research directions.

References

1. George M, Chacko A, Kurien SK (2019) Proactive diabetes management: research directions. In: Proceedings of the 20th international conference on distributed computing and networking, ICDCN '19, pp 486–491. ACM, New York, NY, USA
2. Pivovarov R, Elhadad N (2015) Automated methods for the summarization of electronic health records. J Am Med Inf Assoc 22(5):938–947
3. Dabek F, Jimenez E, Caban J. J (2017) A timeline-based framework for aggregating and summarizing electronic health records. In: 2017 IEEE workshop on visual analytics in healthcare (VAHC), pp 55–61. IEEE
4. Alsentzer E, Kim A (2018) Extractive summarization of ehr discharge notes. In: arXiv:1810.12085v1 [cs.IR]
5. Liang J, Tsou CH, Poddar A (2019) A novel system for extractive clinical note summarization using ehr data. In: Proceedings of the 2nd clinical natural language processing workshop, pp 46–54
6. Si Y, Roberts K (2018) A frame-based nlp system for cancer-related information extraction. In: AMIA annual symposium proceedings, pp 1524–1533
7. Gong JJ, Guttag JV(2018) Learning to summarize electronic health records using cross-modality correspondences. In: Machine learning for healthcare conference, pp 551–570
8. Plaza L, Stevenson M, D´ıaz A (2012) Resolving ambiguity in biomedical text to improve summarization. Inf Process Manag 48(4):755–766
9. Rossiello G, Basile P, Semeraro G (2017) Centroid-based text summarization through compositionality of word embeddings. In: Proceedings of the multiling 2017 workshop on summarization and summary evaluation across source types and genres, pp 12–21
10. Chengzhang X, Dan L (2018) Chinese text summarization algorithm based on word2vec. J Phys 976(1):012006
11. Arora S, Liang Y, Ma T (2016) A simple but tough-to-beat baseline for sentence embeddings. In: ICLR 2017 conference

12. Hirao T, Isozaki H, Maeda E, Matsumoto Y (2002) Extracting important sentences with support vector machines. In: Proceedings of the 19th international conference on computational linguistics, vol 1, pp 1–7
13. Beam AL, Kompa B, Fried I, Palmer NP, Shi X, Cai T, Kohane IS (2018) Clinical concept embeddings learned from massive sources of multimodal medical data. arXiv preprint arXiv:1804.01486
14. Chen Q, Peng Y, Lu Z (2019) Biosentvec: creating sentence embeddings for biomedical texts. In: IEEE international conference on health informatics (ICHI). IEEE
15. Bodenreider O (2004) The unified medical language system (umls): integrating biomedical terminology. Nucleic Acids Res 32(1):D267–D270
16. Aronson AR, Lang FM (2010) An overview of metamap: historical perspective and recent advances. J Am Med Inf Assoc 17(3):229–236
17. Johnson AE, Pollard TJ, Shen L, Li-wei HL, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, Mark RG (2016) Mimic-iii, a freely accessible critical care database. Sci Data 3:160035