CS 421 - NATURAL LANGUAGE PROCESSING

RESEARCH HIGHLIGHT STUDY

AUTOMATIC SUMMARIZATION OF MEDICAL CONVERSATIONS USING NLP

BY APURVA RAGHUNATH & NANDANA SHIMOGA PRASAD

1. PROBLEM AND MOTIVATION

Verbal medical dialogues between doctors and patients are common. Such administrative tasks usually take excessive time and are done at the expense of patient care. Also, documenting these medical conversations is a burden for doctors. This is where automatic gathering of such conversions becomes crucial. For example, a patient could review the diagnosis later if a medical history is generated containing all relevant information from these dialogues. However, capturing the entire conversation is not productive. Correspondingly, this poses a challenge for Natural Language Processing in acquiring automatic summaries from medical conversations. With the unfolding innovations of artificial intelligence and natural language processing technologies, modular methods that develop automatic summaries which can help improve the recall and understanding of a patient's treatment course can be achieved. We explore five different research papers that exploit NLP techniques in automatically summarizing such medical conversations.

2. CONTRIBUTIONS AND METHODLOGIES

2.1 Research 1 – Automatic summarization of medical conversations, a review [1]

This research aims to develop an automatic summarization method for medical conversations between patients and doctors using convolutional and recurrent neural networks. The objective is to detect the applicable segments of dialogue and to condense them for inclusion in a summary through deep learning. Important clinical issues could thus be recognized by discarding other unimportant fragments like greetings, acknowledgments, social concerns, etc. Two important criterions were used. One based on frequency of words using **TF-IDF**. TF (Term-Frequency) is the frequency of a word in the document and IDF (Inverse Document Frequency) reduces the effect of words that occur too often in a collection of documents. Second, the **surface feature** where the features such as title, sentence position and length, proper noun and term weight are used to identify relevant sentences.

The research reviews the usage of the below models in automatically summarizing medical conversations:

a. **Probabilistic Models** - Probabilistic Models for speech summarization such as Context Free Grammars and Markov Models were used to define syntactic structures, inspect, and understand the content.
b. **Optimization methods -** Integer Linear Program is another useful method to get summaries from the text. It can extract inference in automatic summarization and sentence compression under a maximum coverage model and can be used for multi document update summarization.
c. **Graph-based methods -** The prime sentences and keywords to be used in extractive summarization of text documents are identified and extracted using graph-based ranking algorithms.
d. **Machine Learning Approaches –** Models like Naive Bayes can be used to choose if a sentence belongs to a summary or not. Clustering algorithms are helpful in getting extractive summaries. Finally, Support Vector Machines can rank all sentences in the topic cluster thus accommodating for summarization.
e. **Neural Networks -** Neural Networks are most widely used for automatic summarization of text. In NLP, **L**ong **S**hort-**T**erm **M**emory and **G**ated **R**ecurrent **U**nit are the relevant NN's. LSTM uses forget, input, candidate, and output gates to add or remove information. GRU combines the forget and input gates into a single update gate.

The research further aims to develop a system that provides an abstractive summary from medical dialogues and a hybrid pointer-generator network that is useful on medical domain.

2.2 Research 2 – Extracting relevant information from physician – patient dialogues for automated clinical note taking [2]

This research presents a system that extracts information from dialogues and generates a patient note automatically that could be assessed and modified by the clinician. The goal is to liberate the medico's valuable time and switch their focus on interacting with the patients rather than operating the EMR. The study also focuses on linguistic context and time information to determine the medically relevant parts in a conversation. This increases the accuracy of the generated report by providing consistent and neat reports that support clinical decisions.

The research proposed a pipeline with the following components to automatically summarize medical conversations:

a. **Preprocessing and data splitting** - Converts the transcripts text into lowercase and removes punctuations using NLTK modules. To represent each word as a word embedding, ELMo and word2vec embeddings were adapted.
b. **Utterance type classification** - To understand the conversational context, the utterance type is classified as question, statement, positive & negative answer, vague/incomplete utterances (huh, um, yeah). Two-layer bidirectional gated

recurrent unit neural network was used to implement this component. AMI and switchboard corpora were used to map the utterance labels and add the data to the training set.

c. **Entity Extraction**

   Time phrase extraction - The time, duration, frequencies, and quantities of events in patient history was determined using HeidelTime.

   Clinical entity extraction - Medical lexicons such as BioPortal symptom lexicon, SNOMED-CT, CHV and RxNorm were searched to identify and extract symptoms, diagnosis, medications, investigations, and therapies.

   Attribute Classification - To classify attributes into modality (events experienced) and pertinent (disease to which the entities are relevant), SVM classifier was used.

d. **Clinical note generation -** The structured data from the previous steps was classified using the SOAP entity classifier and converted into free text clinical notes.

e. **Primary diagnosis classification -** TF-IDF is applied on the cleansed text followed by logistic regression, SVM and random forest to classify and identify the main diagnosis. The accuracy was measured by computing the F1-score of the assigned labels available in the transcriptions.

f. **Topic Modelling -** LDA (with gensim package) was adapted to form k topics (clusters of words) occurring together. Topic modelling helps to track each visit, word usage distribution and relevant text extraction.

g. **Relevant utterance extraction -** From the LDA model, the dominant topic for each class is picked using a topic weight matrix.

The research presented a system to extract relevant entities from dialogues using linguistic context there by saving clinician valuable time.

2.3 Research 3 – Medical Dialogue Summarization for Automated Reporting in Healthcare [3]

The goal of this research is to generate an automatic medical report called Care2Report of patient-doctor dialogues which aims to support speech and text processing in healthcare. This report combines computational linguistics and AI techniques. The proposed pipeline consists of many sequentially connected components. If an error is introduced in a component, overall pipeline quality is affected. Defining quality metrics such as precision, recall, F-Score thus become central in assessing the performance of individual components. Along these lines, the measures of quality are devised in each component of the pipeline which is the focal point of this research.

The proposed dialogue summarization pipeline consists of a series of computational components that transform output from one system into input for another.
These components are described below:

a. **Speech transcription** - uses audio of the dialogue as input
b. **Triple extraction** - identifies and extracts semantic triples from the clinical notes. Subjects, predicates, and objects make up these semantic triples.
c. **Triple matching** - If the extracted triples match with triples in the ontology (domain-specific information), then such triples are selected. Triples thus matched and selected are included in the report and stored in a graph. Outline of the patient's symptoms, the diagnosis and the treatment are obtained from this graph.
d. **Report generation** - transforms the categorized triples back into natural language to make them comprehensible. The report generated complies with the SOAP format and defines the following sections for reporting on a consultation: Subjective (S), Objective (O), Assessment (A) and Plan (P).

The research's POC indicate that the proposed pipeline structure achieves dialogue summarization with an optimized implementation.

2.4 Research 4 – Towards an Automated SOAP Note: Classifying utterances from Medical Conversations [4]

The research proposes an approach to transcribe the conversations by capturing an audio recording of the medical conversations and using automatic speech recognition & NLU to summarize relevant information. A systematic analysis that adapts deep learning architecture to classify utterances forms the focus of this research.

The research adapts the below mentioned architecture to automate the summarization of medical conversations:

- Bag of words (BoW) and deep learning models were employed to determine natural language understanding for the medical conversations.
- Majority-class (MC), Multinomial Naive Bayes (MNB), logistic regression (LR), and random forest (RF) classifiers with a BoW encoding of each utterance were used for the BoW baselines.
- The most basic model used was a deep learning baseline which learns the weighted average of the three ELMo layers and averages all the word embeddings in-order to yield a sentence embedding. ELMo is a word embedding model that embodies context of a sentence into word meaning representations.
- One specific layer was then added by each of the models. A bi-LSTM is added for generating the utterance embeddings and then an LSTM decoder to give the outputs.
- Finally, all these models formed a dense layer with a softmax activation function that is used for classification. The softmax layer then yields the final sequence which is decoded into speaker labels and SOAP note sections, respectively.

The research indicates capturing both word and utterance level context leads to substantial improvements in classification tasks.

2.5 Research 5 – Summarization from medical documents: A Survey [5]

The research is a survey of probable summarization technologies that have been used in the medical domain. These techniques are built on the existing and modern document types and summarization applications in the medical field. The foremost contribution of this research is to inspect the issues that arise in the usage of these summarization techniques and not just a study of various techniques. Consideration of characteristics of the medical domain is another important factor that makes this research interesting.

This research reviews automatic summarization techniques classified into three broad categories mentioned below:

a. **Extractive summarization** – Aims to select sentences, paragraphs to be included in the summary verbatim.
   - Extractive techniques can be employed for both single and multi-documents.
   - An approach to exploit extractive technique uses a cluster signature of the document to rank the extracted sentences. The system takes as its input, the medical documents, and forms groups of clusters. These clustered are analyzed for key features to form a cluster signature that best characterizes each group of documents.
   - Next, the automatic summarization generation step uses the cluster signature to match each sentence of the document that needs to be summarized. A vector space model is used for representing the cluster signature and the sentence.
   - Finally, the ranked sentences are selected to be outputted in the report summary.

b. **Abstractive summarization** – Aims to select the most prime concepts prevalent in the document.
   - Abstractive techniques can be employed on both single and multi-documents.
   - In this technique, once the sentences are selected, they are represented in the form of a predicate-argument structure. During this, tokenization, morphological analysis, shallow syntactic parsing, chunking, dependency analysis and mapping to the internal representation are also done.
   - Natural language generation system called the lexigen was used to create the summaries of these sentences. Such summaries include meta statements about the document.

c. **Cognitive model-based summarization -** Aims to develop an empirical model for summarization by imitating professional human summaries.
   - One such model is a query-based summarization system. The first step here is to identify a search scenario using domain ontology concepts and mapping this scenario to a biomedical database query. If the query results have some journals, then the interesting pieces of text in them are identified and summarized to the query scenario.

This survey provided the potential of summarization technology in medical domain by making use of medical documents and summarization applications.

3. COMPARE AND CONTRAST OF RESEARCH PAPERS

We can see that ELMo word embedding method was adapted in both research study 2 and 3. ELMo is the word embedding model used to embody context of a sentence into word meaning representations. The advantage of ELMo is that it creates vectors on-the-fly by passing the text through the deep learning model. Also, it takes entire input sentence into equation for calculating the word embeddings. Hence ELMo word embedding seems to have more perks than other word embedding techniques in generating better report summaries.

With SOAP note structure one can easily access patient's records and identify important information. It is also important to structure a report in a clear and concise manner so that it is less likely to result in miscommunication between healthcare providers. Thus, an automated report complying to SOAP format is necessary. This is the key similarity that can be observed in research study 3 and 4 where the focus is generating SOAP compliant automated summaries of medical conversations.

Bidirectional Recurrent Neural Networks connect two hidden layers of opposite directions to the same output. With this form of generative deep learning, the output layer can get information from past and future states simultaneously. Also, it solves the problem of fixed sequence prediction. An added advantage is that it can be used in machine translation where input and output have different sizes or in case of text summarization where input and output are of different length. Research study 1 and 2 have adapted bi-LSTM to take advantage of the same.

Extracting relevant utterances to be included in summary is crucial for a complete a meaningful summary. To achieve this the research studies have employed different techniques including entity extraction, semantic extraction, extractive, and abstractive methods. Each of these has its own advantages and usefulness for respective proposed system and helps in improving the performance of the proposed system.

4. KEY TAKEAWAY FINDINGS

- After reviewing five research papers that focus on automating the process of summarizing medical conversations, we realized that automatic dialogue summarization is achievable with sufficient engineering efforts using different NLP techniques thus reducing administrative burden of the health care providers and improving the patient care.
- This literature study showed that several NLP tools have been developed focusing on medical domain that generate summaries of better quality through identification and extraction of relevant medical terms. Deeping leaning and machine learning models in particular have been useful in classifying dialogue turns and summarizing them.
- We learnt that automatic dialogue summarization is a challenging task because, unlike in journalistic texts, the most important sentence is not the first one in each paragraph. Also,

in such oral conversations, aspects such as the speakers' turns, fillers, repetitions, repairs, or unfinished clauses need to be considered. Developing a system which handles all these factors while generating a meaningful and complete summary without modifying the original context of a patient-doctor conversation is necessary but quite challenging.

- We also find that the summaries generated might be a shallow representation and thus not producing informative or critical summaries. The use of sophisticated natural language processing techniques therefore becomes essential for generating summaries and to scale to large size documents.
- Further we observe that research should be continued in developing a portable technology that generates summaries which are both user and domain oriented. This becomes essential as the summaries must cover information needs of various types of users and sub-domains of medicine.

5. BIBILOGRAPHY

[1]. https://hal.archives-ouvertes.fr/hal-02611210/document
[2]. https://www.aclweb.org/anthology/D19-6209.pdf
[3]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7225507/pdf/978-3-030-49165-9_Chapter_7.pdf
[4]. https://arxiv.org/pdf/2007.08749.pdf
[5]. https://www.researchgate.net/publication/220103096_Summarization_from_Medical_Documents_A_Survey