DigitalOwl

THE EVOLUTION OF AUTOMATED MEDICAL SUMMARIZATION

Different technological approaches and how they can be applied

By Amit Man, Co-founder and CTO

Medical summarization is the process of reducing large amounts of medical information, such as patient records or research papers, into shorter, easier-to-digest portions. Throughout this post, I treat the term summarization quite loosely, including methods that facilitate the understanding of large corpora of text, even those not formally considered summaries. Summarization methods have been evolving and will likely continue to evolve in upcoming years. Below are a few key phases within this evolution.

KEYWORD DETECTION

One early approach to medical summarization involved simply extracting key terms or phrases from a document and using those as a summary. This method was simple, but it often resulted in summaries that were incomplete or difficult to understand. This is because it did not take into account the context or structure of the original document.

In addition, the same keyword might be used to indicate a condition, negated condition or family condition with no way of distinguishing between them without this additional context.

ENTITY RECOGNITION

Entity recognition is a technique that identifies and extracts specific pieces of information - entities such as disease or treatment names - from a document. While this can be useful for extracting specific pieces of information that are relevant to a particular task, alone - it does not always provide a comprehensive understanding of the content of the document as a whole.

To address this, the concept of relationship extraction was introduced. This technique allows the additional identification of relationships between entities, such as the relationship between a disease and the date it was diagnosed. By considering these relationships, it is possible to gain a more comprehensive understanding of the content of a document.

NORMALIZATION TO MEDICAL CONCEPTS AND DATA ENRICHMENT

As each medical concept can be represented using multiple wordings, it is common practice to normalize the extracted entities into “medical concepts”. Normally, this is done by matching the entity text against a curated medical database. The normalized concepts are then coded, so that they can be managed and analyzed by machines. By adding data enrichment to the process, more information can be added to the summary, such as severity and clinical field for the different extracted conditions. In this way, we reduced the amount of redundancy in the summary and ensured a better understanding of the content and its implications.

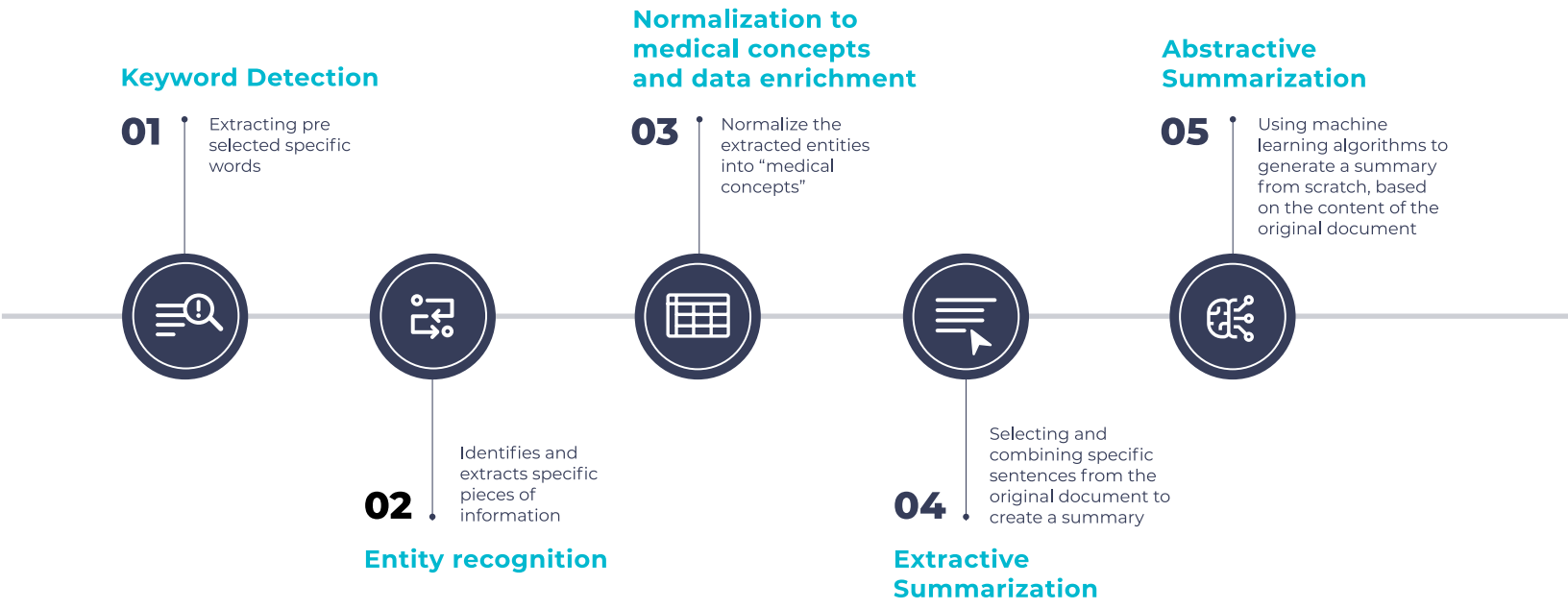
EXTRACTIVE SUMMARIZATION

A different approach to medical summarization is extractive summarization, which involves selecting and combining specific sentences or passages from the original document to create a summary - similar to a human copy-paste action. This can be done automatically using algorithms that identify the most significant or relevant sentences based on factors such as frequency of occurrence or position in the document. Extractive summarization has some advantages over keyword extraction or entity recognition because it takes into account the context and structure of the original document. However, there are many limitations to using the original wording of the text.

ABSTRACTIVE SUMMARIZATION

The most advanced and recent approach to medical summarization is abstractive summarization based on generative models. This involves using machine learning algorithms to generate a summary from scratch, based on the content of the original document. Generative models are typically very large models that are trained on vast amounts of data. While more complex and resource-intensive to develop, these models are able to capture the key ideas and themes of the source and produce texts that are coherent and flow naturally, making it difficult for humans to distinguish the automatically generated text from one written by a person.

- EVOLUTION OF SUMMARIZATION METHODS -



Limitations of medical summarization based on entity recognition

SUPPORTS ONLY A PREDEFINED SET OF ENTITY TYPES

Entity recognition systems rely on a predefined set of entity types, such as the names of drugs, diseases, or treatments. This can limit the scope of the summary, as entities that are not included in the pre-defined set may not be identified.

Adding new labels to entity types is a laborious task. As an example, adding the "mm increase in hypertrophy" condition to the summary requires creating a new entity type, labeling it based on real data, and re-training the model.

STRUGGLES WITH AMBIGUOUS OR UNCLEAR LANGUAGE

Entity recognition systems may struggle with ambiguous or unclear language, which can lead to incorrect or incomplete identification of entities. 'Heavy bleeding' can refer both to menstruation or a serious abdominal injury. 'ARF' can mean both acute renal failure, acute respiratory failure or acute rheumatic fever. Examining the context of the acronym is required in order to determine the correct option. Misspelling and OCR errors aggravate this situation. One wrong letter due to a doctor's mistype or an OCR error could make the difference between ITP (Immune Thrombocytopenia) and TTP (Thrombotic Thrombocytopenic Purpura).

LIMITED GRANULARITY OF MEDICAL CONCEPTS

As mentioned, it is typical for entity recognition models to use a medical database in order to assign coding and meaning to the various extractions. These DBs are ever growing. ICD-10 has 14,000 codes in it. Its extension, ICD-10-CM, has about 70,000 codes. SNOMED reached even higher numbers at 352,567 different codes! But even with the most detailed medical databases, entity recognition systems can struggle to provide a fine enough level of granularity. Take for example head trauma. As detailed as the medical database may be, it is impossible to list all the different types of head injury. While both a hit to the head by a tennis ball and a hit to the head by a brick are considered head trauma, they are different medical situations and their impact on the individual's health is very different. Simply listing "Head trauma" in the summary will not convey that information.

COMPLEX RELATIONS BETWEEN ENTITIES ARE NOT EXTRACTED

While simple relationships between entities, such as a condition and its diagnosis date, are certainly possible, complex relationships are becoming increasingly untangled. For example, it is very difficult for an entity recognition engine to capture the following relations:

Pneumonia was suspected because of a chest x-ray and therefore amoxicillin was given. Because the patient experienced gastrointestinal side effects from the amoxicillin, azithromycin was given instead, resulting in improvement of the pneumonia.

THE CONTEXT OF ENTITIES IS NOT REPRESENTED

Entity recognition systems do not capture the context from which entities were retrieved. The shoulder injury might be mentioned, but whether it occurred during a gym workout, a slip on the ice, or a car accident will not be mentioned. This additional context may also have medical implications. The drug taken by the patient might be mentioned, but whether the patient followed the instructions to the letter or "missed a pill here and there" will not be revealed.

THE LIMITATIONS OF A BINARY APPROACH WHEN REALITY ISN'T BLACK OR WHITE:

Entity extraction is the process of extracting entities and classifying them. This is a condition, this is a procedure, this is a medication, etc. One comes to realize very quickly that those entities come in many flavors and new classes of entities are frequently added: past condition, family condition, negated condition, improved condition, suspected condition etc. But even after constantly increasing the number of entity types, some items do not fit nicely into a bucket.

- 'Smoker' is a clear (social history) condition.
- 'Not a smoker' is a clear negated (social history) condition.

But what about 'nonsmoker'? Is it a condition or a negated condition?
Well, kind of both, depending on the context. This issue is not purely academic.
Your summary might treat negated conditions very differently than regular conditions, to the point of excluding them from the summary.

Most will treat the pair of words 'leg pain' as a body part and a condition, where the 'pain' condition is associated with the 'leg' body part. Same for 'arm fractures'.

But what about the phrase 'heart attack'? Even though 'heart' is a body part, in this context, most people treat heart attacks as one unified condition. And then it becomes even trickier, what do you do with items such as 'lung infection' or 'eye strain'. Which is it? A condition associated with a body part or one condition? Well, kind of both. And if your summary relies on those body part extractions in any way, it can influence the summary content.

CHALLENGING TEXT FOR ENTITY EXTRACTION BASED SUMMARY

EXAMPLE 1 - **Supports only a predefined set of entity types**

> The test revealed that the patient has stage III breast cancer, with the tumor measuring **2.5cm** in diameter and showing **moderate differentiation** of the cancer cells. Additionally, the test revealed that the cancer had **not yet spread to the lymph** nodes. The test also revealed that the cancer cells were **ER/PR positive, HER2 negative**, which suggests that the cancer **may respond to hormone therapy.**

EXAMPLE 2 - **Struggles with ambiguous or unclear language**

Patient is currently experiencing **ARF** as a result of dehydration and sepsis. Despite aggressive fluid and antibiotic therapy, the patient's creatinine levels have not improved and they will need to be started on dialysis. We will closely monitor the patient's kidney function and continue to work towards stabilizing their condition.

Patient is displaying symptoms consistent with **ARF** including joint pain, fever, and a characteristic rash. They have been started on appropriate antibiotics and anti-inflammatory medication. Close monitoring of cardiac function will be required as **ARF** can lead to long-term damage to the heart. We will work closely with the patient to ensure prompt treatment and manage any potential complications.

EXAMPLE 3 - **Limited granularity of medical concepts**

A patient presents with a **head injury** resulting from being struck by a falling brick from the third floor.

A patient presents with a **head injury** caused by impact from a tennis ball.

EXAMPLE 4 - **Complex relations between entities are not extracted**

Pneumonia was **suspected because of** a chest x-ray and **therefore** amoxicillin was given. **Because** the patient **experienced** gastrointestinal **side effects** from the amoxicillin, azithromycin was **given instead**, **resulting in improvement** of the pneumonia.

EXAMPLE 5 - **The Context of entities is not represented**

Patient is currently experiencing significant symptoms of **depression** following the recent loss of her husband. She reports feelings of overwhelming sadness and difficulty with activities of daily living.

Patient is displaying symptoms of **depression** related to his chronic conditions, including chronic pain, diabetes, and heart disease. He reports feelings of hopelessness and lack of motivation.

EXAMPLE 6 - **The Limitations of a Binary Approach to Gray Reality**

Patient is a 70-year-old male with a history of heart disease who had a **pacemaker** implanted four years ago. The device has been functioning well and providing appropriate cardiac pacing.

(Example 1) Much of the important and relevant information can not be associated with a major entity classification such as conditions/procedures etc. (Example 2) The context of the abbreviation ARF (Acute Renal Failure or Acute Rheumatic Fever) is crucial to understanding its meaning. (Example 3) While in both paragraphs the extraction condition is head injury, those are in fact very different kinds of head injury. (Example 4) There are at least 6 entity relations in the paragraph. All of different nature. (Example 5) While in both there are symptoms of depression, the context around it paints a very different story. (Example 6) Is the pacemaker a past procedure or part of the patient's list of conditions? Or both?

The power of generative summarization

Generative model based summarization is a type of automated summarization that uses machine learning models to generate natural language text that summarizes a given document or set of documents. These models are trained on large datasets of human-written text and are able to generate human-like summaries that capture the main points and key details of the original text.

Generative model based summarization overcomes many of the limitations of entity recognition based summarization.

First, generative models do not rely on predefined entity types and can identify and summarize relevant information regardless of its specific type. They are also able to understand and use context and nuance in language, which can help them disambiguate unclear or ambiguous language and capture complex relations between entities.

Second, generative models are able to generate summaries in a natural language format, which can be more easily understood and digested by readers. This can be especially useful for users who are not familiar with the specific entity types or technical language used in a given document.

Overall, generative based summarization offers a more comprehensive approach to automated summarization, as it is able to consider the context, relationships, and language usage in the original text and generate a coherent summary in a natural language format.

EXAMPLE:

Abstractive summary based on a generative model:

The patient is a 68-year-old male with a weight of 187 lbs and a history of depression and sleep apnea. The patient underwent colonoscopies and biopsies over the years, which revealed a sessile serrated polyp, ulcerative colitis, and diverticulosis in the colon, for which Humira was prescribed. Additionally, he suffers from hypertension, hyperlipidemia, and stable coronary heart disease. He underwent cardiac catheterization and stent placement in 2016 and was treated with a variety of medications. Recently, he experienced musculoskeletal pain and neuropathy in the elbow and foot regions. This is managed by Gabapentin, Acetaminophen, and CBD oil.

A summary based on entity recognition:

- Sleep Apnea Syndrome 2020
- Depression 2020
- Continuous Positive Airway Pressure 2020
- Obesity 2020
- Surgical Pathology (referral) 2019
- Adenoma Of Large Intestine 2019
- Diverticulosis 2019
- Hyperlipidemia pre 2019
- Coronary Heart Disease 2017
- Hypertension pre 2017
- Gastrointestinal Hemorrhage 2017
- Alcohol Current User 2017
- Angina Pectoris 2017
- Angioplasty 2017
- Hospitalization 2017
- Cardiovascular Disease 2016
- Cardiac Surgery 2016
- Coronary Stent Placement 2016
- Cardiac Catheterization 2016
- Ulcerative Colitis 2010
- Former Smoker 1970

Complementary Techniques

Are generative-based summaries going to replace entity extractions? Probably not. Entity recognition brings its own set of advantages into the mix.

- The link between the extraction and the evidence is very clear and straightforward
- It translates very easily to structure data format used by different systems.
- As the entities are pre-defined and relatively structured, it is easier to apply data enrichment to them. For example assigning severities to a generative model is less predictive as the content it would produce is not bounded.
- The scope of the generative base paragraphs are more limited and hard to define. Because generative models reduce a large amount of text into a paragraph or two, by that definition, not all the information can be included. Hence the scope is limited. While whatever gets in and whatever is left out can be fine-tuned, it is hard to control and even harder to rationalize retrospectively when things go differently than expected.
- Finally, the normalized nature of the extracted medical concepts allows them to be readily used by different analytics, statistical and prediction tools as well as rule engines

At DigitalOwl we see the two techniques as complementary to each other. Together, they produce a very effective method to review medical records. The generative paints a picture full of nuance, context and complex relations. The entity based summary establishes a full coverage of all relevant medical information in the case, allows for clear evidence by referencing the source text and enables the generation of normalized structure data for automations and analytics.

DigitalOwl uses proprietary technology for its NLP and Generative Text capabilities and does not rely on third-party services