International Journal of Engineering, Science and Mathematics
Vol.13 Issue 05, May2024,
ISSN: 2320-0294 Impact Factor: 6.765
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Survey on using Natural Language Processing (NLP) on Electronic Health Records

Radhika Kanubaddhi*
Saidaiah Yechuri**
Venkata Ramana Kandula***

Abstract (12pt)

The main purpose of this survey is to understand the critical steps and challenges in applying Natural Language Processing in the healthcare industry. Natural Language Processing could eliminate many inefficiencies in the delivery of healthcare. However certain key challenges need to be understood and problem solved before full value of Natural Language Processing is realized.

Each step of applying Natural Language Processing to health record data has its own challenges to resolve. None of them are simple to address. Especially the pragmatic use of language in healthcare setting is markedly different from other industries.

Copyright © 2024 International Journals of Multidisciplinary Research Academy.*All rights reserved.*

Keywords:

NLP;
EHR;
Health Data;
Semantics;
Indexing.

Author correspondence:

Radhika Kanubaddhi,
AI and Database Specialist, Amazon Web Services, Dallas, TX
r.kanubaddhi@gmail.com

1. Introduction

In the ever-evolving landscape of healthcare, the deluge of data generated daily presents both a challenge and an opportunity. Natural Language Processing (NLP) stands at the forefront of this revolution, offering powerful tools to unlock the potential within unstructured clinical narratives. This article delves into the transformative impact of NLP on healthcare data, exploring how it enhances patient outcomes, streamlines operations, and paves the way for personalized medicine. By harnessing NLP, healthcare professionals can sift through vast repositories of text, extract meaningful insights, and make data-driven decisions that were once beyond reach. Join us as we explore the intersection of artificial intelligence and healthcare, where NLP is not just a technological advancement but a beacon of hope for the future of medicine.

2. Steps in Natural Language Processing (NLP)

Natural Language Processing (NLP) involves several steps to understand and generate human language. Here's a brief description of the steps involved in NLP [1]:

- Lexical Analysis: This step involves breaking down the input text into its constituent parts, known as tokens, which could be words, phrases, or symbols.
- Syntactic Analysis (Parsing): The tokens are analyzed for their grammatical structure and how they are organized into sentences. This step checks the syntax or arrangement of words.
- Semantic Analysis: Here, the meaning of the words and sentences is determined. The goal is to understand the literal meaning of the tokens.

**Software Development Engineer, Amazon Web Services (AWS), Denver, CO.
**Cloud Solutions Architect, Microsoft, Dallas, TX

- Discourse Integration: This step considers the context beyond individual sentences, ensuring that the interpretation of one sentence can influence the meaning of the next.
- Pragmatic Analysis: Finally, the language is interpreted for its intended effect, taking into account real-world knowledge and the goals of the communication.
- These steps help machines process and understand human language, enabling applications such as translation, sentiment analysis, and chatbots.

3. A Closer Look

Morphology

A word is made up of a root and possibly other morphemes (prefixes or suffixes). NLP system reads the electronic form of a text and separates it into individual units called tokens and the process is called tokenization. Morphemes can also modify the meaning or change part of speech [2].

For example, the term 'phosphorylates' consists of the root 'phosphorylate' plus the third person singular present tense morpheme '-s'.

Another example, the suffix '-tion' changes the verb ('phosphorylate') into the related noun ('phosphorylation'), and the prefix 'de-' can negate the meaning, as in 'deactivation'.

Normally, word boundaries in English are indicated by white space, and a sentence boundary is indicated by '.' (period or full stop). There are many complications, particularly in the health records.

For example, use of '.' in decimals ('1.09'), use of '/' to link multiple gene names ('waf/cip-1'), or variable use of white space in gene names, such as 'BRCA 2' versus 'BRCA-2'and in order to retrieve all of the mentions of the BRCA2 gene in the literature, a gene mention retrieval system would need to capture at least the following typographical variants: Brca2, Brca-2, BRCA-2, and BRCA 2.

Lexicography

NLP system needs to perform lexical look up to identify the words or multiword terms known to the system and determine their categories and canonical forms.

For example, 'abdominal' is an adjective where the canonical form is 'abdomen', and activation is a noun that is the nominal form of the verb 'activate'.

Syntax

The syntax or grammar of a language controls how words are grouped into meaningful phrases and eventually into sentences.

Let's take an example. In the sentence 'VRK1 phosphorylates c-Jun', the noun 'VRK-1' is the subject (and actor) for the verb 'phosphorylate', whereas 'c-Jun' is the object (recipient of the action).

Semantics

Semantic relations capture meaning.
For example, 'c-Jun is activated by VRK1' can be represented as an operator (the verb 'activate) operating on two arguments - 'activate (VRK1, c-Jun)'. Thus the semantics capture the fact that VRK1 does the activation, and c-Jun is activated.

Pragmatics

Pragmatic or discourse relations capture the larger context and its contribution to meaning.

In a mammography report, mass generally denotes breast mass, in a radiological report of the chest it denotes mass in lung whereas in a religious journal it is likely to denote a ceremony.

Similarly, in a health care setting, he drinks heavily and is assumed to be referring to alcohol and not water.

4. Challenges in healthcare data

Healthcare data faces significant challenges, including data quality and integration, where diverse medical records lack standardized formats. Privacy and security concerns arise due to sensitive patient information. Standardization issues affect interoperability, while data storage and transfers pose logistical problems. Additionally, data structure issues complicate analysis, and infrastructure scalability struggles to keep up with the growing volume of data. These challenges hinder the effective use of big data in healthcare, impacting decision-making and patient care.

1. Determining types of information to capture
2. Heterogeneous formats
3. Large number of different clinical domains
4. Lack of a standardized set of domains
5. Interpreting clinical information
6. Compactness of text
7. Limited available of digital records

To resolve an ambiguous word or abbreviation like pvc in a chest X-ray denotes pulmonary vascular congestion whereas in an electrocardiogram it denotes premature ventricular complexes.

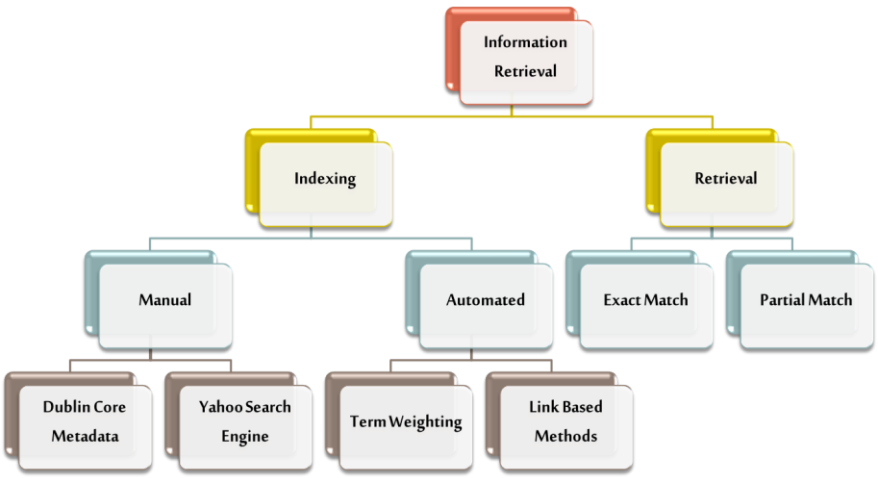
Here is a common healthcare note one could see in electronic health record of a patient.

Admit 10/23 71 yo woman h/o DM, HTN, Dilated CM/CHF, Afib s/p embolic event,
chronic diarrhea, admitted with SOB. CXR pulm edema. Rx'd Lasix.
All: none
Meds Lasix 40mg IVP bid, ASA, Coumadin 5, Prinivil 10,
glucophage 850 bid, glipizide 10 bid, immodium prn Hospitalist = Smith PMD = Jones Full Code, Cx >101

This type of text makes it very challenging to parse the patient specific information, and clinical information to accurately label for the consumers of data.

5. NLP information retrieval process

NLP information retrival has two main steps – indexing and then retrieval.



Automated Indexing

Here is the indexing of documents by the words they contain.Word indexing is typically done by taking all consecutive alphanumeric sequences between white space.Stop words are taken care by few systems by

maintaining a list called negative dictionary. E.g. and, an, by, from, of, the, with. Also stemming is done in order to avoid plurals and suffixes of the words and only consider stem forms.E.g. sleep, sleeps, sleeping

Term Weighting

Approach is TF*IDF weighting, which combines the inverse document frequency (IDF) and term frequency (TF).

Formulae[3]

$$\text{IDF(term)} = \log \frac{\text{number of documents in database}}{\text{number of documents with term}} + 1$$

$$\text{TF(term,document)} = \text{frequency of term in document}$$

$$\text{WEIGHT(term,document)} = \text{TF(term,document)}^{*}\text{IDF(term)}$$

Consider a document containing 100 words wherein the word gene appears 3 times. The TF for gene is then 0.03 (3 / 100). Now, assume we have 10 million documents and gene appears in 1000 times. Then, IDF is calculated as log(10 000 000 / 1 000) = 4. The TF-IDF score is the product of these quantities: 0.03 × 4 = 0.12.

Link based methods

Linkbased methods, started with the success of the Google search engine. This approach gives weight to pages based on how often they are cited by other pages. The PageRank (PR) algorithm is mathematically complex but can be viewed as giving more weight to a Web page based on the number of other pages that link to it.E.g. Thus, the home page of the NLM is likely to have a very high PR, whereas a more obscure page will have a lower PR.

Exact Match Retrieval

This system gives the user all documents that exactly match the criteria specified in the search statement.Also called Boolean Searching and set based searching.Used in early 1950s to 1970s operation IR systems.Tends to be associated with retrieval from bibliographic databases.E. g. To retrieve info from MEDLINE.

Firstly, in exact-match retrieval is to select terms to build sets. Other attributes, such as the author name, publication type, or gene identifier may be selected to build sets as well.Once the search term(s) and attribute(s) have been selected, they are combined with the Boolean operators.Some systems also allow wild-card characters in the search like * within or at end of the word.

Partial Match Retrieval

Widely spread after advent of web search engines in 1990s.Useful for novice searchers.The most common use of this type of searching is with a query of a small number of words, also known as a natural language query.Documents are typically ranked by their closeness of fit to the query, hence called relevance ranking.The entire approach has also been called lexical–statistical retrieval.Give a score based on the sum of the weights of terms common to the document and query.

6. Conclusions

Healthcare data faces significant challenges, including data quality and integration, where diverse medical records lack standardized formats. Privacy and security concerns arise due to sensitive patient information. Standardization issues affect interoperability, while data storage and transfers pose logistical problems. Additionally, data structure issues complicate analysis, and infrastructure scalability struggles to keep up with the growing volume of data. These challenges hinder the effective use of big data in healthcare, impacting decision-making and patient care.

About Authors

Radhika Kanubaddhi

Radhika Kanubaddhi has over 12 years of experience as a software developer, architect, and solutions provider. She is currently working as AI specialist at Amazon Web Services. Previously, she worked as Cloud Solutions Architect at Microsoft. Before that, she developed and deployed machine learning models to improve revenues, profits, increase conversions for clients from various industries such as airlines, banks, pharma, retail, and hospitality. She is an expert in assembling the right set of services to solve client needs.

Radhika has worked with almost all technical innovations and services in the last decade - including Internet of Things, cloud application development, ML models, AI apps, Azure, AWS etc.

Radhika also creates content to share her knowledge and enthusiasm for the latest technical innovations.

Radhika has Master's in Computer Science and Bachelor's in Information Technology.

SaidaiahYechuri

SaidaiahYechuri is an engineer with more than 12 years of experience in applying cutting edge innovations. He has done original research work on medical data anonymization published in Accosiation for Computing Machinary.

Venkata Ramana Kandula

Venkata Ramana Kandual has over a decade of experience as a software developer, product manager and architect in a variety of industries. His most recent as a product manager at a Fortune 5 healthcare company involved application of NLP on healthcare data. He has designed and delivered highly impactful products throughout his career.

References

[1] "Turing.com," [Online].

[2] "Krallinger et al. Genome Biology 2008 9(Suppl 2):S8 doi:10.1186/gb-2008-9-s2-s8".

[3] "nlp.stanford.edu," [Online].

[4] "Biomedical Informatics: Natural Language and Text Processing in Biomedicine, CAROL FRIEDMAN AND STEPHEN B. JOHNSON.".

[5] "Information Retrieval and Digital Libraries, WILLIAM HERSH, P. ZOË STAVRI, AND WILLIAM M. DETMER".