MACHINE LEARNING AND AI FOR HEALTHCARE

By Arjun Panesar

Book Review

NATURAL LANGUAGE PROCESSING

#1

What is NLP?

- NLP refers to the use of AI to understand, analyze, and generate human language in both text and speech.
- It's part of computational linguistics and helps systems work with structured and unstructured data.

Applications of NLP

- Data Retrieval: Searching clinical notes by keyword.
- Social Media Monitoring: Tracking trends or sentiment.
- Virtual Assistants: Like Alexa or Siri answering questions.
- Machine Translation: Translating one language to another.
- Sentiment Analysis: Deciphering mood from tweets or reviews.
- Image to text recognition: for instance, reading a sign or menu

Challenges in NLP

- Ambiguity: Words and sentences can have multiple meanings.
- Complex Rules: Languages have nuanced rules that vary in usage.
- **Example:**
  1. *Read a book*
  2. *Book a ticket*

Strength of NLP

- With most data generated existing in the form of unstructured data, NLP is a powerful tool to interpret and understand natural language.

TEXT PROCESSING

#2

Tokenization and Tokens

- This means breaking down a large chunk of text into smaller pieces called tokens.
- Think of a paragraph being cut into sentences, sentences into words, or even smaller units like characters.

- **Example**:
  - *Original text: "Cats love milk."*
  - *Tokens: ["Cats", "love", "milk", "."]*

Text Object

- A text object can be anything from a single word, a sentence, a paragraph, or even an entire article.

- **Example:**
  - *Text object: "Reading is fun." This sentence is a text object.*

Stems and Stemming

- This is a way to simplify words by removing common endings like “-ing,” “-ly,” or “-s.”
- It’s not always perfect, but it helps group words with similar meanings together.

- **Example:**
  - *Words: "playing," "played," "player.“*
  - *Stems: "play."*

Lemmatization

- Unlike stemming, lemmatization looks for the actual root word (or dictionary form) using rules and language analysis.
- It gives a more accurate result than stemming

- **Example:**
  - *Word: "running.“*
  - *Lemma (root): "run."*

Morpheme

- A morpheme is the smallest part of a word that has meaning.

- **Example:**
  - *In the word "unhappiness," there are three morphemes:*
  - *"un-" (a prefix meaning “not”),*
  - *"happy" (the base word),*
  - *"-ness" (a suffix meaning “state of”).*

Syntax

- Syntax is about how words are arranged to make a sentence that makes sense.
- It’s like grammar rules for putting words in the right order

- **Example:**
  - *Correct syntax: "The dog chased the cat.*
  - *"Incorrect syntax: "Chased the dog cat the."*

Semantics and Pragmatics

- Semantics

  This is about the meaning of words and how we combine them to form meaningful phrases or sentences. It focuses on what the words mean.

- Pragmatics

  Pragmatics is about how we use and interpret sentences in real-life situations. It considers tone, context, and how meaning can change depending on the situation. It's less about the literal meaning of words and more about the intended meaning.

Semantics and Pragmatics

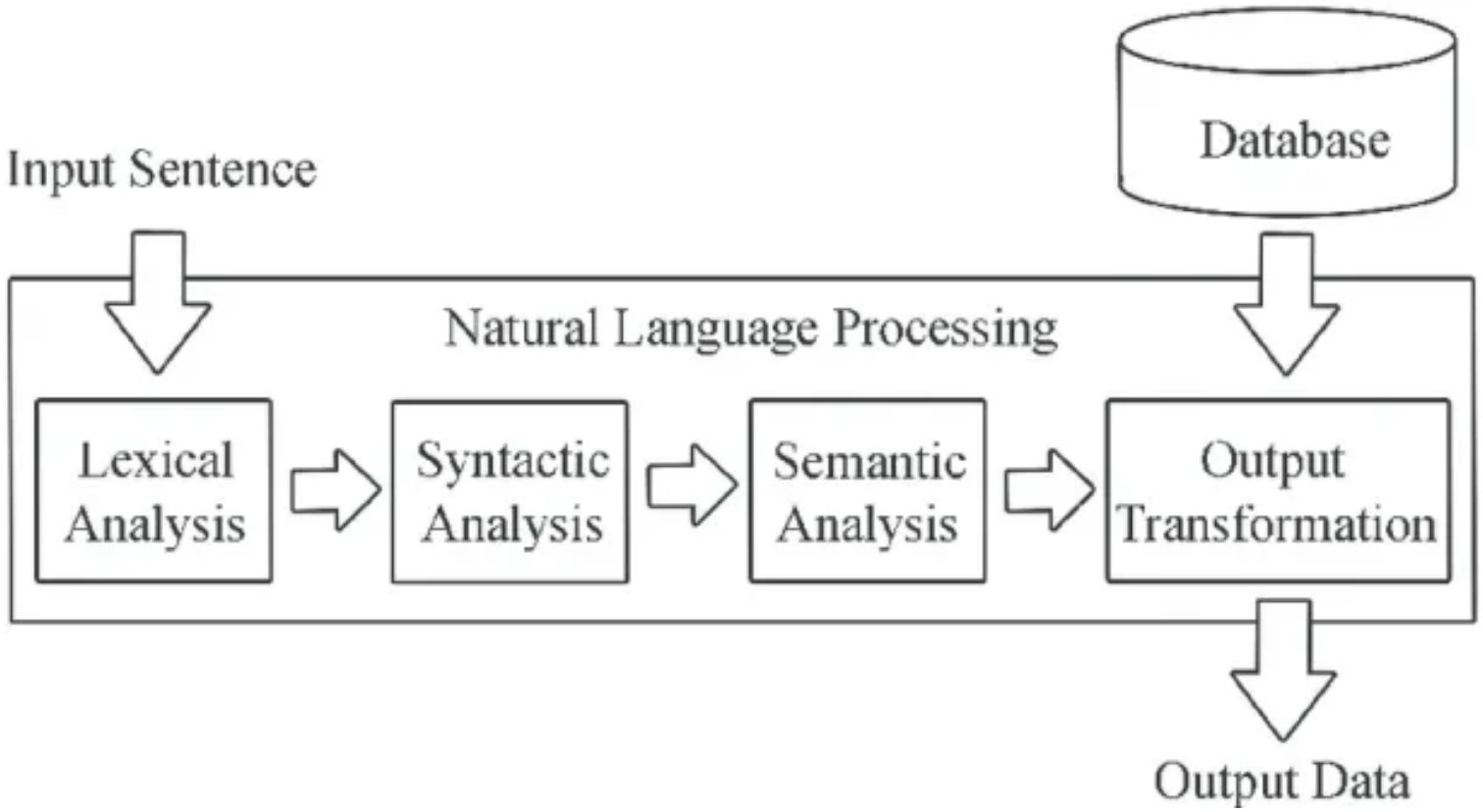
- "Wow, it’s cold in here."
- **Semantics**: The words mean that the temperature is low.
- **Pragmatics**: The speaker might not just be stating a fact but hinting that they want the heater turned on.

- Semantics helps understand *what is being said*, while pragmatics helps understand *why it is being said*.

GETTING STARTED WITH NLP

#3

How NLP works



LEXICAL ANALYSIS

#4

Lexical Analysis or Preprocessing

- Before analyzing text, it's important to clean it up and standardize it. This step helps remove unnecessary or irrelevant parts of the text, often called **"noise,"** so the data is ready for further analysis.

1. Noise Removal

- This step removes unnecessary words or symbols that don't add value to the analysis. For example, common words like "the," "a," "of," and "this" are often removed because they appear frequently but don't provide much context.

- **Example:**

- *Original Text: "The cat is on the tree."*
- *After Noise Removal: "cat on tree"*

2. Lexicon Normalization

- Similar words that have slight variations are simplified to a single form. This makes the analysis easier by grouping these variations under one root word.
- **Two Techniques:**
  - ***Stemming:*** *Removes prefixes and suffixes from words.*
    Example: "running," "runner," and "ran" become "run."
  - ***Lemmatization:*** *Converts words to their dictionary base form.*
    Example: "better" becomes "good," and "was" becomes "be."
- **Why It's Useful:**
  It reduces the number of unique words in the text without losing much meaning.

3. Porter Stemmer Algorithm

- This is a specific method for stemming that groups related words by finding their common root.
- **Example:**
  - *Original Text: "I felt troubled by the trouble my friend was facing."*
  - *After Stemming: "I felt trouble by the trouble my friend was face."*
- **Benefits:**

- Reduces similar words to a single form.
- Helps count words accurately.
  For instance, instead of counting "troubled," "trouble," and "troubling" as three separate words, they are all counted as "trouble."

- **Drawback:**
  Sometimes, the meaning of the word may get slightly distorted.

4. Object Standardization

- Modern text often contains informal words or symbols, especially on platforms like social media. Words like "RT" (retweet), "DM" (direct message), or slang need to be converted to standard forms or removed.
- **Example:**
  - *Original Text: "DM me if you liked my RT of the tweet!"*
  - *After Standardization: "Message me if you liked my share of the post!"*
- **How It's Done:**
  Using dictionaries of slang and abbreviations or regular expressions

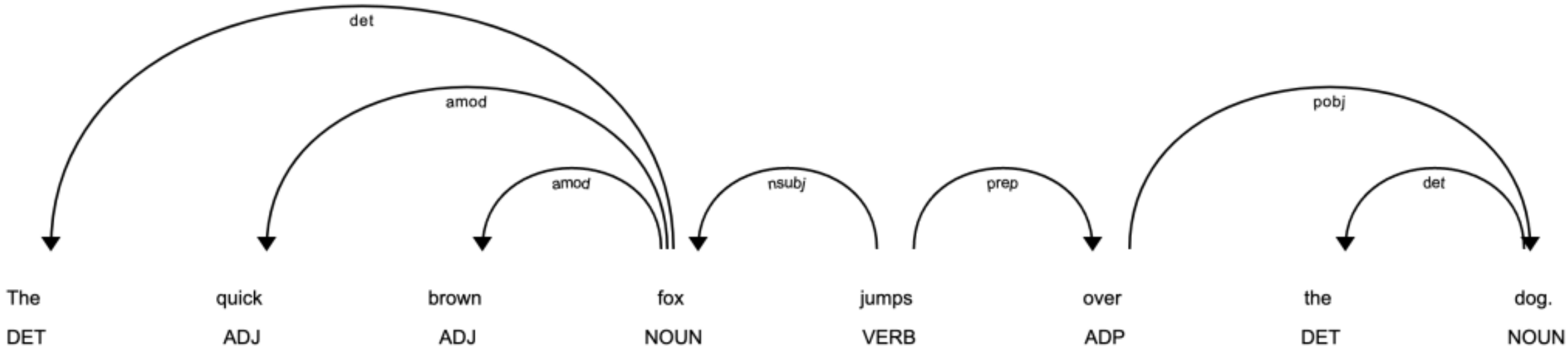
SYNTACTIC ANALYSIS

#5

Syntactic Analysis in Text Processing

- Syntactic analysis focuses on understanding the structure of sentences by identifying the relationships between words.
- It converts sentences into features that can be analyzed. One common method for this is Context-Free Grammar, a straightforward and widely used system for analyzing sentence structure.

Syntactic Analysis in Text Processing



1. Dependency Parsing

- Dependency parsing breaks a sentence into its parts (like the subject, verb, and object) and shows how they are related. It organizes this information into a tree-like structure.
- **Example:**
  For the sentence, “David saw a patient with uncontrolled type 2 diabetes,”
  - The root is the verb **"saw."**
  - **"David"** is the subject.
  - **"Patient"** is the object.

- Dependency parsing visually maps out these relationships.

2. Part of Speech (POS) Tagging

- POS tagging assigns each word in a sentence to a specific category like noun, verb, adjective, or adverb.
- **Example:**
- In the sentence, “I managed to read my book on the train,”
  - ***"Read"*** *is a verb.*
  - ***"Book"*** *is a noun.*
- In “Can you please book my train tickets?”
  - ***"Book"*** *is a verb.*
- **Why It’s Useful:**
- Resolves ambiguity (e.g., recognizing the different meanings of "book").
- Helps with tasks like sentiment analysis or answering questions.

Key Differences

Aspect	POS Tagging	Dependency Parsing
Level of Analysis	Word-level	Sentence-level
Goal	Assign grammatical category to words	Identify syntactic relationships
Complexity	Relatively simpler	More complex, involves sentence structure
Output	Sequence of POS tags	Dependency tree
Example Task	Tagging "run" as a verb or noun	Identifying "run" as the root verb with "quickly" as its modifier

SEMANTIC ANALYSIS

#6

Semantic Analysis

- Semantic analysis goes beyond structure to find the exact meaning of words and sentences. It considers context to interpret text accurately.

- **Example:**
  - *"Bat" could mean an animal or a piece of sports equipment. Semantic analysis figures out which meaning is correct based on the context of the sentence.*

TRANSFORMING TEXT INTO USABLE DATA

#7

After Preprocessing

- Once data preprocessing, lexical, syntactical, and semantic analysis of corpus has taken place, we are required to transform text into mathematical representations for evaluation, comparison, and retrieval.
- For instance, searching a collection of patient profiles for users with “hypertension” should only bring out those with hypertension. This is achieved through transforming documents into the vector space model with scoring and term weighting essential for query ranking and search retrieval.
- Documents can be in the form of patient records, web pages, digitalized books, and so forth.

1. N-grams

- N-grams are sequences of words from a sentence.

- **Example:**
  - ***Sentence:*** *David reversed his metabolic syndrome”*
  - ***Unigram (N = 1):*** *Single words. Example: "David," "reversed," "syndrome."*
  - ***Bigram (N = 2):*** *Two-word sequences. Example: "David reversed," "reversed his," “metabolic syndrome"*
  - ***Trigram (N = 3):*** *Three-word sequences. Example: "David reversed his." “his metabolic syndrome“*

- **Why It’s Useful:**
  N-grams help with tasks like spelling correction or text summarization.

2. TF-IDF Vectors

- **Term Frequency-Inverse Document Frequency** assigns importance to words in a document based on how often they appear in the document versus how common they are in other documents.

- **Example:**
  - *In a medical report, the word “hypertension” might have a high score if it appears often in that report but rarely in other documents. This shows it’s important in that context.*

Latent Semantic Analysis (LSA)

- LSA identifies relationships between words by analyzing large amounts of text. It uses mathematical techniques to group related terms.

- **Example:**
  - *If a document contains the word **"sea,"** it might also include **"beach"** because these terms are often related. LSA finds these patterns automatically.*

Latent Semantic Analysis (LSA)

- To do this, LSA creates a large table, called a matrix, where each row represents a document and each column represents a word. The values in the table show how often each word appears in each document.

- Then, LSA uses a mathematical technique called **Singular Value Decomposition (SVD)** to break down this matrix. This helps to uncover hidden relationships or patterns between words. Through this process, LSA can discover which words tend to appear together in documents, and what they mean in context, even if the word itself isn't explicitly mentioned.

Latent Semantic Analysis (LSA)

- Document 1: "The sea is beautiful.“
- Document 2: "The beach is relaxing.“
- Document 3: "I love swimming in the sea.“

	sea	beach	is	beautiful	relaxing	love	swimming
Document 1	1	0	1	1	0	0	0
Document 2	0	1	1	0	1	0	0
Document 3	1	0	1	0	0	1	1

AIHRC

THANK YOU