Natural Language Processing

Lecture-1

Introduction + Course outline

Introduction

According to industry estimates, more than 80% of the data being generated is in an unstructured format, maybe in the form of text(natural language), image, audio, video

Introduction

Data is getting generated as we

- *Speak*
- *Write*
- *Tweet*
- *Use Social Media platforms*
- *Send messages on various messaging platforms*
- *Use e-commerce for shopping*

……..

Introduction

The majority of this data exists in the textual(natural language) form

Unstructured data

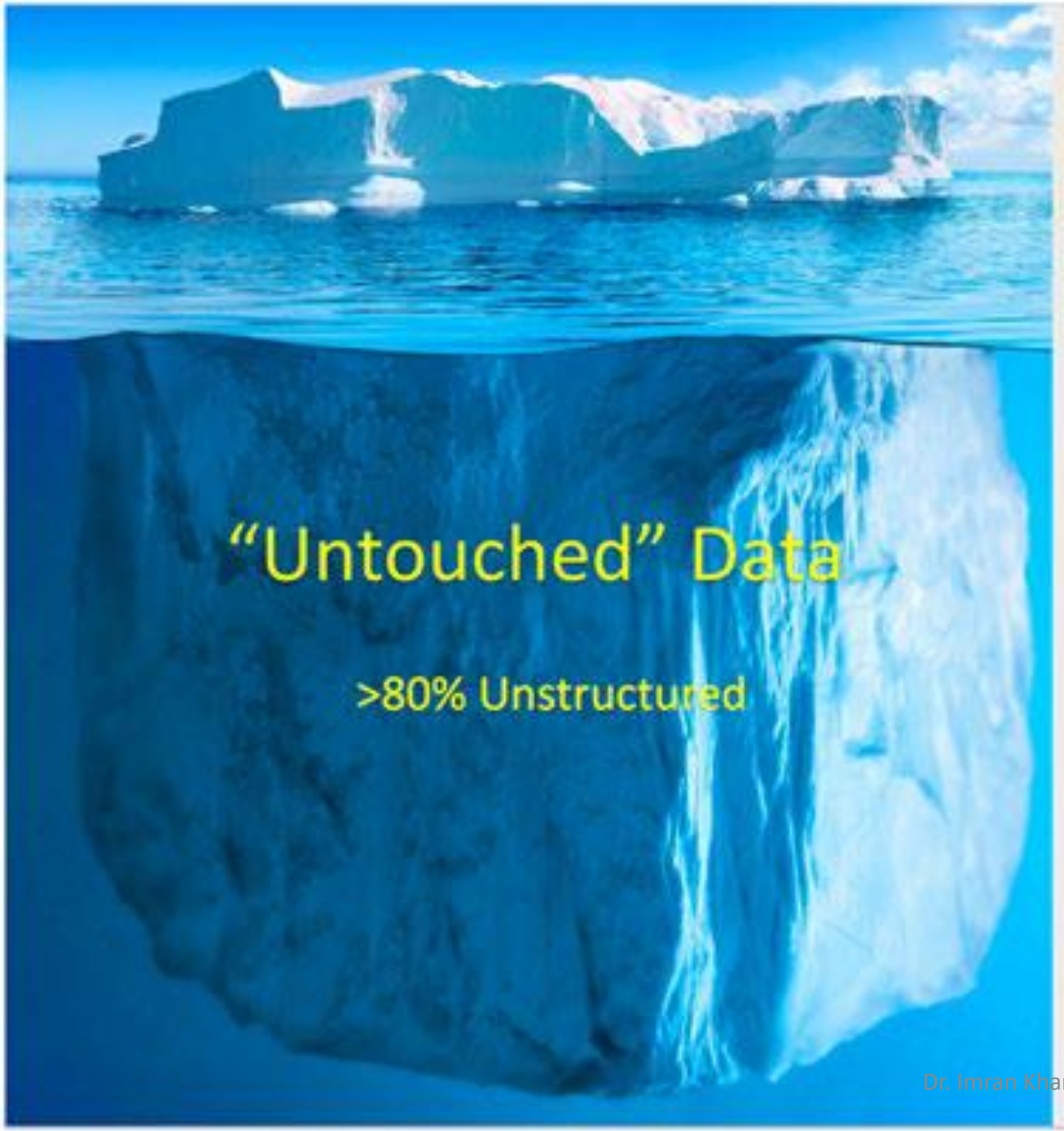
Unstructured data is the information that doesn't reside in a traditional relational database.

Examples include

- Documents, blogs, social media feeds, pictures, and videos

Why Analyzing Unstructured Data

- Most of the insight is locked in unstructured data.
- Text data is most common and covers more than 50% of the unstructured data
- Unlocking it plays a vital role in every organization to make improved and better decisions.



80%

of data generated is unstructured in nature, and growing exponentially

40%

of business executives complain that they have too much unstructured text data and are unable to interpret them

Use of Text Analytics can help extract Insights from unstructured text

Imagine being able to extract insights from Unstructured text and use it to make business decisions. It can be a huge differentiator in this competitive environment

Natural Language Processing

In order to produce significant and actionable insights from natural language(text data), we use Natural Language Processing coupled with machine learning and deep learning.

Natural Language Processing

Definition

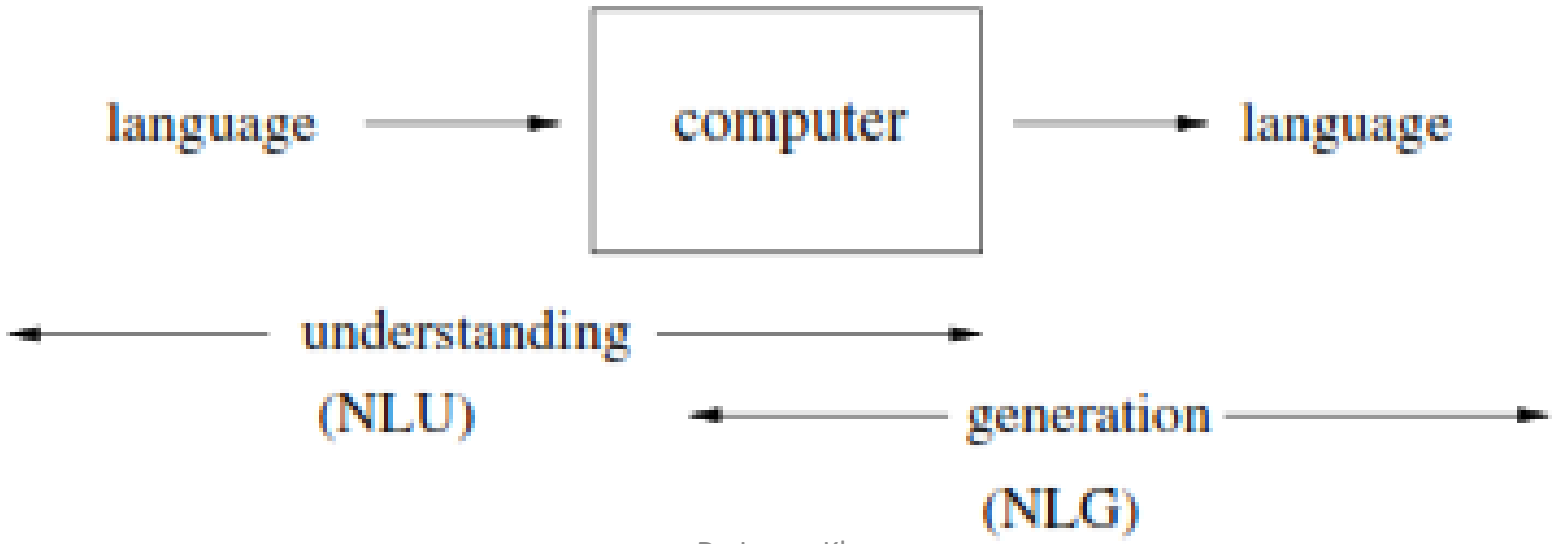
what is Natural Language Processing - NLP?

We all know that machines/algorithms cannot understand texts or characters, so it is very important to convert these **text data** into machine understandable format (like numbers or binary) to perform any kind of analysis on text data

Natural Language Processing

Definition

Natural language processing is an area of research in computer science and artificial intelligence (AI) concerned with processing natural languages such as English



Natural Language Processing

Goal

The goal of NLP is to make machines understand our spoken and written languages…….

more recent ones include voice-driven bots like **Siri**, **Alexa**

Natural Language Processing

What we Learn

- You will learn how to efficiently use a wide range of NLP packages and implement
- text classification,
- identify parts of speech,
- topic modeling,
- text summarization,
- text generation,
- sentiment analysis,
- and many more applications of NLP

Natural Language Processing

What we Learn

You will learn

- ways of extracting text data along with web scraping
- how to clean and preprocess text data and ways to analyze
- explore the semantic as well as syntactic analysis of the text
- text normalization,
- advanced preprocessing methods,
- POS tagging,
- text similarity,
- text summarization,
- sentiment analysis,
- topic modeling,
- word2vec, seq2seq,

Natural Language Processing

What we Learn

Most Important for Implementation

- Working in Python with NLP Packages / Libraries

  i.e. NLTK, TextBlob, SpaCy, Stanford CoreNLP

- Implementing text preprocessing and feature engineering

  i.e. like word embedding.

Working with Different Data Sets

- Implementing an end-to-end pipeline of the NLP life cycle, which
  - includes framing the problem,
  - finding the data,
  - collecting,
  - preprocessing the data,
  - solving it using state-of-the-art techniques.

Natural language pipeline

*A **natural language processing system** is often referred to as a **pipeline***

Natural language Libraries

- **NLTK: Natural language toolkit and commonly called the mother of all** NLP libraries
- **SpaCy: SpaCy is recently a trending library, as it comes with the added** flavors of a deep learning, While SpaCy doesn't cover all of the NLP functionalities
- **TextBlob: This is one of the data scientist's favorite library when it** comes to implementing NLP tasks. It is based on both NLTK and Pattern. However, TextBlob certainly isn't the fastest or most complete library.
- **CoreNLP: It is a Python wrapper for Stanford CoreNLP. The toolkit** provides very robust, accurate, and optimized techniques for tagging, parsing, and analyzing text in various languages

There are hundreds of NLP libraries

NLP Course Output

By the end of the course you will be able to do

- Sentiment analysis: Customer's emotions toward products offered by the business.
- Topic modeling: Extract the unique topics from the group of documents.
- Complaint classifications/Email classifications/ E- commerce product classification, etc.
- Document categorization/management using different clustering techniques.
- Resume shortlisting and job description matching using similarity methods.

NLP Course Output

- Advanced feature engineering techniques (word2vec and fastText) to capture context..
- Information/Document Retrieval Systems, for example, search engine.
- Chatbot, Q & A, and Voice-to-Text applications like Siri and Alexa
- Language detection and translation using neural networks.
- Text summarization using graph methods and advanced techniques
- Text generation/predicting the next sequence of words using deep learning algorithms.

Steps in Text (Natural Language) Analysis

- **Data collection**
- **Text Preprocessing**
- **Text to feature**
- **Machine learning / Deep learning**

Data Source Freely Available

Huge amount of data is freely available over the internet

- start exploring multiple free data sources

- Free APIs like Twitter, Facebook, Amazon etc.
- Wikipedia
- Government data (e.g. http://data.gov)
- Census data (e.g. http://www.census.gov/data.html)
- Health care claim data (e.g. https://www.healthdata.gov/)

Other Data Source

Client Data (Own data that is already present)

- SQL databases
- Hadoop clusters
- Cloud storage
- Flat files

Web scraping

- Extracting the content/data from websites, blogs, forums, and retail websites for reviews with the permission from the respective sources using web scraping packages

lot of other sources like crime data, accident data, and economic data

NLP Case Study

Virtual Assistants (VAs)

- Google Assistant
- Cortana
- Apple Siri,

are largely NLP systems.

NLP Case Study

Asks a Virtual Assistant (VA)

“Can you show me a good Italian restaurant nearby?".

VA will perform various NLP tasks to process our query

NLP Case Study

NLP Tasks Performed by VA

- Convert the sound to text (that is, speech-to-text).
- Understand the semantics of the request and formulate a structured request (for example, cooking = Italian, rating = 3-5, distance< 10 km).
- Search for restaurants filtering by the location and cooking, and then, sort the restaurants by the ratings received.

NLP Case Study

NLP Tasks Performed by VA

- Calculate an overall rating for a restaurant by both the rating and text description provided by each user.

- Finally, once the user is at the restaurant, the VA might assist the user by translating various menu items from Italian to English.

Other NLP Systems

- Searching for today's weather on Google
- Google Translate to find out how to say, "How are you?" in French

..... *and the list continues*

Moral of the lesson

A good NLP system is that which performs many NLP tasks

ATIS example

User: I need a flight from Boston to Washington, arriving by 10 pm.

System: What day are you flying on?

User: Tomorrow

System: Returns a list of flights

What Is a Corpus?

- The plural form of corpus is corpora.
- The corpus may be composed of written language, spoken language or both. Spoken corpus is usually in the form of audio recordings.
- Corpora(plural of corpus) are collections of related documents that contain natural language.
- A corpus can be large or small, though generally they consist of dozens or even hundreds of gigabytes of data inside of thousands of documents.
- Some popular corpora are British National Corpus (BNC), COBUILD/Birmingham Corpus, IBM/Lancaster Spoken English Corpus.

Monolingual and Bilingual Corpora

- Monolingual corpora represent only one language while bilingual corpora represent two languages.

For example

- European Corpus Initiative (ECI) corpus is multilingual having 98 million words in Turkish, Japenese, Russian, Chinese, and other languages.

Open or closed Corpus

- An *open corpus* is one which does not claim to contain all data from a specific area while a *closed corpus* does claim to contain all or nearly all data from a particular field.

- *Historical corpora*, for example, are closed as there can be no further input to an area.

Most ubiquitous NLP tasks

Tokenization

Tokenization is the task of separating a text corpus into atomic units (for example, words).

Word-sense Disambiguation (WSD):

- WSD is the task of identifying the correct meaning of a word.

For example,

- *The dog barked at the mailman*,

and

- *Tree bark is sometimes used as a medicine*,

The word *bark* has two different meanings. WSD is critical for tasks such as question answering.

Named Entity Recognition (NER):

- NER attempts to extract entities(For example, person, location, and organization). from a given body of text or a text corpus.

For example, the sentence,

- *John gave Mary two apples at school on Monday*

will be transformed to

- *[John]name gave [Mary]name [two]number apples at [school]organization on [Monday.]time*.

NER is an imperative topic in fields such as information retrieval and knowledge representation.

Part-of-Speech (PoS) tagging

- PoS tagging is the task of assigning words to their respective parts of speech.

For example

- It can either be basic tags such as noun, verb, adjective, adverb, and preposition

OR

- It can be granular such as proper noun, common noun, phrasal verb, verb, and so on.

Sentence/Synopsis classification

Sentence or synopsis classification has many use cases such as

- Spam detection
- News article classification (for example, political, technology, and sport)
- Product review ratings (that is, positive or negative).

Language generation

- Predict new text based on previous text .

Question Answering (QA)

QA techniques are found at the foundation of chatbots and VA (for example, Google Assistant and Apple Siri).

Machine Translation (MT)

- MT is the task of transforming a sentence/phrase from a source language (for example, German) to a target language (for example, English)

Finally, to develop a system that can assist a human in day-to-day tasks (for example, VA or a chatbot) many of these tasks need to be performed together.

Chatbot

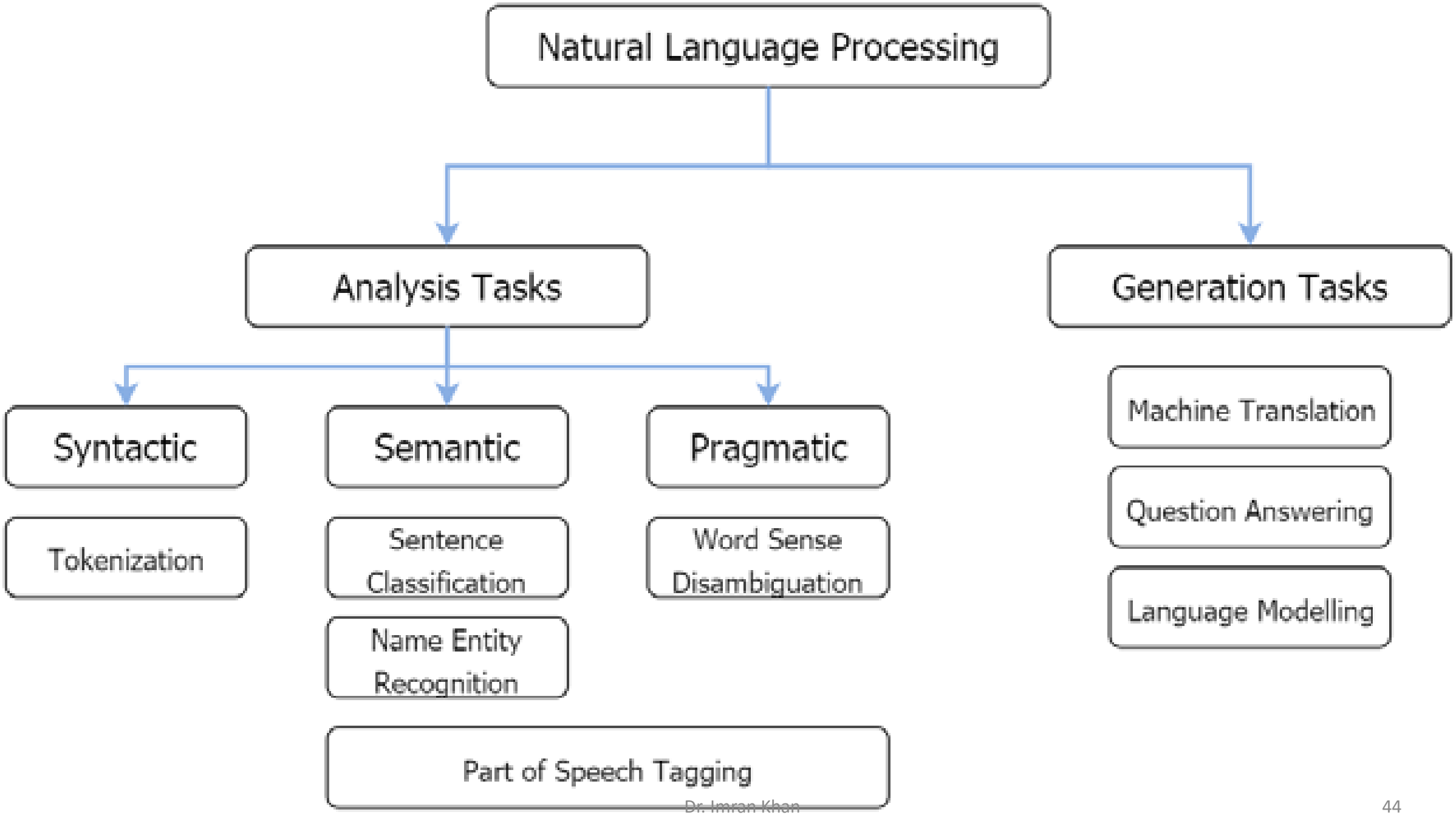
- Chatbots have been adopted by many companies for customer support.

- Chatbots can be used to answer and resolve straightforward customer concerns, which can be solved without human intervention

For example

- Changing a customer's monthly mobile plan.

Hierarchical taxonomy of different NLP tasks


Natural Language Processing
Analysis Tasks
Generation Tasks
Syntactic
Semantic
Pragmatic
Machine Translation
Question Answering
Language Modelling
Tokenization
Sentence Classification
Word Sense Disambiguation
Name Entity Recognition
Part of Speech Tagging

Taxonomy

We first have two broad categories:

1. Analysis (analyzing existing text)
2. Generation (generating new text) tasks.

Then we divide analysis into three different categories: syntactic (language structure-based tasks),semantic (meaning-based tasks), and pragmatic (open problems difficult to solve):

Final Output of the Course

Pick a technical paper and reproduce their results

Make sure the model is reasonably technically demanding

Pick an existing algorithm or learning model and design a new enhanced version

Apply an existing model to a new domain or an application

Make sure to provide rigorous analysis and/or experiment with new model variations

Make a new dataset and conduct annotation studies

Make sure to provide baseline results

Python
Programming language

Java
Programming language

R
Programming language