Introduction to Natural Language Processing

Unit 1: Introduction to NLP

Rupak Raj Ghimire

School of Mathematical Sciences
IOST, TU

July 18, 2025

Table of contents

- Introduction
- Level of Linguistics
- Challenges of NLP
- Terminologies
- NLP Pipeline
- NLP problems

Introduction

What is Language?

Defination by Britannica:
Language is a **system** of **conventional spoken, manual (signed), or written** symbols by means of which human beings, as members of a social group and participants in its culture, express themselves.

The function of language include - communication, the expression of identity, play, imaginative expression, and emotional release

Based on this definition we can divide the language data into

- Written (Text)
- Spoken (Speech)

What is Language?

Defination by Britannica:
Language is a **system** of **conventional spoken, manual (signed), or written** symbols by means of which human beings, as members of a social group and participants in its culture, express themselves.

The function of language include - communication, the expression of identity, play, imaginative expression, and emotional release

Based on this definition we can divide the language data into

- Written (Text)
- Spoken (Speech)

What is Language?

Defination by Britannica:
Language is a **system** of **conventional spoken, manual (signed), or written** symbols by means of which human beings, as members of a social group and participants in its culture, express themselves.

The function of language include - communication, the expression of identity, play, imaginative expression, and emotional release

Based on this definition we can divide the language data into

- Written (Text)
- Spoken (Speech)

What is Language?

Defination by Britannica:
Language is a **system** of **conventional spoken, manual (signed), or written** symbols by means of which human beings, as members of a social group and participants in its culture, express themselves.

The function of language include - communication, the expression of identity, play, imaginative expression, and emotional release

Based on this definition we can divide the language data into

- Written (Text)
- Spoken (Speech)

What do we use language for?

- We **communicate** using language
- We **think** (partly) with language
- We **tell stories** in language
- We build **Scientific Theories** with language
- We make friends/build **relationships**

What do we use language for?

- We **communicate** using language
- We **think** (partly) with language
- We **tell stories** in language
- We build **Scientific Theories** with language
- We make friends/build **relationships**

What do we use language for?

- We **communicate** using language
- We **think** (partly) with language
- We **tell stories** in language
- We build **Scientific Theories** with language
- We make friends/build **relationships**

What do we use language for?

- We **communicate** using language
- We **think** (partly) with language
- We **tell stories** in language
- We build **Scientific Theories** with language
- We make friends/build **relationships**

What do we use language for?

- We **communicate** using language
- We **think** (partly) with language
- We **tell stories** in language
- We build **Scientific Theories** with language
- We make friends/build **relationships**

What is Language? - another pespective

- **Language:** A system of symbols and rules used for communication.
- **Types of Language:**
  - Natural Language (e.g., English, Nepali)
  - Programming Language (e.g., Python, C++)

  What make Natural Language different over programming language?
- **Characteristics of Natural Language:**
  - Ambiguity
  - Context dependence
  - Evolving and diverse

What is Language? - another pespective

- **Language:** A system of symbols and rules used for communication.
- **Types of Language:**
  - Natural Language (e.g., English, Nepali)
  - Programming Language (e.g., Python, C++)

  What make Natural Language different over programming language?

- **Characteristics of Natural Language:**
  - Ambiguity
  - Context dependence
  - Evolving and diverse

What is Language? - another pespective

- **Language:** A system of symbols and rules used for communication.
- **Types of Language:**
  - Natural Language (e.g., English, Nepali)
  - Programming Language (e.g., Python, C++)

  What make Natural Language different over programming language?

- **Characteristics of Natural Language:**
  - Ambiguity
  - Context dependence
  - Evolving and diverse

What is NLP?

It is a sub-field of artificial intelligence that is concerned with the human computer interaction using natural language. It is **interdisciplinary** field Computer and Linguistics.

We can 1) **Analyze** and 2) **Produce** the natural language using NLP techniques

Goals:

- Understand, interpret, and generate human language
- Automate language-related tasks

What is NLP?

It is a sub-field of artificial intelligence that is concerned with the human computer interaction using natural language. It is **interdisciplinary** field Computer and Linguistics.

We can 1) **Analyze** and 2) **Produce** the natural language using NLP techniques

Goals:

- Understand, interpret, and generate human language
- Automate language-related tasks

What is NLP?

It is a sub-field of artificial intelligence that is concerned with the human computer interaction using natural language. It is **interdisciplinary** field Computer and Linguistics.

We can 1) **Analyze** and 2) **Produce** the natural language using NLP techniques

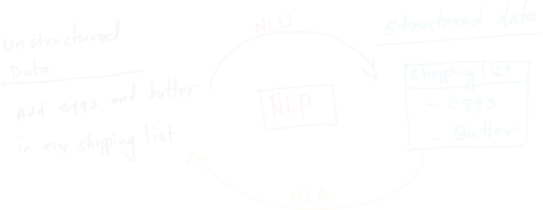
Goals:

- Understand, interpret, and generate human language
- Automate language-related tasks

What is NLP?(cont.)

If we take any form of the language (written or spoken), the format is always **unstructured**

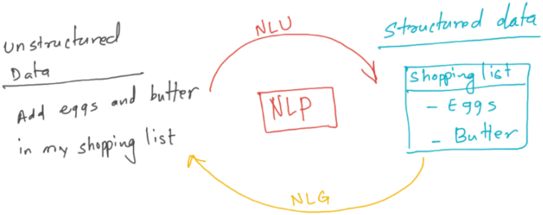
- NLU: Natural Language Understanding
- NLG: Natural Language Generation



What is NLP?(cont.)

If we take any form of the language (written or spoken), the format is always **unstructured**

- NLU: Natural Language Understanding
- NLG: Natural Language Generation



What is NLP?(cont.)

The NLP can be divided into two categories: Text and Speech Processing.

Text Processing

- Machine translation
- Spam email detection
- Document classification
- Text summary generation
- Sentiment Analysis

Speech Processing

- Text-to-speech Generation (Speech Synthesis)
- Automatic Speech Recognition

What is NLP?(cont.)

The NLP can be divided into two categories: Text and Speech Processing.

Text Processing

- Machine translation
- Spam email detection
- Document classification
- Text summary generation
- Sentiment Analysis

Speech Processing

- Text-to-speech Generation (Speech Synthesis)
- Automatic Speech Recognition

What is NLP? (cont.)

The NLP can be divided into two categories: Text and Speech Processing.

Text Processing

- Machine translation
- Spam email detection
- Document classification
- Text summary generation
- Sentiment Analysis

Speech Processing

- Text-to-speech Generation (Speech Synthesis)
- Automatic Speech Recognition

Why NLP?

Some use case of the NLP are:

Access Knowledge

- search engine, recommend system

Communicate

- Translation, synthesis, recognition

Linguistics and Cognitive Sciences

- Analyse Languages themselves

Why NLP?

Some use case of the NLP are:

Access Knowledge

- search engine, recommend system

Communicate

- Translation, synthesis, recognition

Linguistics and Cognitive Sciences

- Analyse Languages themselves

Why NLP?

Some use case of the NLP are:

Access Knowledge

- search engine, recommend system

Communicate

- Translation, synthesis, recognition

Linguistics and Cognitive Sciences

- Analyse Languages themselves

Why NLP?

Some use case of the NLP are:

Access Knowledge

- search engine, recommend system

Communicate

- Translation, synthesis, recognition

Linguistics and Cognitive Sciences

- Analyse Languages themselves

Why NLP?(cont.)

Some products

- Search: Google, Baidu
- Social Media: Facebook, Instagram, WeChat, Twitter
- Machine Translation: google translate
- Voice Assistant: Alexa, Siri, Google Assistant
- Chat Bots: using conventional agents

Why NLP?(cont.)

Some products

- Search: Google, Baidu
- Social Media: Facebook, Instagram, WeChat, Twitter
- Machine Translation: google translate
- Voice Assistant: Alexa, Siri, Google Assistant
- Chat Bots: using conventional agents

Why NLP?(cont.)

Some products

- Search: Google, Baidu
- Social Media: Facebook, Instagram, WeChat, Twitter
- Machine Translation: google translate
- Voice Assistant: Alexa, Siri, Google Assistant
- Chat Bots: using conventional agents

Why NLP?(cont.)

Some products

- Search: Google, Baidu
- Social Media: Facebook, Instagram, WeChat, Twitter
- Machine Translation: google translate
- Voice Assistant: Alexa, Siri, Google Assistant
- Chat Bots: using conventional agents

Why NLP?(cont.)

Some products

- Search: Google, Baidu
- Social Media: Facebook, Instagram, WeChat, Twitter
- Machine Translation: google translate
- Voice Assistant: Alexa, Siri, Google Assistant
- Chat Bots: using conventional agents

Why NLP? (cont.)

Some products

- Search: Google, Baidu
- Social Media: Facebook, Instagram, WeChat, Twitter
- Machine Translation: google translate
- Voice Assistant: Alexa, Siri, Google Assistant
- Chat Bots: using conventional agents

Brief History of NLP

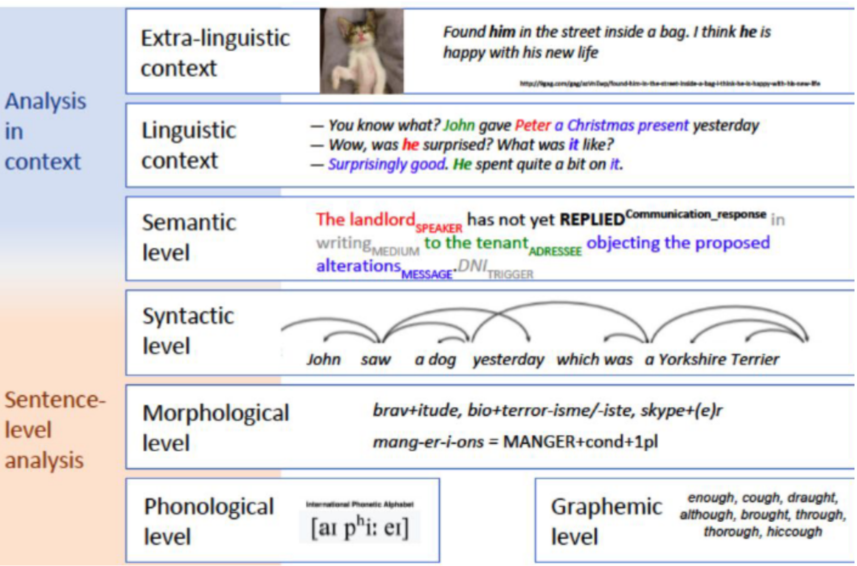
Era	Key Developments
1950s-60s	Rule-based systems (ELIZA, symbolic AI)
1980s-90s	Statistical NLP (HMMs, n-grams)
2000s	Machine Learning (SVMs, CRFs)
2010s-Now	Deep Learning and Transformers (BERT, GPT)

- Transition from rules → statistics → deep learning
- Rise of large pre-trained language models

Level of Linguistics

Five level of linguistics

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics (context)
- Extra-linguistic - other material along with language



Analysis in context

Extra-linguistic context

*Found **him** in the street inside a bag. I think **he** is happy with his new life*

http://9gag.com/gag/aVn1wp/found-him-in-the-street-inside-a-bag-i-think-he-is-happy-with-his-new-life

Linguistic context

— *You know what? John gave Peter a Christmas present yesterday*
— *Wow, was **he** surprised? What was **it** like?*
— *Surprisingly good. **He** spent quite a bit on **it**.*

Semantic level

The landlord$_{\text{SPEAKER}}$ has not yet **REPLIED**$^{\text{Communication\_response}}$ in writing$_{\text{MEDIUM}}$ to the tenant$_{\text{ADRESSEE}}$ objecting the proposed alterations$_{\text{MESSAGE}}$. *DNI*$_{\text{TRIGGER}}$

Syntactic level

John saw a dog yesterday which was a Yorkshire Terrier

Sentence-level analysis

Morphological level

brav+itude, bio+terror-isme/-iste, skype+(e)r

mang-er-i-ons = MANGER+cond+1pl

Phonological level

International Phonetic Alphabet

[aɪ p^{h}i: eɪ]

Graphemic level

enough, cough, draught, although, brought, through, thorough, hiccough

Phonology

It is a study of the **sounds** of a language

Every language has its own **inventory of sounds** and logical rules for combining those sounds to create words.

The phonology of a language essentially refers to its **sound system** and the processes used to combine sounds in **spoken language**.

Phonology

It is a study of the **sounds** of a language

Every language has its own **inventory of sounds** and logical rules for combining those sounds to create words.

The phonology of a language essentially refers to its **sound system** and the processes used to combine sounds in **spoken language**.

Phonology

It is a study of the **sounds** of a language

Every language has its own **inventory of sounds** and logical rules for combining those sounds to create words.

The phonology of a language essentially refers to its **sound system** and the processes used to combine sounds in **spoken language**.

Morphology

Study of the **internal structure of the words** of a language

There are many words to which a speaker can add a **suffix, prefix, or infix** to create a new word

The morphology of a language refers to the **word-building rules** speakers use to create new words or alter the meaning of existing words in their language

Morphology

Study of the **internal structure of the words** of a language

There are many words to which a speaker can add a **suffix, prefix, or infix** to create a new word

The morphology of a language refers to the **word-building rules** speakers use to create new words or alter the meaning of existing words in their language

Morphology

Study of the **internal structure of the words** of a language

There are many words to which a speaker can add a **suffix, prefix, or infix** to create a new word

The morphology of a language refers to the **word-building rules** speakers use to create new words or alter the meaning of existing words in their language

Syntax

Study of **sentence structure**

Every language has its **own rules for combining words** to create sentences

Syntactic analysis attempts to **define and describe** the rules that speakers use to put words together to create meaningful phrases and sentences

Syntax

Study of **sentence structure**

Every language has its **own rules for combining words** to create sentences

Syntactic analysis attempts to **define and describe** the rules that speakers use to put words together to create meaningful phrases and sentences

Syntax

Study of **sentence structure**

Every language has its **own rules for combining words** to create sentences

Syntactic analysis attempts to **define and describe** the rules that speakers use to put words together to create meaningful phrases and sentences

Semantics

Study of **meaning** in language

Linguists attempt to identify not only how speakers of a language distinguish the meanings of words in their language, but also **how the logical rules speakers apply to determine the meaning of phrases, sentences, and entire paragraphs**

The meaning of a given word can depend on the context in which it is used, and the definition of a word may vary slightly from speaker to speaker

Semantics

Study of **meaning** in language

Linguists attempt to identify not only how speakers of a language distinguish the meanings of words in their language, but also **how the logical rules speakers apply to determine the meaning of phrases, sentences, and entire paragraphs**

The meaning of a given word can depend on the context in which it is used, and the definition of a word may vary slightly from speaker to speaker

Semantics

Study of **meaning** in language

Linguists attempt to identify not only how speakers of a language distinguish the meanings of words in their language, but also **how the logical rules speakers apply to determine the meaning of phrases, sentences, and entire paragraphs**

The meaning of a given word can depend on the context in which it is used, and the definition of a word may vary slightly from speaker to speaker

Pragmatics

Study of the **social use** of language

All speakers of a language use different registers, or different **conversational styles**, depending on the places

A linguistic analysis that focuses on pragmatics may describe the **social aspects of the language** sample being analyzed

Such as how the status of the individuals involved in the speech act could affect the meaning of a given utterance.

Pragmatics

Study of the **social use** of language

All speakers of a language use different registers, or different **conversational styles**, depending on the places

A linguistic analysis that focuses on pragmatics may describe the **social aspects of the language** sample being analyzed

Such as how the status of the individuals involved in the speech act could affect the meaning of a given utterance.

Pragmatics

Study of the **social use** of language

All speakers of a language use different registers, or different **conversational styles**, depending on the places

A linguistic analysis that focuses on pragmatics may describe the **social aspects of the language** sample being analyzed

Such as how the status of the individuals involved in the speech act could affect the meaning of a given utterance.

Pragmatics

Study of the **social use** of language

All speakers of a language use different registers, or different **conversational styles**, depending on the places

A linguistic analysis that focuses on pragmatics may describe the **social aspects of the language** sample being analyzed

Such as how the status of the individuals involved in the speech act could affect the meaning of a given utterance.

Challenges of NLP

Challenges of NLP

1. Productivity
2. Ambiguous
3. Variability
4. Diversity
5. Sparsity

Productivity

Definition: "property of the language-system which enables native speakers to **construct and understand** an **indefinitely large number of utterances**, including utterances that they have never previously encountered." (Lyons, 1977)

New **words, senses, structure** are introduced in languages all the time

Examples: social distance were added to the Oxford Dictionary in 2021

Why Sanskrit is not in common use?

Productivity

Definition: "property of the language-system which enables native speakers to **construct and understand** an **indefinitely large number of utterances**, including utterances that they have never previously encountered." (Lyons, 1977)

New **words, senses, structure** are introduced in languages all the time

Examples: social distance were added to the Oxford Dictionary in 2021

Why Sanskrit is not in common use?

Productivity

Definition: “property of the language-system which enables native speakers to **construct and understand** an **indefinitely large number of utterances**, including utterances that they have never previously encountered.” (Lyons, 1977)

New **words, senses, structure** are introduced in languages all the time

Examples: social distance were added to the Oxford Dictionary in 2021

Why Sanskrit is not in common use?

Productivity

Definition: "property of the language-system which enables native speakers to **construct and understand** an **indefinitely large number of utterances**, including utterances that they have never previously encountered." (Lyons, 1977)

New **words, senses, structure** are introduced in languages all the time

Examples: **social distance** were added to the Oxford Dictionary in 2021

Why Sanskrit is not in common use?

Ambiguous

Most linguistic observations (speech, text) are open to several **interpretations**

How We (Humans) disambiguate?
i.e. find the correct interpretation

- using all kind of signals
- linguistic and extra linguistic signals

Ambiguity can appear at all levels

phonology, graphemics, morphology, syntax, semantics

Ambiguous

Most linguistic observations (speech, text) are open to several **interpretations**

How We (Humans) disambiguate?
i.e. find the correct interpretation

- using all kind of signals
- linguistic and extra linguistic signals

Ambiguity can appear at all levels

phonology, graphemics, morphology, syntax, semantics

Ambiguous

Most linguistic observations (speech, text) are open to several **interpretations**

How We (Humans) disambiguate?
i.e. find the correct interpretation

- using all kind of signals
- linguistic and extra linguistic signals

Ambiguity can appear at all levels

phonology, graphemics, morphology, syntax, semantics

Ambiguous

Most linguistic observations (speech, text) are open to several **interpretations**

How We (Humans) disambiguate?
i.e. find the correct interpretation

- using all kind of signals
- linguistic and extra linguistic signals

Ambiguity can appear at all levels

phonology, graphemics, morphology, syntax, semantics

Ambiguous

Most linguistic observations (speech, text) are open to several **interpretations**

How We (Humans) disambiguate?
i.e. find the correct interpretation

- using all kind of signals
- linguistic and extra linguistic signals

Ambiguity can appear at all levels

phonology, graphemics, morphology, syntax, semantics

Semantic Ambiguity

Polysemy:

eg. set, arm, head - Head of New-Zealand is woman **Name Entity:**

eg. Michael Jordan - Michael Jordan is a professor at Berkeley **Object/Color:**

eg. cherry - Your cherry coat

Variability

Laguage varies at all levels

- Phonetics (accent)
- Morphological, Lexical (spelling)
- Syntaic
- Semantic

Variation Determiners

Who is talking?

- To Whom?
- Where? Work, Home, Restaurant
- When? 19th century, 2008, 2022...
- About what? Specialised domain, the Weather,...

Essentially, the Variability of a language depends on

- Social Context
- Geography
- Sociology
- Date
- Topic

Diversity

About **7000 languages** spoken in the world. About **60% are found in the written form**. Diversity are in

- Phonologic Diversity
- Graphemic Diversity (latin, arabic, devanagari, greek)

Syntatic Diversity

A key characteristics of the syntax of a given language is the word order

- Word order differs across languages
- Word order degree of freedom also differs across languages
- We characterize word orders with: `Subject (S), Verb (V), Object (O) order`

Terminologies

Tokenization

Splits longer strings of text into **smaller pieces, or tokens**

Larger chunks of text can be tokenized into sentences

Sentences can be tokenized into words, etc.

Further processing is generally performed after a piece of text has been appropriately tokenized

Tokenization

Splits longer strings of text into **smaller pieces, or tokens**

Larger chunks of text can be tokenized into sentences

Sentences can be tokenized into words, etc.

Further processing is generally performed after a piece of text has been appropriately tokenized

Tokenization

Splits longer strings of text into **smaller pieces, or tokens**

Larger chunks of text can be tokenized into sentences

Sentences can be tokenized into words, etc.

Further processing is generally performed after a piece of text has been appropriately tokenized

Tokenization

Splits longer strings of text into **smaller pieces, or tokens**

Larger chunks of text can be tokenized into sentences

Sentences can be tokenized into words, etc.

Further processing is generally performed after a piece of text has been appropriately tokenized

Tokenization

Splits longer strings of text into **smaller pieces, or tokens**

Larger chunks of text can be tokenized into sentences

Sentences can be tokenized into words, etc.

Further processing is generally performed after a piece of text has been appropriately tokenized

Normalization (Pre-processing)

Normalization generally refers to a series of related tasks

- converting all text to the same case (upper or lower)
- removing punctuation
- expanding contractions
- converting numbers to their word equivalents, and so on.

Normalization puts all words on equal footing, and allows processing to proceed uniformly.

Stemming

Stemming is the process of **eliminating affixes**
- suffixed, prefixes, infixes, circumfixes

Stemmer

Tool to obtain a word stem.
- Running – run
- Unmanage - manage

Stemming

Stemming is the process of **eliminating affixes**
- suffixed, prefixes, infixes, circumfixes

Stemmer

Tool to obtain a word stem.
- Running – run
- Unmanage - manage

Lemmatization

Lemmatization is related to stemming, differing in that lemmatization is able to **capture canonical forms** based on a **word's lemma**.

For example, stemming the word "better" would fail to return its citation form (another word for lemma); however, lemmatization would result in the following:

`better → good`

Implementation of a stemmer would be the less difficult

Lemmatization

Lemmatization is related to stemming, differing in that lemmatization is able to **capture canonical forms** based on a **word's lemma**.

For example, stemming the word ”better” would fail to return its citation form (another word for lemma); however, lemmatization would result in the following:

```
better → good
```

Implementation of a stemmer would be the less difficult

Lemmatization

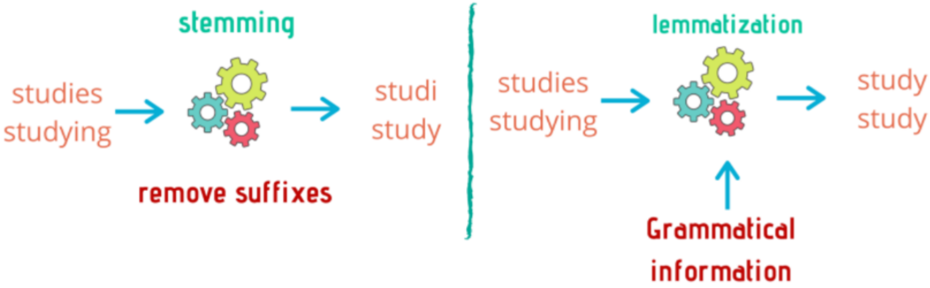
Lemmatization is related to stemming, differing in that lemmatization is able to **capture canonical forms** based on a **word's lemma**.

For example, stemming the word "better" would fail to return its citation form (another word for lemma); however, lemmatization would result in the following:

`better → good`

Implementation of a stemmer would be the less difficult

Stemming vs. Lemmatization



Corpus / Corpora

In Latin corpus means **body** refers to a **collection of texts**

Sources: Single, Multiple, Multilingual

Corpora are generally solely used for statistical linguistic analysis and hypothesis testing.

Corpus / Corpora

In Latin corpus means **body** refers to a **collection of texts**

Sources: Single, Multiple, Multilingual

Corpora are generally solely used for statistical linguistic analysis and hypothesis testing.

Corpus / Corpora

In Latin corpus means **body** refers to a **collection of texts**

Sources: Single, Multiple, Multilingual

Corpora are generally solely used for statistical linguistic analysis and hypothesis testing.

Stop Words

Stop words are those words which are **filtered out before further processing** of text, since these words **contribute little to overall meaning**, given that they are generally the most common words in a language.

A, an, the, in , on etc.

The quick brown fox jumps over the lazy dog.

Stop Words

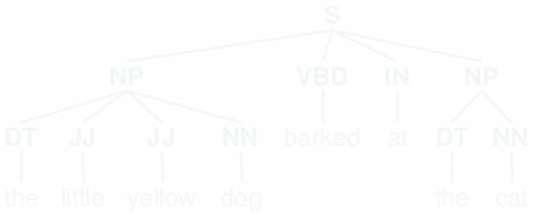
Stop words are those words which are **filtered out before further processing** of text, since these words **contribute little to overall meaning**, given that they are generally the most common words in a language.

A, an, the, in , on etc.
The quick brown fox jumps over the lazy dog.

Parts-of-speech (POS) Tagging

POS tagging consists of assigning a category tag to the tokenized parts of a sentence.

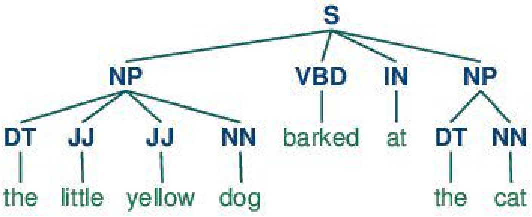
The most popular POS tagging would be identifying words as nouns, verbs, adjectives, etc.



Parts-of-speech (POS) Tagging

POS tagging consists of assigning a category tag to the tokenized parts of a sentence.

The most popular POS tagging would be identifying words as nouns, verbs, adjectives, etc.



Bag of Words

Bag of words is a particular representation model used to **simplify the contents of a selection of text**.

iThe bag of words model omits grammar and word order, but is interested in the number of occurrences of words within the text.

```
"Well, well, well," said John. "There, there," said James. "There, there."
```

```
{ 'well': 3, 'said': 2, 'john': 1, 'there': 4, 'james': 1 }
```

Bag of Words

Bag of words is a particular representation model used to **simplify the contents of a selection of text**.

iThe bag of words model omits grammar and word order, but is interested in the number of occurrences of words within the text.

```
"Well, well, well," said John. "There, there," said James. "There, there."
```

```
{ 'well': 3, 'said': 2, 'john': 1, 'there': 4, 'james': 1 }
```

Bag of Words

Bag of words is a particular representation model used to **simplify the contents of a selection of text**.

iThe bag of words model omits grammar and word order, but is interested in the number of occurrences of words within the text.

```
"Well, well, well," said John. "There, there," said James. "There, there."
```

```
{ 'well': 3, 'said': 2, 'john': 1, 'there': 4, 'james': 1 }
```

Bag of Words

Bag of words is a particular representation model used to **simplify the contents of a selection of text**.

iThe bag of words model omits grammar and word order, but is interested in the number of occurrences of words within the text.

```
"Well, well, well," said John. "There, there," said James. "There, there."
```

```
{ 'well': 3, 'said': 2, 'john': 1, 'there': 4, 'james': 1 }
```

n-grams

n-grams is another representation model for simplifying text selection contents

As opposed to the orderless representation of bag of words, n-grams modeling is **interested in preserving contiguous sequences of N items** from the text selection

`("There, there," said James. "There, there.")`

Appears as a list representation below with 3-gram model:

```
 [
"there there said",
"there said james",
"said james there",
"james there there",
]
```

n-grams

n-grams is another representation model for simplifying text selection contents

As opposed to the orderless representation of bag of words, n-grams modeling is **interested in preserving contiguous sequences of N items** from the text selection

```
("There, there," said James. "There, there.")
```

Appears as a list representation below with 3-gram model:

```
 [
"there there said",
"there said james",
"said james there",
"james there there",
]
```

n-grams

n-grams is another representation model for simplifying text selection contents

As opposed to the orderless representation of bag of words, n-grams modeling is **interested in preserving contiguous sequences of N items** from the text selection

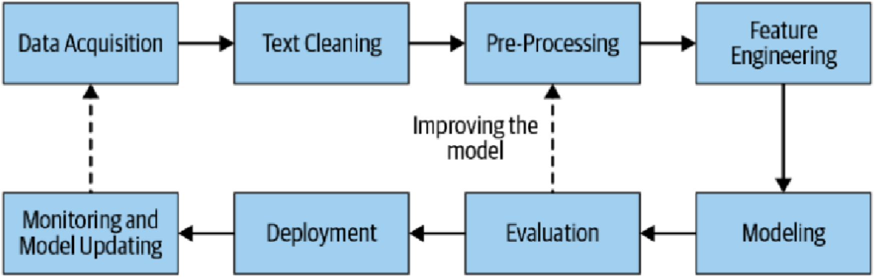
`("There, there," said James. "There, there.")`

Appears as a list representation below with 3-gram model:

```
 [
"there there said",
"there said james",
"said james there",
"james there there",
]
```

NLP Pipeline

Standard NLP Pipeline



Performing NLP research

Assume we have a Research, Engineering, Product Problem

- Define a **NLP System** to solve it. Split into modules, each one performing a task
- Define **Evaluation Metric(s)** for your system and sub-modules
- **Collect Data** to build/train your models
- Build **Baseline Models** (i.e. most simple model you can think of that have a non trivial performance metric)
- Build **Better Models** using **symbolic/statistical/DL** methods

NLP problems

Sentiment Analysis

Definition: Identify the emotional tone (e.g., positive, negative, neutral) in a piece of text.

Example:

- Input: ”The movie was absolutely fantastic!”
- Output: `Positive`

$$f_{\text{sentiment}} : \mathcal{X} \to \mathcal{Y}, \quad \mathcal{Y} = \{\text{positive}, \text{negative}, \text{neutral}\}$$

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y \mid x; \theta)$$

Sentiment Analysis

Definition: Identify the emotional tone (e.g., positive, negative, neutral) in a piece of text.

Example:

- Input: ”The movie was absolutely fantastic!”
- Output: `Positive`

$$f_{\text{sentiment}} : \mathcal{X} \to \mathcal{Y}, \quad \mathcal{Y} = \{\text{positive}, \text{negative}, \text{neutral}\}$$

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(y \mid x; \theta)$$

Document Summarization

Definition: Automatically generate a short and coherent summary from a long document.

Example:

- Input: A 300-word news article on COVID-19 spread ...(long text body) ...
- Output: "COVID-19 cases continue to rise in South Asia."

$$f_{\text{summarize}} : x = (x_1, x_2, \ldots, x_n) \rightarrow y = (y_1, y_2, \ldots, y_m), \quad m \ll n$$

$$\hat{y} = \arg\max_{y} P(y \mid x; \theta)$$

Document Summarization

Definition: Automatically generate a short and coherent summary from a long document.

Example:

- Input: A 300-word news article on COVID-19 spread ...(long text body) ...
- Output: "COVID-19 cases continue to rise in South Asia."

$$f_{\text{summarize}} : x = (x_1, x_2, \ldots, x_n) \rightarrow y = (y_1, y_2, \ldots, y_m), \quad m \ll n$$

$$\hat{y} = \arg\max_{y} P(y \mid x; \theta)$$

Document Classification

Definition: Assign one or more predefined labels or topics to a given document.

Example:

- Input: ”The government passed the new tax reform bill.”
- Output: `Politics`, `Economy`

$$f_{\text{classify}} : \mathcal{X} \to \mathcal{C}, \quad \text{where } \mathcal{C} = \{c_1, c_2, \ldots, c_k\}$$

$$\hat{y} = \arg\max_{c \in \mathcal{C}} P(c \mid x; \theta)$$

Document Classification

Definition: Assign one or more predefined labels or topics to a given document.

Example:

- Input: "The government passed the new tax reform bill."
- Output: `Politics`, `Economy`

$$f_{\text{classify}} : \mathcal{X} \to \mathcal{C}, \quad \text{where } \mathcal{C} = \{c_1, c_2, \ldots, c_k\}$$

$$\hat{y} = \arg\max_{c \in \mathcal{C}} P(c \mid x; \theta)$$

Machine Translation

Definition: Translate text from a source language to a target language automatically.

Example:

- Input: ”I love you.”
- Output: "म तिमीलाई माया गर्छु।"

$$f_{\text{translate}} : x = (x_1, \ldots, x_n) \to y = (y_1, \ldots, y_m)$$

$$\hat{y} = \arg\max_{y} \prod_{t=1}^{m} P(y_t \mid y_{<t}, x; \theta)$$

Machine Translation

Definition: Translate text from a source language to a target language automatically.

Example:

- Input: ”I love you.”
- Output: "म तिमीलाई माया गर्छु।"

$$f_{\text{translate}} : x = (x_1, \ldots, x_n) \rightarrow y = (y_1, \ldots, y_m)$$

$$\hat{y} = \arg\max_{y} \prod_{t=1}^{m} P(y_t \mid y_{<t}, x; \theta)$$

Automatic Speech Recognition (ASR)

Definition:

Convert spoken audio signals into written text. **Example:**

- Input: (audio) WAV file with ”Hello, how are you?”
- Output: ”Hello, how are you?”

$$f_{\text{ASR}} : a = (\mathbf{a}_1, \ldots, \mathbf{a}_T) \to y = (y_1, \ldots, y_m)$$

$$\hat{y} = \arg\max_{y} P(y \mid a; \theta)$$

Automatic Speech Recognition (ASR)

Definition:

Convert spoken audio signals into written text. **Example:**

- Input: (audio) WAV file with "Hello, how are you?"
- Output: "Hello, how are you?"

$$f_{\text{ASR}} : a = (\mathbf{a}_1, \dots, \mathbf{a}_T) \to y = (y_1, \dots, y_m)$$

$$\hat{y} = \arg\max_{y} P(y \mid a; \theta)$$

Questions?

Questions?

Discussion