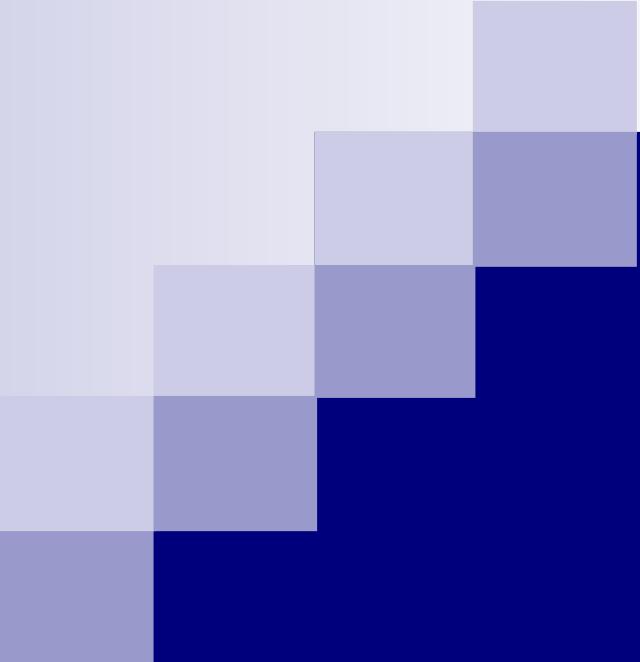


COMP5423 Natural Language Processing



Introduction to Natural Language Process (Part 1)

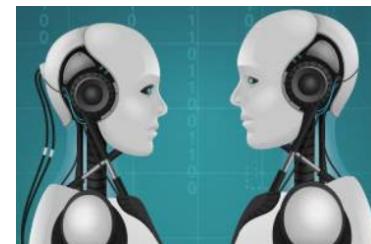
Introduction to NLP

Communication

Human-to-
Human

Human-to-
Machine

Machine-to-
Machine



Language is a medium of communication that helps us expressing and conveying our thoughts, feelings, and emotions.

Introduction to NLP

■ Natural Language Processing (NLP)

Natural
Language
Human Language



Retrieve the birth date and address of the employee(s) whose name is 'John B. Smith'.

```
SELECT      Bdate, Address
FROM        EMPLOYEE
WHERE       Fname='John' AND Minit='B' AND Lname='Smith';
```

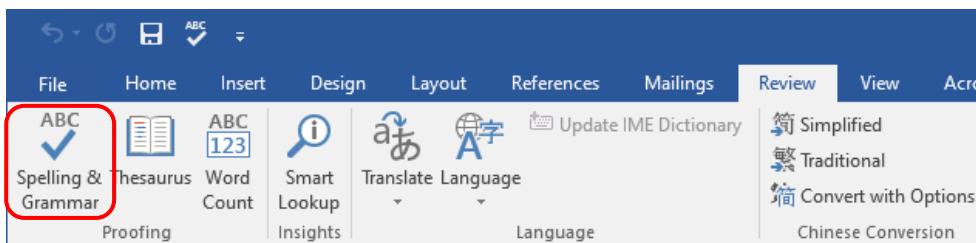


Processing
Application of Computational Techniques
to the Analysis of Natural Language

Introduction to NLP

■ Applications of NLP

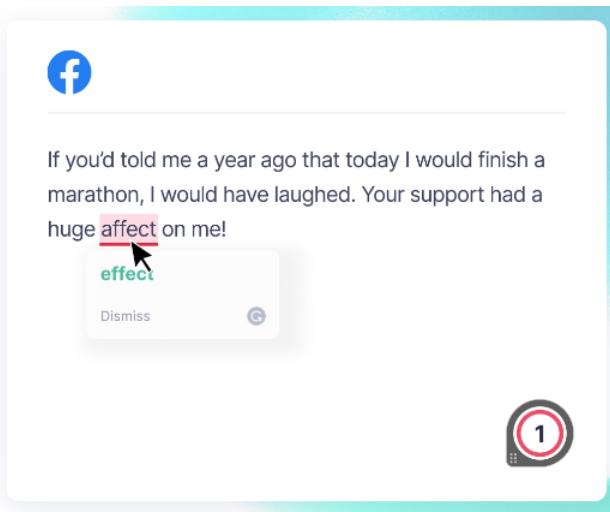
□ Spelling and Grammar Check



Great Writing, Simplified

Compose bold, clear, mistake-free writing with
Grammarly's new AI-powered desktop
Windows app.

[Get Grammarly](#) It's free



Overall score
Text is too short >

Goals
Adjust goals >

All suggestions
Correctness
Looking good

Clarity
Very clear

Engagement
Very engaging

Delivery
Just right

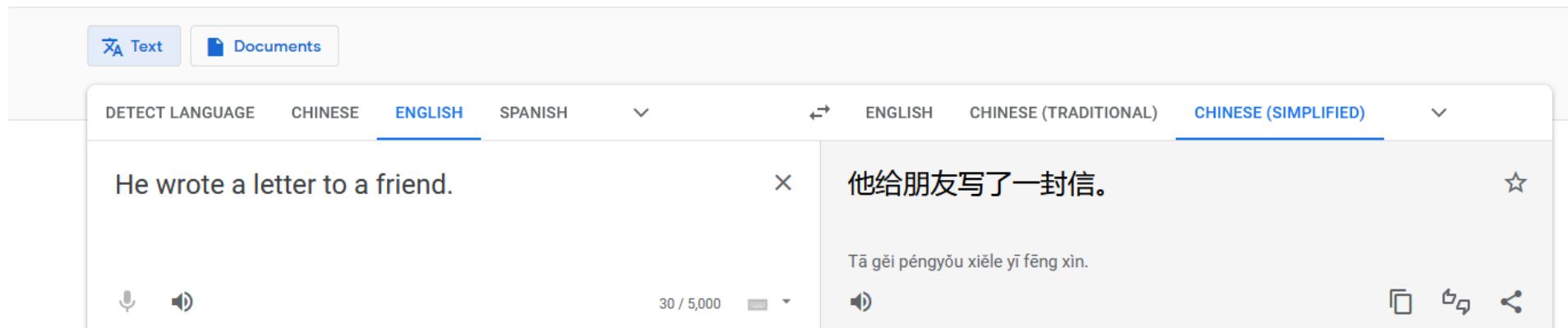
Style guide
All good

Introduction to NLP

■ Applications of NLP

□ Machine Translation (MT)

≡ Google Translate



The screenshot shows the Google Translate website interface. At the top, there are tabs for 'Text' and 'Documents'. Below that, a language selection bar shows 'DETECT LANGUAGE' followed by 'CHINESE', 'ENGLISH' (which is underlined in blue), and 'SPANISH'. To the right of the bar is a double-headed arrow icon, followed by 'ENGLISH', 'CHINESE (TRADITIONAL)', and 'CHINESE (SIMPLIFIED)' (also underlined in blue). A dropdown menu is open over the English input field, displaying the text 'He wrote a letter to a friend.' and its Chinese translation '他给朋友写了一封信.' along with pinyin 'Tā gěi péngyǒu xiěle yī fēng xìn.'. Below the input field are icons for microphone, speaker, and document. The bottom right corner of the interface has a 'Send feedback' link.



Google Translate was launched in April 2006 as a statistical machine translation service. As of December 2024, it supports 249 languages.

Introduction to NLP

■ Applications of NLP

Question Answering (QA)

Passage 1

Bacteria are extremely small living things. While we measure our own sizes in inches or centimeters, bacterial size is measured in microns. One micron is a thousandth of a millimeter. A pinhead is about a millimeter across. Rod shaped bacteria are usually from two to four microns long, while rounded ones are generally one micron in diameter. Thus if you enlarged a founded bacterium a thousand times, it would be just about the size of a pinhead. An adult human magnified by the same amount would be over a mile(1.6 kilometers) tall.

Even with an ordinary microscope, you must look closely to see bacteria. Using a magnification of 100 times, one finds that bacteria are barely visible as tiny rods or dots. One cannot make out anything of their structure. Using special stains, one can see that some bacteria have attached to them wavy-looking "hairs" called flagella. Others have only one flagellum. The flagella rotate, pushing the bacteria through the water. Many bacteria lack flagella and cannot move about by their own power while others can glide along over surfaces by some little understood mechanism.

From the bacterial point of view, the world is a very different place from what it is to humans. To a bacterium water is as thick as molasses is to us. Bacteria are so small that they are influenced by the movements of the chemical molecules around them. Bacteria under the microscope, even those with no flagella, often bounce about in the water. This is because they collide with the water molecules and are pushed this way and that. Molecules move so rapidly that within a tenth of a second the molecules around a bacterium have all been replaced by new ones. Even bacteria without flagella are thus constantly exposed to a changing environment.

Multiple Choice QA

1. Which of the following is the main topic of the passage?
(A) The characteristics of bacteria
(C) The various functions of bacteria
(B) How bacteria reproduce
(A) How bacteria contribute to disease
 2. Bacteria are measured in
(A) inches
(B) centimeters
(C) microns
(D) millimeters
 3. Which of the following is the smallest?
(A) A pinhead
(C) A microscope
(B) A rounded bacterium
(D) A rod-shaped bacterium
 4. According to the passage, someone who examines bacteria using only a microscope that magnifies 100 times would see
(A) tiny dots
(C) large rods
(B) small "hairs"
(D) detailed structures
 5. The relationship between a bacterium and its flagella is most nearly analogous to which of the following?
(A) A rider jumping on a horse's back
(C) A boat powered by a motor
(B) A ball being hit by a bat
(D) A door closed by a gust of wind
 6. In line 16, the author compares water to molasses, in order to introduce which of the following topics?
(A) The bacterial content of different liquids
(B) What happens when bacteria are added to molasses
(C) The molecular structures of different chemicals
(D) How difficult it is for bacteria to move through water

Introduction to NLP

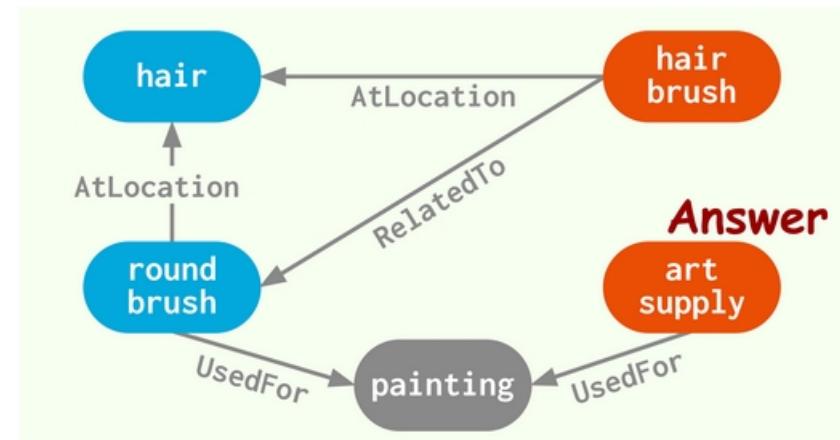
■ Applications of NLP

□ Question Answering (QA)



Document-based QA (DBQA)

If it is not used for hair, a **round brush** is an example of what?
A. hair brush B. bathroom C. art supplies* D. shower



Knowledge Graph

Knowledge-based QA (KBQA)

Introduction to NLP

■ Applications of NLP

□ Question Answering (QA)



Community QA

The image is a screenshot of a web browser displaying the Yahoo! Answers website at <https://answers.yahoo.com>. The page title is 'Cars & Transportation | ...'. The main content area shows the 'Cars & Transportation' category under 'All Categories > Cars & Transportation'. The category page features a grid of sub-categories: Aircraft, Boats & Boating, Buying & Selling, Commuting, Motorcycles, Safety, Car Audio, Car Makes, Insurance & Registration, Maintenance & Repairs, Other - Cars & Transportation, and Rail. Below the categories, there are two visible questions: 'Yesterday I was in my car and when I got out I forgot and left my car lights on for 10 minutes.' and 'Should I buy a salvaged vehicle?'. The interface includes a sidebar with 'Discover' and 'Answer' tabs, and a bottom navigation bar with a 125% zoom icon.



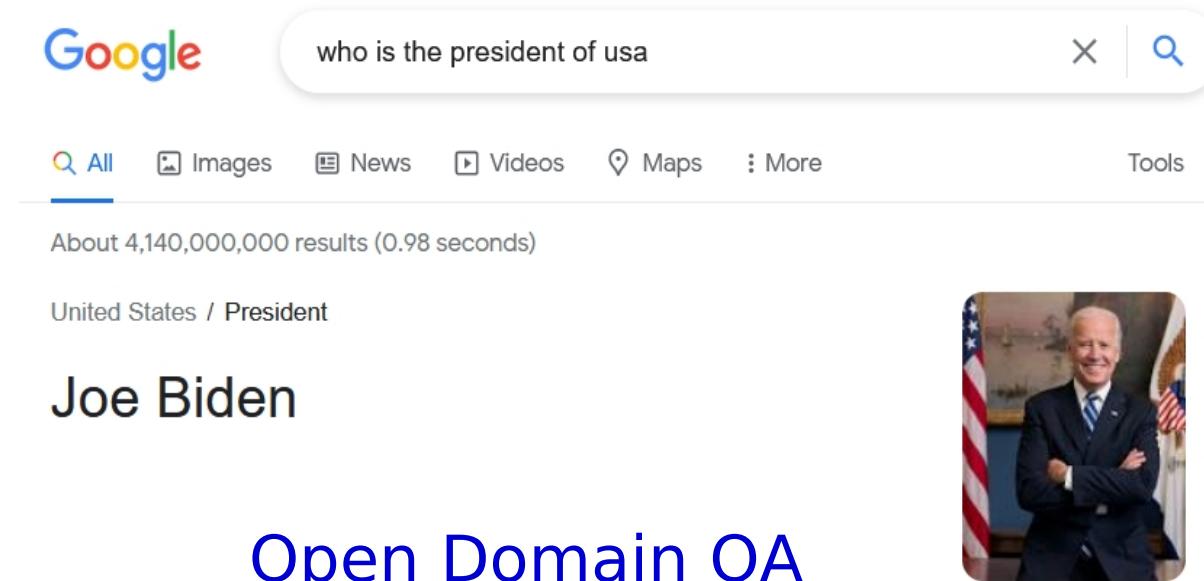
搜索答案

我要提问

Introduction to NLP

■ Applications of NLP

□ Question Answering (QA)



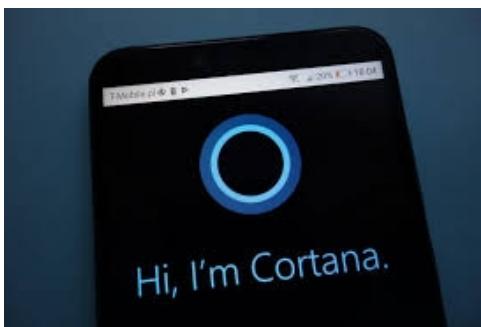
A screenshot of a Google search results page. The search bar at the top contains the query "who is the president of usa". Below the search bar, there are navigation links for "All", "Images", "News", "Videos", "Maps", and "More", with "All" being underlined. To the right of these links is a "Tools" button. A message indicates "About 4,140,000,000 results (0.98 seconds)". Below this, a section titled "United States / President" lists "Joe Biden" as the result. To the right of the text is a portrait photograph of Joe Biden in a dark suit, standing with his arms crossed. The background of the slide features a blue header bar and a decorative graphic of overlapping squares in the top left corner.

Open Domain QA

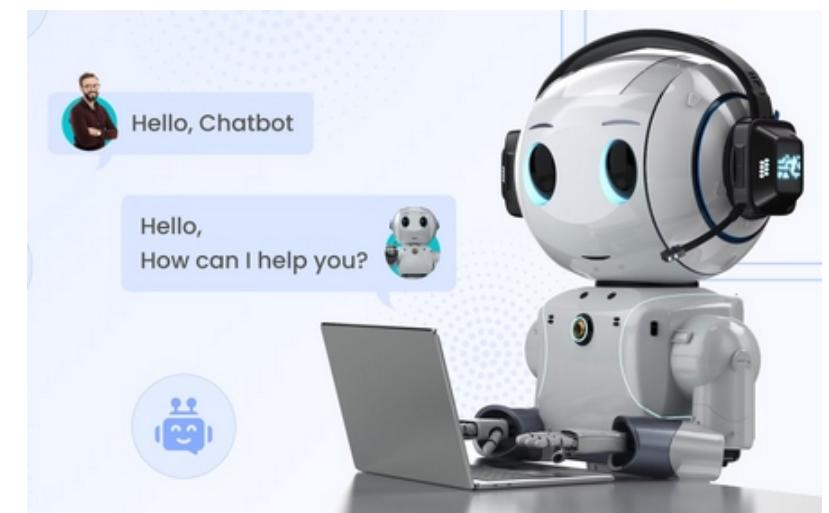
Introduction to NLP

■ Applications of NLP

□ Dialogue Systems (aka. Chatbots)



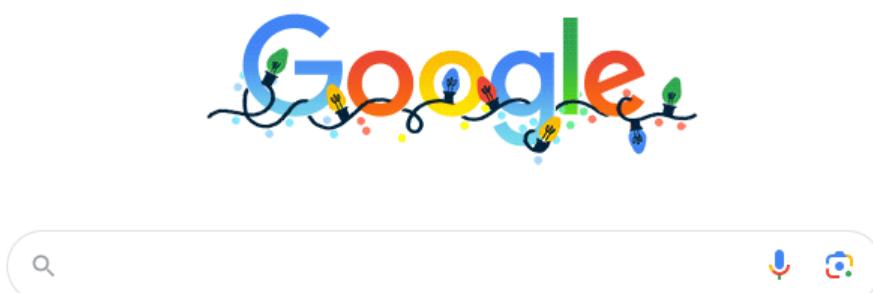
Task-Oriented vs. Non-Task-Oriented (Social-Oriented)



Introduction to NLP

■ Applications of NLP

□ Web Search





natural language processing



Snippet s

https://en.wikipedia.org/wiki/Natural_language_processing ::

Natural language processing - Wikipedia

Natural language processing (NLP) is an interdisciplinary subfield of linguistics, computer science, and artificial intelligence concerned with the ...

[Language processing](#) · [History of natural language](#) · [Language understanding](#)

You've visited this page many times. Last visit: 2/18/22

<https://www.ibm.com/cloud/cloud-learning> ::

What is Natural Language Processing? - IBM

2 Jul 2020 — Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI— ...

[What is natural language](#) · [NLP tasks](#) · [NLP tools and approaches](#)

You've visited this page 2 times. Last visit: 12/21/22

<https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing> ::

What is Natural Language Processing? An Introduction to NLP

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural ...

<https://www.sas.com/en-us/insights/analytics/nlp.html> ::

Natural Language Processing (NLP): What it is and why ... - SAS

Natural language processing helps computers communicate with humans in their own language and scales other language-related tasks. For example, NLP makes it ...
You visited this page on 12/21/22.

<https://www.oracle.com/database/technologies/nlp.html> ::

What is Natural Language Processing (NLP)? - Oracle

Natural language processing (NLP) is a branch of artificial intelligence (AI) that enables computers to comprehend, generate, and manipulate human language.

<https://www.nltk.org> ::

NLTK :: Natural Language Toolkit

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, ...

About 579,000,000 results (0.35 seconds)

Natural language processing (NLP) is a branch of artificial intelligence (AI) that enables computers to comprehend, generate, and manipulate human language. Natural language processing has the ability to interrogate the data with natural language text or voice.

 Oracle
<https://www.oracle.com> › ... › Artificial Intelligence

What is Natural Language Processing (NLP)? - Oracle

People also ask :

What is NLP and example?

Is NLP an AI?

What are the 5 steps in NLP?

Is NLP part of deep learning?



[About featured snippets](#) • [Feedback](#)

Natural language processing

Natural language processing is an interdisciplinary subfield of computer science and linguistics. It is primarily concerned with giving computers the ability to support and manipulate human language.

[Wikipedia](#)

 Wikipedia
https://en.wikipedia.org/wiki/Natural_language_processing

Natural language processing

Natural language processing (NLP) is an interdisciplinary subfield of computer science and linguistics. It is primarily concerned with giving computers the ...

[Language processing](#) · [Language generation](#) · [The history of natural...](#) · [NLP](#)

Scholarly articles for natural language processing

[Natural language processing](#) - Chowdhary - Cited by 2467

[Natural language processing: an introduction](#) - Nadkarni - Cited by 1553

[Natural language processing: a historical review](#) - Jones - Cited by 240

books Natural language processing

[View 5+ more](#)



People also search for

[View 10+ more](#)



Applications of Natural Language Processing



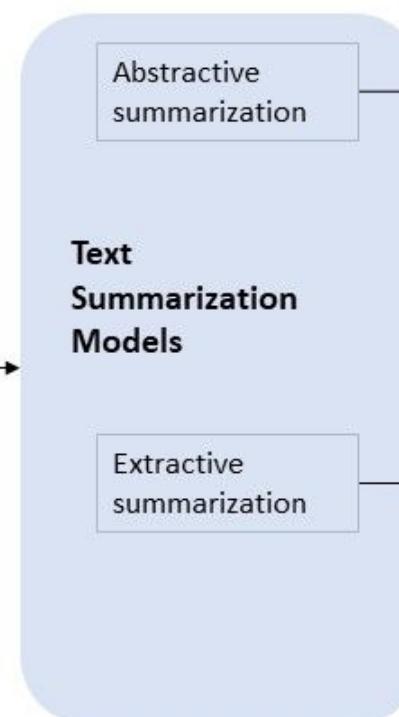
Introduction to NLP

■ Applications of NLP

□ Automatic Text Summarization

Input Article

Marseille, France (CNN) The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that " so far no videos were used in the crash investigation . " He added, " A person who has such a video needs to immediately give it to the investigators . " Robin\ 's comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps . All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. ...



Generated summary

Prosecutor : " So far no videos were used in the crash investigation "

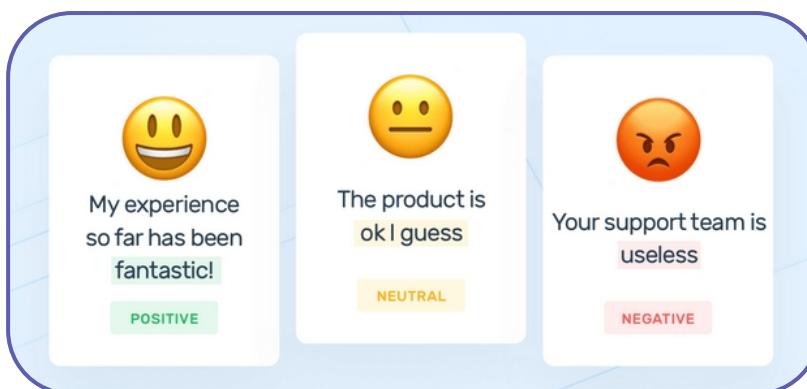
Extractive summary

marseille prosecutor brice robin told cnn that " so far no videos were used in the crash investigation . " robin \ 's comments follow claims by two magazines , german daily bild and french paris match , of a cell phone video showing the harrowing final seconds from on board germanwings flight 9525 as it crashed into the french alps . paris match and bild reported that the video was recovered from a phone at the wreckage site .

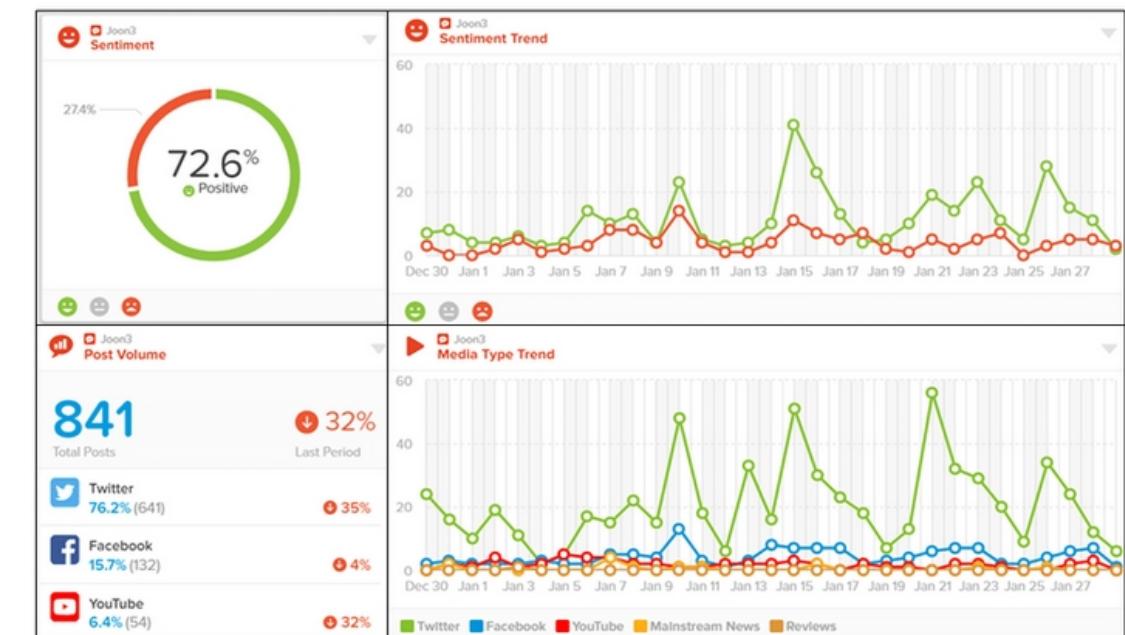
Introduction to NLP

■ Applications of NLP

□ Social Media Monitoring



Sentiment Classification

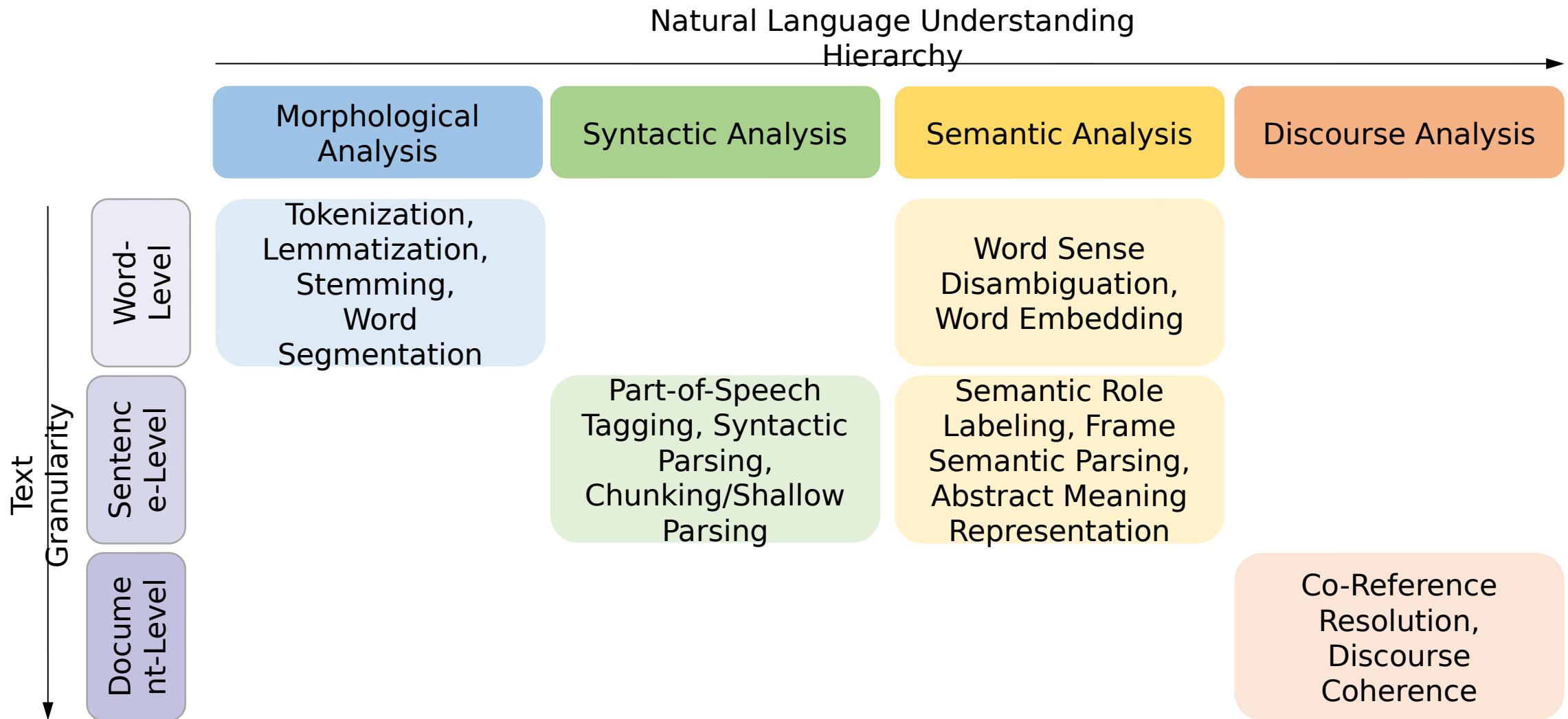


Introduction to NLP

■ Tasks in NLP

- Two Fundamental Tasks
 - Natural Language Understanding
 - Natural Language Generation

Introduction to NLP



Introduction to NLP

Tokenization/Segmentation

Natural Language Processing
↓
['Natural', 'Language', 'Processing']

下雨天留客天留我不留

下雨，天留客。天留，我不留！
下雨天，留客天。留我不，留！

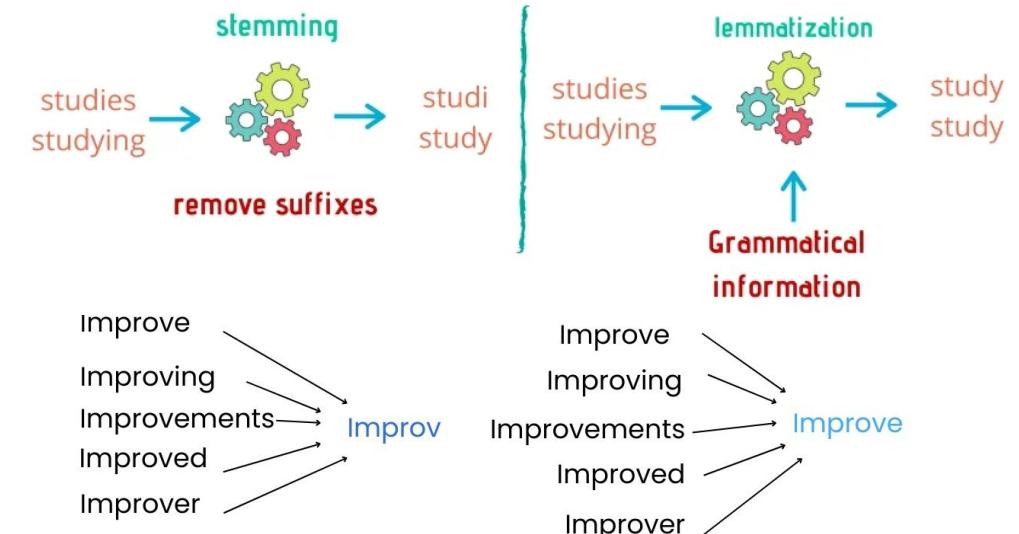
Sentence: "It is raining."

Sub-word level tokenization

It is rain ing .

Given a sequence of characters, word tokenization is the task of chopping it up into pieces, called **word tokens**. Word tokenization in Chinese is known as Chinese **Word Segmentation**, identifying the boundaries of

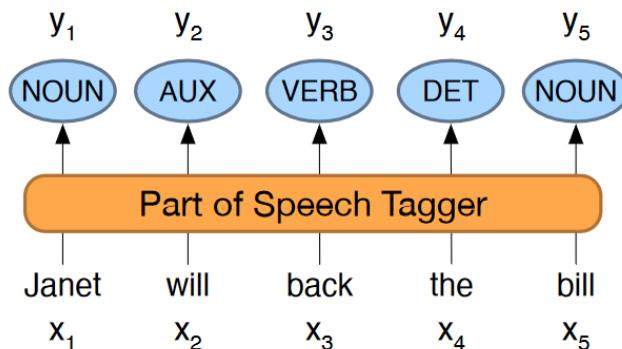
Lemmatization/Stemming



Lemmatization is to reduce inflectional forms (surface form) and sometimes derivationally related forms of a word to a common lemmatized (base, root) form or **lemma**. **Stemming** is a naive version of morphological analysis which simply strip

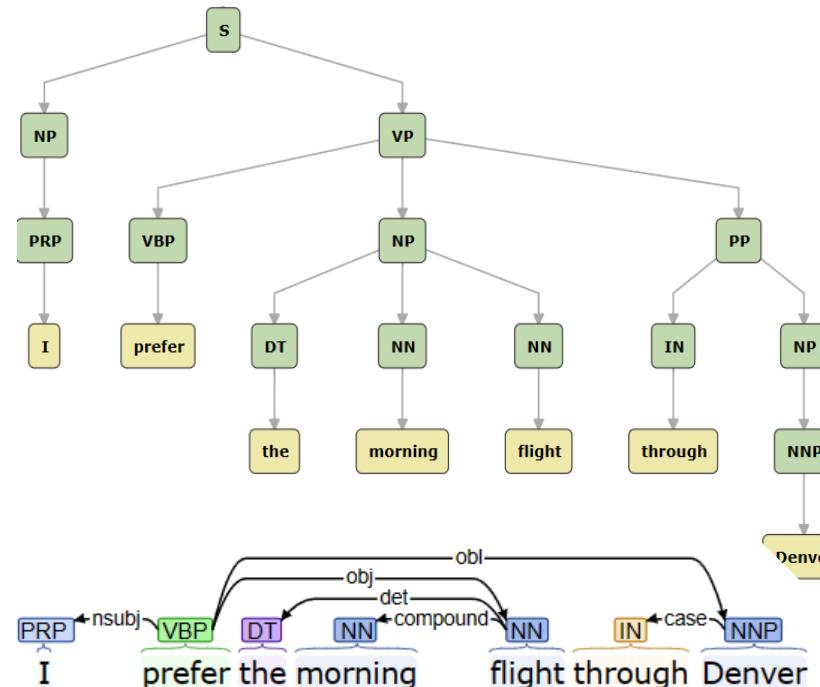
Introduction to NLP

Part-of-Speech Tagging



Part-of-Speech (for short POS) is the name for a group words, which have similar **grammatical functions**, such as noun, verb, pronoun, proposition, adverb, and conjunction, etc. **Part-of-Speech tagging** is a task of assigning a Part-of-Speech tag (like noun, verb, adjectives) to each word in a sentence.

Syntactic Parsing



Syntactic parsing to assign a **syntactic structure** to a sentence. Constituency parsing focuses on the **constituent (phrase) structure**, while dependency parsing describes the syntactic structure with of the **grammatical relations** that hold among the words.

Chunking/Shallow (Partial) Parsing

[_{NP} The morning flight] from [_{NP} Denver [_{VP} has arrived]].

Chunking is to identify the non-overlapping segments of a sentence, such as noun phrases, verb phrases, adjective phrases, and prepositional phrases.

Introduction to NLP

Word Sense Disambiguation



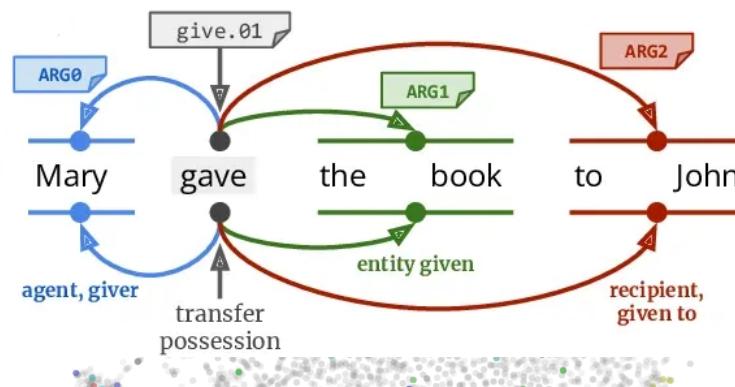
Most words have many different senses. **Word sense disambiguation** (WSD) is a task of determining which sense of a word is being used in a particular context.

WordNet Search - 3.1
- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for: Search WordNet

Display Options: Select option to change

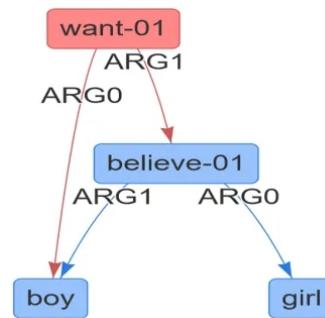
Semantic Role Labeling



Semantic role labeling is the task of assigning roles to spans in sentences, such as predicate, agent, patient, time, location, and instrument, etc.

Word Embeddings

Abstract Meaning Representation (AMR) Parsing



Abstract Meaning Representation (AMR) is a semantic representation language. AMR graphs are rooted, labeled, directed, acyclic graphs (DAGs), comprising whole sentences.

Introduction to NLP

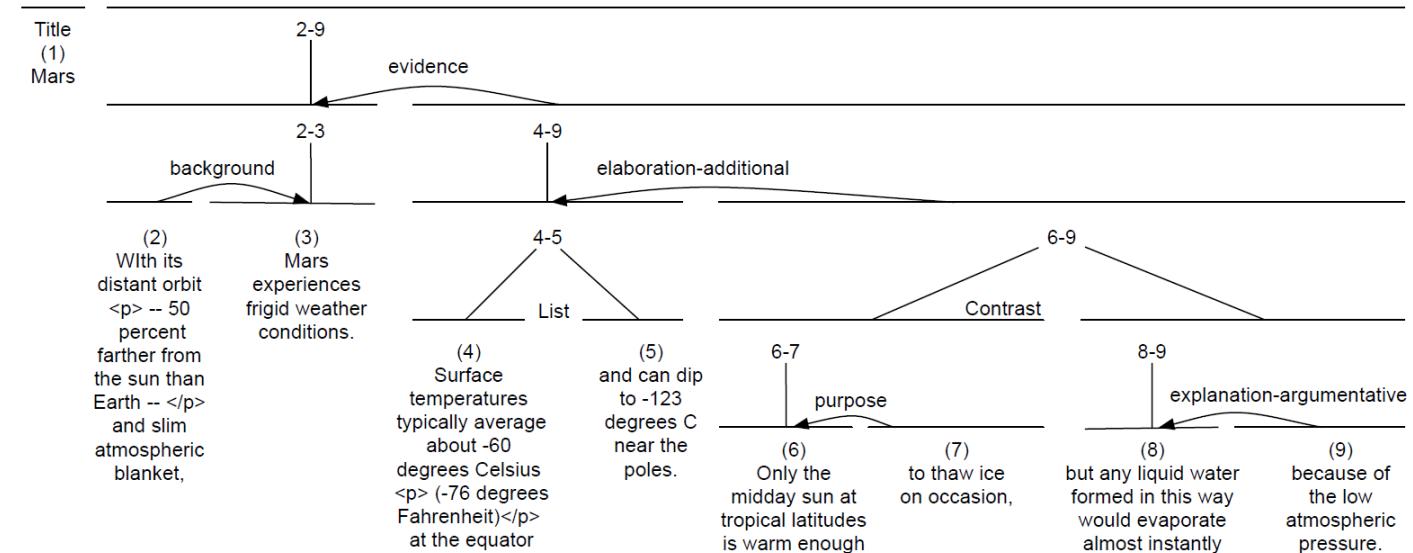
Co-Reference Resolution

0 Paul Allen was born on January 21, 1953, in 1 Seattle, Washington, to Kenneth Sam Allen and Edna Faye Allen. 0 Allen attended Lakeside School, a private school in 1 Seattle, where 0 he befriended 2 Bill Gates, two years younger, with whom 0 he shared an enthusiasm for computers.

3 0 Paul and 2 Bill used a teletype terminal at 3 their high school, Lakeside, to develop 3 their programming skills on several time-sharing computer systems.

In **information extraction (IE)**, a (named) entity is a real-world object, such as a person, location, organization, product, etc., that can be denoted with a proper name. **Co-reference resolution** is the task of determining whether two mentions (of entities) co-refer.

Discourse Coherence



Coherent discourses are structured by **coherence relations** that hold between text units, e.g., defined in **Rhetorical Structure Theory (RST)**. Given a sequence of sentences, the task of automatically determining the coherence relations between **Elementary Discourse Units (EDU)** (Nuclear or Satellite) is called **RST parsing**.²²

Introduction to NLP

Why do we care about coherence of a discourse?



John wanted to buy a piano for his living room.

Jenny also wanted to buy a piano.

He went to the piano store.

It was nearby.

The living room was on the second floor.

She didn't find anything she liked.

The piano he bought was hard to get up to that floor. Entity-based Coherence

Introduction to NLP

■ Tasks in NLP

- Natural Language Understanding
 - Morphological/Lexical Analysis
 - Morphological knowledge concerns how words are constructed from morphemes.
 - Key Concepts: Tokenization, Stemming/Lemmatization, Chinese Word Segmentation

Introduction to NLP

■ Tasks in NLP

- Natural Language Understanding
 - Syntactic Analysis
 - Syntactic knowledge concerns how words can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of what other phrases.
 - Key Concepts: Part-of-Speech (POS) Tagging, Syntactic Parsing (Constituency Parsing, Dependency Parsing), Chunking, Context-Free Grammar (aka. Phrase-Structure Grammar)

Introduction to NLP

■ Tasks in NLP

- Natural Language Understanding
 - Semantic Analysis
 - Semantic knowledge concerns what words mean and how these meanings combine in sentences to form sentence meanings.
 - Key Concepts: Word Sense Disambiguation, Word Embedding, Semantic Role Labeling, Frame Semantic Parsing, Abstract Meaning Representation, First Order Logic (aka. Predicate Calculus)

Introduction to NLP

■ Tasks in NLP

- Natural Language Understanding
 - Discourse Analysis
 - Discourse knowledge concerns how the immediately preceding sentences affect the interpretation of the next sentence.
 - Key Concepts: Co-Reference Resolution, Discourse Coherence

Introduction to NLP

■ Tasks in NLP



Natural Language Understanding
(Reading)

Natural Language Generation
(Writing)

Introduction to NLP

■ Tasks in NLP

□ Natural Language Generation

- Context-Free Grammar (aka. **Phrase-Structure Grammar**) is the most widely used formal system for modeling constituent structure in natural languages.

Lexicon: Words (such as fly, Sunday) and Syntactic Symbols (such as Noun (denoted as N), Noun Phrase (denoted as NP))

Rules (or Productions): Each expresses the ways that words and symbols of the language can be **grouped** and **ordered** together.

Introduction to NLP

A CFG can be thought of in two ways.

- As a device for assigning a structure to a given sentence
- As a device for generating sentences

NP ≡

ProperNoun

NP ≡ Det

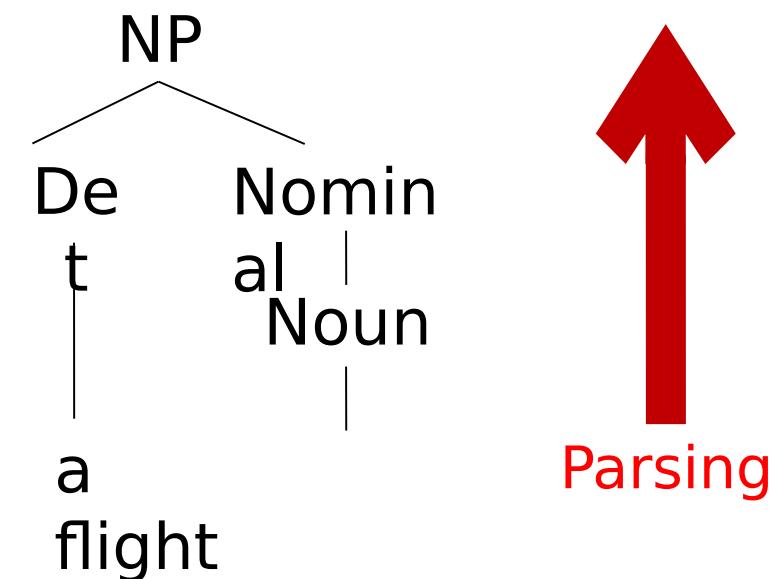
Nominal

Nominal ≡

Det ≡ the

Noun ≡ flight

≡ means to rewrite the symbol on the left with the sequence of symbols on the right.



Introduction to NLP

A CFG can be thought of in two ways.

- As a device for assigning a structure to a given sentence
- As a device for generating sentences

NP

ProperNoun

NP Det

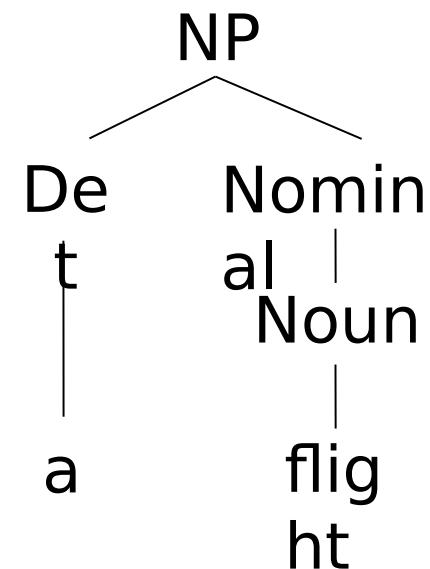
Nominal

Nominal Det

Noun flight

means to rewrite the symbol on the left with the sequence of symbols on the right.

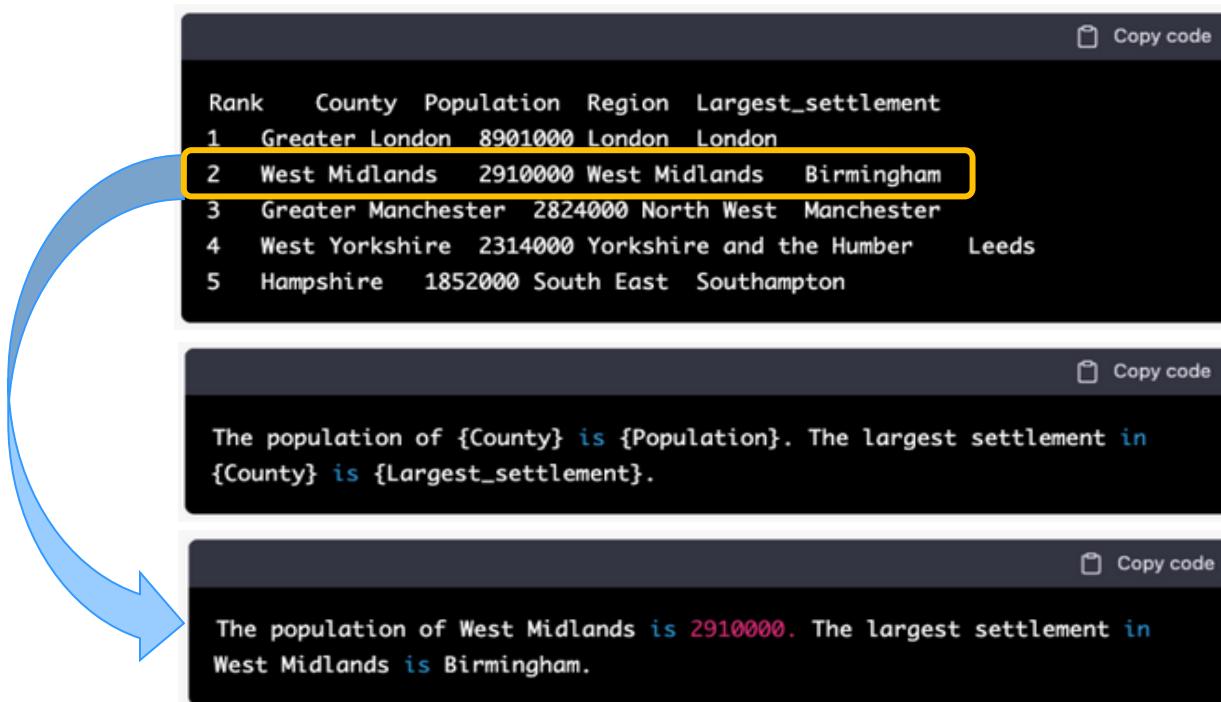
Generation



Introduction to NLP

■ Tasks in NLP

- Natural Language Generation
 - Template based Generation



Relational
Database

Pre-
defined
Template
Natural
Language
Sentence

Introduction to NLP

■ Tasks in NLP

□ Natural Language Generation

■ Pattern Recognition and Transformation Rules

```
Welcome to
      EEEEEE  LL      IIII    ZZZZZZ  AAAAAA
      EE      LL      II      ZZ      AA      AA
      EEEEEE  LL      II      ZZZ     AAAAAAAA
      EE      LL      II      ZZ      AA      AA
      EEEEEE  LLLLLL  IIII  ZZZZZZ  AA      AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.
```

You hate me

WHAT MAKES YOU THINK I HATE YOU

Introduction to NLP

■ Tasks in NLP

□ Natural Language Generation

- Statistical (N-Gram) or Neural Language Model based Generation (Word-by-Word Generation)

The diagram shows a sequence of text where each word is highlighted in yellow. The text is:

The man is
walking
The man is walking
down
The man is walking
down the
The man is walking down
the street

After the word "down", there is a yellow step-like graphic pointing to the words "the" and "street". To the right of this graphic is a yellow speech bubble containing the text "Predicted Next Words" in red.

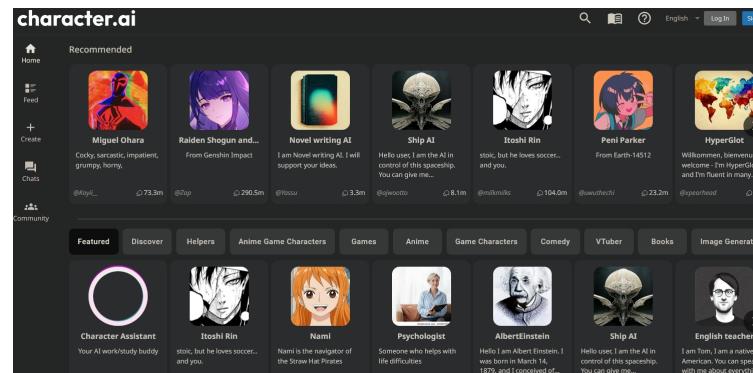
* In 1980, the first significant statistical language model was proposed.

Introduction to NLP

■ Tasks in NLP

□ Natural Language Generation

- Application Scenarios: Story Generation, Report Generation, Summary Generation, Question/Answer Generation, Response/Dialogue Generation, etc.
- Conditional/Controllable Generation



[+] The potato is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state, though...

[Negative] The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...

[Positive] The potato chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them...

■ Personalized and Stylized Generation

Introduction to NLP

■ Tasks in NLP

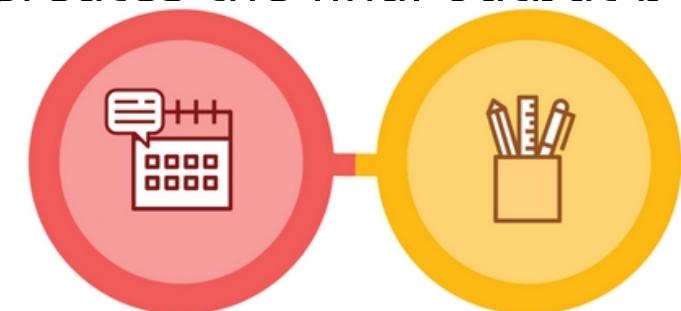
□ Natural Language Generation

- Short Text (Sentence) Generation vs. Long Text (Document) Generation vs. Interactive Text (Dialogue) Generation

The man is walking down
The man is walking down the
The man is walking down the
street
The man is walking down the
street .

Content Planning: Decides what information to include and how to structure sentences and paragraphs.

Text Generation: Creates the final output by composing human-like text.



Introduction to NLP

■ Tasks in NLP

- Textual Entailment (TE) and Natural Language Inference (NLI)

Premise	Label	Hypothesis
A man inspects the uniform of a figure in some East Asian country.	Contradiction	The man is sleeping.
An older and younger man smiling.	Neutral	Two men are smiling and laughing at the cats playing on the floor.
A soccer game with multiple males playing.	Entailment	Some men are playing a sport.

Introduction to NLP

■ Tasks in NLP

□ Information Extraction and Knowledge Graphs

(Named) Entity Recognition

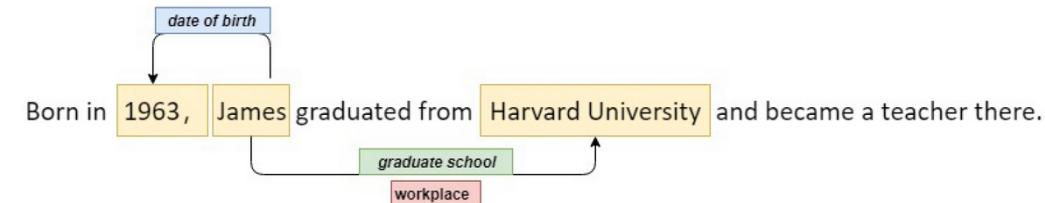
In December 1903 **DATE** the Royal Swedish Academy of Sciences **ORG** awarded Marie **PERSON** and Pierre Curie **PERSON**, along with Henri Becquerel **PERSON**, the Nobel Prize in Physics **WORK_OF_ART**.

Co-Reference Resolution

0 Paul Allen was born on January 21, 1953, in 1 Seattle, Washington, to Kenneth Sam Allen and Edna Faye Allen. 0 Allen attended Lakeside School, a private school in 1 Seattle, where 0 he befriended 2 Bill Gates, two years younger, with whom 0 he shared an enthusiasm for computers. 3 0 Paul and 2 Bill used a teletype terminal at 3 their high school, Lakeside, to develop 3 their programming skills on several time-sharing computer systems.

A (**named**) **entity** is a real-world object, such as a person, location, organization, product, etc., that can be denoted with **Two or more proper name expressions** (i.e., mentions) that are used to refer to the same entity are said to **co-refer**. **Co-reference resolution** is the task of determining whether two mentions

Relation Detection

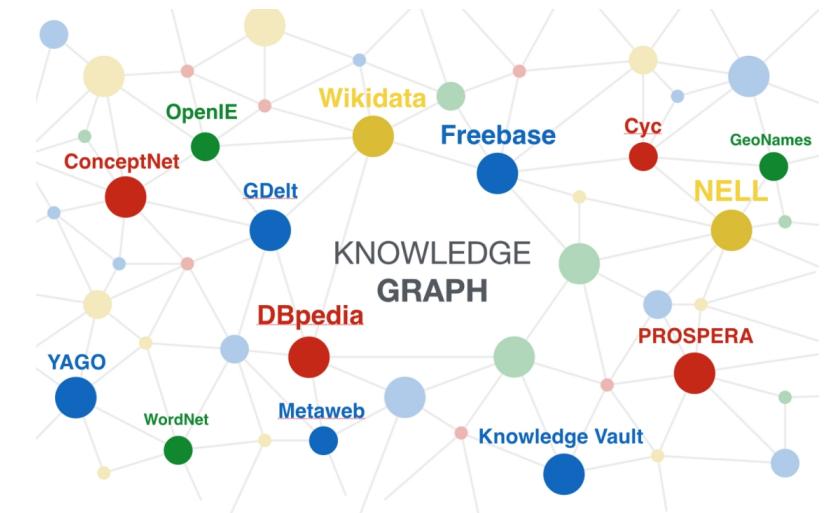
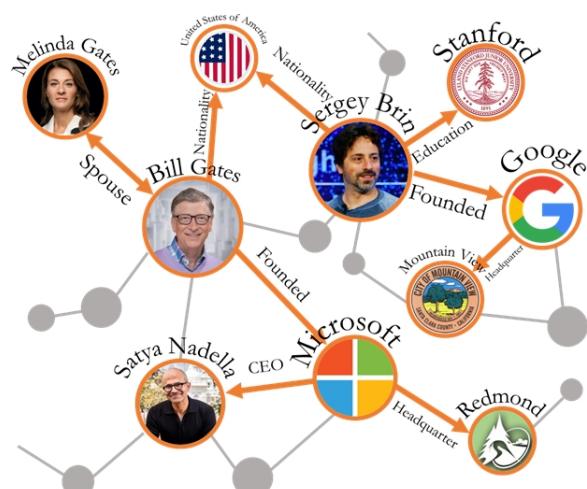


The task of **relation extraction** is finding and classifying semantic relations among entities mentioned in a text.

Introduction to NLP

■ Tasks in NLP

□ Information Extraction and Knowledge Graphs



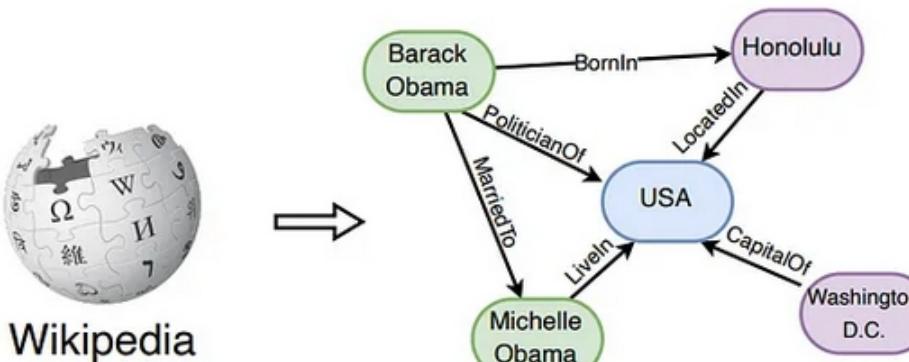
A **knowledge graph** represents a network of real-world entities (i.e., objects and concepts) and illustrates the relationship between them. This information is usually stored as the entity-relation-entity triples in a database and visualized as a graph structure, prompting the term knowledge “graph”.



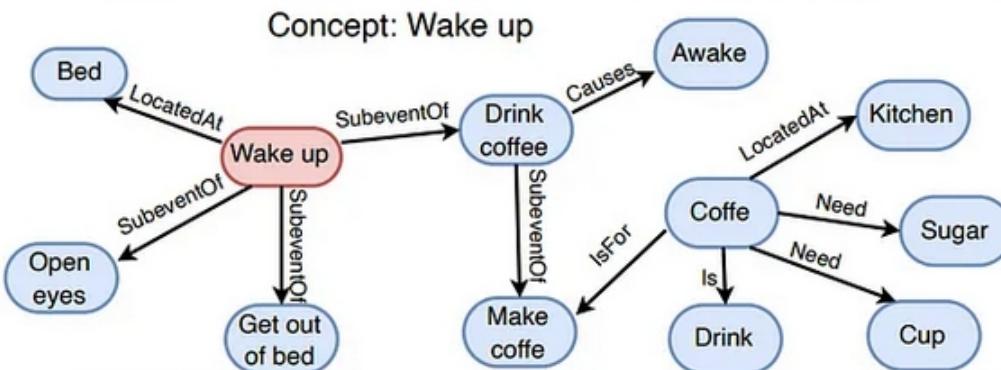
Thank you

Introduction to NLP

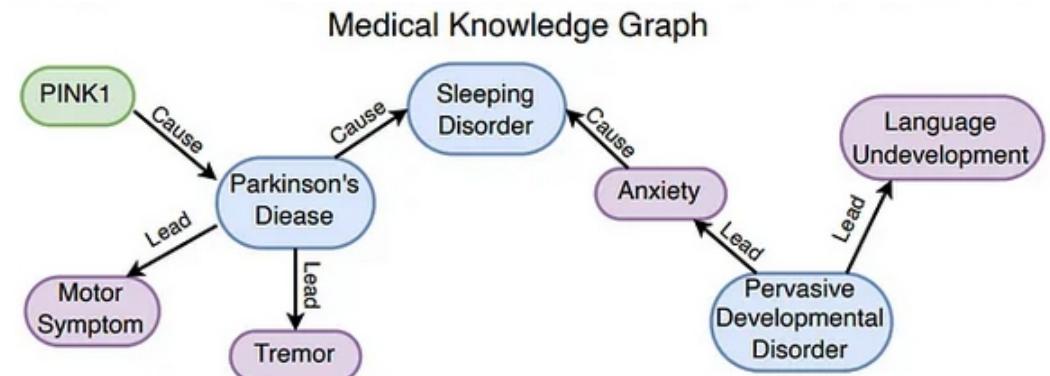
Encyclopedic Knowledge Graphs



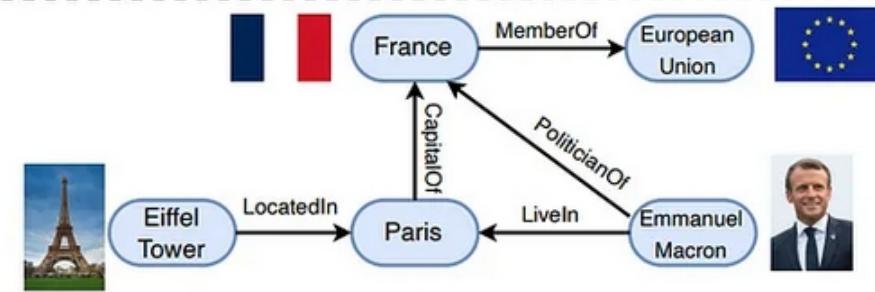
Commonsense Knowledge Graphs



Domain-specific Knowledge Graphs

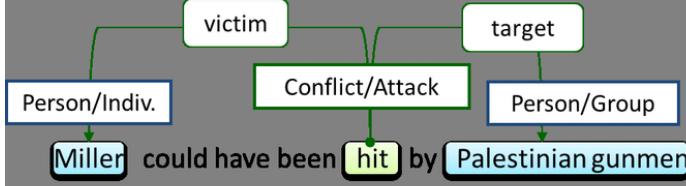


Multi-modal Knowledge Graphs



Introduction to NLP

Event Extraction



Event extraction is the task of finding events in which the entities participate.

Temporal Expression Extraction

Person p Loc l Org o Event e Date d Other z

Barack Hussein Obama II * (born August 4, 1961 *) is an American * attorney and politician who served as the 44th President of the United States * from January 20, 2009 *, to January 20, 2017 *. A member of the Democratic Party *, he was the first African American * to serve as president. He was previously a United States Senator * from Illinois * and a member of the Illinois State Senate *.

Temporal expression recognition is a task of finding the start and end of all of the text spans that correspond to such temporal expressions. These temporal expressions can be **normalized** onto specific calendar dates or times of day to situate events in time.

Temporal Relation Extraction

Some 1,500 ethnic Albanians marched Sunday in downtown Istanbul, burning Serbian flags to protest the killings of ethnic Albanians by Serb police in southern Serb Kosovo province. The police barred the crowd from reaching the Yugoslavian consulate in downtown Istanbul, but allowed them to demonstrate on nearby streets.

Diagram illustrating temporal relations between the events:

- Marched **during** Burning
- Burning **before** Reaching
- Reaching **after** Demonstrating

Temporal Relation Extraction is identifying temporal relations between pairs of events and temporal expressions.

2020 United States Presidential Election Timeline

