Natural Language Processing

A Machine Learning Perspective

A unified machine learning perspective on understanding and generating human language.

Core Focus

How algorithms can systematically understand, process, and generate human languages.

Methodological Shift

Moving from hand-crafted linguistic rules to unified mathematical and statistical modeling.

Based on Chapter 1 of 'Introduction to NLP from a Machine Learning Perspective'.

The Core Challenge of Language

Prevalent ambiguity and extreme flexibility make hand-coded rules fail.

Contextual Ambiguity

The same sentence can express different meanings depending on context.

"They can fish here."

Permission to fish in this place

The act of canning (preserving) fish

Syntactic & Lexical Flexibility

Language bends and adapts, making rigid rules insufficient.

Nouns easily become verbs

"She **pillowed** his head."

Slang bypasses standard spelling

"niiiiiiiiiice!"

Historical failure of rule-based translation in the 1960s

Example: "The spirit is strong, but the flesh is weak" translated to **"The vodka is delicious, but the meat tastes bad"**.

What is NLP?

An interdisciplinary field bridging linguistics, computer science, and artificial intelligence.

Linguistics & CS

Investigating computational methods to model languages and address formal linguistic questions.

Artificial Intelligence

Aiming to equip inteligent systems with human-like language understanding and generation capabilities.

Data Science

Processing large-scale unstructured text data to extract structured, actionable knowledge.

Intersections also extend to psychology, cognitive science, and neuroscience.

The Landscape of NLP Tasks

NLP tasks are broadly categorized into linguistics, extraction, and generation.

Computational Linguistics

Extracting syntax and semantics from words, sentences, and documents.

(e.g. parsing, POS tagging)

provides structure to

Information Extraction

Obtaining structured, tabular knowledge from unstructured text.

(e.g. entities, relations, events)

provides data to

Natural Language Generation

Automatically synthesizing human language from data, trees, or other text.

(e.g. translation, summarization)

These three sub-areas form the complete pipeline of modern language technologies.

Word-Level Processing

Morphology, tokenization, and segmentation form the foundation of text preprocessing.

Task	Input	Output
Morphological Analysis Decomposing words into morphemes.	**German:** “Wochenarbeitszeit”	**Wochen** + **arbeits** + **zeit**
Tokenization Separating words from punctuation while preserving integral markers.	**English:** “Mr. Smith is here.” “Wendy’s book arrived.”	**Mr.** | Smith | is | here | . **Wendy’s** | book | arrived | .
Word Segmentation Crucial for non-spaced languages like Chinese.	**Chinese:** “其中国外企业”	**其中** / **国外** / **企业**
POS Tagging Assigning part-of-speech labels to tokens.	**English:** The quick brown fox jumps.	The | quick | brown | fox | jumps | . DET ADJ ADJ NOUN VERB PUNCT

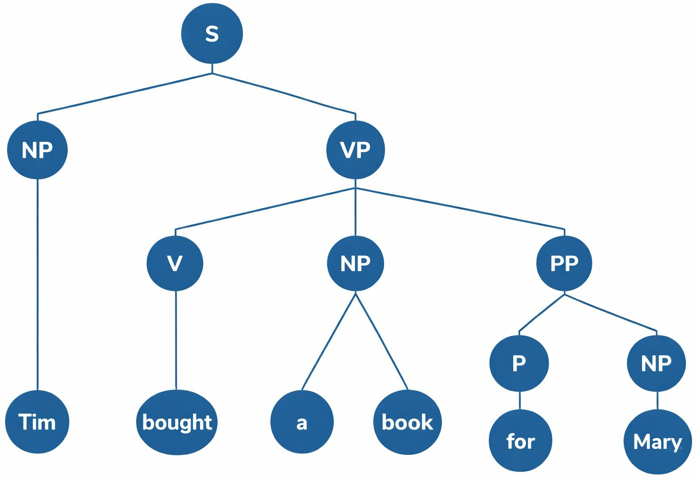
Table 1.1: Morphological analysis, tokenisation, word segmentation, and POS tagging examples.

Syntactic Structures

Constituent parsing models hierarchy, while dependency parsing models word-pair relations.

1 Cyntaituent Parsing

Analyzes sentences into hierarchical phrase structures (S → NP, VP) using phrase-structure grammars.



2 Dependency Parsing

Analyzes directed, labeled relations between head words and dependent words (e.g., nsubj, obj).

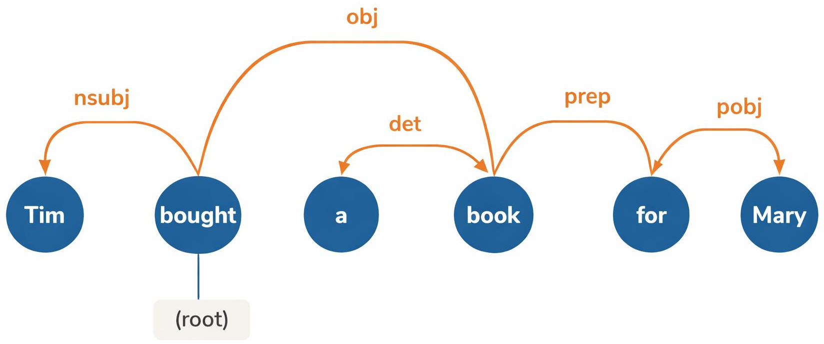


Figure 1.1: Constituent tree vs. Dependency tree for ‘Tim bought a book for Mary’.

Transition to Semantics

From Structure to Meaning

Moving from structural syntax to the extraction of meaning.

Syntax vs. Semantics

Syntax defines the form and structure; semantics defines the actual meaning of words and sentences.

The Semantic Spectrum

Ranges from word-level senses to sentence-level logic and document-level discourse.

Linguistic structures serve as the vehicle for conveying semantic meaning.

Lexical Semantics

Word meaning is defined by senses, context, and structured relations.

Word Sense Disambiguation (WSD)

Resolving polysemy by using context to determine the intended sense.

e.g., "book" as a reservation vs. a physical object.

Metaphor Detection

Identifying non-literal usage of words based on context.

e.g., "The machine ate my dollar."

Table 1.2: Four types of sense relations

(e.g. car-vehicle, leaf-tree).

Relation Type	Definition	Example
Synonymy	Words with similar meaning.	car – automobile
Antonymy	Words with opposite meaning.	hot – cold
Hypernymy	A general term (superordinate) refers to a more specific term (hyponym).	vehicle – car
Meronymy	A part (holonym) is part of a whole.	leaf – tree

Sentence-Level Semantics

Meaning can be represented via predicate-arguments, formal logic, or semantic graphs.

Predicate-Argument Structure

Identifies semantic roles (Agent, Patient) associated with a predicate.

Role	Relation	Argument
Agent	ARG0	Tim
Patient	ARG1	this book
Value	ARG2	$1

Predicate: bought

First-Order Logic

Translates sentences into symbolic formulas to facilitate formal mathematical inference.

$$\exists x\ (\ book(x)\ \wedge\ buy(Tim,\ x)\ \wedge\ price(x,\ \$1)\ )$$

First-order logic representation of the event.

Abstract Meaning Representation (AMR)

Encodes sentences as directed graphs where nodes are concepts and edges are semantic relations.

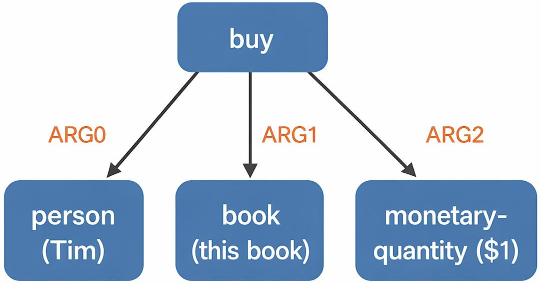


Figure 1.2: Semantic representations for ‘Tim bought this book for $1’.

Semantic Inference

Evaluating directional truth relations between texts.

Textual Entailment

Deciding if a premise text entails a hypothesis.

Premise	Tim went to the Riverside for dinner.
Hypothesis	Tim had dinner.

Entailment

Natural Language Inference (NLI)

Classifying the relationship between two sentences into one of three categories.

Entailment
The premise supports the hypothesis.

Contradiction
The premise contradicts the hypothesis.

Neutral
The premise is unrelated to the hypothesis.

NLI serves as a benchmark task for testing an AI’s true semantic understanding.

Discourse Analysis

Discourse parsing maps coherence relations between multi-sentence sub-topics.

Discourse Segmentation

Breaking text into sub-sentence units (elementary discourse units) separated by punctuation or markers.

Rhetorical Structure Theory (RST)

Hierarchically structures text using coherence relations such as “contrast”, “explain”, and “parallel”.

Nucleus vs. Satellite

Distinguishes central sub-topics (nucleus) from supporting information (satellite).

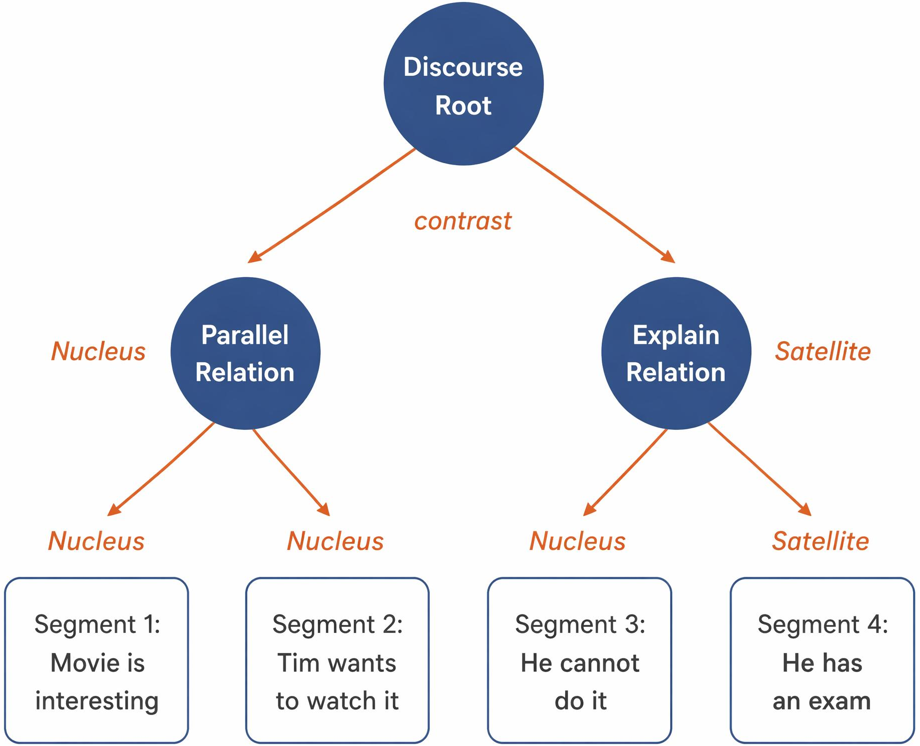


Figure 1.3: Discourse tree structure showing parallel, contrast, and explain relations.

Information Extraction

Unstructured Text to Structured Knowledge

Converting unstructured natural language into structured, queryable databases.

The Core Objective

Automatically extracting entities, relations, events, and sentiments from raw text.

Downstream Value

Enables structured search, knowledge graph construction, and automated decision-making.

IE bridges the gap between human-readable text and machine-readable databases.

Entity Extraction

Identifying named entities and resolving their coreferences across documents.

Named Entity Recognition (NER)

Extracting and classifying open-set entities (e.g., Person, Org, Location, GPE).

Anaphora Resolution

Identifying what a pronoun refers to (e.g., mapping “He” back to “Tim”).

Coreference Resolution

Grouping all expressions that refer to the same entity (e.g., {‘Tim’, ‘he’}, {‘the series’, ‘movies’}).

Table 1.3 & Table 1.4: Named entity and coreference resolution examples.

Table 1.3 Named Entity Recognition Examples

Text Span	Entity Tag	Entity Type	Description
Michael Jordan	PER	Person	Person name
Chicago Bulls	ORG	Organization	Sports team
NBA	ORG	Organization	League
United States	GPE	Geopolitical Entity	Country
Paris	GPE	Geopolitical Entity	City

Table 1.4 Coreference Resolution Examples

Document Excerpt	Coreference Sets
Tim met Mary yesterday. He said he would call her.	{Tim, He} {Mary, her}
The series was a hit. The movies received mixed reviews.	{the series} {The movies}

Relation Extraction

Extracting structured facts to build and complete Knowledge Graphs.

Relation Extraction

Identifying semantic links between entities (e.g. social, affiliation, physical location).

Knowledge Graphs (KG)

Storing entities as nodes and relations as directed edges in a graph database.

Link Prediction

Inferring missing relations mathematically (e.g. predicting “John is from Italy” via logical transitivity).

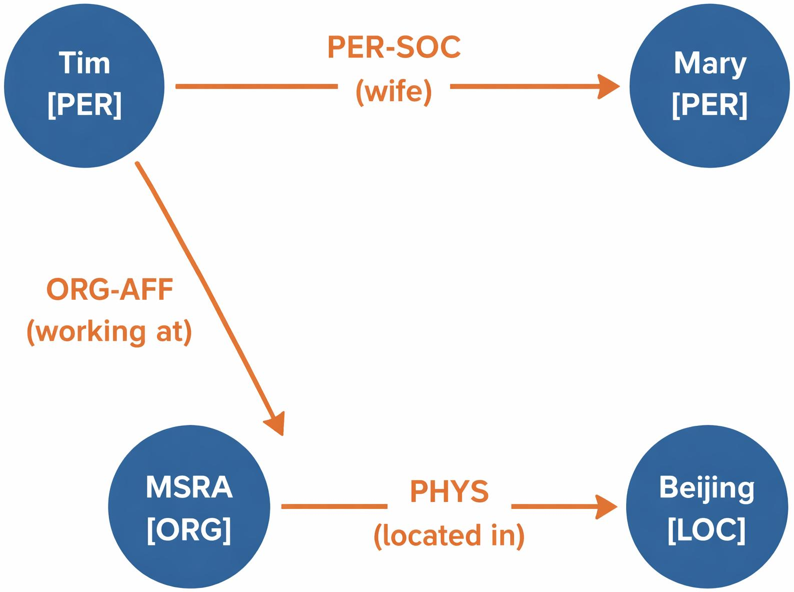


Figure 1.4(a): Relation extraction graph mapping social, affiliation, and physical relations.

Event Extraction

Identifying event triggers and extracting their semantic arguments.

Event Triggers

Detecting words that signify an event occurrence, which can be verbs (e.g., “visited”) or nouns (e.g., “visit”).

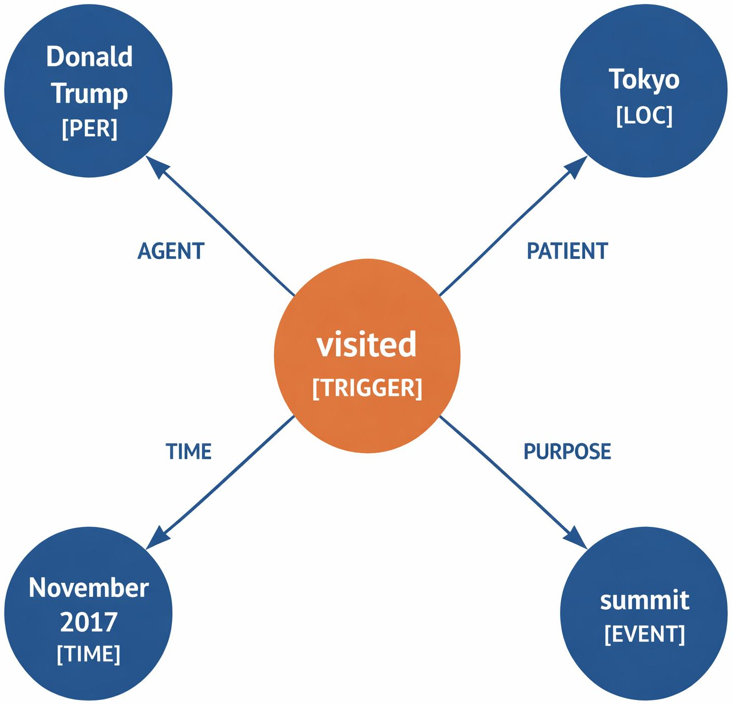
Argument Extraction

Mapping entities to specific event roles (e.g., Visitor, Destination, Time, Purpose).

Temporal & Causal Ordering

Determining the chronological sequence and cause-and-effect relations of events.

Figure 1.4(b): Event extraction mapping arguments (Agent, Patient, Time, Purpose) to the trigger “visited”.



Sentiment Analysis

Sentiment analysis ranges from document polarity to fine-grained aspect extraction.

Task	Input	Structured Output
Sentiment Classification	“I love this product. It works perfectly and the quality is great.”	Polarity: **Positive** *(Positive / Negative / Neutral)*
Targeted Sentiment	The battery life is amazing, but the screen is too dim.”	Target: battery life → **Positive** Target: screen → **Negative**
Aspect Sentiment	“Great keyboard, but the trackpad is not smooth.”	Keyboard: **Positive** Trackpad: **Negative**
Fine-Grained Extraction	I think the USB receiver is useful, but itis too small.”	Holder: **I** Target: **USB receiver** Expression: **useful** Polarity: **Positive**

Table 1.5: Sentiment analysis task variations and structured outputs.

Natural Language Generation

The Art of Synthesizing Language

Synthesizing coherent, fluent human language from structured or unstructured inputs.

The Generation Challenge

Mapping abstract representations, data tables, or source texts into natural, grammatically correct sentences.

Core Applications

- Machine translation
- Text summarization
- Question answering
- Interactive dialogue systems

Evolution: NLG has transitioned from template-based systems to end-to-end neural generation.

Fundamental NLG Tasks

Language modeling and structural transformation form the core of text generation.

Language Modeling (LM)

Predicting the next word (autoregressive) or filling in missing words (masked) based on context.

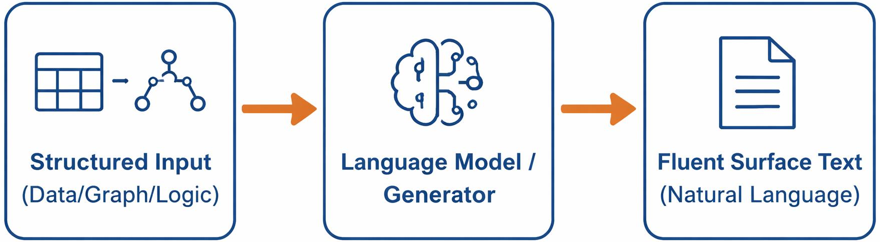
Surface Realization

Converting abstract semantic graphs or logical forms into fluent, natural surface text.

Data-to-Text

Transforming structured tables (e.g., sports statistics, financial data) into coherent textual reports.

> LMs serve as the foundational engine for almost all modern generative AI systems.

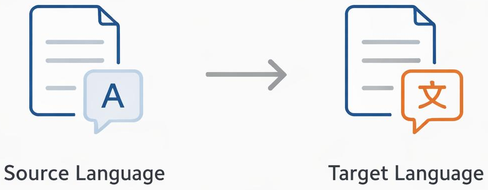


NLG Applications

Practical applications of NLG drive global communication and productivity.

Machine Translation (MT)

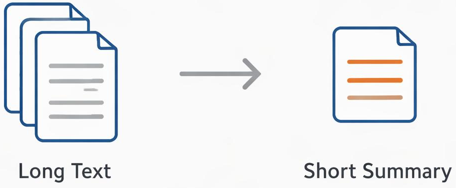
Translating text across languages, moving from sentence-level to document-level coherence.



Equal length

Text Summarization

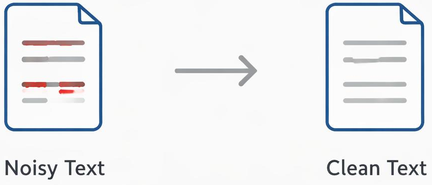
Condensing documents via extractive (selecting key sentences) or abstractive (rewriting) methods.



Compression

Grammar Error Correction

Detecting and correcting grammatical and spelling errors to output fluent, standard text.



Editing

These applications **form the backbone** of commercial writing assistants and translation engines.

Question Answering & Dialogues

Interactive systems combine retrieval, reading comprehension, and state tracking.

Question Answering (QA)

Retrieving answers from structured databases (KBQA) or unstructured text passages (Reading Comprehension).

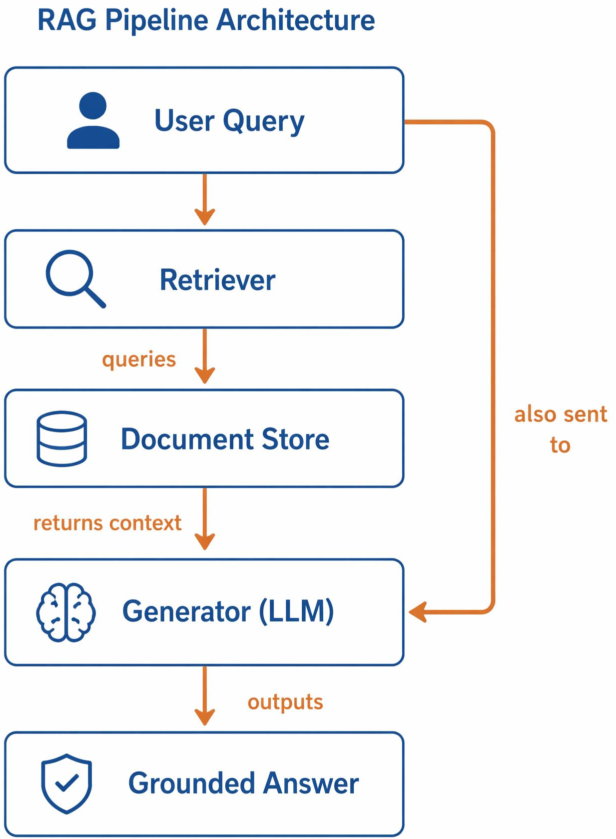
Retrieval-Augmented Generation (RAG)

Combining information retrieval with generative models to provide grounded, factual answers.

Dialogue Systems

Task-oriented systems (e.g., booking flights) track dialogue states, while chit-chat systems generate open-ended responses.

Table 1.6: LLM generation tasks (poems, mathematical proofs, Python code).



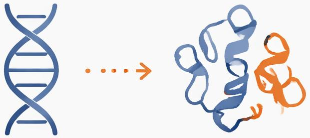
Related Fields

NLP technologies empower multimodal AI, bioinformatics, and text analytics.

Multimodal AI

Combining language with vision and robotics.

- e.g. image captioning, text-to-video, embodied AI.



Bioinformatics

Applying language models to biological sequences.

- e.g. protein structure prediction like AlphaFold.

Text Analytics

Predicting real-world outcomes based on web text and sentiment.

- e.g. stock market trends, movie revenues.

NLP methods are increasingly used as universal sequence-modeling tools across science.

NLP from a Machine Learning Perspective

Unifying NLP via Machine Learning

How unified mathematical modeling reshaped the history and practice of NLP.

The Paradigm Shift

Moving away from task-specific linguistic rules toward general-purpose statistical and neural algorithms.

The Core Thesis

Diverse linguistic tasks are mathematically identical when cast into vector spaces and probability graphs.

The evolution of NLP is a reflection of **growing computing power** and **algorithmic unification**.

Historical Evolution

The evolution of NLP is driven by exponential growth in computing power and data.

30 years · 20 years · 10 years · < 5 years

1 Rule-Based Era
(1950s–1980s)

Relied on hand-crafted dictionaries and grammar rules. Failed due to language ambiguity.

2 Statistical Era
(Late 1980s–2000s)

Algorithms learned probability distributions of linguistic patterns from annotated corpora.

3 Deep Learning Era
(Late 2000s–2018)

Stacked neural networks learned complex functions end-to-end, bypassing manual feature engineering.

4 Pre-training Era
(2018–Present)

Foundation models trained on massive data, capturing rich representations of language.

5 LLM Era
(2022–Present)

Large language models exhibit emerging capabilities like in-context learning and reasoning.

The pace of paradigm shifts has **rapidly accelerated:** 30 years to transition to statistical, 20 years to deep learning, 10 years to pre-training, and under 5 to LLMs.

Keyword Evolution

Conference keywords reveal the shifting focus of NLP research over 50 years.

1975–1994

Dominated by formal grammars, syntax, semantics, and database queries.

1995–2014

Shifted to statistical resources, Hidden Markov Models, SVMs, CRFs, and parsing.

2015–Present

Overwhelmingly focused on neural networks, transformers, pre-training, and LLMs.

Table 1.7: Frequent keywords in paper titles across five distinct historical spans.

Time Span	Frequent Keywords in Paper Titles
1975–1979	language, analysis, semantic(s), context(s), synthesis, grammar(s), sentence, rule(s), syntactic(s)
1980–1989	language, linguistic(s), machine, system(s), natural, processing, text(s), model(s)
1990–1994	language, model(s), system(s), natural, processing, knowledge, mapping, text(s)
1995–2014	model(s), system(s), language, learning, statistical, information, based, data
2015–2024	model(s), neural, learning, language, network(s), using, based, text(s)

The Core Insight

Linguistically diverse tasks map to identical mathematical formulations.

Sequence Labighg Unification

POS tagging, word segmentation, and NER are linguistically distinct but **mathematically identical** sequence labeling tasks.

POS Tagging

Word Segmentation

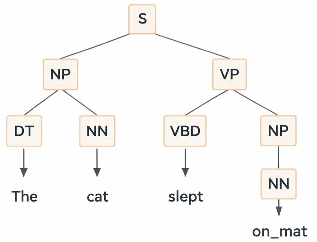
Named Entity Recognition

Different linguistic goals, **same sequence labeling formulation**.

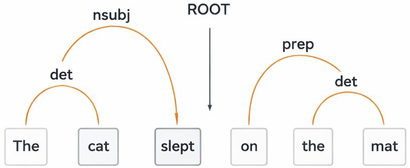
Tree Structure Unification

Constituent and dependency trees are structurally different but modeled using the **same parsing algorithms**.

Constituent Tree



Dependency Tree



Different tree structures, **same parsing algorithms**.

A single algorithmic advance (e.g. Transformers) routinely improves performance across dozens of distinct NLP tasks.

Categorizing by ML Nature

From an ML perspective, NLP tasks fall into three mathematical categories.

Classification Tasks

Predicting a discrete label from a predefined set.

Examples:
sentiment polarity,
rumor detection

Structured Prediction

Predicting complex outputs with interdependent sub-structures.

Examples:
parsing trees,
sequence labeling

Regression Tasks

Predicting continuous, real-valued numbers.

Examples:
stock price forecasting,
essay scoring

★ Structured prediction is the most dominant and challenging paradigm in NLP.

Book Structure

A unified, chronological journey through the foundations of NLP.

Part I
Basis

Mathematical modeling, vector spaces, and foundational probability graphs.

Part II
Structures

Supervised and unsupervised structured prediction algorithms.

Part III
Deep Learning

Pretrained deep neural networks, transformers, and foundation models.

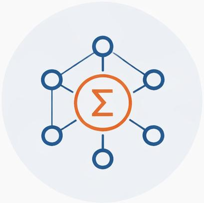
Uses a **unified mathematical framework and notation** throughout the entire book.

Summary

Re-establishing common sense: NLP is the study of mapping language to mathematics.

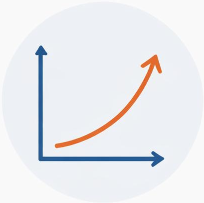
1. Task Spectrum

Linguistics, extraction, and generation form a continuous pipeline of language processing.



2. Algorithmic Unification

Machine learning provides the unified mathematical framework to solve diverse linguistic tasks.



3. Accelerating Evolution

The field is evolving at an exponential pace, driven by the synergy of compute, data, and unified models.

Chapter 1 establishes the foundational taxonomy and motivation for the rest of the book.