INTRODUCTION TO NATURAL LANGUAGE PROCESSING

The History and Evolution of NLP

From Symbolic Logic to Autonomous Agents

Victor Zhong

Two Philosophical Views on Language

The history of NLP is a 70-year pendulum swing between two schools of thought:

Rationalism

(Chomskyan)

Language is an **innate, biological structure**. To master it, we must hard-code the rules of syntax and logic.

Empiricism

(Shannon/Firth)

Language is a **learned behavioral pattern**. To master it, we must observe frequencies and context.

Historically, Empiricism has "eaten" Rationalism as compute has scaled.

The Representation Ladder

Every era of NLP is defined by how a word is represented:

1. Symbolic

```
cat = Integer ID #452
```

(Atomic, discrete)

2. Statistical

(Probabilistic)

3. Representation Learning

```
cat = [0.12, -0.5, ...]
```

(Dense vector)

4. Deep Learning

cat = A pathway to an action or a visual grounding.

(Performative)

Speech Act Theory (J.L. Austin, 1962)

Constative

Describes the world.

"The sky is blue."

Performative

Changes the world.

"I bet you five dollars."

Legacy: Early NLP focused on constative language (translation/summary). Modern NLP focuses on performative language (agents/actions).

The Turing Test (1950)

Alan Turing's Insight

Intelligence isn't defined by what is inside the box, but by how the box behaves in conversation.

Language was chosen as the ultimate benchmark because it requires reasoning, world knowledge, and intent.

The Birth of the Field (1954)

The Georgetown Experiment

The first public demonstration of Machine Translation (Russian to English).

Method

Six simple grammar rules and a dictionary of 250 words.

Outcome

Researchers predicted translation would be "solved" in 5 years. They were off by 60 years.

Formal Syntax and Generative Grammar

Context-Free Grammars (CFGs)

The Goal

To "understand" was to construct a perfect parse tree.

Limitation

Real language is too ambiguous for rigid rules (e.g., "I saw the man with the telescope").

The Pattern Matching Era (ELIZA, 1966)

ELIZA

A "chatbot" that simulated a therapist using simple string substitution.

Mechanism

If input contains "I am X", output "Why are you X?".

The ELIZA Effect

Humans projected deep intelligence onto a system with zero understanding.

Grounded Micro-Worlds (SHRDLU, 1970)

SHRDLU: The first true "agent." It lived in a virtual "Blocks World."

Capability

Could move blocks and answer questions based on logic.

Lesson

It achieved perfect accuracy in its sandbox but hit a "scaling wall"—it couldn't handle the complexity of the real world.

The First AI Winter (1966)

The ALPAC Report

Concluded that MT was slower and less accurate than humans.

Consequence

Government funding collapsed. The field learned a painful lesson: **Don't overpromise.**

Logic and Knowledge Engineering

Cyc (1984)

An attempt to manually code all "common sense" knowledge.

The Failure of Logic

Symbolic systems couldn't handle the "fuzziness" of human thought. Logic is or 1; 0 language is a spectrum.

Summary of Era 1

Philosophy

Rationalism.

Tools

LISP, Logic, Parse Trees.

Legacy

We defined the structure of language but failed to scale to its usage.

The Empiricist Turn

Fred Jelinek (IBM)

"Every time I fire a linguist, the performance of the speech recognizer goes up."

The Pivot

Stop writing rules. Start counting events in large corpora (e.g., The Wall Street Journal).

The Noisy Channel Model

Borrowed from Claude Shannon's Information Theory.

$$\hat{E} = \arg\max_{E} P(F|E)\, P(E)$$

Metaphor

A French sentence is just an English sentence that has been corrupted by "noise."

Goal

Find the English sentence E that maximizes translation into the French sentence.

N-gram Language Models

Insight

The next word is determined by the previous words.

Standard

Trigrams (3-grams) became the backbone of Speech Recognition for 20 years.

The Sparsity Problem

If a word pair never appeared in the training data, the probability was zero. We needed "Smoothing".

IBM Translation Models

Replaced grammar rules with **Word Alignment**.

> How often does "chat" (French) align with "cat" (English) in a million parallel sentences?

This data-driven approach quickly outperformed all symbolic systems.

Note: these weren't just one monolithic model, but a progressively more complex statistical sequence that defined a whole generation of research.

Machine Learning Classifiers

NLP was reframed as a series of classification tasks:

Sentiment

(+) or (-).

Spam

Yes or No.

Tools

Naive Bayes, SVMs.

The Era of Feature Engineering

The Shift

The human's job shifted from writing rules to writing **features**.

Example

"Does the word end in '-ing'?"
"Is the previous word a determiner?"

The machine optimized the weights, but the features were still hand-crafted.

Evaluation: The Rise of Metrics

To scale, we needed objective scores.

BLEU (2002)

For translation.

Perplexity

For language modeling.

The Result

NLP became an experimental science. If the score went up, the model was better.

Summary of Era 2

Philosophy

Empiricism.

Tools

HMMs, CRFs, N-grams.

Legacy

Data became the most important asset. The "Bitter Lesson" began: compute + data > human intuition.

Words as Vectors (Word2Vec)

2013: Mikolov et al. release Word2Vec.

The Idea

Words are no longer discrete symbols. They are **dense vectors** in a high-dimensional space.

King - Man + Woman = Queen.

The End of Feature Engineering

Neural Networks allowed for **Representation Learning**.

Mechanism

The model learns its own features directly from the raw text.

Impact

Human intuition was removed from the loop of defining "what matters" in a sentence.

Recurrence and Memory (RNNs/LSTMs)

RNNs

Language is sequential. RNNs processed text word-by-word, maintaining a "hidden state."

LSTMs (1997/2014)

Solved the "Vanishing Gradient" problem: gradients become too small to update weights in early layers during backprop, allowing models to remember context from the beginning of a paragraph.

Sequence-to-Sequence (Seq2Seq)

2014: Sutskever et al. introduce the Encoder-Decoder.

Encoder

Compresses a sentence into a single "context vector."

Decoder

Unpacks that vector into a new language.

Outcome: Neural Machine Translation (NMT) replaced IBM models overnight.

The Attention Mechanism (2015)

The Bottleneck

Compressing a 50-word sentence into one vector is too hard.

Attention

Allowed the decoder to "look back" at specific parts of the input sentence at every step.

This was the first step toward the Transformer.

Summary of Era 3

Philosophy

Connectionism.

Tools

Word2Vec, LSTMs, Attention.

Legacy

NLP moved from discrete symbols to continuous vector spaces.

The Transformer Revolution (2017)

"Attention is All You Need"

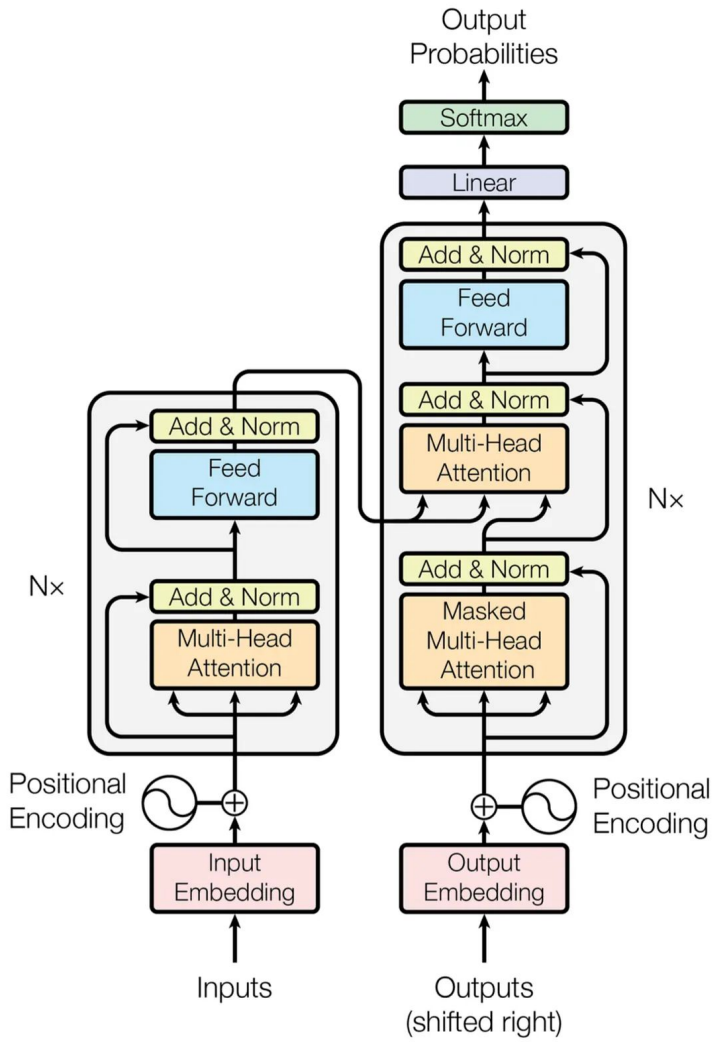
Discarded Recurrence

Replaced LSTMs entirely with Self-Attention.

Parallelism

Processed all words in a sentence simultaneously.

This allowed for massive scaling on GPUs.



BERT and Contextualization (2018)

Pre-2018: "Bank" (river) and "Bank" (money) had the same vector.

BERT

Used a Masked Language Model (MLM) objective to create dynamic embeddings based on context.

Proved that **"Pretraining + Fine-tuning"** was the winning formula.

The Rise of Generative Pretraining (GPT)

GPT-1 & 2: Showed that predicting the next word is a "universal" task.

If a model can predict the next word well, it must understand syntax, facts, and reasoning.

GPT-3 and In-Context Learning (2020)

Zero-shot/Few-shot

Proved you don't need to "fine-tune" weights for new tasks. You just need to "prompt" the model.

Emergent Abilities

Scaling from 1B to 175B parameters unlocked abilities that small models didn't have (e.g., translation, coding).

Scaling Laws (Kaplan et al., 2020)

Performance is a predictable power law of three factors:

Compute

(FLOPs)

Dataset Size

(Tokens)

Model Size

(Parameters)

Architecture details matter less than pure scale.

The Shift from Chat to Agency

Foundation Models

Good at "thinking" (text in → text out).

Agents

Good at "acting" (observation → environment action).

We are now climbing the "Ladder of Agency."

Level 1 - Structured Tool Use (APIs)

Gorilla (2023)

Proved that LLMs could be fine-tuned to use thousands of real APIs without hallucination.

APIBench

A transition from "Language as description" to "Language as a command for a tool."

Level 2 - Unstructured Code Action

SWE-bench (2024): Challenged agents to fix real bugs in GitHub repositories.

The Insight

To fix code, an agent must explore directories, read documentation, and verify via execution.

Execution-Based Eval

Success is defined by the code working, not by looking like human text.

Level 3 - OS and GUI Control

OSWorld (2024): A benchmark for agents to control a full desktop (Ubuntu/Windows).

Multimodality

Agents must now "see" (Screenshots) and "act" (Mouse/Keyboard).

The Grounding Problem

Linking the command "Save file" to precise pixel coordinates (x, y).

Planning and World Modeling

The Limits of Reactivity

Basic agents are "short-sighted." They click without thinking about the consequence.

WebDreamer (2025)

Introduced **Model-Based Planning**.

Before acting, an agent simulates the outcome in a latent "World Model" to choose the safest, most efficient path.

Scaling Agentic Data

The Data Bottleneck

We have trillions of text tokens, but few examples of "how to use a computer."

OpenCUA (2025)

Introduced **AgentNet**, a massive dataset of real human computer-use trajectories.

Reflective CoT

Training models to "talk to themselves" to detect and recover from errors.

Summary of Era 4

Philosophy

Generalist Foundation Models and Agents.

Tools

LLMs, GPUs, external interfaces as tools.

Legacy

TBD

Agents: We are learning the consequences.

The Arc of NLP Progress

Symbols: We defined the atoms.

Probability: We learned the patterns.

Vectors: We learned the relationships.

Foundations: We learned the context.

Agency: We learn to act in context.

The Bitter Lesson Revisited

The history of NLP has consistently rewarded **Search** and **Learning** over human-designed structure.

Agents are the ultimate test of this: we don't code the "browser logic"; we let the model learn to navigate the pixels.

Ethics and Societal History

Early Era

Fear of "Deceptive machines" (ELIZA).

Statistical Era

Concerns about "Data Privacy" and "Copyright."

Representation Era

Concerns about “Biases and Fairness”.

Deep Learning Era

Fears of "Autonomous Harm" (AI acting without oversight).

The Hardware Lottery in History

NLP history is inseparable from hardware:

1950s

Mainframes (Symbolic).

1990s

CPUs (Statistical).

2010s

GPUs (Neural/Transformer).

2025+

Clusters/Interconnects (Scaling Agents).

From Meaning to Intent

The Past

The field used to ask: "What does this sentence mean?" (Semantics).

The Present

The field now asks: "What does the user want me to do?" (Pragmatics/Agency).

The End of "NLP" as a Subfield?

- NLP is no longer isolated from Vision or Robotics.
- Language has become the **Universal Operating System** for all AI interaction.
- The "Natural Language" is now the programming language for the physical world.

Conclusion: The Trajectory of Intelligence

We have moved from machines that **calculate** to machines that **generate**, and now to machines that **act**.

The future of NLP is the future of Autonomous Problem Solving.