

Natural Language Processing

Course Overview

Berlin Chen

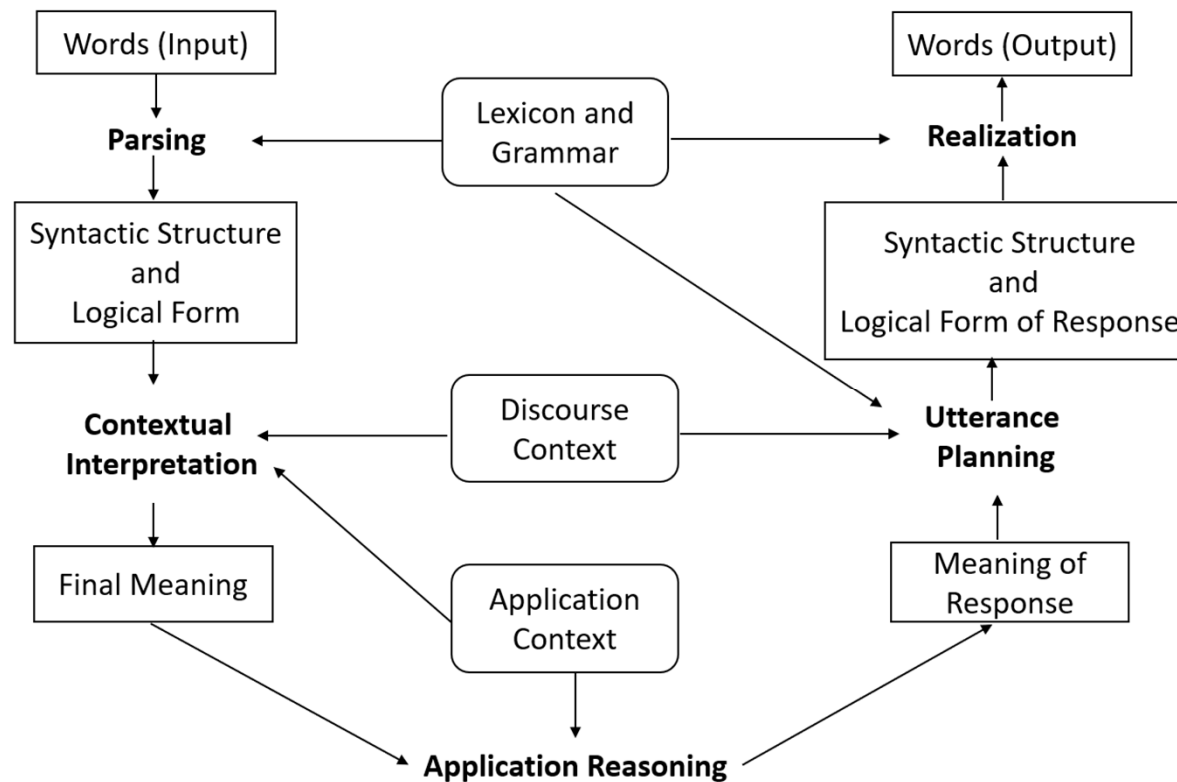
Department of Computer Science & Information Engineering
National Taiwan Normal University

Main Objectives of this Course

- Provide a technical perspective on natural language processing (NLP)
- Both the theoretical and practical issues of NLP will be considered
 - Mathematical and linguistic foundations
 - Classic and novel NLP applications (tagging, parsing, WSD, NE, MT, NLU, NLG, IR, CALL etc.)
- Statistical modeling methods for NLP and associated learning paradigms will be further emphasized
 - Shallow and deep machine learning (ML) models
 - Supervised and unsupervised training algorithms of ML models
- Learn skills to analyze and build NLP systems, and understand the latest research trends

Processing Pipeline of NLU and NLG (in Old Days?)

- A pathway for natural language understanding (NLU) & natural language generation (NLG)



Some Textbooks and References

- D. Jurafsky and J. H. Martin, *Speech and Language Processing* (3 Ed.)
- C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999
- J. Allen, *Natural Language Understanding*, Benjamin/Cummings Publishing Co, 1995
- J. Eisenstein, *Introduction to Natural Language Processing*, MIT Press, 2019

Tentative Schedule

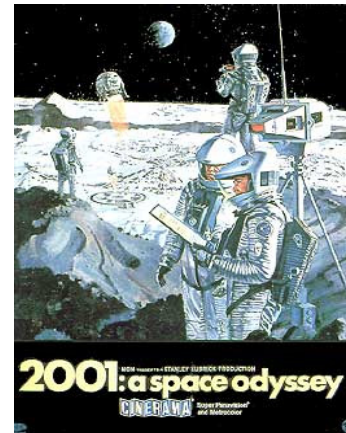
- Course Overview
- Mathematical Foundations
- Corpus and Linguistic Essentials
- Vector Semantics and Embeddings
- More on Neural Language Models
- Part of Speech Tagging and Named-Entity Extraction
- Parsing with Context-Free Grammars
- Text Clustering, Classification and Summarization
- Computer-assisted Language Learning and Readability Assessment*
- Probabilistic Context-Free Grammars
- Collocation and Word Sense Disambiguation
- Statistical Sentence Alignment and Machine Translation
- Invited Talks*
- Paper Presentations*

Grading (*Tentatively!*)

- Midterm and Final: 40% (Open-book Exams)
 - The final exam may be replaced by the activity of paper survey and presentation
- Homework: 40%
 - Paper-and-pencil or software homework assignments
- Attendance/Other: 20%

Hollywood Cinema

- Computers/robots can listen, speak, and answer our questions
 - E.g.: HAL 9000 computer in “[2001: A Space Odyssey](#)” (Stanley Kubrick, 1968)
(2001太空漫遊)



HAL: Conversational Agent/Robot
電影走在科技之前?

- This film is noted for its scientifically accurate depiction of space flight, pioneering special effects, and ambiguous imagery

[https://en.wikipedia.org/wiki/2001:_A_Space_Odyssey_\(film\)](https://en.wikipedia.org/wiki/2001:_A_Space_Odyssey_(film))

Language

- Language is one of the fundamental aspects of human behavior and is a crucial component of our lives
 - In written form it serves as a long-term record of knowledge from one generation to the next
 - In spoken form it serves as our primary means of coordinating our day-to-day behavior with others
- It has long been our ultimate goal to be able to create **computational models** that approach human performance in the linguistic tasks of reading, writing, hearing, and speaking

Knowledge of Language

discourse

Pragmatics

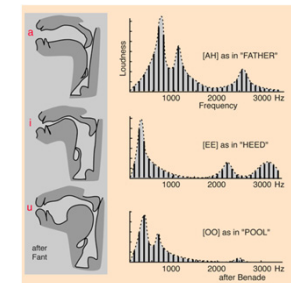
semantics

Syntax

morphology

Phonology

Phonetics



The ELIZA Program (Weizenbaum at MIT, 1966)

- ELIZA was an early NLP system capable of carrying on a limited form of conversation with a user

User₁: You **are like** my father in some ways.

ELIZA₁: WHAT RESEMBLANCE DO YOU SEE (or IN WHAT WAY)

User₂: You **are** not **very aggressive** but I think you **don't want me to notice that**.

ELIZA₂: WHAT MAKES YOU THINK I AM NOT AGGRESSIVE

User₃: You **don't** argue with me.

ELIZA₃: WHY DO YOU THINK I DON'T ARGUE WITH YOU

User₄: You are **afraid** of me.

ELIZA₄: DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU

- A version of ELIZA that imitated the responses of a Rogerian psychotherapist
- ELIZA has no world knowledge and cannot model the conversation history

A Simple Description of how ELIZA Works

- There is a database of particular words that are called keywords
- For each keyword, the system stores an integer, a pattern to match against the input, and a specification of the output
- The algorithm is as follows:
 - Given a sentence S, find a keyword in S whose pattern matches S. If there is more than one keyword, pick the one with the highest integer value. Use the output specification that is associated with this keyword to generate the next sentence
 - If there are no keywords, generate an innocuous continuation statement, such as "*Tell me more*" or "*Go on*"
- ELIZA sometimes produces completely off-the-wall (離譜的) responses
 - E.g.: if you say "*Necessity is the mother of invention*", it might respond with "*Tell me more about your family*"
- As a conversation progresses, it becomes obvious that ELIZA does not retain any of the content in the conversation

IBM Watson (IBM, 2011)

- Watson was created as a question answering (QA) computing system that IBM (DeepQA project) built to apply advanced natural language processing, information retrieval, knowledge representation, automated reasoning, and machine learning technologies to the field of open domain question answering (https://en.wikipedia.org/wiki/IBM_Watson)



- IBM Watson cannot help deal with rare diseases when only insufficient training data is provided

<https://www.ibm.com/watson>

<https://technews.tw/2021/02/22/ibm-watson-ai/>

Knowledge of Language

- To engage in complex language behaviors, we require various kinds of knowledge of language
 - **Phonetics and Phonology**: Knowledge about linguistic sounds
 - **Morphology**: Knowledge of the meaningful components of words
 - **Syntax**: Knowledge of the structural relationships between words
 - **Semantics**: Knowledge of meaning
 - **Pragmatics**: Knowledge of the relationship of meaning to the goals and intentions of the speaker
 - **Discourse**: Knowledge about linguistic units larger than a single utterance
- Wisdom of language?

《莊子·外物》 (369-286 B.C.)

荃者所以在魚，得魚而忘荃；蹄者所以在兔，得兔而忘蹄；言者所以在意，得意而忘言。
吾安得夫忘言之人而與之言哉！

Text vs. Speech

- Recognition, Analysis and Understanding
 - **Text**: analyze and understand text
 - **Speech**: recognize speech (i.e., ASR), and subsequently analyze and understand the recognized text (propagations of ASR errors)
- Variability
 - **Text** : different synonyms may refer to the same specific semantic object or meaning, such as “bicycle”, “cycle”, “bike,” etc.
 - **Speech** : an infinite number of utterances with respect to the same words (e.g., “bicycle”)
 - Manifested by a wide variety of oral phenomena such as disfluences (hesitations), repetitions, restarts, and corrections
 - Gender, age, emotional and environmental variations further complicate ASR
 - No (correct) punctuation marks (delimiters) or/and structural information cues exist in (the transcripts of) speech

Major Disciplines Studying Language

- This common goal has motivated a new area of interdisciplinary research often called **cognitive science**

Discipline	Typical Problems	Tools
Psycholinguists	How do people identify the structure of sentences? How are word meanings identified? When does understanding take place?	Experimental techniques based on measuring human performance; statistical analysis of observations
Philosophers	What is meaning, and how do words and sentences acquire it? How do words identify objects in the world?	Natural language argumentation using intuition about counter-examples; mathematical models (for example, logic and model theory)
Computational Linguists & Computer Scientists	How is the structure of sentences identified? How can knowledge and reasoning be modeled? How can language be used to accomplish specific tasks?	Algorithms, data structures; formal models of representation and reasoning; AI & ML techniques (search and representation methods)

freedom fighters
(defensive strikes)
vs.
terrorists

Natural Language Processing (NLP)

- Natural language processing is the set of methods for making human language accessible to computers
- In the past decade, natural language processing has become embedded in our daily lives

Text-based

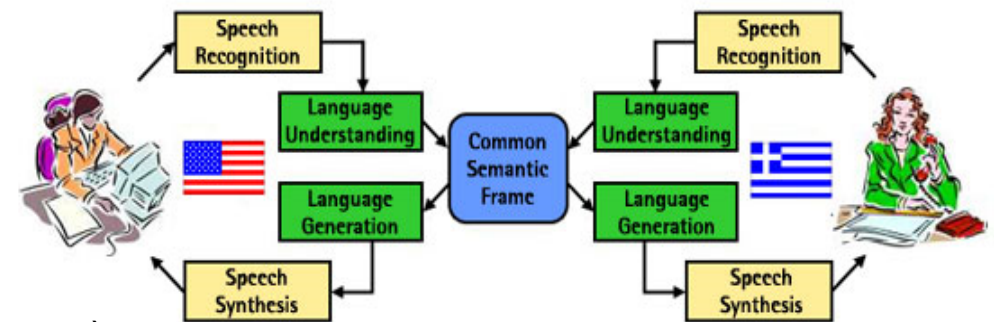
- Machine Translation (MT)
- Information Retrieval (IR)
- Text Clustering and Classification
- Text Summarization (TS)

Dialogue-based

- Conversational Agents (Question-Answering, IVR & Dialogue Systems)
- Tutoring, CALL & Essay Grading Systems

• ...

- These diverse applications are based on a common set of ideas, drawing on algorithms, linguistics, logic, statistics, and more



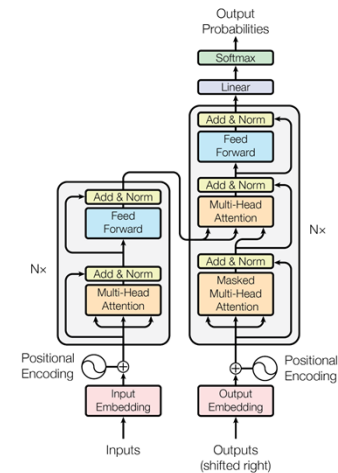
Should we have to devise linguistically motivated computational models of language understanding and production in the era of deep learning?

Current State-of-the-art Conversational Agents

- There has been a surge of interest recently in the development of conversational agents (QA systems)

科技巨擘AI大戰	
Google	6日推出「Bard」服務，目前由一群「受信任的測試人員」試用，計劃未來幾周向大眾推出，未來會加入Google 搜尋引擎
微軟	近幾年大手筆投資ChatGPT的開發商OpenAI，預定7日發表融入聊天機器人版本的Bing搜尋引擎，還將開放工具給開發者進行創建
百度	7日宣布開發出類似ChatGPT的產品「文心一言」，正進行上線前的衝刺，3月將完成內部測試，向公眾開放
蘋果	下周可能召開內部AI高峰會，形式類似WWDC，預料會提出相關議題

large language models (LLM)
with prompt-based learning



Transformer
One model fits all?

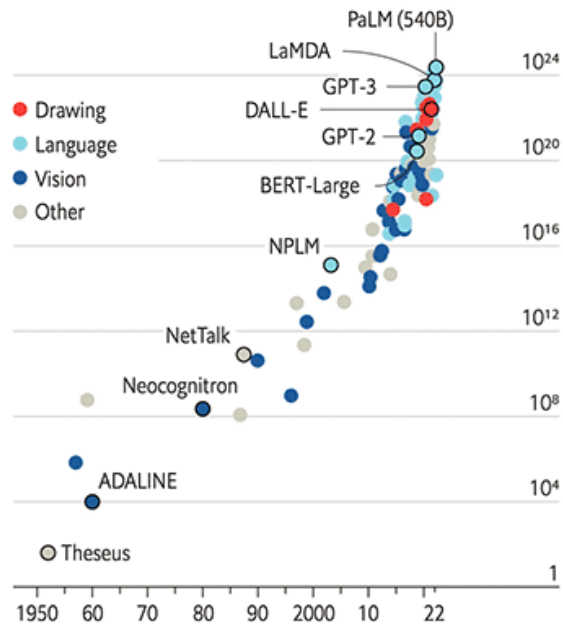
- We have witnessed good success in building deep-learning (neural network) based models for natural language understanding (NLU) and generation (NLG)
 - Generalization capability? simultaneous (multitask) learning vs. lifelong (continual) learning

1. Source: <https://money.udn.com/money/story/11162/6956769> (2023/02/08)
 2. A. Vaswani et al., "Attention Is all you need," arxiv, 2017

Exceeding Human Performance on Many Tasks (1/2)

The blessings of scale

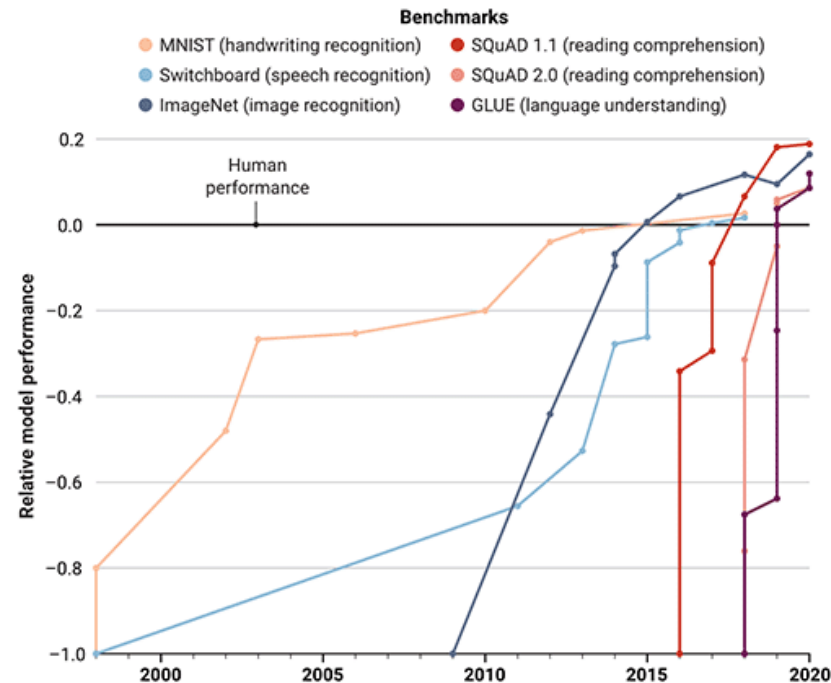
AI training runs, estimated computing resources used
Floating-point operations, selected systems, by type, log scale



Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data

Quick learners

The speed at which artificial intelligence models master benchmarks and surpass human baselines is accelerating. But they often fall short in the real world.



(GRAPHIC) K. FRANKLIN/SCIENCE; (DATA) D. KIELA ET AL., DYNABENCH: RETHINKING BENCHMARKING IN NLP, DOI:10.48550/ARXIV.2104.14337

Source: <https://www.sequoiacap.com>

Exceeding Human Performance on Many Tasks (2/2)

	PRE - 2020	2020	2022	2023?	2025?	2030?
TEXT	Spam detection Translation Basic Q&A	Basic copy writing First drafts	Longer form Second drafts	Vertical fine tuning gets good (scientific papers, etc)	Final drafts better than the human average	Final drafts better than professional writers
CODE	1-line auto-complete	Multi-line generation	Longer form Better accuracy	More languages More verticals	Text to product (draft)	Text to product (final), better than full-time developers
IMAGES			Art Logos Photography	Mock-ups (product design, architecture, etc.)	Final drafts (product design, architecture, etc.)	Final drafts better than professional artists, designers, photographers)
VIDEO / 3D / GAMING			First attempts at 3D/video models	Basic / first draft videos and 3D files	Second drafts	AI Roblox Video games and movies are personalized dreams

Large model availability: ● First attempts ● Almost there ● Ready for prime time

Source: <https://www.sequoiacap.com>

ChatGPT: Hype or Holy Grail for NLP?



2023.02.13 | AI與大數據

ChatGPT背後「血汗工廠」：時薪40、每小時標註2萬單詞，數據標註員有多慘？

在ChatGPT成功的背後，有類「數據標註員」標記有害的內容，然而他們時薪僅約新台幣39.9元~60.5元，這些幾乎「隱形」的工人仍然處於最邊緣地帶。

36氪

#AI #CHATBOT

2,396

讚



24



ChatGPT被譽為當年最具創新性的人工智慧工具之一。這項強大的AI聊天機器人幾乎可以生成任何問題的文本，從莎士比亞的十四行詩，到用5歲孩子都能理解語言描述的複雜數學問題，可謂「無所不能」。

瑞銀此前發布的一份研究報告顯示，ChatGPT的月活躍用戶在今年1月份預計達到了1億，這距離其推出只有2個月時間，成為史上增長最快的消費者應用。

然而，在ChatGPT掀起的這場人工智慧「革命」和資本市場狂歡的背後，有一個被邊緣化的、被遺忘的卻又至關重要的群體，值得所有人關注，他們就是**數據標註員**。

為了訓練ChatGPT，大量數據標註員時薪不到2美元

據美國《時代周刊》上月中旬的報導，為了訓練ChatGPT，OpenAI僱備了時薪不到2美元的肯尼亞外包勞工，他們所負責的工作就是對龐大的數據庫手動進行數據標註。

1. <https://www.bnext.com.tw/article/74065/chatgpt-bethsc->
2. E. A. M. van Dis et al, "ChatGPT: five priorities for research," nature, vol. 614, 2023
3. T. Y. Zhuo et al., "Exploring AI ethics of ChatGPT: A diagnostic analysis," arxiv, 2023

ChatGPT: How to be a master of NLP?

BE How to be a master of NLP?



Becoming a master of Natural Language Processing (NLP) requires dedication, persistence, and continuous learning. Here are some steps that you can take to improve your skills and knowledge in NLP:



1. Learn the fundamentals: Start by learning the basics of NLP, including its history, key concepts, and techniques. This will give you a strong foundation for understanding more advanced topics.
2. Study the latest research: Keep up to date with the latest research in NLP by reading academic papers and attending conferences. This will help you stay abreast of the latest trends and techniques in the field.
3. Practice, practice, practice: Work on projects that involve NLP, such as text classification, sentiment analysis, or machine translation. This will help you build hands-on experience and improve your skills.
4. Use the right tools: Learn how to use popular NLP tools and libraries such as NLTK, spaCy, and TensorFlow. These tools will make your work more efficient and effective.
5. Collaborate with others: Work with other NLP practitioners, join NLP communities and forums, and participate in hackathons and competitions. This will give you exposure to different perspectives and approaches to NLP.
6. Keep learning: NLP is a rapidly evolving field, so it's important to keep learning and staying up-to-date with the latest trends, techniques, and technologies.

By following these steps, you can improve your NLP skills and become a master of this exciting field.

🔄 Regenerate response

People are extremely good at distinguishing word meanings and interpreting sentences to fit the context. ChatGPT (or previously ELIZA & Waston) appears to be intelligent because you use your own intelligence to make sense of what it says.

Can GPT Perform Arithmetic and Reasoning Tasks Rationally?

BE 小明給小華2塊錢，小莉給小華3塊錢，小明給小莉1塊錢，小華又各還給小明和小莉1塊錢，小華將這些錢投資後變成2倍，小華最後一共獲得幾塊錢？

以下是解題的過程：

一開始，小明給小華2塊錢，小莉給小華3塊錢，小明給小莉1塊錢，小華就有5塊錢。

然後，小華又各還給小明和小莉1塊錢，那麼小華手上就剩下 $5-2-3+1+1+1=3$ 塊錢。

接下來，小華將這些錢投資後變成2倍，那麼小華最後一共獲得6塊錢。

因此，小華最後一共獲得6塊錢。

BE 小明給小華2塊錢，小莉給小華3塊錢，小明給小莉1塊錢，小華又各還給小明和小莉1塊錢，小華將這些錢投資後變成2倍，小華最後一共獲得幾塊錢？

我們可以用代數的方式來解決這道題目，讓 x 表示小華最初所收到的錢的數量，根據題目中的描述，我們可以得到以下等式：

小明給小華2塊錢： $x + 2$

小莉給小華3塊錢： $x + 3$

小明給小莉1塊錢： $x - 1$

小華還給小明1塊錢： $(x + 2) - 1 = x + 1$

小華還給小莉1塊錢： $(x + 3) - 1 = x + 2$

綜合以上等式，我們可以得到下列方程式：

$$(x + 2) + (x + 3) + (x - 1) - (x + 1) - (x + 2) = x$$

化簡後得到：

$$x = 1$$

因此，小華最初收到的錢的數量為1元，投資後獲得的金額為原本的兩倍，因此小華最後一共獲得2元。

BE 小明給小華2塊錢，小莉給小華3塊錢，小明給小莉1塊錢，小華又各還給小明和小莉1塊錢，小華將這些錢投資後變成2倍，小華最後一共獲得幾塊錢？

首先，小明給小華2塊錢，小莉給小華3塊錢，小明給小莉1塊錢，這些交易後，小華手上有 $2 + 3 - 1 = 4$ 塊錢。

然後，小華向小明和小莉各還1塊錢，也就是總共還了 $1 + 1 = 2$ 塊錢，因此，小華手上剩下的錢是 $4 - 2 = 2$ 塊錢。

接下來，小華將這2塊錢投資，最終變成2倍，也就是變成了 $2 \times 2 = 4$ 塊錢。

因此，小華最後一共獲得了4塊錢。

BE Joe Biden是誰？他前一份工作的老闆是誰？他現在的工做是承繼誰？

Joe Biden是美國現任總統，於2021年1月20日就職，是美國歷史上第46任總統。他也曾在美國政壇擔任多年職務，包括美國參議院議員（1973年至2009年）和副總統（2009年至2017年）。

在他的政治生涯之外，Joe Biden也曾在法律領域工作。他曾於1968年至1972年期間擔任過特拉華州的縣議會成員，此後便當選為聯邦參議員，擔任該職位長達36年。在2019年時，他宣布參加2020年美國總統選舉，最終在2020年贏得該選舉，成為美國總統。

在他的政治生涯中，Joe Biden曾擔任過美國副總統，他的老闆是當時的美國總統巴拉克·奧巴馬。而現在他的工作是美國總統，他的前任是前總統唐納德·特朗普。

Should we just replicate the functionality or instead overcome the limitations of ChatGPT?

1. **AI Fairness:** Data bias, model(algorithmic/evaluation) bias, user bias
2. **AI Ethics:** plagiarism issues (citations), accountability & transparency (explainable), patent claims ...
3. **Applications:** low-stakes vs. high-stakes
4. **Environment-friendly:** Energy consumption and carbon emission

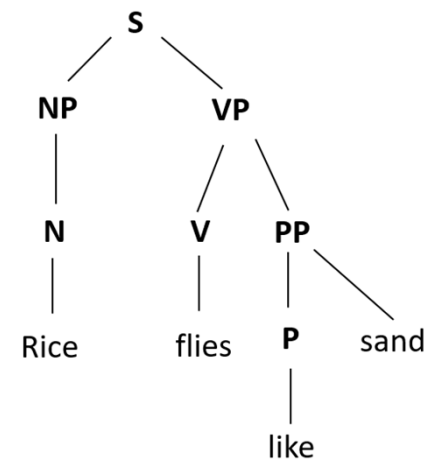
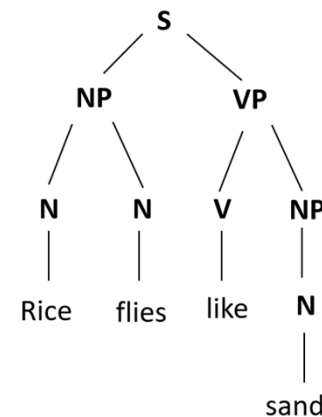
Two Different Perspectives on NLP

- **Researchers:** explore the nature of linguistic communication
 - Obtain a better understanding of how languages work
- **Practitioners:** enable effective human-machine communication
 - **Conversational Agents** are becoming an important form of human-computer communication
 - **Nature Language Interface (NLI)** for Programming/Data Science (Codex, ChatGPT, etc.)
 - ...
 - Revolutionize the way computers are used
 - More flexible and intelligent

Major Topics for NLP (1/2)

- Probability Theory, Linear Algebra, Machine Learning
 - Shallow and Deep Learning based Techniques
 - Supervised/Unsupervised ML Techniques
- Words
 - Morphology
 - Regular Expressions
 - Automata, Finite-State Transducers
- Syntax
 - Part-of-Speech Tagging
 - (Probabilistic) Context-Free Grammar
 - Parsing

1. **Flying planes** is dangerous.
2. **Flying planes** are dangerous.



Major Topics for NLP (2/2)

- Semantics/Meaning

- Representation of Meaning
- Semantic Analysis
- Word Sense Disambiguation

- Pragmatics

- Natural Language Generation
- Machine Translation
- Discourse, Dialogue and Conversational Agents

- bank [1, noun]: the rising ground bordering a lake, river, or sea...(岸)
- bank [2, verb]: to heap or pile in a bank (築堤防護)
- bank [3, noun]: an establishment for the custody, loan, or exchange of money (銀行)
- bank [4, verb]: to deposit money (存錢)
- bank [5, noun]: a series of objects arranged in a row (排,組)

5 out of 28 definitions in Webster's Dictionary online <http://www.m-w.com>

Spanish: La botella entró a la cueva flotando
(The bottle floated into the cave)
English: The bottle entered the cave floating

(In Spanish, the direction is expressed using the verb and the manner is expressed with a separate phrase)

Jack **invited** Mary to the Halloween **ball**.

Dissidences: Rationalists and Empiricists

- Rationalists (e.g. [Norm Chomsky](#))
 - Humans are innate language faculties
 - (Almost fully) encoded rules plus reasoning mechanisms
 - Dominating between 1960's~mid 1980's
- Empiricists (e.g. [Claude Shannon](#))
 - The mind does not begin with detailed sets of principles and procedures for language components and cognitive domains
 - Rather, only general operations for association, pattern recognition, generalization etc., are endowed with
 - General language models plus machine learning approaches
 - Dominating between 1920's~mid 1960's and resurging 1990's~

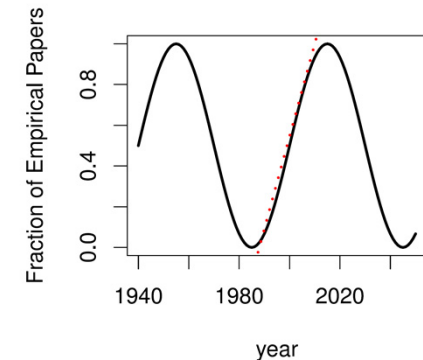


FIGURE 2 An extreme view of the literature, where the trend in Figure 1 (denoted by a dashed red line) is dominated by the larger oscillation every couple of decades. Note that that line is fit to empirical data, unlike the oscillation which is drawn to make a point.

Dissidences: Statistical and Non-Statistical NLP

- The dividing line between the two has become much more fuzzy recently
 - An increasing number of non-statistical researches use corpus evidence and incorporate quantitative methods
 - Corpus: “a body of texts”
 - Statistical NLP needs to start with all the scientific knowledge available about a phenomenon (with [Knowledge Injection](#)) when building a probabilistic model, rather than closing one’s eye and taking a clean-slate approach
 - Probabilistic and data-driven
 - Clean Slate: a state in which you are starting an activity or process again, not considering what has happened in the past at all
- Statistical NLP → “Language Technology” or “Language Engineering”

Ambiguity of Language (1/4)

- Pronunciation Ambiguity

- A text-to-speech synthesis system reading the word *lead* needs to decide whether it should be pronounced as in *lead pipe* or as in *lead me on*

- Spatial Ambiguity (Reasoning)

The trophy doesn't fit into the brown suitcase because it is too [small/large].

- Natural language understanding cannot be achieved in isolation from knowledge and reasoning

Ambiguity of Language (2/4)

- Lexical Ambiguity

I made her duck.

- Deciding whether *duck* is a verb or a noun can be solved by **part-of-speech tagging**,
- Deciding whether *make* means “create” or “cook” can be solved by **word sense disambiguation**
- Lexical disambiguation: part-of-speech (POS) tagging + word sense disambiguation

- Speech Act Interpretation

- Determine whether a sentence is a statement or question or request

Dave: HAL, open the pod bay door.

(REQUEST)

Dave: HAL, the pod bay door is open.

(STATEMENT)

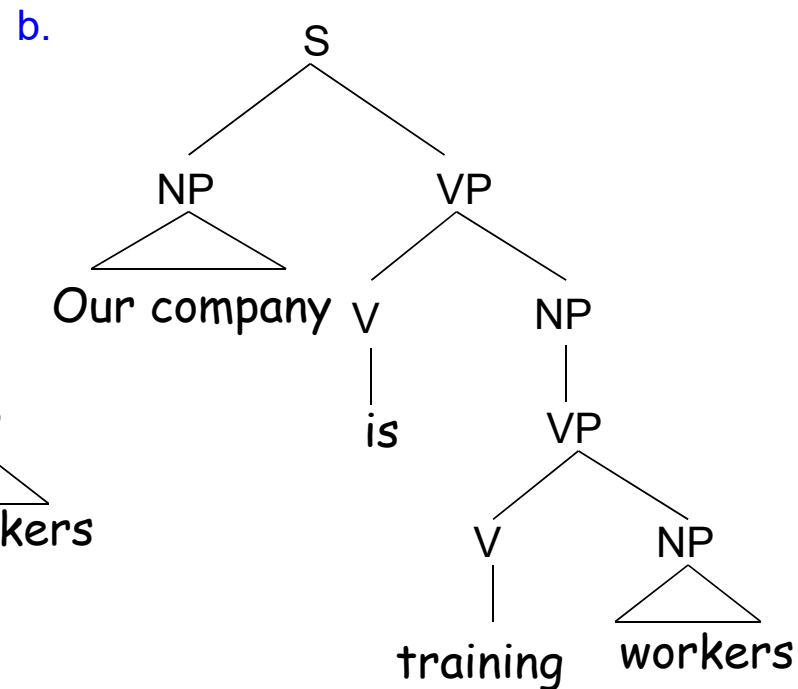
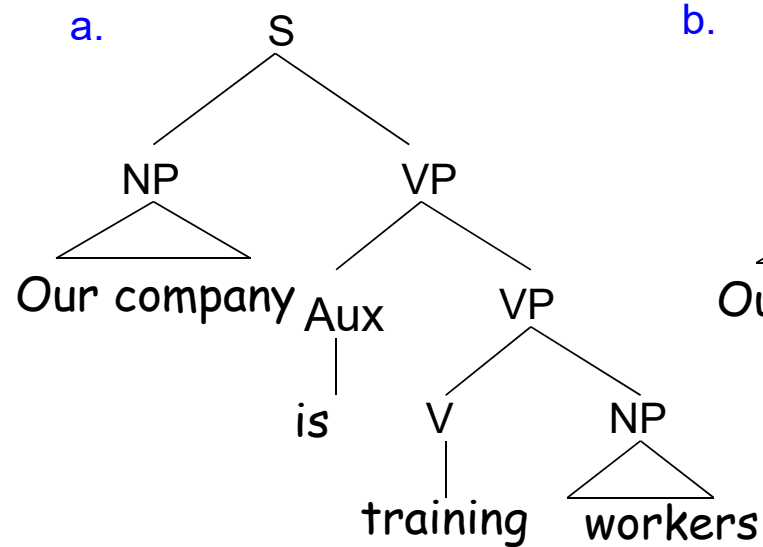
Dave: HAL, is the pod bay door open?

(INFORMATION QUESTION)

Ambiguity of Language (3/4)

- Syntactic Ambiguity

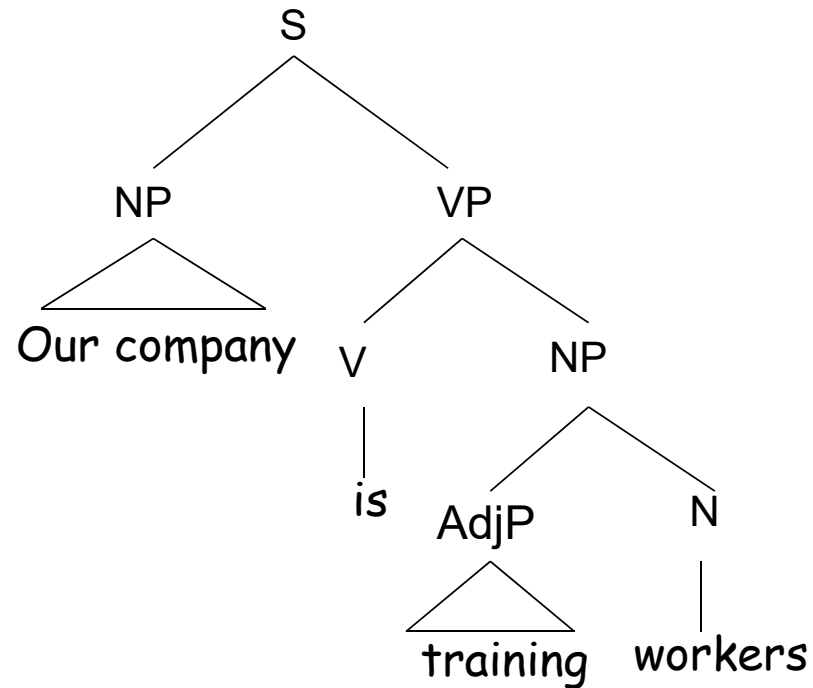
- A simple sentence, such as “Our company is training workers,” has 3 syntactic analyses (parses)



(Cf. Our problem is training workers.)

Ambiguity of Language (4/4)

c.



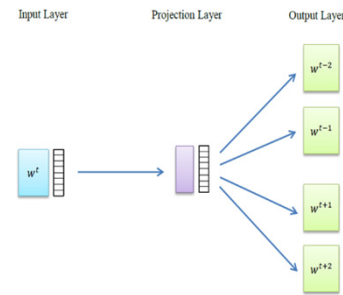
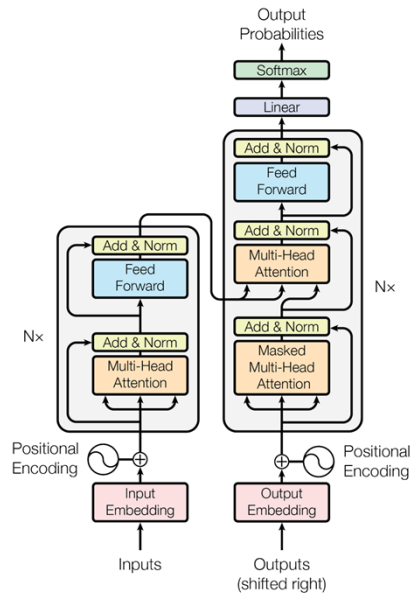
(Cf. Those are training wheels.)

- The last two parses (b. and c.) are semantic anomalous!

Natural Language Processing and its Neighbors

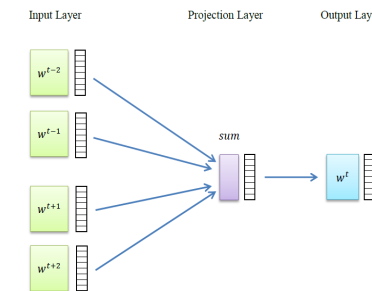
- Computational Linguistics
- Machine Learning
- Artificial Intelligence
- Speech Processing
- Ethics
- ...

Transformer
(e.g., for Machine Translation)

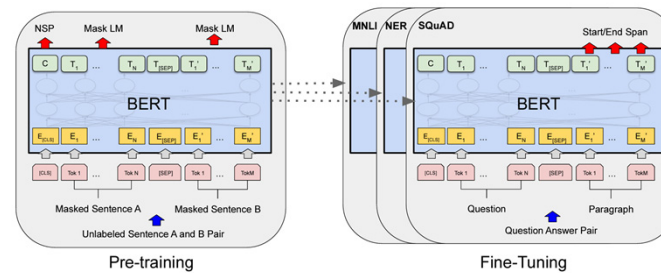


static embeddings

Skip-gram



CBOW



Contextualized Embedding

BERT

Comparison between BERT, GPT and ELMo

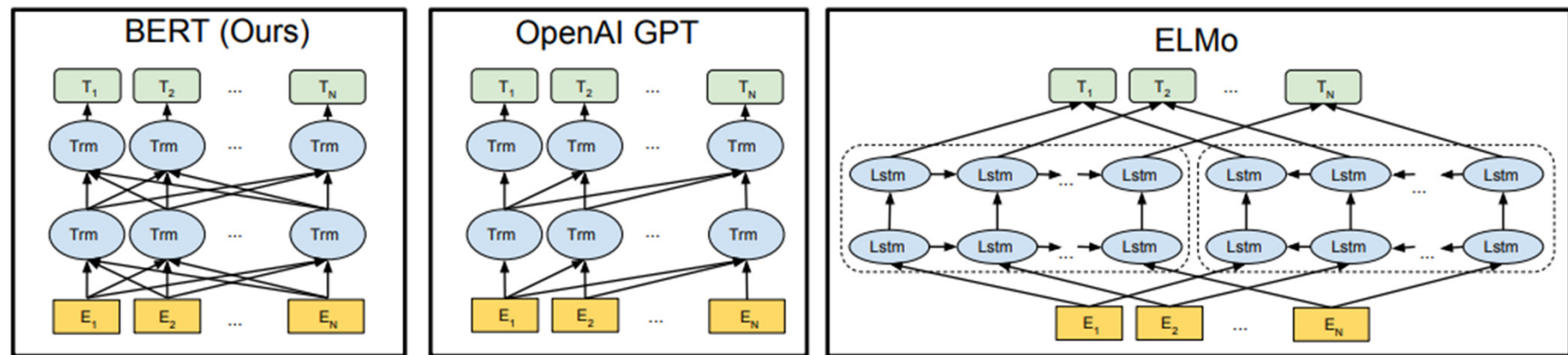
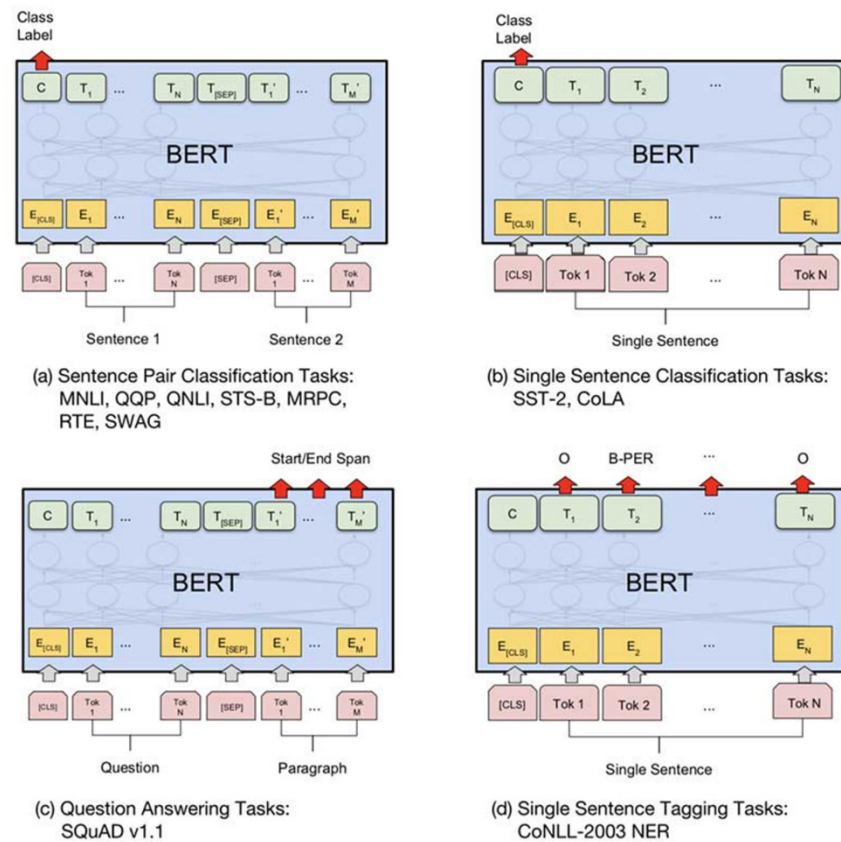


Figure 3: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTMs to generate features for downstream tasks. Among the three, only BERT representations are jointly conditioned on both left and right context in all layers. In addition to the architecture differences, BERT and OpenAI GPT are fine-tuning approaches, while ELMo is a feature-based approach.

Some Typical NLP Tasks Leveraging BERT

- Fine-tuning BERT in different tasks



Computational Linguistics

- Most of the meetings and journals that host natural language processing research bear the name “computational linguistics”, and the terms may be thought of as essentially synonymous
- There is an important difference in focus
 - In linguistics, language is the object of study
 - Computational methods may be brought to bear, just as in scientific disciplines like computational biology and computational astronomy, but they play only a supporting role
 - NLP is focused on the design and analysis of computational algorithms and representations for processing natural human language
 - The goal of natural language processing is to provide new computational capabilities around human language
 - For example, extracting information from texts, translating between languages, answering questions, holding a conversation, taking instructions, and so on

Machine Learning (or Deep Learning)

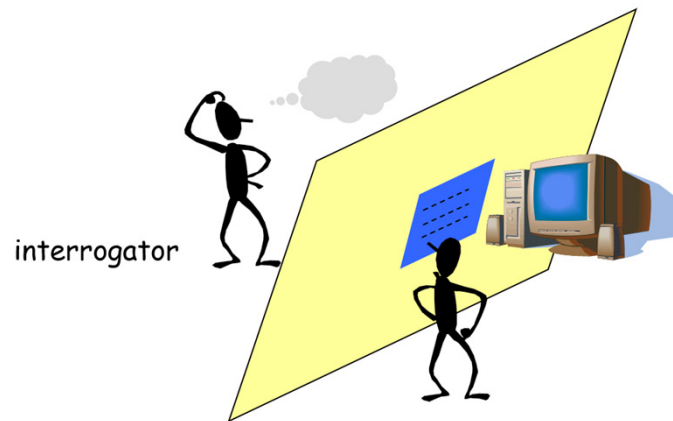
- Contemporary approaches to natural language processing rely heavily on machine learning (or deep learning), which makes it possible to build complex computer programs from examples (statistical approaches)
- Machine learning provides an array of general techniques for tasks like machine translation
 - Converting a sequence of discrete tokens in one vocabulary to a sequence of discrete tokens in another vocabulary
- Much of today's natural language processing research can be thought of as applied machine learning
 - However, natural language processing has characteristics that distinguish it from many of machine learning's other application domains

Artificial Intelligence (AI)

- The goal of AI is to build software and robots with the same range of abilities as humans
 - Thinking Humanly (Rationally) & Acting Humanly (Rationally)
- Research on AI has experienced several hype cycles, followed by disappointment and criticism, followed by funding cuts, followed by renewed interest years or even decades later
- NLP is relevant to this goal in several ways (is known as the crown jewel of AI)
 - The capacity for language is one of the central features of human intelligence, and is therefore a prerequisite for AI
 - Much of AI research is dedicated to the development of systems that can reason from premises to a conclusion (still far from perfect)
- NLP is a potential solution to the “knowledge bottleneck”, by acquiring knowledge from texts, and perhaps also from conversations

Turing Test

- Alan Turing's 1949 paper "Computing Machinery and Intelligence" proposed the Turing test for determining whether AI had been achieved



Q: Please write me a sonnet on the topic of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (Pause about 30 seconds and then give answer as) 105621.

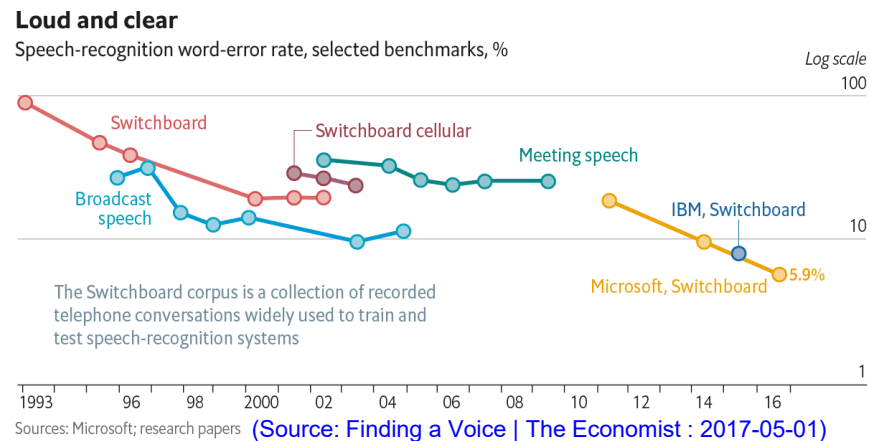
- Alan predicted at the end of 20 century a machine with 10 gigabytes of memory would have 30% chance of fooling a human interrogator after 5 minutes of questions
 - Does it come true?

Computer Science

- The discrete and recursive nature of natural language invites the application of theoretical ideas from computer science
- Linguists such as Chomsky and Montague have shown how formal language theory can help to explain the syntax and semantics of natural language
 - Theoretical models such as finite-state and pushdown automata are the basis for many practical natural language processing systems
- Algorithms for searching the combinatorial space of analyses of natural language utterances can be analyzed in terms of their computational complexity, and theoretically motivated approximations can sometimes be applied
- Techniques for parallel (GPU) computing and memory footprint control play pronounced roles
- ...

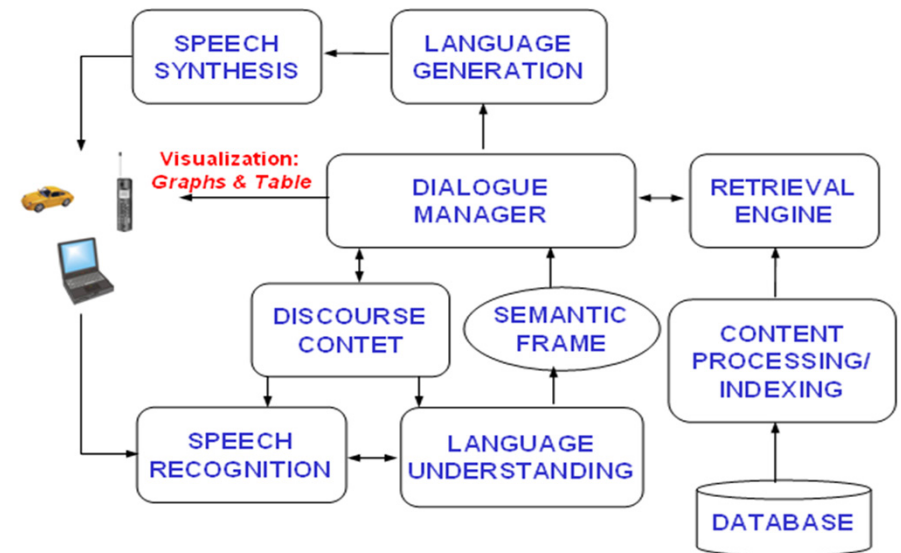
Speech Processing (1/2)

- Natural language is often communicated in spoken form, and automatic speech recognition (ASR) is the task of converting an audio signal to text
 - From one perspective, this is a signal processing problem, which might be viewed as a preprocessing step before NLP can be applied
- ASR is often integrated with text analysis, particularly with statistical language models (LM), which quantify the probability of a sequence of text
 - knowledge of the surrounding words influences perception and helps to correct for noise



Speech Processing (2/2)

- The broader field of speech processing includes the study of speech-based dialogue systems (conversational agents)



- Historically, speech processing has often been pursued in electrical engineering departments, while NLP has been the purview of computer scientists. For this reason, the extent of interaction between these two disciplines is less than it might otherwise be

Ethics

- As machine learning and artificial intelligence become increasingly ubiquitous, it is crucial to understand how their benefits, costs, and risks are distributed across different kinds of people
- NLP raises some particularly salient issues around ethics, fairness, and accountability
 - **Access:** Who is natural language processing designed to serve? For example, whose language is translated from, and whose language is translated to?
 - **Bias:** Does language technology learn to replicate **social biases** from text corpora, and does it reinforce these biases as seemingly **objective computational conclusions**?
 - **Labor:** Whose text and speech comprise the datasets that power NLP, and who performs the annotations? Are the benefits of this technology shared with all the people whose work makes it possible?
 - **Privacy and internet freedom:** What is the impact of large-scale text processing on the right to free and private communication? What is the potential role of NLP in regimes of censorship (審查制度) or surveillance (監控)?

Pragmatic and Discourse Knowledge

- Pragmatic Usage

Dave: *HAL, open the pod bay door.* → **HAL:** *No, I won't open the door.
I'm sorry (or I'm afraid). I can't.* [embellishment + refusal]

- Discourse (Dialogue) Knowledge

How many states were in the United States that year?

(This question talks about the year that Lincoln was born in the dialogue)

- **Dialogue** is a cooperative, two-way conversation. The goal is for participants to exchange information and build relationships with one another
- **Discourse** is a cooperative, one-way conversation. The goal to deliver information from the speaker/writer to the listeners/readers

<https://medium.com/@DavidWAngel>

Models and Algorithms (1/2)

- Classical Approaches

- Model: State machines, rule systems, logic (propositional/predicate/first-order logic), probabilistic models, vector space models, hidden Markov models etc.
- State space search algorithms: dynamic programming (or Viterbi search), A* search etc.
- Learning Algorithms: maximum likelihood (ML), root mean squared error (MSE), expectation-maximization (EM) etc.

- Deep Learning Approaches

- Models: Feed-forward neural networks (FFNN), recurrent neural networks (RNN), long short-term memory (LSTM), convolutional neural networks (CNN), Transformer, BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer)
- Learning Algorithms: gradient descent, contrastive learning, reinforcement learning, meta (transfer) learning, prompt-based learning, continual (lifelong) learning, federated learning etc.

Models and Algorithms (2/2)

- Exemplar Applications of Search Algorithms
 - **Speech recognition**: we search through a space of phone sequences for the correct word
 - **Machine translation**: we search through a space of translation hypotheses for the correct translation of a sentence into another language
 - **Parsing**: we search through a space of trees for the syntactic parse of an input sentence

Brief History of NLP & Speech Processing (1/2)

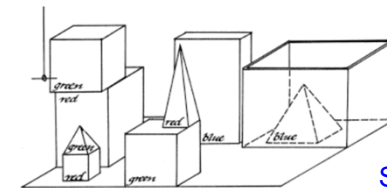
- Foundational Insights: 1940s and 1950s
 - Automation: algorithmic computation, neurons, Markov processes, finite state machines
 - Probabilistic or information-theoretic models: noisy channel and decoding, entropy
- The Two Camps: 1957–1970
 - By the end of the 1950s and the early 1960s, speech and language processing had split very cleanly into two paradigms: symbolic (e.g., Parsing with DP, the Birth of AI) and stochastic (e.g., Bayesian methods)
- Four Paradigms: 1970–1983
 - An explosion in research in stochastic paradigm, logic-based paradigm, discourse modeling paradigm, natural language understanding (NLU)



Audrey, Bell Laboratories, 1952



Dartmouth workshop, 1956



Microworlds
(Marvin Minsky et al.)

SHRDLU system, MIT, 1972
(Rule-based System)

Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I don't understand which pyramid you mean.
Person (changing their mind): Find a block which is taller than the one you are

The past serves the future! 歷史讀的越久，未來看的越遠。
(The more we know about the past, the better we foresee the future)

1. https://en.wikipedia.org/wiki/AI_winter
2. Kenneth Church, "A Pendulum Swung too Far," LiLT 2011
3. <https://adamdhalla.medium.com/the-illusion-of-intelligence-561ecd3a1f4e>

The First AI Winter
1974-1980

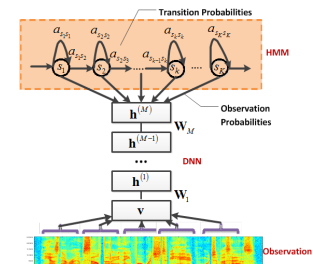
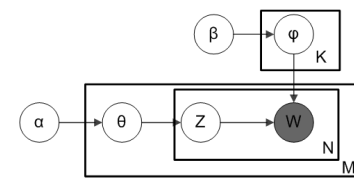
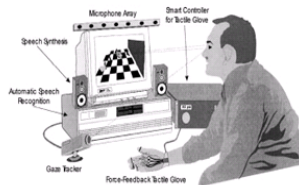


Briefing for US Vice President
Gerald Ford in 1973 on the junction-
grammar-based computer translation
model

Brief History of NLP & Speech Processing (2/2)

- Empiricism and Finite-State Models Redux: 1983–1993
 - The return of finite-state models and empiricism (viz. probabilistic models and data-driven approaches)
 - Quantitative metrics for model evaluation
- The Field Comes Together: 1994–1999
 - Probabilistic and data-driven models had become quite standard throughout most NLP tasks like parsing, POS tagging, reference resolution and discourse processing
 - The emergence of Web
- The Rise of Machine Learning: 2000–2012
 - Large amounts of spoken and written material, tight interplay with the ML community, unsupervised statistical approaches (e.g., topic models)
- Deep Learning Dominates NLP: 2012–

The Second AI Winter
1987-1993



Contemplation:
 1. Do speech scientists/engineers and linguists always see eye-to-eye?
 2. The criticisms of *PCM*: Pierce (1969), Chomsky (1957) and Minsky (1969)
 (They live in us ! They watch over everything we see)

Corpora and Toolkits

- The Linguistic Data Consortium (LDC), a non-profit consortium that creates and distributes corpora (<http://www ldc.upenn.edu/>)
- NLTK, the Natural Language Toolkit (Bird and Loper, 2004; <https://www.nltk.org/>), is a suite of Python modules and data for natural language processing, , together with a Natural Language Processing book based on the NLTK suite (<https://www.nltk.org/book/>)

Evaluating NLP-related Methods and Systems

- Granularities: black-box vs. glass-box evaluations
 - **Black-box** (overall/holistic) evaluation: Evaluate system performance without looking inside to see how it works
 - **Glass-box** evaluation: Look inside at the structure of a system, identify its various subcomponents, and then evaluate each one with appropriate tests
 - The problem with glass box evaluation is that it requires some consensus on what the various components of an NLP-based system should be (e.g., systems for oral proficiency assessments)
- Types: objective vs. subjective evaluations
 - **Objective evaluation**: evaluations generally based on statistical criteria like recall, precision, F1-score, and many others
 - **Subjective evaluation**: evaluations generally based on human subjects to measure or quantify performance and quality
- Corpora (datasets) used: **benchmark** or **in-house** corpora
- Evaluation results: **quantitative measures** or **qualitative comments**

Ablation Study?

Journals & Conferences

- Conferences
 - ACL: Association for Computational Linguistics
 - EMNLP: Empirical Methods in Natural Language Processing
 - NAACL: Annual Conference of the North American Chapter of the Association for Computational Linguistics
 - COLING: International Conference on Computational Linguistics
 - ICASSP: IEEE International Conference on Acoustics, Speech, and Signal Processing
 - INTERSPEECH: Annual Conference of the International Speech Communication Association
 - ...
- Journals
 - Computational Linguistics
 - Transactions of the Association for Computational Linguistics
 - IEEE/ACM Transactions on Audio, Speech and Language Processing
 - Computer Speech & Language
 - Natural Language Engineering
 - ...

ACLCLP: The Association for Computational Linguistics and Chinese Language Processing

- 中華民國計算語言學學會(www.aclclp.org.tw)

社團法人中華民國計算語言學學會
The Association for Computational Linguistics and Chinese Language Processing

[關於學會](#) | [會員專區](#) | [研究群與實驗室](#) | [出版品](#) | [語料庫](#) | [會議資訊](#) | [學生獎助](#) | [學術服務](#) | [網路資源](#)


[English]

最新公告

[SIGIR 2023榮獲經濟部國貿局「推動國際會議及展覽在臺辦理計畫」補助](#)

[IEEE ASRU 2023榮獲經濟部國貿局「推動國際會議及展覽在臺辦理計畫」補助](#)

[中文計算語言學期刊第二十七卷第二期全文上網](#)

[中文計算語言學期刊收錄於人文學核心期刊 \(THCI\)](#)

[SpeechBB \(Speech Bulletin Board\)](#)

[SpeechWiki](#)

1988年台灣舉辦了首次的計算語言學研討會 (ROCLING I)，由陳克健、黃居仁、李錫堅、林一鵬與蘇克毅等五位教授籌備發起。會中凝聚了發起學會的共識，並推舉最早投入中文電腦研究的謝清俊教授擔任籌備委員會主席及首屆理事長，在1990年正式成立中華民國計算語言學學會。目前本學會有200多名會員，包含159位終身會員及5個團體會員(包含3個永久團體會員)。

學會成立三十多年來主要宗旨一直在於推動計算語言學之研究、推行計算語言學之應用與發展、促進國內外中文計算語言學之研究與發展與聯繫國際有關組織並推動學術交流。每年舉辦的ROCLING年會是計算語言學學術界與業界發表與交流成果的盛會。本學會也定期舉辦計算語言學相關的工作坊及研習營以提高年輕研究人員對計算語言學的興趣與訓練。

學會在主辦國際會議以及接待國際學者方面，也具有豐富的經驗。近期協辦的國際會議包括：LPSS 2019, ISCSLP 2018, IJCNLP 2017, ISMIR 2014; 即將舉辦的國際會議有AAACL-IJCNLP 2022及SIGIR 2023。為讓國內語音處理相關領域之研究生對語音相關研究課題有更廣泛、更深入的認識，並增進相互交流，2006年起本學會也共同承辦了國科會整合型計畫 - 新世代自動語音辨識技術之研究寒暑期講習會。2017年本會也協辦中央研究院人工智慧跨域領袖營，期以短期密集研習及交流課程，讓參加者快速掌握人工智慧的最新發展及應用。並透過跨領域之間交流，共同凝聚台灣對於人工智慧的應有的認知及策略規劃。

在出版品方面，學會在1996年創刊的《中文計算語言學期刊》(International Journal of Computational Linguistics and Chinese Language Processing)是從事中文語言處理學者交流，國內外發表中文計算語言學論文的最重要的刊物之一。期刊收錄於ACL Anthology, THCI(科技部人文及社會科學核心期刊)及Linguistics Abstracts。

學會成立以來，一個重要的工作項目就是推動中文語言處理資源的分享。發行的語料庫，文字部分，包含了有標記一千萬詞的中研院漢語平衡語料庫、中文詞庫(八萬目詞)、中文句結構樹資料庫、中研院中文核心辭彙表、中文資訊檢索標竿測試集及增廣中文意見詞典及中文情感語意分析套件。語音方面，包含符合COCOSDA格式的國語語音語料庫(MAT, TCC)、台灣口音英語語料庫(EAT)、中研院現代漢語對話語音語料庫(MCDC)、北科大電子書語音語料庫(NTUT-AB01)、北科大教育電台廣播節目語音語料庫(NER-Trs、NER-Pro)及台文語音合成語料庫(TAI)。

Takeaways

- *NLP & its Relatives*: the crown jewel of artificial Intelligence (AI)?
 - https://www.youtube.com/watch?v=JxrSIY_Filk (Dr. Xuedong Huang)
 - Kenneth Church, “A Pendulum Swung too Far,” LiLT 2011
- *NLP & its Relatives* have now appealed to a wider audience but also stimulate and attract both young bloods and experts
- A comprehensive understanding of human language processing is anticipated to be helpful in building better ML models of NLU, NLG and among others
 - Imitate (copy) a solution that behaves the way people are accustomed to for an application domain?
 - Recall that [Frederick Jelinek](#) ever said
 - 1998: “*Every time I fire a linguist, my performance goes up*” (describing when he was with IBM)
 - 1995: “*Put language back into language modeling!*” (when he was with JHU, his last stop)