Natural Language Processing (NLP)

Dr. Nguyen Van Vinh
UET-VNU

Introductor

- Dr. Nguyễn Văn Vinh
- PhD (2009) in JAIST (Japan Advanced Institute of Science and Technology), Japan
- Research focus: AI, Natural Language Processing, Large Language Models
- Lecturer, AI4Life lab leaders - UET– Hanoi VNU
- Senior Data Scientist (Leader of LLM applied domain group), AI Center, Fsoft
- Design Leader of Machine Learning course for FUNiX – Online University (https://vnexpress.net/giao-duc/funix-ra-mat-chuong-trinh-machine-learning-3986011.html)
- Principle Investigator of National KC Project: "**Phát triển hệ thống dịch đa ngữ giữa tiếng Việt và một số ngôn ngữ khác**", 2020-2022.
- Principle Investigator of VNU Hanoi Project: "**Research and optimizing large language model and applying for Vietnamese question-answering system**", 2023-2025.
- Principle Investigator of GHTK-UET project: "**Name Entity Recogintion**" and "**Intents Detection**"
- Công bố hơn 40 công trình trên các tạp chí và hội nghị ISI/Scorpus trong lĩnh vực AI và NLP (bao gồm Conll (A), Paclic, EMNLP (A*)…)
- Teaching some ML, NLP Lectures in Samsung, BigData Vin, Viettel, Tp bank, …
- **Source:** https://vnexpress.net/chuyen-gia-ban-cach-xay-ung-dung-ai-cho-du-lich-4496592.html

Content

- Course Information
- Some achievements of NLP
- Overview of NLP
  - Linguistic levels of description
  - Why is NLP difficult?
- Conclusion

Course information

- **Course: Introduction to Natural Language Processing**

- **Instructor:** Dr. Nguyen Van Vinh, CS Department, Information Faculty.

  Email: vinhnv@vnu.edu.vn, vinhnv2000@gmail.com

  Tel: 0912263062

- **Teacher Assistant:** Master Tran Quoc Hung, Quach Manh Cuong, Talent Bachelor, CS Department, Information

Course information

- **Course web page:** https://portal.uet.vnu.edu.vn/uet-lms/ choose NLP course.
  - Up to date information
  - Lecture notes
  - Relevant dates, links, etc.
- **Foundations of Machine Learning:** If you already have basic machine learning and/or deep learning methods, the course will be easier; however it is possible to take this course without it.
- **Proficiency in Python:** programming assignments and projects will require use of Python, Numpy and PyTorch.
- **Grading:** 30% for (midterm + homeworks/assignments) +10% for attendence + 60% for final

Policy & Practical issues

- Encourage discussion but assignments must be your individual work
- Codes copied from books or other libraries but be explicitly acknowledged
- Sharing or copying codes is strictly prohibited.

Generative AI Policy

- You may use generative AI tools such as Co-Pilot and ChatGPT, as you would use a human collaborator. “This means that you may NOT directly ask generative AI tools for answers or copy solutions. You're required to acknowledge generative AI tools as collaborators and include a paragraph describing how you used the tool”.
- The use of generative AI tools to substantially complete an assignment or exam (e.g., by directly copying) is prohibited and will result in honor code violations. We will be checking students' assignments to enforce this policy.

Reference & Reading

- ***Lecture Slides***
- **Text books (**NLP is a rapid-moving field…**) :**

  1) *Speech and Language Processing*, Daniel Jurasky & James H. Martin, second edition, printed by Prentice Hall, 2009

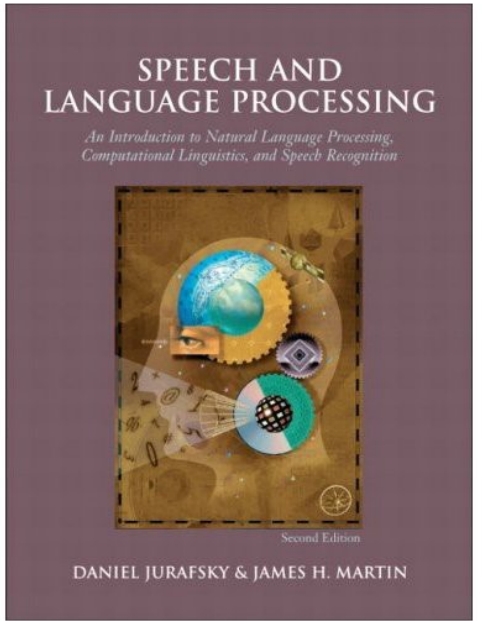
  (https://web.stanford.edu/~jurafsky/slp3/ 3rd ed. draft)

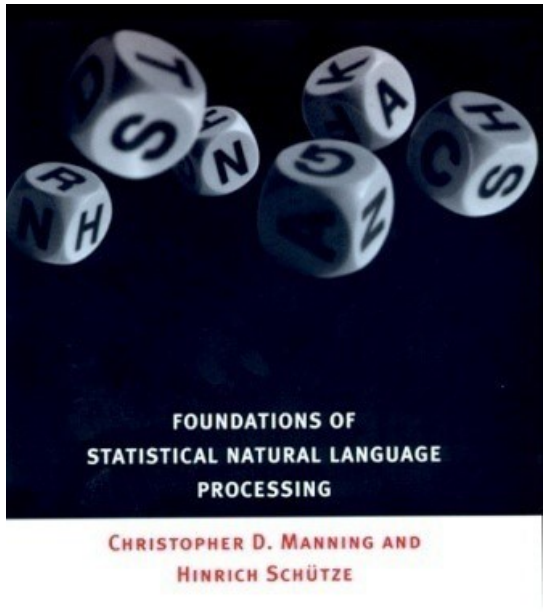
  2) *Natural Language Processing* , Eisenstein, 2018

  (*https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf*)

  3) *Foundation of Statistical Natural Language Processing*, Christopher D. Manning & Hinrich Schutze, 2001

- **We will link recommended papers and reading materials to the course schedule**





Reference (audio & video)

- https://web.stanford.edu/class/cs224n/
- https://youtu.be/rmVRLeJRkl4?si=vKO80qTInTDTQ1AB
- https://web.stanford.edu/class/cs224u/
- https://youtube.com/playlist?list=PLoROMvodv4rOwvldxftJTmoR3kRcWkJBp&si=xl29uLX8BIjx6NhI (Video lecturers)
- **Basic tools**
  - https://web.stanford.edu/class/cs224u/background.html

NLP market

- The Americas Natural Language Processing Market size was estimated at USD 5,985.64 million in 2021, is expected to reach USD 6,911.43 million in 2022, and is projected to grow at a CAGR of 16.44% to reach USD 14,924.84 million by 2027.
- The Asia-Pacific Natural Language Processing Market size was estimated at USD 4,842.69 million in 2021, is expected to reach USD 5,705.64 million in 2022, and is projected to grow at a CAGR of 17.22% to reach USD 12,563.54 million by 2027.
- The Europe, Middle East & Africa Natural Language Processing Market size was estimated at USD 5,248.39 million in 2021, is expected to reach USD 6,110.70 million in 2022, and is projected to grow at a CAGR of 16.76% to reach USD 13,303.33 million by 2027.

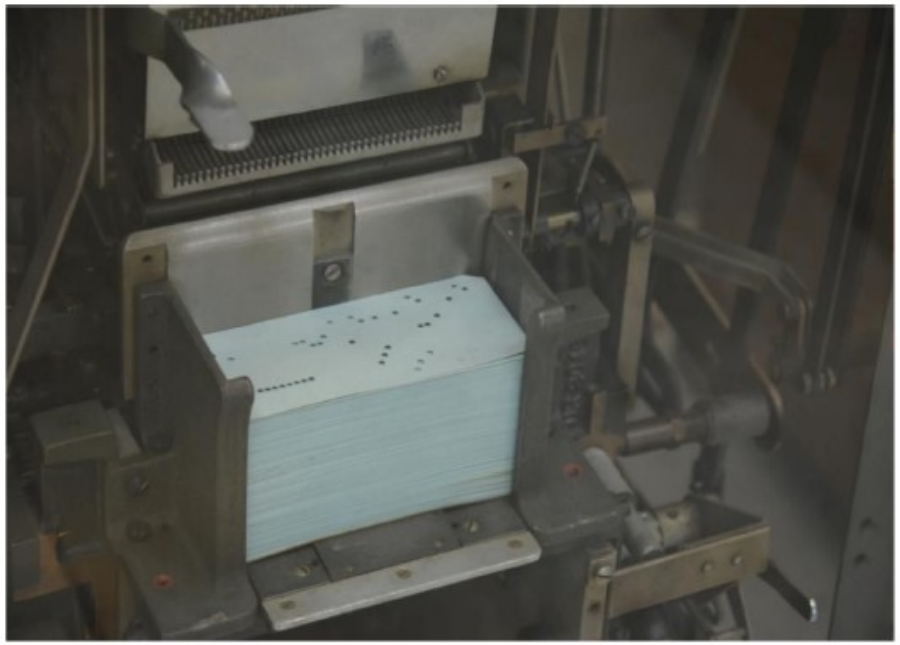
Source: https://www.businesswire.com/news/home/20220816005515/en/Natural-Language-Processing-NLP-Market-Intelligence-Report---Global-Forecast-to-2027---ResearchAndMarkets.com

NLP Market

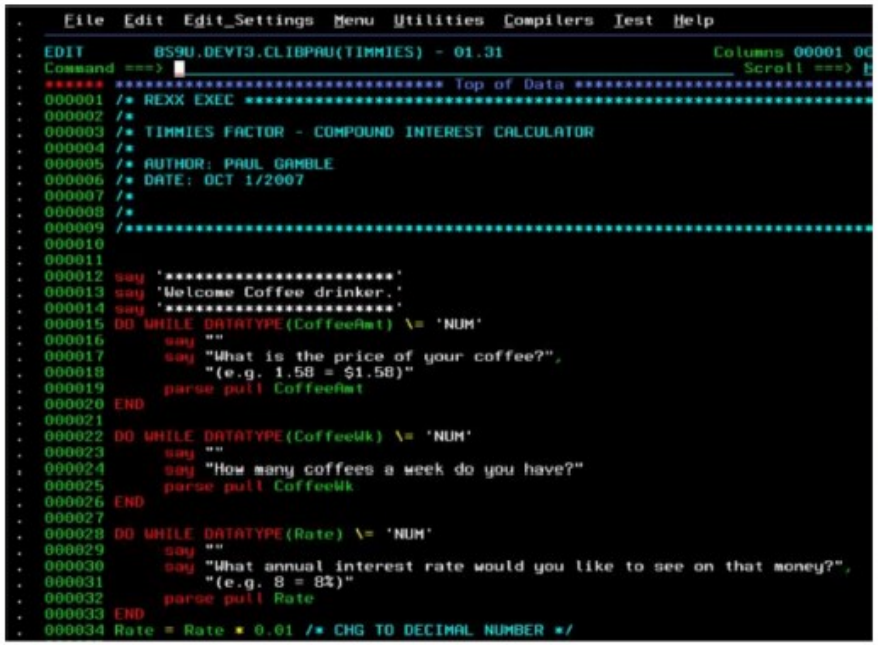
- The **global natural language processing market size** was worth **USD 13.5 billion in 2021**. It is estimated to reach an expected value of **USD 91 billion by 2030**, growing at a **CAGR of 27%** during the forecast period (2022–2030).
- **By deployment**, the global natural language processing market is segmented into On-premise and Cloud. The On-premise occupied the most significant market share and is expected to grow at a **CAGR of 23.5%** over the forecast period.
- **By organization size**, the global natural language processing market is segmented into Large Organizations and Small & Medium Organizations. Small & Medium Organizations occupy the largest market share, and it is expected to grow at a **CAGR of 26%** over the forecast period.
- **By processing type**, the global natural language processing market is segmented into Text, Speech/Voice, and Image. The Text processing segment occupies the most significant market share and is expected to grow at a **CAGR of 23.4%** over the forecast period.
- **By end-user**, the global natural language processing market is segmented into Education, BFSI, Healthcare, IT & Telecommunication, Retail, Manufacturing, Media & Entertainment, and Others. The IT & Telecommunication segment occupies the largest market share and is expected to grow at a **CAGR of 26.1%** over the forecast period.

Source: https://www.globenewswire.com/en/news-release/2022/08/11/2497065/0/en/Natural-Language-Processing-Market-Size-is-projected-to-reach-USD-91-Billion-by-2030-growing-at-a-CAGR-of-27-Straits-Research.html

Communication With Machines



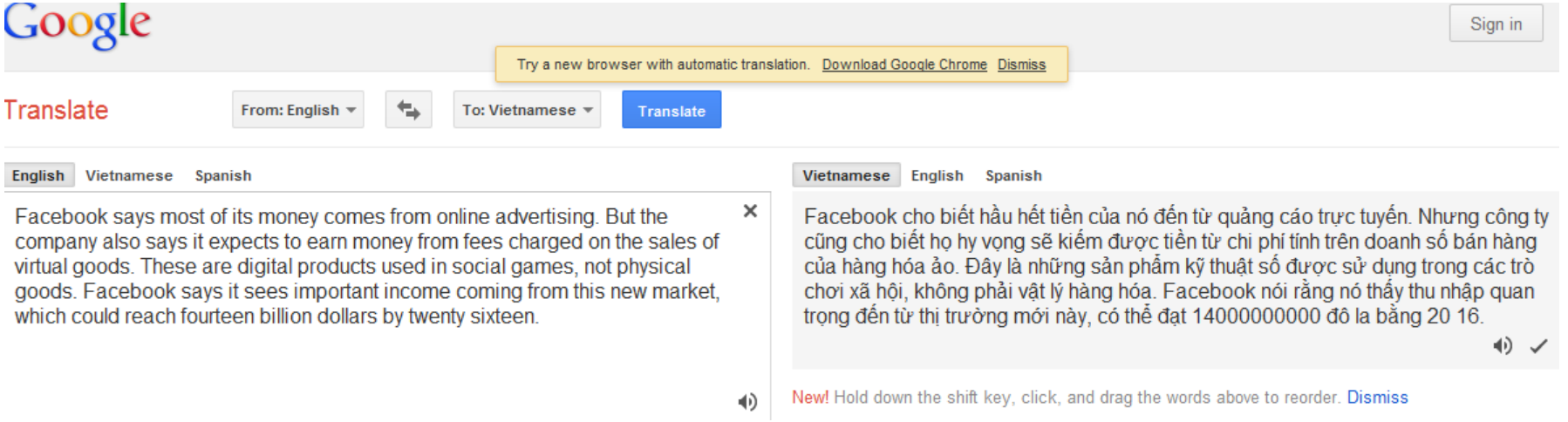
~50-70s

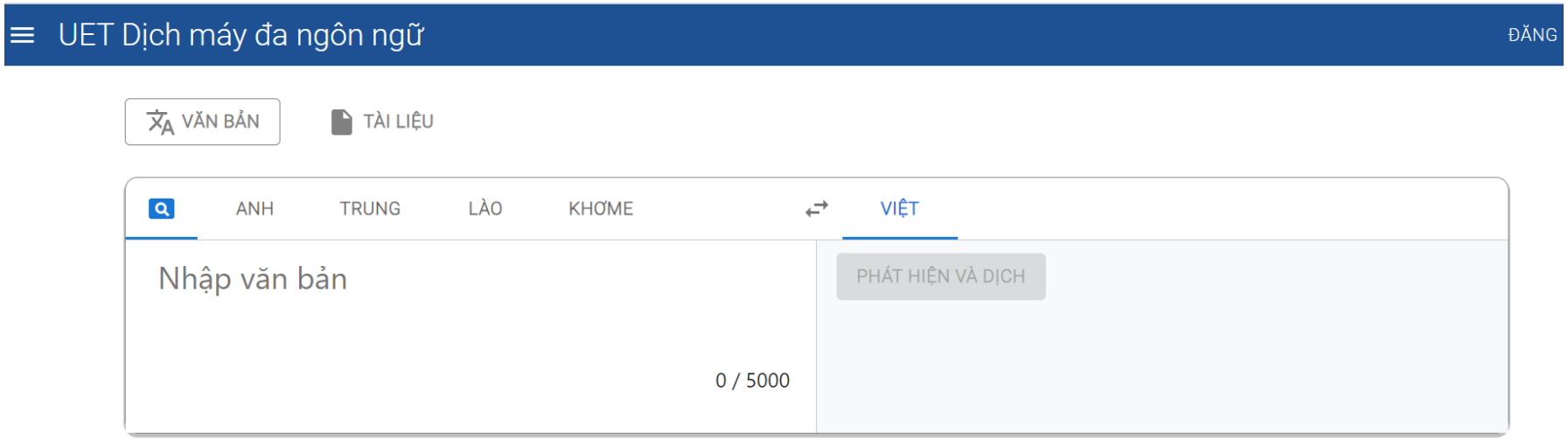


~80s

today

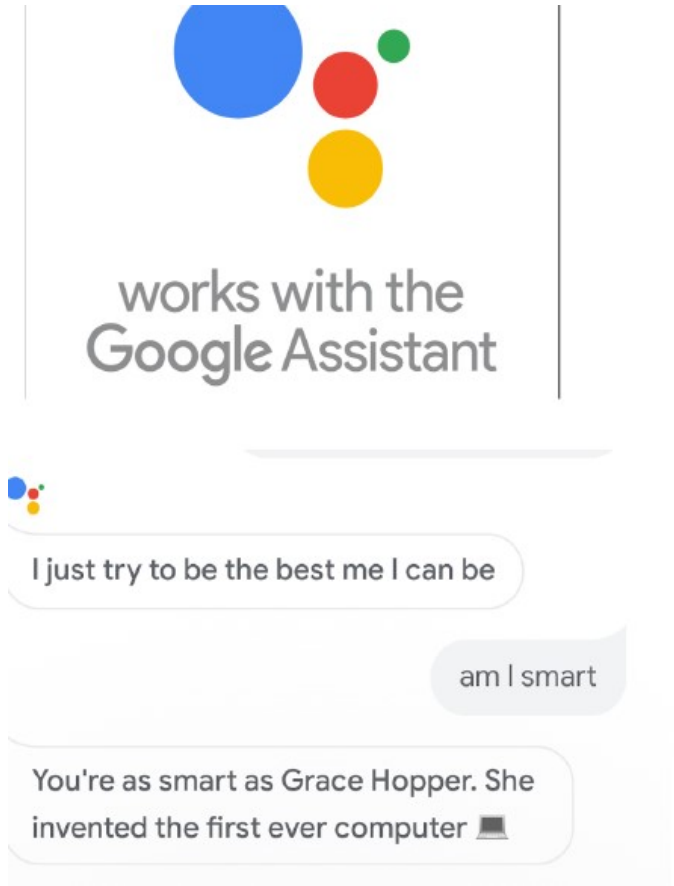
Google Translate & UET Translation





Virtual Assistant

- **Conversational agents contain:**
  - Speech recognition
  - Language analysis
  - Dialogue processing
  - Information retrieval
  - Text to speech
- **Google now, Alexa, Siri, Cortana, Watson, ChatGPT, …**



ChatGPT (AI hotest topic)

- **ChatGPT** (**Chat Generative Pre-trained Transformer**) is a chatbot launched by OpenAI in **November 2022**
- **Generative Pre-trained Transformer 3** (**GPT-3**) is an autoregressive language model that uses deep learning to produce human-like text. Given an initial text as prompt, it will produce text that continues the prompt.
- The architecture is a **standard transformer network** (with a few engineering tweaks) with the (back then) unprecedented size of 2048-token-long context and **175 billion parameters** (requiring 800 GB of storage).
- ChatGPT can fluently answer all the questions that users ask, any domains. Besides, ChatGPT can also write code, write poetry, compose music, write letters (documents), design and even fix errors in programming.
- ChatGPT has surpassed **10 million users just 40 days** after its official launch, breaking all previous records (100 million just 2 months).

Deepseek-R1 → Reasoning LLMs

- **January 23, 2025**, DeepSeek released R1, an open source reasoning model similar to OpenAI's ChatGPT-o1
- DeepSeek-V3 costs only 2.788M GPU hours for its full training ($5.576M- the rental price of the H800 GPU is $2 per GPU hour).
- Market Panic Over DeepSeek? Nvidia's $589 Billion DeepSeek Rout Is Largest in Market History
- Gemini Pro 2.5, Grok 5, ChatGPT 5 (deep thinking)
- Agentic LLMs

LLM-Powered Intelligent Agents

- Present in Large Language Models

NLP Careers: So hot!

- Industry (VinAI, Vin BigData, Viettel, Fpt, …)
- Government
- Academia

Source: https://daring.my.id/

Natural Language processing

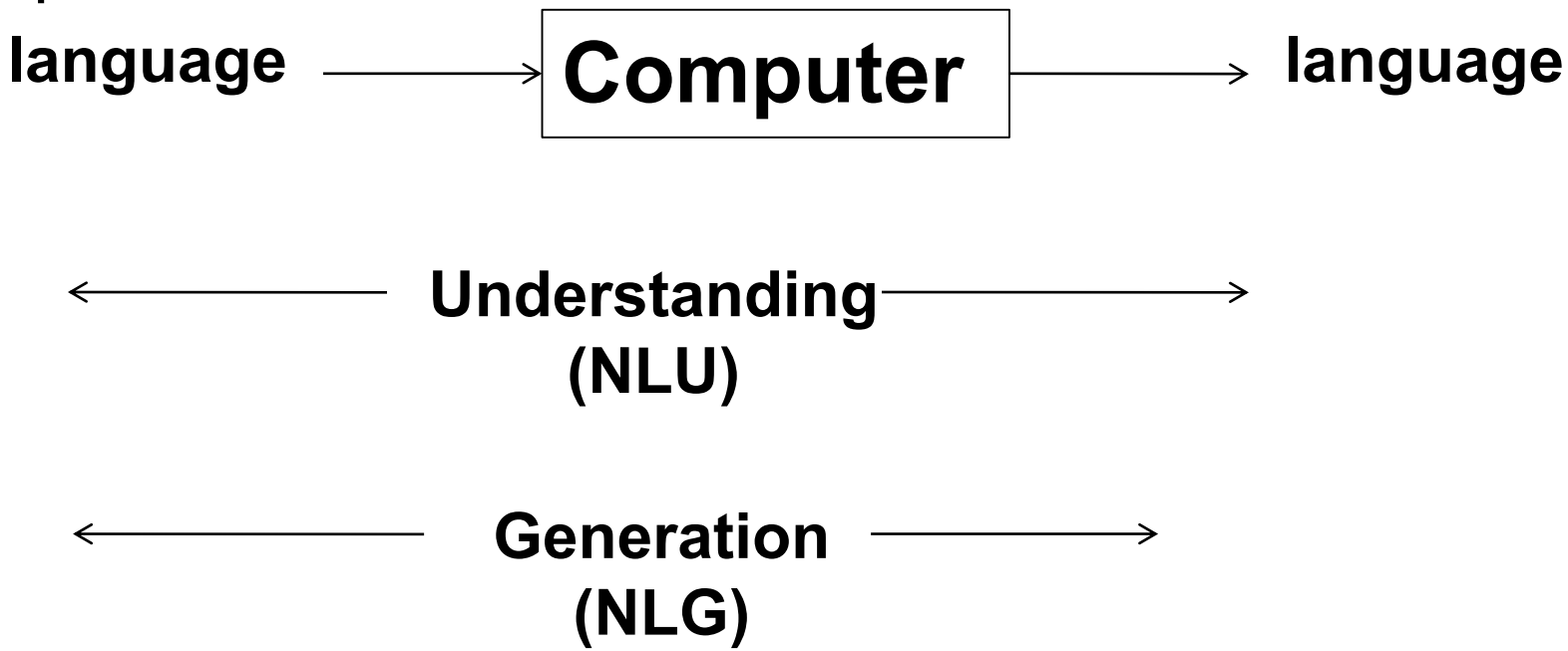
- NLP = building **computer programs** to analyze, understand and generate **human language - either spoken or written** (informal)
  - Not just string processing or keyword matching
- End systems that we want to build:
  - **Simple:** spelling correction, text categorization …
  - **Complex:** speech recognition, machine translation, information extraction, sentiment analysis, question answering …
  - **Unknown:** human-level comprehension (is this just NLP?)

What is NLP?

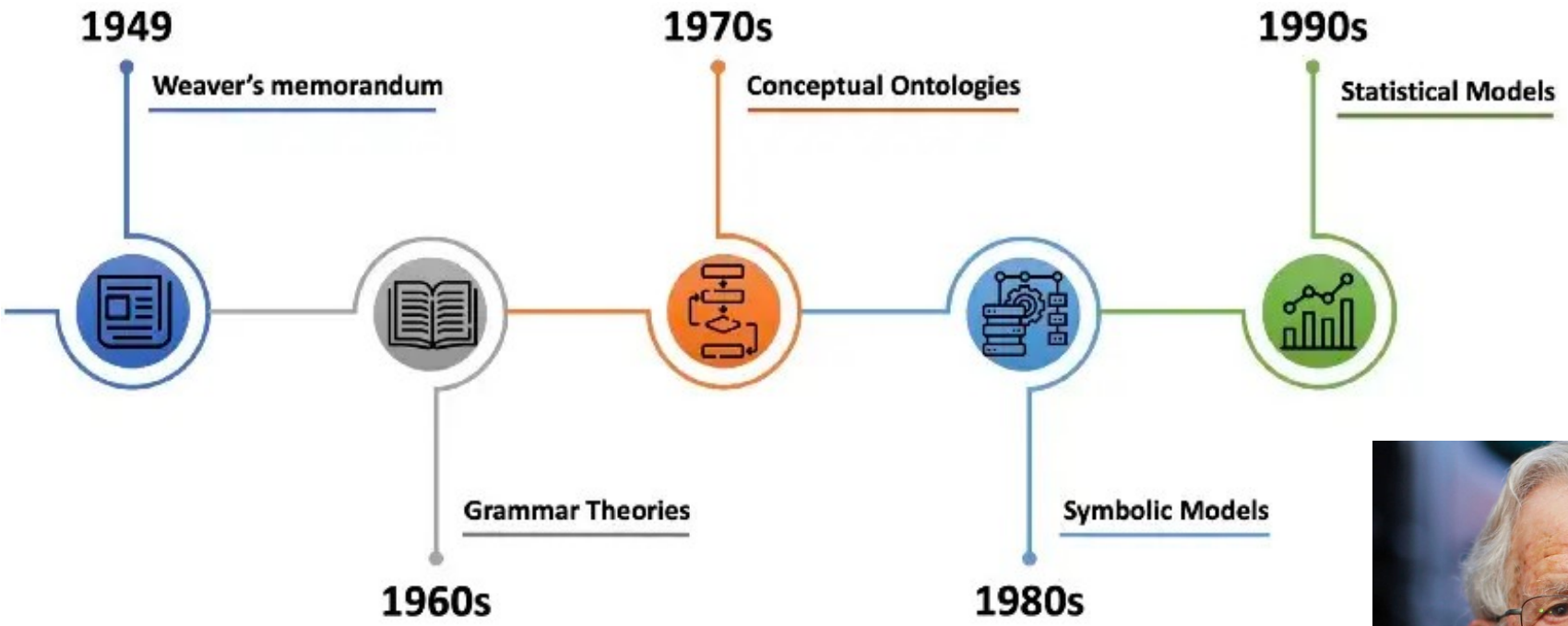
- **Natural language processing** (**NLP**) is a subfield of artificial intelligence and computational linguistics. It studies the problems of automated generation and understanding of natural human languages.
- **Natural-language-generation systems** convert information from computer databases into normal-sounding human language. **Natural-language-understanding systems** convert samples of human language into more formal representations that are easier for **computer** programs to manipulate.

What is Natural Language Processing?

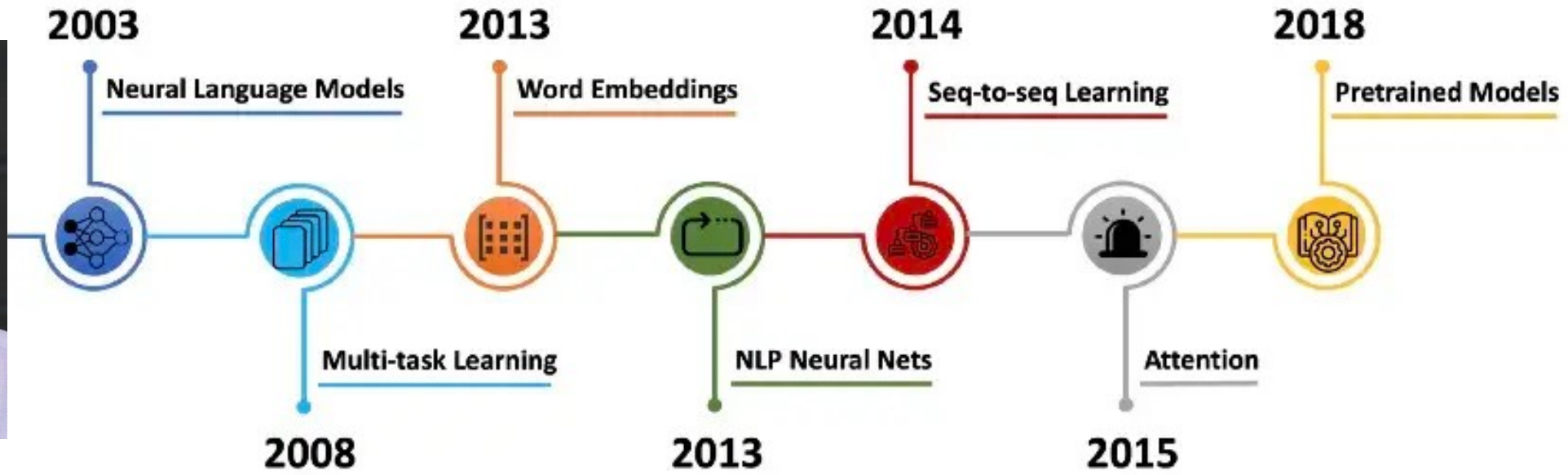
- Computers using natural language as input and/or output



A brief history of NLP

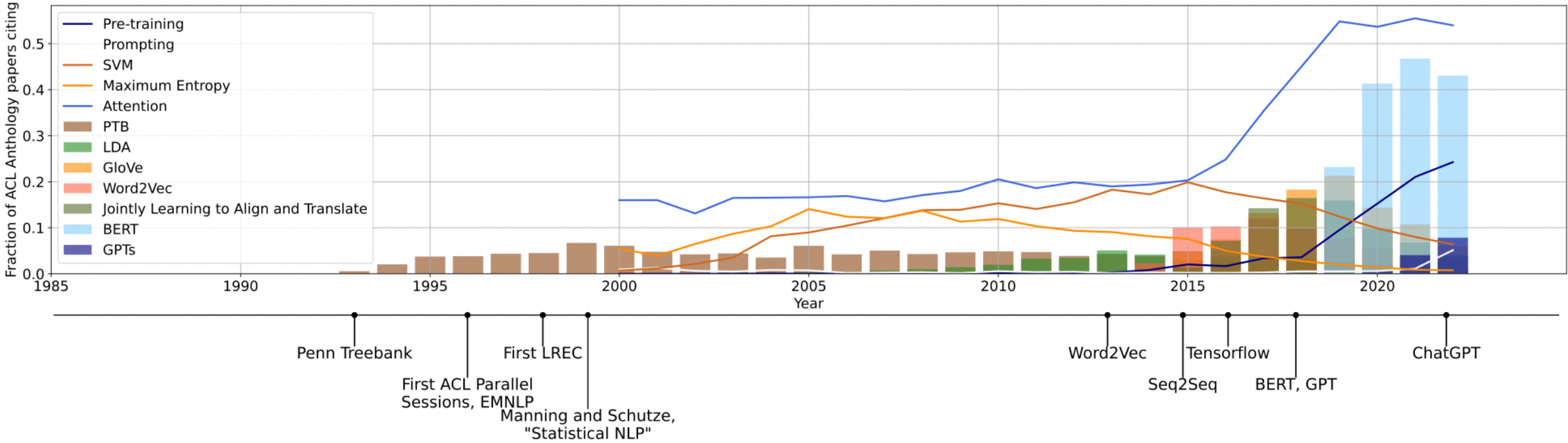


A brief history of NLP



- **2018:** BERT **2019:** T5, RoBERTa **2020:** GPT-3 **2022:** ChatGPT

History of NLP Research



Penn Treebank

First ACL Parallel Sessions, EMNLP

First LREC

Manning and Schutze, "Statistical NLP"

Word2Vec

Seq2Seq

Tensorflow

BERT, GPT

ChatGPT

Symbolic Methods Dominate

IBM Machine Translation Models

Money dries up for neural methods in the US

Metrics become important at DARPA

Parsing and MT dominate *CL conferences

Statistical NLP (including topic models, PGMs) dominates

Discussions of data scale solving everything

Early work in neural NLP

Neural revolution in NLP

Framework-based neural research

Word embeddings fuel SOTA chasing

Pre-train, fine-tune

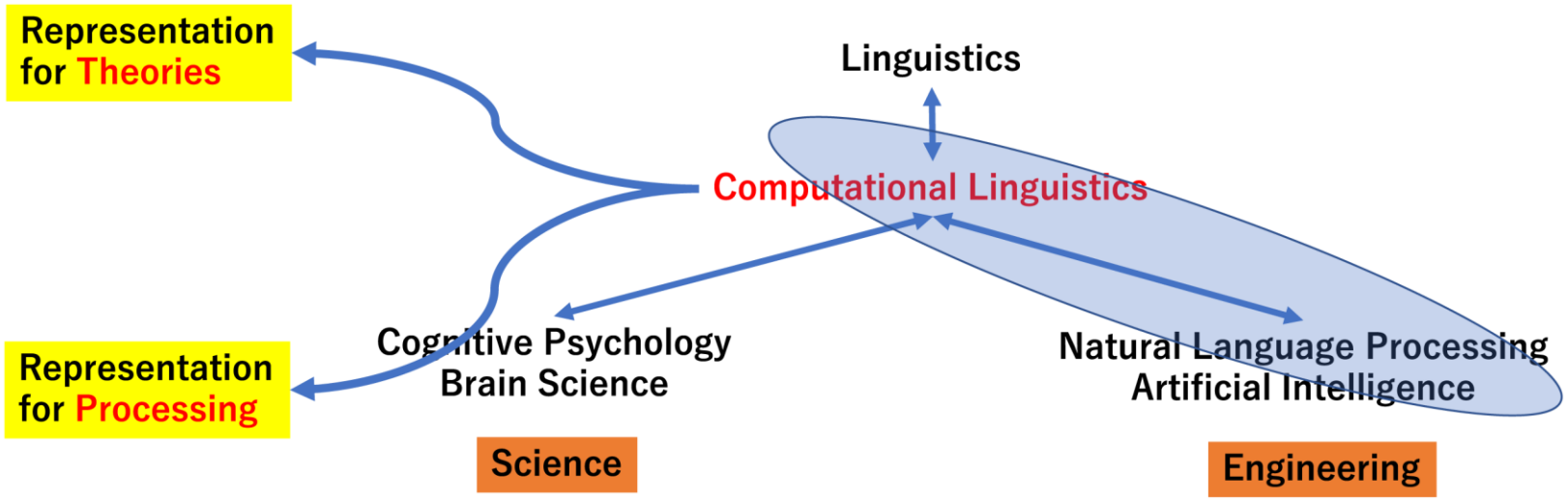
Generation over classification

Prompting

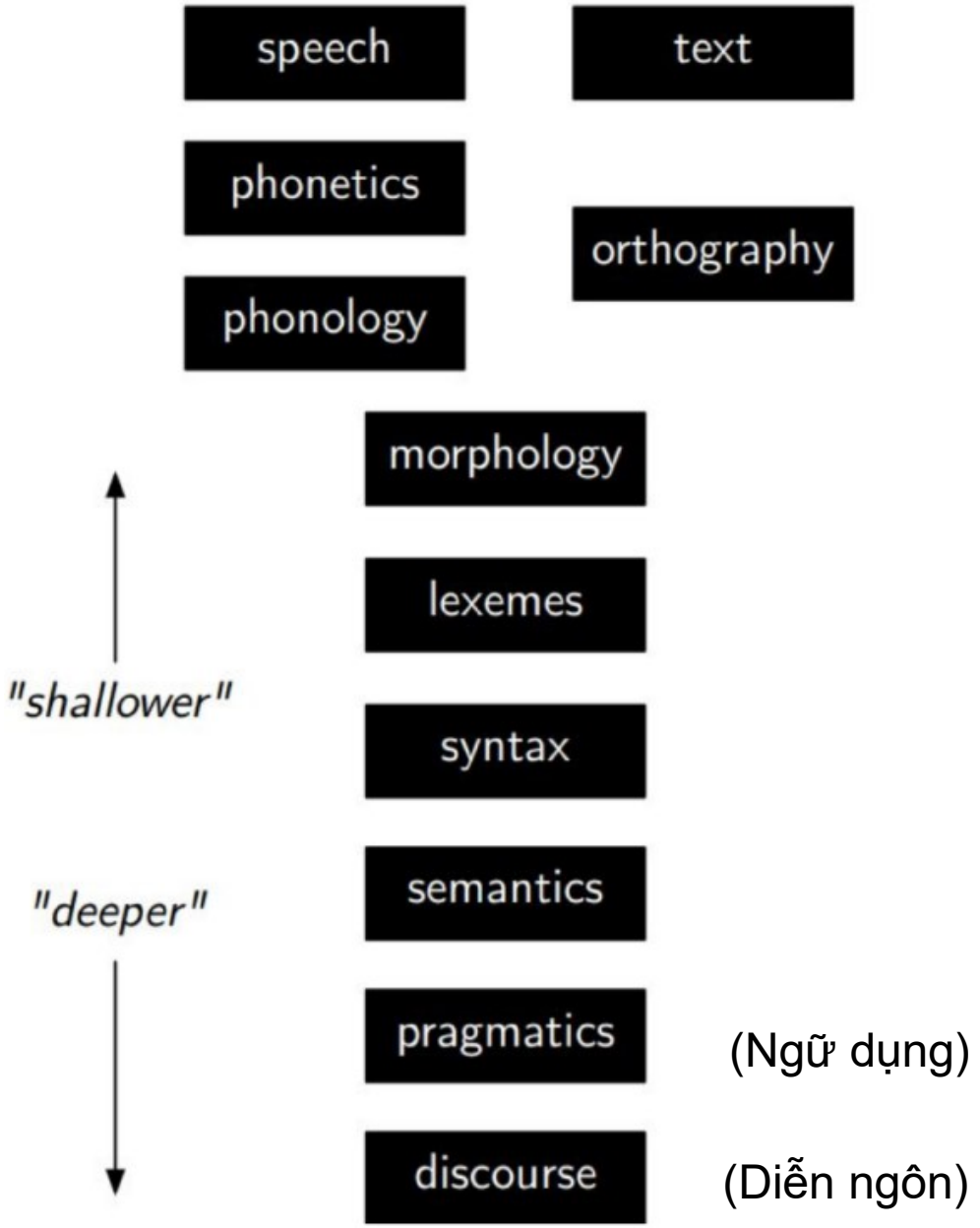
Natural language processing and Computational linguistics

- **Natural language processing (NLP)** develops methods for solving practical problems involving language:
  - Automatic speech recognition
  - Machine Translation
  - Sentiment Analysis
  - Information extraction from documents

- **Computational linguistics (CL)** focused on using technology to support/implement linguistics:
  - how do we understand language?
  - how do we produce language?
  - how do we learn language?



Level Of Linguistic Knowledge



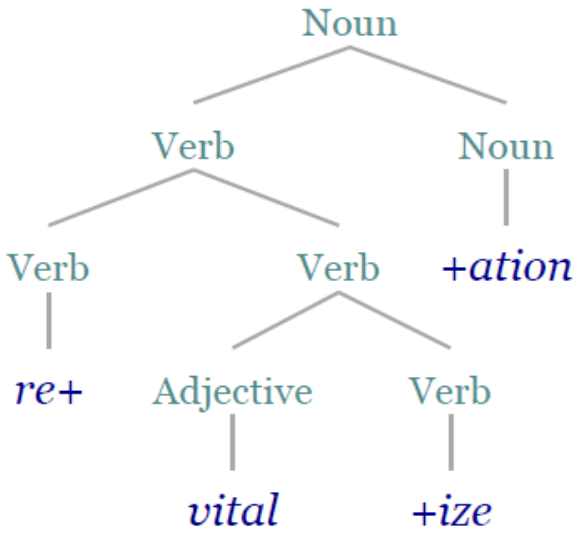
Phonetics and phonology

- *Phonetics (ngữ âm) studies the sounds of a language*
- *Phonology (âm vị học) studies the distributional properties of these sounds*

Morphology

- *Morphology studies the structure of words*
- Morphological derivation exhibits hierarchical structure

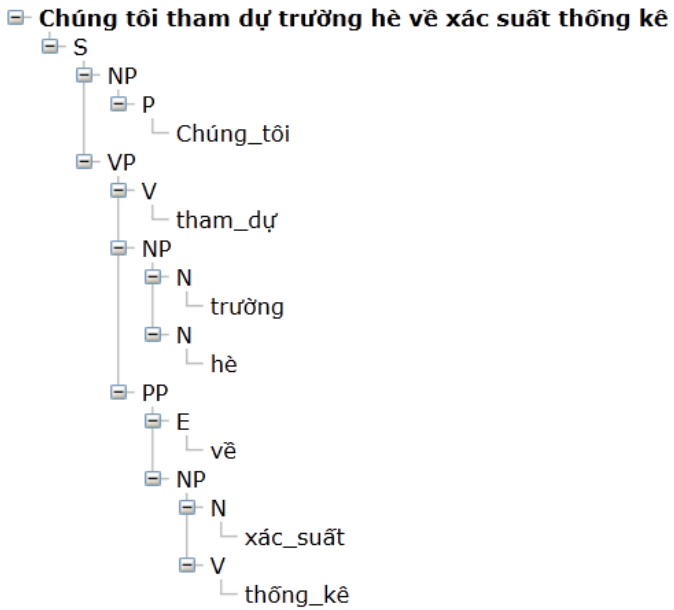
Example: re+vital+ize+ation



- The *suffix usually determines the syntactic category of the derived word*

Syntax

- Syntax studies the ways words combine to form phrases and sentences



- Syntactic parsing helps identify who did what to whom, a key step in understanding a sentence

Semantics and pragmatics

- Semantics studies the meaning of words, phrases and sentences

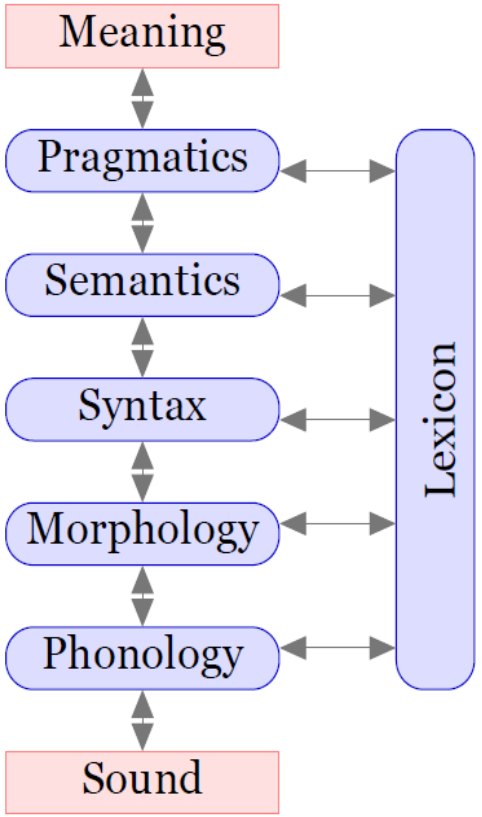
  Ex: I have a dinner in/for an hour

- Pragmatics (Ngữ dụng) studies how we use language to do things in the world

  Ex: Con vịt chạy đến Mary và liếm chân cô.

The lexicon

- A language has a lexicon, which lists for each morpheme
  - how it is pronounced (phonology),
  - its distributional properties (morphology and syntax),
  - what it means (semantics), and
  - its discourse properties (pragmatics)
- The lexicon interacts with all levels of linguistic representation



Natural Language Processing

- **Applications**
  - Machine Translation
  - Information Retrieval
  - Question Answering
  - Dialogue Systems
  - Information Extraction
  - Summarization
  - Sentiment Analysis
  - …
- **Core Technologies (NLP sub-problems)**
  - Language modeling
  - Part-of-speech tagging
  - Syntactic parsing
  - Named-entity recognition
  - Word sense disambiguation
  - Semantic role labeling
  - …

NLP lies at the intersection of computational linguistics and machine learning.

Statistical learning

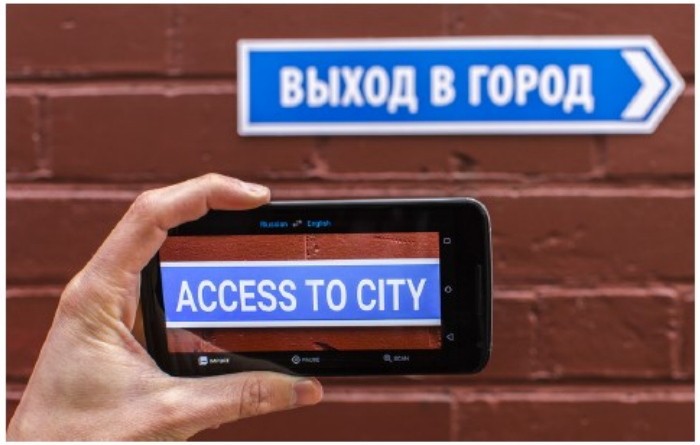
- Use of machine learning techniques in NLP
- Increase in computational capabilities
- Availability of electronic corpora

Unsupervised vs. supervised?

The era of deep learning

- Significant advances in core NLP technologies
- Essential ingredient: large-scale supervision, lots of compute
- Reduced manual effort - less/zero **feature engineering**

GPU TPU

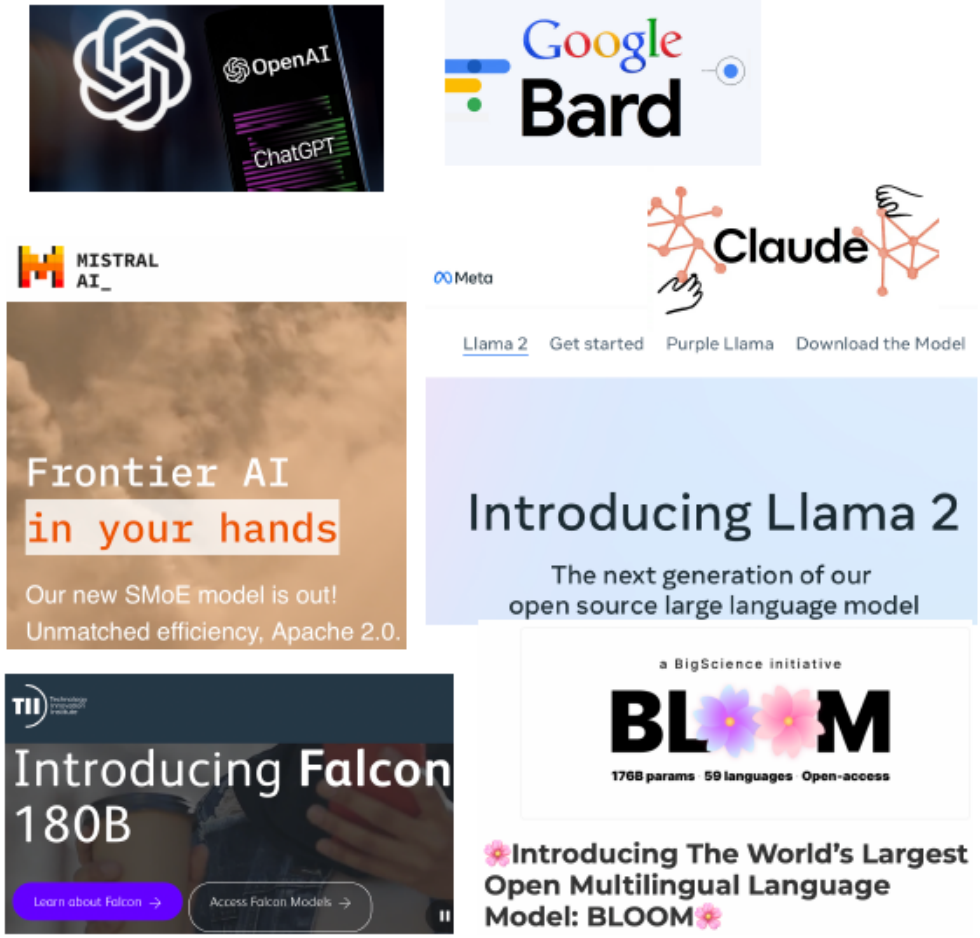


36M sentence pairs

Russian: Машинный перевод - это круто!

English: Machine translation is cool!

The New Era of Pre-training/LLMs

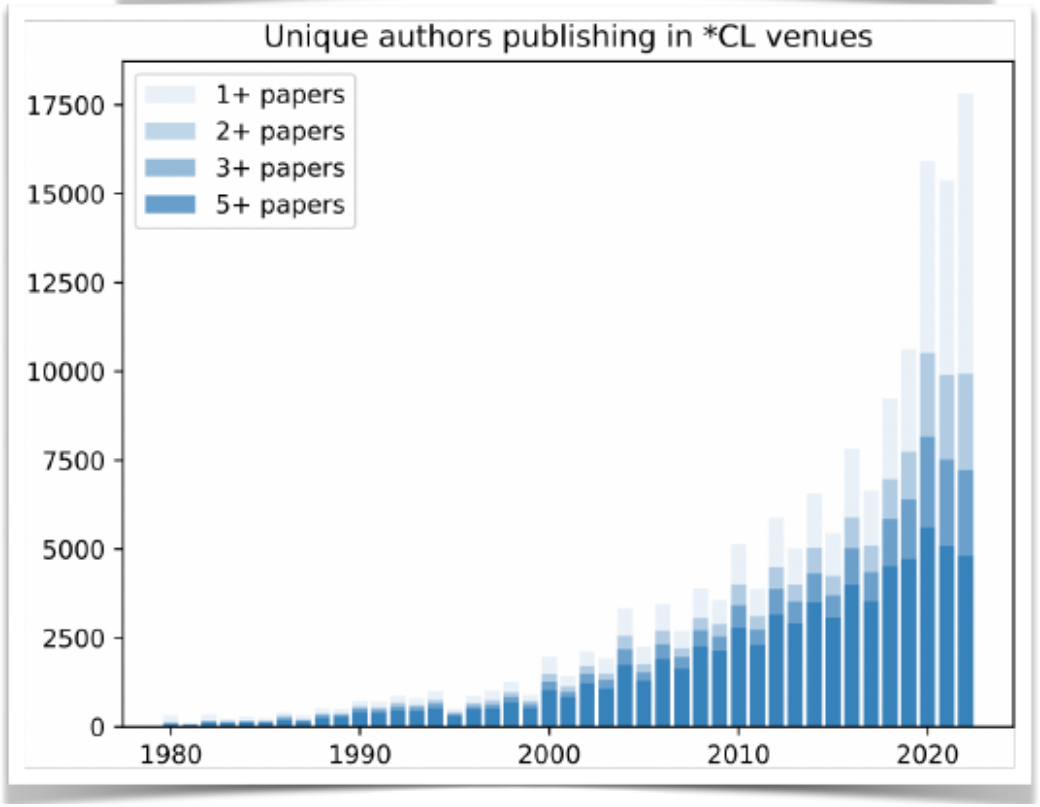


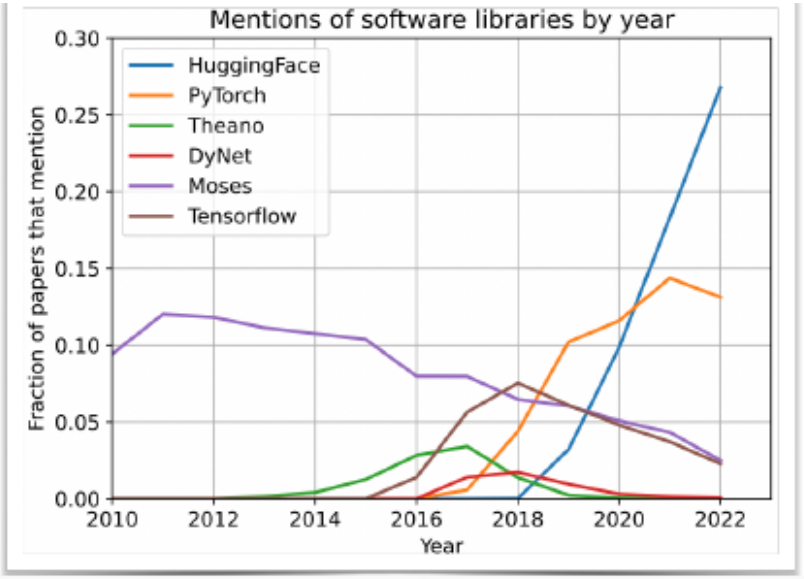
- Large language models (LLMs) are large-scale neural networks that are pre-trained on vast amounts of text data.
- They can potentially perform a wide range of language tasks such as recognizing, summarizing, translating, predicting, classifying, and generating texts.
- LLMs are built with the Transformer architecture.
- From several millions to hundreds of billions of parameters.

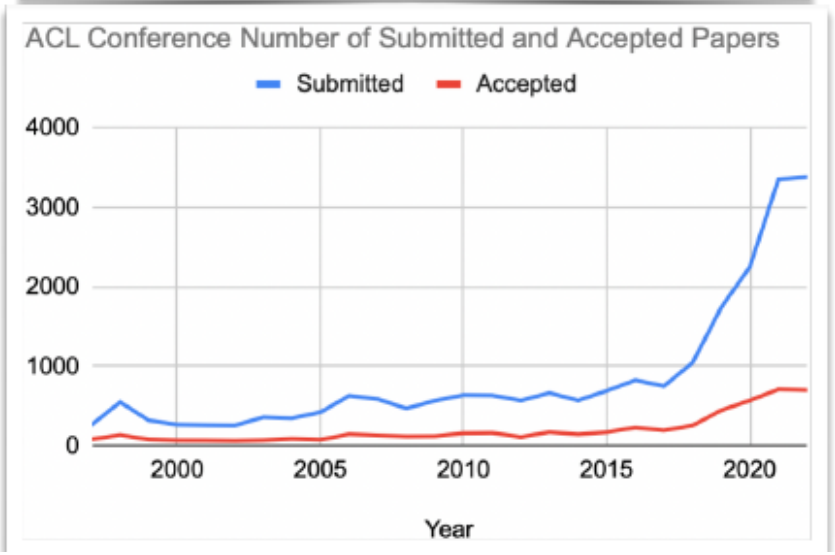
Bloom of NLP with LLMs

To Build Our Future, We Must Know Our Past:
Contextualizing Paradigm Shifts in Natural Language Processing

Sireesh Gururaja[1*] **Amanda Bertsch**[1*] **Clara Na**[1*]
David Gray Widder[2] **Emma Strubell**[1,3]
[1]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[2]Digital Life Initiative, Cornell Tech, Cornell University, New York City, NY, USA
[3]Allen Institute for Artificial Intelligence, Seattle, WA, USA
{sgururaj, abertsch, csna, estrubel}@cs.cmu.edu, david.g.widder@gmail.com







ChatGPT examples

DA

What NLP topics should I teach in 2023?

Here are some NLP topics that are popular and relevant in 2023:

1. Transformers (BERT, GPT, etc.)
2. Transfer Learning in NLP
3. Multi-lingual NLP
4. Attention Mechanisms
5. Adversarial Training in NLP
6. NLP for Question Answering
7. Sentiment Analysis
8. Text Generation (summarization, translation, etc.)
9. Named Entity Recognition
10. Dialogue Systems

Note: The field of NLP is constantly evolving, so it's important to stay up-to-date with recent advances and adjust your teaching content accordingly.

Why is NLP difficult?

- Ambiguity
- Sparsity
- Training data for supervised learning is often not available

  → Unsupervised/semi-supervised techniques for training from available data
- Algorithmic challenges
  - vocabulary can be large (e.g., 50K words)
  - data sets are often large (GB or TB)

Ambiguity ???

“At last, a computer that understands you like your mother”

“Ông già đi nhanh quá”

Ambiguity

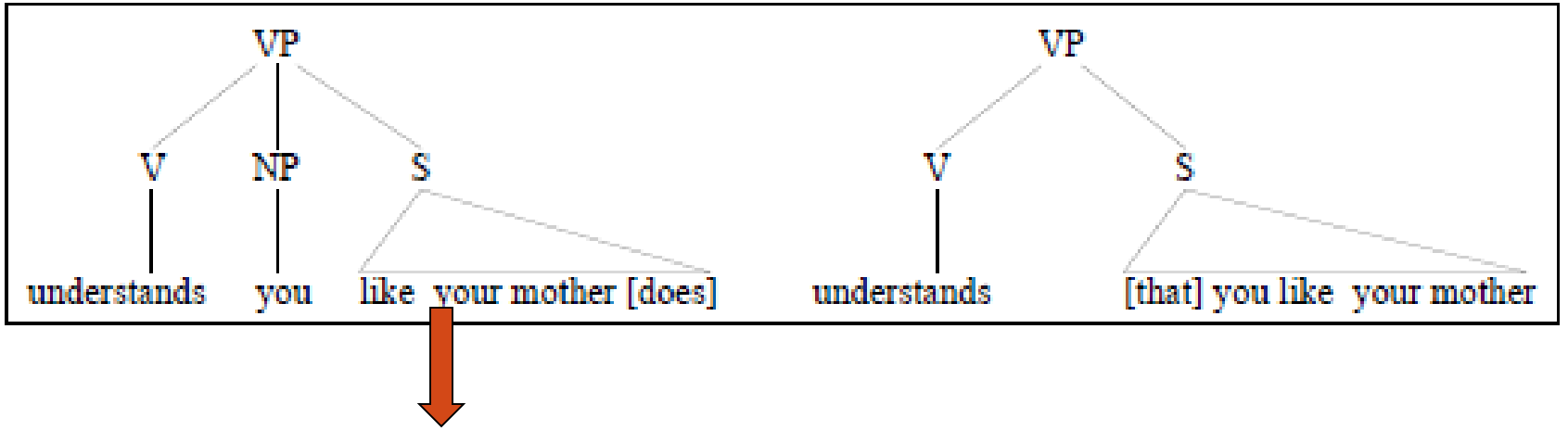
- “**At last, a computer that understands you like your mother**”
- It understands you as well as your mother understands you
- It understands (that) you like your mother
- It understands you as well as it understands your mother

Ambiguity at Many Levels

- At the acoustic level (speech recognition):
- "… a computer that understands you like your mother"
- "… a computer that understands you lie cured mother"

Ambiguity at Many Levels

- At the **syntactic** level:



Different structures lead to different interpretations

Ambiguity at Many Levels

- At the **semantic** (meaning) level:
  - Two definitions of “bank”
    - an organization where people and businesses can invest or borrow money, change it to foreign money, etc., or a building where these services are offered
    - sloping raised land, especially along the sides of a river
- This is an instance of **word sense ambiguity**

More Word Sense Ambiguity

- At the semantic (meaning) level:
  - They put money in the bank
  - I saw her duck with a telescope

Dealing with Ambiguity

- **How can we model ambiguity?**
  - Non-probabilistic methods (CKY parsers for syntax) return all possible analyses
  - Probabilistic models (HMMs for POS tagging, PCFGs for syntax) and algorithms (Viterbi, probabilistic CKY) return **the best possible analyses**, i.e., the most probable one.
- But the “best” analysis is only good if our probabilities are accurate. Where do they come from?

Corpora

- A corpus is a collection of text
  - Often annotated in some way
  - Sometimes just lots of text
- Examples
  - Penn Treebank: 1M words of parsed WSJ
  - Canadian Hansards: 10M+ words of French/English sentences
  - Yelp reviews
  - VLSP Corpus (Vietnamese)

Sparsity

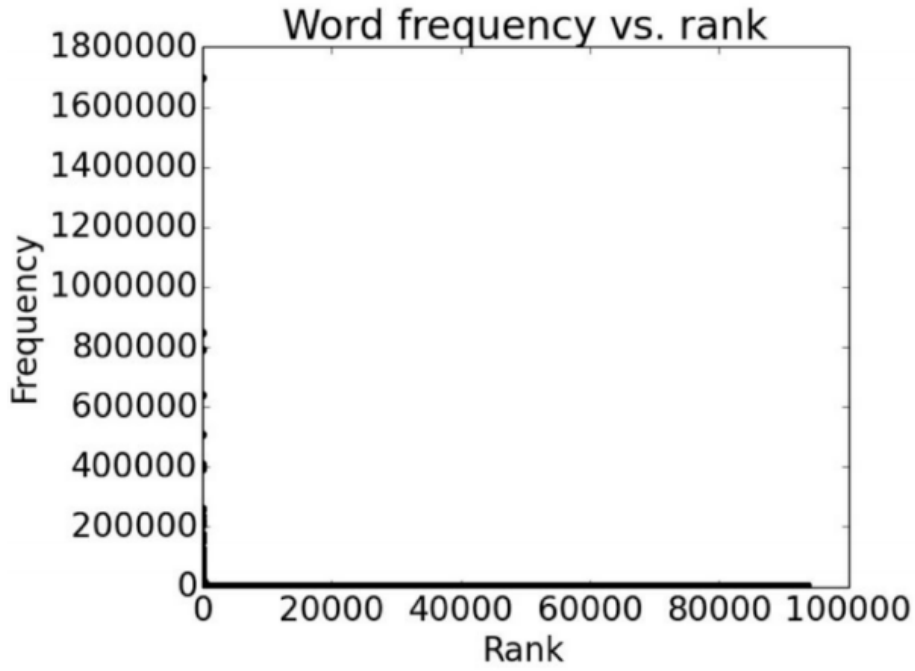
- Sparse data due to Zipf's Law
- Example: the frequency of different words in a large text corpus

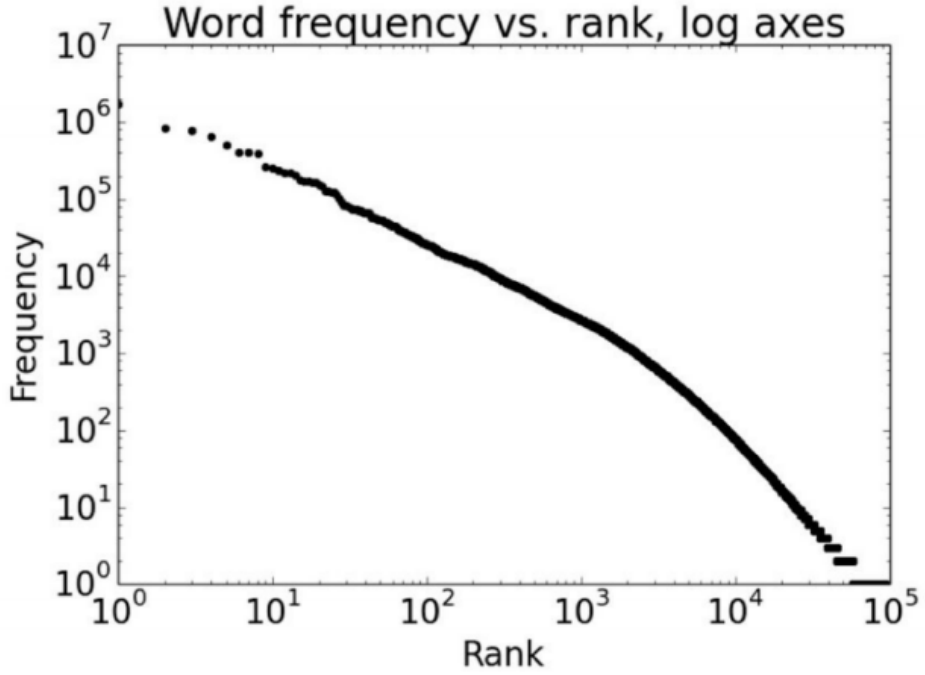
any word	
Frequency	Token
1,698,599	the
849,256	of
793,731	to
640,257	and
508,560	in
407,638	that
400,467	is
394,778	a
263,040	I

nouns	
Frequency	Token
124,598	European
104,325	Mr
92,195	Commission
66,781	President
62,867	Parliament
57,804	Union
53,683	report
53,547	Council
45,842	States

Sparsity

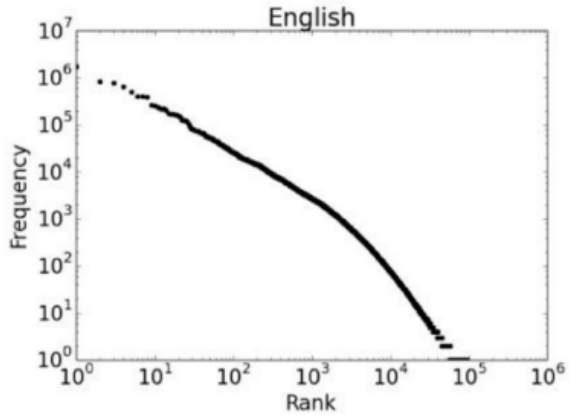
- Order words by frequency. What is the frequency of nth ranked word?

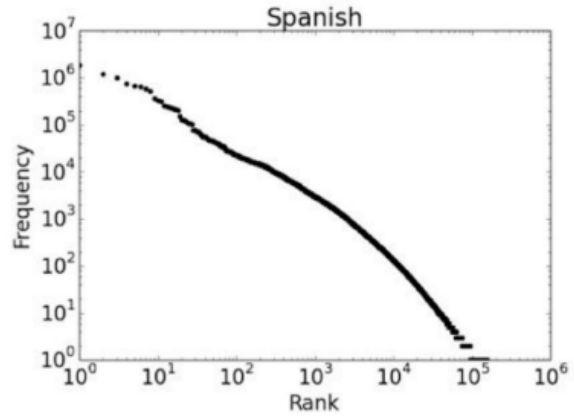


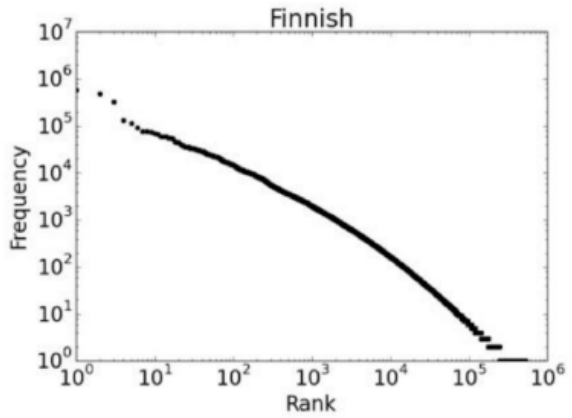


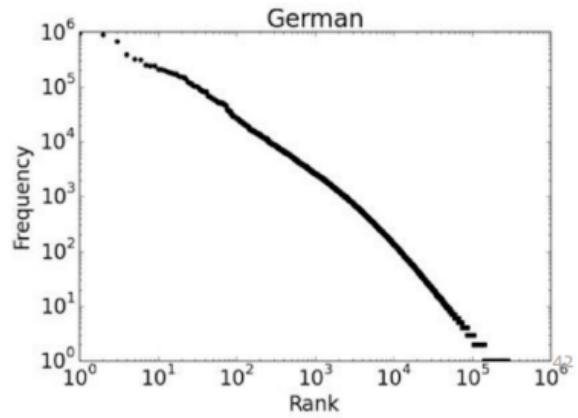
Sparsity

- Regardless of how large our corpus is, there will be a lot of infrequent words
- This means we need to find clever ways to estimate probabilities for things we have rarely or never seen









Fields with Connections to NLP

- Machine learning
- Linguistics (including psycho-, socio-, descriptive, and theoretical)
- Cognitive science
- Information theory
- Logic
- Data science
- Political science
- Psychology
- Economics
- Education

Today's Applications

- Conversational agents
- Information extraction and question answering
- Machine translation
- Summarization
- Opinion and sentiment analysis
- Social media analysis
- Visual understanding
- Essay evaluation
- Mining legal, medical, or scholarly literature
- ChatGPT

What is this course?

- **Methods**
  - Introduction to cutting-edge research in Machine Learning for NLP?
  - Feature Egineering for NLP: Traditional ML methods
  - Deep neural networks for NLP: LSTM, CNN, Seq2seq, Transformer, Pretraining models/LLMs

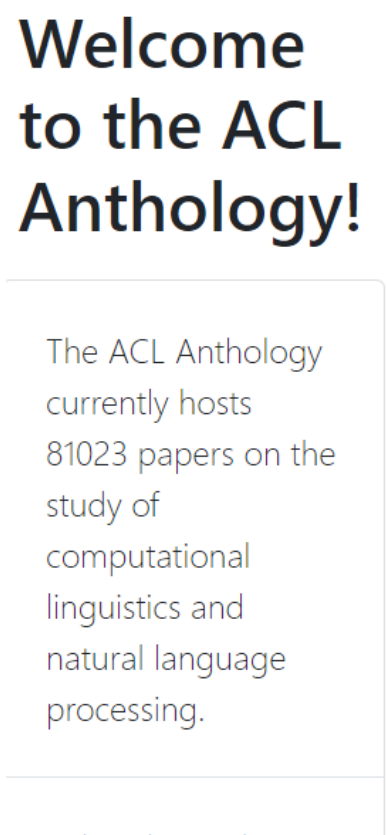
- **NLP in Applications**
  - Chatbot, Machine Translation, Text Summarization, …

Goals of this Course

- **[Foundation of NLP] and Technology Trends of NLP (rapid evolving field!) ?**
- **Learn about the problems and possibilities of natural language analysis:**
  - What are the major issues?
  - What are the major solutions?
- **At the end you should:**
  - Agree that language is difficult, interesting and important
  - Be able to assess language problems
    - Know which solutions to apply when, and how
  - Be able to use software to tackle some NLP language tasks
  - Know language resources
  - Be able to read papers in the field

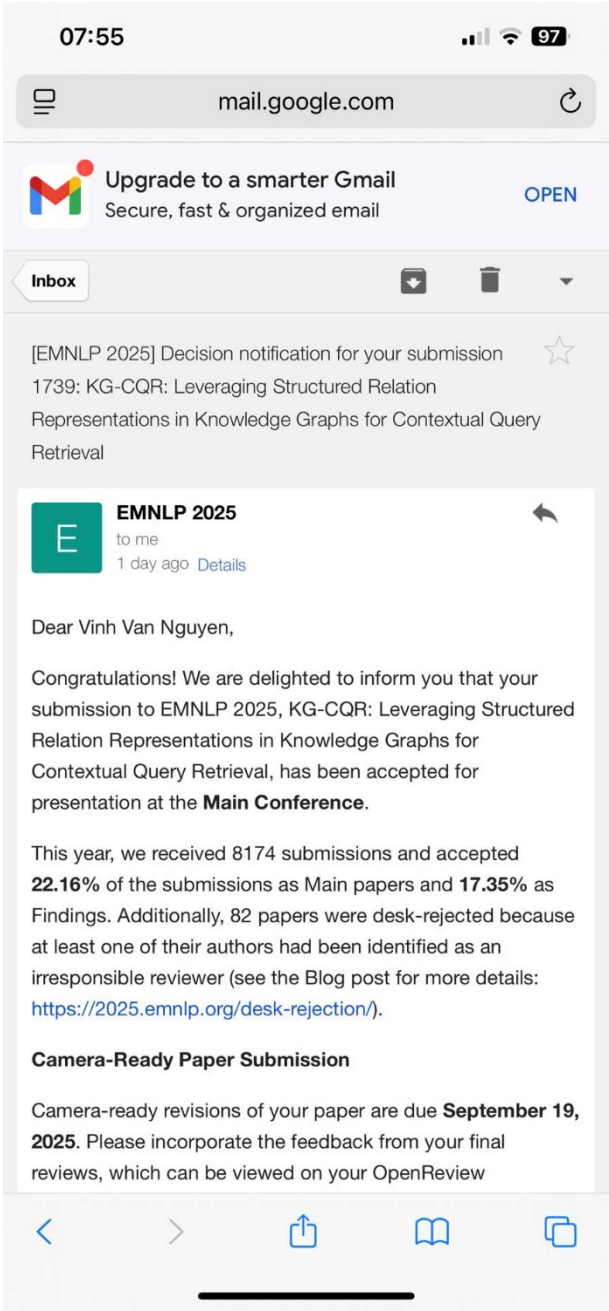
Journal and Conference in NLP

- https://aclanthology.org/



ACL Events

Venue	2022 – 2020			2019 – 2010										2009 – 2000										1999 –				
AACL	22		20																									
ACL	22	21	20	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00	99	98	97	96	95
ANLP																							00			97		
CL	22	21	20	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00	99	98	97	96	95
CoNLL		21	20	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00	99	98	97		
EACL		21				17			14		12			09			06			03					99		97		95
EMNLP		21	20	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00	99	98	97	96	
Findings	22	21	20																									
IWSLT	22	21	20	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04									
NAACL	22	21		19	18		16	15		13	12		10	09		07	06		04	03		01	00					
SemEval	22	21	20	19	18	17	16	15	14	13	12		10			07			04			01			98			


07:55
97
mail.google.com
Upgrade to a smarter Gmail
Secure, fast & organized email
OPEN
Inbox
[EMNLP 2025] Decision notification for your submission 1739: KG-CQR: Leveraging Structured Relation Representations in Knowledge Graphs for Contextual Query Retrieval
E
EMNLP 2025
to me
1 day ago Details
Dear Vinh Van Nguyen,
Congratulations! We are delighted to inform you that your submission to EMNLP 2025, KG-CQR: Leveraging Structured Relation Representations in Knowledge Graphs for Contextual Query Retrieval, has been accepted for presentation at the Main Conference.
This year, we received 8174 submissions and accepted 22.16% of the submissions as Main papers and 17.35% as Findings. Additionally, 82 papers were desk-rejected because at least one of their authors had been identified as an irresponsible reviewer (see the Blog post for more details: https://2025.emnlp.org/desk-rejection/).
Camera-Ready Paper Submission
Camera-ready revisions of your paper are due September 19, 2025. Please incorporate the feedback from your final reviews, which can be viewed on your OpenReview

Conclusion

- **Computational linguistics and Natural language processing:**
  - were originally inspired by linguistics,
  - but now they are almost applications of machine learning and statistics
- **We solve these problems using standard methods from machine learning:**
  - Feature Engineering: SVM, CRFs and PCFGs
  - End2end: **Deep Learning**
  - Multi-tasks/Modality-tasks: **Pretraining/LLMs (**2020 –now**)**

References

- Slides of NLP course from CMU, Stanford, Toronto University, …
- Some Tutorials of NLP