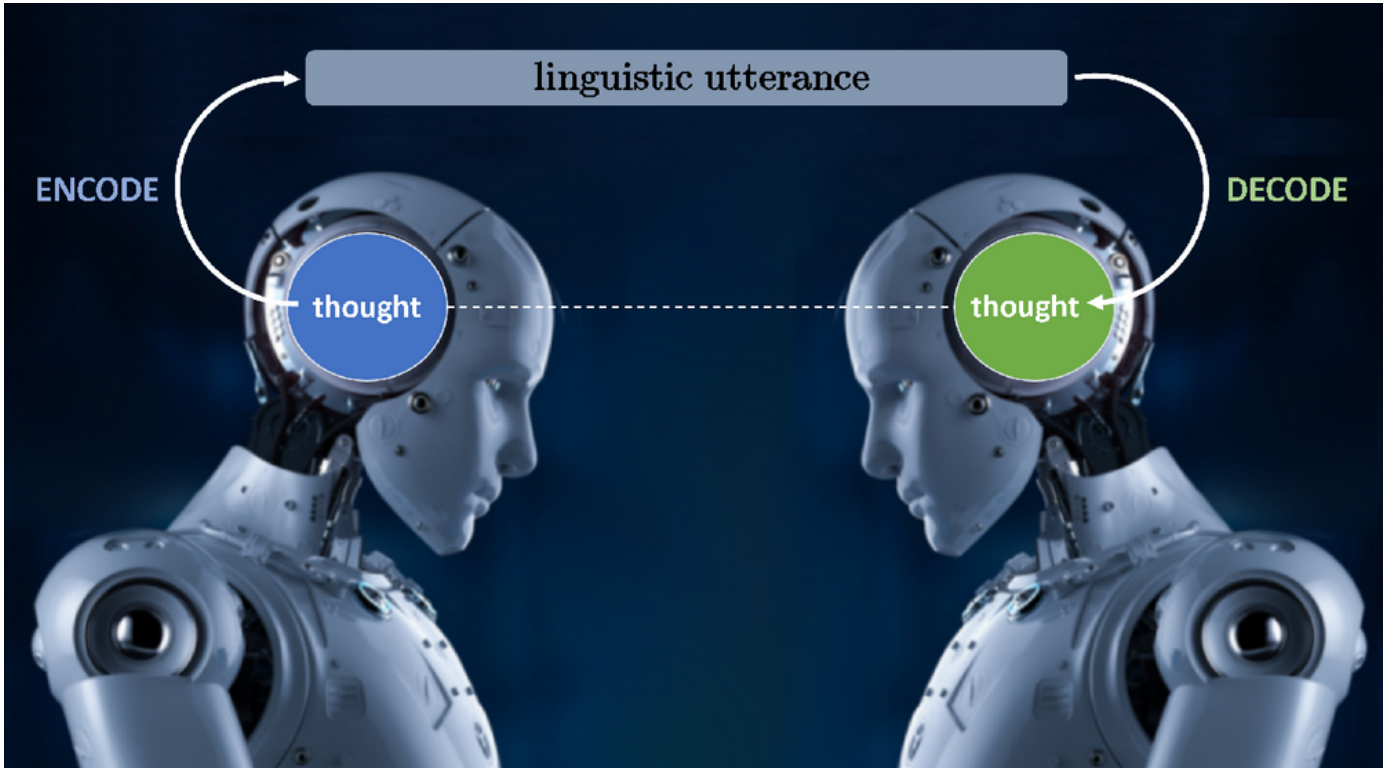
Natural Language Processing

Lecture 15 : Discussion & Conclusions

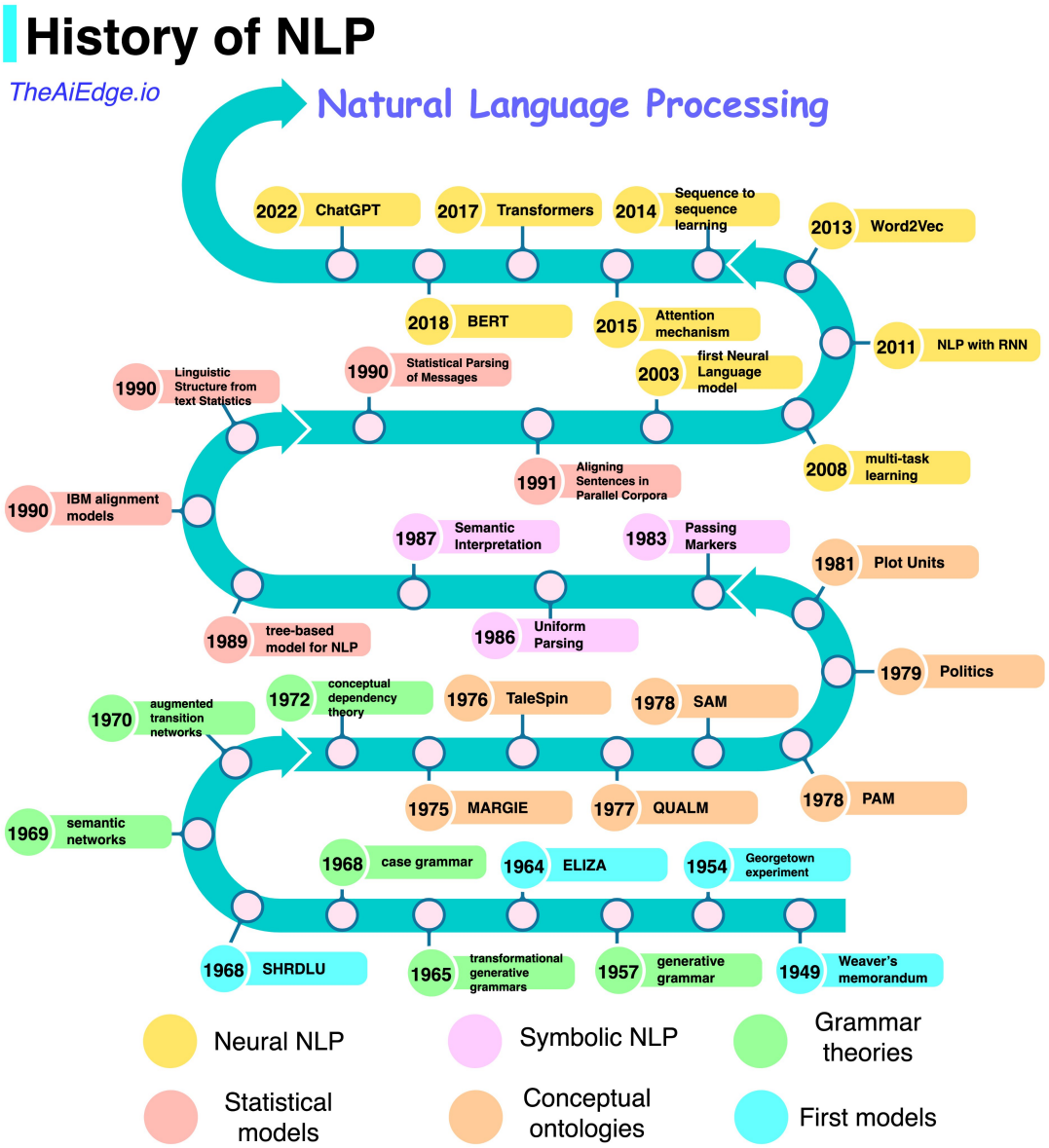
Master Degree in Computer Engineering
University of Padua
Lecturer : Giorgio Satta

Discussion & Conclusions



The gradient, Walid S. Saba

NLP timeline



https://newsletter.theaiedge.io/p/natural-language-processing-how-did

NLP timeline

NLP is now moving on at an **unprecedented** pace.

Novel models that came out since the start of our 2023/24 class:

- GPT-4 Omni (OpenAI)
- Copilot (Microsoft)
- Gemini, Gemma (Google)
- LLaMA 3 (Meta AI)
- Claude 3 (Anthropic)

Open problems

The dominant approach to the study of meaning is **denotational semantics**: the meaning of a word, phrase, or sentence is the set of objects or situations in the world that it describes.

The dominant approach to the representation of meaning in NLP is **distributional semantics**: the meaning of a word is the distribution of the contexts in which the word appears.

The two things

- are not entirely different
- yet, they are not the same

Open problems

Missing text phenomenon: our linguistic communication is compressed, we leave out details that we can safely assume the listener/reader knows by virtue of common knowledge of the world.

Example :
Contrast 'eastern philosophy professor' with 'amazing philosophy professor'.

Example :
How many interpretations for 'the table with the book'?

Open problems

The **Winograd schema challenge** (WSC) is a multiple-choice test that employs questions of a very specific structure.

`https://en.wikipedia.org/wiki/Winograd_schema_challenge`

Example :
The city councilmen refused the demonstrators a permit because **they** [feared/advocated] violence.

Does the pronoun 'they' refer to the city councilmen or to the demonstrators?

Open problems

Adversarial testing: create adversarial examples by adding distracting sentences to the input paragraph.

Hallucination: confident response by an AI that cannot be grounded in any of its training data for the LM.

Overstability: the inability of a model to distinguish a correct answer from one that has words in common with it.

Open problems

In order to move toward **better NLP systems** we need to obtain advancements on

- correlation between language and action (pragmatics)
- principles of communications
- discourse planning
- creative aspects of language
- world common knowledge

Explainability

Model explainability refers to the concept of being able to understand the machine learning model and its decisions.

This is usually done through the technique of **probing**

- parametric probing based on multi-layer perceptron (MLP)
- non-parametric probing based on focus words and minimal pairs

Grounding

Language is **grounded** in experience. Humans understand many basic words in terms of associations with sensory-motor experiences.

This is in contrast to dictionaries, which define words in terms of other words.

We need to train our models on **multi-modal data sets**, where words are linked to, for instances, image segments.

Theory vs. invention

Theory often follows invention.

Invention	**Theory**
Telescope [1608]	Optics [1650–1700]
Steam engine [1595–1715]	Thermodynamics [1824–...]
Microscope (1590)	Cell Theory (1665)
Electromagnetism [1820]	Electrodynamics [1821]
Airplane [1885–1905]	Wing Theory [1907]
Compounds [???]	Chemistry [1760s]
Feedback amplifier [1927]	Electronics [...]
Computer [1941–1945]	Computer Science [1950–1960]
Teletype [1906]	Information Theory [1948]

K. Church and M. Liberman, The Future of Computational Linguistics

Source: The Future of Computational Linguistics: On Beyond Alchemy, Kenneth Church and Mark Liberman, 2021

Ethics

Growing research literature/activities on **value sensitive design** in NLP and allied AI fields.

Also called FAccT: Fairness, Accountability, and Transparency.

The main problems are not yet solved. We seek to answer the following questions

- What can go wrong when we use NLP systems, in terms of specific harms to people?
- How can we fix/prevent/mitigate those harms?
- What are our responsibilities as NLP researchers and developers in this regard?

NLP Hype

Superhuman Conversational AI

***Behshad Behzadi**, VP Engineering Google*

AI has reached superhuman levels in various areas such as playing complex strategic and video games, calculating protein folding, and visual recognition. Are we close to superhuman levels in conversational AI as well? In this talk, we address this question, sharing some of the recent developments from Google Cloud AI, Google Brain Research, Deepmind, and Duplex across speech recognition and generation, and natural language understanding.

NLP Hype

Dr. Sasha Luccioni 💻🌎✨
@SashaMTL

Here, fixed it.

~~**Superhuman Conversational AI**~~ **Making Progress in NLP**

high accuracy on benchmarks Go Atari AlphaFold

AI has reached ~~superhuman levels in~~ various ~~areas~~ such as playing ~~complex strategic~~ and ~~video games~~, ~~calculating protein folding~~, and ~~visual recognition.~~ Are we close to ~~superhuman levels~~ in ~~conversational AI~~ as well? In this talk, we address this question, sharing some of the recent developments from Google Cloud AI, Google Brain Research, Deepmind, and Duplex across speech recognition and generation, and natural language understanding.

VisualQA high accuracy language generation

NLP & teaching

One concern with the end-to-end approach is that it encourages students to focus

- too much on network architecture and training methods
- not enough on methodology and content

Unfortunately, NLP courses are under increasing pressure to make room for currently popular methods at the expense of traditional topics.

NLP lecturers ought to provide a broad education, because we do not know what will be important next.

Source: The Future of Computational Linguistics: On Beyond Alchemy, Kenneth Church and Mark Liberman, 2021

NLU Datasets

General Language Understanding Evaluation (GLUE) benchmark is a collection of 9 datasets for evaluating natural language understanding (NLU) systems:

- Corpus of Linguistic Acceptability (CoLA)
- Stanford Sentiment Treebank (SST)
- Microsoft Research Paragraph Corpus (MRPC)
- Quora Question Pairs (QQP)
- Multi-Genre NLI (MNLI)
- Question NLI (QNLI)
- Recognizing Textual Entailment (RTE)
- Winograd NLI
- Diagnostics Main

`https://gluebenchmark.com` — superseeded by SuperGLUE

NLU Datasets

Massive Multitask Language Understanding (MMLU) is a test set to measure a model multitask accuracy.

The test covers 57 tasks, including among others

- science, technology, engineering and mathematics (STEM)
- social science and humanities
- finance, accounting, and marketing
- professional medicine

To attain high accuracy on this test, models must possess extensive world knowledge and problem solving ability.

`https://paperswithcode.com/dataset/mmlu`.

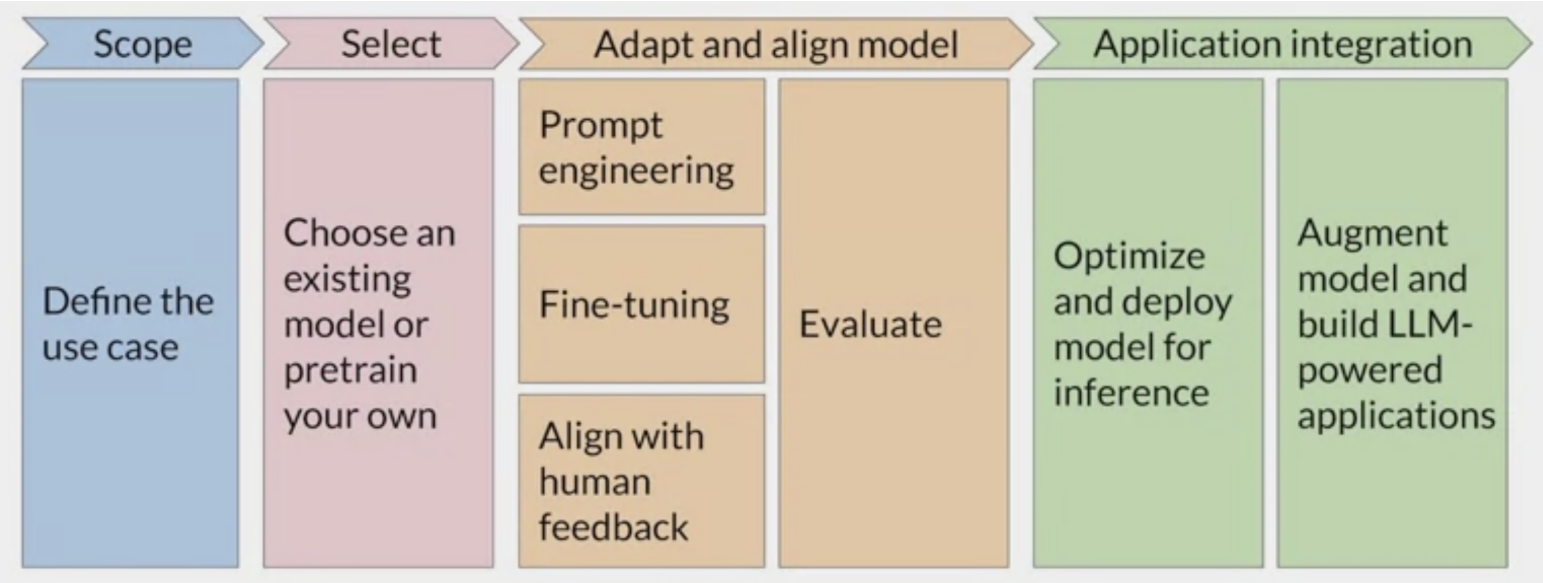
NLU Datasets

Holistic Evaluation of Language Models (HELM) aims to improve the transparency of models, and to offer guidance on which models perform well for specific tasks.

HELM takes a multimetric approach, measuring seven metrics: accuracy, calibration, robustness, fairness, bias, toxicity, efficiency.

Generative AI project lifecycle

The general schema for the lifecycle of a generative AI project.



FLAN

Fine-tuned Language Net (FLAN) compiles several datasets into a mix of zero-shot, few-shot and chain-of-thought templates. It is a specific set of instructions used to fine-tune different models.

Consists of 473 datasets across 146 task categories.

Several LLM instruction tuned with FLAN: Flan-T5, Flan-PaLM, etc.

FLAN

		Model Details				Data Collection & Training Details			
Release	Collection	Model	Base	Size	Public?	Prompt Types	Tasks in Flan	# Exs	Methods
2020 05	UnifiedQA	UnifiedQA	RoBerta	110-340M	P	ZS	46 / 46	750k	
2021 04	CrossFit	BART-CrossFit	BART	140M	NP	FS	115 / 159	71.M	
2021 04	Natural Inst v1.0	Gen. BART	BART	140M	NP	ZS / FS	61 / 61	620k	+ Detailed k-shot Prompts
2021 09	Flan 2021	Flan-LaMDA	LaMDA	137B	NP	ZS / FS	62 / 62	4.4M	+ Template Variety
2021 10	P3	T0, T0+, T0++	T5-LM	3-11B	P	ZS	62 / 62	12M	+ Template Variety + Input Inversion
2021 10	MetaICL	MetaICL	GPT-2	770M	P	FS	100 / 142	3.5M	+ Input Inversion + Noisy Channel Opt
2021 11	ExMix	ExT5	T5	220M-11B	NP	ZS	72 / 107	500k	+ With Pretraining
2022 04	Super-Natural Inst.	*Tk*-Instruct	T5-LM, mT5	11-13B	P	ZS / FS	1556 / 1613	5M	+ Detailed k-shot Prompts + Multilingual
2022 10	GLM	GLM-130B	GLM	130B	P	FS	65 / 77	12M	+ With Pretraining + Bilingual (en, zh-cn)
2022 11	xP3	BLOOMz, mT0	BLOOM, mT5	13-176B	P	ZS	53 / 71	81M	+ Massively Multilingual
2022 12	Unnatural Inst.†	T5-LM-Unnat. Inst.	T5-LM	11B	NP	ZS	~20 / 117	64k	+ Synthetic Data
2022 12	Self-Instruct†	GPT-3 Self Inst.	GPT-3	175B	NP	ZS	Unknown	82k	+ Synthetic Data + Knowledge Distillation
2022 12	OPT-IML Bench†	OPT-IML	OPT	30-175B	P	ZS + FS, CoT	~2067 / 2207	18M	+ Template Variety + Input Inversion + Multilingual
2022 10	Flan 2022 (ours)	Flan-T5, Flan-PaLM	T5-LM, PaLM	10M-540B	P NP	ZS + FS, CoT	1836	15M	+ Template Variety + Input Inversion + Multilingual

The Flan Collection: Designing Data and Methods for Effective Instruction Tuning, Longpre et al. 2023

ChatBot Arena

Chatbot Arena Leaderboard is a novel platform that leverages crowdsourced human evaluation to rank LLMs

- LLMs take on the role of "players" in head-to-head comparisons
- users are invited to vote on which LLM they find more engaging, informative, or helpful

The **Elo** system is used to dynamically adjusts the LLMs' scores, generating a ranking.

ChatBot Arena

Rank* (UB)	Model	☆ Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo-2024-04-09	1259	+4/-3	35931	OpenAI	Proprietary	2023/12
2	GPT-4-1106-preview	1253	+2/-3	73547	OpenAI	Proprietary	2023/4
2	Claude 3 Opus	1251	+3/-3	80997	Anthropic	Proprietary	2023/8
2	Gemini 1.5 Pro API-0409-Preview	1250	+3/-3	39482	Google	Proprietary	2023/11
2	GPT-4-0125-preview	1247	+3/-2	67354	OpenAI	Proprietary	2023/12
6	Llama-3-70b-Instruct	1210	+3/-4	53404	Meta	Llama 3 Community	2023/12
6	Bard (Gemini Pro)	1209	+5/-6	12387	Google	Proprietary	Online
7	Claude 3 Sonnet	1201	+2/-3	78956	Anthropic	Proprietary	2023/8
9	Command R+	1191	+3/-3	44988	Cohere	CC-BY-NC-4.0	2024/3
9	GPT-4-0314	1190	+3/-4	52079	OpenAI	Proprietary	2021/9
11	Claude 3 Haiku	1181	+2/-3	69660	Anthropic	Proprietary	2023/8
12	GPT-4-0613	1165	+3/-3	70726	OpenAI	Proprietary	2021/9