Linguistics in the Age of Data: Introspection vs. Observation

James Pustejovsky
CS 115B

Brandeis University
January 14, 2025

Outline

Investigate hypotheses by consulting native speakers? intuitions

- Most linguists assume that people can distinguish strings of words that are sentences of their language from strings of words that are not sentences of their language.
- So imagine that you are a machine or a classifier that takes a sentence as input, and returns accept or reject as output.

Speakers as automata: accept/reject strings of words

- ☺ The student read a book.
- ☹ Student the a read book.
- ☹ Of to dog every walks.
- ☺ Every dog walks fast.

Grammaticality

- A string of words that you recognize as a sentence in your native language is grammatical.
- A string of words that you do not recognize as a sentence in your native language is ungrammatical.
- When you decide whether a sentence is grammatical or ungrammatical, this is called giving a grammaticality judgment.
- Ungrammatical sentences are preceded by an asterisk or star (*). Sometimes they are called starred sentences.

Grammatical ≠ meaningful

- 1. It is unlikely that Pat will succeed.
  2. It is improbable that Pat will succeed.
  3. Pat is unlikely to succeed.
  4. *Pat is improbable to succeed.
     This could be meaningful, but most people consider it to be ungrammatical
- 1. They saw Pat with Chris.
  2. They saw Pat and Chris.
  3. Who did they see Pat with?
  4. *Who did they see Pat and?
     Again, this could be meaningful, but it is ungrammatical.

Foundational Insights: 1940s - 1950s

automata theory and information-theoretic models

- Turing's work led first to the McCulloch-Pitts neuron (McCulloch and Pitts, 1943), a simplified model of the neuron as a kind of computing element that could be described in terms of propositional logic,
- Kleene (1951) and (1956) develops models finite automata and regular expressions.
- Shannon (1948) applied probabilistic models of discrete Markov processes to automata for language. (continued)

Foundational Insights: 1940s - 1950s

formal language theory

- Chomsky (1956), drawing the idea of a finite state Markov process from Shannon's work, first considered finite-state machines as a way to characterize a grammar, and defined a finite-state language as a language generated by a finite-state grammar.
- These early models led to the field of formal language theory, which used algebra and set theory to define formal languages as sequences of symbols.
- This includes the context-free grammar, first defined by Chomsky (1956) for natural languages

Foundational Insights: 1940s - 1950s

Noisy Channel Model

- Shannon's metaphor of the noisy channel and decoding for the transmission of language through media like communication channels and speech acoustics.
- Shannon's use of entropy from thermodynamics measures the information capacity of a channel, or the information content of a language; performed the first measure of the entropy of English using probabilistic techniques.
- Sound spectrograph was developed (Koenig et al., 1946), and instrumental phonetics that laid the groundwork for later work in speech recognition.
- First machine speech recognizers in the early 1950s.

- Language Theory: The work of Chomsky and others on formal language theory and generative syntax throughout the late 1950s and early to mid 1960s, and the work of many linguistics and computer scientists on parsing algorithms, initially top-down and bottom-up and then via dynamic programming.
- One of the earliest complete parsing systems was Zelig Harris's Transformations and Discourse Analysis Project (TDAP), which was implemented between 1958-1959 at the University of Pennsylvania (Harris, 1962).

Two Cultures Emerge: 1957 - 1970

Speech and Language Processing: Symbolic

- Artificial Intelligence: In the summer of 1956 John McCarthy, Marvin Minsky, Claude Shannon, and Nathaniel Rochester brought together a group of researchers for a two-month workshop on what they decided to call artificial intelligence (AI).
- The major focus of the new field was the work on reasoning and logic typified by Newell and Simon's work on the Logic Theorist and the General Problem Solver.
- At this point early natural language understanding systems were built.
- These were simple systems that worked in single domains mainly by a combination of pattern matching and keyword search with simple heuristics for reasoning and question-answering. By the late 1960s more formal logical systems were developed.

- Bayesian method was beginning to be applied to the problem of optical character recognition.
- Bledsoe and Browning (1959) built a Bayesian system for text-recognition that used a large dictionary and computed the likelihood of each observed letter sequence given each word in the dictionary by multiplying the likelihoods for each letter.
- Mosteller and Wallace (1964) applied Bayesian methods to the problem of authorship attribution on The Federalist papers.
- The first on-line corpora: the Brown corpus of American English, a 1 million word collection of samples from 500 written texts from different genres (newspaper, novels, non-fiction, academic, etc.), which was assembled at Brown University in 1963-64)

Paradigms Develop

Stochastic Models

- Development of speech recognition algorithms, particularly the use of the Hidden Markov Model and the metaphors of the noisy channel and decoding, developed independently by Jelinek, Bahl, Mercer, and colleagues at IBM, and by Baker at Carnegie Mellon University

- Finite-state models began to receive attention again after work on finite-state phonology and morphology by Kaplan and Kay (1981) and finite-state models of syntax by Church (1980).
- Return of empiricism; most notably the rise of probabilistic models throughout speech and language processing, influenced strongly by the work at the IBM Thomas J. Watson Research Center on probabilistic models of speech recognition.
- These probabilistic methods and other such data-driven approaches spread into part-of-speech tagging, parsing and attachment ambiguities, and connectionist approaches from speech recognition to semantics.

Paradigm Integration

The Field Changes: 1994-2010

- Probabilistic and data-driven models had become quite standard throughout natural language processing.
- Algorithms for parsing, part-of-speech tagging, reference resolution, and discourse processing all began to incorporate probabilities, and employ evaluation methodologies borrowed from speech recognition and information retrieval.
- The increases in the speed and memory of computers had allowed commercial exploitation of a number of subareas of speech and language processing, in particular speech recognition and spelling and grammar checking.

Rationalism and Empiricism 1/3

- Rationalism:
  the source of knowledge is reason.
- Empiricism:
  the source of knowledge is data.

Rationalism and Empiricism 2/3

- Chomsky emphasized *creativity* of language, as manifested in recursive generative rules :
  $S \rightarrow NP\ VP$, $VP \rightarrow VP\ NP$
- Empiricists emphasize *common language patterns* (e.g. collocations) and *predictability* of language
  "About half of the letters or words we choose in writing or speaking (although we are not ordinarily aware of it) are really controlled by the statistical structure of the language. ",
  Warren Weaver, pioneer of MT (1949)

Rationalism and Empiricism 3/3

- Chomsky (1957): *I think that we are forced to conclude that grammar is autonomous and independent of meaning.*
- Corpus linguist John Sinclair (1991):
  It is folly to decouple lexis and syntax, or either of those and semantics. The realization of meaning is far more explicit than is suggested by abstract grammars. By far the majority of text is made of the occurrence of common words in common patterns. Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multiword patterns that make up text.

Views of linguistic data

Rationalism/Internalism

- Language is something in people's minds: a set of rules and principles that allows them to make grammaticality judgments and produce and understand sentences that they have never heard before (i-language or internal language)
- We study i-language by asking people to give grammaticality judgments.
- A corpus (a collection of texts or speech) is e-language, or external language. It is not the object of study.

Views of linguistic data

Rationalism/Internalism

- Accounts for creativity: People can produce and understand sentences that have never been uttered before. They are not repeating memorized patterns, but applying productive rules.
- Probability is not grammaticality: grammatical sentences may have very low probability.
- Probability reflects facts about the world, but grammaticality is independent of context.

Views of linguistic data

Empiricist

- Corpora are the objects of study.
- We study language by examining patterns in corpora (collections of texts or speech).
- Frequency of occurrence in a corpus is easier to measure reliably than a grammaticality judgment.
- Many ungrammatical sentences turn out to be acceptable in the right context.
- Identifying the right context turns out to be an interesting question in its own right.

Studying Language Diversity

Franz Boas and Students

- Notions and concepts of traditional grammar were useless against totally new languages with no historical relation to them.
- The study of language to study indigenous cultures made them aware of the diversity in human languages.
- According to unlimited diversity principle, universalist beliefs fell out of favor.
- Each language has its own way of organizing concepts of grammar hence the grammar of each language is different.
- There was no guarantee that techniques, methods or tools developed for language will work on very next language.
- Thus to study the grammar of a language a linguist had to have a number of tools (for description and analysis).

Discovery Procedures and Grammar Induction

Bloomfield, others, 1940s - 1950s

- Adopt Hempel's and Popper's notion of *science as pattern discovery*
- Language can be explained in probabilistic, behavioral terms
- Languages are diverse systems learned from the environment
- The aim was to describe the diversity of linguistic behavior; analyze linguistic structure in formal terms - producing formal descriptions of grammar (including phonetics, morphology, syntax, etc.)

Early Computational Models

- Empirical and statistical methods were popular the 1950s
- Shannon's information theoretic approach to language
- *"All of us were convinced that speech, in English or any other language, was a Markov process. From this to the conviction that ... the set of all English sentences can be generated by a Markov source was only a small step."* (Bar-Hillel, 1975)
- Early machine translation efforts of 1950s and 1960s.

Chomsky's Critique of Finite-state Methods

- *"I had no personal interest in the experimental studies and technological advances. [...] As for machine translation and related enterprises, this seemed to me pointless, as well as probably quite hopeless. [...] I could not fail to be aware of the ferment and excitement [in the early 1950s]. But I felt myself no part of it."* (Chomsky, 1975)

Syntactic Structures (Chomsky, 1957) 1/3

- *From now on I will consider a language to be a set (finite or infinite) of sentences, each finite in length and constructed out of a finite set of elements.*
- *The fundamental aim in the linguistic analysis of a language L is to separate the grammatical sequences which are the sentences of L from the ungrammatical sequences which are not sentences of L and to study the structure of the grammatical sequences.*
- On what basis do we actually go about separating grammatical sequences from ungrammatical sequences?

Syntactic Structures (Chomsky, 1957) 2/3

- First, it is obvious that the set of grammatical sentences cannot be identified with any finite and somewhat accidental corpus of observed utterances.
- Second, the notion "grammatical" cannot be identified with "meaningful" or "significant" in any semantic sense.
- Sentences (1) and (2) are equally nonsensical, but any speaker of English will recognize that only the former is grammatical.
    1. *Colorless green ideas sleep furiously.*
    2. *Furiously sleep ideas green colorless.*

Syntactic Structures (Chomsky, 1957) 3/3

- Third, the notion "grammatical in English" cannot be identified in any way with the notion "high order of statistical approximation to English." It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse.
- Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English. Yet (1), though nonsensical, is grammatical, while (2) is not.
- One's ability to produce and recognize grammatical utterances is not based on notions of statistical approximation and the like. . I think that we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure.

Carnap's Logical Structure of Language (1934)

- Pirots karulize elatically.

Umformungsregeln von gleicher Art sind, und daß beide formal gefaßt werden können. Daß z.B. die Wortreihe "Piroten karulieren elatisch" ein Satz ist, kann, wenn eine geeignete Regel aufgestellt ist, festgestellt werden, sofern nur bekannt ist, daß "Piroten" ein *Substantivum* (im Plural), "karulieren" ein *Verbum* (in der 3. Pers. Plur, Ind.) und "elatisch" ein *Adverbium* ist (was übrigens in einer gut konstruierten Wortsprache, wie z.B. in Esperanto, aus der *Form* der Wörter zu ersehen sein würde); die *Bedeutung* der Wörter braucht hierfür nicht bekannt zu sein. Ferner kann, wenn eine geeignete Regel aufgestellt ist, aus dem genannten Satz und dem Satz "A ist ein Pirot" der Satz "A karuliert elatisch" erschlossen werden, wenn nur wieder die Wortarten der einzelnen Wörter bekannt sind; auch hierfür braucht ihre Bedeutung und der Sinn der drei Sätze nicht bekannt zu sein.

Creating a Language Model 1/5

- Language models are core component of many applications where task is to recover an utterance from an input. This is done through Bayes' rule:

$$\underbrace{p(X \mid Y)}_{\text{posterior probability}} = \frac{\overbrace{p(Y \mid X)}^{\text{likelihood}} * \overbrace{p(X)}^{\text{prior probability}}}{\underbrace{p(Y)}_{\text{marginal likelihood (probability)}}}$$

Creating a Language Model 2/5

- Goal: predict sentence given acoustics

$$w^* = \arg\max_X P(X \mid a)$$

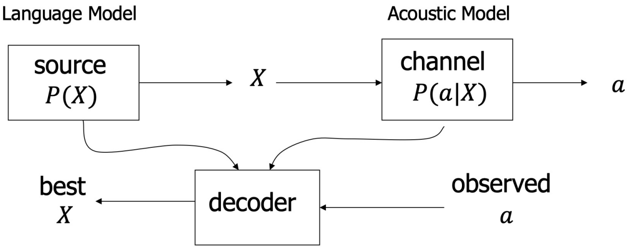
- The noisy channel approach:

Language model: Distributions over sequences of words (sentences)

$$\begin{aligned} w^* &= \arg\max_X P(X \mid a) \\ &= \arg\max_X P(a \mid X) P(X) / P(a) \\ &= \arg\max_X P(a \mid X) P(X) \end{aligned}$$

Acoustic model: Distributions over acoustic waves given a sentence

Creating a Language Model 3/5



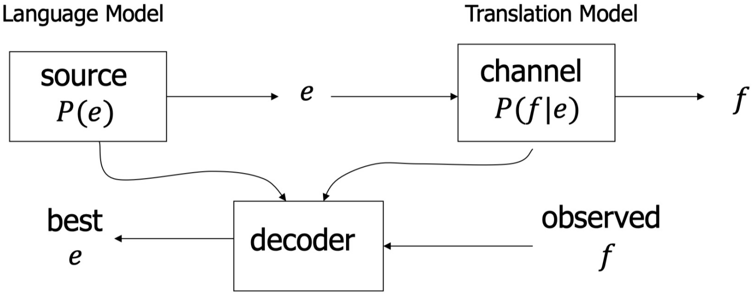
$$\arg\max_X P(X \mid a) = \arg\max_X P(a \mid X)P(X)$$

Creating a Language Model 4/5

"Also knowing nothing official about, but having guessed and inferred considerable about, the powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. **When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'** "

Warren Weaver
(1955:18, quoting a letter he wrote in 1947)

Creating a Language Model 5/5



$$\arg\max_e P(e \mid f) = \arg\max_e P(f \mid e)P(e)$$

Shannon's Language Model 1/5

- How to compute this joint probability:
    - P(its, water, is, so, transparent, that)
- Intuition: let's rely on the Chain Rule of Probability

Shannon's Language Model 2/5

- Recall the definition of conditional probabilities

Rewriting:

- More variables:

  $P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$
- The Chain Rule in General

$$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

Shannon's Language Model 3/5

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

P("its water is so transparent") =

P(its) × P(water|its) × P(is|its water)

× P(so|its water is) × P(transparent|its water is so)

Shannon's Language Model 4/5

- Could we just count and divide?

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{Count(\text{its water is so transparent that the})}{Count(\text{its water is so transparent that})}$$

- No! Too many possible sentences!
- We'll never see enough data for estimating these

Shannon's Language Model 5/5

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{that})$$

- Or maybe

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{transparent that})$$

Deconstructing Chomsky's Argument 1/4

- Neither of these sentences has ever appeared before in human discourse. None of the substrings has either.
  1. Colorless green ideas sleep furiously.
  2. Furiously sleep ideas green colorless.
- Hence the probabilities $P(W_1)$ and $P(W_2)$ cannot be relevant to a native English speaker's knowledge of the difference between these sentences.
- "... in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English. "

Deconstructing Chomsky's Argument 1/4

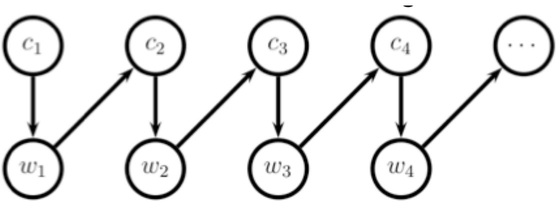
- Neither of these sentences has ever appeared before in human discourse. None of the substrings has either.
  1. Colorless green ideas sleep furiously.
  2. Furiously sleep ideas green colorless.
- Hence the probabilities $P(W_1)$ and $P(W_2)$ cannot be relevant to a native English speaker's knowledge of the difference between these sentences.
- "... in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English. "
- This claim is false: all modern statistical models of language can assign probabilities to previously unseen utterances.

Deconstructing Chomsky's Argument 2/4

- Language users must be able to generalize beyond their input.
- Statistical inference is the study of such generalization
- Pereira (2000) used a class-based bigram model to model (1) and (2):
  Maybe probabilities are not such a useless model of language knowledge after all.

$$\frac{P(\text{Colorless green ideas sleep furiously})}{P(\text{Furiously sleep ideas green colorless})} \approx 2 \times 10^5$$

class-based model



- The c_i are unseen variables that have to be introduced into the model, either through annotation or unsupervised learning
- For now, focus on the basic problem of the language model: $P(W)$

Big Data Corpora

- It is quite typical for researchers to use any collection of texts for linguistic analysis.
- Often proceed opportunistically: whatever data comes in handy is used.
- However, the term corpus usually implies the following characteristics:
    - sampling/representativeness
    - finite size
    - machine-readable form
    - a standard reference (time-bound)

1950-1990 - The Absence of Data

- Chomsky liberated the field of linguistics in the 1950s
- Generation through recursive functions allows one to create your own corpus
- Experimenting with new datasets that are not attested in actual "found data"

1990-2024 - The Absence of Theory

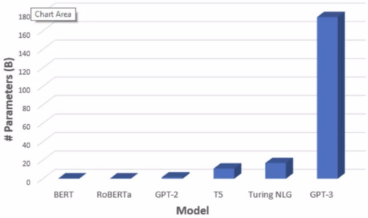
- Big Data and statistical modeling has largely dominated the fields of CL and AI, both theoretical and applied.
- Deep Learning seems positioned to obviate theory completely: RNNs, CNNs,
- This will not happen: machine learning and deep learning make theoretical assumptions in both the data preparation and feature selection and engineering phase of training.
- Theory is more relevant than ever before.

2018 - ... - Challenges requiring Theory and Data

- Dialogue Systems
- Situated and Contextualized Meaning
- Multimodal Communication

Overview of GPT-3

- Generative Pre-trained Transformer – 3
- Developed by OpenAI in May 2020
- Largest neural network ever created



- Philosophy: the bigger, the better

What is the motivation around it?

Pre-trained language models

- Current state-of-the-art models in NLP
- Trained in a semi-supervised learning with large corpora
  - Both X and Y are extracted from the text without having a prior labeled dataset
  - Acquire high capability for modeling the natural language (with task-agnostic architecture)
- Limitation (i): need for downstream task-specific datasets and fine-tuning
  - Difficult to collect large supervised training datasets for every new task
- Limitation (ii): non-correlation with humans
  - Humans do not require large supervised datasets to learn most language tasks, but only brief directives or a tiny number of demonstrations are needed
  - So, why give models a large dataset of labeled examples for every new task?

Why not try to create NLP systems to have the same fluidity and generality as humans?

LLMs and GPT-3

Solution: more parameters and in-context learning

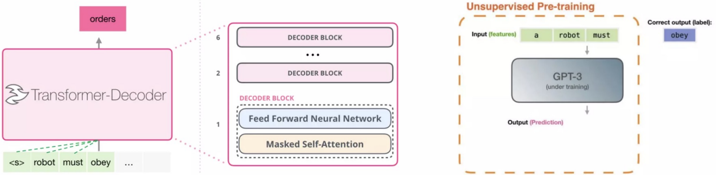
- Let models develop a broad set of skills and pattern recognition abilities during pre-training and use them at inference time to adapt to the desired task rapidly

- Since in-context learning involves absorbing many skills and tasks within the model's parameters, it is plausible that learning abilities correlate with model size.

OpenAI creates GPT-3 to show that very large unsupervised language models trained with a lot of data can multitask to the level of fine-tuned state-of-the-art models

Model Architecture and Training Process

- The GPT-3 model architecture is the same as its GPT-2 predecessor
  - **Transformer-based**, built using **only decoder blocks** (BERT opposite)
  - Stronger in natural language generation (NLG), instead of creating contextual embeddings
- An **auto-regressive** language model
  - GPT-3 is trained using **next word prediction**, outputting one token (wordpiece) at a time
  - Differently from bidirectional models like BERT, the prediction at each step is conditioned only on the **left context** (masked self-attention)

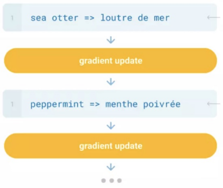


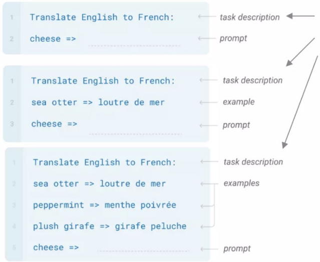
- From an architecture perspective, GPT-3 is not actually very novel!
  - … So, what makes it so special and magical? **It's really big**

Zero-, one-, few-shot vs fine-tuning

- GPT-3 can perform specific tasks without any special tuning
  - **Most other pre-trained language models require an elaborate fine-tuning (also with architectural changes) on thousands of samples to perform well on downstream tasks**
  - **GPT-3 doesn't need a fine-tuning step and directly uses a single pre-trained model for all downstream tasks (plug-and-play), demonstrating even superior performance**
- **Three different evaluation settings focused on task-agnostic performance**, which allows zero, one, or a few examples to be prefixed to the input model

Fine-tuning (repeated gradient updates using a large corpus of example tasks) → postponed





(i) Zero-shot
(ii) One-shot
(iii) Few-shot

Context to better inform the model about what it is expected to do

"If I were to see this text somewhere on the Web, what will be the most likely next word?"

Results - ii

- **GPT-3,** the giant version, **was compared to the SOTA solutions** in different datasets.
  - **LAMBADA, StoryCloze, HellaSwag, TriviaQA, BLUE...**
- In many cases, **it reaches and outperforms the previous SOTA**, which are neural models fine-tuned on the dataset.

Setting	LAMBADA (acc)	LAMBADA (ppl)	StoryCloze (acc)	HellaSwa (acc)
SOTA	68.0[a]	8.63[b]	**91.8**[c]	**85.6**[d]
GPT-3 Zero-Shot	**76.2**	**3.00**	83.2	78.9
GPT-3 One-Shot	**72.5**	**3.35**	84.7	78.1
GPT-3 Few-Shot	**86.4**	**1.92**	87.7	79.3

Setting	En→Fr	Fr→En	En→De	De→En	En→Ro
SOTA (Supervised)	**45.6**[a]	35.0 [b]	**41.2**[c]	40.2[d]	**38.5**[e]
XLM [LC19]	33.4	33.3	26.4	34.3	33.3
MASS [STQ+19]	<u>37.5</u>	34.9	28.3	35.2	<u>35.2</u>
mBART [LGG+20]	-	-	<u>29.8</u>	34.0	35.0
GPT-3 Zero-Shot	25.2	21.2	24.6	27.2	14.1
GPT-3 One-Shot	28.3	33.7	26.2	30.4	20.6
GPT-3 Few-Shot	32.6	<u>39.2</u>	29.7	<u>40.6</u>	21.0

```
Poor English input:  The patient was died.
Good English output:  The patient died.
```

```
Poor English input:  We think that Leslie likes ourselves.
Good English output:  We think that Leslie likes us.
```

Limits

- Despite the great improvement of GPT-3, **it still has notable weakness,** it has difficulty with **common sense physics** and in **long text generation**.
  - Lose of coherence
  - Contradiction
  - Useless semantical repetition
- Large language model are not grounded in other domains of experience as video or real-world physic interaction, **lacking a large amount of context.**
- GPT-3 works better after pretrain, **it is still far from human level.**
  - Humans show strong zero-shot capabilities
- By now, **it is impossible to say how GPT-3 learns** during train
  - **It is even more complex understand what it learns in inference time**
    - Does it learn the new task from scratch? Does it reshape similar task it learned?
- Finally, it shares some limitations common to most deep learning models
  - The knowledge learned is no interpretable
  - It requires large resources and time for the train
  - It is strongly affected by biases in the data

Conclusion

- Linguistics is now both a theoretical and experimental discipline
- The scope of observed data for language study and theorizing is richer and broader than ever.
- Linguistic Corpora and captured media datasets will enable contextualized and embodied interpretation of linguistic utterances
- This will enable the development of more expressive and broader theories of language and communication

Temporary page!

LATEX was unable to guess the total number of pages correctly.
there was some unprocessed data that should have been added
the final page this extra page has been added to receive it.
If you rerun the document (without altering it) this surplus page
will go away, because LATEX now knows how many pages to expe
for this document.