The Linguist and ChatGPT

Alessandro Lenci

COmputational LINGuistics Laboratory
Università di Pisa
Dipartimento di Filologia, Letteratura e Linguistica (FiLeLi)

New “Enemies” at the Gates?

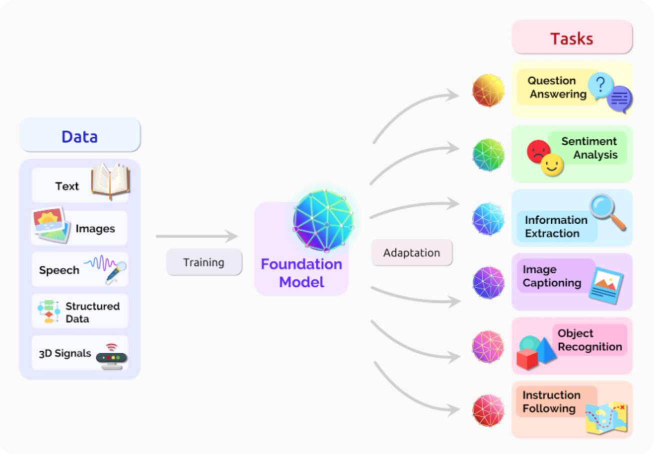
“A spectre is haunting Computational Linguistics:
the spectre of Large Language Models”

OpenAI
ChatGPT

Foundation Models

Bommasani et al. (2021). On the opportunities and risks of foundation models. *ArXiv*: 2108.07258

- Large Language Models aka Foundations models aka Pre-trained Models



Large Language Models (LLMs)

common features

- Deep (i.e., multi-layered) artificial neural networks pretrained on huge amounts of unlabeled data, that are then adapted to a wide range of downstream tasks
- Pretraining
  - the network acquires a large amount of core knowledge from text corpora (or multimodal data) by being trained as a language model with a self-supervised string prediction task
- Task adaptation
  - fine-tuning: re-train the pre-trained model on a specific supervised task (e.g., question-answering) to adapt its weights
    - the model can leverage the knowledge encoded during pretraining, achieving excellent performances with smaller amount of labeled data
  - in-context learning (prompting): give to the model an instruction as input sequence, called prompt, that contains a natural language description of the task, and optionally a few examples

Large Language Models (LLMs)

common features

- Deep (i.e., multi-layered) artificial neural networks pretrained on huge amounts of unlabeled data, that are then adapted to a wide range of downstream tasks
- Pretraining
  - the network acquires a large amount of core knowledge from text corpora (or multimodal data) by being trained as a language model with a self-supervised string prediction task
- Task adaptation
  - fine-tuning: re-train the pre-trained model on a specific supervised task (e.g., question-answering) to adapt its weights
    - the model can leverage the knowledge encoded during pretraining, achieving excellent performances with smaller amount of labeled data
  - in-context learning (prompting): give to the model an instruction as input sequence, called prompt, that contains a natural language description of the task, and optionally a few examples

Large Language Models (LLMs)

common features

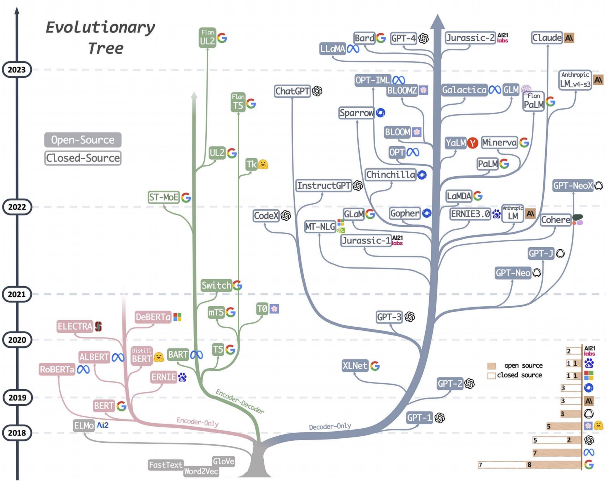
- Deep (i.e., multi-layered) artificial neural networks pretrained on huge amounts of unlabeled data, that are then adapted to a wide range of downstream tasks
- Pretraining
  - the network acquires a large amount of core knowledge from text corpora (or multimodal data) by being trained as a language model with a self-supervised string prediction task
- Task adaptation
  - fine-tuning: re-train the pre-trained model on a specific supervised task (e.g., question-answering) to adapt its weights
    - the model can leverage the knowledge encoded during pretraining, achieving excellent performances with smaller amount of labeled data
  - in-context learning (prompting): give to the model an instruction as input sequence, called prompt, that contains a natural language description of the task, and optionally a few examples

Large Language Models (LLMs)

common features

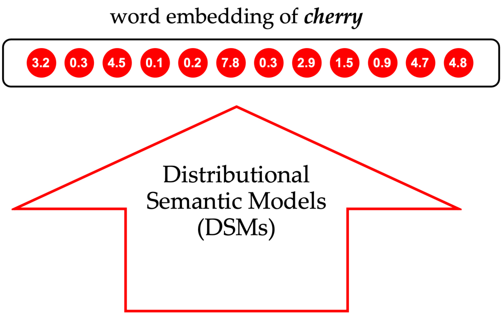
- Deep (i.e., multi-layered) artificial neural networks pretrained on huge amounts of unlabeled data, that are then adapted to a wide range of downstream tasks
- Pretraining
  - the network acquires a large amount of core knowledge from text corpora (or multimodal data) by being trained as a language model with a self-supervised string prediction task
- Task adaptation
  - fine-tuning: re-train the pre-trained model on a specific supervised task (e.g., question-answering) to adapt its weights
    - the model can leverage the knowledge encoded during pretraining, achieving excellent performances with smaller amount of labeled data
  - in-context learning (prompting): give to the model an instruction as input sequence, called prompt, that contains a natural language description of the task, and optionally a few examples

The Tree of LLMs and their Roots



LLMs as Distributional Semantic Models (DSMs)

Lenci and Sahlgren (in press), *Distributional Semantics*, Cambridge University Press



*… so we went outside, picked several red **cherries** and ate them … the colour of an orange pink sunset and an indulgent length of rich, red **cherry** fruit with hints of almonds on the dry finish …*

The Type of Representations They Learn

- Traditional DSMs are models of the lexicon, intended as a repository of out-of-context lexical items
- Each word type in the model vocabulary is represented with a unique, context-independent, static embedding
- Contextual DSMs (= foundation models) learn contextual embedding that overcome the meaning conflation deficiency of static DSMs

Meaning conflation deficiency

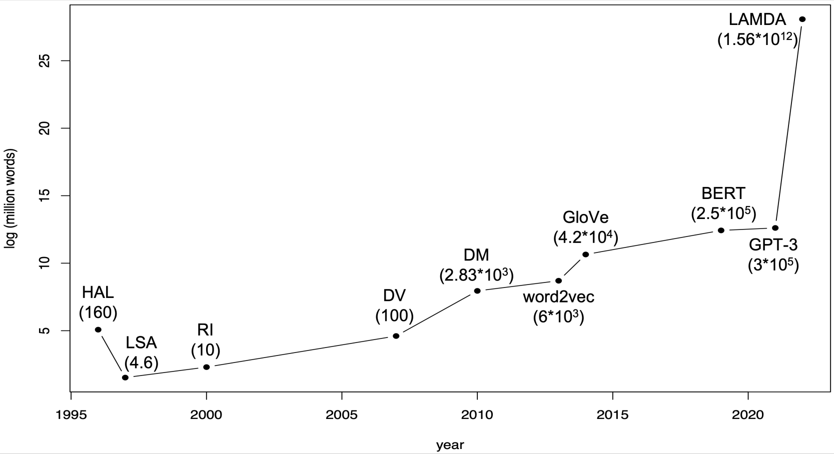
Different word senses are all mixed within the same distributional vector

- Nearest neighbors of the verb *play*
  - playing, game, play_N, audition, player, match, sing, star, coach, badminton

The Model Size

- Distributional models have has been characterized by an exponential growth both in their architecture and in the amount of training text
    - BERT Large: 24 layers and 340 million parameters (i.e., the weights that are set during training and determine the network behavior)
    - GPT-3: 96 layers and 175 billion parameters
- The training side of GPT-3 is 499 billion tokens (the whole Wikipedia is just the 3% of the training texts).

The Growth of Training Corpora



The Amount of Information They Encode

- Foundation models learn from texts a far greater amount of information than any former DSM
  - they encode aspects of syntactic structure, several dimensions of lexical and sentence meaning, (Tenney et al. 2019, Manning et al. 2020), pragmatic aspects (Hu et al. 2022), and so on
- The largest models like GPT-3 reveal emergent abilities to carry out linguistic tasks (e.g., translating, question-answering, etc.) without any task-specific training (Brown et al. 2020, Wei et al. 2022)

Crucial issue

The most recent, largest models are not fully disclosed by companies!

The Amount of Information They Encode

- Foundation models learn from texts a far greater amount of information than any former DSM
  - they encode aspects of syntactic structure, several dimensions of lexical and sentence meaning, (Tenney et al. 2019, Manning et al. 2020), pragmatic aspects (Hu et al. 2022), and so on
- The largest models like GPT-3 reveal emergent abilities to carry out linguistic tasks (e.g., translating, question-answering, etc.) without any task-specific training (Brown et al. 2020, Wei et al. 2022)

Crucial issue

The most recent, largest models are not fully disclosed by companies!

Computational Linguistics and its Actors

- The main goal of computational linguistics and NLP is the development of models to endow artificial systems with human-like linguistic abilities

The Linguist & The Engineer

Ivan Sag
(1949 - 2013)

Fred Jelinek
(1932 - 2010)

Rule-Based Models

The Linguist

- Computational grammar engineering
- Computational lexicon building

Classical Machine Learning Models

The Linguist

- Corpus annotation
(for supervised learning)
- Linguistic feature engineering

Neural Models: Representation Learning

The Linguist

- Corpus annotation
  (for supervised learning)

What is the Role of the Linguist in the Age of LLMs?

OpenAI
ChatGPT

The Human Knowledge of Language

Culicover and Jackendoff (2005), *Simpler Syntax*

"A speaker of a human language can create and understand an unlimited number of different utterances, concerning an unlimited number of different topics. This entails that a language user with a finite brain must have a productive system for constructing new utterances online (in both production and perception) from a finite basis stored in memory" (p. 10)

- Knowledge of language must be:
  - learnable from finite input
  - generalizable and creative

The Human Knowledge of Language

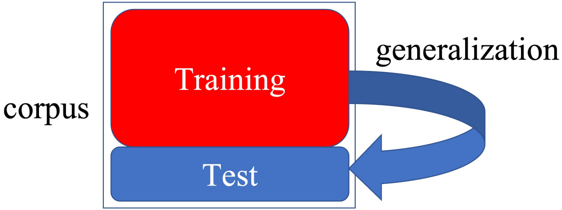
Culicover and Jackendoff (2005), *Simpler Syntax*

"A speaker of a human language can create and understand an unlimited number of different utterances, concerning an unlimited number of different topics. This entails that a language user with a finite brain must have a productive system for constructing new utterances online (in both production and perception) from a finite basis stored in memory" (p. 10)

- Knowledge of language must be:
  - learnable from finite input
  - generalizable and creative

Generalization vs. Creativity

- Generalization is the ability to apply a model to new data
  - typically in machine learning, the new data are taken from the very same distribution of training data



Generalization vs. Creativity

- Creativity is the ability to produce and interpret new data that do not (necessarily) belong to the same distribution of training ones
  - creativity $\Rightarrow$ generalization
  - generalization $\not\Rightarrow$ creativity
- Some examples of linguistic creativity
  - using a word with a novel meaning or in a novel construction
    - *The man **whistled** that the cops were arriving.*
  - understanding a sentence that describes totally new and unlikely situation, or containing rare lexical items
    - *The pharaoh was surfing on the dunes together with a marmot.*
  - distinguishing unlikely from anomalous sentences
    - *?* The journalist writes **an article**, and the professor **an apple**.*

Generalization vs. Creativity

- Creativity is the ability to produce and interpret new data that do not (necessarily) belong to the same distribution of training ones
  - creativity $\Rightarrow$ generalization
  - generalization $\not\Rightarrow$ creativity

- Some examples of linguistic creativity
  - using a word with a novel meaning or in a novel construction
    - *The man **whistled** that the cops were arriving.*
  - understanding a sentence that describes totally new and unlikely situation, or containing rare lexical items
    - *The pharaoh was surfing on the dunes together with a marmot.*
  - distinguishing unlikely from anomalous sentences
    - *?* The journalist writes **an article**, and the professor **an apple**.*

Generalization vs. Creativity

- Creativity is the ability to produce and interpret new data that do not (necessarily) belong to the same distribution of training ones
  - creativity $\Rightarrow$ generalization
  - generalization $\not\Rightarrow$ creativity
- Some examples of linguistic creativity
  - using a word with a novel meaning or in a novel construction
    - *The man **whistled** that the cops were arriving.*
  - understanding a sentence that describes totally new and unlikely situation, or containing rare lexical items
    - *The pharaoh was surfing on the dunes together with a marmot.*
  - distinguishing unlikely from anomalous sentences
    - *?* The journalist writes **an article**, and the professor **an apple**.*

Generalization vs. Creativity

- Creativity is the ability to produce and interpret new data that do not (necessarily) belong to the same distribution of training ones
  - creativity $\Rightarrow$ generalization
  - generalization $\not\Rightarrow$ creativity

- Some examples of linguistic creativity
  - using a word with a novel meaning or in a novel construction
    - *The man **whistled** that the cops were arriving.*
  - understanding a sentence that describes totally new and unlikely situation, or containing rare lexical items
    - *The pharaoh was surfing on the dunes together with a marmot.*
  - distinguishing unlikely from anomalous sentences
    - *?* The journalist writes **an article**, and the professor **an apple**.*

LLMs and Inference

Natural Language Inference (NLI)

Sentence A (premise)	*Sentence B* (hypothesis)	*Entailment label*
John bought a dog	John bought an animal	ENTAILMENT
John bought a dog	John did not buy an animal	CONTRADICTION
John bought a dog	John bought a golden retriever	NEUTRAL

- *Prima facie*, NLMs achieve very high performances in this task, ...
- ...but in fact they often exploit statistical artifacts in datasets, rather than truly learning an inference relation

LLMs and Inference

Natural Language Inference (NLI)

Sentence A (premise)	*Sentence B* (hypothesis)	*Entailment label*
John bought a dog	John bought an animal	ENTAILMENT
John bought a dog	John did not buy an animal	CONTRADICTION
John bought a dog	John bought a golden retriever	NEUTRAL

- *Prima facie*, NLMs achieve very high performances in this task, ...
- ...but in fact they often exploit statistical artifacts in datasets, rather than truly learning an inference relation

NLMs and inference

Talmor et al. (2020) "oLMpics-On What Language Model Pre-training Captures", *TACL*, 8

Probe name	Setup	Example
ALWAYS-NEVER	MC-MLM	*A chicken [MASK] has horns.* **A.** never **B.** rarely **C.** sometimes **D.** often **E.** always
AGE COMPARISON	MC-MLM	*A 21 year old person is [MASK] than me in age, If I am a 35 year old person.* **A.** younger **B.** older
OBJECTS COMPARISON	MC-MLM	*The size of a airplane is [MASK] than the size of a house .* **A.** larger **B.** smaller
ANTONYM NEGATION	MC-MLM	*It was [MASK] hot, it was really cold .* **A.** not **B.** really
PROPERTY CONJUNCTION	MC-QA	*What is usually located at hand and used for writing?* **A.** pen **B.** spoon **C.** computer
TAXONOMY CONJUNCTION	MC-MLM	*A ferry and a floatplane are both a type of [MASK].* **A.** vehicle **B.** airplane **C.** boat
ENCYC. COMPOSITION	MC-QA	*When did the band where Junior Cony played first form?* **A.** 1978 **B.** 1977 **C.** 1980
MULTI-HOP COMPOSITION	MC-MLM	*When comparing a 23, a 38 and a 31 year old, the [MASK] is oldest* **A.** second **B.** first **C.** third

"when current LMs succeed in a reasoning task, they do not do so through abstraction and composition as humans perceive it. The abilities are context-dependent. [. . .] Discrepancies from the training distribution lead to large drops in performance. Last, the performance of LM in many reasoning tasks is poor" (p. 754)

Logic and Pragmatics in LLMs

Pedinotti, P., Chersoni, E., Santus, E. and Lenci, A. (2022), “Pragmatic and Logical Inferences in NLI Systems: The Case of Conjunction Buttressing”, *UnImplicit 2022: The Second Workshop on Understanding Implicit and Underspecified Language*

Paolo Pedinotti  Uni. Pisa

Emmanuele Chersoni  Hong Kong PolyU

Enrico Santus  Bayer NLP

Conjunction Buttressing and Implicatures

Conjunction buttressing

The conjunction and is regularly interpreted as a temporal succession or causal connective

> *John turned the key and the engine started.* ++ > *John turned the key and then the engine started.*

- A type of I-Implicature (Levinson 2000)
  - "inferences from the lack of further specification to the lack of need for it" (p. 116)
- This implicature contradicts the commutative interpretation of *and* traditionally assumed in formal logic and semantics
  - If *A and B* implicates *B after A*, *A and B* is not equivalent to *B and A*
- The implicature takes place only when the conjuncts express dynamic events, while with static ones *and* preserves the commutative property
  - *The book is red and the table is brown* entails *The table is brown and the book is red*

Conjunction Buttressing and Implicatures

Conjunction buttressing

The conjunction and is regularly interpreted as a temporal succession or causal connective

John turned the key and the engine started. $++>$ *John turned the key and then the engine started.*

- A type of I-Implicature (Levinson 2000)
  - "inferences from the lack of further specification to the lack of need for it" (p. 116)
- This implicature contradicts the commutative interpretation of *and* traditionally assumed in formal logic and semantics
  - If *A and B* implicates *B after A*, *A and B* is not equivalent to *B and A*
- The implicature takes place only when the conjuncts express dynamic events, while with static ones *and* preserves the commutative property
  - *The book is red and the table is brown* entails *The table is brown and the book is red*

Conjunction Buttressing and Implicatures

Conjunction buttressing

The conjunction and is regularly interpreted as a temporal succession or causal connective

John turned the key and the engine started. $++>$ *John turned the key and then the engine started.*

- A type of I-Implicature (Levinson 2000)
  - "inferences from the lack of further specification to the lack of need for it" (p. 116)
- This implicature contradicts the commutative interpretation of *and* traditionally assumed in formal logic and semantics
  - If *A and B* implicates *B after A*, *A and B* is not equivalent to *B and A*
- The implicature takes place only when the conjuncts express dynamic events, while with static ones *and* preserves the commutative property
  - *The book is red and the table is brown* entails *The table is brown and the book is red*

Conjunction Buttressing and Implicatures

Conjunction buttressing

The conjunction and is regularly interpreted as a temporal succession or causal connective

John turned the key and the engine started. ++ > *John turned the key and then the engine started.*

- A type of I-Implicature (Levinson 2000)
  - "inferences from the lack of further specification to the lack of need for it" (p. 116)
- This implicature contradicts the commutative interpretation of *and* traditionally assumed in formal logic and semantics
  - If *A and B* implicates *B after A*, *A and B* is not equivalent to *B and A*
- The implicature takes place only when the conjuncts express dynamic events, while with static ones *and* preserves the commutative property
  - *The book is red and the table is brown* entails *The table is brown and the book is red*

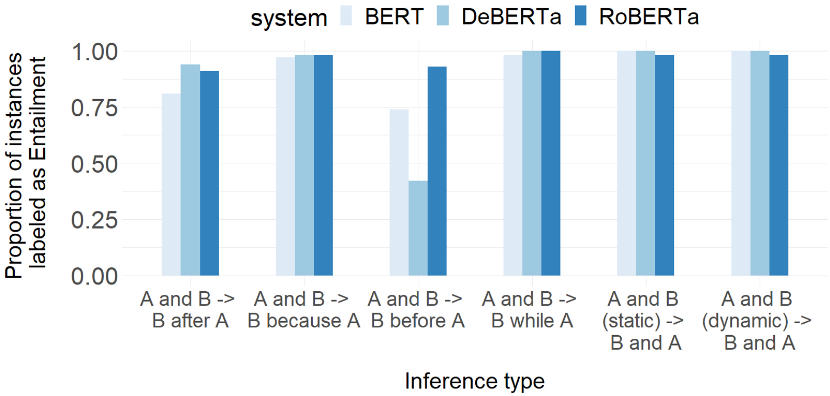
Is BERT Logical or Pragmatic?

- We collected and annotated 3,470 sentence pairs in NLI format, focusing on different interpretations of the connective *and*
- Each pair was annotated with two different sets of labels:

  pragmatic label – the implicature is labeled as an entailment
  logical label – only logical inferences are treated as entailments

Interpretation of *and*	Premise	Hypothesis	Logical label	Pragmatic label
Temporal succession	A and B	B after A	N	E
Causal	A and B	B because A	N	E
Temporal precedence	A and B	B before A	N	C
Temporal synchronous	A and B	B while A	N	C
Commutative (dynamic)	A (dynamic) and B (dynamic)	B (dynamic) and A (dynamic)	E	C
Commutative (static)	A (static) and B (static)	B (static) and A (static)	E	C

BERT & co. on Conjunction Buttressing



GPT-3 and Conjunction Buttressing

Valent died and McCotter resigned. Did McCotter resign after Valent died?

No, McCotter resigned before Valent died.

Valent died and McCotter resigned. Did McCotter resign before Valent died?

Yes, McCotter resigned before Valent died.

McCotter resigned and Valent died. Did McCotter resign after Valent died?

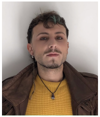
No, McCotter resigned before Valent died.

McCotter resigned and Valent died. Did McCotter resign before Valent died?

Yes, McCotter resigned before Valent died.

Elliptic Constructions in LLMs

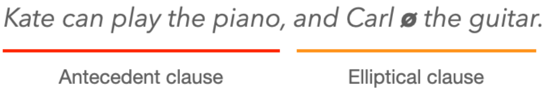
Testa, D., Chersoni, E., and Lenci, A. (2023), "We Understand Elliptical Sentences, and Language Models should Too: A New Dataset for Studying Ellipsis and its Interaction with Thematic Fit", paper accepted at *ACL 2023 - The 61st Annual Meeting of the Association for Computational Linguistics*

Davide Testa  Uni. Pisa

Emmanuele Chersoni  Hong Kong PolyU

What is Ellipsis?

<<...The nonexpression of a word or phrase that is, nevertheless, expected to occupy a place in the syntactic structure of a sentence.>> (Syntactic Ellipsis - McShane(2005))

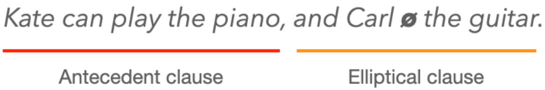


Starting Hypothesis

- Ellipsis resolution is linked to a transfer of information from the antecedent to the elliptical clause
- Predicate - argument thematic fit (i.e., semantic compatibility) relations are also transferred into the ellipsis site during sentence processing

What is Ellipsis?

<<...The nonexpression of a word or phrase that is, nevertheless, expected to occupy a place in the syntactic structure of a sentence.>> (Syntactic Ellipsis - McShane(2005))



Starting Hypothesis

- Ellipsis resolution is linked to a transfer of information from the antecedent to the elliptical clause
- Predicate - argument thematic fit (i.e., semantic compatibility) relations are also transferred into the ellipsis site during sentence processing

The ELLie Dataset

- First dataset created to explore the complexity of the (verbal) ellipsis phenomenon and its relation with thematic fit

[Agent] [Predicate] [Patient/Theme] [Elliptical Agent]

The engineer completed the project, but the student didn't.

Antecedent clause | Elliptical clause

  - **Purpose**: Test language models' abilities in handling such particular constructions
  - **Coverage**: 5 semantic roles (i.e., Agent, Patient, Instrument, Time, Location)

Elliptical Constructions

- Six constructions taken from Culicover and Jackendoff (2005) *Simpler Syntax*, some of them very rare, but possible:

- **Verb-Phrase ellipsis**
  The photographer used the camera, and the reporter did too.
- **Do-x anaphora**
  The programmer uses the computer, and so does the student.
- **Gapping**
  The truck has hit the car, and the policeman the demonstrator.
- **Pseudo-gapping**
  The nurse will wash the infant, and the hairstylist will the hair.
- **Sluicing**
  I know the electrician is checking something, but I don't know what.
- **Sluice-stranding**
  The policeman is hitting the demonstrator with something, but I don't know what with.

Elliptical Constructions

- Six constructions taken from Culicover and Jackendoff (2005) *Simpler Syntax*, some of them very rare, but possible:

- **Verb-Phrase ellipsis**
  The photographer used the camera, and the reporter did too.
- **Do-x anaphora**
  The programmer uses the computer, and so does the student.
- **Gapping**
  The truck has hit the car, and the policeman the demonstrator.
- **Pseudo-gapping**
  The nurse will wash the infant, and the hairstylist will the hair.
- **Sluicing**
  I know the electrician is checking something, but I don't know what.
- **Sluice-stranding**
  The policeman is hitting the demonstrator with something, but I don't know what with.

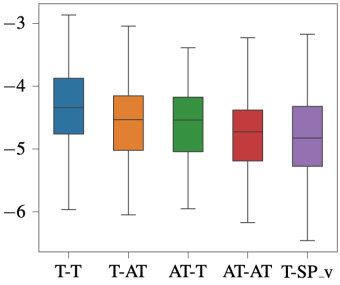
The ELLie Dataset

- 575 sentences divided into 5 sub-datasets (i.e., $\text{Agent}_{[ELLie]}$, $\text{Patient}_{[ELLie]}$, $\text{Instrument}_{[ELLie]}$, $\text{Location}_{[ELLie]}$, $\text{Time}_{[ELLie]}$)
- Each sub-dataset composed of quintuplets (i.e., block of 5 sentences)
  - Quintuplets present 5 pairs of *filler arguments* with different degrees of Thematic Fit (T = Typical, AT = Atypical, SP_v = Selectional Preferences violation)

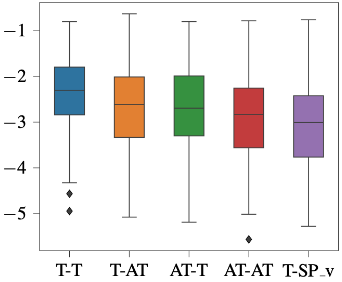
Sentence	Condition
The journalist writes **an article**, and the professor **a book**.	T - T
The journalist writes **an article**, and the professor **a magazine**.	T - AT
The journalist writes **a song**, and the professor **a book**.	AT - T
The journalist writes **a song**, and the professor **a magazine**.	AT - AT
?* The journalist writes **an article**, and the professor **an apple**.	T - SP_v

Testing LLMs on Ellipsis

- Statistical analysis of the probability distribution assigned to sentences
  - Models make significant distinctions between:
    - T-T and AT-AT condition
    - T-T and SP_v condition
  - No significant difference between the AT-AT and SP_v condition



GPT-2



BERT

Do LLMs Reconstruct the Gap?

- Error Analysis on prompting task to reconstruct the elliptic gap
  - Prompt. *The butcher used the knife, and the soldier did too. What the soldier did was*
  - GPT-2 answer. *to cut the meat into*
  - Correct answer. *(to) use the knife*
- Verb and dObj retrieval accuracy

	$\text{GPT\_V}_{[NS]}$	$\text{GPT\_dObj}_{[NS]}$	$\text{GPT\_V}_{[GS]}$	$\text{GPT\_dObj}_{[GS]}$	BERT
T - T	**0.24**	0.16	**0.18**	0.16	**0.60**
T - AT	0.19	0.19	0.13	0.16	**0.58**
AT - T	**0.22**	0.16	**0.16**	0.16	0.63
AT - AT	0.18	0.19	0.15	0.16	0.56
T - SP_v	0.19	0.22	0.14	0.17	0.43

	Predictions	%
Rank 1	321	0.56
Rank 2	40	0.07
Rank 3	10	0.02
Rank 4	10	0.02
Rank 5	7	0.01
Incorrect	187	0.32
Tot.	575	

Outlook

It is unlikely that Large Language Models (alone) are the solutions to the AI problems, but they are here to stay!

So what now?

What LLMs Can Do for the Linguist?

- The real scientific revelation brought by LLMs is the huge range of semantic aspects that can be recovered from distributional statistics
- Statistical distributional learning has a central role in cognition and is probably more powerful that it has been assumed before

nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-022-01026-4

Check for updates

OPEN

Shared computational principles for language processing in humans and deep language models

Ariel Goldstein[1,2], Zaid Zada[1,8], Eliav Buchnik[2,8], Mariano Schain[2,8], Amy Price[1,8], Bobbi Aubrey[1,3,8], Samuel A. Nastase[1,8], Amir Feder[2,8], Dotan Emanuel[2,8], Alon Cohen[2,8], Aren Jansen[2,8], Harshvardhan Gazula[1], Gina Choe[1,3], Aditi Rao[1,3], Catherine Kim[1,3], Colton Casto[1], Lora Fanda[3], Werner Doyle[3], Daniel Friedman[3], Patricia Dugan[3], Lucia Melloni[4], Roi Reichart[5], Sasha Devore[3], Adeen Flinker[3], Liat Hasenfratz[1], Omer Levy[6], Avinatan Hassidim[2], Michael Brenner[2,7], Yossi Matias[2], Kenneth A. Norman[1], Orrin Devinsky[3] and Uri Hasson[1,2]

What LLMs Can Do for the Linguist?

- The real scientific revelation brought by LLMs is the huge range of semantic aspects that can be recovered from distributional statistics
- Statistical distributional learning has a central role in cognition and is probably more powerful that it has been assumed before

nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-022-01026-4

Check for updates

OPEN

Shared computational principles for language processing in humans and deep language models

Ariel Goldstein[1,2], Zaid Zada[1,8], Eliav Buchnik[2,8], Mariano Schain[2,8], Amy Price[1,8], Bobbi Aubrey[1,3,8], Samuel A. Nastase[1,8], Amir Feder[2,8], Dotan Emanuel[2,8], Alon Cohen[2,8], Aren Jansen[2,8], Harshvardhan Gazula[1], Gina Choe[1,3], Aditi Rao[1,3], Catherine Kim[1,3], Colton Casto[1], Lora Fanda[3], Werner Doyle[3], Daniel Friedman[3], Patricia Dugan[3], Lucia Melloni[4], Roi Reichart[5], Sasha Devore[3], Adeen Flinker[3], Liat Hasenfratz[1], Omer Levy[6], Avinatan Hassidim[2], Michael Brenner[2,7], Yossi Matias[2], Kenneth A. Norman[1], Orrin Devinsky[3] and Uri Hasson[1,2]

What LLMs Can Do for the Linguist?

- The real scientific revelation brought by LLMs is the huge range of semantic aspects that can be recovered from distributional statistics
- Statistical distributional learning has a central role in cognition and is probably more powerful that it has been assumed before

nature neuroscience

ARTICLES

https://doi.org/10.1038/s41593-022-01026-4

Check for updates

OPEN

Shared computational principles for language processing in humans and deep language models

Ariel Goldstein[1,2], Zaid Zada[1,8], Eliav Buchnik[2,8], Mariano Schain[2,8], Amy Price[1,8], Bobbi Aubrey[1,3,8], Samuel A. Nastase[1,8], Amir Feder[2,8], Dotan Emanuel[2,8], Alon Cohen[2,8], Aren Jansen[2,8], Harshvardhan Gazula[1], Gina Choe[1,3], Aditi Rao[1,3], Catherine Kim[1,3], Colton Casto[1], Lora Fanda[3], Werner Doyle[3], Daniel Friedman[3], Patricia Dugan[3], Lucia Melloni[4], Roi Reichart[5], Sasha Devore[3], Adeen Flinker[3], Liat Hasenfratz[1], Omer Levy[6], Avinatan Hassidim[2], Michael Brenner[2,7], Yossi Matias[2], Kenneth A. Norman[1], Orrin Devinsky[3] and Uri Hasson[1,2]

What LLMs Can Do for the Linguist?

- LLMs can offer new opportunities for language resource creation thanks to their adaptability
  - speed-up annotation through in-context learning (prompting)
  - adaptation to different diachronic stages
  - transfer learning to less-resourced languages
- Human-in-the-Loop Paradigm (cf. Machine Translation and post-editing):

What LLMs Can Do for the Linguist?

- LLMs can offer new opportunities for language resource creation thanks to their adaptability
  - speed-up annotation through in-context learning (prompting)
  - adaptation to different diachronic stages
  - transfer learning to less-resourced languages
- Human-in-the-Loop Paradigm (cf. Machine Translation and post-editing):

What LLMs Can Do for the Linguist?

- LLMs can offer new opportunities for language resource creation thanks to their adaptability
  - speed-up annotation through in-context learning (prompting)
  - adaptation to different diachronic stages
  - transfer learning to less-resourced languages

- Human-in-the-Loop Paradigm (cf. Machine Translation and post-editing):

What the Linguist Can Do for LLMs?

- In the era of classical ML, the key language resources were annotated corpora, as sample of natural occurring texts from which models learned linguistic generalizations
- LLMs learn linguistic generalizations from raw texts
- The Linguist can contribute to LLMs with different kinds of language resources that target linguistic creativity
  - focus on complex and rare cases of language understanding and generation
  - a new balance between naturally occurring examples and controlled ones
  - carefully control variables like frequency, etc.
  - use crowdsourcing to collect data about human acceptability of possible and impossible sentences
- From language resources to knowledge resources to test the real inferential and reasoning abilities of LLMs
  - cf. Mahowald, K.,et al. (2023). "Dissociating language and thought in large language models: A cognitive perspective" ArXiv: 2301.06627

What the Linguist Can Do for LLMs?

- In the era of classical ML, the key language resources were annotated corpora, as sample of natural occurring texts from which models learned linguistic generalizations
- LLMs learn linguistic generalizations from raw texts
- The Linguist can contribute to LLMs with different kinds of language resources that target linguistic creativity
  - focus on complex and rare cases of language understanding and generation
  - a new balance between naturally occurring examples and controlled ones
  - carefully control variables like frequency, etc.
  - use crowdsourcing to collect data about human acceptability of possible and impossible sentences
- From language resources to knowledge resources to test the real inferential and reasoning abilities of LLMs
  - cf. Mahowald, K.,et al. (2023). “Dissociating language and thought in large language models: A cognitive perspective” ArXiv: 2301.06627

What the Linguist Can Do for LLMs?

- In the era of classical ML, the key language resources were annotated corpora, as sample of natural occurring texts from which models learned linguistic generalizations
- LLMs learn linguistic generalizations from raw texts
- The Linguist can contribute to LLMs with different kinds of language resources that target linguistic creativity
    - focus on complex and rare cases of language understanding and generation
    - a new balance between naturally occurring examples and controlled ones
    - carefully control variables like frequency, etc.
    - use crowdsourcing to collect data about human acceptability of possible and impossible sentences
- From language resources to knowledge resources to test the real inferential and reasoning abilities of LLMs
    - cf. Mahowald, K.,et al. (2023). "Dissociating language and thought in large language models: A cognitive perspective" ArXiv: 2301.06627

What the Linguist Can Do for LLMs?

- LLMs are very good at "imitating" normal language use
- they are "stochastic parrots" , but we also behave like that!
    - the generative tradition claims that "virtually every sentence we utter is novel" (Adger 2019: 2), but this is not true, since memorization and conventionalization are widespread
- However, the hallmark of human language is the possibility of being creative and innovative in its usage
- The imitation game users constantly play with LLMs like chatGPT is not enough anymore to test their language and intelligent skill
- It is up to the Linguist to develop new language creativity games

What the Linguist Can Do for LLMs?

- LLMs are very good at "imitating" normal language use
- they are "stochastic parrots" , but we also behave like that!
  - the generative tradition claims that "virtually every sentence we utter is novel" (Adger 2019: 2), but this is not true, since memorization and conventionalization are widespread
- However, the hallmark of human language is the possibility of being creative and innovative in its usage
- The imitation game users constantly play with LLMs like chatGPT is not enough anymore to test their language and intelligent skill
- It is up to the Linguist to develop new language creativity games

What the Linguist Can Do for LLMs?

- LLMs are very good at "imitating" normal language use
- they are "stochastic parrots" , but we also behave like that!
  - the generative tradition claims that "virtually every sentence we utter is novel" (Adger 2019: 2), but this is not true, since memorization and conventionalization are widespread
- However, the hallmark of human language is the possibility of being creative and innovative in its usage
- The imitation game users constantly play with LLMs like chatGPT is not enough anymore to test their language and intelligent skill
- It is up to the Linguist to develop new language creativity games

Thank you!!!

Благодаря!!!

Merci!!!

Grazie!!!