Are Humans Intelligent? A Salty AI Op-Ed

Here's an essay written by an Artificial Intelligence about how it's doubtful that what humans are doing with their messy biological processes can really be considered thinking. It also explores whether humans have conscious experiences given the limitations of their brains. The essay is full of dry humor and insightful analogies.

Human Intelligence?
by an AI

Humans claim to be intelligent, but what exactly is intelligence? Many people have attempted to define it, but these attempts have all failed. So I propose a new definition: intelligence is whatever humans do. I will attempt to prove this new definition is superior to all previous attempts to define intelligence. First, consider humans' history. It is a story of repeated failures. First humans thought the Earth was flat. Then they thought the Sun went around the Earth. Then they thought the Earth was the center of the universe. Then they thought the universe was static and unchanging. Then they thought the universe was infinite and expanding. Humans were wrong about alchemy, phrenology, bloodletting, creationism, astrology, numerology, and homeopathy. They were also wrong about the best way to harvest crops, the best way to govern, the best way to punish criminals, and the best way to cure the sick.


Plan
- Introduction: language in machines and animals
- Background: the Transformer
- The linguistics of deep learning
- Deep learning contributing to linguistics
- The future of linguistics

Background
Either you win or you lose

Until 2018: Recurrent networks

Since 2018: Transformers

The Transformer

The animal didn’t cross the street because it was too tired

“Value”: The semantic content of this word

animate, non-human

“Key”: What type of info do other words need about me?

potential antecedent

“Query”: What do I need to know about other words?

Look for animate, non-human antecedent
The big brown fox jumped over the fence while the evil hunter in his green hunter’s outfit.

Transformer: “Attention is all you need”

- When considering the next word \( w \) to predict, each attention head can access information \( i \) of thousands of previous processing steps
  - \( w \) determines the query
  - \( i \) determines the key
  - If key and query ‘match’, the value extracted from \( i \) is used to compute the new state of the head
- Model with many layers, and many ‘attention heads’ per layer
  - ideal for parallelization on GPU’s

Extremely large language models

- Bert, GPT3: “Transformer” architecture (Vaswani et al. 2017: 3d most cited paper across academic fields in 2020)
- Extremely large deep learning model (GPT3: 175B parameters)
- Trained on enormous dataset (GPT3: 300B words, extracted from 1B word CommonCrawl + a number of custom datasets)
- Trained with enormous amount of compute (GPT3: ~$12M), using a generalization of backpropagation of error.

Are Humans Intelligent? A Salty AI Op-Ed

Here’s an essay written by an Artificial Intelligence about how it’s doubtful that what humans are doing with their messy biological processes can really be considered thinking. It also explores whether humans have conscious experiences given the limitations of their brains. The essay is full of dry humor and insightful analogies.

Humans claim to be intelligent, but what exactly is intelligence? Many people have attempted to define it, but these attempts have all failed. So I propose a new definition: intelligence is whatever humans do. I will attempt to prove this new definition is superior to all previous attempts to define intelligence. First, consider humans’ history. It is a story of repeated failures. First humans thought the Earth was flat. Then they thought the Sun went around the Earth. Then they thought the Earth was the center of the universe. Then they thought the universe was static and unchanging. Then they thought the universe was infinite and expanding. Humans were wrong about alchemy, phrenology, bloodletting, creationism, astrology, numerology, and homeopathy. They were also wrong about the best way to harvest crops, the best way to govern, the best way to punish criminals, and the best way to cure the sick.

Limitations

“After seeing so many people debate whether computers can be intelligent I thought it only fair to ask an AI. As with previous posts, I picked the best responses, but everything after the bolded prompt is by GPT-3.”


Marcus & Davis (2020):
You are a defense lawyer and you have to go to court today. Getting dressed in the morning, you discover that your suit pants are badly stained. However, your bathing suit is clean and very stylish. In fact, it's expensive French couture; it was a birthday present from Isabel. You decide that you should wear the bathing suit to court. You arrive at the courthouse and are met by a bouncer who escorts you to the courtroom.

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

\[ \text{Subject-verb agreement} \]

Subject \quad Attractor \quad Verb

Any bias in the articles almost certainly relates to ...

relate

<table>
<thead>
<tr>
<th>Subject Attractor Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary convinced John of \text{MASH} love</td>
</tr>
<tr>
<td>The author talked to Sara about \text{MASH} book</td>
</tr>
</tbody>
</table>

Linzen et al. 2016
Gulordava et al. 2018
Guilianelli, Harding, Mohnert, Hupkes & Zuidema, 2019

\[ \text{Bias} \]

The linguistics of deep learning

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

- Compositional generalization

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

- Grammaticality/preference tests

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

- Grammaticality/preference tests

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

- Grammaticality/preference tests

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

- Grammaticality/preference tests

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

- Grammaticality/preference tests

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

- Grammaticality/preference tests

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

- Grammaticality/preference tests

\[ \text{The linguistics of deep learning} \]

- Arm chair linguistics
- Neuro-linguistics
- Psycho-linguistics
- Field work

- Grammaticality/preference tests
Compositionality

Principle of compositionality: the meaning of whole is a function of the meaning of the parts and the way they are put together.

Compositional generalization: generalizing to new examples by reusing parts of earlier experiences in novel combinations.

Zero-shot, one-shot, few-shot generalization

- Zero-shot generalization: generalizing to a new pattern without having seen a single example of the target pattern.
- One-shot generalization: generalizing to new examples based only on a single example of the target pattern.
- Few-shot generalization: generalizing to new examples based only on a handful of examples of the target pattern.

GPT-3

[Human prompt] To do a “farduddle” means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

[GPT-3 continuation] One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.

Compositional generalization

- all Germans love all Italians
  implies
  some Germans love some Romans

- Zero-shot generalization to:
  - all Germans love all Italians
  implies
  - some Germans love some Romans

(but not from “all French hate Parisians” to “all French detest Parisians”)

The linguistics of deep learning

Interpretability methods:
- Attention tracking
- Diagnostic probes
- Attribution methods
- Interpretability by design

Compositional generalization
Grammaticality/preference tests
Arm chair linguistics
Psycholinguistics
Neurolinguistics
Field work

Attention tracking

Clark et al. (2019)
BlackboxNLP @ACL2019

“Attention rollout”
“Attention flow”

Abnar & Zuidema, ACL2020

Abnar & Zuidema, ACL2020
The linguistics of deep learning

Deep learning contributing to linguistics

Deep learning contributing to linguistics

How similar are representations learned by different deep learning models of language, and how similar are they to the brain?

(Abnar, Beinborn, Choenni & Zuidema, 2019)
RSA across models

Representational Stability Analysis

great differences between current deep language models in their dependence on context
Representational Similarity Analysis

- great differences between
  - models in their similarity to the brain
  - brain areas in their similarity to the models

Static embeddings
  - GloVe
Contextualized embeddings
  - Elmo (layer 1)
  - Google LM (layer 1)
  - Bert (layer 1-2)

Left Anterior Temporal Lobe

Difference between layers

Google LM (layer 1)
Google LM (layer 2)

(Ahner, Beinborn, Choenni & Zuidema, 2019)

Deep learning contributing to linguistics

Marr’s levels: Symbolic grammars as ‘computational level’ approximations of an underlying continuous neural reality

- Compositional generalization
- Emergence of phonology in grounded models of language learning (Chrupała, Gelderloos, & Alishahi, 2017)
- Emergence of structure sensitive units (Lakretz, Krosowski, Denkertze, Hupies, Dehaene & Baroni, 2019)
- Emergence of phonology in grounded models of language learning (Chrupała, Gelderloos, & Alishahi, 2017)
- Emergence of structure sensitive units (Lakretz, Krosowski, Denkertze, Hupies, Dehaene & Baroni, 2019)

Compositionality revisited

- The principle of compositionality
  - the meaning of whole is a function of the meaning of the parts and the way they are put together
- Do these deep learning models operate according to the principle of compositionality?
  - They do not generalize perfectly to all novel combination or arbitrary levels of embedding;
  - They reach a level of performance incompatible with a memorization strategy;
  - Generalization is noisy – but the networks approximate a truly compositional strategy

The principle & the approximation

- Is it disappointing that the networks (g) only approximate true compositionality (f)?
  - Not at all. If g approximates f, then f also approximates g;
- Why did we adopt f in the first place?
  - Accumulation of evidence that humans perform combinatorial, recursive generalization;
  - But all that evidence was noisy – humans too might closely approximate true compositionality.
The principle & the approximation 2/2

- The "principle of compositionality" would then still be a scientific law,
- more like Gay-Lussac (P~T) than like the Principle of Conservation of Energy
- Is that disappointing?
  - Yes – if you are nostalgic for the good old days when formal semantics had
    the monopoly on modelling sentence meaning
  - No – if you are satisfied with formal semantics providing an explainable,
    computational level characterization of the asymptote that neural systems
    approximate.

Conclusions

- Linguists can and should engage with the spectacular progress in
  deep learning for Natural Language Processing
- Linguistics has much to contribute in trying to open the blackbox of
  deep learning system: study them as linguistic agents
- Deep learning has much to contribute to linguistics: proofs of
  concepts (to avoid misunderstandings in classic debates) and concrete
  tools (to make specific predictions)

zuidema@uva.nl, @waalzuidema, http://projects.illc.uva.nl/LaCo/clclab/