People in parameter space: Finding meaning in machine learning models

TACCSTER 2021

Katrin Erk

University of Texas at Austin

Artificial intelligence and linguistics

- Immense advances in machine learning in recent years
- Big improvements in many areas
- Natural language processing: big improvements too, but natural language understanding remains difficult

- How can we use machine learning better?
- What does this tell us about language?

Machine learning for computational linguistics

Natural language processing tasks, like sentiment analysis:
given a review,
determine positive
or negative sentiment.
Learn this from training data: text plus known sentiment.

screenshots from https://www.rogerebert.com/

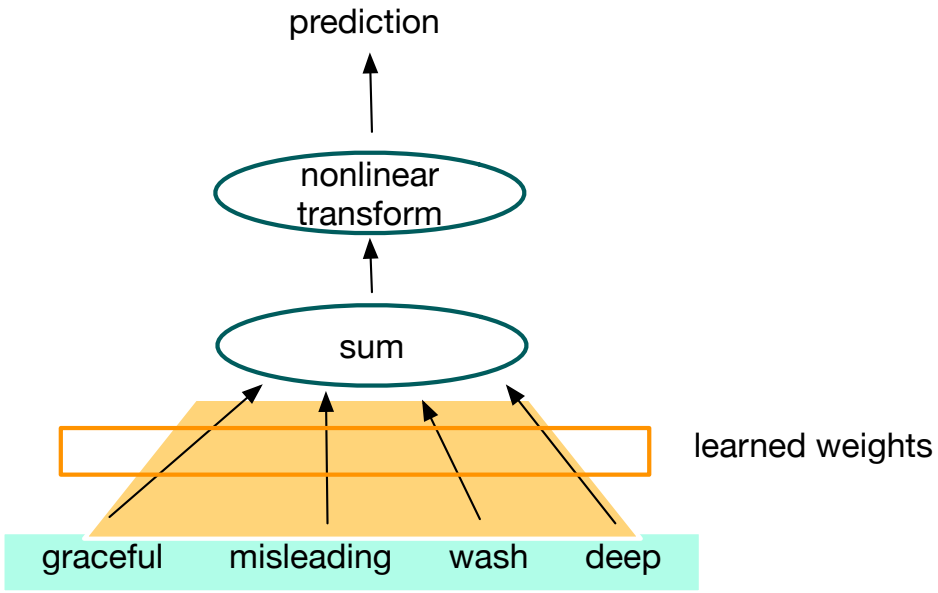
predict this

from that

There is a graceful ease to Mia Hansen-Løve's cinematic prose, one that can feel misleadingly simple at times. But once you allow her placid beats wash over you, the intricacy of her ideas rises to the surface with little effort, revealing the deep thinker and feeler Hansen-Løve always has been. Just think of "Eden" and the serenity

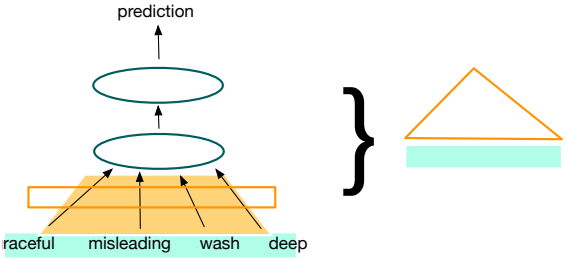
Machine learning for computational linguistics

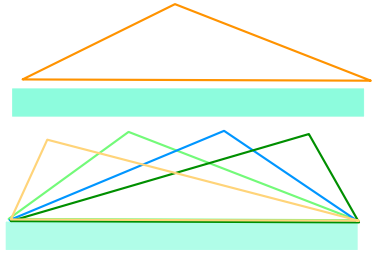
Learning from training data: Learn to assign weights to features of the input

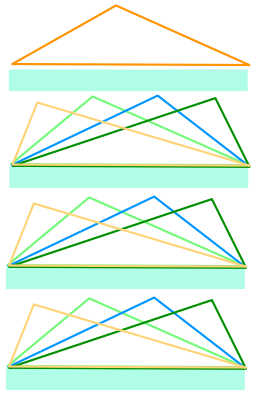


Recent machine learning advances in artificial intelligence

- **Powerful idea #1: deep models**







Make many weighted combinations of features:
Learn you own feature combinations.
Then combine those.

And repeat…

Universal function approximator: Given function inputs and outputs, approximate the function

Recent machine learning advances in artificial intelligence

- **Powerful idea #2: attention**
  - Given the current piece of data, how much should I take others into account?
  - Learned **attention weights**
  - Example: machine translation

- Recent model architecture: **Transformers**
  - Read many inputs in parallel
  - Combine and re-combine based on attention weights

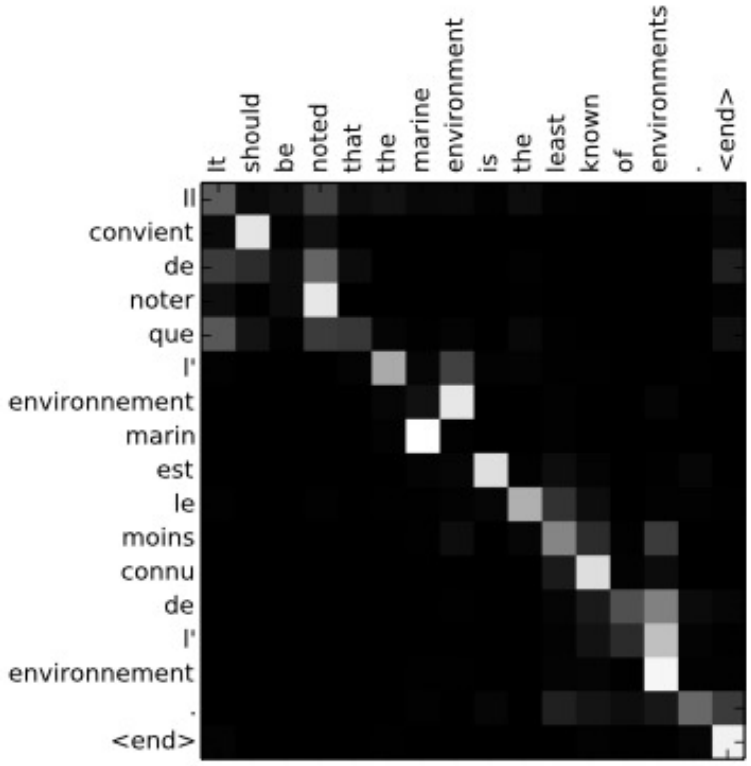


image from: Bahdanau, Cho, Bengio ICLR 2015

Recent machine learning advances in artificial intelligence

- **Powerful idea #3: pre-training**
  - "We just want your weights!"
  - Train a model on a general language task: What is the missing word?

    Hi how are _____?
  - Learned weights: **embeddings** of the input data into a weight space
  - Use the learned embeddings on a different task:
    machine translation, question answering, …



The many hats of the computational linguist

- **Building language technology**

  For example: How can we use machine learning for better question answering?

- **Understanding how language works**

  Build a computational model of human language understanding:
  Does it match human reactions?

- **Probing texts**

  With machine learning, extract statistical patterns from text.
  For example: What emotions can you find in tweets right after a hurricane?
  (Desai/Caragea/Li 2020)

People in parameter space

- Machine learning for language technology today: statistical patterns in text

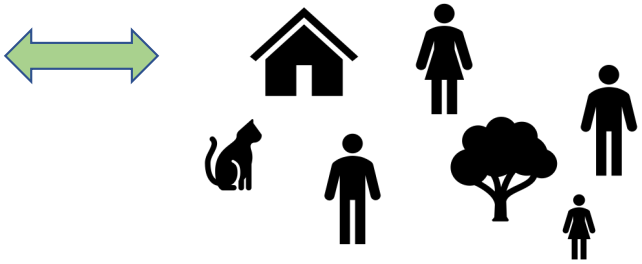
> Holmes was certainly not a difficult man to live with. He was quiet in his ways, and his habits were regular. It was rare for him to be up after ten at night, and he had invariably breakfasted and gone out before I rose in the morning. Sometimes he spent his day at the chemical laboratory, sometimes in the dissecting-rooms, and occasionally in long walks, which appeared to take him into the lowest portions of the City. Nothing could exceed his energy when the working fit was upon him; but now and again a reaction would seize him, and for days on end he would lie upon the sofa in the sitting-room, hardly uttering a word or moving a muscle from morning to night. On these occasions I have noticed such a dreamy, vacant expression in his eyes, that I might have suspected him of being addicted to the use of some narcotic, had not the temperance and cleanliness of his whole life forbidden such a notion.

→ prediction

- When a human reads a text, they know it is about people, and things, and events

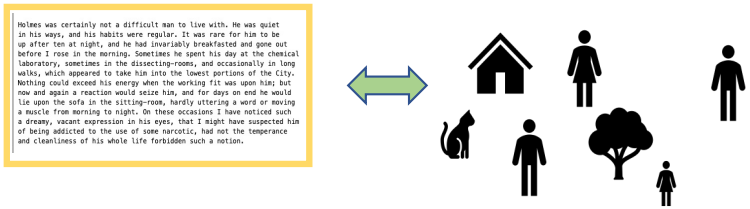
> Holmes was certainly not a difficult man to live with. He was quiet in his ways, and his habits were regular. It was rare for him to be up after ten at night, and he had invariably breakfasted and gone out before I rose in the morning. Sometimes he spent his day at the chemical laboratory, sometimes in the dissecting-rooms, and occasionally in long walks, which appeared to take him into the lowest portions of the City. Nothing could exceed his energy when the working fit was upon him; but now and again a reaction would seize him, and for days on end he would lie upon the sofa in the sitting-room, hardly uttering a word or moving a muscle from morning to night. On these occasions I have noticed such a dreamy, vacant expression in his eyes, that I might have suspected him of being addicted to the use of some narcotic, had not the temperance and cleanliness of his whole life forbidden such a notion.

In linguistics: Model-theoretic semantics (Montague 1970)



People in parameter space

- Representing people in machine learning models
  - Can we do better in language technology tasks?
  - How would we build a computational knowledge that humans have about other people in their minds?
  - Say we compute people-embeddings from text. What do they tell us about the texts?



Study 1: Cheng and Erk, Attending to entities for better text understanding, AAAI 2020

Can a model understand text better when it is aware of entities?

LAMBADA dataset (Paperno et al 2016): Really hard cases of missing words

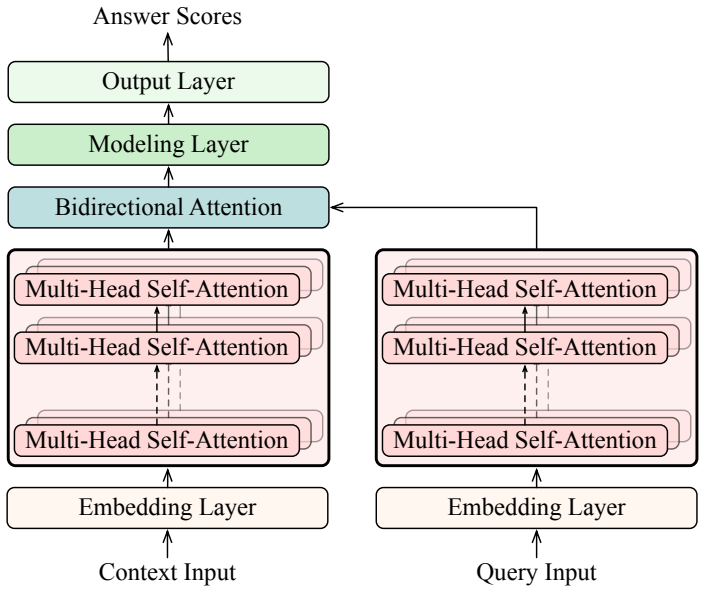
Context: "By the way, Elizabeth asked if I'd seen you," Tony lied. He wanted Jon to leave so he could talk with Ezekiel alone. There was something that aunt Casey, Patella and Gabriella had said about Tom that had bothered him ever since meeting Ezekiel earlier that afternoon.

Target sentence: "I'm sure she'll find me," Jon remarked curtly, trying to cut short the conversation with ___ .

Target word: Tony

Study 1: Cheng and Erk, Attending to entities for better text understanding, AAAI 2020

- Tell the model where attention should be high:
  - When reading “Tony”, pay attention to the underlined “he” and “him”



Context: "By the way, Elizabeth asked if I’d seen you," Tony lied. He wanted Jon to leave so he could talk with Ezekiel alone. There was something that aunt Casey, Patella and Gabriella had said about Tom that had bothered him ever since meeting Ezekiel earlier that afternoon.

Target sentence: "I’m sure she’ll find me," Jon remarked curtly, trying to cut short the conversation with ___ .

Target word: Tony

Study 1: Cheng and Erk, Attending to entities for better text understanding, AAAI 2020

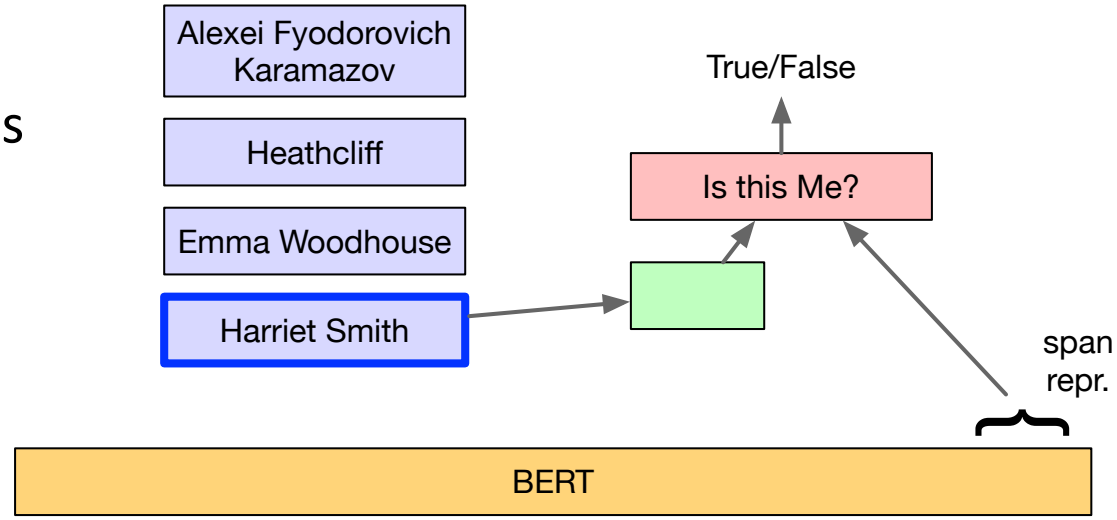
- Result:
  - Strong improvement over base model (+4.5 pts. F-score)
  - Performance matches huge text-pattern model with 1000x as many parameters
  - Improvement particularly when missing word is a pronoun (he, she, …)
- Follow-up study: Can entity information also help a Transformer model?
  - Many more parameters (billions)
  - No significant improvement
  - Transformers do better on most tasks – but how can we combine them with other information?

Study 2: Holgate and Erk, “Politeness, you simpleton,” retorted [MASK]. IWCS 2021

- What is a good computational representation of all we know about a person?
- How can machine learning models deal with really long texts?
- How can we combine Transformers (text statistics) with other information? Here: representations of individuals

Study 2: Holgate and Erk, “Politeness, you simpleton,” retorted [MASK]. IWCS 2021

- Can we build up “people embeddings” over long stretches of text?
- 62 classic novels: Austen, Dostoyevsky, Dumas, Carroll, …
- Model reads books front to back
- Build person embeddings from a task: When a person is mentioned, guess who.
- Base representation of text: Transformer



[MASK], in fact, was one of the few people who could see faults in [MASK].

Study 2: Holgate and Erk, “Politeness, you simpleton,” retorted [MASK]. IWCS 2021

- Results: good accuracy, 74%
- Stop updating person embeddings after 1/3 of novel:
  accuracy drops to 57%
- Different genre, Wikipedia novel summaries: acc. just above chance
- Training on what a person says, rather than what they do:
  accuracy rises several points
  - “This is not to be borne. Miss Bennet, I insist on being satisfied. Has he, has my nephew, made you an offer of marriage?” (Lady Catherine, Pride and Prejudice)
- Next up: CSI transcripts: Can the model guess the murderer?

People in parameter space

- Large questions, which we study on fun data:
  - What is a good computational representation of all we know about a person?
    - What can we infer from such a representation?
  - How can machine learning models deal with really long texts?
  - How can we combine Transformers (text statistics) with other information? Here: representations of individuals