Natural Language Understanding

CSC401/2511 – Natural Language Computing – Winter 2023

University of Toronto

Logistics

- A3 is released.
  - The lectures this week and next week will cover most materials for doing Assignment 3.

  - In the test() function for classifier.py: change args.use_cuda to args.use_gpu.

  - To ask questions: please post on piazza, or email the A3 team at csc401-2023-01-a3@cs.toronto.edu.

UNIVERSITY OF TORONTO

Contents

- What is NLU?
- Quantitative testing of language understanding.
- Building AI systems that “understand” languages.

NLU of AI systems

Natural language understanding (NLU) is a metaphorical term.

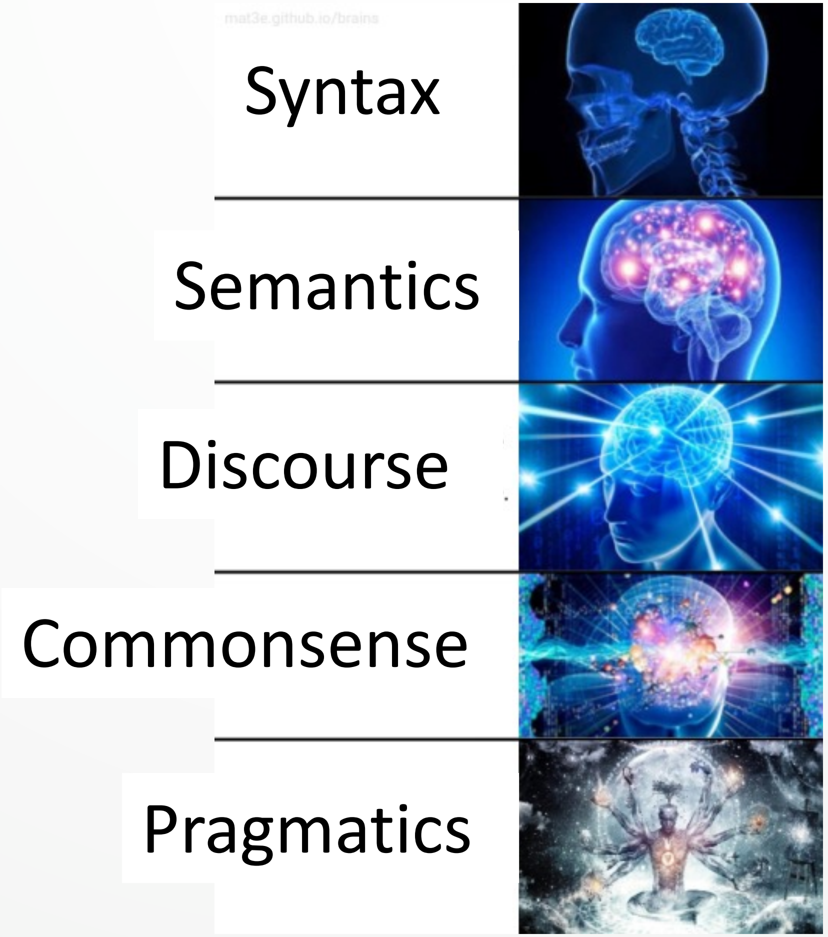
- NLU is a framework that studies AI systems similar to humans.
  - Humans can understand languages. There is rich resource regarding:
  - (1) the taxonomy of human language understanding, and
  - (2) how to evaluate human language understanding.

- NLU is considered a holy grail problem of NLP.

“Hierarchies” of understanding

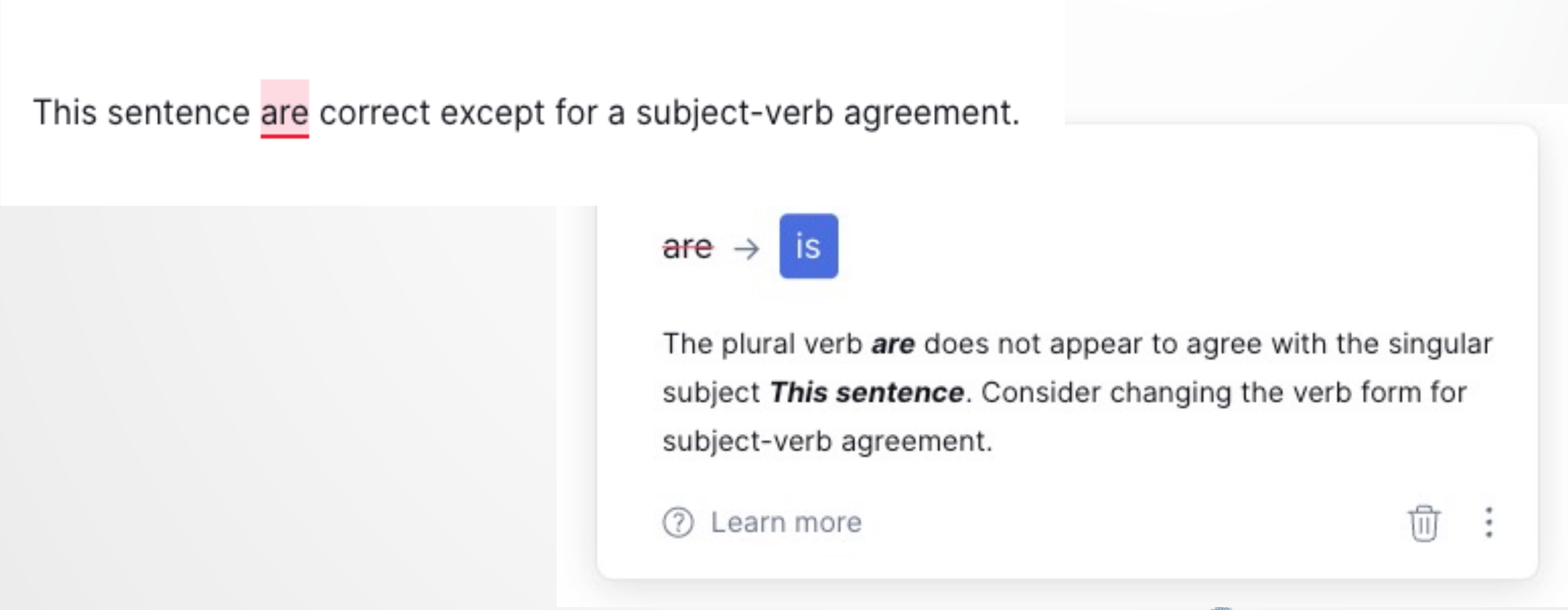
Humans can understand languages on multiple “levels”.

- These levels are *not* mutually exclusive!
- Many NLP tasks require understanding at multiple levels (e.g., translation).
- These levels are here for **organizing** the tasks.



Syntax: sequence tagging

- Example: grammatical error detection / correction.



Semantics: vector embedding

- Machines operate on vectors.
- Vector representations (embeddings) are key to the "understanding" of the meanings.

- Example: Quora question pair
  - Are the two questions paraphrases of each other?

Understand the discourse

- Example: A movie review on IMDB.

This is not a movie for fans of the usual eerie Lynch stuff. Rather, it's for those who either appreciate a good story, or have grown tired of the run-of-the-mill stuff with overt sentimentalism [...]

The story unfolds flawlessly, and we are taken along a journey that, I believe, most of us will come to recognize at some time. A compassionate, existentialist journey where we make amends for our past when approaching ourt inevitable demise.

Acting is without faults, cinematography likewise (occasionally quite brilliant!), and the dialogue leaves out just enough for the viewer to grasp the details od the story.

A warm movie. Not excessively sentimental.

Understand the discourse

- What can we attribute the polarity of this review to?

This is not a movie for fans of the usual eerie Lynch stuff. Rather, it's for those who either appreciate a good story, or have grown tired of the run-of-the-mill stuff with overt sentimentalism […]

The story unfolds flawlessly, and we are taken along a journey that, I believe, most of us will come to recognize at some time. A compassionate, existentialist journey where we make amends for our past when approaching ourt inevitable demise.

Acting is without faults, cinematography likewise (occasionally quite brilliant!), and the dialogue leaves out just enough for the viewer to grasp the details od the story.

A warm movie. Not excessively sentimental.

Understand the discourse

- Discourse structure of the movie review.

This is not a movie for fans of the usual eerie Lynch stuff. Rather, it's for those who either appreciate a good story, or have grown tired of the run-of-the-mill stuff with overt sentimentalism [...]

The story unfolds flawlessly, and we are taken along a journey that, I believe, most of us will come to recognize at some time. A compassionate, existentialist journey where we make amends for our past when approaching ourt inevitable demise.

Acting is without faults, cinematography likewise (occasionally quite brilliant!), and the dialogue leaves out just enough for the viewer to grasp the details od the story.

A warm movie. Not excessively sentimental.

Discourse structure in Abstract

NLP tasks, such as question-answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset – matching or exceeding the performance of 3 out of 4 baseline systems without using the 170,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

Language Models are Unsupervised Multitask Learners (Radford et al., 2019)

UNIVERSITY OF TORONTO

Discourse structure in Abstract

NLP tasks, such as question-answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. *When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset – matching or exceeding the performance of 3 out of 4 baseline systems without using the 170,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text.* These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations.

Problem

Our solution

Our system's performance

Significance

Language Models are Unsupervised Multitask Learners (Radford et al., 2019)

UNIVERSITY OF TORONTO

Understand the commonsense

- AI systems should have some “commonsense”.
  - Broadly speaking, commonsense is everything that is outside of the context.

- Example: Winograd Schema Challenge.
  - The elephant is too large to fit into the fridge because it is too big. What is too big?
  - The elephant is too large to fit into the fridge because it is too small. What is too small?

UNIVERSITY OF TORONTO

Understand the commonsense

- A popular AI system, GPT-3 davinci-002, shows some evidence of understanding on commonsense:
  - *Prompt* the model with a question. Let it generate.

- *An elephant can't be put into the fridge because it is too large. What is it?* A fridge is a household appliance that is used to store food and keep it fresh. An elephant is a land mammal that is too large to fit inside a fridge.

Understand the pragmatics

- Same sentence, different effects to different audience.
  - Example: High-context culture vs. low-context culture.

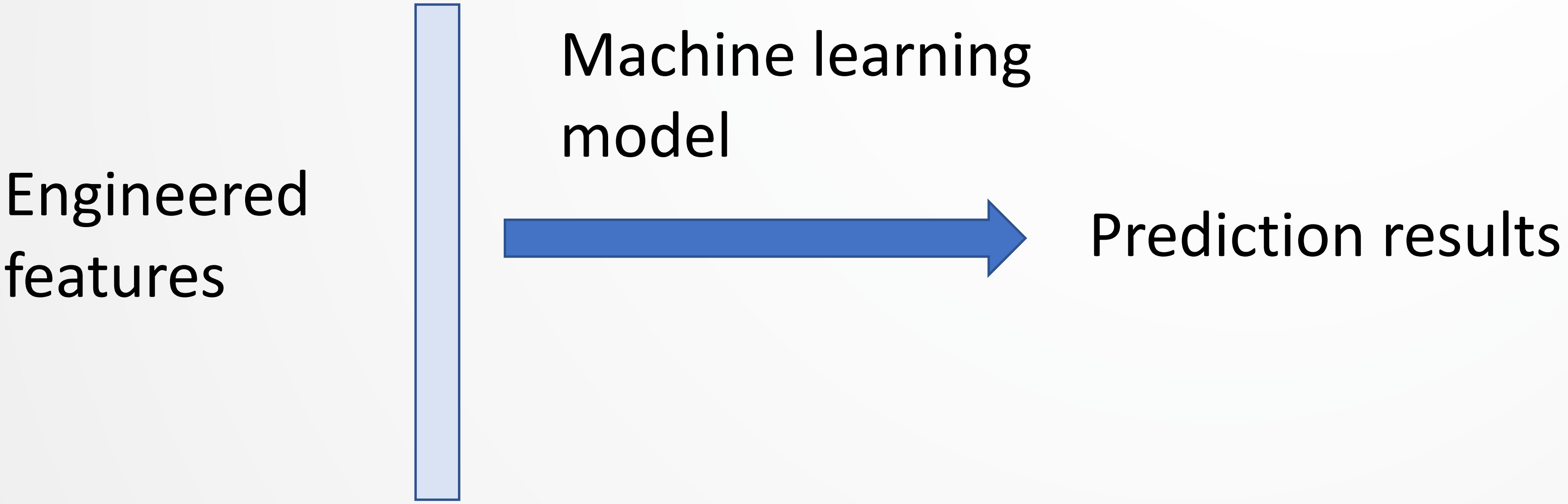
- Example: Identify political stance on Reddit. (Assignment 1)
- Aside: playing Diplomacy
  - https://ai.facebook.com/research/cicero/

Building AIs for NLU

- Here we will describe three types of systems. For each system, we will discuss:
  - **How well** they can understand languages.
  - **Why** they can understand languages.
  - And where do they usually **fail**.

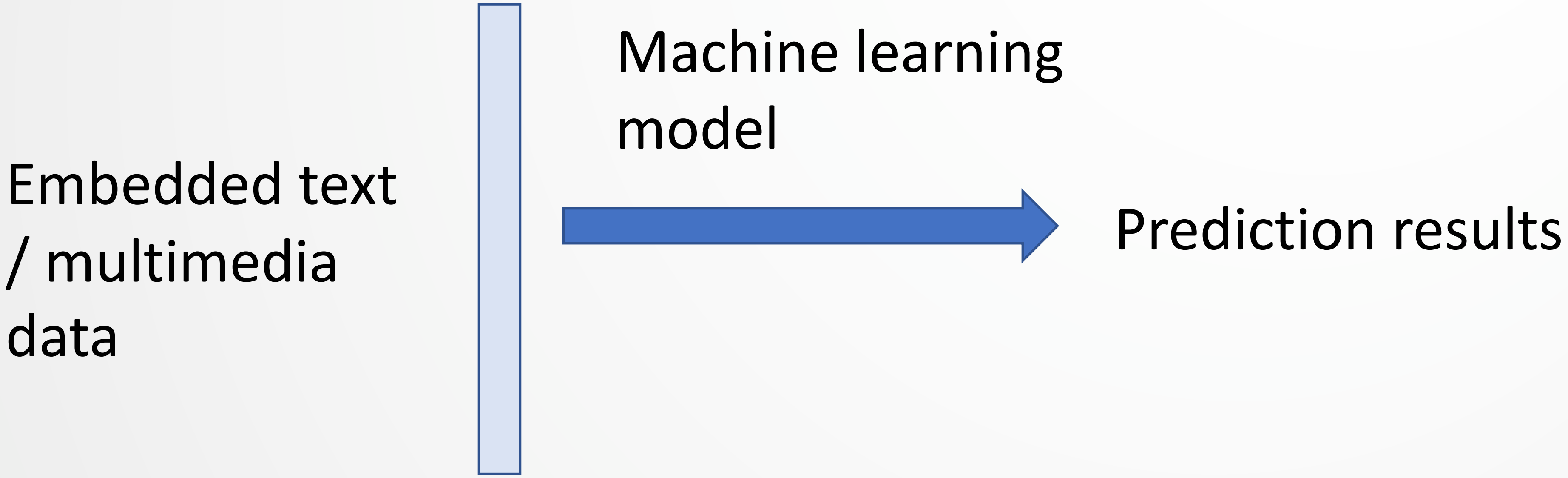
- Then, we will compare these systems.

Feature-based NLU



"How much artificial work in the engineered features, how much intelligence there is."

Embedding-based NLU



Embedding handles the low-level features.

[A Unified Architecture for NLP](). Collobert & Weston (2008)

DNN-based NLU

Pre-trained

Language Model

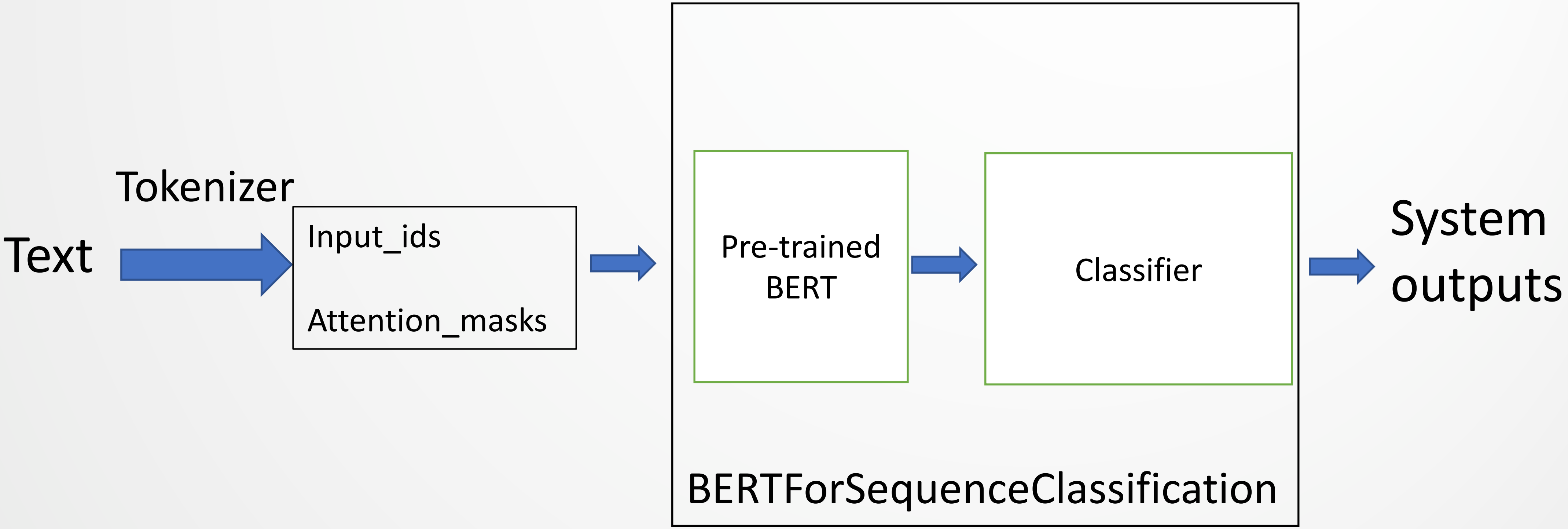
Downstream task data → Prediction results

Language modeling requires language understanding.

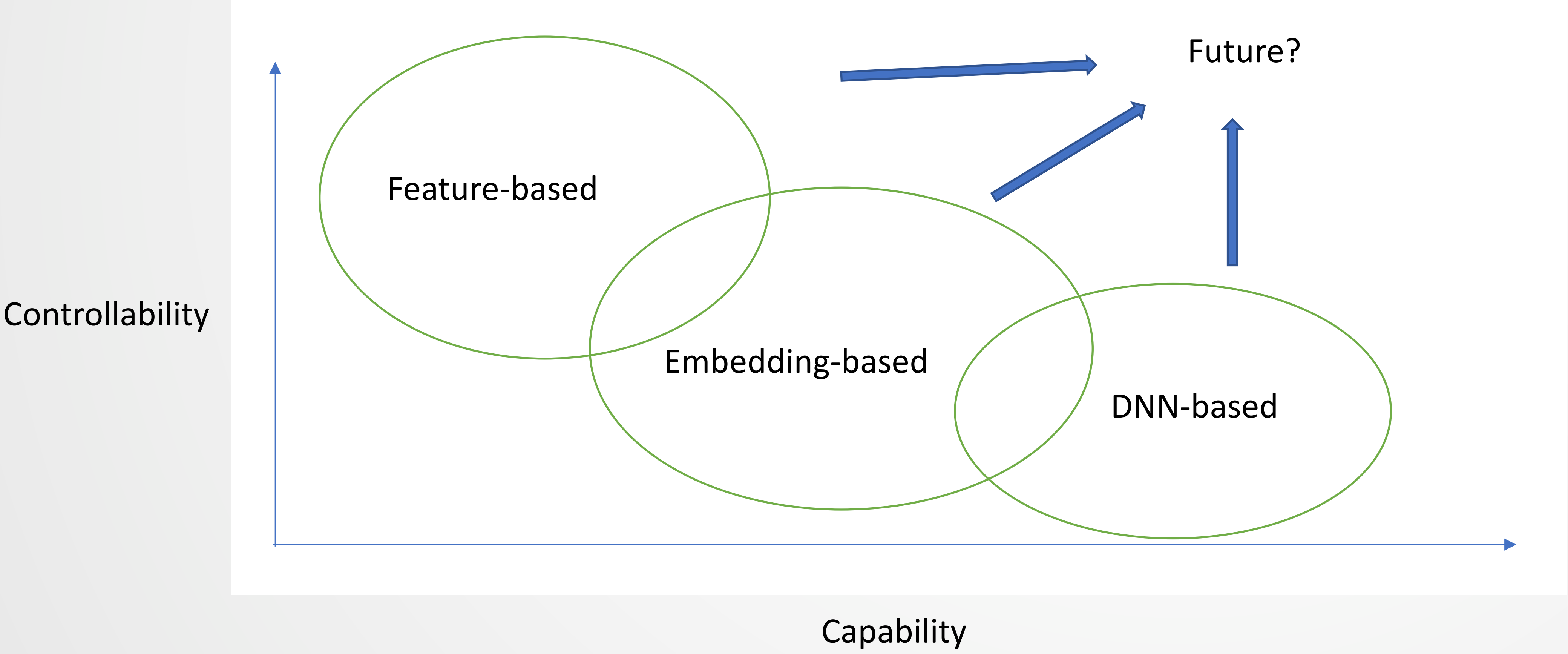
[Deep contextualized word representations](). Peters et al., (2018)

BERT for Sequence Classification

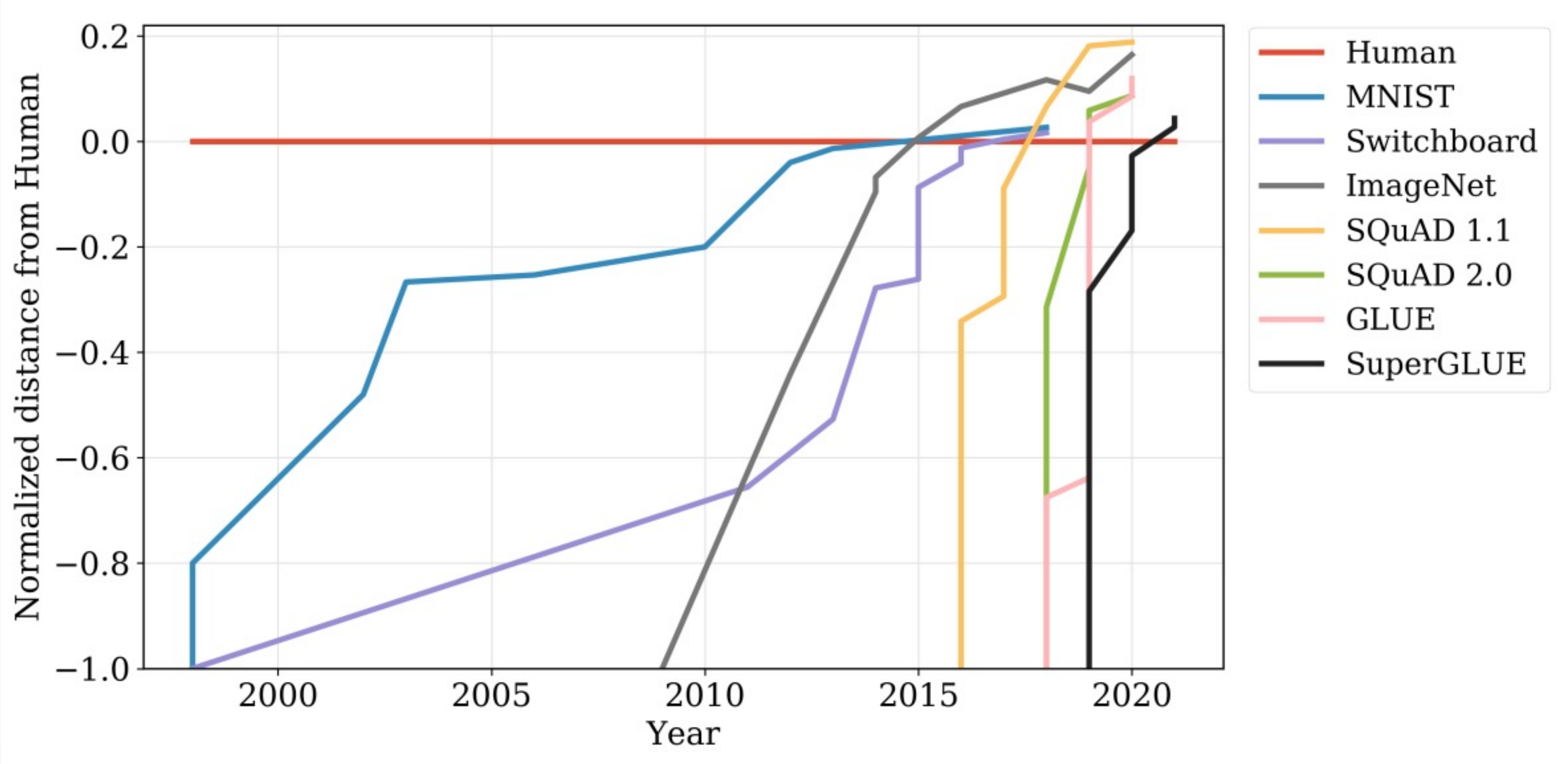
- A question in Assignment 3



Comparing these three systems



The benchmark saturation problem



Dynabench: Rethinking Benchmarking in NLP (Kiela et al., NAACL 2021)

Aside: Simulator Hypothesis

- An AI that can predict what happens next based on what happens now in a system is effectively a **simulator** of the system.
- Read more at this blog post.

Does that sound similar to the language model?

Aside: Mary's Room

- Mary is a brilliant scientist who lives in a room with only black and white colors.
- She knows all text in the world.

- One day she decides to step out of the room.
- Will she learn something new?

NLU needs to ground onto the real world!

Lecture review questions

By the end of this lecture, you should be able to:

- Describe NLU.
- Describe an NLU system in each of the three categories:
  - Feature-based
  - Embedding-based
  - DNN-based
- Identify a deployment scenario where each category of NLU system is suitable.

Anonymous feedback form: https://forms.gle/W3i6AHaE4uRx2FAJA