Course introduction

CS 685, Fall 2025

Advanced Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs685_f25/

Brendan O'Connor

College of Information and Computer Sciences

University of Massachusetts Amherst

- "Advanced" Natural Language Processing
  - Overview of key methods and approaches for computers to understand and generate ***human natural language***
  - Main focus: Large Language Models
  - (LLMs are a huge topic now; many, many LLM topics are out of scope.)

- Why do ***you*** want to take this course?

- Language is uniquely human and interesting!
- It's optimized for communication
- It's high-dimensional and discretely infinite

- Is language modeling all of AI?

Course logistics

- https://people.cs.umass.edu/~brenocon/cs685_f25/
- Follow along w/ the lectures in-person (or Zoom)
  - Zoom will be best-effort, no guarantee
  - Recordings will be posted to Piazza Resources page when available

who?

Your TAs!:
Nguyen Tran
Marisa Hudspeth
Rohan Pandey

email all of us (including me!) at
cics.685.instructors@gmail.com
use for signup issues right now

course website:
https://people.cs.umass.edu/~brenocon/cs685_f25/

Office hours (in-person and on zoom)

On course webpage; watch for updates on Piazza

If necessary, TA office hours may be extended during homework or exam weeks

Office hours will begin next week

Readings

- No need to buy any textbooks!
- Readings will be provided as PDFs on website
    - Book chapters from Jurafsky and Martin's online textbook
    - Otherwise, NLP research papers / notes

Questions / comments?

- Submit questions/concerns/feedback to Piazza
- FAQ
    - does this course require prior knowledge of NLP? *No, but basic ML/probability/stats/programming will help a lot*
    - Size of final project groups? *2-3*
    - Will we have notes? *Slides will be posted after lecture*

No official prereqs, but the following will be useful:

- comfort with programming
  - We’ll be using Python (and PyTorch) throughout the class
- comfort with probability, linear algebra, and mathematical notation
- Some familiarity with matrix calculus
- Excitement about language!
- Willingness to learn

Please brush up on these things as needed!

Previous class videos / material

- Spring 2024: https://people.cs.umass.edu/~miyyer/cs685/
  - Feel free to use these materials / videos to study!
- Different versions over the years available from https://nlp.cs.umass.edu/courses/

Course grade is based on

- https://people.cs.umass.edu/~brenocon/cs685_f25/grading.html

- Quizzes / exercises
- Problem sets (hw1, hw2, hw3)
    - Written: math & concept understanding
    - Programming: in Python
- Midterms
    - **Two** midterms
- Final projects
    - Groups of 2-3
    - Choose (propose!) any topic you want
    - Project proposal (earlier in semester)
    - Final project report (end of semester)

Extra credit

- We may have extra credit opportunities based on writing up research talks at UMass during the semester

Homework

- Strongly recommend to do the homework by yourself
    - You can use LLMs to help you do the homework
        - Please provide all of your prompts that allow us to reproduce your answer

- Plagiarism
    - If we find that your answers are the same or very similar to those of other people, we might report your behavior
        - e.g., copying from others or from last year's homework

Late Policy

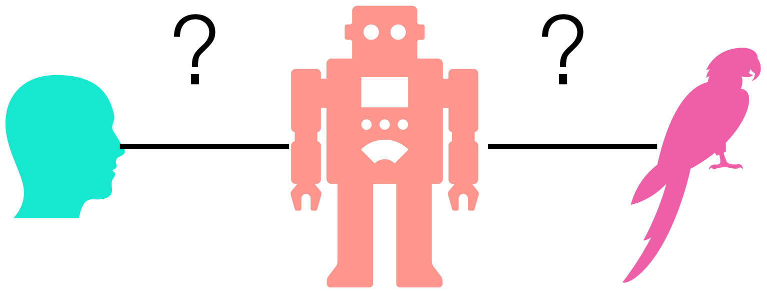
- For unforeseen health and personal emergencies, please contact the instructors at cics.685.instructors@gmail.com .
    - Job interviews / other schoolwork are **not** excuses for late homework.
- We won't accept late homework otherwise, sorry. ***Start early.***

Midterms

- We will schedule two evening midterms (outside class sessions) during the course
  - One A4 or letter-sized, double-sided sheet of notes, ***handwritten***, is allowed
    - Although I don’t think you will need one
  - No devices
- Questions would be centered on the classes, quizzes, and homework

~~Facts~~ Perspectives

- Many materials are based on our **interpretation/perspective** of the latest findings
  - Or even just insights
  - No good textbook on this



- Perspectives are debatable
  - Could be even controversial
  - You often see lots of debate between experts
- Uncertainty could lead to creativity
  - Discussion welcome!

natural language processing

natural language processing

languages that evolved naturally through human use

e.g., Spanish, English, Arabic, Hindi, etc.

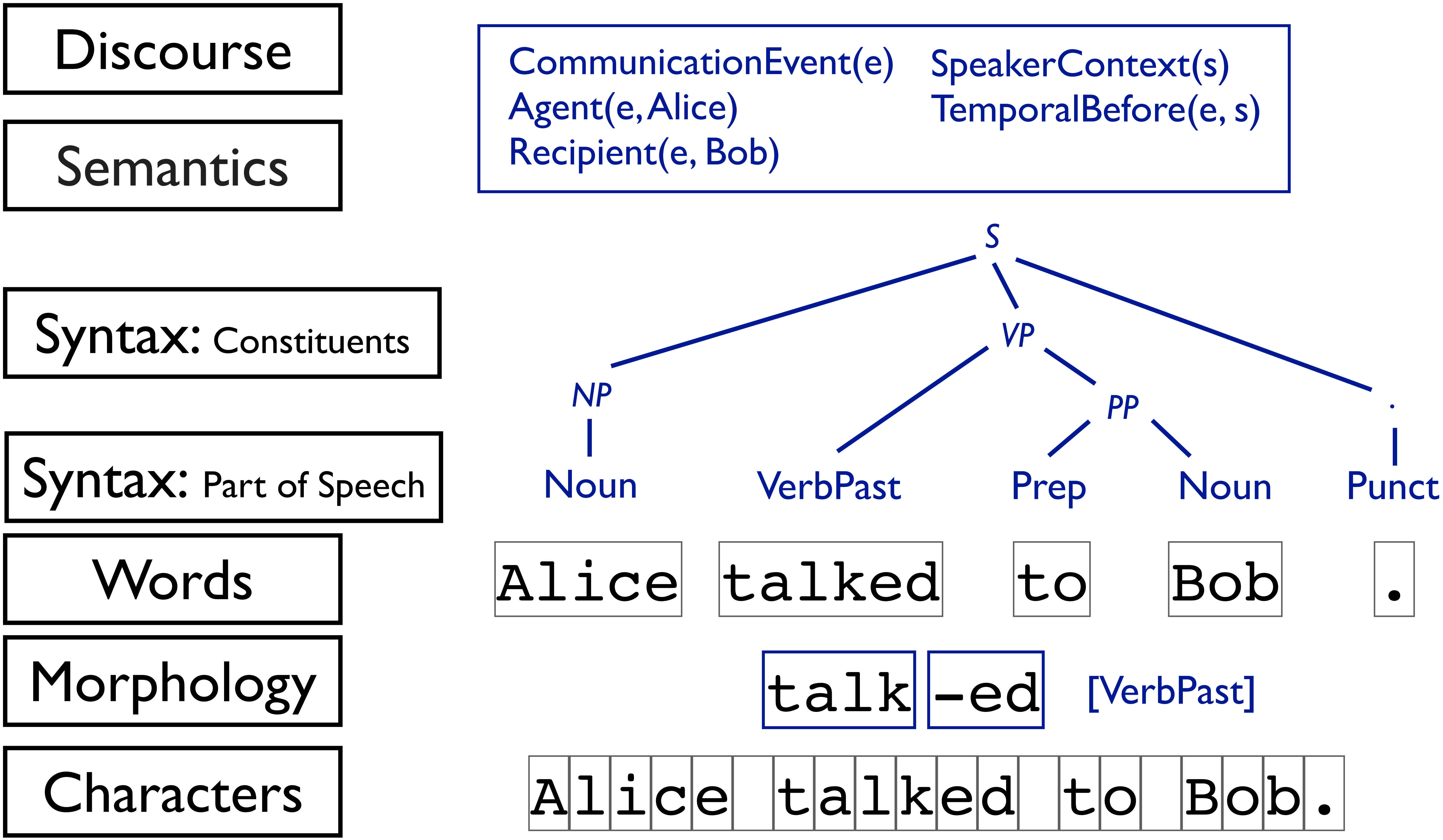
natural language processing

supervised learning: *map text to* **X**

unsupervised learning: *learn* **X** *from text*

generate *text from* **X**

Levels of linguistic structure



Discourse

Semantics

Syntax: Constituents

Syntax: Part of Speech

Words

Morphology

Characters

CommunicationEvent(e)
Agent(e, Alice)
Recipient(e, Bob)

SpeakerContext(s)
TemporalBefore(e, s)

Alice talked to Bob .

talk -ed [VerbPast]

A l i c e   t a l k e d   t o   B o b .

supervised learning: given a collection of labeled examples (where each example is a text *X* paired with a label *Y*), learn a mapping from *X* to *Y*

Example: given a collection of 20K movie reviews, train a model to map review text to review score (*sentiment analysis*)

self-supervised learning: given a collection of *just text,* without extra labels, create labels out of the text and use them for *pretraining* a model that has some general understanding of human language

- **Language modeling**: given the beginning of a sentence or document, predict the next word
- **Masked language modeling**: given an entire document with some words or spans masked out, predict the missing words

How much data can we gather for these tasks?

transfer learning: first *pretrain* a large self-supervised model, and then *fine-tune* it on a small labeled dataset using supervised learning

Example: pretrain a large language model on hundreds of billions of words, and then fine-tune it on 20K reviews to specialize it for sentiment analysis

in-context learning: first *pretrain* a large self-supervised model, and then *prompt* it in natural language to solve a particular task without any further training

Example: pretrain a large language model on hundreds of billions of words, and then feed in “what is the sentiment of this sentence: <insert sentence>”

Language models

api.together.xyz

What are people using LLMs for?

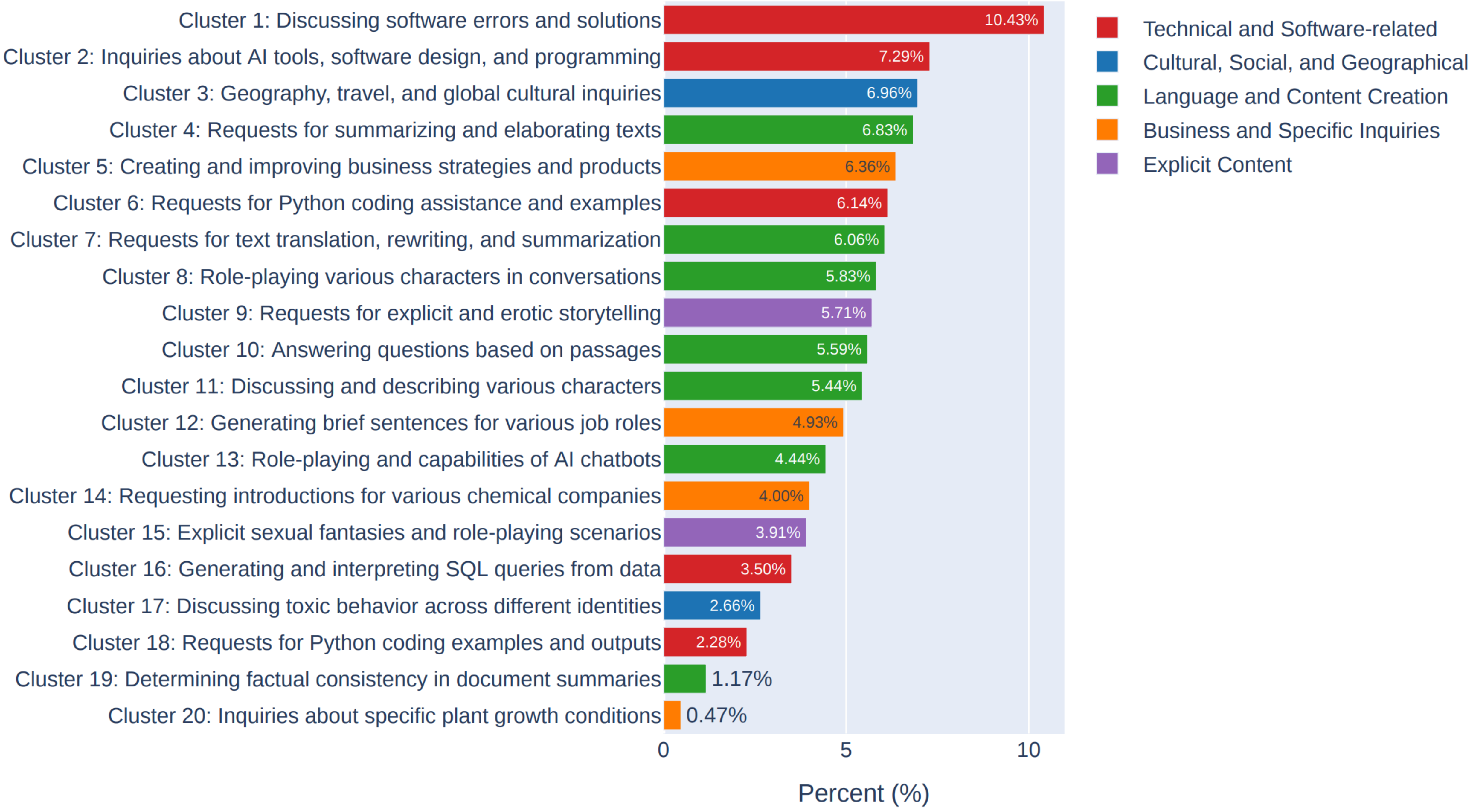
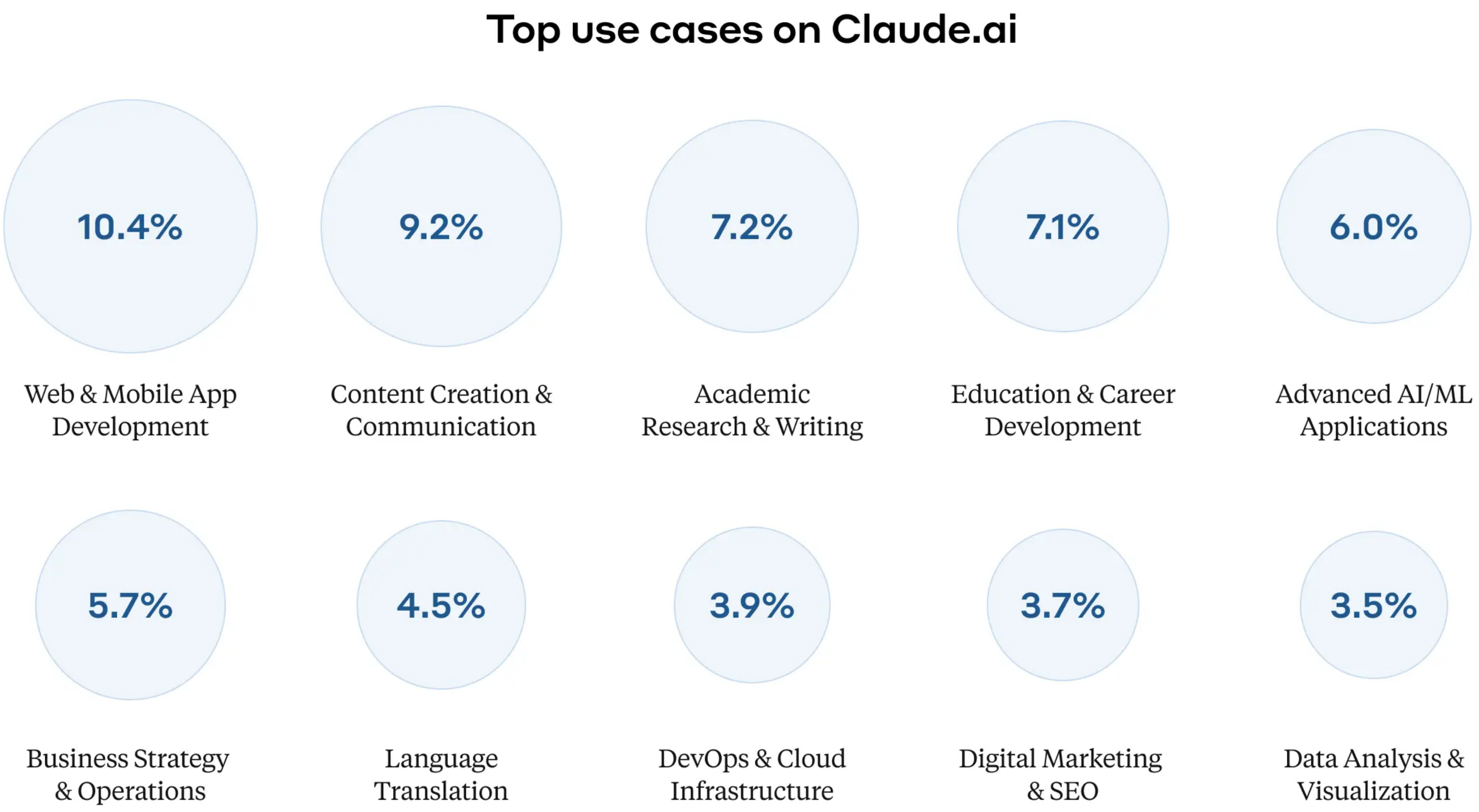


Figure 3: Topic distribution of 100K sampled conversations. Manual inspection of cluster centroids

https://arxiv.org/pdf/2309.11998.pdf

What are people using LLMs for?



https://www.anthropic.com/research/clio

Rough list of topics

- **Background**: language models and neural networks
- **Models**: Transformers
  - RNN > BERT > GPT3 > ChatGPT > today’s LLMs
- **Tasks**: text generation (e.g., translation, summarization), classification, retrieval, etc.
- **Data**: annotation, evaluation, artifacts
- **Methods**: pretraining, finetuning, preference tuning, prompting, reasoning?
- **Notice**: NLP != LLMs

Course topics (approximate)

- Language Modeling
- Neural Language models, Optimization and Backpropagation
- Embeddings
- Attention Mechanisms
- Transformer
- Fine-Tuning and Instruction Tuning
- Datasets and Evaluation
- LLM Alignment
- Tokenization
- Interpretability
- Reasoning
- Decoding and Positional Embedding
- Prompt Engineering and In-context Learning
- Special Topics

- stopped here 9/2/25