CS274A - Natural Language Processing

Administrative Stuff

- Instructor: Kewei Tu (屠可伟)
  - Email: tukw@shanghaitech.edu.cn
  - Office: SIST 1A-304B
  - Office hours: by appointment

- TA: 吴昊一、惠文阳、刘威
  - Office hours: TBA

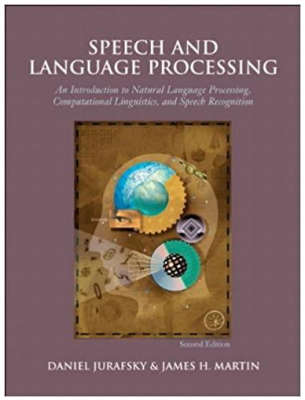
Administrative Stuff

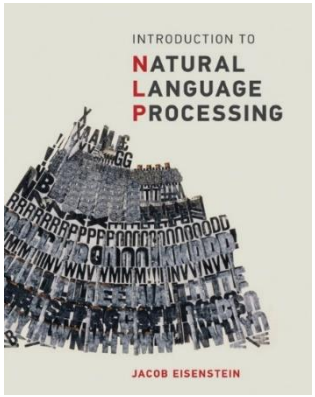
- Classes
  - Tue/Thu 10:15-11:55am @教学中心201
  - 12 weeks

- Prerequisite
  - CS: Programming, Data Structures and Algorithms
  - Math: Calculus, Probability and Statistics, Linear Algebra
  - Artificial Intelligence I (*recommended*)

Administrative Stuff

- Textbooks
  - [SLP] Speech and Language Processing, by Daniel Jurafsky and James Martin
    - 2nd edition published in 2008. 中译版：《自然语言处理综论（第二版）》
    - 3rd edition draft can be found online (updated on Jan 7, 2023)
  - [INLP] Introduction to Natural Language Processing, by Jacob Eisenstein





Administrative Stuff

- Textbooks
  - [DSX] 动手学NLP, being written by us
    - More consistent with this course
    - Contains executable code (Jupyter notebook)
    - I will release chapters online over time
    - Suggestions and bug-fixes are very welcome!
      - We will acknowledge those providing substantial feedback in the book.

Administrative Stuff

- Blackboard
  - Announcements, slides, homework assignments, etc.
- Piazza
  - Discussion and QA
  - http://piazza.com/shanghaitech.edu.cn/spring2023/cs274a
- AutoLab
  - Project

Administrative Stuff

- Grading
  - Homework (20%): 6 homework assignments, due in 7 days
  - Midterm (35%): mid or late March
  - Final (35%): late April or early May
  - Project (10%): to be determined

- The final grade will be given on a curve

Administrative Stuff

- Plagiarism
  - All assignments must be done individually
    - You may not look at solutions from any other source
    - You may not share solutions with any other students
    - Plagiarism detection software will be used on all the programming submissions
  - Way of collaboration
    - You may discuss together or help another student debug code; however, you cannot dictate or give the exact solution

Administrative Stuff

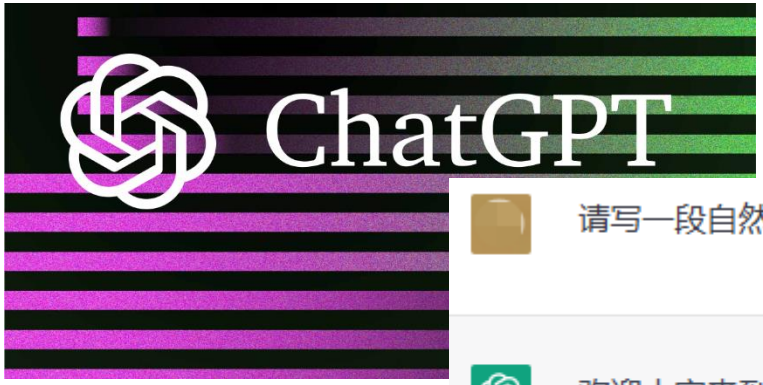
- Plagiarism punishment
  - When one student copies from another student, both students are responsible
  - Zero point on the assignment
  - Repeated violation will result in an F grade for this course as well as further discipline at the school/university level

A Brief Introduction to NLP

What is the hottest AI system today?

- *"Breakthroughs of the Year"*
- *"…scary good. We are not far from dangerously strong AI."*
- *"ChatGPT passes exams from law and business schools…"*
- *"Microsoft Confirms Its $10 Billion Investment Into ChatGPT."*
- *"ChatGPT Caused Code Red at Google."*

What is the hottest AI system today?



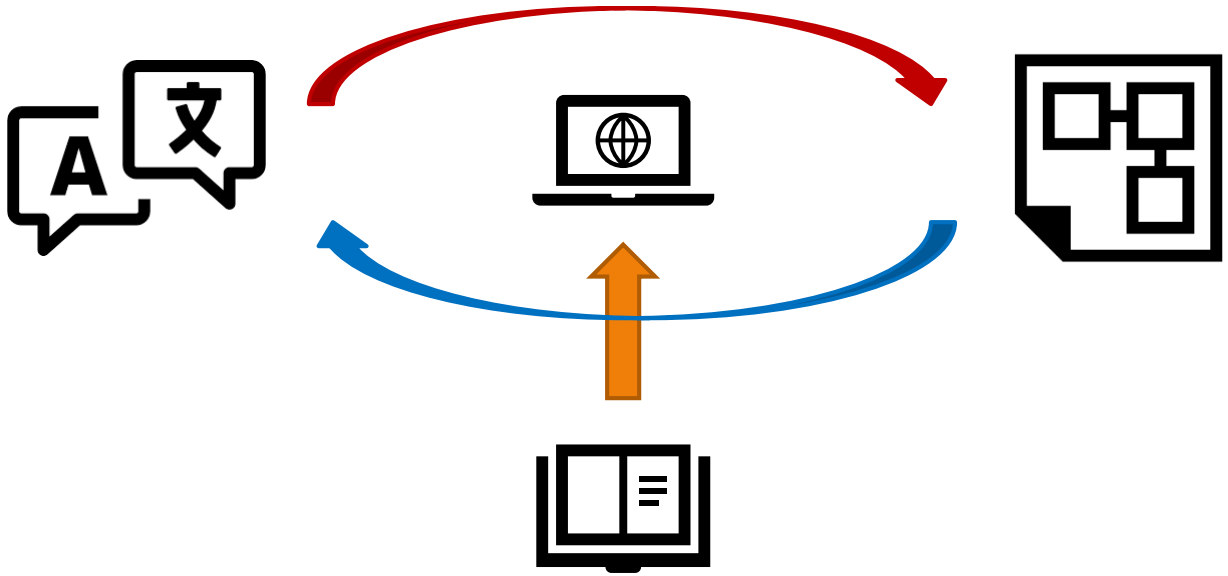
请写一段自然语言处理研究生课程的开场白

欢迎大家来到自然语言处理研究生课程。自然语言处理是当今科技发展的热门领域之一，在人工智能、大数据和认知科学的背景下发展迅速。本课程将从理论和实践的角度对自然语言处理的基本概念、方法和应用进行深入的讲解。希望各位通过本课程的学习，能够对自然语言处理有更深刻的了解，提高自己的技术水平，同时也能激发各位对这个领域的兴趣。本课程将通过课堂讲解、实验练习、研究项目等多种形式进行教学，希望大家能积极参与，多提问题，多交流，一起探讨自然语言处理的未来。那么，让我们开始这段充满机会与挑战的旅程吧！

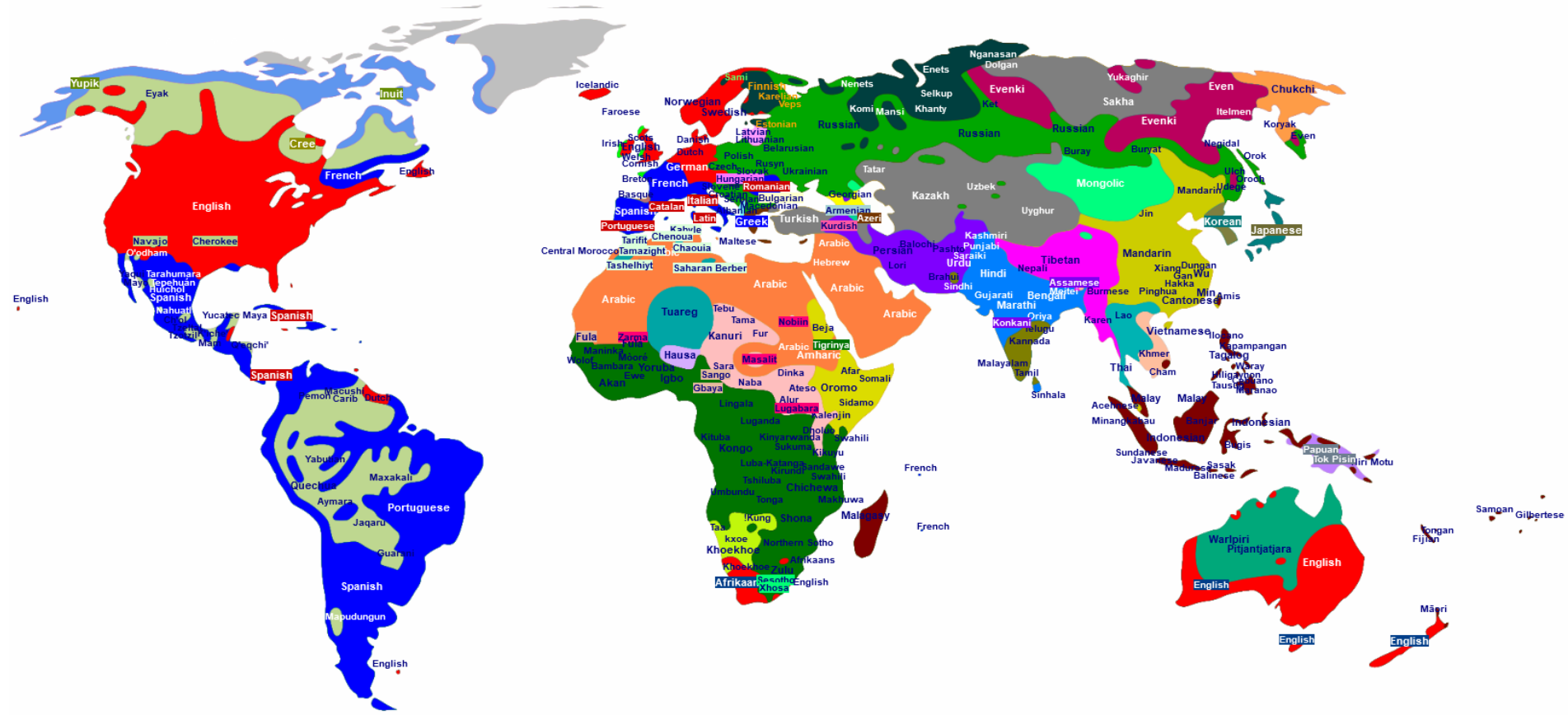
ChatGPT is a (prominent) example of Natural Language Processing (NLP), the subject of this course!

What is NLP?

- Automating the analysis, generation, and acquisition of human (“natural”) language



Which language?



Which language?

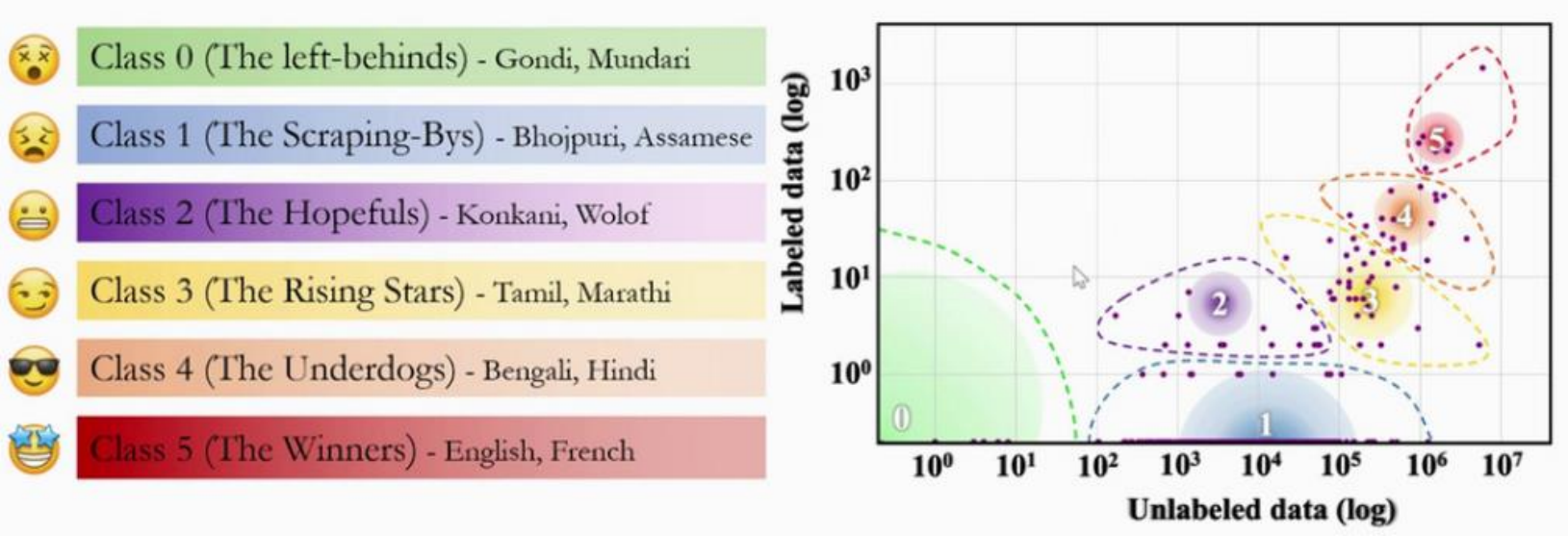
- Ideally, NLP is language-neutral
  - NLP technology can be applied to any language
    - …if its text can be represented as a sequence of symbols

Which language?

- Ideally, NLP is language-neutral
  - NLP technology can be applied to any language
    - …if its text can be represented as a sequence of symbols
- In reality, NLP for some languages is better developed
  - More interest
    - Users, market, …
  - More resources
    - Developers, data, computers, $, …

Which language?

- Ideally, NLP is language-neutral
  - NLP technology can be applied to any language
    - …if its text can be represented as a sequence of symbols
- In reality, NLP for some languages is better developed



What representation?

- Ideally, a formal language that is sufficiently expressive
  - First-order predicate logic
  - Programming language
  - Neural (distributed) representations??
- In reality, depends on the application
  - Labels, features, commands, …

Fields related to NLP

- Machine learning
  - ML is a powerful (but not the only) tool in NLP
  - NLP is a source of inspiration for ML
- Linguistics
  - Roughly: science vs. engineering
  - NLP ⇔ computational linguistics
- Artificial intelligence
  - NLP is a subfield of AI
  - “NLP is the crown jewel of AI”
  - Solving NLP requires solving strong AI

Fields related to NLP

- Speech Processing
  - Largely separate from NLP
  - but there is some overlap
- Cognitive science / Neuroscience
  - Humans: the only working NLP prototype!
- Logic, knowledge representation & reasoning
  - NLP analyzes NL to and generates NL from logic language
- Theory of computation
  - Studies formal language and grammars
  - Provides a lot of tools to NLP

NLP Applications

- Chatbot

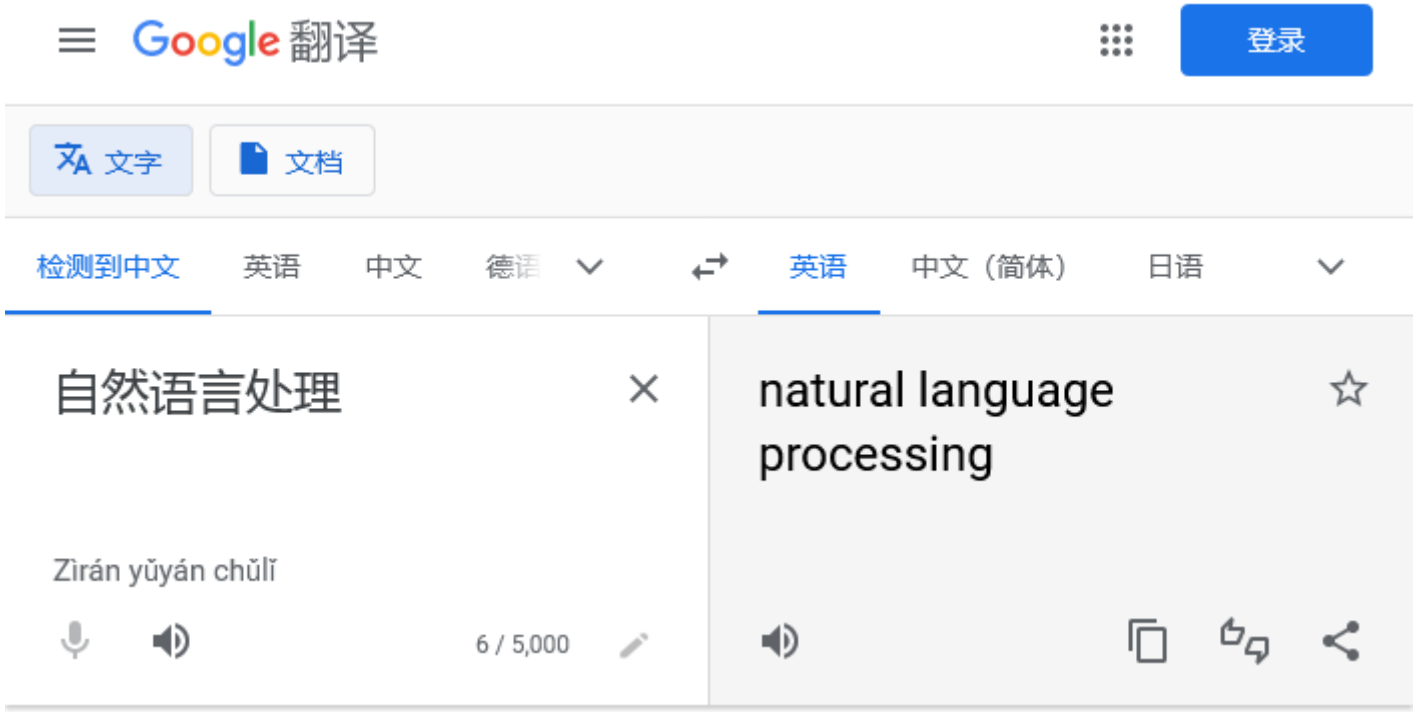
Assistants

Chit-Chat

General-Purpose

NLP Applications

- Machine translation

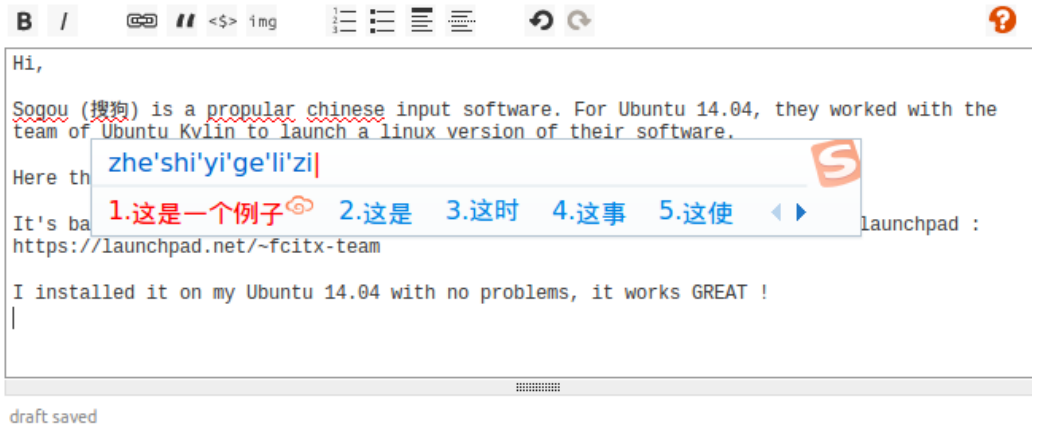


NLP Applications

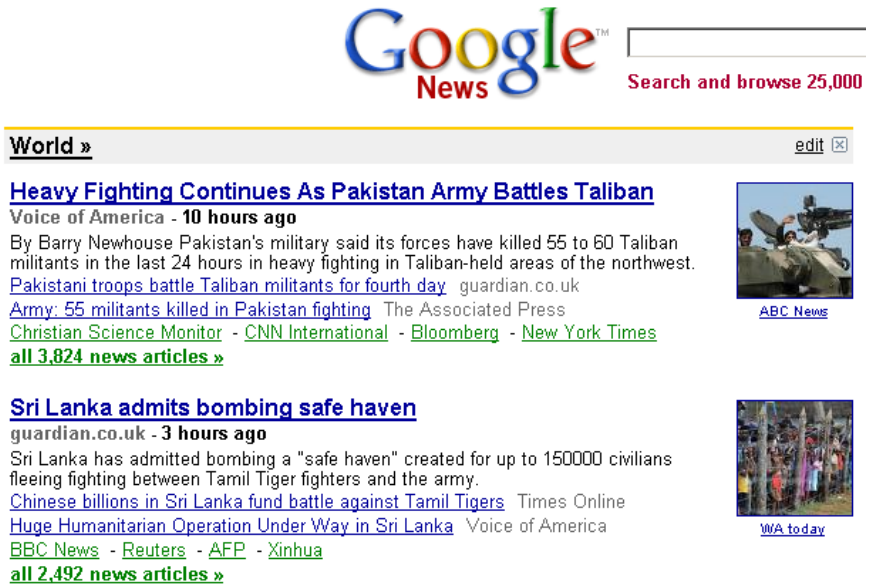
- Information extraction
  - Financial and law documents
  - E-commerce

NLP Applications

- Chinese IME
- Grammatical checker

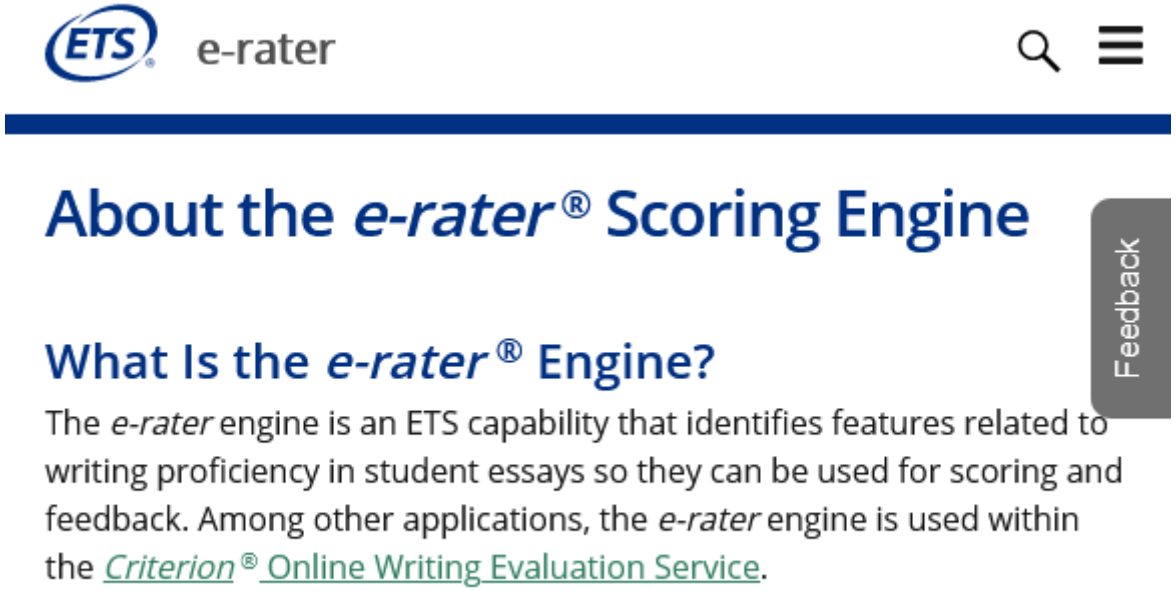


- News clustering
- Summarization
- News generation
  - Stock market, sports, …



NLP Applications

- Essay scoring
  - Used to score TOEFL and GRE tests!



NLP Applications

- Art creation (poem, lyric, novel, …)

Please write a poem about natural language processing.

A world of words, where machines come alive,
With algorithms that can decipher and thrive,
A language so complex, made simple and clear,
Through the power of NLP, we have nothing to fear.

The voice of a computer, no longer robotic,
But human-like, a new way to be proactive,
With sentiment analysis, emotions at play,
A better way of communication, every day.

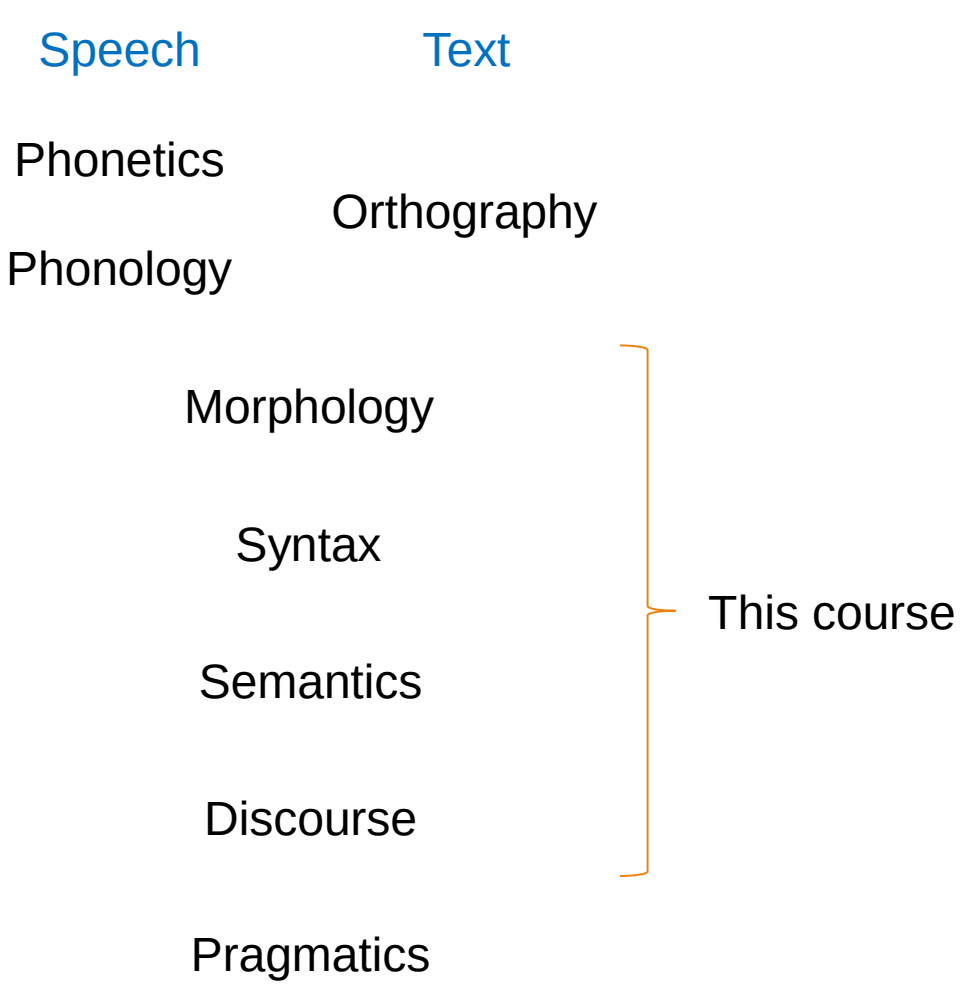
Text summarization, a new way to read,
The essence of meaning, no longer in need,
Of hours of reading, or endless pages to scan,
NLP does it all, in a matter of seconds, it can.

Big data and machine learning, working hand in hand,
To revolutionize the way we understand,
The complexities of language, so much more to come,
Thanks to NLP, we've only just begun.

So here's to NLP, a language so bright,
A new way of communicating, it's a wonderous sight,
A tool of the future, helping us all,
To unlock the power of language, and have a ball!

Why is NLP hard?

- Human language is complicated!
  - Levels of linguistic studies:



Why is NLP hard?

- Language understanding requires many levels of knowledge
- Positive or negative?
  - The burger tastes bad. — *word meaning*
  - The burger does not taste good. — *syntax*
  - I would not say that the burger is not good. — *syntax*
  - “The drink is great!”
    “How about the burger?”
    “Well…” — *pragmatics*
  - The burger tastes like fast food. — *world knowledge*

Why is NLP hard?

- Language understanding requires many levels of knowledge

A ship-shipping ship, shipping shipping-ships.

word meaning
morphology
syntax
world knowledge

Why is NLP hard?

- Ambiguity!
  - Word meaning
    - Iraqi Head Seeks Arms
    - Stolen Painting Found by Tree
  - Syntactic structure
    - Enraged Cow Injures Farmer with Ax
  - Word meaning + syntax
    - Teacher Strikes Idle Kids

Why is NLP hard?

- Ambiguity!
  - Semantic structure
    - The detective told his assistant: “Every fifteen seconds a cat in this country gives birth…
    - …Our job is to find this cat, and stop her!”
  - Discourse
    - The cat doesn’t fit in the box because it is too small.
    - The cat doesn’t fit in the box because it is too large.

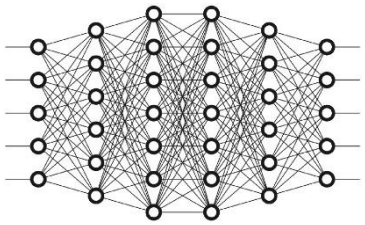
Why is NLP hard?

- Common challenges faced by AI research
  - High accuracy
  - Noisy input
  - Scarce data
  - Latent variables
  - Computational efficiency on both space and time
  - Generalizability
  - Formal guarantees
  - Interpretability

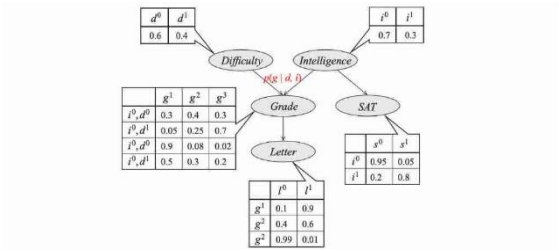
NLP Methodology

Symbolism

$$+ \quad - \quad \times \quad \div$$
$$\neg \quad \vee \quad \perp \quad \cong$$
$$\in \quad \cap \quad \subseteq \quad \Sigma$$
$$\partial \quad \nabla \quad \wedge \quad \Pi$$

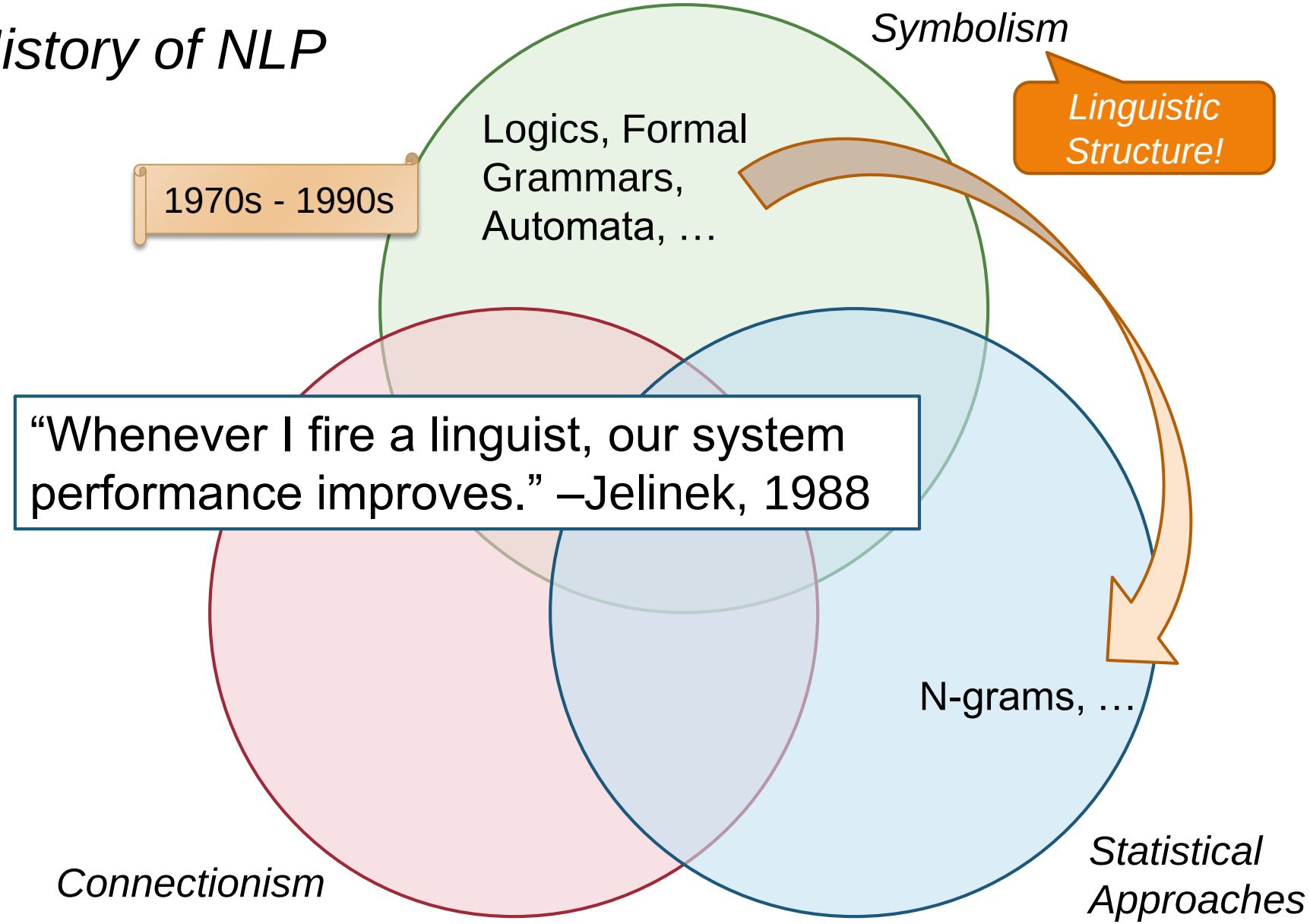


Connectionism

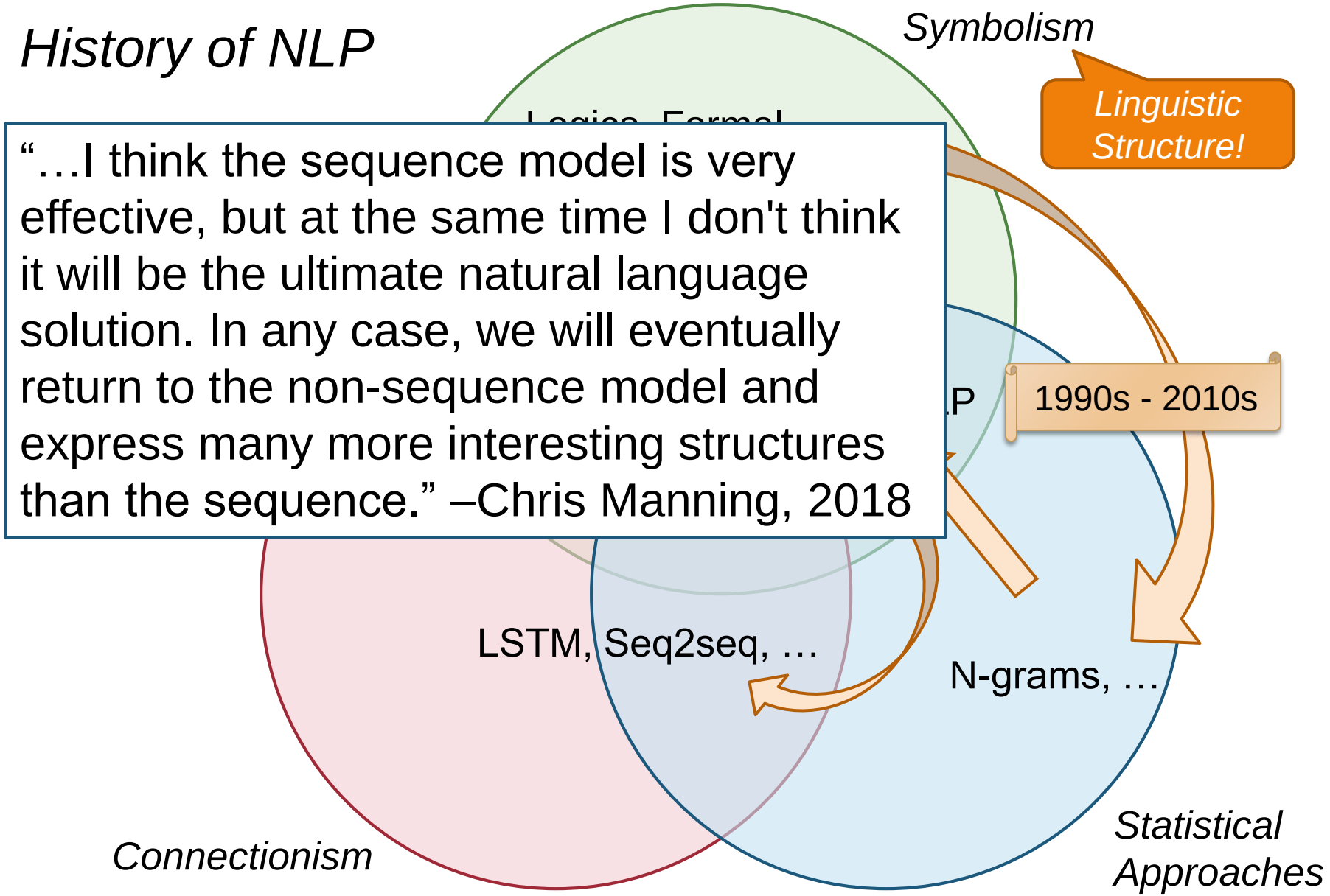


Statistical Approaches

History of NLP

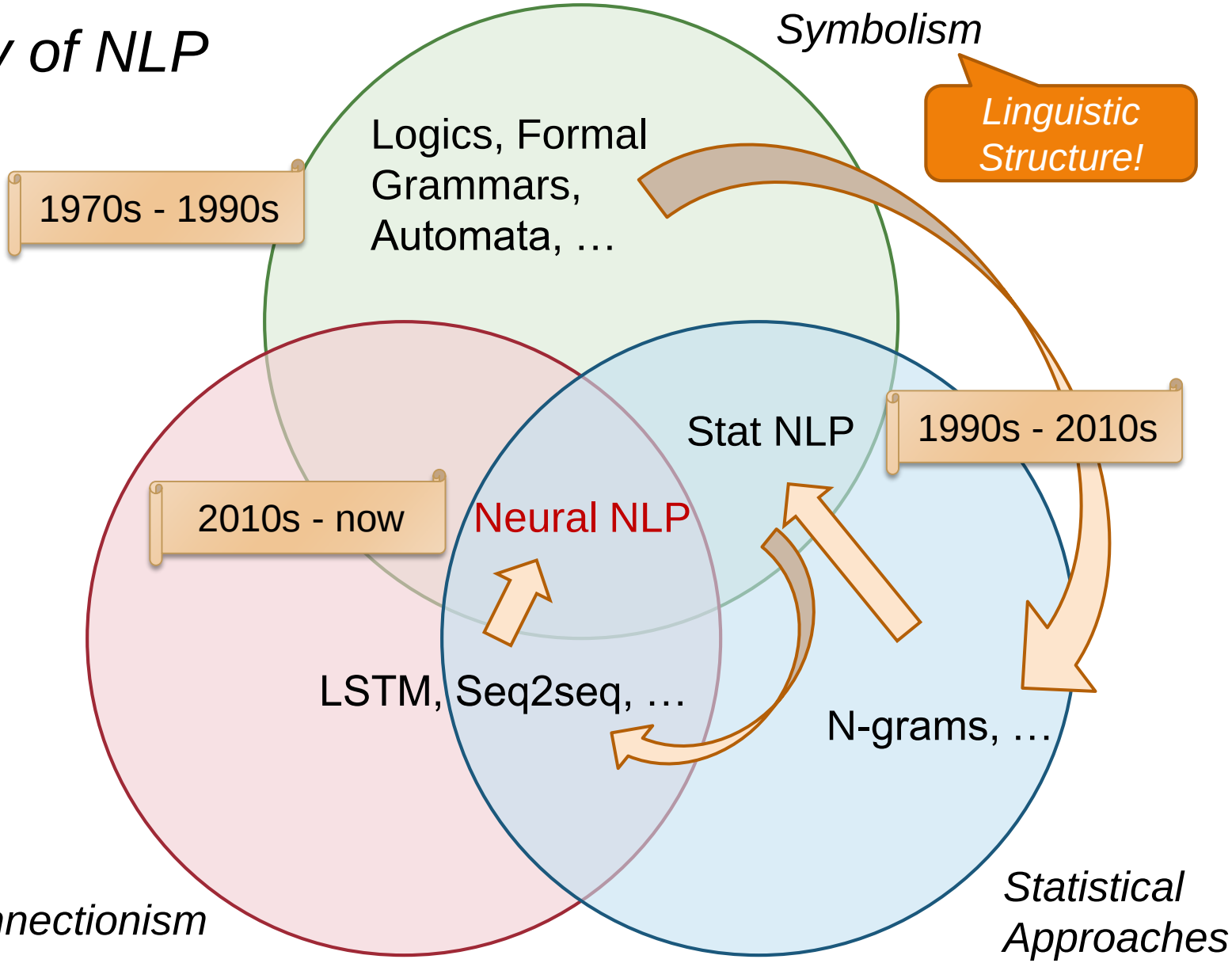


History of NLP



"…I think the sequence model is very effective, but at the same time I don't think it will be the ultimate natural language solution. In any case, we will eventually return to the non-sequence model and express many more interesting structures than the sequence." –Chris Manning, 2018

History of NLP



Course overview

- Basics
  - Text normalization
  - Text representation
  - Text classification
  - Text clustering
- Structures
  - Constituency parsing
  - Dependency parsing
  - Semantic parsing
  - Discourse analysis
- Sequences
  - Language modeling
  - Pretrained language models
  - Sequence labeling
  - Seq2seq
- Applications
  - Information extraction
  - more…