# CS274A - Natural Language Processing

# Administrative Stuff

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

- TA: 蒋承越、杨松霖、楼超、吴昊一
  - Office hours: TBA

# Administrative Stuff

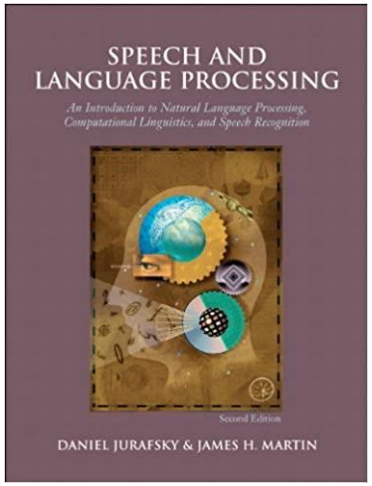
- Enrollment quota increased!

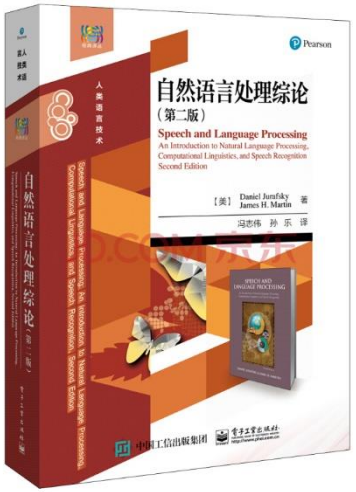
- Classes
  - Mon/Wed 8:15-9:55am @教学中心202
  - 12 weeks

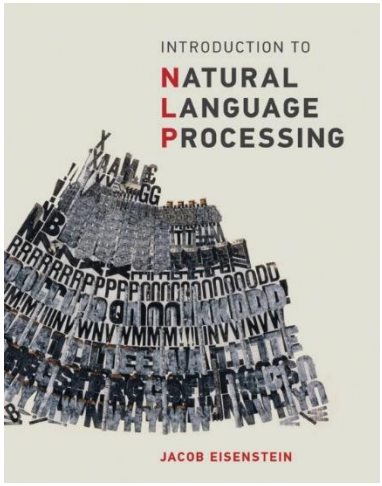
- Prerequisite
  - CS: Programming, Data Structures and Algorithms
  - Math: Calculus, Probability and Statistics, Linear Algebra
  - Artificial Intelligence I (recommended)

# Administrative Stuff

- Textbooks
  - [SLP] Speech and Language Processing, 2nd Edition, by Daniel Jurafsky and James Martin
    - [中译版] 自然语言处理综论（第二版）
    - 3rd edition draft can be found online
  - [INLP] Introduction to Natural Language Processing, by Jacob Eisenstein




# Administrative Stuff

- Blackboard
  - Announcements, homework assignments, slides, etc.
- Piazza
  - Discussion and QA
  - http://piazza.com/shanghaitech.edu.cn/spring2022/cs274a
- AutoLab
  - Project

# Administrative Stuff

- Grading
  - Homework (10%): 4 homework assignments, due in 5 days
  - Midterm (40%): possibly in the 6-7th week
  - Final (40%): possibly in early or mid 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 assignments
  - Way of collaboration
    - You may discuss together or help another student debug code; however, you cannot dictate or give the exact solution

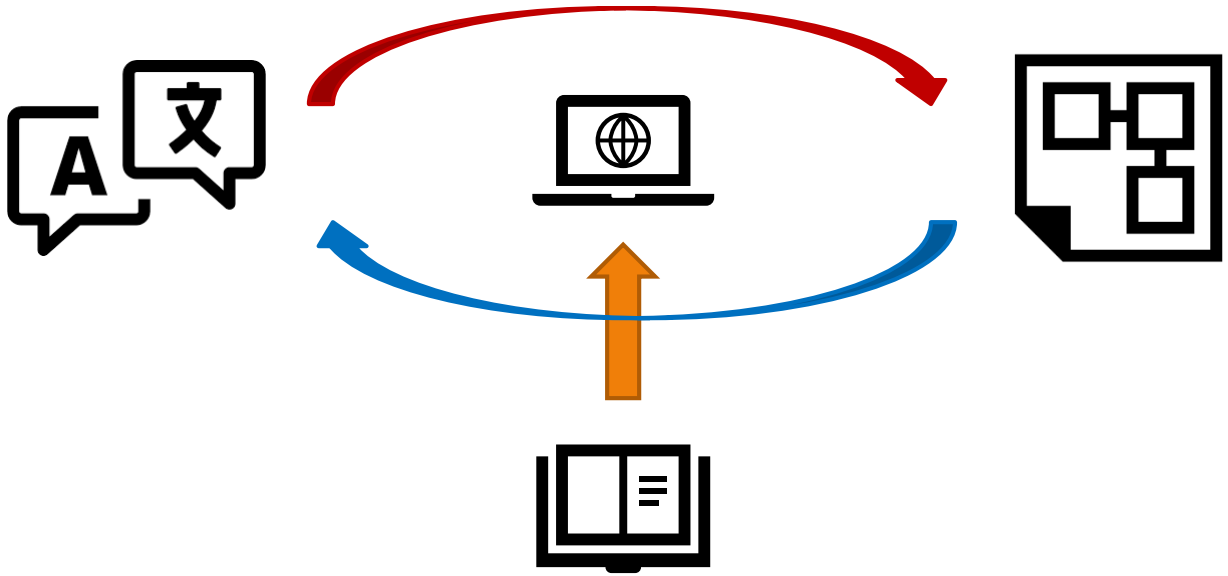
# Administrative Stuff

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

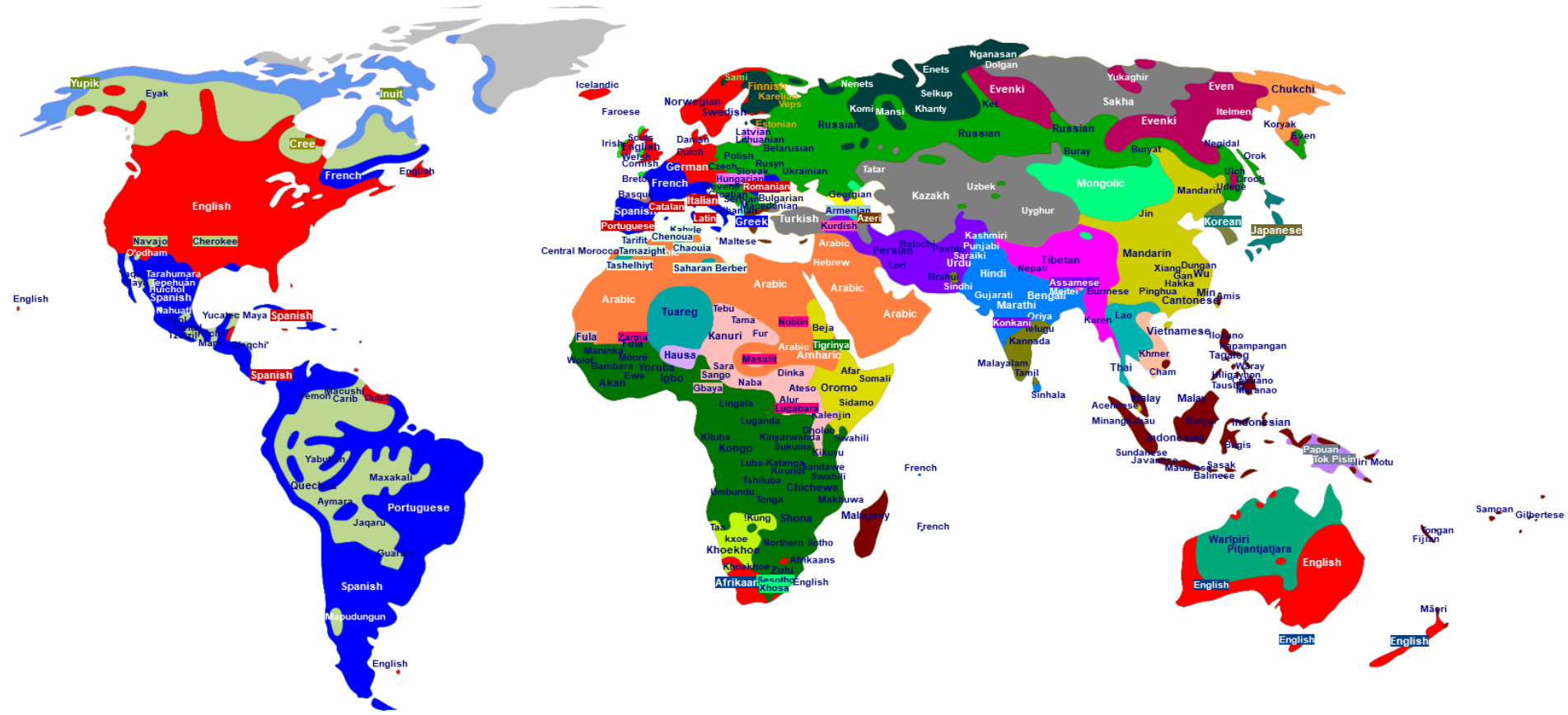
# A Brief Introduction to NLP

# What is NLP?

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




# Which language?

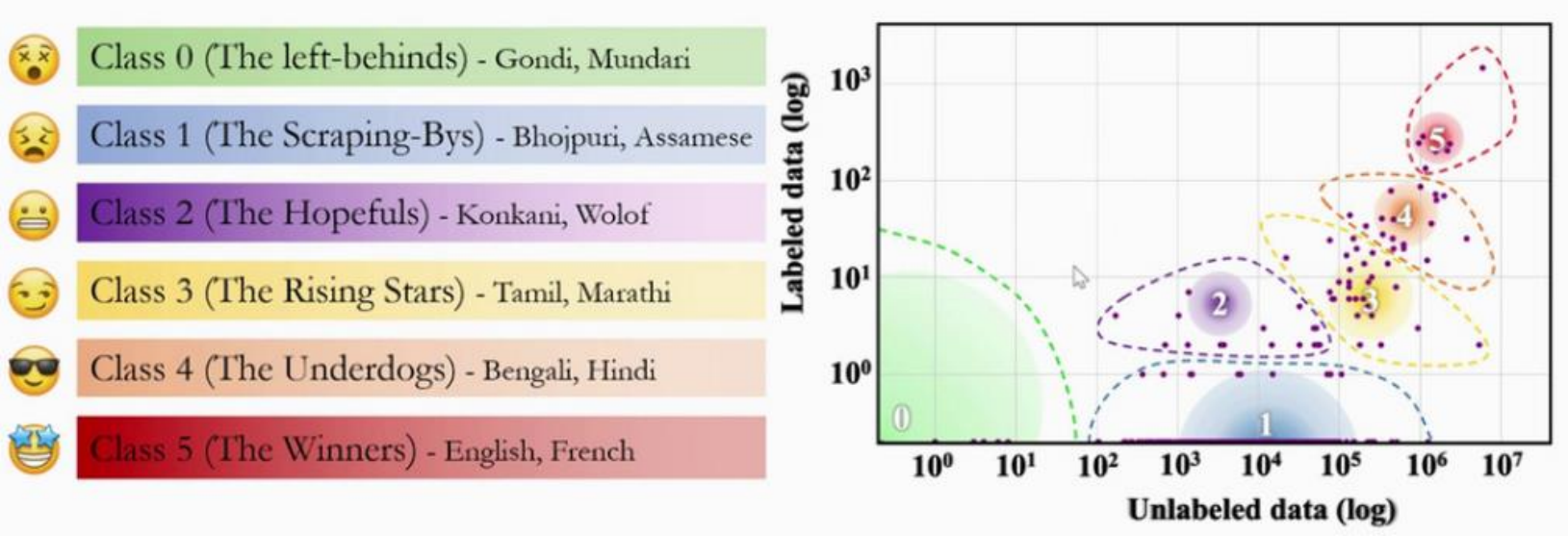


# Which language?

- 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



# 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, $, …

# What representation?

- Ideally, a formal language that is sufficiently expressive
  - First-order predicate logic
  - Programming language
  - Neural 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

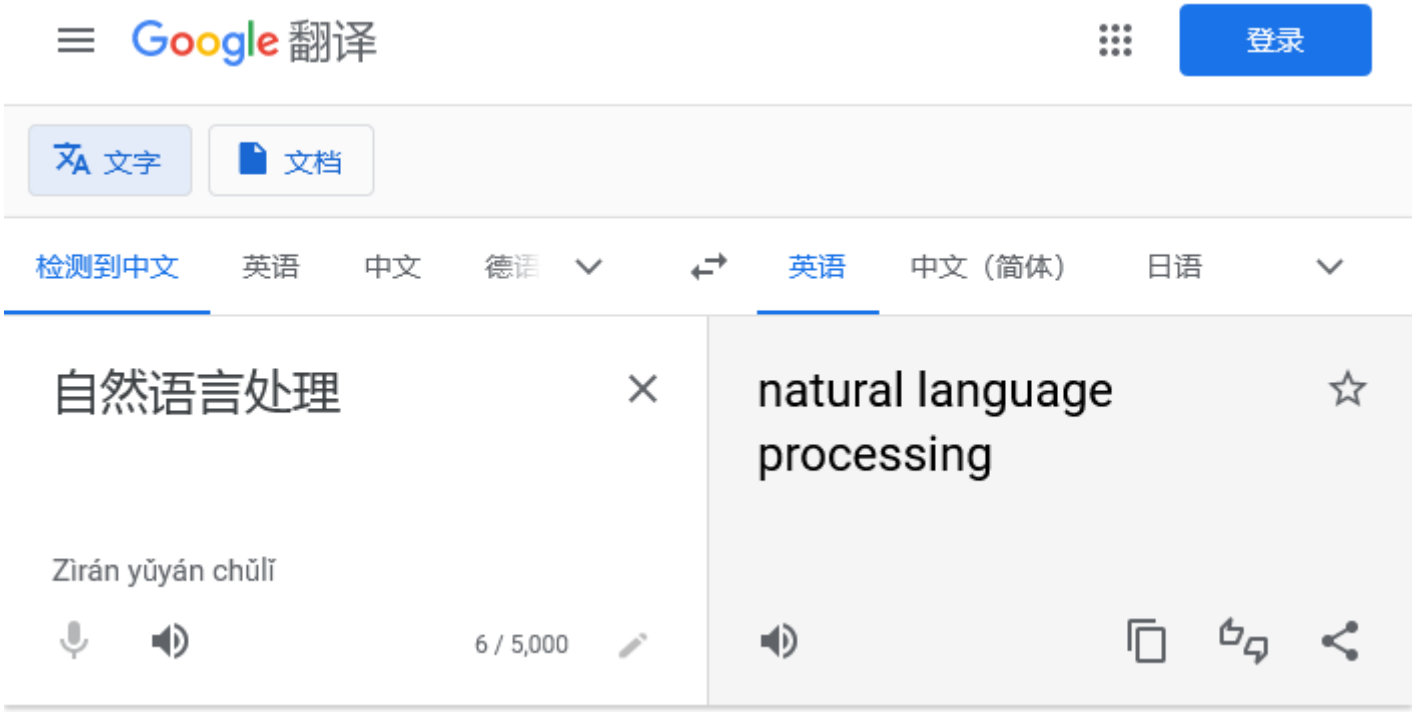



amazon alexa

Google ASSISTANT

# NLP Applications

- Machine translation

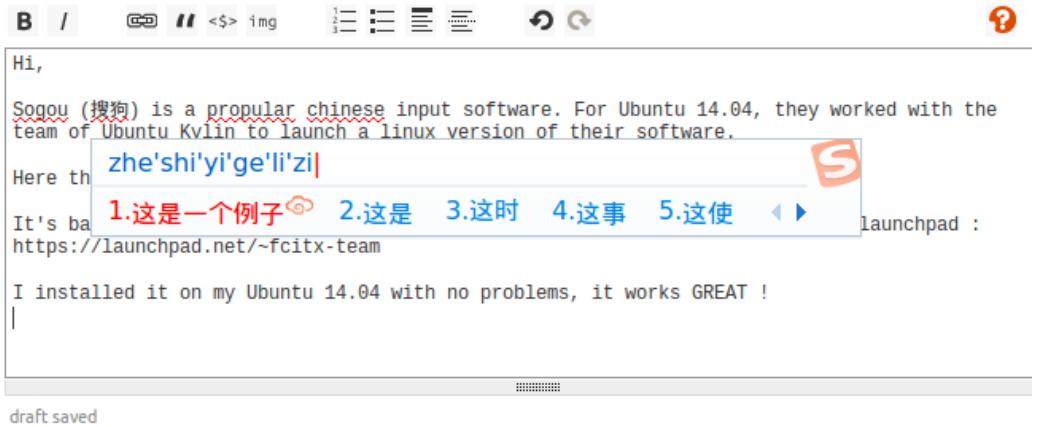


# NLP Applications

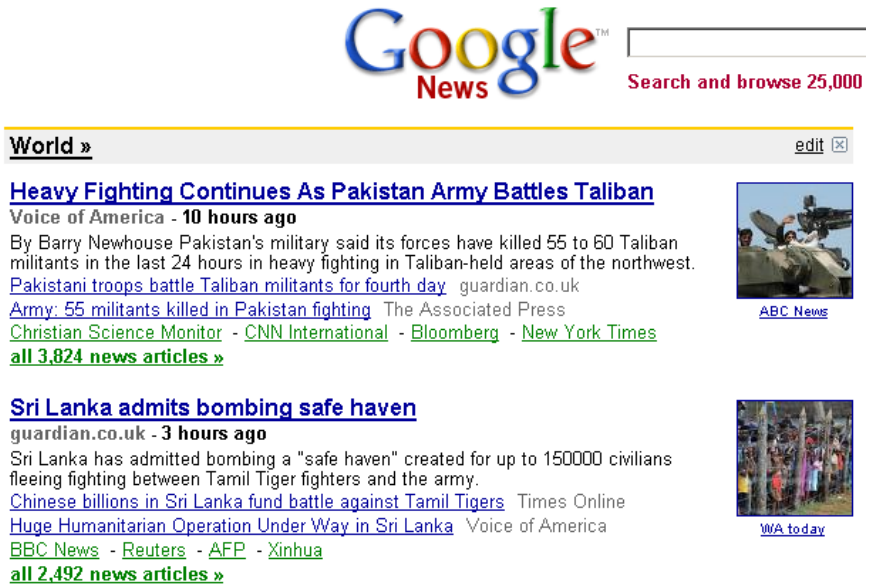
- Information extraction
  - Financial and law documents
  - E-commerce

# NLP Applications

- Chinese IME
- Grammatical checker

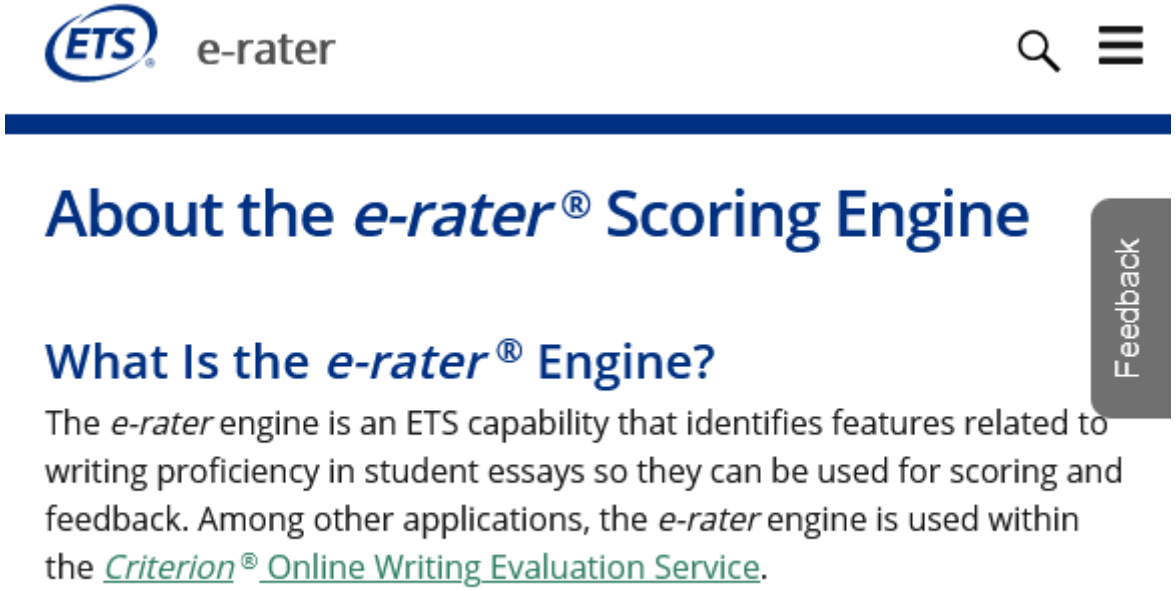



- News clustering
- Summarization
- News generation
  - Stock market, sports, …



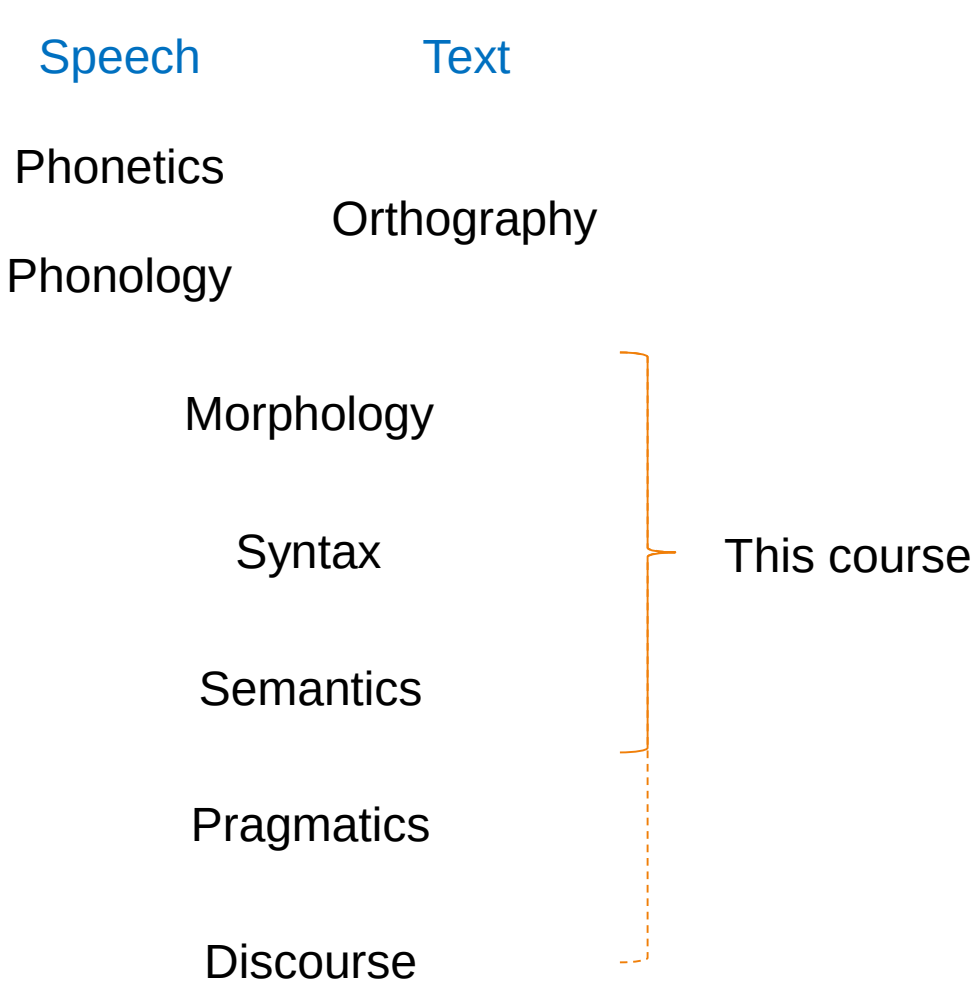
# NLP Applications

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




# 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 it!”
  - 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.

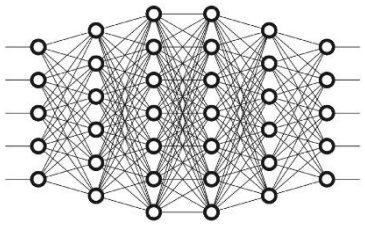
# Why is NLP hard?

- 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

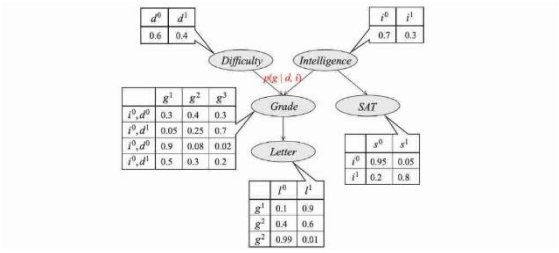
# NLP Methodology

*Symbolism*

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

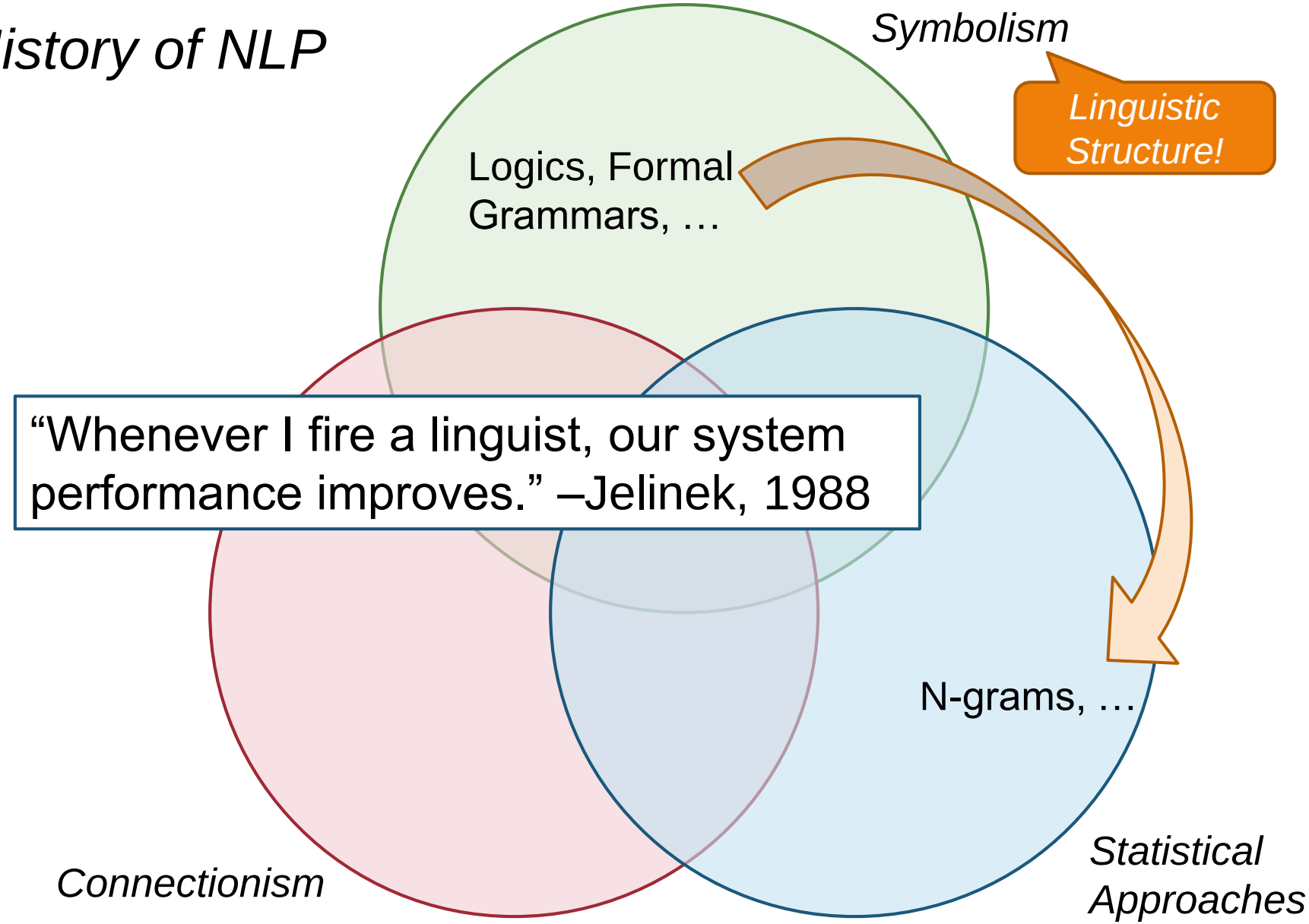


*Connectionism*




*Statistical Approaches*

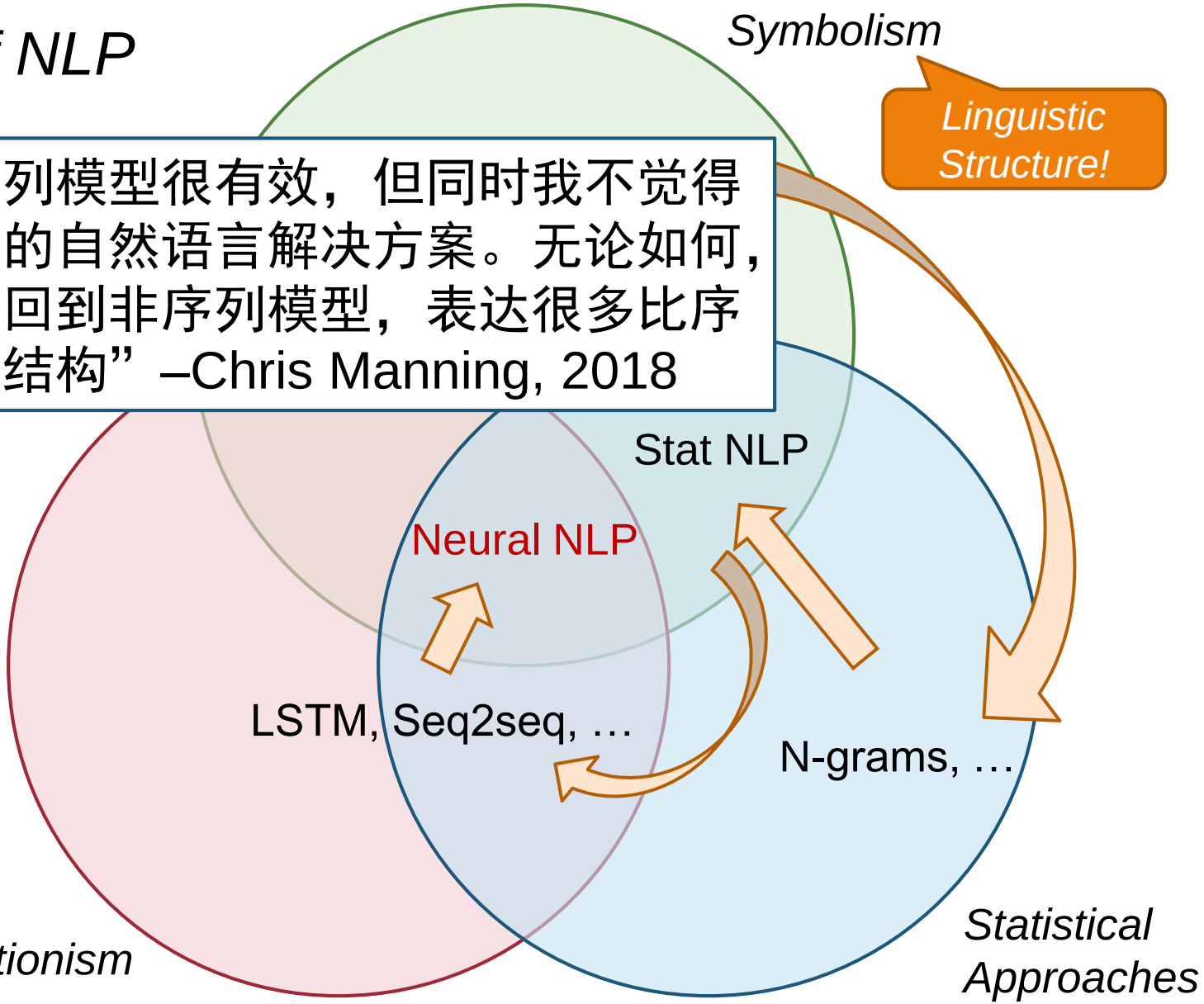
# History of NLP




"Whenever I fire a linguist, our system performance improves." –Jelinek, 1988

# History of NLP

“我认为序列模型很有效，但同时我不觉得它会是最终的自然语言解决方案。无论如何，最终我们会回到非序列模型，表达很多比序列更有趣的结构”–Chris Manning, 2018



# Course overview

- Text normalization
- Word representation
- Text classification
- Text clustering
- Language modeling
- Contextual word representation
- Sequence labeling
- Constituency parsing
- Dependency parsing
- Word senses
- Sentence semantics
- Information extraction
- Discourse analysis
- Machine translation & seq2seq
- Generation, question answering and dialog