NLP Special Topics

Ruihong Huang
Texas A&M University

- "An Aggie does not lie, cheat, or steal or tolerate those who do." For additional information, please visit: http://aggiehonor.tamu.edu.

Upon accepting admission to Texas A&M University, a student immediately assumes a commitment to uphold the Honor Code, to accept responsibility for learning, and to follow the philosophy and rules of the Honor System. Students will be required to state their commitment on examinations, research papers, and other academic work. Ignorance of the rules does not exclude any member of the TAMU community from the requirements or the processes of the Honor System.

- The Americans with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires that all students with disabilities be guaranteed a learning environment that provides for reasonable accommodation of their disabilities. If you believe you have a disability requiring an accommodation, please contact Disability Services, currently located in the Disability Services building at the Student Services at White Creek complex on west campus or call 979-845-1637. For additional information, visit http://disability.tamu.edu.

- Piazza: CSCE 689, Special Topics in Natural Language Processing: Information Extraction

- course page: http://faculty.cse.tamu.edu/huangrh/Spring16_nlp_ie_v3.html

Paper presentations: 25%

Paper summaries: 10%

Class participation: 10%

Mid-term Project: 25%

Final-term Project: 30%

Submitting Paper Summaries

- Due before each class
- Email Subject Line: paper_summary_week_m_day_n (m: 1, 2, 3...; n: 2 or 4)
- Email content: 2 lines (line 1: First Name, Last Name, line 2: UIN)
- paper summary: a plain text file attachment, at most a page.

Paper Summaries and Presentations

- Problem defination and motivation.
- Possible Downsteam Applications (as described in the paper and your thoughts)
- Proposed solution/algorithm/method.
- Strengths of the proposed method.
- Weakness of the proposed method.
- Your thoughts to improve on the proposed method.

- We should meet and talk about the paper a few days before the presentation.

Tentative Project Timeline

- **2/16 -- Proposal report and presentation**. report due before class, report length: 1 page, presentation: 3 minutes.
- **3/22 -- Midterm project report and presentation**. report due before class, report length: at most 3 pages, presentation: 5 minutes.
- **4/26 -- Final report and presentation**. report due before class, report length: at most 8 pages, presentation: 15 minutes.

Project Submission

- By the end of the semester: code and data.

Submitting Project Reports

- Email Subject Line: project_proposal or project_mid_term or project_final_term
- Late Policy: 20% reduction per day.

Basic Recipe of Forming a Project

- Choose a Topic and do a quick survey
- Prepare data
- Think about evaluation methods
- Start to work on it

Applications

- Sentiment Analysis
- Question-Answering
- Text Summarization

Why sentiment analysis?

- *Movie*: is this review positive or negative?
- *Products*: what do people think about the new iPhone?
- *Public sentiment*: how is consumer confidence? Is despair increasing?
- *Politics*: what do people think about this candidate or issue?
- *Prediction*: predict election outcomes or market trends from sentiment

Dan Jurafsky

Scherer Typology of Affective States

- **Emotion**: brief organically synchronized … evaluation of a major event
  - *angry, sad, joyful, fearful, ashamed, proud, elated*
- **Mood**: diffuse non-caused low-intensity long-duration change in subjective feeling
  - *cheerful, gloomy, irritable, listless, depressed, buoyant*
- **Interpersonal stances**: affective stance toward another person in a specific interaction
  - *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*
- **Attitudes**: enduring, affectively colored beliefs, dispositions towards objects or persons
  - *liking, loving, hating, valuing, desiring*
- **Personality traits**: stable personality dispositions and typical behavior tendencies
  - *nervous, anxious, reckless, morose, hostile, jealous*

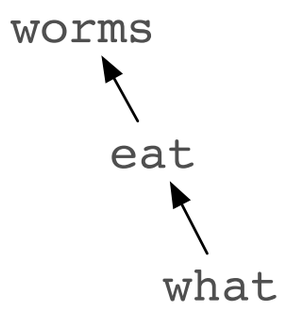
Dan Jurafsky

Question Answering

One of the oldest NLP tasks (punched card systems in 1961)

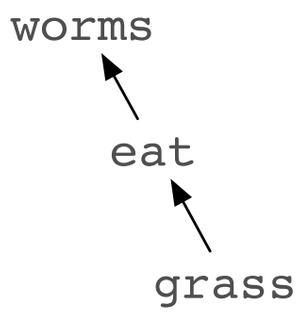
Simmons, Klein, McConlogue. 1964. Indexing and Dependency Logic for Answering English Questions. American Documentation 15:30, 196-204

Question: What do worms eat?

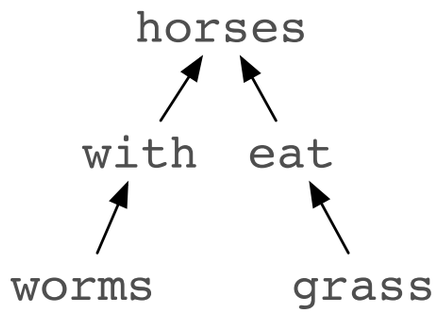


Potential Answers:

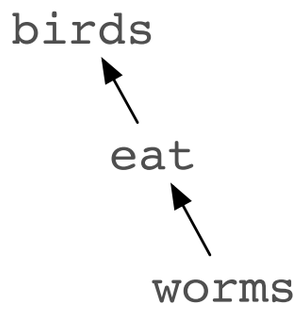
Worms eat grass



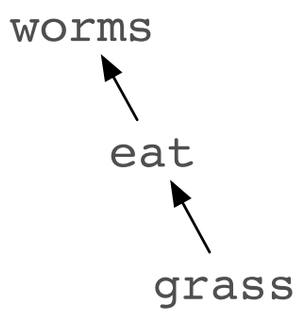
Horses with worms eat grass



Birds eat worms



Grass is eaten by worms



Dan Jurafsky

Question Answering: IBM's Watson

WILLIAM WILKINSON'S
"AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA"
INSPIRED THIS AUTHOR'S
MOST FAMOUS NOVEL

Bram Stoker

on Februa

Dan Jurafsky

Apple's Siri

Dan Jurafsky

Types of Questions in Modern Systems

- Factoid questions
  - *Who wrote "The Universal Declaration of Human Rights"?*
  - *How many calories are there in two slices of apple pie?*
  - *What is the average age of the onset of autism?*
  - *Where is Apple Computer based?*
- Complex (narrative) questions:
  - *In children with an acute febrile illness, what is the efficacy of acetaminophen in reducing fever?*
  - *What do scholars think about Jefferson's position on dealing with pirates?*

Text Summarization

- **Goal**: produce an abridged version of a text that contains information that is important or relevant to a user.

- **Summarization Applications**
  - **outlines or abstracts** of any document, article, etc
  - **summaries** of email threads
  - **action items** from a meeting
  - **simplifying** text by compressing sentences

What to summarize? Single vs. multiple documents

- **Single-document summarization**
  - Given a single document, produce
    - abstract
    - outline
    - headline
- **Multiple-document summarization**
  - Given a group of documents, produce a gist of the content:
    - a series of news stories on the same event
    - a set of web pages about some topic or question

Dan Jurafsky

Query-focused Summarization & Generic Summarization

- Generic summarization:
  - Summarize the content of a document
- Query-focused summarization:
  - summarize a document with respect to an information need expressed in a user query.
  - a kind of complex question answering:
    - Answer a question by summarizing a document that has the information to construct the answer

Information Extraction Tasks

- Named Entity Recognition
- Relation Extraction
- Event Extraction
- Coreference Resolution

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:

  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Christopher Manning

Named Entity Recognition (NER)

- A very important sub-task: find and classify names in text, for example:

  - The decision by the independent MP Andrew Wilkie to withdraw his support for the minority Labor government sounded dramatic but it should not further threaten its stability. When, after the 2010 election, Wilkie, Rob Oakeshott, Tony Windsor and the Greens agreed to support Labor, they gave just two guarantees: confidence and supply.

Person
Date
Location
Organi-
zation

Why Relation Extraction?

- Create new structured knowledge bases, useful for any app
- Augment current knowledge bases
  - Adding words to WordNet thesaurus, facts to FreeBase or DBPedia
- Support question answering
  - The granddaughter of which actor starred in the movie "E.T."?

```
(acted-in ?x "E.T.")(is-a ?y actor)(granddaughter-of ?x ?y)
```

- But which relations should we extract?

Event Extraction

Event Roles in Terrorism Domain:

- Perpetrators, victims, targets, weapons

Lopez Albujar, former defense minister, was killed with pistols by three young individuals. His two bodyguards were also shot to death.

→

Perp: *three young individuals.*
Vict: *Lopez Albujar*, *His two bodyguards*
Weap: *pistols*

Event Roles in Management Succession Domain:

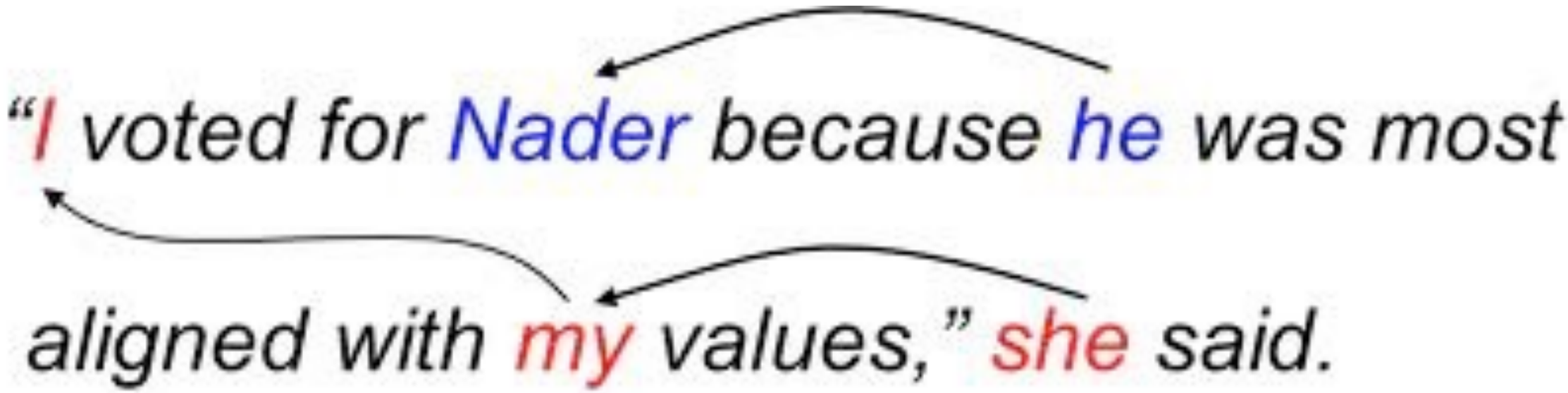
- People and companies in corporate management changes.

Daniel Glass was named president of EMI Records Group, He succeeds Charles Koppelman.

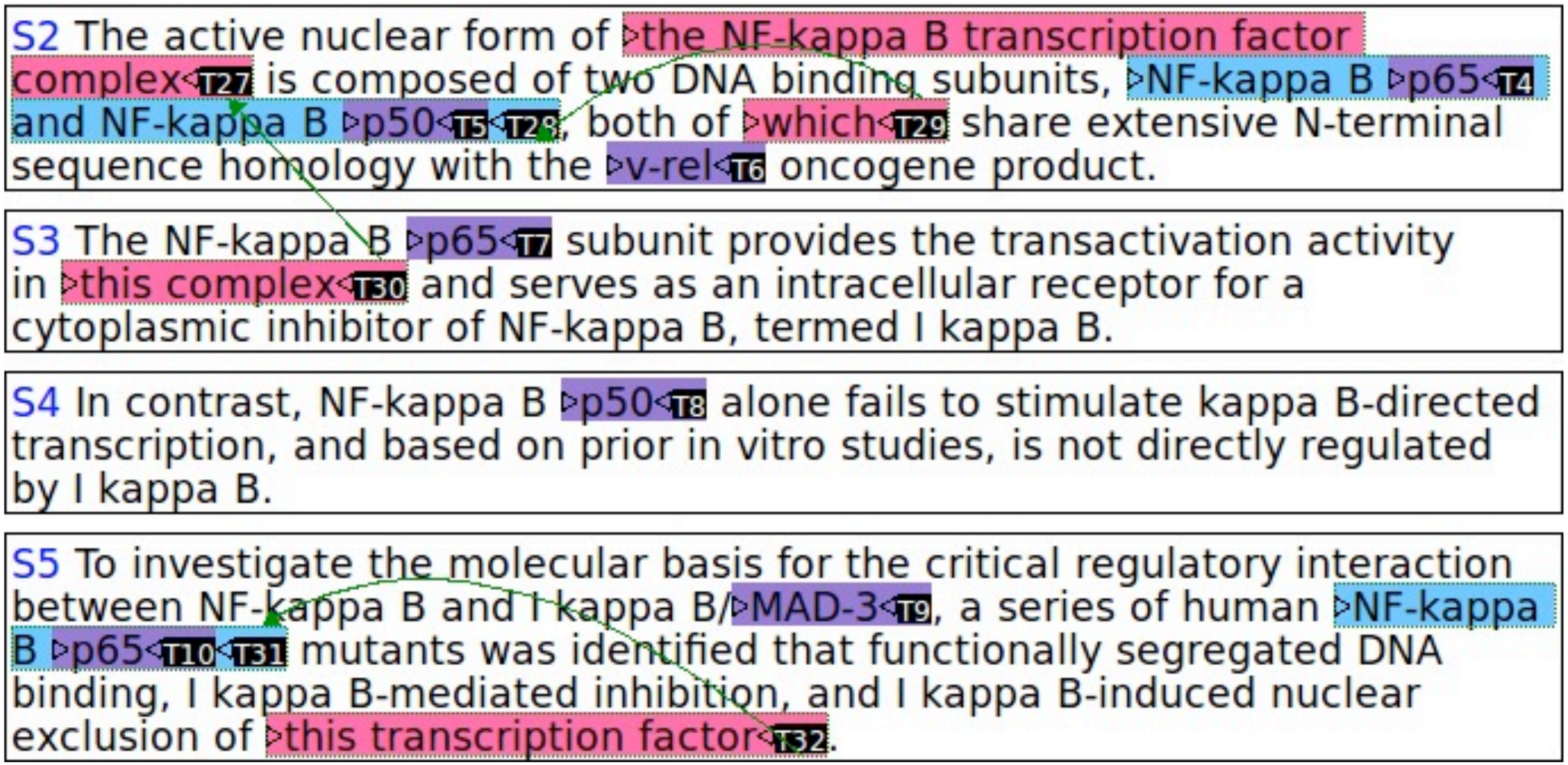
→

Per_In: *Daniel Glass*
Per_Out: *Charles Koppelman.*
Post: *president*
Org: *EMI Records Group*

Coreference Resolution



Coreference Resolution



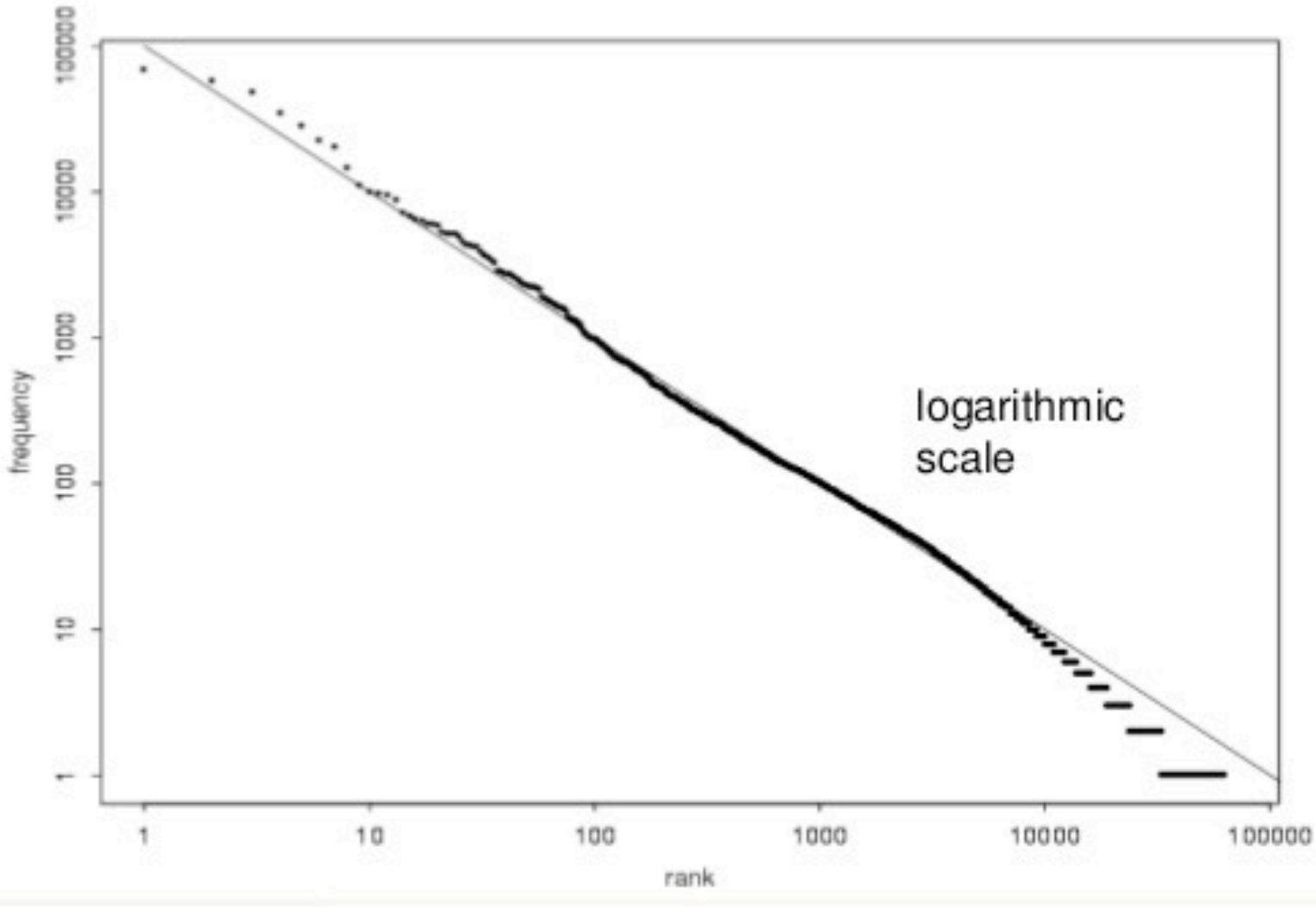
Ambiguities inherent in Language

- Language is succinct and expressive.
- Human resolve ambiguities naturally.
- Human own and can effortlessly resort to their rich common sense knowledge, domain knowledge and contexts.

Zipf's Law

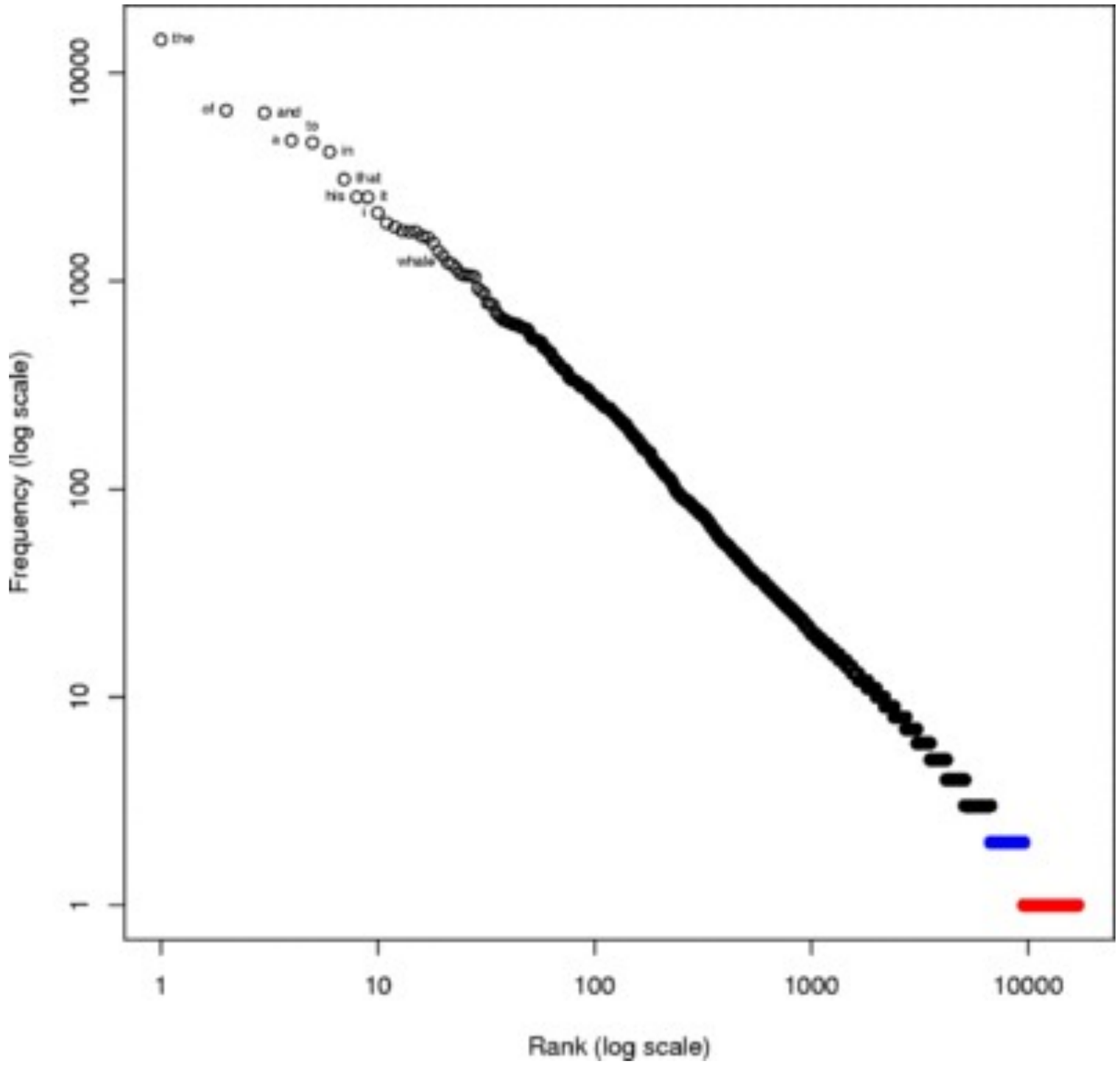
- the frequency of any word is inversely proportional to its rank: f = K / r
- fat-tail, most words occur only a couple of times
- high lexical diversity -> data sparseness

Illustration of Zipf's Law



(Brown Corpus, from M&S p. 30)

- Brown Corpus: A balanced corpus of written American English in 1960 (except poetry!), 1 million words.



- the novel: "The Whale", 44% words : one time

Following lectures

- IE summary: tasks and approaches
- NLP basics: syntax and semantic
- Machine Learning basics