Natural Language Processing
Introduction
Part 1

Sudeshna Sarkar

17 July 2019

Natural Language Processing

- NLP is focused on developing systems that allow computers to communicate with people using natural language.
- Also concerns how computational methods can aid the understanding of human language.
- Automating Language
  - Analysis Language → Representation
  - Generation Representation → Language
  - Acquisition Obtaining the representation and necessary algorithms, from knowledge and data

Language Processing

- Goals can be very ambitious
  - True text understanding
  - Good quality translation
- Or goals can be practical
  - Web search engines
  - Question Answering
  - Machine Translation services on the Web
  - Speech synthesis
  - Voice recognition
  - Conversational Agents
  - Summarization
- Natural language technology not yet perfected

  But still good enough for several useful applications

Logistics

- Moodle / Piazza forum for slides and discussion
- Moodle for assignment upload
- Textbook: Dan Jurafsky and James Martin Speech and Language Processing
- Instructor: Sudeshna Sarkar
- Teaching Assistants:
  - Debanjana Kar
  - Ishani Mondal
  - Sukannya Purakayastha

Logistics

- Attendance: Compulsory (5 marks)
- Assignments / Projects/ Quiz : 40 marks
- Midterm : 25 marks
- Endterm : 30 marks

Pre-requisites

- Probability and Statistics
- Machine learning
- Neural networks

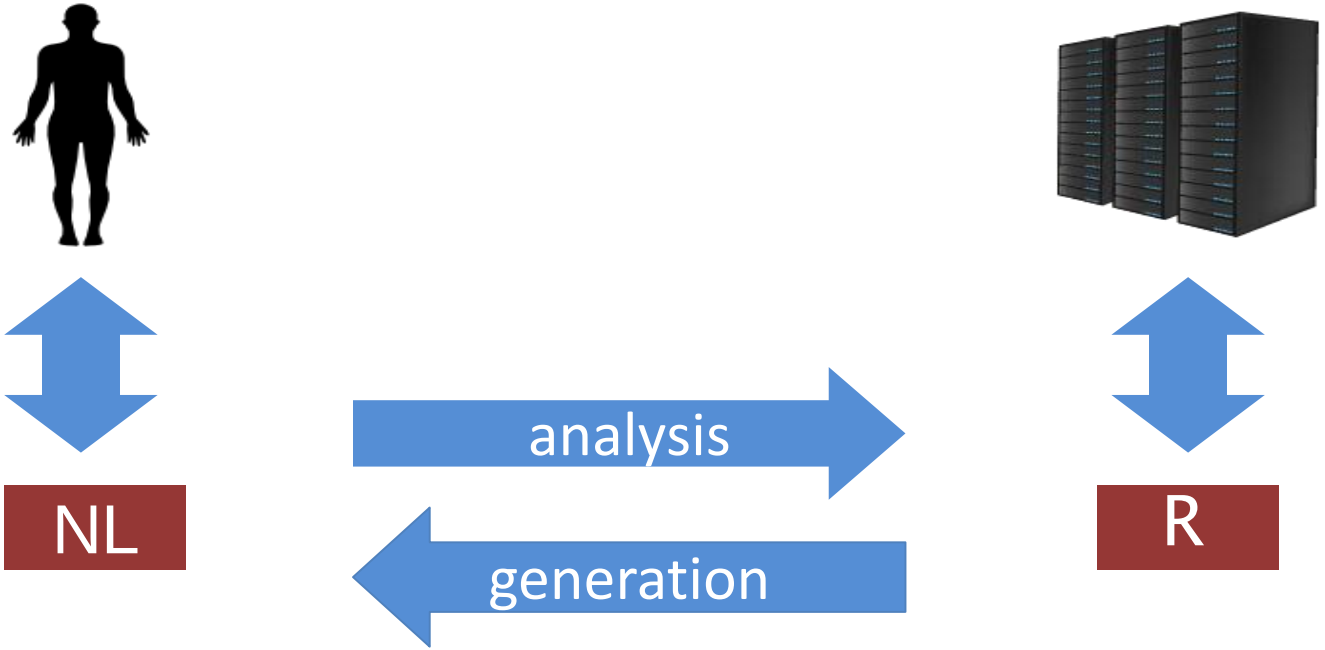
Course Focus

- Linguistic Issues
- Modeling Techniques
  - Probabilistic
  - ML
- Engineering Methods
- Multilinguality

- Science Goal: Understand the way language operates

- Engineering Goal: Build systems that analyse and generate language

What does it mean to “know” a language?



Examples of End Systems

- Text classification
- Machine translation, information extraction, dialog interfaces, question answering…
- human-level comprehension

Brief History of NLP

- 1940s –1950s: Foundational Insights
  - Two foundational paradigms
    - Automaton
    - Probabilistic / Information-Theoretic Models
- 1957-1970: The two camps
  - Symbolic paradigm: Chomsky and others on formal language theory and generative syntax
  - Stochastic paradigm

NLP History

- 1970 – 1983: FOUR paradigms
  - Stochastic
  - Logic-based
  - Natural language understanding
  - Discourse modeling
- 1983-1993: Empiricism and Finite State Models
- 1994-1999: The fields come together. Probabilistic and data-driven models
- Rise of ML: 2000 –
  - Lots of data and compute

Three Generations of NLP

- Hand-crafted Systems –Knowledge Engineering [1950s–]
- Automatic, Trainable (Machine Learning) Systems with engineered features [1985s–2012]
- Automatic, Trainable Neural architectures with no/limited engineered features [2012--]

NLP is Hard

- Ambiguity
- Ill-defined problems

- AI-complete

- The city council refused the demonstrators a permit because they _____ violence

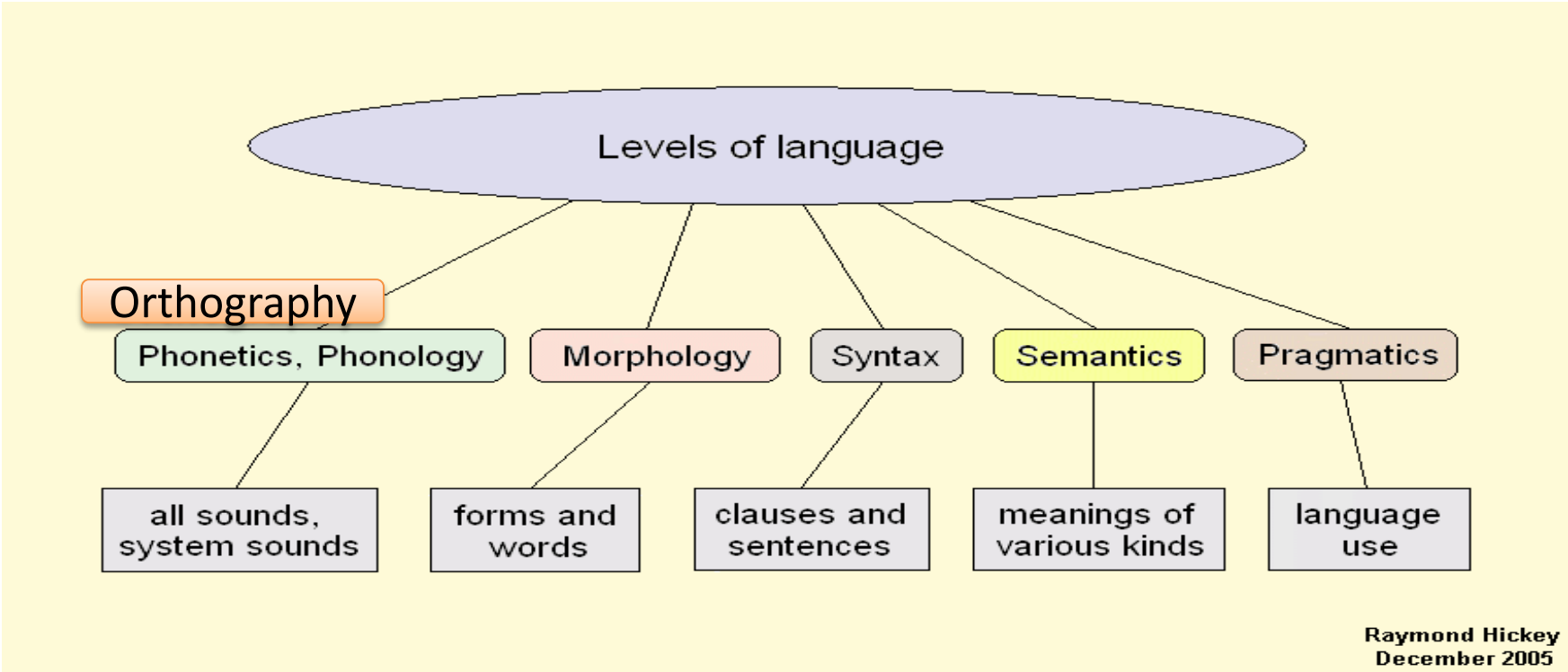
- The city council refused the demonstrators a permit because they feared violence
- The city council refused the demonstrators a permit because they advocated violence

- Headlines

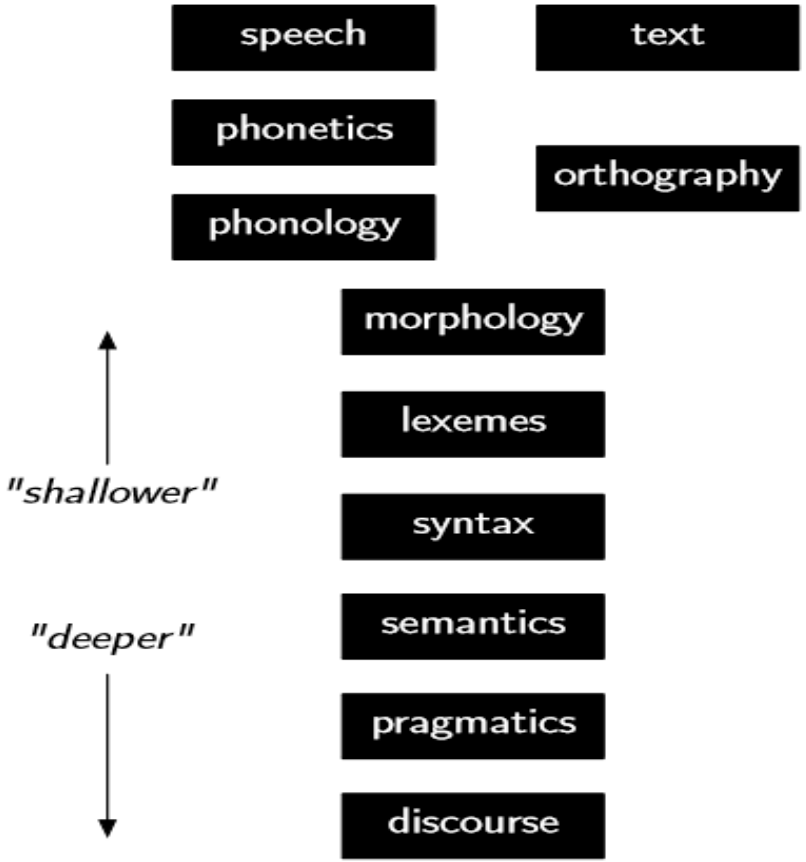
  Syntactic/semantic ambiguity: parsing needed to resolve these, but need context to figure out which parse is correct

  - Teacher Strikes Idle Kids
  - Hospitals Sued by 7 Foot Doctors
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half

slide credit: Dan Klein


Levels of language
Orthography
Phonetics, Phonology
Morphology
Syntax
Semantics
Pragmatics
all sounds, system sounds
forms and words
clauses and sentences
meanings of various kinds
language use
Raymond Hickey
December 2005

Levels of Linguistic Representation



Complexity of Linguistic Representations

- Richness: there are many ways to express the same meaning, and immeasurably many meanings to express.
- Each level interacts with the others.
- There is tremendous diversity in human languages.
  - Languages express the same kind of meaning in different ways
  - Some languages express some meanings more readily/often

Natural language understanding

- Uncovering the mappings between the linear sequence of words and the meaning that it encodes.

- Representing this meaning in a useful (usually symbolic) representation.

- By definition - heavily dependent on the target task
  - Words and structures mean different things in different contexts
  - The required target representation is different for different tasks.
- Appropriateness of a representation depends on the application.

Why is NLP Hard?

- The mappings between words, their linguistic structure and the meaning that they encode is extremely complex and difficult to model and decompose.
- Natural language is very ambiguous
  - **Lexical (word level) ambiguity** -- different meanings of words
  - **Syntactic ambiguity** -- different ways to parse the sentence
  - **Interpreting partial information** -- how to interpret pronouns
  - **Contextual information** -- context of the sentence may affect the meaning of that sentence.
- Noisy Input

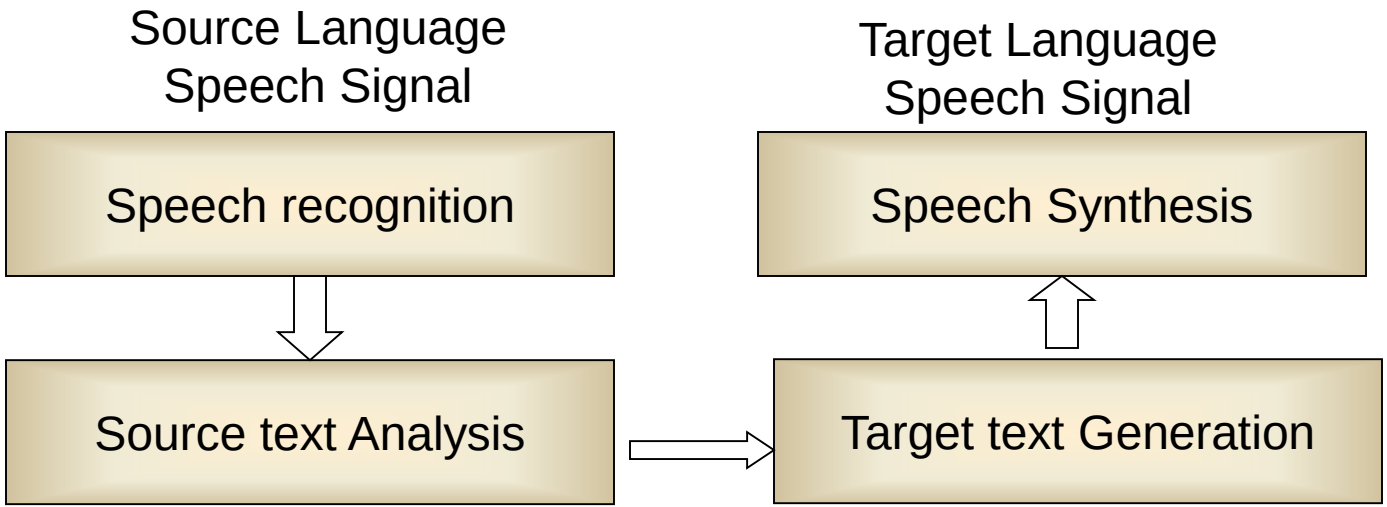
Complexity of Linguistic Representations

- Richness: there are many ways to express the same meaning, and immeasurably many meanings to express.
- Each level interacts with the others.
- There is tremendous diversity in human languages.
  - Languages express the same kind of meaning in different ways
  - Some languages express some meanings more readily/often

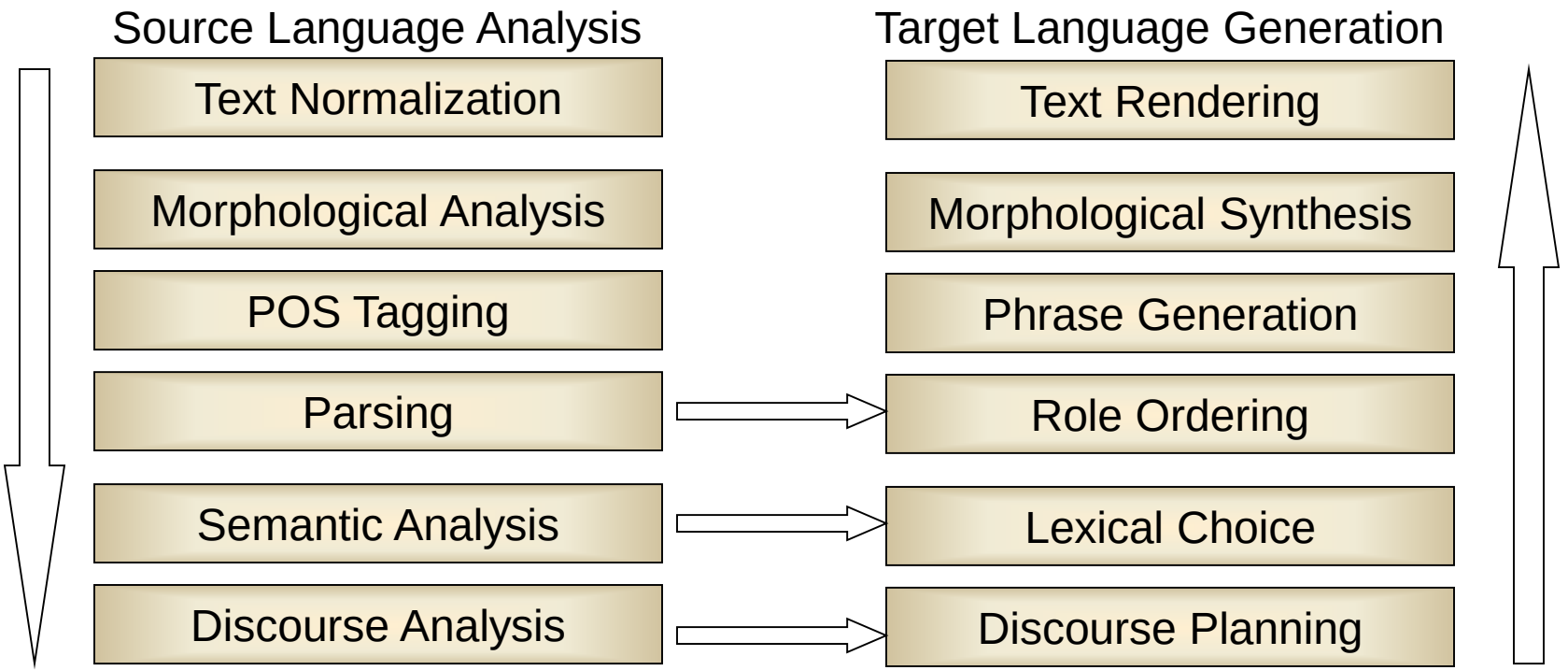
Components of NLP

- **Natural Language Understanding**
  - Mapping the given input in the natural language into a useful representation.
  - Different level of analysis required: morphological, syntactic, semantic, discourse
- **Natural Language Generation**
  - Producing output in the natural language from some internal representation.
  - Different level of synthesis required:
    - deep planning (what to say),
    - syntactic generation

The Big Picture



The Reductionist Approach



Knowledge of Language

- **Phonology** – concerns how words are related to the sounds that realize them.
- **Morphology** – concerns how words are constructed from more basic meaning units called morphemes. A morpheme is the primitive unit of meaning in a language.
- **Syntax** – concerns how can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of other phrases.
- **Semantics** – concerns what words mean and how these meaning combine in sentences to form sentence meaning. The study of context-independent meaning.

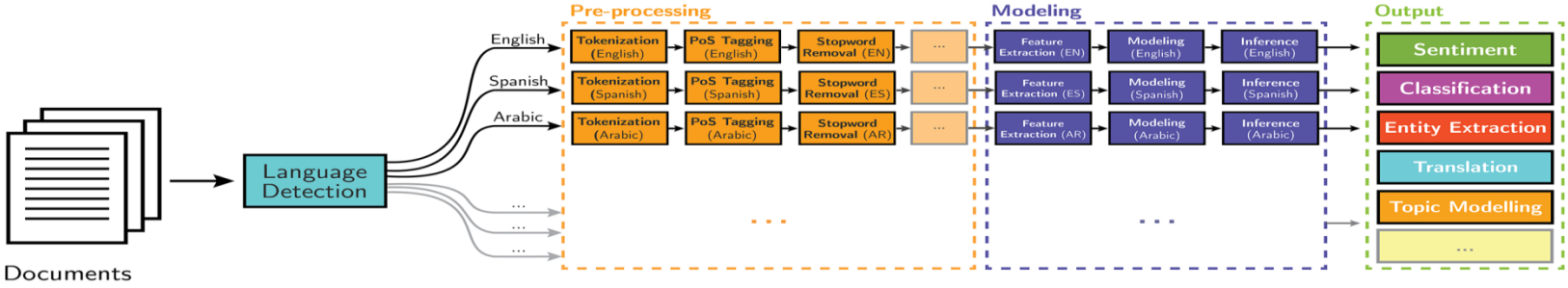
Knowledge of Language

- **Pragmatics** – concerns how sentences are used in different situations and how use affects the interpretation of the sentence.
- **Discourse** – concerns how the immediately preceding sentences affect the interpretation of the next sentence.For example, interpreting pronouns and interpreting the temporal aspects of the information.
- **World Knowledge** – includes general knowledge about the world. What each language user must know about the other's beliefs and goals.

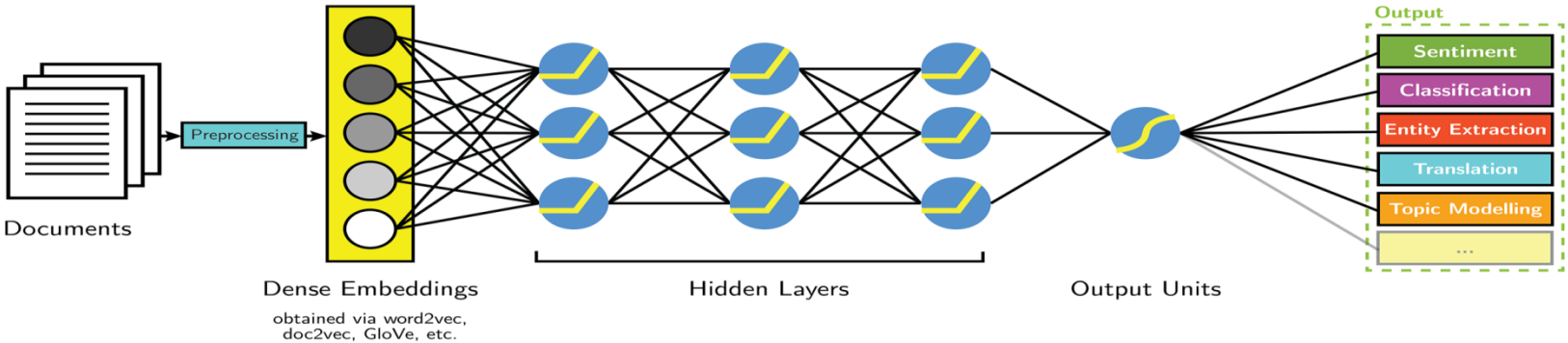
Two (or three) views of NLP

1. Classical View: Layered Processing; Various Ambiguities
2. Statistical/Machine Learning View
3. Deep Learning View

Classical NLP



Deep Learning-based NLP



AYLIEN