Introduction

Lecturers: Chenhua Lin

With guest lectures on NLG from Arria Data2Text

Reading: Chapter 1 (Jurafsky&Martin)

Definition of NLP

- Natural-language processing (NLP) systems are computer programs that process texts in human languages.
  - Written vs. spoken language
  - Understanding vs. generation vs. dialogue
  - English vs. French vs. Japanese vs…. (dialects?)
  - Domain Variation: social media vs newspapers

Example Applications

- Your phone converts speech into text (Siri, etc)
- MS Word can correct your grammar
- Your phone uses text prediction
- Google translates Web pages to different languages
- Arria Data2Text generate reports for oil companies from sensor data
- Text analytics (sentiment analysis, etc) is a billion dollar industry

Applications

Other applications include

- Generating weather reports - SumTime
- Opinion Mining
- Information extraction
- Text and speech summarization
- Computer-aided language learning

In general, NLP is now mainstream, but mixed success in the real-world

- Better understanding of language and “language engineering” still needed!

What is Language?

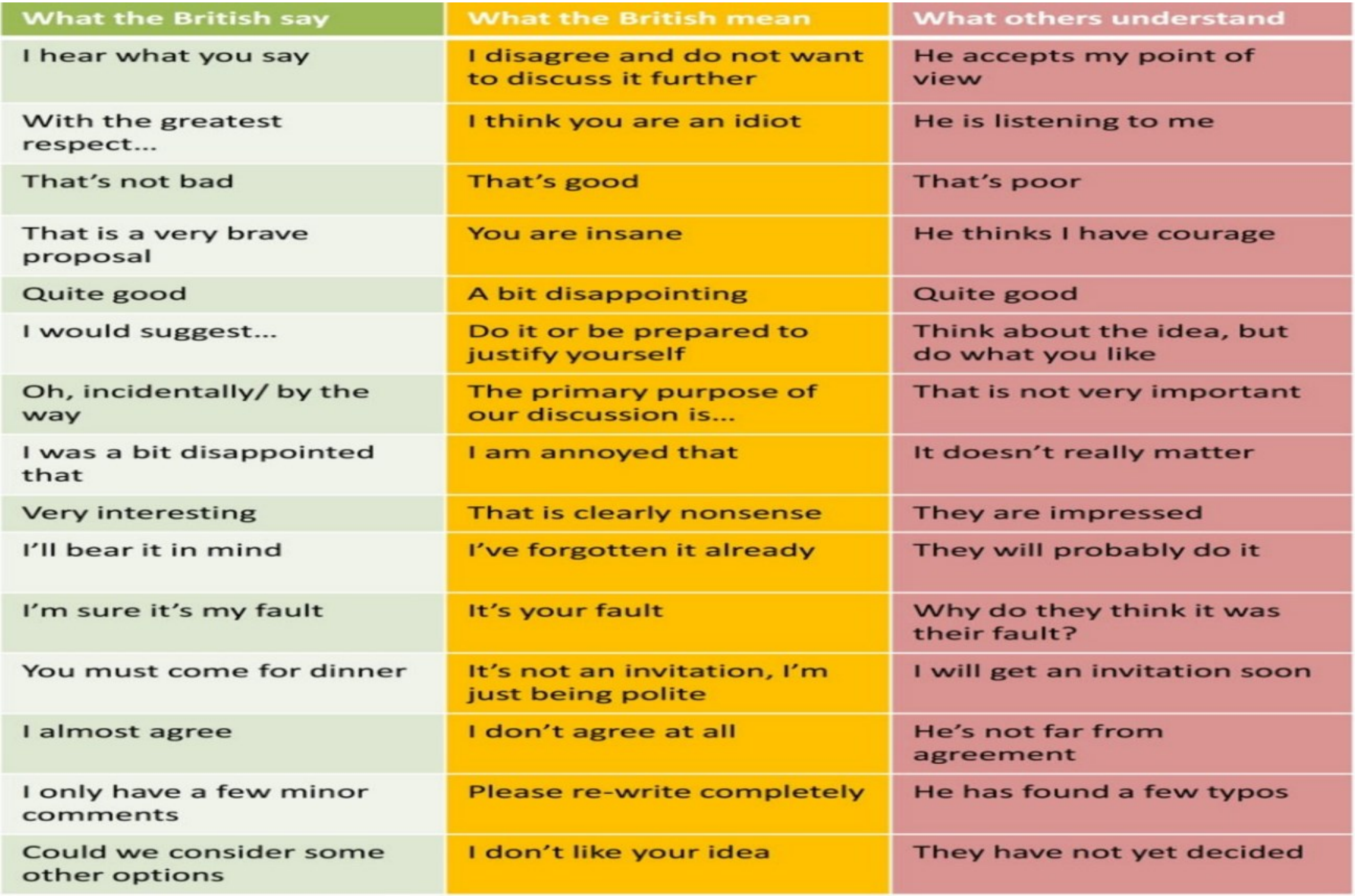
The study of language breaks down into a number of fields:

- Phonetics
  - sound signal <-> phonemes
- Morphology
  - eat, eating, eats, eaten, ate
- Syntax
  - the dog ate the cat
  - the cat ate the dog

What is Language

- Semantics
  - Delete all text files -> rm *.txt
- Pragmatics
  - Do you know what time it is?
  - Can I have some cake?

Pragmatics



What the British say	What the British mean	What others understand
I hear what you say	I disagree and do not want to discuss it further	He accepts my point of view
With the greatest respect…	I think you are an idiot	He is listening to me
That's not bad	That's good	That's poor
That is a very brave proposal	You are insane	He thinks I have courage
Quite good	A bit disappointing	Quite good
I would suggest…	Do it or be prepared to justify yourself	Think about the idea, but do what you like
Oh, incidentally/ by the way	The primary purpose of our discussion is…	That is not very important
I was a bit disappointed that	I am annoyed that	It doesn't really matter
Very interesting	That is clearly nonsense	They are impressed
I'll bear it in mind	I've forgotten it already	They will probably do it
I'm sure it's my fault	It's your fault	Why do they think it was their fault?
You must come for dinner	It's not an invitation, I'm just being polite	I will get an invitation soon
I almost agree	I don't agree at all	He's not far from agreement
I only have a few minor comments	Please re-write completely	He has found a few typos
Could we consider some other options	I don't like your idea	They have not yet decided

Natural Language Processing

- Use “black-box” models based on statistics or machine learning
- Implement algorithms and data structures based on linguistic theories
- Create linguistic resources which describe a language
  - dictionaries, grammars, corpora, …

Approaches to NLP

- Implement algorithms and data structures based on linguistics theories (grammar and parsing, etc)
- Use “black box” models based on statistics or machine learning (often classification problems, like sentiment analysis)

Example

- How to interpret a speech signal?

1) I scream is delicious

2) Ice cream is delicious

Example

- How to interpret a speech signal?
  1) I scream is delicious
  2) Ice cream is delicious
- Linguistic model
  - (2) is grammatical, (1) isn’t
- Statistical model
  - “Ice cream is” occurs much more often than “I scream is”

Example 2

- How to extract relationships from:
  - The word of the Lord came to Zechariah, son of Berekiah, son of Iddo, the prophet.

Example 2

The word of the Lord came to Zechariah, son of Berekiah, son of Iddo, the prophet.

son_of (Zecharia, Berekiah)

son_of(Zecharia, Iddo)

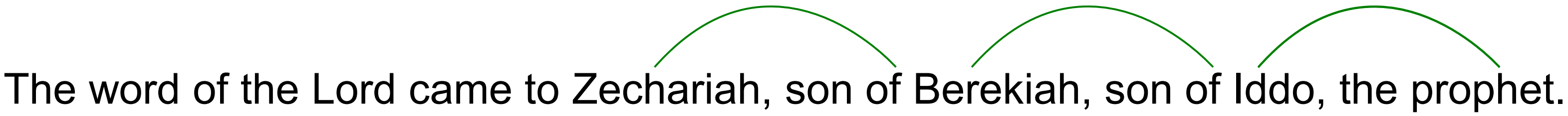
son_of(Berekiah, Iddo)

prophet(Iddo)

prophet(Berekiah)

prophet(Zechariah)

Example 2: Local Attachment Heuristic



The word of the Lord came to Zechariah, son of Berekiah, son of Iddo, the prophet.

son_of (Zecharia, Berekiah) ✓

son_of(Zecharia, Iddo)

son_of(Berekiah, Iddo) ✓

prophet(Iddo) ✓

prophet(Berekiah)

prophet(Zechariah)

History – 1940’s and 1950’s

- Fundamental theoretical developments:
  - Formal language theory (e.g. Chomsky)
  - Noisy channel model for transmission of language (Shannon and Weaver)
- Naïve optimism about Machine Translation
  - The spirit is willing but the flesh is weak
  - The Vodka is strong but the meat is rotten
- The beginnings of Information Retrieval:
  - Luhn (1957): “the frequency of word occurrence in an article furnishes a useful measurement of word significance”

History – 1960s

- Clear division between speech and language processing communities
- Symbolic models inspired by Chomsky’s context-free and transformational grammar
- Simple conversation systems (e.g. Weizenbaum’s ELIZA) and understanding systems using pattern matching
- Statistical methods used for OCR and authorship attribution
- First online Corpora:
  - Brown corpus of American english (1 million words)

History – 1970s

- Explosion of activity in NLP, stimulated partly by Winograd’s SHRDLU program which simulated a robot in a domain of toy blocks
- Explicit use of grammars and parsing
- Development of hidden Markov models for speech recognition
- Logic-based approaches to syntax and reasoning
  - Prolog, definite clause grammars
  - Lunar question-answering system
- Start of the study of discourse structure (Grosz)
- The need for knowledge (the Yale School)

History – 1980s

- Construction of Question-Answering systems for small domains (PHLIQA, Core language Engine)
- Revival of work on finite-state models, e.g. for morphology
- Revival of probabilistic models based on IBM models of speech recognition – part-of-speech tagging, statistical parsing, connectionist approaches.
- Start of serious work in natural language generation

History – 1990s to present

- Standard use of probabilistic and data-driven models throughout the field, informed by theoretical insights
- Increasingly rigorous evaluation methodologies
- Commercial exploitation aided by increasingly powerful hardware
- Beginning of work in information extraction (JASPER: real time extraction of financial news)
- Commercial exploitation (Billion $ business)
  - e.g. Machine Translation, Sentiment Analysis and Opinion mining, Information Extraction, Text summarization

Challenges for NLP: Ambiguity

Perhaps the most significant problem for language recognition/interpretation/understanding:

- Many sentences are ambiguous
  - Time flies like an arrow
  - I made her duck
  - Jack invited Mary to the Halloween ball.
- Computer sees ambiguities we don’t
  - Visiting parents can be a pain.
  - Visiting museums can be a pain.
- Resolve with knowledge
  - world knowledge, contextual knowledge, statistical knowledge

Challenges for NLP: Ambiguity

Resolving ambiguity can require us to cross sentence boundaries:

- Pronoun Resolution:
  - Merck & Co. formed a joint venture with Ache Group, of Brazil. It will be called Prodome Ltd.
  - Merck & Co. formed a joint venture with Ache Group, of Brazil. It will own 50% of the new company to be called Prodome Ltd.
  - Merck & Co. formed a joint venture with Ache Group, of Brazil. It had previously teamed up with Merck in two unsuccessful pharmaceutical ventures.

Challenges for NLP: Coreference

- Some interesting examples of co-reference:
  - Perhaps the key was under a flowerpot. He looked under them.
  - Frank was angry, and so was I.
  - Because he was very cold, David put on his coat
  - John asked Mary to go to the party. They arrived at the same time.
  - The plane landed and the pilot got out.

Challenges for NLG: Choice

The “analogue” of ambiguity for language *generation.*

- Choosing a text structure, syntactic construction, word or intonation. E.g.
  - John made a mistake/ John made a blunder
  - John made a mistake/ **He** made a mistake
  - John ate the cake/ The cake was eaten by John
  - John gave Susan the book/ John gave the book to Susan
  - I saw the elderly man. He was sleeping/ I saw the elderly man, who was sleeping

Research

We will be discussing

- State-of-the-art systems which don’t work perfectly, but often well enough for some practical purpose
- Theories and models which are the best we can do but might still have many problems

NLP is a research area!

Content of this course

- Words – spelling, morphology, sequences (n-grams)
- Syntax - grammars, POS tagging, DCGs
- Parsing – bottom-up, top-down, charts, statistics
- Semantics – compositionality, logics, statistical
- Argumentation
- Summarisation
- Information Extraction
- Pragmatics
- Discourse Theories
- Natural Language Generation

Recommended Textbook

- Daniel Jurafsky and James Martin: “Speech and Language Processing”, Prentice Hall, 2000.
- Copies in the library
- An excellent book, which covers much more than we need but is the nearest thing to a single book covering the whole course content

Structure of this Course

- Lectures (1hr): 2 per week
- Practicals (2hr): 1 per week (You will build and evaluate real systems)
- Assessment: 75% exam (in December); 25% continuous assessment