Course Overview & Introduction

COMP90042

Natural Language Processing

Lecture 1

Semester 1 Week 1
Jey Han Lau

Prerequisites

- Machine learning basics (COMP30027, COMP90049, COMP90051)
  - Modules → Welcome → Machine Learning and Linguistics Readings
- Python
- No knowledge of linguistics or advanced mathematics is assumed
- Caveats – Not “vanilla” computer science
  - Involves some basic linguistics, e.g., syntax and morphology
  - Requires maths, e.g., optimisation, linear algebra, dynamic programming

Expectations and outcomes

- Expectations
  - develop Python skills
  - keep up with readings
  - lecture/discussion board participation
- Outcomes
  - Practical familiarity with range of text analysis technologies
  - Understanding of theoretical models underlying these tools
  - Competence in reading research literature

Assessment

- **Assignments** (25% total for 3 activities)
  - 2 programming exercises
    - Released in week 3 and 5; 1 week to complete
  - 1 peer review
    - Released in week 11/12
- **Group Project** (35%)
  - Released in week 6/7; 4 weeks to complete
- **Exam** (40%)
  - 2 hours, on-campus (format to be determined)
  - Covers content from lectures, workshop and prescribed reading
- **Hurdle** >50% exam (20/40), and >50% for assignments + project (30/60)

Subject Coordinators

Jey Han Lau

Caren Han

Tutors

- Bryan Chen (Head Tutor)
- Huygaa Batsuren
- Nate Carpenter
- Rena Gao
- Rahmad Mahendra
- Jinrui Yang
- Rongxin Zhu

Recommended Texts

- Texts:
  - *Jurafsky and Martin*, *Speech and Language Processing*, 3rd ed., Prentice Hall. draft
  - *Eisenstein*; *Natural Language Processing*, Draft 15/10/18
  - Goldberg; *A Primer on Neural Network Models for Natural Language Processing*
- Recommended for learning python:
  - *Steven Bird, Ewan Klein and Edward Loper*, *Natural Language Processing with Python*, O'Reilly, 2009

Contact hours

- Lectures
  - Tue 12:00-13:00 Carrillo Gantner Theatre
  - Friday 11:00-12:00 Carrillo Gantner Theatre
- Workshops: several across the week
  - Worksheets & programming exercises
- Method of contact — ask questions on the Canvas discussion board

Lecture Slides

- Preliminary version (v1) of some lecture slides have been published **(Modules > Lectures > Slides)**
- Lecture slides will continuously be updated
- Lecture recordings will be available after each lecture

Python

- Making extensive use of python
  - workshops feature programming challenges
  - provided as interactive ‘notebooks’
    - Modules → Using Jupyter Notebook and Python
  - assignment and project in python
- Using several great python libraries
  - NLTK (basic text processing)
  - Numpy, Scipy, Matplotlib (maths, plotting)
  - Scikit-Learn (machine learning tools)
  - keras, pytorch (deep learning)

Python

- New to Python?
  - Expected to pick this up during the subject, on your own time
  - Learning resources on worksheet

Natural Language Processing

- Interdisciplinary study that involves linguistics, computer science and artificial intelligence.
- Aim of the study is to understand how to design algorithms to process and analyse human language data.

Why process text?

- Masses of information ‘trapped’ in unstructured text
- How can we find or analyse this information?
- Let computers automatically reason over this data?
- First need to understand the structure, find important elements and relations, etc…
- Over 1000s of languages....

Language Generation Demo

Why are you interested in NLP?

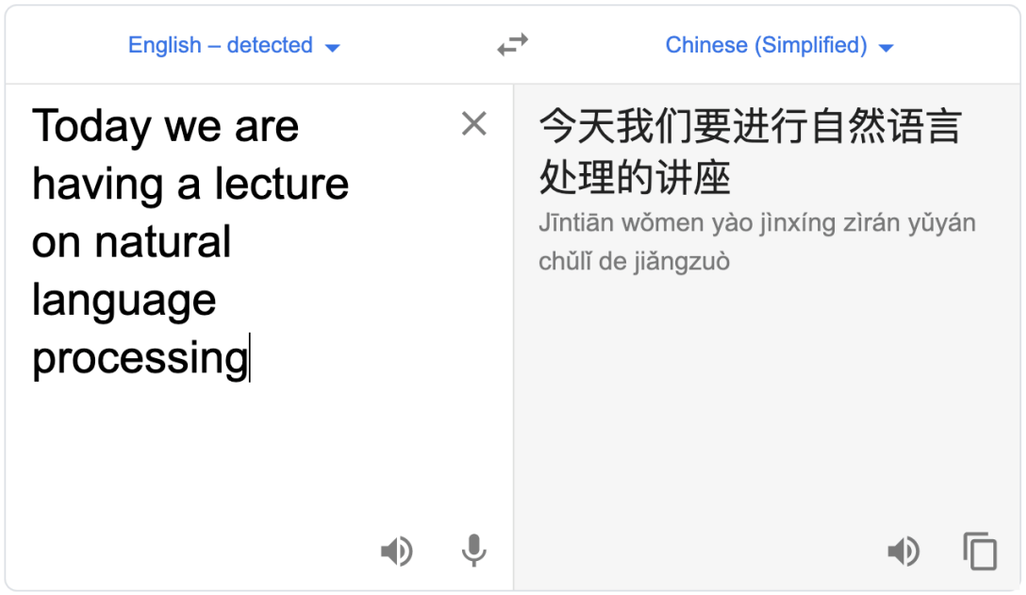
PollEv.com/jeyhanlau569

Motivating Applications (Sci-fi)

- Intelligent conversational agent, e.g. TARS in Interstellar (2014)
  - https://www.youtube.com/watch?v=wVEfFHzUby0
  - Speech recognition
  - Natural language understanding
  - Speech synthesis

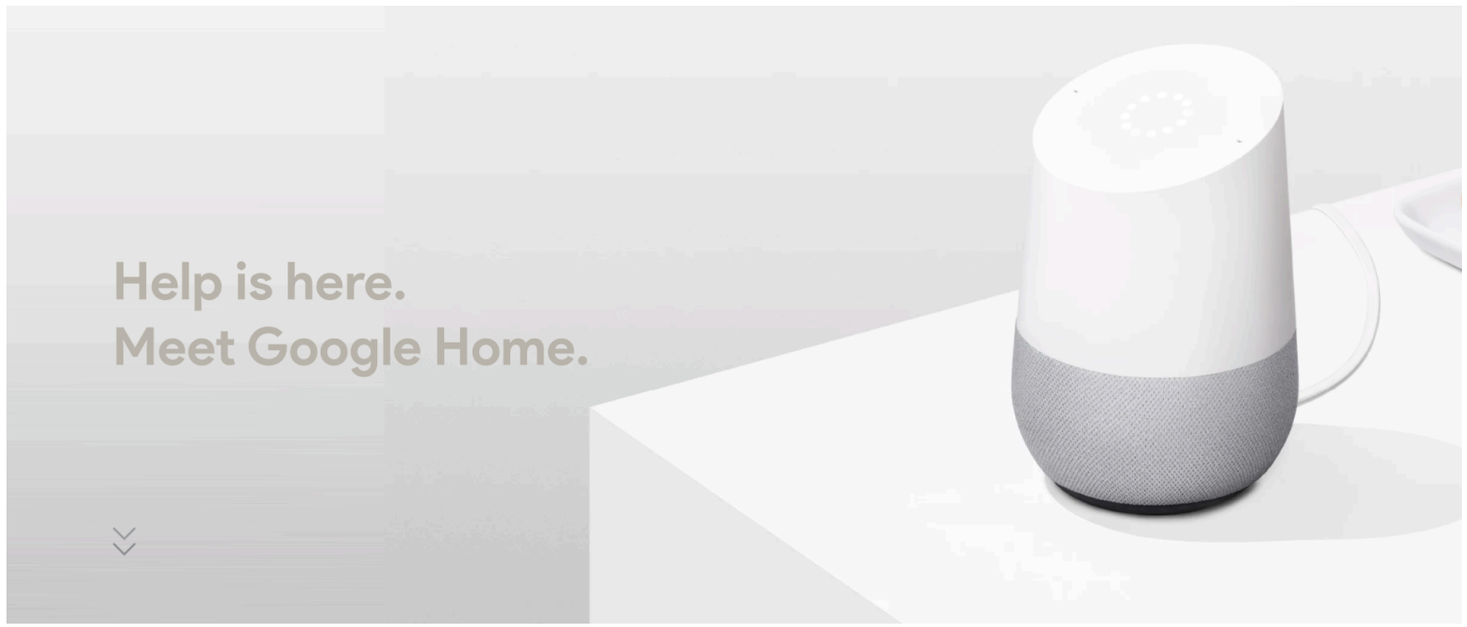
Motivating Applications (Real-world)

- ChatGPT
  - A very large language model trained on web-scale data
  - Lots of fine-tuning to get it to interact with humans


English – detected
Chinese (Simplified)
Today we are having a lecture on natural language processing
今天我们要进行自然语言处理的讲座
Jīntiān wǒmen yào jìnxíng zìrán yǔyán chǔlǐ de jiǎngzuò
Open in Google Translate
Feedback


google translate
google translate english to spanish
google translate audio
google translate english to french
google translate website
google translate statistics
translate to hindi
translate to english
inside google translate


who is the first australian prime minister
All
News
Images
Videos
Maps
More
Settings
Tools
About 78,100,000 results (1.18 seconds)
Prime Minister of Australia (1)
Edmund Barton
Australia's first prime minister, Edmund Barton at the central table in the House of Representatives in 1901.
en.wikipedia.org/wiki/Prime_Minister_of_Australia
Prime Minister of Australia - Wikipedia


Help is here.
Meet Google Home.

Course Overview

- **Word, sequences, and documents**
  - Text preprocessing
  - Language models
  - Text classification
- **Structure learning**
  - Sequence tagging (e.g. part-of-speech)
- **Deep learning for NLP**
  - Feedforward and recurrent models

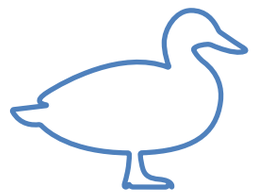
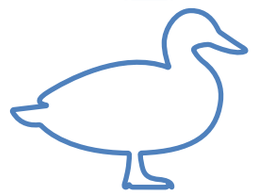
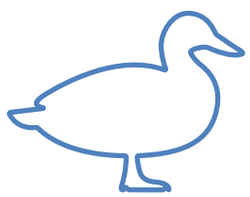
Course Overview

- **Semantics**
  - How words form meaning
- **Pretrained models**
  - Transformer, large language models
- **Applications**
  - Named Entity Recognition
  - Reading Comprehension
  - Dialog System

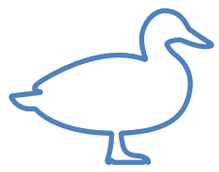
Are Machines Intelligent Yet?

- Alan Turing, famously proposed the **Turing test**, to assess whether a machine is intelligent
- Alan Turing predicted in 1950 that by 2000 a machine with 10 gigabytes of memory has 30% of fooling the human interrogator.
- ~~The smartest conversational agent we have today are far away from being truly intelligent…~~

Challenges of Language: Ambiguity

- *I made her duck:*
  - *I cooked*  *for her*
  - *I cooked*  *belonging to her*
  - *I caused her to quickly lower her head or body*
  - *I waved my magic wand and turned her into a* 
- Why so many possible interpretations?

Challenges of Language: Ambiguity

- *Duck* can mean:
  - Noun: 
  - Verb: move head or body quickly down (e.g. to dodge something)
- *Her* can be a dative pronoun (i.e. indirect object to a verb) or possessive pronoun
- *Make* is syntactically ambiguous:
  - Transitive (takes one object: *duck*)
  - Ditransitive (1st object: *her*; 2nd object: *duck*)
  - Can take a direct object and verb: object (*her*) is caused to perform the verbal action (*duck*)

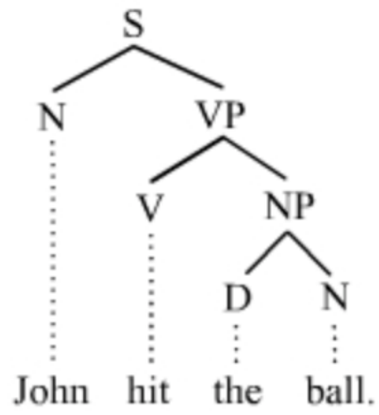
What are other challenges that made language processing difficult?

PollEv.com/jeyhanlau569

A brief history of NLP: 1950s

- "Computing Machinery and Intelligence", Alan Turing
  - Turing test: measure machine intelligence via a conversational test
- "Syntactic Structures", Noam Chomsky
  - Formal language theory: uses algebra and set theory to define formal languages as sequences of symbols
  - *Colourless green ideas sleep furiously*
    - Sentence doesn't make sense
    - But its grammar is fine
    - Highlights the difference between semantics (meaning) and syntax (sentence structure)

1960-1970s



- Symbolic paradigm
  - Generative grammar
    - Discover a system of rules that generates grammatical sentences
  - Parsing algorithms
- Stochastic paradigm
  - Bayesian method for optical character recognition and authorship attribution
- First online corpus: Brown corpus of American English
  - 1 million words, 500 documents from different genres (news, novels, etc)

1970-1980s

- Stochastic paradigm
  - Hidden Markov models, noisy channel decoding
  - Speech recognition and synthesis
- Logic-based paradigm
  - More grammar systems (e.g. Lexical functional Grammar)
- Natural language understanding
  - Winograd's SHRDLU
  - Robot embedded in a toy blocks world
  - Program takes natural language commands (*move the red block to the left of the blue block*)
  - Motivates the field to study semantics and discourse

1980-1990s

- Finite-state machines
  - Phonology, morphology and syntax
- Return of empiricism
  - Probabilistic models developed by IBM for speech recognition
  - Inspired other data-driven approaches on part-of-speech tagging, parsing, and semantics
  - Empirical evaluation based on held-out data, quantitative metrics, and comparison with state-of-the-art

1990-2000s: Rise of Machine Learning

- Better computational power
- Gradual lessening of the dominance of Chomskyan theories of linguistics
- More language corpora developed
  - Penn Treebank, PropBank, RSTBank, etc
  - Corpora with various forms of syntactic, semantic and discourse annotations
- Better models adapted from the machine learning community: support vector machines, logistic regression

2000s: Deep Learning

- Emergence of very deep neural networks (i.e. networks with many many layers)
- Started from the computer vision community for image classification
- Advantage: uses raw data as input (e.g. just words and documents), without the need to develop hand-engineered features
- Computationally expensive: relies on GPU to scale for large models and training data → large language models like chatgpt
- Contributed to the AI wave we now experience:
  - Home assistants, generative AI

Future of NLP

- Are NLP problems solved?
  - Machine translation still far from perfect
  - Summarise a novel
  - Conversational agent can be ‘smarter’
  - Smaller, parameter-efficient models
  - Not all NLP problems are generation problems (which large language models are particularly good at)