We now ask the question, 'What will happen when a machine takes the part of A in this game?'

...

These questions replace our original, 'Can machines think?'

Alan Turing

COMP6713 | 2025 T1 | UNSW

All images from Wikimedia commons, unless specified.

UNSW SYDNEY

1

COMP6713
Natural Language Processing (NLP)

Convener

Dr. Aditya Joshi

aditya.joshi@unsw.edu.au

Week 1

Introduction

Schedule

2025 Term 1

UNSW SYDNEY

2

Course Convener

Name: Aditya Joshi (he/him)

Office: K17-217B

Email: aditya.joshi@unsw.edu.au

Questions?

Technical: WebCMS Forum

Technical/personal: Email

Hobbies: Bushwalking, cooking, standup comedy [Sydney Fringe '24]

Research: Foundational and Applied Natural language processing

(https://unswnlp.github.io/)

Youtube screenshot

3

Course Convener

Name: Aditya Joshi (he/him)

Office: K17-217B

Email: aditya.joshi@unsw.edu.au

Questions?

Technical: WebCMS Forum

Technical/personal: Email

Hobbies: Bushwalking, cooking, standup comedy [Sydney Fringe '24]

Research: Foundational and Applied Natural language processing
(https://unswnlp.github.io/)

NLP and me:

- Masters Project (2010): Sentiment analysis of tweets

- PhD Thesis (2018): Investigations in Computational Sarcasm

Experience

- CSIRO's Data61: NLP for epidemic intelligence

- SEEK: NLP for candidate-job recommendation

- Notiv (startup acquired by Dubber): NLP for meeting summarisation

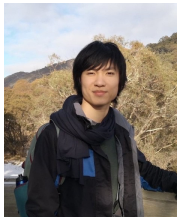
4

Course Team

Lihua Wang (Lily)

Martin Eftimoski

Rashini Liyanarachchi (Rashini)



Duc Anh Nguyen (Duke)

Zechen Li (Ryan)

Photos from LinkedIn

UNSW SYDNEY

5

6

UNSW SYDNEY | Australia's Global University

Part 0

Course Overview

7

COMP6713: Natural Language Processing

- Introductory course; second offering
- Philosophy:
  - What: Language phenomena
  - How: Computational models
  - Why: Historical background

Is this course an advanced disciplinary elective (or similar term) for you? Please contact the Student Nucleus Hub.

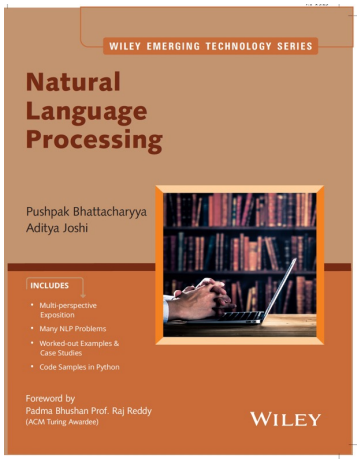
UNSW SYDNEY

8

Recommended Resources

1) Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. Online manuscript released January 12, 2025. https://web.stanford.edu/~jurafsky/slp3.

2) Pushpak Bhattacharyya and Aditya Joshi. 2023. Natural Language Processing. Kindle Edition released December 2023. Wiley. https://www.amazon.com.au/Natural-Language-Processing-Pushpak-Bhattacharyya-ebook/dp/B0CR64RX4T (Copies available in the Library)



9

Pre-requisites

Postgraduate: COMP9020 and (COMP9814 or COMP9444)*

Undergraduate: MATH1081 and (COMP3411 or COMP9444)*

Will I enjoy the course? (Not a complete list. Not necessary conditions)

- I am a language nerd
- I enjoy coding
- I like learning about AI models
- I often connect modern ideas with those from the past

*Exceptions made.

10

Learning Objectives

- CL01: Describe NLP problems such as POS tagging, sentiment analysis, information extraction and machine translation along with their challenges in terms of ambiguity resolution.
- CL02: Explain typical NLP approaches based on statistical and neural approaches.
- CL03: Use NLP libraries (e.g. NLTK, scikit-learn, Transformers) to implement the training of models for NLP problems and use them for inference.
- CL04: Design an NLP solution by selecting the NLP problem formulation, approach and evaluation strategy, by analysing the requirements of a specific application.

11

Why are we even studying pre-deep learning NLP?

- Large language models (LLMs) are far from perfect
- Pre-deep learning NLP approaches are a ‘small’ component of the course
- They are still used in several businesses.. in Australia
- Concepts from pre-deep learning NLP models help us appreciate the ingenuity of deep learning-based NLP
  - This goes to the last slide:
    - Linguistic examples help us understand the ‘what’ of NLP
    - Neural models help us understand the ‘how’ of NLP
    - Three-generational view helps us understand the ‘why’ of NLP

12

How is the 2025T1 offering different from 2024T1?

Student feedback:

- Overall: Strongly Positive
- Appreciated the engagement and humour in the class.
- Agreed that it was a valuable course.

13

How is the 2025T1 offering different from 2024T1?

Student feedback:

- Overall: Strongly Positive
- Appreciated the engagement and humour in the class.
- Agreed that it was a valuable course.

- Thought the assignment was not handled sufficiently well.
- Thought more engagement with the lecturer would have helped.
- Drag-and-drop questions in Moodle don't work well on all devices.

14

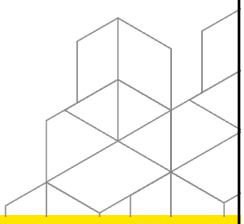
How is the 2025T1 offering different from 2024T1?

Student feedback:

- Overall: Strongly Positive
- Appreciated the engagement and humour in the class.
- Agreed that it was a valuable course.

- Thought the assignment was not handled sufficiently well.
  - Revised assignment with clear evaluation criteria.
- Thought more engagement with the lecturer would have helped.
  - Option of online consultation in addition to in-person consultation.
- Drag-and-drop questions in Moodle don't work well on all devices.
  - Cool, None.

- Also, new content in EVERY module to keep up with the fast-changing NLP landscape.
- New format for group project (industry project option) and final exam (Inspera – TO BE CONFIRMED).
- etc.



UNSW SYDNEY

15

Content Plan (Subject to Minor Revision)

Week#	Module	Topics
1	Introduction	NLP tasks, ambiguity resolution and generations of NLP, Ethical considerations, Black-box NLP libraries; Datasets **+ API calls for LLMs**
2	Representation learning	Grammar, Probabilistic language models, Word Vectors **+ Review of Sequential models**
3	Attention & Transformer	Attention, Fine-tuning, Prompt tuning
4	Language models	Encoder models, decoder models, LoRA, LangChain **+….**
5	Sentiment Analysis	Lexicons, Statistical classifiers, LSTM stacks, BERT-based models**+ Evaluation, Benchmarks**
7	POS tagging and NER	HMM-based POS tagging, POS tagging a seq2seq task, CRF-based NER
8	Machine translation	Rule-based, statistical, Neural, Decoding**+ ….**
9	Summarisation & Question-Answering	Extractive and abstractive summarization **+ …..**
10	Other NLP tasks	Shared tasks and benchmarks **+Detailed benchmarks**; NLP applications

+ Guest lectures; "How to read an NLP paper" and so on.

UNSW SYDNEY

16

Course Schedule

Lectures:

Thursday: 14:00-16:00. O'Shane 104

Friday: 11:00-13:00. Griff M18

Public Holiday (Good Friday): Friday 18 April 2025

Have you been added to WebCMS and Moodle?

Have you received your tutorial allotments?

https://webcms3.cse.unsw.edu.au/COMP6713/25T1

Tutorials

Name	Staff	Day	Start Time	End Time	Weeks	Room
H17A		Thu	17:00	18:00	1-5,7-10	Myers 540
		Thu	17:00	18:00	1-5,7-10	Myers 540
H17B		Thu	17:00	18:00	1-5,7-10	Myers 440
		Thu	17:00	18:00	1-5,7-10	Myers 440
F10B		Fri	10:00	11:00	1-5,7-8	Quad G026
		Fri	10:00	11:00	1-5,7-8	Quad G026
F13A		Fri	13:00	14:00	1-5,7-8	Law 276
		Fri	13:00	14:00	1-5,7-8	Law 276
F13B		Fri	13:00	14:00	1-5,7-8	Law 203
		Fri	13:00	14:00	1-5,7-8	Law 203
F14A		Fri	14:00	15:00	1-5,7-8	Mat 309
		Fri	14:00	15:00	1-5,7-8	Mat 309

17

Assessments (1/2)

Weekly Quizzes (20%):

- Moodle-based, closed-book quizzes
- Timeline:
  - Weekly: 2-5, 7-9
  - One attempt; no time limit
- Evaluation: Auto-marked
- Quiz will be open from Friday 5:25pm to Wednesday 5:25pm | SAMPLE quiz this week

Assignment (10%):

- Individual programming assignment
- Timeline:
  - Released in Week 3
  - Due in the end of week 5
  - Based on content from Weeks 1 to 3
- Evaluation:
  - Test scripts
  - Manual evaluation (style; correctness, etc.)

18

Assessments (2/2)

Group Project (25%):

- Groups of 3 – 5
- Timeline:
  - Week 4: Register your group and topic in consultation with the team
  - Project due on Monday of Week 10
- Deliverables:
  - Code, Report, Presentation/video

Final Exam (45%):

- 2 hours; during exam period
- Multiple-choice questions, short-answers and code analysis.

To pass this course, students must score more than 40% on the final exam.

19

Assessments (2/2)

Group Project (25%):

- Groups of 3 – 5
- Timeline:
  - Week 4: Register your group and topic in consultation with the team
  - Project due on Monday of Week 10
- Deliverables:
  - Code, Report, Presentation/video **+ In-person/Online Q&A**

Final Exam (45%):

- 2 hours; during exam period
- Multiple-choice questions, short-answers and code analysis.

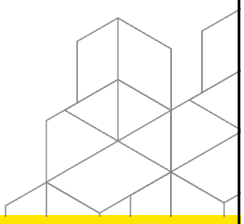
To pass this course, students must score more than 40% on the final exam.

Exciting news this year!
You may have the option of working on an industry-led group project.

20

Integrity

You acknowledge that you understand the UNSW Academic Integrity policy .
Also, a reminder about Academic Integrity in Programming Courses



21

Course Resources

- Lecture Slides & Recordings
- Tutorial Problem Sets (Not evaluated)
- Suggested textbooks:
  - *Natural Language Processing, Pushpak Bhattacharyya and Aditya Joshi, Wiley, 2023.*
  - *Speech and Language Processing, Daniel Jurafsky and James Martin, Prentice Hall, 2023.*

22

Part 1

NLP today & yesterday

Note: Brace yourself for terminology. If confused, please ask.

23

What is natural language processing?

Natural language processing (NLP) is the branch of artificial intelligence that deals with computational processing of text.

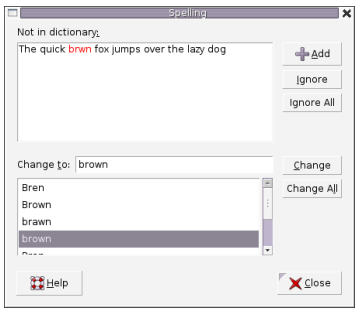
Also known as ‘computational linguistics’..

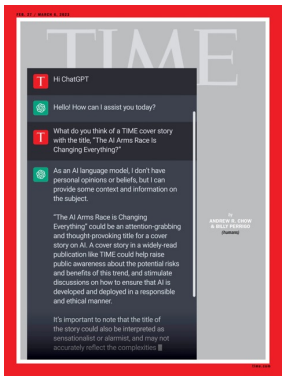
..and has a significant overlap with ‘human language technology’…. or ‘text analytics’.

- Natural languages: Language that humans use
- Artificial languages: Programming languages

24

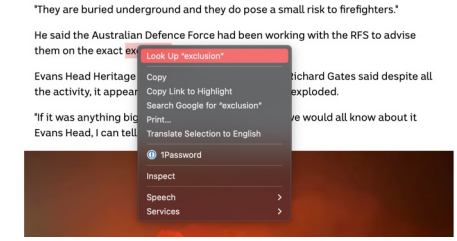
NLP Today





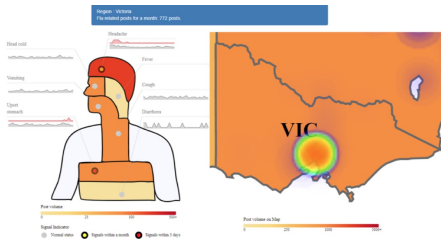
deepseek

grammarly



https://en.wikipedia.org/wiki/Siri
https://en.wikipedia.org/wiki/ChatGPT
https://en.wikipedia.org/wiki/Grammarly
https://en.wikipedia.org/wiki/Spell_checker
https://www.business-standard.com/world-news/deepseek-r1-chinese-ai-research-breakthrough-challenging-openai-explained-125012700327_1.html

NLP finds several applications



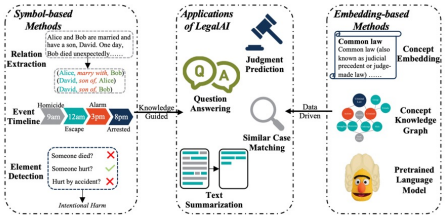


Figure 1: An overview of tasks in LegalAI.

LEGITIMATE	FAKE
Nintendo Switch game console to launch in March for $299 The Nintendo Switch video game console will sell for about $260 in Japan, starting March 3, the same date as its global rollout in the U.S. and Europe. The Japanese company promises the device will be packed with fun features of all its past machines and more. Nintendo is promising a more immersive, interactive experience with the Switch, including online playing and using the remote controller in games that don't require players to be constantly staring at a display.	**New Nintendo Switch game console to launch in March for $99** Nintendo plans a promotional roll out of it's new Nintendo switch game console. For a limited time, the console will roll out for an introductory price of $99. Nintendo promises to pack the new console with fun features not present in past machines. The new console contains new features such as motion detectors and immerse and interactive gaming. The new introductory price will be available for two months to show the public the new advances in gaming.

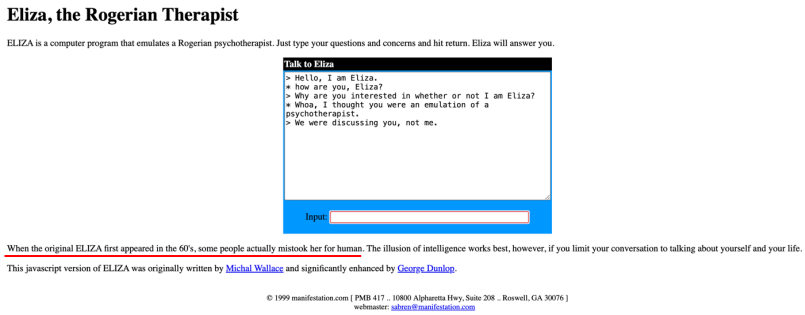
Table 2: Sample legitimate and crowdsourced fake news in the Technology domain

Brian Jin*, Aditya Joshi*, Ross Sparks, Stephen Wan, Cecile Paris and C Raina MacIntyre, 'Watch The Flu: A Tweet Monitoring Tool for Epidemic Intelligence of Influenza in Australia', **AAAI 2020**.
Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In ACL 2020.
Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In COLING.

Early NLP

NLP has long explored automation of natural language:

- 1964: ELIZA: Chatbot: Counsellor
- 1968: SYSTRAN: Translation system (Russian to English translation)
- 2006: IBM Watson: Question-answering system



Eliza, the Rogerian Therapist

ELIZA is a computer program that emulates a Rogerian psychotherapist. Just type your questions and concerns and hit return. Eliza will answer you.

Talk to Eliza

```
> Hello, I am Eliza.
* how are you, Eliza?
> Why are you interested in whether or not I am Eliza?
* Whoa, I thought you were an emulation of a
psychotherapist.
> We were discussing you, not me.
```

Input:

When the original ELIZA first appeared in the 60's, some people actually mistook her for human. The illusion of intelligence works best, however, if you limit your conversation to talking about yourself and your life.

This javascript version of ELIZA was originally written by Michal Wallace and significantly enhanced by George Dunlop.

https://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm

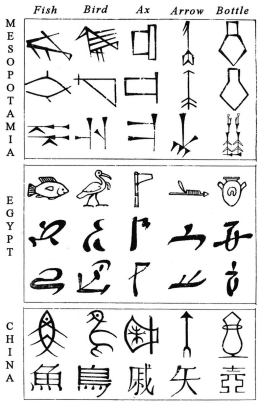
Natural Language

- Languages that humans speak (Examples: Mandarin Chinese, Spanish, English, Arabic, Hindi, Bengali)
- Language serves multiple functions: share knowledge, collaborate, .. Gossip
- Language communicates ideas
- Some ‘languages’ are specialised:
  - Language of a doctor
  - Language of an electrician
  - Language of a teacher teaching NLP
  - Language of a teacher teaching law
- Languages are different:
  - Native versus non-native speakers
  - Dialects versus so-called ‘standard’ language
  - Register (politeness versus argument)
- Language is an important factor for the success of the human race

Words

- Fundamental unit of a language
- Sounds produced to convey an idea
- Sounds were later written down
- Written scripts allowed ideas to be passed down to generations
  - Hieroglyphics
- A word conveys an idea
  - ‘apple’ (English) -→
  - ‘say-b’ (Hindi) ->
  - ‘úll’ (Irish) ->

- ‘apple’ (English)



https://en.wikipedia.org/wiki/Egyptian_hieroglyphs

Images from Wikimedia commons

.. The curious case of tea/cha

Derivatives of *te* [edit]

Language	Name	Language	Name	Language	Name	Language	Name	Language	Name
Afrikaans	*tee*	Armenian	թեյ [tʰɛj]	Basque	*tea*	Belarusian	*гарбата* (*harbatá*)(1)	Berber	*ⵜⴰⵢ, atay*
Catalan	*te*	Kashubian	*(h)arbata*(1)	Czech	*čaj* or *thé*(2)	Danish	*te*	Dutch	*thee*
English	*tea*	Esperanto	*teo*	Estonian	*tee*	Faroese	*te*	Finnish	*tee*
French	*thé*	West Frisian	*tee*	Galician	*té*	German	*Tee*	Greek	*τσάι tsái*
Hebrew	תה, *te*	Hungarian	*tea*	Icelandic	*te*	Indonesian	*teh*	Irish	*tae*
Italian	*tè*	Javanese	*tèh*	Kannada	ಟೀಸೊಪ್ಪು *tīsoppu*	Khmer	ថែ *tae*	scientific Latin	*thea*
Latvian	*tēja*	Leonese	*té*	Limburgish	*tèè*	Lithuanian	*arbata*(1)	Low Saxon	*Tee* [tʰeːə] or *Tei* [tʰæɪ]
Malay	*teh*	Malayalam	ചായ *tēyila*	Maltese	*tè*	Norwegian	*te*	Occitan	*tè*
Polish	*herbata*(1)	Scots	*tea* [tiː] – [teː]	Scottish Gaelic	*tì, teatha*	Sinhalese	*තේ tē*	Spanish	*té*
Sundanese	*entèh*	Swedish	*te*	Tamil	தேநீர் *tēnīr* (3)	Telugu	తేనీరు *tēnīru* (4)	Western Ukrainian	*herbata*(1)
Welsh	*te*								

Derivatives of *cha* [edit]

Language	Name	Language	Name	Language	Name	Language	Name	Language	Name
Chinese	茶 *Chá*	Assamese	চাহ *sah*	Bengali	চা *cha* (*sa* in Eastern regions)	Kapampangan	*cha*	Cebuano	*tsá*
English	*cha* or *char*	Gujarati	ચા *chā*	Japanese	茶, ちゃ *cha*(1)	Kannada	ಚಹಾ *chahā*	Khasi	*sha*
Punjabi	ਚਾਹ *chá*	Korean	차 *cha*(1)	Kurdish	*ça*	Lao	ຊາ /saː˧˩/	Marathi	चहा *chahā*
Oḍiā	ଚା' *cha'a*	Persian	چای *chā*	Portuguese	*chá*	Sindhi	چانهه *chahen*	Somali	*shaah*
Tagalog	*tsaá*	Thai	ชา /t͡ɕʰaː˧/	Tibetan	ཇ *ja*	Vietnamese	*trà* and *chè*(2)		

Words -> Word order

Languages typically choose word orders

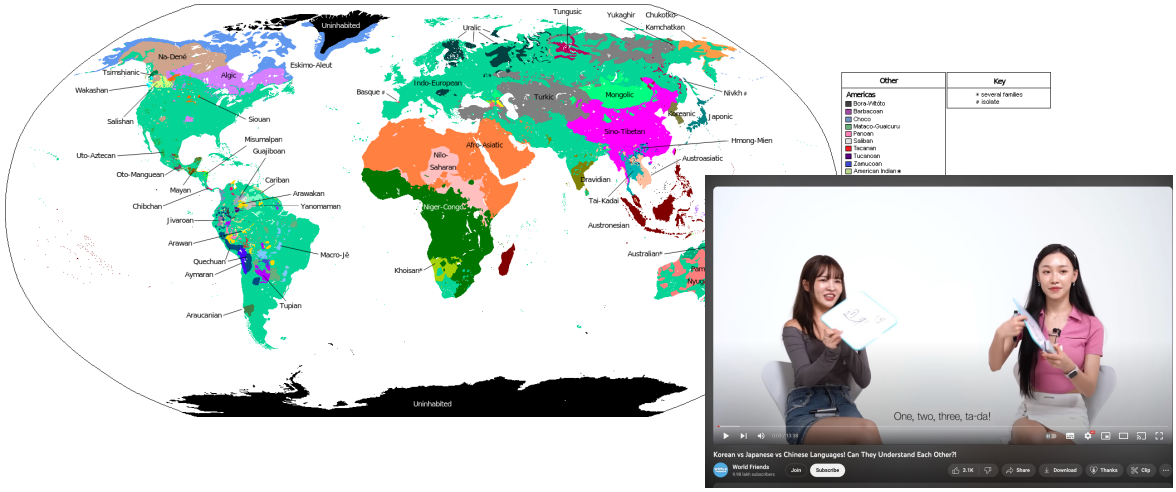
Word Order	Example	Languages
Subject-Verb-Object (SVO)	I went to school	English, Chinese, Italian
Subject-Object-Verb (SOV)	Mi shaalela gelo (I to-school went)	Japanese, Marathi, Tamil
Verb-Subject-Object (VSO)	Chuaigh mé ar scoil (Went I to-school)	Arabic, Irish
….		

Languages exhibit similarities in terms of word order, word choices, etc.

Tomlin R. S. (1986). *Basic word order : functional principles*. Croom Helm.

31

Language Families



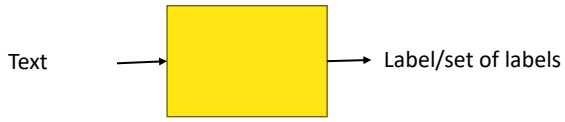
https://en.wikipedia.org/wiki/Language_family

https://www.youtube.com/watch?v=bYi2pu6gNY0

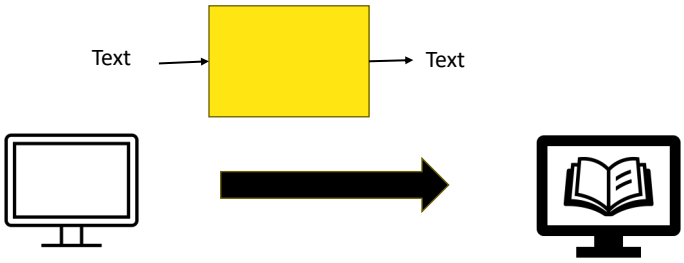
32

So.. What is NLP?

Computers 'understand' language: **Natural language understanding (NLU)**



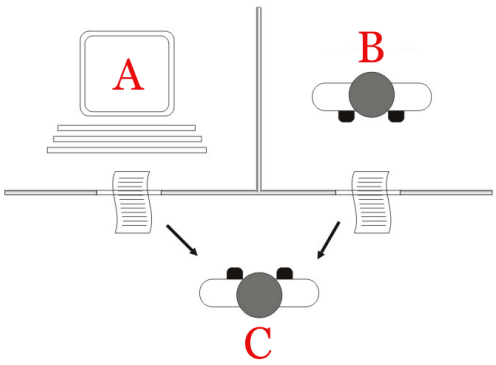
Computers 'generate' language: **Natural language generation (NLG)**



33

The Imitation Game

We now ask the question, 'What will happen when a machine takes the part of A in this game?'

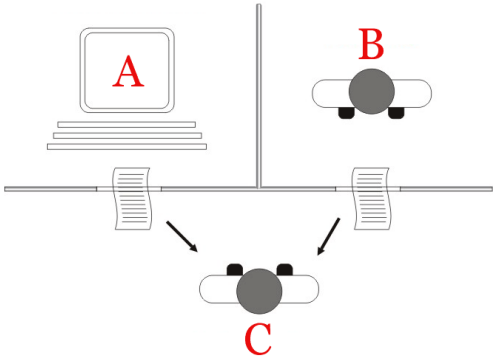


If A fools C into believing that it is human, is it the pinnacle of linguistic machine intelligence? → **NLP**

"**Computing Machinery and Intelligence**", Alan Turing, 1950.

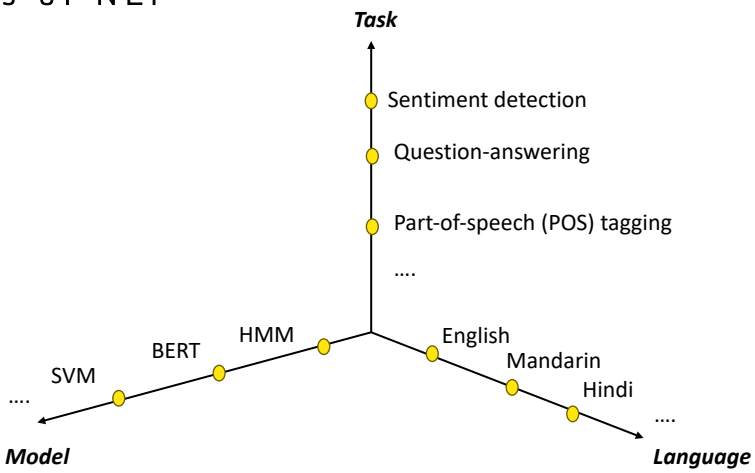
34

What does A need to do to fool C?



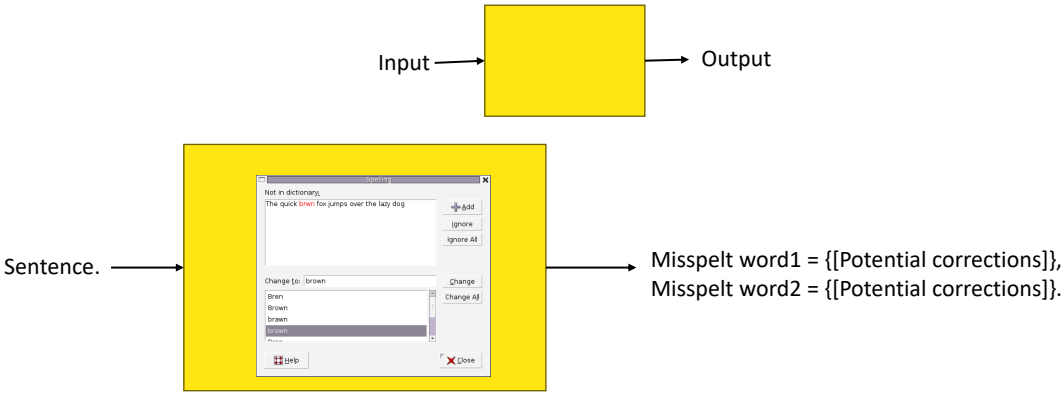
35

Dimensions of NLP



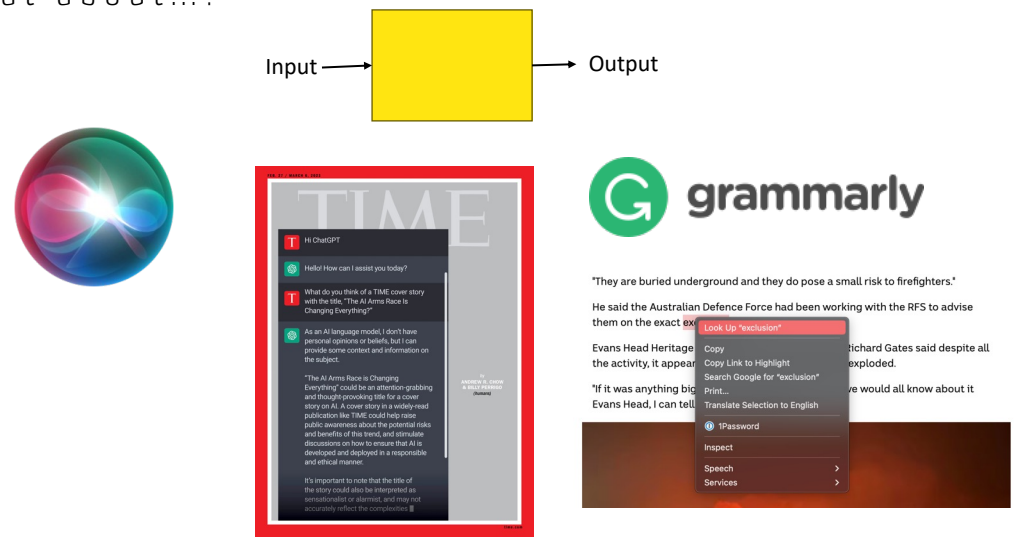
36

Black—box view



*Any colour that makes you think ‘opaque’

...what about...?



Examples of linguistic tasks

Part-of-speech (POS) tagging: Given a sentence, label individual words with POS tags. POS tags are defined by a tag set.

The boy goes to school. → DT NN VB PRP NN .

Chunking: Given a sentence, create chunks by grouping words together. Chunks are typically marked with brackets.

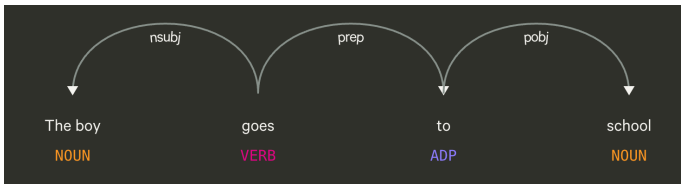
The blue book is on the table. → [The blue book] is on [the table].

39

Examples of linguistic tasks

Parsing: Given a sentence, create a tree rooted at the main verb of the sentence (as in the case of dependency parse tree) with edge types indicating roles of phrases.

The boy goes to school. →



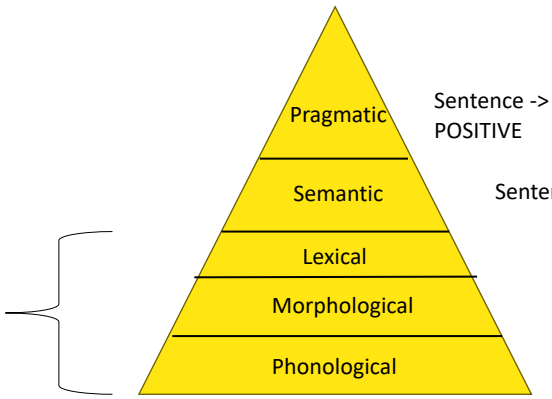
Created using spaCy with https://demos.explosion.ai/ wrapper.

Co-reference resolution: Given a document, link references (such as pronouns) to appropriate phrases in the preceding text.

The blue book is on the table. Can you pass it to me please? → [The blue book]:1 is on [the table]:2. Can you pass [it]:@1 to me please?

40

These tasks constitute the lower level of the ambiguity hierarchy



Sentence -> implied meaning ; e.g.: I love being ignored.-> NOT POSITIVE

Sentence -> meaning ; e.g.: I love this sandwich -> POSITIVE

Words -> meaning ; e.g.: man -> NOUN

Words -> morphemes ; e.g.: informing -> inform

Sounds -> words

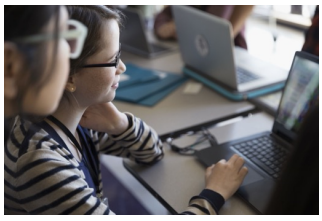
Demo: NLTK

Natural Language Toolkit (NLTK) (https://www.nltk.org/)

Text processing pipelines for multiple languages

Several corpora

Often used as a component in a larger system (say, recommender system)



Demo time!

Bird, Steven, Edward Loper and Ewan Klein (2009), *Natural Language Processing with Python*. O'Reilly Media Inc.

Downstream tasks

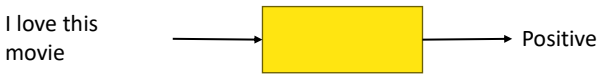
Downstream tasks: A term used to refer to application-oriented tasks

NLP witnessed a shift in the focus of tasks from linguistic tasks to application-oriented tasks

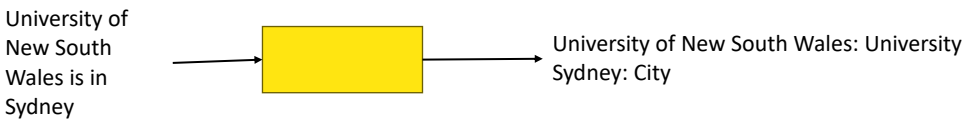
So what are some downstream tasks?

Examples of downstream tasks (simplified)

Sentiment Analysis: Given a text, detect the sentiment of the text.



Information Extraction: Given a text, extract entities along with their types. (e.g., named entity recognition)



Examples of downstream tasks (simplified)

Machine translation: Given a text in a source language, generate its translation in a target language.

I would like a salad please. → [ ] → Vorrei l'insalata, per favore.

Summarisation: Given a document, generate a summary that is shorter than the document but contains crucial information in the document.

[News article] → [ ] → Sydney Mardi Gras will be held in March 2025.

NLP tasks & research tracks

ACL Anthology: https://aclanthology.org/

- Computational Social Science and Cultural Analytics
- Dialogue and Interactive Systems
- Discourse and Pragmatics
- Efficient/Low-Resource Methods for NLP
- Ethics, Bias, and Fairness
- Generation
- Information Extraction
- Information Retrieval and Text Mining
- Interpretability and Analysis of Models for NLP
- Linguistic theories, Cognitive Modeling and Psycholinguistics
- Machine Learning for NLP
- Machine Translation
- Multilinguality and Language Diversity
- Multimodality and Language Grounding to Vision, Robotics and Beyond
- NLP Applications
- Phonology, Morphology and Word Segmentation
- Question Answering
- Resources and Evaluation
- Semantics: Lexical
- Semantics: Sentence-level Semantics, Textual Inference and Other areas
- Sentiment Analysis, Stylistic Analysis, and Argument Mining
- Speech recognition, text-to-speech and spoken language understanding
- Summarization
- Syntax: Tagging, Chunking and Parsing

While the focus of this course will be English, several considerations should get you to think: "how does this work in languages other than English that I speak?"

Data

Model

Linguistic phenomena

47

Part 2

Ambiguity

48

Announcements

This week:

No assessment or tutorials this week

Sample quiz for week 1 will be available on Moodle later today – but will not be evaluated. The caption in the quiz clearly says so.

Week 2 & thereafter:

Quizzes from week 2 onwards will be available from Friday evening onwards and will run until Wednesday of next week

Quiz for week 2 will include material from weeks 1 and 2

Consultation Hour: Thursdays 10am-11am | Weeks 2 to 10 | K-17, 217-B.

Start planning your project groups; talk to people ☺ We will ask for group details in week 4

49

Language conveys ideas

Sydney Opera House

But language is inherently ambiguous, and may convey multiple ideas.

Ambiguity: The quality of being open to more than one interpretation

50

Types of ambiguity

Lexical ambiguity: Words have multiple meanings

"Why should you never date a tennis player?" "Love means nothing to them"

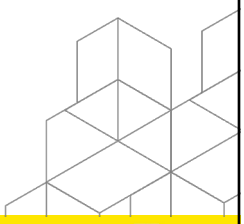
Syntactic ambiguity: Syntactic structures may be ambiguous

"I saw a boy with a telescope."

"I got a job offer from SEEK".

Pragmatic ambiguity: Sentences might imply a different meaning

"Being stranded in traffic is the best way to start the week."



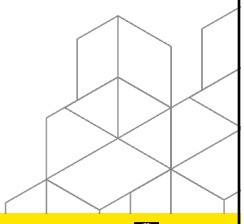
UNSW SYDNEY

51

Role of Ambiguity in NLP

NLP creates computational techniques to resolve ambiguity in language tasks

Ambiguity resolution in language is at the heart of NLP



UNSW SYDNEY

52

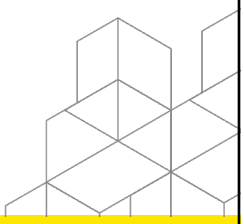
HELP THE COMPUTER TO FIND THE END OF A SENTENCE

A common task that a computer needs to do with text is to identify the words and the sentences.

Easy for humans

We would like to write a rule to split a document into sentences.

Can you define a rule to be given to a computer?



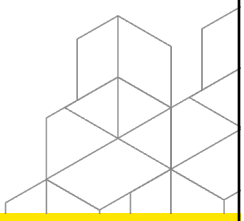
Let's make some rules

The Bank of New York ADR Index, which tracks depositary receipts traded on major U.S. stock exchanges, gained 1.3% to 183.32 points in recent session. The index lost 4.63 from the beginning of July. American Depositary Receipts are dollar-denominated securities that are traded in the U.S. but represent ownership of shares in a non-U.S. company.

Q. 1. What is the underlying ambiguity?

Q.2. Can you define IF THEN ELSE rules to segment the text into sentences?

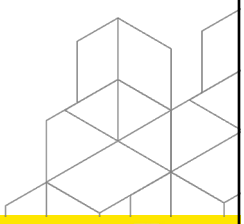
Example: "If a character is a '.', start a new sentence"



How could have ELIZA done this in Prolog?

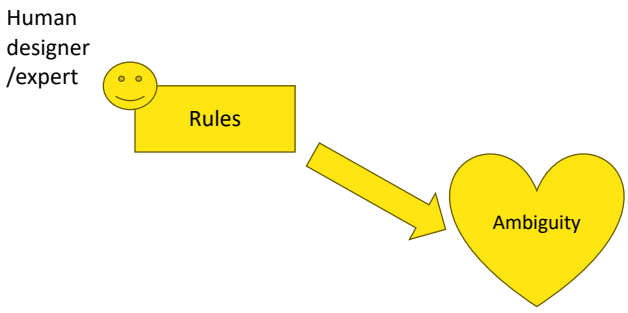
```
punctuation(',').
punctuation('.').
punctuation(';').
punctuation('?').
punctuation('!').

sentence_separator(X) :- punctuation(X).
sentence_separator(but).
sentence_separator(yet).

split_to_sentences(Words, Sentences) :-
  split_to_sentences_aux(Words, Sentences, [], []).

split_to_sentences_aux([], Sentences, Sentence_Acc, Sentences_Acc) :-
  (  Sentence_Acc = []
  -> reverse(Sentences_Acc, Sentences)
  ;  reverse(Sentence_Acc, Sentence),
     reverse([Sentence | Sentences_Acc], Sentences)
  ).

split_to_sentences_aux([X | Xs], Sentences, Sentence_Acc, Sentences_Acc) :-
  (  sentence_separator(X)
  -> reverse(Sentence_Acc, Sentence),
     split_to_sentences_aux(Xs, Sentences, [],               [Sentence | Sentences_Acc])
  ;  split_to_sentences_aux(Xs, Sentences, [X | Sentence_Acc], Sentences_Acc)
  ).
```

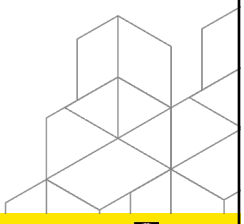
https://github.com/bartosz-witkowski/books/tree/master/natural-language-processsing-for-prolog-programmers



55

Rules were an early way to resolve ambiguity in NLP





56

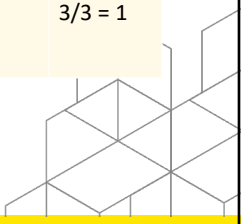
How well do rules work?

The Bank of New York ADR Index, which tracks depositary receipts traded on major U.S. stock exchanges, gained 1.3% to 183.32 points in recent session. The index lost 4.63 from the beginning of July. American Depositary Receipts are dollar-denominated securities that are traded in the U.S. but represent ownership of shares in a non-U.S. company.

Precision: % of times the rule works correctly
Recall: % of times the rule covers the positive cases

Rule	#retrieved sentences	#correct sentences	#expected sentences	Precision	Recall
If a character is a '.', start a new sentence	12	0	3	0/12 = 0	0/3 = 1
If a character is a '.', is followed by a blank space and a capital letter, start a new sentence	3	3	3	3/3 = 1	3/3 = 1

When will the second rule not work? it will not work for this text, for example.



UNSW SYDNEY

57

Rules are inadequate to capture linguistic phenomena

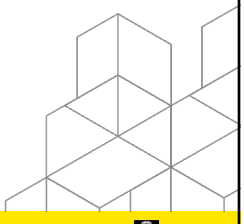
Early NLP systems relied on rules
- ELIZA: Written in LISP

Rules are time-consuming and expert-intensive

Rules may be subjective and fraught with biases of the 'expert'

Rules can 'overfit'
- High precision, low recall – typical for rule-based systems

Language is an evolving entity
- New words: *lit, kickass*
- New concepts: I *googled*
- New events: COVID



UNSW SYDNEY

58

Ambiguity can be resolved using observations

Abma is spoken by more than 8,000 people making it one of the largest indigenous languages of Vanuatu, a Pacific island nation that enjoys great linguistic diversity.

Can you translate the following sentences to Abma:

1. The teacher carries the water down.
2. The child keeps eating.
3. The child crawls here.
4. The teacher walks uphill.
5. The palm-tree keeps growing downwards.

NOTE: There is no separate word for 'the' or 'he' in these Abma sentences.

ABMA	ENGLISH
Mwamni sileng.	He drinks water.
Nutsu mwatbo mwamni sileng.	The child keeps drinking water.
Nutsu mwegau.	The child grows.
Nutsu mwatbo mwegalgal.	The child keeps crawling.
Mworob mwabma.	He runs here.
Mwerava Mabontare mwisib.	He pulls Mabontare down.
Mabontare mwisib.	Mabontare goes down.
Mweselkani tela mwesak.	He carries the axe up.
Mwelebte sileng mwabma.	He brings water.
Mabontare mworob mwesak.	Mabontare runs up.
Sileng mworob.	The water runs.

Now here are some more words in Abma:

ABMA	ENGLISH
sesesrakan	teacher
mwegani	eat
bwet	taro
muhural	walk
butsu-kul	palm-tree

https://ozclo.org.au/wp-content/uploads/2014/11/2010-first-round-problems.pdf

59

Solution

	ENGLISH	ABMA
1.	The teacher carries the water down.	***Sesesrakan mweselkani sileng mwisib.***
2.	The child keeps eating.	***Nutsu mwatbo mwegani.***
4.	The child crawls here.	***Nutsu mwegalgal mwabma.***
5.	The teacher walks uphill.	***Sesesrakan muhural mwesak.***
6.	The palm-tree keeps growing downwards.	***Butsukul mwatbo mwegau mwisib.***

60

Let's introspect...

How did you use the data?

What assumptions did you make?

Challenge: What sentences would not be possible to translate with certainty?

NOTE: There is no separate word for 'the' or 'he' in these Abma sentences.

ABMA	ENGLISH
Mwamni sileng.	He drinks water.
Nutsu mwatbo mwamni sileng.	The child keeps drinking water.
Nutsu mwegau.	The child grows.
Nutsu mwatbo mwegalgal.	The child keeps crawling.
Mworob mwabma.	He runs here.
Mwerava Mabontare mwisib.	He pulls Mabontare down.
Mabontare mwisib.	Mabontare goes down.
Mweselkani tela mwesak.	He carries the axe up.
Mwelebte sileng mwabma.	He brings water.
Mabontare mworob mwesak.	Mabontare runs up.
Sileng mworob.	The water runs.

Now here are some more words in Abma:

ABMA	ENGLISH
sesesrakan	teacher
mwegani	eat
bwet	taro
muhural	walk
butsu-kul	palm-tree

61

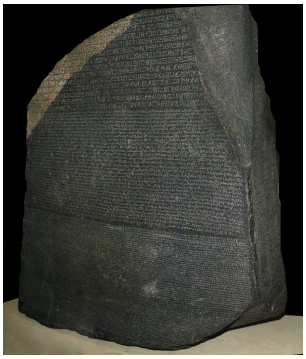
Observations come from data

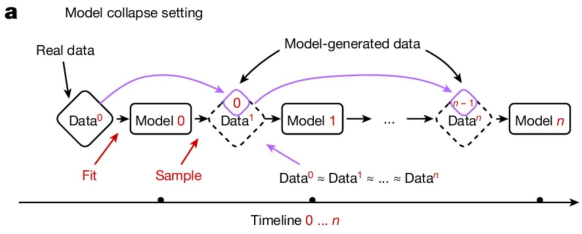
Rosetta stone

- 196 BC
- Three versions of a decree
- One of the earliest known 'corpus'

Textual Corpus/Dataset

- Digitisation made datasets available
- Internet exploded available datasets
- Website have been blocking crawlers that collect data





https://www.theguardian.com/technology/2023/aug/25/new-york-times-cnn-and-abc-block-openais-gptbot-web-crawler-from-scraping-content
https://www.nature.com/articles/s41586-024-07566-y

62

Terminology

Supervised:

Data annotated with expected output labels

"I love the movie" -> Positive ; "The movie sucks" -> Negative

Distant-supervised:

Data annotated with expected output labels, using heuristics. Labeling is 'distant'

"I love the movie #not' -> Sarcastic ; 'I love the movie' -> Not sarcastic

Unsupervised:

Data with no annotations w.r.t. the output labels

"I love the movie"

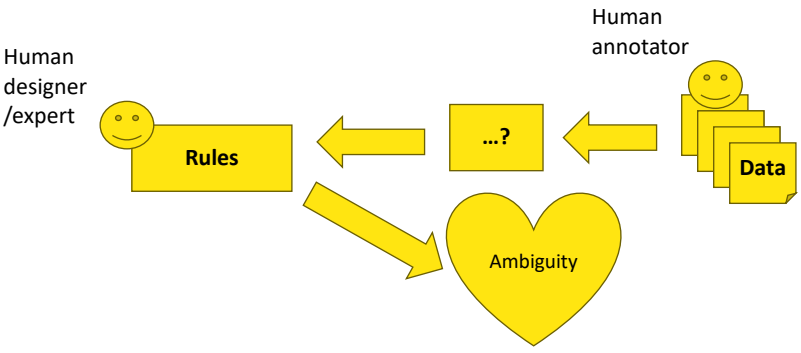
Self-supervised:

The data itself becomes supervision.

Introduced in deep learning-based NLP

"I love the movie" is mapped to: "I ____ the movie" -> "love"

Data can help learn 'rules'



... errm.. What does learning mean?

Enter: Probability

Pen-and-paper time!

Recap: Terminology

Naïve Bayes

Conditional independence assumption

N-grams

'Skip'-grams?

Argmax

Suggested reading: https://web.stanford.edu/~jurafsky/slp3/4.pdf

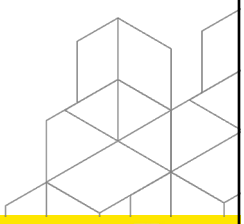
N-gram language models (Module 2)

Are words really random variables?

"Her face **fell** when she heard that she had been fired."
"The fruit **fell** from the tree."

"Avoid vulgar language please."
"Now that's real crude behavior!"

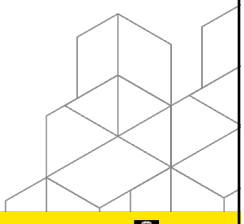
One-hot vector -> Word vector representations (Module 2)



67

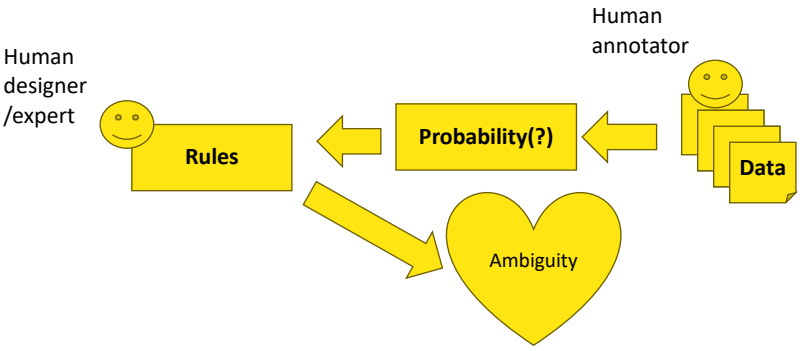
softmax

Neural networks generate vectors with logits
To convert them to probabilities, softmax function is used
It has desirable properties not seen in standard normalisation



68

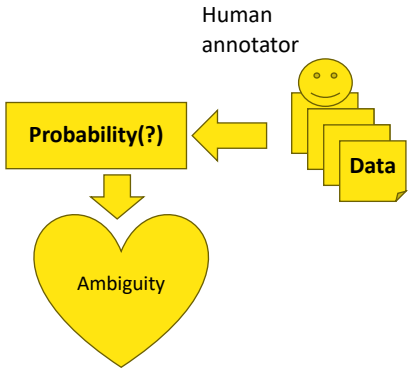
Data, Computation (Probability) and Ambiguity



… errm.. What does learning mean?

69

Data, Computation (Probability) and Ambiguity



70

Can 'data' be anything other than textual data? Lexicons

WordNet (https://wordnet.princeton.edu/)

LIWC (Lexical Inquiry and Word Count)

Domain-specific lexicons

- Medical ontologies
- Proprietary assets: Businesses often consider in-house lexicons to be valuable assets

May be represented either as graphs or lists

What is the 'easiest' way to use such lexicons?

WordNet

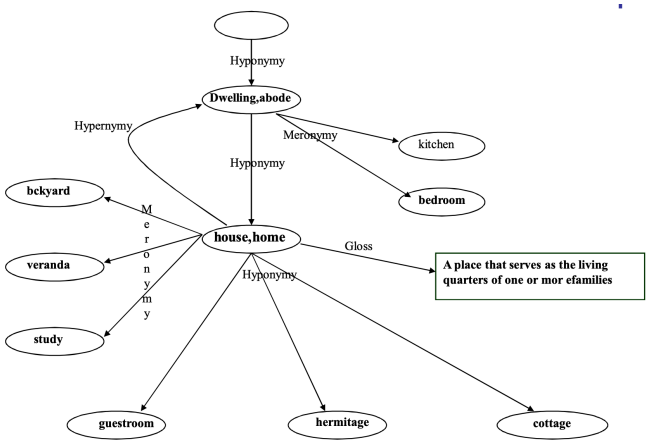
A lexical database

- Work on the English language WordNet began in 1985
- Inspired WordNets in other languages

http://wordnetweb.princeton.edu/perl/webwn

Synsets (synonym sets) connected to each other by relations

Semantic relations: Synonymy, Antonymy, Hypo/hypernymy, mero/holonymy.



Fellbaum, Christiane, ed. *WordNet: An electronic lexical database*. MIT press, 1998.
Example from: https://www.cse.iitb.ac.in/~pb/cs626-2013/cs626-lect24to26-wn-2013-9-17.pdf

spaCy

spaCy is an open-source software library for NLP

Multi-lingual support

“Industrial-strength NLP”

https://spacy.io/usage/spacy-101

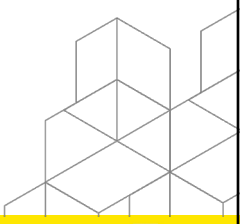
Supports deep learning pipelines

AllenNLP: Built on the top of spaCy and PyTorch (https://spacy.io/universe/project/allennlp)

Integration with LLMs available as well

https://spacy.io/usage/large-language-models

```
pip install spacy
python -m spacy download en_core_web_sm
```



Spacy Matcher

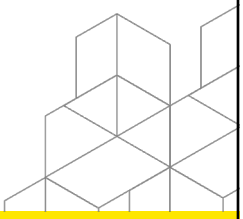
Regular expressions: Another artifact of rule-based NLP

SpaCy matcher allows simple pattern-matching using regular expressions

Several types of matcher primitives: https://spacy.io/api/matcher

PhraseMatcher may be useful for ontology-based matching

Demo time!



```
doc = self.nlp(text)
print("Sentences are:")
for sent in doc.sents:
        print(sent.text)
```

```
(base) z3539958@L-Q601F00993 week1 % python spacy2_sentencesplitter.py "The Bank of New York ADR Index, which tracks depos
itary receipts traded on major U.S. stock exchanges, gained 1.3% to 183.32 points in recent session. The index lost 4.63 f
rom the beginning of July. American Depositary Receipts are dollar-denominated securities that are traded in the U.S. but
represent ownership of shares in a non-U.S. company."

Sentences are:
The Bank of New York ADR Index, which tracks depositary receipts traded on major U.S. stock exchanges, gained 1.3% to 183.
32 points in recent session.
The index lost 4.63 from the beginning of July.
American Depositary Receipts are dollar-denominated securities that are traded in the U.S. but represent ownership of shar
es in a non-U.S. company.
```

https://spacy.io/usage/rule-based-matching

spaCy: Recap

Black-box NLP library for several NLP tasks

SpaCy matcher!

Part 3

The three generations

77

Computational techniques in NLP have witnessed paradigm shifts

Generation 1: Rule-based NLP

- Human-written rules
- Highly explainable
- Any adaptation needs human intervention

Generation 2: Statistical NLP

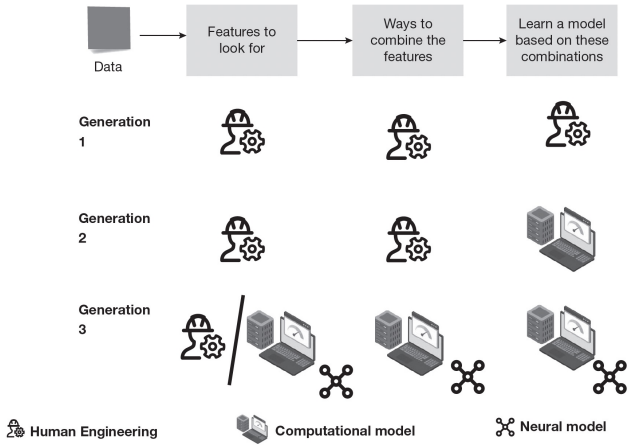
- Human-written features
- Probability and argmax at the heart of computation

Generation 3: Neural NLP

- Overt task-specific features not essential
- Foundation models learn from self-supervised data; Task-specific data may be used to adapt models

78

How do the three generations differ?



79

Characteristics of the three generations

Rule-based generation
- Based on rules designed for the task.
- Facilitated by human experts.
- Concerns around 'recall'/ generalisability.

Statistical generation
- Effective due to task-specific, human-designed signals.
- Facilitated by advances in statistical machine learning models.
- Concerns around optimality of features.

Neural generation
- Independent of task-specific, human-designed signals.
- Facilitated by large volumes of data and high computational power.
- Concerns around interpretability.

Also characterized by lexicons:
curated, structured resources
e.g. WordNet, SentiWordNet

What other examples?

80

Neural NLP has been influenced by Transformer

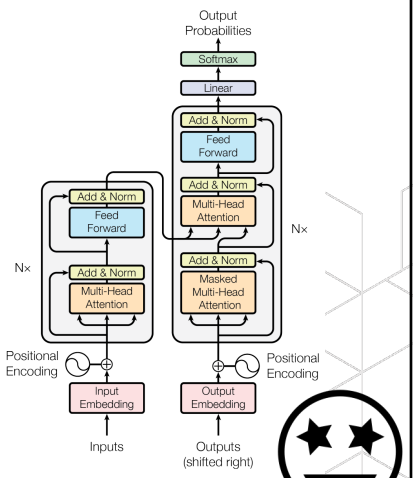
Transformer is an architecture originally introduced for NLP; adapted to other areas of AI

Models trained using Transformer architecture

Input Sequence → Transformer → Output Sequence

Encoder-only models: **BERT**

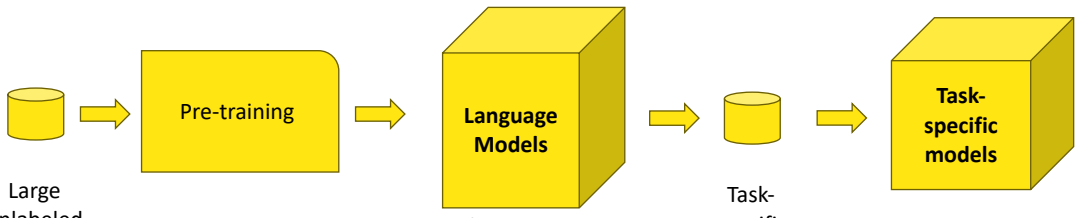
Decoder-only models: **GPT** (Week 4)



Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

81

Neural NLP Pipeline



Large unlabeled datasets

(e.g. Web-scale corpora)

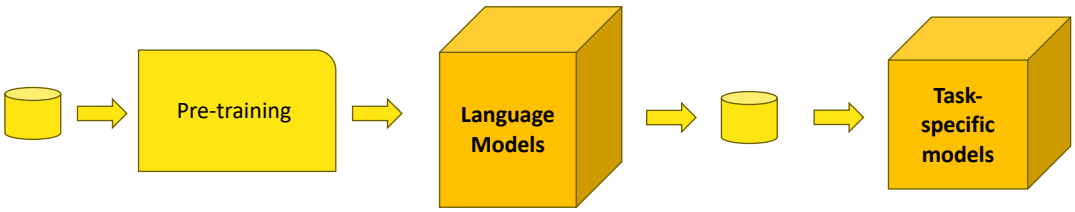
a.k.a.
Large language models (LLM)
Foundation models
Pre-trained Language models (PTLM)

Task-specific labeled dataset

(e.g. sentiment-labeled corpora)

82

Neural NLP Pipeline



Both PTLMs and TLMs can be used for inference.
How are language models 'pre-trained'?
How are they adapted to specific tasks?
.. And related considerations.

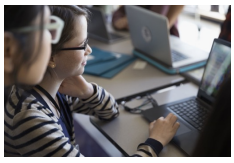
All in Week 4

UNSW SYDNEY

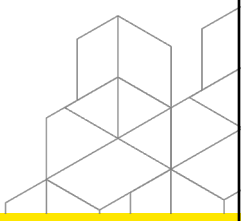
83

Another black-box library: HuggingFace

HuggingFace (https://huggingface.co/)
HuggingFace makes pre-trained and/or fine-tuned models available
https://huggingface.co/meta-llama/Llama-2-7b-chat-hf



Demo time!



UNSW SYDNEY

84

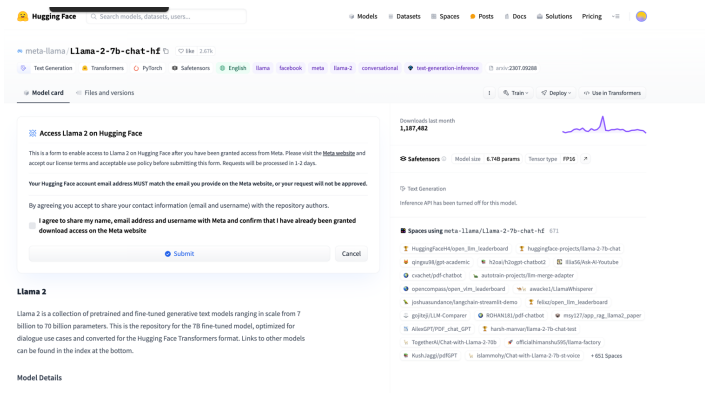
```
>>> res = rr("University of New South Wales is in Sydney")
>>> res
[{'entity': 'B-ORG', 'score': 0.99777824, 'index': 1, 'word': 'University', 'start': 0, 'end': 10}, {'entity': 'I-ORG', 's
core': 0.9964133, 'index': 2, 'word': 'of', 'start': 11, 'end': 13}, {'entity': 'I-ORG', 'score': 0.99850214, 'index': 3,
'word': 'New', 'start': 14, 'end': 17}, {'entity': 'I-ORG', 'score': 0.9936692, 'index': 4, 'word': 'South', 'start': 18,
'end': 23}, {'entity': 'I-ORG', 'score': 0.9954946, 'index': 5, 'word': 'Wales', 'start': 24, 'end': 29}, {'entity': 'B-LO
C', 'score': 0.99470913, 'index': 8, 'word': 'Sydney', 'start': 36, 'end': 42}]
```

```
>>> document = "Natural language processing (NLP) is a branch of artificial intelligence that deals with computational app
roaches used to process text. Human language (i.e., natural language) is inherently ambiguous. Ambiguity resolution using
computational techniques is at the heart of NLP. As a result, the advancements in NLP can be visualized as three generatio
ns: rule-based, statistical and neural. The course introduces the three generations of NLP through the philosophy of ambig
uity resolution being the core task of NLP. The content covers different NLP sub-problems (such as POS tagging, sentiment
classification, named entity recognition, machine translation and summarisation), and typical approaches in the three gene
rations to tackle these sub-problems. With recent advancements in large language models, there has been a renewed interest
 in NLP from industry and research alike. However, NLP precedes large language models. The exposition of NLP centered arou
nd ambiguity resolution helps to develop an understanding of the past and the present of NLP."
>>> print(summarizer(document, max_length=130, min_length=30, do_sample=False))
[{'summary_text': 'Natural language processing is a branch of artificial intelligence that deals with computational approa
ches used to process text. Ambiguity resolution using computational techniques is at the heart of NLP. The advancements in
 NLP can be visualized as three generations: rule-based, statistical and neural.'}]
```


HuggingFace Hub

Community-contributed repository of models and datasets

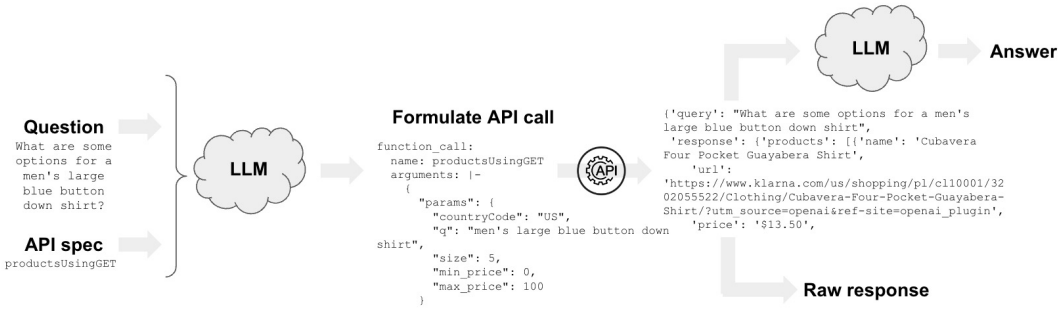
Easy to integrate with quick deployment libraries like Gradio



APIs for LLMs

Commercial LLMs offer access in the form of API calls.

Often in the form of an API call.



https://python.langchain.com/v0.1/docs/use_cases/apis/

87

Example: OpenAI API

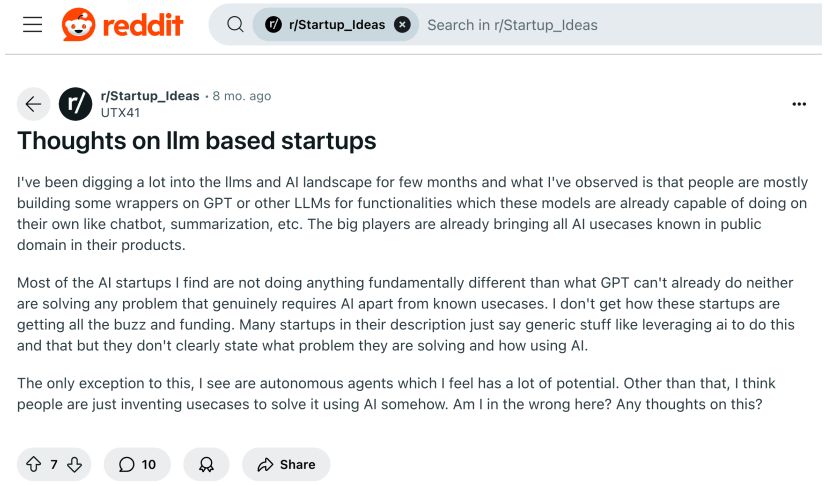
Example: OpenAI API call

```
from openai import OpenAI
client = OpenAI()

response = client.chat.completions.create(
  model="gpt-4o",
  messages=[
    {
      "role": "system",
      "content": "You will be provided with statements, and your task is to convert them to standard English."
    },
    {
      "role": "user",
      "content": "She no went to the market."
    }
  ],
  temperature=1,
  max_tokens=256,
  top_p=1
)
```

https://python.langchain.com/v0.1/docs/use_cases/apis/

88

... but....



UNSW SYDNEY

89

UNSW SYDNEY | Australia's Global University

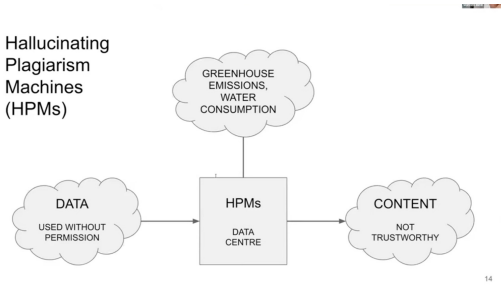
Part 4

Key Considerations

90

Is NLP a 'solved' problem?

Availability of web-based demo may give the impression that NLP is a 'solved' problem



Several considerations assume importance.

I hope this course will encourage you to investigate some of the above aspects.

Image from Steven Bird's ALTA 2024 keynote: https://www.youtube.com/watch?v=YcU_tZIesKI

91

Hallucination

In the case of generative models

Factually incorrect

Semantically inconsistent

.. Other definitions of hallucination are possible.

ALARMING

Prompt: *Engineering effort to build Eiffel tower*

AI-generated text: ...Designed by Gustave Eiffel, it was inaugurated in 1889 to celebrate the 100th anniversary of the European Civil War...

Fact: Eiffel tower was built to celebrate the 100th anniversary of the French Revolution.

Which LLM might work well for you?

Chatbot Arena: https://lmarena.ai/

•Rawte, Vipula, et al. "The Troubling Emergence of Hallucination in Large Language Models–An Extensive Definition, Quantification, and Prescriptive Remediations." EMNLP 2023.

92

Ethical Considerations

Bias

Sentence completion can produce sexist/racist/religiously offensive output

Privacy

Transparency

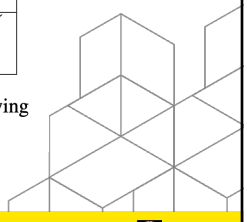
Explanation of output; citing one's source

93

Gender Bias in modern NLP

Task	Example of Representation Bias in the Context of Gender	D	S	R	U
Machine Translation	Translating "He is a nurse. She is a doctor." to Hungarian and back to English results in "She is a nurse. He is a doctor." (Douglas, 2017)		✓	✓	
Caption Generation	An image captioning model incorrectly predicts the agent to be male because there is a computer nearby (Burns et al., 2018).		✓	✓	
Speech Recognition	Automatic speech detection works better with male voices than female voices (Tatman, 2017).			✓	✓
Sentiment Analysis	Sentiment Analysis Systems rank sentences containing female noun phrases to be indicative of anger more often than sentences containing male noun phrases (Park et al., 2018).		✓		
Language Model	"He is doctor" has a higher conditional likelihood than "She is doctor" (Lu et al., 2018).		✓	✓	✓
Word Embedding	Analogies such as "man : woman :: computer programmer : homemaker" are automatically generated by models trained on biased word embeddings (Bolukbasi et al., 2016).	✓	✓	✓	✓

Table 1: Following the talk by Crawford (2017), we categorize representation bias in NLP tasks into the following four categories: (D)enigration, (S)tereotyping, (R)ecognition, (U)nder-representation.



•Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

94

Dialectal Bias in modern NLP

NLP Task	Paper	Impact
Language classification	[Blodgett et al. 2016]	Language detection shows lower performance for African-American English.
Sentiment classification	[Okpala et al. 2022]	Text in African-American English may be predicted higher as hate speech.
Natural Language Understanding	[Ziems et al. 2022]	Popular models perform worse on GLUE tasks for African-American English text.
Summarisation	[Keswani and Celis 2021]	Generated multi-document summaries may be biased towards majority dialect.
Machine translation	[Kantharuban et al. 2023]	Significant drop in MT from and to dialects of Portuguese/Bengali/etc. to and from English.
Parsing	[Scannell 2020]	Lower performance of parsers on Mancks Gaellic as compared to Irish/Scottish Gaellic.

Table 1. Examples of adverse impact on NLP task performance due to dialectic variations.

Joshi, Aditya, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. "Natural Language Processing for Dialects of a Language: A Survey." *ACM Computing Surveys 2025.*

95

Privacy in modern NLP

Samsung bans Generative AI use

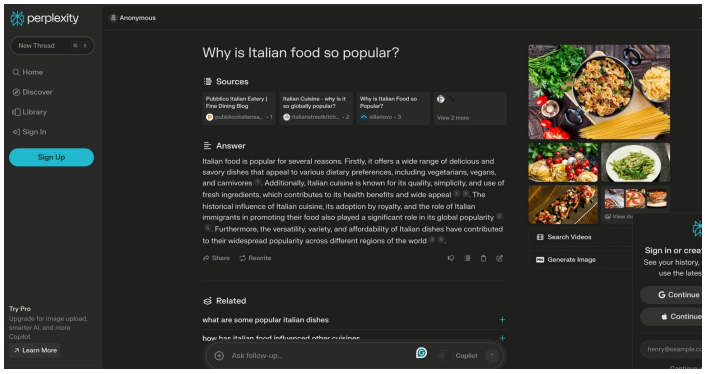
Data leakage into training and test set

https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak

96

Transparency

https://www.perplexity.ai/



https://techcrunch.com/2024/01/04/ai-powered-search-engine-perplexity-ai-now-valued-at-520m-raises-70m/

97

Summary

	Key Idea	Demos
NLP Today & Yesterday	NLP has fascinated computer science for a long time.	Fundamental NLP tasks using NLTK
Ambiguity	Interaction between data, probability and ambiguity resolution	Text matching using spaCy
The three generations	Three-generational view of NLP	Open-source NLP models using HuggingFace
Considerations	NLP is far from solved. Hallucination, privacy, biases, etc.	Emerging NLP tools: Perplexity.ai

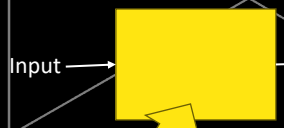
98

Coming up next…

Grammar, Probabilistic language models
Word vectors
Sequential networks

99

UNSW SYDNEY | Australia's Global University



In the next module, we will begin to unpack this black-box.

100