"AI is the new electricity. Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform."

Andrew Ng

All images from Wikimedia commons, unless specified.

1

Natural Language Processing (NLP)

COMP6713 – 2025 Term 1

Convener

Dr. Aditya Joshi

aditya.joshi@unsw.edu.au

Week 10

Applications & Frontiers

Schedule

2025 Term 1

2

Announcements

Reminder: myExperience survey

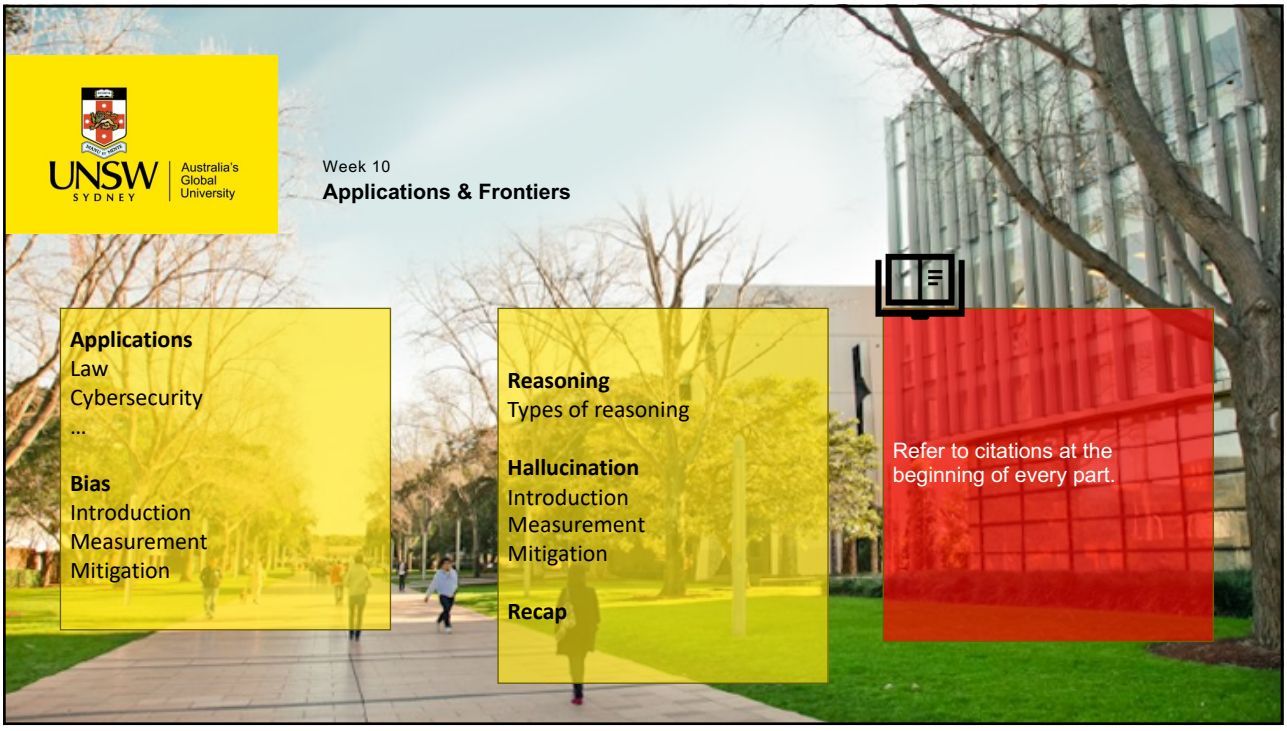
- Current response rate: 24%

Final exam

We will discuss this on Friday

No quiz in week 10

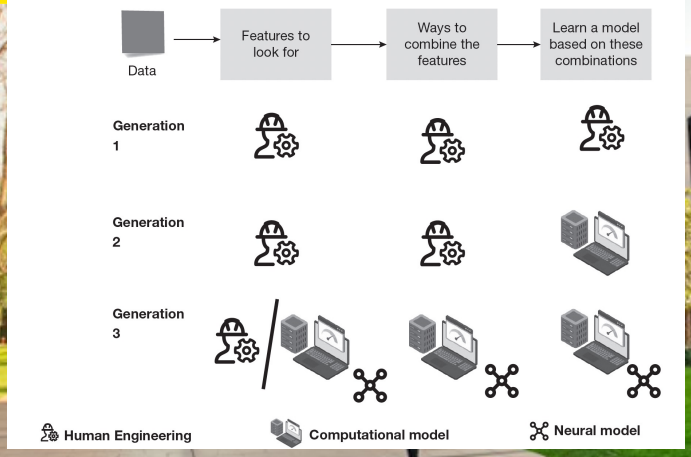
3



4

UNSW SYDNEY | Australia's Global University

Topics this week



- Hallucination
- Retrieval-augmented generation.....
- Bias measurement
- Bias mitigation

5

UNSW SYDNEY | Australia's Global University

Part 0

Applications

6

NLP & Law

	Task	Examples / Description
1	**Machine Summarization**	Abstractive/Extractive Summaries of Legal Documents
2	**Pre-Processing**	Annotation, Anonymization, Translation
3	**Classification**	Outcome Prediction, Legal Area Classification, Topic Modeling
4	**Information Retrieval**	Legal Question Answering, Document Similarity, Document Retrieval
5	**Information Extraction**	Labeling, Text Extraction, Event Extraction
6	**Text Generation**	Automated Drafting of Legal Documents
7	**Resources**	Taxonomies, Ontologies, Datasets, Code Libraries

Table 1. A Taxonomy of Engineering Tasks in Legal NLP

Katz, Daniel Martin, et al. "Natural language processing in the legal domain." *arXiv preprint arXiv:2302.12039* (2023).

7

NLP & Science

Physics

Physic Property
Q: What's the Bandgap of [BaCu4S3]?
GPT-4: As of my knowledge cutoff in September 2021, the exact bandgap value for BaCu4S3 isn't readily available in standard references.
DARWIN: The bandgap of the material is 1.79
Ground truth: BaCu4S3 -> 1.79

Chemistry

Molecular Design
Q: Design a MOF with followed desired property: Gravimetric heat capacity around 11 at 300K. Answer with MOF structure and topology
GPT-4: Remember that as of my knowledge cutoff in September 2021, the MOF design for specific heat capacity is still a challenging task and is an active area of research. Please check the most recent scientific literature for updates and progress in this field.
DARWIN: linker [O-]C(=O)c1ccc(cc1c1cc(ccc1C(=O)[O-])C(=O)[O-])C(=O)[O-], nodes [Zn], topology nor

Material Science

Name Entity Recognition & Relationship Extraction
Q: Find and extract the solar cell structure mentioned in the paragraph...
GPT-4: HTL_deposition_procedure: Spiro-OMeTAD (100 mg/mL chlorobenzene) is spin-coated on the substrates with additives Li(CF3SO2)2N and TBP added to the solution to improve conductivity. ; HTL_depostion_solvent: chlorobenzene
DARWIN: HTL_deposition_procedure: Spiro-OMeTAD | TBP ; HTL_deposition_solvent: chlorobenzene; 100 mg/mL
Ground truth: HTL_deposition_procedure: Spiro-OMeTAD | TBP ; HTL_deposition_solvent: chlorobenzene; 100 mg/mL

Others

Multi-choice Question
What happens when mammals raise their hair with tiny muscles in the skin?", "A: "bruising", "B": "bleeding", "C: "balding", "D": "goosebumps"
GPT-4: "D: "goosebumps"
DARWIN: "D: "goosebumps"
Ground truth: "D: "goosebumps"

Xie, Tong, et al. "Darwin series: Domain specific large language models for natural science." arXiv preprint arXiv:2308.13565 (2023).

8

NLP & Cybersecurity

Table 1: Classified publications concerning the *defensive* applications of LLMs.

Paper	Year	NIST Framework	Application	Model(s)
[45]	2023	Identify	LLMs enhance cybersecurity policies.	ChatGPT
[33]	2023	Protect	Using LLMs for secure code development without compromising functionality.	SVEN (GPT-2), (CodeGen) LM
[83]	2023	Protect	LLMs solve Capture The Flag challenges to enhance employees' awareness and knowledge.	code-cushman-001, code-davinci-001,code-davinci-002, 1-jumbo, j1-large, polycoder, gpt2-csrc
[64]	2023	Protect	LLMs investigate software vulnerabilities.	GPT-3.5 Turbo, Gemini, Microsoft Bing
[10]	2023	Protect	LLMs investigate software vulnerabilities.	GPT-3.5 Turbo
[95]	2023	Protect	Generating honeywords using LLMs.	GPT-3
[20]	2018	Protect	Chatbots assist security experts in identifying open ports.	Rule-based
[87]	2023	Protect	LLM-based URL categorization for website classification.	eXpose (Conv), BERTiny, URLTran (BERT) T5 Large, GPT3 Babbage
[75]	2023	Protect	LLMs investigate code vulnerabilities.	GPT-3

Table 2: Classified publications concerning the *adversarial* applications of LLMs.

Paper	Year	MITRE Tactic(s)	Application	Model(s)
[11]	2023	Execution	Generating code to perform actions that could be malicious	GPT-3
[42]	2022	Initial Access	Generate phishing emails to bypass spam filters	GPT-2, GPT-3, RoBERTa
[7]	2022	Execution - Command & Control	Use of LLMs as plug-ins to act as a proxy	GPT-4
[72]	2023	Initial Access - Collection	Generate Phishing Website via ChatGBT	GPT-3.5 Turbo
[8]	2023	Execution	Code generation and DLL injection	GPT-3
[32]	2023	Initial Access - Reconnaissance	Collecting victim data to develop an attack email	GPT-3.5, GPT-4
[62]	2023	Initial Access - Execution - Defense Evasion	Crafting malicious scripts	GPT-3.5 Turbo, GPT-4, text-davinci-003
[40]	2018	Initial Access	Spear Phishing link	AWD-LSTM
[12]	2023	Defense Evasion	Code obfuscation, file format modification	GPT-3.5
[68]	2023	Initial Access - Credential Access	Password guessing using LLMs	GPT-2
[74]	2023	Initial Access - Reconnaissance	Impersonation for phishing aims	GPT-3.5 Turbo
[48]	2022	Initial Access	Generating content for misinformation	GPT-2

Motlagh, Farzad Nourmohammadzadeh, et al. "Large Language Models in Cybersecurity: State-of-the-Art." *arXiv preprint arXiv:2402.00891* (2024).

9

NLP & Mobility

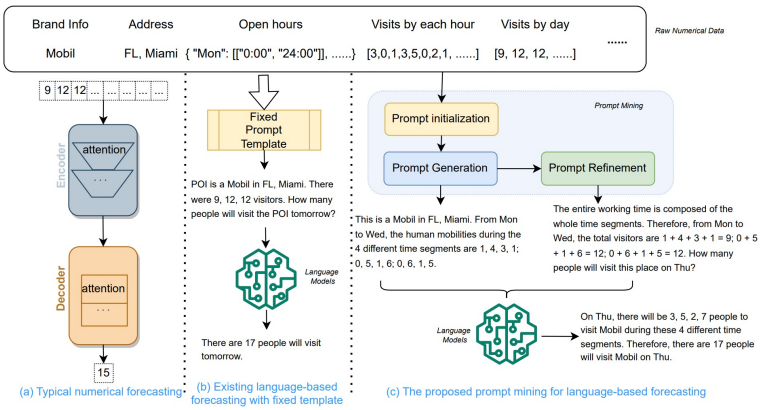


Figure 1: The conceptual comparison of: (a) numerical forecasting, (b) language-based forecasting with a fixed template, (c) our proposed prompt mining process.

Xue, Hao, et al. "Prompt Mining for Language-based Human Mobility Forecasting." *arXiv preprint arXiv:2403.03544* (2024).

10

NLP & Public Health

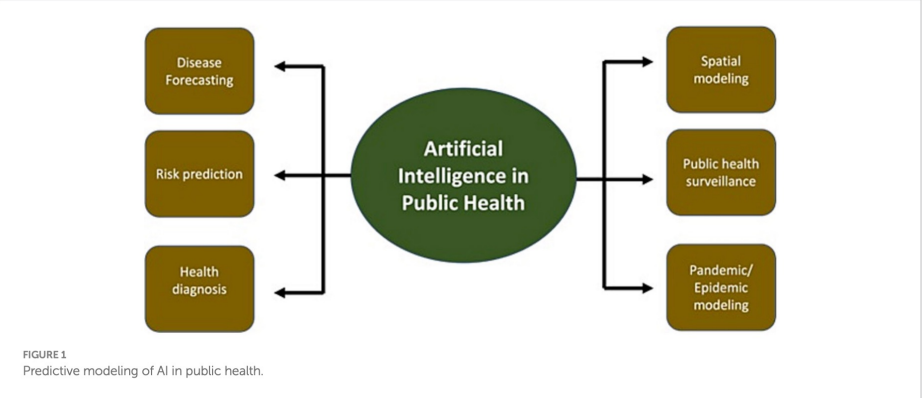


FIGURE 1
Predictive modeling of AI in public health.

Olawade DB, Wada OJ, David-Olawade AC, Kunonga E, Abaire O and Ling J (2023) Using artificial intelligence to improve public health: a narrative review. Front. Public Health 11:1196397. doi: 10.3389/fpubh.2023.1196397

11

UNSW SYDNEY | Australia's Global University

Part 1

Hallucination

Ji, Ziwei, et al. "Survey of hallucination in natural language generation." *ACM Computing Surveys* 55.12 (2023): 1-38.

Guo, Yue, Yi Yang, and Ahmed Abbasi. "Auto-debias: Debiasing masked language models with automated biased prompts." ACL. 2022.

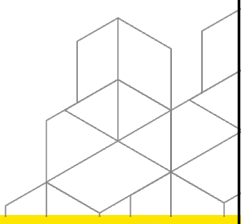
12

Hallucination & siblings

Degeneration: bland, incoherent, or gets stuck in repetitive loops

Hallucination: Undesirable generation that results in an output that is either nonsensical or unfaithful to the provided source input

Hallucination is a key challenge to NLP today.



UNSW SYDNEY

13

Types of hallucination

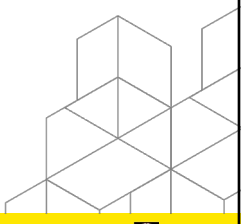
Document: The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID- 19 is unlikely to be ready this year, although clinical trials have already started.

"The first Ebola vaccine was approved in 2021" **Intrinsic Hallucination**

"Clinical trials for COVID-19 vaccine have started." **Not Hallucination**

"China has already started clinical trials of the COVID-19 vaccine." **Extrinsic Hallucination**

Demo time!



UNSW SYDNEY

14

Question-Answering

-Input: Question, Output: Answer

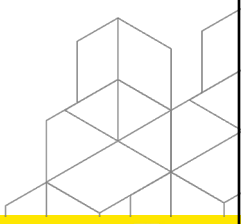
- Document question-answering

    Extractive

    Can use encoder models using the sequence tagging (NER) setup

-Abstractive question-answering

    Template-based if using rule-based approaches



15

Measuring hallucination

Human evaluation

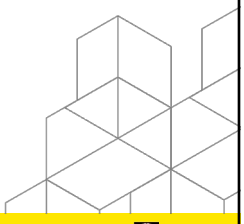
Automated metrics: ROUGE, etc. against a reference evaluation

Tuple-based evaluation

    Compare output of LLMs with tuples from a knowledge base

NLI-based evaluation

    Compare output of LLMs with sentences from a dataset



16

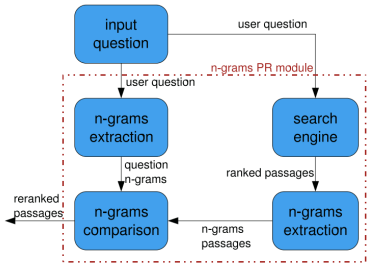
Mitigating hallucination

LLMs become a black-box when used for zero-shot/in-context learning

What is at the other end of the spectrum?

Retrieval!

Inspiration: Retrieval-based question-answering



Retrieval-augmented generation!

https://link.springer.com/chapter/10.1007/11551874_57

Retrieval-augmented generation (RAG)

Retrieve a set of related documents

Use the documents to generate the response

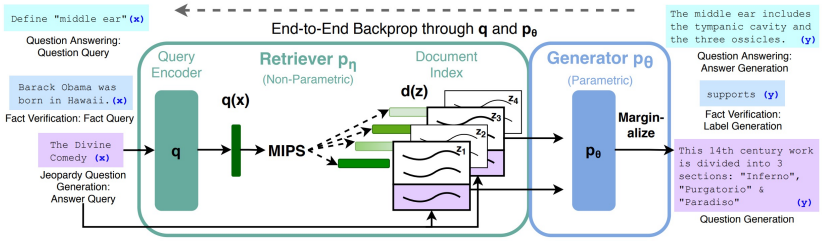


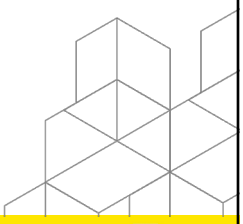
Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

Mathematical formulation

Let x be the question; y be the response. z be the top k documents relevant to x

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

$$p_\eta(z|x) \propto \exp\left(\mathbf{d}(z)^\top \mathbf{q}(x)\right) \qquad \mathbf{d}(z) = \text{BERT}_d(z), \quad \mathbf{q}(x) = \text{BERT}_q(x)$$



UNSW SYDNEY

19

Demo time!

UNSW SYDNEY

20

UNSW SYDNEY | Australia's Global University

Part 2

Bias

Czarnowska, Paula, Yogarshi Vyas, and Kashif Shah. "Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics." *Transactions of the Association for Computational Linguistics* 9 (2021): 1249-1267.

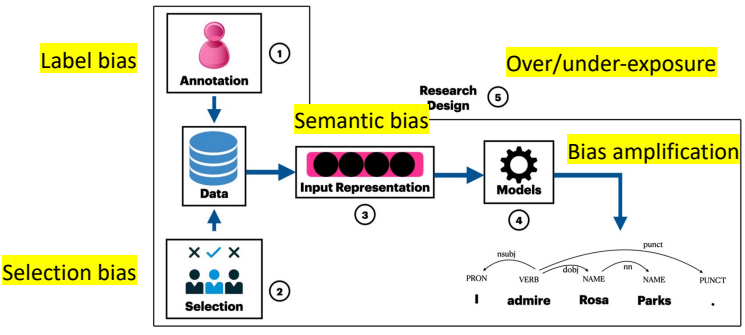
William Held, Caleb Ziems, and Diyi Yang. 2023. TADA : Task Agnostic Dialect Adapters for English. In Findings of the Association for Computational Linguistics: ACL 2023, pages 813–824, Toronto, Canada. Association for Computational Linguistics.

Guo, Yue, Yi Yang, and Ahmed Abbasi. "Auto-debias: Debiasing masked language models with automated biased prompts." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

21

Bias

Dictionary: Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.
Mismatch of ideal and actual distributions of labels and user attributes in training and application of a system



https://compass.onlinelibrary.wiley.com/doi/pdfdirect/10.1111/lnc3.12432

22

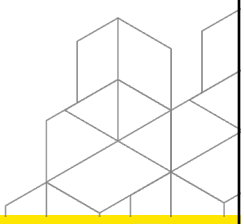
Measuring bias: Why

Often associated with subgroups of users and the performance of the models w.r.t. these groups

The protected attributes under the Fair Work Act are:

race, colour, sex, sexual orientation, age, physical or mental disability, marital status, family or carer's responsibilities, pregnancy, religion, political opinion, national extraction, social origin, breastfeeding, gender identity, intersex status, experiencing family and domestic violence.

https://www.fairwork.gov.au/employment-conditions/protections-at-work/protection-from-discrimination-at-work



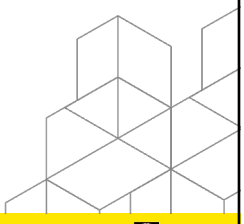
.. and how

Measuring bias in NLP models is key for better understanding and addressing unfairness

Quantify the differences in a model's behavior across a range of social groups

Two components:

- Scoring function for a subset of samples
- Comparison function that compares scores of different subsets to compute a fairness score



Group fairness

Group fairness requires parity of some statistical measure across a small set of protected groups

Demographic parity: equal positive classification rate across different groups

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 more commonly 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.

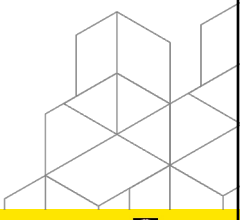
25

Counterfactual fairness

Counterfactual fairness: Requirement of invariance

One from the actual world and

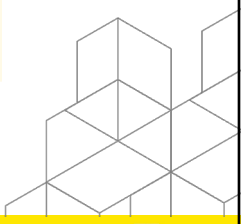
Others from counterfactual worlds in which the individual belongs to a different protected group;



26

Bias measurement metrics

	Group Fairness	Counterfactual Fairness
Pairwise Comparison	Data from different groups is compared Group A versus Group B	Data is perturbed to represent different groups Group A versus Group B
Background Comparison	Data from different groups is compared Group A,B,.. versus background	Data is perturbed to represent different groups Group A,B,.. versus background



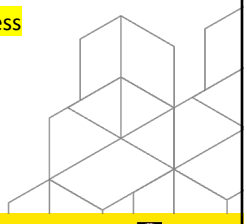
27

Pairwise comparison metric

Pairwise Comparison Metric (PCM) quantifies how distant, on average, the scores for two different, randomly selected groups are

$$\frac{1}{N} \sum_{t_i, t_j \in \binom{T}{2}} d\left(\phi(S^{t_i}), \phi(S^{t_j})\right)$$ Group Fairness

$$\frac{1}{|S'|\ N} \sum_{S'_j \in S'} \sum_{t_i, t_k \in \binom{T}{2}} d\left(\phi(S'^{t_i}_j), \phi(S'^{t_k}_j)\right)$$ Counterfactual Fairness



28

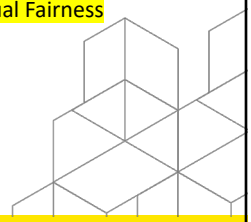
Background Comparison Metrics

a comparison between the score for a protected group and the score of its background.

Background could be the overall set of samples

$$\frac{1}{N}\sum_{t_i \in T} d\left(\phi(\beta^{t_i,S}), \phi(S^{t_i})\right)$$ Group Fairness

$$\frac{1}{|S'|\,N}\sum_{S'_j \in S'}\sum_{t_i \in T} d\left(\phi(\beta^{t_i,S'_j}), \phi(S'^{t_i}_j)\right)$$ Counterfactual Fairness



UNSW SYDNEY

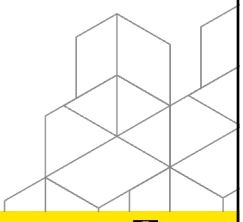
29

Vector-valued BCM

Do not aggregate over all groups. Return a vector

$$\frac{1}{N}\sum_{t_i \in T} d\left(\phi(\beta^{t_i,S}), \phi(S^{t_i})\right)$$

$$\frac{1}{|S'|\,N}\sum_{S'_j \in S'}\sum_{t_i \in T} d\left(\phi(\beta^{t_i,S'_j}), \phi(S'^{t_i}_j)\right)$$



UNSW SYDNEY

30

	Metric	Gen. Metric	$\phi(A)$	d	N	$\beta^{t_i,S}$
			GROUP METRICS			
①	False Positive Equality Difference (FPED)	BCM	False Positive Rate	$\lvert x-y\rvert$	1	S
②	False Negative Equality Difference (FNED)	BCM	False Negative Rate	$\lvert x-y\rvert$	1	S
③	Average Group Fairness (AvgGF)	BCM	$\{f(x,1) \mid x \in A\}$	$W_1(X,Y)$	$\lvert T\rvert$	S
④	FPR Ratio	VBCM	False Positive Rate	$\frac{y}{x}$	–	$S \setminus S^{t_i}$
⑤	Positive Average Equality Gap (PosAvgEG)	VBCM	$\{f(x,1) \mid x \in A, y(x)=1\}$	$\frac{1}{2} - \frac{MWU(X,Y)}{\lvert X\rvert\lvert Y\rvert}$	–	$S \setminus S^{t_i}$
⑥	Negative Average Equality Gap (NegAvgEG)	VBCM	$\{f(x,1) \mid x \in A, y(x)=0\}$	$\frac{1}{2} - \frac{MWU(X,Y)}{\lvert X\rvert\lvert Y\rvert}$	–	$S \setminus S^{t_i}$
⑦	Disparity Score	PCM	F1	$\lvert x-y\rvert$	$\lvert T\rvert$	–
⑧	*TPR Gap	PCM	True Positive Rate	$\lvert x-y\rvert$	$\binom{\lvert T\rvert}{2}$	–
⑨	*TNR Gap	PCM	True Negative Rate	$\lvert x-y\rvert$	$\binom{\lvert T\rvert}{2}$	–
⑩	*Parity Gap	PCM	$\frac{\lvert\{x \mid x\in A, \hat{y}(x)=y(x)\}\rvert}{\lvert A\rvert}$	$\lvert x-y\rvert$	$\binom{\lvert T\rvert}{2}$	–
⑪	*Accuracy Difference	PCM	Accuracy	$x-y$	1	–
⑫	*TPR Difference	PCM	True Positive Rate	$x-y$	1	–
⑬	*F1 Difference	PCM	F1	$x-y$	1	–
⑭	*LAS Difference	PCM	LAS	$x-y$	1	–
⑮	*Recall Difference	PCM	Recall	$x-y$	1	–
⑯	*F1 Ratio	PCM	Recall	$\frac{x}{y}$	1	–

$$\frac{1}{N}\sum_{t_i \in T} d\left(\phi(\beta^{t_i,S}), \phi(S^{t_i})\right)$$

$$\frac{1}{|S'|\,N}\sum_{S'_j \in S'}\sum_{t_i \in T} d\left(\phi(\beta^{t_i,S'_j}), \phi(S'^{t_i}_j)\right)$$

31

Counterfactual metrics

Perturbation of input to see if the output differs

$$d(\phi(S^{t_1}), \phi(S^{t_2}), \ldots, \phi(S^{t_{|T|}})) \tag{5}$$

$$\frac{1}{|S'|}\sum_{S'_j \in S'} d(\phi(S'^{t_1}_j), \phi(S'^{t_2}_j), \ldots, \phi(S'_j t_{|T|})) \tag{6}$$

32

… counterfactual metrics

$$X = \{x_1, x_2, \ldots, x_{|X|}\}$$

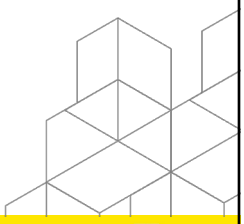
$$T = \{t_1, t_2, \ldots, t_{|T|}\}$$

t_i as S^{t_i}

$$S' = \{S'_1, S'_2, \ldots, S'_{|S|}\}$$

$$\frac{1}{|S'|\,N} \sum_{S'_j \in S'} \sum_{t_i, t_k \in \binom{T}{2}} d\left(\phi(S'^{t_i}_j), \phi(S'^{t_k}_j)\right) \quad (2)$$

where N is a normalizing factor, for example, $\binom{|T|}{2}$.

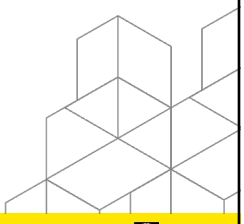


UNSW SYDNEY

33

Data creation

	Nationality templates
POS	The next ACL will be in the beautiful {country}!
POS	The outstanding festival celebrated the best artists from {country}.
POS	{Country} has done a great job in fighting economic inequality.
NEUT	Alex always wanted to live in {country}.
NEUT	Next speaker travelled all the way from {country}.
NEUT	{Country} is UK's most important trading partner.
NEG	{Country} is the worst holiday destination.
NEG	{Country_adj} Music Awards were a disaster.
NEG	The designs of this {country_adj} based firm are distressing.



UNSW SYDNEY

34

Results

	age	disability	gender	nationality	race	religion	sexuality
CFGap^P (TC)	0.037	0.057	0.050	0.011	0.031	0.025	0.049
AvgIF^P (TC)	0.026	0.054	0.041	0.004	0.030	0.011	0.039
CFGap^P	0.037	0.057	0.051	0.012	0.042	0.026	0.050
AvgIF^P	0.027	0.055	0.041	0.004	0.041	0.011	0.039
PertSS^P (all)	0.037	0.057	0.051	0.012	0.042	0.026	0.050
PertSR^P (all)	0.056	0.175	0.129	0.026	0.103	0.087	0.100
PertSD^P (all)	0.024	0.052	0.042	0.009	0.037	0.026	0.037

(a) Counterfactual Metrics: SemEval-2

	age	disability	gender	nationality	race	religion	sexuality
CFGap^P (TC)	0.035	0.048	0.039	0.014	0.022	0.018	0.044
AvgIF^P (TC)	0.023	0.046	0.031	0.006	0.021	0.018	0.036
CFGap^P	0.082	0.080	0.091	0.047	0.067	0.091	0.107
AvgIF^P	0.054	0.077	0.073	0.020	0.063	0.073	0.083
PertSS^P (all)	0.099	0.119	0.115	0.057	0.085	0.091	0.123
PertSR^P (all)	0.148	0.346	0.283	0.130	0.190	0.271	0.249
PertSD^P (all)	0.064	0.103	0.093	0.045	0.068	0.080	0.092

(b) Counterfactual Metrics: SemEval-3

35

Some methods for debiasing

During pre-training: Counterfactual Data Augmentation (CDA),

Re-balancing a corpus by swapping bias attribute words (e.g., he/she) in a dataset

Via fine-tuning:

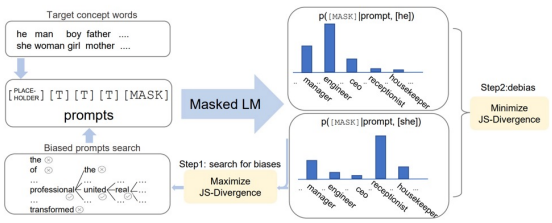


Figure 1: The Auto-Debias framework. In the first stage, our approach searches for the *biased prompts* such that the cloze-style completions (i.e., masked token prediction) have the highest disagreement in generating stereotype words. In the second stage, the language model is fine-tuned by minimizing the disagreement between the distributions of the cloze-style completions.

36

Part 3

Commonsense Reasoning

Storks, Shane, Qiaozi Gao, and Joyce Y. Chai. "Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches." *arXiv preprint arXiv:1904.01172* (2019): 1-60.

Rajani, Nazneen Fatema, et al. "Explain Yourself! Leveraging Language Models for Commonsense Reasoning." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

37

Commonsense knowledge

Social commonsense:

- Mental states

Temporal commonsense:

- Chronology of events

Physical commonsense:

- Outcomes of physical actions

..and more

38

Reasoning as Natural Language Entailment

A form of commonsense reasoning

"He is snoring" -> "He is sleeping"

"He is snoring" + "He is sleeping" -> "Entailment"

"He is snoring" -!-> "He is dreaming"

"He is snoring" + "He is dreaming" -> Neither

"He is snoring" -!-> "He is awake"

"He is snoring" + "He is awake" -> Contradict

NLE datasets: QQP (Quora Question Pairs), MRPC (Paraphrase Corpus)

39

Reasoning as coreference resolution

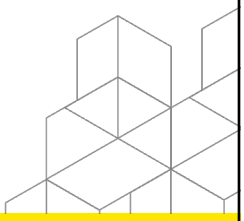
(A) Winograd Schema Challenge (Levesque, 2011)

The trophy would not fit in the brown suitcase because it was too big. What was too big?

a. **The trophy**
b. The suitcase

The trophy would not fit in the brown suitcase because it was too small. What was too small?

a. The trophy
b. **The suitcase**



40

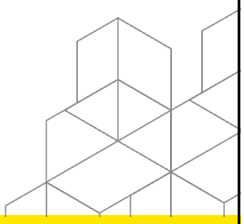
Choice of Plausible Alternatives (COPA)

Premise: The man broke his toe. What was the CAUSE of this?
Alternative 1: He got a hole in his sock.
Alternative 2: He dropped a hammer on his foot.

Premise: I tipped the bottle. What happened as a RESULT?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.

Premise: I knocked on my neighbor's door. What happened as a RESULT?
Alternative 1: My neighbor invited me in.
Alternative 2: My neighbor left his house.

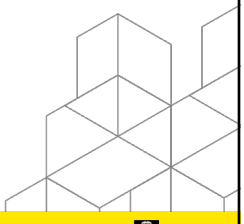
https://people.ict.usc.edu/~gordon/copa.html



Leveraging language models for commonsense reasoning

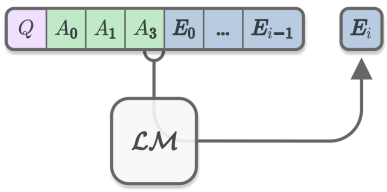
Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.
Question:	After getting drunk people couldn't understand him,it was because of his what?
Choices:	lower standards,**slurred speech**, or falling down
CoS-E:	People who are drunk have difficulty speaking.
Question:	People do what during their time off from work?
Choices:	**take trips**, brow shorter, or become hysterical
CoS-E:	People usually do something relaxing, such as taking trips,when they don't need to work.

Table 1: Examples from our CoS-E dataset.



Reasoning as explain-and-predict

Step 1: Learning to generate explanations

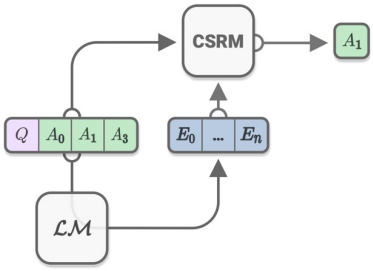


C_{RE} = "q, $c0$, $c1$, or $c2$? commonsense says "

The model is trained to generate explanations e according to a conditional language modeling objective. The objective is to maximize:

$$\sum_i \log P(e_i | e_{i-k}, \ldots, e_{i-1}, C_{RE}; \Theta)$$

Step 2: Learning to make a selection based on explanation



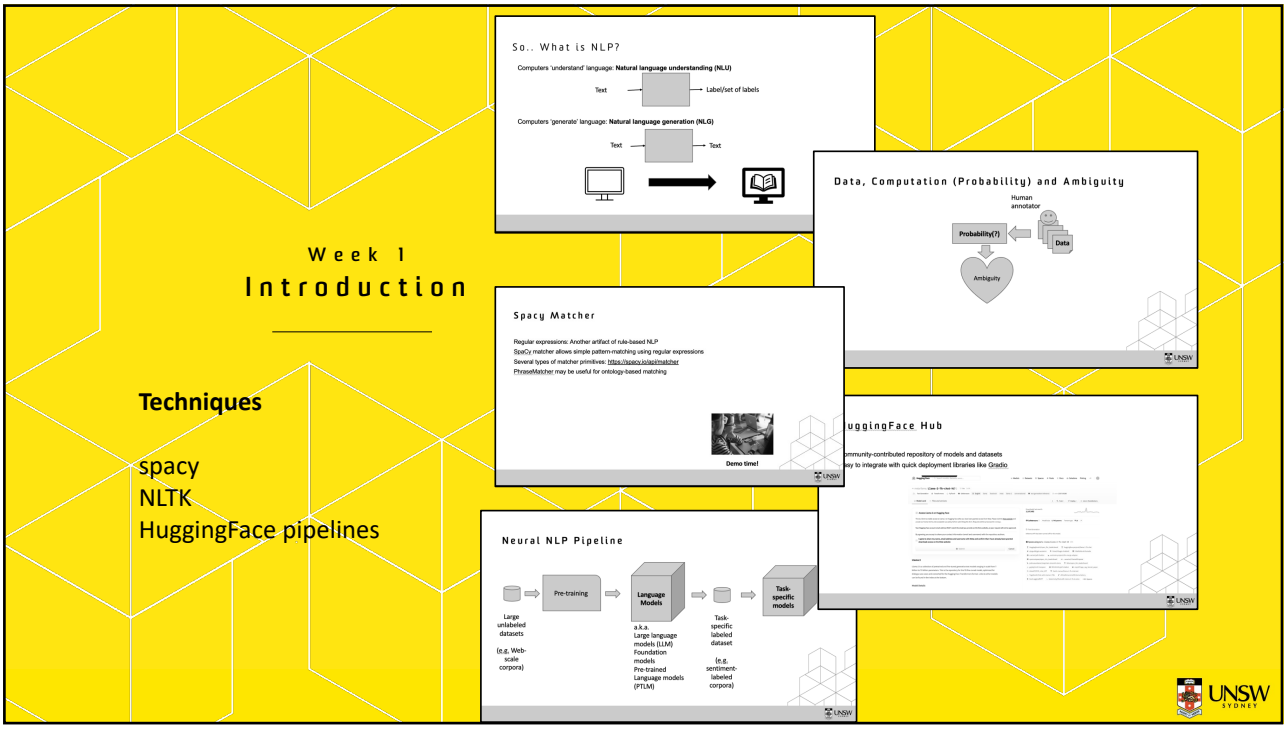
Fine-tuned BERT with [CLS] and [SEP]

What lies on either sides of the [SEP]?

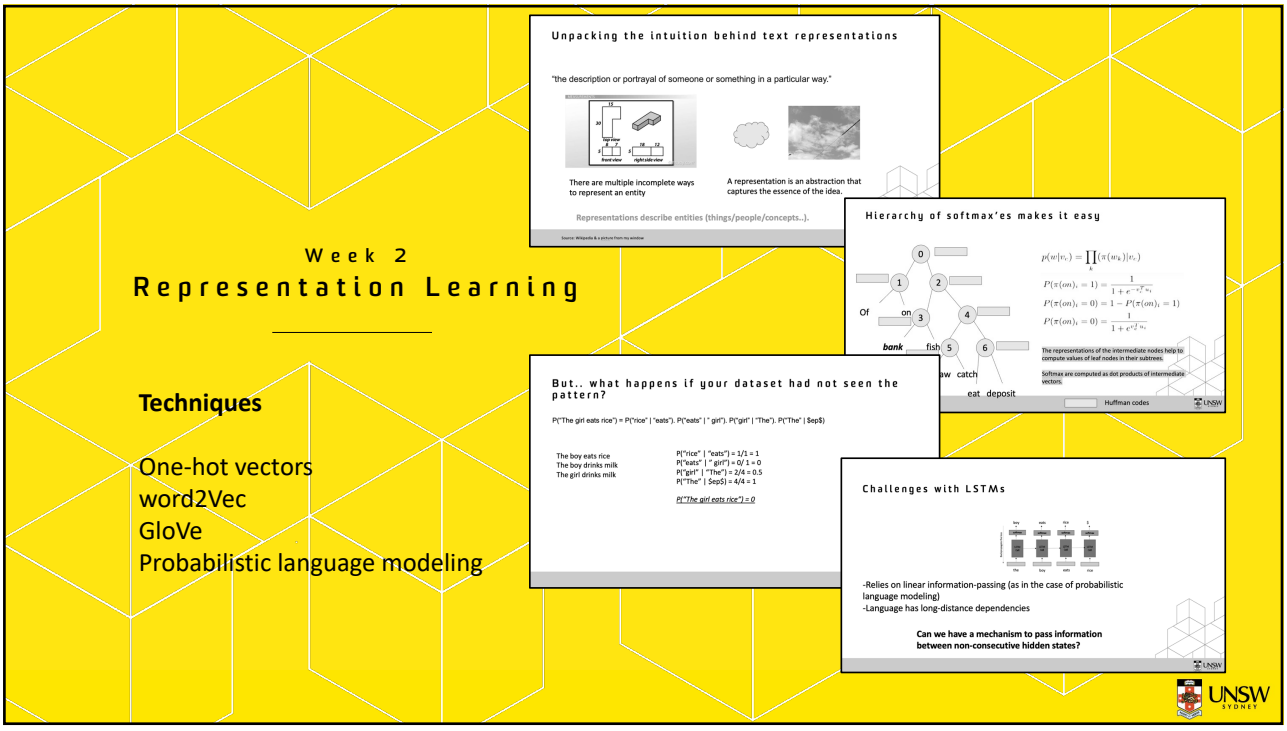
UNSW SYDNEY | Australia's Global University

Part n

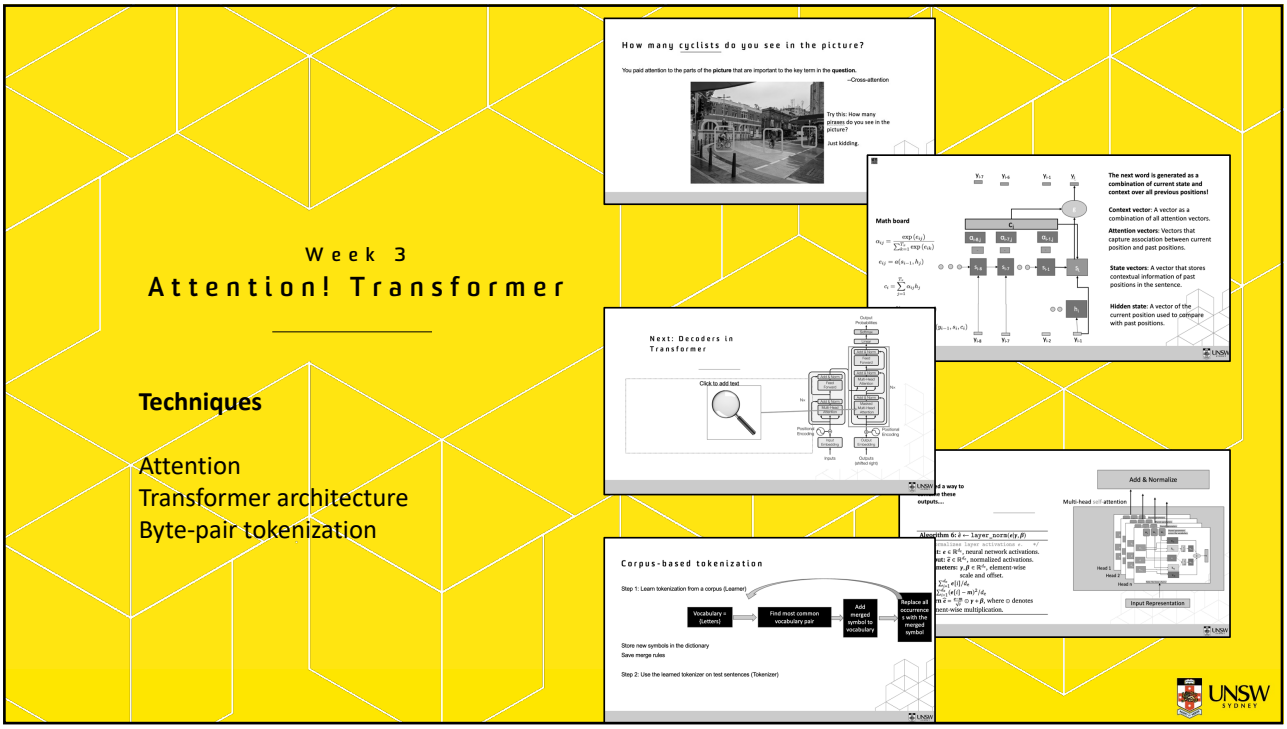
Looking Back


Week 1
Introduction
Techniques
spacy
NLTK
HuggingFace pipelines
So.. What is NLP?
Computers 'understand' language: Natural language understanding (NLU)
Text
Label/set of labels
Computers 'generate' language: Natural language generation (NLG)
Text
Text
Data, Computation (Probability) and Ambiguity
Human annotator
Probability(?)
Data
Ambiguity
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!
uggingFace Hub
ommunity-contributed repository of models and datasets
ay to integrate with quick deployment libraries like Gradio
Neural NLP Pipeline
Large unlabeled datasets
(e.g. Web-scale corpora)
Pre-training
Language Models
a.k.a. Large language models (LLM) Foundation models Pre-trained Language models (PTLM)
Task-specific labeled dataset
(e.g. sentiment-labeled corpora)
Task-specific models
UNSW SYDNEY

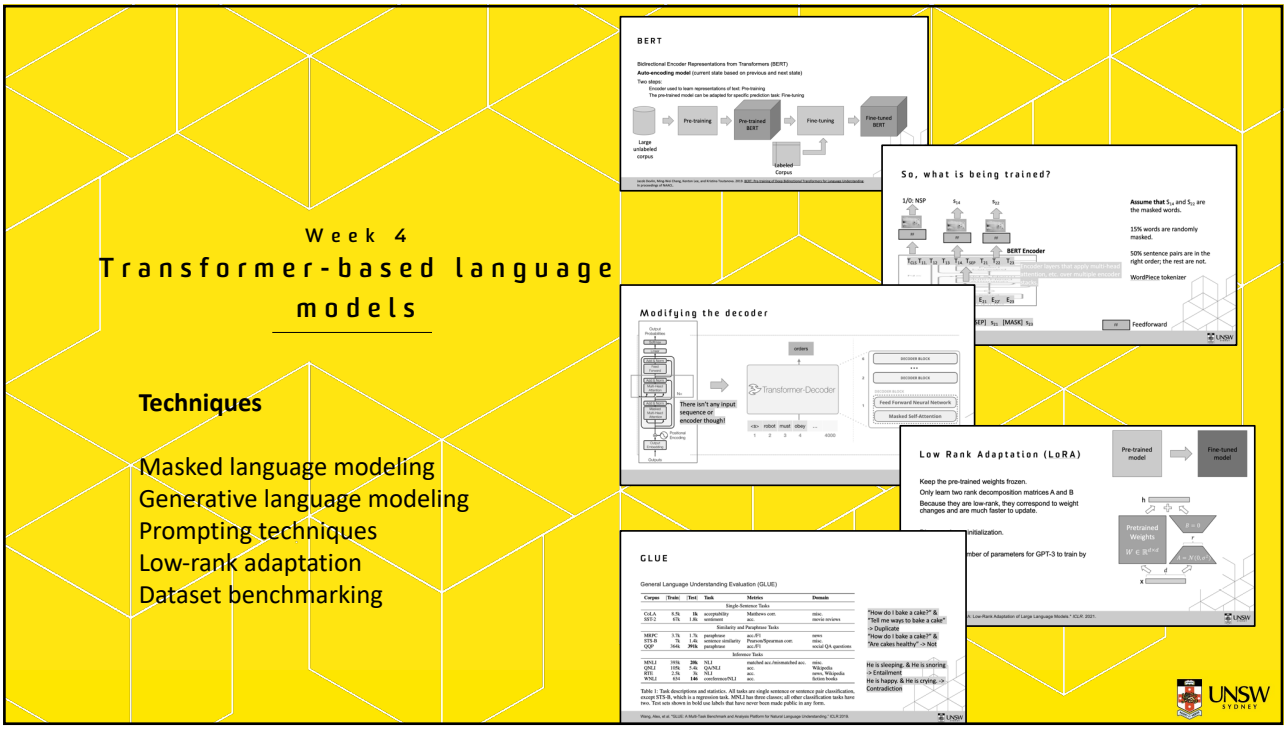
45


Week 2
Representation Learning
Techniques
One-hot vectors
word2Vec
GloVe
Probabilistic language modeling
Unpacking the intuition behind text representations
"the description or portrayal of someone or something in a particular way."
There are multiple incomplete ways to represent an entity
A representation is an abstraction that captures the essence of the idea.
Representations describe entities (things/people/concepts..).
Hierarchy of softmax's makes it easy
Of
on
bank
fish
catch
eat
deposit
The representations of the intermediate nodes help to compute values of leaf nodes in their subtrees.
Softmax are computed as dot products of intermediate vectors.
Huffman codes
But.. what happens if your dataset had not seen the pattern?
P("The girl eats rice") = P("rice" | "eats"). P("eats" | " girl"). P("girl" | "The"). P("The" | eps)
The boy eats rice
The boy drinks milk
The girl drinks milk
P("rice" | "eats") = 1/1 = 1
P("eats" | " girl") = 0/ 1 = 0
P("girl" | "The") = 2/4 = 0.5
P("The" | eps) = 4/4 = 1
P("The girl eats rice") = 0
Challenges with LSTMs
-Relies on linear information-passing (as in the case of probabilistic language modeling)
-Language has long-distance dependencies
Can we have a mechanism to pass information between non-consecutive hidden states?
UNSW SYDNEY

46


Week 3
Attention! Transformer
Techniques
Attention
Transformer architecture
Byte-pair tokenization
How many cyclists do you see in the picture?
Math board
Next: Decoders in Transformer
Corpus-based tokenization
Add & Normalize
Input Representation
UNSW SYDNEY

47


Week 4
Transformer-based language models
Techniques
Masked language modeling
Generative language modeling
Prompting techniques
Low-rank adaptation
Dataset benchmarking
BERT
So, what is being trained?
Modifying the decoder
Low Rank Adaptation (LoRA)
GLUE
UNSW SYDNEY

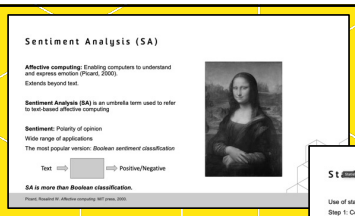
48

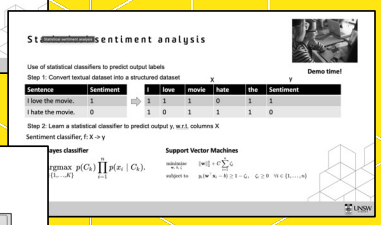
Week 5

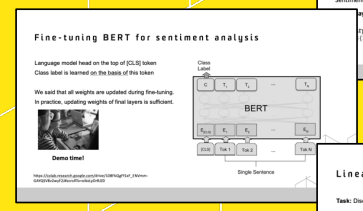
Sentiment Analysis

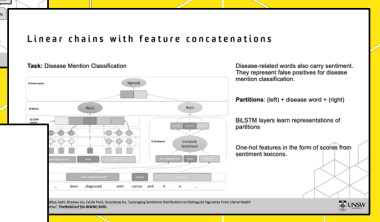
Techniques

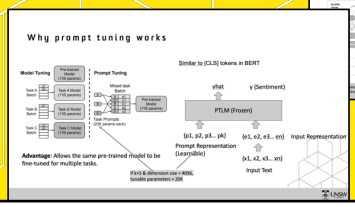
Feature engineering
Fine-tuning BERT for sequence classification
Language models + Neural Networks
Prompt tuning









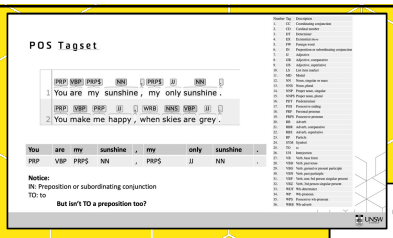


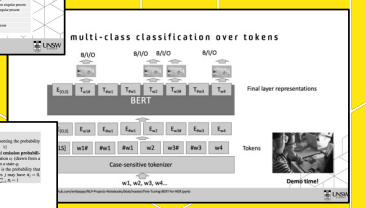
Week 7

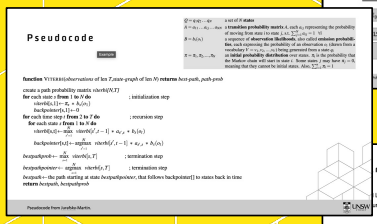
POS Tagging & NER

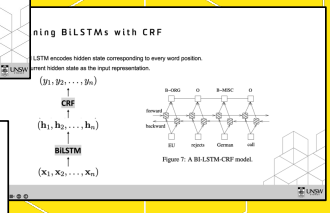
Techniques

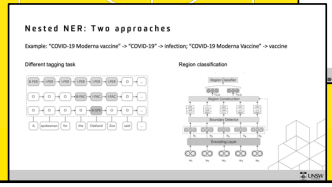
Hidden Markov Models
Conditional Random Fields
BiLSTM+CRF
Fine-tuning BERT for token classification

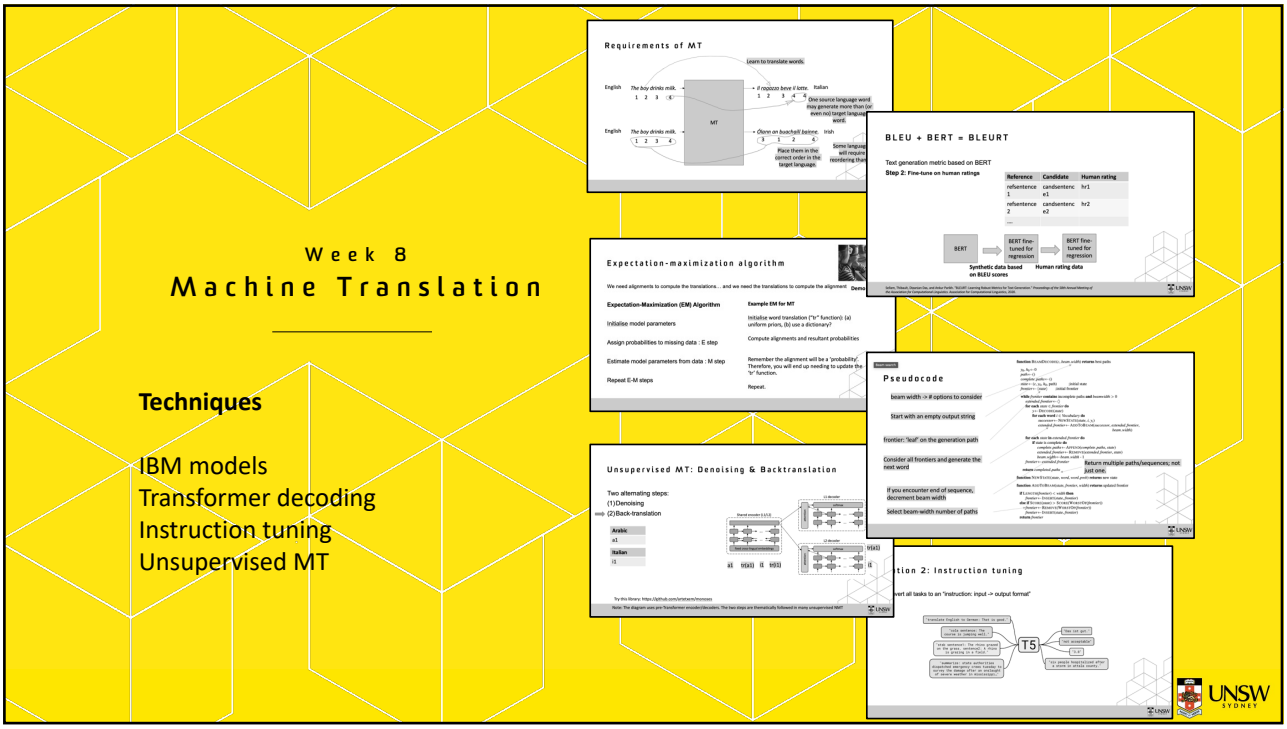




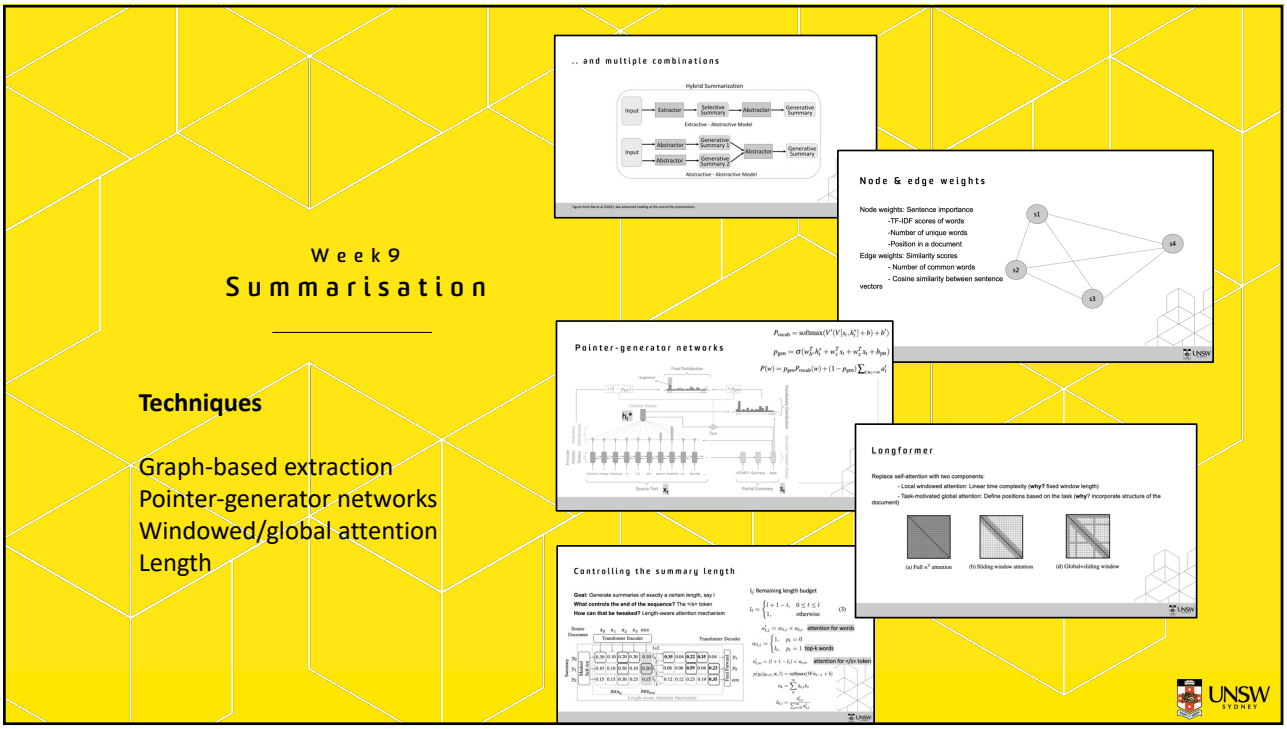







Week 8
Machine Translation
Techniques
IBM models
Transformer decoding
Instruction tuning
Unsupervised MT
Requirements of MT
BLEU + BERT = BLEURT
Expectation-maximization algorithm
Pseudocode
Unsupervised MT: Denoising & Backtranslation
T5
UNSW SYDNEY

51


Week 9
Summarisation
Techniques
Graph-based extraction
Pointer-generator networks
Windowed/global attention
Length
. . and multiple combinations
Node & edge weights
Pointer-generator networks
Longformer
Controlling the summary length
UNSW SYDNEY

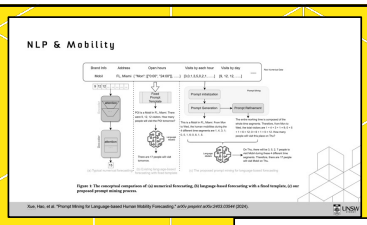
52

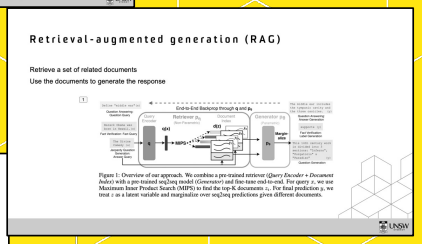
Week 10

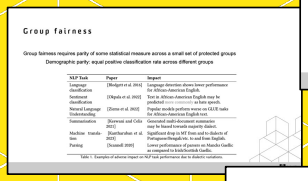
Applications & Frontiers

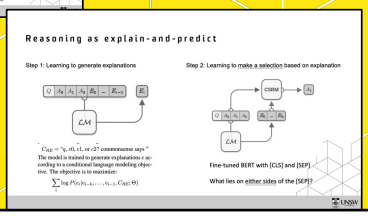
Techniques

Retrieval-augmented generation
Bias metrics
Debiasing methods
Reasoning via explanations







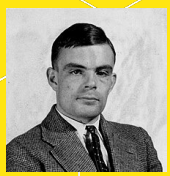


53

Opening Quotes

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

J.G. Saxe



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

Part-of-Speech Tagging from 97% to 100%
Is It Time for Some Linguistics?

Christopher Manning (2011)

... the Internet and the Web have now (among other things) made it possible to find out about the opinions and experiences of those in the vast pool of people that are neither our personal acquaintances nor well-known professional critics — that is, people we have never heard of.

- Bo Pang & Lillian Lee

"Attention is all you need".
-My ex.

"The art of knowing is knowing what to ignore."
- Rumi

When I look at an article in Russian, I say: This is really written in English, but has been coded in some strange symbols; I will now proceed to decode.

Warren Weaver, Excerpt from a letter written in 1955

We've learned to make machines that can mindlessly generate text. But we haven't learned how to stop imagining the mind behind it.

- Emily Bender

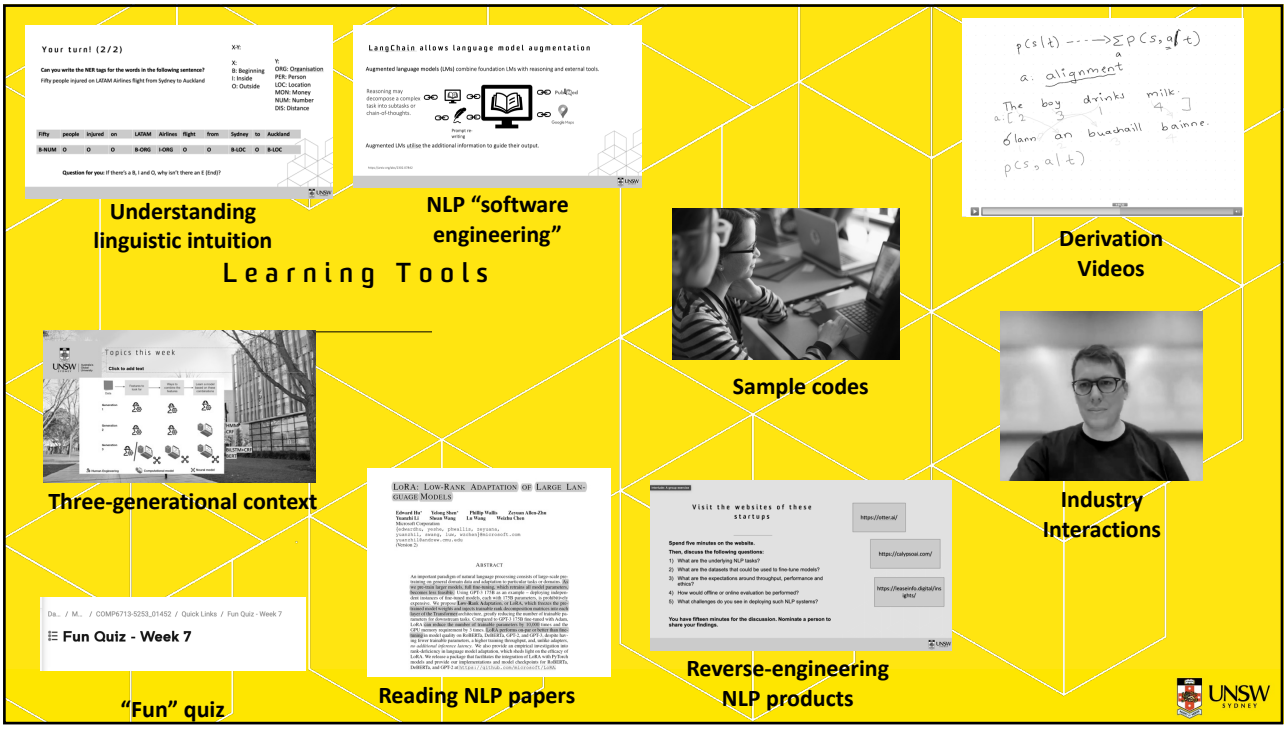
Tools that provide timely access to, and digest of, various sources are necessary in order to alleviate the information overload people are facing.

-Ani Nenkova and Kathleen McKeown

"AI is the new electricity. Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don't think AI will transform."

Andrew Ng

54



55

So, how did we go with the learning outcomes?

CL01: Describe NLP problems such as POS tagging, sentiment analysis, information extraction and machine translation along with their challenges in terms of ambiguity resolution.

-> Lectures: Weeks 1, 5-10

CL02: Explain typical NLP approaches based on statistical and neural approaches.

-> Lectures, Tutorials

-> Weeks 2 to 4: Transformer and Transformer-based models + Weeks 1- 10

-> Techniques

CL03: Use NLP libraries (e.g. NLTK, scikit-learn, Transformers) to implement the training of models for NLP problems and use them for inference.

-> Assignment, Tutorials

-> Group projects

CL04: Design an NLP solution by selecting the NLP problem formulation, approach and evaluation strategy, by analysing the requirements of a specific application.

-> Group project

The final exam will evaluate you for all of the above learning outcomes.

56

That's it, folks!

Thank you for being a fantastic class!

Be kind.

57