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

All images from Wikimedia commons, unless specified.

1

COMP6713 Natural Language Processing

Student

https://go.blueja.io/DIp4NklFqUaTJH7kFAZAZQ

To access the evaluation, scan this QR code with your mobile phone.

Please give us your feedback. ☺

If you liked the course and/or our teaching, please say so.

It will really help us.

If the response rate reaches 70%, I will be releasing additional information about the questions to expect in the final exam.

Note: Thursday 17th April Lecture will run as usual:

RL for summarisation
Talk by Dr. Ben Hutchinson (Google)

2

Natural Language Processing (NLP)

COMP6713 – 2025 Term 1

Convener

Dr. Aditya Joshi

aditya.joshi@unsw.edu.au

Week 9

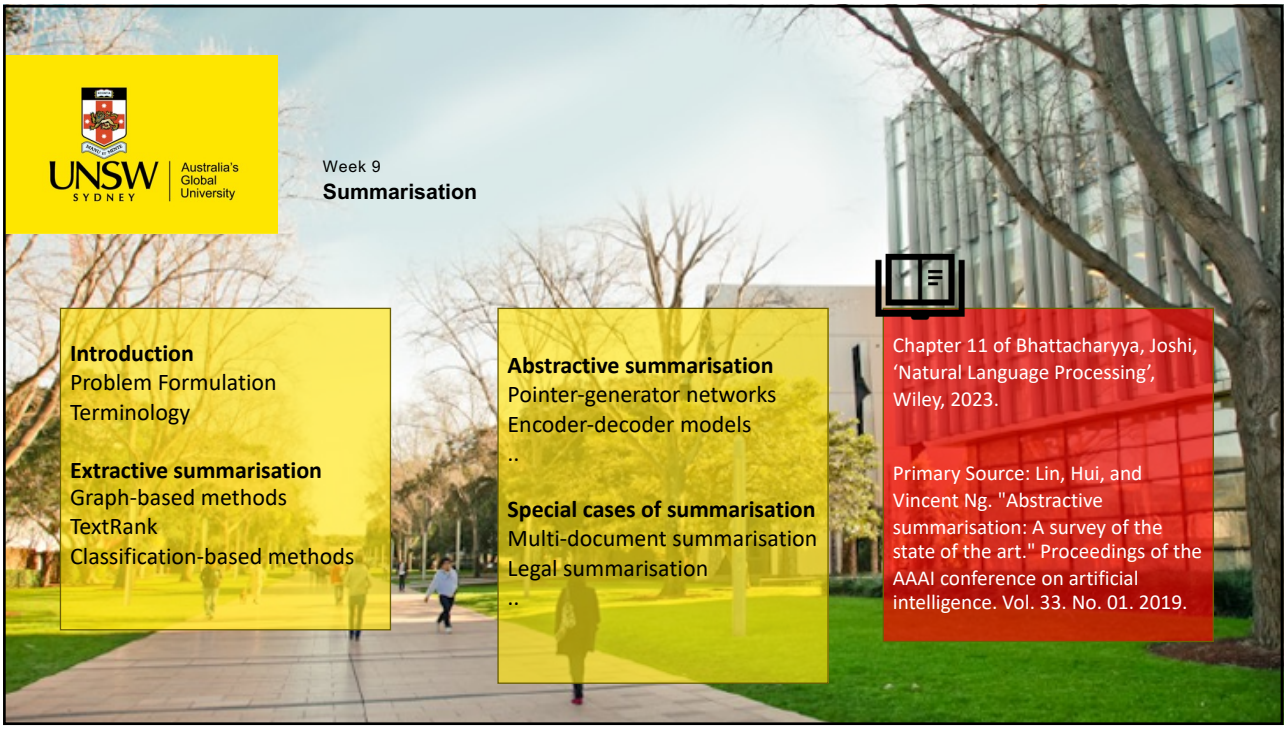
Summarisation

Schedule

2025 Term 1

UNSW SYDNEY

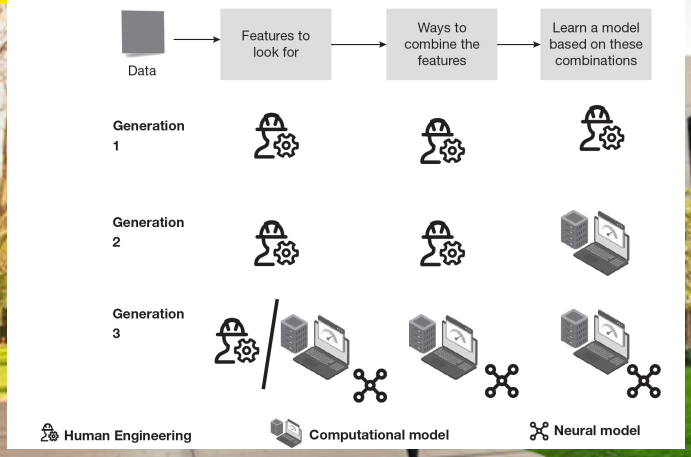
3



4

UNSW SYDNEY | Australia's Global University

Topics this week



Features for extractive summarisation

TextRank

Pointer-generative networks

Encoder-decoder models

Multi-document summarisation

5

UNSW SYDNEY | Australia's Global University

Part 0

What and why

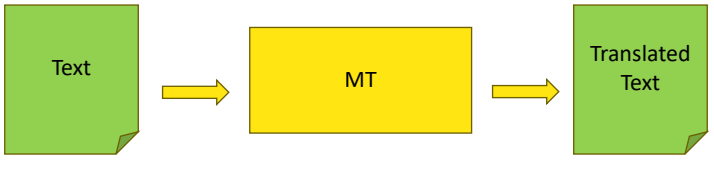
6

Looking back…

	Input	Output
Week 5: Sentiment Analysis	Sentence or document	Sequence Classification
Week 7: POS Tagging & NER	Sentence	Token Classification
Week 8: Machine Translation	Sentence or document	Sequence-to-sequence generation
Week 9: summarisation	Document	Sequence-to-sequence generation

7

MT & summarisation: seq2seq but different



- Different vocabulary as the input
- Almost the same length as the input

- Same vocabulary as the input
- Smaller length than the input

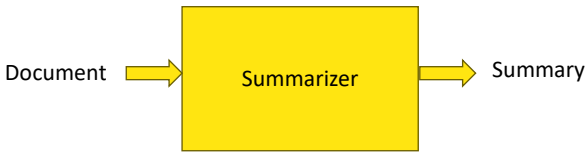
8

What is a summary?

the act of expressing the most important facts or ideas about something or someone in a short and clear form, or a text in which these facts or ideas are expressed.

Cambridge Dictionary.

..therefore, summarisation:



9

Document structure may sometimes bear a summary

Journalism

Tasmania has elected a hung parliament. So what does that mean, and how do minority governments work?

By Ashleigh Barraclough

Posted Tue 26 Mar 2024 at 5:19am, updated Tue 26 Mar 2024 at 4:40pm

As the numbers for the Tasmanian election rolled in on Saturday night, it quickly became clear neither Labor nor the Liberals would be able to form majority government.

But the Liberal party is projected to win more seats, with Jeremy Rockliff declaring victory on election night.

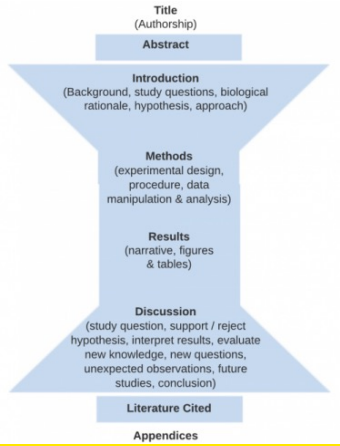
On Sunday, Labor leader Rebecca White conceded it was now up to Mr Rockliff to form government.

https://www.abc.net.au/news/2024-03-26/tasmania-election-minority-government-hung-parliament-explainer/103627670

10

Document sections may sometimes bear a summary

Research



Natural Language Processing for Dialects of a Language: A Survey

ADITYA JOSHI, University of New South Wales, Australia
RAJ DABRE, National Institute of Information and Communications Technology, Japan
DIPTESH KANOJIA, Institute for People-Centred AI, University of Surrey, United Kingdom
ZHUANG LI, Monash University, Australia
HAOLAN ZHAN, Monash University, Australia
GHOLAMREZA HAFFARI, Monash University, Australia
DORIS DIPPOLD, University of Surrey, United Kingdom

State-of-the-art natural language processing (NLP) models are trained on massive training corpora, and report a superlative performance on evaluation datasets. This survey delves into an important attribute of these datasets: the dialect of a language. Motivated by the performance degradation of NLP models for dialectic datasets and its implications for the equity of language technologies, we survey past research in NLP for dialects in terms of datasets, and approaches. We describe a wide range of NLP tasks in terms of two categories: natural language understanding (NLU) (for tasks such as dialect classification, sentiment analysis, parsing, and NLU benchmarks) and natural language generation (NLG) (for summarisation, machine translation, and dialogue systems). The survey is also broad in its coverage of languages which include English, Arabic, German among others. We observe that past work in NLP concerning dialects goes deeper than mere dialect classification, and . This includes early approaches that used sentence transduction that lead to the recent approaches that integrate hypernetworks into LoRA. We expect that this survey will be useful to NLP researchers interested in building equitable language technologies by rethinking LLM benchmarks and model architectures.

https://wisc.pb.unizin.org/biocorewriting/chapter/structure-of-a-scientific-research-paper/

UNSW SYDNEY

11

What makes a good summary?

A summary must be…	From the NLP perspective..
Comprehensive: Must contain all important points	… cover important words
Concise: Must not repeat points	… do not repeat phrases during generation
Independent: Use your own words	… abstractive summaries are favoured
Coherent: Must not be disjoint points	… the discourse must be coherent

https://www.hunter.cuny.edu/rwc/handouts/the-writing-process-1/invention/Guidelines-for-Writing-a-Summary

UNSW SYDNEY

12

Summarisation

Objective: Given a reference document, create a summary that is: (a) shorter in length, (b) covers the core content of the reference document.

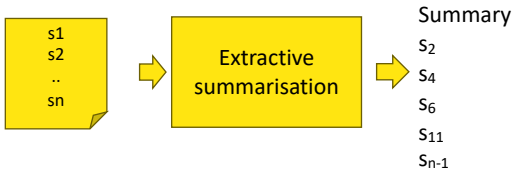
sequence-to-sequence task

May not always be modeled as that.

Instruction tuning for summarisation: https://huggingface.co/docs/transformers/en/tasks/summarisation

Two types of summarisation

Extractive summarisation

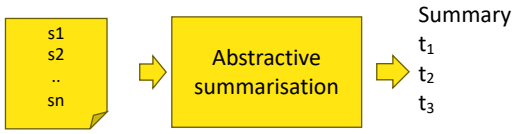


The summary is a subset of sentences in the original document.

The number of sentences to be retrieved may be a parameter.

Task formulations: Retrieval, Classification

Abstractive summarisation



The summary is a set of sentences that may not exactly occur in the original document.

The number of sentences to be generated may be a parameter.

Task formulations: (Long-) Seq2seq

.. and multiple combinations

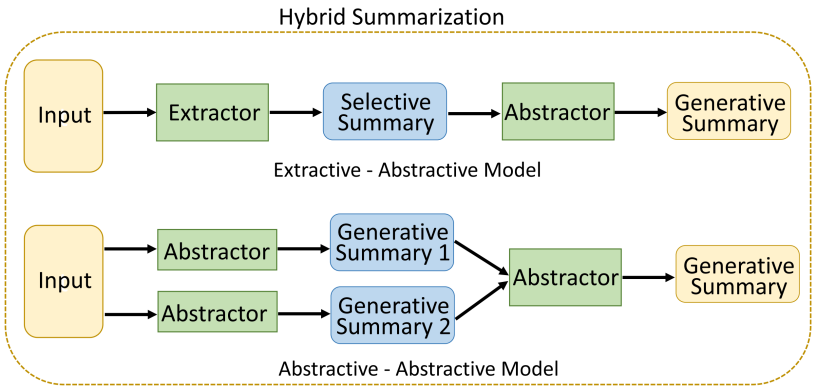
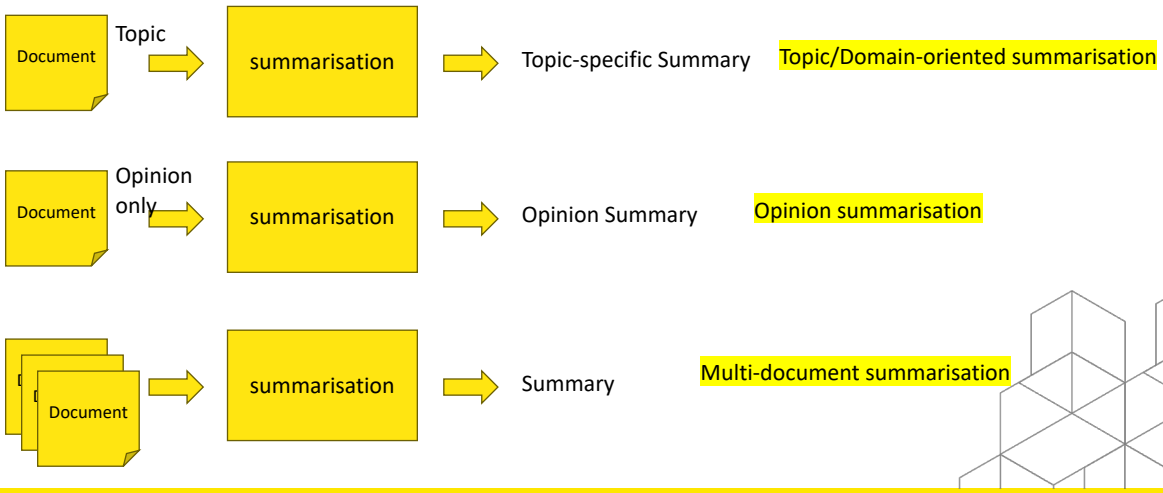


Figure from Ma et al (2022). See advanced reading at the end of the presentation.

15

Other types of summarisation



16

Applications

News summarisation
Opinion summarisation
Meeting summarisation

Think: What aspects of data will influence modeling choices?

Demo time!

17

Pop question

How would you empirically evaluate summarisation?
Using ROUGE, BLEU, etc. with respect to a reference summary.

Source: Microsoft Stock Images

18

UNSW SYDNEY | Australia's Global University

Part 1

Extractive summarisation

Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into text." *EMNLP 2004*.

19

Two types of approaches

Graph-based algorithms

- Inspired from HITS and PageRank algorithms used in search engines
- Intuition: Decide the importance of a vertex in a graph

Classification-based algorithms

- Selection decision for every sentence
- Feature engineering

UNSW SYDNEY

20

Extractive summary: Example

s1: A New York lawyer is facing a court hearing of his own after his firm used AI tool ChatGPT for legal research.

s2: A judge said the court was faced with an "unprecedented circumstance" after a filing was found to reference example legal cases that did not exist.

s3: The lawyer who used the tool told the court he was "unaware that its content could be false".

s4: ChatGPT creates original text on request, but comes with warnings it can "produce inaccurate information".

s5: The original case involved a man suing an airline over an alleged personal injury. His legal team submitted a brief that cited several previous court cases in an attempt to prove, using precedent, why the case should move forward.

s6: A judge said the court was faced with an "unprecedented circumstance" after a filing was found to reference example legal cases that did not exist.

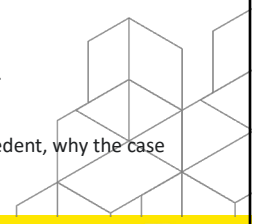
s7: The lawyer who used the tool told the court he was "unaware that its content could be false".

s8: ChatGPT creates original text on request, but comes with warnings it can "produce inaccurate information".

s9: The original case involved a man suing an airline over an alleged personal injury.

s10: His legal team submitted a brief that cited several previous court cases in an attempt to prove, using precedent, why the case should move forward.

https://www.bbc.com/news/world-us-canada-65735769



21

Extractive summary: Example

s1: A New York lawyer is facing a court hearing of his own after his firm used AI tool ChatGPT for legal research.

s2: A judge said the court was faced with an "unprecedented circumstance" after a filing was found to reference example legal cases that did not exist.

s3: The lawyer who used the tool told the court he was "unaware that its content could be false".

s4: ChatGPT creates original text on request, but comes with warnings it can "produce inaccurate information".

s5: The original case involved a man suing an airline over an alleged personal injury. His legal team submitted a brief that cited several previous court cases in an attempt to prove, using precedent, why the case should move forward.

s6: A judge said the court was faced with an "unprecedented circumstance" after a filing was found to reference example legal cases that did not exist.

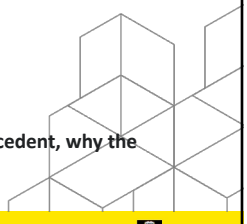
s7: The lawyer who used the tool told the court he was "unaware that its content could be false".

s8: ChatGPT creates original text on request, but comes with warnings it can "produce inaccurate information".

s9: The original case involved a man suing an airline over an alleged personal injury.

s10: His legal team submitted a brief that cited several previous court cases in an attempt to prove, using precedent, why the case should move forward.

https://www.bbc.com/news/world-us-canada-65735769



22

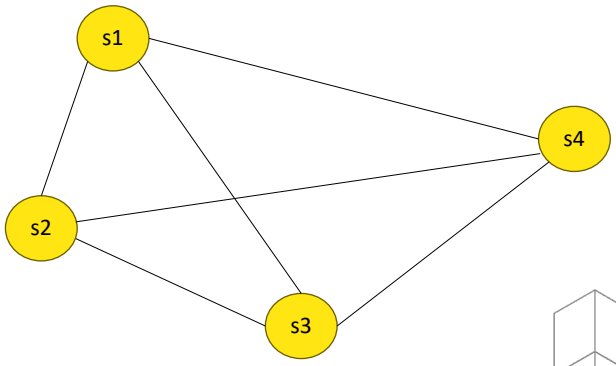
Graph-based summarisation

A document is represented as a complete graph.

Nodes: sentences

Extractive summarisation: Select a subset of nodes

What kind of graph algorithms can you think of?



23

Node & edge weights

Node weights: Sentence importance

-TF-IDF scores of words

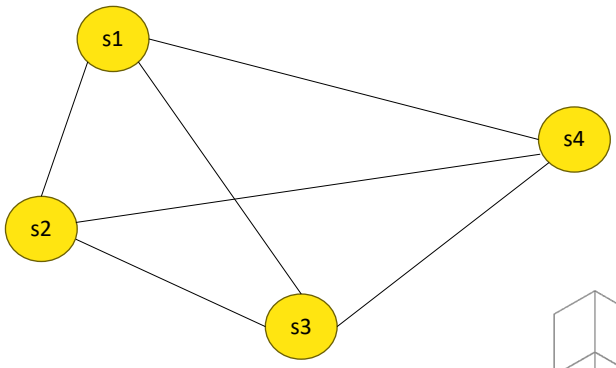
-Number of unique words

-Position in a document

Edge weights: Similarity scores

- Number of common words

- Cosine similarity between sentence vectors



24

TextRank

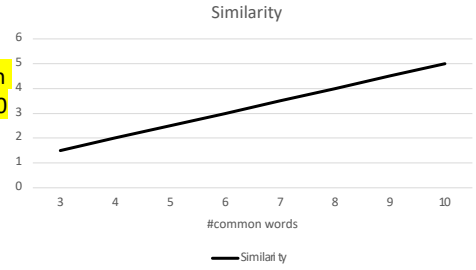
A seminal algorithm in extractive summarisation
Unsupervised algorithm
Derives salience scores for sentences
Summaries are based on a clipped set based on the ranked list of sentences

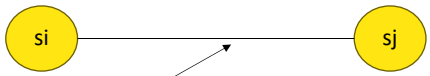
summa (Python library)

25

Edges in TextRank

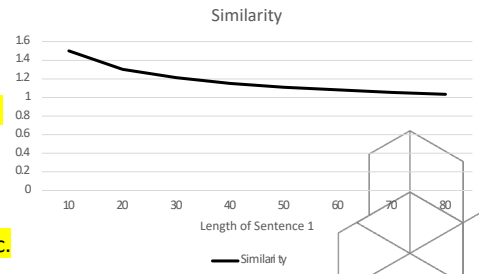
Length of both sentences = 10





$$Similarity(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)}$$

#common words = 3



Alternatives: Longest common subsequence, cosine similarity, etc.

26

Nodes in TextRank

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

```python
def pagerank_weighted(graph, initial_value=None, damping=0.85):
    """Calculates PageRank for an undirected graph"""
    if initial_value == None: initial_value = 1.0 / len(graph.nodes())
    scores = dict.fromkeys(graph.nodes(), initial_value)

    iteration_quantity = 0
    for iteration_number in range(100):
        iteration_quantity += 1
        convergence_achieved = 0
        for i in graph.nodes():
            rank = 1 - damping
            for j in graph.neighbors(i):
                neighbors_sum = sum(graph.edge_weight((j, k)) for k in graph.neighbors(j))
                rank += damping * scores[j] * graph.edge_weight((j, i)) / neighbors_sum

            if abs(scores[i] - rank) <= CONVERGENCE_THRESHOLD:
                convergence_achieved += 1

            scores[i] = rank

        if convergence_achieved == len(graph.nodes()):
            break

    return scores
```

Source: summa library

Algorithm

1. Initialise similarity matrix
2. Compute node scores
3. Update node scores over multiple iterations ------ Remember HITS algorithm?

Demo time!


```python
def pagerank_weighted(graph, initial_value=None, damping=0.85):
    """Calculates PageRank for an undirected graph"""
    if initial_value == None: initial_value = 1.0 / len(graph.nodes())
    scores = dict.fromkeys(graph.nodes(), initial_value)

    iteration_quantity = 0
    for iteration_number in range(100):
        iteration_quantity += 1
        convergence_achieved = 0
        for i in graph.nodes():
            rank = 1 - damping
            for j in graph.neighbors(i):
                neighbors_sum = sum(graph.edge_weight((j, k)) for k in graph.neighbors(j))
                rank += damping * scores[j] * graph.edge_weight((j, i)) / neighbors_sum

            if abs(scores[i] - rank) <= CONVERGENCE_THRESHOLD:
                convergence_achieved += 1

            scores[i] = rank

        if convergence_achieved == len(graph.nodes()):
            break

    return scores
```

Source: summa library

summarisation as classification

Supervised methods

Sentence selection as a **skewed*** classification task

Typical statistical classifiers

Document	Summary
s1	s3
s2	s8
s3	s9
..	
s10	

Document text	Sentence	Appears in the summary?
[s1, s2… s10]	s1	0
	s2	0
	s3	1
	s4	0
	s5	0
	s6	0
	s7	0
	s8	1
	s9	1
	s10	0

* Why?

29

Features

Document text	Sentence	Appears in the summary?
[s1, s2… s10]	s1	0
	s2	0
	s3	1
	s4	0
	s5	0
	s6	0
	s7	0
	s8	1
	s9	1
	s10	0

Sentence features:
Unigrams
Modified TF-IDF (IDF -> ISF!)

Document-Sentence features:
Position in the document
Unique words
…

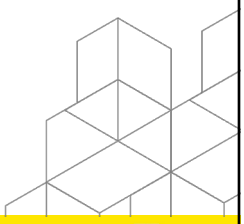
30

Exercise

You are building a summarizer that creates an extractive summary of a research paper (without the abstract).

How will you create a labeled dataset?

What features would you use?



UNSW SYDNEY

31

COMP6713 Natural Language Processing

Student

https://go.blueja.io/DIp4NklFqUaTJH7kFAZAZQ

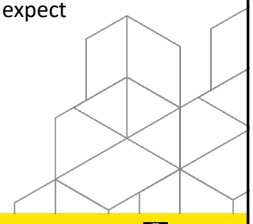
To access the evaluation, scan this QR code with your mobile phone.

Please give us your feedback. ☺

If you liked the course and/or our teaching, please say so.

It will really help us.

If the response rate reaches 70%, I will be releasing additional information about the questions to expect in the final exam.



UNSW SYDNEY

32

Part 2

Abstractive summarisation

Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The long-document transformer." *arXiv preprint arXiv:2004.05150* (2020).
Lewis, Mike, et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." ACL 2020.
See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get To The Point: summarisation with Pointer-Generator Networks." *ACL*, 2017.

33

Quick note: Pre-deep learning abstractive summarisation

Source: the sri lanka government on wednesday announced the closure of government schools with immediate effect as a military campaign against tamil separatists escalated in the north of the country.

Summary: sri lanka **closes** schools as **war escalates**.

Can you think of ways to create summaries using statistical/rule-based techniques?

➔ Hint: Probabilistic language modeling

➔ Hint: Templates

Step 1: Identify important phrases

Step 2: Use a probabilistic language model to 'stitch together' the sentence

34

Neural summarisation

Pre-Transformer

- Pointer-generator networks

Post-Transformer

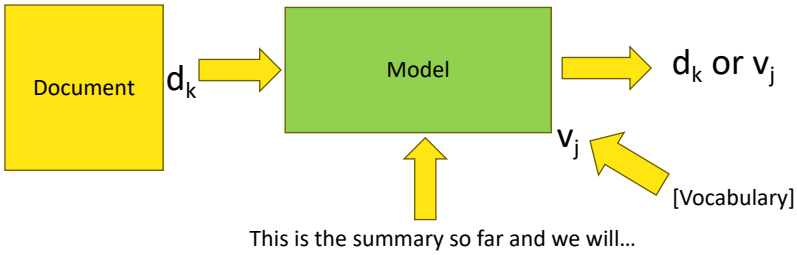
- Denoising as a heuristic for summarisation
- Encoder-decoder models

35

Underlying operations in abstractive summarisation

Inspired from statistical approaches, choose:

1. Select words from the document
2. Generate new words from the underlying document



36

Pointer-generator networks

Pointer networks

- Conditional language modeling corresponding to positions in input sequence [1]
- … only specific words in the vocabulary
- Select words from the source text

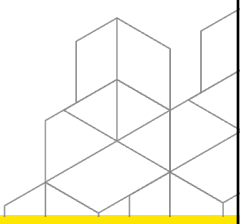
Generator networks

- Conditional generation of words over the vocabulary
- Generate new words

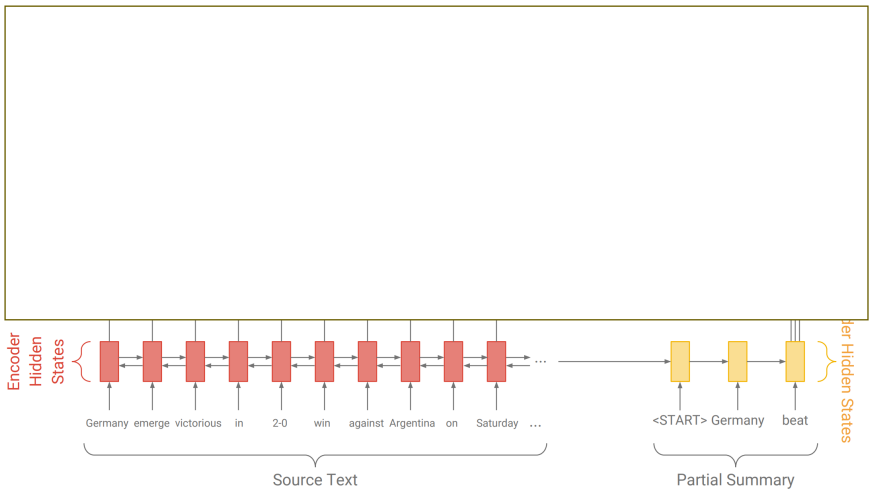
Abstractive summarisation involves the choice between:

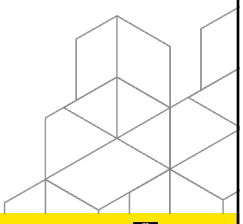
- 'Point' &
- 'Generate'

[1] https://arxiv.org/abs/1506.03134

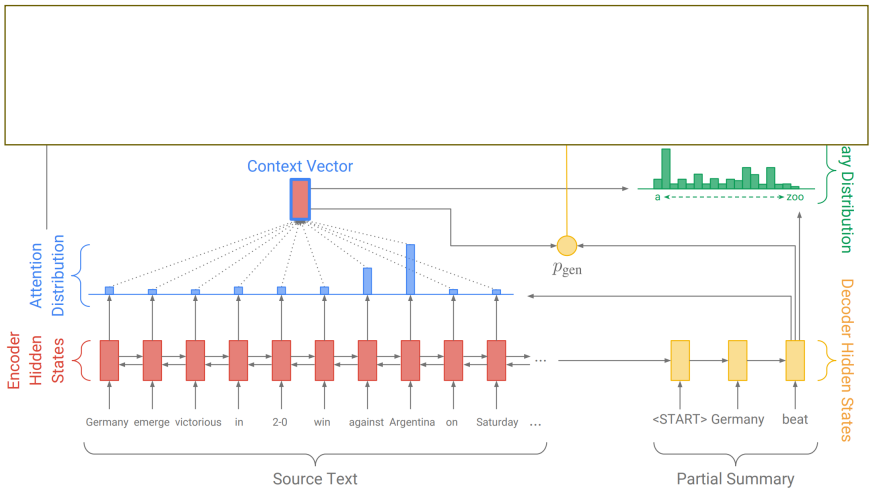


Pointer-generator networks

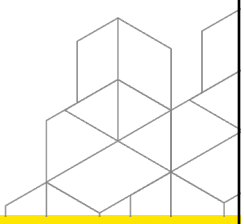




Pointer-generator networks



pgen controls the variability between pointer & the generator



UNSW SYDNEY

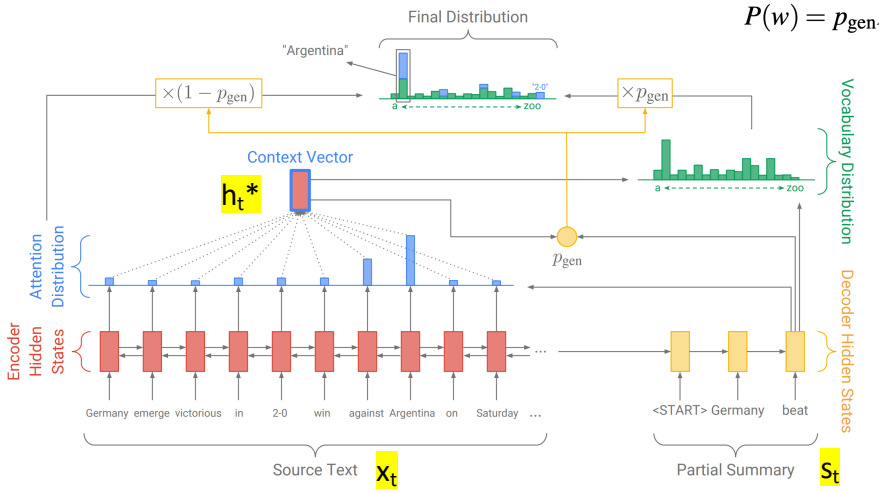
39

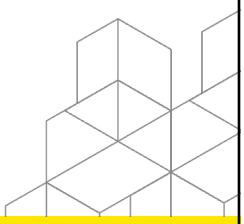
Pointer-generator networks

$$P_{\text{vocab}} = \text{softmax}(V'(V[s_t, h_t^*] + b) + b')$$

$$p_{\text{gen}} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}})$$

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i = w} a_i^t$$





UNSW SYDNEY

40

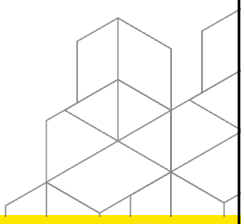
Post-Transformer abstractive summarisation

Can pointer-generation choice be implicitly modeled in a decoder?

Encoder-decoder models

… but aren't Transformers encoder-decoder models?!

How are encoder-decoder models different?

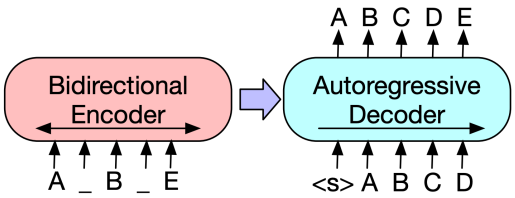


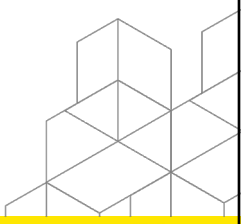
BART: Transformer-based model

Bidirectional and Auto-Regressive Transformers (BART)

- Bidirectional
- …. and auto-regressive
- Huh?

Encoder-decoder model that combines MLM and next-word-prediction

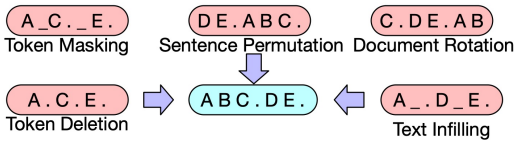




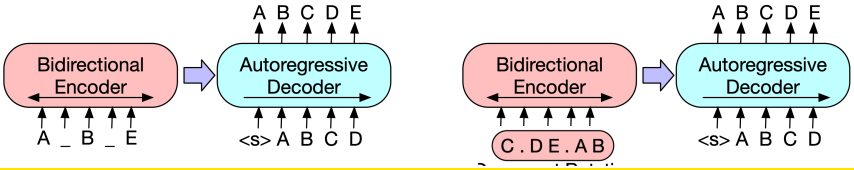
BART

Core concept: Denoising (Remember unsupervised NMT?)

Step 1: Corrupt the original text with an arbitrary noising function



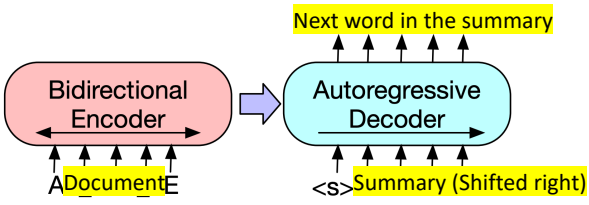
Step 2: Learning a model to reconstruct the original text.



...and other options.

Fine-tuning BART for summarisation

Vanilla seq2seq fine-tuning



…but vanilla encoder/decoder may be really slow.

Self-attention has quadratic time complexity – why?

This restricts the use of Transformer to short documents

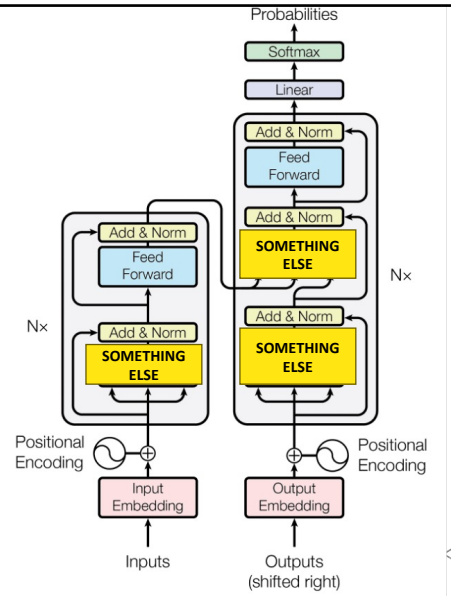
512 tokens

summarisation depends on models that can handle long documents

Therefore, Longformer: a model that works on long documents

- **-**Sliding window attention
- **-**Global attention

https://github.com/allenai/longformer

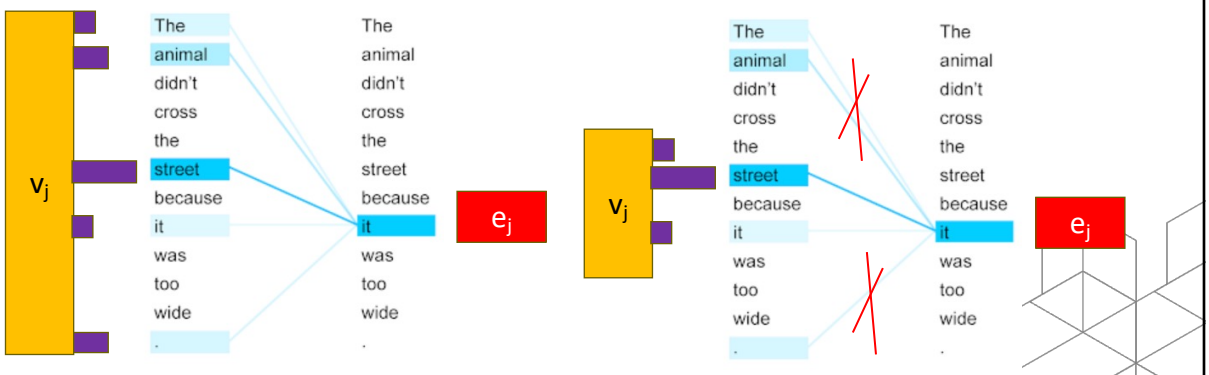


45

Window attention

Self-attention (in Transformer): Attention for a token with respect to all other tokens in the same sequence.

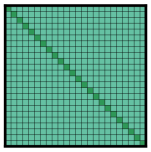
(Sliding) Window attention (in Longformer): Attention for a token with respect **to tokens within a context window.**



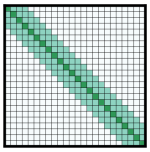
*Dilated windowed attention: Similar to skip-grams in Word2Vec; not shown to work as effectively for this model.

46

Do we need more?



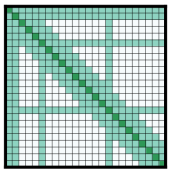
(a) Full n^2 attention



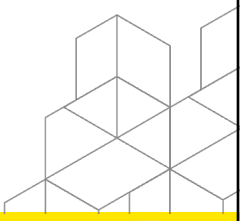
(b) Sliding window attention

Did sliding window attention optimize too much?

Can we compute attention over ALL positions for some FIXED token positions?



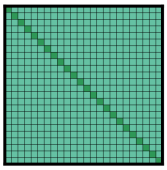
(d) Global+sliding window



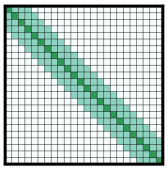
Longformer

Replace self-attention with two components:

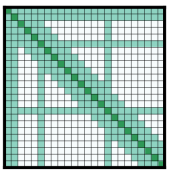
- Local windowed attention: Linear time complexity (**why?** fixed window length)
- Task-motivated global attention: Define positions based on the task (**why**? incorporate structure of the document)



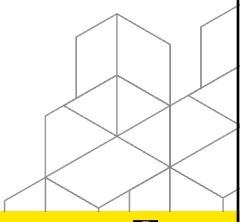
(a) Full n^2 attention



(b) Sliding window attention



(d) Global+sliding window



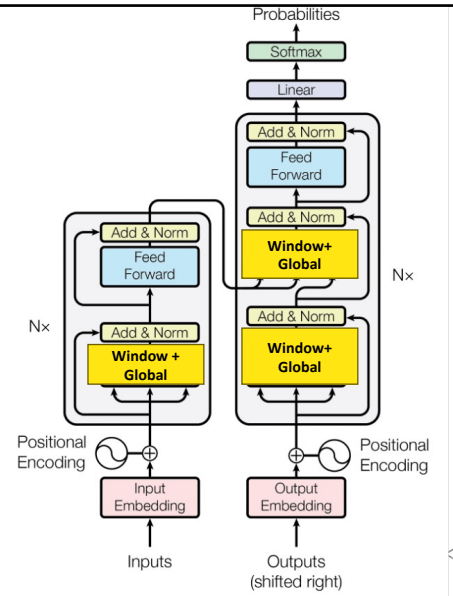
Longformer Encoder-Decoder (LED) summarisation

Initialise the parameters with another model: RoBERTa or BART

Local attention: 1024 tokens

Global attention (for encoder): The first <s> token

Demo time!



49

Part 3

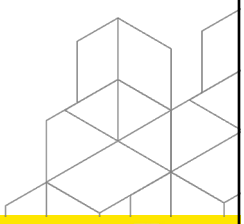
Special Cases of summarisation

Ma, Congbo, et al. "Multi-document summarisation via deep learning techniques: A survey." *ACM Computing Surveys* 55.5 (2022): 1-37.
Wang, Qiqi, et al. "Towards Legal Judgment summarisation: A Structure-Enhanced Approach." *ECAI2023* (2023).
Liu, Yizhu, Qi Jia, and Kenny Zhu. "Length control in abstractive summarisation by pretraining information selection." ACL 2022.

50

Special cases

1. Multi-document summarisation
   - Information from multiple documents needs to be combined
2. Controlling the structure
   - Legal summaries follow a specific structure
3. Controlling the length
   - The summary must be exactly of a certain length



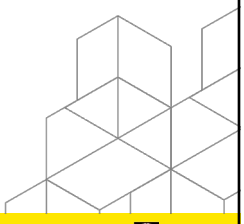
51

Multi-document summarisation

Input consists of multiple documents

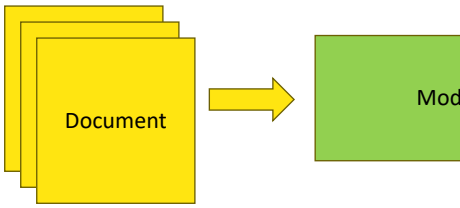
Requirements:

- Summarize
-Avoid repetition
-Ensure representation



52

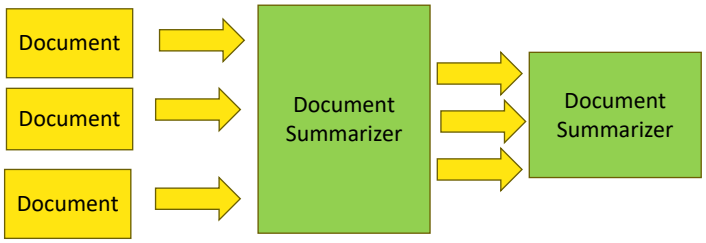
Alternatives



Summary

How would you extend pointer-generator networks?

How would you extend encoder-decoder models?



Summary

53

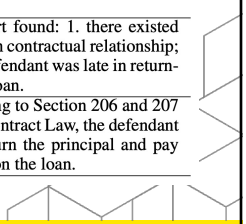
Controlling the summary structure: Legal documents

Table 1. Summary structure for New Zealand judgments

Part	Description	Example
Decision	What the court decides	unsuccessful application by r for leave to commence proceeding.
Fact	Essential factual information of the case	r a vexatious litigant; r seeking to commence proceeding seek 'exemplary damage ' of ' $ 900 trillion ' against the high court for allegedly fail to process an application for leave to appeal a judgment from 2013.
Reason	Reason for the decision	Held, proposed proceeding frivolous and vexatious.

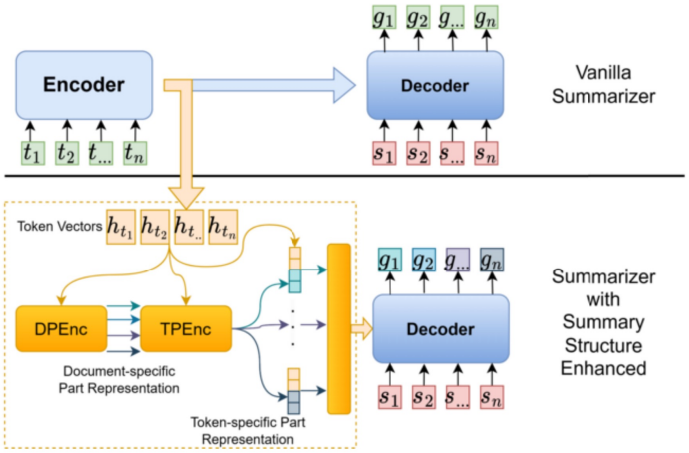
Table 2. Summary structure for Chinese judgments

Part	Detail	Sample
Type	Category of case	The case is a dispute over a loan contract.
Plaintiff's Claims	Claims from plaintiff(s)	The plaintiff filed a claim for return of the principal amount of a loan together with normal interest and penalty interest.
Defendant's Response	Response from defendant(s)	The defendant admitted that the loan was genuine and agreed to return it.
Reasons	Why make the decision	The court found: 1. there existed a the loan contractual relationship; 2. the defendant was late in returning the loan.
Decision	What the court decides	According to Section 206 and 207 of the Contract Law, the defendant is to return the principal and pay interest on the loan.



54

Architecture

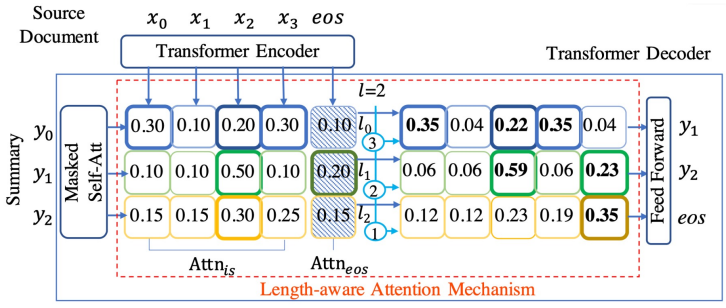


Controlling the summary length

Goal: Generate summaries of exactly a certain length, say l

What controls the end of the sequence? The </s> token

How can that be tweaked? Length-aware attention mechanism



l_t: Remaining length budget

$$l_t = \begin{cases} l+1-t, & 0 \le t \le l \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

$$a'_{t,i} = w_{t,i} \times a_{t,i}$$ attention for words

$$w_{t,i} = \begin{cases} 1, & p_i = 0 \\ l_t, & p_i = 1 \end{cases}$$ top-k words

$$a'_{t,m} = (l+1-l_t) \times a_{t,m}$$ attention for </s> token

$$p(y_t|y_{i<t}, \mathbf{x}, l) = \text{softmax}(W\mathbf{c}_{t-1} + b)$$

$$\mathbf{c_t} = \sum_0^m \tilde{a}_{t,i} h_i$$

$$\tilde{a}_{t,i} = \frac{a'_{t,i}}{\sum_{i=0}^m a'_{t,i}}$$

Summary

Part	Key Concepts	Demos
What and why	Terminology: Abstractive/Extractive; What is a good summary?	SparkNLP
Extractive summarisation	Graph-based and classification-based methods	TextRank
Abstractive summarisation	Pointer-generator networks, Denoising using encoder-decoder models (BART); windowed attention in Longformer	Longformer
Special cases of summarisation	Multi-document summarisation; domain-specific summarisation; length-specific summaries (modified attention) **(Summarisation using Reinforcement Learning)**	-

-Modifying vanilla attention for specific summarisation requirements is a typical observation in approaches for summarisation.

57

Advanced/Optional Reading

Dated but good summary of summarisation: [PDF] from nowpublishers.com

Multi-document summarisation: Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2022. Multi-document summarisation via Deep Learning Techniques: A Survey. ACM Comput. Surv. 55, 5, Article 102 (May 2023), 37 pages. https://doi.org/10.1145/3529754

Dialogue summarisation: Zhao, Lulu, et al. "Domain-Oriented Prefix-Tuning: Towards Efficient and Generalizable Fine-tuning for Zero-Shot Dialogue summarisation." *NAACL*. 2022.

58


Week 9 (Thursday): We will focus on summarization using reinforcement learning
We have now covered some typical formulations of NLP tasks.
->(Week 10: Applications & Frontiers)
UNSW SYDNEY

59