Summarisation

COMP90042

Natural Language Processing

Lecture 21

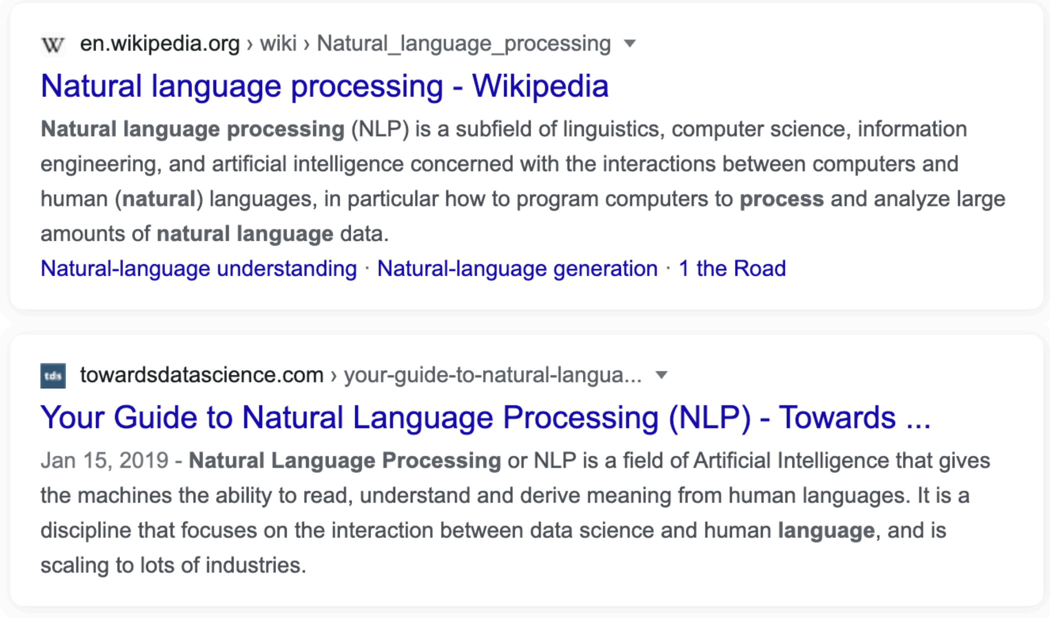
Semester 1 2022 Week 11

Jey Han Lau

THE UNIVERSITY OF MELBOURNE

Summarisation

- Distill the most important information from a text to produce shortened or abridged version
- Examples
  - **outlines** of a document
  - **abstracts** of a scientific article
  - **headlines** of a news article
  - **snippets** of search result



W en.wikipedia.org › wiki › Natural_language_processing ▾

Natural language processing - Wikipedia

Natural language processing (NLP) is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (**natural**) languages, in particular how to program computers to **process** and analyze large amounts of **natural language** data.

Natural-language understanding · Natural-language generation · 1 the Road

tds towardsdatascience.com › your-guide-to-natural-langua... ▾

Your Guide to Natural Language Processing (NLP) - Towards ...

Jan 15, 2019 - **Natural Language Processing** or NLP is a field of Artificial Intelligence that gives the machines the ability to read, understand and derive meaning from human languages. It is a discipline that focuses on the interaction between data science and human **language**, and is scaling to lots of industries.

What to Summarise?

- **Single-document summarisation**
  - Input: a single document
  - Output: summary that characterise the content
- **Multi-document summarisation**
  - Input: multiple documents
  - Output: summary that captures the gist of all documents
  - E.g. summarise a news event from multiple sources or perspectives

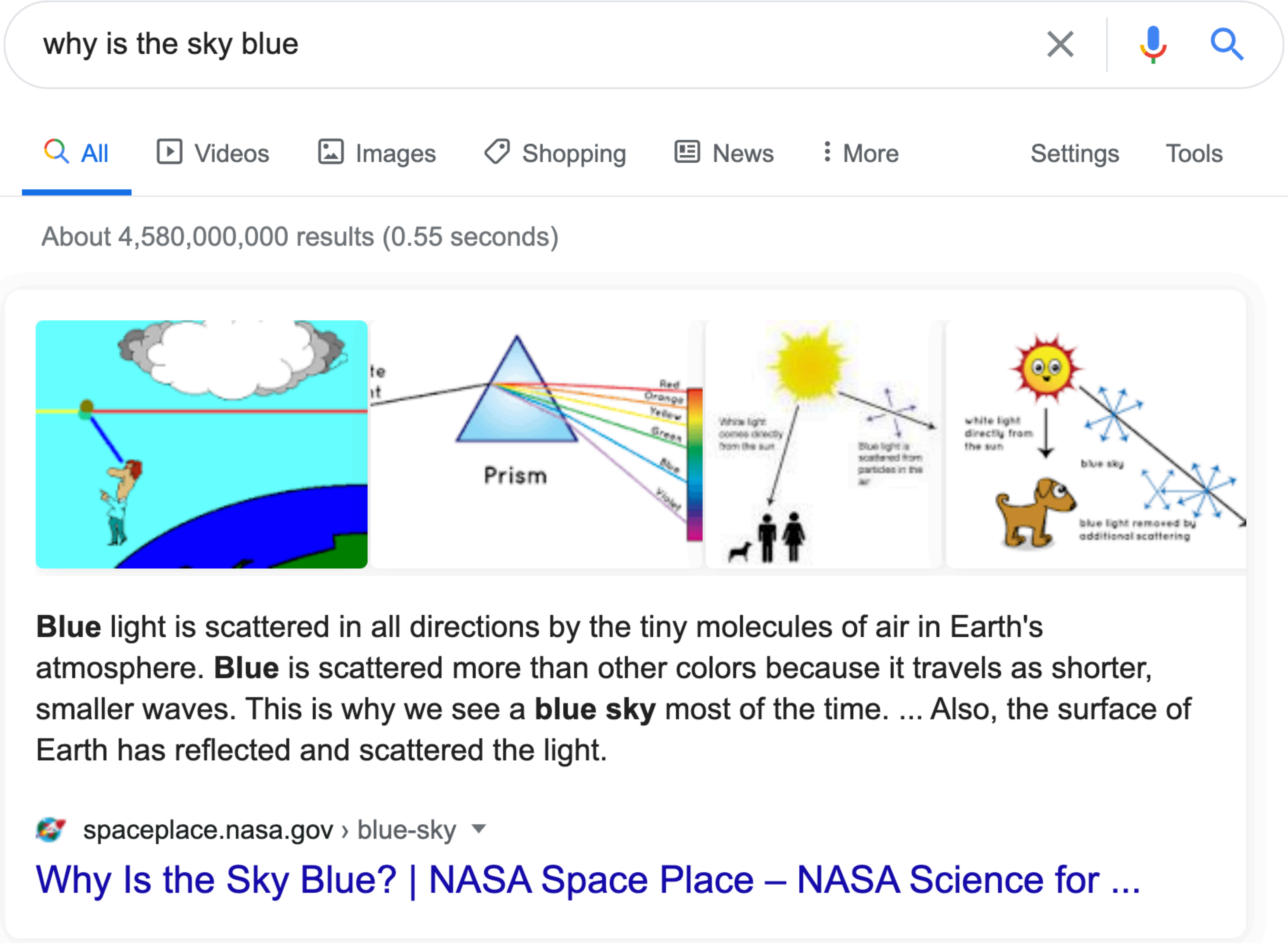
How to Summarise?

- **Extractive summarisation**
  - Summarise by selecting representative sentences from documents
- **Abstractive summarisation**
  - Summarise the content in your own words
  - Summaries will often be paraphrases of the original content

Goal of Summarisation?

- **Generic summarisation**
  - Summary gives important information in the document(s)
- **Query-focused summarisation**
  - Summary responds to a user query
  - "Non-factoid" QA
  - Answer is much longer than factoid QA

Query-Focused Summarisation



why is the sky blue

About 4,580,000,000 results (0.55 seconds)

Blue light is scattered in all directions by the tiny molecules of air in Earth's atmosphere. **Blue** is scattered more than other colors because it travels as shorter, smaller waves. This is why we see a **blue sky** most of the time. ... Also, the surface of Earth has reflected and scattered the light.

spaceplace.nasa.gov › blue-sky

Why Is the Sky Blue? | NASA Space Place – NASA Science for ...

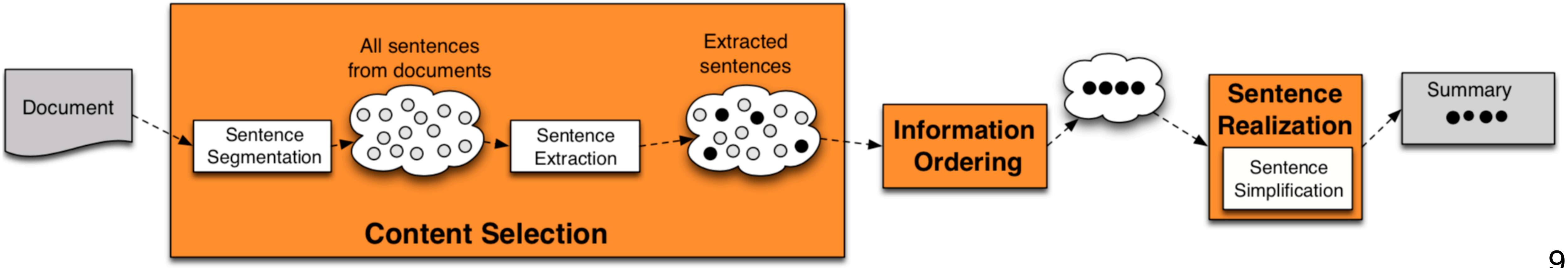
Outline

- Extractive summarisation
    - Single-document
    - Multi-document
- Abstractive summarisation
    - Single-document (deep learning models!)
- Evaluation

Extractive: Single-Doc

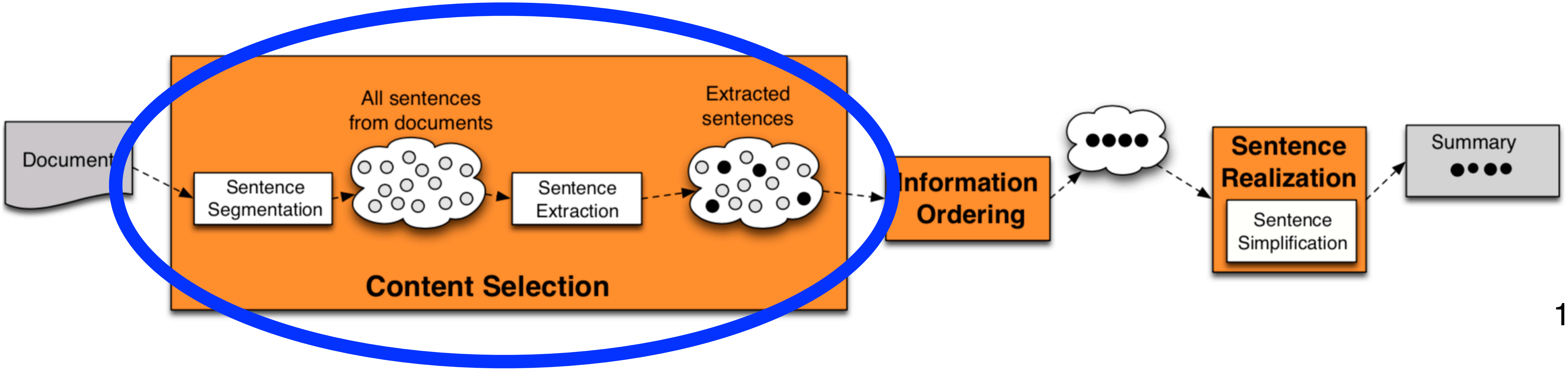
Summarisation System

- **Content selection:** select what sentences to extract from the document
- **Information ordering:** decide how to order extracted sentences
- **Sentence realisation:** cleanup to make sure combined sentences are fluent



Summarisation System

- We will focus on **content selection**
- For single-document summarisation, information ordering not necessary
  - present extracted sentences in original order
- Sentence realisation also not necessary if they are presented in dot points



Content Selection

- Not much data with ground truth extractive sentences
- Mostly unsupervised methods
- **Goal:** Find sentences that are important or **salient**

Method 1: TF-IDF

- Frequent words in a doc → salient
- But some generic words are very frequent but uninformative
    - function words
    - stop words
- Weigh each word w in document d by its inverse document frequency:
    - weight$(w) = tf_{d,w} \times idf_w$

Method 2: Log Likelihood Ratio

- Intuition: a word is salient if its probability in the **input corpus** is very different to a **background corpus**

$$\binom{N_I}{x} p^x (1-p)^{N_I - x}$$

- weight$(w) = \begin{cases} 1, & \text{if } -2 log \lambda(w) > 10 \\ 0, & \text{otherwise} \end{cases}$

$$\binom{N_B}{y} p^y (1-p)^{N_B - y}$$

- $\lambda(w)$ is the ratio between:
  - P(observing w in I) and P(observing w in B), assuming $P(w \mid I) = P(w \mid B) = p$

$$\frac{x+y}{N_I + N_B}$$

  - P(observing w in I) and P(observing w in B), assuming $P(w \mid I) = p_I$ and $P(w \mid B) = p_B$

$$\binom{N_I}{x} p_I^x (1-p_I)^{N_I - x} \qquad \frac{x}{N_I} \qquad \frac{y}{N_B} \qquad \binom{N_B}{y} p_B^y (1-p_B)^{N_B - y}$$

Saliency of A Sentence?

- $$\text{weight}(s) = \frac{1}{|S|} \sum_{w \in S} \text{weight}(w)$$
- Only consider non-stop words in S

Method 3: Sentence Centrality

- Alternative approach to ranking sentences
- Measure distance between sentences, and choose sentences that are closer to other sentences
- Use tf-idf BOW to represent sentence
- Use cosine similarity to measure distance
- $$\text{centrality}(s) = \frac{1}{\#\text{sent}} \sum_{s'} \cos_{tfidf}(s, s')$$

Final Extracted Summary

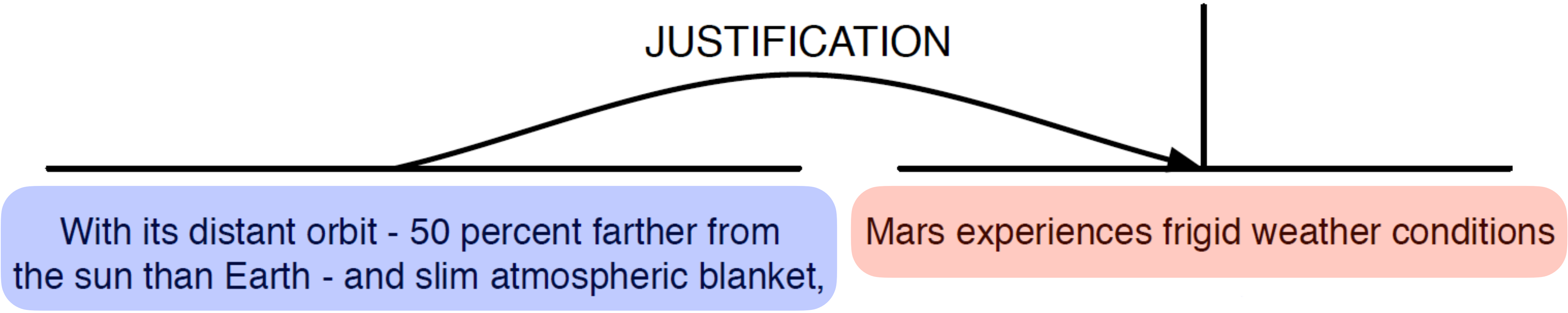
- Use top-ranked sentences as extracted summary
    - Saliency (tf-idf or log likelihood ratio)
    - Centrality

Method 4: RST Parsing

With its distant orbit – 50 percent farther from the sun than Earth – and slim atmospheric blanket, Mars experiences frigid weather conditions. Surface temperatures typically average about -70 degrees Fahrenheit at the equator, and can dip to -123 degrees C near the poles. Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion, but any liquid water formed in this way would evaporate almost instantly because of the low atmospheric pressure. Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop, most Martian weather involves blowing dust or carbon dioxide.

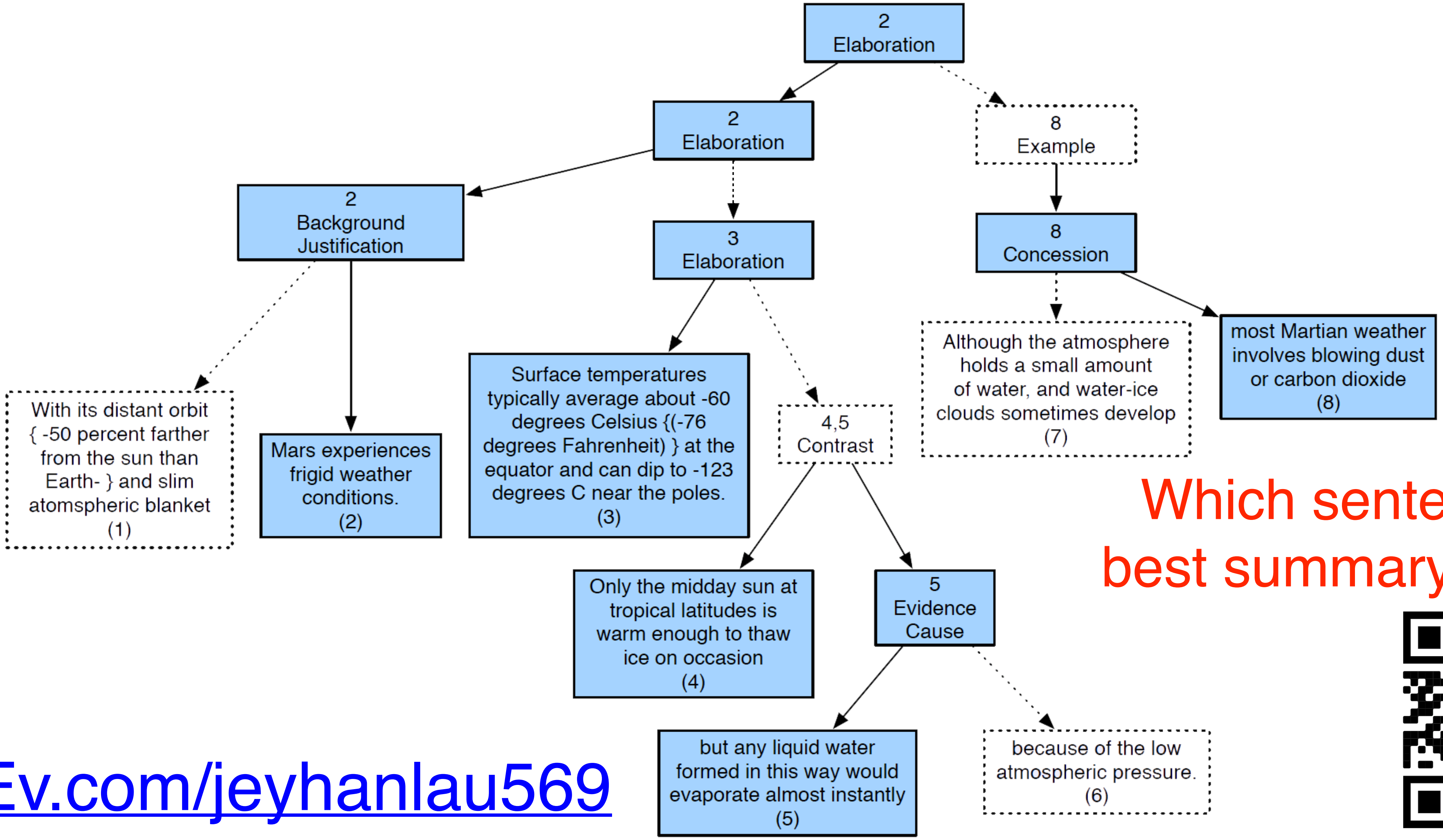
Method 4: RST Parsing

- Rhetorical structure theory (L12, Discourse): explain how clauses are connected
- Define the types of relations between a **nucleus** (main clause) and a **satellite** (supporting clause)


JUSTIFICATION
With its distant orbit - 50 percent farther from the sun than Earth - and slim atmospheric blanket,
Mars experiences frigid weather conditions

Method 4: RST Parsing

- Nucleus more important than satellite
- A sentence that functions as a nucleus to more sentences = more salient


2 Elaboration
2 Elaboration
8 Example
2 Background Justification
3 Elaboration
8 Concession
With its distant orbit { -50 percent farther from the sun than Earth- } and slim atomspheric blanket (1)
Mars experiences frigid weather conditions. (2)
Surface temperatures typically average about -60 degrees Celsius {(-76 degrees Fahrenheit) } at the equator and can dip to -123 degrees C near the poles. (3)
4,5 Contrast
Although the atmosphere holds a small amount of water, and water-ice clouds sometimes develop (7)
most Martian weather involves blowing dust or carbon dioxide (8)
Only the midday sun at tropical latitudes is warm enough to thaw ice on occasion (4)
5 Evidence Cause
but any liquid water formed in this way would evaporate almost instantly (5)
because of the low atmospheric pressure. (6)

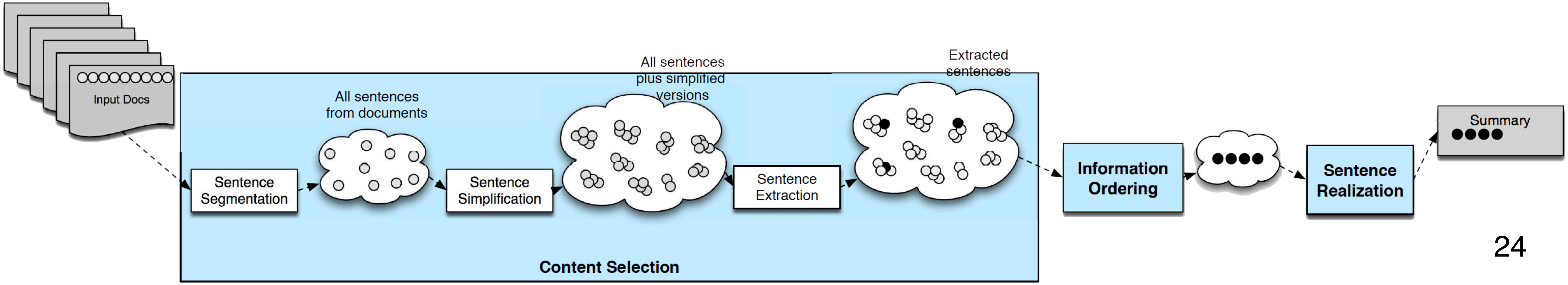
Which sentence is the best summary sentence?

PollEv.com/jeyhanlau569

Extractive: Multi-Doc

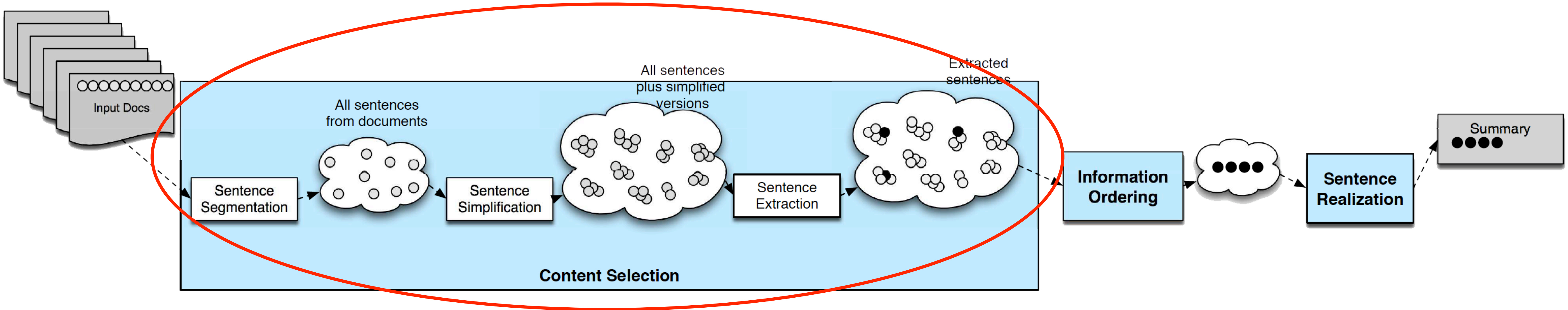
Summarisation System

- Similar to single-document extractive summarisation system
- Challenges:
  - Redundancy in terms of information
  - Sentence ordering



Content Selection

- We can use the same unsupervised content selection methods (tf-idf, log likelihood ratio, centrality) to select **salient sentences**
- But ignore sentences that are **redundant**



Maximum Marginal Relevance

- Iteratively select the best sentence to add to summary

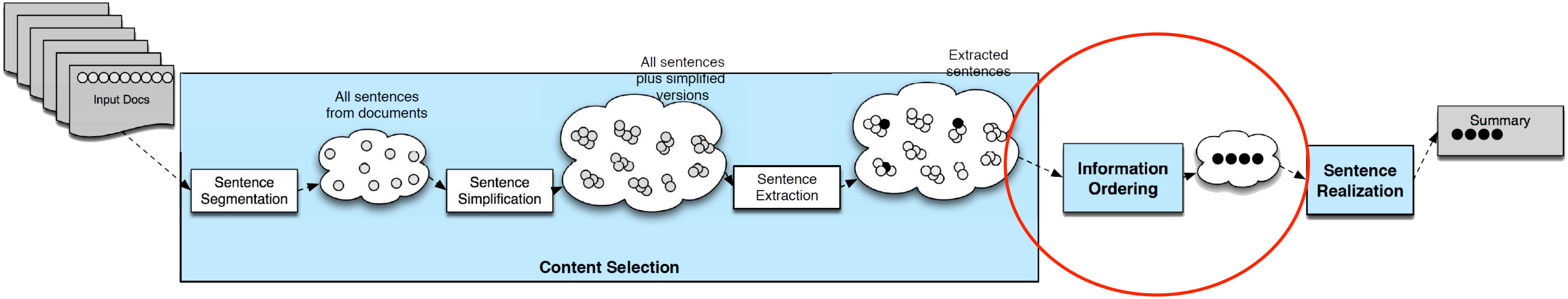
- Sentences to be added must be **novel**

- Penalise a candidate sentence if it's similar to extracted sentences:
  - MMR-penalty$(s) = \lambda \max_{s_i \in \mathcal{S}} sim(s, s_i)$

- Stop when a desired number of sentences are added

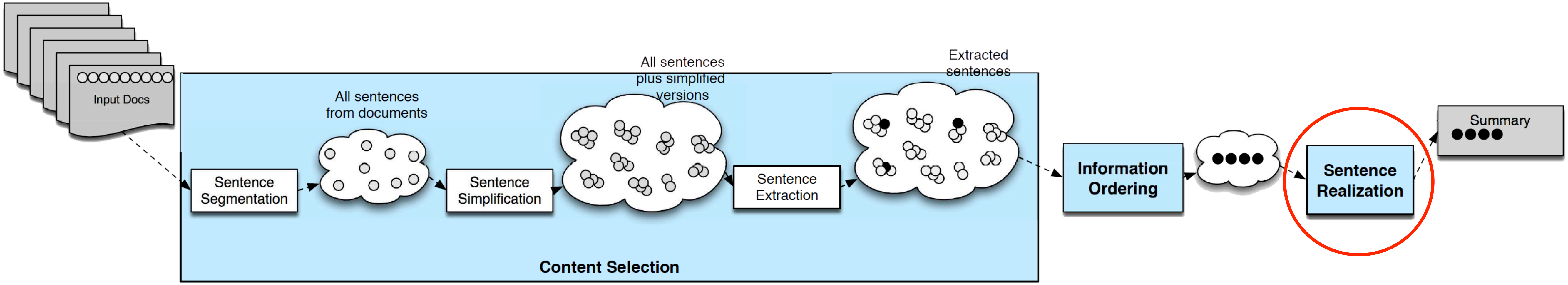
Information Ordering

- **Chronological ordering:**
  - Order by document dates
- **Coherence:**
  - Order in a way that makes adjacent sentences similar
  - Order based on how entities are organised (centering theory, L12)

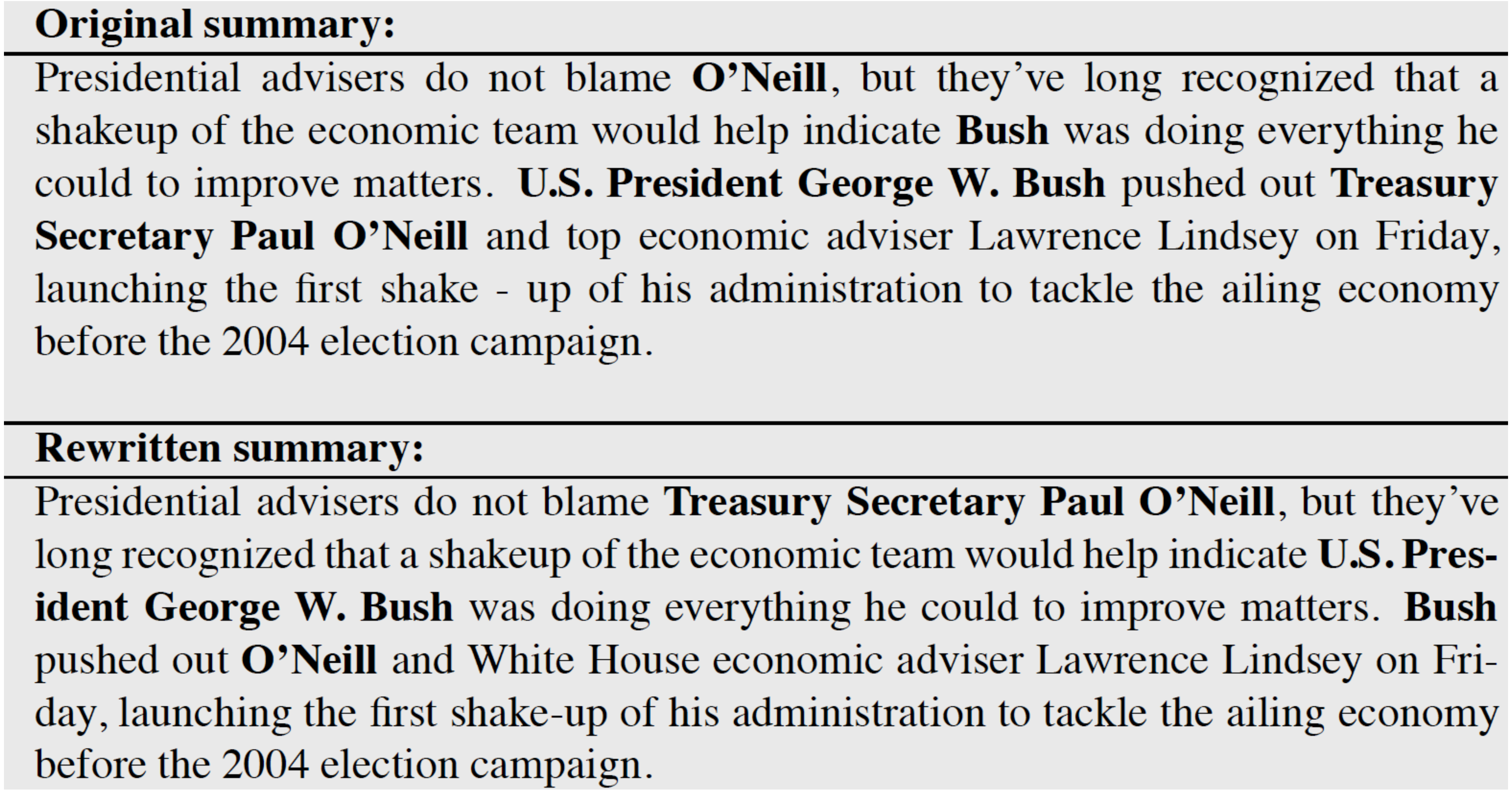


Sentence Realisation

- Make sure entities are referred coherently
    - Full name at first mention
    - Last name at subsequent mentions
- Apply coreference methods to first extract names
- Write rules to clean up



Sentence Realisation

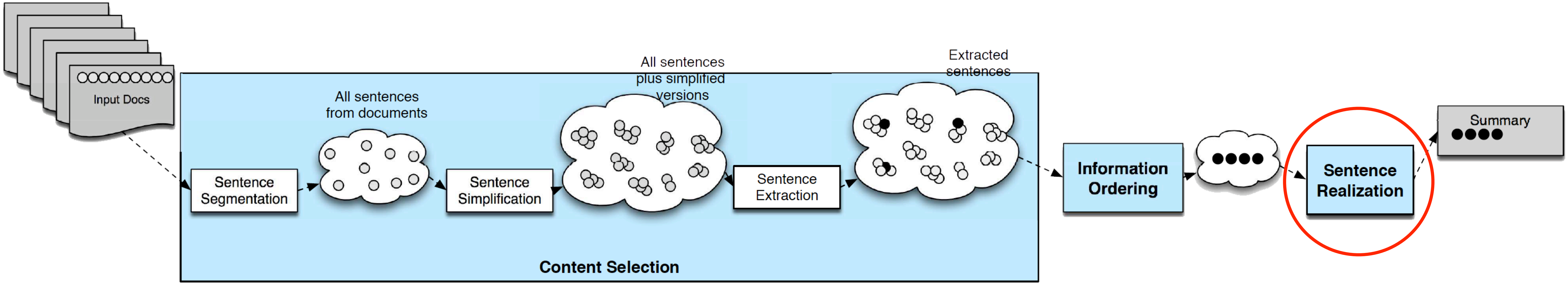


Original summary:

Presidential advisers do not blame **O’Neill**, but they’ve long recognized that a shakeup of the economic team would help indicate **Bush** was doing everything he could to improve matters. **U.S. President George W. Bush** pushed out **Treasury Secretary Paul O’Neill** and top economic adviser Lawrence Lindsey on Friday, launching the first shake - up of his administration to tackle the ailing economy before the 2004 election campaign.

Rewritten summary:

Presidential advisers do not blame **Treasury Secretary Paul O’Neill**, but they’ve long recognized that a shakeup of the economic team would help indicate **U.S. President George W. Bush** was doing everything he could to improve matters. **Bush** pushed out **O’Neill** and White House economic adviser Lawrence Lindsey on Friday, launching the first shake-up of his administration to tackle the ailing economy before the 2004 election campaign.



Abstractive: Single-Doc

Example

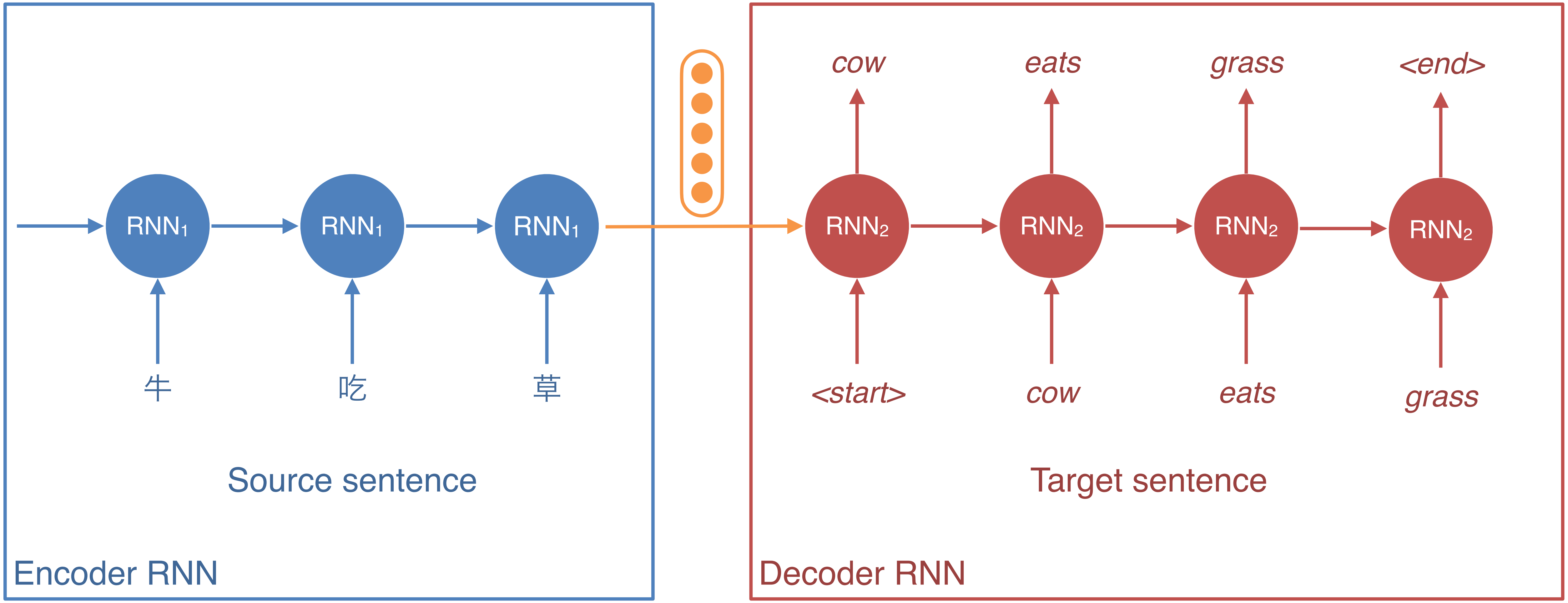
*a detained **iranian-american academic** accused of acting against national security has been **released** from a **tehran** prison after a hefty **bail** was posted, a top judiciary official said tuesday*

↓

iranian-american academic** held in **tehran released** on **bail

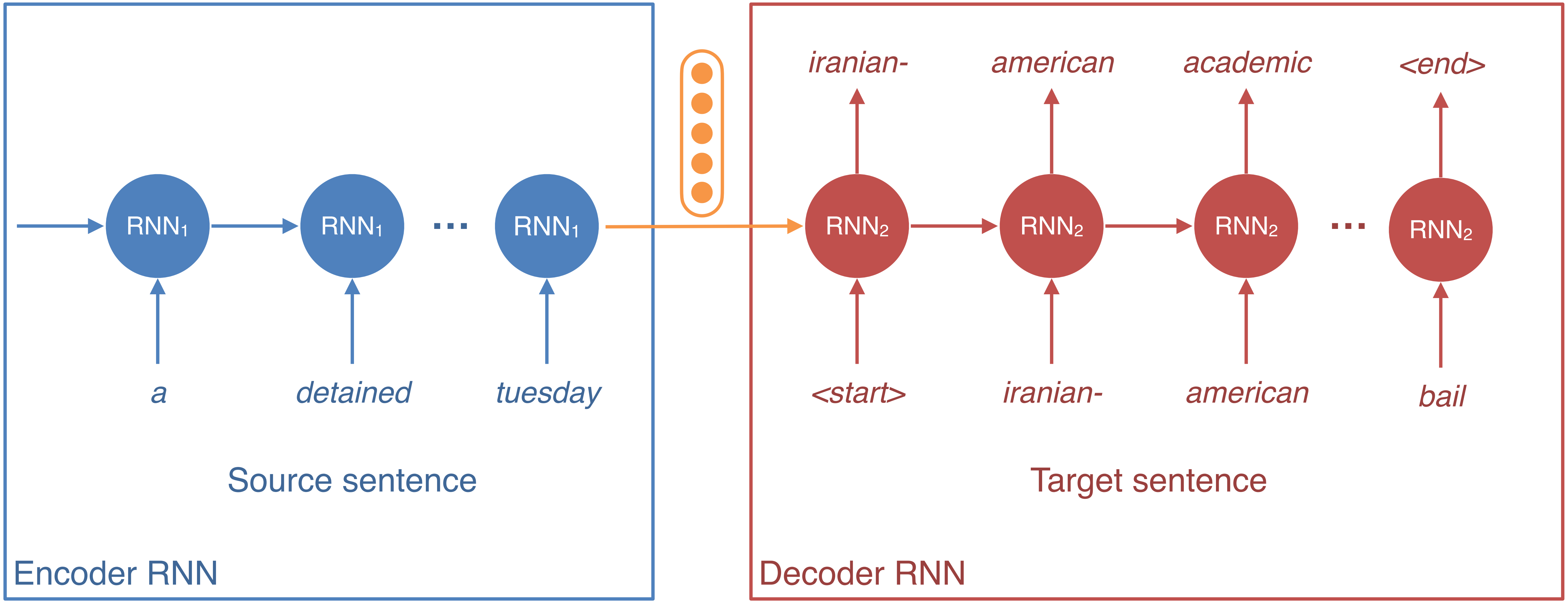
- Paraphrase
- A very difficult task
- Can we train a neural network to generate summary?

Encoder-Decoder?



- What if we treat:
  - Source sentence = “document”
  - Target sentence = “summary”

Encoder-Decoder?



*a detained **iranian-american academic** accused of acting against national security has been **released** from a **tehran** prison after a hefty **bail** was posted, a top judiciary official said tuesday*

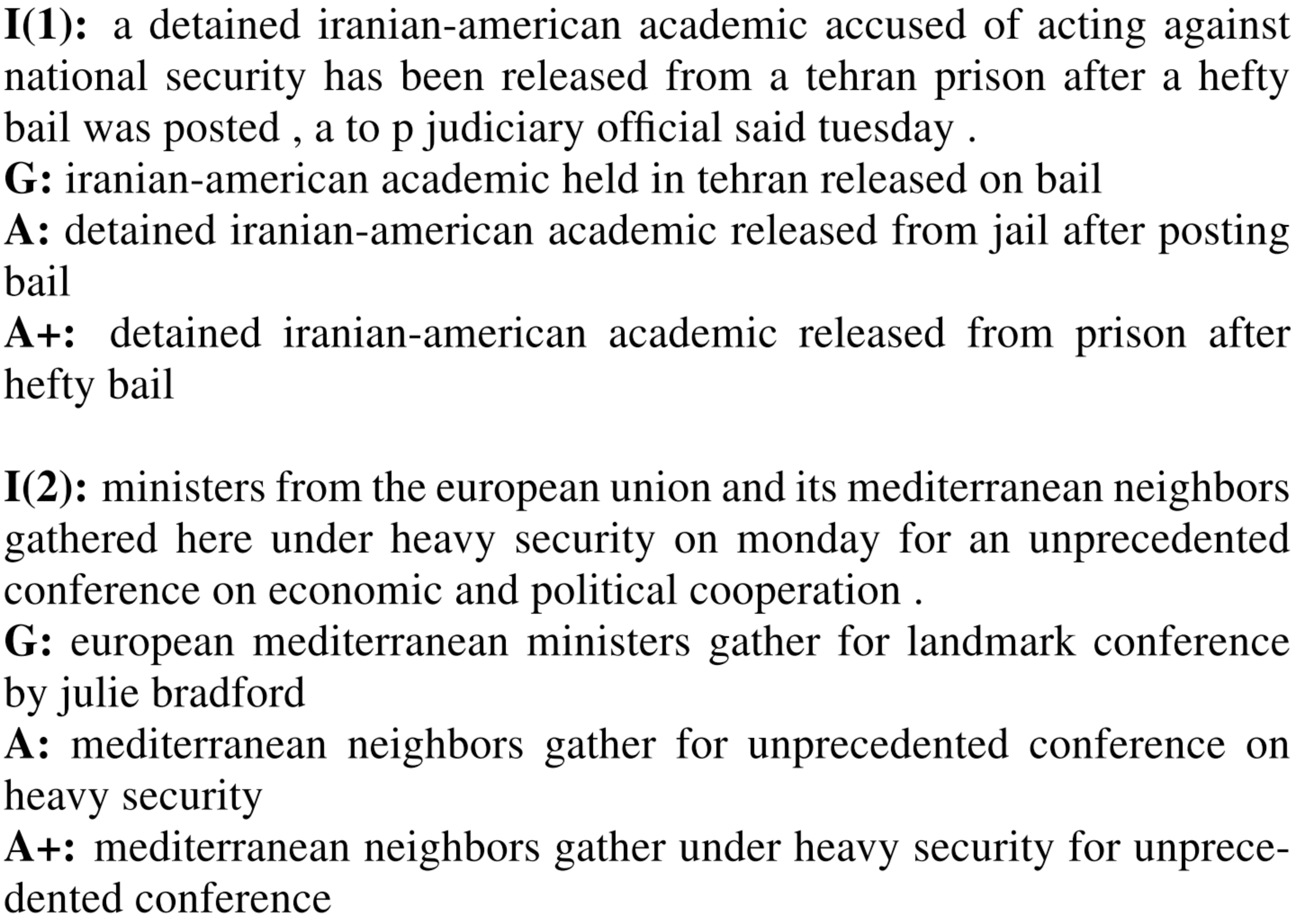
↓

iranian-american academic** held in **tehran released** on **bail

Data

- News headlines
- Document: First sentence of article
- Summary: News headline/title
- Technically more like a “headline generation task”

And It Kind of Works…



I(1): a detained iranian-american academic accused of acting against national security has been released from a tehran prison after a hefty bail was posted , a to p judiciary official said tuesday .
G: iranian-american academic held in tehran released on bail
A: detained iranian-american academic released from jail after posting bail
A+: detained iranian-american academic released from prison after hefty bail

I(2): ministers from the european union and its mediterranean neighbors gathered here under heavy security on monday for an unprecedented conference on economic and political cooperation .
G: european mediterranean ministers gather for landmark conference by julie bradford
A: mediterranean neighbors gather for unprecedented conference on heavy security
A+: mediterranean neighbors gather under heavy security for unprecedented conference

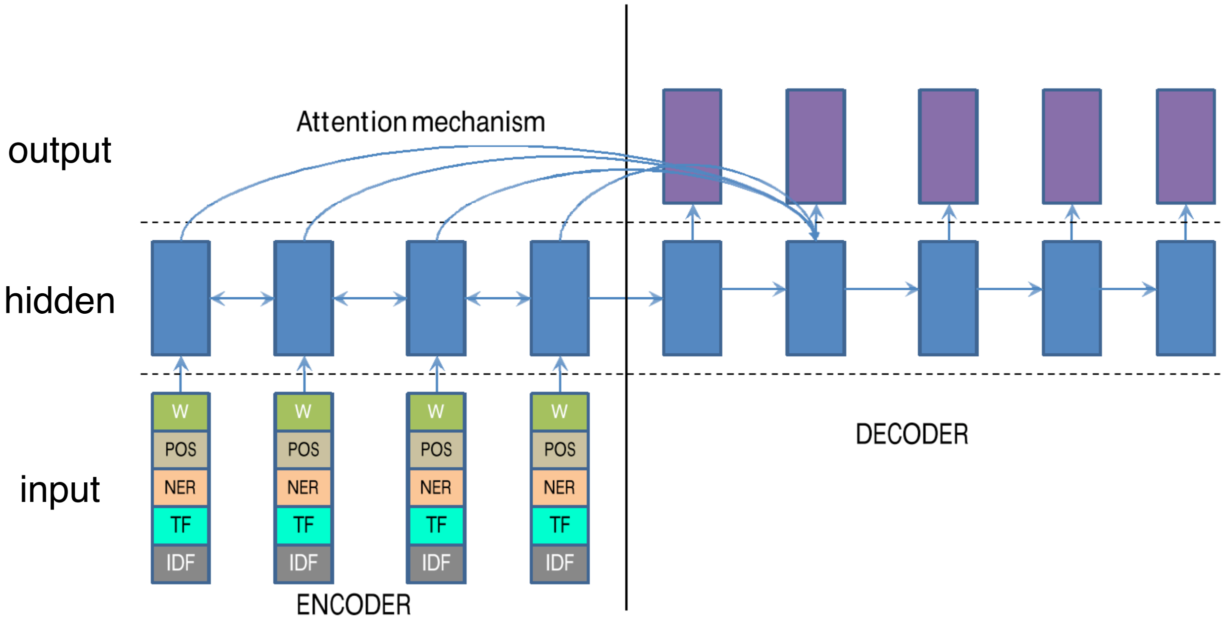
Rush et al. (2015): A Neural Attention Model for Abstractive Sentence Summarisation

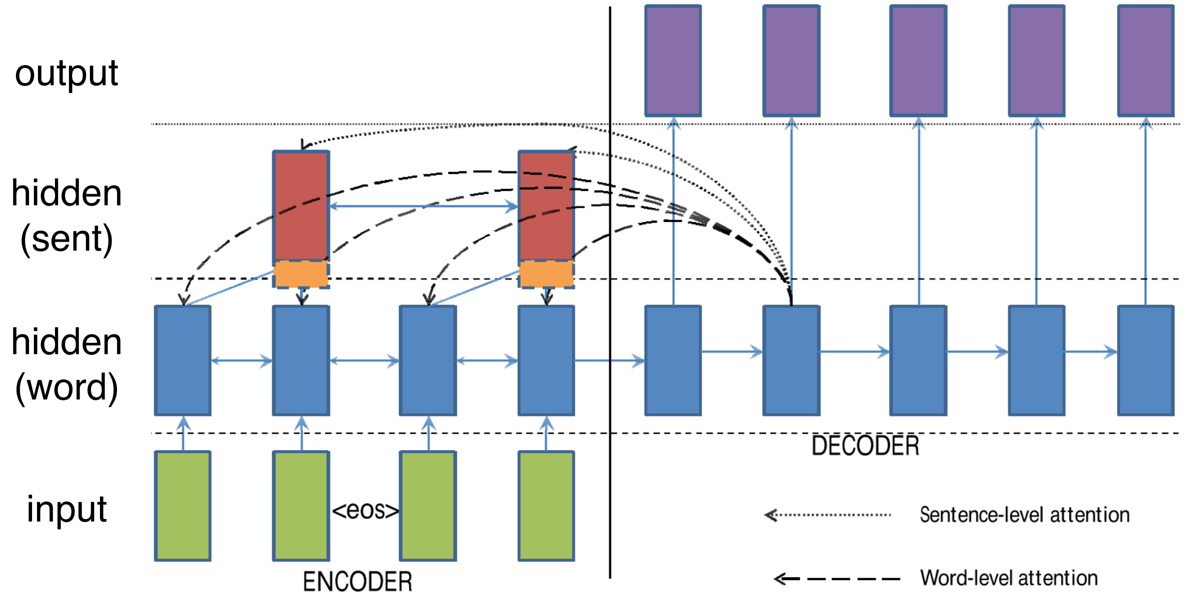
More Summarisation Data

- But headline generation isn't really exciting…
- Other summarisation data:
  - **CNN/Dailymail:** 300K articles, summary in bullets
  - **Newsroom:** 1.3M articles, summary by authors
    - Diverse; 38 major publications
  - **XSum:** 200K BBC articles
    - Summary is more abstractive than other datasets

Improvements

- Attention mechanism
- Richer word features: POS tags, NER tags, tf-idf
- Hierarchical encoders
  - One LSTM for words
  - Another LSTM for sentences



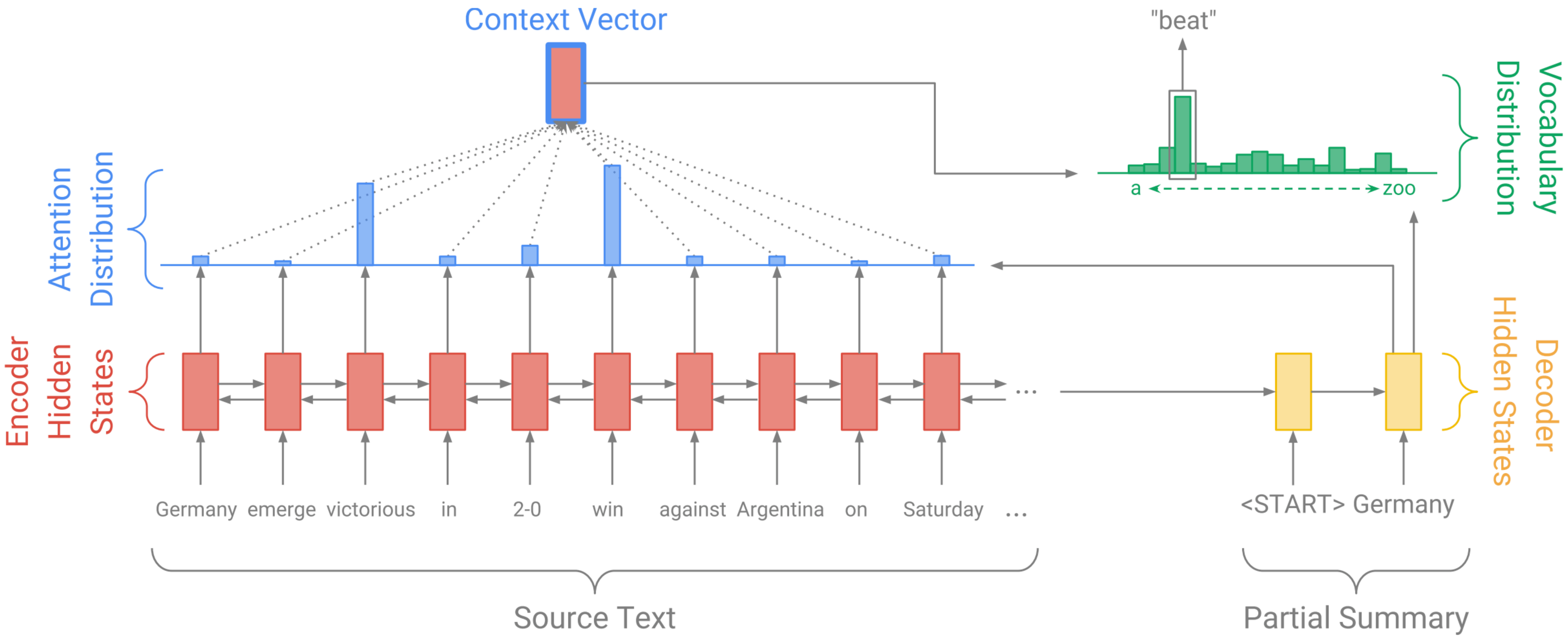


Nallapati et al. (2016): Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

Potential issues of an attention encoder-decoder summarisation system?

- Has the potential to generate new details not in the source document
- Unable to handle unseen words in the source document
- Information bottleneck: a vector is used to represent the source document
- Can only generate one summary

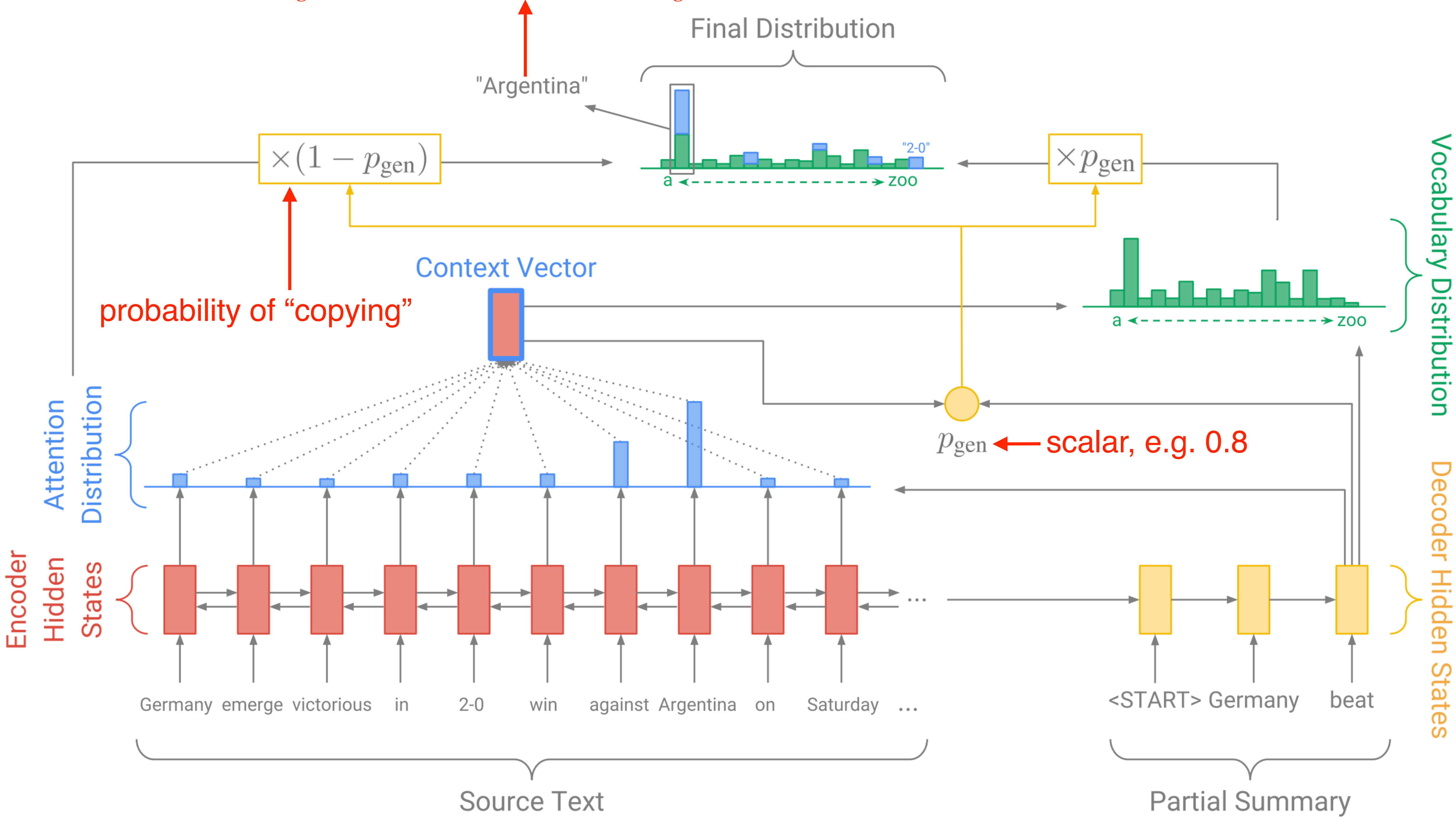
PollEv.com/jeyhanlau569



Encoder-decoder with Attention

See et al. (2017): Get To The Point: Summarization with Pointer-Generator Networks

$$P(\text{Argentina}) = (1 - p_{gen}) \times P_{attn}(\text{Argentina}) + p_{gen} \times P_{voc}(\text{Argentina})$$



Encoder-decoder with Attention + Copying

See et al. (2017): Get To The Point: Summarization with Pointer-Generator Networks

Copy Mechanism

- Generate summaries that reproduce details in the document
- Can produce out-of-vocab words in the summary by copying them in the document
  - e.g. *smergle* = out of vocabulary
  - p(*smergle*) = attention probability + generation probability = attention probability

Latest Development

- State-of-the-art models use transformers instead of RNNs
- Lots of pre-training
- Note: BERT not directly applicable because we need a unidirectional decoder (BERT is only an encoder)

Evaluation

ROUGE

(Recall Oriented Understudy for Gisting Evaluation)

- Similar to BLEU, evaluates the degree of word overlap between **generated summary** and **reference/human summary**
- But recall oriented
- Measures overlap in *N*-grams separately (e.g. from 1 to 3)
- ROUGE-2: calculates the percentage of bigrams from the reference that are in the generated summary

ROUGE-2: Example

$$\text{ROUGE-2} = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{bigram \in S} \text{Count}_{\text{match}}(bigram)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{bigram \in S} \text{Count}(bigram)}$$

- **Ref 1:** *Water spinach is a green leafy vegetable grown in the tropics.*
- **Ref 2:** *Water spinach is a commonly eaten leaf vegetable of Asia.*
- **Generated summary:** *Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.*
- *ROUGE-2* $= \frac{3+6}{10+9}$

A Final Word

- Research focus on single-document abstractive summarisation
  - Mostly news data
- But many types of data for summarisation:
  - Images, videos
  - Graphs
  - Structured data: e.g. patient records, tables
- Multi-document abstractive summarisation