Summarisation

COMP90042

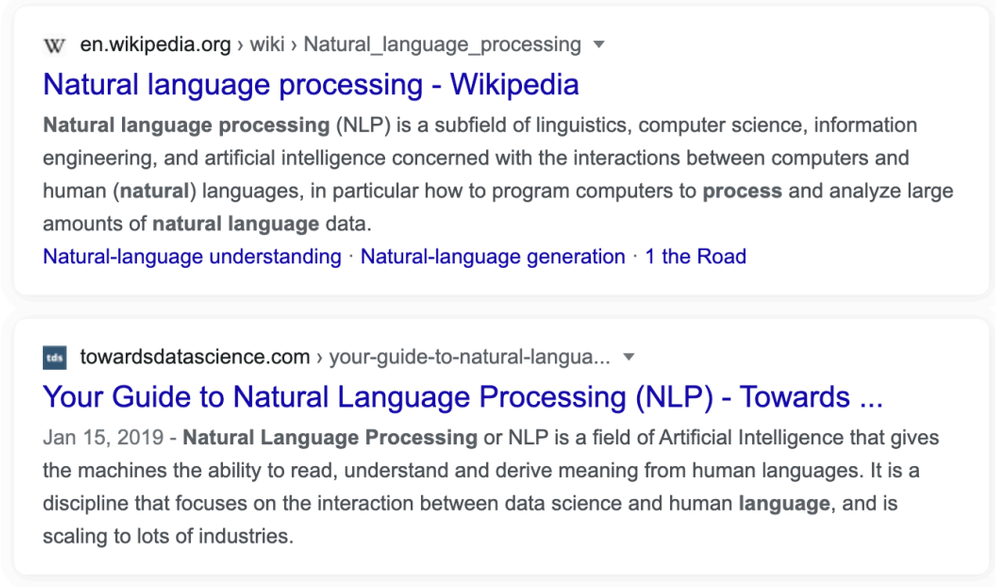
Natural Language Processing

Lecture 21

THE UNIVERSITY OF MELBOURNE

Summarisation

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



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

Fourscore and seven years ago our fathers brought forth on this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle- field of that war. We have come to dedicate a portion of that field as a final resting-place for those who here gave their lives that this nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we cannot dedicate...we cannot consecrate...we cannot hallow... this ground. The brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. The world will little note nor long remember what we say here, but it can never forget what they did here. It is for us, the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us...that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion; that we here highly resolve that these dead shall not have died in vain; that this nation, under God, shall have a new birth of freedom; and that government of the people, by the people, for the people, shall not perish from the earth.

Extractive

Extract from the Gettysburg Address:

Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation can long endure. We are met on a great battlefield of that war. We have come to dedicate a portion of that field. But the brave men, living and dead, who struggled here, have consecrated it far above our poor power to add or detract. From these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion — that government of the people, by the people for the people shall not perish from the earth.

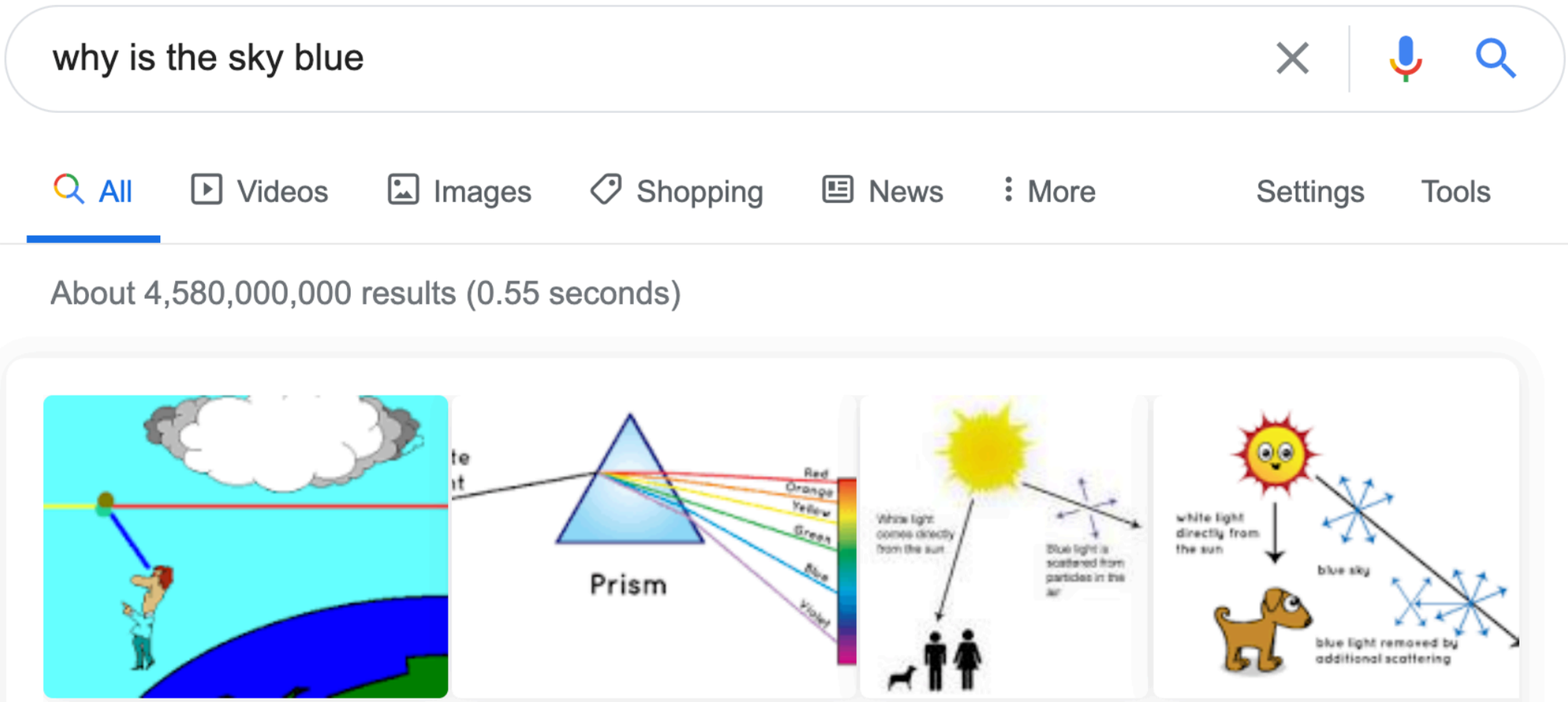
Abstract of the Gettysburg Address:

This speech by Abraham Lincoln commemorates soldiers who laid down their lives in the Battle of Gettysburg. It reminds the troops that it is the future of freedom in America that they are fighting for.

Why Summarise?

- **Generic summarisation**
  - Summary gives important information in the document(s)
- **Query-focused summarisation**
  - Summary responds to a user query
  - Similar to question answering
  - But answer is much longer (not just a phrase)

Query-Focused Summarisation



long answers

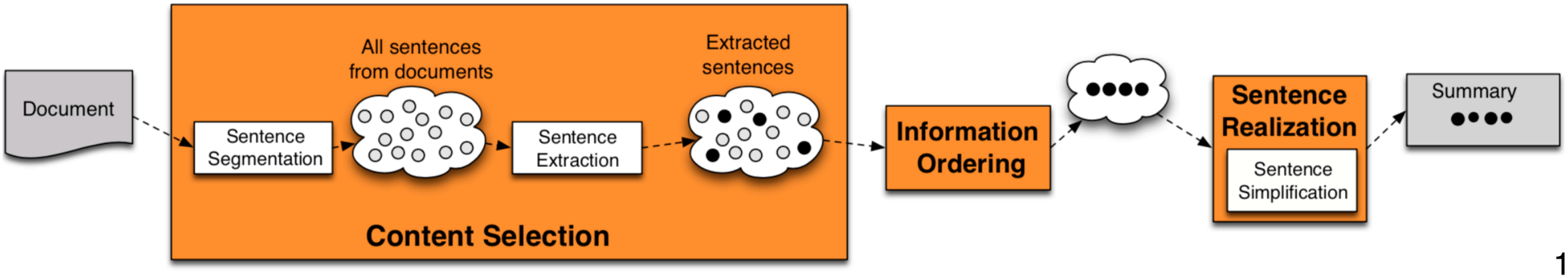
Outline

- 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 reorder sentences
- **Sentence realisation:** cleanup to make sure combined sentences are fluent § sentence simplification NER cleanup.



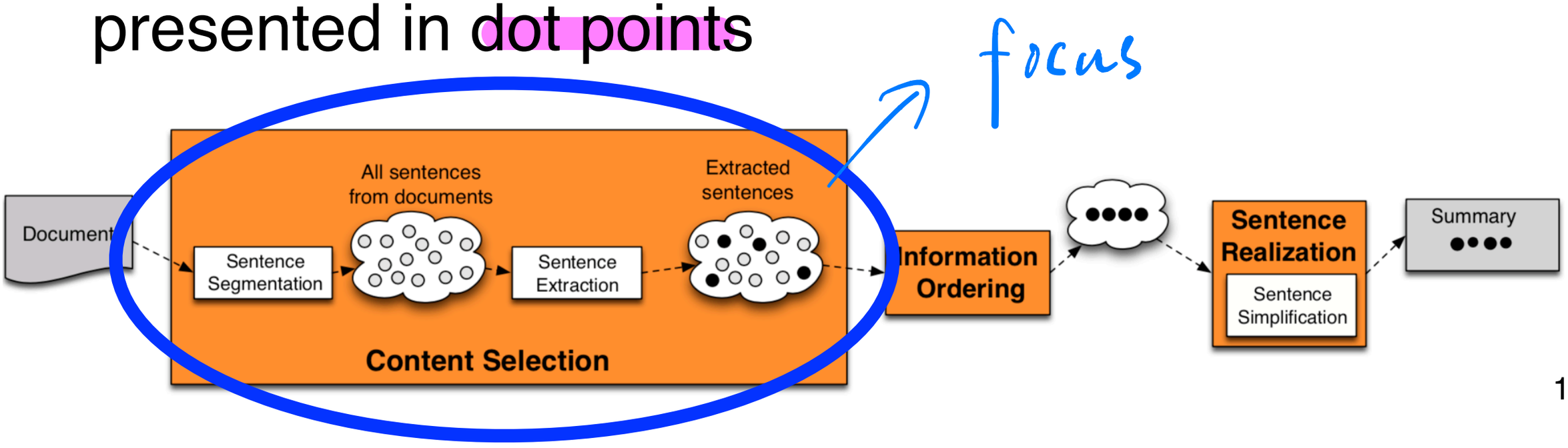
Summarisation System

- We will focus on **content selection**
- For single-document summarisation, information ordering not necessary

Not much need. enough

  - 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:
  - $\text{weight}(w) = tf_{d,w} \times idf_w$

penalize the words appear in many documents!

term frequency

inversed document frequency.

Method 2: Log Likelihood Ratio

How many times you see w N trials given p_I.

- Intuition: a word is salient if its probability in the **input corpus** is very different to a **background corpus**
- weight$(w) = \begin{cases} 1, & \text{if } -2log\lambda(w) > 10 \\ 0, & \text{otherwise} \end{cases}$

Binomial Distribution

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

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

- $\lambda(w)$ is the ratio between:
  - P(observing w in I) and P(observing w in B), assuming $P(w|I) = P(w|B) = p$ ← $\frac{x+y}{N_I + N_B}$
  - P(observing w in I) and P(observing w in B), assuming $P(w|I) = p_I$ and $P(w|B) = p_B$

Input corpus. background corpus → all occurrance.

separately

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

Saliency of A Sentence?

Average of weight over all words

- $$\text{weight}(s) = \frac{1}{|S|} \sum_{w \in S} \text{weight}(w)$$
- 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

more centroid

- Use tf-idf to represent sentence
- Use cosine similarity to measure distance
- $$\text{centrality}(s) = \frac{1}{\#\text{sent}} \sum_{s'} \cos_{tfidf}(s, s')$$

tfidf

Average cos sim with other sentences

Final Extracted Summary

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

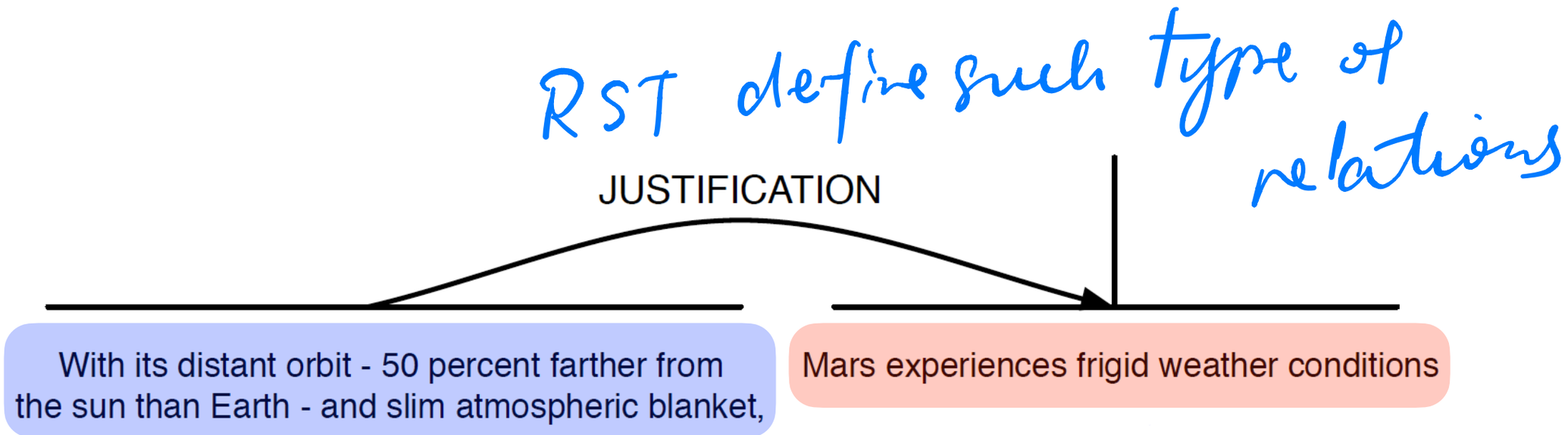
simple rank.

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)

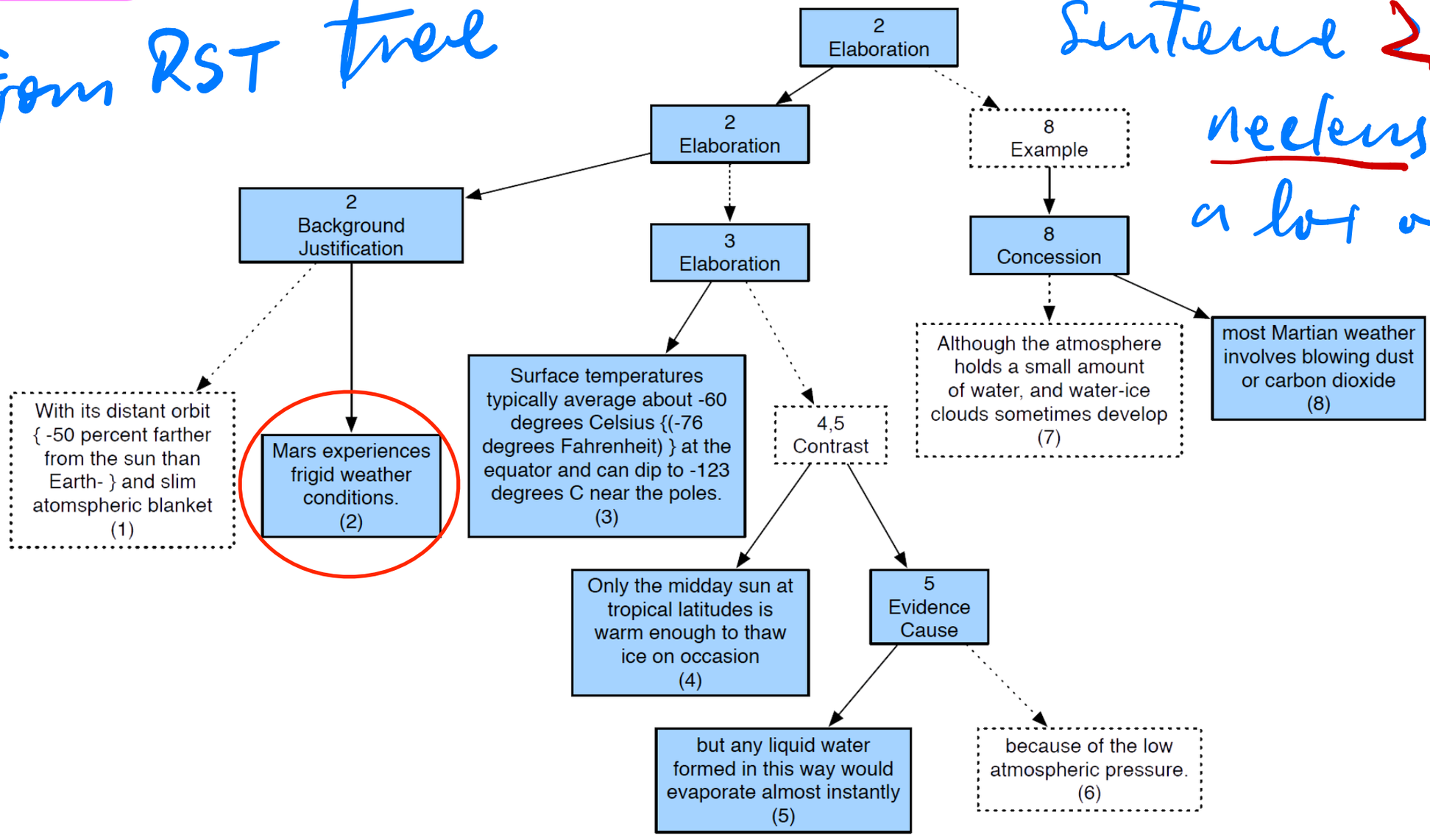


Method 4: RST Parsing

main point

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

From RST tree



Sentence 2 is nucleus for a lot of sentences

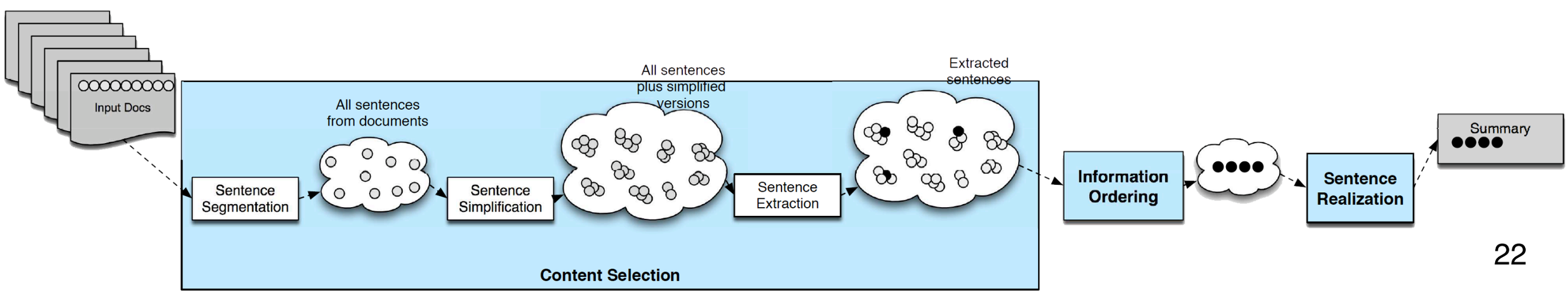
Extractive: Multi-Doc

Summarisation System

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

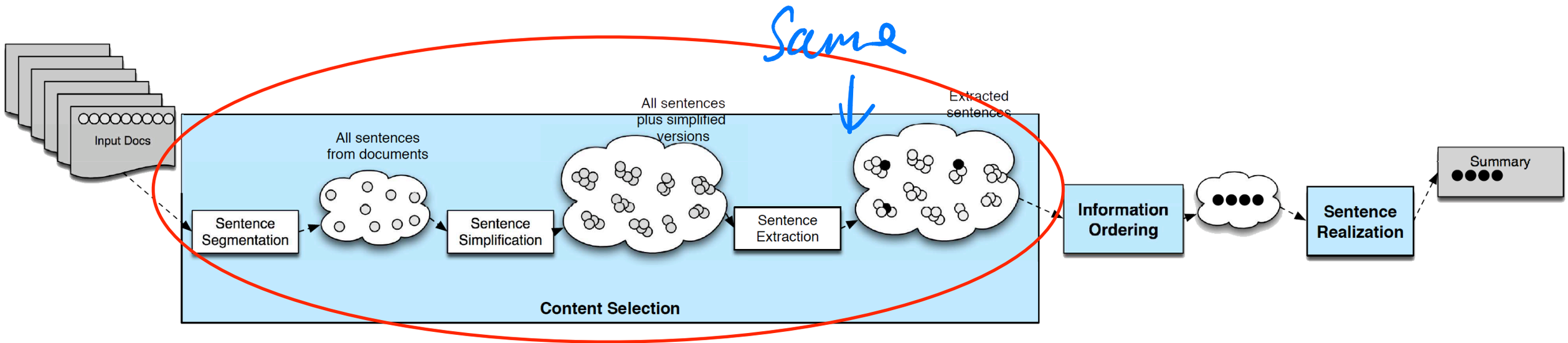
say same things for two salient sentences

not a problem for single document sum.



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**

make sure salient sentences is unique.

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

all selected S. already extracted

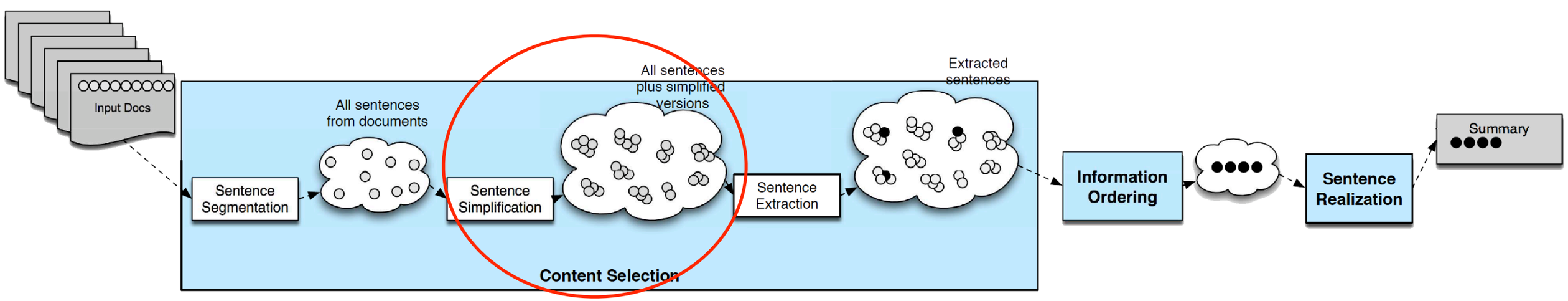
- Stop when a desired number of sentences are added

Sentence Simplification

- Create multiple simplified versions of sentences before extraction
- *Former Democratic National Committee finance director Richard Sullivan faced more pointed questioning from Republicans during his second day on the witness stand in the Senate's fund-raising investigation*
  - *Richard Sullivan faced pointed questioning*
  - *Richard Sullivan faced pointed questioning from Republicans during day on stand in Senate fundraising investigation*

parse three sentences
pick one by MMR.

- MMR to make sure only non-redundant sentences are selected



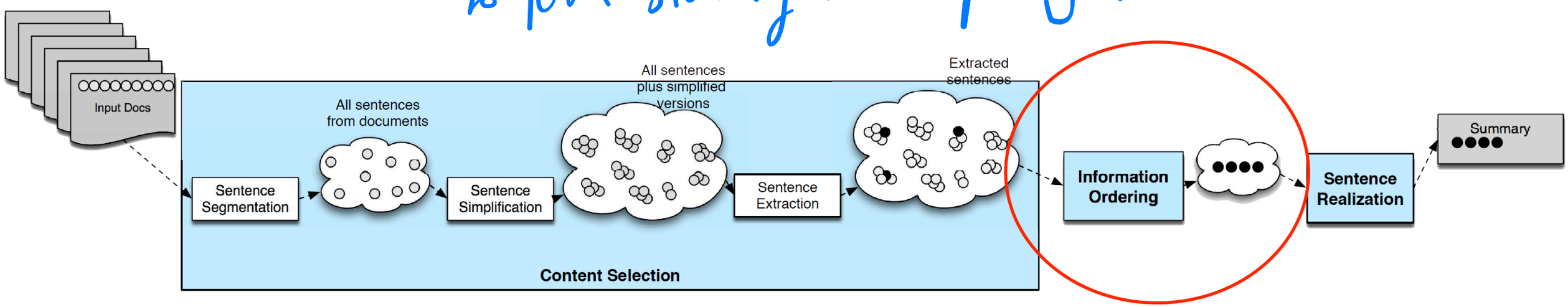
Information Ordering

- **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)

high sim

make permutations . pick one with

focus on entity of previous sentence before slowly transfering to other sentences.



Sentence Realisation

Original summary:

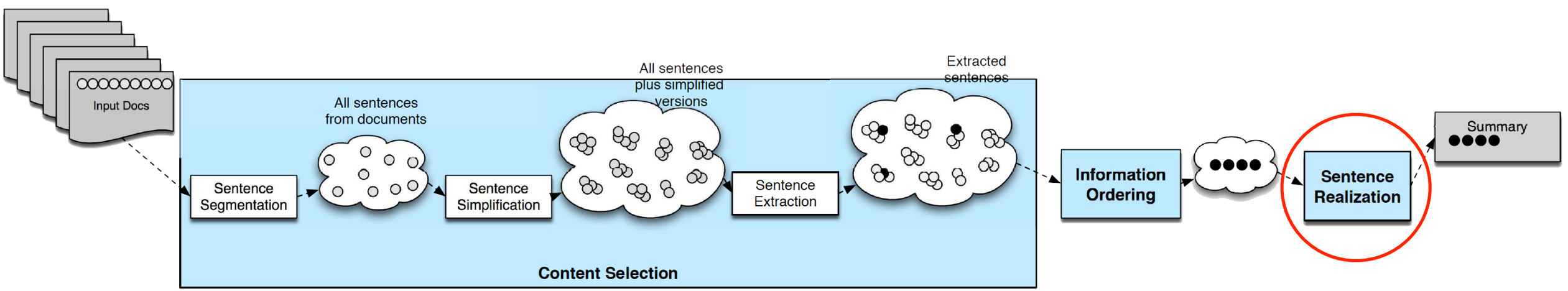
Presidential advisers do not blame **O'Neill**, but they've long recognized that a shakeup of the economic team would help indicate **Bush** was doing everything he could to improve matters. **U.S. President George W. Bush** pushed out **Treasury Secretary Paul O'Neill** and top economic adviser Lawrence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

Rule based cleanup

Rewrite long names first then short name after

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.



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

Abstractive: Single-Doc

Example

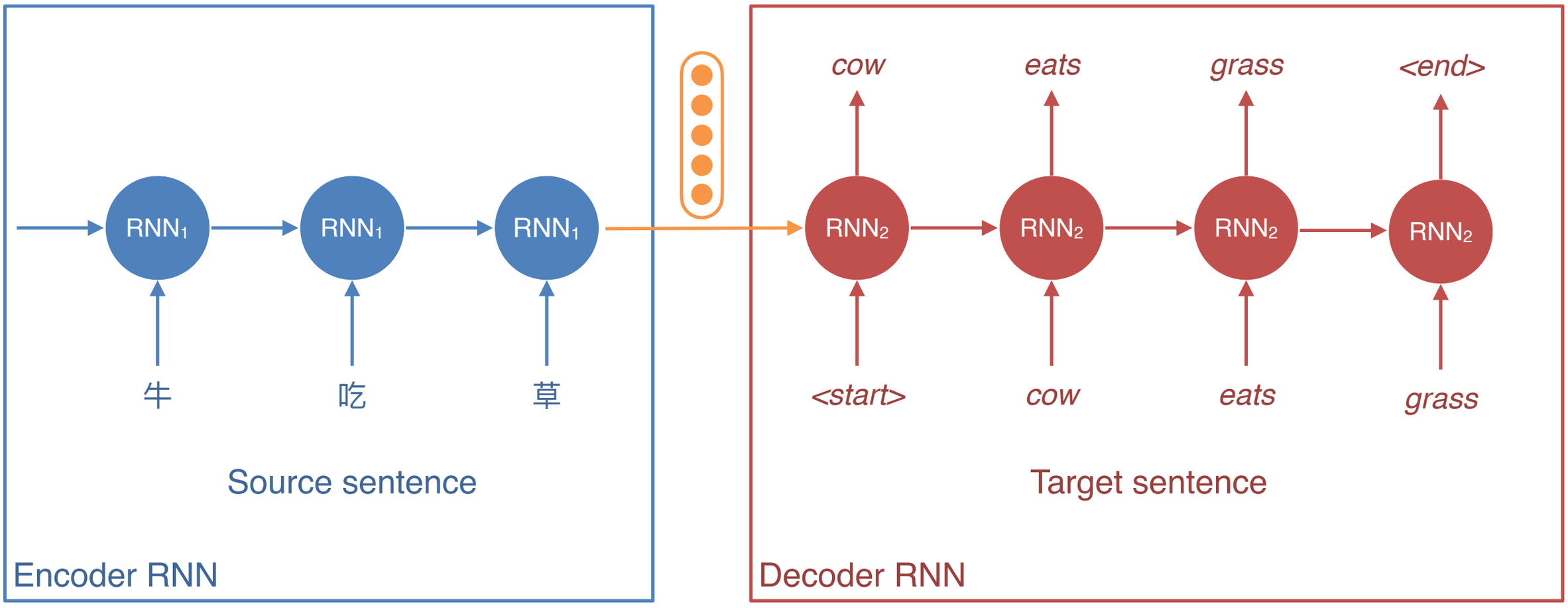
*a detained **iranian-american academic** accused of acting against national security has been **released** from a **tehran** prison after a hefty **bail** was posted, a top judiciary official said tuesday*

↓ paraphrase.

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

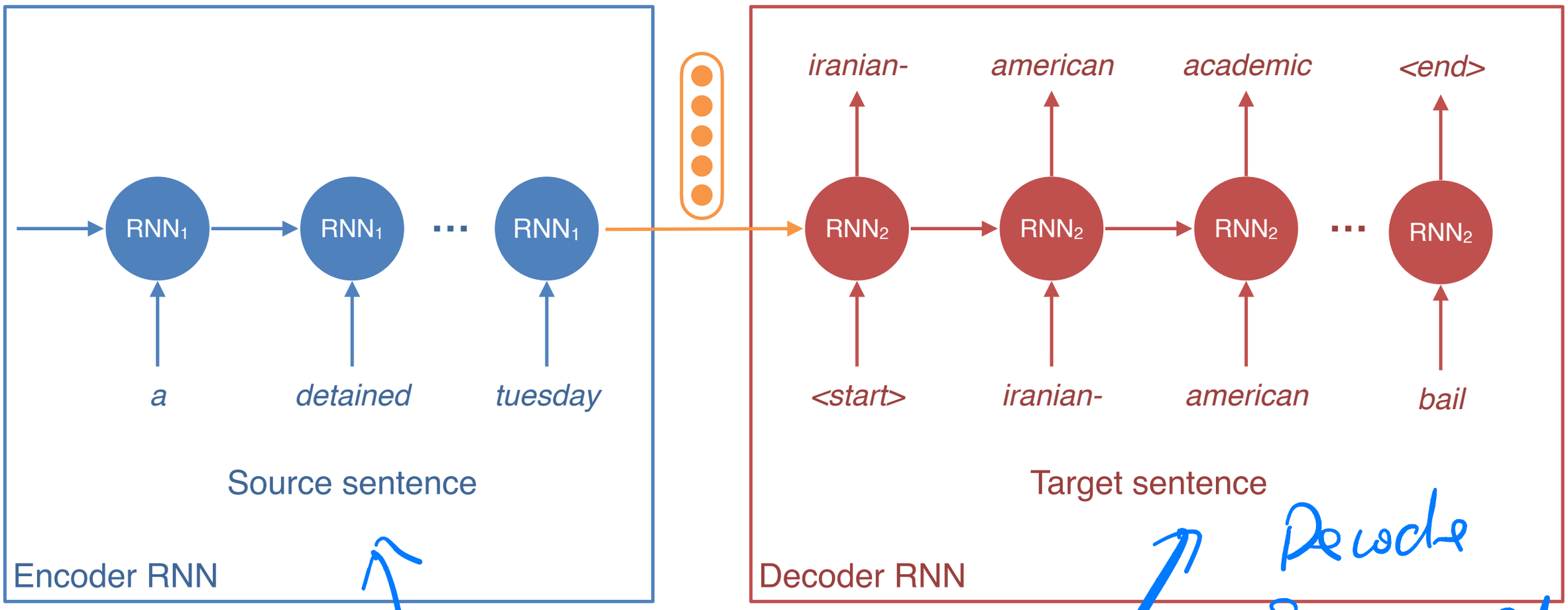
- Paraphrase
- A very difficult task
- Can we train a neural network to generate summary?

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”

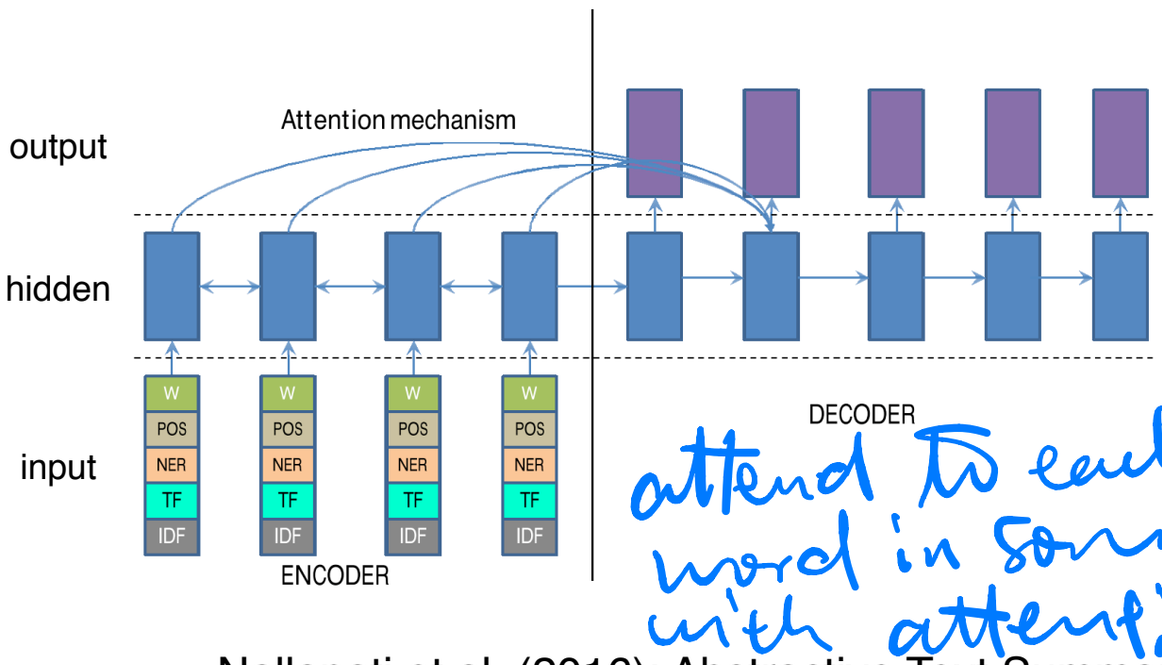
And It Kind of Works…

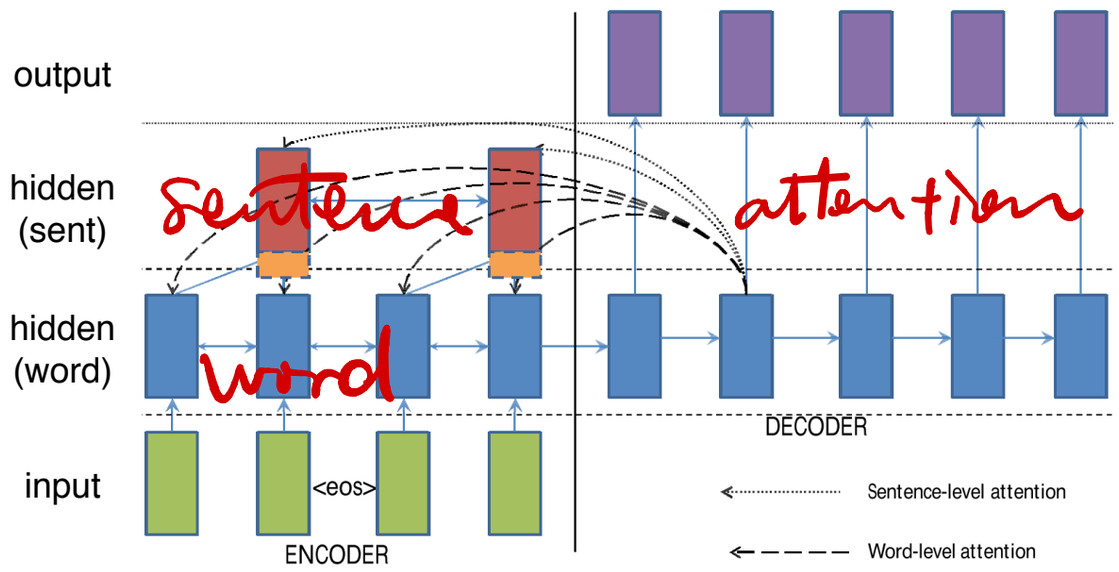
I(1): a detained iranian-american academic accused of acting against national security has been released from a tehran prison after a hefty bail was posted , a to p judiciary official said tuesday .
G: iranian-american academic held in tehran released on bail
A: detained iranian-american academic released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

Improvements

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

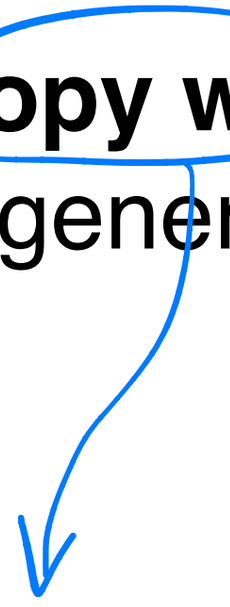


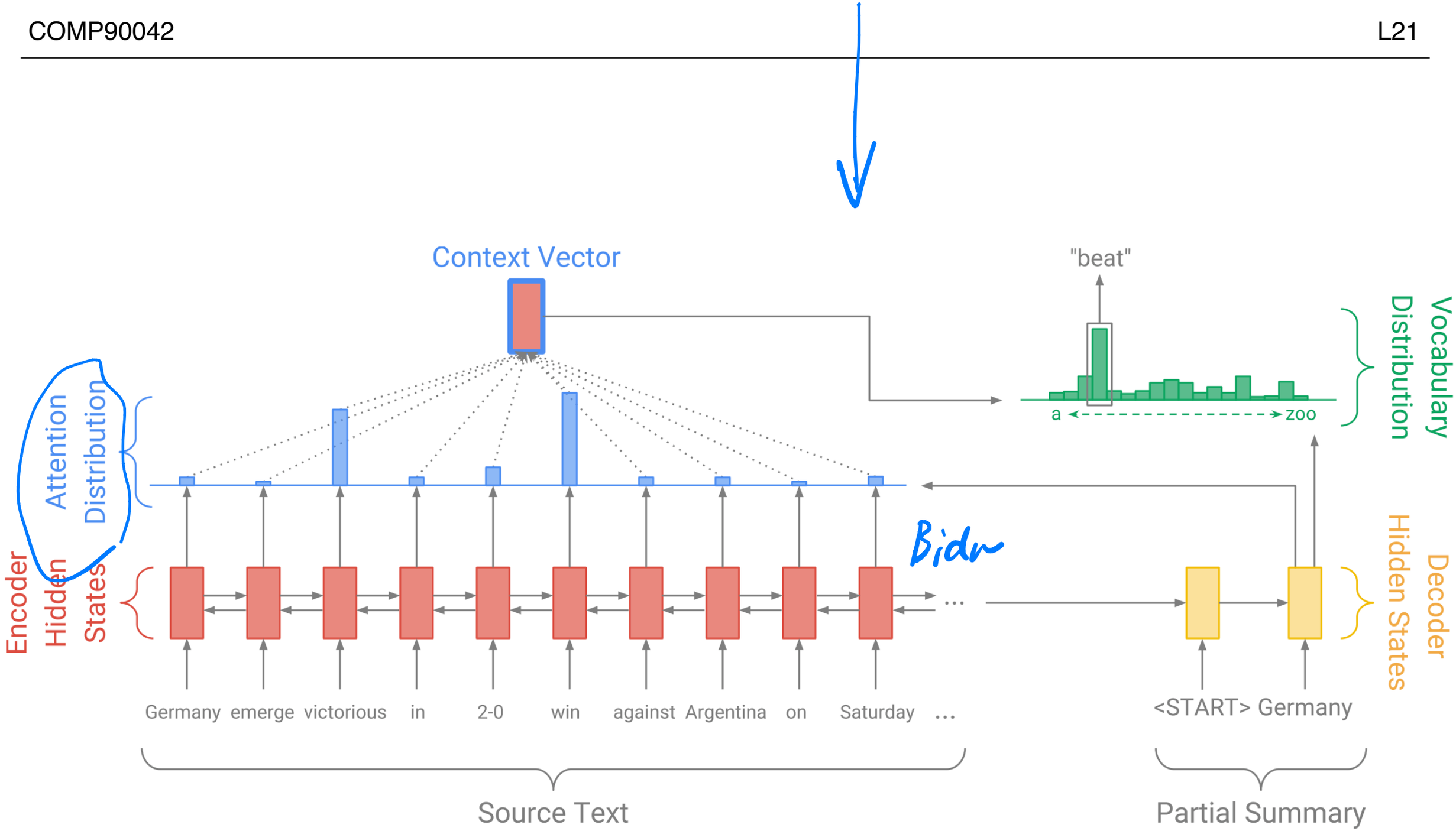


Nallapati et al. (2016): Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond

Issues

- Occasionally reproduce statements incorrectly (hallucinate new details!)
- Unable to handle out-of-vocab words in document
  - Generate UNK in summary
  - E.g. new names in test documents
- Solution: allow decoder to **copy words** directly from input document during generation





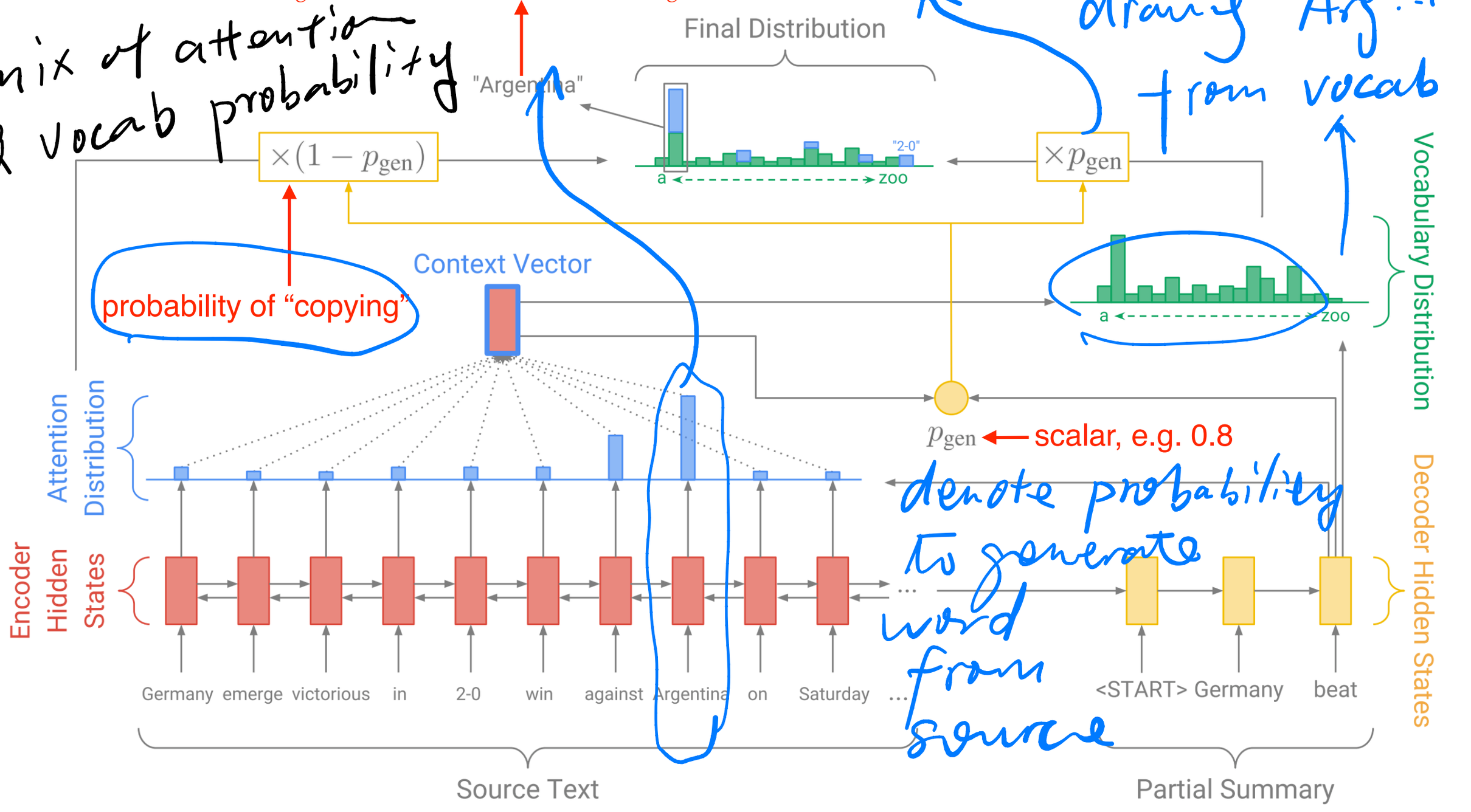
Encoder-decoder with Attention

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

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

mix of attention & vocab probability

drawing "Arg..." from vocab



Encoder-decoder with Attention + Copying

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

still chance to generate such word.

no generation prob.

Generated Summaries

Article: smugglers lure arab and african migrants by offering discounts to get onto overcrowded ships if people bring more potential passengers, a cnn investigation has revealed. (...)
Summary: cnn investigation **uncovers** the **business inside** a **human smuggling ring**.

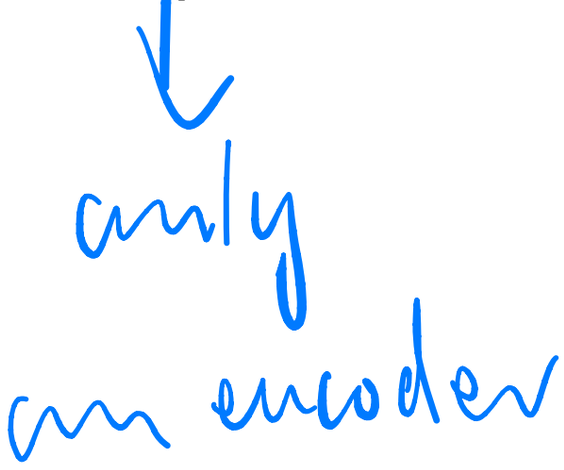
Article: andy murray (...) is into the semi-finals of the miami open , but not before getting a scare from 21 year-old austrian dominic thiem, who pushed him to 4-4 in the second set before going down 3-6 6-4, 6-1 in an hour and three quarters. (...)
Summary: andy murray **defeated** dominic thiem 3-6 6-4, 6-1 in an hour and three quarters.

More Summarisation Data

- But headline generation isn't really exciting…
- Latest 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

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 (e.g. from 1 to 3)

- ROUGE-2: calculates the percentage of bigrams from the reference that are in the generated summary

rather than in the summary

ROUGE-2: Example

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

generate summary

total bigram in #

- **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* $= \frac{3 + 6}{10 + 9}$

total bigram in reference.

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

much more difficult

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