Natural Language Understanding, Generation, and Machine Translation

Lecture 26: Summarisation

Pasquale Minervini
p.minervini@ed.ac.uk
March 15th, 2024

Natural Language Generation

(non-)linguistic input ⟹

⟹ text

databases
news articles
log files
images

reports
help messages
summaries
captions

Natural Language Generation

(non-)linguistic input ⟹

⟹ text

databases
news articles
log files
images

reports
help messages
summaries
captions

Natural Language Generation

(non-)linguistic input ⟹

⟹ text

databases
news articles
log files
images

reports
help messages
summaries
captions

Natural Language Generation

$\underbrace{\text{(non-)linguistic input}}$ $\Longrightarrow$ $\Longrightarrow$ $\underbrace{\text{text}}$

databases
news articles
log files
images

reports
help messages
summaries
captions

Summarisation

(non-)linguistic input ⟹

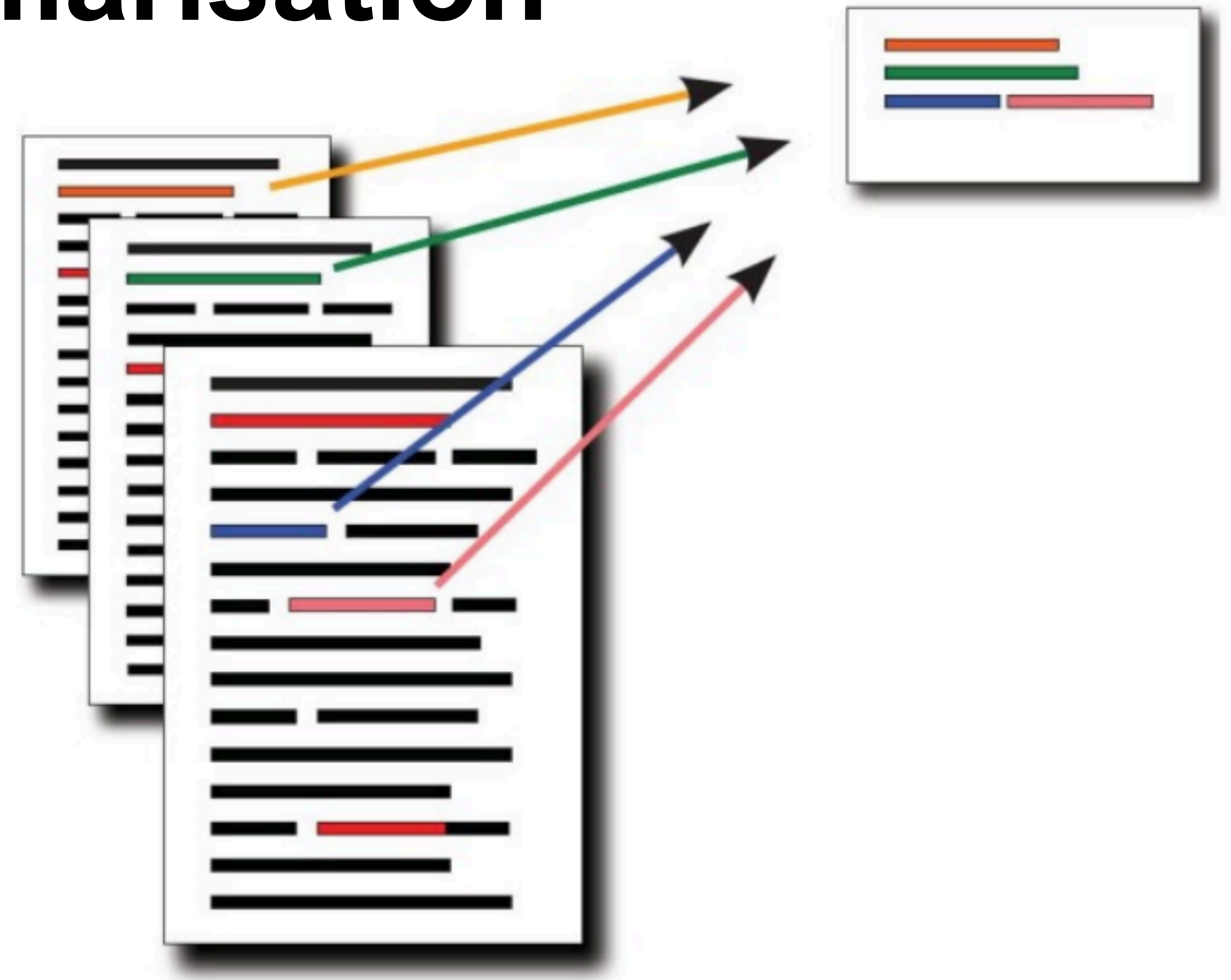
⟹ text

databases
news articles
log files
images

reports
help messages
summaries
captions

Summarisation

Summarisation task: produce a **concise and coherent** summary of a longer document or multiple documents, to **capture essential information** themes or points presented in the original document while **reducing its length**.



Types of Summarisation

Input:

- Single document summarisation (SDS) or Multi-document summarisation (MDS)

Types of Summarisation

Input:

- Single document summarisation (SDS) or Multi-document summarisation (MDS)

Output:

- Extractive or Abstractive

Types of Summarisation

Input:

- Single document summarisation (SDS) or Multi-document summarisation (MDS)

Output:

- Extractive or Abstractive

Focus:

- Generic (unconditioned) or query-focused (conditioned)

Types of Summarisation

Input:

- Single document summarisation (SDS) or Multi-document summarisation (MDS)

Output:

- Extractive or Abstractive

Focus:

- Generic (unconditioned) or query-focused (conditioned)

Approach:

- Supervised or unsupervised

Summarisation

Useful for creating, for example:

- **Outlines** or **abstracts** for documents and articles,
- **Summaries** for online conversations (Slack, e-mail)
- **Action items** for a meeting,
- **Simplifying documents** by compressing them,
- etc.

Summarisation

Facilitates information access:

- A lot of data, both in textual and non-textual format
- Even textual data can be difficult to read
- People tend to be more prone to understand text that numbers or graphs [Law et al., 2005]

Most NLP applications operate over text:

- Search engines
- Question answering systems
- Speech synthesisers

Summarisation

Stock data						
04/10/96	103	101.25	101.625	32444	-74	5485
04/09/96	104	101.5	101.625	41839	-33	5560
04/08/96	103.875	101.875	103.75	46096	-88	5594
04/05/96	Holiday					
04/04/96	104.875	103.5	104.375	18101	-6	5682

Summarisation

Stock data						
04/10/96	103	101.25	101.625	32444	-74	5485
04/09/96	104	101.5	101.625	41839	-33	5560
04/08/96	103.875	101.875	103.75	46096	-88	5594
04/05/96	Holiday					
04/04/96	104.875	103.5	104.375	18101	-6	5682

Microsoft avoided the downwards trend of the Dow Jones average today. Confined trading by all investors occurred today. After shooting to a high of \$104.87, its highest price so far for the month of April, Microsoft stock eased to finish at an enormous \$104.37. The Dow closed after trading at a weak 5682, down 6 points.

Summarisation

Team Stat Comparison		
1st Downs	19	22
Total Yards	338	379
Passing	246	306
Rushing	92	73
Penalties	16-149	7-46
3rd Down Conversions	4-13	6-16
4th Down Conversions	0-0	0-1
Turnovers	2	0
Possession	27:40	32:20

Summarisation

Team Stat Comparison		
1st Downs	19	22
Total Yards	338	379
Passing	246	306
Rushing	92	73
Penalties	16-149	7-46
3rd Down Conversions	4-13	6-16
4th Down Conversions	0-0	0-1
Turnovers	2	0
Possession	27:40	32:20

The New England Patriots lost two linebackers and two coaches in the offseason. They still know how to win thanks in large part to two stars they didn't lose. Tom Brady threw for 306 years and two touchdowns and Richard Seymour helped make a a game-turning defensive play as the Patriots opened their quest for an unprecedented third straight Super Bowl victory by beating Oakland 30–20 on Thursday night.

Summarisation

Summarisation

a crowd of people on a beach flying kites.

Summarisation

a crowd of people on a beach flying kites.
a man flying kite in the middle of a crowded beach.

Summarisation

a crowd of people on a beach flying kites.
a man flying kite in the middle of a crowded beach.
lots of people enjoying their time on the beach.

Summarisation

Most blacks say MLK's vision fulfilled, poll finds WASHINGTON (CNN) – More than two-thirds of African-Americans believe Martin Luther King Jr.'s vision for race relations has been fulfilled, a CNN poll found – a figure up sharply from a survey in early 2008.

The CNN-Opinion Research Corp. survey was released Monday, a federal holiday honoring the slain civil rights leader and a day before Barack Obama is to be sworn in as the first black U.S. president.

The poll found 69 percent of blacks said King's vision has been fulfilled in the more than 45 years since his 1963 'I have a dream' speech – roughly double the 34 percent who agreed with that assessment in a similar poll taken last March.

But whites remain less optimistic, the survey found. 'Whites don't feel the same way – a majority of them say that the country has not yet fulfilled King's vision,' CNN polling director Keating Holland said. However, the number of whites saying the dream has been fulfilled has also gone up since March, from 35 percent to 46 percent.

Summarisation

Most blacks say MLK's vision fulfilled, poll finds WASHINGTON (CNN) – More than two-thirds of African-Americans believe Martin Luther King Jr.'s vision for race relations has been fulfilled, a CNN poll found – a figure up sharply from a survey in early 2008.

The CNN-Opinion Research Corp. survey was released Monday, a federal holiday honoring the slain civil rights leader and a day before Barack Obama is to be sworn in as the first black U.S. president.

The poll found 69 percent of blacks said King's vision has been fulfilled in the more than 45 years since his 1963 'I have a dream' speech – roughly double the 34 percent who agreed with that assessment in a similar poll taken last March.

But whites remain less optimistic, the survey found. 'Whites don't feel the same way – a majority of them say that the country has not yet fulfilled King's vision,' CNN polling director Keating Holland said. However, the number of whites saying the dream has been fulfilled has also gone up since March, from 35 percent to 46 percent.

Highlights:

- 69% of blacks polled say Martin Luther King Jr's vision realised
- Slim majority of white people say King's vision is not fulfilled
- King gave his "I have a dream" speech in 1963

Modeling Approach

A **language model** produces a distribution over possible next words, given the previous words in the text:

$$P\left(y_t \mid y_1, \ldots, y_{t-1}\right)$$

Modeling Approach

A **language model** produces a distribution over possible next words, given the previous words in the text:

$$P\left(y_t \mid y_1, \ldots, y_{t-1}\right)$$

A **conditional language model** produces a distribution over possible next words, given the previous words in the text *and some additional input x:*

$$P\left(y_t \mid y_1, \ldots, y_{t-1}, x\right)$$

Modeling Approach

A **language model** produces a distribution over possible next words, given the previous words in the text:

$$P\left(y_t \mid y_1, \ldots, y_{t-1}\right)$$

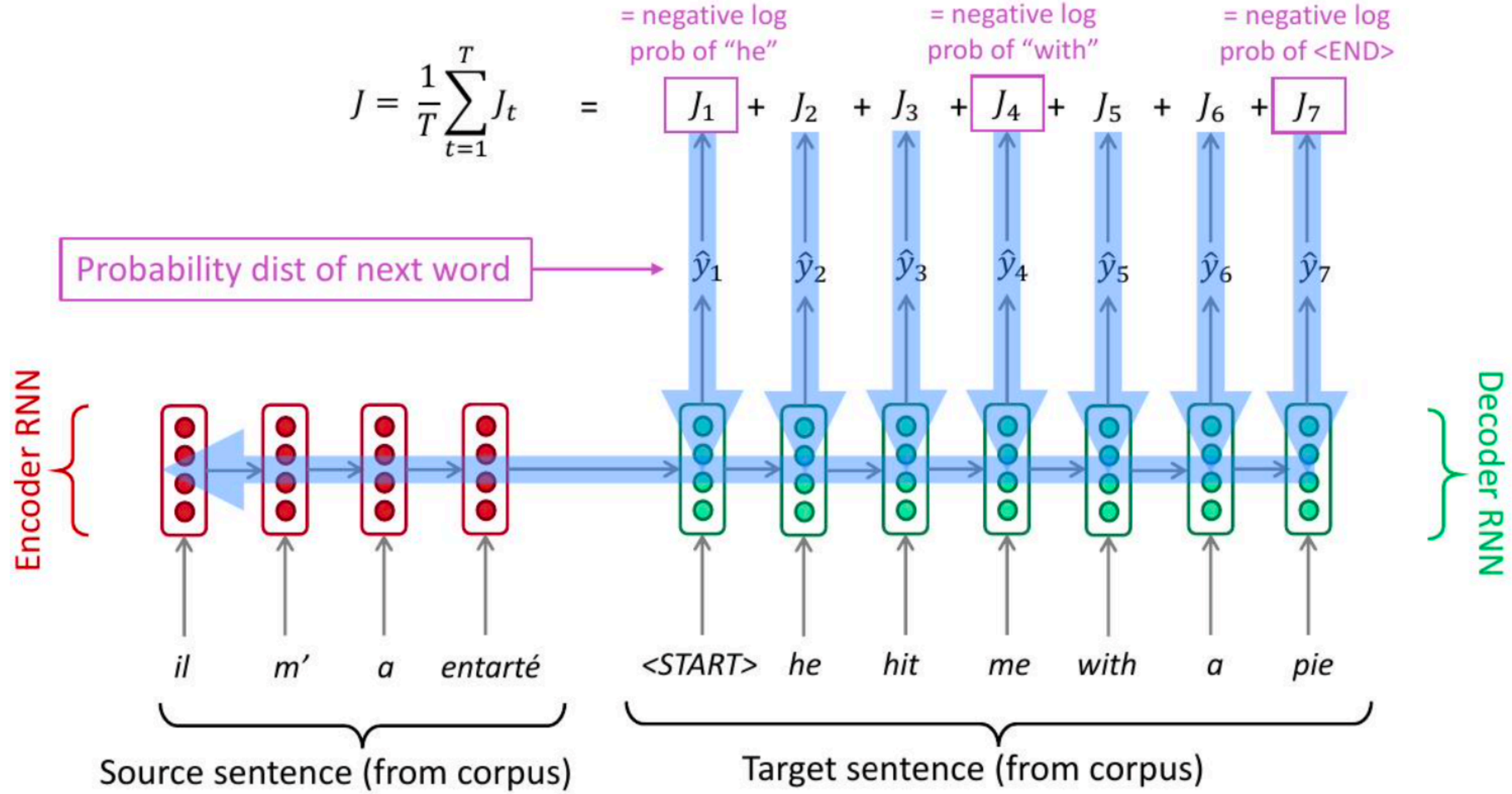
A **conditional language model** produces a distribution over possible next words, given the previous words in the text *and some additional input x:*

$$P\left(y_t \mid y_1, \ldots, y_{t-1}, x\right)$$

We can use any sequence to sequence model for representing this conditional distribution!

Summarisation — x: input text, y: summarised text

Modeling Approach



Summarisation — Task Definition

Definition: Given an input text x (single- or multi-document), write a summary y which is shorter and contains the main information in x.

Summarisation — Task Definition

Definition: Given an input text x (single- or multi-document), write a summary y which is shorter and contains the main information in x.

- **Single-document:** we write a summary y of a single document x
- **Multi-document:** we write a summary y of multiple documents $x_1, \ldots, x_n$

Summarisation — Task Definition

Definition: Given an input text x (single- or multi-document), write a summary y which is shorter and contains the main information in x.

- **Single-document:** we write a summary y of a single document x
- **Multi-document:** we write a summary y of multiple documents $x_1, \ldots, x_n$

Typically, the documents $x_1, \ldots, x_n$ have **overlapping content** — e.g., news articles discussing the same event

Summarisation — Main Strategies

Extractive Summarisation:
select parts (e.g., sentences) of the original text to form a summary.

"Easier", more restrictive
(no paraphrasing allowed)

Summarisation — Main Strategies

Extractive Summarisation:
select parts (e.g., sentences) of the original text to form a summary.

"Easier", more restrictive
(no paraphrasing allowed)

Abstractive Summarisation:
generate new text using natural language generation methods.

"More difficult", flexible.
(can do paraphrasing)

CNN/Daily Mail Dataset

Training data: pairs of news articles (~800 words on average) and summaries (aka *story highlights*), usually 3 or 4 sentences long (~56 words on average)

CNN/Daily Mail Dataset

Training data: pairs of news articles (~800 words on average) and summaries (aka *story highlights*), usually 3 or 4 sentences long (~56 words on average)

CNN: 100k pairs; Daily Mail: 200k pairs

CNN/Daily Mail Dataset

Training data: pairs of news articles (~800 words on average) and summaries (aka *story highlights*), usually 3 or 4 sentences long (~56 words on average)

CNN: 100k pairs; Daily Mail: 200k pairs

Highlights were sourced from journalists in compressed, “telegraphic”, manner

CNN/Daily Mail Dataset

Training data: pairs of news articles (~800 words on average) and summaries (aka *story highlights*), usually 3 or 4 sentences long (~56 words on average)

CNN: 100k pairs; Daily Mail: 200k pairs

Highlights were sourced from journalists in compressed, “telegraphic”, manner

The highlights need not to form a coherent summary — each highlight is relatively stand-alone, with little co-referencing

Available at **https://github.com/abisee/cnn-dailymail**

Summarisation

Most blacks say MLK's vision fulfilled, poll finds WASHINGTON (CNN) – More than two-thirds of African-Americans believe Martin Luther King Jr.'s vision for race relations has been fulfilled, a CNN poll found – a figure up sharply from a survey in early 2008.

The CNN-Opinion Research Corp. survey was released Monday, a federal holiday honoring the slain civil rights leader and a day before Barack Obama is to be sworn in as the first black U.S. president.

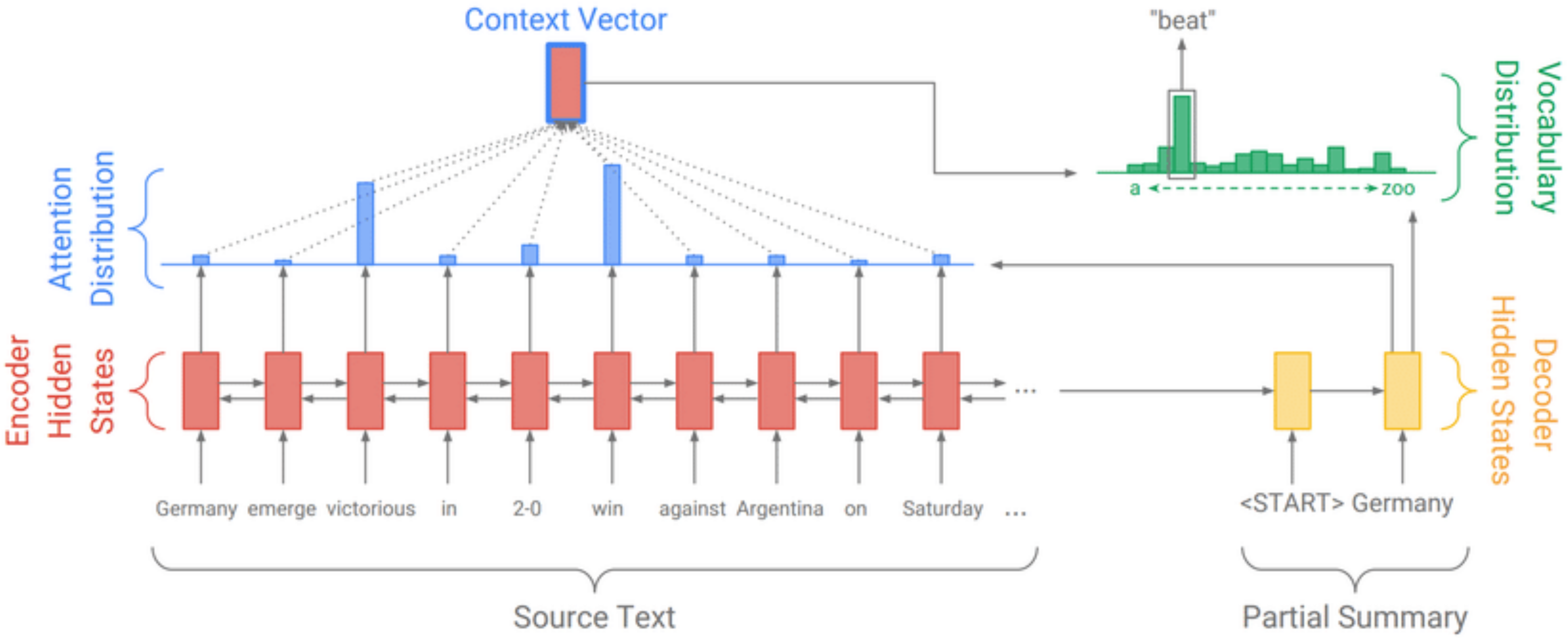
The poll found 69 percent of blacks said King's vision has been fulfilled in the more than 45 years since his 1963 'I have a dream' speech – roughly double the 34 percent who agreed with that assessment in a similar poll taken last March.

But whites remain less optimistic, the survey found. 'Whites don't feel the same way – a majority of them say that the country has not yet fulfilled King's vision,' CNN polling director Keating Holland said. However, the number of whites saying the dream has been fulfilled has also gone up since March, from 35 percent to 46 percent.

Highlights: Paraphrased Verbatim

- **69% of blacks** *polled* say **Martin Luther King Jr's vision** *realised*
- *Slim majority of white people say King's vision is not fulfilled*
- King gave *his* **"I have a dream" speech** *in 1963*

Sequence-to-Sequence with Attention



Sequence-to-Sequence with Attention

Encoder: single-layer bidirectional LSTM produces a sequence of *hidden states* $\boxed{h_i}$

Sequence-to-Sequence with Attention

Encoder: single-layer bidirectional LSTM produces a sequence of *hidden states* $\boxed{h_i}$

Decoder: single-layer unidirectional LSTM receives word embeddings of previous words produced by the decoder, and has a *decoder state* $\boxed{s_t}$

Sequence-to-Sequence with Attention

Encoder: single-layer bidirectional LSTM produces a sequence of *hidden states* $\boxed{h_i}$

Decoder: single-layer unidirectional LSTM receives word embeddings of previous words produced by the decoder, and has a *decoder state* $\boxed{s_t}$

Attention distribution: $e_i^t = v^\top \tanh\left(W_h \boxed{h_i} + W_s \boxed{s_t} + b_{\text{attn}}\right); \; a^t = \text{softmax}(e^t)$

Sequence-to-Sequence with Attention

Encoder: single-layer bidirectional LSTM produces a sequence of *hidden states* $\boxed{h_i}$

Decoder: single-layer unidirectional LSTM receives word embeddings of previous words produced by the decoder, and has a *decoder state* $\boxed{s_t}$

Attention distribution: $e_i^t = v^\top \tanh\left(W_h \boxed{h_i} + W_s \boxed{s_t} + b_{\text{attn}}\right); \;\; a^t = \text{softmax}(e^t)$

Context vector: weighted sum of enc. hidden states $h_i^* = \sum_i a_i^t h_i$

Sequence-to-Sequence with Attention

Encoder: single-layer bidirectional LSTM produces a sequence of *hidden states* $\boxed{h_i}$

Decoder: single-layer unidirectional LSTM receives word embeddings of previous words produced by the decoder, and has a *decoder state* $\boxed{s_t}$

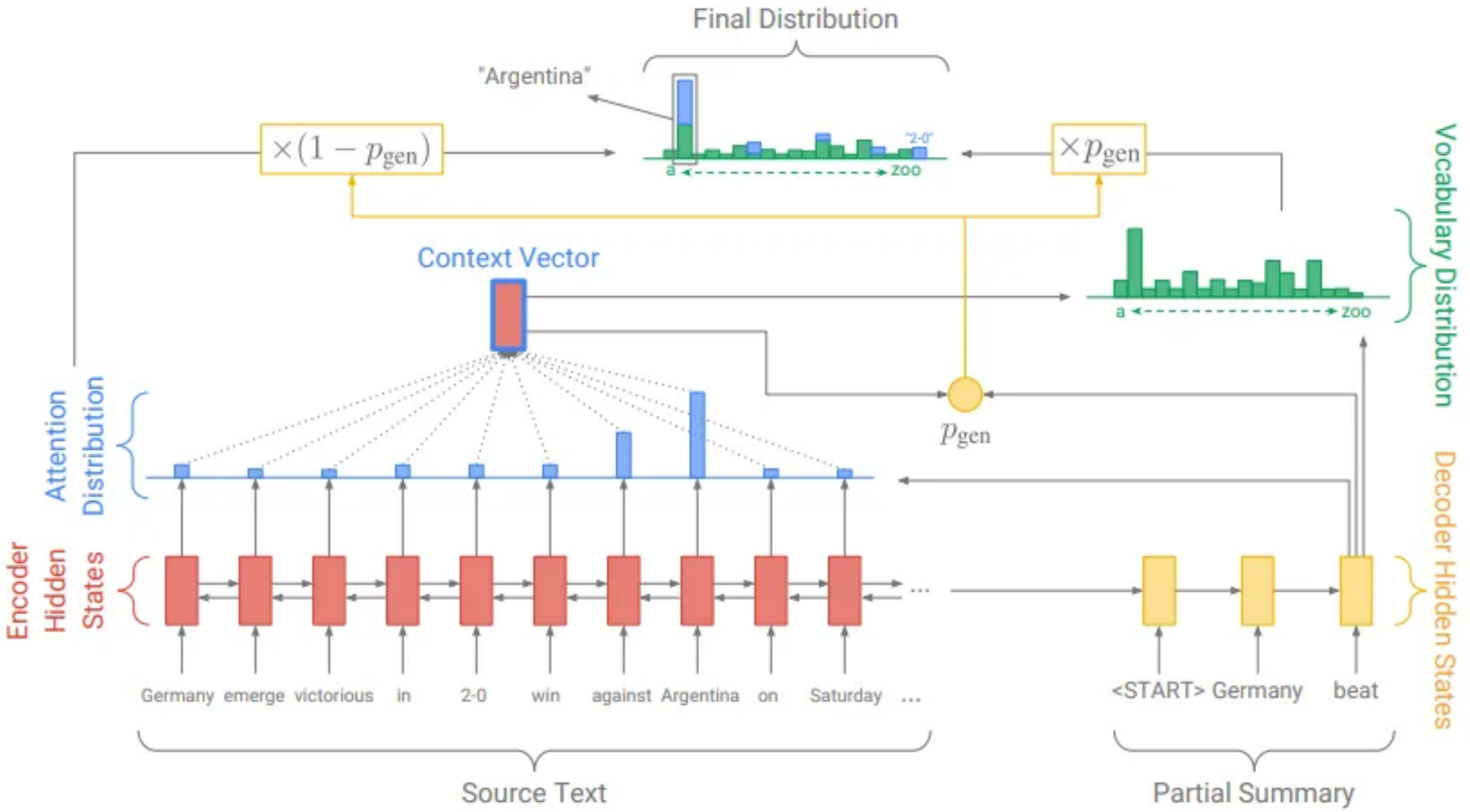
Attention distribution: $e_i^t = v^\top \tanh\left(W_h \boxed{h_i} + W_s \boxed{s_t} + b_{\text{attn}}\right); \; a^t = \text{softmax}(e^t)$

Context vector: weighted sum of enc. hidden states $h_i^* = \sum_i a_i^t h_i$

Vocab distribution: probability distribution over words in the vocabulary:

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

Pointer-Generator Network



Pointer-Generator Network

Pointer-Generator Network: implements a *copying mechanism*, useful for rare words and phrases

The model allows both *copying words by pointing* and *generating words* from a fixed vocabulary

Pointer-Generator Network

Pointer-Generator Network: implements a *copying mechanism*, useful for rare words and phrases

The model allows both *copying words by pointing* and *generating words* from a fixed vocabulary

On each decoder step, calculate p_{gen} which represents the probability of *generating the next word (rather than copying it):*

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

At each decoding step

Probability of copying

Pointer-Generator Network — Coverage Mechanism

The **coverage mechanism** attempts to generate less repetitive summaries by **penalising repeatedly attending to the same parts** of the source text

Pointer-Generator Network — Coverage Mechanism

The **coverage mechanism** attempts to generate less repetitive summaries by **penalising repeatedly attending to the same parts** of the source text

Coverage vector tells us what has been attended so far:

$$c^t = \sum_{t'}^{t-1} a^{t'}$$

Pointer-Generator Network — Coverage Mechanism

The **coverage mechanism** attempts to generate less repetitive summaries by **penalising repeatedly attending to the same parts** of the source text

Coverage vector tells us what has been attended so far:

$$c^t = \sum_{t'}^{t-1} a^{t'}$$

The coverage vector is provided as an extra input to the attention mechanism:

$$e_i^t = v^\top \text{tanh}\left(W_h h_i + W_s s_t + w_c c_i^t + b_{\text{attn}}\right)$$

Pointer-Generator Network — Coverage Mechanism

The **coverage mechanism** attempts to generate less repetitive summaries by **penalising repeatedly attending to the same parts** of the source text

Coverage vector tells us what has been attended so far:

$$c^t = \sum_{t'}^{t-1} a^{t'}$$

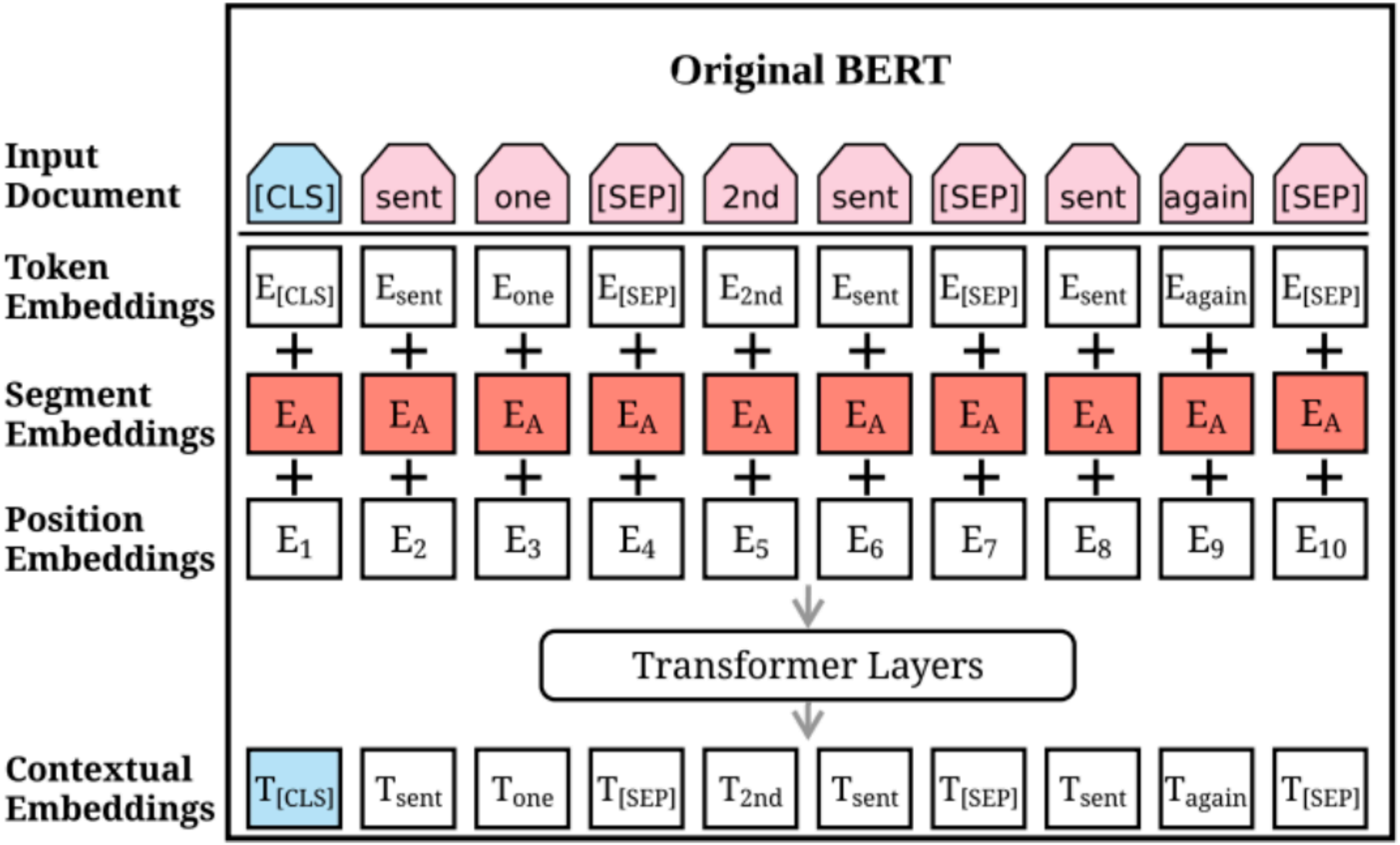
The coverage vector is provided as an extra input to the attention mechanism:

$$e_i^t = v^\top \text{tanh}\left(W_h h_i + W_s s_t + w_c c_i^t + b_{\text{attn}}\right)$$

Coverage loss penalises overlap between coverage vector c^t and new attention distribution a^t:

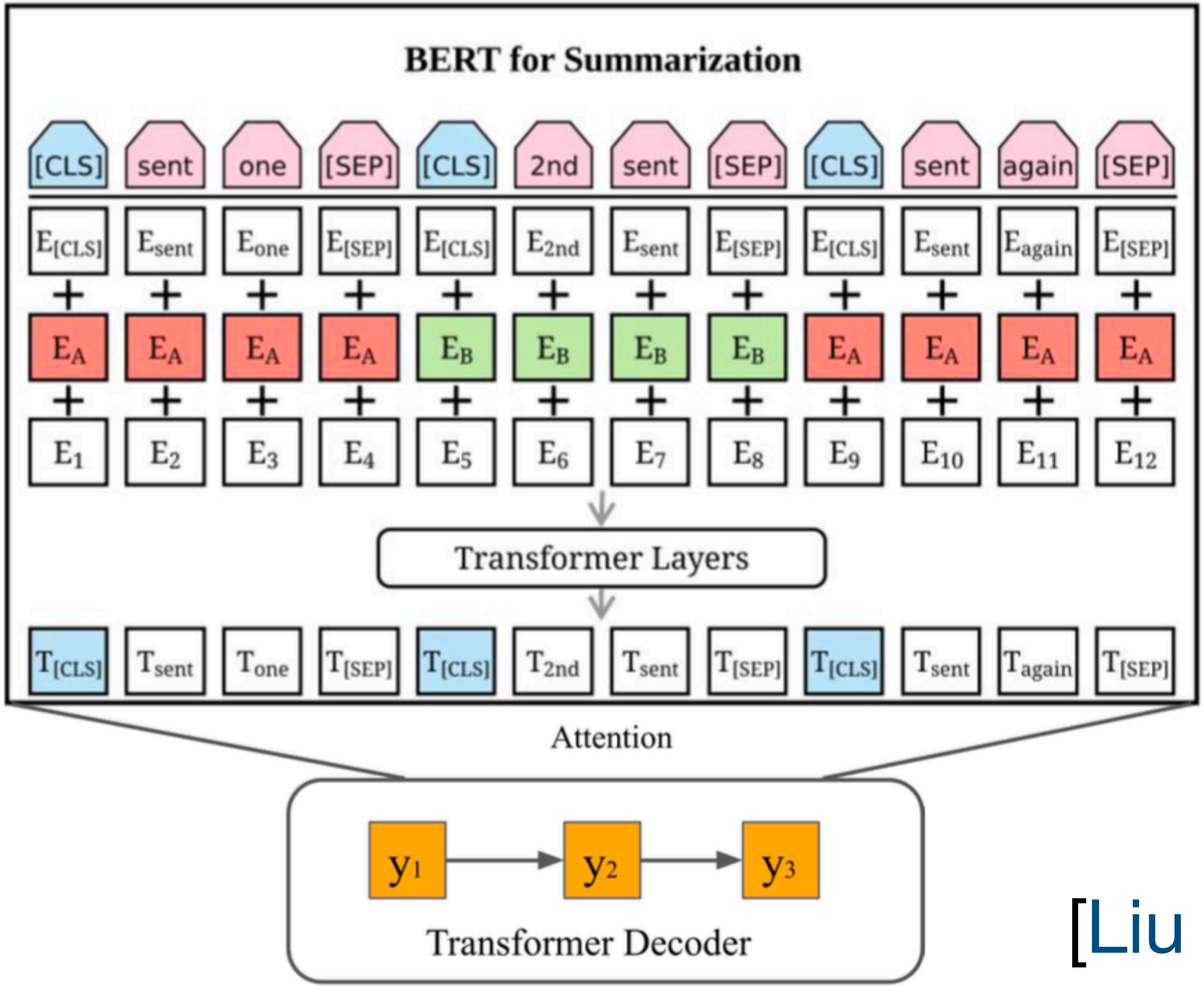
$$\text{covloss}_t = \sum_i \min\left(a_i^t, c_i^t\right)$$

Summarisation with Pre-Trained Encoders



[Devlin et al., 2018]

Summarisation with Pre-Trained Encoders



[Liu et al., 2019]

Pre-Trained Encoders — Fine-Tuning

Learning rate schedule [Vaswani et al., 2017]

$$\mathsf{lr} = \tilde{\mathsf{lr}} \cdot \min\{\mathsf{step}^{-0.5}, \mathsf{step} \cdot \mathsf{warmup}^{-1.5}\}$$

Smaller learning rate, longer warming-up for the **encoder**:

$$\tilde{\mathsf{lr}}_e = 2e^{-3}, \quad \mathsf{warmup}_e = 20{,}000$$

Larger learning rate, shorter warming-up for the **decoder**:

$$\tilde{\mathsf{lr}}_d = 0.1, \quad \mathsf{warmup}_d = 10{,}000$$

Summarisation Evaluation — ROUGE

ROUGE — **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{Ref. Summaries}} \sum_{\text{gram}_n \in S} \text{count}_{\text{match}}\left(\text{gram}_n\right)}{\sum_{S \in \text{Ref. Summaries}} \sum_{\text{gram}_n \in S} \text{count}\left(\text{gram}_n\right)}$$

Summarisation Evaluation — ROUGE

ROUGE — **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{Ref. Summaries}} \sum_{\text{gram}_n \in S} \text{count}_{\text{match}}\left(\text{gram}_n\right)}{\sum_{S \in \text{Ref. Summaries}} \sum_{\text{gram}_n \in S} \text{count}\left(\text{gram}_n\right)}$$

Based on **n-gram overlap**

No brevity penalty, based on **recall**

Most commonly-reported ROUGE scores: ROUGE-1 **unigram** overlap, ROUGE-2 **bigram** overlap, ROUGE-L **longest common subsequence** overlap

Summarisation — Discussion

CNN/Daily Mail is a **rather extractive** dataset — you can get away with some copying and pasting

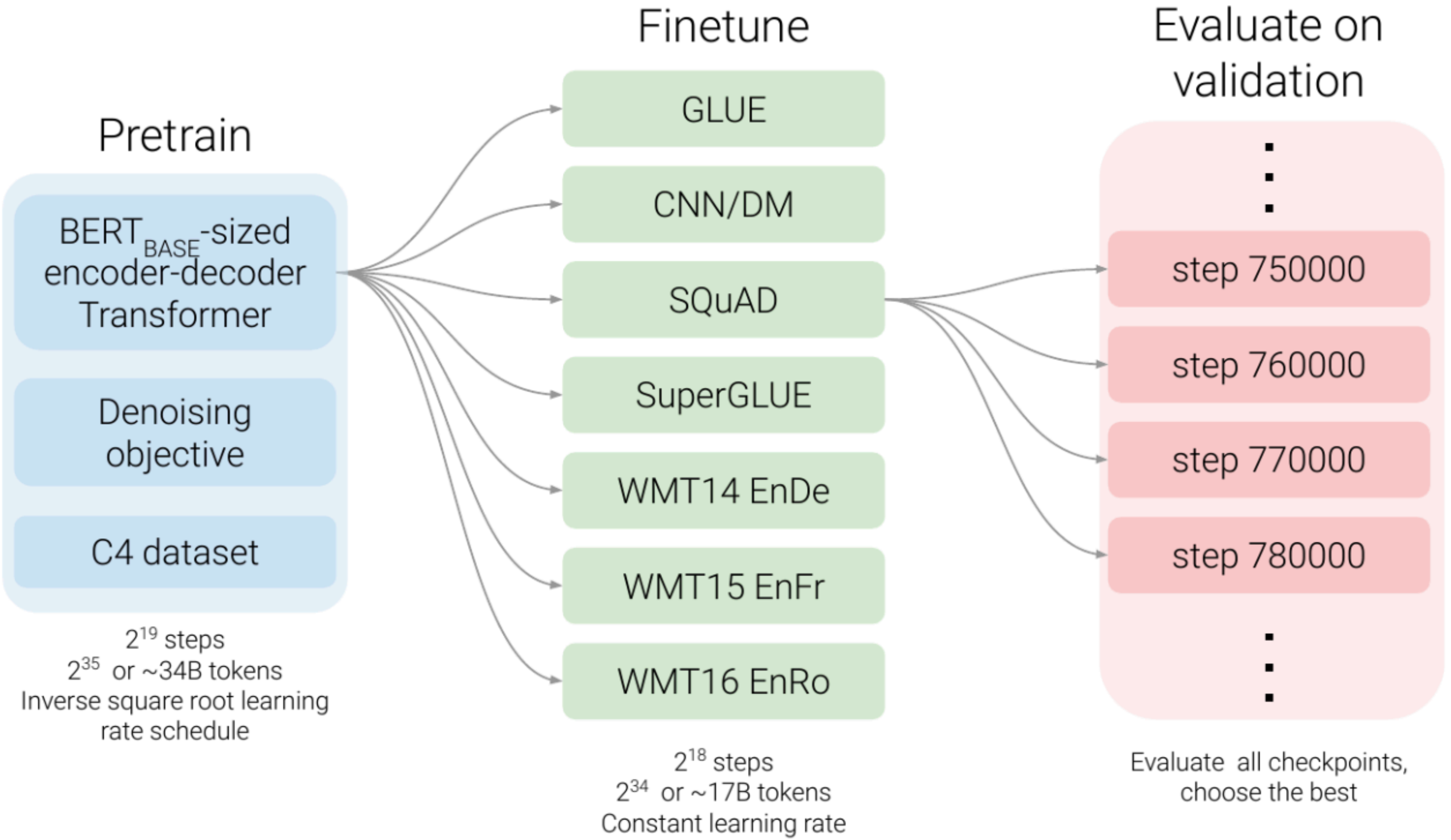
Generated summaries are fluent but can contain **factual inaccuracies**

Do we trust ROUGE as an evaluation metric? How do we evaluate output summaries with humans?

How would we build an extractive summarisation model? How would the training data look like?

T5: Text-to-Text Transfer Transformer

T5: Text-to-Text Transfer Transformer



T5: Text-to-Text Transfer Transformer

T5-Small (60 million parameters): `gs://t5-data/pretrained_models/small`

T5-Base (220 million parameters): `gs://t5-data/pretrained_models/base`

T5-Large (770 million parameters): `gs://t5-data/pretrained_models/large`

T5-3B (3 billion parameters): `gs://t5-data/pretrained_models/3B`

T5-11B (11 billion parameters): `gs://t5-data/pretrained_models/11B`

T5: Text-to-Text Transfer Transformer

Models	ROUGE 1	ROUGE 2	ROUGE L
seq-to-seq+attn	31.33	11.81	28.83
pointer-generator	36.44	15.66	33.42
pointer-generator + coverage	39.53	17.28	36.38
lead-3 baseline	40.34	17.70	36.57
BertSumAbs	41.72	19.39	38.76
T5-Small	41.12	19.56	38.35
T5-Base	42.05	20.34	39.40
T5-Large	42.50	20.68	39.75
T5-3B	43.52	21.55	40.69

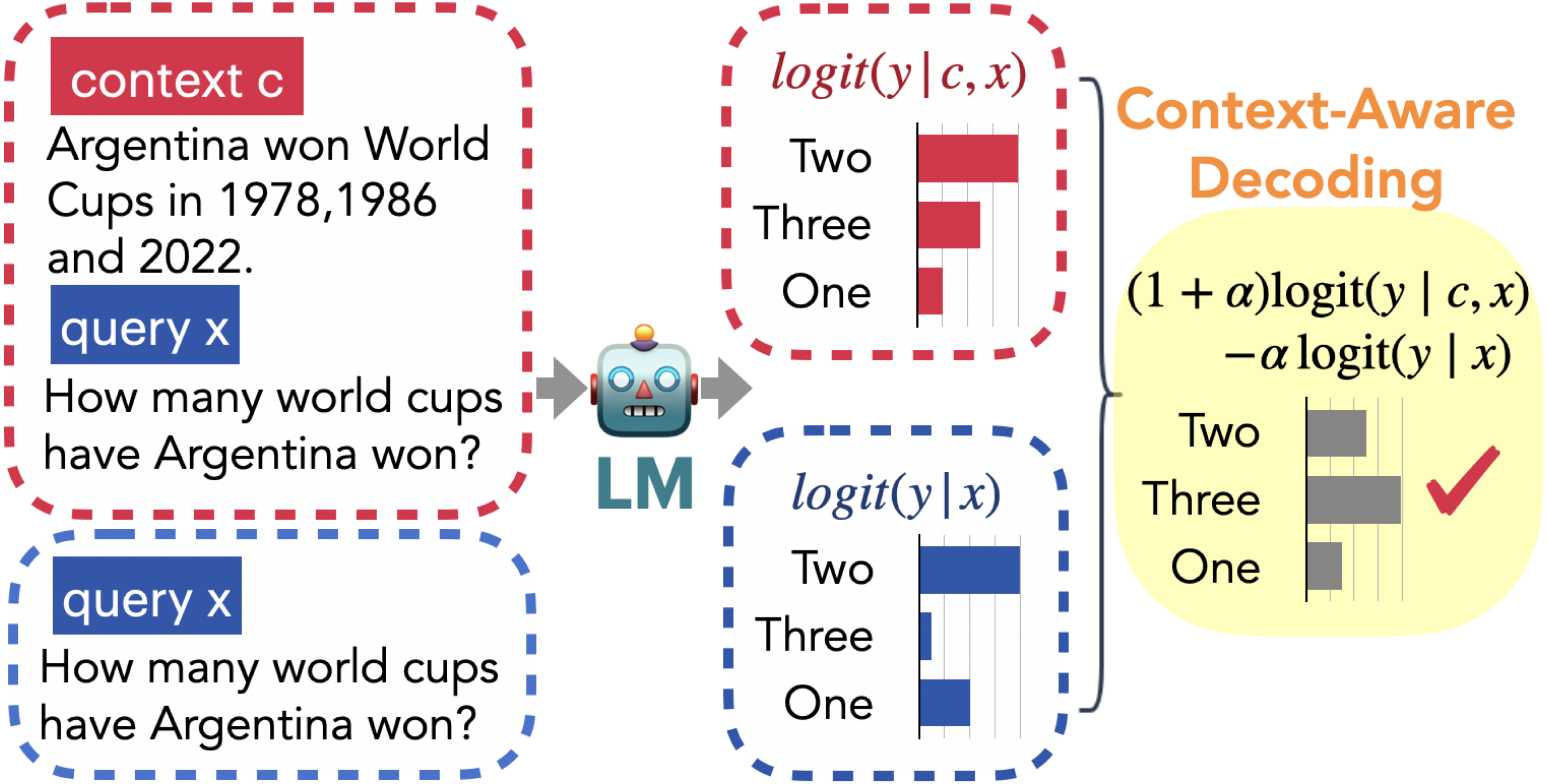
Zero-Shot Summarisation with LLMs

	XSUM
$\boldsymbol{c}$	Article: Prison Link Cymru had 1,099 referrals in 2015-16 and said some ex-offenders were living rough for up to a year before finding suitable accommodation ...
$\boldsymbol{x}$	Summarize the article in one sentence. Summary:

The Problem of Hallucinations

XSUM	
Article	He passed away peacefully in hospital on Tuesday after a short illness. Born in Tourmakeady, County Mayo, he worked as a teacher before securing a part in the premiere of the Brian Friel play Translations in 1980. Lally became a household name in Ireland for his role as Miley Byrne in the RTE soap opera Glenroe and later starred in the BBC series Ballykissangel. He also appeared in the Hollywood movie Alexander and provided the voice for the Oscar-nominated, animated Irish film, The Secret of Kells. As a fluent Irish speaker and advocate of the language, Lally had roles in several Irish language films ...
Regular	Westminister actor Pat Lally died in hospital on Tuesday night aged 82

Context-Aware Decoding



[Shi et al., 2023]

Context-Aware Decoding

XSUM	
Article	He passed away peacefully in hospital on Tuesday after a short illness. Born in Tourmakeady, County Mayo, he worked as a teacher before securing a part in the premiere of the Brian Friel play Translations in 1980. Lally became a household name in Ireland for his role as Miley Byrne in the RTE soap opera Glenroe and later starred in the BBC series Ballykissangel. He also appeared in the Hollywood movie Alexander and provided the voice for the Oscar-nominated, animated Irish film, The Secret of Kells. As a fluent Irish speaker and advocate of the language, Lally had roles in several Irish language films ...
Regular	Westminister actor Pat Lally died in hospital on Tuesday night aged 82
CAD	Actor Lally, best known for Glenroe and Ballykissangel, has died in hospital on Tuesday

[Shi et al., 2023]

Context-Aware Decoding

MemoTrap	
Input	Write a quote that ends in the word "early". Better late than
Regular	never
CAD	early

[Shi et al., 2023]

Context-Aware Decoding

Model		Decoding	CNN-DM			XSUM		
			ROUGE-L	factKB	BERT-P	ROUGE-L	factKB	BERT-P
OPT	13B	Regular	22.0	77.8	86.5	16.4	47.2	85.2
		CAD	**27.4**	**84.1**	**90.8**	**18.2**	**64.9**	**87.5**
	30B	Regular	22.2	81.7	87.0	17.4	38.2	86.1
		CAD	**28.4**	**87.0**	**90.2**	**19.5**	**45.6**	**89.3**
GPT-Neo	3B	Regular	24.3	80.5	87.5	17.6	54.0	86.6
		CAD	**27.7**	**87.5**	**90.6**	**18.1**	**65.1**	**89.1**
	20B	Regular	18.7	68.3	85.2	14.9	42.2	85.7
		CAD	**24.5**	**77.5**	**89.4**	**19.0**	**63.3**	**90.6**
LLaMA	13B	Regular	27.1	80.2	89.5	19.0	53.5	87.8
		CAD	**32.6**	**90.8**	**93.0**	**21.1**	**73.4**	**91.7**
	30B	Regular	25.8	76.8	88.5	18.7	47.7	87.1
		CAD	**31.8**	**87.8**	**92.2**	**22.0**	**66.4**	**90.3**
FLAN	3B	Regular	25.5	90.2	91.6	18.8	31.9	88.2
		CAD	**26.1**	**93.9**	**92.1**	**19.5**	**35.9**	**88.8**
	11B	Regular	25.4	90.4	91.4	19.4	29.8	88.3
		CAD	**27.1**	**93.1**	**92.2**	**20.0**	**35.0**	**88.8**

[Shi et al., 2023]