Lecture 26: Summarisation

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

(non-)linguistic input \(\rightarrow\) 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 $\rightarrow$ text

- databases
- news articles
- log files
- images

- reports
- help messages
- summaries
- captions
Natural Language Generation

(non-)linguistic input $\rightarrow$ text

- databases
- news articles
- log files
- images

- reports
- help messages
- summaries
- captions
Summarisation

(non-)linguistic input $\rightarrow$ text

databases
news articles
log files
images

reports
help messages
summaries
captions
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

<table>
<thead>
<tr>
<th>Date</th>
<th>Stock Price</th>
<th>Opening Price</th>
<th>Previous Close</th>
<th>Volume</th>
<th>Change</th>
<th>Adjusted Close</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/10/96</td>
<td>103</td>
<td>101.25</td>
<td>101.625</td>
<td>32444</td>
<td>-74</td>
<td>5485</td>
</tr>
<tr>
<td>04/09/96</td>
<td>104</td>
<td>101.5</td>
<td>101.625</td>
<td>41839</td>
<td>-33</td>
<td>5560</td>
</tr>
<tr>
<td>04/08/96</td>
<td>103.875</td>
<td>101.875</td>
<td>103.75</td>
<td>46096</td>
<td>-88</td>
<td>5594</td>
</tr>
<tr>
<td>04/05/96</td>
<td>Holiday</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04/04/96</td>
<td>104.875</td>
<td>103.5</td>
<td>104.375</td>
<td>18101</td>
<td>-6</td>
<td>5682</td>
</tr>
</tbody>
</table>
**Summarisation**

<table>
<thead>
<tr>
<th>Date</th>
<th>Share</th>
<th>Price 1</th>
<th>Price 2</th>
<th>Price 3</th>
<th>Price 4</th>
<th>Price 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>04/10/96</td>
<td>103</td>
<td>101.25</td>
<td>101.625</td>
<td>32444</td>
<td>-74</td>
<td>5485</td>
</tr>
<tr>
<td>04/09/96</td>
<td>104</td>
<td>101.5</td>
<td>101.625</td>
<td>41839</td>
<td>-33</td>
<td>5560</td>
</tr>
<tr>
<td>04/08/96</td>
<td>103.875</td>
<td>101.875</td>
<td>103.75</td>
<td>46096</td>
<td>-88</td>
<td>5594</td>
</tr>
<tr>
<td>04/05/96</td>
<td>Holiday</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04/04/96</td>
<td>104.875</td>
<td>103.5</td>
<td>104.375</td>
<td>18101</td>
<td>-6</td>
<td>5682</td>
</tr>
</tbody>
</table>

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

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

\[
P \left( y_t | y_1, \ldots, y_{t-1} \right)
\]
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

\[ J = \frac{1}{T} \sum_{t=1}^{T} J_t \]

\[ = J_1 + J_2 + J_3 + J_4 + J_5 + J_6 + J_7 \]

- \( J_1 \): negative log prob of "he"
- \( J_2 \): negative log prob of "with"
- \( J_7 \): negative log prob of \(<\text{END}>\)

Probability dist of next word

Source sentence (from corpus)

Source sentence (from corpus)

\(<\text{START}>\) he hit me with a pie
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 $x_1, ..., x_n$ which have overlapping content—e.g., news articles discussing the same event $x_1, ..., x_n$. 
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 $x_1, \ldots, x_n$
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)

Available at https://github.com/abisee/cnn-dailymail
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
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**
Sequence-to-Sequence with Attention
Sequence-to-Sequence with Attention

Encoder: single-layer bidirectional LSTM produces a sequence of hidden states $h_i$
Sequence-to-Sequence with Attention

**Encoder:** single-layer bidirectional LSTM produces a sequence of *hidden states* \( h_i \)

**Decoder:** single-layer unidirectional LSTM receives word embeddings of previous words produced by the decoder, and has a *decoder state* \( S_t \)
Sequence-to-Sequence with Attention

**Encoder:** single-layer bidirectional LSTM produces a sequence of hidden states $[h_i]$

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

**Attention distribution:** $e_i^t = v^\top \tanh (W_h h_i + W_s s_t + b_{\text{attn}}); \quad a^t = \text{softmax}(e^t)$
Sequence-to-Sequence with Attention

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

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

**Attention distribution:** $e_t^i = v^\top \tanh(W_h h_i + W_s s_t + b_{attn}); \quad a_t = \text{softmax}(e_t)$

**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 \( h_i \)

Decoder: single-layer unidirectional LSTM receives word embeddings of previous words produced by the decoder, and has a decoder state \( s_t \)

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

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:
\[
P_{\text{vocab}} = \text{softmax} \left( V' \left( V[s_t, h_i^*] + b \right) + b' \right)\]
Pointer-Generator Network

![Diagram of Pointer-Generator Network]

- Encoder Hidden States
- Attention Distribution
- Context Vector
- Final Distribution
- Vocabulary Distribution
- Decoder Hidden States
- Source Text
- Partial Summary

Key Elements:
- $(1 - p_{gen})$
- $p_{gen}$
- "Argentina"
- "2-0"
- "zoo"
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 $p_{\text{gen}}$ which represents the probability of *generating the next word* (rather than copying it):

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

At each decoding step, the probability of copying is $1 - p_{\text{gen}}$. If $p_{\text{gen}} = 1$, then the model only copies from the source, and if $p_{\text{gen}} = 0$, it only generates from the vocabulary.
The coverage mechanism attempts to generate less repetitive summaries by penalising repeatedly attending to the same parts of the source text.
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 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:

\[ e^t_i = v^\top \tanh \left( W_h h_i + W_s s_t + w_c c^t_i + b_{\text{attn}} \right) \]
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:

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

**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]
\[
\text{lr} = \text{lr} \cdot \min\{\text{step}^{-0.5}, \text{step} \cdot \text{warmup}^{-1.5}\}
\]
Smaller learning rate, longer warming-up for the encoder:
\[
\text{lr}_e = 2e^{-3}, \quad \text{warmup}_e = 20,000
\]
Larger learning rate, shorter warming-up for the decoder:
\[
\text{lr}_d = 0.1, \quad \text{warmup}_d = 10,000
\]
Summarisation Evaluation — ROUGE

ROUGE — Recall-Oriented Understudy for Gisting Evaluation

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

ROUGE — Recall-Oriented Understudy for Gisting Evaluation

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

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

- "translate English to German: That is good."
- "cola sentence: The course is jumping well."
- "stsrb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
- "summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi..."
- "Das ist gut."
- "not acceptable"
- "six people hospitalized after a storm in attala county."
- "3.8"
T5: Text-to-Text Transfer Transformer

Pretrain
- BERT\textsubscript{BASE}-sized encoder-decoder Transformer
- Denoising objective
- C4 dataset

Finetune
- GLUE
- CNN/DM
- SQuAD
- SuperGLUE
- WMT14 EnDe
- WMT15 EnFr
- WMT16 EnRo

Evaluate on validation
- step 750000
- step 760000
- step 770000
- step 780000

Evaluate all checkpoints, choose the best
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

<table>
<thead>
<tr>
<th>Models</th>
<th>ROUGE 1</th>
<th>ROUGE 2</th>
<th>ROUGE L</th>
</tr>
</thead>
<tbody>
<tr>
<td>seq-to-seq+attn</td>
<td>31.33</td>
<td>11.81</td>
<td>28.83</td>
</tr>
<tr>
<td>pointer-generator</td>
<td>36.44</td>
<td>15.66</td>
<td>33.42</td>
</tr>
<tr>
<td>pointer-generator + coverage</td>
<td>39.53</td>
<td>17.28</td>
<td>36.38</td>
</tr>
<tr>
<td>lead-3 baseline</td>
<td>40.34</td>
<td>17.70</td>
<td>36.57</td>
</tr>
<tr>
<td>BERTSUMABS</td>
<td>41.72</td>
<td>19.39</td>
<td>38.76</td>
</tr>
<tr>
<td>T5-Small</td>
<td>41.12</td>
<td>19.56</td>
<td>38.35</td>
</tr>
<tr>
<td>T5-Base</td>
<td>42.05</td>
<td>20.34</td>
<td>39.40</td>
</tr>
<tr>
<td>T5-Large</td>
<td>42.50</td>
<td>20.68</td>
<td>39.75</td>
</tr>
<tr>
<td>T5-3B</td>
<td>43.52</td>
<td>21.55</td>
<td>40.69</td>
</tr>
</tbody>
</table>
Zero-Shot Summarisation with LLMs

---

**XSUM**

\(c\) 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 ...

\(x\) Summarize the article in one sentence. Summary:
The Problem of Hallucinations

| 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 | Westminster actor Pat Lally died in hospital on Tuesday night aged 82 |
Context-Aware Decoding

context $c$

query $x$
How many world cups have Argentina won?

[Shi et al., 2023]
<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article</td>
<td>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 ...</td>
</tr>
<tr>
<td>Regular</td>
<td>Westminster actor Pat Lally died in hospital on Tuesday night aged 82</td>
</tr>
<tr>
<td>CAD</td>
<td>Actor Lally, best known for Glenroe and Ballykissangel, has died in hospital on Tuesday</td>
</tr>
</tbody>
</table>
# Context-Aware Decoding

## MemoTrap

<table>
<thead>
<tr>
<th>Input</th>
<th>Write a quote that ends in the word “early”. Better late than never early</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular CAD</td>
<td>never</td>
</tr>
<tr>
<td>CAD</td>
<td>early</td>
</tr>
</tbody>
</table>

[Shi et al., 2023]
## Context-Aware Decoding

<table>
<thead>
<tr>
<th>Model</th>
<th>13B</th>
<th>3B</th>
<th>30B</th>
<th>11B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CNN-DM</td>
<td></td>
<td>XSUM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ROUGE-L</td>
<td>factKB</td>
<td>BERT-P</td>
<td>ROUGE-L</td>
</tr>
<tr>
<td>OPT</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular</td>
<td>22.0</td>
<td>77.8</td>
<td>86.5</td>
<td>16.4</td>
</tr>
<tr>
<td>CAD</td>
<td>27.4</td>
<td>84.1</td>
<td>90.8</td>
<td>18.2</td>
</tr>
<tr>
<td>Regular</td>
<td>22.2</td>
<td>81.7</td>
<td>87.0</td>
<td>17.4</td>
</tr>
<tr>
<td>CAD</td>
<td>28.4</td>
<td>87.0</td>
<td>90.2</td>
<td>19.5</td>
</tr>
<tr>
<td>GPT-Neo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular</td>
<td>24.3</td>
<td>80.5</td>
<td>87.5</td>
<td>17.6</td>
</tr>
<tr>
<td>CAD</td>
<td>27.7</td>
<td>87.5</td>
<td>90.6</td>
<td>18.1</td>
</tr>
<tr>
<td>Regular</td>
<td>18.7</td>
<td>68.3</td>
<td>85.2</td>
<td>14.9</td>
</tr>
<tr>
<td>CAD</td>
<td>24.5</td>
<td>77.5</td>
<td>89.4</td>
<td>19.0</td>
</tr>
<tr>
<td>LLaMA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular</td>
<td>27.1</td>
<td>80.2</td>
<td>89.5</td>
<td>19.0</td>
</tr>
<tr>
<td>CAD</td>
<td>32.6</td>
<td>90.8</td>
<td>93.0</td>
<td>21.1</td>
</tr>
<tr>
<td>Regular</td>
<td>25.8</td>
<td>76.8</td>
<td>88.5</td>
<td>18.7</td>
</tr>
<tr>
<td>CAD</td>
<td>31.8</td>
<td>87.8</td>
<td>92.2</td>
<td>22.0</td>
</tr>
<tr>
<td>FLAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regular</td>
<td>25.5</td>
<td>90.2</td>
<td>91.6</td>
<td>18.8</td>
</tr>
<tr>
<td>CAD</td>
<td>26.1</td>
<td>93.9</td>
<td>92.1</td>
<td>19.5</td>
</tr>
<tr>
<td>Regular</td>
<td>25.4</td>
<td>90.4</td>
<td>91.4</td>
<td>19.4</td>
</tr>
<tr>
<td>CAD</td>
<td>27.1</td>
<td>93.1</td>
<td>92.2</td>
<td>20.0</td>
</tr>
</tbody>
</table>

[Shi et al., 2023]