Automatic Text Summarization

Arthur Bražinskas
The University of Edinburgh, Scotland

THE UNIVERSITY OF EDINBURGH

SCHOOL OF DATA ANALYSIS

Yandex

About me

MSc in Artificial Intelligence

UNIVERSITEIT VAN AMSTERDAM

Theoretical machine learning and natural language processing

University of Amsterdam
Amsterdam, Netherlands

ML experience

SAXO BANK

Copenhagen
Denmark

ML experience

SAXO BANK

Copenhagen
Denmark

ELSEVIER

Amsterdam
Netherlands

ML experience

SAXO BANK

Copenhagen
Denmark

ELSEVIER

Amsterdam
Netherlands

zalando

Berlin
Germany

ML experience

SAXO BANK

Copenhagen
Denmark

ELSEVIER

Amsterdam
Netherlands

zalando

Berlin
Germany

amazon

Berlin; Seattle
Germany; USA

Ph.D. in NLP

The University of Edinburgh
Scotland

Supervisors

Ivan Titov

Mirella Lapata

Research topic

- Work on: **abstractive opinion summarization** in **high-** and **low-** resource settings
- Also interested in:
  - reinforcement learning
  - variational inference
  - latent graphical models

Agenda

Agenda

- Introduction to summarization
- How to evaluate summarizers
- News summarization
- Opinion summarization

Summarization: Different Perspectives

Summarization

‘The act of expressing the most **important facts or ideas** about something or someone in a **short and clear form**.’ - *Cambridge dictionary*

Summarization

‘Importance-driven data reduction’

Statistics

Data summarization

- Say we have some continuous data
- Instead of storing the whole dataset
- We can store its ‘**summary**’
- E.g., **sufficient statistics** (Wasserman, 2005), **moments** or **learned parameters**
- **Preference** is given to parameters that allow us to better predict the data

Information Theory

Lossy compression

- Want to **binary represent** and **compress** i.i.d. discrete observations: X ~ F
- Want **reduce** the expected length of the binary string below **H(X)** (optimal code)
- Ok with not being able to decode **some** symbols

Lossy compression

- Represent in the binary format **‘the most important’** symbols or a ‘**summary**’ of symbols
- What symbols are important?
- The ones that are **frequent**

Psychology

Empathic paraphrasing

*A form of responding empathically to the emotions of another person by **repeating in other words** what this person said while **focusing on the essence** of what they feel and **what is important to them**.* (Seehause et al., 2012)

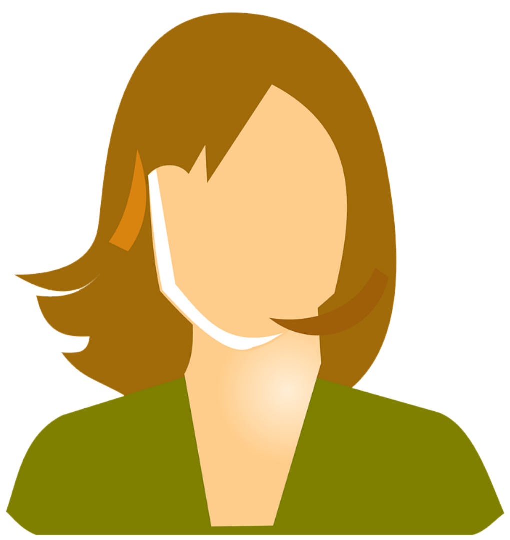
Conceptually similar to **abstractive summarization (reduce, paraphrase, retain what is important)**

Therapy

Therapy

- **Goal**: interpersonal conflict resolution
- Framed as a **dialog game**
- Two persons speak in turns
- Each needs to **summarize** what has been said before continuing the conversation

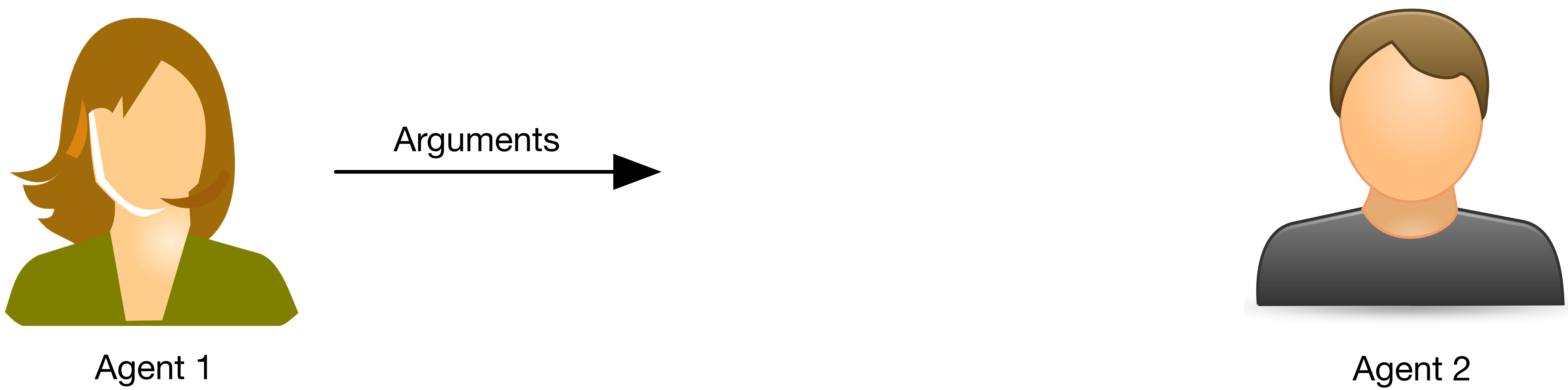
Therapy



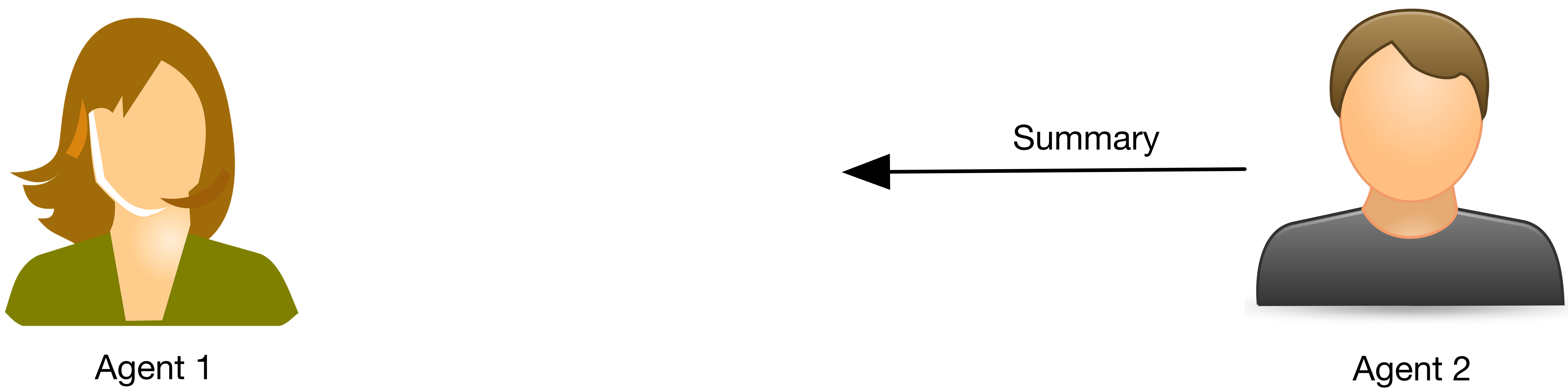
Agent 1

Agent 2

Therapy

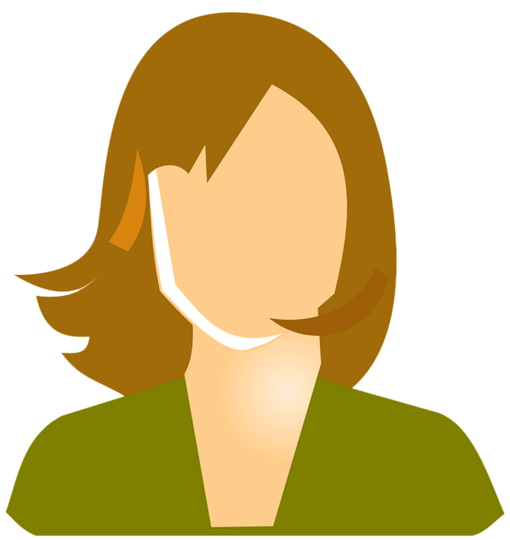


Therapy



Therapy

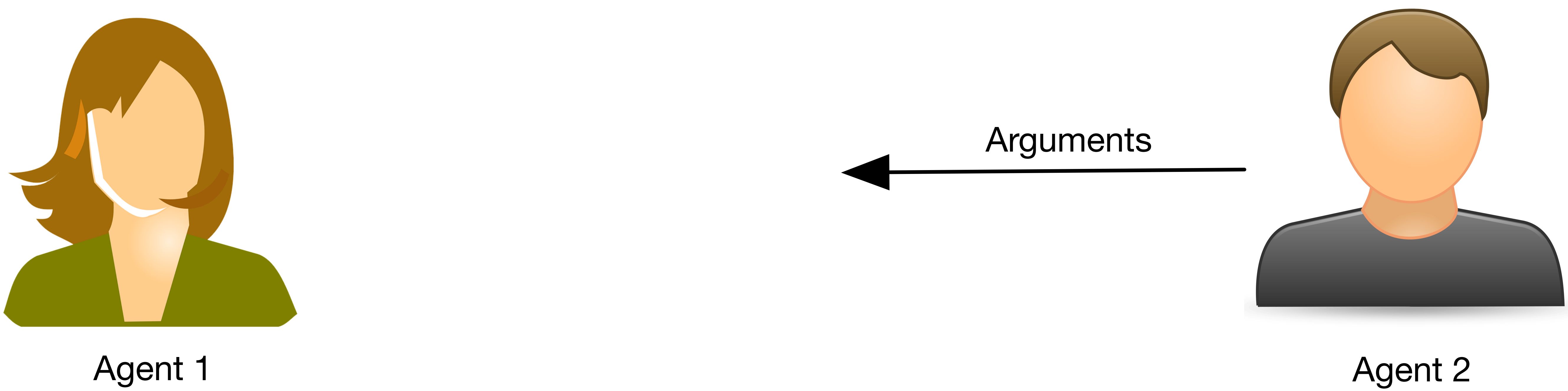
Yes, you’ve understood me



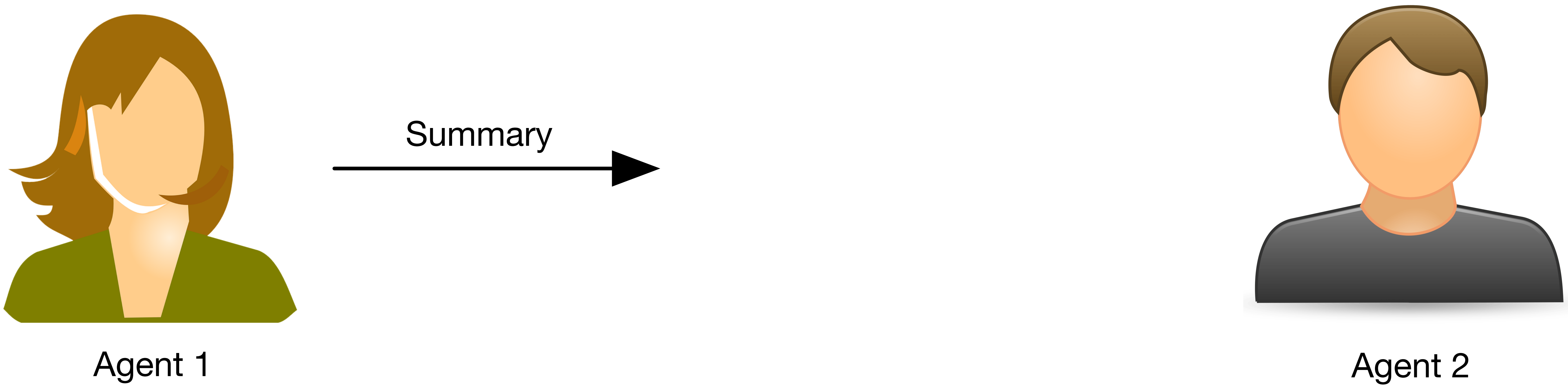
Agent 1

Agent 2

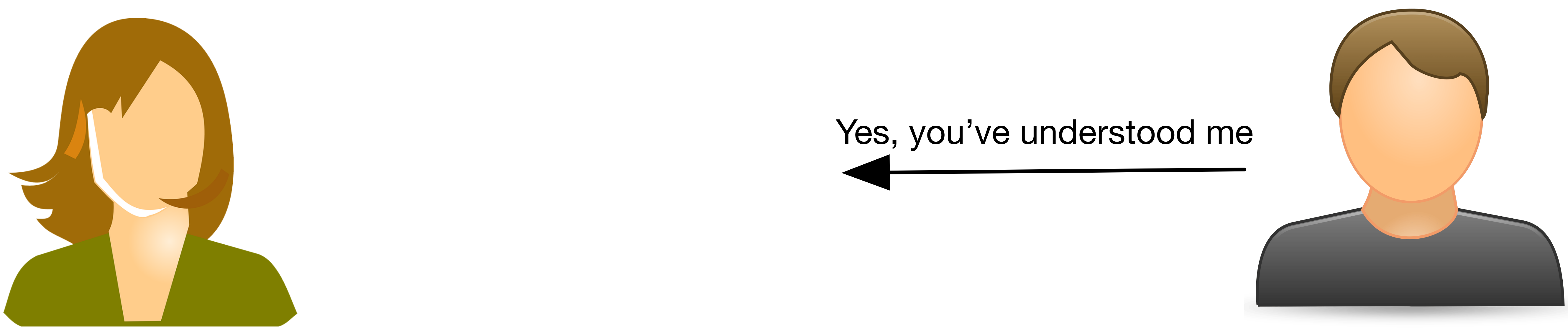
Therapy



Therapy



Therapy



Agent 1

Agent 2

Text Summarization

Summarization applications

- Summarize a **100-page book** to **10 pages**
- Get an overview of a specific event based on recent news articles
- Condense a **wikipedia article** to a **short paragraph** based on a **query**
- Get **contrastive summaries** of multiple products based on user reviews

Summary input types

- Single-document
- Multi-document
- Sentence

Summary Output Types

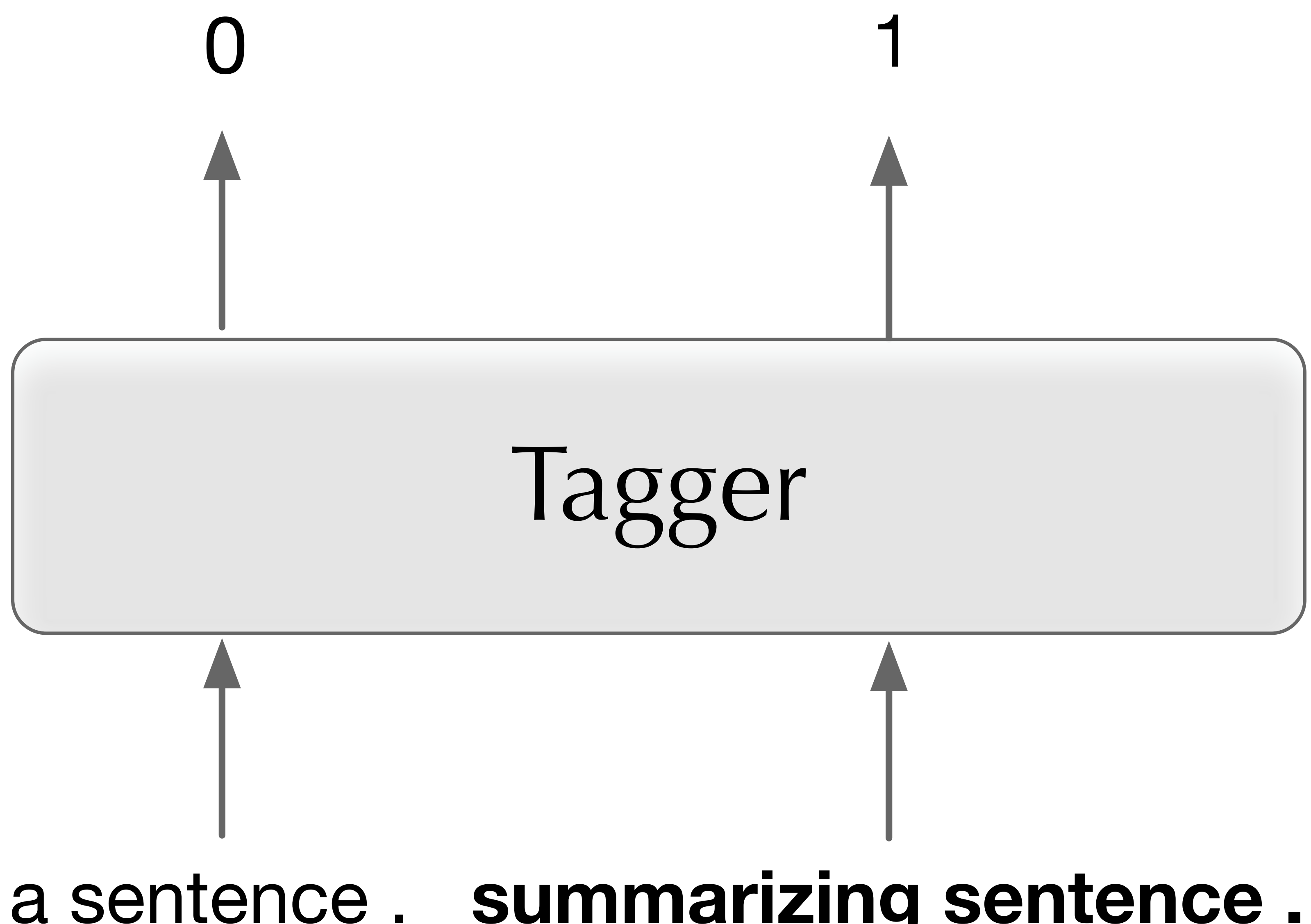
- Extractive
- Abstractive

Extract or Abstract?

Extractive methods

- Well studied across different summarization tasks
- Usually framed as a **tagging problem**:
  - Given a document (s)
  - Select **K summarizing fragments** (e.g., sentences)
  - Concatenate to form a summary

Extractive methods



Extractive methods

- The central challenge is **how to represent sentences**
- We want **powerful** semantic representations that can be used for **accurate** binary classification

Extractive methods

- The tagger is usually a **neural encoder** that produces **sentence semantic representations**
- Such as a Transformer (Vaswani et al., 2017)
- Often pre-trained (Liu and Lapata, 2019)

Extractive methods

- Binary predictions:
  - **linear transformations** of sentence representations
  - sigmoid function

Extractive datasets

- In most cases, we don’t have ‘extractive’ datasets
- Instead, we **utilize abstractive reference summaries** to produce training datasets
- We **select sentences** from the input document that have the **maximum ROUGE score** to the **summary** (Nallapati et al., 2016)
- These are **summarizing sentences**
- Train the extractive summarizer to correctly tag

Extractive methods

Pros:

- Easy-to-build models
- Always factually correct summaries
- Faster training and inference
- Less data demanding

Extractive methods

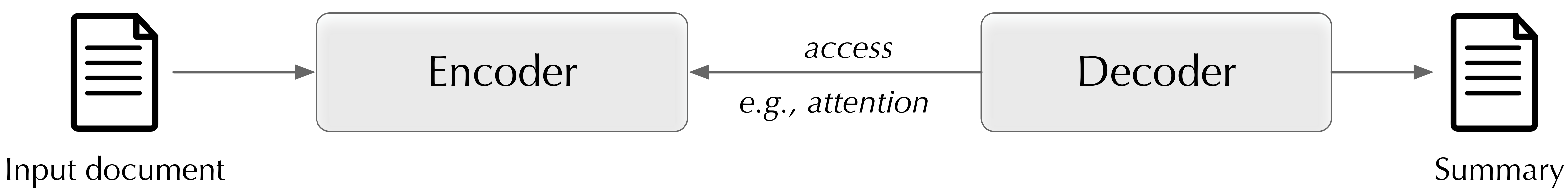
Cons:

- Incoherent output
- ‘Jammed’ unimportant details
- Inability to abstract information
- Limited vocabulary of words

Abstractive methods

- Based on the **encoder-decoder** architecture
- Generate text (Paulus et al., 2017; See et al., 2017; Liu et al., 2018)

Abstractive methods



Abstractive methods

- **Pros:**
  - Can use a **richer vocabulary** of words
  - Can **rephrase** and **abstract**
  - Can deal with **conflicting information**
- **Cons:**
  - Often require **large annotated datasets** for training
  - Prone to **hallucinations** (iPhone vs iPad)

Example

DAGOSTINO'S

Italian

Example

The stake was cold, and the bread was sour. The server forgot about our order.

The waitress was very rude. The pasta was too dry, would not recommend it.

Example

The **stake was cold**, and the **bread was sour**. The **server forgot about our order**.

The **waitress was very rude**. The **pasta was too dry**, would not recommend it.

Example

The **stake was cold**, and the **bread was sour**. The **server forgot about our order**.

The **waitress was very rude**. The **pasta was too dry**, would not recommend it.

Extractive summary: ?

Example

The **stake was cold**, and the **bread was sour**. The **server forgot about our order**.

The waitress was very rude. The **pasta was too dry**, would not recommend it.

Extractive summary: The **server forgot about our order**. The **pasta was too dry**, would not recommend it.

Example

The **stake was cold**, and the **bread was sour**. The **server forgot about our order**.

The **waitress was very rude**. The **pasta was too dry**, would not recommend it.

Abstractive summary: Both the service and food are terrible.

Example

The **stake was cold**, and the **bread was sour**. The **server forgot about our order**.

The **waitress was very rude**. The **pasta was too dry**, would not recommend it.

Abstractive summary: Both the **service** and **food** are terrible.

Evaluation

ROUGE

- The status-quo metric (Lin, 2004)
- N-gram overlap between the **reference** and **hypothesis** summary

ROUGE-N

- Recall: $$\frac{|\text{ngrams}(ref)\ \&\ \text{ngrams}(hyp)|}{|\text{ngrams}(ref)|}$$

- Precision: $$\frac{|\text{ngrams}(ref)\ \&\ \text{ngrams}(hyp)|}{|\text{ngrams}(hyp)|}$$

- F1: $$2\frac{P * R}{R + P}$$

ROUGE-N

- Recall: $\dfrac{|\text{ngrams}(ref) \ \& \ \text{ngrams}(hyp)|}{|\text{ngrams}(ref)|}$

- Precision: $\dfrac{|\text{ngrams}(ref) \ \& \ \text{ngrams}(hyp)|}{|\text{ngrams}(hyp)|}$

- F1: $2\dfrac{P * R}{R + P}$ **(reported results are in F1)**

ROUGE-L

- Based on the longest common subsequence
- Gaps are allowed
- **The most important sub-metric** in summarization
- **Correlated with fluency** (harder for extractive systems to score highly)

ROUGE: shortcomings

- Not sensitive to **factual mistakes** (Falke et al., 2019; Maynez et al., 2020; Bražinskas et al., 2020)

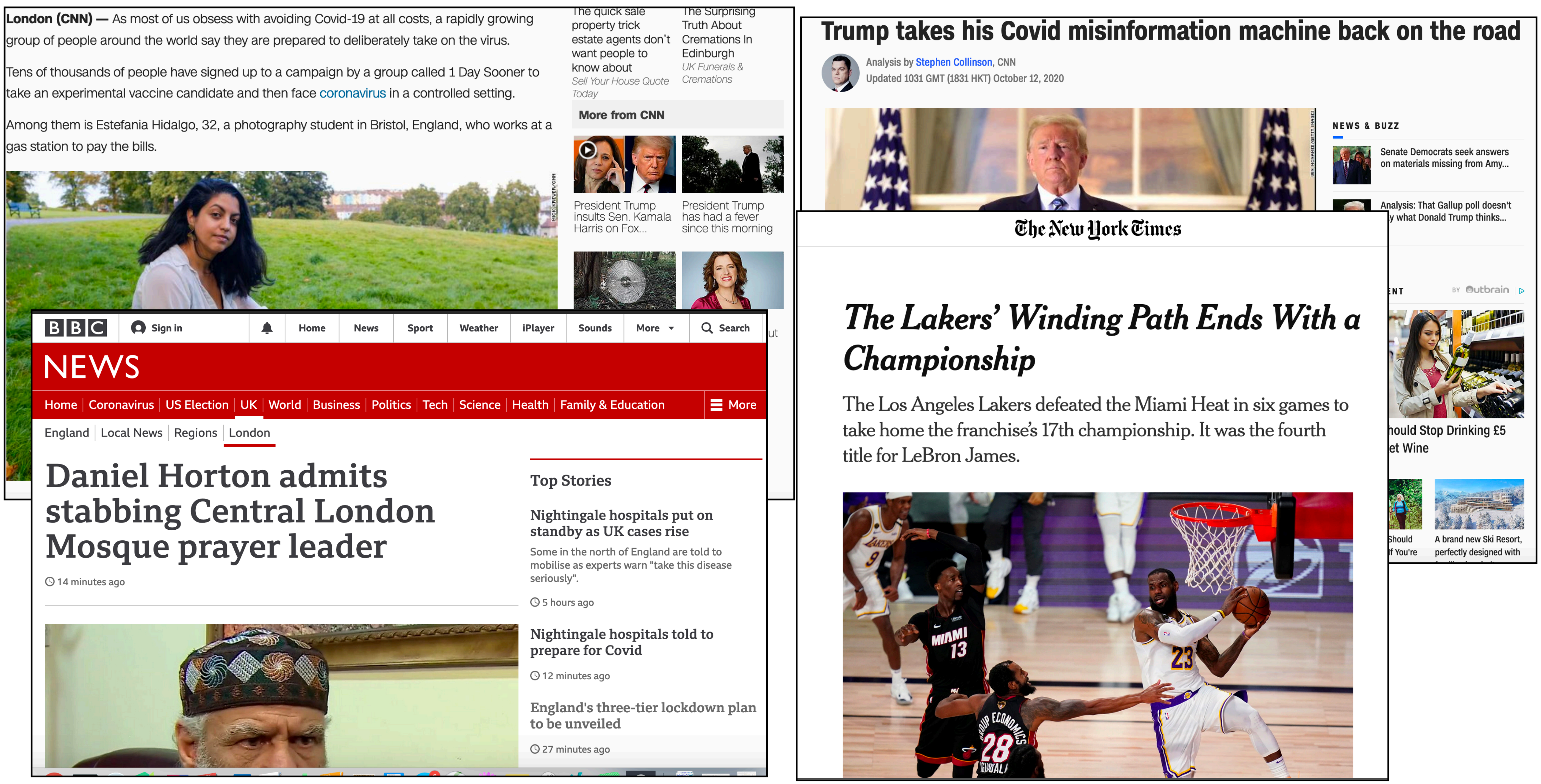
- Not sensitive to **flipped sentiment** (Tay et al., 2019)

Human Evaluation

- Often used to address the ROUGE shortcomings
- Hired workers (e.g., AMT) assess summaries based on various criteria
- Extensively used in opinion summarization

News Summarization

News


London (CNN) — As most of us obsess with avoiding Covid-19 at all costs, a rapidly growing group of people around the world say they are prepared to deliberately take on the virus.
Tens of thousands of people have signed up to a campaign by a group called 1 Day Sooner to take an experimental vaccine candidate and then face coronavirus in a controlled setting.
Among them is Estefania Hidalgo, 32, a photography student in Bristol, England, who works at a gas station to pay the bills.
The quick sale property trick estate agents don't want people to know about
Sell Your House Quote Today
The Surprising Truth About Cremations In Edinburgh
UK Funerals & Cremations
More from CNN
President Trump insults Sen. Kamala Harris on Fox...
President Trump has had a fever since this morning
BBC
Sign in
Home
News
Sport
Weather
iPlayer
Sounds
More
Search
NEWS
Home | Coronavirus | US Election | UK | World | Business | Politics | Tech | Science | Health | Family & Education
More
England | Local News | Regions | London
Daniel Horton admits stabbing Central London Mosque prayer leader
14 minutes ago
Top Stories
Nightingale hospitals put on standby as UK cases rise
Some in the north of England are told to mobilise as experts warn "take this disease seriously".
5 hours ago
Nightingale hospitals told to prepare for Covid
12 minutes ago
England's three-tier lockdown plan to be unveiled
27 minutes ago
Trump takes his Covid misinformation machine back on the road
Analysis by Stephen Collinson, CNN
Updated 1031 GMT (1831 HKT) October 12, 2020
NEWS & BUZZ
Senate Democrats seek answers on materials missing from Amy...
Analysis: That Gallup poll doesn't ...y what Donald Trump thinks...
The New York Times
The Lakers' Winding Path Ends With a Championship
The Los Angeles Lakers defeated the Miami Heat in six games to take home the franchise's 17th championship. It was the fourth title for LeBron James.
A brand new Ski Resort, perfectly designed with

Summarization of news

Input article

Summarization of news

Input article

~700 words

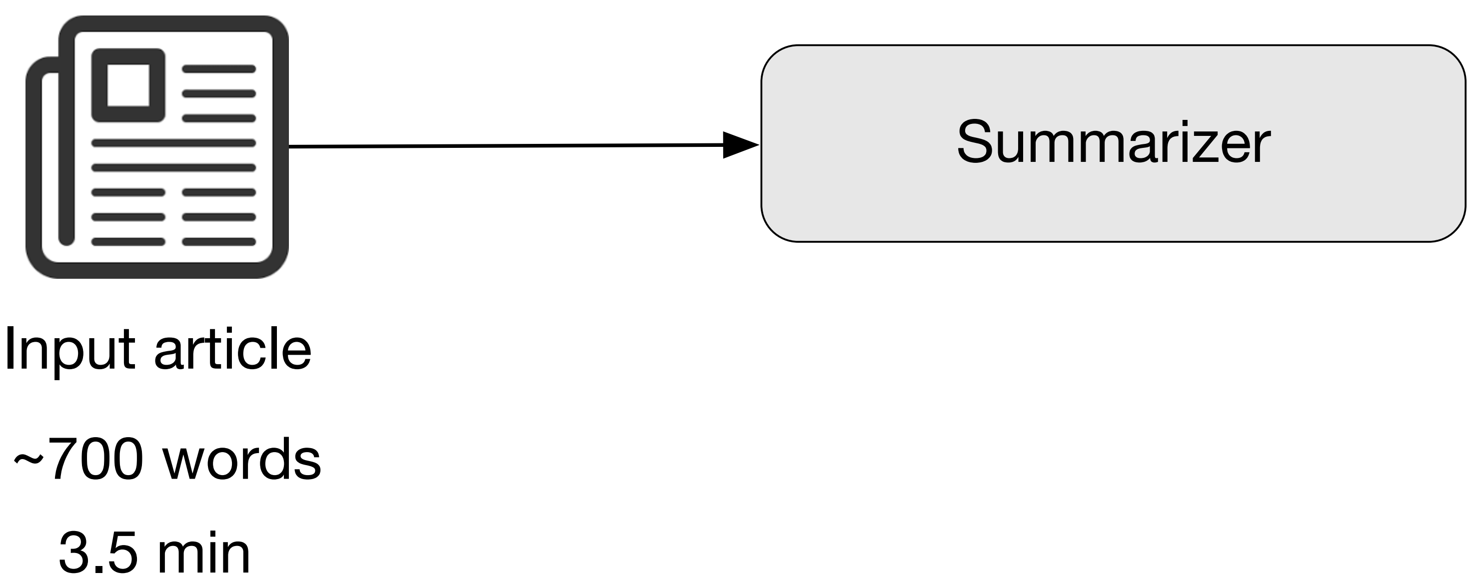
Summarization of news

Input article

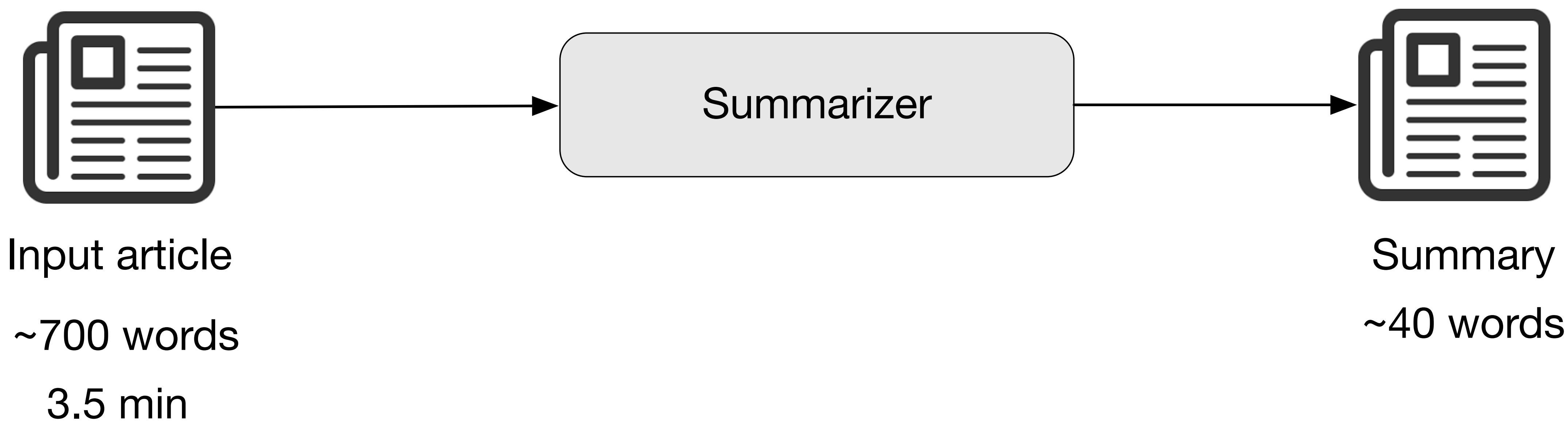
~700 words

3.5 min

Summarization of news



Summarization of news



Summarization of news

Input article

~700 words

3.5 min

Summary

~40 words

12 sec

News summarization

- Often synonymous to summarization
- A well established branch
- Large datasets for supervised training
- A large body of research (models and theories)
- Mostly **single document**

Datasets

Name	Multidoc?	# pairs	#words summary	Note
CNN/DM	No	312k	56.20	Main one; highly extractive
NYT	No	654k	45.54	Highly extractive; behind the pay wall
XSum	No	230k	23.26	Abstractive; issues with content support
Newsroom	No	1.3M	26.7	Diverse; noisy; scraped from the web
Multi-news	Yes	56k	263.66	First large multi-doc

CNN Example

CNN politics 2020 Election Facts First Election 101

What we learned from Donald Trump in 2015

By Stephen Collinson, CNN
Updated 0051 GMT (0851 HKT) December 31, 2015

How Donald Trump proved critics wrong in 2015 02:08

STORY HIGHLIGHTS

- Trump insists he is not a politician, but he was the most accomplished politician in the Republican field for much of 2015
- Trump's not just a master of social media; he also plays the traditional media establishment like no one else

Washington (CNN) — He's churned up torrents of insults, incited grass-roots Republican fury, fearlessly flouted taboos on gender, race and religion and confounded the pundits again and again.

In a riotous six-month carnival of political incorrectness, Donald Trump has fused his message to the mood of his seething supporters like no other candidate and defied

CNN Example

CNN politics 2020 Election Facts First Election 101

What we learned from Donald Trump in 2015

By Stephen Collinson, CNN
Updated 0051 GMT (0851 HKT) December 31, 2015

How Donald Trump proved critics wrong in 2015 02:08

source document

STORY HIGHLIGHTS

- Trump insists he is not a politician, but he was the most accomplished politician in the Republican field for much of 2015
- Trump's not just a master of social media; he also plays the traditional media establishment like no one else

Washington (CNN) — He's churned up torrents of insults, incited grass-roots Republican fury, fearlessly flouted taboos on gender, race and religion and confounded the pundits again and again.

In a riotous six-month carnival of political incorrectness, Donald Trump has fused his message to the mood of his seething supporters like no other candidate and defied

CNN Example

CNN politics 2020 Election Facts First Election 101

What we learned from Donald Trump in 2015

By Stephen Collinson, CNN
Updated 0051 GMT (0851 HKT) December 31, 2015

How Donald Trump proved critics wrong in 2015 02:08

summary

STORY HIGHLIGHTS

- Trump insists he is not a politician, but he was the most accomplished politician in the Republican field for much of 2015
- Trump's not just a master of social media; he also plays the traditional media establishment like no one else

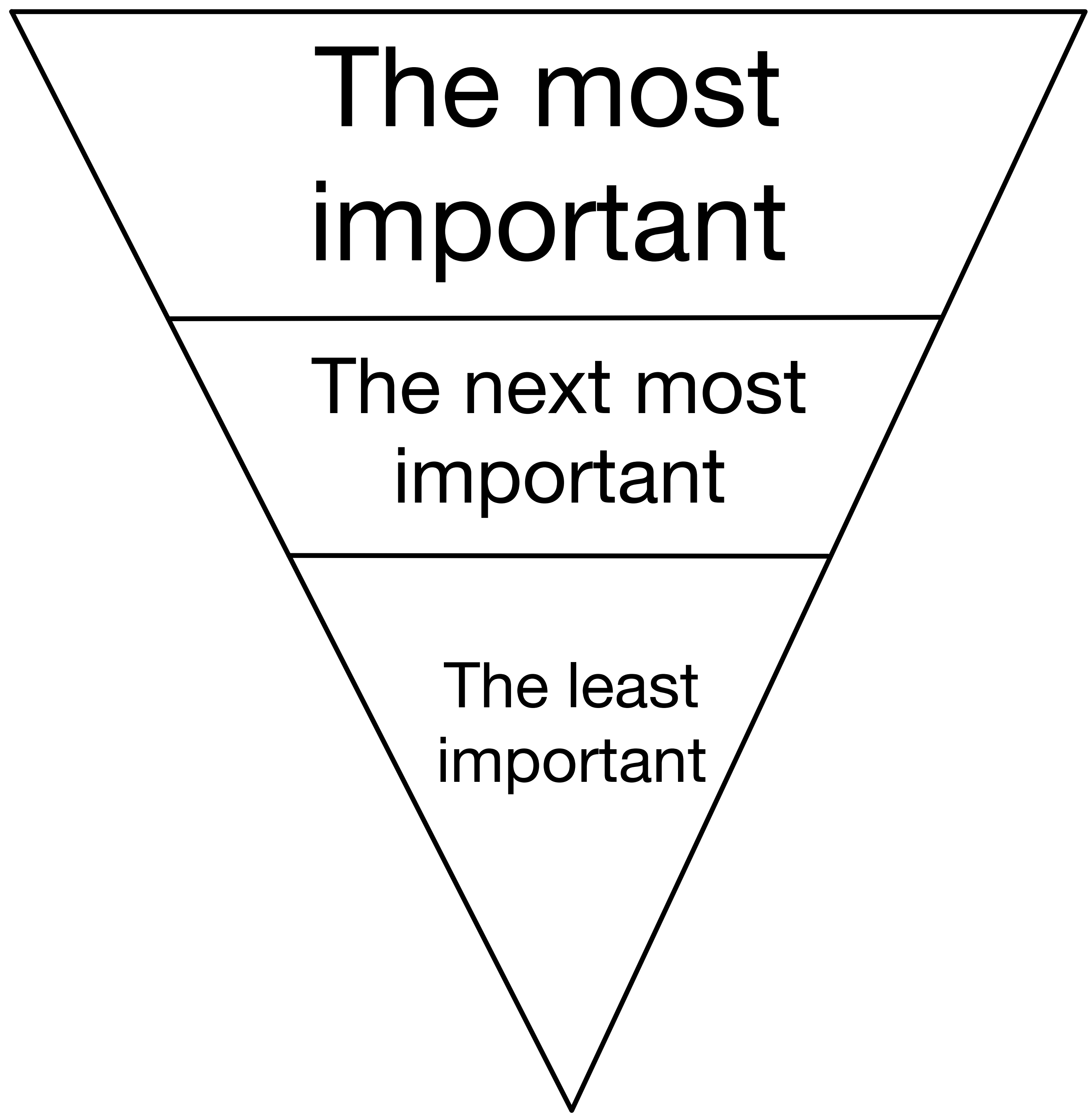
Washington (CNN) — He's churned up torrents of insults, incited grass-roots Republican fury, fearlessly flouted taboos on gender, race and religion and confounded the pundits again and again.

In a riotous six-month carnival of political incorrectness, Donald Trump has fused his message to the mood of his seething supporters like no other candidate and defied

Single document summarization

- The model needs to learn **a notion** of **importance**
- For example, to attend important text segments
- Often **can’t** take an advantage of **redundancies**

Inverted pyramid of importance



LEAD-3

- Can select **top-3 sentences** and form a summary (*LEAD-3*)
- For a long time, *LEAD-3* was an **unbeatable baseline** across different datasets

CNN/DM

Model	Type	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	Ext	**40.42**	**17.62**	**36.67**
SummaRunner (Nallapati et al., 2016)	Abs	37.50	14.50	33.40
SummaRunner (Nallapati et al., 2016)	Ext	39.60	16.20	35.30

Pointer-Generator Network

Abigail See, Peter Liu, and Christopher Manning

Pointer-generator network

- Addresses two main problems:
  - Inaccurate generation of details
  - Repetitions
- **Augment** the **standard attention module**
- Introduces a loss for coverage (*not covered in details*)

Attention mechanism

- Introduced as a way to alleviate the inability of seq2seq models to accurately decode **target sequences** from continues representations of **source sequences** (Bahdanau et al., 2014)
- The **decoder** gets access to a **context vector**
- The context vector is a **weighted sum of the encoder hidden states**

Attention mechanism

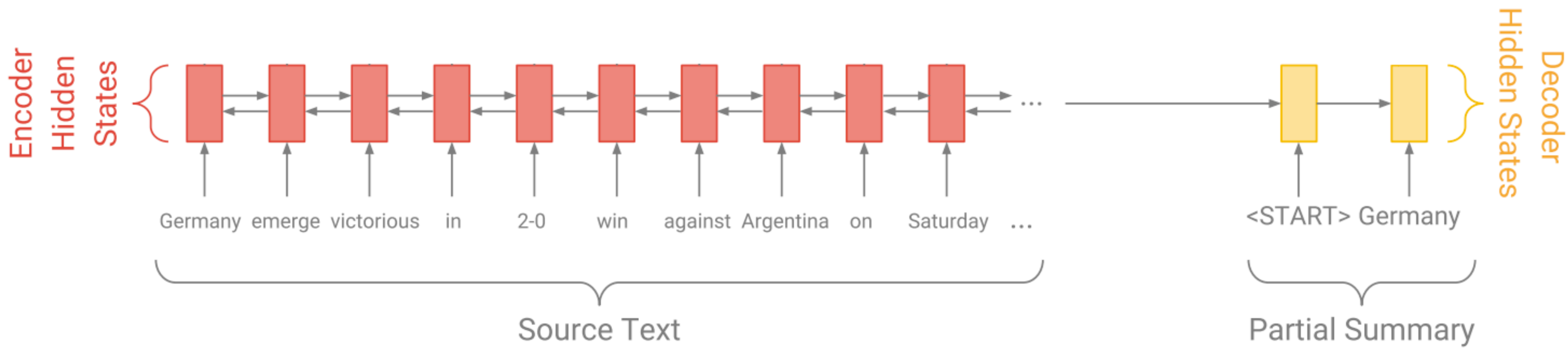
Germany emerge victorious in 2-0 win against Argentina on Saturday ...

Source Text

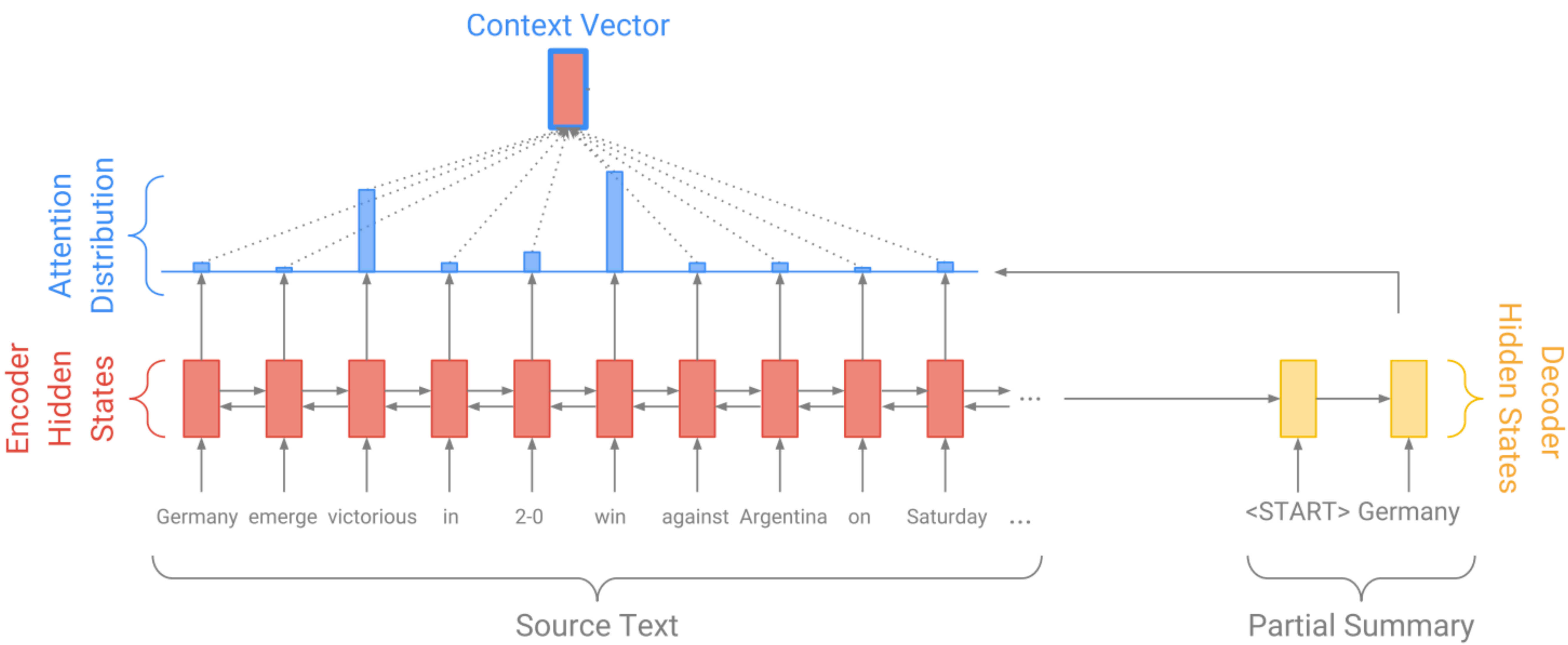
<START> Germany

Partial Summary

Attention mechanism



Attention mechanism



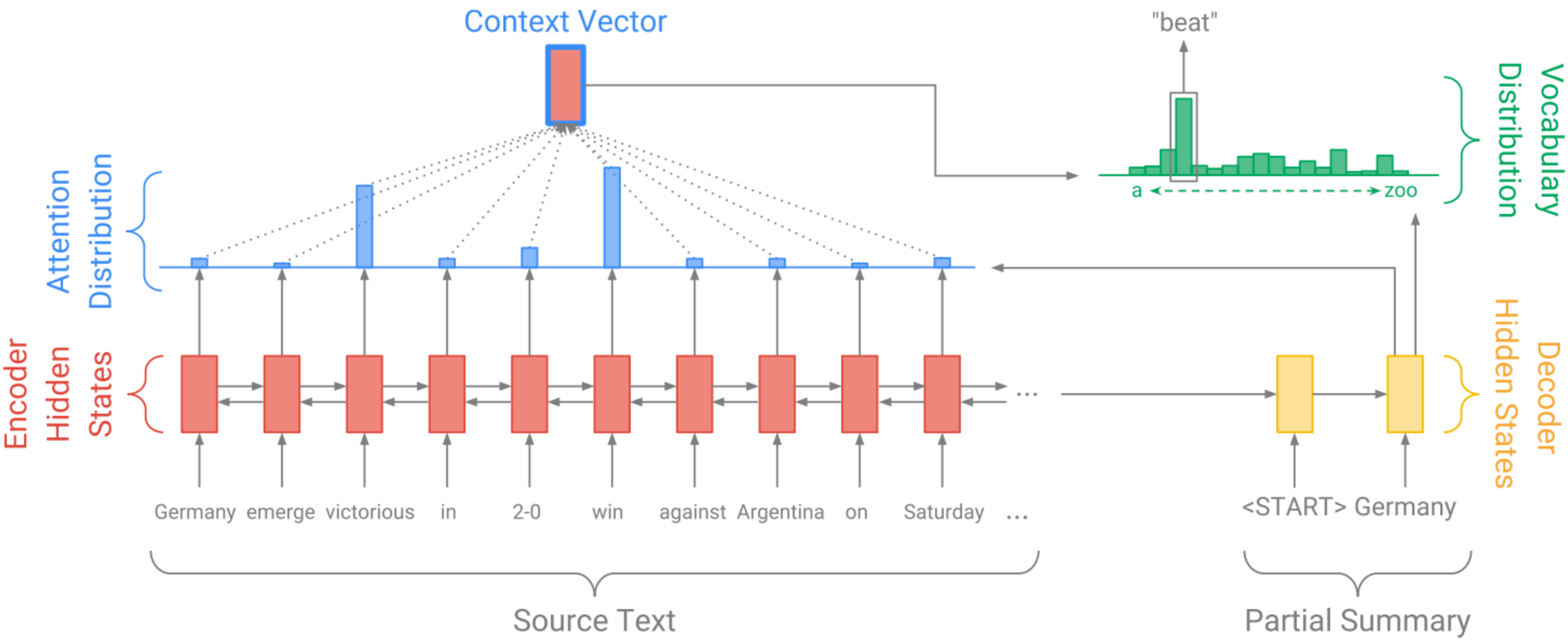
Context vector

Encoder hidden states

$$h_t^* = \sum_i a_i^t h_i$$

Attention weights

Attention mechanism



Attention mechanism

Decoder hidden states

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

Attention mechanism

Decoder hidden states

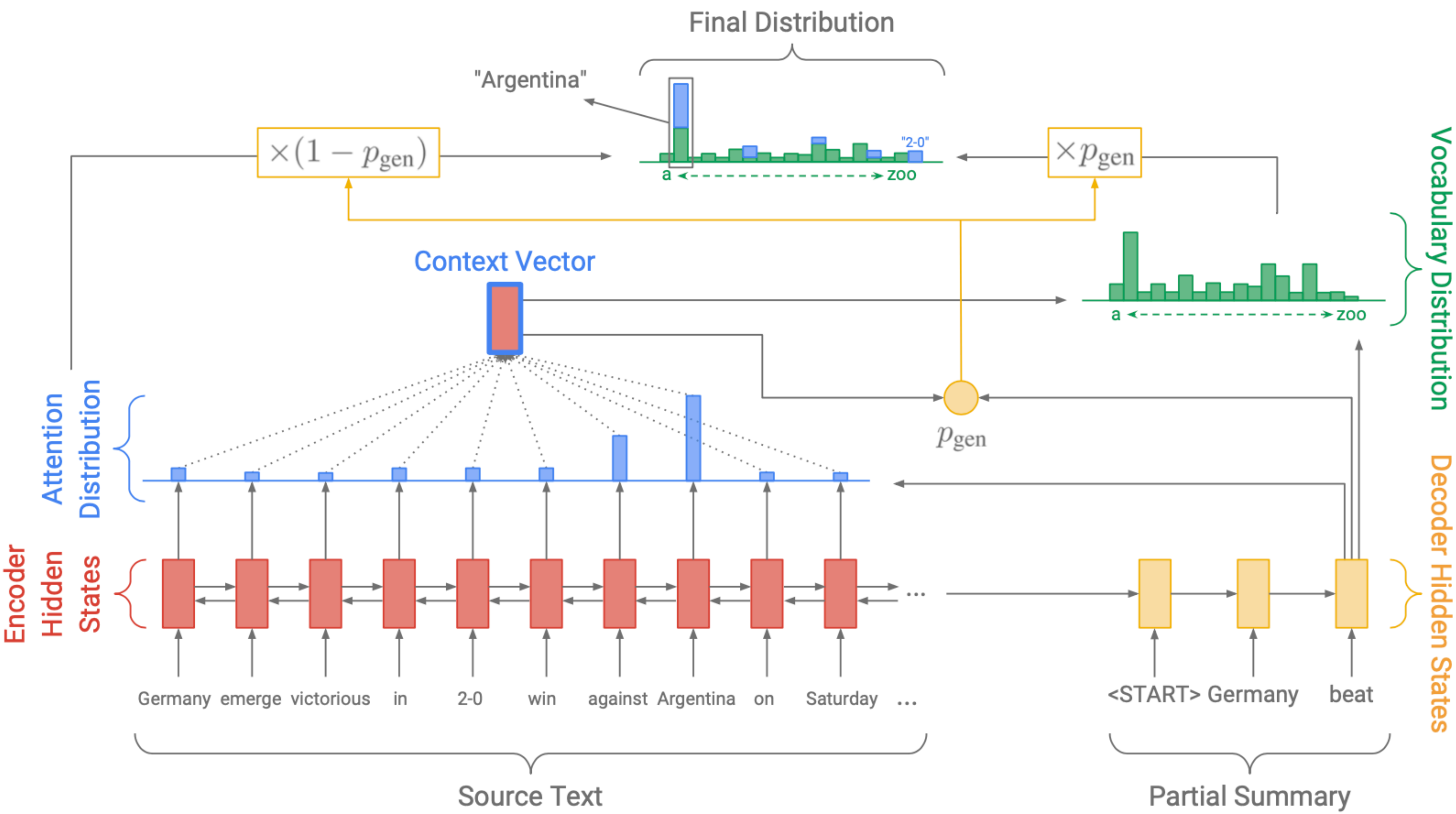
$$P_{\text{vocab}} = \text{softmax}(V'(V[s_t, h_t^*] + b) + b')$$

Context vector

Copy mechanism

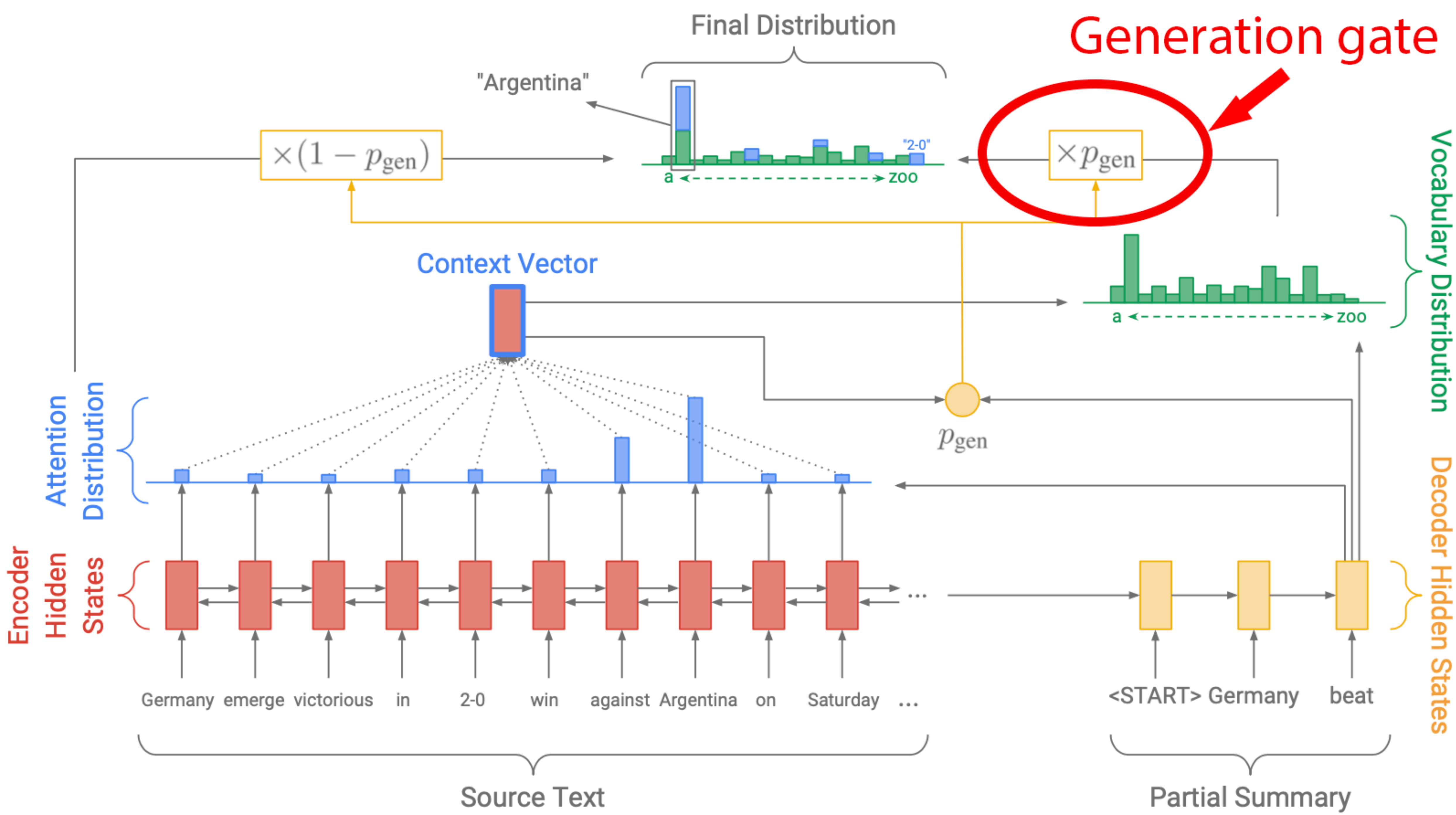
- Directly copies words from the source via a **pointer network** (Vinyals et al., 2015)
- Reuses attention weights
- Useful for the **OOV** words problem
- The final word distribution combines **generation** and **‘copy’** word distributions

Full model



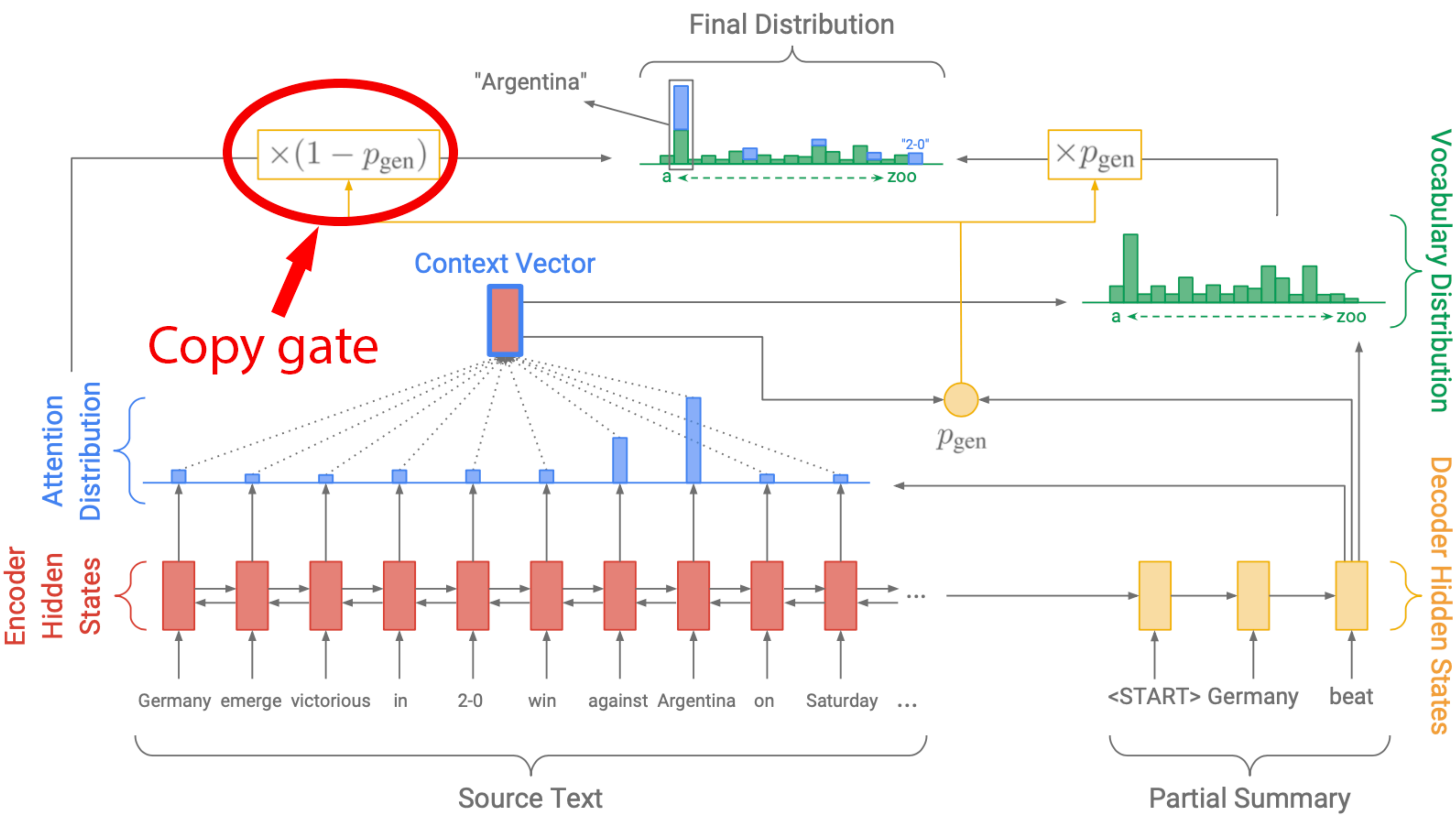
Final Distribution
"Argentina"
"2-0"
a
zoo
$\times(1 - p_{\text{gen}})$
$\times p_{\text{gen}}$
Context Vector
Vocabulary Distribution
a
zoo
p_{gen}
Attention Distribution
Encoder Hidden States
Decoder Hidden States
Germany emerge victorious in 2-0 win against Argentina on Saturday ...
<START> Germany beat
Source Text
Partial Summary

Full model



Final Distribution
"Argentina"
"2-0"
a
zoo
$\times(1-p_{\text{gen}})$
$\times p_{\text{gen}}$
Generation gate
Context Vector
Vocabulary Distribution
a
zoo
p_{gen}
Attention Distribution
Encoder Hidden States
Decoder Hidden States
Germany
emerge
victorious
in
2-0
win
against
Argentina
on
Saturday
...
<START>
Germany
beat
Source Text
Partial Summary

Full model



Final Distribution
"Argentina"
"2-0"
a
zoo
$\times(1-p_{\text{gen}})$
$\times p_{\text{gen}}$
Copy gate
Context Vector
Vocabulary Distribution
a
zoo
p_{gen}
Attention Distribution
Encoder Hidden States
Decoder Hidden States
Germany
emerge
victorious
in
2-0
win
against
Argentina
on
Saturday
...
<START>
Germany
beat
Source Text
Partial Summary

Gate

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

Context vector

Gate

Decoder hidden state

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

Context vector

Gate

Decoder hidden state

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

Context vector

Current word embedding

Gate

Decoder hidden state

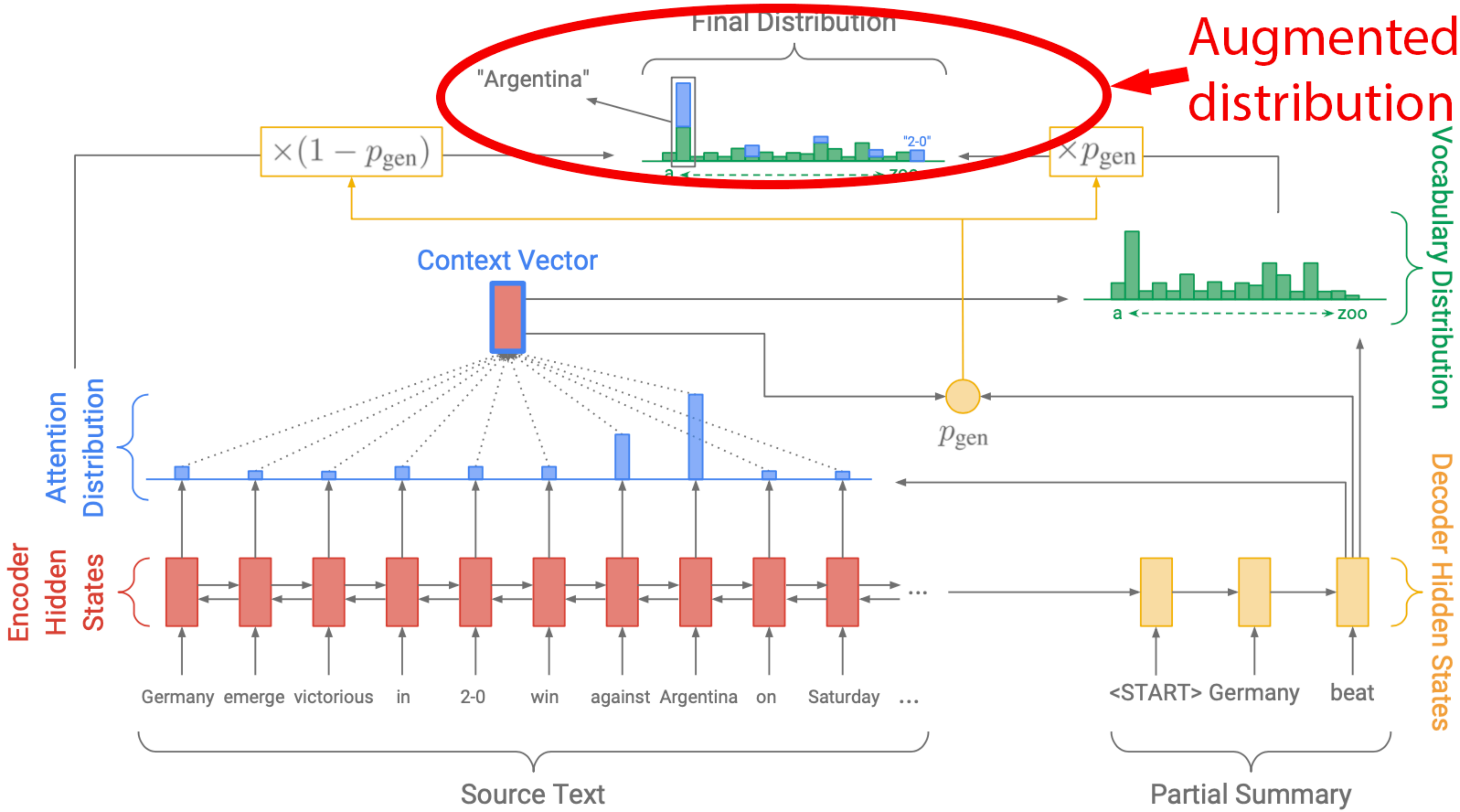
Bias

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

Context vector

Current word embedding

Full model



Final Distribution
"Argentina"
"2-0"
a
zoo
$\times(1-p_{\text{gen}})$
$\times p_{\text{gen}}$
Augmented distribution
Context Vector
Vocabulary Distribution
a
zoo
p_{gen}
Attention Distribution
Encoder Hidden States
Decoder Hidden States
Germany emerge victorious in 2-0 win against Argentina on Saturday ...
<START> Germany beat
Source Text
Partial Summary

Final distribution

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

Final distribution

Generation distribution

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

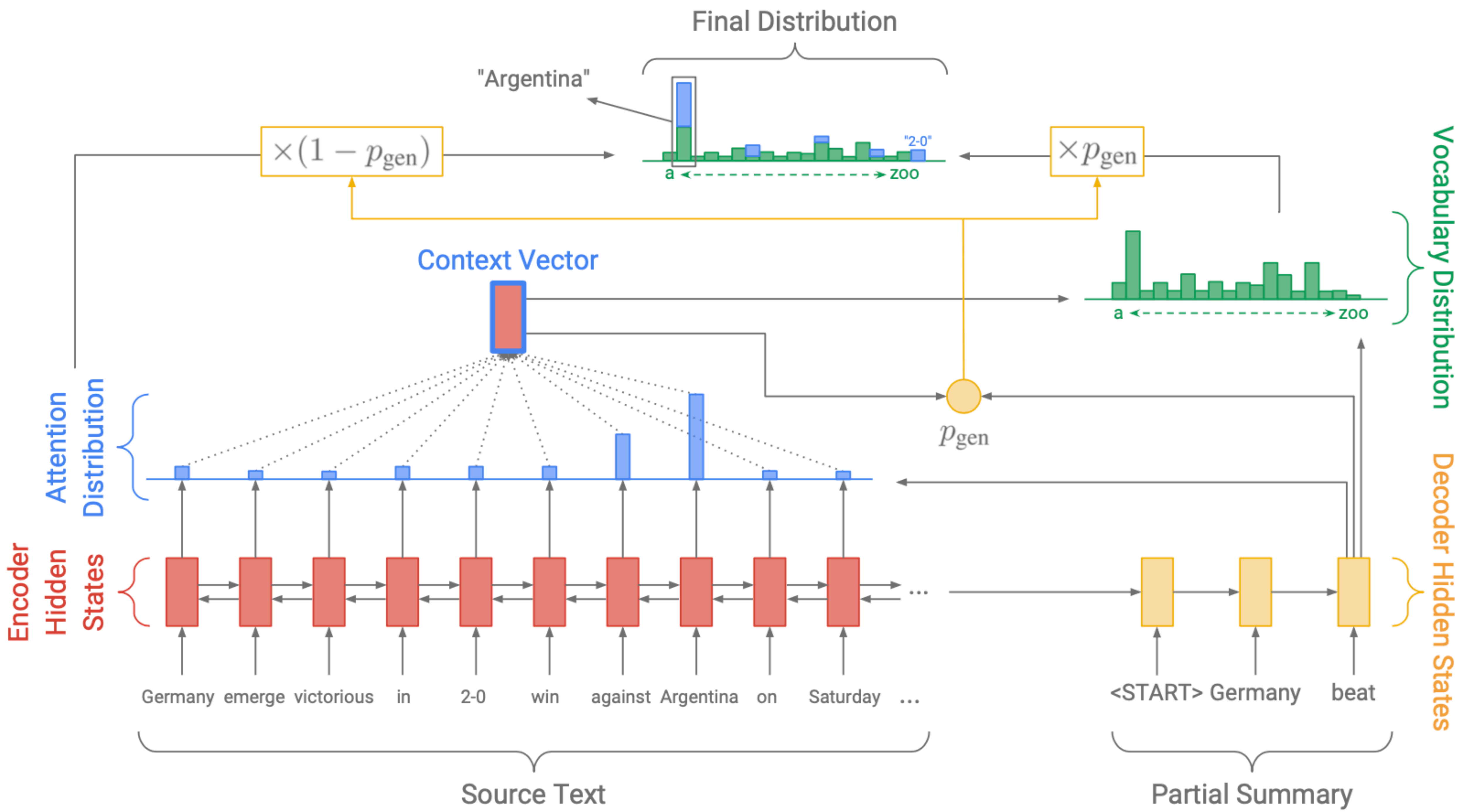
Final distribution

Generation distribution

Copy distribution

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

Full model



Final Distribution
"Argentina"
"2-0"
a
zoo
$\times(1-p_{\text{gen}})$
$\times p_{\text{gen}}$
Context Vector
Vocabulary Distribution
a
zoo
p_{gen}
Attention Distribution
Encoder Hidden States
Decoder Hidden States
Germany
emerge
victorious
in
2-0
win
against
Argentina
on
Saturday
...
<START>
Germany
beat
Source Text
Partial Summary

CNN/DM

Model	Type	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	Ext	**40.42**	**17.62**	**36.67**
SummaRunner (Nallapati et al., 2016)	Abs	37.50	14.50	33.40
SummaRunner (Nallapati et al., 2016)	Ext	39.60	16.20	35.30
PTGEN+COV (See et al., 2017)	Abs	39.53	17.28	36.38

Bottom-Up Abstractive Summarization

Sebastian Gehrmann, Yuntian Deng, Alexander Rush

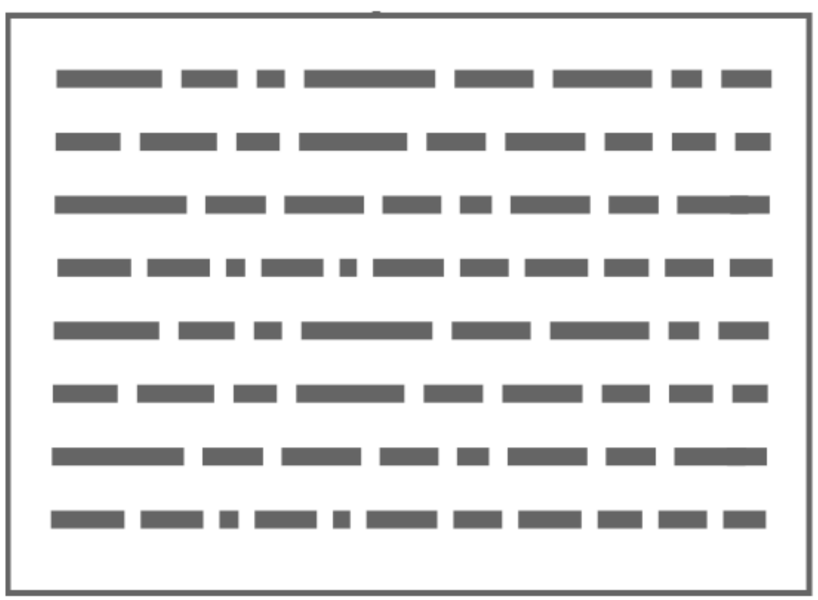
BottomUP

- Builds on top of the PGN model
- Address the **poor selection of words** via the attention
- Train **a separate content selector** of words
- **Hard mask** not important words
- **Augment the copy attention distribution** at test time (inference) to copy only words that are not masked

Models

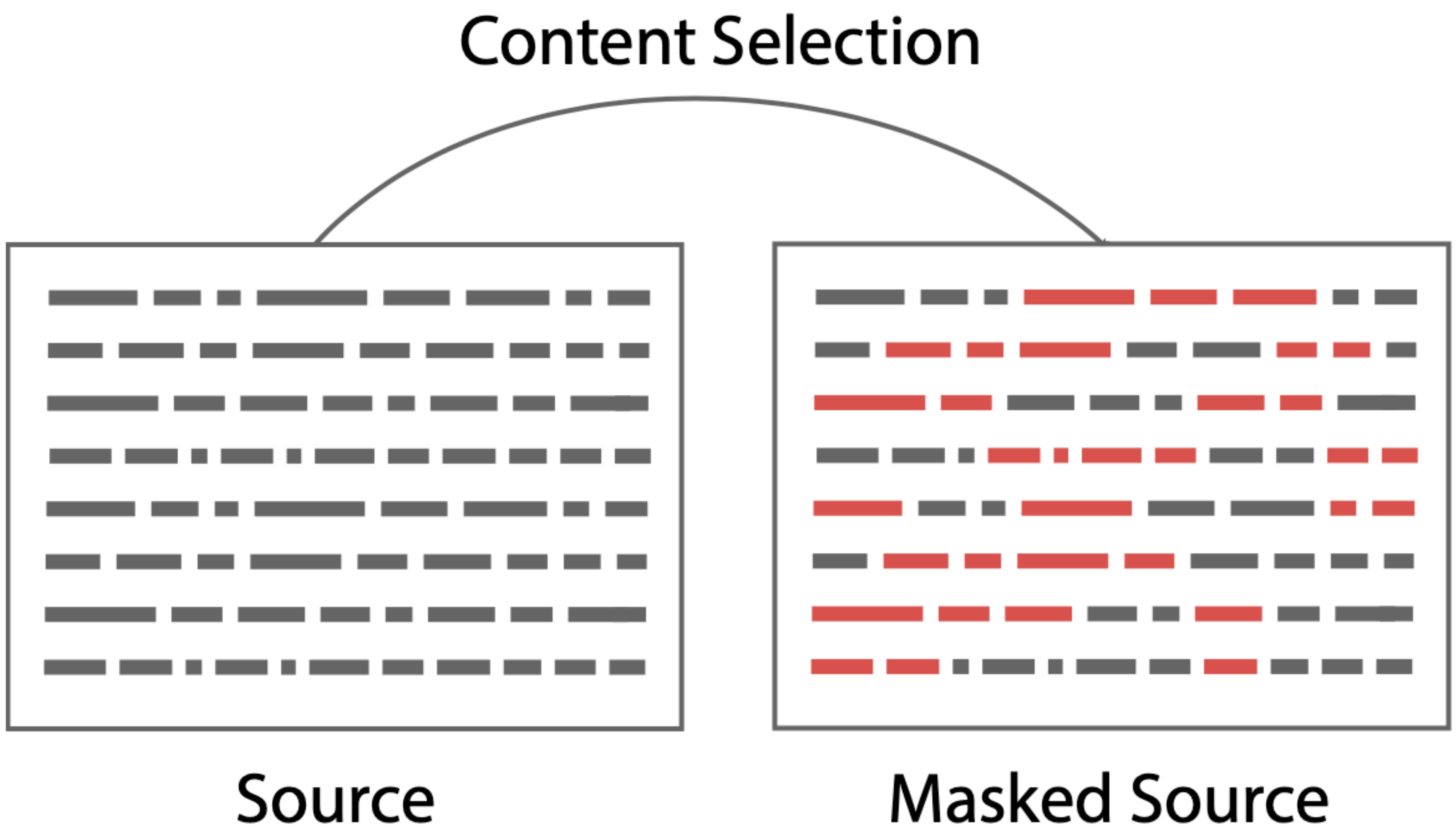
- **Content selector:**
  - GloVe (Pennington et al., 2014)
  - ELMo (character-aware token embeddings + bi-LSTM layers) (Peters et al., 2018)
  - bi-LSTM
  - Linear projection + sigmoid
- **Main model:**
  - Pointer-generator network (See et al., 2018)

Two-step procedure

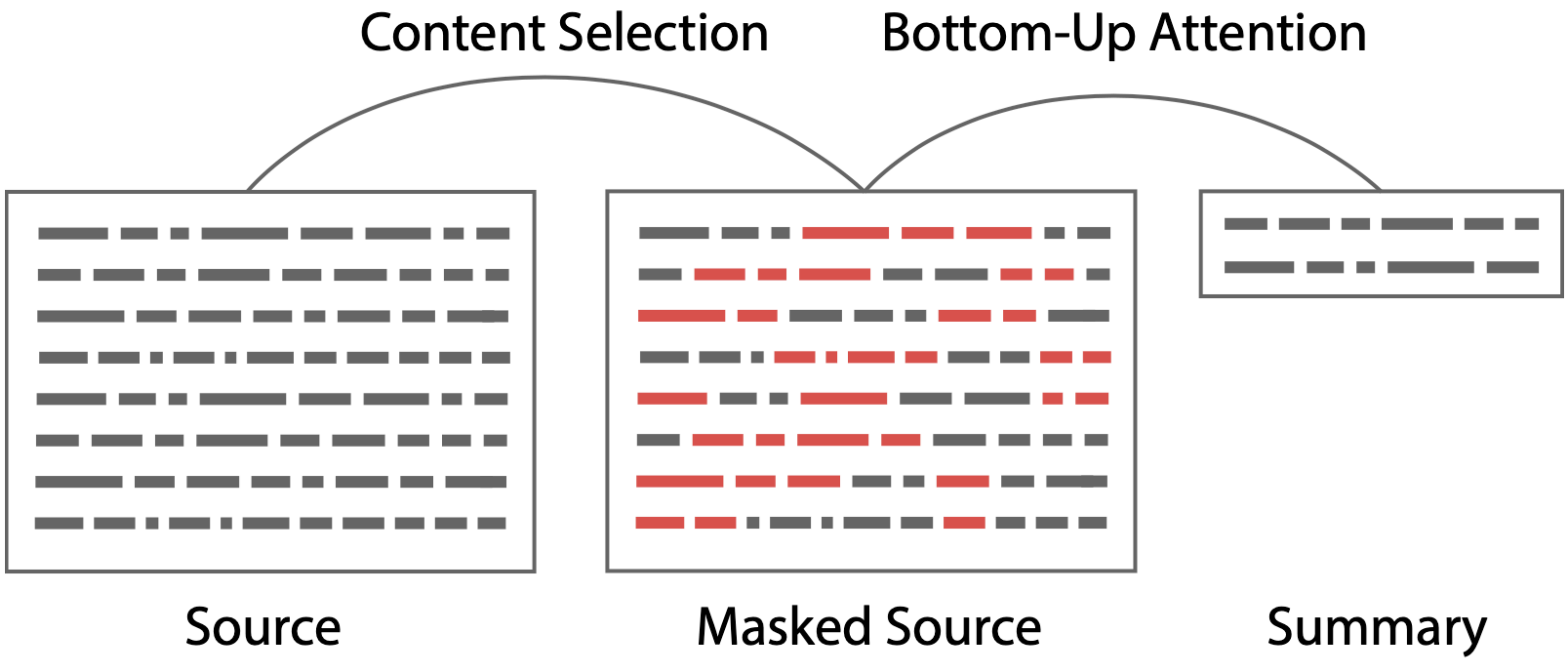


Source

Two-step procedure



Two-step procedure



Augmented copy distribution

$$p(\tilde{a}_j^i | x, y_{1:j-1}) = \begin{cases} p(a_j^i | x, y_{1:j-1}) & q_i > \epsilon \\ 0 & \text{ow.} \end{cases}$$

Augmented copy distribution

$$p(\tilde{a}_j^i|x, y_{1:j-1}) = \begin{cases} p(a_j^i|x, y_{1:j-1}) & q_i > \epsilon \\ 0 & \text{ow.} \end{cases}$$

source words

Augmented copy distribution

current prefix words

$$p(\tilde{a}_j^i|x, y_{1:j-1}) = \begin{cases} p(a_j^i|x, y_{1:j-1}) & q_i > \epsilon \\ 0 & \text{ow.} \end{cases}$$

source words

Augmented copy distribution

current prefix words

$$p(\tilde{a}_j^i|x, y_{1:j-1}) = \begin{cases} p(a_j^i|x, y_{1:j-1}) & q_i > \epsilon \\ 0 & \text{ow.} \end{cases}$$

source words

attention probability

Augmented copy distribution

selector probability

current prefix words

$$p(\tilde{a}^i_j | x, y_{1:j-1}) = \begin{cases} p(a^i_j | x, y_{1:j-1}) & q_i > \epsilon \\ 0 & \textbf{ow.} \end{cases}$$

source words

attention probability

Augmented copy distribution

selector probability

current prefix words

$$p(\tilde{a}_j^i | x, y_{1:j-1}) = \begin{cases} p(a_j^i | x, y_{1:j-1}) & q_i > \epsilon \\ 0 & \text{ow.} \end{cases}$$

threshold

source words

attention probability

Augmentation at inference

- This augmentation is performed **at inference**
- Show that **joint training** does not substantially improve the performance

CNN/DM

Model	Type	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	Ext	40.42	17.62	36.67
SummaRunner (Nallapati et al., 2016)	Abs	37.50	14.50	33.40
SummaRunner (Nallapati et al., 2016)	Ext	39.60	16.20	35.30
PTGEN+COV (See et al., 2017)	Abs	39.53	17.28	36.38
BottomUP (Gehrmann et al., 2018)	Abs	**41.22**	**18.68**	**38.34**

News Summarization: Modern Approach

Two-step paradigm

- **Pre-training:**
  - Large language models trained on **unannotated** datasets
  - **Unsupervised objectives**, such as masked predictions (Devlin et al., 2018; Radford et al., 2018; Lewis et al., 2020)
- **Fine-tuning:**
  - Task specific datasets
  - Supervised learning

BertSum

- Based on a pre-trained encoder (Liu and Lapata, 2019)
- Use a pre-trained **BERT encoder** (Devlin et al., 2019)
- Transformer **encoder-decoder** architecture
- The **decoder** is **trained** from **scratch**

CNN/DM

Model	Type	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	Ext	40.42	17.62	36.67
BottomUP (Gehrmann et al., 2018)	Abs	41.22	18.68	38.34
\wo BERT (Liu and Lapata, 2019)	Abs	40.21	17.76	37.09
\w BERT (Liu and Lapata, 2019)	Abs	**41.72**	**19.39**	**38.76**

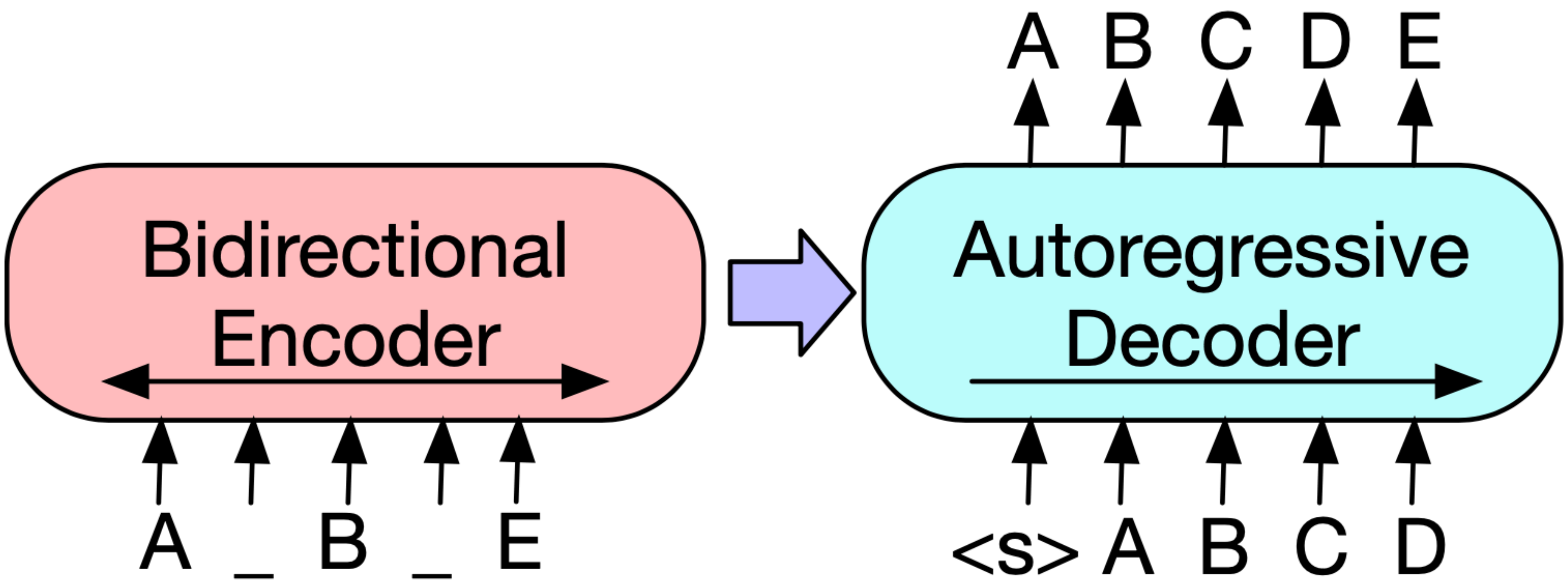
Pre-trained decoder?

- BertSum has only a **pre-trained encoder**
- But the **decoder** is trained from **scratch**
- Can we **pre-train** the **decoder** too?

BART

- Encoder-decoder model (Lewis et al., 2020)
- Also based on Transformers (Vaswani et al., 2017)
- Uses an unsupervised **denoising objective**
- **Fine-tuned** on end task datasets (incl. summarization)

BART



A B C D E
Bidirectional
Encoder
A _ B _ E
Autoregressive
Decoder
<s> A B C D

CNN/DM

Model	Type	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	Ext	40.42	17.62	36.67
BottomUP (Gehrmann et al., 2018)	Abs	41.22	18.68	38.34
BertSum large (Liu and Lapata, 2019)	Abs	42.13	19.60	39.18
BART* (Lewis et al., 2020)	Abs	**44.16**	**21.28**	**40.90**

Transductive Summarization

Transductive summarization

- Test set input might contain details that the model is not ‘familiar’ with
- For example, if fine-tuned on **dated news** and tested on **recent ones**
- Hypothesized that the model can learn various details from the **test set’s** input
- Will lead to better summaries

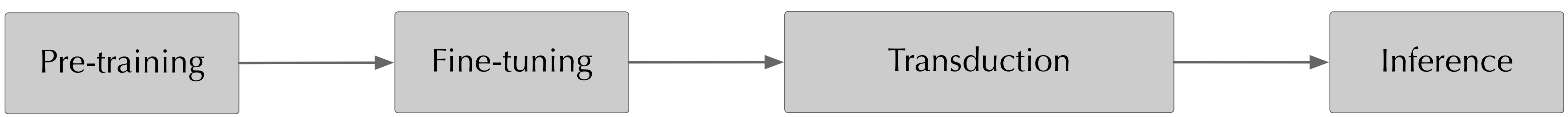
Transductive summarization

- To train the model on the test set we need references (summaries) that are not available

- Use an **extractive summarizer** to create extractive summaries

- Use these summaries as references for training

Transductive summarization



TrSum

- Used a pre-trained BART model
- Jointly fine-tuned* on the CNN/DM dataset
- Trained an extractive model to create **extractive references** on the test set
- Architecture remained exactly the same
- **Transducted** on the test set

TrSum

- Used a pre-trained BART model
- Jointly fine-tuned* on the CNN/DM dataset
- In training, used extractive and abstractive references

TrSum

- Trained an extractive model to create **extractive references** on the test set
- Architecture remained exactly the same
- **Transducted** on the test set

CNN/DM

Model	Type	ROUGE-1	ROUGE-2	ROUGE-L
LEAD-3	Ext	40.42	17.62	36.67
BottomUP (Gehrmann et al., 2018)	Abs	41.22	18.68	38.34
BertSum large (Liu and Lapata, 2019)	Abs	42.13	19.60	39.18
BART* (Lewis et al., 2020)	Abs	44.16	21.28	40.90
TrSum (Bražinskas et al., 2021)	Abs	**44.96**	**21.89**	**41.86**

Opinion Summarization

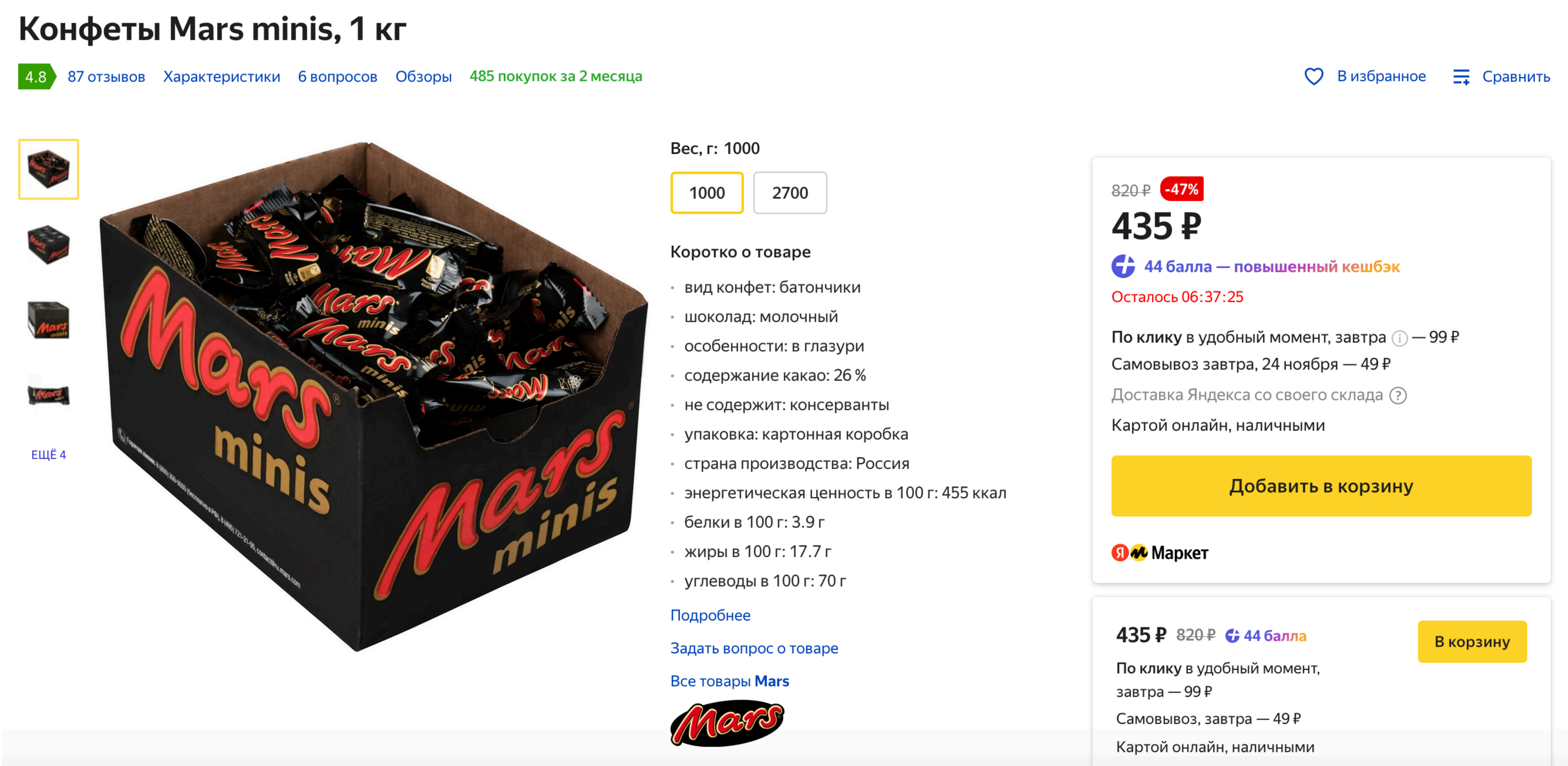
Customer reviews

- Users often purchase products **online** (e.g., from Yandex.Market or Amazon)
- Seek **opinions** of other **users** expressed in **reviews**
- Use this information for **better purchasing decisions**

Motivation

As per a New York Times study, Amazon reported a near 200-percent rise in profits, accelerated by much of North America's swift shift to exclusively online shopping. Amazon's sales were US$96.1 billion, up 37% from 2019, with profits rising to a jaw-dropping US$6.3 billion. The pandemic hasn't only increased the company's profits but also its expansion. Amazon expanded its fulfillment infrastructure by 50% in 2020, adding more than 250,000 employees in the process. For the first time in the company's history, Amazon now employs more than one million workers around the world.

Motivation


Конфеты Mars minis, 1 кг
4.8
87 отзывов
Характеристики
6 вопросов
Обзоры
485 покупок за 2 месяца
В избранное
Сравнить
ЕЩЁ 4
Mars minis
Вес, г: 1000
1000
2700
Коротко о товаре
вид конфет: батончики
шоколад: молочный
особенности: в глазури
содержание какао: 26 %
не содержит: консерванты
упаковка: картонная коробка
страна производства: Россия
энергетическая ценность в 100 г: 455 ккал
белки в 100 г: 3.9 г
жиры в 100 г: 17.7 г
углеводы в 100 г: 70 г
Подробнее
Задать вопрос о товаре
Все товары Mars
Mars
820 ₽
-47%
435 ₽
44 балла — повышенный кешбэк
Осталось 06:37:25
По клику в удобный момент, завтра — 99 ₽
Самовывоз завтра, 24 ноября — 49 ₽
Доставка Яндекса со своего склада
Картой онлайн, наличными
Добавить в корзину
Маркет
435 ₽
820 ₽
44 балла
В корзину
По клику в удобный момент,
завтра — 99 ₽
Самовывоз, завтра — 49 ₽
Картой онлайн, наличными

Motivation

Кристина Михайлова 4
Отличный товар Товар куплен на Маркете

Конфет много., качество и доставка порадовали

Комментировать 2 месяца назад

Алекс 7
Отличный товар Товар куплен на Маркете

Вкусные конфеты, удобный формат батончиков 👍

Комментировать 2 месяца назад

Елена З. 5
Отличный товар Товар куплен на Маркете

Опыт использования: менее месяца

Достоинства: вкусные

Недостатки: нет

Комментарий: срок хороший

Комментировать 2 месяца назад, Саратов

Имя скрыто
Отличный товар Товар куплен на Маркете

Опыт использования: более года

Достоинства: Много любимого шоколада в удобной упаковке

Недостатки: Цена

Комментарий: Вкусный любимый шоколад) хватает надолго

Комментировать Месяц назад, Серебряные Пруды

Motivation

- Reviews contain **useful information** for decision making
- Can be **condensed** to **short texts** to help the user in making informative decision

Challenges

- Products have **hundreds** or **even thousands** of reviews (hard to encode using standard neural encoders)
- Lack of annotated data (especially for abstractive models)
- Consequently, often approached using **extractive-** or **frequency-based** methods

Differences from news

- In news summarization, we summarize **objective information**
- In opinion summarization, we summarize **subjective information**
- Opinion summarization is relatively new direction

Example

Example

Вот о чём пишут чаще всего

Этот отзыв написал наш умный алгоритм — он всё прочитал и выделил главное

Достоинства

«Вкусные.» «Качества продукта.» «Вкус.» «Любимый вкус детства, большая упаковка.»

Недостатки

«Не пробовал ещё.» «Очень сладкие.» «Быстро съедаю.»

Полезная информация? **Да** **Нет**

Example: repetitions

Вот о чём пишут чаще всего

Этот отзыв написал наш умный алгоритм — он всё прочитал и выделил главное

Достоинства

«Вкусные.» «Качества продукта.» «Вкус.» «Любимый вкус детства, большая упаковка.»

Недостатки

«Не пробовал ещё.» «Очень сладкие.» «Быстро съедаю.»

Полезная информация? Да Нет

Example: uninformative

Вот о чём пишут чаще всего

Этот отзыв написал наш умный алгоритм — он всё прочитал и выделил главное

Достоинства

«Вкусные.» «Качества продукта.» «Вкус.» «Любимый вкус детства, большая упаковка.»

Недостатки

«Не пробовал ещё.» «Очень сладкие.» «Быстро съедаю.»

Полезная информация? Да Нет

Abstractive Models

Advantages of abstractive summarize

- Can use a **richer vocabulary of words**
- Can **rephrase** and **abstract**
- Can deal with **conflicting information**

Scarce annotated data

- Datasets with reviews-summary pairs are **very limited**
- Large quantities of reviews without summaries (**millions**)

Opinion summarization (unannotated data)

amazon.com

233 million reviews

yelp

8 million reviews

Abstractive models

- MeanSum (ICML 2019)
- Copycat (ACL 2020)
- FewSum (EMNLP 2020)
- SelSum (EMNLP 2021)

MeanSum: A Model for Unsupervised Neural Multi-Document Abstractive Summarization

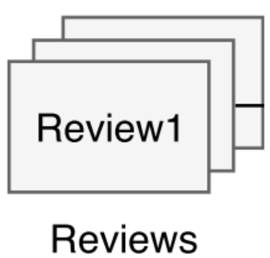
Eric Chu, Peter Liu

ICML 2019

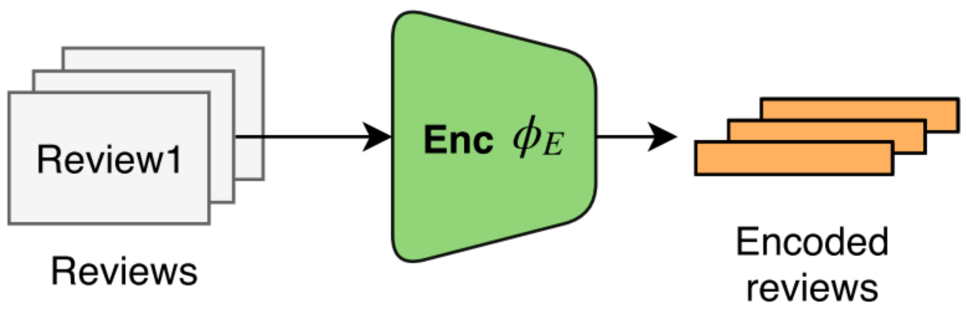
MeanSum

- **Unsupervised** abstractive summarizer of reviews (Chu and Liu, 2019)
- **Summary:**
  - Represented as sequence of latent categorical variables
  - **Differentiable** samples via **Gumbel-softmax trick** (Jang et al., 2016)
- Based on **multi-tasking:**
  - **Auto-encoding** of reviews
  - **Semantic similarly** of the sampled summary and input reviews

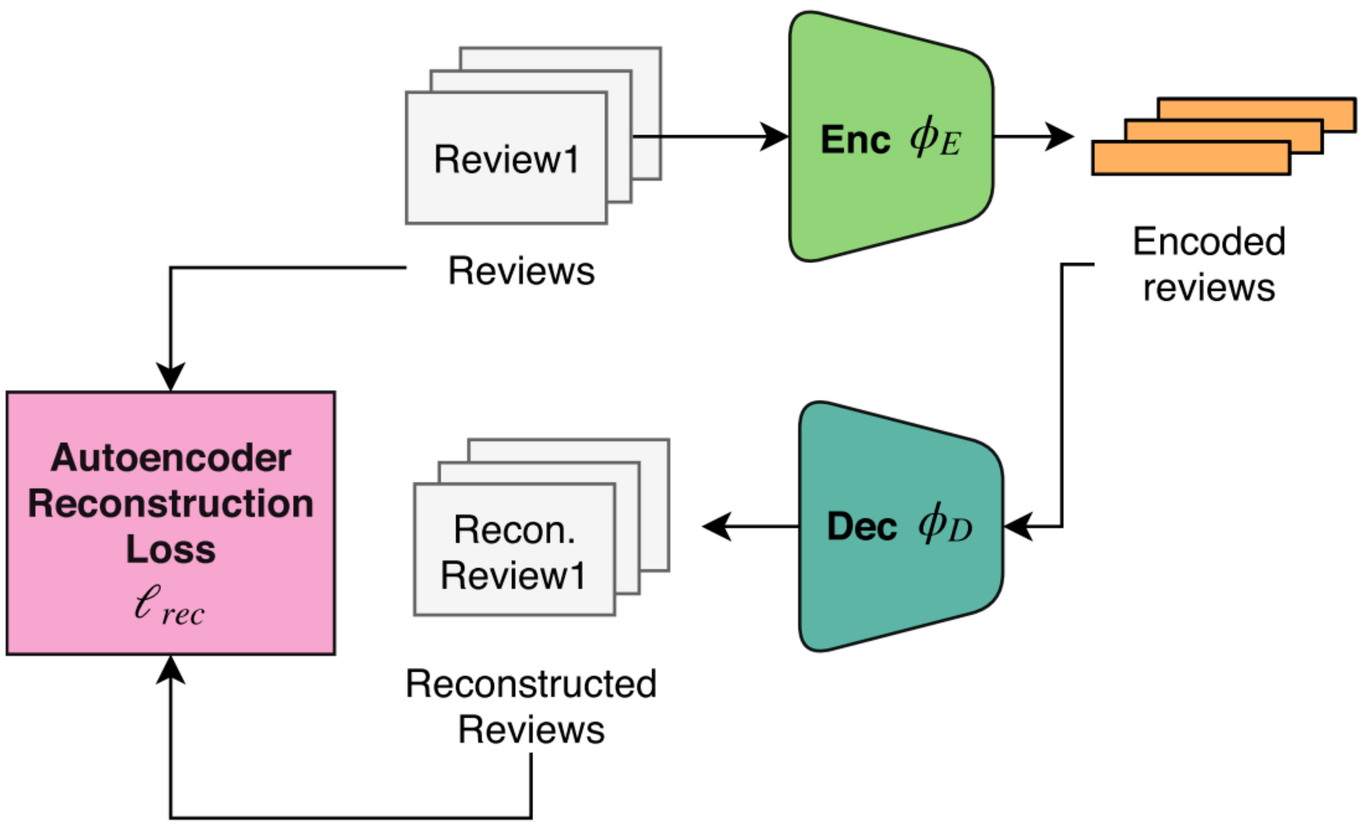
MeanSum



MeanSum



MeanSum



Review1
Reviews
Enc ϕ_E
Encoded reviews
Autoencoder Reconstruction Loss
ℓ_{rec}
Recon. Review1
Dec ϕ_D
Reconstructed Reviews

Reconstruction loss

ϕ_E - encoder x_i - review document

ϕ_D - decoder

$$l_{rec}(\{x_1, x_2, ..., x_N\}, \phi_E, \phi_D) = \sum_{i=1}^{N} CE(x_i, \phi_D(\phi_E(x_i)))$$

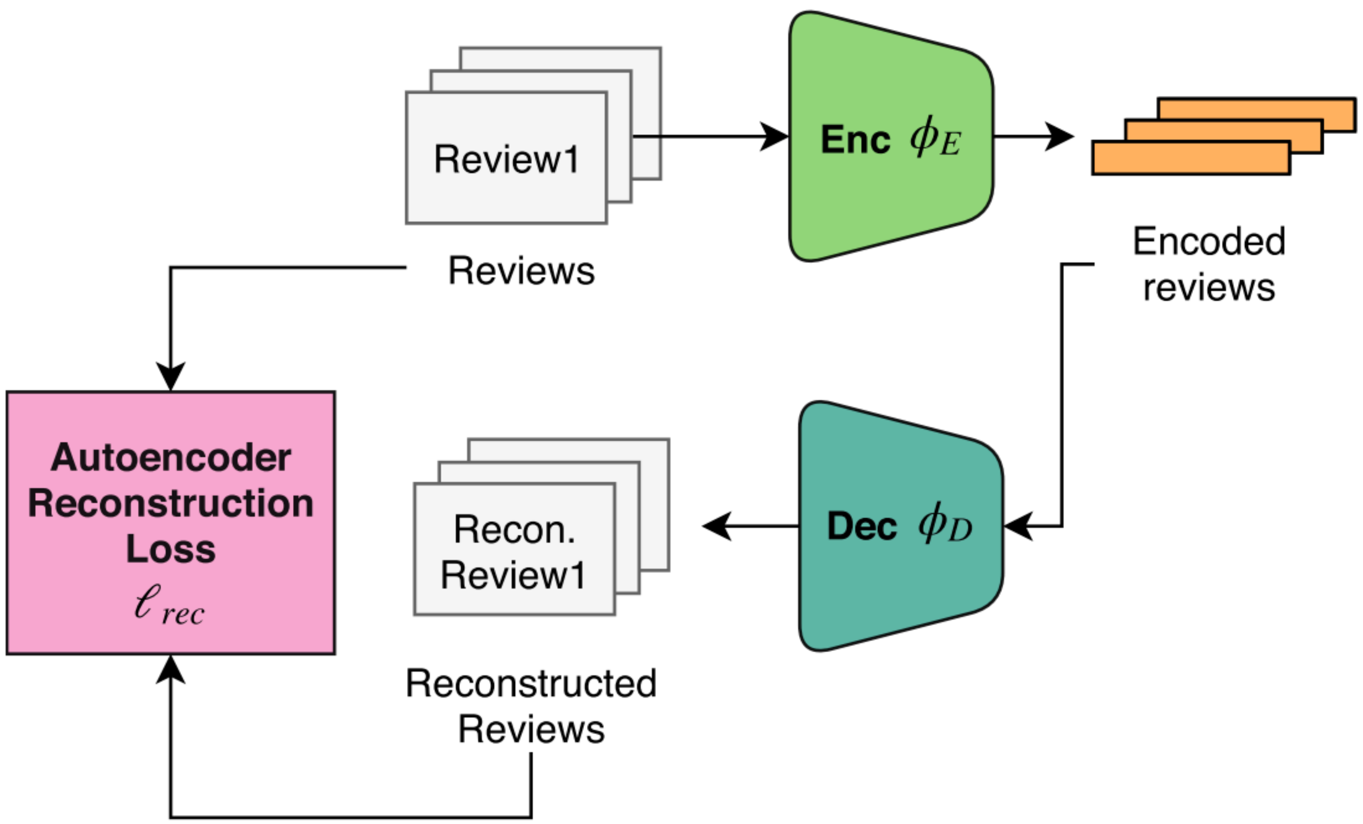
Reconstruction loss

ϕ_E - encoder $\quad x_i$ - review document

ϕ_D - decoder (**use Teacher Forcing**)

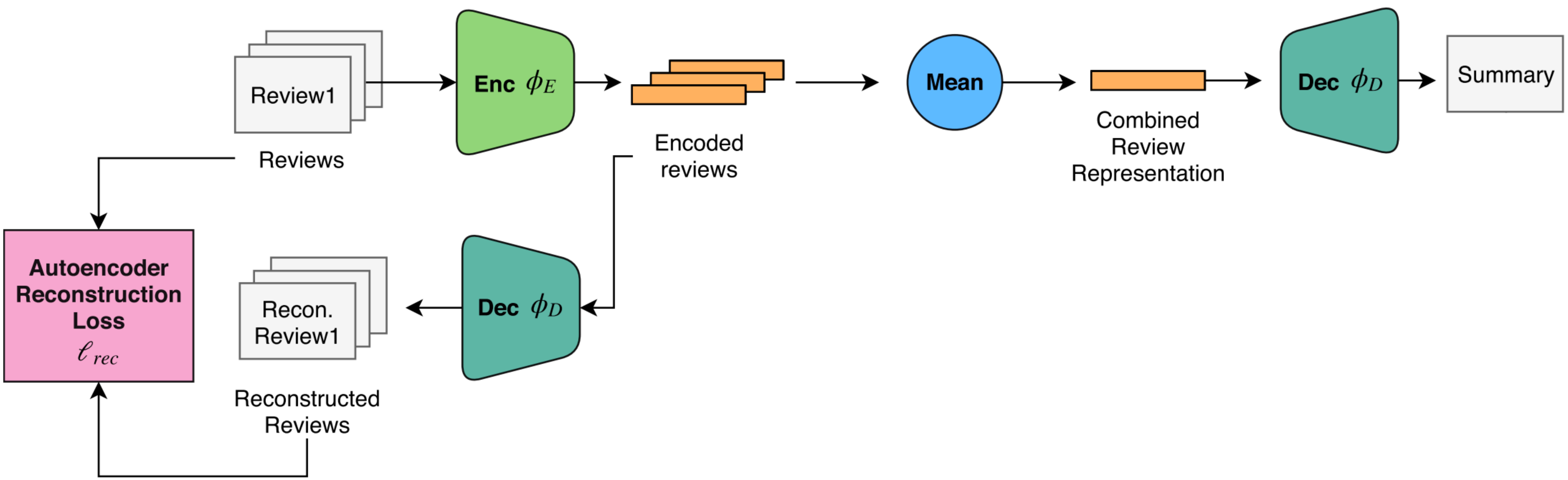
$$l_{rec}(\{x_1, x_2, ..., x_N\}, \phi_E, \phi_D) = \sum_{i=1}^{N} CE(x_i, \phi_D(\phi_E(x_i)))$$

MeanSum



Review1
Reviews
Enc ϕ_E
Encoded reviews
Autoencoder Reconstruction Loss ℓ_{rec}
Recon. Review1
Dec ϕ_D
Reconstructed Reviews

MeanSum



Review1
Reviews
Enc ϕ_E
Encoded reviews
Mean
Combined Review Representation
Dec ϕ_D
Summary
Autoencoder Reconstruction Loss
ℓ_{rec}
Recon. Review1
Reconstructed Reviews
Dec ϕ_D

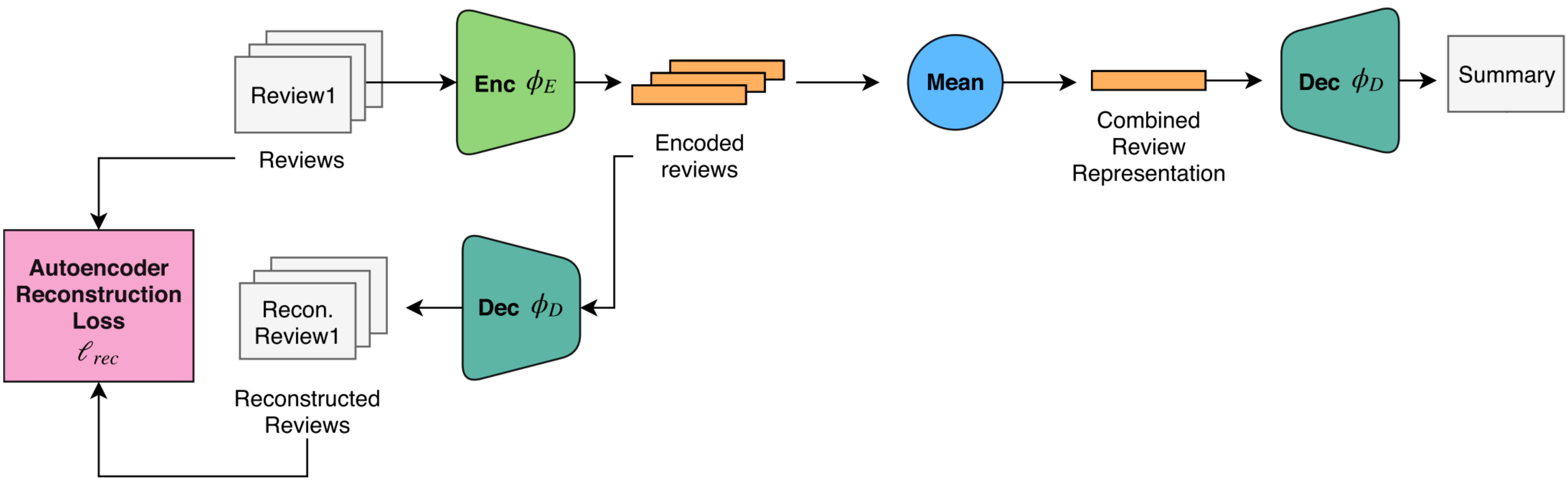
Summary sampling

- Decoder ϕ_D assigns **probabilities** to words
- Can obtain a differentiable sample using **Gumbel-softmax re-parametrizaiton trick** (Jang et al., 2016)
- Can backprop through the sample
- Notice that we **can't use Teacher Forcing** (no gold prefixes)

Semantic similarity loss

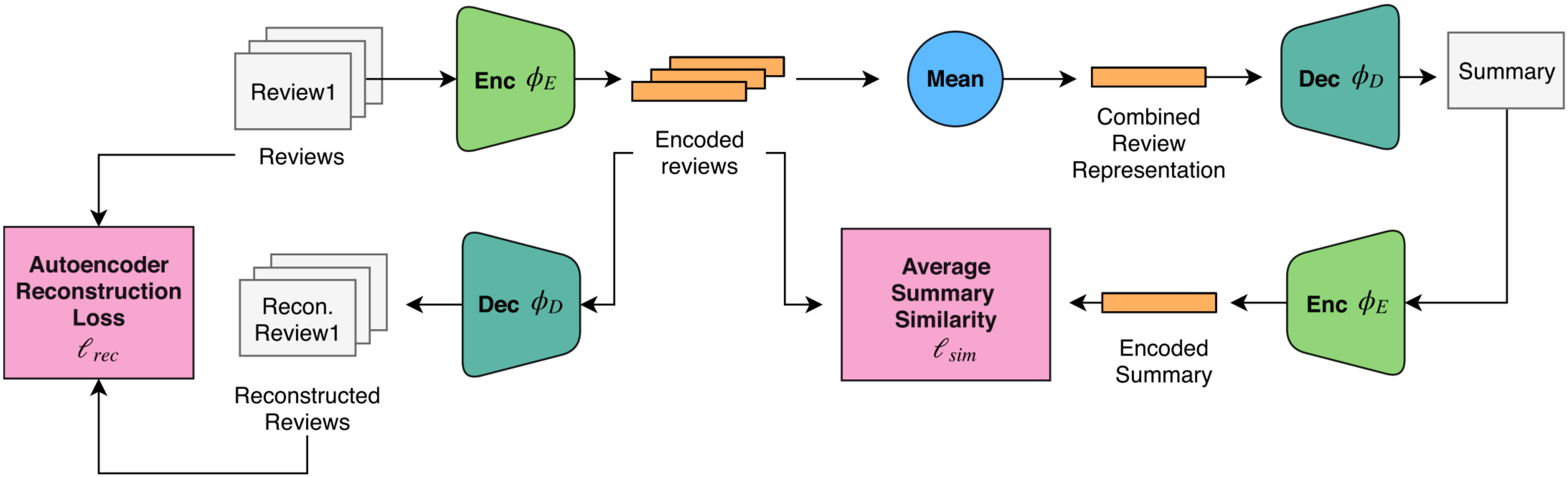
$$s \sim \phi_D(\frac{1}{N}\sum_{i=1}^{N}\phi_E(x_i))$$

MeanSum



Review1
Reviews
Enc ϕ_E
Encoded reviews
Mean
Combined Review Representation
Dec ϕ_D
Summary
Autoencoder Reconstruction Loss
ℓ_{rec}
Recon. Review1
Reconstructed Reviews
Dec ϕ_D

MeanSum



Review1
Reviews
Enc ϕ_E
Encoded reviews
Mean
Combined Review Representation
Dec ϕ_D
Summary
Autoencoder Reconstruction Loss ℓ_{rec}
Recon. Review1
Reconstructed Reviews
Dec ϕ_D
Average Summary Similarity ℓ_{sim}
Encoded Summary
Enc ϕ_E

Semantic similarity loss

$$s \sim \phi_D(\frac{1}{N}\sum_{i=1}^{N}\phi_E(x_i))$$

$$l_{sim}(\{x_1, x_2, ..., x_N\}) = \frac{1}{N}\sum_{i=1}^{N} d_{cos}(\phi_E(x_i), \phi_E(s))$$

Final loss

$$l_{rec}(\{x_1, x_2, ..., x_N\}, \phi_E, \phi_D) + l_{sim}(\{x_1, x_2, ..., x_N\}, \phi_E, \phi_D)$$

Results on Amazon

	ROUGE-1	ROUGE-2	ROUGE-L

Results on Amazon

	ROUGE-1	ROUGE-2	ROUGE-L
Lead	27.00	4.92	14.95

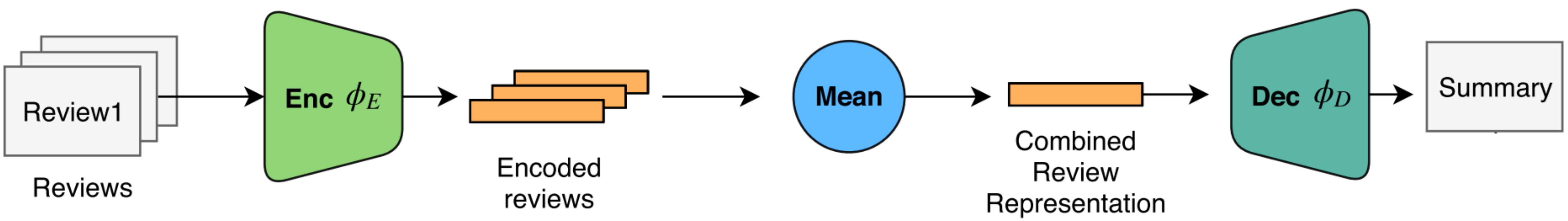
Results on Amazon

	ROUGE-1	ROUGE-2	ROUGE-L
MeanSum	26.63	4.89	17.11
Lead	27.00	4.92	14.95

Averaged representations?

Why would **the averaged review representations** correspond to a **summary** and not another **review**?

Averaged representations?



MeanSum

The shirt is very soft and comfortable. I bought a size larger than I normally wear and it fits fine. I'm 5 '4 and the top is a bit short. I guess I just got a good deal.

MeanSum

problem: superficial, unimportant details

The shirt is very soft and comfortable. ***I bought a size larger than I normally wear and it fits fine.*** *I'm 5 '4 and the top is a bit short. I guess I just got a good deal.*

MeanSum

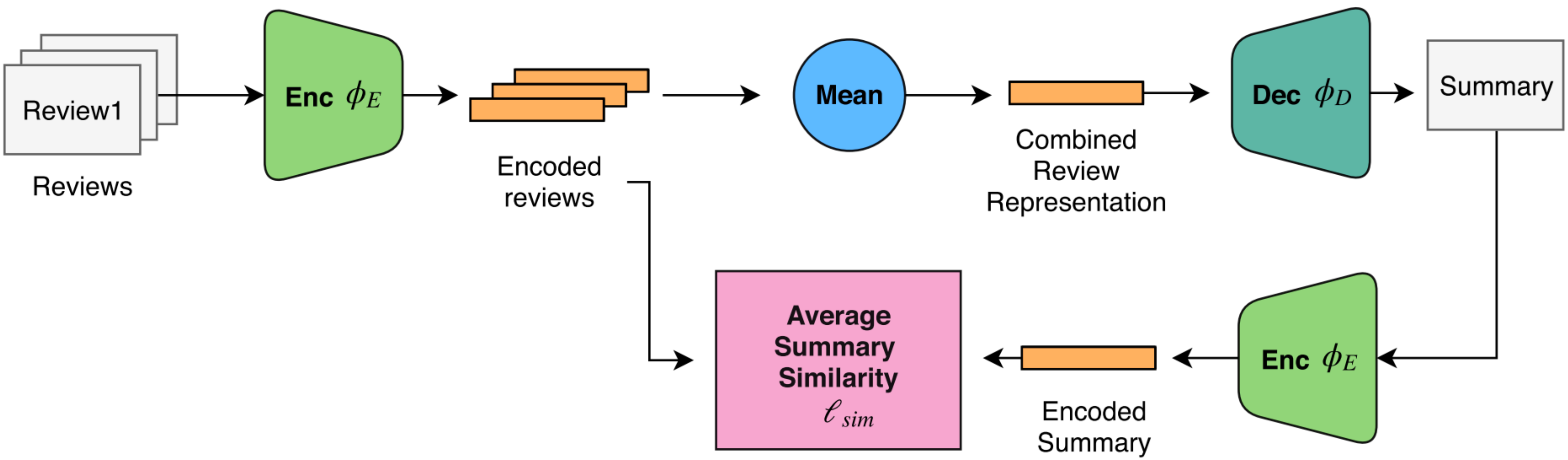
problem: writing style

The shirt is very soft and comfortable. ***I*** *bought a size larger than* ***I*** *normally wear and it fits fine.* ***I'm*** *5 '4 and the top is a bit short.* ***I*** *guess* ***I*** *just got a good deal.*

No prior?

- Is it possible to guarantee fluency of summaries without using a prior?
- What restricts the decoder from not producing degenerate summaries? E.g., a sequence of keywords.

No prior?



Review1
Reviews
Enc ϕ_E
Encoded reviews
Mean
Combined Review Representation
Dec ϕ_D
Summary
Enc ϕ_E
Encoded Summary
Average Summary Similarity ℓ_{sim}

No prior?

$$s \sim \phi_D(\frac{1}{N}\sum_{i=1}^{N}\phi_E(x_i))$$

No prior distribution restricts what **the summary** should be

We observed that the model can **diverge** to generation of **not fluent text**

MeanSum

Pros:

- Simple model
- Does not require annotated summaries

MeanSum

Cons:

- Generates summaries that look like reviews
  - Informal writing style
  - Unimportant details
- Poor content support (hallucinations)

Unsupervised Opinion Summarization as Copycat-Review Generation

Arthur Bražinskas, Mirella Lapata, Ivan Titov

ACL 2020

Approach

- Unsupervised latent model (continuous variables)
- Learns **latent semantic representations** of products and individual reviews
- Generates summaries from **‘summarizing’** latent representations

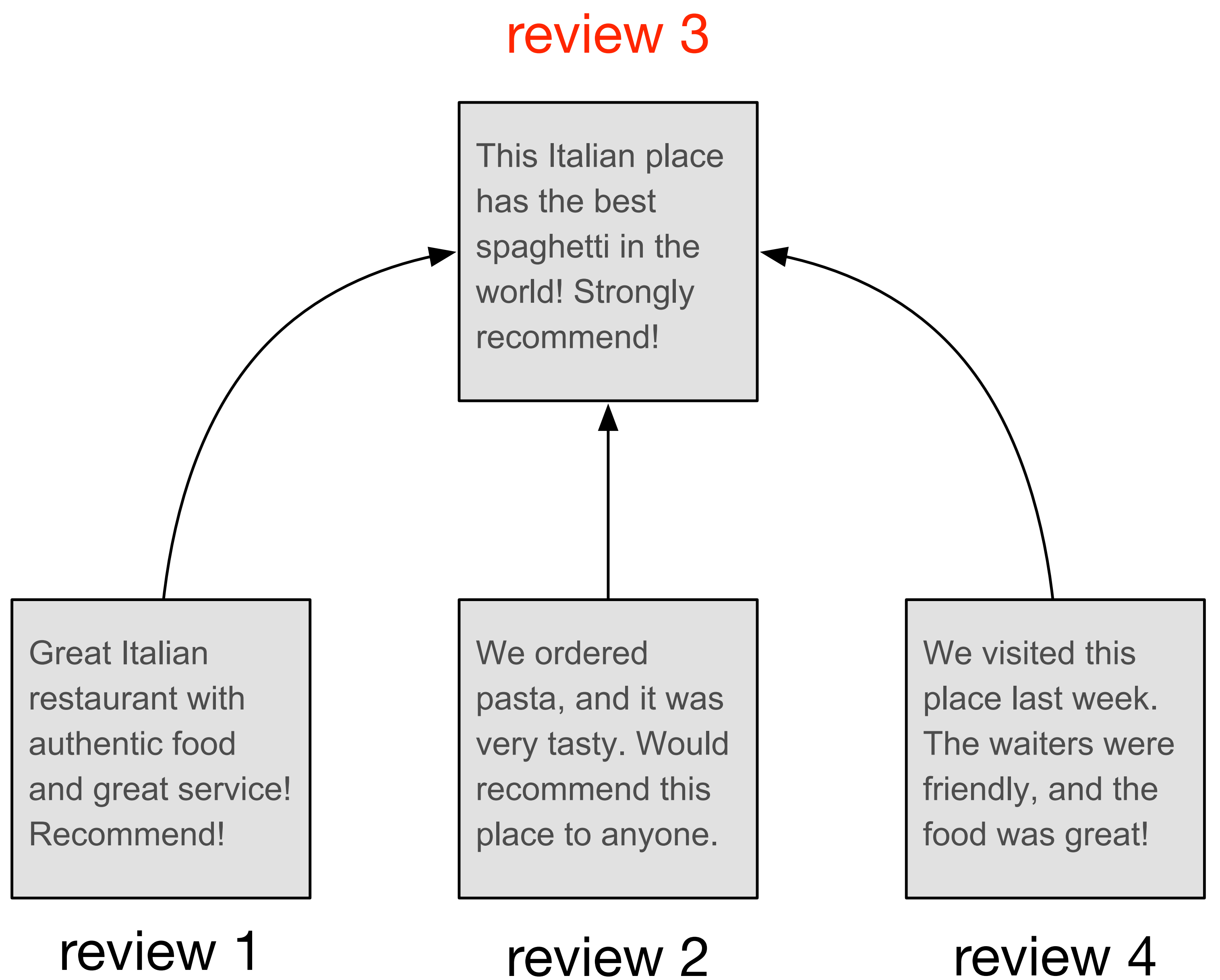
Conditional LM

- Formulate a **conditional language model (CLM)**
- Predicts a review conditioned on the **other** reviews of a product (**leave-one-out**)
- Intuitively, similar to the **pseudolikelihood** estimation (Besag, 1975)

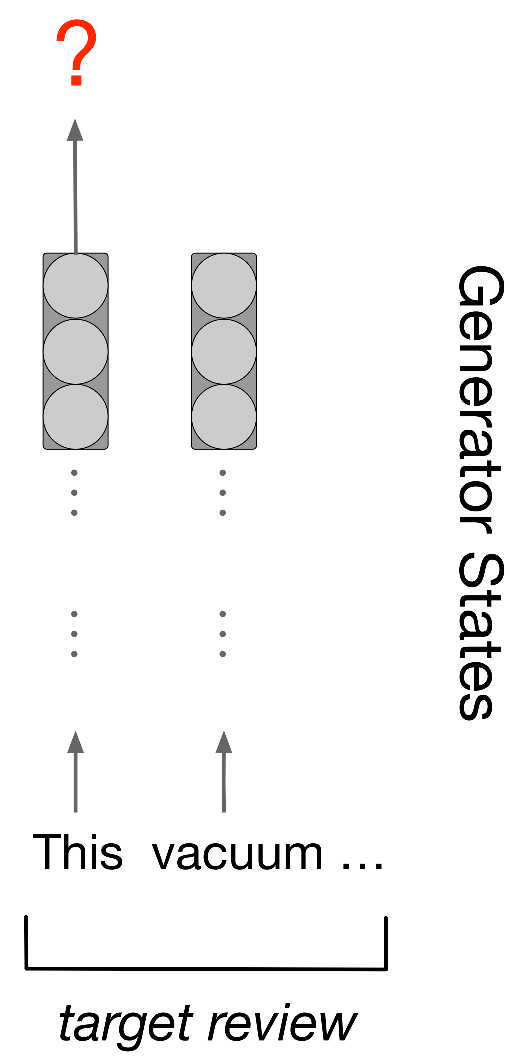
Leave-one-out

review 1	review 2	review 3	review 4
Great Italian restaurant with authentic food and great service! Recommend!	We ordered pasta, and it was very tasty. Would recommend this place to anyone.	This Italian place has the best spaghetti in the world! Strongly recommend!	We visited this place last week. The waiters were friendly, and the food was great!

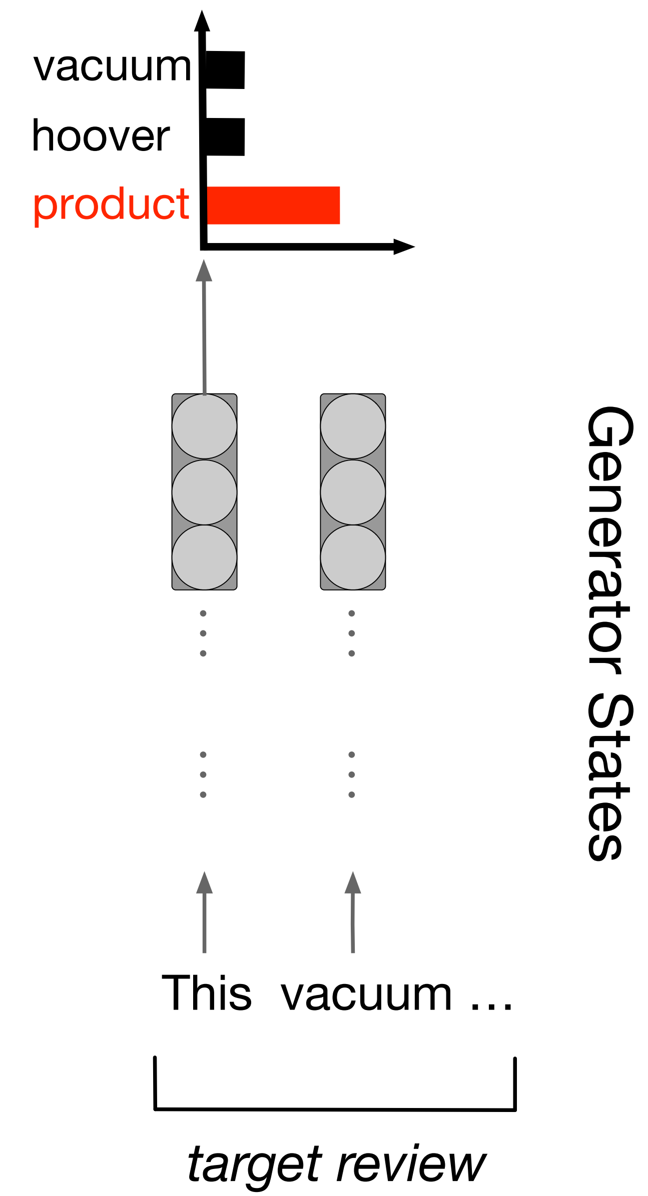
Leave-one-out



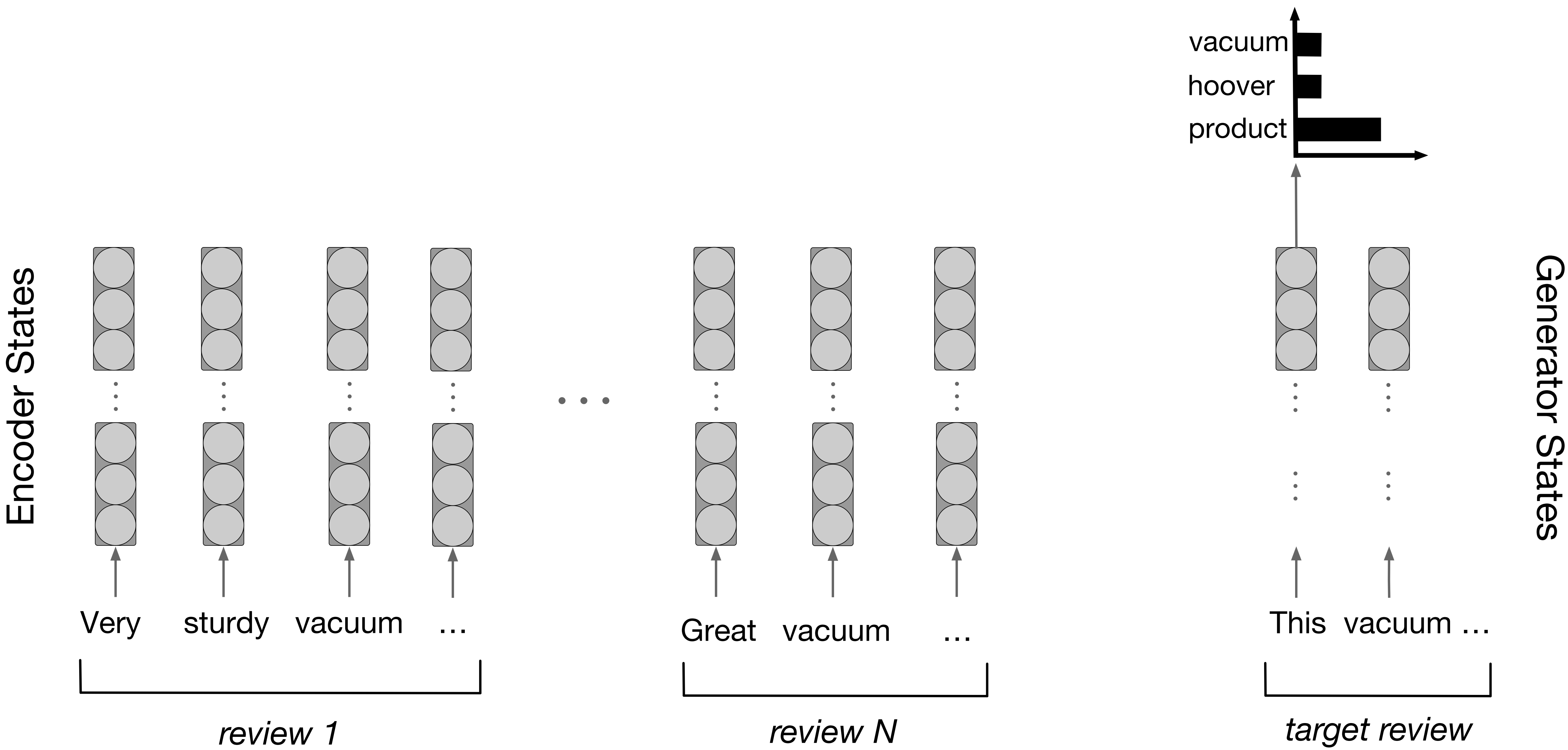
Leave-one-out



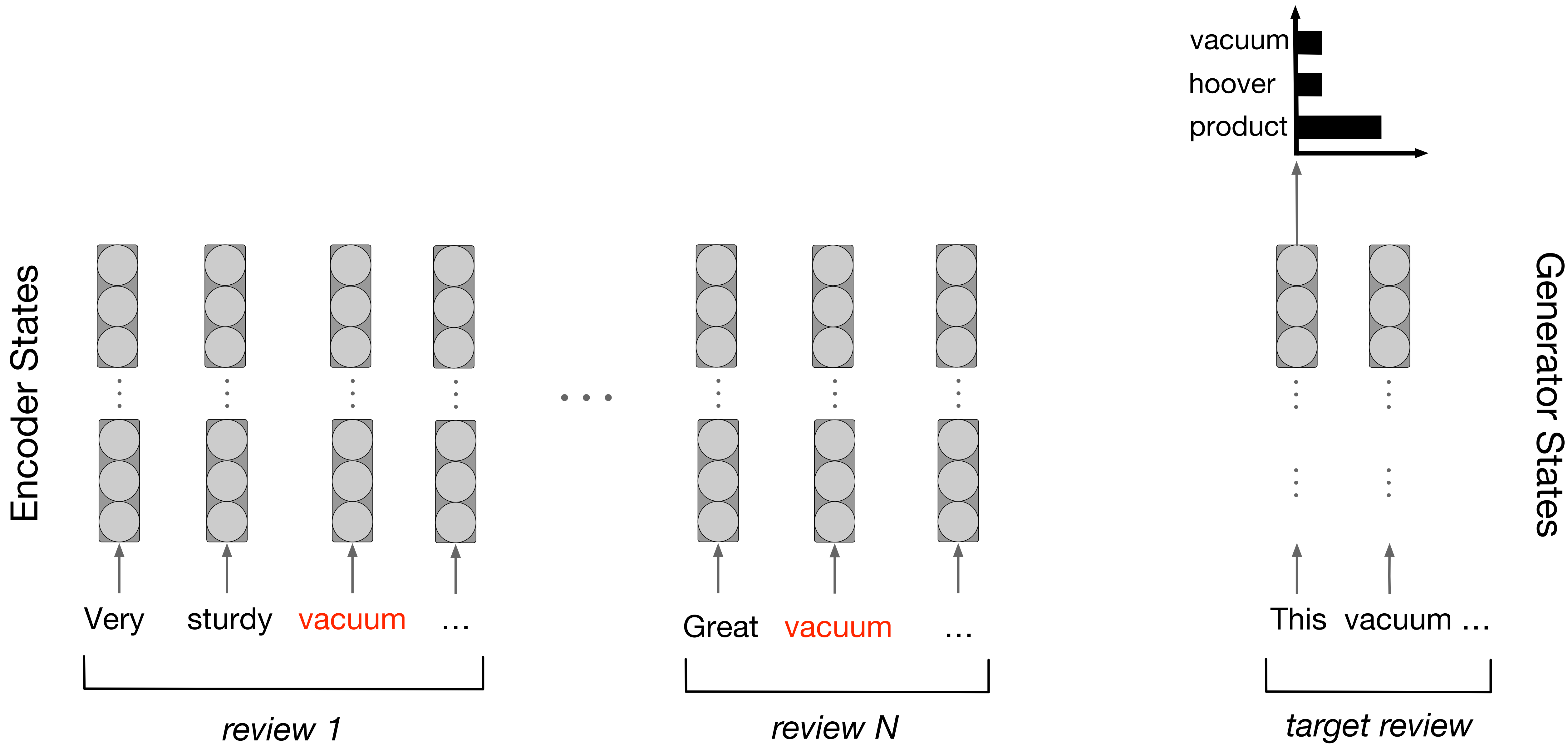
Leave-one-out



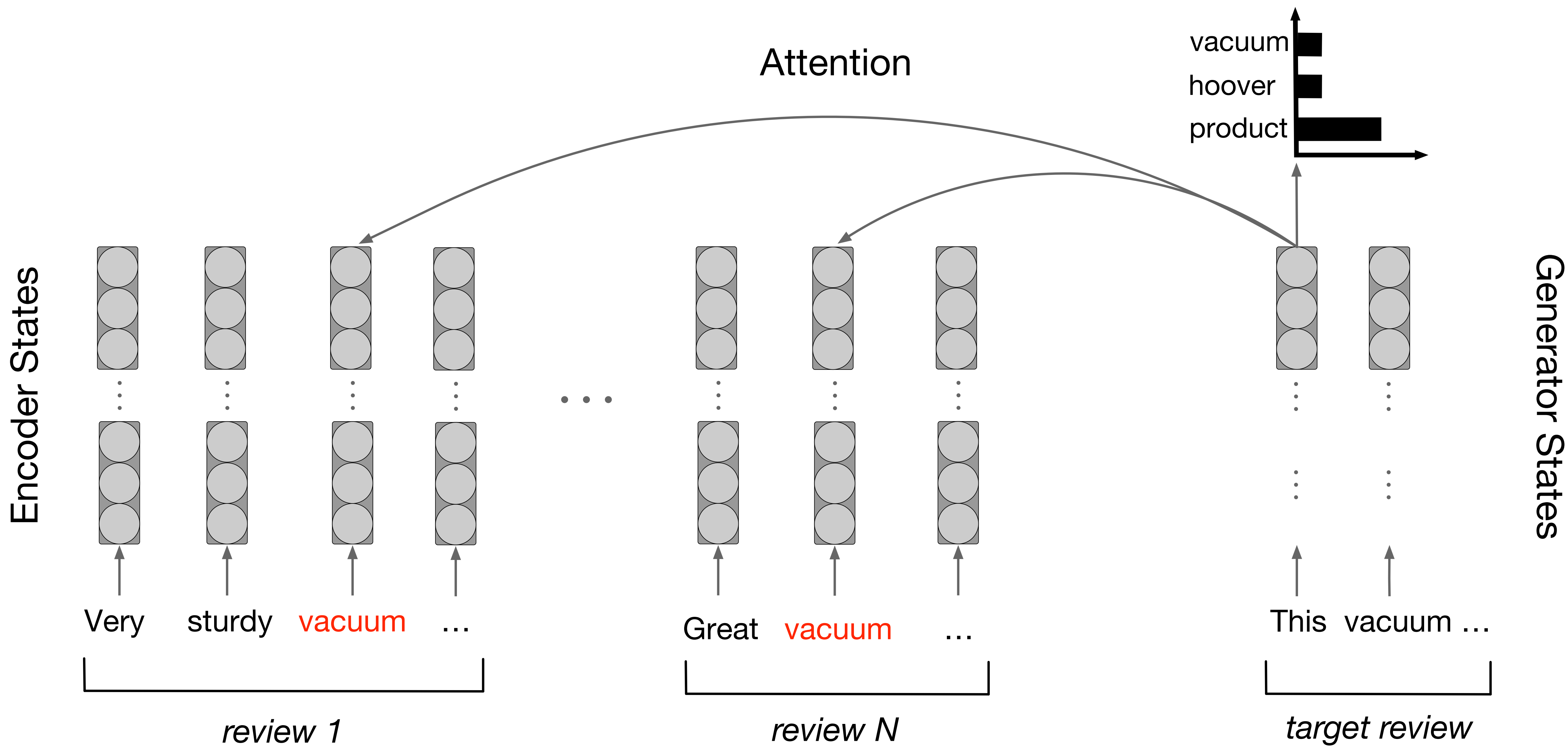
Leave-one-out



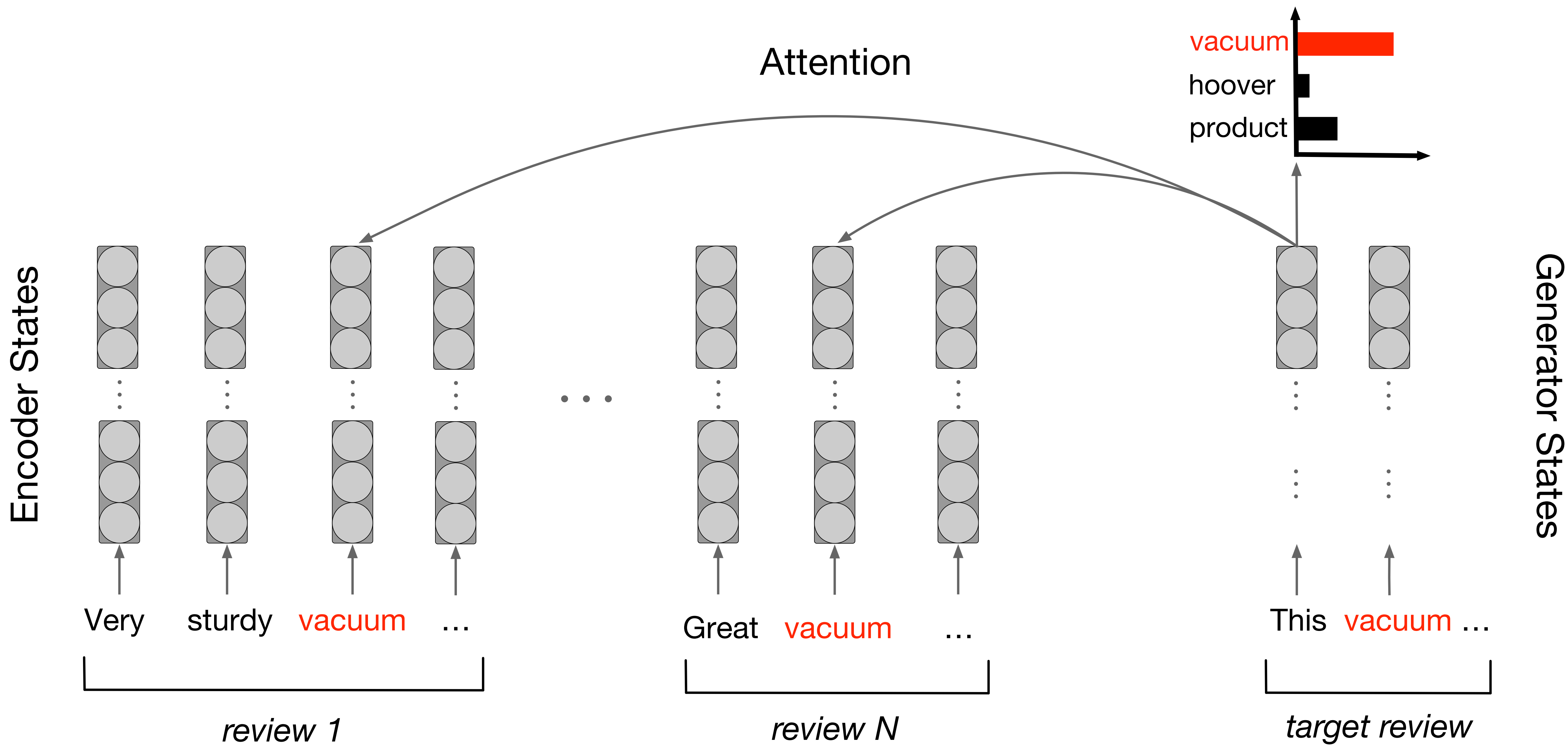
Leave-one-out



Leave-one-out



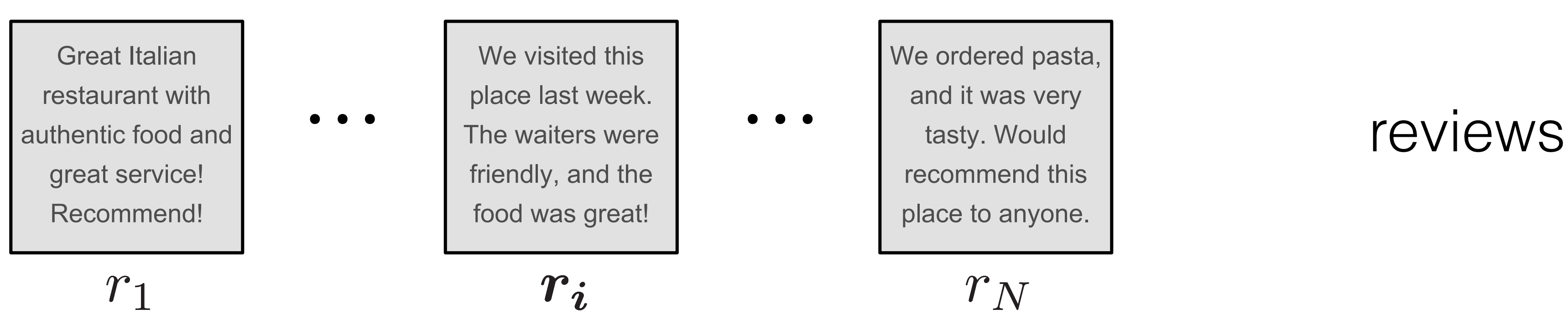
Leave-one-out



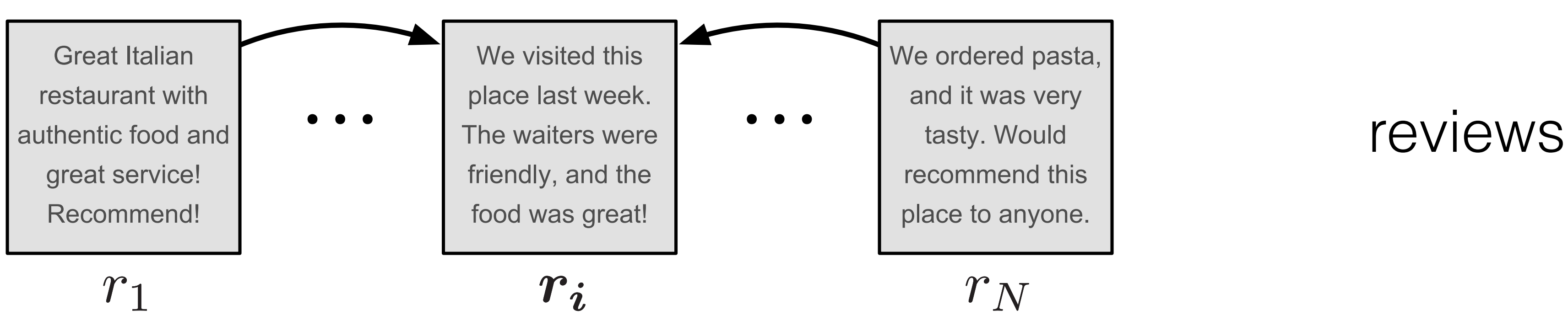
Novelty reduction

- Model is trained to predict reviews
- Summaries are different from reviews in content
- Summaries do not have **novel content**
- Control the amount of ‘novelty’ via **latent variables**

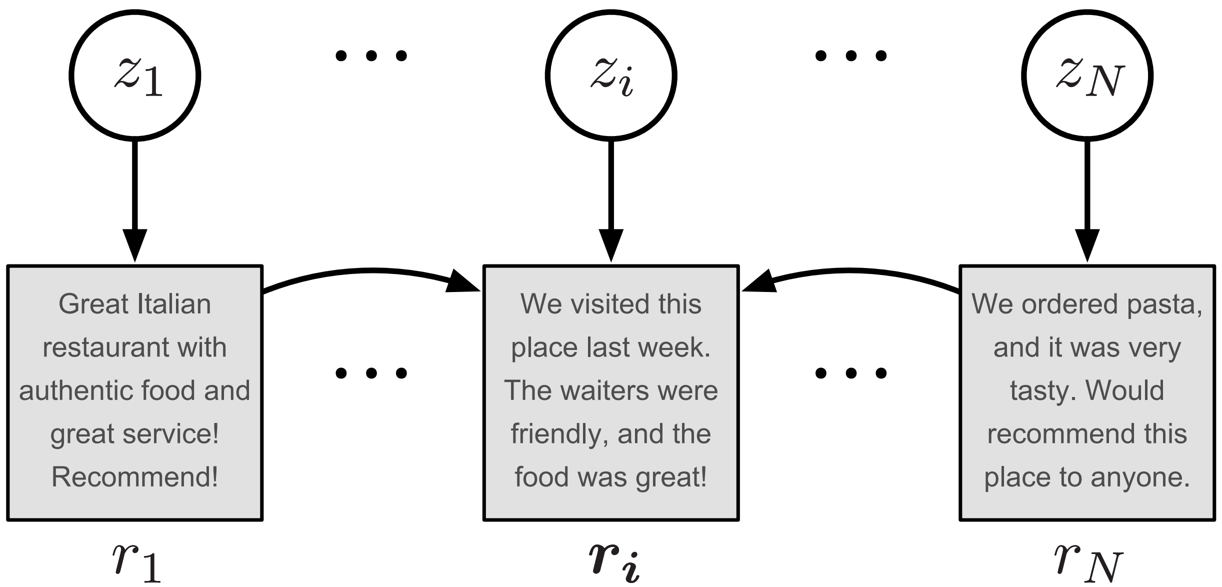
Latent model



Latent model



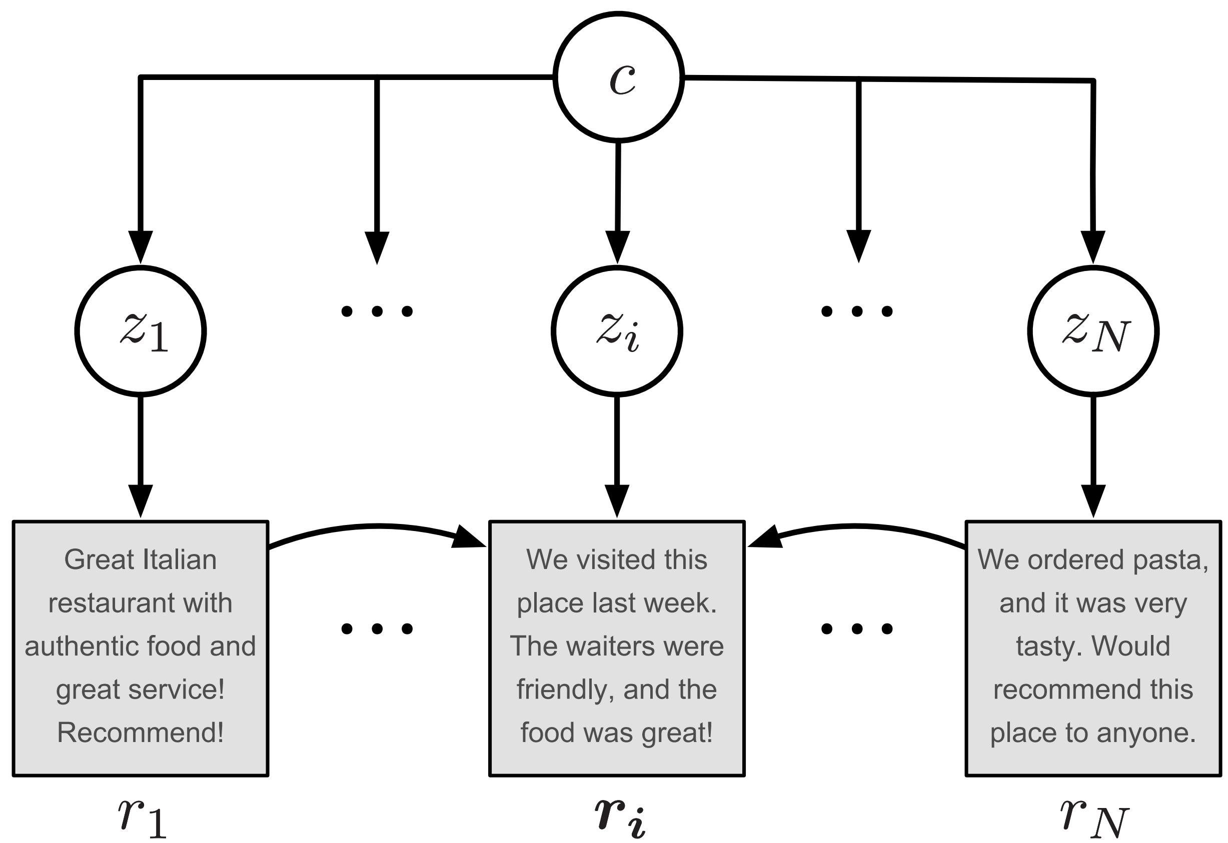
Latent model



review representations

reviews

Latent model



product representation

review representations

reviews

Model training

Variational Auto-encoders (Kingma and Welling, 2013) via differentiable sampling

Summary generation

- Use **mean values** of the latent variables to **limit novelty**

- Show that they correspond to **summarizing reviews**

Summary generation

1. Infer **the mean** representation of the product:

$$c^* = \mathbb{E}_{c \sim q_\phi(c|r_{1:N})}[c]$$

Summary generation

1. Infer **the mean** representation of the product:

$$c^* = \mathbb{E}_{c \sim q_\phi(c|r_{1:N})}[c]$$

2. Infer **the mean** representation of the review:

$$z^* = \mathbb{E}_{z \sim p_\theta(z|c^*)}[z]$$

Summary generation

1. Infer **the mean** representation of the product:

$$c^* = \mathbb{E}_{c \sim q_\phi(c|r_{1:N})}[c]$$

2. Infer **the mean** representation of the review:

$$z^* = \mathbb{E}_{z \sim p_\theta(z|c^*)}[z]$$

3. Generate **the summarizing review**:

$$r^* = \arg\max_{r} p_\theta(r|z^*, r_{1:N})$$

Example Summary

Summary	This restaurant is a hidden gem in Toronto. The food is delicious, and the service is impeccable. Highly recommend for anyone who likes French bistro.
Reviews	We got the steak frites and the chicken frites both of which were very good ... Great service ... || I really love this place ... Côte de Boeuf ... A Jewel in the big city ... || French jewel of Spadina and Adelaide , Jules ... They are super accommodating ... moules and frites are delicious ... || Food came with tons of greens and fries along with my main course , thumbs uppp ... || Chef has a very cool and fun attitude ... || Great little French Bistro spot ... Go if you want French bistro food classics ... || Great place ... the steak frites and it was amazing ... Best Steak Frites ... in Downtown Toronto ... || Favourite french spot in the city ... crème brule for dessert

Summary	This restaurant is a hidden gem in Toronto. The food is delicious, and the service is impeccable. Highly recommend for anyone who likes French bistro.
Reviews	We got the steak frites and the chicken frites both of which were very good ... Great service ... || I really love this place ... Côte de Boeuf ... A Jewel in the big city ... || French jewel of Spadina and Adelaide , Jules ... They are super accommodating ... moules and frites are delicious ... || Food came with tons of greens and fries along with my main course , thumbs uppp ... || Chef has a very cool and fun attitude ... || Great little French Bistro spot ... Go if you want French bistro food classics ... || Great place ... the steak frites and it was amazing ... Best Steak Frites ... in Downtown Toronto ... || Favourite french spot in the city ... crème brule for dessert

Summary	This restaurant is a hidden gem in Toronto. The food is delicious, and the service is impeccable. Highly recommend for anyone who likes French bistro.
Reviews	We got the steak frites and the chicken frites both of which were very good ... Great service ... || I really love this place ... Côte de Boeuf ... A Jewel in the big city ... || French jewel of Spadina and Adelaide , Jules ... They are super accommodating ... moules and frites are delicious ... || Food came with tons of greens and fries along with my main course , thumbs uppp ... || Chef has a very cool and fun attitude ... || Great little French Bistro spot ... Go if you want French bistro food classics ... || Great place ... the steak frites and it was amazing ... Best Steak Frites ... in Downtown Toronto ... || Favourite french spot in the city ... crème brule for dessert

Summary	This restaurant is a hidden gem in Toronto. The food is delicious, and the service is impeccable. Highly recommend for anyone who likes French bistro.
Reviews	We got the steak frites and the chicken frites both of which were very good ... Great service ... || I really love this place ... Côte de Boeuf ... A Jewel in the big city ... || French jewel of Spadina and Adelaide , Jules ... They are super accommodating ... moules and frites are delicious ... || Food came with tons of greens and fries along with my main course , thumbs uppp ... || Chef has a very cool and fun attitude ... || Great little French Bistro spot ... Go if you want French bistro food classics ... || Great place ... the steak frites and it was amazing ... Best Steak Frites ... in Downtown Toronto ... || Favourite french spot in the city ... crème brule for dessert

Results on Amazon

	ROUGE-1	ROUGE-2	ROUGE-L
MeanSum	26.63	4.89	17.11
Lead	27.00	4.92	14.95

Results on Amazon

	ROUGE-1	ROUGE-2	ROUGE-L
Copycat	27.85	4.77	18.86
MeanSum	26.63	4.89	17.11
Lead	27.00	4.92	14.95

Pitfalls

- The model is **never exposed** to the **actual requirements** for a **good summary**
- Can produce fragments that are:
  - Written in the informal writing style
  - Not all details are important

Example summary

These are the tights **I’ve ever worn**. They fit well and are comfortable to wear. I wish they were a little bit thicker, but I’m sure they will last a long time.

Example summary

These are the tights **I’ve ever worn**. They fit well and are comfortable to wear. **I wish they were** a little bit thicker, but I’m sure they will last a long time.

Example summary

These are the tights **I’ve ever worn**. They fit well and are comfortable to wear. **I wish they were** a little bit thicker, **but I’m sure they will last a long time**.

Few-Shot Learning for Opinion Summarization

Arthur Bražinskas, Mirella Lapata, Ivan Titov

EMNLP 2020

Approach

- Proposed a **few-shot learning** framework (FewSum)
- Utilizes a **handful of human-written summaries** for training
- Effectively **switch** an **unsupervised model** to a **summarizer**
- Summaries are written in the **formal writing style** with more **informative content**

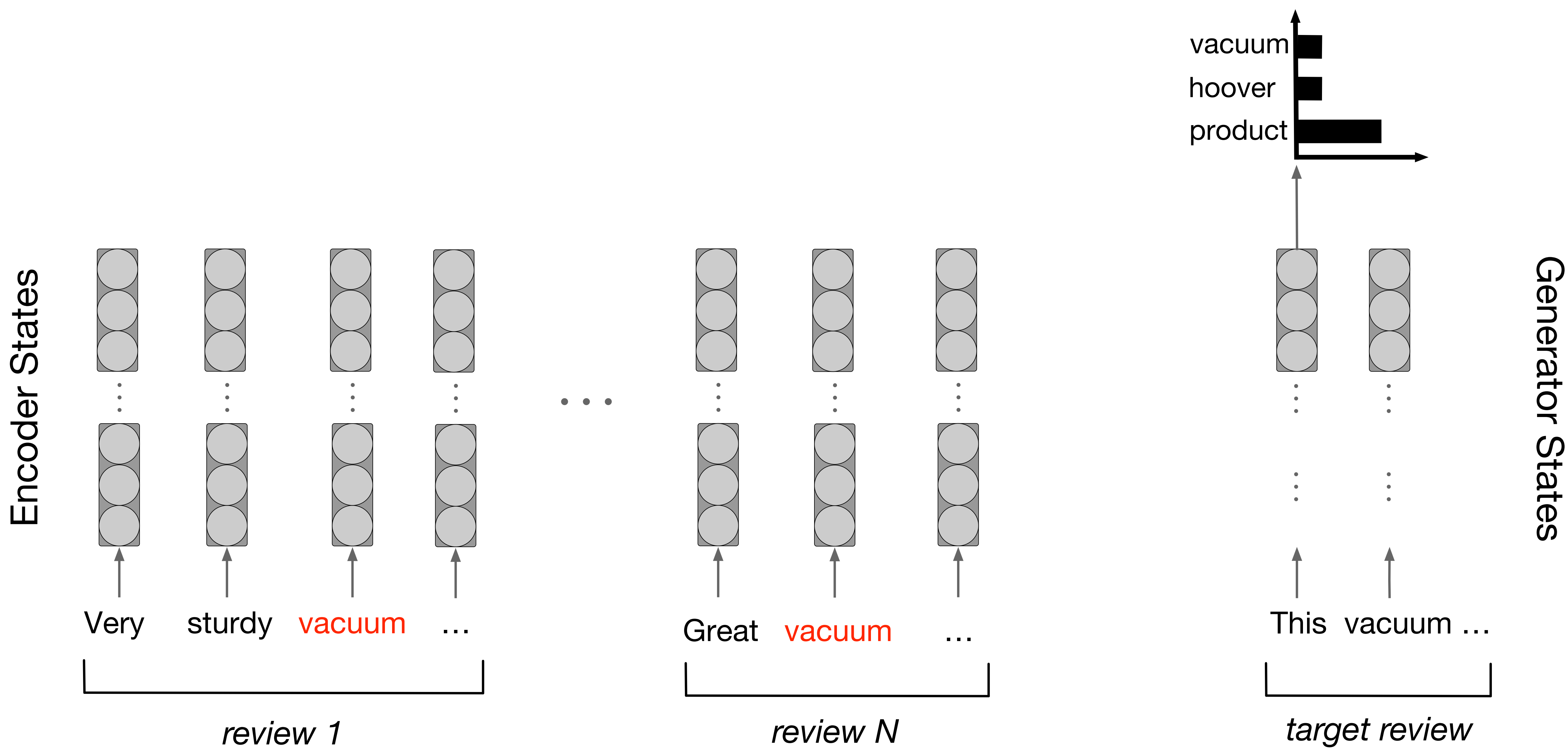
Annotated data

- Fine-tuning, in most cases, is performed on **hundreds of thousands** of **summaries**
- CNN/DM **~ 300k** article-summary pairs
- In our case, we have **~30 annotated products** for fine-tuning
- Yet, we show that they can be **efficiently utilized** in a **few-shot fashion**

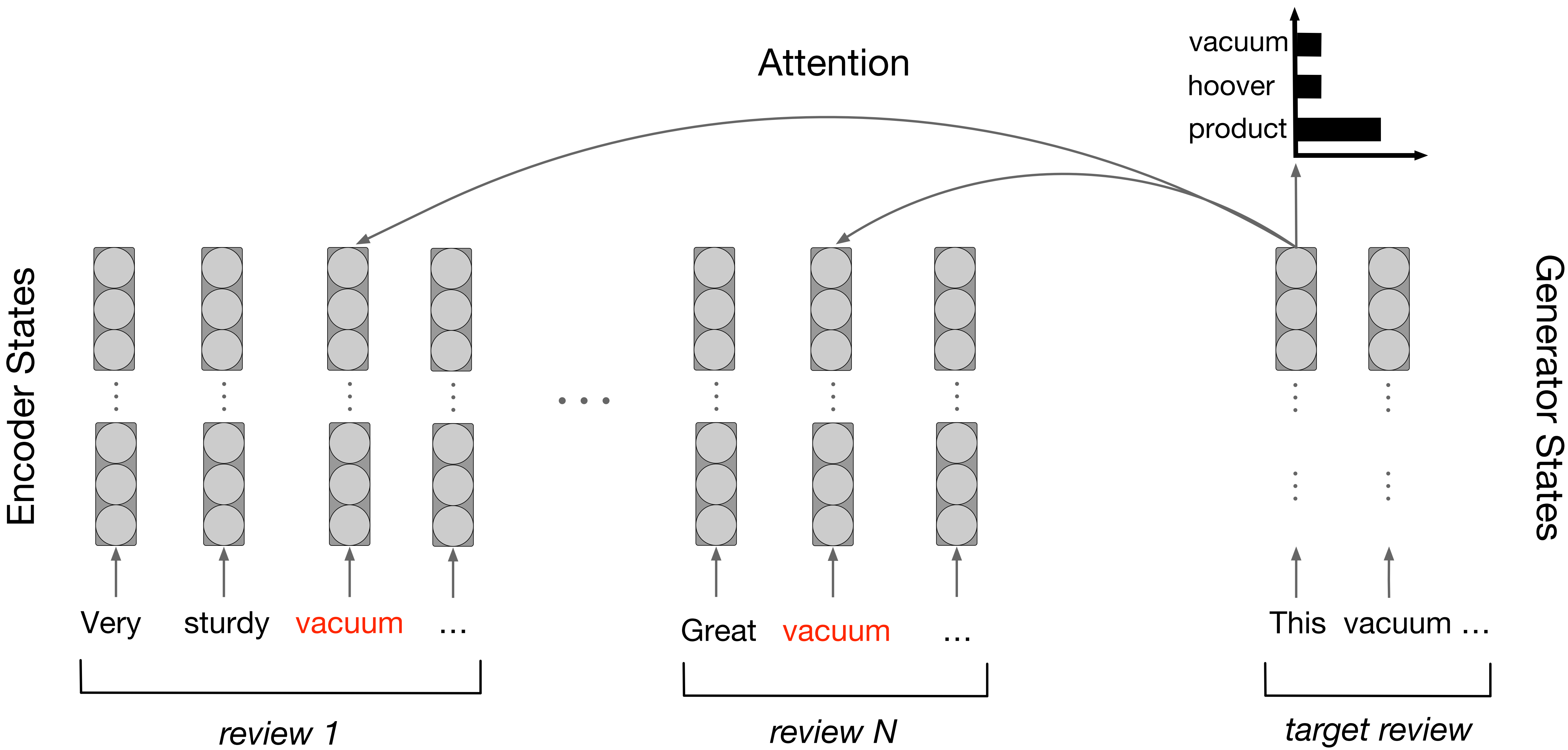
Conditional language model

- Same as in Copycat
- Conditional language model (CLM)
- Encoder-generator architecture
- Training on a large collection of customer reviews
- Using the **leave-one-out objective**

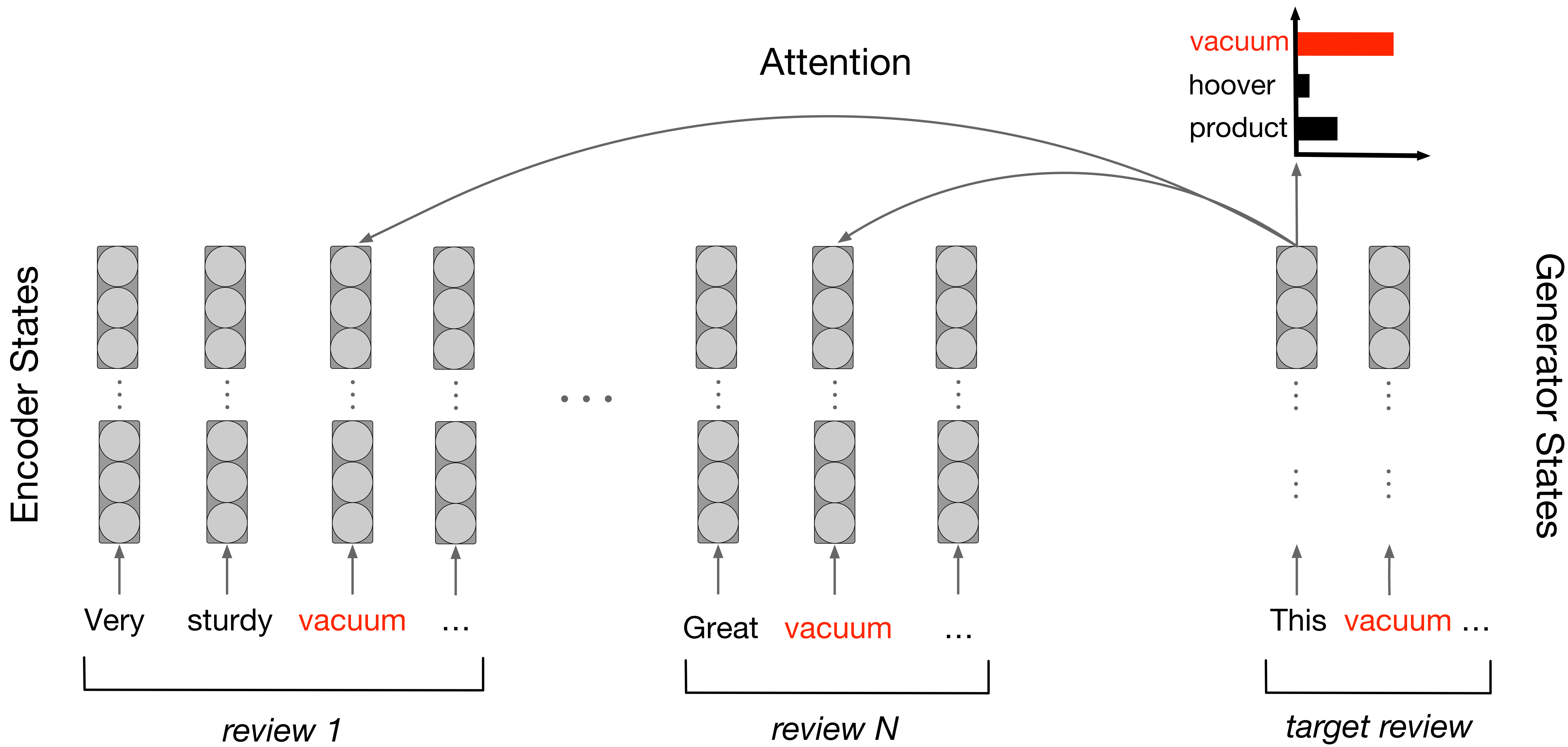
Leave-one-out



Leave-one-out



Leave-one-out



Review properties

- Observation:
  - Some reviews are more like summaries
  - Some are less

Review 1

Varys

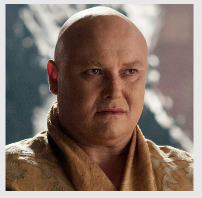
When I first got diabetes I got this. It has a lot of what we need. But later I have switched to another brand.

Review 1

Varys

When I first got diabetes I got this. It has a lot of what we need. But later I have switched to another brand.

Review 1



Varys

When I first got diabetes I got this. It has a lot of what we need. But later I have switched to another brand.

Review 2

Jon Snow

★★★★★

These capsules are a natural alternative to other over-the-counter medications. They are easy to swallow and have a great taste. Overall, great value for money.

Review 2

Jon Snow

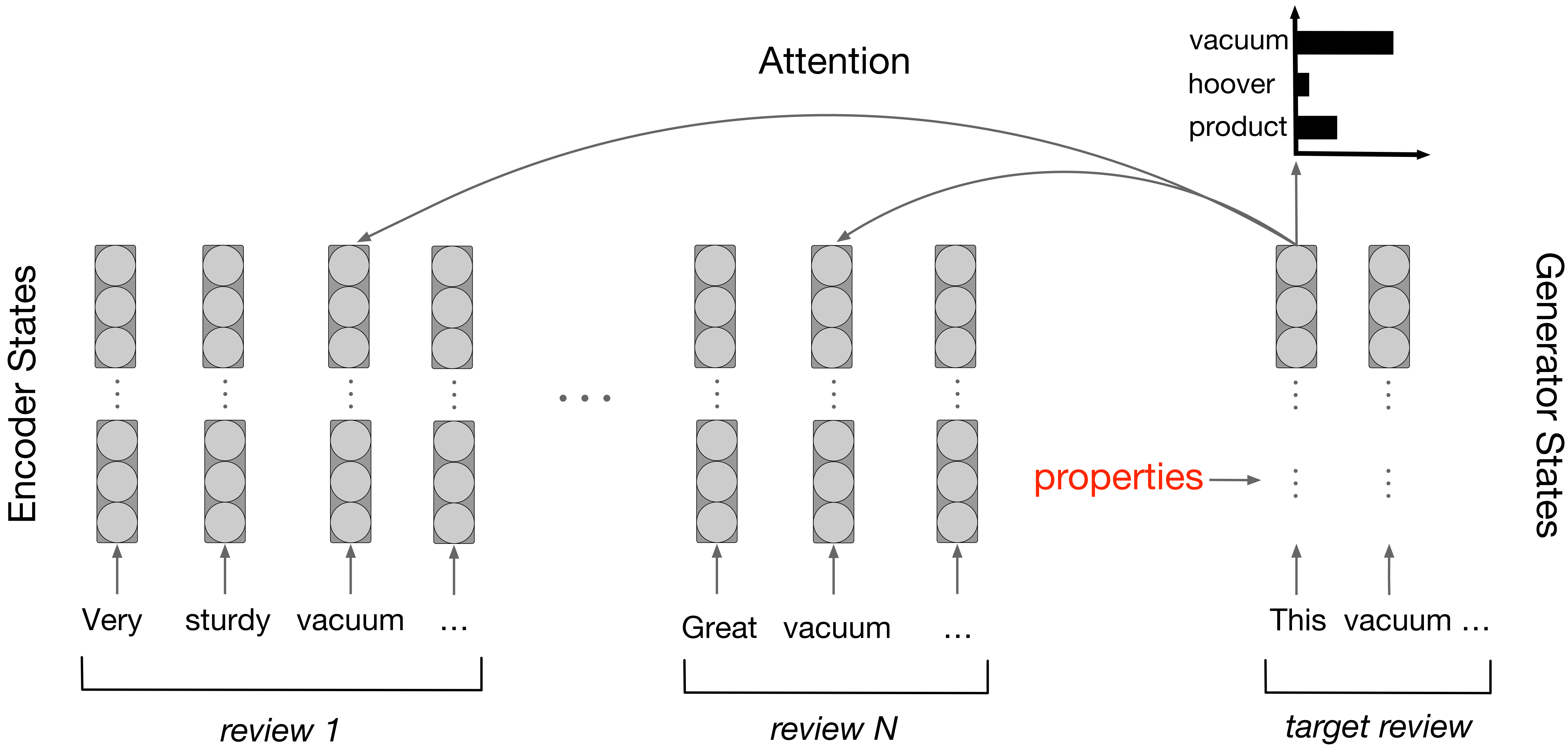
These capsules are a natural alternative to other over-the-counter medications. They are easy to swallow and have a great taste. Overall, great value for money.

Review 2

Jon Snow

These capsules are a natural alternative to other over-the-counter medications. They are easy to swallow and have a great taste. Overall, great value for money.

Properties



Attention
vacuum
hoover
product
Encoder States
Generator States
properties
Very
sturdy
vacuum
…
Great
vacuum
…
This
vacuum …
review 1
review N
target review

Property types

Type	Reviews	Summaries	Implementation
Information coverage	Uncommon	Common	ROUGE scores
Writing style	Informal	Formal	Pronoun counts
…	…	…	…

Writing style

- We found that conditioning **on pronoun counts** is a simple yet effective way to control the style of writing
- We categorized pronouns to the 1st, 2nd, 3rd points-of-view.
- One more class if a review has no pronouns

1st POV: personal experiences

- **I** bought this as a gift for my husband.
- I’ve been using Drakkar Noir Balm for over twenty years.
- I purchased these for my son as a kind of a joke.

2nd POV: recommendations

- This is the best product you can buy!
- You get what you pay for.
- Please do yourself a favor and avoid this product.

3rd POV:
formal writing style

- This is his every work day scent.
- It's very hard to buy the balm separately.
- It smells like Drakkar, but it is hard to find

No pronouns: aspects/utilization

- Very nice, not too overpowering.
- This product has no smell what ever.
- Nice to use for hardwood floors

Oracle

- Automatically computes **property values** based on:
  - target review
  - source reviews
- $q(r_{target}, \{r_1, ..., r_N\})$

Plug-in network

- At test time, want to generate **summaries**
- Have access only to source reviews - **can’t use the oracle**
- Might **not know** what **property values** are needed
- Replace the **oracle** by a **trainable neural network**

Plug-in network

- Using a **handful** of summaries (~30 data-points)
- Can train the **plug-in network**
- Learns what property values lead to **generation of summaries**

Recap

- **Pre-train**
  - Large corpus of reviews
  - Leave-one-out objective
  - Oracle that computes property values
- **Fine-tune**
  - Replace the oracle by the **plug-in network**
  - Fine-tune it on a **handful** of **human-written summaries**

Gold	These shoes run true to size, do a good job supporting the arch of the foot and are well-suited for exercise. They're good looking, comfortable, and the sole feels soft and cushioned. Overall they are a nice, light-weight pair of shoes and come in a variety of stylish colors.
FewSum	These running shoes are great! They fit true to size and are very comfortable to run around in. They are light weight and have great support. They run a little on the narrow side, so make sure to order a half size larger than normal.

Results on Amazon

	ROUGE-1	ROUGE-2	ROUGE-L
FewSum	**33.56**	**7.16**	**21.49**
Copycat	27.85	4.77	18.86
MeanSum	26.63	4.89	17.11
Lead	27.00	4.92	14.95

Alternative adaptation methods

Alternative adaptation

- Few-shot learning is not the only way to adapt to the target dataset
- Experimented with a number of alternatives

Amazon results

	ROUGE-1	ROUGE-2	ROUGE-L
Unsupervised learning	21.45	3.15	15.23

Unsupervised learning

Gold	These shoes run true to size, do a good job supporting the arch of the foot and are well-suited for exercise. They're good looking, comfortable, and the sole feels soft and cushioned. Overall they are a nice, light-weight pair of shoes and come in a variety of stylish colors.
USL	This is my second pair of Reebok running shoes and I love them. They are the most comfortable shoes I have ever worn.

Amazon results

	ROUGE-1	ROUGE-2	ROUGE-L
Unsupervised learning	21.45	3.15	15.23
Unsupervised learning + fine-tuning	28.23	6.24	19.64

Unsupervised learning + fine-tuning

Gold	These shoes run true to size, do a good job supporting the arch of the foot and are well-suited for exercise. They're good looking, comfortable, and the sole feels soft and cushioned. Overall they are a nice, light-weight pair of shoes and come in a variety of stylish colors.
USL+F	This is my second pair of Reebok running shoes and they are the best running shoes I have ever owned. They are lightweight, comfortable, and provide great support for my feet.

Amazon results

	ROUGE-1	ROUGE-2	ROUGE-L
Unsupervised learning	21.45	3.15	15.23
Unsupervised learning + fine-tuning	28.23	6.24	19.64
FewSum	**33.56**	**7.16**	**21.49**

FewSum

Gold	These shoes run true to size, do a good job supporting the arch of the foot and are well-suited for exercise. They're good looking, comfortable, and the sole feels soft and cushioned. Overall they are a nice, light-weight pair of shoes and come in a variety of stylish colors.
FewSum	These running shoes are great! They fit true to size and are very comfortable to run around in. They are light weight and have great support. They run a little on the narrow side, so make sure to order a half size larger than normal.

Human evaluation

- We asked AMT workers to judge summaries based on a number of criteria (fluency, informativeness, etc)
- The results suggest **a substantial preference** for FewSum

Learning Opinion Summarizers by Selecting Informative Reviews

Arthur Bražinskas, Mirella Lapata, Ivan Titov

EMNLP 2021

Motivation

Motivation

- Datasets in the domain are **very scarce**
- Makes it hard to **develop** and **train** models
- Supervised models often require large datasets for training

Available Datasets

	#Entities	#Summaries	Domain
MeanSum (Chu and Liu, 2019)	200	200	Yelp
Copycat (Bražinskas et al., 2020)	60	180	Amazon
FewSum (Bražinskas et al., 2020)	60	180	Amazon
SpaCe (Angelidis et al., 2021)	50	1,050	TripAdvisor

Unsupervised Abstractive Methods

- **MeanSum** (Chu and Liu, 2019)
- **Copycat** (Bražinskas et al. 2020)
- **OpinionDigest** (Suhara et al. 2020)
- **DenoiseSum** (Amplayo et al., 2020)
- **SelfSum** (Elsahar et al., 2020)
- **RecurSum** (Isonuma et al., 2020)
- **MultimodalSum** (Im et al., 2021)
- …

Low-resource Methods

- **FewSum** (Bražinskas et al. 2020)
- **PASS** (Oved and Levy, 2021)

Contributions

- We provide the **largest dataset** for multi-document abstractive opinion summarization
- A novel model that **selects** and **summarizes** reviews from large collections **end-to-end**

AmaSum

AmaSum

- More than **33,000 summaries** for more than **31,000** Amazon products
- Each paired with more than **320 reviews**, on average
- Human-written by **professional product reviewers**
- Extracted from popular web portals

AmaSum

	# Entities	Rev/Ent	# Summaries	Domain
AmaSum (this work)	**31,483**	**326**	**33,324**	**Amazon**
SpaCe (Angelidis et al., 2020)	50	100	1,050	Tripadvisor
Copycat (Bražinskas et al., 2020)	60	8	180	Amazon
FewSum (Bražinskas et al., 2020)	60	8	180	Amazon
MeanSum (Chu and Liu, 2019)	200	8	200	Yelp

AmaSum

- Summaries consist of:
  - Verdicts
  - Pros and cons

Example

Olympus E-500 EVOLT

Verdict

The Olympus Evolt E-500 is a compact, easy-to-use digital SLR camera with a broad feature set for its class and very nice photo quality overall.

Pros

- Compact design
- Strong autofocus performance
- Intuitive and easy-to-navigate menu system

Cons

- Unreliable automatic white balance
- Slow start-up time when dust reduction is enabled

Challenges

- Each summary is paired with more than **320 reviews**, on average
- Standard encoding-decoding can be challenging
- Not all **reviews content** covers the **summary content**
- Training on **random review subsets** leads to **hallucinations in test time** (show in this work)
- We address these challenges by introducing SelSum

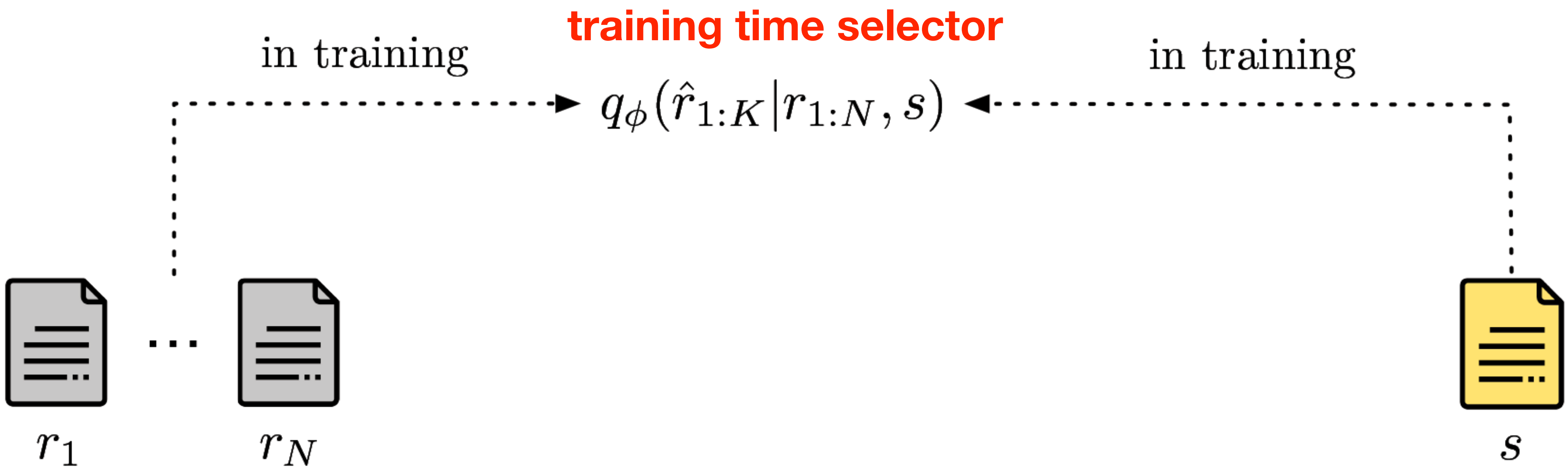
SelSum

SelSum

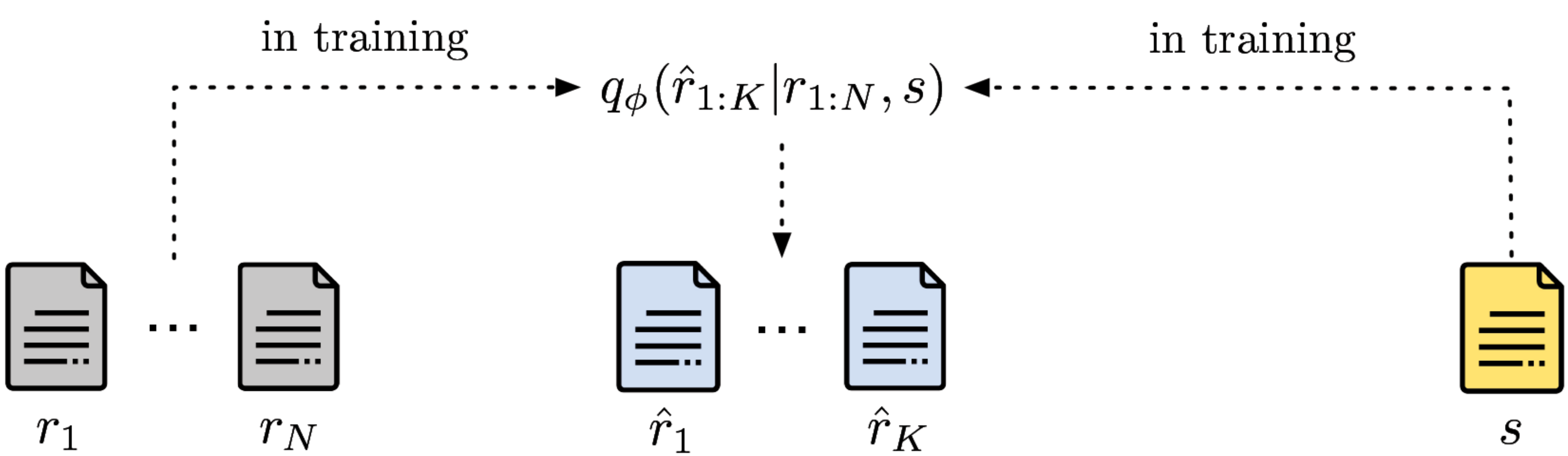
- A probabilistic latent model that **selects** and **summarizes** reviews end-to-end
- Learns to select **subsets** of **summary relevant reviews** in training

Review Selection

Review Selection

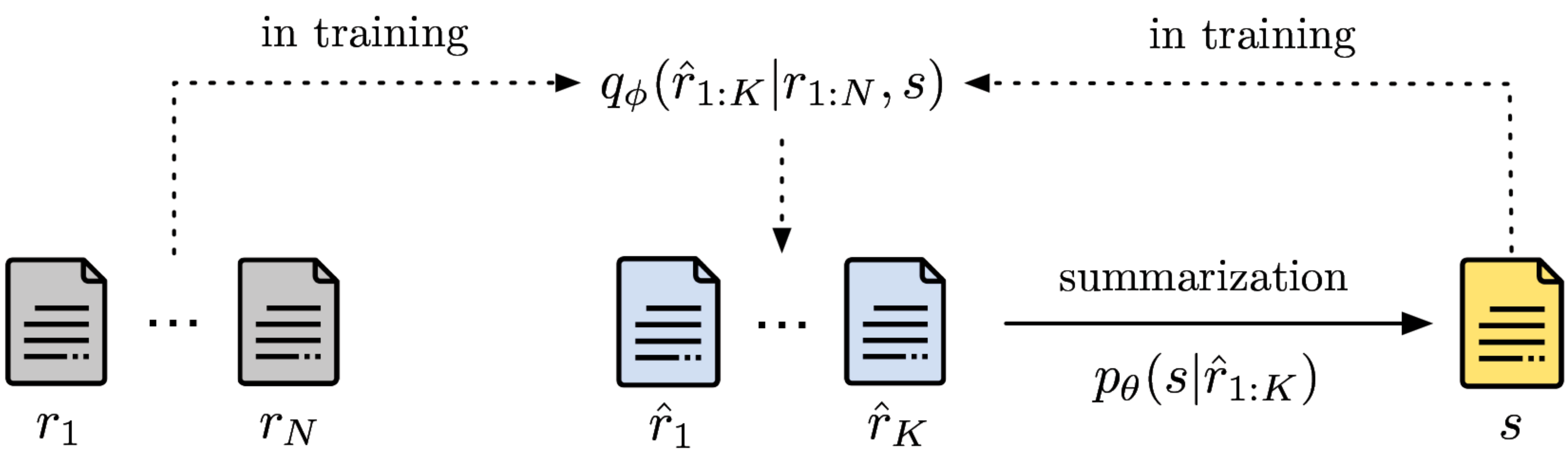


Review Selection



N > K

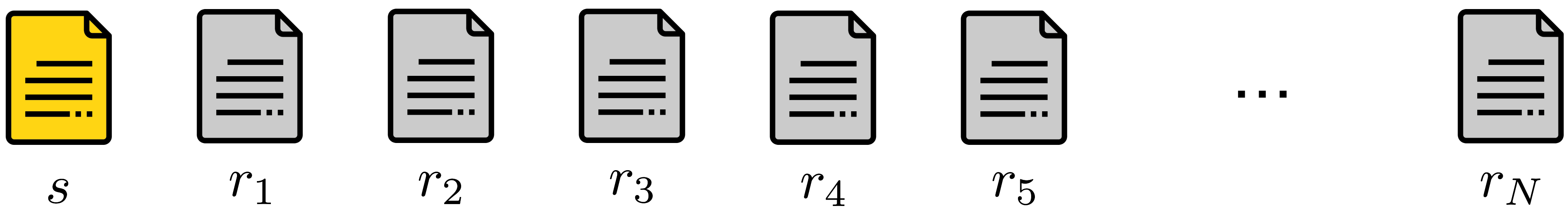
Review Selection



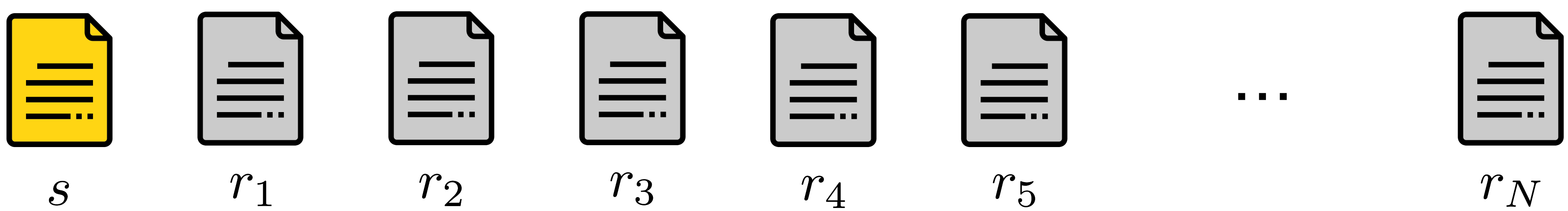
Training Time Selector

- Review subsets are treated as vectors of **categorical variables** (K slots)
- **Sampling without replacement**

Training Time Selector

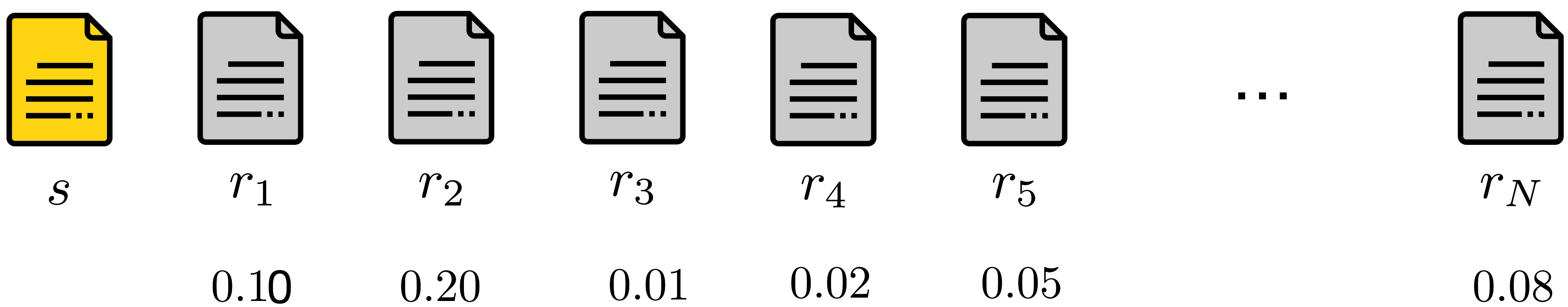


Training Time Selector



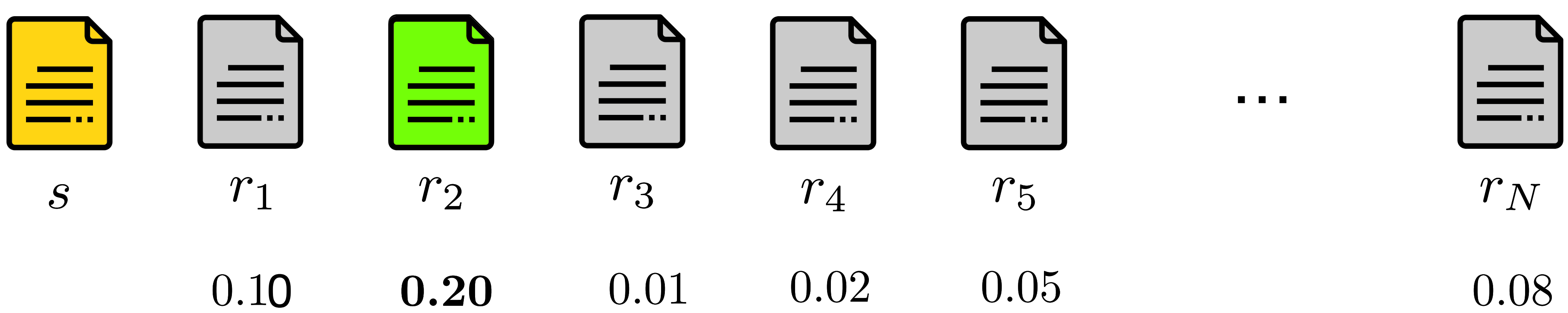
$$q_\phi(\hat{r}_1 | r_{1:N}, s)$$

Training Time Selector



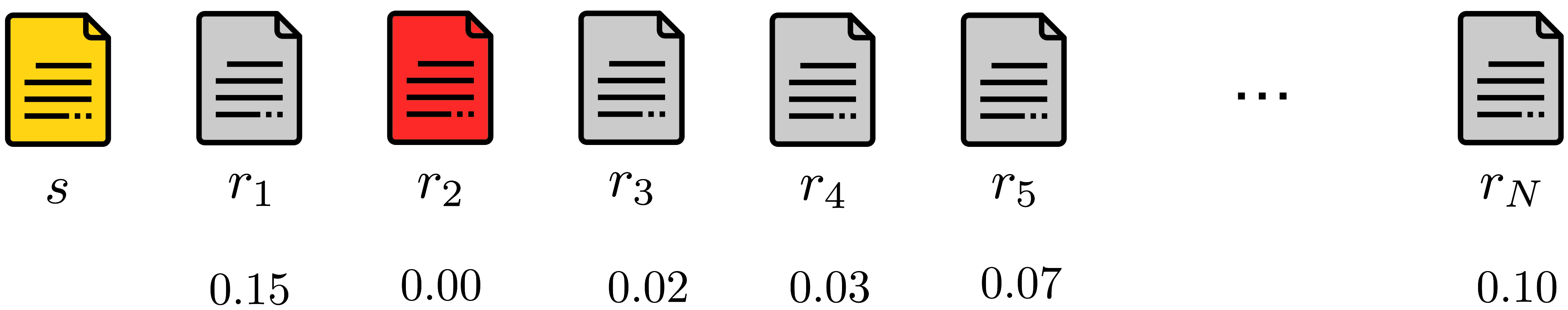
$$q_\phi(\hat{r}_1 | r_{1:N}, s)$$

Training Time Selector



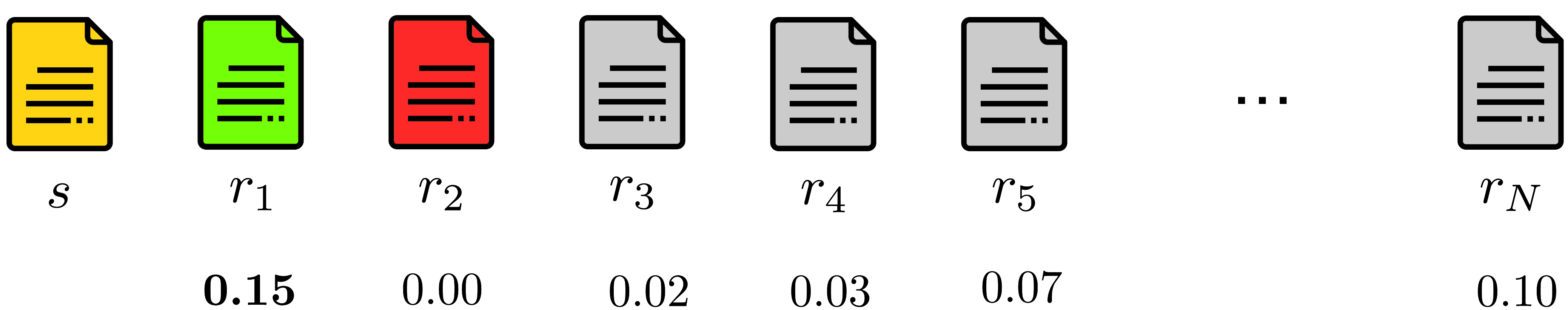
$$\hat{r}_1 \sim q_\phi(\hat{r}_1 | r_{1:N}, s)$$

Training Time Selector



$$q_\phi(\hat{r}_2 | r_{1:N}, \hat{r}_1, s)$$

Training Time Selector



$$\hat{r}_2 \sim q_\phi(\hat{r}_2 | r_{1:N}, \hat{r}_1, s)$$

Model Training

- Sampling categorical variable assignments is **not differentiable**
- To train the selector and summarizer **end-to-end** we use:
  - Amortized variational inference (Kingma and Welling, 2013; Cremer et al., 2018)
  - REINFORCE (Williams, 1992)

Review Selection

- **Computational** and **memory savings**
    - Only the subset is encoded using the deep encoder
- Better **interpretability** of the generated output
- **Fewer hallucinations** (as we show)

Lexical Features

- Training time selector inputs **review representations**
- Represent each review in the collection with **pre-computed 23 features**
- Feed to a tiny non-linear neural network (< 0.1% params of the model)
- Minimal computational burden in training

Feature Examples

- ROUGE scores between a **review** and **summary**
- ROUGE scores between a **review** and the **other ones in the collection** (measures uniqueness)
- Aspect keyword-based scores
  - Used a vocabulary of aspect keywords
  - Counted their occurrence in reviews and summaries
  - Computed recall and precision scores
- …

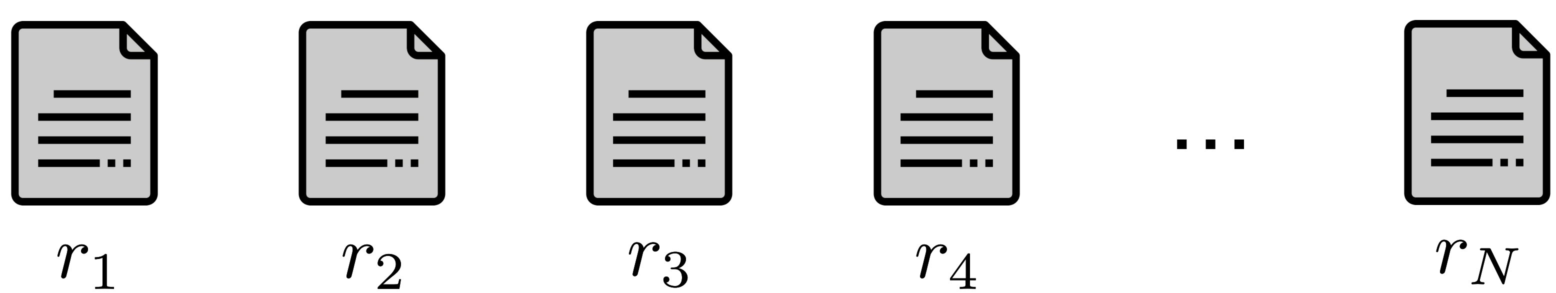
Test Time Selector

- In test time would like to select and summarize **informative reviews**
- Can’t use the **training time selector**
    - summary is **not available** in **test time**
    - fit a **test time selector** that relies only on reviews (Razavi et al., 2019)

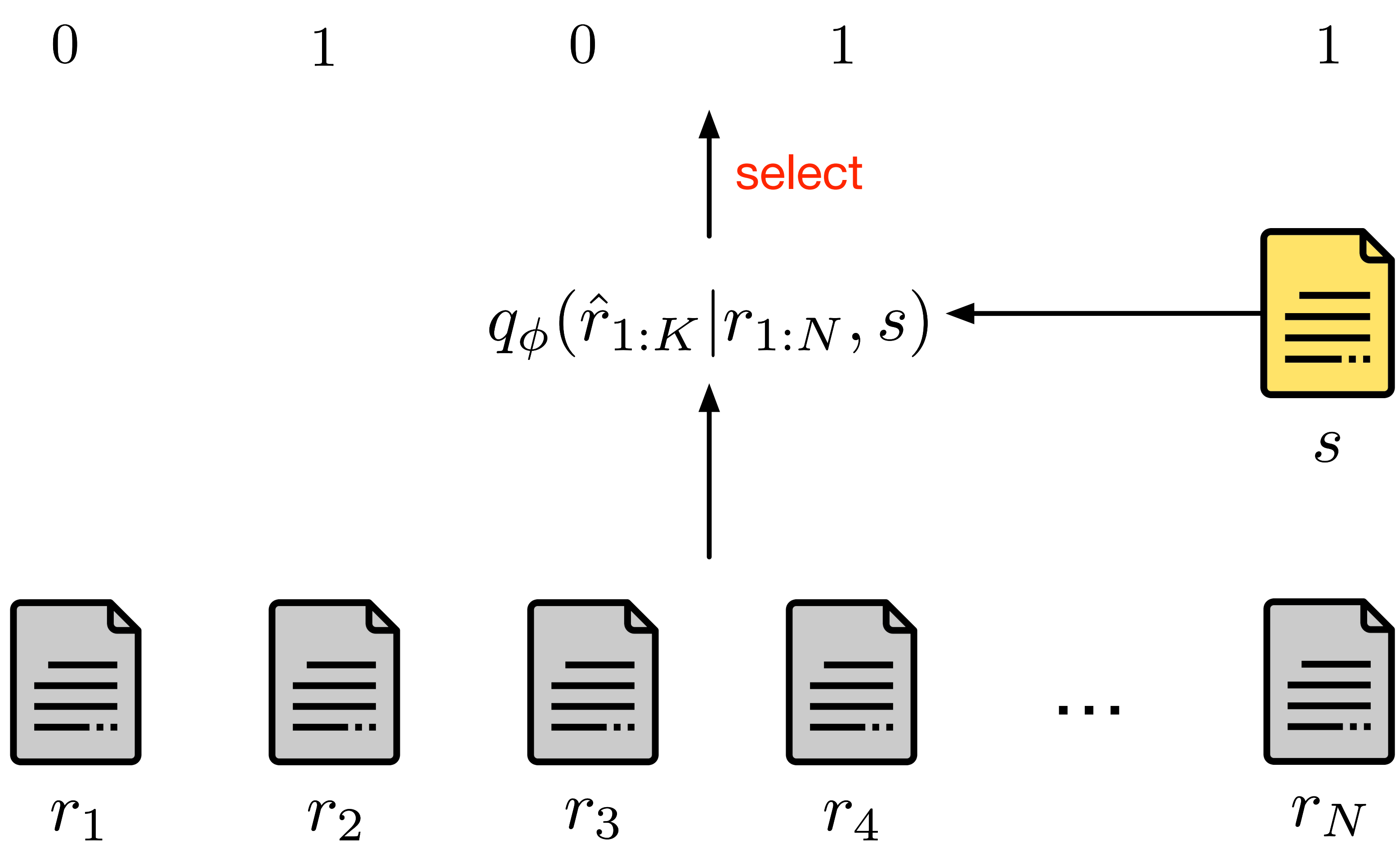
Test Time Selector

- Select reviews using the **training time selector**
- Fit the **test time selector** to predict the selected reviews

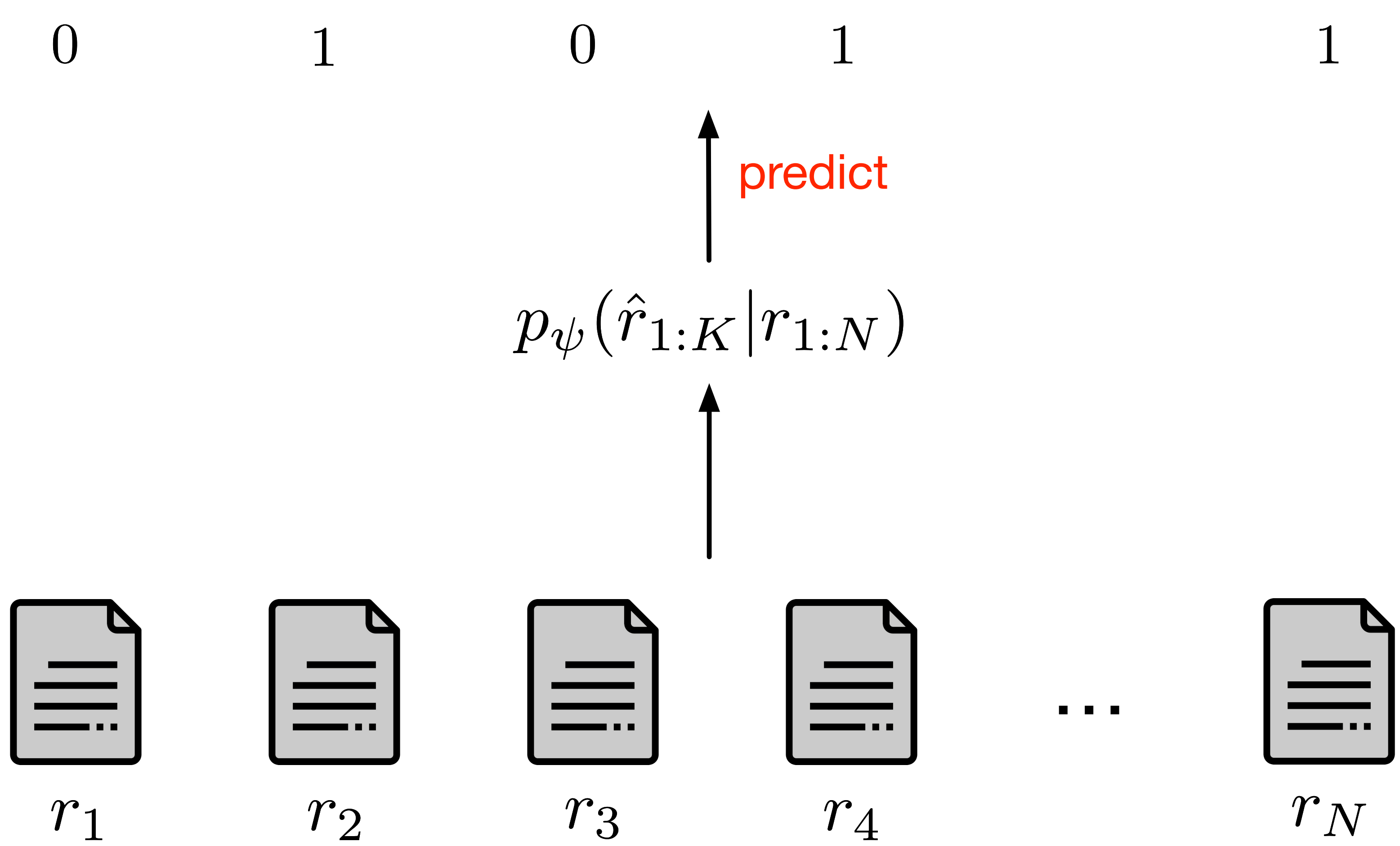
Test Time Selector



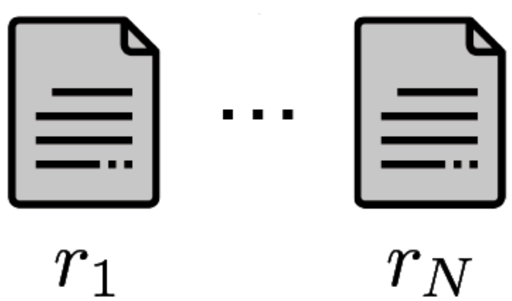
Test Time Selector



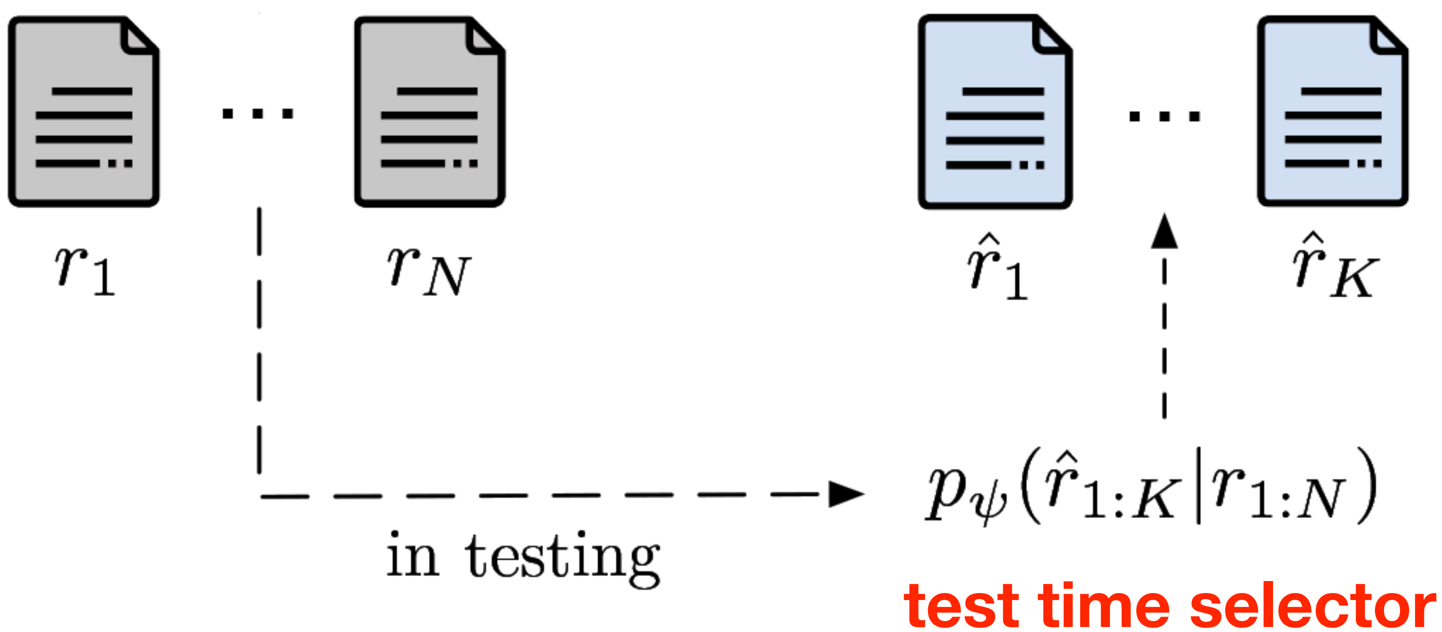
Test Time Selector



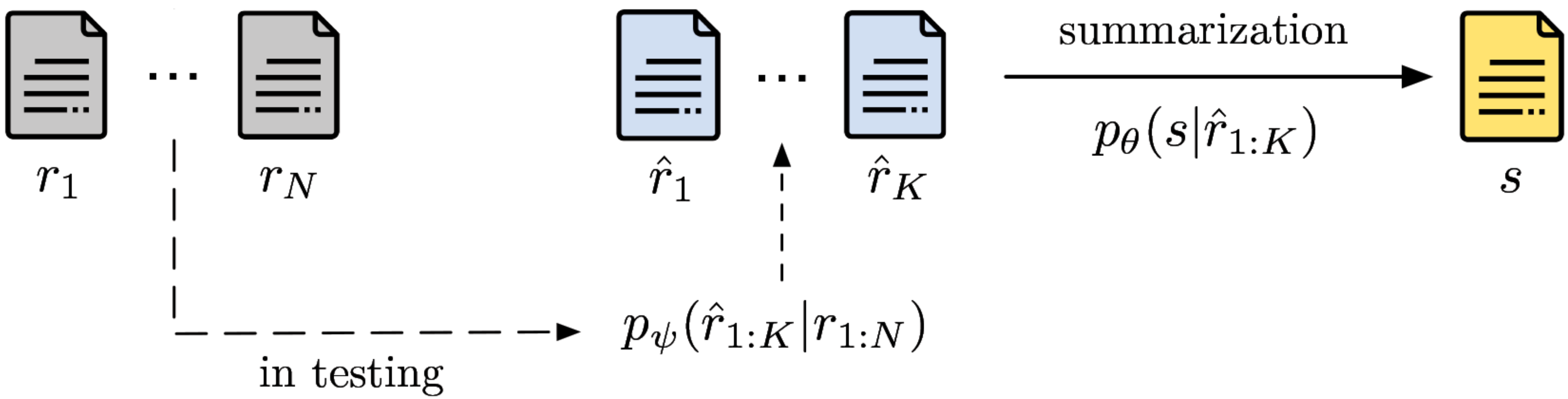
Test Time Selector



Test Time Selector



Test Time Selector



Setup and Results

Splits

- **Training**: 26,660 summaries
- **Validation**: 3,302 summaries
- **Testing**: 3,362 summaries

Summarizer

- Pre-trained BART (Lewis et al, 2020) **encoder-decoder**
- Verdicts, pros and cons were **concatenated** together as one string

Training Time Selector

- Feed-forward network inputing static features
- Selecting **10** out of **100** reviews

Test Time Selector

- Pre-trained BART encoder on the end-task to **represent reviews**
- Feed-forwards to tag reviews

Baseline Models

- **Random**: random sentences from reviews
- **Oracle**: greedy selection of sentences with maximum ROUGE-1 and -2 scores to the summary
- **LexRank** (Erkan and Radev, 2004): unsupervised extractive
- **MeanSum** (Chu and Liu, 2019): unsupervised abstractive
- **Copycat** (Bražinskas et al, 2020): unsupervised abstractive
- **ExtSum** (ours): supervised extractive summarizer

Review Selectors

- Experimented with review selectors (**non-learned**)
- **RandSel:**
  - Random selection of reviews
- **R1 top-K**:
  - **K highest scored** reviews based on ROUGE-1 with respect to the **summary**
  - Before test time, fit the test time selector

Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77

Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24

Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29

Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89

Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44

Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44
RANDSEL	23.25	4.75	17.82	20.26	3.60	18.52	13.59	2.32	11.86

Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44
RANDSEL	23.25	4.75	17.82	20.26	3.60	18.52	13.59	2.32	11.86
R1 TOP-K	23.43	4.94	18.52	22.01	3.94	19.84	14.93	2.57	12.96

Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44
RANDSEL	23.25	4.75	17.82	20.26	3.60	18.52	13.59	2.32	11.86
R1 TOP-K	23.43	4.94	18.52	22.01	3.94	19.84	14.93	2.57	12.96
SELSUM	24.33	5.29	18.84	21.29	4.00	19.39	14.96	2.60	13.07

Automatic Evaluation

	Verdict			Pros			Cons		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
ORACLE	38.14	11.76	31.50	37.22	10.53	33.50	34.09	10.75	29.66
RANDOM	13.12	0.82	10.85	14.29	1.04	13.02	9.91	0.72	8.77
LEXRANK	15.12	1.84	12.60	14.12	1.50	12.81	8.28	0.82	7.24
MEANSUM	13.78	0.93	11.70	10.44	0.63	9.55	5.95	0.45	5.29
COPYCAT	17.05	1.78	14.50	15.12	1.48	13.85	6.81	0.82	5.89
EXTSUM	18.74	3.01	15.74	19.06	2.47	17.49	11.63	1.19	10.44
RANDSEL	23.25	4.75	17.82	20.26	3.60	18.52	13.59	2.32	11.86
R1 TOP-K	23.43	4.94	18.52	**22.01**	3.94	**19.84**	14.93	2.57	12.96
SELSUM	**24.33**	**5.29**	**18.84**	21.29	**4.00**	19.39	**14.96**	**2.60**	**13.07**

Content Support

- **ROUGE is not always reliable** for assessing how **input faithful** summaries are (Tay et al., 2019; Bražinskas et al., 2020)

- Generation of **input faithful** summaries is **crucial** for practical applications

- Remains an **open problem** (Maynez et al., 2020; Fabbri et al., 2020; Want et al., 2020)

- Performed **human evaluation** via Amazon Mechanical Turk (AMT)

Content Support

- Evaluated different selectors
- Summarizer remained exactly the same

Content Support

- Asked AMT workers to assess **faithfulness** of each summary sentence to input reviews by marking them as:
  - Fully supported
  - Partially supported
  - Not supported
- Normalized these to percentages

Content Support

	Verdict			**Pros**			**Cons**		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RANDSEL	28.96	45.90	25.14	38.62	29.10	32.28	14.92	14.60	70.48

Content Support

	Verdict			Pros			Cons		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RANDSEL	28.96	45.90	25.14	38.62	29.10	32.28	14.92	14.60	70.48
R1 TOP-K	55.21	31.77	13.02	56.07	26.61	17.31	33.33	27.78	38.89

Content Support

	Verdict			Pros			Cons		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RANDSEL	28.96	45.90	25.14	38.62	29.10	32.28	14.92	**14.60**	70.48
R1 TOP-K	55.21	31.77	13.02	56.07	26.61	17.31	33.33	27.78	38.89
SELSUM	**66.08**	**25.15**	**8.77**	**70.21**	**17.99**	**11.80**	**38.41**	29.21	**32.38**

Content Support

- Investigated the role of **better review subsets** in test time
- We selected reviews using the **SelSum's test time selector**
- Input them to the summarizer trained on **random review subsets (RandSel)**
- Indicated by *

Content Support

	Verdict			Pros			Cons		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RANDSEL	28.96	45.90	25.14	38.62	29.10	32.28	14.92	14.60	70.48
RANDSEL*	50.79	31.75	17.46	50.62	22.96	26.42	16.84	**13.75**	69.42

Content Support

	Verdict			Pros			Cons		
	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓	Full↑	Partial↓	No↓
RandSel	28.96	45.90	25.14	38.62	29.10	32.28	14.92	14.60	70.48
RandSel*	50.79	31.75	17.46	50.62	22.96	26.42	16.84	**13.75**	69.42
R1 top-k	55.21	31.77	13.02	56.07	26.61	17.31	33.33	27.78	38.89
SelSum	**66.08**	**25.15**	**8.77**	**70.21**	**17.99**	**11.80**	**38.41**	29.21	**32.38**

Take Away

- Random review subsets **might not cover well the content of summaries**
- A summarizer trained on these reviews **learns to hallucinate**
- Evident when better review subsets are provided in **test time**

Conclusions

- We contribute the **largest dataset** for **multi-document opinion summarization** (more than 33,000 summaries)
- Propose an end-to-end model **selecting** and **summarizing** reviews
- Show that learned review selection leads to generation of **input faithful summaries**

Dataset and Codebase

Publicly available:

https://github.com/abrazinskas/SelSum

Example Summary

Verdict	If you like the idea of a glass feeder, this is the one to get. It has a lot to offer for the price.
Pros	• Has a large opening that makes it easy to get in and out of the feeder • Has a nice design that's easy to clean
Cons	• The lid is a little flimsy, and it's not as durable as some of the other models
Reviews	... looks just as nice as the glass feeders ||... Very happy with the value, quality and function ... ||... the cheapest most flexible "jar" I've ever seen ... ||... Nice large opening so it's easy to pour the sugar water ||... This feeder has a nice large opening ... ||... this is the perfect design and size ... ||The hummingbirds liked it and had no trouble feeding or perching.... || ... The main compartment is easy to clean... ||... The top is a little flimsy ... ||... it fell out of the hanger it broke for good ... there are so many other nice ones out there that have glass "jar's" or at least sturdier plastic ... ||... The tray is easy to clean ...

Wrap up

Overview

- News summarization
  - Objective facts
  - Mostly single-document
- Opinion summarization
  - Subjective information
  - Multi-document

Open Problems

- **Hallucinations** are one of the central problems in **summarization**
- Hard to evaluate automatically
- Multi-document encoding is hard

Contact me

abrazinskas@ed.ac.uk