Automatic Text Summarization

Arthur Bražinskas
The University of Edinburgh, Scotland

SCHOOL OF DATA ANALYSIS

Yandex

About me

Born in Lithuania

Klaipeda

BSc in Computer Science

Programming languages and algorithms

Aarhus Tech,
Aarhus, Denmark

MSc in Computer Science

Classical AI, data mining, theoretical CS algorithms

IT University of Copenhagen,
Copenhagen, Denmark

MSc (exchange)

Evolutionary algorithms, neural networks, data mining

Victoria University of Wellington
Wellington, New Zealand

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

Copenhagen
Denmark

Amsterdam
Netherlands

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 text summarization** in **low-resource settings**
- Also interested in:
    - deep generative models
    - variational inference
    - latent graphical models

Agenda of this lecture

- Overview of **models** and **methods** in **text summarization**
- Overview of two main domains:
  - news articles
  - customer reviews (opinions)
- Datasets
- Open problems

What is Summarization?

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’

Summarization: Different Perspectives

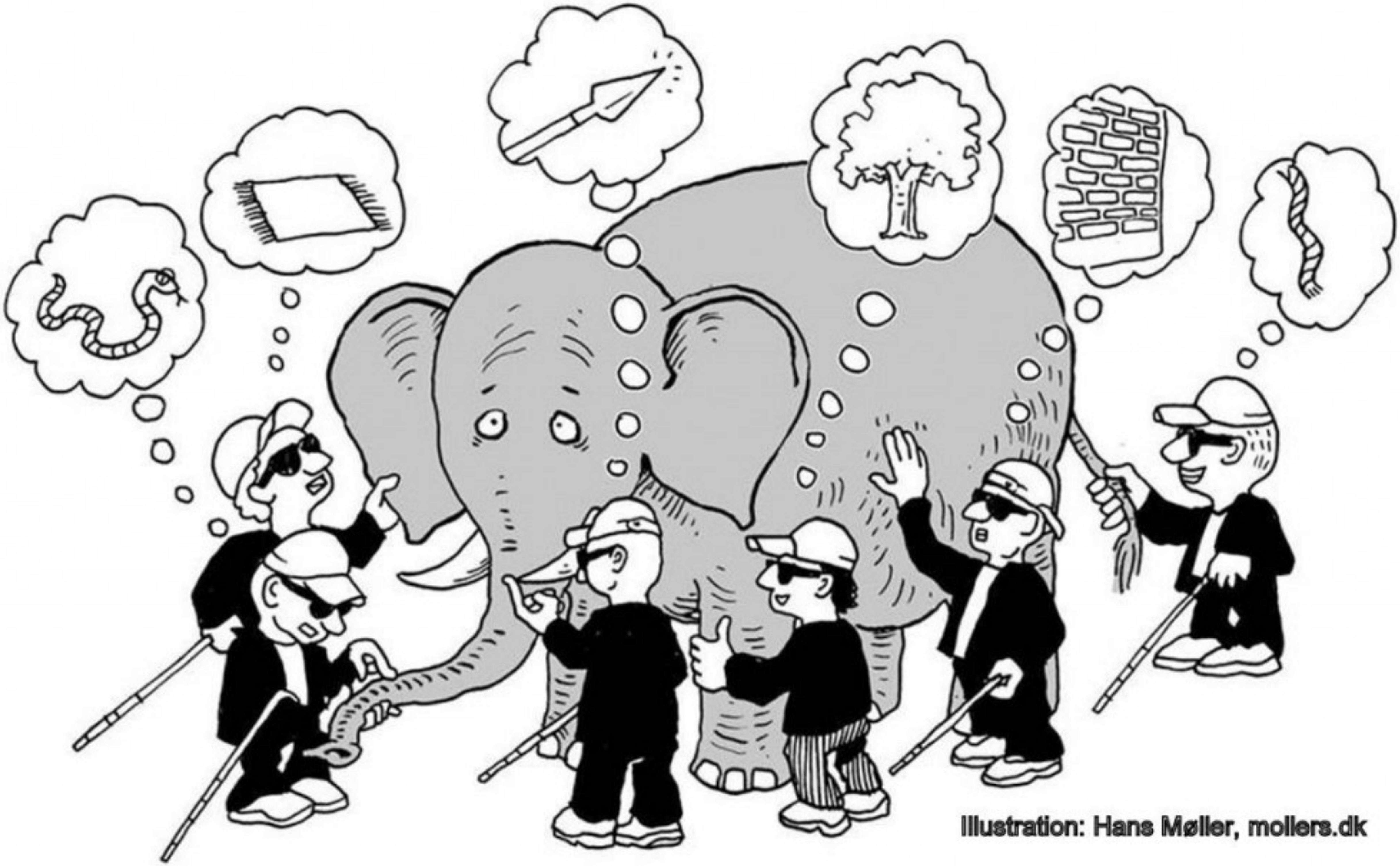


Illustration: Hans Møller, mollers.dk

Statistics

Data summarization

- Say we have some continues data
- Instead of storing the whole dataset
- We can store its ‘**summary**’
- E.g., **sufficient statistics** (Wasserman, 2005), **moments** or **learned parameters**
- **Preference/importance** is given to parameters that capture dynamics of the true model

Information Theory

Lossy compression

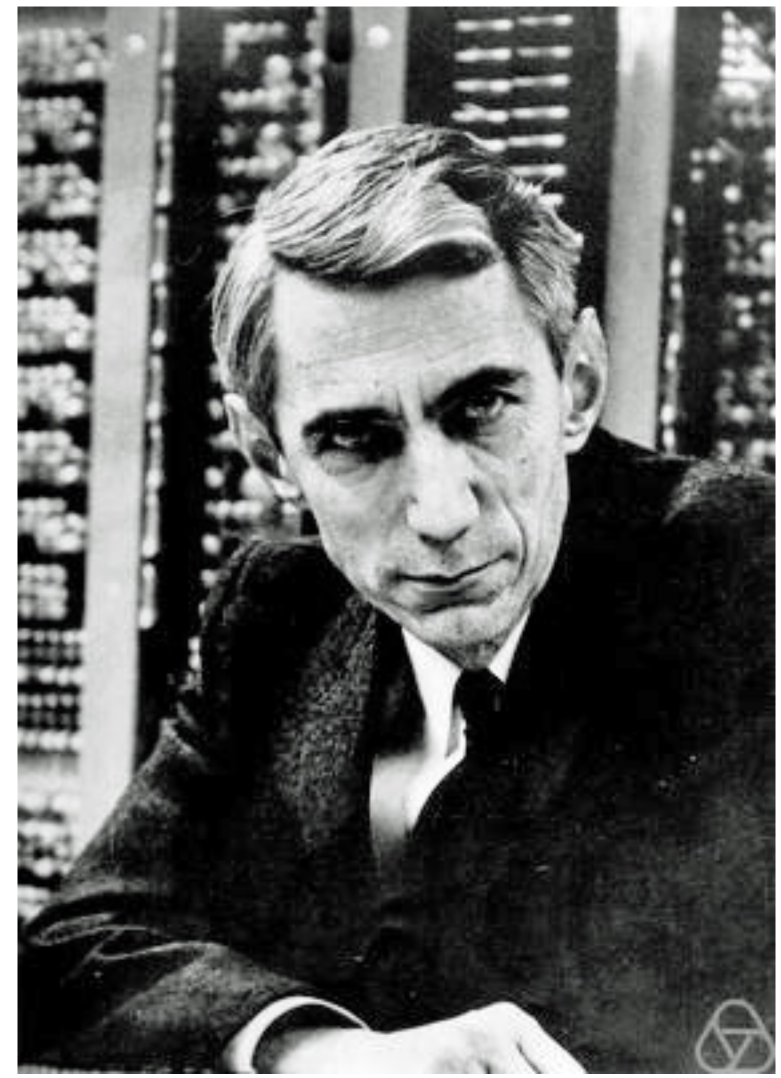
- Want to compress (+binary represent) i.i.d. discrete observations: $X \sim F$
- Want **reduce** the expected length of the binary string below **H(X)** (optimal code)
- Ok with not being able to decode **some** symbols

Lossy compression

- One way to think about lossy compression is that we perform binary representation of **‘the most important’** symbols or a ‘**summary**’ of symbols
- Don’t care about the rest
- What symbols are important?
- The ones that **are frequent**

The noisy-channel coding theorem

Error-free communication over a discrete channel is achievable by **a block code encoder-decoder** with a **rate** up to the **channel capacity**.



Claude Shannon

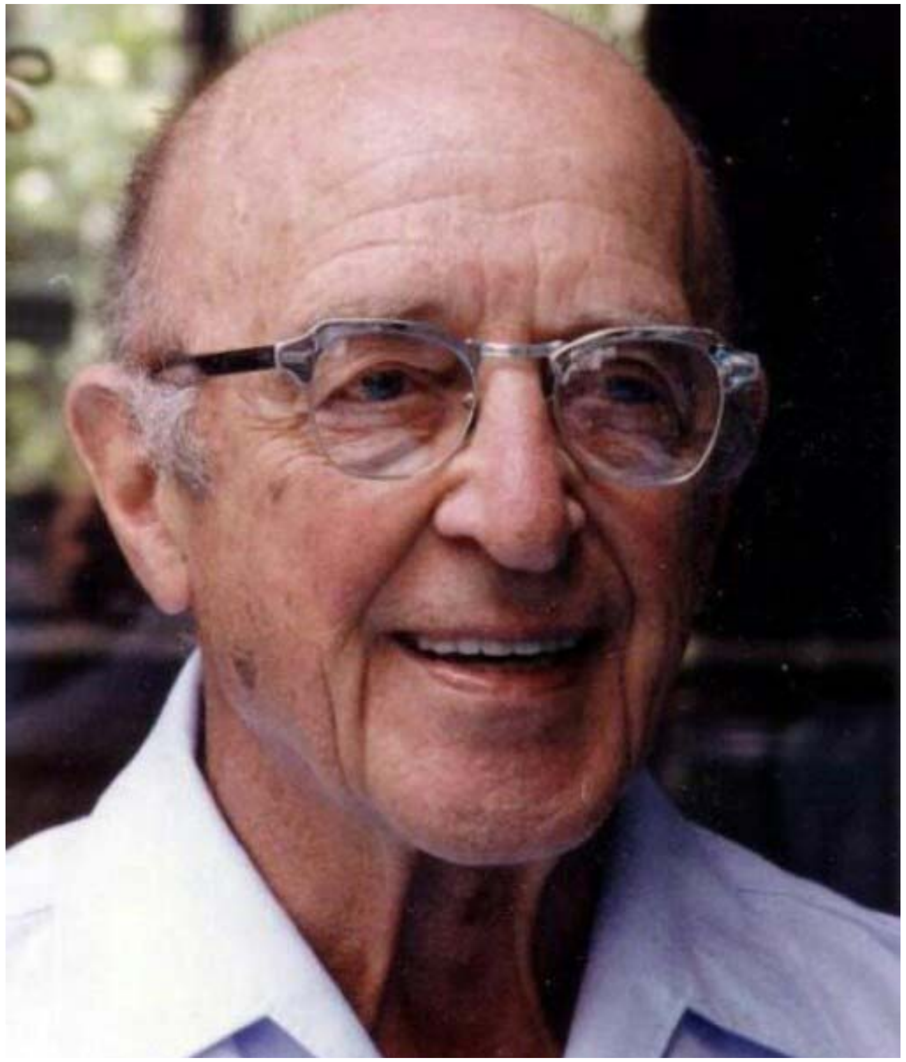
The noisy-channel coding theorem

- The proof builds on a summarizing subset of block codes (**typical set**) (McKay, 2003)
- $$T_{N\beta} \equiv \left\{ \mathbf{x} \in \mathcal{A}_X^N : \left| \frac{1}{N} \log_2 \frac{1}{P(\mathbf{x})} - H \right| < \beta \right\}$$

Psychology

Carl Rogers

- American psychologist (1902-1987)
- The founder of **client-centered approach**
- Emphasizes the individual's inherent drive toward **self-actualization**



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)

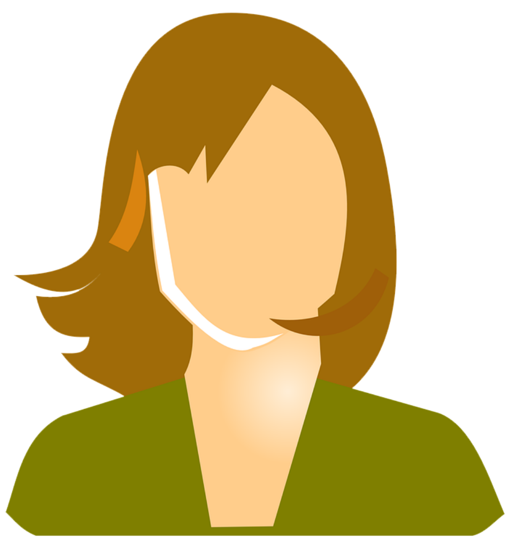
Conceptually similar to **abstractive summarization (reduce, paraphrase, retain what is important)**

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

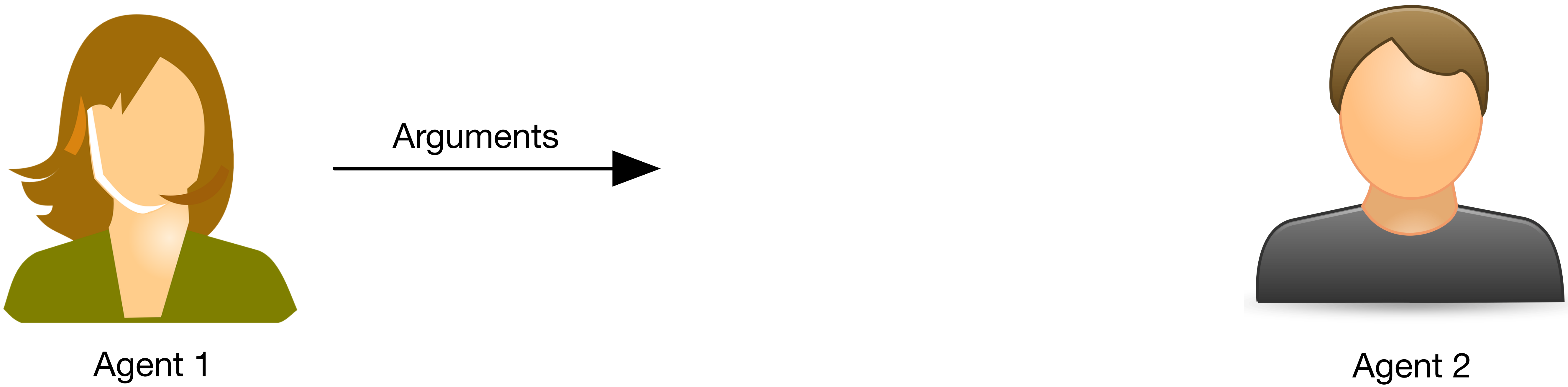
Therapy



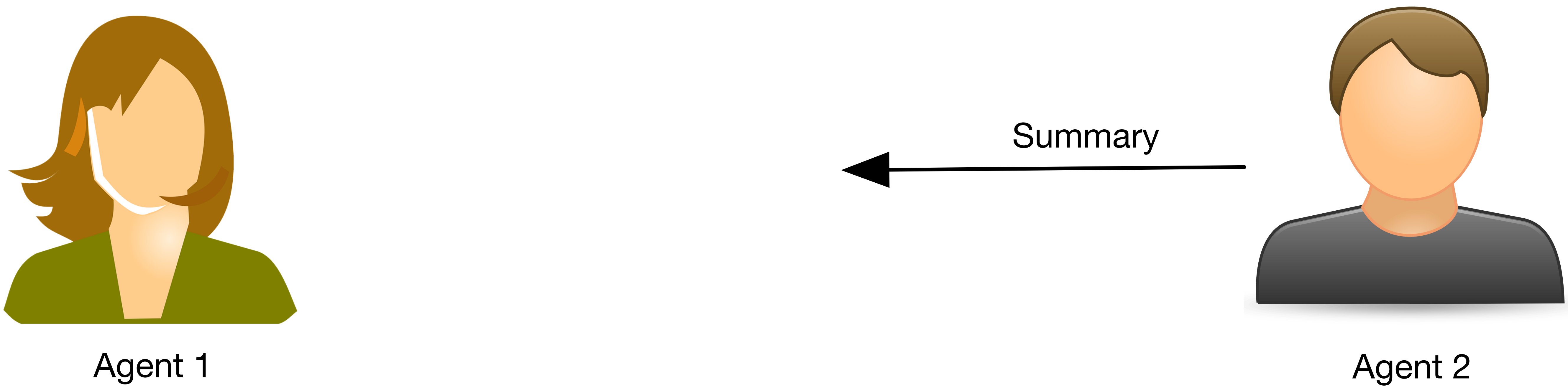
Agent 1

Agent 2

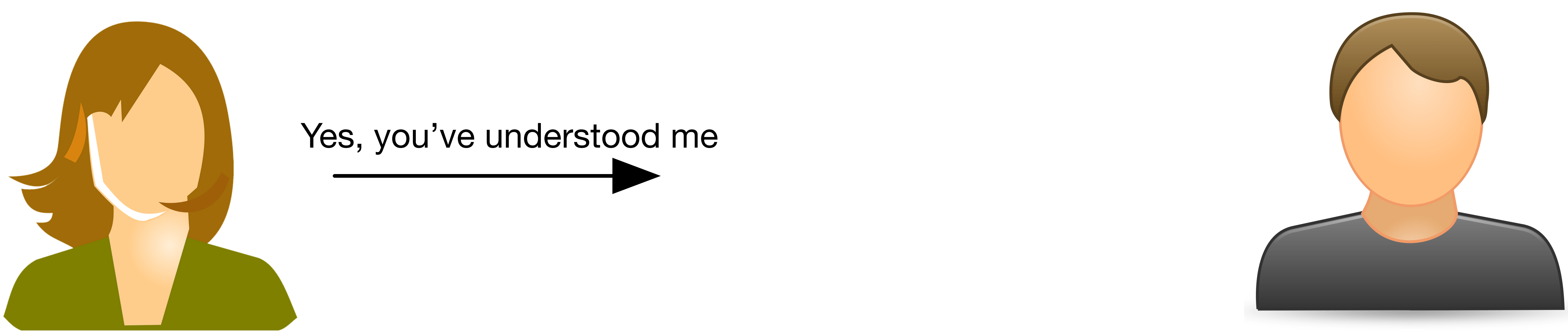
Therapy



Therapy



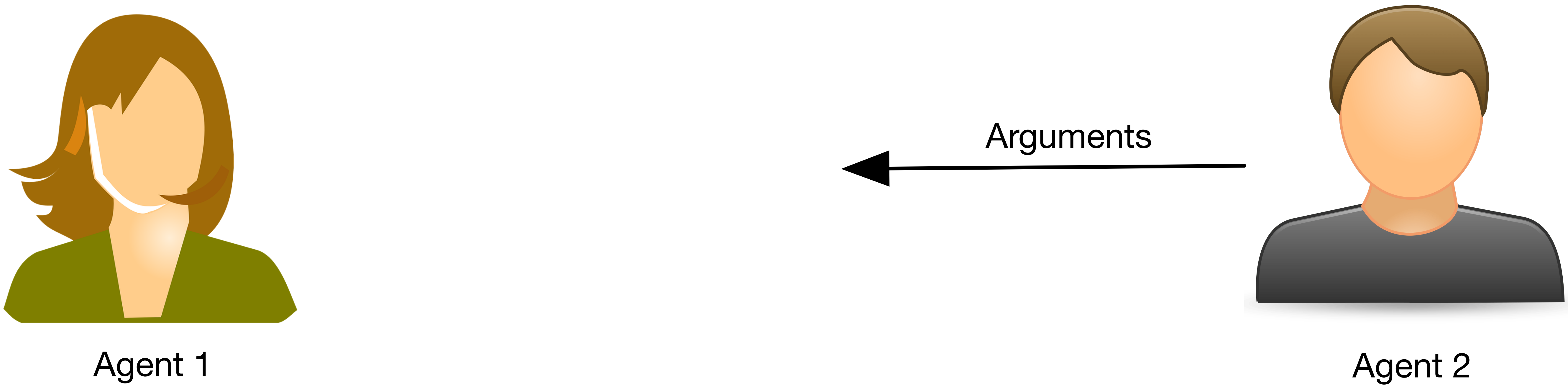
Therapy



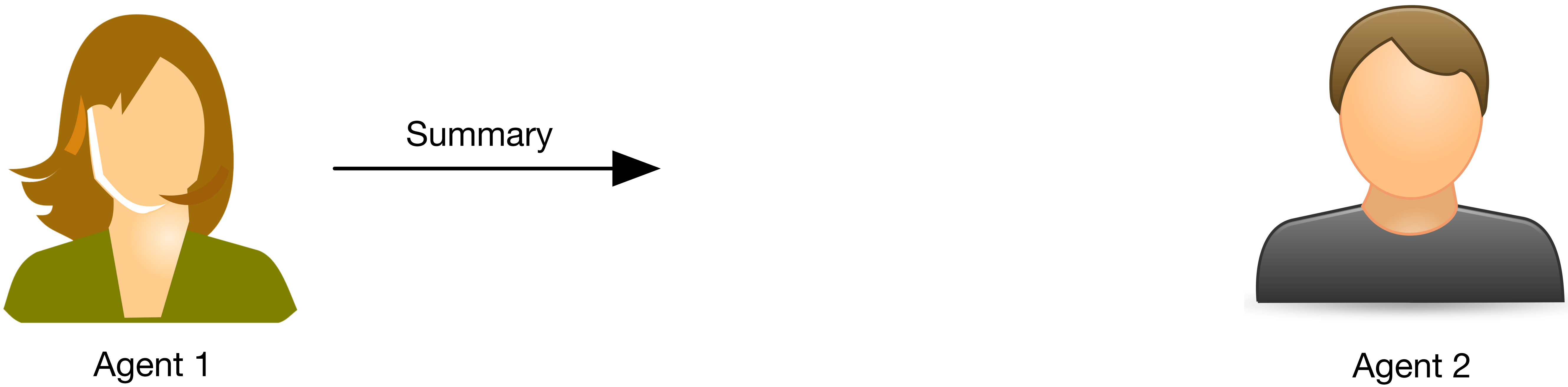
Agent 1

Agent 2

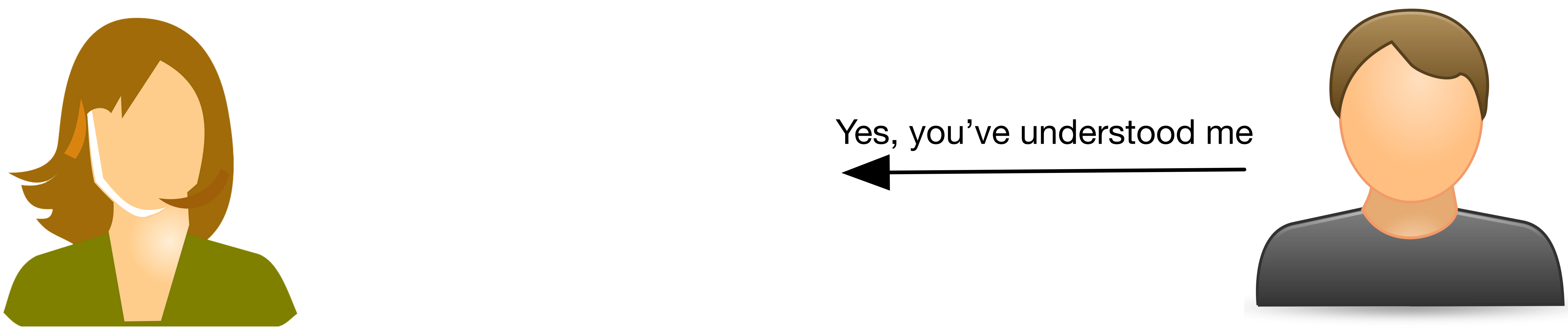
Therapy



Therapy



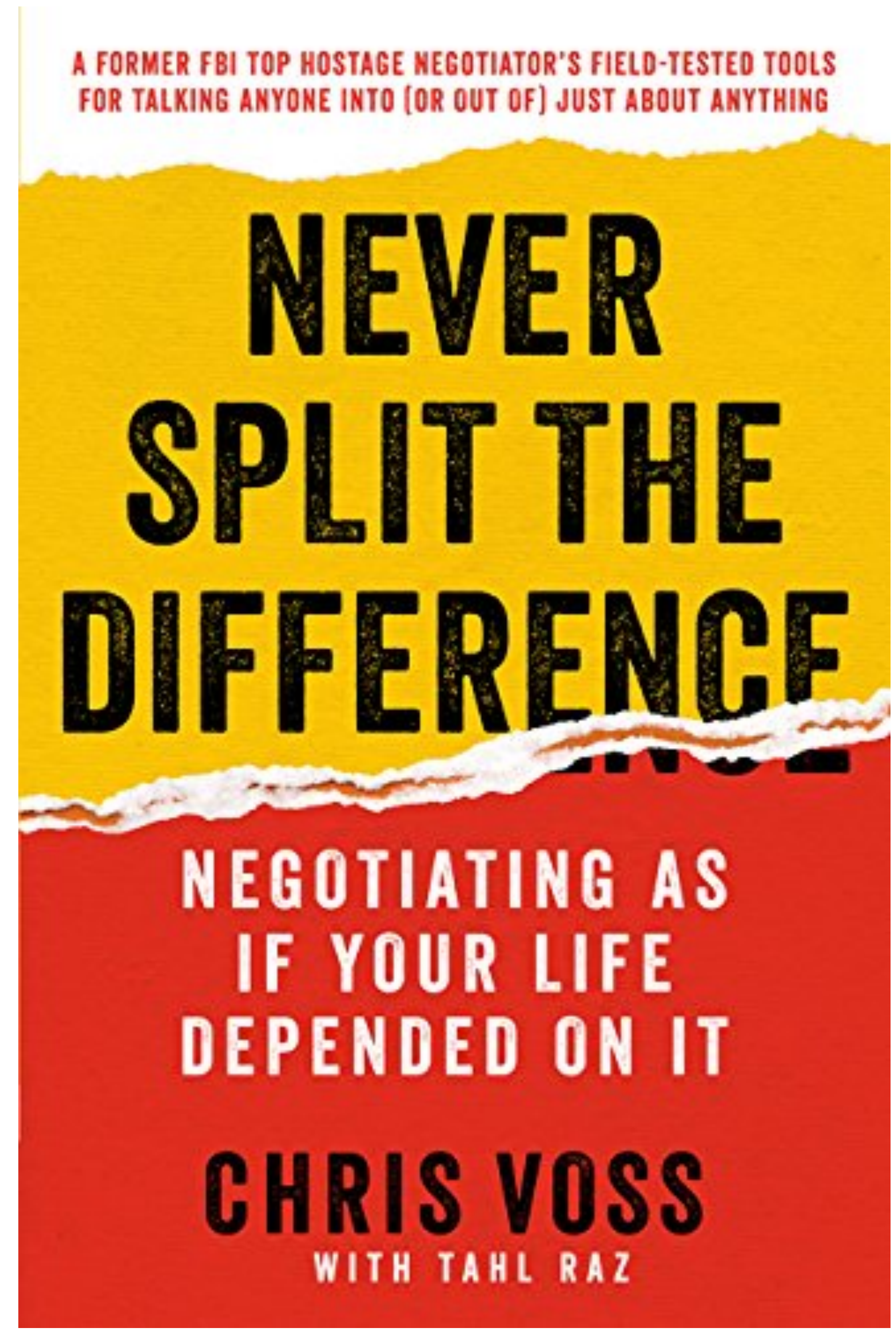
Therapy



Agent 1

Agent 2

Negotiations



Schema

- **Input data:** visual and auditory signal
- **Summarizer:** (one or multiple) agents
- **What to preserve?** what is important to the oner person
- **Goal:** conflict resolution / negotiations

Text Summarization

Why summarization

- The amount of text documents available online is **enormous**
- **Summarization allows for:**
  - Fast information **skimming/consumption**
  - Faster **decision making**
  - Downstream utilization (analysis)

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

Summarization flavors

Summarization flavors

Boring vanilla

Summarization flavors

Boring vanilla

Extractive

Summarization flavors

Boring vanilla

Birthday cake

Extractive

Summarization flavors

Boring vanilla

Extractive

Birthday cake

Abstractive

Summarization flavors

Boring vanilla

Extractive

Birthday cake

Abstractive

Methods

Summarization flavors

Boring vanilla

Birthday cake

Salt & caramel

Extractive

Abstractive

Summarization flavors

Boring vanilla

Extractive

Birthday cake

Abstractive

Salt & caramel

Contrastive

Summarization flavors

Boring vanilla | Birthday cake | Salt & caramel | Chocolate & vodka




Extractive | Abstractive | Contrastive

Summarization flavors

Boring vanilla

Extractive

Birthday cake

Abstractive

Salt & caramel

Contrastive

Chocolate & vodka

Extreme

Summarization flavors

Boring vanilla

Extractive

Birthday cake

Abstractive

Salt & caramel

Contrastive

Chocolate & vodka

Extreme

Fruity blend

Summarization flavors

Boring vanilla

Extractive

Birthday cake

Abstractive

Salt & caramel

Contrastive

Chocolate & vodka

Extreme

Fruity blend

Consensus

Summarization flavors

Boring vanilla

Extractive

Birthday cake

Abstractive

Salt & caramel

Contrastive

Chocolate & vodka

Extreme

Fruity blend

Consensus

Summary formats

Summarization flavors

Boring vanilla

Extractive

Birthday cake

Abstractive

Salt & caramel

Contrastive

Chocolate & vodka

Extreme

Fruity blend

Consensus

Methods

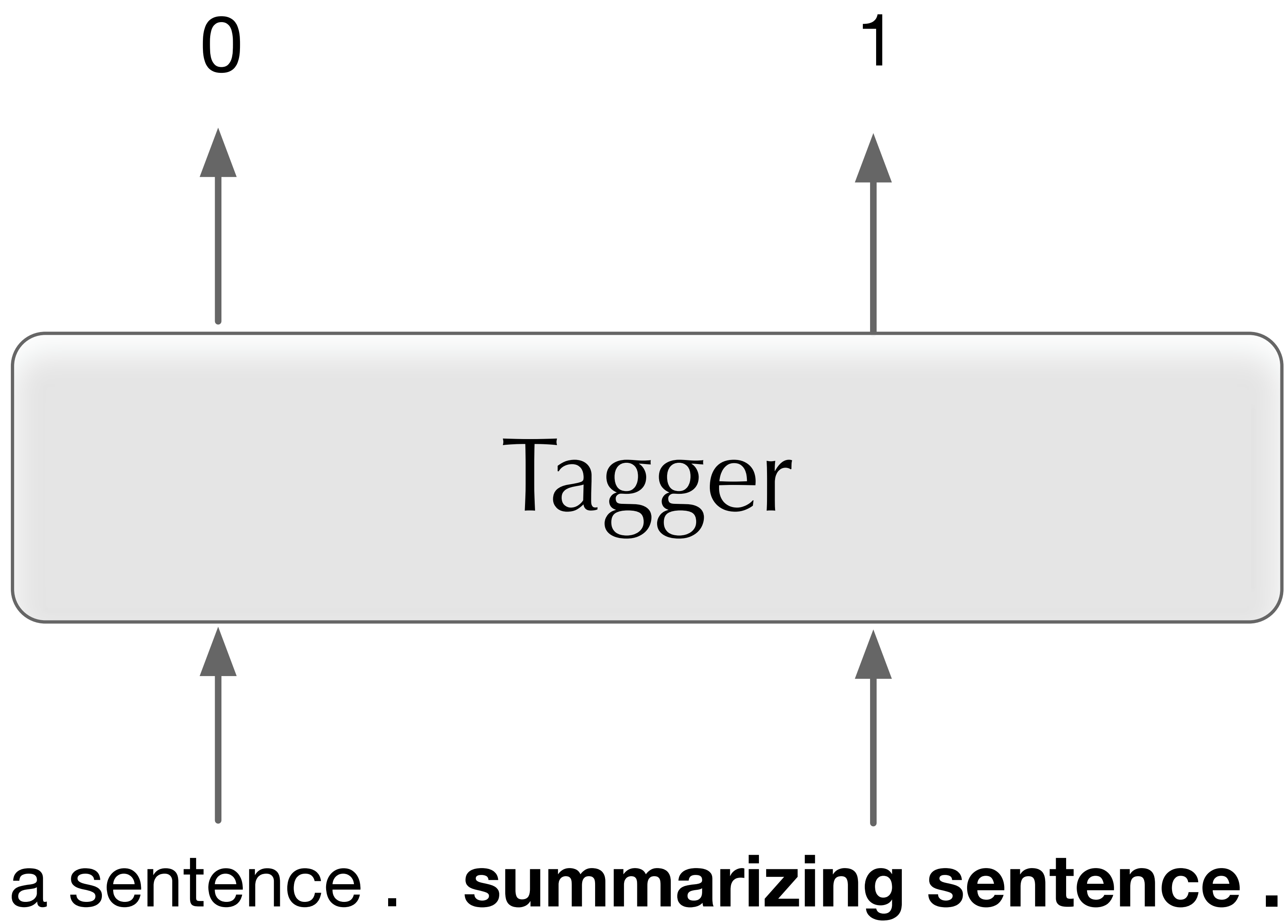
Summary formats

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 it’s pre-trained before the start (Liu and Lapata, 2019)

Extractive methods

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

Extractive data

- In most cases, we don’t have explicit ‘extractive’ datasets
- Instead, we can **utilize abstractive reference summaries** to produce the training dataset
- 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

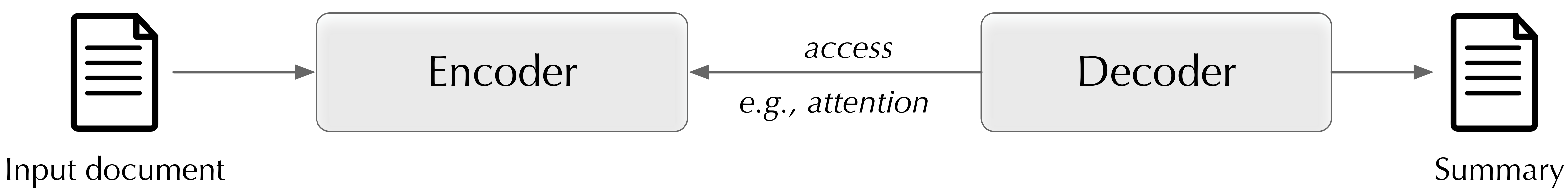
Extractive methods

- **Pros:**
  - Easy-to-build models
  - Always factually correct summaries
  - Fast training and inference
  - Less data demanding
- **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:**
  - Require **large annotated datasets** for training
  - Prone to **hallucinations** (iPhone vs iPad)

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

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

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

ROUGE-L

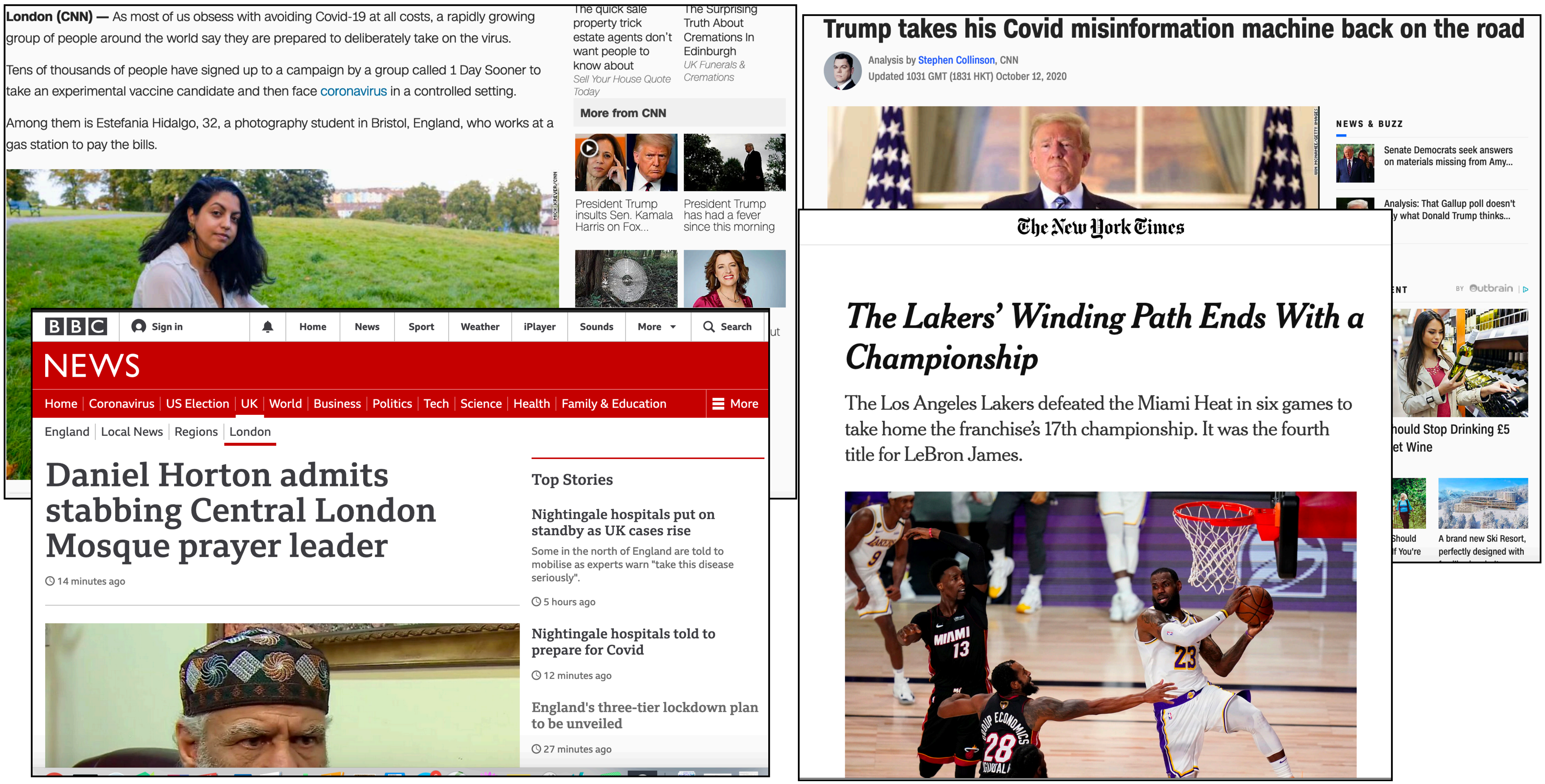
- Based on the longest common subsequence
- Gaps are allowed
- **The most important sub-metric** in summarization
- **Correlated with fluency** (harder for extractive systems to score highly)

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)

News Summarization: Basics

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.

...hould Stop Drinking £5 ...et Wine

A brand new Ski Resort, perfectly designed with

Summarization of news

Input article

Summarization of news

Input article

~700 words

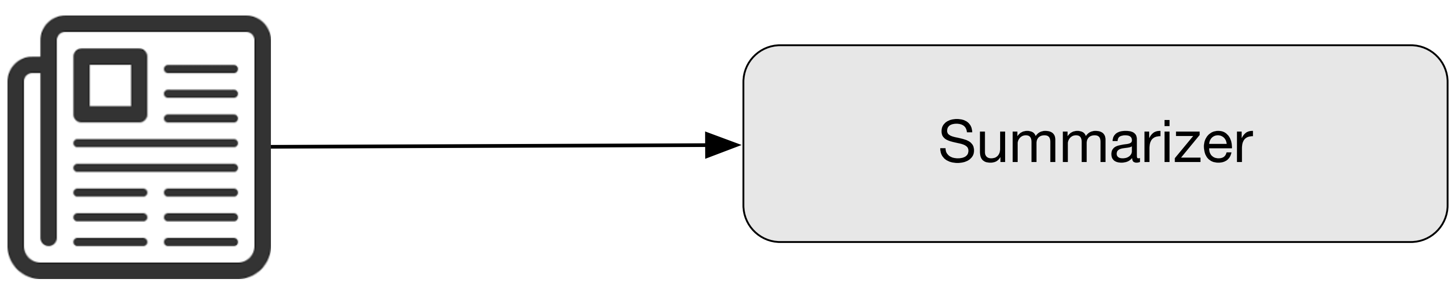
Summarization of news

Input article

~700 words

3.5 min

Summarization of news

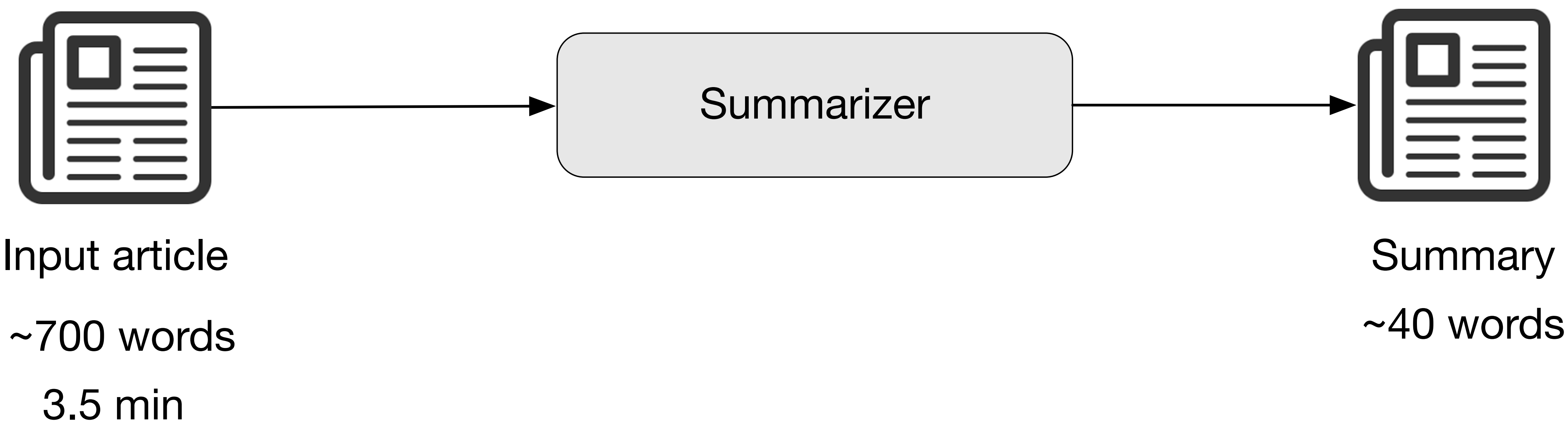


Input article

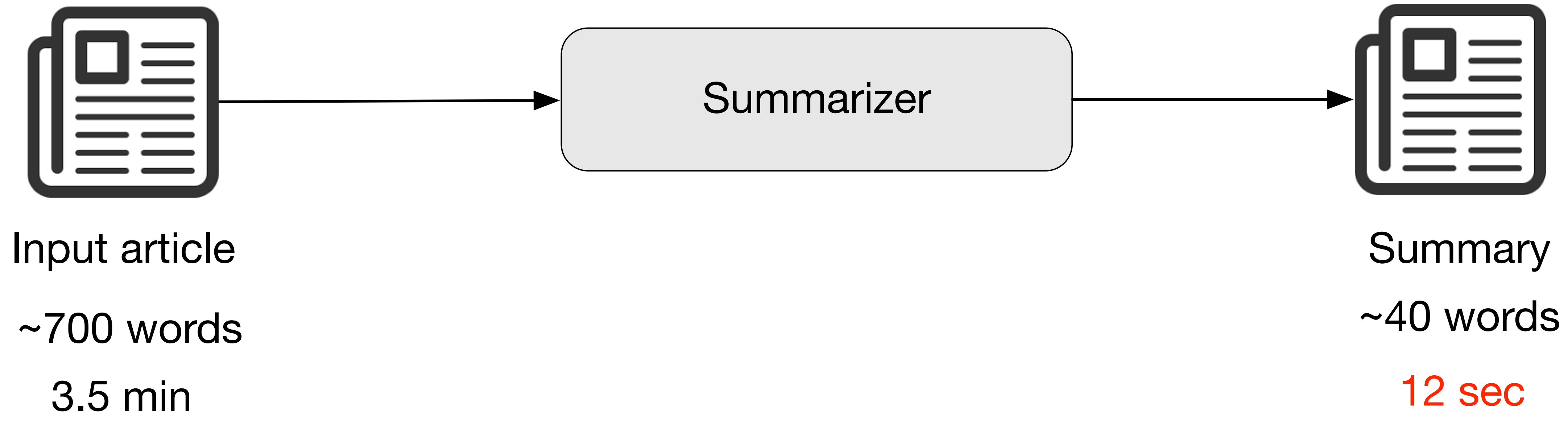
~700 words

3.5 min

Summarization of news



Summarization of news



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

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

source document

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

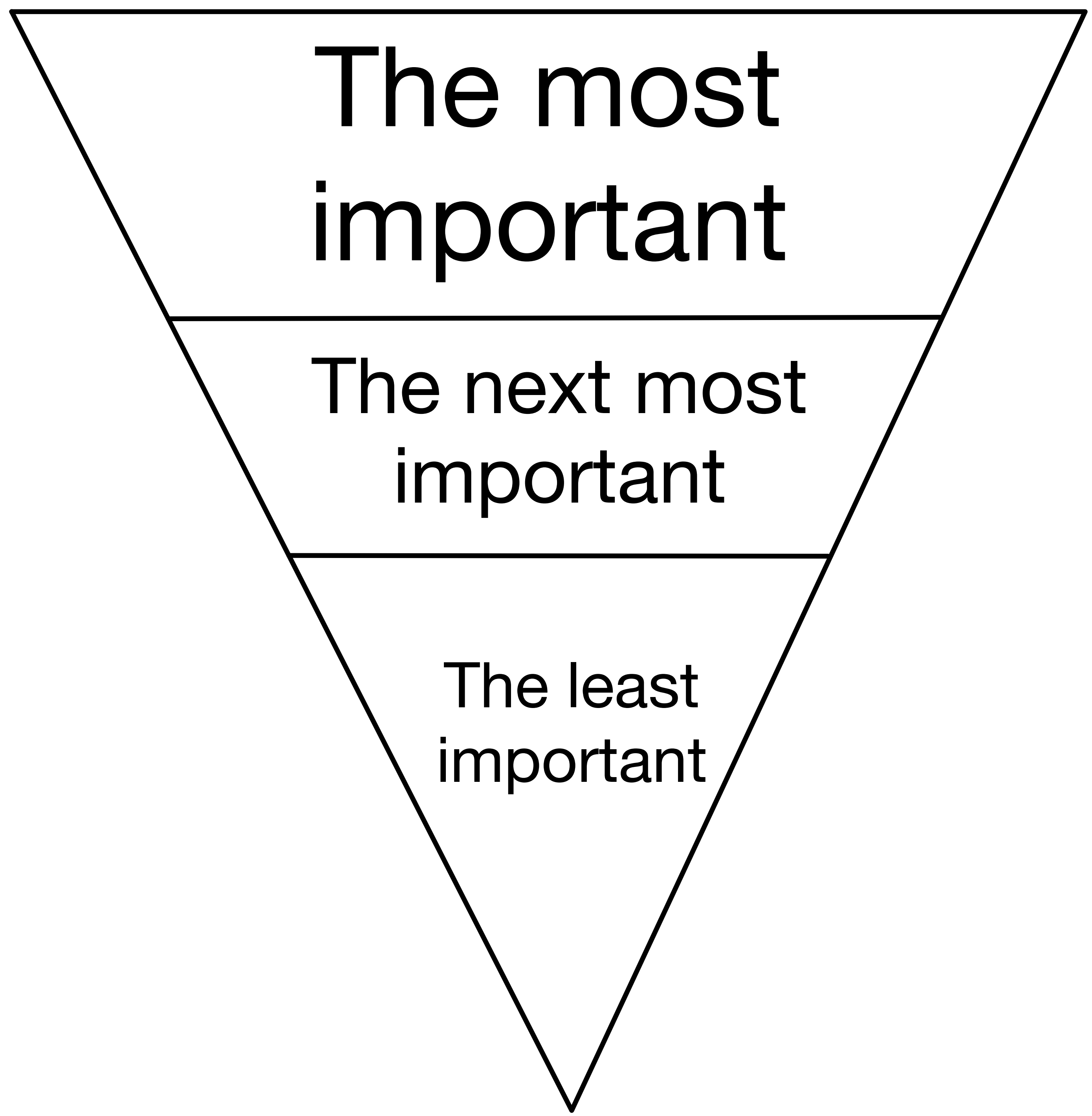
Washington (CNN) — He's churned up torrents of insults, incited grass-roots Republican fury, fearlessly flouted taboos on gender, race and religion and confounded the pundits again and again.

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 machine 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 reproduction of details
  - Repetitions
- **Augment** the **standard attention module**
- Introduce 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

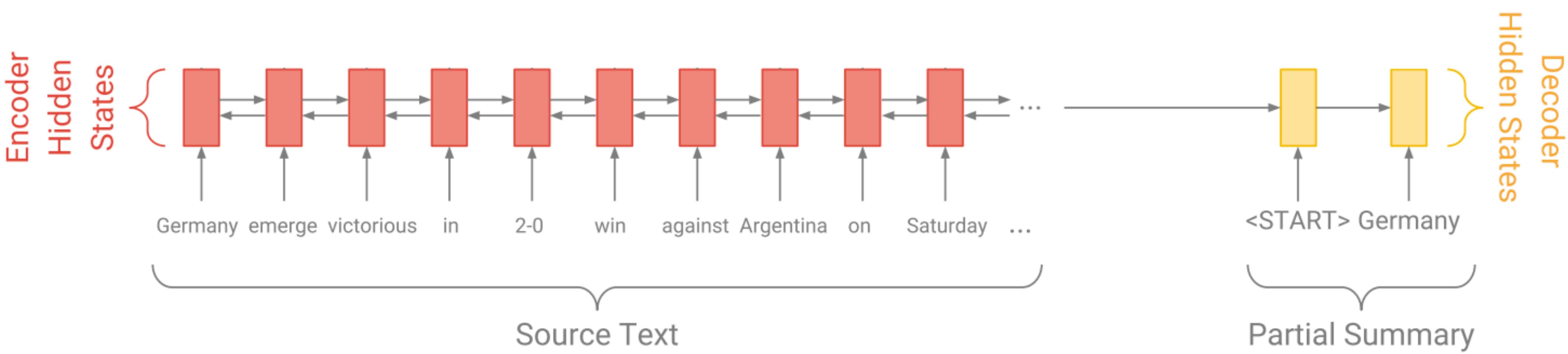
Germany emerge victorious in 2-0 win against Argentina on Saturday ...

Source Text

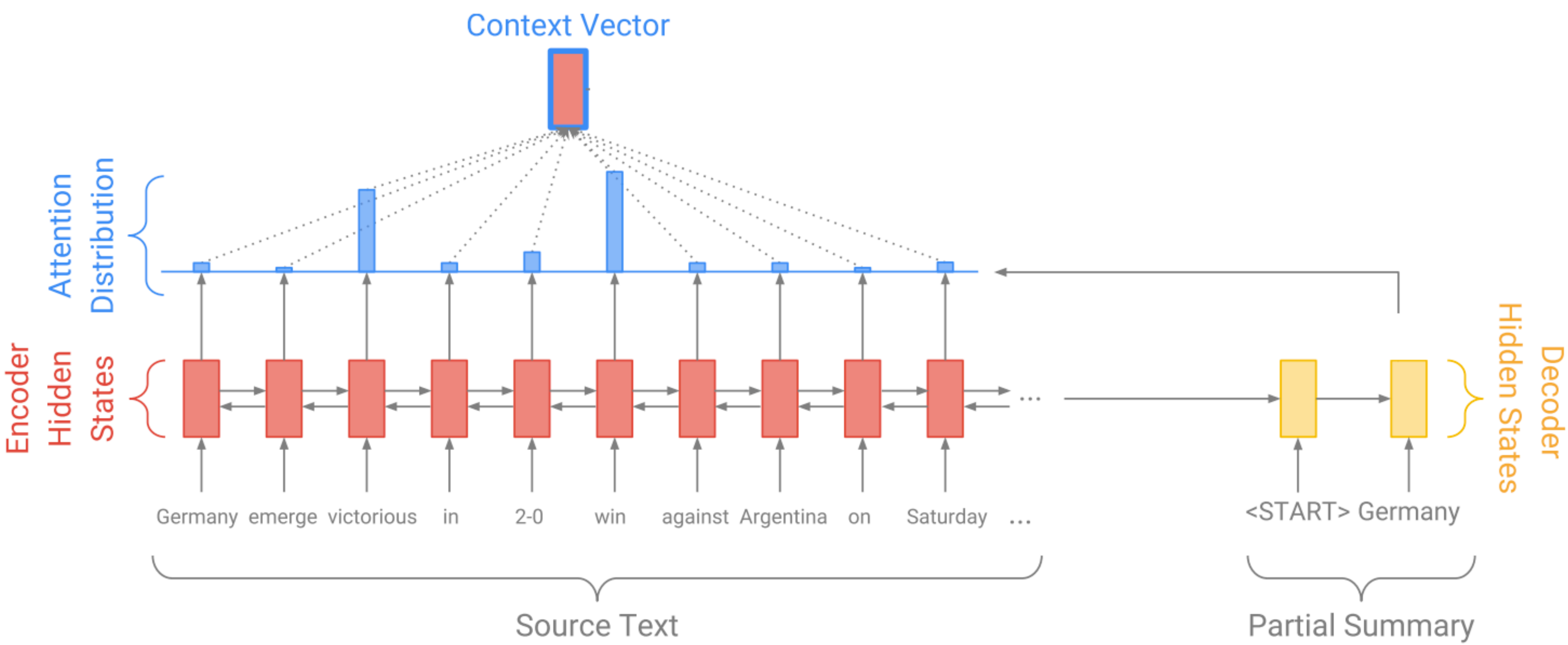
<START> Germany

Partial Summary

Attention mechanism



Attention mechanism



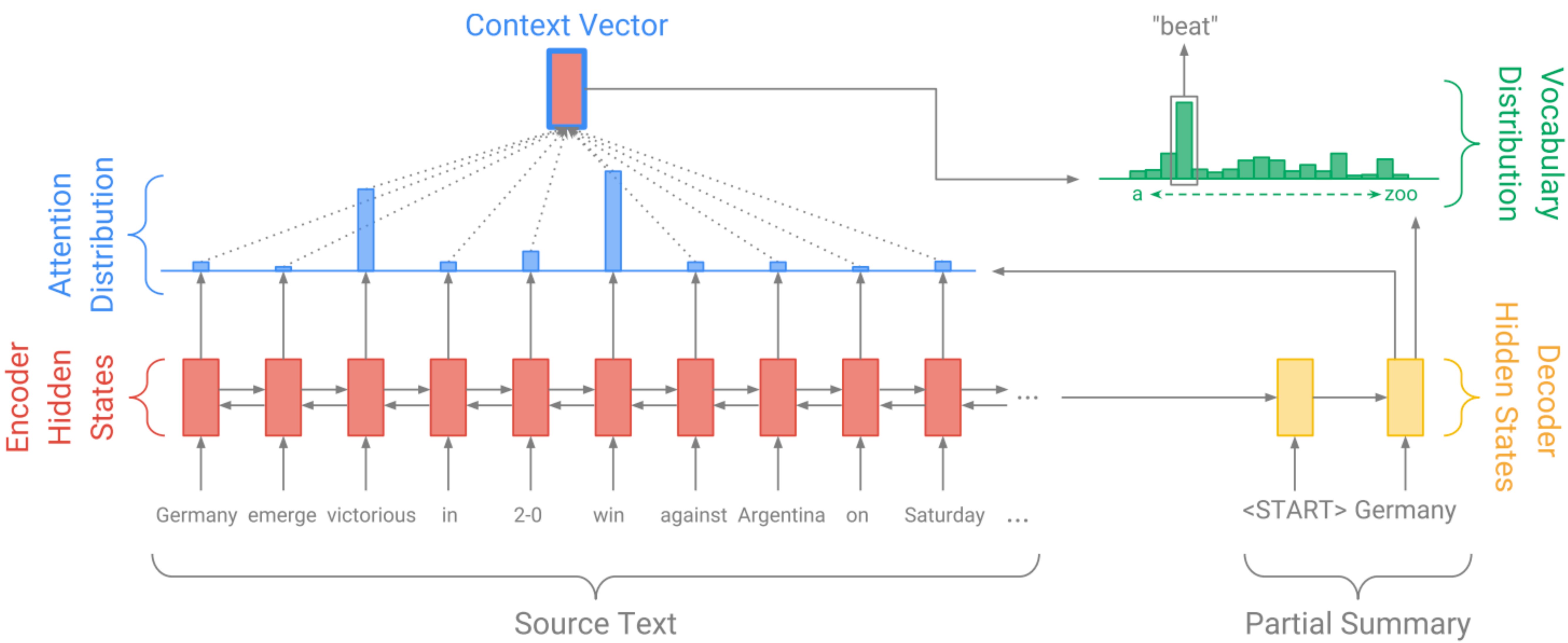
Context vector

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

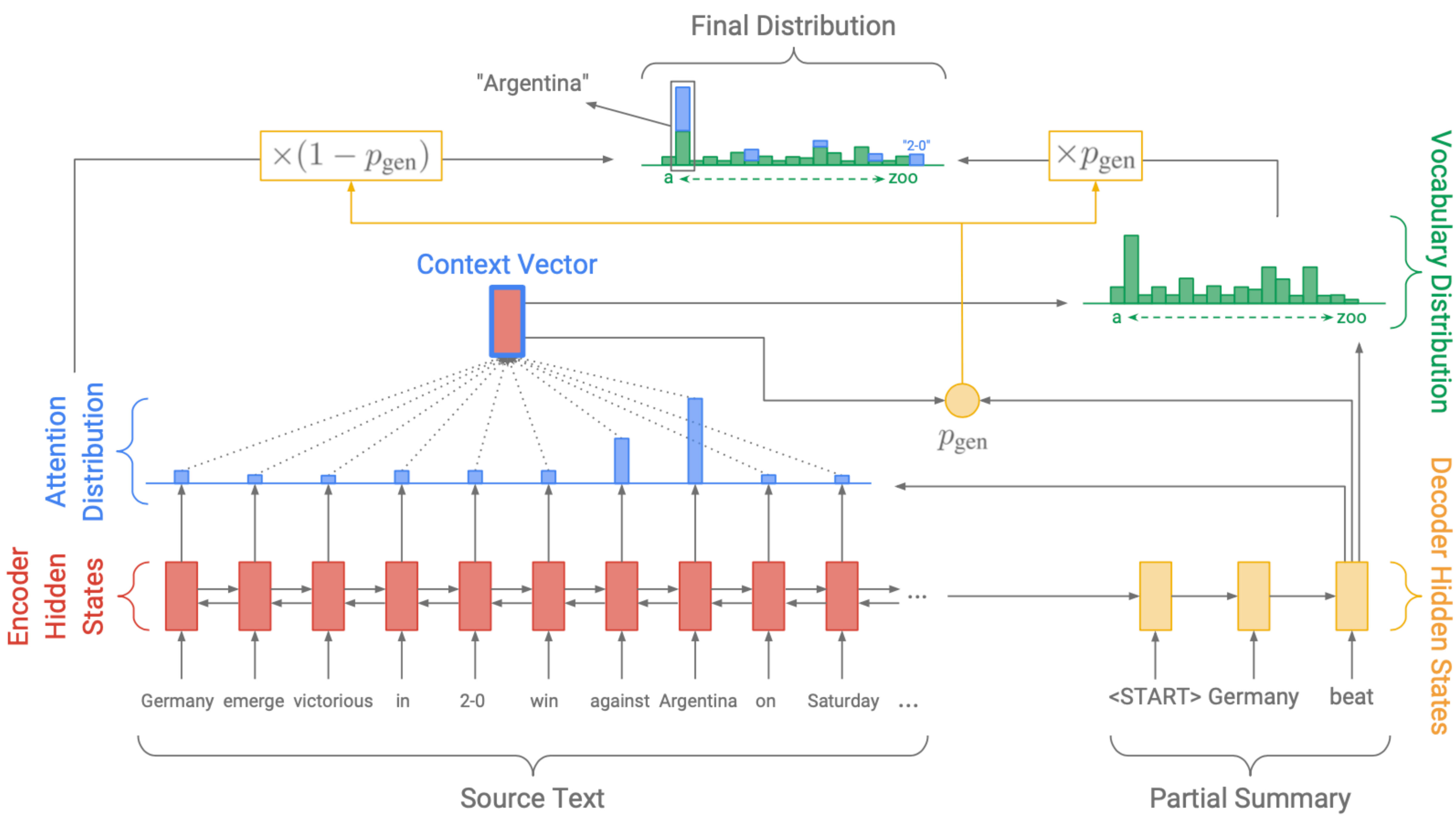
$$P_{\text{vocab}} = \text{softmax}(V'(V[s_t, h_t^*] + b) + b')$$

Context vector

Copy mechanism

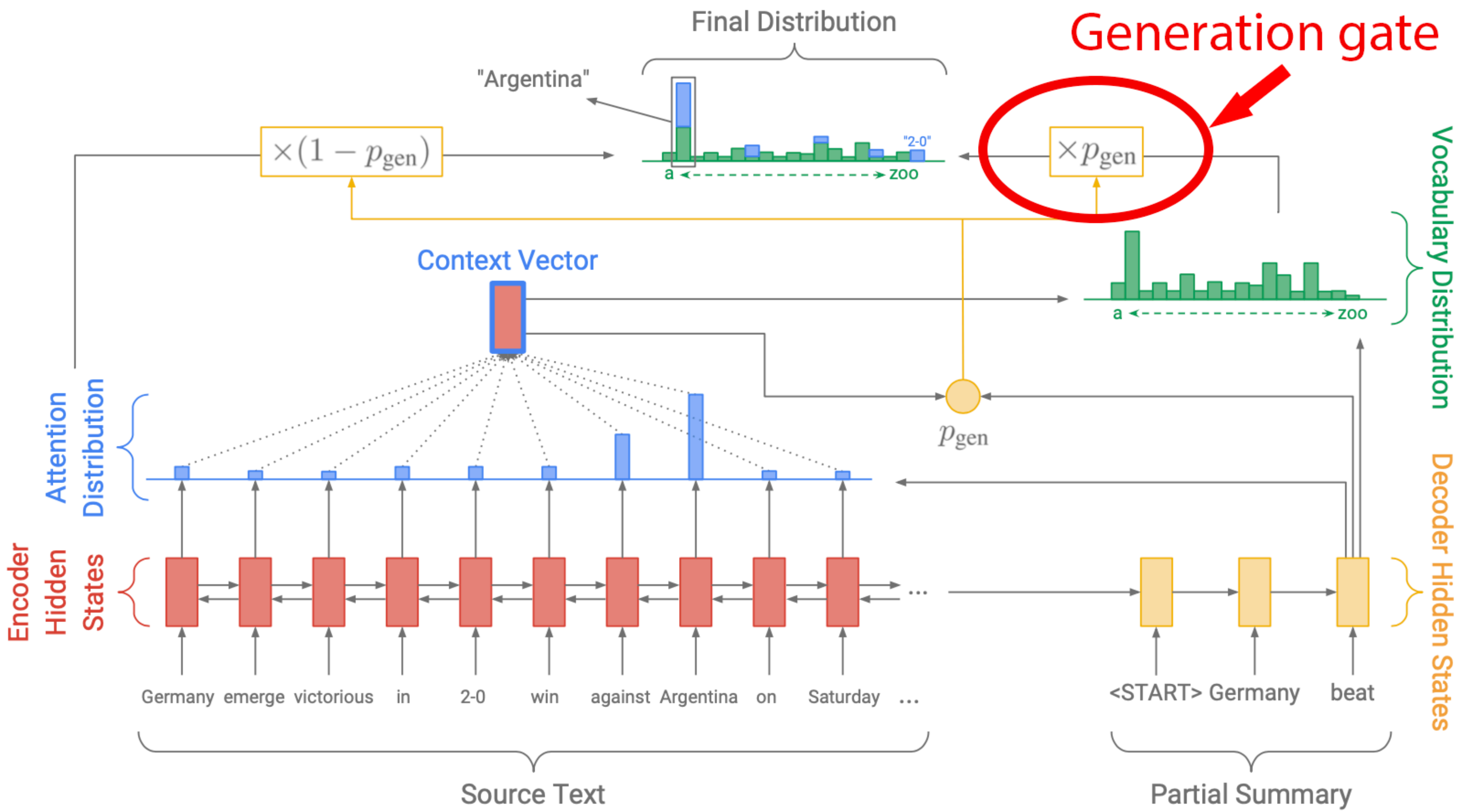
- Directly copies words from the source via a **pointer network** (Vinyals et al., 2015)
- Reuses attention weights
- Useful for the **OOV** words problem
- The final word distribution combines **generation** and **‘copy’** word distributions

Full model



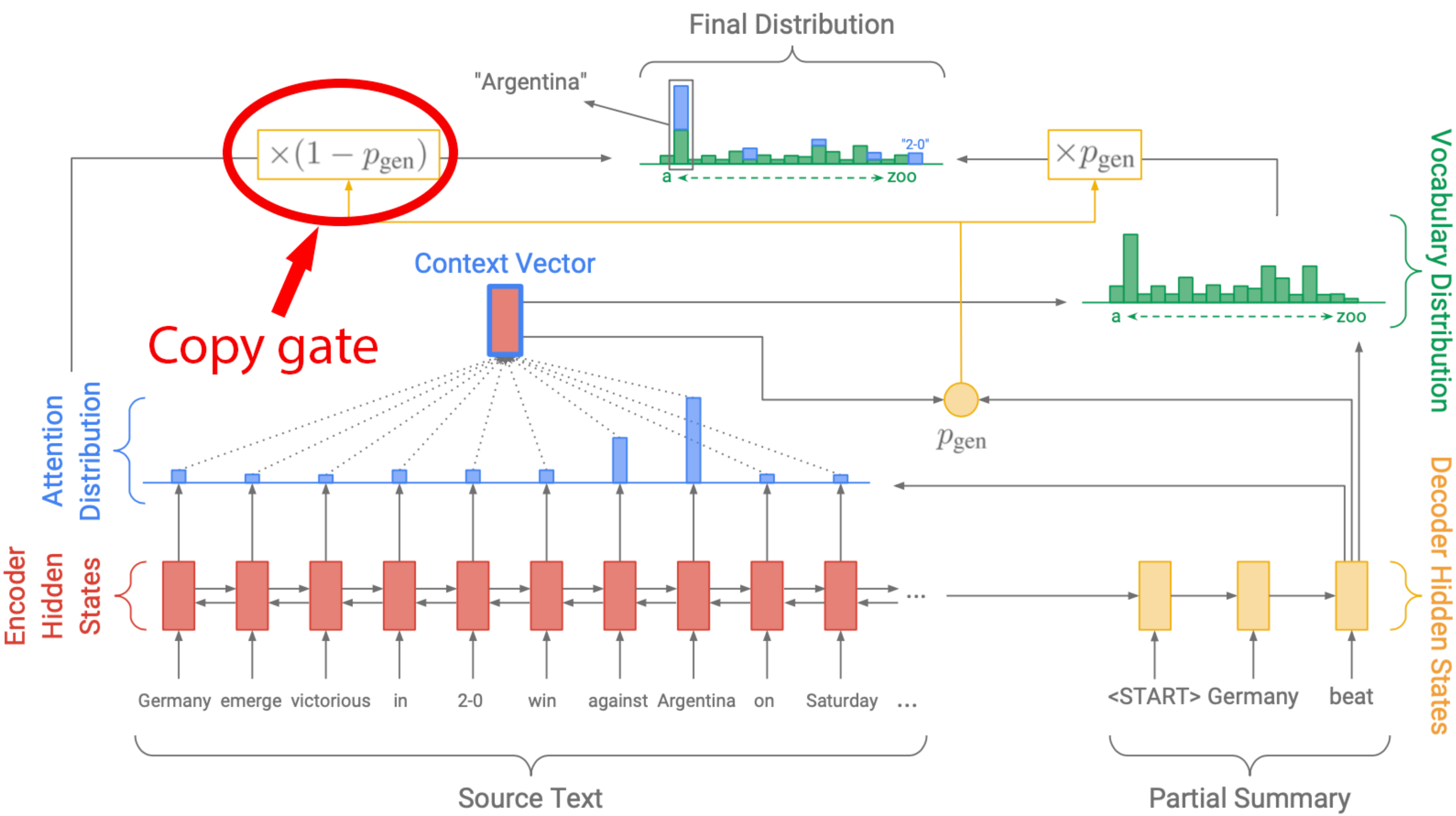
Final Distribution
"Argentina"
"2-0"
a
zoo
$\times(1 - p_{gen})$
$\times p_{gen}$
Context Vector
Vocabulary Distribution
a
zoo
Attention Distribution
p_{gen}
Decoder Hidden States
Encoder Hidden States
Germany emerge victorious in 2-0 win against Argentina on Saturday ...
<START> Germany beat
Source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

Full model



Final Distribution
Generation gate
"Argentina"
$\times(1-p_{\text{gen}})$
"2-0"
$\times p_{\text{gen}}$
a
zoo
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"
$\times(1-p_{\text{gen}})$
"2-0"
a
zoo
$\times p_{\text{gen}}$
Vocabulary Distribution
Context Vector
Copy gate
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

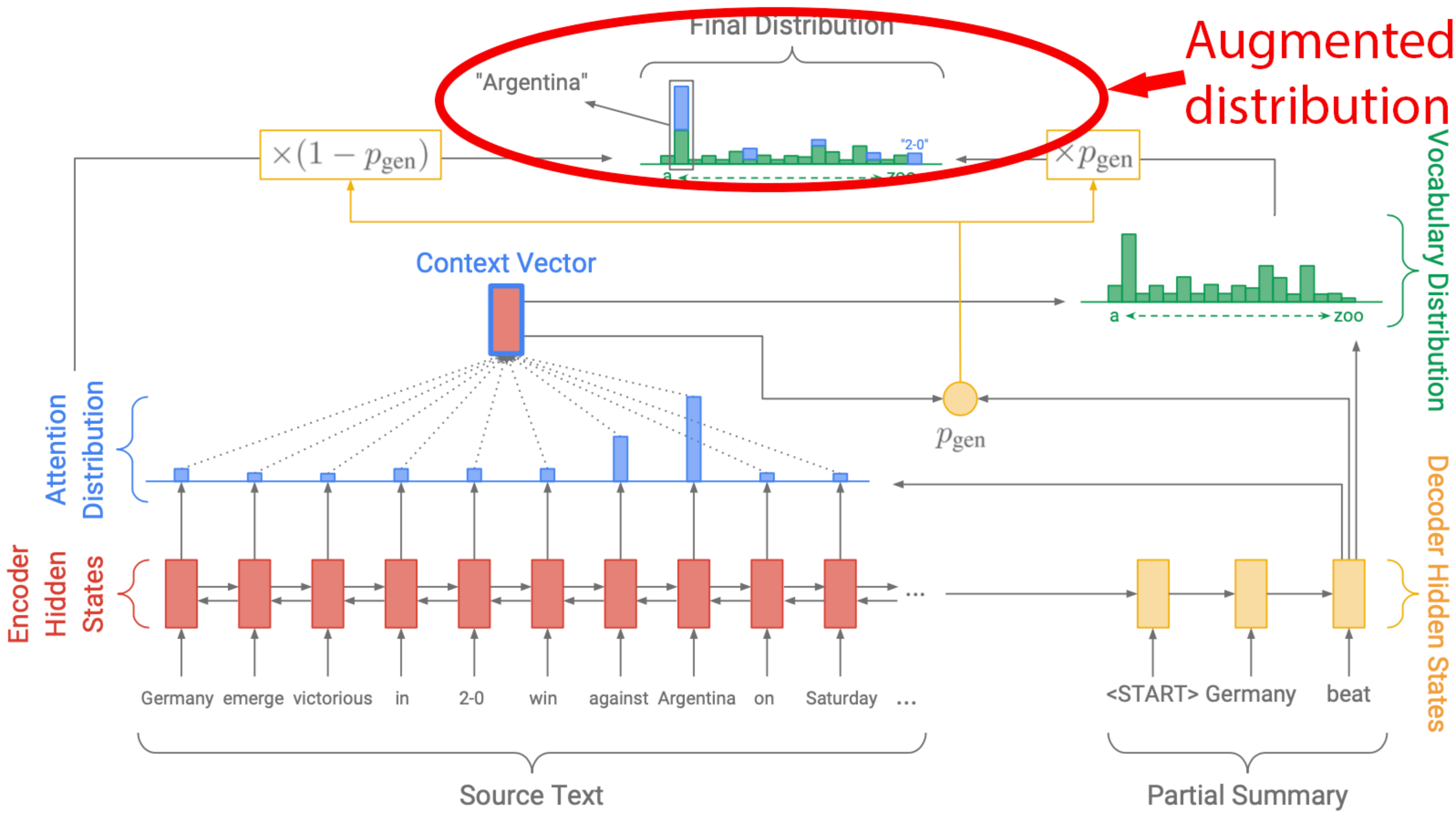
Bias

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

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

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

Final distribution

Generation distribution

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

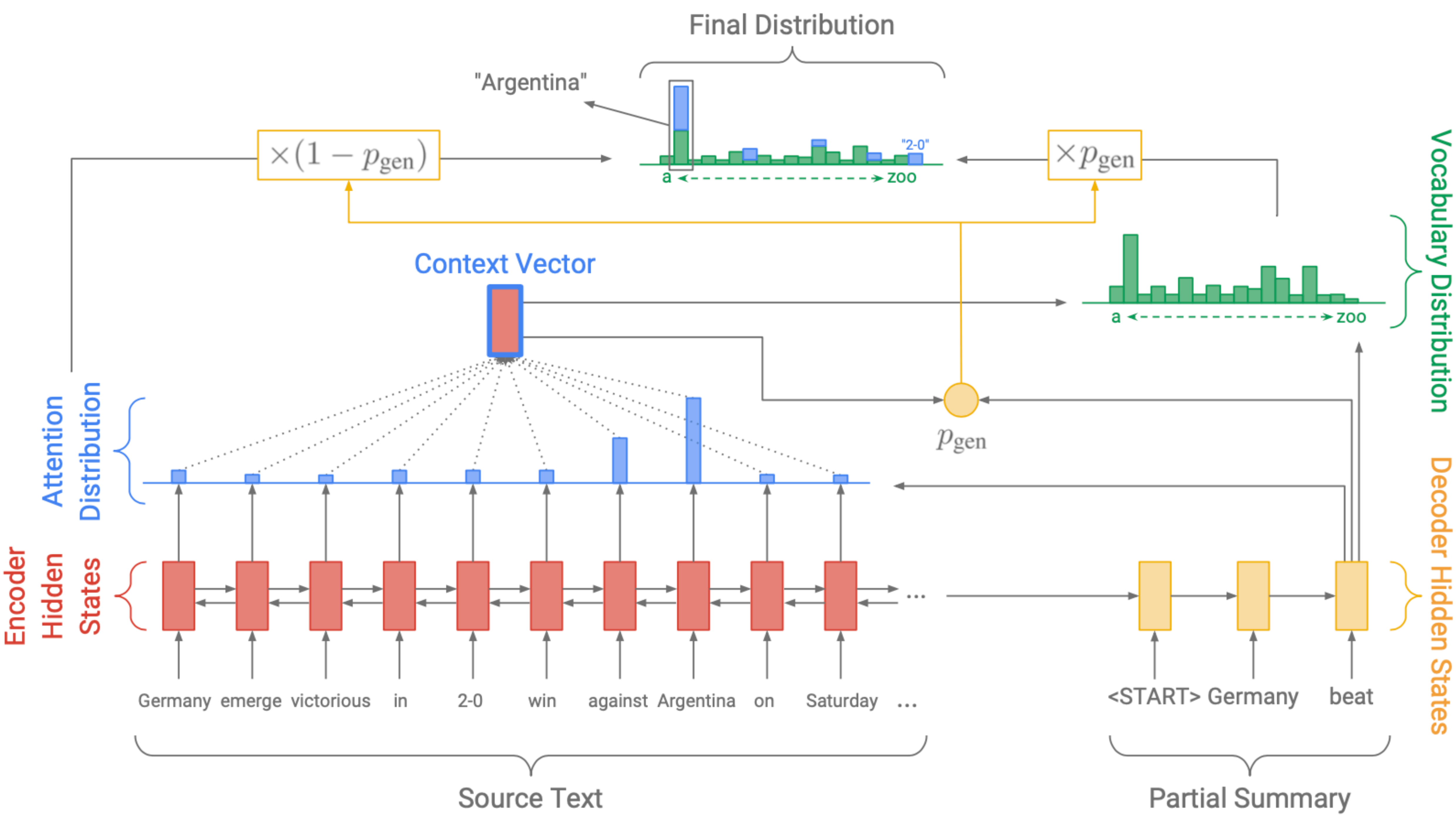
Final distribution

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"
$\times(1-p_{\text{gen}})$
$\times p_{\text{gen}}$
a
zoo
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

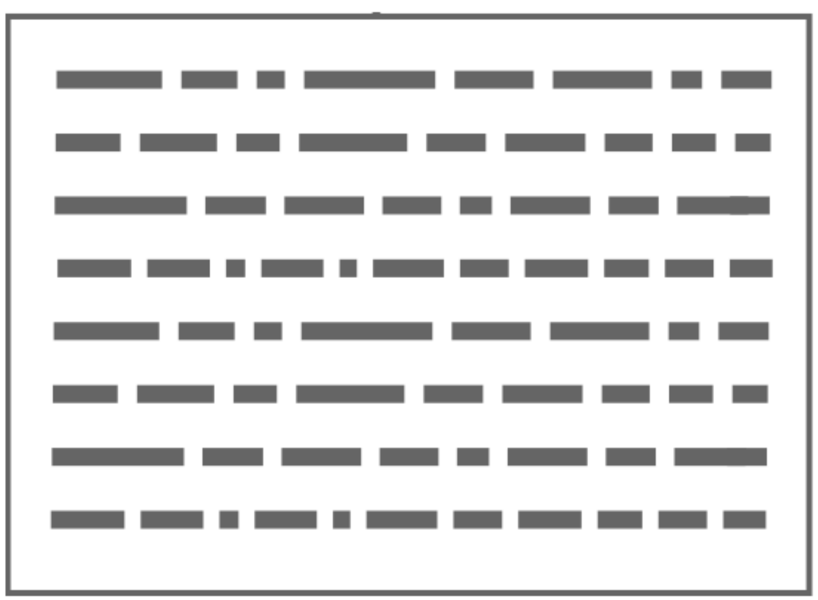
BottomUP

- Builds on top of the PGN model
- Address the problem of **poor content selection**
- 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

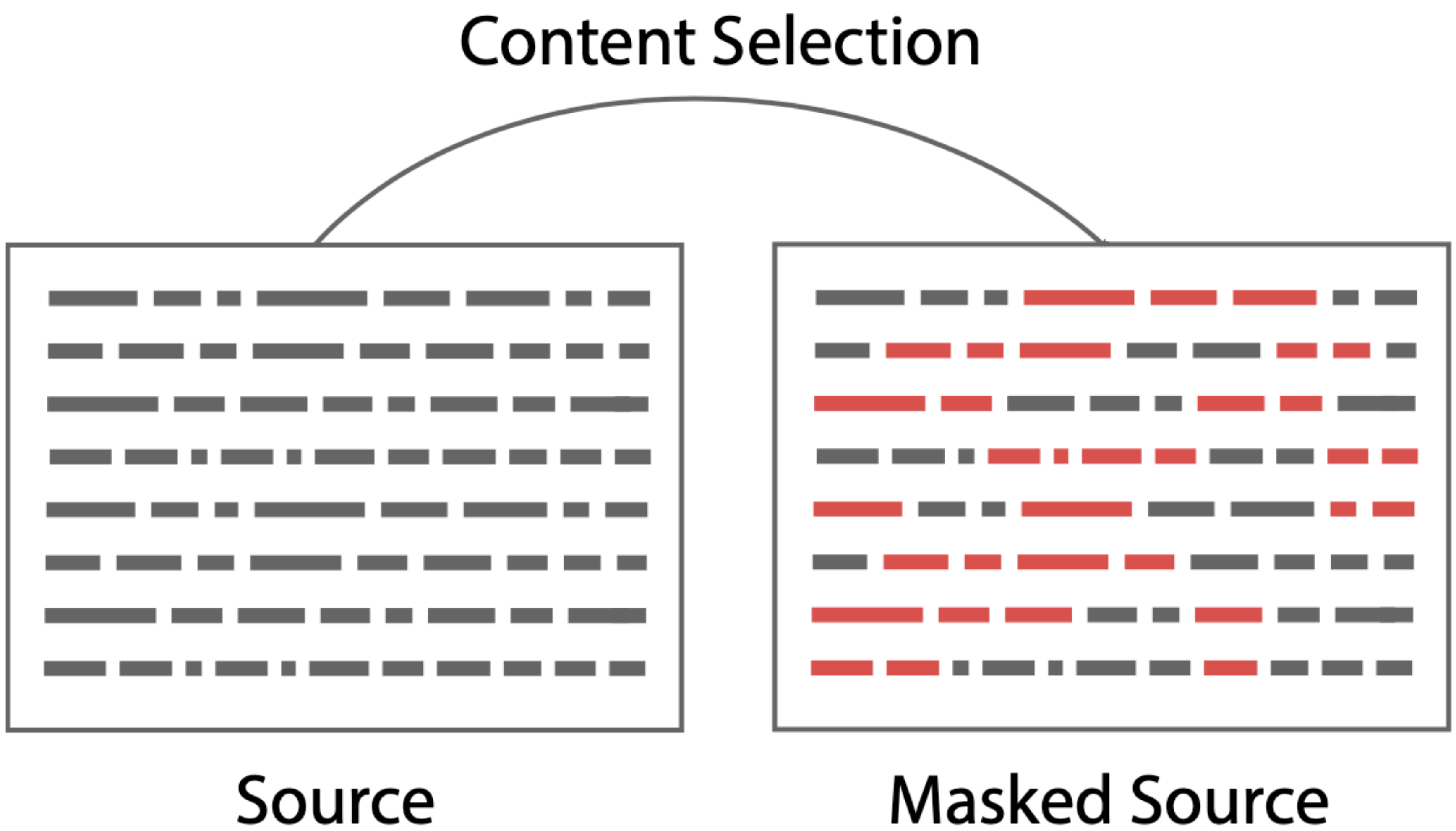
- **Content selector:**
  - GloVe (Pennington et al., 2014)
  - ELMo (character-aware token embeddings + bi-LSTM layers) (Peters et al., 2018)
  - bi-LSTM
  - Linear projection + sigmoid
- **Main model:**
  - Pointer-generator network (See et al., 2018)

Two-step procedure

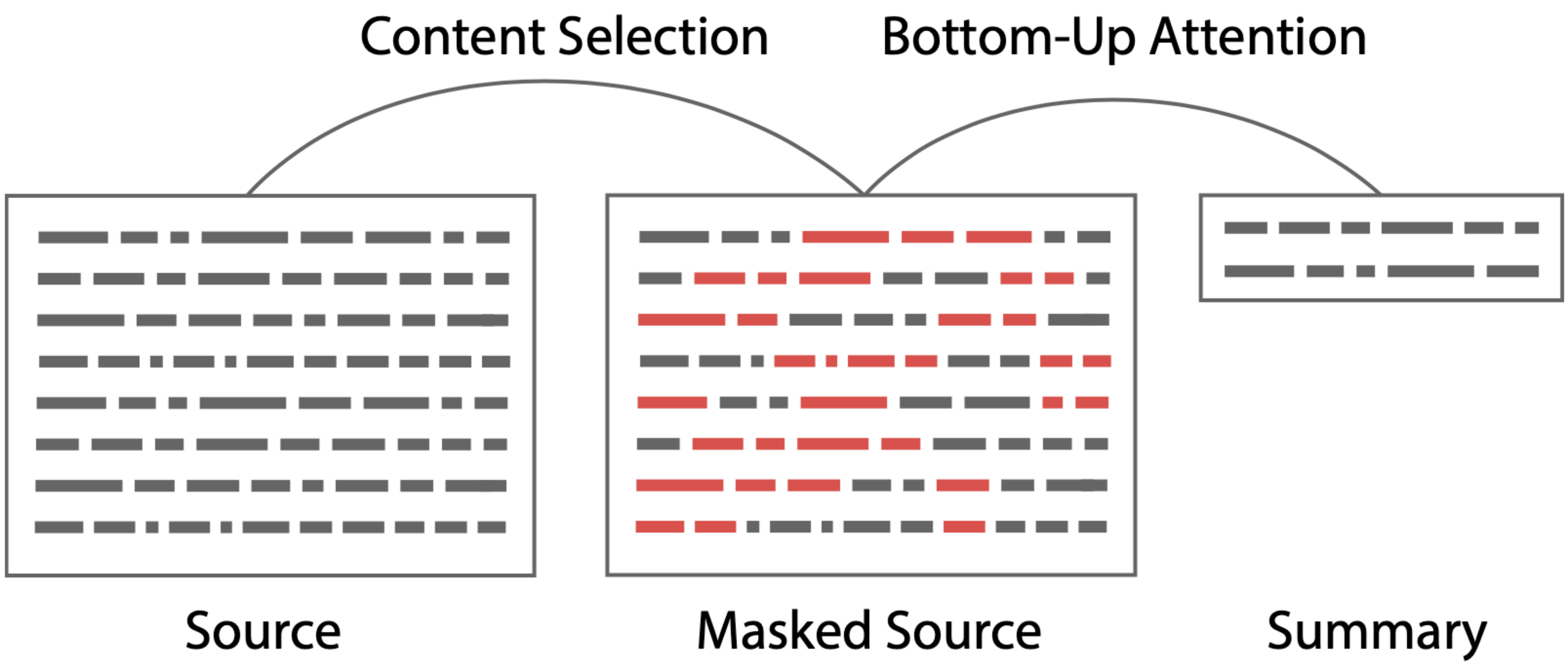


Source

Two-step procedure



Two-step procedure



Augmented copy distribution

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

Augmented copy distribution

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

source words

Augmented copy distribution

current prefix words

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

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

current prefix words

selector probability

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

source words

attention probability

Augmented copy distribution

current prefix words

selector probability

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

source words

attention probability

threshold

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 (conditional) 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**

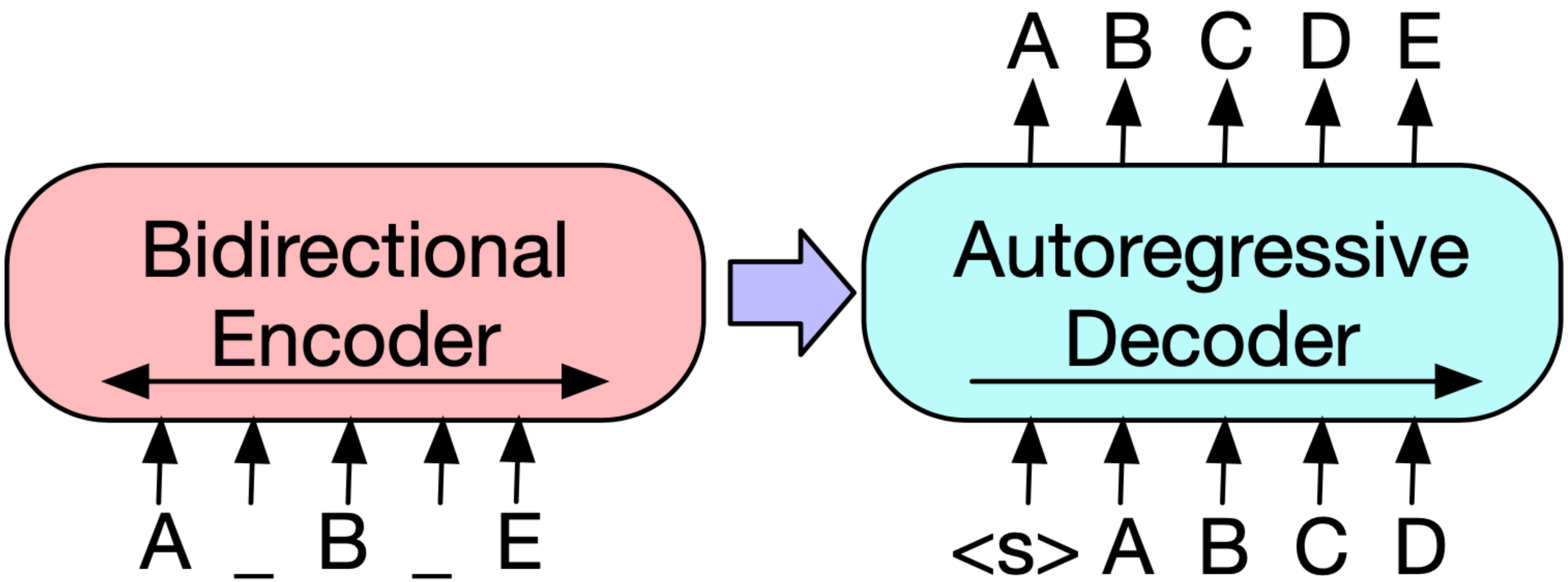
Pre-trained decoder?

- 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
Autoregressive Decoder
A _ B _ E
<s> A B C D

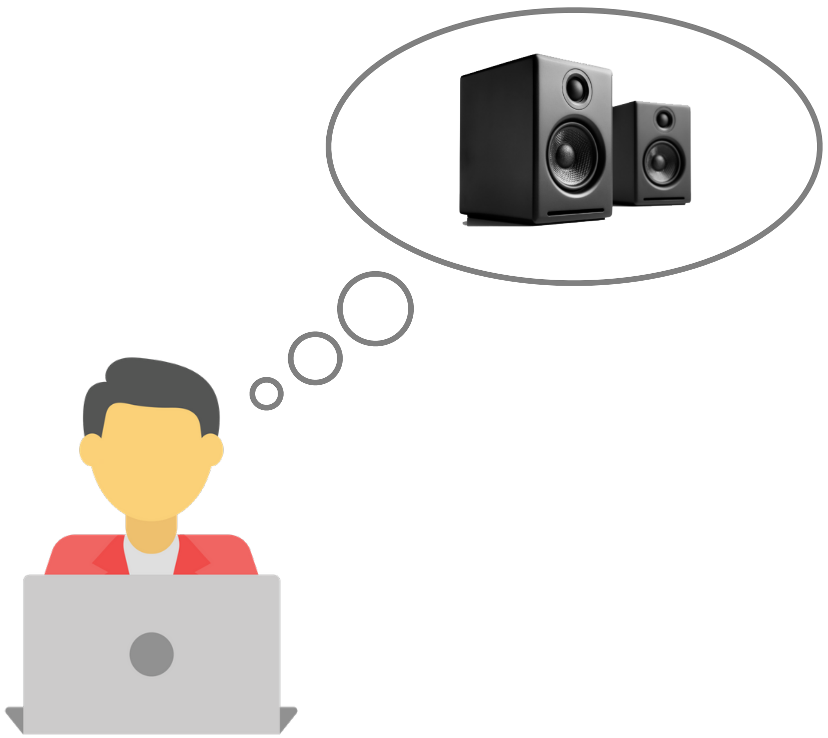
CNN/DM

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**

Opinion Summarization

James

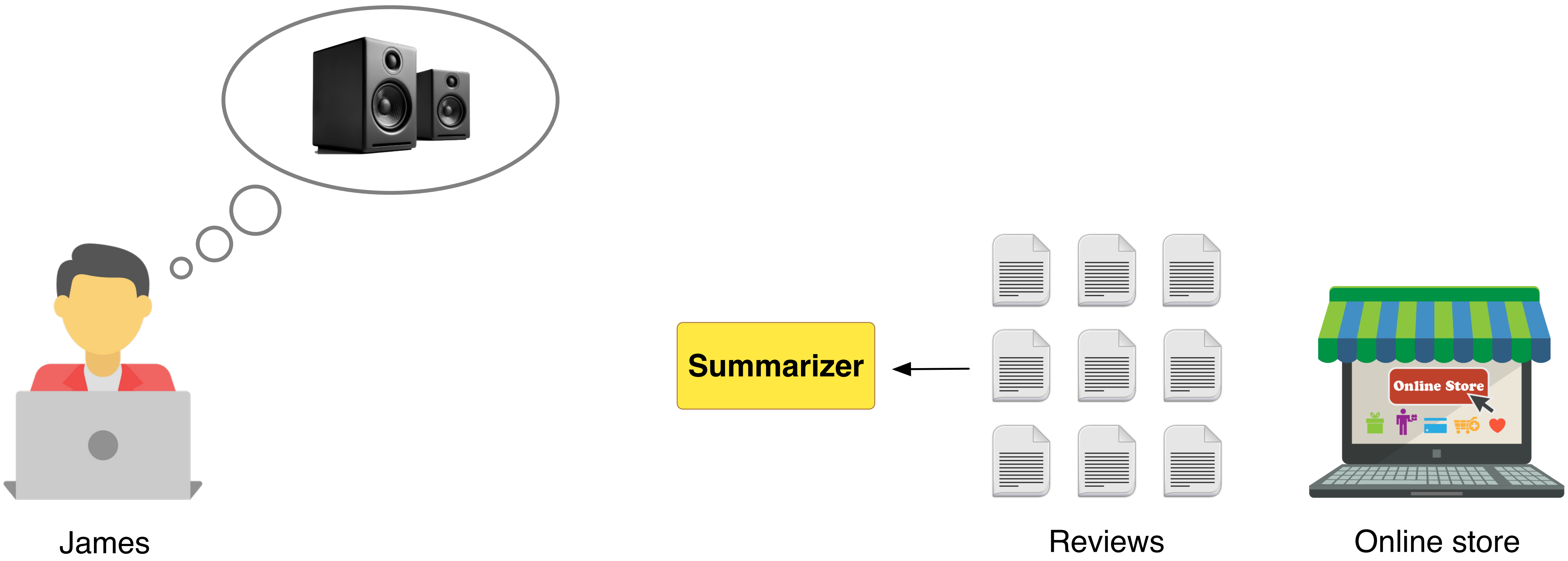

James

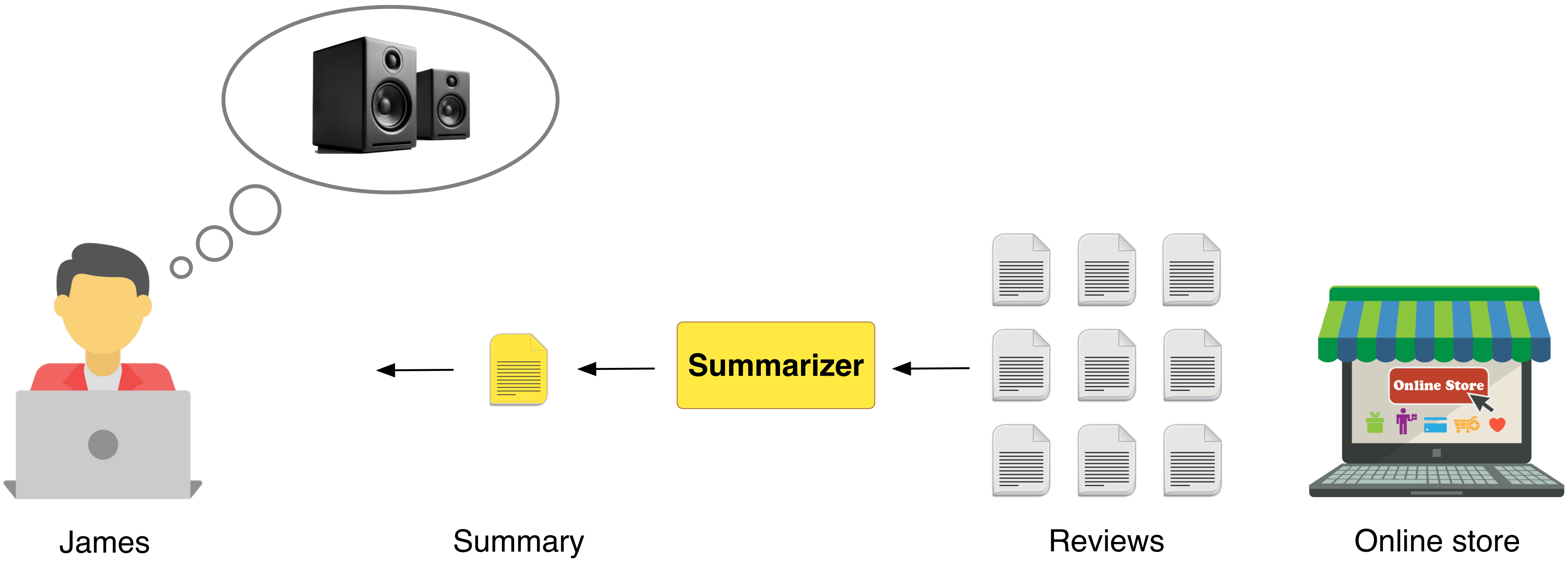


James

Online store


Online Store
James
Reviews
Online store


Summarizer
Online Store
James
Reviews
Online store


Online Store
Summarizer
James
Summary
Reviews
Online store

Extractive summarizers

- Are commonly used for the task (Ganesa et. al, 2010; Angelidis and Lapata, 2018; Isonuma et al., 2019)
- Mostly unsupervised or weakly-supervised
- Select summarizing input fragments
- Concatenate to form a summary
- Can be **incoherent** and contained **unimportant details**

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.

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**
- The largest one:**100 pairs with summaries** (Chu and Liu, 2019)
- Large quantities of reviews without summaries (**millions**)

Opinion and news summarization

	News	**Opinion**
Setup	Single-document	Multi-document
Task	Objective facts	Subjective opinions
Annotated abstractive data	1M+ (Grusky et. al. 2018)	100 (Chu and Liu, 2019)

Opinion summarization (unannotated data)

amazon.com

233 million reviews

yelp

8 million reviews

Abstractive summarizers

- Next, we're going to take a look at 3 models for abstractive opinion summarization
  - **MeanSum** (Chu and Liu, 2019)
  - **Copycat** (Bražinskas et al., 2020)
  - **FewSum** (Bražinskas et al., 2020)
- Each alleviates **the annotated data scarcity** in its own way
- Generate **consensus summaries**

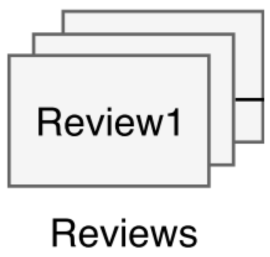
MeanSum: A Model for Unsupervised Neural Multi-Document Abstractive Summarization

Eric Chu, Peter Liu

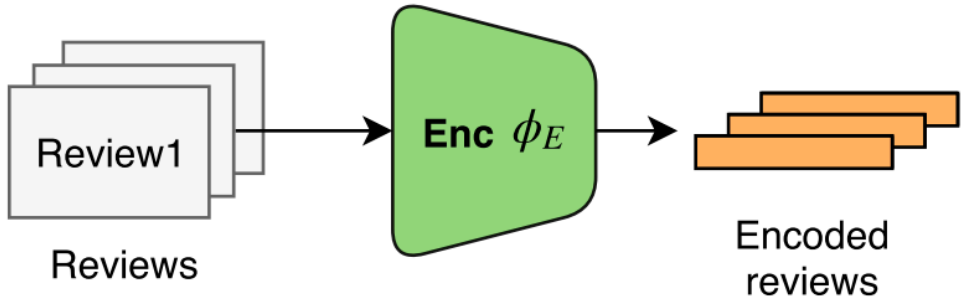
MeanSum

- Recent **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

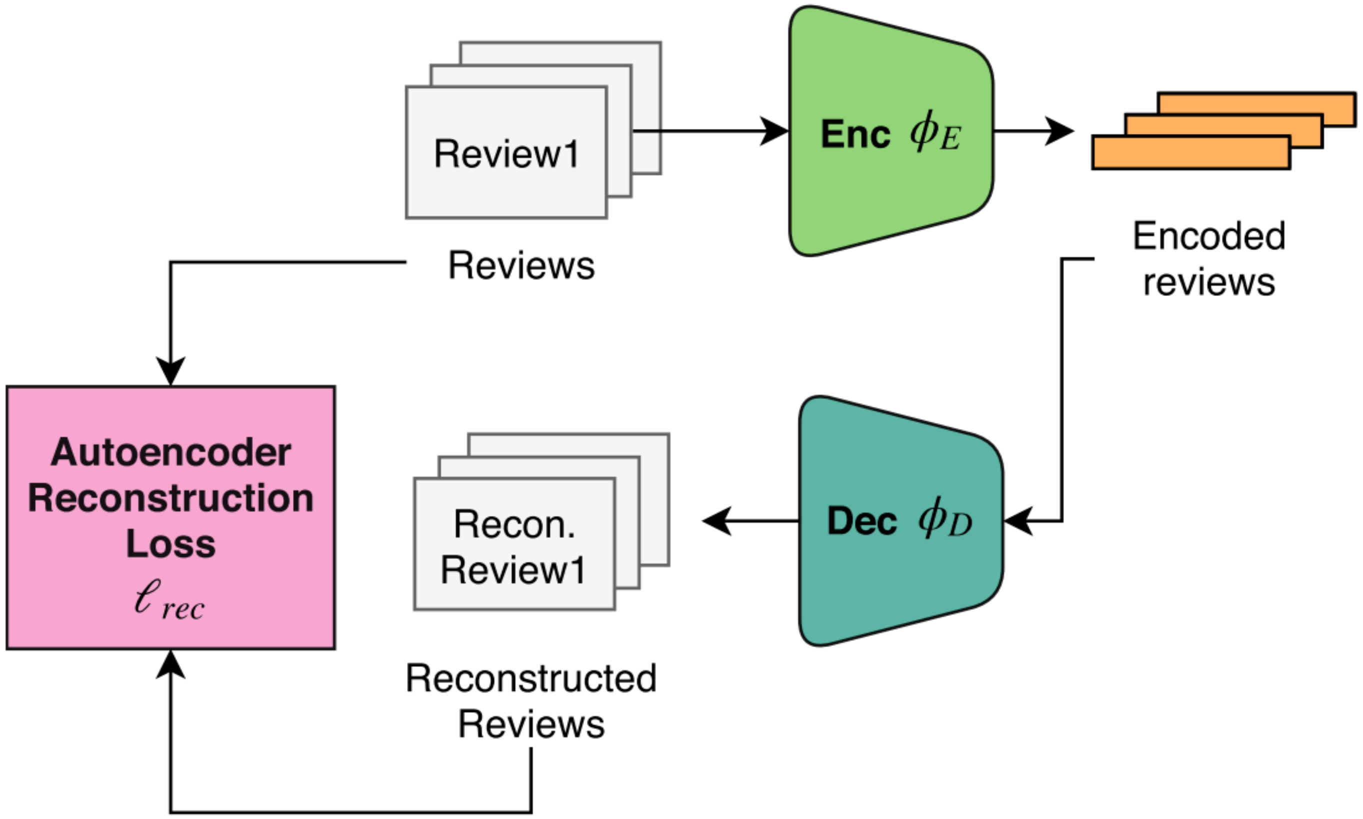
MeanSum



MeanSum



MeanSum



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

Reconstruction loss

ϕ_E - encoder $\quad x_i$ - review document

ϕ_D - decoder

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

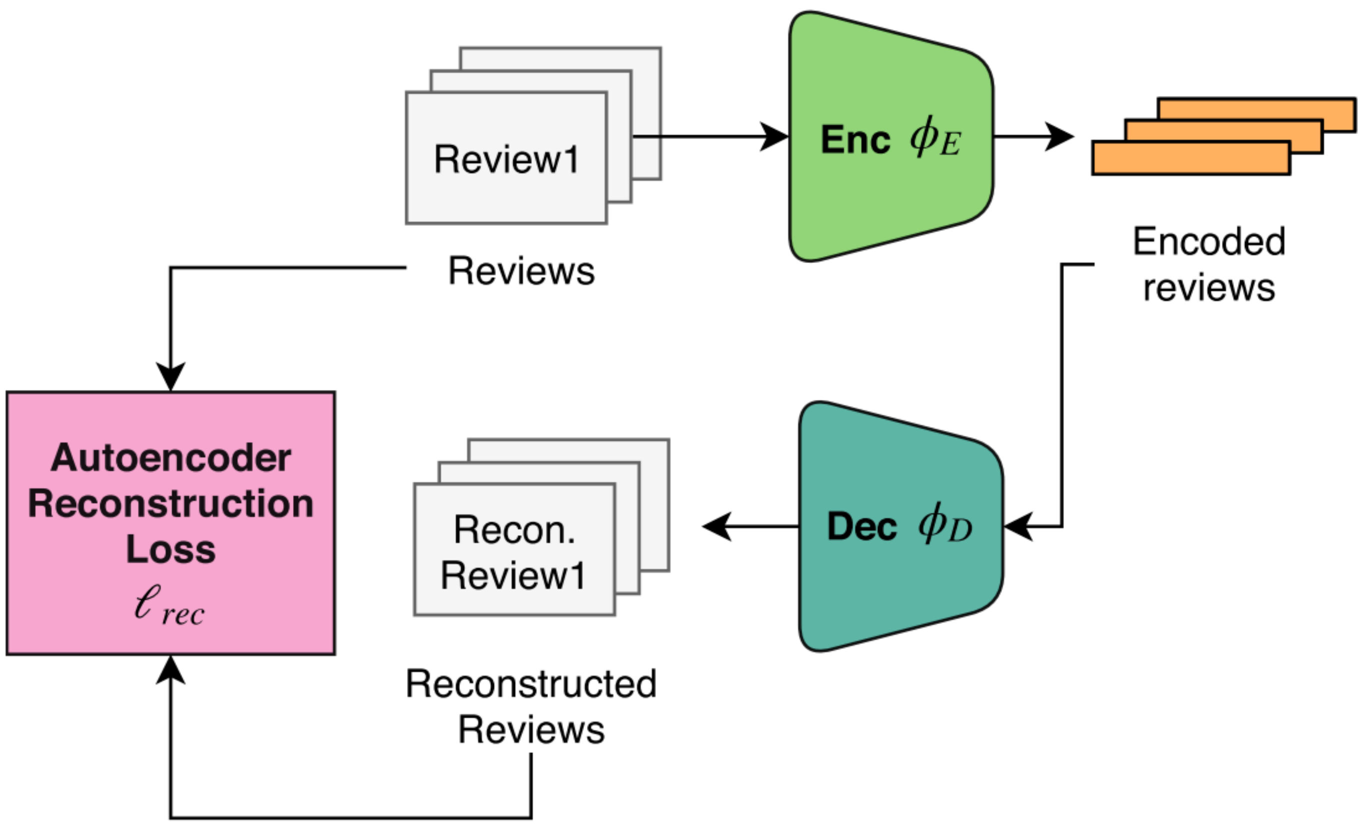
Reconstruction loss

ϕ_E - encoder $\quad x_i$ - review document

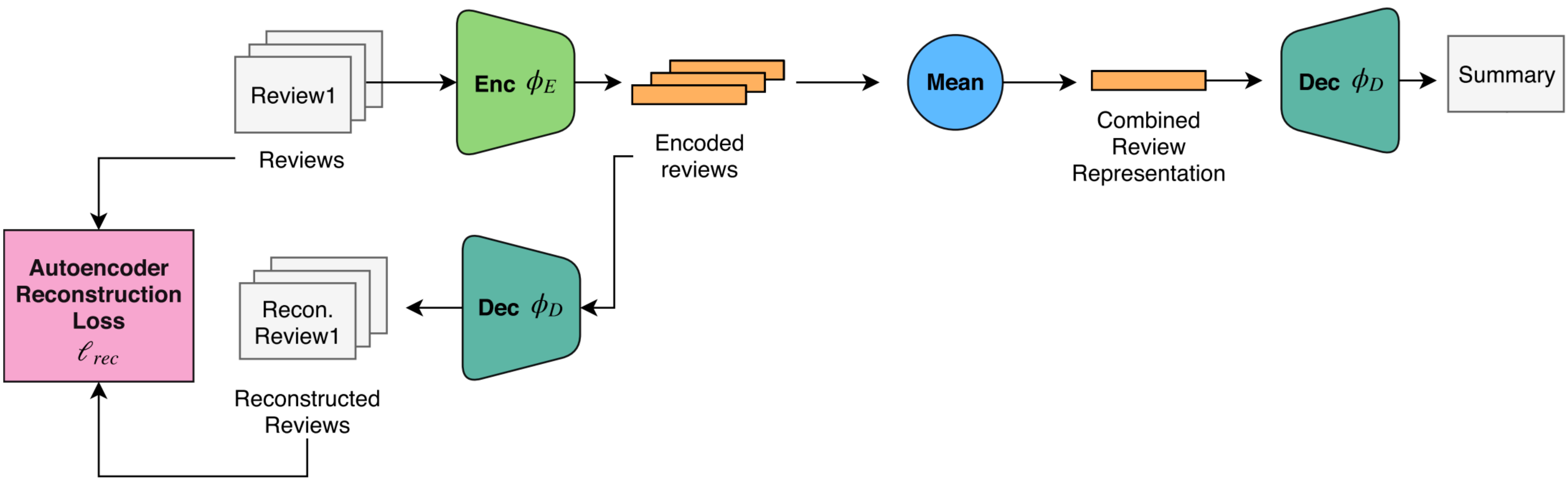
ϕ_D - decoder (**use Teacher Forcing**)

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

MeanSum


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

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

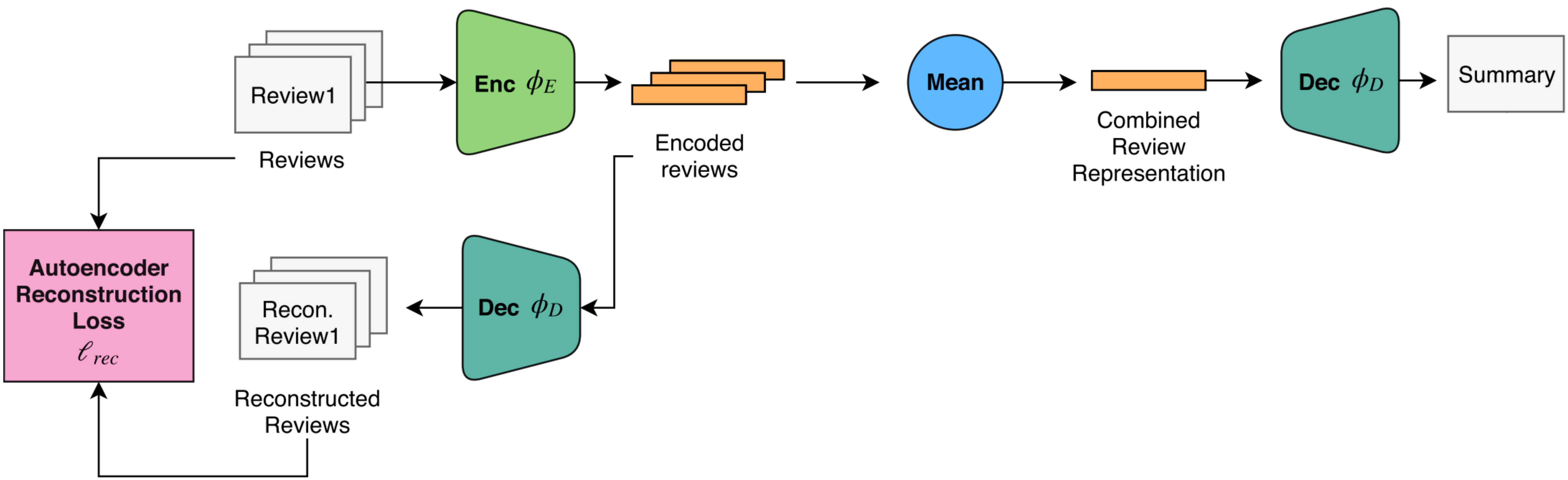
Summary sampling

- 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)

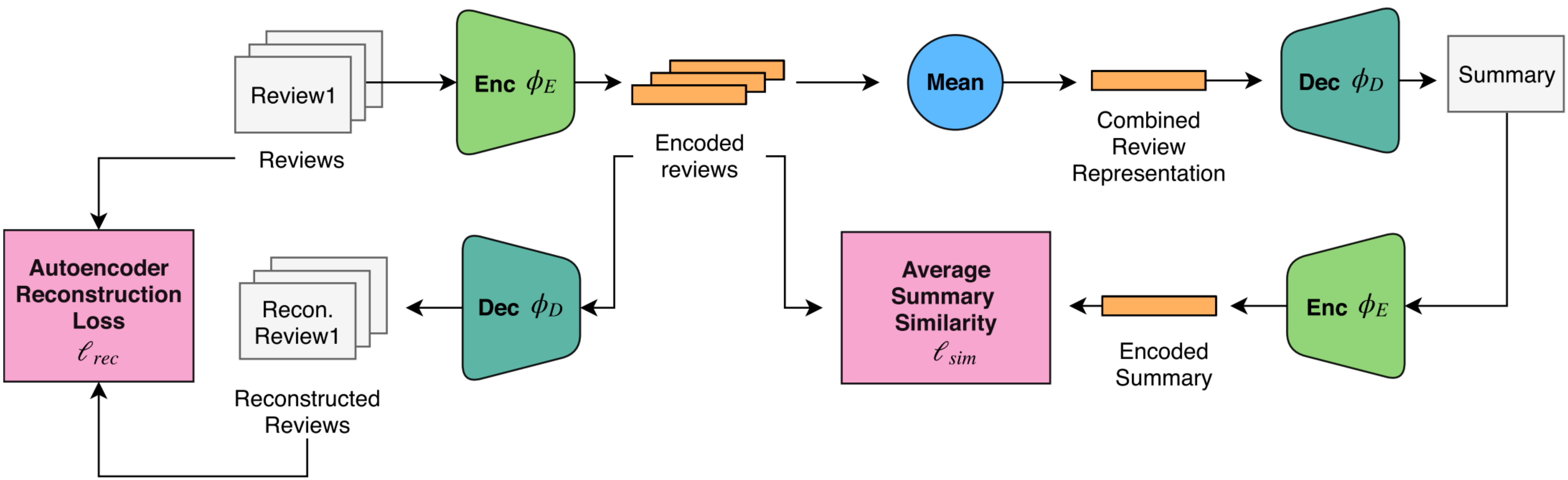
Semantic similarity loss

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

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

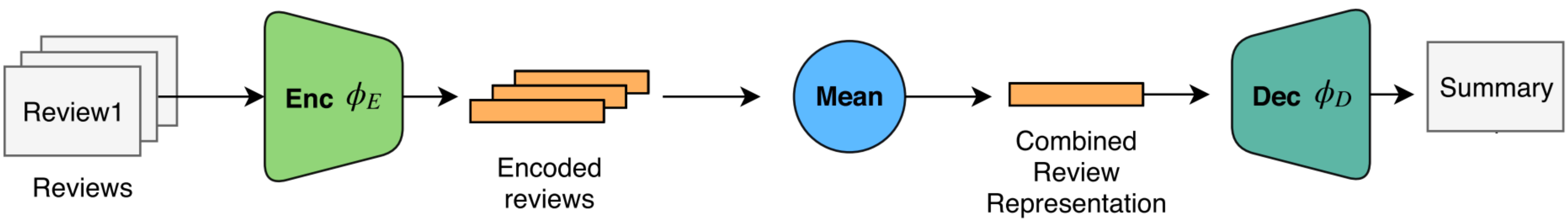
Results on Amazon

	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

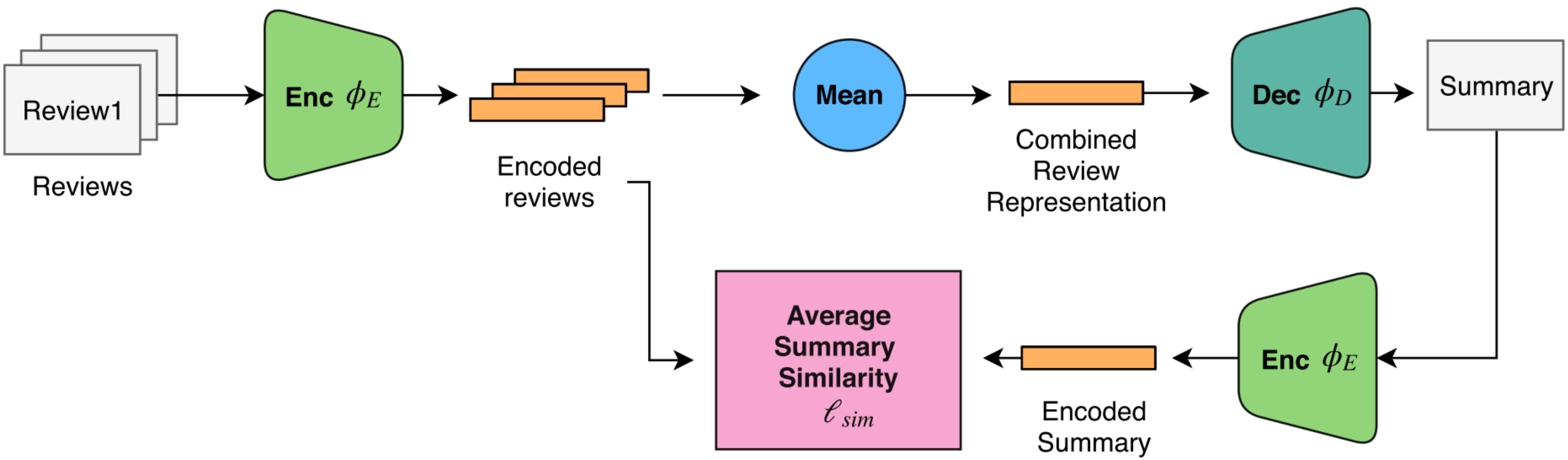
problem: writing style

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
- **Cons:**
  - Generates summaries that look like reviews
    - Informal writing style
    - Unimportant details
  - Poor content support

Unsupervised Opinion Summarization as Copycat-Review Generation

Arthur Bražinskas, Mirella Lapata, Ivan Titov

ACL 2020

Approach

- Unsupervised latent model (continues 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

Great Italian restaurant with authentic food and great service! Recommend!

review 1

We ordered pasta, and it was very tasty. Would recommend this place to anyone.

review 2

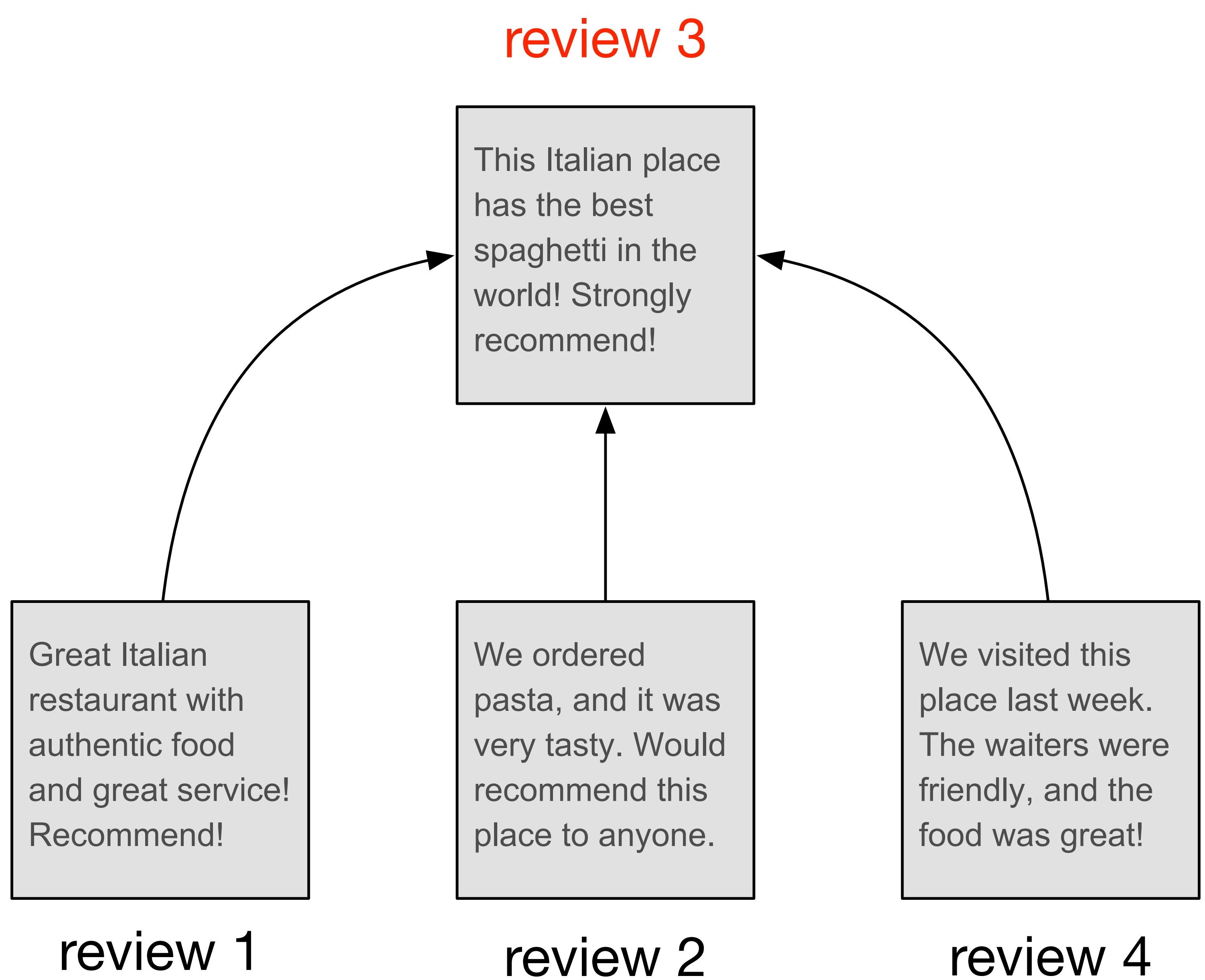
This Italian place has the best spaghetti in the world! Strongly recommend!

review 3

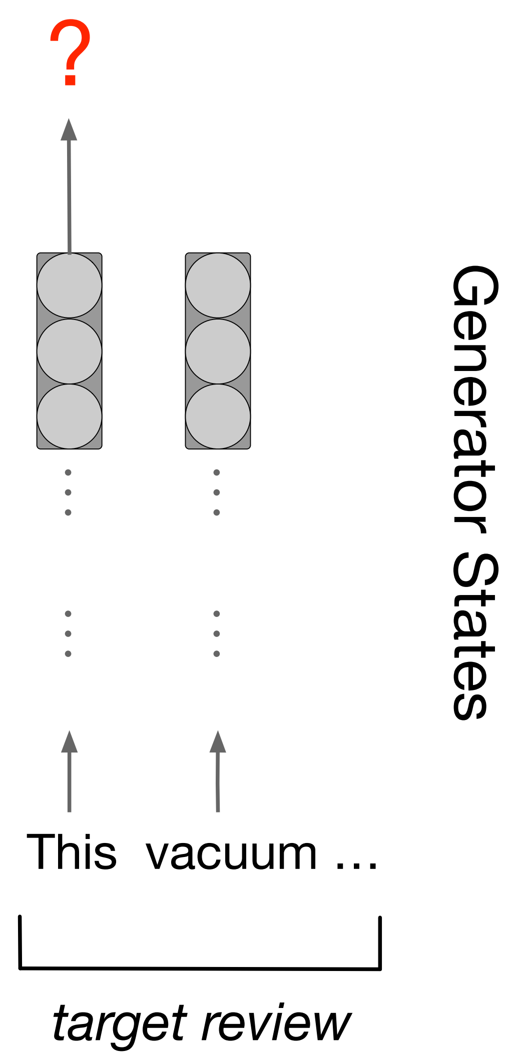
We visited this place last week. The waiters were friendly, and the food was great!

review 4

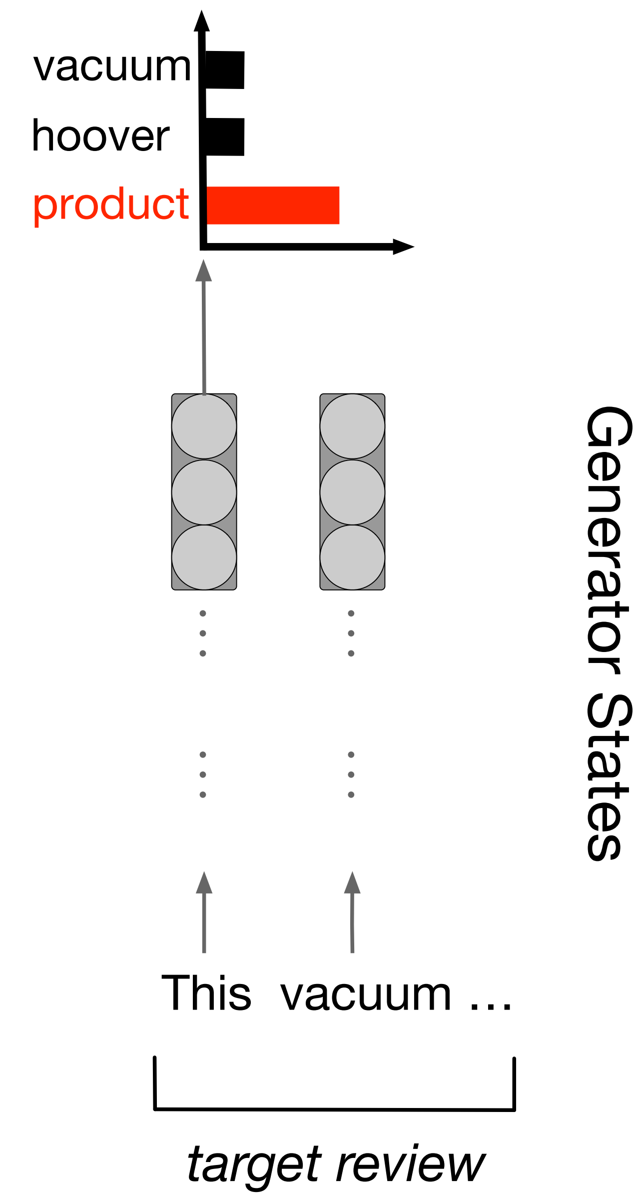
Leave-one-out



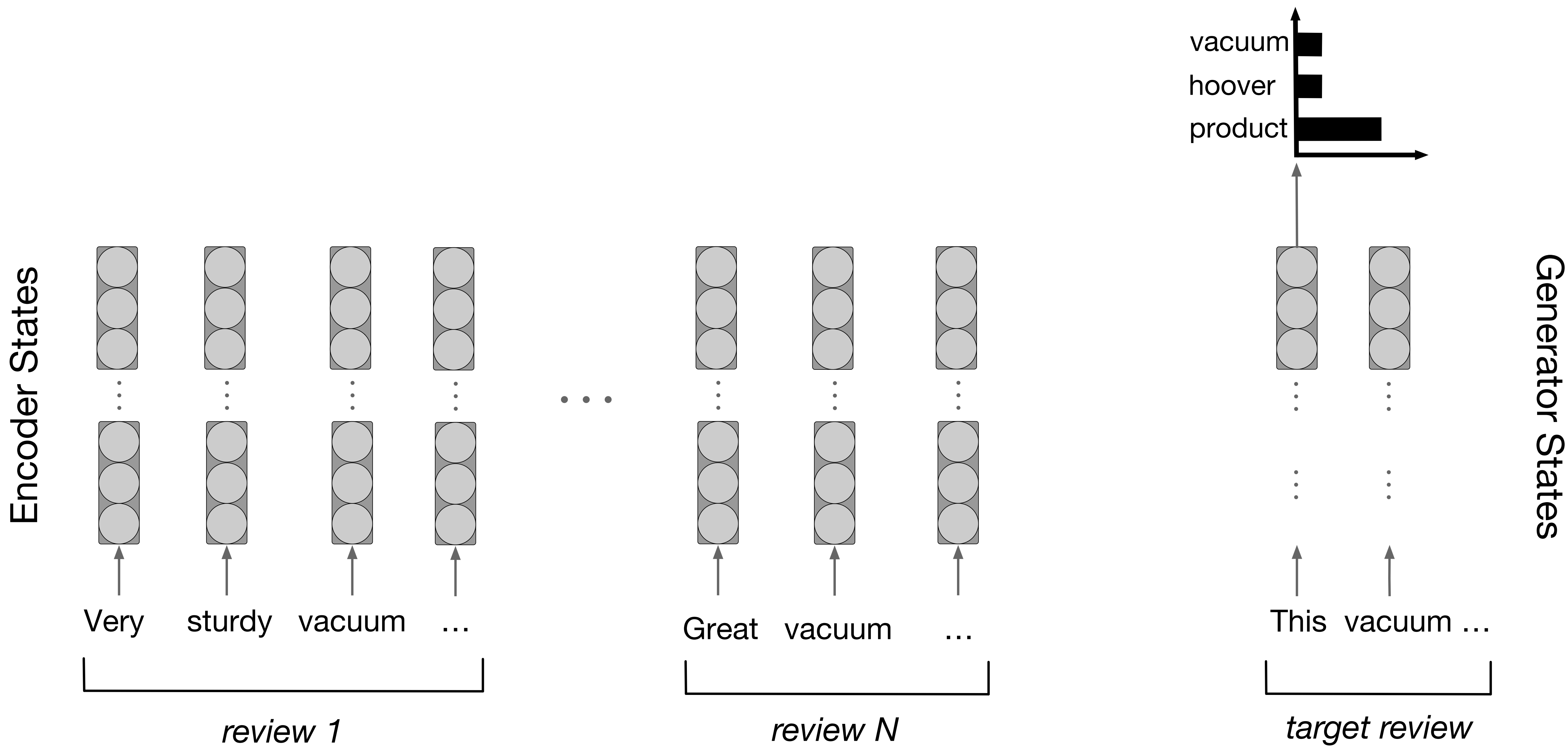
Leave-one-out



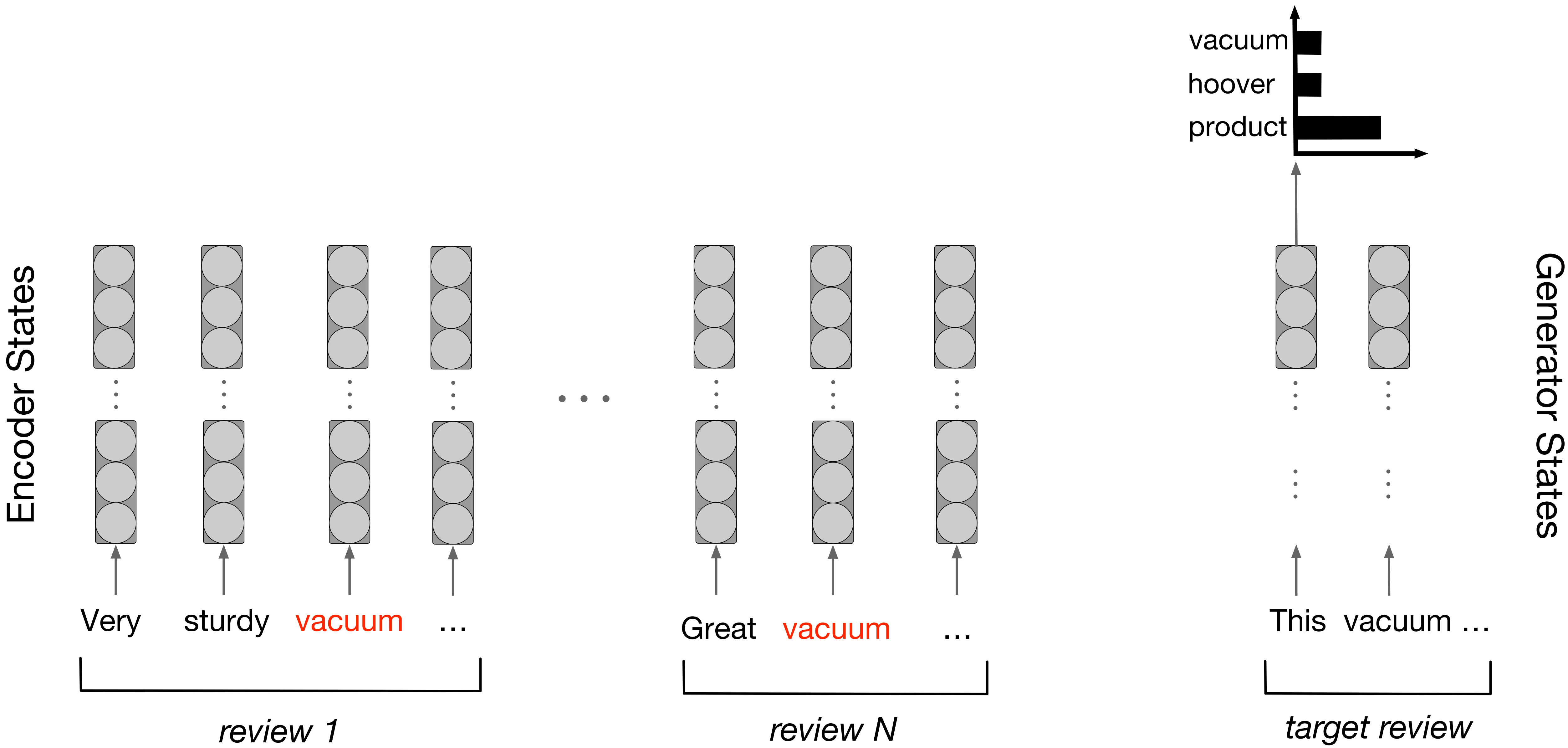
Leave-one-out



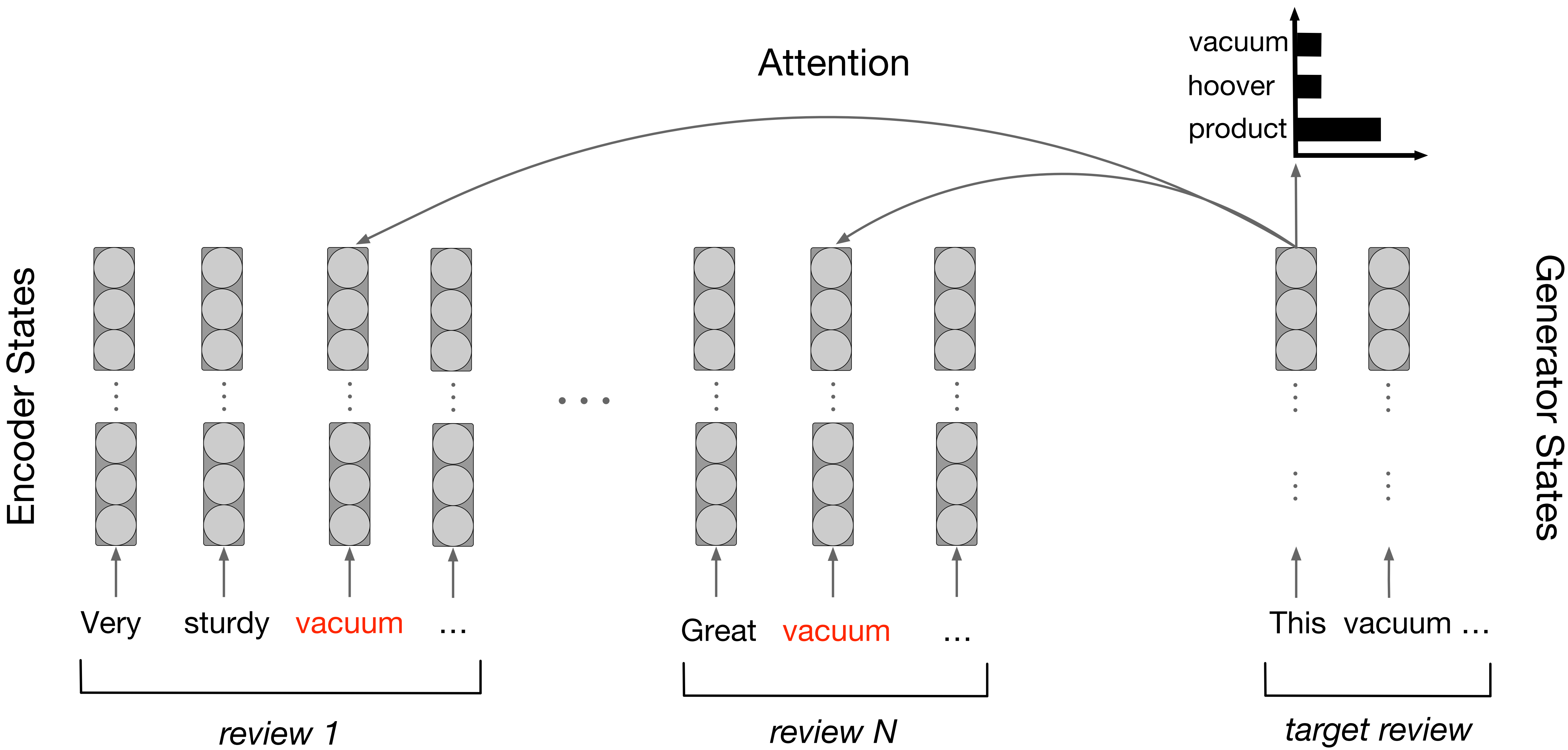
Leave-one-out



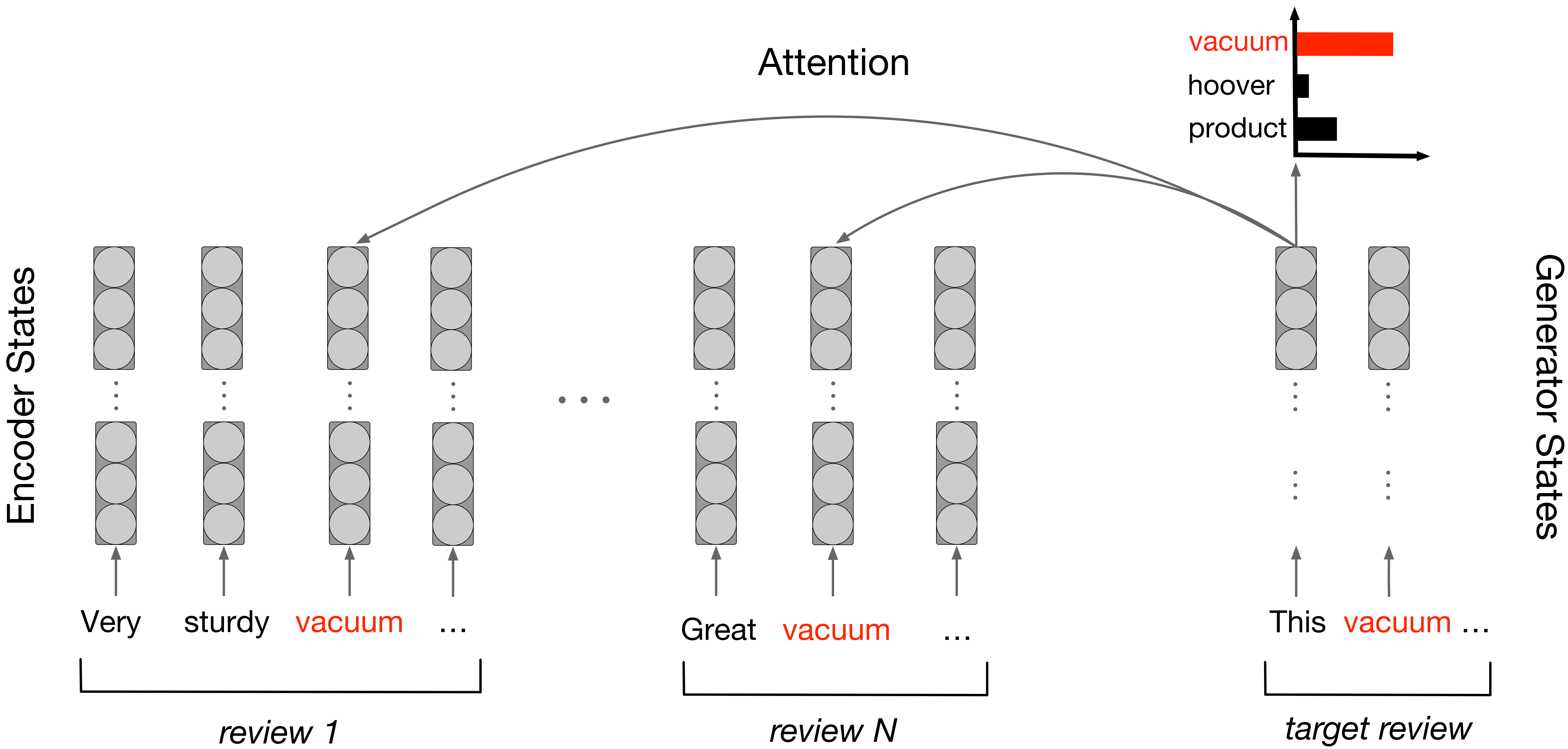
Leave-one-out



Leave-one-out



Leave-one-out



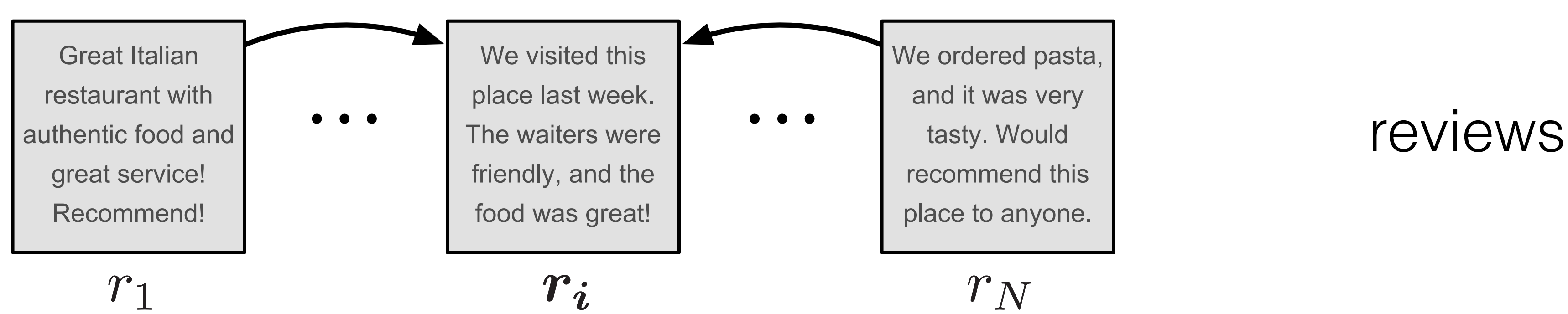
Novelty reduction

- 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**

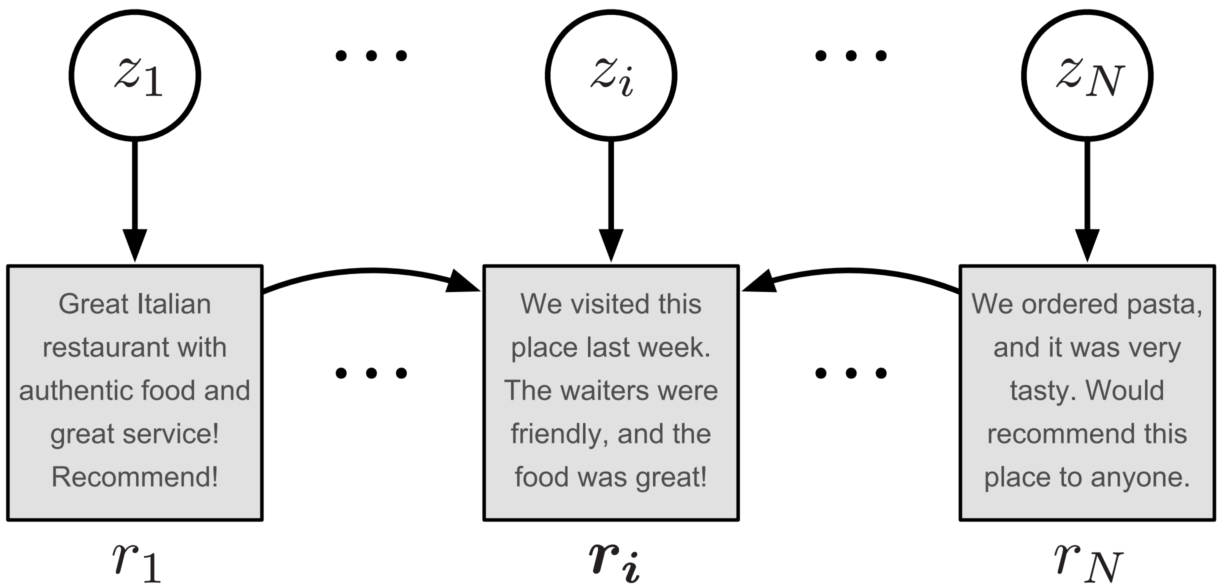
Latent model

reviews

Latent model



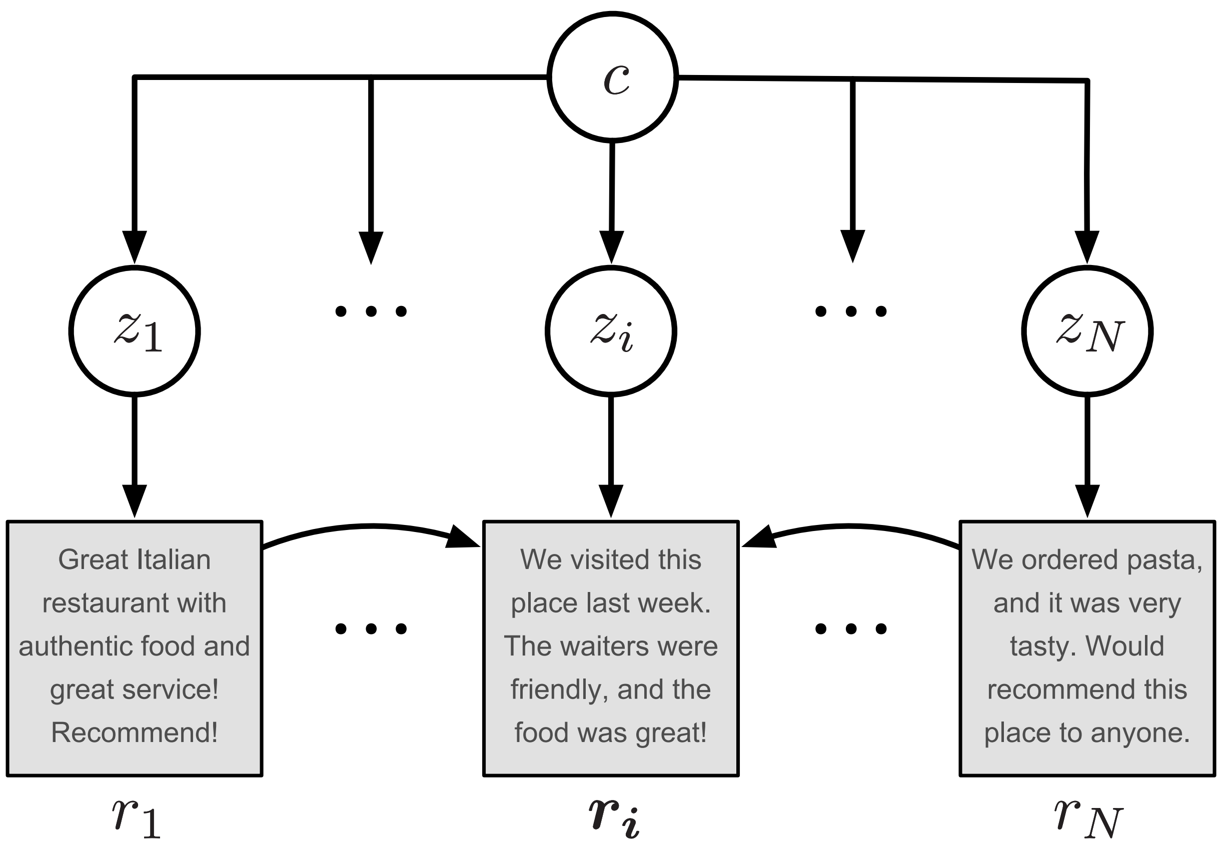
Latent model



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 informally
  - 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)
- the first in opinion summarization
- Utilizes **a handful of human-written summaries**
- Effectively **switch** an **unsupervised model** to a **summarizer**
- Summaries are written **formally** with more **informative content**

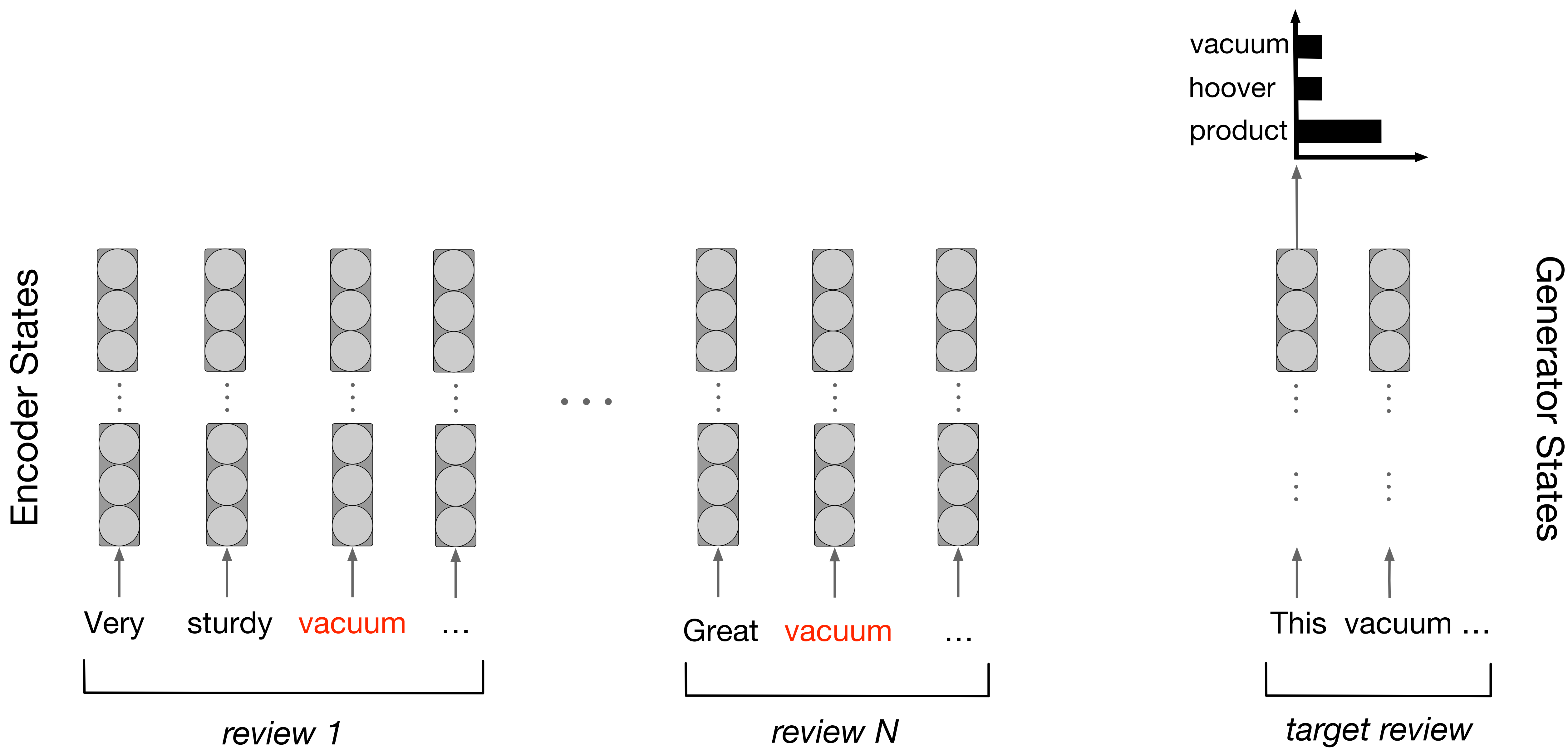
Annotated data

- 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**

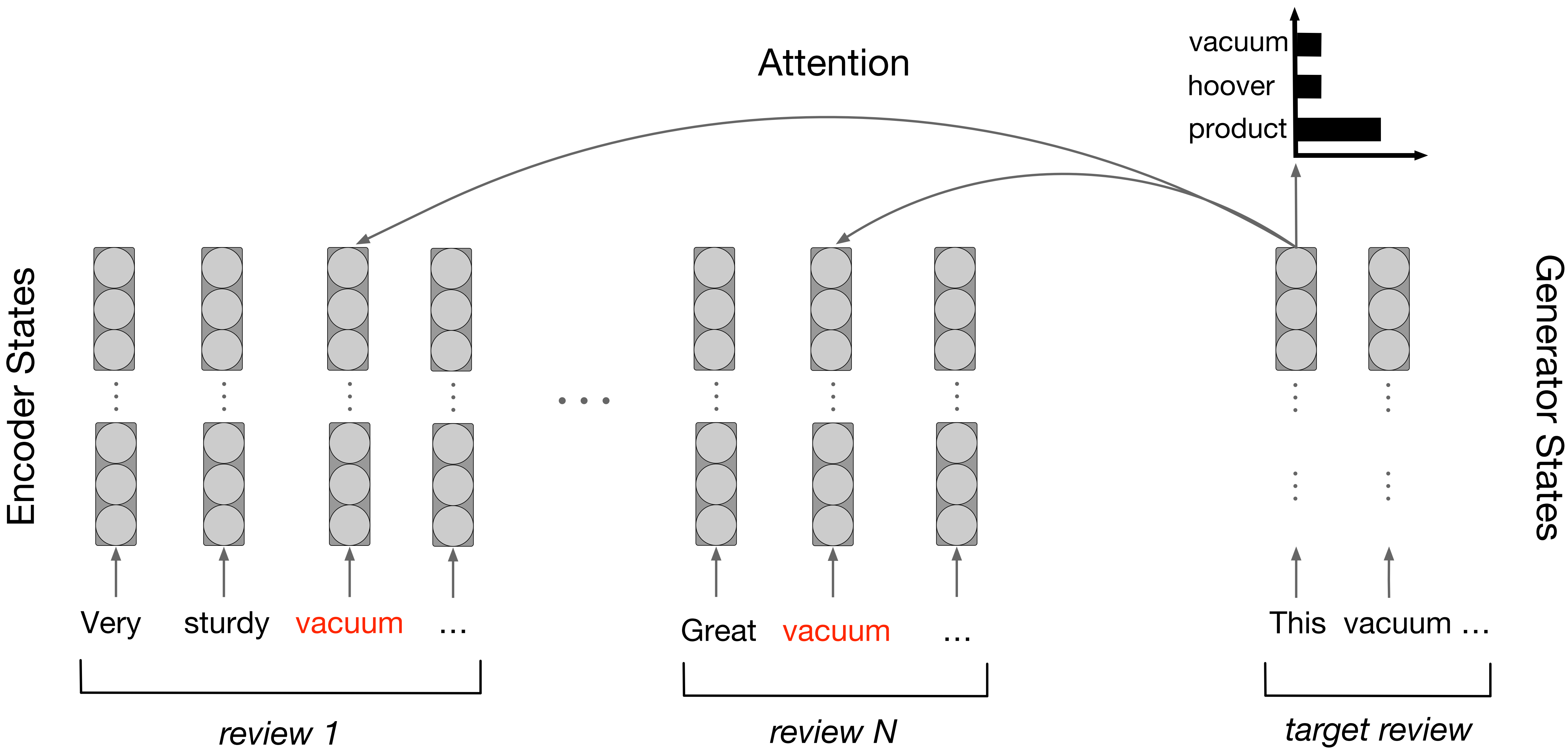
Conditional language model

- 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**

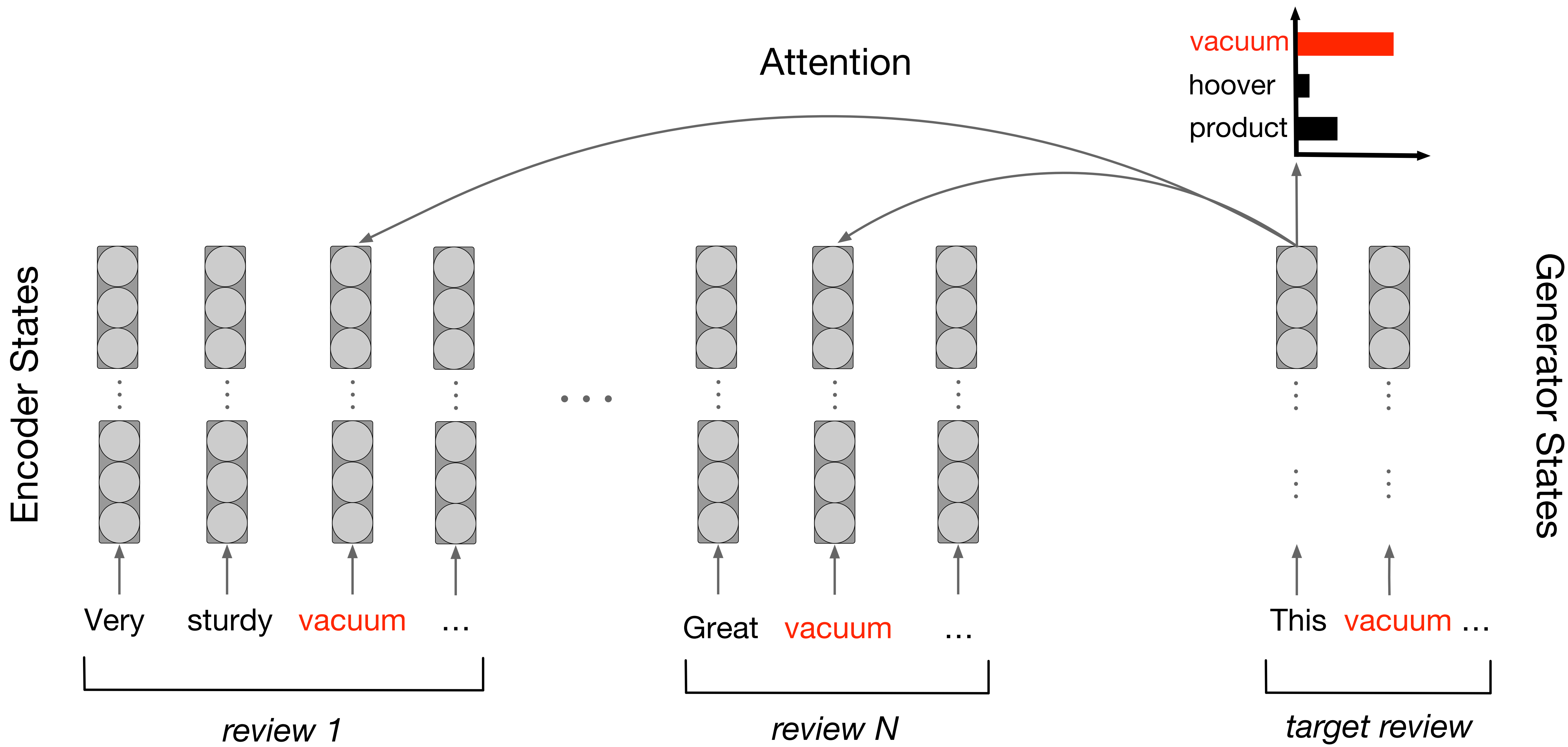
Leave-one-out



Leave-one-out



Leave-one-out



Review properties

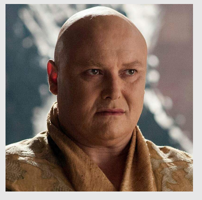
- Observation:
    - Some reviews are more like summaries
    - Some are less

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

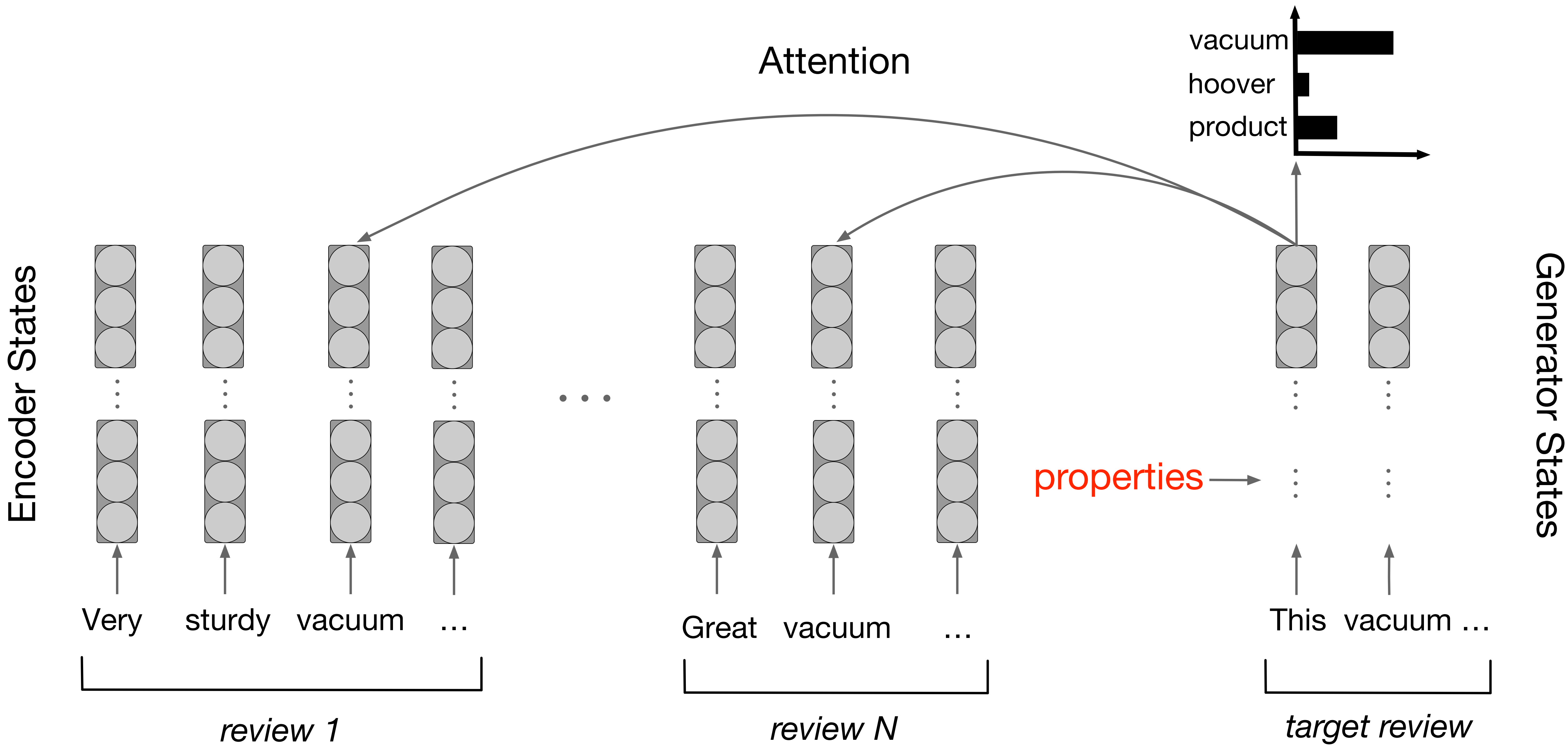
These capsules are a natural alternative to other over-the-counter medications. They are easy to swallow and have a great taste. Overall, great value for money.

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 point-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

Open Problems in Summarization

Hallucinations

- Neural generators are prone to hallucinations (Falke et al., 2019; Bražinskas et al., 2020; Krysci nski et al. 2020)
- We don’t have good metrics to capture the phenomenon (Wang et al., 2020)

Data scarcity

- Multi-document abstractive summaries are very **expensive** to produce
- The datasets are very **scarce**
- An open field for unsupervised, semi-supervised, and few-shot learning approaches

Multi-document summarization

- In multi-document review summarization we might need to summary 500+ reviews
- Infeasible due to memory constraints

Final Thoughts

Unsupervised learning

- Unsupervised learning (UL) for the **end-task** is **HARD**
- UL heavily relies on **unsupervised hypotheses:**
    - distributional hypothesis (word embeddings)
    - Hierarchical word generation process hypothesis (topic models)
    - left-right statistical text regularities (LMs)
- The hypothesis ideally needs to **substitute** what can't be learned directly from data (no annotated data)

Unsupervised learning

- In NLP we have a number of powerful classes of unsupervised models:
  - word embeddings (Mikolov et al., 2013)
  - topic models (Blei et al., 2003)
  - language models (Devlin et al., 2018; Radford et al. 2018)

Fine-tuning

These days most success is attained in NLP by further **fine-tuning** these models instead of directly using them for the end-task

Fine-tuning

- Fine-tuning can be performed in the few-shot mode yet the problem is overfitting

- Large models (millions of parameters, e.g., BART 400M)

- We observed that in our few-shot framework overfitting is alleviated as the plug-in is very parameter-compact

<END>

Contact

If any questions, contact me:

abrazinskas@ed.ac.uk