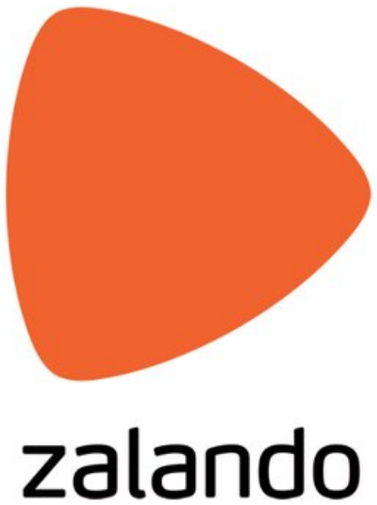
Abstractive Opinion Summarization of Customer Reviews

Arthur Bražinskas

The University of Edinburgh, Scotland



About me

Ph.D. in NLP

The University of Edinburgh
Scotland

Supervisors

Ivan Titov

Mirella Lapata

Focus on

- Abstractive opinion summarization
- Latent models (Bayesian ML methods)
- Variational inference

ML experience

Copenhagen
Denmark

ML experience

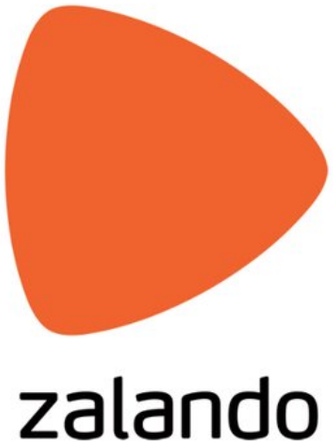
Copenhagen
Denmark

Amsterdam
Netherlands

ML experience

Copenhagen
Denmark

Amsterdam
Netherlands



Berlin
Germany

ML experience

Copenhagen
Denmark

Amsterdam
Netherlands

Berlin
Germany

Berlin; Seattle
Germany; USA

Text Summarization

Why summarization

- The amount of text documents available online is **enormous**
- Summarization useful for:
  - Faster **information consumption** for the user
  - Faster **decision making** for the user
  - Downstream utilization (analysis)

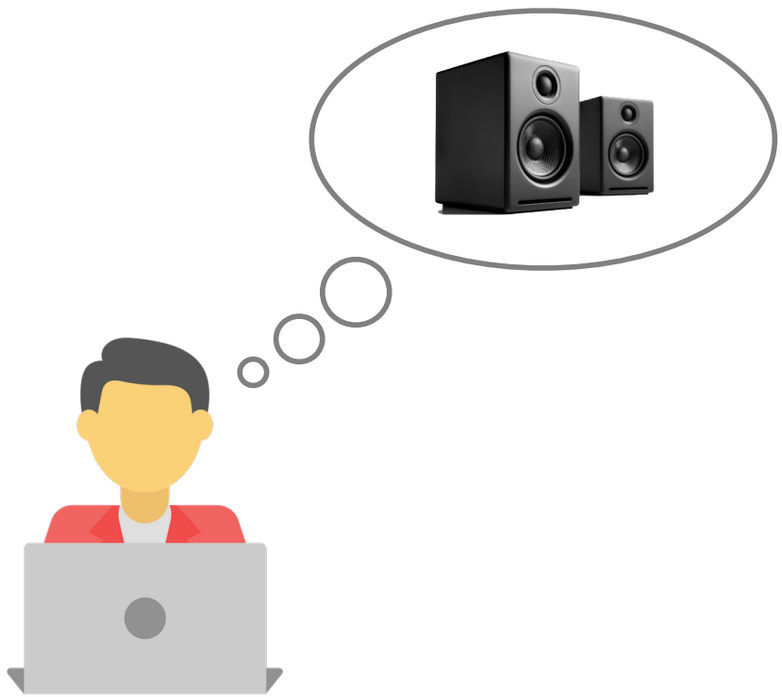
Applications

- 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 a **summary of opinions** based on user review

Opinion

Summarization

James

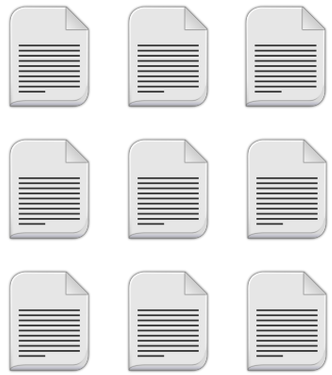


James

James

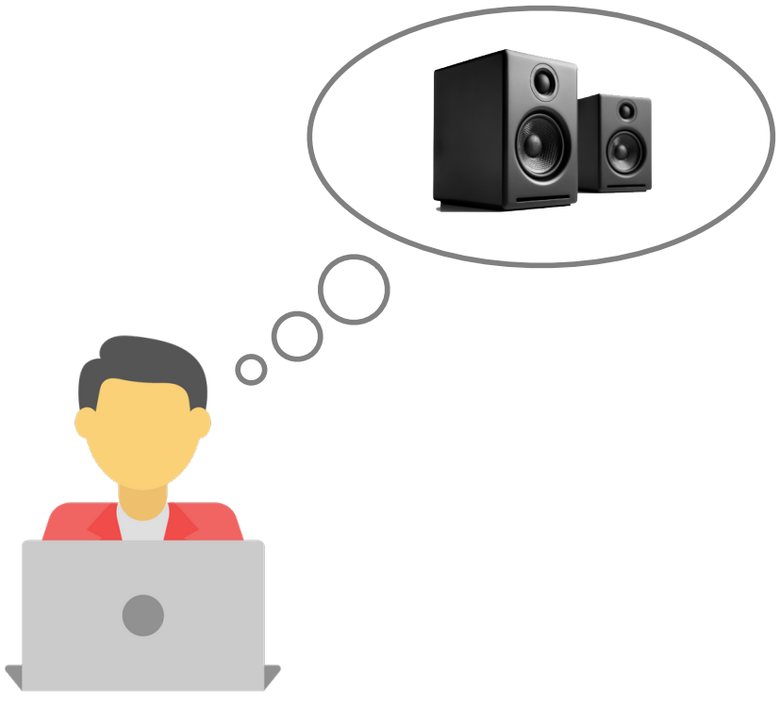
Online store

James

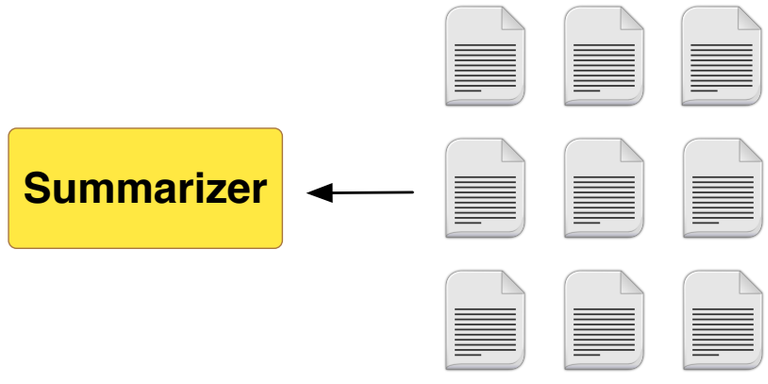


Reviews

Online store

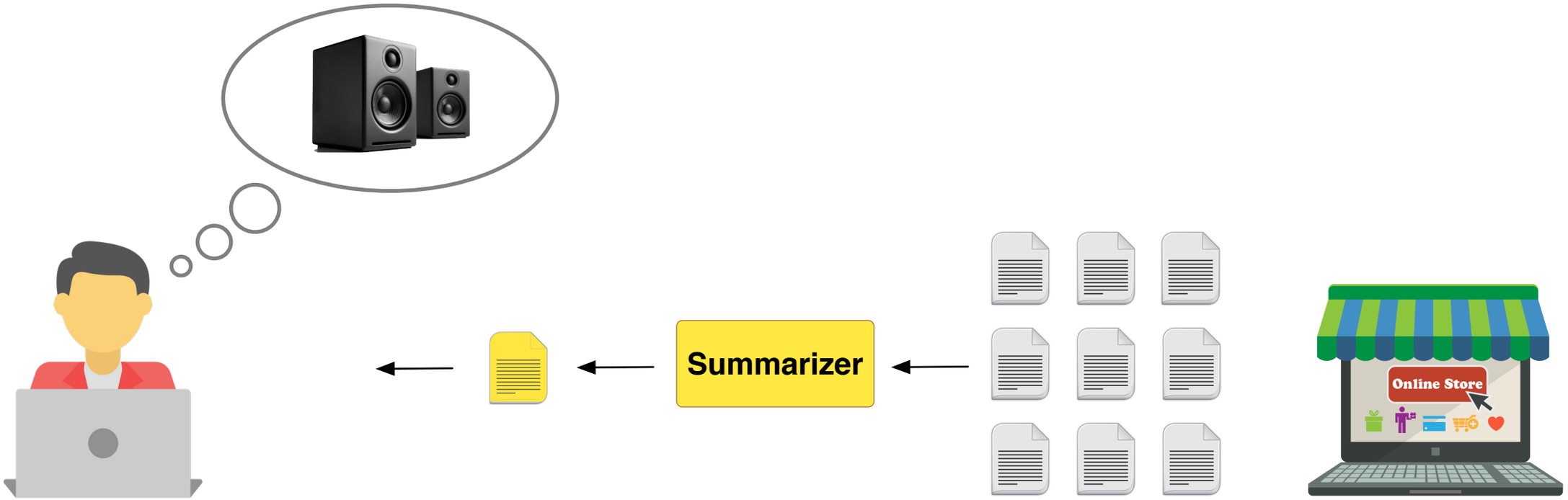


James



Reviews

Online store



James | Summary | Reviews | Online store

Extractive summarizers

- 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**

- Concatenated to form a summary

- Can be **incoherent** and contain **unimportant details**

Abstractive summarizers

- Based on the **encoder-decoder architecture**
- Generate text (Paulus et al., 2017; See et al., 2017; Liu et al., 2018)
- Can use a **rich vocabulary** of words
- Can **compress** and **fuse** input fragments

Abstractive summarizers

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

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**)

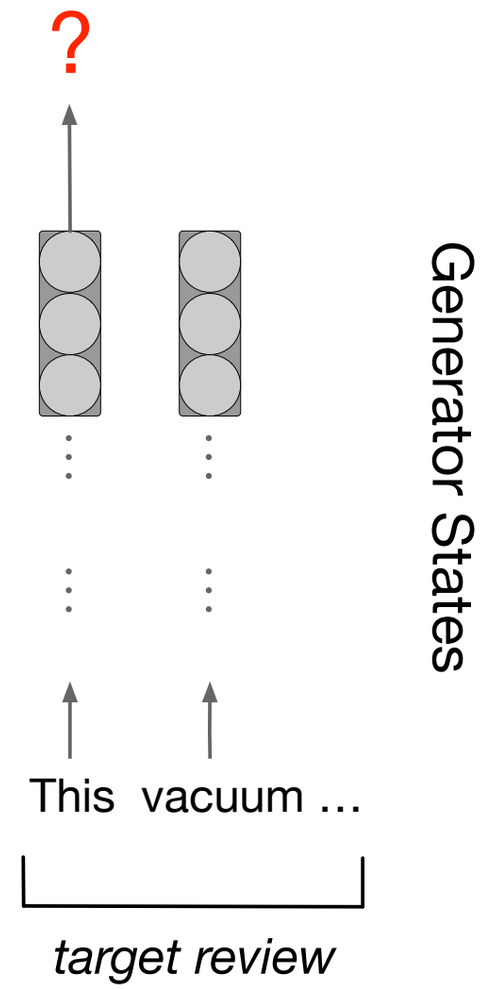
- Intuitively similar to the **pseudolikelihood estimation** (Besag, 1975)

Leave-one-out

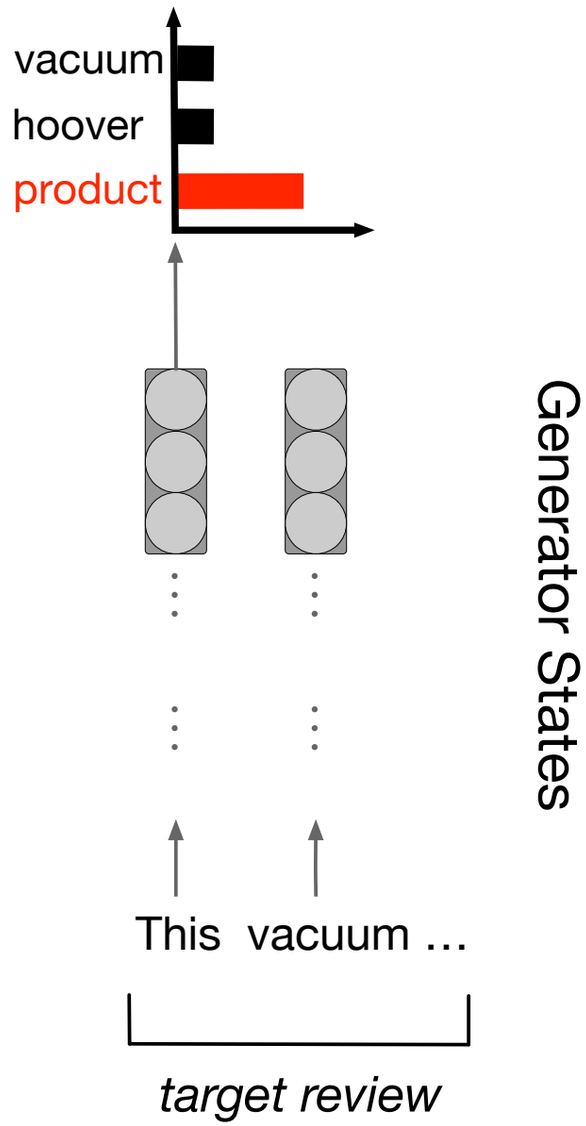
Great Italian restaurant with authentic food and great service! Recommend!	We ordered pasta, and it was very tasty. Would recommend this place to anyone.	This Italian place has the best spaghetti in the world! Strongly recommend!	We visited this place last week. The waiters were friendly, and the food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

Leave-one-out

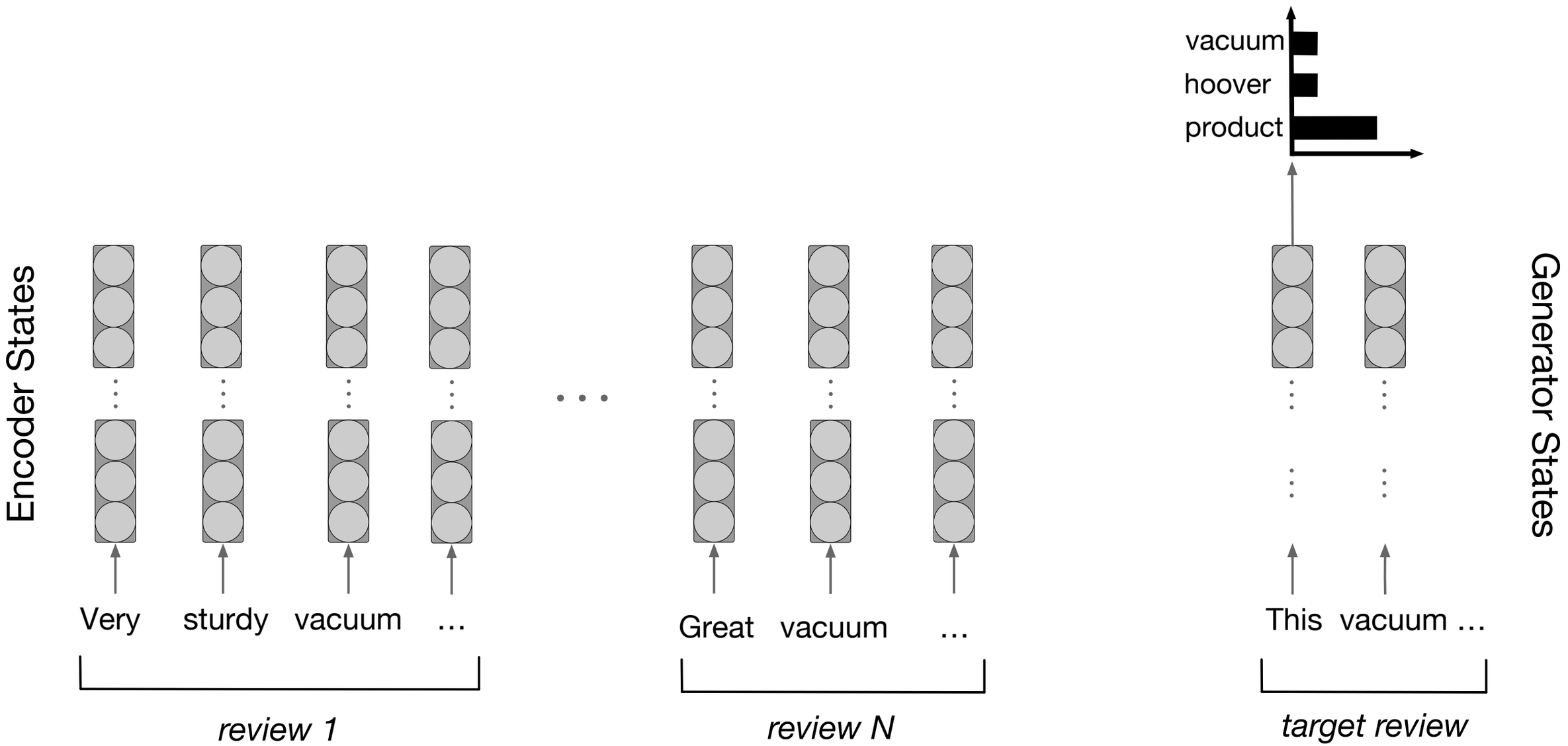
Leave-one-out



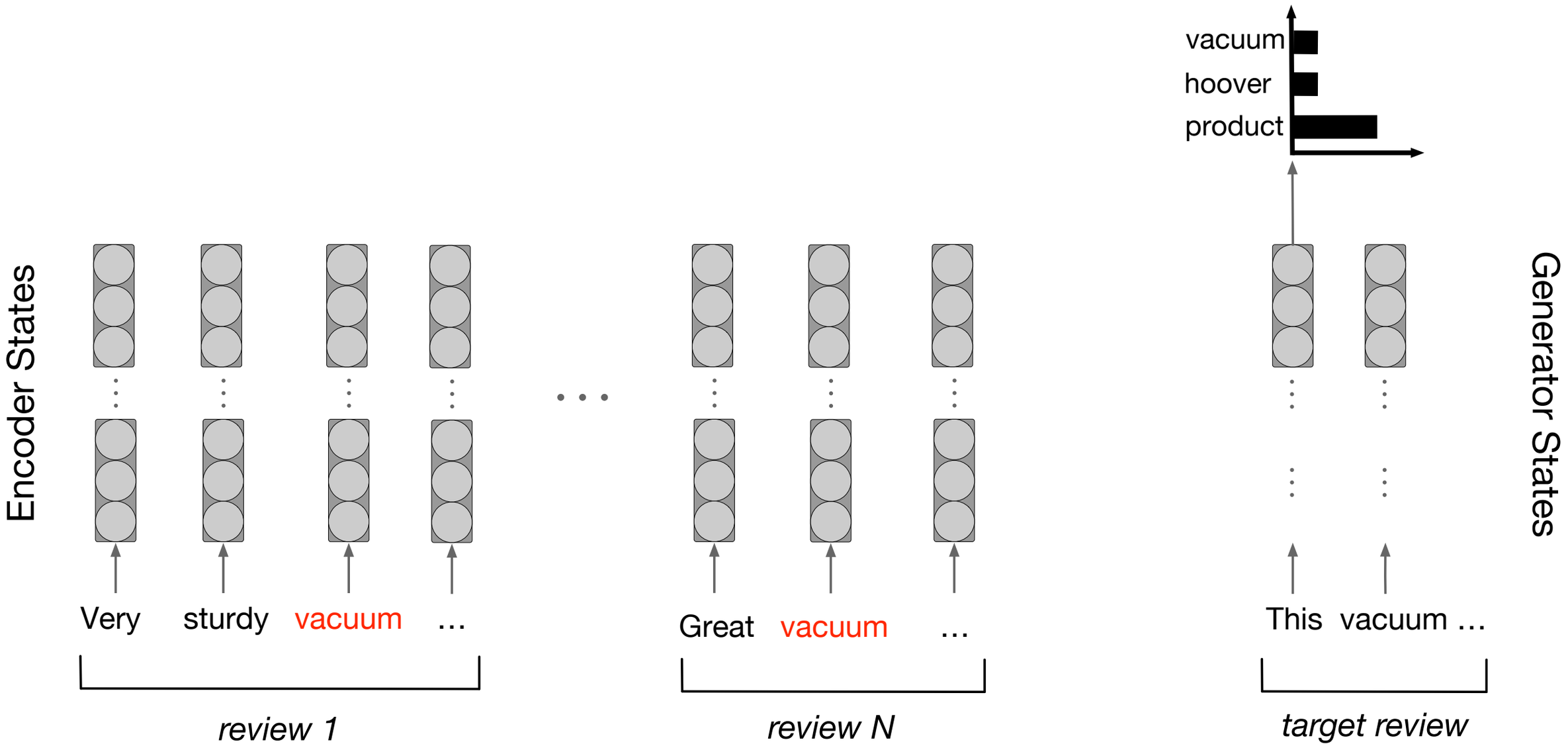
Leave-one-out



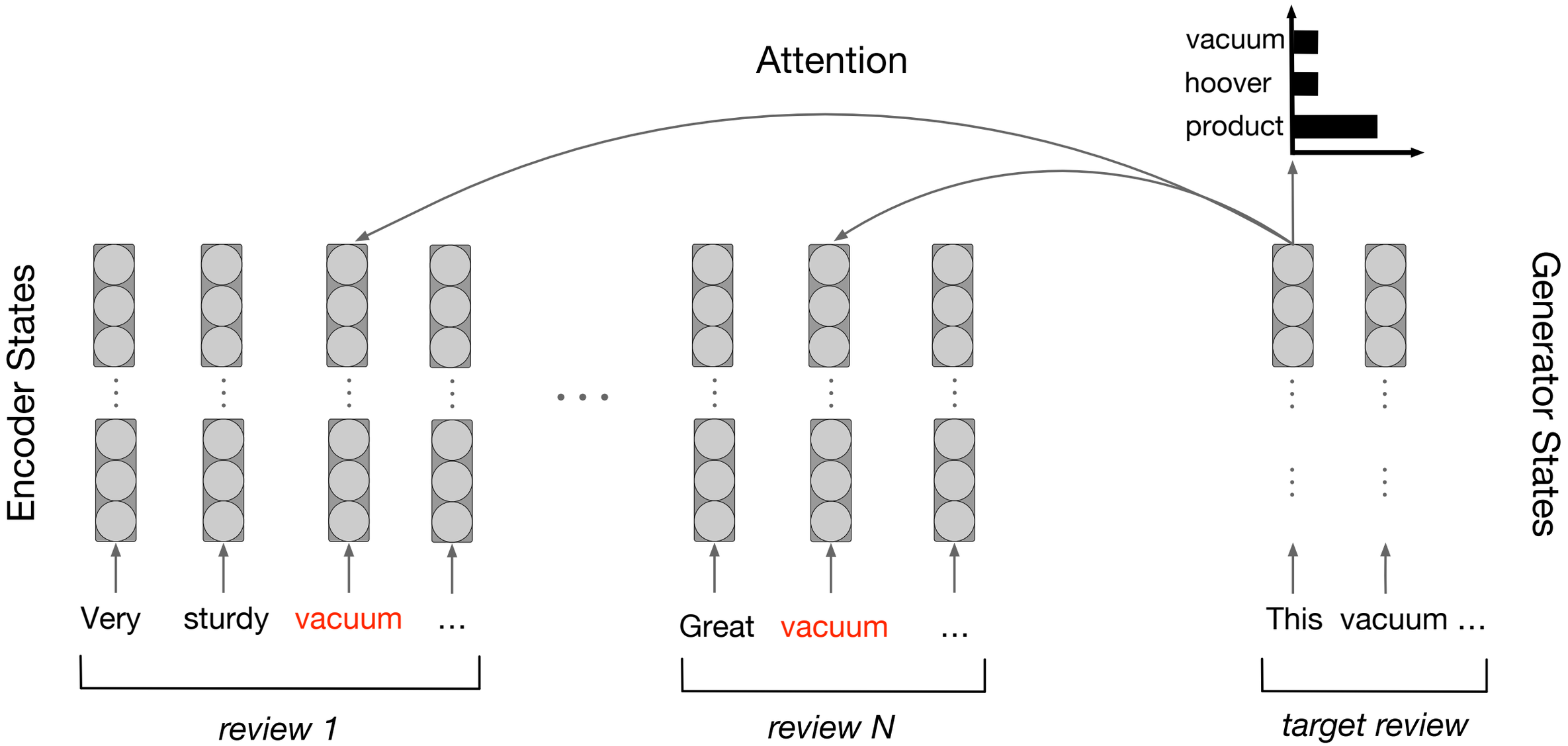
Leave-one-out



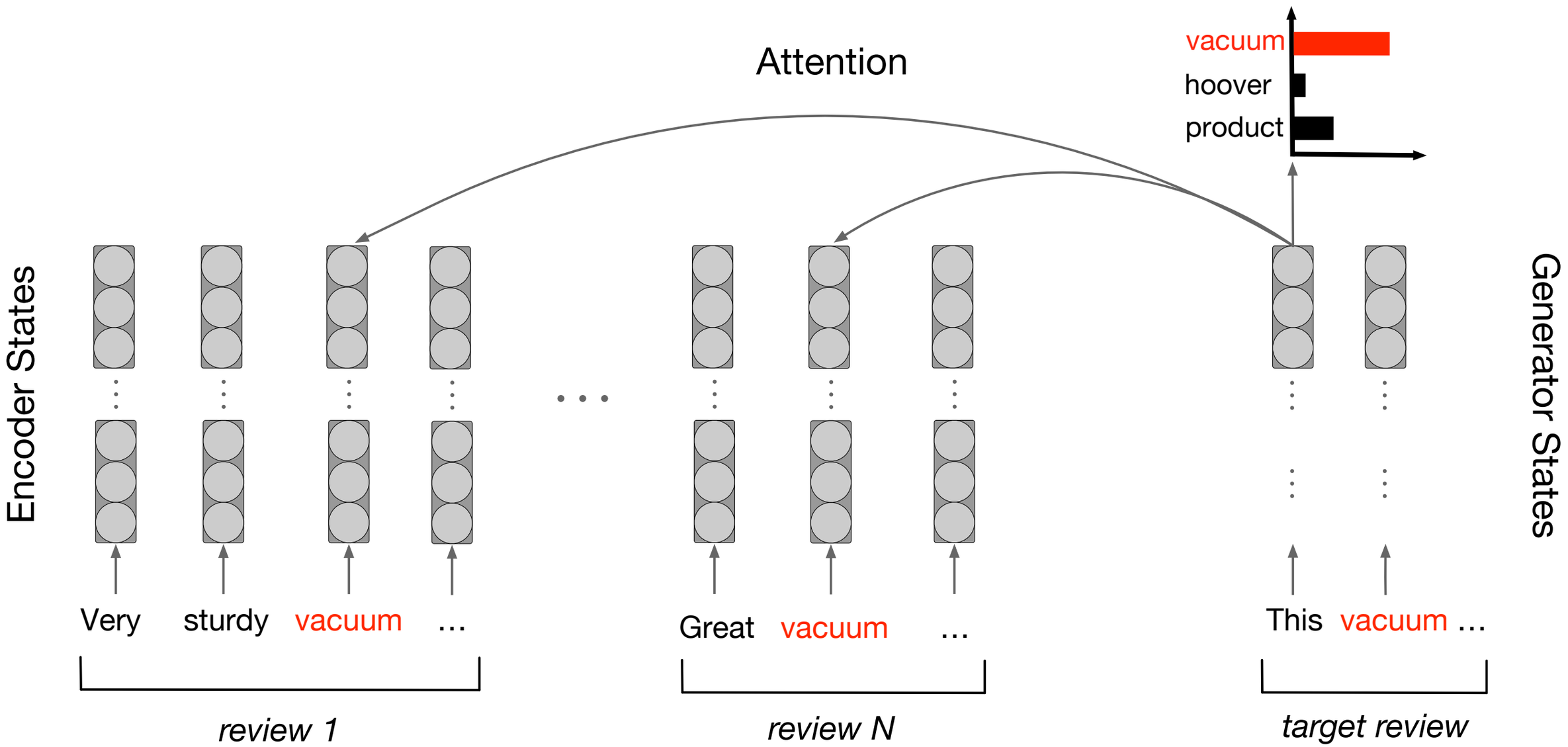
Leave-one-out



Leave-one-out



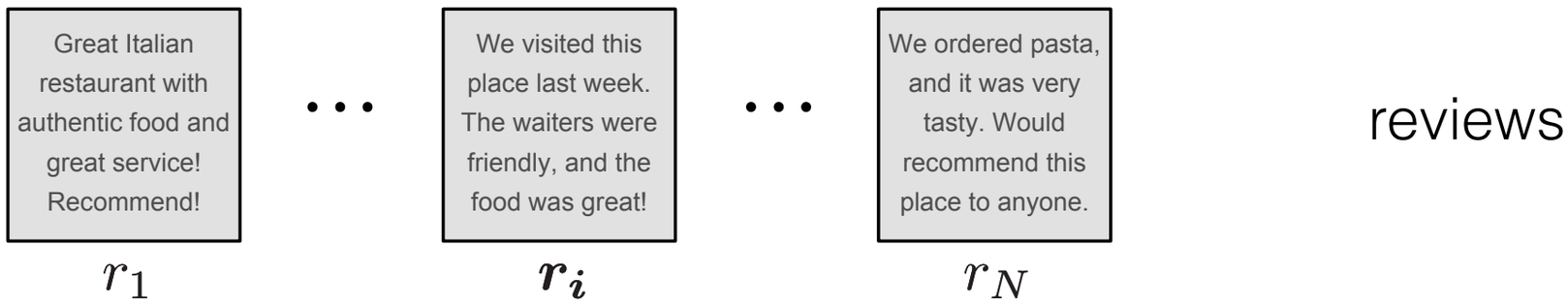
Leave-one-out



Novelty reduction

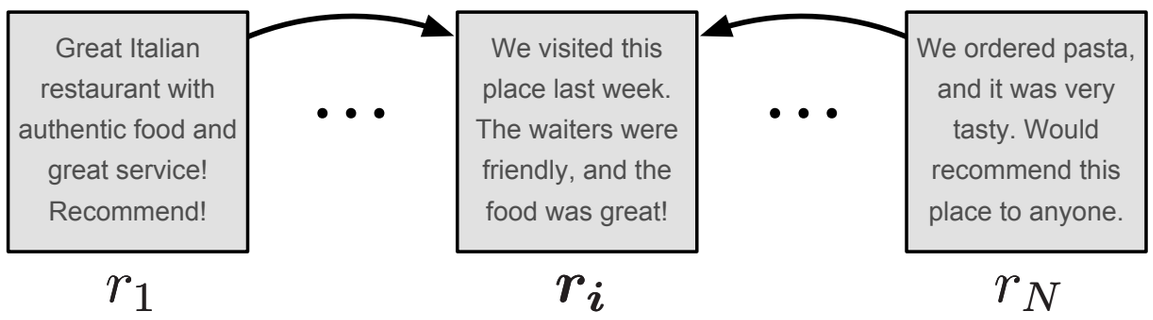
- Model is trained to predict reviews
- Summaries are different from reviews in content
- Summaries do not have **novel content**
- Control the amount of ‘novelty’ via **latent variables**

Latent model



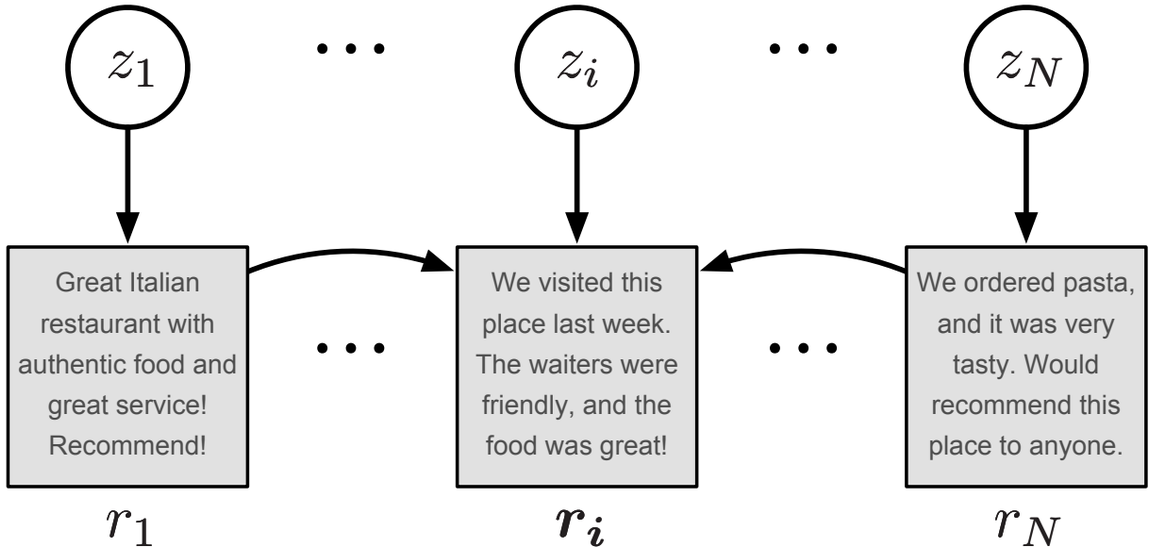
reviews

Latent model



reviews

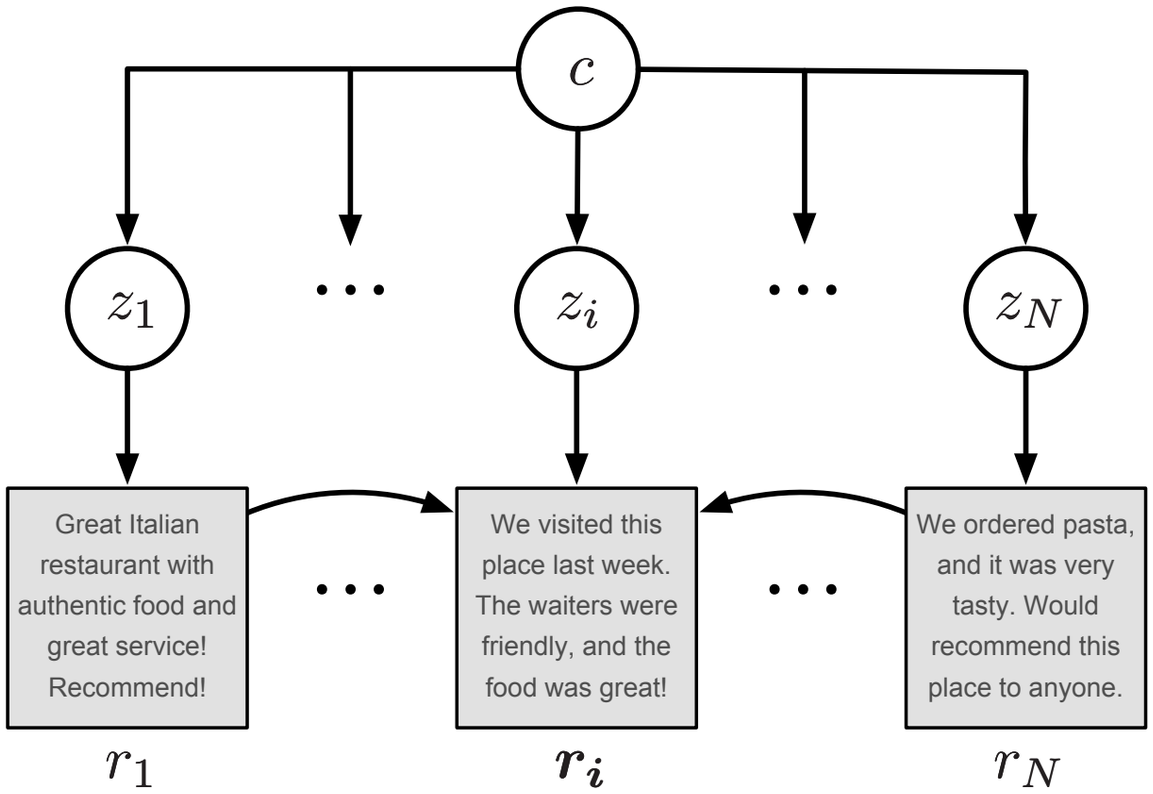
Latent model



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 the generator maps them 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

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: $2\frac{P * R}{R + P}$ **(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)

Baselines

- **Lead**: leading sentences in each review used as a summary

- **MeanSum** (Chu and Liu, 2019) is an encoder-decoder unsupervised abstractive summarizer

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 summaries
- 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 the first **few-shot learning** framework (FewSum)

- Utilizes **a handful of human-written summaries**

- Effectively **switch** an **unsupervised model** to a **summarizer**

- Summaries are written **formally** with more **informative content**

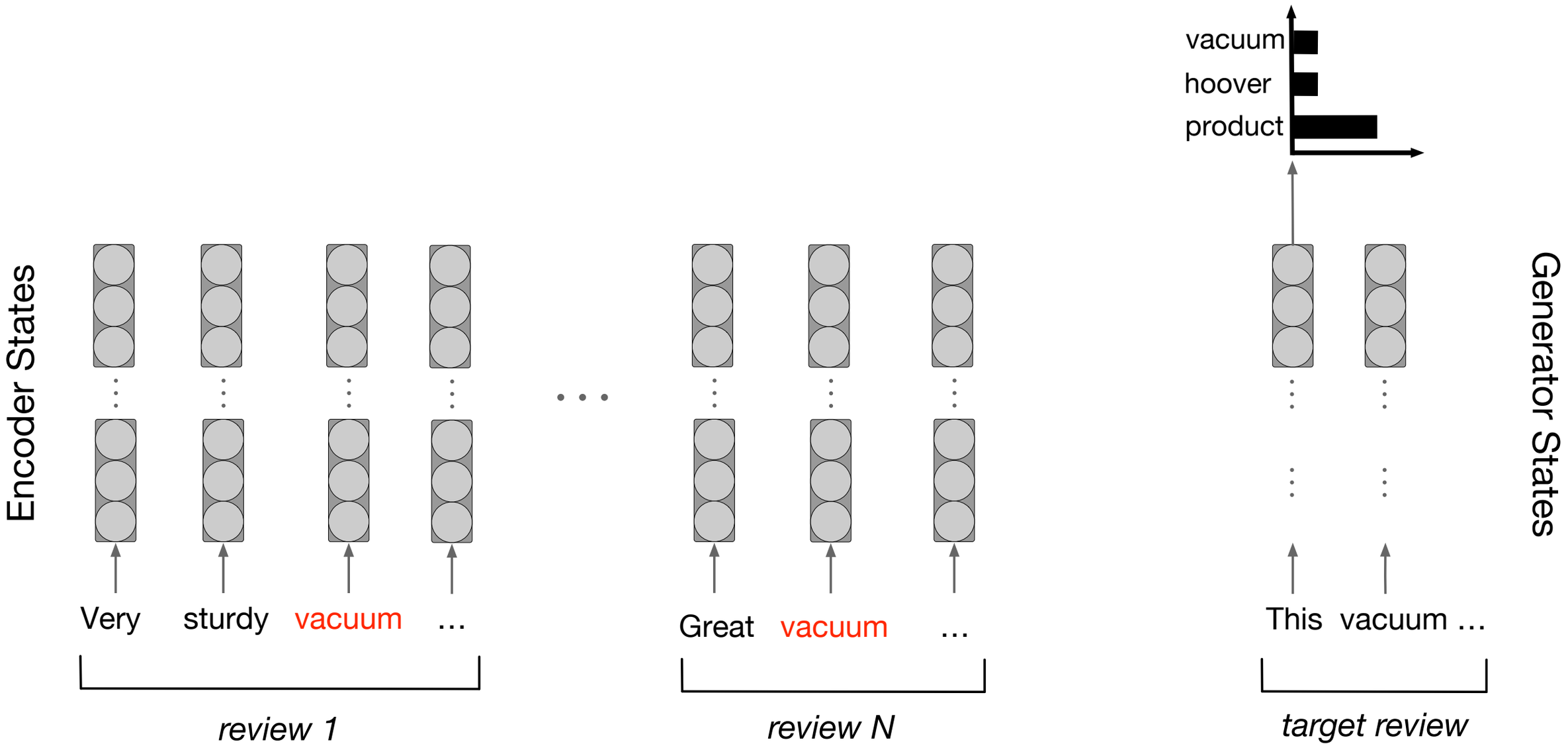
Annotated data

- 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**

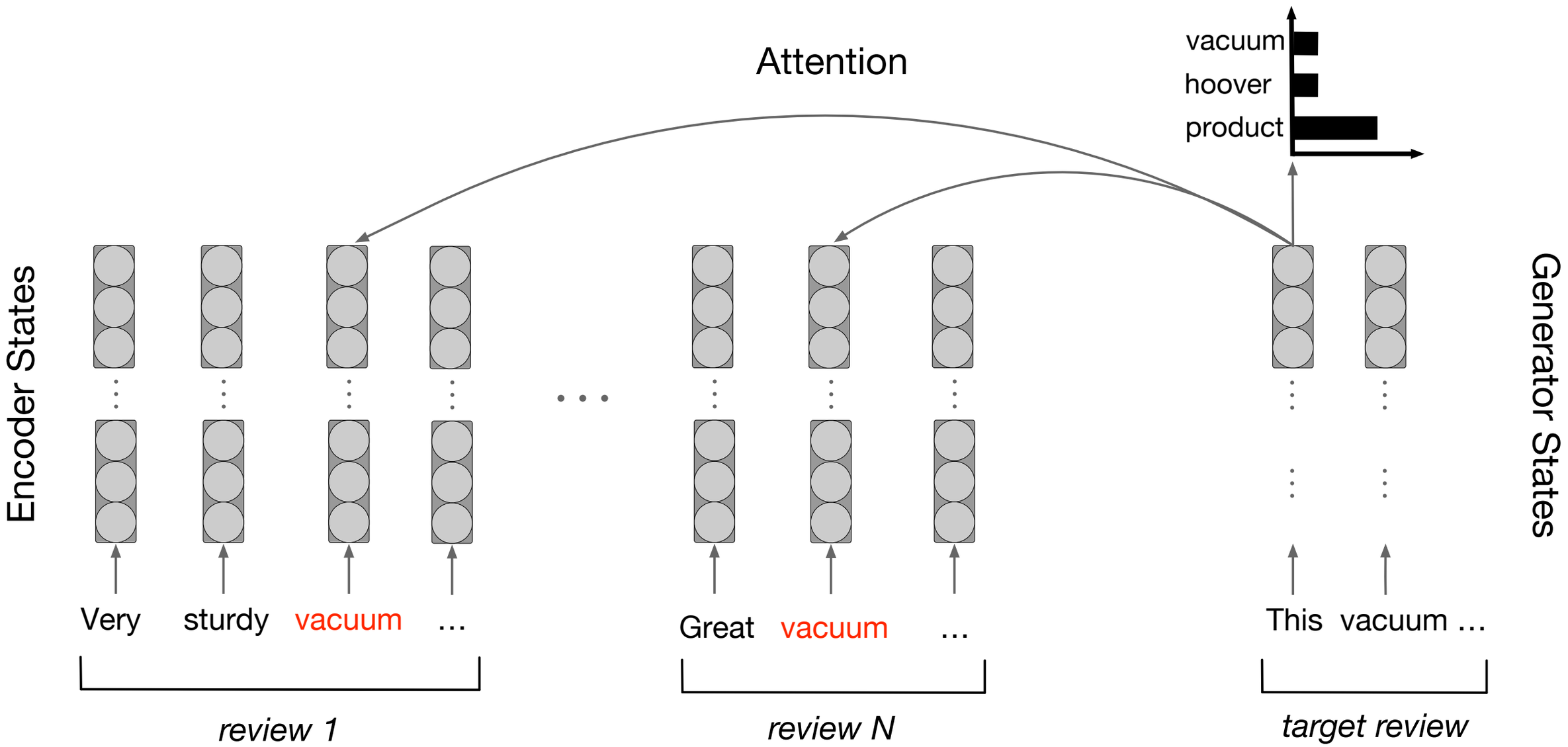
Conditional language model

- 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**

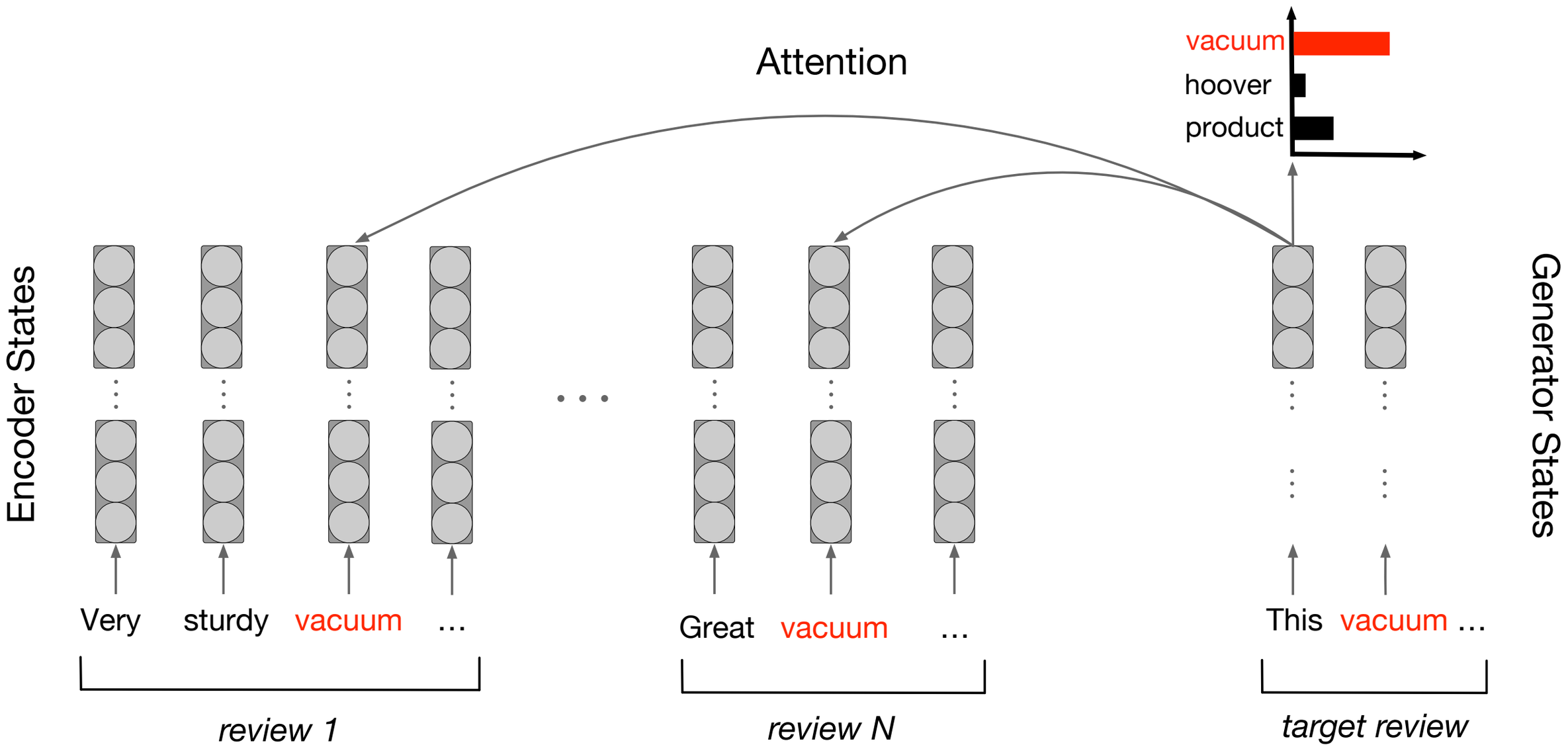
Leave-one-out



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.

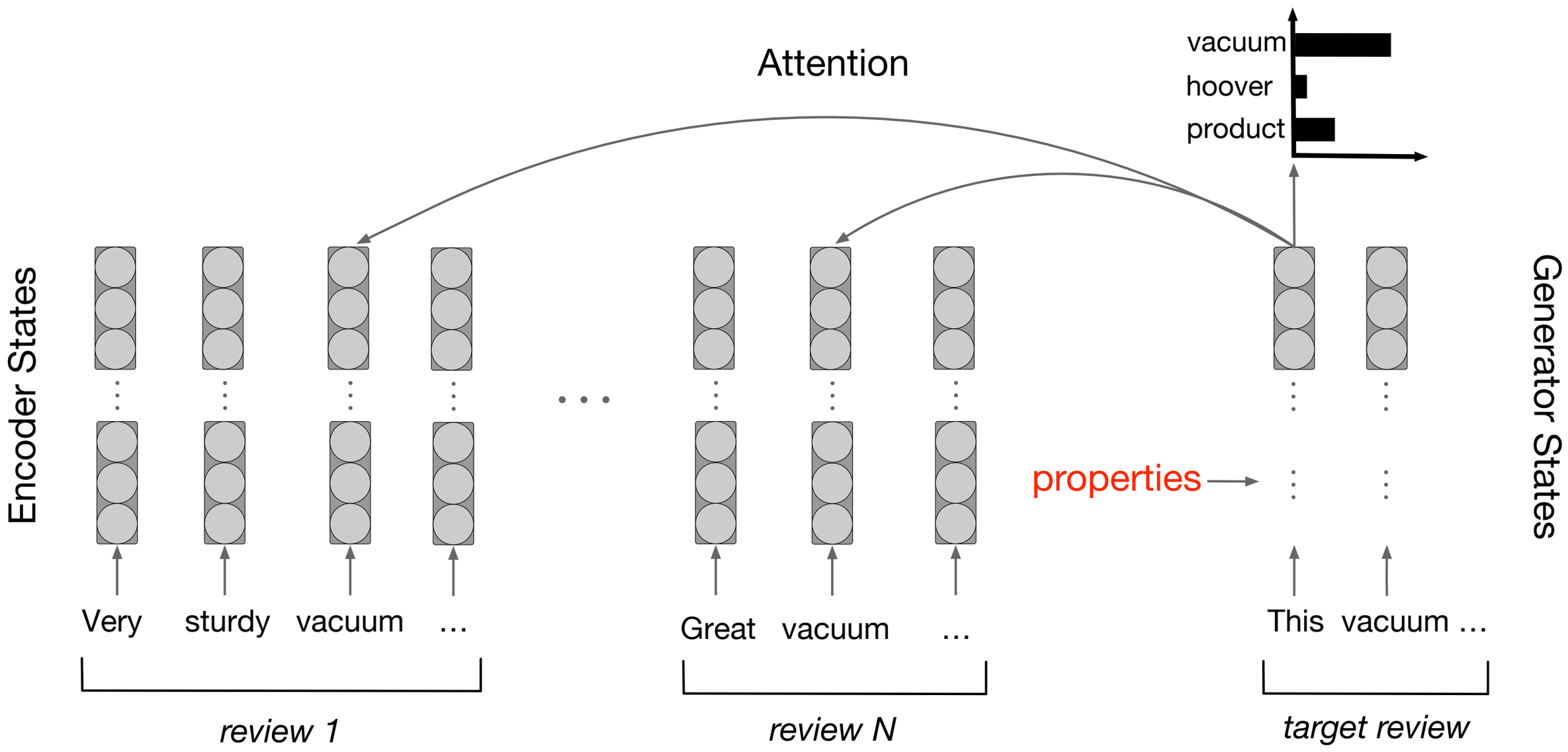
Review 2

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
Very
sturdy
vacuum
…
review 1
Great
vacuum
…
review N
properties
This
vacuum
…
target review
Generator States

Properties

- Are like features
- Used as an additional input to the generator/ decoder
- Are calculated using **an oracle** for tuples (source, **target**)

Property types

Type	Reviews	Summaries	Implementation
Content coverage	Can contain novel content	Only content present in reviews	ROUGE scores
Writing style	Informal	Formal	Pronoun counts
…	…	…	…

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 assignments** 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 computes properties for (source, target)
- **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.

Bottom-line

- FewSum produces summaries that are:
  - Written in the **formal writing style**
  - **Informative** and **sentiment** aligned to reviews
  - Better in automatic evaluation (ROUGE)
  - Substantially more **preferred by humans**

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 **well established metrics** to capture the phenomenon (Wang et al., 2020)

Multi-document summarization

- In multi-document review summarization we might need to summary 500+ reviews

- Computationally **infeasible** via the standard encoder-decoder architecture due to memory constraints

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

<END>

Contact

If any questions, please contact me:

abrazinskas@ed.ac.uk