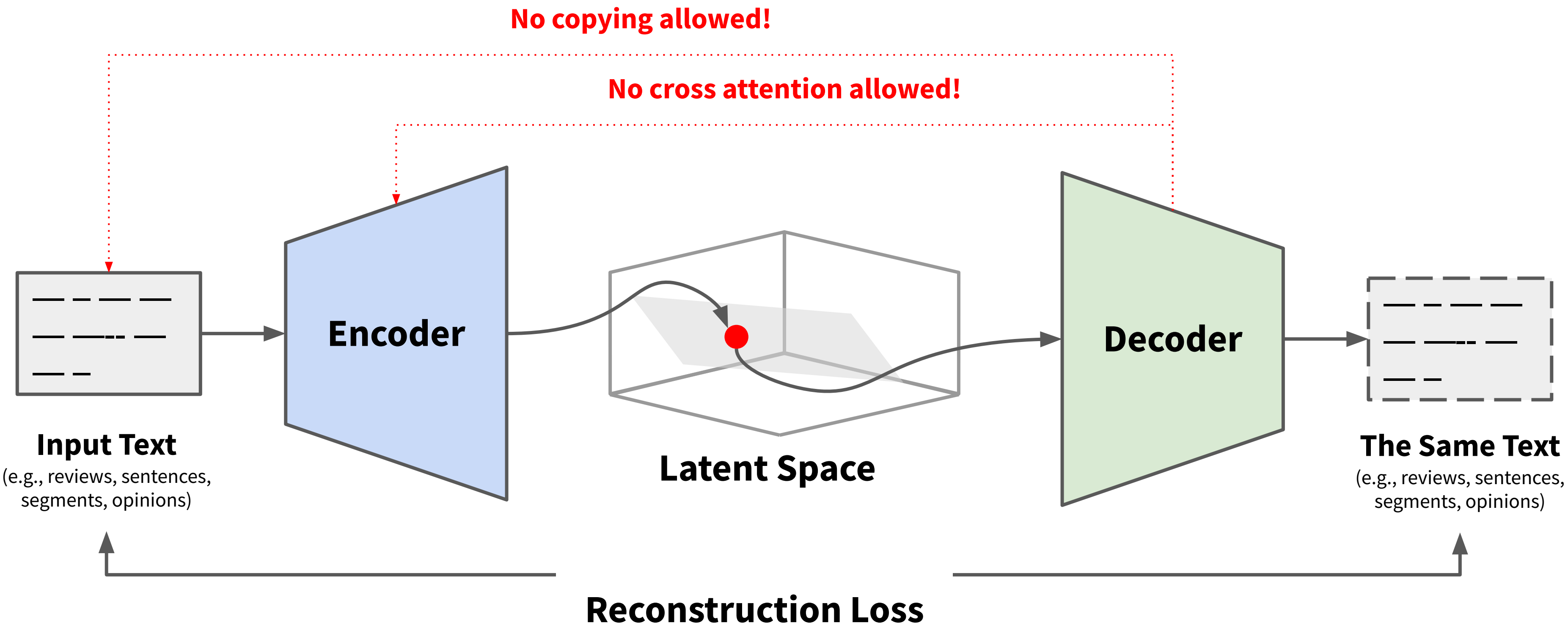
Synthetic Dataset Creation

Major Limitation of Autoencoder Approaches



“The Decoder Dilemma”

Use “strong” decoders?

- Model learns to use shortcuts
- Latent space ends up not being used

Use “weak” decoders?

- No encoder-decoder cross attention, no copying
- Sequence prediction performance is limited

Synthetic Dataset Creation

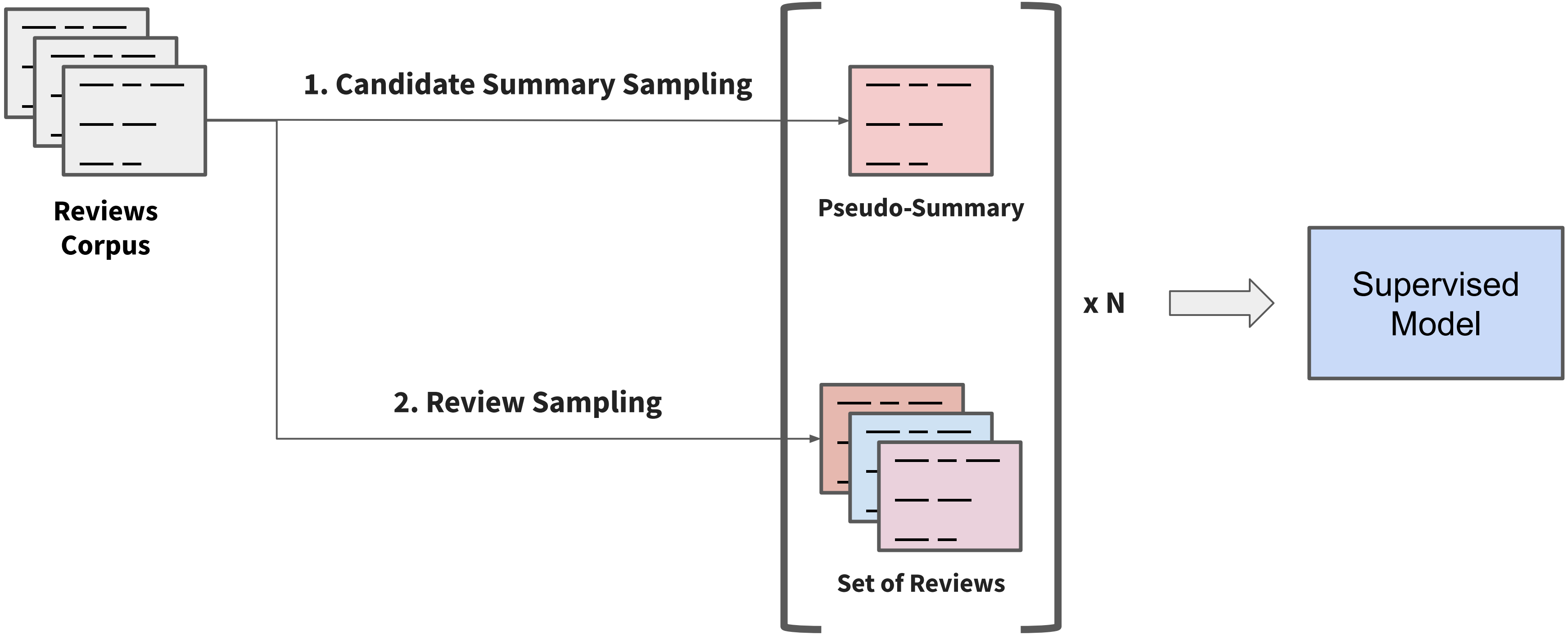
Motivation: We need supervised methods to leverage “strong” decoders.

What if we can synthesize supervised datasets from review corpora?

Advantages:

1. Allows the use of **supervised models**, which is most of the literature
2. **Large-scale dataset** creation is possible (given large-scale review datasets)
3. Related to **self-supervised learning**, which has shown to improve recent NLP models

Synthetic Dataset Creation



Candidate Summary Sampling

How should a good opinion summary look like?

- Contains **informative** opinions
  - The location is great.
  - The location is close to attractions.
- Written in the **third person**
  - I stayed with my dog in this hotel.
  - This is a dog-friendly hotel.
- Written in a **brief manner**
  - This hotel is perfect!
  - The hotel is close to attractions. There is a famous monument nearby, a museum in 2-min walk, and an airport in 15 mins bus ride. There are also … (100 more tokens)
  - The hotel is close to attractions. The staff are friendly and the rooms are air-conditioned.

Candidate Summary Sampling

Candidate summaries should be informative, brief, and written in third person.

Can we select such summaries without gold data? No…

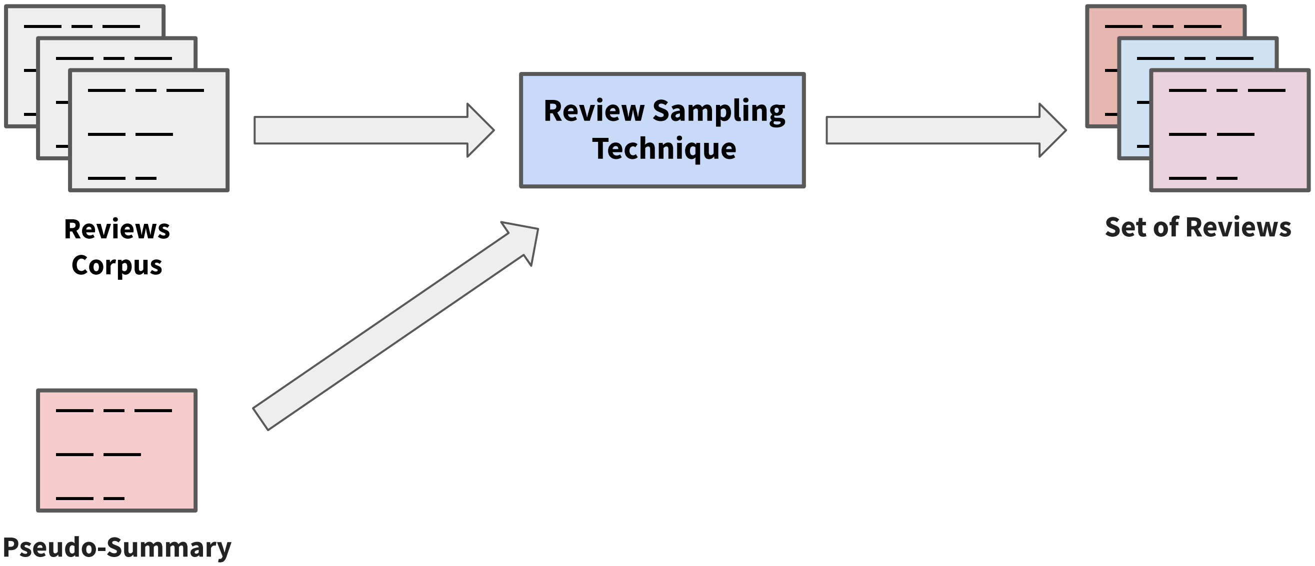
But we can use heuristics!

- Select only high-IDF reviews[1]
- Select only reviews within a specific range of length[2]
- Filter out sentences that have first-person pronouns[3]

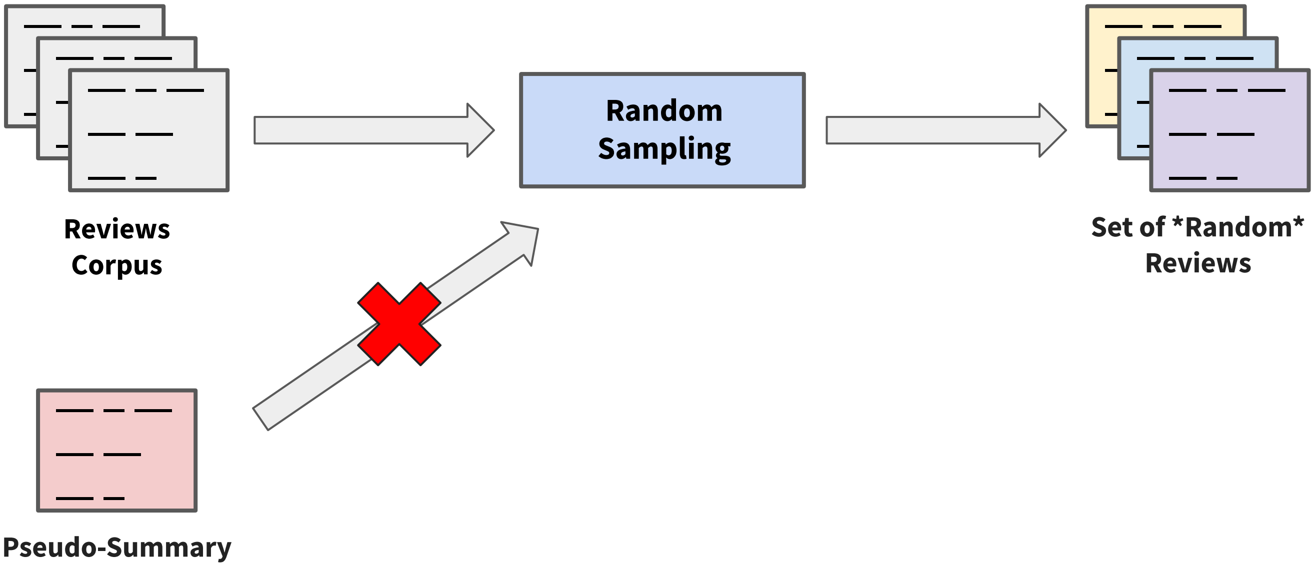
What if we have few of such data?

1. Elsahar, Hady, Maximin Coavoux, Jos Rozen, and Matthias Gallé. "Self-Supervised and Controlled Multi-Document Opinion Summarization." In *EACL*, pp. 1646-1662. 2021.
2. Amplayo, Reinald Kim and Mirella Lapata. “Unsupervised Opinion Summarization with Noising and Denoising.” In *ACL*, pp. 1934–1945. 2020.
3. Amplayo, Reinald Kim, Stefanos Angelidis, and Mirella Lapata. "Aspect-Controllable Opinion Summarization." In *EMNLP*, pp. 6578-6593. 2021.

Review Sampling



Random Review Sampling

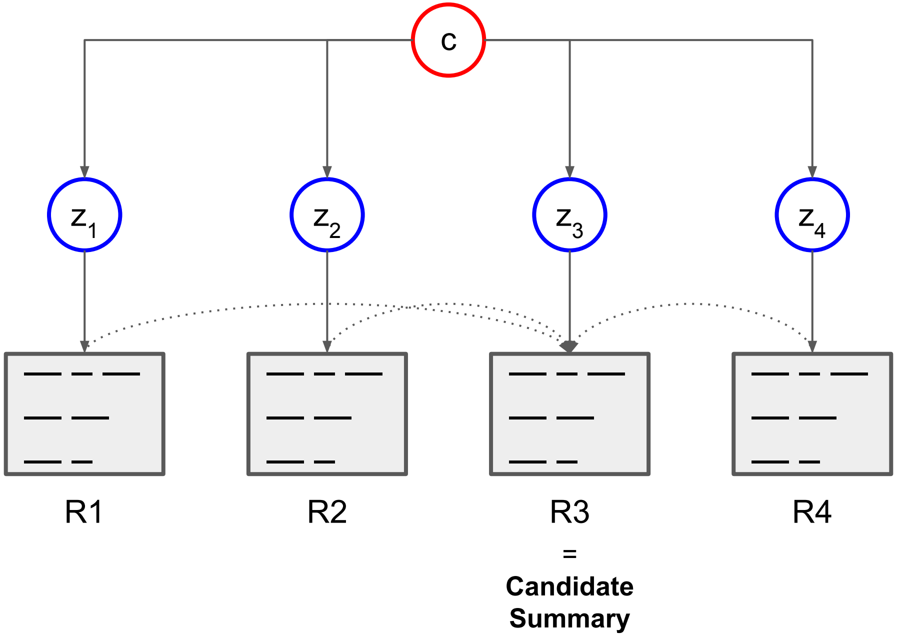


Copycat[1]

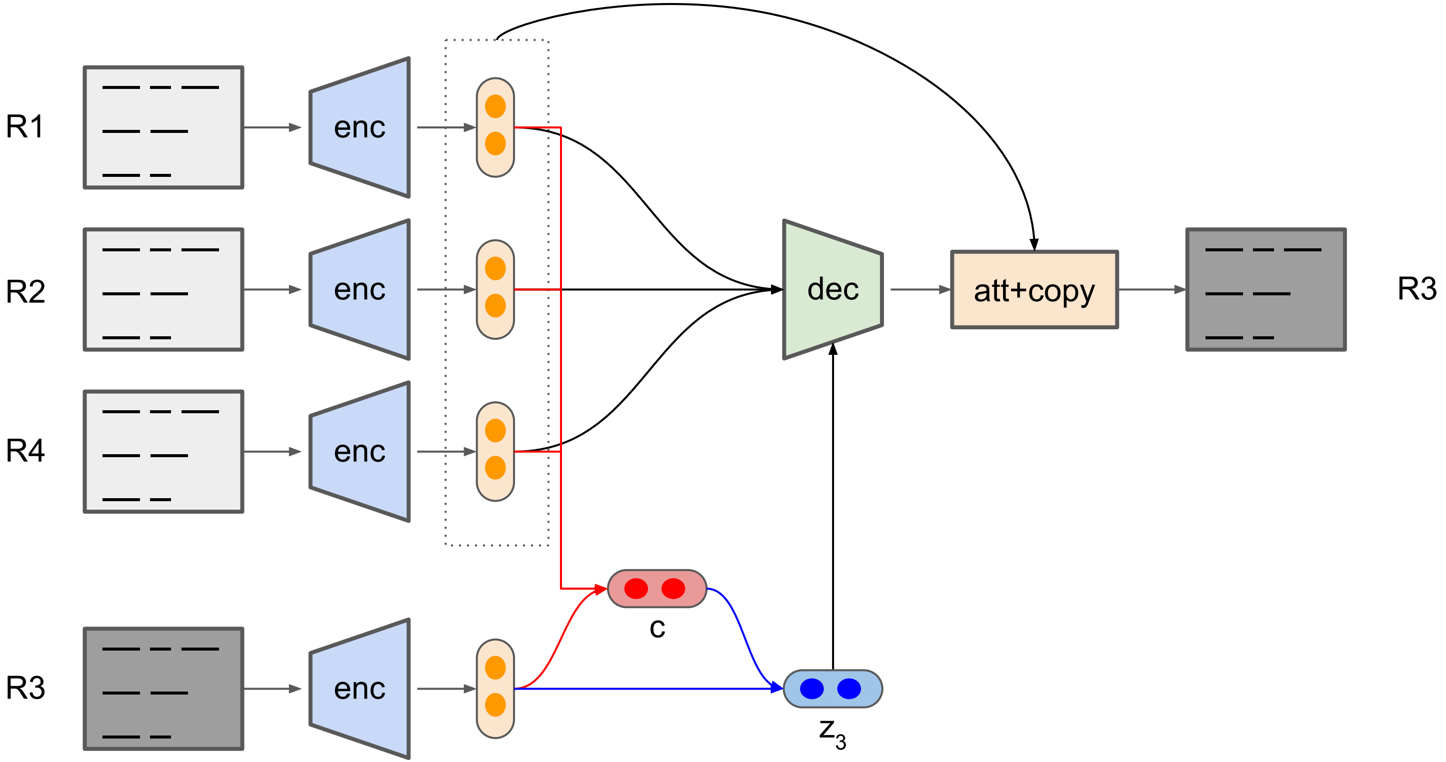
- Use “leave-one-out” strategy: Reconstruct a review (=candidate summary) using N random reviews
- Related to skip-gram representations[2] and masked language modeling[3]
- Use variational inference to infer a latent code of the candidate summary
- Attend and copy mechanisms can now be used

1. Bražinskas, Arthur, Mirella Lapata, and Ivan Titov. "Unsupervised Opinion Summarization as Copycat-Review Generation." In *ACL*, pp. 5151-5169. 2020.
2. Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In *NIPS*. 2013.
3. Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *NAACL-HLT*, pp. 4171-4186. 2019.

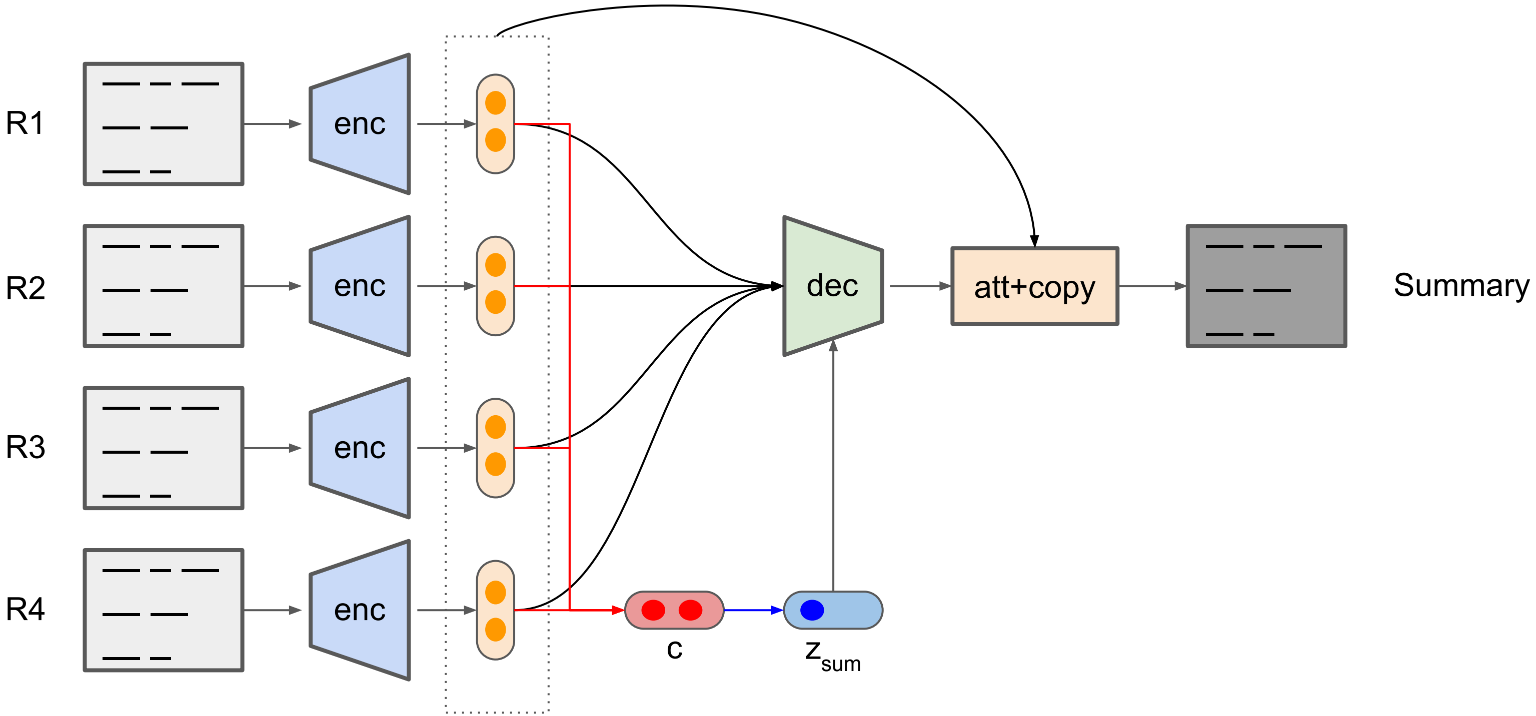
Copycat: Graphical Representation



Copycat: Training Time



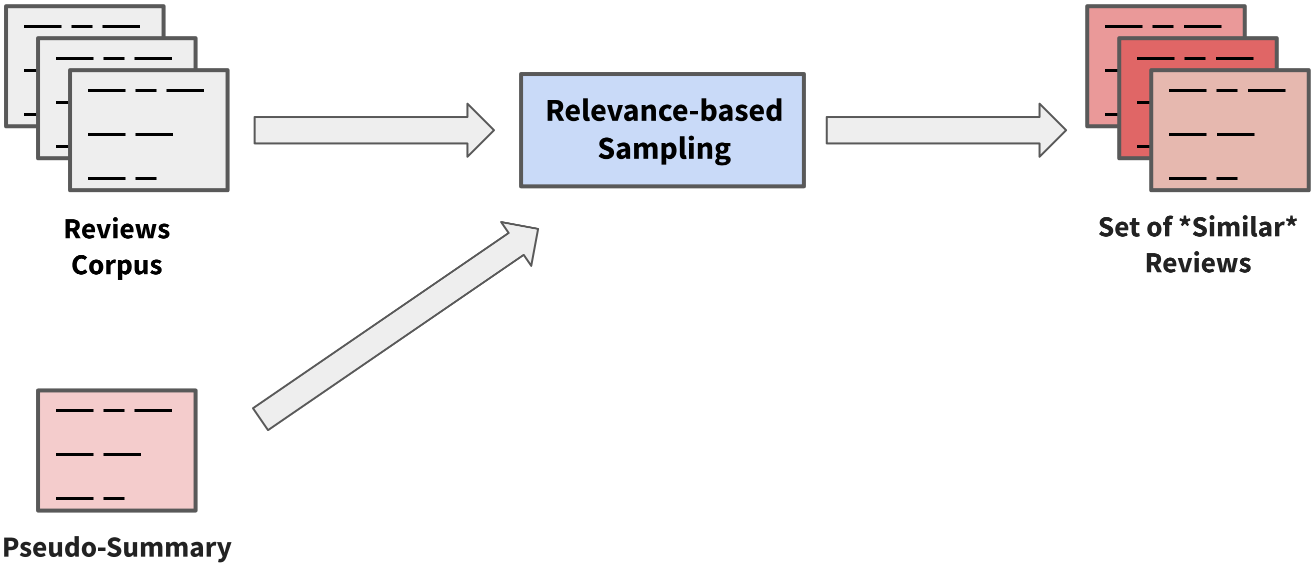
Copycat: Inference Time



Summary

Review Sampling Method	Advantages	Disadvantages
Random Sampling	● Unlimited Training Data	● Encourages hallucination
Relevance-based Sampling	● Model has a better understanding of what to learn	● Does not capture real-world opinion variance in reviews
Review Noising	● Introduces phrase-level variation	● Encourages grammatical errors
Planned Sampling	● Can capture real-world opinion variance in reviews	● Planning stage may propagate errors

Relevance-based Review Sampling



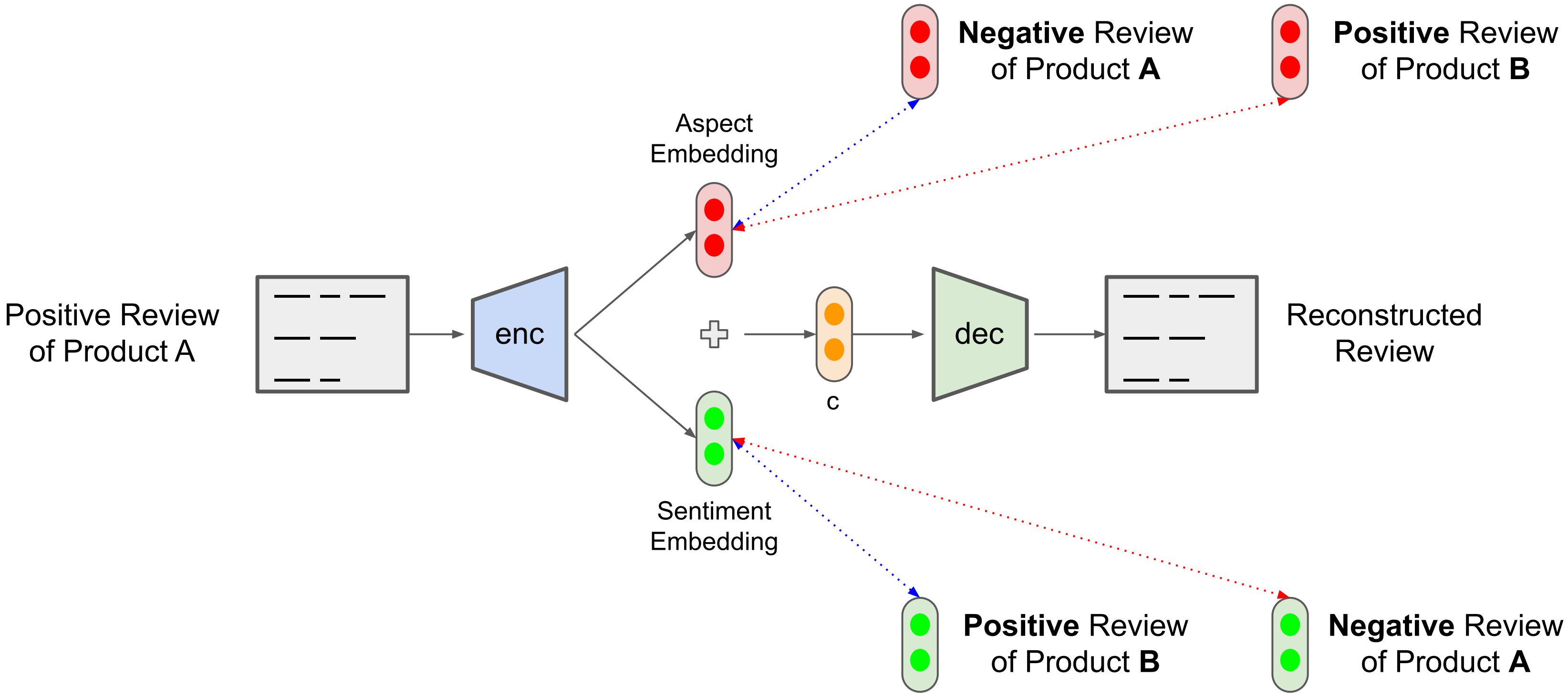
TransSum[1]

Consists of two components

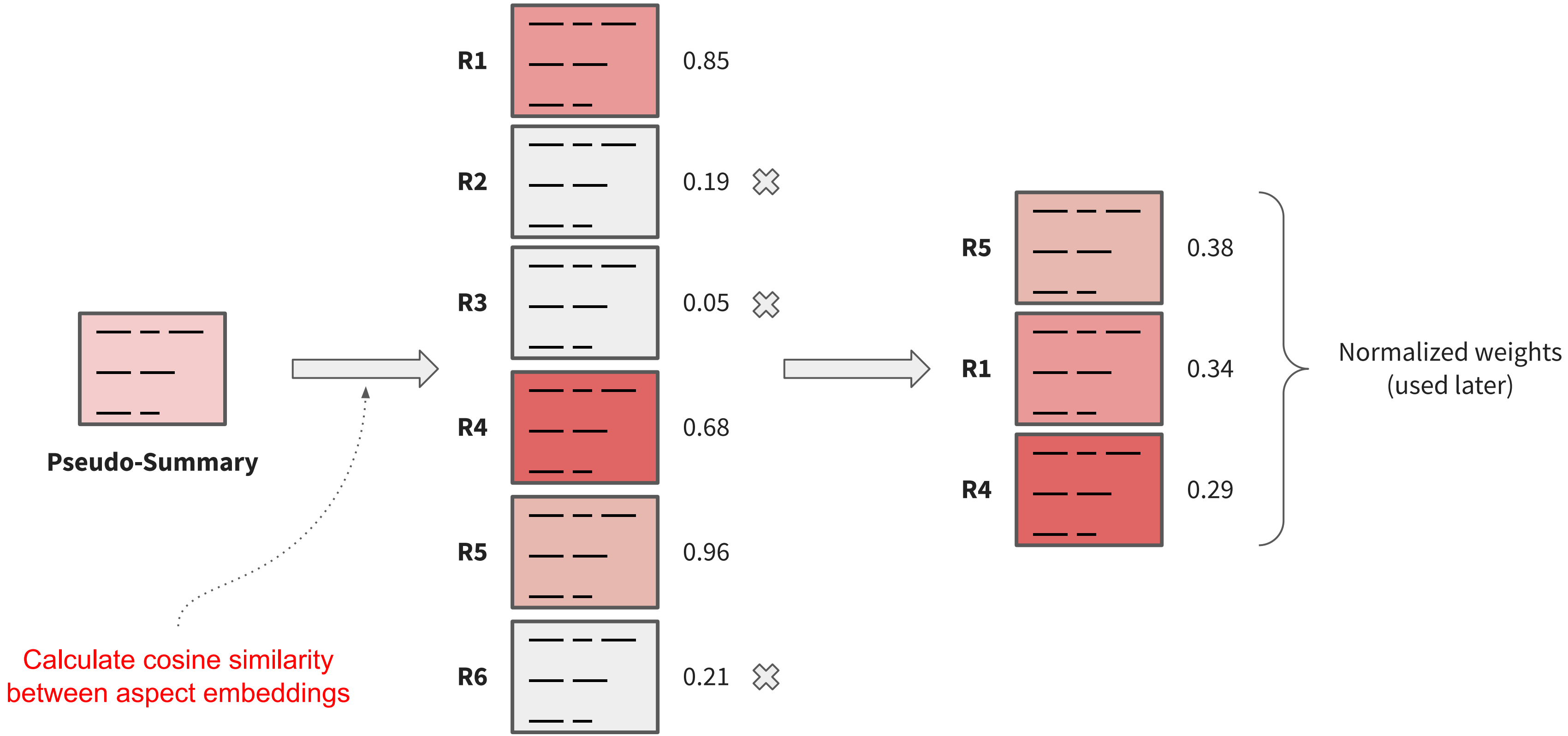
1. Review reconstruction component
   > learns aspect- and sentiment-specific embeddings
   > uses autoencoders with contrastive learning
2. Opinion summarization component
   > creates synthetic data using aspect-specific relevance-based sampling
   > uses relevance weights to aggregate multiple reviews

1. Wang, Ke, and Xiaojun Wan. "TransSum: Translating aspect and sentiment embeddings for self-supervised opinion summarization." In *ACL*, pp. 729-742. 2021.

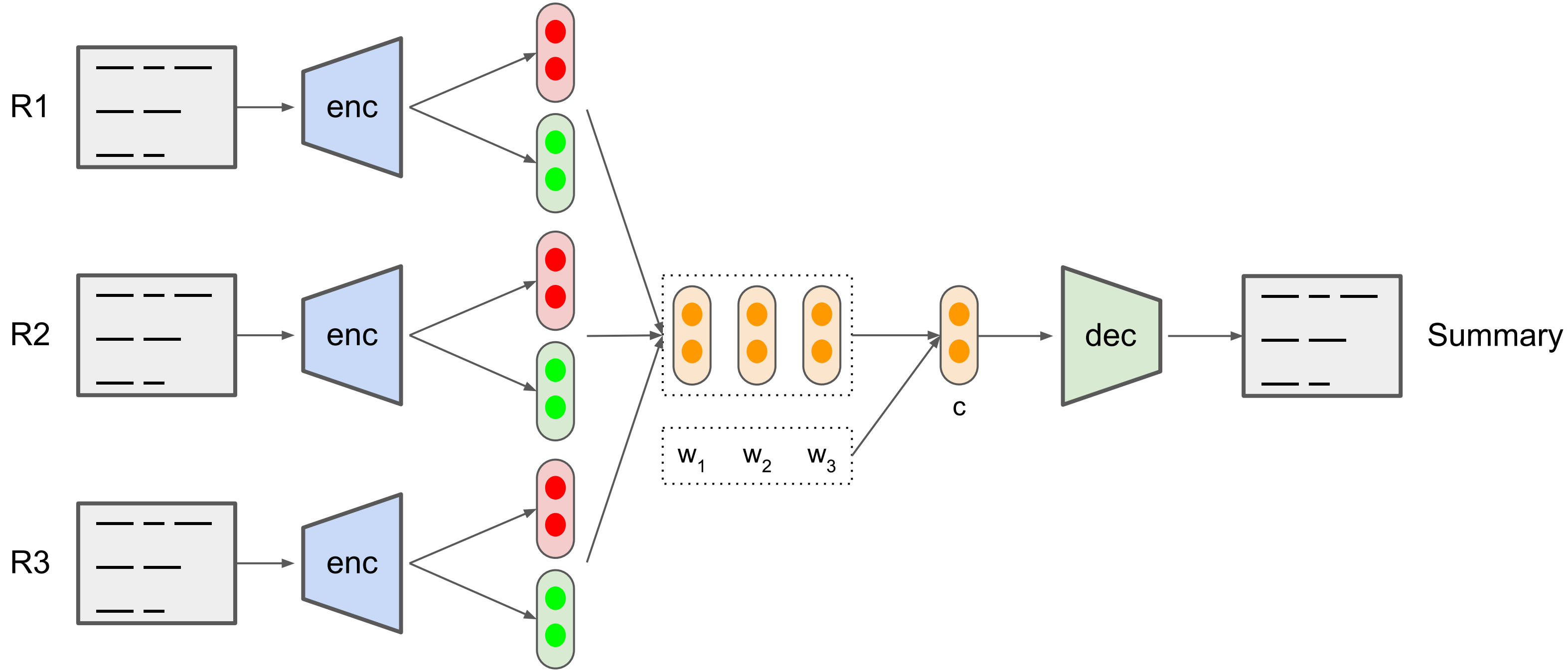
TransSum: Translation-based Review Modeling



TransSum: Dataset Creation



TransSum: Multi-input Opinion Summarization



Why could relevance-based sampling be suboptimal?

Burgers here are very delicious, but they were too expensive.

The waiter was rude to me. Too bad since the food was great…

I did not like the food here, and the staff as well! Not recommended

Worst food ever! There is also no parking and the location is bad.

These reviews exist in real-world setting!

Set of Reviews

Delicious food

Rude service

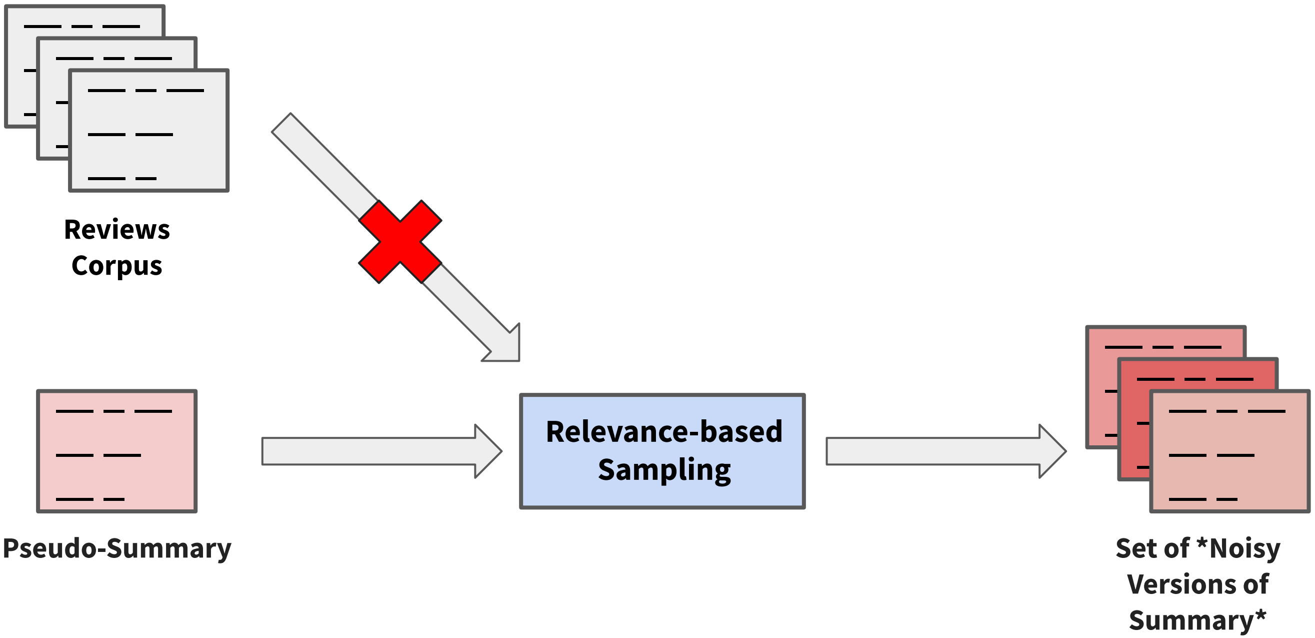
The food here is delicious, especially the burger. The service is awful, and the staff is rude.

Human-Written Summary

Summary

Review Sampling Method	Advantages	Disadvantages
Random Sampling	● Unlimited Training Data	● Encourages hallucination
Relevance-based Sampling	● Model has a better understanding of what to learn	● Does not capture real-world opinion variance in reviews
Review Noising	● Introduces phrase-level variation	● Encourages grammatical errors
Planned Sampling	● Can capture real-world opinion variance in reviews	● Planning stage may propagate errors

Review ~~Sampling~~ Noising



DenoiseSum[1]

- Treat opinion summarization as “denoising” reviews
  - Non-salient information in reviews are “noise” that needs to be denoised
- Create a synthetic dataset by introducing different noise to the reviews
  - Segment noising: Adding, removing, and replacing tokens and chunks
  - Document noising: Replacing the whole document entirely (≅ relevance-based sampling)
- Introduce a denoising module that explicitly corrects noised reviews at the encoding-level

1. Amplayo, Reinald Kim and Mirella Lapata. "Unsupervised Opinion Summarization with Noising and Denoising." In *ACL*, pp. 1934-1945. 2020.

DenoiseSum: Segment and Document Noising

the movie is a fun comedy with fine performances .

Token-level Noising

the **film** is a **nice** comedy with **stellar cast** .

Chunk-level Noising

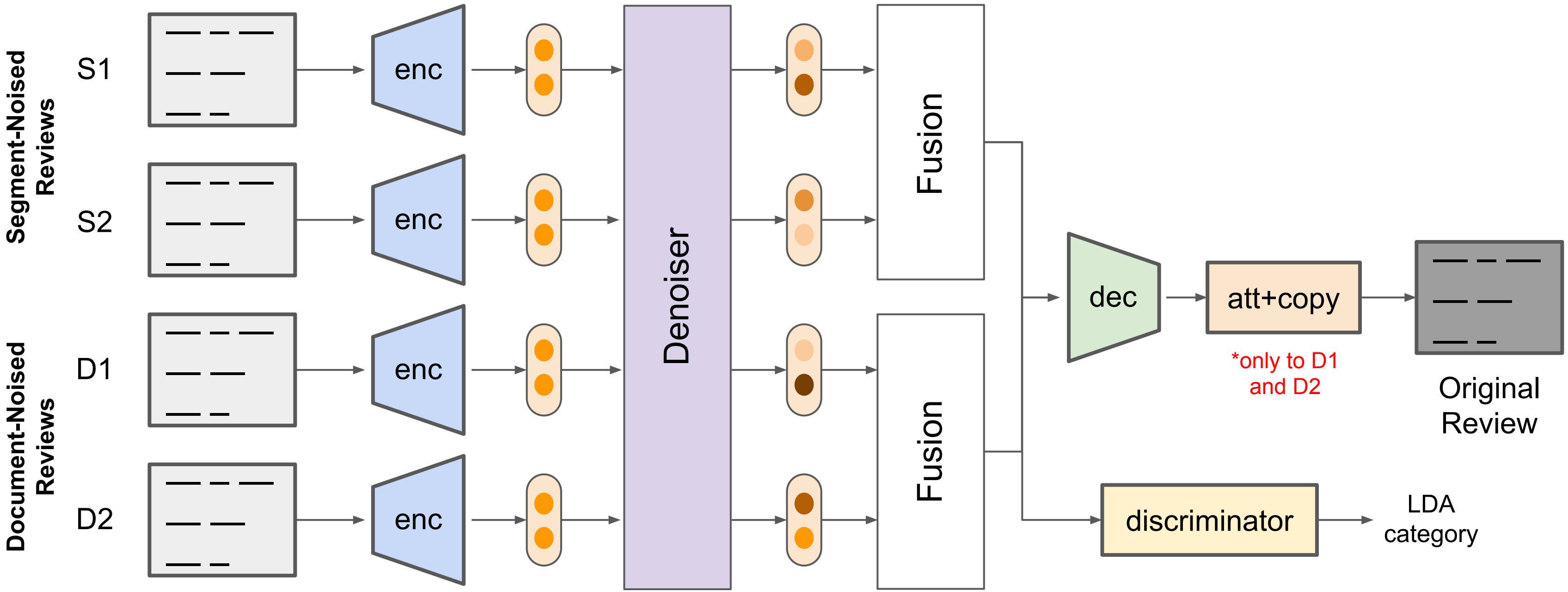
with **good production and zohan ,** a **nice** comedy **by** the **film** .

PP NP CC NP ,
NP PP NP .

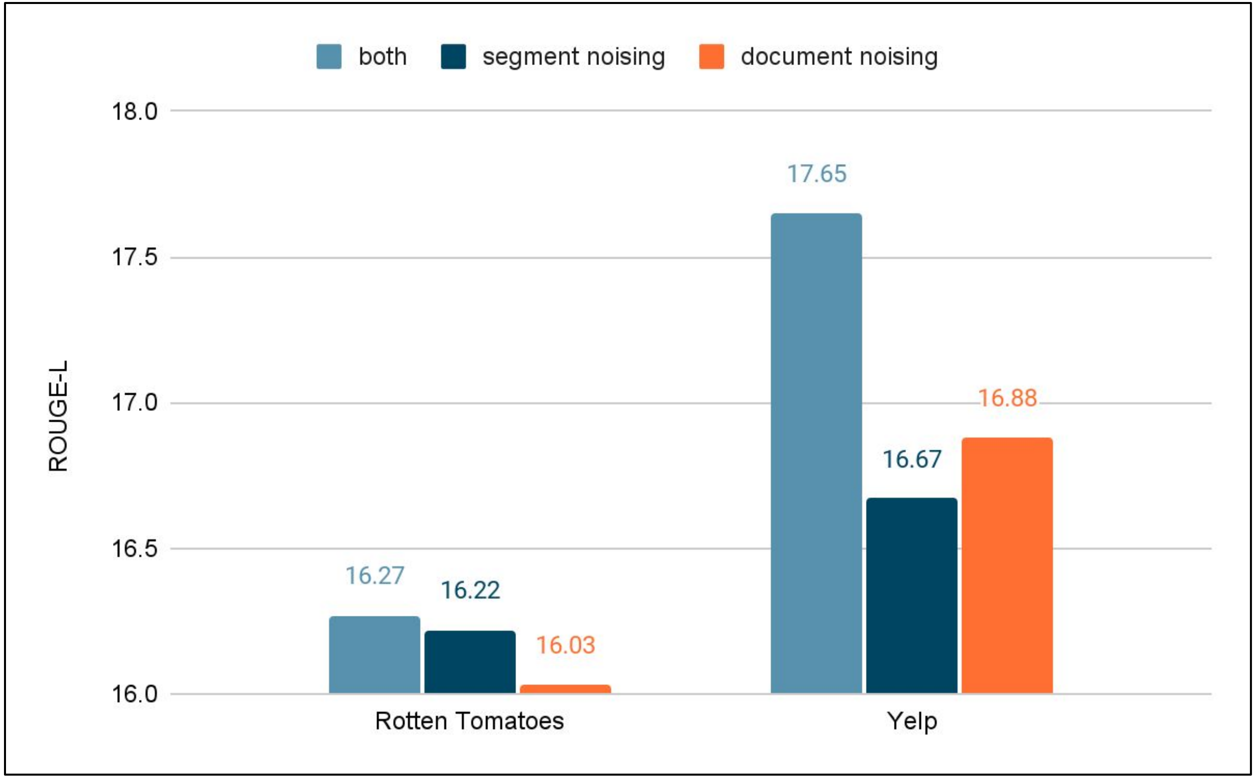
Document-level Noising = relevance-based sampling

the high-handed premise does not always work in zohan but you have to admire the chutzpah in trying it .	0.05
the fine performance of sandler as zohan in this very funny **comedy** makes this **movie** special .	**0.67**
the latest in a long line of underwhelming adam sandler **comedies** .	0.12

DenoiseSum: Summarization via Review Denoising



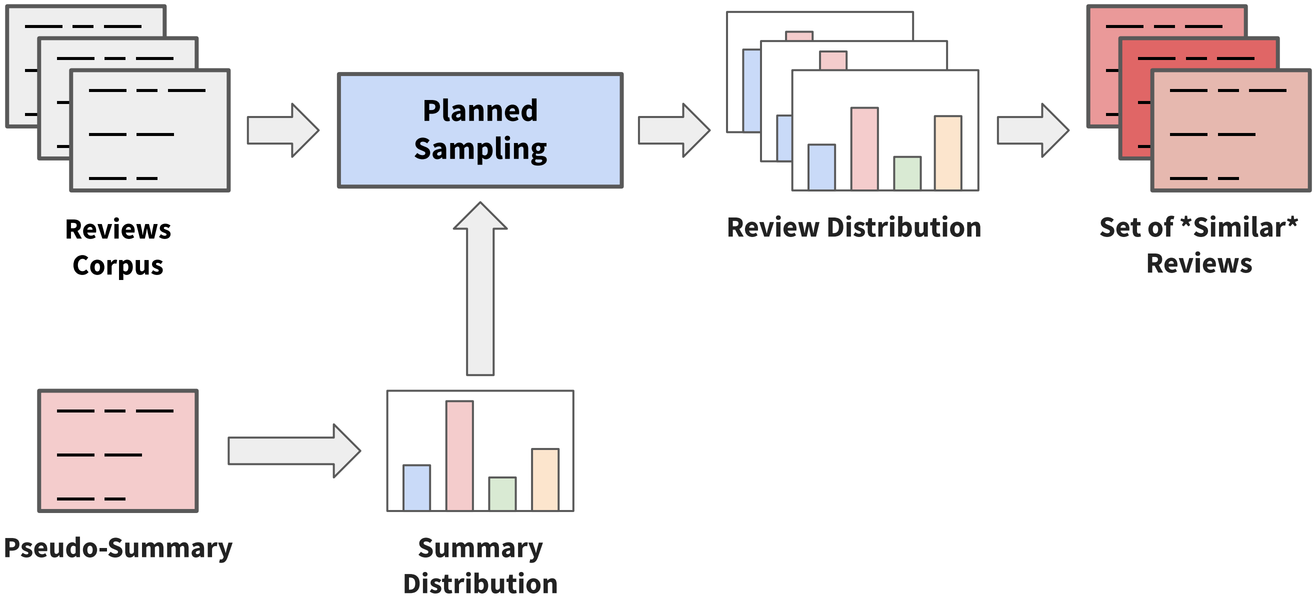
Segment vs Document Noising



Summary

Review Sampling Method	Advantages	Disadvantages
Random Sampling	• Unlimited Training Data	• Encourages hallucination
Relevance-based Sampling	• Model has a better understanding of what to learn	• Does not capture real-world opinion variance in reviews
Review Noising	• Introduces phrase-level variation	• Encourages grammatical errors
Planned Sampling	• Can capture real-world opinion variance in reviews	• Planning stage may propagate errors

Planned Review Sampling

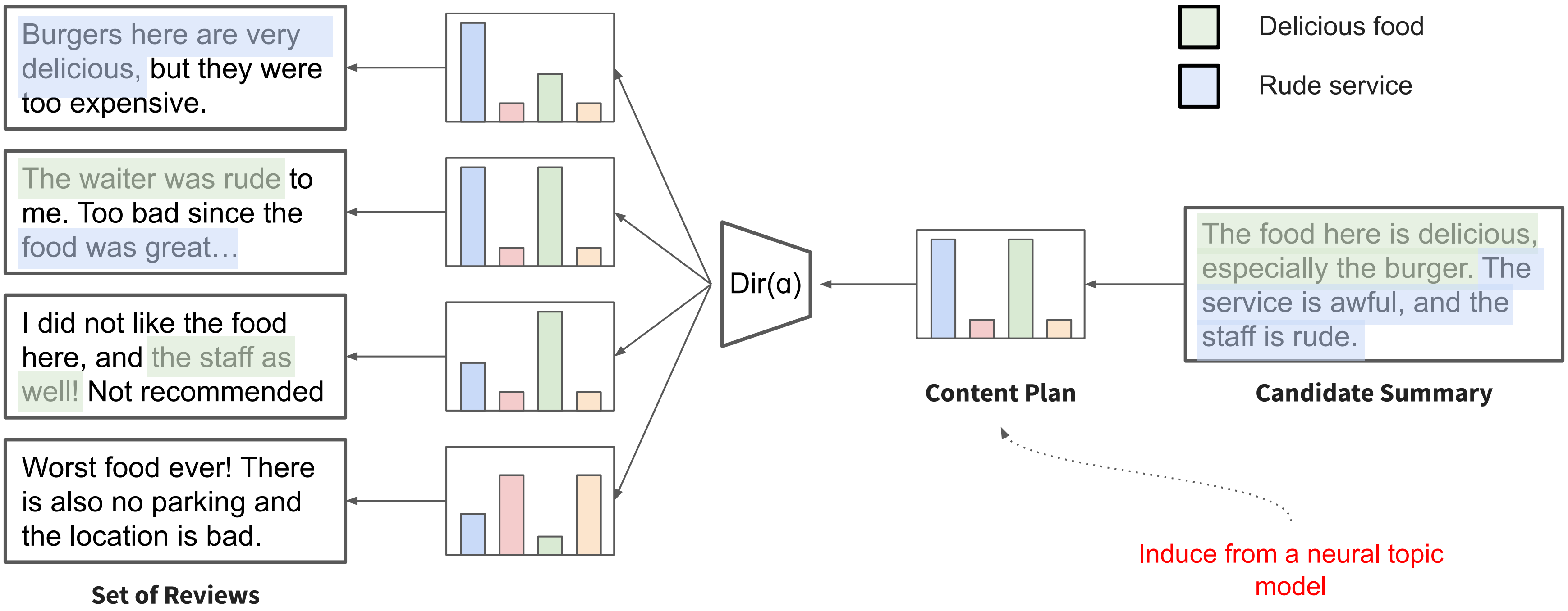


PlanSum[1]

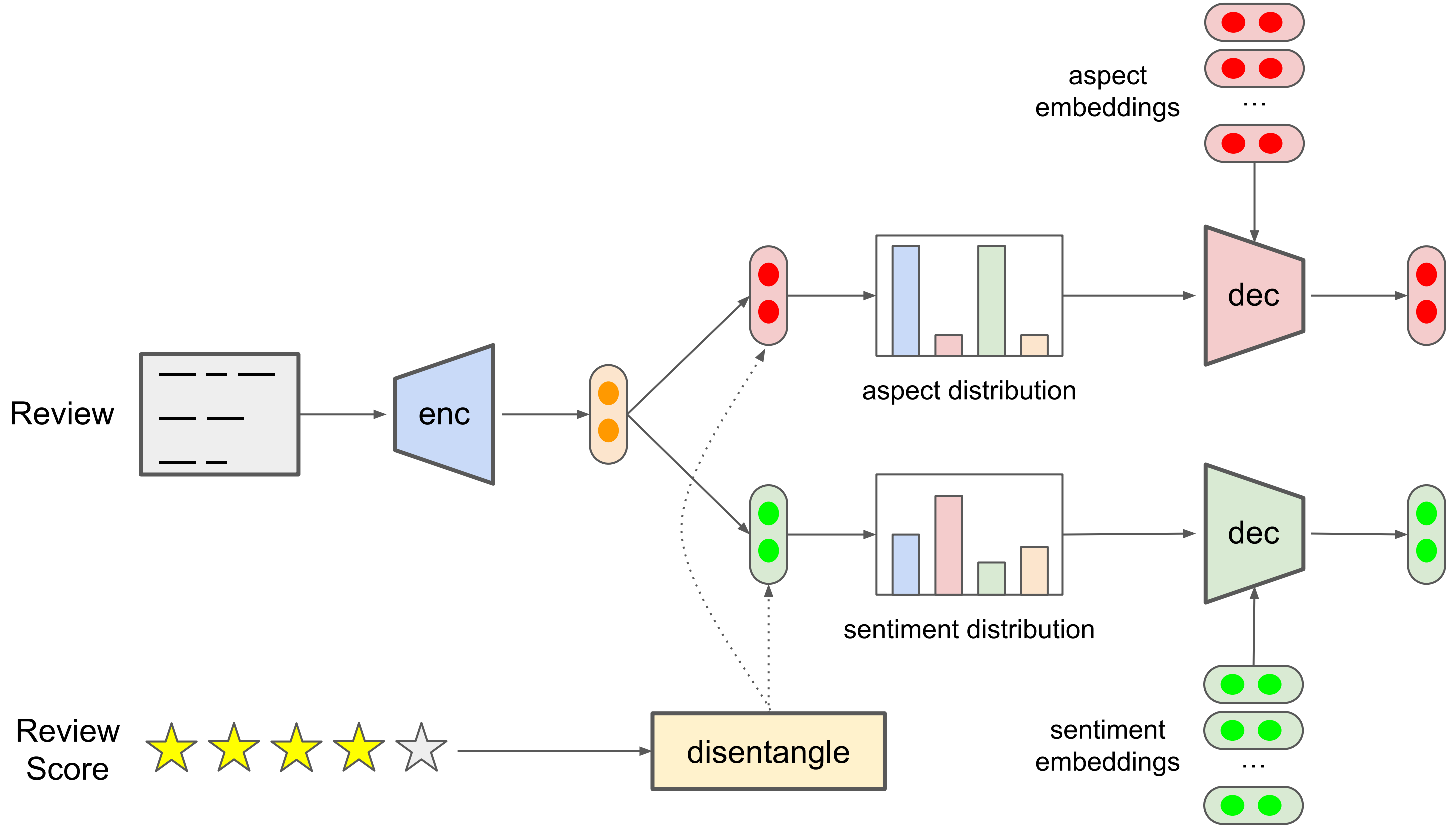
- Incorporate content planning[2] in opinion summarization
- Content plans are represented as aspect- and sentiment-specific distributions
- Content plans are used to create synthetic datasets with reviews that resemble real-world data
- The model also leverage content plans to guide/ground generation towards the right content

1. Amplayo, Reinald Kim, Stefanos Angelidis, and Mirella Lapata. "Unsupervised opinion summarization with content planning." In *AAAI*, pp. 12489-12497. 2021.
2. Kukich, Karen. "Design of a knowledge-based report generator." In *ACL*, pp. 145-150. 1983.

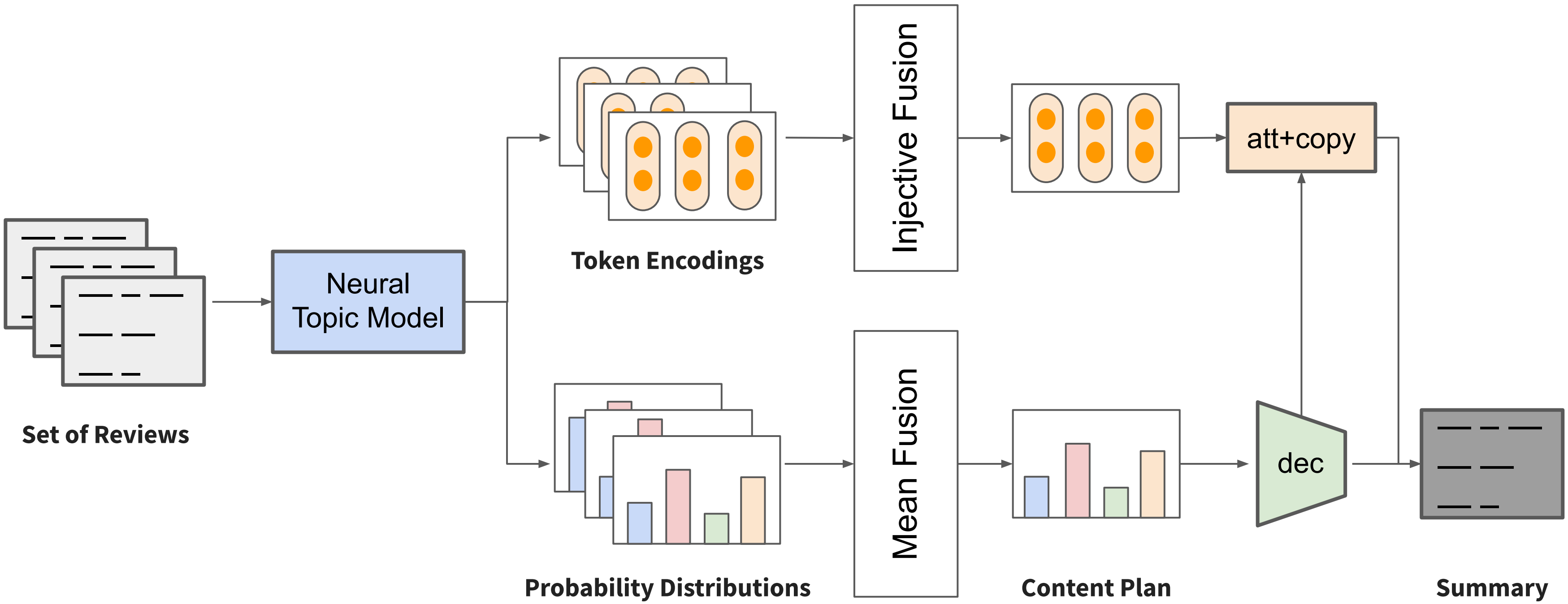
PlanSum: Sampling through Content Planning



PlanSum: Content Plan Induction



PlanSum: Summarization with Content Planning



Content planning results to better opinion variation

Gold:
If you're looking for a comfortable and inviting bar this is a great place to go. They have a lot of unique beers on tap that you will not find anywhere else. The staff here is extremely friendly, and after just a couple of minutes it feels like you are chatting with an old friend. The next time you want to head out for some drinks give them a shot.

Planned Sampling:
This is a great place to hang out with friends. The staff is very friendly and helpful. They have a lot of different beers to choose from and the beer selection is great. I'm not a big fan of beers but this place has some good selections. If you're in the mood for a beer and a fun atmosphere, this will be the place for you.

Random Sampling:
This is a great place to hang out with friends and family. The beer selection is great, and the atmosphere is very nice. I've been here a few times and have never had a bad experience. It's a fun place for a group of friends or groups.

Similarity Sampling:
This is a great place to go if you're in the area. It's a cool place for a night out, *but it is well worth it*. The atmosphere is great and the staff is always friendly. I'm not sure if I will go back.

Review Sampling: Summary

Review Sampling Method	Advantages	Disadvantages
Random Sampling	● Unlimited Training Data	● Encourages hallucination
Relevance-based Sampling	● Model has a better understanding of what to learn	● Does not capture real-world opinion variance in reviews
Review Noising	● Introduces phrase-level variation	● Encourages grammatical errors
Planned Sampling	● Can capture real-world opinion variance in reviews	● Planning stage may propagate errors

Takeaways and Future Work

Creating synthetic training data can be used for supervised learning

- Careful design of such dataset creation method is essential for the performance of the supervised model

Future work (and personal desires)

- Explore different (non-heuristic) methods to sample (and revise) reviews as candidate summaries
- Clear evaluation of dataset creation method, e.g. using the same model but trained on different synthetic datasets