Evaluation and Resources

Evaluation Approaches

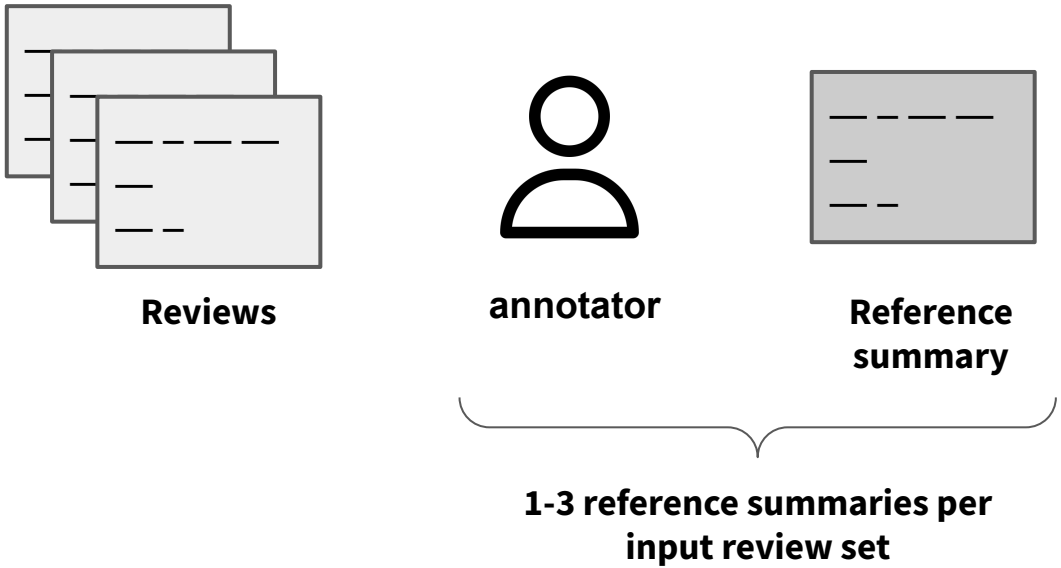
- 1) Automatic evaluation
  - Reference-based methods (e.g., ROUGE, BERTScore)
  - Classification-based methods (e.g., review rating classifier)
- 2) Human evaluation
  - Evaluation criteria
  - Best-Worst Scaling

Evaluation Approaches

- 1) Automatic evaluation
  - Reference-based methods (e.g., ROUGE, BERTScore)
  - Classification-based methods (e.g., review rating classifier)
- 2) Human evaluation
  - Evaluation criteria
  - Best-Worst Scaling

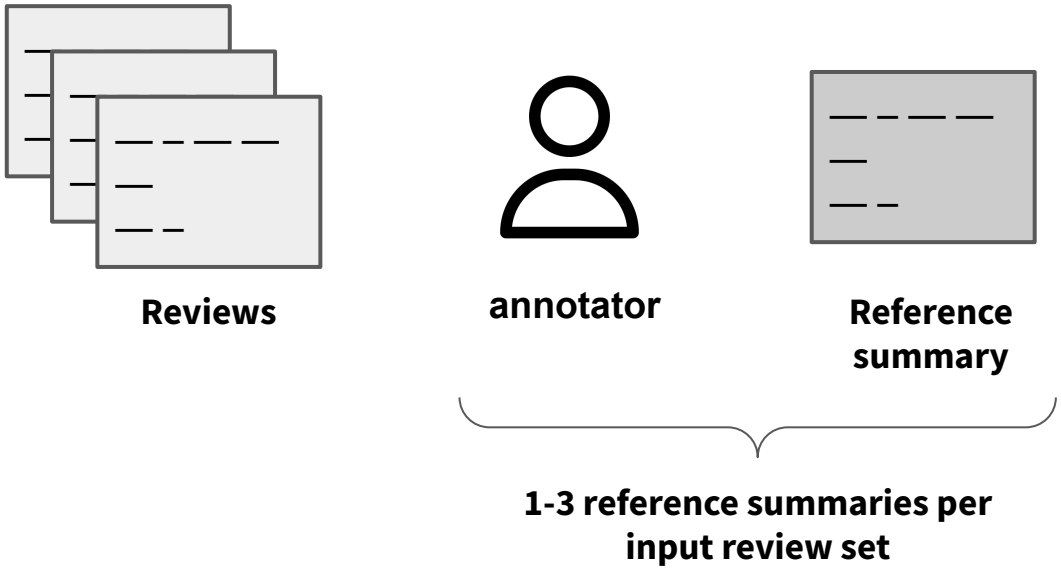
Collecting Reference Summaries

- Manual annotation by human workers



Collecting Reference Summaries

- Manual annotation by human workers



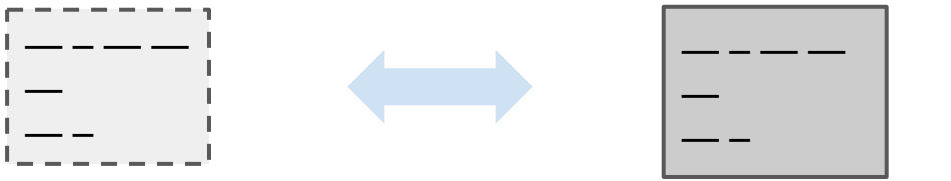
Opinion summarization benchmarks
= input review sets + reference summaries

Opinion Summarization Benchmarks

	# of ent.	# input reviews	# ref sum	Type	Review source
OpoSum	60	10	180	Extractive	Amazon
MeanSum	200	8	200	Abstractive	Yelp
CopyCat	60	8	180	Abstractive	Amazon, Yelp
FewSum	60	8	180	Abstractive	Amazon, Yelp
Space	50	100	1050	Abstractive + Aspect	TripAdvisor
AmaSum	31483	326	33324	Abstractive	Amazon

Reference-based Evaluation: ROUGE [Lin 2004]

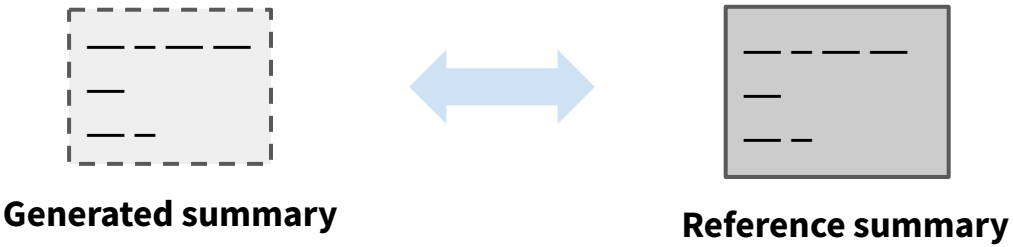
- **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation
- The de facto standard **set of metrics** for summarization
  - Tools includes a Python library rouge-score that replicates results by the official Perl script
- The metrics compare generated and reference summaries
  - ROUGE-N (N=1, 2) and ROUGE-L are commonly used for opinion summarization
    - ROUGE-N originally defined only Recall but F1 is commonly used



Chin-Yew Lin, “ROUGE: A Package for Automatic Evaluation of Summaries”, 2004

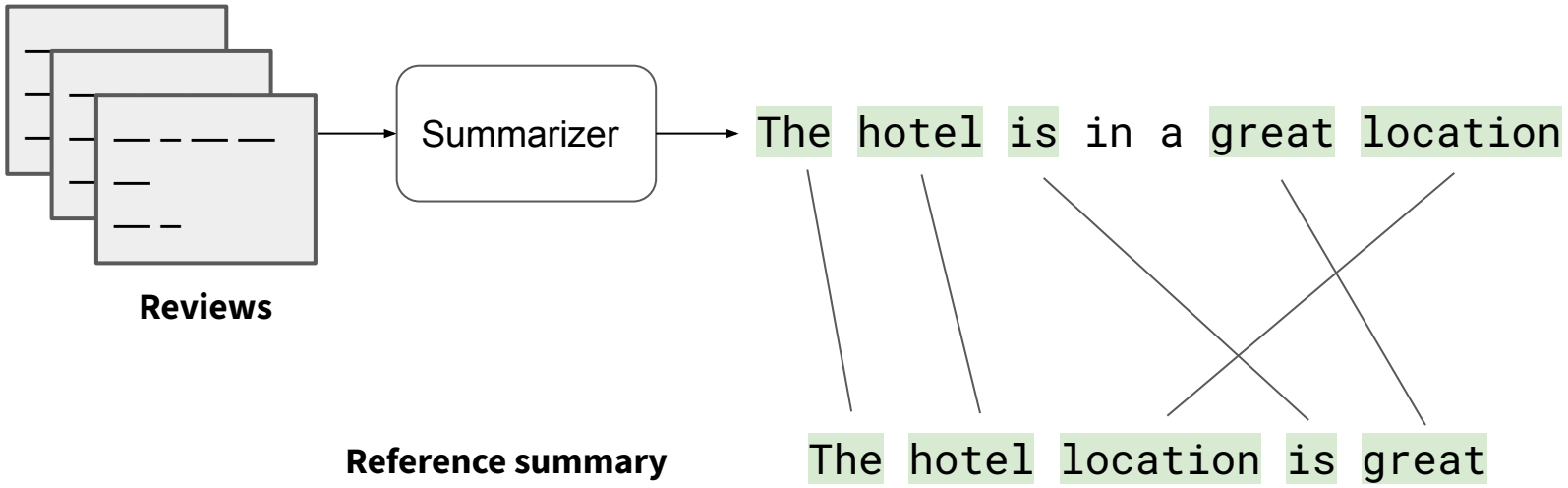
Reference-based Evaluation: ROUGE [Lin 2004]

- **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation
- ROUGE is **a set of metrics**
  - ROUGE-N/L/W/S
- F1 measure is a common choice for ROUGE-* metrics
  - ROUGE-N originally defined only Recall metric

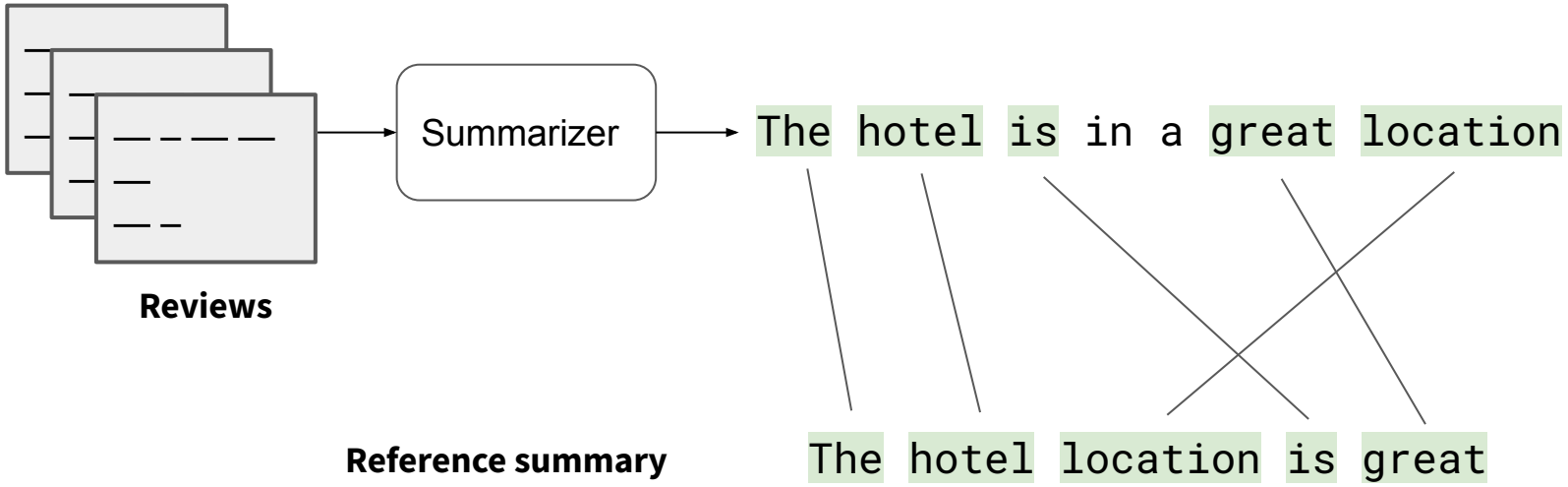


Chin-Yew Lin, “ROUGE: A Package for Automatic Evaluation of Summaries”, 2004

ROUGE-1: Unigram matching

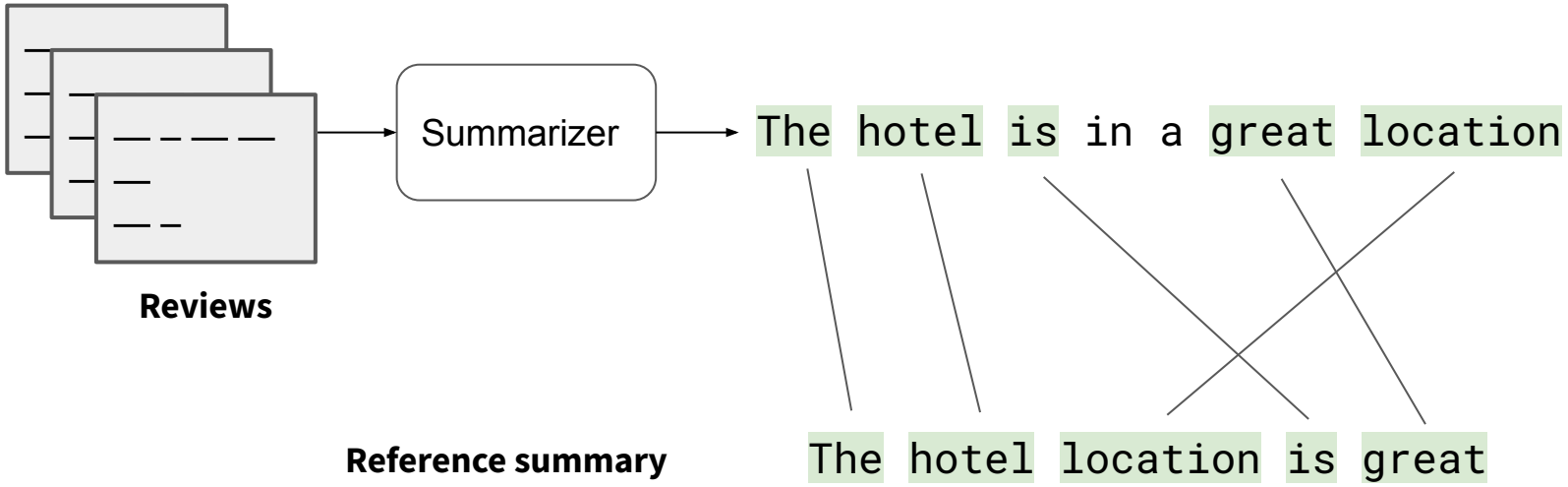


ROUGE-1: Unigram matching



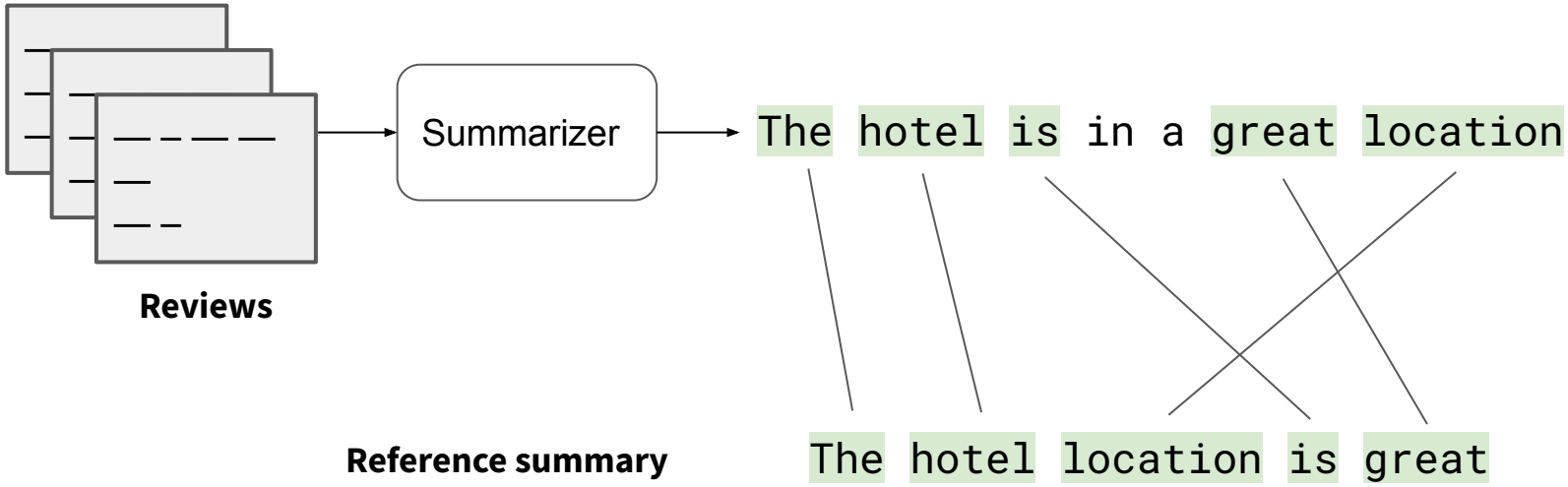
$$\text{Precision} = \frac{\text{\# of matched unigrams}}{\text{\# of total unigrams in generated summary}}$$

ROUGE-1: Unigram matching



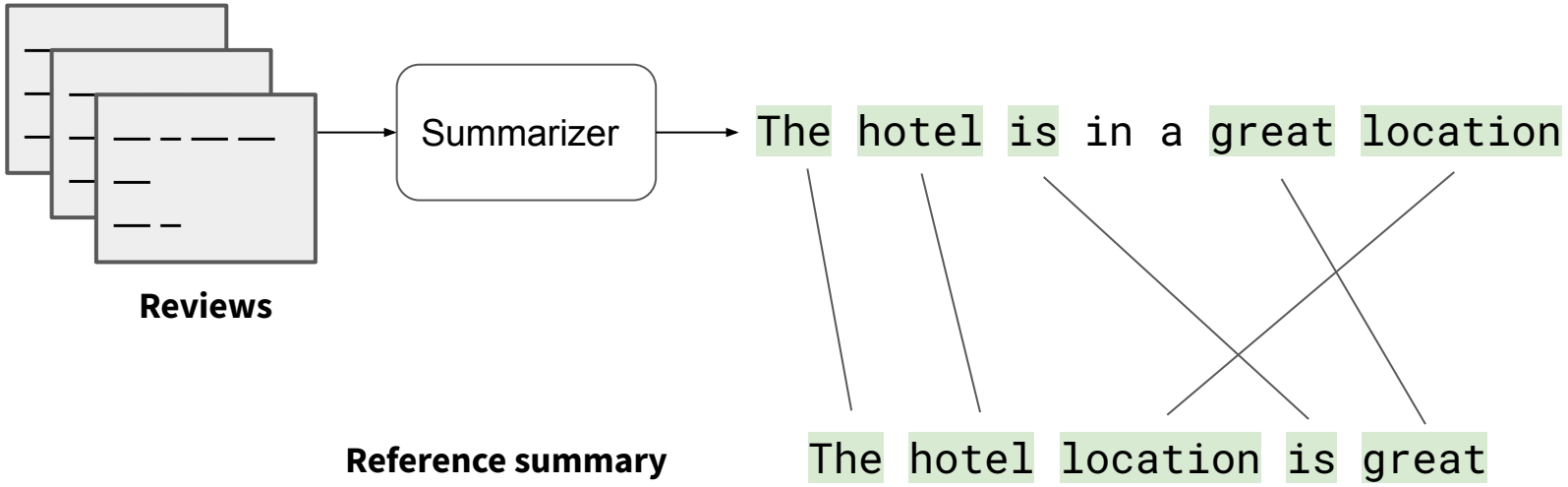
$$\text{Precision} = \frac{5}{7} = 0.71$$

ROUGE-1: Unigram matching



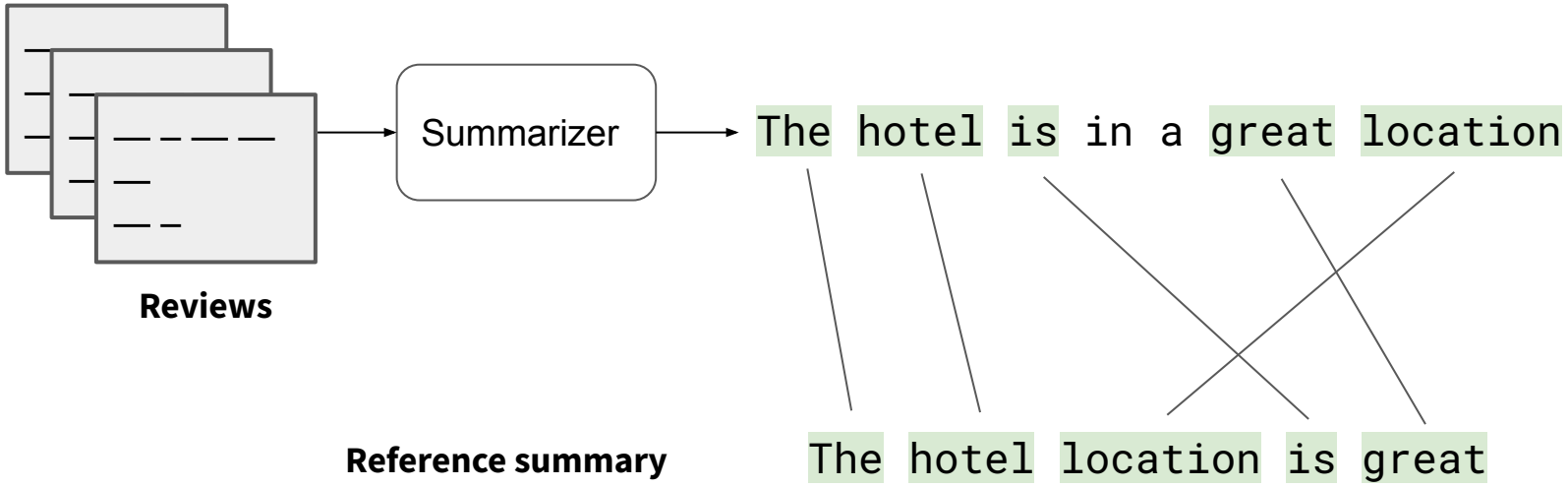
$$\text{Precision} = \frac{5}{7} = 0.71 \qquad \text{Recall} = \frac{\text{\# of matched unigrams}}{\text{\# of total unigrams in reference summary}}$$

ROUGE-1: Unigram matching



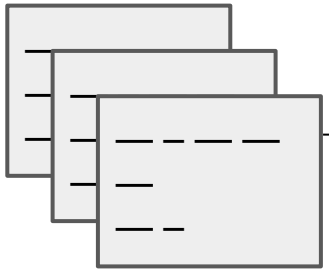
$$\text{Precision} = \frac{5}{7} = 0.71 \qquad \text{Recall} = \frac{5}{5} = 1$$

ROUGE-1: Unigram matching



$$\text{Precision} = \frac{5}{7} = 0.71 \qquad \text{Recall} = \frac{5}{5} = 1 \qquad \text{F1} = \frac{2PR}{P+R} = 0.83$$

ROUGE-2: Bigram matching



Reviews

Summarizer

The hotel is in a great location

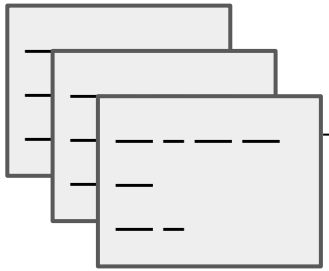
Reference summary

The hotel location is great

$$\text{Precision} = \frac{1}{6} = 0.17 \qquad \text{Recall} = \frac{1}{4} = 0.25 \qquad \text{F1} = 0.20$$

ROUGE-L: Longest common subsequence

(not substring, which only considers contiguous tokens)



Reviews

Summarizer

The hotel is in a great location

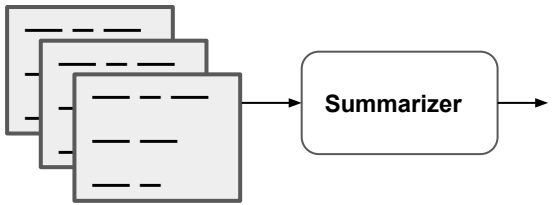
Reference summary

The hotel location is great

$$\text{Precision} = \frac{4}{7} = 0.57 \qquad \text{Recall} = \frac{4}{5} = 0.80 \qquad \text{F1} = 0.67$$

Limitations of ROUGE scores

- ROUGE scores are based on **token-level (exact) matching**
  - e.g., “great” != “excellent”



`The hotel is in an excellent location`

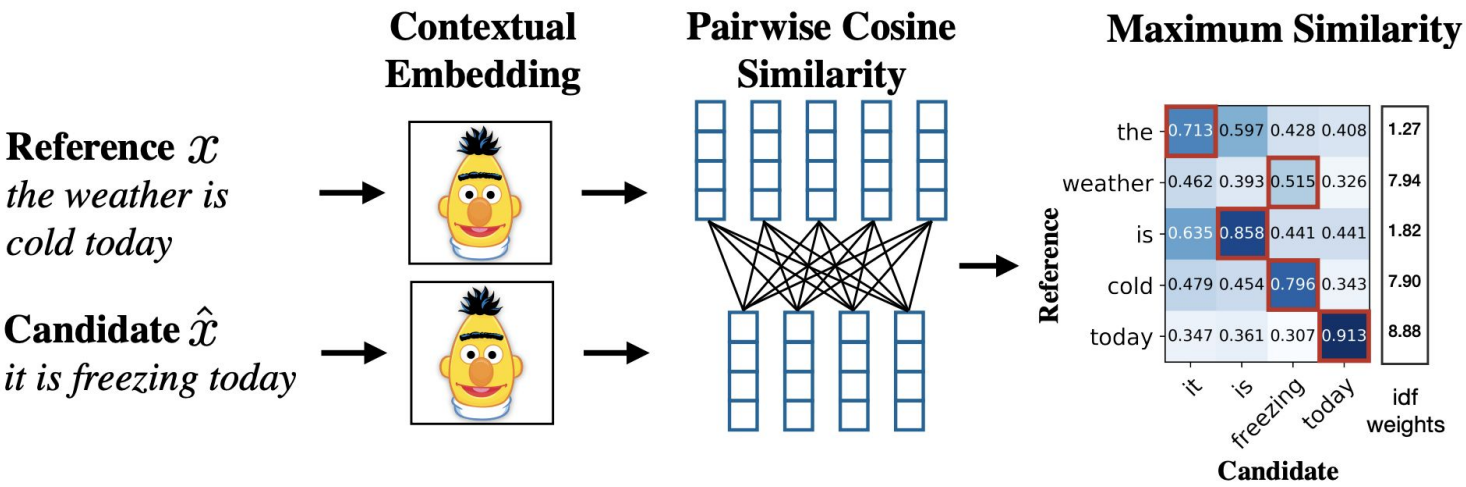
`The hotel is in a bad location`

Reference summary `The place is great`

Can we take into account semantic similarity into token matching?
(e.g., sim(“great”, “excellent”) > sim(“great”, “bad”))

BERTScore

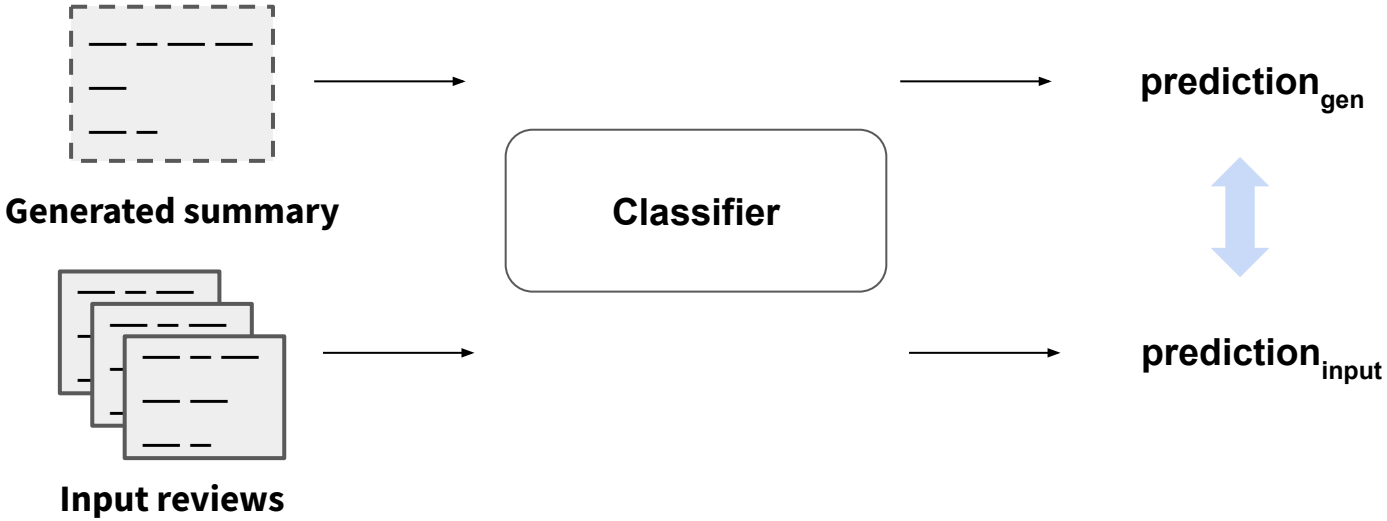
- Use BERT embeddings to calculate **semantic similarity** between reference summaries and generated summaries
  - The official Python library: [bert-score]



Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi, “BERTScore: Evaluating Text Generation with BERT”, ICLR 2020.

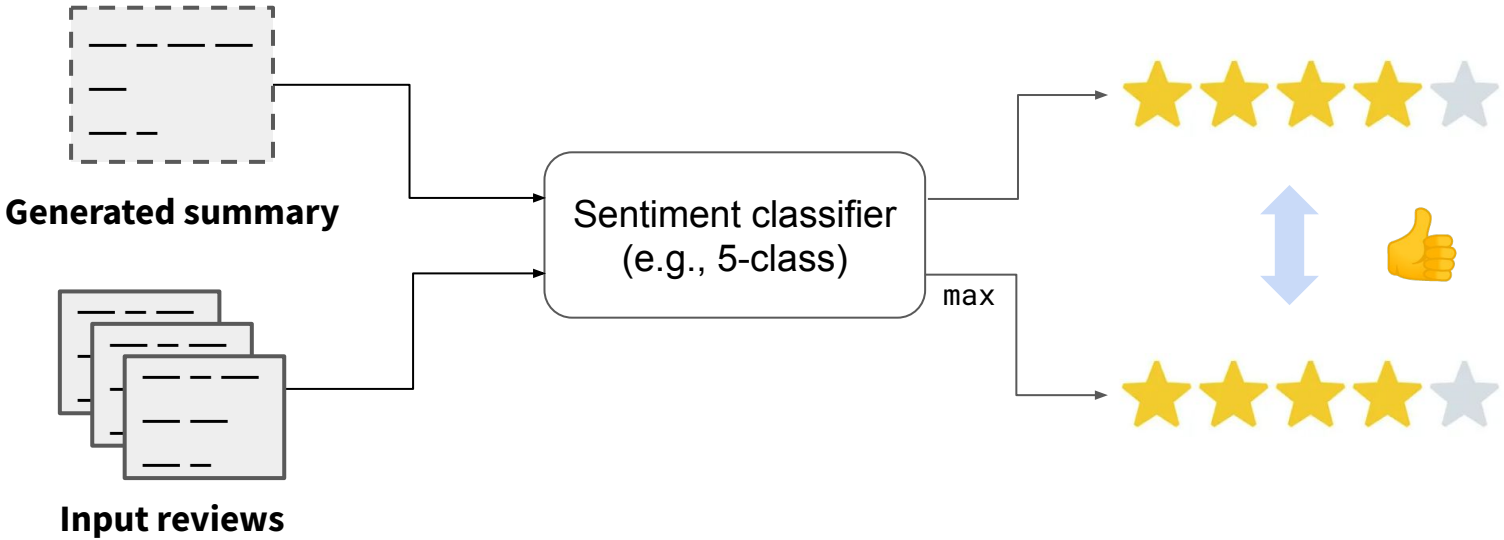
Classification-based evaluation

- Calculate **consistency** between generated summaries and the input reviews (or the reference summary) using **a classifier trained on another task**
  - Sentiment classification (e.g., MeanSum, SelfSum)
  - Aspect-category classification (e.g., SelfSum, QT, LSARS)



Sentiment Accuracy [Chu and Liu 2020][Elsahar et al. 2021]

- Sentiment consistency
  - Generated summary vs Input reviews (reference summary can be also used)



Other options include aspect-category classifiers [Elsahar et al. 2021][Angelidis et al. 2021]

Automatic Evaluation: Summary & Pros/Cons

- 1) Reference-based evaluation
  - ROUGE 1/2/L and BERTScore are commonly used automatic evaluation
- 2) Classification-based evaluation
  - Aspect-based Sentiment Analysis models to evaluate the consistency b/w generated summaries and the input reviews

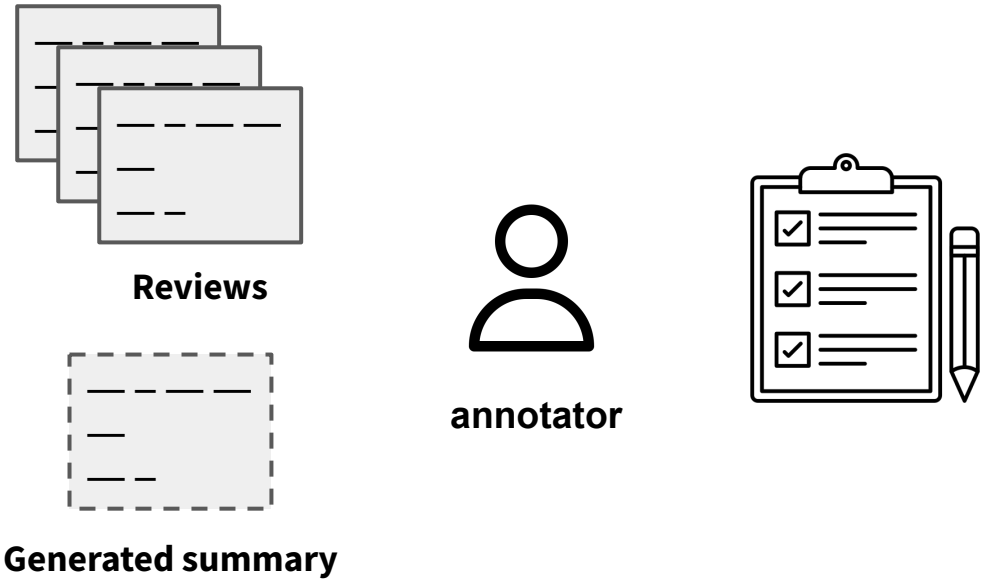
- Pros 👍
  - Reproducible
- Cons 👎
  - Evaluation heavily relies on the quality of reference summaries/classifiers

Evaluation Approaches

- 1) Automatic evaluation
  - Reference-based methods (e.g., ROUGE, BERTScore)
  - Classification-based methods (e.g., review rating classifier)
- 2) Human evaluation
  - Evaluation criteria
  - Best-Worst Scaling

Human Evaluation

- Ask human annotators to assess specific performance characteristics
  - Informativeness, Coherence, Redundancy, Content Support



Commonly-used Evaluation Criteria

- Informativeness
- Fluency
- Conciseness
- Coherence
- (Non-)Redundancy

Limitations of Multi-point Rating Scale

- Rating scales may not be discriminating. Every method can get 5 out of 5
- Different annotators use scale differently

	1	2	3	4	5
Method A					✓
Method B					✓
Method C					✓
Method D					✓

	1	2	3	4	5
Method A			✓		
Method B		✓			
Method C			✓		
Method D			✓		

Limitations of Multi-point Rating Scale

- Rating scales may not be discriminating. Every method can get 5 out of 5
- Different annotators use scale differently

	1	2	3	4	5
Method A					✓
Method B					
Method C					
Method D					✓

	1	2	3	4	5
Method A			✓		
Method B					
Method C					
Method D			✓		

Pairwise comparisons should address the issues,
but it requires nCk (k=2) judgements :(

Best-Worst Scaling

- Effective pairwise annotation schema without directly asking pairwise judgements
- With Best-Worst scaling, the annotator just chooses **the best and worst methods**

	Best	Worst
Method A	✓	
Method B		
Method C		
Method D		✓

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. Best-worst scaling: Theory, methods and applications. Cambridge University Press. 2015.

Best-Worst Scaling

- Effective pairwise annotation schema without directly asking pairwise judgements
- With Best-Worst scaling, the annotator just chooses **the best and worst methods**

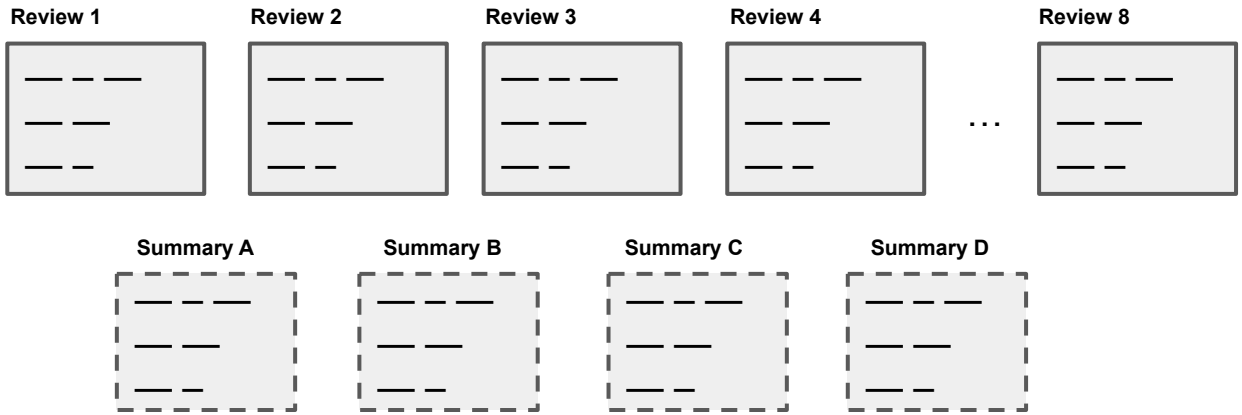
	Best	Worst
Method A	✓	
Method B		
Method C		
Method D		✓

Method A > Method B
Method A > Method C
Method A > Method D
Method B ? Method C
Method B > Method D
Method C > Method D

5 pairwise judgements from 2 annotations :)

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. Best-worst scaling: Theory, methods and applications. Cambridge University Press. 2015.

Best-Worst Scaling Task Example



Task: Read those summaries carefully and select the best and worst one for each of the following criteria:

Informativeness:

Best: ○ A ○ B ○ C ○ D

Worst: ○ A ○ B ○ C ○ D

Coherence:

Best: ○ A ○ B ○ C ○ D

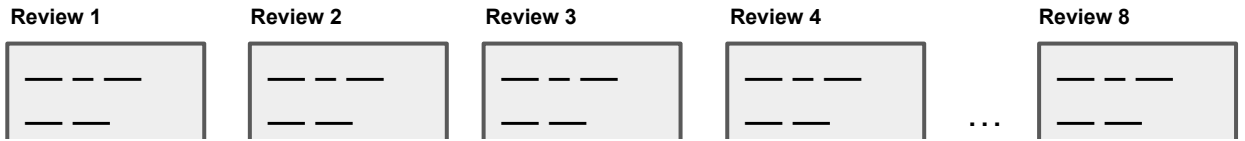
Worst: ○ A ○ B ○ C ○ D

Non-redundancy:

Best: ○ A ○ B ○ C ○ D

Worst: ○ A ○ B ○ C ○ D

Best-Worst Scaling Task Example



	Inform.	Coherent	Concise	Redund.
Centroid	+36.0	−57.3	−60.7	−12.7
LexRank	−52.7	−38.0	−44.7	−1.3
MeanSum	−23.3	+26.7	+28.7	+3.3
Copycat	−10.7	+**34.7**	+38.0	−3.3
QT	+**50.7***	+34.0†	+**38.7**†	+**18.0***

Best: ○ A ○ B ○ C ○ D
Worst: ○ A ○ B ○ C ○ D

Best: ○ A ○ B ○ C ○ D
Worst: ○ A ○ B ○ C ○ D

Best: ○ A ○ B ○ C ○ D
Worst: ○ A ○ B ○ C ○ D

Best-Worst Scaling: Count Analysis

- A simple method that converts Best-Worst scaling annotations into individual scores

(% of times method X was selected as *best*) − (% of times method X was selected as *worst*)

	Best	Worst
Method A	✓	
Method B		
Method C		
Method D		✓

	Best	Worst
Method A	✓	
Method B		
Method C		✓
Method D		

	Best	Worst
Method A		
Method B	✓	
Method C		
Method D		✓

	Best	Worst
Method A	✓	
Method B		
Method C		
Method D		✓

Bryan Orme, MaxDiff Analysis: Simple Counting, Individual-Level Logit, and HB, Sawtooth Software

Best-Worst Scaling: Count Analysis

- A simple method that converts Best-Worst scaling annotations into individual scores

(% of times method X was selected as *best*) − (% of times method X was selected as *worst*)

	Best	Worst
Method A	✓	
Method B		
Method C		
Method D		✓

	Best	Worst
Method A	✓	
Method B		
Method C		✓
Method D		

	Best	Worst
Method A		
Method B	✓	
Method C		
Method D		✓

	Best	Worst
Method A	✓	
Method B		
Method C		
Method D		✓

	Best	**Worst**
Method A	**75%**	**0%**
Method B	**25%**	**0%**
Method C	**0%**	**25%**
Method D	**0%**	**75%**


```
Method A:  0.75
Method B:  0.25
Method C: -0.25
Method D: -0.75
```

Human Evaluation: Summary and Pros/Cons

- Human evaluation
  - Informativeness, Fluency, Coherence, Redundancy
  - Best-Worst scaling

- Pros 👍
  - More detailed assessment than automatic evaluation
- Cons 👎
  - Relies on annotators' subjective judgments
  - Expensive (Time & $$$)

Error/Qualitative Analysis

Reviews
1. Birdsong is a gem. A true gem! I was over at noda and wandered back and around to birdsong. The staff were very friendly and I found the bar a bit like home. *They have a great outdoor area* and, most importantly, their beer is quality. I'm generally not a fan of flavored beers. Ipa por vida! But! Their jalapeno pale ale!? Hello deliciousness. Seriously. Give it a try.
2. Great beer to try! Fun flavors like jalapeno pale ale. The staff inside is nice and friendly. *I was able to get a t-shirt with no hassle at all. The outdoor seating area is wonderful. Birdsong is next door to noda*, so you should definitely check it out!
3. Had the extra pale ale and loved it. In fact I loved everything about this place. The vibe was ideal for a long night of serious causal drinking. From *the peanuts on the table* to the friendly bartenders, this place just felt homey as soon as you sat on a stool. But unlike other dive this bar has delicious beer and an a chill atmosphere that really makes the beer go down quick and easy. I am looking forward to visiting again!
4. This is a hiden gem.... Reminds me of Asheville, nc nice happy laid bk plp and great beer. The jalepeno pale ale was amazen..... It drove my sinses in overload. The smell and taste wrk great for it, you have got to try!
5. Jalapeno pale ale.... maybe a little crazy.... but so good. I have been going to birdsong since they first opened. I have always enjoyed their free will. They have made a couple new brews as of late that I sampled and all are really good. *I love that they are right across the way from noda brewery* and tend to always go to both of them during my visits. *I love the games and the free peanuts. For the quality of the beer, I feel the prices are really good.* Hoping to see some additional brews in the future.
6. We thought that we would give this place a shot considering the positive reviews. However our visit was all but positive. First of all *I don't believe that the air conditioning was working*. I know that in the past some places do this to encourage the beer consumption. I don't necessarily have a problem with this but if thats the case at least make sure that the beer is cold. I spoke to someone up at bar and she basically said that because the beer is brewed there its not served cold but just slightly colder than room temperature. If i wanted a warm pint of bitter I'll go to the U.K. needless to say we won't be back.
7. *Liked the outdoor space* and the beer is really good. *Wish they had more seating*
8. Great atmosphere! Was a band on the saturday night we were there that was excellent! The beers were wonderful as well, would be back!

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!
BERTCENT	*Had the extra pale ale and loved it.* In fact I loved everything about this place. The vibe was ideal for a long night of serious causal drinking. From *the peanuts on the table* to the friendly bartenders, this place just felt homey as soon as you sat on a stool. But unlike other dive *this bar has delicious beer* and an a chill atmosphere that really makes the beer go down quick and easy. I am looking forward to visiting again!
DENOISESUM	Great beer and *most of the outdoor seating is wonderful.* The beer is really good and the staff was friendly. I wish found this place to go out of my way to go to.
COPYCAT	This place is awesome! I had the jalapeno pale ale and it was delicious. *It was a little pricey but worth the trip.* The staff was friendly and helpful as well. I will be back!
RANDOM	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	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.
NOPLAN	This is a great place to hang out with friends. The staff is very friendly and the beer selection is great. I've had a couple of beers and they have a good selection of beer and beer. *It's a little pricey but it is worth the wait.*
PLANSUM	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.

atmosphere, staff, beers, and *others*

Summary

- Automatic evaluation
  - 1) Reference-based metrics: ROUGE and BERTScore
  - 2) Classification-based metrics: Rating/aspect-category classification
- Human evaluation
  - Evaluation criteria
  - Best-Worst Scaling
- Error/Qualitative analysis

No single evaluation metric is perfect!

Evaluation should be comprehensive:
automatic evaluation + human evaluation + qualitative analysis

Challenges and Opportunities