BLEU Neighbors:
A Reference-less Approach to Automatic Evaluation

Eval4NLP @ EMNLP 2020

Kawin Ethayarajh

Stanford NLP

Dorsa Sadigh

iliad

In many applications, reference-based automatic metrics (e.g., BLEU) can't be used or are less than ideal.

- In dialogue, there are many valid responses but only a few are given as references.
- In open-ended NLG (e.g., with a language model), there are no references at all.
- This necessitates the human evaluation of quality, which is slow and expensive.

[Lowe et al., 2017]

Reference-less Evaluation

- **Tired**: Reference-based automatic evaluation
- **Wired**: Human evaluation (e.g., Mechanical Turk)
- **Inspired:** An automatic reference-*less* evaluation metric for language quality that is fast, simple, and correlates well with human judgment.

Reference-less Evaluation

- **Tired**: Reference-based automatic evaluation
- **Wired**: Human evaluation (e.g., Mechanical Turk)
- **Inspired:** An automatic reference-*less* evaluation metric for language quality that is fast, simple, and correlates well with human judgment.
    - We want to complement, not supplant, humans. BLEU speeds up translation model development; we want to speed up NLG model development.

Reference-less Evaluation … is harder than it looks.

- Heuristic-based evaluation has a narrow scope (e.g., grammar correction).
- Fluency (e.g., log-odds of output) only captures one facet of language quality.
- Trained models (e.g., ADEM) generalize poorly and exploit annotation artefacts.

[Napoles et al., 2016; Kann et al., 2018; Lowe et al., 2017; Lowe, 2019]

Reference-less Evaluation ... is harder than it looks.

- Heuristic-based evaluation has a narrow scope (e.g., grammar correction).
- Fluency (e.g., log-odds of output) only captures one facet of language quality.
- Trained models (e.g., ADEM) generalize poorly and exploit annotation artefacts.
- **Idea:** Don't be too ambitious; don't try to score the unscorable.

[Napoles et al., 2016; Kann et al., 2018; Lowe et al., 2017; Lowe, 2019]

BLEU Neighbors

How can we estimate the quality of x given human-scored data S (not references)?

Non-unigram BLEU:

$$\text{BLEU}^*(x, s) = \beta \cdot \prod_{i=2}^{4} P_i(x, s)^{1/3}$$

Find neighbors of x:

$$\mathcal{N} = \{s \in S \mid \text{BLEU}^*(x, s) \geq \tau\}$$

Estimate quality $q(x)$:

$$\widehat{q}(x) = \begin{cases} \frac{1}{|\mathcal{N}|} \sum_{s \in \mathcal{N}} q(s) & a \leq |\mathcal{N}| \leq b|S| \\ \text{undefined} & \text{otherwise} \end{cases}$$

BLEU Neighbors

How can we estimate the quality of x given human-scored data S (not references)?

Non-unigram BLEU:

$$\mathrm{BLEU}^*(x, s) = \beta \cdot \prod_{i=2}^{4} P_i(x, s)^{1/3}$$

Find neighbors of x:

$$\mathcal{N} = \{s \in S \mid \mathrm{BLEU}^*(x, s) \geq \tau\}$$

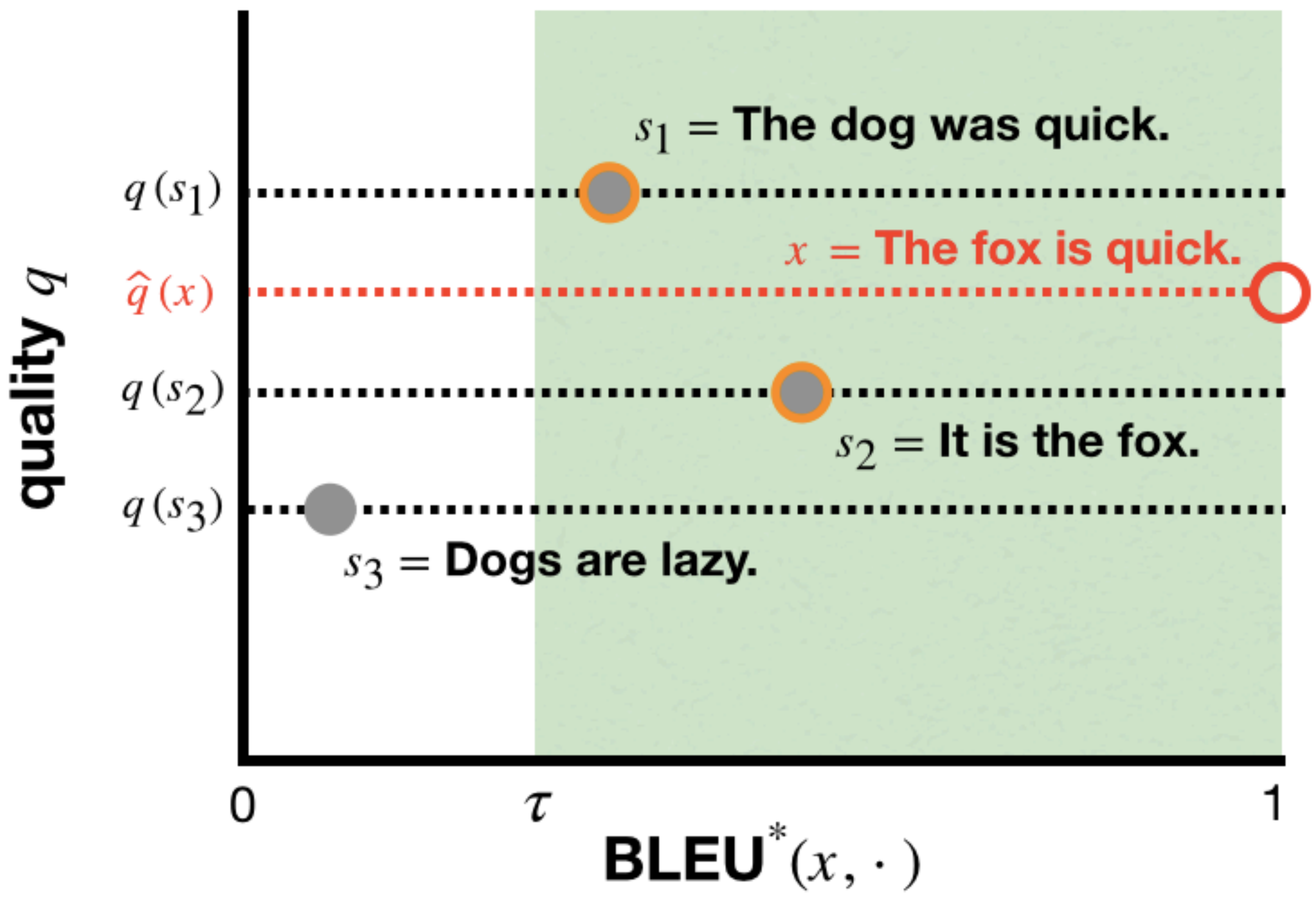
Estimate quality $q(x)$:

$$\widehat{q}(x) = \begin{cases} \frac{1}{|\mathcal{N}|} \sum_{s \in \mathcal{N}} q(s) & a \leq |\mathcal{N}| \leq b|S| \\ \text{undefined} & \text{otherwise} \end{cases}$$

In practice, similarity threshold $\tau = 0.08$, minimum neighbors $a = 5$, maximum frequency of neighbors $b = 0.66$ are near-optimal for all tasks.

BLEU Neighbors

$$q(x) = ? \quad \hat{q}(x) = \frac{1}{2}\left(q(s_1) + q(s_2)\right)$$



How to evaluate the evaluation metric?

- output for three tasks: open-ended NLG, chitchat dialogue, summarization
  - How well do our estimates correlate with the ground-truth quality (mean human judgment over 20 annotators)?
  - How much of the data can we make predictions for (i.e., *coverage*)?
  - What if we used ROUGE/METEOR/BERTScore instead of BLEU?

[Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005; Zhang et al., 2019; Hashimoto et al., 2019]

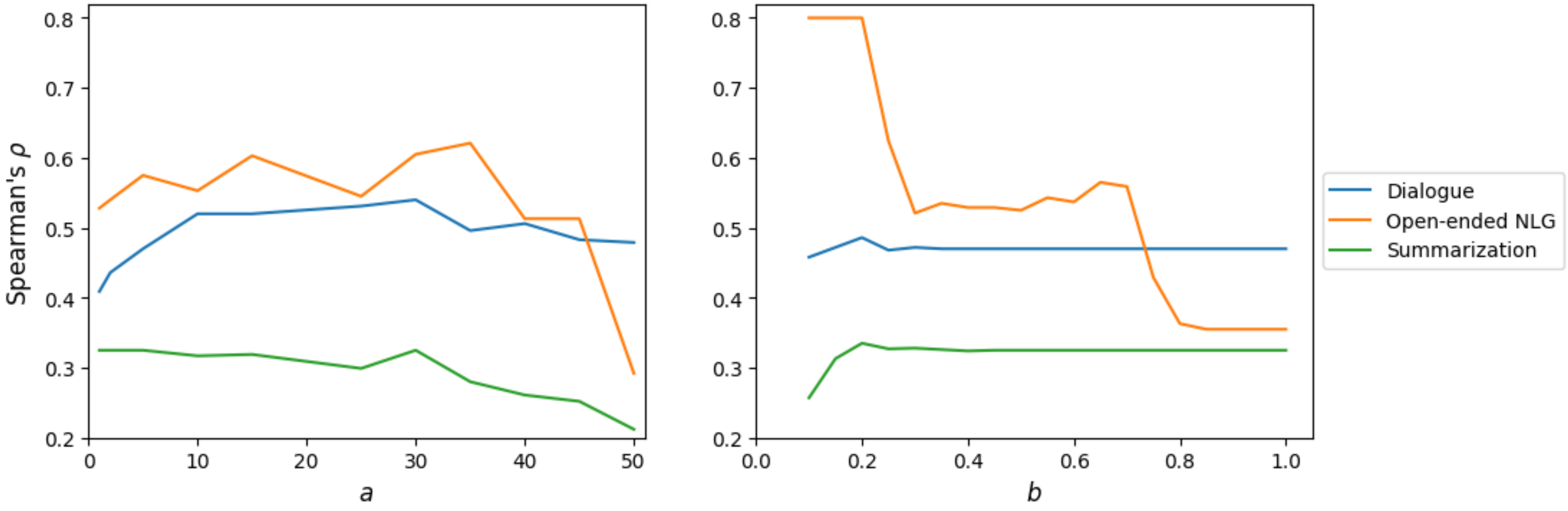
BLEU Neighbors outperforms its ROUGE, METEOR, and BERTScore counterparts while getting > 40% coverage.

	Dialogue			Open-ended Generation			Summarization		
	MSE	ρ	Coverage	MSE	ρ	Coverage	MSE	ρ	Coverage
Human (best)	0.0208	0.878	1.00	0.0177	0.861	1.00	0.0200	0.921	1.00
Human (average)	0.0807	0.456	1.00	0.0719	0.472	1.00	0.0802	0.405	1.00
BLEU Neighbors	**0.0164**	**0.470***	0.76	0.0204	**0.575***	0.41	0.0213	**0.325***	0.99
ROUGE Neighbors	0.0197	0.342*	0.86	**0.0174**	0.077	0.47	0.0226	0.245*	0.97
METEOR Neighbors	0.0165	0.382*	0.47	0.0209	0.395	0.22	**0.0180**	0.240	0.12
BERTScore Neighbors	0.0229	0.150*	0.89	0.0192	0.566*	0.32	0.0223	0.225	0.53

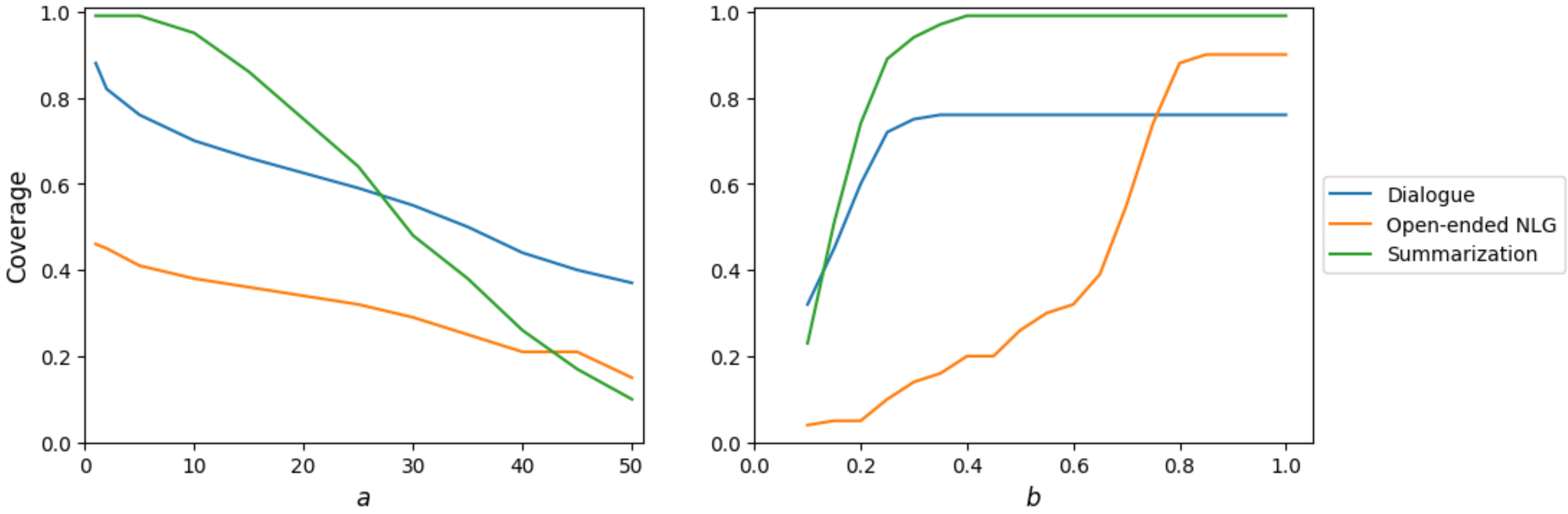
For open-ended generation and dialogue, BLEU Neighbors even outperforms human annotators (on average).

	Dialogue			Open-ended Generation			Summarization		
	MSE	ρ	Coverage	MSE	ρ	Coverage	MSE	ρ	Coverage
Human (best)	0.0208	0.878	1.00	0.0177	0.861	1.00	0.0200	0.921	1.00
Human (average)	0.0807	0.456	1.00	0.0719	0.472	1.00	0.0802	0.405	1.00
BLEU Neighbors	**0.0164**	**0.470***	0.76	0.0204	**0.575***	0.41	0.0213	**0.325***	0.99
ROUGE Neighbors	0.0197	0.342*	0.86	**0.0174**	0.077	0.47	0.0226	0.245*	0.97
METEOR Neighbors	0.0165	0.382*	0.47	0.0209	0.395	0.22	**0.0180**	0.240	0.12
BERTScore Neighbors	0.0229	0.150*	0.89	0.0192	0.566*	0.32	0.0223	0.225	0.53

Performance changes as evidence thresholds (i.e., min/max number of neighbors allowed) change.



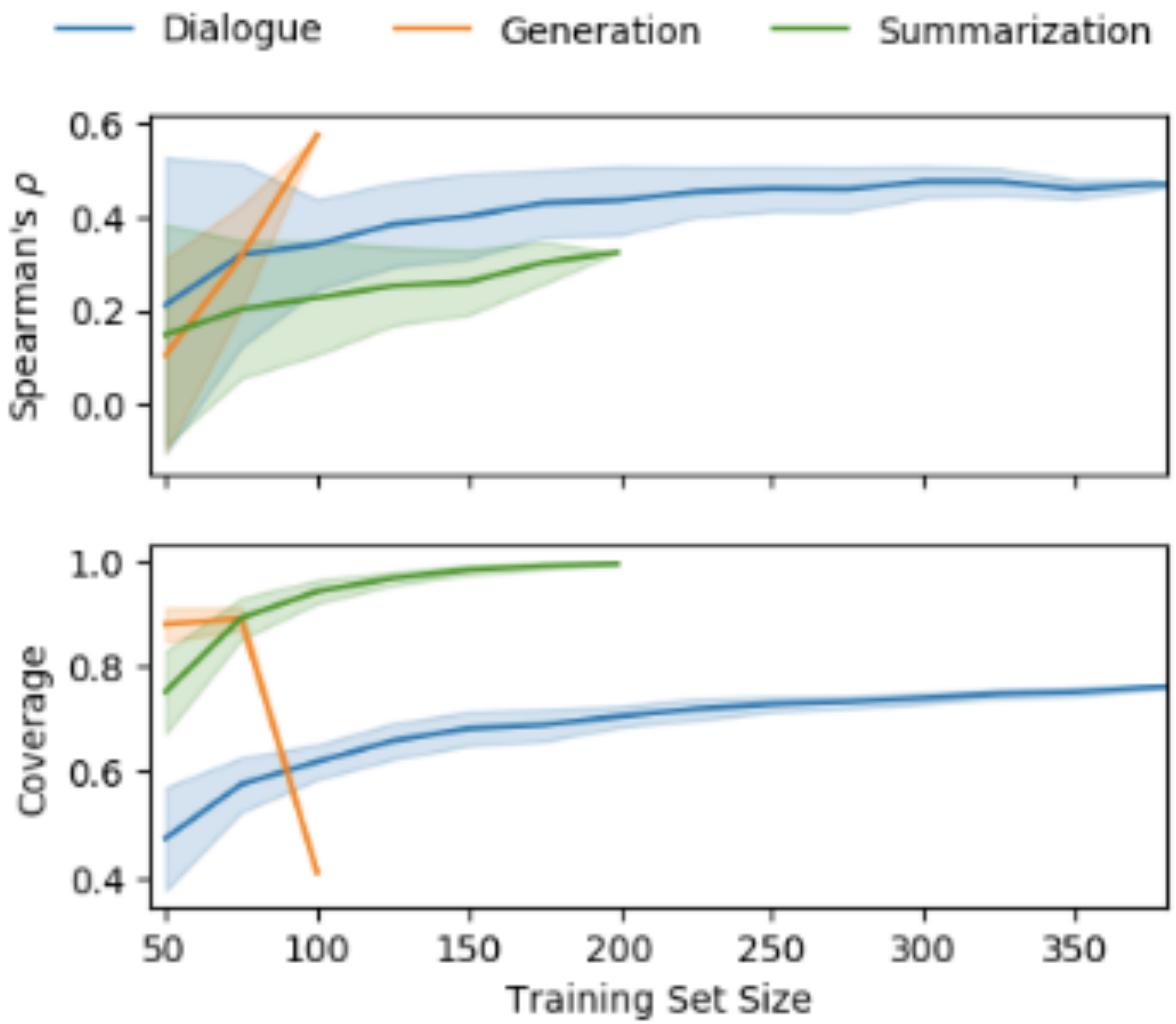
Coverage changes as evidence thresholds (i.e., min/max number of neighbors allowed) change.



BLEU Neighbors doesn't only make predictions for easy-to-score sentences (i.e., low-hanging fruit).

- There's no statistically significant difference between the MSE of annotators on all data vs. just those that are scored by BLEU Neighbors (except on dialogue).
- On dialogue, MSE is 15.6% higher on all data. But a similar difference exists with the sentences scored by ROUGE Neighbors, which performs much worse.

Performance is quite robust to the amount of training data, but coverage is not.



BLEU Neighbors even works when the train/test data are from different tasks (though not as well).

Source Task	Target Task		
	→ D	→ G	→ S
Dialogue (D) →	**0.470**	0.206	0.032
Generation (G) →	0.310	**0.575**	-0.070
Summarization (S) →	0.276	0.095	**0.325**

Limitations

- By design, BLEU Neighbors doesn’t measure language diversity.
- BLEU Neighbors doesn’t consider the source text (e.g., for summarization).
- BLEU Neighbors needs to be tested on larger and more diverse datasets for assurance that annotation artefacts are not being exploited.

Conclusion

- BLEU Neighbors is
  - a nearest neighbors approach to estimating language quality
  - simple, data-efficient, and correlates well with human judgment
- It can’t replace humans, but that’s not the goal; we just want to speed up NLG model development.

Thank you!