The Thirty-sixth International Conference on Machine Learning

Empirical Analysis of Beam Search Performance Degradation in Neural Sequence Models

Eldan Cohen J. Christopher Beck

Poster: Pacific Ballroom #47

Motivation

- Most commonly used inference algorithm for neural sequence decoding
- Intuitively, increasing beam width should lead to better solutions
- In practice, performance degradation for larger beams
  - While the search finds solutions that are more probable, they tend to have lower evaluation
- One of six main challenges in machine translation (Koehn & Knowles, 2017)

Beam Search Performance Degradation

Task	Dataset	Metric	B=1	B=3	B=5	B=25	B=100	B=250
Translation	En-De	BLEU4	25.27	26.00	**26.11**	25.11	23.09	21.38
	En-Fr	BLEU4	40.15	40.77	**40.83**	40.52	38.64	35.03
Summarization	Gigaword	R-1 F	33.56	**34.22**	34.16	34.01	33.67	33.23
Captioning	MSCOCO	BLEU4	29.66	**32.36**	31.96	30.04	29.87	29.79

- Different tasks: translation, summarization, image captioning

- Previous works highlighted potential explanations:
  - Machine translation: source copies (Ott et al., 2018)
  - Image captioning: training set predictions (Vinyals et al., 2017)

Analytical Framework: Search Discrepancies

- Inspired by search discrepancies in combinatorial search (Harvey & Ginsberg, 1995)

- Search discrepancy at sequence position t

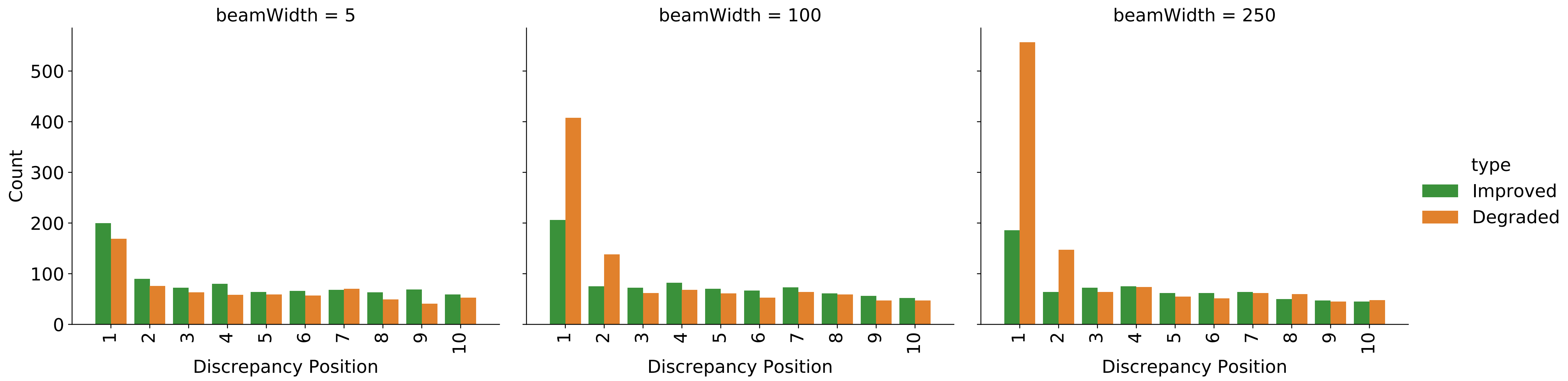
$$log P_\theta(y_t \mid \mathrm{x}; \{y_0, ..., y_{t-1}\}) < \max_{y \in \mathcal{V}} log P_\theta(y \mid \mathrm{x}; \{y_0, ..., y_{t-1}\}).$$

- Discrepancy gap for position t

$$\max_{y \in \mathcal{V}} \log P_\theta(y \mid \mathrm{x}; \{y_0, ..., y_{t-1}\}) - \log P_\theta(y_t \mid \mathrm{x}; \{y_0, ..., y_{t-1}\}).$$

Empirical Analysis (WMT'14 En-De)

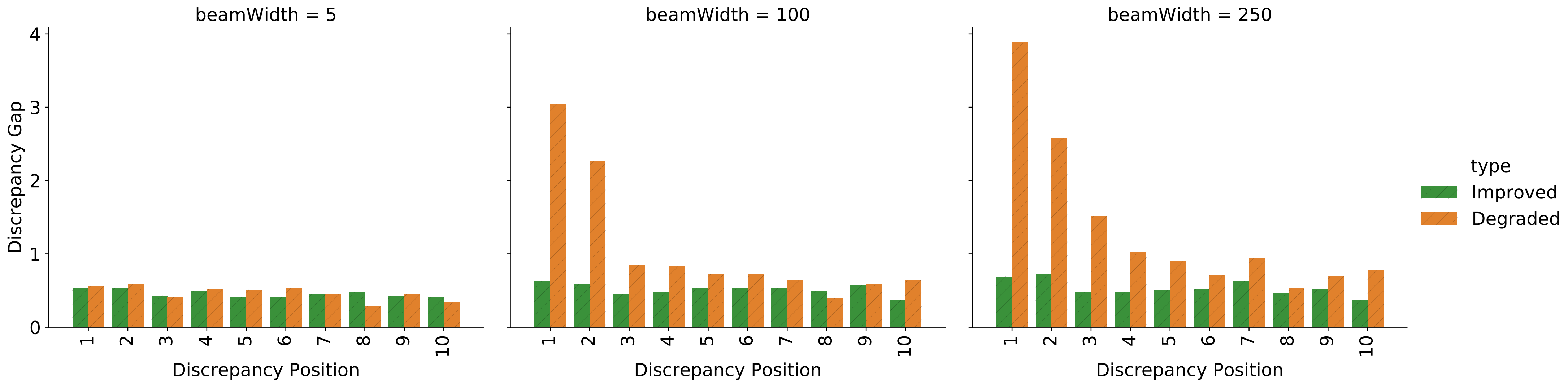
Search discrepancies vs. sequence position



- Increasing the beam width leads to more, early discrepancies
- For larger beam widths, these discrepancies are more likely to be associated with degraded solutions

Empirical Analysis (WMT'14 En-De)

Discrepancy gap vs. sequence position



- As we increase the beam width, the gap of early discrepancies in degraded solutions grows

Discrepancy-Constrained Beam Search

<sos> comment

vas [-0.69] | est [-0.92] | venu [-2.99] | ...

Discrepancy gap:	0	0.23	2.30	...	$\leq \mathcal{M}$
Candidate rank:	1	2	3	...	$\leq \mathcal{N}$

- M and N are hyper-parameters, tuned on a held-out validation set.
- **The methods successfully eliminate the performance degradation**

Summary

- Analytical framework based on search discrepancies
  - Performance degradation is associated with early large search discrepancies
- Propose two heuristics based on constraining the search discrepancies
  - Successfully eliminate the performance degradation.
- In the paper:
  - Detailed analysis of the search discrepancies
  - Our results generalize previous observations on copies (Ott et al., 2018) and training set predictions (Vinyals et al., 2017)
  - Discussion on the biases that can explain the observed patterns

Mechanical & Industrial Engineering
UNIVERSITY OF TORONTO

Toronto Intelligent Decision
TIDEL
Engineering Laboratory

The Thirty-sixth International Conference on Machine Learning

Empirical Analysis of Beam Search Performance Degradation in Neural Sequence Models

Eldan Cohen J. Christopher Beck

Poster: Pacific Ballroom #47