A. Full Empirical Analysis

A.1. Machine Translation on WMT'14 En-Fr

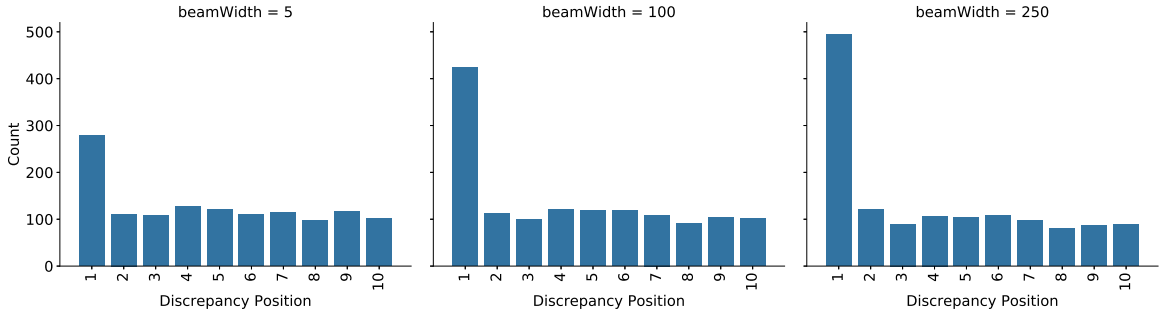

Figure 7. WMT'14 En-Fr: Distribution of discrepancy positions for different beam widths.

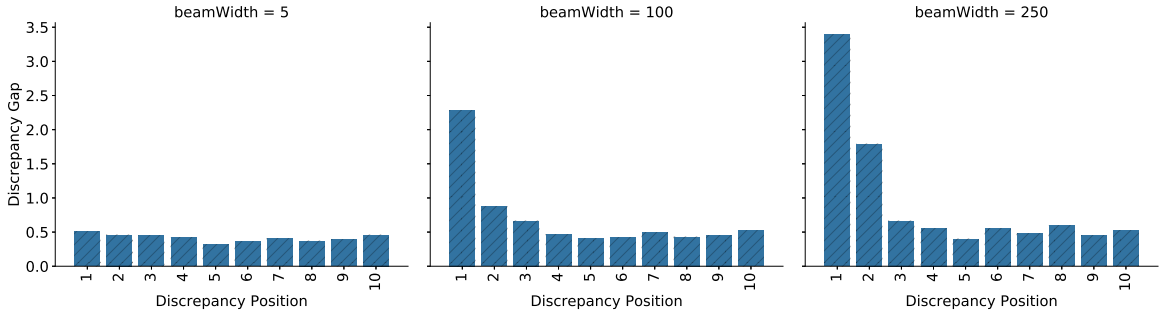

Figure 8. WMT'14 En-Fr: Mean discrepancy gap per position for different beam widths.

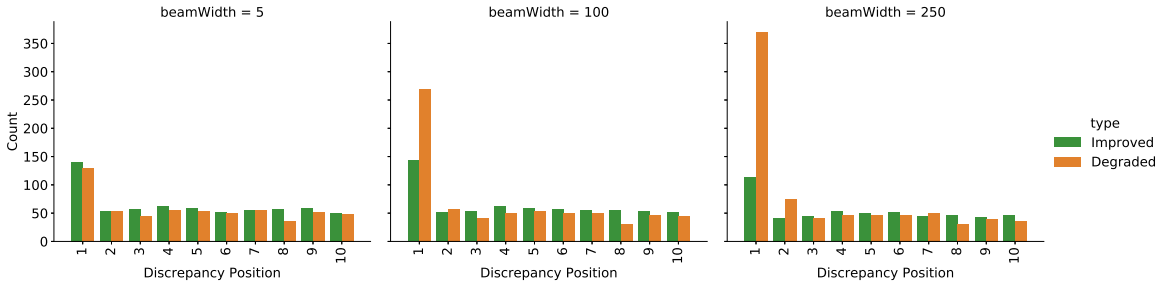

Figure 9. WMT'14 En-Fr: Distribution of discrepancy positions for different beam widths.

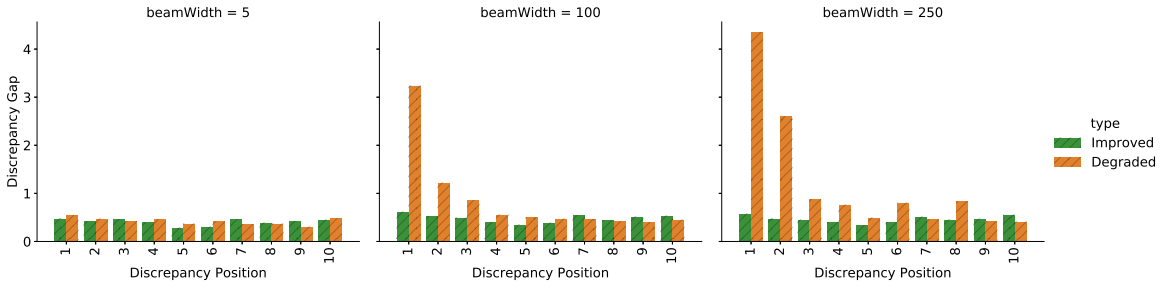

Figure 10. WMT'14 En-Fr: Mean discrepancy gap per position for different beam widths.

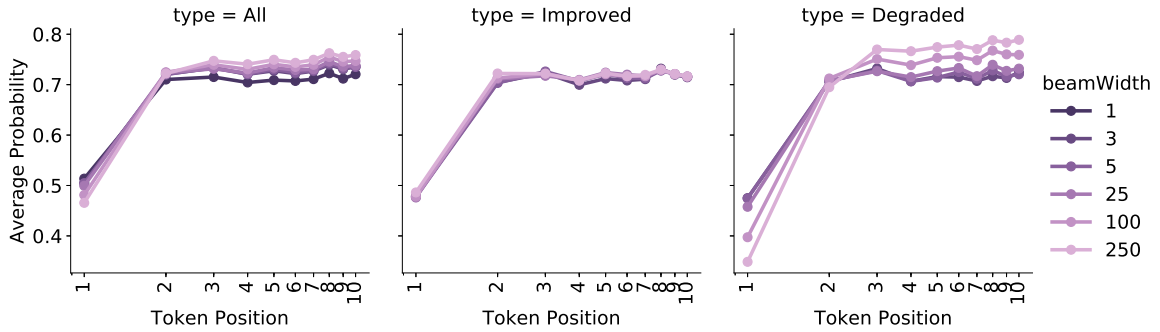

Figure 11. WMT'14 En-Fr: Average token probability per position for different beam widths.

A.2. Summarization on Gigaword

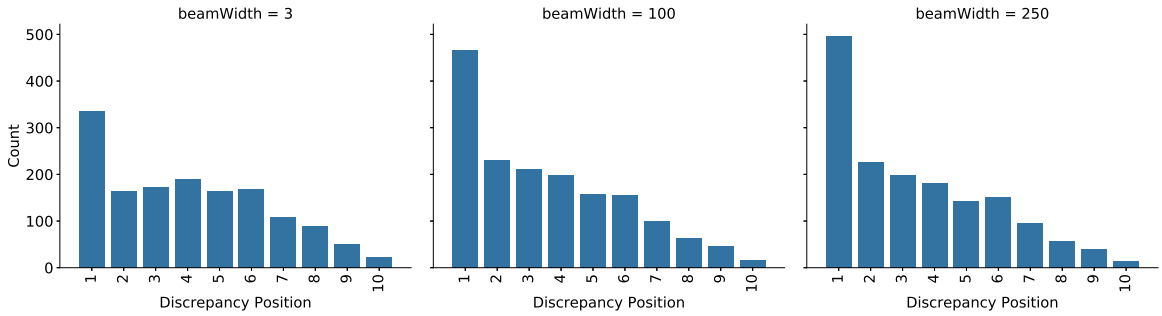

Figure 12. Gigaword: Distribution of discrepancy positions for different beam widths.

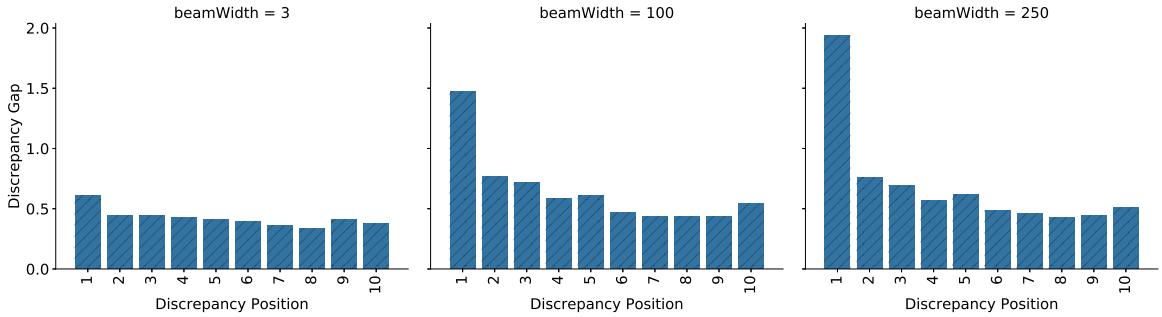

Figure 13. Gigaword: Mean discrepancy gap per position for different beam widths.

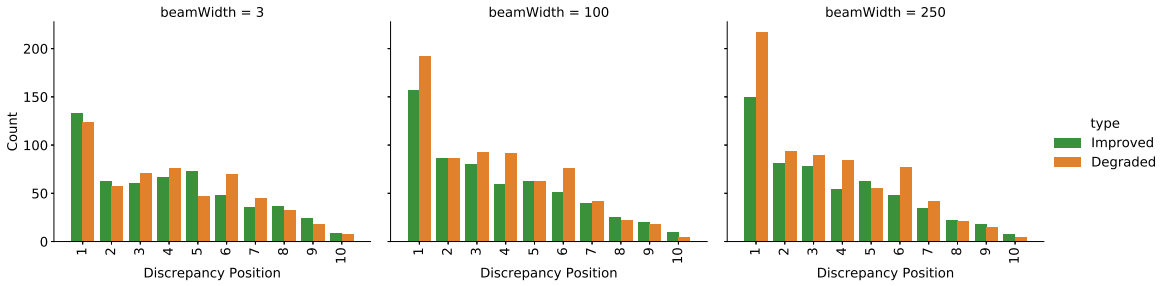

Figure 14. Gigaword: Distribution of discrepancy positions for different beam widths.

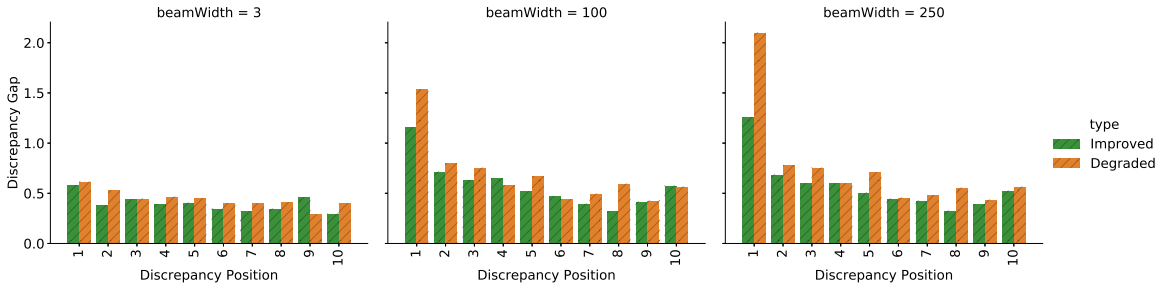

Figure 15. Gigaword: Mean discrepancy gap per position for different beam widths.

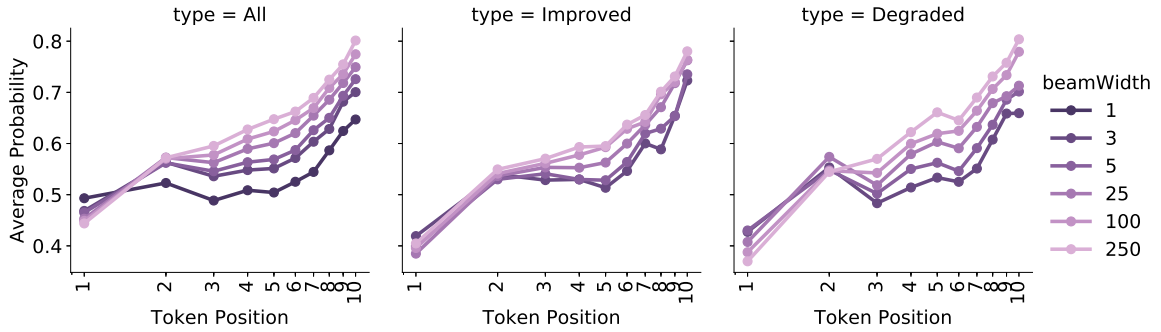

Figure 16. Gigaword: Average token probability per position for different beam widths.

A.3. Image Captioning on MSCOCO

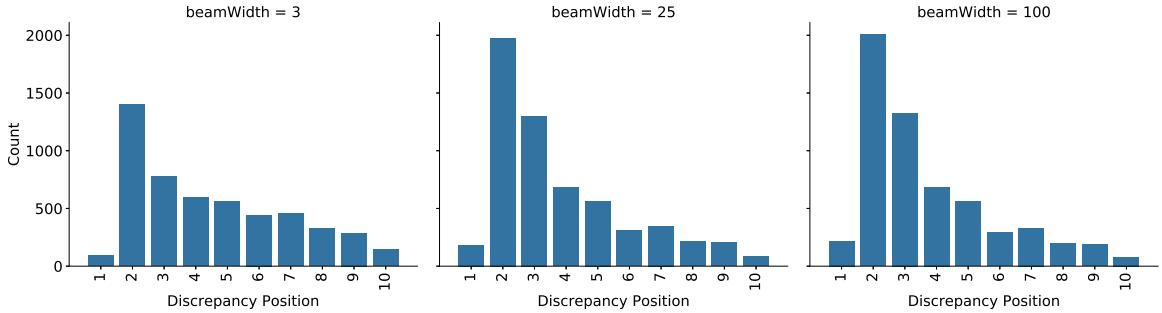

Figure 17. MSCOCO: Distribution of discrepancy positions for different beam widths.

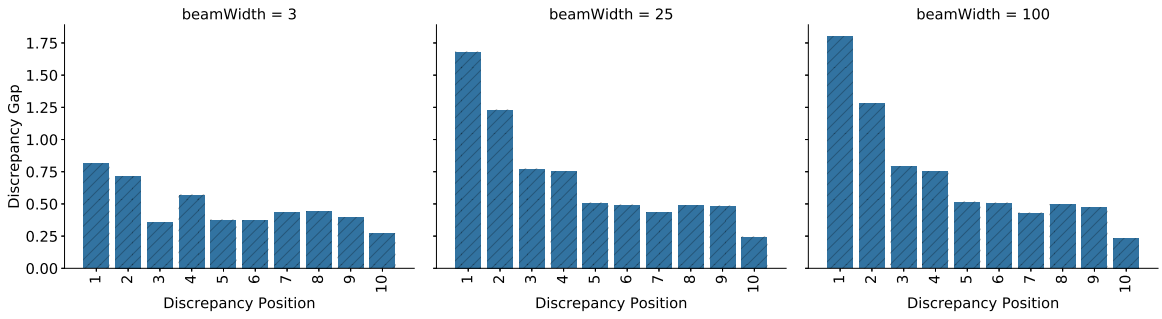

Figure 18. MSCOCO: Mean discrepancy gap per position for different beam widths.

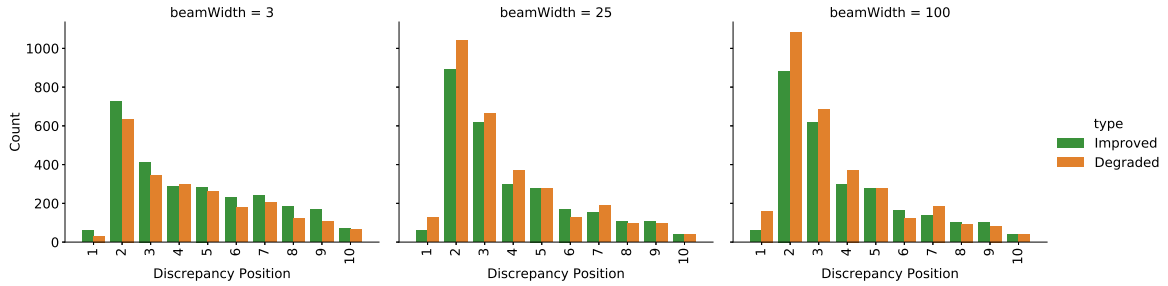

Figure 19. MSCOCO: Distribution of discrepancy positions for different beam widths.

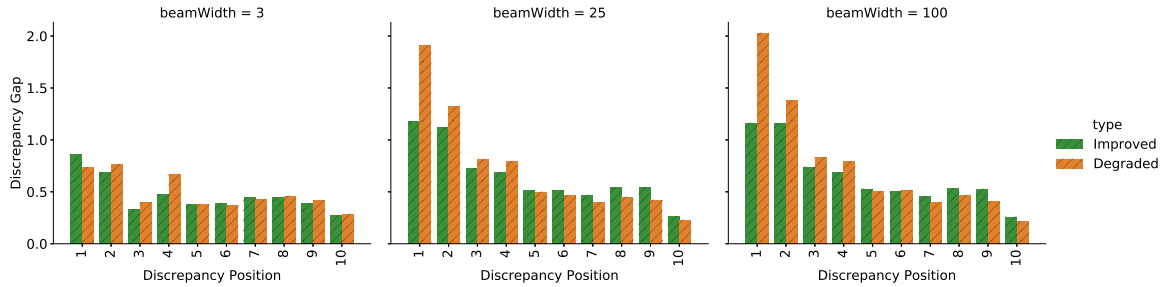

Figure 20. MSCOCO: Mean discrepancy gap per position for different beam widths.

Table 5. MSCOCO: Probability of the early (first two) tokens vs. the probability of the rest.

	All		Improved		Degraded	
Beam	Early	Rest	Early	Rest	Early	Rest
B=1	-1.48	-6.98	N/A	N/A	N/A	N/A
B=3	-1.68	-5.11	-1.76	-5.32	-1.78	-5.09
B=25	-2.02	-4.09	-2.04	-4.26	-2.21	-3.99
B=100	-2.07	-4.02	-2.06	-4.21	-2.30	-3.91
B=250	-2.08	-4.01	-2.06	-4.20	-2.31	-3.90

B. Copies and Training Set Predictions in Discrepancy-constrained Beam Search

Table 6 compares the number of copies in the baseline vs. the discrepancy-constrained methods for the machine translation tasks for each beam width. For the baseline, we can see that as we increase the beam width, the number of copies grows significantly. However, both discrepancy-constrained methods significantly reduce this growth.

Table 6. Number of copies in machine translations for the baseline and the two types of discrepancy-constrained beam search for different beam widths.

Dataset	Method	Parameter	B=1	B=3	B=5	B=25	B=100	B=250
En-De	Baseline		23	40	49	179	385	567
En-De	Constr. Gap	$\mathcal{M} = 1.5$	23	39	42	50	53	55
En-De	Constr. Rank	$\mathcal{N} = 2$	23	38	44	46	54	55
En-Fr	Baseline		25	28	41	89	227	358
En-Fr	Constr. Gap	$\mathcal{M} = 2.0$	25	27	37	43	46	45
En-Fr	Constr. Rank	$\mathcal{N} = 3$	25	28	38	42	42	46

Table 7 compares the number of training set predictions in the baseline vs. the discrepancy-constrained methods for the summarization and image captioning tasks for each beam width. For the baseline, we can see that as we increase the beam

width, the number of training set predictions grows significantly. However, as with copies, both discrepancy-constrained methods significantly reduce the growth in training set predictions.

Table 7. Number of predictions that are in the training set for the baseline and the two types of discrepancy-constrained beam search for different beam widths.

Dataset	Method	Parameter	B=1	B=3	B=5	B=25	B=100	B=250
Gigaword	Baseline		81	86	86	115	163	224
Gigaword	Constr. Gap	$\mathcal{M} = 0.85$	81	81	77	79	78	78
Gigaword	Constr. Rank	$\mathcal{N} = 2$	81	81	79	79	79	79
MSCOCO	Baseline		163	260	371	588	582	576
MSCOCO	Constr. Gap	$\mathcal{M} = 0.45$	163	265	271	271	271	271
MSCOCO	Constr. Rank	$\mathcal{N} = 2$	163	242	262	231	231	231

C. Results for Constrained Beam Search on WMT'14 En-De

C.1. Results for Discrepancy Gap Constrained Beam Search ($\mathcal{M} = 1.5$)

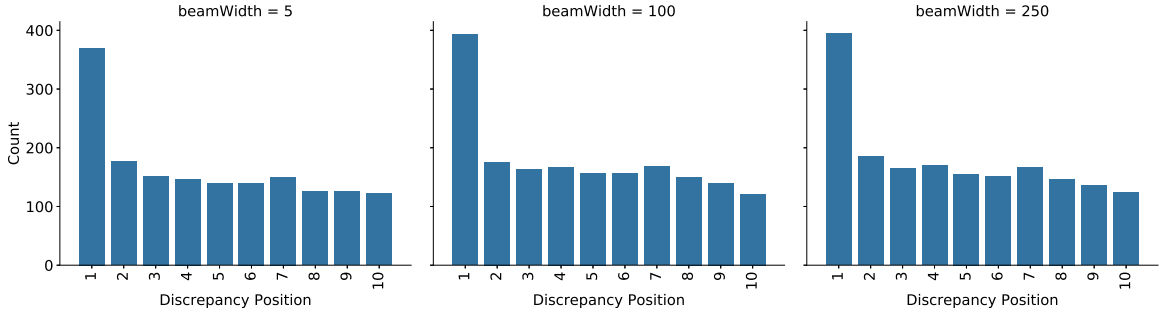

Figure 21. WMT'14 En-De: Distribution of discrepancy positions ($\mathcal{M} = 1.5$).

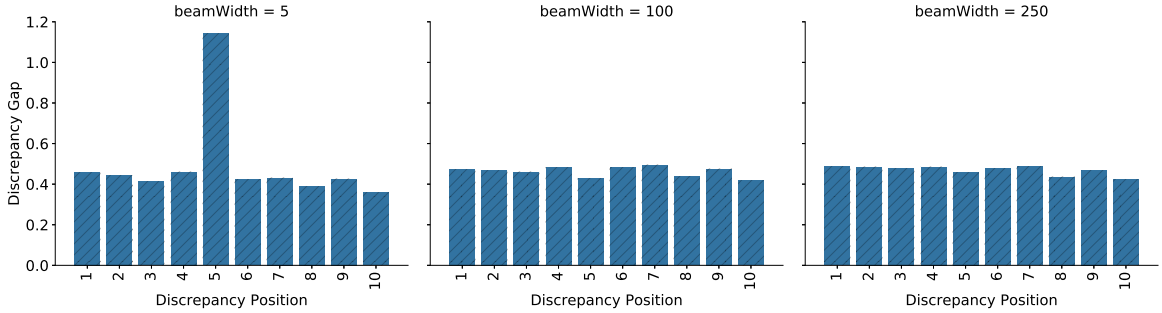

Figure 22. WMT'14 En-De: Mean discrepancy gap per position ($\mathcal{M} = 1.5$).

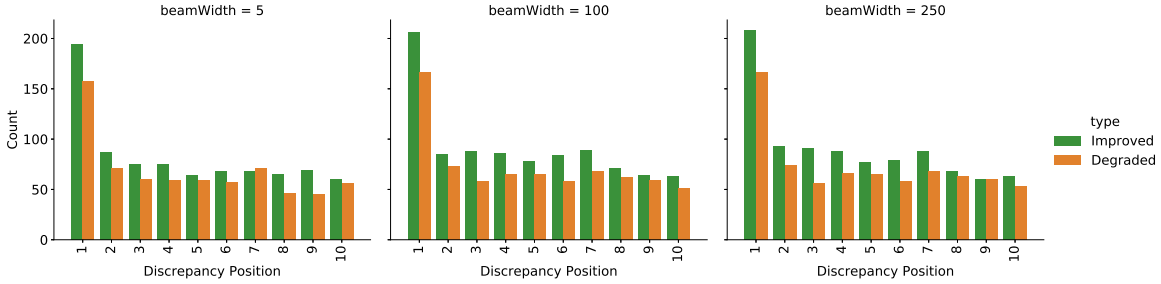

Figure 23. WMT'14 En-De: Distribution of discrepancy positions ($\mathcal{M} = 1.5$).

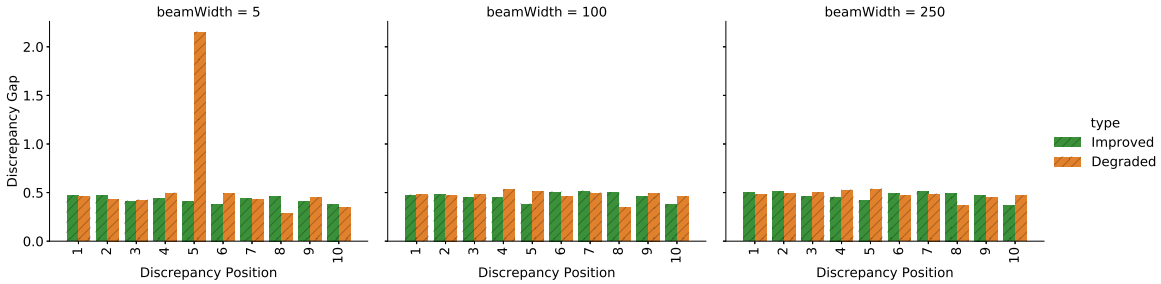

Figure 24. WMT'14 En-De: Mean discrepancy gap per position ($\mathcal{M} = 1.5$).

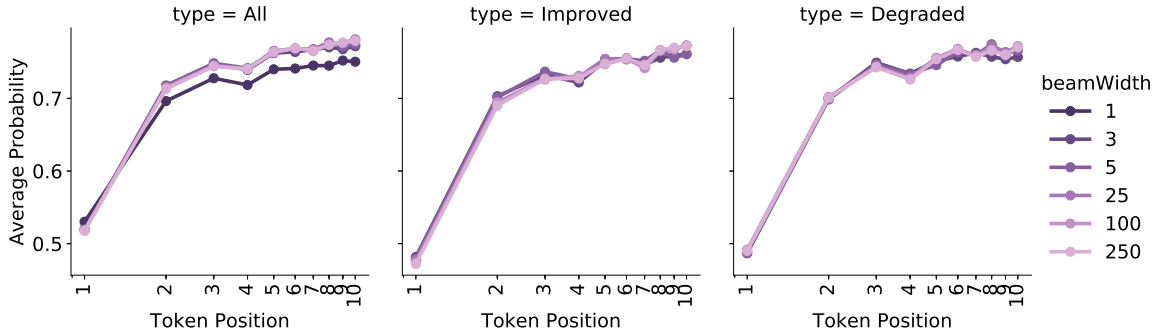

Figure 25. WMT'14 En-De: Average token probability per position ($\mathcal{M} = 1.5$).

C.2. Results for Rank Constrained Beam Search ($\mathcal{N} = 2$)

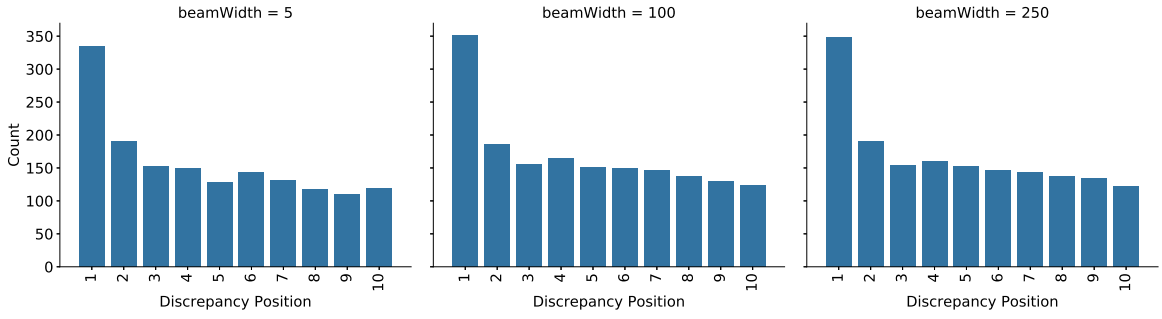

Figure 26. WMT'14 En-De: Distribution of discrepancy positions ($\mathcal{N} = 2$).

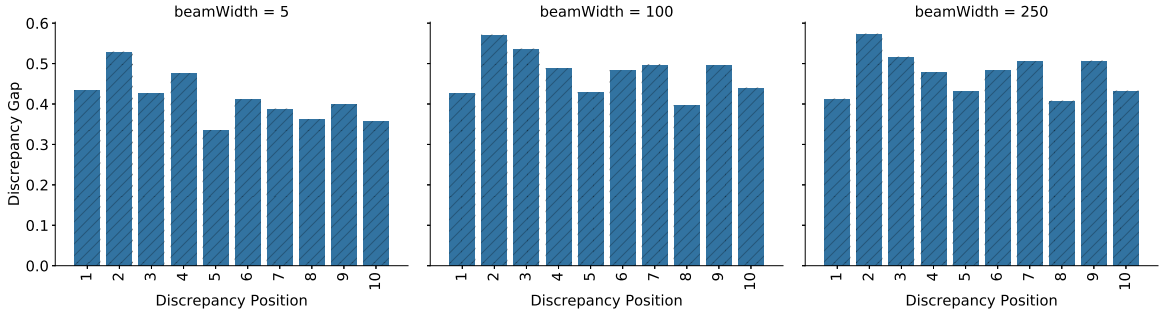

Figure 27. WMT'14 En-De: Mean discrepancy gap per position ($\mathcal{N} = 2$).

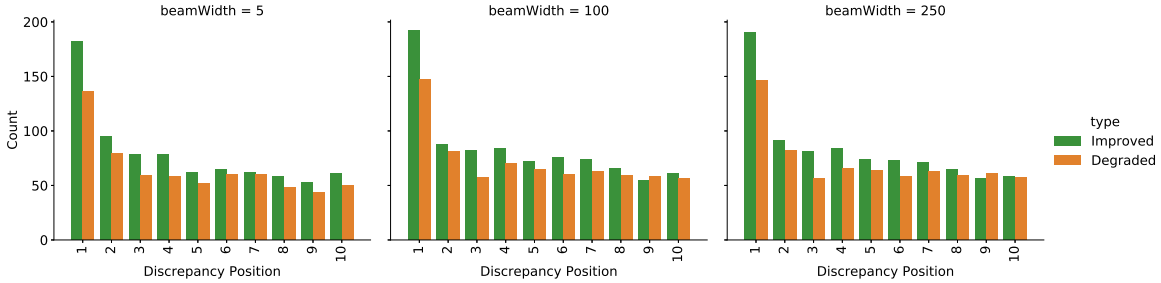

Figure 28. WMT'14 En-De: Distribution of discrepancy positions ($\mathcal{N} = 2$).

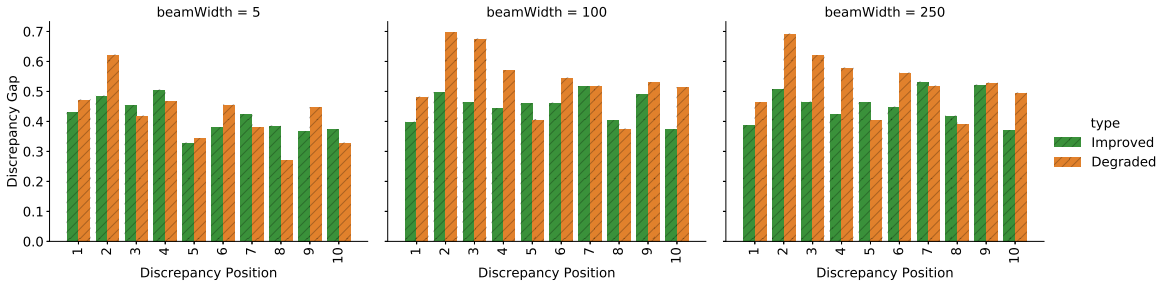

Figure 29. WMT'14 En-De: Mean discrepancy gap per position ($\mathcal{N} = 2$).

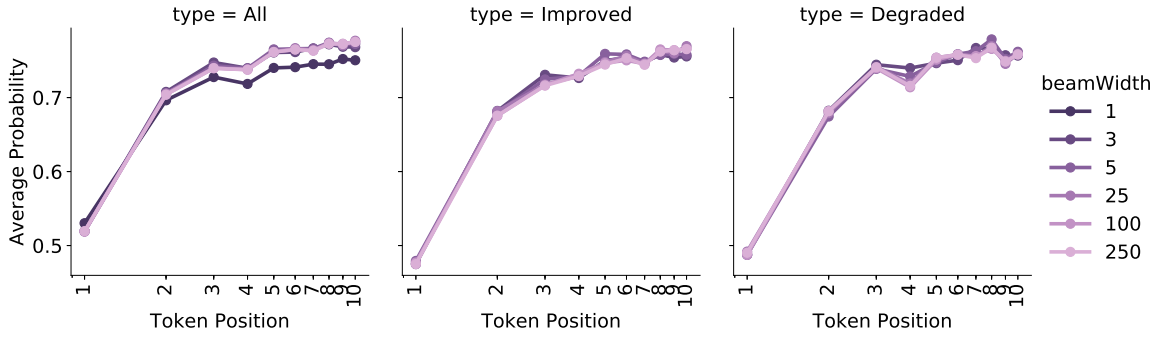

Figure 30. WMT'14 En-De: Average token probability per position ($\mathcal{N} = 2$).

D. Analysis of Length Bias

Table 8 shows the mean length of generated sentences for different beam widths for the baseline, normalized to the best tested beam width. All values are very close 1.0, which suggest that the observed performance degradation is not due to length bias. We note that for machine translation and summarization this is due to the use of length normalization on the hypotheses log-likelihood, as suggested by Koehn & Knowles (2017) (without normalization, the performance degradation would have been worse).[6] In image captioning, however, there is no observed length bias even when length normalization is not used.[7]

In Section 4.1, we showed substantial performance degradation as we increase the beam width. As the results in Table 8 demonstrate that there is no significant change in the length of generated sequences, the observed performance degradation cannot be attributed to length bias.

Table 8. Analysis of the mean length, normalized to best test width (in bold).

Task	Dataset	B=1	B=3	B=5	B=25	B=100	B=250
Translation	En-De	0.99	1.0	**1.0**	1.0	0.99	0.98
	En-Fr	0.99	1.0	**1.0**	1.0	0.99	0.91
Summarization	Gigaword	1.03	**1.0**	0.99	0.99	1.0	1.01
Captioning	MSCOCO	1.04	**1.0**	0.99	0.98	0.98	0.98

E. Image Captioning: CIDEr and SPICE

Table 9 compares the baseline vs. the constrained beam search methods on the MSCOCO image caption task using the metrics CIDEr and SPICE. The results show similar trends to those observed for BLEU in Section 6. In particular, we see that the performance degradation for larger beams also occurs for CIDEr and SPICE in the baseline, and that our gap constraint method eliminates this degradation. Similar to our results for BLEU, we note that our rank constraint is not as effective as our gap constraint for the image captioning task.

Table 9. Evaluation of image captioning on MSCOCO dataset using the CIDEr and SPICE metrics (higher values are better, best baseline in bold).

Dataset	Method	Threshold	B=1	B=3	B=5	B=25	B=100	B=250
CIDEr	Baseline		0.974	**1.018**	1.005	0.953	0.946	0.945
	Constr. Gap	$\mathcal{M} = 0.4$	0.974	1.016	1.018	1.016	1.016	1.016
	Constr. Rank	$\mathcal{N} = 2$	0	1.022	1.006	0.978	0.977	0.977
SPICE	Baseline		18.13	**18.54**	18.43	17.76	17.68	17.64
	Constr. Gap	$\mathcal{M} = 0.45$	18.13	18.41	18.44	18.43	18.43	18.43
	Constr. Rank	$\mathcal{N} = 2$	0	18.60	18.51	18.15	18.15	18.15

[6] This is consistent with Ott et al.'s (2018) results on performance degradation even when using length normalization.

[7] In fact, as we stated earlier, we found that length normalization reduces the overall performance.

F. Results for WMT'17 En-Zh

To show that our results extend beyond generating text in English or in European languages that share some similarity to English, we present results for the WMT'17 En-Zh dataset, using the Nematus toolkit (Sennrich et al., 2017). Table 10 shows the results for the baseline vs. the constrained methods for different beam widths. Consistent with WMT'17 instructions for evaluating Chinese output, we report BLEU-4 scores computed on characters.[8] The results show a clear performance degradation for the baseline, with BLEU-4 score dropping by more than 3 points. Similar to the other tasks, our constrained methods successfully eliminate the performance degradation and even lead to a higher evaluation.

Table 10. A comparison of the baseline results vs. the constrained beam search methods for the WMT'17 En-Zh dataset based on the BLEU-4 metric (higher values are better; best baseline result in bold).

Dataset	Method	Parameter	B=1	B=3	B=5	B=25	B=100	B=250
En-Zh	Baseline		32.41	**33.17**	33.16	33.01	31.33	29.61
En-Zh	Constr. Gap	$\mathcal{M} = 1.0$	32.41	33.15	33.22	33.43	33.45	33.50
En-Zh	Constr. Rank	$\mathcal{N} = 2$	32.41	33.20	33.17	33.35	33.28	33.30

Figure 31 shows the distribution of discrepancy positions for different beam widths and Figure 32 shows the mean discrepancy gap per position for different beam widths. We can see that the results for WMT'17 En-Zh exhibit the same phenomena observed for the other datasets in Section 4.

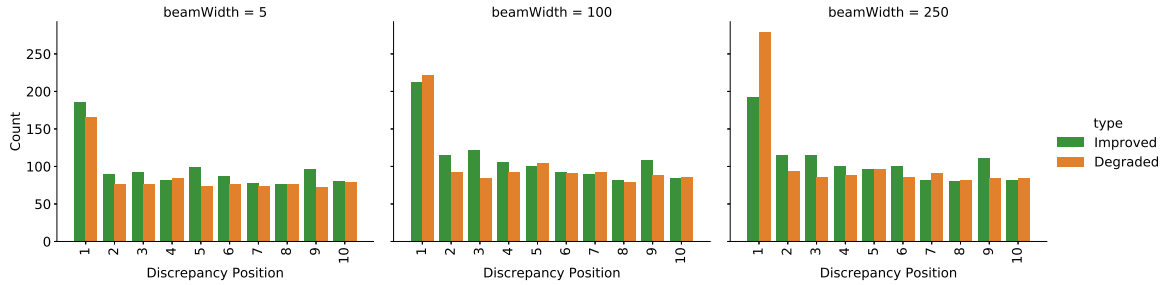

Figure 31. WMT'17 En-Zh: Distribution of discrepancy positions for different beam widths.

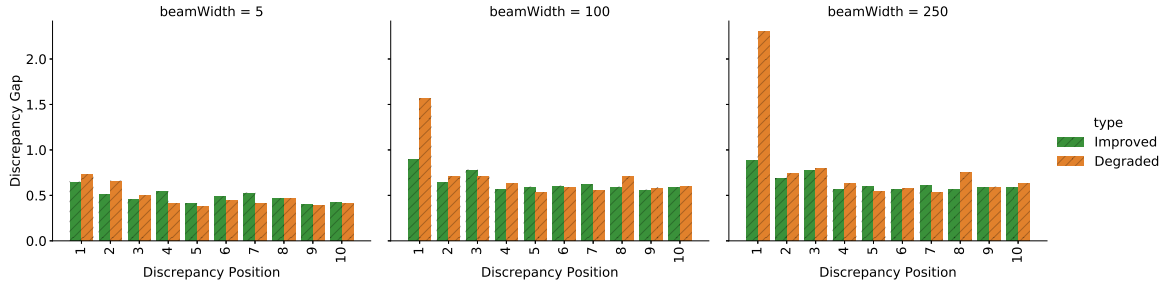

Figure 32. WMT'17 En-Zh: Mean discrepancy gap per position for different beam widths.

[8] http://www.statmt.org/wmt17/translation-task.html