# Empirical Analysis of Beam Search Performance Degradation in Neural Sequence Models

**Eldan Cohen** [1] **J. Christopher Beck** [1]

## Abstract

Beam search is the most popular inference algorithm for decoding neural sequence models. Unlike greedy search, beam search allows for non-greedy local decisions that can potentially lead to a sequence with a higher overall probability. However, work on a number of applications has found that the quality of the highest probability hypothesis found by beam search degrades with large beam widths. We perform an empirical study of the behavior of beam search across three sequence synthesis tasks. We find that increasing the beam width leads to sequences that are disproportionately based on early, very low probability tokens that are followed by a sequence of tokens with higher (conditional) probability. We show that, empirically, such sequences are more likely to have a lower evaluation score than lower probability sequences without this pattern. Using the notion of search discrepancies from heuristic search, we hypothesize that large discrepancies are the cause of the performance degradation. We show that this hypothesis generalizes the previous ones in machine translation and image captioning. To validate our hypothesis, we show that constraining beam search to avoid large discrepancies eliminates the performance degradation.

## 1. Introduction

Neural sequence models are among the most popular tools for modeling sequential data and have been applied to a range of applications including machine translation (Gehring et al., 2017), summarization (Chopra et al., 2016), image captioning (Vinyals et al., 2017), and conversation modeling (Vinyals & Le, 2015). The most commonly used inference algorithm for decoding neural sequence models is beam search, a search algorithm that generates the sequence tokens one-by-one while keeping a fixed number of active candidates (beam size) at each step.

Recently, several works have reported the problem of performance degradation in beam search. In machine translation, Koehn & Knowles (2017) found that beam search "only improves translation for narrow beams and deteriorates when exposed to a larger search space". While length-normalization can alleviate the problem somewhat, it does not eliminate it. Koehn & Knowles chose this problem as one of six central challenges in machine translation.

Ott et al. (2018) proposed the existence of training pairs in which the target is a copy of the source as an explanation for the performance degradation in length-normalized machine translation models. For larger beams, they found that more predictions can be classified as "copies"[1] and that filtering these copies reduces the performance degradation.

In image captioning, Vinyals et al. (2017) observed performance degradation for wider beams and highlighted the use of a narrower beam search as one of the most significant improvements in their model. They hypothesized that the degradation is either due to overfitting or that the objective used in training (likelihood) is not aligned with human judgement. Their analysis found that wider beams exhibited more predictions that repeat training captions and fewer novel ones. This observation is used to support the hypothesis that the model is overfitted and therefore they propose the use of smaller beam width as "another way to regularize".

In this work, we analyze the performance of beam search across multiple tasks: machine translation, abstractive summarization, and image captioning. We present an explanatory model that is based on the concept of *search discrepancies* (deviations from greedy choices) and perform an empirical study of the distribution of such discrepancies. We make the following contributions:

1. We show that increasing the beam width leads to solutions with more and larger discrepancies early in the sequence. These sequences often have lower evaluation

[1]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada. Correspondence to: Eldan Cohen <ecohen@mie.utoronto.ca>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

[1]"Copies" are predictions that share at least 50% of their unigrams with their source (Ott et al., 2018).

score, leading to the observed performance degradation. As we increase the beam width, the differences in the distribution of discrepancies that are associated with improved vs. degraded solutions grow substantially.

2. We show that our explanatory model generalizes the previously observed "copies" and predictions that repeat training set targets and accounts for more of the degraded predictions.

3. We demonstrate that modifying beam search to prevent it from considering large search discrepancies eliminates the performance degradation.

## 2. Preliminaries

### 2.1. Neural Sequence Models

Given a model parameterized by $\theta$ and an input $x$, the problem of sequence generation consists of finding a sequence $\hat{\mathrm{y}}$ such that $\hat{\mathrm{y}} = \arg\max_{\mathrm{y}\in Y} P_\theta(\mathrm{y}|\mathrm{x})$, where $Y$ is the set of all sequences. y is a sequence of tokens $\mathrm{y} = \{y_0, ...y_{T-1}\}$ from vocabulary $\mathcal{V}$, where $T$ is the length of the sequence y. The expression $P_\theta(\mathrm{y} \mid \mathrm{x})$ can then be factored as $P_\theta(\mathrm{y} \mid \mathrm{x}) = \prod_{t=0}^{T-1} P_\theta(y_t \mid \mathrm{x}; \{y_0, ..., y_{t-1}\})$, or for convenience using log-probability as $\sum_{t=0}^{T-1} \log P_\theta(y_t \mid \mathrm{x}; \{y_0, ..., y_{t-1}\})$.

It is common to model $\log P_\theta(y_t \mid \mathrm{x}; \{y_0, ..., y_{t-1}\})$ using a Recurrent Neural Network (RNN), where the sequence $\{y_0, ..., y_{t-1}\}$ conditioned on is expressed by a fixed length hidden state $h_t$. This hidden state is updated using a non-linear function $f$: $h_{t+1} = f(h_t, y_t)$.

Exhaustive search to find the globally optimal sequence is not tractable. A greedy algorithm that selects the best candidate at each time step $y_t = \arg\max_{y\in\mathcal{V}} \log P_\theta(y|\mathrm{x}; \{y_0, ..., y_{t-1}\})$ makes a sequence of locally optimal decisions, but can lead to a globally sub-optimal sequence. Beam search extends the $B$ most probable partial solutions at each step, where $B$ is called *beam width*. Following Vijayakumar et al. (2018), we denote the set of $B$ solutions held by the beam search at step $t-1$ as $Y_{[t-1]}=\{\mathrm{y}_{1,[t-1]}, ..., \mathrm{y}_{B,[t-1]}\}$. At each step, beam search selects the top scoring $B$ candidates from the set of all possible one token extensions of its beams $\mathcal{Y}_t=\{\mathrm{y}_{[t]} \mid \mathrm{y}_{[t-1]} \in Y_{[t-1]} \wedge y_t \in \mathcal{V}\}$. Formally, the beam search candidates are updated as follows:

$$Y_{[t]} = \underset{\mathrm{y}_{[1,t]},...,\mathrm{y}_{[B,t]}\in\mathcal{Y}_t}{\arg\max} \sum_{b\in[1..B]} \log P_\theta(\mathrm{y}_{[b,t]} \mid \mathrm{x}) \quad s.t. \quad \mathrm{y}_i \neq \mathrm{y}_j \quad \forall i \neq j; \quad i,j \in [1..B] \tag{1}$$

### 2.2. Search Discrepancies in Neural Sequence Generation

In combinatorial search, a search discrepancy is a decision made by the search algorithm that is not the most highly rated one according to the heuristic (Harvey & Ginsberg, 1995). In the context of search for neural sequence generation, we define a search discrepancy as extending a partial sequence with a token that is not the most probable one. More formally, a sequence y is considered to have a search discrepancy at time step $t$ if

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

We denote the difference in log-probability between the most likely token and the chosen token as *discrepancy gap*. At time step $t$, the discrepancy gap is defined as

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

To demonstrate how the discrepancy gap is computed, Figure 1 shows the extension of a partial hypothesis in machine translation. The candidate with the highest conditional probability has a discrepancy gap of zero, by definition, while the gap of the other candidates is the difference in log-probability.

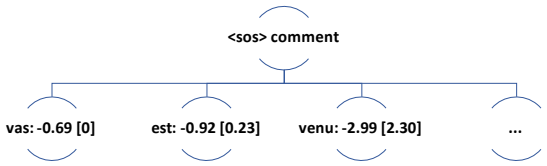


*Figure 1.* Example: expanding a partial hypothesis in the translation of "How are you?" to French. Discrepancy gap in brackets.

## 3. Experimental Setup

We perform an extensive empirical evaluation over multiple tasks, models, datasets, and evaluation metrics. Following is a description of the experimental setup for each task.

**Machine Translation.** We use the convolutional model by Gehring et al. (2017) implemented in the *fairseq-py* toolkit. We present results for two models, trained on WMT'14 En-Fr and En-De datasets and evaluated on newstest2014 En-Fr and En-De, respectively, with a vocabulary based on byte pair encoding (BPE; Sennrich et al., 2016).

**Summarization.** We use the abstractive summarization model by Chopra et al. (2016) implemented in the OpenNMT toolkit (Klein et al., 2017). The model is trained and evaluated using Rush et al.'s (2015) test split of the Gigaword corpus (Graff et al., 2003).

**Image Captioning.** We use the model by Vinyals et al. (2017), trained on the MSCOCO dataset (Lin et al., 2014). We present results for a test set of 5000 images based on Karpathy & Fei-Fei's (2015) splits.

*Table 1.* Baseline results for different beam widths (higher values are better, best results in bold).

| Task | Dataset | Size (Test) | Metric | $B$=1 | $B$=3 | $B$=5 | $B$=25 | $B$=100 | $B$=250 |
|---|---|---|---|---|---|---|---|---|---|
| Translation | En-De | 3003 | BLEU4 | 25.27 | 26.00 | **26.11** | 25.11 | 23.09 | 21.38 |
| | En-Fr | 3003 | BLEU4 | 40.15 | 40.77 | **40.83** | 40.52 | 38.64 | 35.03 |
| Summarization | Gigaword | 1751 | R-1 F | 33.56 | **34.22** | 34.16 | 34.01 | 33.67 | 33.23 |
| Captioning | MSCOCO | 5000 | BLEU4 | 29.66 | **32.36** | 31.96 | 30.04 | 29.87 | 29.79 |

In machine translation and summarization, we apply length normalization on the hypotheses log-likelihood, as it was shown to reduce the performance degradation by not prioritizing short sentences (Koehn & Knowles, 2017; Gehring et al., 2017). For image captioning, consistent with previous works, we do not use length normalization (we also found that such normalization reduces the overall performance).

### 3.1. Evaluation Metrics

While beam search finds the (approximately) most probable sequence, the quality of a sequence is evaluated based on human references using a task-specific evaluation metric. For machine translation and image captioning we use BLEU-$n$ (Papineni et al., 2002), a geometric average of precision over 1- to $n$-grams multiplied by a brevity penalty for short sentences. As in recent literature, we present results for BLEU-4. Corpus-level BLEU is reported without smoothing, while for sentence-level BLEU we use smoothed $n$-gram counts for $n > 1$ (Lin & Och, 2004). For image captioning, we also evaluated the performance using CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016) and report these metrics in Appendix E.[2]

For summarization, we use ROUGE (Lin, 2004), the $n$-gram recall between candidate summary and a reference. We report the F-score of ROUGE-1, however similar trends were observed for the F-score of ROUGE-L (for longest common subsequence).

[2]All the appendices are in the supplementary material.

## 4. Empirical Analysis of Search Discrepancies in Beam Search

We analyze and compare the most likely hypotheses found by a beam search for the following beam widths: {1, 3, 5, 25, 100, 250}. Due to space, we present detailed results for one of the tasks and summarize the results for the others. The results for all tasks and metrics are in Appendix A.

### 4.1. Baseline Results

Table 1 presents the performance of beam search with different beam widths, based on the chosen evaluation metrics. The performance degradation for larger beam widths appears for all tested tasks based on their task-specific evaluation metric. These results are consistent with the existing reports of such performance degradation (Koehn & Knowles, 2017; Ott et al., 2018; Vinyals et al., 2017).

### 4.2. The Distribution of Search Discrepancies

Figure 2 shows the number of discrepancies per position (index) for the most likely hypotheses generated by a beam search on the WMT'14 En-De test set for different beam widths (all graphs are based on the same number of solutions, however the total number of discrepancies in the generated solutions is not necessarily the same for different beam widths). In general, the majority of discrepancies happen in early positions. More interestingly, for larger beams, the number of early discrepancies grows significantly while the number of later discrepancies stays approximately constant. Larger beams allow the search to find solutions with

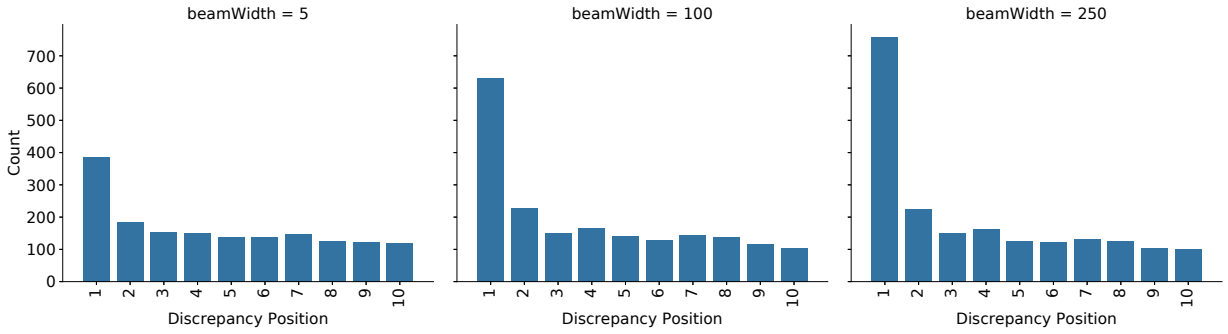


*Figure 2.* WMT'14 En-De: Distribution of discrepancy positions for different beam widths.

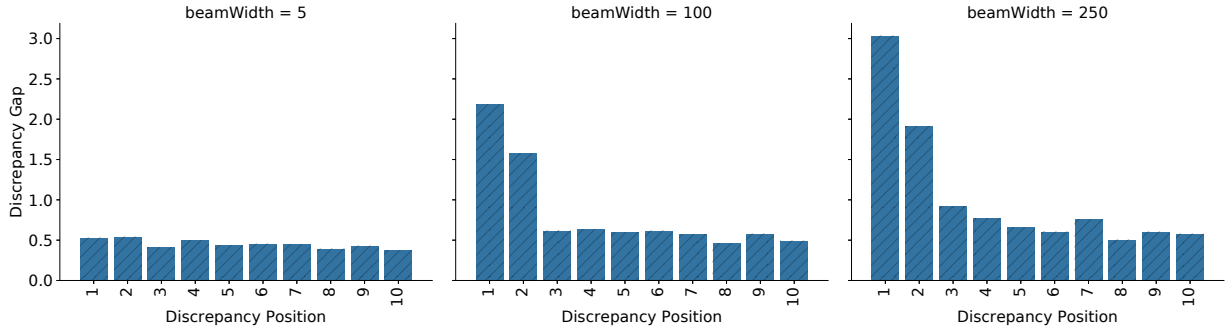


*Figure 3.* WMT'14 En-De: Mean discrepancy gap per position for different beam widths.

higher overall probability by exploring less probable early tokens, however they do not seem to lead to more probable sequences that share a prefix with solutions found for a smaller beam width. Similar results for the other tasks are reported in Appendix A. For image captioning (MSCOCO), we find the majority of early search discrepancies appear on the second token due to the first token being "a" with high probability in almost all sentences (in greedy search, for example, 99% of the generated captions start with "a").

Next, we analyze the discrepancy gap vs. sequence position. Figure 3 presents the mean gap per position for WMT'14 En-De for different beam widths. Again, we can see that the changes are mainly in the early positions: for larger beams, the search tends to find solutions with larger early discrepancy gap, i.e., the early tokens are relatively less likely. The gap of the other tokens remains similar. Similar results for the other tasks are reported in Appendix A.

The increase in number and size of early discrepancies for larger beams means that the search manages to find solutions with higher overall probability when starting from a large discrepancy. However, these solutions are not necessarily better according to the evaluation metric. The observed performance degradation suggests that the more probable solutions found by larger beams are, in fact, worse. Identifying discrepancies that are likely to lead to a worse solution is therefore a key task in addressing the degradation.

### 4.3. Discrepancies in Improved vs. Degraded Solutions

We now compare the solutions generated by a greedy search with the solutions generated by beam search with different widths. We then analyze the discrepancies in solutions that were improved by increasing the beam width (with respect to the evaluation metric) vs. solutions that were degraded.

Figure 4 shows the number of discrepancies per position for WMT'14 En-De, comparing solutions that were improved vs. solutions that were degraded. For $B$=5 there are 386 solutions in which the first token is not based on a greedy decision. Of those, 200 have a better evaluation than the greedy solution and 169 have a worse evaluation. However, as we increase the beam width, the increase in early discrepancies observed in Figure 2 is associated almost entirely with degraded solutions. This result explains the observed performance degradation for larger beam widths. Similar results for the other tasks are reported in Appendix A.

Next, we compare the discrepancy gaps in degraded vs. improved solutions. Figure 5 presents the mean discrepancy gap per position for both the improved and the degraded solutions. Interestingly, we find that the additional early

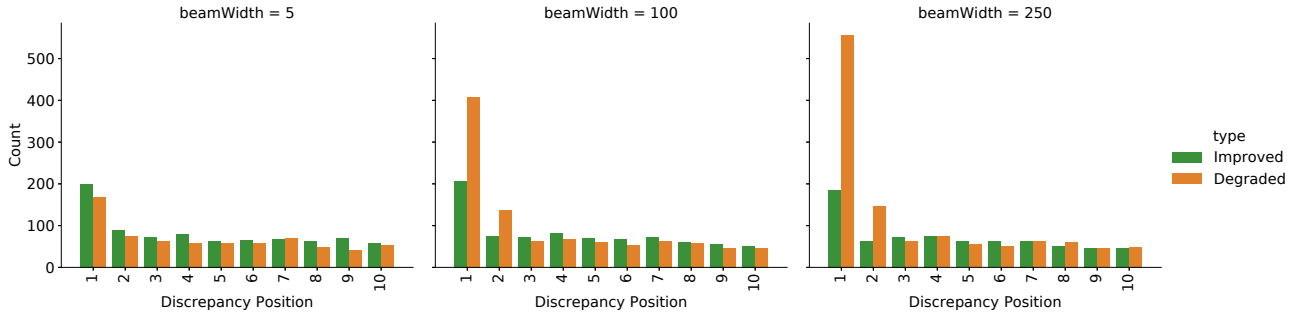


*Figure 4.* WMT'14 En-De: Distribution of discrepancy positions for improved vs. degraded solutions.

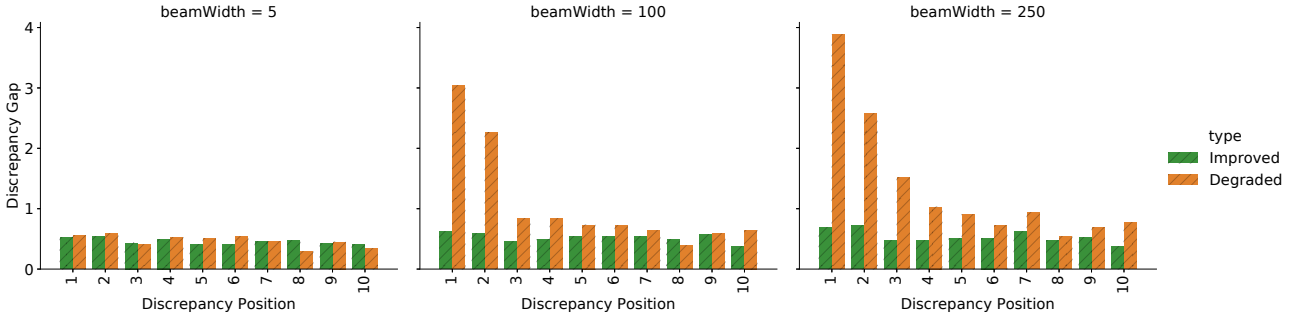


*Figure 5.* WMT'14 En-De: Mean discrepancy gap per position for improved vs. degraded solutions.

discrepancies that are associated with degraded solutions tend to have a much higher discrepancy gap compared to the ones associated with improved solutions. Similar results for the other tasks are reported in Appendix A.

### 4.4. Discrepancies and the Most Likely Hypothesis

In order for a sequence with an early large discrepancy to be selected by a beam search as (approximately) the most likely hypothesis, it must be followed by tokens with higher (conditional) probability. Figure 6 shows the average (conditional) token probability for WMT'14 En-De (we use log-scale on the x axis to highlight the early positions). For larger beams, the average probability of early tokens decreases (due to larger discrepancy gaps) while the average probability of later tokens increases explaining the overall higher probability.[3] Figure 6 also shows the same graph for the improved vs. degraded solutions (compared to greedy search). For improved solutions, we do not see significant change as we increase the beam width. For degraded solutions, however, as we increase the beam width we find more and more early discrepancies that lead to an overall higher probability but a worse evaluation metric value. For all tasks, we found that the changes in the tokens average probability for increased beam width are larger in the case of degraded solutions than for improved solutions (see Appendix A).

Ott et al. (2018) observed the same pattern for copies, i.e., they have low first token probability and higher probabilities for subsequent tokens. Our analysis accounts for this behavior and suggests that copies are one instance of a more general pattern that leads to degraded sequences. In the next section, we show that our analysis generalizes copies, as well as training set predictions in image captioning, and even accounts for additional degraded sequences.

### 4.5. Generalizing Copies and Training Set Predictions

Table 2 shows the number of copies in machine translation and training set predictions in summarization and image captioning. For larger beams, the number of copies and training set predictions grows. Table 2 also reports the mean discrepancy gap of the first token (second token for MSCOCO, see Section 4.2). As our analysis predicts, the early gap of these predictions also grows significantly.

Note that copies and training set predictions only partially

[3] When length normalization is not used, we compare the product of token probabilities rather than the average token probabilities. See Appendix A.3 for results on the unnormalized image captioning task.

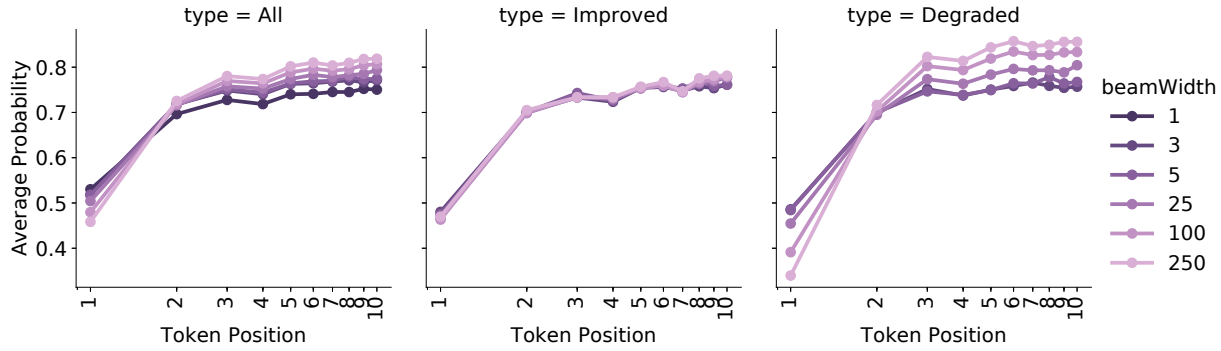


*Figure 6.* Average token probability per position for different beam widths.

*Table 2.* Number of copies and training set examples and the average first token discrepancy gap.

| | | $B$=1 | $B$=3 | $B$=5 | $B$=25 | $B$=100 | $B$=250 |
|---|---|---|---|---|---|---|---|
| En-De | # Copies | 23 | 40 | 49 | 179 | 385 | 567 |
| En-De | First token gap (copies) | 0.0 | 0.12 | 0.28 | 1.79 | 3.05 | 3.71 |
| En-De | First token gap (all) | 0.0 | 0.05 | 0.07 | 0.18 | 0.46 | 0.77 |
| En-Fr | # Copies | 25 | 28 | 41 | 89 | 227 | 358 |
| En-Fr | First token gap (copies) | 0.0 | 0.12 | 0.31 | 1.69 | 3.68 | 4.38 |
| En-Fr | First token gap (all) | 0.0 | 0.04 | 0.05 | 0.10 | 0.32 | 0.60 |
| Gigaword | # Training set predictions | 81 | 86 | 86 | 115 | 163 | 224 |
| Gigaword | First token gap (train pred.) | 0.0 | 0.07 | 0.07 | 0.98 | 1.84 | 2.61 |
| Gigaword | First token gap (all) | 0.0 | 0.12 | 0.12 | 0.29 | 0.39 | 0.55 |
| MSCOCO | # Training set predictions | 163 | 260 | 371 | 588 | 582 | 576 |
| MSCOCO | Second token gap (train pred.) | 0.0 | 0.39 | 0.87 | 1.76 | 1.82 | 1.82 |
| MSCOCO | Second token gap (all) | 0.0 | 0.20 | 0.29 | 0.49 | 0.51 | 0.51 |

account for the performance degradation. In WMT'14 En-De translation with $B$=25, we find that copies account for $\approx 40\%$ of degraded solutions with first token gap. In Gigaword summarization with similar beam width, training set examples account for $\approx 68\%$ of degraded solutions with first token gap. Furthermore, in MSCOCO, since many of the improved sequences are training set captions, eliminating them all together is not desired. Instead, we are interested in avoiding the training captions in the larger beams that led to the performance degradation. These, as Table 2 shows, have a larger difference in the discrepancy gap.

### 4.6. An Illustrative Example

Consider the following example of training set predictions in Gigaword. As we increase the beam width, we find more predictions with the structure: "⟨weekday⟩'s sports scoreboard" (Table 3).[4] As expected, these predictions have a large early discrepancy, followed by highly (conditionally) probable tokens. For $B$=100, the average first token discrepancy gap for these summaries is ≈3.6 compared to ≈0.4 in the full test set. As none of the test references includes "sports scoreboard", these summaries have low evaluation.

*Table 3.* Number of "⟨weekday⟩'s sports scoreboard" predictions.

| $B=3$ | $B=5$ | $B=25$ | $B=100$ | $B=250$ |
|---|---|---|---|---|
| 0 | 1 | 17 | 19 | 19 |

As a potential explanation for this phenomenon, we find that all texts that were summarized as "⟨weekday⟩'s sports scoreboard" included the corresponding weekday. In the training set, we found that in 2962 of the 2971 texts that were summarized to "⟨weekday⟩'s sports scoreboard" included the corresponding weekday. This can lead to the ⟨weekday⟩ token suggested as a first token with a low, but sufficiently high, probability to get into the top $B$ tokens. Followed by high probability tokens, it can, in some cases, have an overall probability that is higher than the alternatives.

## 5. Search Discrepancies and the Performance Degradation in Beam Search

Our baseline results support the observations that performance degradation is a significant problem that occurs across different neural sequence tasks, using different models and evaluation metrics. Consistent with previous results we find this problem even in length-normalized models.[5]

Based on our empirical analysis, we hypothesize that large search discrepancies are the cause of the previously reported performance degradation in beam search. To test this hypothesis, we modify the beam search algorithm to prevent it from considering large discrepancies, with the prediction that it will eliminate the observed performance degradation.

## 6. Discrepancy-Constrained Beam Search

We evaluate two heuristic methods of constraining the beam search from considering large search discrepancies.

**Discrepancy gap:** Given a threshold $\mathcal{M}$, we modify beam search to only consider candidates with a discrepancy gap smaller or equal to $\mathcal{M}$. Formally, we modify Eq. 1 to include the constraint

$$\max_{y \in \mathcal{V}} logP_\theta(y|\mathrm{x};\{y_0,...,y_{t-1}\}) - logP_\theta(y_t|\mathrm{x};\{y_0...y_{t-1}\}) \leq \mathcal{M}$$

[4] Without length normalization, the numbers are higher as this sequence is shorter than most summaries.

[5] We further show that this problem is not due to length bias in Appendix D.

**Beam candidate rank:** Given a threshold $\mathcal{N}$, we modify $\mathcal{Y}_t$ to only include the top $\mathcal{N}$ one-token extensions in each beam. Note that the beam search still retains the top $B$ candidates, however it will not consider more than $\mathcal{N}$ candidates from the each beam.

Using the setup in Section 4.1, we compare these methods to the baseline. Although the analysis in Section 4 was done on the test set (to account for the performance degradation that was previously observed on the test set), $\mathcal{M}$ and $\mathcal{N}$ are tuned on a held-out validation set and no information from the test set was used to tune our methods.

As shown in Table 4, both methods significantly reduce, and in some cases completely eliminate, the performance degradation. In machine translation and summarization, we improve performance compared to the baseline with the best test beam width. In general, the discrepancy gap constraint seems to perform better (most notably, for MSCOCO). The gap constraint allows for a finer-grained control over the accepted search discrepancies, however the rank constraint is simpler and easier to tune.

Ott et al. (2018) proposed to add an inference constraint that prunes copies in the beam search and showed that it significantly reduces the performance degradation in machine translation. However, their empirical analysis still found a drop of approximately a point in the BLEU evaluation for $B = 200$, consistent with our observation that copy predictions do not fully account for the performance degradation in machine translation (Section 4.5). Our inference constraints, more general and not limited to copies, completely eliminate the performance degradation (and even slightly improve the evaluation) in machine translation.

We compared the number of copies and training set predictions in the baseline vs. the two discrepancy-constrained variants of beam search. We find that our methods reduce the growth in the number of both copies and training set predictions, supporting the claim that our hypothesis is a generalization of the previous explanations. The detailed comparison can be found in Appendix B.

We also repeated the analysis above and found that both constrained beam search variations substantially reduce the discrepancy phenomena detailed in Section 4. Complete results for both constrained methods on WMT'14 En-De are in Appendix C (other tasks exhibited similar results).

Finally, to show that these results are not limited to text generated in English or European languages that share some similarity to English, we performed experiments on the WMT'17 En-Zh dataset (that involves generating translations in Chinese). We observe a similar performance degradation, associated with large early discrepancies. Furthermore, the performance degradation on this dataset is not due to copies, as there are none in the translations. Our constrained variants of beam search successfully eliminate the performance degradation and lead to improved evaluation. We report the results in Appendix F.

## 7. Discussion

Our results show that larger beam width leads to increasingly large early discrepancies. These very unlikely early tokens are later compensated by subsequent tokens with a much higher (conditional) probability compared to the subsequent tokens of the more probable early tokens. The large difference in the conditional probability of the subsequent tokens is at the heart of the observed performance degradation. Previous work has highlighted two potential biases that can account for this difference. *Exposure bias* (Ranzato et al., 2016) occurs since the model is only exposed to the

*Table 4.* A comparison of the baseline results vs. the constrained beam search methods (higher values are better, best baseline results in bold).

| Dataset | Method | Threshold | $B$=1 | $B$=3 | $B$=5 | $B$=25 | $B$=100 | $B$=250 |
|---|---|---|---|---|---|---|---|---|
| En-De (BLEU-4) | Baseline | | 25.27 | 26.00 | **26.11** | 25.11 | 23.09 | 21.38 |
| | Constr. Gap | $\mathcal{M} = 1.5$ | 25.27 | 26.00 | 26.18 | 26.18 | 26.22 | 26.29 |
| | Constr. Rank | $\mathcal{N} = 2$ | 25.27 | 26.07 | 26.01 | 26.08 | 26.10 | 26.10 |
| En-Fr (BLEU-4) | Baseline | | 40.15 | 40.77 | **40.83** | 40.52 | 38.64 | 35.03 |
| | Constr. Gap | $\mathcal{M} = 2.0$ | 40.15 | 40.78 | 40.86 | 40.98 | 41.05 | 41.06 |
| | Constr. Rank | $\mathcal{N} = 3$ | 40.15 | 40.77 | 40.81 | 40.99 | 41.05 | 41.02 |
| Gigaword (R-1 F) | Baseline | | 33.56 | **34.22** | 34.16 | 34.01 | 33.67 | 33.23 |
| | Constr. Gap | $\mathcal{M} = 0.85$ | 33.56 | 34.27 | 34.29 | 34.43 | 34.33 | 34.32 |
| | Constr. Rank | $\mathcal{N} = 2$ | 33.56 | 34.48 | 34.45 | 34.25 | 34.23 | 34.32 |
| MSCOCO (BLEU-4) | Baseline | | 29.66 | **32.36** | 31.96 | 30.04 | 29.87 | 29.79 |
| | Constr. Gap | $\mathcal{M} = 0.45$ | 29.66 | 32.24 | 32.33 | 32.36 | 32.35 | 32.35 |
| | Constr. Rank | $\mathcal{N} = 2$ | 29.66 | 32.52 | 31.97 | 30.88 | 30.87 | 30.87 |

training data and can be biased towards the training set distribution (our illustrative example demonstrates such bias due to a repetitive pattern in the training data). *Label bias* (Wiseman & Rush, 2016) occurs since token probabilities at each time step are locally normalized and therefore the successors of incorrect histories receive the same probability mass as the successors of a correct history.

These biases help explain the observed behavior with large beam width: a biased (conditional) probability that concentrates high probability mass on one token and is locally normalized to sum to one compensates for earlier low probability tokens. The negative effects of these biases have been discussed before (Ranzato et al., 2016; Wiseman & Rush, 2016), however the connection to the performance degradation in beam search and the explanatory framework to allow such analysis is, to our best knowledge, novel.

While beam search is used to perform (approximate) inference, it is a heuristic search algorithm (Bisiani, 1987). We therefore believe it is natural to address the performance degradation from a heuristic search perspective. Our analysis, based on the search discrepancy concept from heuristic and combinatorial search, views the probabilities predicted by the neural network as a heuristic value to guide search for a solution. Early mistakes have been shown to have a large negative effect on search performance (Gent & Walsh, 1994) and substantial work has analyzed and proposed techniques to mitigate the phenomenon (Gomes et al., 2005; Cohen & Beck, 2018), including limited discrepancy search (Harvey & Ginsberg, 1995). Further investigation of the connection between such work and neural sequence decoding may lead to further insight.

In this work, we study the previously reported beam search performance degradation in the most commonly used neural sequence models that are based on an RNN decoder and are trained to maximize the word-level likelihood, conditioned on the input sequence and the reference history (Sutskever et al., 2014; Bahdanau et al., 2014). Recently, there have been several proposals for alternative models and training schemes that rely on sequence level losses and that can potentially mitigate the effects of the biases described above (Ranzato et al., 2016; Wiseman & Rush, 2016; Edunov et al., 2018). These works only report the results for small beam widths, and it is not clear if they reduce, or even eliminate, the performance degradation in beam search. Analyzing the performance of these models for larger beam widths, and the associated distribution of search discrepancies in these models is a direction for future work. We are also interested in analyzing the recent Transformer model (Vaswani et al., 2017).

## 8. Related Work

Search discrepancies have been the base of many search techniques in combinatorial search and optimization (e.g., Harvey & Ginsberg, 1995; Walsh, 1997; Beck & Perron, 2000). Furcy & Koenig (2005) proposed BULB, a complete variant of beam search that backtracks based on search discrepancies, as a memory-efficient alternative to best-first heuristic search for path-finding problems.

Several works have modified or constrained beam search for different purposes. Vijayakumar et al. (2018) changed the search objective to allow diverse decoding. Hokamp & Liu (2017) proposed grid beam search to support lexical constraints. Anderson et al. (2017) proposed a constrained beam search that forces inclusion of selected tokens in the output. Freitag & Al-Onaizan (2017) analyzed pruning techniques for a beam search decoder in machine translation. Their strategy of limiting "maximum candidates per node" is similar to the rank constraint in our work, however their analysis is focused on speeding up beam search rather than addressing the phenomenon of performance degradation.

A recent line of work in machine translation suggested the performance degradation is due to length bias (Yang et al., 2018; Murray & Chiang, 2018). For larger beams, an end-of-sentence token with a lower probability that leads to an overall more probable hypothesis is more likely to be considered by the beam search. However, we showed performance degradation above even when using length normalization and in tasks where length bias does not appear (see Appendix D for more details).

## 9. Conclusion

In this work, we perform an empirical analysis of the performance degradation in beam search across three neural sequence decoding tasks. We find that the performance degradation for large beam widths is associated with increasing number of early and large search discrepancies. We hypothesize that the fact that beam search exhibits large discrepancies is the cause of the performance degradation and that avoiding such discrepancies will eliminate the performance degradation. We show that this hypothesis generalizes previous results including the existence of copy predictions in machine translation and the training set predictions in image captioning, and accounts for additional degraded sequences. To validate this hypothesis, we show that methods that prevent the search from considering large search discrepancies eliminate the performance degradation in beam search.

## Acknowledgements

We thank the anonymous reviewers for their valuable feedback. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

## References

Anderson, P., Fernando, B., Johnson, M., and Gould, S. SPICE: Semantic propositional image caption evaluation. pp. 382–398, 2016.

Anderson, P., Fernando, B., Johnson, M., and Gould, S. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, pp. 936–945, 2017.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Beck, J. C. and Perron, L. Discrepancy-bounded depth first search. In *CPAIOR*, pp. 8–10, 2000.

Bisiani, R. Beam search. In Shapiro, S. (ed.), *Encyclopedia of Artificial Intelligence*, pp. 56–58. Wiley & Sons, 1987.

Chopra, S., Auli, M., and Rush, A. M. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL-HLT*, pp. 93–98, 2016.

Cohen, E. and Beck, J. C. Local minima, heavy tails, and search effort for GBFS. In *IJCAI*, pp. 4708–4714, 2018.

Edunov, S., Ott, M., Auli, M., Grangier, D., et al. Classical structured prediction losses for sequence to sequence learning. In *NAACL*, pp. 355–364, 2018.

Freitag, M. and Al-Onaizan, Y. Beam search strategies for neural machine translation. In *WNMT*, pp. 56–60, 2017.

Furcy, D. and Koenig, S. Limited discrepancy beam search. In *IJCAI*, pp. 125–131, 2005.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. Convolutional sequence to sequence learning. In *ICML*, pp. 1243–1252, 2017.

Gent, I. P. and Walsh, T. Easy problems are sometimes hard. *Artificial Intelligence*, 70(1-2):335–345, 1994.

Gomes, C. P., Fernández, C., Selman, B., and Bessière, C. Statistical regimes across constrainedness regions. *Constraints*, 10(4):317–337, 2005.

Graff, D., Kong, J., Chen, K., and Maeda, K. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4 (1):34, 2003.

Harvey, W. D. and Ginsberg, M. L. Limited discrepancy search. In *IJCAI*, pp. 607–615, 1995.

Hokamp, C. and Liu, Q. Lexically constrained decoding for sequence generation using grid beam search. In *ACL*, pp. 1535–1546, 2017.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pp. 3128–3137, 2015.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. Opennmt: Open-source toolkit for neural machine translation. *ACL (System Demonstrations)*, pp. 67–72, 2017.

Koehn, P. and Knowles, R. Six challenges for neural machine translation. In *WNMT*, pp. 28–39, 2017.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

Lin, C.-Y. and Och, F. J. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING*, pp. 501, 2004.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

Murray, K. and Chiang, D. Correcting length bias in neural machine translation. *arXiv preprint arXiv:1808.10006*, 2018.

Ott, M., Auli, M., Grangier, D., and Ranzato, M. Analyzing uncertainty in neural machine translation. In *ICML*, pp. 3953–3962, 2018.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pp. 311–318, 2002.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. Sequence level training with recurrent neural networks. In *ICLR*, 2016.

Rush, A. M., Chopra, S., and Weston, J. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *ACL*, pp. 1715–1725, 2016.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. Nematus: a toolkit for neural machine translation. In *EACL (Software Demonstrations)*, pp. 65–68, 2017.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *NIPS*, pp. 3104–3112, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. CIDEr: Consensus-based image description evaluation. In *CVPR*, pp. 4566–4575, 2015.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D. J., and Batra, D. Diverse beam search for improved description of complex scenes. In *AAAI*, 2018.

Vinyals, O. and Le, Q. V. A neural conversational model. In *ICML Deep Learning Workshop*, 2015.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE TPAMI*, 39(4):652–663, 2017.

Walsh, T. Depth-bounded discrepancy search. In *IJCAI*, pp. 1388–1393, 1997.

Wiseman, S. and Rush, A. M. Sequence-to-sequence learning as beam-search optimization. In *EMNLP*, pp. 1296–1306, 2016.

Yang, Y., Huang, L., and Ma, M. Breaking the beam search curse: A study of (re-) scoring methods and stopping criteria for neural machine translation. In *EMNLP*, pp. 3054–3059, 2018.