

# An Empirical Investigation of Global and Local Normalization for Recurrent Neural Sequence Models Using a Continuous Relaxation to Beam Search

**Kartik Goyal**  
School of Computer Science  
Carnegie Mellon University  
kartikgo@cs.cmu.edu

**Chris Dyer**  
Deepmind  
cdyer@google.com

**Taylor Berg-Kirkpatrick**  
University of California, San Diego  
tberg@ucsd.edu

## Abstract

Globally normalized neural sequence models are considered superior to their locally normalized equivalents because they may ameliorate the effects of label bias. However, when considering high-capacity neural parametrizations that condition on the whole input sequence, both model classes are theoretically equivalent in terms of the distributions they are capable of representing. Thus, the practical advantage of global normalization in the context of modern neural methods remains unclear. In this paper, we attempt to shed light on this problem through an empirical study. We extend an approach for search-aware training via a continuous relaxation of beam search (Goyal et al., 2017b) in order to enable training of *globally normalized* recurrent sequence models through simple backpropagation. We then use this technique to conduct an empirical study of the interaction between global normalization, high-capacity encoders, and search-aware optimization. We observe that in the context of inexact search, globally normalized neural models are still more effective than their locally normalized counterparts. Further, since our training approach is sensitive to warm-starting with pre-trained models, we also propose a novel initialization strategy based on self-normalization for pre-training globally normalized models. We perform analysis of our approach on two tasks: CCG supertagging and Machine Translation, and demonstrate the importance of global normalization under different conditions while using search-aware training.

## 1 Introduction

Neural encoder-decoder models have been tremendously successful at a variety of NLP tasks, such as machine translation (Sutskever et al., 2014; Bahdanau et al., 2015), parsing (Dyer et al., 2016, 2015), summarization (Rush et al., 2015),

dialog generation (Serban et al., 2015), and image captioning (Xu et al., 2015). With these models, the target sequence is generated in a left-to-right step-wise manner with the predictions at every step being conditioned on the input sequence and the whole prediction history. This long-distance memory precludes exact search for the maximally scoring sequence according to the model and therefore, approximate algorithms like greedy search or beam search are necessary in practice during decoding. In this scenario, it is natural to resort to search-aware learning techniques for these models which makes the optimization objective sensitive to any potential errors that could occur due to inexact search in these models.

This work focuses on comparison between search-aware locally normalized sequence models that involve projecting the scores of items in the vocabulary onto a probability simplex at each step and globally normalized/unnormalized sequence models that involve scoring sequences without explicit normalization at each step. When conditioned on the the full input sequence and the entire prediction history, both locally normalized and globally normalized conditional models should have same expressive power under a high-capacity neural parametrization in theory, as they can both model same set of distributions over all finite length output sequences (Smith and Johnson, 2007). However, locally normalized models are constrained in how they respond to search errors during training since the scores at each decoding step must sum to one. To let a search-aware training setup have the most flexibility, abandoning this constraint may be useful for easier optimization.

In this paper, we demonstrate that the interaction between approximate inference and non-convex parameter optimization results in more robust training and better performance for models with global normalization compared to those with

the more common locally normalized parametrization. We posit that this difference is due to *label bias* (Bottou, 1991) arising from the interaction of approximate search and search-aware optimization in locally normalized models. A commonly understood source of label bias in *locally normalized* sequence models is an effect of conditioning only on partial input (for example, only the history of the input) at each step during decoding (Andor et al., 2016; Lafferty et al., 2001; Wiseman and Rush, 2016). We discuss another potential source of label bias arising from approximate search with locally normalized models that may be present even with access to the full input at each step. To this end, we train search-aware globally and locally normalized models in an end-to-end (sub)-differentiable manner using a continuous relaxation to the discontinuous beam search procedure introduced by Goyal et al. (2017b). This approach requires initialization with a suitable globally normalized model to work in practice. Hence, we also propose an initialization strategy based upon self-normalization for pre-training globally normalized models.

We demonstrate the effect of both sources of label bias through our experiments on two common sequence tasks: CCG supertagging and machine translation. We find that label bias can be eliminated by both, using a powerful encoder, and using a globally normalized model. We observe that global normalization yields performance gains over local normalization and is able to ameliorate label bias especially in scenarios that involve a very large hypothesis space.

## 2 Recurrent Sequence Models and Effects of Normalization

We now introduce the notation that we will use in the remainder of the paper for describing locally and globally normalized neural sequence-to-sequence models. We are interested in the probability of output sequence,  $\mathbf{y}$ , conditioned on input sequence,  $\mathbf{x}$ . Let  $s(\mathbf{x}, y_{1:i-1}, y_i)$  be a non negative score of output label  $y$  at time-step  $i$  for the input  $x$  and the prediction history  $y_{1:i-1}$ , let  $V$  be the label space, and let  $\mathcal{Y}_{\mathbf{x}}$  be the space of all finite sequences for  $\mathbf{x}$ .<sup>1</sup> A neural encoder (e.g. a bidirectional LSTM) encodes information about  $\mathbf{x}$  and a recurrent neural decoder generates the output  $\mathbf{y}$

<sup>1</sup>For notational convenience we suppress the dependence of the score  $s$  on model parameters  $\theta$ .

(typically step-by-step from left-to-right) conditioned on the encoder.

### 2.1 Locally normalized models

Under a locally normalized model  $\mathcal{M}_L$ , the probability of  $\mathbf{y}$  given  $\mathbf{x}$  is:

$$p_{\mathcal{M}_L}(\mathbf{y} | \mathbf{x}) = \prod_{i=1}^n p(y_i | \mathbf{x}, y_{1:i-1}) = \prod_{i=1}^n \frac{s(\mathbf{x}, y_{1:i-1}, y_i)}{Z_{L,i}(\mathbf{x}, y_{1:i-1})}$$

where  $Z_{L,i}(\mathbf{x}, y_{1:i-1}) = \sum_{y \in V} s(\mathbf{x}, y_{1:i-1}, y)$ , is the local normalizer at each time step and  $n$  is the number of prediction steps. Since, the local normalizer is easy to compute, likelihood maximization based training is a standard approach for training these models.

### 2.2 Globally normalized models

In contrast, under a globally normalized model  $\mathcal{M}_G$ , the probability of  $\mathbf{y}$  given  $\mathbf{x}$  is:

$$p_{\mathcal{M}_G}(\mathbf{y} | \mathbf{x}) = \frac{\prod_{i=1}^n s(\mathbf{x}, y_{1:i-1}, y_i)}{Z_G(\mathbf{x})}$$

where  $Z_G(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}} \prod_{i=1}^n s(\mathbf{x}, y_{1:i-1}, y_i)$ , is the global log-normalizer.  $Z_G(\mathbf{x})$  is intractable to estimate for most problems of interest due to the large search space therefore, an exact likelihood maximization training approach is intractable for these models.

### 2.3 Label Bias with partial input

It was shown in Andor et al. (2016); Lafferty et al. (2001), locally normalized conditional models *with access to only partial input*,  $x_{1:i-1}$ , at each decoding step are biased towards labeling decisions with low-entropy transition probabilities at each decoding step and, as a result, suffer from a weakened ability to revise previous decisions based upon future input observations. This phenomenon has been referred to as *label bias*, and presents itself as an arbitrary allocation of probability mass to unlikely or undesirable label sequences despite the presence of well-formed sequences in training data. Andor et al. (2016) prove that this class of locally normalized models that relies on the structural assumption of access to only left-to-right partial input at each step,

$$\prod_{i=1}^n p(y_i | \mathbf{x}, y_{1:i-1}) = \prod_{i=1}^n p(y_i | x_{1:i-1}, y_{1:i-1}),$$

is strictly less expressive than its globally normalized counterpart.

However, the standard sequence-to-sequence models used most often in practice and presented in this paper actually condition the decoder on a summary representation of the *entire input sequence*,  $\mathbf{x}$ , computed by a neural encoder. Hence, depending on the power of the encoder, it is commonly thought that such models avoid this type of label bias. For these models, both locally normalized and globally normalized conditional models are equally expressive, in principle, with a sufficiently powerful encoder.

However, as we suggest in the next section and show empirically in experiments, this does not necessarily mean that both parametrizations are equally amenable to gradient-based training in practice, particularly when the search space is large and search-aware training techniques are used. We will argue that they suffer from a related, but distinct, form of bias introduced by inexact decoding.

## 2.4 Search-aware training

To improve performance with inexact decoding methods (e.g. beam search), search-aware training techniques take into account the decoding procedure that will be used at test time and adjust the parameters of the model to maximize prediction accuracy under the decoder. Because of the popularity of beam search as a decoding procedure for sequence models, in this paper we focus on beam search-aware training. While many options are available, including beam-search optimization (BSO) (Wiseman and Rush, 2016), in Section 3.1 we will describe the particular search-aware training strategy we use in experiments (Goyal et al., 2017b), chosen for its simplicity.

## 2.5 Label Bias due to approximate search

We illustrate via example how optimization of locally normalized models may suffer from a new kind of label bias when using beam search-aware training, and point to reasons why this issue might be mitigated by the use of globally normalized models. While the scores of successors of a single candidate under a locally normalized model are constrained to sum to one, scores of successors under a globally normalized model need only be positive. Intuitively, during training, this gives the globally normalized model more freedom to

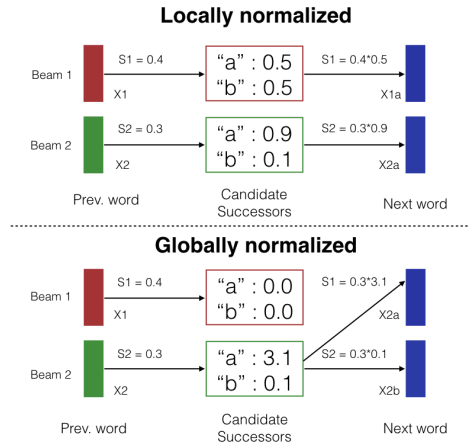


Figure 1: Illustrative example of bias arising in locally normalized models due to beam search. Red indicates the candidate that optimization should learn to discard and green indicates the candidate that should be propagated. Locally normalized models are constrained to return normalized scores for the successors of each candidate, while globally normalized models are unconstrained and can more easily learn to drop successors of the red candidate.

downweight undesirable intermediate candidates in order avoid search errors.

In the example beam search decoding problem in Figure 1, we compare the behavior of locally and globally normalized models at a single time step for a beam size of two. In this example, we assume that the score for beams in both the models is exactly the same until the step shown in Figure 1. Suppose that the lower item on the beam ( $X_2$ ) is correct, and thus, for more effective search, we would prefer the models scores to be such that only successors of the lower beam item are present on the beam at the next step. However, since, the scores at each step for a locally normalized model are constrained to sum to one, the upper beam item ( $X_1$ ) generates successors with scores comparable to those of the lower beam item. As we see in the example, due to the normalization constraint, search-aware training of the locally normalized model might find it difficult to set the parameters to prevent extension of the poorer candidate. In contrast, because the scores of a *globally normalized* model are not constrained to sum to one, the parameters of the neural model can be set such that *all* the successors of the bad candidate have a very low score and thus do not compete for space on the beam. This illustrates a mechanism by which search-aware training of globally normalized models in a large search spaces might

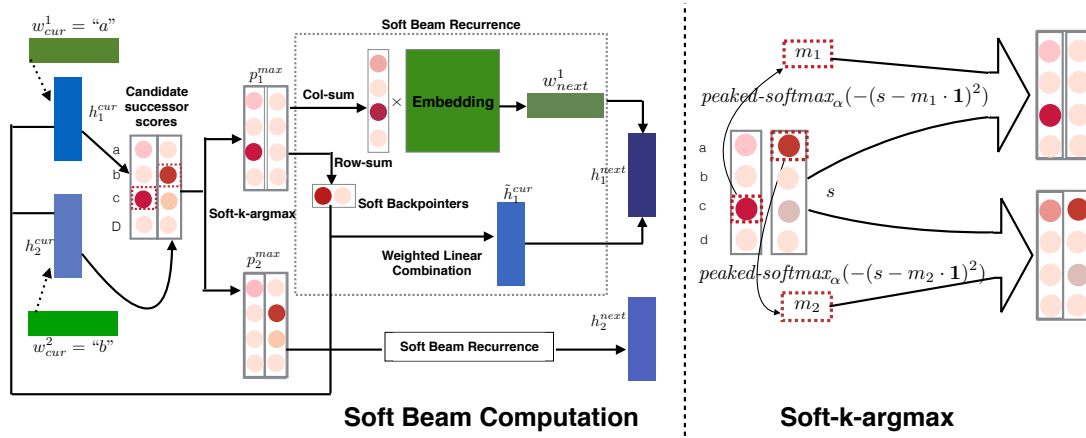


Figure 2: **Left:** Computing LSTM hidden states at a subsequent step using continuous relaxation to beam search for beam size of 2. **Right:** Illustration of top-k-argmax.  $m_1$  and  $m_2$  are the maximum and second maximum candidate scores.

be more effective. However as discussed earlier, if we can perform *exact search* then this *label bias* ceases to exist because both the models have the same expressive power with a search-agnostic optimization scheme. In experiments, we will explore this trade-off empirically.

### 3 Search-aware Training for Globally Normalized Models

In order to conduct an empirical study with meaningful comparisons, we devise an extension of the relaxed beam-search based optimization proposed by Goyal et al. (2017b) that allows us to train both the search-aware globally and locally normalized models in a similar manner with the same underlying architecture.

#### 3.1 Continuous Relaxation to Beam Search

Following Goyal et al. (2017b), we train a beam-search aware model by optimizing a continuous surrogate approximation to a *direct loss* objective,  $J$ , defined as a function of the output of beam search and the ground truth sequence  $\mathbf{y}^*$ :

$$\min_{\theta} J(\mathbf{x}, \theta, \mathbf{y}^*) = \min_{\theta} \ell(\text{Beam}(\mathbf{x}, \mathcal{M}(\theta)), \mathbf{y}^*)$$

Here  $\ell$  is a function that computes the loss of the model’s prediction produced by beam search, and  $\mathcal{M}$  refers to the model parametrized by  $\theta$ . While this objective is search-aware, it is discontinuous and difficult to optimize because beam search involves discrete *k-argmax* operations. Therefore, Goyal et al. (2017b) propose a continuous surrogate,  $\tilde{J}$ , by defining a continuous approximation (*soft-k-argmax*) of the discrete *k-argmax* and using

this to compute an approximation to a composition of the loss function and the beam search function.

$$\min_{\theta} \tilde{J}(\mathbf{x}, \theta, \mathbf{y}^*) \approx \min_{\theta} (\ell \circ \text{Beam})(\mathbf{x}, \mathcal{M}(\theta), \mathbf{y}^*)$$

The *soft-k-argmax* procedure involves computing distances between the scores of the successors and the  $k^{\text{th}}$ -max score and using the temperature based argmax operation (Maddison et al., 2017; Jang et al., 2016; Goyal et al., 2017a) to get an output peaked on the  $k^{\text{th}}$ -max value as shown in the right panel of Figure 2. The temperature is a hyperparameter which is typically annealed toward producing low entropy distributions during optimization. As shown in the left panel of Figure 2, the *soft candidate vectors* and the *soft backpointers* are computed at every decoding step using this *soft-k-argmax* operation in order to generate the embeddings and recurrent hidden states of the LSTM at each step of the *soft beam search* procedure. With a locally decomposable loss like Hamming loss, both *soft loss* and *soft scores* for the relaxed procedure are iteratively computed so that the end-to-end objective computation can be described by a computation graph that is amenable to backpropagation.

Using this relaxation, point-wise convergence of the surrogate objective to the original objective can be established ( $\alpha$  is the inverse temperature):

$$\tilde{J}_{\alpha}(\mathbf{x}, \theta, \mathbf{y}^*) \xrightarrow[\rho]{\alpha \rightarrow \infty} J(\mathbf{x}, \theta, \mathbf{y}^*)$$

Goyal et al. (2017b) demonstrated empirically that optimizing the surrogate objective,  $\tilde{J}$  – which can be accomplished via simple backpropagation for decomposable losses like Hamming distance – leads to improved performance at test time.

In experiments, for training locally normalized models, we use *log-normalized* successor scores. However, for training globally normalized models, we will directly use *unnormalized scores*, which are  $\in \mathbb{R}_+$ .

### 3.2 Initialization for training globally normalized models

Goyal et al. (2017b) reported that *initialization* with a locally normalized model pre-trained with teacher-forcing was important for their continuous beam search based approach to be stable and hence they used the locally normalized log-scores for their search-aware training model. In this work, we experimented with the unnormalized candidate successor scores and found that initializing the optimization for a globally normalized objective with a cross-entropy trained locally normalized model resulted in unstable training. This is expected because the locally normalized models are parametrized in a way such that using the scores before the softmax normalization results in a very different outcome than using scores after local normalization. For example, the locally normalized Machine Translation model in Table 1, that gives a BLEU score of 27.62 when decoded with beam search using locally normalized scores, results in BLEU of 4.30 when beam search decoding is performed with unnormalized scores. Pre-training a truly globally normalized model for initialization is not straightforward because no exact likelihood maximization techniques exist for globally normalized models as the global normalizer is intractable to compute.

Therefore, we propose a new approach to initialization for search-aware training of globally normalized models: we pre-train a locally normalized model that is parametrized like a globally normalized model. More specifically, we train a locally normalized model with its distribution over the output sequences denoted by  $p_L(\mathcal{Y})$  such that we can easily find a globally normalized model with a distribution  $p_G(\mathcal{Y})$  that matches  $p_L(\mathcal{Y})$ . Following the notation in Section 2, for a locally normalized model, the log-probability of a sequence is:

$$\sum_{i=1}^n [\log s(\mathbf{x}, y_{1:i-1}, y_i) - \log Z_{L,i}(\mathbf{x}, y_{1:i-1})]$$

and for a globally normalized model it is:

$$\left[ \sum_{i=1}^n \log s(\mathbf{x}, y_{1:i-1}, y_i) \right] - \log Z_G(\mathbf{x})$$

#### 3.2.1 Self Normalization

One way to find a locally normalized model that is parametrized like a globally normalized model is to ensure that the local normalizer at each step,  $\log Z_{L,i}(\mathbf{x}, y_{1:i-1})$ , is 0. With the local normalizer being zero it is straightforward to see that the log probability of a sequence under a locally normalized model can easily be interpreted as log probability of the sequence under a globally normalized model with the global log-normalizer,  $\log Z_G(\mathbf{x}) = 0$ . This training technique is called *self-normalization* (Andreas and Klein, 2015) because the resulting models’ unnormalized score at each step lies on a probability simplex. A common technique for training self-normalized models is L2-regularization of local log normalizer which encourages learning a model with  $\log Z = 0$  and was found to be effective for learning a language model by Devlin et al. (2014)<sup>2</sup>. The L2-regularized cross entropy objective is given by:

$$\min_{\theta} \sum_{\mathbf{x}, \mathbf{y}^* \in \mathcal{D}} - \sum_{i=1}^n \log p(y_i^* | \mathbf{x}, y_{1:i-1}) + \lambda \cdot (\log Z_{L,i}(\mathbf{x}, y_{1:i-1}))^2$$

In Table 1, we report the mean and variance of the local log normalizer on the two different tasks using L2-regularization (*L2*) based self normalization and no self normalization (*CE*). We observe that *L2* models are competitive performance-wise to the cross-entropy trained locally normalized models while resulting in a much smaller local log-normalizer on average. Although, we couldn’t minimize  $\log Z$  exactly to 0, we observe in Section 4 that this is sufficient to train a reasonable initializer for the search-aware optimization of globally normalized models. It is important to note that these approaches yield a *globally normalized* model that is equivalent to a locally normalized model trained via teacher-forcing and hence these are only used to *warm-start* the search-aware optimization of globally normalized models. Our search-aware training approach is

<sup>2</sup>Noise Contrastive Estimation (Mnih and Teh, 2012; Gutmann and Hyvärinen, 2010) is also an alternative to train unnormalized models but our experiments with NCE were unstable and resulted in worse models.

free to adjust the parameters of the models such that the final globally normalized model has a non-zero log-normalizer  $Z_G$  over the data.

		Train logZ		Dev logZ		Acc/ BLEU
		Mean	Var	Mean	Var	
CCG	CE	21.08	9.57	21.96	9.18	93.3
	L2	0.6	0.29	0.26	0.08	91.9
MT	CE	24.7	115.4	25.8	129.1	27.62
	L2	0.65	0.18	0.7	0.29	26.63

Table 1: Comparison of logZ between cross entropy trained models (CE) and self normalized models (L2) for CCG supertagging and Machine Translation tasks.

Other possible approaches to project locally normalized models onto globally normalized models include distribution matching via knowledge distillation (Hinton et al., 2015). We leave exploration of warm-starting of search aware optimization with this approach to future work.

## 4 Experiments and Empirical Analysis

To empirically analyze the interaction between label bias arising from different sources, search-aware training, and global normalization, we conducted experiments on two tasks with vastly different sizes of output space: CCG supertagging and Machine Translation. As described in the next section, the task of tagging allows us to perform controlled experiments which explicitly study the effect of amount of input information available to the decoder at each step, we analyze the scenarios in which search aware training and global normalization are expected to improve the model performance.

In all our experiments, we report results on training with standard teacher forcing optimization and self-normalization as our baselines. We report results with both search-aware locally and globally normalized models (Section 3.1) after warm starting with both cross entropy trained models and self-normalized models to study the effects of search-aware optimization and global normalization. We follow Goyal et al. (2017b) and use the decomposable Hamming loss approximation with search-aware optimization for both the tasks and decode via *soft beam search decoding* method which involves continuous beam search with soft backpointers for the LSTM Beam search dynamics as described in Section 3, but using identifiable backpointers and labels (using MAP estimates of soft backpointers and labels) to decode.

We tune hyperparameters like learning rate and annealing schedule by observing performance on development sets for both the tasks. We performed at least three random restarts for each class and report results based on best development performance.

### 4.1 CCG supertagging

We used the standard splits of CCG bank (Hockenmaier and Steedman, 2002) for training, development, and testing. The label space of supertags is 1,284 and the labels are correlated with each other based on their syntactic relations. The distribution of supertag labels in the training data exhibits a long tail distribution. This task is sensitive to the long range sequential decisions because it encodes rich syntactic information about the sentence. Hence, this task is ideal to analyze the effects of label bias and search effects. We perform minor preprocessing on the data similar to the preprocessing in Vaswani et al. (2016). For experiments related to search aware optimization, we report results with beam size of 5.<sup>3</sup>

#### 4.1.1 Tagging model for ablation study

We changed the standard sequence-to-sequence model to be more suitable for the tagging task. This change also lets us perform controlled experiments pertaining to the amount of input sequence information available to the decoder at each time step.

In a standard encoder-decoder model with attention, the initial hidden state of the decoder is often some function of the final encoder state so that the decoder’s predictions can be conditioned on the full input. For our tagging experiments, instead of influencing the initial decoder state with the encoder, we set it to a vector of zeros. Thus the information about input for prediction is *only* available via the attention mechanism. In addition to the change above, we also forced the model to attend to only the  $i^{th}$  input representation while predicting the  $i^{th}$  label. This is enforceable because the output length is equal to the input length and it is also a more suitable structure for a tagging model. With these changes in the decoder, we can precisely control the amount of information about the input available to the decoder at each prediction step. For example, with a unidirectional LSTM encoder, the decoder at  $i^{th}$  step

<sup>3</sup>We observed similar results with beam size 10

only has access to input till the  $i^{th}$  token and the prediction history:

$$p(y_i | \mathbf{x}, y_{1:i-1}) = p(y_i | x_{1:i}, y_{1:i-1})$$

This setting lets us clearly explore the classical notion of *label bias* arising out of access to partial input at each prediction step (Section 2.3). A bidirectional LSTM encoder, however provides access to all of the input information to the decoder at all the prediction steps.

	Unidirectional	Bidirectional
pretrain-greedy	76.54	92.59
pretrain-beam	77.76	93.29
locally normalized	83.9	<b>93.76</b>
globally normalized	<b>83.93</b>	93.73

Table 2: **Accuracy results on CCG supertagging when initialized with a regular teacher-forcing model.** Reported using *Unidirectional* and *Bidirectional* encoders respectively with fixed attention tagging decoder. *pretrain-greedy* and *pretrain-beam* refer to the output of decoding the initializer model. *locally normalized* and *globally normalized* refer to search-aware soft-beam models

	Unidirectional	Bidirectional
pretrain-greedy	73.12	91.23
pretrain-beam	73.83	91.94
locally normalized	83.35	<b>92.78</b>
globally normalized	<b>85.50</b>	92.63

Table 3: **Accuracy results on CCG supertagging when initialized with a self normalized model.**

## 4.2 Machine Translation

We use the same dataset (the German-English portion of the IWSLT 2014 machine translation evaluation campaign (Cettolo et al., 2014)), preprocessing and data splits as Ranzato et al. (2016) for our Machine Translation experiments. The output label/vocabulary size is 32000 and unlike tagging, the length of output sequences cannot be deterministically determined from the length of the input sequence. Moreover, the output sequence does not necessarily align monotonically with the input sequence. Hence the output sequence space for MT is much larger than that for tagging and the effects of inexact search on optimization are expected to be even more apparent for MT. We use a standard LSTM-based encoder/decoder model with a standard attention mechanism (Bahdanau et al., 2016) for our MT experiments. For search-aware opti-

Init-scheme $\rightarrow$	Regular	Self-normalized
pretrain-greedy	26.24	25.42
pretrain-beam	27.62	26.63
locally-normalized	<b>29.28</b>	27.71
globally-normalized	26.24	<b>29.27</b>

Table 4: **BLEU results on de-en Machine Translation.** *Regular* and *Self-normalized* refer to the initialization scheme for soft beam search training. *pretrain-greedy* and *pretrain-beam* refer to the output of decoding the initializer model. *locally normalized* and *globally normalized* refer to search-aware soft-beam models

mization experiments, we report results with beam size 3.<sup>4</sup>

## 4.3 Results and Analysis

The results reported in Tables 2, 3 and 4 allow us to analyze the effect of interaction of label bias, inexact search and global normalization in detail.

### 4.3.1 Label bias with partial input

First, we analyze the effect of label bias that arises from conditioning on partial input (Section 2.3) during decoding on optimization of the models. The unidirectional encoder based tagging experiments suggest that conditioning on partial input during decoding results in poor models when trained with cross entropy based methods. Interestingly, all techniques improve upon this: (i) search-aware locally and globally normalized models are able to train for accuracy directly and eliminate exposure bias that arises out of the mismatch between train-time and test-time prediction methods, and, (ii) the bidirectional tagging model which provides access to all of the input is powerful enough to learn a complex relationship between the decoder and the input representations for the search space of the CCG supertagging task and results in a much better performance.

### 4.3.2 Initialization of search-aware training

Next, we analyze the importance of appropriate initialization of search-aware optimization with pretrained models. Across all the results in Tables 2, 3 and 4, we observe that search-aware optimization for locally normalized models always improves upon the pre-trained locally

<sup>4</sup>We observed similar results beam size of 5.

normalized models used for initialization. But when the search-aware optimization for globally normalized models is initialized with locally normalized CE models, the improvement is not as pronounced and in the case of MT, the performance is actually *hurt* by the improper initialization for training globally normalized models – probably a consequence of large search space associated with MT and incompatibility between unnormalized scores for search-aware optimization and locally normalized scores of the CE model used for pre-training. When the *self-normalized* models are used for initialization, optimization for globally normalized models always improves upon the pre-trained self-normalized model. It is interesting to note that we see improvements for the globally normalized models even when  $\log Z$  is not exactly reduced to 0 indicating that the scores used for search-aware training initially are comparable to the scores of the pre-trained self-normalized model. We also observe that self-normalized models perform slightly worse than CE-trained models but search aware training for globally normalized models improves the performance significantly.

### 4.3.3 Search-aware training

Next, we analyze the effect of search-aware optimization on the performance of the models. Search-aware training with locally normalized models improves the performance significantly in *all* our experiments which indicates that accounting for exposure bias and optimizing for predictive performance directly is important. We also observe that the bidirectional model for tagging is quite powerful and seems to account for both *exposure bias* and *label bias* to a large extent. We reckon that this may be because the greedy decoding itself is very close to *exact search* for this well-trained tagging model over a search space that is much simpler than that associated with MT. Therefore, the impact of search-aware optimization on the bidirectional tagger is marginal. However, it is much more pronounced on the task of MT.

### 4.3.4 Global normalization and label bias

We analyze the importance of training globally normalized models. In the specific setup for tagging with the unidirectional encoder, globally

normalized models are actually *more expressive* than the locally normalized models (Andor et al., 2016) as described in Section 2.3 and this is reflected in our experiments (table 3) with tagging. The globally normalized model (warm started with a self-normalized model) performs the best among all the models in the unidirectional tagger case which indicates that it is ameliorating something beyond exposure bias which is fixed by the search-aware locally normalized model. For MT (table 4), both globally normalized and locally normalized models are equally expressive in theory because the decoder is conditioned on the full input information at each step, but we still observe that the globally normalized model improves significantly over the self-normalized pre-trained model and the search-aware locally normalized model. This indicates that it might be ameliorating the label bias associated with inexact search (discussed in Section 2.5). As discussed in Section 3.2, the globally normalized model, when initialized with a CE trained model, performs worse because of improper initialization of the search aware training. The self-normalized model starts off 1 BLEU point worse than the CE model point but global normalization, initialized with the self-normalized model improves the performance and is competitive with the best model for MT. This suggests that a better technique for initializing the optimization for globally normalized models should be helpful in improving the performance.

### 4.3.5 Global normalization and sentence length

In tables 5 and 6, we analyze the source of improvement from global normalization for MT. In table 5, we report the ngram overlap scores and ratio of length of the predictions to length of hypothesis for the case when the search-aware training is initialized with a self-normalized model. We observe that the globally normalized model produces longer predictions than the locally normalized model. More interestingly, it seems to have better 3 and 4-gram overlap and slightly worse unigram and bigram overlap score than the locally normalized model. These observations suggest that globally normalized models are better able to take longer range effects into account and are also cautious about predicting the *end-of-sentence* symbol too soon. Moreover, in table 6, we observe



	N-gram overlap	Length ratio
pretrain-beam	63.5/35.7/21.8/13.7	0.931
locally-normalized	66.9/39.4/22.7/14.0	0.918
globally-normalized	65.0/39.1/23.2/14.7	0.959

Table 5: **Breakdown of BLEU results on de-en Machine Translation dev set.** Reported on Self-normalized initialization

Src sent-length →	0-20	20-30	30-40	40+
pretrain-beam	29.36	25.73	24.71	24.50
locally-normalized	32.35	26.95	25.39	25.2
globally-normalized	33.21	28.08	26.75	26.41

Table 6: **BLEU scores with different length inputs on dev set** Reported on Self-normalized initialization. The header specifies the range of length of the input sentences

that globally normalized models perform better on all the length ranges but especially so on long sentences.

## 5 Related Work

Much of the existing work on search-aware training of globally normalized neural sequence models uses some mechanism like early updates (Collins and Roark, 2004) that relies on explicitly tracking if the gold sequence falls off the beam and is not end-to-end continuous. Andor et al. (2016) describe a method for training globally normalized neural feedforward models, which involves optimizing a CRF-based likelihood where the normalizer is approximated by the sum of the scores of the final beam elements. They describe label bias arising out of conditioning on partial input and hence focused on the scenario in which locally normalized models can be less expressive than globally normalized models, whereas we also consider another source of label bias which might be affecting the optimization of equally expressive locally and globally normalized conditional models. Wiseman and Rush (2016) also propose a beam search based training procedure that uses unnormalized scores similar to our approach. Their models achieve good performance over CE baselines – a pattern that we observe in our results as well. In this work, we attempt to empirically analyze the factors affecting this boost in performance with end-to-end continuous search-aware training (Goyal et al., 2017b) for globally normalized models.

Smith and Johnson (2007) proved that locally normalized conditional PCFGs and unnormalized

conditional WCFGs are equally expressive for finite length sequences and posit that Maximum Entropy Markov Models (MEMMs) are weaker than CRFs because of the structural assumptions involved with MEMMs that result in label bias.

Recently, energy based neural structured prediction models (Amos et al., 2016; Belanger and McCallum, 2016; Belanger et al., 2017) were proposed that define an energy function over candidate structured output space and use gradient based optimization to form predictions making the overall optimization search aware. These models are designed to model global interactions between the output random variables without specifying strong structural assumptions.

## 6 Conclusion

We performed empirical analysis to analyze the interaction between label bias, search-aware optimization and global normalization in various scenarios. We proposed an extension to the continuous relaxation to beam search proposed by Goyal et al. (2017b) to train search-aware globally normalized models and comparable locally normalized models. We find that in the context of inexact search over large output spaces, globally normalized models are more effective than the locally normalized models in spite of them being equivalent in terms of their expressive power.

## Acknowledgement

This project is funded in part by the NSF under grant 1618044. We thank the three anonymous reviewers for their helpful feedback.

## References

- Brandon Amos, Lei Xu, and J Zico Kolter. 2016. Input convex neural networks. *arXiv preprint arXiv:1609.07152*.
- Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. 2016. Globally normalized transition-based neural networks. In *Association for Computational Linguistics*.
- Jacob Andreas and Dan Klein. 2015. When and why are log-linear models self-normalizing? In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 244–249.

- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.
- Dzmitry Bahdanau, Dmiry Serdyuk, Philémon Brakel, Nan Rosemary Ke, Jan Chorowski, Aaron Courville, and Yoshua Bengio. 2016. Task loss estimation for structured prediction.
- David Belanger and Andrew McCallum. 2016. Structured prediction energy networks. In *International Conference on Machine Learning*, pages 983–992.
- David Belanger, Bishan Yang, and Andrew McCallum. 2017. End-to-end learning for structured prediction energy networks. *arXiv preprint arXiv:1703.05667*.
- Léon Bottou. 1991. *Une Approche théorique de l'Apprentissage Connexionniste: Applications à la Reconnaissance de la Parole*. Ph.D. thesis, Université de Paris XI, Orsay, France.
- Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam*.
- Michael Collins and Brian Roark. 2004. Incremental parsing with the perceptron algorithm. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 111. Association for Computational Linguistics.
- Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1370–1380.
- Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. *arXiv preprint arXiv:1505.08075*.
- Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. 2016. Recurrent neural network grammars. *arXiv preprint arXiv:1602.07776*.
- Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2017a. Differentiable scheduled sampling for credit assignment. *arXiv preprint arXiv:1704.06970*.
- Kartik Goyal, Graham Neubig, Chris Dyer, and Taylor Berg-Kirkpatrick. 2017b. A continuous relaxation of beam search for end-to-end training of neural sequence models. *arXiv preprint arXiv:1708.00111*.
- Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Julia Hockenmaier and Mark Steedman. 2002. Acquiring compact lexicalized grammars from a cleaner treebank. In *LREC*.
- Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*.
- John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*.
- Andriy Mnih and Yee Whye Teh. 2012. A fast and simple algorithm for training neural probabilistic language models. *arXiv preprint arXiv:1206.6426*.
- Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*.
- Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Empirical Methods in Natural Language Processing*.
- Iulian V Serban, Alessandro Sordani, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI’16 Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*.
- Noah A Smith and Mark Johnson. 2007. Weighted and probabilistic context-free grammars are equally expressive. *Computational Linguistics*, 33(4):477–491.
- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.
- Ashish Vaswani, Yonatan Bisk, Kenji Sagae, and Ryan Musa. 2016. Supertagging with lstms. In *Proceedings of NAACL-HLT*, pages 232–237.
- Sam Wiseman and Alexander M Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Empirical Methods in Natural Language Processing*.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81.