Is MAP Decoding All You Need? The Inadequacy of the Mode in Neural Machine Translation

Bryan Eikema
University of Amsterdam
b.eikema@uva.nl

Wilker Aziz
University of Amsterdam
w.aziz@uva.nl

Abstract

Recent studies have revealed a number of pathologies of neural machine translation (NMT) systems. Hypotheses explaining these mostly suggest there is something fundamentally wrong with NMT as a model or its training algorithm, maximum likelihood estimation (MLE). Most of this evidence was gathered using maximum *a posteriori* (MAP) decoding, a decision rule aimed at identifying the highest-scoring translation, *i.e.* the mode. We argue that the evidence corroborates the inadequacy of MAP decoding more than casts doubt on the model and its training algorithm. In this work, we show that translation distributions do reproduce various statistics of the data well, but that beam search strays from such statistics. We show that some of the known pathologies and biases of NMT are due to MAP decoding and not to NMT's statistical assumptions nor MLE. In particular, we show that the most likely translations under the model accumulate so little probability mass that the mode can be considered essentially arbitrary. We therefore advocate for the use of decision rules that take into account the translation distribution holistically. We show that an approximation to minimum Bayes risk decoding gives competitive results confirming that NMT models do capture important aspects of translation well in expectation.

1 Introduction

Recent findings in neural machine translation (NMT) suggest that modern translation systems have some serious flaws. This is based on observations such as: *i)* translations produced via beam search typically under-estimate sequence length (Sountsov and Sarawagi, 2016; Koehn and Knowles, 2017), the *length bias*; *ii)* translation quality generally deteriorates with better approximate search (Koehn and Knowles, 2017; Murray and Chiang, 2018; Ott et al., 2018; Kumar and Sarawagi, 2019), the *beam search curse*; *iii)* the true most likely translation under the model (*i.e.*, the mode of the distribution) is empty in many cases (Stahlberg and Byrne, 2019) and a general negative correlation exists between likelihood and quality beyond a certain likelihood value (Ott et al., 2018), we call this the *inadequacy of the mode problem*.

A number of hypotheses have been formulated to explain these observations. They mostly suggest there is something fundamentally wrong with NMT as a model (*i.e.*, its factorisation as a product of locally normalised distributions) or its most popular training algorithm (*i.e.*, regularised maximum likelihood estimation, MLE for short). These explanations make an unspoken assumption, namely, that identifying the mode of the distribution, also referred to as maximum *a posteriori* (MAP) decoding (Smith, 2011), is in some sense the obvious decision rule for predictions. While this assumption makes intuitive sense and works well in unstructured classification problems, it is less justified in NMT, where oftentimes the most likely translations together account for very little probability mass, a claim we shall defend conceptually and provide evidence for in experiments. Unless the translation distribution is extremely peaked about the mode for every plausible input, criticising the model in terms of properties of its mode can at best say something about the adequacy of MAP decoding. Unfortunately, as previous research

Proceedings of the 28th International Conference on Computational Linguistics, pages 4506–4520
Barcelona, Spain (Online), December 8-13, 2020

has pointed out, this is seldom the case (Ott et al., 2018). Thus, pathologies about the mode cannot be unambiguously ascribed to NMT as a model nor to MLE, and inadequacies about the mode cannot rule out the possibility that the model captures important aspects of translation well in expectation.

In this work, we criticise NMT models as probability distributions estimated via MLE in various settings: varying language pairs, amount of training data, and test domain. We observe that the induced probability distributions represent statistics of the data well in expectation, and that some length and lexical biases are introduced by approximate MAP decoding. We demonstrate that beam search outputs are rare events, particularly so when test data stray from the training domain. The empty string, shown to often be the true mode (Stahlberg and Byrne, 2019), too is an infrequent event. Finally, we show that samples obtained by following the model's own generative story are of reasonable quality, which suggests we should base decisions on statistics gathered from the distribution holistically. One such decision rule is minimum Bayes risk (MBR) decoding (Goel and Byrne, 2000; Kumar and Byrne, 2004). We show that an approximation to MBR performs rather well, especially so when models are more uncertain.

To summarise: we argue that *i)* MAP decoding is not well-suited as a decision rule for MLE-trained NMT; we also show that *ii)* pathologies and biases observed in NMT are not necessarily inherent to NMT as a model or its training objective, rather, MAP decoding is at least partially responsible for many of these pathologies and biases; finally, we demonstrate that *iii)* a straight-forward approximation to a sampling-based decision rule known as minimum Bayes risk decoding gives good results, showing promise for research into decision rules that take into account the distribution holistically.

2 Observed Pathologies in NMT

Many studies have found that NMT suffers from a *length bias*: NMT underestimates length which hurts the adequacy of translations. Cho et al. (2014a) already demonstrate that NMT systematically degrades in performance for longer sequences. Sountsov and Sarawagi (2016) identify the same bias in a chat suggestion task and argue that sequence to sequence models underestimate the margin between correct and incorrect sequences due to local normalisation. Later studies have also confirmed the existence of this bias in NMT (Koehn and Knowles, 2017; Stahlberg and Byrne, 2019; Kumar and Sarawagi, 2019).

Notably, all these studies employ beam search decoding. In fact, some studies link the length bias to the *beam search curse*: the observation that large beam sizes hurt performance in NMT (Koehn and Knowles, 2017). Sountsov and Sarawagi (2016) already note that larger beam sizes exacerbate the length bias. Later studies have confirmed this connection (Blain et al., 2017; Murray and Chiang, 2018; Yang et al., 2018; Kumar and Sarawagi, 2019). Murray and Chiang (2018) attribute both problems to local normalisation which they claim introduces label bias (Lafferty et al., 2001) to NMT. Yang et al. (2018) show that likelihood negatively correlates with translation length. These findings suggest that the mode might suffer from length bias, likely thereby failing to sufficiently account for adequacy. In fact, Stahlberg and Byrne (2019) show that oftentimes the true mode is the empty sequence.

The connection with the length bias is not the only reason for the beam search curse. Ott et al. (2018) find that the presence of copies in the training data cause the model to assign too much probability mass to copies of the input, and that with larger beam sizes this copying behaviour becomes more frequent. Cohen and Beck (2019) show that translations obtained with larger beam sizes often consist of an unlikely prefix with an almost deterministic suffix and are of lower quality. In open-ended generation, Zhang et al. (2020) correlate model likelihood with human judgements for a fixed sequence length, thus eliminating any possible length bias issues. They find that likelihood generally correlates positively with human judgements, up until an inflection point, after which the correlation becomes negative. An observation also made in translation with BLEU rather than human judgements (Ott et al., 2018). We call this general failure of the mode to represent good translations in NMT the *inadequacy of the mode problem*.

3 NMT and its Many Biases

MT systems are trained on sentence pairs drawn from a parallel corpus. Each pair consists of a sequence x in the source language and a sequence y in the target language. Most NMT models are conditional

models (Cho et al., 2014b; Bahdanau et al., 2015; Vaswani et al., 2017),[1] that is, only the target sentence is given random treatment. Target words are drawn in sequence from a product of locally normalised Categorical distributions without Markov assumptions: $Y_j|\theta, x, y_{<j} \sim \mathrm{Cat}(f(x, y_{<j}; \theta))$. At each step, a neural network $f(\cdot; \theta)$ maps from the source sequence x and the prefix sequence $y_{<j}$ to the parameters of a Categorical distribution over the vocabulary of the target language. These models are typically trained via regularised maximum likelihood estimation, MLE for short, where we search for the parameter θ_{MLE} that assigns maximum (regularised) likelihood to a dataset of observations $\mathcal{D}$. A local optimum of the MLE objective can be found by stochastic gradient-based optimisation (Robbins and Monro, 1951; Bottou and Cun, 2004).

For a trained model with parameters θ_{MLE} and a given input x, a translation is predicted by searching for the mode of the distribution: the sequence $y^\star$ that maximises $\log p(y|x, \theta_{\mathrm{MLE}})$. This is a decision rule also known as maximum *a posteriori* (MAP) decoding (Smith, 2011).[2] Exact MAP decoding is intractable in NMT, and the beam search algorithm (Sutskever et al., 2014) is employed as a viable approximation.

It has been said that due to certain design decisions NMT suffers from a number of biases. We review those biases here and then discuss in Section 4 one bias that has received very little attention and which, we argue, underlies many biases in NMT and explains some of the pathologies discussed in Section 2.

Exposure bias. MLE parameters are estimated conditioned on observations sampled from the training data. Clearly, those are not available at test time, when we search through the learnt distribution. This mismatch between training and test, known as exposure bias (Ranzato et al., 2016), has been linked to many of the pathologies of NMT and motivated modifications or alternatives to MLE aimed at exposing the model to its own predictions during training (Bengio et al., 2015; Ranzato et al., 2016; Shen et al., 2016; Wiseman and Rush, 2016; Zhang et al., 2019). While exposure bias has been a point of critique mostly against MLE, it has only been studied in the context of approximate MAP decoding. The use of MAP decoding and its approximations shifts the distribution of the generated translations away from data statistics (something we provide evidence for in later sections), thereby exacerbating exposure bias.

Non-admissible heuristic search bias. In beam search, partial translations are ranked in terms of log-likelihood without regards to (or with crude approximations of) their future score, which may lead to good translations being pruned too early. This corresponds to searching with a non-admissible heuristic (Hart et al., 1968), that is, a heuristic that may underestimate the likelihood of completing a translation. This biased search affects statistics of beam search outputs in unknown ways and may well account for some of the pathologies of Section 2, and has motivated variants of the algorithm aimed at comparing partial translations more fairly (Huang et al., 2017; Shu and Nakayama, 2018). This problem has also been studied in parsing literature, where it's known as imbalanced probability search bias (Stanojević and Steedman, 2020).

Label bias. Where a conditional model makes independence assumptions about its inputs (*i.e.*, variables the model does not generate), local normalisation prevents the model from revising its decisions, a problem known as *label bias* (Bottou, 1991; Lafferty et al., 2001). This is a model specification problem which limits the distributions a model can represent (Andor et al., 2016). While this is the case in incremental parsing (Stern et al., 2017) and simultaneous translation (Gu et al., 2017), where inputs are incrementally available for conditioning, this is *not* the case in standard NMT (Sountsov and Sarawagi, 2016, Section 5), where inputs are available for conditioning in all generation steps. It is plausible that local normalisation might affect the kind of local optima we find in NMT, but that is orthogonal to label bias.

[1] Though fully generative accounts do exist (Shah and Barber, 2018; Eikema and Aziz, 2019).

[2] The term MAP decoding was coined in the context of generative classifiers and their structured counterparts, where the posterior probability $p(y|x, \theta) \propto p(y|\theta)p(x|y, \theta)$ updates our prior beliefs about y in light of x. This is not the case in NMT, where we do not express a prior over target sentences, and $p(y|x, \theta)$ is a direct parameterisation of the likelihood, rather than a posterior probability inferred via Bayes rule. Nonetheless, we stick to the conventions used in the MT literature.

4 Biased Statistics and the Inadequacy of the Mode

In most NMT research, criticisms of the model are based on observations about the mode, or an approximation to it obtained using beam search. The mode, however, is not an unbiased summary of the probability distribution that the model learnt. That is, properties of the mode say little about properties of the learnt distribution (*e.g.*, a short mode does not imply the model underestimates average sequence length). MAP decoding algorithms and their approximations bias the statistics by which we criticise NMT. They restrict our observations about the model to a single or a handful of outcomes which on their own can be rather rare. To gain insight about the model as a distribution, it seems more natural to use all of the information available to us, namely, all samples we can afford to collect, and search for frequent patterns in these samples. Evidence found that way better represents the model and its beliefs.

On top of that, the sample space of NMT is high-dimensional and highly structured. NMT models must distribute probability mass over a massive sample space (effectively unbounded). While most outcomes ought to be assigned negligible mass, for the total mass sums to 1, the outcomes with non-negligible mass might still be too many. The mode might only account for a tiny portion of the probability mass, and can actually be extremely unlikely under the learnt distribution. Using the mode for predictions makes intuitive sense in unstructured problems, where probability distributions are likely very peaked, and in models trained with large margin methods (Vapnik, 1998), since those optimise a decision boundary directly. With probability distributions that are very spread out, and where the mode represents only a tiny bit of probability mass, targeting at the mode for predictions is much less obvious, an argument that we shall reinforce with experimental results throughout this analysis.[3]

At the core of our analysis is the concept of an unbiased sample from the model, which we obtain by ancestral sampling: iteratively sampling from distributions of the form $\mathrm{Cat}(f(x, y_{<j}; \theta))$, each time extending the generated prefix $y_{<j}$ with an unbiased draw, until the end-of-sequence symbol is generated. By drawing from the model's probability distribution, unlike what happens in MAP decoding, we are imitating the model's training procedure. Only we replace samples from the data by samples from the model, thus shedding light onto the model's fit. That is, if these samples do not reproduce statistics of the data, we have an instance of poor fit.[4] Crucially, ancestral sampling is not a pathfinding algorithm, thus the non-admissible heuristic search bias it not a concern. Ancestral sampling is *not* a decision rule either, thus returning a single sample as a prediction is not expected to outperform MAP decoding (or any other rule). Samples can be used to diagnose model fit, as we do in Section 6, and to approximate decision rules, as we do in Section 7.4. In sum, we argue that MAP decoding is a source of various problems and that it biases conclusions about NMT. Next, we provide empirical evidence for these claims.

5 Data & System

We train our systems on German-English (de-en), Sinhala-English (si-en), and Nepali-English (ne-en), in both directions. For German-English we use all available WMT'18 (Bojar et al., 2018) parallel data, except for Paracrawl, amounting to about 5.9 million sentence pairs, and train a Transformer base model (Vaswani et al., 2017). For Sinhala and Nepali, for which very little parallel data are available, we mimic the data and system setup of Guzmán et al. (2019). As we found that the data contained many duplicate sentence pairs, we removed duplicates, but left in those where only one side (source or target) of the data is duplicate to allow for paraphrases. For all language pairs, we do keep a portion of the training data ($6,000$ sentence pairs) separate as held-out data for the analysis. In this process we also removed any sentence that corresponded exactly to either the source or target side of a held-out sentence from the training data. To analyse performance outside the training domain, we use WMT's *newstest2018* for German-English, and the FLORES datasets collected by Guzmán et al. (2019) for the low-resource pairs. Our analysis is focused on MLE-trained NMT systems. However, as Transformers are commonly trained with label smoothing (LS) (Szegedy et al., 2016), we do additionally report automatic quality assessments of beam search outputs on LS-trained systems.

[3]This perhaps non-intuitive notion that the most likely outcomes are rare and do not summarise a model's beliefs well enough is related to an information-theoretic concept, that of typicality (MacKay, 2003, Section 4.4).

[4]Where one uses (approximate) MAP decoding instead of ancestral sampling this is known as exposure bias.

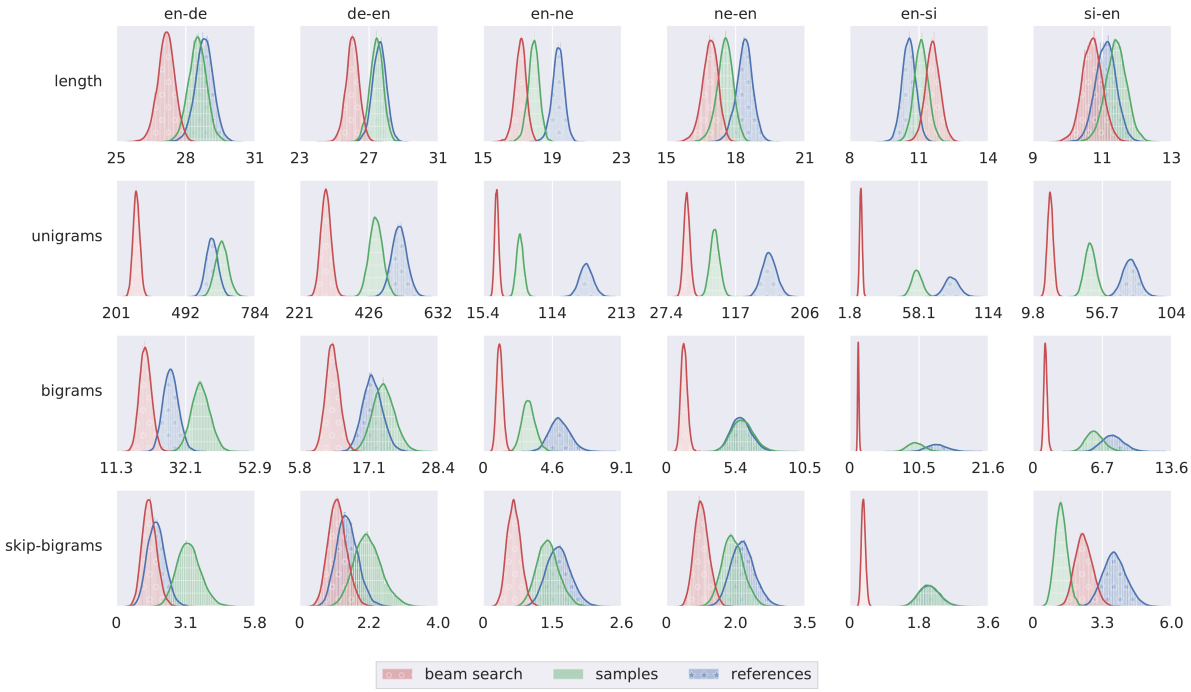

Figure 1: A comparison using hierarchical Bayesian models of statistics extracted from beam search outputs, samples from the model and gold-standard references. We show the posterior density on the y-axis, and the mean Poisson rate (length) and agreement with training data (unigrams, bigrams, skip-bigrams) on the x-axis for each group and language pair.

6 Assessing the Fit of MLE-Trained NMT

We investigate the fit of the NMT models of Section 5 on a held-out portion of the training data. This allows us to criticise MLE without confounders such as domain shift. We will turn to data in the test domain (*newstest2018*, FLORES) in Section 7. We compare unbiased samples from the model to gold-standard references and analyse statistics of several aspects of the data. If the MLE solution is good, we would expect statistics of sampled data to closely match statistics of observed data.

We obtain statistics from reference translations, ancestral samples, and beam search outputs and model them using hierarchical Bayesian models. For each type of statistic, we formulate a joint model over these three groups and inspect the posterior distribution over the parameters of the analysis model. We also include statistics extracted from the training data in our analysis, and model the three *test groups* as a function of posterior inferences based on training data statistics. Our methodology follows that advocated by Gelman et al. (2013) and Blei (2014). In particular, we formulate separate hierarchical models to inspect length, lexical, and word order statistics: sequence length, unigram and bigram counts, and skip-bigram counts, respectively.[5] In Appendix A, we describe in detail all analysis models, inference procedures, and predictive checks that confirm their fit.

For length statistics, we look at the expected posterior Poisson rate for each group, each rate can be interpreted as that group's average sequence length. Ideally, the expected Poisson rates of predicted translations are close to those of gold-standard references. Figure 1 (top row) shows the inferred posterior distributions for all language pairs. We observe that samples generated by NMT capture length statistics reasonably well, overlapping a fair amount with the reference group. In almost all cases we observe that beam search outputs stray away from data statistics, usually resulting in shorter translations.

For unigrams, bigrams, and skip-bigrams, we compare the posterior agreement with training data of

[5]Skip-bigrams are pairs of tokens drawn in the same order as they occur in a sentence, but without enforcing adjacency.

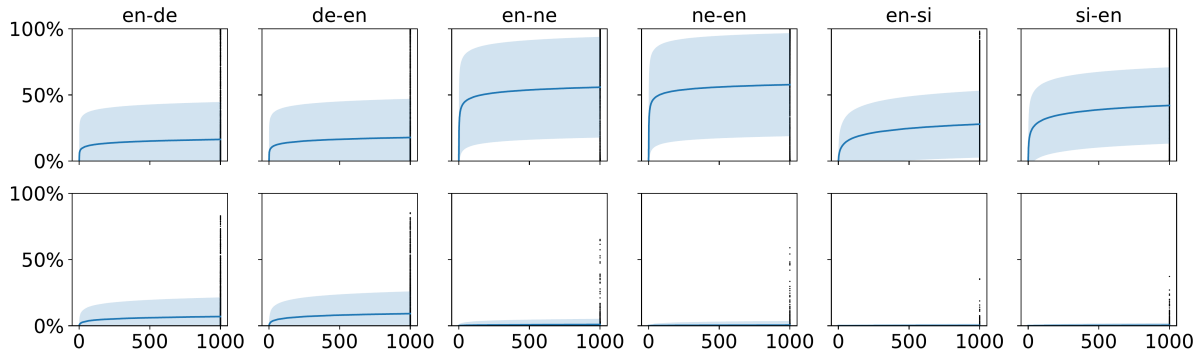

Figure 2: Cumulative probability of the unique translations in 1,000 ancestral samples on the held-out (top), and *newstest2018* / FLORES (bottom) test sets. The dark blue line shows the average cumulative probability over all test sentences, the shaded area represents 1 standard deviation away from the average. The black dots to the right show the final cumulative probability for each individual test sentence.

each group (this is formalised in terms of a scalar concentration parameter whose posterior we can plot). Higher values indicate a closer resemblance to training data statistics. For each statistic, the posterior distribution for gold-standard references gives an indication of ideal values of this agreement variable. Figure 1 (rows 2–4) show all posterior distributions. In most cases the gold-standard references agree most with the training data, as expected, followed by samples from the model, followed by beam search outputs. For nearly all statistics and language pairs beam search outputs show least agreement with the training data, even when samples from the model show similar agreement as references do. Whereas samples from the model do sometimes show less similarity than references, in most cases they are similar and thus lexical and word order statistics are captured reasonably well by the NMT model. Beam search on the other hand again strays from training data statistics, compared to samples from the model.

7 Examining the Translation Distribution

The NMT models of Section 5 specify complex distributions over an unbounded space of translations. Here we examine properties of these distributions by inspecting translations in a large set of unbiased samples. To gain further insight we also analyse our models in the test domain (*newstest2018*, FLORES).

7.1 Number of Likely Translations

NMT, by the nature of its model specification, assigns probability mass to each and every possible sequence consisting of tokens in its vocabulary. Ideally, however, a well-trained NMT model assigns the bulk of its probability mass to good translations of the input sequence. We take $1,000$ unbiased samples from the model for each input sequence and count the cumulative probability mass of the unique translations sampled. Figure 2 shows the average cumulative probability mass for all test sentences with 1 standard deviation around it, as well as the final cumulative probability values for each input sequence. For the held-out data we observe that, on average, between 16.4% and 57.8% of the probability mass is covered. The large variance around the mean shows that in all language pairs we can find test sentences for which nearly all or barely any probability mass has been covered after $1,000$ samples. That is, even after taking $1,000$ samples, only about half of the probability space has been explored. The situation is much more extreme when translating data from the test domain (see bottom half of Figure 2).[6] Naturally, the NMT model is much more uncertain in this scenario, and this is very clear from the amount of probability mass that has been covered by $1,000$ samples: on average, only between 0.2% and 0.9% for the low-resource pairs and between 6.9% and 9.1% for English-German of the probability space has been explored. This shows that the set of likely translations under the model is very large and the probability

[6]For English-German and German-English the test domain would not be considered out-of-domain here, as both training and test data concern newswire.

	Training Domain					Test Domain				
Task	LS	beam	sample	MBR	Oracle	LS	beam	sample	MBR	Oracle
en-de	19.5	19.5	15.3	19.2	22.7	35.2	34.9	20.5	31.5	35.7
de-en	26.6	26.8	21.9	26.2	29.4	39.6	39.4	26.6	37.3	41.0
en-ne	36.7	37.3	36.1	40.5	43.4	32.5	31.3	30.6	34.9	37.0
ne-en	30.6	29.8	26.7	30.2	35.4	19.2	17.2	12.8	16.6	20.1
en-si	34.2	34.3	31.0	36.3	41.5	31.8	30.3	30.3	34.8	36.8
si-en	29.1	28.9	24.3	29.1	36.2	20.0	18.4	13.7	17.7	21.6
High-resource	23.1	23.1	18.6	22.7	26.0	37.4	37.1	23.6	34.4	38.3
Low-resource	32.7	32.6	29.5	34.0	39.1	25.9	24.3	21.8	26.0	28.9
All	29.5	29.4	25.9	30.2	34.8	29.7	28.6	22.4	28.8	32.0

Table 1: METEOR scores under different strategies for prediction: beam search, single sample, MBR, and an oracle rule. MBR and the oracle both use 30 ancestral samples and sentence-level METEOR as utility, but the oracle has access to the reference. To show that our MLE-trained systems are competitive with LS-trained systems, we list the LS column (using beam search). The sample columns show average scores of 30 independent samples from the model. All standard deviations were below 0.2.

distribution over those sentences mostly quite flat, especially so in the test domain. In fact, if we look at each input sequence individually, we see that for 37.0% (en-de), 35.5% (de-en), 18.5% (en-ne), 15.7% (ne-en), 9.2% (en-si), and 3.3% (si-en) of them all $1,000$ samples are unique. On the test domain data these numbers increase to 46.7% (en-de), 41.5% (de-en), 52.1% (en-ne), 86.8% (ne-en), 84.6% (en-si), and 87.3% (si-en). For these input sequences, the translation distributions learnt are so flat that in these $1,000$ samples no single translation stands out over the others.

7.2 Sampling the Mode

As the predominant decision rule in NMT is MAP decoding, which we approximate via beam search, it is natural to ask how frequently it is that the beam search output is observed amongst unbiased samples. We find that the beam search output is contained within $1,000$ unbiased samples for between 54.3% and 92.2% of input sequences on the held-out data. In the test domain, we find that on English-German for between 44.3% and 49.3%, and in the low-resource setting for between 4.8% and 8.4% of the input sequences the beam search output is contained in the set. This shows that for a large portion of the input sequences, the beam search solution is thus quite a rare outcome under the model.

Recently, Stahlberg and Byrne (2019) showed that oftentimes the true mode of a trained NMT system is the empty sequence. This is worrying since NMT decoding is based on mode-seeking search. We find that for between 7.2% and 29.1% of input sequences for held-out data and between 2.8% and 33.3% of input sequences in the test domain an empty sequence is sampled at least once in $1,000$ samples. When an empty sequence is sampled it only occurs on average 1.2 ± 0.5 times. Even though it could well be, as the evidence that Stahlberg and Byrne (2019) provide is strong, that often the true mode under our models is the empty sequence, the empty string remains a rather unlikely outcome under the models.

7.3 Sample Quality

The number of translations that an NMT model assigns non-negligible mass to can be very large as we have seen in Section 7.1. We now investigate what the average quality of these samples is. For quality assessments, we compute METEOR (Denkowski and Lavie, 2011) using the `mteval-v13a` tokeniser.[7] We translate the test sets using a single ancestral sample per input sentence and repeat the experiment 30 times to report the average in Table 1 (sample). We also report beam search scores (beam). We see that, on average, samples of the model always perform worse than beam search translations.

[7]For our analysis, it is convenient to use a metric defined both at the corpus and at the segment level. We use METEOR, rather than BLEU (Papineni et al., 2002), for it outperforms (smoothed) BLEU at the segment-level (Ma et al., 2018).

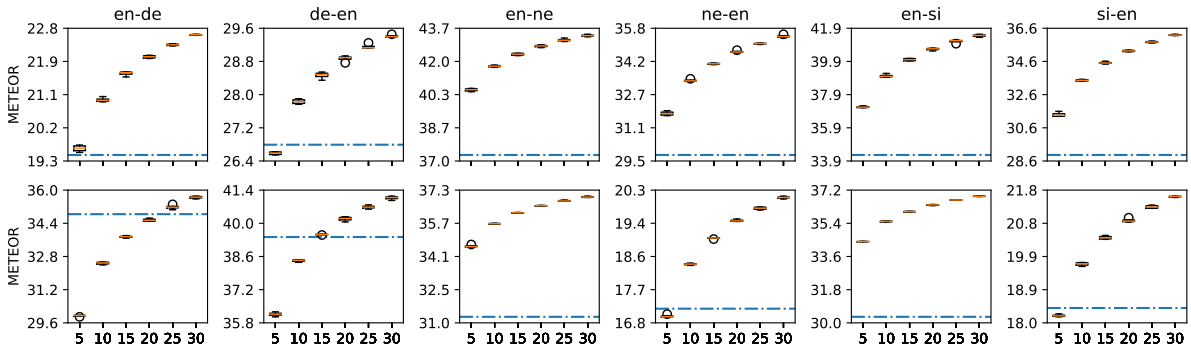

Figure 3: METEOR scores for oracle-selected samples as a function of sample size on the held-out data (top) and *newstest2018* / FLORES (bottom) test sets. For each sample size we repeat the experiment 4 times and show a box plot per sample size. Dashed blue lines show beam search scores.

This is no surprise, of course, as ancestral sampling is not a fully fledged decision rule, but simply a technique to unbiasedly explore the learnt distribution. Moreover, beam search itself does come with some adjustments to perform well (such as a specific beam size and length penalty). The gap between sampling and beam search is between 0 and 14.4 METEOR. The gap can thus be quite large, but overall the quality of an average sample is reasonable compared to beam search. We also observe that the variance of the sample scores is small with standard deviations below 0.2.

Next, we investigate the performance we would achieve if we could select the best sample from a set. For that, we employ an oracle selection procedure using sentence-level METEOR with the reference translation to select the best sample from a set of samples. We vary sample size from 5 to 30 samples and repeat each experiment four times. Figure 3 plots the results in terms of corpus-level METEOR. Average METEOR scores for oracle selection out of 30 samples are shown in Table 1. METEOR scores steadily increase with sample size. For a given sample size we observe that variance is generally very small. Only between 5 and 10 samples are required to outperform beam search in low-resource language pairs and English-German in the training domain, but surprisingly 15 to 25 samples are necessary for English-German in the test domain. Still, this experiment shows that samples are of reasonable and consistent quality with respect to METEOR. For fewer than 30 random samples the model could meet or outperform beam search performance in most cases, if we knew how to choose the best sample from the set. This is a motivating result for looking into sampling-based decision rules.

7.4 Minimum Bayes Risk Decoding

We have seen that translation distributions spread mass over a large set of likely candidates, oftentimes without any clear preference for particular translations within the set (Section 7.1). Yet, this set is not arbitrary, it captures various statistics of the data well (Section 6) and holds potentially good translations (Section 7.3). Even though the model does not single out one clear winner, the translations it does assign non-negligible mass to share statistics that correlate with the reference translation. This motivates a decision rule that exploits all information we have available about the distribution. In this section we explore one such decision rule: minimum Bayes risk (MBR) decoding.

For a given *utility function* $u(y, h)$, which assesses a hypothesis h against a reference y, statistical decision theory (Bickel and Doksum, 1977) prescribes that the optimum decision $y^\star$ is the one that maximises expected utility (or minimises expected loss) under the model: $y^\star = \text{argmax}_{h \in \mathcal{H}(x)} \; \mathbb{E}_{p(y|x,\theta)}[u(y, h)]$, where the maximisation is over the entire set of possible translations $\mathcal{H}(x)$. Note that there is no need for a human-annotated reference, expected utility is computed by having the model *fill in* reference translations. This decision rule, known as MBR decoding in the NLP literature (Goel and Byrne, 2000), is especially suited where we trust a model in expectation but not its mode in particular (Smith, 2011,

Section 5.3).[8] MBR decoding, much like MAP decoding, is intractable. We can at best obtain unbiased estimates of expected utility via Monte Carlo (MC) sampling, and we certainly cannot search over the entirety of $\mathcal{H}(x)$. Still, a tractable approximation can be designed, albeit without any optimality guarantees. We use MC both to approximate the support $\mathcal{H}(x)$ of the distribution and to estimate the expected utility of a given hypothesis. In particular, we maximise over the support $\bar{\mathcal{H}}(x)$ of the empirical distribution obtained by ancestral sampling:

$$y^\star = \underset{h \in \bar{\mathcal{H}}(x)}{\operatorname{argmax}} \frac{1}{S} \sum_{s=1}^{S} u(y^{(s)}, h) \quad \text{for } y^{(s)} \sim p(y|x, \theta) \,, \tag{1}$$

which runs in time $\mathcal{O}(S^2)$. Though approximate, this rule has interesting properties: MC improves with sample size, occasional pathologies in the set pose no threat, and there is no need for incremental search.

Note that whereas our translation distribution might be very flat over a vast number of translations, not showing a clear ordering in terms of relative frequency within a large set of samples, this need not be the case under expected utility. For example, in Section 7.2 we found that for some input sequences the empty sequence is contained within the 1,000 samples in our set and appears in there roughly once on average. If all the 1,000 samples are unique (as we found to often be the case in Section 7.1), we cannot distinguish the empty sequence from the other 999 samples in terms of relative frequency. However, under most utilities the empty sequence is so unlike the other sampled translations that it would score very low in terms of expected utility.

We chose METEOR as utility function for it, unlike BLEU, is well-defined at the sentence level.[9] We estimate expected utility using $S = 30$ ancestral samples, and use the translations we sample to make up an approximation to $\mathcal{H}(x)$. Results are shown in Table 1. As expected, MBR considerably outperforms the average single sample performance by a large margin and in many cases is on par with beam search, consistently outperforming it in low-resource pairs. For English-German in the test domain, we may need more samples to close the gap with beam search. Whereas an out-of-the-box solution based on MBR requires further investigation, this experiment shows promising results. Crucially, it shows that exploring the model as a probability distribution holds great potential.

8 Related Work

Some of our observations have been made in previous work. Ott et al. (2018) observe that unigram statistics of beam search stray from those of the data, while random samples do a better job at reproducing them. Holtzman et al. (2020) find that beam search outputs have disproportionately high token probabilities compared to natural language under a sequence to sequence model. Our analysis is more extensive, we include richer statistics about the data, more language pairs, and vary the amount of training resources, leading to new insights about MLE-trained NMT and the merits of mode-seeking predictions.

Ott et al. (2018) also observe that NMT learns flat distributions, they analyse a high-resource English-French system trained on 35.5 million sentence pairs from WMT'14 and find that after drawing $10,000$ samples from the WMT'14 validation set less than 25% of the probability space has been explored. Our analysis shows that even though NMT distributions do not reveal clear winners, they do emphasise translations that share statistics with the reference, which motivates us to look into MBR.

MBR decoding is old news in machine translation (Kumar and Byrne, 2004; Tromble et al., 2008), but it has received little attention in NMT. Previous approximations to MBR in NMT employ beam search to define the support and to evaluate expected utility (with probabilities renormalised to sum to 1 in the beam), these studies report the need for very large beams (Stahlberg et al., 2017; Blain et al., 2017; Shu and Nakayama, 2017). They claim the inability to directly score better translations higher is a

[8] MAP decoding is in fact MBR with a very strict utility function which evaluates to 1 if a translation exactly matches the reference, and 0 otherwise. As a community, we acknowledge by means of our evaluation strategies (manual or automatic) that exact matching is inadequate for translation, unlike many unstructured classification problems, admits multiple solutions.

[9] Even though one can alter BLEU such that it is defined at the sentence level (for example, by adding a small positive constant to n-gram counts), this "smoothing" in effect biases BLEU's sufficient statistics. Unbiased statistics are the key to MBR, thus we opt for a metric that is already defined at the sentence level.

deficiency of the model scoring function. We argue this is another piece of evidence for the inadequacy of the mode: by using beam search, they emphasise statistics of high-scoring translations, potentially rare and inadequate ones. Very recently, Borgeaud and Emerson (2020) present a voting-theory perspective on decoding for image captioning and machine translation. Their proposal is closely-related to MBR, but motivated differently. Their decision rule too is guided by beam search, which may emphasise pathologies of highest-likelihood paths, but they also propose and investigate stronger utility functions which lead to improvements w.r.t. length, diversity, and human judgements.

The only instance that we are aware of where unbiased samples from an NMT model support a decision rule is the concurrent work by Naskar et al. (2020). The authors make the same observation that we make in Section 7.3, namely that an oracle selection from a small set of samples of an NMT model shows great potential, greatly outperforming beam search. Motivated by this observation, the authors propose a re-ranking model that learns to rank sampled translations according to their oracle BLEU. Using the trained model to re-rank a set of 100 samples from the NMT model they find strong improvements over beam search of up to 3 BLEU points, again showing the potential of sampling-based decision rules.

9 Conclusion

In this work, we discuss the inadequacy of the mode in NMT and question the appropriateness of MAP decoding. We show that for such a high dimensional problem as NMT, the probability distributions obtained with MLE are spread out over many translations, and that the mode often does not represent any significant amount of probability mass under the learnt distribution. We therefore argue that MAP decoding is not suitable as a decision rule for NMT systems. Whereas beam search performs well in practice, it suffers from biases of its own (*i.e.*, non-admissible heuristic search bias), it shifts statistics away from those of the data (*i.e.*, exposure bias and other lexical and length biases), and in the limit of perfect search it falls victim to the inadequacy of the mode. Instead, we advocate for research into decision rules that take into account the probability distribution more holistically. Using ancestral sampling we can explore the learnt distribution in an unbiased way and devise sampling-based decision rules. Ancestral sampling does not suffer from non-admissibility, and, if the model fit is good, there is no distribution shift either.

We further argue that criticisms about properties of the mode of an NMT system are not representative of the probability distributions obtained from MLE training. While this form of criticism is perfectly reasonable where approximations to MAP decoding are the only viable alternative, there are scenarios where we ought to criticise models as probability distributions. For example, where we seek to correlate an observed pathology with a design decision, such as factorisation, or training algorithm. In fact, we argue that many of the observed pathologies and biases of NMT are at least partially due to the use of (approximate) MAP decoding, rather than inherent to the model or its training objective.

Even though NMT models spread mass over many translations, we find samples to be of decent quality and contain translations that outperform beam search outputs even for small sample sizes, further motivating the use of sampling-based decision rules. We show that an approximation to one such decision rule, MBR decoding, shows competitive results. This confirms that while the set of likely translations under the model is large, the translations in it share many statistics that correlate well with the reference.

MLE-trained NMT models admit probabilistic interpretation and an advantage of the probabilistic framework is that a lot of methodology is already in place when it comes to model criticism as well as making predictions. We therefore advocate for criticising NMT models as probability distributions and making predictions using decision rules that take into account the distributions holistically. We hope that our work paves the way for research into scalable sampling-based decision rules and motivates researchers to assess model improvements to MLE-trained NMT systems from a probabilistic perspective.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825299 (GoURMET). We also thank Khalil Sima'an, Lina Murady, Miloš Stanojević, and Lena Voita for comments and helpful discussions. A Titan Xp card used for this research was donated by the NVIDIA Corporation.

References

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 *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2442–2452, Berlin, Germany, August. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR, 2015*, San Diego, USA.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1171–1179. Curran Associates, Inc.

Peter J Bickel and Kjell A Doksum. 1977. *Mathematical statistics: basic ideas and selected topics*. Holden-Day Inc., Oakland, CA, USA.

Frédéric Blain, Lucia Specia, and Pranava Madhyastha. 2017. Exploring hypotheses spaces in neural machine translation. *Asia-Pacific Association for Machine Translation (AAMT), editor, Machine Translation Summit XVI. Nagoya, Japan*.

David M Blei. 2014. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October. Association for Computational Linguistics.

Sebastian Borgeaud and Guy Emerson. 2020. Leveraging sentence similarity in natural language generation: Improving beam search using range voting. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 97–109, Online, July. Association for Computational Linguistics.

Léon Bottou and Yann L. Cun. 2004. Large scale online learning. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 217–224. MIT Press.

Bottou. 1991. Une approche théorique de lapprentissage connexioniste; applications à la reconnaissance de la parole.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar, October. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Eldan Cohen and J. Christopher Beck. 2019. Empirical analysis of beam search performance degradation in neural sequence models. In *ICML*.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of WMT, 2011*, pages 85–91, Edinburgh, Scotland, July.

Bryan Eikema and Wilker Aziz. 2019. Auto-encoding variational neural machine translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 124–141, Florence, Italy, August. Association for Computational Linguistics.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2013. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3rd ed. edition.

Vaibhava Goel and William J. Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Comput. Speech Lang.*, 14(2):115–135.

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O.K. Li. 2017. Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1053–1062, Valencia, Spain, April. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6100–6113, Hong Kong, China, November. Association for Computational Linguistics.

P. E. Hart, N. J. Nilsson, and B. Raphael. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Liang Huang, Kai Zhao, and Mingbo Ma. 2017. When to finish? optimal beam search for neural text generation (modulo beam size). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2134–2139, Copenhagen, Denmark, September. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML 01, page 282289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels, October. Association for Computational Linguistics.

David J. C. MacKay. 2003. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press.

Kenton Murray and David Chiang. 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium, October. Association for Computational Linguistics.

Subhajit Naskar, Amirmohammad Rooshenas, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2020. Energy-based reranking: Improving neural machine translation using energy-based models. *CoRR*, abs/2009.13267.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965, Stockholmsmssan, Stockholm Sweden, 10–15 Jul. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL, 2002*, pages 311–318, Philadelphia, USA, July.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Herbert Robbins and Sutton Monro. 1951. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407.

Harshil Shah and David Barber. 2018. Generative neural machine translation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1346–1355. Curran Associates, Inc.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany, August. Association for Computational Linguistics.

Raphael Shu and Hideki Nakayama. 2017. Later-stage minimum bayes-risk decoding for neural machine translation. *CoRR*, abs/1704.03169.

Raphael Shu and Hideki Nakayama. 2018. Improving beam search by removing monotonic constraint for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 339–344, Melbourne, Australia, July. Association for Computational Linguistics.

Noah A. Smith. 2011. *Linguistic Structure Prediction*. Morgan and Claypool.

Pavel Sountsov and Sunita Sarawagi. 2016. Length bias in encoder decoder models and a case for global conditioning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1525, Austin, Texas, November. Association for Computational Linguistics.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3347–3353, Hong Kong, China, November. Association for Computational Linguistics.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. 2017. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 362–368, Valencia, Spain, April. Association for Computational Linguistics.

Miloš Stanojević and Mark Steedman. 2020. Max-Margin Incremental CCG Parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Mitchell Stern, Daniel Fried, and Dan Klein. 2017. Effective inference for generative neural parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1695–1700, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. 2014. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *NIPS, 2014*, pages 3104–3112. Montreal, Canada.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice Minimum Bayes-Risk decoding for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 620–629, Honolulu, Hawaii, October. Association for Computational Linguistics.

Vladimir Vapnik. 1998. Statistical learning theory Wiley. *New York*, 1:624.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 6000–6010.

Sam Wiseman and Alexander M. Rush. 2016. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, Austin, Texas, November. Association for Computational Linguistics.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059, Brussels, Belgium, October-November. Association for Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy, July. Association for Computational Linguistics.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2020. Trading off diversity and quality in natural language generation. *CoRR*, abs/2004.10450.

A Analysis Models

A.1 Length Analysis

We model length data from the training group using a hierarchical Gamma-Poisson model. Each target sequence length is modelled as being a draw from a Poisson distribution with a Poisson rate parameter specific to that sequence. All Poisson rates share a common population-level Gamma prior with population-level parameters α and β. The population-level parameters are given fixed Exponential priors set to allow for a wide but reasonable range of Poisson rates *a priori*.

$$\alpha \sim \text{Exp}(1) \qquad \beta \sim \text{Exp}(10)$$
$$\lambda_i \sim \text{Gamma}(\alpha, \beta) \qquad y_i \sim \text{Poisson}(\lambda_i)$$

Here, i indexes one particular data point. This model is very flexible, because we allow the model to assign each datapoint its own Poisson rate. We model test groups as an extension of the training group. Test group data points are also modelled as draws from a Gamma-Poisson model, but parameterised slightly differently.

$$\mu = \mathbb{E}\left[\text{Gamma}(\alpha, \beta | \mathcal{D}_T)\right] \qquad \eta \sim \text{Exp}(1.)$$
$$s_g \sim \text{Exp}(\eta) \qquad t_g = 1/\mu$$
$$\lambda_{gi} \sim \text{Gamma}(s_g, t_g) \qquad y_{gi} \sim \text{Poisson}(\lambda_{gi})$$

Here, i again indexes a particular data point, g a group in $\{\text{reference}, \text{sampling}, \text{beam}\}$, and $\mathcal{D}_T$ denotes the data of the training group. All Poisson rates are individual to each datapoint in each group. The Poisson rates do share a group-level Gamma prior, whose parameters are s_g and t_g. s_g shares a prior among all test groups and therefore ties all test groups together. t_g is derived from posterior inferences on the training data by taking the expected posterior Poisson rate in the training data and inverting it. This is done such that the mean Poisson rate for each test group is $s_g \cdot \mu$, where s_g can be seen as a parameter that scales the expected posterior training rate for each test group individually. We infer Gamma posterior approximations for all unknowns using stochastic variational inference (SVI). After inferring posteriors, we compare predictive samples to the observed data in terms of first to fourth order moments to verify that the model fits the observations well.

A.2 Lexical & Word Order Analyses

We model unigram and (skip-)bigram data from the training group using a hierarchical Dirichlet-Multinomial model as shown below:

$$\alpha \sim \text{Gamma}(1, 1) \qquad \beta \sim \text{Gamma}(1, 1)$$
$$\theta \sim \text{Dir}(\alpha) \qquad \psi_u \sim \text{Dir}(\beta)$$
$$u \sim \text{Multinomial}(\theta) \qquad b|u \sim \text{Multinomial}(\psi_u)$$

Here, we have one Gamma-Dirichlet-Multinomial model to model unigram counts u, and a separate Dirichlet-Multinomial model for each u (the first word of a bigram) that b (the second word of a bigram) conditions on, sharing a common Gamma prior that ties all bigram models. This means that we effectively have $V + 1$ Dirichlet-Multinomial models (where V is BPE vocabulary size) in total to model the training group, where the V bigram models share a common prior.

We model the three test groups using the inferred posterior distributions on the data of the training group $\mathcal{D}_T$. We compute the expected posterior concentration of the Dirichlets in the training group models and normalise it such that it sums to 1 over the entire vocabulary. The normalisation has the effect of

spreading the unigram and bigram distributions. The test groups are modelled by scaling this normalised concentration parameter using a scalar. In order for test-groups to recover the training distribution the scaling variable needs to be large to undo the normalisation. This scalar, s_g for unigrams or m_g for bigrams, can be interpreted as the amount of agreement of each test group with the training group. The higher this scalar is, the more peaked the test group Multinomials will be about the training group lexical distribution.

$$
\begin{aligned}
\mu(\alpha) &= \mathbb{E}\left[\alpha|\mathcal{D}_T\right]. & \mu(\beta) &= \mathbb{E}\left[\beta|\mathcal{D}_T\right] \\
\eta_s &\sim \mathrm{Gamma}(1, 0.2) & \eta_m &\sim \mathrm{Gamma}(1, 0.2) \\
s_g &\sim \mathrm{Gamma}(1, \eta_s) & m_g &\sim \mathrm{Gamma}(1, \eta_m) \\
\theta_g &\sim \mathrm{Dir}(s_g \cdot \mu(\alpha)) & \psi_g &\sim \mathrm{Dir}(m_g \cdot \mu(\beta)) \\
u_g &\sim \mathrm{Multinomial}(\theta_g) & b_g|u_g &\sim \mathrm{Multinomial}(\psi_g) \\
g &\in \{\text{reference}, \text{sampling}, \text{beam}\}
\end{aligned}
$$

We do collapsed inference for each Dirichlet-Multinomial (as we are not interested in assessing θ_g or ϕ_g), and infer posteriors approximately using SVI with Gamma approximate posterior distributions. To confirm the fit of the analysis model, we compare posterior predictive samples to the observed data in terms of absolute frequency errors of unigrams and bigrams as well as ranking correlation.