Analyzing Uncertainty in Neural Machine Translation

Myle Ott[1] **Michael Auli**[1] **David Grangier**[1] **Marc'Aurelio Ranzato**[1]

Abstract

Machine translation is a popular test bed for research in neural sequence-to-sequence models but despite much recent research, there is still a lack of understanding of these models. Practitioners report performance degradation with large beams, the under-estimation of rare words and a lack of diversity in the final translations. Our study relates some of these issues to the inherent uncertainty of the task, due to the existence of multiple valid translations for a single source sentence, and to the extrinsic uncertainty caused by noisy training data. We propose tools and metrics to assess how uncertainty in the data is captured by the model distribution and how it affects search strategies that generate translations. Our results show that search works remarkably well but that models tend to spread too much probability mass over the hypothesis space. Next, we propose tools to assess model calibration and show how to easily fix some shortcomings of current models. As part of this study, we release multiple human reference translations for two popular benchmarks.

1. Introduction

Machine translation (MT) is an interesting task not only for its practical applications but also for the formidable learning challenges it poses, from how to transduce variable length sequences, to searching for likely sequences in an intractably large hypothesis space, to dealing with the multimodal nature of the prediction task, since typically there are several correct ways to translate a given sentence.

The research community has made great advances on this task, recently focusing the effort on the exploration of several variants of neural models (Bahdanau et al., 2014; Luong et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) that have greatly improved the state of the art performance on

---

[1]Facebook AI Research, USA. Correspondence to: Myle Ott <myleott@fb.com>.

public benchmarks. However, several open questions remain (Koehn & Knowles, 2017). In this work, we analyze top-performing trained models in order to answer some of these open questions. We target better understanding to help prioritize future exploration towards important aspects of the problem and therefore speed up progress.

For instance, according to conventional wisdom neural machine translation (NMT) systems under-estimate rare words (Koehn & Knowles, 2017), why is that? Is the model poorly calibrated? Is this due to exposure bias (Ranzato et al., 2016), i.e., the mismatch between the distribution of words observed at training and test time? Or is this due to the combination of uncertainty in the prediction of the next word and inference being an $\arg\max$ selection process, which always picks the most likely/frequent word? Similarly, it has been observed (Koehn & Knowles, 2017) that performance degrades with large beams. Is this due to poor fitting of the model which assigns large probability mass to bad sequences? Or is this due to the heuristic nature of this search procedure which fails to work for large beam values? In this paper we will provide answers and solutions to these and other related questions.

The underlying theme of all these questions is *uncertainty*, i.e. the one-to-many nature of the learning task. In other words, for a given source sentence there are several target sequences that have non negligible probability mass. Since the model only observes one or very few realizations from the data distribution, it is natural to ask the extent to which an NMT model trained with token-level cross-entropy is able to capture such a rich distribution, and whether the model is calibrated. Also, it is equally important to understand the effect that uncertainty has on search and whether there are better and more efficient search strategies.

Unfortunately, NMT models have hundreds of millions of parameters, the search space is exponentially large and we typically observe only one reference for a given source sentence. Therefore, measuring fitness of a NMT model to the data distribution is a challenging scientific endeavor, which we tackle by borrowing and combining tools from the machine learning and statistics literature (Kuleshov & Liang, 2015; Guo et al., 2017). With these tools, we show that search works surprisingly well, yielding highly likely sequences even with relatively narrow beams. Even if we

consider samples from the model that have similar likelihood, beam hypotheses yield higher BLEU on average. Our analysis also demonstrates that although NMT is well calibrated at the token and set level, it generally spreads too much probability mass over the space of sequences. This often results in individual hypotheses being under-estimated, and overall, poor quality of samples drawn from the model. Interestingly, systematic mistakes in the data collection process also contribute to uncertainty, and a particular such kind of noise, the target sentence being replaced by a copy of the corresponding source sentence, is responsible for much of the degradation observed when using wide beams.

This analysis – the first one of its kind – introduces tools and metrics to assess fitting of the model to the data distribution, and shows areas of improvement for NMT. It also suggests easy fixes for some of the issues reported by practitioners. We also release the data we collected for our evaluation, which consists of ten human translations for 500 sentences taken from the WMT'14 En-Fr and En-De test sets.[1]

2. Related Work

In their seminal work, Zoph et al. (2015) frame translation as a compression game and measure the amount of information added by translators. While this work precisely quantifies the amount of uncertainty, it does not investigate its effect on modeling and search. In another context, uncertainty has been considered for the design of better evaluation metrics (Dreyer & Marcu, 2012; Galley et al., 2015), in order not to penalize a model for producing a valid translation which is different from the provided reference.

Most work in NMT has focused on improving accuracy without much consideration for the intrinsic uncertainty of the translation task itself (Bahdanau et al., 2014; Luong et al., 2015; Gehring et al., 2017; Vaswani et al., 2017) (§3). Notable exceptions are latent variable models (Blunsom et al., 2008; Zhang et al., 2016) which explicitly attempt to model multiple modes in the data distribution, or decoding strategies which attempt to predict diverse outputs while leaving the model unchanged (Gimpel et al., 2013; Vijayakumar et al., 2016; Li & Jurafsky, 2016; Cho, 2016). However, none of these works check for improvements in the match between the model and the data distribution.

Recent work on analyzing machine translation has focused on topics such as comparing neural translation to phrase-based models (Bentivogli et al., 2016; Toral & Sanchez-Cartagena, 2017). Koehn & Knowles (2017) presented several challenges for NMT, including the deterioration of accuracy for large beam widths and the under-estimation of

---

[1]Additional reference translations are available from: https://github.com/facebookresearch/analyzing-uncertainty-nmt.

rare words, which we address in this paper. Isabelle et al. (2017) propose a new evaluation benchmark to test whether models can capture important linguistic properties. Finally, Niehues et al. (2017) focus on search and argue in favor of better translation modeling instead of improving search.

3. Data Uncertainty

Uncertainty is a core challenge in translation, as there are several ways to correctly translate a sentence; but what are typical sources of uncertainty found in modern benchmark datasets? Are they all due to different ways to paraphrase a sentence? In the following sections, we answer these questions, distinguishing uncertainty inherent to the task itself (§3.1), and uncertainty due to spurious artifacts caused by the data collection process (§3.2).

3.1. Intrinsic Uncertainty

One source of uncertainty is the existence of several *semantically equivalent* translations of the same source sentence. This has been extensively studied in the literature (Dreyer & Marcu, 2012; Padó et al., 2009). Translations can be more or less literal, and even if literal there are many ways to express the same meaning. Sentences can be in the active or passive form and for some languages determiners and prepositions such as 'the', 'of', or 'their' can be optional.

Besides uncertainty due to the existence of distinct, yet semantically equivalent translations, there are also sources of uncertainty due to *under-specification* when translating into a target language more inflected than the source language. Without additional context, it is often impossible to predict the missing gender, tense, or number, and therefore, there are multiple plausible translations of the same source sentence. Simplification or addition of cultural context are also common sources of uncertainty (Venuti, 2008).

3.2. Extrinsic Uncertainty

Statistical machine translation systems, and in particular NMT models, require lots of training data to perform well. To save time and effort, it is common to augment high quality human translated corpora with lower quality web crawled data (Smith et al., 2013). This process is error prone and responsible for introducing additional uncertainty in the data distribution. Target sentences may only be partial translations of the source, or the target may contain information not present in the source. A lesser-known example are target sentences which are entirely in the source language, or which are primarily *copies of the corresponding source*. For instance, we found that between 1.1% to 2.0% of *training* examples in the WMT'14 En-De and WMT'14 En-Fr datasets (§4.2) are "copies" of the source sentences, where a target sentence is labeled as "copy" if the intersection over

the union of unigrams (excluding punctuation and numbers) is at least 50%. Source copying is particularly interesting since we show that, even in small quantities, it can significantly affect the model output (§5.3). Note that *test* sets are manually curated and never contain copies.

4. Experimental Setup

This section describes the models and the datasets used in our analysis.

4.1. Sequence to Sequence Model

Our experiments rely on the publicly available pre-trained models of the *fairseq-py* toolkit (Gehring et al., 2017), although we obtained similar findings with other architectures as well. These models achieve competitive performance on the datasets we consider. Formally, let $\mathbf{x}$ be an input sentence with m words $\{x_1, \ldots, x_m\}$, and $\mathbf{t}$ be the ground truth target sentence with n words $\{t_1, \ldots, t_n\}$. The model is composed of an *encoder* and a *decoder*. The encoder takes $\mathbf{x}$ through several convolutional layers to produce a sequence of hidden states, $\mathbf{z} = \{z_1, \ldots, z_m\}$, one per input word. At time step k, the decoder takes a window of words produced so far (or the ground truth words at training time), $\{t_{k-1}, \ldots, t_{k-i}\}$, the set of encoder hidden states $\mathbf{z}$ and produces a distribution over the current word: $p(t_k|t_{k-1}, \ldots, t_{k-i}, \mathbf{z})$. More precisely, at each time step, an *attention* module (Bahdanau et al., 2014) summarizes the sequence $\mathbf{z}$ with a single vector through a weighted sum of $\{z_1, \ldots, z_m\}$. The weights depend on the source sequence $\mathbf{x}$ and the decoder hidden state, h_k, which is the output of several convolutional layers taking as input $\{t_{k-1}, \ldots, t_{k-i}\}$. From the source attention vector, the hidden state of the decoder is computed and the model emits a distribution over the current word as in: $p(t_k|h_k) = \text{softmax}(Wh_k + b)$. Gehring et al. (2017) provides further details. To train the translation model, we minimize the cross-entropy loss: $\mathcal{L} = -\sum_{i=1}^{n} \log p(t_i|t_{i-1}, \ldots, t_1, \mathbf{x})$, uaing Nesterov's momentum (Sutskever et al., 2013).

At test time, we aim to output the most likely translation given the source sentence, according to the model estimate. We approximate such an output via *beam search*. Unless otherwise stated, we use beam width $k = 5$, where hypotheses are compared to one another in terms of length-normalized log-likelihood. Some experiments consider *sampling* from the model conditional distribution $p(t_i|t_{i-1}, \mathbf{h}_{i-1}, \mathbf{x})$, one token at a time, until the special end of sentence symbol is sampled.

4.2. Datasets and Evaluation

We consider the following datasets:
WMT'14 English-German (En-De): We use the same

	En-Fr	En-De
Automatic evaluation		
train PPL	2.54	5.14
valid PPL	2.56	6.36
test BLEU	41.0	24.8
Human evaluation (pairwise)		
Ref > Sys	42.0%	80.0%
Ref = Sys	11.6%	5.6%
Ref < Sys	46.4%	14.4%

Table 1. Automatic and human evaluation on a 500 sentence subset of the WMT'14 En-Fr and En-De test sets. Models generalize well in terms of perplexity and BLEU. Our human evaluation compares (reference, system) pairs for beam 5.

setup as Luong et al. (2015) which comprises 4.5M sentence pairs for training and we test on newstest2014. We build a validation set by removing 44k random sentence-pairs from the training data. As vocabulary we use 40k sub-word types based on a joint source and target byte pair encoding (BPE; Sennrich et al., 2016).

WMT'17 English-German (En-De): The above pre-processed version of WMT'14 En-De did not provide a split into sub-corpora which we required for some experiments. We therefore also experiment on the 2017 data where we test on newstest2017. The full version of the dataset (*original*) comprises 5.9M sentence pairs after length filtering to 175 tokens. We then consider the news-commentary portion with 270K sentences (*clean*), and a *filtered* version comprising 4M examples after removing low scoring sentence-pairs according to a model trained on the cleaner news-commentary portion.

WMT'14 English-French (En-Fr): We remove sentences longer than 175 words and pairs with a source/target length ratio exceeding 1.5 resulting in 35.5M sentence pairs for training. The source and target vocabulary is based on 40k BPE types. Results are reported on both newstest2014 and a validation set held-out from the training data comprising 26k sentence pairs.

We evaluate with tokenized BLEU (Papineni et al., 2002) on the corpus-level and the sentence-level, after removing BPE splitting. Sentence-level BLEU is computed similarly to corpus BLEU, but with smoothed n-gram counts (+1) for $n > 1$ (Lin & Och, 2004).

5. Uncertainty and Search

In this section we start by showing that the models under consideration are well trained (§5.1). Next, we quantify the amount of uncertainty in the model's output and compare two search strategies: beam search and sampling (§5.2). Finally we investigate the influence of a particular kind of extrinsic uncertainty in the data on beam search, and provide an explanation for the performance degradation observed with wide beams (§5.3).
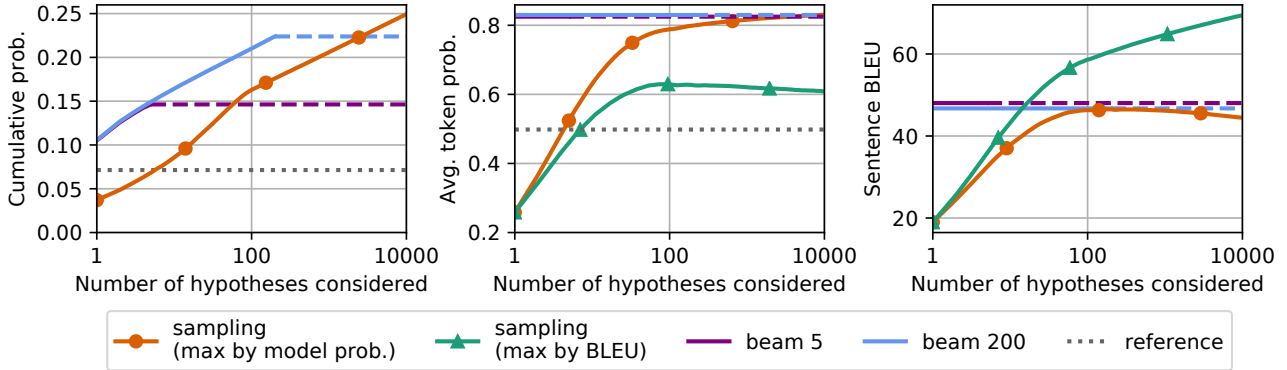
Figure 1. Left: Cumulative sequence probability of hypotheses obtained by beam search and sampling on the WMT'14 En-Fr valid set; *Center*: same, but showing the average per-token probability as we increase the number of considered hypotheses, for each source sentence we select the hypothesis with the maximum probability (orange) or sentence-level BLEU (green); *Right*: same, but showing averaged sentence-level BLEU as we increase the number of hypotheses.

5.1. Preliminary: Models Are Well Trained

We start our analysis by confirming that the models under consideration are well trained. Table 1 shows that the models, and particularly the En-Fr model, achieve low perplexity and high BLEU scores.

To further assess the quality of these models, we conducted a human evaluation with three professional translators. Annotators were shown the source sentence, reference translation, and a translation produced by our model through *beam search*—a breadth-first search that retains only the k most likely candidates at each step. Here, we consider a relatively narrow beam of size $k = 5$. The reference and model translations were shown in random order and annotators were blind to their identity. We find that model translations roughly match human translations for the En-Fr dataset, while for the En-De dataset humans prefer the reference over the model output 80% of the time. Overall, the **models are well trained**—particularly the En-Fr model— and beam search can find outputs that are highly rated by human translators.

5.2. Model Output Distribution Is Highly Uncertain

How much uncertainty is there in the model's output distribution? What search strategies are the most *effective* (i.e., produce the highest scoring hypotheses) and *efficient* (i.e., require generating the fewest candidate hypotheses)? To answer these questions we sample 10k translations and compare the hypotheses to those produced by beam search with $k = 5$ and $k = 200$.

Figure 1 (*Left*) shows that **the model's output distribution is highly uncertain**: even after drawing 10k samples we cover only 24.9% of the sequence-level probability mass. And while beam search is much more efficient at searching this space, covering 14.6% of the output probability mass with $k = 5$ and 22.4% of the probability mass with $k = 200$,
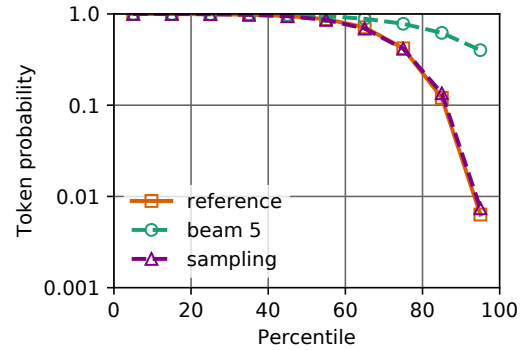

Figure 2. Probability quantiles for tokens in the reference, beam search hypotheses ($k = 5$), and sampled hypotheses for the WMT'14 En-Fr validation set.

these finding suggest that most of the probability mass is spread elsewhere in the space (see also §6.2).

Figure 1 also compares the average sentence-level BLEU and model scores of hypotheses produced by sampling and beam search. Sampling results for varying sample size $n = 1, \ldots, 10k$ are on two curves: orange reports probability (*Center*) and sentence BLEU (*Right*) for the sentence with the highest probability within n samples, while green does the same for the sentence with the highest sentence BLEU in the same set (Sokolov et al., 2008). We find that sampling produces hypotheses with similar probabilities as beam search (*Center*), however, for the same likelihood beam hypotheses have higher BLEU scores (*Right*). We also note that **BLEU and model probability are imperfectly correlated**: while we find more likely translations as we sample more candidates, BLEU over those samples eventually decreases (*Right*, orange curve). Vice versa, hypotheses selected by BLEU have lower likelihood score beyond 80 samples (*Center*, green curve). We revisit this surprising finding in §5.3.

Finally, we observe that the model on average assigns much

lower scores to the reference translation compared to beam hypotheses (Figure 1, *Center*). To better understand this, in Figure 2 we compare the token-level model probabilities of the reference translation, to those of outputs from beam search and sampling. We observe once again that **beam search is a very effective search strategy**, finding hypotheses with very high average token probabilities and rarely leaving high likelihood regions; indeed only 20% of beam tokens have probabilities below 0.7. In contrast, the probabilities for sampling and the human references are much lower. The high confidence of beam is somewhat surprising if we take into account the exposure bias (Ranzato et al., 2016) of these models, which have only seen gold translations at training time. We refer the reader to §6.2 for discussion about how well the model actually fits the data distribution.

5.3. Uncertainty Causes Large Beam Degradation

In the previous section we observed that repeated sampling from the model can have a negative impact on BLEU, even as we find increasingly likely hypotheses. Similarly, we observe lower BLEU scores for beam 200 compared to beam 5, consistent with past observations about performance degradation with large beams (Koehn & Knowles, 2017).

Why does the BLEU accuracy of translations found by larger beams deteriorate rather than improve despite these sequences having higher likelihood? To answer this question we return to the issue of extrinsic uncertainty in the training data (§3.2) and its impact on the model and search. One particularly interesting case of noise is when target sentences in the training set are simply a copy of the source.

In the WMT'14 En-De and En-Fr dataset between 1.1% and 2.0% of the training sentence pairs are "copies" (§3.2). How does the model represent these training examples and does beam search find them? It turns out that **copies are overrepresented in the output of beam search**. On WMT'14 En-Fr, beam search outputs copies at the following rates: 2.6% (beam=1), 2.9% (beam=5), 3.2% (beam=10) and 3.5% (beam=20).

To better understand this issue, we trained models on the news-commentary portion of WMT'17 English-German which does not contain copies. We added synthetic copy noise by randomly replacing the true target by a copy of the source with probability p_{noise}. Figure 3 shows that larger beams are much more affected by copy noise. Even just 1% of copy noise can lead to a drop of 3.3 BLEU for a beam of $k = 20$ compared to a model with no added noise. For a 10% noise level, all but greedy search have their accuracy more than halved.

Next, we examine model probabilities at the token-level. Specifically, we plot the average per position log-probability
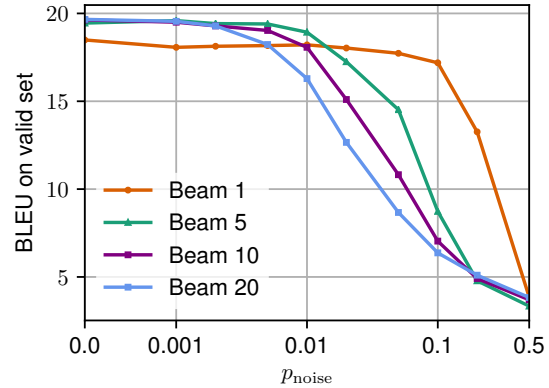

Figure 3. Translation quality of models trained on WMT'17 English-German news-commentary data with added synthetic copy noise in the training data (x-axis) tested with various beam sizes on the validation set.
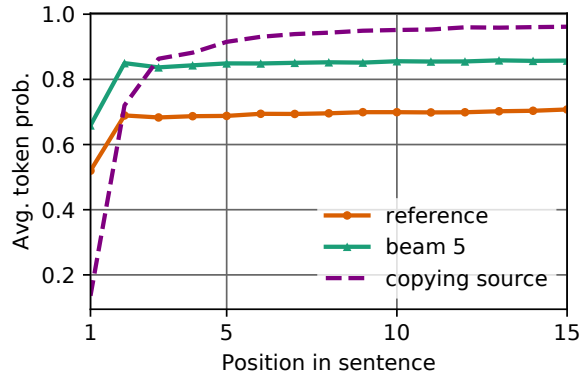

Figure 4. Average probability at each position of the output sequence on the WMT'14 En-Fr validation set, comparing the reference translation, beam search hypothesis ($k = 5$), and copying the source sentence.

assigned by the En-Fr model to each token of: (i) the reference translation, (ii) the beam search output with $k = 5$, and (iii) a synthetic output which is a copy of the source sentence. Figure 4 shows that the probability of copying the first source token is very unlikely according to the model (and actually matches the ground truth rate of copy noise). However, after three tokens the model switches to almost deterministic transitions. Because beam search proceeds in strict left-to-right manner, the copy mode is only reachable if the beam is wide enough to consider the first source word which has low probability. However, once in the beam, the copy mode quickly takes over. This explains why large beam settings in Figure 3 are more susceptible to copy noise compared to smaller settings. Thus, while larger beam widths are effective in finding higher likelihood outputs, such sequences may correspond to copies of the source sentence, which explains the drop in BLEU score for larger beams. Deteriorating accuracy of larger beams has been previously observed (Koehn & Knowles, 2017), however, it has not until now been linked to the presence of copies in the training data or model outputs.
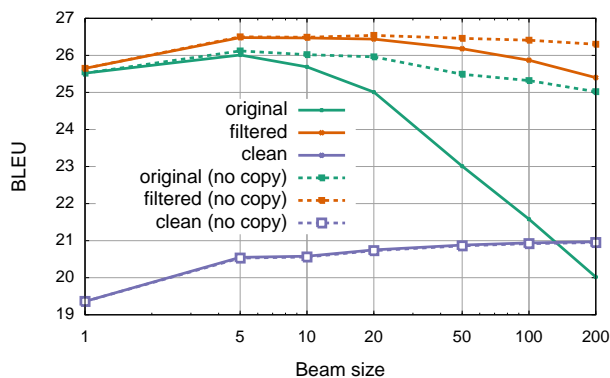
Figure 5. BLEU on newstest2017 as a function of beam width for models trained on all of the WMT'17 En-De training data (*original*), a filtered version of the training data (*filtered*) and a small but clean subset of the training data (*clean*). We also show results when excluding copies as a post-processing step (*no copy*).

Note that this finding does not necessarily imply a failure of beam nor a failure of the model to match the data distribution. Larger beams do find more likely hypotheses. It could very well be that the true data distribution is such that no good translation individually get a probability higher than the rate of copy. In that case, even a model perfectly matching the data distribution will return a copy of the source. We refer the reader to §6.2 for further analysis on this subject. The only conclusion thus far is that **extrinsic uncertainty is (at least partially) responsible for the degradation of performance of large beams**.

Finally, we present two simple methods to mitigate this issue. First, we pre-process the training data by removing low scoring sentence-pairs according to a model trained on the news-commentary portion of the WMT'17 English-German data (*filtered*; §4.2). Second, we apply an inference constraint that prunes completed beam search hypotheses which overlap by 50% or more with the source (*no copy*). Figure 5 shows that BLEU improves as beam gets wider on the clean portion of the dataset. Also, the performance degradation is greatly mitigated by both filtering the data and by constraining inference, with the best result obtained by combining both techniques, yielding an overall improvement of 0.5 BLEU over the *original* model. Appendix A describes how we first discovered the copy noise issue.

6. Model Fitting and Uncertainty

The previous section analyzed the most likely hypotheses according to the model distribution. This section takes a more holistic view and compares the estimated distribution to the true data distribution. Since exact comparison is intractable and we can only have access to few samples from the data distribution, we propose several necessary conditions for the two distributions to match. First, we inspect the
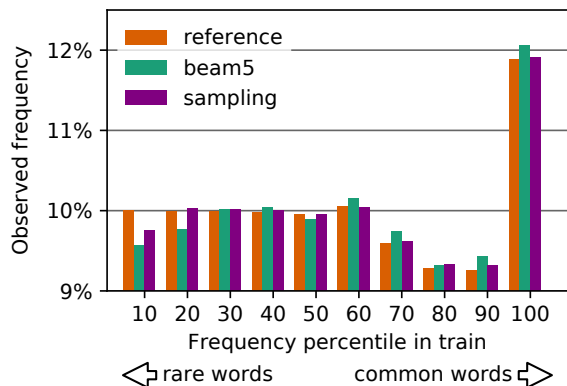

Figure 6. Unigram word frequency over the human references, the output of beam search ($k = 5$) and sampling on a random subset of 300K sentences from the WMT'14 En-Fr training set.

match for unigram statistics. Second, we move to analyze calibration at the set level and design control experiments to assess probability estimates of sentences. Finally, we compare in various ways samples from the model with human references. We find uncontroversial evidence that the model spreads too much probability mass in the hypothesis space compared to the data distribution, often under-estimating the actual probability of individual hypothesis. Appendix B outlines another condition.

6.1. Matching Conditions at the Token Level

If the model and the data distribution match, then unigram statistics of samples drawn from the two distributions should also match (not necessarily vice versa). This is a particularly interesting condition to check since NMT models are well known to under-estimate rare words (Koehn & Knowles, 2017); is the actual model poorly estimating word frequencies or is this just an artifact of beam search? Figure 6 shows that samples from the model have roughly a similar word frequency distribution as references in the training data, except for extremely rare words (see Appendix C for more analysis of this issue). On the other hand, beam search over-represents frequent words and under-represents more rare words, which is expected since high probability sequences should contain more frequent words.

Digging deeper, we perform a synthetic experiment where we select 10 target word types $w \in W$ and replace each w in the training set with either w_1 or w_2 at a given *replacement rate* $p(w_1|w)$.[2] We train a new model on this modified data and verify whether the model can estimate the original replacement rate that determines the frequency of w_1 and w_2. Figure 7 compares the replacement rate in the data (prior) to the rate measured over the output of either beam search or sampling. Sampling closely matches the data distribution

---

[2]Each replaced type has a token count between 3k-7k, corresponding to bin 20 in Fig. 6. $|W| = 50k$.
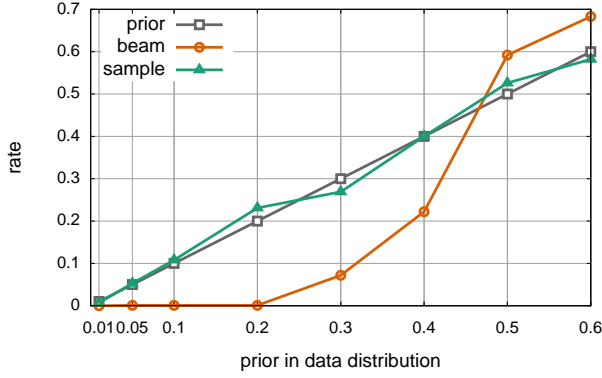
Figure 7. Comparison of how often a word type is output by the model with beam search or sampling compared to the data distribution; prior is the data distribution. Values below prior underestimate the data distribution and vice versa.

for all replacement rates but beam greatly overestimates the majority class: it either falls below the prior for rates of 0.5 or less, or exceeds the prior for rates larger than 0.5. These observations confirm that the **model closely matches unigram statistics except for very rare words, while beam prefers common alternatives to rarer ones**.

6.2. Matching Conditions at the Sequence Level

In this section, we further analyze how well the model captures uncertainty in the data distribution via a sequence of necessary conditions operating at the sequence level.
Set-Level Calibration. Calibration (Guo et al., 2017; Kuleshov & Liang, 2015) verifies whether the model probability estimates p_m match the true data probabilities p_d. If p_d and p_m match, then for any set S, we observe:

$$\mathbb{E}_{x \sim p_d}\left[\mathbb{I}\{x \in S\}\right] = p_m(S).$$

The left hand side counts the expected number of times we observe a sample in S, when sampling from the data distribution; the right hand side sums the model probability estimates over S.

In Figure 8, we plot the left hand side against the right hand side where S is a set of 200 beam search hypotheses on the WMT'14 En-Fr validation set, covering an average of 22.4% of the model's probability mass. Points are binned so that each point represents 10% of sentences in the validation or test set (Nguyen & O'Connor, 2015). For instance, the rightmost point in the figure corresponds to sentences for which beam collects nearly the entire probability mass, typically very short sentences. This experiment shows that the **model matches the data distribution remarkably well at the set level** on both the validation and test set.

Control Experiment. To assess the fit to the data distribution further, we re-consider the models trained with varying levels of copy noise (p_{noise}, cf. §5.3) and check whether
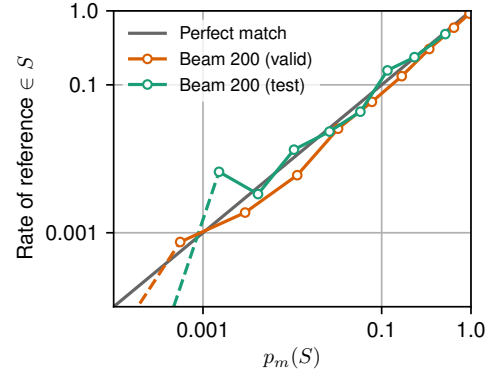

Figure 8. Matching distributions at the set level using 200 beam search hypotheses on the WMT'14 En-Fr valid and test set. Points are binned so that each represents 10% of sentences. The lowest probability bin (not shown) has value 0 (reference never in S).
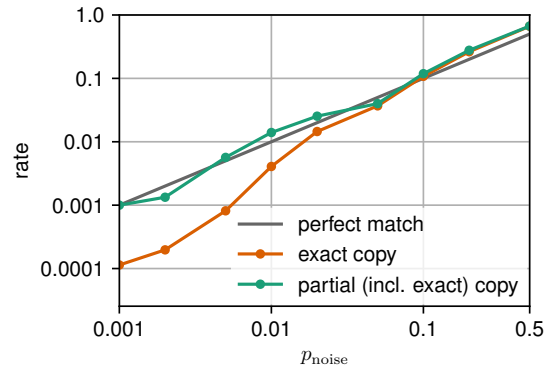

Figure 9. Rate of copy of the source sentence (exact and partial) as a function of the amount of copy noise present in the model's train data (§5.3). Results on WMT'17 En-De validation set.

we reproduce the correct amount of copying (evaluated at the sequence level) when sampling from the model. Figure 9 shows a large discrepancy: at low p_{noise} the model underestimates the probability of copying (i.e., too few of the produced samples are *exact copies* of the source), while at high noise levels it overestimates it. Moreover, since our model is smooth, it can assign non-negligible probability mass to partial copies[3] which are *not* present in the training data. Indeed, when we consider both partial and exact copies, the model correctly reproduces the amount of copy noise present in the training data. Therefore, although the model appears to under-estimate some hypotheses at low copy rates, it actually **smears probability mass in the hypothesis space**. Overall, this is the first concrete evidence of the model distribution not perfectly fitting the data distribution.

Expected Inter-Sentence BLEU is defined as

$$\mathbb{E}_{x \sim p, x' \sim p}\left[BLEU(x, x')\right]$$

---

[3]Partial copies are identified using the IoU at 50% criterion, described in §3.2.

which corresponds to the expected BLEU between two translations sampled from a distribution p where x is considered as the hypothesis and x' as the reference. If the model matches the data distribution, then the expected BLEU computed with sentences sampled from the model distribution p_m should match the expected BLEU computed using two independent reference translations (see (§6.3) for further details on data collection).

We find that the expected BLEU is 44.5 and 32.1 for human translations on the WMT'14 En-Fr and WMT'14 En-De datasets, respectively. However, the expected BLEU of the model is only 28.6 and 24.2, respectively. This large discrepancy provides further evidence that **the model spreads too much probability mass across sequences**, compared to what we observe in the actual data distribution.

6.3. Comparing Multiple Model Outputs to Multiple References

Next we assess whether model outputs are similar to those produced by multiple human translators. We collect 10 additional reference translations from 10 distinct humans translators for each of 500 sentences randomly selected from the WMT'14 En-Fr and En-De test sets. Similarly, we collect a large set of translations from the model via beam search ($k = 200$) or sampling. We then compute two versions of *oracle BLEU* at the sentence-level: (i) *oracle reference* reports BLEU for the most likely hypothesis with respect to its best matching reference (according to BLEU); and (ii) *average oracle* computes BLEU for every hypothesis with respect to its best matching reference and averages this number over all hypotheses. Oracle reference measures whether one of the human translations is similar to the top model prediction, while average oracle indicates whether most sentences in the set have a good match among the human references. The average oracle will be low if there are hypotheses that are dissimilar from all human references, suggesting a possible mismatch between the model and the data distributions.

Table 2 shows that beam search (besides its degradation due to copy noise) produces not only top scoring hypothesis that are very good (single reference scoring at 41 and oracle reference at 70) but most hypotheses in the beam are close to a reference translation (as the difference between oracle reference and average oracle is only 5 BLEU points). Unfortunately, beam hypotheses lack diversity and they are all close to very few references as indicated by the coverage number, which measures how many *distinct* human references are matched to at least one of the hypotheses. In contrast, hypotheses generated by sampling exhibit opposite behavior: the quality of the top scoring hypothesis is lower, several hypotheses poorly match references (as indicated by the 25 BLEU points gap between oracle refer-

	beam		sampling
	$k = 5$	$k = 200$	$k = 200$
Prob. covered	4.7%	11.1%	6.7%
Sentence BLEU			
single reference	41.4	36.2	38.2
oracle reference	70.2	61.0	64.1
average oracle	65.7	56.4	39.1
- # refs covered	1.9	5.0	7.4
Corpus BLEU (`multi-bleu.pl`)			
single reference	41.6	33.5	36.9
10 references	81.5	65.8	72.8

Table 2. Sentence and corpus BLEU for beam search hypotheses and 200 samples on a 500 sentence subset of the WMT'14 En-Fr test set. "Single reference" uses the provided reference and the most likely hypothesis, while oracle reference and average oracle are computed with 10 human references.

ence and average oracle) but coverage is much higher. This finding is again consistent with the previous observation that the **model distribution is too spread in hypothesis space**. We conjecture that the excessive spread may also be partly responsible for the lack of diversity of beam search, as probability mass is spread across similar variants of the same sequence even in the region of high likelihood. This over-smoothing might be due to the function class of NMT; for instance, it is hard for a smooth class of functions to fit a delta distribution (e.g., a source copy), without spreading probability mass also in nearby hypotheses (e.g., partial copies), or to assign exact 0 probability in space, resulting in an overall under-estimation of hypotheses present in the data distribution.

7. Conclusions and Final Remarks

In this study we investigate the effects of uncertainty in NMT model fitting and search. We found that search works remarkably well. While the model is generally well calibrated both at the token and sentence level, it tends to diffuse probability mass too much. We have not investigated the causes of this, although we surmise that it is largely due to the class of smooth functions that NMT models can represent. We instead investigated some of the effects of this mismatch. In particular, excessive probability spread causes poor quality samples from the model. It may also cause the "copy mode" to become more prominent once the probability of genuine hypotheses gets lowered. We show that this latter issue is linked to a form of extrinsic uncertainty which causes deteriorating accuracy with larger beams. Future work will investigate even better tools to analyze distributions and leverage this analysis to design better models.

References

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

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. Neural versus phrase-based machine translation quality: a case study. In *EMNLP*, 2016.

Blunsom, P., Cohn, T., and Osborne, M. A discriminative latent variable model for statistical machine translation. In *ACL-HLT*, 2008.

Cho, K. Noisy parallel approximate decoding for conditional recurrent language model. *arXiv:1605.03835*, 2016.

Dreyer, M. and Marcu, D. Hyter: Meaning-equivalent semantics for translation evaluation. In *ACL*, 2012.

Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., and Dolan, B. deltableu: A discriminative metric for generation tasks with intrinsically diverse targets. *arXiv preprint arXiv:1506.06863*, 2015.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. Convolutional Sequence to Sequence Learning. In *Proc. of ICML*, 2017. URL github.com/facebookresearch/fairseq-py.

Gimpel, K., Batra, D., Dyer, C., and Shakhnarovich, G. A systematic exploration of diversity in machine translation. In *Conference on Empirical Methods in Natural Language Processing*, 2013.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Isabelle, P., Cherry, C., and Foster, G. A challenge set approach to evaluating machine translation. *arXiv:1704.07431*, 2017.

Koehn, P. and Knowles, R. Six challenges for neural machine translation. *arXiv:1706.03872*, 2017.

Kuleshov, V. and Liang, P. Calibrated structured prediction. In *Neural Information Processing Systems*, 2015.

Li, J. and Jurafsky, D. Mutual information and diverse decoding improve neural machine translation. *arXiv:1601.00372*, 2016.

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

Luong, M.-T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. In *Proc. of EMNLP*, 2015. URL http://nlp.stanford.edu/projects/nmt.

Nguyen, K. and O'Connor, B. Posterior calibration and exploratory analysis for natural language processing models. In *Proc. of EMNLP*, 2015.

Niehues, J., Cho, E., Ha, T.-L., and Waibel, A. Analyzing neural mt search and model performance. *arXiv:1708.00563*, 2017.

Padó, S., Cer, D., Galley, M., Jurafsky, D., and Manning, C. D. Measuring machine translation quality as semantic equivalence: A metric based on entailment features. *Machine Translation*, 23:181–193, 2009.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

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

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Association for Computational Linguistics*, 2016.

Smith, J. R., Saint-Amand, H., Plamada, M., Koehn, P., Callison-Burch, C., and Lopez, A. Dirt cheap web-scale parallel text from the common crawl. In *ACL*, 2013.

Sokolov, A., Wisniewski, G., and Yvon, F. Lattice bleu oracles in machine translation. *ACM Transactions on Speech and Language Processing*, 2008.

Sutskever, I., Martens, J., Dahl, G. E., and Hinton, G. E. On the importance of initialization and momentum in deep learning. In *ICML*, 2013.

Toral, A. and Sanchez-Cartagena, V. M. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *ACL*, 2017.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *arXiv*, 2017.

Venuti, L. *The Translator's Invisibility*. Routledge, 2008.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv:1610.02424*, 2016.

Zhang, B., Xiong, D., Su, J., Duan, H., and Zhang, M. Variational neural machine translation. In *EMNLP*, 2016.

Zoph, B., Ghazvininejad, M., and Knight, K. How much information does a human translator add to the original? In *EMNLP*, 2015.