Appendices

In this supplementary material, we report three additional findings. First, we discuss the experiments that made us realize that larger beam degradation was due to copy noise in the training data. Second, we introduce another necessary condition for the model and the data distribution to match, which is based on the observation that for some source sentences we do have access to hundreds of references, and therefore, we can directly check whether the two distributions match over the set of unique references. And finally, we provide a more in depth analysis of the unigram statistics mis-match.

A. How We Discovered Copy Noise

In this section we report the initial experiment which led us to the realization that degradation of large beams is due to noise in the training data, as the process may be instructive also for other researchers working in this area.

A nice visualization of samples drawn from the model is via a scatter plot of log-probability VS. BLEU, as shown in Figure **??** for four sentences picked at random from the test set of WMT'14 En-Fr.

First, this plot shows that while high BLEU implies high log-likelihood, the vice versa is not true, as low BLEU scoring samples can have wildly varying log-likelihood values.

Second, the plot makes very apparent that there are some outlier hypotheses that nicely cluster together.

For instance, there are two clusters corresponding to sentence id 2375, marked with (2) and (3) in Figure **??**. These clusters have relatively high log-likelihood but very different BLEU score. The source sentence is:
"*Should this election be decided two months after we stopped voting?*"
The target reference is:
"*Cette élection devrait-elle ëtre décidé deux mois après que le vote est terminé?*"
while a sample from cluster (2) is:
"*Ce choix devrait-il ëtre décidé deux mois après la fin du vote?*"
and a sample from cluster (3) is:
"*Cette élection devrait-elle ëtre décidée deux mois aprës l'arrët du scrutin?*"
This example shows that translation (2), which is a valid translation, gets a low BLEU because of a choice of a synonym word with different gender which causes all subsequent words to be inflected differently, yielding overall a very low n-gram overlap with the reference, and hence a low BLEU score. This is an example of the model nicely capturing (intrinsic) uncertainty, but the metric failing at acknowledging that.

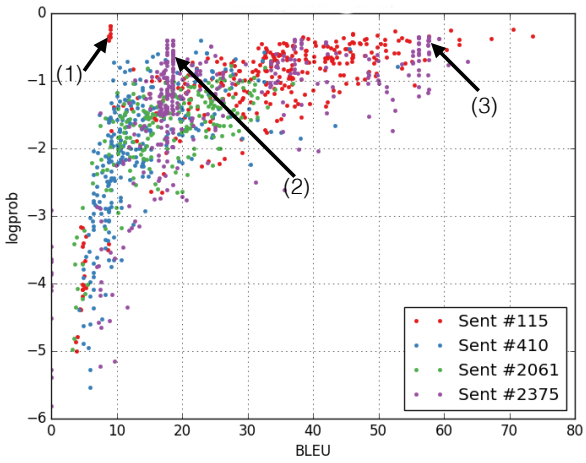

Figure 10. Scatter plot showing log-probability and BLEU of samples drawn from the model for four sentences taken from the test set of WMT'14 En-Fr (each color corresponds to a different test sentence). (1) shows samples where the model copied the source sentence, yielding very large likelihood but low BLEU. (2) and (3) are valid translations of the same source sentence, except that (2) is a cluster of samples using different choice of words.

Let's now look at cluster (1) of sentence id 115. This cluster achieves extremely high log-likelihood but also extremely low BLEU score. The source sentence is:
"*The first nine episodes of Sheriff [unk]'s Wild West will be available from November 24 on the site [unk] or via its application for mobile phones and tablets.*"
The target reference is:
"*Les neuf premiers épisodes de [unk] [unk] s Wild West seront disponibles à partir du 24 novembre sur le site [unk] ou via son application pour téléphones et tablettes.*"
while a sample from cluster (1) is:
"*The first nine episodes of Sheriff [unk] s Wild West will be available from November 24 on the site [unk] or via its application for mobile phones and tablets.*"
In this case, the model copies almost perfectly the source sentence. Examples like these made us discover the "copy issue", and led us to then link beam search degradation to systematic mistakes in the data collection process.

In conclusion, lots of artifacts and translation issues can be easily spotted by visualizing the data and looking at clusters of outliers.

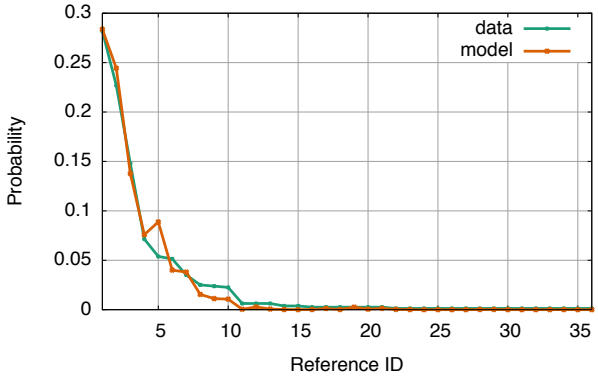

Figure 11. Comparison between the data and the model distributions for the source sentence "(The president cutoff the speaker)". The data distribution is estimated over 798 references of which 36 are unique. The hypotheses of the data distribution (x-axis) are sorted in descending order of empirical probability mass. The model matches rather well the data distribution.

B. Another Necessary Condition: Matching the Full distribution for a Given Source

In §6 we have investigated several necessary conditions for the model distribution to match the data distribution. Those conditions give an aggregate view of the match and they are mostly variants of calibration techniques, whereby the data distribution is approximated via Monte Carlo samples (human translations), since that is all we have access to.

Ideally, we would like to check the two distributions by evaluating their mass at every possible target sequence, but this is clearly intractable and not even possible since we do not have access to the actual data distribution.

However, there are sentences in the training set of WMT'14 En-Fr (EuroParl corpus) that appear several times. For instance, the source sentence "(The President cut off the speaker)." appears almost 800 times in the training set with 36 unique translations. For such cases, we can then have an accurate estimate of the ground truth data distribution (for that given source sentence) and check the match with the model distribution. This is yet another necessary condition: if the model and data distribution match, they also match for a particular source sentence.

Figure **??** shows that for this particular sentence the model output distribution closely matches the data distribution.

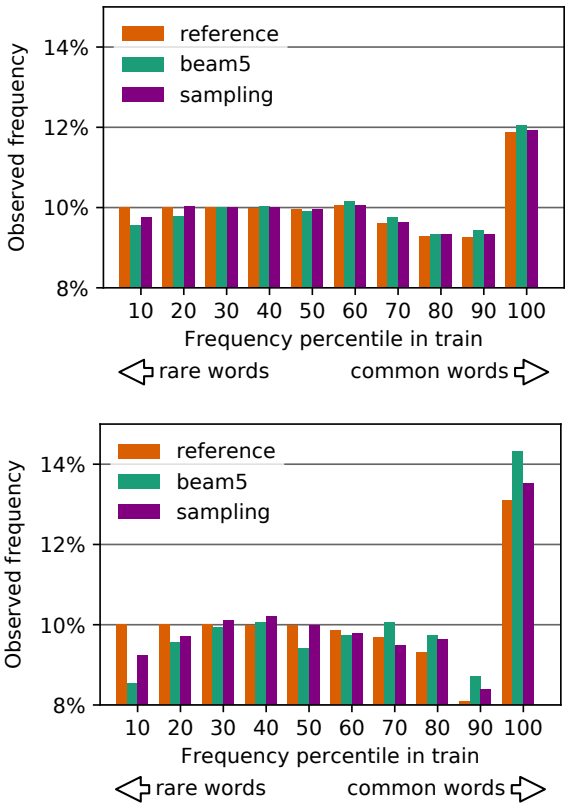

Figure 12. Unigram word frequency over the human references, the output of beam search ($k = 5$) and sampling on the WMT'14 En-Fr (top) and WMT'17 En-De news-commentary portion (bottom) of the training set.

C. Does More Data Help?

The findings reported in this paper are quite robust to the choice of architecture as well as dataset. For instance, we compare in Figure **??** the binned unigram word frequencies on the smaller news-commentary portion of the WMT'17 En-De dataset with the larger WMT'14 En-Fr dataset (which was already reported in Figure 6). The En-Fr data is about 100 times bigger than the En-De news-commentary dataset, as described in §4.2 and the En-Fr model performs much better than the En-De model, with a BLEU of 41 versus only 21 (see Table 1 and Figure 5). We observe the same tendency of the model to under-estimate very rare words (compare `beam5` vs. `reference` in the 10 percentile bin). However, the under-estimation is much more severe in the En-De model, nearly 1.5% as opposed to only 0.4%. Note that the median frequency of words in the 10 percentile bin is only 12 for the En-De dataset, but is 2552 for the En-Fr dataset. The NMT model clearly needs more data to better estimate its parameters and fit the data distribution.