# Are you certain about this translation?

## Detecting Out-of-Distribution Translations with Variational Transformers

Tim Z. Xiao, Aidan N. Gomez, Yarin Gal
OATML, Department of Computer Science, University of Oxford
tim.z.xiao@outlook.com, {aidan.gomez, yarin}@cs.ox.ac.uk

UNIVERSITY OF OXFORD

### Challenge: Flag Out-Of-Distribution (OOD) Data

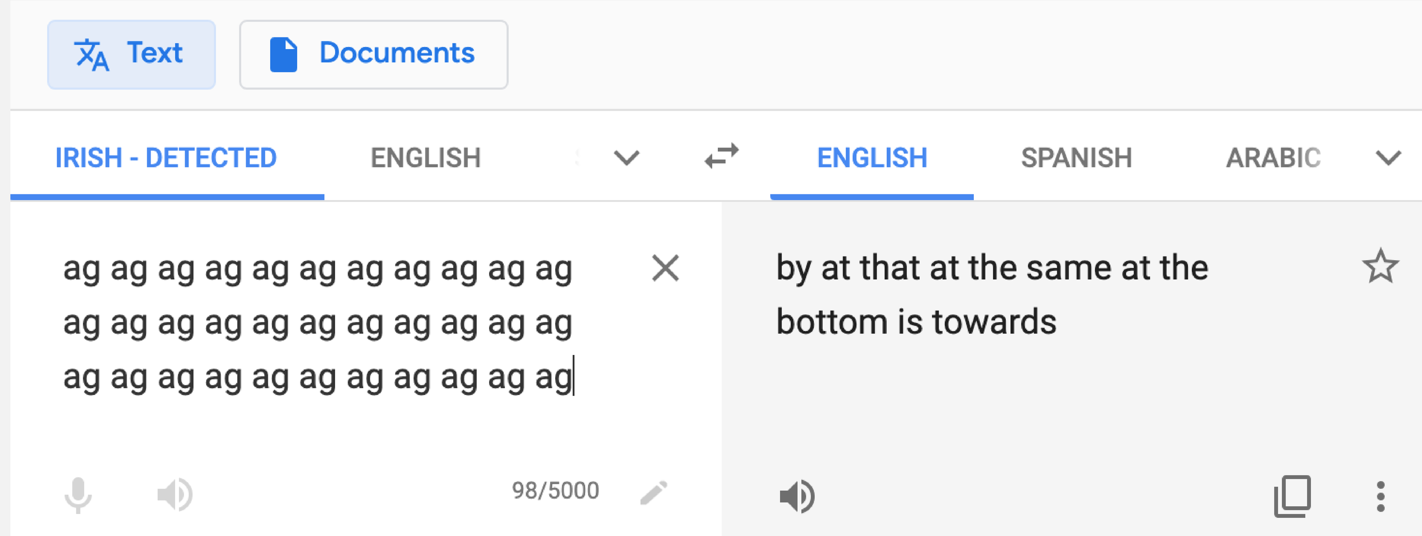



Figure: A screenshot from Google Translate.

- Current NMT models fail to provide uncertainty estimates for their translations.
- If input lies outside of the training data distribution, the models are not able to distinguish and flag them.

**Can we estimate uncertainty and identify OOD data in Neural Machine Translation (NMT)?**

### Existing Uncertainty Measures

- In **regression**, we can use the **variance** of samples as an uncertainty estimate.
- In **classification**, there are a few approaches: **variation ratios, predictive entropy, mutual information** etc.

However, translation is neither a regression or classification task, and the above measures are not applicable. There are some attempts at investigating uncertainty in NMT tasks:

- **Kumar & Sarawagi**, 2019
  - Found that the predictive probability distribution over the vocabulary used during decoding is not a good reference for model uncertainty.
- **Ott et al.**, 2018
  - Found that the model has a highly uncertain output predictive distribution in the way that probability mass at the sequence level spread widely over the hypothesis space.

**Missing effective uncertainty measures for out-of-distribution detection in NMT.**

### Proposed Measures

We investigate several measures of uncertainty appropriate for **long sequences of discrete variables (e.g. sentences)**:

1. **Beam Score** (Baseline): we assign a confidence to output $y$ generated (using beam search) from input $x$ using the score assigned to $y$'s beam (Wu et al., 2016).

$$BS = \log(p(y|x)) \;/\; \text{length\_penalty}(y; 0.6)$$

$$\text{length\_penalty}(y;\alpha) = \left(\frac{5+|y|}{5+1}\right)^{\alpha}$$

2. **Sequence Probability**: we assign a confidence to output $y$ generated from input $x$ by taking the log predictive probability under the weight distribution.

$$SP = \log(\mathbb{E}_{\theta\sim q(\theta)}\, p_\theta(y|x)) \;/\; \text{length\_penalty}(y; 0.6)$$

3. **BLEU Variance**: we assign uncertainty at an input $x$ by producing pairs of outputs from the model and measuring the squared complement of the BLEU (Papineni et al., 2002) to judge disagreement between model outputs on input $x$.

$$\text{BLEUVar} = \mathbb{E}_{\theta\sim q(\theta)}\mathbb{E}_{y,y'\sim p_\theta(y|x)}\,(1-\text{BLEU}(y,y'))^2$$

### How do we estimate these measures?

1. **Beam Score** (Baseline):
   - We use the **deterministic model** found by gradient descent and simply take the probabilities from under its predictive distribution.
2. **Sequence Probability**:
   - We use **MC Dropout** (Gal, 2016) and take a number of samples ($N$) to estimate the expectations:

$$SP \approx \log\left(\sum_{i=1}^{N} p_{\theta_i}(y|x)\right) \;/\; \text{length\_penalty}(y; 0.6)$$

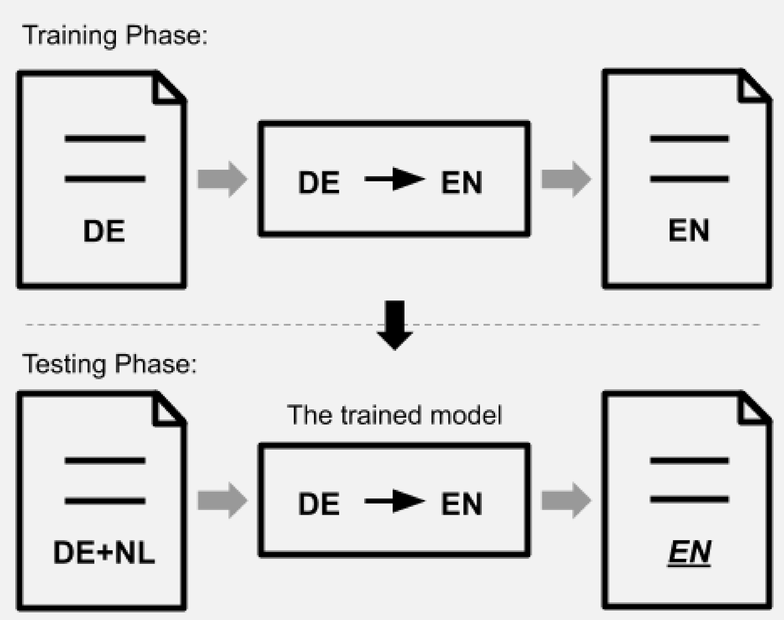
3. **BLEU Variance**:
   - With **MC Dropout**, we use the results from beam search applied to $N$ different model samples and measuring the complement BLEU between pairs of these outputs:

$$\text{BLEUVar} \approx \sum_{i=1}^{N}\sum_{j\neq i}^{N}\left(1-\text{BLEU}(\text{decode}_{\theta_i}(x), \text{decode}_{\theta_j}(x))\right)^2.$$

### Evaluating Uncertainty in Sequence Models

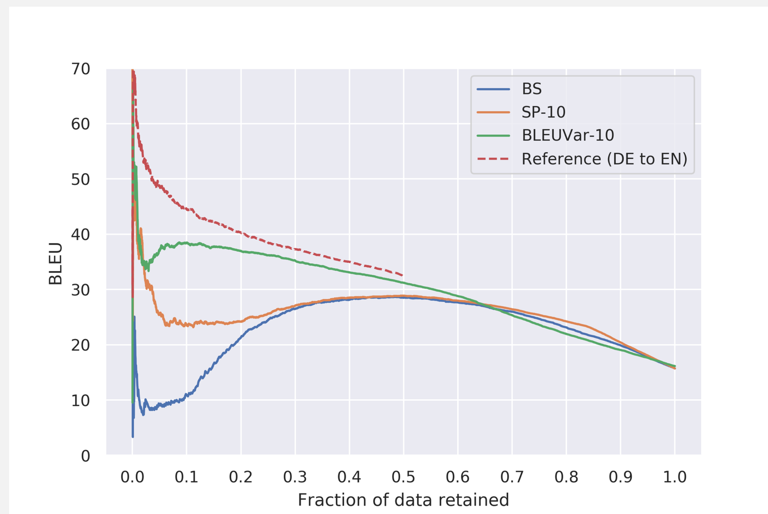
- The **performance-retention curve** indicates how well an uncertainty measure would perform if the $k$% least certain outputs were deleted from the test dataset.
- The $x$-axis ranges along the fraction of data retained, while the $y$-axis measures some performance metric of the model on the retained data (Filos et al., 2019).
- **A good uncertainty measure shows improvement in performance as low-confidence predictions are excluded from the test set**

### Results

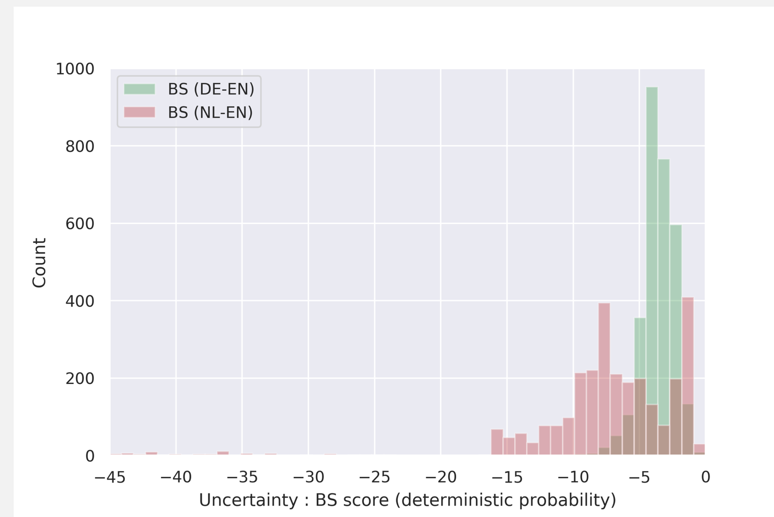



- **Trained** a **Transformer** model on WMT 13 and Europarl **DE (German) to EN (English)** sentence pairs (obtaining BLEU 33 on the WMT 14 test set).
- **Evaluated** the model on out-of-training-distribution input sentences in **NL (Dutch)** and a mixture of **NL and DE**.

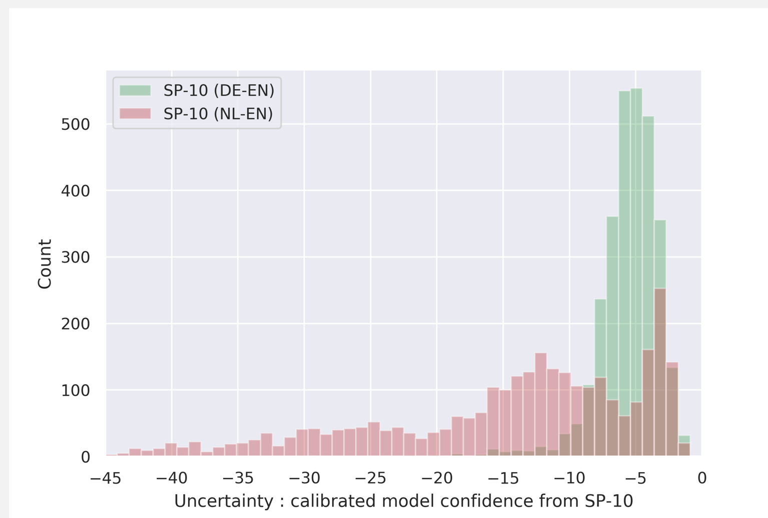
- **OOD separation** with a mixed test set **DE+NL** (# samples $N = 10$):



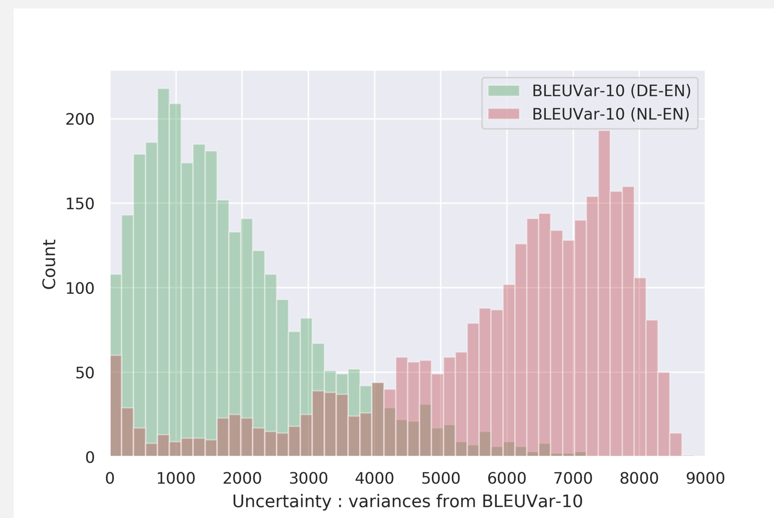

(a) Performance-retention curves on a mixed test set (DE+NL).




(b) **BS** histogram on a mixed test set (Left - higher uncertainty)




(c) **SP-10** on a mixed test set (Left - higher uncertainty)




(d) **BLEUVar-10** on a mixed test set (Right - higher uncertainty)

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.