# Is MAP Decoding All You Need?

## The Inadequacy of the Mode in Neural Machine Translation

Bryan Eikema  Wilker Aziz

University of Amsterdam

COLING 2020, Barcelona, Spain

# Key Takeaways

We question the use of MAP decoding in NMT

We show that:

- MAP decoding **introduces biases**
- The mode is a very **rare event**
- NMT models **capture data statistics well**

We argue that:

- MAP decoding is **not suitable** for NMT
- We should base model criticism and predictions on **unbiased samples**

# Neural Machine Translation (NMT)

NMT is trained as **probabilistic model** using maximum likelihood estimation (MLE)

We generate translations using ***maximum a-posteriori*** (MAP) decoding:

$$y^{mode} = \text{argmax}_y \, P(y|x, \theta_{MLE})$$

Finding the exact MAP is **intractable**, so we use an approximation: **beam search**

# Pathologies and Biases of NMT

- Length bias
- Beam search curse
- Inadequacy of the mode
- Exposure bias
- Non-admissible heuristic search bias

Many works blame NMT as a **model or its training algorithm**

But note: all these observations are using **approximate MAP decoding**

# Biased Statistics & The Inadequacy of the Mode

We use the mode for **model criticism**, but:

- The mode is **no unbiased statistic** of the learnt distribution
  - e.g. a short mode does not imply that the model underestimates average sequence length!

We target the mode for **making predictions**, but:

- The mode could still be a **very rare event**
- Focusing on the mode alone **throws away a lot of valuable information** learnt by the model

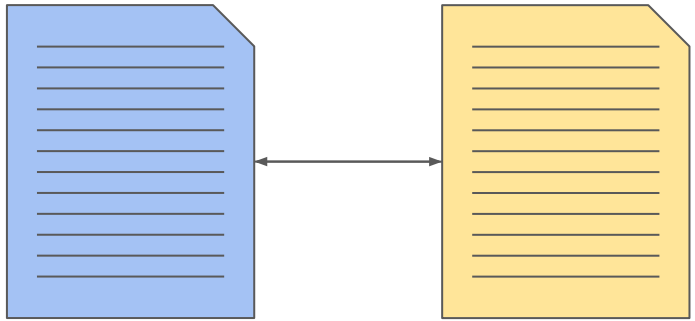
# Experiments

We will be answering:

1. Does the NMT model fit the data well?
2. What do the learnt distributions look like?
3. Can we make predictions using all of the information available?

# Experiments

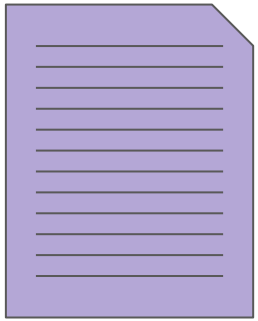
Train on:



English

German (5.9M)
Nepali (573k)
Sinhala (235k)

Test on:



newstest2018
Flores
Flores

Model:

# Assessing Data Fit

# Assessing Data Fit: Methodology

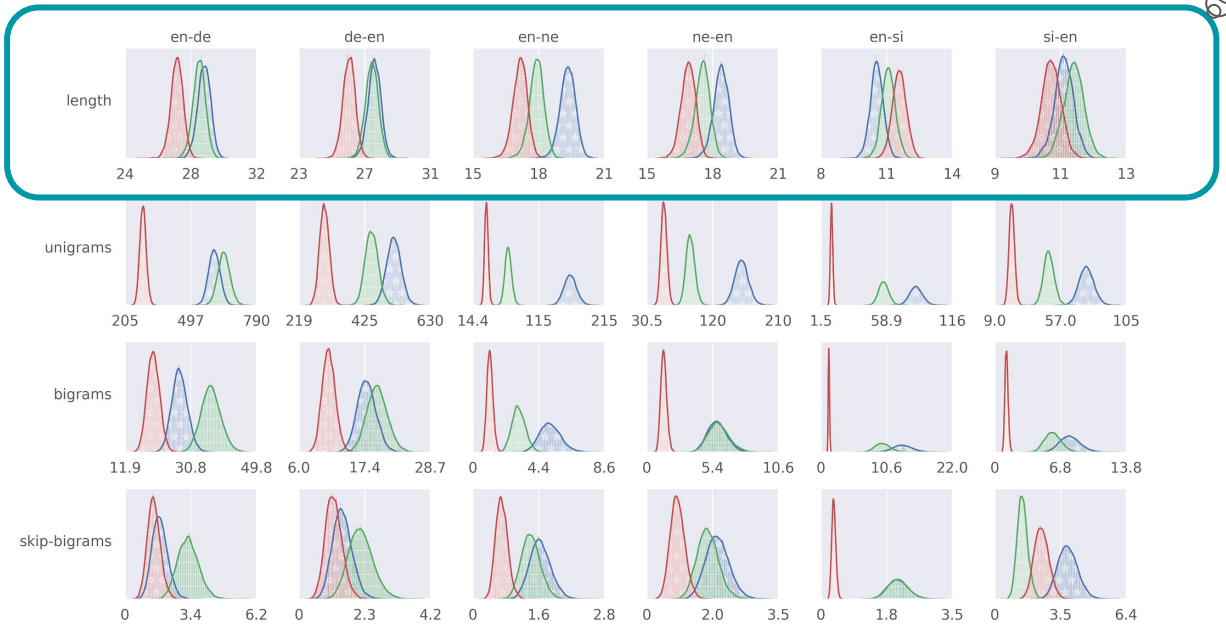
1. Gather statistics from **data**, **unbiased samples**, and **beam search** outputs
2. Model all data in a hierarchical Bayesian model
3. Compare posteriors between data and model output

We compare:

- Length
- Lexical properties: unigram and bigram counts
- Word order: skip-bigram counts

# Assessing Data Fit: Length

x-axis shows "average length"

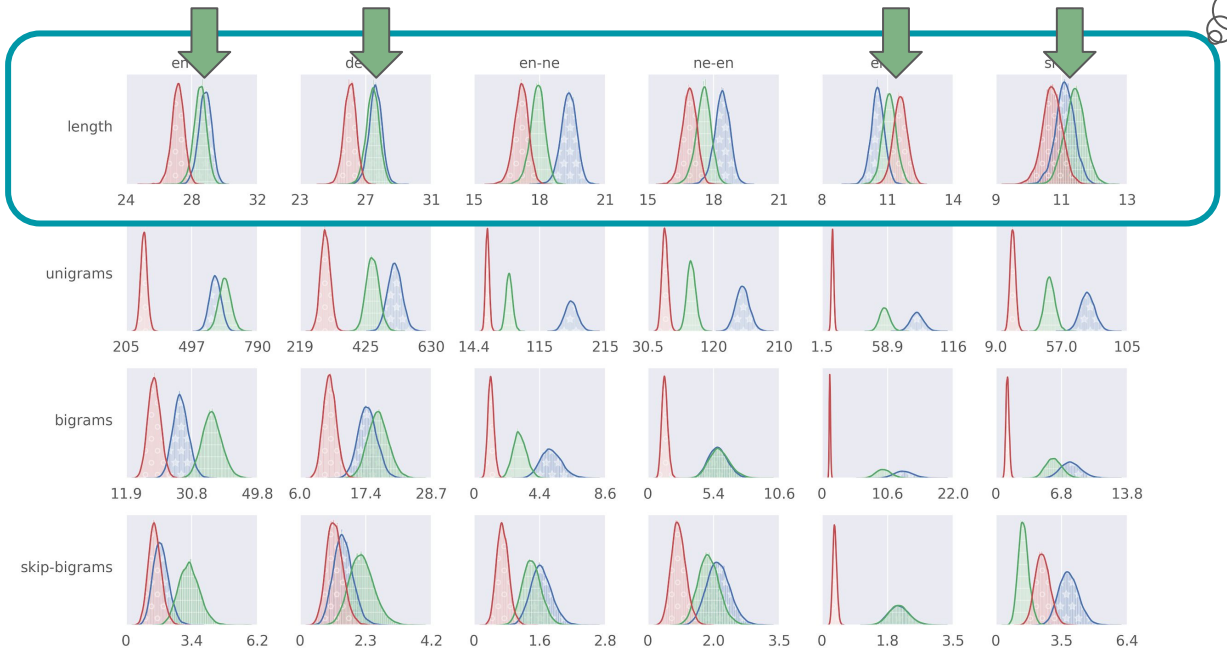



Beam Search

Samples

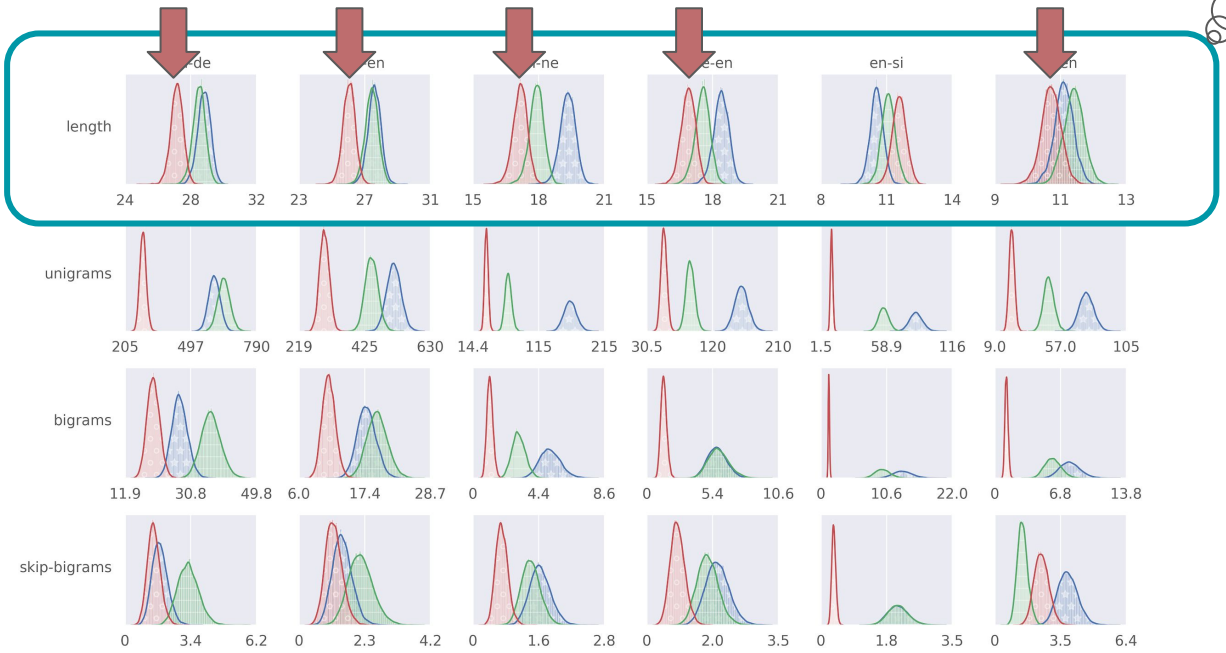
References

# Assessing Data Fit: Length



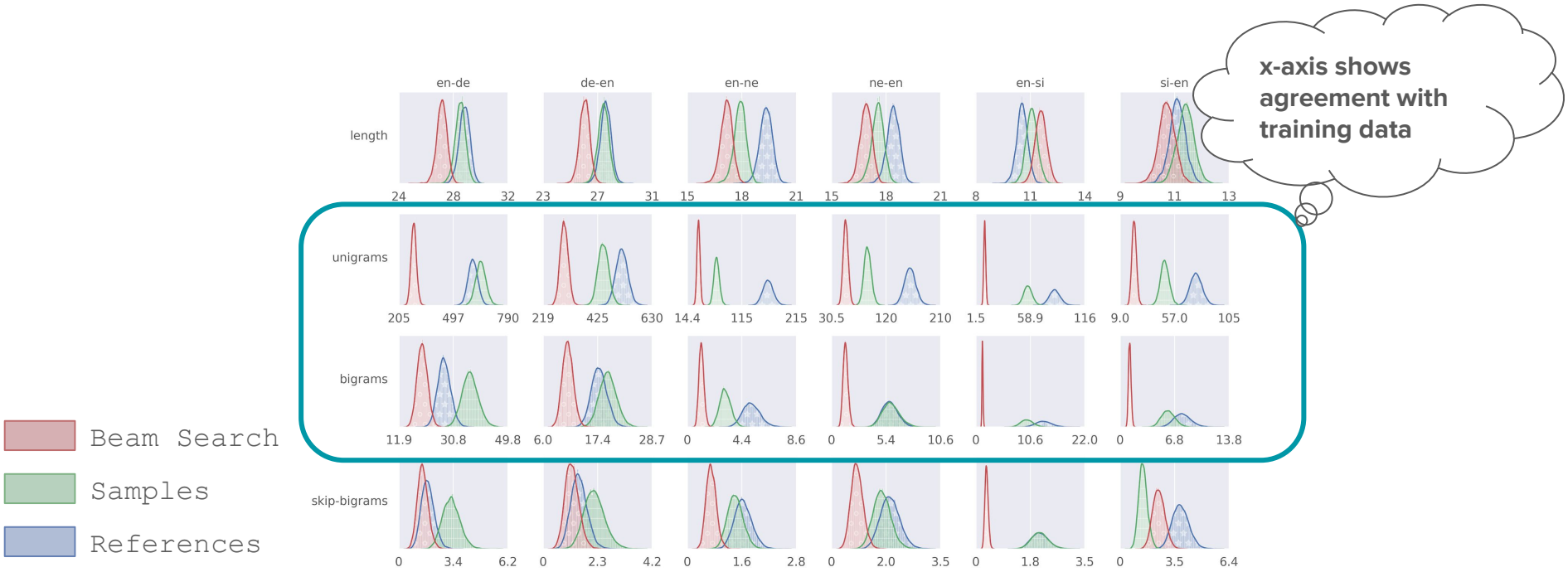

In most cases the **model captures length reasonably well**

# Assessing Data Fit: Length

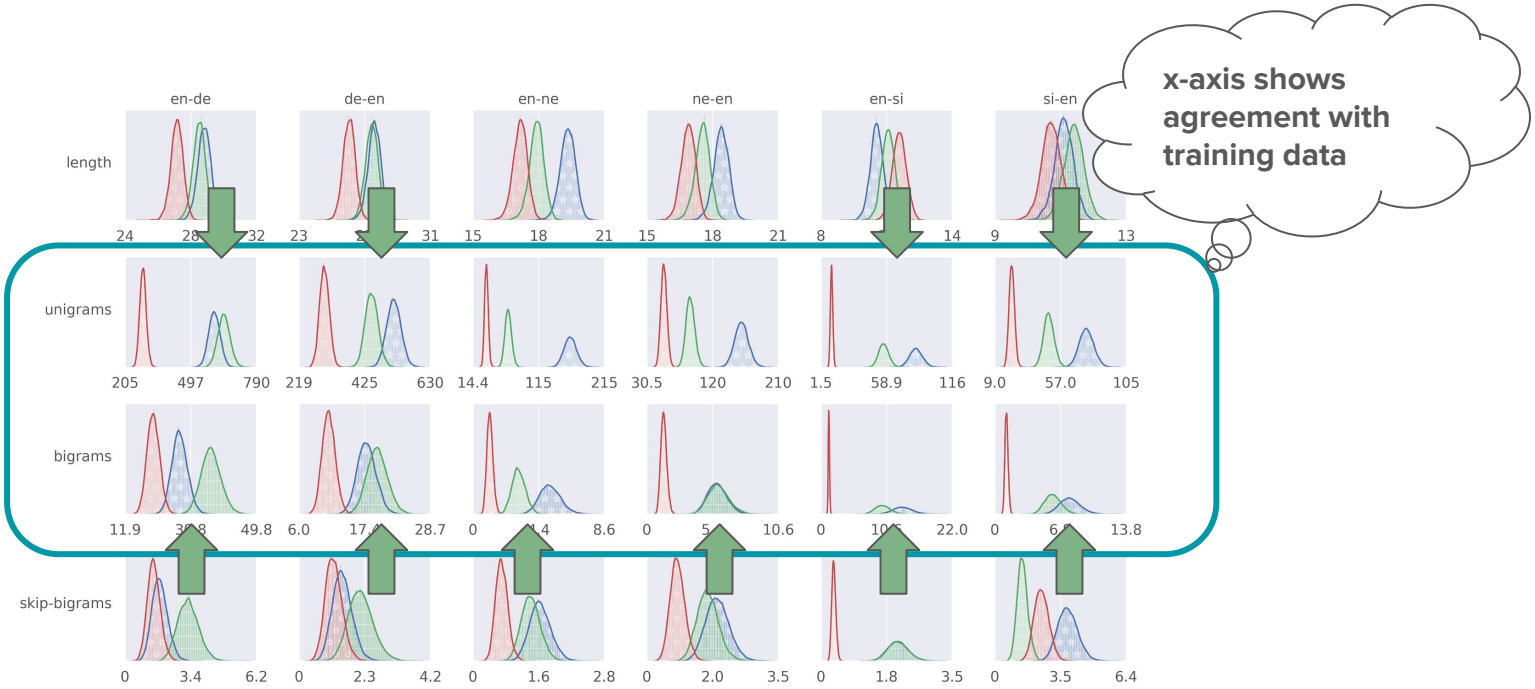



**Beam search shifts from data statistics**, underestimating length

# Assessing Data Fit: Lexical Statistics

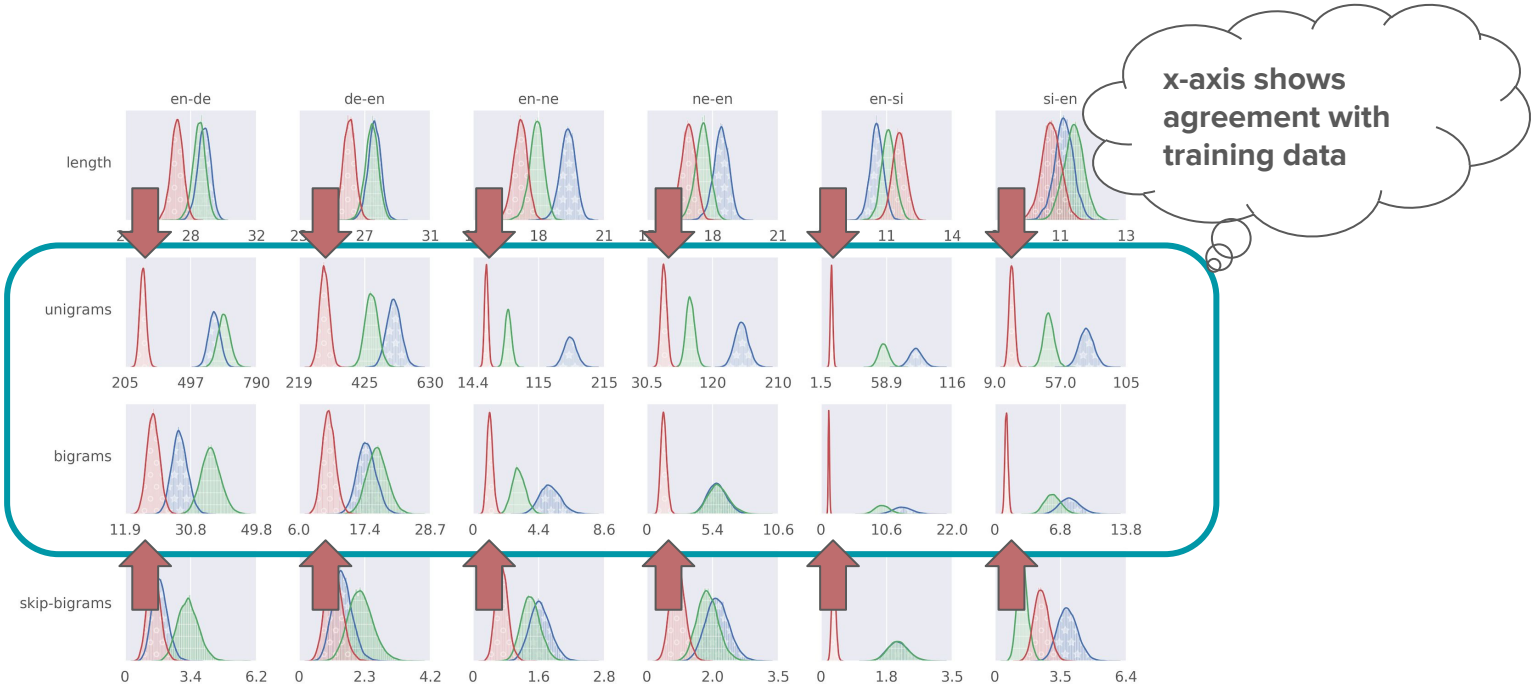


# Assessing Data Fit: Lexical Statistics



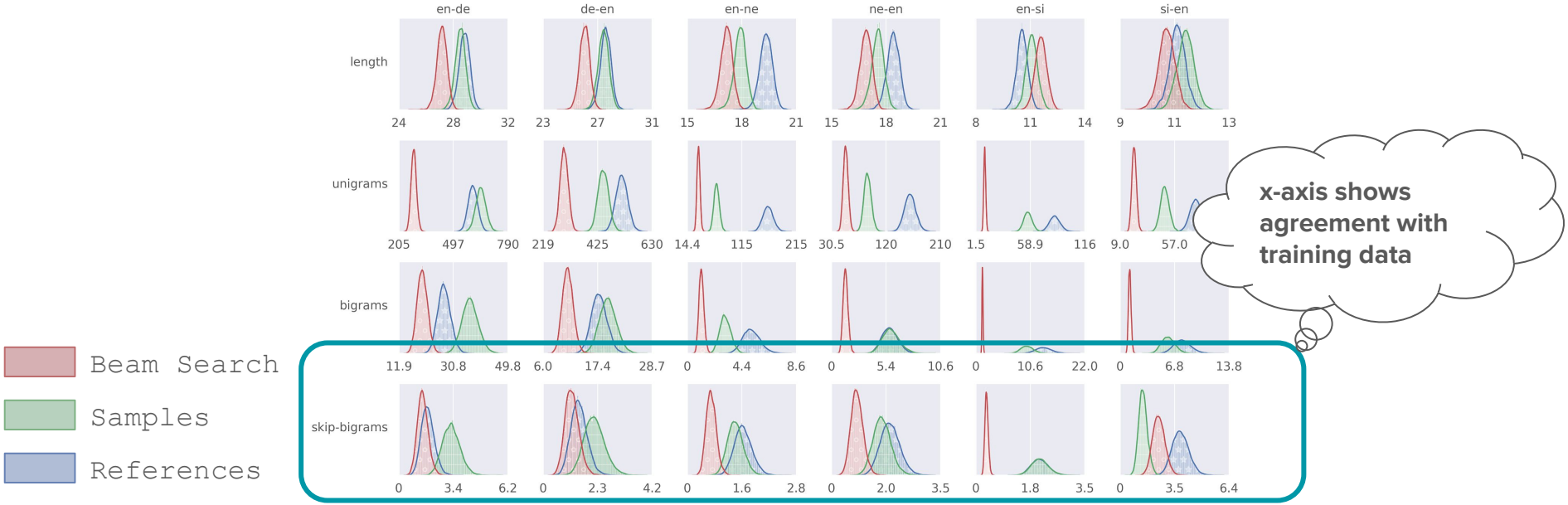

In most cases the **model captures lexical statistics reasonably well**

# Assessing Data Fit: Lexical Statistics

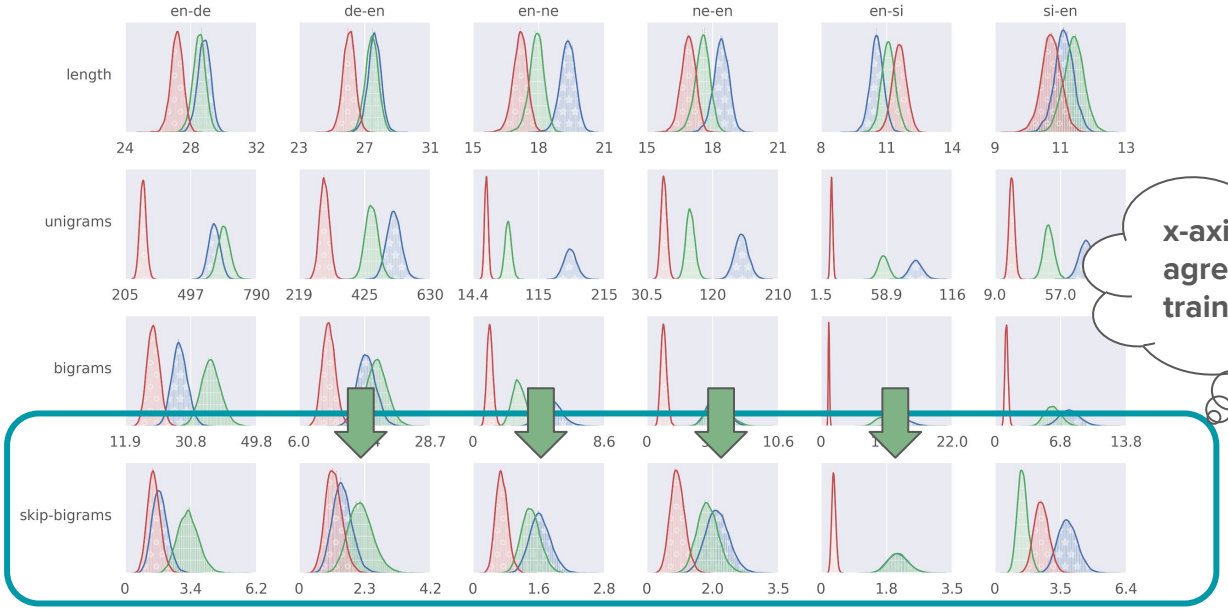



**Beam search shifts from data statistics**, changing lexical characteristics

# Assessing Data Fit: Word Order

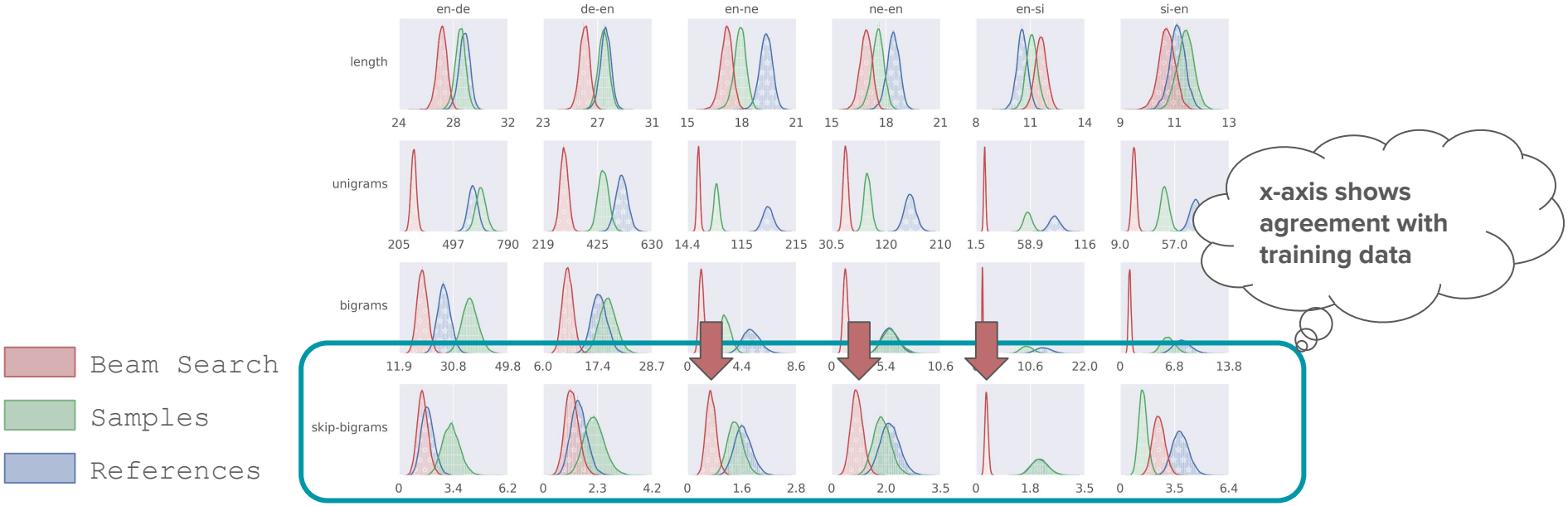


# Assessing Data Fit: Word Order




In most cases the **model captures word order statistics reasonably well**

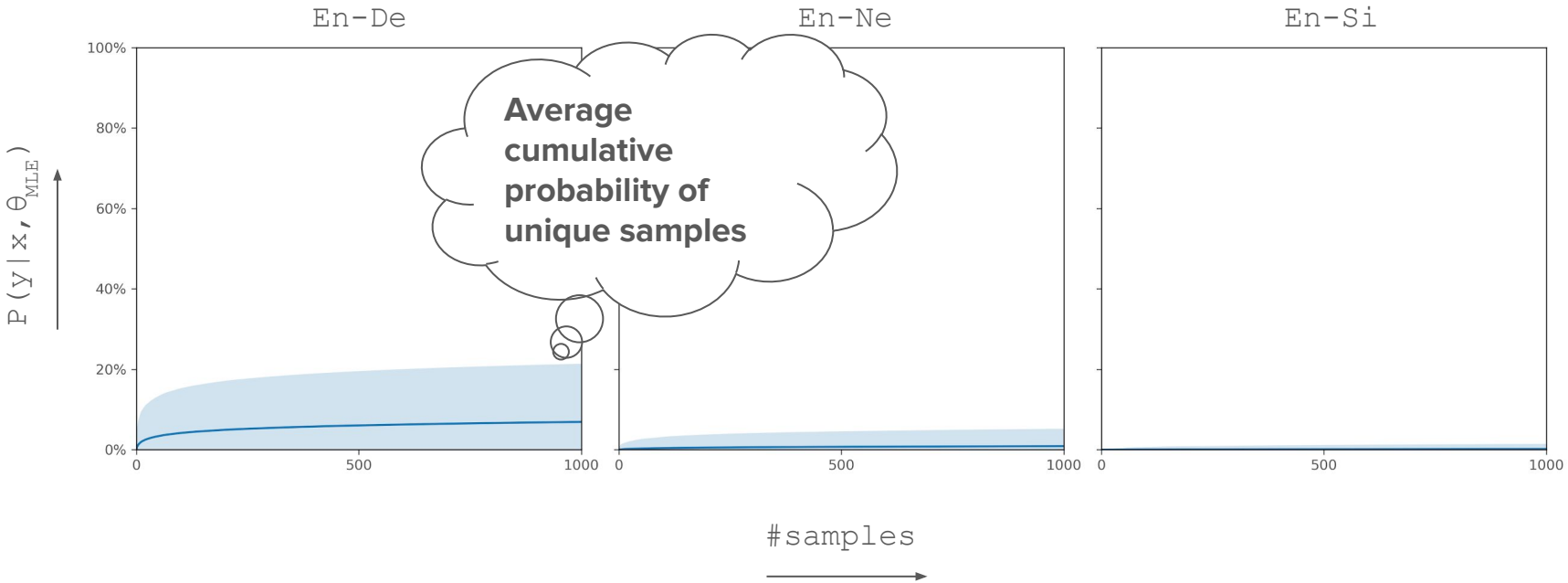
# Assessing Data Fit: Word Order




**Beam search shifts from data statistics**, affecting word order

# Properties of Translation Distributions

# Spread of the Translation Distribution




NMT **spreads mass** over many translations

# Sampling the Mode

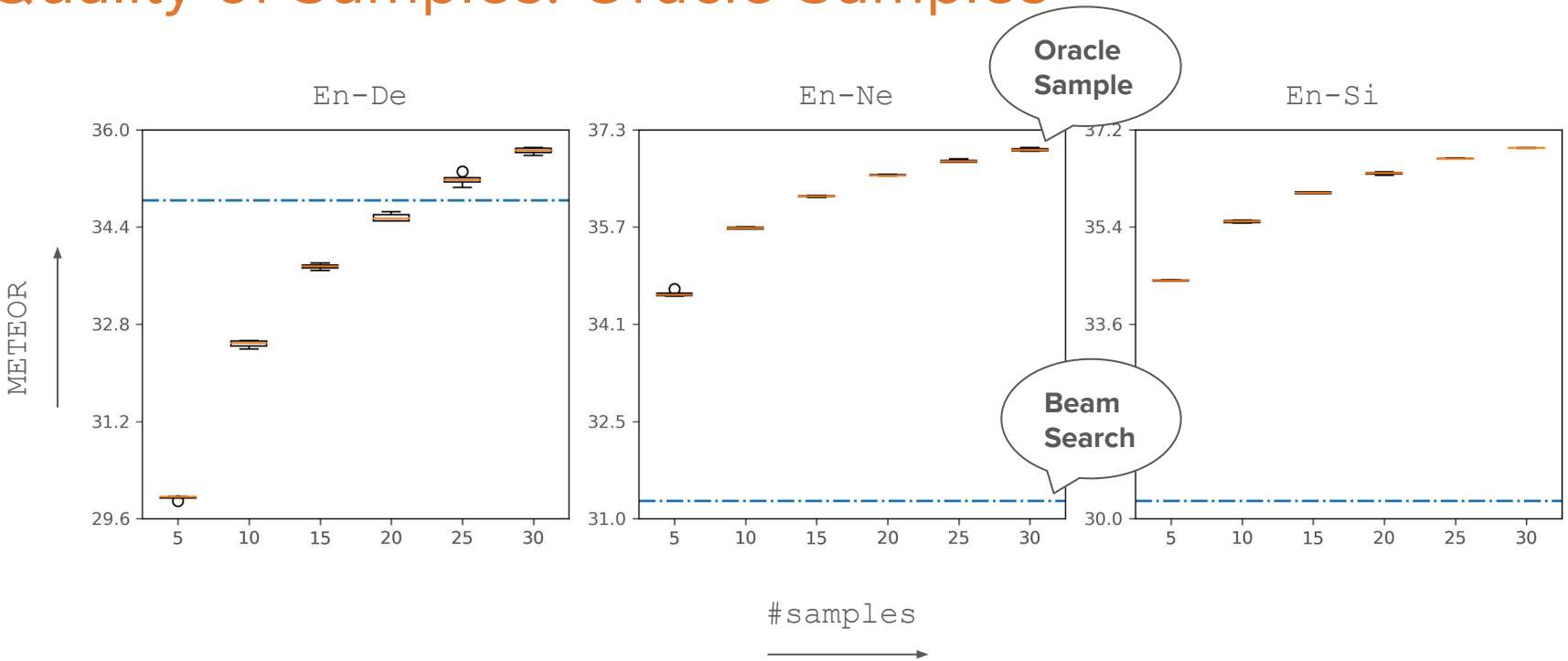
Beam search:

> For most input sequences, the beam search output was **not drawn after 1,000 samples** (>50% high-resource, >90% low-resource)

Empty Sequence:

> In fewer than 35% of input sequences the empty string is drawn, but if drawn it **only occurs roughly once in 1,000 samples**

# Quality of Samples: Oracle Samples




A **small number of samples** contains **good translations**

# A Sampling-Based Decoding Method

# Minimum Bayes Risk (MBR) Decoding

- Maximize expected utility, e.g. METEOR
- But we don’t have the reference
- Use the translation distribution to **fill in the reference**
- Use **unbiased samples** to approximate the objective

Properties:

- Makes use of the **translation distribution as a whole**
- Approximation gets better with more samples
- Doesn’t suffer from the aforementioned pathologies and biases, only poor fit

# Minimum Bayes Risk

Using 30 samples:

| | Beam Search | MBR Decoding | Oracle Decoding |
|---|---|---|---|
| High-Resource | **37.1** | 34.4 | 38.3 |
| Low-Resource | 24.3 | **26.0** | 28.9 |
| All | 28.6 | **28.8** | 32.0 |

# Minimum Bayes Risk

Using 30 samples:

| | Beam Search | MBR Decoding | Oracle Decoding |
|---|---|---|---|
| High-Resource | **37.1** | 34.4 | 38.3 |
| Low-Resource | 24.3 | **26.0** | 28.9 |
| All | 28.6 | **28.8** | 32.0 |

Beam search outperforms MBR in high-resource setting

# Minimum Bayes Risk

Using 30 samples:

| | Beam Search | MBR Decoding | Oracle Decoding |
|---|---|---|---|
| High-Resource | **37.1** | 34.4 | 38.3 |
| Low-Resource | 24.3 | **26.0** | 28.9 |
| All | 28.6 | **28.8** | 32.0 |

MBR decoding outperforms beam search in low-resource settings

# Minimum Bayes Risk

Using 30 samples:

| | Beam Search | MBR Decoding | Oracle Decoding |
|---|---|---|---|
| High-Resource | **37.1** | 34.4 | 38.3 |
| Low-Resource | 24.3 | **26.0** | 28.9 |
| All | 28.6 | **28.8** | 32.0 |

The gap with oracle decoding shows there is a lot of room for improvement

# Conclusion

We should **not be doing MAP decoding** in NMT

MAP decoding **introduces biases** to NMT

Translation distributions **do capture data statistics well**

**Sampling-based decision rules** show great potential