Is MAP Decoding All You Need?

The Inadequacy of the Mode in Neural Machine Translation

Bryan Eikema Wilker Aziz

University of Amsterdam

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 a **probabilistic model**:

- Learn **conditional probability distributions**
- Distributions over **all possible sequences**
- Factorisation into **locally normalised** Categorical distributions
- Estimate parameters using **maximum likelihood estimation** (MLE)

Making Predictions in NMT

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

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

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
- Non-admissible heuristic search bias
- Exposure 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

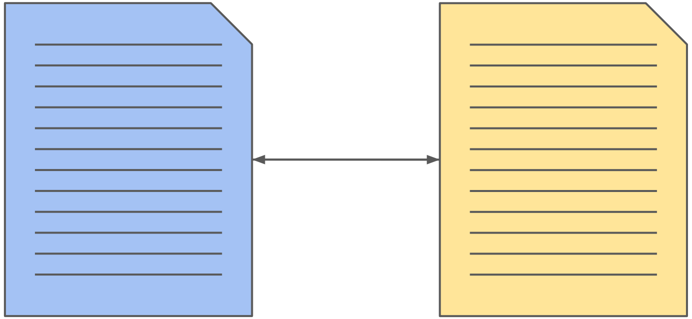
Experiments

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

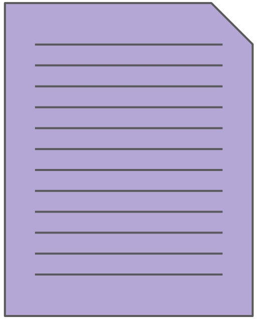
Train on:



English

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

Test on:



newstest2018
Flores
Flores

Model:

Assessing Data Fit

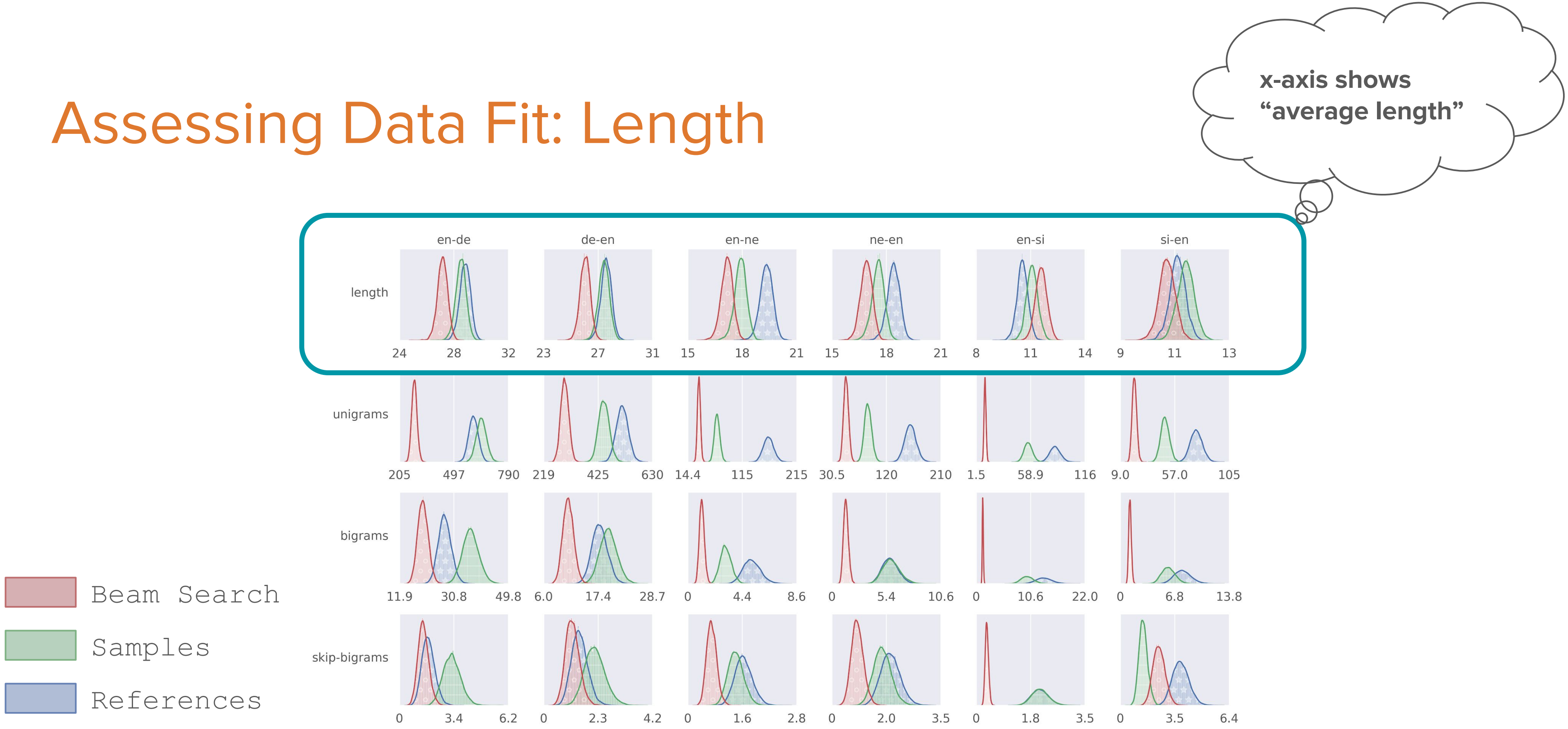
Assessing Data Fit: Methodology

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

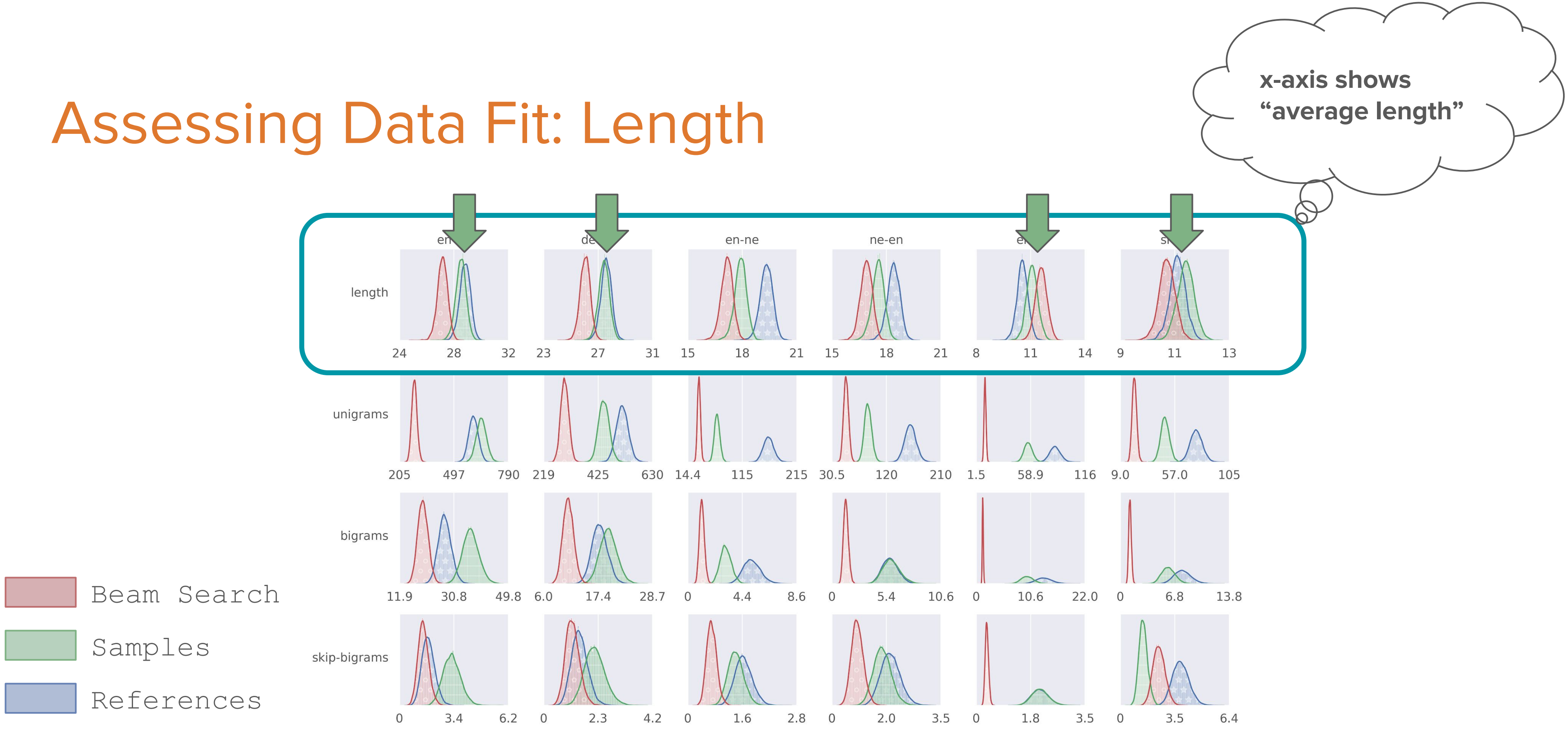
We compare:

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

Assessing Data Fit: Length



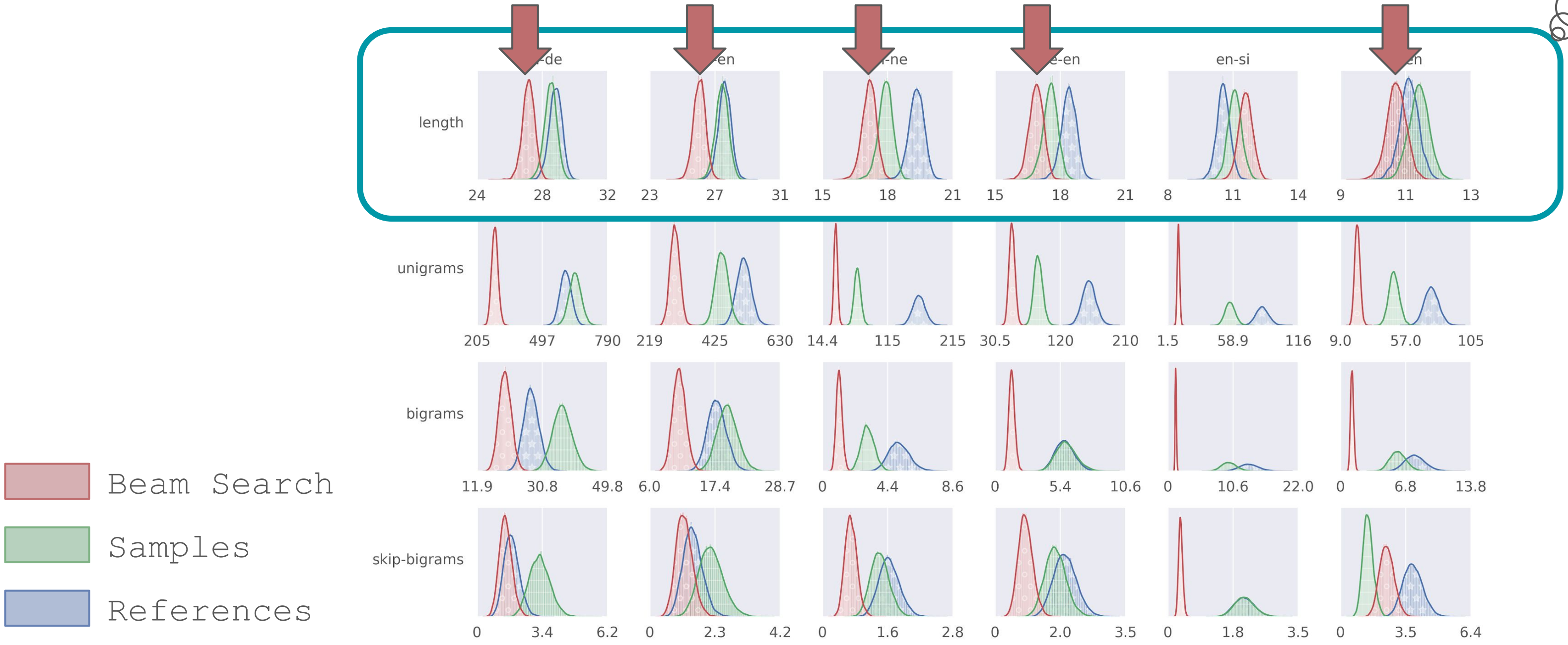
Assessing Data Fit: Length



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

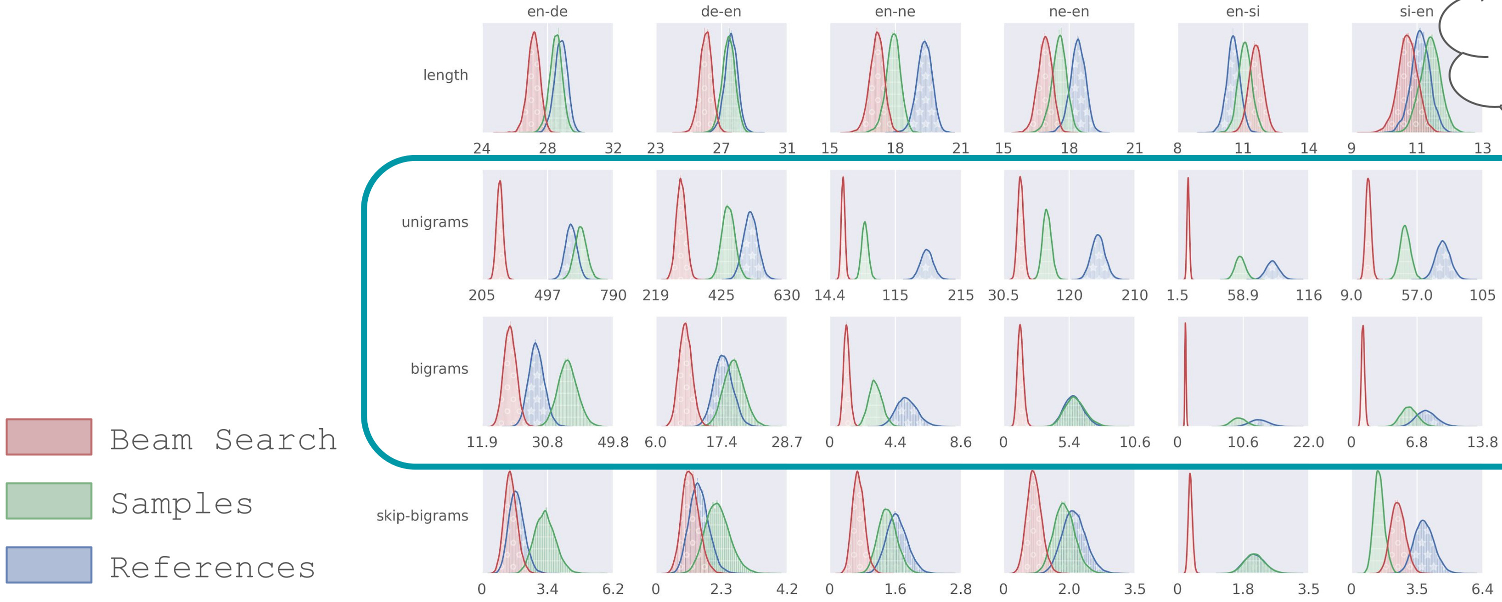
Assessing Data Fit: Length

x-axis shows "average length"



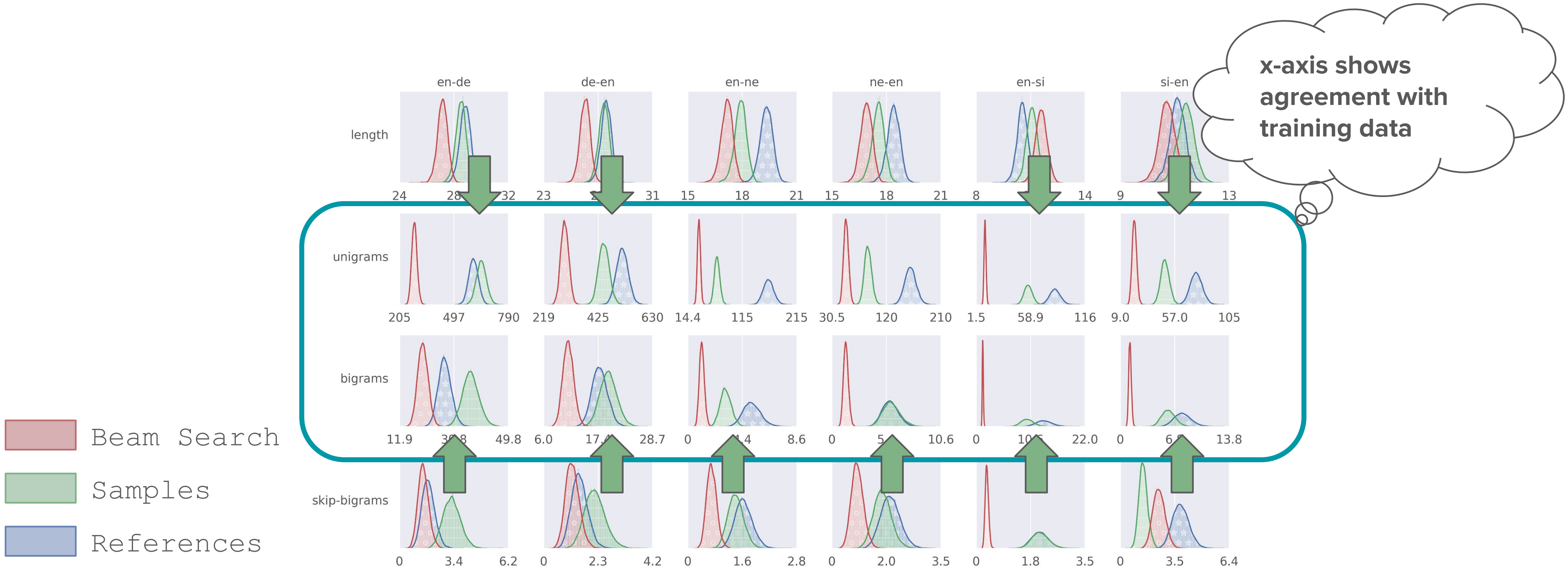
Beam search shifts from data statistics, underestimating length

Assessing Data Fit: Lexical Statistics



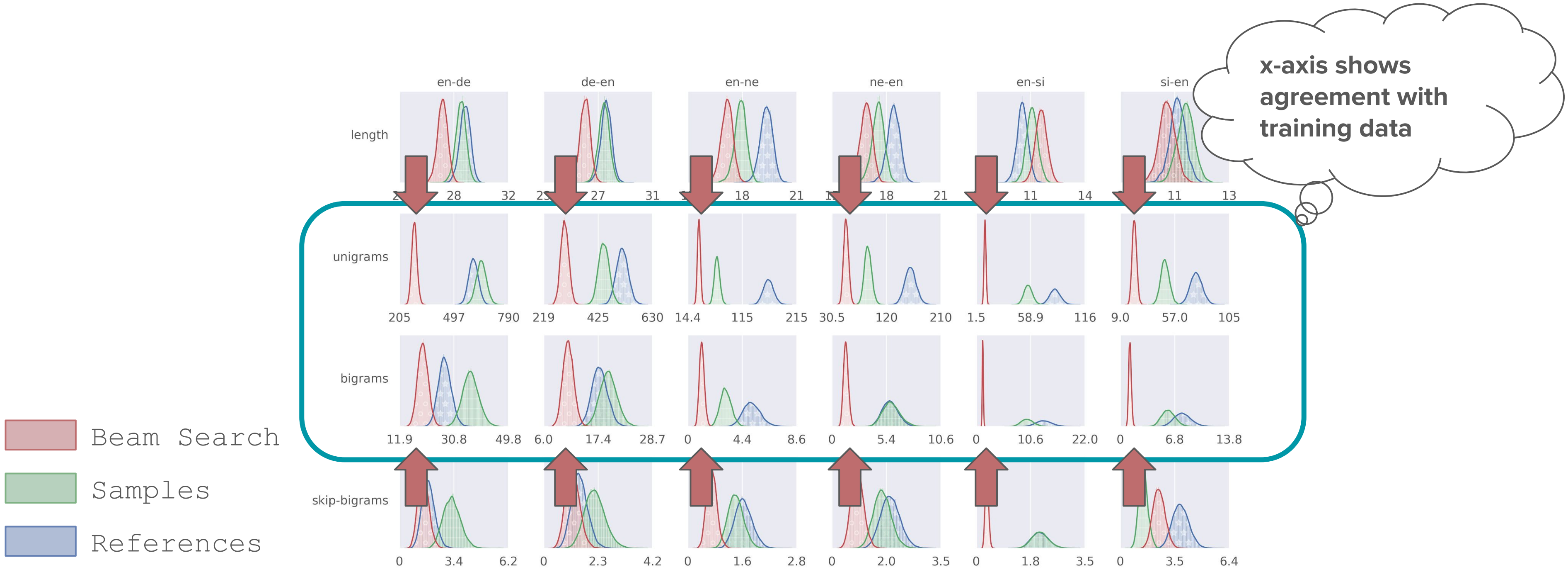
x-axis shows agreement with training data

Assessing Data Fit: Lexical Statistics



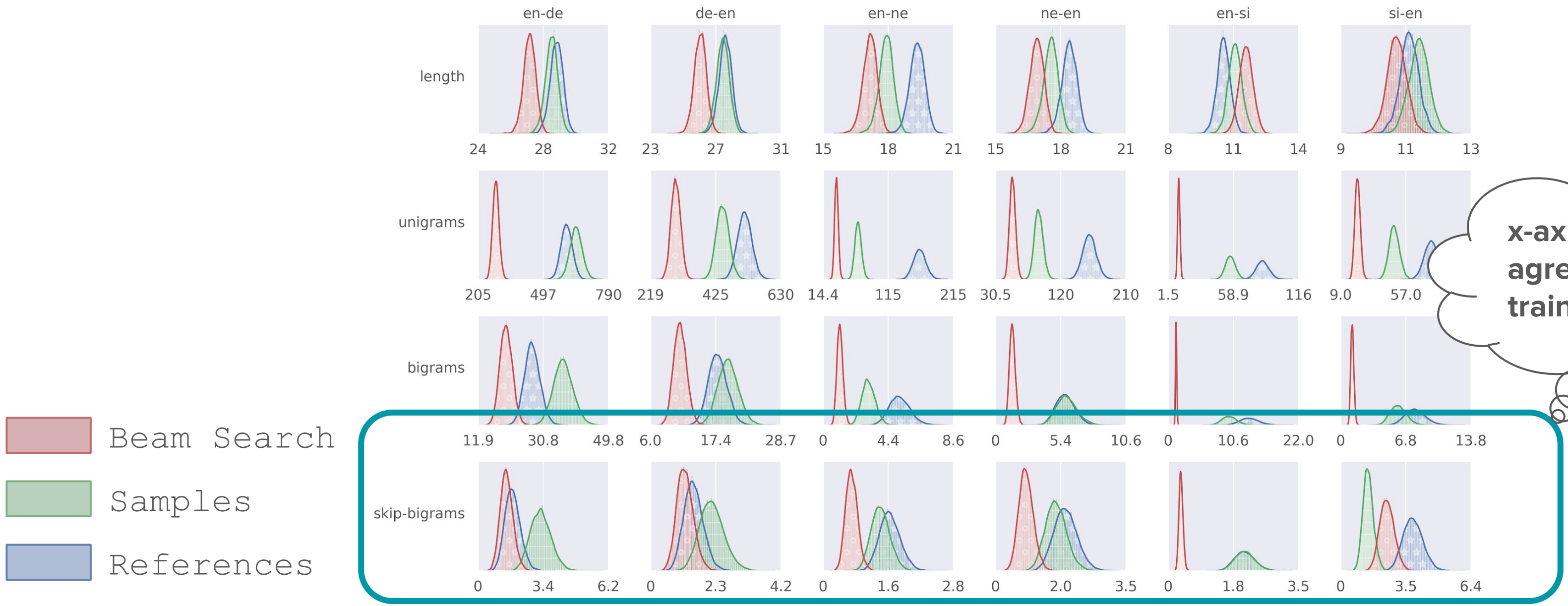
In most cases the **model captures lexical statistics reasonably well**

Assessing Data Fit: Lexical Statistics



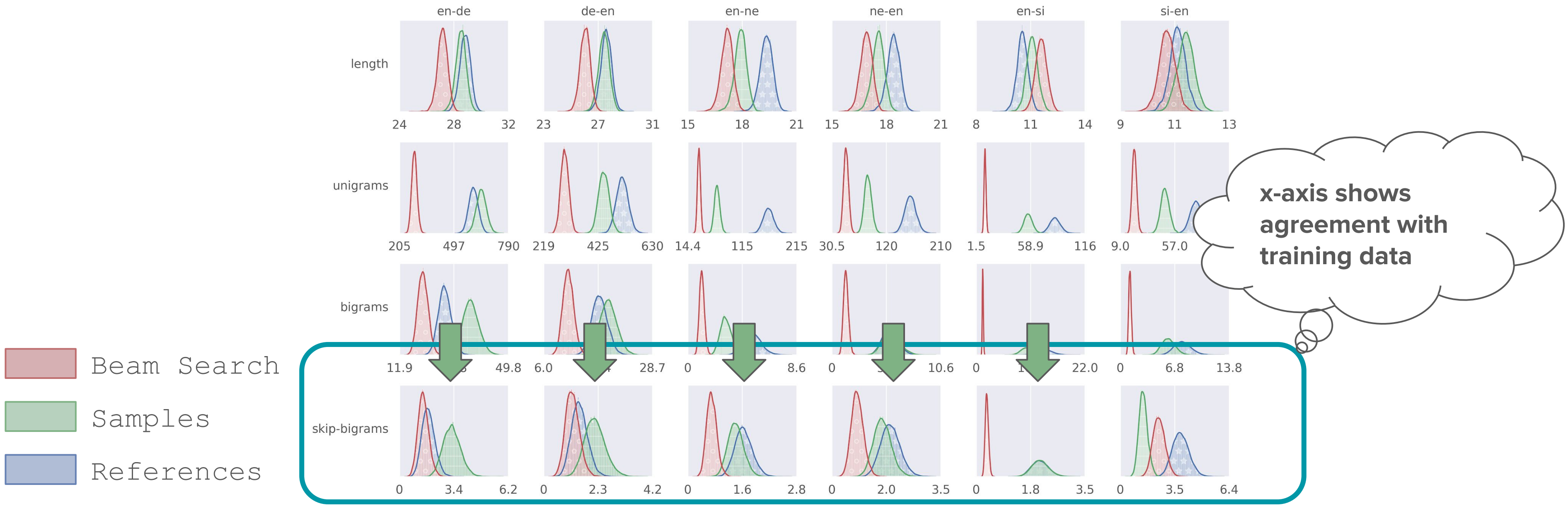
Beam search shifts from data statistics, changing lexical characteristics

Assessing Data Fit: Word Order



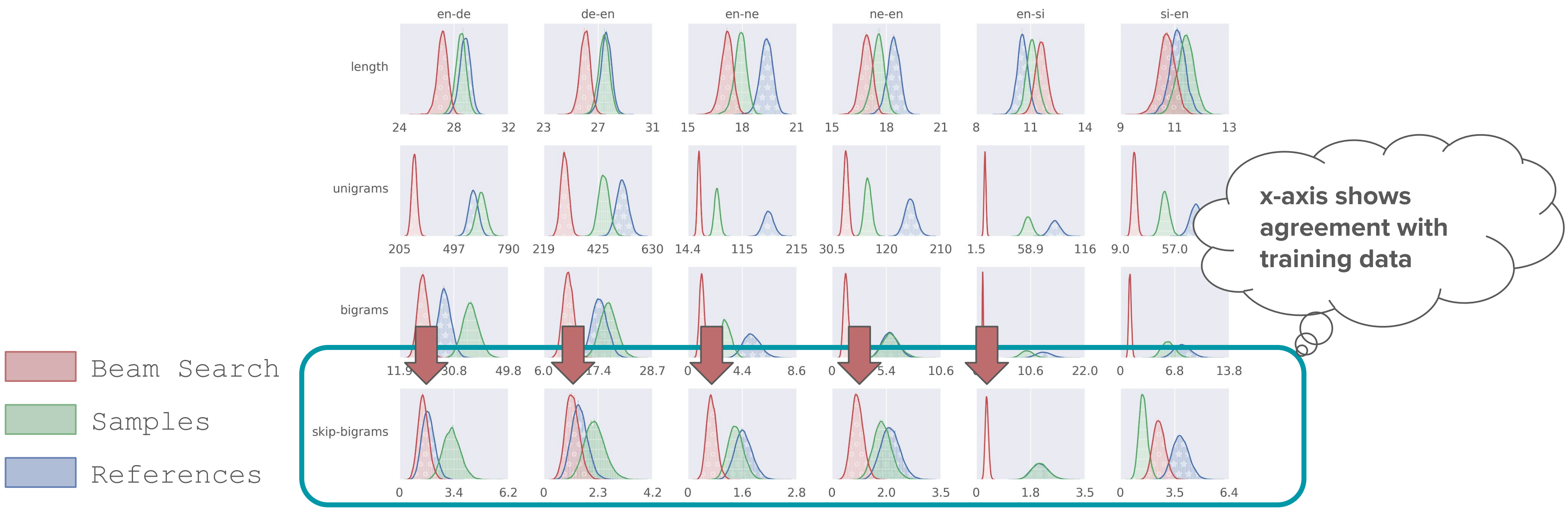
x-axis shows agreement with training data

Assessing Data Fit: Word Order



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

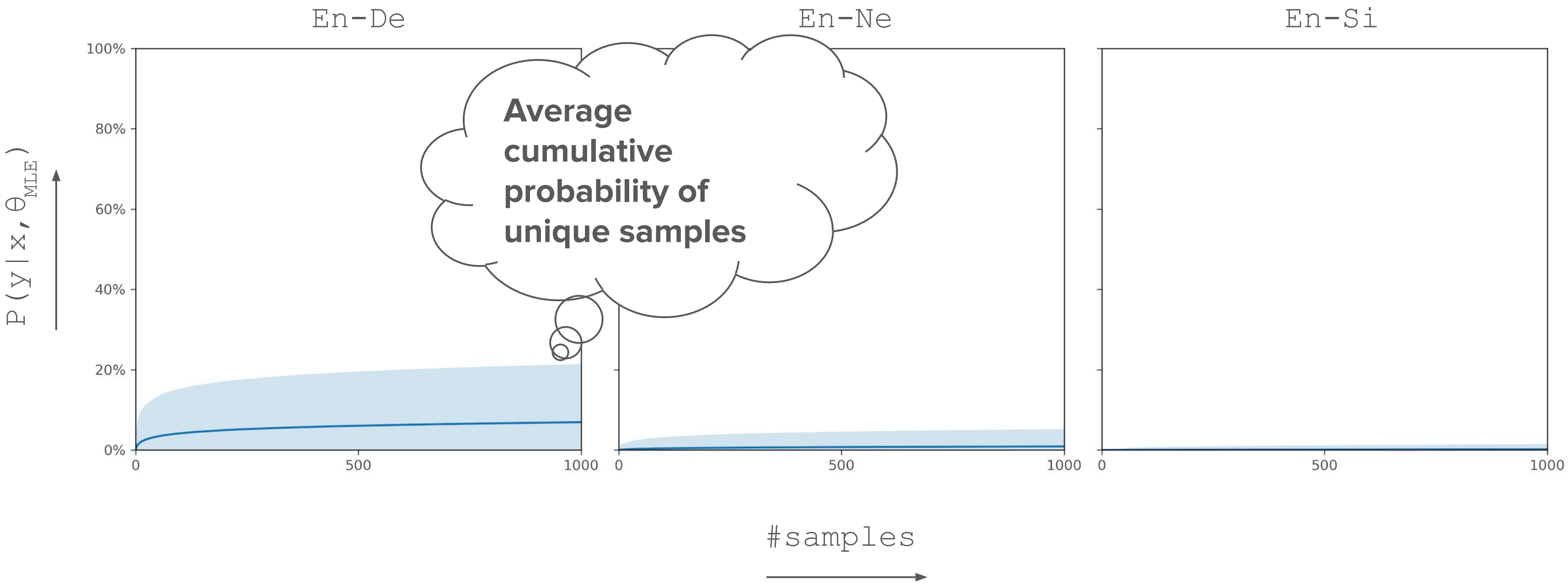
Assessing Data Fit: Word Order



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

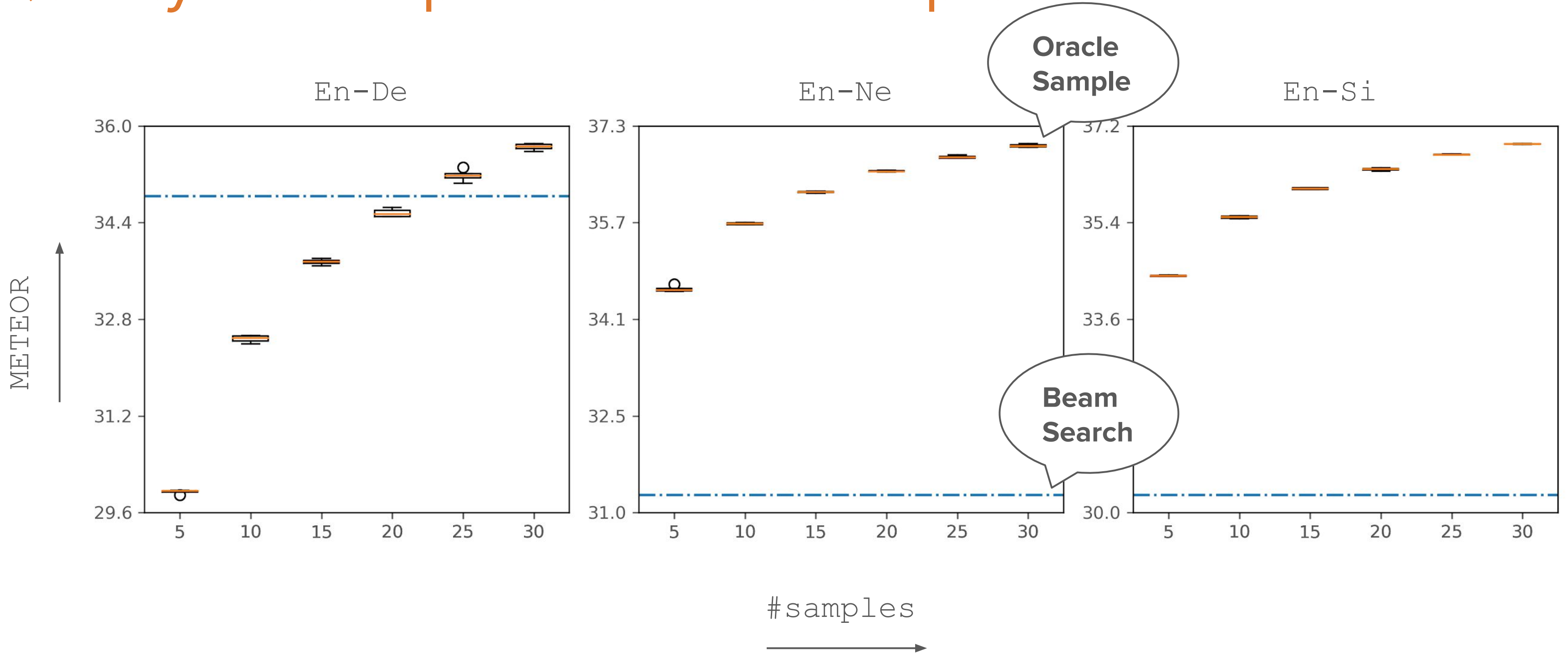
Beam search:

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

$$y^{MBR} = \text{argmax}_h \; E_{P(y|x,\theta)}[U(h, y)]$$

Minimum Bayes Risk (MBR) Decoding

$$\mathtt{y}^{\mathtt{MBR}} = \mathtt{argmax}_{\mathtt{h}}\ \mathtt{E}_{\mathtt{P(y|x,\theta)}}[\mathbf{U}(\mathtt{h},\ \mathtt{y})]$$

- Find hypothesis $\mathtt{h}$ that maximises utility $\mathbf{U}$, e.g. METEOR

Minimum Bayes Risk (MBR) Decoding

$$\mathrm{y}^{\mathrm{MBR}} = \mathrm{argmax}_{\mathrm{h}}\ \mathrm{E}_{\mathrm{P}(\mathrm{y}|\mathrm{x},\theta)}[\mathrm{U}(\mathrm{h},\ \mathrm{y})]$$

- Find hypothesis `h` that maximises utility `U`, e.g. METEOR
- But we don’t have access to the reference

Minimum Bayes Risk (MBR) Decoding

$$y^{MBR} = \mathrm{argmax}_h \; \mathbf{E}_{\mathbf{P(y|x,\theta)}}[U(h, \mathbf{y})]$$

- Find hypothesis h that maximises utility U, e.g. METEOR
- But we don’t have access to the reference
- Use the translation distribution to **fill in the reference** using $\mathbf{P(y|x,\theta)}$

Minimum Bayes Risk (MBR) Decoding

$$y^{\mathrm{MBR}} = \mathbf{argmax}_{\mathbf{h}}\ \mathrm{E}_{P(y|x,\theta)}[U(\mathbf{h},\ y)]$$

- Find hypothesis h that maximises utility U, e.g. METEOR
- But we don’t have access to the reference
- Use the translation distribution to fill in the reference using $P(y|x,\theta)$
- Pick the hypothesis **h** with **highest expected utility**

Minimum Bayes Risk (MBR) Decoding

$$y^{MBR} = \text{argmax}_h \; E_{P(y|x,\theta)}[U(h, y)]$$

- Find hypothesis h that maximises utility U, e.g. METEOR
- But we don’t have access to the reference
- Use the translation distribution to fill in the reference using $P(y|x,\theta)$
- Pick the hypothesis h with highest expected utility

Properties:

- Makes use of the translation distribution as a whole
- We can approximate it using unbiased samples
- Doesn’t suffer from many of the aforementioned pathologies and biases

Approximate MBR with Unbiased Samples

Given trained model $P(y|x, \theta_{MLE})$, input x, utility U, sample size S

$$y^{MBR} = \mathrm{argmax}_{h \in H} \ 1/S \ \sum_s U(h, y^{(s)})$$

Approximate MBR with Unbiased Samples

Given trained model $P(y|x, \theta_{MLE})$, input x, utility U, sample size S

$$y^{MBR} = \mathrm{argmax}_{h \in H}\ 1/S \sum_s U(h, \mathbf{y}^{(s)})$$

1. Sample S unbiased samples: $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(S)} \sim P(y|x, \theta_{MLE})$

Approximate MBR with Unbiased Samples

Given trained model $\mathtt{P(y|x,\ \theta_{MLE})}$, input $\mathtt{x}$, utility $\mathtt{U}$, sample size $\mathtt{S}$

$$y^{MBR} = \mathrm{argmax}_{h \in \mathbf{H}}\ 1/S\ \sum_s U(h,\ y^{(s)})$$

1. Sample $\mathtt{S}$ unbiased samples: $y^{(1)},\ldots,y^{(S)} \sim P(y|x,\ \theta_{MLE})$
2. Use samples as hypotheses as well: $\mathbf{H} = \mathrm{unique}(y^{(1)},\ldots,y^{(S)})$

Approximate MBR with Unbiased Samples

Given trained model $P(y|x, \theta_{MLE})$, input x, utility U, sample size S

$$y^{MBR} = \text{argmax}_{h \in H} \ 1/S \ \sum_s \mathbf{U(h, y^{(s)})}$$

1. Sample S unbiased samples: $y^{(1)},\ldots,y^{(S)} \sim P(y|x, \theta_{MLE})$
2. Use samples as hypotheses as well: $H = \text{unique}(y^{(1)},\ldots,y^{(S)})$
3. **Compute a matrix of utilities** between all pairs of hypotheses and samples

Approximate MBR with Unbiased Samples

Given trained model $P(y|x, \theta_{MLE})$, input x, utility U, sample size S

$$y^{MBR} = \text{argmax}_{h \in H} \mathbf{1/S} \, \boldsymbol{\Sigma_s} \, U(h, y^{(s)})$$

1. Sample S unbiased samples: $y^{(1)}, \ldots, y^{(S)} \sim P(y|x, \theta_{MLE})$
2. Use samples as hypotheses as well: $H = \text{unique}(y^{(1)}, \ldots, y^{(S)})$
3. Compute a matrix of utilities between all pairs of hypotheses and samples
4. Compute the **sample average** utility for each hypothesis

Approximate MBR with Unbiased Samples

Given trained model $P(y|x,\ \theta_{MLE})$, input x, utility U, sample size S

$$y^{MBR} = \mathbf{argmax}_{h \in H}\ 1/S\ \sum_s U(h,\ y^{(s)})$$

1. Sample S unbiased samples: $y^{(1)},\ldots,y^{(S)} \sim P(y|x,\ \theta_{MLE})$
2. Use samples as hypotheses as well: $H = \text{unique}(y^{(1)},\ldots,y^{(S)})$
3. Compute a matrix of utilities between all pairs of hypotheses and samples
4. Compute the sample average utility for each hypothesis
5. Pick the hypothesis with **highest average utility**

Pathologies and Biases for MBR Decoding

MBR doesn’t suffer from many of the aforementioned pathologies and biases:

- Length bias: model fit
- Beam search curse: estimates improve with more samples
- Inadequacy of the mode: not mode-seeking
- Exposure bias: model fit
- Non-admissible heuristic search bias: no search

Results of MBR Decoding

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

Results of MBR Decoding

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

Results of MBR Decoding

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

Results of MBR Decoding

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

The Way Forward

Better sampling-based decision rules:

- More efficient approximations to MBR
- Utilities that better reflect our ideas of quality
- Other sampling-based decision rules

Change the model:

- Sparsifying output distributions
- Learning a decision boundary during training

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