Decoding is deciding under uncertainty

The case of neural machine translation

Wilker Aziz

University of Amsterdam

https://Probabll.github.io

Gourmet

A bit of context

In this talk I assume a trained NMT model is available.

By model, I mean a mechanism to map from a source sentence to a space of translation candidates, with each candidate being assigned a probability mass.

Decoding is then the process of electing one candidate in this space as our preferred translation.

Decoding in MT

We enumerate the translation candidates that are assigned highest probability

then, regardless of

- probability values
- rest of distribution
- properties of the outcomes

  (other than length)

we decide to output the **mode**.

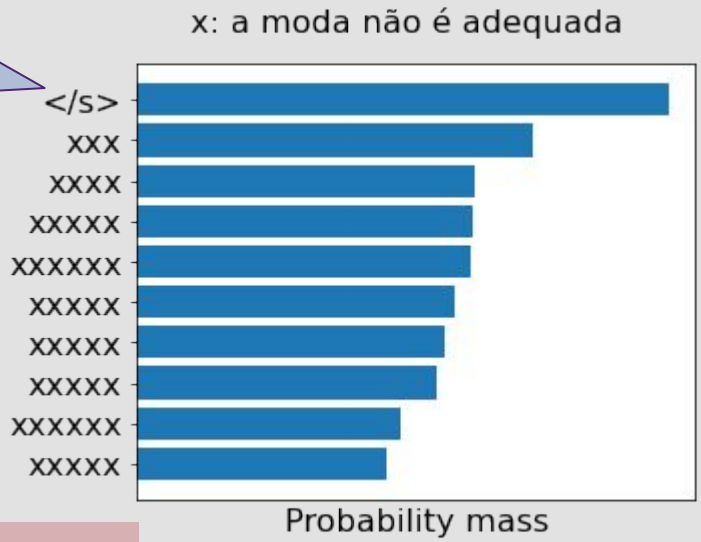
Decoding in MT - what could go wrong?

Sometimes the mode is obviously **inadequate**.

Possible conclusion:

- *Bad model! Else why would the empty string be preferred over all other possible translations?*
  - *Maybe we need to scale this up XD*

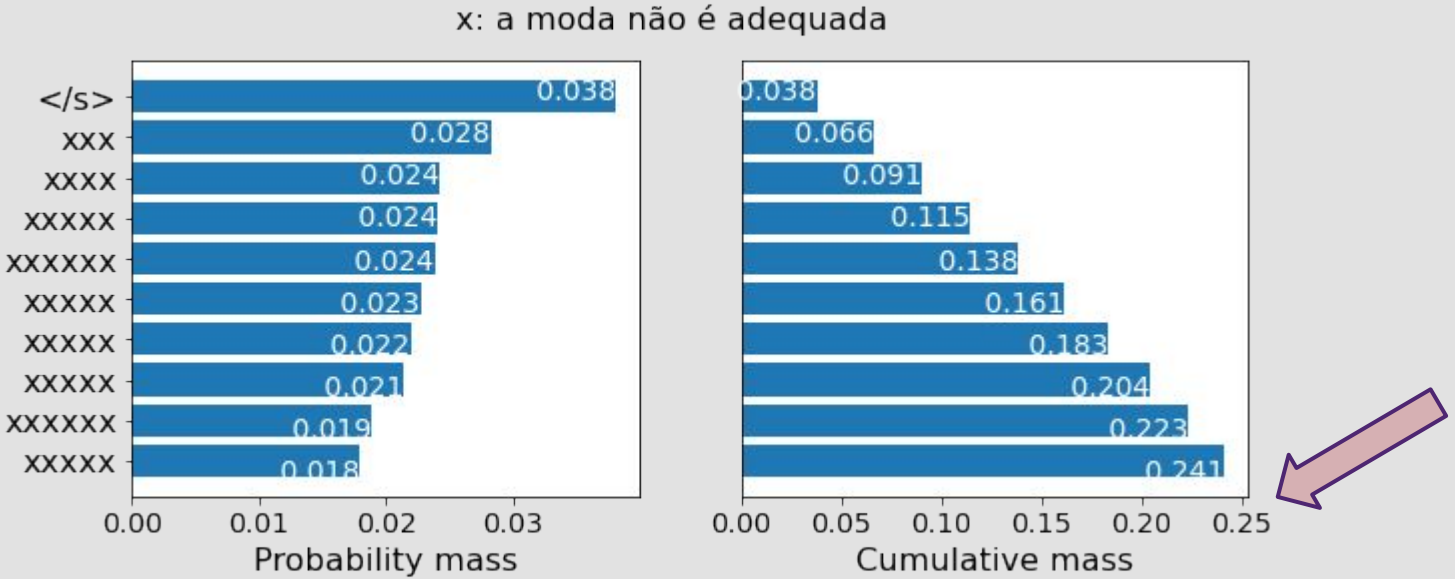
Game over hun?



But hold on, is the empty string “preferred”?

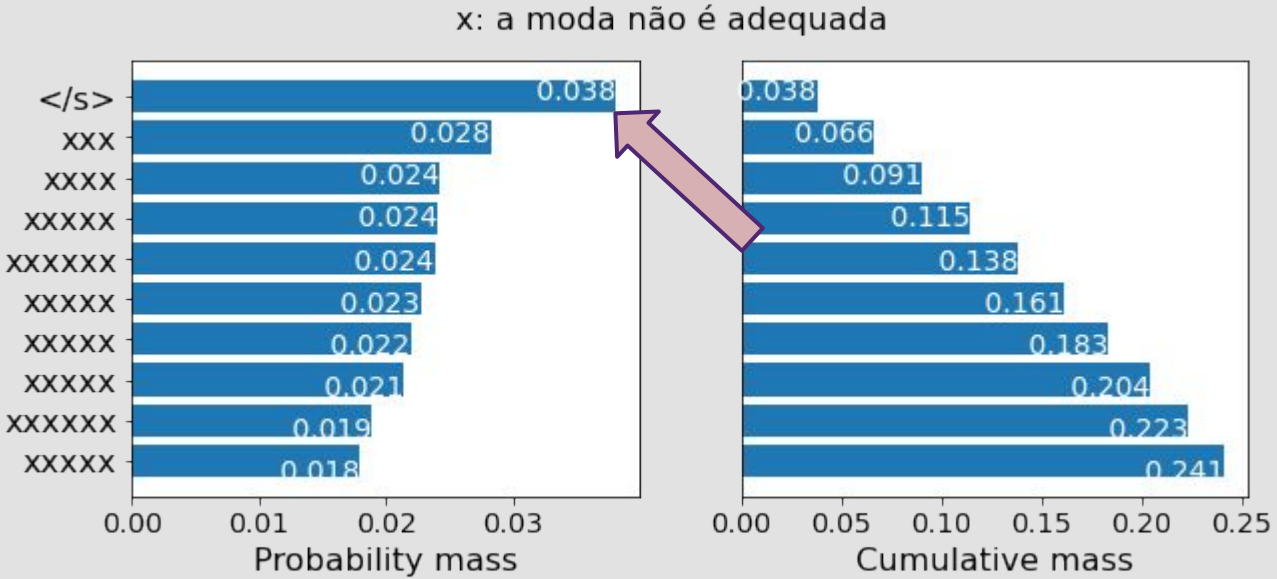
Let’s see what the decision maker chose to ignore

First, we barely covered enough ground. The top 10 translations cover only about 25% of the probability space.



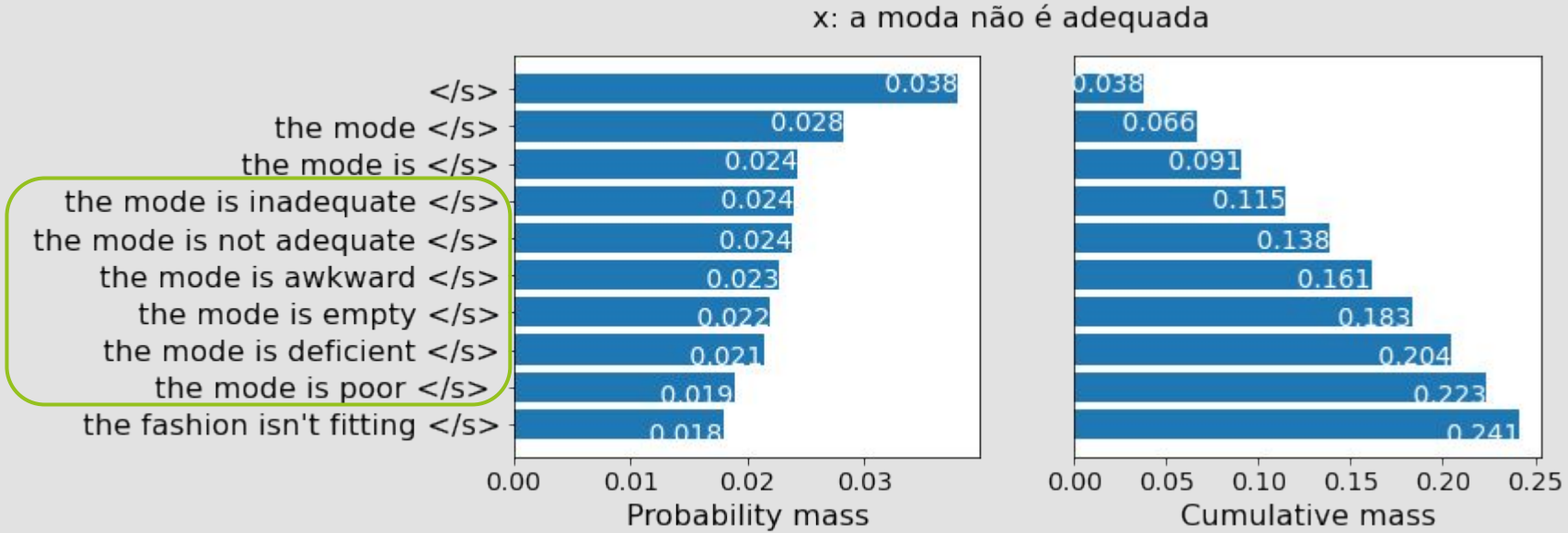
Let’s see what the decision maker chose to ignore

Second, the mode gets less than 4% of the mass. The evidence against the empty sequence is overwhelming.



Let’s see what the decision maker chose to ignore

Third, most sentences are structurally and semantically similar to one another, many are fairly adequate translations of the source.



Deciding under uncertainty

We tend to think of NMT models as predicting the correct translation of $\mathbf{x}$, but, as far as the model is concerned, there is no such a thing as a single correct translation.

NMT packs its beliefs in an entire distribution over candidates. To pick a translation, we (not the model) decide to place all of ours bets on a single outcome (e.g., the mode).

- To decide under uncertainty, we need a criterion (i.e., a decision rule).
- An NMT model is not a decision rule, it cannot tell you how to decide.
- But we can use the uncertainty NMT quantifies to make an informed decision.

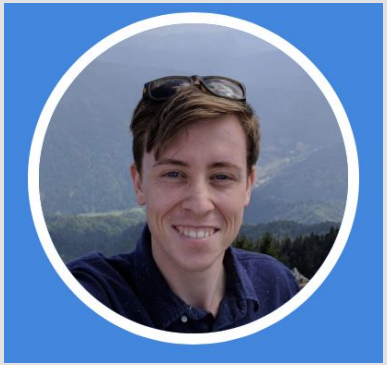
Outline

1. NMT as Markov processes
2. Statistical model criticism
3. Deciding under uncertainty
4. An origin story
5. What next?

Credits

Many of the points I will discuss today were developed in collaboration with my brilliant PhD student, Bryan Eikema.

Some of the results here were presented in Is MAP decoding all you need? The inadequacy of the mode in NMT.

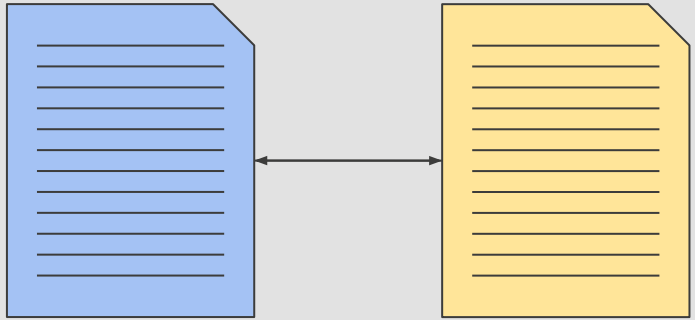


https://roxot.github.io

You can find Bryan's excellent talk on underline.

Experiments

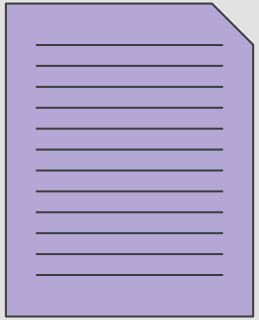
Train on:



English

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

Test on:



newstest2018
Flores
Flores

Model:

NMT as Markov processes

NMT prescribes a Markov process

Given a source sentence x, an NN $f(.;\theta)$ parameterises a Markov process:

1. We start from the state $s_0 = (x, \langle s \rangle)$ and predict a Categorical distribution for the first target word Y_1
2. With probability $f(s_0, y_1; \theta)$ we draw (or observe) an outcome $Y_1 = y_1$ and move to a new state $s_1 = (x, \langle s \rangle y_1)$
3. From s_1 we predict a Categorical distribution for the second target word Y_2
4. With probability $f(s_1, y_2; \theta)$ we draw/observe $Y_2 = y_2$ and move to state $s_2 = (x, \langle s \rangle y_1 y_2)$.
5. This goes on until we draw/observe a terminating symbol ($\langle /s \rangle$), which prompts us to reset the process to its start state s_0.

Whatever sequence of symbols is emitted in a trajectory between two occurrences of s_0 is treated as a possible translation of x.

Remarks

The state resets whenever we draw `</s>`

- Trajectories are independent of one another and the invariant measure is exactly the conditional distribution with pmf $p_\theta(y|x) = \prod_j f(s_{j-1}, y_j; \theta)$

We can draw i.i.d. samples from the model (e.g., via ancestral sampling)

- good for estimating expectations

and we can assess the probability of any given outcome

- good for parameter estimation

Asymptotic Equipartition Property (AEP)

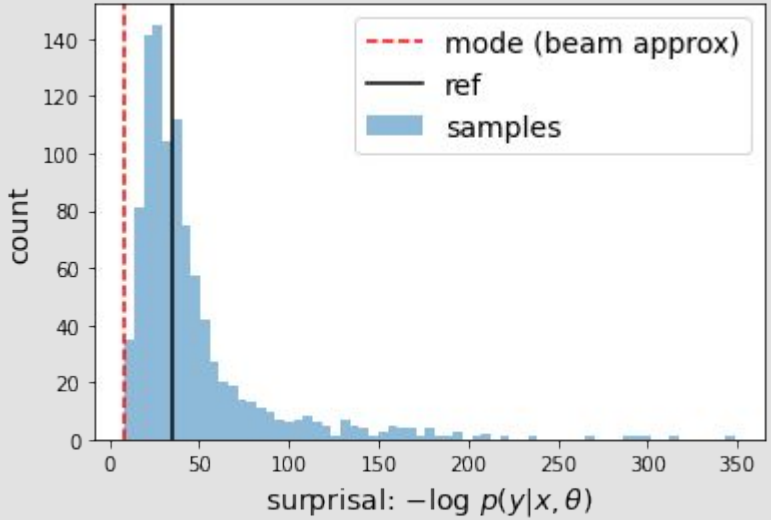
AEP *(Certain)* Markov processes exhibit a ‘concentration of surprisal’ property

- the surprisal of a sampled trajectory is typically within a margin of the entropy rate of the process;
- this is more likely the case, the longer the trajectory;
- effectively, sampled outcomes live in a ‘small’ set which does not include outcomes that have extreme surprisal values (eg, the mode).

If an AEP would hold for NMT: the longer we expect a reasonable translation to be, the less likely it is that the mode of the distribution is in this *typical set*.

Concentration in surprisal

The reference is not at the mode, the mode is not even in the sample



This observation had been made before, but it was perceived as a defect in the model and/or its training algorithm.

Regularised maximum likelihood estimation

The NN parameters are chosen via (regularised) **MLE**

$$\theta_{MLE} = \text{argmax}_{\theta} \sum_{(x,y)} \log p_{\theta}(y|x) - R(\theta)$$

This objective is in no way connected to the mode of the distribution.

MLE identifies a Markov process from where we could have sampled the observations:

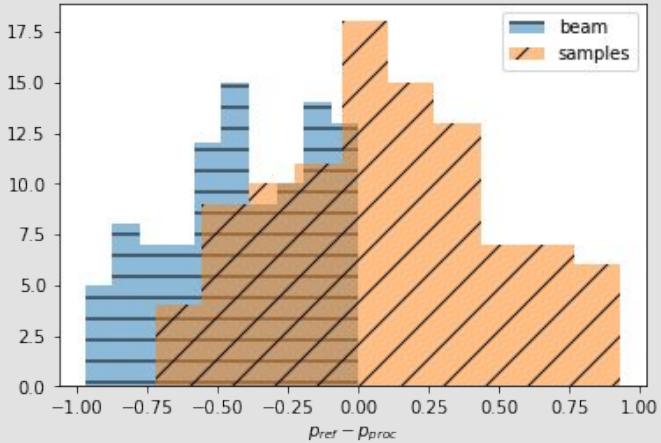
- a dataset of N observations is **a very long trajectory**
- very long trajectories likely live in typical sets ➜ the observations are not going to be modes

Theoretical vs empirical properties

The state can store an arbitrary amount of information

- Standard forms of AEP/typicality cannot be guaranteed

Yet, model samples and data samples are fairly similar in terms of surprisal, while modes (rather, beam search outputs) are not.



Sampling the Mode

We cannot guarantee an AEP, but we can look for signs of it

Draw a 1,000 samples from the model. These samples are a faithful to the statistics packed in the output distribution.

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

- Low surprisal (most probable) outcomes are effectively rare, by focusing on them we are exaggerating statistics that are not faithful to the model distribution.

Summary

- NMT prescribes a Markov process
  - Tractable sampling
  - Tractable pmf
- MLE-trained NMT models exhibit
  - Concentration in surprisal
  - Modes are unlikely to be sampled (≅ atypical)
  - References are better supported by the model
- This view explains various pathologies of mode-seeking decoding

[Ranzato et al, 2016; Sountsov and Sarawagi, 2016; Huang et al ,2017; Koehn and Knowles, 2017; Murray and Chiang, 2018; Kumar and Sarawagi, 2019; Stahlberg and Byrne, 2019]

Statistical model criticism

References are ‘better supported by the model’?

MLE training is trying to recover a process whose samples are faithful to statistics of observed data.

For any statistic `t`, `t(Y)` where `Y` is a model sample should distribute roughly like `t(R)` where `R` is a data sample (an observation).

In statistics this is just an assessment of goodness of fit. In the paper we show that

- samples from the model better reproduce statistics such as length, unigram counts, bigram counts, and skip gram counts;
- samples from the model share structural similarity with observations (as measured by BLEU and METEOR)

Summary

Beam search **shifts** the distribution of statistics such as length, unigram/bigram, and skip-bigram counts away from human references.

Unbiased samples better reproduce those statistics.

The model fits the data better than beam search outputs would have suggested.

Deciding under uncertainty

Decision Rules

We train models in order to support a decision maker (e.g., someone who wants to generate a piece of text about a given input).

A probabilistic model can power an algorithm to make decisions under uncertainty. For example:

- Output the most probable outcome or an approximation thereof (e.g., beam search)

These algorithms are generally called **decision rules**.

MAP decoding

Predicting the mode also goes by the name of maximum-a-posterior (MAP) decoding.

We have found no theoretical reason to support MAP decoding and committed to it following an intuition (one that need not hold for Markov processes).

Let’s take a moment to look for other such axiomatic ways to make decisions, and then get back to MAP decoding with more tools for analysis.

Utility

If we interpret a translation candidate as atomic and unrelated to any other outcome, all NMT does is to express a preference over complete translations. This preference is often very weak.

Interpreted as combinatorial structures, we can appreciate structural similarity.

A utility function quantifies this similarity in a way that matters for a decision maker.

We say that `u(y, h; x)` quantifies the benefit in choosing `h` as the translation of `x` when `y` is known to be a plausible translation of it.

- Examples: METEOR, BEER, ChrF, COMET, BLEURT, human judgement, etc.

Uncertainty about utility

When deciding whether or not `h` is a reasonable translation of `x`, we do not have access to translations we already know to be plausible choices.

But we have NMT models as a representation of what we know about translation (at least as exemplified by a training data set).

Expected utility

If all I know is that y translates x with probability $p(y|x)$, then my expectation on h's utility is the weighted average utility against every valid translation under the model:

$$u(h;x) = p(y^1|x)u(y^2, h;x) + p(y^3|x)u(y^3, h;x) + \ldots$$

$$= \sum_y p(y|x)\ u(y, h;\ x)$$

$$= E[u(Y, h;\ x)]$$

where, in turn and with some probability, each and every translation is assumed to be a reference.

h	y	p(y|x)	u(y, h;x)	p(y|x) * u(y, h;x)
</s>	</s>	0.0067	100.00	0.67
	the mode </s>	0.0051	29.71	0.15
	the mode is </s>	0.0045	24.93	0.11
	the mode is inadequate </s>	0.0038	13.84	0.05
	the mode is not adequate </s>	0.0037	13.25	0.05
	the mode is awkward </s>	0.0036	15.97	0.06
	the mode is empty </s>	0.0035	17.79	0.06
	the mode is deficient </s>	0.0034	14.48	0.05
	the mode is poor </s>	0.0033	18.87	0.06
	the fashion isn't fitting </s>	0.0033	12.21	0.04
	[...]			
	[SUM]			22.98
the mode isn't adequate </s>	</s>	0.0067	37.93	0.25
	the mode </s>	0.0051	58.62	0.30
	the mode is </s>	0.0045	62.16	0.28
	the mode is inadequate </s>	0.0038	77.17	0.30
	the mode is not adequate </s>	0.0037	82.98	0.30
	the mode is awkward </s>	0.0036	45.80	0.17
	the mode is empty </s>	0.0035	49.20	0.17
	the mode is deficient </s>	0.0034	44.47	0.15
	the mode is poor </s>	0.0033	49.81	0.16
	the fashion isn't fitting </s>	0.0033	23.08	0.08
	[...]			
	[SUM]			31.63

Minimum Bayes Risk (MBR) Decoding

Find hypothesis h that maximises utility u, in expectation under the model distribution

$$y^{MBR} = \text{argmax}_h \; \mathbf{E}[u(\mathbf{Y}, h; x)]$$

Properties

- Makes use of the translation distribution as a whole
- Exploits similarity to redistribute beliefs

Goodman (1996), Sima'an (2003), Goel and Byrne (2000), Kumar and Byrne (2002, 2004)

Intractability of MBR decoding

In general, MBR decoding is intractable and there are two sources of intractability

$$y^{MBR} = \text{argmax}_h \; E[u(Y, h)]$$

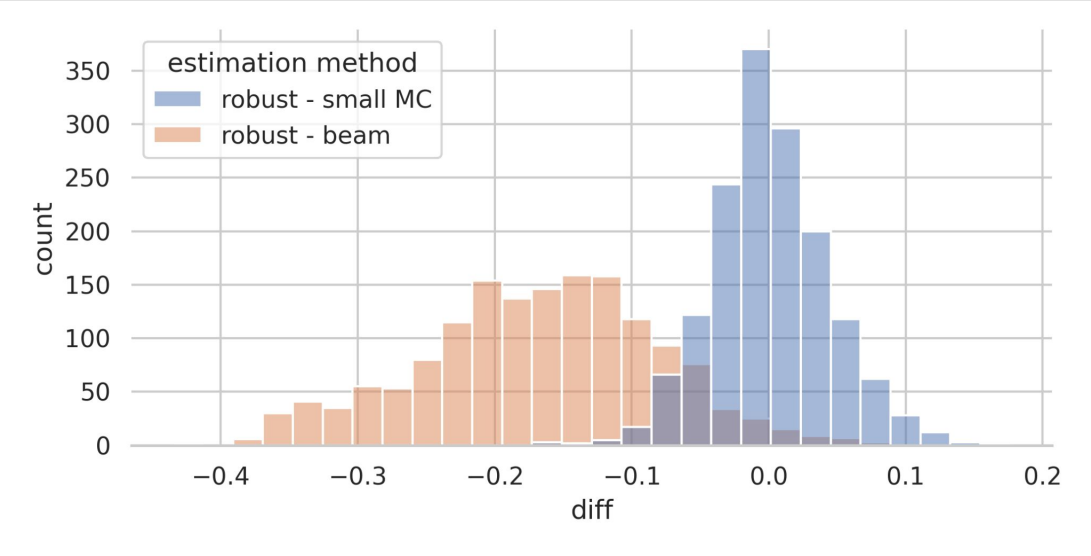
- As in MAP decoding, the hypothesis space is unbounded
  - But we can enumerate a subset
- The objective function (expected utility) requires an intractable sum
  - But we can obtain an **unbiased estimate through Monte Carlo**

$$\mu(h;x) = E[u(Y, h; x)] \cong 1/S \sum_s u(y^s, h; x)$$

Beam-based approximations [Shu and Nakayama 2017, Stahlberg et al, 2017, Blain et al, 2017]

Why MC?

Unbiased estimates of the objective function (expected utility)



Approximate MBR with Unbiased Samples

y ~ Y|x	u(y, "</s>";x)	u(y, "the mode isn't adequate </s>";x)
aren't adequate </s>	17.79	75.46
uncool mode </s>	23.07	29.39
uncool </s>	32.88	18.16
rare rare rare rare </s>	15.97	16.90
NMT does strange things </s>	13.25	15.62
NMT does strange things </s>	13.25	15.62
the is </s>	36.82	31.67
mode is weird </s>	21.48	35.96
fashion isn't a thing </s>	14.48	29.31
sometimes NMT does strange things </s>	9.59	14.09
the mode is empty </s>	17.79	49.20
the mode is not very probable </s>	11.33	41.56
mode is strange </s>	18.87	38.55
unfashionable </s>	18.87	21.98
weird mode </s>	24.93	33.37
mode is a mode </s>	21.48	42.70
the mode is awkward </s>	15.97	45.80
aren't adequate </s>	17.79	75.46
unfashionable </s>	18.87	21.98
well I told you so didn't I ? </s>	12.21	16.69
[AVG]	18.83	33.48

Sample Size and Utility

De-En

Samples	METEOR	BEER	BLEU	METEOR	BEER	BLEU
30	34.4	60.2	26.5	36.1	63.1	31.3
75	37.1	63.1	30.7	37.7	64.8	33.9
105	37.6	63.5	31.1	38.1	65.3	34.5
210	38.4	64.3	32.3	38.4	65.7	35.5
300	38.6	64.4	32.5	38.6	65.9	35.7
Beam search	38.5	64.9	36.4	38.5	64.9	36.4

Qualitative remarks

Scales with computation (unlike beam search).

Sensitive to choice of utility (e.g., BEER-MBR leads to better BLEU/METEOR than BLEU-MBR or METEOR-MBR).

Other observations:

- Less bias towards short translations, robustness to copying noise and hallucination [Müller and Sennrich, 2021].
- Surprisal closer to that of references [Meister et al, 2022].
- Improves substantially with modern neural utilities [Freitag et al, 2022].

An origin story

MAP decoding is MBR

Consider the exact match utility $\mathbf{1}(y, h)$, which returns 1 when y and h are the same and 0 otherwise.

Its expected value under the model is $E[\mathbf{1}(Y, h)] = p(h|x)$.

Thus $\text{argmax } E[1(Y, h)] = \text{argmax } p(h|x)$ which is MAP decoding!

When we decide via MAP decoding, we implicitly decide via MBR using exact match as utility.

This has been known since MBR's introduction – but I guess we forgot about it =O

Great, right?

No, not really!

- From a task point of view. In MT we certainly expect multiple correct translations (e.g., [Dreyer and Marcu 2012], [Khayrallah et al, 2020]).
- From a practical/statistical point of view. We know since at least [Ott et al, 2018] that NMT models are relatively high-entropy (in a large sample, most sequences appear once).

Summary

Model's beliefs are expressed in terms of expectations of quantities of interest

- There is no principled reason to rank candidates in terms of model probability

In MBR, a rational decision maker acts as to maximise expected utility, a criterion that combines a utility and the model distribution:

- this is axiomatic (call it a principle),
- it includes MAP decoding as special case

Whether or not we pick it consciously, decision making requires a utility function.

What next?

A whole bunch of new knobs to turn

Understanding the role of the utility function (Müller and Sennrich, 2021)

Deciding with neural utilities (Freitag et al, 2022)

Quality-aware decoders (Fernandes et al, 2022) and other axioms for decision making (Borgeaud and Emerson, 2020)

Enumerating better candidates (Meister et al, 2022)

Role of intrinsic uncertainty (Forster et al 2021, Stahlberg et al 2022)

Decision-aware training and/or learn to search (Leblond et al 2021, Ling et al 2022)

Uncertainty estimation in structured prediction (Malinin and Gales, 2021)

Key Takeaways

We question the use of MAP decoding in NMT

- MAP decoding **introduces biases that aren't controlled for**
- The mode is a very **rare outcome**
- NMT models **capture data statistics well**

We argue that:

- Models convey beliefs through expectations (not modes)
- **Unbiased samples** power an additional angle for model criticism
- Decision making requires a **utility function**

Thanks!

Additional slides

Addressing or Circumventing the Inadequacy of the Mode

Regularising beam search

- less mode-seeking [Holtzman et al, 2019]
- exploit patterns in surprisal and entropy [Meister et al, 2020]
- mutual information [Holtzman et al, 2021]

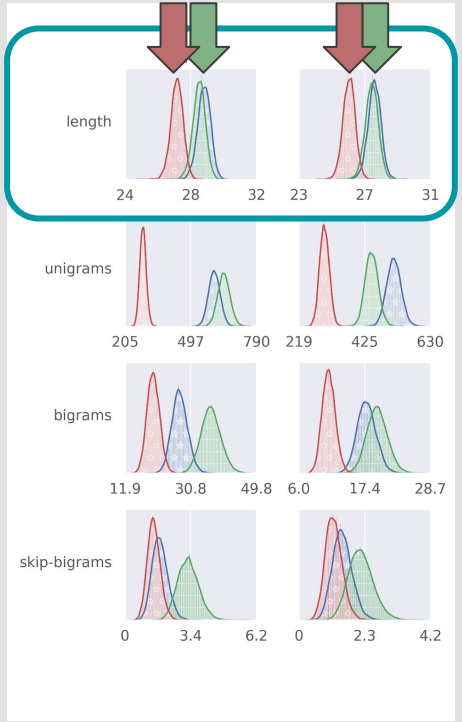
External utility

- Voting [Borgeaud and Emerson, 2020]
- Energy-based re-raking [Naskar et al, 2020]

Change the model or its training:

- Learning a decision boundary during training [Wiseman and Rush, 2016; Shen et al, 2016]
- Sparsifying output distributions [Peters et al, 2019; Peters and Martins, 2021]

Assessing Data Fit: Length



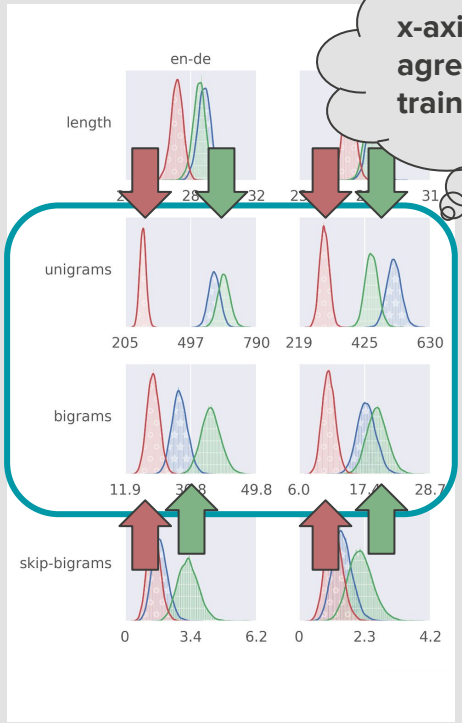
x-axis shows
“average length”

Beam Search

Samples

References

Assessing Data Fit: Length



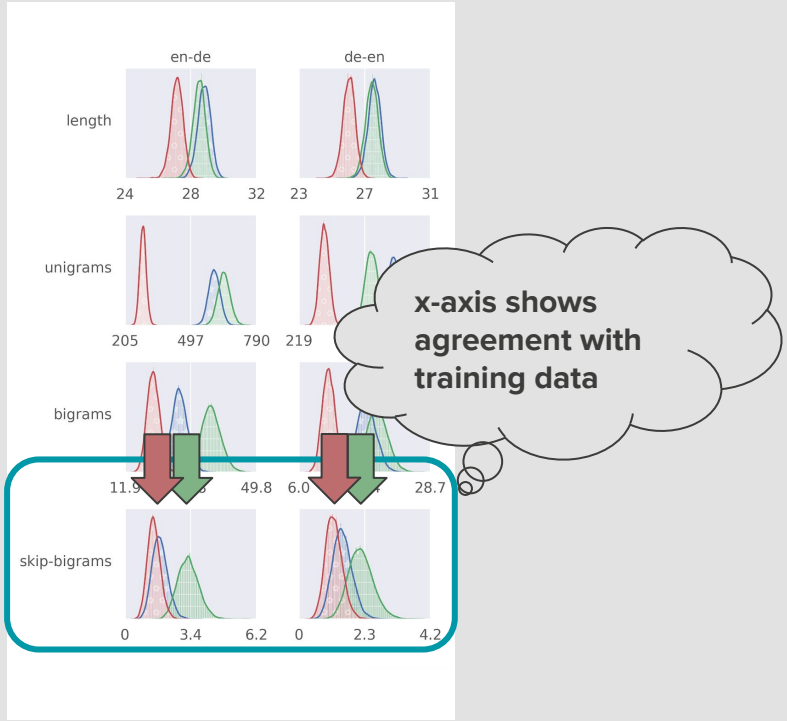
x-axis shows agreement with training data

Beam Search

Samples

References

Assessing Data Fit: Length

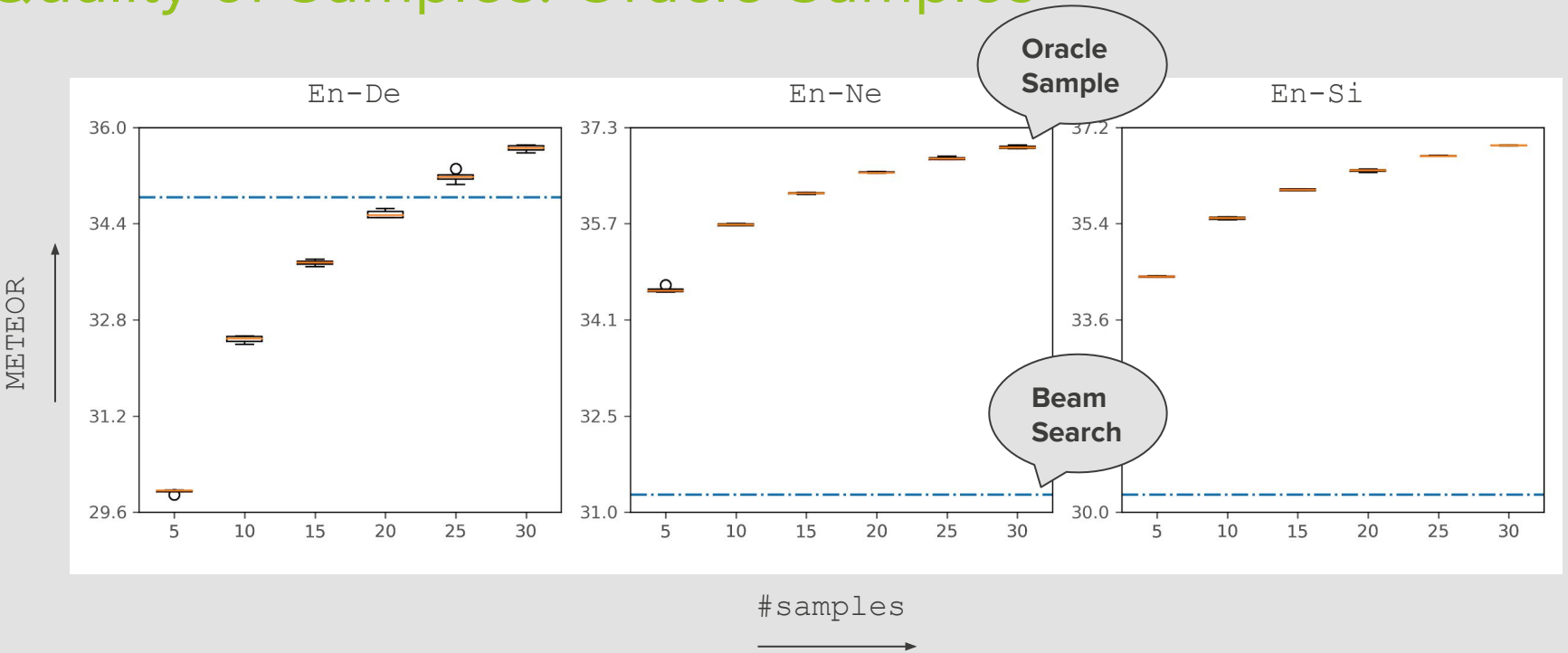


Beam Search

Samples

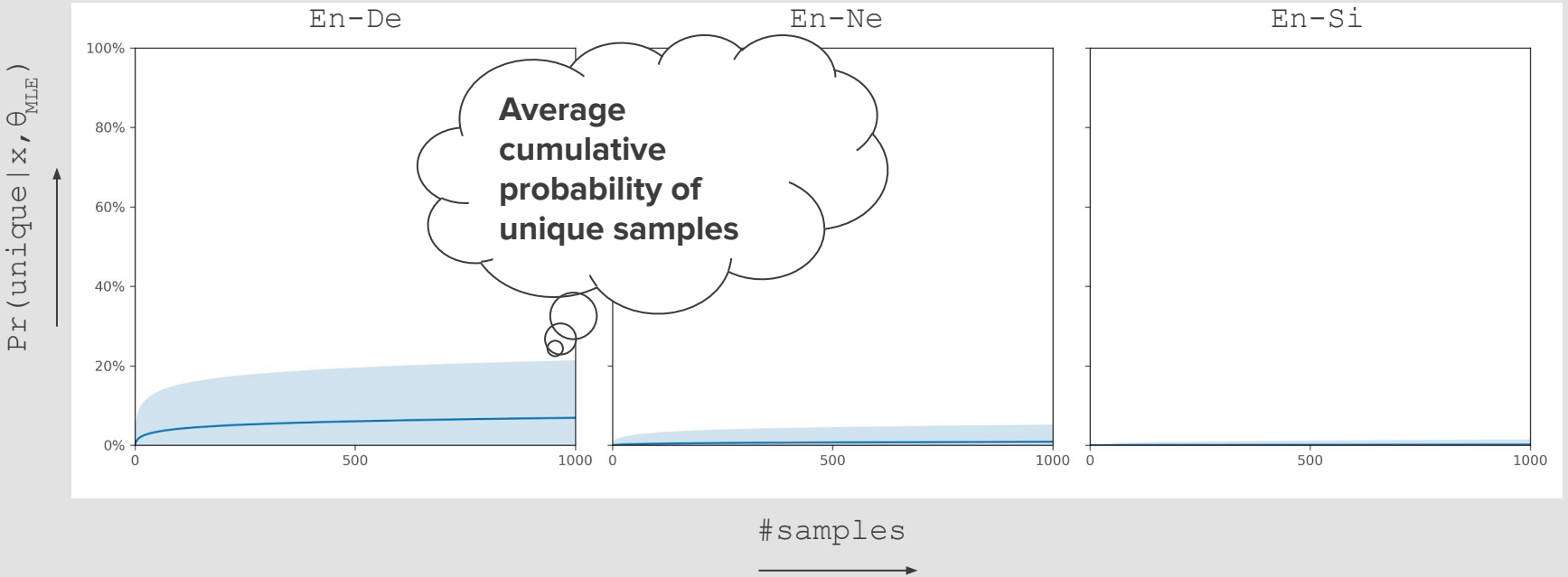
References

Quality of Samples: Oracle Samples



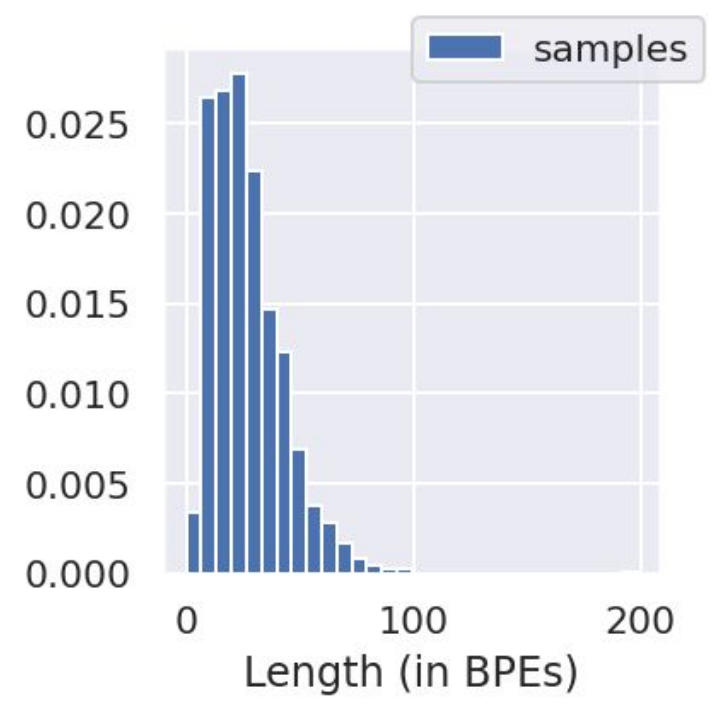
A **small number of samples** contains **good translations**

Spread of the Translation Distribution



NMT **spreads mass** over many translations

Finite length (almost surely?)



Beam vs MBR in terms of surprisal