Decoding is deciding under uncertainty

The case of neural machine translation

Bryan Eikema   Wilker Aziz

University of Amsterdam

https://Probabll.github.io

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

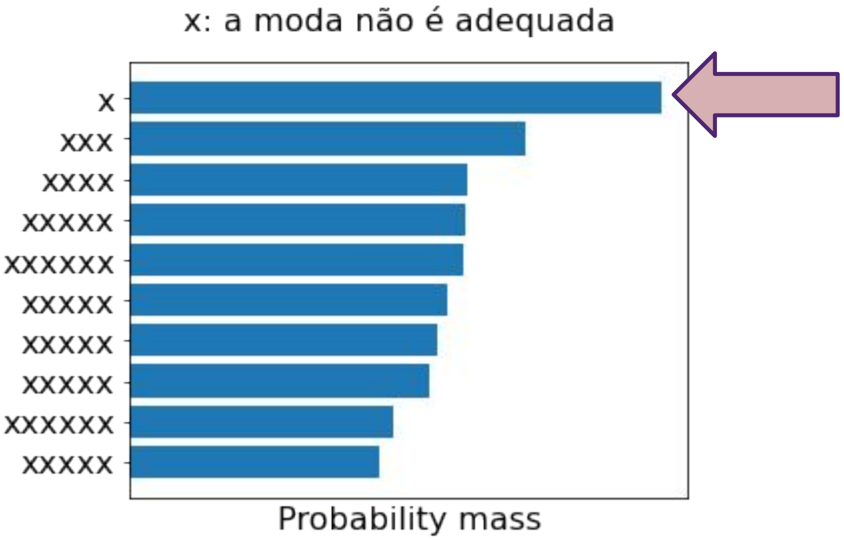
We enumerate the translation candidates that are assigned highest probability

then, regardless of

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

  (other than length)

we decide to output the **mode**.



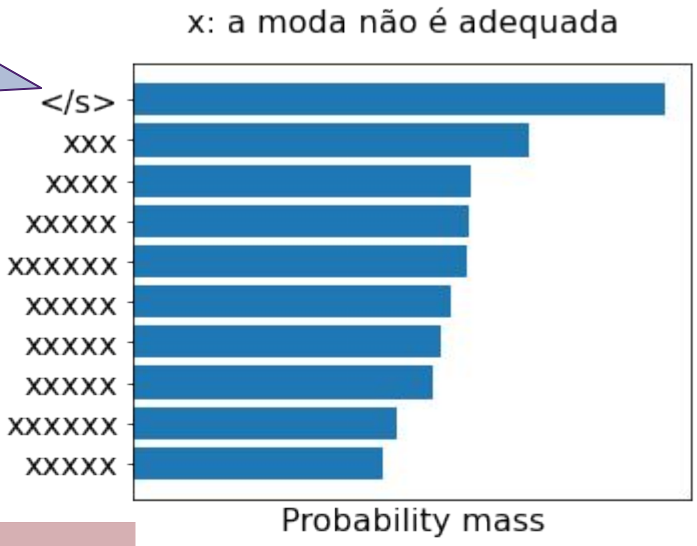
Decoding in MT - what could go wrong?

Sometimes the mode is obviously **inadequate**.

Possible conclusion:

- *Bad model! Else why would the empty string be preferred over all other possible translations?*

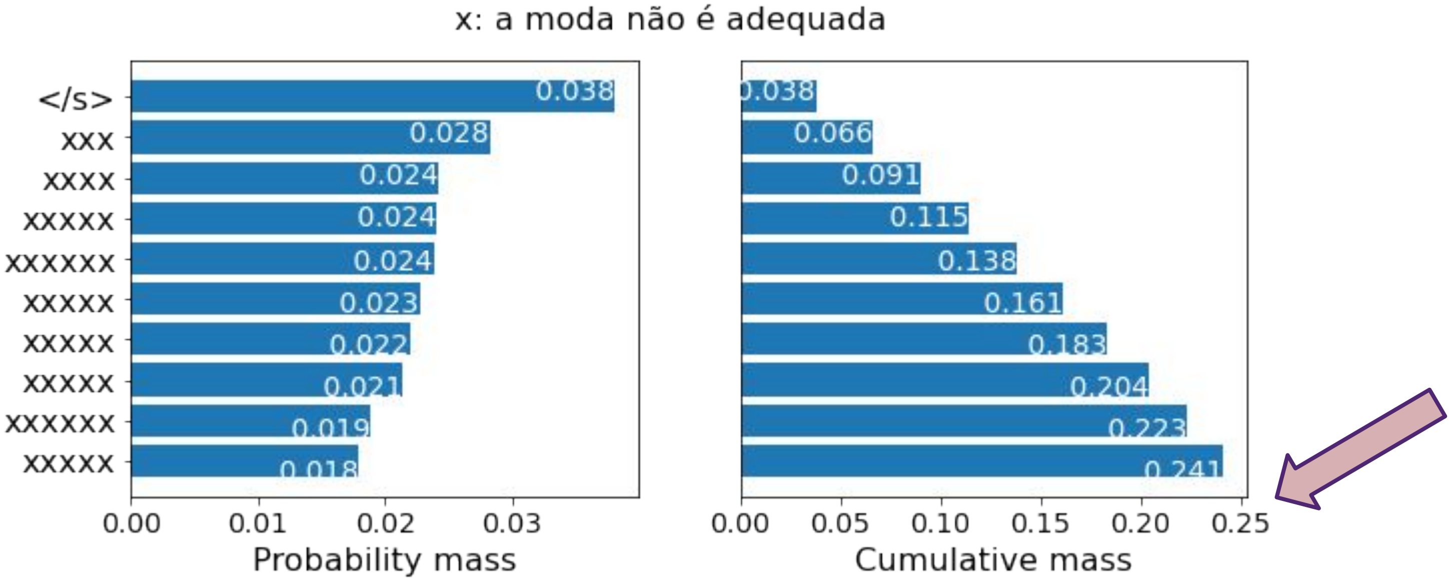
Seriously?!



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

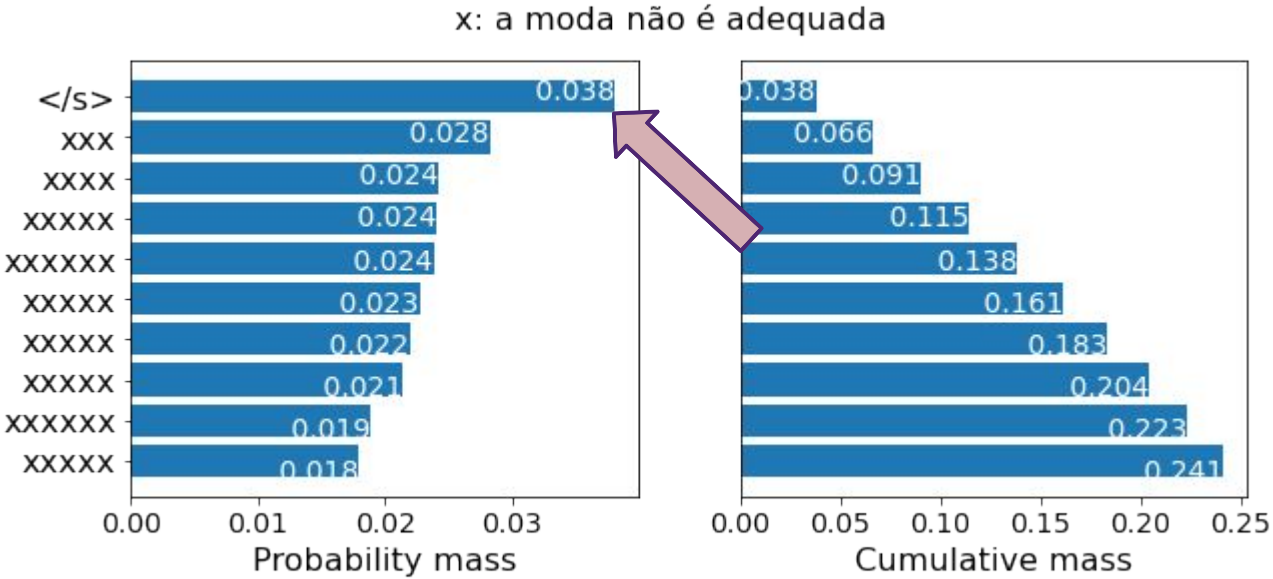
Let’s see what the decision maker chose to ignore

First, **we barely covered enough ground**. The top 10 translations cover only about 25% of the probability space (i.e., 75% of times a sample is not in this ‘top set’).



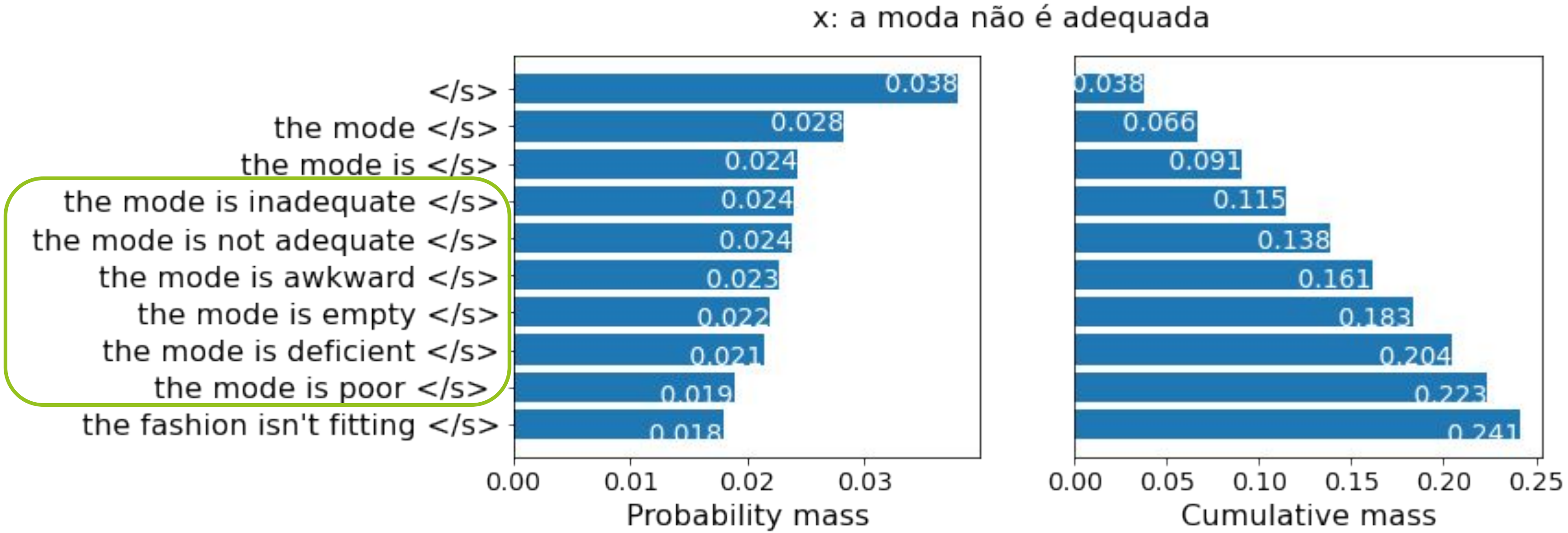
Let’s see what the decision maker chose to ignore

Second, **the mode gets less than 4% of the mass** (i.e., 96% of times a sample is non-empty). The evidence against the empty sequence is overwhelming .



Let’s see what the decision maker chose to ignore

Third, **candidates are structurally and semantically similar**. The set of adequate translations of the source may well be more probable than its complement.



Deciding under uncertainty

We tend to think of NMT models as predicting the correct translation of 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 our 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 us how to decide.
- But we can use the uncertainty NMT quantifies to make an informed decision.

Outline

1. NMT as a probabilistic model
2. Deciding under uncertainty
3. An origin story
4. What next?

NMT as a probabilistic model

NMT as a probabilistic model

NMT is trained as a **probabilistic model**:

- Given a sentence x, an NN predicts a distribution over its possible translations

$$Y \mid \theta, x$$

NMT as a probabilistic model

NMT is trained as a **probabilistic model**:

- Given a sentence x, an NN predicts a distribution over its possible translations

$$Y \mid \theta, x$$

- This distribution is factorized into **locally normalized** Categorical distributions

$$Y_j \mid \theta, x, y_{<j} \sim \mathrm{Cat}(\mathrm{NN}(x, y_{<j}; \theta))$$

NMT as a probabilistic model

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

Regularised maximum likelihood estimation

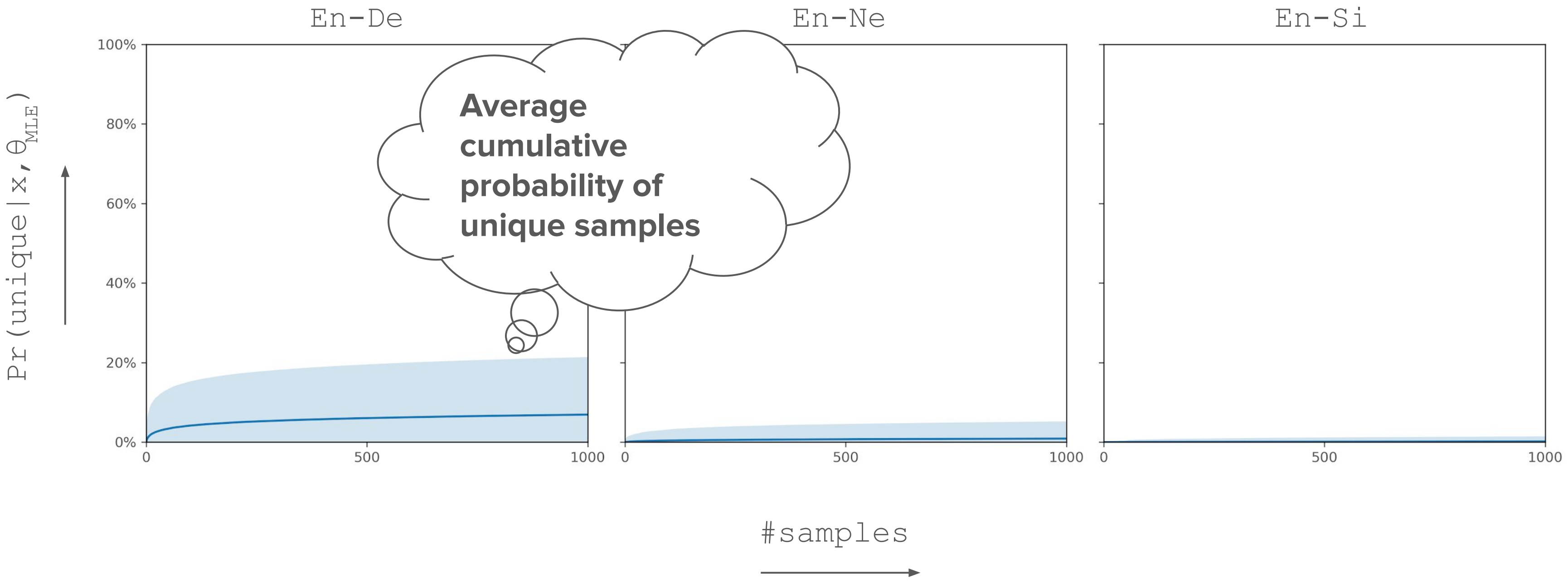
The NN parameters are chosen via (regularised) **MLE**

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

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

MLE identifies a probability distribution from where we could have sampled the observations

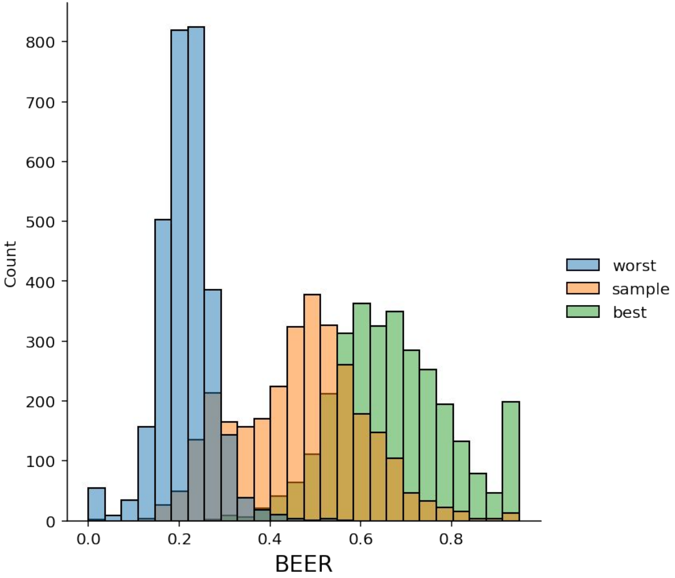
Spread of the Translation Distribution

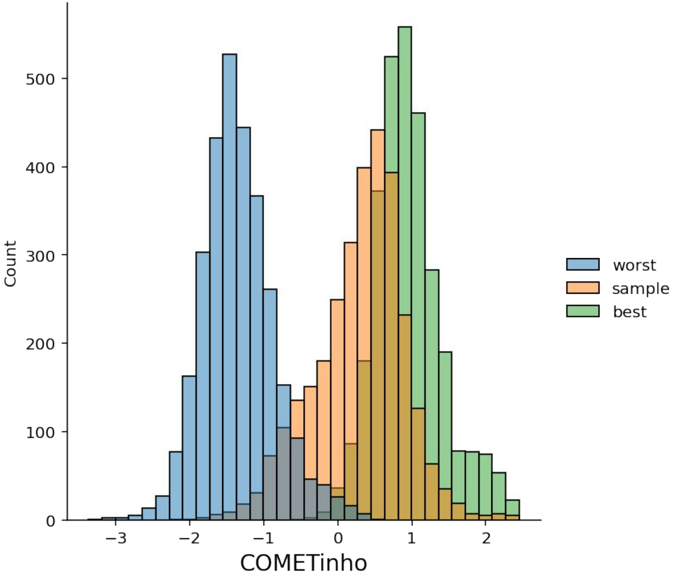


Average cumulative probability of unique samples

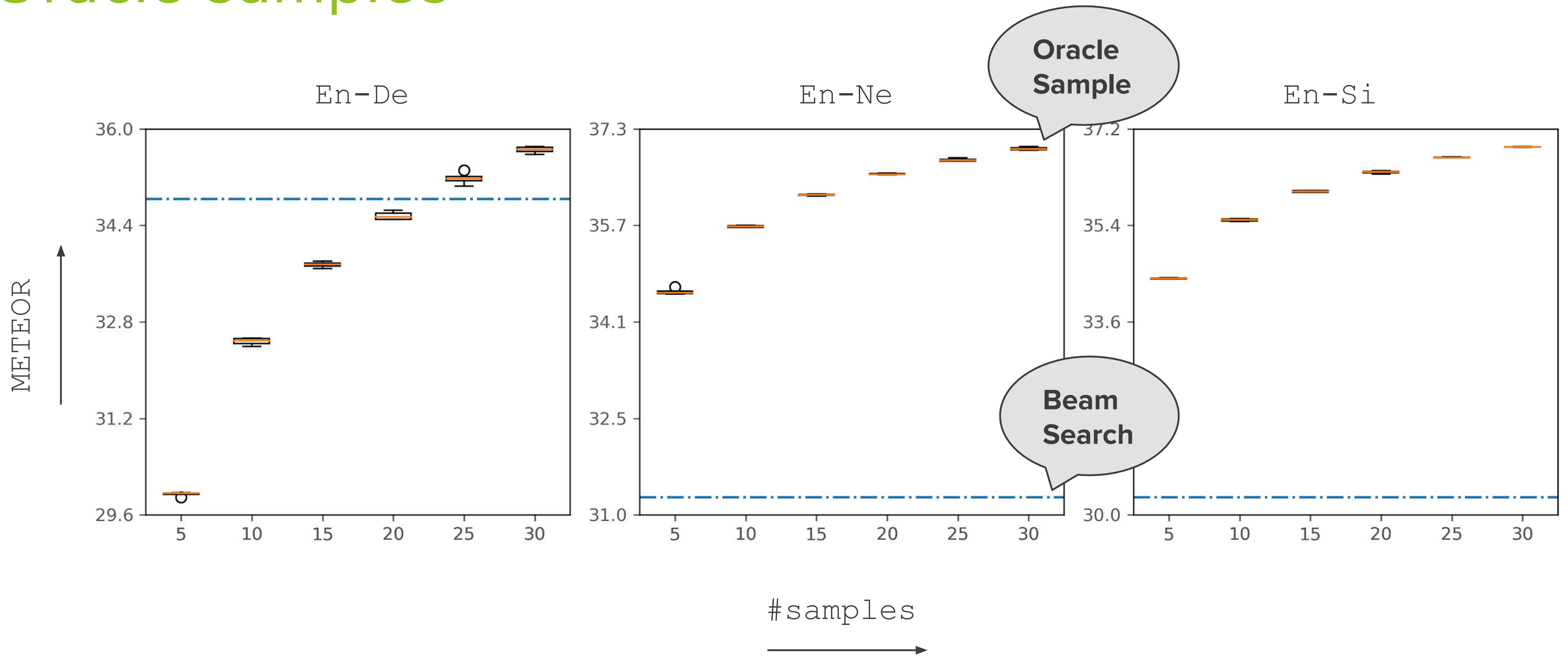
Quality of samples

Model samples are also similar to the reference translation (data samples), as measured by automatic metrics.





Oracle samples



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

Sampling the mode

Draw 1,000 samples from the model. These samples are 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 probabilistic model over translations
    - Tractable sampling
    - Tractable pmf
- Translation distributions tend to show
    - Spread-out distributions
    - But showing appreciable structural similarity with references
    - Modes are unlikely to be sampled

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-posteriori (MAP) decoding.

We have found no theoretical reason to support MAP decoding and committed to it following an intuition.

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 degrees of 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

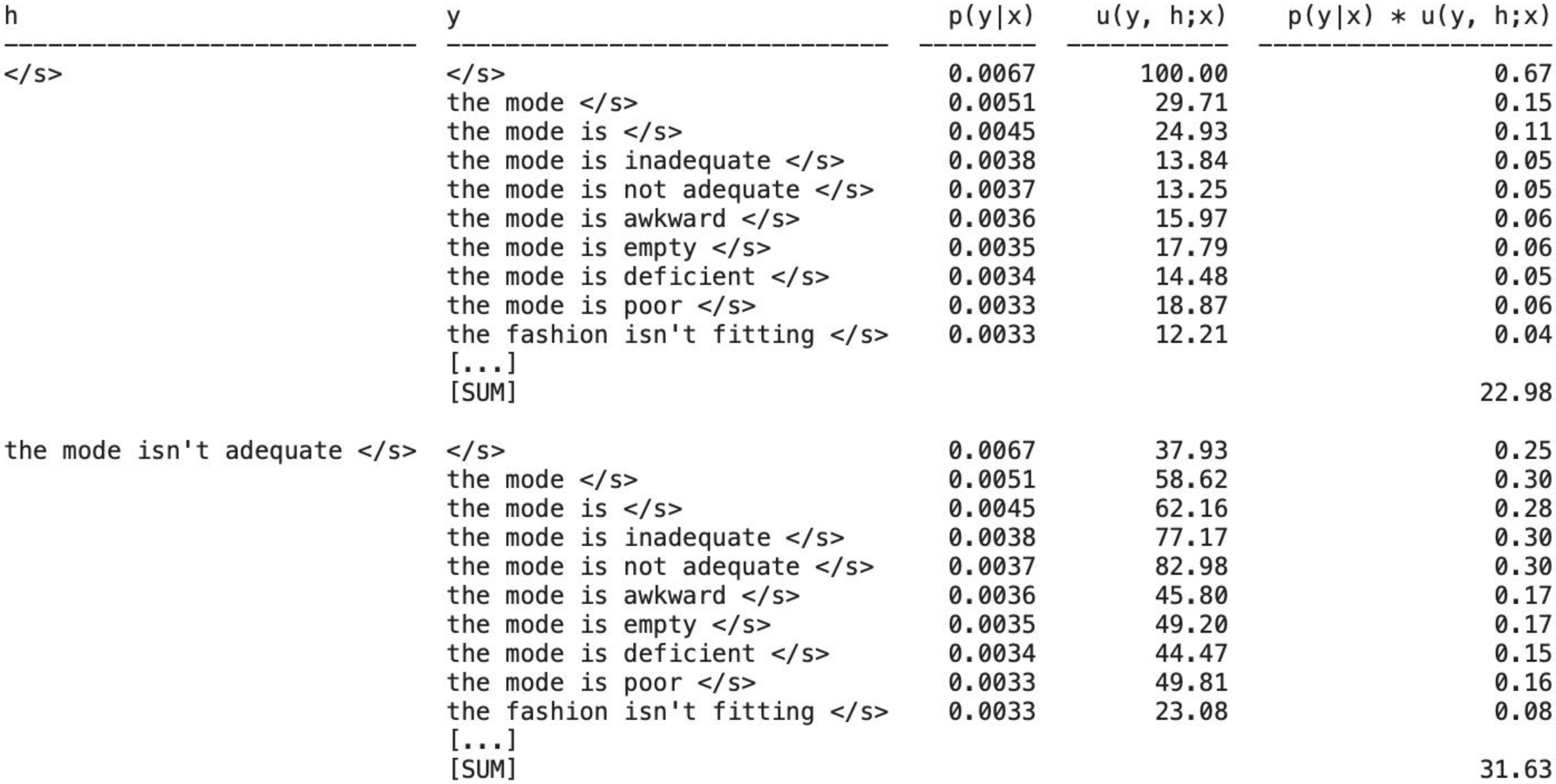
If all I know is that $\mathtt{y}$ translates $\mathtt{x}$ with probability $\mathtt{p(y|x)}$, then my expectation on $\mathtt{h}$'s utility is the weighted average utility against every possible translation under the model:

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

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

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

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

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

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

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

- 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**

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

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)

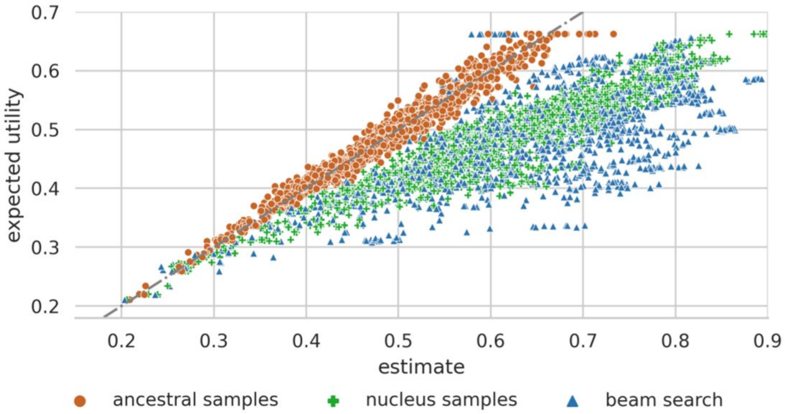
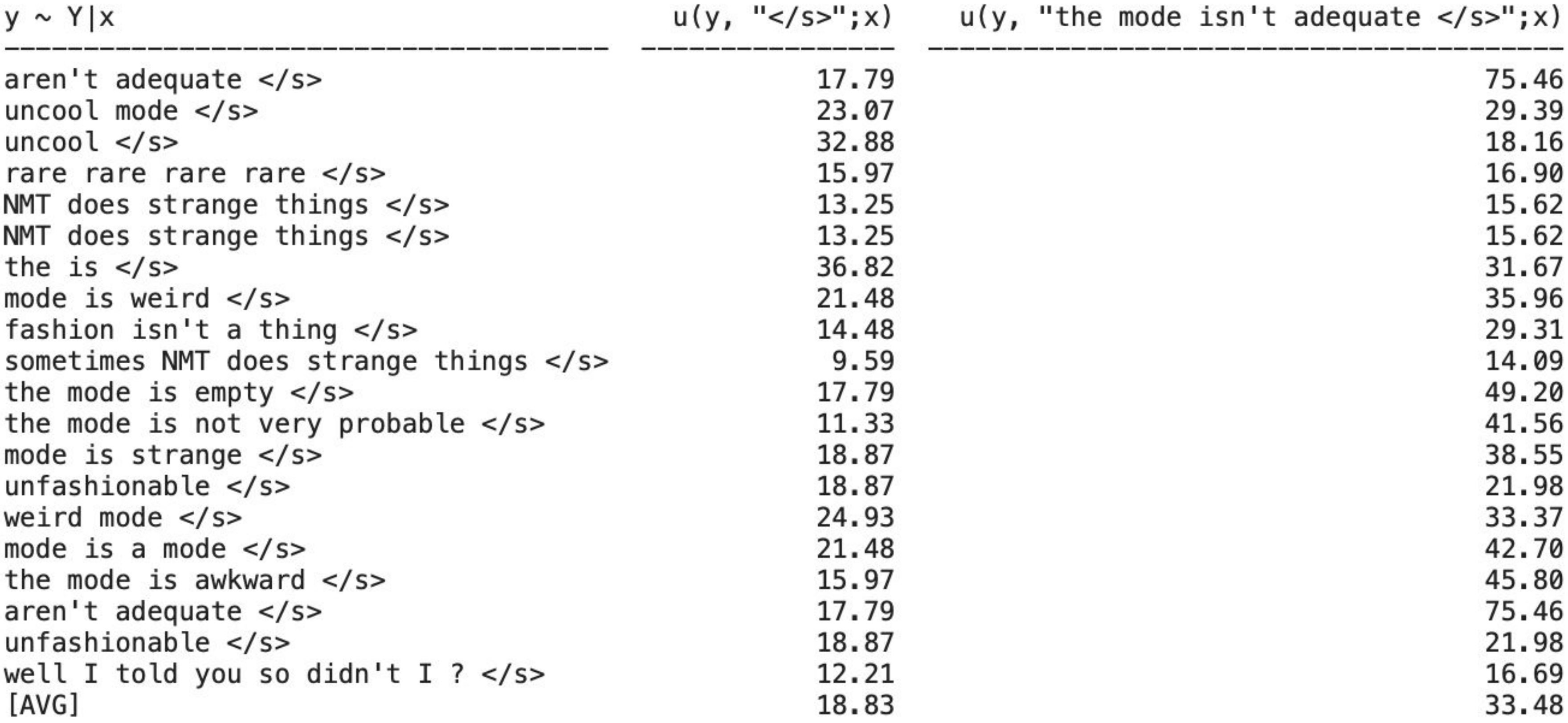


Figure 3: Estimates of expected utility for various hypotheses. We plot practical estimates of expected utility (x-axis) using either ancestral, nucleus or ‘beam’ samples against an accurate MC estimate using 1,000 ancestral samples. The gray line depicts a perfect estimator.

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
```

How about the hypothesis space?

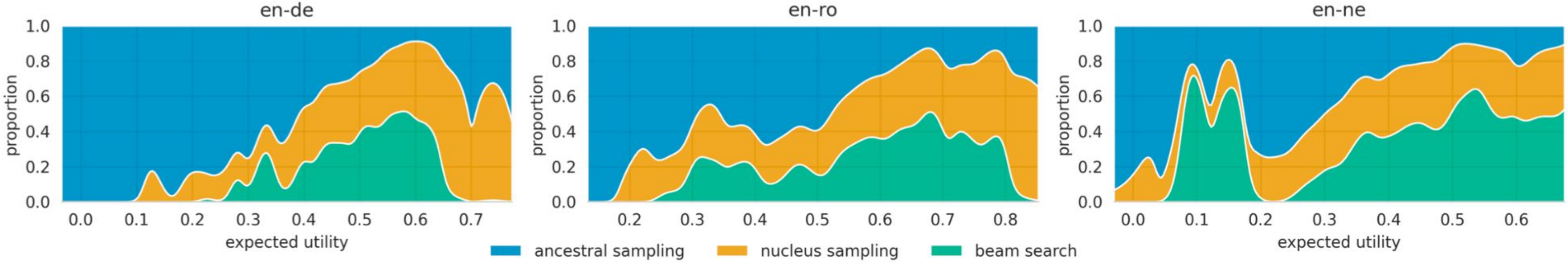
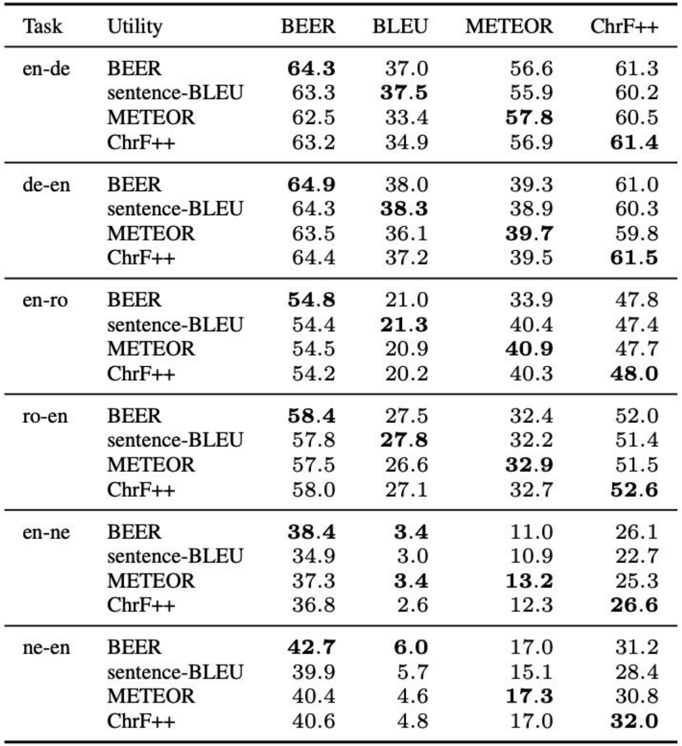


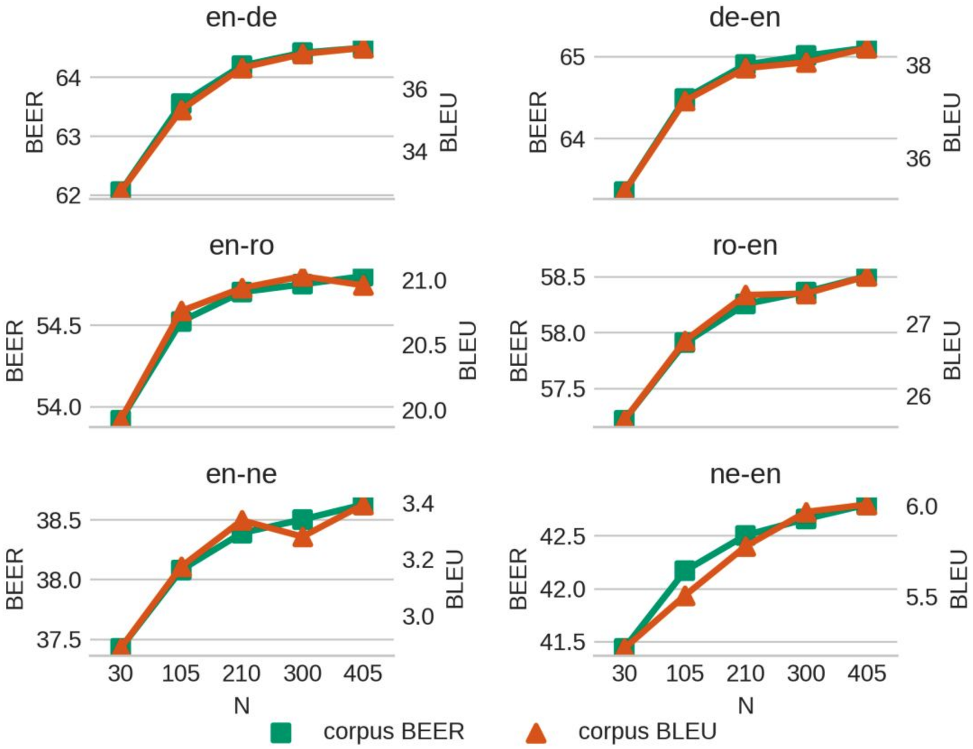
Figure 6: Proportion plots of expected utility for 3 strategies for constructing $\bar{\mathcal{H}}(x)$, using 100 translation candidates per strategy. We estimate expected utility using 1,000 samples. Results are aggregated over 100 source sentences.

Varying Utility



Task	Utility	BEER	BLEU	METEOR	ChrF++
en-de	BEER	**64.3**	37.0	56.6	61.3
	sentence-BLEU	63.3	**37.5**	55.9	60.2
	METEOR	62.5	33.4	**57.8**	60.5
	ChrF++	63.2	34.9	56.9	**61.4**
de-en	BEER	**64.9**	38.0	39.3	61.0
	sentence-BLEU	64.3	**38.3**	38.9	60.3
	METEOR	63.5	36.1	**39.7**	59.8
	ChrF++	64.4	37.2	39.5	**61.5**
en-ro	BEER	**54.8**	21.0	33.9	47.8
	sentence-BLEU	54.4	**21.3**	40.4	47.4
	METEOR	54.5	20.9	**40.9**	47.7
	ChrF++	54.2	20.2	40.3	**48.0**
ro-en	BEER	**58.4**	27.5	32.4	52.0
	sentence-BLEU	57.8	**27.8**	32.2	51.4
	METEOR	57.5	26.6	**32.9**	51.5
	ChrF++	58.0	27.1	32.7	**52.6**
en-ne	BEER	**38.4**	**3.4**	11.0	26.1
	sentence-BLEU	34.9	3.0	10.9	22.7
	METEOR	37.3	**3.4**	**13.2**	25.3
	ChrF++	36.8	2.6	12.3	**26.6**
ne-en	BEER	**42.7**	**6.0**	17.0	31.2
	sentence-BLEU	39.9	5.7	15.1	28.4
	METEOR	40.4	4.6	**17.3**	30.8
	ChrF++	40.6	4.8	17.0	**32.0**

Increasing Sample Size



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, Fernandes 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.

- 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 properties and biases in MBR decoding (Müller and Sennrich, 2021)

Deciding with neural utilities (Freitag et al, 2022, Fernandes et al, 2022)

Other axioms for decision making (Borgeaud and Emerson, 2020)

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

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

Better Approximations to Expected Utility (Eikema and Aziz, 2022)

Key takeaways

We question the use of MAP decoding in NMT

- The mode is not a particularly informative summary of the NMT models' beliefs
- Regularised MLE-training is unaware of the MAP decoder
- MAP decoding is a special case of a more general principle and makes a suboptimal choice of utility.

We argue that:

- Models convey beliefs through **expectations** (not modes)
- **Unbiased samples** summarise such beliefs
- Decision making requires a **utility function**

Thanks!