Analyzing Uncertainty in Neural Machine Translation

Myle Ott
myleott@fb.com

Michael Auli

David Grangier

Marc'Aurelio Ranzato

Facebook AI Research

facebook
Artificial Intelligence Research

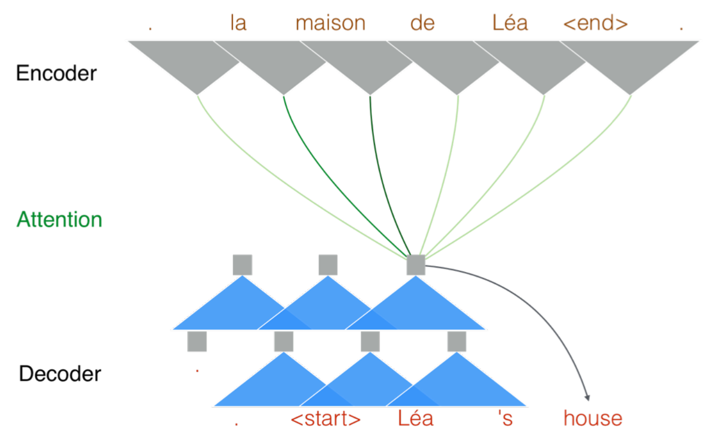
Background

Neural Machine Translation

Input: source sentence $X = \{x_1, ..., x_N\}$

Output: target translation $Y = \{y_1, ..., y_T\}$

$$p(Y|X;\theta) = \prod_{t=1}^{T} p(y_t|y_{1:t-1}, X;\theta)$$



(Gehring et al., 2017)

Background

Training: maximum likelihood (autoregressive) with cross entropy loss

$$\mathcal{L}_{\mathrm{ML}} = \sum_{t=1}^{T} \log p(y_t | y_{1:t-1}, X; \theta)$$

Inference: sampling or MAP

$$\hat{y}_{\mathrm{MAP}} = \arg\max_{w_{1:T}} \sum_{t} \log p(w_t | w_{1:t-1}, X; \theta)$$

Background

Training: maximum likelihood (autoregressive) with cross entropy loss

$$\mathcal{L}_{\mathrm{ML}} = \sum_{t=1}^{T} \log p(y_t | y_{1:t-1}, X; \theta)$$

Inference: sampling or MAP

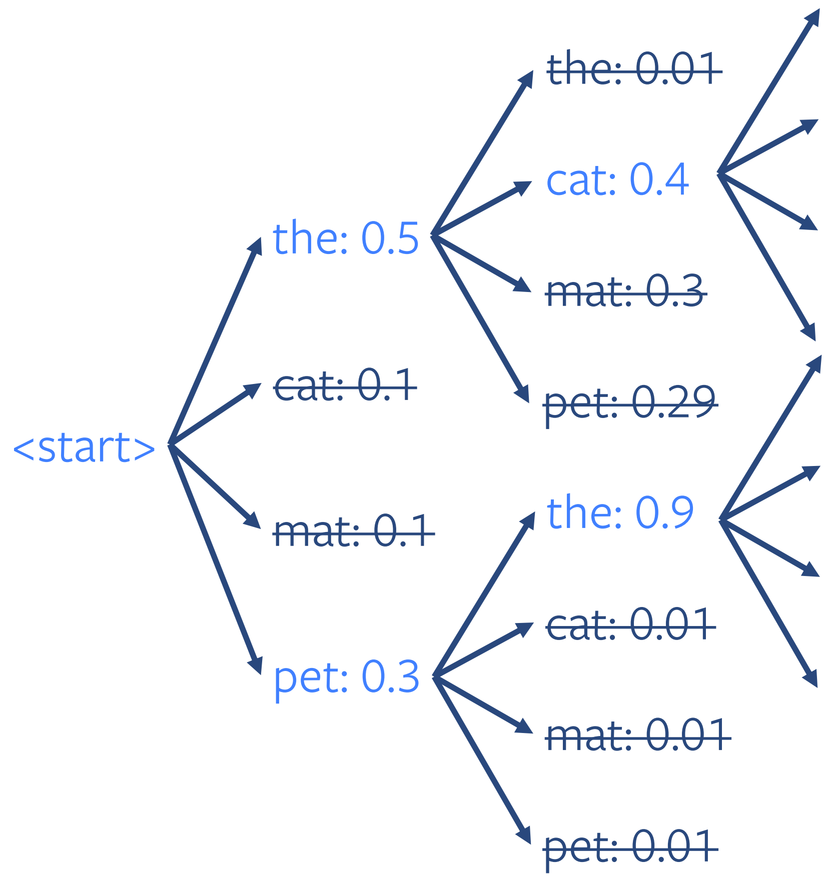
Intractable to enumerate

$$\hat{y}_{\mathrm{MAP}} = \boxed{\arg\max_{w_{1:T}}} \sum_{t} \log p(w_t | w_{1:t-1}, X; \theta)$$

.\ I Background

Approximate inference with beam search

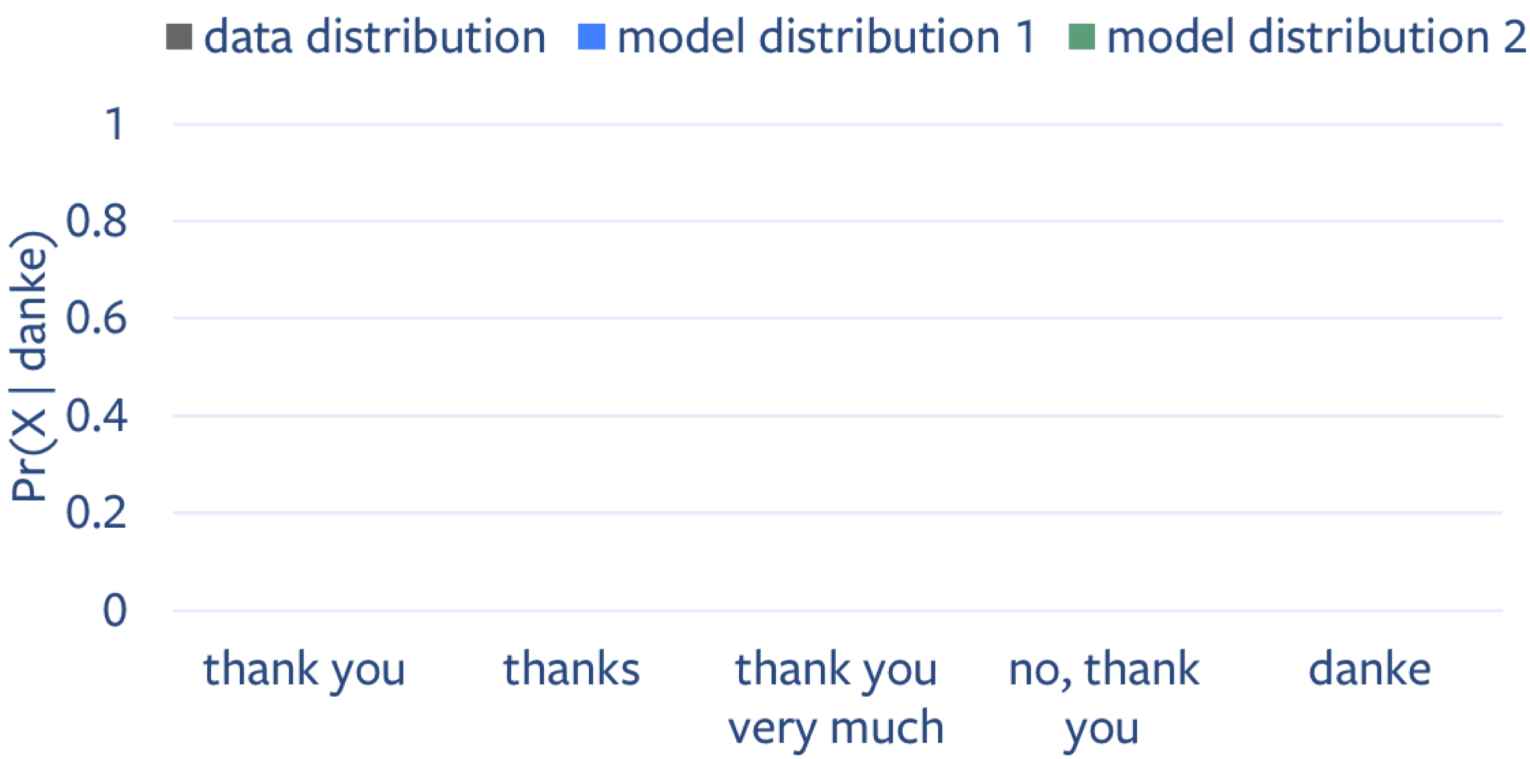
- Decode sequence left-to-right and keep K best hypotheses at each step
- Equiv. to greedy search when the beam width (K) = 1



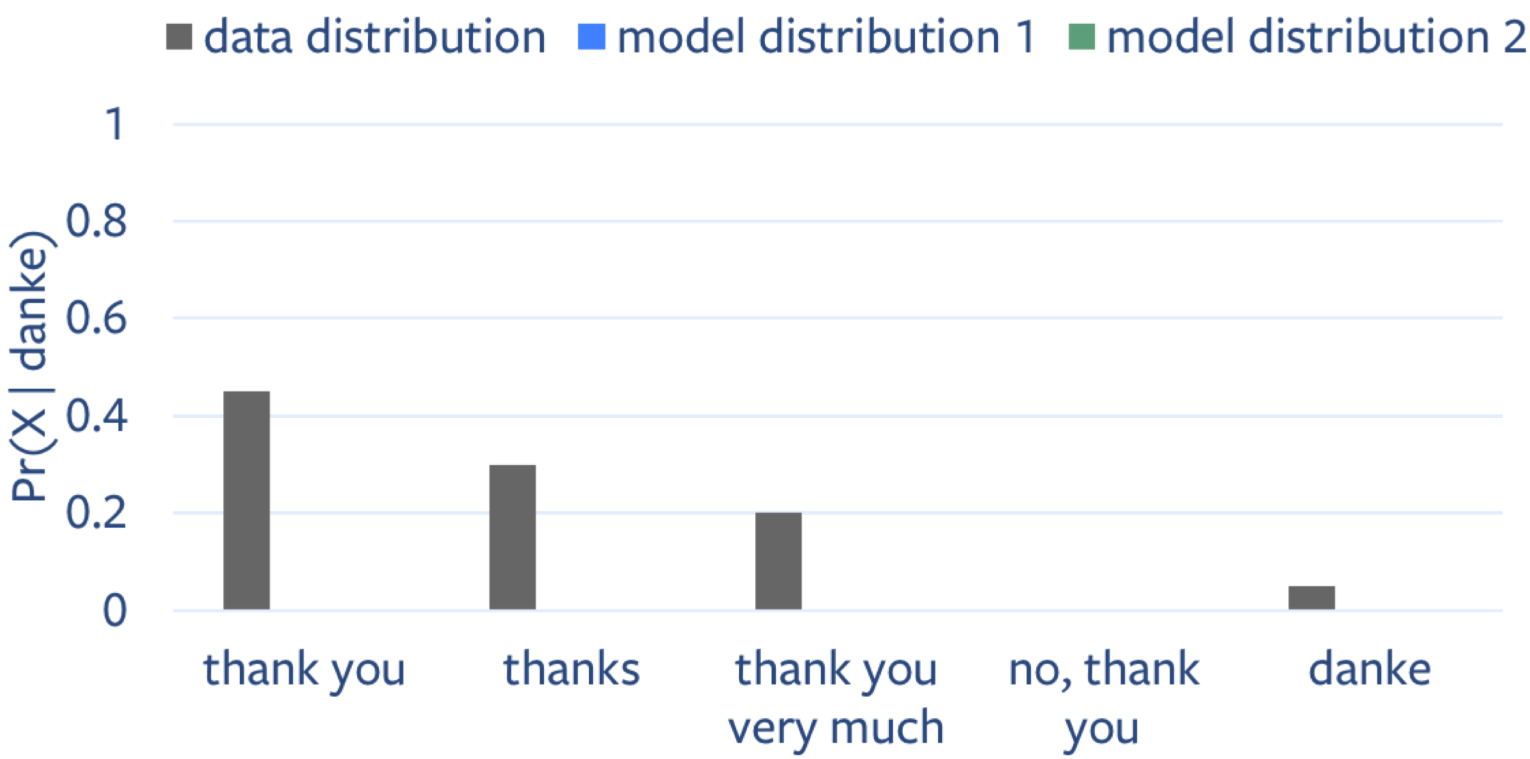
This work

Goal: Investigate the effects of uncertainty in NMT model fitting and search

This work

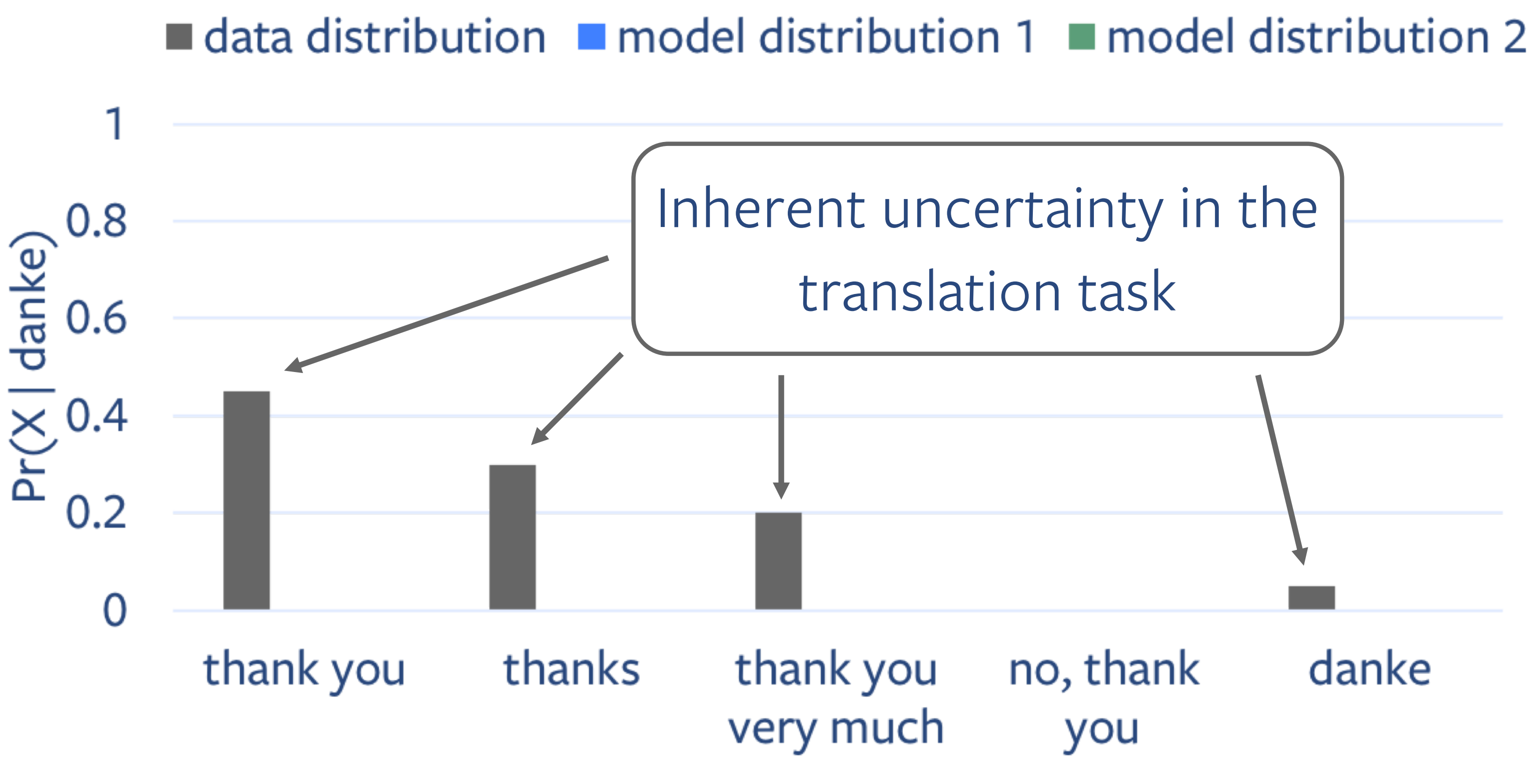

data distribution
model distribution 1
model distribution 2
Pr(X | danke)
1
0.8
0.6
0.4
0.2
0
thank you
thanks
thank you very much
no, thank you
danke

This work

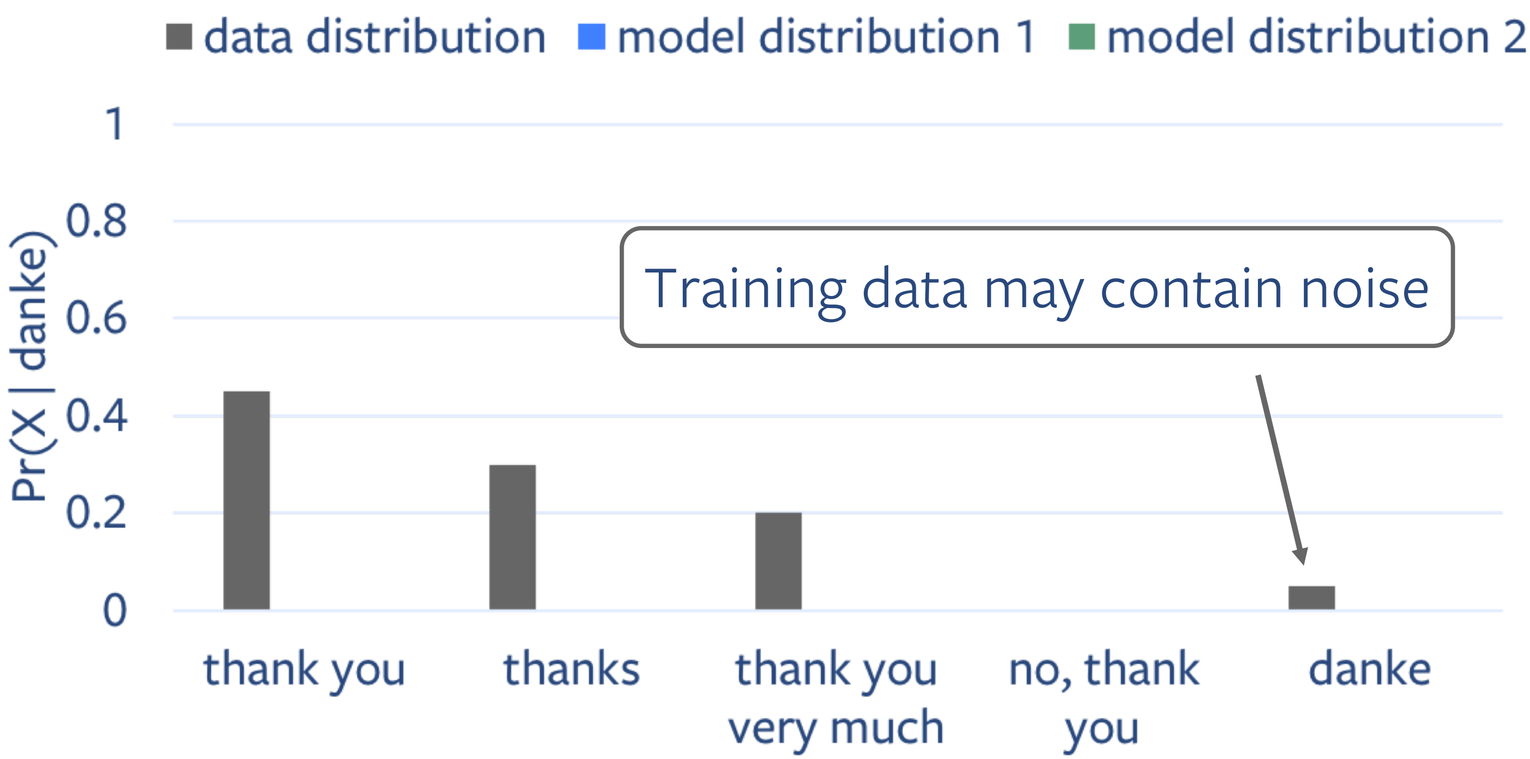


data distribution
model distribution 1
model distribution 2
1
0.8
0.6
0.4
0.2
0
Pr(X | danke)
thank you
thanks
thank you very much
no, thank you
danke

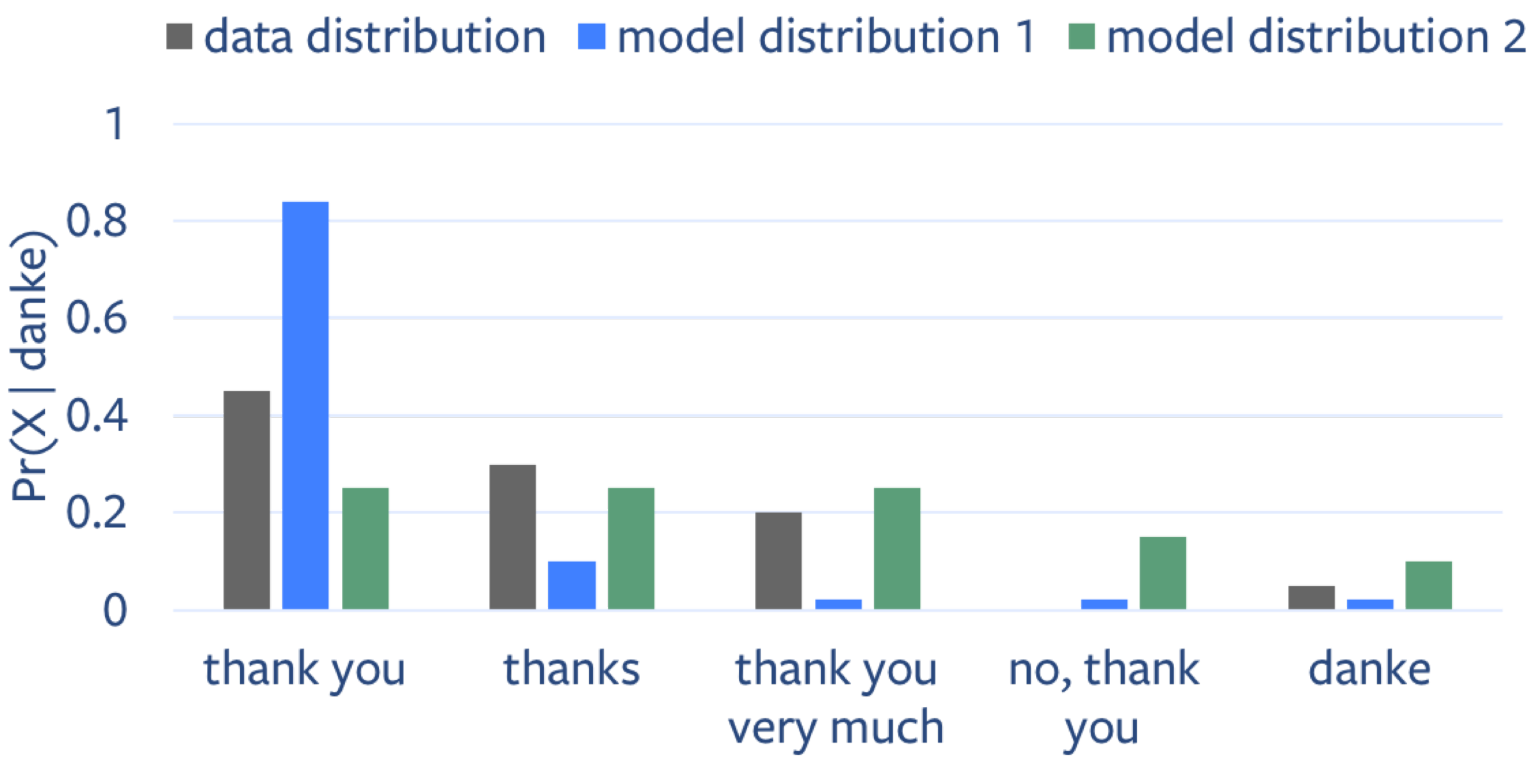
This work


data distribution
model distribution 1
model distribution 2
1
0.8
0.6
0.4
0.2
0
Pr(X | danke)
Inherent uncertainty in the translation task
thank you
thanks
thank you very much
no, thank you
danke

This work

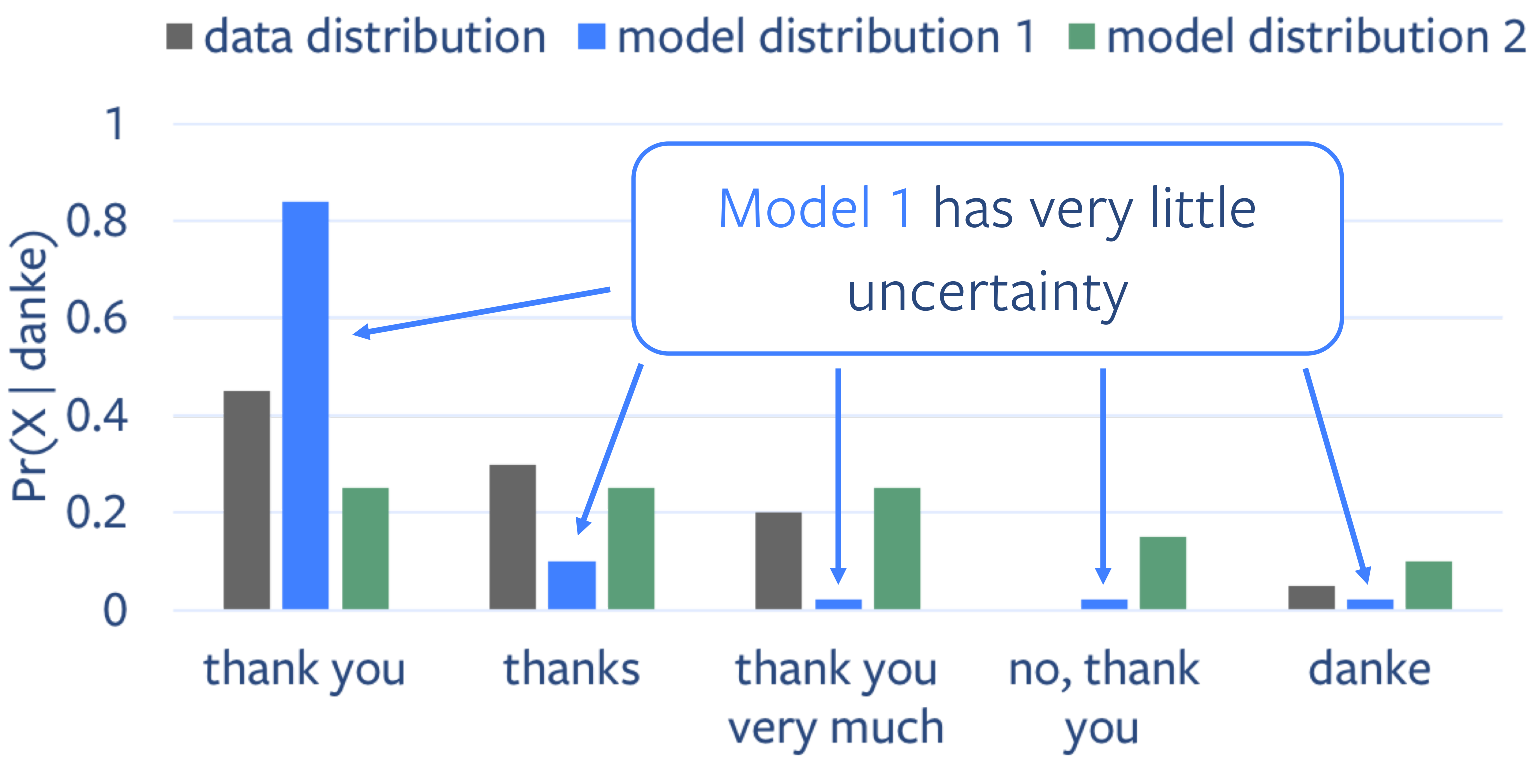

data distribution
model distribution 1
model distribution 2
Pr(X | danke)
1
0.8
0.6
0.4
0.2
0
Training data may contain noise
thank you
thanks
thank you very much
no, thank you
danke

This work



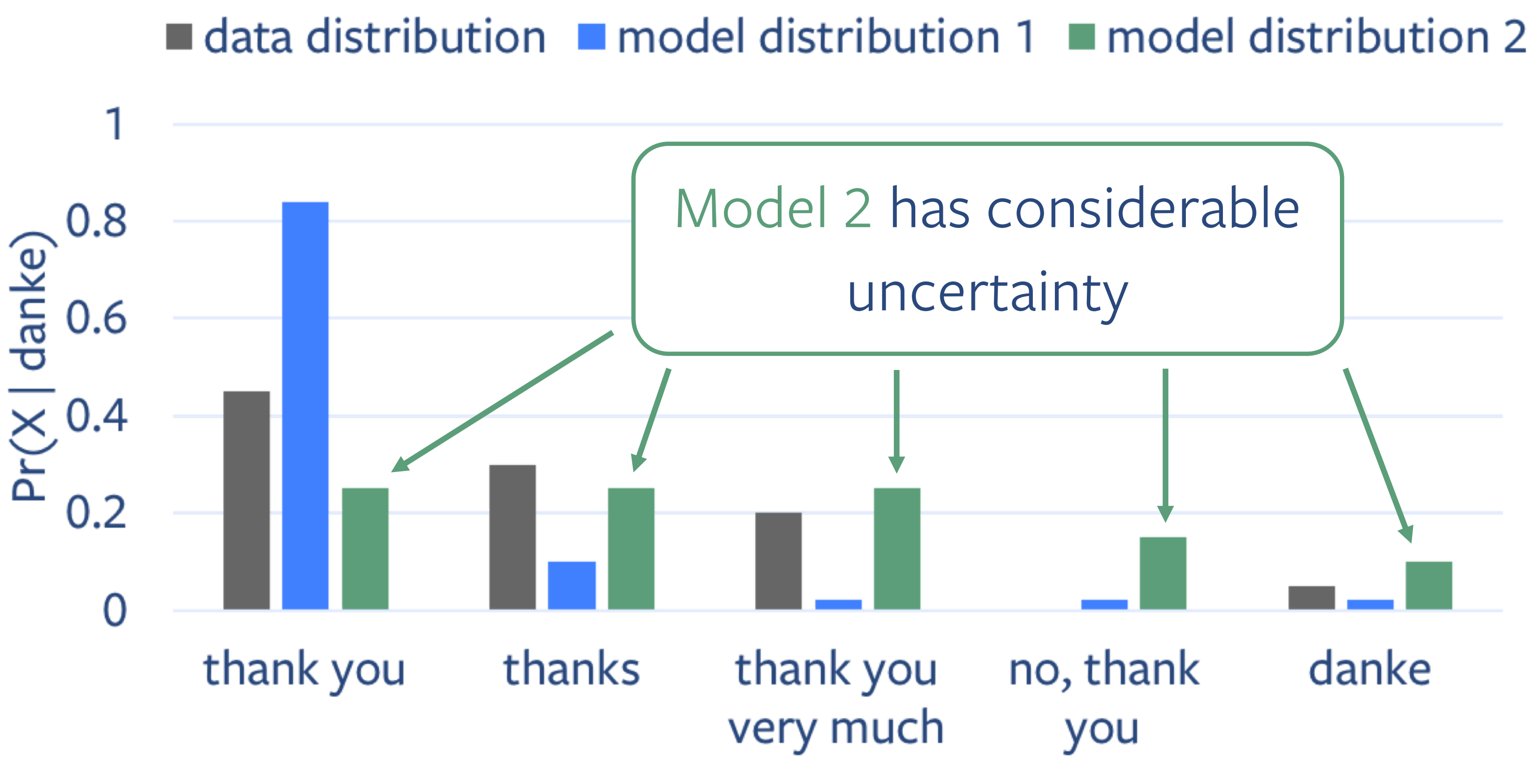
data distribution
model distribution 1
model distribution 2
Pr(X | danke)
1
0.8
0.6
0.4
0.2
0
thank you
thanks
thank you very much
no, thank you
danke

This work



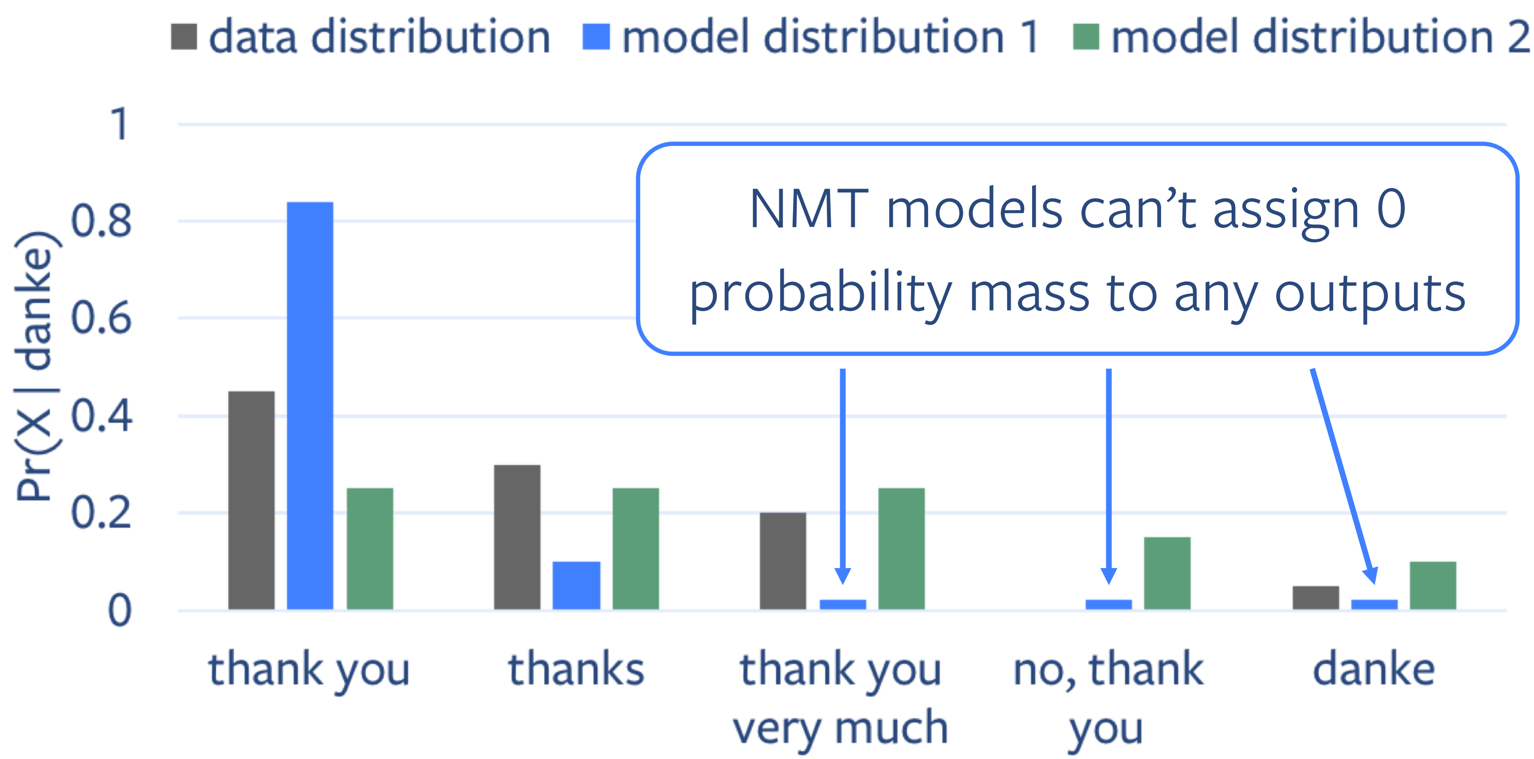
data distribution
model distribution 1
model distribution 2
1
0.8
0.6
0.4
0.2
0
Pr(X | danke)
Model 1 has very little uncertainty
thank you
thanks
thank you very much
no, thank you
danke

This work



data distribution
model distribution 1
model distribution 2
1
0.8
0.6
0.4
0.2
0
Pr(X | danke)
Model 2 has considerable uncertainty
thank you
thanks
thank you very much
no, thank you
danke

This work



data distribution
model distribution 1
model distribution 2
1
0.8
0.6
0.4
0.2
0
Pr(X | danke)
NMT models can't assign 0 probability mass to any outputs
thank you
thanks
thank you very much
no, thank you
danke

This work

Goal: Investigate the effects of uncertainty in NMT model fitting and search

- Do NMT models capture uncertainty, and how is this uncertainty represented in the model's output distribution?
- How does uncertainty affect search?
- How closely does the model distribution match the data distribution?
- How do we answer these questions with (typically) only a single reference translation per source sentence?

Experimental setup

Convolutional sequence-to-sequence models* (Gehring et al., 2017)

Evaluation: compare translations with BLEU (Papineni et al., 2002)

- Modified n-gram precision metric, values from 0 (worst) to 100 (best)

Datasets: WMT14 English-French and English-German

- Mixture of news, parliamentary and web crawl data

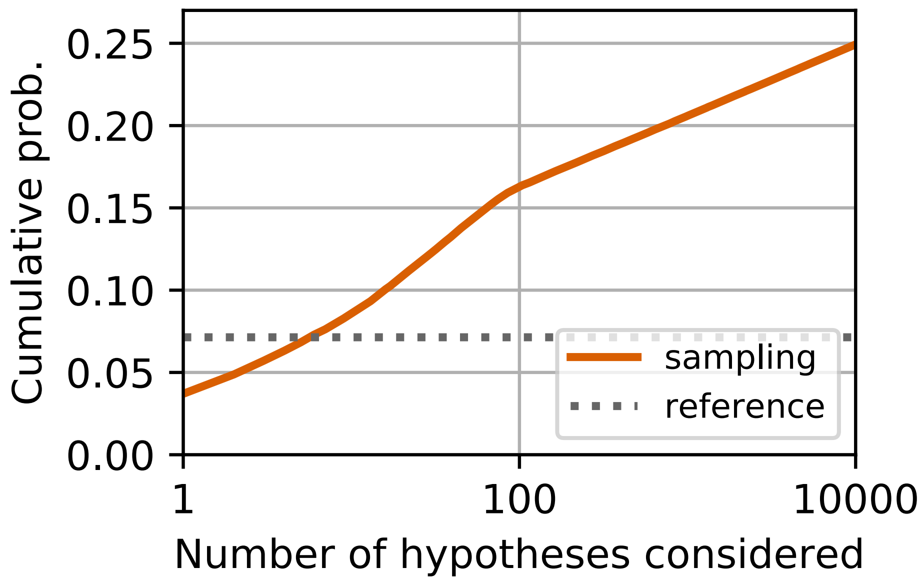
* Results hold for other tested architectures too, e.g., LSTM

Do NMT models capture uncertainty?

Question: How much uncertainty is there in the model's output distribution?

Experiment: How many independent samples does it take to cover most of the sequence-level probability mass?

Do NMT models capture uncertainty?



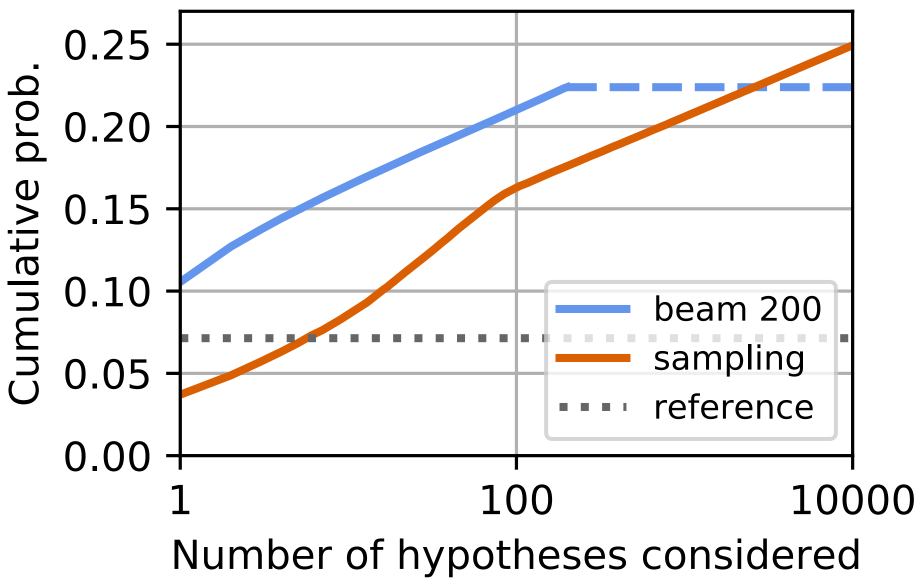
(WMT14 En-Fr)

Model's output distribution is **highly uncertain**!

- Even after 10K samples we cover only 25% of sequence-level probability mass

What about beam search?

Do NMT models capture uncertainty?



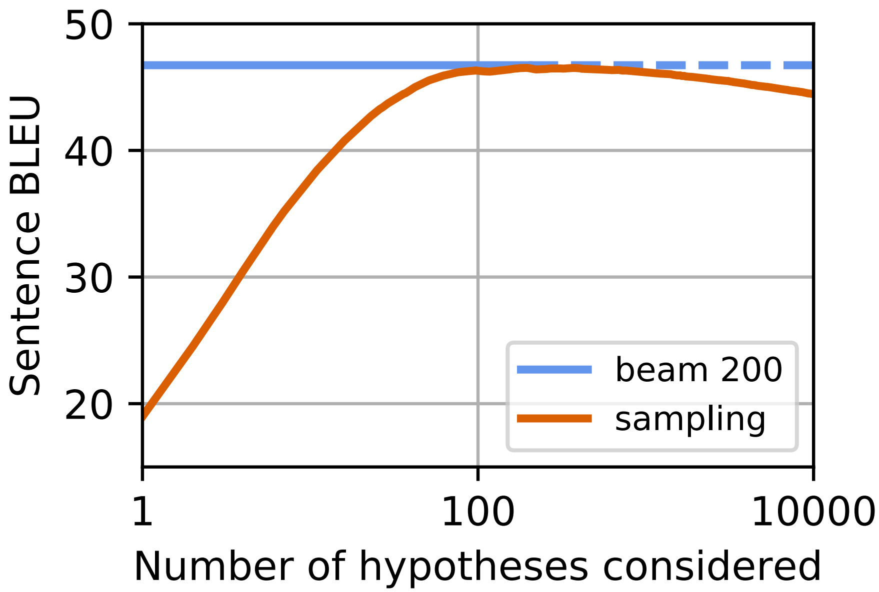
(WMT14 En-Fr)

Beam search is very efficient!

The reference score (····)
is lower than beam hypotheses

What is the quality (BLEU) of
these translations?

Uncertainty and Search



(WMT14 En-Fr)

Beam search is efficient **and produces accurate translations**

Sampling produces increasingly likely hypotheses, but these get worse BLEU after ~200

.\| Uncertainty and Search

Source: The first nine episodes of Sheriff Callie ’s Wild West will be available (…)

Reference: Les neuf premiers épisodes de shérif Callie’ s Wild West seront disponibles (…)

Hypothesis: The first nine episodes of Sheriff Callie ’s Wild West will be available (…)

Uncertainty and Search

Source: The first nine episodes of Sheriff Callie 's Wild West will be available (…)

log probs: -4.53 -0.02 -0.28 -0.11 -0.01 -0.001 -0.004 -0.002

Hypothesis: The first nine episodes of Sheriff Callie 's Wild West will be available (…)

Uncertainty and Search

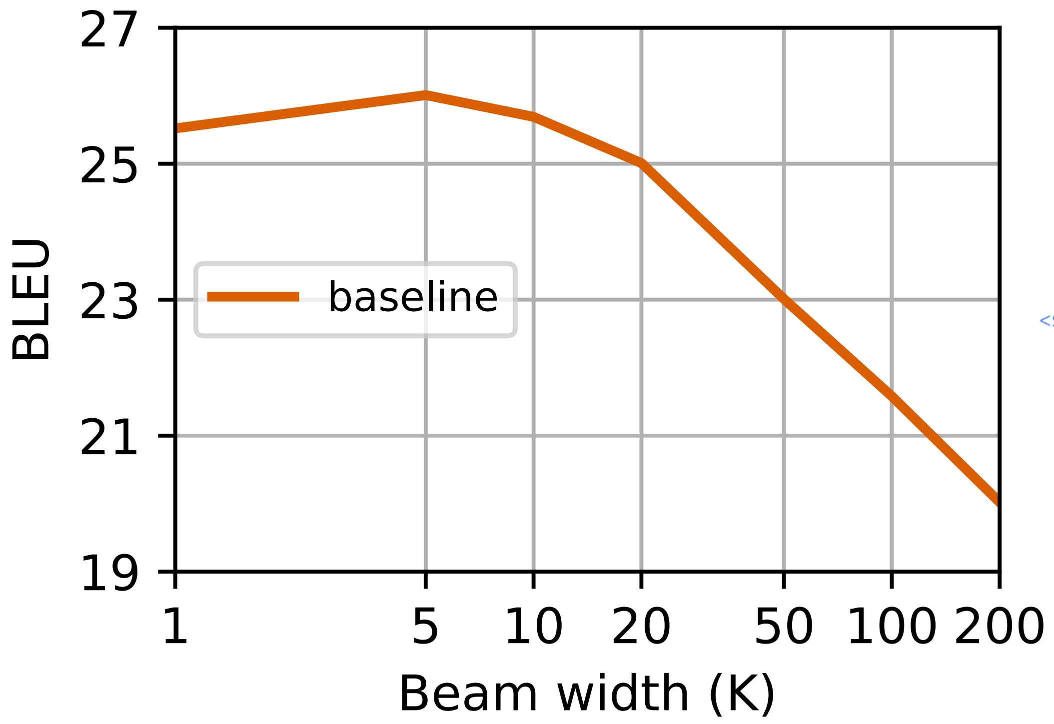
Copies* make up 2.0% of the WMT14 En-Fr training set, but are **over-represented in the output of beam search**

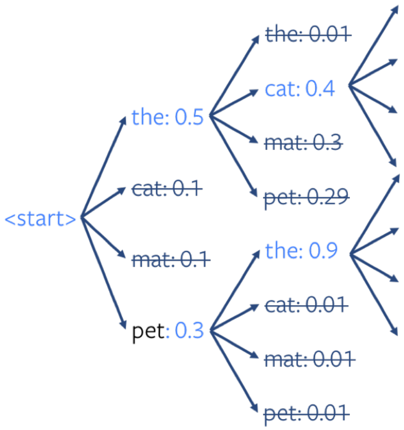
- Among beam hypotheses, copies account for:

Beam=1: 2.6% Beam=5: 2.9% Beam=20: 3.5%

* a copy is a translation that shares >= 50% of its unigrams with the source

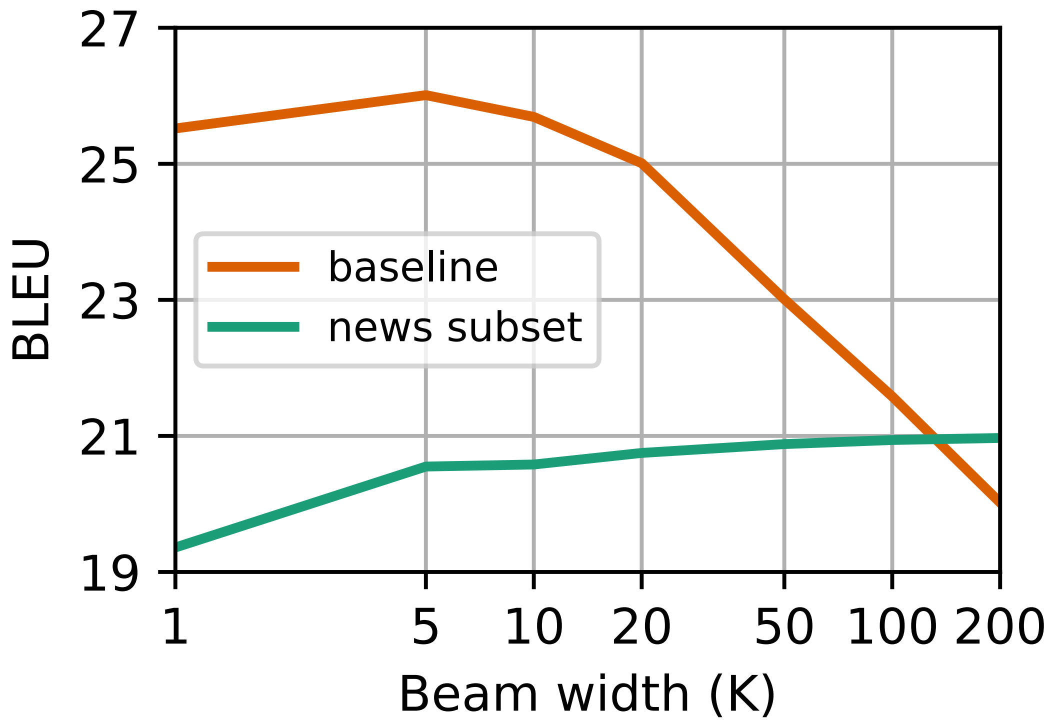
Uncertainty and Search





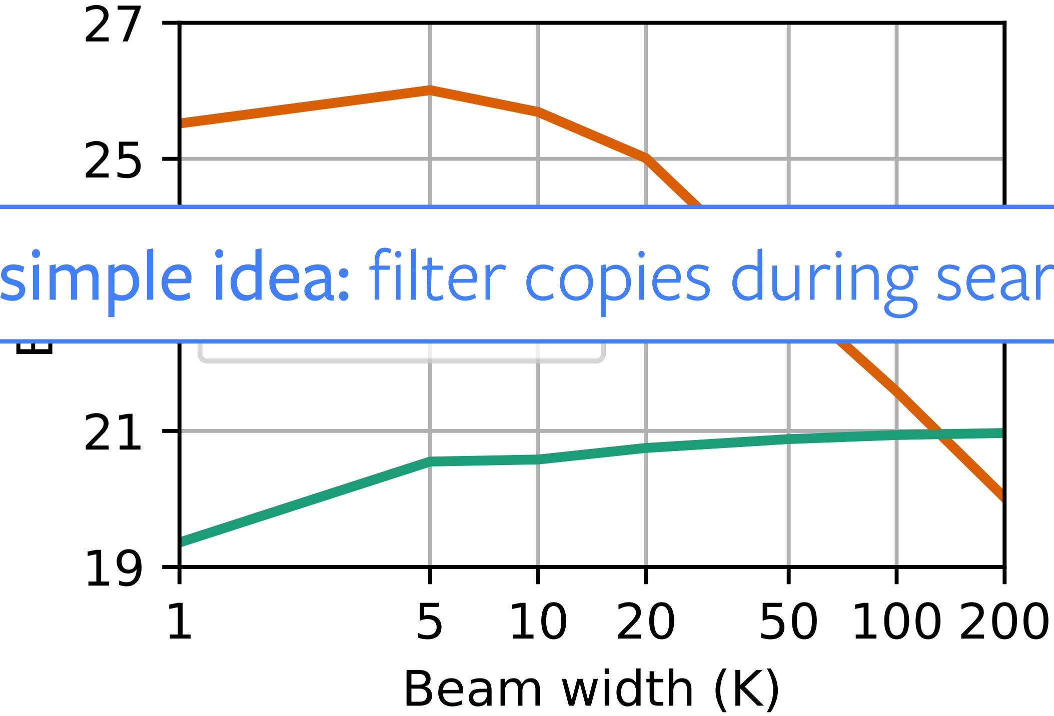
(WMT17 En-De)

.\ | Uncertainty and Search



(WMT17 En-De)

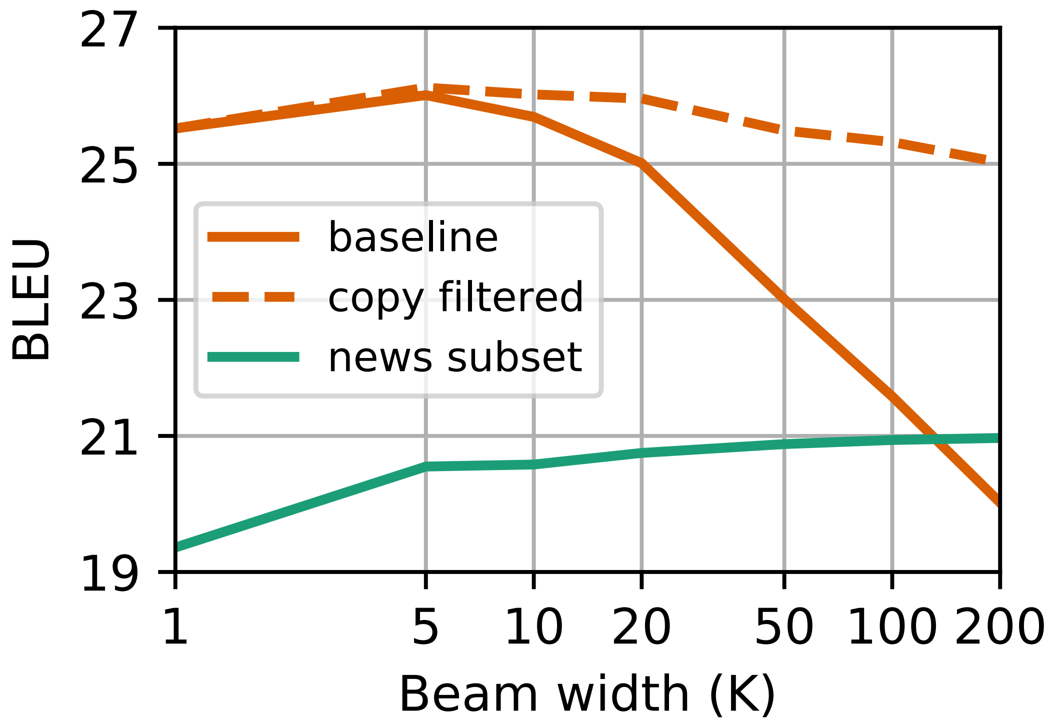
Uncertainty and Search



A simple idea: filter copies during search

(WMT17 En-De)

Uncertainty and Search



(WMT17 En-De)

Do NMT models capture uncertainty?

Yes, with interesting effects on search!

Follow-up: How is it represented? Does it match the data distribution?

Challenging because:

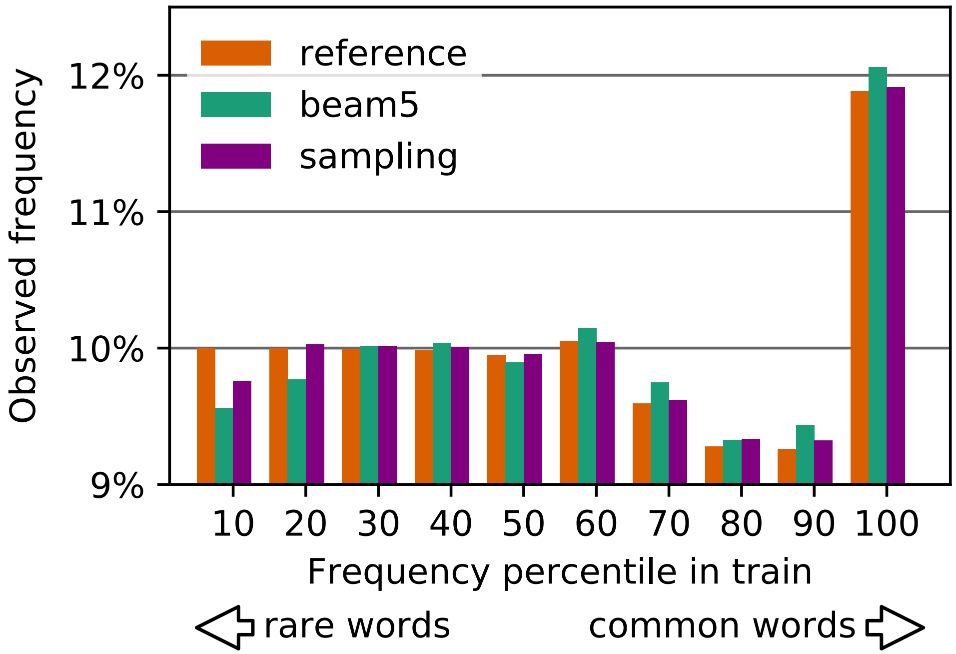
- We typically observe only a single sample from the data distribution for each source sentence (i.e., one reference translation)
- The model distribution is intractable to enumerate

Necessary matching conditions

What are the necessary conditions for the model distribution to match the data distribution:

- ...at the token level?
- ...at the sequence level?
- ...when considering multiple reference translations?

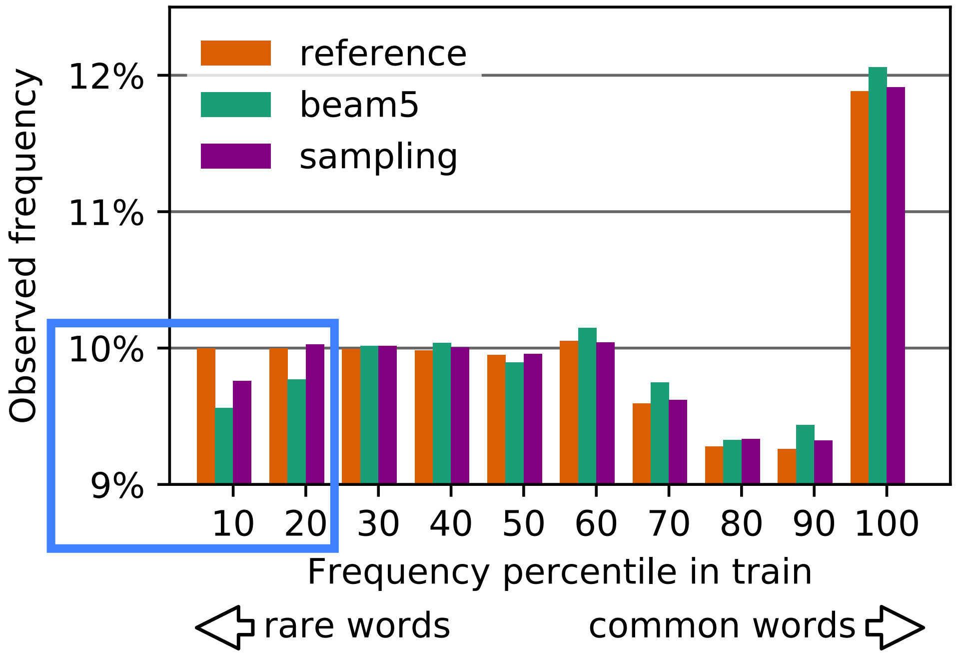
Necessary matching conditions—Token Level



(WMT14 En-Fr)

Histogram of unigram frequencies

Necessary matching conditions—Token Level

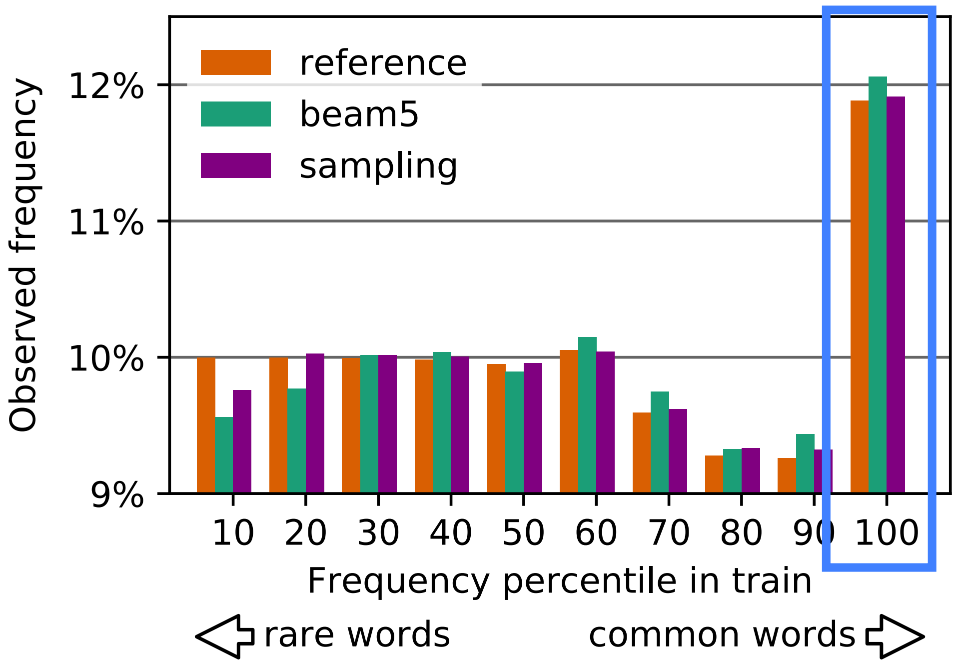


(WMT14 En-Fr)

Histogram of unigram frequencies

Beam under-estimates the rarest words

Necessary matching conditions—Token Level



(WMT14 En-Fr)

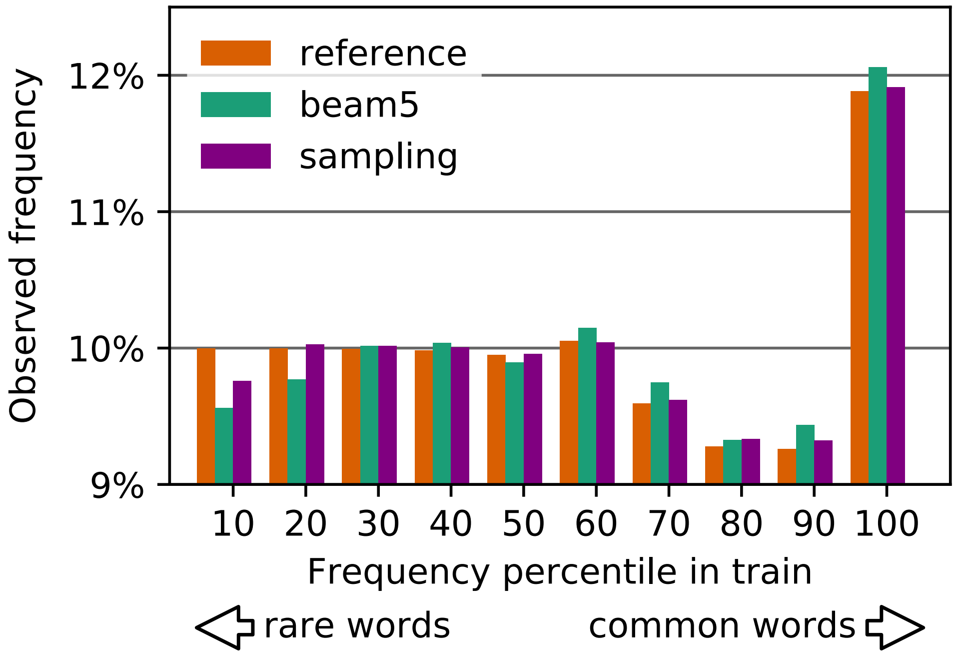
Histogram of unigram frequencies

Beam under-estimates the rarest words

Beam over-estimates frequent words.

We should expect this!

Necessary matching conditions—Token Level



(WMT14 En-Fr)

Histogram of unigram frequencies

Beam under-estimates the rarest words

Beam over-estimates frequent words.
We should expect this!

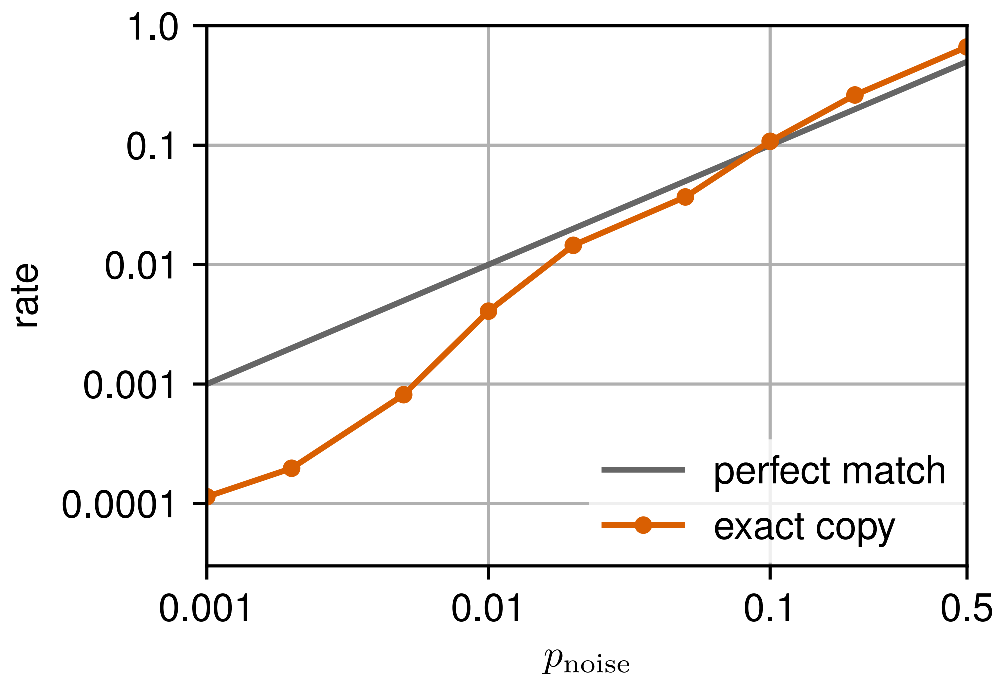
Sampling mostly matches the reference data distribution

Necessary matching conditions—Sequence Level

Synthetic experiment:

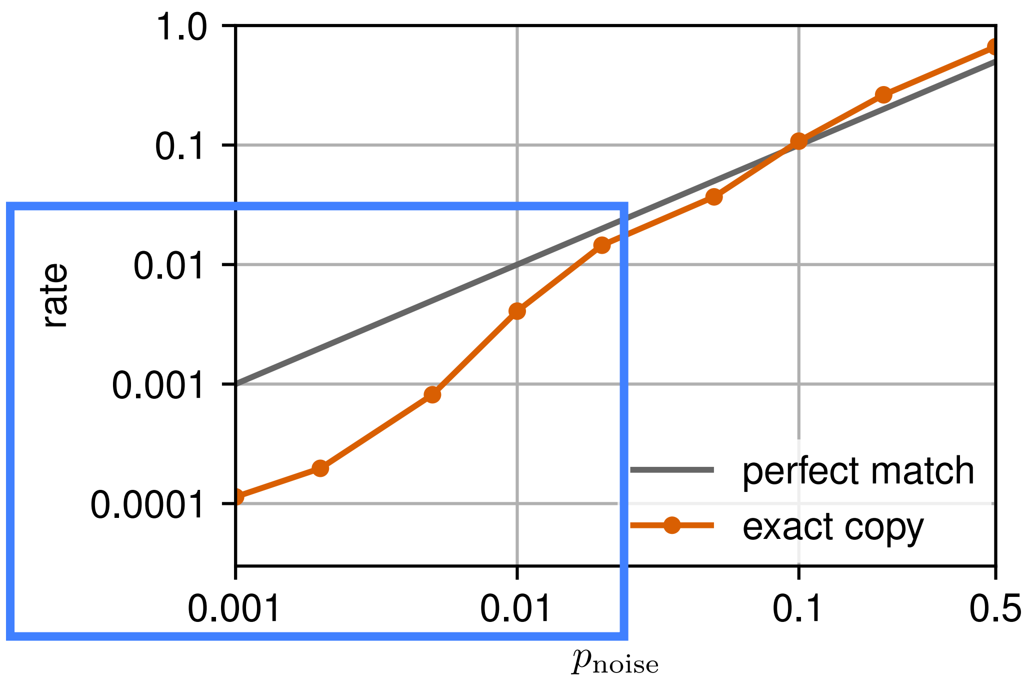
- Retrain model on news subset of WMT, which does not contain copies
- Artificially introduce copies in the training data with probability p_{noise}
- Measure rate of copies among sampled hypotheses

Necessary matching conditions—Sequence Level



(WMT17 En-De)

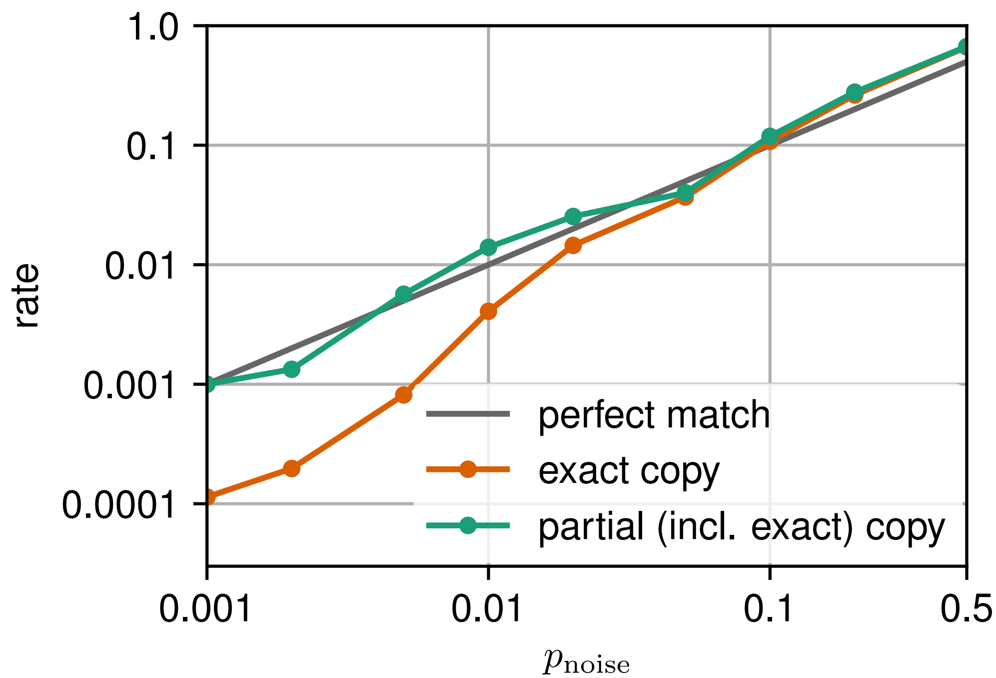
Necessary matching conditions—Sequence Level



Model under-estimates copies at a sequence level

(WMT17 En-De)

Necessary matching conditions—Sequence Level



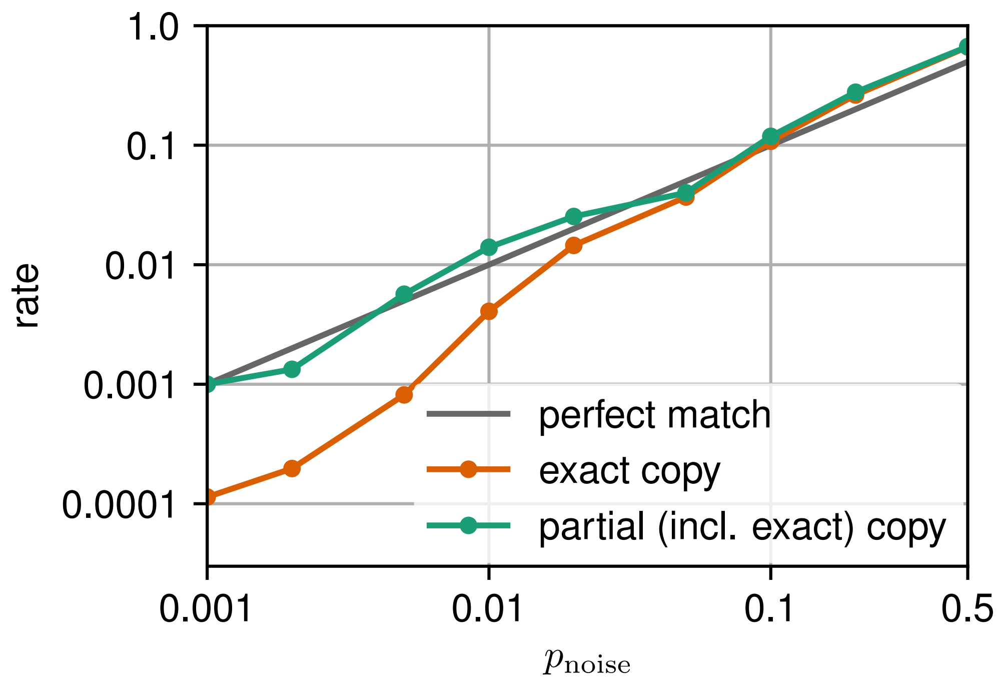
(WMT17 En-De)

p_{noise} controls rate of **exact copies**

Partial copies* do not appear in training, yet…

* A partial copy has a unigram overlap of >= 50% with the source

.\ I Necessary matching conditions—Sequence Level



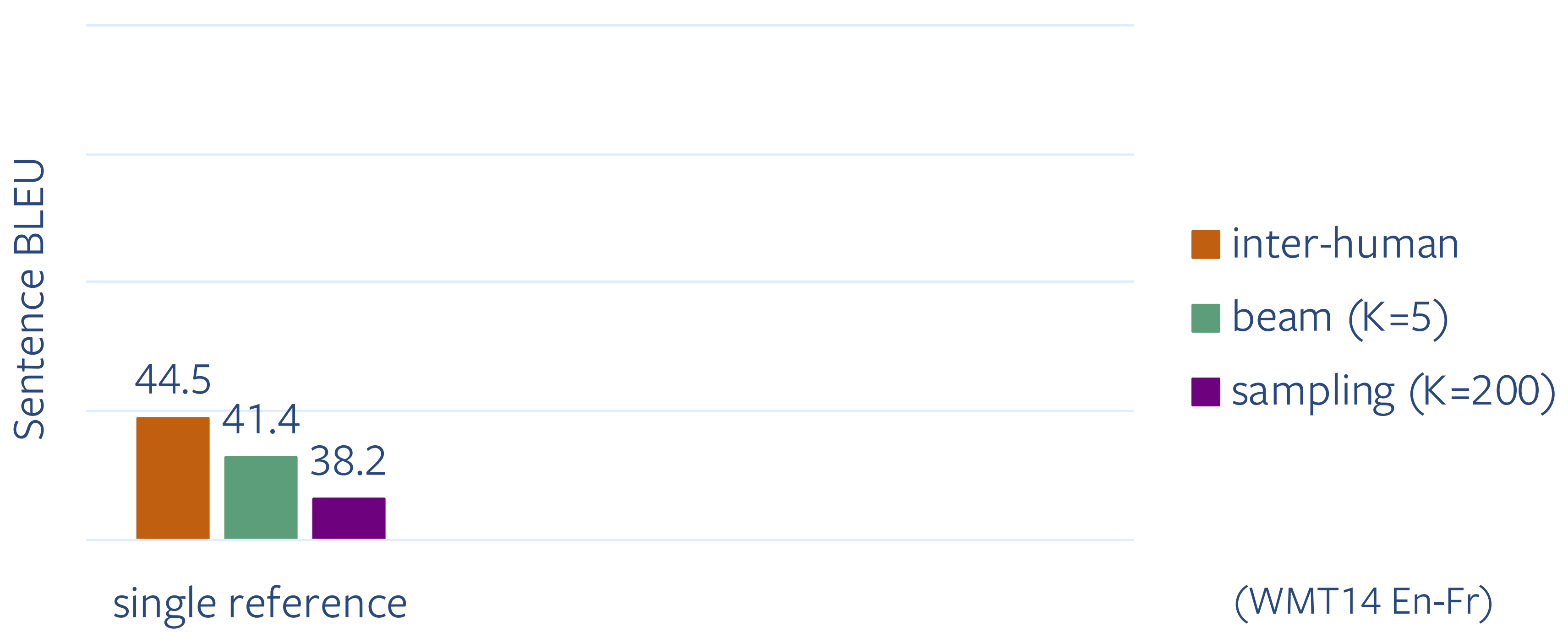
(WMT17 En-De)

The model smears probability mass in hypothesis space!

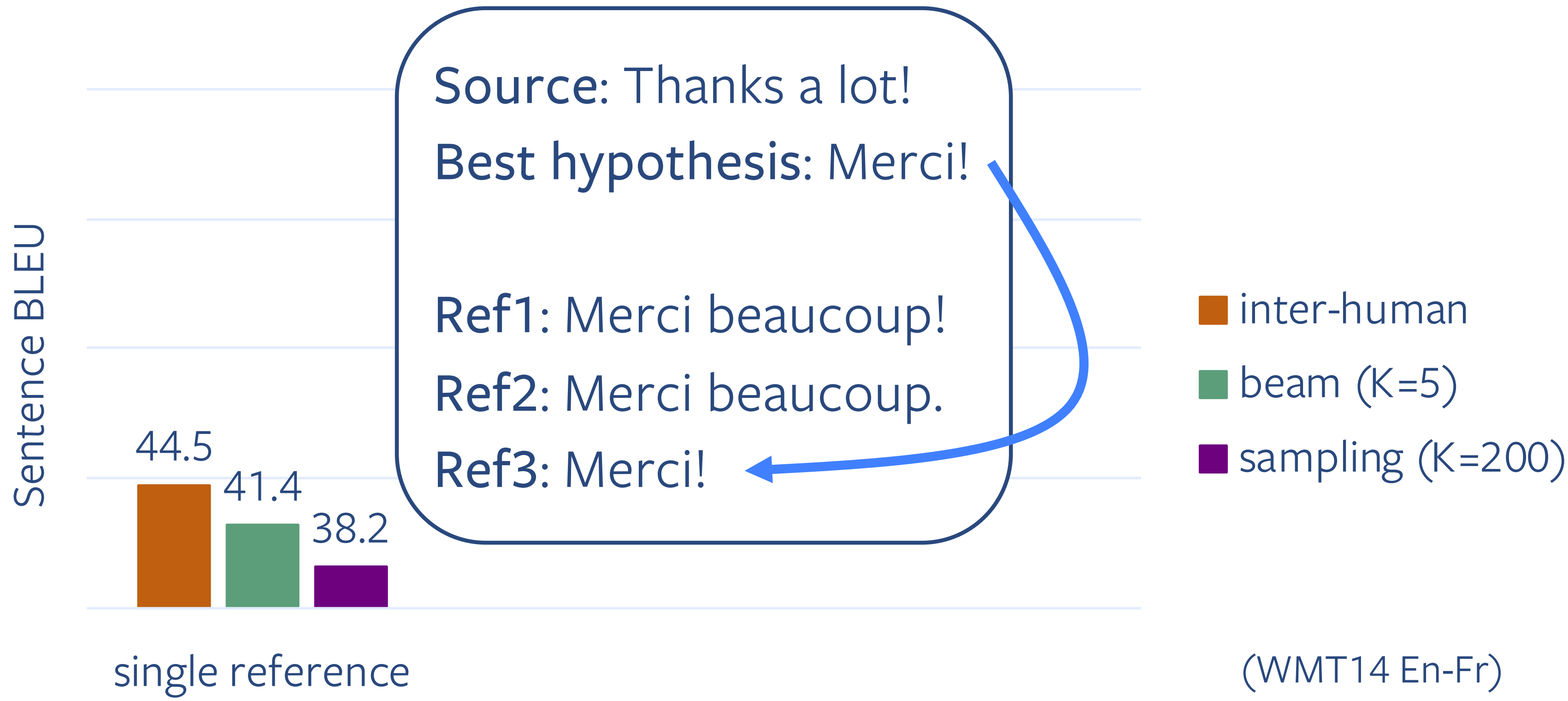
Necessary matching conditions—with Mult. References

Question: Can we use BLEU to assess how well the model distribution matches the data distribution?

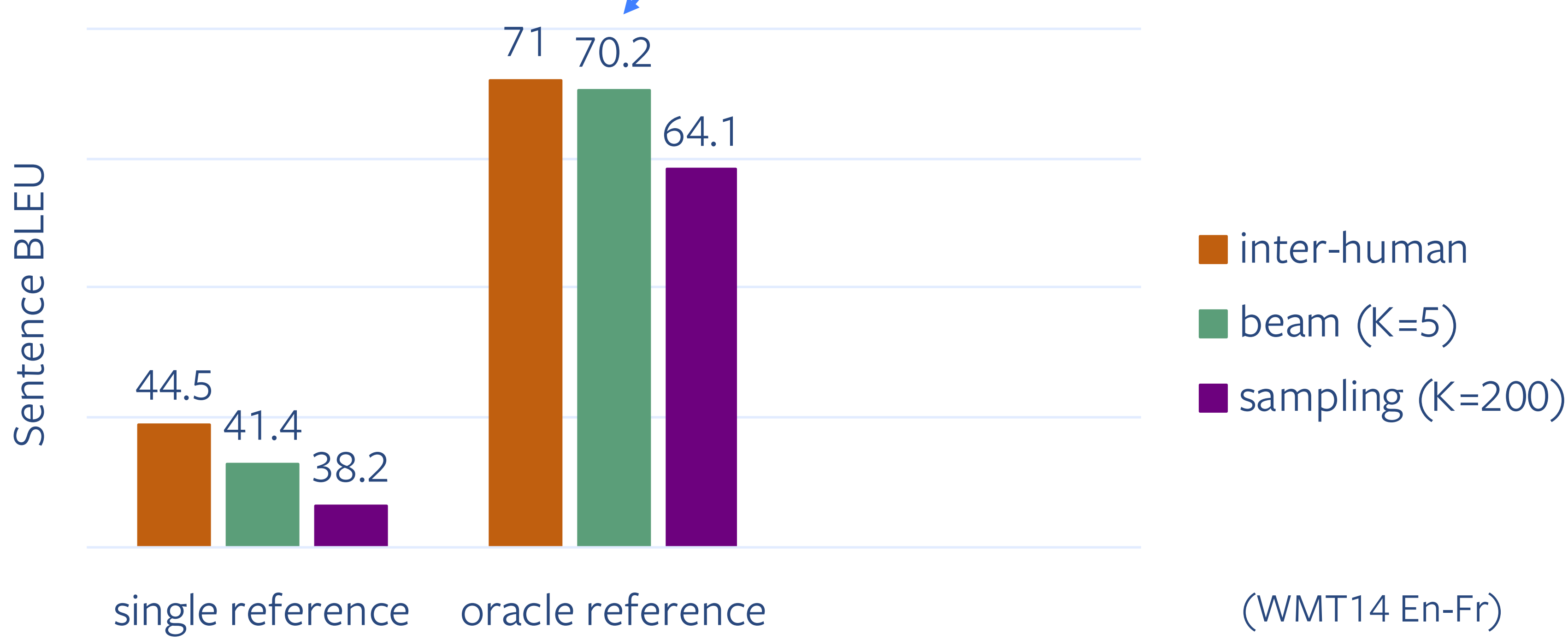
- Collect 10 additional reference translations from distinct human translators
- 500 sentences (En-Fr) and 500 sentences (En-De)
- 10K sentences total
- Available at: github.com/facebookresearch/analyzing-uncertainty-nmt


Sentence BLEU
44.5
41.4
38.2
single reference
inter-human
beam (K=5)
sampling (K=200)
(WMT14 En-Fr)

oracle reference: BLEU w.r.t. best matching reference

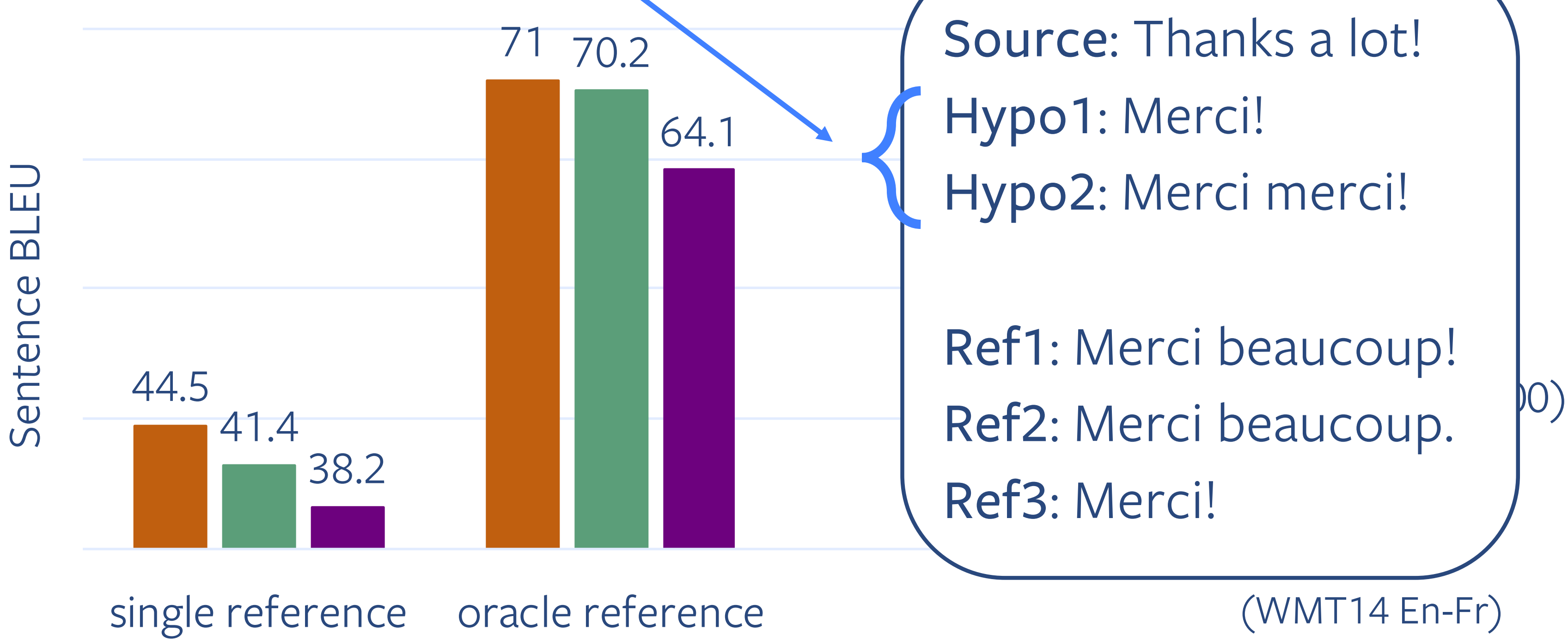


The best beam hypothesis is very close to a reference

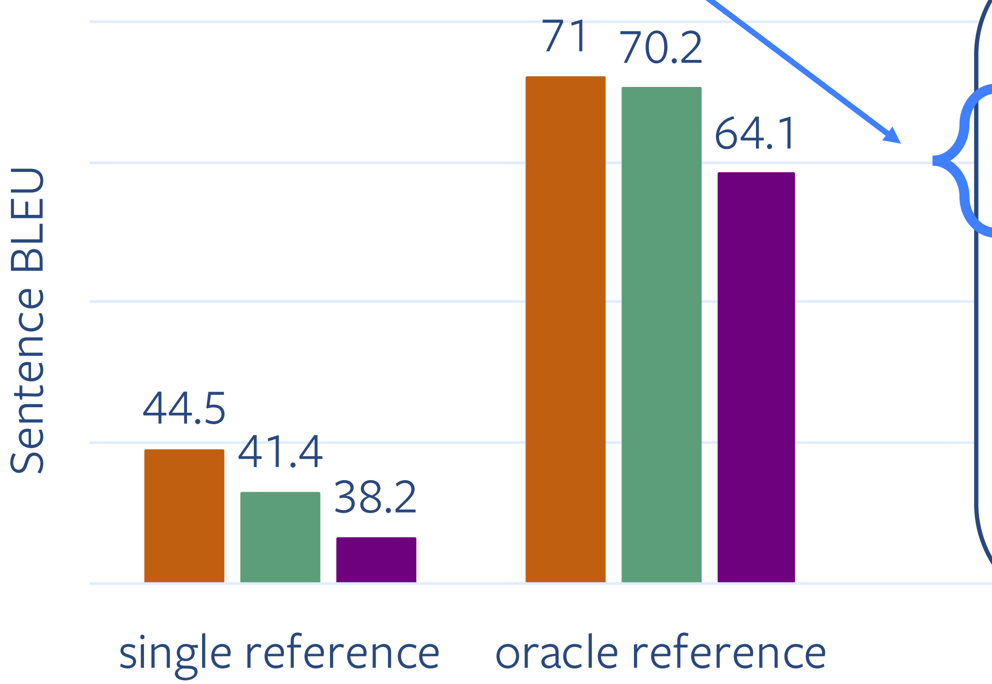


average oracle:

average oracle reference BLEU over top-K hypotheses



average oracle:
average oracle reference BLEU over top-K hypotheses



Source: Thanks a lot!

Hypo1: Merci!

Hypo2: Merci merci!

Ref1: Merci beaucoup!

Ref2: Merci beaucoup.

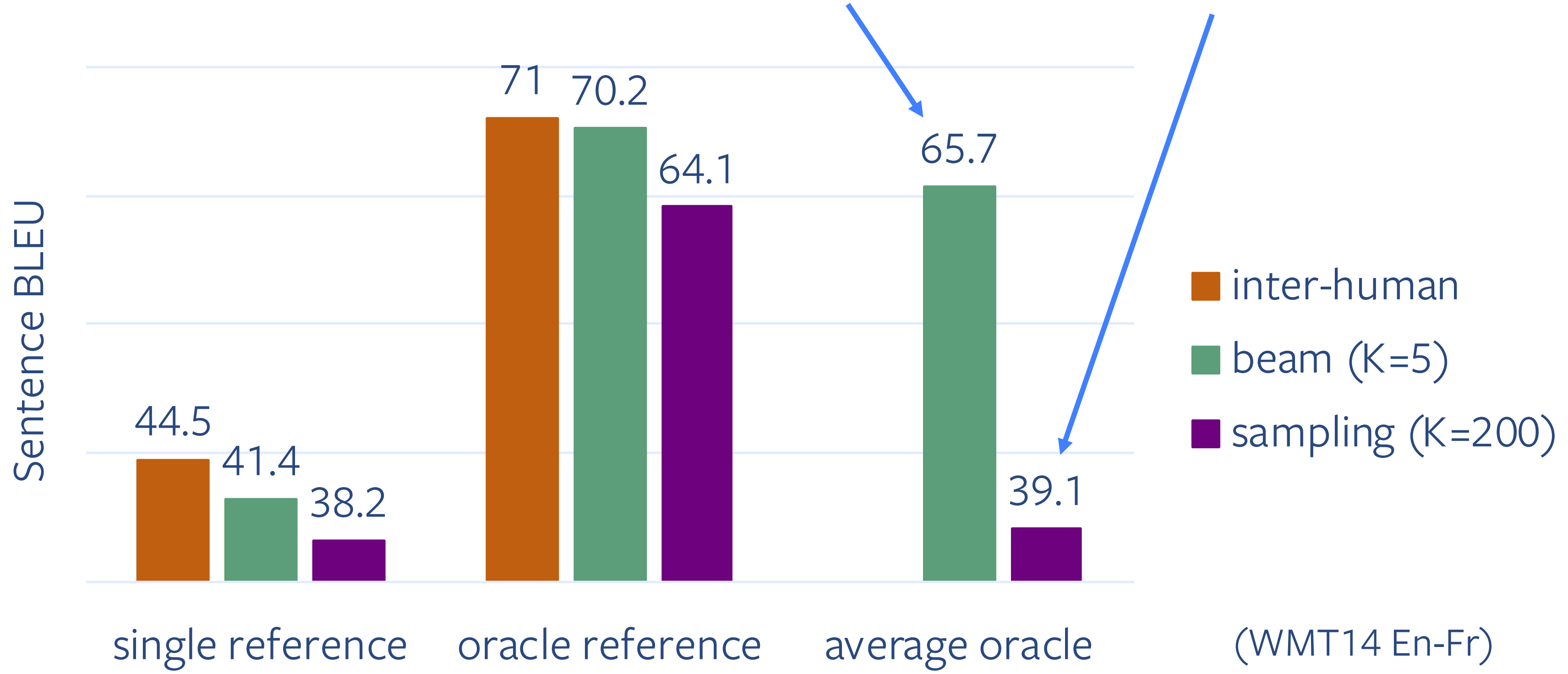
Ref3: Merci!

0)

(WMT14 En-Fr)

Most beam hypotheses are close to a reference

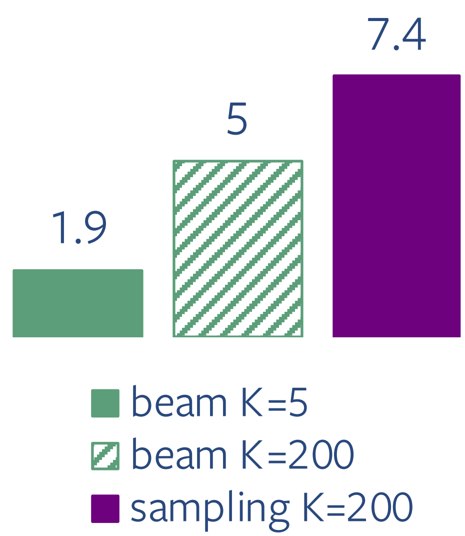
Most sampled hypotheses are far from a reference



(WMT14 En-Fr)

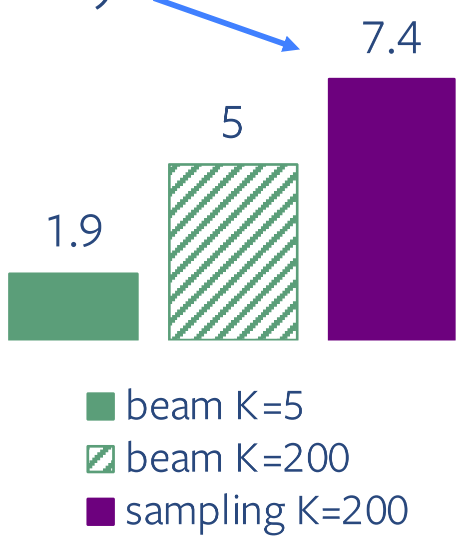
.\ I

 # refs covered: number of distinct references (out of 10) matched to at least one hypothesis



 # refs covered: number of distinct references (out of 10) matched to at least one hypothesis

Sampling covers more hypotheses (is more diverse) than beam search

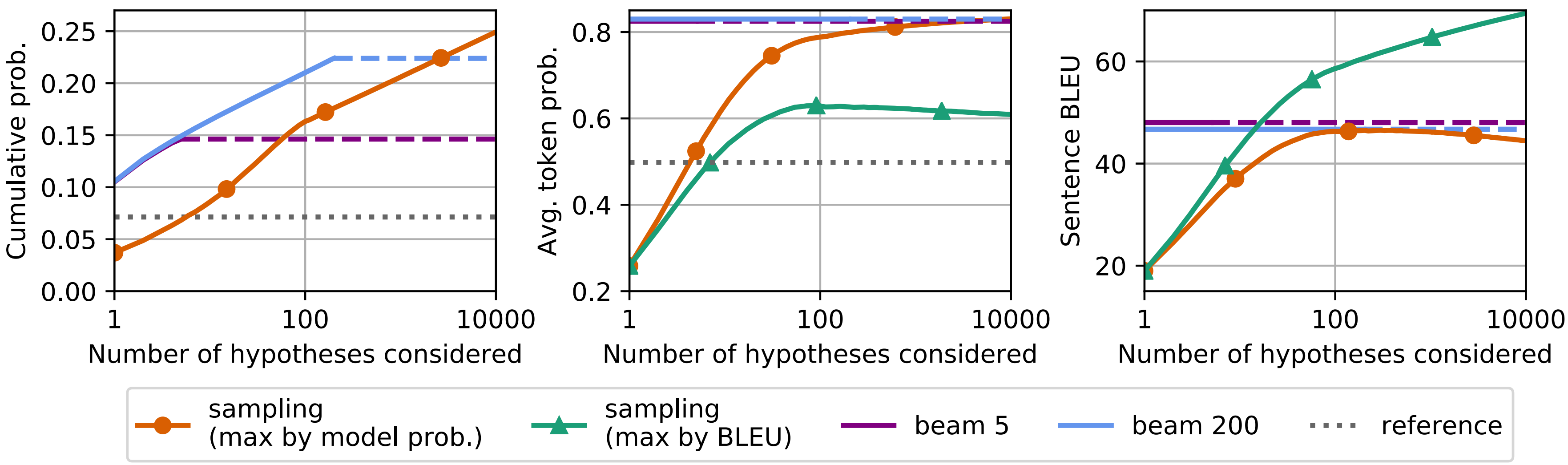


Conclusion

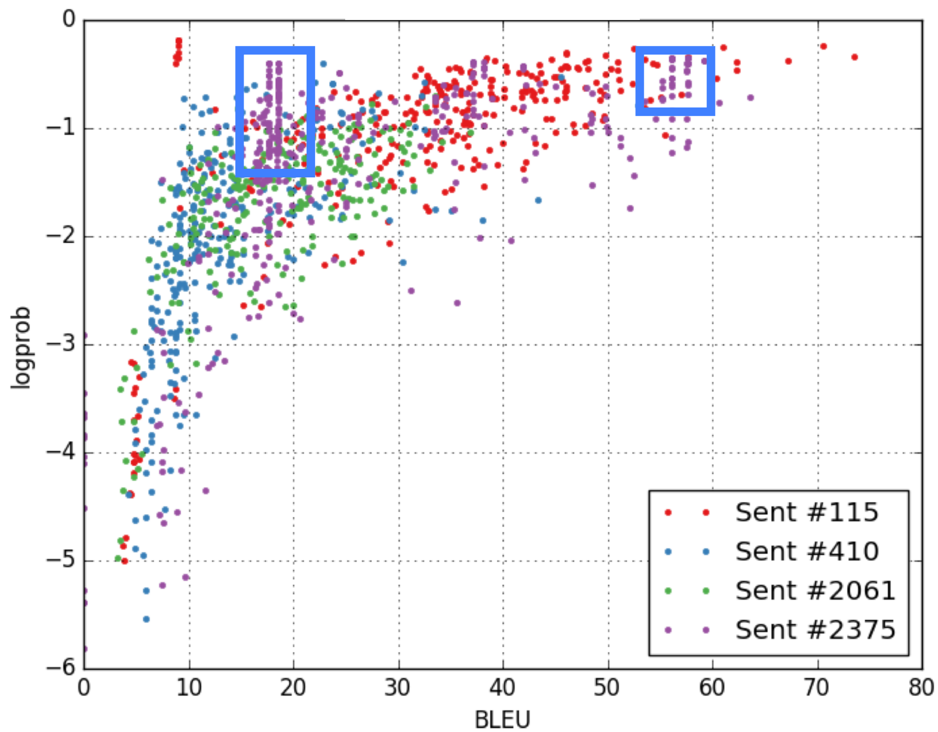
Poster #163

- NMT **models capture uncertainty** in their output distributions
- Beam search is **efficient** and **effective**, but prefers frequent words
- Degradation with large beams is mostly due to **copying**, but this can be mitigated by **filtering**
- Models are well calibrated at the token level, but **smear probability mass** at the sequence level
- Smearing may be responsible for **lack of diversity** in beam search outputs

Dataset link: github.com/facebookresearch/analyzing-uncertainty-nmt



Cumulative prob.
0.25
0.20
0.15
0.10
0.05
0.00
1
100
10000
Number of hypotheses considered
Avg. token prob.
0.8
0.6
0.4
0.2
1
100
10000
Number of hypotheses considered
Sentence BLEU
60
40
20
1
100
10000
Number of hypotheses considered
sampling (max by model prob.)
sampling (max by BLEU)
beam 5
beam 200
reference



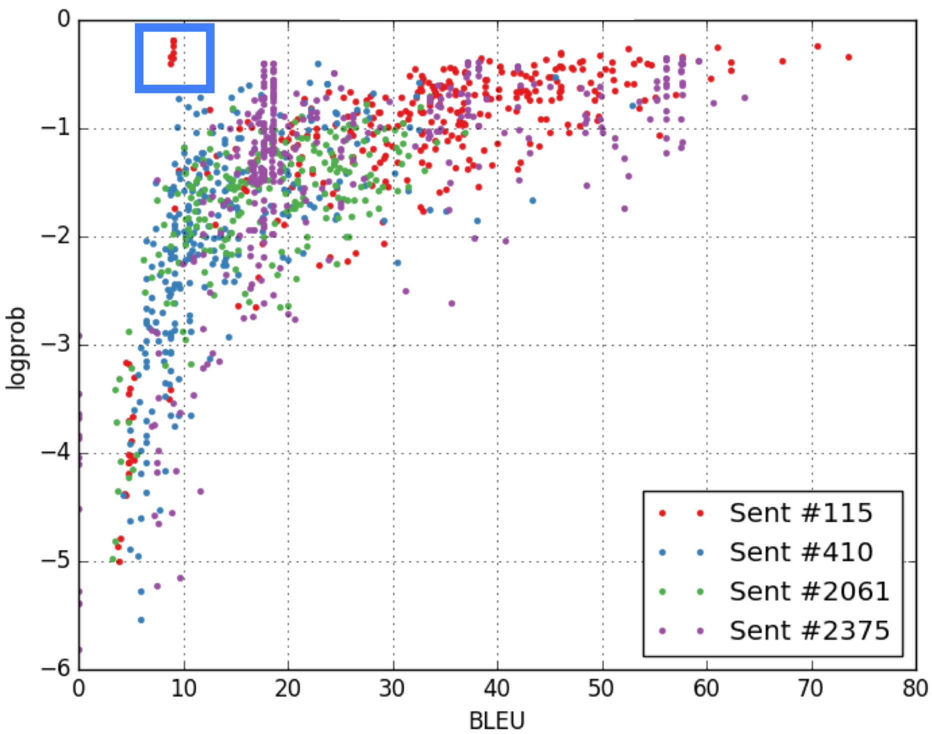
Source: Should this election be decided two months after we stopped voting?

Ref: Cette élection devrait-elle être décidée deux mois après que le vote est terminé?

Low BLEU: Ce choix devrait-il être décidé deux mois après la fin du vote?

High BLEU: Cette élection devrait-elle être décidée deux mois après l'arrêt du scrutin?

.\ I

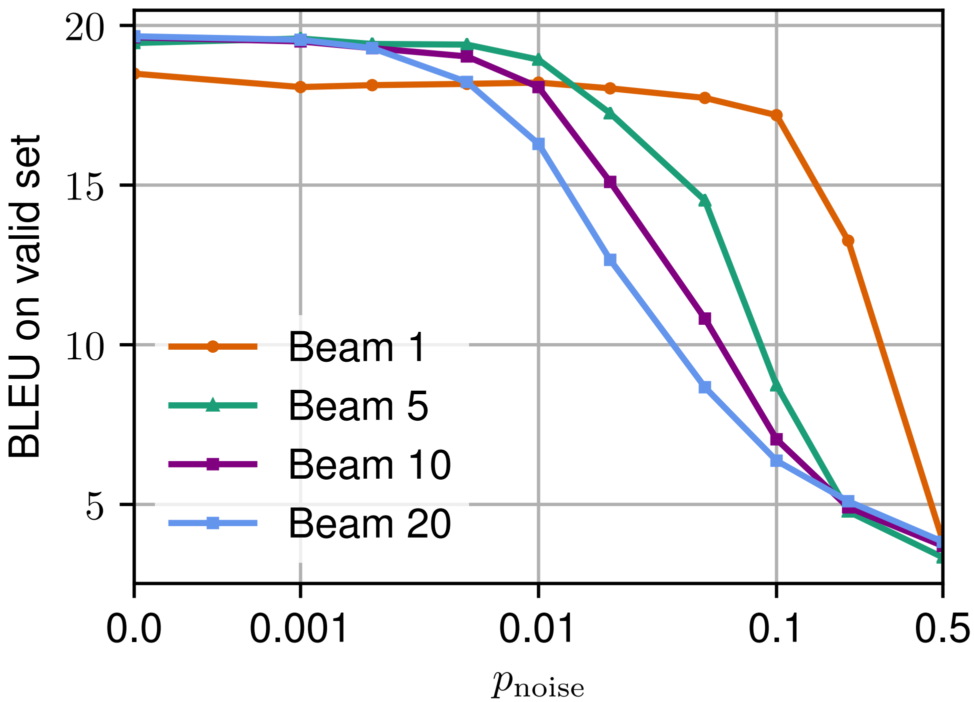


Source: The first nine episodes of Sheriff <unk> 's Wild West will be available (...)

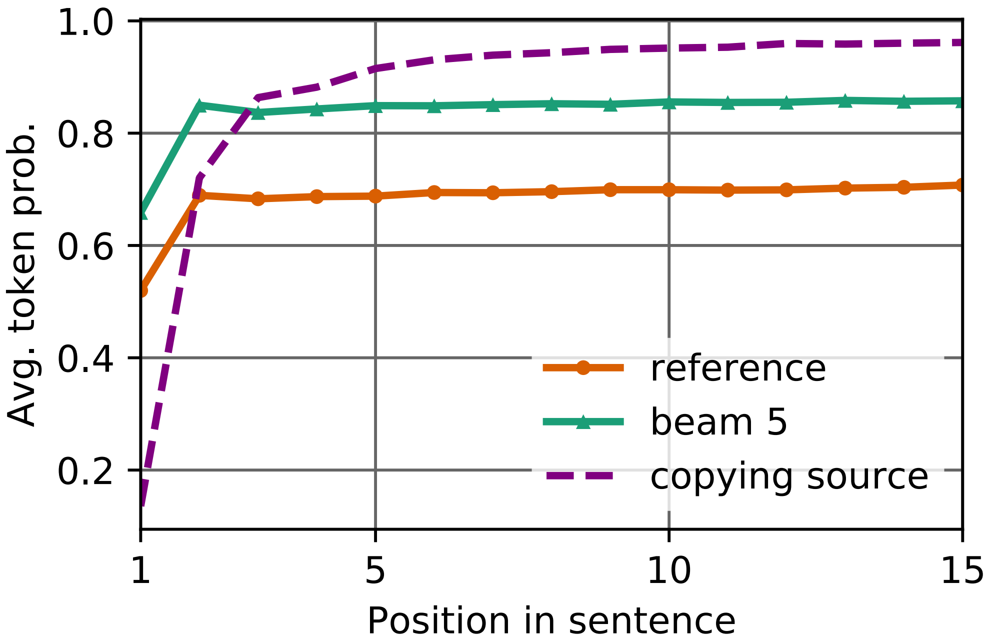
Ref: Les neuf premiers épisodes de <unk> <unk> s Wild West seront disponibles (...)

Low BLEU: The first <unk> <unk> of <unk> <unk> s Wild West will be available (...)

Output is a "copy" in the source language!

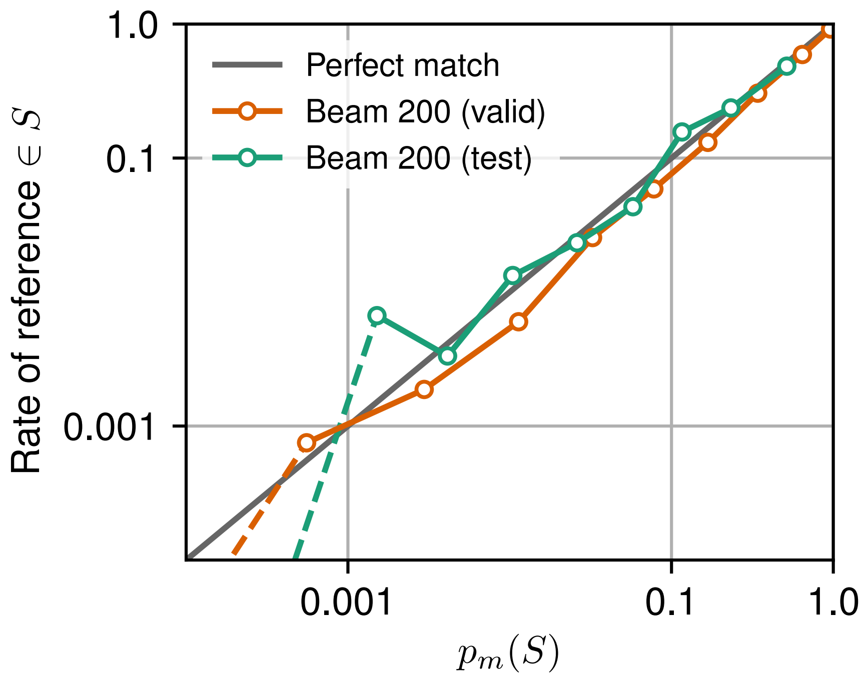


- Train model on news subset of WMT, which does not contain copies
- Artificially introduce copies in training data with p_{noise}
- Small amounts of copy noise lead to a large drop in BLEU for beam k=20



- During decoding we pay a large penalty for the first copied word
- Subsequently, there is little uncertainty—just continue copying
- Large beams increase chance of reaching the “copy” mode

Necessary matching conditions—Sequence Level



Set-level calibration

(Guo et al., 2017; Kuleshov & Liang, 2015)

$$\mathop{\mathbb{E}}_{x \sim p_d}[\mathbb{I}\{x \in S\}] = p_m(S)$$

- x-axis: model score of 200 beam hypos
- y-axis: rate at which reference translation is among beam hypos

.\ | Necessary matching conditions—with Mult. References

	beam $k = 5$	beam $k = 200$	sampling $k = 200$
Prob. covered	4.7%	11.1%	6.7%
Sentence BLEU			
single reference	41.4	36.2	38.2
oracle reference	70.2	61.0	64.1
average oracle	65.7	56.4	39.1
- # refs covered	1.9	5.0	7.4
Corpus BLEU (`multi-bleu.pl`)			
single reference	41.6	33.5	36.9
10 references	81.5	65.8	72.8