University of Zurich

Computational Linguistics

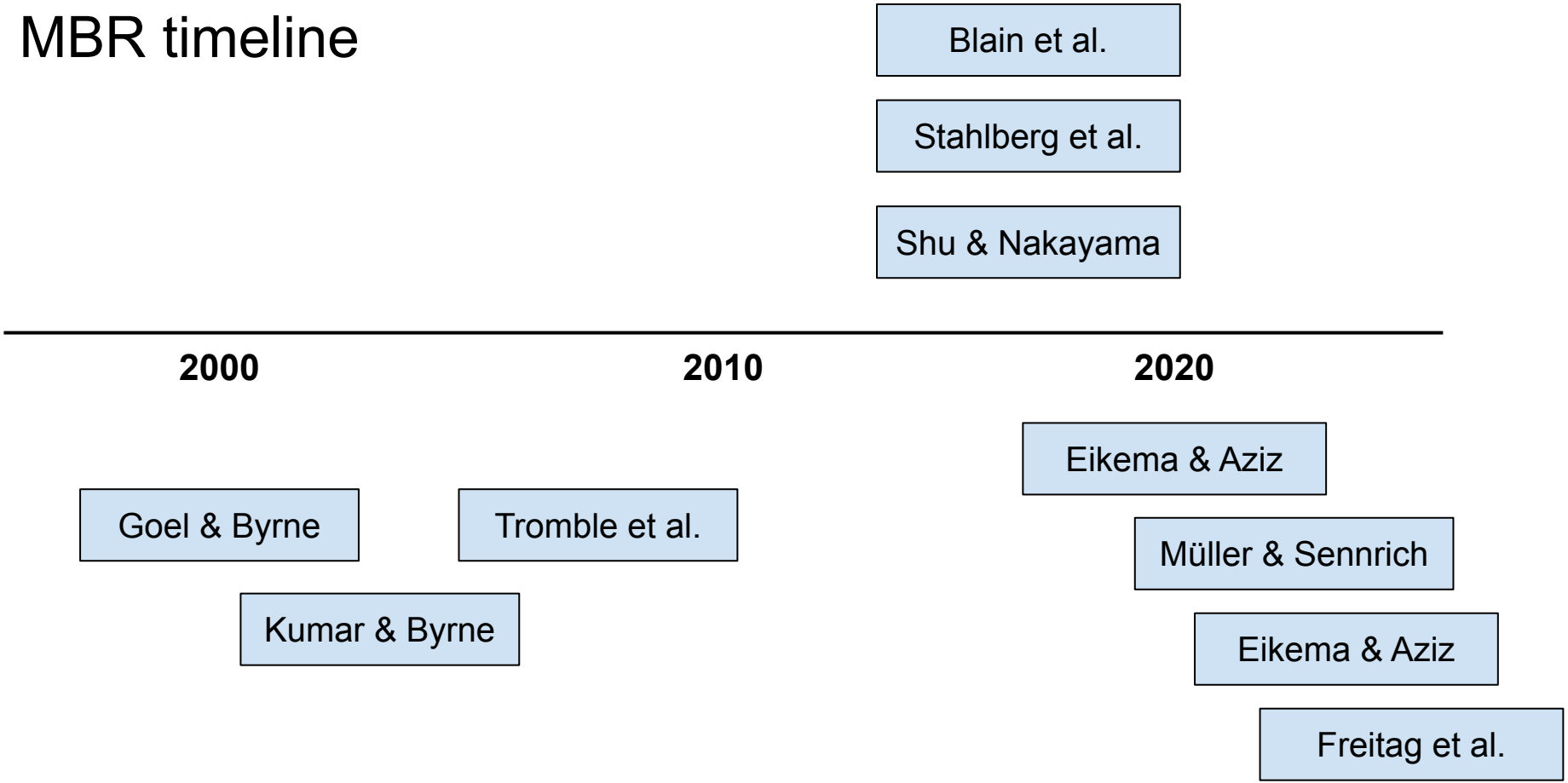
Recent works on Minimum Bayes Risk decoding in machine translation

Mathias Müller

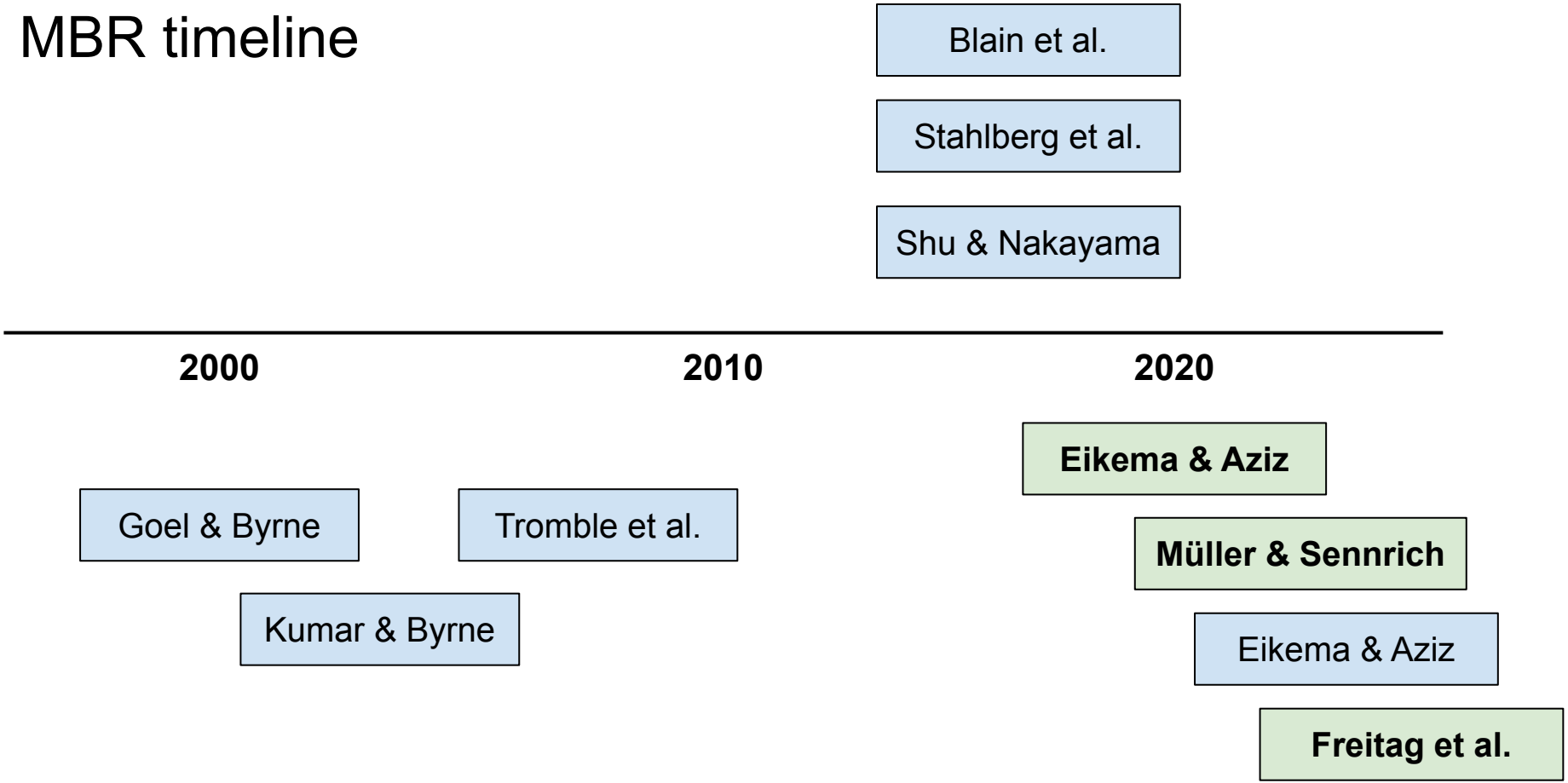
Beam search

Minimum Bayes Risk decoding

MBR timeline



MBR timeline



Reviving MBR in MT research

No-nonsense explanation of MBR algorithm

Risks and opportunities of MBR

Combining MBR with learned metrics

Reviving MBR in MT research

Reviving MBR in MT research

No-nonsense explanation of MBR algorithm

Risks and opportunities of MBR

Combining MBR with learned metrics

Why bring MBR decoding back in MT?

1) NMT has well-known shortcomings with no clear way forward
2) There is evidence that beam search is at least partially to blame

Well-known shortcomings of NMT

1) Length bias
2) Low robustness to noise in the training data

Length bias

	Average # tokens
Reference	11.91
Beam 5	11.61

Müller and Sennrich (2021)
Danish-Esperanto

Length bias

	Average # tokens
Reference	11.91
Beam 5	11.61

Müller and Sennrich (2021)
Danish-Esperanto

Low robustness to copy noise

AR ⟶ DE

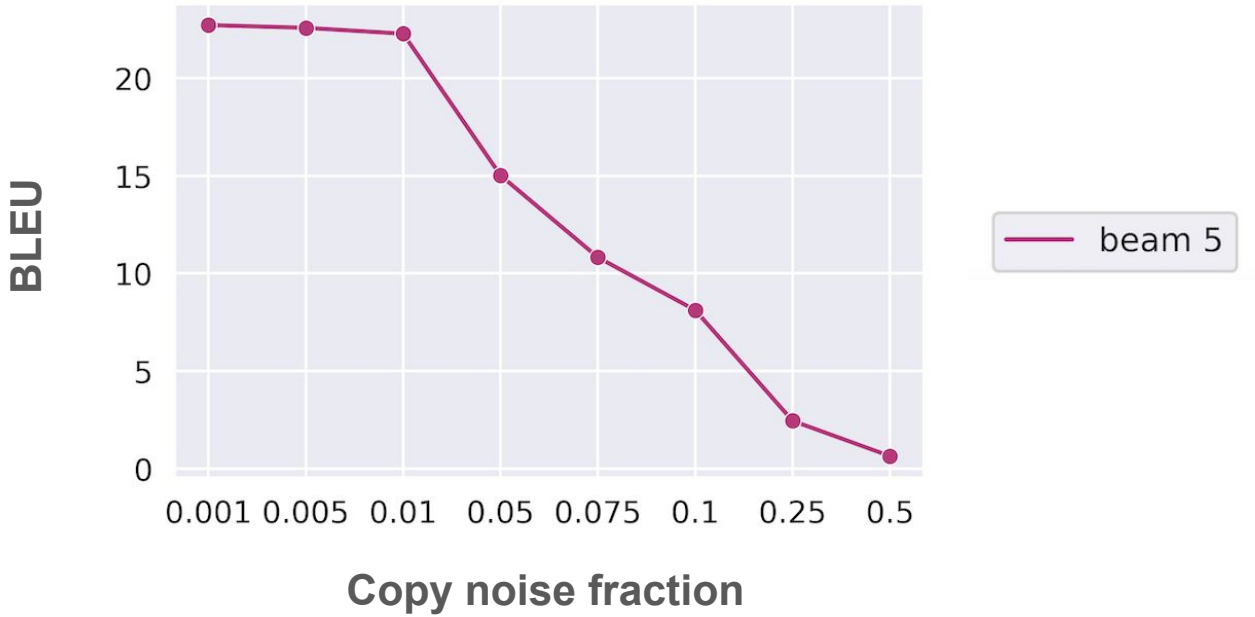
Source

أسئلة حول خطط الفيفا لإقامة كأس العالم كل عامين

Target

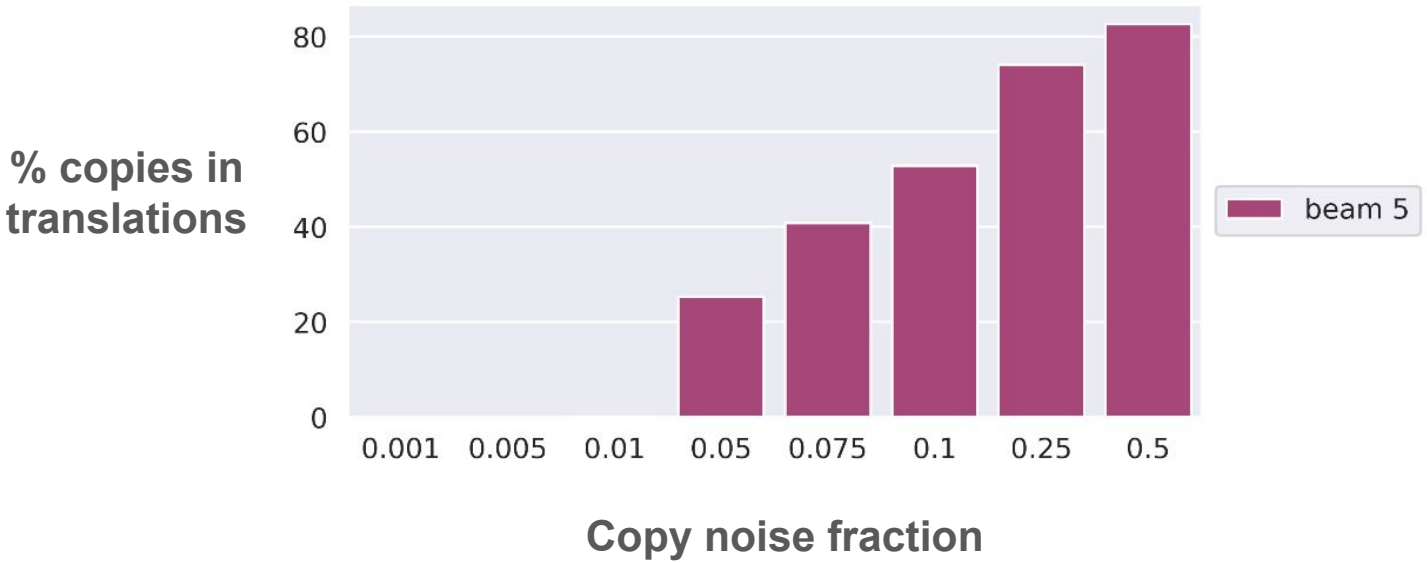
أسئلة حول خطط الفيفا لإقامة كأس العالم كل عامين

Low robustness to copy noise



Müller and Sennrich (2021)
Arabic-German

Low robustness to copy noise



Müller and Sennrich (2021)
Arabic-German

Why bring MBR decoding back in MT?

1) NMT has well-known shortcomings with no clear way forward
2) **There is evidence that beam search is at least partially to blame**

Is MAP Decoding All You Need?
The Inadequacy of the Mode in Neural Machine Translation

Bryan Eikema
University of Amsterdam
b.eikema@uva.nl

Wilker Aziz
University of Amsterdam
w.aziz@uva.nl

Main arguments of Eikema & Aziz (2020)

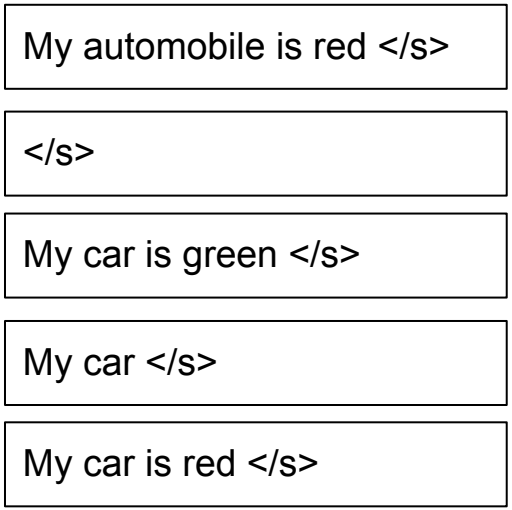
1) Beam search connects seemingly unrelated failure cases
2) Beam search is at odds with current training methods
3) Model samples fit the data well

A medical joke

Decoding objective

Training objective

Spread of probability mass

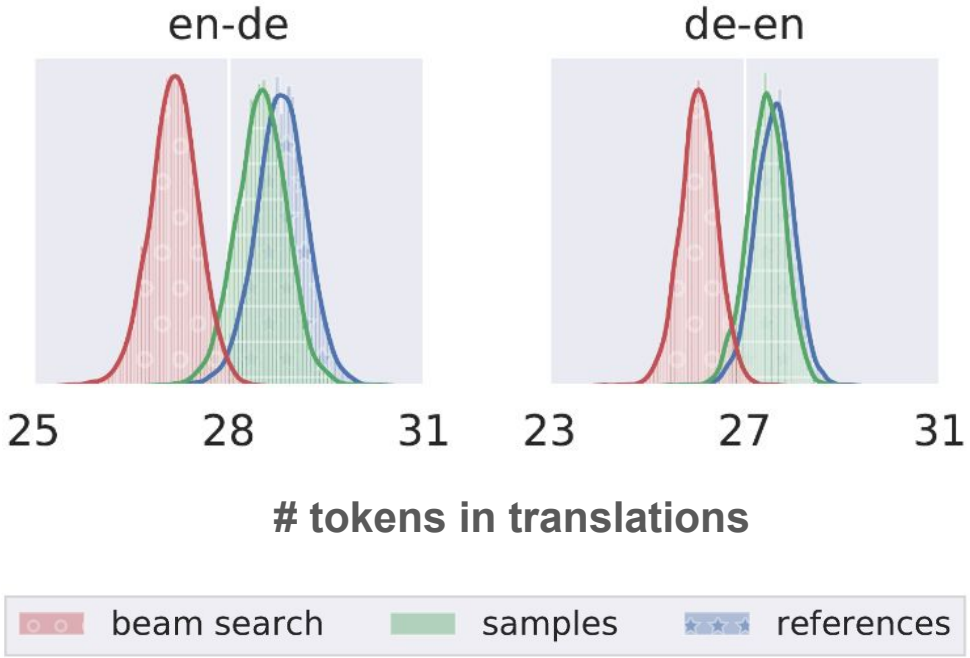


expectation

reality

Eikema and Aziz (2021)

Models fit the data well



Eikema and Aziz (2020)

Summary

- NMT has some well-known and well-documented failure cases. Examples: length bias and sensitivity to copy noise.

- Beam search is at least partially to blame. One kind of evidence for this: Judging from model samples, models fit the data well.

- Eikema & Aziz (2020) propose MBR decoding based on samples as an alternative to beam search.

No-nonsense explanation of MBR algorithm

Reviving MBR in MT research

No-nonsense explanation of MBR algorithm

Risks and opportunities of MBR

Combining MBR with learned metrics

Minimum Bayes Risk (MBR) decoding ≈ choose a translation from a set of samples

Generate a pool of samples $\mathcal{S}$:

$$\mathcal{S} = (s_1, ..., s_n) \sim p(y|x, \theta)$$

$\mathcal{S} =$

The quick brown fox

The quick

The fast brown fox

For every sample s_i in the pool compute its utility:

$$\frac{1}{n}\sum_{s_j=1}^{n} u(s_i, s_j)$$

$$\mathcal{S} =$$

The quick brown fox	1.37
The quick	0.04
The fast brown fox	0.68

For every sample s_i in the pool compute its utility:

$$\frac{1}{n}\sum_{s_j=1}^{n} u(s_i, s_j)$$

$\mathcal{S} =$

The quick brown fox	1.37
The quick	0.04
The fast brown fox	0.68

$$y^\star = \underset{s_i \in \mathcal{S}}{\operatorname{argmax}} \frac{1}{n} \sum_{s_j=1}^{n} u(s_i, s_j)$$

$$\mathcal{S} =$$

The quick brown fox	1.37
The quick	0.04
The fast brown fox	0.68

Summary

- Sampling-based MBR decoding roughly works as follows:
  - Draw pool of samples from model
  - Compute utility of each sample
  - Select sample with highest utility as final translation

- Important hyperparameters: size of sample pool, utility function

Risks and opportunities of MBR

Reviving MBR in MT research

No-nonsense explanation of MBR algorithm

Risks and opportunities of MBR

Combining MBR with learned metrics

Understanding the Properties of Minimum Bayes Risk Decoding in Neural Machine Translation

Mathias Müller[1] and **Rico Sennrich**[1,2]
[1]Department of Computational Linguistics, University of Zurich
[2]School of Informatics, University of Edinburgh

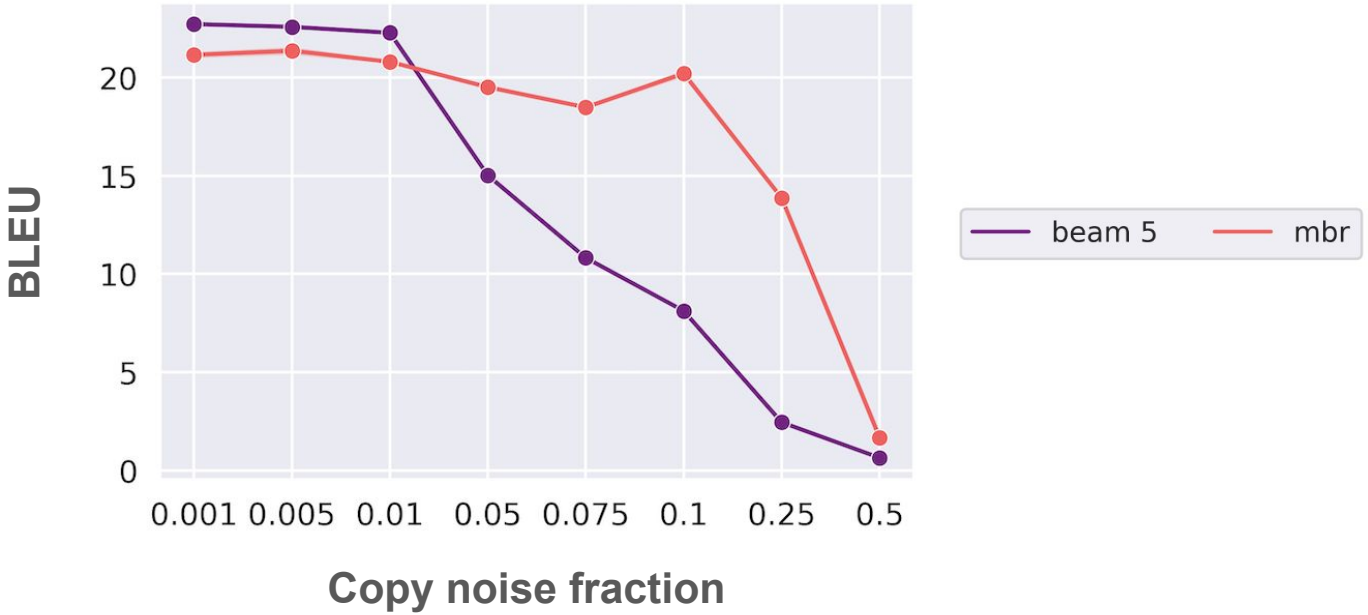
Length bias

	Average # tokens
Reference	11.91
Beam 5	11.61
Sample	11.73

MBR + BLEU	11.51
MBR + METEOR	12.23
MBR + CHRF	12.50

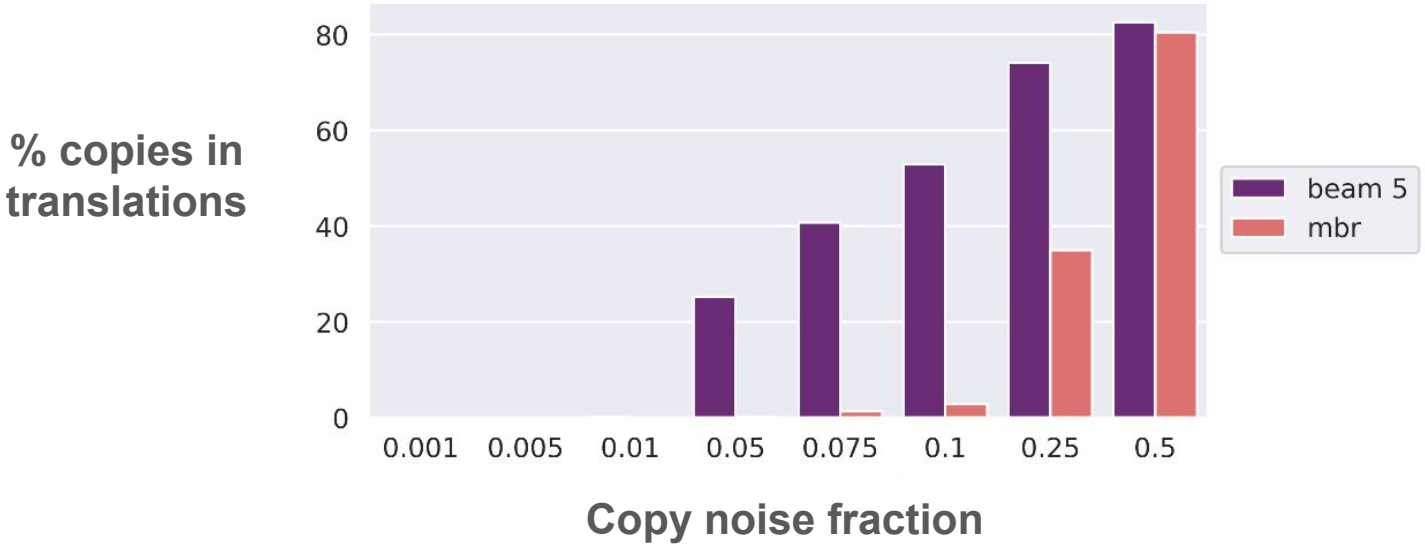
Müller and Sennrich (2021)
Danish-Esperanto

Robustness to copy noise



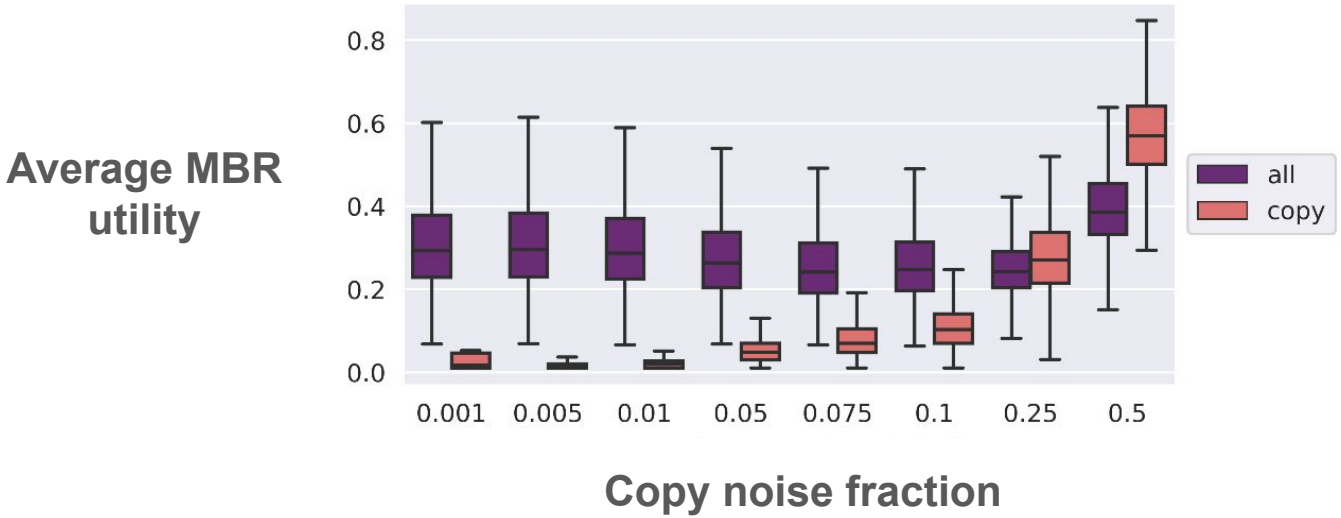
Müller and Sennrich (2021)
Arabic-German

Robustness to copy noise



Müller and Sennrich (2021)
Arabic-German

Robustness to copy noise



Müller and Sennrich (2021)
Arabic-German

Summary

- Minimum Bayes Risk != Minimum Bias Risk

- MBR increases robustness to copies in the training data

Combining MBR with learned metrics

Reviving MBR in MT research

No-nonsense explanation of MBR algorithm

Risks and opportunities of MBR

Combining MBR with learned metrics

For future work, we are interested in exploring more sophisticated similarity metrics to be used as utility functions, including trainable metrics such as COMET (Rei et al., 2020), and investigating how these utility functions affect the overall quality and biases of translations.

Müller & Sennrich (2021)

Learned metrics clearly superior to string metrics

system pairs:	3344
COMET	83.4
COMET-source	83.2
Prism	80.6
BLEURT	80.0
ESIM	78.7
BERTScore	78.3
ChrF	75.6
TER	75.6
CharacTER	74.9
BLEU	74.6
Prism-source	73.4

Kocmi et al. (2021)

Minimum Bayes Risk Decoding with Neural Metrics of Translation Quality

Markus Freitag, David Grangier, Qijun Tan, Bowen Liang
Google Research
{freitag, grangier, qijuntan, bowenl}@google.com

Learned metric as utility function

	Human evaluation score (MQM ↓)
Human translation	0.388
Beam search	2.030
MBR + BLEU	1.855
MBR + CHRF	2.139
MBR + BLEURT	1.571

Freitag et al. (2021)
English-German

Summary

- We have good evidence that learned metrics are better predictors of human judgement.

- Using learned metrics as utility functions for MBR seems promising.

Outlook

- More emphasis on role of decoding

- Desirable: general awareness that the decoding algorithm is a replaceable component of a system

Thank you!

Relevant papers

Bibliography

Eikema & Aziz (2020) https://arxiv.org/abs/2005.10283

Eikema & Aziz (2021) https://arxiv.org/abs/2108.04718

Müller et al. (2020) https://arxiv.org/abs/1911.03109

Müller & Sennrich (2021) https://arxiv.org/abs/2105.08504

Leblond et al. (2021) https://arxiv.org/abs/2104.05336

Kocmi et al. (2021) https://arxiv.org/abs/2107.10821

Freitag et a. (2021) https://arxiv.org/pdf/2111.09388.pdf

Adams et al. (2021) https://www.nature.com/articles/s41586-021-03380-y

Bonus material (things I would have loved to talk about as well!)

More results comparing beam search and MBR

Low domain robustness

DE → EN

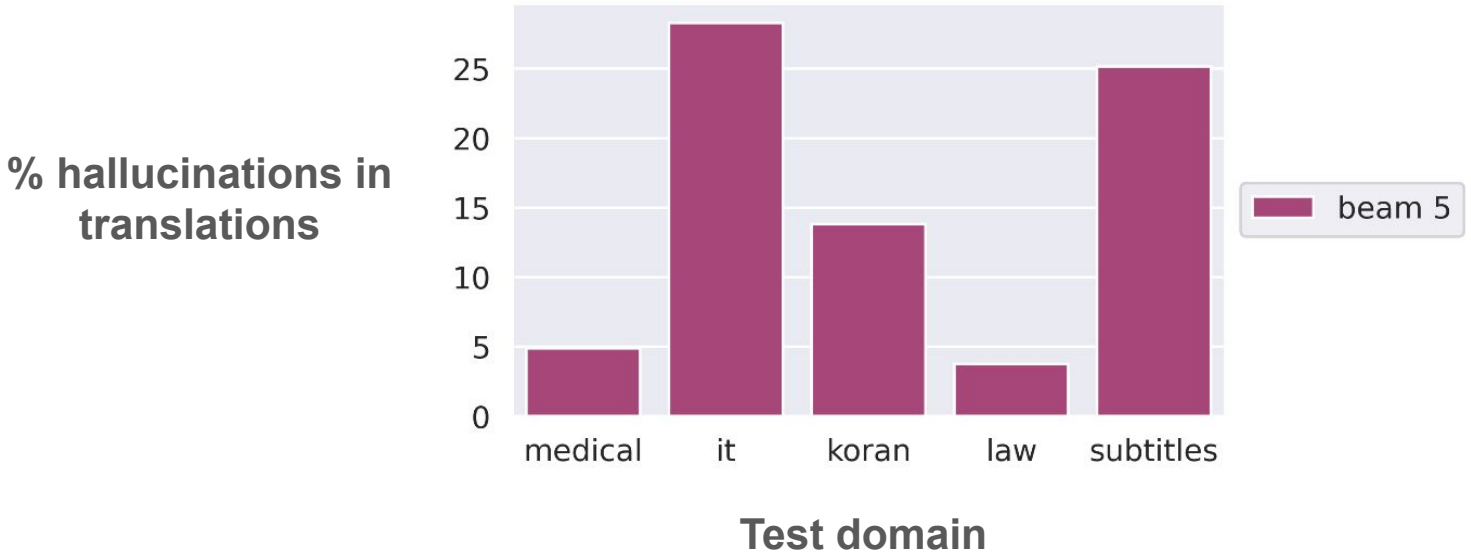
Training domain	Test domains
medical	law
	it
	koran
	subtitles

Hallucination

Input	Reference	System
Aber geh subtil dabei vor.	But be subtle about it.	Pharmacokinetic parameters are not significantly affected in patients with renal impairment (see section 5.2).

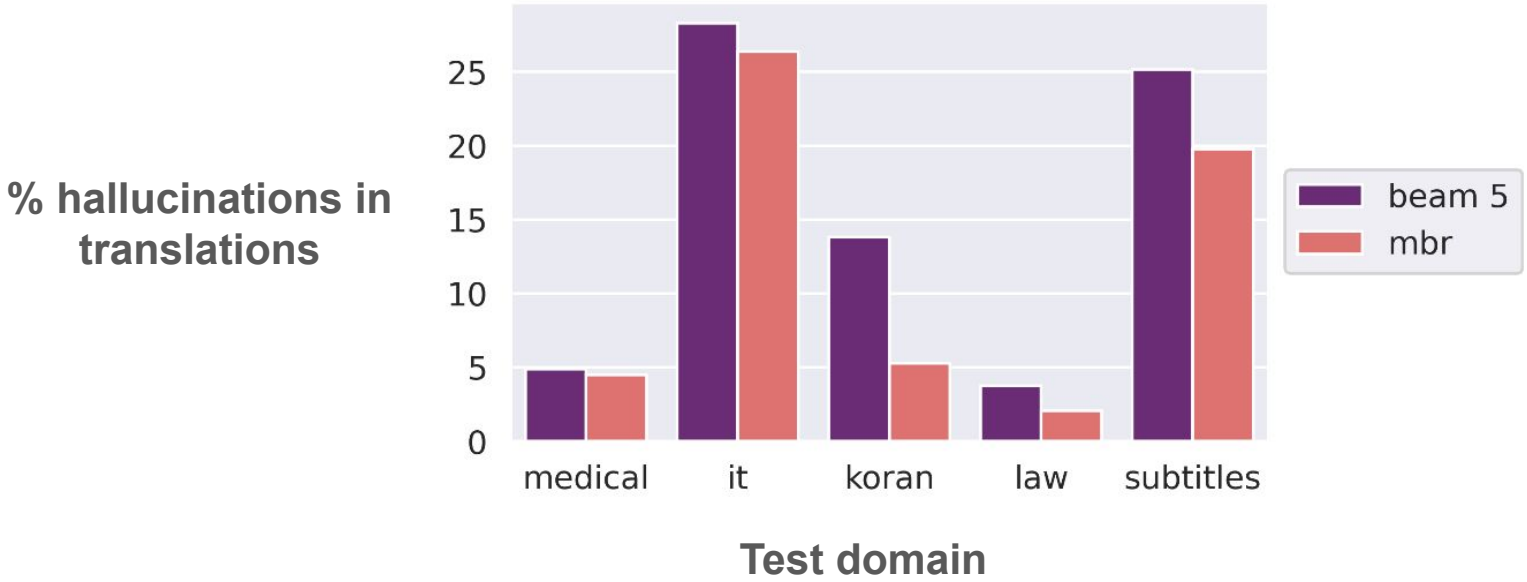
Müller et al. (2020)
German-English

Low domain robustness



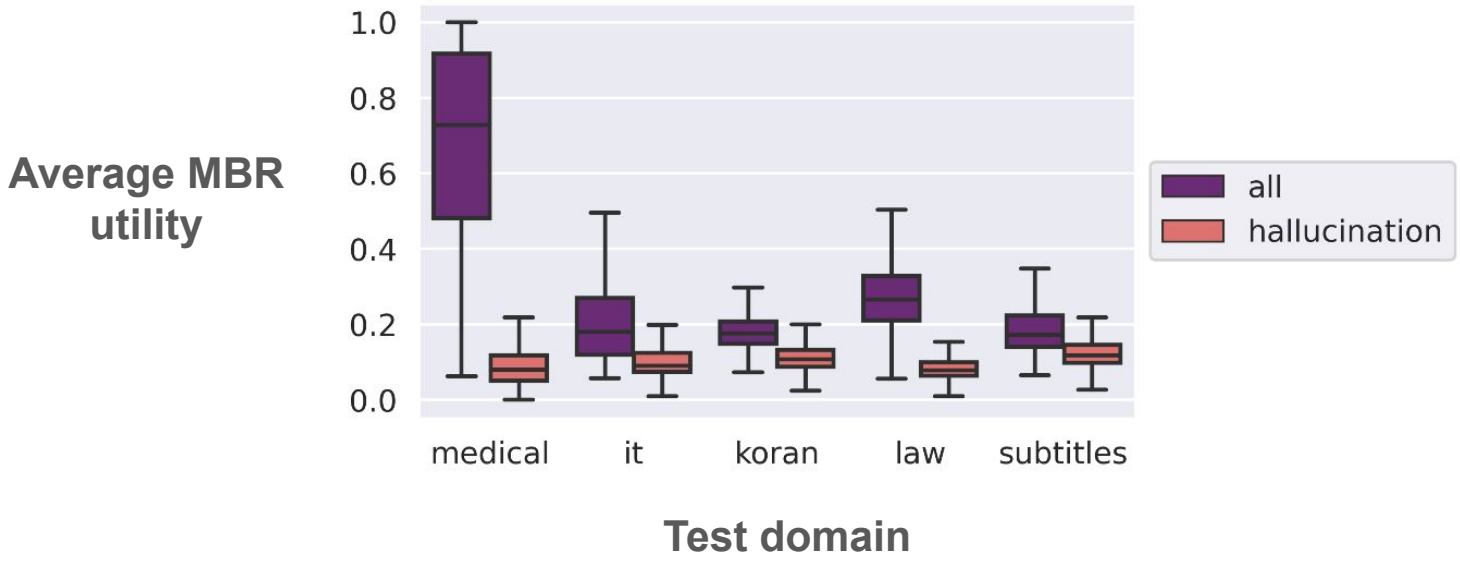
Müller and Sennrich (2021)
German-English

Domain robustness



Müller and Sennrich (2021)
German-English

Domain robustness



Müller and Sennrich (2021)
German-English

MBR example pools and utility

Domain robustness:

https://files.ifi.uzh.ch/cl/archiv/2020/clcontra/deu-eng.domain_robustness.it.html

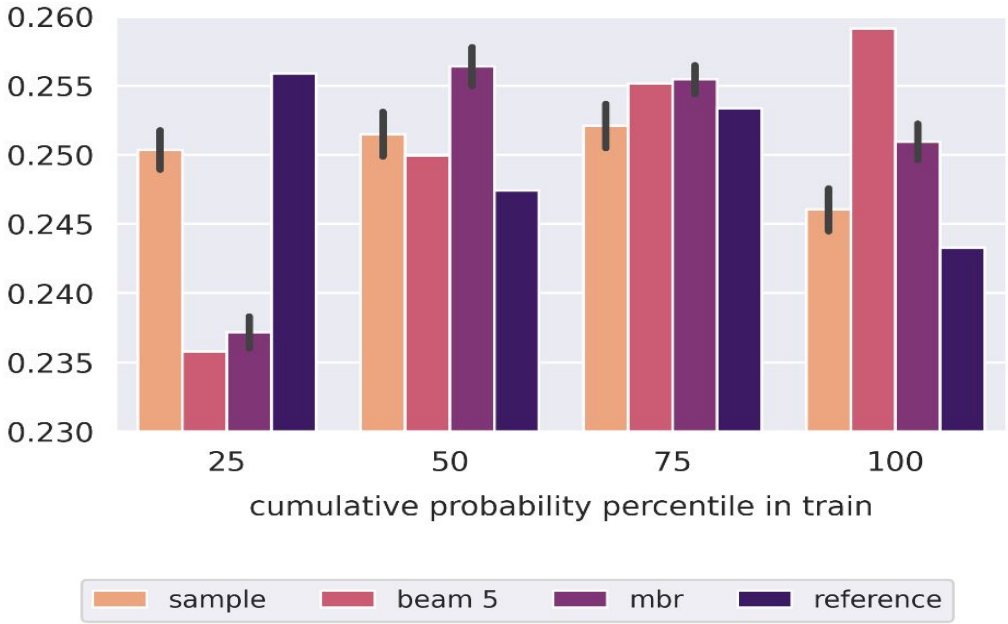
Copy noise:

https://files.ifi.uzh.ch/cl/archiv/2020/clcontra/ara-deu.copy_noise.0.1.slice-test.html

More here:

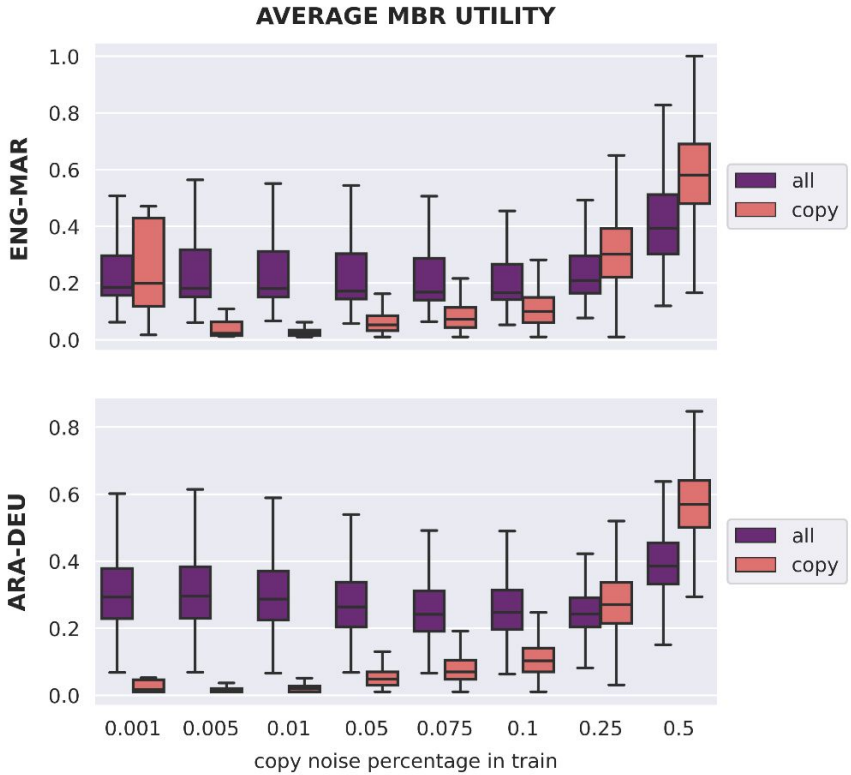
https://github.com/ZurichNLP/understanding-mbr#browse-mbr-samples

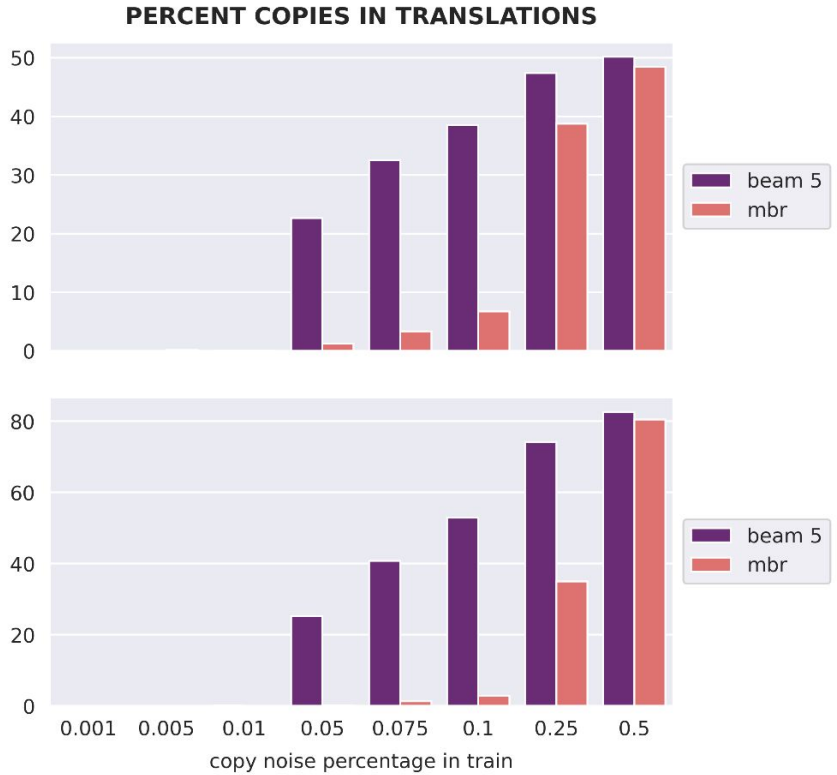
Frequency bias



Müller and Sennrich (2021)

Average utility of copies





Unsolved problems and potential future work

Well-known shortcomings of NMT

Role of beam search in causing these problems

Investigating if an alternative to beam search (MBR) causes fewer problems

Unsolved problems and potential future work

Going forward

1) Fix problems of MBR
2) Align training and decoding objectives

Unsolved issues with MBR

1) Efficiency
2) Idiosyncratic biases of utility function

Efficiency

For every sample s_i in the pool compute its utility:

$$\frac{1}{n}\sum_{s_j=1}^{n} u(s_i, s_j)$$

Sampling-Based Minimum Bayes Risk Decoding for Neural Machine Translation

Bryan Eikema
University of Amsterdam
b.eikema@uva.nl

Wilker Aziz
University of Amsterdam
w.aziz@uva.nl

Idiosyncratic biases of utility function

MBR + BLEU	11.51
MBR + METEOR	12.23
MBR + CHRF	12.50

Aligning training with decoding

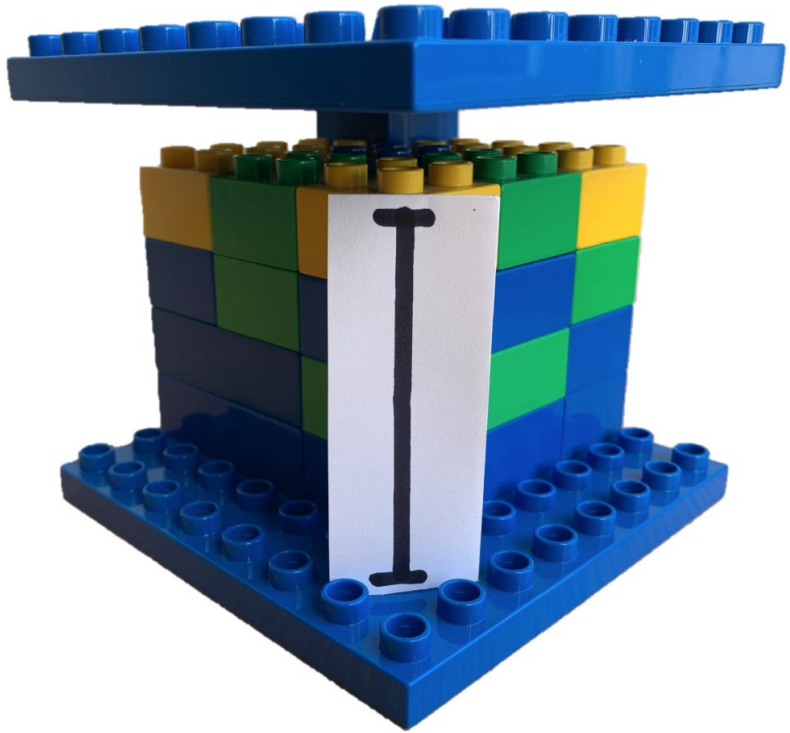
Machine Translation Decoding beyond Beam Search

Rémi Leblond [1] **Jean-Baptiste Alayrac** [1] **Laurent Sifre** [1] **Miruna Pislar** [1] **Jean-Baptiste Lespiau** [1]
Ioannis Antonoglou [1] **Karen Simonyan** [1] **Oriol Vinyals** [1]

Significance of replacing beam search in an MT experiment

Broad categories of solutions

Problem	Solutions typically about	Example
Length bias	Modifying beam search	Heuristic length normalization
Robustness to training noise	Data preprocessing	Corpus filtering
Domain robustness	Model architecture	Reconstruction model



Adams et al (2021)