



University of
Zurich^{UZH}

MT Marathon 2023: Decoding

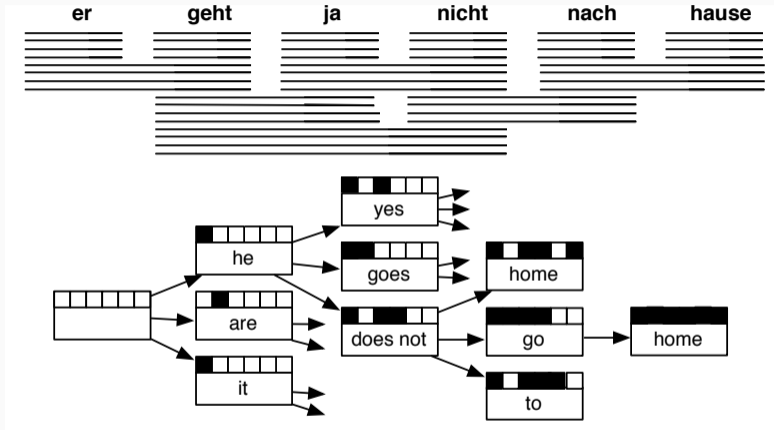
Rico Sennrich

29 August 2023

University of Zurich

University of Edinburgh

Decoding for Phrase-based Statistical Machine Translation



<http://www.statmt.org/book/slides/06-decoding.pdf>

today's lecture

- what are the standard decoding algorithms for neural MT?
- problems with beam search
- some advanced decoding algorithms:
 - constrained decoding
 - simultaneous translation
 - Minimum Bayes Risk decoding

Basic Decoding Algorithms for Neural Machine Translation

Modelling Translation

- Suppose that we have:
 - a source sentence S of length m (x_1, \dots, x_m)
 - a target sentence T of length n (y_1, \dots, y_n)
- We can express translation as a probabilistic model

$$T^* = \arg \max_T p(T|S)$$

- Expanding using the chain rule gives

$$\begin{aligned} p(T|S) &= p(y_1, \dots, y_n | x_1, \dots, x_m) \\ &= \prod_{i=1}^n p(y_i | y_1, \dots, y_{i-1}, x_1, \dots, x_m) \end{aligned}$$

Application of Encoder-Decoder Model

Scoring (a translation)

T = La croissance économique s'est ralentie ces dernières années.

S = Economic growth has slowed down in recent years.

$p(T|S) = ?$

Decoding (a source sentence)

Generate the most probable translation of a source sentence

S = Economic growth has slowed down in recent years.

$T^* = \operatorname{argmax}_T p(T|S)$

naive algorithm:

- generate every possible sentence T in target language
- compute score $p(T|S)$ for each
- pick best one

intractable: $|\text{vocab}|^N$ translations for output length N

better exact search [Stahlberg and Byrne, 2019, Meister et al., 2020]:

- probability of hypothesis monotonically decreases as it is extended
→ we can safely discard any partial hypothesis that is less probably than *most probable completed hypothesis*
- build tree of translation hypotheses depth-first, or with Dijkstra's algorithm

still impractically slow, and only used for analysis

Decoding for Neural Machine Translation: Sampling/Greedy Search

- at each time step, compute probability distribution $P(y_i|S, y_{<i})$
- select y_i according to some heuristic:
 - sampling: sample from $P(y_i|S, y_{<i})$
 - greedy search: pick $\operatorname{argmax}_y p(y_i|S, y_{<i})$
- continue until we generate $\langle \text{eos} \rangle$



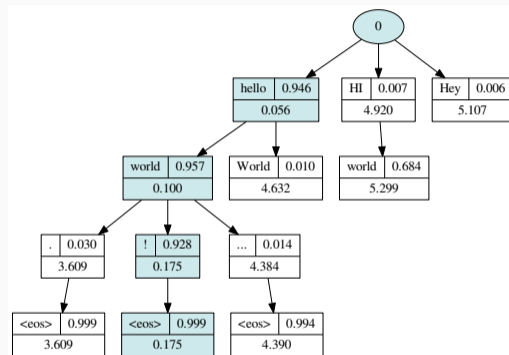
efficient, but suboptimal

Decoding for Neural Machine Translation: Beam Search

- maintain list of K hypotheses (beam)
- at each time step:
 - expand each hypothesis k : $p(y_i^k | S, y_{<i}^k)$
 - select K hypotheses with highest total probability, and add to new beam:

$$\prod_i p(y_i^k | S, y_{<i}^k)$$

- remove hypotheses ending in $\langle eos \rangle$ from beam (to final list)
- when beam is empty, select hypothesis (in final list) with highest total probability



$K = 3$

minibatches allow more parallelism at training time

at inference time, similar strategy possible:

- predict continuations of hypotheses in beam in parallel
- process different source sentences in parallel
- do a mix of both

Decoding Efficiency: Further Pointers

- **prune** model parameters
efficiency gains especially when whole structures (layers, attention heads, ...) can be pruned
- **quantize** model parameters to 4-bit or 8-bit
better memory efficiency; faster computation (depending on hardware)
- **knowledge distillation** improves quality with small models and beam search
- predict different time-steps in parallel:
 - non-autoregressive translation**: all time-steps predicted in parallel
 - semi-autoregressive translation**: multiple time-steps predicted in parallel

For further reading, check out the efficiency shared task of WMT!

basic idea

- combine decision of multiple classifiers by voting
- ensemble will reduce error if these conditions are met:
 - base classifiers are accurate
 - base classifiers are diverse (make different errors)

- vote at each time step to explore same search space
(better than decoding with one, reranking n-best list with others)
- voting mechanism: typically average (log-)probability

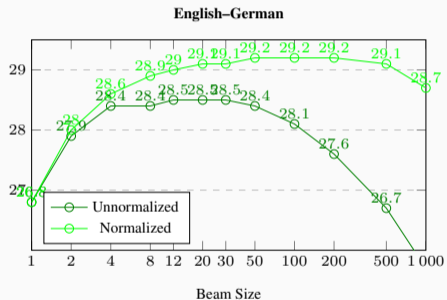
$$\log P(y_i|S, y_{<i}) = \frac{\sum_{m=1}^M \log P_m(y_i|S, y_{<i})}{M}$$

- requirements for voting at each time step:
 - same output vocabulary
 - same factorization of Y
 - but: internal network architecture may be different
- individual models can be checkpoints of same training run
(cheap, but less diversity)

Problems with Beam Search

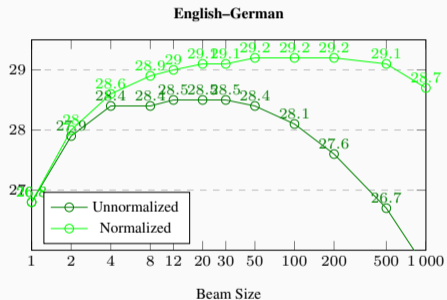
Beam Size

- small beam ($K \approx 10$) offers good speed-quality trade-off
- larger beams can even hurt quality! [Koehn and Knowles, 2017]



Beam Size

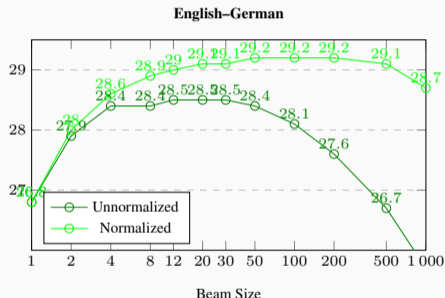
- small beam ($K \approx 10$) offers good speed-quality trade-off
- larger beams can even hurt quality! [Koehn and Knowles, 2017]



how can beam search perform worse with larger beams?

Beam Size

- small beam ($K \approx 10$) offers good speed-quality trade-off
- larger beams can even hurt quality! [Koehn and Knowles, 2017]



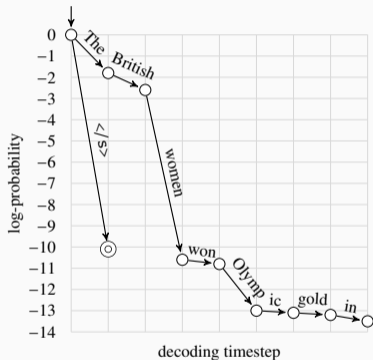
how can beam search perform worse with larger beams?

hint: search errors not to blame: exact search gives even worse results

[Stahlberg and Byrne, 2019, Meister et al., 2020]

Length Bias

locally normalized models
have label bias



heuristic solutions:

- divide total cost by length (length normalization):
$$score(Y, X) = \frac{\log(P(Y|X))}{|Y|}$$
- more complex normalisation term parametrised by α
$$score(Y, X) = \frac{\log(P(Y|X))}{\frac{(5+|Y|)^\alpha}{(6)^\alpha}}$$
- regularize towards *uniform information density*
e.g. squared regularizer:
$$score(Y, X) = \sum_{i=1}^n (-\log p(y_i|y_{<i, X}))^2$$

[Murray and Chiang, 2018]:
<eos> as low-entropy state

- **copy mode:** if prefix is copy of input, highly probable that next token continues pattern.
- **repetition loop:** what are likely continuations of this hypothesis?
oh my god ! ! ! ! ! ! ! !
- **hallucination:** certain prefixes (unrelated to source) that happen to be low entropy (frequent during training?)

Advanced Decoding Algorithms

why?

- force translation of terminology
- interactive machine translation

2

Contributors: (this should be a list of wo

Mitarbeiter:

Mitarbeiter: (das sollte eine Liste von v

3

Donate link: <http://example.com/>

Spenden Link: |

Spenden Link: <http://example.com/>

Prefix-Constrained Decoding

- cumbersome in phrase-based SMT
- very natural in neural MT
- standard decoding:

$$p(T|S) = \prod_{i=1}^n p(y_i | y_1, \dots, y_{i-1}, x_1, \dots, x_m)$$

- prefix-constrained decoding:

$$\text{PRE} = y_1, \dots, y_j$$

$$p(T|S, \text{PRE}) = \prod_{i=j+1}^n p(y_i | y_1, \dots, y_{i-1}, x_1, \dots, x_m)$$

- simple change to decoding algorithm; no changes to model/training

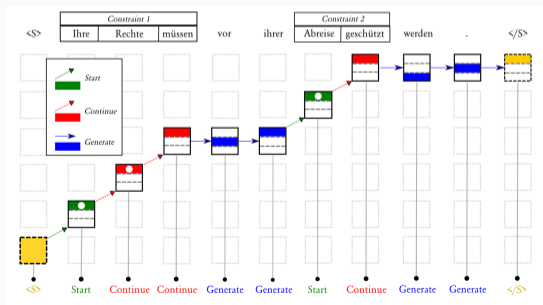
arbitrary constraints

- how can we decode with more general constraints?
- keep track of how many constraints hypothesis fulfills
- finished hypothesis is only valid if all constraints are fulfilled
- challenge: hypotheses that fulfill constraints must survive pruning

Constrained Decoding

Grid Beam Search [Hokamp and Liu, 2017]

- core idea: eliminate competition between hypos that fulfill different number of constraints
- 2d grid (each box is one beam):
 - x axis: number of time steps
 - y axis: number of constraint tokens matched



Input: Rights protection should begin before their departure .

Grid Beam Search [Hokamp and Liu, 2017]

- very general:
 - agnostic to model architecture
 - requires no source-side information
 - requires no retraining
- constraints must be in-vocabulary: use subword-level model
- problem: high computational complexity: $O(|V|kct)$
(k : beam size; t : length; c : # constraint tokens)
→ [Post and Vilar, 2018] use single, shared beam

Soft Constraints: Sentence-Level

motivation: controlling politeness/formality

T-V distinction

language	informal (T)	formal (V)
Latin	tu	vos
Chinese	你 (nǐ)	您 (nín)
French	tu	vous
German	du	Sie

Soft Constraints: Sentence-Level

motivation: controlling politeness/formality

T-V distinction		
language	informal (T)	formal (V)
Latin	tu	vos
Chinese	你 (nǐ)	您 (nín)
French	tu	vous
German	du	Sie
Early Modern English	thou	ye
Modern English		you

- inconsistency in T-V choice is a “limitation of MT technology” that is “often frustrat[ing]” to post-editors [Etchegoyhen et al., 2014]

Soft Constraints: Sentence-Level

motivation: controlling politeness/formality

T-V distinction

language	informal (T)	formal (V)
Latin	tu	vos
Chinese	你 (nǐ)	您 (nín)
French	tu	vous
German	du	Sie
Early Modern English	thou	ye
Modern English		you

What users want



- inconsistency in T-V choice is a “limitation of MT technology” that is “often frustrat[ing]” to post-editors [Etchegoyhen et al., 2014]

Core idea

- additional input feature that is based on target-side information
→ extra word at end of source sentence
- mark in English text if German translation is formal or not (+noise)
 - Are you ok?
 - Sind Sie in Ordnung?
 - are you ok?
 - Bist du in Ordnung?

At test time

- we can control level of formality by adding side constraints to input

Core idea

- additional input feature that is based on target-side information
→ extra word at end of source sentence
- mark in English text if German translation is formal or not (+noise)
 - Are you ok?
 - Sind **Sie** in Ordnung?
 - are you ok?
 - Bist **du** in Ordnung?

At test time

- we can control level of formality by adding side constraints to input

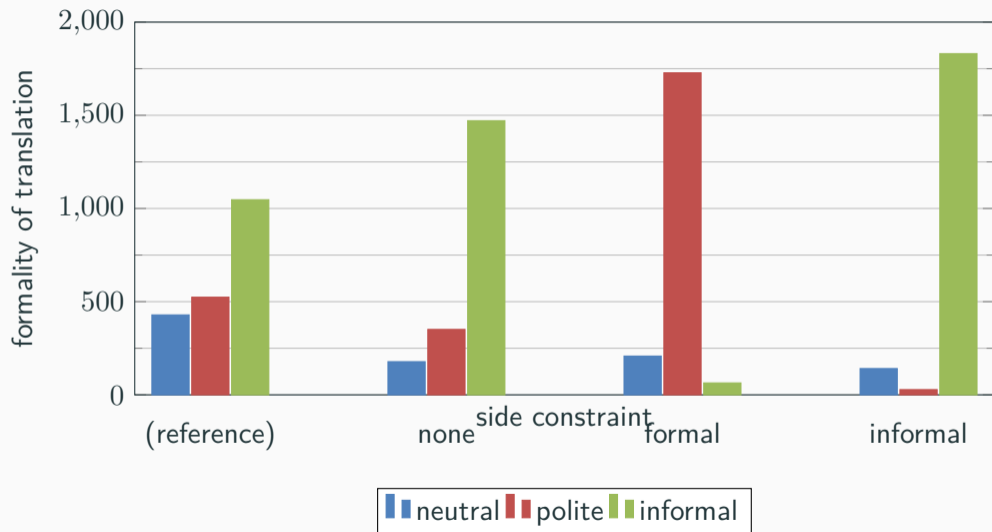
Core idea

- additional input feature that is based on target-side information
→ extra word at end of source sentence
- mark in English text if German translation is formal or not (+noise)
 - Are you ok? <formal>
 - are you ok? <informal>
 - Sind Sie in Ordnung?
 - Bist du in Ordnung?

At test time

- we can control level of formality by adding side constraints to input

Results: formality as a function of soft constraint



- control production of other information missing from source text
 - gender marking
 - tense
 - evidentiality
 - ...
- domain adaptation
- control output language

Soft Constraints: Token-Level

recipe for terminological constraints [Dinu et al., 2019]:

project target-side information (translations) into source; mark them with token-level features

- at training time, copy target words to source based on terminology match.
(but not always, so that model still works without constraints)
- extra embedding indicates whether word should be:
 - 0 translated normally
 - 1 not translated, but used for disambiguation
 - 2 copied
- with the right training data augmentation (fuzzy matching between target word and terminology entry), model also learns to inflect terms.

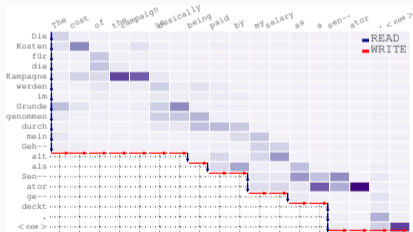
src (orig)	all alternates shall be elected for one term.
ref	alle Stellvertreter werden für eine Amtszeit gewählt.
src-app	all ₀ alternates ₁ Stellvertreter ₂ shall ₀ be ₀ elected ₀ for ₀ one ₀ term ₀ .

Simultaneous Translation

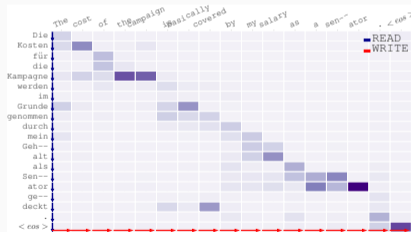
objectives in simultaneous translation:

- 1 maximize translation quality
- 2 minimize latency

to minimize latency, start translating before full input has been seen



(a) Simultaneous Neural Machine Translation



(b) Neural Machine Translation

- model that predicts translation based on partial input
- policy that decides whether to output translation or wait for more input
- metrics to measure latency and translation quality to optimize policy

Translation Model

- 😊 no need to change model architecture
some authors use unidirectional encoders and greedy search for efficiency [Gu et al., 2017, Cho and Esipova, 2016]
- 😞 sentence-level systems perform poorly when input is sentence fragment
- solution: train on sentence fragments [Niehues et al., 2018]
→ simply use prefix of parallel sentences (proportionally to length)
- recent tip: use distilled training data [Sen et al., 2023]
→ more monotonic, so less need to “guess”

Source	For more than 30 years , Josef Winkler has been writing from the heart , telling of the hardships of his childhood and youth .
Distilled Target	Seit mehr als 30 Jahren schreibt Josef Winkler aus dem Herzen und erzählt von der Not seiner Kindheit und Jugend .
Real Target	Josef Winkler schreibt sich seit mehr als 30 Jahren die Nöte seiner Kindheit und Jugend von der Seele .

actions

- simplest action space:
 - **read** a source token
 - **write** a target token
- more complex policies can allow overwriting past decisions

policy

- policy can be learned as function of state (input read so far / output produced so far)
[Grissom II et al., 2014, Gu et al., 2017]
- simple heuristic policies can work

Heuristic Policy: Wait- k [Ma et al., 2019]

- 1 read k source tokens
- 2 write a target token, then read the next source token
- 3 repeat 2 until we reach EOS (in source)
- 4 write target tokens until we produce EOS

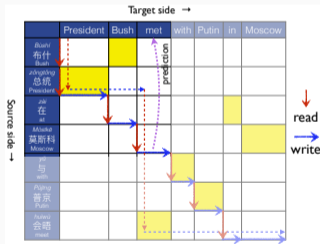


Figure 1: Our wait- k model emits target word y_t given source-side prefix $x_1 \dots x_{t+k-1}$, often before seeing the corresponding source word (here $k=2$, outputting y_3 ="met" before x_7 ="huiwù"). Without anticipation, a 5-word wait is needed (dashed arrows). See also Fig. 2.

a simplified measure of lag:

$$LAG_g(x, y) = \frac{1}{|y|} \sum_{t=1}^{|y|} g(t) - (t - 1)$$

$g(t)$: how many source words have we read at time step t ?

Measuring Latency

getting rid of two simplifications:

- we don't average lag over all positions y , but only until we have read full source:

$$\tau_g(|x|) = \min\{t \mid g(t) = |x|\}$$

- if $|x| \neq |y|$, take into account length ratio

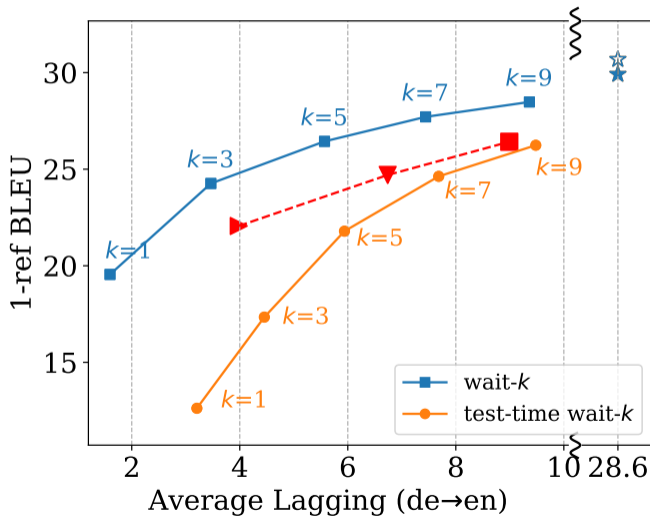


Figure 4: Illustration of our proposed Average Lagging latency metric. The left figure shows a simple case when $|x| = |y|$ while the right figure shows a more general case when $|x| \neq |y|$. The red policy is wait-4, the yellow is wait-1, and the thick black is a policy whose AL is 0.

Average Lag (AL) [Ma et al., 2019]:

$$AL_g(x, y) = \frac{1}{\tau_g(|x|)} \sum_{t=1}^{\tau_g(|x|)} g(t) - (t-1) \cdot \frac{|x|}{|y|}$$

Trading Off Latency and Translation Quality



Policies Allowing Corrections: Re-translation

- 😊 with corrections, we match (final) quality of sentence-level system
- 😊 effect even larger in SLT when transcript is updated
- 😞 frequent corrections lead to 'flicker' and poor user experience
- simplest policy: re-translate each fragment, allowing unlimited corrections

some results [Niehues et al., 2018]

system	BLEU (EN→ES; tst2010)	corrections (words)
trained on sentence-level	26.0	182 000
trained on sentence fragments	25.5	98 000
trained on both	26.0	101 000

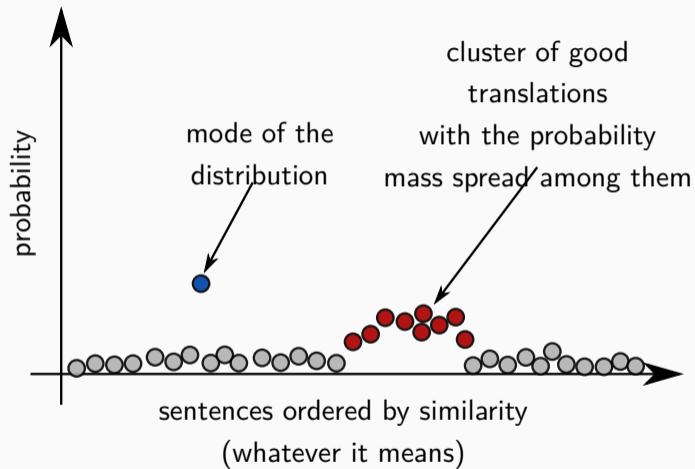
Re-translation: Trading off Flicker, Lag, and Quality

- higher lag, lower flicker: test-time wait- k
 - lower BLEU, lower flicker: bias beam search towards prefix:
interpolate between
 - probability distribution from model
 - probability distribution that assigns 100% probability to prefix
- interpolation weight controls trade-off

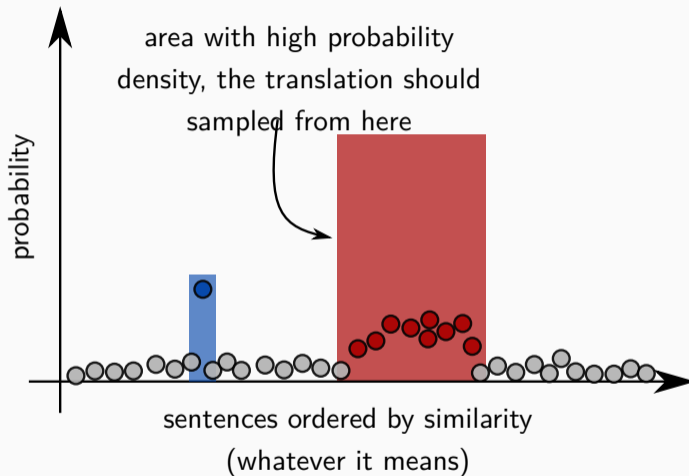
System	β	k	BLEU	Translation Lag	Normalized Erasure
Baseline	0.0	0	20.40	4.13	2.11
+ Bias	0.5	0	20.03	3.00	0.72
+ Mask- k	0.0	10	20.40	5.98	0.53
+ Both	0.5	5	20.17	4.11	0.12

Table 1: English-to-German results on our TED test set. Translation Lag is the time delay (in seconds) between when a source word was spoken versus when a corresponding output word was finalized. A word is finalized when the word and any words before it remain unchanged. Normalized Erasure is measured in number of erased partial target tokens per final target token. [Arivazhagan et al., 2020]

Minimum Bayes Risk Decoding



Minimum Bayes Risk Decoding



$$y^* = \operatorname{argmin}_{y_i \in \mathcal{Y}} \sum_{y_j \in \mathcal{Y}} P(y_j|x) \Delta(y_i, y_j)$$

- use some user-defined risk function Δ (here: -BLEU)
- approximate \mathcal{Y} via sampling or beam search
- Δ is not defined in respect to reference, but other hypotheses
→ consensus decoding

$$y^* = \operatorname{argmin}_{y_i \in \mathcal{Y}} \sum_{y_j \in \mathcal{Y}} P(y_j|x) \Delta(y_i, y_j)$$

efficiency considerations:

- increasing $|\mathcal{Y}|$ leads to empirically better results
 - no beam search curse 😊
 - high computational cost 😞
 - produce $|\mathcal{Y}|$ hypotheses
 - score $|\mathcal{Y}|^2$ pairs with metric
- active work on faster approximations [Eikema and Aziz, 2021]:
 - conceptually, we can use subsets of $|\mathcal{Y}|$ for candidate hypotheses (C) and as pseudo-references (support S):

$$y^* = \operatorname{argmin}_{y_i \in C} \sum_{y_j \in S} P(y_j|x) \Delta(y_i, y_j)$$

Why Could MBR Decoding Be More Robust?

some pathological translations (like copying source) can amass high average probability over time, but will be unlike other probable translations.

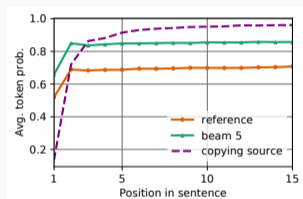


Figure 4. Average probability at each position of the output sequence on the WMT 14 En-Fr validation set, comparing the reference translation, beam search hypothesis ($k = 5$), and copying the source sentence.

[Ott et al., 2018]

MBR decoding reduces the generation of some high-probability deficient translations like copying and “hallucinations” [Müller and Sennrich, 2021]

Minimum Bayes Risk Decoding with Neural Metrics

exciting recent results [Freitag et al., 2022]:

MBR with neural metric as risk function (BLEURT)

- have much lower model probability than beam search outputs
- are significantly better according to human error annotation
30% reduction in mistranslations DE→EN

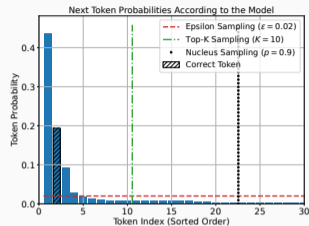
source	Das Lagern auf den Wiesen ist laut Parkordnung untersagt [...]
human	Camping on the grassland is omitted according to park ordinance [...]
beam search	Storing on the meadows is prohibited according to the park regulations [...]
MBR BLEURT	The park rules prohibit camping in the meadows [...]

Minimum Bayes Risk Decoding: On Sampling Strategies

standard (ancestral) sampling may produce poor hypotheses.

Heuristics to focus on most probable predictions:

- **top-k sampling**: only sample from k most probable tokens
- **nucleus sampling**: only sample from smallest set of tokens that has cumulative probability mass $\geq p$
- **epsilon sampling**: only sample from tokens whose probability $\geq \epsilon$



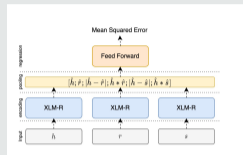
[Freitag et al., 2023]

[Freitag et al., 2023] report that MBR with epsilon sampling performs better than other sampling methods.

Do Biases of Metrics Affect MBR?

neural metric: COMET

- high correlation with human judgments [Kocmi et al., 2021]
- independent encoding of hyp/ref/src allows re-use: → better efficiency



observation: more name/number corruptions than with surface-level chrF++ [Amrhein and Sennrich, 2022]

src	Schon drei Jahre nach der Gründung verließ Green die Band 1970 .
ref	Green left the band three years after it was formed, in 1970 .
MBR _{chrF++}	Already three years after the foundation, Green left the band in 1970 .
MBR _{COMET}	Three years after the creation, Green left the band in 1980 .

src	[...] Mahmoud Guemama's Death - Algeria Loses a Patriot [...], Says President Tebboune .
ref	[...] Mahmoud Guemamas Tod - Algerien verliert einen Patrioten [...], sagt Präsident Tebboune .
MBR _{chrF++}	[...] Mahmoud Guemamas Tod - Algerien verliert einen Patriot [...], sagt Präsident Tebboune .
MBR _{COMET}	[...] Mahmud Guemamas Tod - Algerien verliert einen Patriot [...], sagt Präsident Tebboene .

Do Biases of Metrics Affect MBR?

F1-score for translation of numbers and named entities (EN→DE)

	Numbers		Named Entities	
reference	93.46		n/a	
alternative human	95.66	+ 2.20	77.66	
beam search	95.73	+ 2.27	70.03	- 7.63
MBR bleu	91.37	- 2.09	62.50	-15.16
MBR wmt20-comet-da	89.14	- 4.32	54.17	-23.49
MBR wmt21-comet-mqm	77.10	-16.36	53.31	-24.35
MBR retrain-comet-da	90.17	- 3.29	60.48	-17.18

- MBR has more corruptions than beam search; worse with COMET
- retraining COMET with synthetically corrupted data helps, but gap remains

Take-home messages

- simple beam search with modest beam size sufficient to find most probable translation...
...but most probable translation is not always good
- common fixes:
 - length normalisation
 - data cleaning (e.g. no source language text on target side to reduce copy problem)
- decoding becomes more complex if you want to:
 - output translations during speaking/typing to minimize latency
 - control output (e.g. terminology constraints)
- active research on alternatives to mode-seeking decoding

I'm hiring a post-doctoral researcher

join my group to work on low-resource NLP and transfer learning across tasks, languages and/or modalities!

great research group in Zurich (among top-ranked for quality of life!), and generous employment conditions

funding secured until 31/7/2025; apply by 15/9/2023



<https://www.cl.uzh.ch/senrich>



Amrhein, C. and Sennrich, R. (2022).

Identifying Weaknesses in Machine Translation Metrics Through Minimum Bayes Risk Decoding: A Case Study for COMET.

[arXiv e-prints.](#)



Arivazhagan, N., Cherry, C., Te, I., Macherey, W., Baljekar, P., and Foster, G. (2020).

Re-translation strategies for long form, simultaneous, spoken language translation.

In [ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing \(ICASSP\)](#), pages 7919–7923.



Cho, K. and Esipova, M. (2016).

Can neural machine translation do simultaneous translation?

CoRR, abs/1606.02012.



Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019).


Training neural machine translation to apply terminology constraints.

In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.



Eikema, B. and Aziz, W. (2021).

Sampling-based approximations to minimum bayes risk decoding for neural machine translation.

-  Etchegoyhen, T., Bywood, L., Fishel, M., Georgakopoulou, P., Jiang, J., Loenhout, G. V., Pozo, A. D., Maucec, M. S., Turner, A., and Volk, M. (2014).

Machine Translation for Subtitling: A Large-Scale Evaluation.

In

Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Reykjavik, Iceland. European Language Resources Association (ELRA).

-  Freitag, M., Ghorbani, B., and Fernandes, P. (2023).

Epsilon sampling rocks: Investigating sampling strategies for minimum bayes risk decoding for machine translation.



Freitag, M., Grangier, D., Tan, Q., and Liang, B. (2022).

High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics.

Transactions of the Association for Computational Linguistics, 10:811–825.




Grissom II, A., He, H., Boyd-Graber, J., Morgan, J., and Daumé III, H. (2014).

Don't Until the Final Verb Wait: Reinforcement Learning for Simultaneous Machine Translation.

In


Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (pages 1342–1352, Doha, Qatar. Association for Computational Linguistics.

 Gu, J., Neubig, G., Cho, K., and Li, V. O. (2017).

Learning to Translate in Real-time with Neural Machine Translation.

In


Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 1053–1062, Valencia, Spain. Association for Computational Linguistics.

 Hokamp, C. and Liu, Q. (2017).

Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search.

In

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pages 1535–1546.

 Kocmi, T., Federmann, C., Grundkiewicz, R., Junczys-Dowmunt, M., Matsushita, H., and Menezes, A. (2021).

To ship or not to ship: An extensive evaluation of automatic metrics for machine translation.

In Proceedings of the Sixth Conference on Machine Translation, pages 478–494, Online. Association for Computational Linguistics.

 Koehn, P. and Knowles, R. (2017).

Six Challenges for Neural Machine Translation.

In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver.



Libovický, J. (2020).

Jindřich's blog – machine translation weekly 63: Maximum a posteriori vs. minimum bayes risk decoding.

Online, Accessed: 03.10. 2021.



Ma, M., Huang, L., Xiong, H., Zheng, R., Liu, K., Zheng, B., Zhang, C., He, Z., Liu, H., Li, X., Wu, H., and Wang, H. (2019).

STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework.

In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3025–3036, Florence, Italy. Association for Computational Linguistics.

 Meister, C., Cotterell, R., and Vieira, T. (2020).

If beam search is the answer, what was the question?

In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2173–2185, Online. Association for Computational Linguistics.

 Müller, M. and Sennrich, R. (2021).

Understanding the properties of minimum Bayes risk decoding in neural machine translation.

In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing

(Volume 1: Long Papers), pages 259–272, Online. Association for Computational Linguistics.

 Murray, K. and Chiang, D. (2018).

Correcting Length Bias in Neural Machine Translation.

In Proceedings of the Third Conference on Machine Translation, pages 212–223, Belgium, Brussels.

 Niehues, J., Pham, N.-Q., Ha, T.-L., Sperber, M., and Waibel, A. (2018).

Low-latency neural speech translation.

In 19th Annual Conference of the International Speech Communication, INTERSPEECH 2018; Hyderabad International Convention Centre (HICC)Hyderabad; India; 2 September 2018 through 6 September 2018. Ed.: C.C. Sekhar, volume 2018-September of

Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pages 1293–1297. ISCA, Lous Tourils.



Ott, M., Auli, M., Grangier, D., and Ranzato, M. (2018).

Analyzing uncertainty in neural machine translation.

In International Conference on Machine Learning.



Post, M. and Vilar, D. (2018).

Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation.

In

Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, pages 1314–1324. Association for Computational Linguistics.



Sen, S., Sennrich, R., Zhang, B., and Haddow, B. (2023).

Self-training reduces flicker in retranslation-based simultaneous translation.



In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3734–3744, Dubrovnik, Croatia. Association for Computational Linguistics.



Stahlberg, F. and Byrne, B. (2019).

On NMT search errors and model errors: Cat got your tongue?

In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3356–3362, Hong Kong, China. Association for Computational Linguistics.

-  Wang, C., Zhang, J., and Chen, H. (2018).
Semi-autoregressive neural machine translation.
In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 479–488, Brussels, Belgium. Association for Computational Linguistics.
-  Zhang, B., Titov, I., and Sennrich, R. (2020).
Fast interleaved bidirectional sequence generation.
In Proceedings of the Fifth Conference on Machine Translation, pages 426–438, Online. Association for Computational Linguistics.
-  Zhou, C., Neubig, G., and Gu, J. (2019).
Understanding knowledge distillation in non-autoregressive machine translation.