Global Under-Resourced MEedia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D3.2 – GoURMET Final report on learning structural models

Nature	Report	**Work Package**	WP3
Due Date	31/03/2021	**Submission Date**	31/03/2021
Main authors	Wilker Aziz (UVA)		
Co-authors	Ivan Titov (UEDIN)		
Reviewers	Alexandra Birch		
Keywords	survey, languages, resources		
Version Control			
v0.1	**Status**	Draft	10/06/2022
v1.0	**Status**	Final	29/06/2022

Contents

1 Introduction **6**

2 Task 3.1 – Modelling Latent Alignments **6**

2.1 Analysing the Source and Target Contributions to Predictions in NMT 7

2.2 The Training Process of NMT through the Lens of Classical SMT 9

2.3 Improving Knowledge Distillation for NAT-MT 12

3 Task 3.2 – Structured Sentence Models **12**

3.1 Mixed Random Variables 13

3.2 Sparse Encoders 15

3.3 Exploiting Context Beyond Sentence Level 16

4 Task 3.3 – Probabilistic Neural Machine Translation **18**

4.1 Inadequacy of the Mode in Markov Processes 19

4.2 Sampling-Based MBR Decoding 21

4.3 Scalable MBR Approximations 24

4.4 Sample-Efficient Approximations of Expected Utility 25

4.5 Loss-Calibrated Machine Translation 26

5 Conclusion **29**

List of Figures

1 (a) contribution of the whole source at each generation step; (b) total contribution of source tokens at each position to the whole target sentence. 8

2 For each generation step, the figure shows entropy of (a) source, (b) target contributions to the prediction. 8

3 Evolution of source and target contributions to prediction over the course of training. 9

4 (a) KenLM scores (horizontal dashed lines are the scores for the references); (b) proportion of tokens of different frequency ranks in model translations. En-Ru. . . 9

5 Translations at different steps during training. En-De. 10

6 (a) BLEU score; (b) token-level accuracy (the proportion of cases where the correct next token is the most probable choice). WMT En-Ru. 10

7 (a) fuzzy reordering score (for references: 0.6), (b) Kendall tau distance (for references: 0.06); WMT En-Ru. The arrows point in the direction of less monotonic alignments (more complicated reorderings). 11

8 Translations at different training steps. Same-colored chunks are approximately aligned to each other. 11

9 (a) BLEU score of the autoregressive (AT) Transformer-base (teacher for distillation); (b) fuzzy reordering score for the distilled training data obtained from checkpoints of the AT teacher; (c) BLEU scores for the vanilla NAT-MT model trained on different distilled data. 13

10 Multivariate distributions over the probability simplex. Standard distributions, like the Logistic-Normal (left), assign zero probability to all faces except to the relative interior of the simplex. Mixed distributions support assigning probability to the full simplex, including its boundary: the Gaussian-Sparsemax (right) induces a distribution over the 1-dimensional edges (shown as a histogram), and assigns $\Pr\{(1, 0, 0)\} = 0.022$. 14

11 Encoder-decoder attention distribution of target words (vertical axis) over source words (horizontal axis) for vanilla attention (Vaswani et al., 2017), sparse attention (Correia et al., 2019) (which aims at sparsity with respect to individual decoder steps) and our model (Zhang et al., 2021). Darker color indicates larger attention weight, and the white blocks denote an attention weight of 0. The source words whose encoding is pruned by L0DROP (receiving zero weight) are highlighted in red. 15

12 Distribution of surprisal based on 1,000 samples from the model. Source: *Der Großbrand in einem als besonders gefährlich geltenden Chemiewerk in der nordfranzösischen Stadt Rouen ist gelöscht.* Reference: *The large fire was put out at a chemical plant considered to be particularly hazardous and located in the northern French city of Rouen.* Beam search output: *The big fire in a particularly dangerous chemical plant in the northern French town of Rouen has been extinguished.* 21

13 Quality of sampling-based MBR output for various sizes of N using BEER as target utility. We report both BEER and BLEU scores. 23

14 Proportion plots of expected utility for 3 strategies for constructing $\mathcal{H}(x)$, using 100 translation candidates per strategy. We estimate expected utility using 1,000 samples. Results are aggregated over 100 source sentences. 23

15 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. 24

16 Performance of MBR-N-by-S: we estimate the expected utility of N hypotheses using S samples. We show average performance over 3 runs with 1 standard deviation. The dashed line shows the performance at $N = S = 405$. 25

17 Performance of Bayesian Monte Carlo against Monte Carlo estimates of expected utility as a function of sample size S (horizontal axis). We measure mean-squared error (vertical axis) against a very robust estimate of expected utility obtained using 1,000 independent samples. The blue curve is the error of MC using S samples, the orange curve is the error of BMC using S samples. 27

18 Summary of research output. 30

Abstract

This deliverable reports the work conducted within WP3 on structure induction at sentence level for low-resource neural machine translation (NMT). It focuses on three main tasks: *inducing word alignments*, *learning structured sentence models*, and *exploiting the probabilistic framework for better decisions and data-efficient NMT*. We report on progress in the second half of the project.

1 Introduction

WP3 provides scientific advances in machine translation for the low-resource setting by developing probabilistic learning algorithms which induce and exploit structured representations of sentences and documents. WP3 has four main goals:

- Develop methods which explicitly model inter-dependencies between terms in the source and target sentences as latent alignments, and induce them in such a way as to be beneficial for the translation quality;
- Develop algorithms which induce structured representations of sentences and documents from parallel and monolingual data;
- Develop both NMT methods which exploit these induced representations and methods which optimise for translation and structure induction in an end-to-end fashion.
- Exploit and advance implications of the probabilistic formulation of neural machine translation models, in particular, where this will lead to advances in low-resource settings.

To cover these goals we proposed 3 tasks, namely,

T3.1 Modelling latent alignments (Section 2);

T3.2 Structured sentence models (Section 3);

T3.3 Probabilistic neural machine translation (Section 4)

The first half of the project has seen considerable progress in all three fronts of the work package with some emphasis on T3.2 and this work has been described in the M18 Deliverable D3.1 Initial progress report on learning structural models. The second half of the project similarly advanced along all three tasks, with some emphasis on T3.1 and T3.3. The work reported here has appeared in conference publications, MSc theses, pre-prints under review, and has led to the release of open-source software and data. This document is an overview of this research output, in particular, it highlights research challenges and progress due to GoURMET.

2 Task 3.1 – Modelling Latent Alignments

Proposal highlights:

- induce alignments as latent variable jointly with a simpler NMT system (one that makes stronger independence assumptions than standard NMT does);
- overcome intractability with variational inference and investigate both discrete and approximately discrete alignments;
- combine alignments with NMT aiming at improved translation quality.

As the field commits to standardised architectures and pre-trained models, largely due to their potential for multilingual training and transfer learning, modifying the internal structure of the model to exploit explicit word alignments is not as relevant a path as it seemed at the beginning of the project. One of the advantages of word alignments is that they provide a transparent and human-interpretable lens into what information models exploit to perform predictions. Here we report on work that retains that exact motivation, but uses alignment as analysis—rather than modelling—tools to advance our understanding of NMT from training to generation.

Summary of work done. In Section 2.1, we use a post-hoc explanation technique to investigate the influence of source and target context to each generation step in NMT. This can be thought of viewing the implict way in which a prediction is aligned to or influenced by the different types of contex. This analysis tells us more about NMT training and generation and informs strategies that mitigate certain known biases of NMT models (*e.g.*, exposure bias). In Section 2.2, we then analyse NMT models in terms of components of a statistical machine translation pipeline, including components based on word alignments. Such an analysis sheds light onto what's expected of an NMT model throughout its training in terms of performance along dimensions such as fluency, adequacy and word order. This increased understanding can be used, for example, in settings where an NMT teacher model guides the training of NMT student model, as it is common in the training of non-autoregressive machine translation models, which we investigate in Section 2.3.

2.1 Analysing the Source and Target Contributions to Predictions in NMT

In NMT, the generation of a target token is influenced by two types of context: the source and the prefix of the target sequence. While many attempts to understand the internal workings of NMT models have been made, none of them explicitly evaluates relative source and target contributions to a generation decision.

In (Voita et al., 2021a), we argue that this relative contribution can be evaluated by adopting a variant of layerwise relevance propagation (LRP), a type of model explanation technique. We extend LRP to the Transformer and conduct an analysis of NMT models which explicitly evaluates the source and target relative contributions to the generation process. We analyze changes in these contributions when conditioning on different types of prefixes, when varying the training objective or the amount of training data, and during the training process. We find that models trained with more data tend to rely on source information more and to have more sharp token contributions; the training process is non-monotonic with several stages of different nature.

Data. We use random subsets of the WMT14 En-Fr dataset of different size: 1m, 2.5m, 5m, 10m, 20m, 30m sentence pairs.

Model. We follow the setup of Transformer base model (Vaswani et al., 2017) with the standard training setting (for complete details see (Voita et al., 2021a)).

Findings. During the generation process, the influence of source decreases (or, equivalently, the influence of the prefix increases)—see Figure 1a. This is expected: with a longer prefix, the model has less uncertainty in deciding which source tokens to use, but needs to control more for fluency. On average, source tokens at earlier positions influence translations more than tokens at later

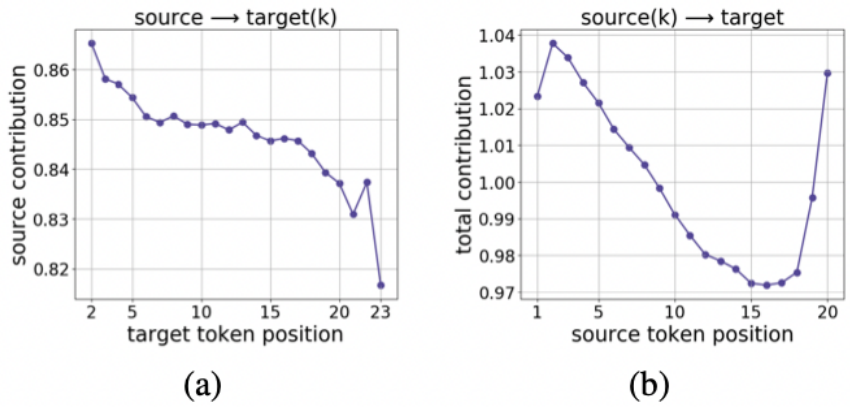

Figure 1: (a) contribution of the whole source at each generation step; (b) total contribution of source tokens at each position to the whole target sentence.

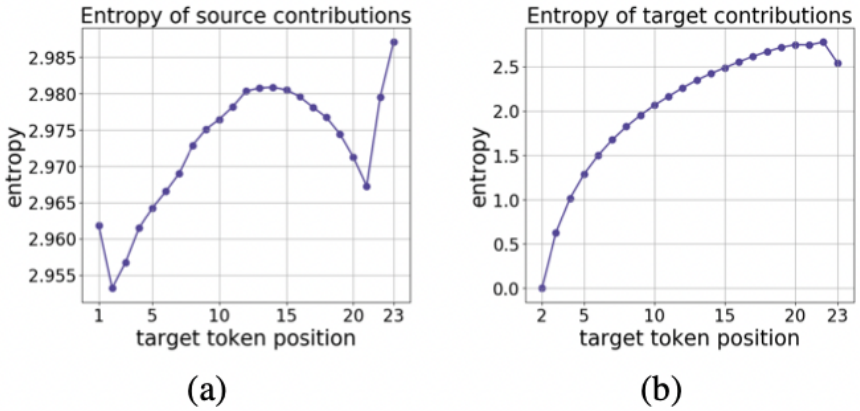

Figure 2: For each generation step, the figure shows entropy of (a) source, (b) target contributions to the prediction.

ones—see Figure 1b. This may be because the alignment between English and French languages is roughly monotonic. We leave for future work investigating the changes in this behavior for language pairs with more complex alignment. We now look at how 'sharp' contributions of source or target tokens are at different generation steps by evaluating entropy of normalised source or target contributions. Figure 2a shows that during generation, entropy increases until approximately 2/3 of the translation is generated, then decreases when generating the remaining part. Figure 2b shows that entropy of target contributions is higher for longer prefixes. This means that the model does use longer contexts in a non-trivial way. Now, rather than generation from a converged model, we turn to analyzing the training process of an NMT model. Specifically, we look at the changes in how the predictions are formed (*e.g.*, changes in the amount of source/target contributions and in the entropy of these contributions) over the course of training. Our main findings are summarized in Figure 3, in particular, the training process is non-monotonic with several distinct stages. These stages agree with the ones found in previous work focused on validating the lottery ticket hypothesis (Frankle and Carbin, 2019; Frankle et al., 2020), which suggests future investigation of this connection. In the paper, we also connect over-reliance on target history to exposure bias and hallucination. In future work, our methodology can be used to measure the effects of different and novel training regimes on the balance of source and target contributions.

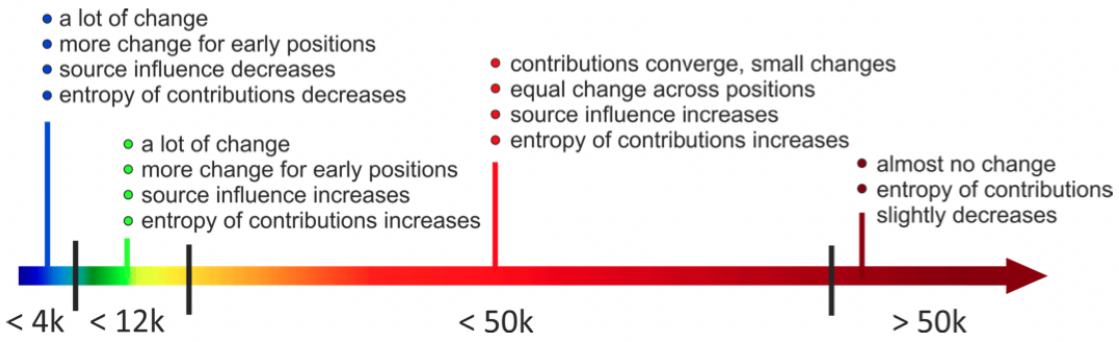

Figure 3: Evolution of source and target contributions to prediction over the course of training.

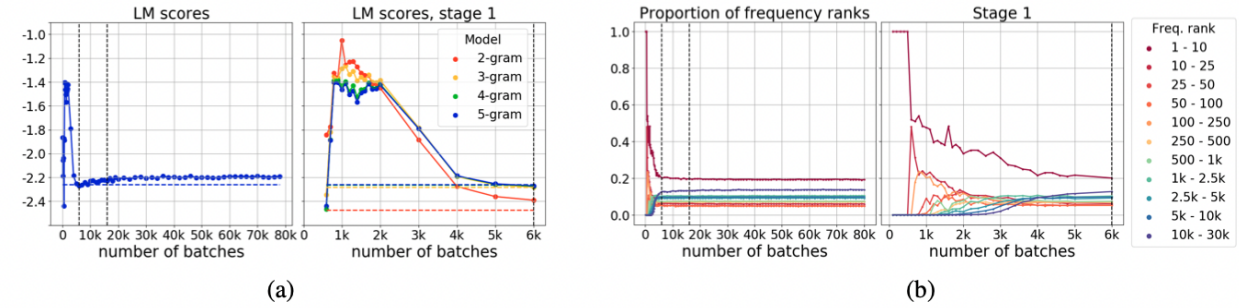

Figure 4: (a) KenLM scores (horizontal dashed lines are the scores for the references); (b) proportion of tokens of different frequency ranks in model translations. En-Ru.

2.2 The Training Process of NMT through the Lens of Classical SMT

Differently from the traditional statistical MT that decomposes the translation task into distinct separately learned components, neural machine translation uses a single neural network to model the entire translation process. While this has led to improved translation quality and transfer, it limits our understanding of the model and hence our ability to improve specific aspects of NMT's design and training.

In (Voita et al., 2021b), we analyse the output of NMT over the course of training relating its competences to three core SMT components and find that during training, NMT first focuses on learning target-side language modeling, then improves translation quality approaching word-by-word translation, and finally learns more complicated reordering patterns.

Data. We use the WMT14 news translation shared task for English-German (5.8m sentence pairs) and English-Russian (2.5m sentence pairs).

Methodology. We train Transformer base models (Vaswani et al., 2017) and analyse model outputs after a number of batches of training along a number of dimensions. We investigate fluency in terms of n-gram LMs (Heafield et al., 2013), we analyse word order in terms of fuzzy reordering score (Talbot et al., 2011) and Kendall tau distance, for which we word-align model outputs and reference using `fast align` (Dyer et al., 2013).

Source sentence: six months of construction works , that 's brutal .

Step	Model translation (German)	English (approximate translation from German)
100	, die ,	, the ,
300	, die , die , die , die , die , die , die , die , die , die , die ,	, the , the , the , the , the , the , the , the , the , the , the , the ,
700	die kommission , dass die kommission die kommission , die kommission , die kommission .	the commission , that the commission, the commission , the commission , the commission .
900	es ist , dass die kommission , dass die kommission .	it is that the commission that the commission .
1000	die europäische union ist , dass die eu .	the european union is that the eu .
1200	im rahmen der welt , die sich in der welt der welt .	within the framework of the world which is in the world of the world .
1700	zwei monaten , die sich in den letzten jahren ist .	two months, which is in recent years .
2000	sechs monate von fünf monaten , die die aus der ersten zeit .	six months of five months, which the from the first time .
3000	sechs monate des bauwerkes , das ist , dass der kampf .	six months of the building, that is that the struggle .
4000	sechs monate der bauarbeiten , die brutalen .	six months of construction, the brutal .
6000	sechs monate bauwerke , das brutale .	six months of buildings, the brutal .

Figure 5: Translations at different steps during training. En-De.

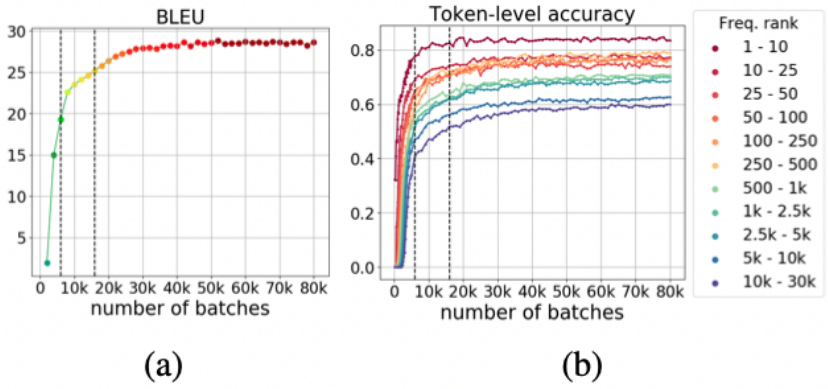

Figure 6: (a) BLEU score; (b) token-level accuracy (the proportion of cases where the correct next token is the most probable choice). WMT En-Ru.

Findings. The beginning of training is mostly devoted to target-side language modeling. We see huge changes in the LM scores (Figure 4a) with generated outputs performing better under simpler LMs (*e.g.*, 2-gram) than more complicates ones (*e.g.*, 5-grams), which shows that early in training translations tend to consist of frequent words and bigrams (but larger sequences are not necessarily fluent). In earlier iterations, all generated tokens are from the top-10 most frequent tokens, then only from the top-50, and only later less frequent tokens are starting to appear—see Figure 4b. Finally, early in training, the model hallucinates frequent n-grams (Figure 5).

Early in training, the model quickly improves its lexical choices, see Figure 6a for BLEU score on the development set during training and Figure 6b for token-level accuracy as a function of token frequency. We see that both the BLEU score and accuracy become large very fast, e.g. after the first 20k iterations (25% of the training process), the scores are already good. During the last half of the training, BLEU scores improve only by 0.5 and accuracy does not seem to change much, especially for rare tokens. Next we will discuss what happens in that phase.

In the second half of the training, the model is slowly refining translations, and, among the three competences we look at, the most visible changes are due to more complicated (i.e. less monotonic) reorderings—see Figure 7. The change in the fuzzy reordering score is only twice smaller than during the preceding stage. Moreover, the alignments keep changing and become less mono-

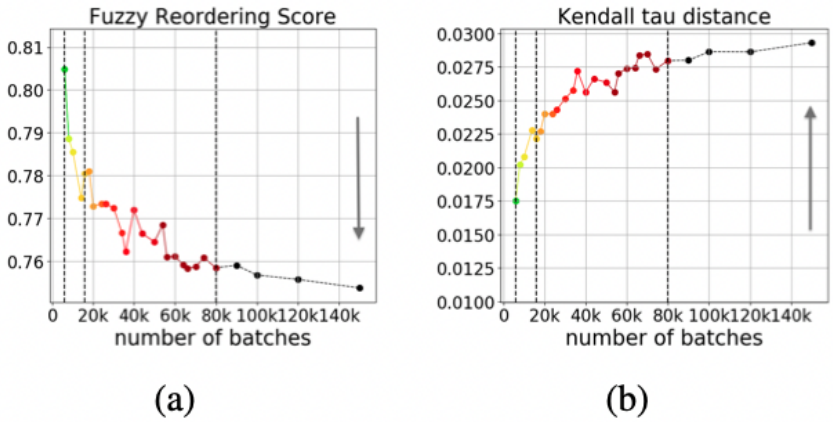

Figure 7: (a) fuzzy reordering score (for references: 0.6), (b) Kendall tau distance (for references: 0.06); WMT En-Ru. The arrows point in the direction of less monotonic alignments (more complicated reorderings).

Source: he was minister of defence from 1994 to 1995 and minister of agriculture and smus from 1995 to 1999 .

Model translations during training:

14k er war verteidigungsminister von 1994 bis 1995 und minister für landwirtschaft und smus von 1995 bis 1999 .

50k von 1994 bis 1995 war er verteidigungsminister und minister für landwirtschaft und smus von 1995 bis 1999 .

100k von 1994 bis 1995 war er verteidigungsminister und von 1995 bis 1999 minister für landwirtschaft und smus .

(a) En-De

Source: simple axis configuration for simultaneous processing of up to three tools , is the main feature of this machine .

Model translations during training:

14k простая ось для одновременной обработки до трех инструментов является основной характеристикой этой машины .

30k простая конфигурация осей для одновременной обработки до трех инструментов является основной функцией этой машины .

80k основная особенность этой машины - простая конфигурация осей для одновременной обработки до трех инструментов .

(b) En-Ru

Figure 8: Translations at different training steps. Same-colored chunks are approximately aligned to each other.

tonic even after both BLEU and token-level accuracy converged, i.e. iterations after 80k (Figure 7). Overall, we interpret this refinement stage as the model slowly learning to reduce interference from the source text and exacerbated even more in NMT: it learns to apply complex reorderings to more closely follow typical word order in the target language. This means that while language modeling improves more prominently during the first training stage, there is a long tail of less frequent and more nuanced patterns that the model learns later. This is additional evidence against using BLEU as a stopping criterion (Voita et al., 2019a). See examples of changes that happen in this last stage in Figure 8.

In summary, we show that during a large part of the training, the translation quality (e.g., BLEU) changes little, but the alignments become less monotonic. Intuitively, the translations become more complicated while their quality remains roughly the same. One way to directly apply our analysis is to consider tasks and settings where data properties such as regularity and/or simplicity are important, *e.g.* in data augmentation. For example, in neural machine translation, higher monotonicity of artificial sources was hypothesized to be a facilitating factor for back-translation;

additionally, complexity of the distilled data is crucial for sequence-level dis- tillation in non-autoregressive machine translation.

2.3 Improving Knowledge Distillation for NAT-MT

Non-autoregressive neural machine translation (NAT-MT; Gu et al., 2018) is different from the traditional NMT in the way it generates target sequences: instead of the standard approach where target tokens are predicted step-by-step by conditioning on the previous ones, NAT models predict the whole sequence simultaneously. This is possible only with an underlying assumption that the output tokens are independent from each other, which is unrealistic for natural language. Fortunately, while this independence assumption is unrealistic for real references, it might be more plausible for simpler sequences, e.g. artificially generated translations. That is why targets for NAT models are usually not references but beam search translations of the standard autoregressive NMT (which, as we already mentioned above, are simpler than references in many aspects). This is called sequence-level knowledge distillation (Kim and Rush, 2016), and it is currently one of the de-facto standard parts of the NAT-MT training pipelines (Gu et al., 2018; Zhou et al., 2020). Recently Zhou et al. (2020) showed that the quality of a NAT model strongly depends on the complexity of the distilled data, and changing this complexity can improve the model. Since distilled data consists of translations from a standard autoregressive teacher, our analysis of Section 2.2 suggests a very simple way of modifying the complexity of this data.

In (Voita et al., 2021b) we propose to use as teachers intermediate check-points during training, rather than a fully converged model, capitalising on the findings of Section 2.2.

Data. The dataset is WMT14 English-German (En-De) with newstest2013 as the validation set and newstest2014 as the test set, and BPE vocabulary of 37,000. We use the preprocessed dataset and the vocabularies released by Zhou et al. (2020).

Model. The NAT-MT model is the re-implemented by Zhou et al. (2020) version of the vanilla NAT by Gu et al. (2018). The teacher is the standard Transformer-base from `fairseq` (Ott et al., 2019). For the baseline distilled dataset, we use the fully converged model (in this case, the model after 200k updates). For other datasets, we use earlier check-points.

Findings. Since during a large part of training, NMT quality (e.g., BLEU) changes little, but the alignments become less monotonic, earlier checkpoints produce simpler and more monotonic translations, which in turn are better targets for NAT-MT. Figure 9c shows the BLEU scores for NAT models trained with distilled data obtained from different teacher's checkpoints; the baseline is the fully converged model (200k iterations). We see that by taking an earlier checkpoint, after 40k iterations, we improve NAT quality by 1.1 BLEU. For this checkpoint, the teacher's BLEU score is not much lower than that of the final model (Figure 9a), but the reorderings are much simpler (a higher fuzzy reordering score in Figure 9b).

3 Task 3.2 – Structured Sentence Models

Proposal highlights:

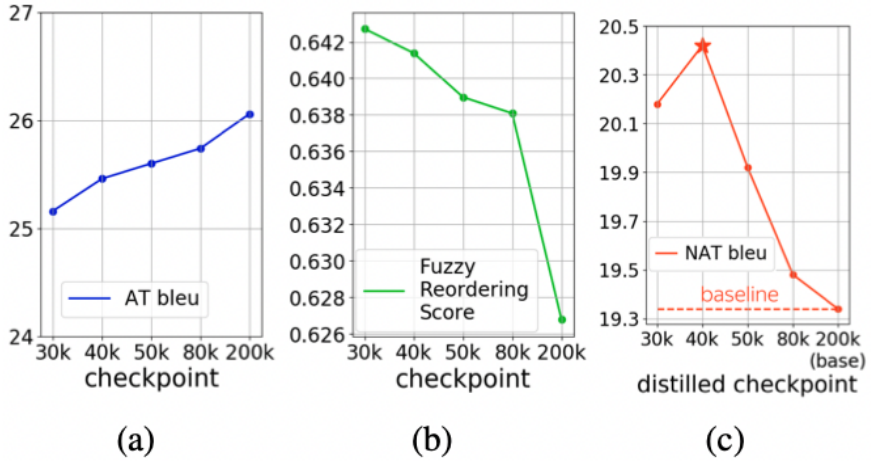

Figure 9: (a) BLEU score of the autoregressive (AT) Transformer-base (teacher for distillation); (b) fuzzy reordering score for the distilled training data obtained from checkpoints of the AT teacher; (c) BLEU scores for the vanilla NAT-MT model trained on different distilled data.

- we aim to develop NMT models that induce structured representations at sentence level (e.g., trees, graphs, latent factors);
- techniques to be investigate include discrete structure via REINFORCE, continuous relaxations, and iterative refinement;
- we will develop joint models representing structure of both source and target sentences, with the goal of achieving better data efficiency;
- we will exploit supervised tree banks for the resource-rich language (English in our case);
- as parallel data is scarce in the lower-resource setting, we will combine parallel and monolingual corpora.

This task focuses on learning structured combinatorial representations that can lead to improved generalisation and/or data efficiency. Examples include: sparse sentence and/or document embeddings, structured attention, syntactic trees and/or semantic graphs.

Summary of work done. We report progress in two fronts, one focused on learning sparse unobserved variables, this aims at advancing technology that will enable latent structure in deep models (including NMT), another focused on making use of context beyond the sentence level. In the first front, we (i) develop a theoretical framework for learning with mixed random variables unifying various previously introduced techniques, including techniques that we developed in the first half of the project, and also proposing novel ones, finally, this technique is applied to sparsify the encoder states of an NMT model; in the second front, we continue on the path started in the first half of the project, and take steps towards compact context-aware translation models.

3.1 Mixed Random Variables

Neural networks and other machine learning models compute continuous representations, while humans communicate mostly through discrete symbols. Reconciling these two forms of com-

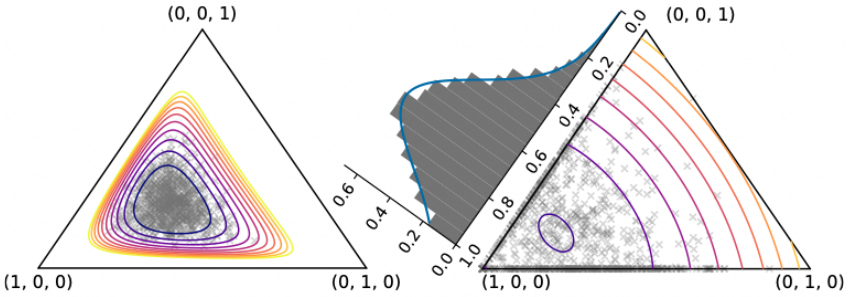

Figure 10: Multivariate distributions over the probability simplex. Standard distributions, like the Logistic-Normal (left), assign zero probability to all faces except to the relative interior of the simplex. Mixed distributions support assigning probability to the full simplex, including its boundary: the Gaussian-Sparsemax (right) induces a distribution over the 1-dimensional edges (shown as a histogram), and assigns $\Pr\{(1, 0, 0)\} = 0.022$.

munication is desirable for generating human-readable interpretations or learning discrete latent variable models, while maintaining end-to-end differentiability. Some existing approaches (such as the Gumbel-Softmax transformation (Jang et al., 2017; Maddison et al., 2017)) build continuous relaxations that are discrete approximations in the zero-temperature limit, while others (such as the Hard Concrete distribution (Louizos et al., 2018) and our own Hard Kumaraswamy distribution (Bastings et al., 2019)) produce discrete/continuous hybrids.

In (Farinhas et al., 2022), we build rigorous theoretical foundations for these hybrids, which we call "mixed random variables". Armed with a better theoretical understanding of these techniques, we extend them to the multivariate case, which was not known before (Figure 10 illustrates a simple example of a mixed random variable for sparse 3-dimensional probability vectors) and various feasible instances of it.

Methodology. Our starting point is a new "direct sum" base measure defined on the face lattice of the probability simplex. From this measure, we introduce new entropy and Kullback-Leibler divergence functions that subsume the discrete and differential cases and have interpretations in terms of code optimality. Our framework suggests two strategies for representing and sampling mixed random variables, an extrinsic ("sample-and-project") and an intrinsic one (based on face stratification). In the paper, we experiment with both approaches on an emergent communication benchmark (Lazaridou and Baroni, 2020), on modeling MNIST (LeCun et al., 2010) and Fashion-MNIST (Xiao et al., 2017) data with variational auto-encoders (Kingma and Welling, 2014) with mixed latent variables, as well as simplex-valued regression towards sparse voting proportions (Gordon-Rodriguez et al., 2020) showing results superior to techniques based on biased gradients (Jang et al., 2017), purely continuous relaxations (Maddison et al., 2017), and noisy policy gradients (Williams, 1992).

Discussion. Mixed random variables enable differentiable representations that retain a combinatorial inductive bias. Earlier instances in NLP have been used for rationale extraction (Bastings et al., 2019), analysis of NMT (Voita et al., 2019b) and other Transformer models (De Cao et al., 2020), they have, however, been limited to the 2-dimensional case which accounts for the inductive bias of a switch/selector (or a collection of independent switches/selectors). Our extensions are a

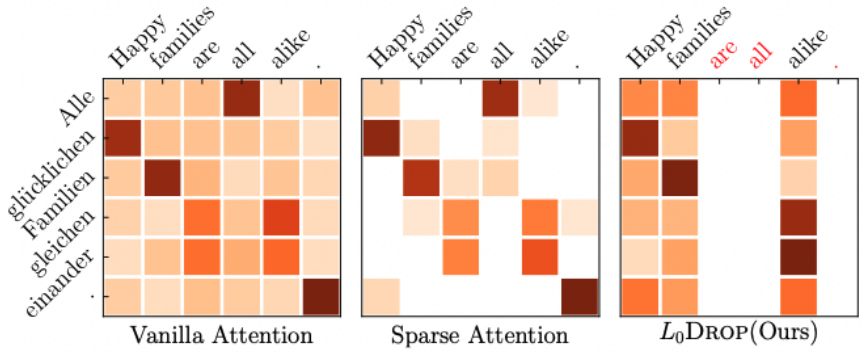

Figure 11: Encoder-decoder attention distribution of target words (vertical axis) over source words (horizontal axis) for vanilla attention (Vaswani et al., 2017), sparse attention (Correia et al., 2019) (which aims at sparsity with respect to individual decoder steps) and our model (Zhang et al., 2021). Darker color indicates larger attention weight, and the white blocks denote an attention weight of 0. The source words whose encoding is pruned by L0DROP (receiving zero weight) are highlighted in red.

pathway to other inductive biases such as categorical or structured variables with dependent parts.

3.2 Sparse Encoders

Sequence-to-sequence models usually transfer all encoder outputs to the decoder for generation. We hypothesize that these encoder outputs can be compressed to shorten the sequence delivered for decoding, in a way somewhat reminiscent of phrase-based models.

In (Zhang et al., 2021), we take Transformer models (Vaswani et al., 2017) as the testbed and introduce a layer of stochastic gates (which we term L_0DROP) in-between the encoder and the decoder. The gates are regularized using the expected value of a sparsity-inducing L0 penalty (Louizos et al., 2018; Bastings et al., 2019),[1] resulting in completely masking-out a subset of encoder outputs. In other words, via joint training, the L_0DROP layer forces the Transformer to route information through a subset of its encoder states. We investigate the effects of this sparsification on two machine translation and two summarization tasks.

Approach. We aim at detecting uninformative source encodings and dropping them to shorten the encoding sequence before generation. To this end, we build on recent work on sparsifying weights (Louizos et al., 2018) and activations (Bastings et al., 2019) of neural networks. Specifically, we insert a differentiable neural sparsity layer (L_0DROP) in-between the encoder and the decoder. The layer can be regarded as providing a multiplicative scalar gate for every encoder output. The gate is a random variable and, unlike standard attention, can be exactly zero, effectively masking out the corresponding source encodings. The sparsity is promoted by introducing an extra term to the learning objective, i.e. an expected value of the sparsity-inducing L0 penalty. By varying the coefficient for the regularizer, we can obtain different levels of sparsity. Importantly, the objective remains fully end-to-end differentiable. Given an encoding sequence of length N, the vanilla attention model attends to it recurrently for M steps at the decoding phase, leading to a computational complexity of $O(NM)$. This could be costly if N or M is very large. With the

[1] This technique is built upon a 2-dimensional special case of our mixed random variables of Section 3.1.

induced sparse structure by L_0DROP, we introduce a specialized decoding algorithm which lowers this complexity to $O(N'M)$ with $N' \leq N$. As a result, L_0DROP can improve decoding efficiency by reducing the encodings' length, especially for long inputs. See Figure 11 for an illustration of L_0DROP and how it differs from existing work.

Datasets. We conduct extensive experiments on WMT translation tasks with two language pairs (WMT14 English-German (Bojar et al., 2014) and and WMT18 Chinese-English (Bojar et al., 2018)) and document summarization tasks (CNN/Daily Mail (Hermann et al., 2015) and WikiSum (Liu* et al., 2018)). We adopt BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) to evaluate the translation and summarization quality, respectively.

Findings. Our main findings are summarized as follows:

- We confirm that the encoder outputs can be compressed, around 40–70% of them can be dropped without large effects on the genera- tion quality.
- The resulting sparsity level differs across word types, the encodings corresponding to function words (such as determiners, prepositions) are more frequently pruned than those of content words (e.g., verbs and nouns).
- L_0DROP can improve decoding efficiency particularly for lengthy source inputs. We achieve a decoding speedup of up to 1.65x on document summarization tasks and 1.20x on character-based machine translation task. Filtering out source encodings with rule-based sparse patterns is feasible, and confirms information-theoretic expectations, although rule-based patterns do not generalize well across tasks.

3.3 Exploiting Context Beyond Sentence Level

In a realistic scenario, an end user is interested in translating *documents*, or some other form of coherent excerpt of text. Given the richness of linguistic phenomena going on in translation already at the sentence level (arguably even within clauses), it is understandable why so much research focuses on independent translation of (shorter) segments such as sentences. The growing need for translation in applied settings where context is crucial is pushing the community to look into solutions to this modelling challenge (Wang et al., 2017; Miculicich et al., 2018; Voita et al., 2018; Zheng et al., 2020). Standard NMT factorises the probability of a document $\mathcal{D} = \langle (x^{(1)}, y^{(1)}), \ldots, (x^{(S)}, y^{(S)}) \rangle$

$$p(\mathcal{D}|\theta) = \prod_{s=1}^{|\mathcal{D}|} p(y^{(s)}|x^{(s)}, \theta) \tag{1}$$

as if sentence pairs in $\mathcal{D}$ were independent of one another. So called *document-level* NMT models the probability of a document $\mathcal{D}$ as follows:

$$p(\mathcal{D}|\theta) = \prod_{s=1}^{S} p(y^{(s)}|x^{(1:S)}, y^{(1:s-1)}, \theta) \,, \tag{2}$$

where $y^{(1:s-1)}$ corresponds to the history of already generated target sentences and $x^{(1:S)}$ is the entire source document. Crucially, document-level (or context-aware) NMT models the document

without making the independence assumptions of standard sentence-level NMT. The model likelihood given a dataset of document pairs then factorises independently over documents (but not independently over sentences), which is far more reasonable. This does require parallel data annotated with document alignments which poses some challenge in the low-resource setting. Moreover, architectures that can condition on information beyond the sentence boundary are typically larger requiring more parameters to be estimated and thus more data. In the first half of the project, we aimed at addressing limitations of document-level NMT more generally including model design, evaluation, and impact of finer-grained document-level annotation. In the second half of the project we attempted to design architectures that exploit context without requiring extending the memory mechanism of a Transformer decoder to accommodate the entire history of sentences.

Methodology. We investigate a combination of prefix tuning (Li and Liang, 2021) and variational NMT (Zhang et al., 2016) which allows us to i) conditionl on large context efficiently and ii) reuse pre-trained NMT components. We start with a setence-level NMT component $p(y^{(s)}|x^{(s)}, \theta)$ and extend it to condition on a prefix embedding $z^{(s)}$, which we pre-prend to the sth target sentence $y^{(s)}$ in the document before sentence-level decoding. This embedding is drawn from a Gaussian distribution

$$\mathcal{N}(\mu(\mathbf{x}, \mathbf{v}_s; \phi), \sigma^2(\mathbf{x}, \mathbf{v}_s; \phi))$$

whose parameters depend on the complete source-language document and the history of already generated sentences through a compact bag-of-words representation of $x^{(1:S)}$, denoted by $\mathbf{x}$, and $y^{(1:s-1)}$, denoted by $\mathbf{v}_s$. Clearly, a Transformer-based encoding of $x^{(1:S)}, y^{(1:s-1)}$ would require processing an ever growing sequence of inputs, we use a bag of words representation instead in order to keep our models compact. Finally, we estimate the parameters λ by optimising the evidence lowerbound (ELBO):

$$\mathbb{E}_{q(z^{(s)}|\mathbf{x},\mathbf{y},\lambda)}\left[\sum_{s=1}^{S} \log p(y^{(s)}|x^{(s)}, z^{(s)}, \theta) + \log p(z^{(s)}|\mathbf{x}, \mathbf{y}, \phi) - \log q(z^{(s)}|\mathbf{x}, \mathbf{y}, \lambda)\right] \tag{3}$$

where $q(z^{(s)}|\mathbf{x}, \mathbf{y}, \lambda)$ is a Gaussian variational approximation to the model's posterior distribution over latent codes, which additional conditions on a bag-of-words representation of the target document $y^{(1:S)}$, denoted $\mathbf{y}$. Importantly, we optimise the ELBO with respect to the Gaussian parameters ϕ and λ, while leaving the NMT parameters θ fixed. The embedding z learns to control the sentence-level component adapting it to the context-level setting, it also informs the decoder of shallow document-level features captured by the bag-of-words representation of the documents. For training, we use reparameterised gradients (Kingma and Welling, 2014) as in variational neural machine translation (Zhang et al., 2016; Eikema and Aziz, 2019).

Findings. We use a Marian NMT (Junczys-Dowmunt et al., 2018) pre-trained component and investigate the effect of training it using the 2to2 objective of Fernandes et al. (2021) as well as our proposed approach. Note that 2to2 optimises a document-level NMT loss while translating essentially two adjacent sentences at a time, this requires extending the prefix of a target sequence with the complete previous sentences. Our approach on the other hand, extends the prefix of Marian with a single state and leaves Marian's parameters untouched. Table 1 show preliminary results on IWSLT14 English-German, here a document is an entire TED talk, which we can condition on efficiently due to our compact latent document embedding model. We probed for improvements along the pronoun dimension using a contrastive evaluation set (Müller et al., 2018). Our approach

Model	BLEU	COMET
NMT	32.63	0.4766
2to2	34.50	0.5259
latent control	34.60	0.5277

Table 1: Translation performance of context-aware MT on IWSLT14 English-German (test).

does not lead to improvements along that dimension, which is somewhat expected given the bag-of-words view is unlikely to help with coreferences. The nature of our document embeddings is more aligned with improvements along lexical cohesion (Bawden et al., 2018), but analysing the model in this dimension is ongoing work.

4 Task 3.3 – Probabilistic Neural Machine Translation

Proposal highlights:

- Revise decision rules in NMT to exploit NMT models as probability distributions. Here we seek to make predictions with a holistic view of the model's beliefs.
- Introduce global statistics to decision rules. This may take the form of n-gram statistics, and other edit operations sensitive to insertion, substitution, and word order differences. This can also accommodate document-level statistics.
- Make use of Bayesian modelling techniques to improve the data efficiency of NMT models. This may take the form of Bayesian priors in parameter and/or function space. Changes to the way NMT factorises the probability of observations with the goal of better uncertainty management and increased data efficiency are also relevant.

Summary of work done. In the first half of the project we uncovered evidence that the inadequacy of high-scoring translations in NMT (Koehn and Knowles, 2017; Murray and Chiang, 2018b; Stahlberg and Byrne, 2019) is not on its own indicative of failure of the model to capture essential properties of translations, but rather a predictable property of Markov processes. In the second half of the project we expand on this point, finding stronger arguments and evidence (Section 4.1), and advance our algorithms for approximate minimum Bayes risk decoding (Section 4.2), making them more scalable and accurate (Sections 4.3 and 4.4). We also investigate so-called loss-calibrated objectives (Section 4.5), in an attempt to optimise the training algorithm for compatibility with the decoding algorithm.

Background

NMT employs neural networks (NNs) to predict a conditional probability distribution $Y|\theta, x$ over translation candidates (*i.e.*, all finite-length sequence of target-language symbols) of any given source sentence x. To predict such a complex object efficiently, NMT factorises the distribution as a chain of random draws from Categorical distributions

$$Y_j|\theta, x, y_{<j} \sim \text{Categorical}(f(x, y_{<j}; \theta)) \tag{4}$$

parameterised in context. The prefix translation $y_{<j}$ starts empty and grows one symbol at a time until a special end-of-sequence symbol is drawn. At each step j, f maps from varying inputs $(x, y_{<j})$ to a probability distribution over the vocabulary. Common choices for f include recurrent networks (Sutskever et al., 2014; Bahdanau et al., 2015), convolutional networks (Gehring et al., 2017), graph convolutional networks (Bastings et al., 2017), and Transformers (Vaswani et al., 2017). Given a dataset $\mathcal{D}$ of translation pairs, the NN parameters θ are estimated to attain a local optimum of the regularised log-likelihood function

$$\theta^{\text{MLE}} = \arg\max_{\theta} \; \mathbb{E}_{(x,y)\sim\mathcal{D}}[\log p_{Y|X}(y|x,\theta)] - \mathcal{R}(\theta), \tag{5}$$

via stochastic gradient back-propagation (here $\mathcal{R}$ is a regulariser). This procedure approximates maximum likelihood estimation (MLE), which aims to mimic the unknown distribution of the training data.

After training, and for a given input, choosing a translation requires a *decision rule* to map from a distribution over translation candidates to a single 'preferred' translation. The most common decision rule in NMT is maximum-a-posterior (MAP) decoding, which outputs the most probable translation under the model (*i.e.*, mode of the conditional distribution):

$$y^{(\text{MAP})} = \arg\max_{h\in\mathcal{Y}} \; \log p_{Y|X}(h|x,\theta)\,. \tag{6}$$

As this is intractable, beam search (Graves, 2012) is used. Beam search is a pruned version of breadth-first search which maintains an active set of k partial translations. For large beam size k, translation quality degrades (Koehn and Knowles, 2017) and the exact $y^{(\text{MAP})}$ is often the empty sequence (Stahlberg and Byrne, 2019). Therefore, in practice, the beam size is kept small and the objective is length normalised to up-rank longer hypotheses (Murray and Chiang, 2018a). Despite the widespread intuition that MAP decoding is an obvious choice, MLE is oblivious to our desire to form predictions using the MAP decoder (or any decoder, for that matter).

4.1 Inadequacy of the Mode in Markov Processes

Certain Markov processes are stationary and ergodic, these two properties together lead to a result known as asymptotic equipartition property (AEP; Gray, 2011). When an AEP holds for a Markov process, the surprisal (negative log probability) of a trajectory sampled from the process is within a margin of the entropy rate of the process, this is true almost surely as the length of the trajectory increases. The set of such samples, which are the samples we are bound to observe if we just run the process, is known as the *typical set*. In simple terms, the AEP tells us that trajectories drawn from a Markov process show a concentration towards average surprisal: they are never too suprising (*i.e.*, we will not sample trajectories whose probabilities are the lowest nor the highest). An AEP for NMT would have strong formal implications. For example, if NMT is a Markov process that does not generate extreme outcomes, an objective such as MLE (Equation 5) will lead to reference translations receiving typical probability, rather than maximum probability. This would cast doubt on MAP decoding and its approximations at a rather formal level.

NMT's generative story does prescribe a Markov process for each source sentence x. We sketch it here:

- Start from a initial state $(x, \triangleright)$ that stores a transformation of the source and a special begin-of-sequence symbol $\triangleright$ (*i.e.*, an empty translation prefix), from there an NN computes a V-dimensional vector $f(x, \triangleright; \theta)$ of transition probabilities.[2]

- With probability $0 < f_a(x, \triangleright; \theta) < 1$ we move to state $(x, \triangleright\, a)$. At this point, we use the same NN to compute a V-dimensional vector $f(x, \triangleright\, a; \theta)$ of transition probabilities, and this time we move to state $(x, \triangleright\, a\, b)$ with probability $f_b(x, \triangleright\, a; \theta)$.

- This goes for each step j: the state of the system stores a transformation of $x, y_{<j}$, the V transition probabilities $f(x, y_{<j}; \theta)$ are predicted in context, and the system moves on to state $f(x, y_{<j}\, w; \theta)$ by drawing a word w with probability $0 < f_w(x, y_{<j}; \theta) < 1$.

- Wheneve we draw a special end-of-sequence symbol $\triangleleft$, we reset the state to $(x, \triangleright)$. The sequence of symbols drawn between two visits to state $(x, \triangleright)$ is interpreted as a translation candidate.

It is clear, by construction, that a translation $y_{1:J}$ is drawn from this Markov process with probability given by $\prod_{j=1}^{J} f_{y_j}(x, y_{<j}; \theta)$, hence this Markov process has as invariant measure the distribution whose probability mass function is shown in Equation (4). The construction above tells us two things. First, that we can draw independent samples from NMT via a simple procedure that takes linear time in sequence length—this procedure is known as ancestral sampling (Robert and Casella, 2010). Second, that the state of NMT is not limited in any way, until, of course, its reset when we hit the end-of-sequence symbol. The latter means that NMT is a non-ergodic Markov process, hence an AEP cannot be trivially stated and proven for NMT. While it is hard to prove that an AEP holds for NMT, it is easy to inspect whether NMT systems exhibit effects that are consistent with it.

Methodology. We analyse the systems developed for the studies reported in (Eikema and Aziz, 2020), these are Transformer base models covering three language pairs with varying amount of resources for training: English into and from German, Nepali and Sinhala. For German-English (de-en) we use all available WMT'18 (Bojar et al., 2018) news data except for Paracrawl, resulting in 5.9 million sentence pairs; for the other two pairs we use the data setup by Guzmán et al. (2019). For a validation set, we draw 1,000 samples from the model and observe that:

- the distribution of surprisals indeed shows behaviour consistent with an AEP;

- the surprisal of reference translations is close to that of the average suprisal (in a thousand samples)—see an example in Figure 12;

- beam search outputs are seldom amongst sampled translations, and are clear outliers in surprisal plots (true for more than 50% of instances in high-resource pair, and more than 90% of instances in low-resource pairs).

[2] We use V for target vocabulary size.

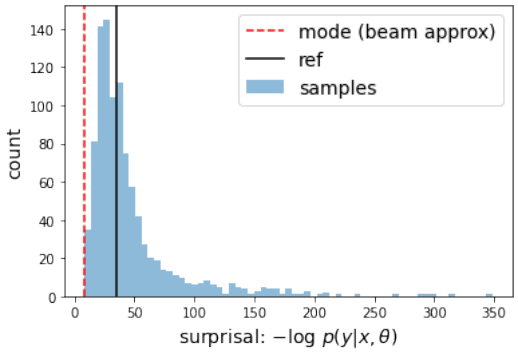

Figure 12: Distribution of surprisal based on 1,000 samples from the model. Source: *Der Großbrand in einem als besonders gefährlich geltenden Chemiewerk in der nordfranzösischen Stadt Rouen ist gelöscht.* Reference: *The large fire was put out at a chemical plant considered to be particularly hazardous and located in the northern French city of Rouen.* Beam search output: *The big fire in a particularly dangerous chemical plant in the northern French town of Rouen has been extinguished.*

Discussion. These observations are significant for they explain why the reference translation is assigned lower probability than beam outputs. Prior literature has interpreted this as a flaw in model design or problem with parameter estimation. An alternative explanation is that nothing about NMT design and training gives the mode special status, and that, to date, mode-seeking search has been employed not for its obvious plausibility but simply following an intuition that need not hold for Markov processes (namely, that the most probable translation is in any sense representative of the model's beliefs). Some of these results have been discussed in our earlier paper (Eikema and Aziz, 2020), which received the best paper award at Coling. An extensive followup is in preparation at the time of writing.

4.2 Sampling-Based MBR Decoding

Minimum Bayes risk (MBR) decoding stems from the principle of maximisation of expected utility (Berger, 1985). A utility function $u(y, h)$ measures the benefit in choosing $h \in \mathcal{Y}$ when $y \in \mathcal{Y}$ is the ideal decision. When forming predictions, we lack knowledge about ideal translations and must decide under uncertainty. MBR lets the model fill in 'ideal decisions' probabilistically as we search through the space of candidates for the one which is assigned highest utility *in expectation*:

$$y^{(\text{MBR})} = \arg\max_{h \in \mathcal{Y}} \underbrace{\mathbb{E}[u(Y, h) \mid \theta, x]}_{=:\mu_u(h;x,\theta)} . \tag{7}$$

MBR has a long history in parsing (Goodman, 1996; Sima'an, 2003), speech recognition (Stolcke et al., 1997; Goel and Byrne, 2000), and MT (Kumar and Byrne, 2002, 2004), including, more recently, neural machine translation (Stahlberg et al., 2017; Blain et al., 2017; Eikema and Aziz, 2020). In MT, u can be a sentence-level evaluation metric (*e.g.*, METEOR (Lavie and Agarwal, 2007), Sentence BLEU (Chen and Cherry, 2014)), BEER (Stanojević and Sima'an, 2014), BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), etc.). Intuitively, whereas the MAP prediction is the translation to which the model assigns highest probability, no matter how idiosyncratic, the MBR prediction is the translation that is closest (under the chosen u) to all other probable translations. Seeking support for a prediction not only in terms of probability but also

in terms of utility makes MBR decoding robust to situations where inadequate translations are assigned high probability, as it often happens with the empty string (Stahlberg and Byrne, 2019), when the training data are noisy (Ott et al., 2018), too small (Eikema and Aziz, 2020) or distant from the test domain (Müller and Sennrich, 2021).

In (Eikema and Aziz, 2021), we develop Monte Carlo approximations to MBR and study their properties.

Methodology. Like in MAP decoding, exhaustive enumeration of the hypotheses is impossible, we must resort to a finite subset $\mathcal{H}(x)$ of candidates. Unlike MAP decoding, the objective function $\mu_u(h; x, \theta)$ *cannot* be evaluated exactly. Most approximations to MBR decoding, from Kumar and Byrne (2004) to recent instances (Stahlberg et al., 2017; Shu and Nakayama, 2017; Blain et al., 2017), use k-best lists from beam search for $\mathcal{H}(x)$ and to form a biased estimate of expected utility. In Eikema and Aziz (2020) we use unbiased samples from the model for both approximations: i) we follow the generative story in Equation (4) to obtain N independent samples $y^{(n)}$, a procedure known as ancestral sampling (Robert and Casella, 2010); then, ii) for a hypothesis h, we compute an MC estimate of $\mu_u(h; x, \theta)$:

$$\hat{\mu}_u(h; x, N) \overset{\text{MC}}{\coloneqq} \frac{1}{N} \sum_{n=1}^{N} u(y^{(n)}, h) \,, \tag{8}$$

which is unbiased for any sample size N. In Eikema and Aziz (2020) use the same N samples as candidates and approximate Equation (7) by

$$y^{(\text{MC})} \coloneqq \underset{h \in \{y^{(1)}, \ldots, y^{(N)}\}}{\arg\max} \; \hat{\mu}_u(h; x, N) \,. \tag{9}$$

We call this class of MBR algorithms using unbiased MC estimation instances of *sampling-based MBR decoding*.

Data. We perform experiments on three language pairs with varying amount of resources for training: English into and from German (Bojar et al., 2018), Romanian (Bojar et al., 2016a) and Nepali Guzmán et al. (2019). We train a Transformer base model (Vaswani et al., 2017) until convergence and average the last 10 epoch checkpoints to obtain our final model. In all models we disable label smoothing, as this has been found to negatively impact model fit, which would compromise the performance of MBR (Eikema and Aziz, 2020). Complete details in (Eikema and Aziz, 2021). For computational efficiency, we opt for non-neural evaluation metrics for use as utility function in MBR. BEER (Stanojević and Sima'an, 2014) is a non-neural trained metric that has shown good correlation with human judgements in previous WMT metrics shared tasks (Bojar et al., 2016b).

Findings. We find that MBR steadily improves across language pairs as N grows larger, see Figure 13. SacreBLEU (Post, 2018) scores improve at a similar rate to that of BEER, showing no signs of overfitting to the utility. This is strong empirical evidence that *sampling-based* MBR has no equivalent to the beam search curse. We see this as an important property of a decoding objective.

The candidate set in our approximation do not need to be obtained using ancestral sampling, in fact, ideally, they would come from a strategy that is biased towards enumerating outcomes with

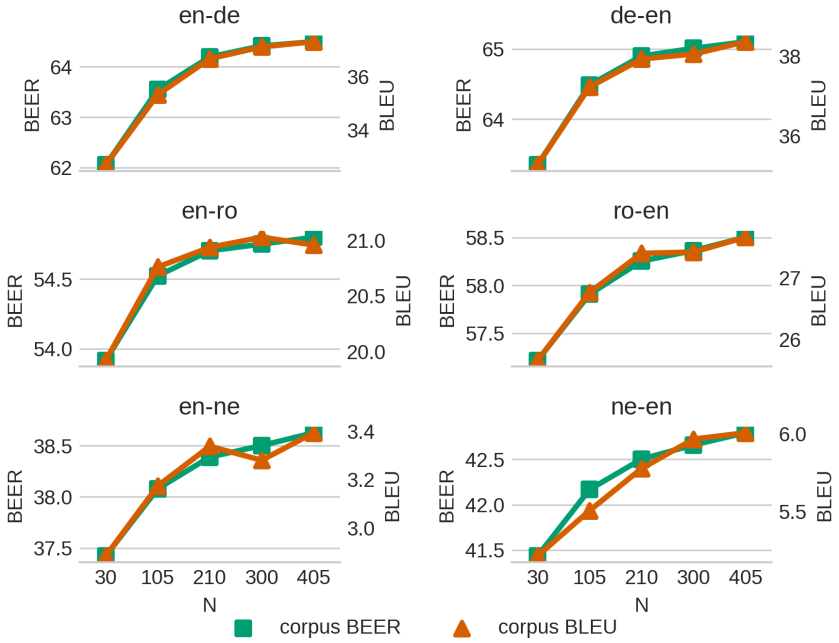

Figure 13: Quality of sampling-based MBR output for various sizes of N using BEER as target utility. We report both BEER and BLEU scores.

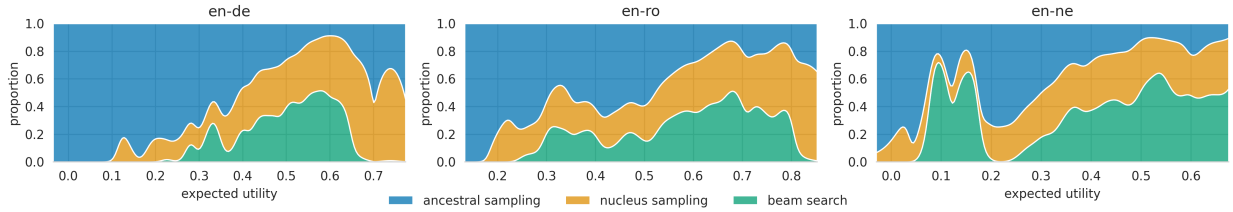

Figure 14: Proportion plots of expected utility for 3 strategies for constructing $\mathcal{H}(x)$, using 100 translation candidates per strategy. We estimate expected utility using 1,000 samples. Results are aggregated over 100 source sentences.

high expected utility first. As no such algorithm is currently available, we explore other strategies that are known to perform well for NMT, namely, nucleus sampling (Holtzman et al., 2020) and beam search. We compare each strategy by the expected BEER values of the translations generated, using accurate estimates of expected BEER (using 1,000 samples for MC estimation) as this shows us which candidate set has higher potential—see Figure 14. We find ancestral sampling to produce hypotheses across the entire range of expected BEER scores. Nucleus sampling and beam search generally produce translations at the higher end of expected BEER. Therefore, these seem more suitable for generating effective $\mathcal{H}(x)$ at smaller N. Nucleus sampling seems to lead to the largest proportion of high expected utility translations across language pairs. Beam search has a noticeably high proportion of poor translations for English-Nepali, a low-resource language pair where mode-seeking search has been observed to be less reliable. Results in the opposite direction were similar.

While we can choose which strategy to use for enumerating candidates, it is important to use ancestral samples for estimation of expected utility of those candidates, for only ancestral sampling is faithful to the model distribution by design and hence yields unbiased estimates. In Figure 15 we illustrate this claim empirically by showing the bias of two alternative strategies: nucleus sampling

(Holtzman et al., 2020) and 'beam sampling' (*i.e.*, using k-best outputs from beam search for estimating expected utility; Blain et al. (2017)).

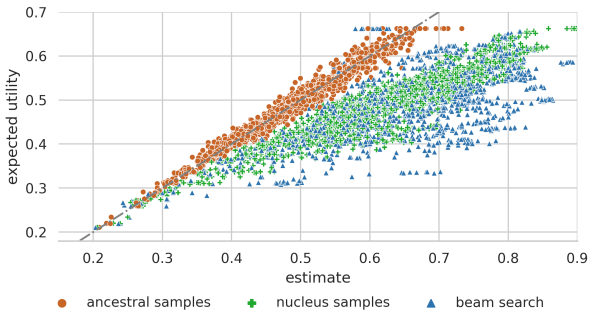

Figure 15: Estimates of expected utility for various hypotheses. We plot practical estimates of expected utility (x-axis) using either ancestral, nucleus or 'beam' samples against an accurate MC estimate using 1,000 ancestral samples. The gray line depicts a perfect estimator.

4.3 Scalable MBR Approximations

Our original algorithm (Eikema and Aziz, 2020) couples two approximations, namely, tractable exploration and unbiased estimation of expected utility are based on the same N ancestral samples. This leads to a large number of assessments of utility, which prevents exploration of even larger hypothesis spaces. Our aim is to learn more about the impact of these two approximations, for which we look into estimating expected utility using fewer (S) samples. We call this approximation MBR-N-by-S. With $N \times S$ assessments of utilities per decoding, rather than $N \times N$, we can also investigate a larger hypothesis space $\mathcal{H}(x)$.

In (Eikema and Aziz, 2021), we explore N (number of candidates) ranging from 210 to 1005, while keeping S (the number of samples used for approximating expected utility of each hypothesis) smaller, with S ranging from 10 to 200. We argue that S does not need to grow at the same pace as N, as MC estimates should stabilize after a certain point. We find that growing N beyong 405 improves translation quality further, even when the estimates of expected utility are less accurate, see Figure 16. Increasing S also steadily improves translation quality, with diminishing returns in the magnitude of improvement. On the other hand, smaller values of S lead to notable deterioration of translation quality and we note higher variance in results. For all language pairs it is possible to improve upon the best MBR-N-by-N results by considering a larger hypothesis spaces and smaller S. This experiment shows that the two approximations can be controlled independently and better results are within reach if we explore more. On top of that, the best setting of MBR-N-by-N takes 164,025 utility assessments per decoding, MBR-N-by-S with $S = 100$ brings this number down to 100,500 for the largest N considered, while improving BEER scores on all language pairs. We note that again increasing either N or S generally improves translation quality in our experiments. This further strengthens our previous finding that sampling-based MBR does not seem to have an equivalent of the beam search curse. We have also developed other approximations aimed at scaling the algorithm up to very large hypothesis spaces, extensive experiments are reported in a pre-print (Eikema and Aziz, 2021).

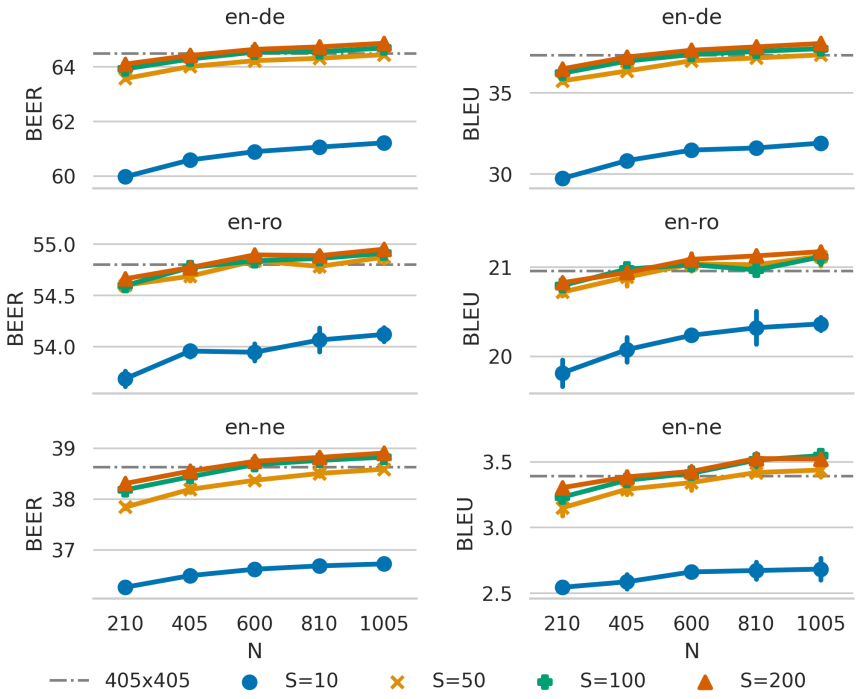

Figure 16: Performance of MBR-N-by-S: we estimate the expected utility of N hypotheses using S samples. We show average performance over 3 runs with 1 standard deviation. The dashed line shows the performance at $N = S = 405$.

4.4 Sample-Efficient Approximations of Expected Utility

The most obvious estimate of expected utility one can obtain is a Monte Carlo estimate, that is, the sample mean utility against a number of independent samples from the model. The potential issue with that is that the quantity used to rank translation candidates is therefore a random variable (*i.e.*, the Monte Carlo estimate). For small S, the variance of this estimate is too high leading to unrealiable ranking. Growing S to reduce variance has two disadvantages: i) we need to sample more from the model, which is costly, and ii) we need to perform more assessments of the utility function, which can be quite expensive for modern neural utility functions such as BLEURT and COMET. Moreover, the variance of MC estimation decays slowly as a function of S, more precisely, it decays only with the squared-root of S. Here we investigate estimators that are potentially more sample-efficient than MC.

Methodology. An alternative to MC estimation is the so-called Bayesian Monte Carlo (BMC) estimators (Rasmussen and Ghahramani, 2003). In BMC, we treat expected utility as a latent variable drawn from a Gaussian process (GP) prior that ties a collection of N hypotheses together. We then observe a number of assessments of utility for each hypothesis against a small number S of samples (for example, $S = 1$ sample per hypothesis). These assessments are assumed to be drawn from a Normal distribution whose mean (expected utility) is an N-dimensional latent variable drawn from the GP. Next, we infer the posterior distribution over the unknown expected utilities in closed-form (since the GP-Normal posterior is a multivariate Gaussian with known mean and covariance). The GP exploits similarity between hypotheses to achieve greater variance reduction than standard MC. Intuitively, whereas in MC we observe S samples per hypothesis and

estimate the expected utility of each hypothesis independently, in BMC we observe S samples per hypothesis and estimate their expected utilities jointly assuming that their estimates correlate (e.g., because the hypotheses themselves are similar or the samples against which we compute utilities are similar). Concretely, we design the following BMC procedure:

$$\mu_1, \ldots, \mu_N \sim \mathcal{N}(\mathbf{m}, \Sigma) \tag{10}$$

$$\text{for } n = 1, \ldots, N \tag{11}$$

$$u(y^{(s)}, h^{(n)}) \sim \mathcal{N}(\mu_n, \sigma^2) \quad \text{for } s = 1, \ldots, S\ . \tag{12}$$

In words:

- we draw the expected utilities for all N hypotheses jointly from a multivariate Gaussian with prior mean $\mathbf{m}$ and prior covariance matrix Σ (we will explain how those are set later);
- then, for each hypothesis $h^{(n)}$, we draw its utility against a sample $y^{(s)}$ from a Normal distribution with mean μ_n and a fixed variance σ (shared across hypotheses).

We fix the prior mean and the observed variance empirically as to have prior predictive samples cover a reasonable range of values (*i.e.*, reasonable for the utility under consideration). The N-by-N covariance matrix Σ is specified through a kernel function (we use the rbf kernel) and a feature representation of the hypotheses. For features we use a vector of word counts (*i.e.*, a bag of words representation of the hypothesis). Let $\mathbf{U}$ denote $N \times S$ observed utilities values (one per hypothesis per sample), the posterior distribution over expected utility $\mu_1, \ldots, \mu_N | \mathbf{U} \sim \mathcal{N}(\mathbf{p}, \mathbf{C})$ is a known Gaussian whose parameters can be computed in closed-form (Rasmussen and Ghahramani, 2003) in time $O(N^3)$, which is reasonable on CPUs for N in the order of hundreds, and on GPU can be extended to thousands.[3]

Findings. Figure 17 assesses the potential of this estimator by comparing it to a robust and expensive MC estimate of expected utility (computed using 1,000 samples). In this demonstration, we aim to rank 100 candidates drawn via ancestral sampling for a portion of the German-English dev set. The figure compares BMC using $S < 1000$ against an S-samples MC estimate. It is clear that BMC exploits correlations very effectively. At only 1 sample per hypothesis the BMC error is as good as MC using 75 samples. A complete evaluation of this approach within an MBR procedure is ongoing work.

4.5 Loss-Calibrated Machine Translation

Loss-calibrated Bayes (Lacoste-Julien et al., 2011) couples the inference problem (i.e., learning from data) to the decision-making algorithm for predictions (this is what the term 'loss' in 'loss-calibrated' refers to, and it is different from the traditional 'training loss' in NMT). The principal claim of the framework is that we can learn to make inferences that are better suited to minimise the decision-making loss. Our notion of decision-making loss (rather, gain, as prefer to express our preferences in terms of utility rather than loss functions) is the same as in minimum Bayes risk decoding, that is, expected utility. Let $h^\star$ denote the optimum decision from MBR:

$$h^\star = \operatorname*{argmax}_{h \in \mathcal{Y}} \mathbb{E}[u(Y, h)|\theta] \tag{13}$$

[3] Moreover, we can exploit scalable GP implementations such as sparse GPs (Snelson and Ghahramani, 2007) if we need even larger hypothesis spaces.

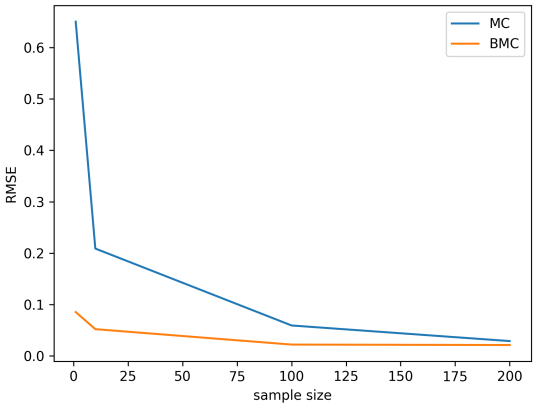

Figure 17: Performance of Bayesian Monte Carlo against Monte Carlo estimates of expected utility as a function of sample size S (horizontal axis). We measure mean-squared error (vertical axis) against a very robust estimate of expected utility obtained using 1,000 independent samples. The blue curve is the error of MC using S samples, the orange curve is the error of BMC using S samples.

loss-calibrated inference regularises the objective for parameter estimation with the so-called q-gain function:

$$\mathcal{G}(h^\star|\lambda) = \int q(\theta|\lambda)\mathbb{E}[u(Y, h^\star)|\theta]\mathrm{d}\theta \tag{14}$$

where $q(\theta)$ is an approximation to the model's true posterior. Intuitively, this regulariser seeks posterior inferences (captured by $q(\theta)$) which lead to the MBR optimum accumulating high expected gain. The total objective is as follows:

$$\underset{\lambda}{\operatorname{argmax}}\ \mathbb{E}_{q(\theta|\lambda)}[\log p(y|x,\theta)] - \mathrm{KL}(q(\theta|\lambda)||p(\theta)) + \mathcal{G}(h^\star|\lambda) \tag{15}$$

In a Bayesian neural network, this q-gain can be estimated via, for example, Monte Carlo dropout (Gal and Ghahramani, 2016), which is the technique we investigate. The gain function (14) is intractable in a number of ways, requiring principled and efficient approximations. We now discuss how we approach these approximations.

Methodology. Let's start with the *intractable posterior expectation.* Suppose we know the optimum decision $h^\star$, assessing the gain function still requires an intractable expectation: average expected utility $\mathbb{E}[u(Y, h^\star)|\theta]$ across all possible assignments of the model parameters θ. There are infinitely many possible assignments for θ, but we can use a finite-time unbiased estimate, for example by performing K stochastic forward passes with MC dropout:

$$\mathcal{G}(h^\star|\lambda) \overset{\text{MC}}{\approx} \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}[u(Y, h^\star)|\theta^{(k)}]\ . \tag{16}$$

Next, we address the *intractable expected utility.* Suppose we know the optimum decision $h^\star$ and have a specific assignment of the model parameters $\theta^{(s)}$, for example, in one of the forward passes of MC dropout. It remains intractable to assess the expected utility of $h^\star$, as we had already

discussed when we introduced MBR. The same solution applies here, we obtain a Monte Carlo estimate using S samples from the NMT component:

$$\mathbb{E}[u(Y, h^\star)|\theta^{(k)}] \overset{\text{MC}}{\approx} \frac{1}{S}\sum_{s=1}^{S} u(y^{(s)}, h^\star) \quad \text{where } y^{(s)} \sim p_{Y|X}(\cdot|x, \theta^{(k)})\,. \tag{17}$$

Finally, let's turn to the *intractable search.* Even though we can approximate expectations as shown above (both under a given assignment of θ and under the posterior distribution of the Bayesian NMT model), we cannot actually solve the search for the optimum $h^\star$. Instead, we rely on an extension of our sampling-based approximation from (Eikema and Aziz, 2020). This extension, approximates MBR not only under a single NMT model (one assignment of θ) but under K sampled NMT models (K assignments of θ, simulated in practice via MC dropout). We call this Q-MBR, to reflect the fact that this is approximation to the MBR solution under the variational approximation of a Bayesian NMT model:

$$h^\star \approx \arg\max_h \frac{1}{K}\sum_{k=1}^{K}\frac{1}{S}\sum_{s=1}^{S} u(y^{(s)}, h^\star) \tag{18}$$

for candidates we use the unique translations found in $K \times S$ ancestral samples from the Bayesian NMT model. Combining this with other findings of ours (*e.g.*, that beam search and nucleus sampling outputs make better candidates, and that BMC can lead to faster convergence of estimates) is left for future work.

Data and models. So far we only experimented with IWSLT14 German-English data (Cettolo et al., 2014). We have 153326 training data points, 6969 validation pairs, and a test set with a total of 6750 pairs of sequences. We have tokenised and BPE preprocessed (Sennrich et al., 2016) our data with 3200 operations per language. The following results are for German into English. For this investigation, we use chrF (Popović, 2015) as utility. The main motivation for using chrF, apart from the low computational cost, is that it is designed to perform well on the sentence level. Modern sentence-level utilities will be used in future work. We compare different decoding methods: MBR, greedy decoding, and beam search, and report different quality metrics to ensure that our results are consistent and robust. For chrF3 and BLEU4 use the sacrebleu package (Post, 2018). There are a number of objectives possible: a baseline, which is simply Transformer NMT (Vaswani et al., 2017), we build upon JoeyNMT (Kreutzer et al., 2019), a Bayesian extension of NMT (BNMT) via MC dropout (Xiao et al., 2019), and loss-calibrated versions of both approaches. First, we look into loss calibration without Bayesian estimation.

Findings. Our results in Table 2 suggest that calibration towards a decision rule does lead to improvements across decoders. Contrary to our expectation, the MBR decoder benefited less, even though the gain function was based on it. We attribute this to the relatively small sample size we used ($S = 10$). Loss calibration affects the entropy of the output distributions, reducing it, which has an impact (even if indirect) on all decoders. Bayesian estimation for this dataset is in fact quite effective, as the BNMT results show. Once again, the Q-MBR decoder is the one to benefit less, given the computation budget invested ($K = 10$, $S = 10$). Finally, loss-calibrated BNMT did not improve on BNMT. We speculate this is due to compounding approximation errors (in loss calibration and in Bayesian estimation). Loss calibration of the NMT baseline seems a

Method	Decoding	chrF	BEER	BLEU
NMT	MBR	47.0	0.67	23.4
	Greedy	47.7	0.67	25.2
	Beam	48.5	0.68	26.4
Loss-calibrated NMT	MBR	47.2	0.67	24.8
	Greedy	49.4	0.69	28.4
	Beam	49.9	0.69	29.2
BNMT	Q-MBR	50.7	0.69	26.9
	Greedy	52.1	0.7	30.3
	Beam	52.7	0.70	31.2
Loss-calibrated BNMT	Q-MBR	49.7	0.68	26.2
	Greedy	51.5	0.70	30.1
	Beam	52.2	0.70	31.1

Table 2: Effects of Bayesian estimation and loss-calibration on Transformer NMT for IWSLT14 German-English (test results).

promising direction moving forward. A complete comparison to MLE alternatives strategies such as minimum risk training (MRT; Shen et al., 2016) and other forms of reinforcement learning (Kreutzer et al., 2017) is left for furture work.

5 Conclusion

This deliverable has reported the work conducted within WP3 on learning structural models, in particular, during the second half of the project. In task 3.1, we probed NMT models to better understand their design, training and generation, in particular, under the lens of the more classic SMT pipeline, leading to concrete recommendations for training objectives and data augmentation. In task 3.2, we have continued to develop machine learning techniques for training deep neural network models whose internal representations are sparse and retain inductive biases typical of discrete and possibly combinatorial structure, we have applied this to compress NMT encoders leading to better and more interpretable models. We have also continued to make parameter efficient use of document-level context by exploiting latent document-informed representations. In task 3.3, we continued to exploit the consequences of the probabilistic formulation of NMT models, with algorithmic advances in decoding and training. In all three fronts there is room for improvement and countless opportunities for original work, which thanks to GoURMET our research groups are well-positioned to develop.

A summary of our research output, in the form of conference publications, MSc theses, pre-prints under review, and open-source software and data can be found in Figure 18.

Conference Papers	**Main Task**
Bastings et al. (2019), Eikema and Aziz (2019)	3.2
Pelsmaeker and Aziz (2020), Correia et al. (2020)	3.2
Zheng et al. (2020) , Lopes et al. (2020), Dobreva et al. (2020)	3.2
Eikema and Aziz (2020)	3.3
Voita et al. (2021a), Voita et al. (2021b)	3.1
Wang et al. (2021), Zhang et al. (2021)	3.2
Farinhas et al. (2022)	3.2
Pre-prints	**Main Task**
Eikema and Aziz (2021) [*under review*]	3.3
Theses	**Main Task**
van Stigt (2019), Murady (2020)	3.2
Bortych (2021)	3.3
Software and Data	**Main Task**
Alignment models https://github.com/Roxot/m-to-n-alignments	3.1
Deep latent language models https://github.com/tom-pelsmaeker/deep-generative-lm	3.2
Sparse approximations to binary variables https://github.com/bastings/interpretable_predictions	3.2
Language models with latent syntax https://github.com/daandouwe/thesis	3.2
Deep latent translation models https://github.com/Roxot/AEVNMT.pt	3.2
Contrastive test sets for document-level machine translation https://github.com/rbawden/Large-contrastive-pronoun-testset-EN-FR	3.2
Training data for document-level machine translation https://github.com/radidd/Doc-substructure-NMT	3.2
Bayesian data analysis of NMT models https://github.com/probabll/bda-nmt	3.3
Constrained optimisation for torch https://github.com/EelcovdW/pytorch-constrained-opt.git	3.*
Probabilistic modules for torch https://github.com/probabll/dgm.pt	3.*
Probability distributions for torch https://github.com/probabll/dists.pt	3.*
Minimum Bayes risk decoding for NMT https://github.com/Roxot/mbr-nmt	3.3

Figure 18: Summary of research output.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR, 2015*, San Diego, USA, 2015. URL http://arxiv.org/abs/1409.0473.

Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1209. URL https://aclanthology.org/D17-1209.

Joost Bastings, Wilker Aziz, and Ivan Titov. Interpretable Neural Predictions with Differentiable Binary Variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1284. URL https://www.aclweb.org/anthology/P19-1284.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1118. URL https://aclanthology.org/N18-1118.

James O Berger. *Statistical decision theory and Bayesian analysis; 2nd ed.* Springer Series in Statistics. Springer, New York, 1985. doi: 10.1007/978-1-4757-4286-2. URL https://cds.cern.ch/record/1327974.

Frédéric Blain, Lucia Specia, and Pranava Madhyastha. Exploring hypotheses spaces in neural machine translation. *Asia-Pacific Association for Machine Translation (AAMT), editor, Machine Translation Summit XVI. Nagoya, Japan*, 2017.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3302. URL https://aclanthology.org/W14-3302.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August 2016a. Association for Computational Linguistics. doi: 10.18653/v1/W16-2301. URL https://aclanthology.org/W16-2301.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared*

Task Papers, pages 199–231, Berlin, Germany, August 2016b. Association for Computational Linguistics. doi: 10.18653/v1/W16-2302. URL https://aclanthology.org/W16-2302.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6401. URL https://aclanthology.org/W18-6401.

Nikita Bortych. Loss calibrating variational neural machine translation. Master's thesis, University of Amsterdam, 2021.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. Report on the 11th IWSLT evaluation campaign. In *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 2–17, Lake Tahoe, California, December 4-5 2014. URL https://aclanthology.org/2014.iwslt-evaluation.1.

Boxing Chen and Colin Cherry. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3346. URL https://aclanthology.org/W14-3346.

Gonçalo M. Correia, Vlad Niculae, and André F. T. Martins. Adaptively sparse transformers. 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 2174–2184, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1223. URL https://aclanthology.org/D19-1223.

Gonçalo M. Correia, Vlad Niculae, Wilker Aziz, and André F. T. Martins. Efficient marginalization of discrete and structured latent variables via sparsity. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.262. URL https://aclanthology.org/2020.emnlp-main.262.

Radina Dobreva, Jie Zhou, and Rachel Bawden. Document Sub-structure in Neural Machine Translation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3657–3667, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.451.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N13-1073.

Bryan Eikema and Wilker Aziz. Auto-Encoding Variational Neural Machine Translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 124–141, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4315. URL https://www.aclweb.org/anthology/W19-4315.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398.

Bryan Eikema and Wilker Aziz. Sampling-based minimum bayes risk decoding for neural machine translation. *arXiv preprint arXiv:2108.04718*, 2021.

António Farinhas, Wilker Aziz, Vlad Niculae, and Andre Martins. Sparse communication via mixed distributions. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WAid50QschI.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. Measuring and increasing context usage in context-aware 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 6467–6478, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.505. URL https://aclanthology.org/2021.acl-long.505.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Hkl1iRNFwS.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, June 2016. PMLR. URL http://proceedings.mlr.press/v48/gal16.html.

Jonas Gehring, Michael Auli, David Grangier, and Yann Dauphin. A convolutional encoder model for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 123–135, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1012. URL https://aclanthology.org/P17-1012.

Vaibhava Goel and William J. Byrne. Minimum bayes-risk automatic speech recognition. *Comput. Speech Lang.*, 14(2):115–135, 2000. doi: 10.1006/csla.2000.0138. URL https://doi.org/10.1006/csla.2000.0138.

Joshua Goodman. Parsing algorithms and metrics. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 177–183, Santa Cruz, California, USA, June 1996. Association for Computational Linguistics. doi: 10.3115/981863.981887. URL https://aclanthology.org/P96-1024.

Elliott Gordon-Rodriguez, Gabriel Loaiza-Ganem, and John Cunningham. The continuous categorical: a novel simplex-valued exponential family. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3637–3647. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/gordon-rodriguez20a.html.

Alex Graves. Sequence transduction with recurrent neural networks. In *ICML Workshop on Representation Learning*, volume abs/1211.3711, 2012.

Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1l8BtlCb.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1632. URL https://aclanthology.org/D19-1632.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. Scalable modified Kneser-Ney language model estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 690–696, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-2121.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28, 2015.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-4020. URL https://aclanthology.org/P18-4020.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3204. URL https://aclanthology.org/W17-3204.

Julia Kreutzer, Artem Sokolov, and Stefan Riezler. Bandit structured prediction for neural sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1503–1513, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1138. URL https://aclanthology.org/P17-1138.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. Joey NMT: A minimalist NMT toolkit for novices. 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): System Demonstrations*, pages 109–114, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-3019. URL https://aclanthology.org/D19-3019.

Shankar Kumar and William Byrne. Minimum Bayes-risk word alignments of bilingual texts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 140–147. Association for Computational Linguistics, July 2002. doi: 10.3115/1118693.1118712. URL https://aclanthology.org/W02-1019.

Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/N04-1022.

Simon Lacoste-Julien, Ferenc Huszár, and Zoubin Ghahramani. Approximate inference for the loss-calibrated bayesian. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 416–424. JMLR Workshop and Conference Proceedings, 2011.

Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0734.

Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era. *preprint arXiv:2006.02419*, 2020.

Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database. Available online: <http://yann.lecun.com/exdb/mnist>, 2010.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. URL https://aclanthology.org/2021.acl-long.353.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Peter J. Liu*, Mohammad Saleh*, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. Generating wikipedia by summarizing long sequences. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hyg0vbWC-.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André T. Martins. Document-level Neural MT: A Systematic Comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisbon, Portugal, 2020.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning Sparse Neural Networks through L_0 Regularization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1Y8hhg0b.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations*, 2017.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1325. URL https://www.aclweb.org/anthology/D18-1325.

Mathias Müller and Rico Sennrich. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.22. URL https://aclanthology.org/2021.acl-long.22.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6307. URL https://aclanthology.org/W18-6307.

Lina Murady. Probabilistic Models for Joint Classification and Rationale Extraction. Master's thesis, University of Amsterdam, 2020.

Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium, October 2018a. Association for Computational Linguistics. doi: 10.18653/v1/W18-6322. URL https://aclanthology.org/W18-6322.

Kenton Murray and David Chiang. Correcting Length Bias in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223, Brussels, Belgium, October 2018b. Association for Computational Linguistics. doi: 10.18653/v1/W18-6322. URL https://www.aclweb.org/anthology/W18-6322.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3956–3965, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/ott18a.html.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-4009. URL https://www.aclweb.org/anthology/N19-4009.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://www.aclweb.org/anthology/P02-1040.

Tom Pelsmaeker and Wilker Aziz. Effective estimation of deep generative language models. In *ACL*, 2020.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

Carl Edward Rasmussen and Zoubin Ghahramani. Bayesian monte carlo. *Advances in neural information processing systems*, pages 505–512, 2003.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Publishing Company, Incorporated, 2010. ISBN 1441919392.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1159. URL https://aclanthology.org/P16-1159.

Raphael Shu and Hideki Nakayama. Later-stage minimum bayes-risk decoding for neural machine translation. *CoRR*, abs/1704.03169, 2017. URL http://arxiv.org/abs/1704.03169.

Khalil Sima'an. On maximizing metrics for syntactic disambiguation. In *Proceedings of the Eighth International Conference on Parsing Technologies*, pages 183–194, Nancy, France, April 2003. URL https://aclanthology.org/W03-3021.

Edward Snelson and Zoubin Ghahramani. Local and global sparse gaussian process approximations. In *Artificial Intelligence and Statistics*, pages 524–531. PMLR, 2007.

Felix Stahlberg and Bill Byrne. 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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1331. URL https://aclanthology.org/D19-1331.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 362–368, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2058.

Miloš Stanojević and Khalil Sima'an. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1025. URL https://aclanthology.org/D14-1025.

Andreas Stolcke, Yochai Konig, and Mitchel Weintraub. Explicit word error minimization in n-best list rescoring. In *Fifth European Conference on Speech Communication and Technology*, 1997.

Ilya Sutskever, Oriol Vinyals, and Quoc V. V Le. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *NIPS, 2014*, pages 3104–3112. Montreal, Canada, 2014.

David Talbot, Hideto Kazawa, Hiroshi Ichikawa, Jason Katz-Brown, Masakazu Seno, and Franz Och. A lightweight evaluation framework for machine translation reordering. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 12–21, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2102.

Daan van Stigt. Neural language models with latent syntax. Master's thesis, University of Amsterdam, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-Aware Neural Machine Translation Learns Anaphora Resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1117. URL https://www.aclweb.org/anthology/P18-1117.

Elena Voita, Rico Sennrich, and Ivan Titov. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1116. URL https://aclanthology.org/P19-1116.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580.

Elena Voita, Rico Sennrich, and Ivan Titov. Analyzing the source and target contributions to predictions 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 1126–1140, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.91. URL https://aclanthology.org/2021.acl-long.91.

Elena Voita, Rico Sennrich, and Ivan Titov. Language modeling, lexical translation, reordering: The training process of NMT through the lens of classical SMT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8478–8491, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.667. URL https://aclanthology.org/2021.emnlp-main.667.

Bailin Wang, Mirella Lapata, and Ivan Titov. Meta-learning for domain generalization in semantic parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–379, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.33. URL https://aclanthology.org/2021.naacl-main.33.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. Exploiting Cross-Sentence Context for Neural Machine Translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1301. URL https://www.aclweb.org/anthology/D17-1301.

Ronald J. Williams. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine Learning*, 8(3-4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *preprint arXiv:1708.07747*, 2017.

Tim Z Xiao, Aidan N Gomez, and Yarin Gal. Wat zei je? detecting out-of-distribution translations with variational transformers. 2019.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. Variational Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 521–530, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1050. URL https://www.aclweb.org/anthology/D16-1050.

Biao Zhang, Ivan Titov, and Rico Sennrich. On sparsifying encoder outputs in sequence-to-sequence models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2888–2900, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.255. URL https://aclanthology.org/2021.findings-acl.255.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. Toward Making the Most of Context in Neural Machine Translation. In *International Joint Conference on Artificial Intelligence*, 2020.

Chunting Zhou, Jiatao Gu, and Graham Neubig. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BygFVAEKDH.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D3.2 GoURMET Final report on learning structural models