Global Under-Resourced MEedia Translation (GoURMET)

H2020 Research and Innovation Action
Number: 825299
D3.1 – Initial Progress Report on Learning Structural Models

Nature	Report	**Work Package**	WP3
Due Date	30/06/2020	**Submission Date**	30/06/2020
Main authors	Wilker Aziz (UVA)		
Co-authors	Bryan Eikema (UVA), Lina Murady (UVA), Rachel Bawden (UEDIN), Radina Dobreva (UEDIN)		
Reviewers	Alexandra Birch (UEDIN)		
Keywords	latent, structure, low-resource, inductive bias		
Version Control			
v0.1	**Status**	Draft	20/06/2020
v1.0	**Status**	Final	30/06/2020

Contents

List of Figures

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* (which replaces one of the initially proposed tasks, namely, *multilingual learning of sentence structure*). We report on progress thus far and our priorities for the second half of the project.

1 Introduction

WP3 provides scientific advances in machine translation for the low-resource setting by developing machine learning algorithms which induce and exploit structured representations of sentences. 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 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.
- Induce sentence structure relying on data for multiple language pairs simultaneously, thus, selectively transferring knowledge about linguistic structure from resource-richer languages to resource-poorer ones.

To cover these goals we originally proposed 3 tasks, namely,

T3.1 Modelling latent alignments (Section 2);

T3.2 Structured sentence models (Section 3);

T3.3 ~~Multilingual learning for sentence structure~~.

Task T3.3 turned out too closely-related to T3.2 as well as to the overall goals of work package 4. Learning sentence-level structure is the main objective of T3.2, and T3.3 does not seem to require new methodology as far as modelling with latent variables goes. The technical advances it does require concern transfer learning more than anything else, which is within the scope of WP4. Besides, as originally formulated, it depended too heavily on T3.2, thus preventing parallel progress in both fronts. We therefore changed the scope of T3.3 from multilingual learning for sentence structure to

T3.3 Probabilistic neural machine translation (Section 4)

whose goals revolve around exploiting and advancing implications of the probabilistic formulation of neural machine translation models, in particular, where this will lead to advances in low-resource settings. See Section 4 for a complete motivation including goals and progress thus far.

The first half of the project has led to considerable progress in all three fronts of the work package as well as cross-package collaborations. 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.

Summary of work done: We present an unsupervised neural model for alignment that builds upon the simple factorisation of a classic statistical model, namely, the IBM model 1 (Brown et al., 1993). This model can be thought of as a very simple NMT model which generates the target sentence one word at a time, each time translating an independently selected subset of source tokens. Because the space of subsets grows exponentially large with source sequence length, we employ variational inference learning via approximate marginalisation with Monte Carlo (MC) methods. We investigate both discrete and approximately discrete alignments and find that by employing some variance reduction techniques discrete alignments are viable and in fact outperform approximately discrete ones. Whereas the model shows to be a good alignment model, initial attempts at combining it with a complete NMT model did not lead to improved translation quality. We are currently developing alternative uses of this model for other work packages, for example, as a model of sentence alignment (in WP1) and for unsupervised discovery of sub-word units (in WP2).

2.1 Many-to-Many Alignments

Whereas alignments are no longer used in modern neural machine translation systems, they still can serve their purpose in low-resource tasks. As alignment models are often much simpler than translation models, they might be simpler to train in low-resource scenarios than NMT systems. The resulting alignments could potentially be used to improve NMT systems with applications such as: supervising the attention component (Liu et al., 2016); supervising every decision from segmentation, to reordering, to phrase translation (Alkhouli et al., 2016; Alkhouli and Ney, 2017); auxiliary supervision (in a multi-task learning fashion) to promote better translation while keeping attention weights interpretable as alignments (Garg et al., 2019); aligning embedding spaces in unsupervised NMT (Artetxe et al., 2019). Alignments are central to SMT (Koehn et al., 2003; Chiang, 2005), and in some low-resource scenarios, SMT might still work better than NMT, provide soft constraints to unsupervised NMT (Ren et al., 2019), or at least contribute valuable synthetic data to train NMT models (Burlot and Yvon, 2018).[1]

Traditional non-neural alignment models (Brown et al., 1993; Och and Ney, 2003; Dyer et al., 2013) make some strong assumptions that simplify their training. Most notably, IBM models 1 and 2, the basis of the most powerful alignment models, assume that target words are generated from at most a single source word. Whereas this makes training feasible, this is by no means a correct assumption about language. It is easy to come up with a counter-example: *"The dog sleeps"*

[1] For example, SMT was used to produce synthetic data to some of our own systems (e.g., English-Amharic and Amharic-English).

translates to Romanian into *"Câinele doarme"*. In Romanian the definite article *"le"* is added as a postfix to the noun for dog, *"câine"*. *"Câinele"* cannot be explained by *"The"* or by *"dog"* alone, but only by the phrase *"The dog"*. Some neural alignment models (Wang et al., 2018; Rios et al., 2018) do not relax these strong assumptions. Others do, but are based on components that do not easily integrate in other probabilistic models (Legrand et al., 2016) or target high-resource settings (Garg et al., 2019; Zenkel et al., 2020).[2]

In this research, we seek a model that is lightweight, that performs well without requiring large architecture blocks, such as a Transformer (Vaswani et al., 2017) encoder or decoder, and that produces many-to-many alignments. Therefore, we formulate a model where target words can be generated from any number of source words. We use neural networks to learn mappings from a selection of source tokens to a distribution over target words, and to select the source tokens from which a target word is generated. To achieve this we introduce latent alignment *bit vectors*, where each target word is augmented with a latent bit vector indicating which positions in the source sentence are responsible for generating that target word. The likelihood of the data is the marginal of this latent variable model, i.e.,

$$p(y|x,\theta) = \prod_{j=1}^{|y|} \sum_{a_j \in \{0,1\}^{|x|}} p(a_j|x, y_{<j}, \theta) p(y_j|x, a_j, \theta) \tag{1}$$

where we denote the source sequence as x of length $|x|$, the target sequence as y of length $|y|$, any individual target words as y_j and its corresponding bit vector as a_j, and the sequence $\langle y_1, \ldots, y_{j-1} \rangle$ as $y_{<j}$. Neural network parameters are contained in θ. A priori we can give preference to certain bit vectors, we can give preference to how many source words explain a target word on average, or we could even prefer that multiple source words selected should ideally be adjacent. In experiments we currently only experiment with independent alignment bits that have a fixed or source sequence dependent prior.

In order to estimate parameters of the model in Equation 1, we need to compute the marginalisation over all possible bit vectors. The number of bit vectors is exponential in source sequence length and therefore generally intractable. At test time we are interested in predicting alignments using the posterior distribution $p(a_j|x, y)$, computing this is intractable due to the same intractable summation. For this reason we resort to variational inference (Jordan et al., 1999; Blei et al., 2017), giving us an approximate posterior with its own set of parameters λ, and a tractable lower-bound on the logarithm of Equation 1. We then optimise this lowerbound with respect to both sets of parameters, namely, θ and λ. In order to compute gradients for the approximate posterior λ, we experiment with the REINFORCE estimator (Williams, 1992), and with relaxing the discrete bits to a continuous-discrete mixture using the HardKuma distribution (Bastings et al., 2019) developed in this work package (see Section 3.1.2).

As a baseline we implement a neural variant of IBM 1, where the translation component of IBM 1 is parameterised by a neural network. Because in IBM 1 we can exactly marginalise the alignments in time $O(|x| \times |y|)$, parameter estimation via EM algorithm (Dempster et al., 1977) is possible. As we have introduced neural components, we use gradient-based training which is more convenient to use with standard auto-grad packages. We also report fast-align (Dyer et al., 2013), a strong

[2] The closest to low-resource setting considered in recent literature is the case of English-Romanian, where about half a million sentence pairs are available. English-Amharic, one of the language pairs we are interested in, counts with only 50 thousand sentence pairs.

model	prior	AER
neural IBM 1	Cat($\frac{1}{M}, \ldots, \frac{1}{M}$)	34%
HardKuma	HardKuma(0.3, 1)	30%
REINFORCE	Bernoulli$\left(\frac{1}{1+M}\right)$	29%
+ baselines		27%
+ self-critic		27%
+ PPO		27%
fast-align	Dir(α)	20%

Table 1: Alignment error rate for the baseline model (neural IBM 1) and several variants of our model. HardKuma uses a HardKuma approximate posterior whereas REINDORCE uses a Bernoulli approximate posterior together with REINFORCE. We only show the prior that worked best in our experiments.

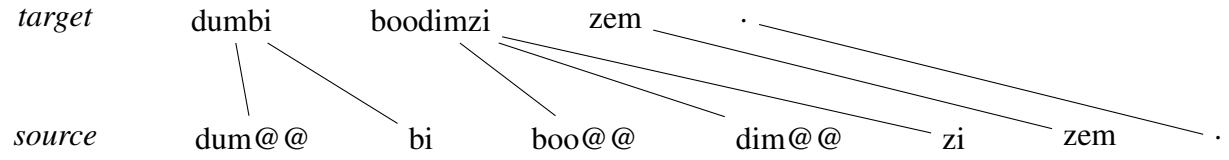

Figure 1: An example sentence pair from the artificial data we generated to demonstrate the capacity of our model. The task is to merge words ending on "@@" into a single compound word. This task requires target words to align to potentially multiple source words by design.

(non-neural) reparameterisation of IBM 2 which employs a sparse Dirichlet prior for lexical distributions. Note that while stronger neural alignment models do exist, at this point our goal is to investigate whether we can induce many-to-one alignments with a small neural model trained end-to-end, not at all whether we can outperform a pipeline of models and symmetrisation strategies.

Small scale experiment We use a small English-French dataset, the sentence-aligned Hansards (Germann, 2001), containing about 230,000 thousand observations. For this pair, we have a small test set annotated with manual alignments, the NAACL English-French hand-aligned data (Mihalcea and Pedersen, 2003), that we can use to evaluate the unsupervised models. We compare induced alignments to gold-standard while for various gradient estimation techniques and priors in terms of alignment error rate (AER; Och and Ney, 2000). Table 1 summarises the results. We only show the prior settings that worked best in our experiments. We see that our model in general has an edge over the baseline, both for the HardKuma as well as the REINFORCE strategy. We see that reducing the variance of the REINFORCE estimator through standard baselines (Williams, 1992) is beneficial. Further attempting to reduce variance using a greedy self-critic baseline (Rennie et al., 2017) and proximal policy optimization (PPO; Schulman et al., 2017) does not seem to further improve results.

Use in NMT Our first attempt to combine this alignment component with a complete NMT system is based on multi-task learning (Caruana, 1997), whereby we have two models explain the data differently while sharing some of their components, in this case, the encoder and the embedding

model	prior	AER
neural IBM 1	$\text{Cat}(\frac{1}{M}, \ldots, \frac{1}{M})$	42%
fast-align	$\text{Dir}(\alpha)$	35%
our model	$\text{Bernoulli}\left(\frac{2}{1+M}\right)$	8%

Table 2: Alignment error rates on the toy data demonstrated in Figure 1. Our model is trained with REINFORCE and baselines.

matrices. We model English→Turkish data using an attention-based NMT model (Bahdanau et al., 2015) and the alignment model developed in this section. Unfortunately, we have not yet observed appreciable improvements, thus other integration strategies will have to be investigated.

Potential of many-to-many alignments We now demonstrate with a toy experiment the potential of the model in discovering groups. We generate artificial data in which the task is to merge groups of words together. Every target word is thus, by design, generated from potentially many source words. We demonstrate the task in Figure 1. Whereas this task is quite artificial and acts as to demonstrate the potential of our model, the concept of many-to-many alignments is not, as we have shown before in an example for English-Romanian. We compare neural IBM 1, our model and fast-align in Table 2. We see that our model is close to solving the task, whereas neural IBM 1 and fast-align miss many alignments. In fact, modelling groups like that is beyond the capabilities of IBM 1 and 2 style models, requiring a pipeline of models and symmetrisation heuristics.

Discussion We see potential for alignment models, especially so when struggling with lack of resources. Moreover, we see great potential in neural end-to-end alignment models. Whereas classic statistical models struggle with feature-rich parameterisations, neural networks are extremely flexible in that regard. Consider for example the case of morphologically complex languages, complex features such as long-distance agreement and sub-word features are arguably valuable and hard to learn with little data. Recent advances in neural variational inference allow us to consider classes of models that would render classical approaches intractable. Still, moving forward in the project, we have chosen to de-prioritise this direction in favour of another with higher impact (see Task 3). We will still investigate a few more options for integration, but we do not plan to investigate new models. The ideas developed here are not being discarded, however. We have recently realised an opportunity for the model presented here, and other models of this kind, in the context of data collection (work package 1), where many-to-many sentence alignment has proven rather challenging. This model may also find space in work package 2, in the context of learning to segment words into sub-word units.

3 Task 3.2 – Structured Sentence Models

Proposal highlights:

- 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.

Summary of work done: we report progress in three fronts, one focused on learning with unobserved variables (Section 3.1), this aims at advancing technology that will enable latent structure in NMT, another focused on inducing sentence-level structure within NMT (Section 3.2), and the last one focused on making use of context beyond the sentence level (Section 3.3). In the first front, we (i) improve variational inference for text generation problems by addressing a common failure mode of deep latent variable models known as posterior collapse, (ii) develop sparse relaxations to binary random variables that admit unbiased reparameterised gradients, (iii) improve variational inference for language models with latent syntactic structure, and (iv) improve variational inference for discrete combinatorial structure. In the second front, we (i) develop a joint generative model that induces latent representations of sentence pairs, and (ii) build discrete latent factors into this model. In the last front, we (i) unify sentence-level and document-level translation, (ii) create evaluation resources and present a systematic evaluation of alternative architectures for representing global context, and (iii) investigate the effect of sub-document information.

3.1 Latent Structure in Neural Network Models

Unobserved variables lead to challenges in statistical learning, and, in particular, in deep learning, where approximate marginalisation of unobserved (latent) variables pose challenges for computation of gradients. The building block of deep latent variable models is the *variational auto-encoder* (VAE; Kingma and Welling, 2014; Rezende et al., 2014), a probabilistic latent variable model parameterised by neural networks. While VAEs have great potential, they have two big limitations. When parameterised with strong neural network architectures, as we typically want in text generation applications, they fail to exploit latent variables, this is known as *posterior collapse* (Bowman et al., 2016; Alemi et al., 2018). VAEs are trained via *reparameterised gradients*, that is, gradient estimates obtained by a differentiable sampling procedure. Unfortunately, differentiable sampling *is not* possible with discrete random variables, and therefore, one needs to either employ a continuous relaxation for which differentiable sampling is possible, or employ a general class of estimators known as score function estimator (Rubinstein, 1986; Williams, 1992), which is cursed with high variance (Paisley et al., 2012; Ranganath et al., 2014). In this section we look

into advancing machine learning technology around these challenges. We start by addressing posterior collapse for deep continuous latent variable language models, and then turn to discrete and combinatorial latent variables.

Summary of contributions:

- Strategies to address posterior collapse in deep latent language models;
- A novel distribution which admits unbiased and differentiable sampling of sparse relaxations to binary random variables;
- Variational posteriors over context-free forests using conditional random fields;
- Deterministic gradients via tractable marginalisation of categorical and combinatorial latent variables.

3.1.1 Avoiding Posterior Collapse in Latent Variable Models

The variational auto-encoder (VAE; Kingma and Welling, 2014; Rezende et al., 2014) is one of the most important building blocks of deep latent variable models (i.e., probabilistic latent variable models parameterised by neural networks). In a latent variable model of the kind $p(z)p(x|z,\theta)$, the likelihood $p(x|z,\theta)$ has the potential to exploit structure in latent space, that is, it has the potential *to correlate specific regions of the latent space to specific patterns in data space*. Such potential is precisely what we seek to exploit. For example, we hope to be able to control generations from that space, or we hope to cluster observations that are similar and thus leverage incomplete supervision. For any of that to happen, we need to learn a joint distribution that actually uses the latent variable for something. Learning a VAE requires approximating the model's true posterior distribution $p(z|x,\theta)$ with a tractable variational approximation $q(z|x,\lambda)$ and we typically train both in tandem. The posterior distribution, or its approximation, can be thought of as a form of interface between the model and the user. Queries to this posterior can reveal neighbourhood in data space and insights into what generalisations the model makes.

In text generation tasks, such as machine translation, the generator $p(x|z,\theta)$ is typically parameterised by a very powerful neural network architecture, such as an autoregressive recurrent neural network (Hochreiter and Schmidhuber, 1997; Cho et al., 2014) or Transformer (Vaswani et al., 2017). Unfortunately, maximum likelihood estimation does not express a preference as to whether the observation x should be modelled by mapping patterns to latent space and back or simply by mapping from structure internal to the observation. For example, rather than predicting the next word x_i from a combination of z and the prefix of observations $x_{<i}$, such a model often conditions on $x_{<i}$ alone. On the one hand this may seem counter-intuitive, on the other hand, recall that $x_{<i}$ is a real observation sampled from the data, unlike z, which is sampled from a distribution undergoing training, and, above all, $x_{<i}$ is surely correlated with x_i, since both co-occur in the sentence x sampled from the data, whereas z is not correlated with neither $x_{<i}$ nor x_i unless the generator $p(x_i|z,x_{<i},\theta)$ decides to do so. This leads to a phenomenon known as *posterior collapse*, a *qualitatively* bad local optimum in maximum likelihood estimation (Alemi et al., 2018), whereby the estimated joint distribution exploits no structure in latent space. This is diagnosed by noticing that the *true posterior* is independent of the data, and thus has "collapsed" to the prior. As we cannot infer the true posterior, we employ a proxy diagnostic tool, namely, the expected KL divergence $\mathbb{E}_X[\mathrm{KL}(q(z|x,\lambda)\|p(z))]$, also known as *rate*. When this quantity is estimated to be approximately

0, we are confident that x is modelled independently of z. It is important to remark that collapsed models are not bad as such, they compete fairly (possibly rival) non-collapsed models in terms of likelihood. There is nothing objectively bad—as far as likelihood is concerned—about a collapsed model. It is our desire to exploit structure in latent space to accommodate appreciable generalisations or patterns in data space that is left unattended. And with it, all the applications we envisioned for the latent space (such as controllable generation or semi-supervised learning).

MDR In Pelsmaeker and Aziz (2020), presented at ACL 2020, we review a number of strategies that combat posterior collapse and introduce a few, most notably, minimum desired rate (or MDR). MDR uses Lagrangian relaxation to find models that satisfy a constraint on the rate. We show that by keeping the rate well above 0 we help the optimiser find generative models that are not collapsed. The idea is to optimise the evidence lowerbound (ELBO) subject to a minimum rate r:

$$\begin{aligned} &\max_{\theta,\lambda} \underbrace{\mathbb{E}_X\left[\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right] - \mathrm{KL}(q(z|x,\lambda)\|p(z))\right]}_{\mathrm{ELBO}(\theta,\lambda)} \\ &\text{s.t. } \mathbb{E}_X\left[\mathrm{KL}(q(z|x,\lambda)\|p(z))\right] > r\ . \end{aligned} \tag{2}$$

Because constrained optimisation is intractable, we optimise the Lagrangian (Boyd et al., 2004)

$$\Phi(\theta,\lambda,u) = \mathrm{ELBO}(\theta,\lambda) - u(r - \mathbb{E}_X[\mathrm{KL}(q(z|x,\lambda)\|p(z))]) \tag{3}$$

where $u \in \mathbb{R}_{\geq 0}$ is a positive Lagrangian multiplier. We define the dual function

$$\phi(u) = \max_{\theta,\lambda} \Phi(\theta,\lambda,u) \tag{4}$$

and solve the dual problem $\min_{u\in\mathbb{R}_{\geq 0}} \phi(u)$. Local minima of the resulting min-max objective can be found by performing stochastic gradient descent with respect to u and stochastic gradient ascent with respect to θ, λ. In our ACL paper we also discuss in detail how MDR relates to existing techniques aimed at addressing the same problem. We report results on language modelling for English. Here we comment on only a subset of the experiments involving the Penn Treebank dataset (PTB; Marcus et al., 1993). Table 3 shows that we can de-collapsed VAEs that employ different priors: Gaussian ($\mathcal{N}/\mathcal{N}$), mixture of Gaussians (MoG/$\mathcal{N}$), and the strong VampPrior (Tomczak and Welling, 2018), which we adapt for language modelling. Perplexity gives an indication of the model's performance and active units is an indication that the model is not collapsed (0 would mean collapsed). Similarly, a non-zero rate indicates the model is not collapsed. The VAEs in the Table evaluate to $R = 5$ because they were optimised to achieve that rate. Without MDR those models would have collapsed, as we show in the paper. Figures 2 and 3 illustrate samples from non-collapsed VAEs. The former shows samples from the model's own prior distribution, while the latter shows samples from the hyperplane connecting the posterior distributions of two given data samples. In a collapsed model, we would not be able to appreciate any relationship amongst those samples.

MDR has been an important development. Since its introduction for language modelling, we have made MDR part of most our VAE models. For example, in D2.1 we present a model for generation of inflected wordforms, where MDR was shown to be crucial. It is also part of other models presented in this deliverable and has been integrated in our deep latent variable translation model (see Section 3.2.1). To make it accessible to a larger audience we have packed it as a simple plug-and-play module in torch.[3]

[3] https://github.com/EelcovdW/pytorch-constrained-opt.git

Model	Distortion ↑	Rate ↓	Perplexity ↓	Active Units ↑
RnnLm	-	-	84.5	-
$\mathcal{N}/\mathcal{N}$	103.5	5.0	81.5	13
MoG/$\mathcal{N}$	103.3	5.0	81.4	32
Vamp/$\mathcal{N}$	103.1	5.0	81.2	22

Table 3: Performance on the PTB test set for different priors ($\mathcal{N}$, MoG, Vamp).

Sample	Closest training instance	TER
For example, the Dow Jones Industrial Average fell almost 80 points to close at 2643.65.	*By futures-related program buying, the Dow Jones Industrial Average gained 4.92 points to close at 2643.65.*	0.38
The department store concern said it expects to report profit from continuing operations in 1990.	*Rolls-Royce Motor Cars Inc. said it expects its U.S. sales to remain steady at about 1,200 cars in 1990.*	0.59
The new U.S. auto makers say the accord would require banks to focus on their core businesses of their own account.	*International Minerals said the sale will allow Mallinckrodt to focus its resources on its core businesses of medical products, specialty chemicals and flavors.*	0.78

Figure 2: Samples from a sentence VAE with an MoG prior trained via MDR: we sample from the prior and decode greedily. We also show the closest training instance in terms of a string edit distance, this allows us to verify that the model is not simply memorising data it has seen. The fact that greedy samples are not always the same demonstrates that the generator is sensitive to the latent space.

Weak generators As we explained, VAEs collapse because strong generators have no incentive to use latent space to accommodate variability in data space. This begs the question, why don't we use weak generators then? Weak generators are models that do not have access to all of the structure present in the data, for example, a bag-of-words generator cannot capture much more than marginal word frequencies. It certainly cannot model intricate dependencies due to collocations, morphological agreement, or word order. In fact, weak generators do not collapse as previous work has shown (Zhao et al., 2017; Semeniuta et al., 2017; Park et al., 2018). The problem with weak generators is that they do not lead to fluent text either. Recall, we want to have good generators $p(x|z,\theta)$ and, in addition, be able to explore posterior queries $q(z|x,\lambda)$ for something interesting like data augmentation, clustering in terms of structural aspects of the data such as syntax and morphology. Access to these posterior queries at the expense of good generations is of little use. With our work on MDR we learnt that we can tackle posterior collapse by imposing a lowerbound on the model's rate. While that is interesting and has been shown to indeed correlate x and z, it remains quite beyond our control how this potential is exploited by the generator. In this research we aim to shape this. Moreover, we aim not to impose a lowerbound on rate directly, but rather impose constraints that promote z being predictive of a particularly useful aspect of the data. Our constraints are based on weak generators, for they are known not to suffer from posterior collapse. Indirectly, not only we promote higher rates, since weak generators lead to higher rates, but also target this rate to accommodate the view of the data the weak generator exposes. As the main

The inquiry soon focused on the judge.
The judge declined to comment on the floor.
The judge was dismissed as part of the settlement.
The judge was sentenced to death in prison.
The announcement was filed against the SEC.
The offer was misstated in late September.
The offer was filed against bankruptcy court in New York.
The letter was dated Oct. 6.

Figure 3: Latent space homotopy from a sentence VAE trained via MDR. Note the smooth transition of topic and grammatically of the samples.

model remains a strong generator, this should not hurt performance on the main downstream task. Again, we formalise this in the language of constrained optimisation:

$$\begin{aligned} &\max_{\theta,\lambda} \underbrace{\mathbb{E}_X\left[\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\theta)\right] - \mathrm{KL}(q(z|x,\lambda)\|p(z))\right]}_{\mathrm{ELBO}(\theta,\lambda)} \\ &\text{s.t. } \mathbb{E}_X\left[\mathbb{E}_{q(z|x,\lambda)}\left[\log p(x|z,\phi_v)\right]\right] > \rho_v \quad \text{for } v = 1,\ldots,K\,, \end{aligned} \tag{5}$$

where we optimise the ELBO of the main generative model, which employs a strong generator $p(x|z,\theta)$, subject to meeting pre-specified performance levels with K weak generators, each based on a VAE $p(z)p(x|z,\phi_v)$. This is an ongoing collaboration between UVA and UA, here we present only preliminary results. Figure 4 shows VAEs trained with free bits (FB; Kingma et al., 2016), which can be thought of as a simple form of MDR without Lagrangian relaxation, and our new strategy. At this point we have experimented with: bag-of-words (BoW) decoders, non-autoregressive decoders (NonAR; Ziegler and Rush, 2019),[4] and a combination of the two. As we can see our models are not at all collapsed, and that without the need for MDR. In particular, NonAR performs quite well (lower-left corner of the plots) in terms of different notions of perplexity. Preliminary analysis suggests that generations from these models differ considerably in terms of statistics of the data, such as length, unigram frequencies, bigram and skip bigram frequencies.[5] This suggests we are managing to use the latent space for specific purposes, unlike with techniques such as FB and MDR which do not target any specific use of the latent space.

This is another important development, with careful choice of weak generators we can induce structured representations via a continuous space, which simplifies training. We are currently investigating weak generators that are sensitive to word order statistics. Being able to shape the latent space of a VAE has implications across work packages, for example, in D1.2 this helps with data augmentation.

3.1.2 Differentiable and Sparse Relaxations to Binary Random Variables

Latent variable models learn to correlate observed variance in a rather complex data space to unobserved variance in a simpler (often lower-dimensional) latent space. This machinery is yet

[4] Like BoW, these models do not condition on a history of already generated tokens, unlike BoW, these models use a product of different distributions, one per generation step, while BoW uses the same distribution for all generation steps.

[5] This uses the methodology developed in Section 4.1.

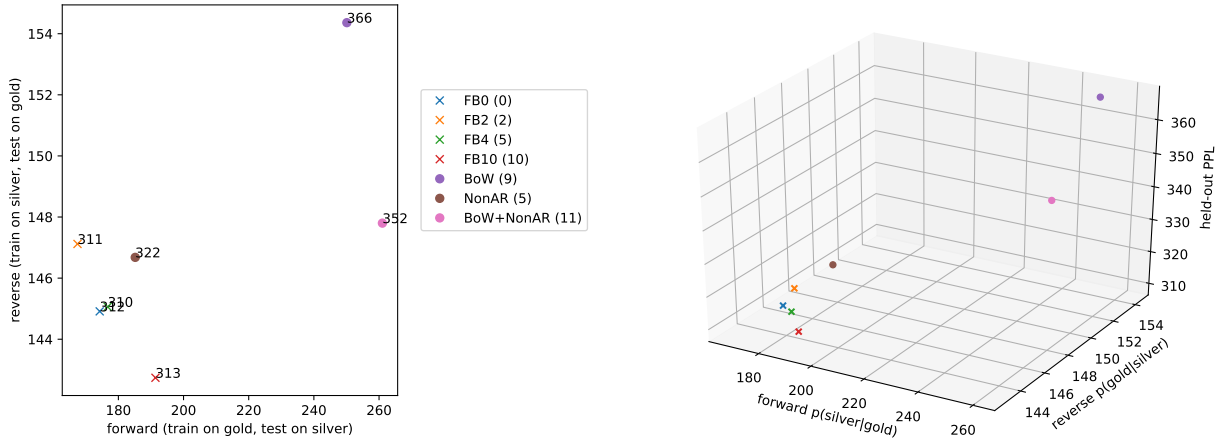

Figure 4: Perplexity of sentence VAE trained to also achieve high likelihood with weak generators. Forward perplexity (x-axis): lower means the generated data cover the gold-standard well. Reverse perplexity (y-axis): lower means the gold-standard data covers the generated data well. Held-out perplexity (z-axis): lower means the model explains gold-standard data well. In brackets we show the rate of the main model at the end of training (0 indicates a collapsed model).

more appealing where the unobserved variables are *discrete*. That is because discrete variables are more amenable to interpretation and can be used to inject certain inductive biases into our models. For example, a categorical variable may be thought of as a cluster, a collection of binary variables may be thought of as a description in terms of attributes. With some amount of supervision, we can go beyond making analogies and learn models that use discrete variables in exactly those ways (see for example Task 3 in D2.3, where we semi-supervise a discrete latent variable model to learn morphological attributes of complex word forms). Training an interesting latent variable model almost always requires approximating crucial intractable quantities, such as marginal likelihood or the evidence lowerbound, via sampling. Training deep learning models, on the other hand, requires computing the gradient of those intractable quantities with respect to the parameters of the model. Sampling and gradient computation are often in conflict, and where we are interested in discrete latent variables our options are greatly reduced.

Sampling discrete random variables is a non-differentiable operation due to the cumulative distribution function (cdf) of discrete variables being discontinuous. Gradient estimation can still be performed, but not through low-variance path derivatives (Kingma and Welling, 2014; Rezende et al., 2014), rather via a more general class of estimators known as the score function estimator (SFE; Rubinstein, 1986), or REINFORCE (Williams, 1992). Whereas SFE is unbiased it suffers from high variance requiring variance reduction techniques to be useful (Paisley et al., 2012; Ranganath et al., 2014). An alternative is to turn to continuous relaxations of discrete variables for which reparameterised gradients are possible (Maddison et al., 2017; Jang et al., 2017), these relaxations however either sacrifice sparse (discrete) samples or sacrifice unbiased gradients by making use of the straight-through estimator (STE; Bengio et al., 2013).

In Bastings et al. (2019), presented at ACL 2019, we introduce the HardKuma distribution, a relaxation of a Bernoulli distribution that supports continuous and discrete outcomes. This distribution supports outcomes in the closed interval [0, 1], but with non-zero mass for sampling exactly 0 and exactly 1. Crucially, sampling is performed via a differentiable reparameterisation and without biased estimators. The distribution is based on a *stretch-and-rectify* technique proposed by Louizos

et al. (2018), whereby one smooths discontinuous gradients via stochasticity. Figure 5 illustrates how HardKuma distributions are constructed. We refer to the published paper for details such as density assessments and reparameterised gradients.

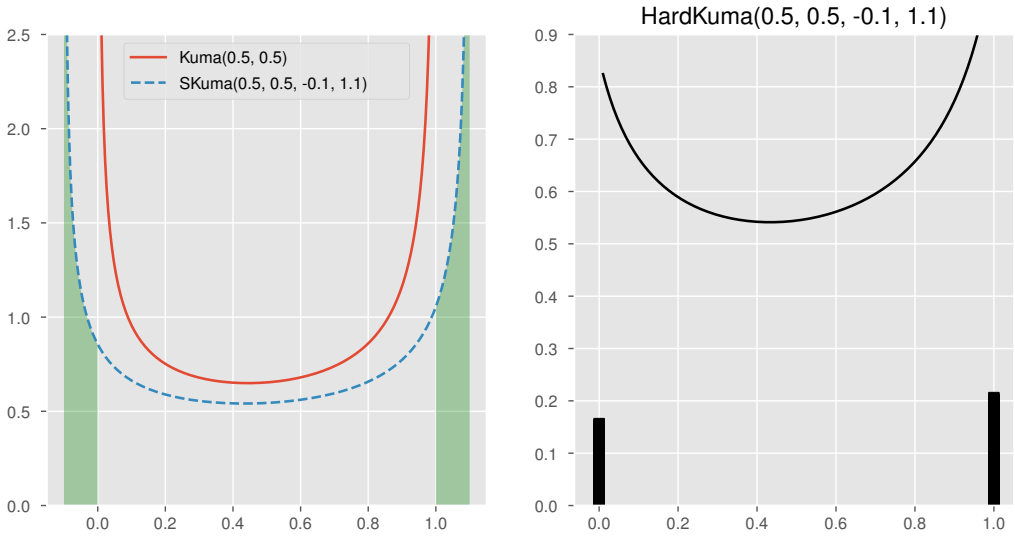

Figure 5: We stretch a base distribution (red solid line) to include 0 and 1 in its support (blue dashed line). We then integrate over the outside regions (green shaded area) collapsing their masses into two point masses, one at 0 and one at 1 (black solid bars). The resulting distribution (right) is a mixture of a truncated density defined over $(0, 1)$ and point masses at $\{0, 1\}$.

Extensions We also developed a relaxation of a Beta-Bernoulli model, which allows for sparsification of Bayesian models, the extension and an application to text classification appears in Murady (2020)'s MSc thesis. To make our distributions accessible to a wider audience, we have extended the library `torch.distributions` with a general stretch-and-rectify interface as well as with the HardKuma distribution.[6]

Applications In Bastings et al. (2019) we demonstrate the effectiveness of HardKuma by comparing it to SFE (REINFORCE) in the context of interpretable text classifiers, where the HardKuma is used to learn latent rationales for classification (Lei et al., 2016).[7] Since its introduction we have also successfully used the HardKuma in our alignment model of Section 2.1 as well as to learn approximately binary latent factors in a joint generative translation model in Section 3.2.2. In work package 2, HardKuma samples power a model of morphological reinflection and a morphologically-aware NMT decoder (Ataman et al., 2020).

[6] https://github.com/probabll/dists.pt
[7] Code available from https://github.com/bastings/interpretable_predictions.

3.1.3 Syntactic Language Models

Language models are trained to estimate the probability $p(x_i|\theta, x_{<i})$ of the next word x_i given a sequence of already generated words $x_{<i}$.[8] Nowadays these models are parameterised by recurrent (Hochreiter and Schmidhuber, 1997; Cho et al., 2014) and Transformer (Vaswani et al., 2017) architectures and are the state of the art in density estimation for text (Melis et al., 2018). There is something common to all of these models, they model sentences from left-to-right disregarding the hierarchical nature of language (Chomsky, 1964). Hierarchical structure is generally compact and lead to data-efficient generalisation. Learning hierarchical structure without any supervision, on the other hand, is generally very hard and cursed with theoretical limitations (Klein, 2005). Informing models with linguistic knowledge is something more realistic.

RNNG Dyer et al. (2016) introduce the recurrent neural network grammar (RNNG), a language model that draws observations from the marginal of a joint distribution over sentences and their syntactic (PTB-style) trees. The model is recurrent, but not along the sentence, rather along its syntactic derivation. Due to its generative formulation, the RNNG can be used in absence of syntactic trees, but training without syntactic trees is difficult, therefore, in its original formulation, the model was trained with complete syntactic supervision. In this research we investigate whether we can train the RNNG semi-supervisedly. Should this be possible, we might be able to address low-resource languages by means of producing some small amount of supervision.[9]

Formally, the RNNG assigns probability

$$p(x|\theta) = \sum_t p(x, t|\theta) = \sum_{t:\text{yield(t)=x}} \prod_{i=1}^{|t|} p(t_i|t_{<i}, x_{\pi(i)}, \theta) \tag{6}$$

to a sentence x, where the summation is over all trees whose yield is the given sentence. Each tree corresponds to a sequence of actions by a transition-based parser whose controller is parameterised by a stack-LSTM (Dyer et al., 2015). Training the model via gradient-based optimisation is straightforward as long as we have an oracle, for example a labelled observation from the Penn Treebank (PTB; Marcus et al., 1993):

$$\arg\max_\theta \sum_{x,t\sim\mathcal{D}_{\text{PTB}}} \log p(x, t|\theta) \,. \tag{7}$$

After training, the model can be used for generating sentences (along with their syntactic trees), or to assess the likelihood of a given sentence (or dataset). For the latter, we need to approximate the marginalisation in (6), which Dyer et al. (2016) do via importance sampling

$$p(x|\theta) = \sum_t q(t|x, \lambda)\frac{p(x, t|\theta)}{q(t|x, \lambda)} \overset{\text{MC}}{\approx} \frac{1}{S}\sum_{s=1}^{S} \frac{p(x, t^{(s)}|\theta)}{q(t^{(s)}|x, \lambda)} \tag{8}$$

using S trees drawn from the independently trained importance distribution $q(t|x, \lambda)$. The importance distribution is itself an RNNG, though trained discriminatively as a parser:

$$\arg\max_\lambda \sum_{x,t\sim\mathcal{D}_{\text{PTB}}} \log q(t|x, \lambda) \,. \tag{9}$$

[8] Masked language models (Devlin et al., 2019) are also called *language models* but they cannot be used to assess likelihood nor generate samples without considerable work and several approximations (Wang and Cho, 2019).

[9] Transferring supervision automatically from English might give us an inexpensive starting point.

CRF proposal We introduce a simpler proposal distribution $q(t|x, \lambda)$, namely, a conditional random field (CRF; Lafferty et al., 2001) parser trained to maximise likelihood of labelled PTB observations. CRFs enjoy certain analytical results that are convenient for a parser, for example, it is tractable to search for the best parse tree, and for a proposal distribution, for example, entropy is available in closed-form. An RNNG parser is not amenable to exact search, and requires greedy approximations, as a proposal it can only yield noisy Monte Carlo estimates of entropy. Our CRF uses the minimal parameterisation of the margin-based parser of Stern et al. (2017) and because of that scales much better than the original neural CRF parser by Durrett and Klein (2015). Table 4 shows how our CRF parser compares to others. Note that it outpeforms the RNNG, while still producing a distribution over parse trees. The margin-based parser outperforms both the RNNG and the CRF, but that is expected, since it does not need to maintain a representation of the entire distribution. Next we evaluate the CRF parser as an importance distribution, note that the margin-

Discriminative parser	**F1**
RNNG	88.58
CRF	90.04
Margin-based (Stern et al., 2017)	91.79

Table 4: Parsing performance measured in terms of F1: the RNNG parser returns the tree found by a greedy approximation to best-tree search, while the CRF parser returns the true highest-scoring tree.

based parser cannot be used for that. Table 5 shows that the much simpler CRF matches the RNNG proposal in terms of F1 (parsing performance), but lags behind in terms of importance sampling estimates. This might be explained by the CRF distributions showing higher entropy, indeed, we flattened an RNNG distribution after training, thus artificially increasing its entropy, and observed worse perplexity (see Figure 6).

Semi-supervised learning We can integrate the training of the proposal (a CRF) and of the generative model (an RNNG) via neural variational inference (Mnih and Gregor, 2014), where we optimise both components jointly to maximise the evidence lowerbound (ELBO):

$$\mathbb{E}_{q(t|x,\lambda)}\left[\log p(x, t|\theta)\right] + \mathbb{H}(T|x, \lambda)\ . \tag{10}$$

With a CRF proposal, we can compute the second term (the entropy) in closed-form, something we could not do with an RNNG proposal. The first term requires gradient estimation, we use the

Proposal	**F1**	**Perplexity**
RNNG	91.1 ± 0.1	108 ± 1
CRF	91.0 ± 0.1	117 ± 2

Table 5: Performance of the language model as a function of type of proposal. Parsing performance is evaluated in terms of F1: the generative model re-ranks the trees sampled from the proposal. Language model performance is evaluated in terms of perplexity: samples from the proposal are used to estimate marginal likelihood. We use 100 samples in both cases.

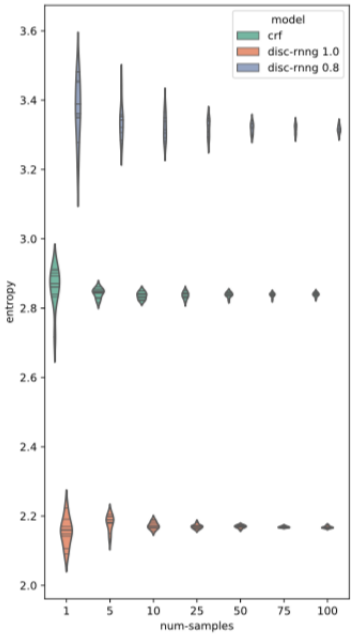

Figure 6: Entropy of a CRF distribution compared to that of RNNG distributions. The CRF (green / middle) shows higher entropy than the RNNG (orange / bottom). We also plot the entropy of an artificially flattened RNNG (blue / top).

score function estimator (or REINFORCE) and baselines (Williams, 1992) including self-critics (Rennie et al., 2017). Whereas this formulation showed to improve the performance of the RNNG as a language model a little, it still turned out rather computationally expensive, which limited our ability to use large unlabelled data. The dynamic programs necessary to sample from the CRF proposal and assess the entropy (Goodman, 1999; Li and Eisner, 2009) do not lend themselves easily to parallelisation in GPUs. Thus we had to resort to online learning with updates based on very few training instances at a time, which greatly increases the variance of gradients.

The work reported here appears in van Stigt (2019)'s MSc thesis,[10] where a much more extensive syntactic evaluation of the models is also reported. Concurrently, Kim et al. (2019) published a very similar formulation focused on fully unsupervised learning. One aspect of their design that allowed it to scale better than ours is that they use binary bracketing grammars. Fixed arity and a single nonterminal category lead to small grammar constants in essential dynamic programs such as the inside-outside algorithm (Goodman, 1999) necessary for sampling and computation of entropy.

Discussion CRFs enjoy properties such as closed-form entropy assessments as well as tractable dynamic programmes such as Viterbi and k-best (Huang and Chiang, 2005). The latter contributed to CRFs outperforming RNNGs as parsers, but the higher entropy of CRF distributions at the end of training made them worse importance sampling distributions. In general, a CRF over n-ary branching labelled trees remains an expensive object to represent and manipulate, posing challenges to gradient-based learning and parallel computations with GPUs. This motivated us to look into making distributions over discrete combinatorial objects more compact such that expensive

[10] Parser: https://github.com/daandouwe/thesis

computations could be avoided and such that high-entropy distributions could be made more confident. This is what we investigate in the next section.

3.1.4 Deterministic Gradients via Tractable Marginalisation

Probability distributions over combinatorial spaces are large objects difficult to represent in a computation graph. Difficulties stem from their combinatorial nature preventing representation by enumeration and rather requiring packed graph-like representations such as hypergraphs (Klein and Manning, 2001; Gallo et al., 1993). Moreover, often the structure of these objects depends on a particular input (for example, different sentences admit different parse trees) and/or the constraints of the domain (for example, projective dependency trees do not admit crossing arcs) in which case instantiating the object requires an efficient combinatorial algorithm (Shieber et al., 1995; Koller and Friedman, 2009). Quantities of interest, such as cumulative probabilities and entropies, require dynamic programmes that factorise their computation along the packed structure without enumeration (Goodman, 1999). Combinatorial sample spaces being so large also implies that one can never reason using all assignments at once, often having to resort to unbiased estimates of quantities of interest, such as marginal probabilities and expectation values, via sub-sampling which is the principle behind Monte Carlo (MC) estimation. Sampling, however, like most selection rules (stochastic or not) poses challenges to computation of gradients. Altogether, representation, computation, and sampling are clear obstacles to gradient-based learning and parallel computation on GPUs, which to a great extent are the backbone of deep learning models.

One way to make distributions over such large sample spaces more manegeable is to make them *sparse*. That is, not all of the possible structures are assigned non-zero probability. A recent line of work has developed differentiable sparse mappings (Martins and Astudillo, 2016; Niculae et al., 2018a; Niculae and Blondel, 2017) that can be used to parameterise sparse probability distributions over discrete sample spaces. These can be used for non-structured objects, such as categories, but also, and crucially, for combinatorial objects, such as sequences and trees. This section reports on a joint effort with researchers from the ERC-funded project DeepSPIN,[11] who pioneered differentiable sparse mappings in deep learning as well as their application to NLP (Niculae et al., 2018b; Peters et al., 2019).[12]

Gradient-based learning of a deep latent variable model $\ell(x, z; \phi)$, where x is observed (e.g., a sentence) and z is latent (e.g., a cluster indicator, a collection of binary factors, a dependency graph, etc.) requires assessing the gradient of the expected value of the loss

$$\mathcal{L}_x(\phi) = \mathbb{E}_{\pi(z|x,\phi)}\left[\ell(x, z; \phi)\right] = \sum_{z \in \mathcal{Z}} \ell(x, z; \phi)\pi(z|x, \phi)\,, \tag{11}$$

with the sum ranging over all possible assignments of the latent variable. For example, VAEs are optimised via maximisation of the evidence lowerbound, which corresponds to having $\ell(x, z; \phi) = -\log p(x, z|\theta) + \log q(z|x, \lambda)$ and $\phi = \theta \cup \lambda$ for a generative model $p(z)p(x|z, \theta)$ and a variational approximation $q(z|x, \lambda)$. Solving the expectation in closed-form, means assessing the loss $\ell(x, z; \phi)$ as many as $|\mathcal{Z}|$ times. Consider the case where z is one of K categories (as in a mixture model), then this involves K assessments of ℓ. In NMT, ℓ is a large encoder-decoder architecture, and maintaining K forward passes through it in memory implies reducing batch size by a factor of K.

[11] https://deep-spin.github.io

[12] At the time of writing, a manuscript presenting this work is under review (Correia et al., 2020).

Typical batch sizes vary from 60 to 100, else special hardware is required, imposing a clear limit on K. Consider the case where z is a binary tree, the number of binary trees is exponential in the length of the input $|x|$, and solving the expectation in closed-form is therefore intractable, no matter how lightweight ℓ might be (and it seldom is). An alternative to assessing the expectation exactly is estimating its gradient via Monte Carlo, this takes the form of

$$\nabla_\phi \mathcal{L}_x(\phi) = \mathbb{E}_{\pi(z|x,\phi)}\left[\ell(x,z;\phi)\nabla_\phi \log \pi(z|x,\phi)\right] \tag{12}$$

$$\overset{\text{MC}}{\approx} \frac{1}{S}\sum_{s=1}^{S} \ell(x,z^{(s)};\phi)\nabla_\phi \log \pi(z^{(s)}|x,\phi) \tag{13}$$

with independent samples $z^{(s)} \sim \pi(z|x,\phi)$. This is known as the score function estimator (SFE; Williams, 1992), or REINFORCE, an estimator whose variance is too high to be useful and which requires variance reduction techniques, some rather complex (Tucker et al., 2017; Grathwohl et al., 2018), and not always very effective. The main advantage of closed-form expectations is that they have deterministic gradients, that is, backpropagation is directly available, without SFE. Clearly, there are very few cases of practical interest where expectations can be computed exactly. To make it possible, we design sparse parameterisations of $\pi(z|x,\phi)$, that is, we make π assign non-zero probability mass to a small subset of outcomes in the sample space of the latent variable. In effect, we use sparse mappings, namely, sparsemax (Martins and Astudillo, 2016) and its structured counterpart, SparseMAP (Niculae et al., 2018a), to parameterise sparse alternatives to Categorical and Gibbs distributions, the former over atomic categories, the latter over combinatorial objects. We also propose a new sparse mapping, namely, top_k sparsemax, convenient for problems where a tractable k-best algorithm is known (such as our CRF parser of the previous section).

Applications We investigate the unstructured case as well as the structured case.[13] In the unstructured case we model a dataset of handwritten digits (MNIST; LeCun et al., 1998) and an emergent communication game (Lazaridou et al., 2017). Here we experiment with mixture models and semi-supervised VAEs (Kingma et al., 2014). In the structured case we again model the MNIST as well as a text auto-complete task (Lee et al., 2019). Here we experiment with a latent factor model, that is, a latent variable model whose latent variable is a collection of binary factors (Mnih and Gregor, 2014). In the MNIST case, this collection has a fixed size (a binary embedding), in the auto-complete case, this collection has variable size (determined by the size of the input). We observe good results across tasks and models managing to achieve the quality of exact marginalisation with the scalability of MC methods. A few highlights: unstructured models achieve such levels of sparsity that only about 2 terms in the expectation needed to be evaluated; structured models required about 10 terms, this is a great reduction of computation effort (from intractably many to only 10 assessments of the downstream loss) for obtaining noise-free gradients.

Discussion Efficient marginalisation is an important development for this work package. It will enable working with distributions over combinatorial objects in a much more scalable way. We are now studying variants of NMT that exploit a combination of ideas presented here and in Section 3.1.3.

[13] A complete paper about this work is currently under review (Correia et al., 2020).

3.2 Latent Structure in Neural Machine Translation

We now turn to NMT with some of the tools developed in the previous section. Our main contribution in this section is a joint generative translation model that allows for induction of sentence-level generalisation. We start with continuous latent variables and then present our first attempts at discrete ones.

3.2.1 Continuous Latent Structure in NMT

We first consider learning continuous latent structure in bitext. Continuous latent structure poses fewer challenges in deep learning systems due to the availability of reparameterisations for efficient gradient estimation (Kingma and Welling, 2014; Rezende et al., 2014), at the cost of being less directly interpretable. In Eikema and Aziz (2019) we introduce AEVNMT: auto-encoding variational neural machine translation. In AEVNMT, source and target sequences are jointly generated from a shared continuous latent space. The latent space can capture hidden variation in translation data. Such variation can be more explicit, such as domain or making a distinction between back-translated data (Sennrich et al., 2016b) and genuine parallel data, or more implicit, such as differences in translator or translation direction.

We model a translation pair (x, y) as being generated from the marginal of a latent variable model:

$$p(x, y) = \int p(z)p(x|z)p(y|x, z)\,\mathrm{d}z \tag{14}$$

The model generates a latent variable z from the prior $p(z)$, followed by a source sentence conditioned on the latent variable (essentially a sentence variational auto-encoder (Bowman et al., 2016)), followed by a translation generated from the source sentence and the latent variable. Following Kingma and Welling (2014) we impose a multivariate diagonal Gaussian prior on z. We train our system using variational inference (Jordan et al., 1999; Blei et al., 2017) with reparameterised gradients. When making predictions with the model at test-time, we infer the latent variable from the source sentence only and condition generations on the mean.

In experiments we test our formulation on English-German data in settings of varying amounts of variation present in the data: in-domain training, mixed-domain training and training with synthetic back-translated data. In the mixed-domain and synthetic data settings we expect the latent variable to capture these larger variations in the data. We find consistent improvements upwards to 0.8 BLEU in almost all settings tested. In later experiments on low-resource Gourmet data, we have found similar consistent marginal improvements on top of traditional NMT.

However, we do not observe a strong correlation between the amount of perceived variation in the data and the improvement in BLEU. This begs the question what the latent variable does capture. Using diagnostic classifiers (Alain and Bengio, 2017; Hupkes et al., 2018) we find that the latent variable does capture domain quite well (accuracy above 90%), and is predictive of the quality of back-translations. However, we find that a traditional NMT model's hidden state performs equally well on this task. We therefore conjecture that the latent variable captures more nuanced variations in the data.

As part of this research we have released multiple codebases. A first one to exactly replicate the experiments in this work,[14] and a second one that acts as a more general purpose codebase for deep

[14] https://github.com/Roxot/AEVNMT

generative modeling for machine translation.[15]

Data augmentation Generative models map similarity in data space to a low-dimensional latent space. This can be used for example to interpolate between different aspects of two sentences. This is an application of sentence-level generalisation to data augmentation, a line of work UVA and UA are currently developing under work package 1.

3.2.2 Discrete Latent Structure in NMT

AEVNMT (Eikema and Aziz, 2019) is a deep latent translation model of bitext that captures the relationship between sentence pairs holistically in a continuous representation. In this section we present strategies to move on toward discrete generalisations.

Mixture Model The first attempt is a change to AEVNMT's prior. Rather than a single Gaussian distribution we attempt to mix K independently parameterised components using a mixture model prior:

$$p(z|\theta, \omega) = \sum_{k=1}^{K} \omega_k p(z|k, \theta) \; . \tag{15}$$

For posterior inference, we may approximate $p(z|x, \theta)$ with a single parameterised component $q(z|x, \lambda)$ or with a mixture model

$$q(z|x, \lambda) = \sum_{k=1}^{K} q(k|x, \lambda) q(z|k, \lambda) \; . \tag{16}$$

The former is simpler as it has no impact in gradient estimation of the ELBO, whereas the latter requires either explicit marginalisation of the component assignment or gradient estimation via MC. For very small K (e.g., less than 10) we may be able to get around with exact marginalisation by incurring a 10-fold decrease of batch size. For larger values of K, we need MC estimation or the technique introduced in Section 3.1.4. For now, we have considered the first two options.

VampPrior The VampPrior is a strong prior that approximates the so called aggregated posterior $q(z) = \sum_{x \in \mathcal{X}} p_\star(x) q(z|x, \lambda)$. This aggregated prior is clearly intractable, since the summation ranges over *all possible sentences* that may ever exist, and the data distribution $p_\star(x)$ is unknown. Tomczak and Welling (2018) propose an approximation to it based on averaging posterior distributions

$$p(z|v) = \frac{1}{K} \sum_{k=1}^{K} q(z|v_k, \lambda) \tag{17}$$

for K learned pseudo-inputs $v_1, \ldots, v_K$. Pseudo-inputs are points in data space, which for Tomczak and Welling (2018) corresponds to fixed-dimension vectors of real values, since they model pixel intensities. In NMT, inputs are sequences of tokens, thus the VampPrior is not directly available. If we see the embedding layer as fixed, we can think of inputs as sequences of token embeddings, which, except for sequence length, are continuous variables. We propose to sample a number of

[15] https://github.com/Roxot/AEVNMT.pt

sequence lengths following a Poisson approximation to the length distribution of the training data, and have those fixed throughout training. Let ℓ_k denote the length of the kth sequence, we define $v_k = \langle \mathbf{v}_1^{(k)}, \ldots, v_{\ell_k}^{(k)} \rangle$ a sequence of continuous vectors which are seen as parameters of the prior. Our adaptation to the VampPrior has been presented in Pelsmaeker and Aziz (2020) in the context of deep language models, where it was amongst the top performing strategies to avoid posterior collapse.

Latent Factors Mixture models struggle with the problem that as the number of components increase, each component will be responsible for explaining less and less data. In a low-resource setting this might be undesirable. A different type of inductive bias is that of a latent feature model or latent factor model (Ghahramani and Griffiths, 2006), where a collection of discrete (typically binary) attributes conspire to generate data points, rather than compete to do so (as components do in mixture models). This can be thought of as a form of unsupervised overlapping clustering. For K binary factors, the number of possible assignments is 2^K, which makes marginalisation clearly intractable. We investigate two possibilities, score function estimation, and the sparse continuous relaxation introduced in Section 3.1.2. To use the HardKuma relaxation we need to design a prior, Murady (2020) introduces the HardUniform prior and derives crucial quantities such as entropies and KL divergences involving HardKuma and HardUniform distributions. In the binary case, we modify AEVNMT to have z be D-dimensional with each $z_k \in \{0, 1\}$, in the relaxed case each $z_k \in [0, 1]$. Then the inference model becomes

$$Z_k | x, \lambda \sim \text{Bernoulli}(\gamma_k) \qquad \gamma = g(x; \lambda) \tag{18}$$

in the binary case, and

$$Z_k | x, \lambda \sim \text{HardKuma}(\alpha_k, \beta_k) \qquad [\alpha, \beta] = g(x; \lambda) \tag{19}$$

in the relaxed base. The parameters γ (or α and β) are predicted by the inference network: D probability values in the Bernoulli case, and $2 \times D$ positive scalars in the HardKuma case.

Word Alignments We investigate the potential of word alignments to inform AEVNMT's inference model as a side loss. This corresponds to optimising the objective

$$\mathbb{E}_{q(z|x,y,\lambda)} \left[\log p(y|z, x, \theta) \right] - \text{KL}(q(z|x, y, \lambda) || p(z|\theta)) + \mathbb{E}_{q(z|x,y,\lambda)} \left[\log p(y|z, x, \phi) \right] \tag{20}$$

where the last term is an IBM 1 factorisation (Brown et al., 1993) of the probability of the target sentence given the source (and the latent code). The component is similar to the L2 decoder of Rios et al. (2018), that is, we transform z to a sequence of $|x|$ Categorical distributions using a recurrent neural network $g(z; \phi)$ and then compute

$$p(y_j | z, x, \phi) = \sum_{i=1}^{|x|} \text{Cat}(y_j | g_i(z; \phi)) \, . \tag{21}$$

The idea here is that we may be able to work with a continuous z while making it predictive of discrete abstractions, word alignments in this case. This would greatly simplify the machinery necessary for training.

Findings Mixture models have shown to disentangle some superficial aspects of the data: whereas some components capture length, other components captured noisy observations.[16] In particular, we noticed that only very few components were active, which suggests that the model struggles to make use of the additional capacity offered by the mixture model formulation. As for the VampPrior, while it has been shown to be a powerful prior in other contexts, it introduces many more parameters that need to be estimated, in particular, $K \times L \times d$ additional parameters, where K is the number of pseudo-inputs (e.g., 100), L is the average sequence length (e.g., 20), and d the dimensionality of word embeddings (e.g., 512). Estimating so many parameters proved challenging leading us to favour more compact formulations, such as binary factor models. In the binary factor model, we have a single inference model that parameterises all D factors, rather than K mostly independent components. So far we have observed factor models perform on par with Gaussian AEVNMT models. Factors, unlike components in mixture models, seem to remain active (we have experimented with up to 64 factors). Preliminary analysis of what these unsupervised factors capture have not shown systematic structure, unfortunately. Next in line for factor models is to supervise them with noisy linguistic structure available for English (for example, predicted by a syntactic or semantic parser). Finally, while the neural IBM 1 side loss converges well, with meaningful word pair associations, we observe no effect on translation performance. This suggests that the NMT model does not find useful learning signal in lexical alignments, at least not via this multi-task learning formulation.

Discussion Our future efforts in this line include to exploit noisy supervision potentially available for English (e.g., predicted linguistic structure) to bias latent factors, and to embrace the methodology put forward in Section 3.1.4. The latter, in particular, enables stable training, and crucially, is amenable to more complex structure allowing, for example, tighter integration of components such as an alignment model or a syntactic language model.

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 dataset $\mathcal{D}$

$$p(\mathcal{D}|\theta) = \prod_{s=1}^{|\mathcal{D}|} p(y^{(s)}|x^{(s)}, \theta) \tag{22}$$

as if sentence pairs in $\mathcal{D}$ were independent of one another. So called *document-level* NMT models the probability of a document $C = \langle (x^{(1)}, y^{(1)}), \ldots, (x^{(|C|)}, y^{(|C|)}) \rangle$

$$p(C|\theta) = \prod_{s=1}^{|C|} p(y^{(s)}|x^{(1:s)}, y^{(1:s-1)}, \theta) \tag{23}$$

[16] Perhaps this could be used as a tool to help clean crawled data.

without making the independence assumptions of standard NMT. The likelihood of a dataset of document pairs then factorises independently over documents, 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 this 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.

Summary of contributions:

- we propose a unified treatment of sentence-level and document-level NMT which leads to improved results in high-resource language pairs;
- we also propose methodology to evaluate document-level NMT with regards to the model's sensitivity to document-level linguistic phenomena such as coreference and conduct a systematic evaluation of existing approaches, again in a high-resource setting;
- we collect parallel data preserving not only document alignment, but also document substructured mined from metadata, and investigate the effects of the latter on document-level NMT; that includes one of the GoURMET language pairs, namely, Bulgarian-English.

3.3.1 Making the most of context in NMT

Document-level machine translation manages to outperform sentence level models by a small margin, but thus far have not been widely adopted. In this section we present work (Zheng et al., 2020) which argues that previous research did not make a clear use of the global context, and we propose a new document-level NMT framework that deliberately models the local context of each sentence with the awareness of the global context of the document in both source and target languages. We specifically design the model to be able to deal with documents containing any number of sentences, including single sentences. This unified approach allows our model to be trained elegantly on standard datasets without needing to train on sentence and document level data separately. Experimental results demonstrate that our model outperforms Transformer baselines and previous document-level NMT models with substantial margins of up to 2.1 BLEU (Papineni et al., 2002) on state-of-the-art baselines. We also provide analyses which show the benefit of context far beyond the neighbouring two or three sentences, which previous studies have typically incorporated.

Figure 7 briefly illustrates typical context-aware models, where the source and/or target document contexts are regarded as an additional input stream parallel to the current sentence, and incorporated into each layer of encoder and/or decoder (Zhang et al., 2018; Tan et al., 2019). More specifically, the representation of each word in the current sentence is a deep hybrid of both *global* document context and *local* sentence context in every layer. We notice that these hybrid encoding approaches have two main weaknesses:

- *Models are context-aware, but do not fully exploit the context*. The deep hybrid makes the model more sensitive to noise in the context, especially when the context is enlarged. This could explain why previous studies show that enlarging context leads to performance degradation. Therefore, these approaches have not taken the best advantage of the entire document context.

- *Models translate documents, but cannot translate single sentences*. Because the deep hybrid requires global document context as additional input, these models are no longer compatible with sentence-level translation based on the solely local sentence context. As a result, these approaches usually translate poorly for single sentence documents without document-level context.

We mitigate the aforementioned two weaknesses by designing a general-purpose NMT architecture which can fully exploit the context in documents of arbitrary number of sentences. To avoid the deep hybrid, our architecture balances *local context* and *global context* in a more deliberate way. More specifically, our architecture independently encodes local context in the source sentence, instead of mixing it with global context from the beginning so it is robust to when the global context is large and noisy. Furthermore our architecture translates in a sentence-by-sentence manner with access to the partially generated document translation as the target global context which allows the local context to govern the translation process for single-sentence documents.

We highlight our contributions:

- We propose a new NMT framework that is able to deal with documents containing any number of sentences, including single-sentence documents, making training and deployment simpler and more flexible.
- We conduct experiments on four document-level translation benchmark datasets, which show that the proposed unified approach outperforms Transformer baselines and previous state-of-the-art document-level NMT models both for sentence-level and document-level translation.
- Based on thorough analyses, we demonstrate that the document context really matters; and the more context provided, the better our model translates. This finding is in contrast to the prevailing consensus that a wider context deteriorates translation quality.

By the definition of local and global contexts, general translation can be seen as a hierarchical natural language understanding and generation problem based on local and global contexts. Accordingly, we propose a general-purpose architecture to exploit context to a better extent.

Figure 8 illustrates the idea of our proposed architecture:

- Given a source document, the encoder builds local context for each individual sentence (local encoding) and then retrieves global context from the entire source document to understand the inter-sentential dependencies (global encoding) and form hybrid contextual representations (context fusion). For single sentence generation, the global encoding will be dynamically disabled and the local context can directly flow through to the decoder to dominate translation.

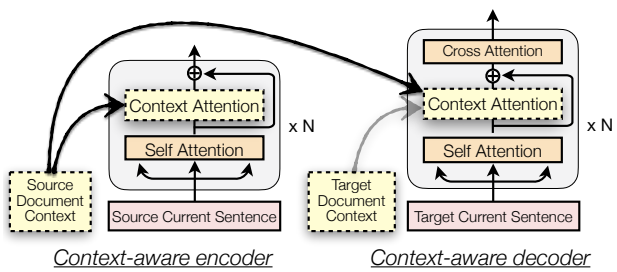

Figure 7: Illustration of typical Transformer-based context-aware approaches (some of them do not consider target context (grey line)).

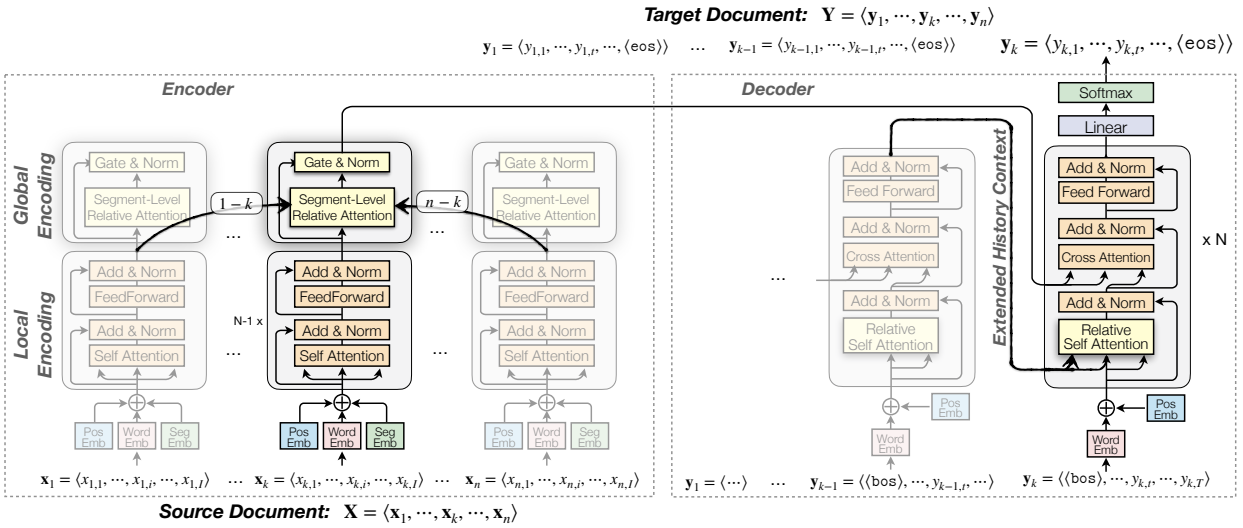

Figure 8: Illustration of the proposed model. The local encoding is complete and independent, which also allows context-agnostic generation.

- Once the local and global understanding of the source document is constructed, the decoder generates target document by sentence basis, based on source representations of the current sentence as well as target global context from previous translated history and local context from the partial translation so far.

What Does Model Learn about Context? A Case Study. Furthermore, we are interested in what the proposed model learns about context. In Figure 9, we visualise the sentence-to-sentence attention weights of a source document based on segment-level relative attention. As shown in Figure 9, we find very interesting patterns (which are also prevalent in other cases): 1) first two sentences (blue frame), which contain the main topic and idea of a document, seem to be a very useful context for all sentences; 2) the previous and subsequent adjacent sentences (red and purple diagonals, respectively) draw dense attention, which indicates the importance of surrounding context; 3) although sounding contexts are crucial, the subsequent sentence significantly outweighs the previous one. This may imply that the lack of target future information but the availability of the past information in the decoder forces the encoder to retrieve more knowledge about the next sentence than the previous one; 4) the model seems not to care about the current sentence. Probably because the local context can flow through the context fusion gate, the segment-level relative attention just focuses on fetching useful global context; 5) the 6-th sentence also gets attraction by all the others (brown frame), which may play a special role in the inspected document.

Discussion Document-level NMT has mostly concentrated on high-resource language pairs, partly due to the need for large Transformer-based architectures. To address this limitation we envision combining the model presented here with ideas from latent variable models of translation (Zhang et al., 2016; Eikema and Aziz, 2019). This takes the form

$$p(C|\theta) = \int p(z|x^{(1)}, \ldots, x^{(|C|)}) \prod_{s=1}^{|C|} \underbrace{p(y^{(s)}|x^{(s)}, z, \theta)}_{\text{simplified doc-level NMT}} \, \mathrm{d}z \tag{24}$$

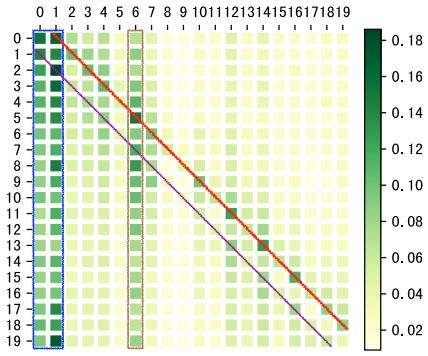

Figure 9: Visualization of sentence-to-sentence attention based on segment-level relative attention. Each row represents a sentence while each column represents another sentence to be attended. The weights of each row sum to 1.

where marginalisation of a document-level latent representation breaks the independence assumptions of the conditional model without necessarily introducing direct conditioning on rich context. Moreover, this can exploit pre-trained architectures for conditioning on entire documents, since now there is no need for a decoder that is document aware, but only an encoder.

3.3.2 Comparison of document-level NMT approaches

This work, published at EAMT (Lopes et al., 2020) and in collaboration with Unbabel, provides a systematic comparison of a representative set of document-level NMT approaches for English-to-French, English-to-German and English-to-Brazilian Portuguese, tested in realistic high-resource settings for two different scenarios: standard document-level translation and chat translation. We evaluate the methods using (i) the standard automatic metric BLEU (Papineni et al., 2002), (ii) contrastive test sets, designed to test the ability of the approaches to use sentence-external context, and (iii) manual annotations using proprietary chat data. Two additional contributions are:

- a new document-level approach, called the Doc-Star-Transformer (inspired by the star transformer's (Guo et al., 2019) efficient method of computing pairwise comparisons), which is capable of incorporating arbitrary long sequences from a document;
- a new large-scale contrastive pronoun test set for English-to-French,[17] similar to the large-scale test set for German pronouns (Müller et al., 2018).

The use of linguistic context (from the surrounding sentences or entire document) in NMT has become a popular area of interest, guided by the development of full-scale neural architectures, which facilitate the inclusion of additional representations such as those based on context. Many different solutions to include additional context (from both the source and target language) have been proposed, for example concatenating multiple sentences together and translating them simultaneously (Tiedemann and Scherrer, 2017), using multiple encoders to encode additional context separately (Zhang et al., 2018) and using a cache mechanism to remember previously decoded target words (Tu et al., 2018).

However, the various approaches have often been evaluated in different settings (evaluated on different language pairs or datasets and sometimes only in artificially low resource settings, where

[17] Freely available at https://github.com/rbawden/Large-contrastive-pronoun-testset-EN-FR under an MIT licence.

performance gains are easier to achieve when adding additional model parameters). Our systematic comparison therefore aims to evaluate the real potential of these approaches and ascertain which strategies work best for exploiting linguistic context.

Full details of the methods, experiments and results can be found in the published article (Lopes et al., 2020), and we report only the more interesting results and conclusions below.

The BLEU score results show that once a strong baseline is used and the methods trained in high-resource settings, few of them score significantly higher than the baseline. However, BLEU scores are a poor indicator of how well the models are actually exploiting context and do not correlate to the performance on the targeted contextual evaluation (Hardmeier, 2012; Bawden, 2018).

The targeted contextual evaluation using contrastive test sets looks at how well each model handles linguistic phenomena that typically require extra-sentential context for the translation to be correct. We test two phenomena:

- Anaphoric pronoun translation, whereby the French/German translations of the neutral English pronuns *it* and *they* are gender-marked, depending on which noun they refer to (which can appear outside of the current sentence). For German, we have *Es* (masc.), *Sie* (fem.) and *Es* (neuter) as translations of English *it*. For French we have *il* (masc.) and *elle* (fem.) for *it* and *ils* (masc.) and *elles* (fem.) for *they*.
- Lexical choice (regrouping lexical coherence lexical cohesion), whereby a source term may be ambiguous in it translation without disambiguating context, such as that provided by a previous sentence.

The idea of these test sets is to evaluate each model on its ability to rank correct translations higher than incorrect contrastive variants in which only the elements being evaluated are modified. An example is shown in Figure 10, where the correct pronoun *elles*, corefering with the feminine noun *roses* is modified with the masculine variant *ils* in the incorrect contrastive example.

Context sentence	
	Some red **roses** for Your Ladyship.
	Des **roses**$_{\text{fem.}}$ pour madame.
Current sentence	
	Who could **they** be from?
✓	De qui peuvent-**elles**$_{\text{fem.}}$ bien être ?
×	De qui peuvent-**ils**$_{\text{masc.}}$ bien être ?

Figure 10: An example from the large-scale EN→FR contrastive test set.

We evaluate on large-scale anaphoric pronoun test sets for both English-to-German and English-to-French using the test set from (Müller et al., 2018) and our large-scale test for French. In addition, we evaluate on two smaller manually crafted test sets form (Bawden et al., 2018), which are designed such that all sentence are ambiguous without preceding context and explicitly testing the models' ability to use preceding context.

The results of the contrastive evaluation is shown in Table 6. The scores are decomposed into target pronouns for each of the large-scale test sets and total scores are given for the two smaller French test sets (Anaphora and Coherence/cohesion). Interestingly, we see that the best scoring model

overall is the simple approach of concatenating the current sentence to the previous sentence on both the source and target side of the data (Concat2to2) and involves no modification to the baseline model architecture. It generally performs better than those that include specific mechanisms for integrating context (cache, multi-source encoder and the proposed Star approach). Notably, the Concat2to2 approach is the only one to achieve scores that are clearly higher than the baseline on the two smaller contrastive sets, showing that it is the only one that can reliably integrate the preceding context. This supports a conclusion made in (Bawden et al., 2018) that this type of model in which the previous translations are channelled through the decoder is more effective than representing the context using a different mechanism.

	EN→DE				EN→FR						
	Total	Es	Sie	Er	Total	it		they		Anaphora	Coherence/ cohesion(%)
						elle	il	elles	ils	All	All
Baseline	45.0	91.9	22.9	20.2	79.7	88.1	82.7	76.1	72.2	50.0	50.0
Concat2to1	48.0	91.6	**27.1**	**25.3**	**80.9**	**88.4**	**83.3**	**77.2**	**73.9**	50.0	**52.5**
Concat2to2	70.8	91.8	**61.9**	**58.7**	**83.2**	**89.2**	**86.2**	**80.4**	**77.6**	**82.5**	**55.0**
Cache	45.2	**92.1**	**23.5**	19.9	79.7	88.0	82.7	76.0	72.0	50.0	50.0
Multi-src Enc	42.6	62.3	**33.9**	**31.5**	59.0	62.0	61.3	57.2	57.3	47.0	46.5
Star, 8 heads	45.9	91.3	**27.0**	19.5	79.6	88.0	82.6	76.1	72.0	50.0	50.0

Table 6: Accuracies (in %) for the contrastive sets. Methods outperforming the baseline are in bold.

The general conclusions of this work are therefore (i) the importance of evaluating document-level models in high-resource scenarios with a strong baseline trained on large quantities of data (as performance gains can disappear in this setting) and (ii) the importance of evaluating these models using targeted contextual evaluation, as BLEU scores do not reveal whether the models' handling of context is being improved. We find that despite many of the models having previously shown improved performance when compared to a baseline, this does not always hold in more realistic scenarios such as the ones we test. Moreover, the simplest approach we test has the best results, suggesting that context is best exploited when integrated into the existing functionalities of the NMT architecture.

3.3.3 Exploiting document sub-structure in NMT

In a work published at LREC 2020 (Dobreva et al., 2020), we investigate the usefulness of document sub-structure as a source of information for NMT. Some documents, such as biographies and encyclopedia entries, have a regular structure, with different articles containing sections which can have similar vocabulary. These similarities can be exploited in MT as has been shown in (Louis and Webber, 2014) in the framework of SMT. We adapt this idea to NMT, exploring two different methods of integrating document structure information into NMT models: one using side constraints (Sennrich et al., 2016a) and the other using a cache-based neural model (Kuang et al., 2018). To this end we develop monolingual and parallel corpora for three language pairs (French-English, Bulgarian-English and Chinese-English) consisting of Wikipedia biographies.[18] The corpora preserve the section headings associated with each sentence.

[18] Freely available at https://github.com/radidd/Doc-substructure-NMT.

To create the corpora, we use Wikimedia dumps to extract Wikipedia articles marked in the page metadata as biographies or as describing certain categories of people, such as "writer" or "politician". Wikipedia metadata also contains information about whether a page is a translation from a corresponding page in another language. For the creation of the parallel corpora we use this information to gather biographies that are translated from the relevant language for each of the language pairs. Document alignment is therefore trivial, as the source page for every translated article is identified via the metadata. Sentence alignment is done using Hunalign (Varga et al., 2005) and Bleualign (Sennrich and Volk, 2011). We preserve information about the structure of documents by preserving the name of the sections with which each sentence in the corpus is associated. We also create monolingual corpora of Wikipedia biographies for each of the four languages, as well as validation and test sets for each of the language pairs. The test sets are separated based on the original language of the text.

We experiment with two methods for integrating document sub-structure information with NMT: side constraints (Sennrich et al., 2016a) and a cache-based neural model (Kuang et al., 2018). Both methods are implemented using the Transformer architecture (Vaswani et al., 2017) and both rely on topic models trained on sections in the articles. The topic models are used to to infer the topic of each section and provide topic representations. This is done since the section titles themselves are not informative enough for the purpose of NMT. We train LDA models (Blei et al., 2003) separately for each of the four languages and use a simple source to target topic projections based on which target topic most commonly occurs with the given source topic. In the side constraints method the topics are integrated in the form of tokens prepended to the source sentence. In the neural cache model method, two caches are passed to the model: one containing the most pertinent vocabulary to the current topic based on the learned topic models, and the other a set number of tokens from the preceding sentences in the current section. The cache model calculates a probability distribution over the vocabulary and is integrated in a shallow way with the NMT model.

We observe that using section level information, as opposed to document level information, does not prove to be advantageous and in some cases appears to degrade model performance. Results differ for the three language pairs, with the side constraints method showing better performance for Bg-En and Zh-En, but not for Fr-En. For all language pairs results differ on the two test sets (originally English text vs. originally {Bg/Fr/En} text), with BLEU scores on the originally English text being consistently higher.

4 Task 3.3 – Probabilistic Neural Machine Translation

NMT models are statistical models trained to output a probability distribution over a large discrete combinatorial space, i.e, the space of all translations of a given input. More often than not, the probabilistic view of NMT is overlooked. While it does motivate a criterion for parameter estimation, namely, maximum likelihood, other implications are seldom exploited.

For example, NMT models are used as if they were discriminative margin-based classifiers, that is, after training we form predictions by searching for a handful of highest-scoring translations. Whereas in the unstructured case, that seems intuitive and works well in practice, in the structured (combinatorial) case, the analogy quickly falls apart, especially in situations of high uncertainty, such as in low-resource settings (Eikema and Aziz, 2020).

An alternative to mode-seeking algorithms that capitalises on the distribution as a whole is a form of probabilistic voting known as minimum Bayes risk (MBR) decoding (Bickel and Doksum, 1977) which was popular in days of statistical machine translation (Kumar and Byrne, 2004; Tromble et al., 2008). MBR allows decisions to be guided by sentence-level (or even document-level) statistics of samples from the model distribution. This is in contrast with decision rules based on path-finding algorithms, like beam search (Sutskever et al., 2014), which are are constrained by the incremental (left-to-right) view with which the model factorises the translation probability.

Another example, we always regularise maximum likelihood training, as otherwise high-capacity neural networks would simply learn to memorise the training data and fail to translate unseen text. Some regularisation techniques are better suited for unstructured classification, and though they have been shown beneficial in high-resource settings, the situation is far less clear in the low-resource case. One such a technique is label smoothing (Szegedy et al., 2016; Pereyra et al., 2017). Understanding what aspects of label smoothing are beneficial in a low-resource setting and may lead to better regularisation and thus better generalisation. One way of doing so is through the language of Bayesian priors on parameter and/or function space. Other regularisation techniques, emphasise a probabilistic account, for example Bayesian dropout (Gal and Ghahramani, 2016), but they are typically not fully exploited after training.

The probabilistic point of view has a major advantage, namely, uncertainty management. Uncertainty is sometimes seen as a problem, but we argue this is mostly so because of the ways in which predictions are traditionally formed, namely, via mode-seeking search algorithms such as beam search. Harnessed well, it can be used to make the most out of little data.

Objectives:

- 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.

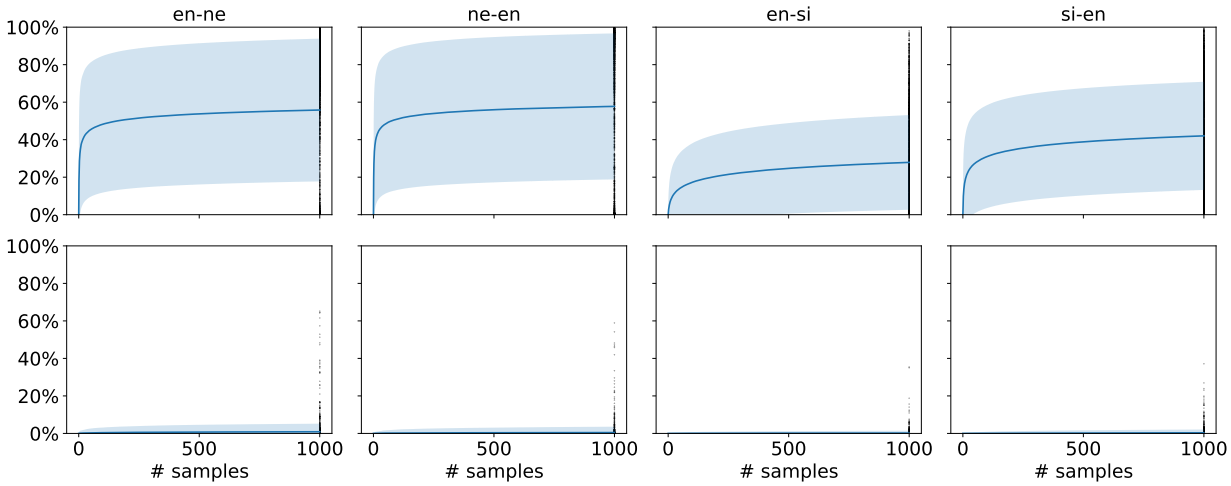

Figure 11: Cumulative probability of 1,000 ancestral samples on the held-out in-domain (top) and FLORES (bottom) test sets. The dark blue line shows the average cumulative probability over all test sentences, the shaded area represents 1 standard deviation away from the average. The black dots to the right show the final cumulative probability for each individual test sentence.

4.1 The Inadequacy of the Mode in NMT

NMT models are explored by deterministic mode-seeking algorithms such as beam search (Graves, 2012; Boulanger-Lewandowski et al., 2012; Sutskever et al., 2014), these algorithms neglect a lot of information encoded in a distribution. In high-resource settings, models may be confident on their predictions and it is plausible that disregarding the distribution as a whole still leads to good results. In a low-resource setting we need to get all we can from the model, and that means we need to use its uncertainty in our favour, rather than ignore it. One way to do so is to explore NMT models as probability distributions, which they are as they are trained on maximum likelihood, and design probabilistic criteria for predictions.

In Eikema and Aziz (2020) we do so by examining the probability distributions learned by NMT systems on English-Nepali and English-Sinhala data. Both these language pairs have very few high quality parallel resources. Moreover, there is little to no in-domain data available for these language pairs. We mimic the training setup of Guzmán et al. (2019) and examine the distributions obtained on the FLORES (Guzmán et al., 2019) test set, as well as on a small held-out portion of the training data.

First of all, we assess the fit of the model on the held-out portion of the training data. We use hierarchical Bayesian models to model several aspects extracted from sampled translations, beam search translations, and the gold-standard data. We look at length, lexical aspects of the data (through unigrams and bigrams), and word order aspects of the data (through skip-bigrams). In all cases we find that samples from the model are able to recover statistics of the data reasonably well, but that beam search translations stray from the statistics of the data.

We also explore the set of translations that are likely under the model, both in the test domain (FLORES) and in the training domain (held-out data). We sample 1,000 translations on both test sets and plot the cumulative probability of unique translations in Figure 11. What we find is that the

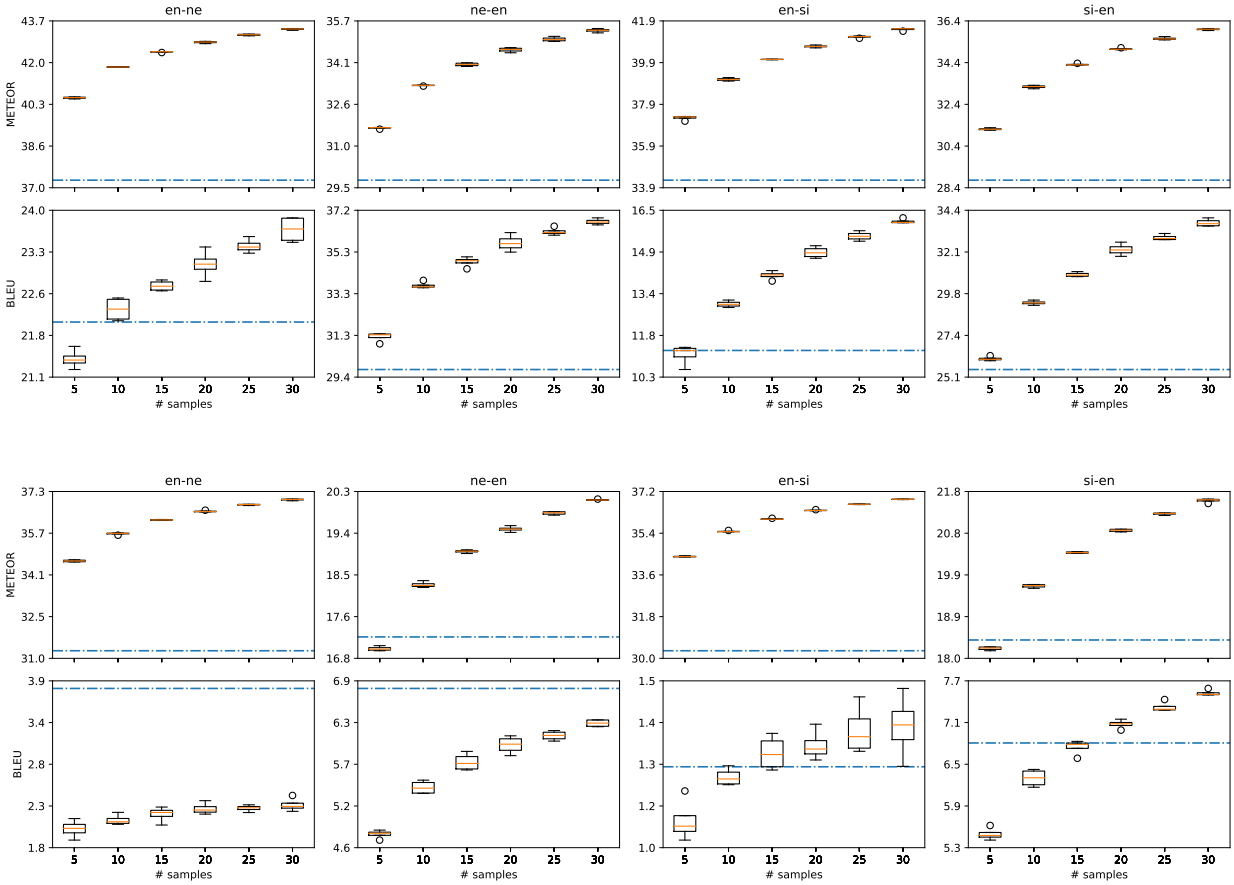

Figure 12: METEOR and BLEU scores for oracle-selected samples as a function of sample size on the held-out in domain (top) and FLORES (bottom) test sets. For each sample size we repeat the experiment 4 times and show a box plot per sample size. The blue lines show beam search scores.

learned probability distributions are extremely flat in most cases, especially in the test domain. For many cases (over half on the inputs in the test domain) we even sample 1,000 unique translations. We also find that the beam search solution is very rarely sampled in the test domain (less than 10% of the inputs). The mode of the distribution therefore seems rather arbitrary, as it represents only a tiny amount of probability mass.

When we look at the average quality of the samples in terms of automatic evaluation metrics like BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2011), we find that samples perform between 0.4 and 8.2 BLEU and 0 and 4.7 METEOR worse than beam search, but that overall the results are still decent. The variance in those results is very small, less than 0.2 BLEU and less than 0.1 METEOR in all cases. Note that the comparison with beam search here is not very fair, as random sampling is no decision rule. We experiment with selecting the best sample out of a set of samples using an oracle decision rule (i.e. we have access to the reference for the selection process). We select on sentence-level METEOR and report the results for varying sample sizes in Figure 12. What we observe is that within a set as small as 10 samples we can already outperform beam search if we knew how to select the best sample.

This motivates us looking into sampling-based decision rules, we experiment with minimum Bayes

risk (MBR) decoding (Bickel and Doksum, 1977; Kumar and Byrne, 2004). Using a straightforward implementation of MBR where we obtain the hypothesis space using 30 samples from the model and use sentence-level METEOR as the utility function. We find that in all comparisons MBR beats or matches beam search performance, improving up to 4.5 METEOR. This is a promising result for future research into practical sampling-based decision rules.

In Eikema and Aziz (2020) we also connect MAP decoding to many pathologies (Sountsov and Sarawagi, 2016; Koehn and Knowles, 2017; Murray and Chiang, 2018; Ott et al., 2018; Kumar and Sarawagi, 2019; Stahlberg and Byrne, 2019) and biases (Ranzato et al., 2016; Eikema and Aziz, 2020) observed in NMT. We show that criticising models in terms of statistics of the mode is a bias in its own right, and argue that criticisms of probabilistically trained NMT models should be based on the entire distribution rather than the mode. The results found in our experiments suggest that the mode is arbitrary in many cases, as the set of likely translation is vast and no single translations stands out over any other. Therefore, it seems that a more logical decision rule would be one that takes into account more properties of the entire distribution rather than just the mode. This is especially so in low-resource scenarios, where models are trained on little data and the test domain often differs much more from the training domain.

4.2 On Probabilistic Alternatives to Label Smoothing

Label smoothing (LS; Szegedy et al., 2016; Pereyra et al., 2017) is a parameter estimation criterion for Categorical-likelihood models which is inspired by, though not equivalent to, maximum likelihood estimation (MLE). In LS, the likelihood of the observed class is weighted by $1 - \epsilon$ and the sum of likelihoods assigned to non-observed classes is weighted by $0 < \epsilon < 1$. In comparison to a standard MLE procedure for a Categorical-likelihood model, label smoothing imposes a stronger preference over hypotheses: for a given probability p of the observed class, standard MLE is not sensitive to how the probability $1 - p$ spreads over the non-observed classes, that is, MLE is not sensitive to the entropy assigned to non-observed classes, whereas LS prefers hypotheses that assign highest entropy over non-observations. While LS has been shown to improve NMT for high-resourced languages (Ott et al., 2018), the situation is far less clear in low-resource settings (Eikema and Aziz, 2020), as we uncovered in Section 4.1. Label smoothing boosts the model's confidence on the observed class at the expense of the ranking over non-observed classes. Whereas this seems to have a positive effective on predictions formed by deterministic algorithms, such as beam search, it has a very negative effect on the probabilistic interpretation of NMT models, leading to distributions that cannot be explored well generatively (that is, via sampling), something our community has already noticed in high-source settings (Graça et al., 2019).

We speculate that there is a relationship between LS, the amount of data, and predictions by local search algorithms. If we have enough data to estimate model parameters, greedy predictions are likely good and there's little risk in outputting distributions that are close to deterministic. Where we have less data, perhaps there is worth in the uncertainty maintained by MLE, though that cannot be exploited by deterministic mode-seeking decision rules such as MAP decoding. In fact, we demonstrate that, see Eikema and Aziz (2020, Appendix B), where a probabilistic decision rule for an MLE-trained model outperforms beam-search for a LS-trained model.

Here, we investigate alternative formulations of label smoothing emphasising MLE-training with the goal of bringing some of the improvements that LS has shown in high-resource settings to the low-resource scenario. Another goal of this research is to tap onto such potential while still exploring the model with probabilistic devices (i.e., ancestral sampling) that enable probabilistic

decision rules. Our general strategy is to let non-observed classes be scored by our objective, as they are in LS, not by flattening the distribution over non-observations, as LS does, but rather by *constraining* output distributions to on average follow a Dirichlet posterior that reflects expected word frequencies conditioned on the available training data.

Formulation Preferences over hypotheses are naturally expressed in the language of priors. For a Categorical likelihood model, a Dirichlet prior can express the kinds of preferences we have discussed so far, namely, a particular ranking (flat or otherwise) over non-observed classes. We are going to use a simple conjugate Bayesian model to operationalise such preferences.

We observe $\mathbf{y}$ a data vector of counts which we model under a Dirichlet-Multinomial model

$$\phi|\alpha \sim \text{Dir}(\alpha\mathbf{1}) \tag{25a}$$

$$\mathbf{Y}|\phi \sim \text{Multinomial}(N, \phi) \tag{25b}$$

where $N = \sum_{i=1}^{v} y_i$ is the total number of observations and α specifies a sparse symmetric Dirichlet prior. The posterior is

$$\phi|y, \alpha \sim \text{Dir}(\alpha\mathbf{1} + \mathbf{y}) \,, \tag{25c}$$

which follows from the conjugacy between Dirichlet prior and Multinomial likelihood.

We want to train a neural network $f(x;\theta)$ to parameterise a Categorical distribution over observations $Y|\theta, x \sim \text{Cat}(f(x;\theta))$, from which the observed class y is sampled. NMT is much like that, where x corresponds to a source sentence and a target prefix, both observed, and y corresponds to the next word in the observed target sentence. The NN $f(x;\theta)$ is the entire encoder-decoder with attention architecture including a softmax output layer that parameterises the Categorical likelihood. In NMT this NN is used repeatedly, with varying inputs, to generate (rather assess the probability of) every token in the observed target sentence. We train this NN by estimating parameters θ that maximise the likelihood of sampling the observed classes in context.

To impose a preference on $f(\cdot;\theta)$ we will think of it as defining an approximation $q(\phi|\theta)$ to the posterior $\text{Dir}(\alpha\mathbf{1} + \mathbf{y})$ distribution over our Bayesian model, where we define $q(\phi|\theta)$ implicitly as the following stochastic procedure

$$X \sim \mathcal{D} \qquad\qquad \phi = f(x;\theta) \,, \tag{26}$$

where $\mathcal{D}$ is a dataset of observations.

We can satisfy both requirements, that is, assigning high likelihood to observations (MLE-training) and following a specific posterior distribution (preference over hypotheses), by framing this as a constrained optimisation problem:

$$\max_{\theta} \; \mathbb{E}_{(x,y)\sim\mathcal{D}}[\log \text{Cat}(y|f(x;\theta))] \tag{27}$$

$$\text{s.t. } \mathbb{E}_{x\sim\mathcal{D}}[f(x;\theta)] \sim \text{Dir}(\alpha + \mathbf{y}) \,, \tag{28}$$

which we can approach via Lagrangian relaxation (Boyd et al., 2004). This constraint in distribution can be expressed in different ways. For example, if $q(\theta)$ were prescribed we could approach MLE subject to

$$\text{KL}(q(\phi))\|\,\text{Dir}(\phi|\alpha\mathbf{1} + \mathbf{y})) = \epsilon \,, \tag{29}$$

where ϵ is a small positive slack (close to zero). Unfortunately, $q(\phi)$ is implicit, thus we cannot assess its value for a sample, we can only obtain samples from it. We can then attempt to satisfy the constraint by matching statistics under the two processes. That is, for some statistic t we constraint the classifier optimisation to satisfy

$$\mathbb{E}_{q(\theta)}[t(\phi)] = \mathbb{E}_{\mathrm{Dir}(\alpha+\mathbf{y})}[t(\phi)] \,, \tag{30}$$

where the left-hand side can be estimated via MC. For example, where $t(\phi) := \log \phi_i$, we have

$$\mathbb{E}_{q(\phi)}[t(\phi)] \overset{\mathrm{MC}}{\approx} \frac{1}{S} \sum_{s=1}^{S} t(\phi^{(s)}) \tag{31}$$

$$\mathbb{E}_{\mathrm{Dir}(\alpha+\mathbf{y})}[\log \phi_i] = \Psi(\alpha + y_i) - \Psi(v\alpha + N) \tag{32}$$

where Ψ is the digamma function and $N = \sum_{i=1}^{v} y_i$. In NMT, this boils down to collecting statistics from the softmax outputs of all steps in a batch, average them, and compare those to the same statistics collected from a sample from the Dirichlet posterior. There are many methods to match moments, this is only one of them and it's an almost naively simple one. Better methods can be found in the literature of two-sample tests. One of the best known tests, which is also very popular in machine learning, is the kernel-based minimum mean discrepancy (MMD; Gretton et al., 2012).

Experiment We use the GoURMET English-Amharic data for this investigation and concentrate on translation into English.[19] Our systems are trained on parallel data only (for this investigation we do not use synthetic data) using fairseq (Ott et al., 2019). We compare our formulation in terms of constrained optimisation formulation to LS using the simple mean/variance matching under the Dirichlet distribution sketched above as well as MMD. In particular, we are interested in obtaining improved performance under a probabilistic (sampling-based) decision rule in low-resource settings, we use minimum Bayes risk (Bickel and Doksum, 1977; Kumar and Byrne, 2004) with the estimator we propose in Section 4.1 (Eikema and Aziz, 2020, see section 7.6) and using sentence-level METEOR as utility function. We experiment with the following posterior constraints:

MS aims at $\mathbb{E}_{\mathrm{Dir}(\alpha+\mathbf{y})}[\log \phi_i]$ with count vector $\mathbf{y}$ derived from training data. This constraint aims at matching the logarithm of the softmax probabilities with the logarithm of the corresponding Dirichlet posterior on average.

MS-B is the same as MS but with count vector $\mathbf{y}$ derived from batch. Because the counts aggregate fewer observations this is more prone to overfitting. Intuititvely this should behave more like LS.

MM aims at matching $\mathbb{E}_{\mathrm{Dir}(\alpha+\mathbf{y})}[\phi_i]$, compared to MS this omits the logarithm, which should make the constraint smaller in magnitude and likely less influential.

MMD Compares softmax outputs in a batch to samples from $\mathbb{E}_{\mathrm{Dir}(\alpha+\mathbf{y})}[\log \phi_i]$ with $\mathbf{y}$ derived from training data in terms of MMD and aims at keeping the difference below an epsilon. This aims at matching all moments of the distribution and thus expresses the stronger preference.

[19] Similar, though yet preliminary, findings hold in the opposite direction, as well as on a different dataset of Nepali-English translations (Guzmán et al., 2019).

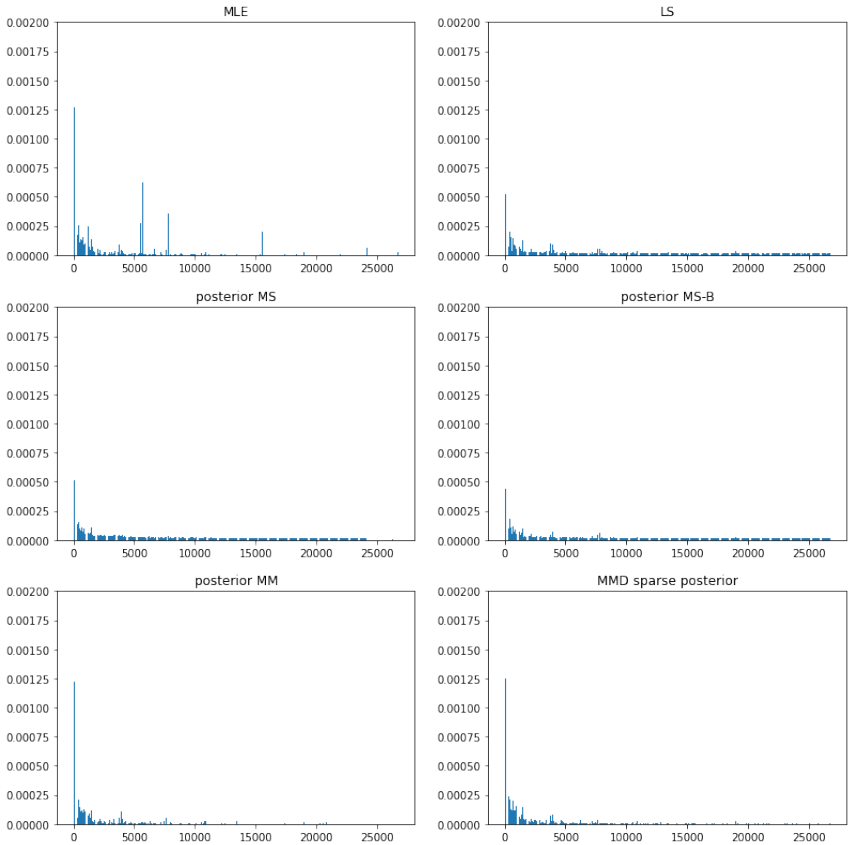

Figure 13: Distribution of softmax outputs in the validation set. Integers along the x-axis stand for entries in the vocabulary ordered by observed frequency in the training set. The y-axis corresponds to output probability on average. Note how LS (top-right) is flat outside the most frequent tokens. Posterior constraints based on $\mathbb{E}_{\mathrm{Dir}(\alpha+\mathbf{y})}[\log \phi_i]$ (middle row) are harsh, perhaps even harsher than LS. Targeting posterior moments (mean in bottom-left, MMD in bottom-right) leads to something close to MLE, but without surprising bumps for infrequent words.

		METEOR ↑		
Task	**Criterion**	**Beam**	**Sampling**	**MBR**
Amharic-English	MLE	11.6	11.0	12.4
	s.t. posterior MS	12.4	3.8	6.2
	s.t. posterior MS-B	12.8	5.1	7.2
	s.t. posterior MM	11.6	10.9	12.3
	s.t. posterior MMD	13.9	11.4	14.28
	LS ($\epsilon = 0.2$)	12.5	7.1	10.4

Table 7: Translation quality measured in METEOR ↑: we use METEOR because MBR needs a metric defined at the sentence level. Beam search uses beam of size 5; sampling is the performance measured on a single ancestral sample; MBR is estimated on 30 ancestral samples. Note that increasing beam size degrades performance.

Findings Table 7 summarises results. First of all, note that MLE with MBR already catches up with LS. This is a strong finding, which we have been observing for different language pairs and with varying amounts of resources. It is in fact part of the motivation for Task 4 altogether, for it shows that the model distribution contains valuable information, which in low-resource settings we very much need to explore. Second, note that is rather difficult to maintain sampling performance, measured on a single sample or via MBR, while constraining MLE. Some of our posterior constraints are too aggressive, like LS. Matching moments via MMD seems the most effective. We are in the process of analysing the translation outputs to determine whether they differ from MLE-training in a meaningful way. See Figure 13 for a visualisation of the effects of these different criteria in the distribution of probabilities assigned by the model to observations in the validation set. This is still ongoing work, but it already illustrates the potential for better regularisation via Bayesian priors.

Related Work Müller et al. (2019) put forward an explanation of the effects of label smoothing in terms of a preference in parameter space. Very recently, Meister et al. (2020) investigated label smoothing under the light of entropy regularisation. They too found that LS and variants seem to favour deterministic exploration of the model distribution. This line of work seems promising and we aim to continue developing it.

5 Conclusion

This section outlines what we have achieved in the first half of the project and how that contributes to achieving the goals we set ourselves in the proposal. 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 14. Finally, this section also outlines our priorities for the second and final half of the project: Table 8 summarises how we plan to allocate resources per task and lists directions that will most likely be explored.

Task	Task Leader	Resources	Potential Directions
T3.1	UVA	20%	Tighter integration with NMT, repurposing models for sentence-alignment (WP1) and discovery of sub-word units (WP2).
T3.2	UVA	40%	Complex structure in NMT via sparse marginalisation, inform variants of latent feature models with linguistic annotation available for English, latent structure in document-level NMT to overcome the need for large architecture blocks.
T3.3	UVA	40%	MBR decoding, utility metrics for MBR decoding, reducing support of distributions, Bayesian regularisation.

Table 8: Allocation of resources per task in the second half of the project.

T3.1 We proposed to induce alignments without explicit supervision and to investigate the viability of continuous relaxations to discrete alignments. Our ultimate goal was to investigate whether low-resource NMT could benefit from these word alignments. We have proposed a model that relaxes strong assumptions made by classical models, and showed that we can train this model reliably with little data. Though continuous relaxations are possible, and they have been show superior for certain problems, for the alignment problem we noticed that discrete variables perform better despite requiring a bit more careful gradient estimation. At this point we have not managed to claim improved translation performance out of an integration between NMT and our unsupervised alignments, but this is not yet a final word on the matter. A tighter integration, the kind pursued in task T3.2, might still reveal opportunities for improvements. Task T3.1 will therefore receive less of our time in the second half of the project, though we will pursue a collaboration with WP1 and WP2 to repurpose the model develop here to meet some of their interests (i.e., data collection and structure at sub-word level).

T3.2 We proposed to develop models that induce structured representations of sentences, examples of which are latent attributes, alignments, graphs, and trees. Our ultimate goal is to improve data efficiency and claim improved translation quality by endowing NMT models with stronger inductive biases. This task relies heavily on machine learning technology that is not readily in place, and because of that, a fair amount of resources went to developing effective machine learning tools for working with unobserved latent variables, especially of discrete nature. We have developed

technology to more reliably induce latent clusters, collection of attributes, and trees (and have left graphs aside for now) within neural models for text classification and language modelling. We have also integrated some of these techniques into NMT showing progress in terms of translation

Conference Papers	**Main Task**
Pelsmaeker and Aziz (2020)	3.2
Bastings et al. (2019)	3.2
Eikema and Aziz (2019)	3.2
Zheng et al. (2020)	3.2
Lopes et al. (2020)	3.2
Dobreva et al. (2020)	3.2
Pre-prints	**Main Task**
Correia et al. (2020) [*under review*]	3.2
Eikema and Aziz (2020) [*under review*]	3.3
Theses	**Main Task**
van Stigt (2019)	3.2
Murady (2020)	3.2
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.*

Figure 14: Summary of research output.

quality. The second half of the project will see more of these techniques be deeply integrated with NMT, in particular, those based on sparsification. Finally, we have identified an aspect of structured sentence models that holds great potential, and that was not sufficiently represented in the original proposal, namely, how a sentence relates to its external context in a document. We have embraced these document-level considerations contributing new models, datasets, as well as evaluation methodology, at this point we have emphasised challenges of document-level NMT more broadly and not necessarily concerning low-resource languages. We expect this to remain an important aspect of task T3.2, and for the next half of the project we will steer its focus towards the low-resource setting. To reduce the size of architectures, enabling these models in low-resource settings, we envision using AEVNMT to introduce weaker forms of independence assumptions.

T3.3 We had originally proposed multilingual extensions to the approaches developed in task T3.2, however, we revisited that decision. The original formulation creates an undesirable bottleneck, namely, the models of T3.2 and the developments of WP4 with regards to transfer learning are crucial dependencies. Moreover, and perhaps more crucially, this hinders the potential for novel contributions from T3.3. While studying discrete variants of AEVNMT in task T3.2, we realised that a lot of valuable information coded in the probability distributions that NMT outputs goes to waste at test time, when predictions are formed as if NMT models were margin-based classifiers, rather than probabilistic models, and that a lot of this potential is hindered by certain regularisation strategies originally developed for unstructured classifiers. We propose to exploit the implications of probabilistic training of NMT models, which we argue have been under-explored. Moreover, we aim to advance NTM's data efficiency by improved probabilistic regularisation. With resources dedicated in the last quarter, we have already shown a great deal of evidence, in particular, in low-resource settings, for improvements due to a better probabilistic account to NMT. In the second half, we expect to continue investing a reasonable amount of resources in this task. Directions of particular interest are: efficient approximations to minimum Bayes risk (MBR) decoding, effective utility metrics for MBR decoding, including those based on statistics gathered from the translation as a whole (rather than incrementally), reducing the amount of probability mass that goes to waste on clearly inadequate translations (such as overly short or overly long sentences), and investigation of Bayesian priors for better regularisation.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a Broken ELBO. 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 159–168, Stockholmsmässan, Stockholm Sweden, July 2018. PMLR. URL http://proceedings.mlr.press/v80/alemi18a.html.

Tamer Alkhouli and Hermann Ney. Biasing Attention-Based Recurrent Neural Networks Using External Alignment Information. In *Proceedings of the Second Conference on Machine Trans-*

lation, pages 108–117, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4711. URL https://www.aclweb.org/anthology/W17-4711.

Tamer Alkhouli, Gabriel Bretschner, Jan-Thorsten Peter, Mohammed Hethnawi, Andreas Guta, and Hermann Ney. Alignment-Based Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 54–65, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2206. URL https://www.aclweb.org/anthology/W16-2206.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. An Effective Approach to Unsupervised Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1019. URL https://www.aclweb.org/anthology/P19-1019.

Duygu Ataman, Wilker Aziz, and Alexandra Birch. A Latent Morphology Model for Open-Vocabulary Neural Machine Translation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxSI1SKDH.

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.

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. *Going beyond the sentence: Contextual Machine Translation of Dialogue*. PhD Thesis, LIMSI, CNRS, Université Paris-Sud, Université Paris-Saclay, Orsay, France, 2018.

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://www.aclweb.org/anthology/N18-1118.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics*. Holden-Day Inc., Oakland, CA, USA, 1977.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=944919.944937. Publisher: JMLR.org.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017. Publisher: Taylor & Francis.

Nicolas Boulanger-Lewandowski, Y. Bengio, and Pascal Vincent. High-dimensional sequence transduction. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2012. doi: 10.1109/ICASSP.2013.6638244.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002. URL https://www.aclweb.org/anthology/K16-1002.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311, 1993. URL https://www.aclweb.org/anthology/J93-2003.

Franck Burlot and François Yvon. Using Monolingual Data in Neural Machine Translation: a Systematic Study. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 144–155, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6315. URL https://www.aclweb.org/anthology/W18-6315.

Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997.

David Chiang. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270. Association for Computational Linguistics, 2005.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

Noam Chomsky. *Syntactic structures*. Mouton, 1964.

Gonçalo M. Correia, Vlad Niculae, Wilker Aziz, and André F. T. Martins. Efficient Marginalization of Discrete and Structured Latent Variables via Sparsity. 2020.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *RSSB*, 39:1–38, 1977.

Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of WMT, 2011*, pages 85–91, Edinburgh, Scotland, July 2011. URL http://www.aclweb.org/anthology/W11-2107.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis,

Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

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.

Greg Durrett and Dan Klein. Neural CRF Parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 302–312, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1030. URL https://www.aclweb.org/anthology/P15-1030.

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.

Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A. Smith. Transition-Based Dependency Parsing with Stack Long Short-Term Memory. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 334–343, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1033. URL https://www.aclweb.org/anthology/P15-1033.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent Neural Network Grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1024. URL https://www.aclweb.org/anthology/N16-1024.

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. *arXiv:2005.10283 [cs]*, May 2020. URL http://arxiv.org/abs/2005.10283. arXiv: 2005.10283.

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.

Giorgio Gallo, Giustino Longo, Stefano Pallottino, and Sang Nguyen. Directed hypergraphs and applications. *Discrete Applied Mathematics*, 42(2-3):177–201, April 1993.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. Jointly Learning to Align and Translate with Transformer Models. 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 4453–4462, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1453. URL https://www.aclweb.org/anthology/D19-1453.

Ulrich Germann. Aligned hansards of the 36th parliament of canada, 2001. URL https://www.isi.edu/natural-language/download/hansard/.

Zoubin Ghahramani and Thomas L. Griffiths. Infinite latent feature models and the Indian buffet process. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 475–482. MIT Press, 2006.

Joshua Goodman. Semiring Parsing. *Computational Linguistics*, 25(4):573–606, 1999. URL https://www.aclweb.org/anthology/J99-4004.

Will Grathwohl, Dami Choi, Yuhuai Wu, Geoff Roeder, and David Duvenaud. Backpropagation through the Void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SyzKd1bCW.

Alex Graves. Sequence Transduction with Recurrent Neural Networks. *ArXiv*, abs/1211.3711, 2012.

Miguel Graça, Yunsu Kim, Julian Schamper, Shahram Khadivi, and Hermann Ney. Generalizing Back-Translation in Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 45–52, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5205. URL https://www.aclweb.org/anthology/W19-5205.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. Star-Transformer. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1315–1325, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1133. URL https://www.aclweb.org/anthology/N19-1133.

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://www.aclweb.org/anthology/D19-1632.

Christian Hardmeier. Discourse in Statistical Machine Translation. a survey and a case study. *Discours*, 11, 2012.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997. Publisher: MIT Press.

Liang Huang and David Chiang. Better k-best Parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 53–64, Vancouver, British Columbia, October 2005. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W05-1506.

Dieuwke Hupkes, Sara Veldhoen, and Willem H. Zuidema. Visualisation and 'Diagnostic Classifiers' Reveal How Recurrent and Recursive Neural Networks Process Hierarchical Structure. *J. Artif. Intell. Res.*, 61:907–926, 2018. doi: 10.1613/jair.1.11196. URL https://doi.org/10.1613/jair.1.11196.

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

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, November 1999. ISSN 1573-0565. doi: 10.1023/A:1007665907178. URL https://doi.org/10.1023/A:1007665907178.

Yoon Kim, Alexander Rush, Lei Yu, Adhiguna Kuncoro, Chris Dyer, and Gábor Melis. Unsupervised Recurrent Neural Network Grammars. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1105–1117, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1114. URL https://www.aclweb.org/anthology/N19-1114.

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.

Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised Learning with Deep Generative Models. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3581–3589. Curran Associates, Inc., 2014.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved Variational Inference with Inverse Autoregressive Flow. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4743–4751. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6581-improved-variational-inference-with-inverse-autoregressive-flow.pdf.

Dan Klein. *The Unsupervised learning of Natural Language Structure*. PhD Thesis, Stanford, March 2005.

Dan Klein and Christopher D. Manning. Parsing and Hypergraphs. In *Seventh International Workshop on Parsing Technologies (IWPT- 2001)*, October 2001.

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://www.aclweb.org/anthology/W17-3204.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 48–54. Association for Computational Linguistics, 2003.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models*. MIT Press, 2009.

Shaohui Kuang, Deyi Xiong, Weihua Luo, and Guodong Zhou. Modeling Coherence for Neural Machine Translation with Dynamic and Topic Caches. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 596–606, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1050.

Aviral Kumar and Sunita Sarawagi. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*, 2019.

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 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N04-1022.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. In *Proceedings of ICLR*, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. Publisher: Ieee.

Mina Lee, Tatsunori B. Hashimoto, and Percy Liang. Learning Autocomplete Systems as a Communication Game. In *Proceedings of the 3rd NeurIPS Workshop on Emergent Communication*, 2019.

Joël Legrand, Michael Auli, and Ronan Collobert. Neural Network-based Word Alignment through Score Aggregation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 66–73, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2207. URL https://www.aclweb.org/anthology/W16-2207.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1011. URL https://www.aclweb.org/anthology/D16-1011.

Zhifei Li and Jason Eisner. First- and Second-Order Expectation Semirings with Applications to Minimum-Risk Training on Translation Forests. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 40–51, Singapore, August 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D09-1005.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Neural Machine Translation with Supervised Attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3093–3102, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL https://www.aclweb.org/anthology/C16-1291.

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.

Annie Louis and Bonnie Webber. Structured and Unstructured Cache Models for SMT Domain Adaptation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 155–163, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1017. URL https://www.aclweb.org/anthology/E14-1017.

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.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330, 1993. Publisher: MIT Press.

Andre Martins and Ramon Astudillo. From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. 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 1614–1623, New York, New York, USA, June 2016. PMLR. URL http://proceedings.mlr.press/v48/martins16.html.

Clara Meister, Elizabeth Salesky, and Ryan Cotterell. Generalized Entropy Regularization or: There's Nothing Special about Label Smoothing. *arXiv:2005.00820 [cs]*, May 2020. URL http://arxiv.org/abs/2005.00820. arXiv: 2005.00820 version: 2.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the State of the Art of Evaluation in Neural Language Models. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByJHuTgA-.

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.

Rada Mihalcea and Ted Pedersen. An Evaluation Exercise for Word Alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 1–10, 2003. URL https://www.aclweb.org/anthology/W03-0301.

Andriy Mnih and Karol Gregor. Neural Variational Inference and Learning in Belief Networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1791–II–1799. JMLR.org, 2014. event-place: Beijing, China.

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 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6322. URL https://www.aclweb.org/anthology/W18-6322.

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://www.aclweb.org/anthology/W18-6307.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4694–4703. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8717-when-does-label-smoothing-help.pdf.

Vlad Niculae and Mathieu Blondel. A Regularized Framework for Sparse and Structured Neural Attention. 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 3338–3348. Curran Associates, Inc., 2017.

Vlad Niculae, Andre Martins, Mathieu Blondel, and Claire Cardie. SparseMAP: Differentiable Sparse Structured Inference. 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 3799–3808, Stockholmsmässan, Stockholm Sweden, July 2018a. PMLR. URL http://proceedings.mlr.press/v80/niculae18a.html.

Vlad Niculae, André F. T. Martins, and Claire Cardie. Towards Dynamic Computation Graphs via Sparse Latent Structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 905–911, Brussels, Belgium, October 2018b. Association

for Computational Linguistics. doi: 10.18653/v1/D18-1108. URL https://www.aclweb.org/anthology/D18-1108.

Franz Josef Och and Hermann Ney. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong, October 2000. Association for Computational Linguistics. doi: 10.3115/1075218.1075274. URL https://www.aclweb.org/anthology/P00-1056.

Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, 2003.

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, July 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.

John Paisley, David M. Blei, and Michael I. Jordan. Variational Bayesian Inference with Stochastic Search. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pages 1363–1370, Madison, WI, USA, 2012. Omnipress. ISBN 978-1-4503-1285-1. event-place: Edinburgh, Scotland.

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.

Yookoon Park, Jaemin Cho, and Gunhee Kim. A Hierarchical Latent Structure for Variational Conversation Modeling. 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 1792–1801. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1162. URL http://aclweb.org/anthology/N18-1162. event-place: New Orleans, Louisiana.

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

Gabriel Pereyra, George Tucker, Jan Chorowski, Lukasz Kaiser, and Geoffrey Hinton. Regularizing Neural Networks by Penalizing Confident Output Distributions. In *ICLR Workshops*, February 2017. URL https://openreview.net/forum?id=HyhbYrGYe.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse Sequence-to-Sequence Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1146. URL https://www.aclweb.org/anthology/P19-1146.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, April 2014. PMLR. URL http://proceedings.mlr.press/v33/ranganath14.html.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence Level Training with Recurrent Neural Networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06732.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. Unsupervised neural machine translation with smt as posterior regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 241–248, 2019.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-Critical Sequence Training for Image Captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.131. URL https://doi.org/10.1109/CVPR.2017.131.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, June 2014. PMLR. URL http://proceedings.mlr.press/v32/rezende14.html. Issue: 2.

Miguel Rios, Wilker Aziz, and Khalil Sima'an. Deep Generative Model for Joint Alignment and Word Representation. 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 1011–1023, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1092. URL https://www.aclweb.org/anthology/N18-1092.

Reuven Y Rubinstein. The score function approach for sensitivity analysis of computer simulation models. *Mathematics and Computers in Simulation*, 28(5):351–379, 1986. Publisher: Elsevier.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. A Hybrid Convolutional Variational Autoencoder for Text Generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 627–637, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1066. URL https://www.aclweb.org/anthology/D17-1066.

Rico Sennrich and Martin Volk. Iterative, MT-based Sentence Alignment of Parallel Texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia, May 2011. Northern European Association for Language Technology (NEALT). URL https://www.aclweb.org/anthology/W11-4624.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1005. URL https://www.aclweb.org/anthology/N16-1005.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://www.aclweb.org/anthology/P16-1009.

Stuart M Shieber, Yves Schabes, and Fernando CN Pereira. Principles and implementation of deductive parsing. *The Journal of logic programming*, 24(1-2):3–36, 1995. Publisher: Elsevier.

Pavel Sountsov and Sunita Sarawagi. Length bias in Encoder Decoder Models and a Case for Global Conditioning. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1516–1525, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1158. URL https://www.aclweb.org/anthology/D16-1158.

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://www.aclweb.org/anthology/D19-1331.

Mitchell Stern, Jacob Andreas, and Dan Klein. A Minimal Span-Based Neural Constituency Parser. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 818–827, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1076. URL https://www.aclweb.org/anthology/P17-1076.

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.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. Hierarchical Modeling of Global Context for Document-Level Neural Machine Translation. 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 1576–1585, Hong Kong, China,

November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1168. URL https://www.aclweb.org/anthology/D19-1168.

Jörg Tiedemann and Yves Scherrer. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4811. URL https://www.aclweb.org/anthology/W17-4811.

Jakub M Tomczak and Max Welling. VAE with a VampPrior. In *AISTATS*, 2018. URL https://arxiv.org/pdf/1705.07120.pdf.

Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. Lattice Minimum Bayes-Risk Decoding for Statistical Machine Translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 620–629, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D08-1065.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. Learning to Remember Translation History with a Continuous Cache. *Transactions of the Association for Computational Linguistics*, 6:407–420, 2018. doi: 10.1162/tacl_a_00029. URL https://www.aclweb.org/anthology/Q18-1029.

George Tucker, Andriy Mnih, Chris J Maddison, John Lawson, and Jascha Sohl-Dickstein. REBAR: Low-variance, unbiased gradient estimates for discrete latent variable models. 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 2627–2636. Curran Associates, Inc., 2017.

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

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 590–596, Borovets, Bulgaria, 2005. URL https://www.kornai.com/Papers/ranlp05parallel.pdf.

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.

Alex Wang and Kyunghyun Cho. BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model. In *Proceedings of the Workshop on Methods for Optimizing and*

Evaluating Neural Language Generation, pages 30–36, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2304. URL https://www.aclweb.org/anthology/W19-2304.

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.

Weiyue Wang, Derui Zhu, Tamer Alkhouli, Zixuan Gan, and Hermann Ney. Neural Hidden Markov Model for Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 377–382, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2060. URL https://www.aclweb.org/anthology/P18-2060.

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.

Thomas Zenkel, Joern Wuebker, and John DeNero. End-to-End Neural Word Alignment Outperforms GIZA++. *arXiv preprint arXiv:2004.14675*, 2020.

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.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. Improving the Transformer Translation Model with Document-Level Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1049. URL https://www.aclweb.org/anthology/D18-1049.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1061. URL https://www.aclweb.org/anthology/P17-1061.

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.

Zachary Ziegler and Alexander Rush. Latent Normalizing Flows for Discrete Sequences. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7673–7682, Long Beach, California, USA, June 2019. PMLR. URL http://proceedings.mlr.press/v97/ziegler19a.html.

ENDPAGE

GoURMET

H2020-ICT-2018-2 825299

D3.1 Initial Progress Report on Learning Structural Models