Multi-way VNMT for UGC: Improving Robustness and Capacity via Mixture Density Networks

Anonymous ACL submission

Abstract

This work presents a novel Variational Neural Machine Translation (VNMT) architecture with enhanced robustness properties, which we investigate through a detailed case-study addressing noisy French user-generated content (UGC) translation to English. We show that the proposed model, with results comparable or superior to state-of-the-art VNMT, improves performance over UGC translation in a zero-shot evaluation scenario while keeping optimal translation scores on in-domain test sets. We elaborate on such results by visualizing and explaining how neural learning representations behave when processing UGC noise. In addition, we show that VNMT enforces robustness to the learned embeddings, which can be later used for robust transfer learning approaches.

1 Introduction

The specificities of UGC (Foster, 2010; Seddah et al., 2012) promote a plethora of vocabulary and grammar variations, which account for the large increase of out-of-vocabulary tokens (OOVs) in UGC corpora with respect to canonical parallel training data and raises many challenges for MT. In particular, UGC productivity (Martínez Alonso et al., 2016) limits the pertinence of 'standard' domain adaptation methods such as fine-tuning, as there will always be new forms that will not have been seen during training and urges the development of robust machine translation models able to cope with out-of-distribution (OOD) texts.

An increasing number of works on Neural Machine Translation, explores the use of latent distributional representations, known as latent-variable (LV-NMT). Such methods were shown to provide higher performance based on their abilities to model unobserved phenomena, such as intrinsic underlying structural information and applied to several NLP tasks (Kim et al., 2018). In this work, we focus on Variational NMT (Zhang et al., 2016) which has been reported to have good performances and interesting adaptability properties compared to other LV-NMT models (Przystupa, 2020).

The goal of this work is to evaluate the performance of VNMT when translating OOD texts, specifically, French social-media noisy UGC. To address the issue of UGC productivity, we consider a highly challenging zero-shot scenario and assume that only canonical texts are available for training the system. We hypothesize and provide experimental evidence supporting that, by leveraging on VNMT, the models can build more robust representations (embeddings and latent vectors) that map OOD observations to more in-distribution instances, which can be thus more easily translated in a zero-shot evaluation setting as shown by our experiments.

Our contributions are fourfold:

- we introduce `VNMT-MDN`, a new extension of VNMT models that relies on Mixture Density Networks (MDN) (Bishop, 1994); each mixture component extract an independent latent space to represent the source sentence and can model a different UGC specificities;
- we study the performance, in a zero-shot UGC translation scenario, of VNMT, `VNMT-MDN-NF` and the recently proposed `VNMT-NF` (Setiawan et al., 2020). This study prompt us to add Normalizing Flows (Rezende and Mohamed, 2015) used in `VNMT-NF` in our model and to introduce a second, better model, `VNMT-MDN-NF`;
- we study the impact of jointly learning source-side reconstruction, which we theorize UGC translation could benefit from, to recover from OOD constructs during evaluation;
- by probing the learned latent representations, we show the importance of using several latent distributions to model UGC and provide insights on the reasons why VNMT outperforms

the baselines.

2 Background and related works

VNMT Variational bayesian methods (Kingma and Ba, 2015) are generative architectures capable, from a distributional perspective, of modeling underlying structures from data. Under supervised settings, such as sequence-to-sequence MT tasks, where $\boldsymbol{x}$ and $\boldsymbol{y}$ are respectively the source and target, VNMT architectures combine a variational posterior approximation mechanism, $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})$, and a neural decoder generative distribution, $p_{\boldsymbol{\theta}}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})$, which are jointly trained to model the output $\boldsymbol{y}$ by looking for the distribution's parameters $(\boldsymbol{\theta},\boldsymbol{\phi})$ that minimize the ELBO for every pair $(\boldsymbol{x},\boldsymbol{y})$ in each training minibatch, as proposed in Zhang et al. (2016):

$$\begin{aligned}&\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})}[\log p_\theta(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{z})]\\ -&D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y})\ ||\ p_\theta(\boldsymbol{z}|\boldsymbol{x}))\end{aligned} \tag{1}$$

In this framework, the latent sentence-level vector $\boldsymbol{z}$ models the implicit structure of data to produce the translation prediction, $\boldsymbol{y}$. More recently, Su et al. (2018) proposed token-level latent representations for the paramemeter vector $\boldsymbol{z}$.

Normalizing Flows One of the major caveats of variational methods is that choosing the prior $q(\boldsymbol{z})$ is a complicated process that requires some *a priori* knowledge of the task. Thus this choice is often eased by selecting a Normal distribution with $\mu = 0.0$ and $\sigma = 1.0$, but such assumption can be restrictive to learn more complex processes. Regarding this issue, Rezende and Mohamed (2015) proposed using Normalizing Flows (NF) (Tabak and Turner, 2013; Tabak and Vanden-Eijnden, 2010) for variational methods by employing a prior distribution that undergoes a series of invertible and smooth transformations $f : \mathbb{R}^d \to \mathbb{R}^d$ (called flows). Then, the random latent variables $\boldsymbol{z}$, associated to a prior distribution $q(\boldsymbol{z})$, are converted to the random variable $\boldsymbol{z}' = f(z)$:

$$q(\boldsymbol{z}') = q(\boldsymbol{z})\left|\det\frac{\partial f^{-1}}{\partial \boldsymbol{z}'}\right| = q(\boldsymbol{z})\left|\det\frac{\partial f}{\partial \boldsymbol{z}}\right|^{-1} \tag{2}$$

Finally, we can build an arbitrarily long K chain of f_k transformations to generate the final prior $\boldsymbol{z}_K$, from the initial random variables (previous $\boldsymbol{z}$, now called $\boldsymbol{z}_0$) with gaussian prior q_0 :

$$\begin{aligned}\boldsymbol{z}_K &= f_K \circ \ldots \circ f_2 \circ f_1(\boldsymbol{z}_0)\\ \ln q_K(\boldsymbol{z}_K) &= \ln q_0(\boldsymbol{z_0}) - \sum_{k=1}^{K}\ln\left|\det\frac{\partial f_k}{\partial \boldsymbol{z}_{k-1}}\right|\end{aligned} \tag{3}$$

This enables higher flexibility of the generative process $p(z|x)$ and, regarding the MT task, was recently showed to improve VNMT models in the order of +0.2 to +1.2 BLEU points on in-domain evaluation (Setiawan et al., 2020). However, their effects over noisy test set haven't not been studied yet. Hence, we adopt this technique to improve the latent code modeling in our variational encoder and evaluate in our noisy ugc scenario.

Mixture Density Networks Much related to variational approaches, MDN, conceived to model multi-modal bayesian models, are a mixture model of M-components variational generative distributions Thus, in MDN, the posterior distribution, is the result on a linear combination of the gaussian kernels:

$$p(\boldsymbol{z}|\boldsymbol{x}) = \sum_{m=1}^{M}\alpha_m(\boldsymbol{x})\cdot q_m(\boldsymbol{z}|\boldsymbol{x}) \tag{4}$$

where α_m are known as the mixing coefficients and are also jointly trained by applying the '*softmax*' function to the corresponding outputs of the network, across the $\boldsymbol{z}_j^\alpha$ random variables to each component m:

$$\alpha_m = \frac{\exp(\boldsymbol{z}_m^\alpha)}{\sum_{j=1}^{M}\exp(\boldsymbol{z}_j^\alpha)} \tag{5}$$

Gumbel-Softmax sampling Regarding the mixing coefficients computation, we also explore employing a categorical probability distribution, for which probabilities are calculated by the network, such as in Ha and Eck (2018). Contrary to them, our supervised end-to-end training requires backpropagating the error gradient through the variational network via reparametrized sampling, which poses optimization challenges because of the discrete random variables used as latent vector for categorical distributions. For this reason, we use the reparametrization of such a distribution via the Gumbel-softmax sampling (Jang et al., 2017), such that, the '*argMax*' function is approximated by using '*softmax*' and generate the relaxed one-hot encoded samples, which correspond to the mixing coefficients:

$$\alpha_m = \frac{\exp(\log(\pi_m) + g_m)/\tau)}{\sum_{j=1}^{M} \exp((\log(\pi_j) + g_j)/\tau)} \quad (6)$$

where $g_m...g_M$ are *i.i.d* samples from the Gumbel(0,1) distribution (Gumbel, 1954; Maddison et al., 2017), π_i the probability associated to the *m*-th MDN's gaussian components, jointly generated by neural networks along with the computations of the corresponding parameters (μ_m, σ_m) for $m...M$; and τ the temperature parameter, which controls variability of the sampling. When $\tau \to 0$, the sampling exhibits a perfectly one-hot encoded output, whereas, conversely, when $\tau \to \inf$, the distribution approaches to an uniform one across all the MDN's components.

Why VNMT for noisy UGC? Variational approaches for NMT have been reported to act as regularizers introducing the prior distribution noise and thus increasing robustness and reducing overfitting (Zhang et al., 2016; Kumar and Poole, 2020). On the other hand, McCarthy et al. (2020) reported higher performance on both low and high resource scenarios, compared to an standard Transformer, as well as improvements when training using noisy data, and notably, using source-side monolingual corpora via a variational reconstruction loss term.

Recently, transformer-based VNMT models have also proved helpful for OOD evaluation, by identifying texts that are out of the training data distribution (Xiao et al., 2020) and improved NMT performance under such evaluation conditions (Setiawan et al., 2020).

In this work we address noisy UGC translation in zero-shot OOD scenarios using VNMT in order to study whether its distributional-shift robustness holds for such texts.

3 Our approach: extending variational methods for robust MT

For this work, we have drawn inspiration from `SketchRNN` (Ha and Eck, 2018) and recurrent World Models (Ha and Schmidhuber, 2018), both featuring a variational encoder-decoder architecture for modeling the input sequences, while employing a recurrent MDN decoder to produce a continuous generative variational posterior. We have adapted to use Transformer layers as encoder and generator, while training the distribution in a end-to-end manner with our usual parallel corpora. To this end, we employ a reparametrized form of the multiple Gaussian priors for sampling (Kingma and Welling, 2014). In addition, we study two mixing coefficient computations, i.e. a vanilla non-latent version using 'softmax' (Equation 5) and a relaxed categorical variational method by the means of Gumbel-softmax sampling (Equation 6).

3.1 Model

`VNMT-MDN`'s architecture in Figure 1, features a variational encoder that trains a latent representation to be fed to the decoder, which in turn, conditions an MDN, that is sampled to obtain the model's output. Backpropagation of the gradients is performed in an encoder-decoder end-to-end training fashion. The models have been integrated to the `OpenNMT-py` (Klein et al., 2018) framework [1]. For all VNMT models, we use a KL annealing schedule as in Ha and Eck (2018). We use the posterior's mean for inference during evaluation.

3.2 Encoder

According to our `Transformer Base` baseline architecture from (Vaswani et al., 2017), the encoder is composed by a 6-layered `Transformer Base` encoder, which output is feed to a 128-dimensional variational network, that estimates the final latent hidden encoded vector.

In Figure 5, we show the Transformer and variational encoding latent state ($\boldsymbol{z}$) as being estimated ($p(\boldsymbol{z}|\boldsymbol{x})$) approximating the posterior's mean and standard deviation, both learned using the reparametrization trick.

In order to be comparable to the recently introduced `VNMT-NF` (Setiawan et al., 2020), we also report results for our VNMT model extending the encoder's variational network with 4-flows Normalizing Planar Flows (PF) (Rezende and Mohamed, 2015)[2]. Other autoregressive normalizing models, such as Sylvester Flows (van den Berg et al., 2018), are available and could prove interesting for higher capacity. However, we decided to only address PF since they are the most simple solution with comparable performance improvement as other more complex flow models (Setiawan et al., 2020).

Similarly to `VNMT-NF`, we mix the last Transformer layer output to the latent vectors using a

[1] Code will be released upon publication

[2] Using the implementation from `https://github.com/riannevdberg/sylvester-flows`

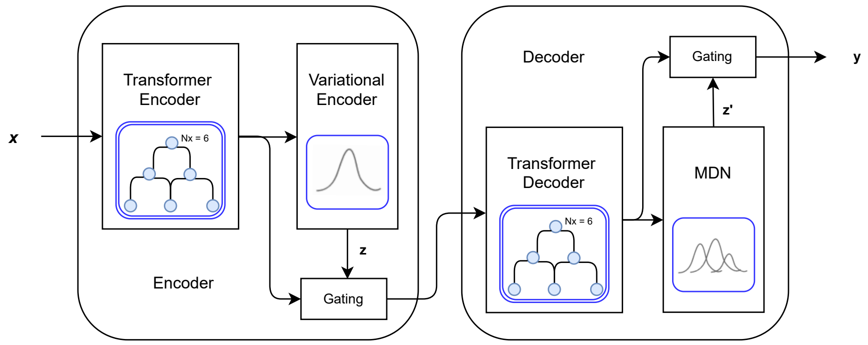


Figure 1: VNMT-MDN architecture overview.

gating mechanism, and a feedforward network in order to upscale the latent representation dimensionality and match the `Transformer Base` decoder number of dimensions (i.e, from 128 to 512); but unlike this model, we do so for both the encoder and decoder blocks' outputs since we introduce variational networks on both sides.

3.3 Decoder

The Transformer decoder's last layer output is passed to a 128-component MDN, with learnable parameters ϕ, encoding the mean and standard deviation of each one of these multivariate gaussian components; and π, which contains the probabilities of the categorical distribution that generates the mixing coefficient for each component. We train the MDN by variational inference using reparametrized sampling, similarly to our variational encoder network. As in our VNMT baseline, `VNMT-NF`, we dropped the contribution of the target, $\boldsymbol{y}$, to the posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{y})$, which has been reported to result in simpler systems with higher performance (Eikema and Aziz, 2019).

3.4 Monolingual reconstruction loss

We use the variational autoencoder (VAE) (Kingma and Welling, 2014; Rezende et al., 2014) as a tool to explore a semi-supervised approach, as done in Zhao et al. (2019), and performed experiments adding a source-side reconstruction loss term, according to Equation 7. This model is trained by sampling the approximated posterior distribution ($p_\theta(z|x)$) via variational inference, represented as the blue arrow in Figure 5 in the appendix.

$$\begin{aligned}\mathcal{L}_{mono} = \mathbb{E}_{z\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log(p_\theta(\boldsymbol{x}|\boldsymbol{z})] \\ -D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \mid\mid p_\theta(\boldsymbol{z}))\end{aligned} \tag{7}$$

Concerning our monolingual source-side data, we only use the source sentences contained in the training datasets to be able to unequivocally assess the advantages of this auxiliary task only, ruling out the impact of supplementary monolingual data, although the latter could arguably be the main interest of such training configuration.

4 Experiments

Datasets We trained our VNMT models, namely, `VNMT-NF`, `VNMT-MDN` and `VNMT-MDN-NF`, and our non-latent backbone architecture, `Transformer Base`, on two different Fr-En canonical parallel corpora: a combination of `WMT` training sets, and `OpenSubtitles`'18 (Lison et al., 2018). We used BPE tokenization (Sennrich et al., 2016) with 16K merge operations for all systems and, as stated in Section 3.4, we constrained our monolingual data, used in the semi-supervised training setup, to the French sources of the parallel training corpora.

We use the `PFSMB` (Rosales Núñez et al., 2019) and `MTNT` (Michel and Neubig, 2018) UGC test sets. In addition, as a complementary evaluation resource employed to probe our neural representations in Section 7, we use the `PMUMT` corpus (Rosales Núñez et al., 2021), which contains 400 annotated and normalized Fr-En UGC sentences. We have used the 400 original French noisy UGC samples and their corresponding fully normalized version, which we refer to `PMUMT Noisy` and `PMUMT Norm`, respectively. For detailed information on training and test corpora, please refer to Section A in the appendix.

Protocols All the MT results are reported using BLUE (Papineni et al., 2002), specifically, SacreBLEU (Post, 2018) using the '`intl`' tokenization, after detokenizing the systems' outputs. We have conducted an ablation test of our proposed `VNMT-MDN-NF` system and show the impact of our architecture's design over MT performance across our different test sets. We also performed experiments by incorporating a reconstruction objective function, according to Section 3.4, and denoted in Table 1 with the prefix `Mono-`. All experiments were done on a single training run and a beam 5 has been used for evaluation.

We have chosen, as initial experimental configuration, $\tau = 1.0$ for the Gumbel softmax sampling, which was selected mainly aiming to avoid artificial gradient scaling during backpropagation, directly caused by this coefficient (c.f. Equation 6), while being relatively larger than 0 in order to introduce variability to the sampling process.

Finally, we present a series of visualizations and metrics to characterize how `VNMT-MDN-NF` behaves when processing UGC during evaluation, in order to give further insights of its robustness capabilities compared to the `VNMT-NF` and `Transformer` baselines, by resorting to the learned latent neural representations' space.

5 Results

In this section we present the main MT results to study MT performance of our methods.

5.1 MT scores

In Table 1, we display the MT BLEU metric scores of our `VNMT-MDN-NF` systems compared to the baselines and to ablated versions of our proposed architecture. We can notice that our complete `VNMT-MDN-NF` system consistently outperforms both `VNMT-NF` and `Transformer` baselines on UGC test sets `PFSMB` and `MTNT`. We obtained mitigated results overall for canonical OOD tests: when training on `OpenSubtitles`, the `newstest'14`, both VNMT systems, `VNMT-MDN-NF` and `VNMT-NF` underperform the `Transformer` baseline; whereas we report a subtle improvement on the `OpenSubTest` canonical OOD in the `WMT` training setup. Regarding in-domain MT performance, we noticed a systematic improvement of VNMT, where our approach, `VNMT-MDN-NF`, seems to perform best, except for `newstest'14` when training on `WMT`, for which `VNMT-NF` achieves +0.1 BLEU more than the full `VNMT-MDN-NF` architecture. We also studied the impact of the latent vector dimensionality, by comparing 512 and 128, for which the former showed higher scores when translating the `MTNT` and `OpenSubTest` when training on `OpenSubtitles`, and unchanged performance for the `WMT` training data conditions.

In order to keep number of parameters and latent dimensions comparable across models, we have chosen `VNMT-MDN -128-128` as a backbone for the final architecture, `VNMT-MDN-NF`. Thus, we kept 128 dimensions as selected by Setiawan et al. (2020).

By looking at the results of the ablated versions of `VNMT-MDN-NF` (indicated in the table by an indentation with respect to the corresponding complete architecture), we can notice that, overall, we obtain the best BLEU scores across all test sets for the full `VNMT-MDN-NF` version. As interesting mixed results, we can highlight the cases for static latent representation (*z static*), where instead of sampling from the learned distributions, we retrieve their mean as output, and which showed slightly better BLEU scores when translating the `MTNT` and `newstest'14` test sets, with +1.2 and +0.1 BLEU improvement, respectively. This might be explained by a more stable training when using the mean of the distribution.

Finally, we can notice the overall highest performance of `VNMT-MDN-NF`, which employs our `VNMT-MDN` architecture and adds 4 normalizing PF, and that we compare to its corresponding `VNMT-NF` version, matching both in latent dimensions number, and number and type of flows. Regarding this BLEU comparison, it is interesting to note that using a categorical variational version of the mixing coefficients, proved to be generally a better design option than the default MDN 'softmax' way of determining such coefficients (π *non-latent*, in the table), only performing better for the `newstest'14` test set when training on the `OpenSubtitles` corpus. Following the same trend, the `WMT` training data configuration also showed improvements when using the Gumbel-softmax version, for which +0.8 and +0.3 BLEU score improvement were obtained for both the `PFSMB` and `MTNT` UGC test sets, respectively.

We have also obtained a consistent loss of performance compared to `Transformer Base` on the `OpenSubtitles` training configuration when translating the canonical OOD `newstest'14`, which could be explained by the considerable longer sentences of the latter compared to the training data (3.5 times on average, c.f. Table 4 in the appendix). These results suggest that the VNMT models used in this work could make bigger the dif-

	WMT				OpenSubtitles				
	PFSMB †	MTNT †	News ◇	OpenSubTest	PFSMB †	MTNT †	News	OpenSubTest ◇	# params.
Transformer Base	15.4	21.2	27.4	16.4	27.6	28.9	26.8	31.4	69M
VNMT-MDN -512-128	15.3	21.6	28.0	16.5	28.2	28.6	26.0	**31.5**	140M
VNMT-MDN -128-128	15.3	**21.6**	**28.0**	16.5	**28.3**	28.8	26.1	31.2	77M
z static	**16.5**	20.9	**28.0**	16.4	28.1	**29.3**	**26.2**	31.1	77M
-MDN	**16.5**	20.9	27.8	**16.6**	27.7	28.7	**26.2**	**31.2**	73M
VNMT-MDN-NF	**16.6**	21.3	27.8	16.5	**28.4**	**29.2**	26.4	**31.5**	77M
π non-latent	15.8	21.0	27.8	16.4	28.1	28.5	**26.6**	31.3	77M
-MDN	13.6	**21.5**	27.7	**16.6**	27.9	28.7	26.2	31.2	74M
Mono-VNMT-MDN-NF	15.8	**21.8**	28.0	16.5	**29.3**	28.7	26.2	**31.6**	77M
VNMT-NF (Setiawan et al., 2020)	15.5	21.4	**27.9**	16.4	28.0	29.0	26.4	31.4	73M

Table 1: BLEU test scores for our models and ablation variants. The † symbol indicates the UGC test sets, and ◇ in-domain test sets. *VNMT-MDN -512-128 stands for the model with a 512-dimensional latent space and 128 MDN components, VNMT-MDN -128-128 to its 128-dimensional latent space version.*

ficulty of translating sentences substantially longer than those of the training data.

As limitations for our models and experimental setup, we cannot generalize our findings for other language pair nor backbone MT architectures.

On the contrary, we achieved slightly better results for the same scenario, when training on WMT and evaluating OpenSubTest, where training sentences are 4 times longer that those of the test.

Posterior collapse Comparing VNMT-MDN-NF and its ablated version system removing the MDN module, both trained on OpenSubtitles and when evaluating the corresponding in-domain test set (OpenSubTest), we have calculated the average KL divergence of the variational decoder's MDN, which resulted in 0.21 and 0.15, respectively. Performing the same analysis for the WMT training and evaluation configuration, the KL divergence resulted in 0.38 for the full VNMT-MDN-NF and 0.33 for its version removing the MDN block. These results suggest that our proposed architecture is less prone to suffer from the posterior collapse phenomenon, and this could be explained by the use of several independent posterior distributions when including MDN in our model. This could also explain why, in Table 1, our systems employing MDN have an overall higher BLEU results than the aforementioned ablated system where we remove this component.

Semi-supervised monolingual joint training In Table 1 we report results with our proposed Mono-VNMT-MDN-NF system, by using source-side monolingual corpora reconstruction loss terms, as discussed in Section 3.4. Both WMT and OpenSubtitles training configurations shown an improvement of +0.2 and +0.1, respectively, when translating their corresponding in-domain test sets. However, for the canonical OOD tests, the latter lost performance on the newstest'14 (from 26.4 to 26.2 BLEU), aggravating this phenomenon reported previously; whereas the former benefited of a slight improvement on OpenSubTest. The results are rather inconsistent across the UGC test sets, which do not show a clear trend of the most performing choice across the two training datasets MT systems. Specifically, when adding the reconstruction loss term, WMT showed a gap of -1.2 and +0.5 BLEU, for PFSMB and MTNT, respectively, whereas OpenSubtitles's performance changed +0.9 and -0.5 BLEU correspondingly.

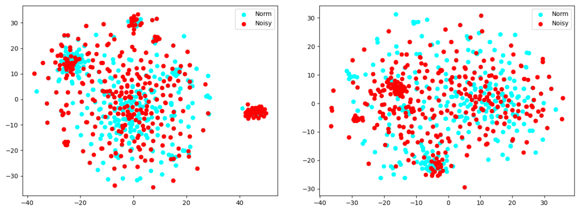
6 Qualitative analysis

In Table 5 in the appendix, we display some UGC translation examples. We notice a general trend of VNMT-MDN-NF (MTX in the table), outperforming the baselines and producing overall longer predictions when rare tokens or letter repetition are present in the input. Such are the cases for ①, with inconsistent-cased tokens, ② contains repeated characters and words, ③ with a out-of-vocabulary (OOV) character ('•'), and ④ presents User mentions and hashtags with the OOD characters '@' and '#'.

7 Learning representations analysis

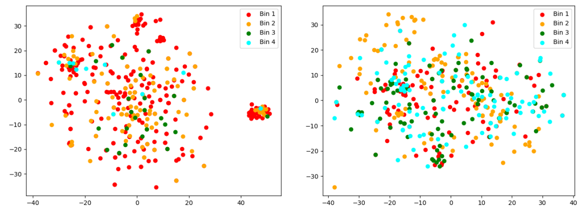
7.1 Latent space

Next, we present supplementary visualization and metrics to assess how VNMT builds more robust learning representations compared to the baseline. In this regard, McCarthy et al. (2020) showed that the learned variational embeddings are not able to separate UGC from canonical texts. This observation follows the reported ability of Deep Learning architectures to implicitly learn to clus-



(a) VNMT-NF (b) VNMT-MDN-NF

Figure 2: T-SNE representation of the latent space for noisy and normalized versions of PMUMT sentences at evaluation.



(a) VNMT-NF (b) VNMT-MDN-NF

Figure 3: T-SNE representation of the latent space for noisy PMUMT sentences during evaluation. Color portrays the bins of cosine similarity between noisy and normalized versions. *Bin 1* contains the samples with the least similarity value.

ter when training specific tasks (Carbonnelle and Vleeschouwer, 2021). We propose another approach to this problem: we report the cosine similarity histogram between FR noisy sentences and their normalized version, taking advantage of the PMUMT presented in Section 4. To obtain these embeddings, we fed its 400 original noisy UGC sentences and their corresponding 400 fully normalized versions to our VNMT baseline, VNMT-NF, and to VNMT-MDN-NF.

Overview The average cosine similarity between both corpus' versions favors VNMT-MDN-NF with 0.36 compared to VNMT-NF, with 0.26, suggesting that the former provides more robustness for the inner learning representations of UGC.

In Figure 2, we show the t-SNE (van der Maaten and Hinton, 2008) 2-dimensional visualization of both VNMT systems, showing the latent encoding of noisy and normalized PMUMT sentences. We can notice that the VNMT-NF latent representations present a series of outliers when noisy sentences abound, contrary to the VNMT-MDN-NF representations. In this set of 43 outlier observations (roughly 5% of the 800 plotted sentences' representation), 88% (37) are the original -noisy- UGC samples of PMUMT.

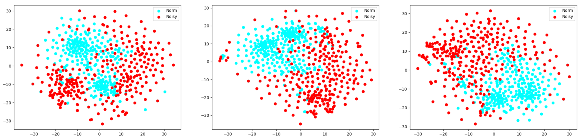
Latent space recovering from noise Since it seems hard to draw conclusions from translation performance distribution in the latent space, in Figure 3 we plot the same dimensional reduced latent space and we encode color for their bins of cosine similarity of the hereby shown noisy sentences to their corresponding normalized version. The bins for both plots were chosen using partitions' delimiters $[0.30, 0.44, 0.57]$. This was done to compare both latent spaces with the same similarity values' bins, however, VNMT-MDN-NF has overall higher metric quantiles ($[0.24, 0.36, 0.45]$) compared to VNMT-NF ($[0.19, 0.30, 0.40]$), which suggests that the VNMT-MDN-NF latent representations are more robust to UGC.

7.2 More robust embeddings for UGC

We also studied the source embeddings of our models to explore how VNMT can contribute to more robust embeddings that could prove valuable for transfer learning methods. We compare noisy and normalized versions of the FR PMUMT source side to assess whether they have a closer representation.

Noisy vs normalized data We now study the embeddings learned by VNMT-MDN-NF and assess how noise affects them compared to those of the baselines. We computed the cosine similarity between corresponding PMUMT noisy and normalized samples for the embedding space learned by Transformer Base, VNMT-NF and VNMT-MDN-NF, which resulted in 0.706, 0.744 and 0.750, respectively. This quantifies how VNMT can enforce learning more robust source representations since noisy UGC sentences are more related to their normalized version than for the baseline. We display the source embeddings for the three NMT systems in Figure 4 and we mark the noisy and normalized corpus's versions in red and blue, respectively. Each observation in the graph corresponds to the embedding of each sentence, computed by taking the average of the token-level embeddings. We can notice how both VNMT systems have a tendency to separate noisy and normalized sentences compared to Transformer Base, while having, higher cosine similarity.

Transfering learning representations As discussed above, in Figure 4 we noticed that VNMT seems to enforce noisy morphology modeling to the Transformer's embeddings in an implicit way. This motivated us to study whether the information in such learning representations can be used by the Transformer Base backbone model and



(a) `Transformer Base` (0.706) (b) `VNMT-NF` (0.744) (c) `VNMT-MDN-NF` (0.750)

Figure 4: T-SNE representation of the encoder embeddings for noisy and corresponding normalized `PMUMT` sentences during evaluation. *Average cosine similarity between corresponding noisy and normalized version of the `PMUMT` evaluation framework are reported between parentheses for each NMT system.*

	`PFSMB`†	`MTNT`†	`News`	`OpenSubTest`◇
`Transformer Base`	27.6	28.9	26.8	31.4
Pretrained init.	**29.0**	28.2	26.2	31.3
Frozen embs.	28.4	**28.9**	**26.8**	31.3
Fine-tuned	28.4	**28.9**	26.5	**31.4**

Table 2: Using VNMT-learned embeddings for transfer robust learned representations to the `Transformer Base` model.

benefit from improved robustness while removing the direct latent space contribution, and notably, with the same number of parameters and architecture as `Transformer Base`. Thus, in Table 2, we report scores for the `Transformer Base` model trained on `OpenSubtitles`, by either initializing the VNMT-pretrained embeddings and fine-tuning (FT) the system. We have performed FT using the same data configuration as in `OpenSubtitles` and continued training for 3 epochs from the `Transformer Base` model in Table 1 while replacing the embeddings by their VNMT-learned version's weights.

Results in Table 2 provide evidence that VNMT enforces more robust embeddings, which perform significantly better over the `PFSMB` UGC test set compared to the baseline, the system `Frozen embs.` giving the most consistent results over UGC. This system also results in keeping goof perfroamnce over the `newstest'14` canonical OOD test set, while taking advantage of an increased robustness to UGC. Such an improvement alleviates the loss of performance over `newstest'14` in our previous results, which was the only test set for which `VNMT-MDN-NF` underperformed the non-VNMT baseline in Table 1. These results indicate that VNMT promotes robustness to the NMT backbone and could be useful for conceiving more robust pretrained embeddings.

	`PFSMB`† (Blind)	`MTNT`† (Blind)	`4Square`
`Transformer Base`	19.7	25.0	21.9
+FT emb.	19.4	25.3	22.0
`VNMT-NF`	**20.0**	25.3	22.0
`VNMT-MDN-NF`	**20.0**	**26.4**	**22.5**

Table 3: Using VNMT FR source embeddings for transfer robust learned representations.

8 Blind test sets scores

We now evaluate our best performing model (`VNMT-MDN-NF` trained on `OpenSubtitles`) on the blind test sets described in Section 4, translating another set of tests to assess whether our approach proves useful for generalization over different types of UGC. We have also included the `4Square` corpus (Berard et al., 2019) to validate our VNMT system on other domain of UGC (restaurant reviews). We also display the results when using the `VNMT-NF` baseline and the `Transformer Base` model to assess improvement of our proposed architecture. We report such results in Table 3, where we can see that `VNMT-MDN-NF` consistently outperforms the baselines for our blind UGC test sets. It is interesting to notice that, although the in-domain performances for these 3 systems are very similar (between 31.4 and 31.5 BLEU in Table 1), the performance gap of blind UGC test sets is considerably larger, i.e. +0.8 BLEU in average to the non-latent baseline.

9 Discussion and perspectives

We introduced a novel VNMT architecture that provides improved performance and robustness over an state-of-the-art VNMT model, specifically when translating French UGC. An ablation study and blind test sets evaluation validate our architecture choice in regards of robustness capabilities for such texts. In addition, by exploring the learning representations trained by our VNMT model, and through conducting transfer learning experiments with such, we investigate the robustness brought to UGC, and show VNMT enforces such property to the backbone model, bringing a promising avenue for more robust pretrained neural learning representations. We report promising results when using an accessory source reconstruction loss to improve robustness, which we plan to study in the future by using other sorts of monolingual data and training protocols, such as denoising autoencoders.

References

Alexandre Berard, Ioan Calapodescu, Marc Dymetman, Claude Roux, Jean-Luc Meunier, and Vassilina Nikoulina. 2019. Machine translation of restaurant reviews: New corpus for domain adaptation and robustness. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019, Hong Kong, November 4, 2019*, pages 168–176. Association for Computational Linguistics.

Christopher M. Bishop. 1994. Mixture density networks. Technical report.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin M. Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 131–198.

Simon Carbonnelle and Christophe De Vleeschouwer. 2021. Intraclass clustering: an implicit learning ability that regularizes dnns. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Bryan Eikema and Wilker Aziz. 2019. Auto-encoding variational neural machine translation. In *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019, Florence, Italy, August 2, 2019*, pages 124–141. Association for Computational Linguistics.

Jennifer Foster. 2010. "cba to check the spelling": Investigating parser performance on discussion forum posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384, Los Angeles, California. Association for Computational Linguistics.

Emil Julius Gumbel. 1954. *Statistical theory of extreme values and some practical applications; a series of lectures*. Applied mathematics series ; 33. U.S. Govt. Print. Office, Washington.

David Ha and Douglas Eck. 2018. A neural representation of sketch drawings. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

David Ha and Jürgen Schmidhuber. 2018. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2455–2467.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Yoon Kim, Sam Wiseman, and Alexander M. Rush. 2018. A tutorial on deep latent variable models of natural language. *CoRR*, abs/1812.06834.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. Opennmt: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, AMTA 2018, Boston, MA, USA, March 17-21, 2018 - Volume 1: Research Papers*, pages 177–184.

Abhishek Kumar and Ben Poole. 2020. On implicit regularization in β-vaes. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 5480–5490. PMLR.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Héctor Martínez Alonso, Djamé Seddah, and Benoît Sagot. 2016. From noisy questions to Minecraft texts: Annotation challenges in extreme syntax scenario. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 13–23, Osaka, Japan. The COLING 2016 Organizing Committee.

Arya D. McCarthy, Xian Li, Jiatao Gu, and Ning Dong. 2020. Addressing posterior collapse with mutual information for improved variational neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8512–8525. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 543–553.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191.

Michael Przystupa. 2020. *Investigating the impact of normalizing flows on latent variable machine translation*. Ph.D. thesis, University of British Columbia.

Danilo Jimenez Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1530–1538. JMLR.org.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1278–1286. JMLR.org.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Comparison between NMT and PBSMT performance for translating noisy user-generated content. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 2–14, Turku, Finland. Linköping University Electronic Press.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2021. Understanding the impact of UGC specificities on translation quality. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 189–198, Online. Association for Computational Linguistics.

Djamé Seddah, Benoît Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The french social media bank: a treebank of noisy user generated content. In *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 2441–2458.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Hendra Setiawan, Matthias Sperber, Udhyakumar Nallasamy, and Matthias Paulik. 2020. Variational neural machine translation with normalizing flows. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7771–7777. Association for Computational Linguistics.

Jinsong Su, Shan Wu, Deyi Xiong, Yaojie Lu, Xianpei Han, and Biao Zhang. 2018. Variational recurrent neural machine translation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5488–5495. AAAI Press.

E. G. Tabak and Cristina V. Turner. 2013. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164.

Esteban G. Tabak and Eric Vanden-Eijnden. 2010. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217 – 233.

Rianne van den Berg, Leonard Hasenclever, Jakub M. Tomczak, and Max Welling. 2018. Sylvester normalizing flows for variational inference. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 393–402. AUAI Press.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. 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.

Tim Z. Xiao, Aidan N. Gomez, and Yarin Gal. 2020. Wat zei je? detecting out-of-distribution translations with variational transformers. *CoRR*, abs/2006.08344.

Biao Zhang, Deyi Xiong, Jinsong Su, Hong Duan, and Min Zhang. 2016. Variational neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* *EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 521–530. The Association for Computational Linguistics.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. 2019. Infovae: Balancing learning and inference in variational autoencoders. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 5885–5892. AAAI Press.

Appendix

A Data

Training Data Because of the lack of a large parallel data set of noisy sentences, we train our systems on 'standard' parallel data sets: `WMT` (Bojar et al., 2016) and `OpenSubtitles` (Lison et al., 2018). A subset of the latter has been randomly sampled to match the former in number of tokens in order to keep the training data quantity conditions comparable for both setups. `WMT` contains canonical texts (2.2M sent.) and `OpenSubtitles` (9.2M sent.) is made of informal dialogues found in popular sitcoms. While `OpenSubtitles` is, intuitively, closer to the kind of content generated by users online, it must be noted that UGC differs significantly from subtitles in many aspects: in UGC emotion are often denoted with repetitions, there are many typographical and spelling errors, and sentences may contain emojis that can even replace some words (e.g. ♥ can stands for the verb 'love' in sentences such as 'I ♥ you').

UGC Test Sets To evaluate the different NMT models, we consider two data sets of manually translated UGC: `MTNT` (Michel and Neubig, 2018) and the Parallel French Social Media Bank corpus (`PFSMB`) (Rosales Núñez et al., 2019)[3] which extends the French Social Media Bank (Seddah et al., 2012) with English translations. These two data sets raise many challenges for MT systems: they notably contain characters that have not been seen in the training data (e.g. emojis), rare character sequences (e.g. inconsistent casing or usernames) as well as many OOVs denoting URL, mentions, hashtags or more generally named entities (NE). Most of the time, OOVs are exactly the same in the source and target sentences.

We certify that we use all data collections in the way they are intended to, following their licence and in aggreement with our Institutional Review Board.

Detailed statistics on our used corpora can be found in Table 4.

B Training models

All systems are trained using a batch size of 4096 tokens, using the Adam optimizer (Kingma and Ba, 2015) and the `Noam` learning rate schedule (Vaswani et al., 2017). Training for, at

[3] https://gitlab.inria.fr/seddah/parallel-french-social-mediabank

Corpus	#sentences	#tokens	ASL	TTR	#chars	Corpus	#sents	#tokens	ASL	TTR	#chars
train set						*UGC test*					
`WMT`	2.2M	64.2M	29.7	0.20	335	`PFSMB`	777	13,680	17.60	0.32	116
`OpenSubtitles`	9.2M	57.7M	6.73	0.18	428	`MTNT`	1,022	20,169	19.70	0.34	122
test set						*UGC blind*					
`OpenSubTest`	11,000	66,148	6.01	0.23	111	`PFSMB`	777	12,808	16.48	0.37	119
`newstest'14`	3,003	68,155	22.70	0.23	111	`MTNT`	599	8,176	13.62	0.38	127

Table 4: Statistics on the French side of the corpora used in our experiments. *TTR stands for Type-to-Token Ratio, ASL for average sentence length.*

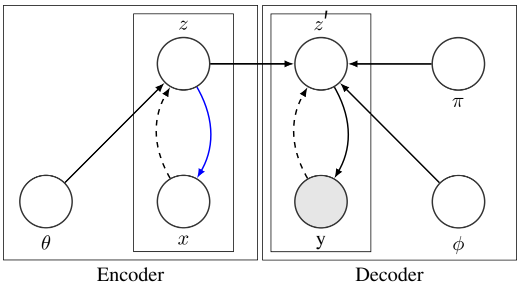


Figure 5: Directed graph of our encoder-decoder model variational inference. Dashed lines represent the variational approximation for the posterior distribution, and solid lines stand for the generative models. The blue arrow depicts the generative networks for source-side monolingual reconstruction distribution $p(\boldsymbol{x}|\boldsymbol{z})$.

most, 300K training iterations on a single Nvidia V100 took roughly 60 hours to converge for the `VNMT-MDN-NF` models.

C Effects of the Gumbel-softmax sampling temperature

In addition, we evaluate the impact of temperature τ, presented in Equation 6, in order to assess whether the sparsity of mixing coefficients favors a given type of texts among out test sets. As we noticed in Figure 7, a main MDN component is consistently the most important for the translations. This can be explained for the temperature we chose by default (1.0) and its desired effect of having less variance in the gradients that are propagated for each components' distribution. In this section we explore the impact of several temperature values to enforce more dense samples from the relaxed categorical mixture distribution. In Table 6, on one hand, we can see that smaller values to temperature τ seem to improve the model for canonical test sets, achieving the best performances for `OpenSubTest` and `newstest'14`, at the cost of negatively impacting performance over UGC test corpora. On the other hand, These results show that correctly setting the temperature parameter can be useful to translate different types of test sets and, for future work in this research track, temperature annealing schedules during training (Jang et al., 2017) or choosing a different value of temperature for evaluation phase, could be promising ideas to work with in order to develop more robust and all-purposed NMT systems.

D How do MDN's components react to UGC?

We now proceed to analyze and visualize how the MDN mixture coefficients react when translating our different test sets. In order to do so, in Figure 6 we report results for the canonical test sets, the normalized `PMUMT` corpus, and its noisy original UGC version. Each bar of the Wind Rose diagram represents one of the 128 independent trained distributions' mixture weights, which have been normalized and scaled across the four graphics, and where the 7th MDN component seems to be consistently the one that drives most of the decoding for the presented experiments. Furthermore, we can notice that most mixing coefficients are, for the most part, have around 50% probability of contributing to the final inference mixture, despite not enforcing this behavior with any specific method (e.g. dropout). On the other hand, the visualization suggests that both yellow (50-60%) and blue components (30-40% of activation) are variable across test sets, being very similar between `PMUMT Norm` and `OpenSubTest`, which could indicate that the mixture components are learning to encode different types of texts, potentially working as an implicit topic modeling module. Regarding the visualization when translating `PMUMT Noisy`, the main MDN component identified above, seems less important even when compared to the out-of-domain `newstest'14` set, which suggests that the MDN uses more dense representations when processing noisy texts.

In parallel, in Table 7 we display the covariance of these coefficients' distributions between the com-

①	src	OOOOOOOUUUUUUUIIIIIII ZLATAN IBRAHIMOVIC SIGNE UN DOUBLÉ À LA 90eme MINUTE ET ENVOI LE #PSG EN FINALE!!!!
	ref	YYYYYYYEEEEEEEESSSSSSSS ZLATAN IBRAHIMOVIC PERFORMS A DOUBLE AT THE 90th MINUTE AND SENDS THE #PSG TO FINAL!!!! 1-2!!!!
	TX	OOOOOOOOOOOOUUULL ZLATAN IBRAHIMOVYSIG A DOUBLED IN THE 90th MINUTE
	NF	OOOOOOOOOOOUUUIIIZLATANIBRAHIMOVIC SIGNS A DOUBL AT THE 90th MINUTE
	MTX	OOOOOOOOOOUUUUIII ZLATAN IBRAIMOV SIGNS A DOUBLED ON THE 90th MINUTE!
②	src	ne spoooooooooilez pas teen wolf non non non et non je dis non
	ref	don't spoooooooooil teen wolf no no no and no I say no
	TX	Don't spoil it.
	NF	No, no, no, no, no, no, no, no.
	MTX	Don't spoooooooil. Don't. Don't.
③	src	Roman Godfrey a regardé Teen Wolf (2011) • S03E17 Silverfinger et retourne dessiner des ronds sur son gitan préféré.
	ref	Roman Godfrey watched Teen Wolf (2011) • S03E17 Silverfinger and is back drawing circles on his favorite gypsy.
	TX	Roman Godfrey looked at Teen Wolf.
	NF	Roman Godfrey looked at Teen Wolf.
	MTX	Roman Godfrey watched Teen Wolf (2011), this is S03E17 Silverfinger and goes back to drawing a round about his favorite gypsy.
④	src	Vient de perdre une grosse heure a #flappybird cc @JohnDoe533 @JohnDoe534 @JohnDoe535
	ref	Just lost a big hour on #flappybird cc @JohnDoe533 @JohnDoe534 @JohnDoe535
	TX	#Flappybird cc #JohnDoe53 #JohnDoe53 #53 #1
	NF	#Flappybird c@JohnDoe5333)@JohnDoe53@JohnDoe53
	MTX	Just lost a huge hour at #flappybird cc at John Doe53 #John John Doe

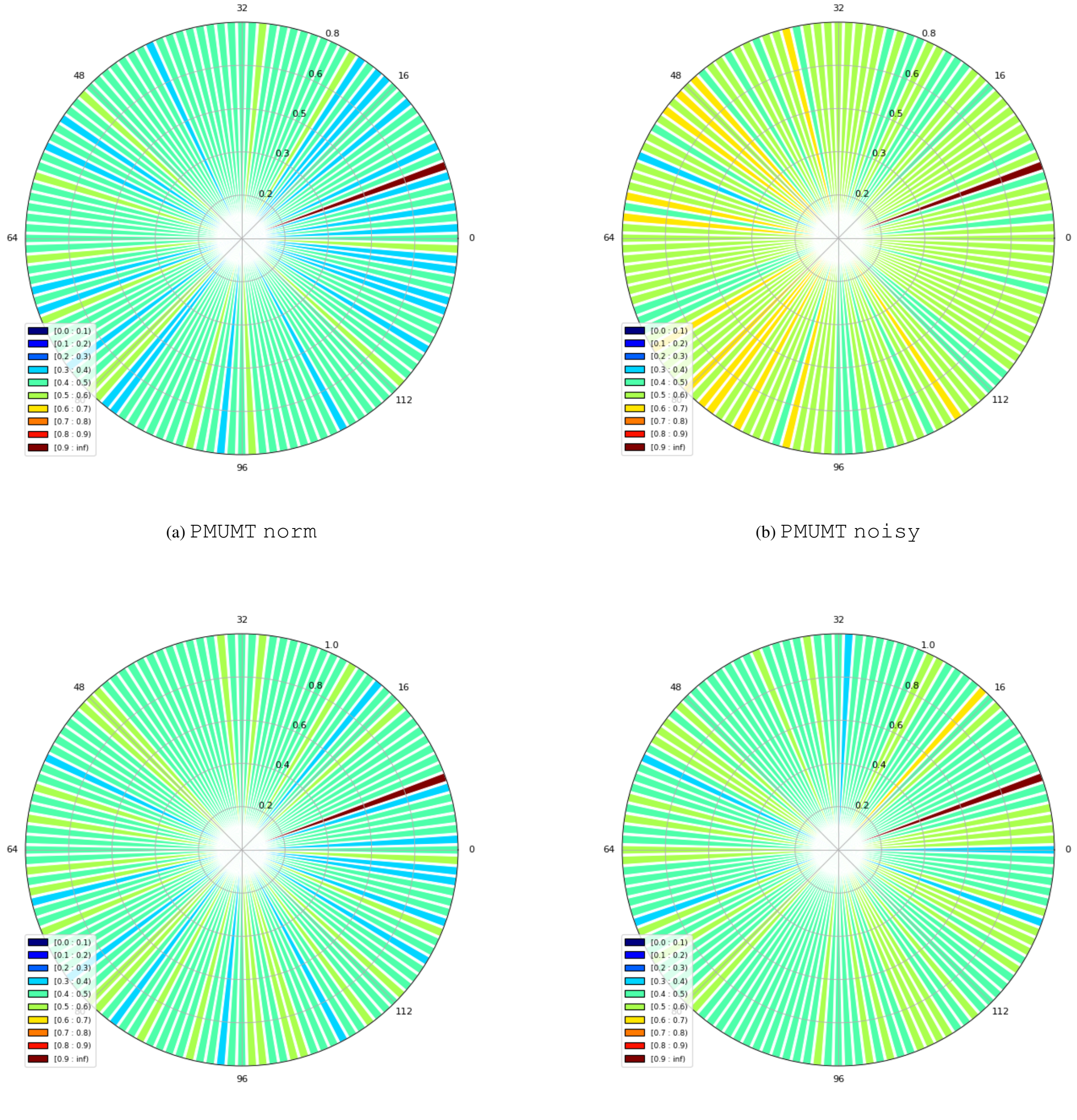
Table 5: b Examples from the PFSMB UGC corpus showing the Transformer, VNMT-NF and our model, VNMT-MDN-NF, predictions. *NF and MTX stand for the VNMT-NF (Setiawan et al., 2020) and VNMT-MDN-NF VNMT systems, respectively.*

			OpenSubtitles	
	PFSMB †	MTNT †	News	OpenSubTest ⋄
VNMT-MDN-NF (τ = 1.0)	**28.4**	**29.2**	26.4	31.5
(τ=0.2)	26.6	28.7	25.9	31.4
(τ=0.5)	28.0	28.7	**26.5**	**31.7**
(τ=2.0)	28.1	28.7	26.0	31.4
(τ=5.0)	27.3	28.3	26.0	31.4
(τ=10.0)	27.4	27.9	26.4	31.4

Table 6: Bleu test scores for our models and ablation variants. The † symbol indicates the UGC test sets, and ⋄ in-domain test sets.

binations of their values when translating different kind of texts, along with the standard deviation and sparsity to describe how the MDN's components behave.

Comparing the visualization in Figure 7, we can notice how the noisy UGC PMUMT and the out-of-domain newstest'14, diverge from the in-domain OpenSubTest and normalized UGC PMUMT corpus. This correlation is evidenced in the results in Table 7, where PMUMT noisy has the lowest score when compared to every other corpus, even if its normalized version seems to be the most correlated to the in-domain evaluation. Specifically, PMUMT Noisy is the least correlated to in-domain OpenSubTest and out-of-domain newstest'14 corpora, which points to the MDN reacting differently to content domain and UGC specificities in the noise; this observation is also supported by the associated figure. It is also interesting to notice that, according to the standard deviation and sparsity values, the active MDN components are more dense and variable for out-of-domain evaluation conditions, for the same Gumbel sampling temperature value.



(a) PMUMT norm (b) PMUMT noisy

(c) OpenSubTest (d) newstest'14

Figure 6: Average MDN mixture weights for test sets of different natures.

	PMUMT Noisy	News	OpenSubTest	std.	sparsity
PMUMT Norm	8.16	9.71	13.05	1.2e-3	0.387
PMUMT Noisy	—	7.72	7.86	1.0e-3	0.382
News	—	—	9.42	1.1e-3	0.384
OpenSubTest	—	—	—	1.1e-3	0.387

Table 7: Covariance between MDN mixture coefficients during inference for different types of test sets and sparsity for each set. *std.* stands for the standard deviation.