Target Conditioning for One-to-Many Generation

Marie-Anne Lachaux
Facebook AI Research
malachaux@fb.com

Armand Joulin
Facebook AI Research
ajoulin@fb.com

Guillaume Lample
Facebook AI Research
glample@fb.com

Abstract

Neural Machine Translation (NMT) models often lack diversity in their generated translations, even when paired with search algorithm, like beam search. A challenge is that the diversity in translations are caused by the variability in the target language, and cannot be inferred from the source sentence alone. In this paper, we propose to explicitly model this one-to-many mapping by conditioning the decoder of a NMT model on a latent variable that represents the domain of target sentences. The domain is a discrete variable generated by a target encoder that is jointly trained with the NMT model. The predicted domain of target sentences are given as input to the decoder during training. At inference, we can generate diverse translations by decoding with different domains. Unlike our strongest baseline (Shen et al., 2019), our method can scale to any number of domains without affecting the performance or the training time. We assess the quality and diversity of translations generated by our model with several metrics, on three different datasets.

1 Introduction

Neural Machine Translation (NMT) models are trained to translate a sentence from a source language into a target language. There are many translations of the same sentence that are both grammatically correct and faithful to the source, but these translations may differ greatly in their vocabulary, style or grammar. Inferring the best translation among them requires to explore a vast output space to cover this variability. This is typically handle as a post-processing step using a search algorithm, like beam search. This procedure is known to produce translations that lack in diversity, often differing only by a punctuation or a word (Kumar and Byrne, 2004; Li et al., 2016). While the search algorithm can certainly be improved, part of the problem resides also in the training of the NMT models; they are trained on 1-to-1 translation datasets without any objective to encourage diverse translations.

There are many ways to model the diversity of translations from data that contain only one translation, such as mixture of experts (Shen et al., 2019) or variational autoencoders (Zhang et al., 2016). A particularity of machine translation is that it is a one-to-many mapping problem. This means that the variability should be encoded by the target sentence and the question is how to combine a NMT system with a target sentence encoder with no posterior collapse.

In this work, we propose to combine the encoder of the NMT with a discrete target encoder. Similar to other discrete autoencoders (Kaiser et al., 2018; van den Oord et al., 2017), each target sentence is assigned to a discrete variable, or domain, and each domain is associated with an embedding. The embeddings from both encoders are then fed to the decoder of the NMT to form a translation. The discrete latent representation follows a categorical distribution that is constrained to be uniform over the dataset to avoid a mode collapse. Since each domain has its own embedding, changing the domain embedding changes the translation. At test time, we can thus condition the generation on each domain embedding to produce multiple translations with high diversity.

Our approach is general and can be applied on top of any model with little computational overhead. An advantage of our approach is that the number of domains can be arbitrarily large without affecting the performance or the running time. Our approach can replace or work with beam search during inference. We assess the quality and diversity of translations generated by our model with several metrics, on three different datasets.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2853–2862
November 16 - 20, 2020.

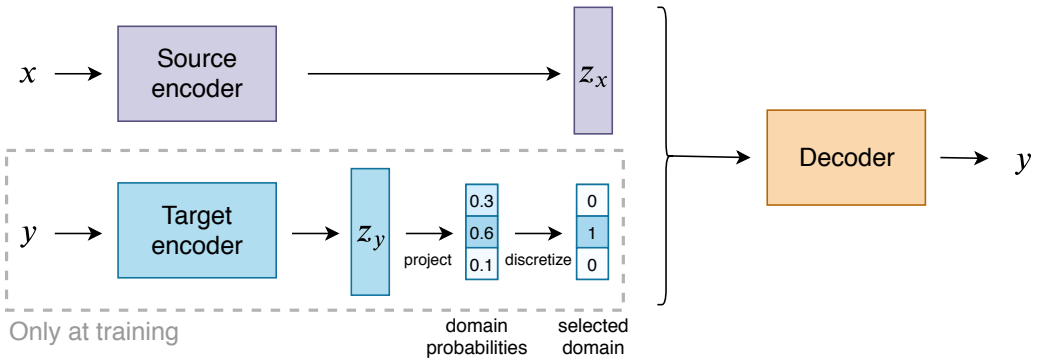

Figure 1: **Illustration of our model.** The model is composed of a source and a target encoder, and a decoder. At training time, a target sentence is encoded with the target transformer encoder to get a latent representation z_y. The latent representation is linearly mapped to a vector of size K on which apply a softmax to obtain domain probabilities. Each domain is associated with an embedding. The decoder is fed with both the source encoding, and the sum of the domain embeddings reweighted by their probabilities. During inference, we can generate K different hypotheses by switching the domain embedding that is fed to the decoder. To prevent a train-test discrepancy, during training we apply an argmax operator on domain probabilities, with probability p_{hard}.

2 Related Work

Several studies have proposed to sample diverse sequences by changing the value of a latent variable. For example, one possibility is to add noise to the latent space of a Variational Auto-Encoder (Kingma and Welling, 2013) to diversify samples in machine translation (Zhang et al., 2016), language modeling (Bowman et al., 2015) or question generation (Jain et al., 2017). In particular, Zhang et al. (2016) also condition the decoder of a NMT Model on a target encoder. As opposed to our work, the output of their encoder is continuous and sampling diverse generation requires to inject random noise, while we obtain diversity by switching between discrete domains. Similar noise injection mechanisms have been investigated to improve the diversity of responses in dialogue (Serban et al., 2017; Cao and Clark, 2017; Wen et al., 2017), and image captioning (Wang et al., 2017; Dai et al., 2017). Closer to our work, (Shen et al., 2019; Shu et al., 2019) and Xu et al. (2018) use domain embeddings to condition their generations. Unlike us, they do not condition the domain on the target, but select the domain which minimizes the reconstruction loss, which becomes expensive as the number of domains increases. Another relevant work is the fast decoder of Kaiser et al. (2018) where they also combine a discrete encoder applied on the target sentence with the NMT encoder. Their goal is to accelerate the decoding process of a machine translation system, while we are interested in efficiently sampling diverse translations.

Another line of work focuses on improving the generation by changing the decoding scheme during inference (Li et al., 2016; Gu et al., 2017) or by matching the training of the model to the decoding scheme (Wiseman and Rush, 2016; Collobert et al., 2019). This is done by either training through a beam search decoder (Wiseman and Rush, 2016; Collobert et al., 2019) or by reframing generation as a reinforcement learning problem (Bengio et al., 2015; Ranzato et al., 2015). These works focus on the decoding scheme to improve generation, but do not address the problem of diversifying the outputs generated from the same input.

3 Model

In this section, we describe our target encoder and how to train it along with a translation model. The target encoder learns to map target sentences to discrete domains, and we show how to use these domains to efficiently sample diverse translations.

3.1 Target encoding

A Neural Machine Translation (NMT) model is composed of a source encoder E_{src}, and a decoder D. Given a dataset $\mathcal{D}$ of pairs (x, y) of source sentences and their target translations, a standard encoder-decoder model is trained to minimize:

$$\mathbb{E}_{(x,y)\in\mathcal{D}}\Big(-\log p_D\big(y \mid E_{\text{src}}(x)\big)\Big)$$

where $p_D\big(y|E_{\text{src}}(x)\big)$ represents the probability given by the decoder D to a target sentence y to be the translation of a source sentence x. In our case, we consider that we also have a target encoder E_{tgt},

and we feed the decoder not only with an encoding of the source sentence, but also with an encoding of the target sentence. As a result, the model is trained to minimize:

$$\mathbb{E}_{(x,y)\in\mathcal{D}}\Big(-\log p_D\big(y \mid E_{\text{src}}(x); E_{\text{tgt}}(y)\big)\Big)$$

Without architectural constraint, the decoder D could trivially learn the identity mapping between the encoding of the target sentence $E_{\text{tgt}}(y)$ and the sentence to generate y. Instead, we propose to use a key-value structure for this embedding where the target encoder provides a probability for a key, or domain, and we feed the associated value to the decoder of the machine translation system. In practice, we constraint the output of the target encoder to represent the domain probability distribution of the target sentence. The output of the target encoder is thus a K-dimensional vector of probabilities $p = E_{\text{tgt}}(y)$. Since the output of the target encoder is not directly fed to the decoder, we bound the amount of information provided by the target encoder, preventing the model from learning a trivial mapping. At test time, we cannot estimate $E_{\text{tgt}}(y)$ since the target sentence y is not available. Instead, we feed the decoder D with any one-hot vector of $\mathbb{R}^K$ to generate K different translations. An illustration of our model is provided in Figure 1.

3.2 Implementation

Our NMT model is the transformer network of Vaswani et al. (2017) with a dimension d, with a transformer encoder E_{src} and a transformer decoder D. The target encoder E_{tgt} that we introduce in this paper is composed of a transformer encoder with the same architecture as the source encoder E_{src} and other components detailed bellow. We refer the reader to Vaswani et al. (2017) for the details of the architecture and describe below the specificites of our target encoder E_{tgt}.

The output $E_{\text{tgt}}(y)$ of the target encoder is a probability vector of size K. To obtain these probabilities, we encode the target through a transformer encoder. We take the first hidden state $h \in \mathbb{R}^d$ of the last layer of the target encoder, corresponding to the start token. We linearly map h to a score vector of dimension K. Finally, we apply a $\mathrm{softmax}$ operator to obtain a vector of domain probabilities:

$$p = E_{\text{tgt}}(y) = \mathrm{softmax}(Mh)$$

In that setting, the decoder is trained with arbitrary probability vectors, which becomes problematic at test time when p is set to a one-hot embedding on which the decoder may never have been trained. To prevent this train-test discrepancy, we apply a temperature on the domain scores s that decreases linearly from 1 to 0 over training. When the temperature reaches 0, we have $p = \mathcal{I}(\mathrm{argmax}(s))$[1] (i.e. the domain with the highest score has probability 1, the others have probability 0) and the target encoder remains frozen during the remaining training time.

Moreover, at each training step, we randomly replace the $\mathrm{softmax}$ by an argmax operator with a probability p_{hard}. In practice, we set $p_{\text{hard}} = 0.25$, which means that 75% of the time the target encoder is trained along with the source encoder and decoder, and 25% of the time the target encoder is only used to predict the domain with the highest probability. Overall, we have:

$$E_{\text{tgt}}(y) = \begin{cases} \mathcal{I}(\mathrm{argmax}(s)), & \text{if } 0 \le X \le p_{\text{hard}} \\ \mathrm{softmax}\left(\frac{s}{T}\right), & \text{otherwise} \end{cases}$$

where X is a random variable from a uniform distribution, i.e., $X \sim \mathcal{U}(0,1)$.

Optimization. When $T > 0$, the model is fully differentiable and the target encoder can be trained in an end-to-end fashion with the rest of the model. We found that it is also possible to use discrete operators like the Gumbel-Softmax (Jang et al., 2016). This way, $E_{\text{tgt}}(y)$ is always a one-hot vector and there is no train-test discrepancy. However, learning the target encoder through a discrete encoding makes optimization more difficult, and we obtained better results with a regular $\mathrm{softmax}$.

Domain input. To feed the target encoder output $E_{\text{tgt}}(y)$ as input to the decoder D, the decoder learns a matrix of embeddings $E = \big[e_0, \ldots, e_{K-1}\big] \in \mathbb{R}^{d\times K}$ where each e_i represents a different domain. Traditionally, the first input of a decoder is an embedding that corresponds to a start symbol $\langle S\rangle$. Instead, we feed as first embedding a vector e, where:

$$e = Ep = \sum_{i=0}^{K-1} p_i e_i, \quad \text{with} \quad p = E_{\text{tgt}}(y)$$

[1] By $\mathcal{I}(j) = (0, \ldots, 0, 1, 0, \ldots, 0)$, we denote the one-hot vector with 1 for j-th coordinate and 0 elsewhere).

The domain embeddings E are learned during training. This process is illustrated in Figure 2.

3.3 Training objective

We denote by θ the parameters of E_{src}, E_{tgt}, and D. Given a mini-batch of source and target sentences $\{(x_i, y_i)\}_{1 \leq i \leq N}$, the model is trained to minimize:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} -\log\Big(p_D\big(y_i | E_{\text{src}}(x_i); E_{\text{tgt}}(y_i)\big)\Big)$$

In practice, we want the decoder to properly leverage $E_{\text{tgt}}(y)$, i.e., the domain information coming from the target encoder. Without additional constraints, nothing prevents the model from collapsing to a mode where the target encoder constantly predicts the same domain, regardless of its input. The model is then perfectly predicting its domain, which means that it receives no gradient to escape this trivial solution.

To address this issue, we add a regularization term to the training objective, to encourage the model to make a uniform usage of available domains. In particular, we define the entropy distribution of selected domains in the mini-batch:

$$\mathcal{L}_{XE}(\theta) = -\widetilde{p}\log(\widetilde{p}), \quad \text{with} \quad \widetilde{p} = \frac{1}{N}\sum_{i=1}^{N} p_i$$

where $p_i = E_{\text{tgt}}(y_i)$ is the probability distribution of domains for the target sentence y_i. Finally, the model is trained to minimize $\mathcal{L}(\theta) - \lambda\mathcal{L}_{XE}(\theta)$, where λ is a hyper-parameter.

3.4 Inference

At inference, we generate one hypothesis per domain, i.e. K hypotheses. To generate the k^{th} hypothesis, we perform decoding by feeding e_k as embedding of the start symbol. We generate translations with greedy decoding, except in Figure 5, where we combine our model with beam search decoding which leads to a different quality vs. diversity trade-off.

4 Experiments

In this section, we describe an evaluation protocol similar to Shen et al. (2019), and compare our approach to several baselines on 3 MT datasets. Then, we show the importance of different components in our model in an ablation study.

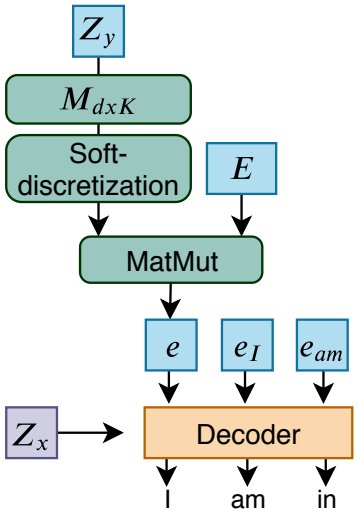

Figure 2: **Detailed illustration of our model.** Z_y is the first hidden state of the output of the target transformer encoder. To obtain $E_{\text{tgt}}(y)$, we linearly map Z_y to a K dimensional vector and perform a "soft-disctretization" by applying either a $\mathrm{softmax}$ or an argmax operator. We then compute the target domain vector e as the sum of the domain embeddings E reweighted by their probabilities contained in $E_{\text{tgt}}(y)$. The vector e is fed to the decoder as the embedding of the first token, along with the source encoding $Z_x = E_{\text{src}}(x)$.

4.1 Evaluation Metrics

To measure both the quality and diversity of our generations, we use an evaluation protocol similar to Shen et al. (2019). The test set has multiple human reference translations which allows to measure diversity. Formally, we denote by $\{s_i, [r_i^1, ..., r_i^P]\}_{1 \leq i \leq N}$ a multi-reference dataset, where each source sentence s_i is provided with P reference translations $[r_i^1, ..., r_i^P]$, and by $[h_i^1, ..., h_i^K]$ the K hypotheses generated by our model for the source sentence s_i.

We denote by $\text{BLEU}\big(\{h_i, [r_i^1, ..., r_i^P]\}_{1 \leq i \leq N}\big)$ the corpus-level BLEU score, with $P \geq 1$ references for each hypothesis. To measure the quality of our generations, we define:

$$\texttt{mBLEU} = \text{BLEU}\Big(\{h_i^j, [r_i^1, \ldots, r_i^P]\}_{1 \leq i \leq N, j \in K}\Big)$$

$\texttt{mBLEU}$ measures the quality of translations for each source sentence, and for each domain. A model that does not generate good translations for each domain will perform poorly. To measure the diversity of translations, we use the $\texttt{pairwise}$ metric of Shen et al. (2019), defined as:

$$\texttt{pairwise} = \text{BLEU}\Big(\{h_{i,j}, [h_{i,k}]\}_{\substack{1 \leq i \leq N \\ (j,k) \in K^2, j \neq k}}\Big)$$

`pairwise` computes the BLEU score between hypotheses of a same source sentence. A low `pairwise` ensures diversity in translations, while a `pairwise` of 100 means that for a given source sentence, the decoder will always generate the same translation. Overall, we want the model to have a low `pairwise` while preserving a high `mBLEU` score.

4.2 Dataset

We train and test our model on three different datasets, following Shen et al. (2019). Each dataset comes with a test set with multiple human reference translations.

WMT'17 English-German (En-De). We follow the same pre-processing protocol as Shen et al. (2019), where we filter all training sentences with more than 80 source or target words, which results in 4.5M sentence pairs. We apply the Moses tokenizer (Koehn et al., 2007) and learn a joint BPE vocabulary with 32k codes (Koehn et al., 2007). We take newstest2013 as a validation set, and test on a subset of 500 sentences of newstest2014 with 10 reference translations.

WMT'14 English-French (En-Fr). We follow the setup of Gehring et al. (2017), which results in 36M training sentence pairs. We use a joint vocabulary of 40k BPE codes. We use newstest2012 and newstest2013 as a validation set, and test on a subset of 500 sentences from newstest2014 with 10 reference translations.

WMT'17 Chinese-English (Zh-En). We follow the pre-processing setup of Hassan et al. (2018). The training set is composed of 20M sentence pairs, with 48k and 32k source and target BPE vocabularies respectively. We develop on devtest2017 and evaluate on a subset of 2000 sentences of newstest2017 that comes with 3 reference translations.

4.3 Experimental details

In all our experiments, we consider transformers with 6 layers, 8 attention heads, and we set the model dimension to $d = 512$. We optimize our model with the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9, \beta_2 = 0.98$ and a learning rate of 3×10^{-4}. We use the same learning rate schedule as Vaswani et al. (2017). We use a dropout (Srivastava et al., 2014) of 0.1 in the source encoder and the decoder. Following Shen et al. (2019), we do not use any dropout in the target encoder. With stochasticity in the target encoder, a same target sentence tends to be mapped to different domains at different iterations, which prevents the decoder from learning the specificity of each domain, and results in identical generations with no diversity.

We use 128 GPUs for the En-Fr experiments, and 16 GPUs for the En-De and Zh-En experiments. For the En-Fr experiments, we train with mini-batches of around 450k tokens, and 55k tokens for En-De and Zh-En. We use float16 operations to speed up training and to reduce the memory usage of our models. We implement our model within the fairseq framework of Ott et al. (2019).

4.4 Baselines

Sampling and Beam. We report results with a sampling and a beam baseline, as well as the diverse beam method (Vijayakumar et al., 2018). We consider a standard NMT system (i.e. an encoder-decoder model, without target encoder or latent variable). At test time, for sampling we sample K translations to generate K hypotheses. For the beam search, we use a beam size of K and return all hypotheses in the beam.

Mixture of Experts. We also compare against the state-of-the-art Mixture of Experts (MoE) model of Shen et al. (2019), with online responsibility update, uniform prior, shared parameters and hard assignment (*hMup* in their paper), which is their overall best setup. MoE model is composed of a source encoder E_{src} and a decoder D. Like our model, the decoder learns a matrix of embeddings $E = \left[e_0, \ldots, e_{K-1}\right] \in \mathbb{R}^{d \times K}$ where each e_i represents a different domain which is fed as first input of the decoder. Unlike us, they do not use a separate target encoder to select the domain, but consider an EM algorithm where the selected domain is the one that minimizes the reconstruction loss of the target sentence. In particular, for a mini-batch of N source and target sentences $\{(x_i, y_i)\}_{1 \leq i \leq N}$, the E-step computes:

$$d_i^* = \underset{d \in [1,K]}{\operatorname{argmax}}\, p_D\big(y_i | E_{\text{src}}(x_i); d\big)$$

Then, the M-step minimizes the negative log-probability of target sentences, given their source encodings, and the selected domains:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} -\log\Big(p_D\big(y_i | E_{\text{src}}(x_i); d_i^*\big)\Big)$$

	mBLEU			pairwise		
	En-De	En-Fr	Zh-En	En-De	En-Fr	Zh-En
Sampling	28.2	43.6	19.1	11.8	21.0	12.0
Beam	66.3	79.3	32.2	74.0	77.7	83.8
Diverse Beam (Vijayakumar et al., 2018)	60.0	72.5	31.6	53.7	64.9	66.5
MoE (Shen et al., 2019)	59.8	72.6	35.7	48.8	64.4	47.5
Our model	55.4	65.9	34.7	46.2	57.3	52.5

Table 1: Results on three WMT datasets: En-De, En-Fr, Zh-En. We use K = 10, 10, 3 domains respectively. We generate the same number of hypotheses as the number of references available in the multi-references datasets. Beam search is computed with beam size of K.

We run all of these baselines with the same transformer architecture as the one used in our model. For fair comparison, we use the same optimizer, learning rate and batch size in all experiments.

4.5 Main results

Table 1 present mBLEU and pairwise scores for different models, on the three considered datasets. We observe that a high mBLEU score is often combined with a high pairwise. For instance, the beam search and sampling baselines fail at generating both diverse and high quality translations. Beam search and diverse beam search hypotheses are accurate, but lack diversity, resulting in a very similar set of hypotheses. On WMT En-De, with $K = 10$, beam search gives a mBLEU score of 66.3 but a pairwise score of 74. On the other hand, the sampling baseline generates very diverse but inaccurate hypotheses, with a pairwise score of 11.8, but a mBLEU of 28.2.

The Mixture of Experts and Target Encoder models have a better trade-off between diversity and quality, as shown in Figure 4. Overall, our method provides more diversity than the MoE method, i.e. it obtains a lower pairwise score, but to the detriment of a lower mBLEU score. In Table 1, we observe that for En-De and En-Fr, our model obtains a lower mBLEU score than beam search decoding and the Mixture of Experts, but provides more diversity, with a pairwise score of 57.3 instead of 64.4 in En-Fr. While both methods perform similarly, our approach is simpler to implement, and can easily scale to an arbitrary number of domains, as shown in the following section.

4.6 Training speed

The training speed of our method is independent of the number of domains. In contrast, the training speed of the MoE model of Shen et al. (2019) decreases drastically when the number of domains increases. Indeed, the MoE model requires to perform K forward passes to determine the best domain. In Figure 3, we compare the training speed of both models for $K = 3$, 5, 10, 20, 50 and 200. Unlike the MoE model, using a target encoder allows generalization to an arbitrary number of domains.

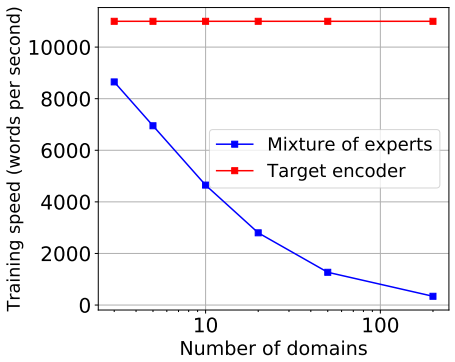

Figure 3: **Training speed.** Measured in number of words per second, for our target encoder model and the Mixture of Experts model of Shen et al. (2019), for different number of domains ($K = 3$, 5, 10, 20, 50, 200). The training speed of the target encoder model is constant while the Mixture of Experts model training speed decreases with the number of domains.

4.7 Ablation study

Beam search. In Figure 5, we study the impact of decoding with beam search instead of greedy decoding. Using beam search improves the quality of translations, but deteriorates the diversity. Combining a target encoder model with a beam search pushes towards the same trade-off of quality-diversity as the greedy MoE model.

Domain regularization. Without any regularization on the domain probabilities, i.e. when $\lambda = 0$,

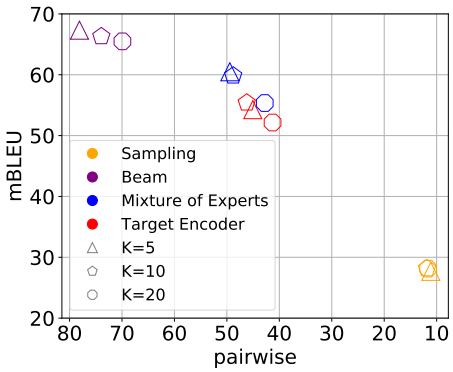

Figure 4: **Impact of the number of domains.** Results on the WMT'17 En-De dataset. We compare beam search, sampling, MoE (Shen et al., 2019) to our Target Encoder. In each case, we report results for $K = 5$, 10 and 20 domains. MoE and Target Encoder provide the best trade-off between quality and diversity. Compared to MoE, Target Encoder provides a lower `mBLEU` score, but also a lower `pairwise` (i.e., more diversity).

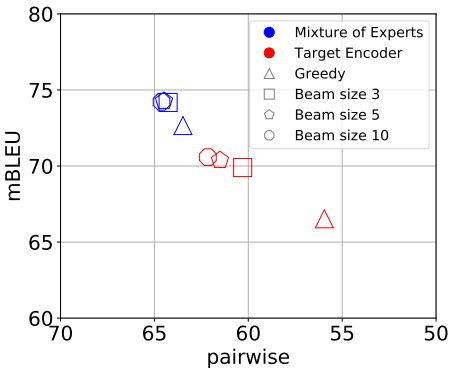

Figure 5: **Beaming search decoding.** Results on the WMT'14 En-Fr dataset for $K = 10$ domains. We study the impact of decoding greedily and beam search, for beam sizes of 3, 5 and 10. Beam search increases both the `mBLEU` and the `pairwise` scores, i.e., it provides higher quality translations, but with lower diversity.

we sometimes encounter the "collapse" scenario where at training time all target sentences are mapped to the same domain. As a result, only the embedding associated to that domain is trained, and at test time, every sentence generated from another (and untrained) domain embedding will be invalid. This means that only one of the K generated hypotheses will be valid, leading to a very poor `mBLEU`. Conversely, when λ is too high, the regularization term becomes predominant and the target encoder primarily focuses on maximizing the domain usage entropy, rather than on minimizing the decoder reconstruction loss. As a result, the target encoder uniformly maps target sentences to all available domains, but the domains do not contain any information about target sentences. This way, the decoder learns to ignore the domain, and will always output the same translation, independently of the input domain, which results in a `pairwise` score close to 100 (i.e. there is no diversity). In practice, we found that setting $\lambda = 0.1$ or $\lambda = 1$ leads to similar results, and is enough to prevent the collapse scenario.

Source versus target encoding. In this experiment, we change the input of our target encoder to probe where the source of diversity in our model comes from. In particular, it is possible that the diversity captured by our model is indirectly coming from the source sentences through the target sentences. We test this hypothesis by replacing the input of the target encoder by the source sentence. This model is identical to ours beside the change in the input of the target encoder. In that setting, on WMT'17 En-De, when using 10 domains, we obtain a `mBLEU` score of 66.5, and a `pairwise` BLEU of 97.2, which means that the model was not able to learn anything specific about each domain,

Source	参与投票的成员中,58%反对该合同交易。	自11月份开始，俄罗斯民意也有所扭转。
Human references	It was rejected by 58% of its members who voted in the ballot.	Russian public opinion has also turned since November.
	Of the members who voted, 58% opposed the contract transaction.	Russian public opinion has started to change since November.
	Of the members who participated in the vote, 58% opposed the contract.	The polls in Russian show a twist turn since the beginning of November.
Beam 3, Top 3	Of those voting, 58 per cent opposed the contract deal.	Since November, Russian public opinion has also turned around.
	Fifty-eight per cent of the members voting opposed the contract deal.	Since November, Russian public opinion has also changed.
	Fifty-eight per cent of the members voting opposed the contract.	Russian public opinion has also changed since November.
Mixture of Experts (Shen et al., 2019)	Of the members who voted, 58% opposed the deal.	Since November, the mood in Russia has also reversed.
	Fifty-eight per cent of the members who voted opposed the contract deal.	Since November, opinion in Russia has also reversed.
	Fifty-eight per cent of the voting members opposed the contract deal.	Opinion in Russia has also shifted since November.
Our Model	Of the members voting, 58 per cent opposed the contract deal.	Since November, Russian public opinion has also reversed.
	Fifty-eight per cent of the members who voted opposed the contract deal.	The mood in Russia has also reversed since November.
	Fifty-eight per cent of those voting had opposed this contract deal.	There has also been a reversal in Russian public opinion since November.

Table 2: Two examples of generations by our model and different baselines on the WMT'17 Zh-En dataset. Beam search generation lack diversity. The target encoder model gives the most diverse sets of translations.

Source	完成该项作业需把北部高架引桥南移700毫米。
Reference	This was done by pulling the northern approach viaduct 700 millimetres southwards.
Our model	Completion of the operation required the southern transfer of 700 mm from the northern elevated bridge.
	This operation will require moving the northern elevated bridge to a further 700 mm south.
	The operation was completed by moving the northern elevated bridge to the south by 700 mm.
Source	尽管该桥的规模和重量都很巨大，但是完成高架桥和桥面之间的合龙却需要精细作业，需要非常精准的装配。
Reference	Although the bridge is very large and heavy, joining the viaduct and roadway is a meticulous operation, requiring precise assembly.
Our model	Despite the scale and weight of the bridge, the completion of the Alloy bridge and the deck requires fine work and very precise assembly.
	Although it is huge in size and weight, it requires fine work and very precise assembly to complete the Alloy between the viaduct and the deck.
	The bridge, though of great scale and weight, requires fine work and a very precise assembly to complete the Alloy between the viaduct and the deck.
Source	据报道，雷杰普·塔伊普·埃尔多安则五次提到“好朋友”普京。
Reference	According to the report, Recep Tayyip Erdoğan mentioned his “good friend” Putin five times.
Our model	It was reported that Recep Tayyip Erdoğan made five references to “good friend” Putin.
	Recep Tayyip Erdoğan reportedly mentioned “good friend” Putin five times.
	The five-time mentions of “good friend” Putin were reported by Recep Tayyip Erdoğan.
Source	叙利亚是关键因素之一。莫斯科近日在叙利亚扮演和事佬的角色，而俄罗斯与土耳其却支持相反派别。
Reference	Syria is a key factor. Moscow has recently played the role of a mediator in Syria. However, Russia and Turkey supported opposing factions.
Our model	One of the key factors is Syria, where Moscow has played a conciliatory role, while Russia and Turkey have supported the opposite.
	Syria is one key factor, with Moscow playing a conciliatory role in Syria, while Russia and Turkey have supported the opposite.
	The role of Moscow as a peacemaker in Syria recently was one of the key factors, while Russia and Turkey supported the opposition.
Source	双方就如何在叙利亚问题上求同存异未达成明确共识。
Reference	There is no clear consensus on where they can seek common ground on Syria.
Our model	There is no clear consensus on how to seek common ground on Syria.
	No clear consensus has been reached on how to find common ground on Syria.
	The two sides did not reach a clear consensus on how to seek common ground on Syria.
Source	莫斯科当时的反应很激烈
Reference	Moscow’s reaction was intense.
Our model	Moscow was reacting violently.
	Moscow’s reaction was very strong.
	The reaction in Moscow was intense.
Source	令人欣慰的是，一切都进展得很顺利。
Reference	It is gratifying that everything is going well.
Our model	It is gratifying to note that everything is going well.
	Thankfully, everything has gone well.
	To the relief of all, everything was going well.

Table 3: Examples of generations by our model on the WMT’17 Zh-En dataset. Translations are sorted by domain index. The model generates high-quality translations with high diversity. Unlike beam search decoding, that tends to return similar hypotheses with only minor differences, our model generates various translations with very different prefixes, even for long sentences.

and the decoder simply ignores the domain information. The fact that learning the domain from the input sentence does not work well is expected, as this information is already encoded in the source encoding z_x. This validates that learning the diversity form the target domains is important. It also suggests that the diversity that our model learns is inherent to the target domain, and does not come from the source domain indirectly. Finally, both models have the same number of parameters, suggesting that the gain in performance is not only caused by the additional parameters.

4.8 Qualitative analysis

Table 3 provides examples of generations by our model on the WMT’17 Zh-En dataset. For each Chinese source sentence, we provide one English human translated reference, and translations by our model for three different domains. We observe that the model generates high-quality translations with high diversity. Unlike beam search decoding, that tends to return similar hypotheses with only minor differences in the suffix (Ott et al., 2018), our model is able to generate diverse translations with very different prefixes, even for long sentences.

5 Conclusion

In this paper, we presented an efficient way to sample diverse translations by adding a discrete target encoder to a NMT model. The discrete representation allows to change the domain of the translation and can be trained without supervision. The advantages of using a discrete encoder is that it is both general and scales with the number of domains with no additional computational time. In the future, we plan to test our discrete target encoder to diversify generations in other domains, such as language modeling, image captioning or image inpainting.

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In Advances in Neural Information Processing Systems, pages 1171–1179.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349.

Kris Cao and Stephen Clark. 2017. Latent variable dialogue models and their diversity. arXiv preprint arXiv:1702.05962.

Ronan Collobert, Awni Hannun, and Gabriel Synnaeve. 2019. A fully differentiable beam search decoder. arXiv preprint arXiv:1902.06022.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In Proceedings of the IEEE International Conference on Computer Vision, pages 2970–2979.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1243–1252. JMLR. org.

Jiatao Gu, Kyunghyun Cho, and Victor OK Li. 2017. Trainable greedy decoding for neural machine translation. arXiv preprint arXiv:1702.02429.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. arXiv preprint arXiv:1803.05567.

Unnat Jain, Ziyu Zhang, and Alexander G Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6485–6494.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144.

Łukasz Kaiser, Aurko Roy, Ashish Vaswani, Niki Parmar, Samy Bengio, Jakob Uszkoreit, and Noam Shazeer. 2018. Fast decoding in sequence models using discrete latent variables. arXiv preprint arXiv:1803.03382.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation.

Shankar Kumar and William Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. Technical report.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. A simple, fast diverse decoding algorithm for neural generation. arXiv preprint arXiv:1611.08562.

Aaron van den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. In Advances in Neural Information Processing Systems, pages 6306–6315.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. arXiv preprint arXiv:1803.00047.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In Proceedings of NAACL-HLT 2019: Demonstrations.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In Thirty-First AAAI Conference on Artificial Intelligence.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. arXiv preprint arXiv:1902.07816.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1823–1827.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In Thirty-Second AAAI Conference on Artificial Intelligence.

Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. 2017. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In Advances in Neural Information Processing Systems, pages 5756–5766.

Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 3732–3741. JMLR. org.

Sam Wiseman and Alexander M Rush. 2016. Sequence-to-sequence learning as beam-search optimization. arXiv preprint arXiv:1606.02960.

Qiongkai Xu, Juyan Zhang, Lizhen Qu, Lexing Xie, and Richard Nock. 2018. D-page: Diverse paraphrase generation. arXiv preprint arXiv:1808.04364.

Biao Zhang, Deyi Xiong, Hong Duan, Min Zhang, et al. 2016. Variational neural machine translation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 521–530.