Unsupervised Machine Translation Using Monolingual Corpora Only

ICLR 2018 Score: 8,6,8

Guillaume Lample, Ludovic Denoyer, Marc'Aurelio Ranzato
Facebook AI Research

Qingchun Bai

2018/1/10

Motivation

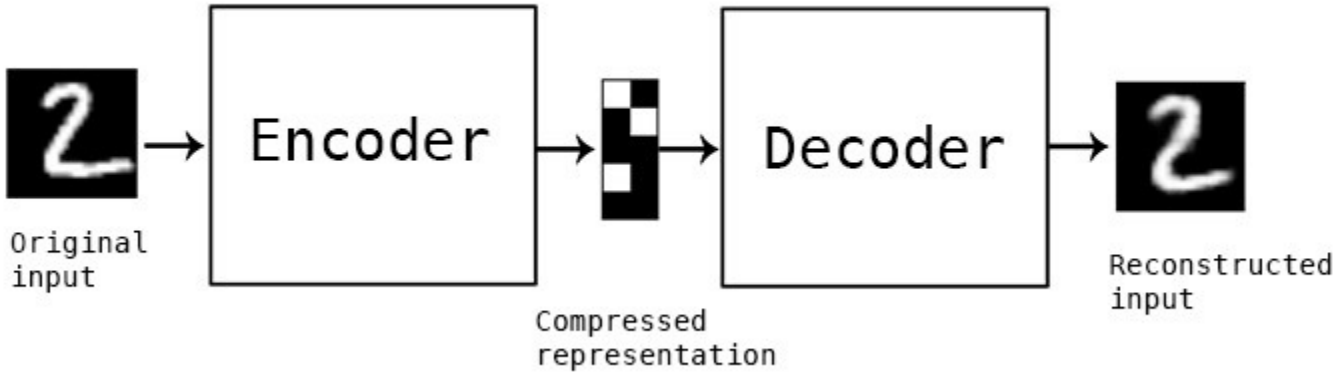
- Extreme and investigate whether it is possible to learn to translate even without any parallel data
- We propose a model that takes sentences from monolingual corpora in two different languages and maps them into the same latent space.

Outline

- *Background*
- *Auto-encoders*
- *Architecture*
- *Experiments*

Background

A Recap on Auto-Encoders



Model

Architecture

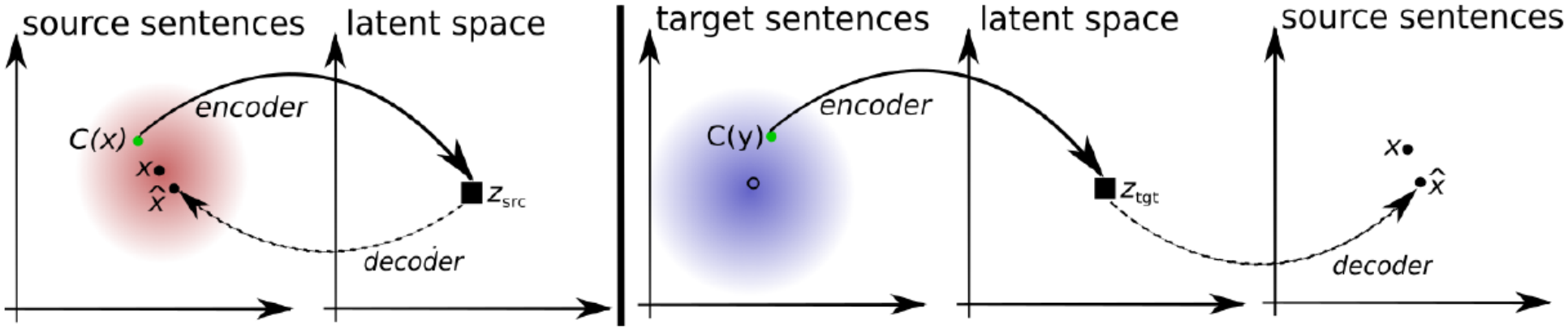


Figure 1: Toy illustration of the principles guiding the design of our objective function. Left (auto-encoding): the model is trained to reconstruct a sentence from a noisy version of it. x is the target, $C(x)$ is the noisy input, $\hat{x}$ is the reconstruction. Right (translation): the model is trained to translate a sentence in the other domain. The input is a noisy translation (in this case, from source-to-target) produced by the model itself, M, at the previous iteration (t), $y = M^{(t)}(x)$. The model is symmetric, and we repeat the same process in the other language. See text for more details.

Model

Architecture

$\mathcal{Z}^S = (z_1^s, ..., z_{|\mathcal{W}_S|}^s)$ $\mathcal{Z}^T = (z_1^t, ..., z_{|\mathcal{W}_T|}^t)$ Z being the set of all the embeddings

encoder : $e_{\theta_{enc}} z(\mathbf{x}, l)$

decoder : $d_{\theta_{dec}} z(\mathbf{z}, l)$

$\mathcal{D}_{src}$, source domain

$\mathcal{D}_{tgt}$ target domain

$l \in \{src, tgt\}$ sentence of m words $\boldsymbol{x} = (x_1, x_2, ..., x_m)$

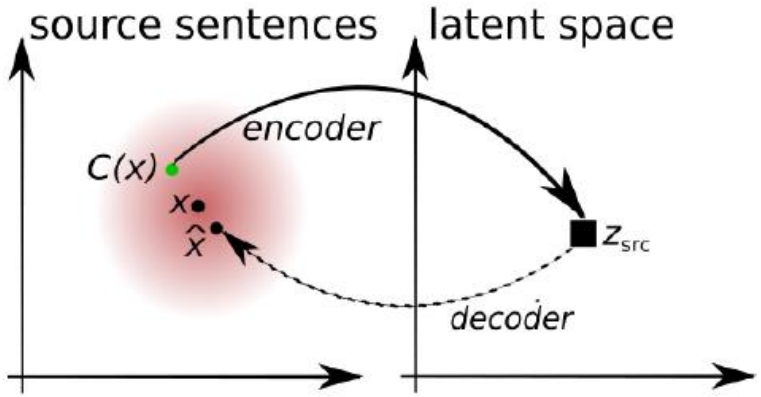
The encoder is a bidirectional-LSTM which returns a sequence of hidden states $\boldsymbol{z} = (z_1, z_2, ..., z_m)$

Model

Denoising auto-encoding

$$\mathcal{L}_{auto}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, \ell) = \mathbb{E}_{x \sim \mathcal{D}_\ell, \hat{x} \sim d(e(C(x),\ell),\ell)} \left[\Delta(\hat{x}, x)\right]$$

$\text{domain } \ell = src \text{ or } \ell = tgt$



Noise model

1. Drop every word in the input sentence with a probability p_{wd}.

2. $\forall i \in \{1, n\}, |\sigma(i) - i| \leq k$

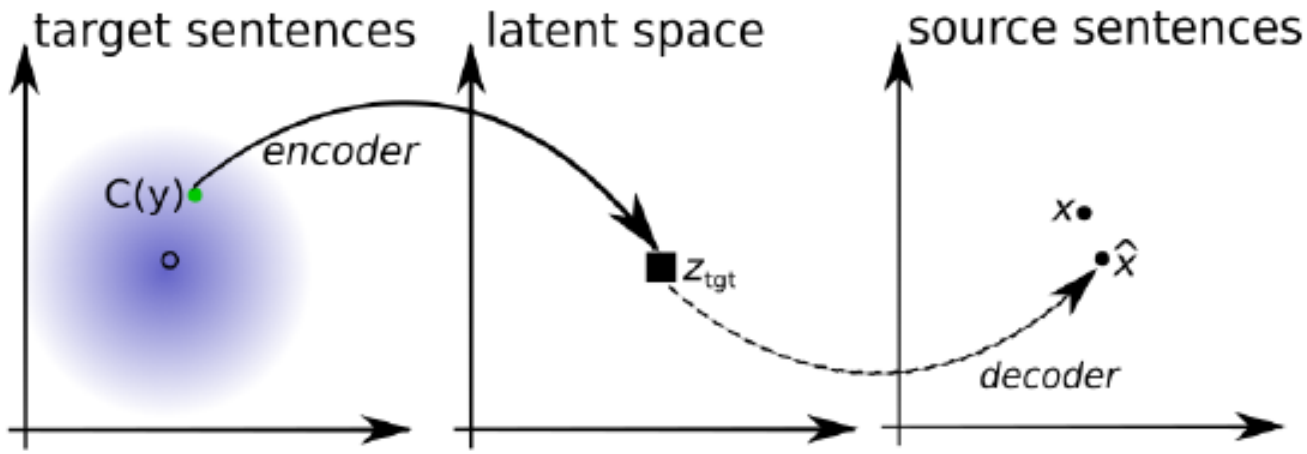
n is the length of the input sentence, and k is a tunable parameter

pwd = 0:1 and k = 3

Model

Cross domain training

back-translation



1 sample a sentence $x \in \mathcal{D}_{\ell_1}$

2 $y = M(x)$.

3 reconstruct x from $C(y)$

$$\mathcal{L}_{cd}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, \ell_1, \ell_2) = \mathbb{E}_{x \sim \mathcal{D}_{\ell_1}, \hat{x} \sim d(e(C(M(x)), \ell_2), \ell_1)} \left[\Delta(\hat{x}, x)\right]$$

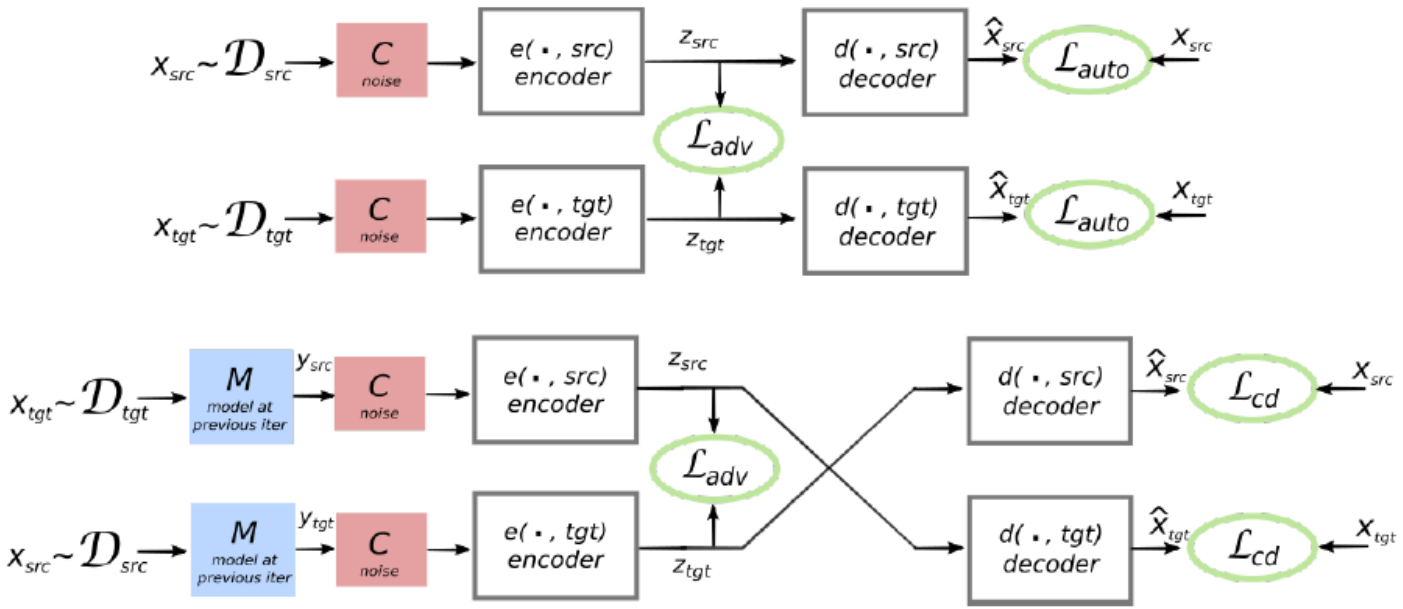
Δ is again the sum of token-level cross-entropy losses.

Model

Adversarial training

auto-encoding

translation



output

$(z_1, ..., z_m)$, with $z_i \in \mathbb{R}^n$

Figure 2: Illustration of the proposed architecture and training objectives. The architecture is a sequence to sequence model, with both encoder and decoder operating on two languages depending on an input language identifier that swaps lookup tables. Top (auto-encoding): the model learns to denoise sentences in each domain. Bottom (translation): like before, except that we encode from another language, using as input the translation produced by the model at the previous iteration (light blue box). The green ellipses indicate terms in the loss function.

$$p_D(l|z_1, ..., z_m) \propto \prod_{j=1}^{m} p_D(\ell|z_j), \text{ with } p_D : \mathbb{R}^n \rightarrow [0; 1].$$

$$\mathcal{L}_{adv}(\theta_{\text{enc}}, \mathcal{Z}|\theta_D) = -\mathbb{E}_{(x_i,\ell_i)}[\log p_D(\ell_j|e(x_i, \ell_i))]$$

Final Objective function

auto-encoding

$$\mathcal{L}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}) = \lambda_{auto}[\mathcal{L}_{auto}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, src) + \mathcal{L}_{auto}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, tgt)] +$$

cross domain training

$$\lambda_{cd}[\mathcal{L}_{cd}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, src, tgt) + \mathcal{L}_{cd}(\theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z}, tgt, src)] +$$

$$\lambda_{adv}\mathcal{L}_{adv}(\theta_{\text{enc}}, \mathcal{Z}|\theta_D)$$

discriminator

where λ_{auto}, λ_{cd}, and λ_{adv} are hyper-parameters weighting the importance of the auto-encoding, cross-domain and adversarial loss. In parallel, the discriminator loss $\mathcal{L}_D$ is minimized to update the discriminator.

Training

Algorithm 1 Unsupervised Training for Machine Translation

procedure TRAINING($\mathcal{D}_{src}, \mathcal{D}_{tgt}, T$)
  Infer bilingual dictionary using monolingual data (Conneau et al., 2017)
  $M^{(1)} \leftarrow$ unsupervised word-by-word translation model using the inferred dictionary
  for $t = 1, T$ **do**
    using $M^{(t)}$, translate each monolingual dataset
    // discriminator training & model training as in eq. 4
    $\theta_{\text{discr}} \leftarrow \arg\min \mathcal{L}_D, \quad \theta_{\text{enc}}, \theta_{\text{dec}}, \mathcal{Z} \leftarrow \arg\min \mathcal{L}$
    $M^{(t+1)} \leftarrow e^{(t)} \circ d^{(t)}$ // update MT model
  end for
  return $M^{(T+1)}$
end procedure

$$MS(e, d, \mathcal{D}_{src}, \mathcal{D}_{tgt}) = \frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}_{src}}\left[BLEU(x, M_{src\to tgt} \circ M_{tgt\to src}(x))\right] +$$

back-translation BLEU

$$\frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}_{tgt}}\left[BLEU(x, M_{tgt\to src} \circ M_{src\to tgt}(x))\right]$$

Dataset

English-French and English-German language pairs

	MMT1 en-fr	MMT1 de-en	WMT en-fr	WMT de-en
Monolingual sentences	14.5k	14.5k	15M	1.8M
Vocabulary size	10k / 11k	19k / 10k	67k / 78k	80k / 46k

Table 1: **Multi30k-Task1 and WMT datasets statistics.** To limit the vocabulary size in the WMT en-fr and WMT de-en datasets, we only considered words with more than 100 and 25 occurrences, respectively.

Test data : newstest2014

Baseline

- Word-by-word translation (WBW)
- Word reordering (WR)
- Oracle Word Reordering (OWR)
- Supervised Learning

Experiments

UNSUPERVISED DICTIONARY LEARNING

	Multi30k-Task1				WMT			
	en-fr	fr-en	de-en	en-de	en-fr	fr-en	de-en	en-de
Supervised	56.83	50.77	38.38	35.16	27.97	26.13	25.61	21.33
word-by-word	8.54	16.77	15.72	5.39	6.28	10.09	10.77	7.06
word reordering	-	-	-	-	6.68	11.69	10.84	6.70
oracle word reordering	11.62	24.88	18.27	6.79	10.12	20.64	19.42	11.57
Our model: 1st iteration	27.48	28.07	23.69	19.32	12.10	11.79	11.10	8.86
Our model: 2nd iteration	31.72	30.49	24.73	21.16	14.42	13.49	13.25	9.75
Our model: 3rd iteration	32.76	32.07	26.26	22.74	15.05	14.31	13.33	9.64

Table 2: **BLEU score on the Multi30k-Task1 and WMT datasets** using greedy decoding.

Experiments

EXPERIMENTAL DETAILS

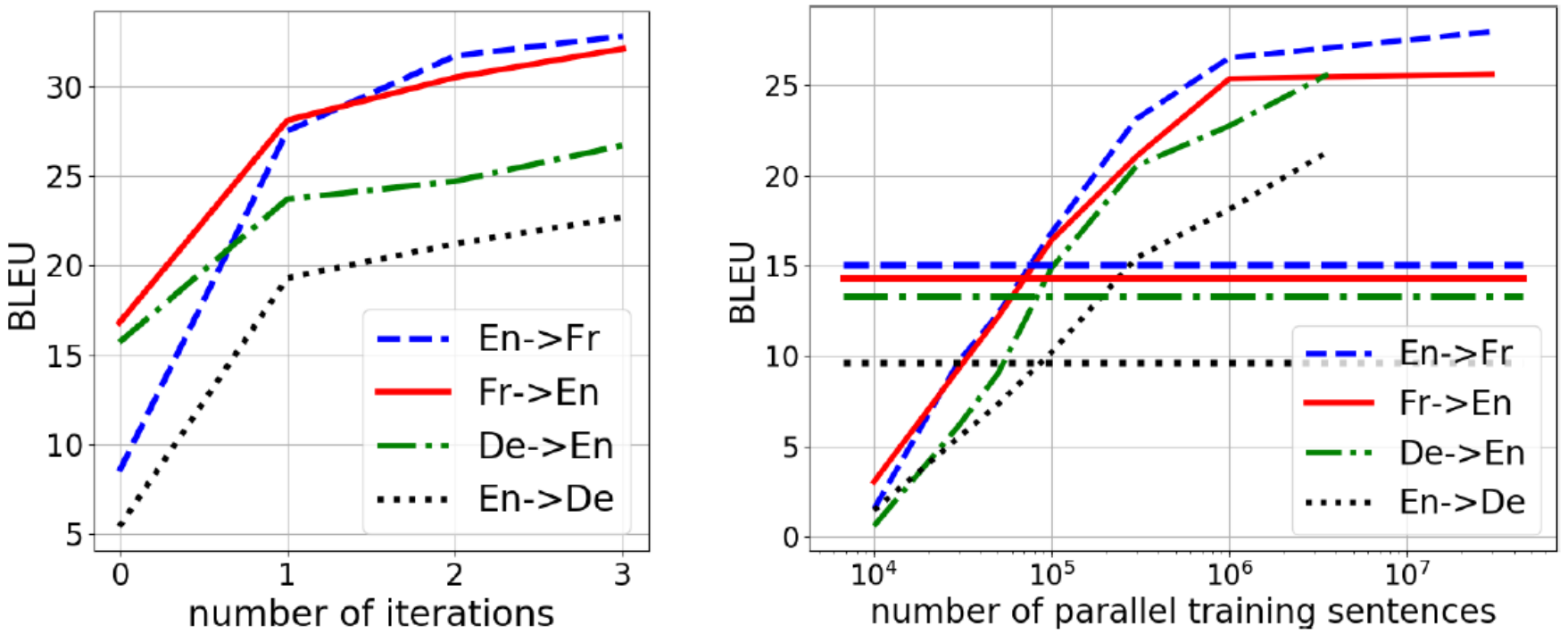


Figure 4: Left: BLEU as a function of the number of iterations of our algorithm on the Multi30k-Task1 datasets. Right: The curves show BLEU as a function of the amount of parallel data on WMT datasets. The unsupervised method which leverages about 15 million monolingual sentences in each language, achieves performance (see horizontal lines) close to what we would obtain by employing 100,000 parallel sentences.

Experiments

Source	un homme est debout près d' une série de jeux vidéo dans un bar .
Iteration 0	a man is seated near a series of games video in a bar .
Iteration 1	a man is standing near a closeup of other games in a bar .
Iteration 2	a man is standing near a bunch of video video game in a bar .
Iteration 3	a man is standing near a bunch of video games in a bar .
Reference	**a man is standing by a group of video games in a bar .**
Source	une femme aux cheveux roses habillée en noir parle à un homme .
Iteration 0	a woman at hair roses dressed in black speaks to a man .
Iteration 1	a woman at glasses dressed in black talking to a man .
Iteration 2	a woman at pink hair dressed in black speaks to a man .
Iteration 3	a woman with pink hair dressed in black is talking to a man .
Reference	**a woman with pink hair dressed in black talks to a man .**
Source	une photo d' une rue bondée en ville .
Iteration 0	a photo a street crowded in city .
Iteration 1	a picture of a street crowded in a city .
Iteration 2	a picture of a crowded city street .
Iteration 3	a picture of a crowded street in a city .
Reference	**a view of a crowded city street .**

Table 3: **Unsupervised translations.** Examples of translations on the French-English pair of the Multi30k-Task1 dataset. Iteration 0 corresponds to word-by-word translation. After 3 iterations, the model generates very good translations.

Experiments

	en-fr	fr-en	de-en	en-de
$\lambda_{cd} = 0$	25.44	27.14	20.56	14.42
Without pretraining	25.29	26.10	21.44	17.23
Without pretraining, $\lambda_{cd} = 0$	8.78	9.15	7.52	6.24
Without noise, $C(x) = x$	16.76	16.85	16.85	14.61
$\lambda_{auto} = 0$	24.32	20.02	19.10	14.74
$\lambda_{adv} = 0$	24.12	22.74	19.87	15.13
Full	**27.48**	**28.07**	**23.69**	**19.32**

Table 4: **Ablation study on the Multi30k-Task1 dataset.**

Conclusion

- New approach to neural machine translation without any alignment between sentences or documents
- The principle of our approach is to start from a simple unsupervised word-by-word translation model, and to iteratively improve this model based on a reconstruction loss, and using a discriminator to align latent distributions of both the source and the target languages.
- Our experiments demonstrate that our approach is able to learn effective translation models without any supervision of any sort.