Unsupervised Machine Translation Using Monolingual Corpora Only

By Guillaume Lample, Ludovic Denoyer, Marc'Aurelio Ranzato

Presentation by: Hassan S. Shavarani

Introduction

- The NMT models work very well only when provided with massive amounts of parallel data (in the order of millions of parallel sentences)
  - parallel corpora are costly to build
  - nonexistent for low-resource languages
- monolingual data is much easier to find, and many languages with limited parallel data still possess significant amounts of monolingual data.

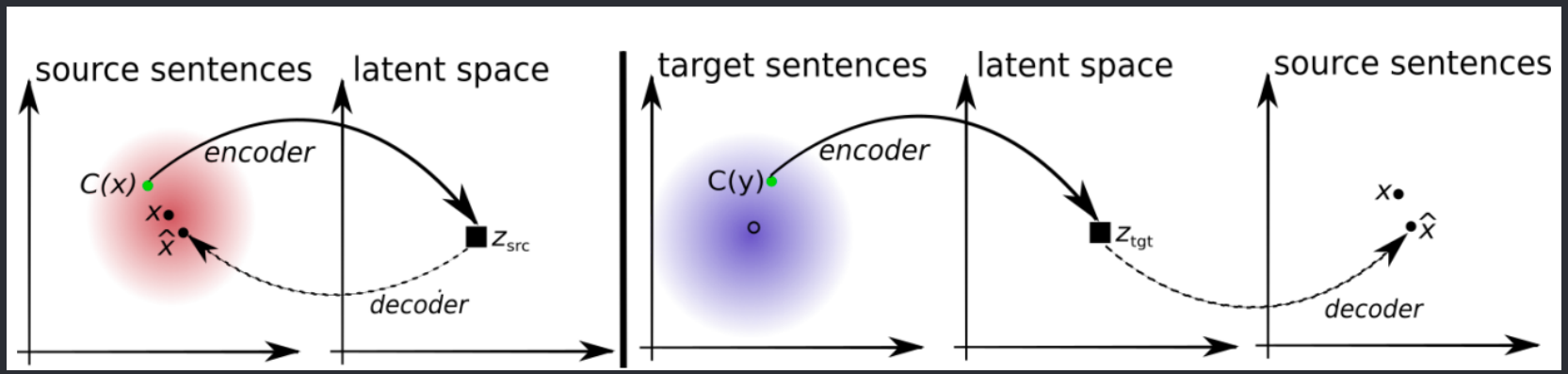
The Problem

- Is it possible to train a general MT system without any form of supervision whatsoever
- Assumption: there exists a monolingual corpus on each language
  - this is applicable whenever we encounter a new language pair for which we have no annotation
  - it provides a strong lower bound performance on what any good semi-supervised approach is expected to yield.

The key idea

- Build a common latent space between the two languages (or domains) and learn to translate by reconstructing in both domains
- Principles:
  - the model has to be able to reconstruct a sentence in a given language from a noisy version of it, as in standard denoising auto-encoders
  - The model also learns to reconstruct any source sentence given a noisy translation of the same sentence in the target domain, and vice versa

Big Picture

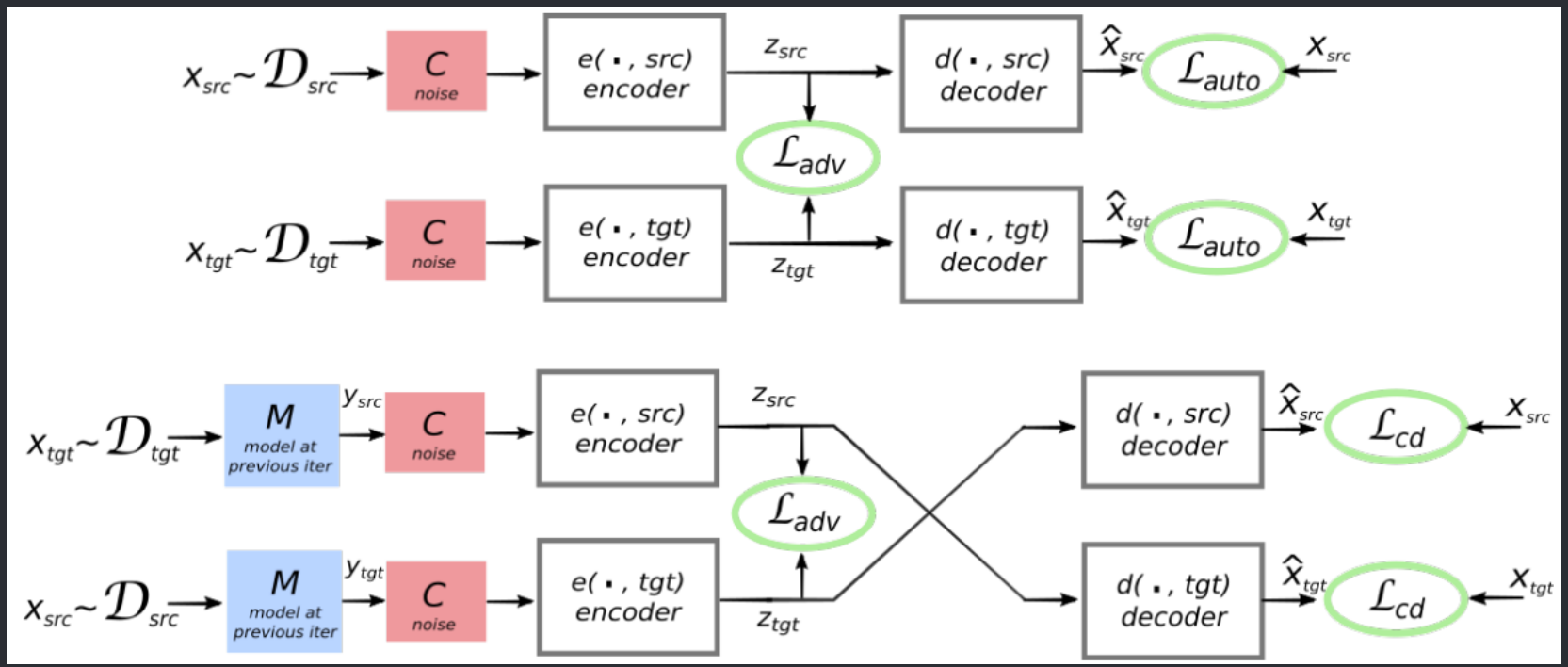

source sentences
latent space
encoder
C(x)
x
x̂
z_src
decoder
target sentences
latent space
source sentences
encoder
C(y)
z_tgt
x
x̂
decoder

Unsupervised NMT Model

- an encoder and a decoder, respectively responsible for encoding source and target sentences to a latent space, and to decode from that latent space to the source or the target domain
- a single encoder and a single decoder for both domains. The only difference when applying these modules to different languages is the choice of lookup tables.

Model Architecture

Auto-Encoding and Translation



Denoising Auto-Encoding

First objective function:

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

- Δ is the sum of token-level cross-entropy losses
- $C(x)$ is a randomly sampled noisy version of sentence x
  - we drop every word in the input sentence with a probability p_{wd}
  - we slightly shuffle the input sentence which makes sure each word stays at most in distance k of its correct position
  - best reported values: $p_{wd} = 0.1$; $k = 3$

Cross Domain Training

Second objective function:

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

- Δ is again the sum of token-level cross-entropy losses
- $y = M(x)$ is the corrupted translation of the sample sentence x generated through applying current translation model (M).

Adversarial Training

Third objective function:

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

$$\mathscr{L}_{adv}(\theta_{enc}, Z|\theta_D) = -\mathbb{E}_{(x_i, \ell_{1/2})}[logp_D(\ell_{2/1}|e(x_i, \ell 1/2))]$$

- we would like our encoder to output features in the same space regardless of the actual language of the input sentence, If such condition is satisfied, our decoder may be able to decode in a certain language regardless of the language of the encoder input sentence

Final Objective Function

Eq. 4 in the paper

$$\mathscr{L}(\theta_{enc}, \theta_{dec}, Z) =$$
$$\lambda_{auto}[\mathscr{L}_{auto}(\theta_{enc}, \theta_{dec}, Z, src) + \mathscr{L}_{auto}(\theta_{enc}, \theta_{dec}, Z, tgt)] +$$
$$\lambda_{cd}[\mathscr{L}_{cd}(\theta_{enc}, \theta_{dec}, Z, src, tgt) + \mathscr{L}_{cd}(\theta_{enc}, \theta_{dec}, Z, tgt, src)] +$$
$$\lambda_{adv}[\mathscr{L}_{adv}(\theta_{enc}, Z|\theta_D)]$$

Algorithm

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 trainingl as in eq. 4
        $\theta_{discr} \leftarrow \arg\min \mathcal{L}_D, \quad \theta_{enc}, \theta_{dec}, \mathcal{Z} \leftarrow \arg\min \mathcal{L}$
        $M^{(t+1)} \leftarrow e^{(t)} \circ d^{(t)}$ // update MT model
    end for
    return $M^{(t+1)}$
end procedure

Validation Measure

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

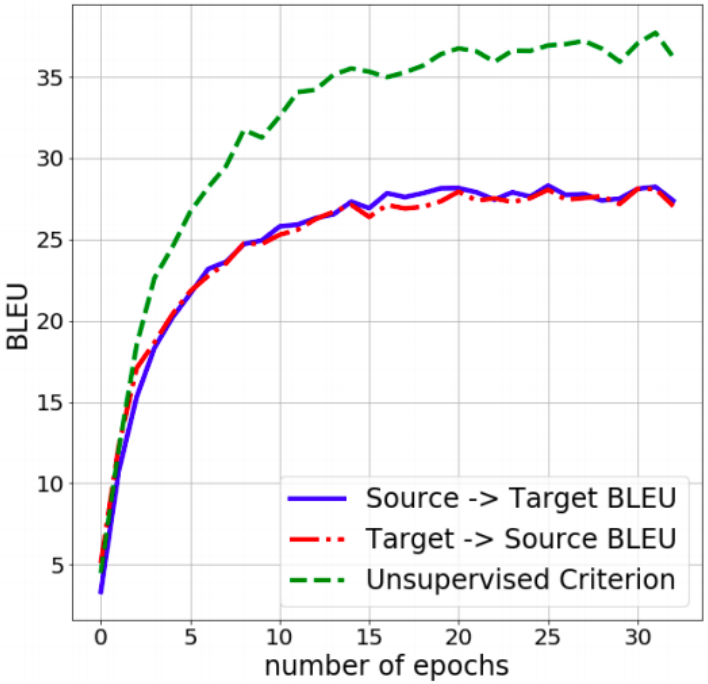


Figure 3: **Unsupervised model selection.** BLEU score of the source to target and target to source models on the Multi30k-Task1 English-French dataset as a function of the number of passes through the dataset at iteration $(t) = 1$ of the algorithm. BLEU correlates very well with the proposed model selection criterion, see Equation 5.

Experimental Setup

Datasets

- Dataset 1
  - train set: **WMT'14 English-French**
    - 30M sentence-pairs with max length of 50 words. devided into two disjoint sets of 15M each.
  - val set: 3000 sentences held out from train data
  - test set: newstest2014
- Dataset 2
  - train set: **WMT'16 English-German**
    - 1.8M sentences processed the same way as WMT'14
  - test set: newstest2016
- Dataset 3
  - Translation annotations of 30K **Multi30k-Task1** images
  - 29K captions as train/1K as test processed the same way as WMT'14

Experimental Setup

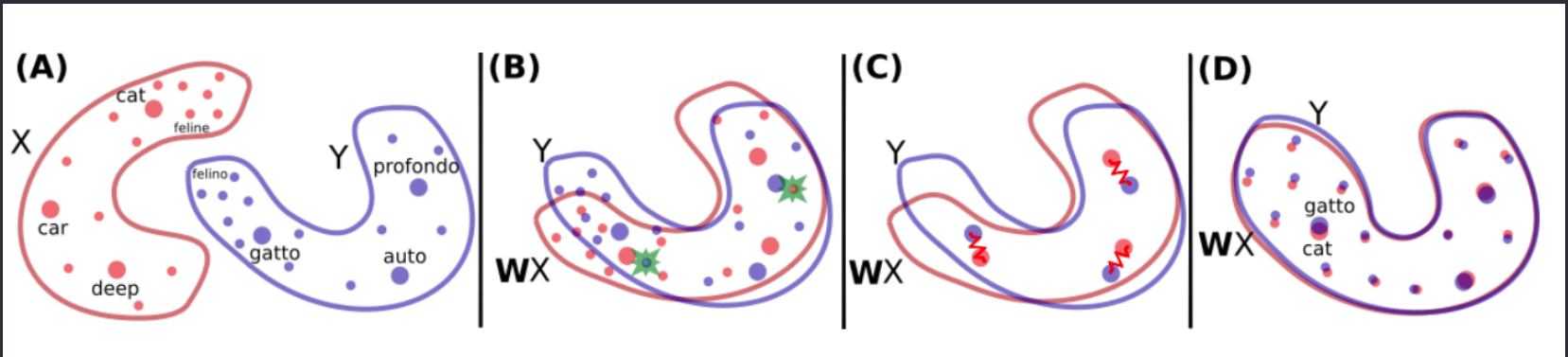
Baselines

- **Word-by-word translation (WBW)**
  - using the inferred bilingual dictionary
- **Word reordering (WR)** [only on WMT dataset]
  - using an LSTM-based language model trained on target side
  - 10 best pairwise swaps of neighbouring words, iteratively
- **Oracle Word Reordering (OWR)**
  - Copy the reordering from the reference
  - use the current translated words (replacement is not allowed)
- **Supervised Learning**
  - Standard cross-entropy loss model trained on data

Unsupervised Dictionary Learning

To initialize the embeddings Z of the model, we

1. train word embeddings on the source and target monolingual corpora using fastText (They need large amounts of data)
2. apply the unsupervised method proposed by (Conneau et al. 2017) to infer a bilingual dictionary which can be use for word-by-word translation [https://arxiv.org/pdf/1710.04087.pdf]



Experimental Details

- **Discriminator Architecture**
  - a MLP with 3 hidden layers of size 1024
  - Leaky-ReLU activation functions
  - Trainer: RMSProp with learning rate of 0.0005
- **Training Details**
  - Trainer: Adam with learning rate of 0.0003; $\beta_1 = 0.5$
  - Mini-Batch size: 32
  - Training: evenly alternate between one encoder-decoder and one discriminator update
  - $\lambda_{auto} = \lambda_{cd} = \lambda_{adv} = 1$

Results [BLEU Scores]

Greedy Decoding Approach

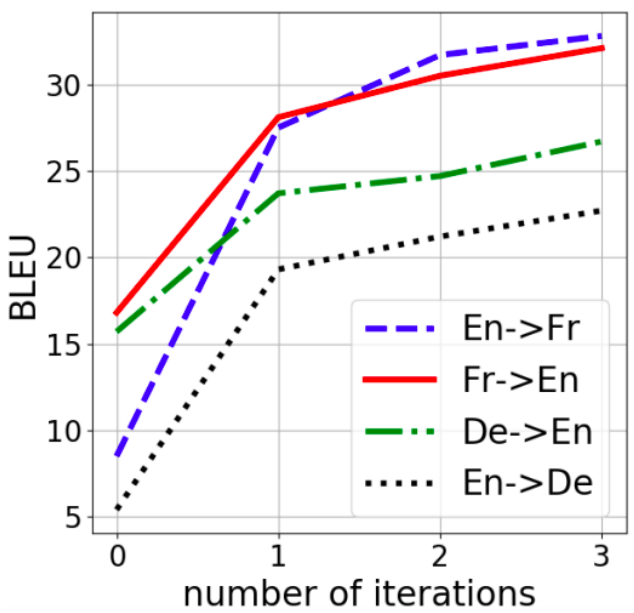
	Multi30k-Task1				WMT			
	en-fr	fr-en	de-en	en-de	en-fr	fr-en	de-en	en-de
Supervised	56.83	50.77	38.38	35.16	27.97	26.13	25.61	21.33
word-by-word	8.54	16.77	15.72	5.39	6.28	10.09	10.77	7.06
word reordering	-	-	-	-	6.68	11.69	10.84	6.70
oracle word reordering	11.62	24.88	18.27	6.79	10.12	20.64	19.42	11.57
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

Examples of unsupervised translations

French-English pair of the Multi30k-Task1 dataset

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 .**

Comparison with supervised approaches



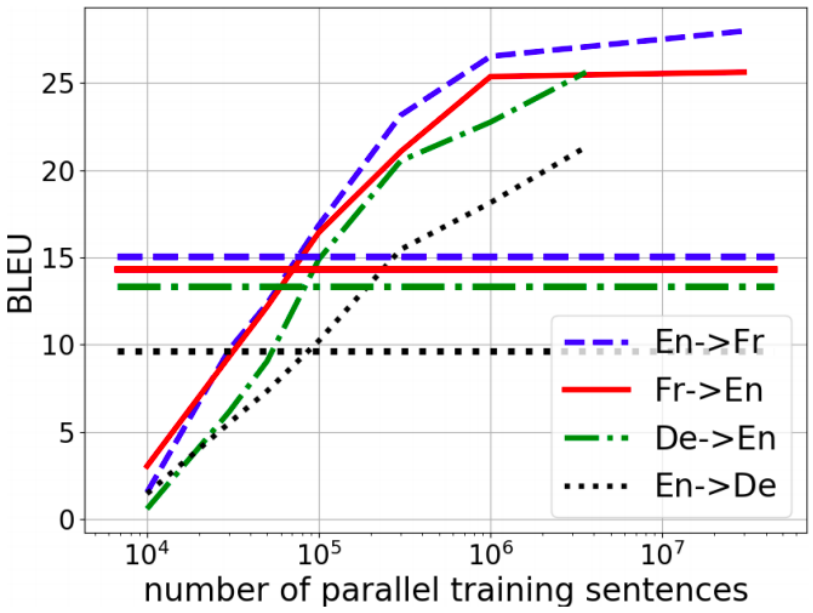


Figure 4: Left: BLEU as a function of the number of iterations of our algorithm on the Multi30k-Task1 datasets. Right: The curves show BLEU as a function of the amount of parallel data on WMT datasets. The unsupervised method which leverages about 10 million monolingual sentences, achieves performance (see horizontal lines) close to what we would obtain by employing 100,000 parallel sentences.

Ablation Study

	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 3: **Ablation study on the Multi30k-Task1 dataset.**

Thanks!