Unsupervised NMT with Weight Sharing

Zhen Yang, Wei Chen, Feng Wang and Bo Xu
Institute of Automation, Chinese Academy of Sciences

2018/07/16

Contents

1 Background

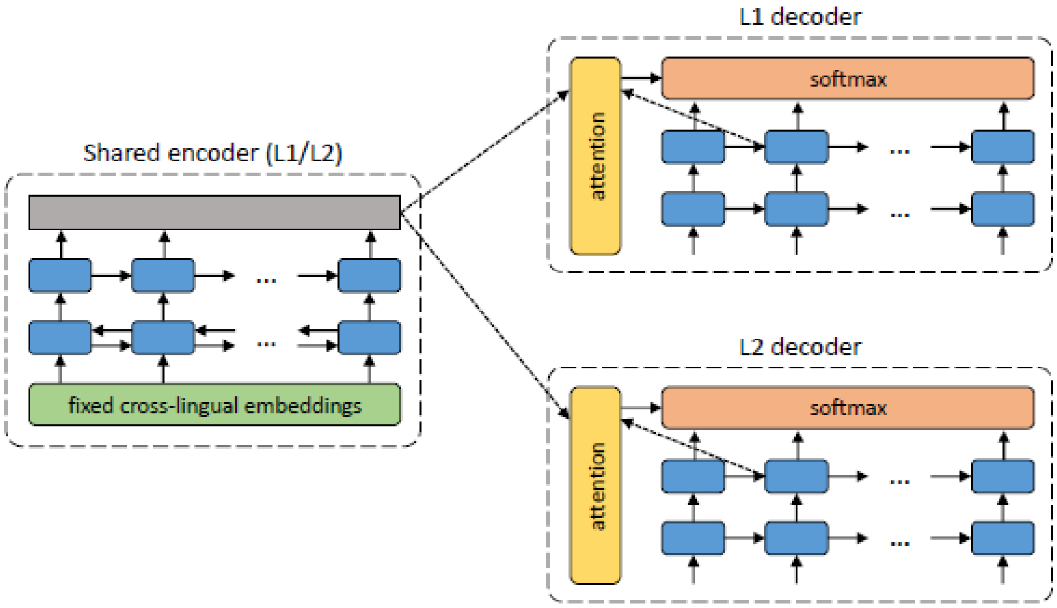
2 The proposed model

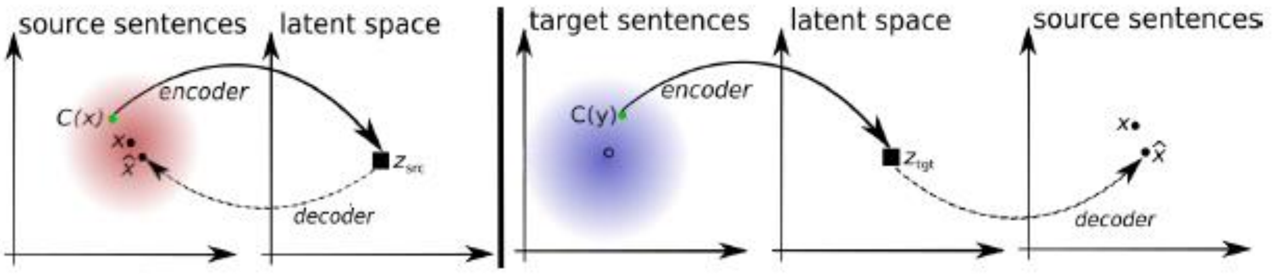
3 Experiments and results

4 Related and future work

Background

Assumption: different languages can be mapped into one shared-latent space





Techniques based on

- Initialize the model with inferred bilingual dictionary

  Unsupervised word embedding mapping

- Learn strong language model

  De-noising Auto-Encoding

- Convert Unsupervised setting into a supervised one

  Back-translation

- Constrain the latent representation produced by encoders to a shared space

  fully-shared encoder   fixed mapped embedding GAN

We find

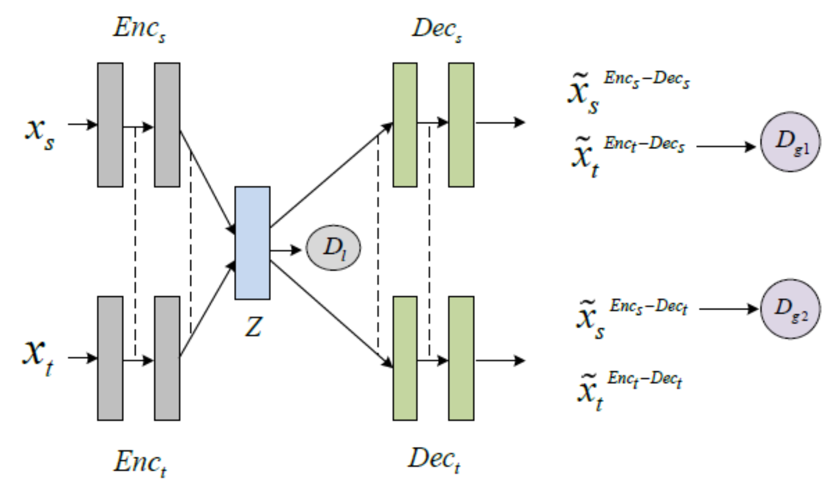
- The shared encoder is a bottleneck for unsupervised NMT

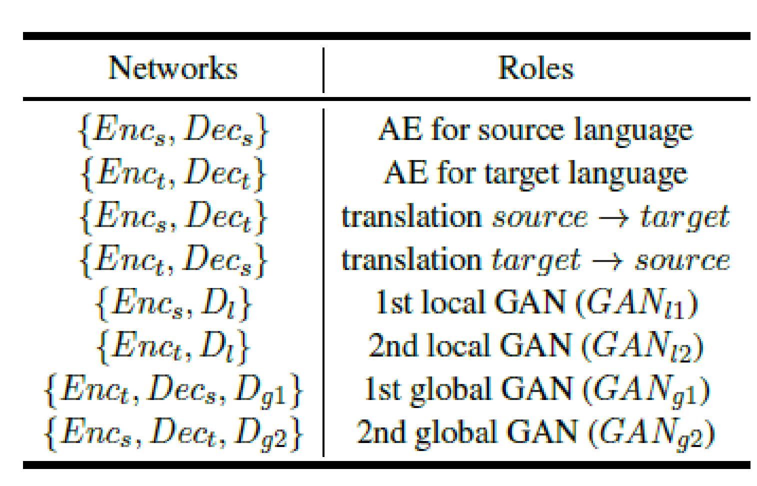
The shared encoder is weak in keeping the unique and internal characteristics of each language, such as the style, terminology and sentence structure. Since each language has its own characteristics, the source and target language should be encoded and learned independently.

- Fixed word embedding also weakens the performance (not included in the paper)

If you are interested about this part, you can find some discussions in our github code: https://github.com/ZhenYangIACAS/unsupervised-NMT

The proposed model:


Enc_s
Dec_s
x_s
x_t
Z
D_l
$\tilde{x}_s^{Enc_s-Dec_s}$
$\tilde{x}_t^{Enc_t-Dec_s}$
D_{g1}
$\tilde{x}_s^{Enc_s-Dec_t}$
D_{g2}
$\tilde{x}_t^{Enc_t-Dec_t}$
Enc_t
Dec_t



Networks	Roles
$\{Enc_s, Dec_s\}$	AE for source language
$\{Enc_t, Dec_t\}$	AE for target language
$\{Enc_s, Dec_t\}$	translation $source \rightarrow target$
$\{Enc_t, Dec_s\}$	translation $target \rightarrow source$
$\{Enc_s, D_l\}$	1st local GAN (GAN_{l1})
$\{Enc_t, D_l\}$	2nd local GAN (GAN_{l2})
$\{Enc_t, Dec_s, D_{g1}\}$	1st global GAN (GAN_{g1})
$\{Enc_s, Dec_t, D_{g2}\}$	2nd global GAN (GAN_{g2})

- The local GAN is utilized to constrain the source and target latent representations to have the same distribution (embedding-reinforced encoder is also designed for this purpose, see our paper for detail).
- The global GAN is utilized to fine tune the whole model.

Experiment setup:

- **Training sets**:

  WMT16En-de, WMT14En-Fr, LDC Zn-En

  Note: The monolingual data is built by selecting the front half of the source language and the back half of the target language.

- **Test sets**:

  newstest2016En-de, newstest2014En-Fr, NIST02En-Zh

- **Model Architecture**:

  4 self-attention layers for encoder and decoder

- **Word Embedding**:

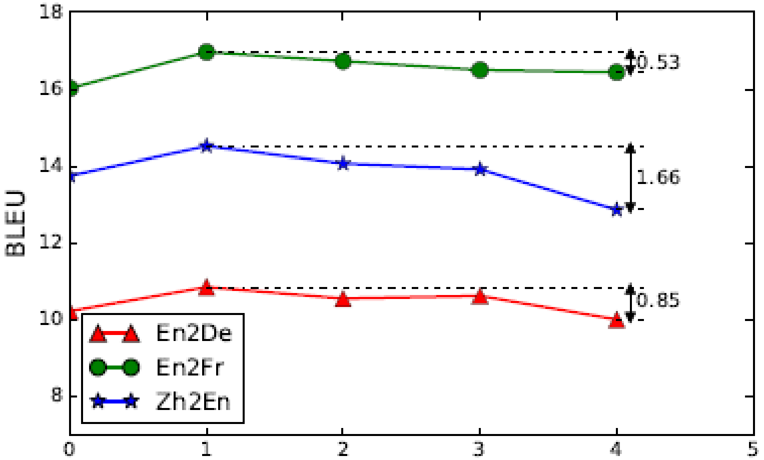
  applying the *Word2vec* to pre-train the word embedding

  utilizing *Vecmap* to map these embedding to a shared-latent space

Experimental results:

- The effects of the weight-sharing layer number

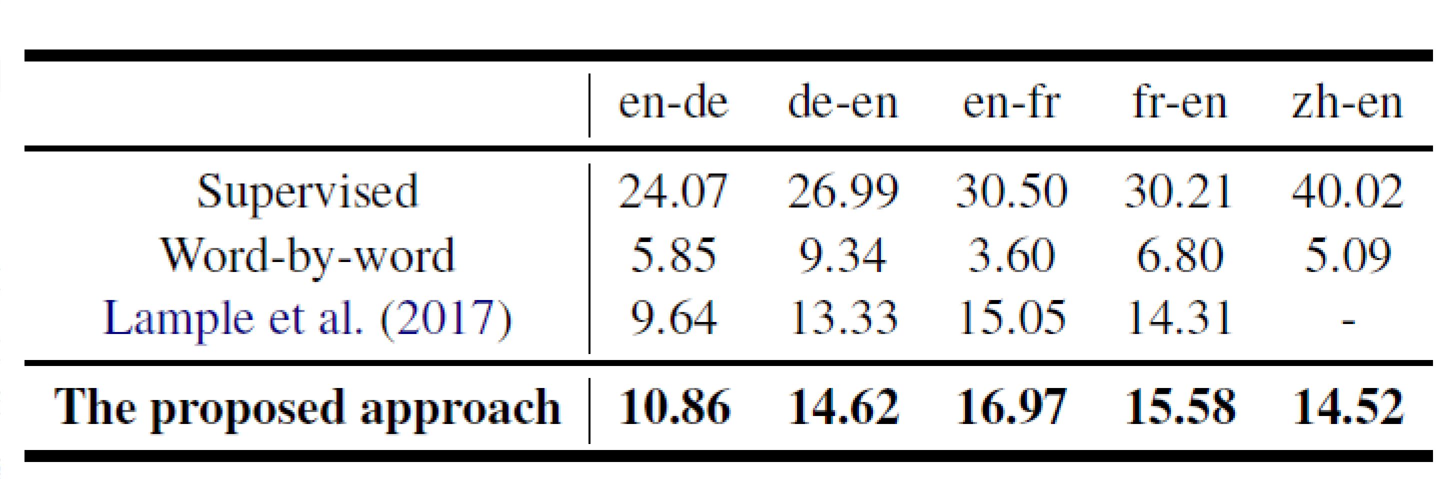
Layers for sharing	En-de	En-Fr	Zh-En
0	10.23	16.02	13.75
1	**10.86**	**16.97**	**14.52**
2	10.56	16.73	14.07
3	10.63	16.50	13.92
4	10.01	16.44	12.86



Sharing one layer achieves the best translation performance.

Experimental results:

- The BLEU results of the proposed model:



	en-de	de-en	en-fr	fr-en	zh-en
Supervised	24.07	26.99	30.50	30.21	40.02
Word-by-word	5.85	9.34	3.60	6.80	5.09
Lample et al. (2017)	9.64	13.33	15.05	14.31	-
The proposed approach	**10.86**	**14.62**	**16.97**	**15.58**	**14.52**

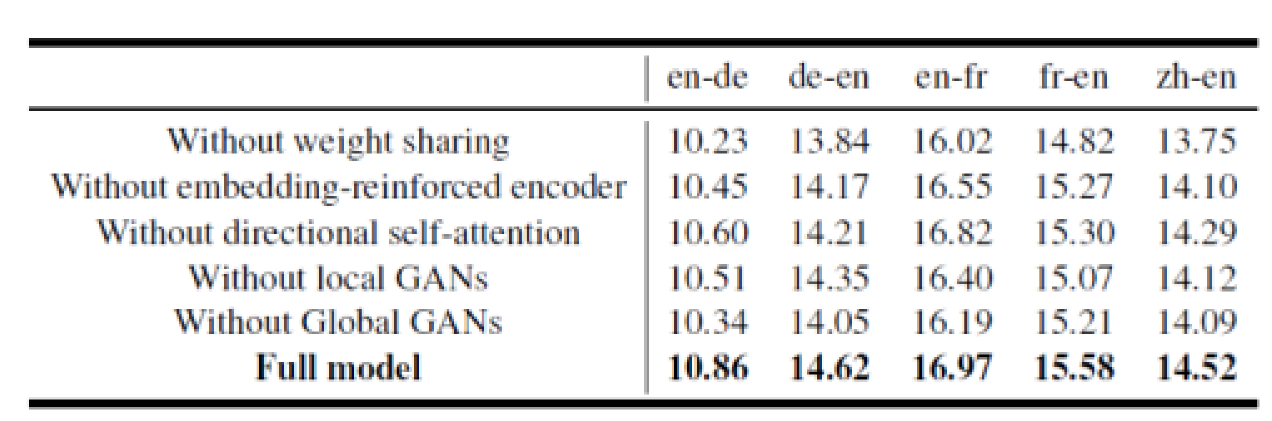
Baseline 1: the word-by-word translation according to the similarity of the word embedding

Baseline 2: “unsupervised NMT with monolingual corpora only” proposed by Facebook.

Upper Bound: the supervised translation on the same model.

Experimental results:

- Ablation study



	en-de	de-en	en-fr	fr-en	zh-en
Without weight sharing	10.23	13.84	16.02	14.82	13.75
Without embedding-reinforced encoder	10.45	14.17	16.55	15.27	14.10
Without directional self-attention	10.60	14.21	16.82	15.30	14.29
Without local GANs	10.51	14.35	16.40	15.07	14.12
Without Global GANs	10.34	14.05	16.19	15.21	14.09
Full model	**10.86**	**14.62**	**16.97**	**15.58**	**14.52**

- We perform an ablation study by training multiple versions of our model with some missing components: the local GAN, global GAN, the directional self-attention, the weight-sharing and the embedding-reinforced encoder.
- We do not test the importance of the auto-encoding, back-translation and the pre-trained embeddings since they have been widely tested in previous works.

Semi-supervised NMT (with 0.2M parallel data)

- Continue training the model after unsupervised training on the parallel data
- From scratch, training the model on monolingual data for one epoch, and then on parallel data for one epoch, and another one on monolingual data, on and on….

Models	BLEU
Only with parallel data	11.59
Fully unsupervised training	10.48
Continuing Training on supervised data	14.51
Jointly training on monolingual and parallel data	**15.79**

Related works:

- G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. 2018. **Unsupervised machine translation using monolingual corpora only**. In International Conference on Learning Representations (ICLR).

- Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. **Unsupervised neural machine translation.** In International Conference on Learning Representations (ICLR).

- G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. 2018 Phrase-Based & Neural Unsupervised Machine Translation (arxiv)

* The newest paper (third one) proposes the shared BPE method for unsupervised NMT, its effectiveness is to be verified (around +10 BLEU points improvement is presented).

Future work:

- Continuing testing the unsupervised NMT and seeking to find its optimal configurations.
- Testing the performance of semi-supervised NMT with a little amount of bilingual data.
- Investigating more effective approach for utilizing the monolingual data in the framework of unsupervised NMT.

Code and new results can be found at:
https://github.com/ZhenYangIACAS/unsupervised-NMT

Thank you !