CycleGAN-like cycle consistency learning for LLM

Seleznyov Mikhail, Sushko Nikita, Kovaleva Maria

Skoltech, 2022

May 26, 2023

1. Problem statement

2. Baseline

3. Cycle consistency

4. Results

Problem

- Traditional methods of Neural Machine Translation (NMT) require a big corpus of paired texts.
- But there is much more unpaired texts on the internet.
- Can we employ this?

Idea

- Observation: if you translate from English to Russian, and then back, you must get the same thing.
- So, having two translation models, we can take unlabeled data, translate it with one model and force other to translate it back.
- And we also can treat labeled data as unlabeled, eliciting more signal from what we have.

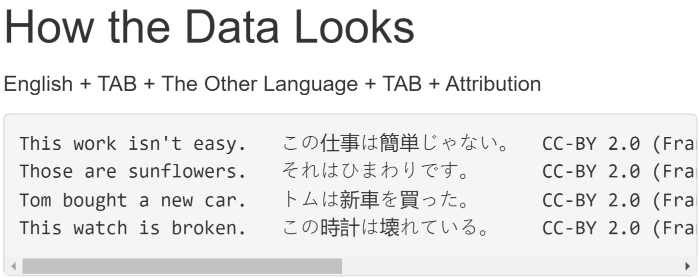
Concrete questions

In our work, we explore three questions:

- Can we benefit from cycle consistency on additional unlabeled data?
- Can we benefit from cycle consistency in low-data regime?
- How does it depend on the model size?

Setup

- To ease validation, we experiment with English-Russian translation.
- We take data from http://www.manythings.org/anki/ – they have a dataset of 467k pairs of short sentences.
- We use BLEU to evalute our models.



How the Data Looks

English + TAB + The Other Language + TAB + Attribution

```
This work isn't easy.    この仕事は簡単じゃない。    CC-BY 2.0 (Fra
Those are sunflowers.    それはひまわりです。        CC-BY 2.0 (Fra
Tom bought a new car.    トムは新車を買った。        CC-BY 2.0 (Fra
This watch is broken.    この時計は壊れている。      CC-BY 2.0 (Fra
```

*We use Russian, but example on the site is for some other language.

Baseline solution

- Machine translation is a seq2seq task, so it's good to use encoder-decoder.
- We tried T5 in two variants, base and small.

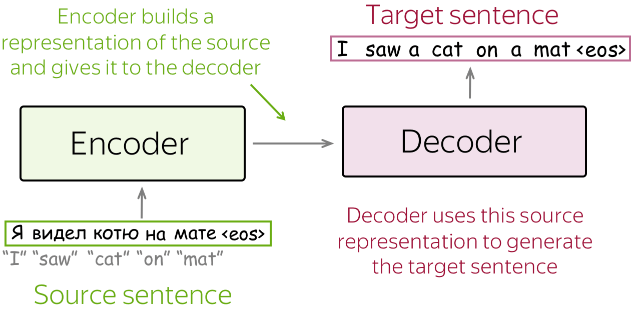
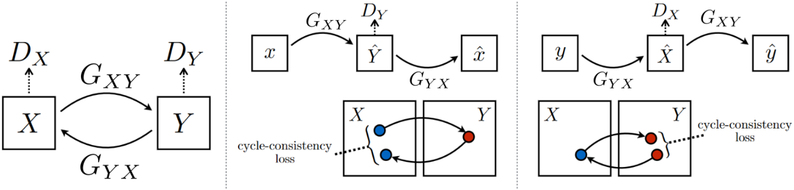

Encoder builds a representation of the source and gives it to the decoder
Target sentence
I saw a cat on a mat <eos>
Encoder
Decoder
Я видел котю на мате <eos>
"I" "saw" "cat" "on" "mat"
Source sentence
Decoder uses this source representation to generate the target sentence

Figure: Seq2seq translation

Cycle consistency

Cycle GAN

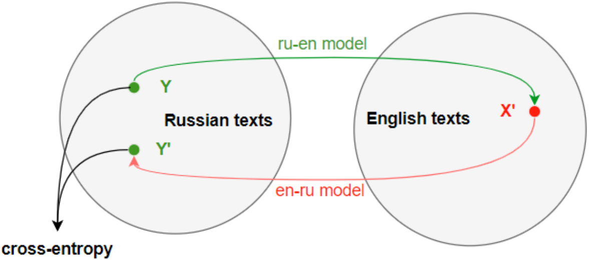
- **Cycle Consistency loss**: $\mathcal{L}_{\text{cyc}}(G_{XY}, G_{YX}) = \mathbb{E}_{x\sim\mathbb{P}}[\|G_{YX}(G_{XY}(x)) - x\|_1] + \mathbb{E}_{y\sim\mathbb{Q}}[\|G_{XY}(G_{YX}(y)) - y\|_1]$
- **Final loss**: $\mathcal{L}(G_{XY}, G_{YX}, D_Y, D_X) = \mathcal{L}_{\text{GAN}}(G_{XY}, D_Y) + \mathcal{L}_{\text{GAN}}(G_{YX}, D_X) + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G_{XY}, G_{YX})$



Cycle consistency

Our loss:

- $\mathcal{L}(G_{XY}, G_{YX}) = \mathbb{E}_{(x,y)\sim\mathbb{P}\times\mathbb{Q}}\mathbf{H}(G_{XY}(x), y) + \mathbb{E}_{(y,x)\sim\mathbb{Q}\times\mathbb{P}}\mathbf{H}(G_{YX}(y), x) + \mathbb{E}_{x\sim\mathbb{P}}\mathbf{H}(G_{YX}(G_{XY}(x)), x) + \mathbb{E}_{y\sim\mathbb{Q}}\mathbf{H}(G_{XY}(G_{YX}(y)), y)$
- **H** - cross-entropy loss



Low-data regime experiments

Models (T5-base) are trained on random subset of 10k samples.

Experiment name	BLEU for en2ru model	BLEU for ru2en model
CrossEntropy loss	4.7844	4.3415
CrossEntropy + Cyclic Loss	**6.1359**	**5.7527**

Table: 10 epochs

Proposed method yields better scores.

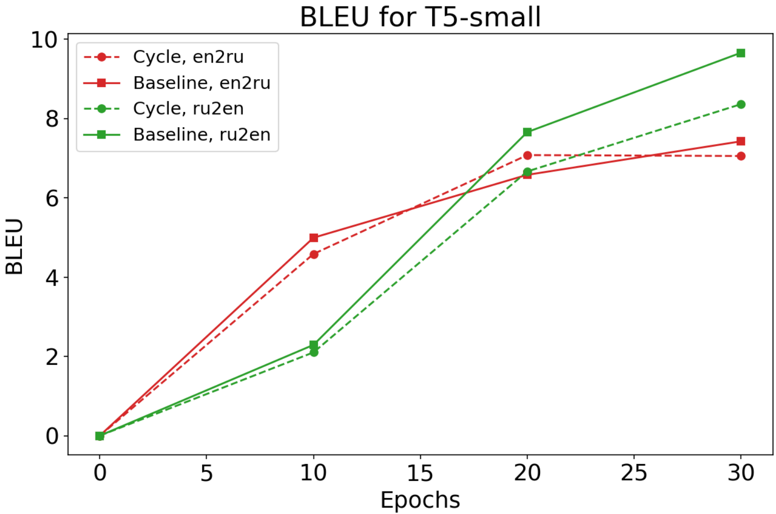
Experiment name	BLEU for en2ru model	BLEU for ru2en model
CrossEntropy loss	**14.7513**	**18.7511**
CrossEntropy + Cyclic Loss	13.7197	17.6676

Table: 30 epochs

Proposed method is marginally worse than baseline.

Small model

Same setup as before, but with T5-small.



Using unpaired data

Multistage: train on paired data (init), then on unpaired data (pretrain), then again on paired data (finetune).
Mixed: alternate paired and unpaired batches.

Experiment name	BLEU for en2ru model	BLEU for ru2en model
CrossEntropy, 10 epochs, subset of labeled data	4.7844	4.3415
CrossEntropy + Cyclic Loss, multistage, 10 + 1 + 10 epochs	**6.7953**	**7.2948**

Experiment name	BLEU for en2ru model	BLEU for ru2en model
CrossEntropy, 30 epochs, subset of labeled data	14.7513	18.7511
CrossEntropy + Cyclic Loss, multistage, 30 + 1 + 10 epochs	15.1079	**20.2587**
CrossEntropy + Cyclic Loss, mixed, 30 epochs	**15.6286**	20.0381

Results

Answers to the questions:

- Can we benefit from additional unlabeled data?
  - Mixed approach brings some benefits. We hypothesize, that with very large amount of epochs it will still be ahead of the baseline, making use of more samples. However, it is dramatically slower (x5-x6).
- Can we benefit from enforcing cycle consistency in low-data regime?
  - Yes, if the amount of training epochs is low enough.
- Does it depend on the model size?
  - In our testing smaller models were harder to train using Cycle Consistency loss.

Team Contributions

- Seleznyov Mikhail
  - Experiment design
  - Implementation of the baseline model
  - Presentation and README.md
  - Experiments
- Sushko Nikita
  - Idea of the project
  - Implementation of helper and utility scripts
  - Presentation and README.md
  - Experiments
- Kovaleva Maria
  - Existing paper research
  - Implementation of Cyclic Loss model
  - Presentation
  - Experiments

Github

https://github.com/Dont-Care-Didnt-Ask/CycleTranslate

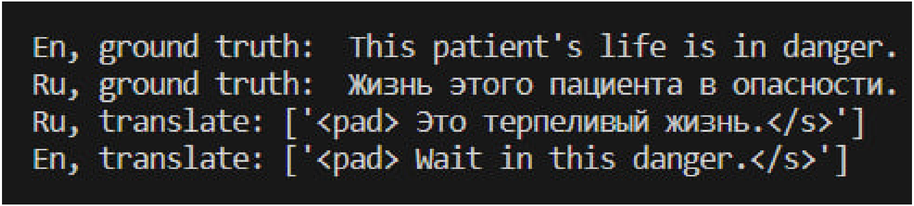
Literature

[1] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[2] Lena Voita. *NLP for You*. 2018. URL: https://lena-voita.github.io/nlp_course.html.

[3] Jun-Yan Zhu et al. *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*. 2020. arXiv: 1703.10593 [cs.CV].

Funny example



```
En, ground truth:  This patient's life is in danger.
Ru, ground truth:  Жизнь этого пациента в опасности.
Ru, translate: ['<pad> Это терпеливый жизнь.</s>']
En, translate: ['<pad> Wait in this danger.</s>']
```

Figure: Instructive mistake