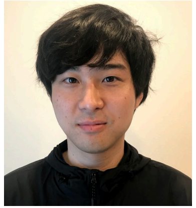
ACL 2019

Effective Adversarial Regularization for Neural Machine Translation

Motoki Sato,[1] Jun Suzuki,[2,3] Shun Kiyono[3,2]



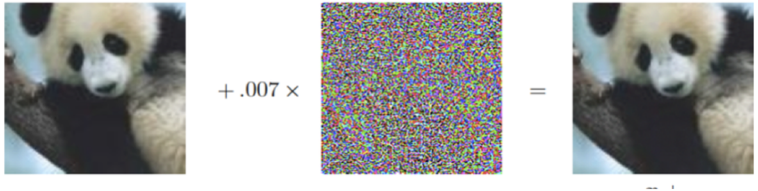
[1] Preferred Networks, Inc.
[2] Tohoku University
[3] RIKEN Center for Advanced Intelligence Project

Preferred Networks

Overview of Our Paper

1. Adversarial Regularization for Image Classification

Image Classification :

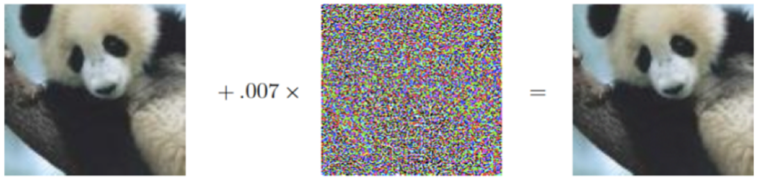


[Goodfellow *et al* .,2015]

Overview of Our Paper

1. Adversarial Regularization for Image Classification

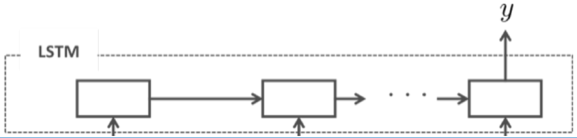
Image Classification :



[Goodfellow *et al* .,2015]

2. Adversarial Regularization for Text Classification

Text Classification :

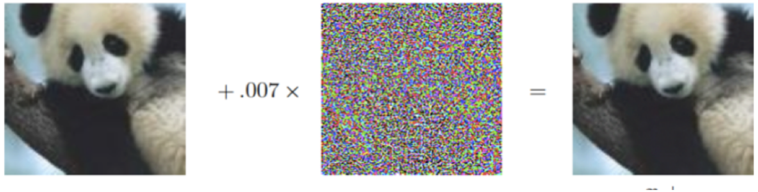


[Miyato et al., 2017]

Overview of Our Paper

1. Adversarial Regularization for Image Classification

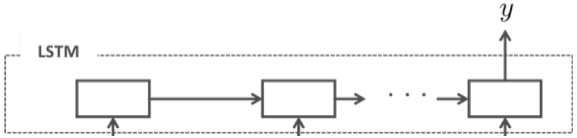
Image Classification :



[Goodfellow *et al* .,2015]

2. Adversarial Regularization for Text Classification

Text Classification :



[Miyato et al., 2017]

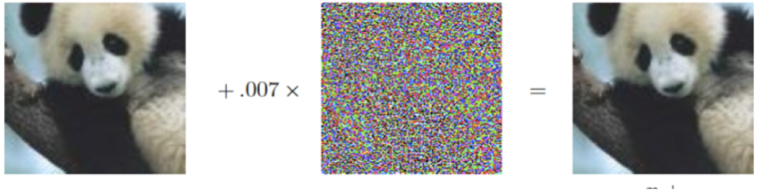
3. Our Main Question

Q. Is "**Adversarial Regularization**" effective for **NMT**?

Overview of Our Paper

1. Adversarial Regularization for Image Classification

Image Classification :



[Goodfellow *et al* .,2015]

2. Adversarial Regularization for Text Classification

Text Classification :

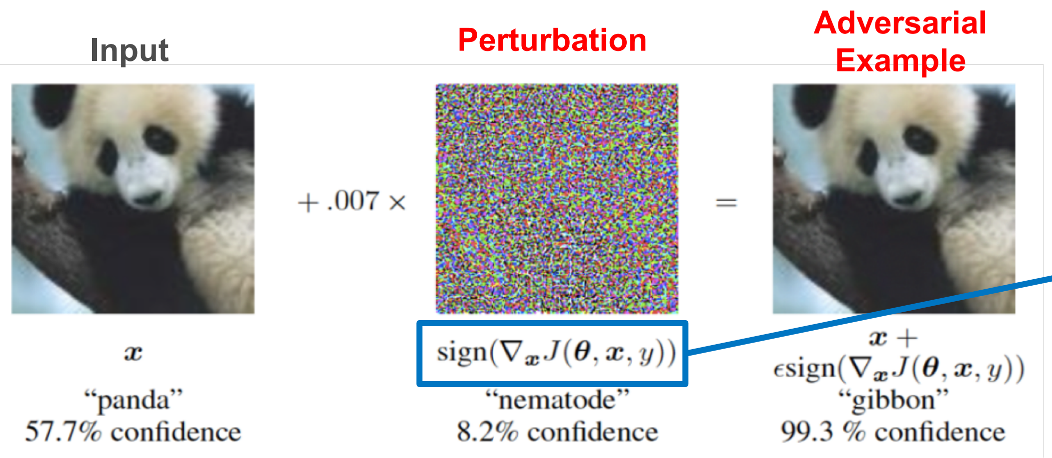
[Miyato et al., 2017]

3. Our Main Question

Q. Is "**Adversarial Regularization**" effective for **NMT**?

1. Adversarial Regularization for Image

[Szegedy et al., 2014, Goodfellow et al .,2015]



[Goodfellow *et al* .,2015]

Perturbation:

The gradient of loss function.

Idea:

$$\tilde{J}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \underbrace{\alpha J(\boldsymbol{\theta}, \boldsymbol{x}, y)}_{\text{Training Example}} + \underbrace{(1-\alpha) J(\boldsymbol{\theta}, \boldsymbol{x} + \epsilon \mathrm{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right))}_{\text{Adversarial Example}}$$

Purpose

Improve generalization performance.

[Goodfellow *et al* .,2015]

Overview of Our Paper

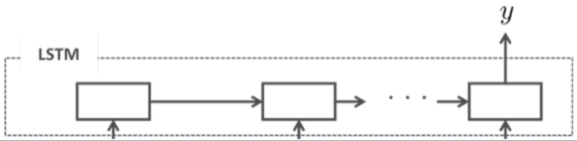
1. Adversarial Regularization for Image Classification

Image Classification :

[Goodfellow *et al* .,2015]

2. Adversarial Regularization for Text Classification

Text Classification :



[Miyato et al., 2017]

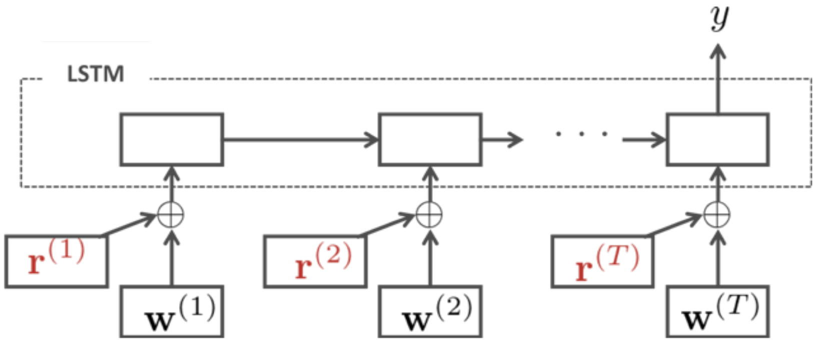
3. Our Main Question

Q. Is "**Adversarial Regularization**" effective for **NMT**? 🤔

2. Adversarial Regularization for Text

[Miyato et al., 2017]

- The perturbation is applied to the **word embedding** layer.
- The adversarial regularization improves the performance on **text classification** task.



w : **Word Embedding**

r : **Adversarial Perturbations**

$$\hat{\boldsymbol{r}}_i = \epsilon \frac{\boldsymbol{a}_i}{||\boldsymbol{a}||_2}, \qquad \boldsymbol{a}_i = \nabla_{\boldsymbol{e}_i} \ell(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Theta})$$

ϵ = 1.0 (hyper-parameter)

$$\underline{\mathcal{A}(\mathcal{D}, \boldsymbol{\Theta})} = -\frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{D}} \ell(\boldsymbol{X}, \hat{\boldsymbol{r}}, \boldsymbol{Y}, \boldsymbol{\Theta})$$

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \left\{ \mathcal{J}(\mathcal{D}, \boldsymbol{\Theta}) + \lambda \underline{\mathcal{A}(\mathcal{D}, \boldsymbol{\Theta})} \right\}$$

Two Options for Computing the Perturbation (How to define "loss function")

2. Adversarial Regularization for Text

[Miyato et al., 2017]

$$\hat{\boldsymbol{r}}_i = \epsilon \frac{\boldsymbol{a}_i}{||\boldsymbol{a}||_2}, \quad \boldsymbol{a}_i = \nabla_{\boldsymbol{e}_i} \ell(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Theta}).$$

Two Options for Computing the Perturbation (how to define "loss function")

① **Adversarial Training (AdvT)** [Goodfellow et al .,2015]

→ compute the loss from the gold label (i.e. target sequence)

$$\ell(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Theta}) = \log\left(p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\Theta})\right)$$

② **Virtual Adversarial Training (VAT)** [Miyato et al., 2016]

→ compute the loss with KL divergence.

$$\ell_{\mathrm{KL}}(\boldsymbol{X}, \hat{\boldsymbol{r}}, \cdot, \boldsymbol{\Theta}) = \mathrm{KL}\left(p(\cdot \mid \boldsymbol{X}, \boldsymbol{\Theta}) || p(\cdot \mid \boldsymbol{X}, \hat{\boldsymbol{r}}, \boldsymbol{\Theta})\right)$$

Overview of Our Paper

1. Adversarial Regularization for Image Classification

Image Classification :

[Goodfellow *et al* .,2015]

2. Adversarial Regularization for Text Classification

Text Classification :

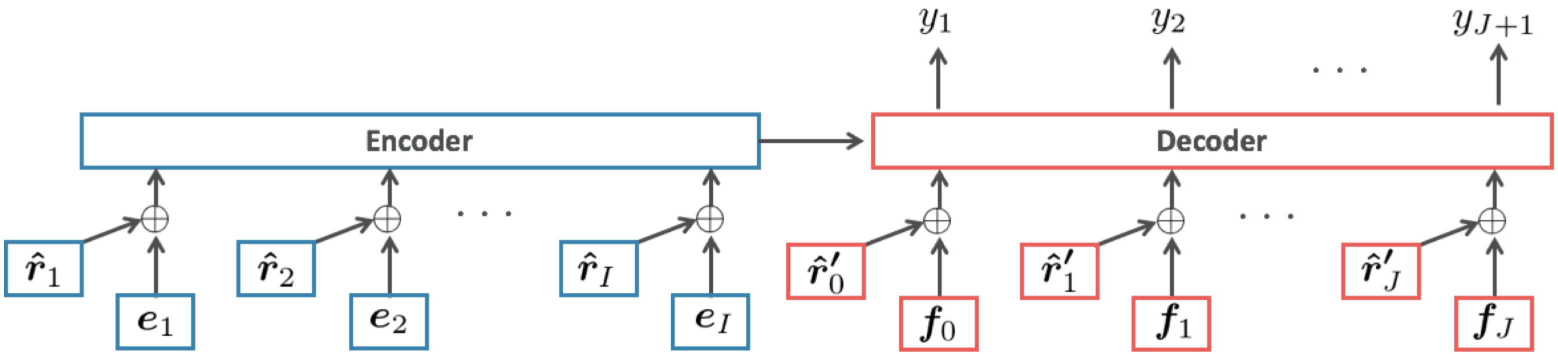
[Miyato et al., 2017]

3. Our Main Question

Q. Is "**Adversarial Regularization**" effective for **NMT**? 🤔

Our Main Question

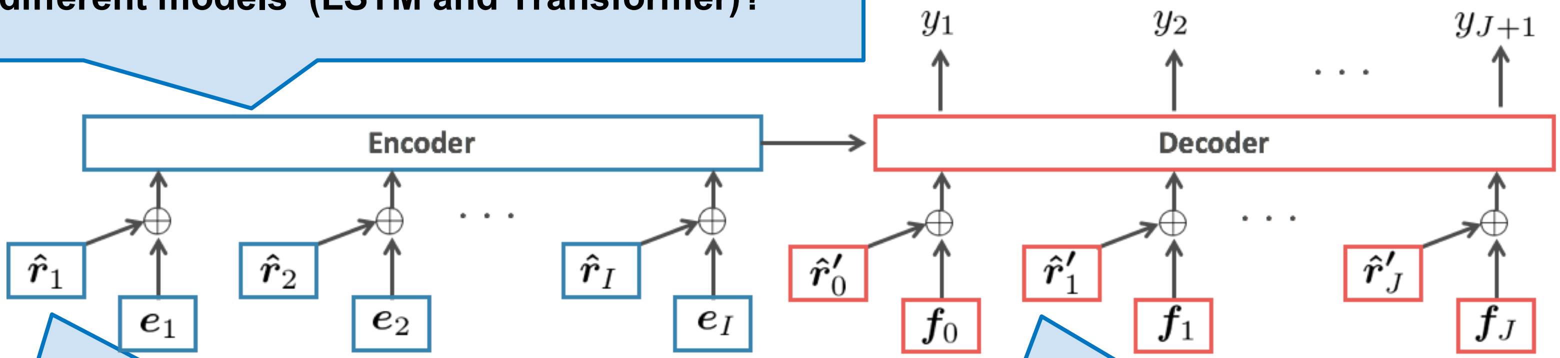
Q. Is "**Adversarial Regularization**" effective for **NMT**? 🤔



Our Main Question

Q. Is "**Adversarial Regularization**" effective for **NMT**? 🤔

① **Is adversarial regularization effective across different models (LSTM and Transformer)?**



② **How should we compute the perturbation?:**
- **Adversarial Training** [Goodfellow et al ., 2015]
- **Virtual Adversarial Training** [Miyato et al., 2017]

③ **Where should we apply the perturbation?:**
Encoder-side, **Decoder-side**, or **Both-side**.

Experimental Setup

- **Dataset**: IWSLT 2016 [Cettolo et al., 2012]
- **Configurations**
  - **1. Model Architecture**
    - LSTM w/ attention [Luong et al., 2015]
    - Transformer [Vaswani et al., 2017]
  - **2. Adversarial regularization techniques**
    - Adversarial Training (AdvT) [Goodfellow et al .,2015]
    - Virtual Adversarial Training (VAT) [Miyato et al., 2017]
  - **3. Perturbation positions**
    - encoder-side, decoder-side, both-side (enc & dec)
- **Language Pairs**
  - **EN→FR, FR→EN,EN→DE, DE→EN**
- **Evaluation**
  - BLEU score [Papineni et al., 2002]

What is the Most Effective Configuration?

		IWSLT (EN→DE)	
Model	Perturbation	test2013	test2014
LSTM	(None)	27.73	23.98
+AdvT	enc	**28.73**	**24.90**
	dec	27.44	23.71
	enc-dec	28.47	24.78
+VAT	enc	29.03	24.75
	dec	27.49	23.20
	enc-dec	**29.47**	**24.92**
Transformer	(None)	29.15	25.19
+AdvT	enc	29.04	25.16
	dec	28.95	25.75
	enc-dec	**29.61**	**25.78**
+VAT	enc	29.95	26.00
	dec	29.62	25.88
	enc-dec	**<u>30.13</u>**	**<u>26.06</u>**

Results

- **Adversarial regularization improves the performance of LSTM & Transformer.**
- **VAT consistently outperforms AdvT.**
- **"enc-dec" is the best position to apply the perturbation.**

Findings

- **Transformer + VAT (Both-side) is the most effective configuration**

Results on four language pair

Model	Perturbation	DE→EN test2013	DE→EN test2014	FR→EN test2013	FR→EN test2014	EN→DE test2013	EN→DE test2014	EN→FR test2013	EN→FR test2014
Transformer	**None**	34.22	30.19	38.87	37.20	29.15	25.19	40.43	37.90
+ VAT	enc-dec	35.06	**31.10**	40.09	37.89	**30.13**	26.06	41.13	38.64
+ VAT + AdvT	enc-dec	**35.50**	30.88	**40.26**	**38.44**	30.04	**26.33**	**41.67**	**38.72**

Findings

- **Transformer+VAT consistently outperformed the baseline (Transformer)**
- **AdvT and VAT can be combined to further improve the performance**

Back-translation + Adversarial Regularization

Q. Is **"Back-translation"** effective with **VAT**? 🤔

[Sennrich et al., 2016]

We incorporated **pseudo-parallel corpora** generated using **back-translation** [Sennrich et al., 2016] as **additional training data**. (we used the WMT14 news translation corpus.)

Back-translation + Adversarial Regularization

Q. Is "**Back-translation**" effective with **VAT**? 🤔
[Sennrich et al., 2016]

We incorporated **pseudo-parallel corpora** generated using **back-translation** [Sennrich et al., 2016] as **additional training data**. (we used the WMT14 news translation corpus.)

		DE→EN		FR→EN		EN→DE		EN→FR	
Model	Perturbation	test2013	test2014	test2013	test2014	test2013	test2014	test2013	test2014
Transformer (baseline)	**None**	34.22	30.19	38.87	37.20	29.15	25.19	40.43	37.90
Transformer + BT	**None**	35.44	31.08	40.44	38.42	30.73	26.02	41.74	39.03
Transformer + BT + VAT	**enc-dec**	**36.43**	**32.53**	**41.29**	**39.76**	**31.99**	**27.20**	**43.41**	**40.15**

Findings

- **Adversarial regularization** can be combined with **back-translation technique.**

Take Home Message of Our Presentation

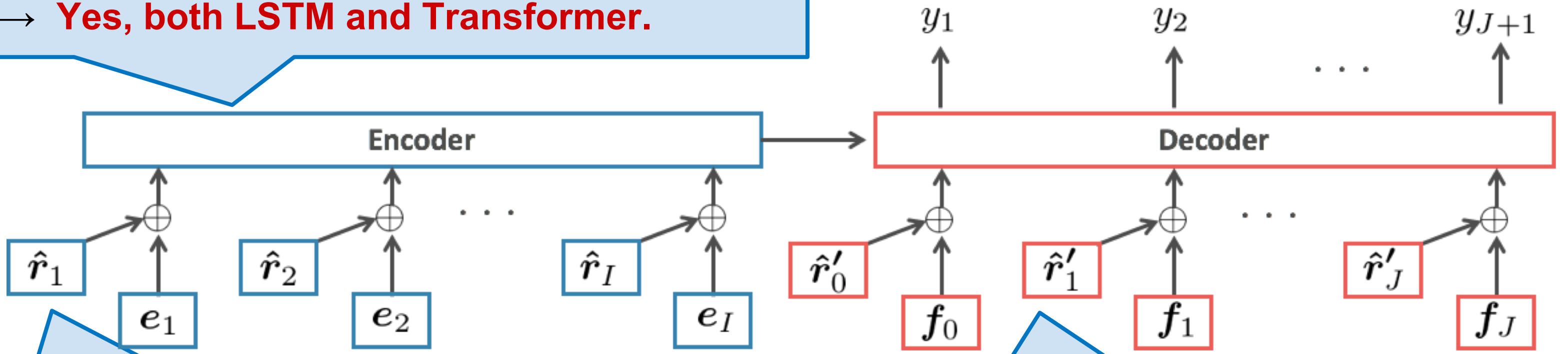
Q. Is “**Adversarial Regularization**” effective for **NMT**? 🤔

→ **YES!!** 😊

code: pfnet-research/vat_nmt

Chainer

① **Is adversarial regularization effective across different models (LSTM and Transformer)?**
→ **Yes, both LSTM and Transformer.**



② **How should we compute the perturbation?:**
Virtual Adversarial Training outperforms Adversarial Training.

③ **Where should we apply the perturbation?:**
Adding perturbation to both embedding layer is the most effective configuration.

code:
pfnet-research/vat_nmt

Motoki Sato

Shun Kiyono

Jun Suzuki

Thank you for your attention!

References

- **[Szegedy et al., 2014]**
  "Intriguing properties of neural networks", ICLR.
- **[Goodfellow et al .,2015]**
  "Explaining and Harnessing Adversarial Examples", ICLR.
- **[Miyato et al., 2016]**
  "Distributional Smoothing with Virtual Adversarial Training", ICLR.
- **[Miyato et al., 2017]**
  "Adversarial Training Methods for SemiSupervised Text Classification", ICLR.
- **[Clark et al., 2018]**
  "Semi-Supervised Sequence Modeling with Cross-View Training", EMNLP.
- **[Vaswani et al., 2017]**
  "Attention is All you Need", NIPS
- **[Luong et al., 2015]**
  "Effective Approaches to Attentionbased Neural Machine Translation", EMNLP
- **[Sennrich et al., 2016]**
  "Improving Neural Machine Translation Models with Monolingual Data", ACL.

Translated Example

Input	meine gebildete Mutter aber wurde Lehrerin .
Reference	but my educated mother became a teacher .
Baseline (Transformer)	my educated mother , though , became a teacher .
Proposed (Transformer+VAT w/ BT)	but my educated mother became a teacher .
Input	aber man kann sehen , wie die Menschen miteinander kommunizieren , zu welchen Zeiten sie einander anrufen , wann sie zu Bett gehen .
Reference	but you can see how your people are communicating with each other , what times they call each other , when they go to bed .
Baseline (Transformer)	but you can see how people talk to each other about what time they call each other when they go to bed .
Proposed (Transformer+VAT w/ BT)	but you can see how people communicate with each other , at which time they call each other , when they go to bed .
Input	wer im Saal hat ein Handy dabei ?
Reference	who in the room has a mobile phone with you ?
Baseline (Transformer)	who in the room has a cell phone in it ?
Proposed (Transformer+VAT w/ BT)	who in the room has a cell phone with me ?

Table 4: Example translation from German→English (test2013).

Virtual Adversarial Training

[Miyato et al., 2016]

$$\ell_{\mathrm{KL}}(\boldsymbol{X}, \hat{\boldsymbol{r}}, \cdot, \boldsymbol{\Theta}) = \mathrm{KL}\big(p(\cdot \mid \boldsymbol{X}, \boldsymbol{\Theta}) || p(\cdot \mid \boldsymbol{X}, \hat{\boldsymbol{r}}, \boldsymbol{\Theta})\big),$$

$$\hat{\boldsymbol{r}}_i = \epsilon \frac{\boldsymbol{a}_i}{||\boldsymbol{a}||_2}, \quad \boldsymbol{a}_i = \nabla_{\boldsymbol{e}_i} \ell(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\Theta}).$$

Experimental Results: Other Directions

	Model	Perturbation	DE→EN test2013	DE→EN test2014	FR→EN test2013	FR→EN test2014	EN→DE test2013	EN→DE test2014	EN→FR test2013	EN→FR test2014
	LSTM	**None**	32.71	28.53	39.09	36.25	27.73	23.98	38.89	36.18
	Transformer	**None**	34.22	30.19	38.87	37.20	29.15	25.19	40.43	37.90
	+ VAT	enc-dec	35.06	**31.10**	40.09	37.89	**30.13**	26.06	41.13	38.64
	+ VAT + AdvT	enc-dec	**35.50**	30.88	**40.26**	**38.44**	30.04	**26.33**	**41.67**	**38.72**
w/ **BT**	Transformer	enc-dec	35.44	31.08	40.44	38.42	30.73	26.02	41.74	39.03
	+ VAT	enc-dec	36.43	**32.53**	41.29	**39.76**	**31.99**	**27.20**	**43.41**	**40.15**
	+ VAT + AdvT	enc-dec	**36.49**	32.39	**41.56**	39.64	31.29	27.05	42.61	39.95

- **AdvT** and **VAT** can be combined to further improve the performance
- **Adversarial regularization** can be combined with **back-translation technique** [Sennrich et al., 2016]

Back-translation

- (Example) EN → DE
  - (x, y) IWSLT
  - (y') WMT 14 corpus (target side unlabeled text)
  - y' → x' (pseudo-parallel corpus)