Unsupervised Machine Translation

CS585, Fall 2019

Simeng Sun (simengsun@umass.edu)

Lots of content from Lample's talk: https://nlpparis.files.wordpress.com/2018/09/talk_meetup_nlp_guillaume_lample.pdf

Quick review: supervised machine translation

- Parallel data

  Fr: Une photo d’ une rue bondée en ville .

  En: A view of a crowded city street .

- (fr→ en)
- Supervised MT objective:

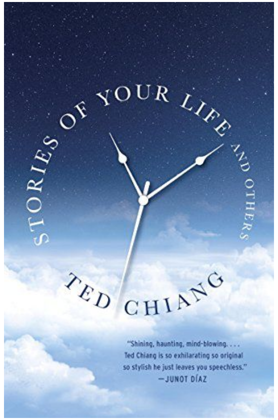
$$\arg\max_{e} p(e \mid f)$$

Do we have enough parallel data?

Parallel Corpus	Sentences	Parallel Corpus	Sentences
Romanian-English	399,375	Greek-English	1,235,976
Bulgarian-English	406,934	Swedish-English	1,862,234
Slovene-English	623,490	Italian-English	1,909,115
Hungarian-English	624,934	German-English	1,920,209
Polish-English	632,565	Finnish-English	1,924,942
Lithuanian-English	635,146	Portuguese-English	1,960,407
Latvian-English	637,599	Spanish-English	1,965,734
Slovak-English	640,715	Danish-English	1,968,800
Czech-English	646,605	Dutch-English	1,997,775
Estonian-English	651,746	French-English	2,007,723

Europarl parallel data: http://www.statmt.org/europarl/

What if we don’t have parallel data?



How we trained a translation model from West African Pidgin to English without a single parallel sentence

"Every act of communication is a miracle of translation."
— Ken Liu

https://towardsdatascience.com/how-we-trained-a-translation-model-from-west-african-pidgin-to-english-without-parallel-sentences-e54efa9f8353 https://arxiv.org/pdf/1804.07755.pdf

Phrase-Based & Neural Unsupervised Machine Translation

Guillaume Lample[†]
Facebook AI Research
Sorbonne Universités
glample@fb.com

Myle Ott
Facebook AI Research
myleott@fb.com

Alexis Conneau
Facebook AI Research
Université Le Mans
aconneau@fb.com

Ludovic Denoyer[†]
Sorbonne Universités
ludovic.denoyer@lip6.fr

Marc'Aurelio Ranzato
Facebook AI Research
ranzato@fb.com

https://arxiv.org/pdf/1711.00043.pdf

UNSUPERVISED MACHINE TRANSLATION USING MONOLINGUAL CORPORA ONLY

Guillaume Lample † ‡ , Alexis Conneau † , Ludovic Denoyer ‡ , Marc'Aurelio Ranzato †
† Facebook AI Research,
‡ Sorbonne Universités, UPMC Univ Paris 06, LIP6 UMR 7606, CNRS
{gl,aconneau,ranzato}@fb.com, ludovic.denoyer@lip6.fr

https://arxiv.org/pdf/1901.07291.pdf

Cross-lingual Language Model Pretraining

Guillaume Lample*
Facebook AI Research
Sorbonne Universités
glample@fb.com

Alexis Conneau*
Facebook AI Research
Université Le Mans
aconneau@fb.com

https://arxiv.org/pdf/1710.11041.pdf

UNSUPERVISED NEURAL MACHINE TRANSLATION

Mikel Artetxe, Gorka Labaka & Eneko Agirre
IXA NLP Group
University of the Basque Country (UPV/EHU)
{mikel.artetxe,gorka.labaka,e.agirre}@ehu.eus

Kyunghyun Cho
New York University
CIFAR Azrieli Global Scholar
kyunghyun.cho@nyu.edu

Outline

- Back-translation (Sennrich et al. 2016)
- Unsupervised word translation (Conneau et al. 2018)
- Unsupervised sentence translation (Lample et al. 2018)
- XLM (Lample & Conneau 2019)

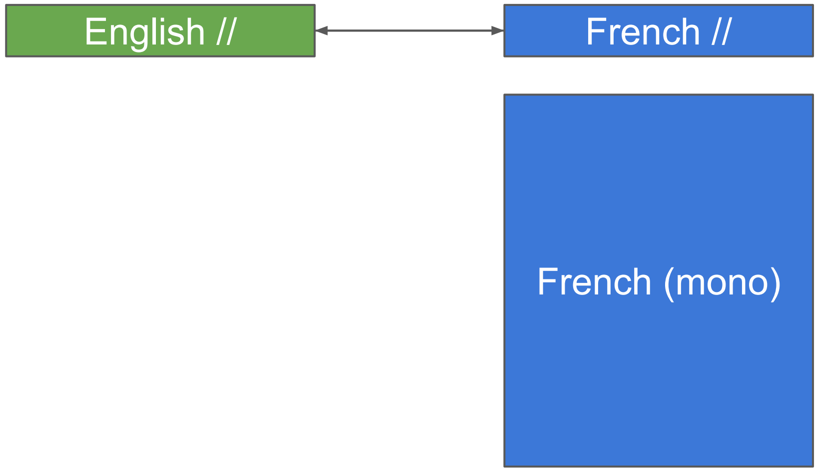
Outline

- **Back-translation (Sennrich et al. 2016)**
- Unsupervised word translation (Conneau et al. 2018)
- Unsupervised sentence translation (Lample et al. 2018)
- XLM (Lample & Conneau 2019)

Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

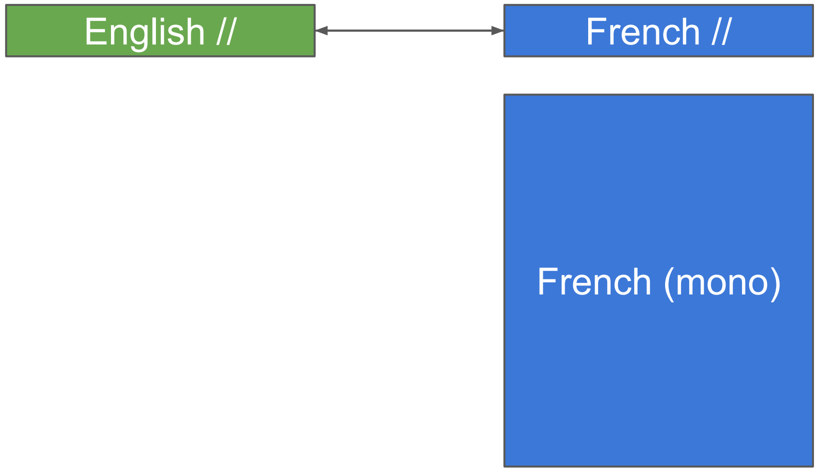
- Small parallel dataset
- Huge monolingual corpus in target language



Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

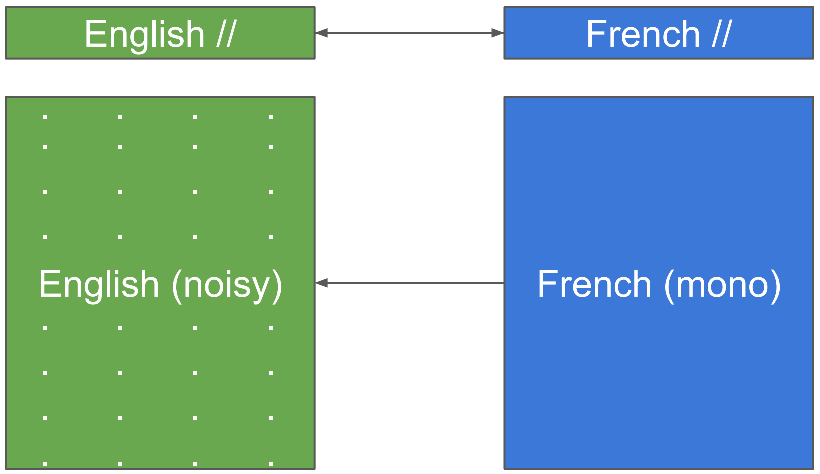
- Small parallel dataset
- Huge monolingual corpus in target language
- Train a (target → source) model $\mathbf{M}_{t2s}$



Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

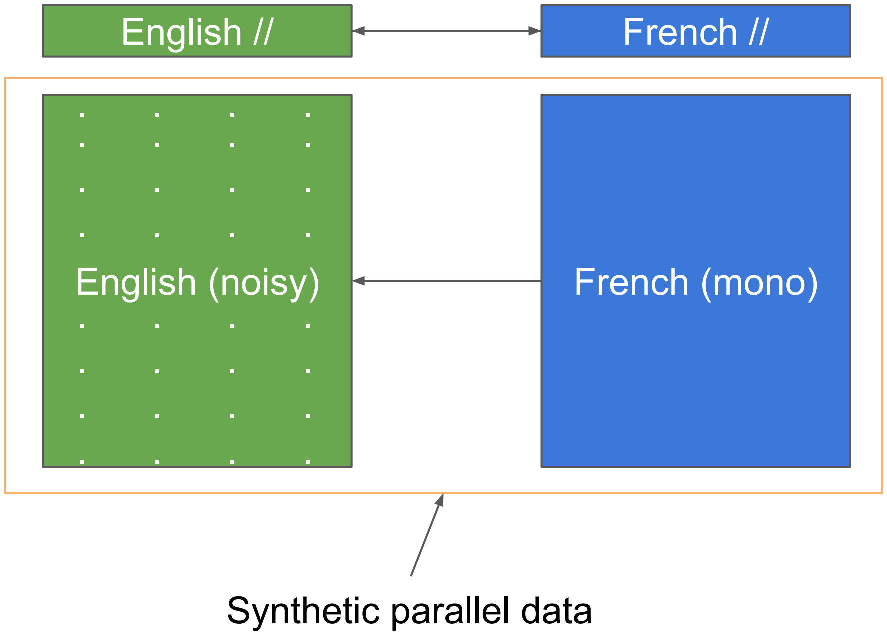
- Small parallel dataset
- Huge monolingual corpus in target language
- Train a (target $\rightarrow$ source) model $\mathbf{M}_{t2s}$
- Use $\mathbf{M}_{t2s}$ to translate target monolingual corpus



Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

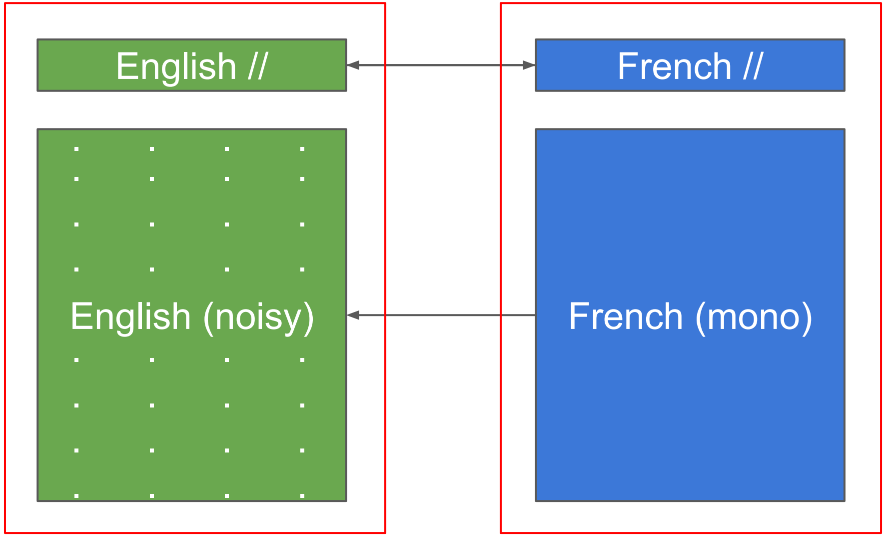
- Small parallel dataset
- Huge monolingual corpus in target language
- Train a (target → source) model $\mathbf{M}_{t2s}$
- Use $\mathbf{M}_{t2s}$ to translate target monolingual corpus



Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

- Small parallel dataset
- Huge monolingual corpus in target language
- Train a (target → source) model $\mathbf{M}_{t2s}$
- Use $\mathbf{M}_{t2s}$ to translate target monolingual corpus
- Use the two parallel datasets to train $\mathbf{M}_{s2t}$



Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

- en-->de WMT14
  - Parallel only: 20.4
  - + back-translation: 23.8
- en-->de WMT15
  - Parallel only: 23.6
  - + back-translation: 26.5

Back-translation (Sennrich et al. 2016)

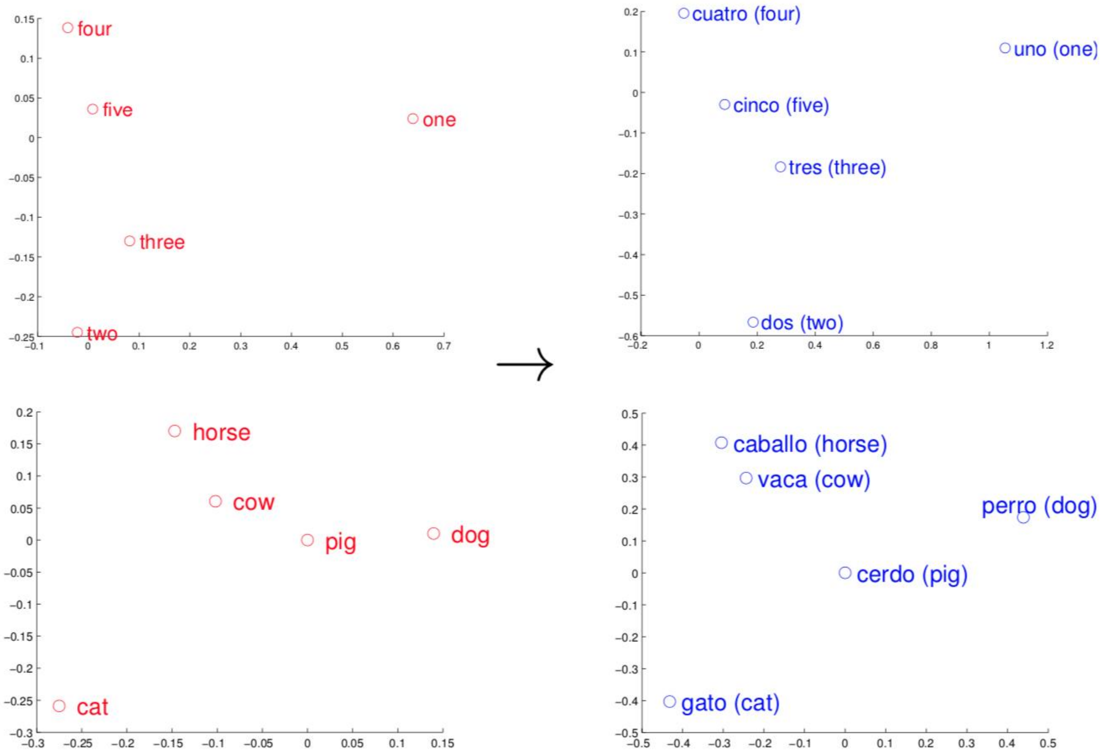
Improving Neural Machine Translation Models with Monolingual Data

- Back-translation can be used for
    - Semi-supervised machine translation
    - Style transfer
    - Domain transfer
    - (small parallel, large unlabeled data)
    - Unsupervised machine translation (later)

Outline

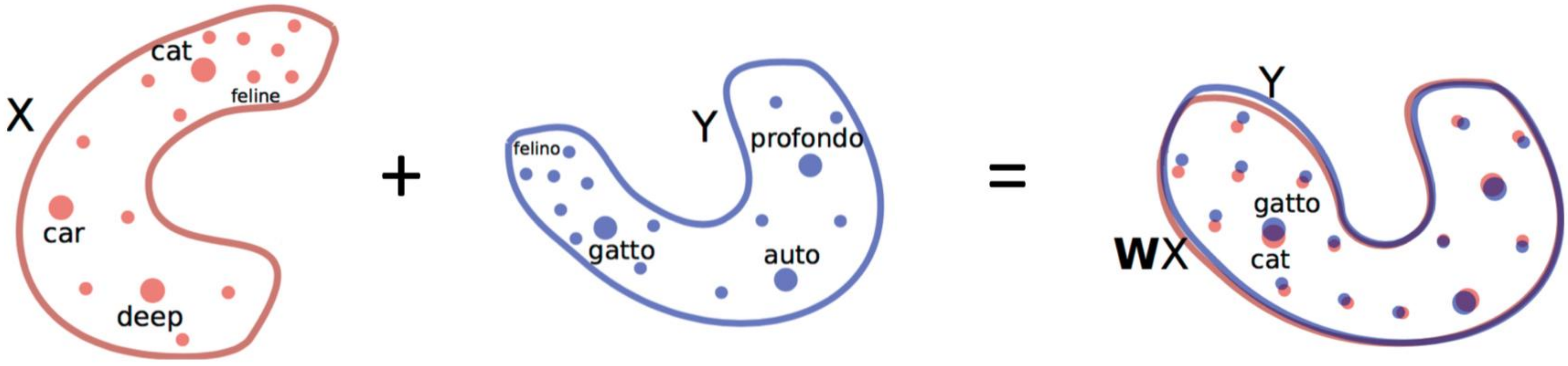
- Back-translation (Sennrich et al. 2016)
- **Unsupervised word translation (Conneau et al. 2018)**
- Unsupervised sentence translation (Lample et al. 2018)
- XLM (Lample & Conneau 2019)

Weakly-supervised word translation (Mikolov et al. 2013b)



- Left English, Right Spanish
- Projected down to 2 dimensions + manually rotated
- Similar geometric arrangements between languages even for distant language pair such as English <-> Vietnamese.

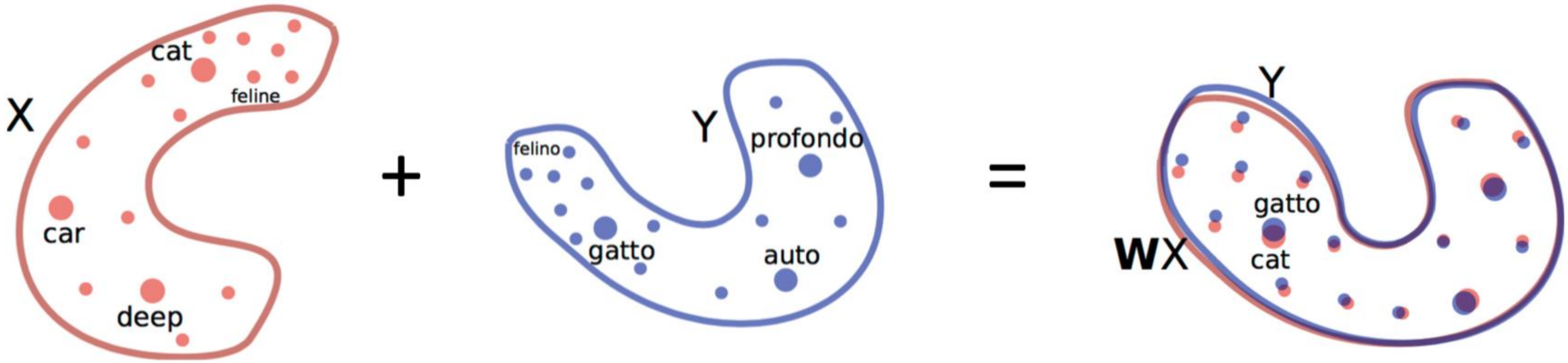
Weakly-supervised word translation (Mikolov et al. 2013b)



X: source embeddings
Y: target embeddings
W: linear transformation matrix
WX: projected embeddings

$$W^\star = \underset{W \in M_d(\mathbb{R})}{\operatorname{argmin}} \|WX - Y\|_{\mathrm{F}}$$

Weakly-supervised word translation (improved)

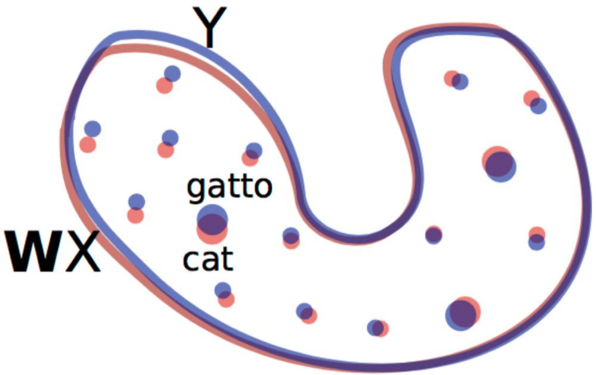


- Orthogonal projection – Xing et al. (2015), Smith et al. (2017) – **Procrustes**

  Add orthogonal constraint to W
  Train time: still need anchor word pairs to compute W*
  Can it be done without seed word pairs?

Unsupervised word translation (Conneau et al. 2018)

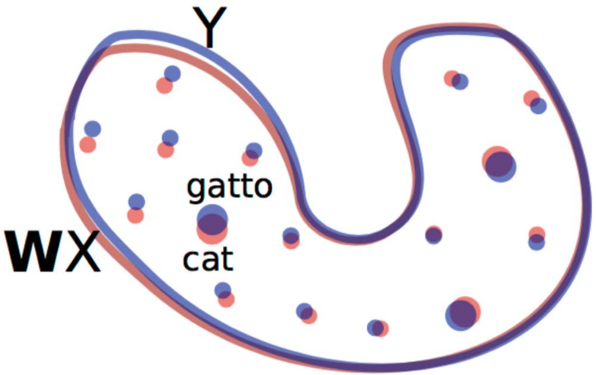
- Adversarial training
  - **W**X and Y should be indistinguishable
  - Train a discriminator D to distinguish **W**X and Y



$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}(\text{source}=1|Wx_i) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}(\text{source}=0|y_i)$$

Unsupervised word translation (Conneau et al. 2018)

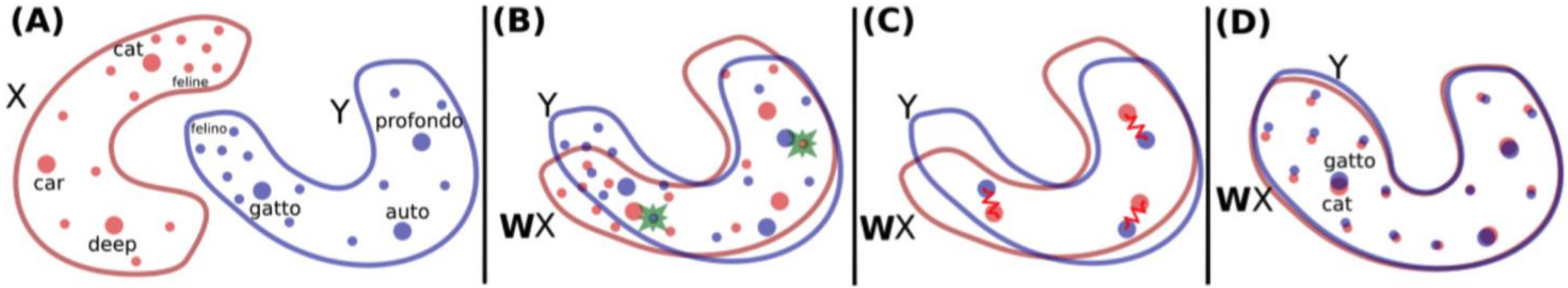
- Adversarial training
  - **W**X and Y should be indistinguishable
  - Train a discriminator D to distinguish **W**X and Y
  - Train **W** to fool the discriminator to make wrong predictions



$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}(\text{source}=1|Wx_i) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}(\text{source}=0|y_i)$$

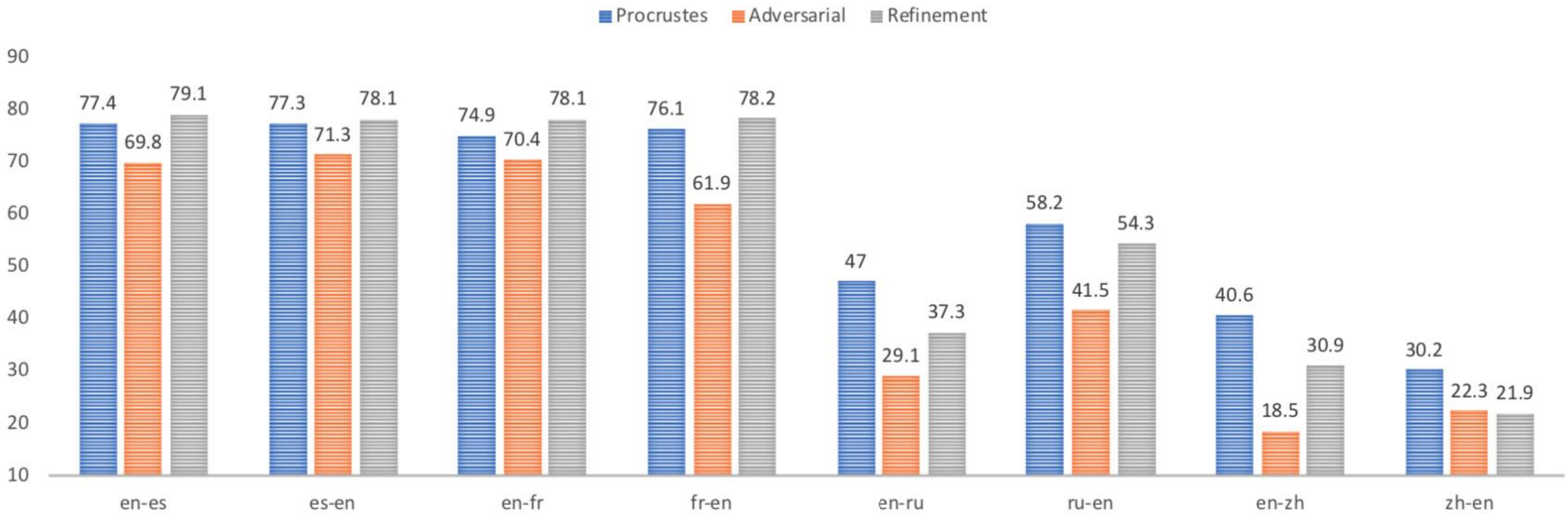
$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}(\text{source}=0|Wx_i) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}(\text{source}=1|y_i)$$

Unsupervised word translation (Conneau et al. 2018)



- **(A)** Train monolingual word embeddings
- **(B)** Align them using adversarial training
- Refinement step
  - **(C)** Select high-confidence translation pairs
  - **(D)** Apply Procrustes on the generated dictionary
- **Generate translations**

Unsupervised word translation (Conneau et al. 2018)



Word translation retrieval – P@1 – Adversarial + Refinement

1.5k source queries, 200k target keys (vocabulary of 200k words for all languages)

Outline

- Back-translation (Sennrich et al. 2016)
- Unsupervised word translation (Conneau et al. 2018)
- **Unsupervised sentence translation (Lample et al. 2018)**
- XLM (Lample & Conneau 2019)

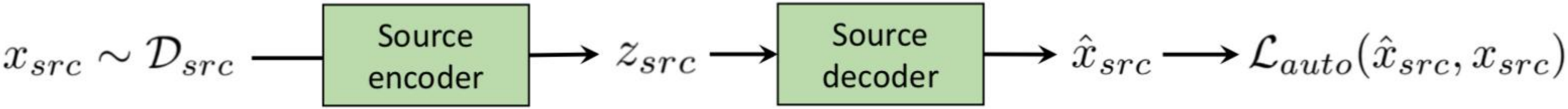
Unsupervised sentence translation (Lample et al. 2018)

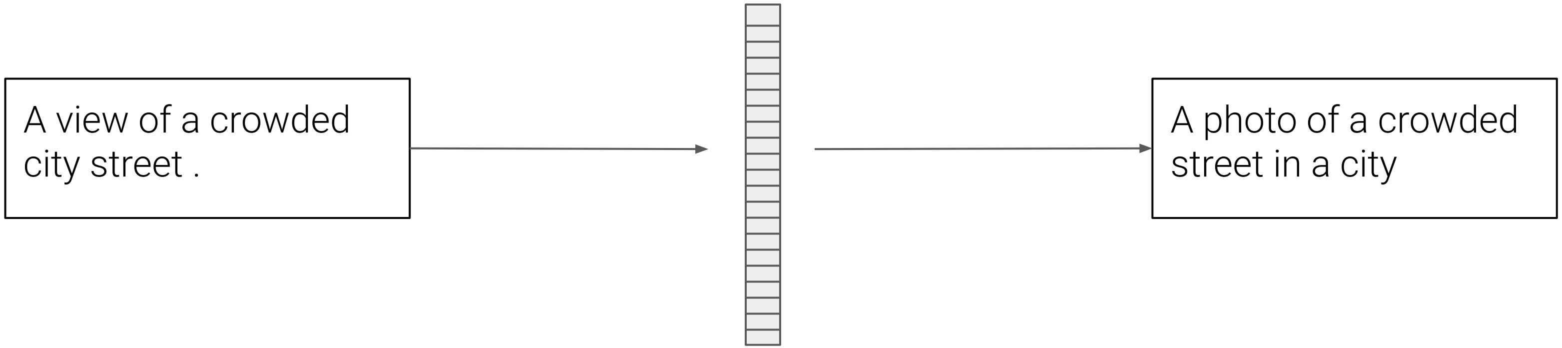
- Important components:
  - Bilingual dictionary (unsupervised)
  - Denoising autoencoding
  - Back-translation

Unsupervised sentence translation (Lample et al. 2018)

- Denoising autoencoding (DAE)
  - Aim to learn how sentences should be read in certain language

Traditional autoencoding: minimize reconstruction error



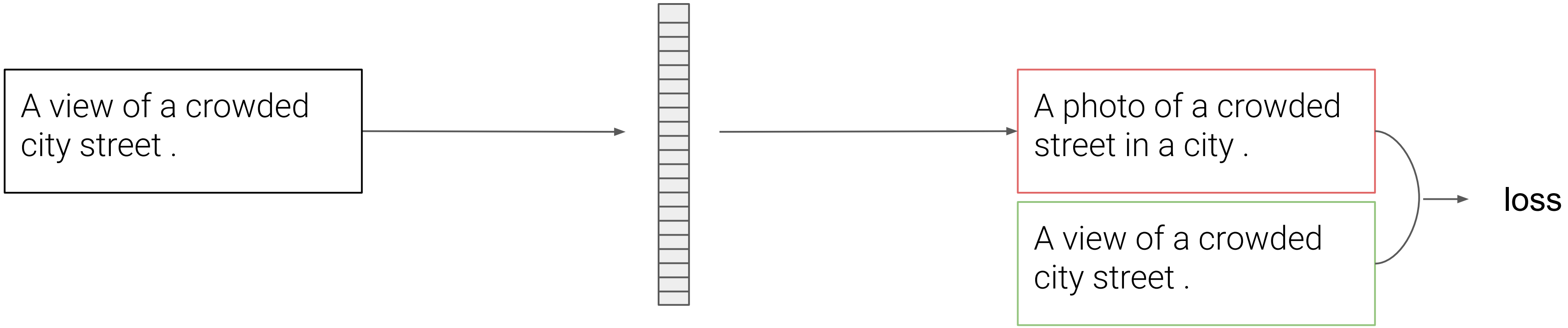


Unsupervised sentence translation (Lampie et al. 2018)

- Denoising autoencoding (DAE)
  - Aim to learn how sentences should be read in certain language

Traditional autoencoding: minimize reconstruction error

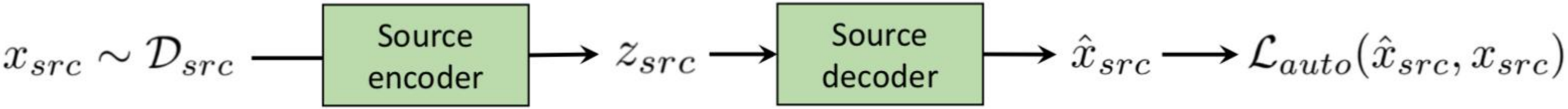
$$x_{src} \sim \mathcal{D}_{src} \longrightarrow \boxed{\text{Source encoder}} \longrightarrow z_{src} \longrightarrow \boxed{\text{Source decoder}} \longrightarrow \hat{x}_{src} \longrightarrow \mathcal{L}_{auto}(\hat{x}_{src}, x_{src})$$



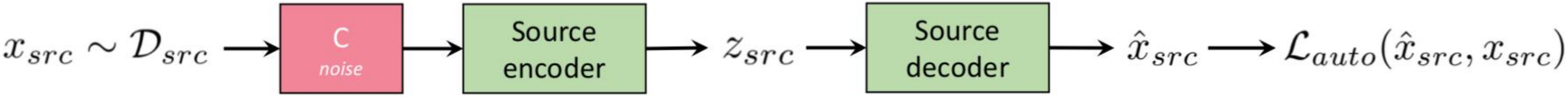
Unsupervised sentence translation (Lample et al. 2018)

- Denoising autoencoding (DAE)
  - Aim to learn how sentences should be read in certain language

Traditional autoencoding: minimize reconstruction error



Denoising autoencoding: add noise to original input to prevent it degenerating to copy-paste



Unsupervised sentence translation (Lample et al. 2018)

- Denoising autoencoding (DAE)
    - Aim to learn how sentences should be read in certain language
    - Noise type:
        - Word dropout: each word is removed with a probability ***p*** (usually 0.1)

Try to reconstruct

Ref: *Arizona was the first to introduce such a requirement .*

encode

→ Arizona was the first to such a requirement .

→ Arizona was first to introduce such a requirement .

Unsupervised sentence translation (Lample et al. 2018)

- Denoising autoencoding (DAE)
  - Aim to learn how sentences should be read in certain language
  - Noise type:
    - Word dropout: each word is removed with a probability ***p*** (usually 0.1)
    - Word shuffle: slightly shuffle words in a sentence

Try to reconstruct

Ref: *Arizona was the first to introduce such a requirement .*

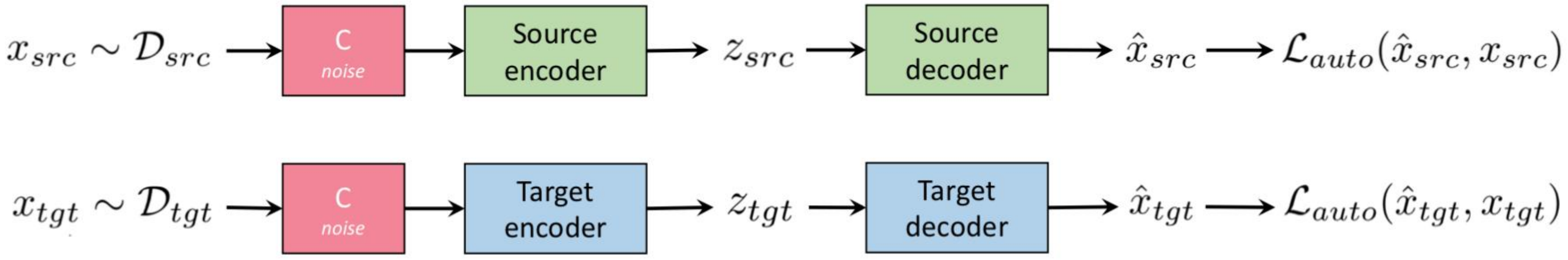
Arizona the first was to introduce a requirement such.

encode

Arizona was the to introduce first such requirement a .

Unsupervised sentence translation (Lample et al. 2018)

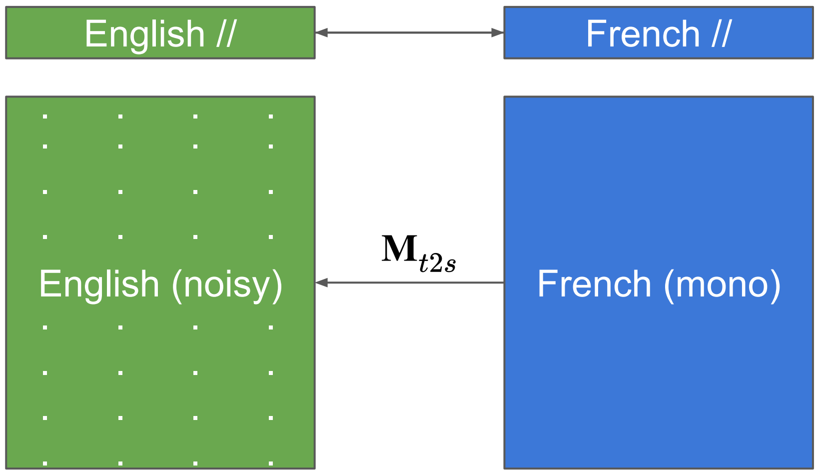
- Denoising autoencoding (DAE)
  - Aim to learn how sentences should be read in certain language
  - DAE for both source and target language



Back-translation (Sennrich et al. 2016)

Improving Neural Machine Translation Models with Monolingual Data

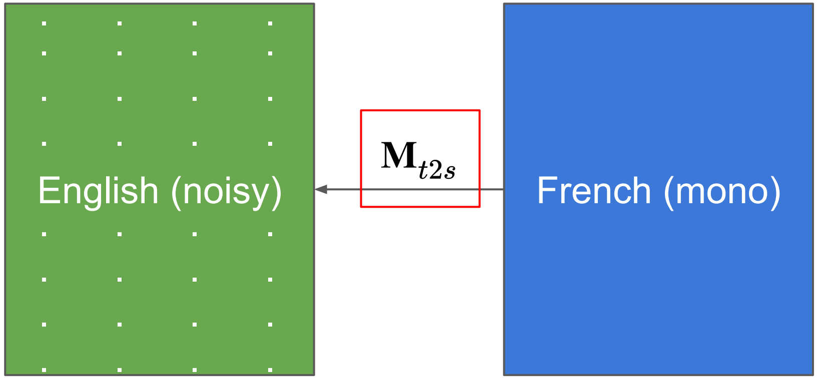
- Small parallel dataset
- Huge monolingual corpus in target language
- Train a (target → source) model $\mathbf{M}_{t2s}$
- Use $\mathbf{M}_{t2s}$ to translate target monolingual corpus



Back-translation (Sennrich et al. 2016)

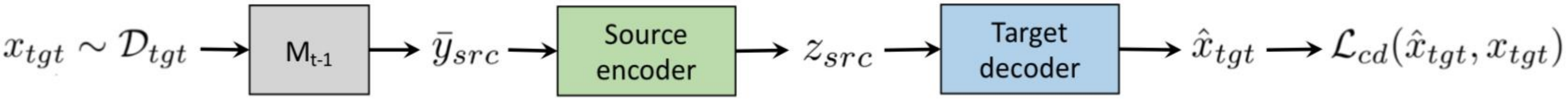
Improving Neural Machine Translation Models with Monolingual Data

- Small parallel dataset
- Huge monolingual corpus in target language
- Train a (target $\rightarrow$ source) model $\mathbf{M}_{t2s}$
- Use $\mathbf{M}_{t2s}$ to translate target monolingual corpus



Unsupervised sentence translation (Lample et al. 2018)

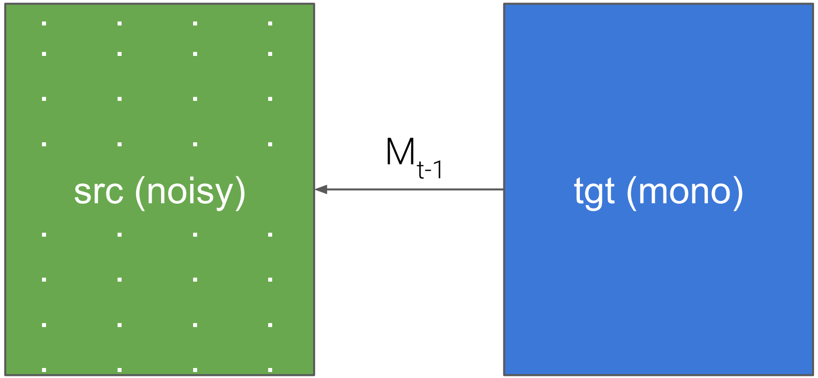
- **Iterative** back-translation



M_0: word-to-word translation model

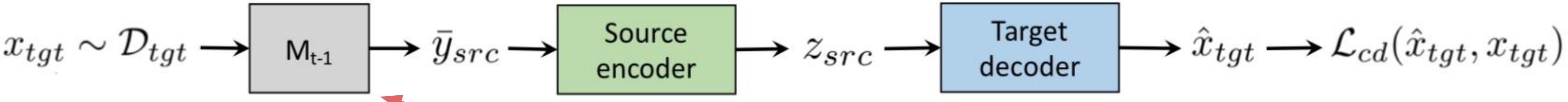
$P_{src\rightarrow tgt}^{(0)}$ and $P_{tgt\rightarrow src}^{(0)}$

M_{t-1}: Previous translation model $P_{tgt\rightarrow src}^{(t-1)}$



Unsupervised sentence translation (Lample et al. 2018)

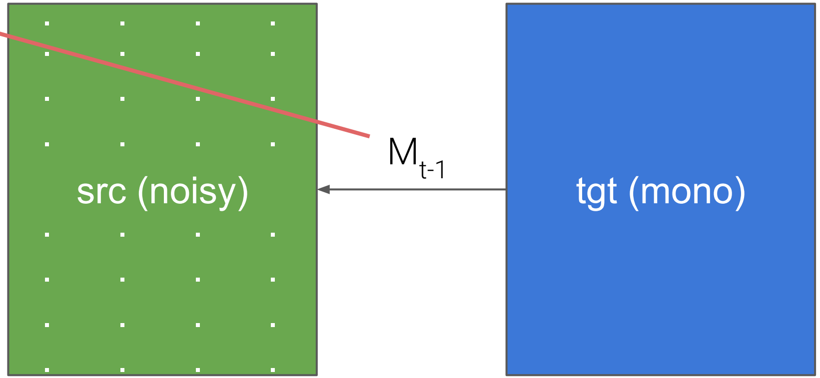
- Iterative back-translation



M_0: word-to-word translation model

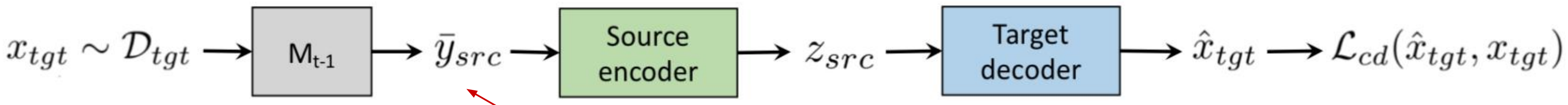
$P_{src \rightarrow tgt}^{(0)}$ and $P_{tgt \rightarrow src}^{(0)}$

M_{t-1}: Previous translation model $P_{tgt \rightarrow src}^{(t-1)}$



Unsupervised sentence translation (Lample et al. 2018)

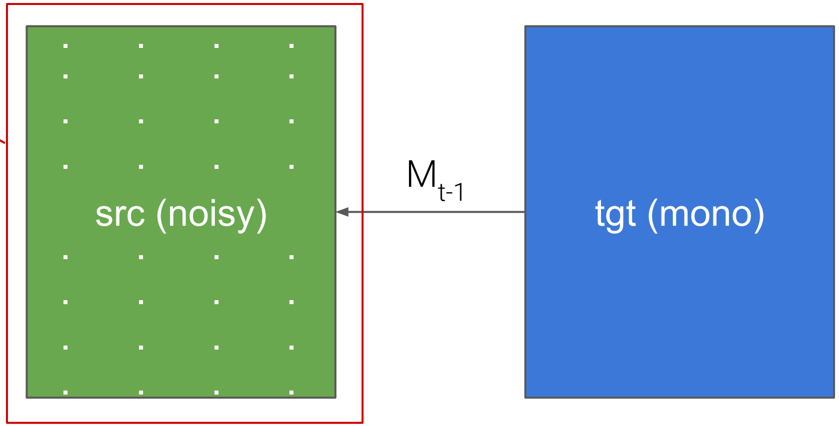
- Iterative back-translation



M_0: word-to-word translation model

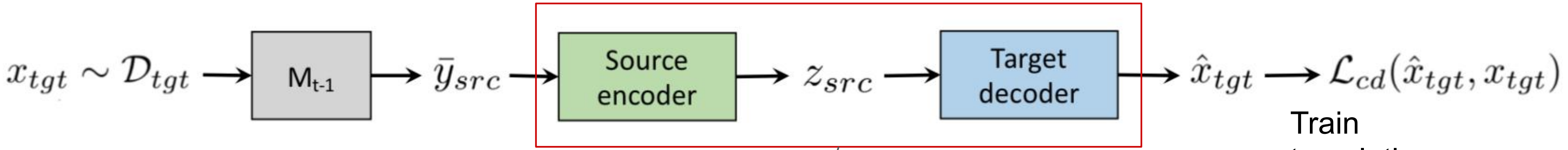
$P_{src\to tgt}^{(0)}$ and $P_{tgt\to src}^{(0)}$

M_{t-1}: Previous translation model $P_{tgt\to src}^{(t-1)}$



Unsupervised sentence translation (Lample et al. 2018)

- Iterative back-translation

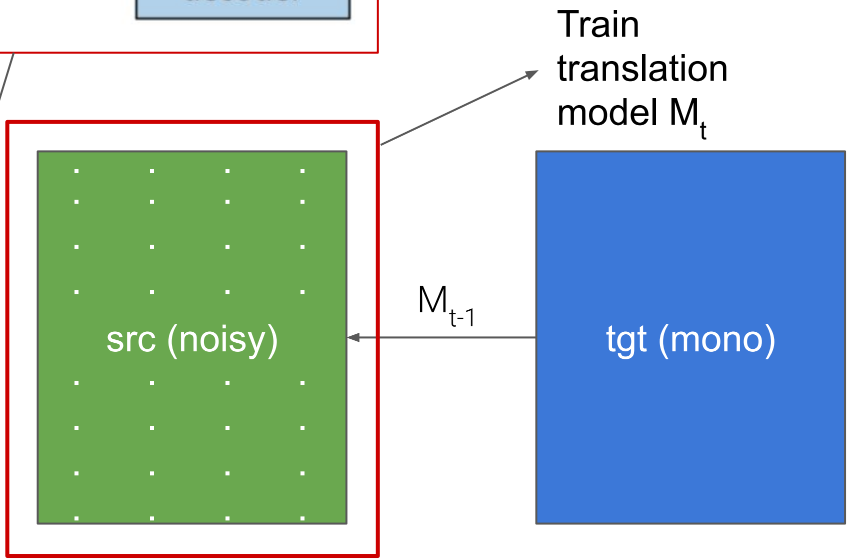


M_0: word-to-word translation model

$P^{(0)}_{src \to tgt}$ and $P^{(0)}_{tgt \to src}$

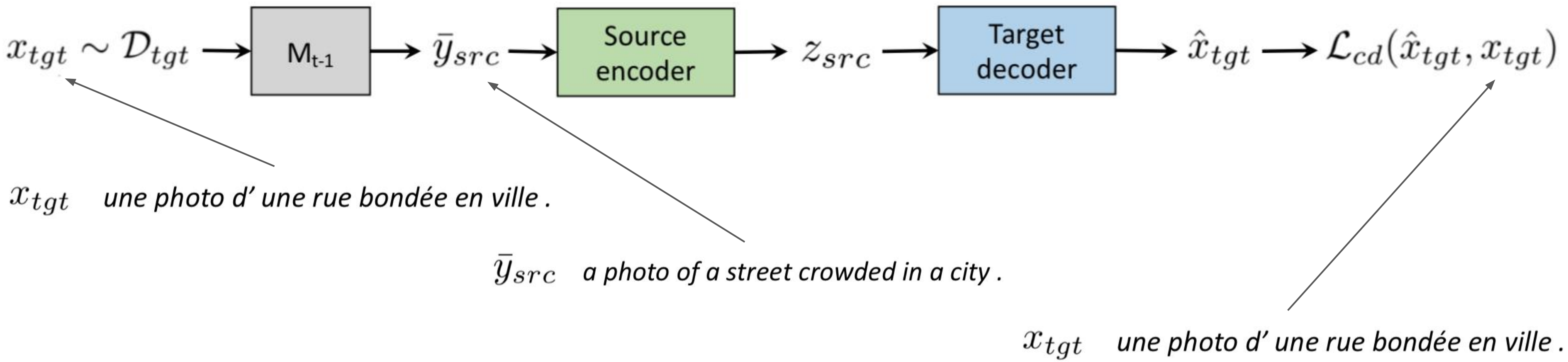
M_{t-1}: Previous translation model $P^{(t-1)}_{tgt \to src}$

Train on noisy parallel data $\{y_{src}, x_{tgt}\} \Rightarrow$ get $P^{(t)}_{src \to tgt}$



Unsupervised sentence translation (Lample et al. 2018)

- Iterative back-translation

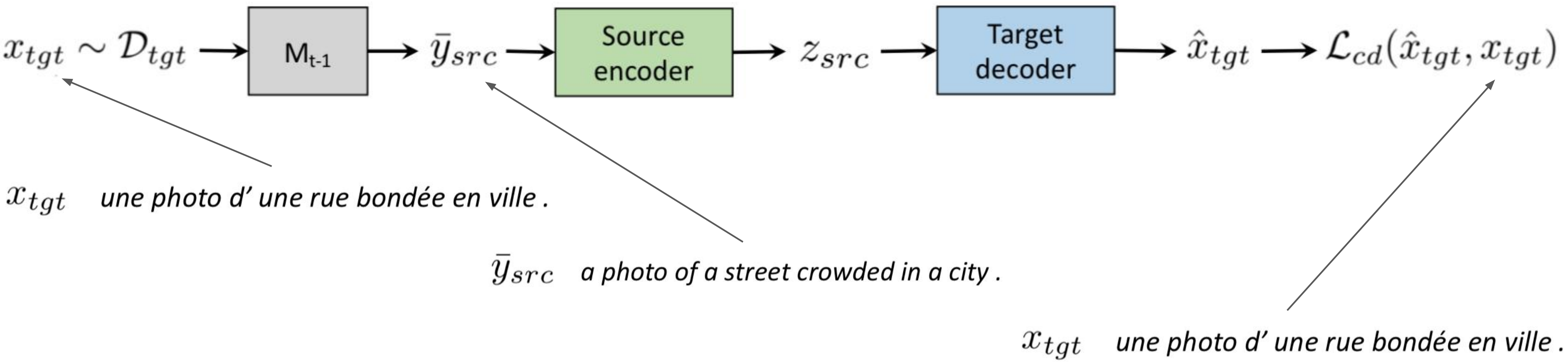


tgt → src → tgt

src → tgt → src

Unsupervised sentence translation (Lample et al. 2018)

- Iterative back-translation

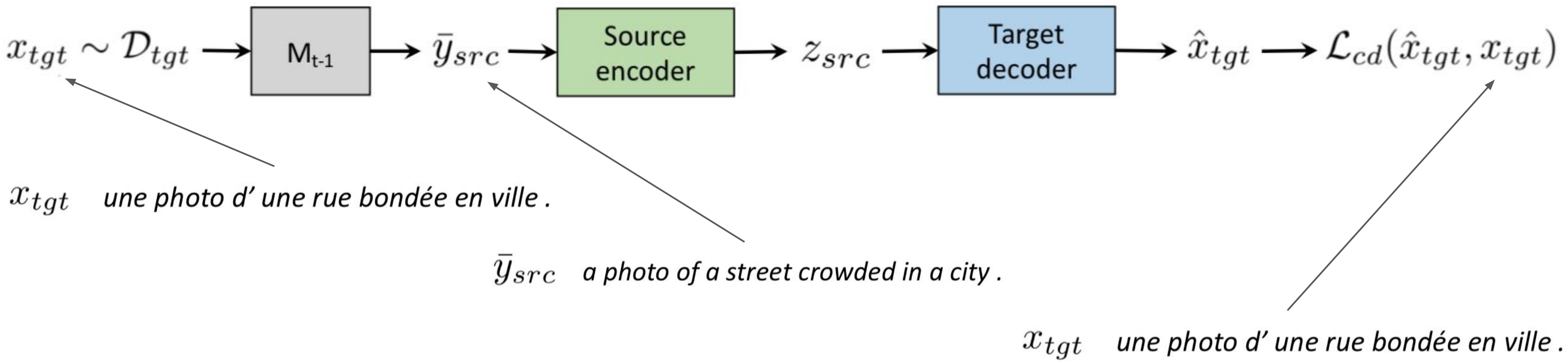


tgt ⟶ src ⟶ tgt

src ⟶ tgt ⟶ src

Unsupervised sentence translation (Lample et al. 2018)

- Iterative back-translation

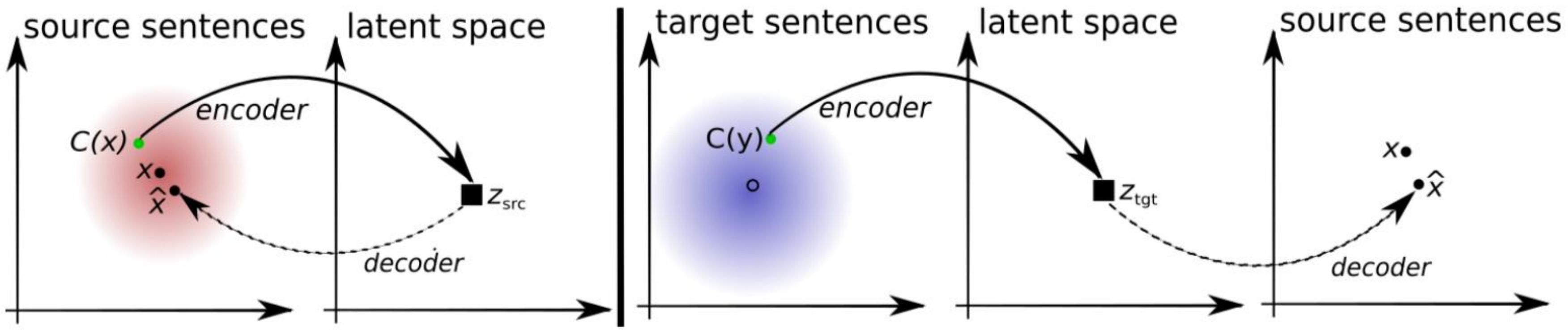


tgt → src → tgt

src → tgt → src

Unsupervised sentence translation (Lample et al. 2018)

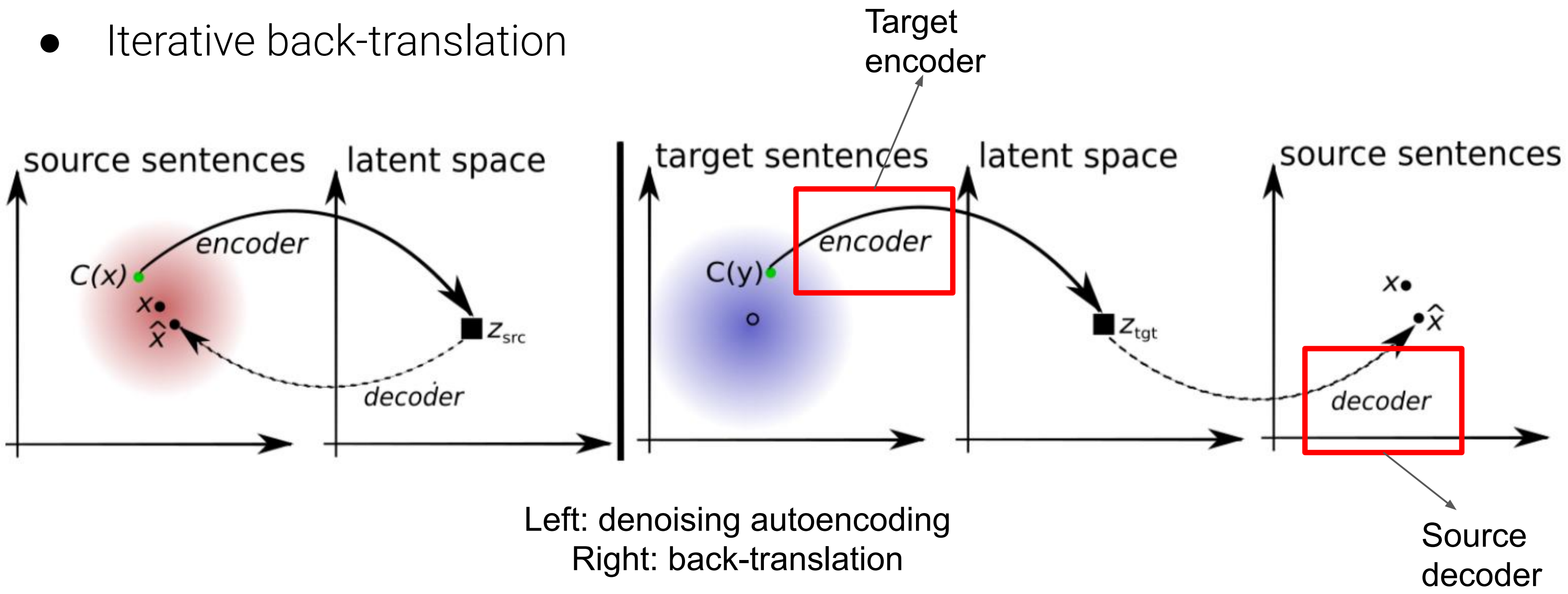
- Denoising autoencoding and Iterative back-translation



Left: denoising autoencoding
Right: back-translation

Unsupervised sentence translation (Lample et al. 2018)

- Iterative back-translation



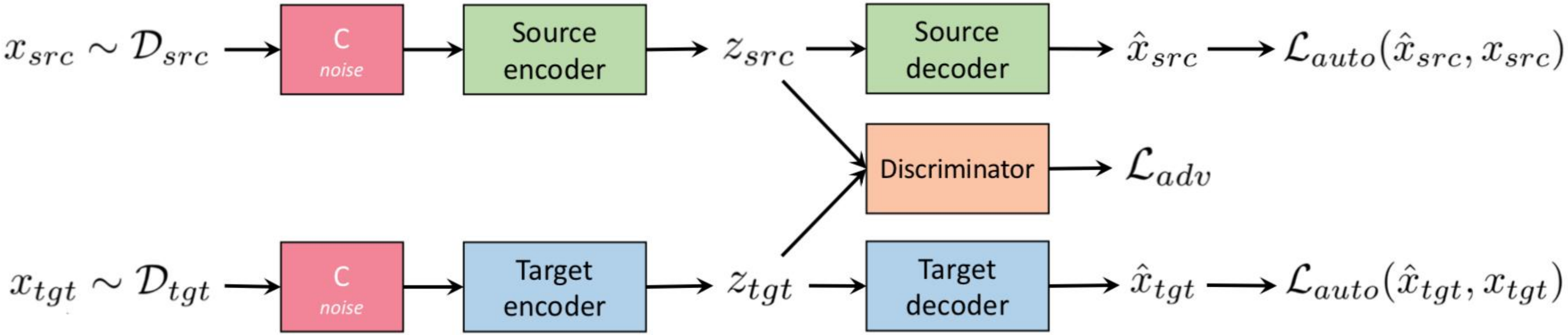
Left: denoising autoencoding
Right: back-translation

Source decoder decode target hidden states

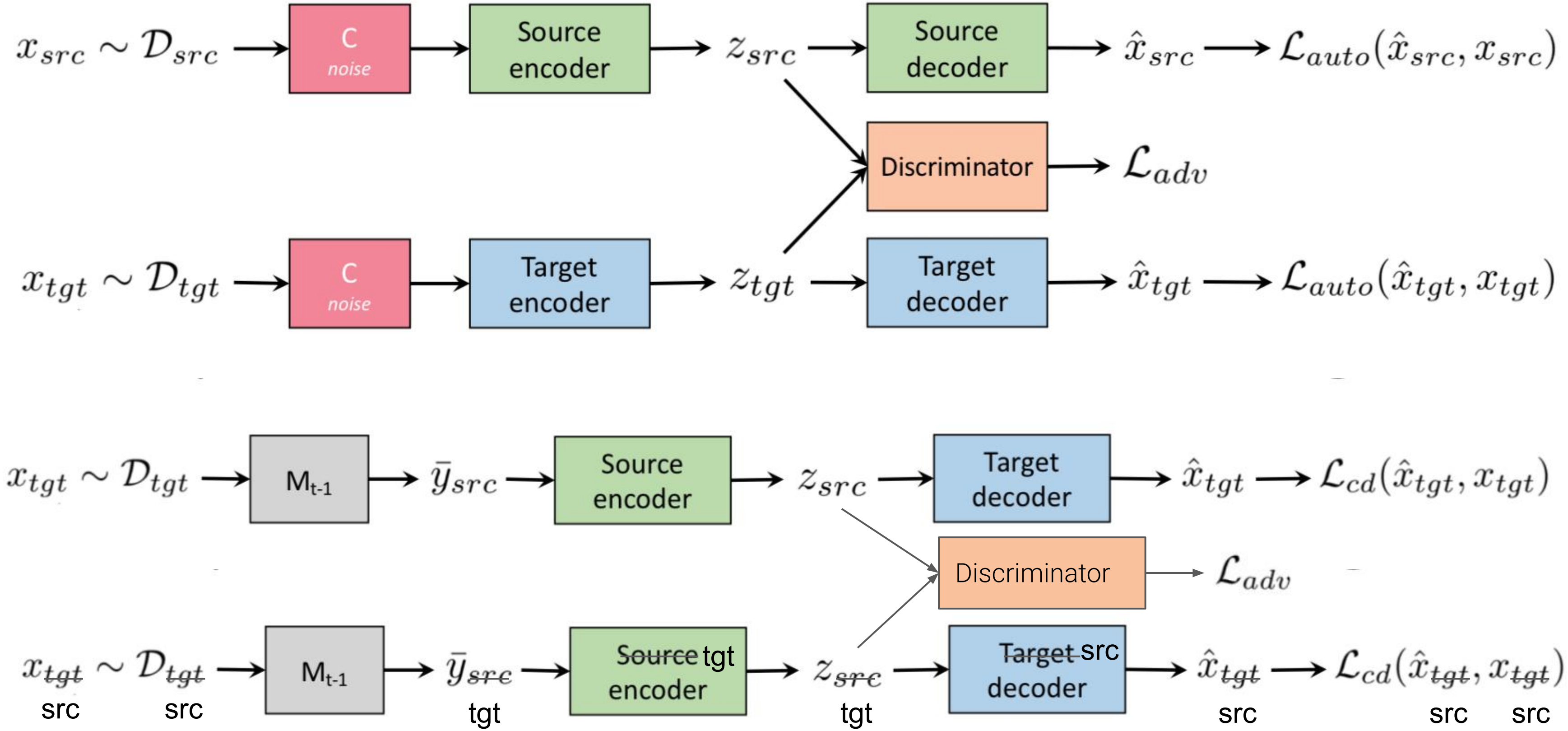
Adversarial training: make z_tgt and z_src indistinguishable

Unsupervised sentence translation (Lample et al. 2018)

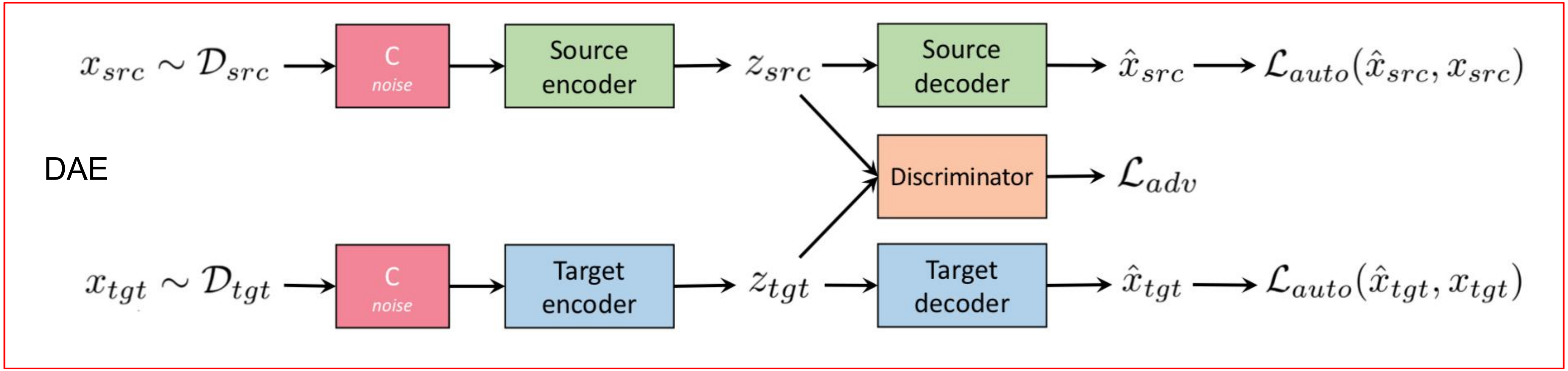
- Adversarial Training
  - Make the hidden states of source and target languages indistinguishable
  - Discriminator target: correctly predict the language given a sequence of hidden states **z**

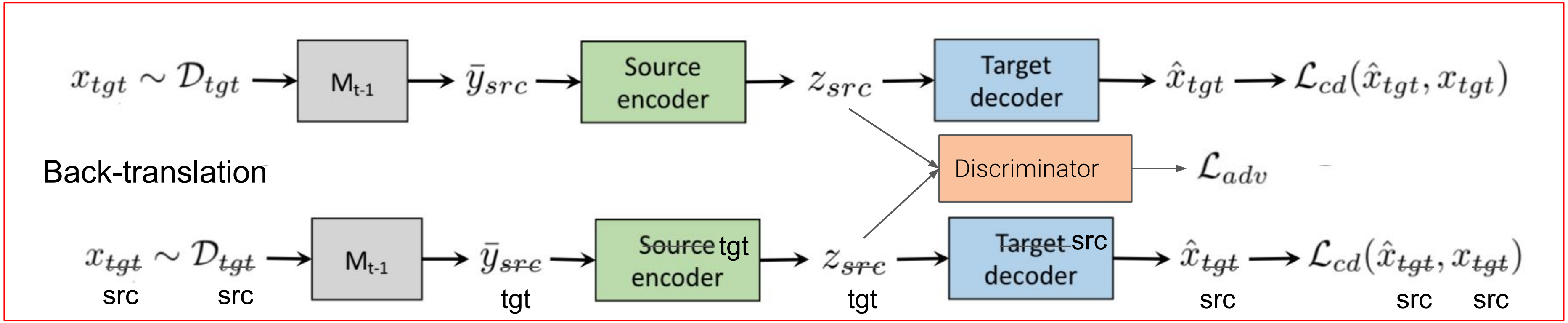


Unsupervised sentence translation (Lample et al. 2018)

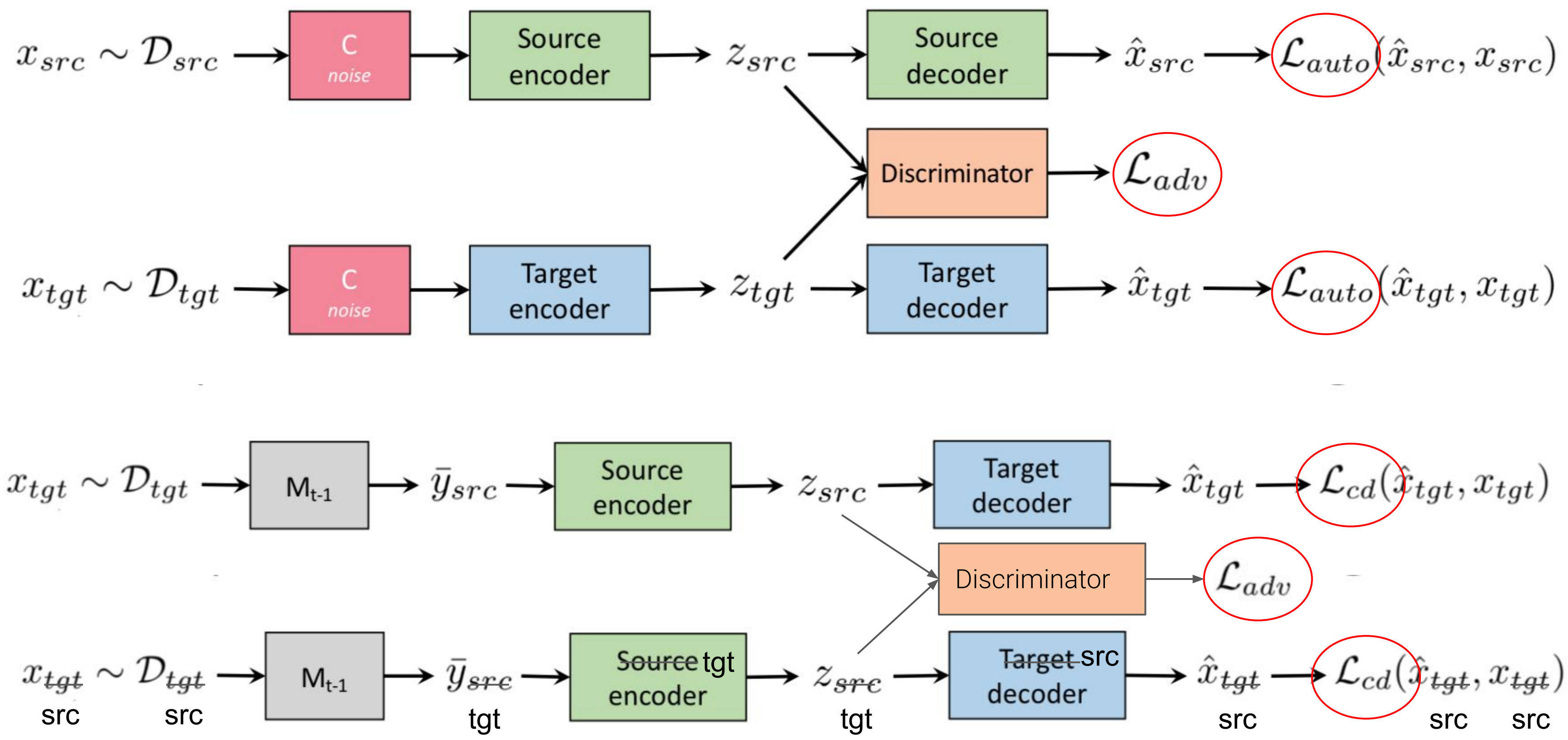


Unsupervised sentence translation (Lample et al. 2018)





Unsupervised sentence translation (Lample et al. 2018)



Unsupervised sentence translation (Lample et al. 2018)

- Training objective
  - $$\mathcal{L} = \lambda_{auto}\mathcal{L}_{auto}(src) + \lambda_{auto}\mathcal{L}_{tgt}(tgt) + \lambda_{bt}\mathcal{L}_{bt}(src, tgt) + \lambda_{bt}\mathcal{L}(tgt, src) + \lambda_{adv}\mathcal{L}_{adv}(z)$$

- Training procedure:
  - Get word translation
  - Initial word-to-word translation model
  - For each iteration:
    - Denoising autoencoding step for {en, fr}
    - Back-translation step for {en-fr-en, fr-en-fr} using translation model from previous iteration

Unsupervised sentence translation (Lample et al. 2018a)

- Infer bilingual dictionary
  -

	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.

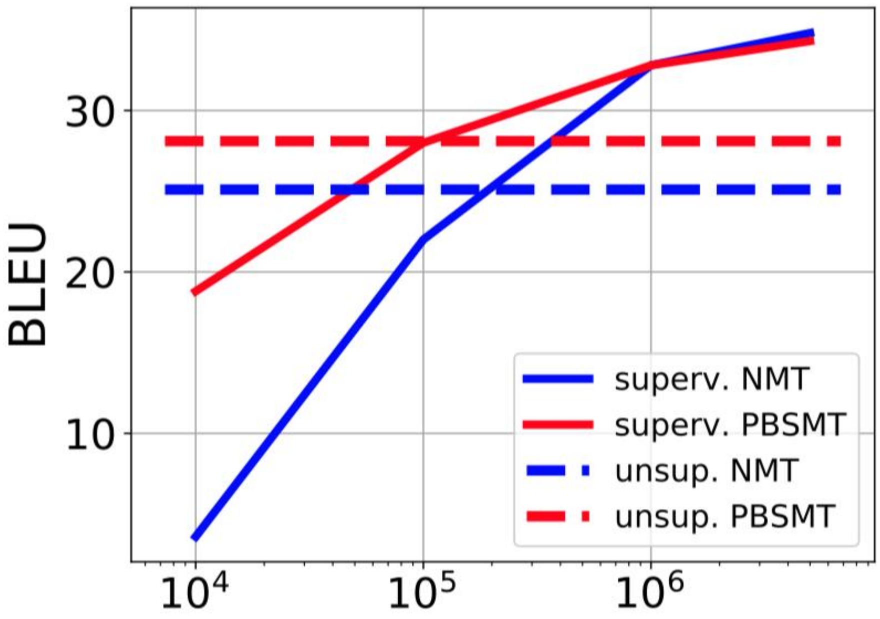
Unsupervised sentence translation (Lample et al. 2018b)

- Previously use bilingual dictionary learned in unsupervised way
- Change to cross-lingual bpe embedding
  - Jointly processing src+tgt corpora
  - Limitation: only for related languages that share a good fraction of BPE tokens

Model	en-fr	fr-en	de-en	en-de
(Artetxe et al., 2018)	15.1	15.6	-	-
(Lample et al., 2018)	15.0	14.3	13.3	9.6
NMT (LSTM)	24.5	23.7	19.6	14.7
NMT (Transformer)	25.1	24.2	21.0	17.2

Unsupervised sentence translation (Lample et al. 2018b)

- Use cross-lingual sub-word embedding
- Comparison with supervised MT



Outline

- Back-translation (Sennrich et al. 2016)
- Unsupervised word translation (Conneau et al. 2018)
- Unsupervised sentence translation (Lample et al. 2018)
- **XLM (Lample & Conneau 2019)**

XLM (Lample & Conneau 2019)

Cross-lingual language model pre-training

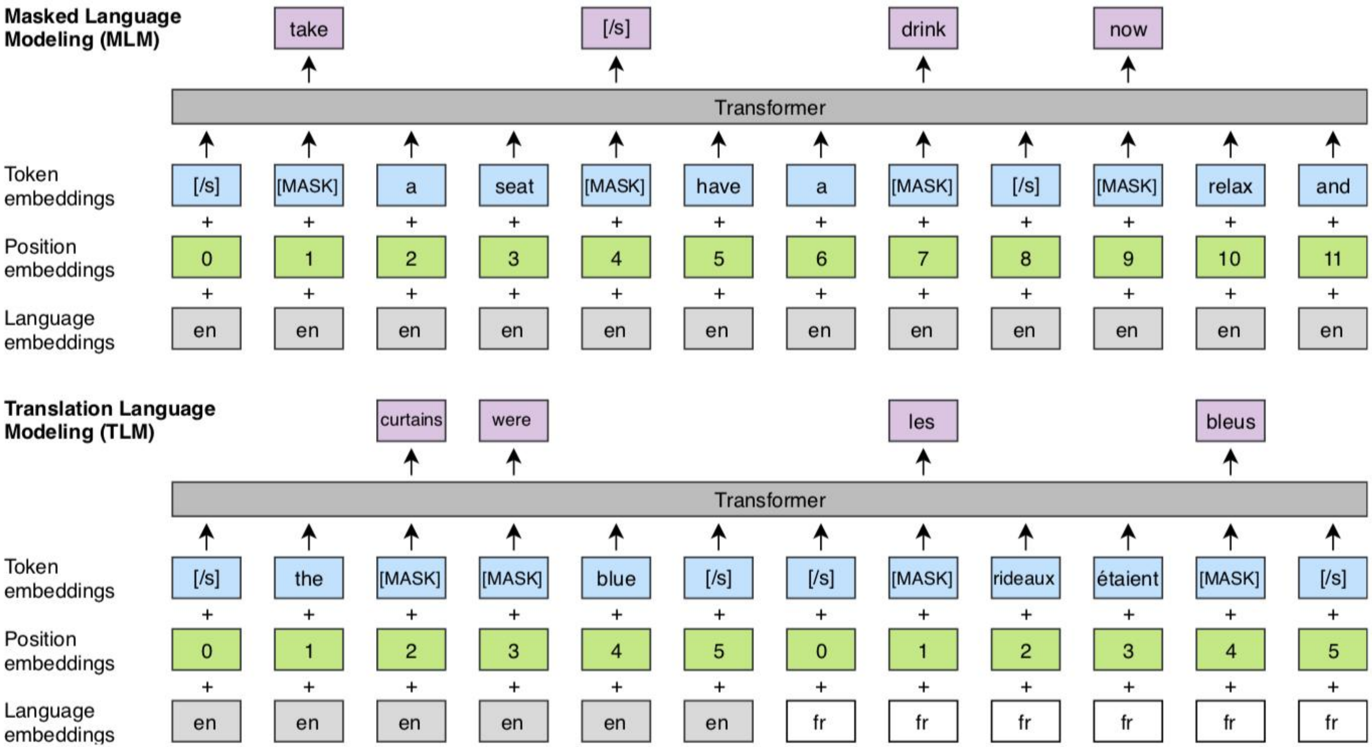
- Motivation:
    - Previous language model pre-training done only on English
    - Extend it to multiple languages
    - Cross-lingual language understanding benchmarks

XLM (Lample & Conneau 2019)

Cross-lingual language model pre-training

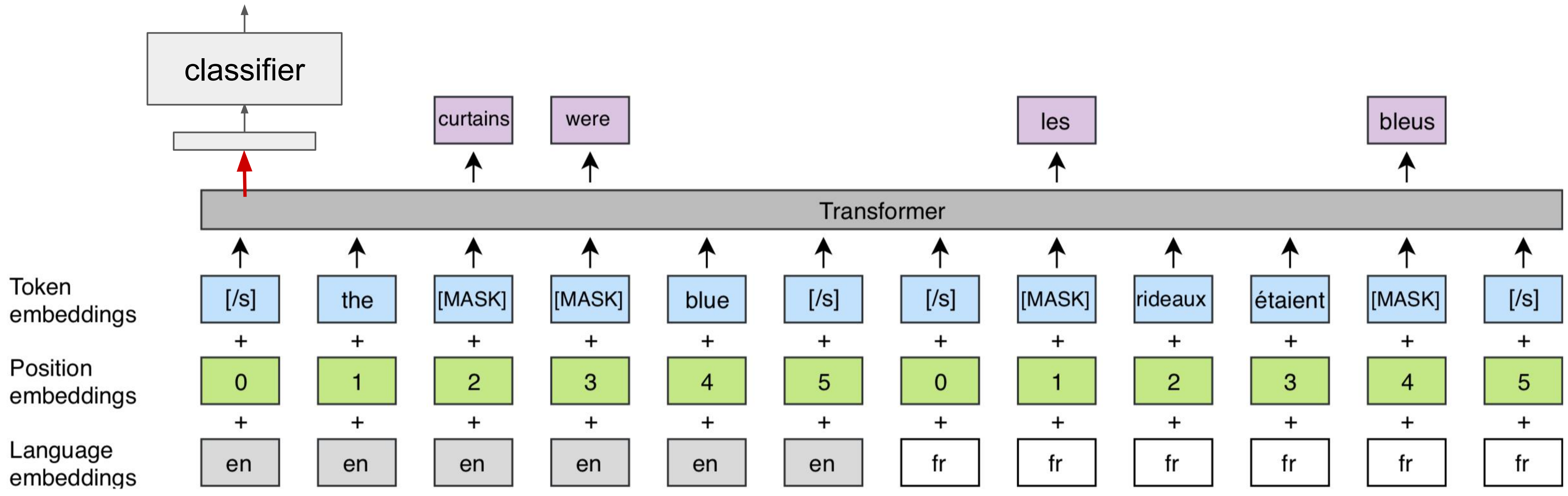
- 3 objectives
  - Causal LM: traditional left to right (monolingual)
  - Masked LM: cloze task on monolingual data (monolingual)
  - Translation LM: cloze task on parallel sentence pair (parallel)
- Pre-training data:
  - *N* languages, *N* monolingual corpora
  - Randomly sample sentences from these N corpora and concatenate to a new one
  - Learn Byte Pair Encoding (BPE) on this new corpus ⇒ shared sub-word vocabulary

XLM (Lample & Conneau 2019)



XLM (Lample & Conneau 2019)

- Better initialization of sentence encoders for cross-lingual classification
  - Add a linear classifier on top of first hidden state of last layer of each sentence, fine-tune on cross-lingual natural language inference (XNLI) dataset



XLM (Lample & Conneau 2019)

- Better initialization of sentence encoders for cross-lingual classification
  - Add a linear classifier on top of first hidden state of XLM, fine-tune on cross-lingual natural language inference (XNLI) dataset
- Better initialization of supervised and unsupervised NMT systems
  - Initialize translation models with the pre-trained sentence encoders

XLM (Lample & Conneau 2019)

		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

XLM (Lample & Conneau 2019)

- Better initialization of sentence encoders for cross-lingual classification
    - Add a linear classifier on top of first hidden state of XLM, fine-tune on cross-lingual natural language inference (XNLI) dataset
- Better initialization of supervised and unsupervised NMT systems
    - Initialize translation models with the pre-trained sentence encoders
- Language models for low-resource languages
    - Train low-resource language model with additional data of similar language
    - E.g.
        - Nepali (low-resource) only: PPL: 157.2
        - Add Hindi, PPL: 109.3

You should now be able to answer:

- What is back-translation? Why and when do we use it?
- Describe a way to find a bilingual dictionary given two monolingual corpora.
- What is a denoising autoencoder and why do we need to add noise?
- Describe a way to do unsupervised machine translation.
- What are the training objectives of XLM?
- How can we use XLM in other cross-lingual tasks?
- ...