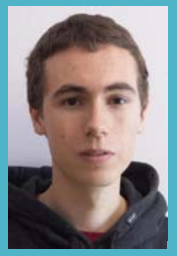
Cross-linguality and machine translation without bilingual data

Eneko Agirre
@eagirre
HiTZ NLP research center - http://hitz.eus
University of the Basque Country (UPV/EHU)




Joint work with: Mikel Artetxe, Gorka Labaka

HiTZ: Basque Center for Language Technology

HiTZ is a reference center on Language Technologies. Its aim is to promote research, training, technological transfer and innovation in this area. We are a multidisciplinary team: computer scientists, linguists and engineers.

Information Extraction and Information Retrieval

Machine Translation

Text Analysis

Speech Synthesis and Speech Recognition

Human-Computer Interaction

Speech and Language Resources

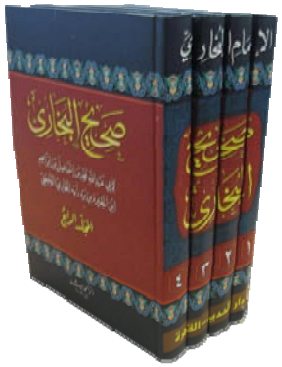
Medical and Legal domains

Digital humanities and education

http://hitz.eus

Overview

Arabic monolingual corpora



Human translations

Chinese monolingual corpora

Arabic representation

Chinese representation

Machine translation

Overview

Arabic monolingual corpora

No bilingual resource

Chinese monolingual corpora

Arabic representation

Chinese representation

Machine translation

Outline

- Initial idea: Bilingual embedding mappings
  - *Introduction embeddings*
  - *Bilingual embedding mappings (AAAI18)*
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions of bilingual embedding mappings*
- Unsupervised neural machine translation
  - *Introduction to NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Self-learning with better initializations (ACL19)*
  - *Conclusions*

Outline

- Initial idea: Bilingual embedding mappings
  - *Introduction embeddings*
  - *Bilingual embedding mappings (AAAI18)*
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions of bilingual embedding mappings*
- Unsupervised neural machine translation
  - *Introduction to NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Self-learning with better initializations (ACL19)*
  - *Conclusions*

Introduction to embeddings

Introduction to embeddings

Geographical space

- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

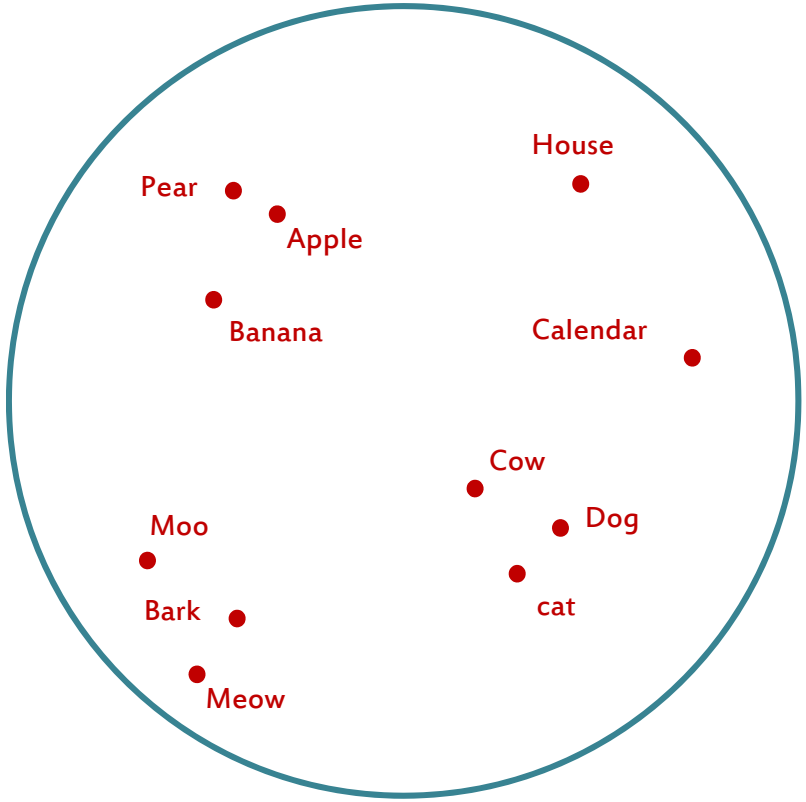
Introduction to embeddings

Geographical space

- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

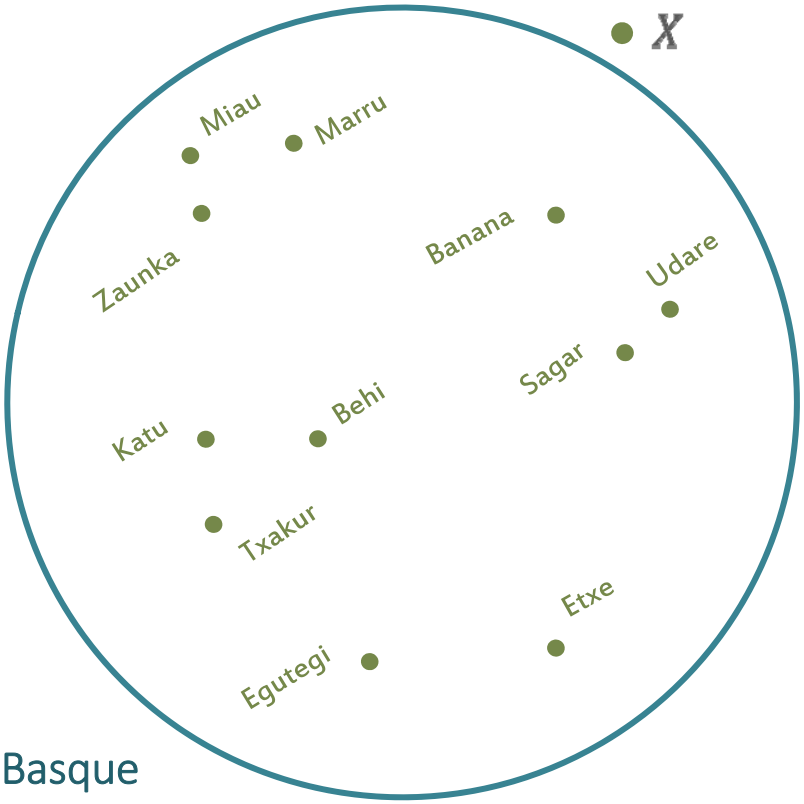
Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Neural networks / linear algebra from co-occurrence counts



Introduction to embedding mappings

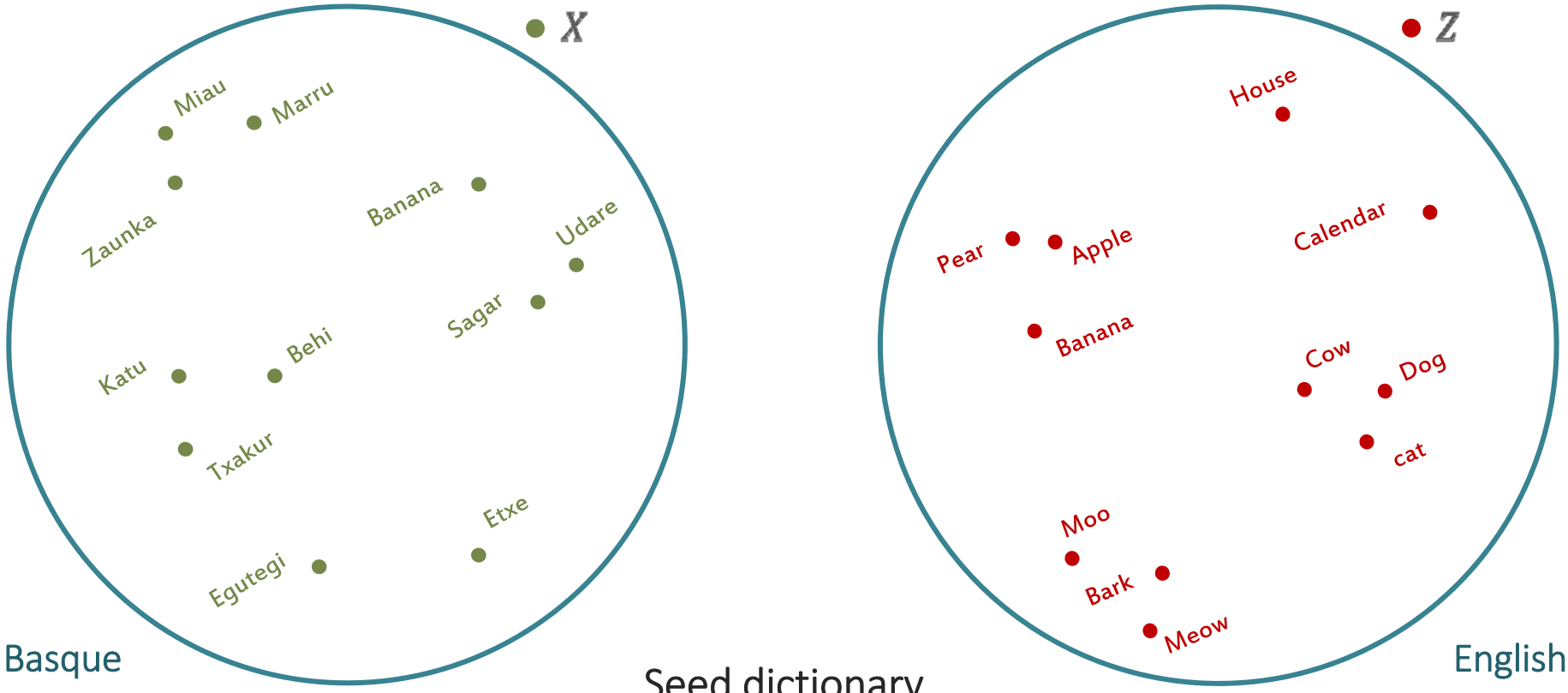
Introduction to embedding mappings



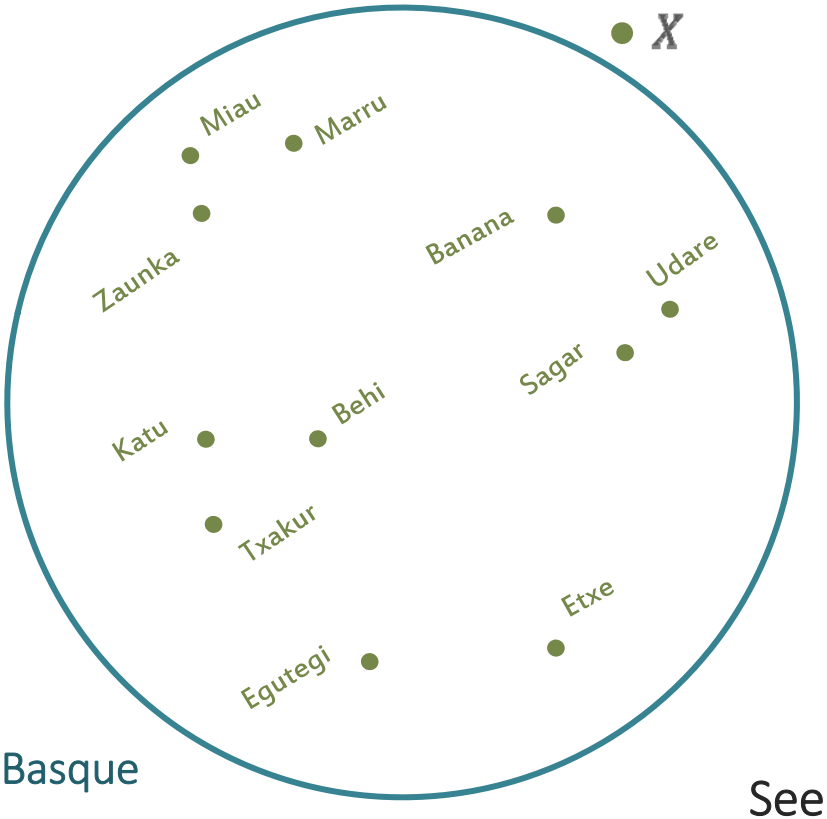
Basque

English

Introduction to embedding mappings



Introduction to embedding mappings



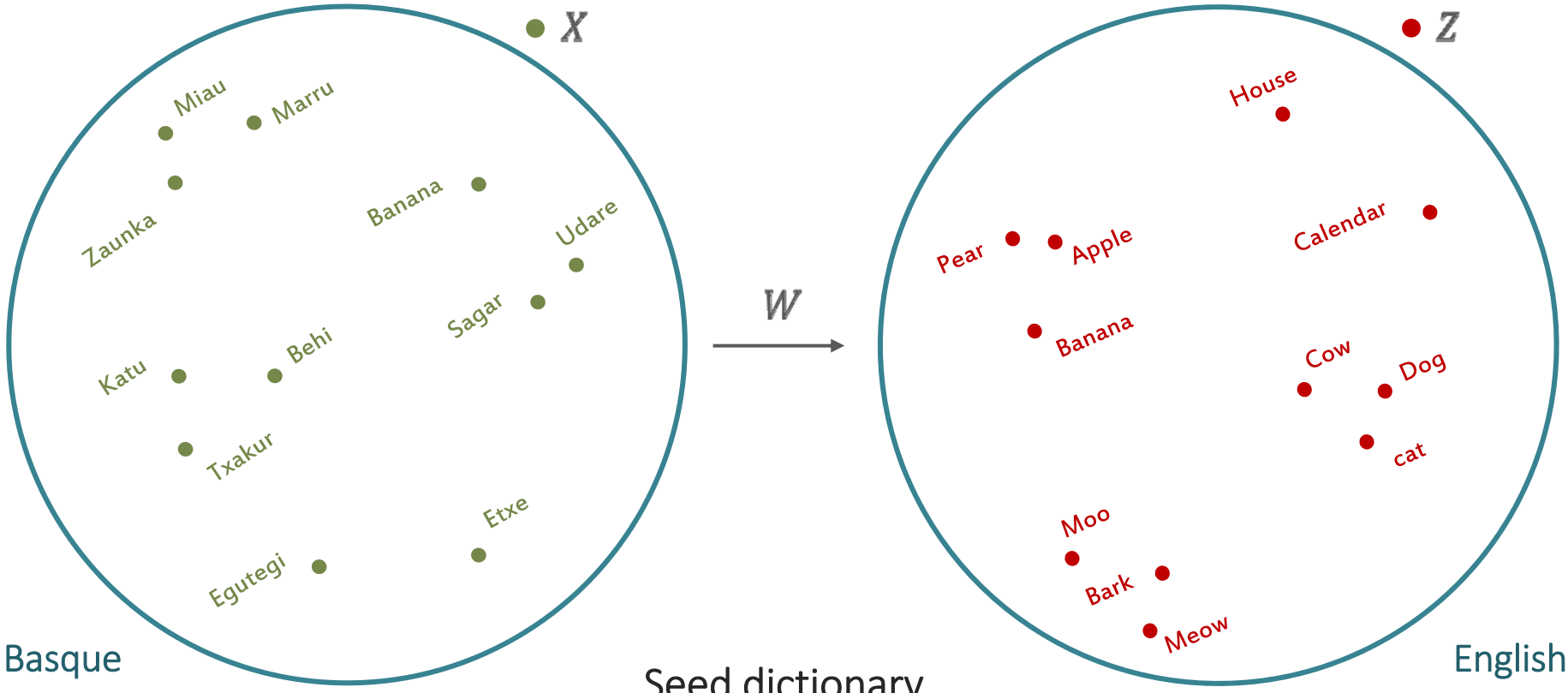
Basque

English

Seed dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

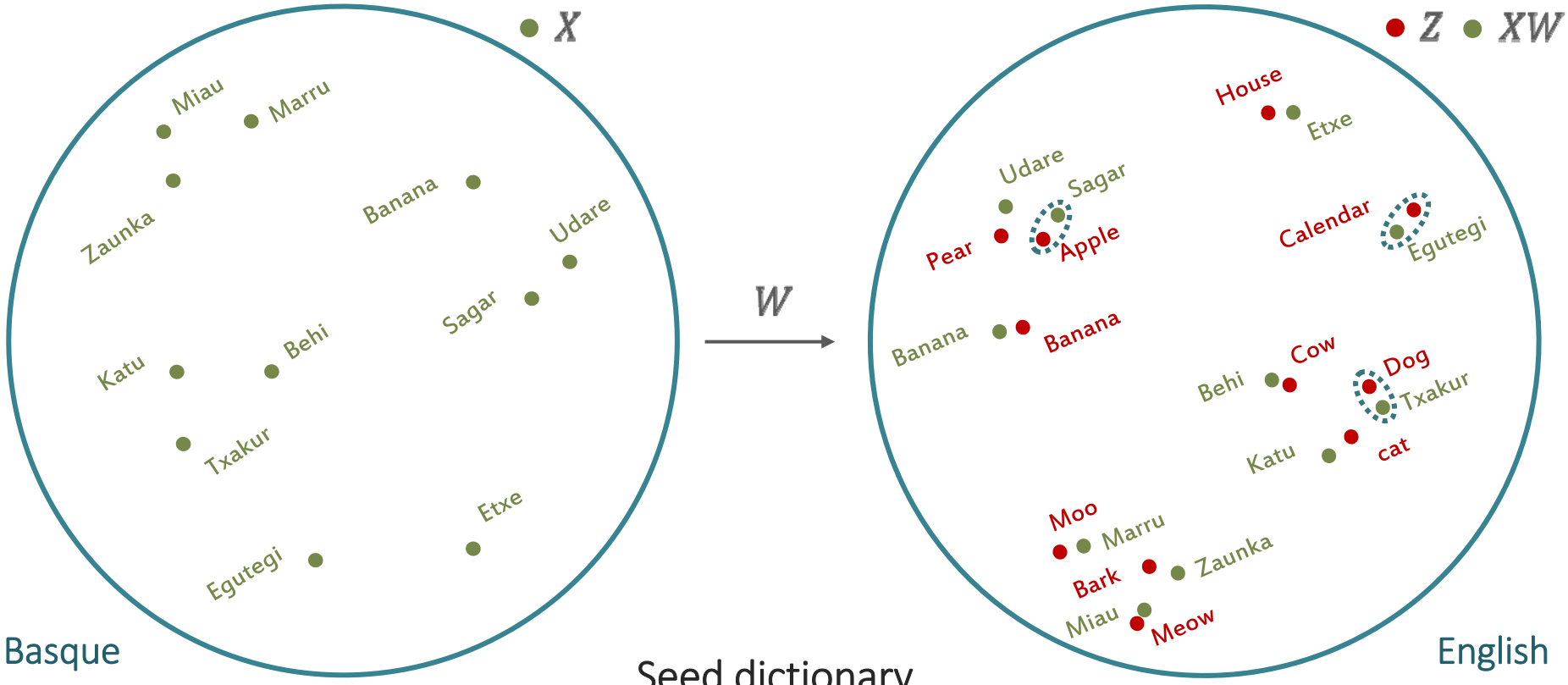
Introduction to embedding mappings



Seed dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

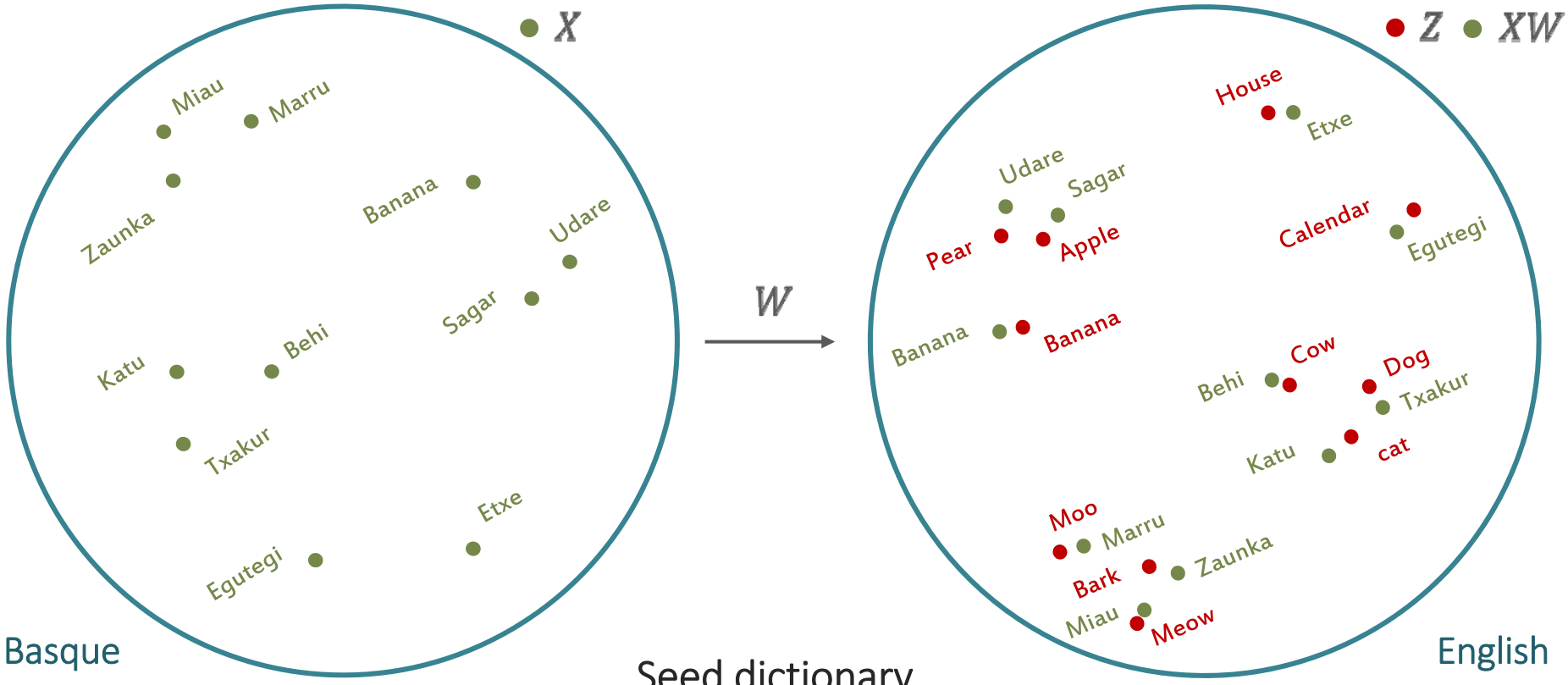
Introduction to embedding mappings



Seed dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

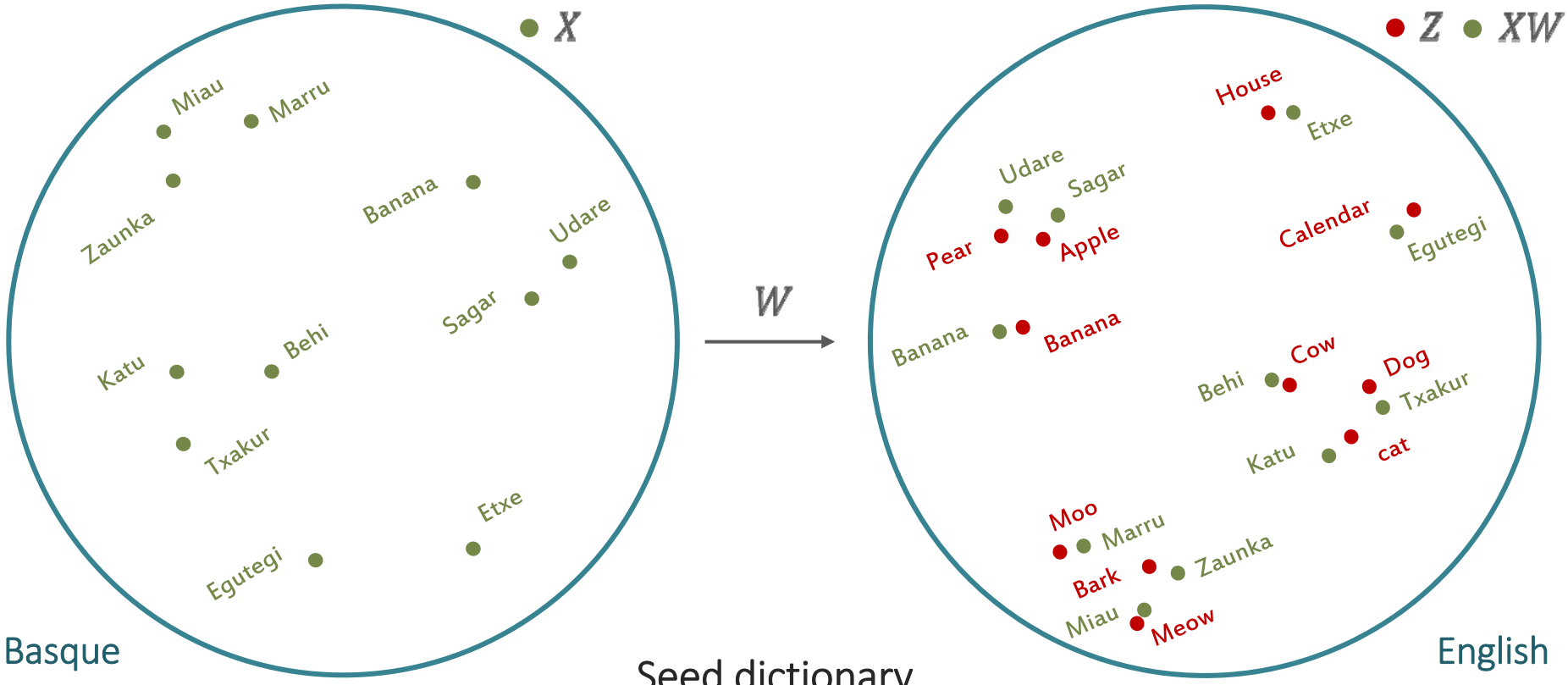
Introduction to embedding mappings



Seed dictionary

$$\begin{matrix} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{matrix} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} \quad \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{matrix} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{matrix}$$

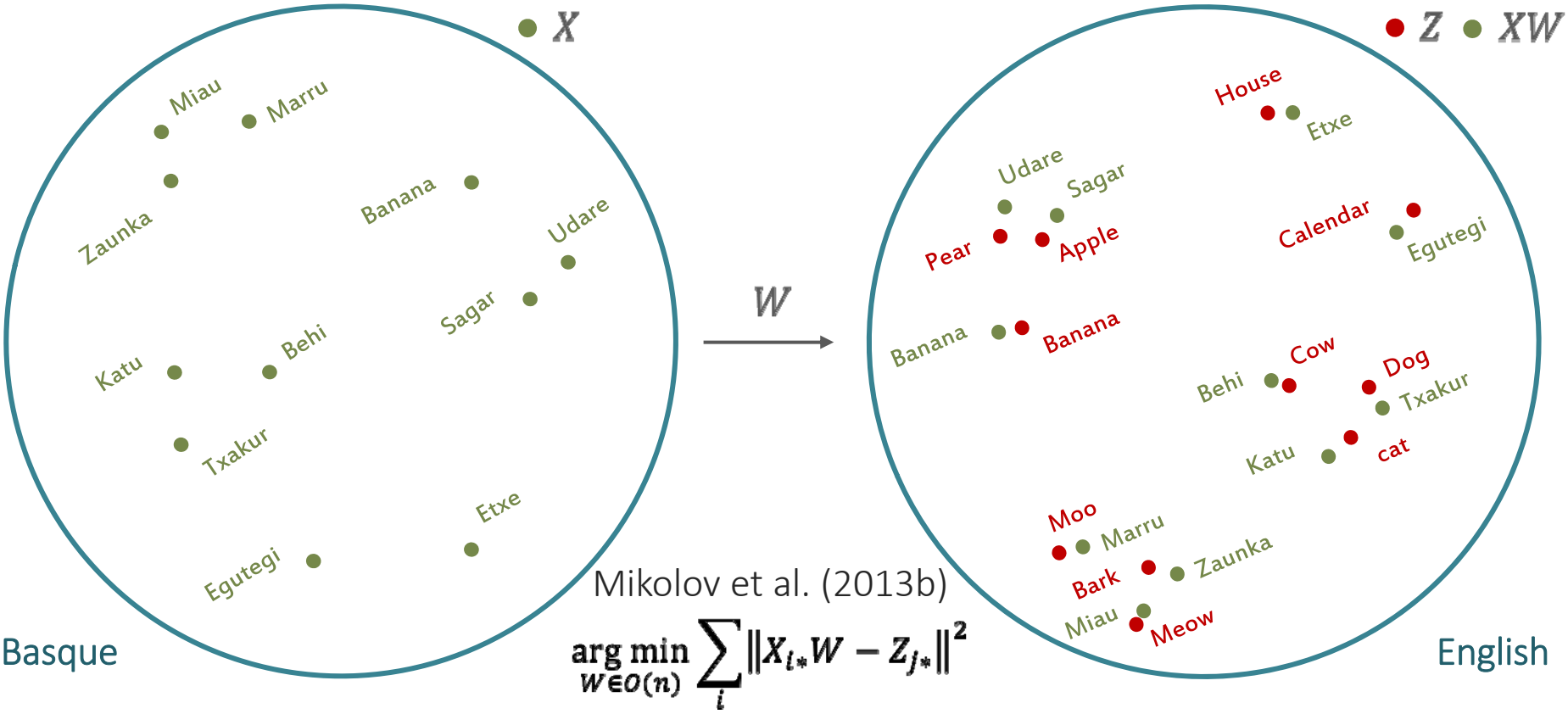
Introduction to embedding mappings



Seed dictionary

$$\begin{matrix} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{matrix} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{matrix} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{matrix}$$

Introduction to embedding mappings

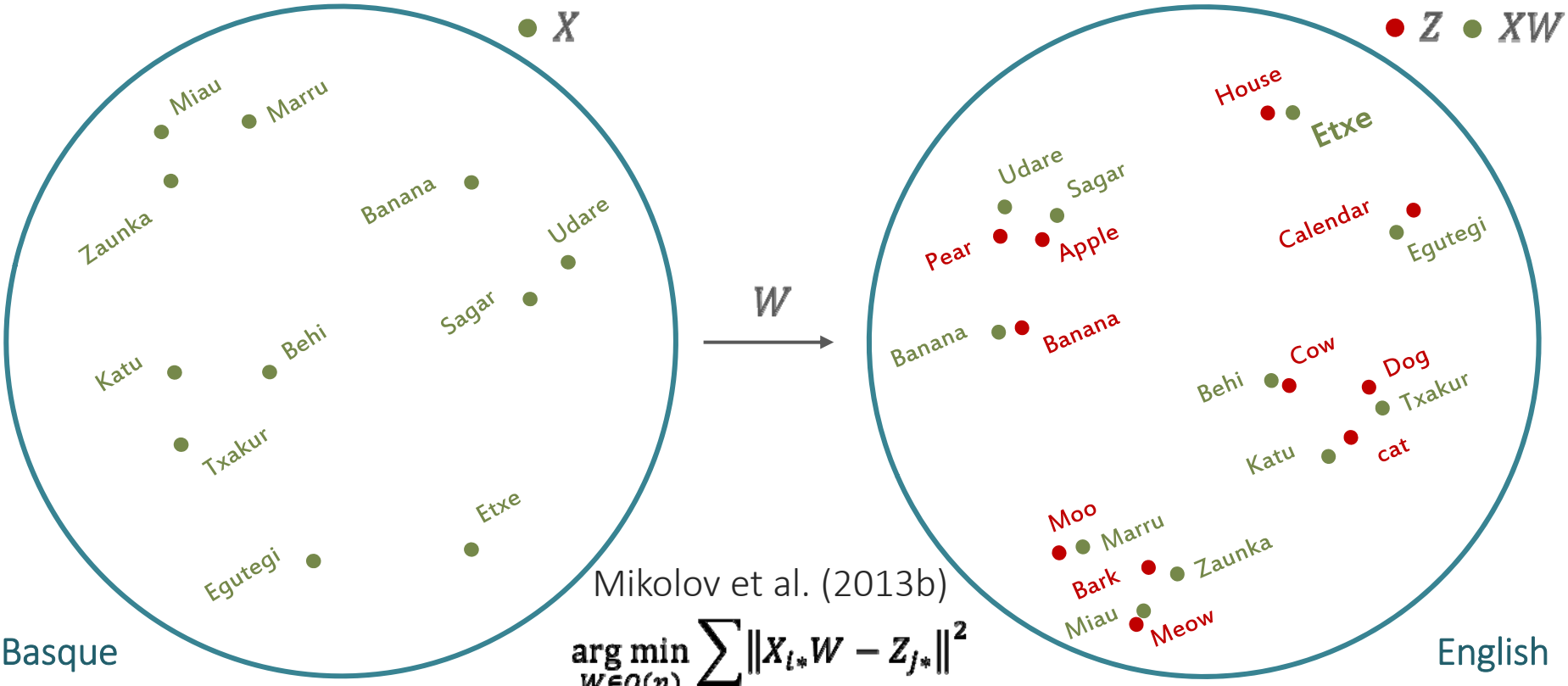


Mikolov et al. (2013b)

$$\arg\min_{W \in O(n)} \sum_i \|X_{i*}W - Z_{j*}\|^2$$

$$\begin{matrix} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{matrix} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{matrix} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{matrix}$$

Introduction to embedding mappings

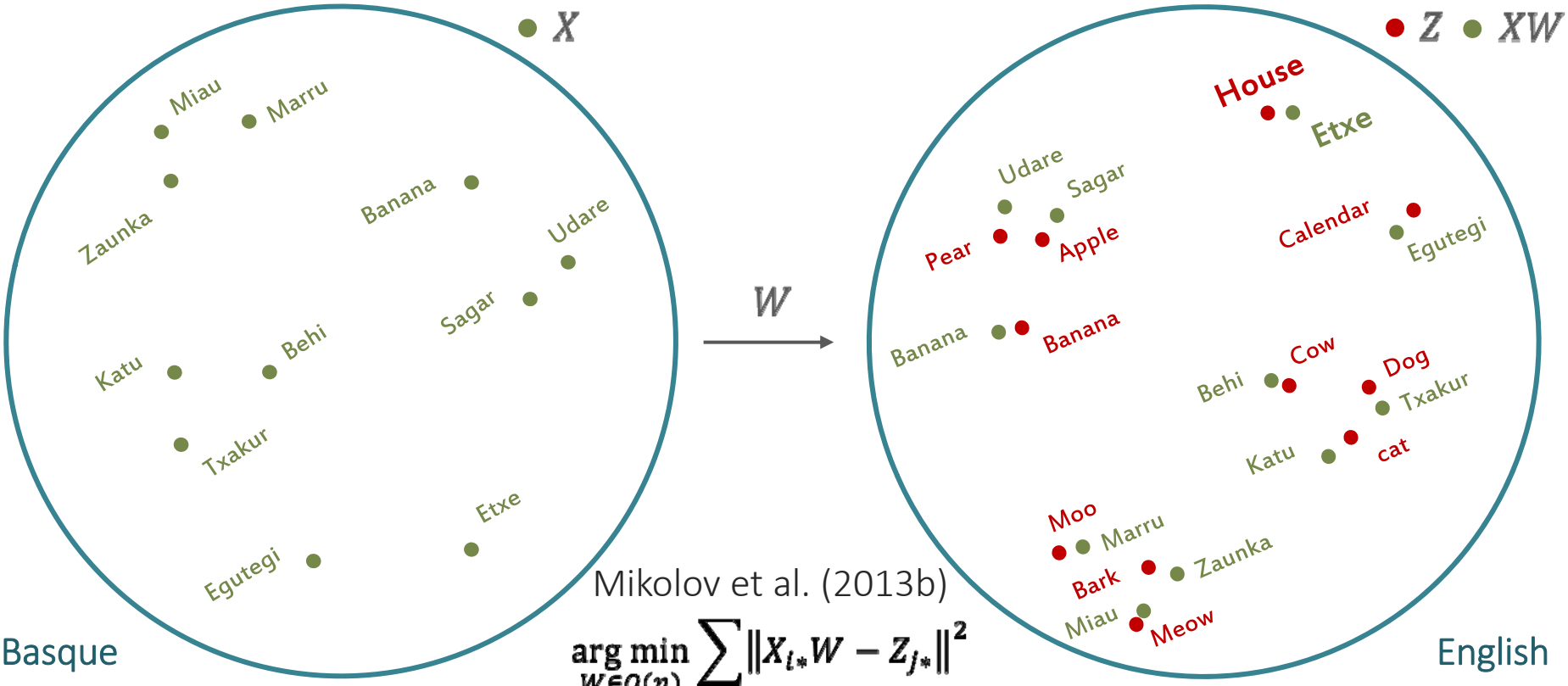


Mikolov et al. (2013b)

$$\arg\min_{W \in O(n)} \sum_{i} \|X_{i*}W - Z_{j*}\|^2$$

$$\begin{matrix} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{matrix} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{matrix} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{matrix}$$

Introduction to embedding mappings

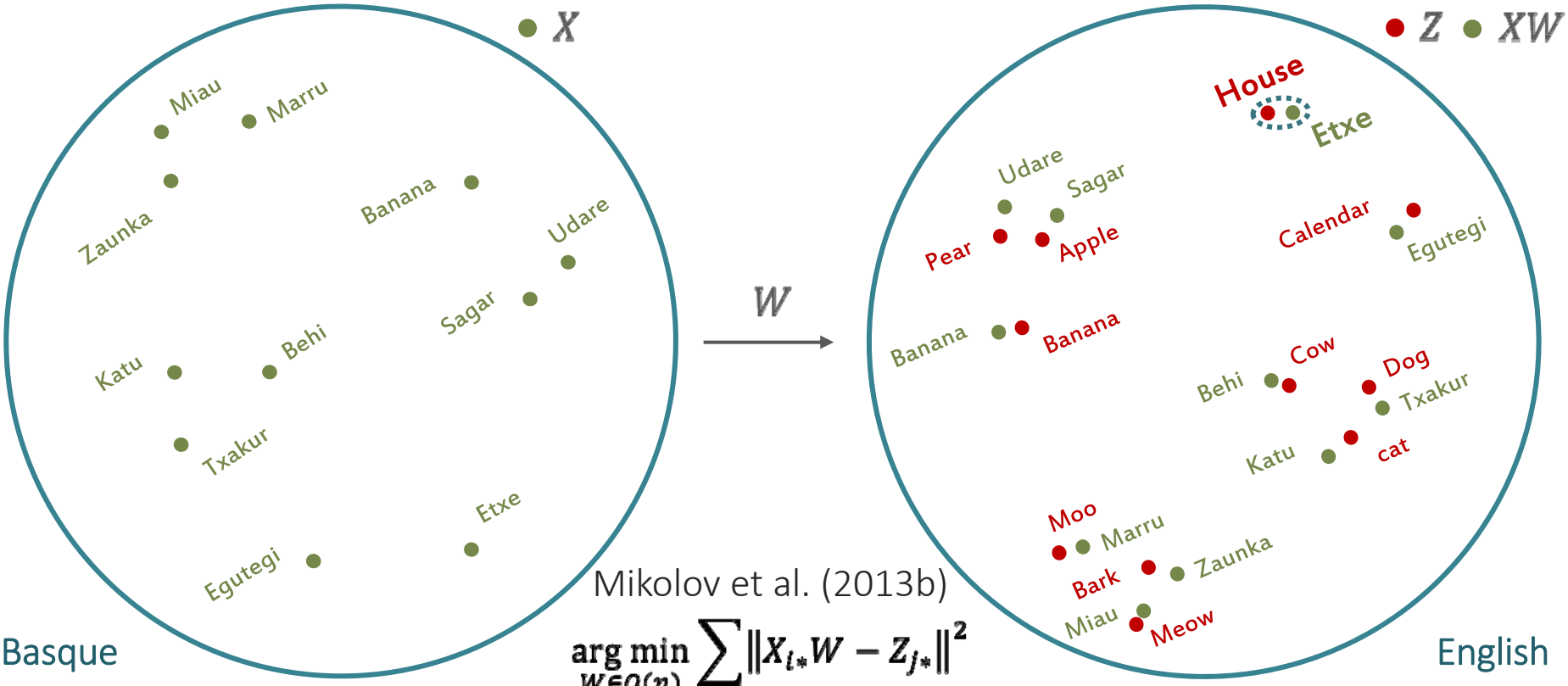


Mikolov et al. (2013b)

$$\arg\min_{W \in O(n)} \sum_i \|X_{i*}W - Z_{j*}\|^2$$

$$\begin{matrix} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{matrix} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{matrix} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{matrix}$$

Introduction to embedding mappings



Mikolov et al. (2013b)

$$\arg\min_{W \in O(n)} \sum_i \|X_{i*}W - Z_{j*}\|^2$$

$$\begin{matrix} \text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi} \end{matrix} \begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix} [W] \approx \begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix} \begin{matrix} \text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar} \end{matrix}$$

Strong system for **supervised** mapping Artetxe et al. AAAI 2018

- Framework subsuming previous work, learns two mappings W_X W_Z as **sequences of (optional) linear mappings**:

  (opt.) Pre-process

  1. (opt.) Whitening
  2. **Orthogonal mapping**
  3. (opt.) Re-weighting
  4. (opt.) De-whitening

- The optional steps, properly combined, bring up to 5 points improvement

Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish

Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish

⇒ Monolingual embeddings (CBOW + negative sampling)

Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish

⇒ Monolingual embeddings (CBOW + negative sampling)

⇒ Seed dictionary: 5,000 word pairs

Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish

⇒ Monolingual embeddings (CBOW + negative sampling)

⇒ Seed dictionary: 5,000 pairs

⇒ Test dictionary: 1,500 pairs (Nearest neighbor, P@1)

Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish

⇒ Monolingual embeddings (CBOW + negative sampling)

⇒ Seed dictionary: 5,000 pairs

⇒ Test dictionary: 1,500 pairs (Nearest neighbor, P@1)

Method	EN-IT	EN-DE	EN-FI	EN-ES

Evaluating via Bilingual Dictionary induction

Dataset by Dinu et al. (2015) extended to German, Finnish, Spanish

⇒ Monolingual embeddings (CBOW + negative sampling)

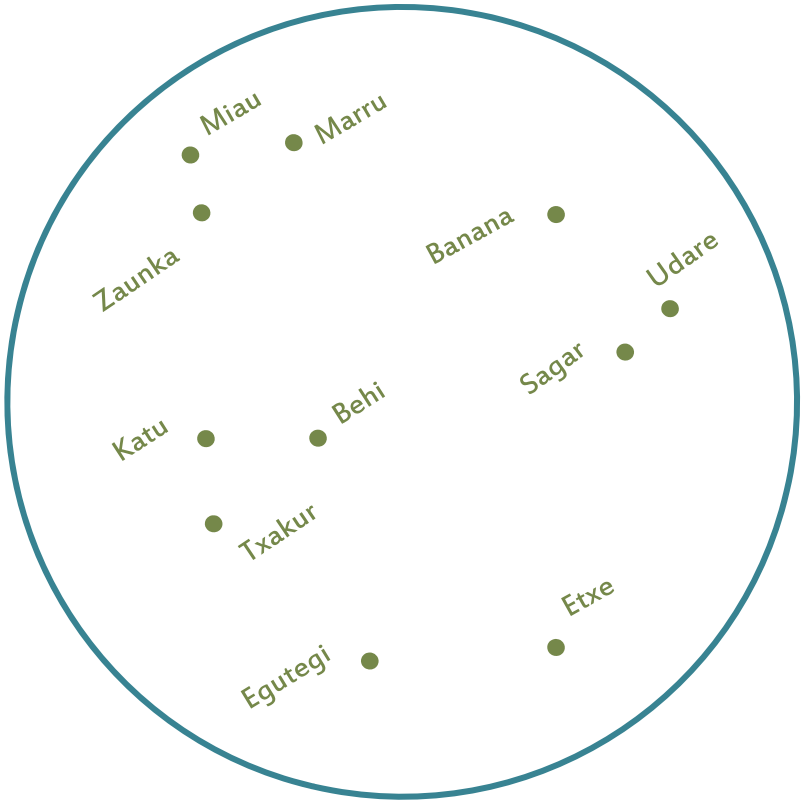
⇒ Seed dictionary: 5,000 pairs

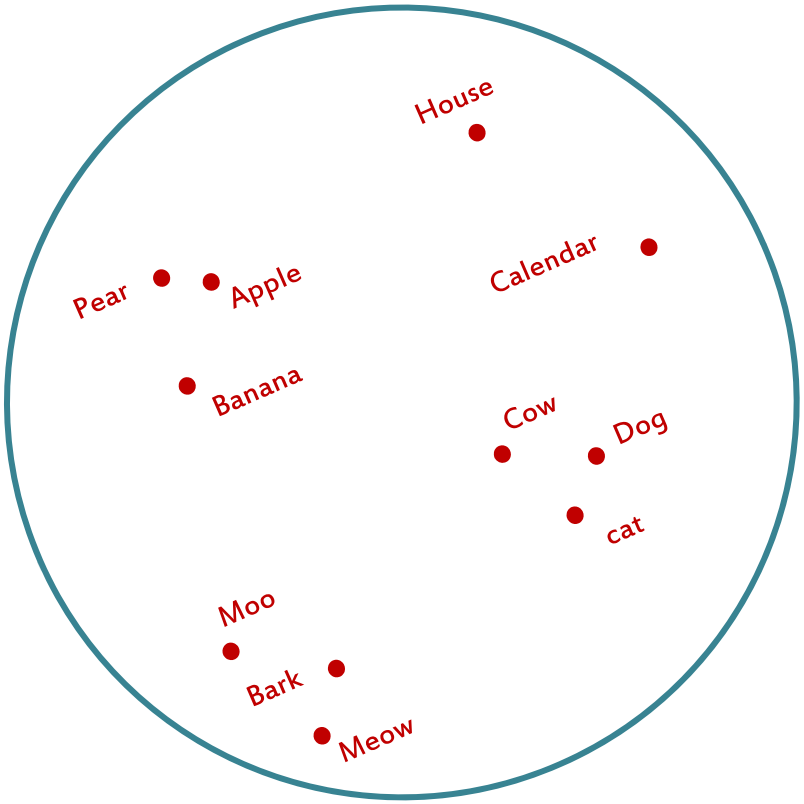
⇒ Test dictionary: 1,500 pairs (Nearest neighbor, P@1)

Method	EN-IT	EN-DE	EN-FI	EN-ES
Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
Faruqui and Dyer (2014)	38.40*	37.13*	27.60*	26.80*
Shigeto et al. (2015)	41.53†	43.07†	31.04†	33.73†
Dinu et al. (2015)	37.7	38.93*	29.14*	30.40*
Lazaridou et al. (2015)	40.2	-	-	-
Xing et al. (2015)	36.87†	41.27†	28.23†	31.20†
Artetxe et al. (2016)	39.27	41.87*	30.62*	31.40*
Zhang et al. (2016)	36.73†	40.80†	28.16†	31.07†
Smith et al. (2017)	43.1	43.33†	29.42†	35.13†
Our method (AAAI18)	**45.27**	**44.13**	**32.94**	**36.60**

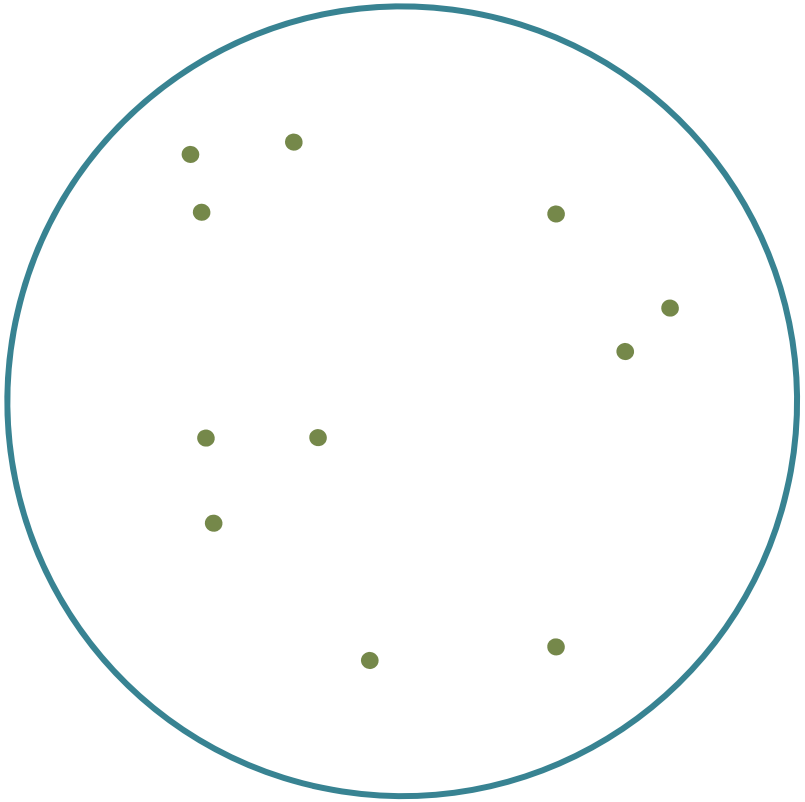
Why does it work?

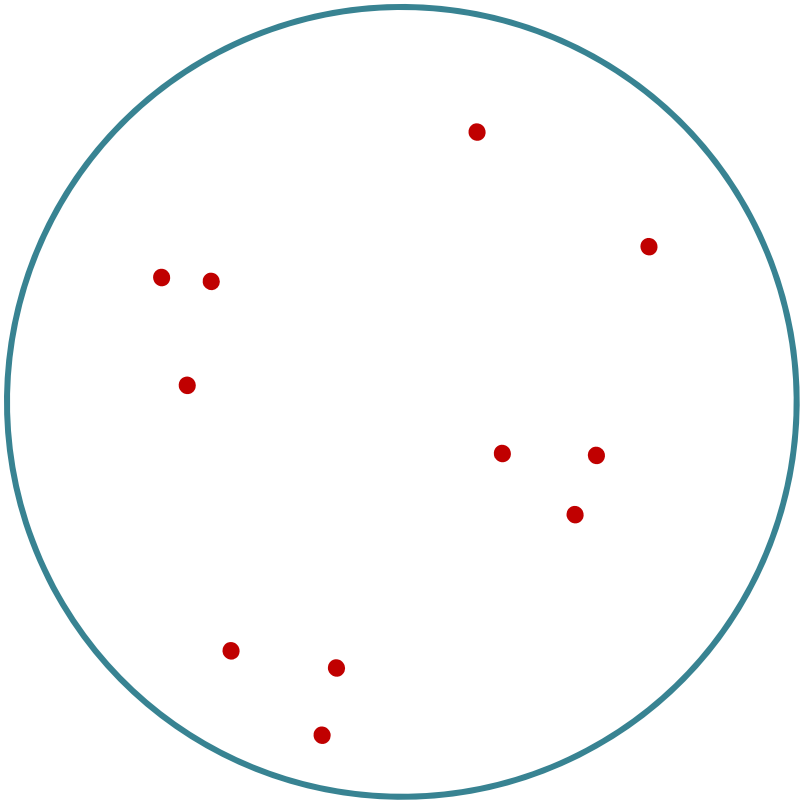
Why does it work?



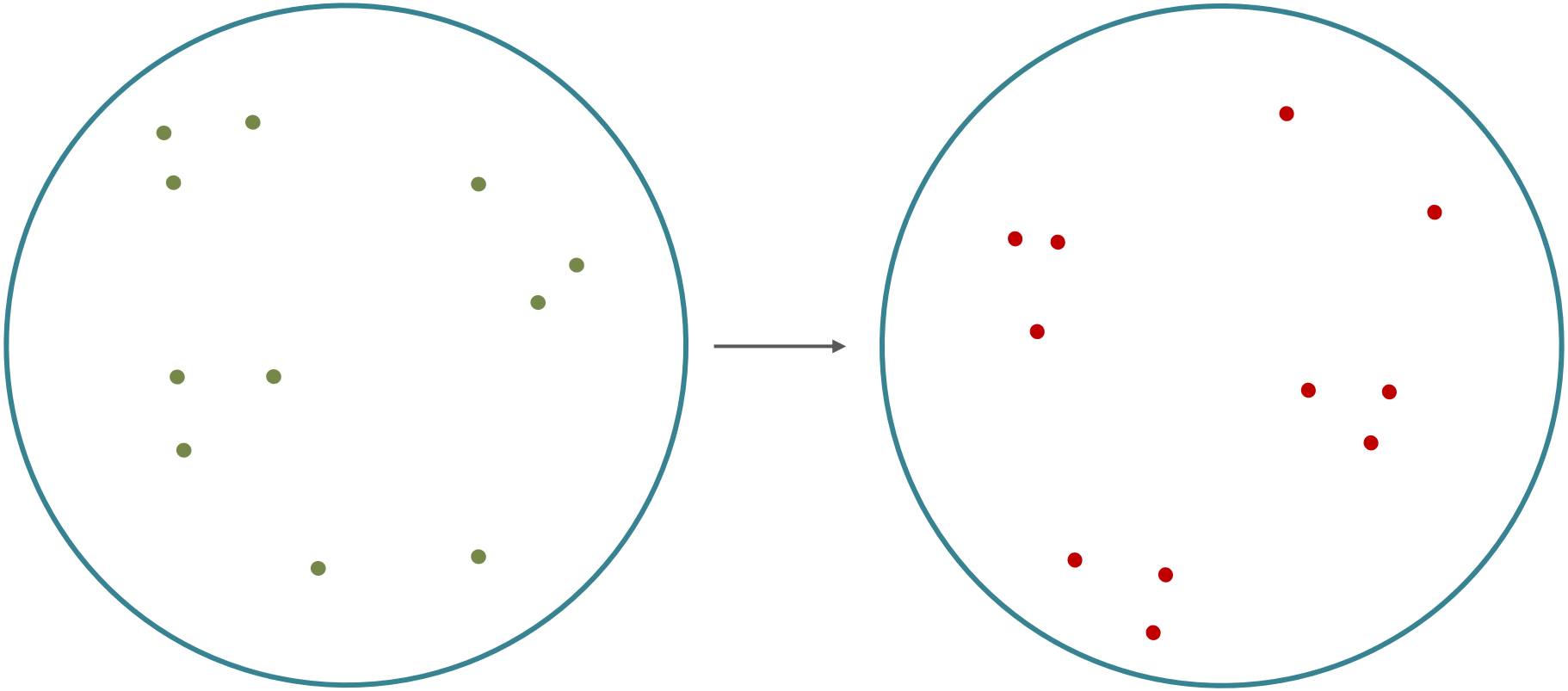


Why does it work?

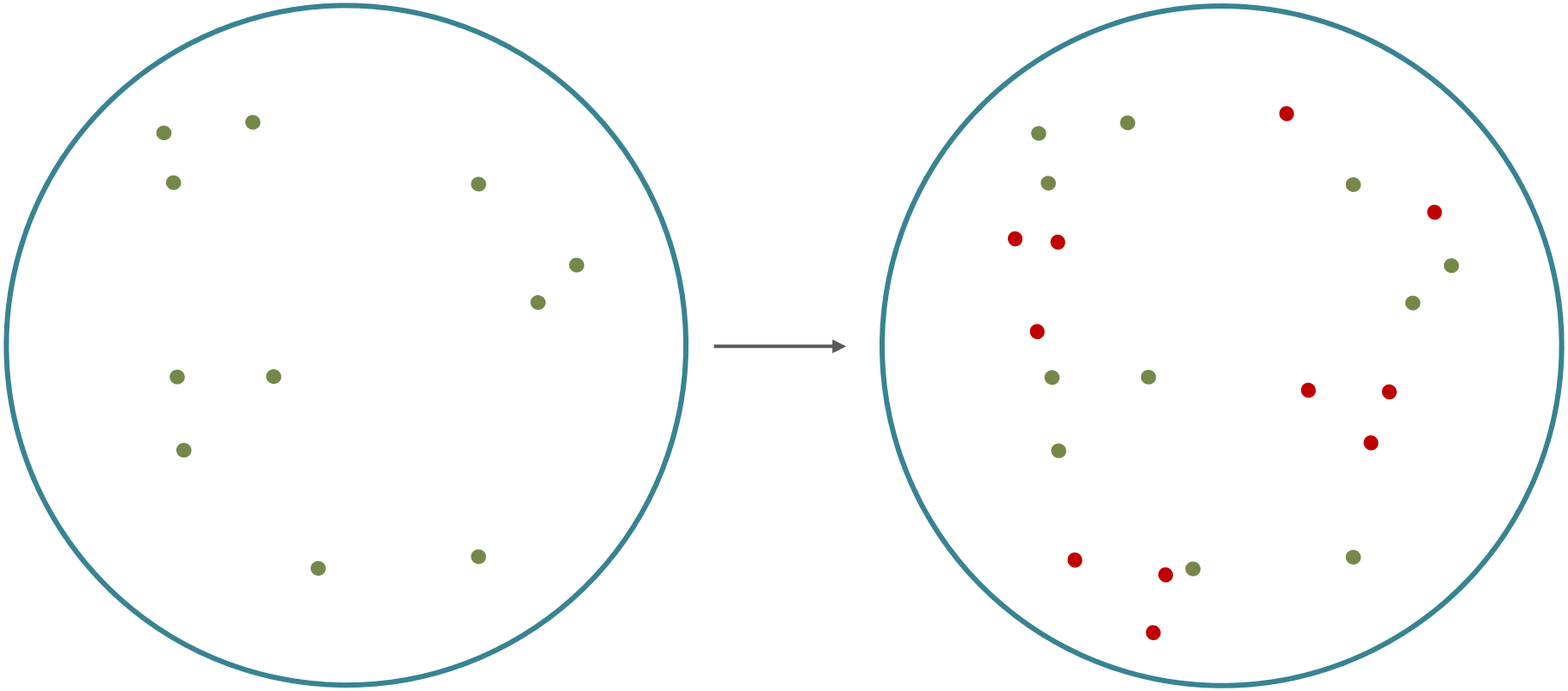




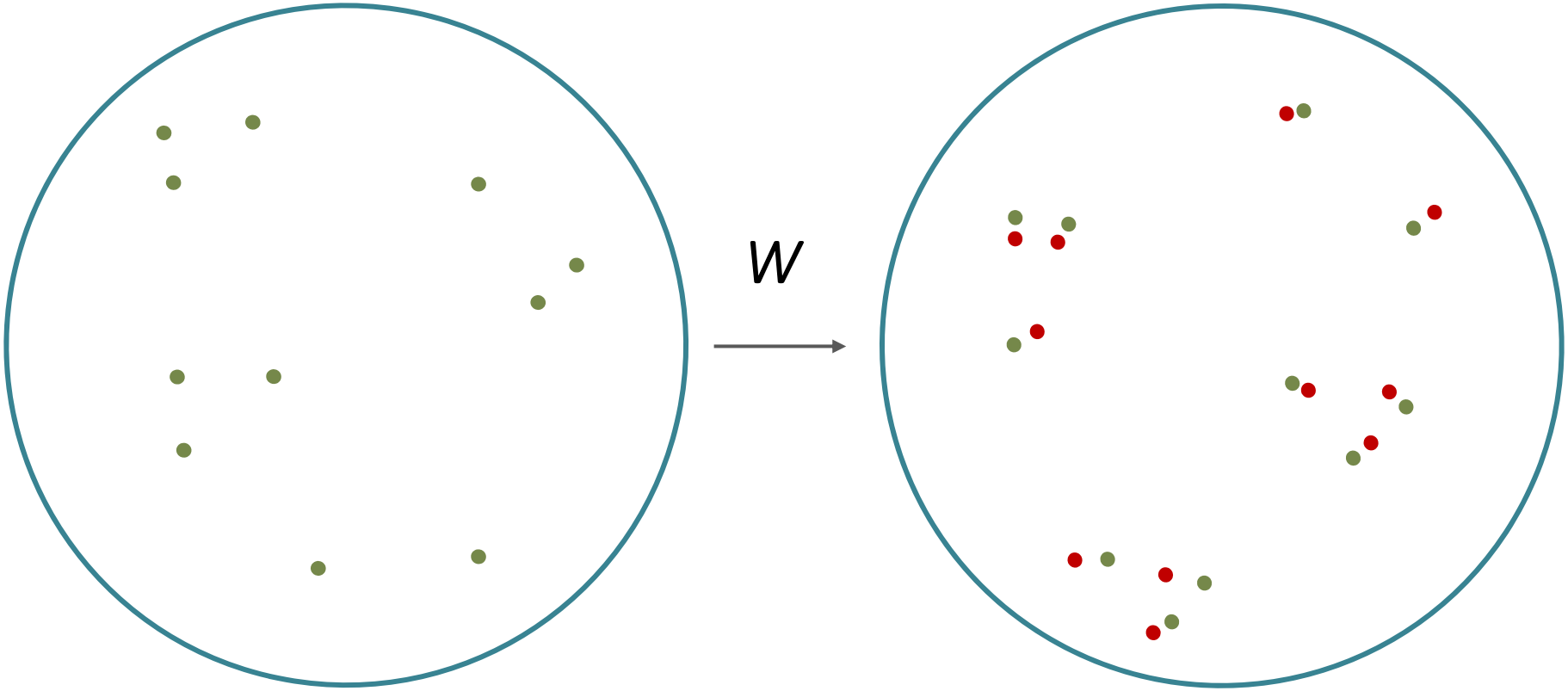
Why does it work?



Why does it work?

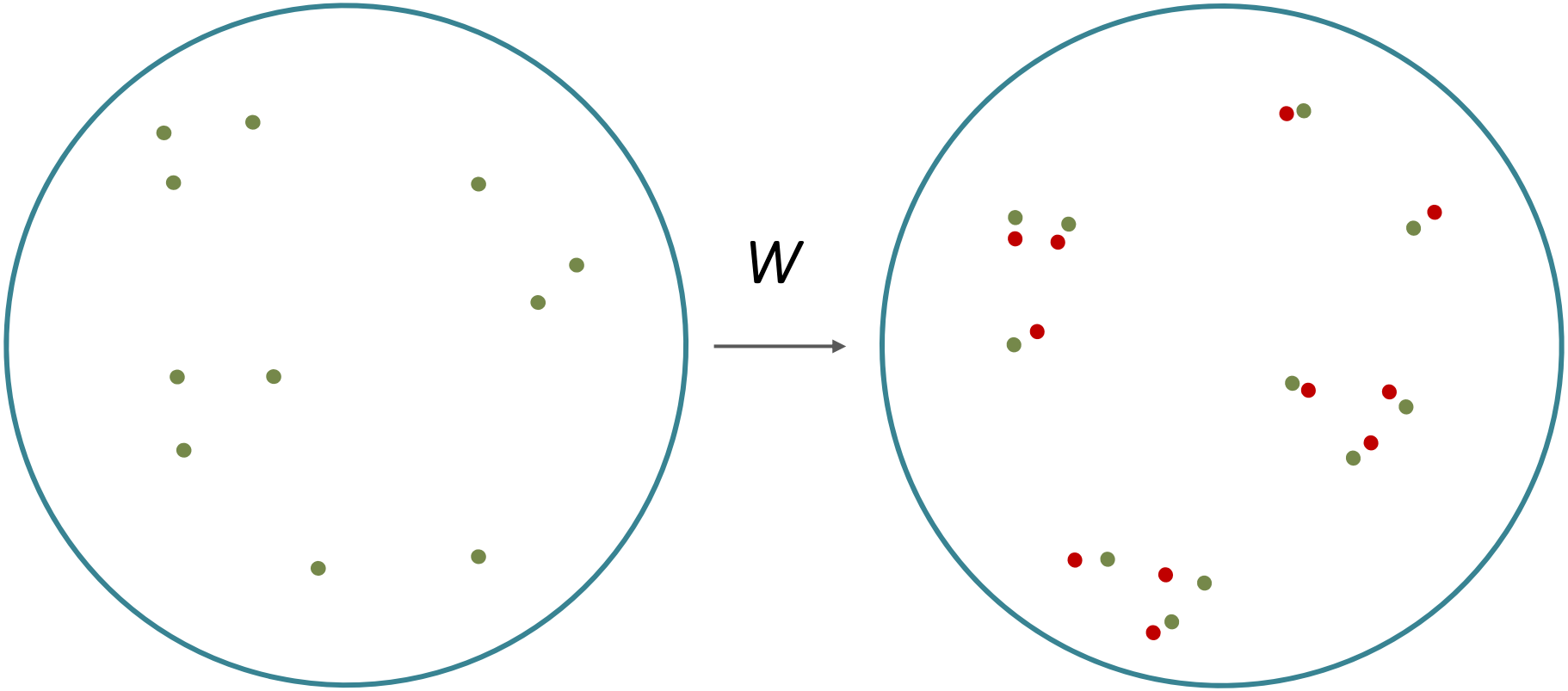


Why does it work?



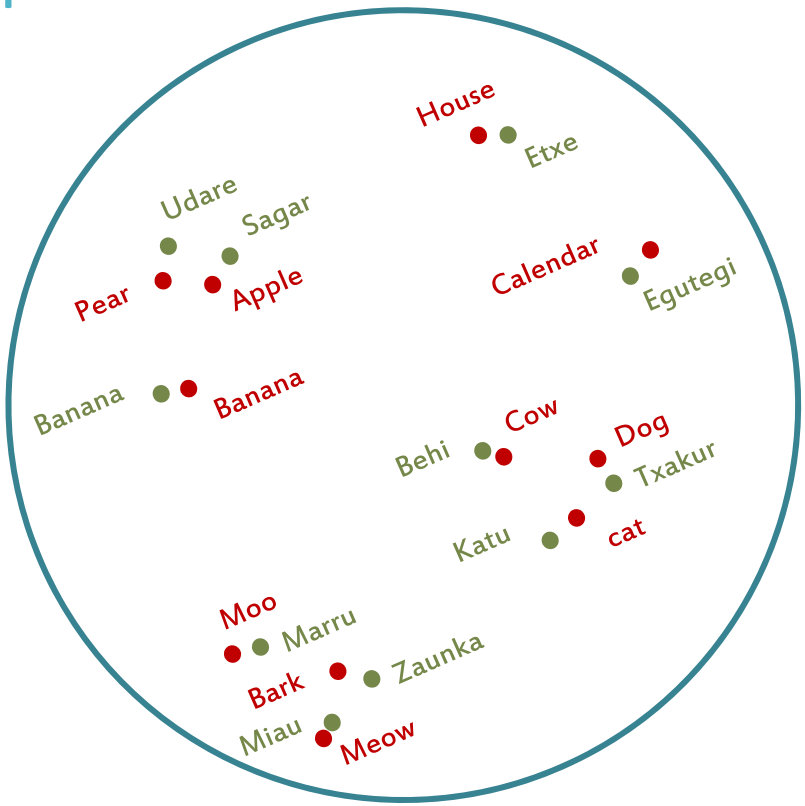
Why does it work?

Languages are (to a large extent)
isometric in word embedding space (!)

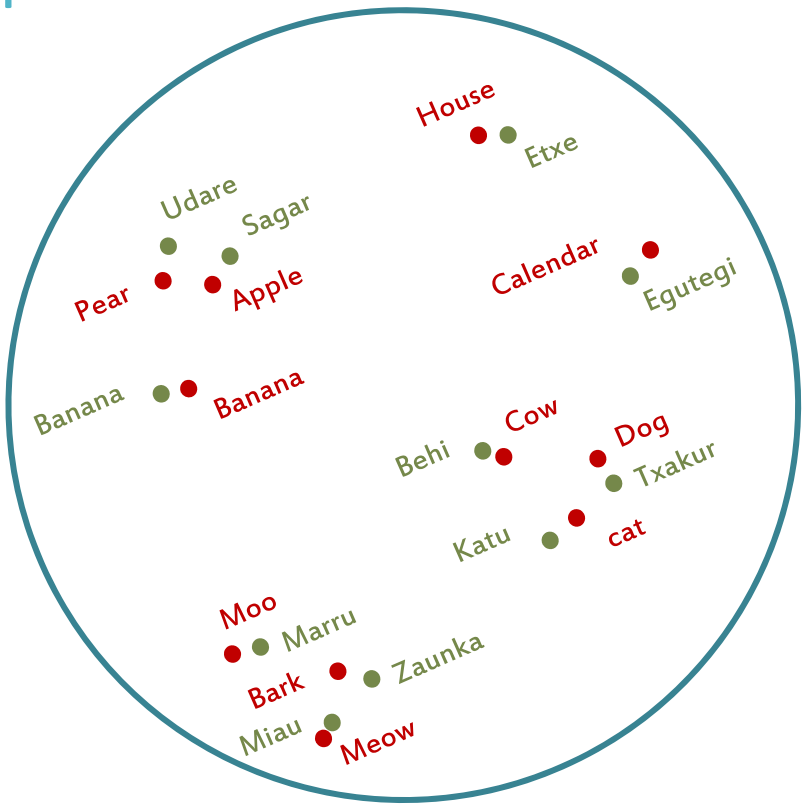


Reducing supervision

Reducing supervision



Reducing supervision

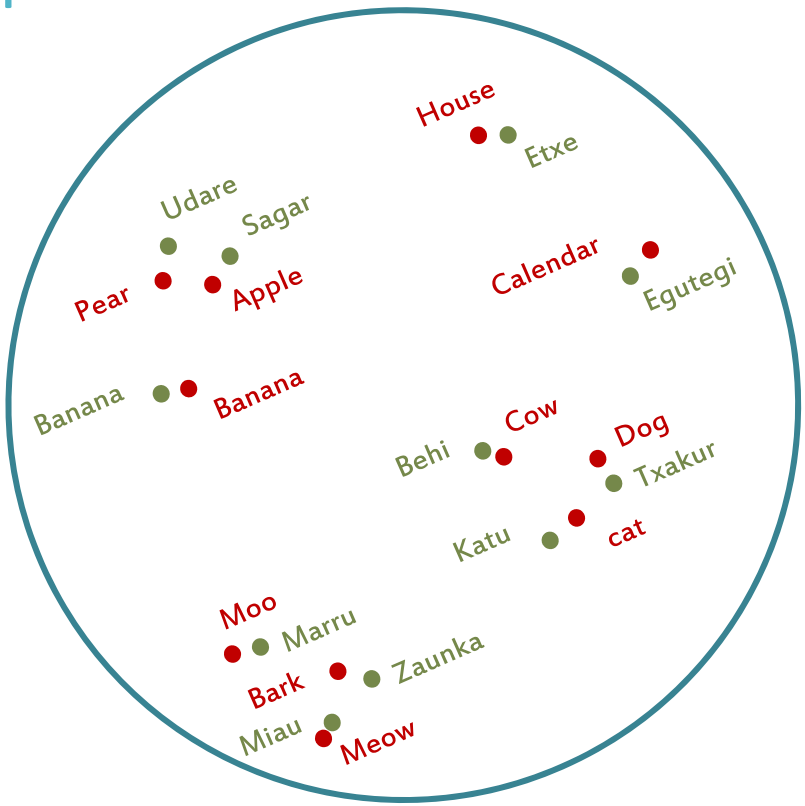


bilingual signal for training

Supervision

- parallel corpora
- comparable corpora
- (big) dictionaries

Reducing supervision



bilingual signal for training

Supervision

- parallel corpora
- comparable corpora
- (big) dictionaries

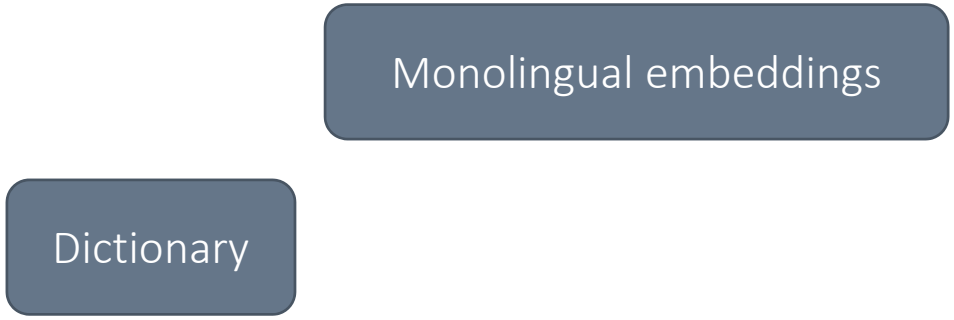
- 25 word dictionary
- numerals (1, 2, 3…)
- nothing

Self-learning

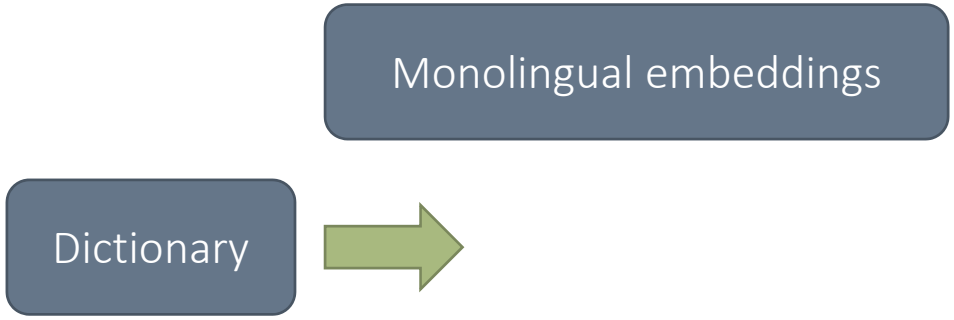
Self-learning

Monolingual embeddings

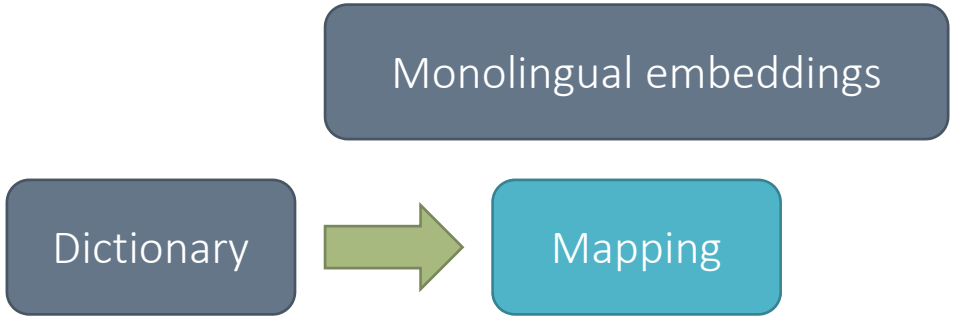
Self-learning



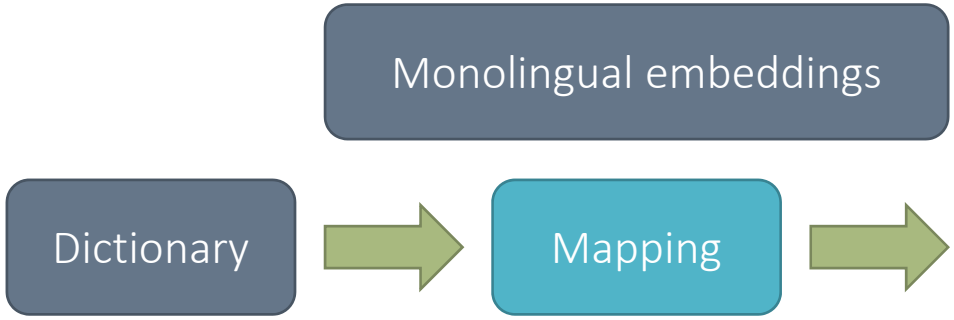
Self-learning



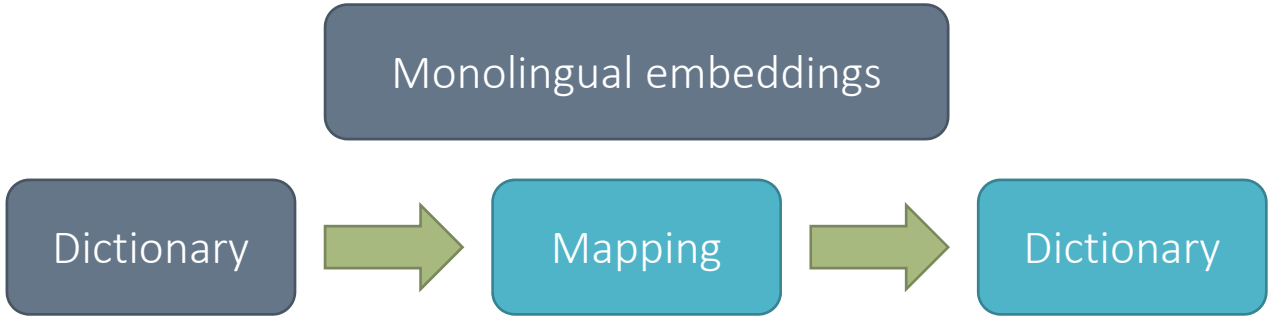
Self-learning



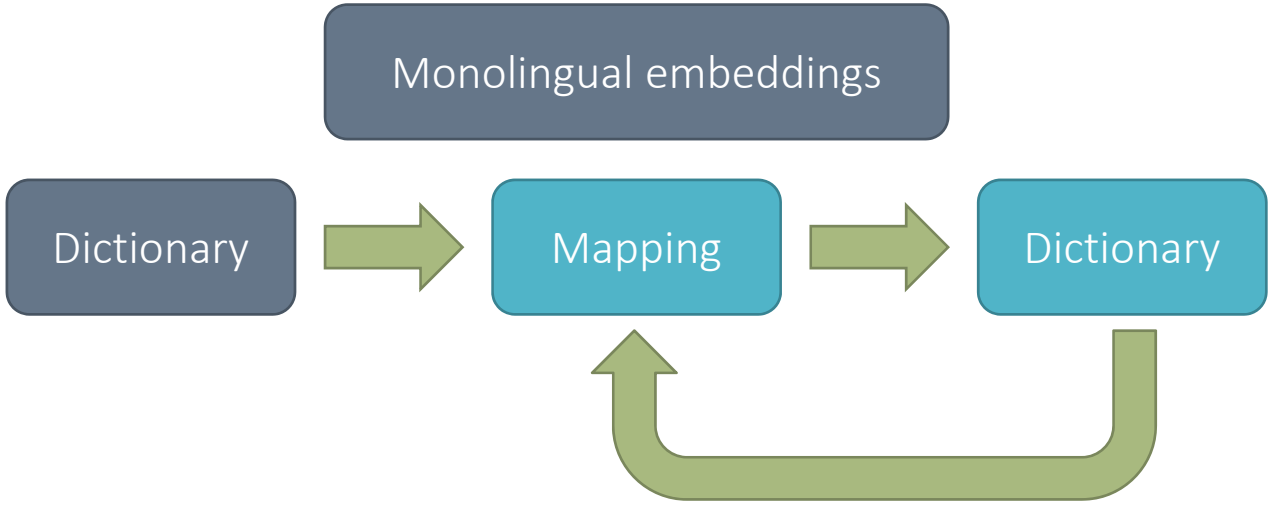
Self-learning



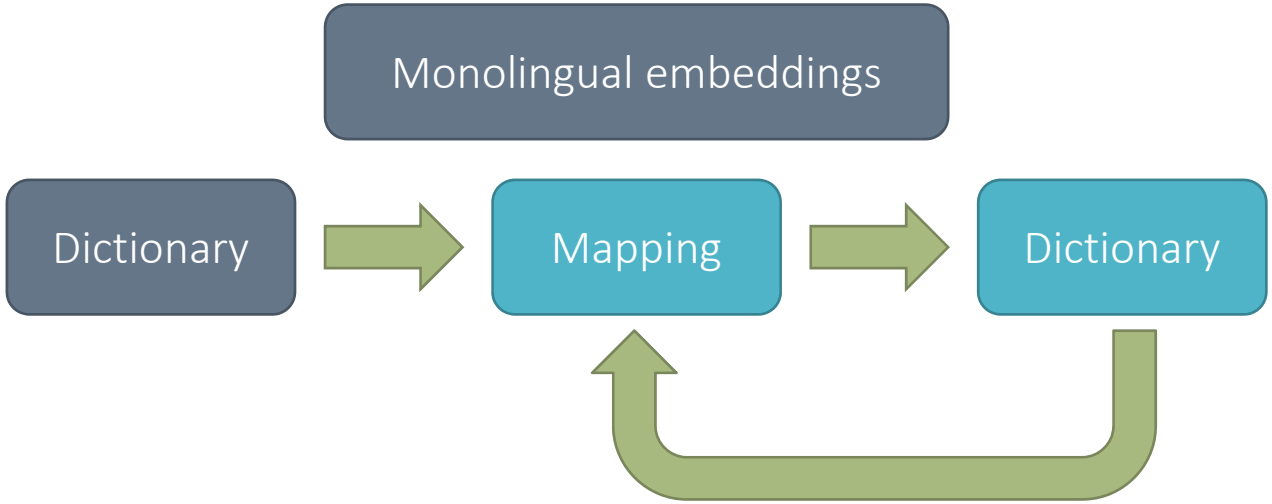
Self-learning



Self-learning



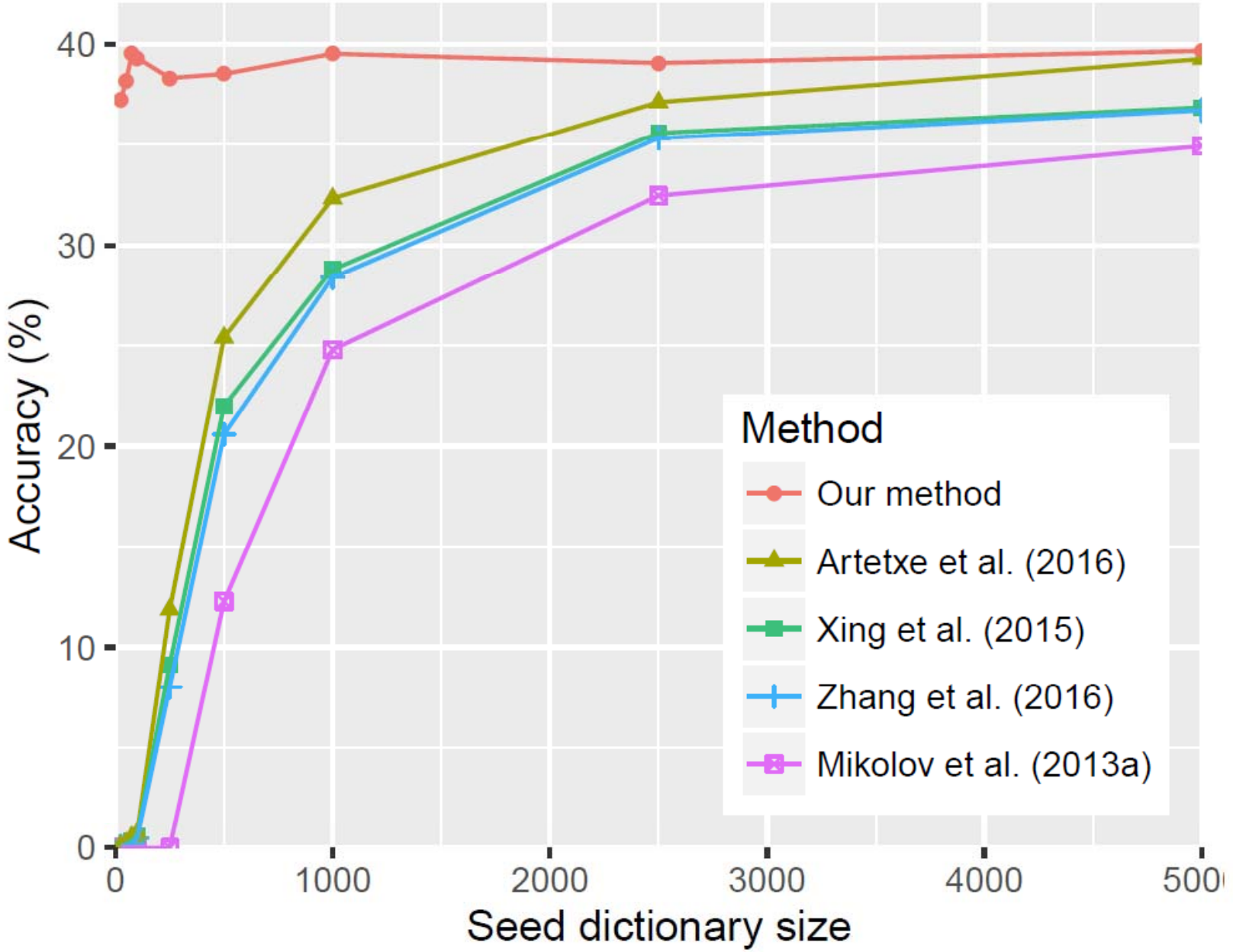
Self-learning



Too good to be true?

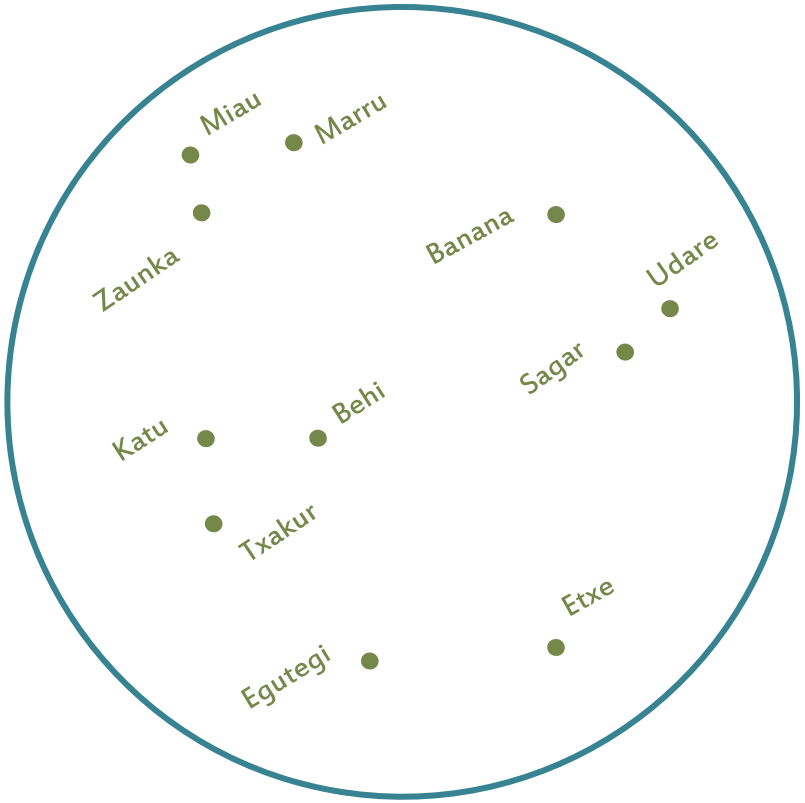
Semi-supervised experiments (ACL17)

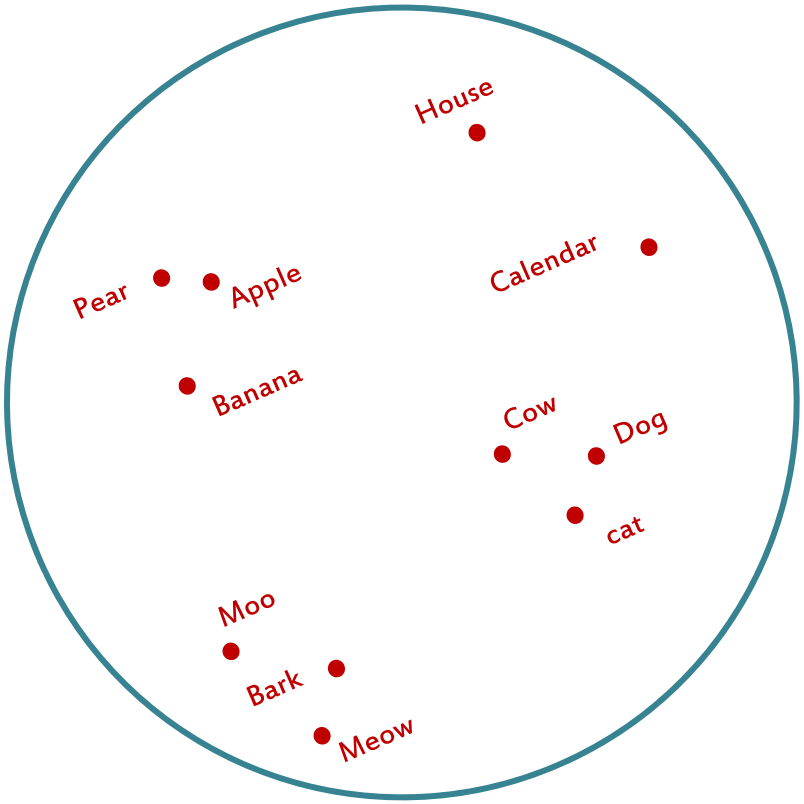
Semi-supervised experiments (ACL17)



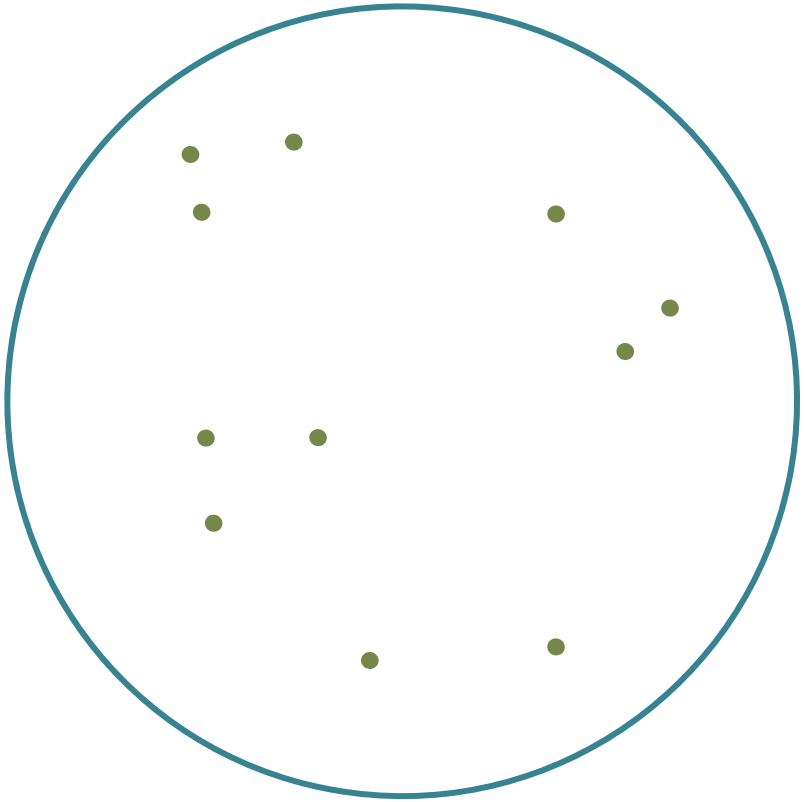
Why does it work?

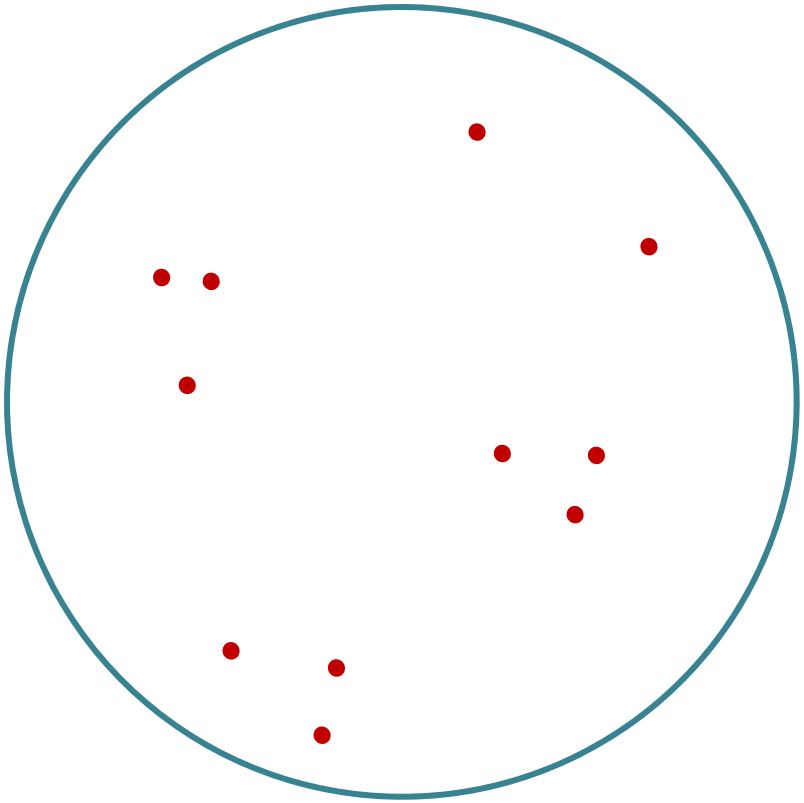
Why does it work?



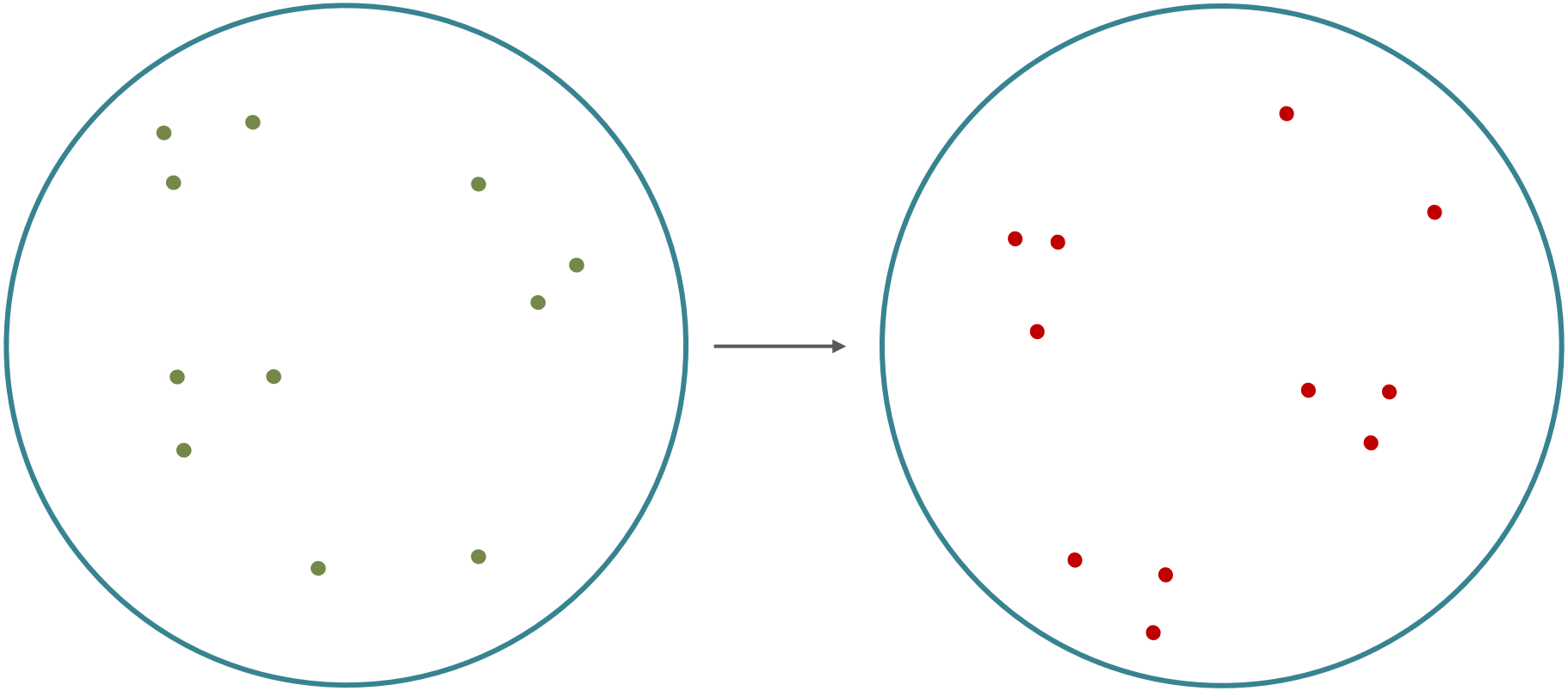


Why does it work?

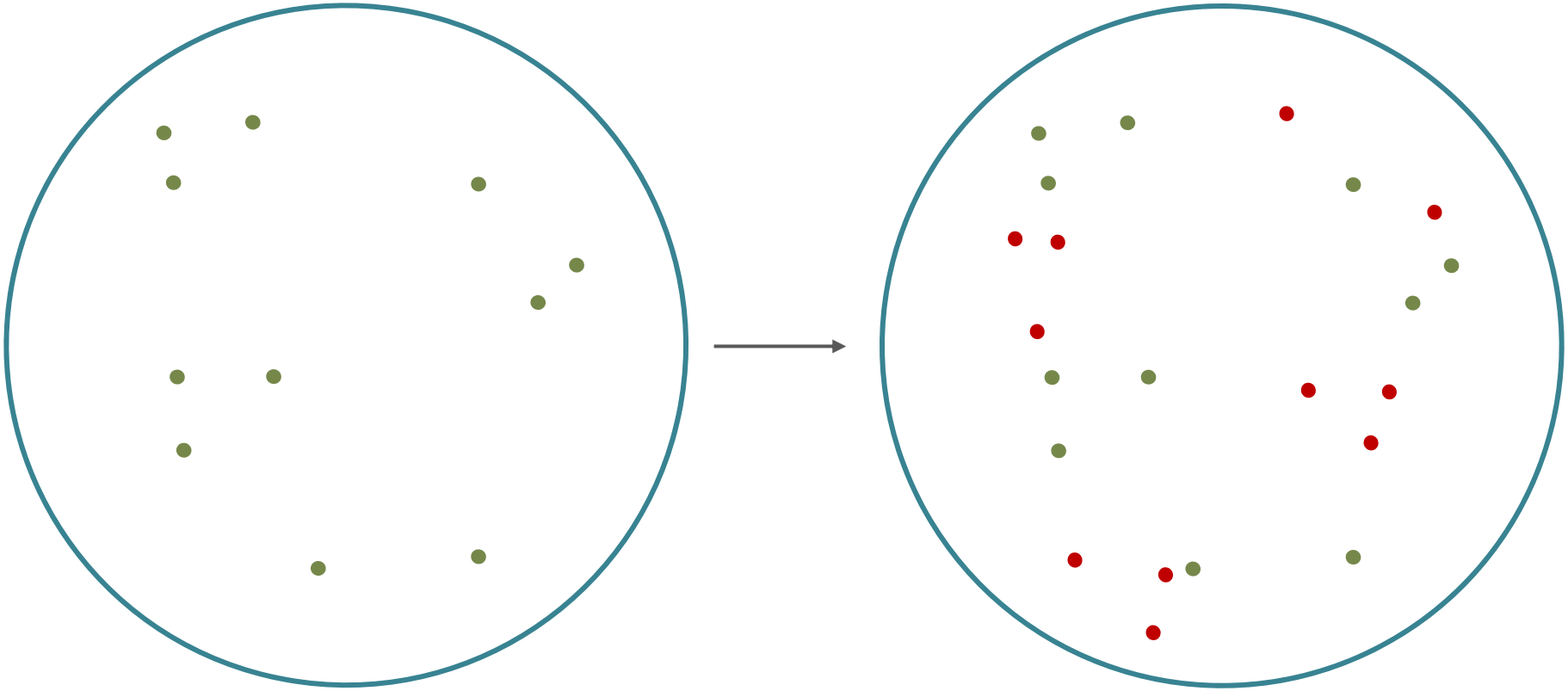




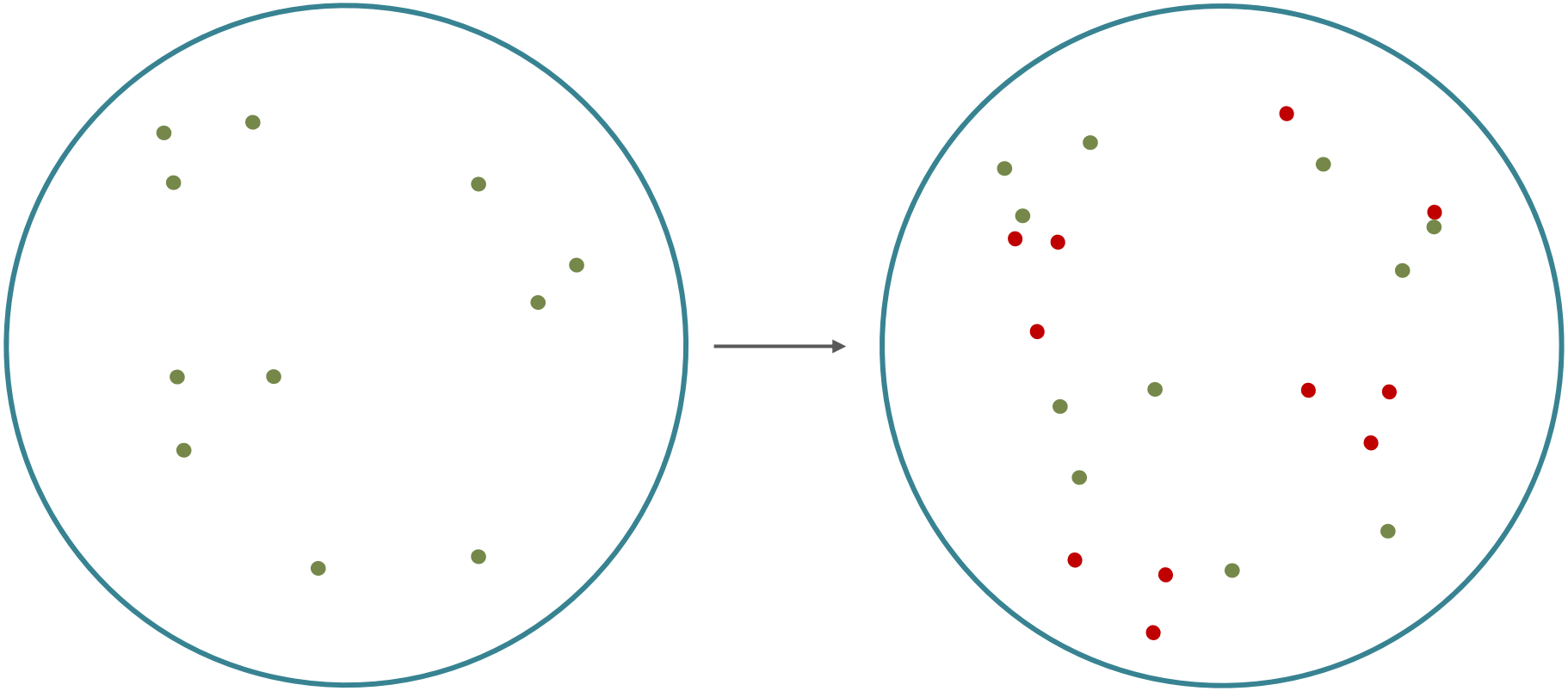
Why does it work?



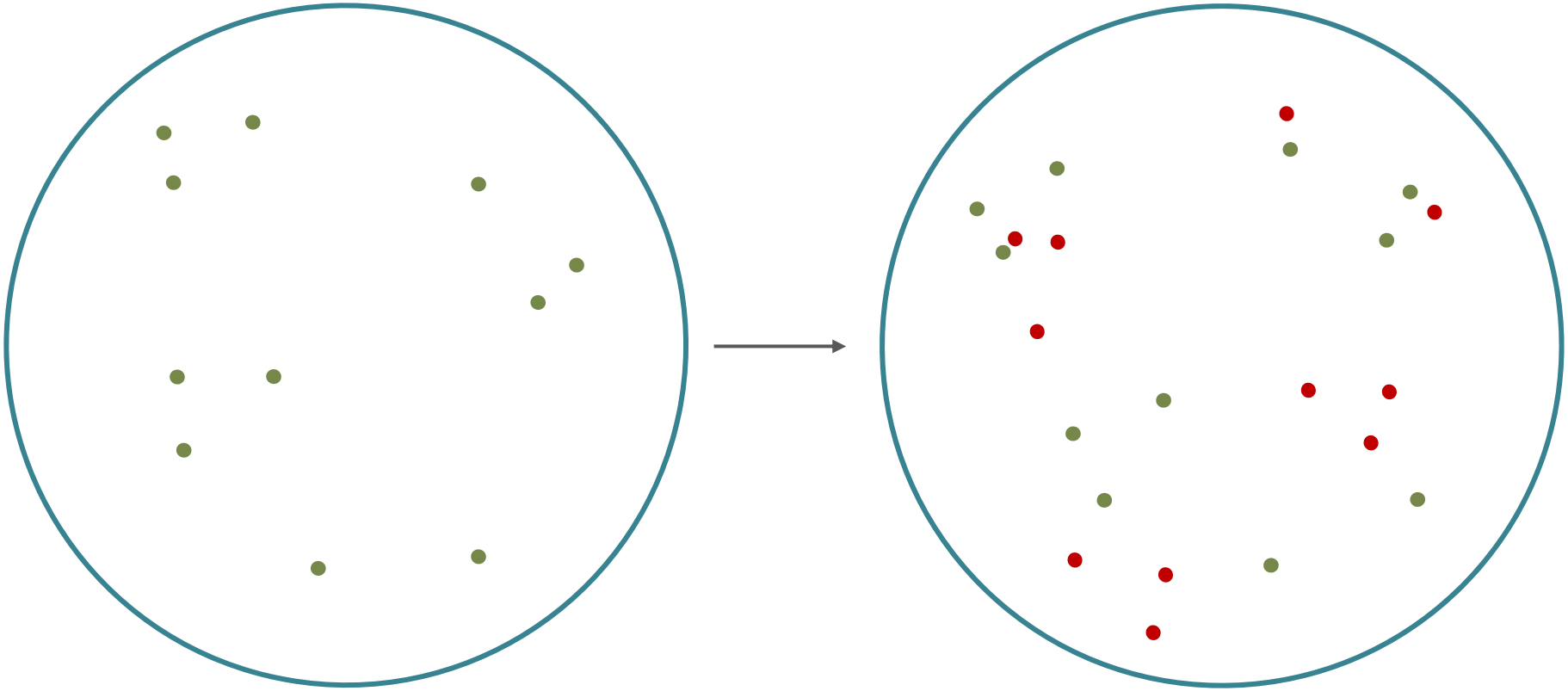
Why does it work?



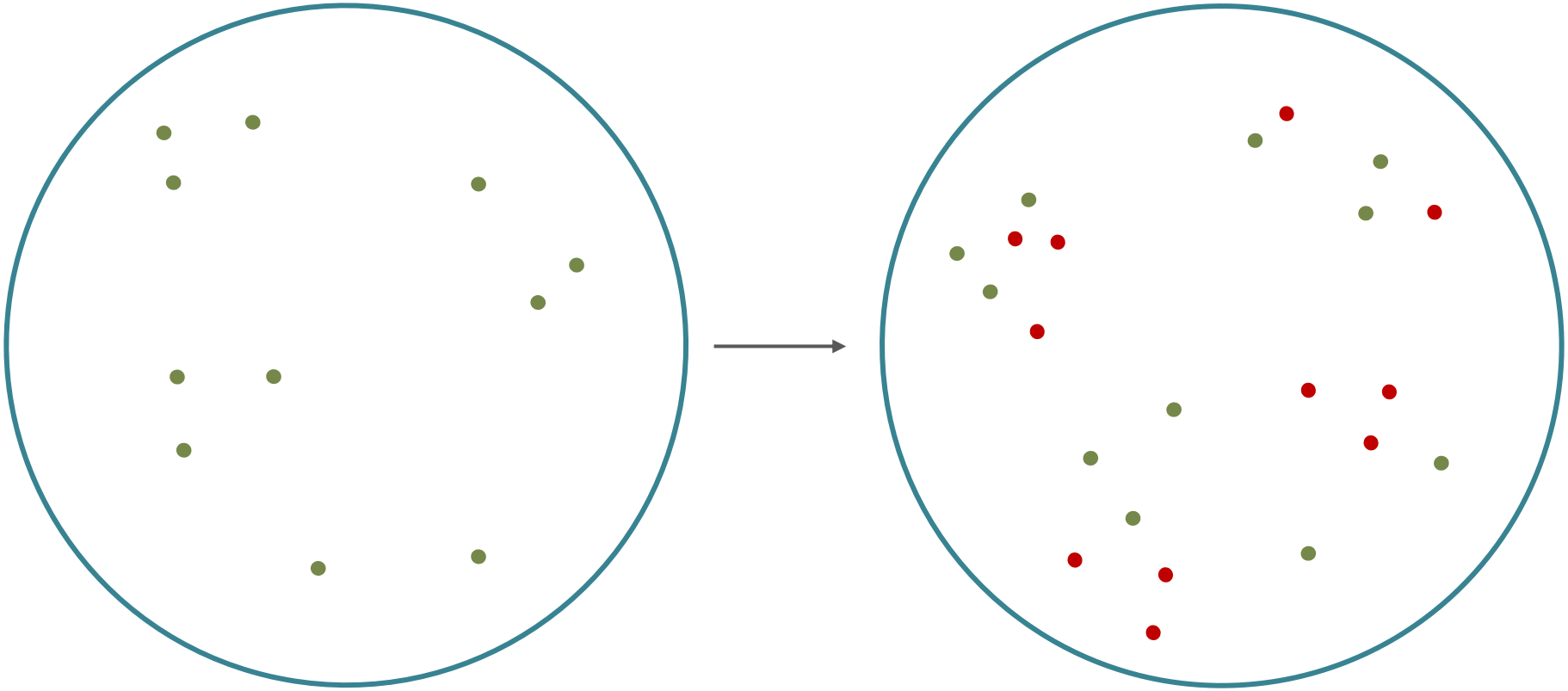
Why does it work?



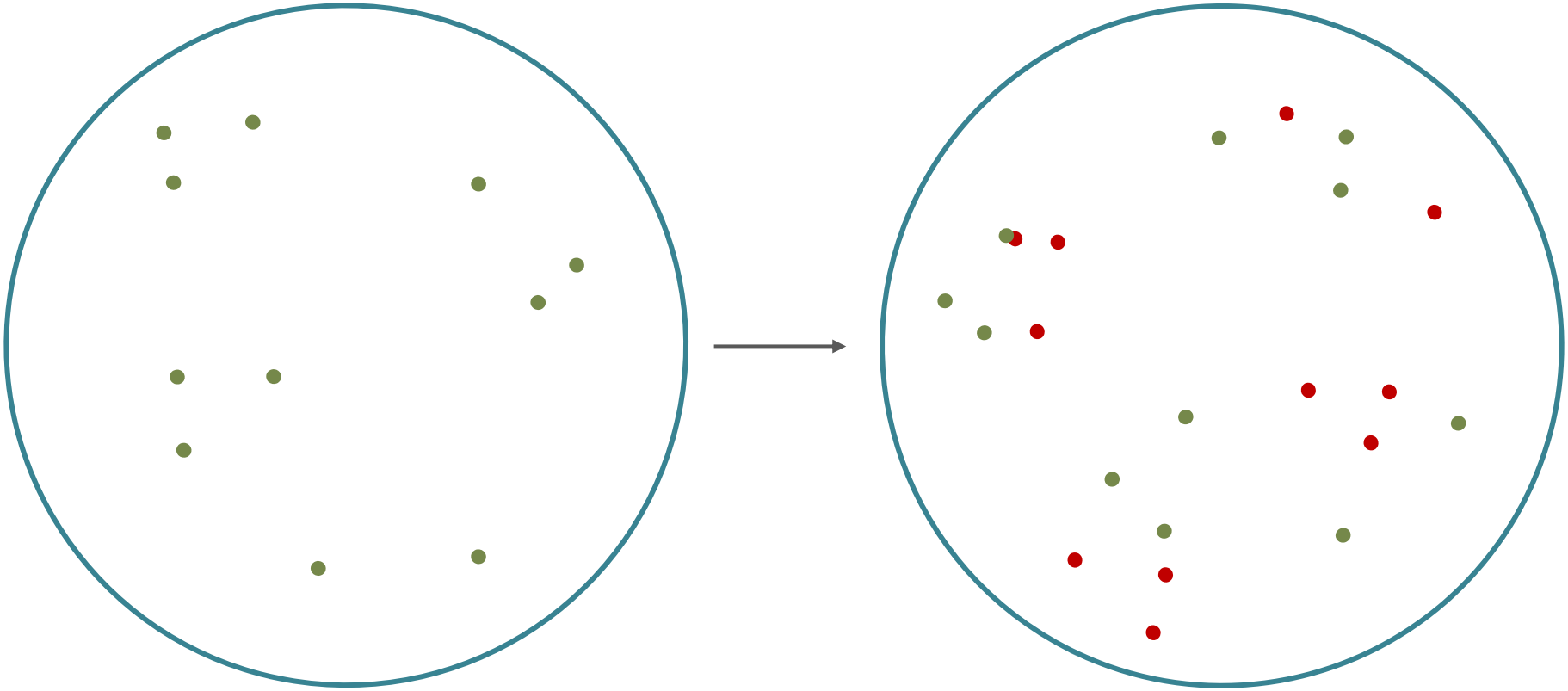
Why does it work?



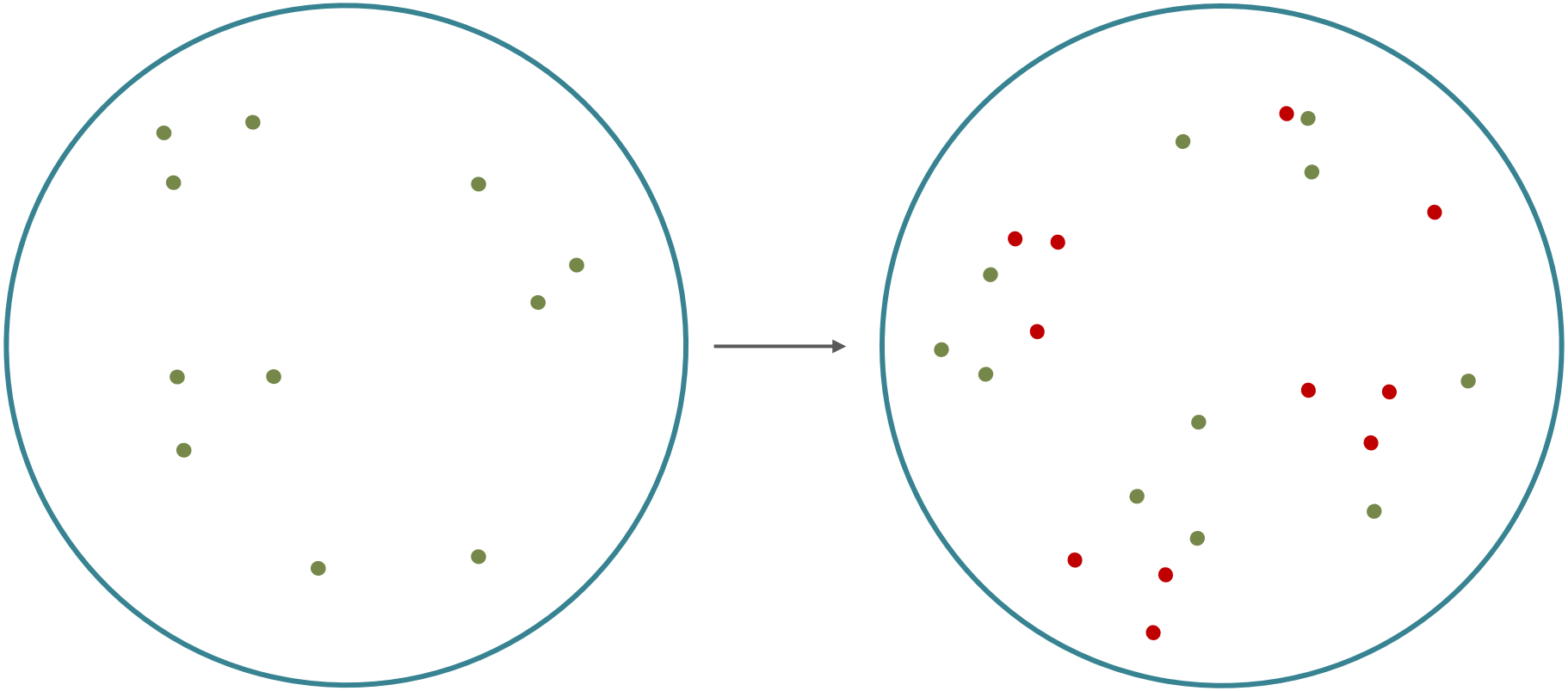
Why does it work?



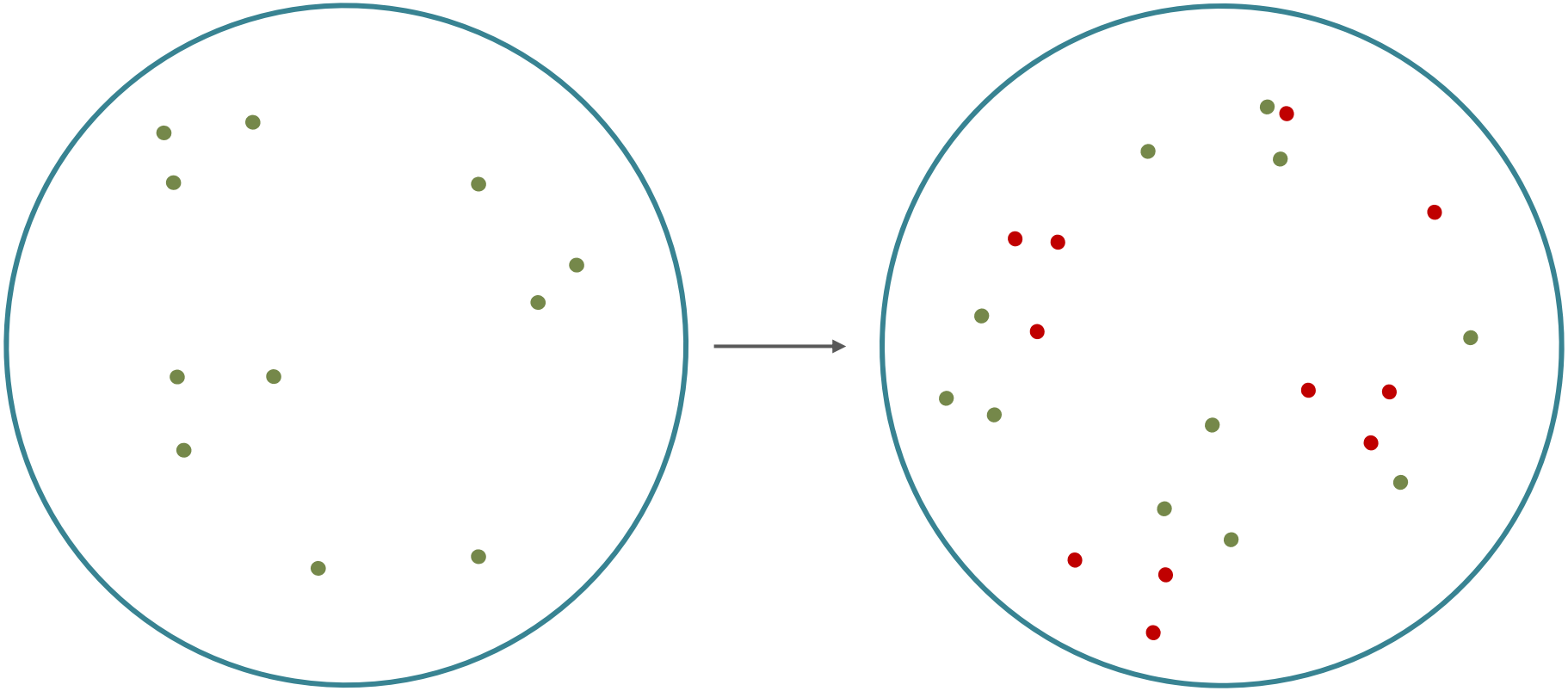
Why does it work?



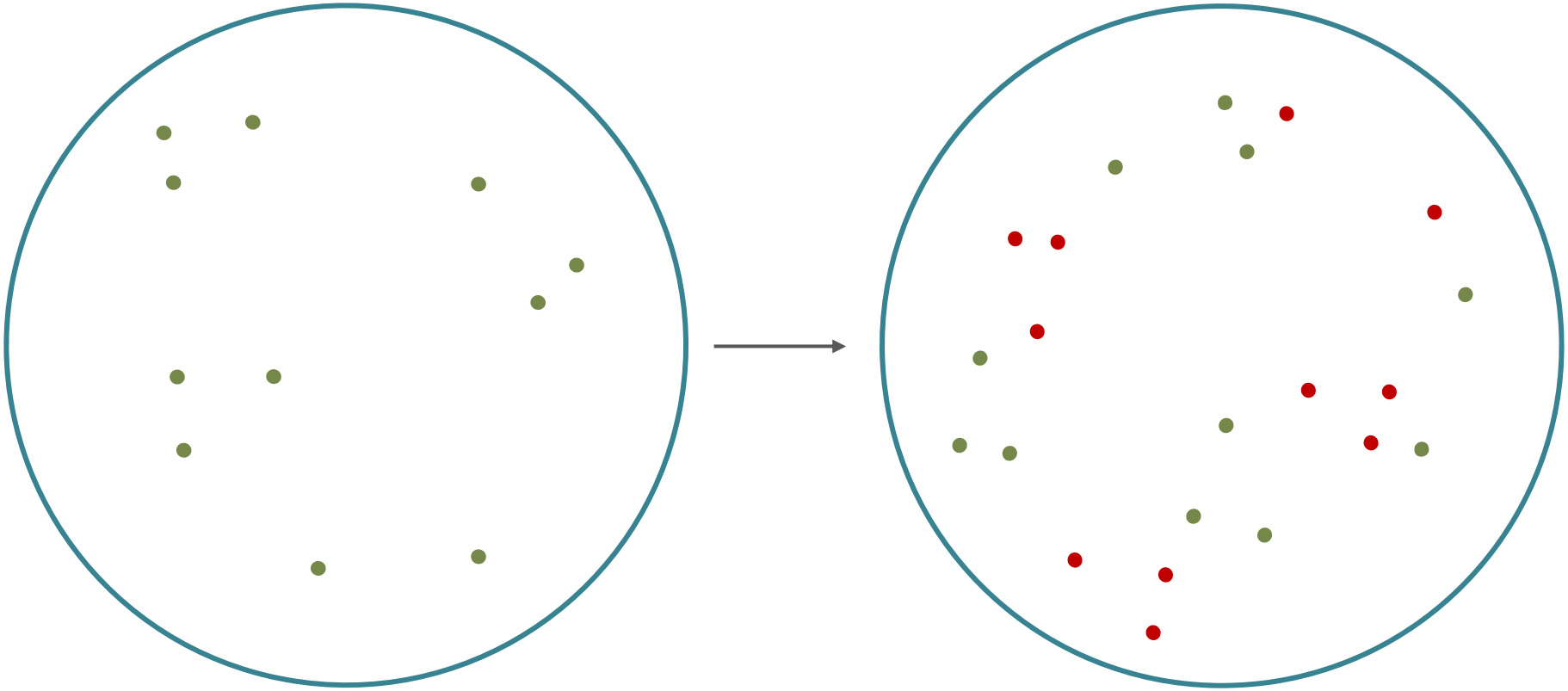
Why does it work?



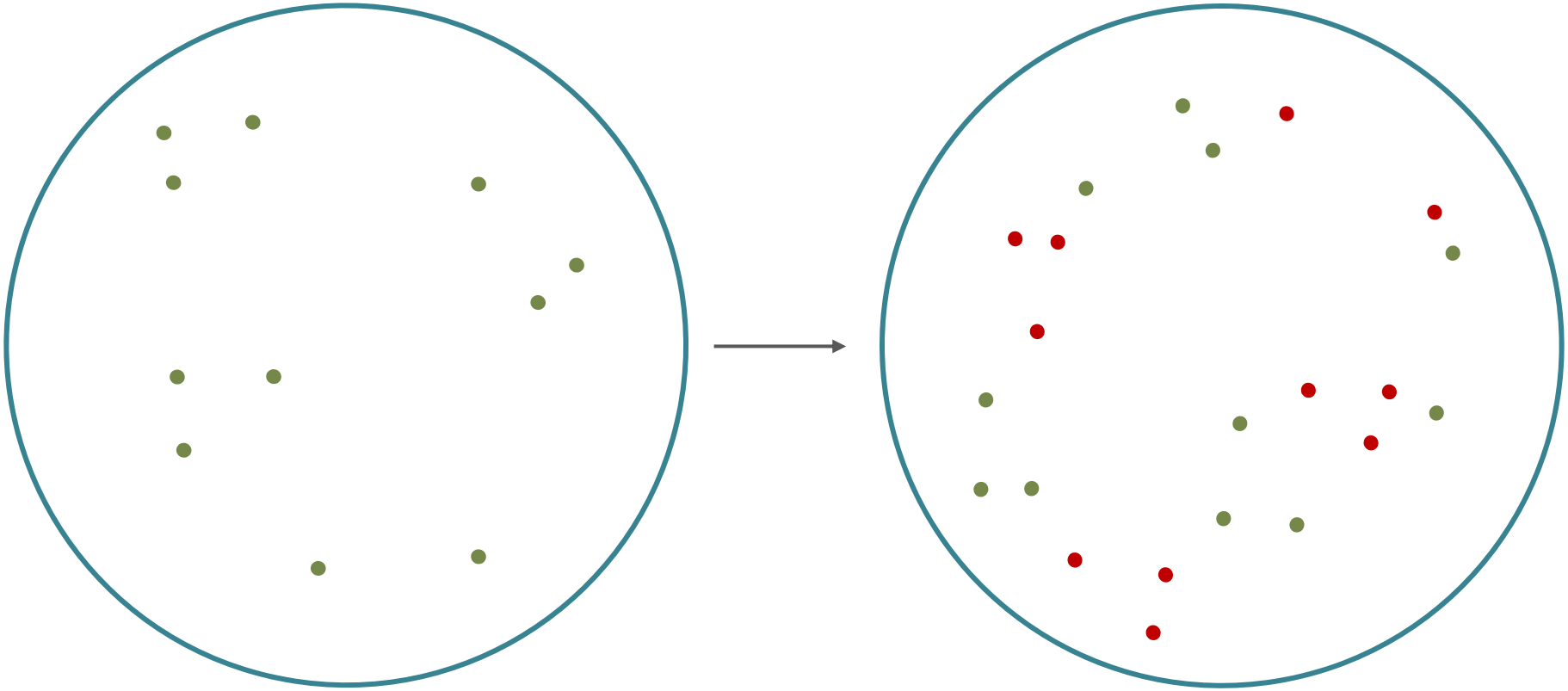
Why does it work?



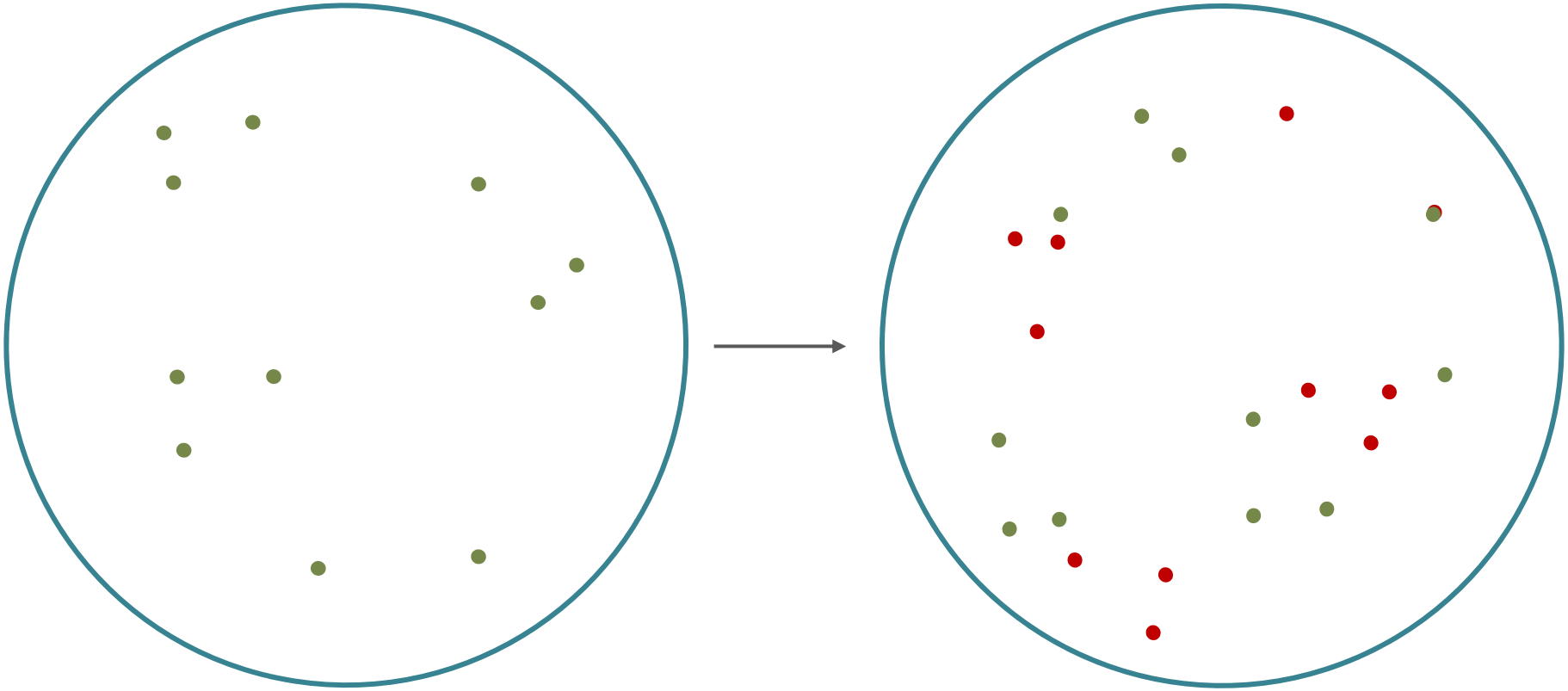
Why does it work?



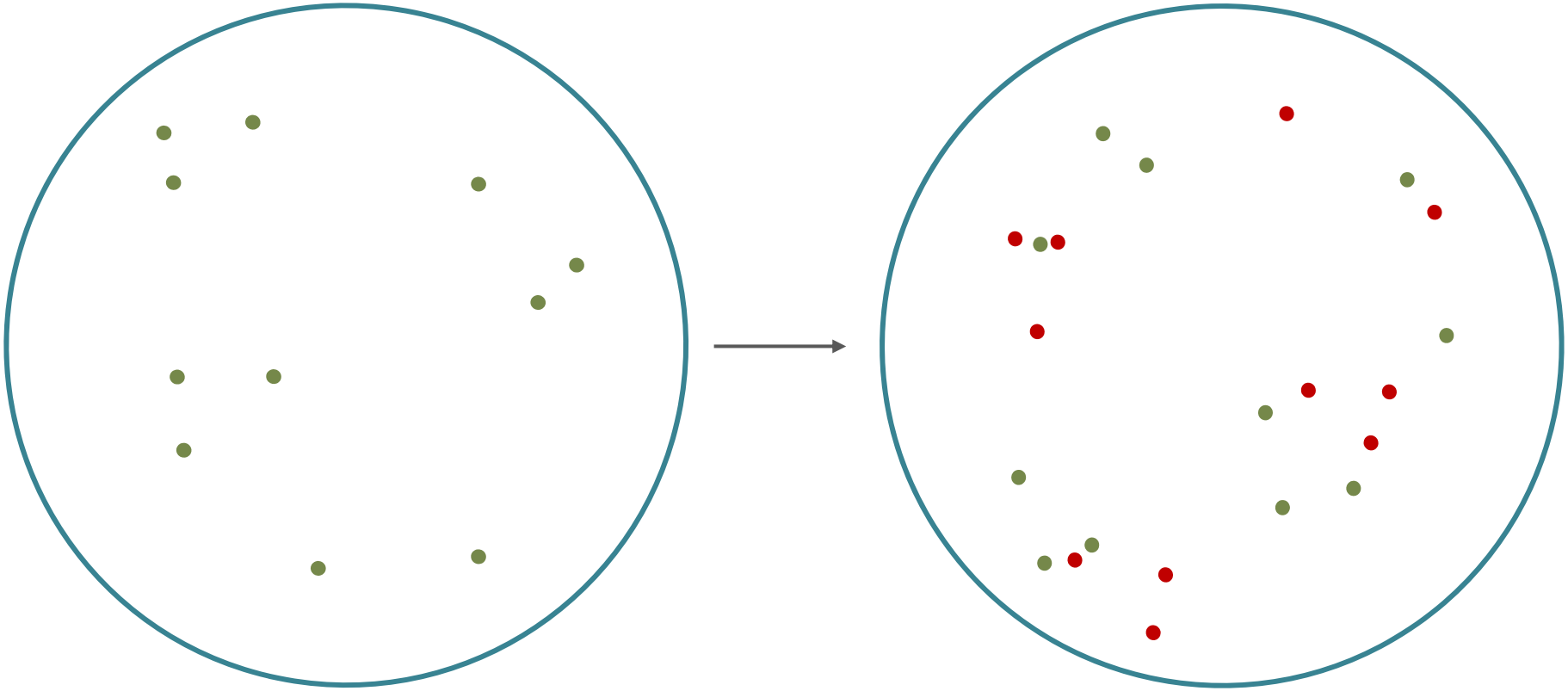
Why does it work?



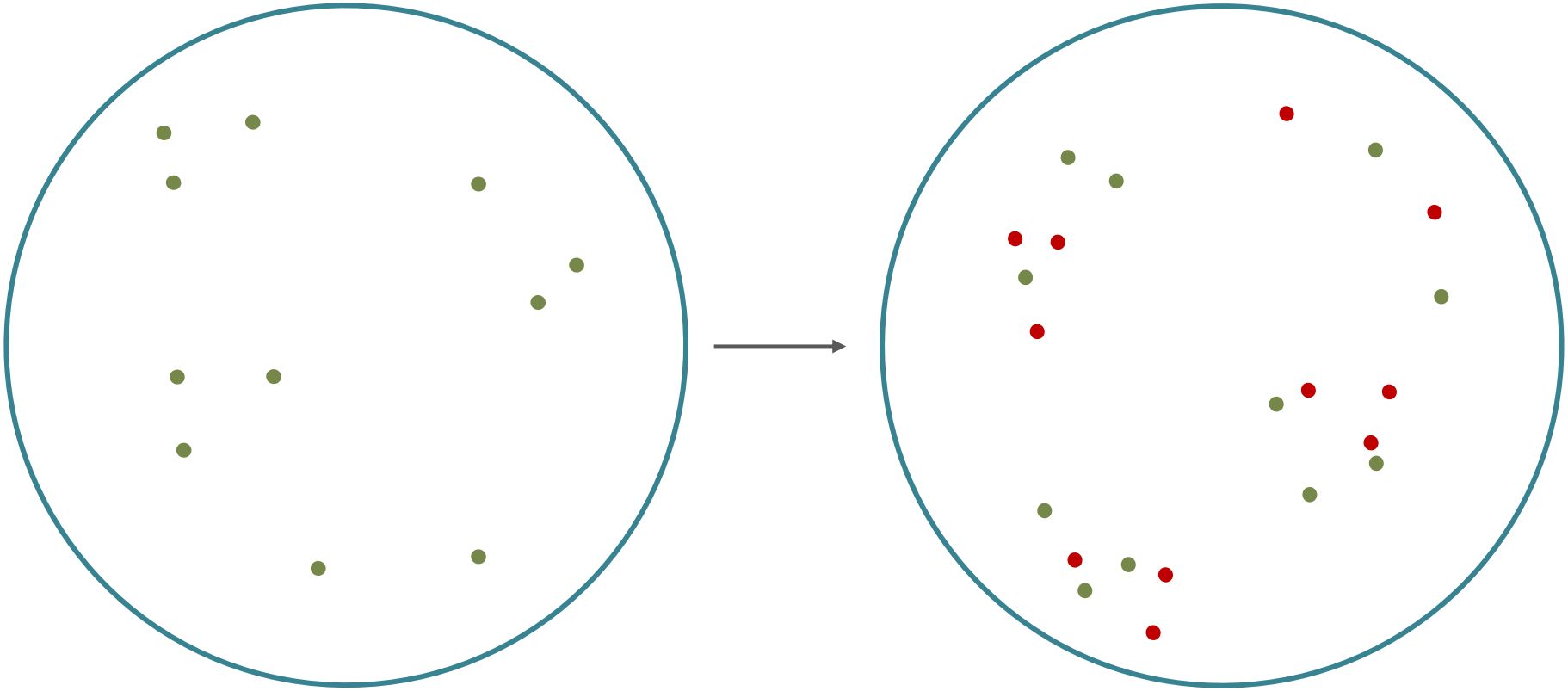
Why does it work?



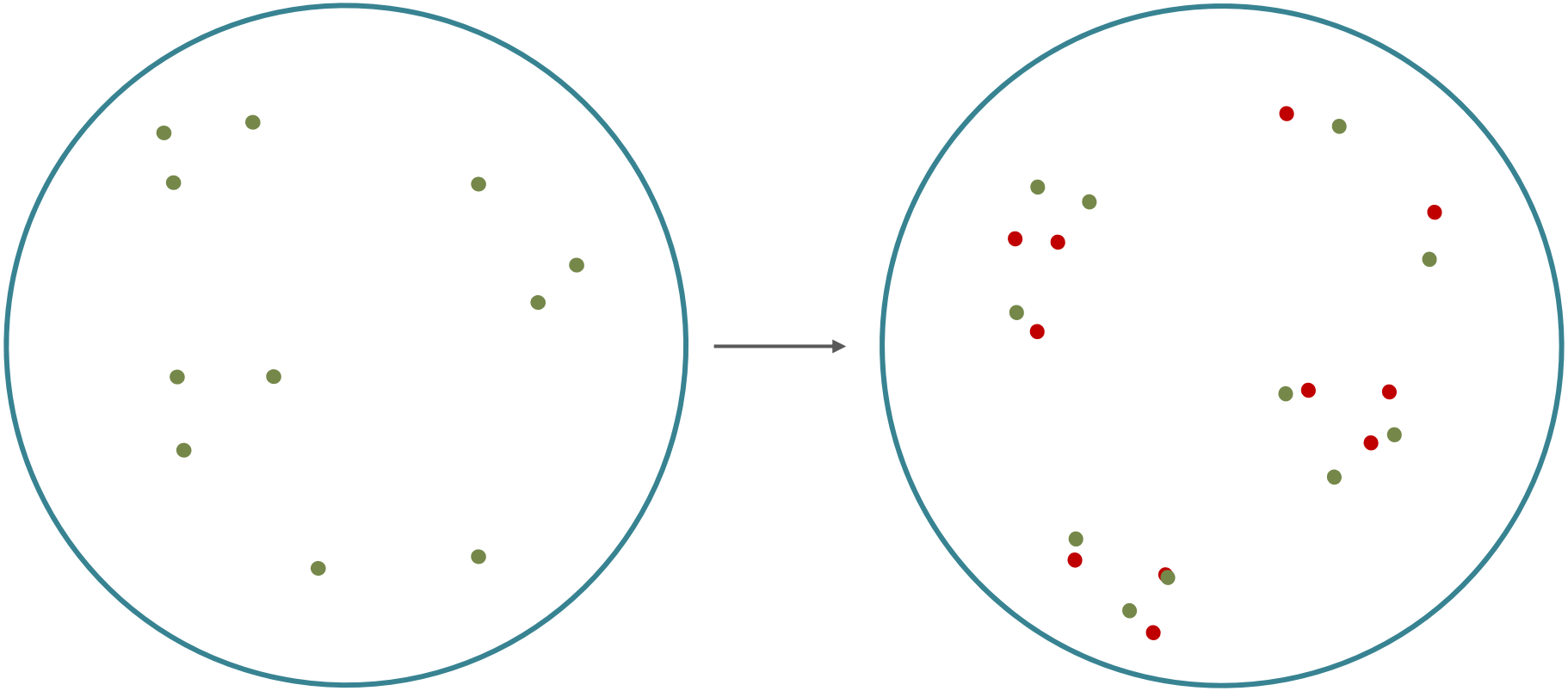
Why does it work?



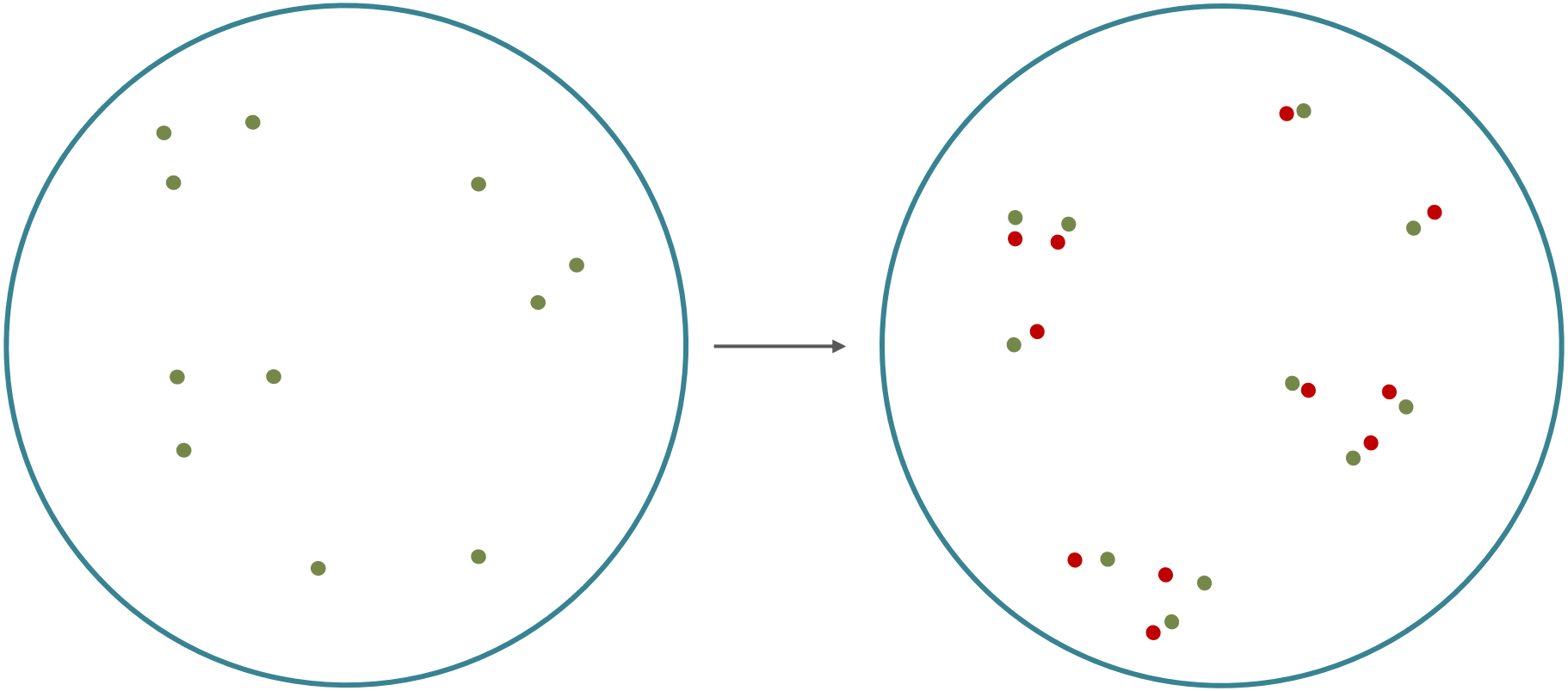
Why does it work?



Why does it work?

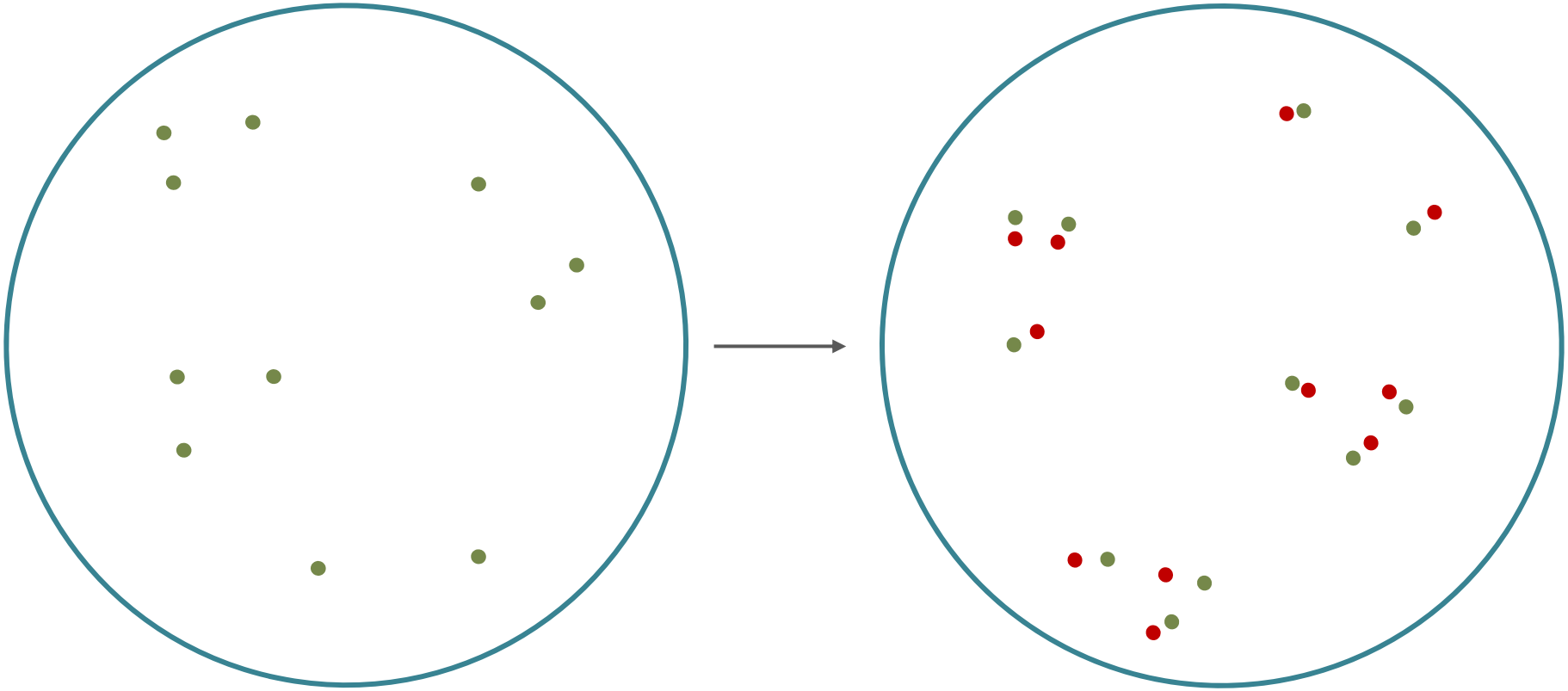


Why does it work?



Why does it work?

Implicit objective: $W^* = \arg\max_W \sum_i \max_j (X_{i*} W) \cdot Z_{j*}$ s.t. $WW^T = W^T W = I$



Why does it work?

Implicit objective: $$W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*} \qquad \text{s.t.} \quad WW^T = W^TW = I$$

Independent from seed dictionary!

Why does it work?

Implicit objective: $$W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*} \qquad \text{s.t.} \quad WW^T = W^TW = I$$

Independent from seed dictionary!

So why do we need a seed dictionary?

Avoid poor local optima!

Next steps

Is there a way we can avoid the seed dictionary?

Would an initial noisy initialization suffice?

Unsupervised experiments (ACL18)

Unsupervised experiments (ACL18)

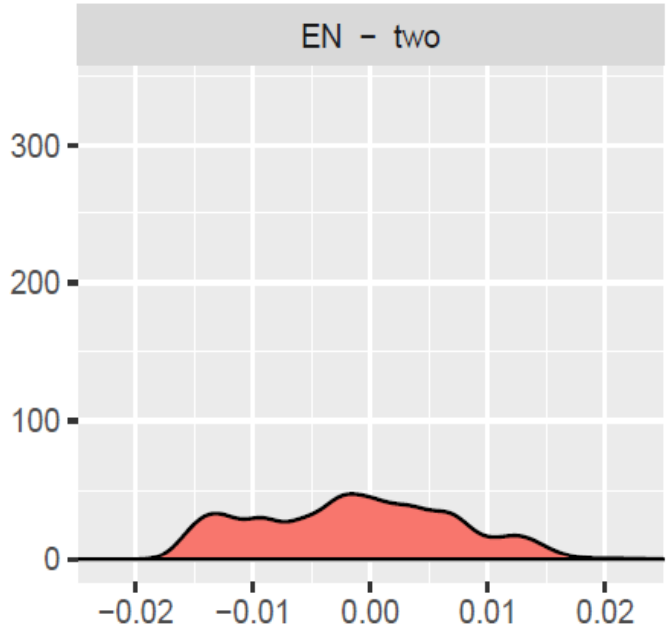
Initial dictionary:

1. Compute intra-language similarity
2. Words which are translations of each other would have analoguous similarity histograms (isometry hyp.)

Unsupervised experiments (ACL18)

Initial dictionary:

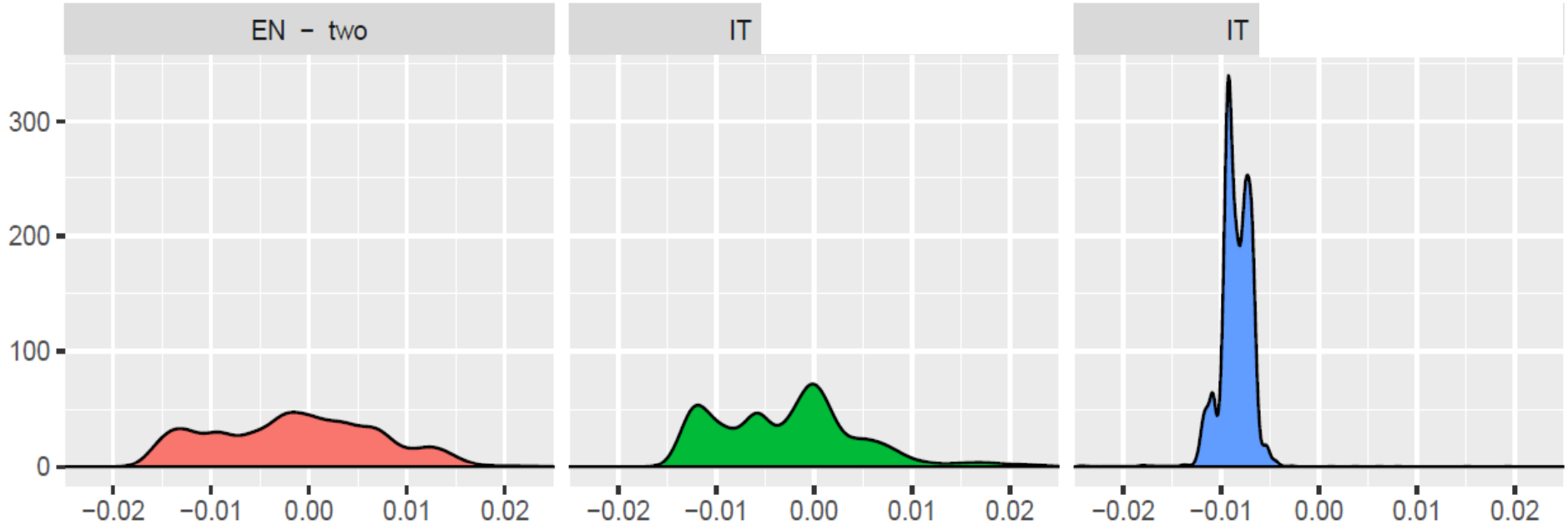
1. Compute intra-language similarity
2. Words which are translations of each other would have analoguous similarity histograms (isometry hyp.)



Unsupervised experiments (ACL18)

Initial dictionary:

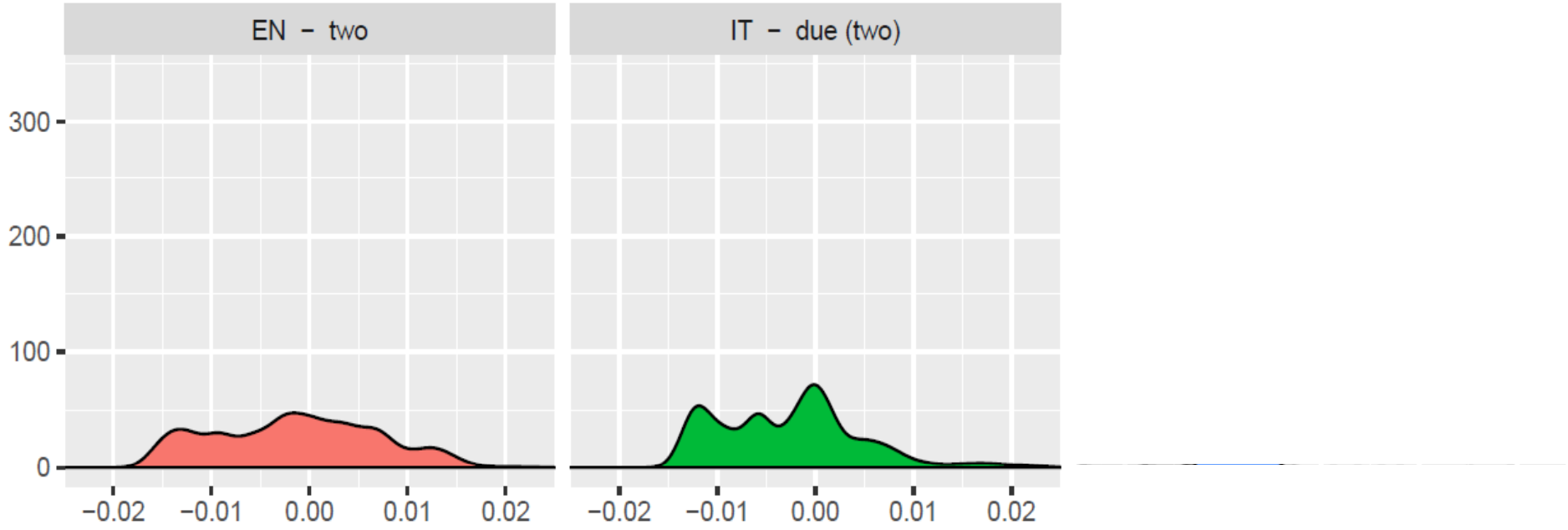
1. Compute intra-language similarity
2. Words which are translations of each other would have analoguous similarity histograms (isometry hyp.)



Unsupervised experiments (ACL18)

Initial dictionary:

1. Compute intra-language similarity
2. Words which are translations of each other would have analoguous similarity histograms (isometry hyp.)



Unsupervised experiments (ACL18)

Initial dictionary:

1. Compute intra-language similarity
2. Words which are translations of each other would have analoguous similarity histograms (isometry hyp.)

It works, but very weak: Accuracy 0.52%

Unsupervised experiments (ACL18)

Initial dictionary:

1. Compute intra-language similarity
2. Words which are translations of each other would have analoguous similarity histograms (isometry hyp.)

It works, but very weak: Accuracy 0.52%

For self-learning to work we had to add:

1. Stochastic dictionary induction
2. Frequency-based vocabulary cut-off
3. Hubness problem: Instead of inducing dictionary with nearest-neighbour use CSLS (Lample et al. 2018)

$$2cos(x, y) - mnn_T(x) - mnn_S(y)$$

$$mnn_T(x) = \frac{1}{K} \sum_{i=1}^{K} cos(x, nn_i)$$

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / **none***
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Artetxe et al. (2918)	45.27	44.13	**32.94**	36.60
None	Zhang et al. (2017)	0.00	0.00	0.01	0.01
	Conneau et al. (2018)	13.55	42.15	0.38	21.23
	Artetxe et al. (2018)	**48.13**	**48.19**	32.63	**37.33**

Conclusions on mapping methods

Conclusions on mapping methods

- Simple self-learning method to train bilingual embedding mappings
- Strong unsupervised results relative to supervised methods
- Implicit optimization objective independent from seed dictionary
- High quality dictionaries:
  Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%
- Shows that languages share “semantic” structure to a large degree

Conclusions on mapping methods

- Simple self-learning method to train bilingual embedding mappings
- Strong unsupervised results relative to supervised methods
- Implicit optimization objective independent from seed dictionary
- High quality dictionaries:
  Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%
- Shows that languages share “semantic” structure to a large degree

- Later work has shown lack of robustness
  on some language pairs / conditions (Vulic et al. 2019)

Conclusions on mapping methods

- Simple self-learning method to train bilingual embedding mappings
- Strong unsupervised results relative to supervised methods
- Implicit optimization objective independent from seed dictionary
- High quality dictionaries:
  Manual analysis shows that real accuracy > 60%
  High frequency words up to 80%
- Shows that languages share “semantic” structure to a large degree

- Later work has shown lack of robustness
  on some language pairs / conditions (Vulic et al. 2019)

- Full reproducibility (including datasets):
  `https://github.com/artetxem/vecmap`

References: cross-lingual mappings

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. **Generalizing and Improving Bilingual Word Embedding Mappings with a Multi-Step Framework of Linear Transformations**. In *AAAI-2018.*

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. **Learning bilingual word embeddings with (almost) no bilingual data**. In *ACL-2017*.

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. **A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings**. In *ACL-2018*.

Outline

- Initial idea: Bilingual embedding mappings
  - *Introduction embeddings*
  - *Bilingual embedding mappings (AAAI18)*
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions of bilingual embedding mappings*
- Unsupervised neural machine translation
  - *Introduction to supervised NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Self-learning with better initializations (ACL19)*
  - *Conclusions*

Introduction to supervised NMT

Introduction to supervised NMT

- Given pairs of sentences with known translation ($x_1 \ldots x_n$, $y_1 \ldots y_m$)

  This is my dearest dog </s>

  Este es mi perro preferido </s>

Introduction to supervised NMT

- Given pairs of sentences with known translation ($x_1 \ldots x_n$, $y_1 \ldots y_m$)

  This is my dearest dog </s>

  Este es mi perro preferido </s>

- Train an **encoder** based on Transformers

  return all hidden states, encoding input $x_1 \ldots x_n$

Introduction to supervised NMT

- Given pairs of sentences with known translation ($x_1 \ldots x_n$, $y_1 \ldots y_m$)

  This is my dearest dog </s>

  Este es mi perro preferido </s>

- Train an **encoder** based on Transformers

  return all hidden states, encoding input $x_1 \ldots x_n$

- Train a **decoder** based on Transformers

  - based on hidden states and last word in translation y_{i-1}
  - plus an **attention** mechanism
  - classifier guesses next word y_i

Introduction to supervised NMT

- Given pairs of sentences with known translation ($x_1 \ldots x_n$, $y_1 \ldots y_m$)

  This is my dearest dog </s>

  Este es mi perro preferido </s>

- Train an **encoder** based on Transformers

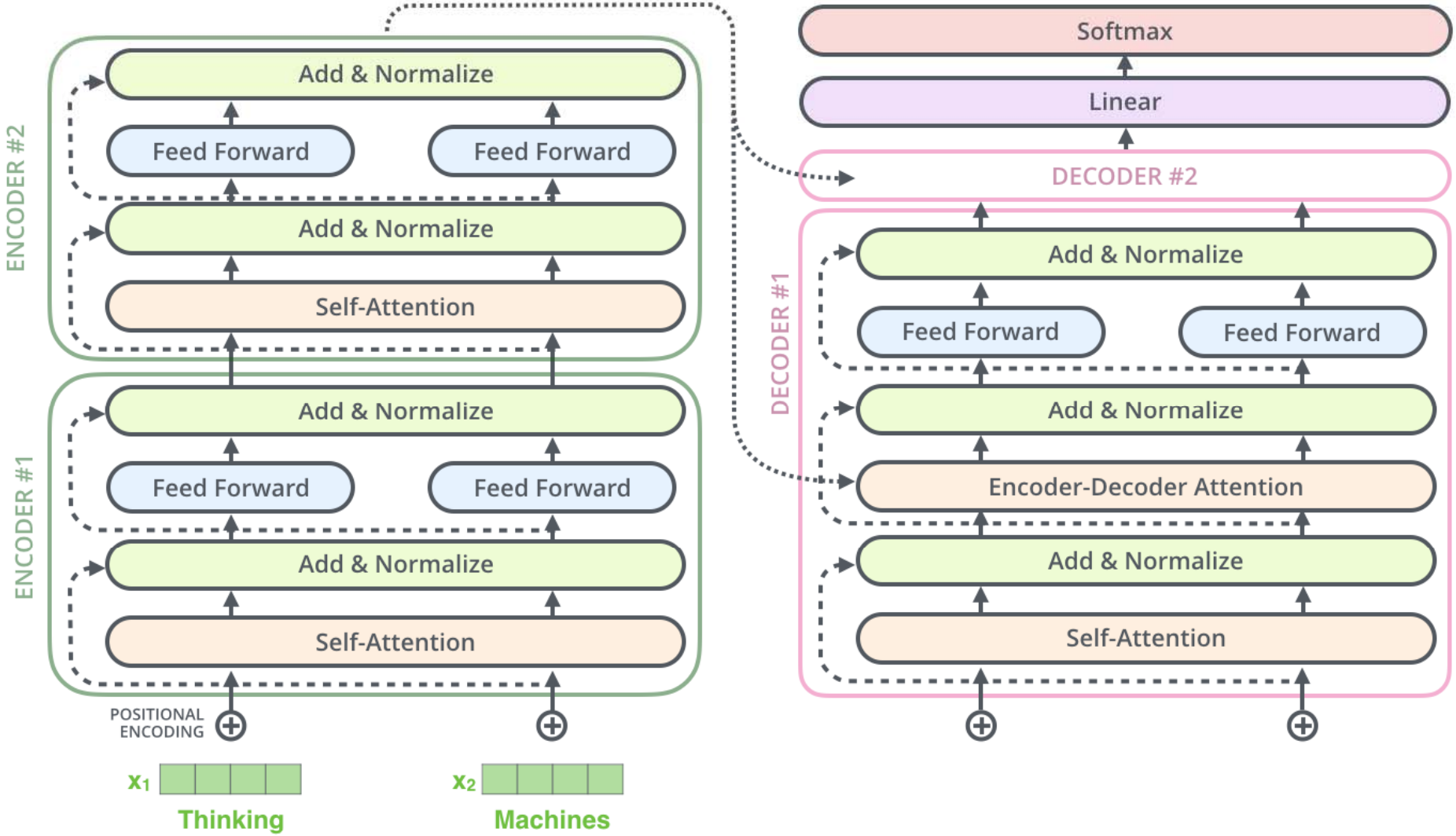
  return all hidden states, encoding input $x_1 \ldots x_n$

- Train a **decoder** based on Transformers

  - based on hidden states and last word in translation y_{i-1}
  - plus an **attention** mechanism
  - classifier guesses next word y_i

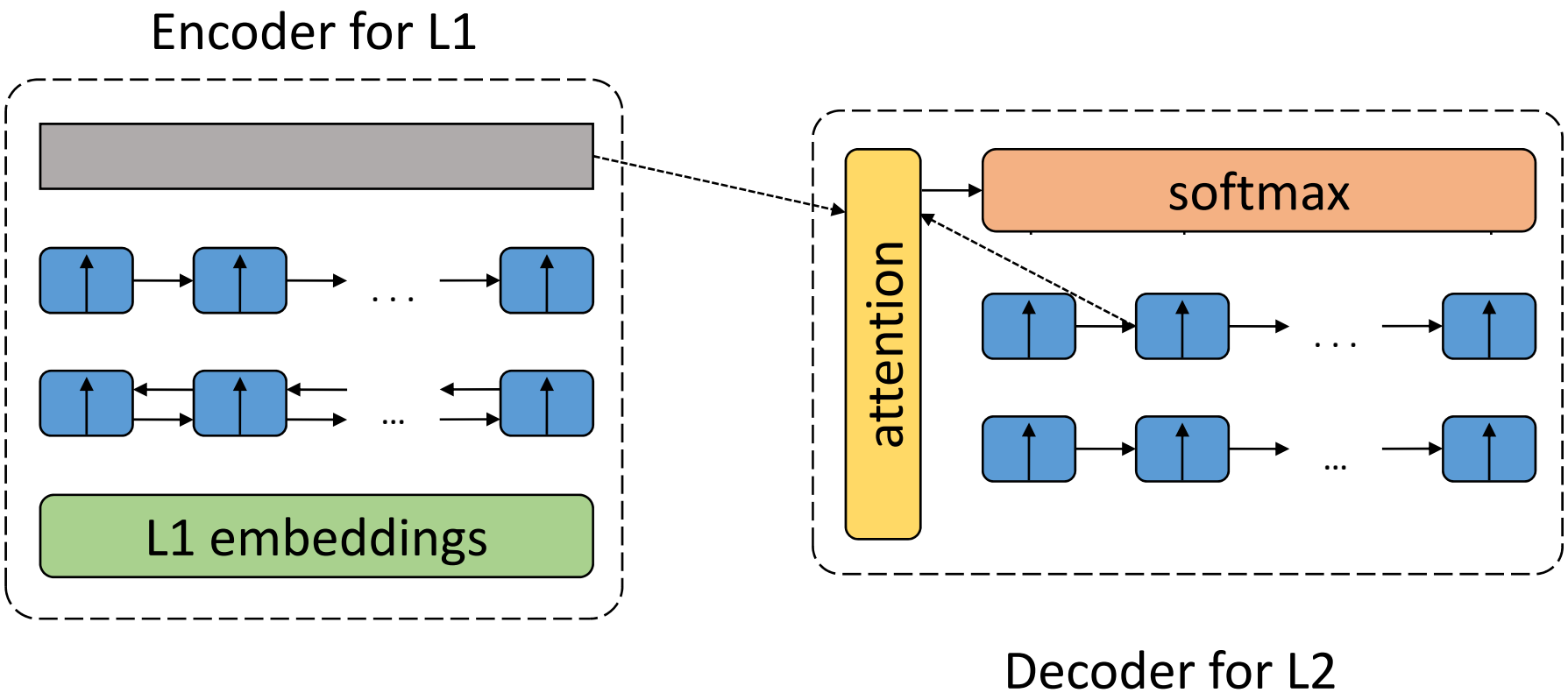
End-to-end training

Introduction to supervised NMT



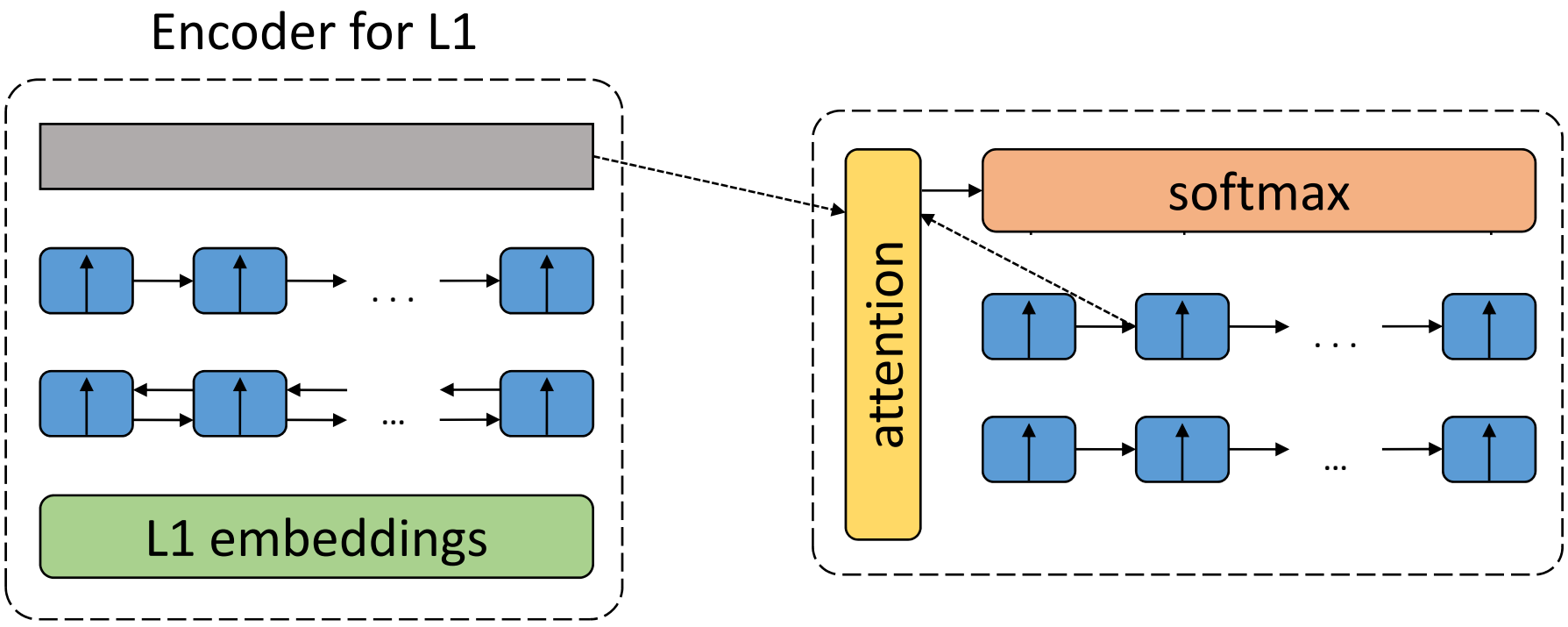
Source: http://jalammar.github.io/illustrated-transformer/

Introduction to supervised NMT



Introduction to supervised NMT

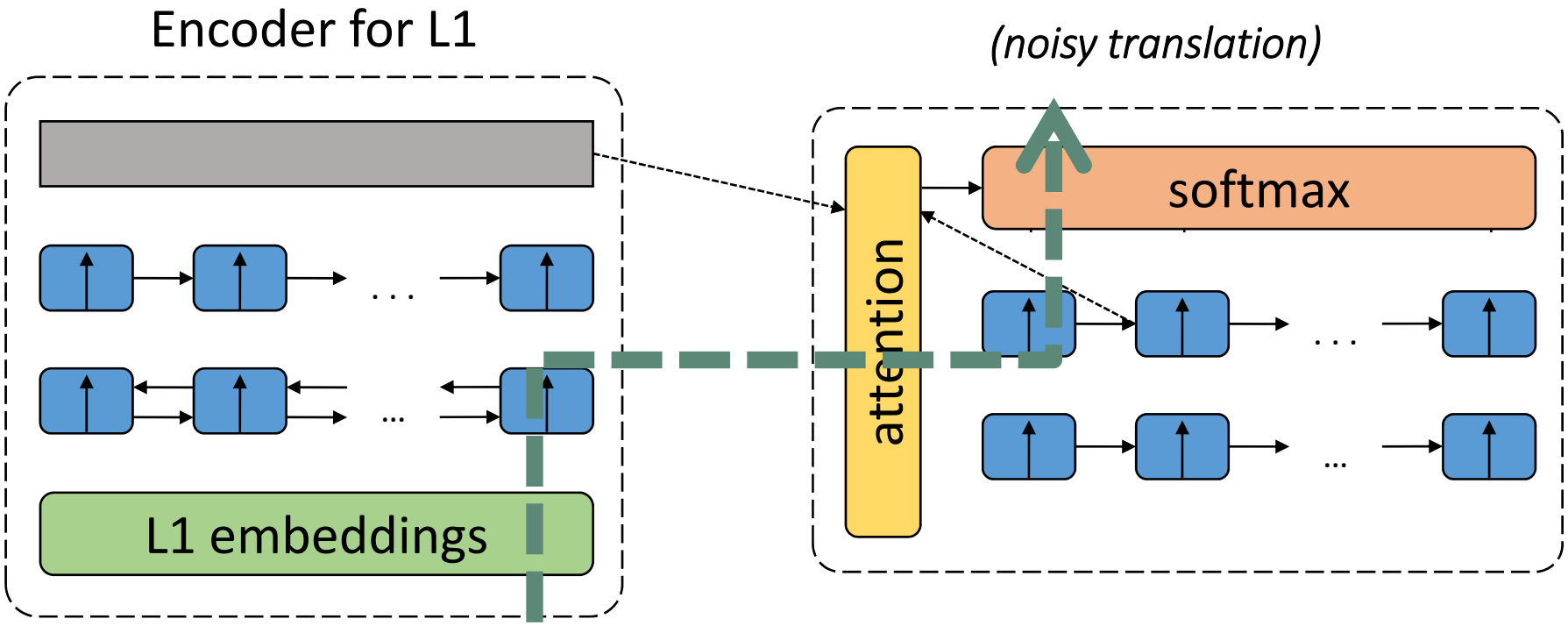
Training



Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Introduction to supervised NMT

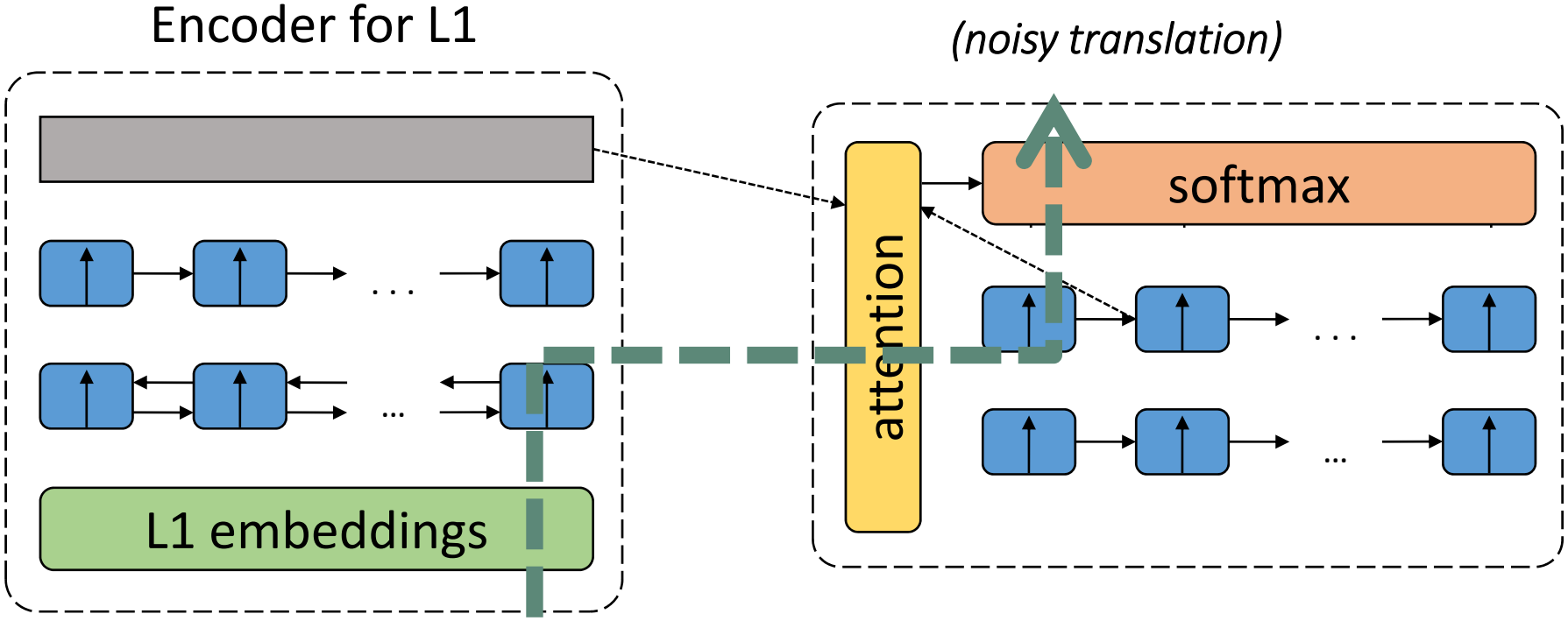
Training



Introduction to supervised NMT

Training

There was a shooting in Los Angeles International Airport.



Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised NMT (ICLR18)

- Given that we **can represent words in two languages in the same embedding space** without bilingual resources…

  … what can we do?

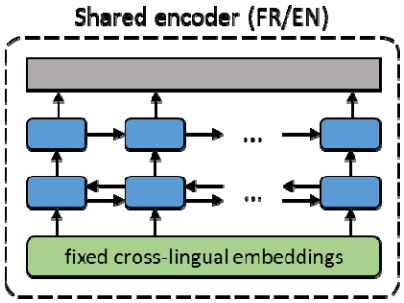
Unsupervised NMT (ICLR18)

- Given that we **can represent words in two languages in the same embedding space** without bilingual resources…

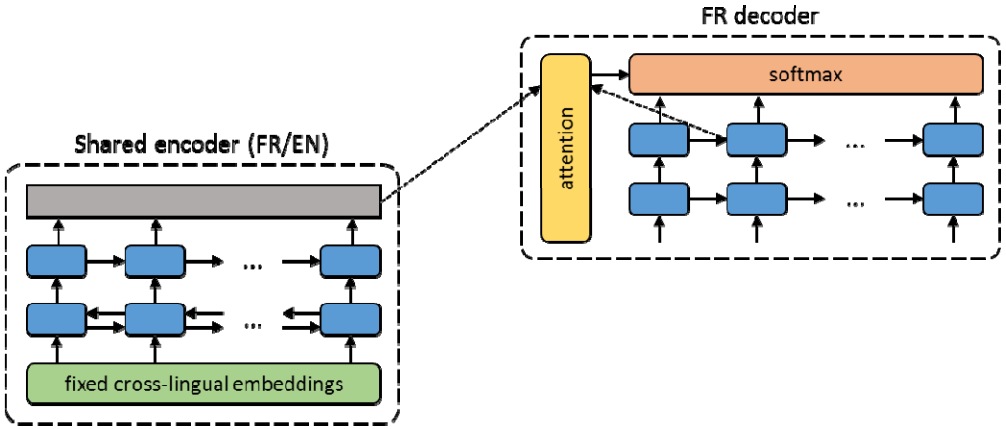
  … what can we do?

- We change the **architecture of the NMT system**:
  - Handle both directions together (L1 -> L2, L2 -> L1)
  - Shared encoder for the two languages (E)
  - Two decoders for each language (D1, D2)
  - Initialize with previously learned cross-lingual embeddings , and freeze

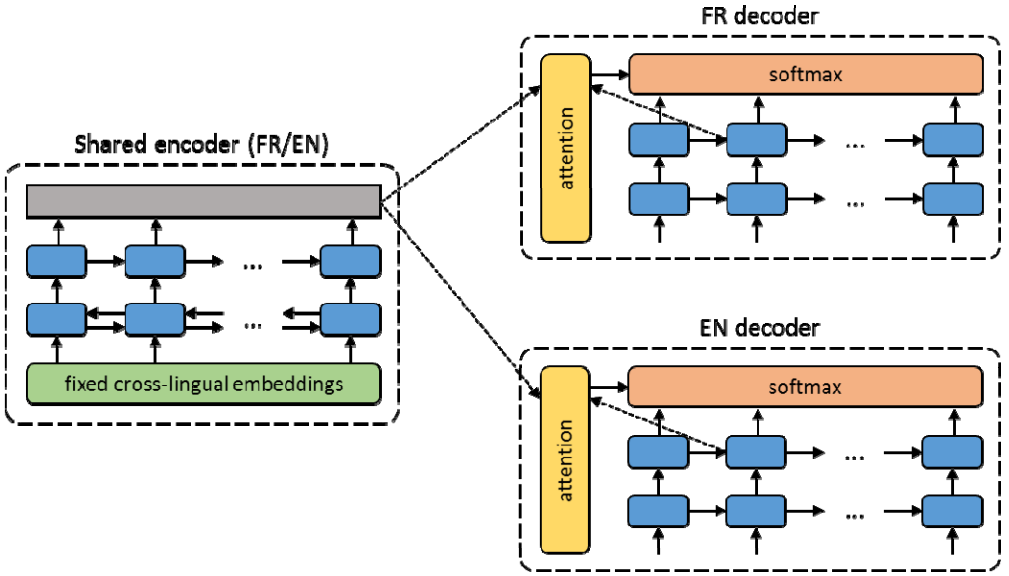
Unsupervised NMT (ICLR18)



Unsupervised NMT (ICLR18)

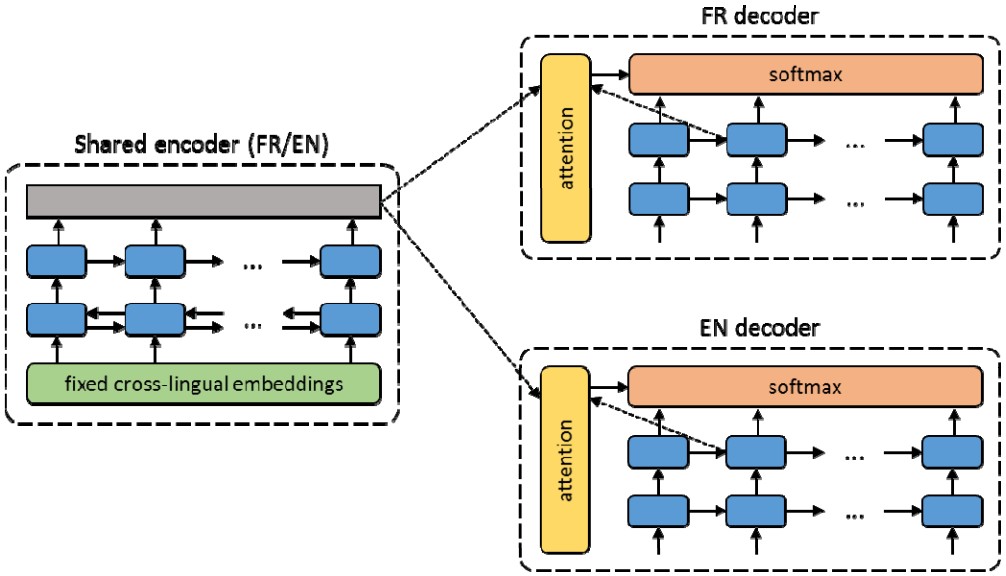


Unsupervised NMT (ICLR18)



Unsupervised NMT (ICLR18)

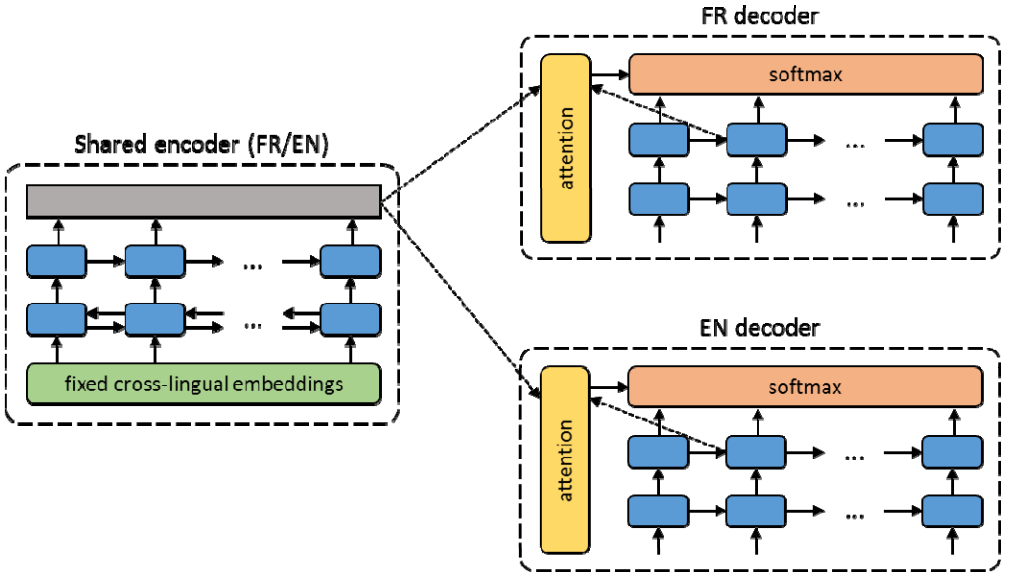
Training



Unsupervised NMT (ICLR18)

Training

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

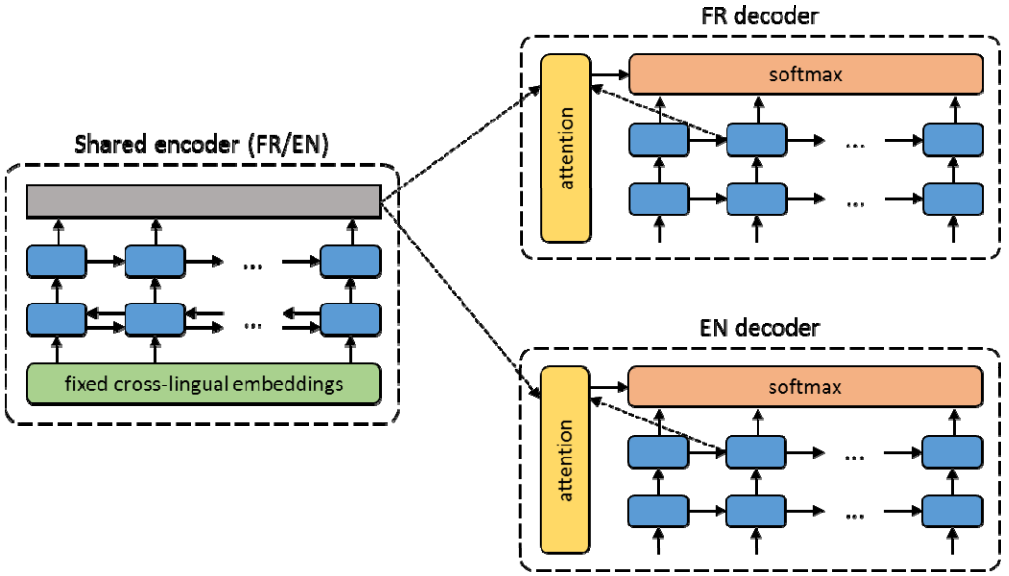


Unsupervised NMT (ICLR18)

Training

- Supervised

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

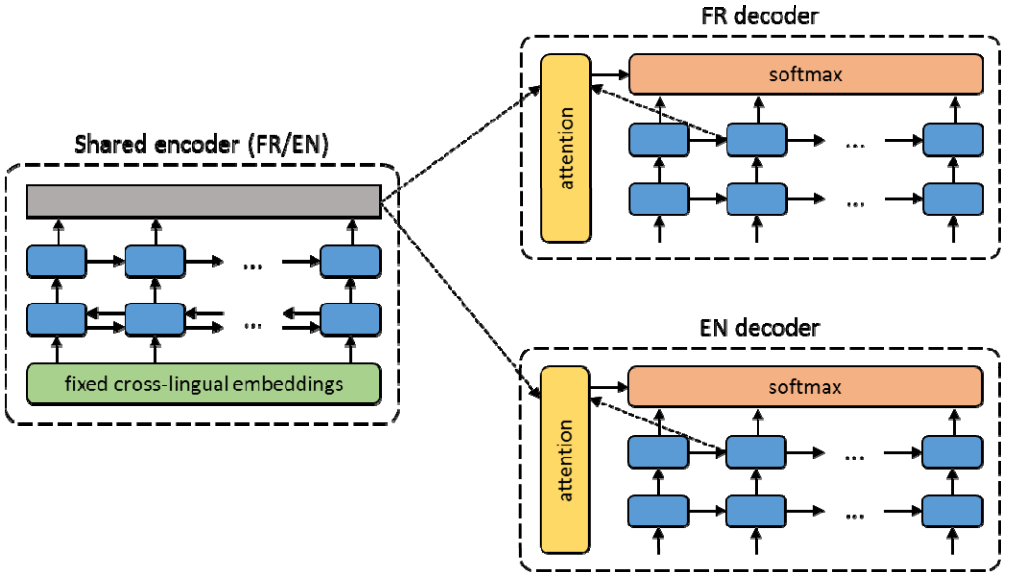


Unsupervised NMT (ICLR18)

Training

- Supervised

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

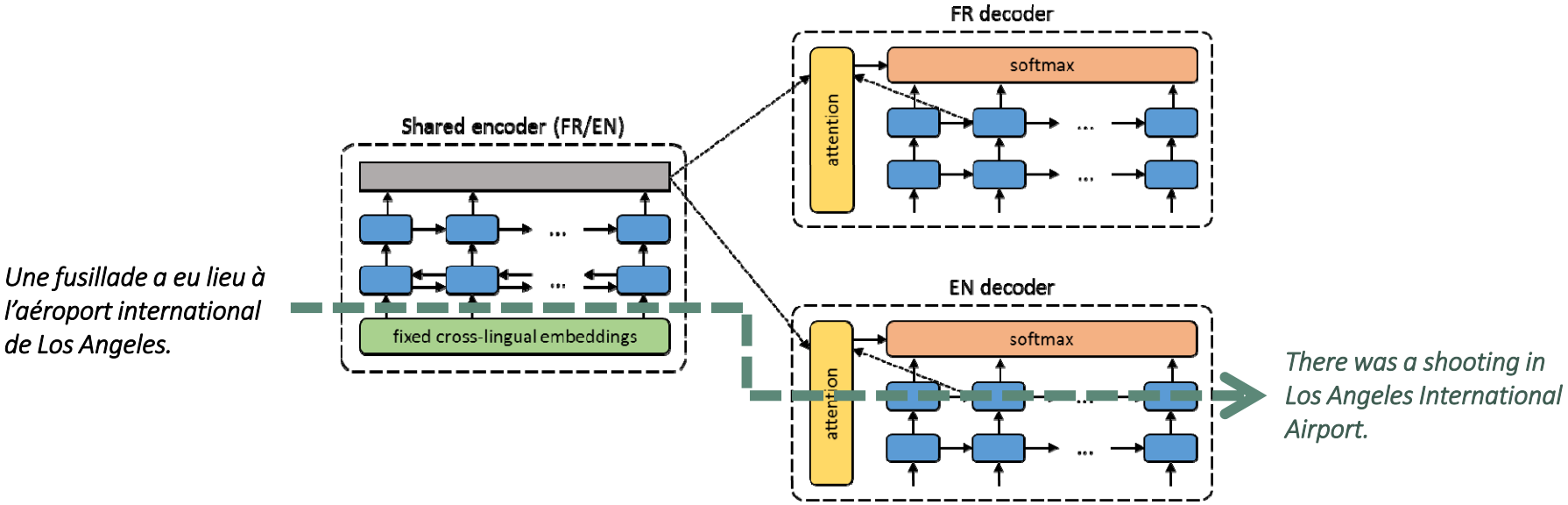


There was a shooting in Los Angeles International Airport.

Unsupervised NMT (ICLR18)

Training

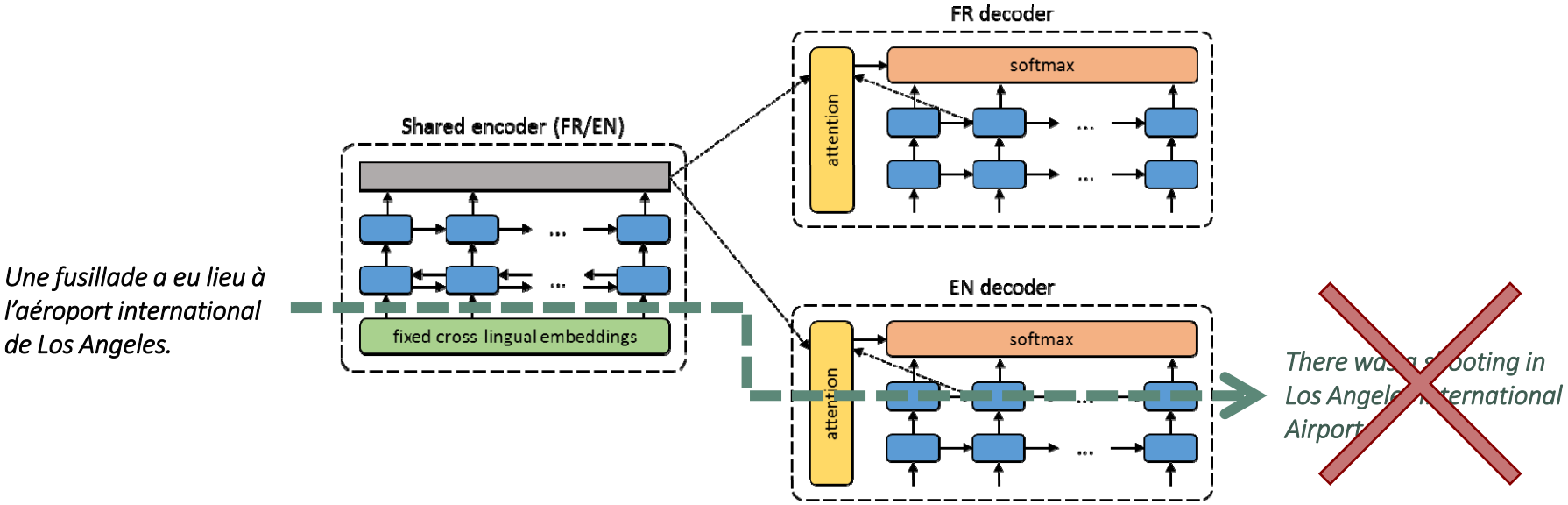
- Supervised



Unsupervised NMT (ICLR18)

Training

- Supervised

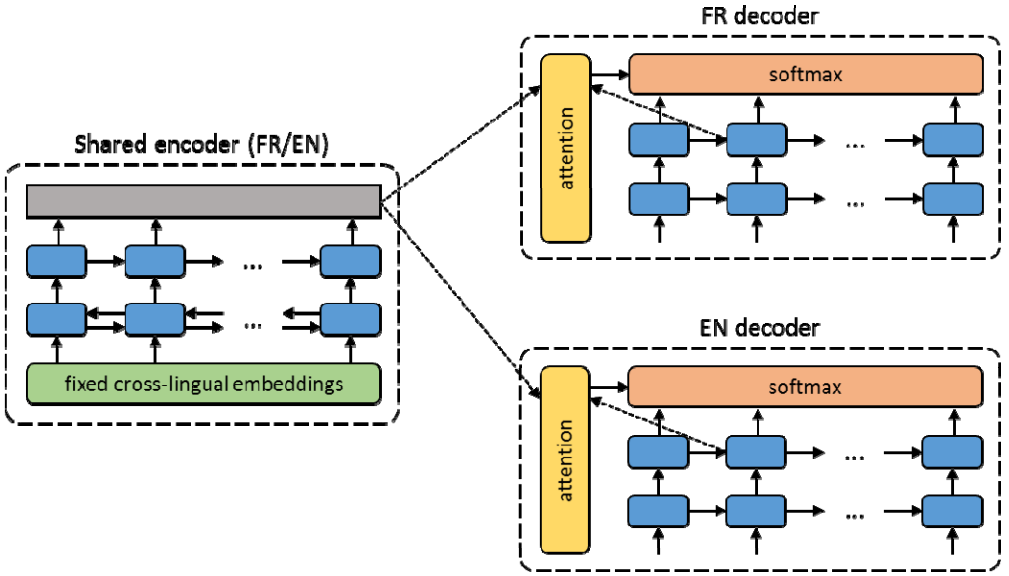


Unsupervised NMT (ICLR18)

Training

- ~~Supervised~~

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

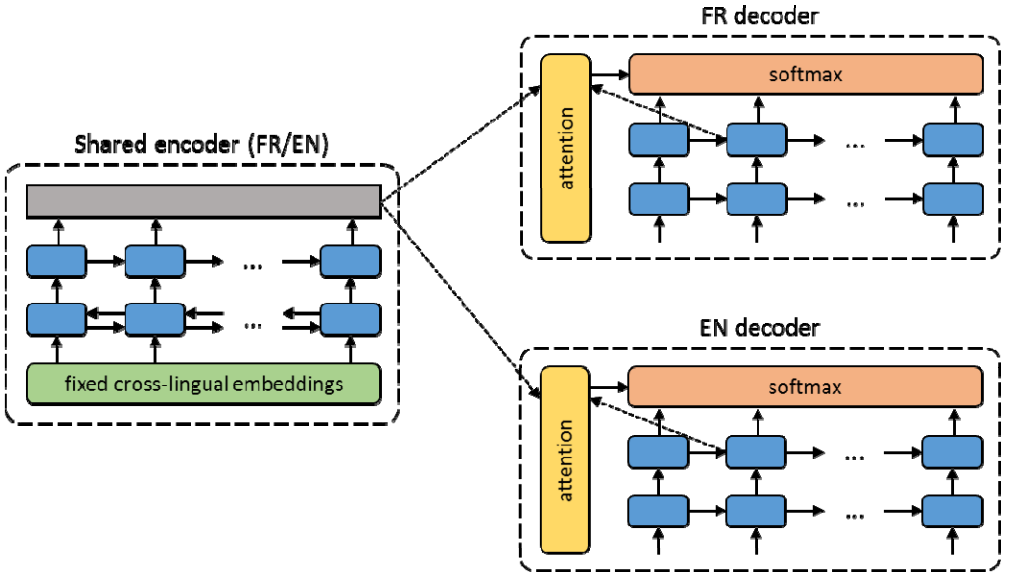


Unsupervised NMT (ICLR18)

Training

- ~~Supervised~~
- Autoencoder

Une fusillade a eu lieu à l'aéroport international de Los Angeles.

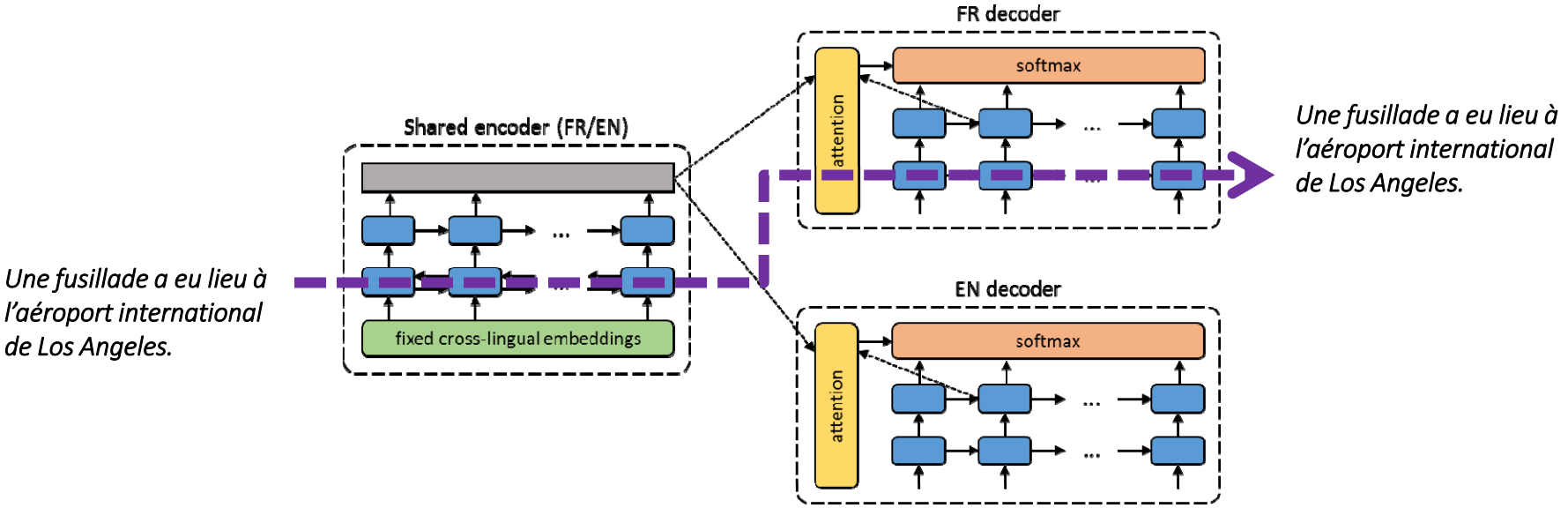


Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised NMT (ICLR18)

Training

- ~~Supervised~~
- Autoencoder

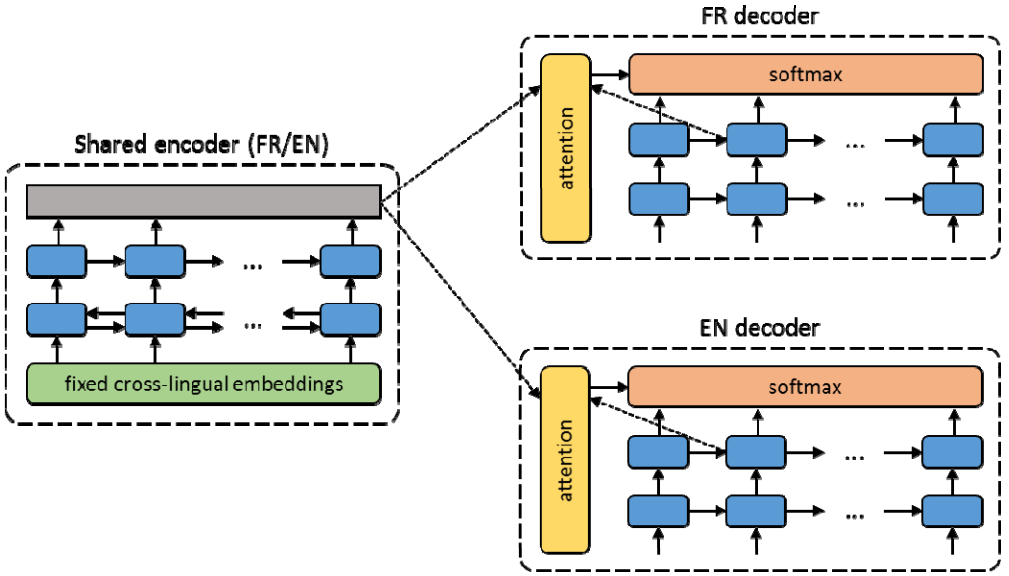


Unsupervised NMT (ICLR18)

Training

- ~~Supervised~~
- Denoising Autoencoder

Une lieu fusillade a eu à l'aéroport de Los international Angeles.

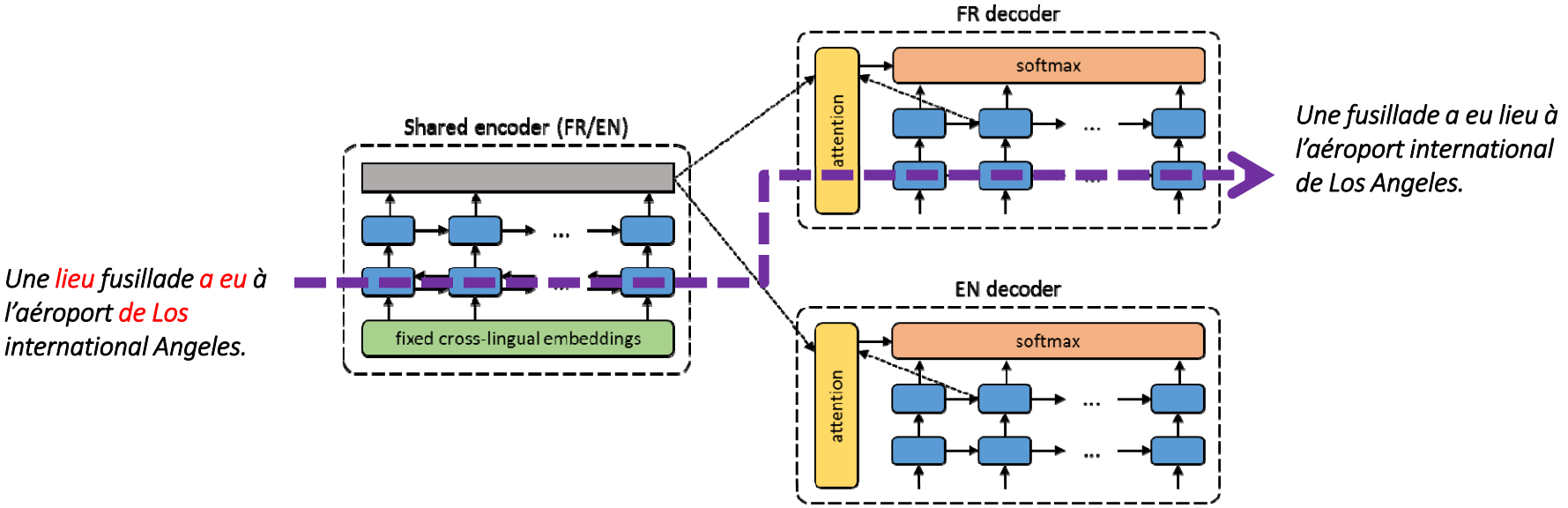


Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised NMT (ICLR18)

Training

- ~~Supervised~~
- Denoising Autoencoder

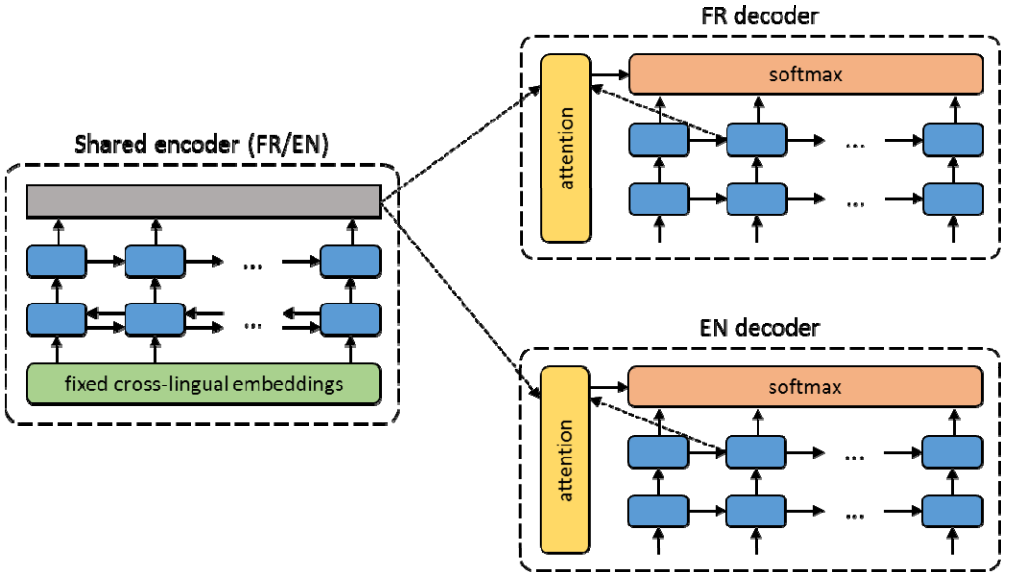


Unsupervised NMT (ICLR18)

Training

- ~~Supervised~~
- Denoising Autoencoder

There a shooting was in Airport Los Angeles International.

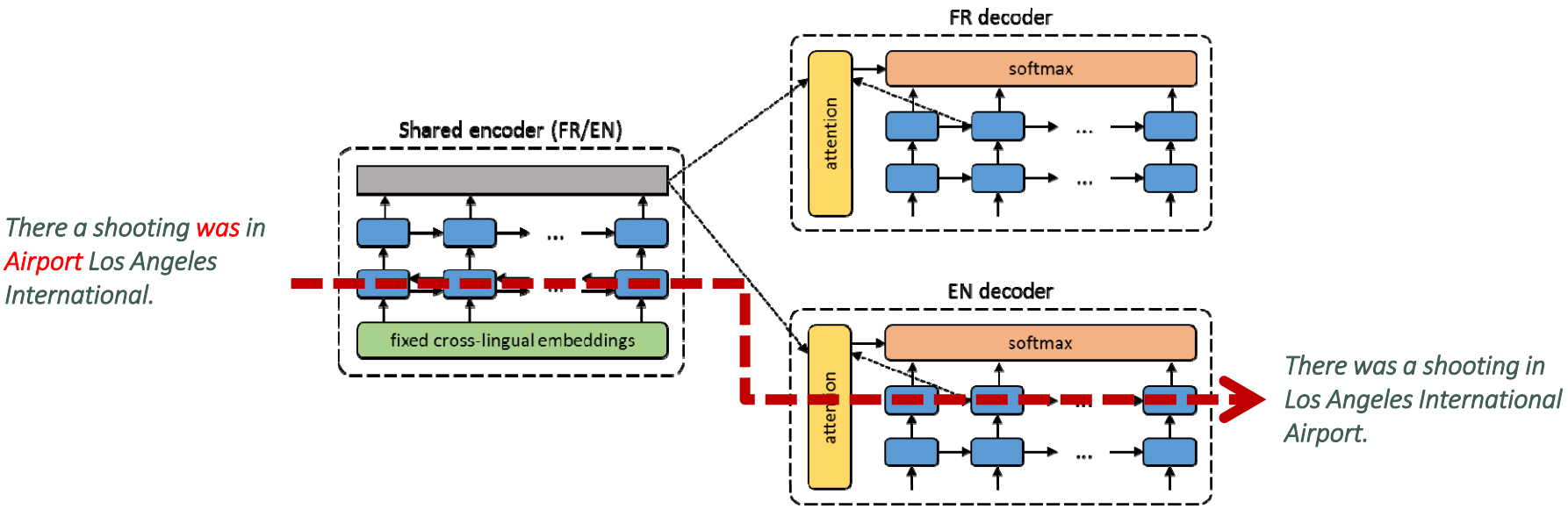


There was a shooting in Los Angeles International Airport.

Unsupervised NMT (ICLR18)

Training

- ~~Supervised~~
- Denoising Autoencoder



Unsupervised NMT (ICLR18)

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

	FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE

Unsupervised NMT (ICLR18)

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

	FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
*Nearest neighbor	9.9	6.3	7.2	4.4		
*Denoising	7.3	5.33	3.64	2.4		

Unsupervised NMT (ICLR18)

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

	FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
*Nearest neighbor	9.9	6.3	7.2	4.4		
*Denoising	7.3	5.33	3.64	2.4		

What else?

Self-learning:

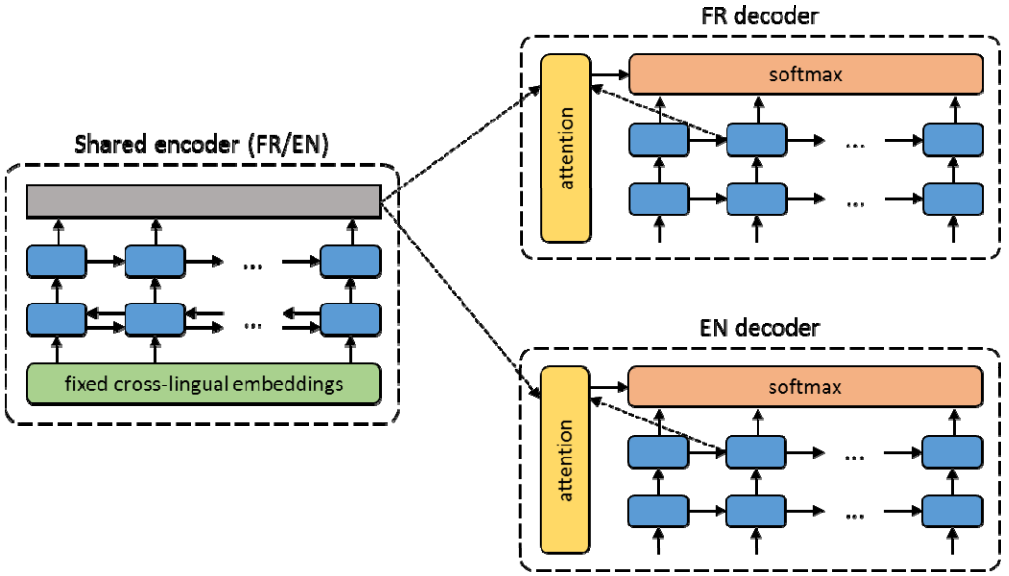
Initialize with denoising, apply back-translation several times

Unsupervised NMT (ICLR18)

Training

- ~~Supervised~~
- Denoising
- Backtranslation

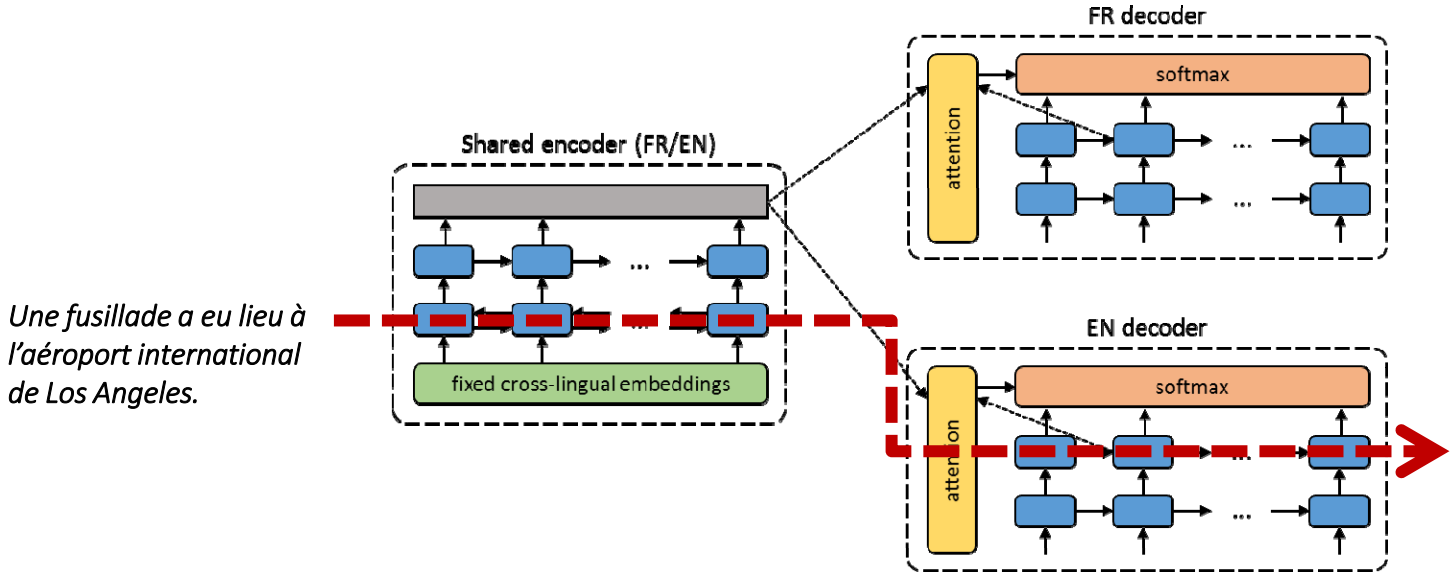
Une fusillade a eu lieu à l'aéroport international de Los Angeles.



Unsupervised NMT (ICLR18)

Training

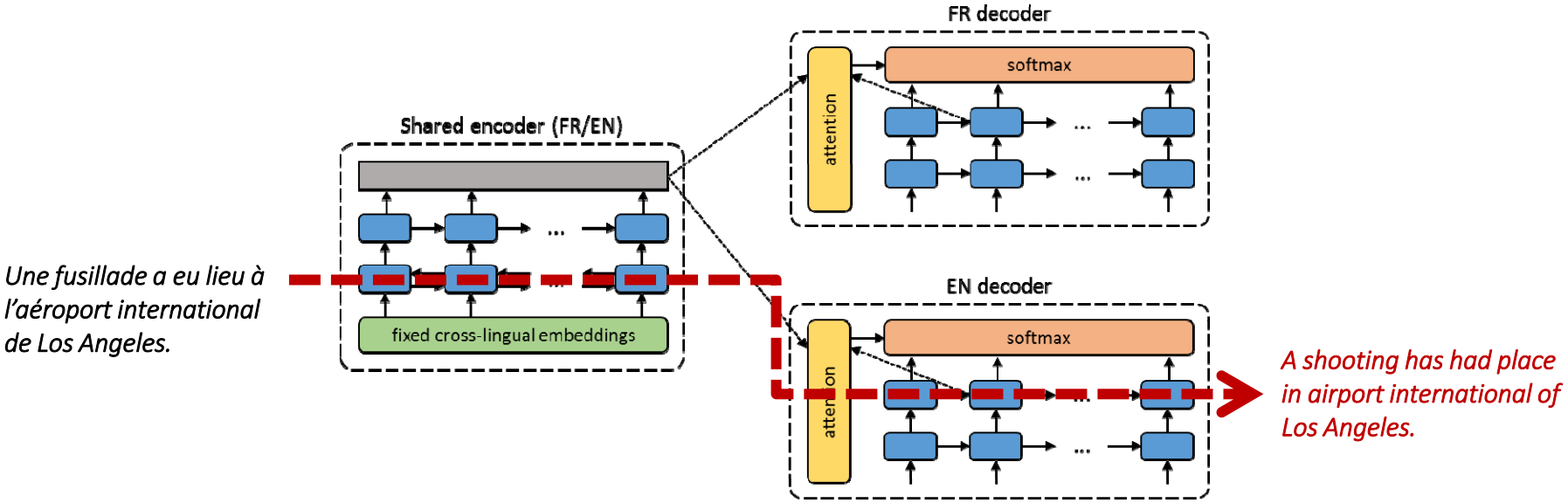
- ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised NMT (ICLR18)

Training

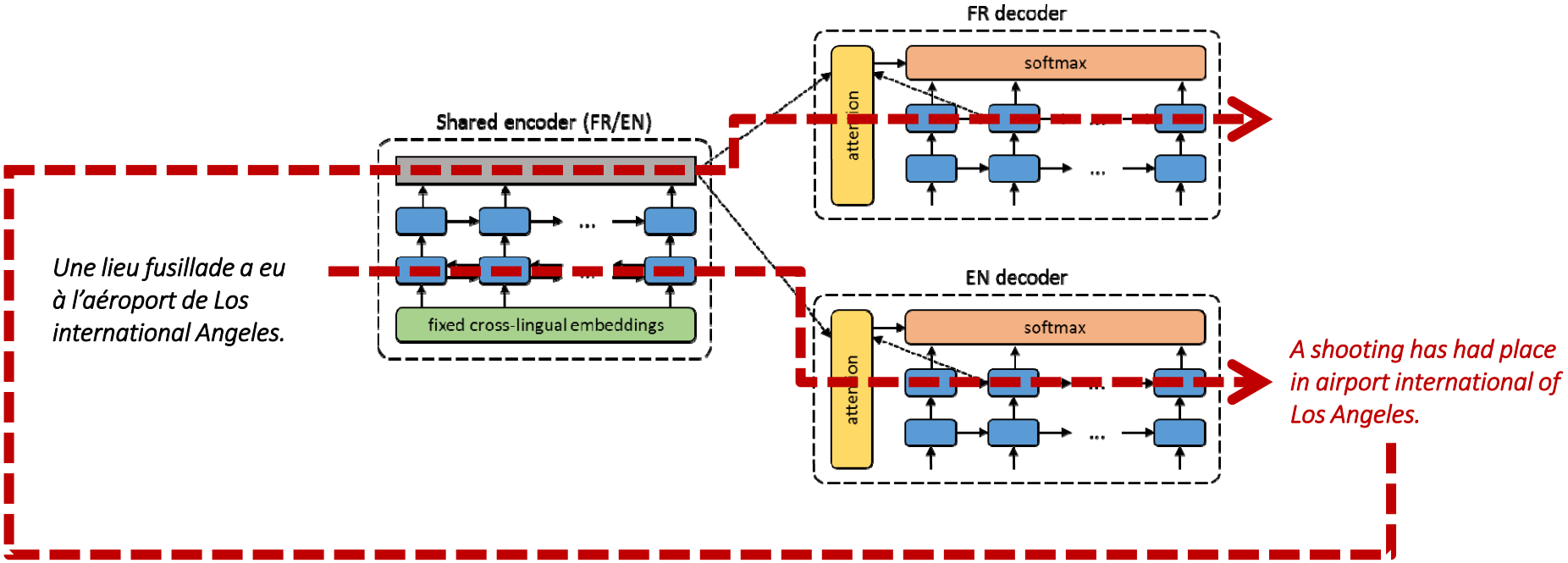
- ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised NMT (ICLR18)

Training

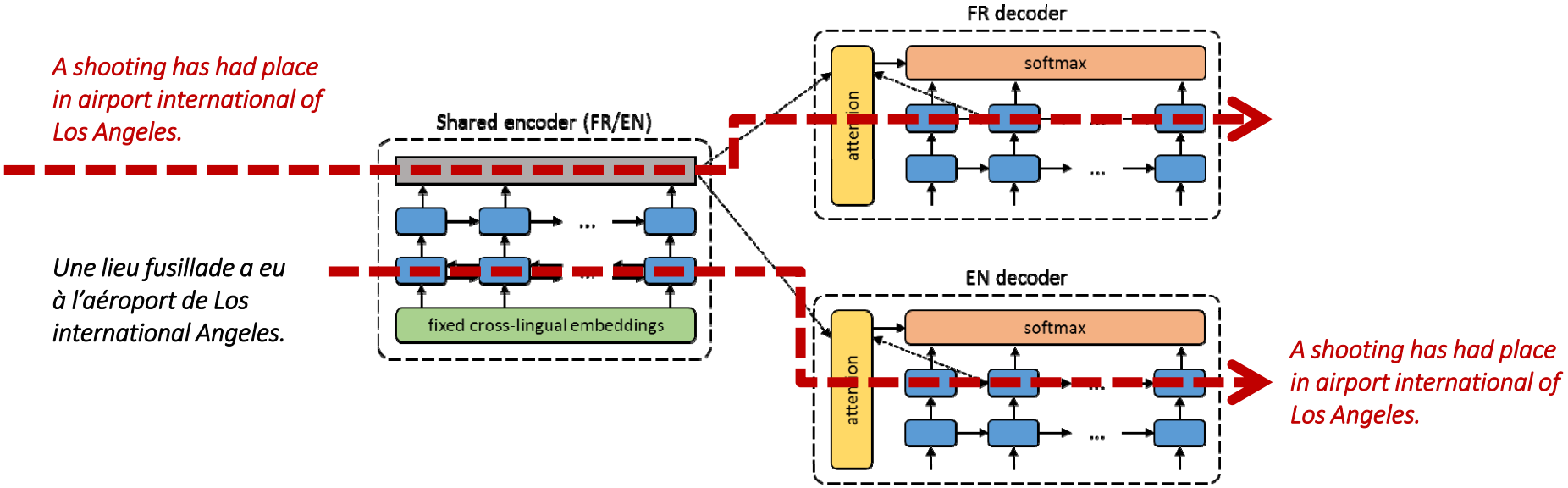
- ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised NMT (ICLR18)

Training

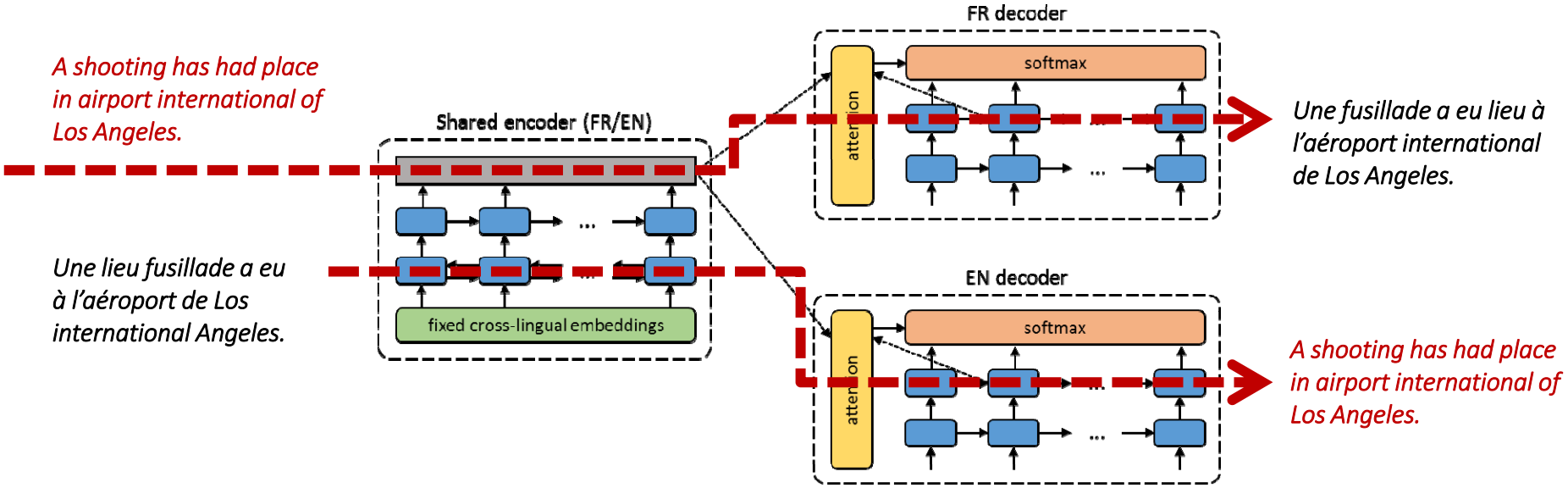
- ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised NMT (ICLR18)

Training

- ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised NMT (ICLR18)

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

Initialization		FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
(Only init.)	*Nearest neighbor	9.9	6.3	7.2	4.4		
	*Denoising	7.3	5.3	3.6	2.4		
Denoising	Artetxe et al. 2018	**15.6**	**15.1**	**10.2**	**6.5**		
	Lample et al. 2018	14.3	**15.1**				

Unsupervised NMT (ICLR18)

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

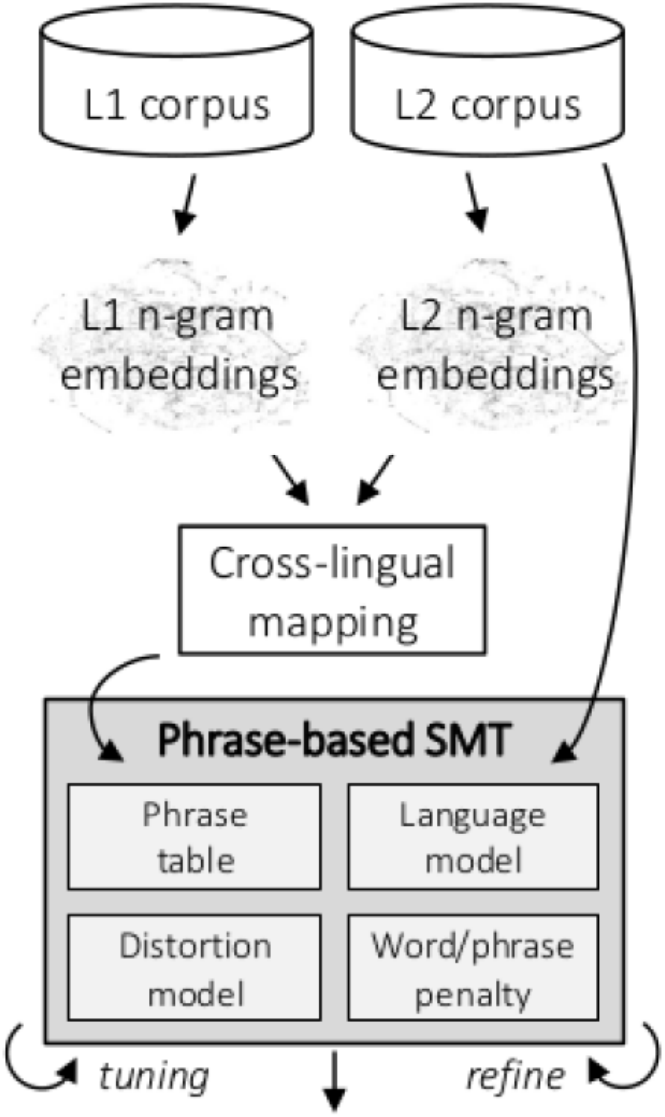
Initialization		FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
(Only init.)	*Nearest neighbor	9.9	6.3	7.2	4.4		
	*Denoising	7.3	5.3	3.6	2.4		
Denoising	Artetxe et al. 2018	**15.6**	**15.1**	**10.2**	**6.5**		
	Lample et al. 2018	14.3	**15.1**				

Self-learning (backtranslation) works!

What if we use other initializations

Unsupervised statistical machine translation (EMNLP 18)

Unsupervised statistical machine translation (EMNLP 18)



Unsupervised machine translation (ACL19)

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

Initialization		FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
(Only init.)	*Nearest neighbor	9.9	6.3	7.2	4.4		
	*Denoising	7.3	5.33	3.64	2.4		
Denoising	Artetxe et al. 2018	15.6	15.1	10.2	6.5		
	Lample et al. 2018	14.3	15.1				
PBMT	Lample et al. 2018b	27.7	27.6			25.2	20.2
	Artetxe et al. 2019	**33.5**	**36.2**	**27.0**	**22.5**	**34.4**	**26.9**

Self-learning (backtranslation) works

Even better with PBMT for initialization

Unsupervised machine translation (ACL19)

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

Initialization		FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
(Only init.)	*Nearest neighbor	9.9	6.3	7.2	4.4		
	*Denoising	7.3	5.33	3.64	2.4		
Denoising	Artetxe et al. 2018	15.6	15.1	10.2	6.5		
	Lample et al. 2018	14.3	15.1				
PBMT	Lample et al. 2018b	27.7	27.6			25.2	20.2
	Artetxe et al. 2019	**33.5**	**36.2**	**27.0**	**22.5**	**34.4**	**26.9**

Any other idea around?

Unsupervised machine translation (ACL19)

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

Initialization		FR-EN	EN-FR	DE-EN	EN-DE	DE	
(Only init.)	*Nearest neighbor	9.9	6.3				
	*Denoising	7.3	5.33				
Denoising	Artetxe et al. 2018	15	1				
	Lample et al. 2018	14					
PBMT	Lample et al. 2018b	27.7	17.6			25.2	
	Artetxe et al. 2019	**33.5**	**36.2**	**27.0**	**22.5**	**34.4**	**26.9**

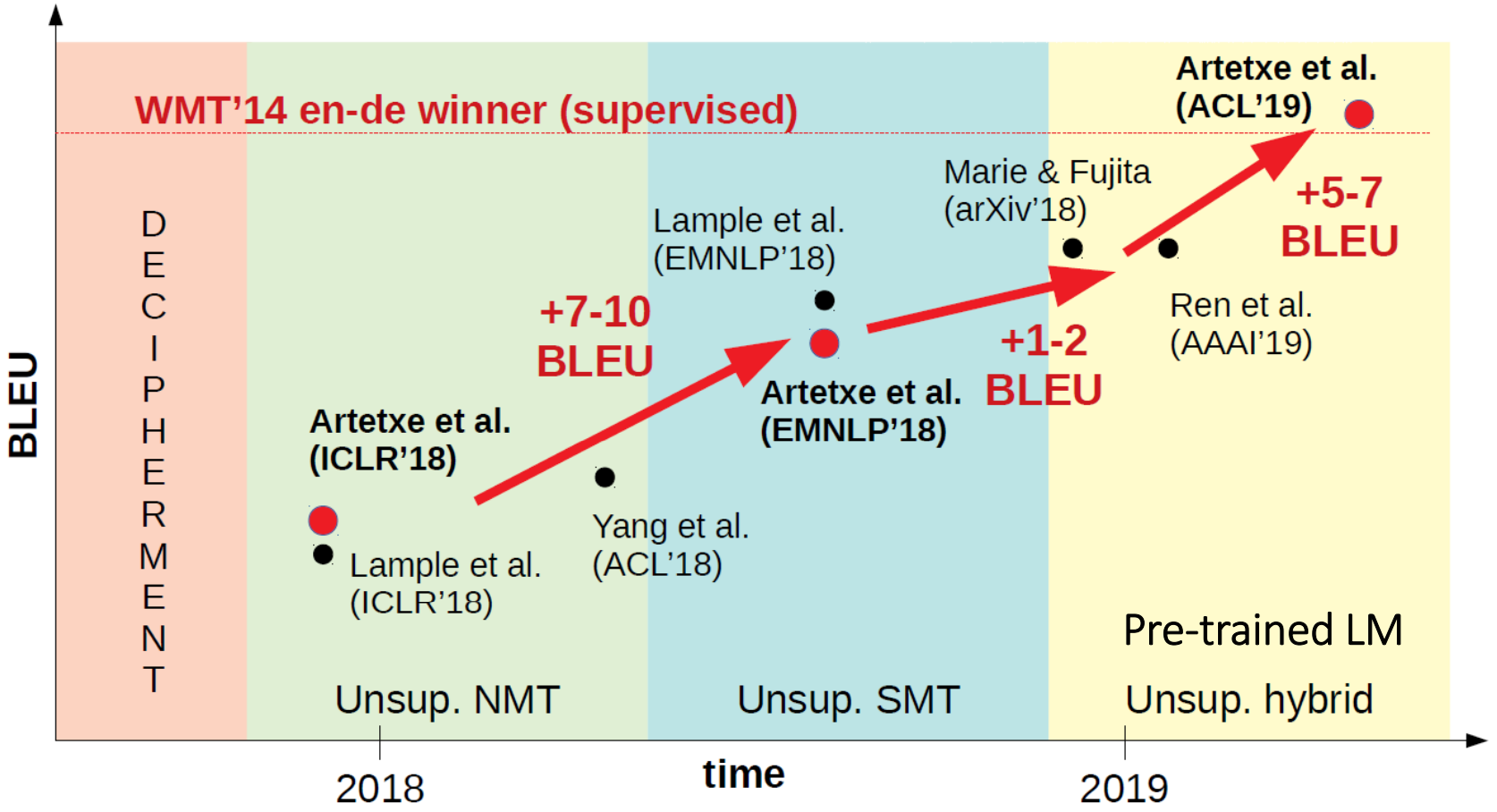
Any other idea around?

Unsupervised machine translation

Only WMT released data (test and monolingual corpora). WMT14 and WMT16

Initialization		FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
(Only init.)	*Nearest neighbor	9.9	6.3	7.2	4.4		
	*Denoising	7.3	5.33	3.64	2.4		
Denoising	Artetxe et al. 2018	15.6	15.1	10.2	6.5		
	Lample et al. 2018	14.3	15.1				
PBMT	Lample et al. 2018b	27.7	27.6			25.2	20.2
	Artetxe et al. 2019	**33.5**	**36.2**	**27.0**	**22.5**	34.4	26.9
Pre-trained LM	Lample and C. 2019	33.3	33.4			34.3	26.4
	Song et al. 2019					**35.2**	28.3
	Liu et al. 2019					34.0	**29.8**

Unsupervised machine translation (ACL19)



UMT is at the level MT was at 2014

Why does it work?

Why does it work?

Early to say… but intuition:

Why does it work?

Early to say… but intuition:

- Mapped embedding space provides information for k-best possible translations
- NMT / PBMT / Back-translation figures out how to best “combine” them

Conclusions

- New research area – unsupervised Machine Translation
- Performance up, WMT'14 En-De in two years
- Self-learning key (back-propagation)
- Room for improvement
- Code for replicability

  https://github.com/artetxem/undreamt
  https://github.com/artetxem/monoses

References: unsupervised MT

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. **Unsupervised Neural Machine Translation**. In *ICLR-2018*.

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. **Unsupervised Statistical Machine Translation**. In *EMNLP-2018*.

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. **An Effective Approach to Unsupervised Machine Translation.** In *ACL-2019.*

Thank you!

@eagirre
http://ixa2.si.ehu.eus/eneko

https://github.com/artetxem/vecmap
https://github.com/artetxem/undreamt
https://github.com/artetxem/monoses