Cross-linguality and machine translation without bilingual data

Eneko Agirre
@eagirre

Joint work with: Mikel Artetxe, Gorka Labaka

IXA NLP group – University of the Basque Country (UPV/EHU)

http://ixa.eus

Motivation

Cross-lingual word representations:

- **Word embeddings key** for Natural Language Processing
- **Mapped embeddings** represent languages in a single space
  - Depend on seed **bilingual dictionaries**
- **Exciting results** in dictionary induction, transfer learning, crosslingual applications, interlingual semantic representations

Motivation

Cross-lingual word representations:

- **Word embeddings key** for Natural Language Processing
- **Mapped embeddings** represent languages in a single space
  - Depend on seed **bilingual dictionaries**
- **Exciting results** in dictionary induction, transfer learning, crosslingual applications, interlingual semantic representations

Our focus: **extend mappings to any pair of languages**

- Most language pairs have **very few bilingual resources**
- Key research area for **wide adoption** of NLP tools

Motivation

Cross-lingual word representations:

- **Word embeddings key** for Natural Language Processing
- **Mapped embeddings** represent languages in a single space
  - Depend on seed **bilingual dictionaries**
- **Exciting results** in dictionary induction, transfer learning, crosslingual applications, interlingual semantic representations

Our focus: **extend mappings to any pair of languages**

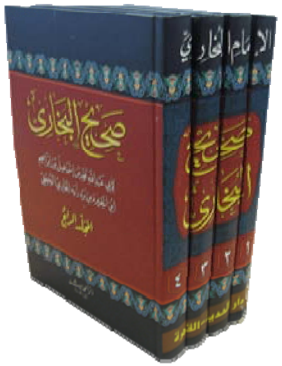
- Most language pairs have **very few bilingual resources**
- Key research area for **wide adoption** of NLP tools

In particular: **no bilingual resources at all**

- **Unsupervised** embedding mappings
- **Unsupervised** neural machine translation

Overview

<u>Arabic monolingual corpora</u>



Arabic embeddings

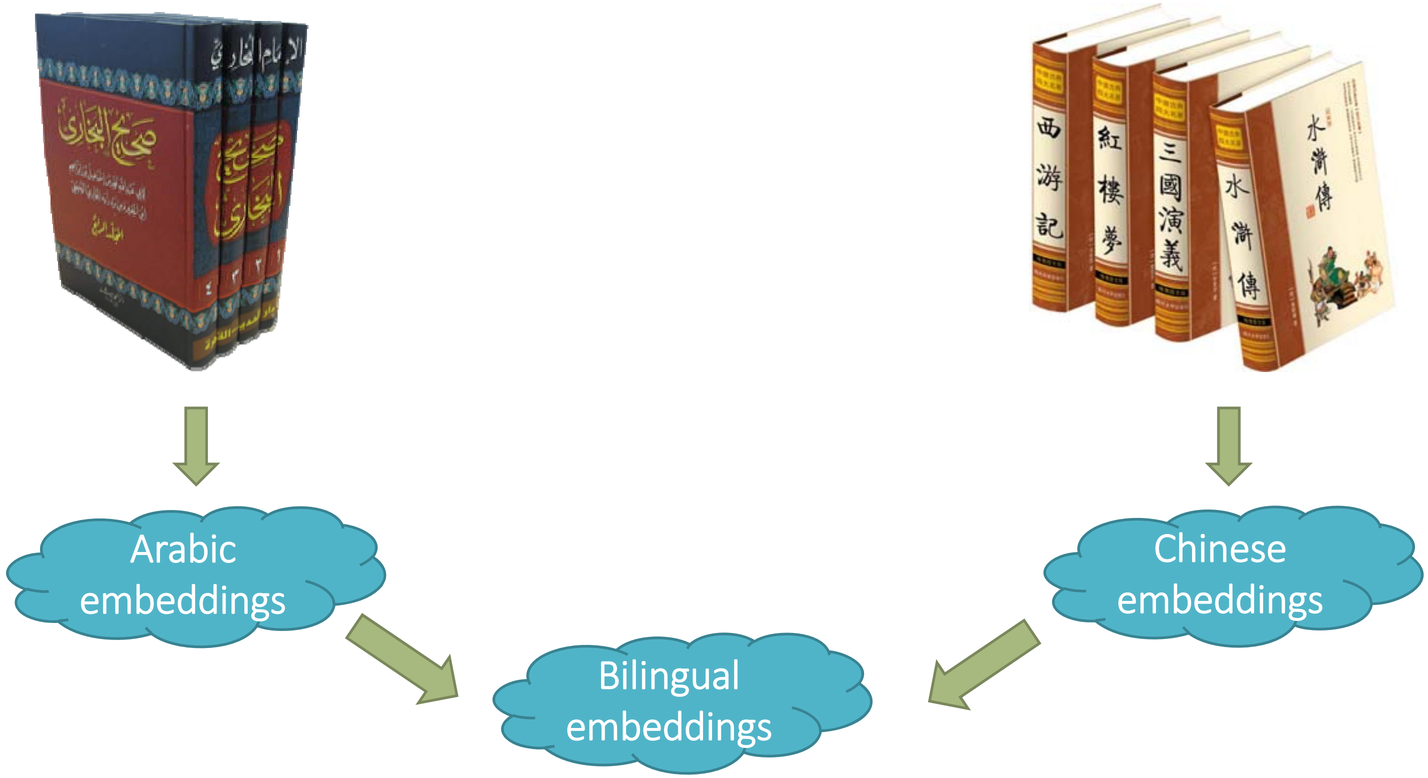
<u>Chinese monolingual corpora</u>

Chinese embeddings

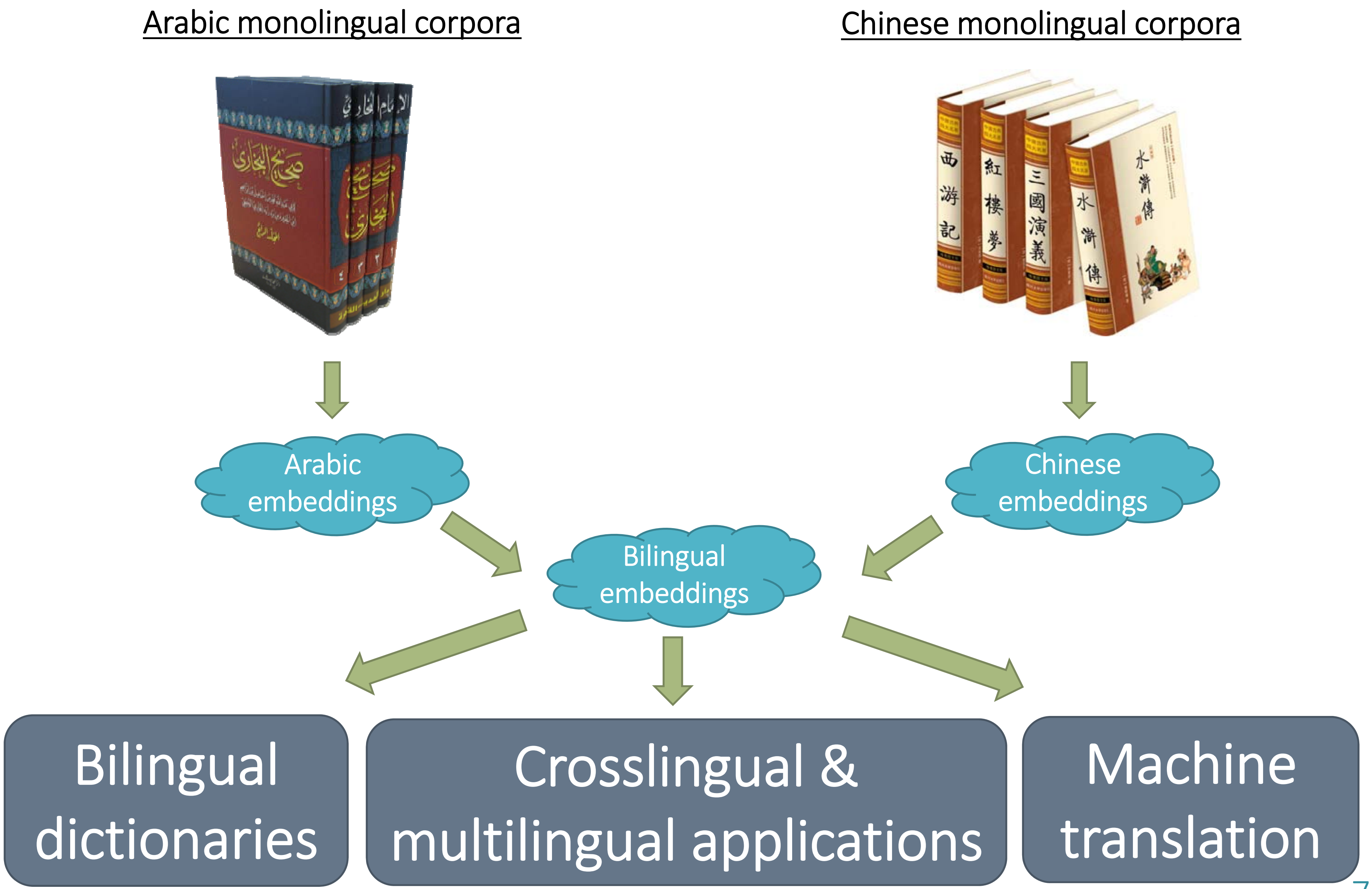
Overview

Arabic monolingual corpora

Chinese monolingual corpora

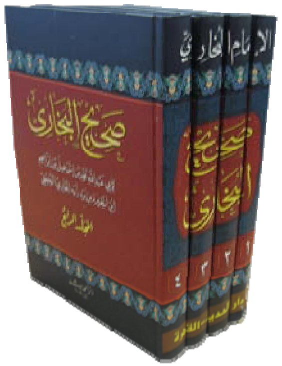


Overview


Arabic monolingual corpora
Chinese monolingual corpora
西游記
紅樓夢
三國演義
水滸傳
Arabic embeddings
Chinese embeddings
Bilingual embeddings
Bilingual dictionaries
Crosslingual & multilingual applications
Machine translation

Overview

Arabic monolingual corpora



Chinese monolingual corpora

No bilingual resource

Arabic embeddings

Chinese embeddings

Bilingual embeddings

Bilingual dictionaries

Crosslingual & multilingual applications

Machine translation

Outline

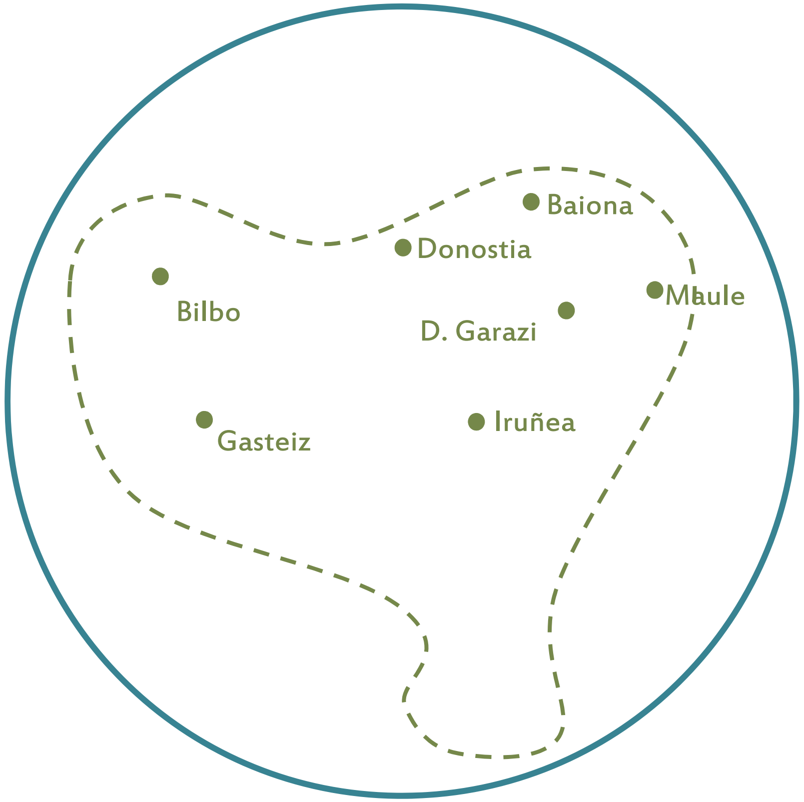
- Bilingual embedding mappings
  - *Introduction to vector space models (embeddings)*
  - *Bilingual embedding mappings (AAAI18)*
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions*
- Unsupervised neural machine translation
  - *Introduction to NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Unsupervised statistical MT (EMNLP18)*
  - *Conclusions*

Outline

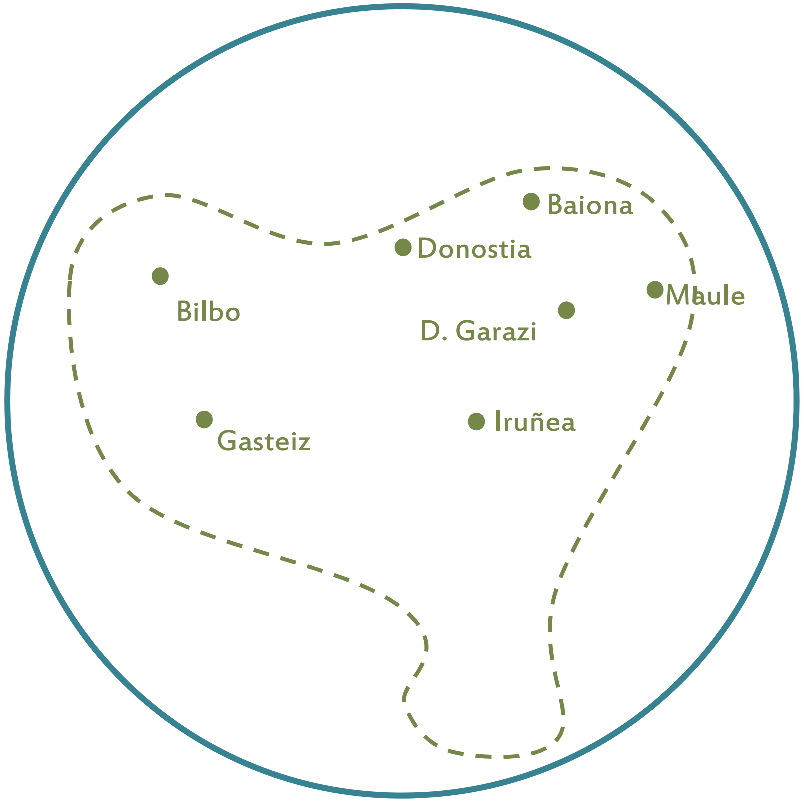
- Bilingual embedding mappings
  - *Introduction to vector space models (embeddings)*
  - *Bilingual embedding mappings* (AAAI18)
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions*
- Unsupervised neural machine translation
  - *Introduction to NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Unsupervised statistical MT (EMNLP18)*
  - *Conclusions*

Introduction to vector space models

Introduction to vector space models



Introduction to vector space models



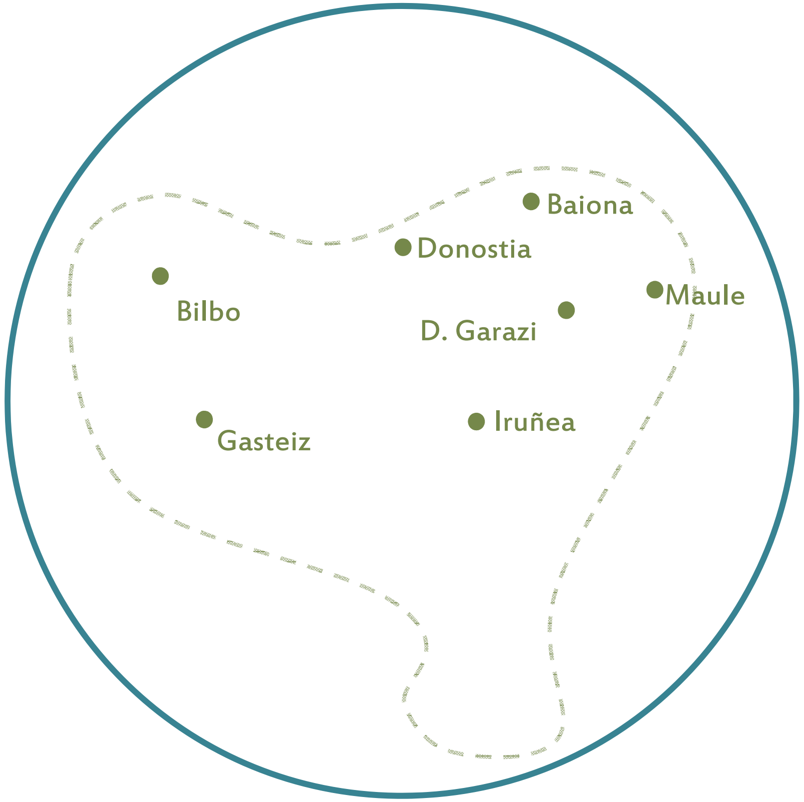
Geographical space

Introduction to vector space models

Geographical space

- Cities

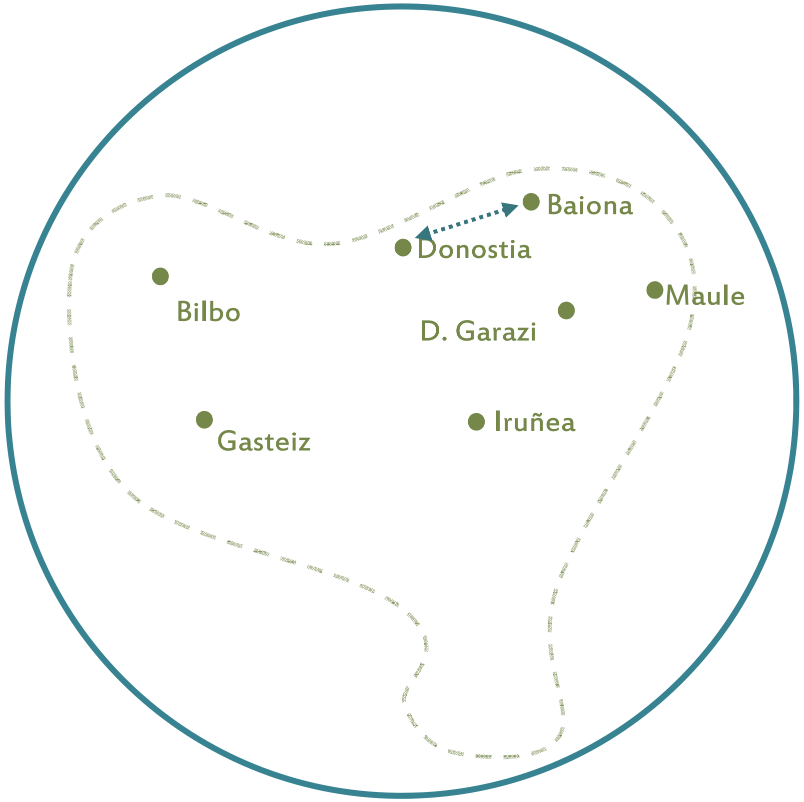
Introduction to vector space models



Geographical space

- Cities
- Meaningful distances

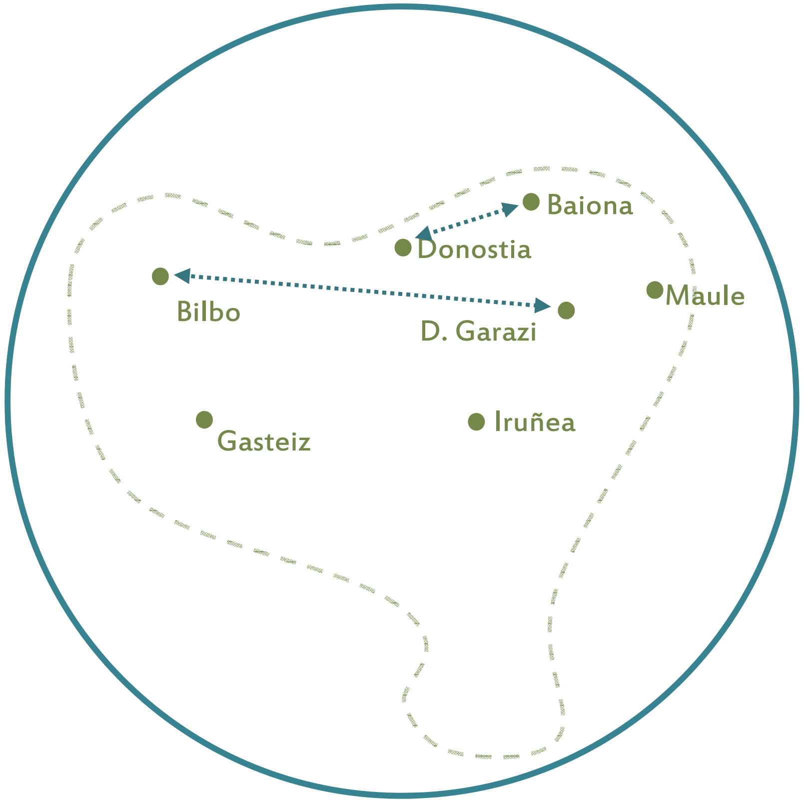
Introduction to vector space models



Geographical space

- Cities
- Meaningful distances

Introduction to vector space models



<u>Geographical space</u>

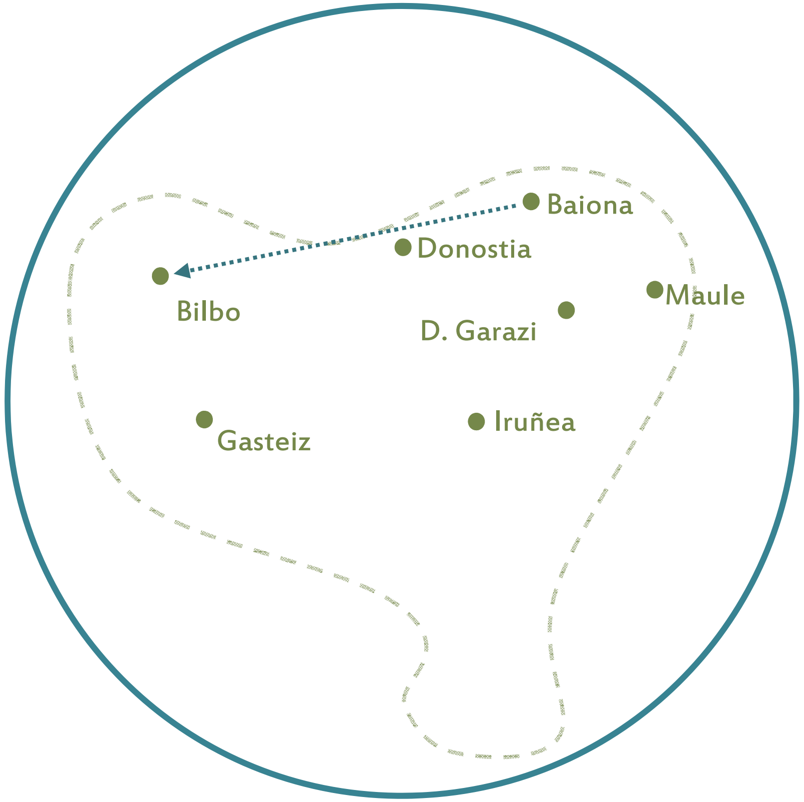
- Cities
- Meaningful distances

Introduction to vector space models

<u>Geographical space</u>

- Cities
- Meaningful distances
- Meaningful relations

Introduction to vector space models



Geographical space

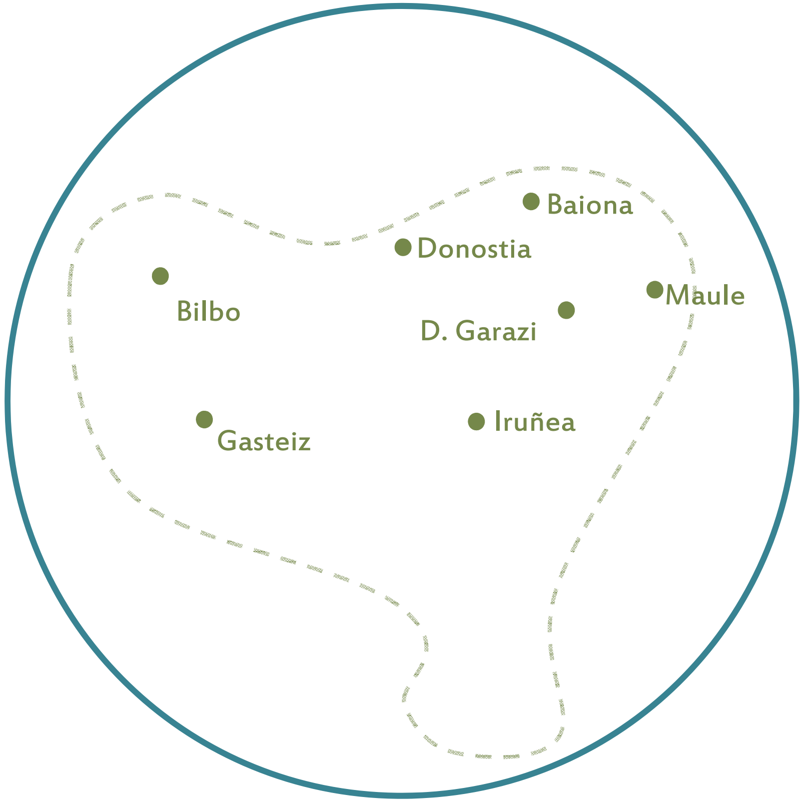
- Cities
- Meaningful distances
- Meaningful relations

Introduction to vector space models

<u>Geographical space</u>

- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions

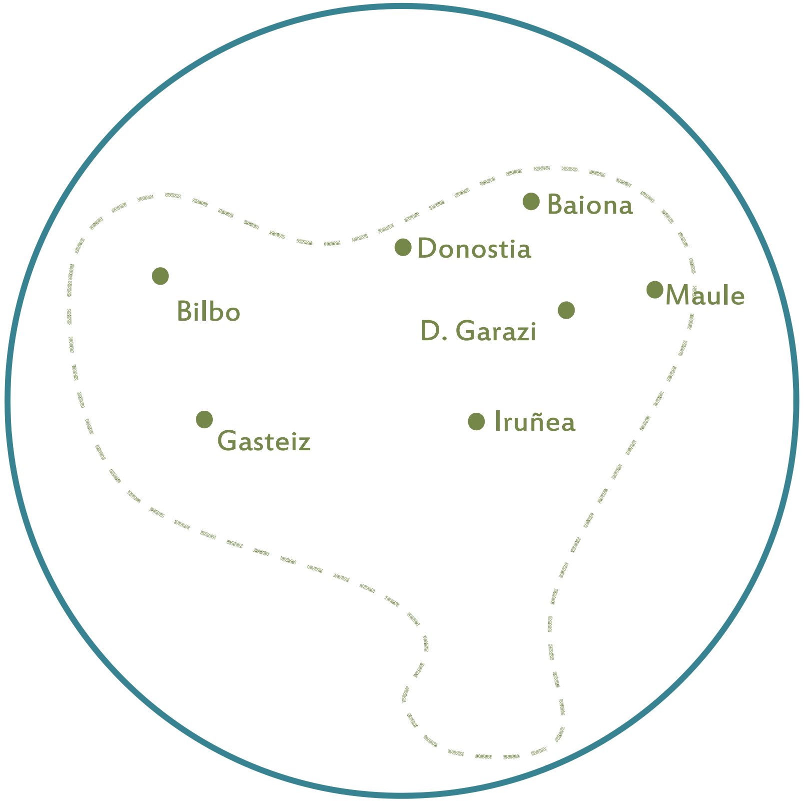
Introduction to vector space models



<u>Geographical space</u>

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

Introduction to vector space models

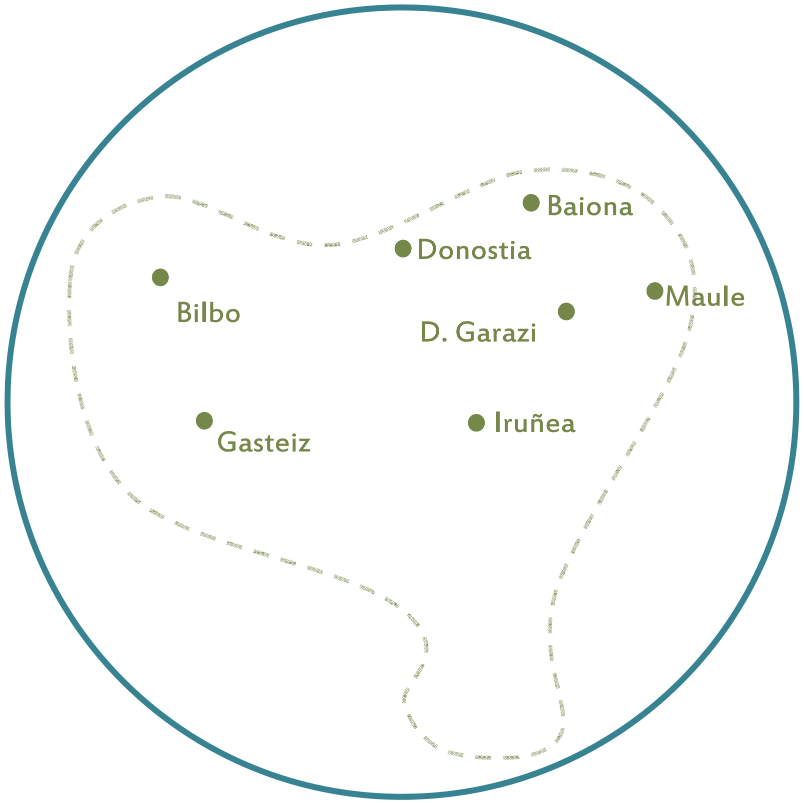


Geographical space

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

Semantic space

Introduction to vector space models



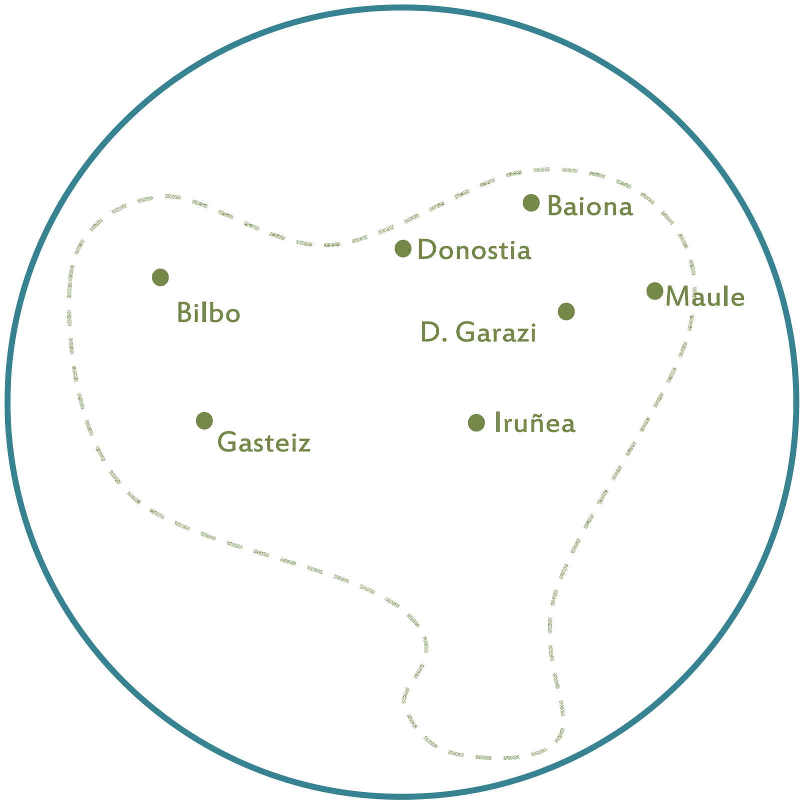
Geographical space

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

Semantic space

- Words

Introduction to vector space models

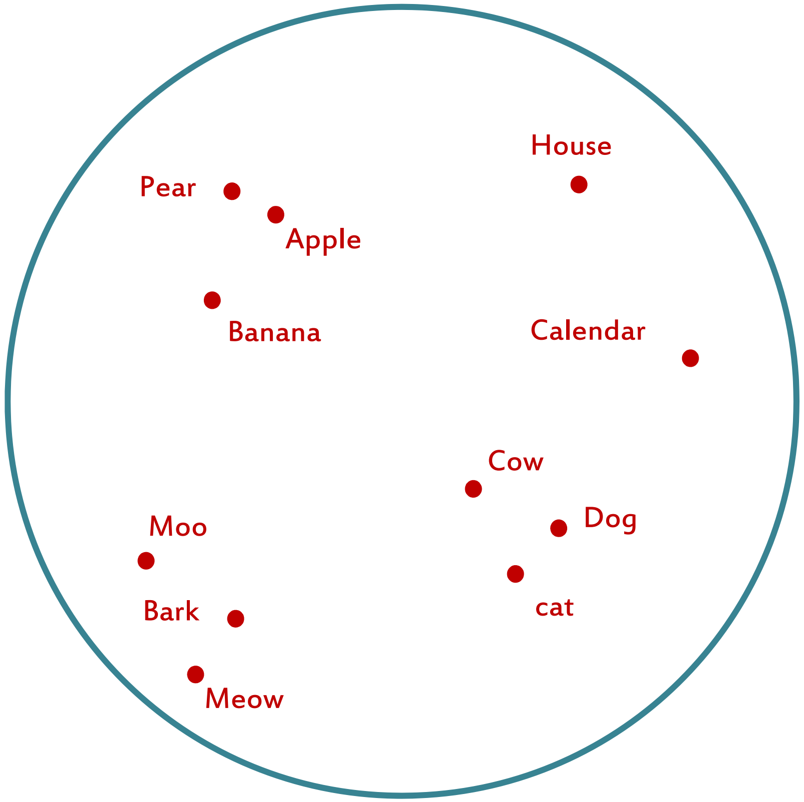


Geographical space

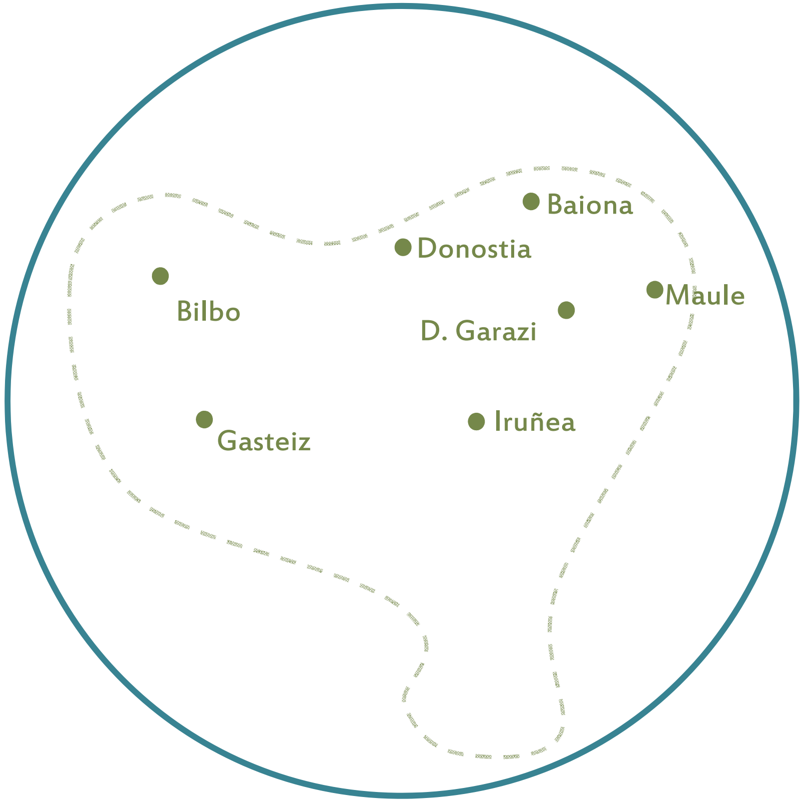
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words



Introduction to vector space models

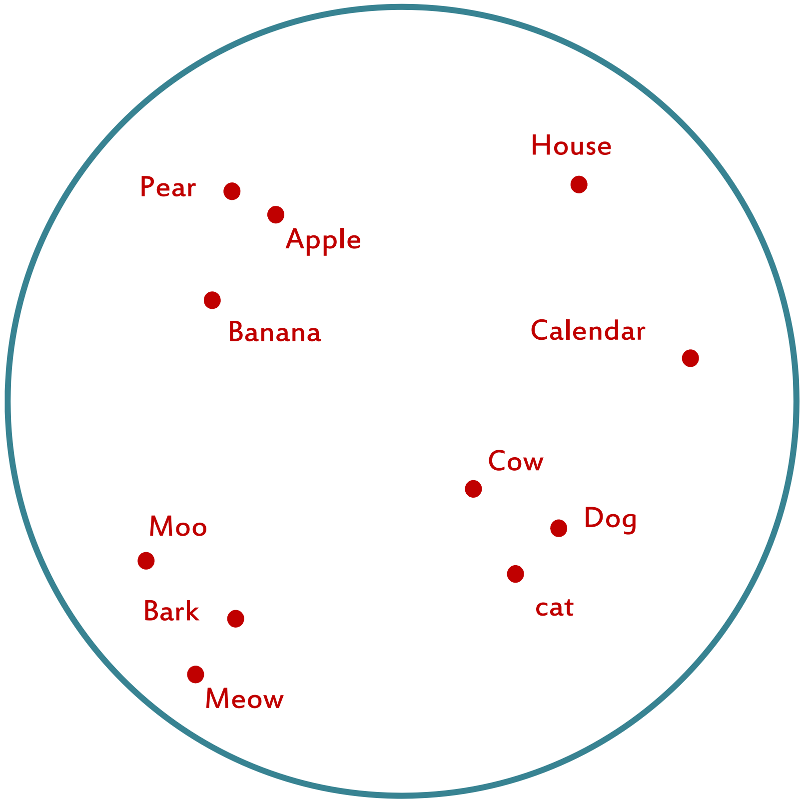


Geographical space

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

Semantic space

- Words
- Meaningful distances



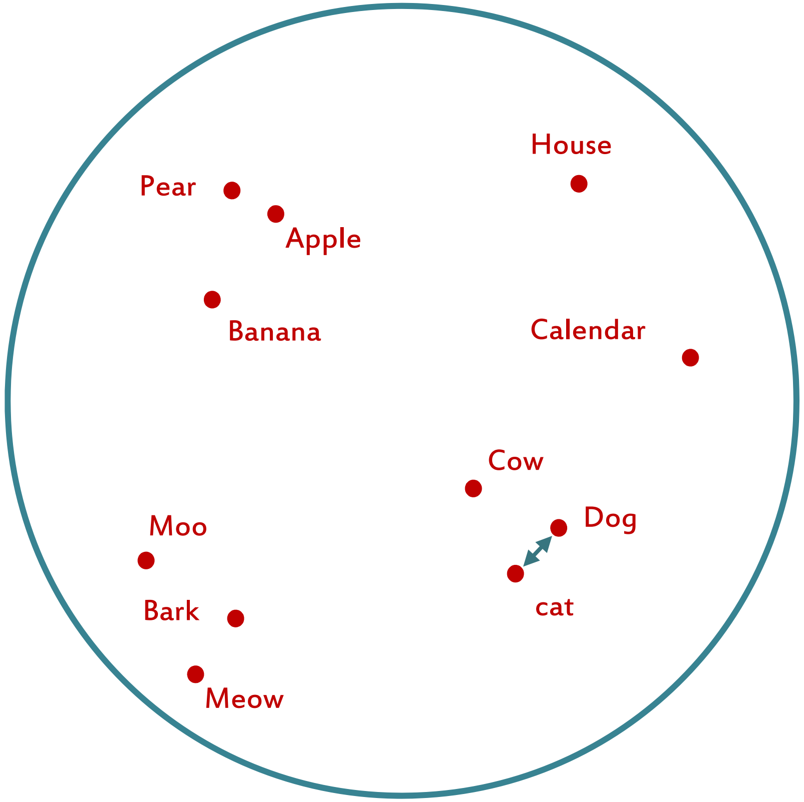
Introduction to vector space models

Geographical space

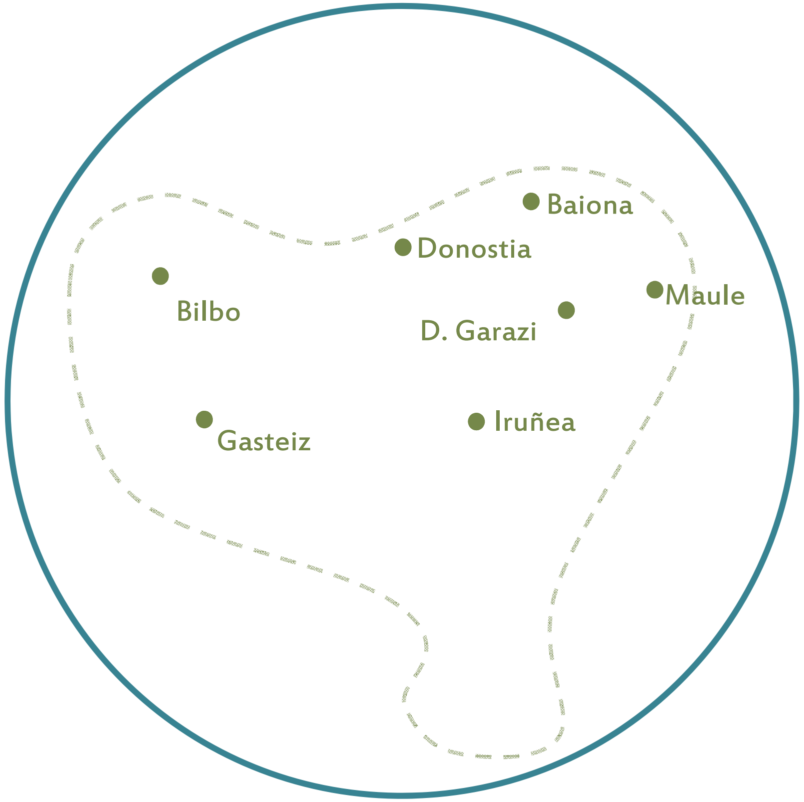
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances



Introduction to vector space models

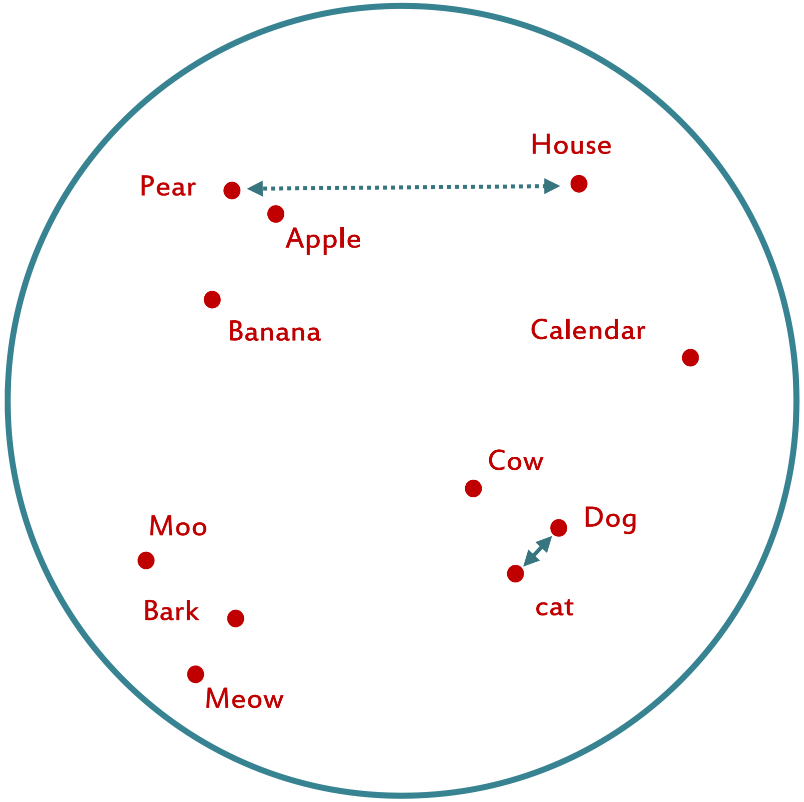


Geographical space

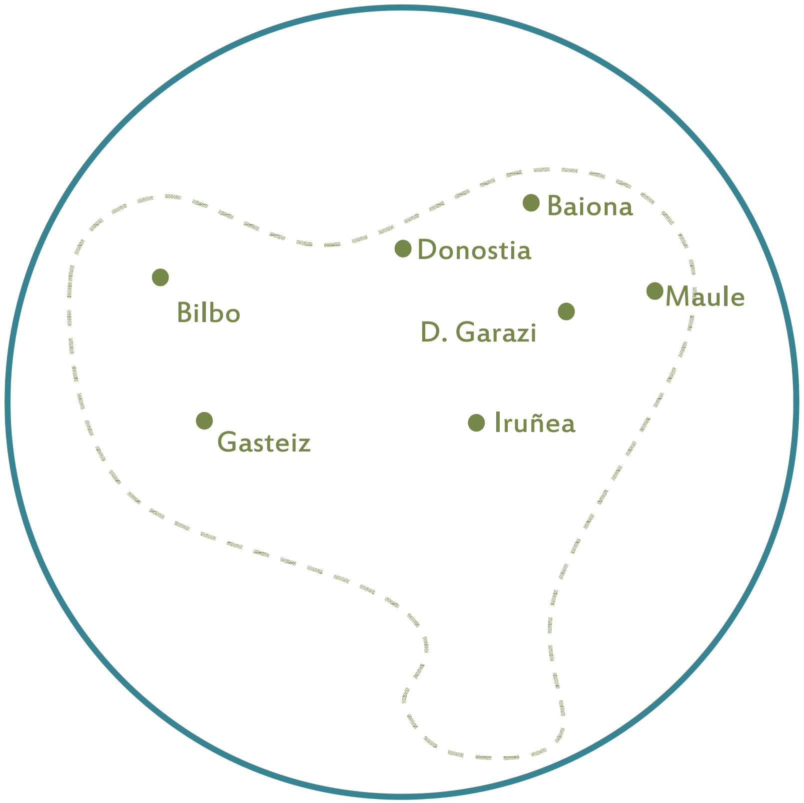
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances



Introduction to vector space models

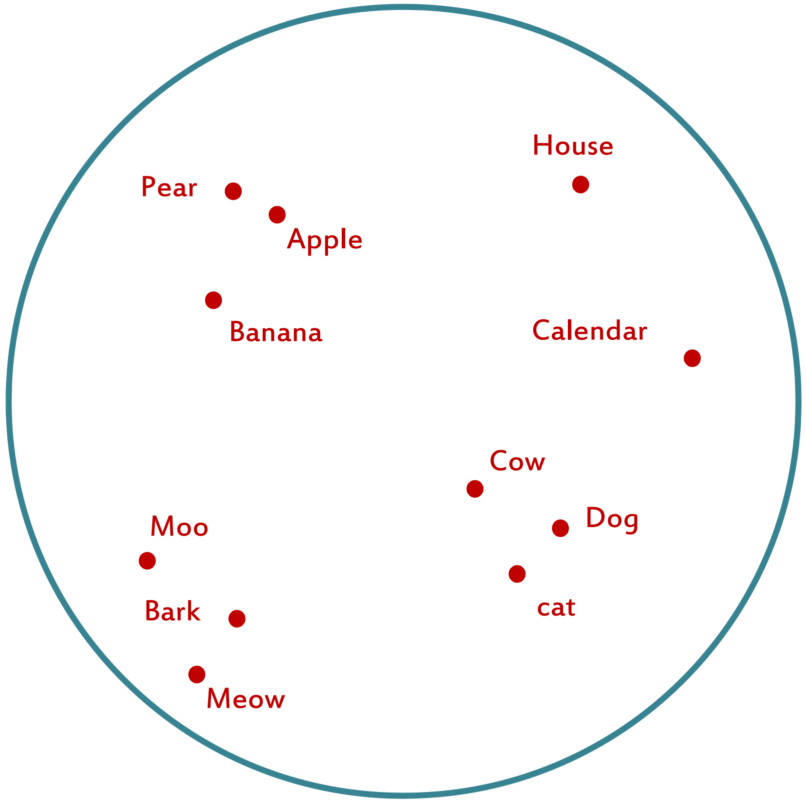


Geographical space

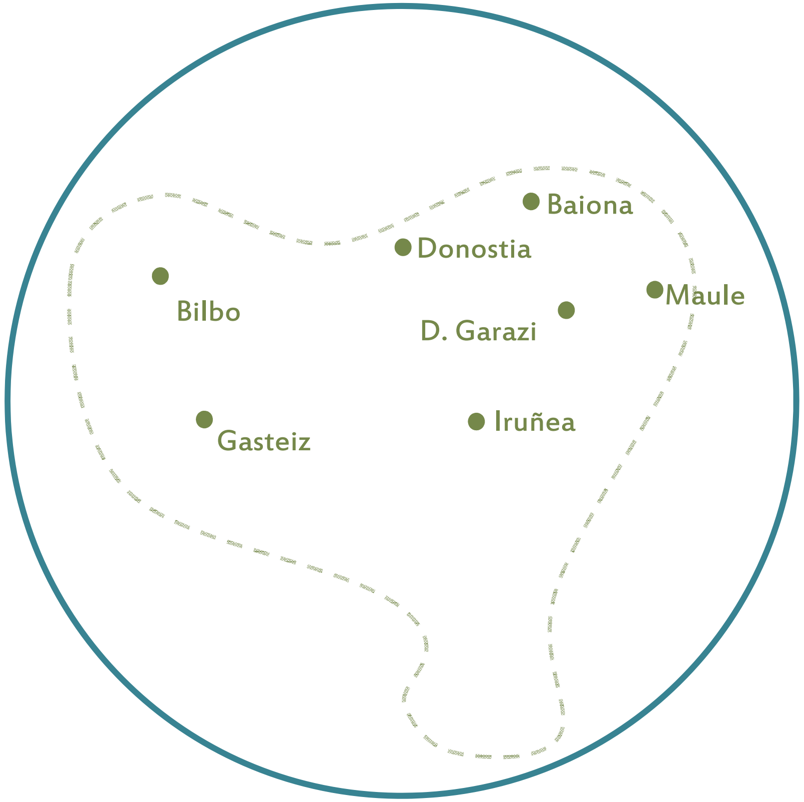
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Introduction to vector space models

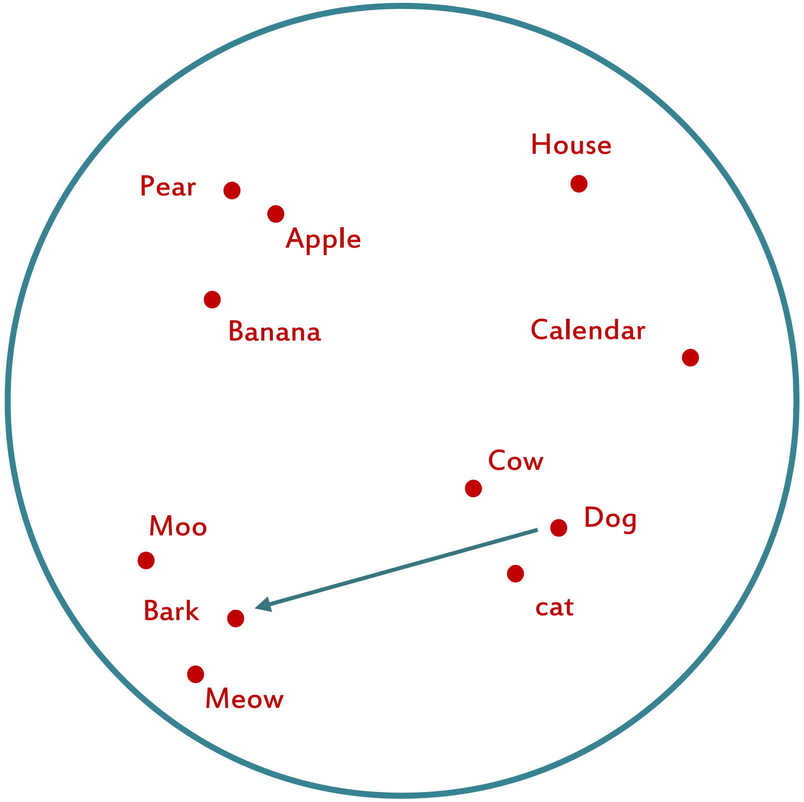


Geographical space

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

Semantic space

- Words
- Meaningful distances
- Meaningful relations



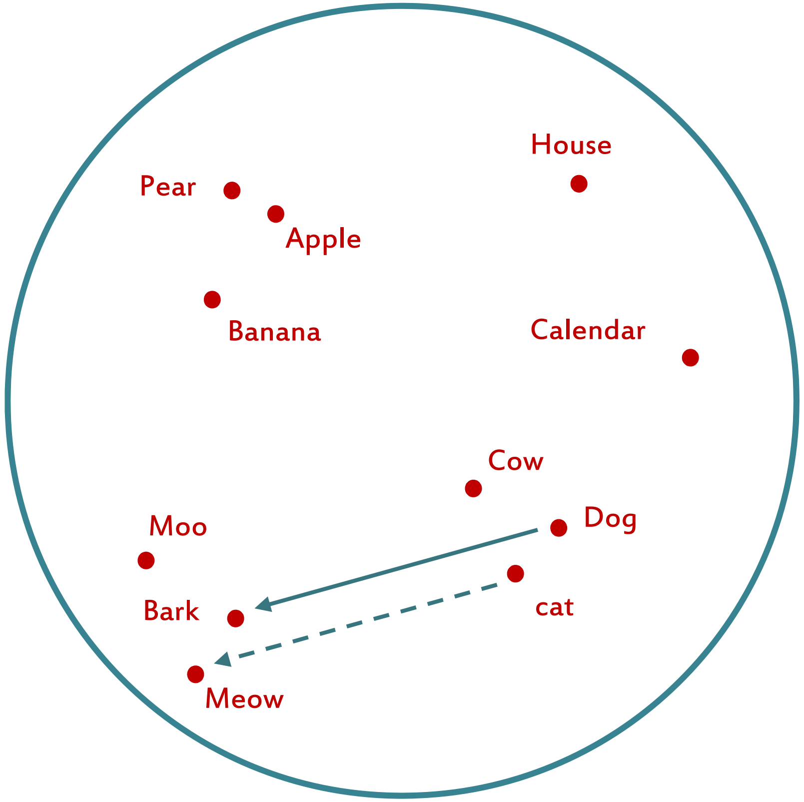
Introduction to vector space models

Geographical space

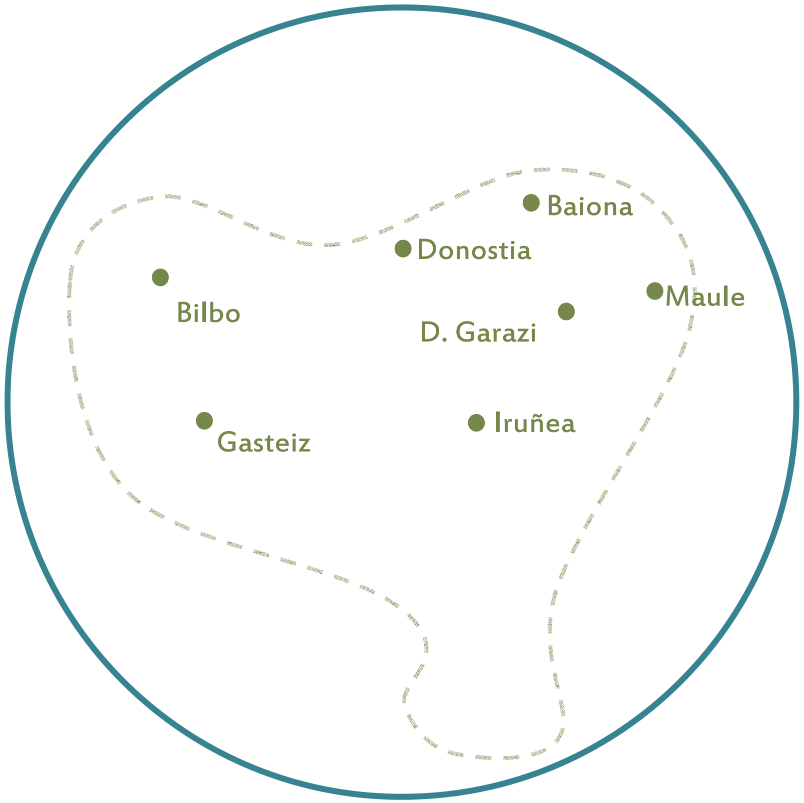
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Introduction to vector space models

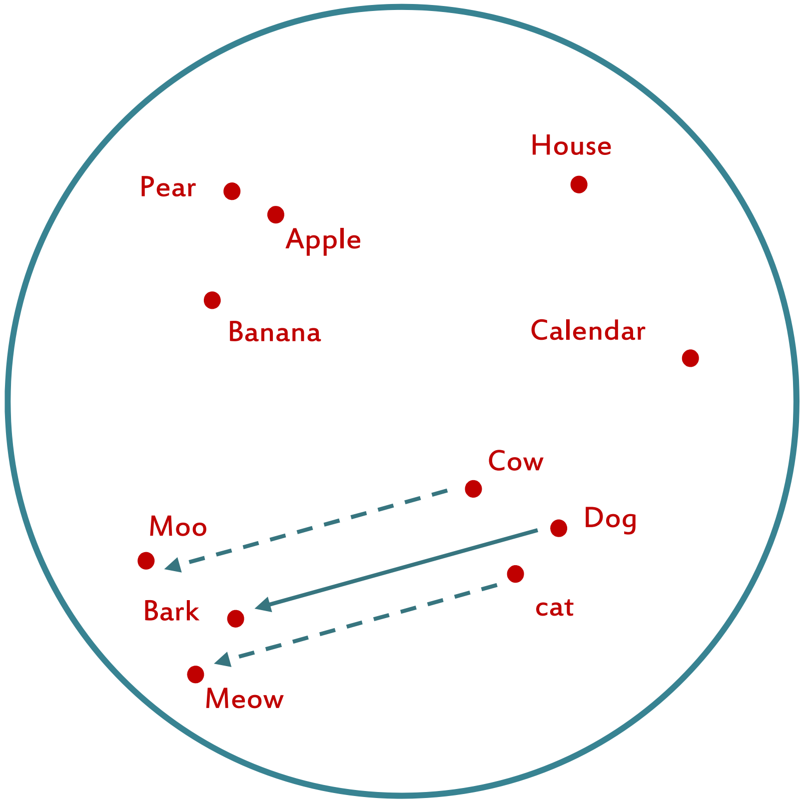


Geographical space

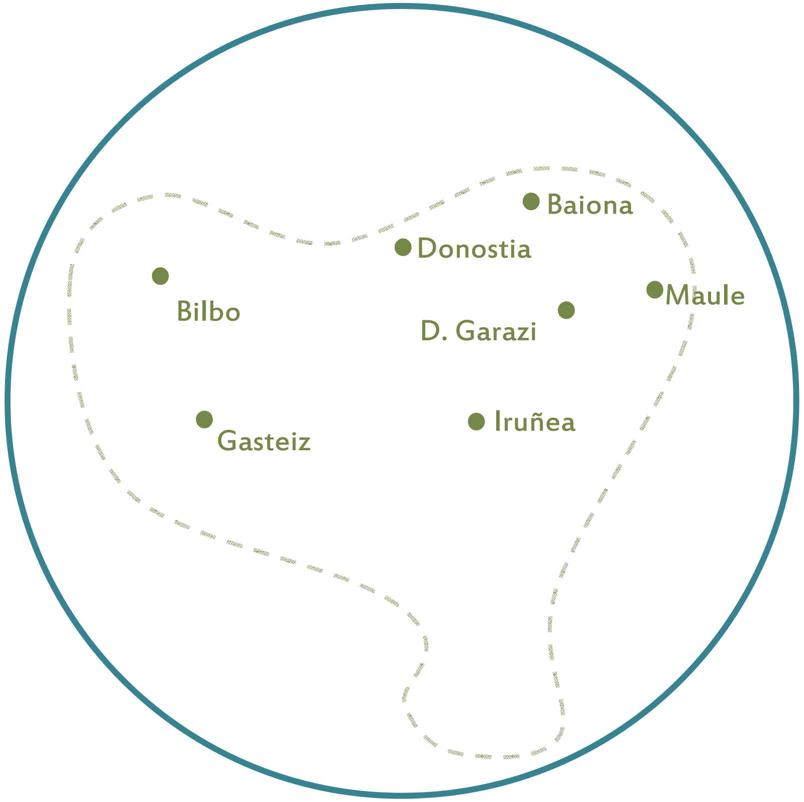
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Introduction to vector space models

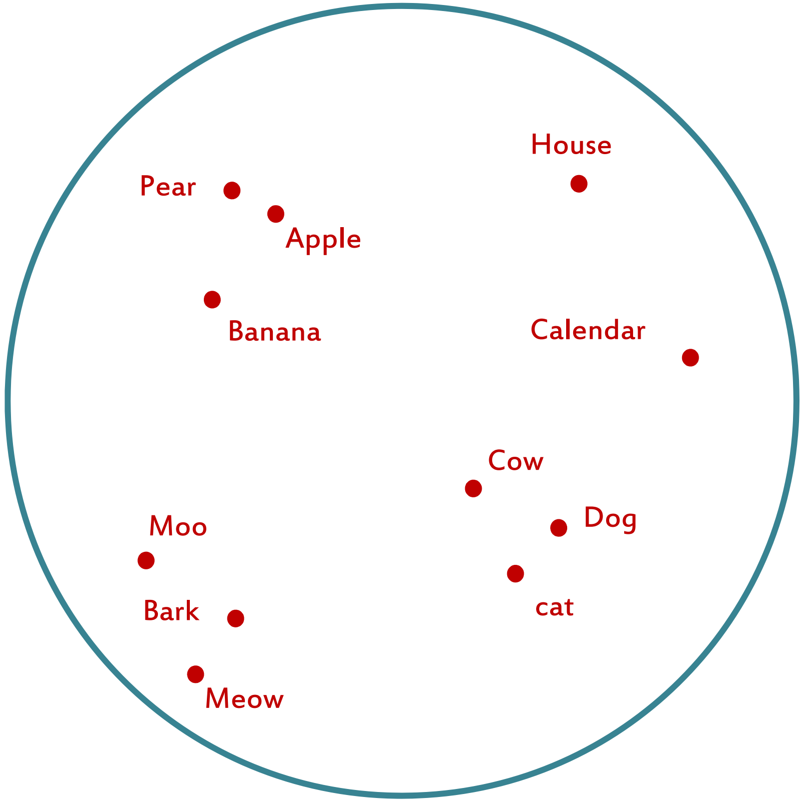


Geographical space

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

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions



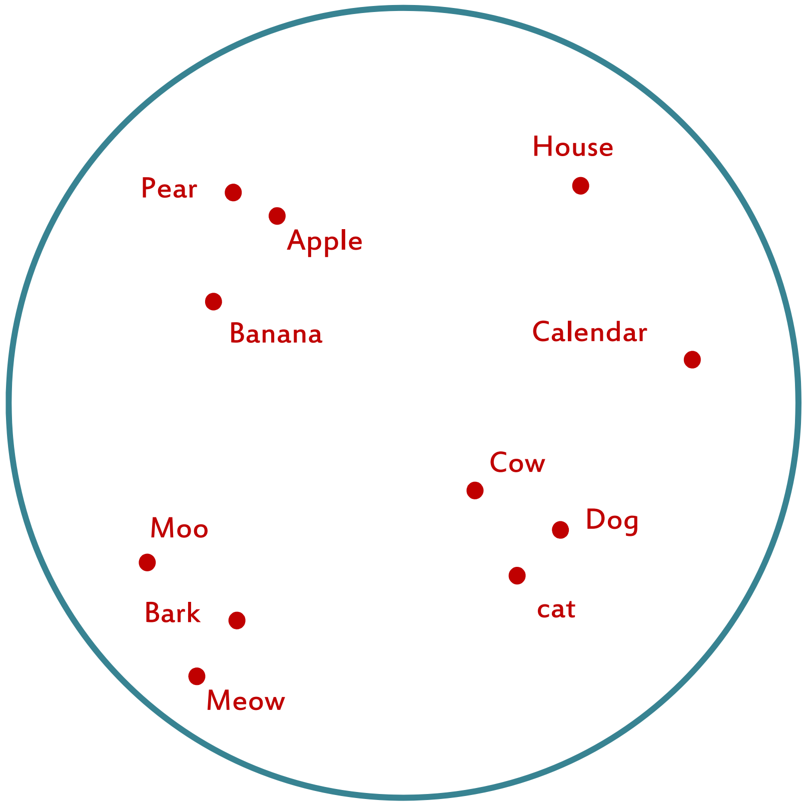
Introduction to vector space models

Geographical space

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

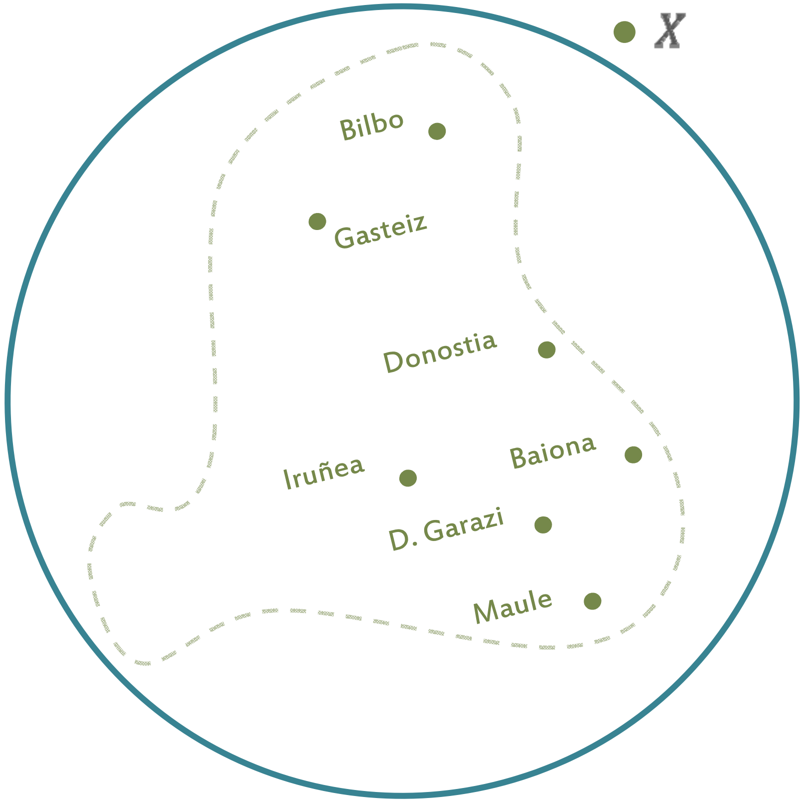
Semantic space

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

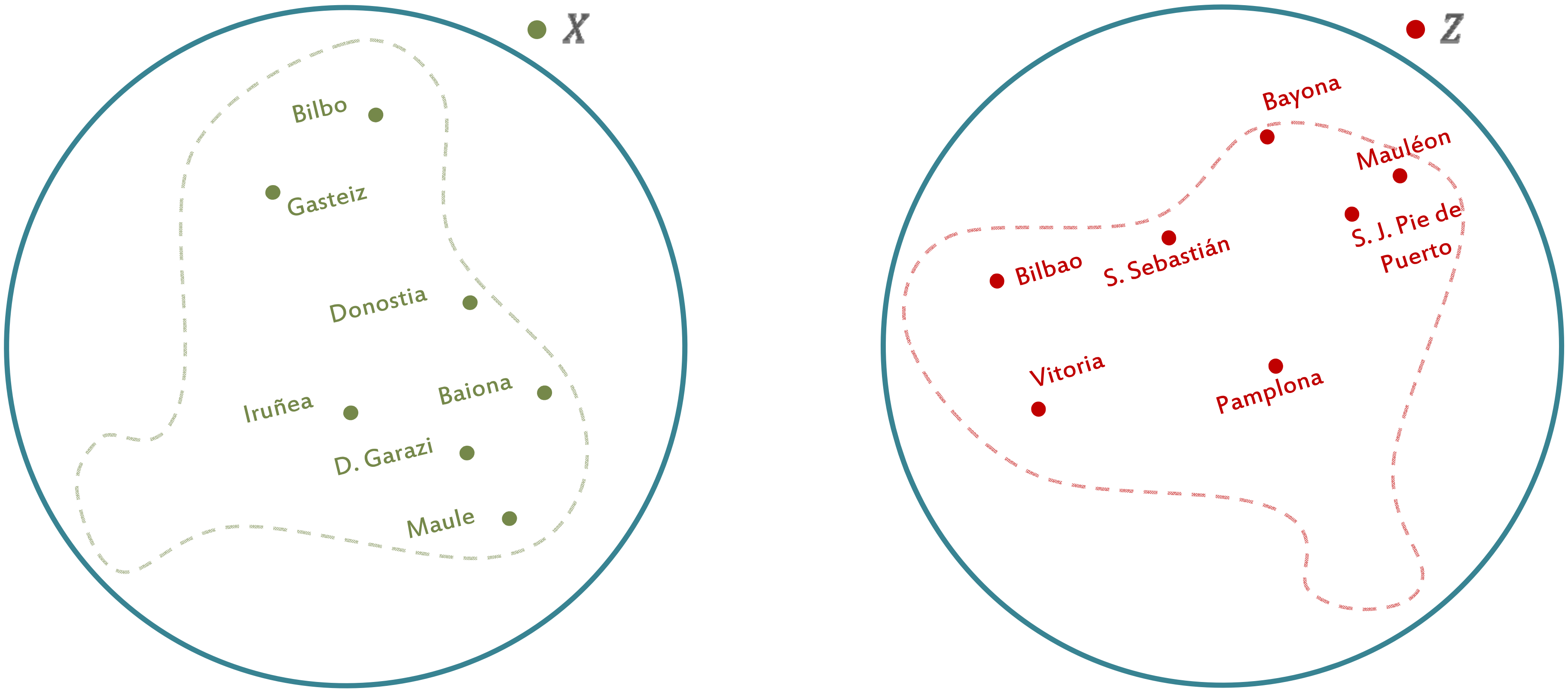


Introduction to embedding mappings

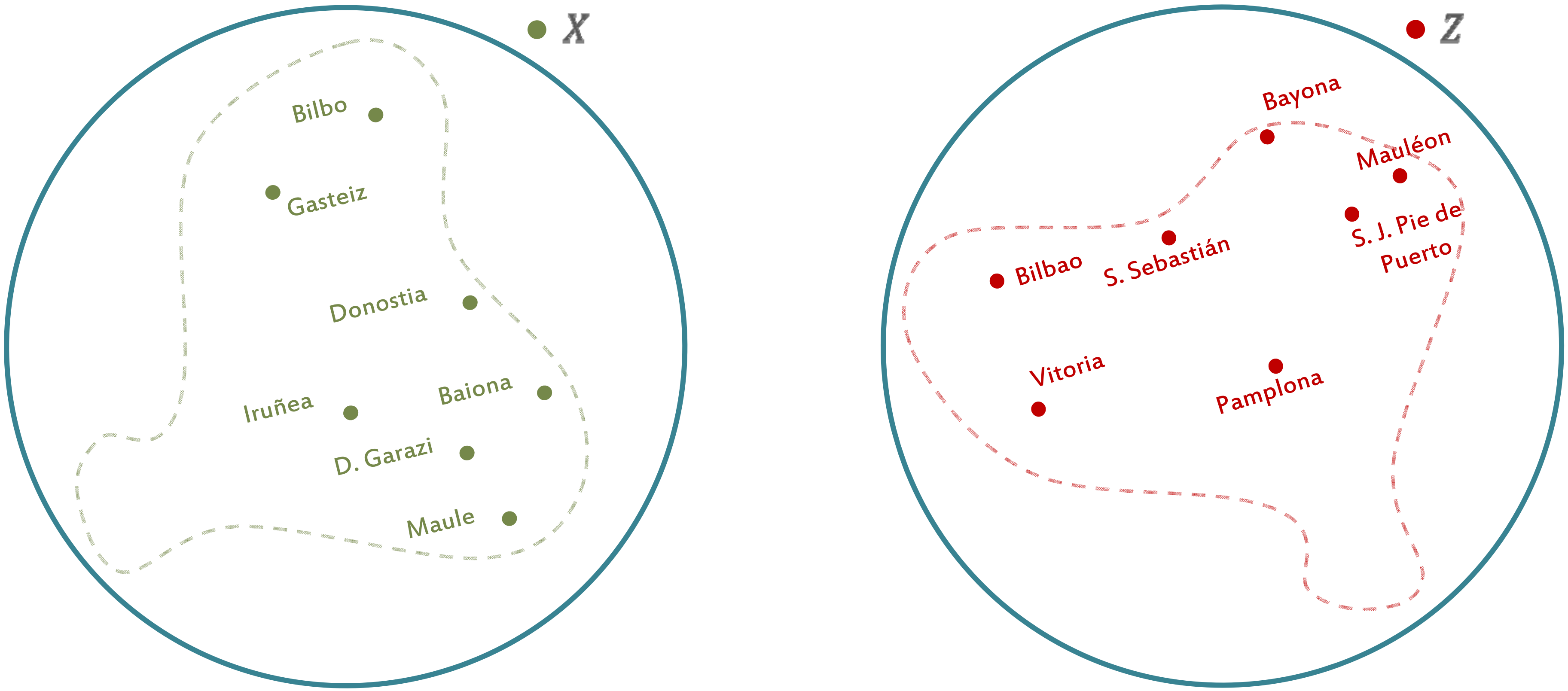
Introduction to embedding mappings



Introduction to embedding mappings

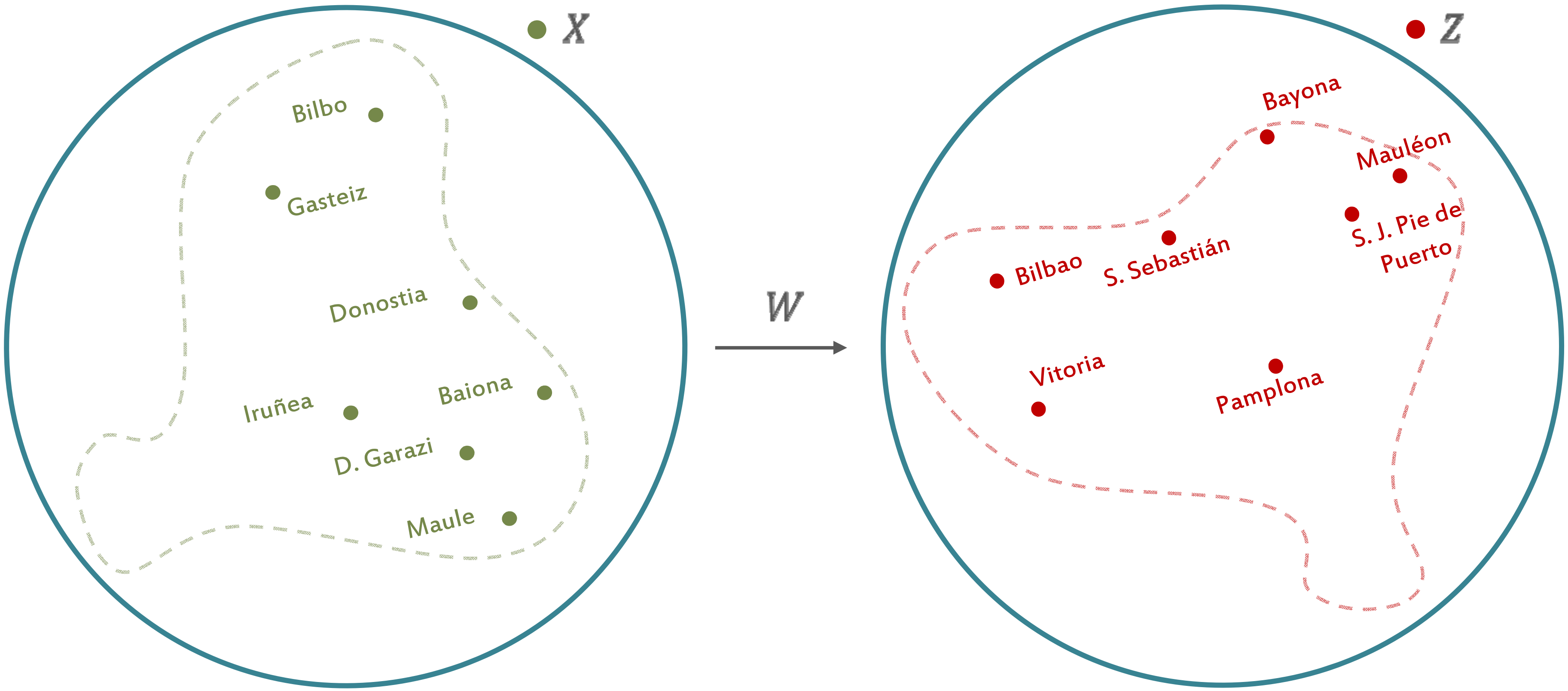


Introduction to embedding mappings



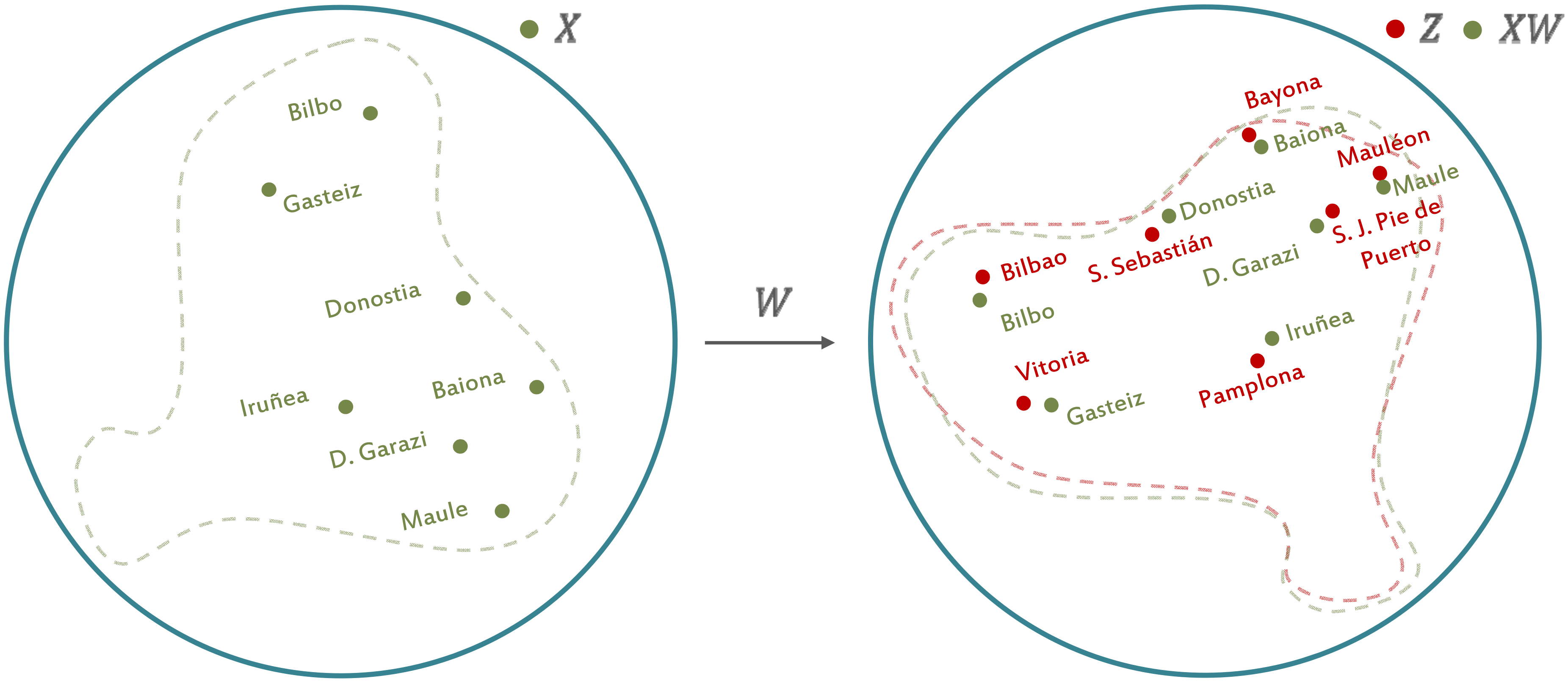
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Introduction to embedding mappings



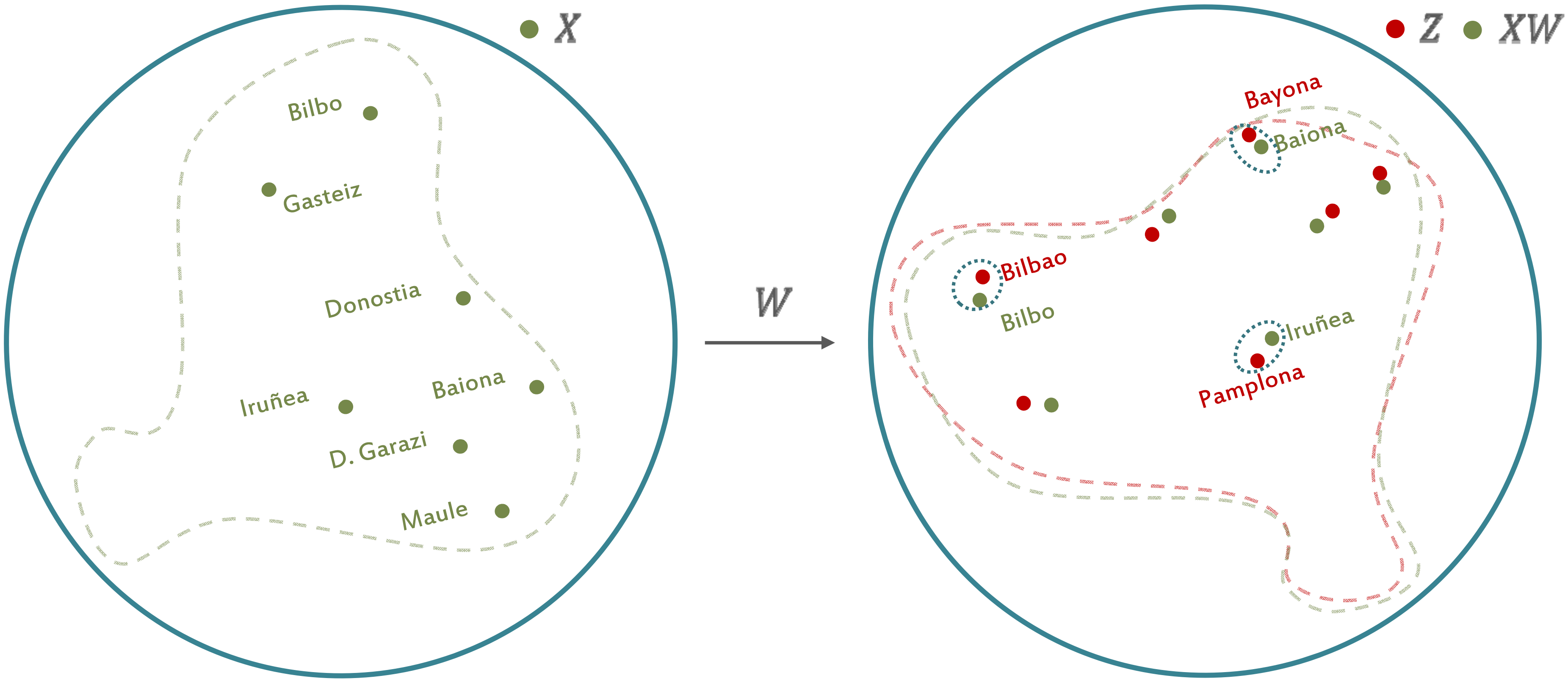
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Introduction to embedding mappings



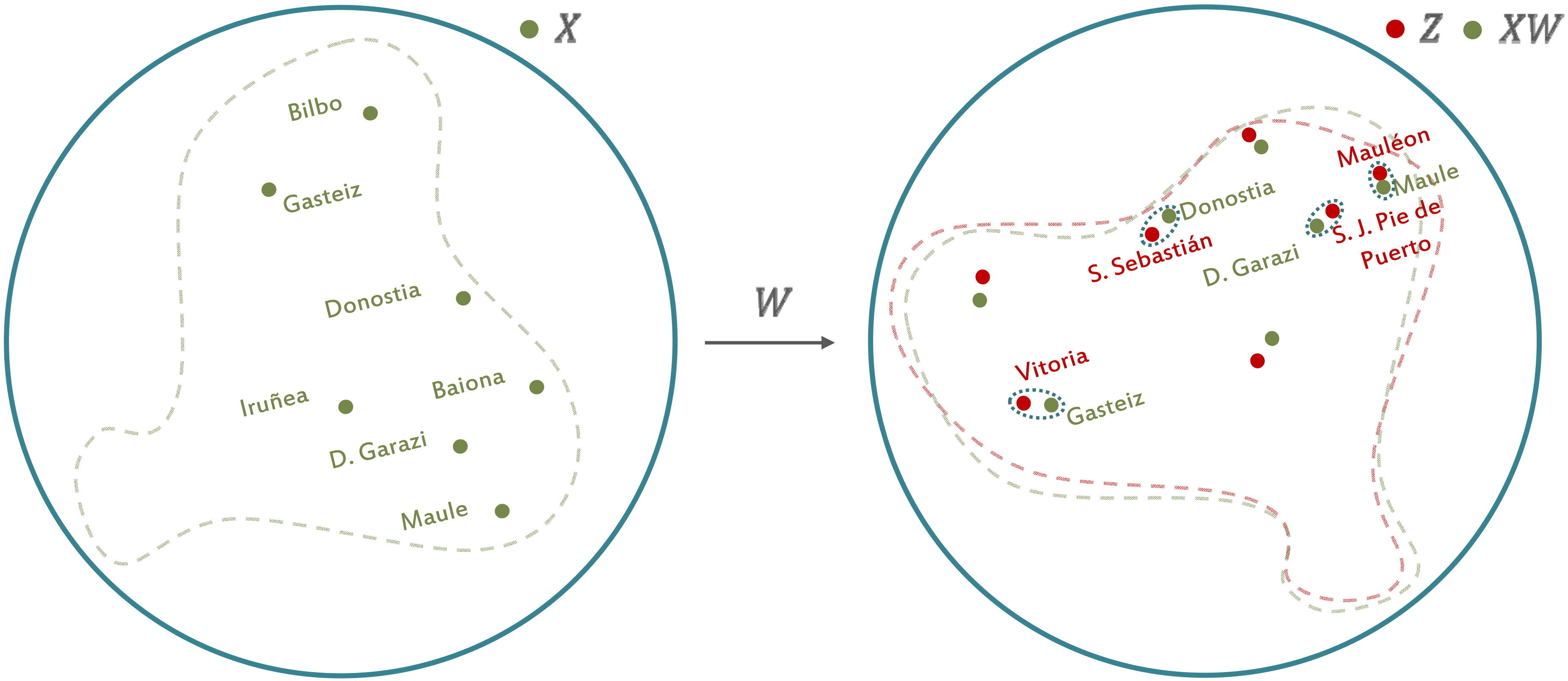
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Introduction to embedding mappings



Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

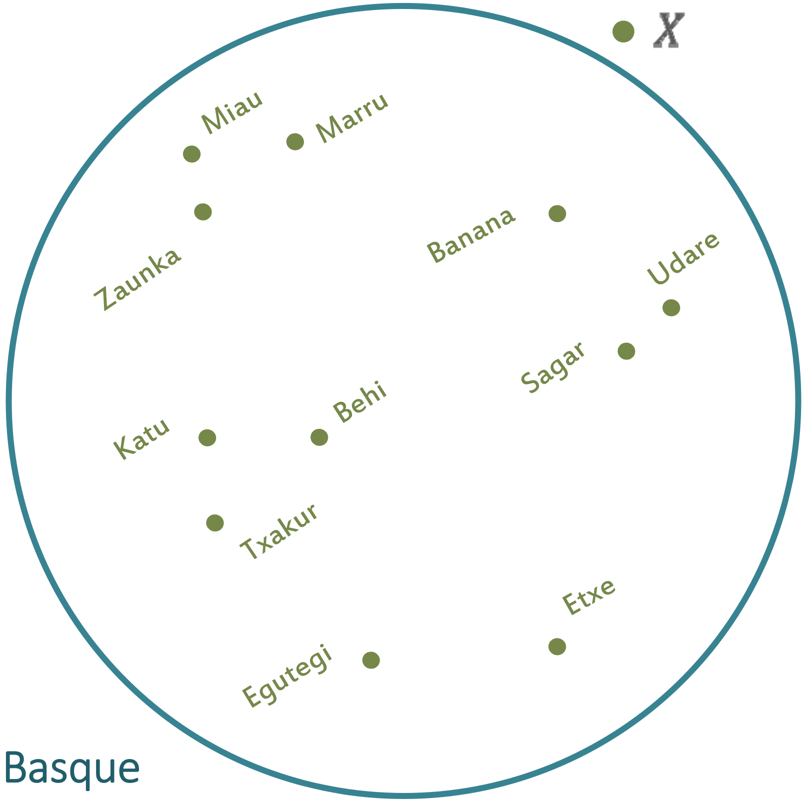
Introduction to embedding mappings



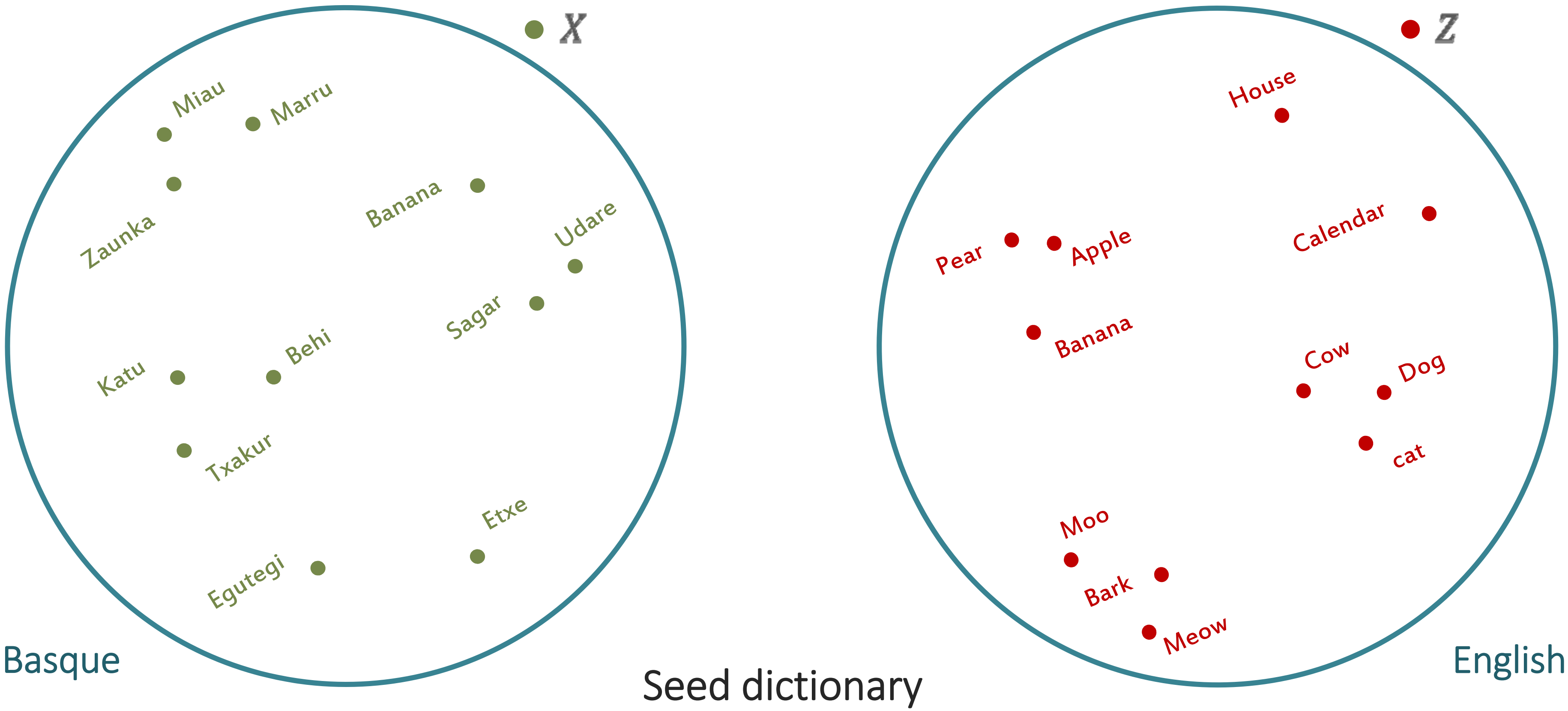
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Introduction to embedding mappings

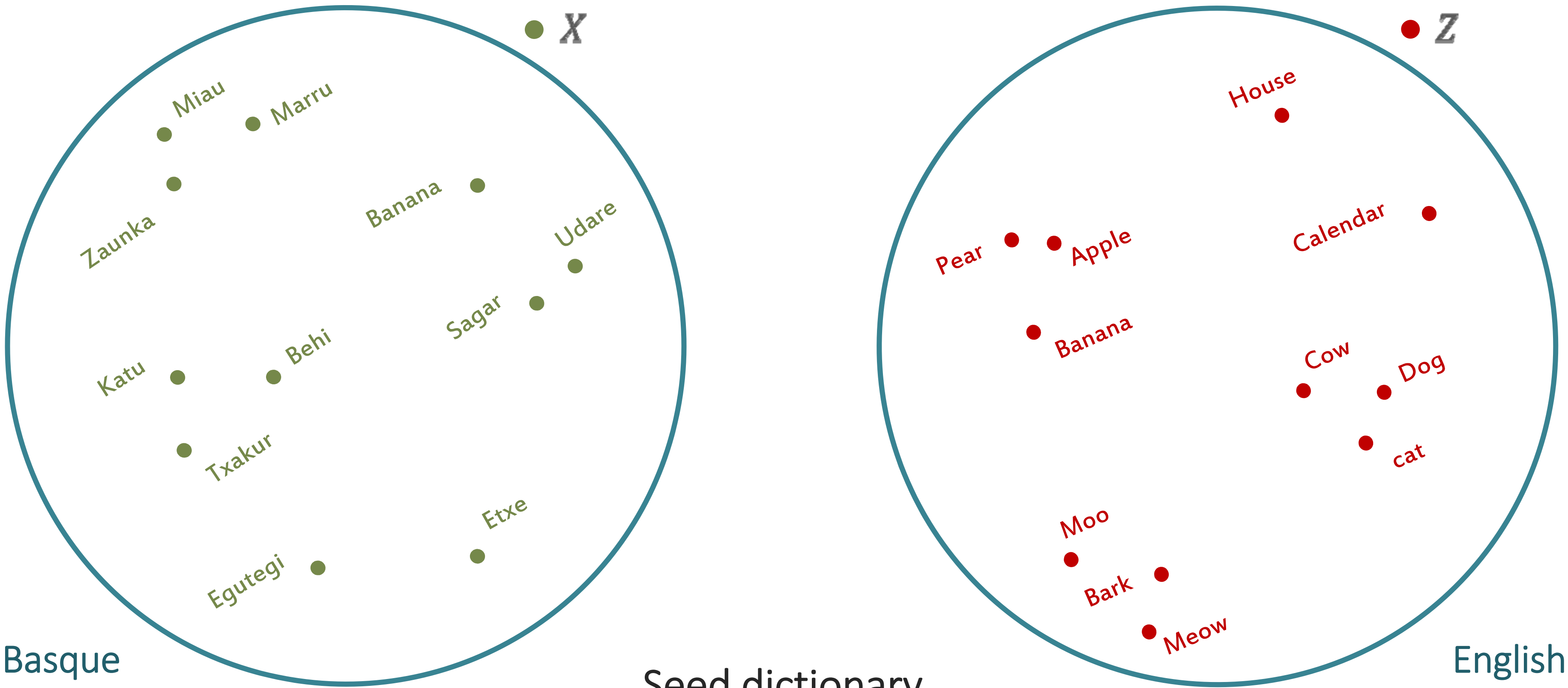
Introduction to embedding mappings



Introduction to embedding mappings



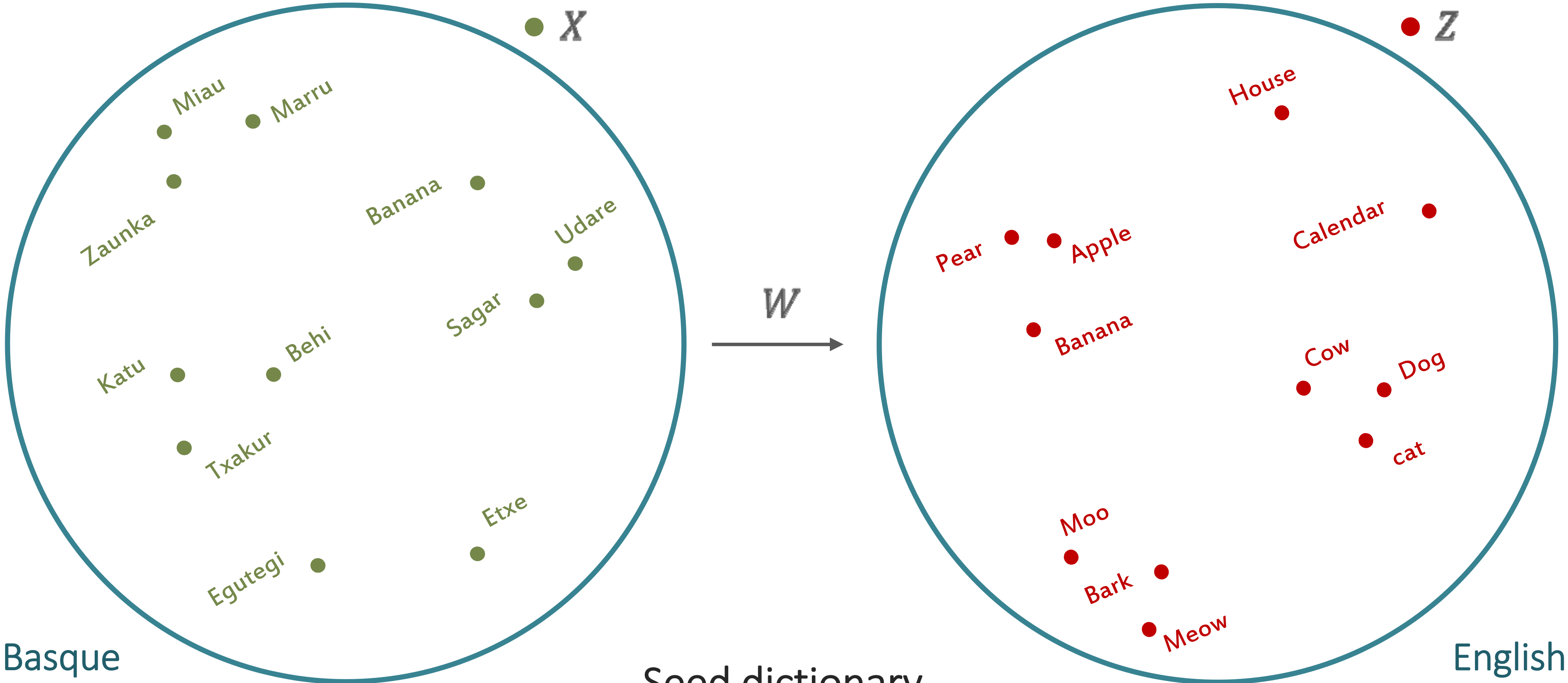
Introduction to embedding mappings



Seed dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

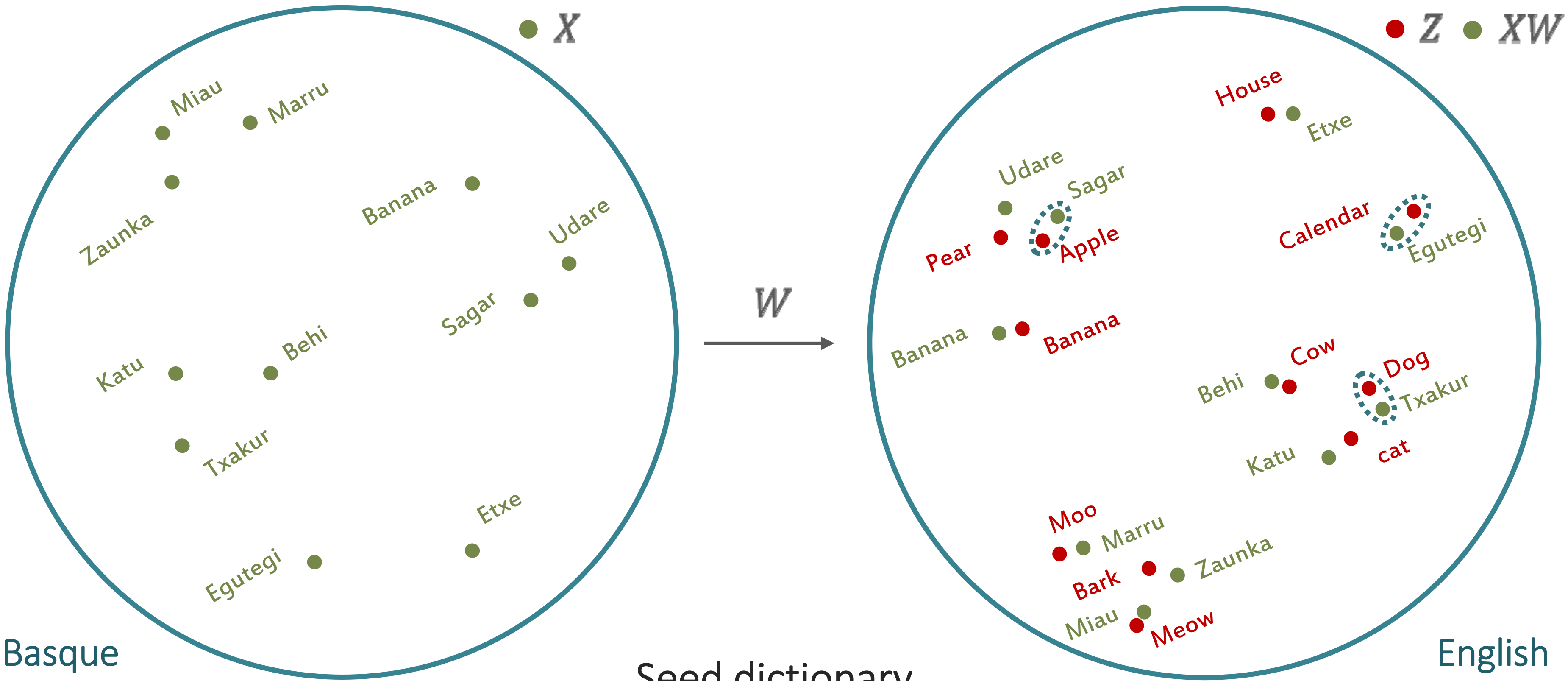
Introduction to embedding mappings



Seed dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

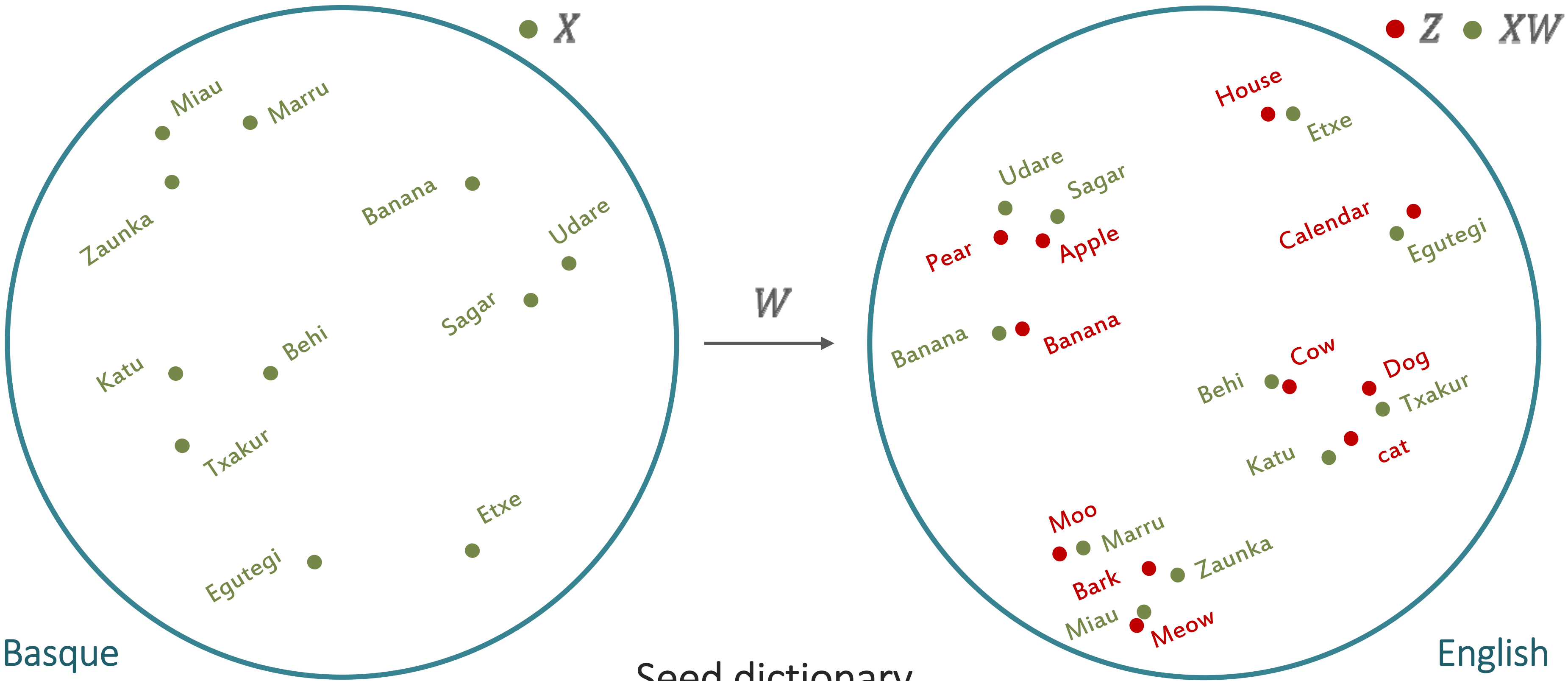
Introduction to embedding mappings



Seed dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

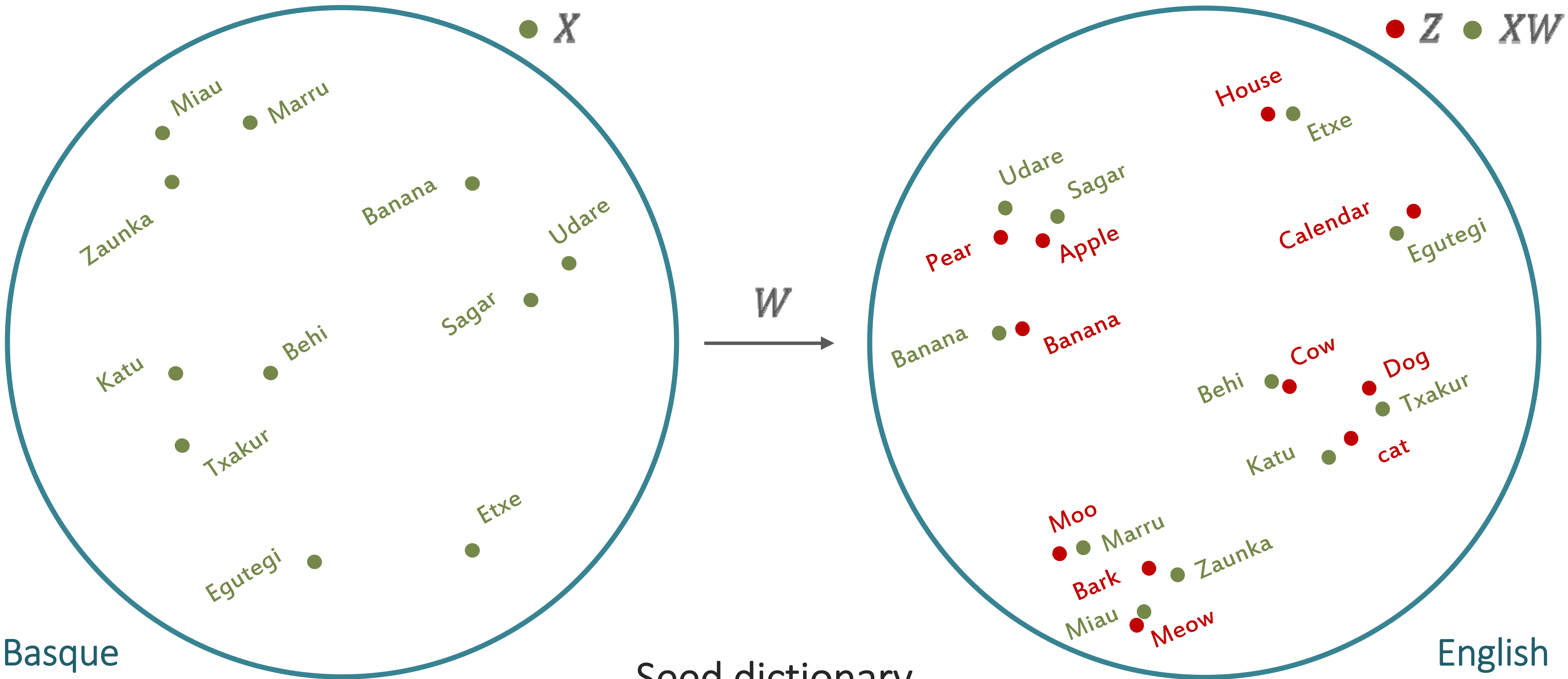
Introduction to embedding mappings



Seed dictionary

$$
\begin{array}{r}
\text{Txakur} \\ \text{Sagar} \\ \vdots \\ \text{Egutegi}
\end{array}
\begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix}
\quad \Bigg| \quad
\begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix}
\begin{array}{l}
\text{Dog} \\ \text{Apple} \\ \vdots \\ \text{Calendar}
\end{array}
$$

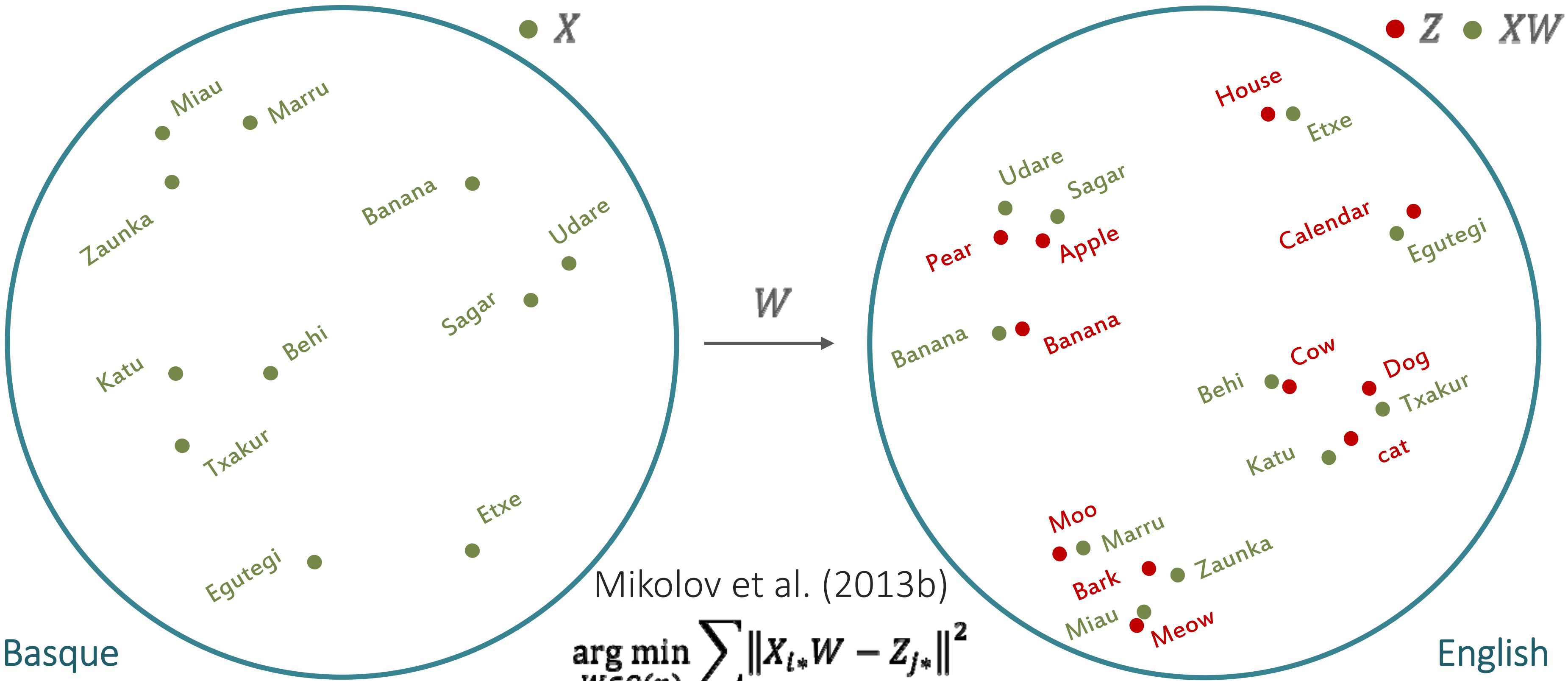
Introduction to embedding mappings



Seed dictionary

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

Introduction to embedding mappings

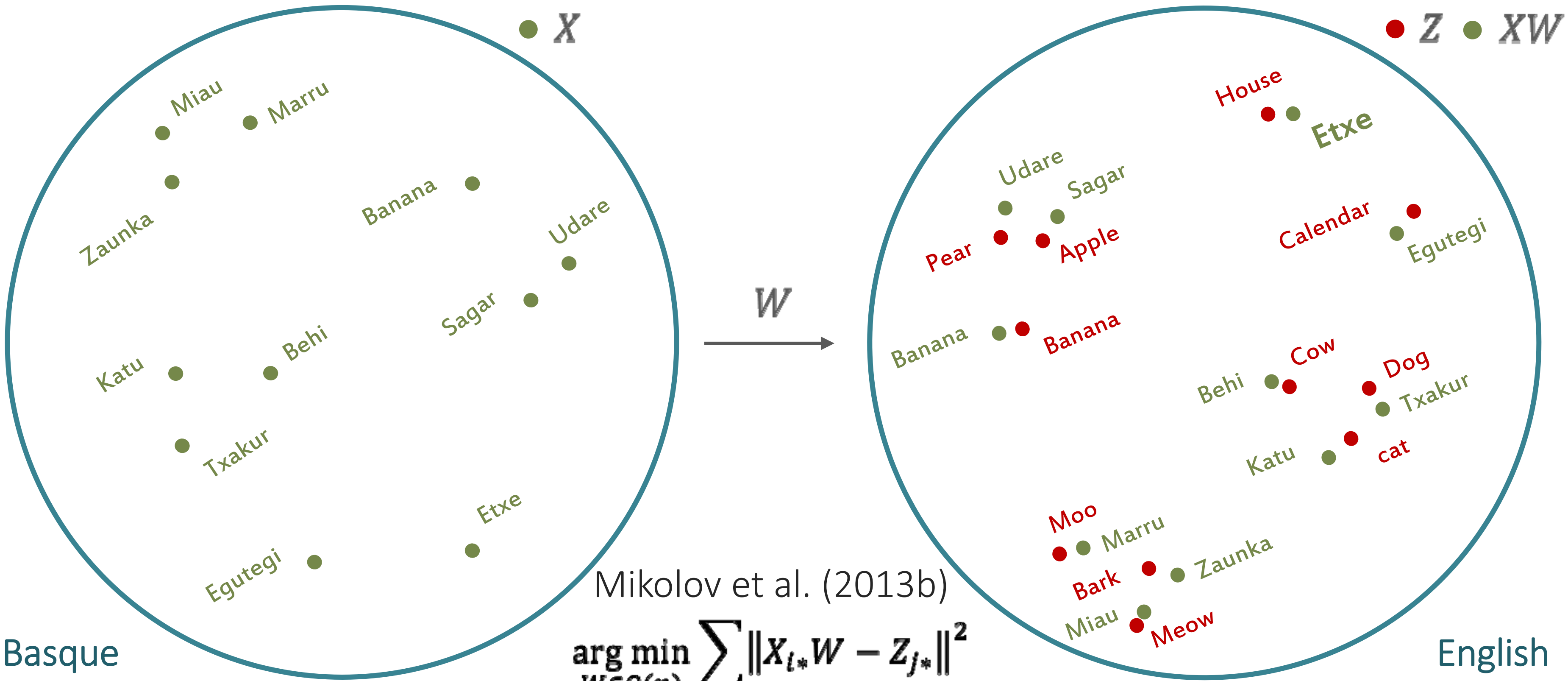


Mikolov et al. (2013b)

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

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

Introduction to embedding mappings

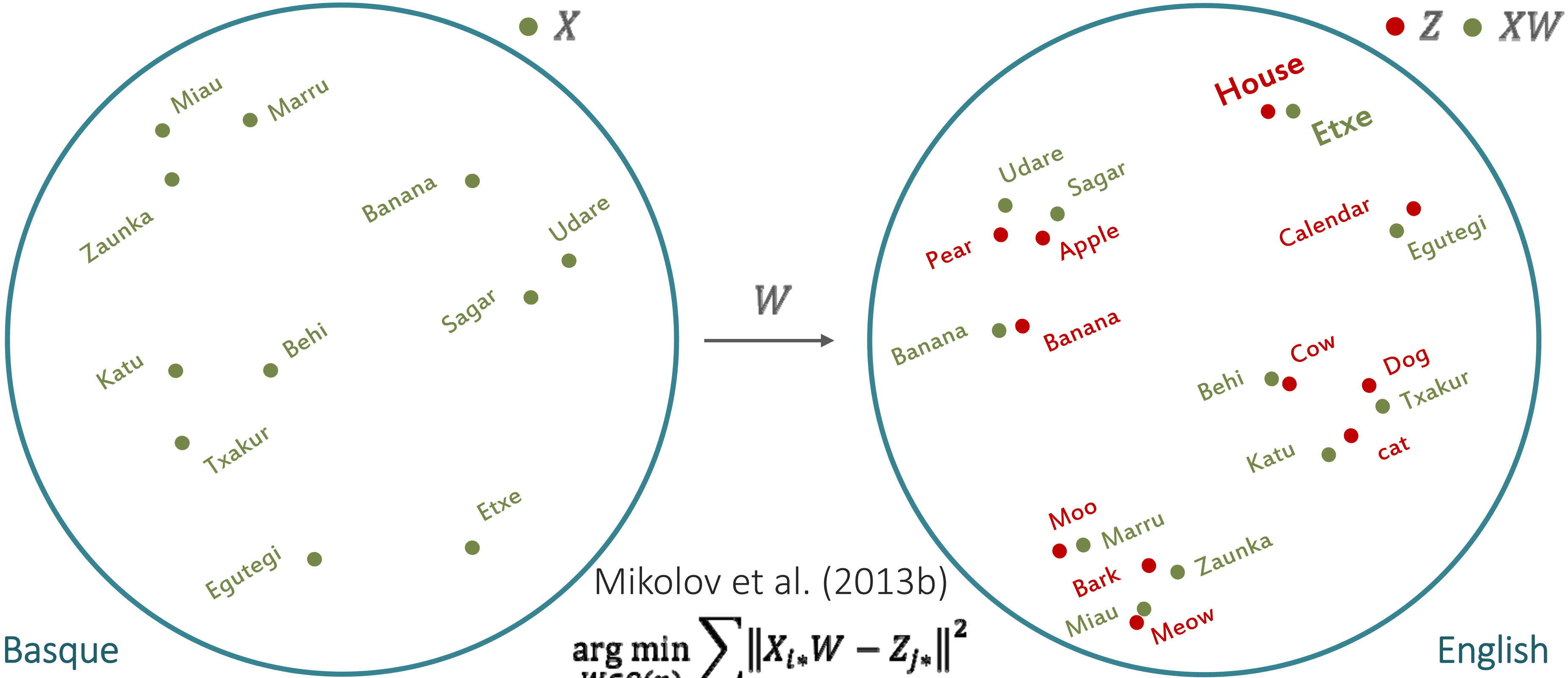


Mikolov et al. (2013b)

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

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

Introduction to embedding mappings

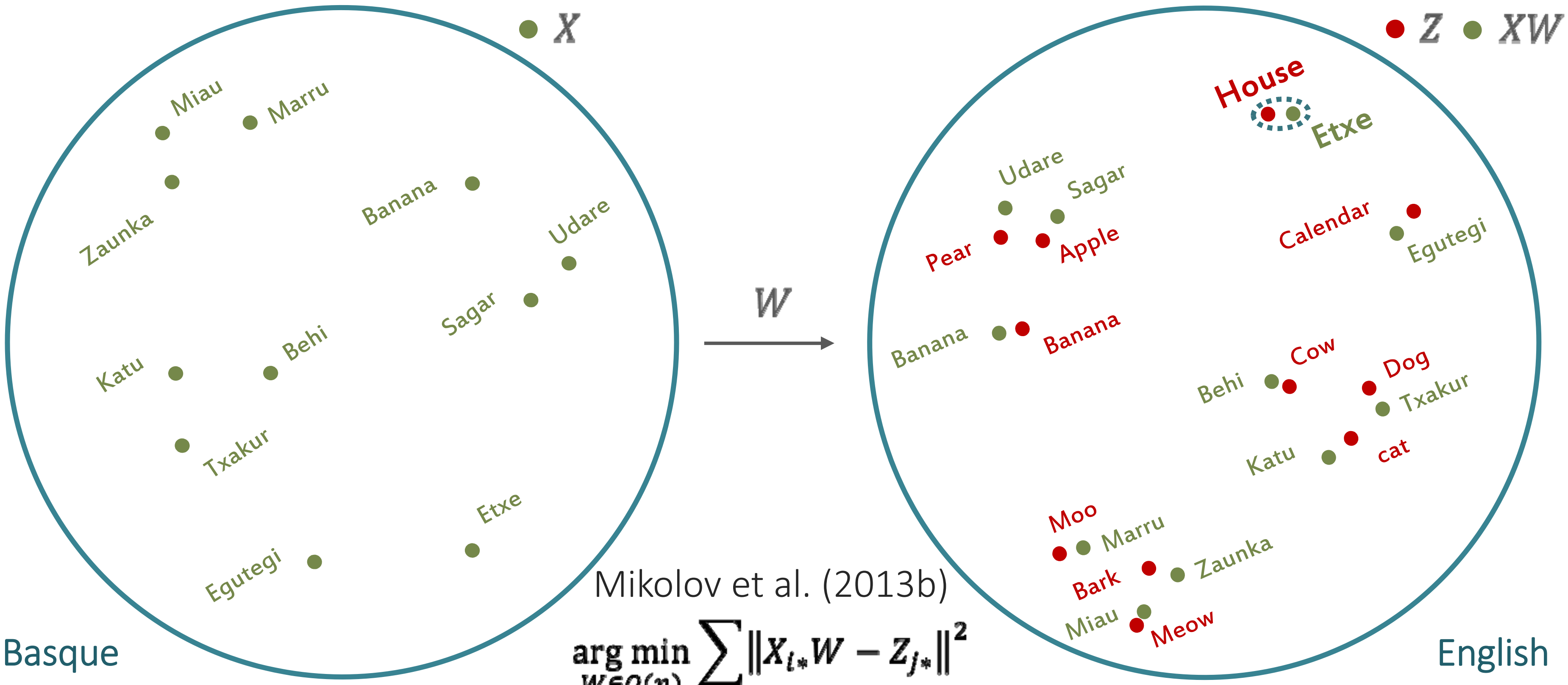


Mikolov et al. (2013b)

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

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

Introduction to embedding mappings



Mikolov et al. (2013b)

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

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

State-of-the-art in **supervised** mappings Artetxe et al. AAAI 2018

- Use 5000 sized seed bilingual dictionary
- Framework subsuming previous work that 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

State-of-the-art in **supervised** mappings

S0 (opt.) Pre-processing: length normalization, mean centering

State-of-the-art in **supervised** mappings

Two sequences of (optional) linear transformations:

$$W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)}$$

S0 (opt.) Pre-processing: length normalization, mean centering

State-of-the-art in **supervised** mappings

Two sequences of (optional) linear transformations:

$$W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)}$$

S0 (opt.) Pre-processing: length normalization, mean centering

S1 (opt.) Whitening : turn covariance matrices into the identity matrix

$$W_{X(1)} = (X^T X)^{-0.5}$$
$$W_{Z(1)} = (Z^T Z)^{-0.5}$$

State-of-the-art in **supervised** mappings

Two sequences of (optional) linear transformations:

$$W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)}$$

S0 (opt.) Pre-processing: length normalization, mean centering

S1 (opt.) Whitening : turn covariance matrices into the identity matrix

$$W_{X(1)} = (X^T X)^{-0.5}$$
$$W_{Z(1)} = (Z^T Z)^{-0.5}$$

S2 Orthogonal mapping: map into a shared space (Procrustes)

$$W_{X(2)} = U \quad W_{Z(2)} = V \quad USV^T = X_{(1)}^T Z_{(1)}$$

State-of-the-art in **supervised** mappings

Two sequences of (optional) linear transformations:

$$W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)}$$

S0 (opt.) Pre-processing: length normalization, mean centering

S1 (opt.) Whitening : turn covariance matrices into the identity matrix $\quad W_{X(1)} = (X^T X)^{-0.5}$, $W_{Z(1)} = (Z^T Z)^{-0.5}$

S2 Orthogonal mapping: map into a shared space (Procrustes) $\quad W_{X(2)} = U$, $W_{Z(2)} = V$ $\quad USV^T = X_{(1)}^T Z_{(1)}$

S3 (opt.) Re-weight each component according to its cross-correlation $\quad W_{X(3)} = S,\ W_{Z(3)} = I$; $W_{X(3)} = I,\ W_{Z(3)} = S$

State-of-the-art in **supervised** mappings

Two sequences of (optional) linear transformations:

$$W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)}$$

S0 (opt.) Pre-processing: length normalization, mean centering

S1 (opt.) Whitening : turn covariance matrices into the identity matrix

$$W_{X(1)} = (X^T X)^{-0.5}$$
$$W_{Z(1)} = (Z^T Z)^{-0.5}$$

S2 Orthogonal mapping: map into a shared space (Procrustes)

$$W_{X(2)} = U \qquad USV^T = X_{(1)}^T Z_{(1)}$$
$$W_{Z(2)} = V$$

S3 (opt.) Re-weight each component according to its cross-correlation

$$W_{X(3)} = S, \; W_{Z(3)} = I$$
$$W_{X(3)} = I, \; W_{Z(3)} = S$$

S4 (opt.) De-whitening: restore original variance in every direction

$$W_{A(4)} = W_{B(2)}^T W_{B(1)}^{-1} W_{B(2)}$$

State-of-the-art in supervised mappings

Two sequences of (optional) linear transformations:

$$W_X = \prod_i W_{X(i)} \quad W_Z = \prod_i W_{Z(i)}$$

S0 (opt.) Pre-processing: length normalization, mean centering

S1 (opt.) Whitening : turn covariance matrices into the identity matrix

$$W_{X(1)} = (X^T X)^{-0.5}$$
$$W_{Z(1)} = (Z^T Z)^{-0.5}$$

S2 Orthogonal mapping: map into a shared space (Procrustes)

$$W_{X(2)} = U \qquad USV^T = X_{(1)}^T Z_{(1)}$$
$$W_{Z(2)} = V$$

S3 (opt.) Re-weight each component according to its cross-correlation

$$W_{X(3)} = S,\ W_{Z(3)} = I$$
$$W_{X(3)} = I,\ W_{Z(3)} = S$$

S4 (opt.) De-whitening: restore original variance in every direction

$$W_{A(4)} = W_{B(2)}^T W_{B(1)}^{-1} W_{B(2)}$$

S5 (opt) Dimensionality reduction: keep the first *n* components only

$$W_{X(5)} = W_{Z(5)} = (I_n\ 0)^T$$

State-of-the-art in **supervised** mappings

		S0 (l)	S0 (m)	S1	S2	S3	S4 (src)	S4 (trg)	S5
OLS	Mikolov et al. (2013)			x	x	src	trg	trg	
	Shigeto et al. (2015)			x	x	trg	src	src	
CCA	Faruqui and Dyer (2014)	x	x	x	x				x
Orth.	Xing et al. (2015)	x			x				
	Artetxe et al. (2016)	x	x		x				
	Zhang et al. (2016)				x				
	Smith et al. (2017)	x			x				x

State-of-the-art in **supervised** mappings

		S0 (l)	S0 (m)	S1	S2	S3	S4 (src)	S4 (trg)	S5
OLS	Mikolov et al. (2013)			x	x	src	trg	trg	
	Shigeto et al. (2015)			x	x	trg	src	src	
CCA	Faruqui and Dyer (2014)	x	x	x	x				x
Orth.	Xing et al. (2015)	x			x				
	Artetxe et al. (2016)	x	x		x				
	Zhang et al. (2016)				x				
	Smith et al. (2017)	x			x				x
	Our method (AAAI18)	x	x	x	x	trg	src	trg	x

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

$\Rightarrow$ 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 publicly available reimplementation

Evaluating via Bilingual Dictionary induction

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

⇒ Monolingual embeddings (CBOW + negative sampling)

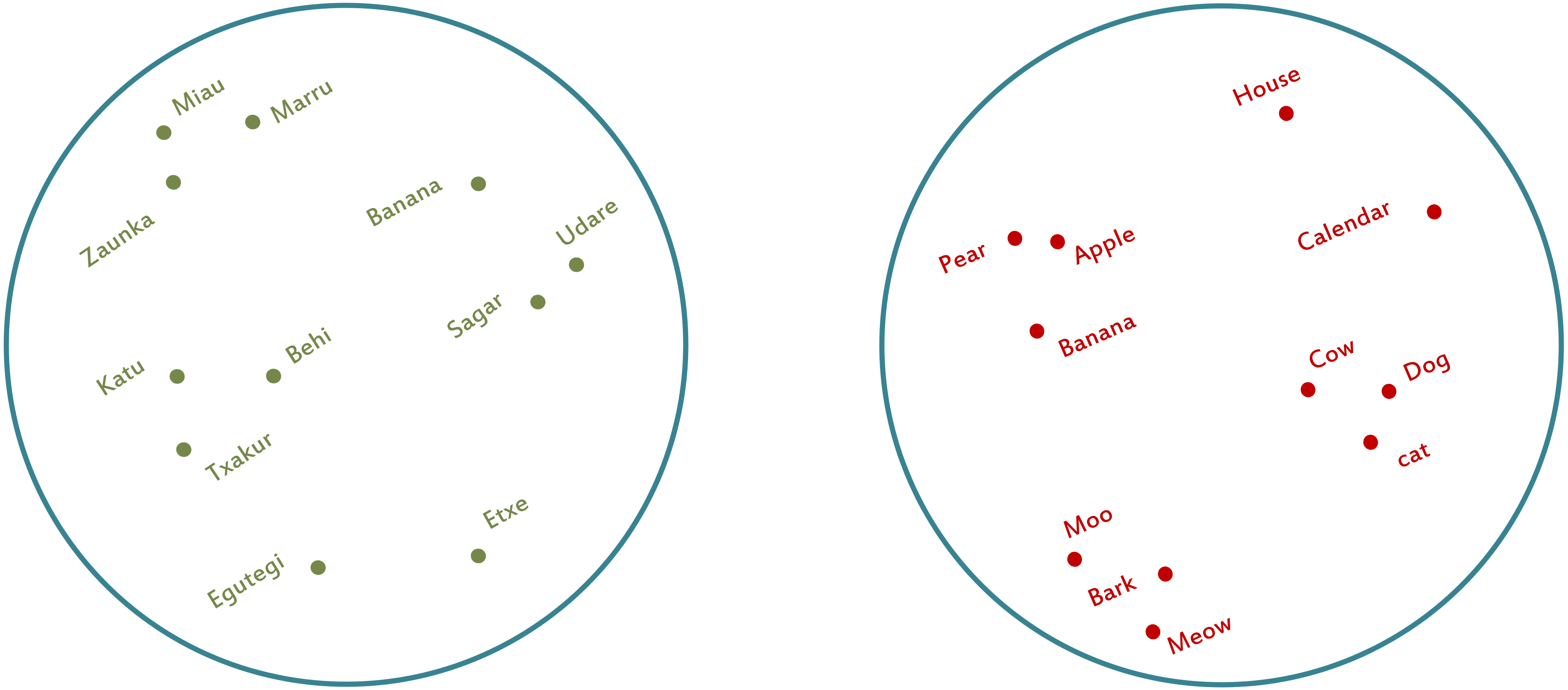
⇒ Seed dictionary: 5,000 pairs

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

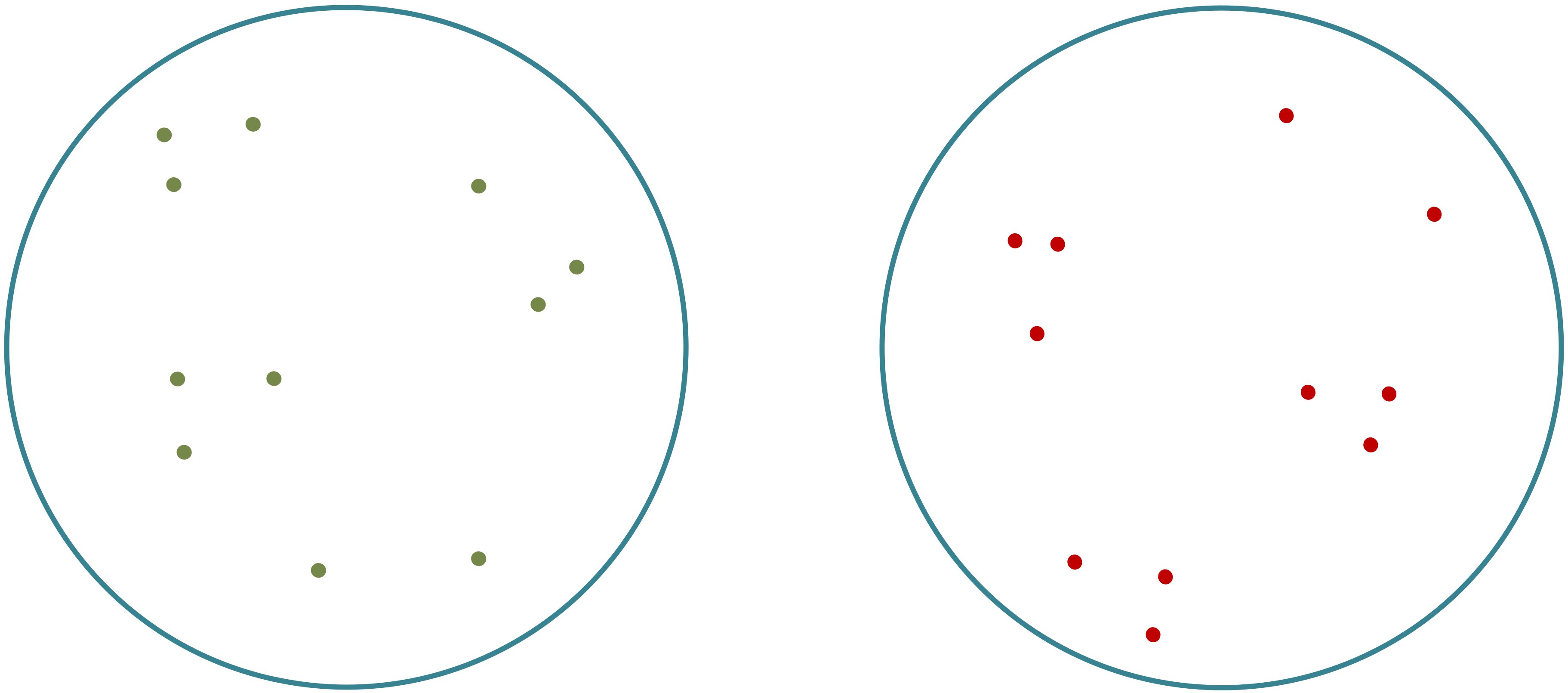
Method	EN-IT	EN-DE	EN-FI	EN-ES
Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
Faruqui and Dyer (2014)	38.40*	37.13*	27.60*	26.80*
Shigeto et al. (2015)	41.53†	43.07†	31.04†	33.73†
Dinu et al. (2015)	37.7	38.93*	29.14*	30.40*
Lazaridou et al. (2015)	40.2	-	-	-
Xing et al. (2015)	36.87†	41.27†	28.23†	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?

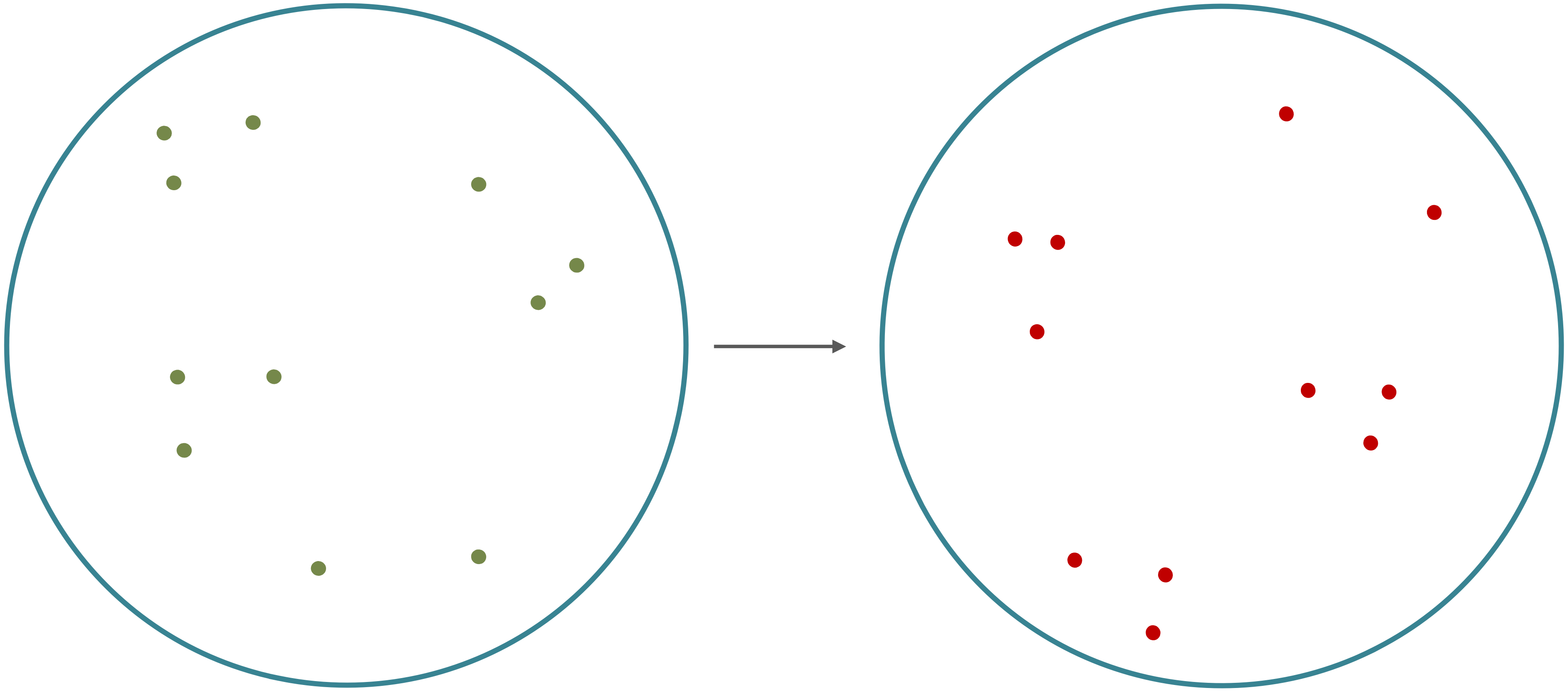
Why does it work?



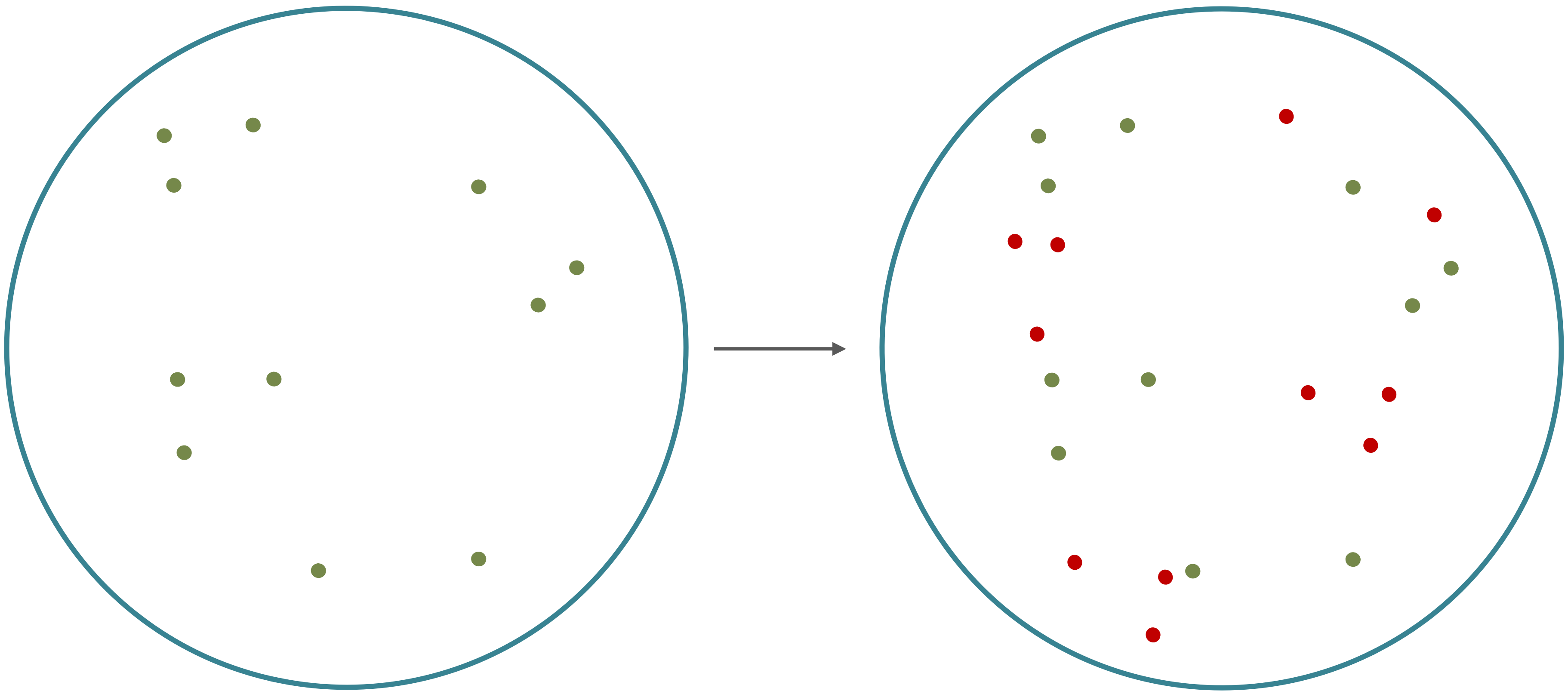
Why does it work?



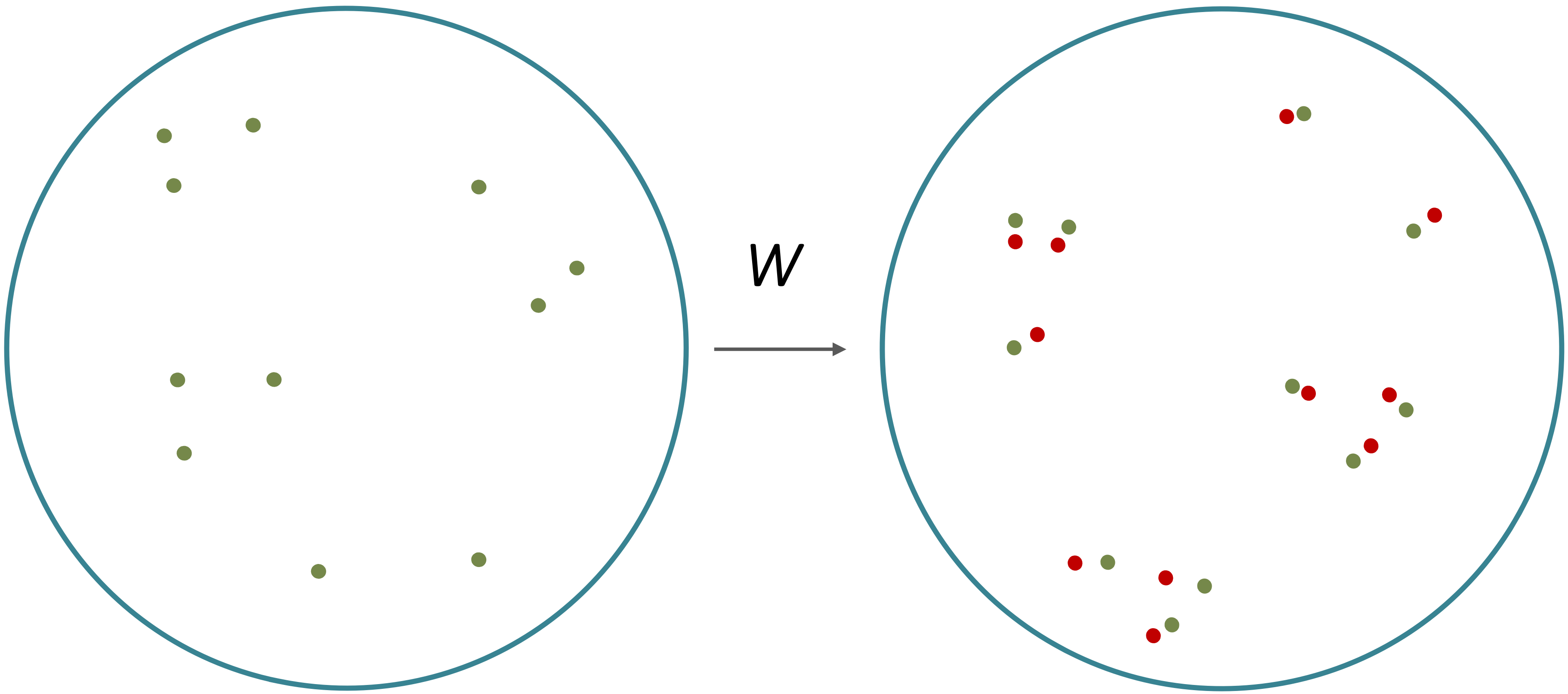
Why does it work?



Why does it work?

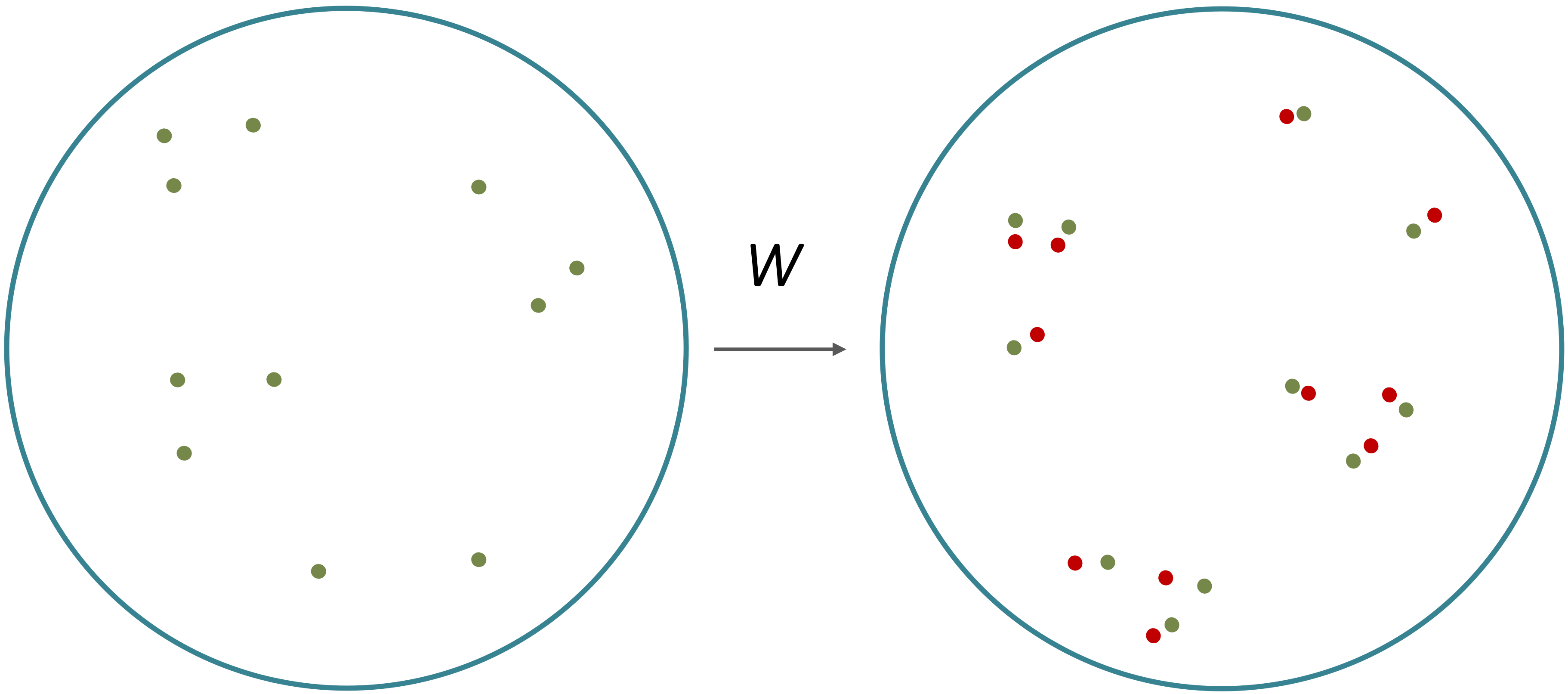


Why does it work?



Why does it work?

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

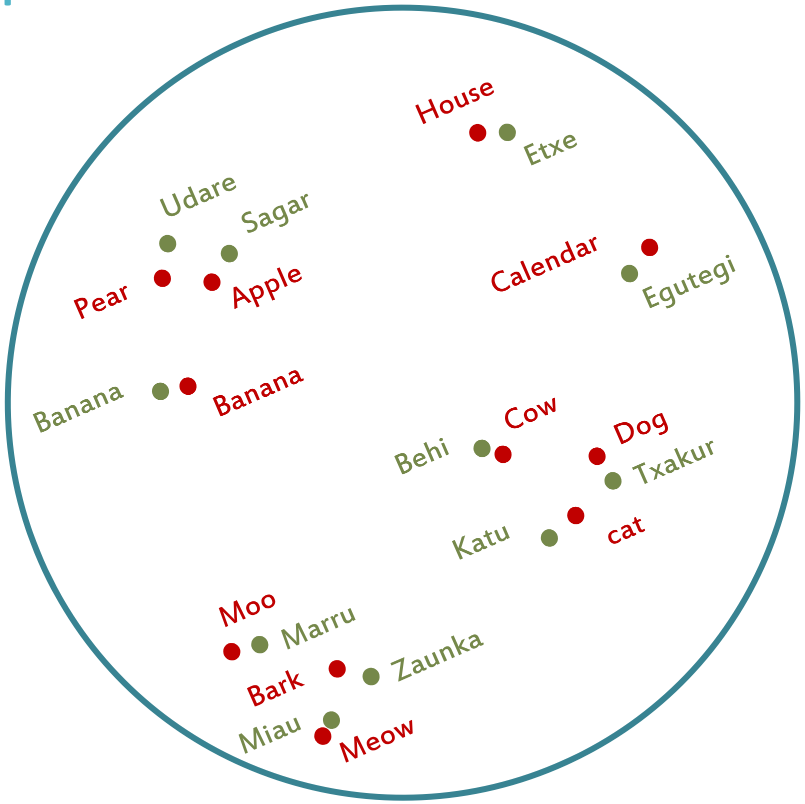


Outline

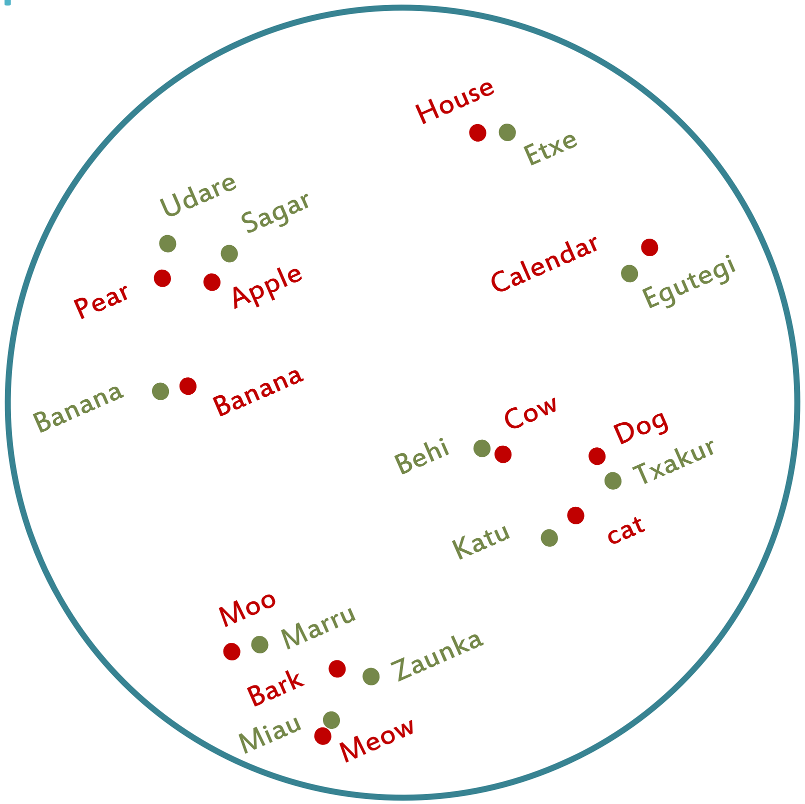
- Bilingual embedding mappings
  - *Introduction to vector space models (embeddings)*
  - *Bilingual embedding mappings (AAAI18)*
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions*
- Unsupervised neural machine translation
  - *Introduction to NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Unsupervised statistical MT (EMNLP18)*
  - *Conclusions*

Reducing supervision

Reducing supervision



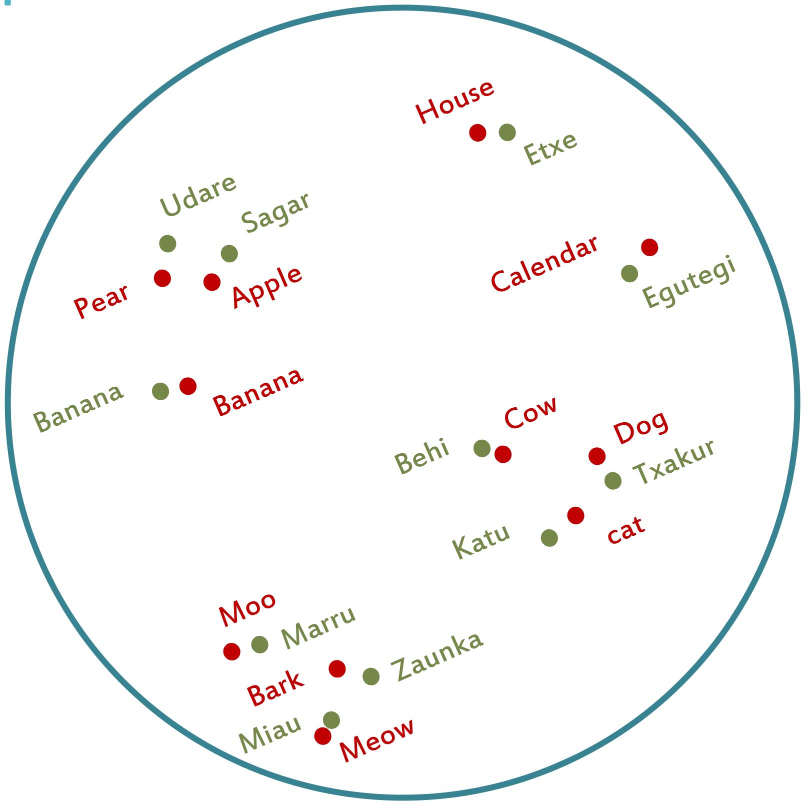
Reducing supervision



bilingual signal for training

Previous work

Reducing supervision

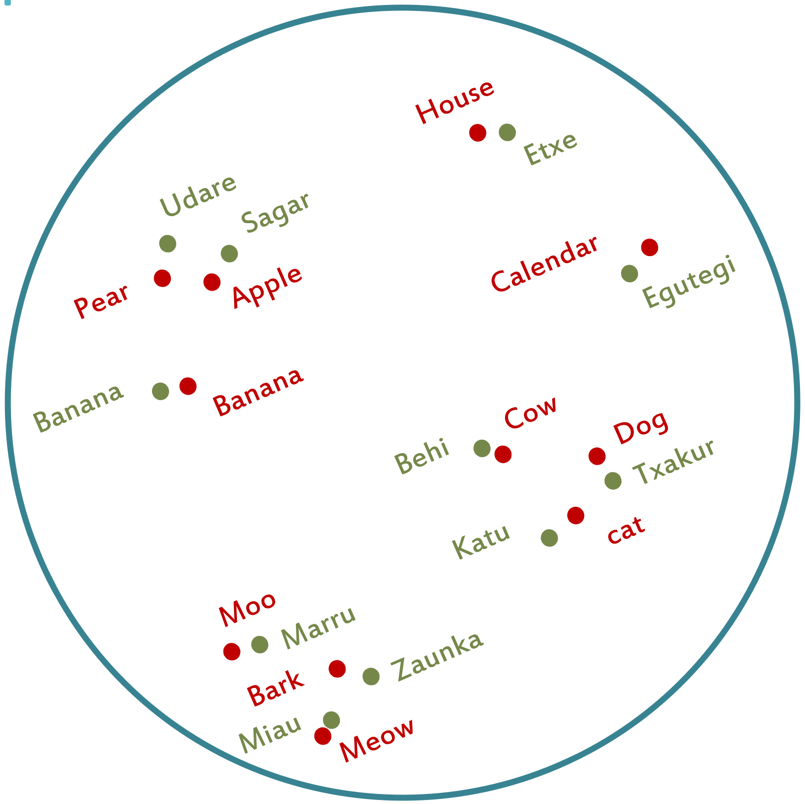


bilingual signal for training

<u>Previous work</u>

- parallel corpora
- comparable corpora
- (big) dictionaries

Reducing supervision

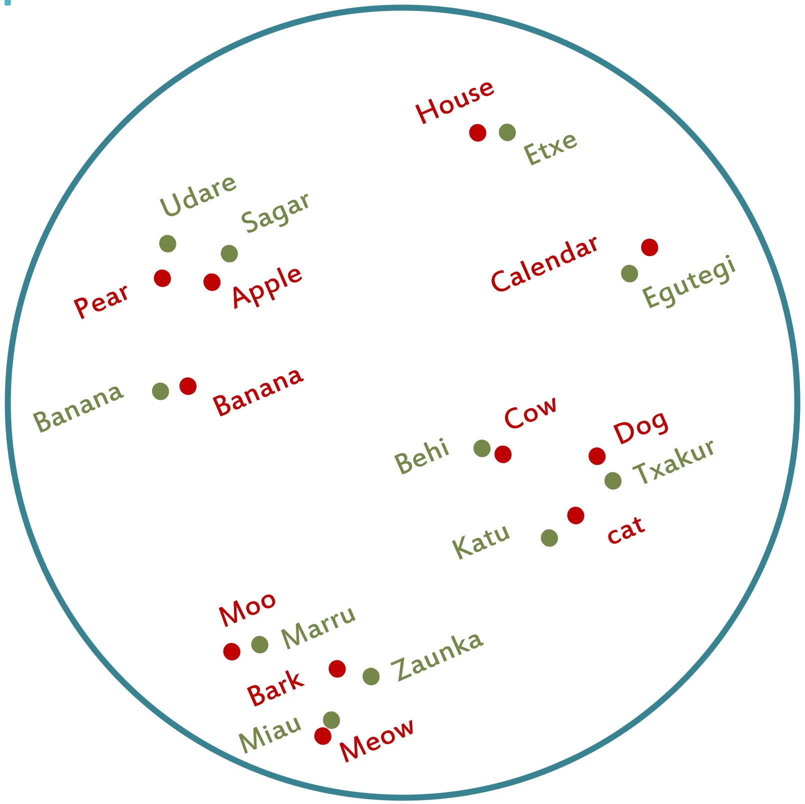


bilingual signal
for training

~~Previous work~~

- ~~parallel corpora~~
- ~~comparable corpora~~
- ~~(big) dictionaries~~

Reducing supervision



bilingual signal for training

~~Previous work~~	Our work
~~- parallel corpora~~	- 25 word dictionary
~~- comparable corpora~~	- numerals (1, 2, 3…)
~~- (big) dictionaries~~	- nothing

Self-learning

Self-learning

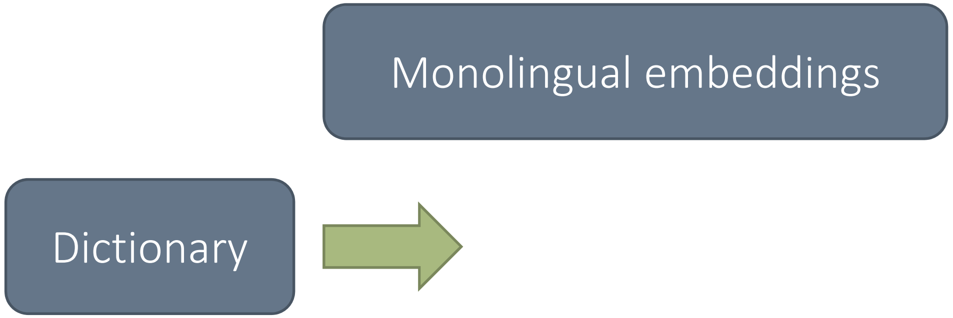
Monolingual embeddings

Self-learning

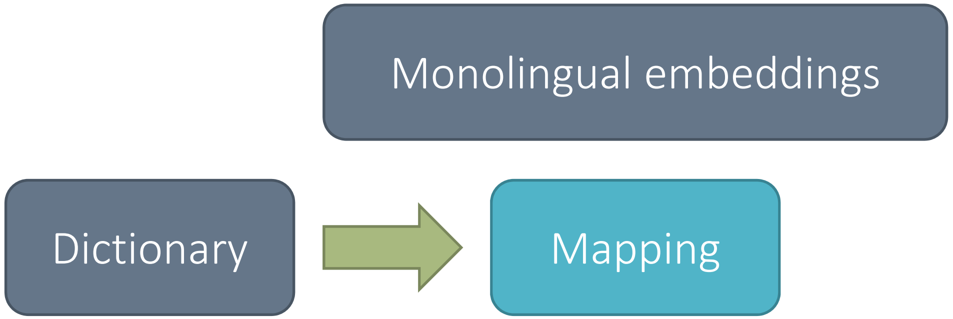
Monolingual embeddings

Dictionary

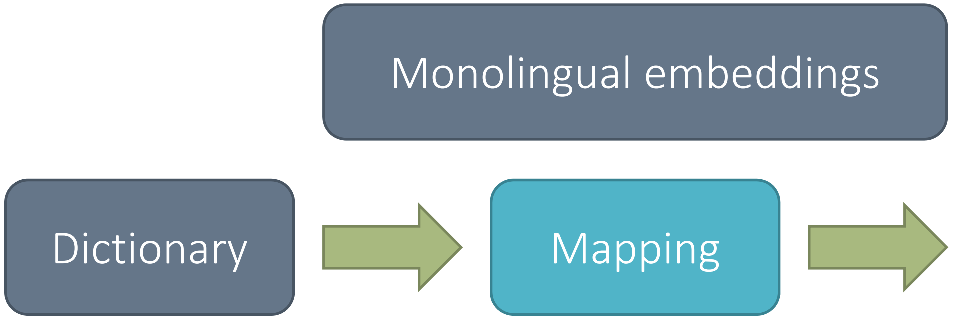
Self-learning



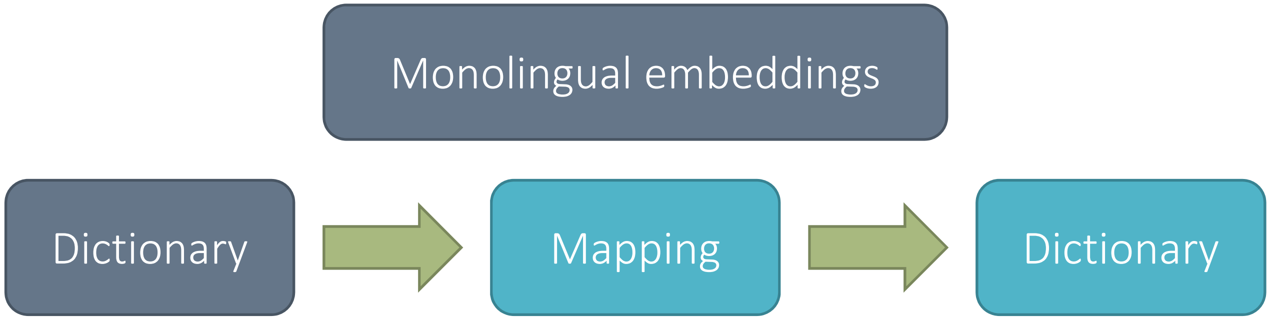
Self-learning



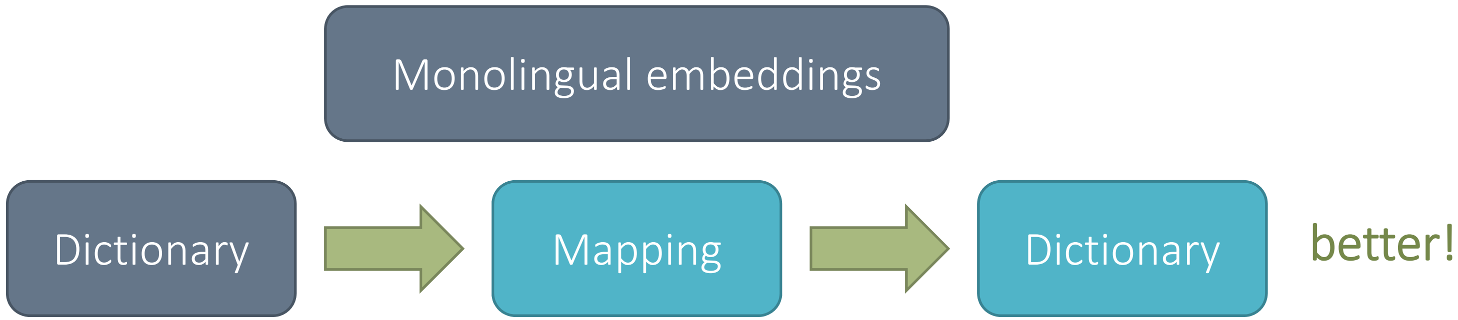
Self-learning



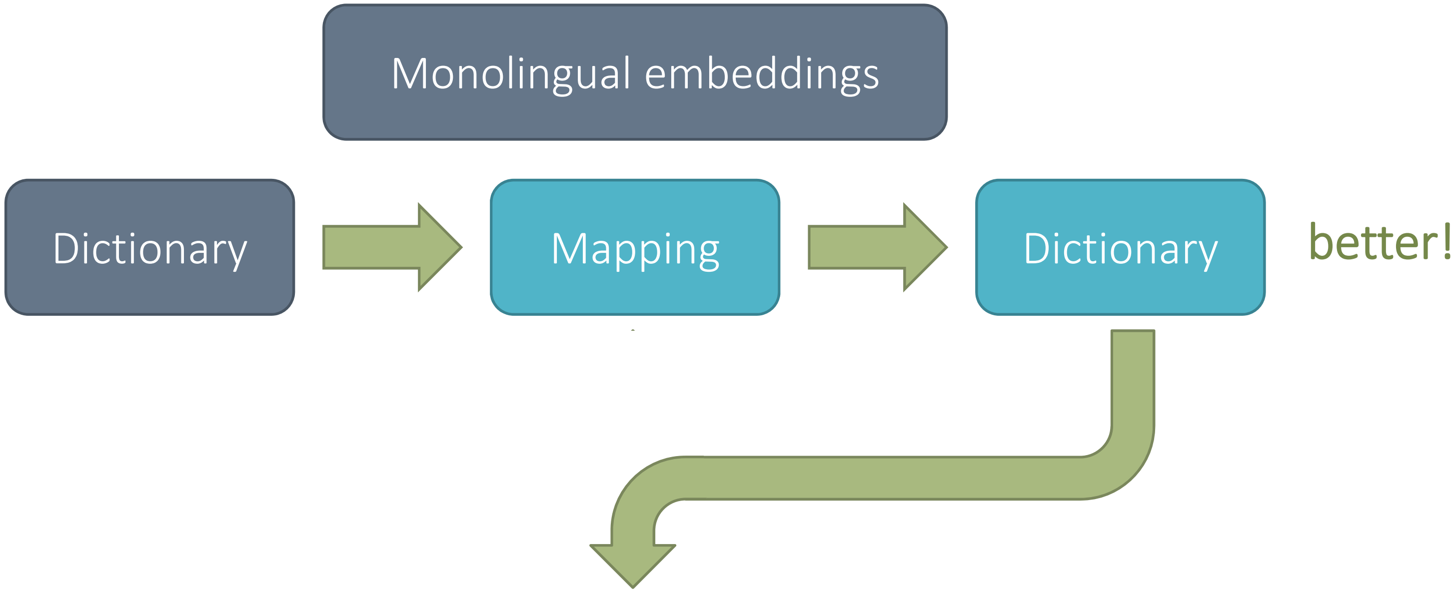
Self-learning



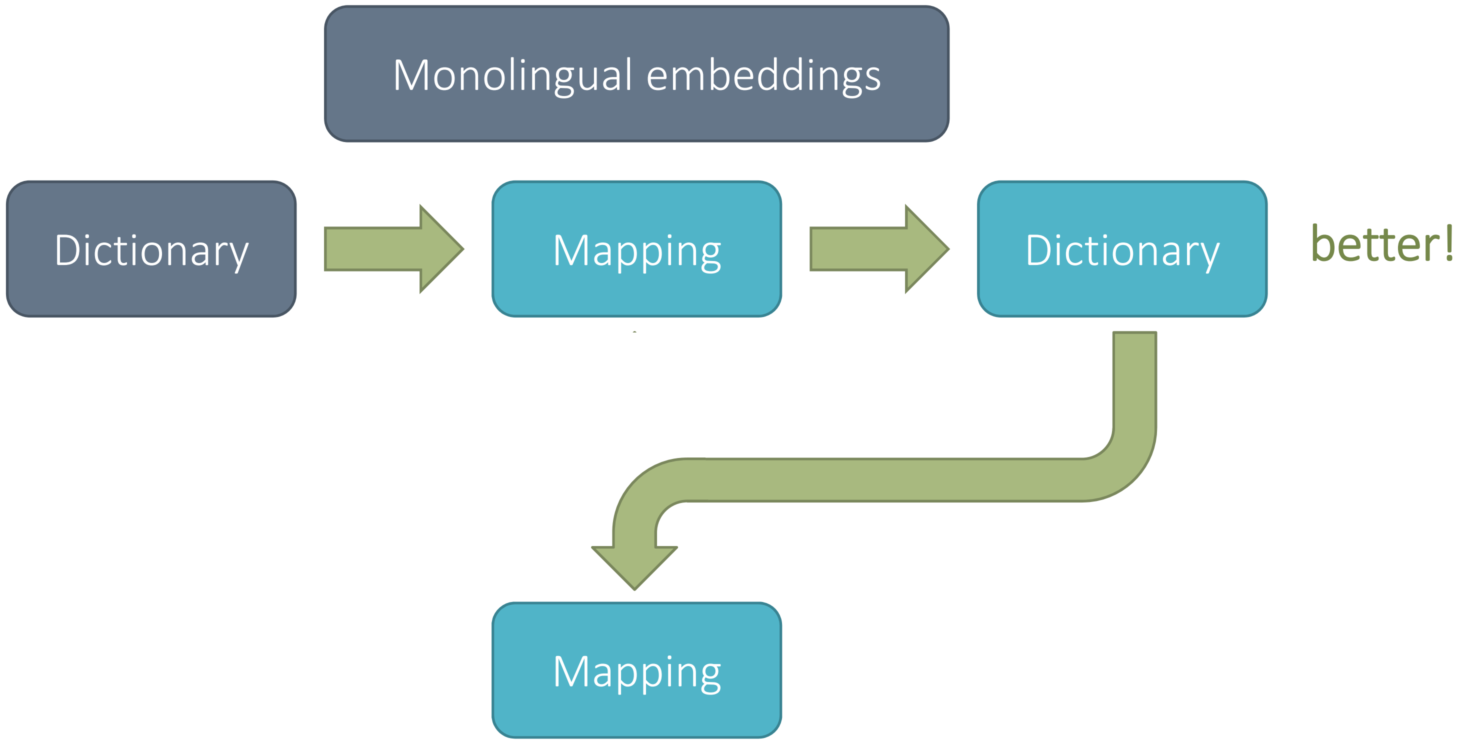
Self-learning



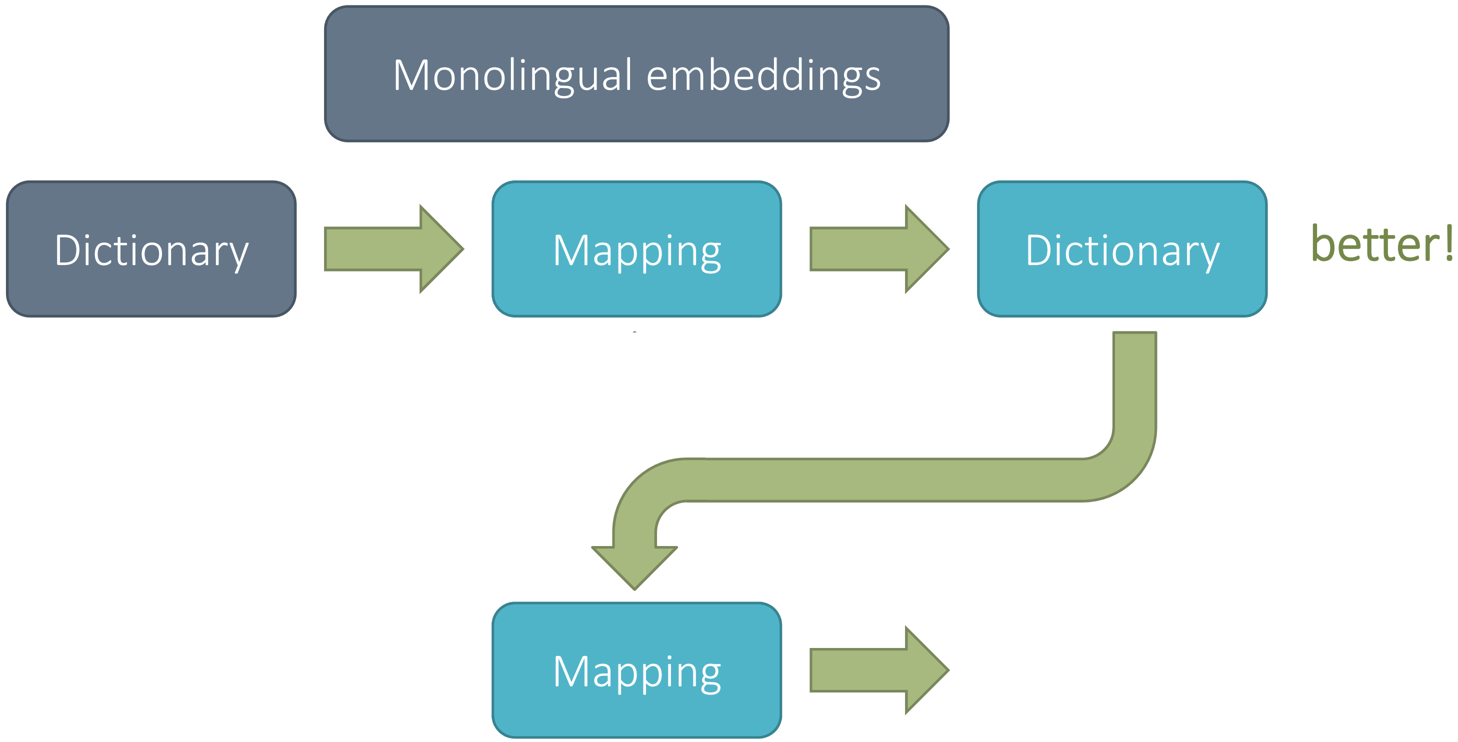
Self-learning



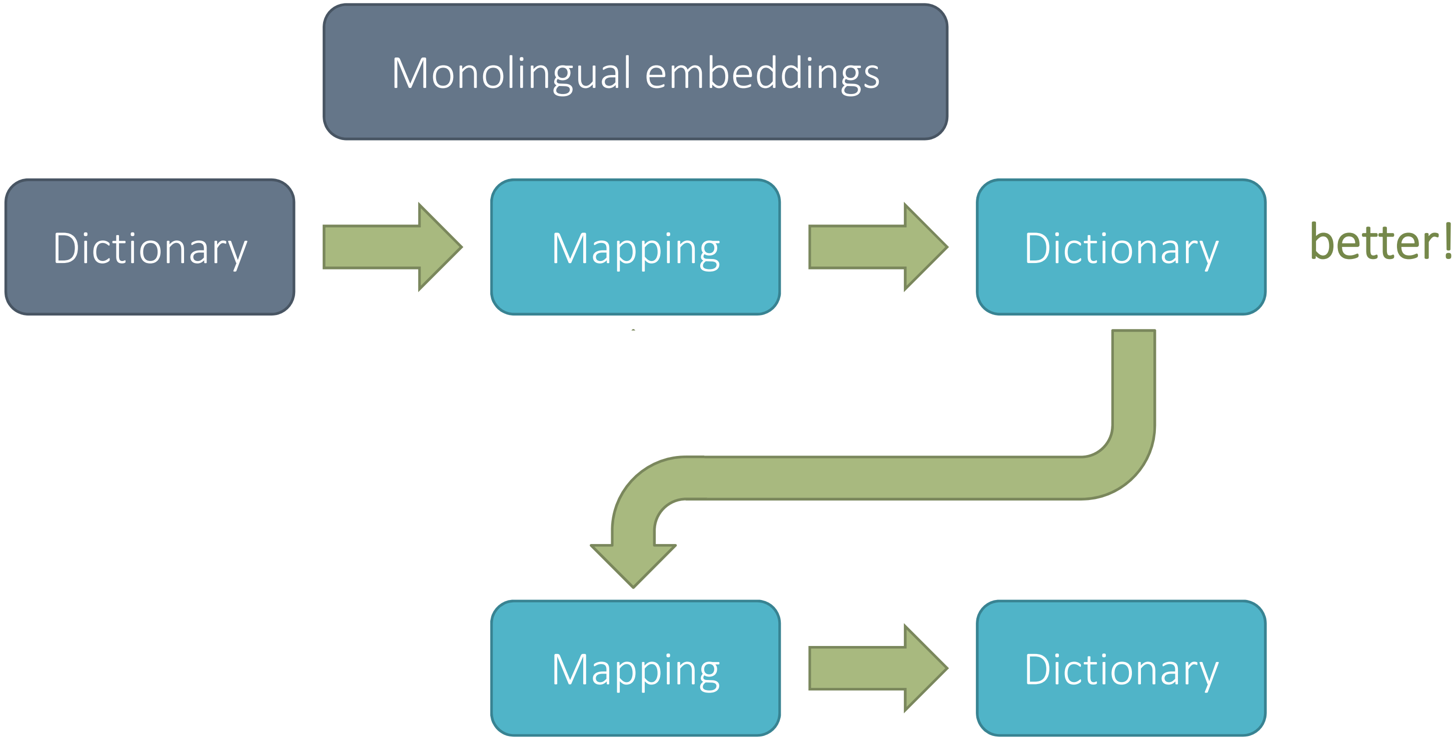
Self-learning



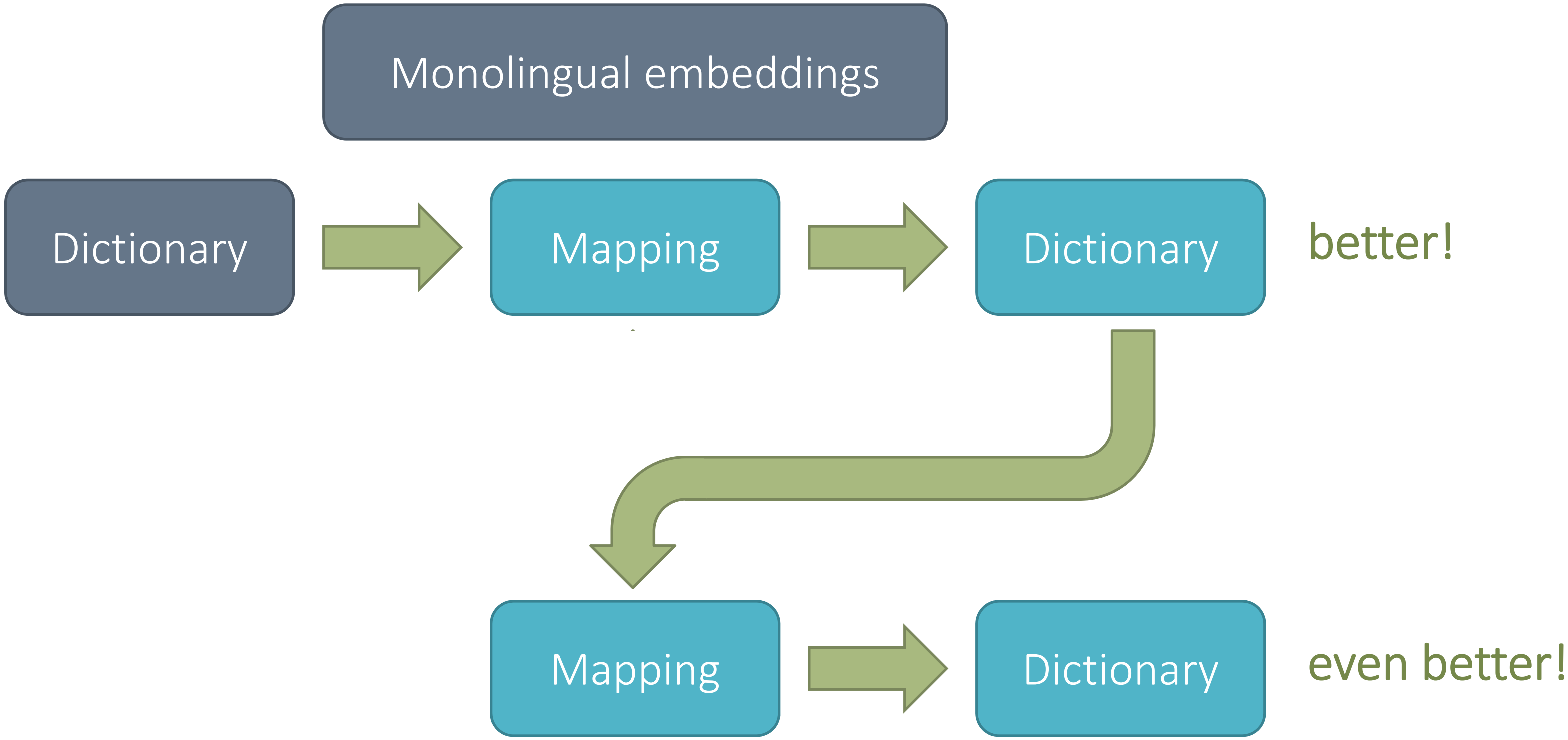
Self-learning



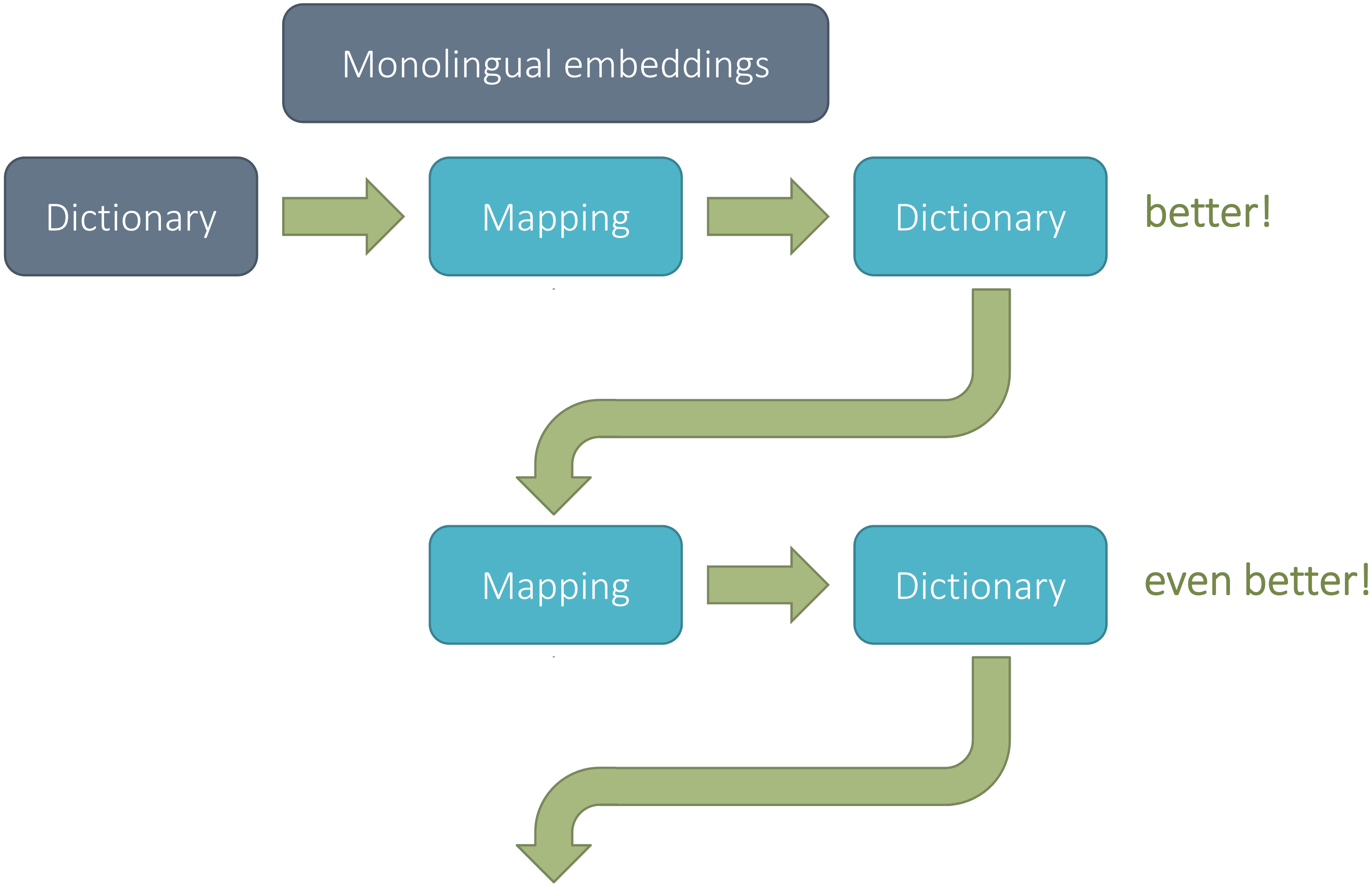
Self-learning



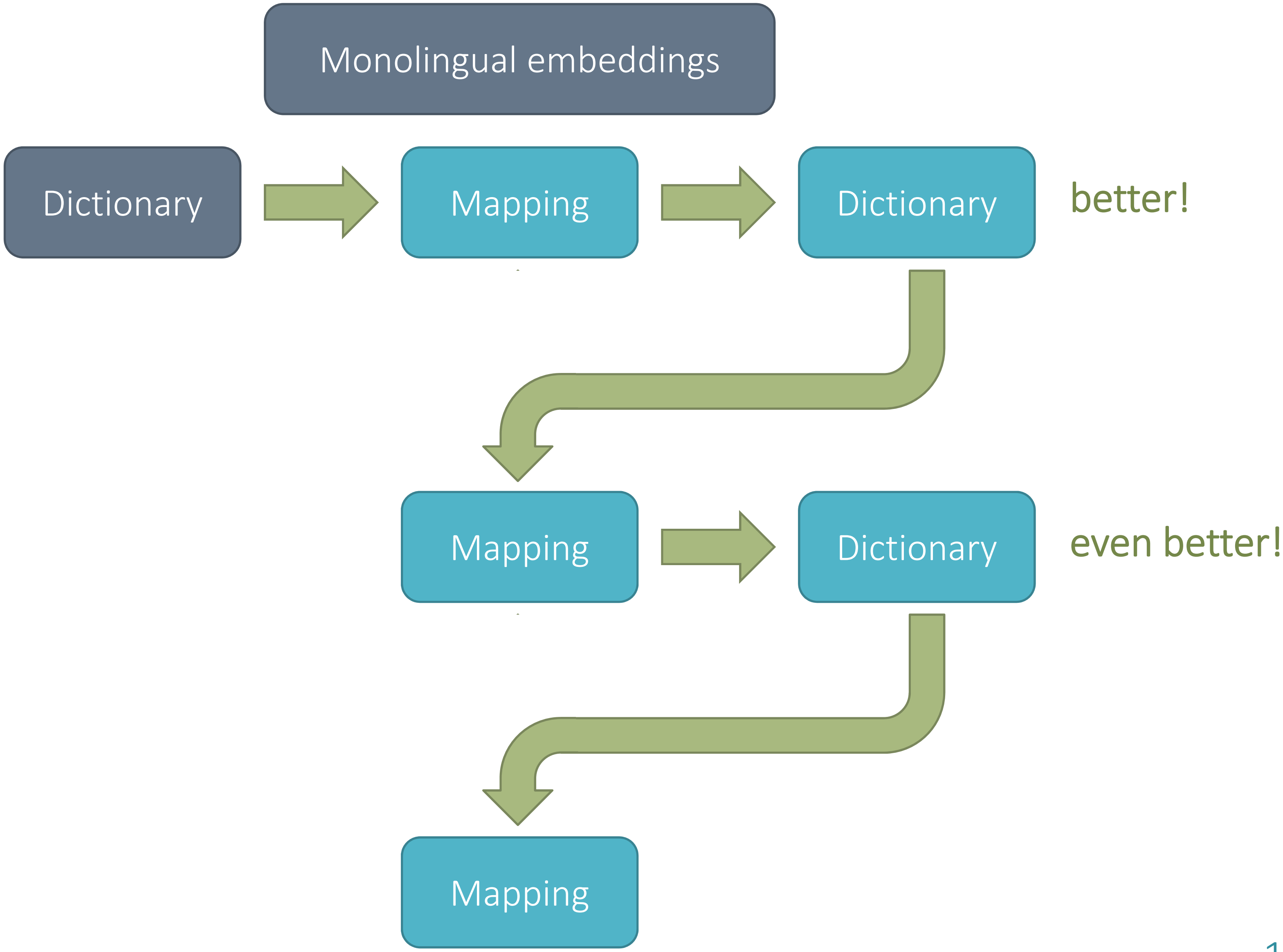
Self-learning



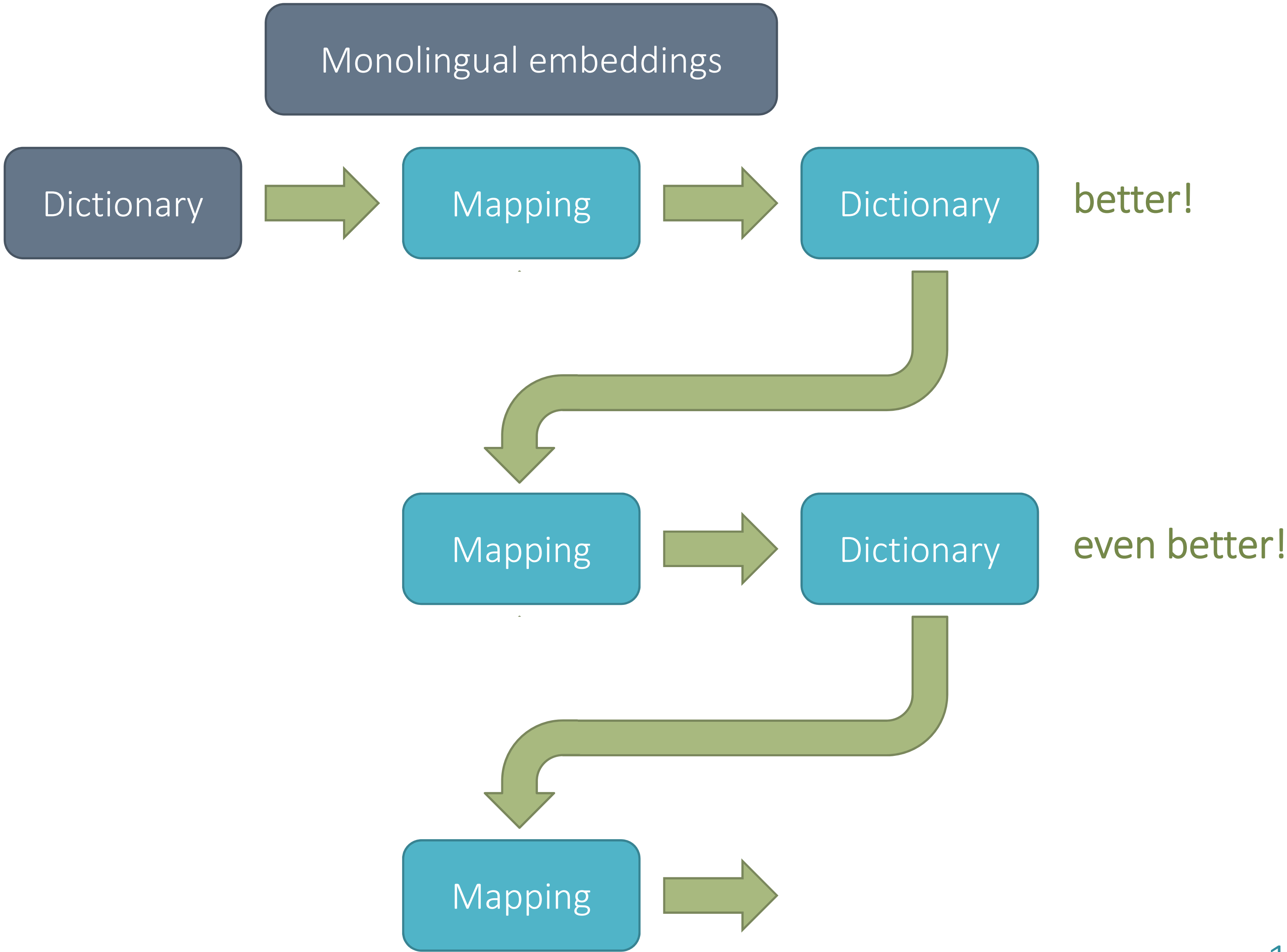
Self-learning



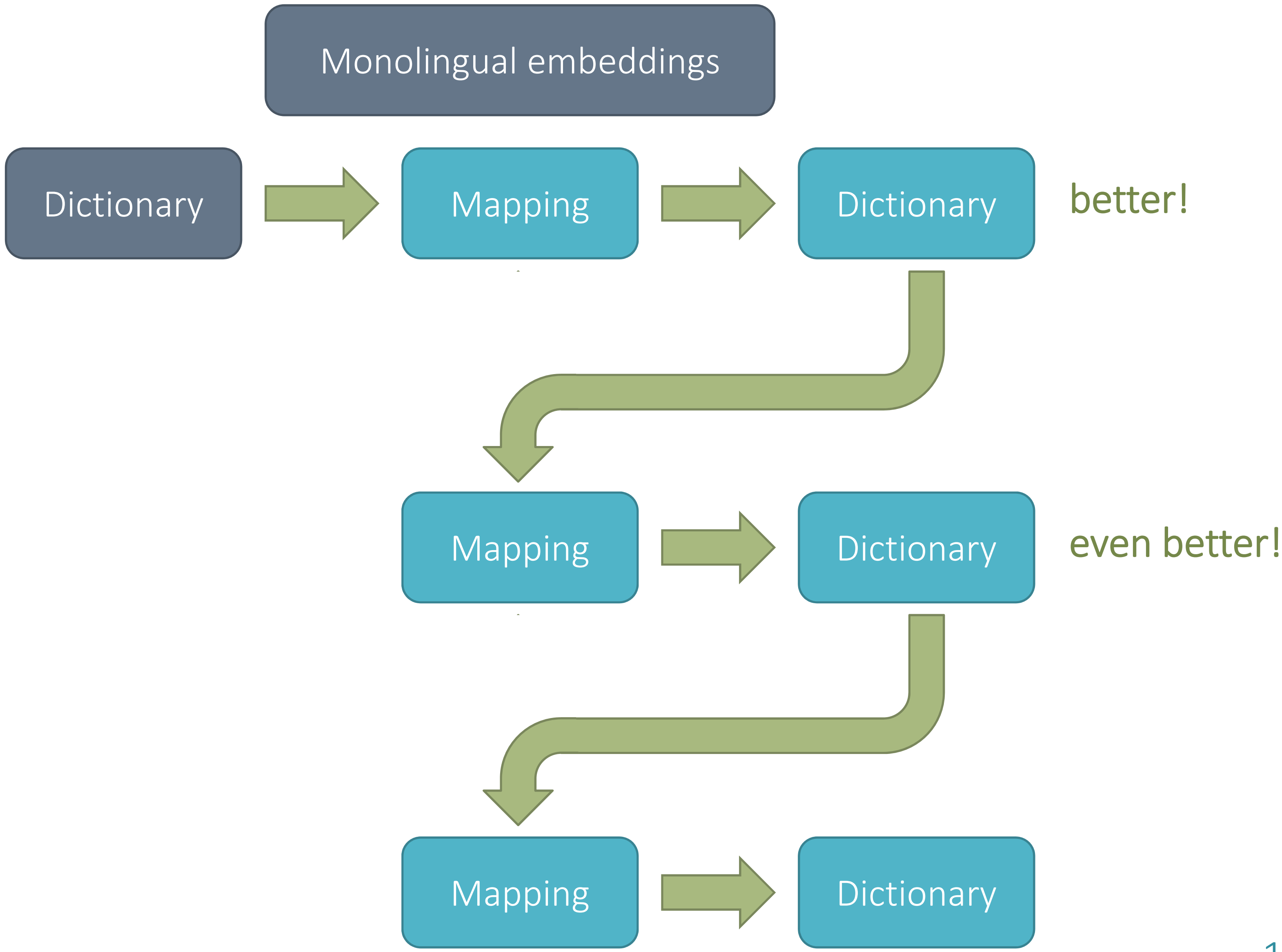
Self-learning



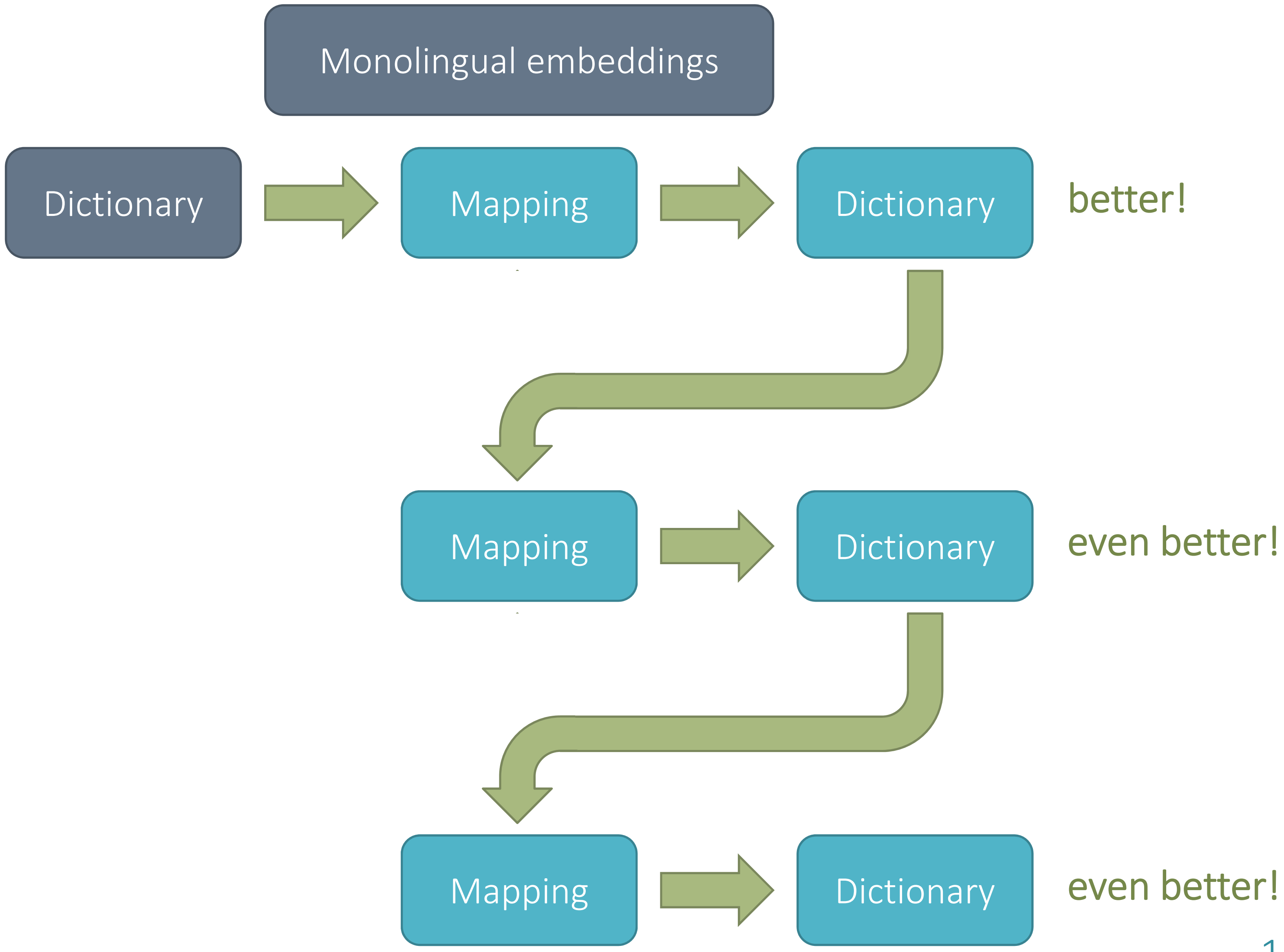
Self-learning



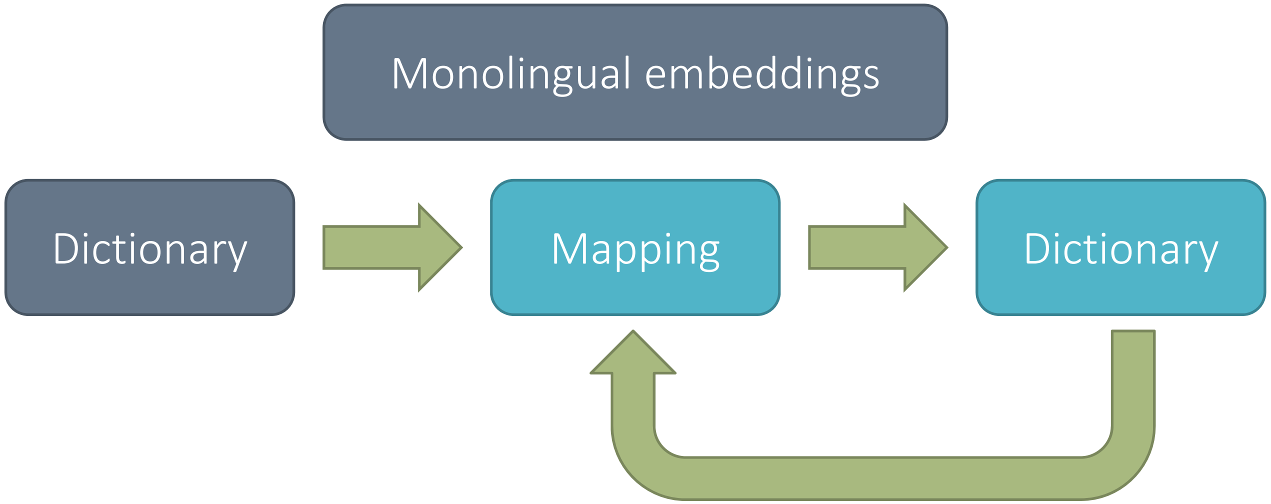
Self-learning



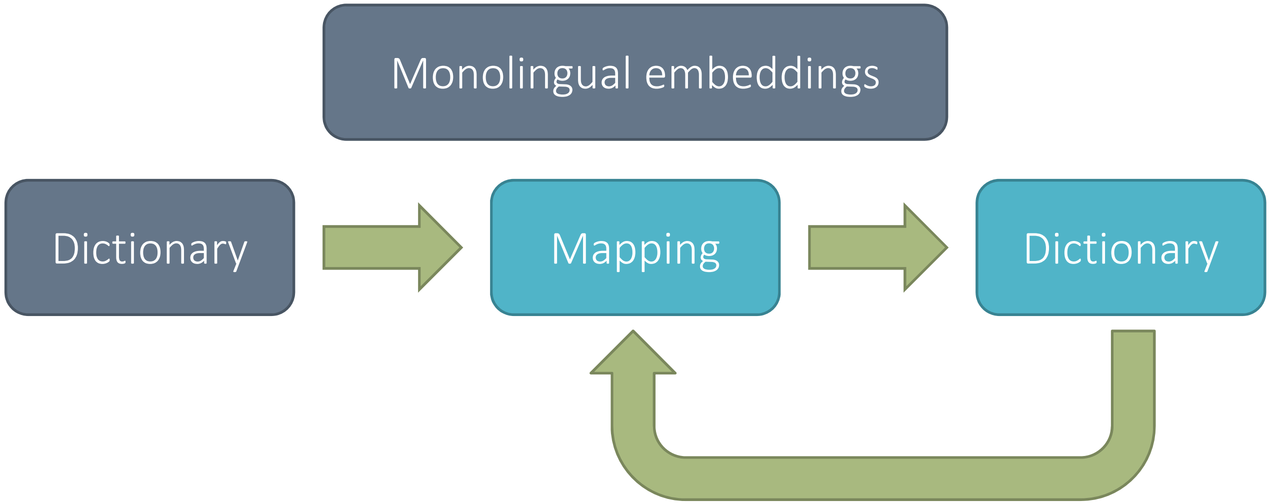
Self-learning



Self-learning



Self-learning



<u>proposed self-learning method</u>

Too good to be true?

Semi-supervised experiments (ACL17)

Semi-supervised experiments (ACL17)

- Given monolingual embeddings
  plus seed bilingual dictionary (*train* dictionary):
  - 25 word pairs
  - Pairs of numerals

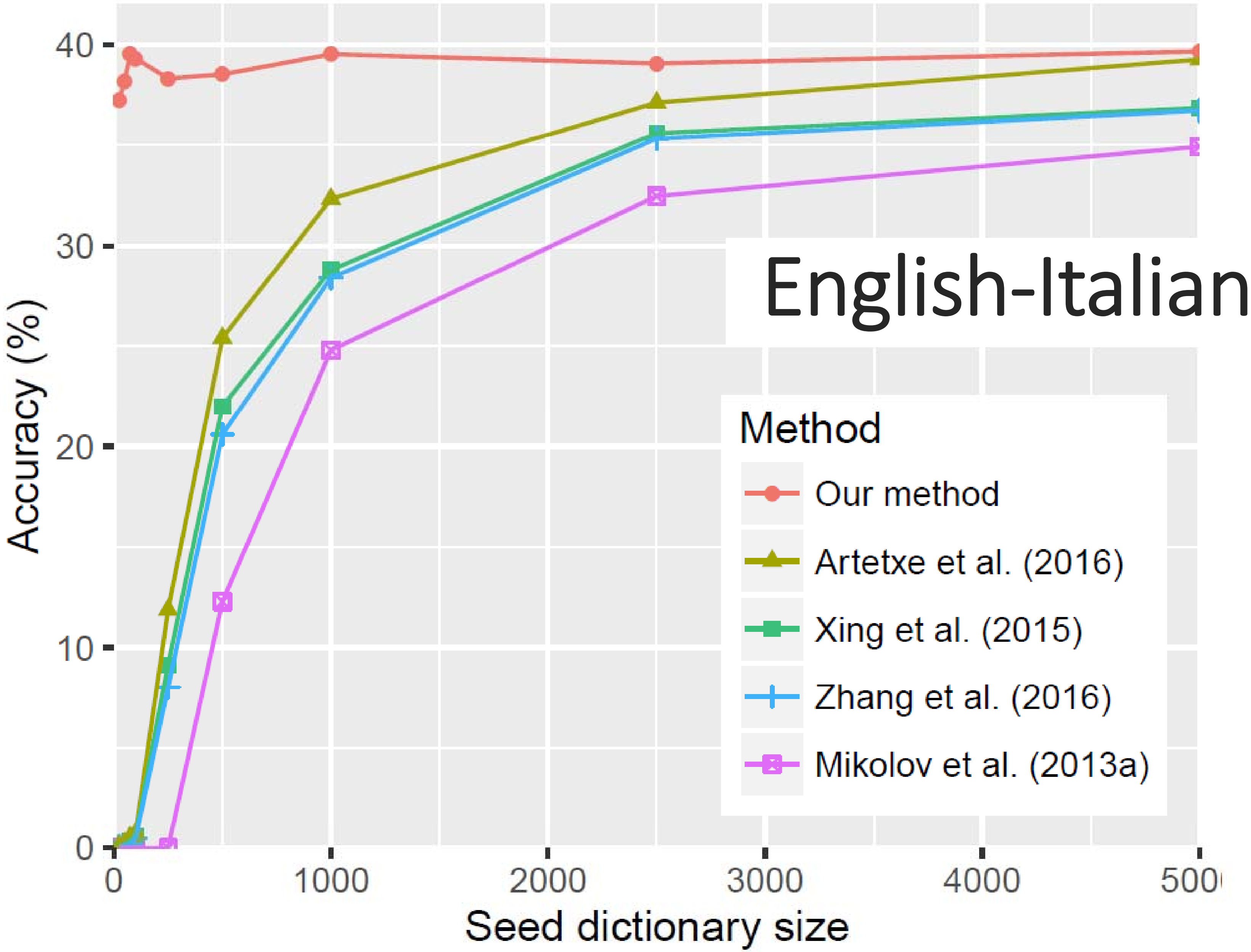
Semi-supervised experiments (ACL17)

- Given monolingual embeddings
  plus seed bilingual dictionary (*train* dictionary):
  - 25 word pairs
  - Pairs of numerals
- Induce bilingual dictionary using self-learning
  for full vocabulary

Semi-supervised experiments (ACL17)

- Given monolingual embeddings plus seed bilingual dictionary (*train* dictionary):
  - 25 word pairs
  - Pairs of numerals
- Induce bilingual dictionary using self-learning for full vocabulary
- Evaluation
  - Compare translations to existing bilingual dictionary (*test* dictionary)
  - Accuracy

Semi-supervised experiments (ACL17)



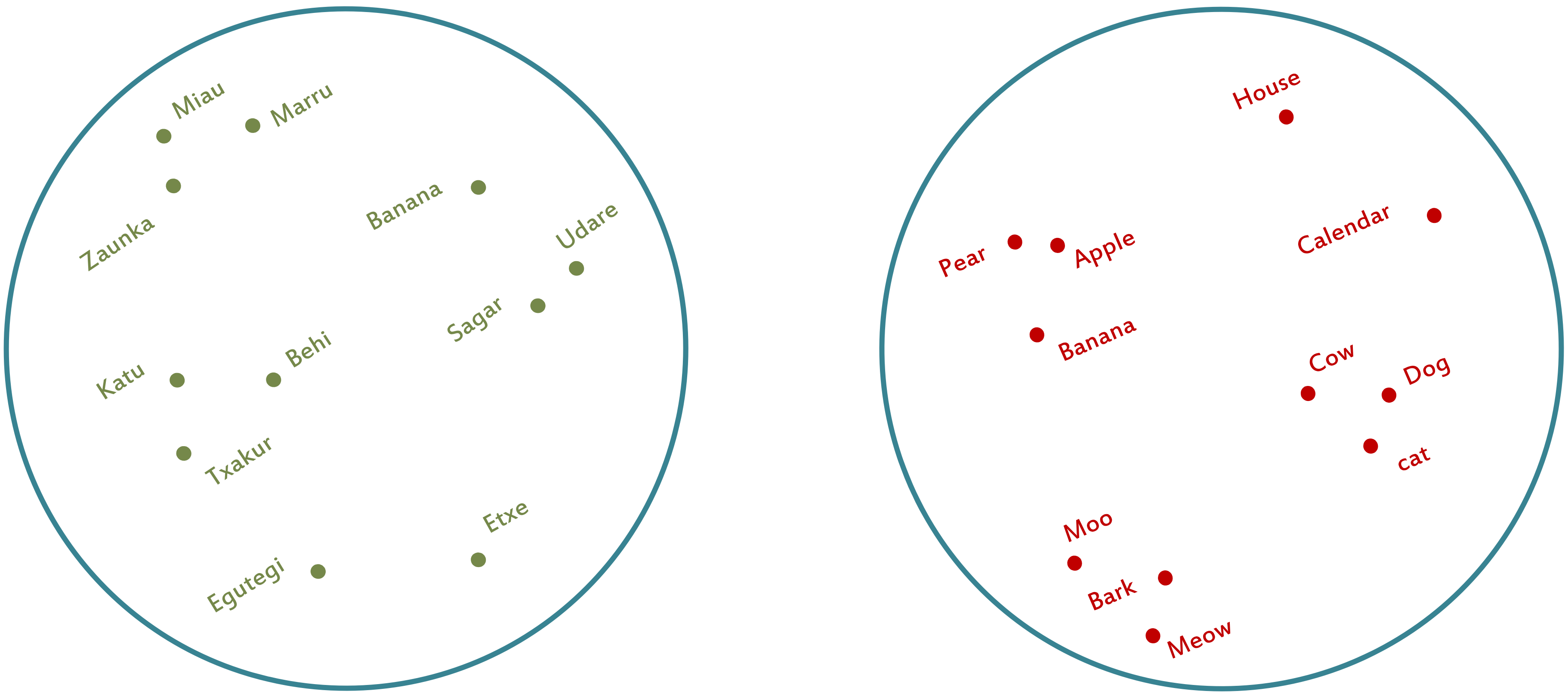
Why does it work?

Why does it work?

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

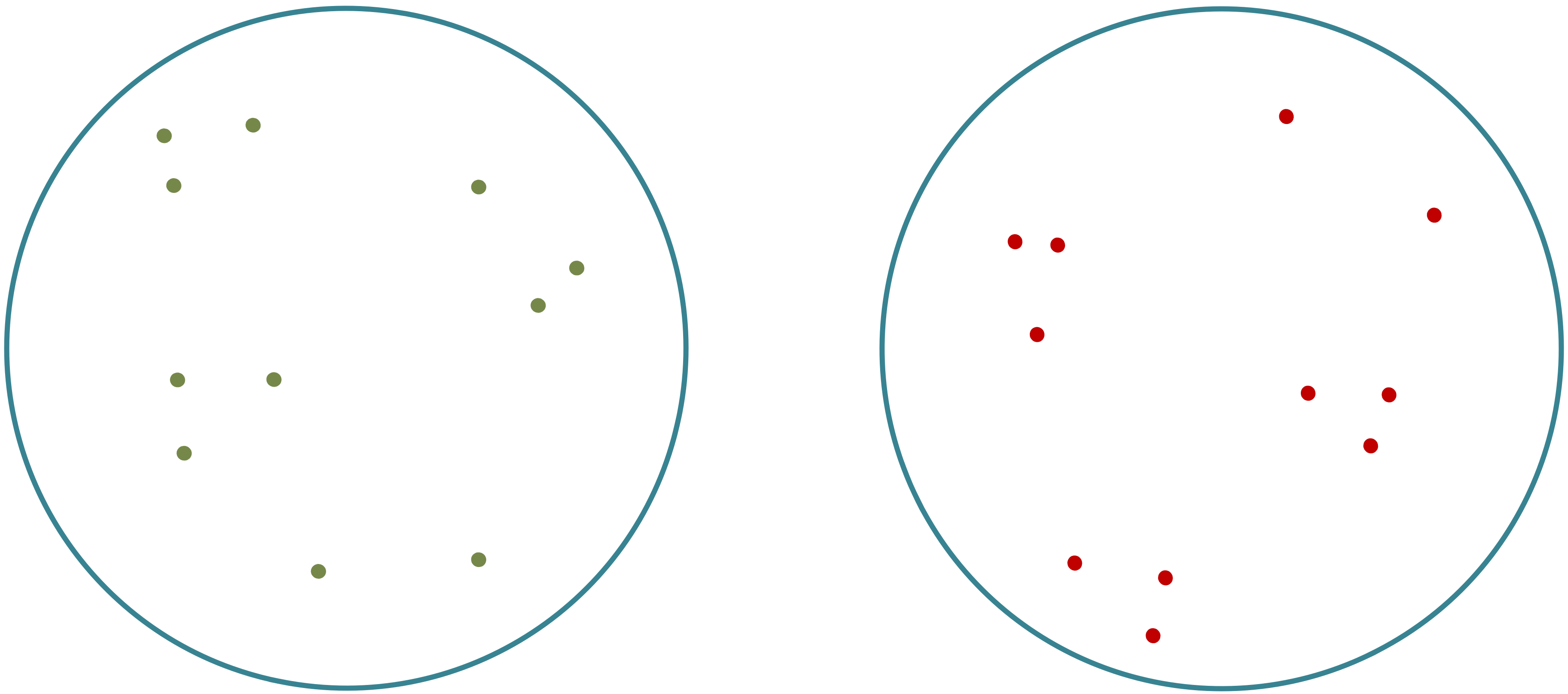
Why does it work?

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



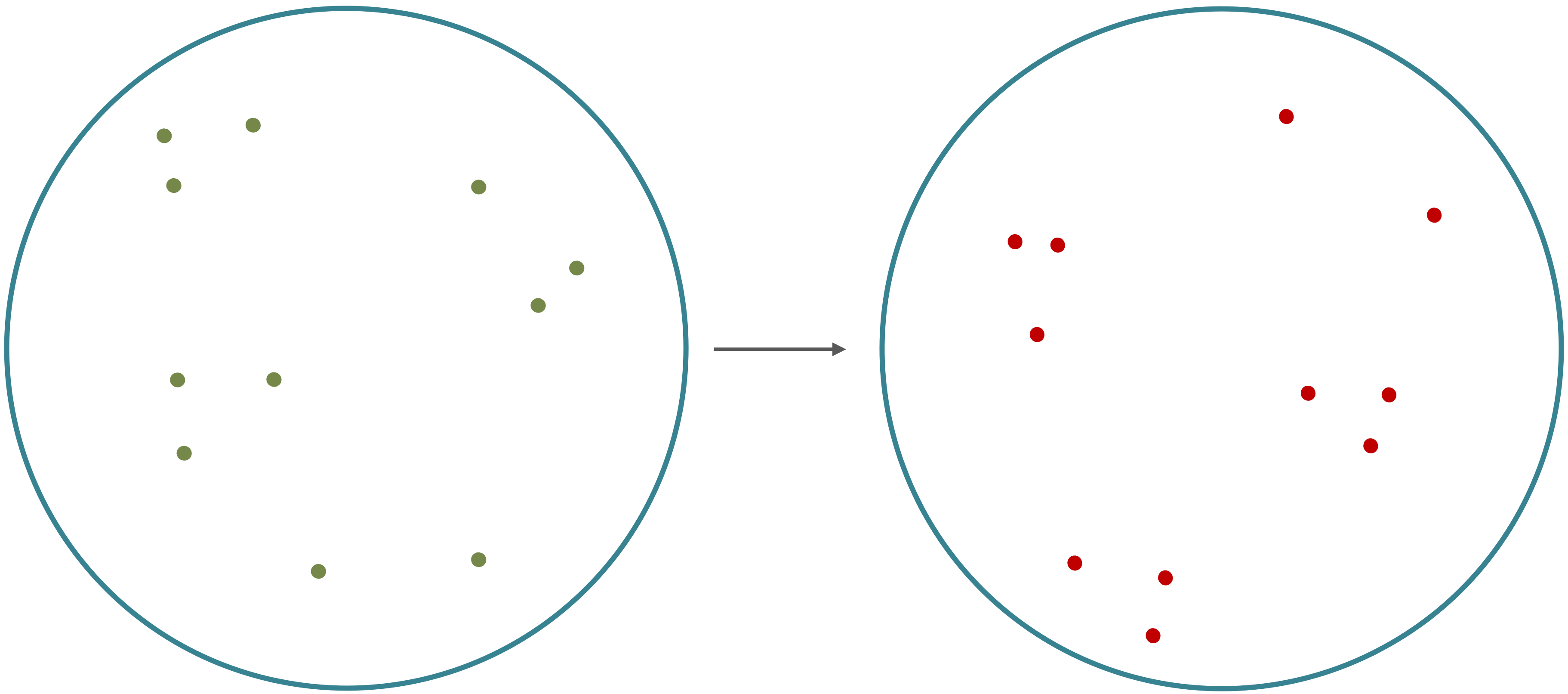
Why does it work?

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



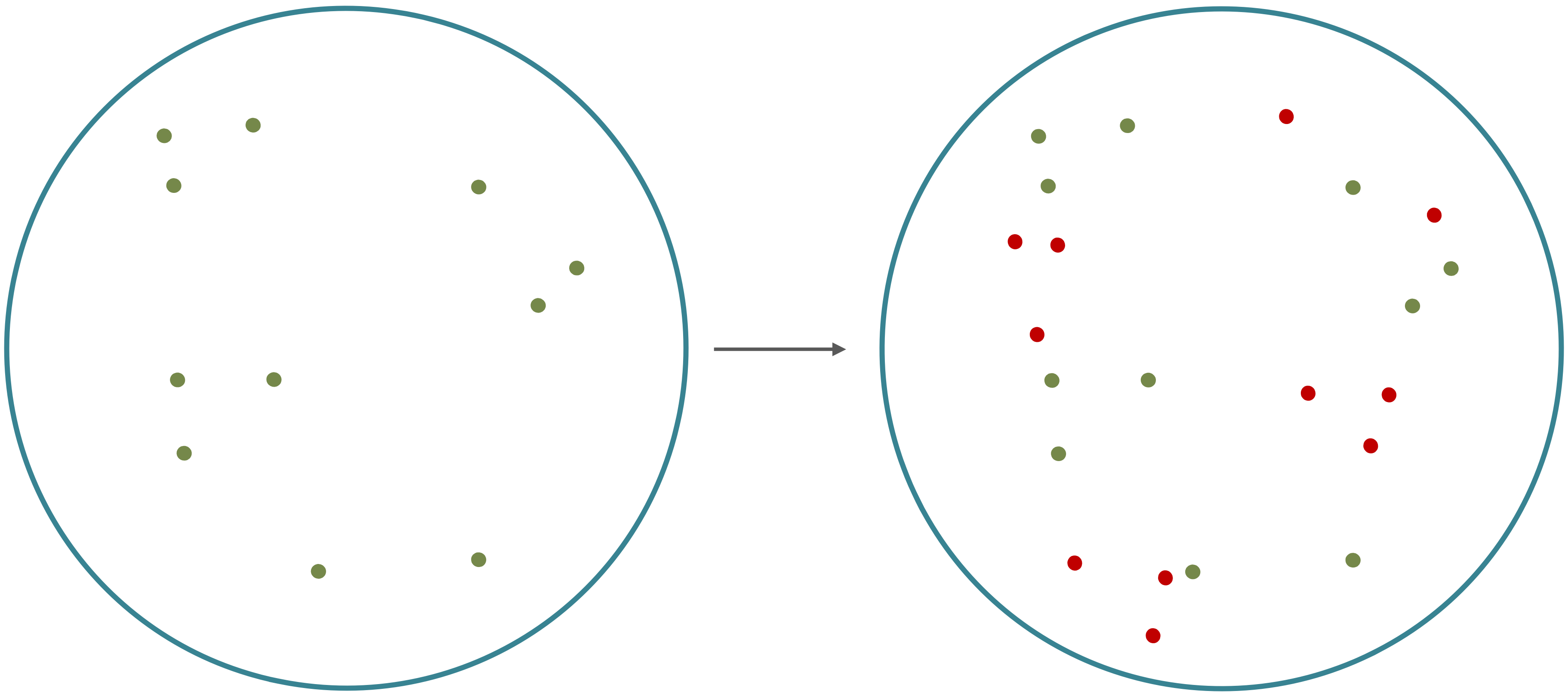
Why does it work?

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



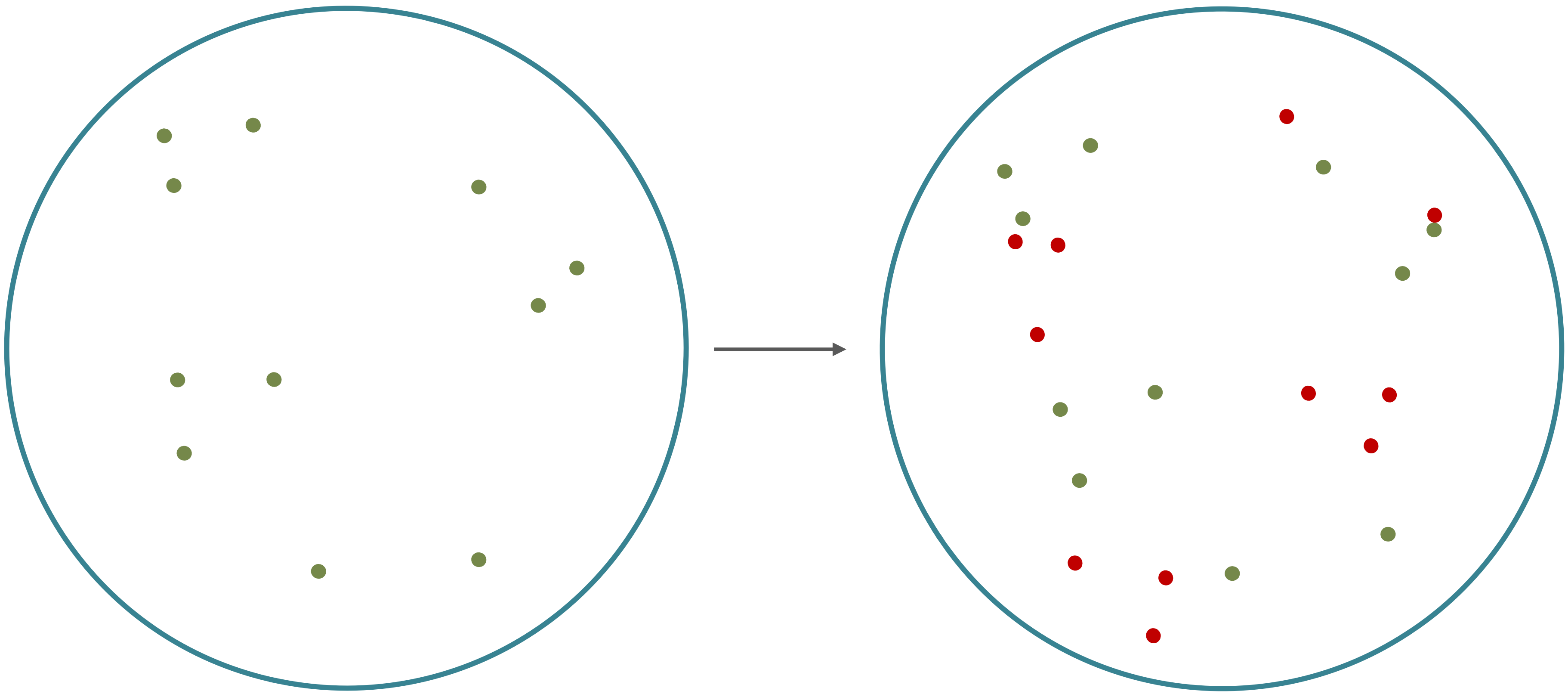
Why does it work?

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



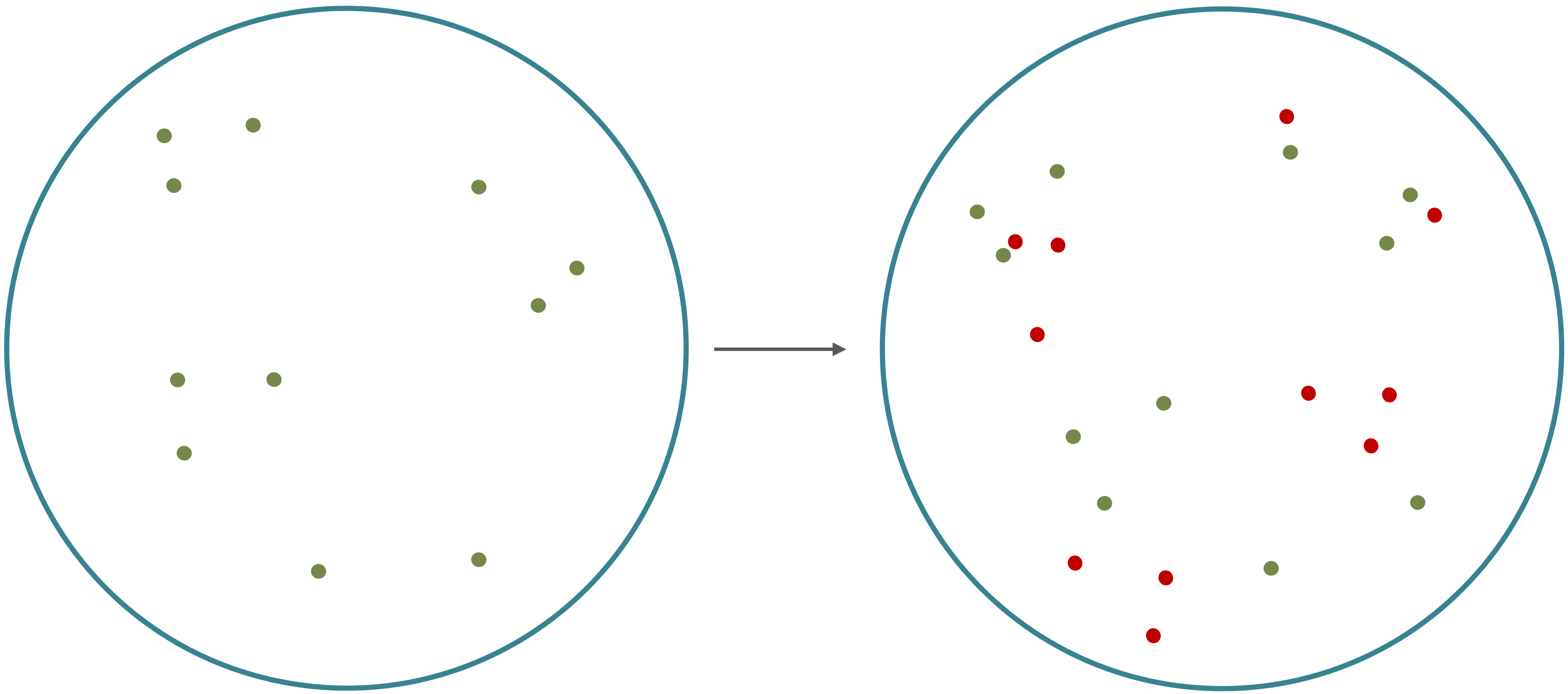
Why does it work?

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



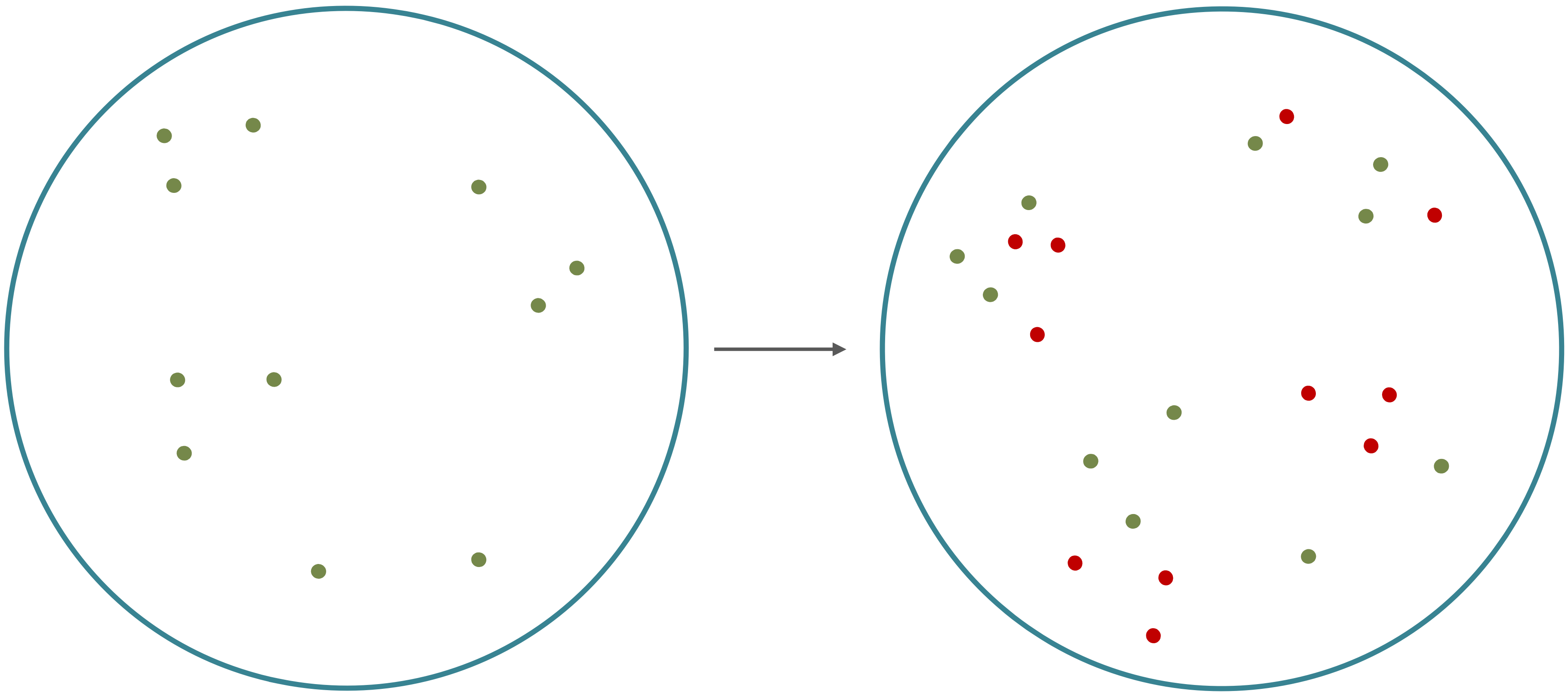
Why does it work?

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



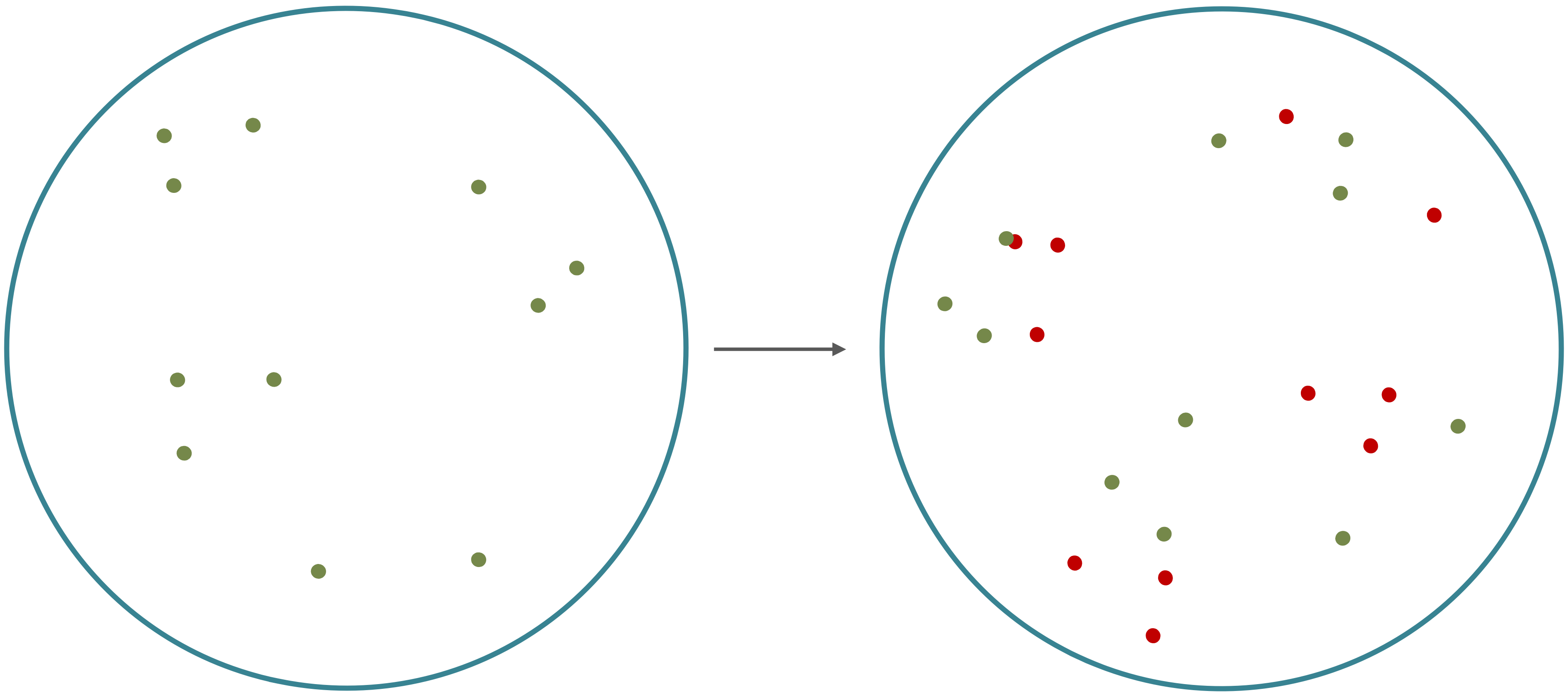
Why does it work?

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



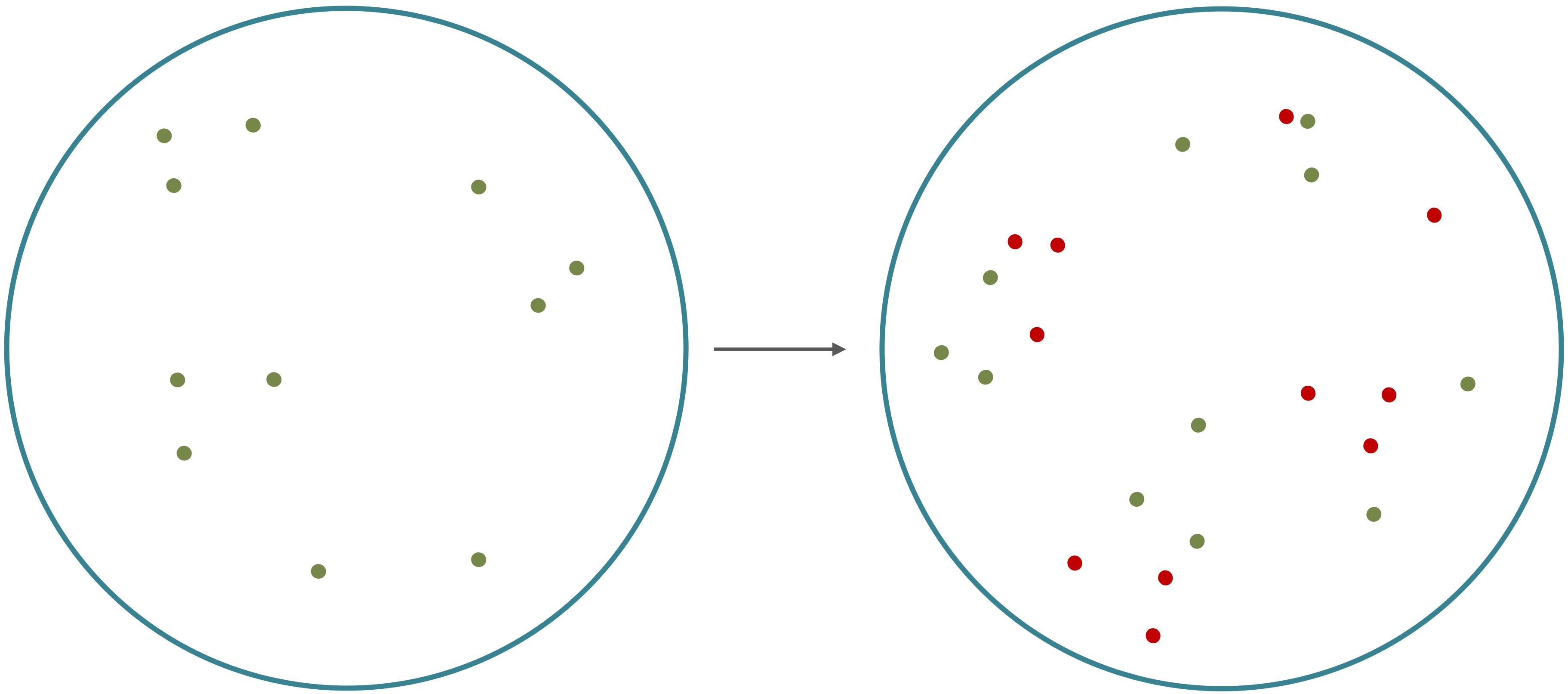
Why does it work?

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



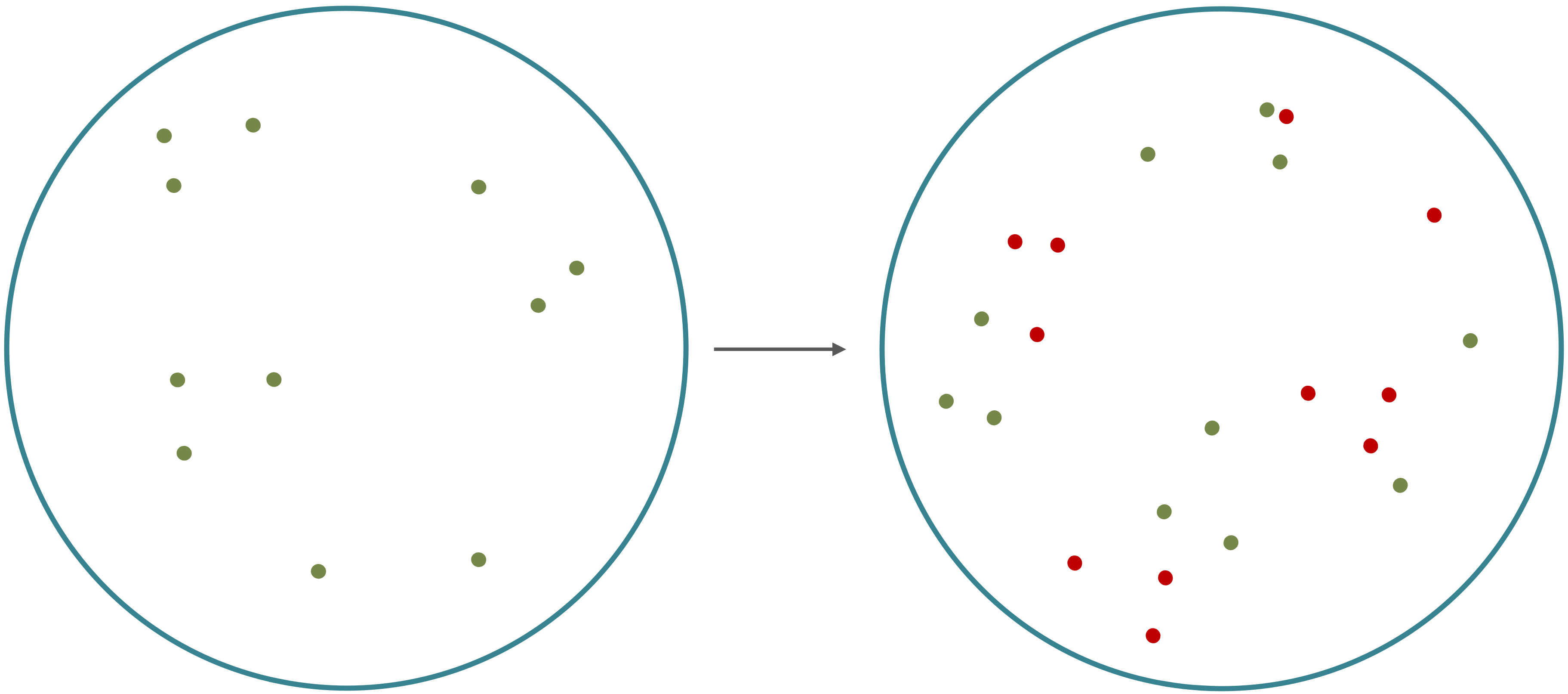
Why does it work?

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



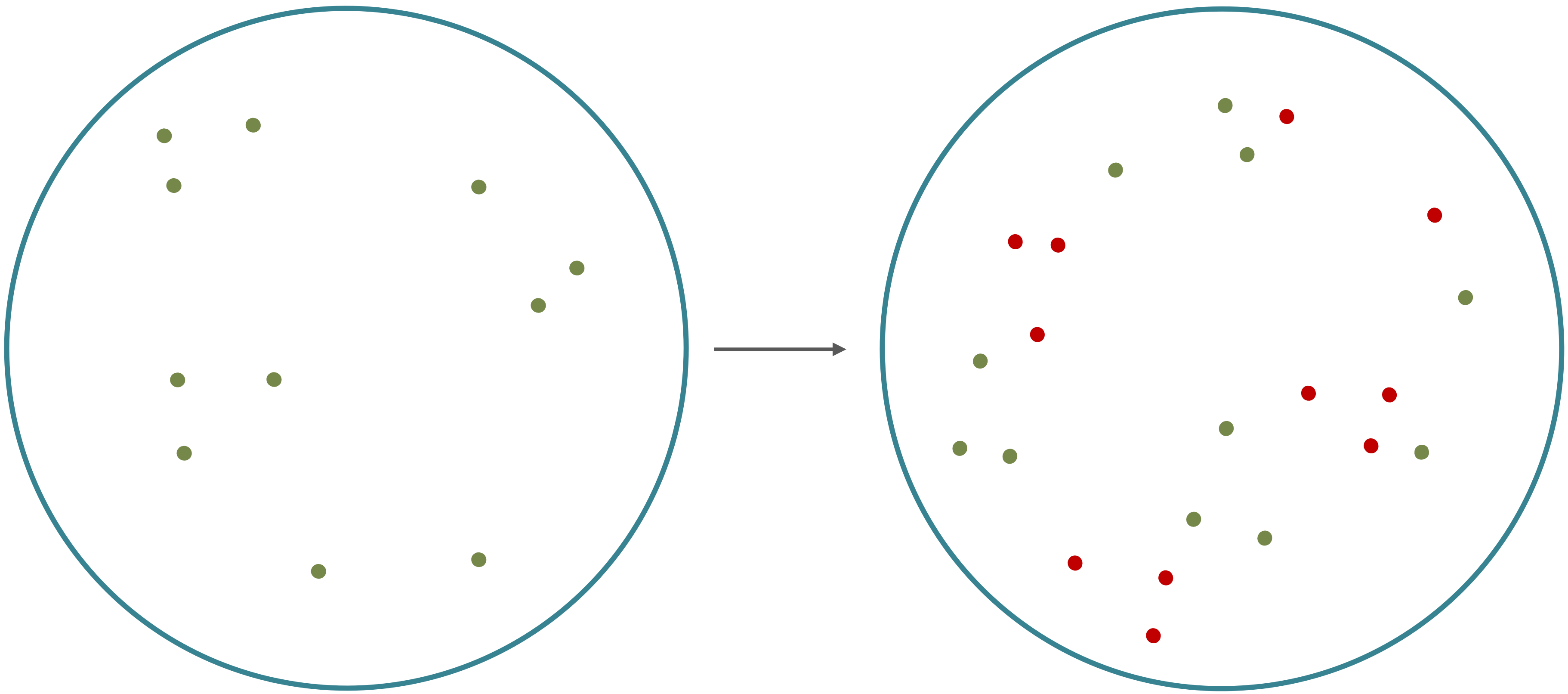
Why does it work?

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



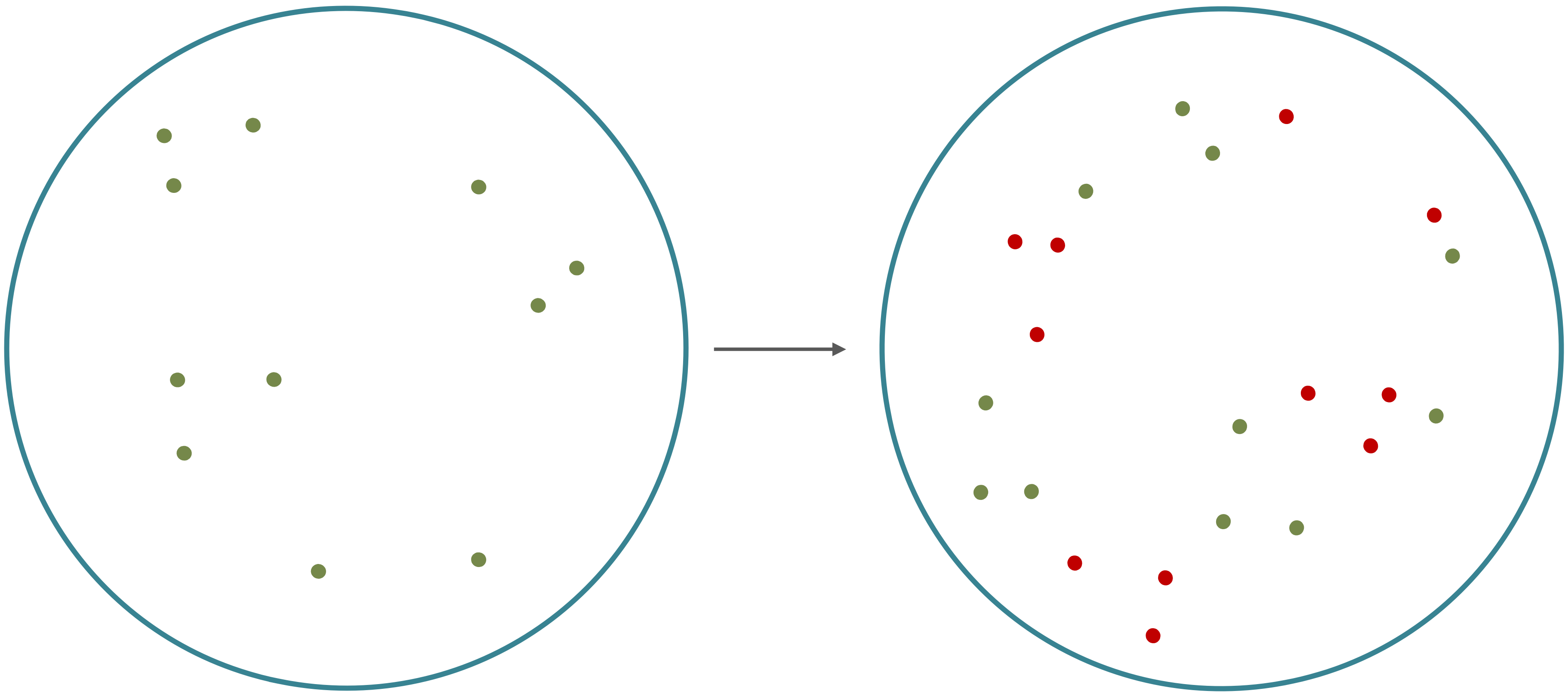
Why does it work?

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



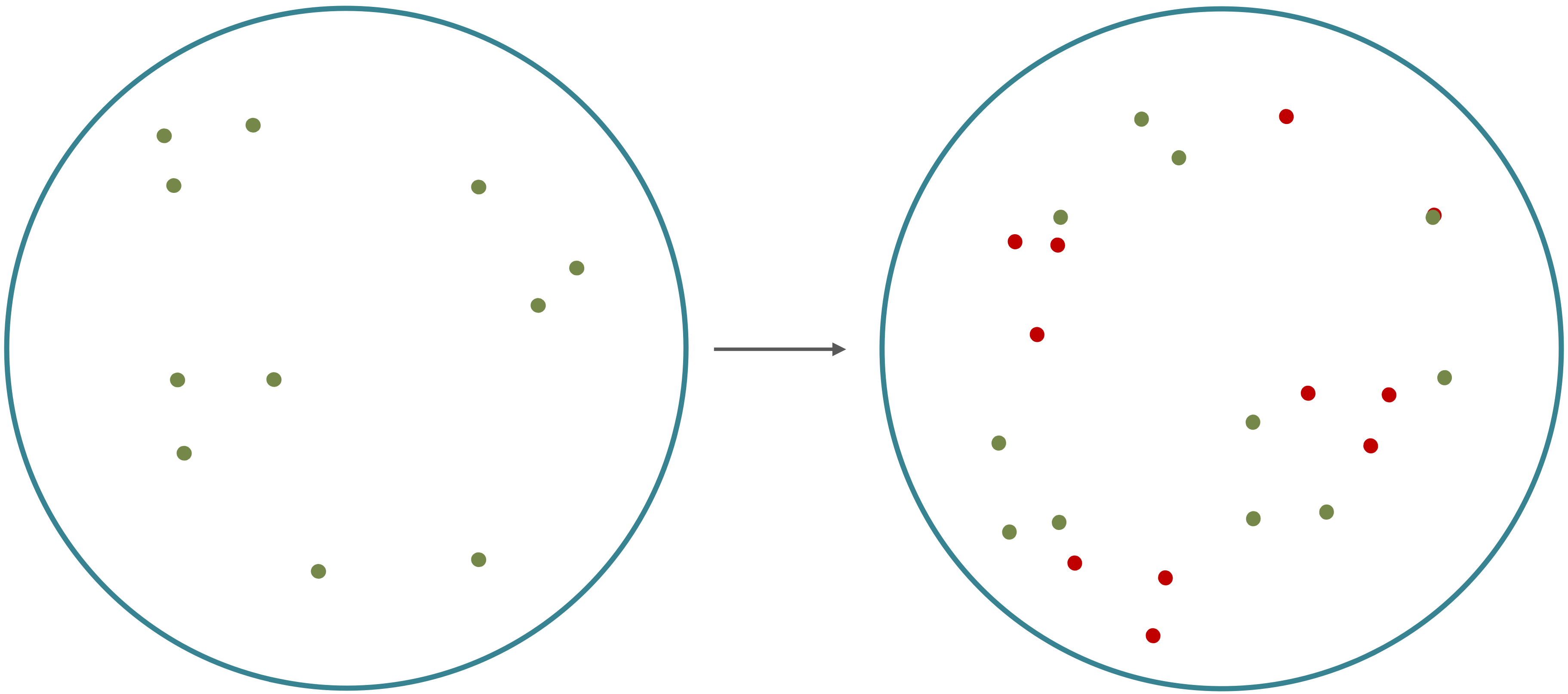
Why does it work?

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



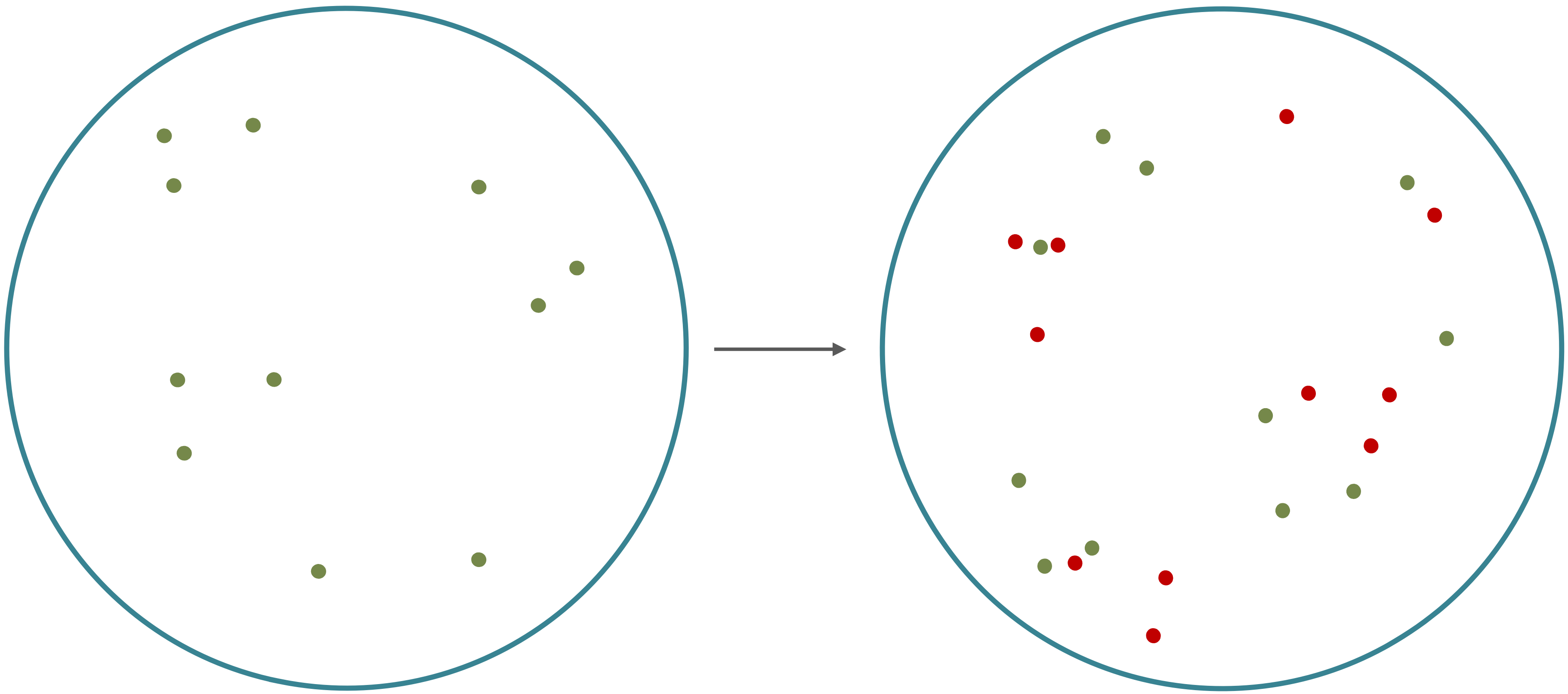
Why does it work?

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



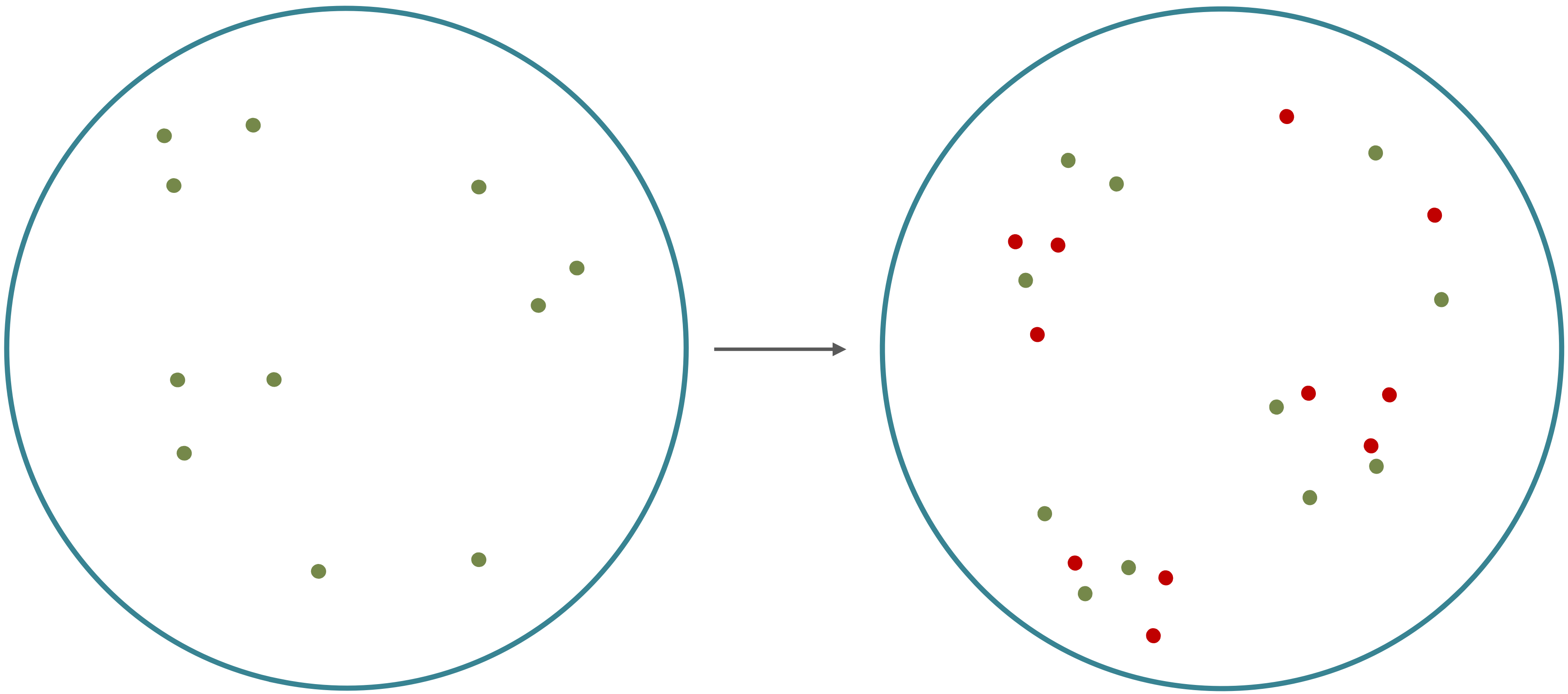
Why does it work?

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



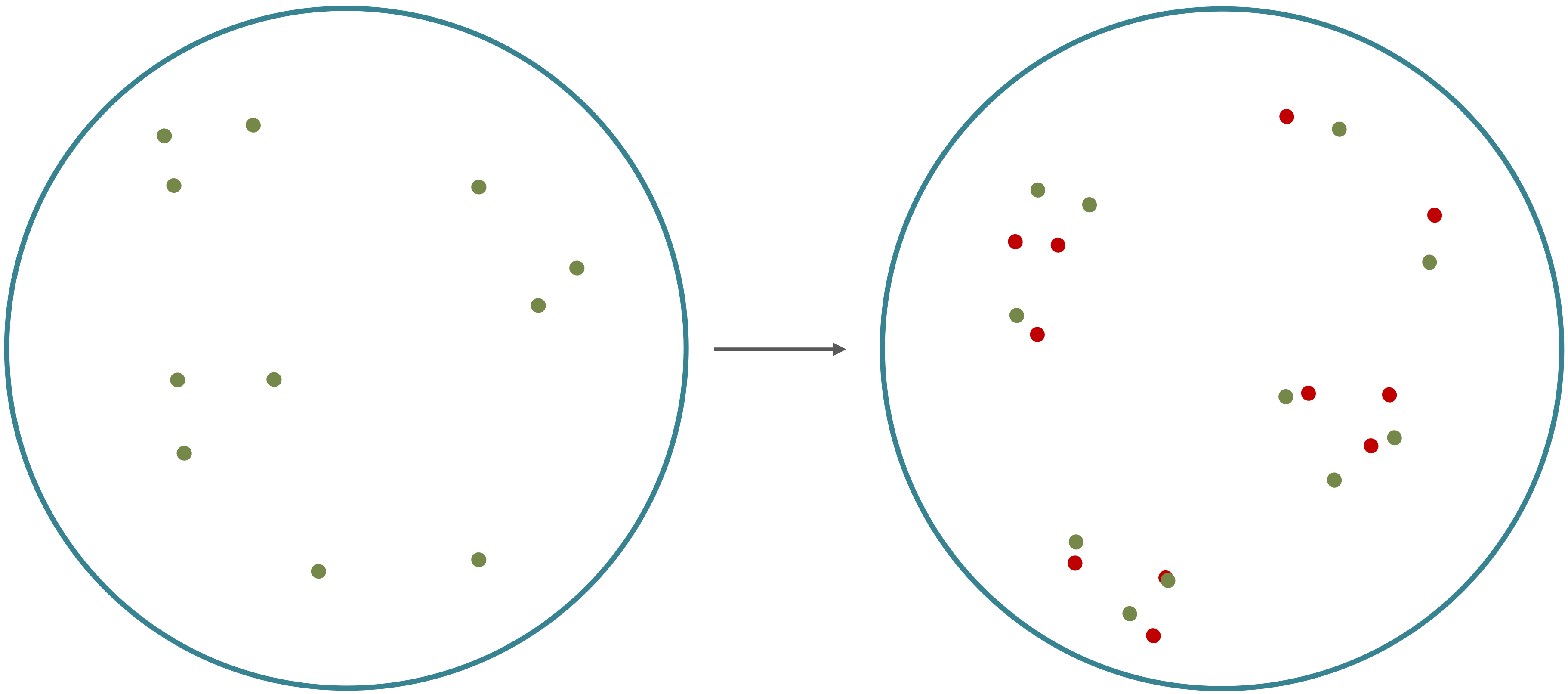
Why does it work?

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



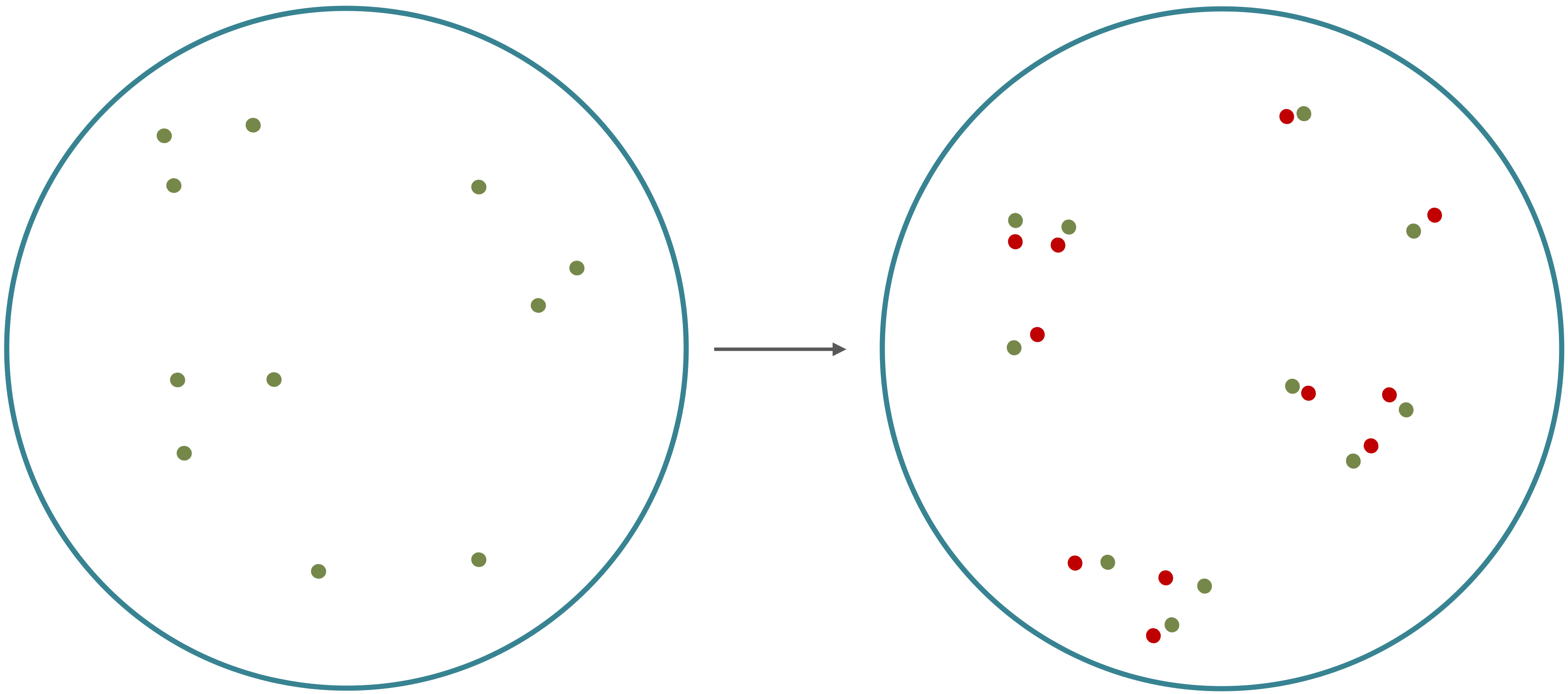
Why does it work?

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



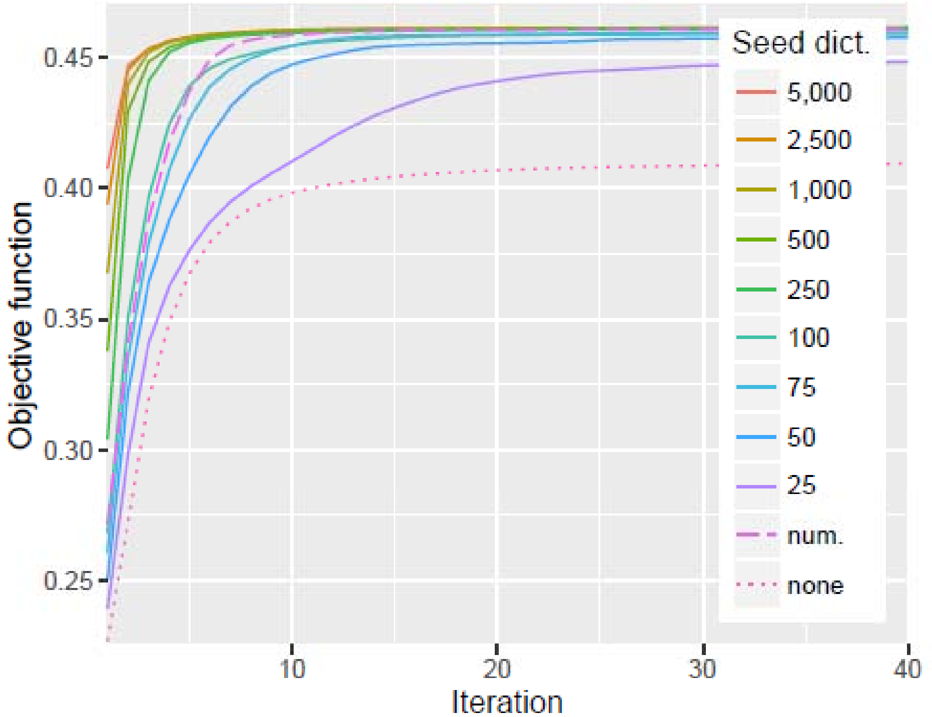
Why does it work?

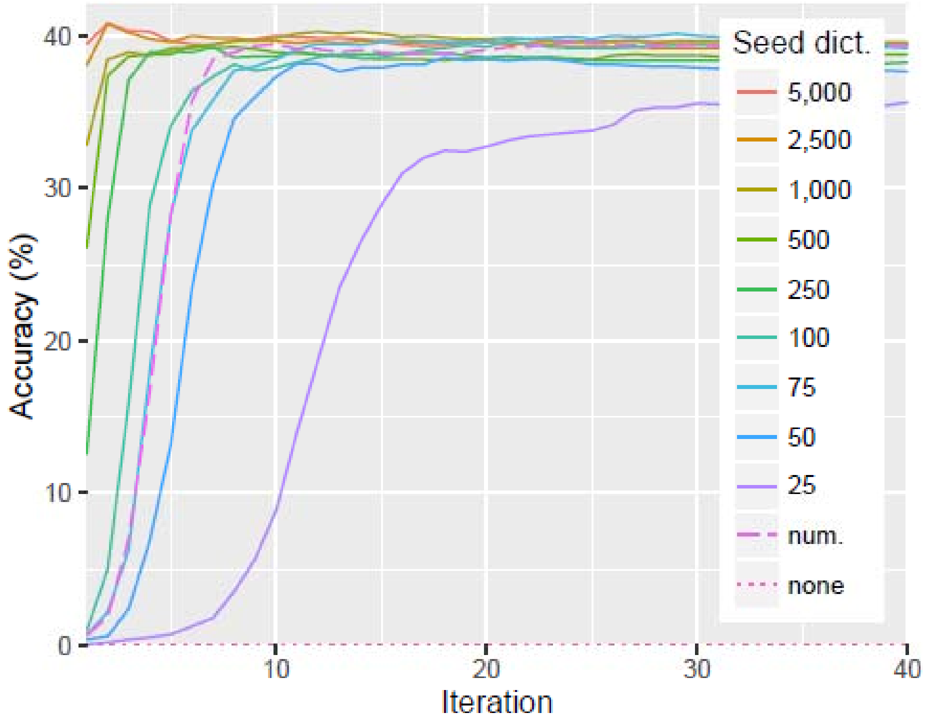
Implicit objective: $W^* = \arg\max_W \sum_i \max_j (X_{i*}W) \cdot Z_{j*} \quad \text{s.t.} \quad WW^T = W^TW = I$

Independent from seed dictionary!

Why does it work?

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





Why does it work?

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

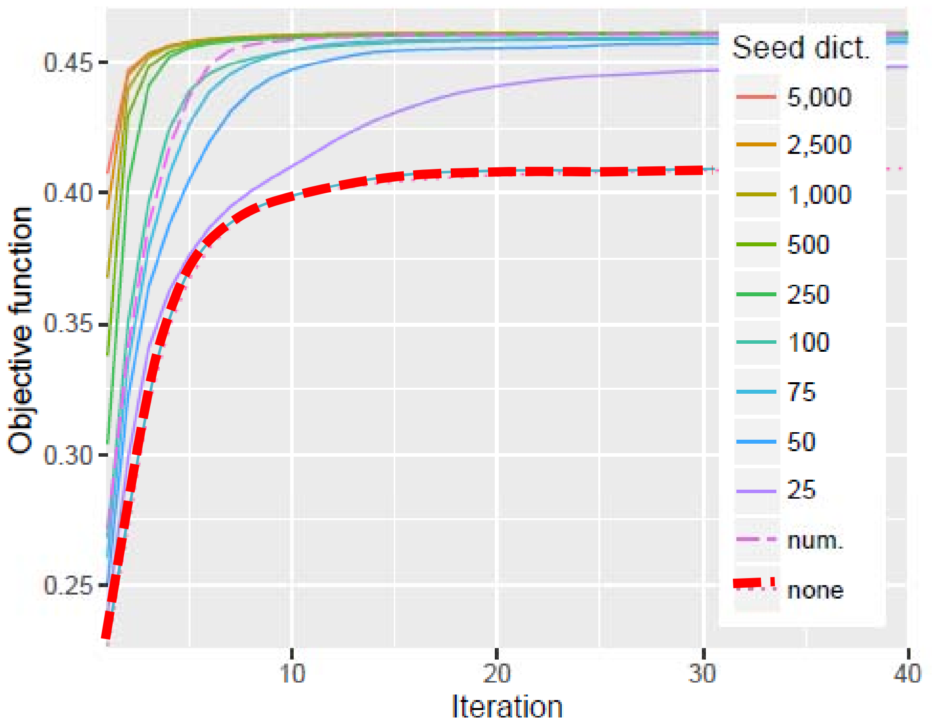
Independent from seed dictionary!

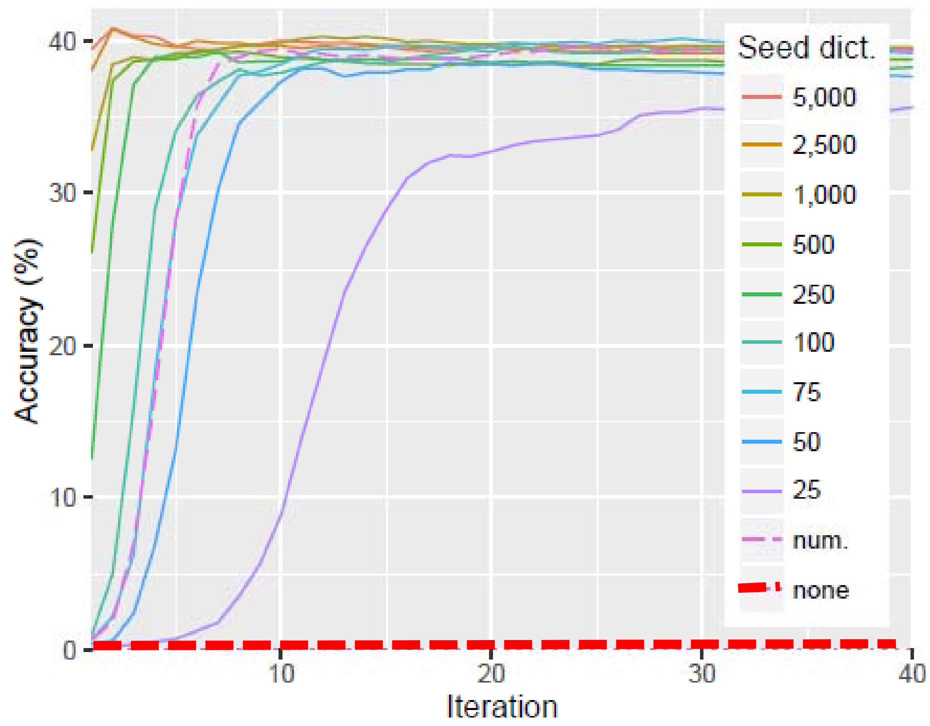
So why do we need a seed dictionary?

Avoid poor local optima!

Why does it work?

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





Next steps

Is there a way we can avoid the seed dictionary?

Would an initial noisy initialization suffice?

Unsupervised experiments (ACL18)

Unsupervised experiments (ACL18)

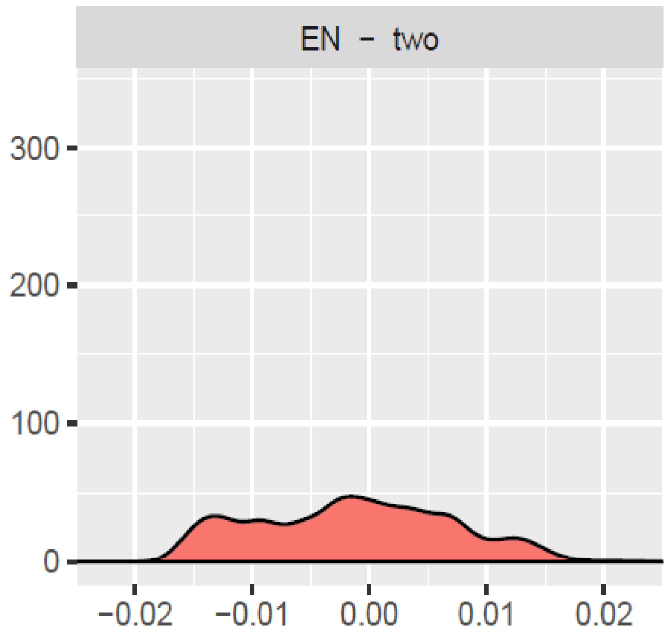
Initial dictionary:

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:

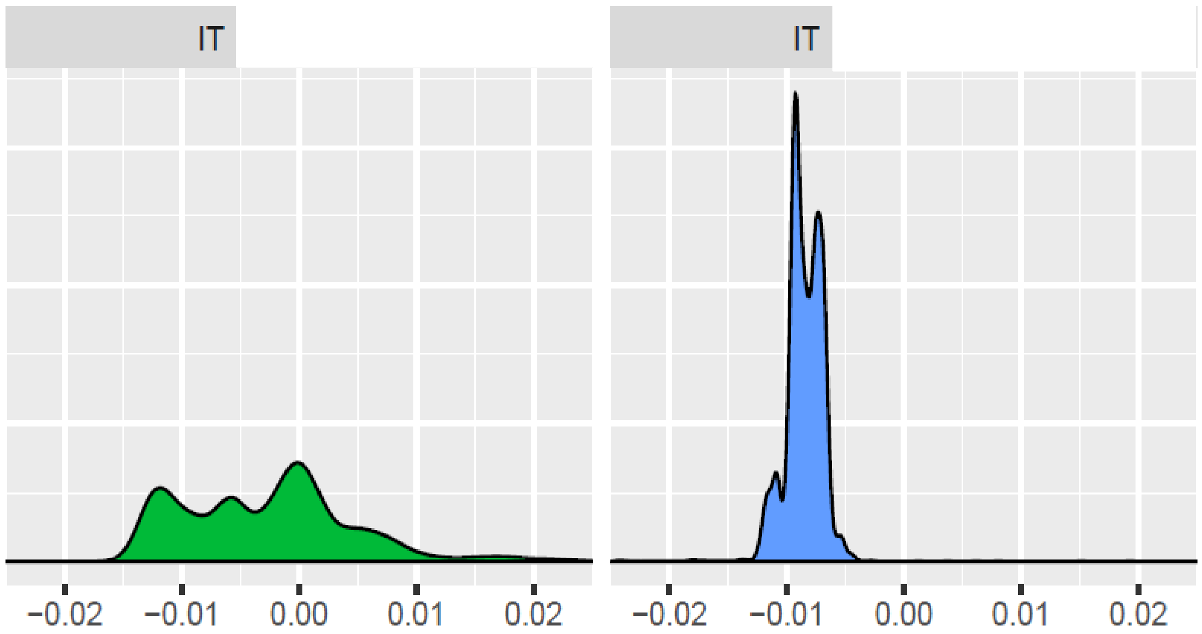
1. Compute intra-language similarity
2. Words which are translations of each other
   would have analoguous similarity histograms (isometry hyp.)



Unsupervised experiments (ACL18)

Initial dictionary:

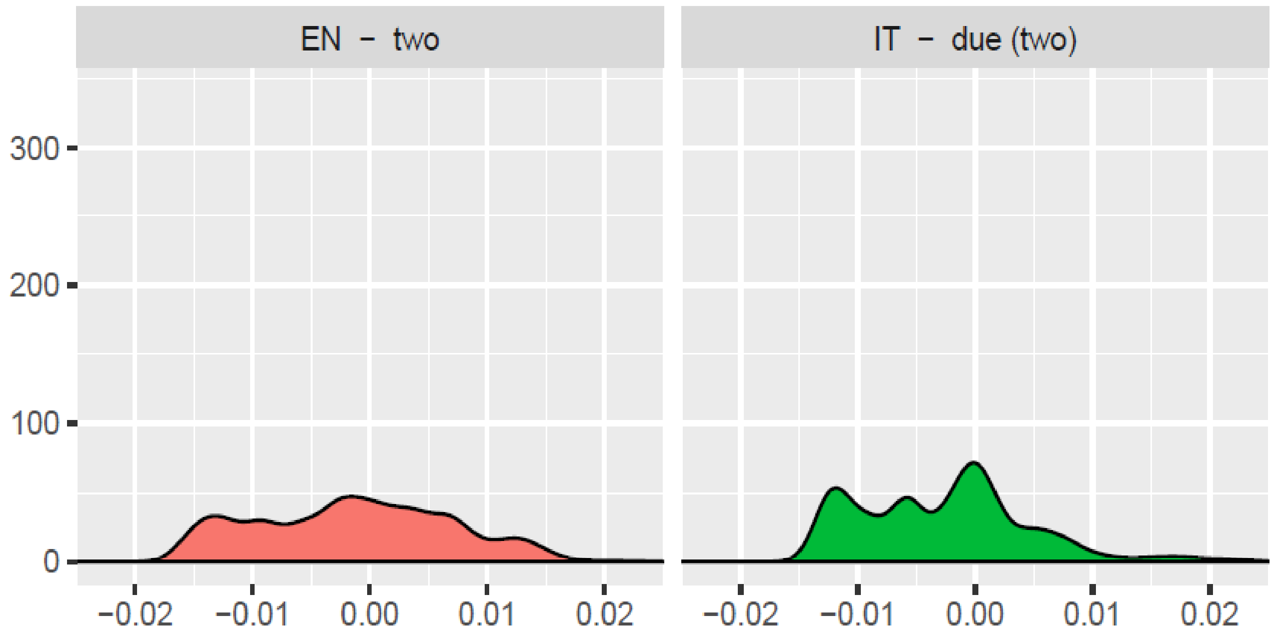
1. Compute intra-language similarity
2. Words which are translations of each other
   would have analoguous similarity histograms (isometry hyp.)



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

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

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ *Monolingual embeddings (CBOW + negative sampling)*

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

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  ⇒ *Seed dictionary: 5,000 word pairs* ***/ 25 word pairs / none***

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.					
25 dict.					
None					

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.					
25 dict.					
None					

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	39.27	41.87*	30.62*	31.40*
	Smith et al. (2017)	43.1	43.33†	29.42†	35.13†
	Our method (AAAI18)	**45.27**	**44.13**	**32.94**	**36.60**
25 dict.					

None

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	**39.27**	**41.87***	**30.62***	31.40*
25 dict.	Our method (ACL17)				
None					

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	**39.27**	**41.87***	**30.62***	31.40*
25 dict.	Our method (ACL17)	37.27	39.60	28.16	-
None					

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.					
25 dict.					
None	Zhang et al. (2017)				
	Conneau et al. (2018)				

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  ⇒ *Test dictionary: 1,500 word pairs*

Previous work convergence problems!
Also observed by Sogard et al. (2018)

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.					
25 dict.					
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

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.					
25 dict.					
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
	Our method (ACL18)	**48.13**	**48.19**	32.63	**37.33**

Unsupervised experiments (ACL18)

- Dataset by Dinu et al. (2015) extended German, Finnish, Spanish
  - ⇒ *Monolingual embeddings (CBOW + negative sampling)*
  - ⇒ *Seed dictionary: 5,000 word pairs / 25 word pairs / none*
  - ⇒ *Test dictionary: 1,500 word pairs*

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Artetxe et al. (2016)	39.27	41.87*	30.62*	31.40*
	Smith et al. (2017)	43.1	43.33†	29.42†	35.13†
	Our method (AAAI18)	45.27	44.13	**32.94**	36.60
25 dict.	Our method (ACL17)	37.27	39.60	28.16	-
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
	Our method (ACL18)	**48.13**	**48.19**	32.63	**37.33**

Conclusions

Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Unsupervised matches results of supervised methods!
- Implicit optimization objective independent from seed dictionary

Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Unsupervised matches results of 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%

Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Unsupervised matches results of 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%
- Full reproducibility (including datasets):
  `https://github.com/artetxem/vecmap`

Conclusions

- Simple self-learning method to train bilingual embedding mappings
- Unsupervised matches results of 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%
- **Full reproducibility (including datasets):**
  `https://github.com/artetxem/vecmap`

- Shows that languages share “semantic” structure to a large degree

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

- Bilingual embedding mappings
  - *Introduction to vector space models (embeddings)*
  - *Bilingual embedding mappings (AAAI18)*
  - *Reduced supervision*
    - Self-learning, semi-supervised (ACL17)
    - Self-learning, fully unsupervised (ACL18)
  - *Conclusions*
- Unsupervised neural machine translation
  - *Introduction to NMT*
  - *From bilingual embeddings to uNMT (ICLR18)*
  - *Unsupervised statistical MT (EMNLP18)*
  - *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 Recurrent Neural Nets

  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 Recurrent Neural Nets

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

- Train a **decoder** based on Recurrent Neural Nets

  - 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 Recurrent Neural Nets

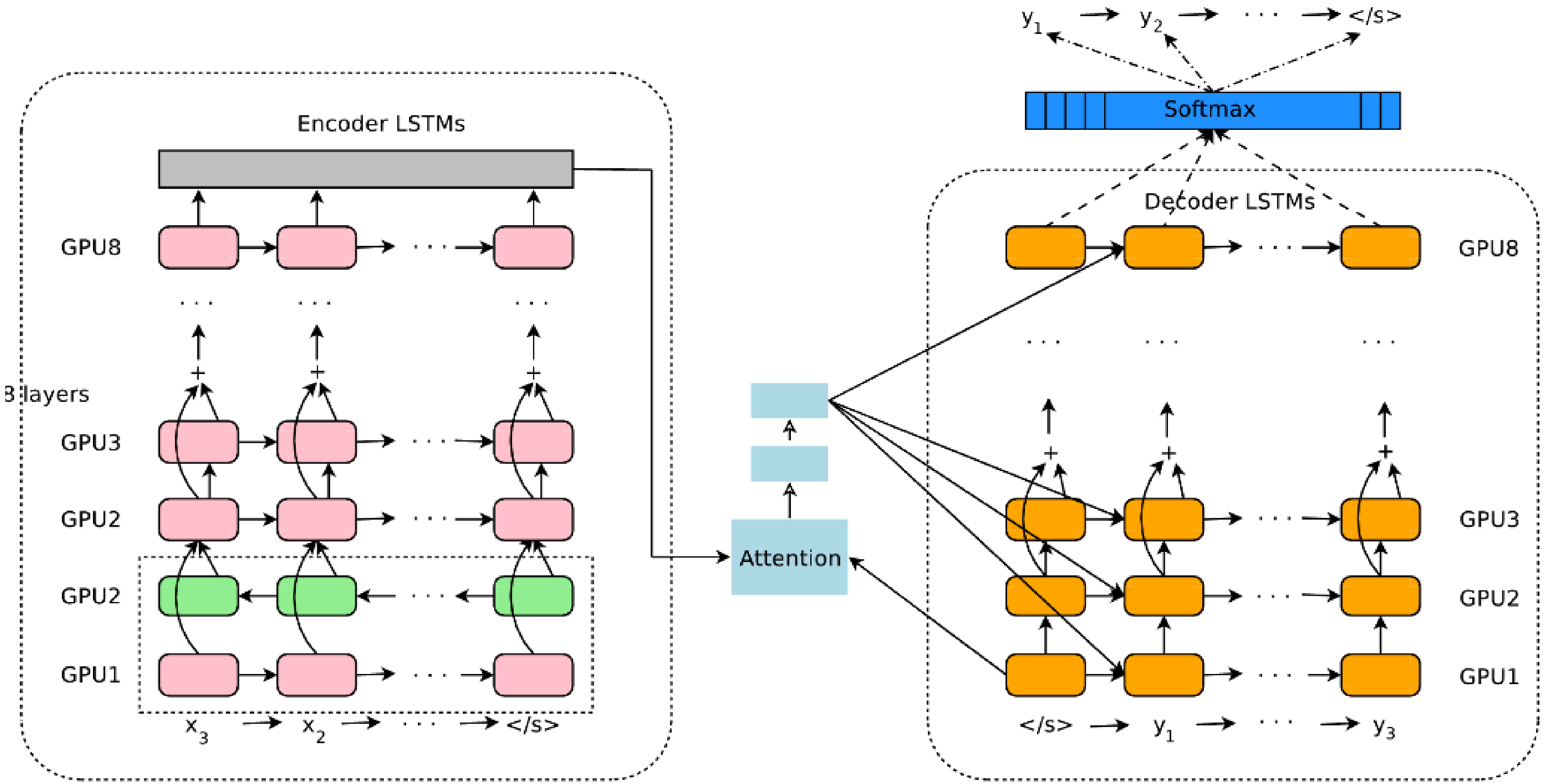
  return all hidden states, encoding input $x_1 \ldots x_n$

- Train a **decoder** based on Recurrent Neural Nets

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

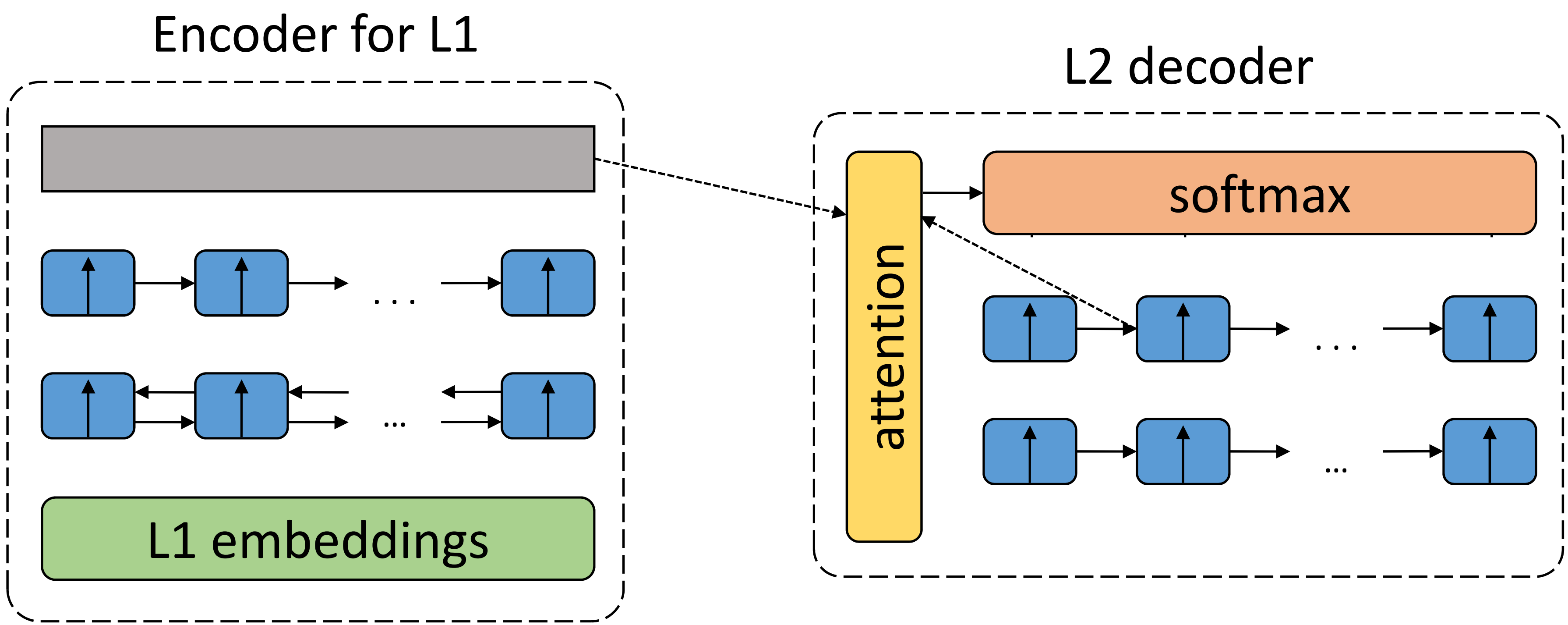
End-to-end training

Introduction to (supervised) NMT



Source: Wu et al. 2016 (~ 30 authors – Also known as Google NMT)

Introduction to (supervised) NMT



Unsupervised neural machine translation

- Now that we **can represent words in two languages in the same embeddings space** without bilingual dictionaries…

  … what can we do?

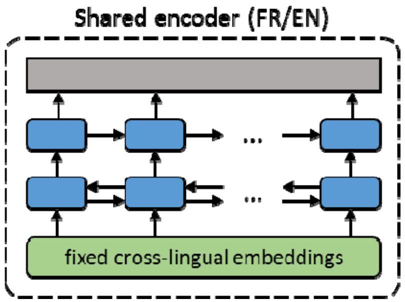
Unsupervised neural machine translation

- Now that we **can represent words in two languages in the same embeddings space** without bilingual dictionaries…

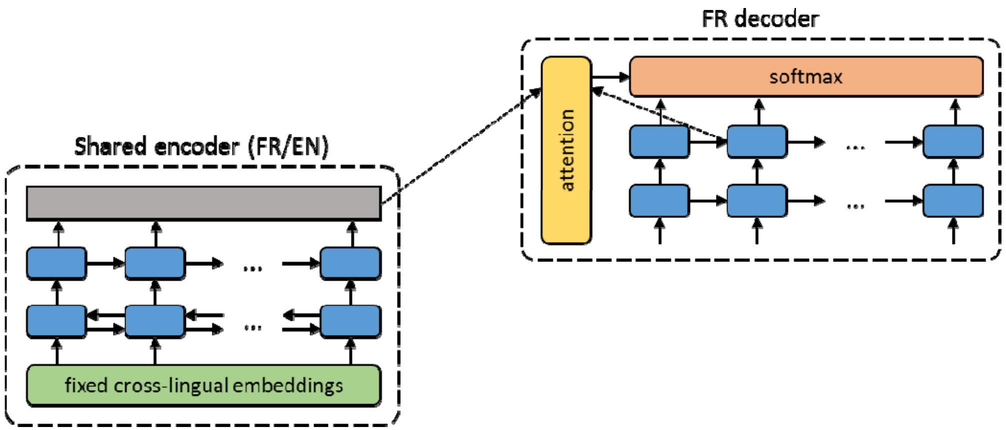
  … what can we do?

- 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)
  - Fixed embeddings

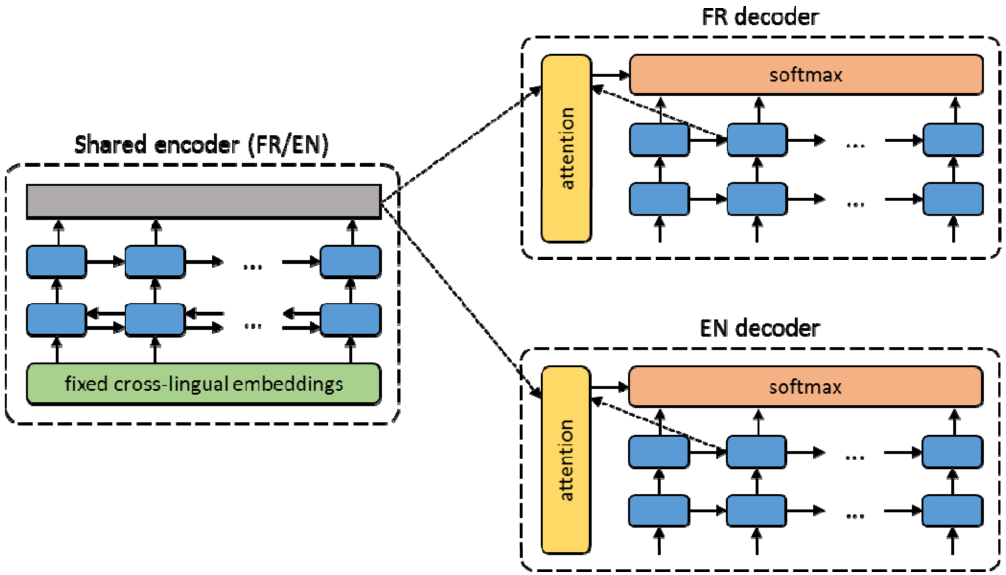
Unsupervised neural machine translation



Unsupervised neural machine translation



Unsupervised neural machine translation



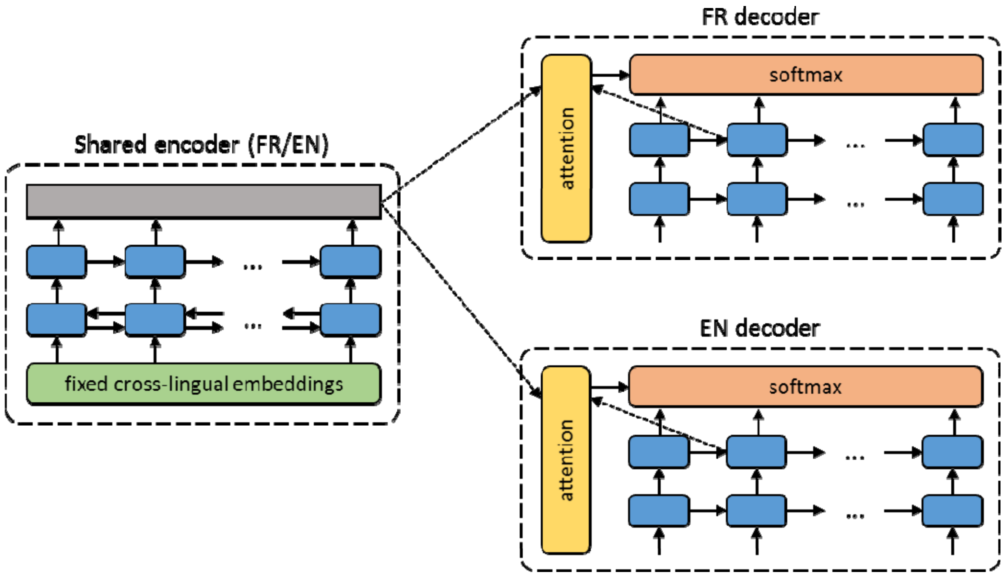
Unsupervised neural machine translation

- 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)
  - Fixed embeddings

- We change the **training regime, mixing mini-batches**:
  - Denoising autoencoder: noisy input in L1, output in the same language (E+D1)
  - Denoising autoencoder: noisy input in L2, output in the same language (E+D2)
  - Backtranslation: input in L1, translate E+D2, translate E+D1, output in L1
  - Backtranslation: input in L2, translate E+D1, translate E+D2, output in L2

Unsupervised neural machine translation

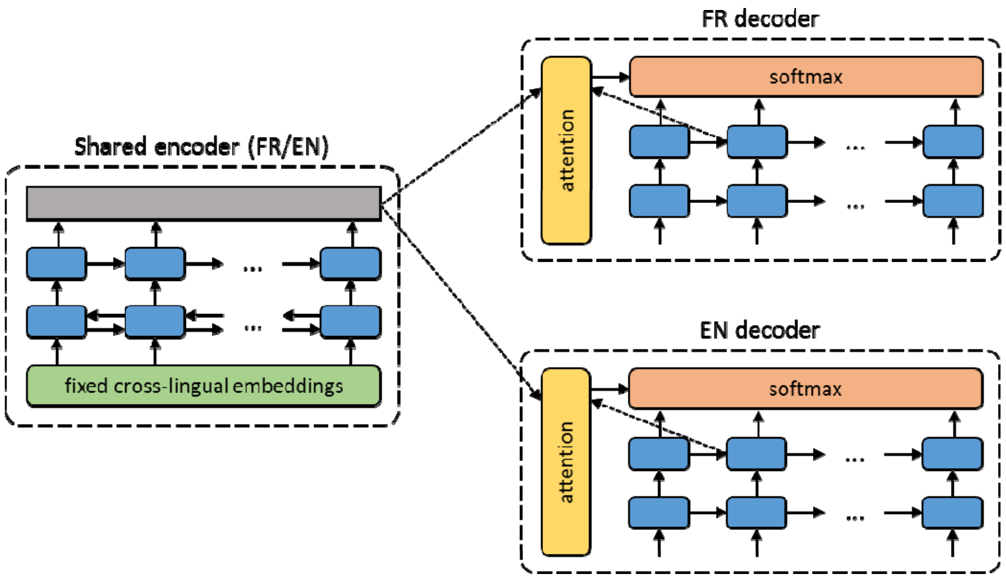
Training



Unsupervised neural machine translation

Training

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

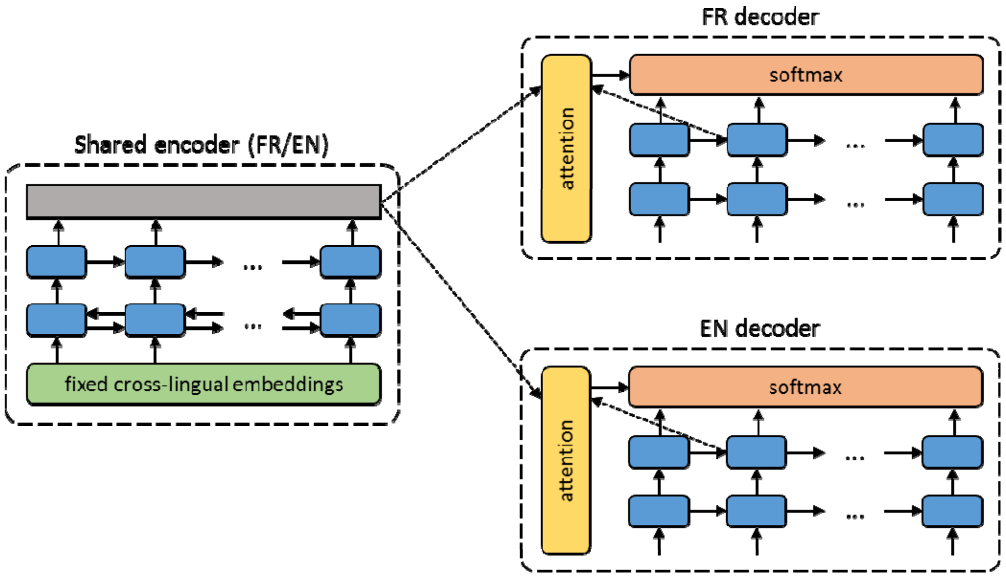


Unsupervised neural machine translation

Training

- Supervised

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

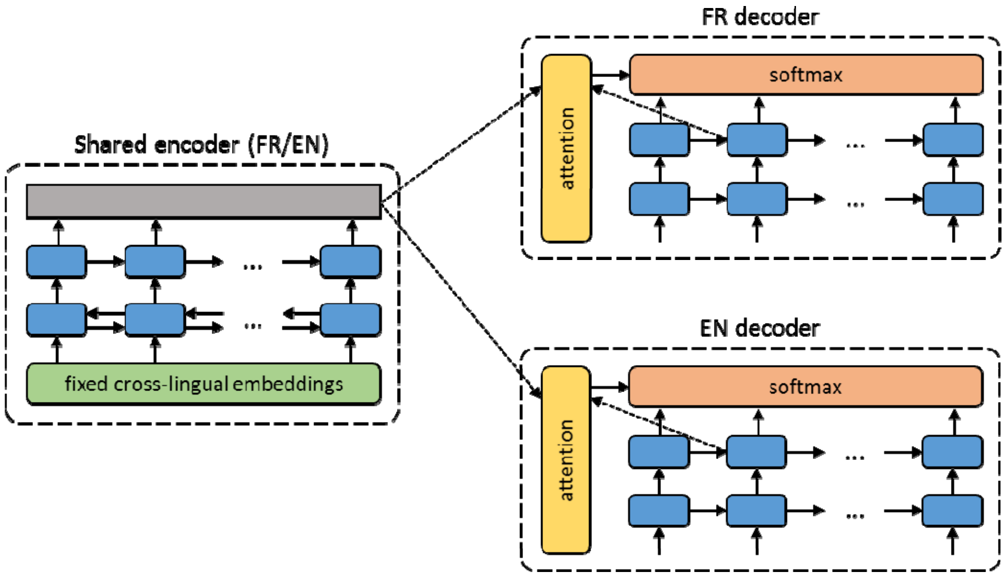


Unsupervised neural machine translation

Training

- Supervised

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



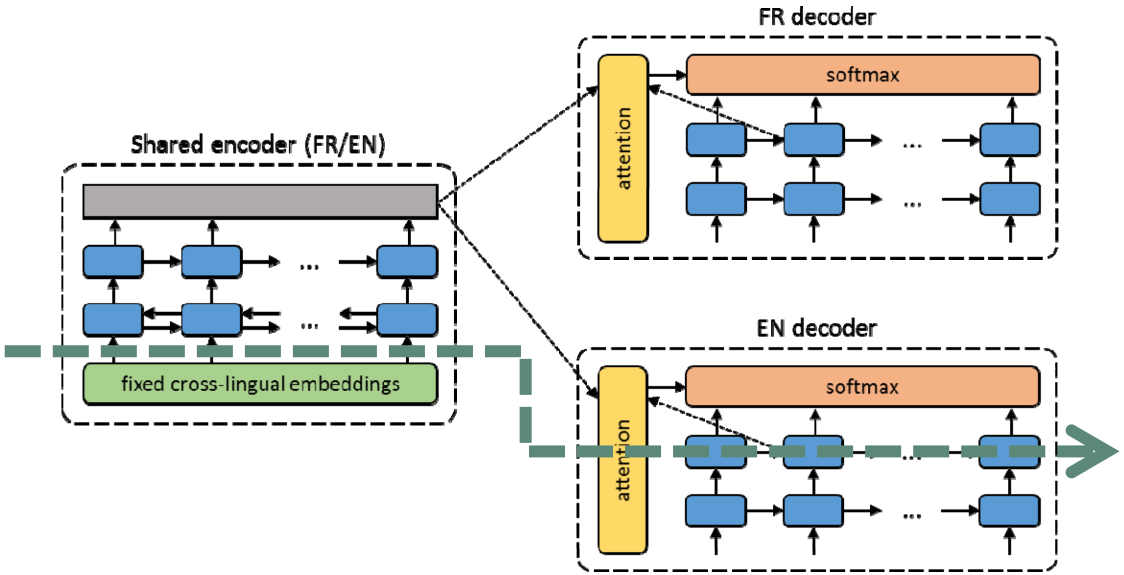
There was a shooting in Los Angeles International Airport.

Unsupervised neural machine translation

Training

- Supervised

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



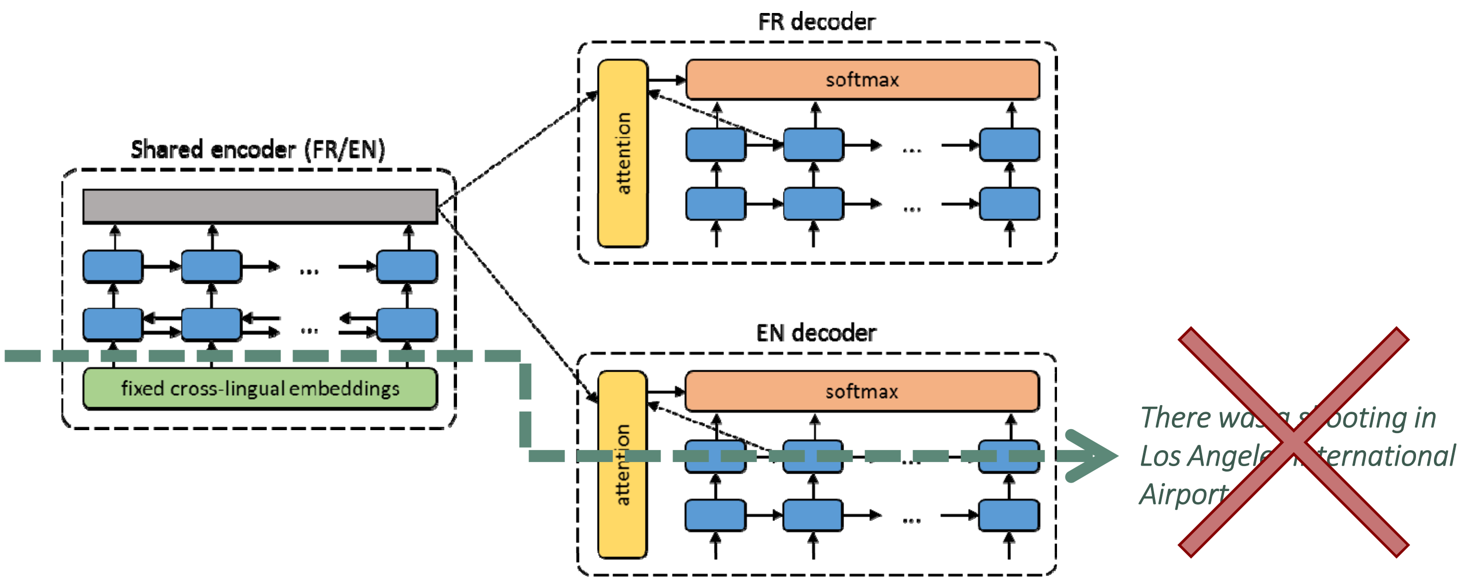
There was a shooting in Los Angeles International Airport.

Unsupervised neural machine translation

Training

- Supervised

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

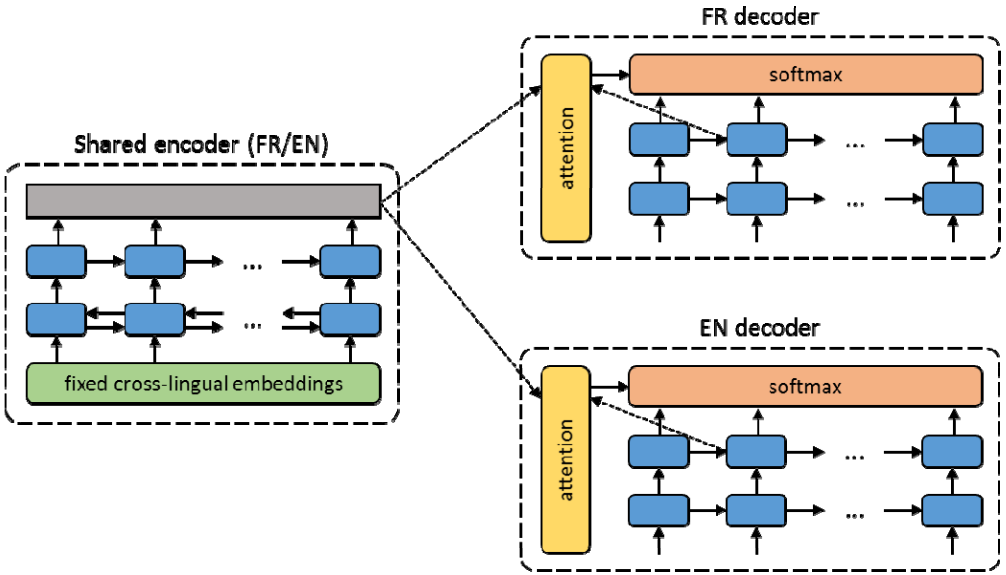


Unsupervised neural machine translation

Training

- ~~Supervised~~

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

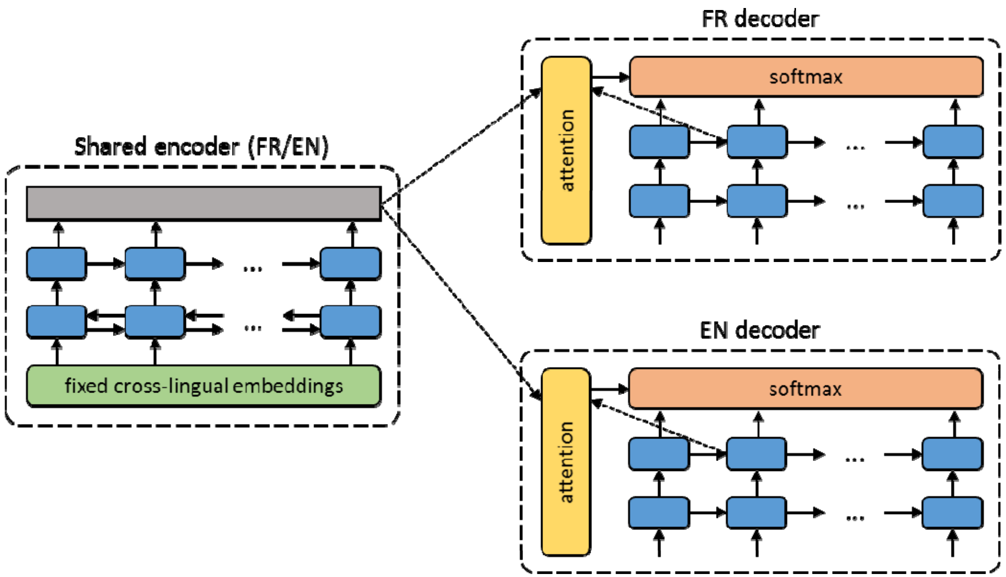


Unsupervised neural machine translation

Training

- ~~Supervised~~
- Autoencoder

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



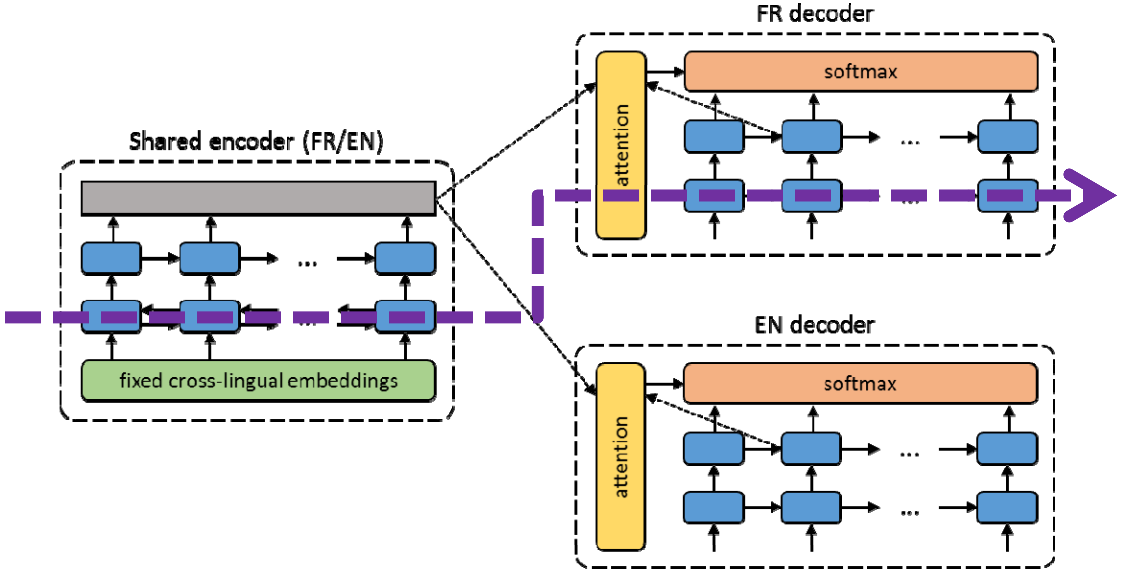
Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Autoencoder

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



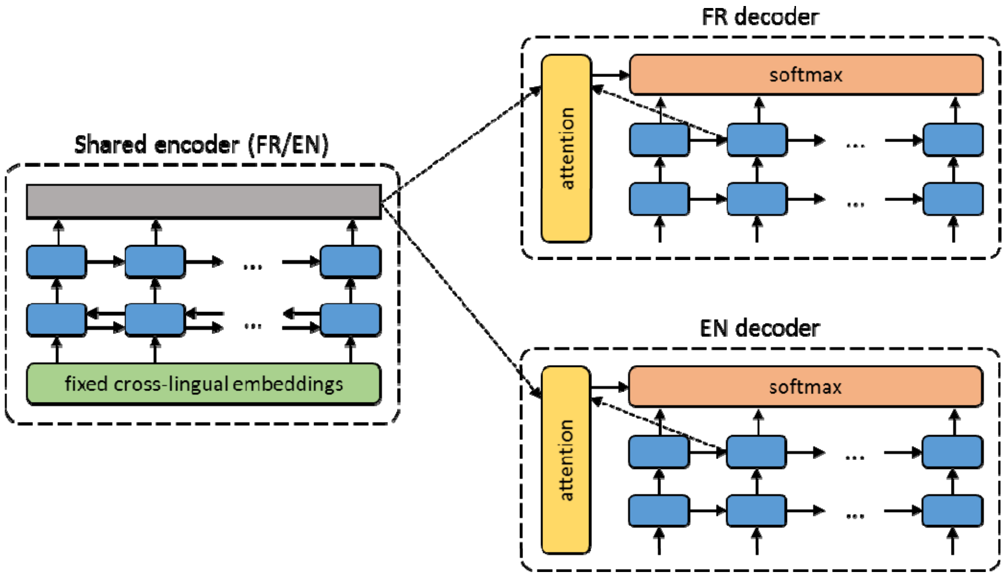
Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising Autoencoder

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

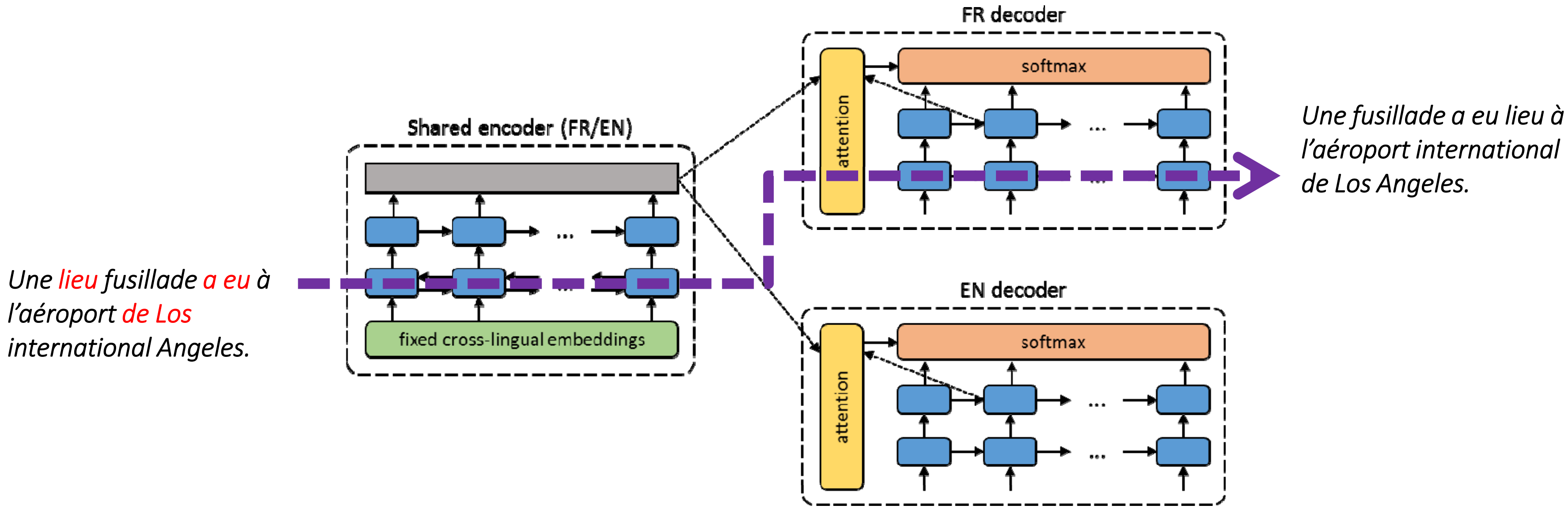


Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising Autoencoder

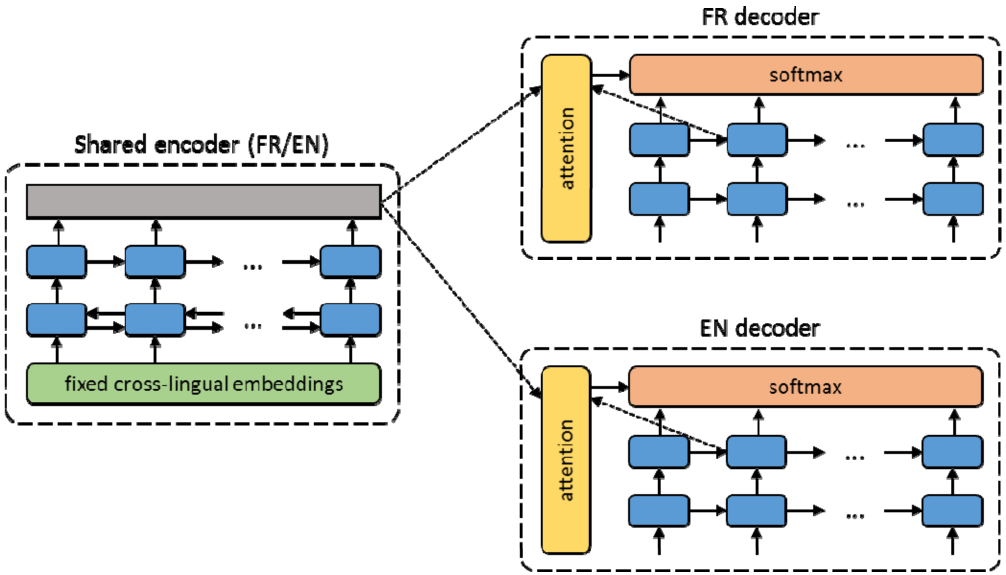


Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising Autoencoder

There a shooting was in Airport Los Angeles International.

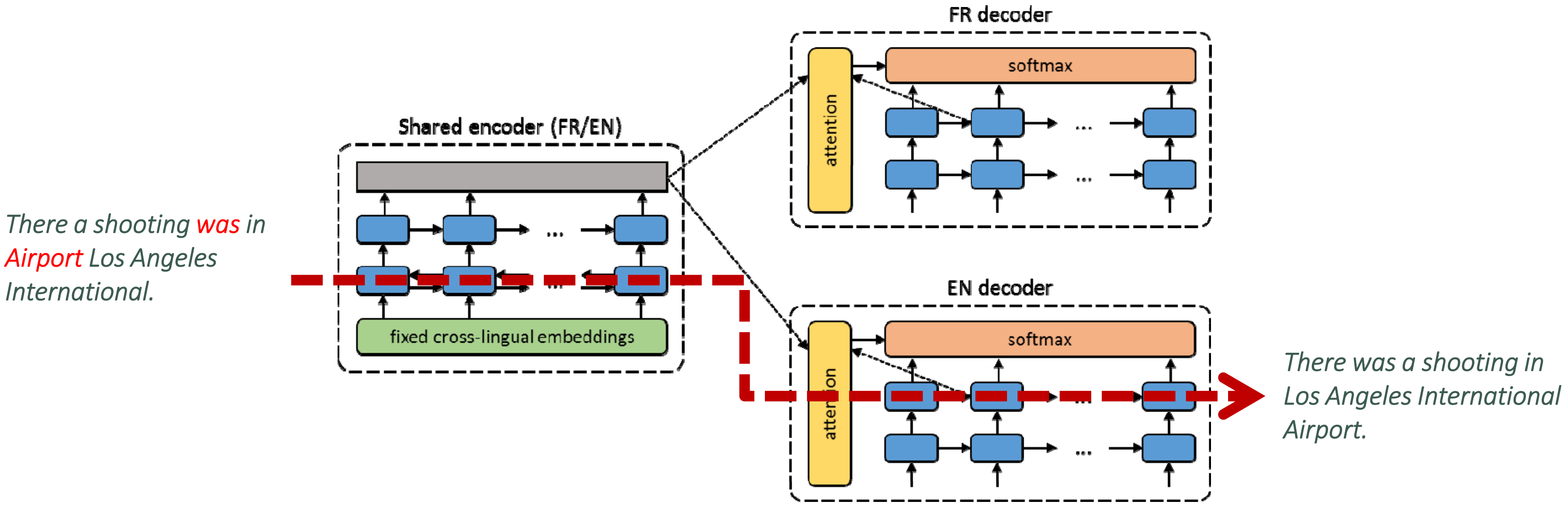


There was a shooting in Los Angeles International Airport.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising Autoencoder

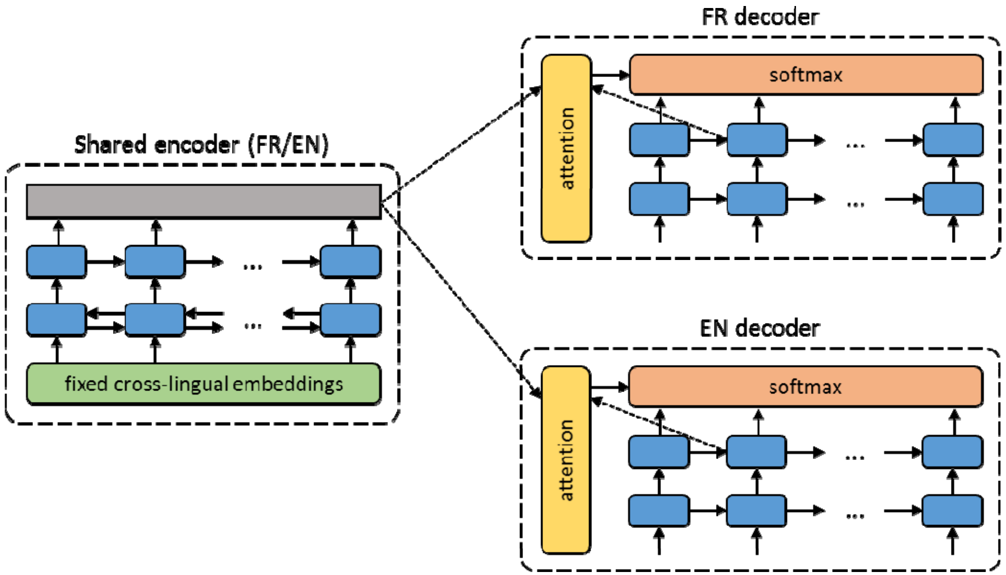


Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising
- Backtranslation

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

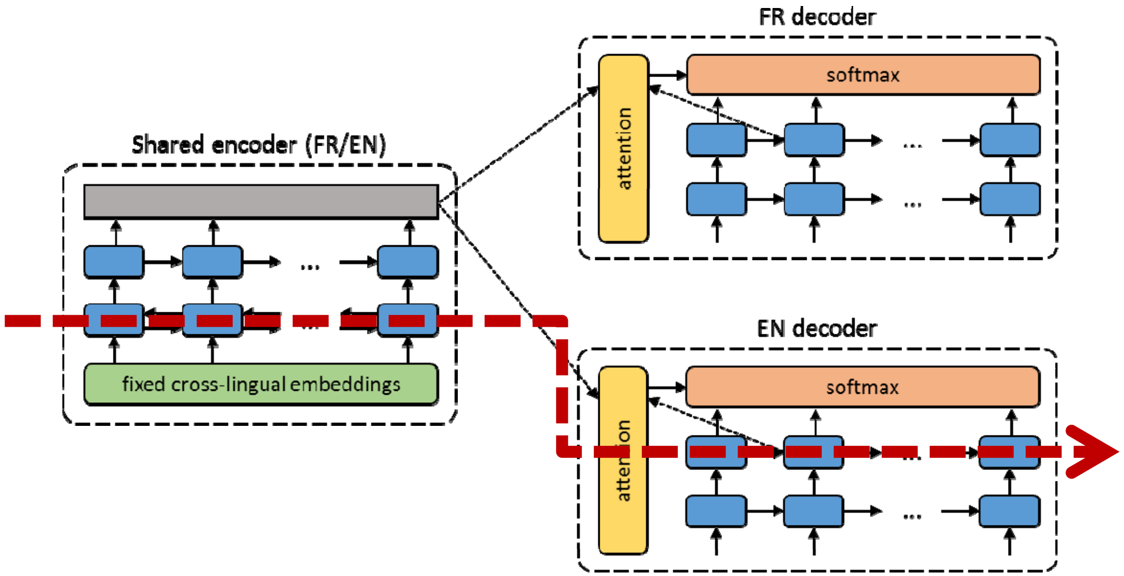


Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising
- Backtranslation

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

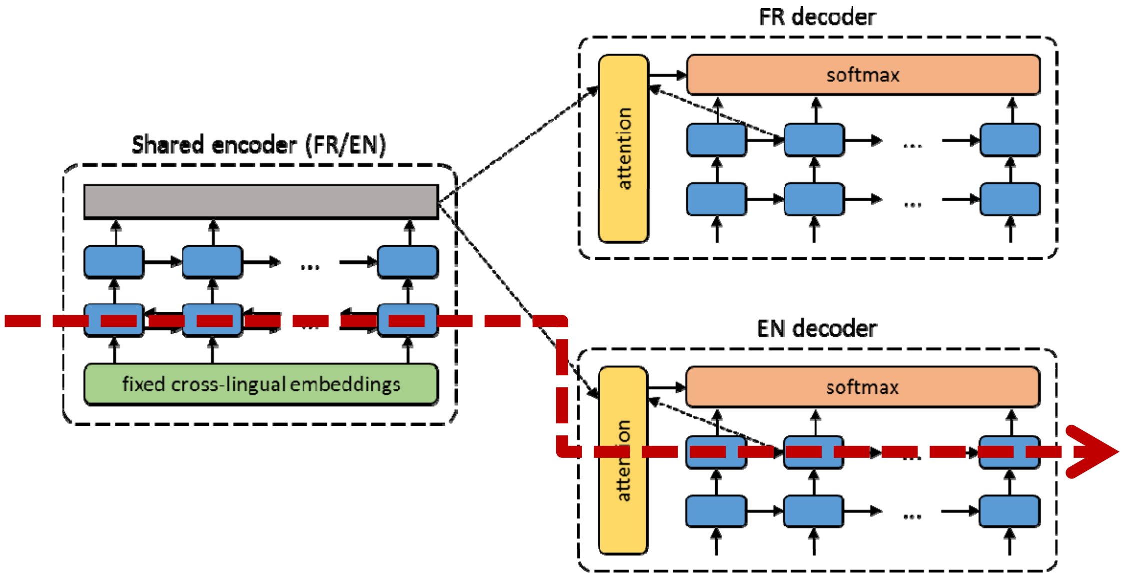


Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising
- Backtranslation

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

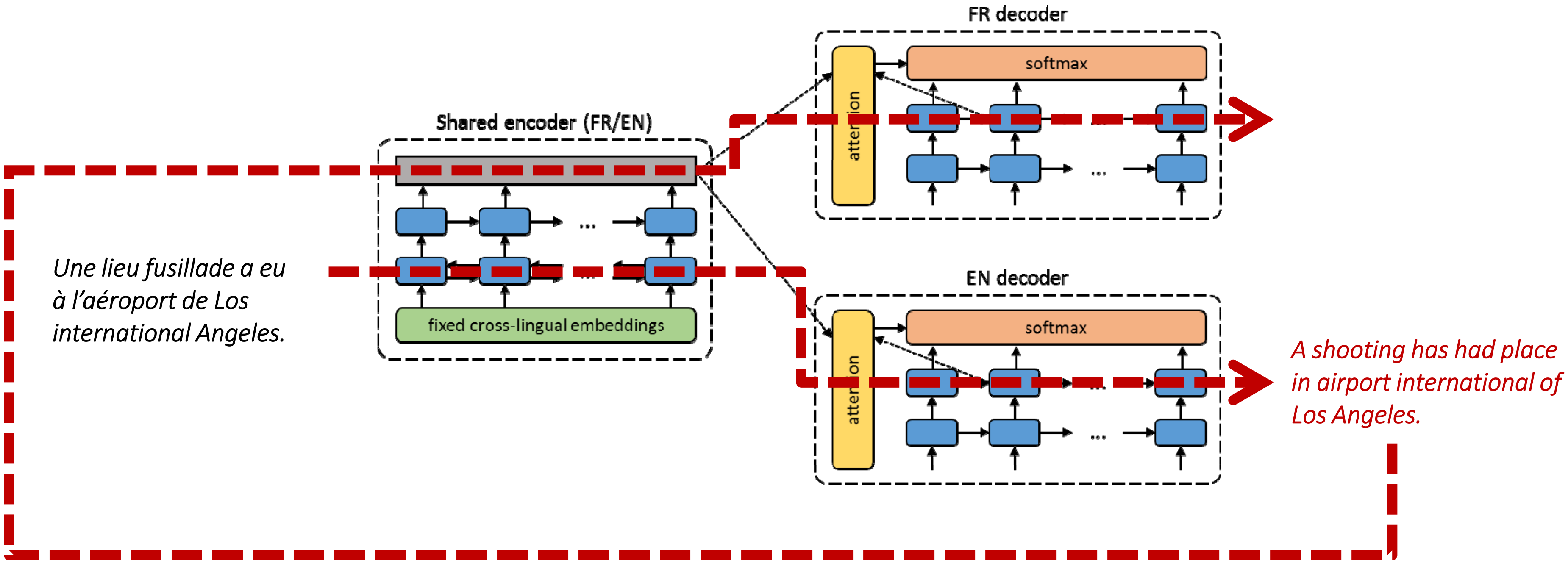


A shooting has had place in airport international of Los Angeles.

Unsupervised neural machine translation

Training

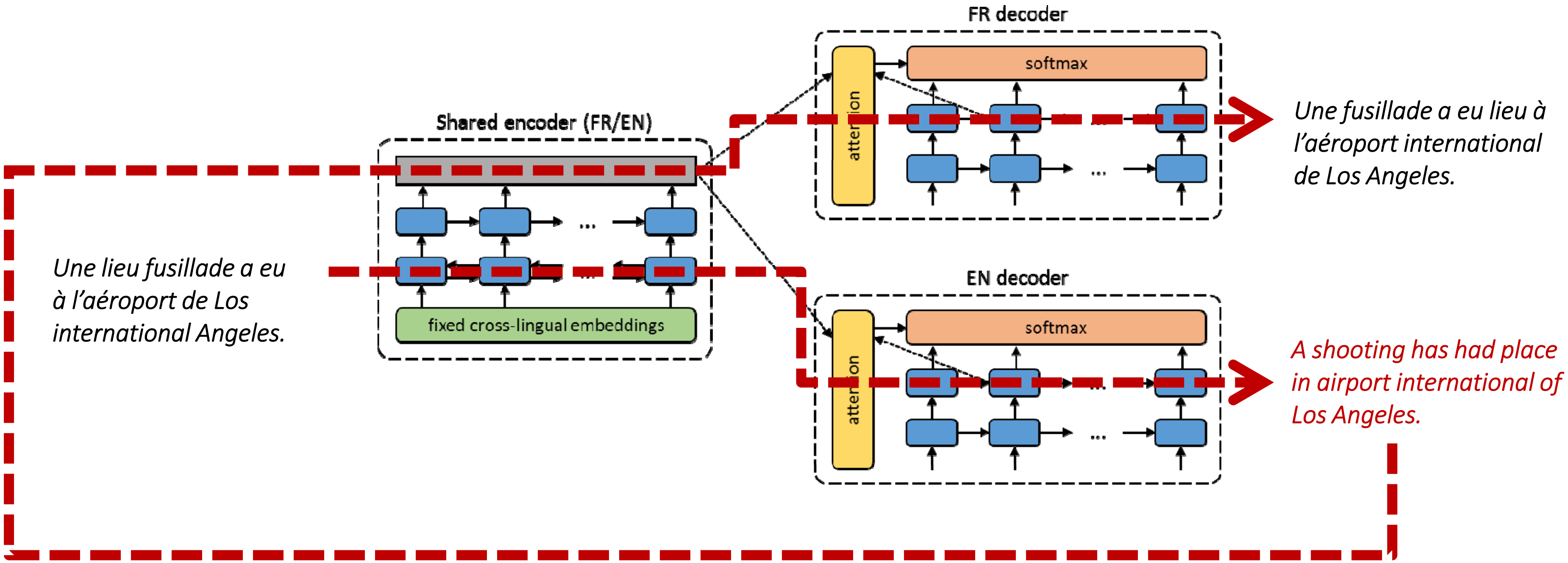
- ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised neural machine translation

Training

- Supervised ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised neural machine translation

- 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)
  - Fixed embeddings

- We change the **training regime, mixing mini-batches**:
  - Denoising autoencoder: noisy input in L1, output in the same language (E+D1)
  - Denoising autoencoder: noisy input in L2, output in the same language (E+D2)
  - Backtranslation: input in L1, translate E+D2, translate E+D1, output in L1
  - Backtranslation: input in L2, translate E+D1, translate E+D2, output in L2

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

	FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised NMT				

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised NMT	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised NMT	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised NMT	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	**15.56**	**15.13**	**10.21**	**6.55**

It works!

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised NMT	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	15.56	15.13	10.21	6.55
Semi-supervised NMT	Proposed (full) + 10k parallel	**18.57**	**17.34**	**11.47**	**7.86**
	Proposed (full) + 100k parallel	**21.81**	**21.74**	**15.24**	**10.95**

It can be easily combined with training data (interesting for low resource MT)

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised NMT	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	**15.56**	**15.13**	**10.21**	**6.55**
	Lample et al. 2018 (Same conference!)	14.31	15.06	-	-

State-of-the-art (not anymore…)

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised NMT	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	**15.56**	**15.13**	**10.21**	**6.55**
	Lample et al. 2018	14.31	15.06	-	-
	Lample et al. 2018b				

Lample et al. 2018b (EMNLP)

- No embedding mappings
- BPE jointly over monolingual corpora. Fails for less related languages (Russian).
- Shared decoder for both languages
- Transformer (instead of LSTM)

Unsupervised neural machine translation

Test on WMT released data (test and monolingual corpora)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised NMT	Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	Proposed (denoising)	7.28	5.33	3.64	2.40
	Proposed (+backtranslation)	15.56	15.13	10.21	6.55
	Lample et al. 2018	14.31	15.06	-	-
	Lample et al. 2018b	**24.2**	**25.1**	**21.0**	**17.2**

Lample et al. 2018b (EMNLP)

- No embedding mappings
- BPE jointly over monolingual corpora. Fails for less related languages (Russian).
- Shared decoder for both languages
- Transformer (instead of LSTM)

Unsupervised statistical machine translation

Unsupervised statistical machine translation

Artetxe et al. 2018b (EMNLP)

- Estimate PBMT parameters
    - Learn monolingual embeddings for bigrams and trigrams
    - Initialize phrase table using prob. estimates from cross-lingual mappings
    - Unsupervised tuning based on back-translation
- Use backtranslation and train reverse PBMT from scratch. Iterate.

Unsupervised statistical machine translation

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

		FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
Unsupervised NMT	Artetxe et al. 2018	15.56	15.13	10.21	6.55		
	Lample et al. 2018b	24.2	25.1			21.0	17.2
Unsupervised PBMT							

Artetxe et al. 2018b (EMNLP)

- Estimate PBMT parameters
    - Learn monolingual embeddings for bigrams and trigrams
    - Initialize phrase table using prob. estimates from cross-lingual mappings
    - Unsupervised tuning based on back-translation
- Use backtranslation and train reverse PBMT from scratch. Iterate.

Unsupervised statistical machine translation

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

		FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
Unsupervised NMT	Artetxe et al. 2018	15.56	15.13	10.21	6.55		
	Lample et al. 2018b	24.2	25.1			21.0	17.2
Unsupervised PBMT	Artetxe et al. 2018b	25.87	26.22	17.43	14.08	23.05	18.23

Artetxe et al. 2018b (EMNLP)

- Estimate PBMT parameters
    - Learn monolingual embeddings for bigrams and trigrams
    - Initialize phrase table using prob. estimates from cross-lingual mappings
    - Unsupervised tuning based on back-translation
- Use backtranslation and train reverse PBMT from scratch. Iterate.

Unsupervised statistical machine translation

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

		FR-EN	EN-FR	DE-EN	EN-DE	DE-EN	EN-DE
Unsupervised NMT	Artetxe et al. 2018	15.56	15.13	10.21	6.55		
	Lample et al. 2018b	24.2	25.1			21.0	17.2
Unsupervised PBMT	Artetxe et al. 2018b	25.87	26.22	17.43	14.08	**23.05**	**18.23**
	Lample et al. 2018b	**27.16**	**28.11**			22.68	17.77

Artetxe et al. 2018b (EMNLP)

- Estimate PBMT parameters
  - Learn monolingual embeddings for bigrams and trigrams
  - Initialize phrase table using prob. estimates from cross-lingual mappings
  - Unsupervised tuning based on back-translation
- Use backtranslation and train reverse PBMT from scratch. Iterate.

Unsupervised machine translation

Getting closer to supervised machine translation!

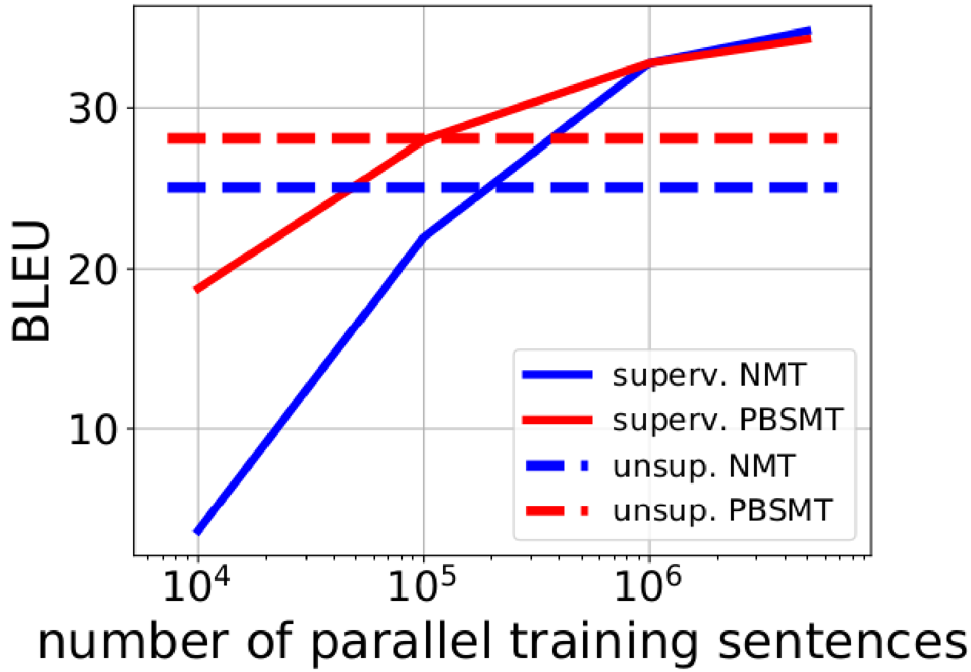


Figure 2: Comparison between supervised and unsupervised approaches on WMT'14 En-Fr, as we vary the number of parallel sentences for the supervised methods.

Source: (Lample et al. 2018)

Unsupervised machine translation

Getting closer to supervised machine translation!

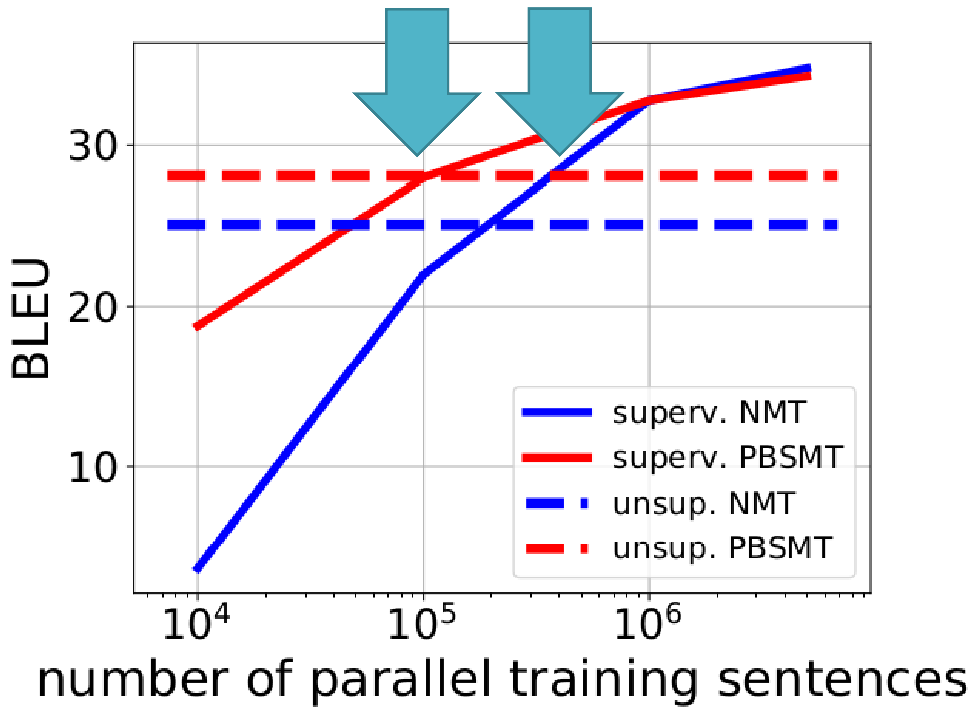


Figure 2: Comparison between supervised and unsupervised approaches on WMT'14 En-Fr, as we vary the number of parallel sentences for the supervised methods.

Source: (Lample et al. 2018)

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 figures out how to best “combine” them

Conclusions

- New research area – unsupervised Machine Translation

  The main Machine Translation competition (WMT18) has now an **unsupervised track**

- Performance up, 28 BLEU En-Fr

- Plenty of margin for improvement

- Code for replicability

  `https://github.com/artetxem/undreamt`
  `https://github.com/artetxem/monoses (soon)`

References: unsupervised MT

- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Unsupervised Neural Machine Translation. In *ICLR-2018*.

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

Final words

- **Word embeddings key** for Natural Language Processing
- Mappings represent **languages in common space**
  - Most of language pairs have **very few resources**
  - New research area: **only monolingual resources**

Final words

- **Word embeddings key** for Natural Language Processing
- Mappings represent **languages in common space**
  - Most of language pairs have **very few resources**
  - New research area: **only monolingual resources**
- **Cross-lingual unsupervised mappings** enabled breakthroughs in
  - Bilingual dictionary induction
  - Unsupervised machine translation
  - Confirmed in (Conneau et al. 2018; Lample et al. 2018)

Final words

- **Word embeddings key** for Natural Language Processing
- Mappings represent **languages in common space**
  - Most of language pairs have **very few resources**
  - New research area: **only monolingual resources**
- **Cross-lingual unsupervised mappings** enabled breakthroughs in
  - Bilingual dictionary induction
  - Unsupervised machine translation
  - Confirmed in (Conneau et al. 2018; Lample et al. 2018)
- Unexplored area in its **infancy**
  - Potential for **MT in low resource languages and domains**
  - Potential for **transforming the NLP landscape**
    - From monolingual NLP (e.g. English) to **multilingual tools**
    - **Universal sentence representations**

Thank you!

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

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