

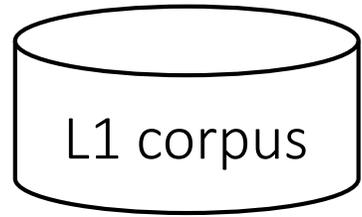
Towards Unsupervised Machine Translation

Mikel Artetxe

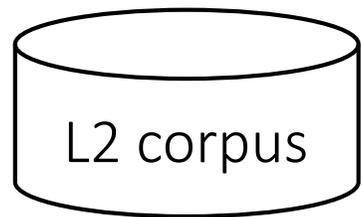
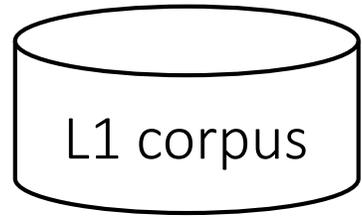
IXA NLP group – University of the Basque Country (UPV/EHU)

Introduction

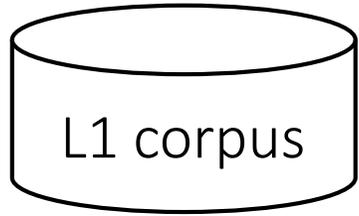
Introduction



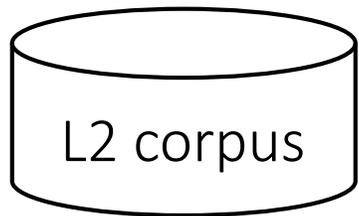
Introduction



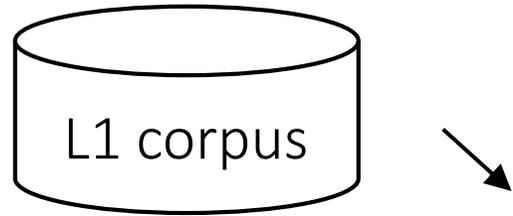
Introduction



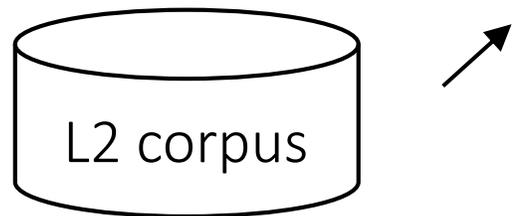
non-parallel



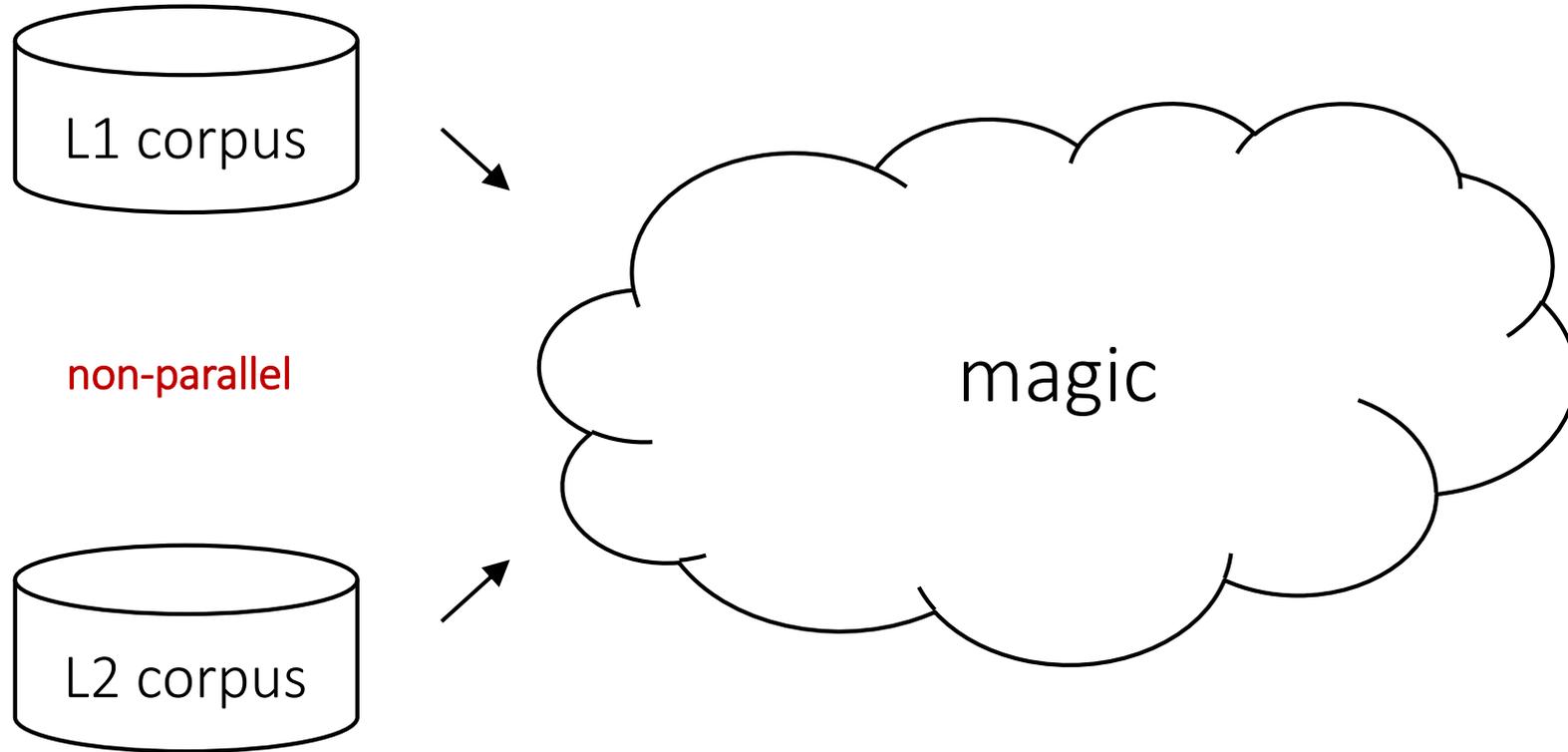
Introduction



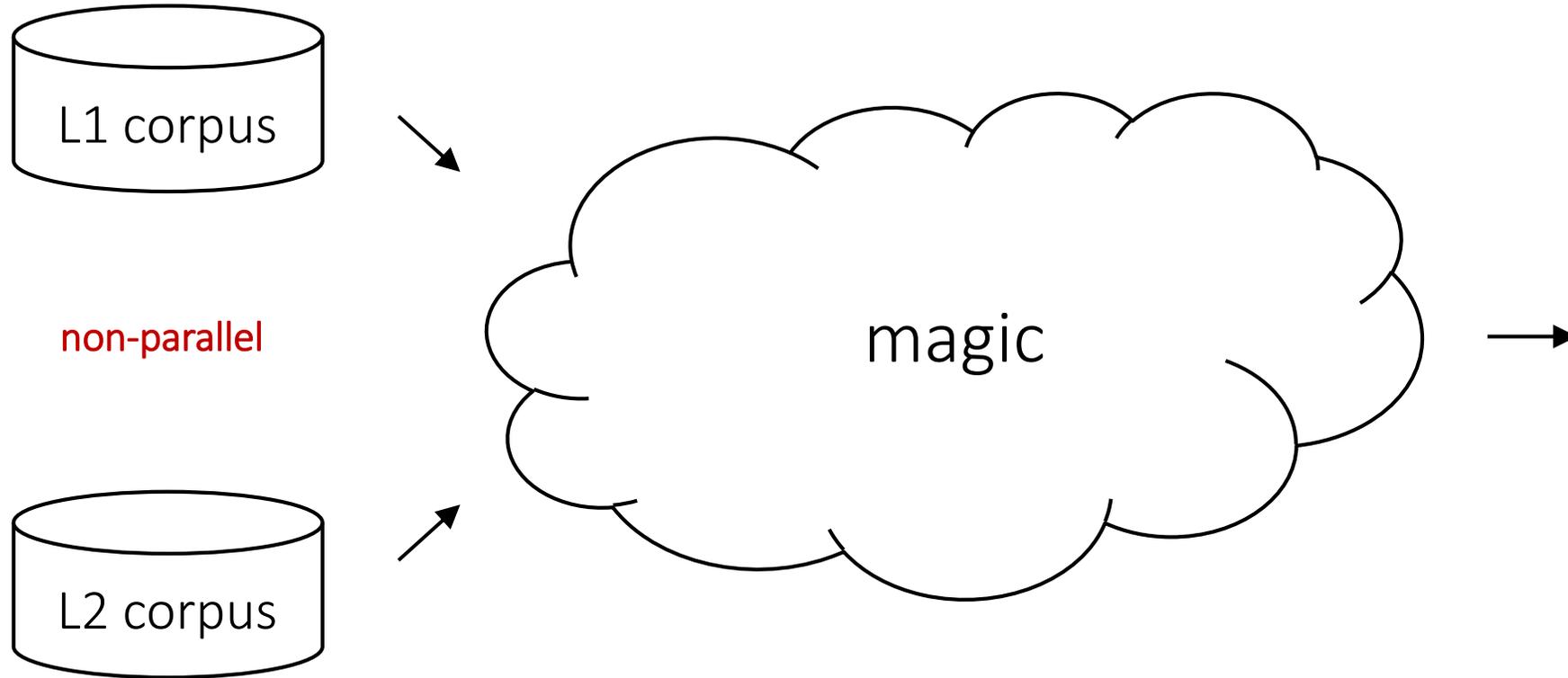
non-parallel



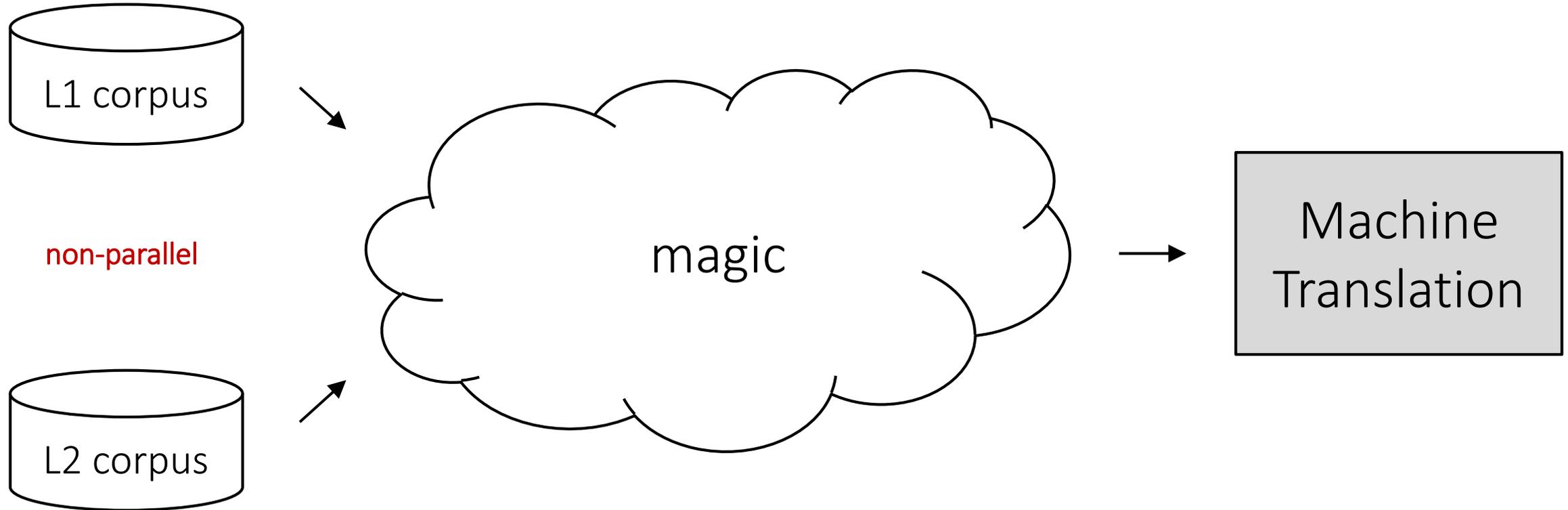
Introduction



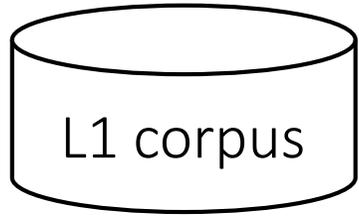
Introduction



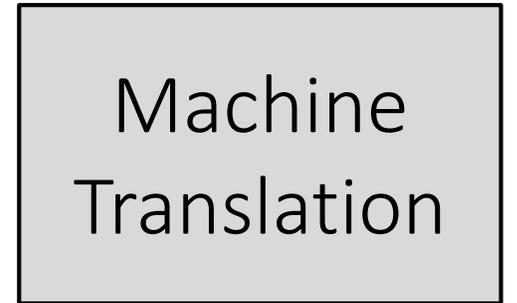
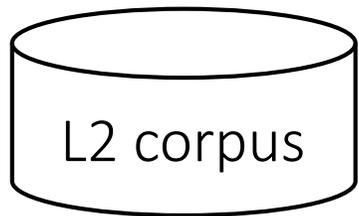
Introduction



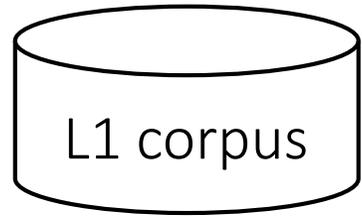
Introduction



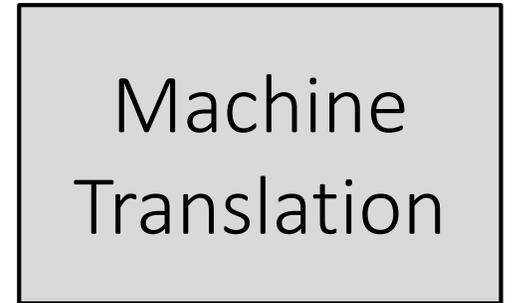
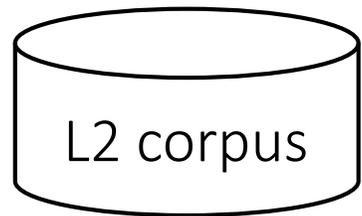
non-parallel



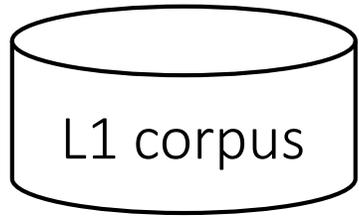
Introduction



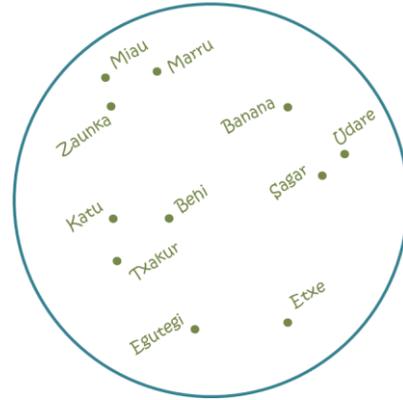
non-parallel



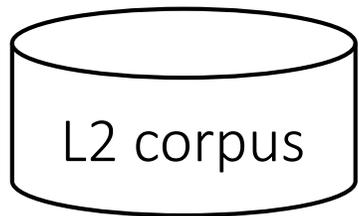
Introduction



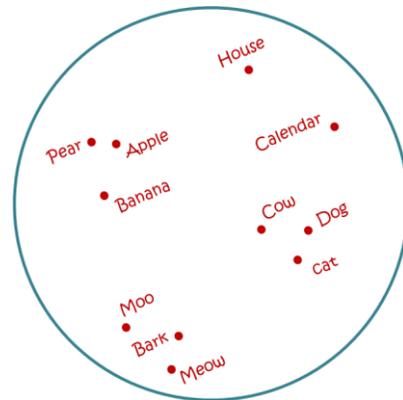
L1 embeddings



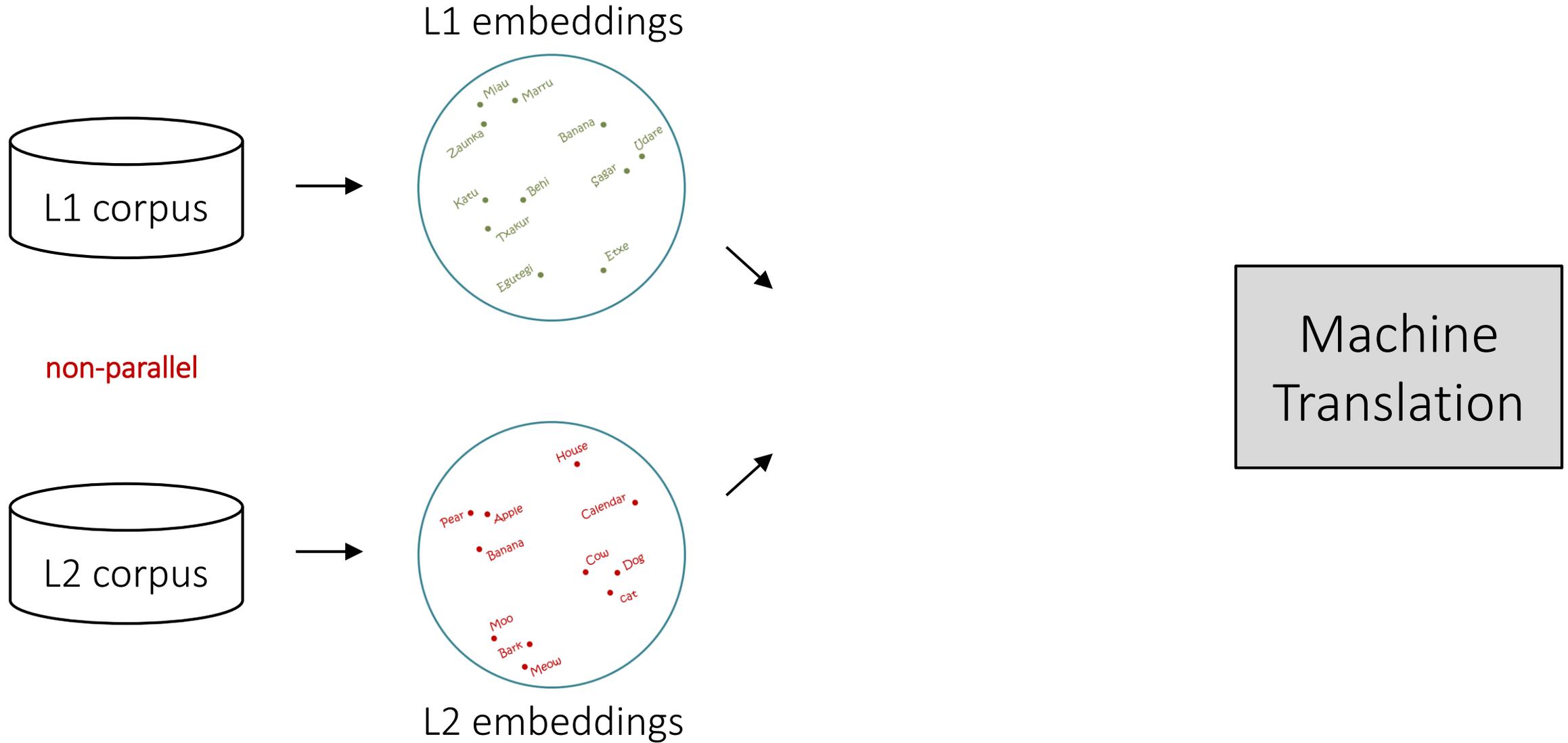
non-parallel



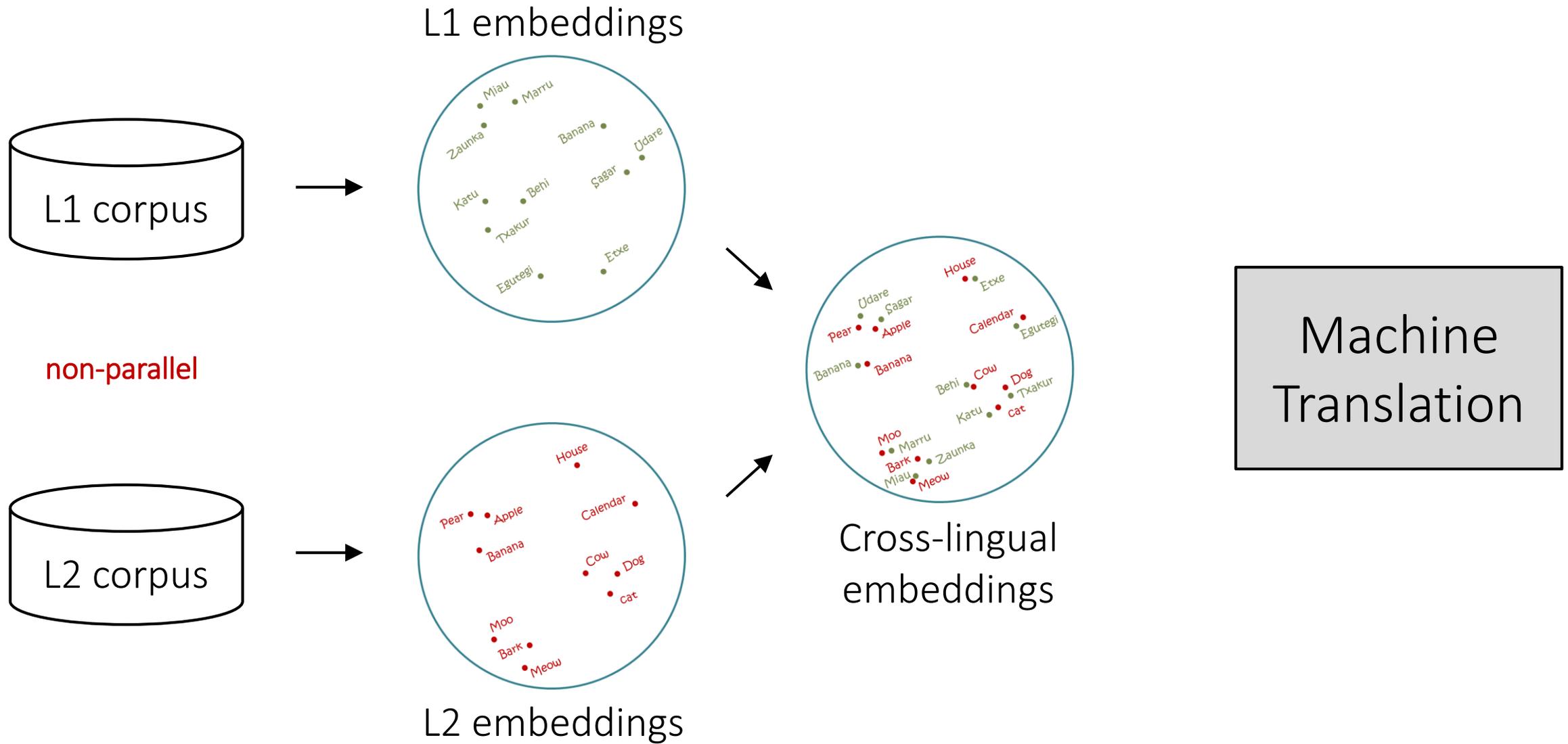
L2 embeddings



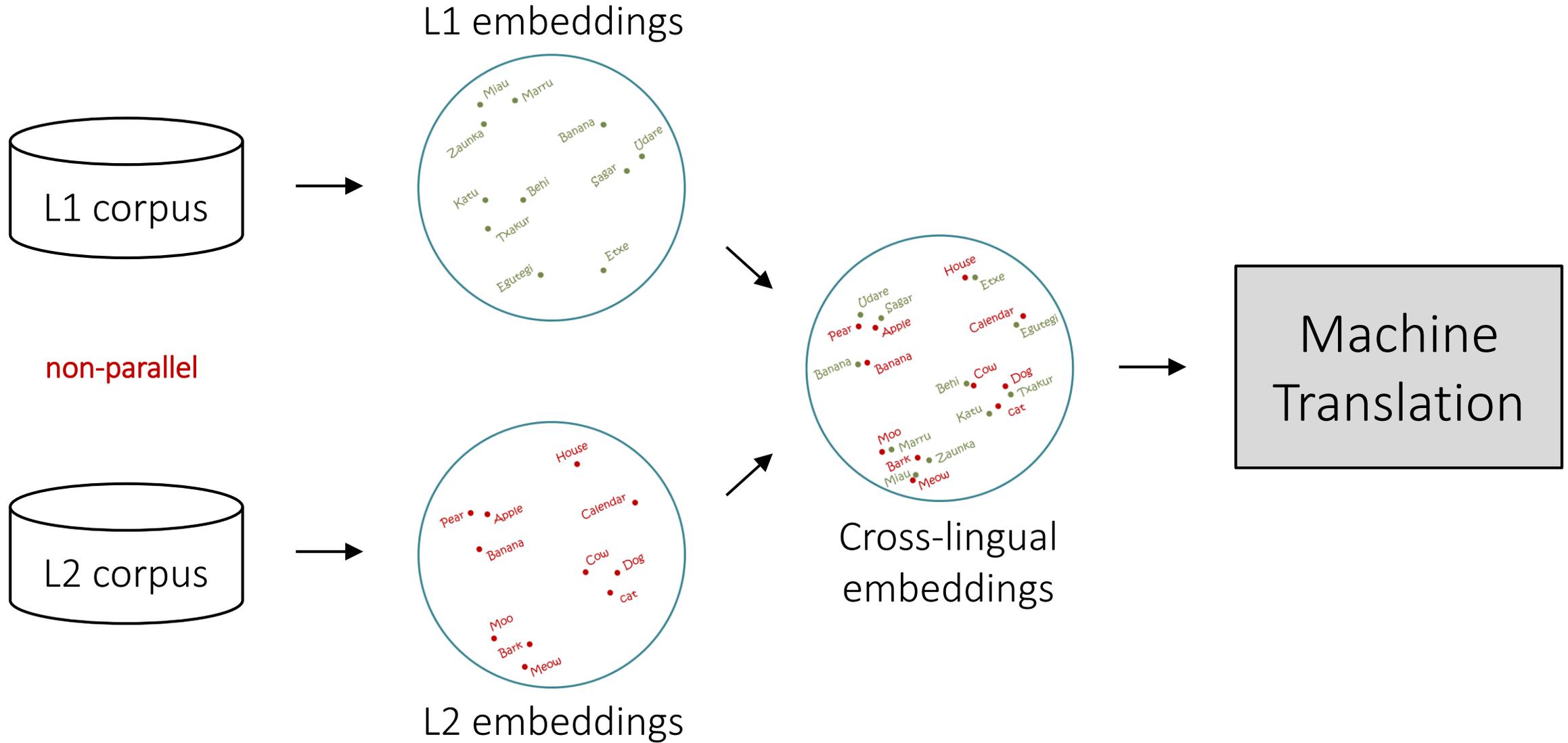
Introduction



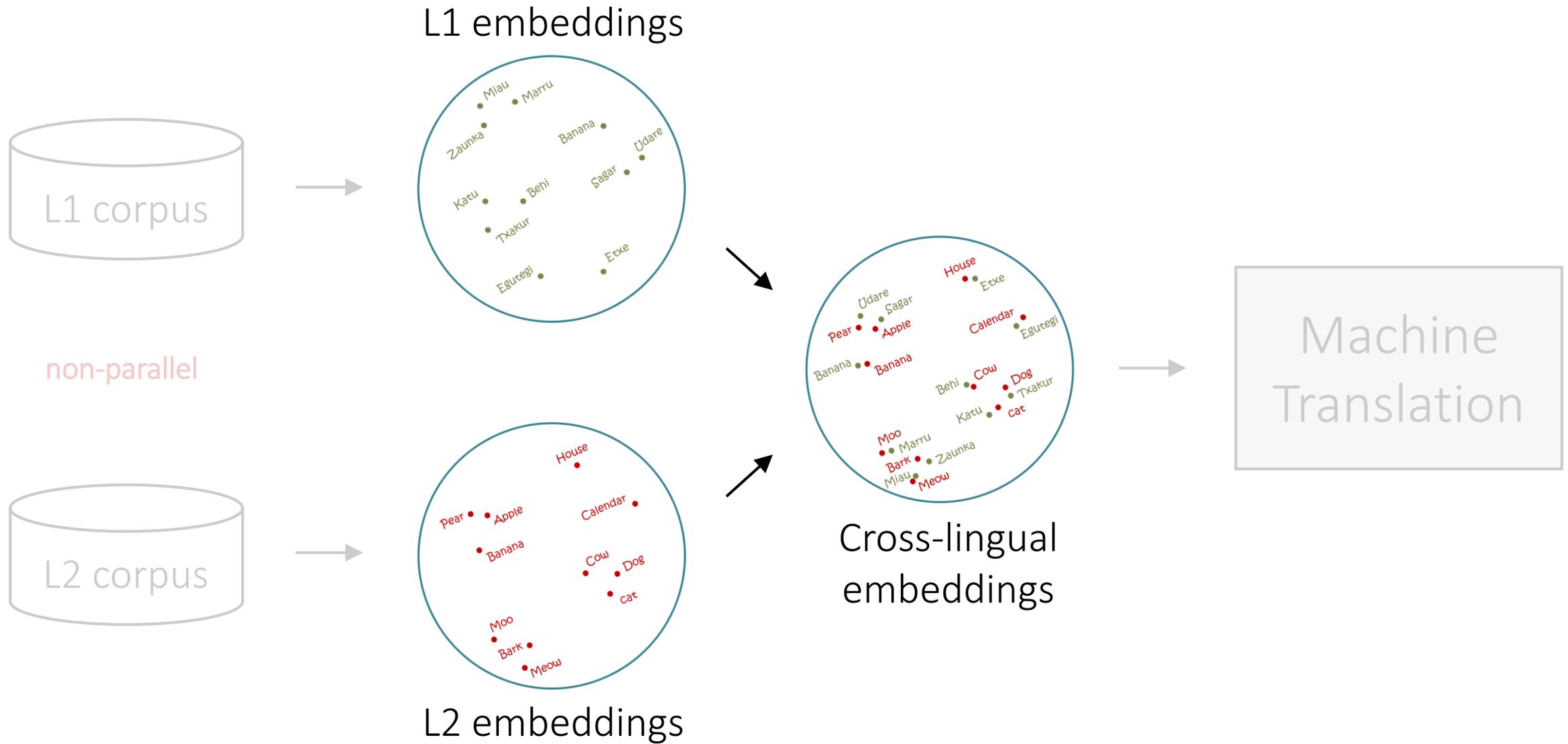
Introduction



Introduction

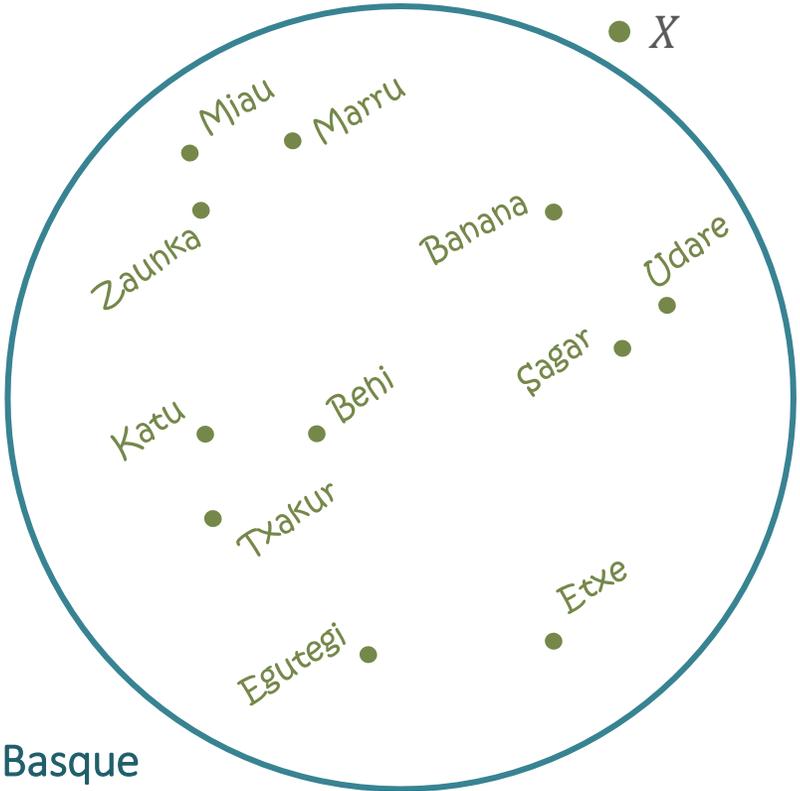


Introduction

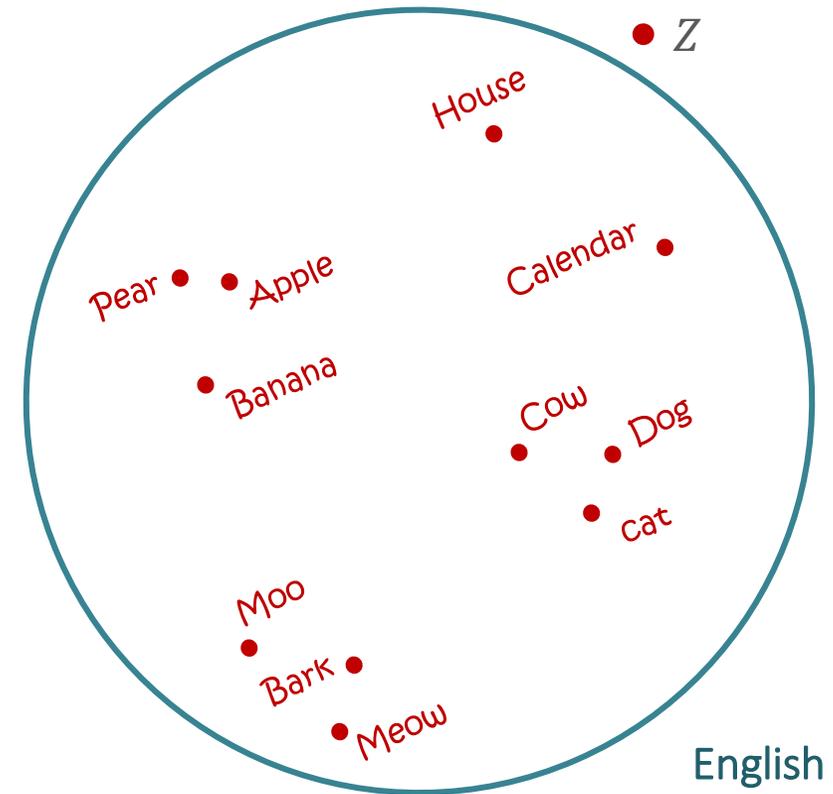
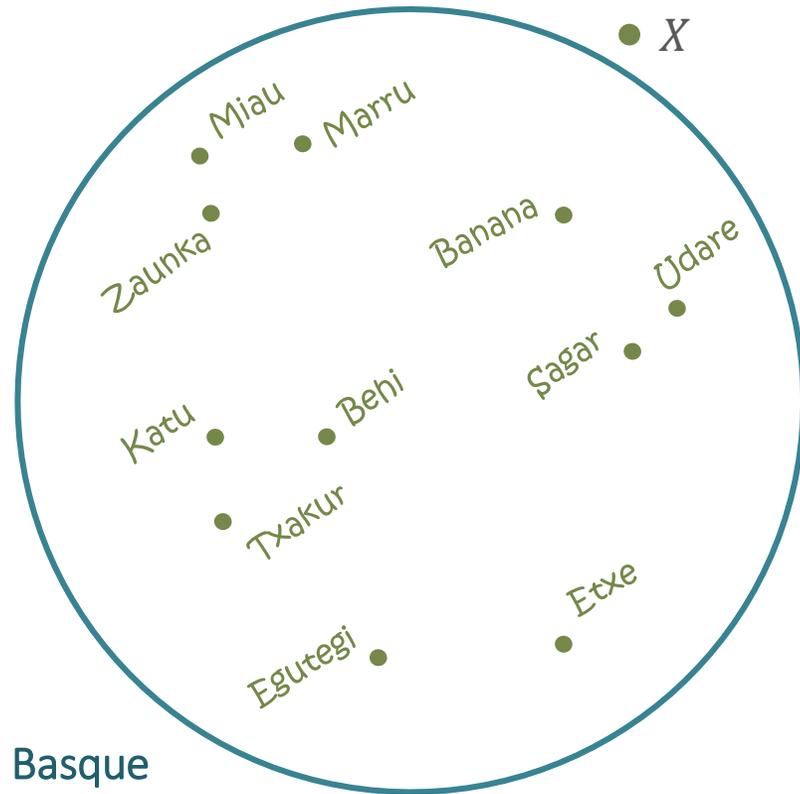


Cross-lingual embedding mappings

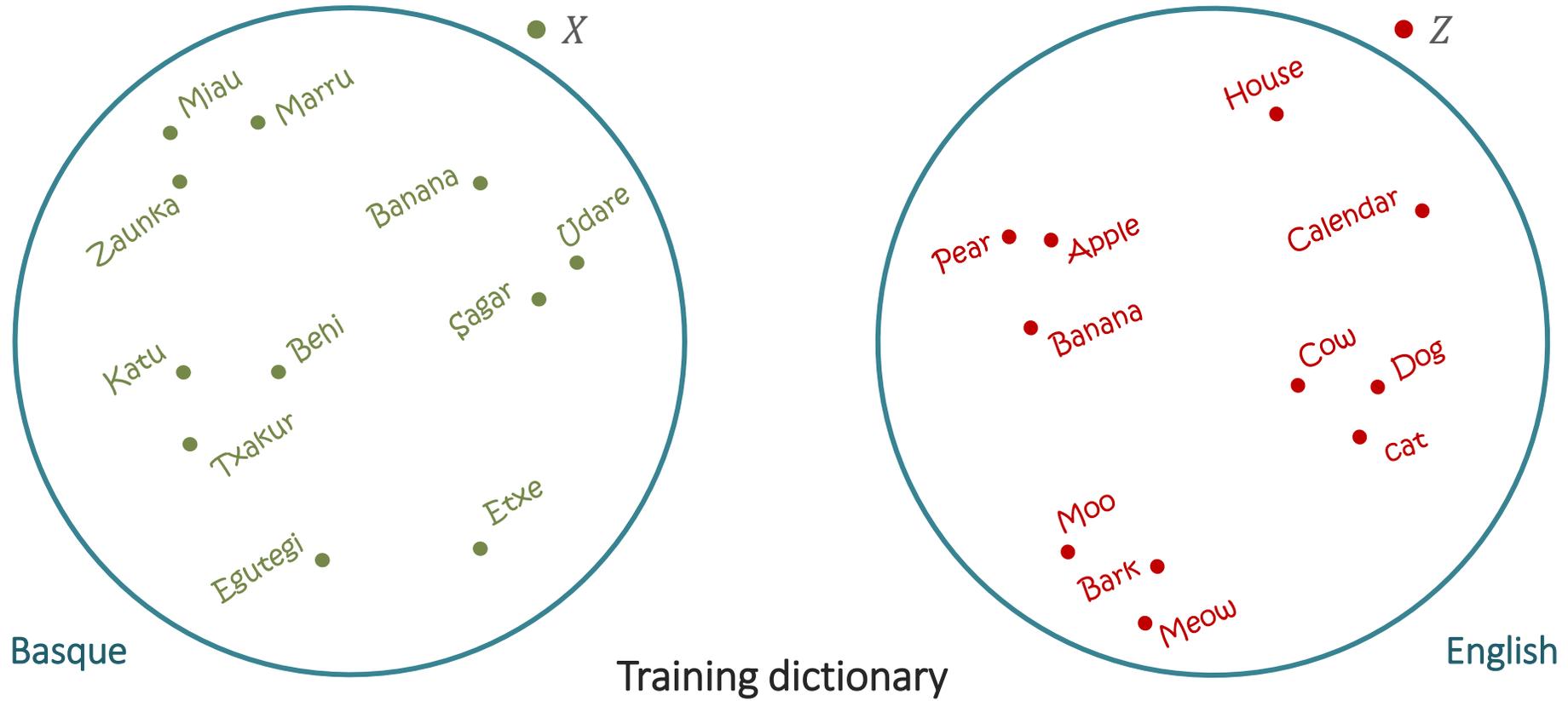
Cross-lingual embedding mappings



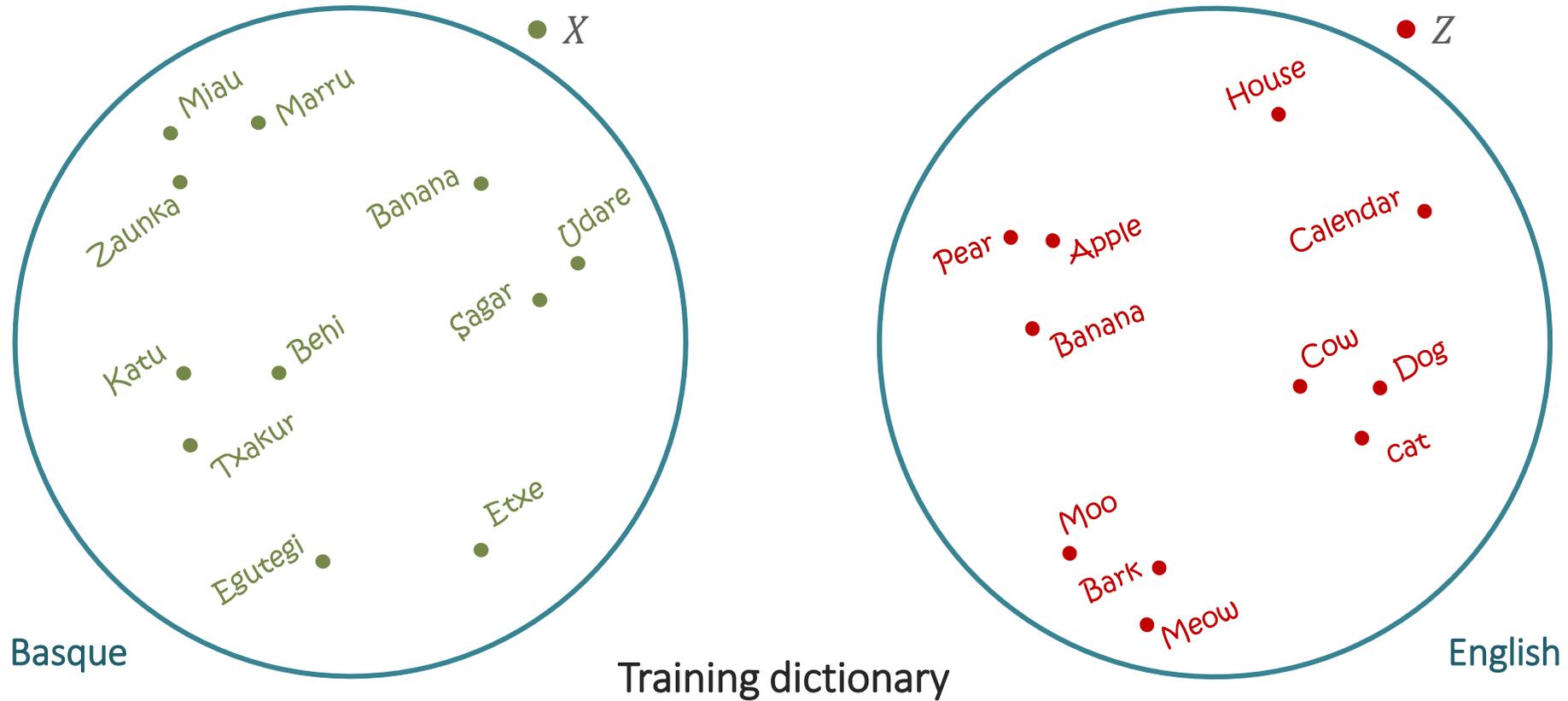
Cross-lingual embedding mappings



Cross-lingual embedding mappings



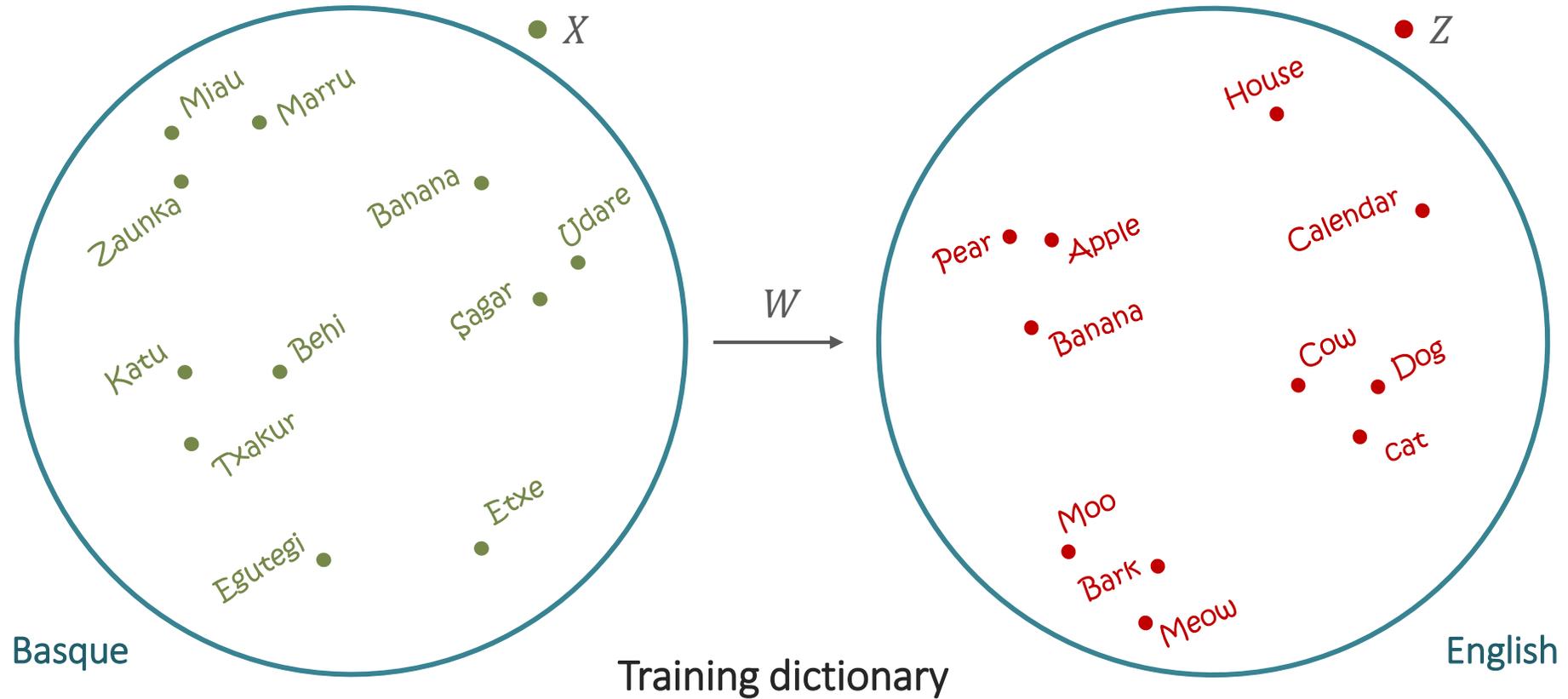
Cross-lingual embedding mappings



Txakur
Sagar
⋮
Egutegi

Dog
Apple
⋮
Calendar

Cross-lingual embedding mappings



Cross-lingual embedding mappings



Txakur
Sagar
⋮
Egutegi

Dog
Apple
⋮
Calendar

Cross-lingual embedding mappings



Txakur $\begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix}$

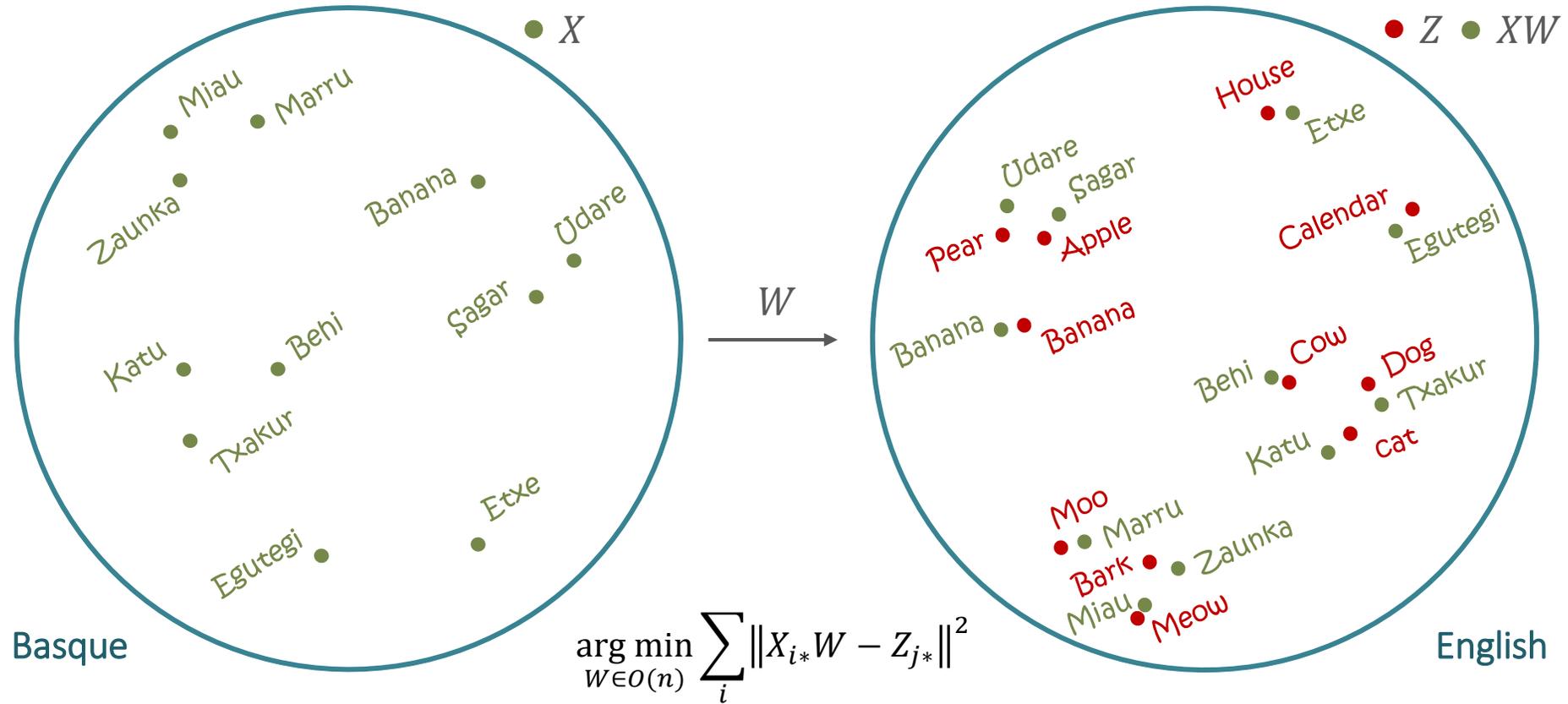
$\begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix}$ Dog
 Apple
 \vdots
 Calendar

Cross-lingual embedding mappings



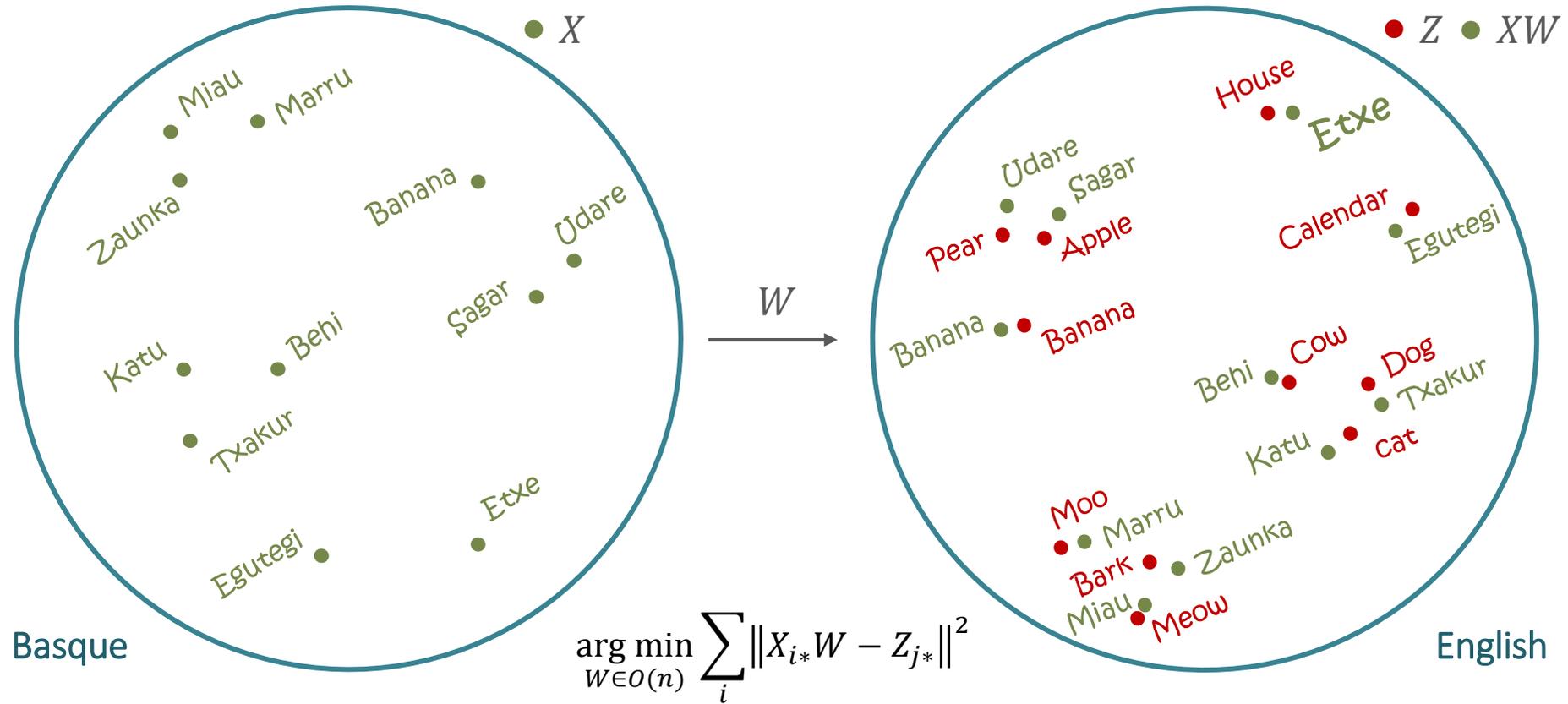
$$\begin{array}{l}
 \text{Txakur} \\
 \text{Sagar} \\
 \vdots \\
 \text{Egutegi}
 \end{array}
 \begin{bmatrix}
 X_{1,*} \\
 X_{2,*} \\
 \vdots \\
 X_{n,*}
 \end{bmatrix}
 [W] \approx
 \begin{bmatrix}
 Z_{1,*} \\
 Z_{2,*} \\
 \vdots \\
 Z_{n,*}
 \end{bmatrix}
 \begin{array}{l}
 \text{Dog} \\
 \text{Apple}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

Cross-lingual embedding mappings



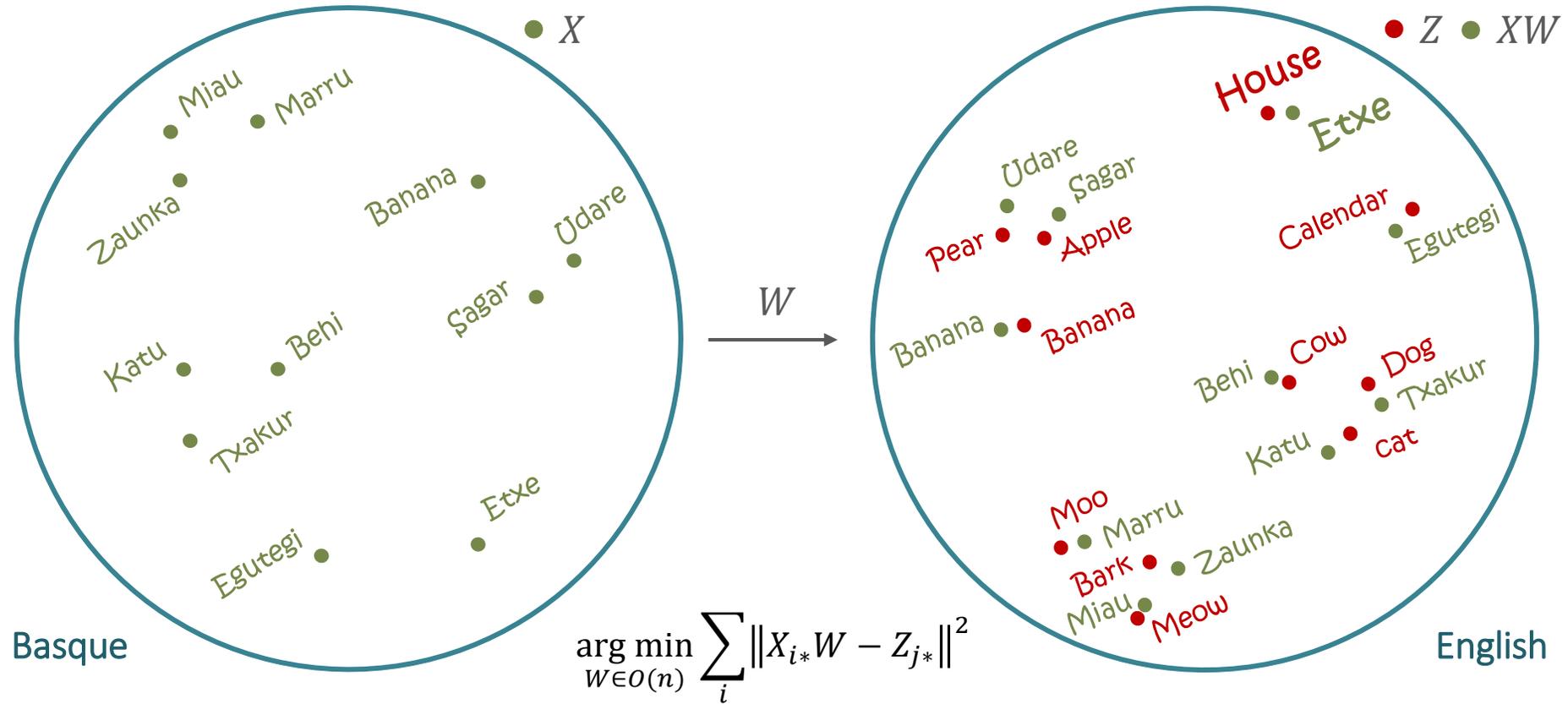
$$\begin{array}{l}
 \text{Txakur} \\
 \text{Sagar} \\
 \vdots \\
 \text{Egutegi}
 \end{array}
 \begin{bmatrix}
 X_{1,*} \\
 X_{2,*} \\
 \vdots \\
 X_{n,*}
 \end{bmatrix}
 [W] \approx
 \begin{bmatrix}
 Z_{1,*} \\
 Z_{2,*} \\
 \vdots \\
 Z_{n,*}
 \end{bmatrix}
 \begin{array}{l}
 \text{Dog} \\
 \text{Apple}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

Cross-lingual embedding mappings



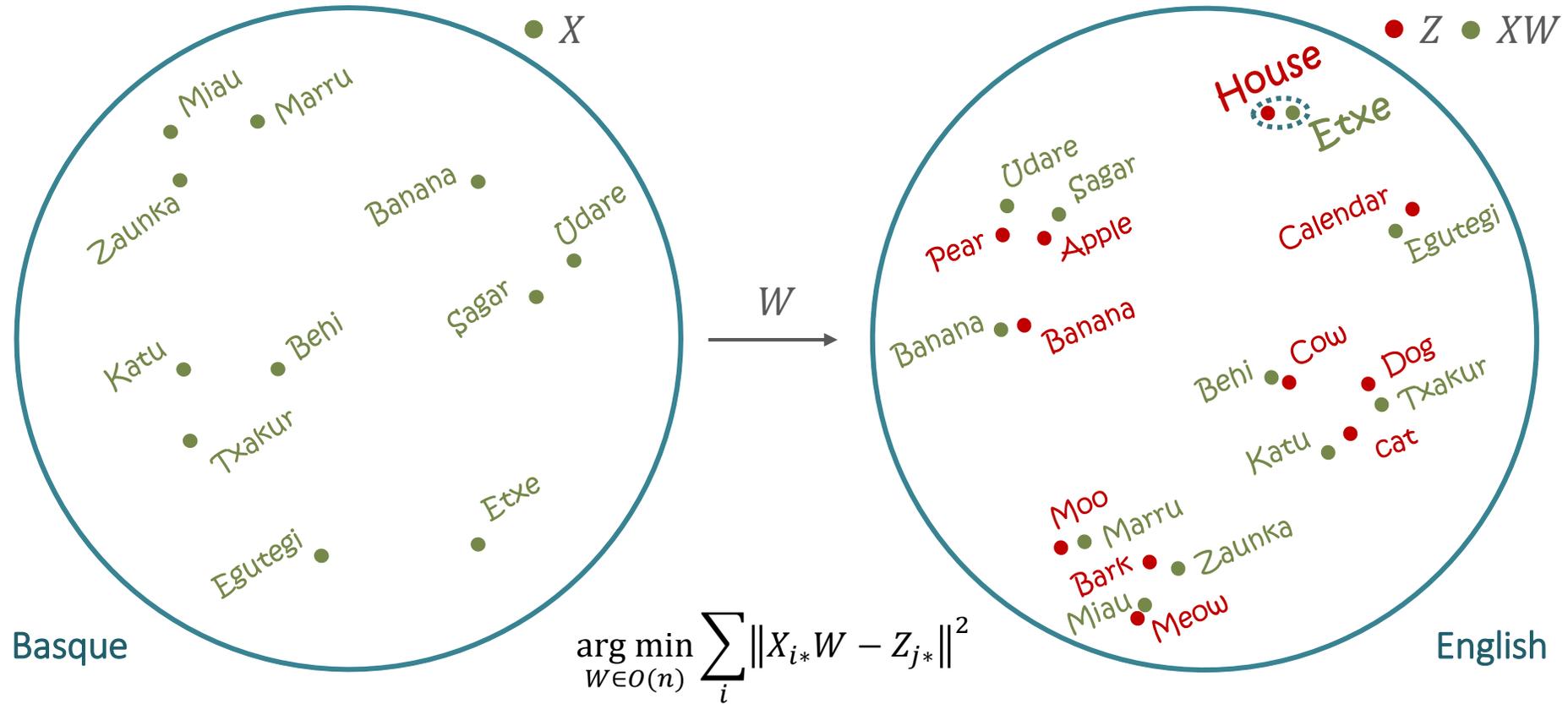
$$\begin{array}{l}
 \text{Txakur} \\
 \text{Sagar} \\
 \vdots \\
 \text{Egutegi}
 \end{array}
 \begin{bmatrix}
 X_{1,*} \\
 X_{2,*} \\
 \vdots \\
 X_{n,*}
 \end{bmatrix}
 [W] \approx
 \begin{bmatrix}
 Z_{1,*} \\
 Z_{2,*} \\
 \vdots \\
 Z_{n,*}
 \end{bmatrix}
 \begin{array}{l}
 \text{Dog} \\
 \text{Apple}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

Cross-lingual embedding mappings



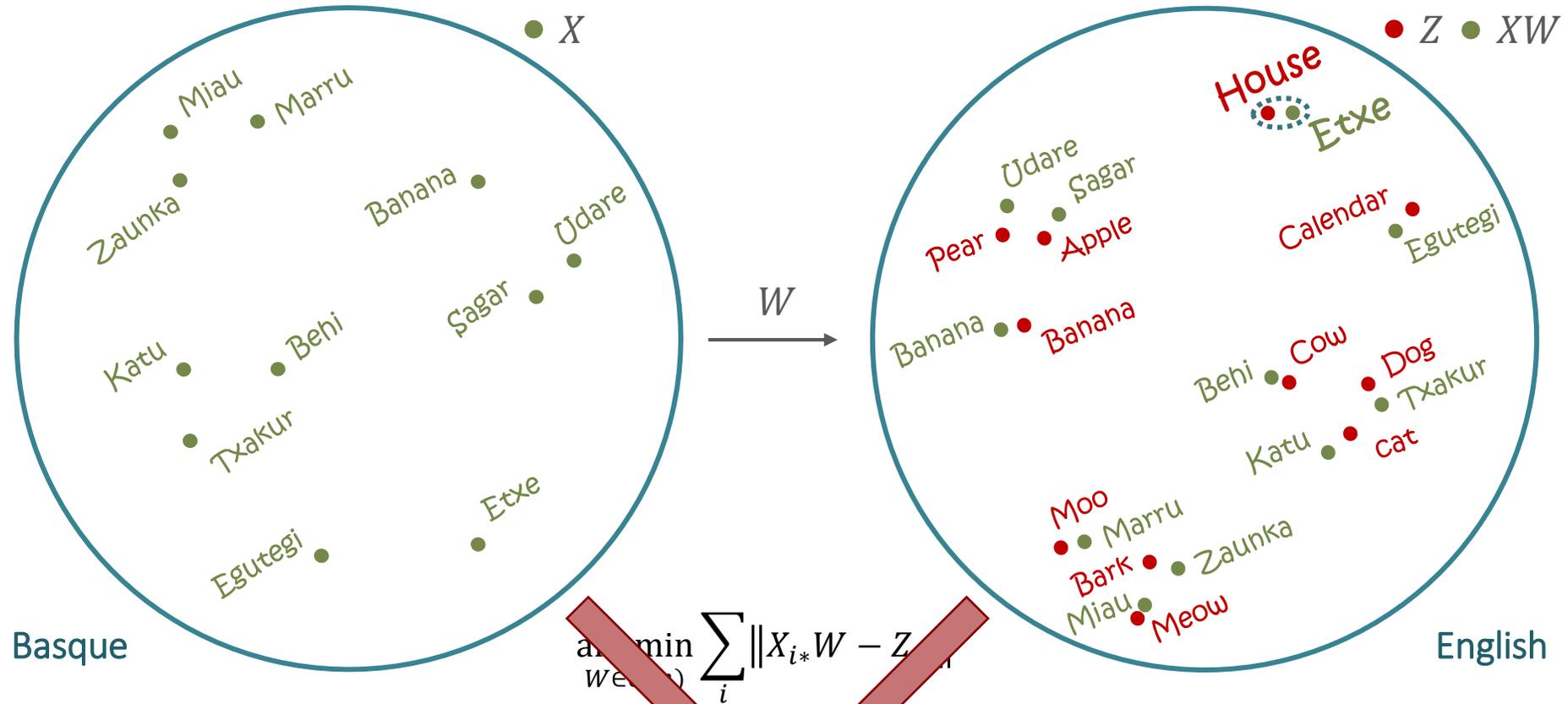
$$\begin{array}{l}
 \text{Txakur} \\
 \text{Sagar} \\
 \vdots \\
 \text{Egutegi}
 \end{array}
 \begin{bmatrix}
 X_{1,*} \\
 X_{2,*} \\
 \vdots \\
 X_{n,*}
 \end{bmatrix}
 [W] \approx
 \begin{bmatrix}
 Z_{1,*} \\
 Z_{2,*} \\
 \vdots \\
 Z_{n,*}
 \end{bmatrix}
 \begin{array}{l}
 \text{Dog} \\
 \text{Apple}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

Cross-lingual embedding mappings



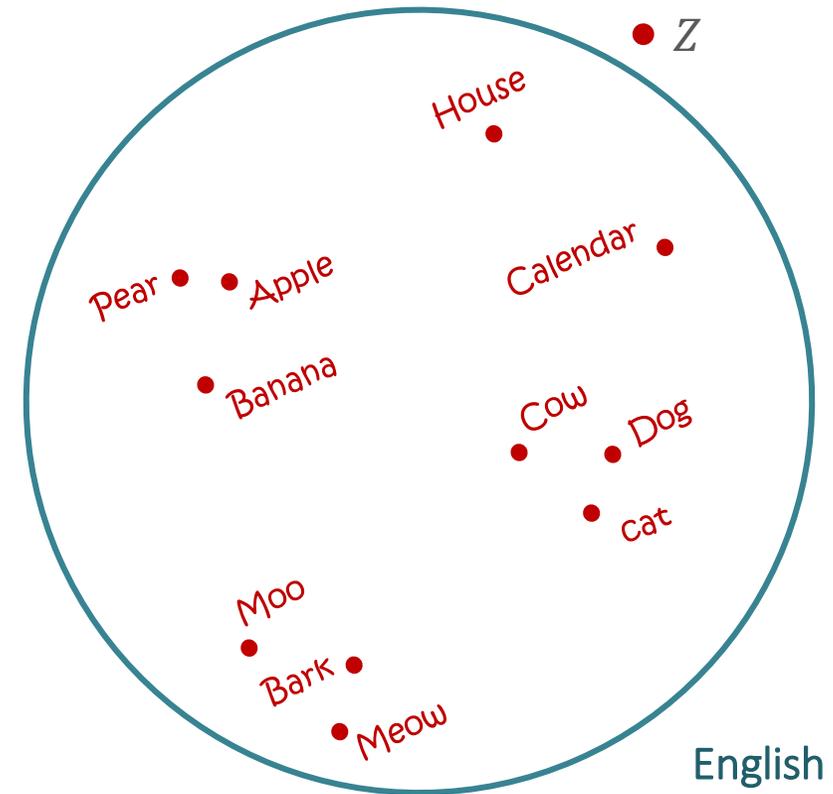
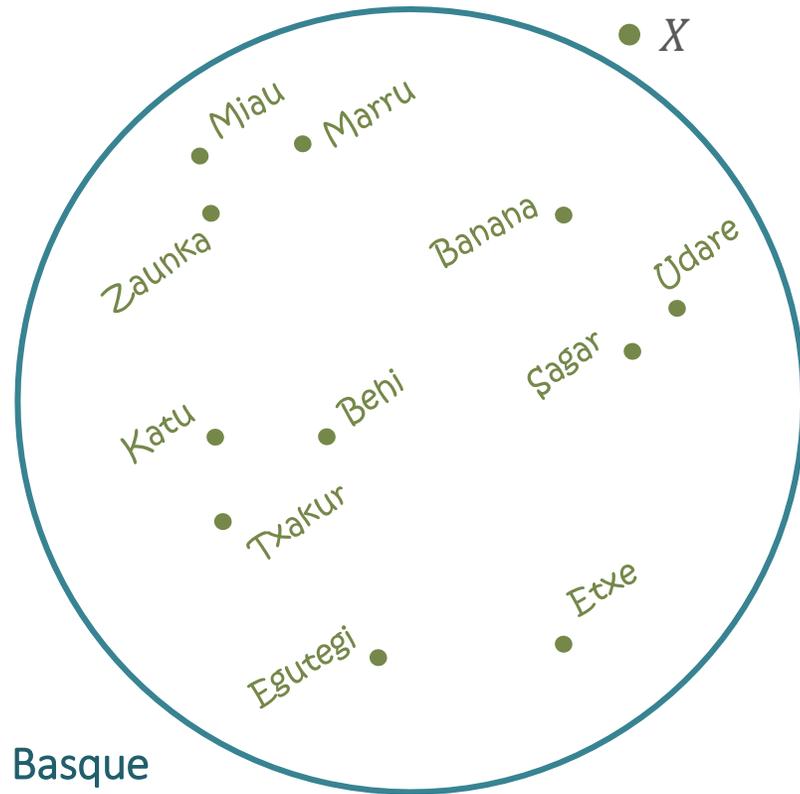
$$\begin{array}{l}
 \text{Txakur} \\
 \text{Sagar} \\
 \vdots \\
 \text{Egutegi}
 \end{array}
 \begin{bmatrix}
 X_{1,*} \\
 X_{2,*} \\
 \vdots \\
 X_{n,*}
 \end{bmatrix}
 [W] \approx
 \begin{bmatrix}
 Z_{1,*} \\
 Z_{2,*} \\
 \vdots \\
 Z_{n,*}
 \end{bmatrix}
 \begin{array}{l}
 \text{Dog} \\
 \text{Apple}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

Cross-lingual embedding mappings

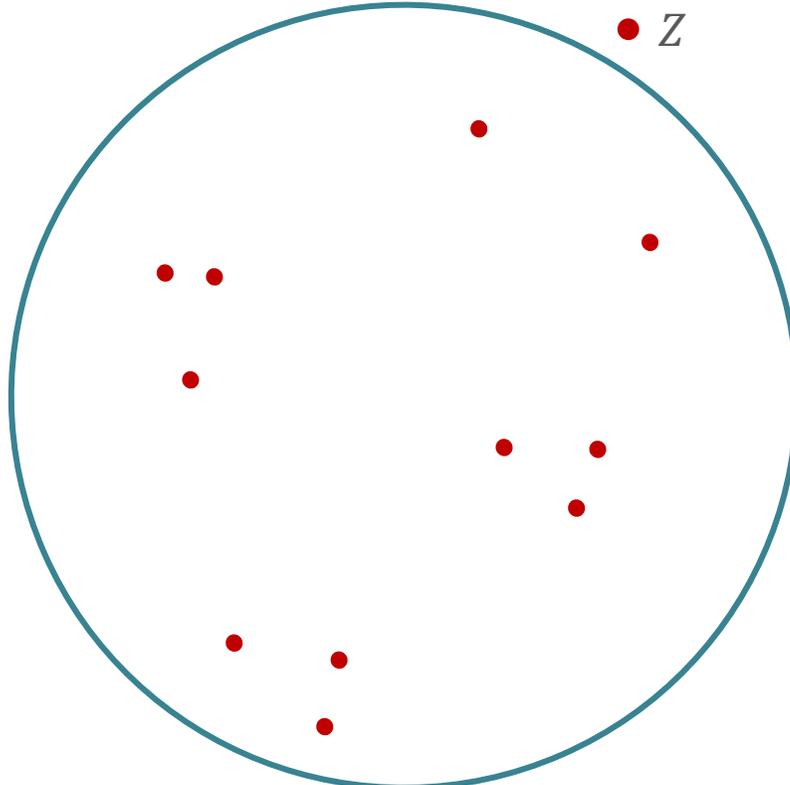
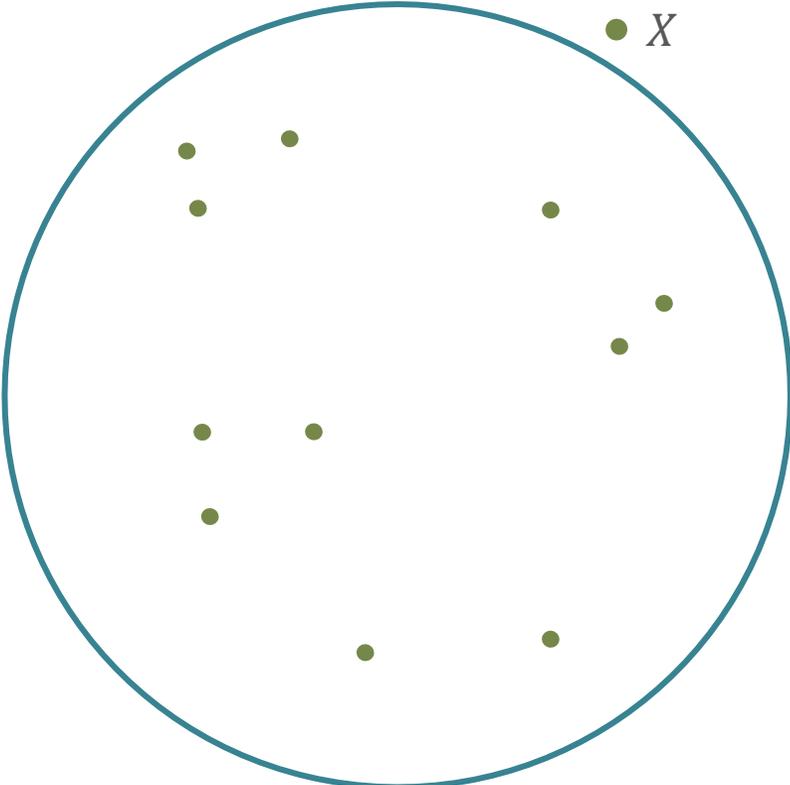


Txakur	$\begin{bmatrix} X_{1,*} \\ X_{2,*} \\ \vdots \\ X_{n,*} \end{bmatrix}$	\approx	$\begin{bmatrix} Z_{1,*} \\ Z_{2,*} \\ \vdots \\ Z_{n,*} \end{bmatrix}$	Dog
Sagar				Apple
⋮				⋮
Egutegi				Calendar

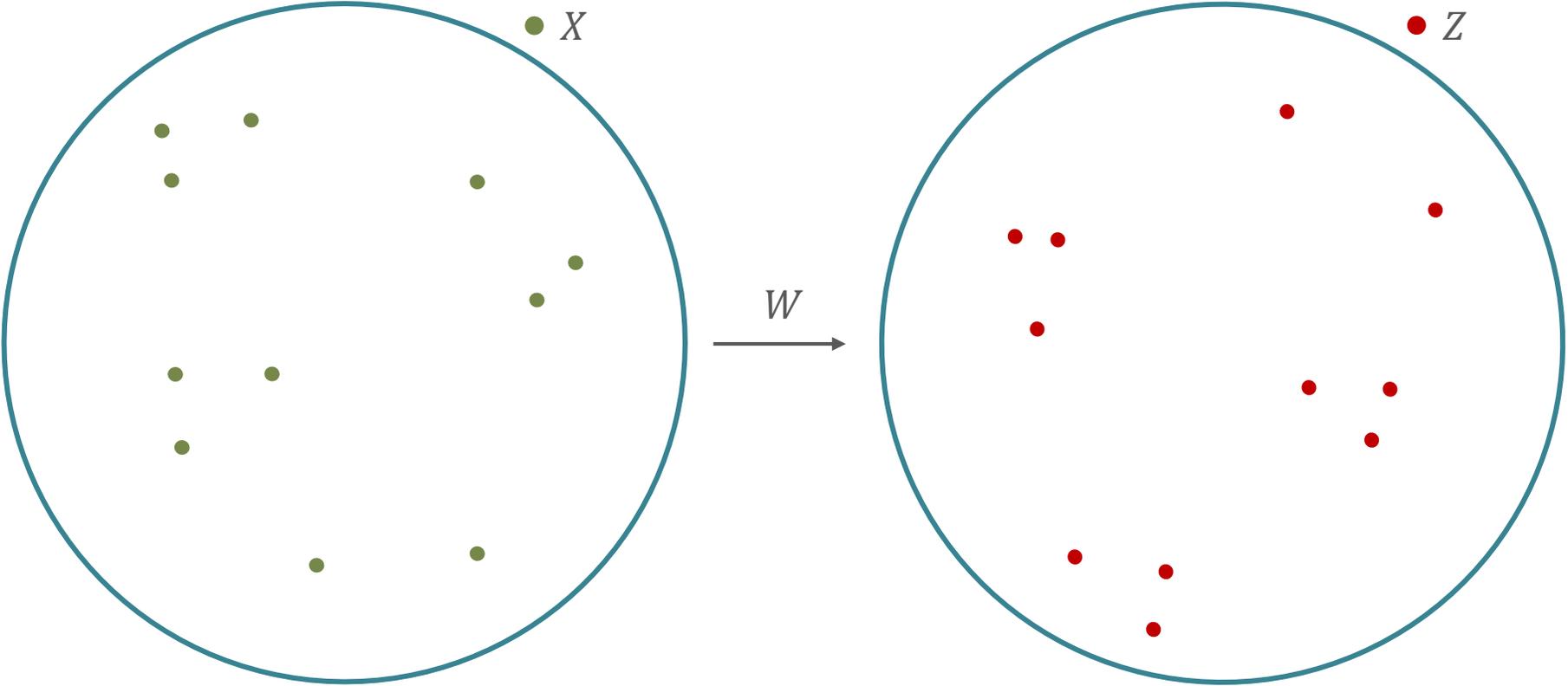
Cross-lingual embedding mappings



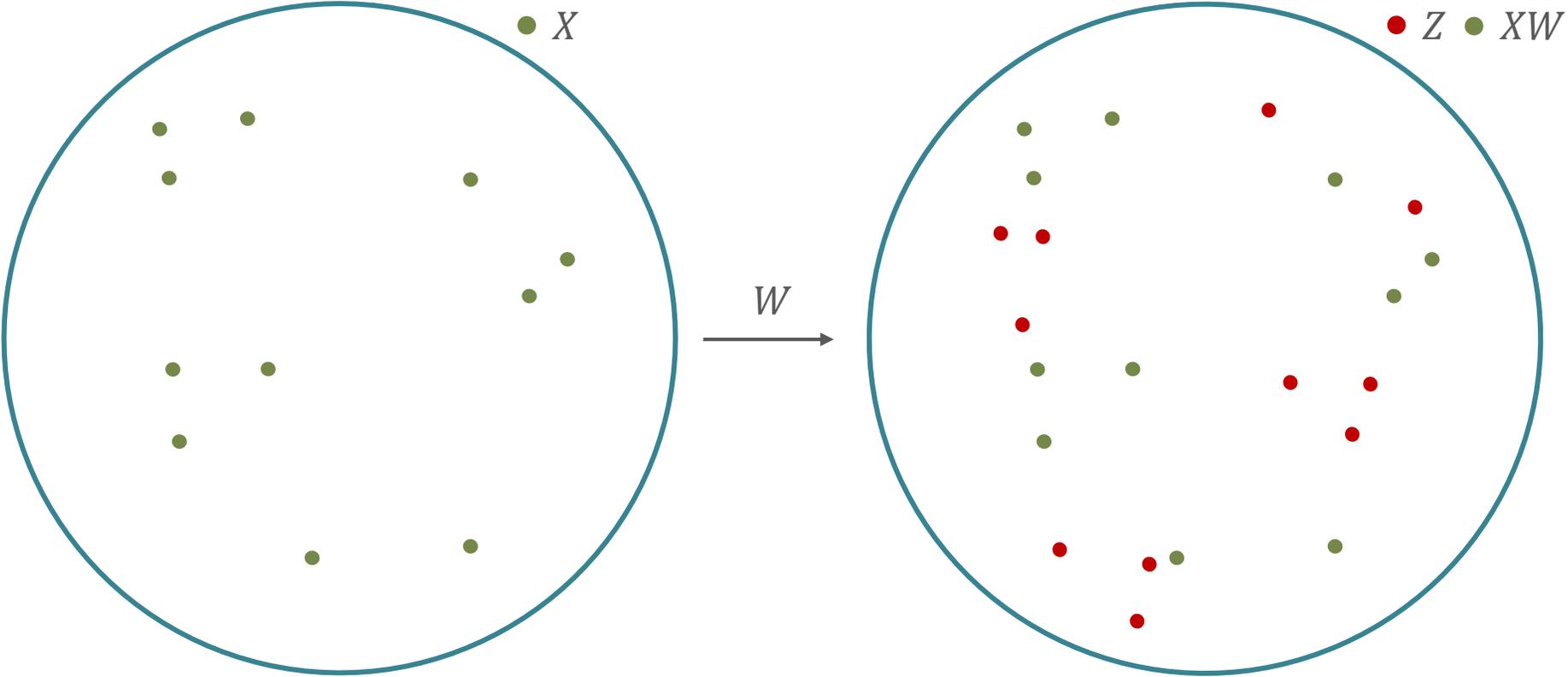
Cross-lingual embedding mappings



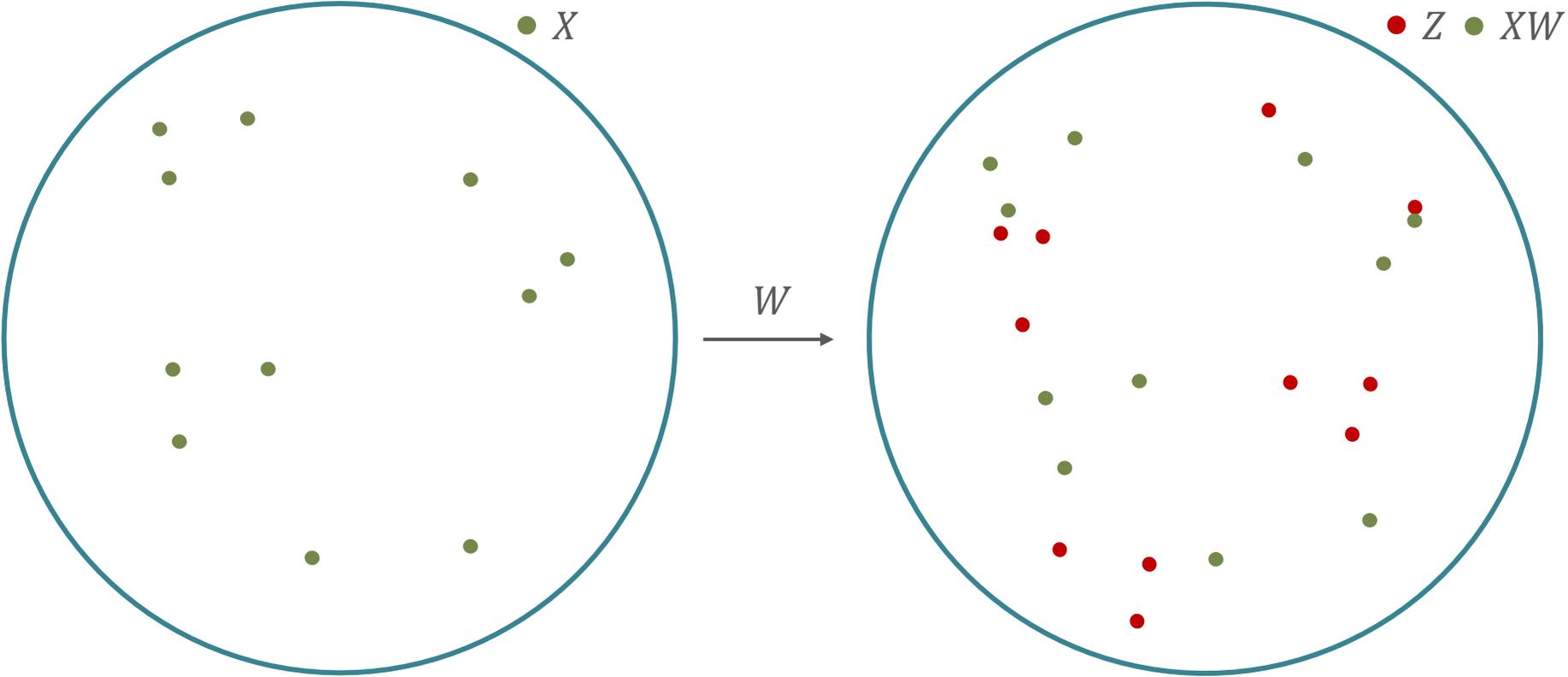
Cross-lingual embedding mappings



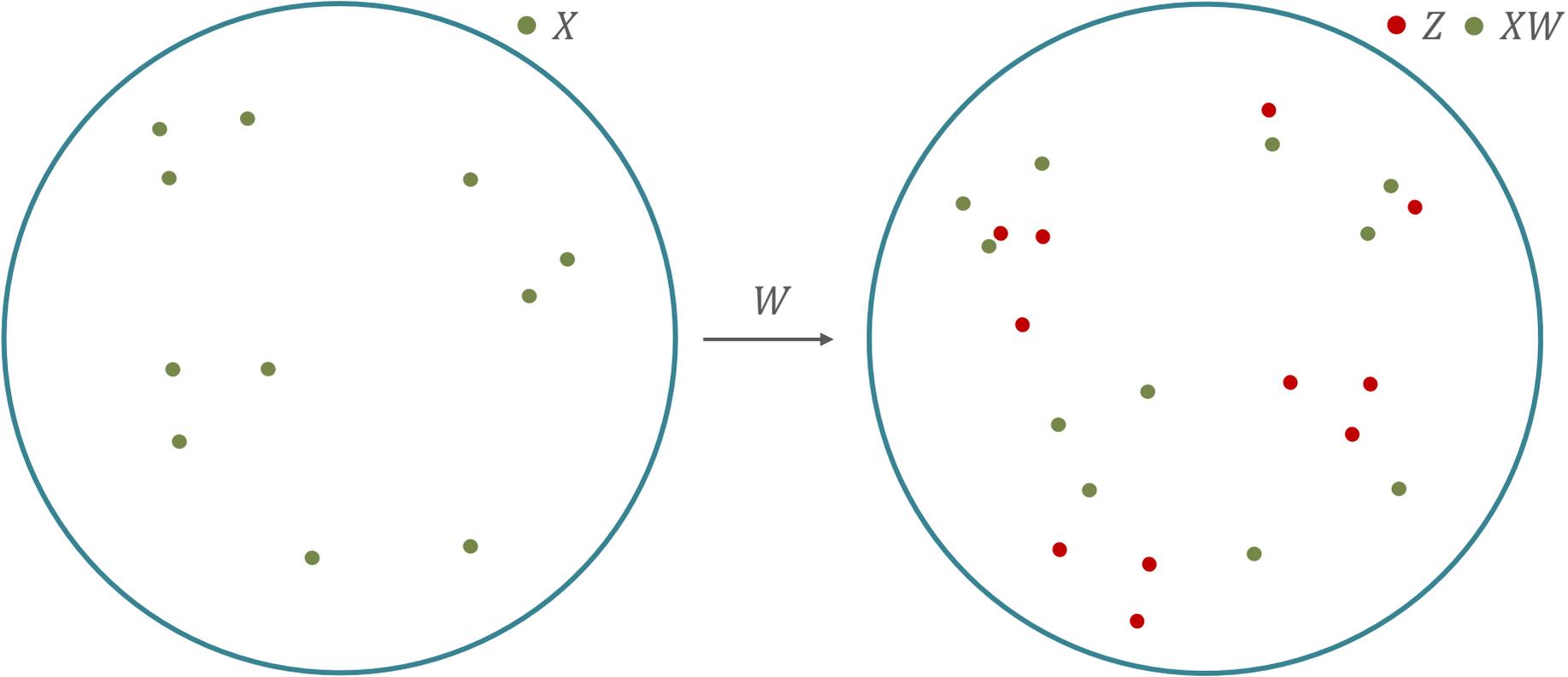
Cross-lingual embedding mappings



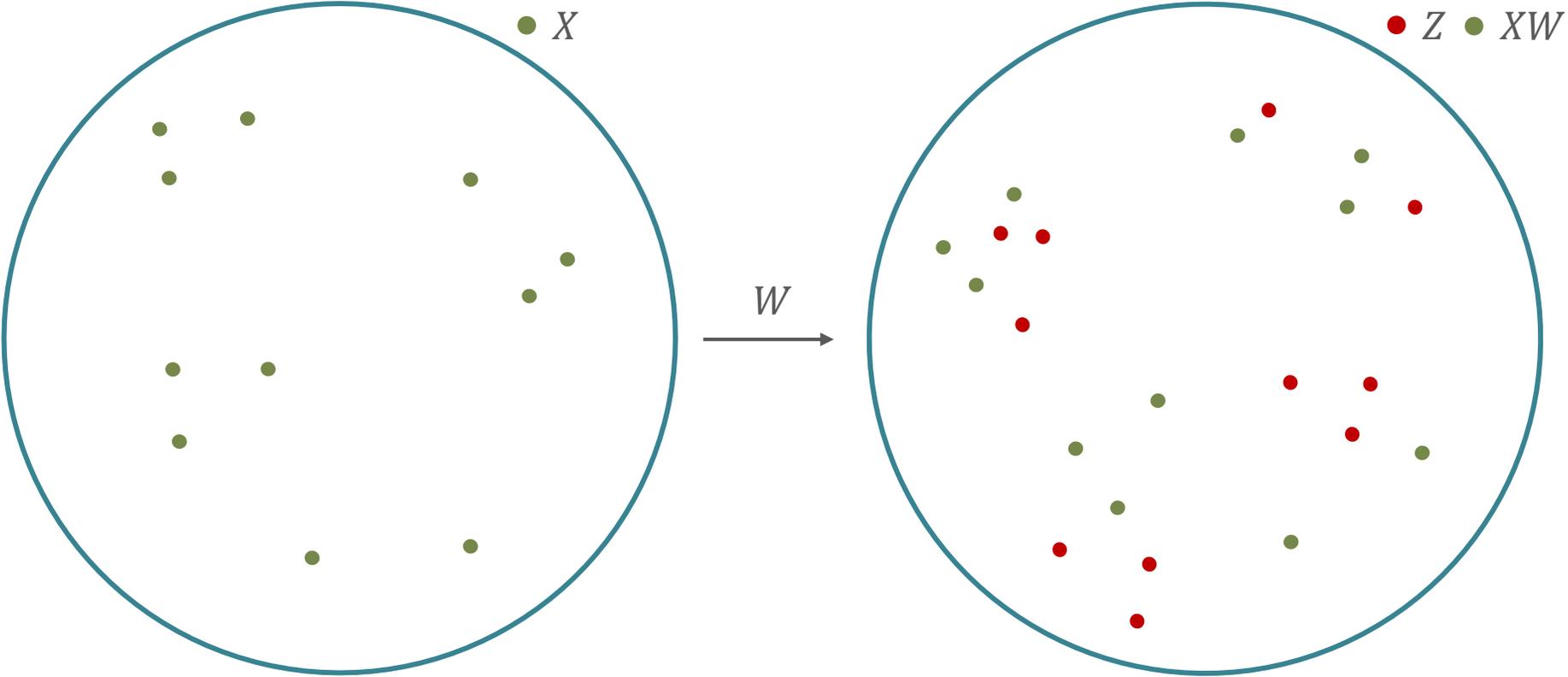
Cross-lingual embedding mappings



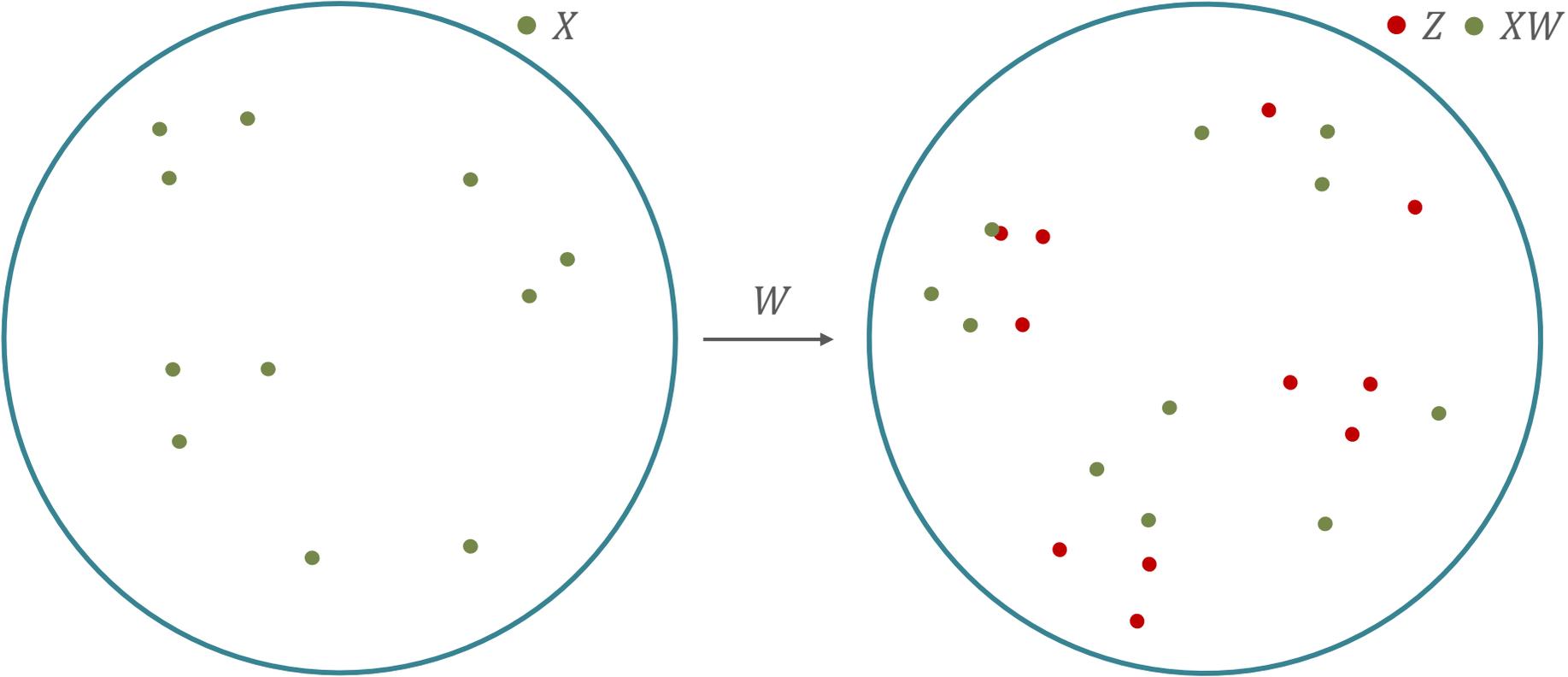
Cross-lingual embedding mappings



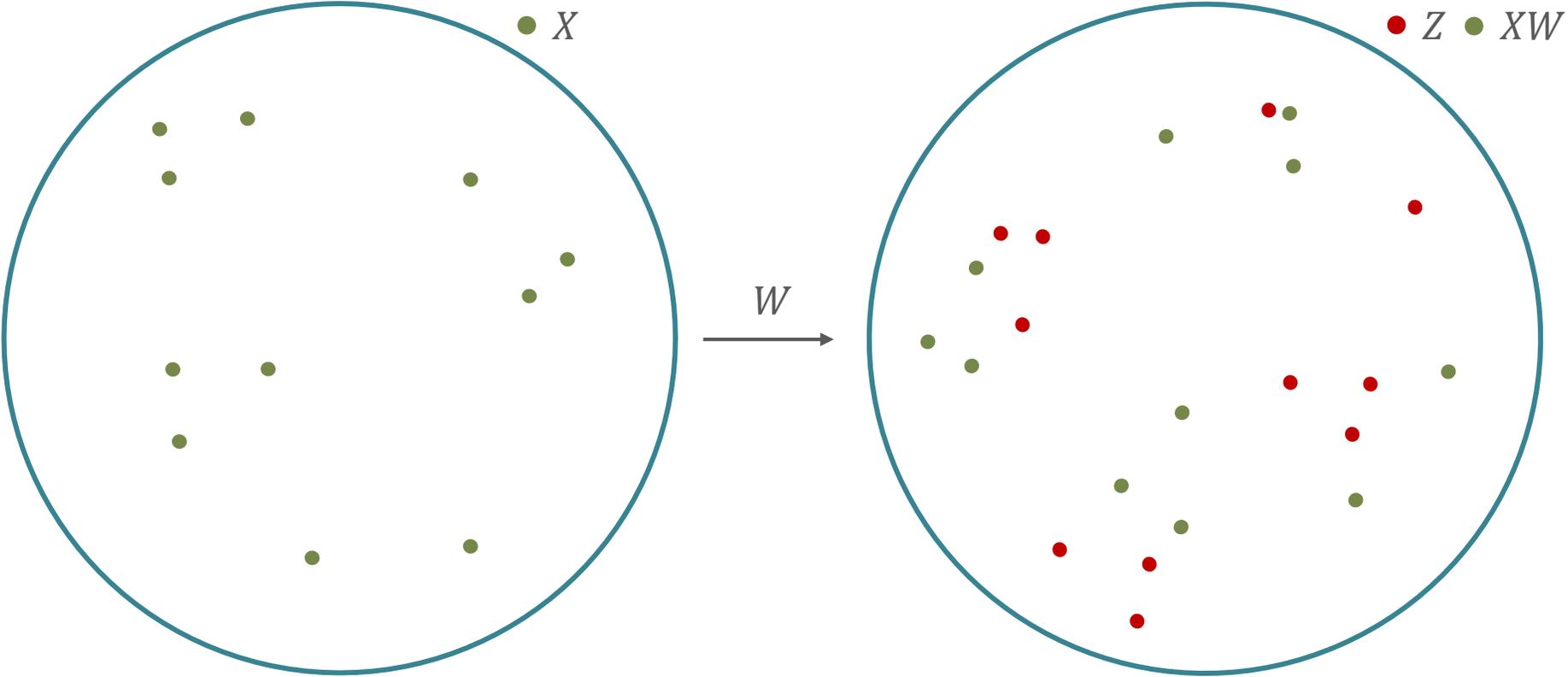
Cross-lingual embedding mappings



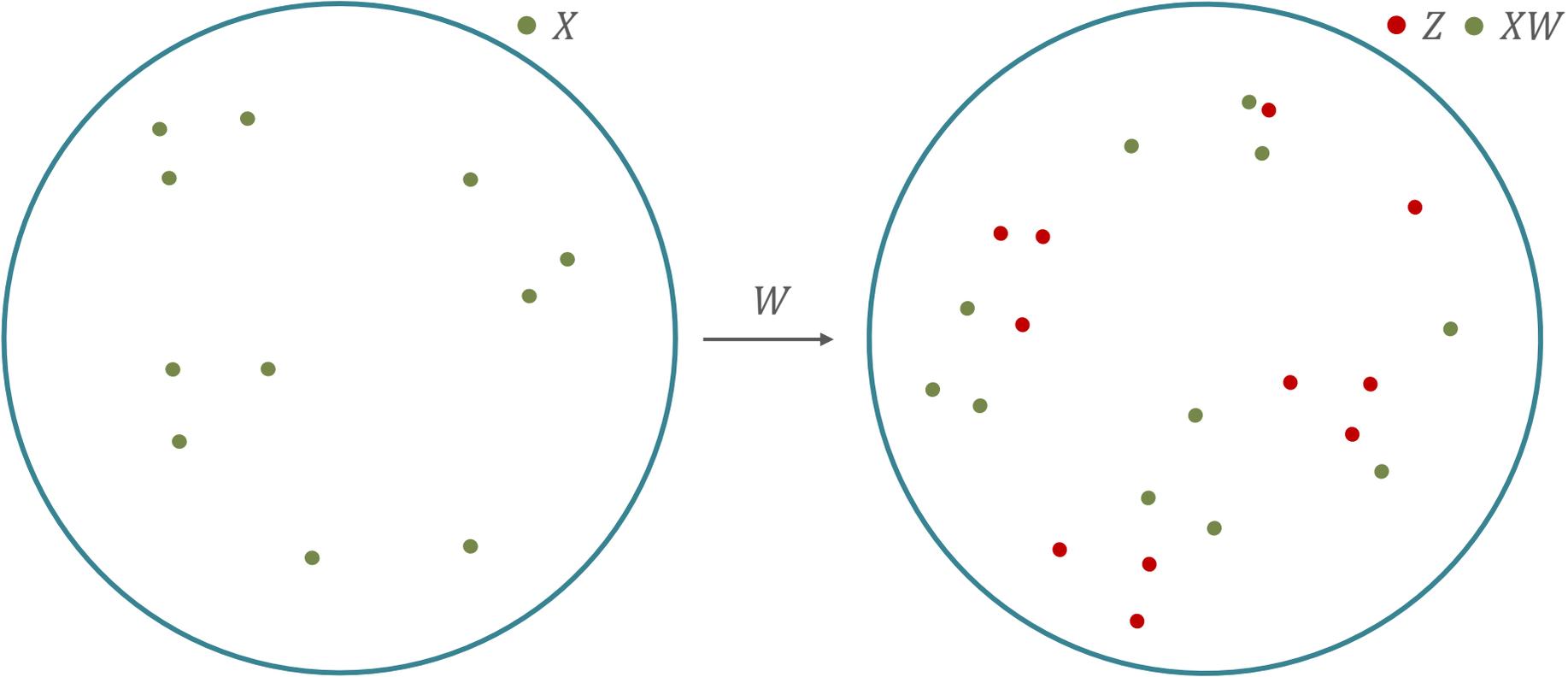
Cross-lingual embedding mappings



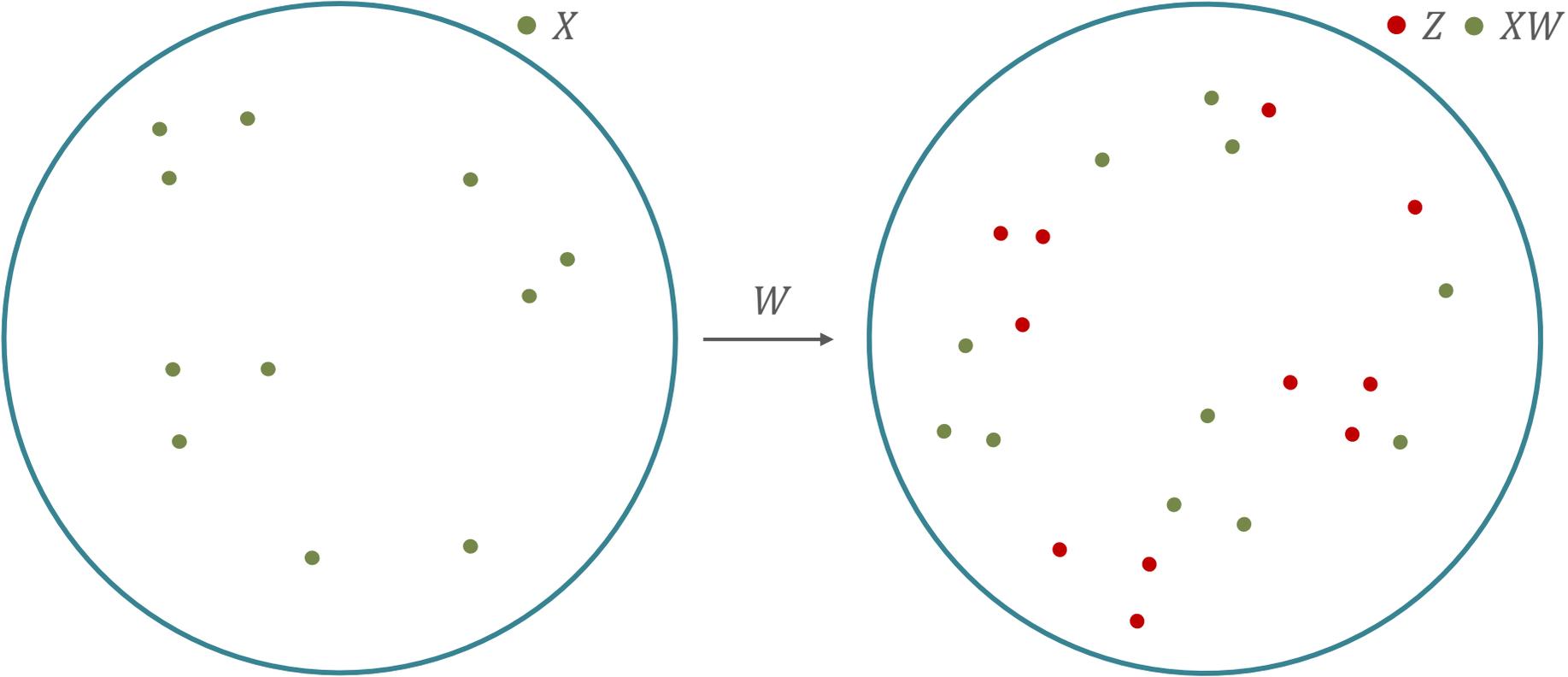
Cross-lingual embedding mappings



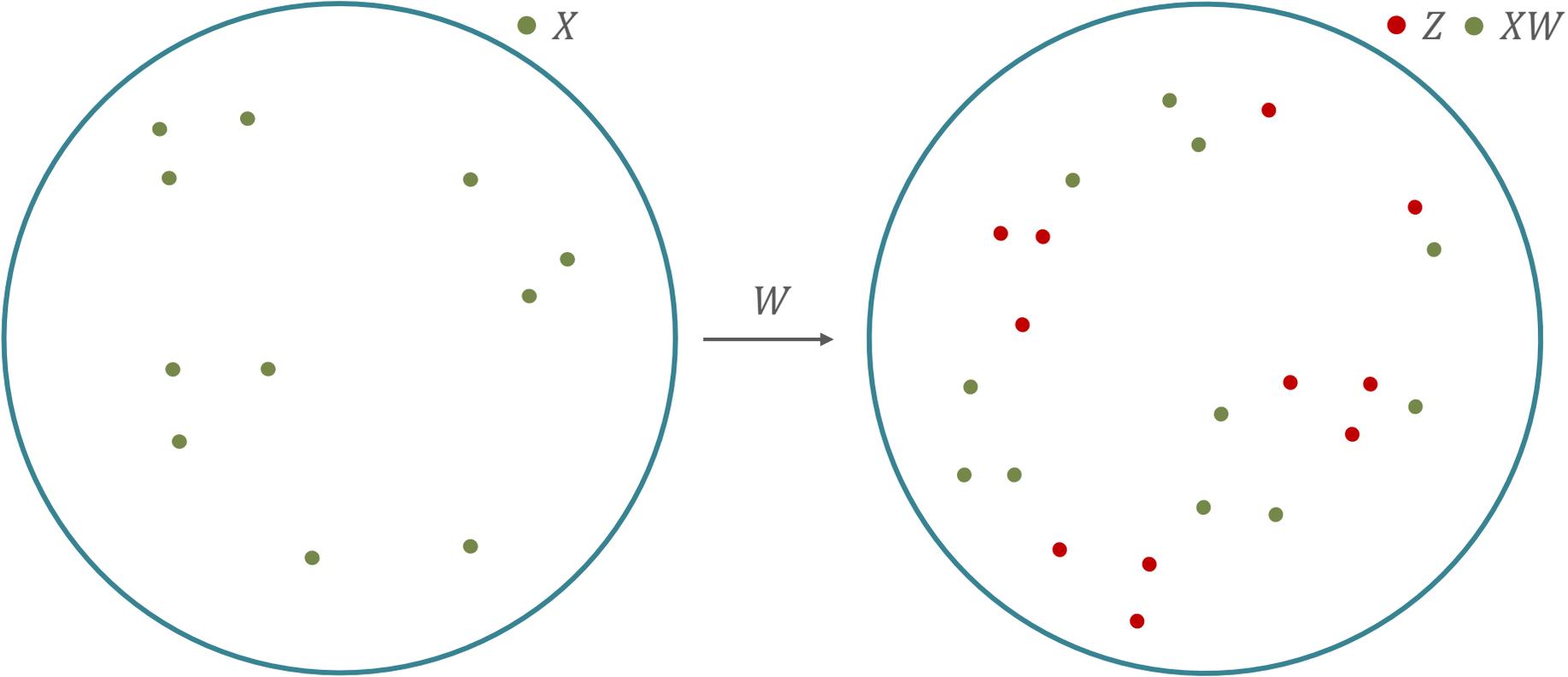
Cross-lingual embedding mappings



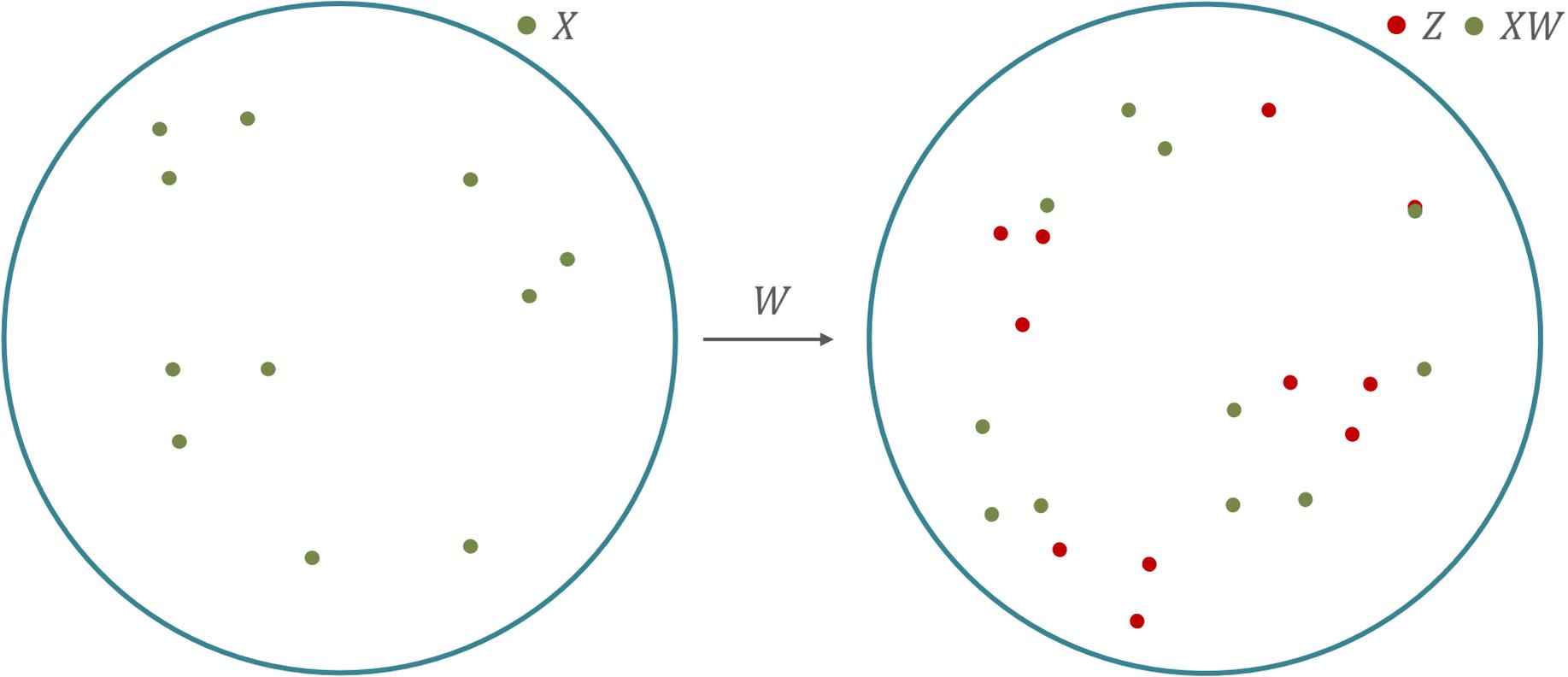
Cross-lingual embedding mappings



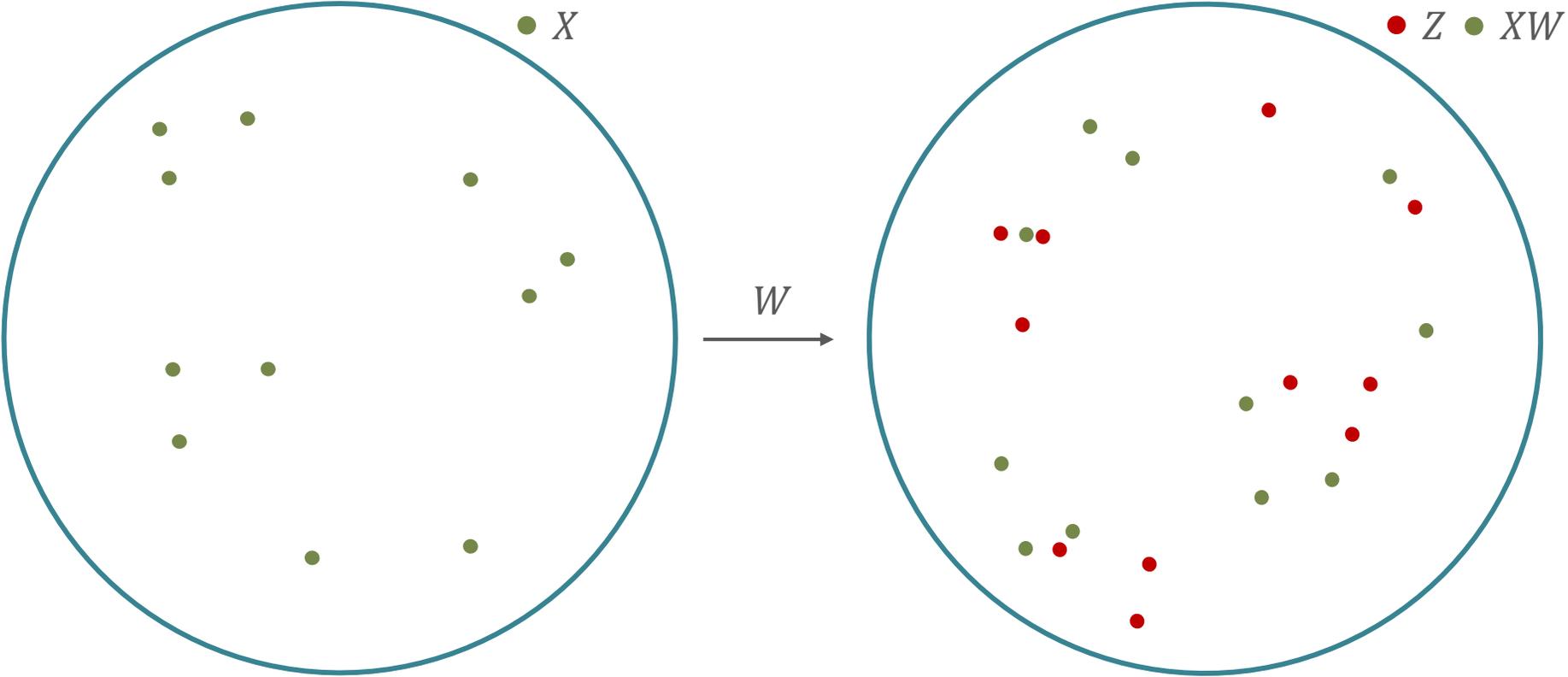
Cross-lingual embedding mappings



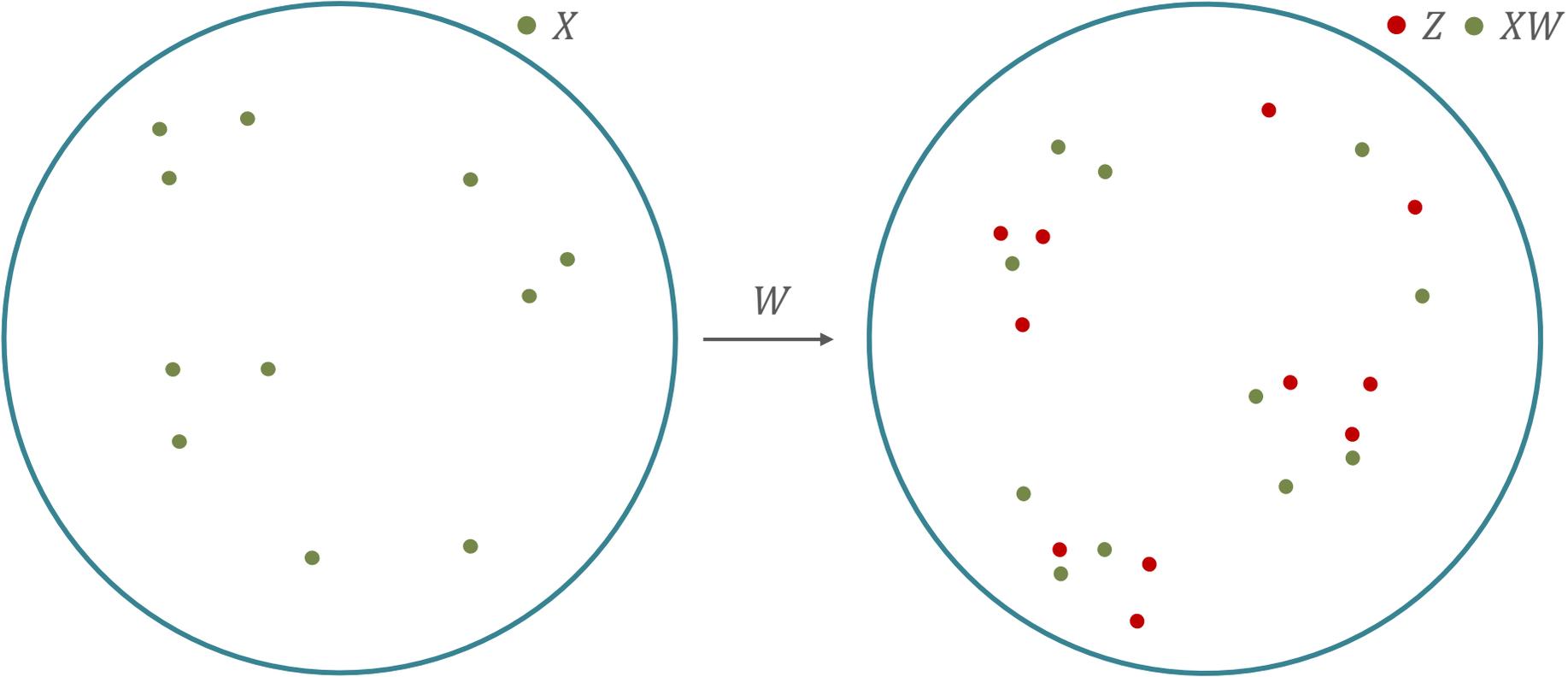
Cross-lingual embedding mappings



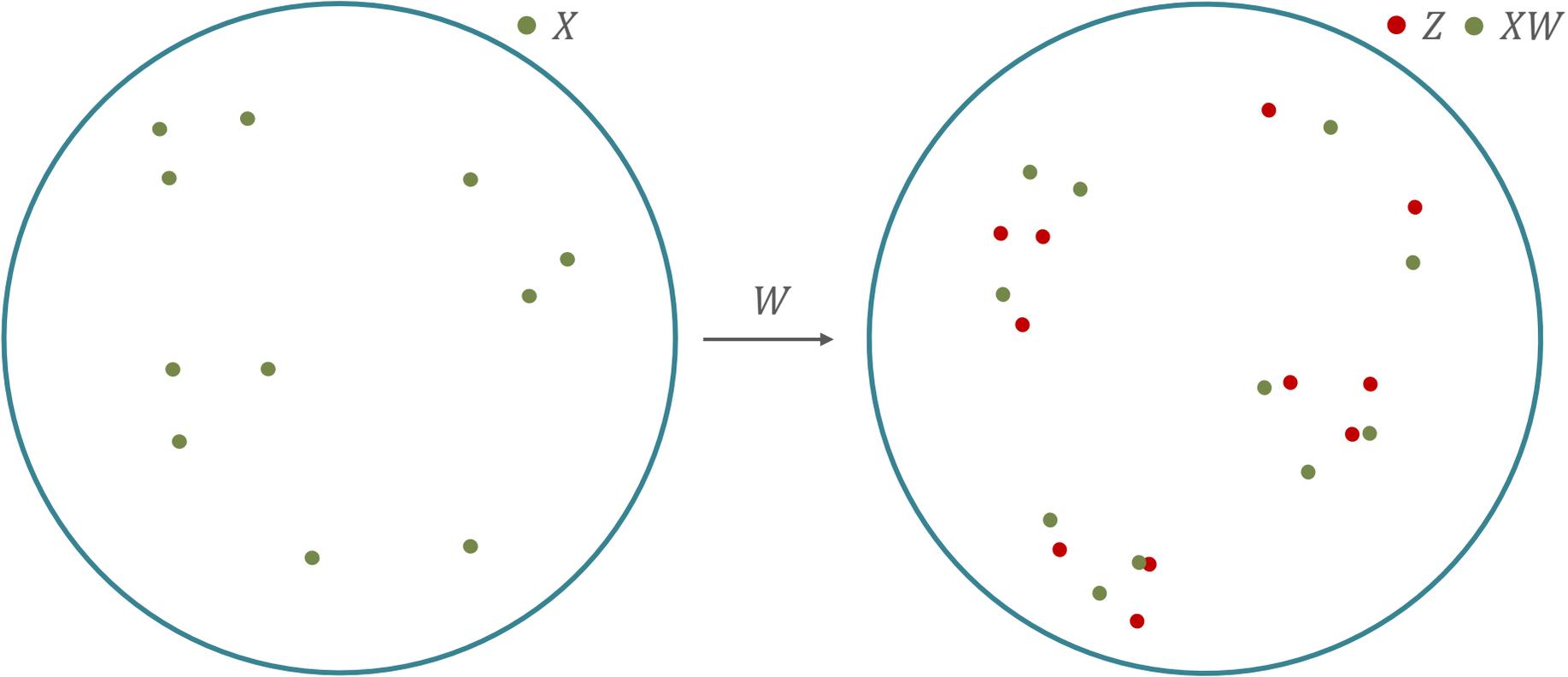
Cross-lingual embedding mappings



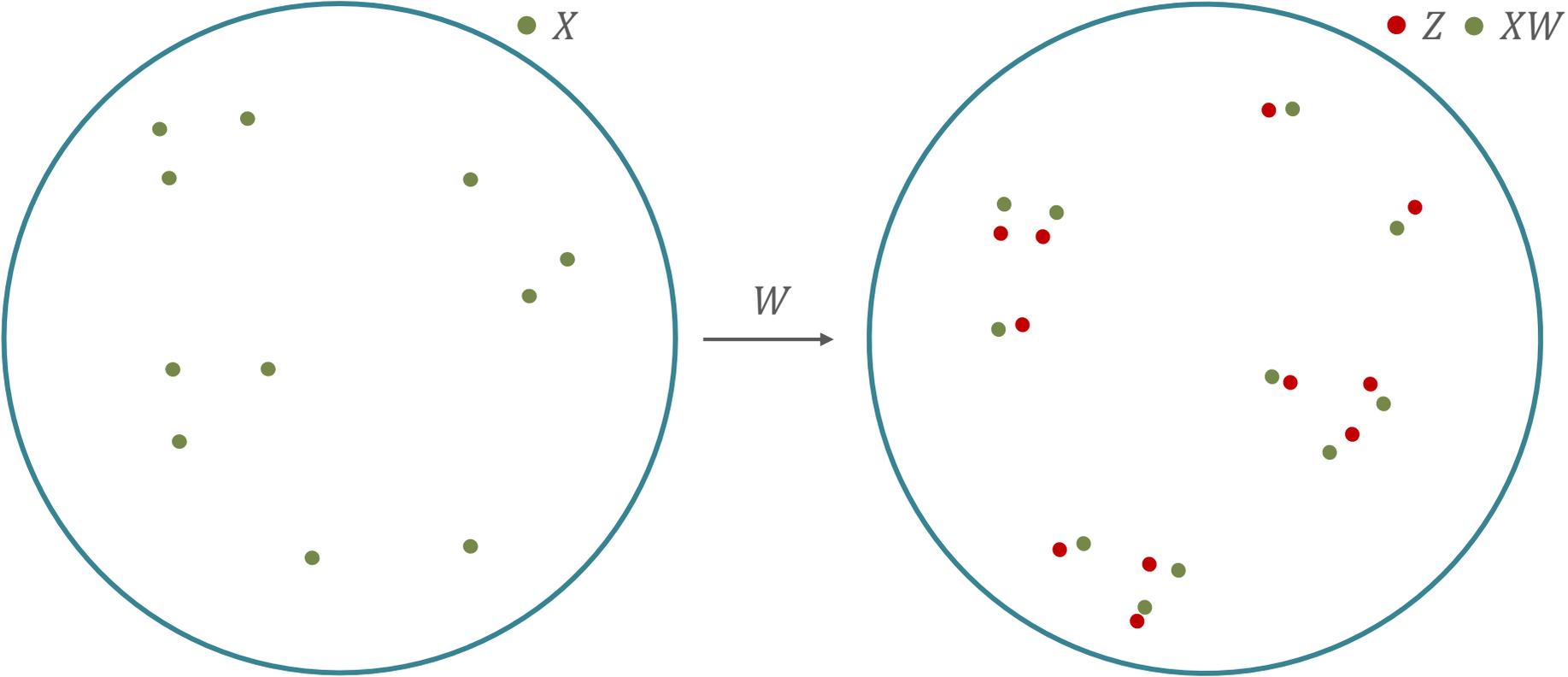
Cross-lingual embedding mappings



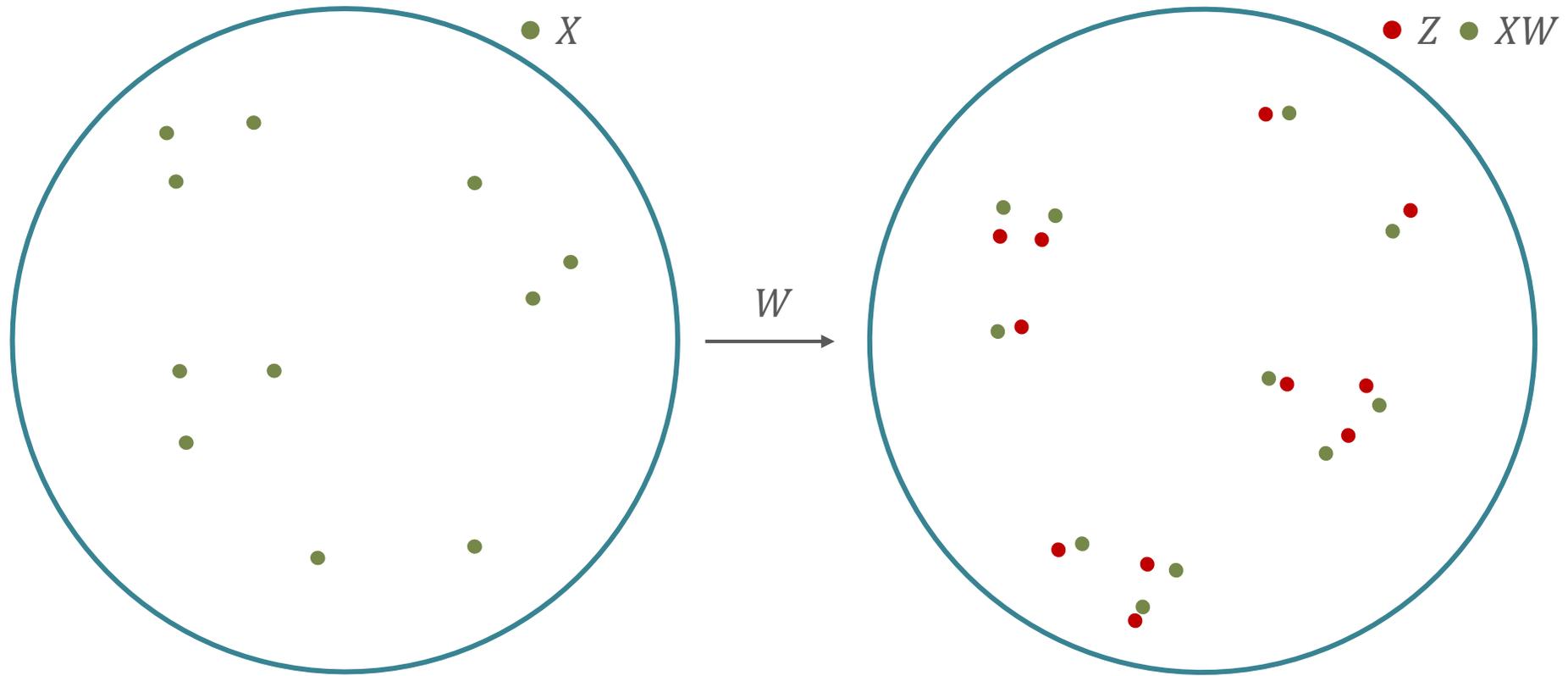
Cross-lingual embedding mappings



Cross-lingual embedding mappings



Cross-lingual embedding mappings



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

Cross-lingual embedding mappings

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

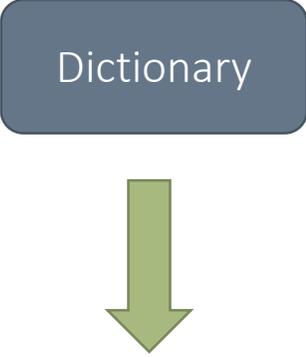
Cross-lingual embedding mappings

Dictionary

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

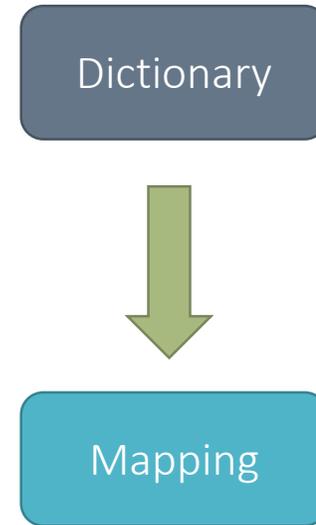
Cross-lingual embedding mappings

Dictionary



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

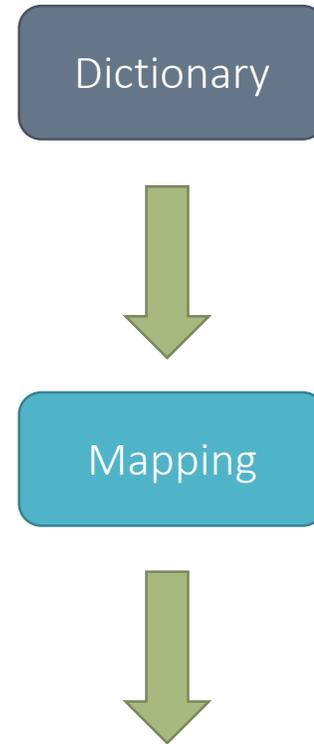
Cross-lingual embedding mappings



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

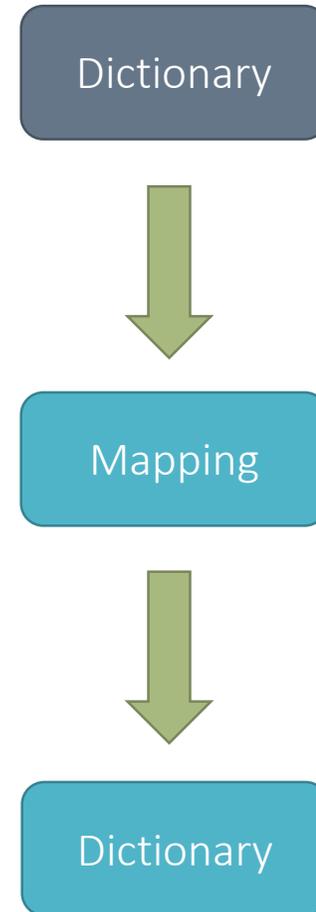
Cross-lingual embedding mappings

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



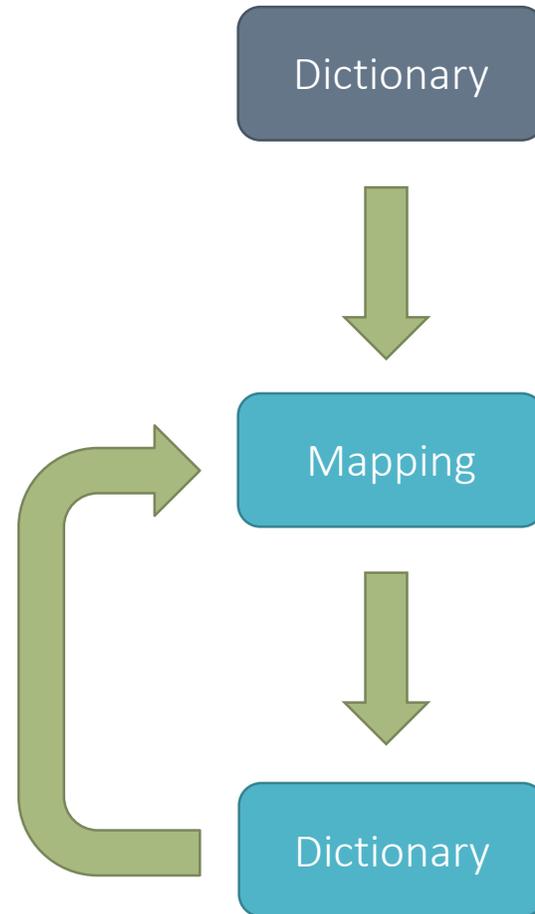
Cross-lingual embedding mappings

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



Cross-lingual embedding mappings

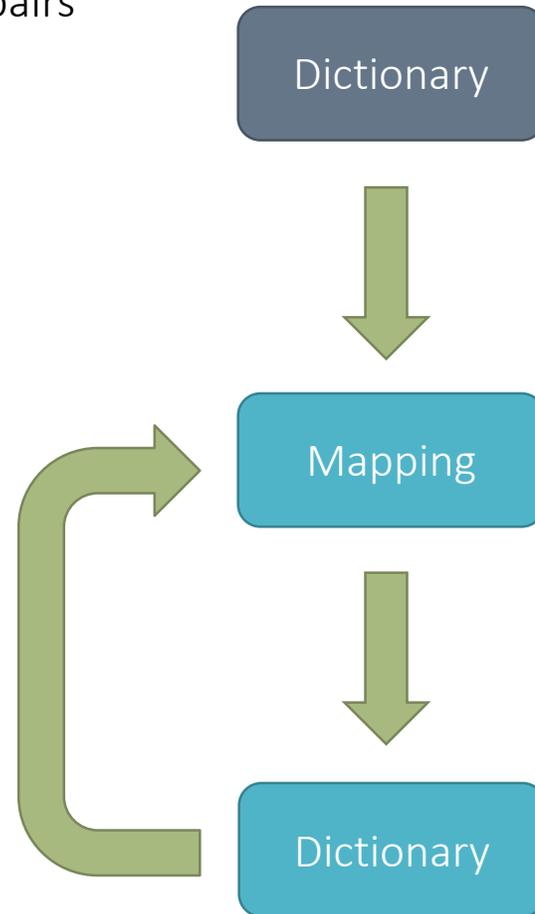
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



Cross-lingual embedding mappings

- 25 word pairs

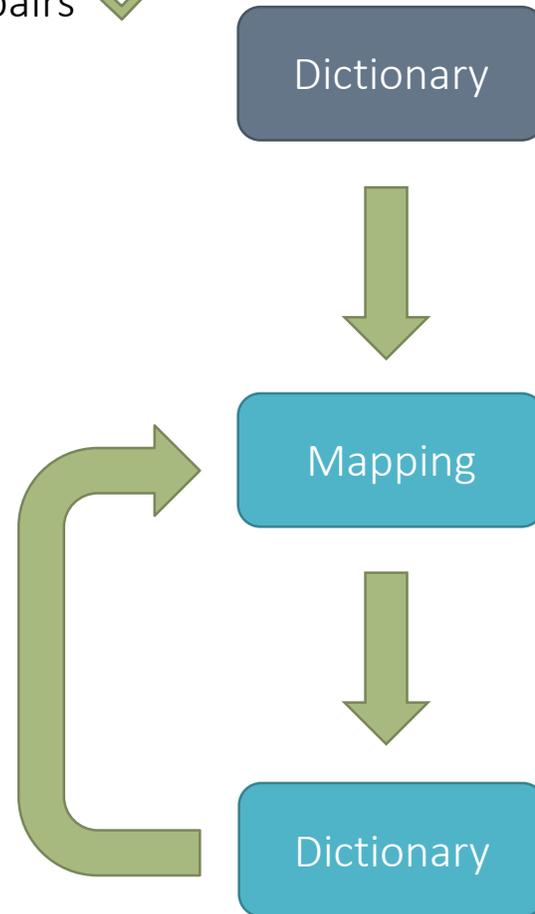
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



Cross-lingual embedding mappings

- 25 word pairs ✓

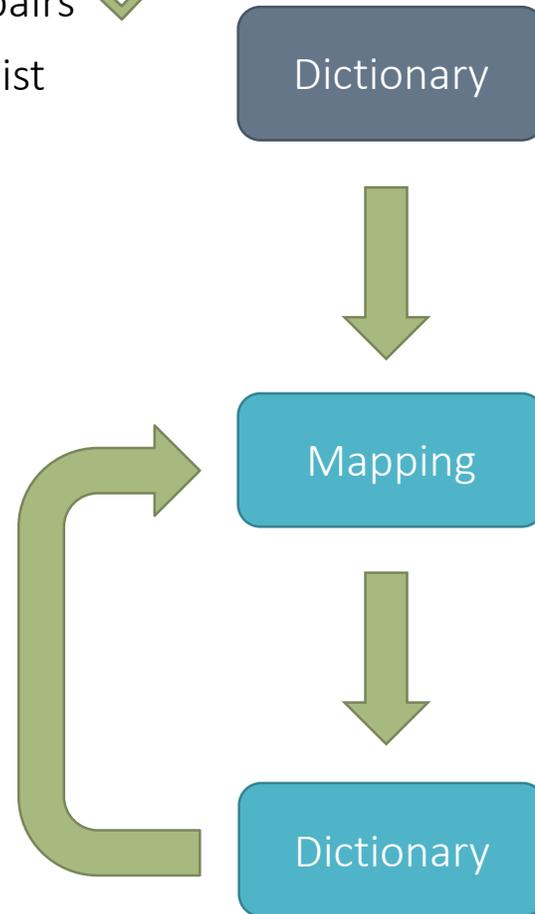
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



Cross-lingual embedding mappings

- 25 word pairs ✓
- Numeral list

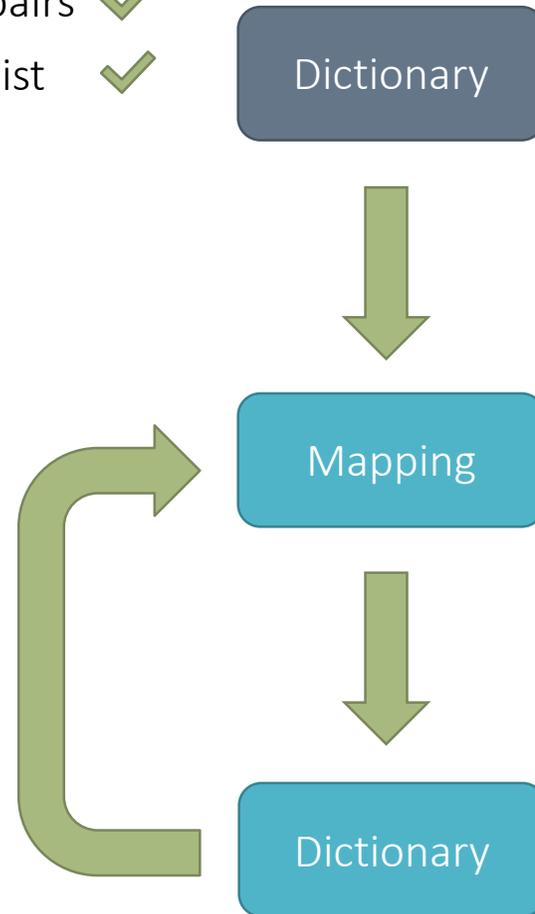
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i*} W - Z_{j*}\|^2$$



Cross-lingual embedding mappings

- 25 word pairs ✓
- Numeral list ✓

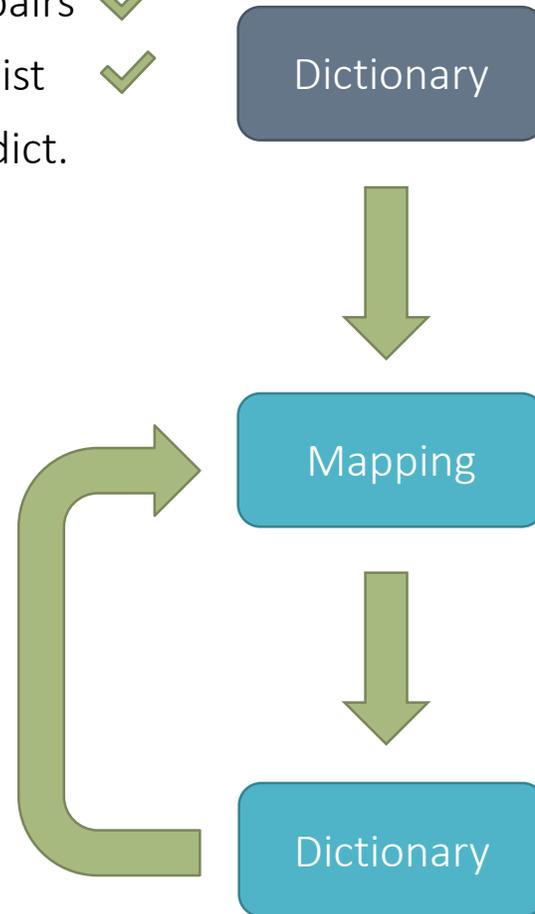
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i*} W - Z_{j*}\|^2$$



Cross-lingual embedding mappings

- 25 word pairs ✓
- Numeral list ✓
- Random dict.

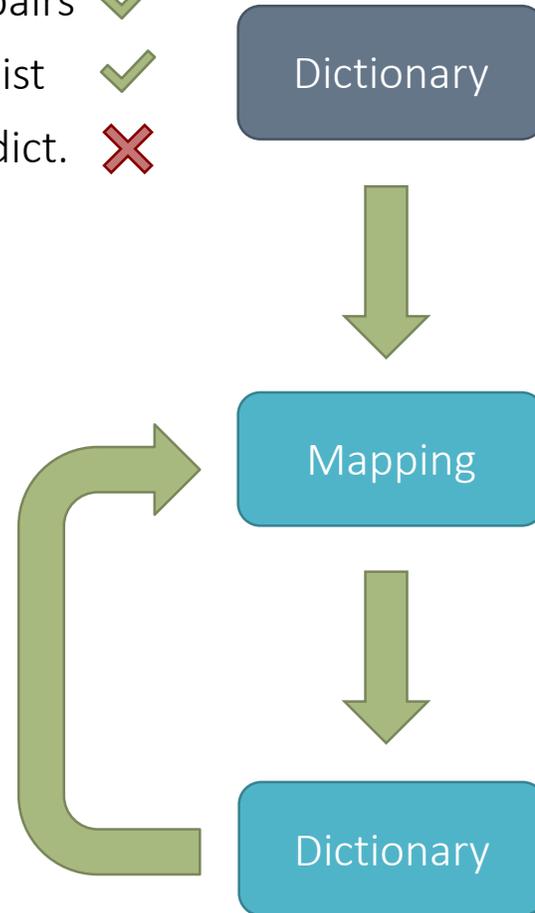
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



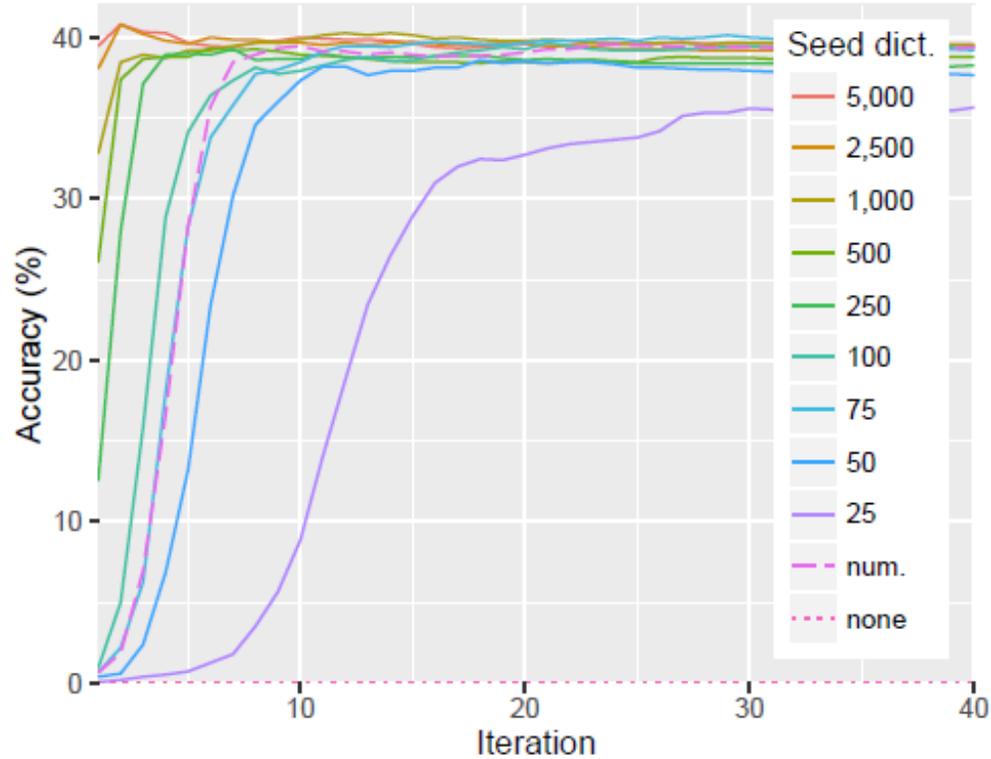
Cross-lingual embedding mappings

- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

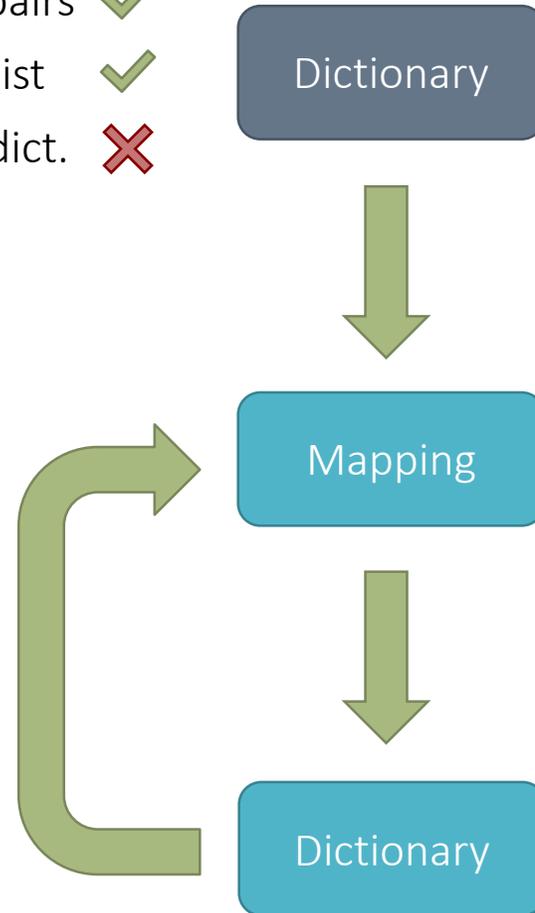
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



Cross-lingual embedding mappings

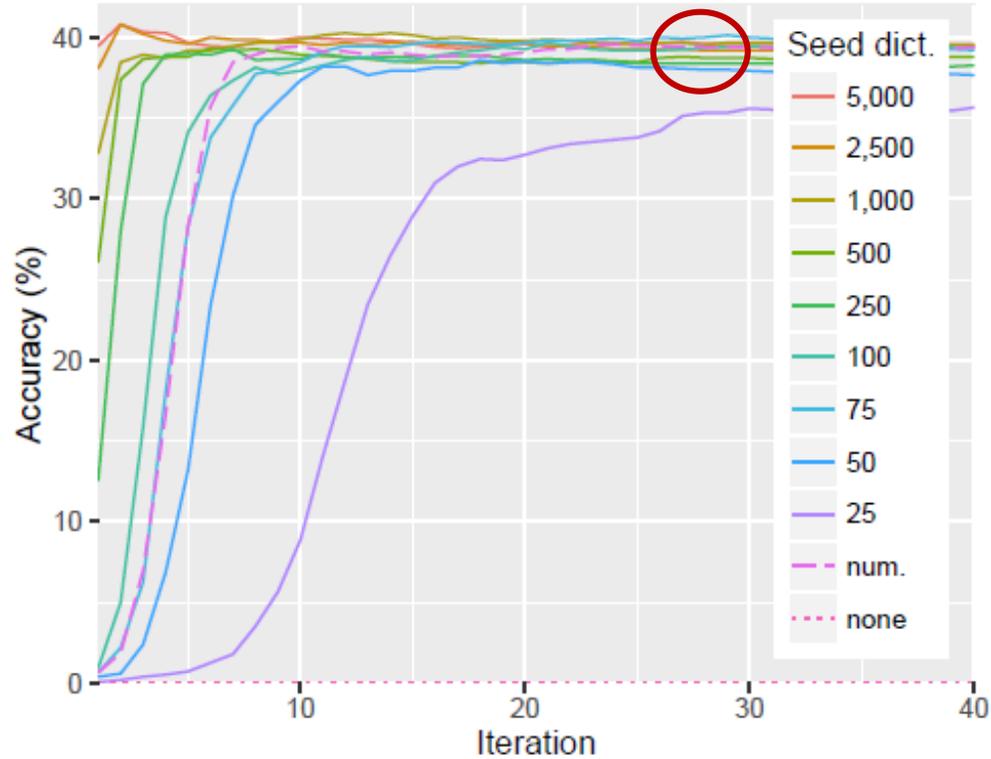


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

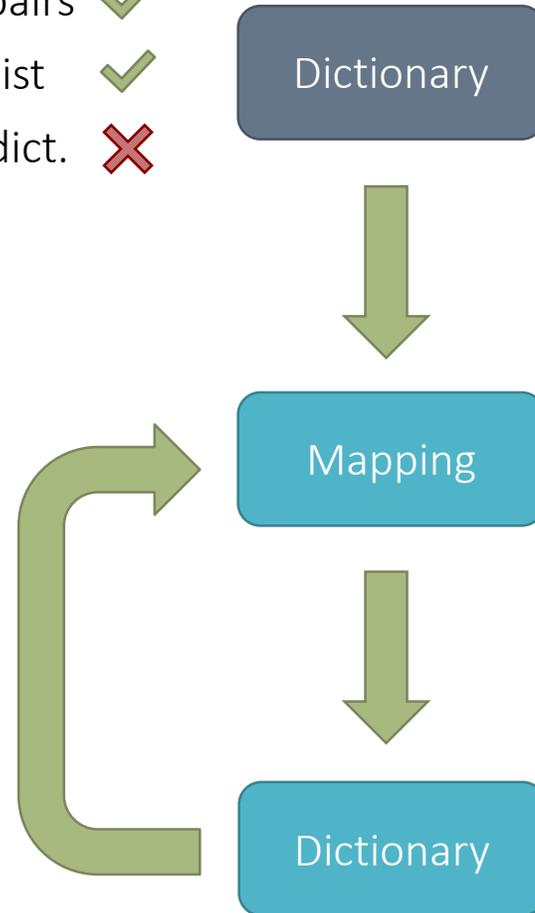


$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$

Cross-lingual embedding mappings

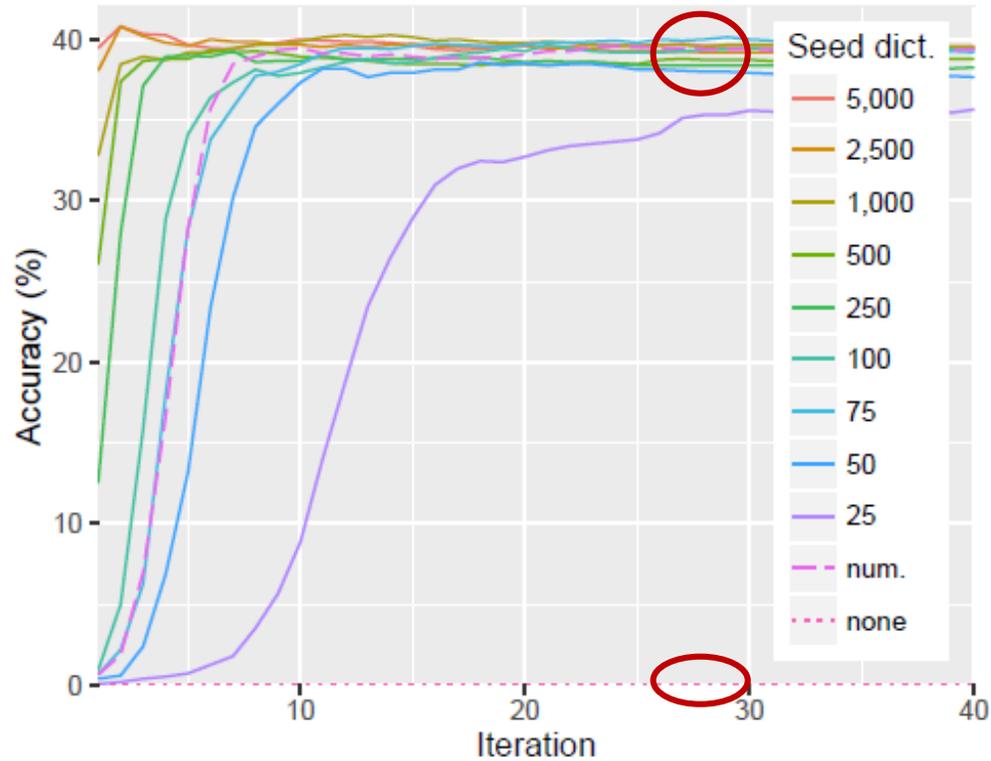


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

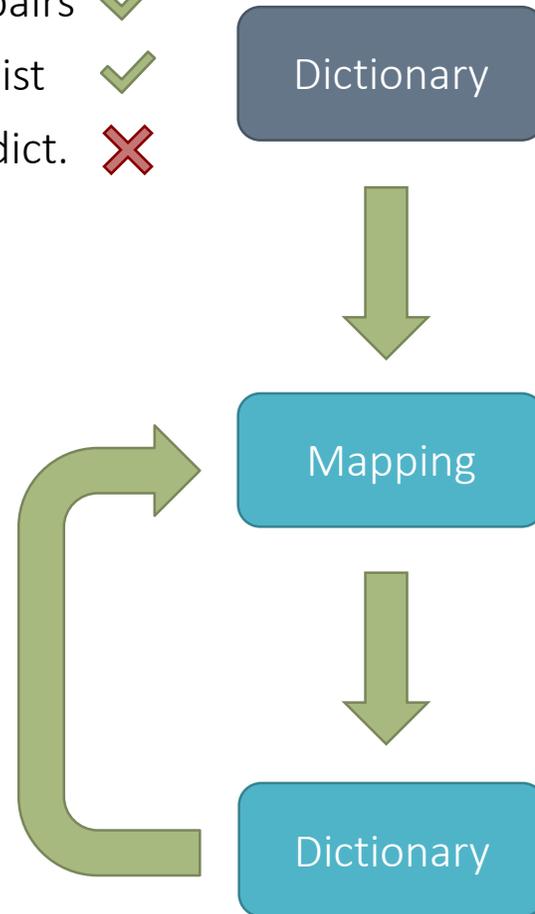


$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$

Cross-lingual embedding mappings

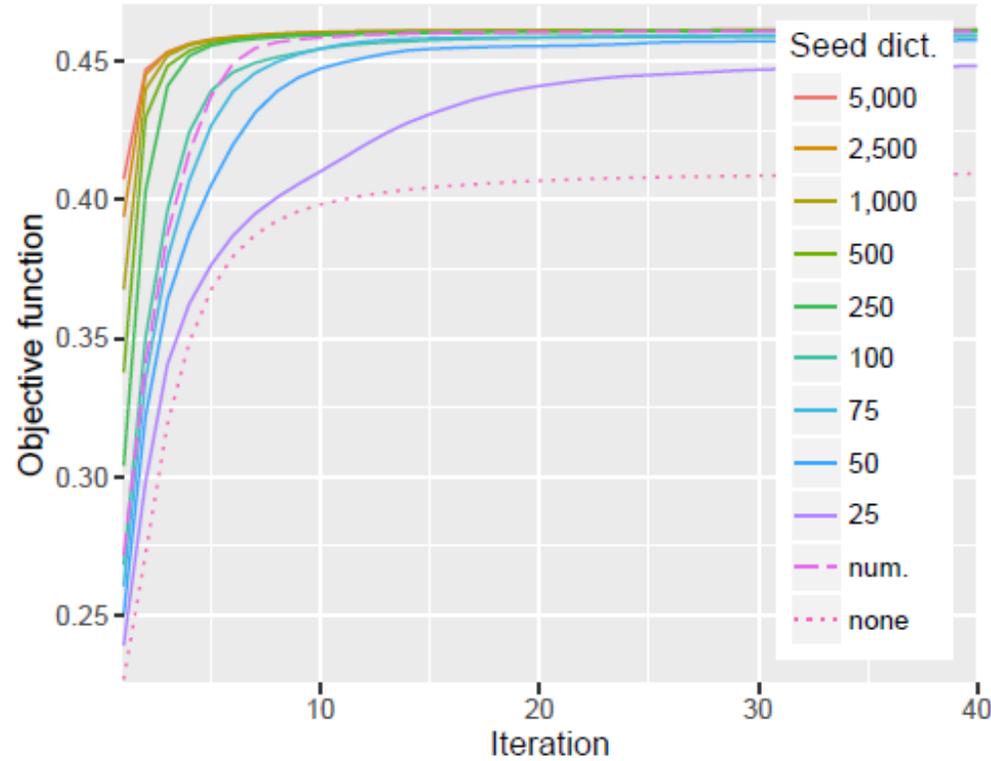


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗



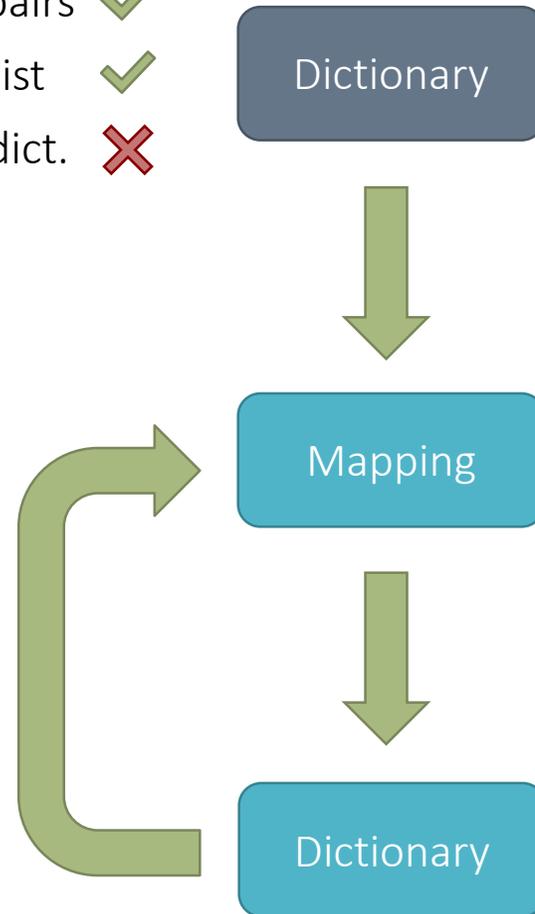
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$

Cross-lingual embedding mappings

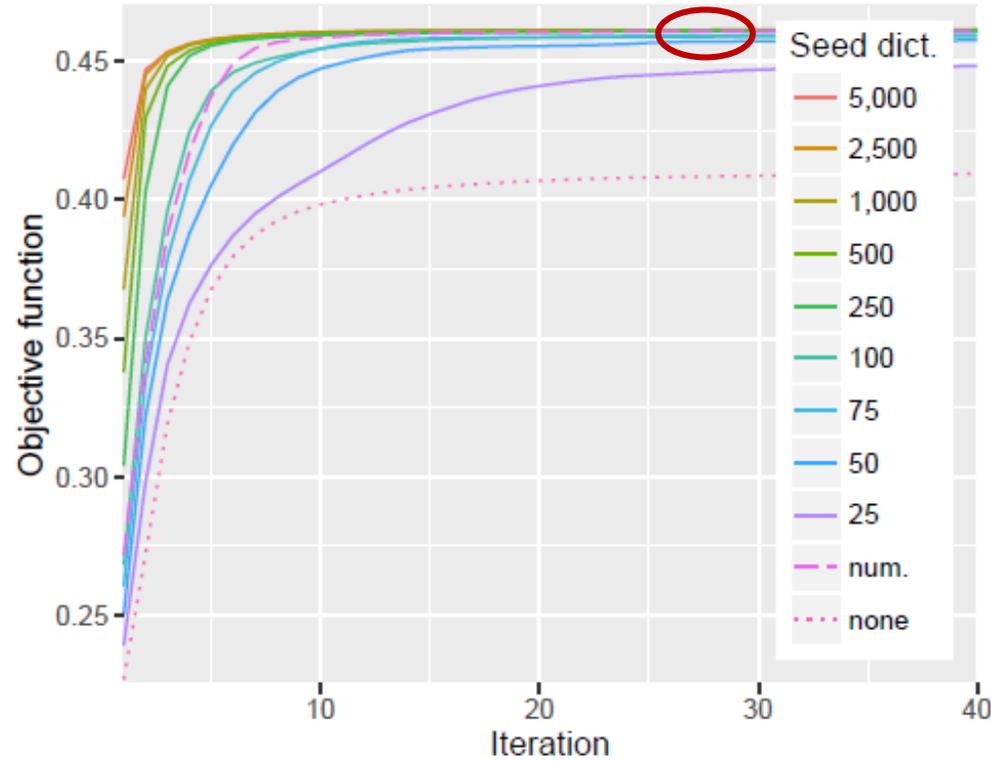


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

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

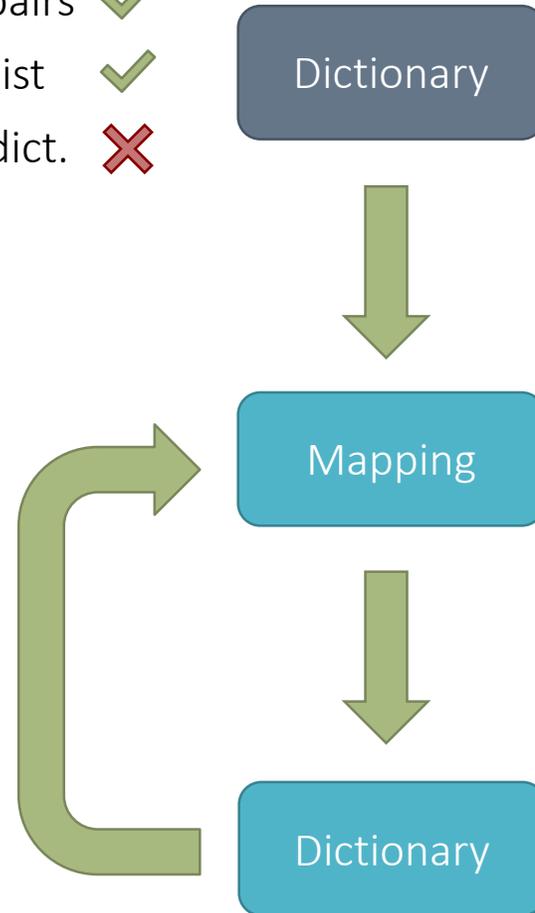


Cross-lingual embedding mappings

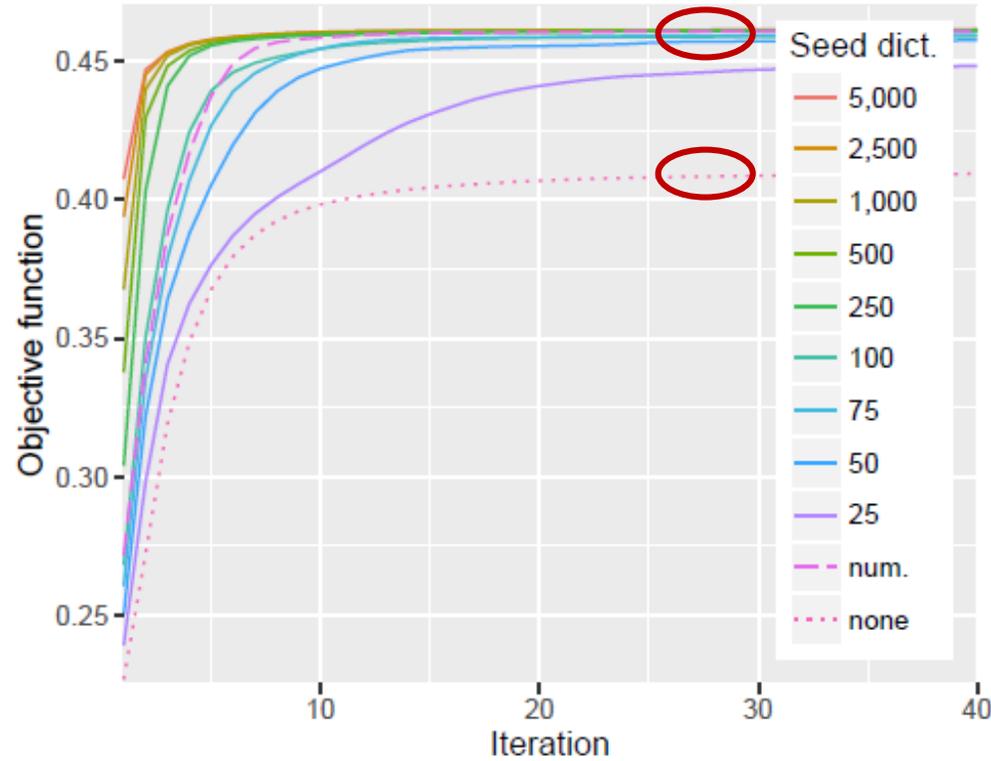


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

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

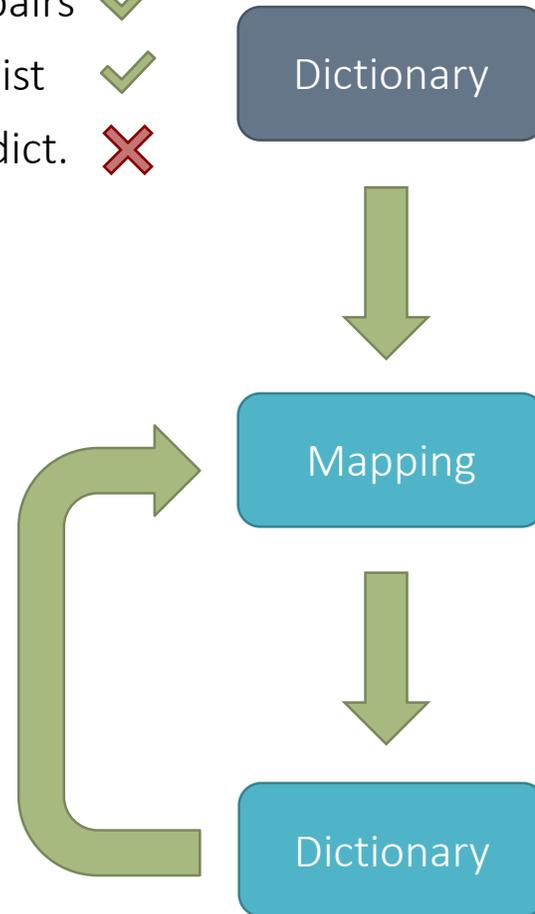


Cross-lingual embedding mappings



- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

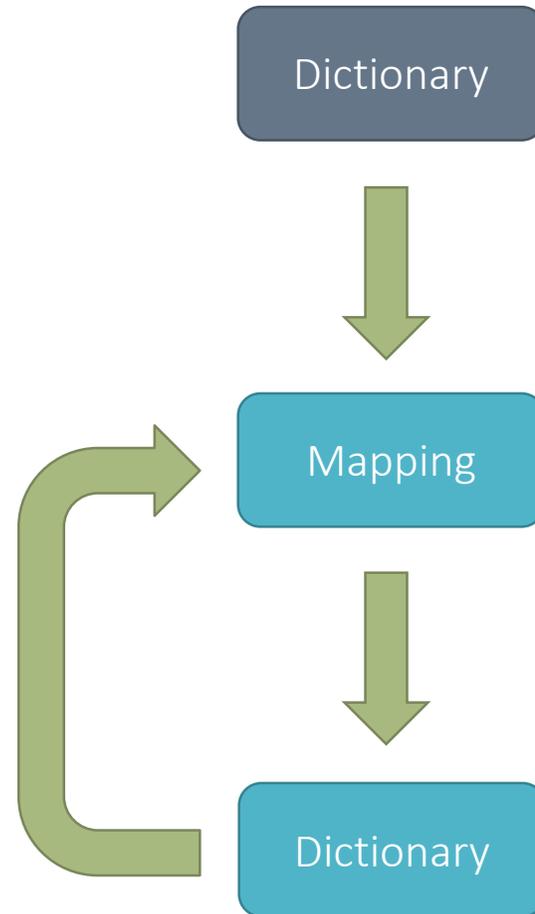
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



Cross-lingual embedding mappings

Intra-lingual similarity distribution

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

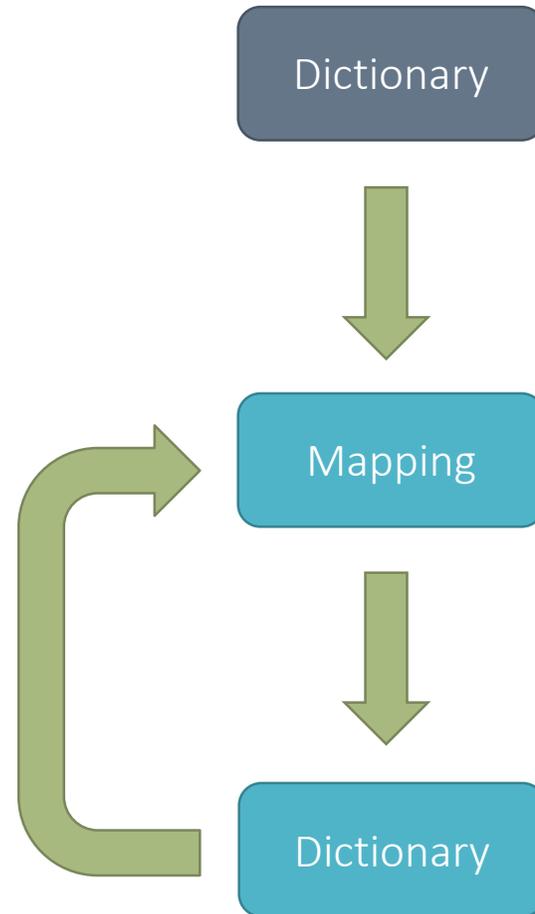


Cross-lingual embedding mappings

English

Intra-lingual similarity distribution

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



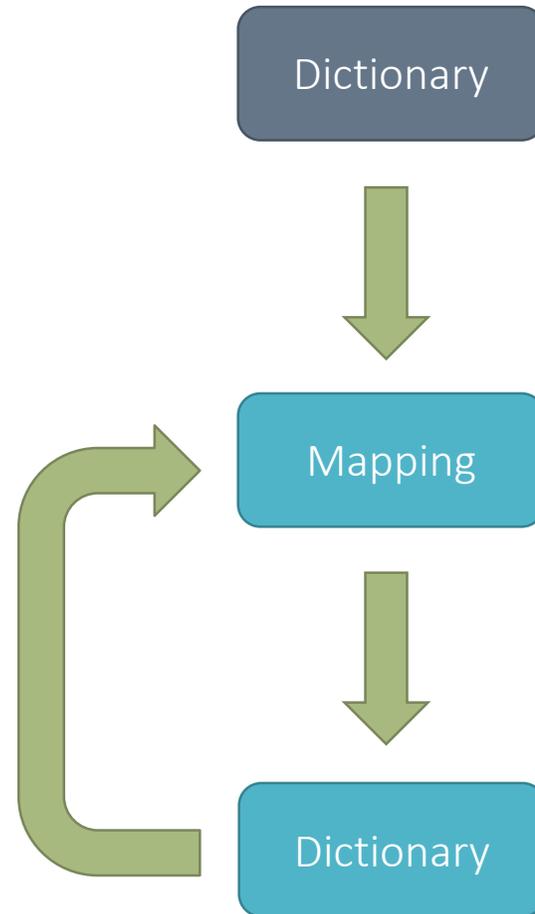
Cross-lingual embedding mappings

English

two

Intra-lingual similarity distribution

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



Cross-lingual embedding mappings

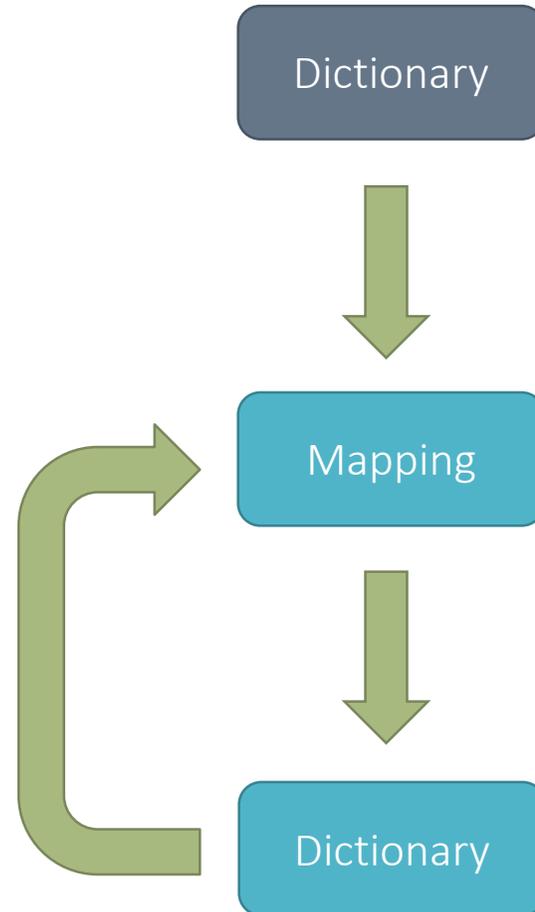
English

```
for x in vocab:  
    sim("two", x)
```

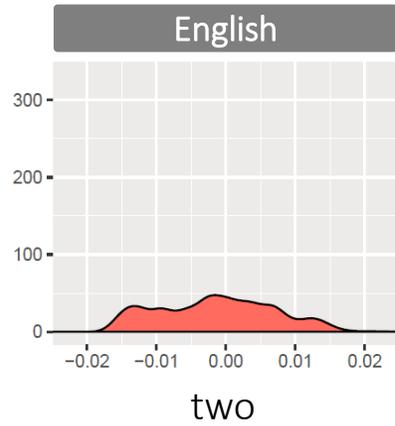
two

Intra-lingual similarity distribution

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

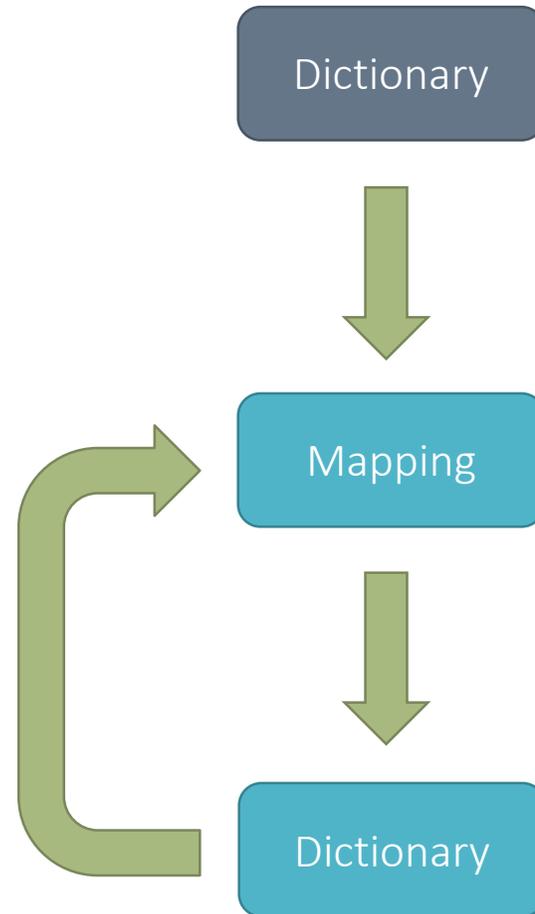


Cross-lingual embedding mappings

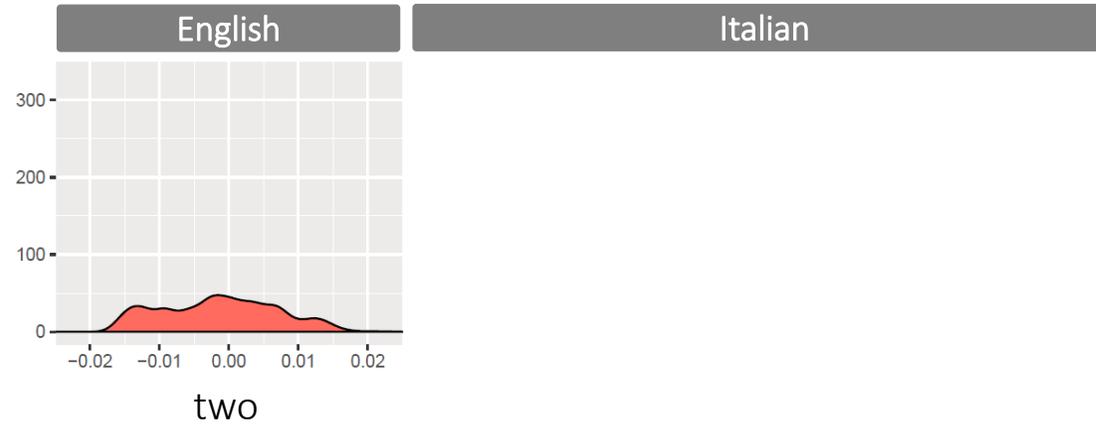


Intra-lingual similarity distribution

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

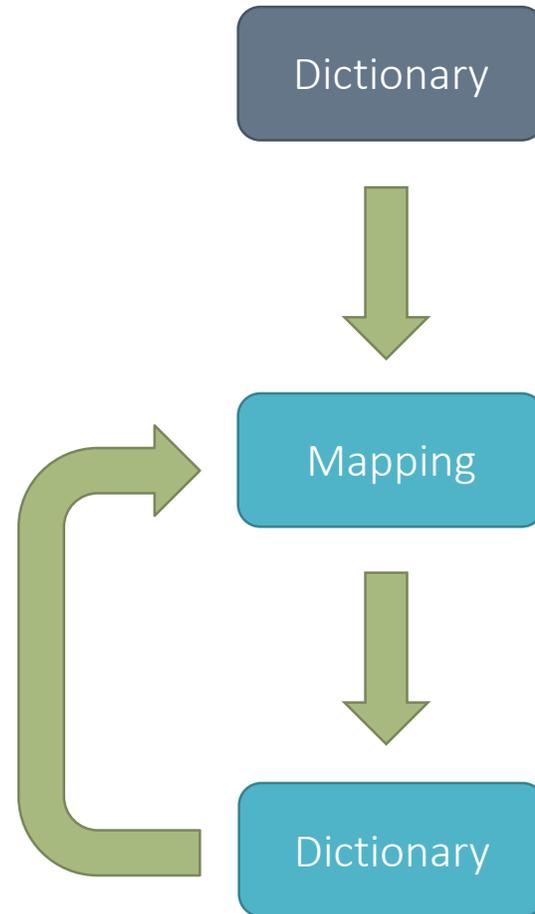


Cross-lingual embedding mappings

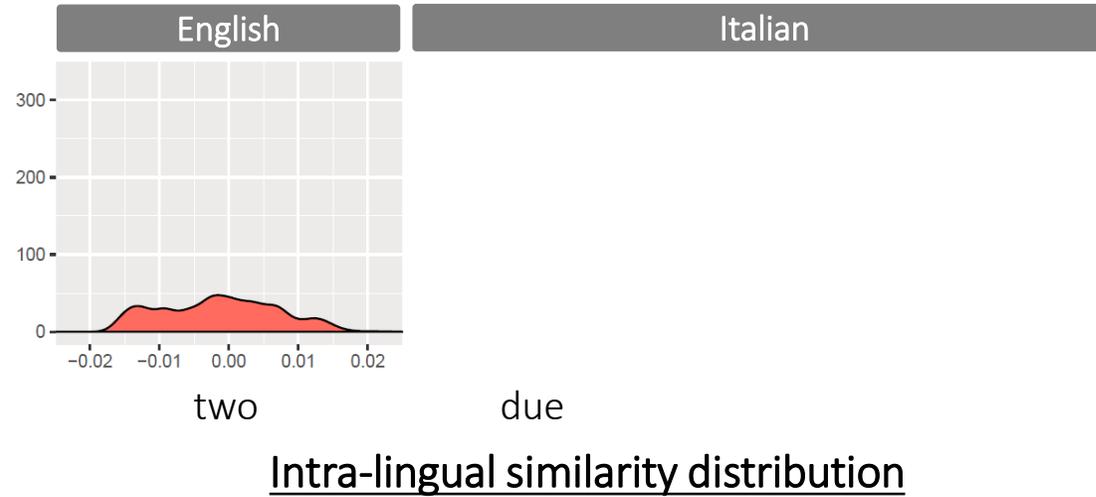


Intra-lingual similarity distribution

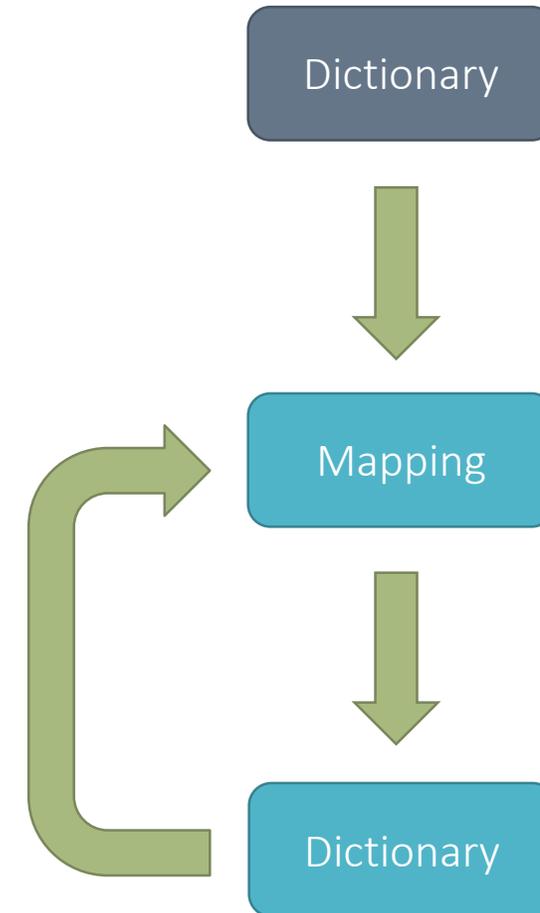
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



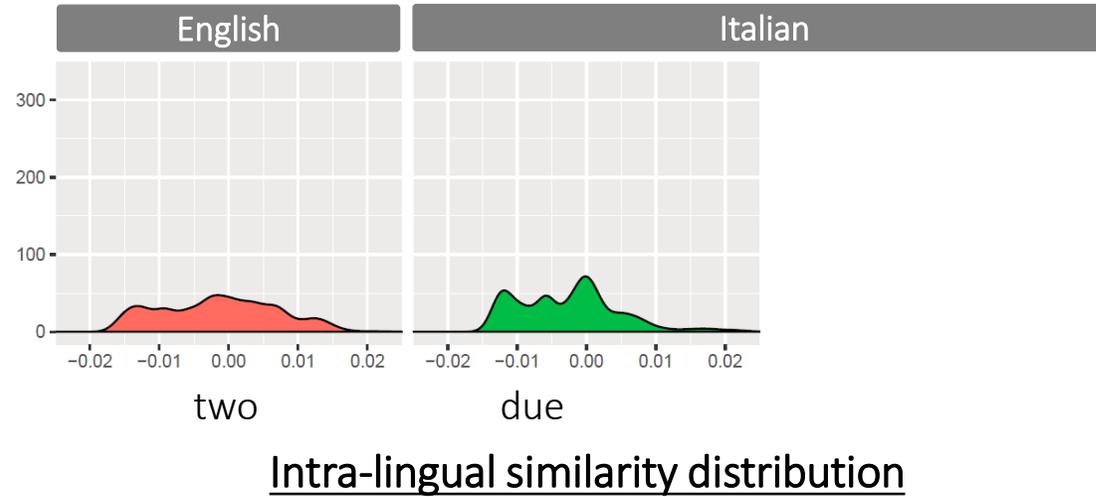
Cross-lingual embedding mappings



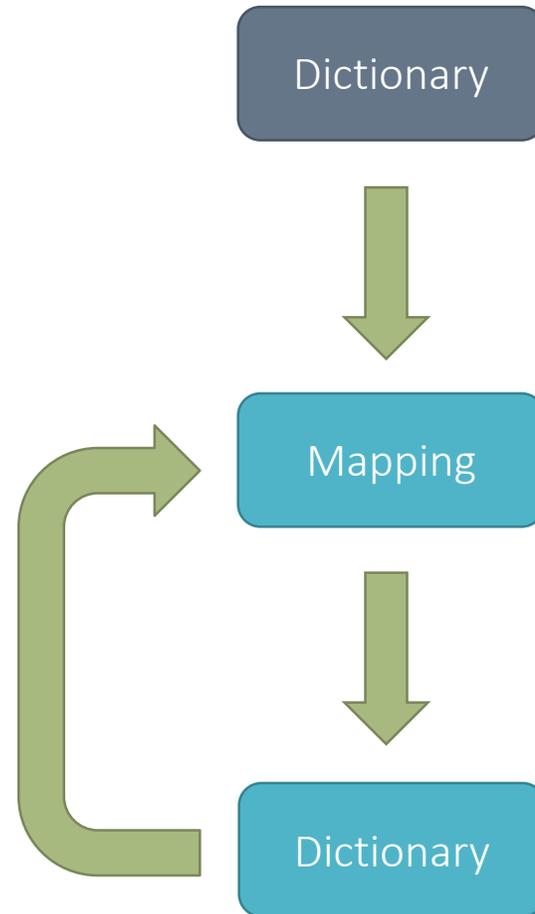
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



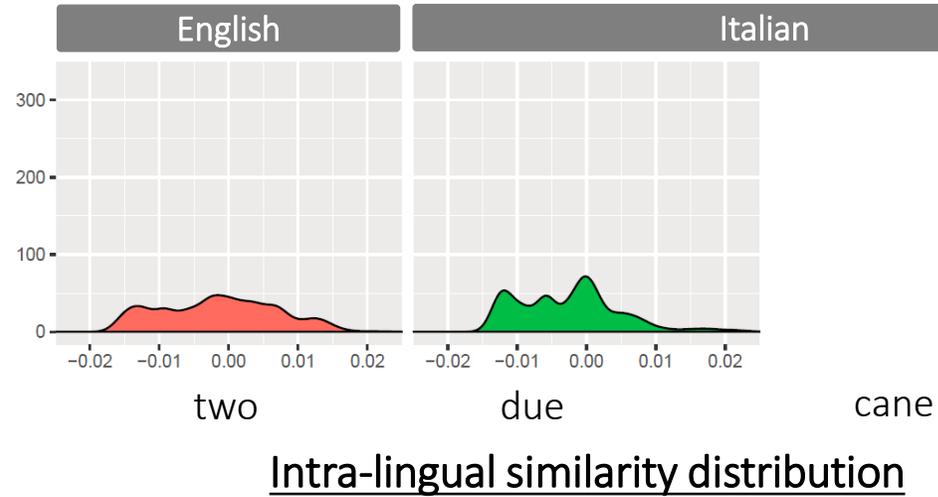
Cross-lingual embedding mappings



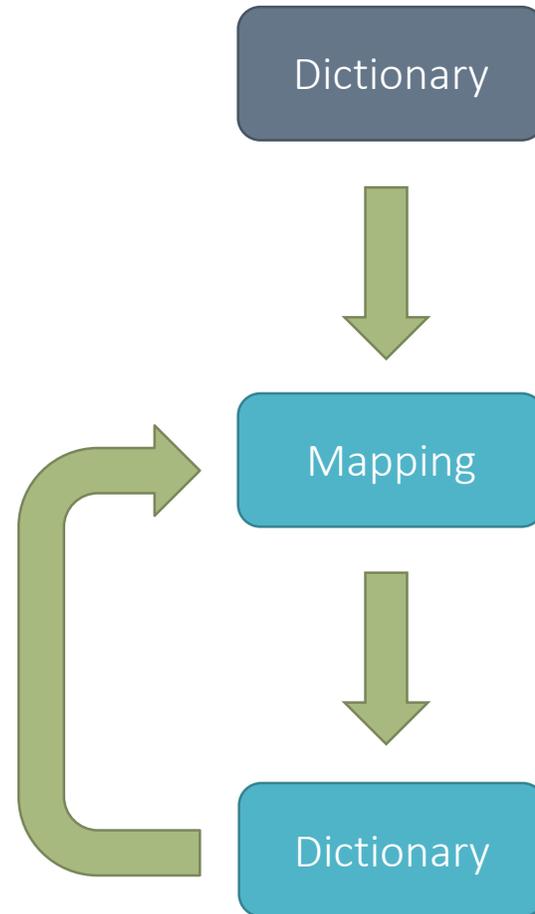
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



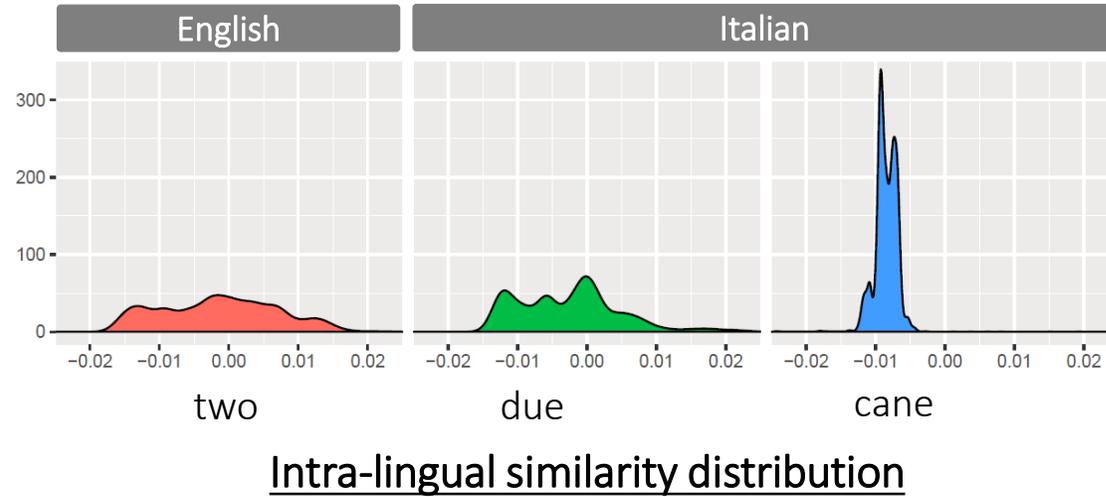
Cross-lingual embedding mappings



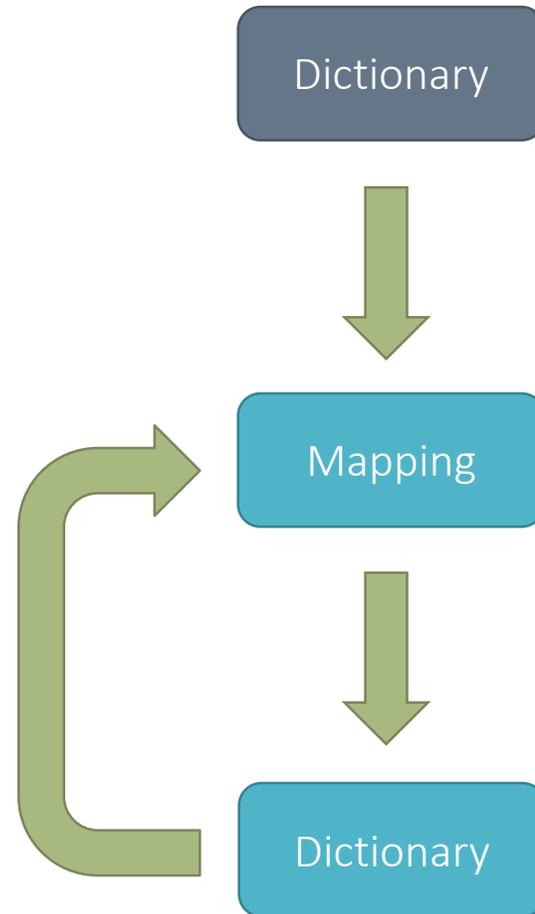
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



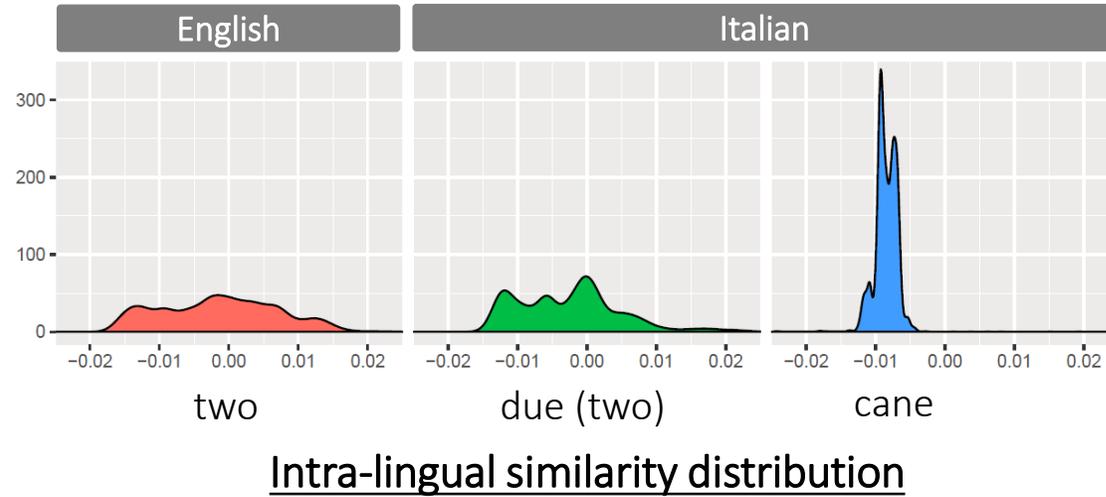
Cross-lingual embedding mappings



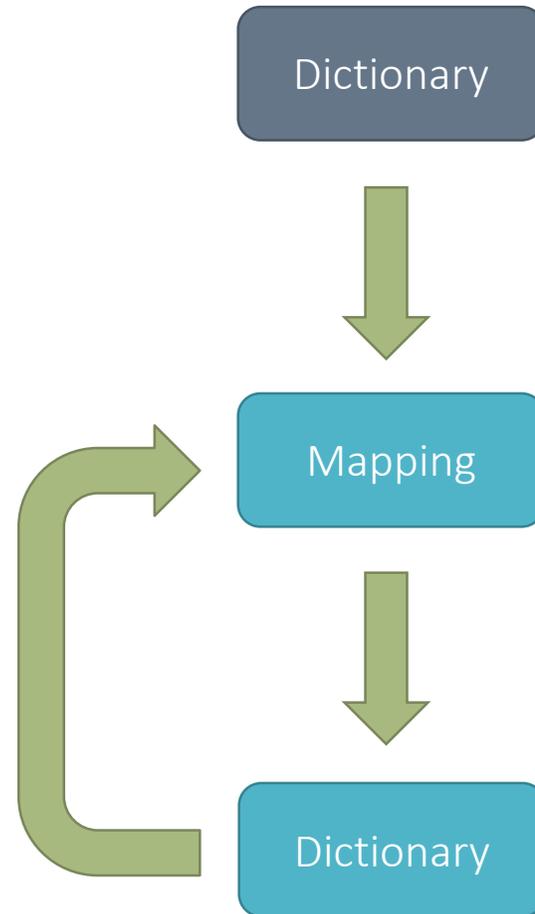
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i*} W - Z_{j*}\|^2$$



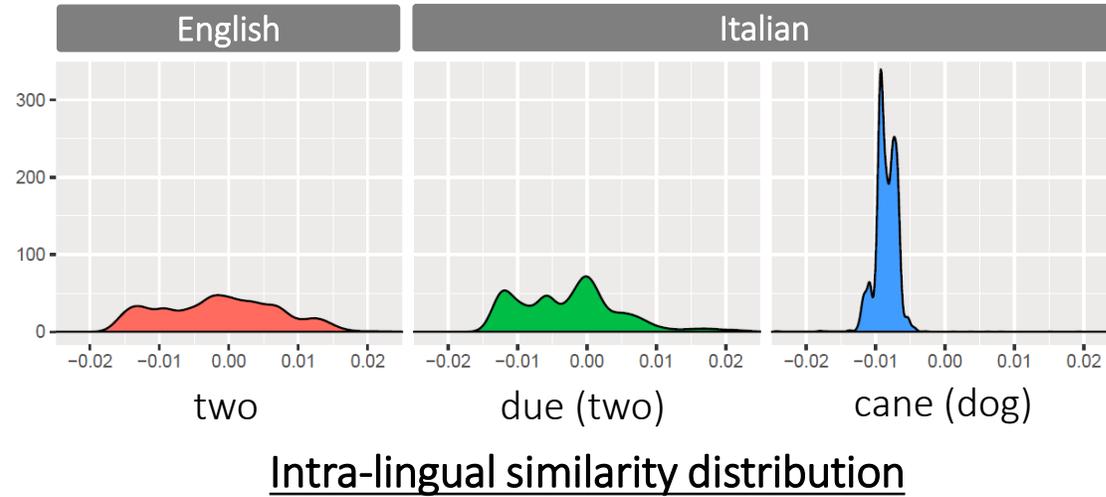
Cross-lingual embedding mappings



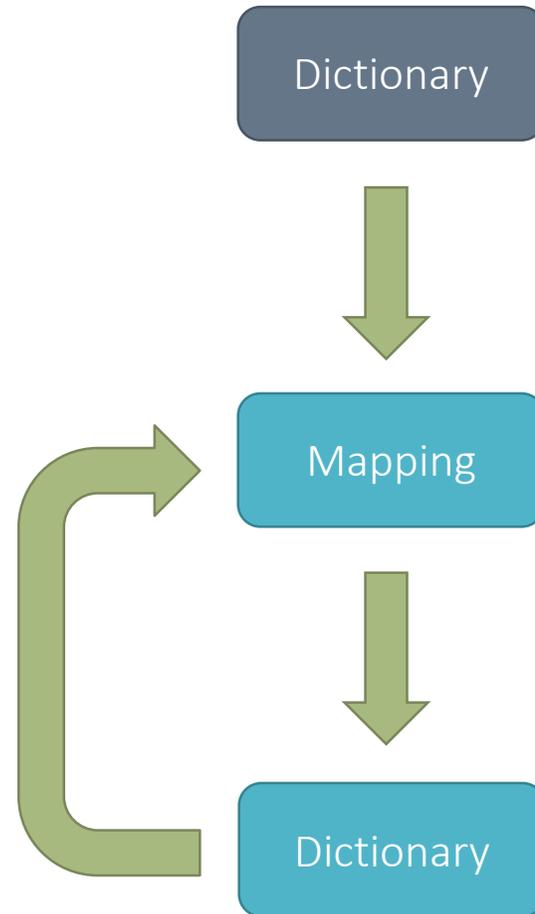
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



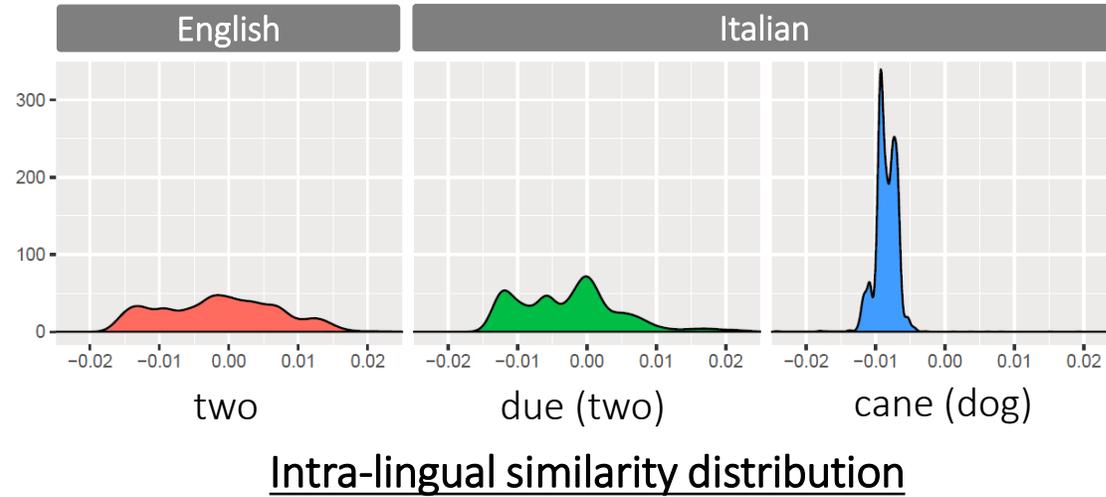
Cross-lingual embedding mappings



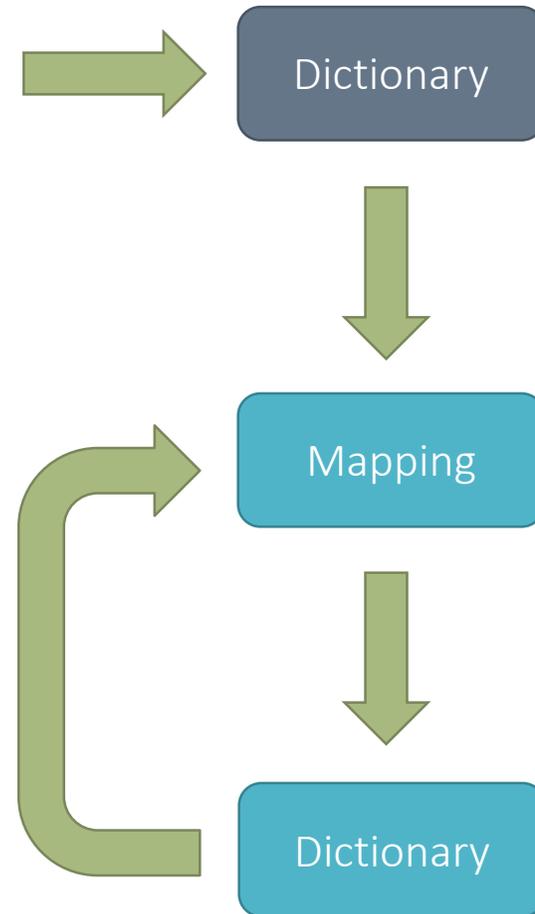
$$W^* = \arg \min_{W \in O(n)} \sum_i \min_j \|X_{i^*} W - Z_{j^*}\|^2$$



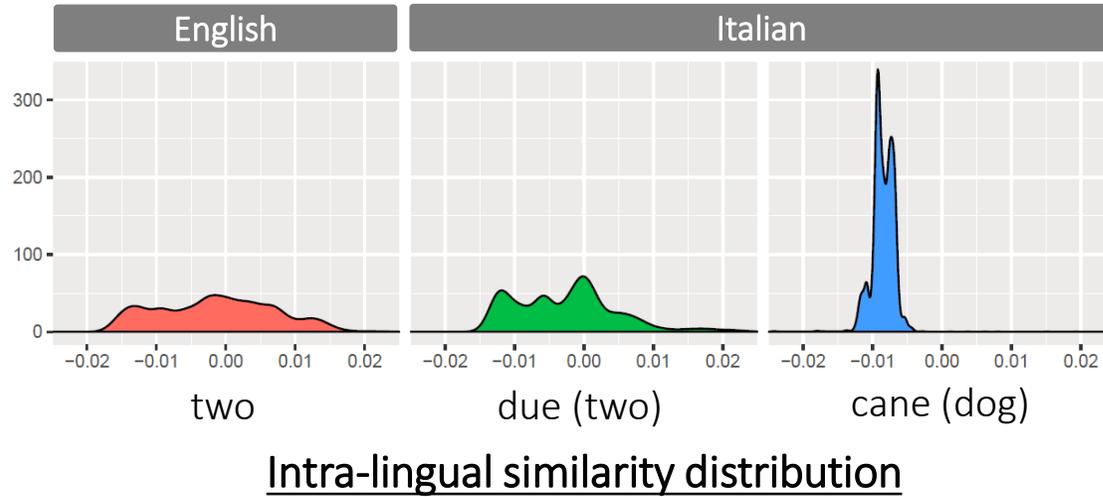
Cross-lingual embedding mappings



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

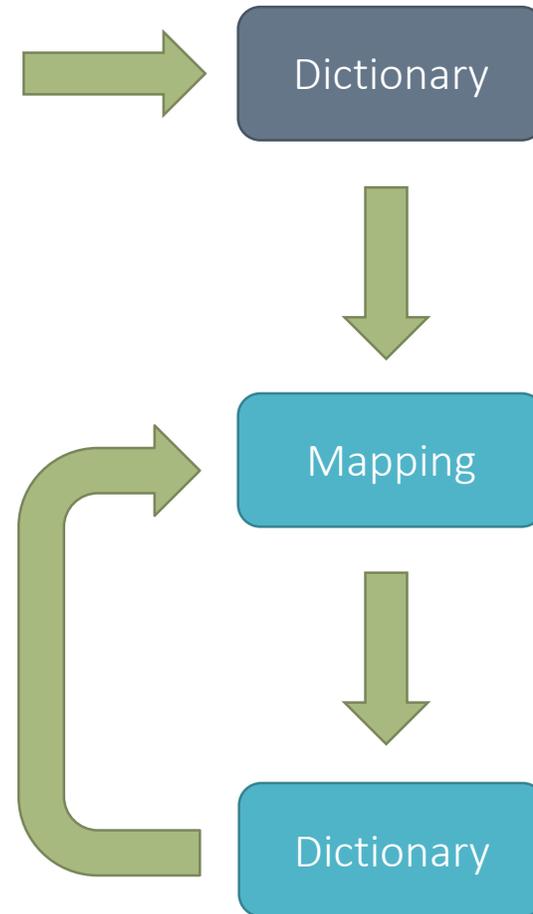


Cross-lingual embedding mappings

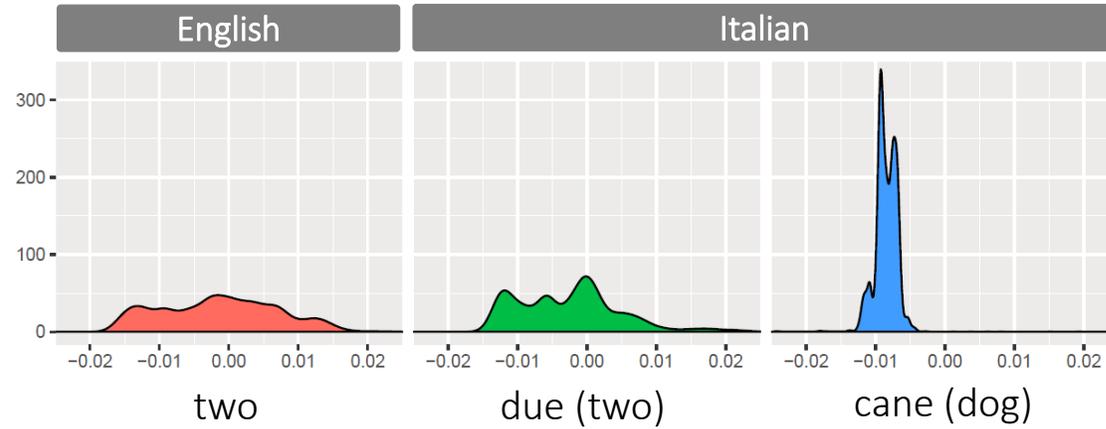


$$X' = \text{sorted}(\sqrt{XX^T})$$

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



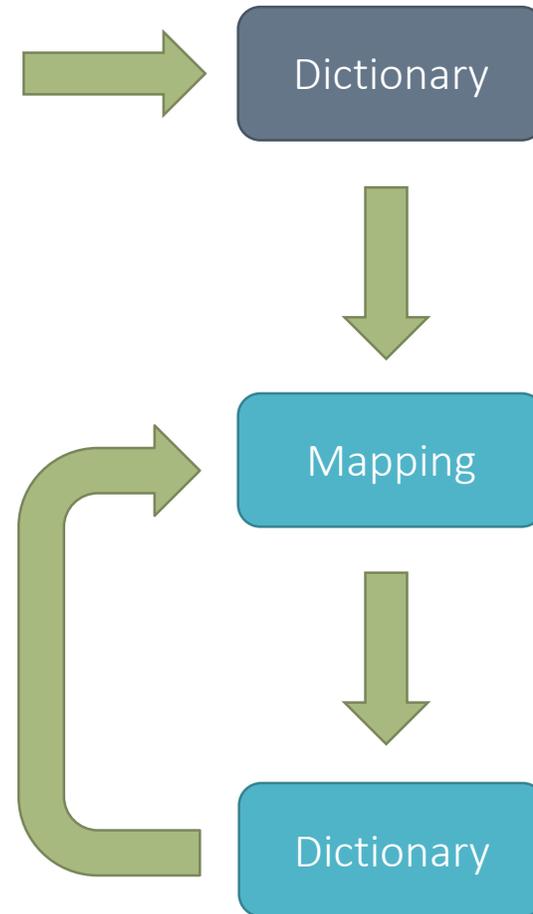
Cross-lingual embedding mappings



Intra-lingual similarity distribution

$$X' = \text{sorted}(\sqrt{XX^T}) \quad Z' = \text{sorted}(\sqrt{ZZ^T})$$

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



Cross-lingual embedding mappings

Cross-lingual embedding mappings

Supervision	Method
-------------	--------

Cross-lingual embedding mappings

Supervision	Method
5k dict.	Mikolov et al. (2013)
	Faruqui and Dyer (2014)
	Shigeto et al. (2015)
	Dinu et al. (2015)
	Lazaridou et al. (2015)
	Xing et al. (2015)
	Zhang et al. (2016)
	Artetxe et al. (2016)
	Smith et al. (2017)
	Artetxe et al. (2018a)

Cross-lingual embedding mappings

Supervision	Method
5k dict.	Mikolov et al. (2013)
	Faruqui and Dyer (2014)
	Shigeto et al. (2015)
	Dinu et al. (2015)
	Lazaridou et al. (2015)
	Xing et al. (2015)
	Zhang et al. (2016)
	Artetxe et al. (2016)
	Smith et al. (2017)
Artetxe et al. (2018a)	
25 dict.	Artetxe et al. (2017)

Cross-lingual embedding mappings

Supervision	Method
5k dict.	Mikolov et al. (2013)
	Faruqui and Dyer (2014)
	Shigeto et al. (2015)
	Dinu et al. (2015)
	Lazaridou et al. (2015)
	Xing et al. (2015)
	Zhang et al. (2016)
	Artetxe et al. (2016)
	Smith et al. (2017)
	Artetxe et al. (2018a)
25 dict.	Artetxe et al. (2017)
Init.	Smith et al. (2017), cognates
heurist.	Artetxe et al. (2017), num.

Cross-lingual embedding mappings

Supervision	Method
5k dict.	Mikolov et al. (2013)
	Faruqui and Dyer (2014)
	Shigeto et al. (2015)
	Dinu et al. (2015)
	Lazaridou et al. (2015)
	Xing et al. (2015)
	Zhang et al. (2016)
	Artetxe et al. (2016)
	Smith et al. (2017)
Artetxe et al. (2018a)	
25 dict.	Artetxe et al. (2017)
Init.	Smith et al. (2017), cognates
heurist.	Artetxe et al. (2017), num.
None	Zhang et al. (2017), $\lambda = 1$
	Zhang et al. (2017), $\lambda = 10$
	Conneau et al. (2018), code [‡]
	Conneau et al. (2018), paper [‡]
	Artetxe et al. (2018b)

Cross-lingual embedding mappings

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)				
	Faruqui and Dyer (2014)				
	Shigeto et al. (2015)				
	Dinu et al. (2015)				
	Lazaridou et al. (2015)				
	Xing et al. (2015)				
	Zhang et al. (2016)				
	Artetxe et al. (2016)				
	Smith et al. (2017)				
Artetxe et al. (2018a)					
25 dict.	Artetxe et al. (2017)				
Init.	Smith et al. (2017), cognates				
heurist.	Artetxe et al. (2017), num.				
None	Zhang et al. (2017), $\lambda = 1$				
	Zhang et al. (2017), $\lambda = 10$				
	Conneau et al. (2018), code [‡]				
	Conneau et al. (2018), paper [‡]				
	Artetxe et al. (2018b)				

Cross-lingual embedding mappings

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	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 [†]
	Zhang et al. (2016)	36.73 [†]	40.80 [†]	28.16 [†]	31.07 [†]
	Artetxe et al. (2016)	39.27	41.87 [*]	30.62 [*]	31.40 [*]
	Smith et al. (2017)	43.1	43.33 [†]	29.42 [†]	35.13 [†]
	Artetxe et al. (2018a)	45.27	44.13	32.94	36.60
25 dict.	Artetxe et al. (2017)	37.27	39.60	28.16	-
Init.	Smith et al. (2017), cognates	39.9	-	-	-
heurist.	Artetxe et al. (2017), num.	39.40	40.27	26.47	-
None	Zhang et al. (2017), $\lambda = 1$	0.00 [*]	0.00 [*]	0.00 [*]	0.00 [*]
	Zhang et al. (2017), $\lambda = 10$	0.00 [*]	0.00 [*]	0.01 [*]	0.01 [*]
	Conneau et al. (2018), code [‡]	45.15 [*]	46.83 [*]	0.38 [*]	35.38 [*]
	Conneau et al. (2018), paper [‡]	45.1	0.01 [*]	0.01 [*]	35.44 [*]
		Artetxe et al. (2018b)	48.13	48.19	32.63

Cross-lingual embedding mappings

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	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 [†]
	Zhang et al. (2016)	36.73 [†]	40.80 [†]	28.16 [†]	31.07 [†]
	Artetxe et al. (2016)	39.27	41.87 [*]	30.62 [*]	31.40 [*]
	Smith et al. (2017)	43.1	43.33 [†]	29.42 [†]	35.13 [†]
	Artetxe et al. (2018a)	45.27	44.13	32.94	36.60
25 dict.	Artetxe et al. (2017)	37.27	39.60	28.16	-
Init.	Smith et al. (2017), cognates	39.9	-	-	-
heurist.	Artetxe et al. (2017), num.	39.40	40.27	26.47	-
None	Zhang et al. (2017), $\lambda = 1$	0.00 [*]	0.00 [*]	0.00 [*]	0.00 [*]
	Zhang et al. (2017), $\lambda = 10$	0.00 [*]	0.00 [*]	0.01 [*]	0.01 [*]
	Conneau et al. (2018), code [‡]	45.15 [*]	46.83 [*]	0.38 [*]	35.38 [*]
	Conneau et al. (2018), paper [‡]	45.1	0.01 [*]	0.01 [*]	35.44 [*]
		Artetxe et al. (2018b)	48.13	48.19	32.63

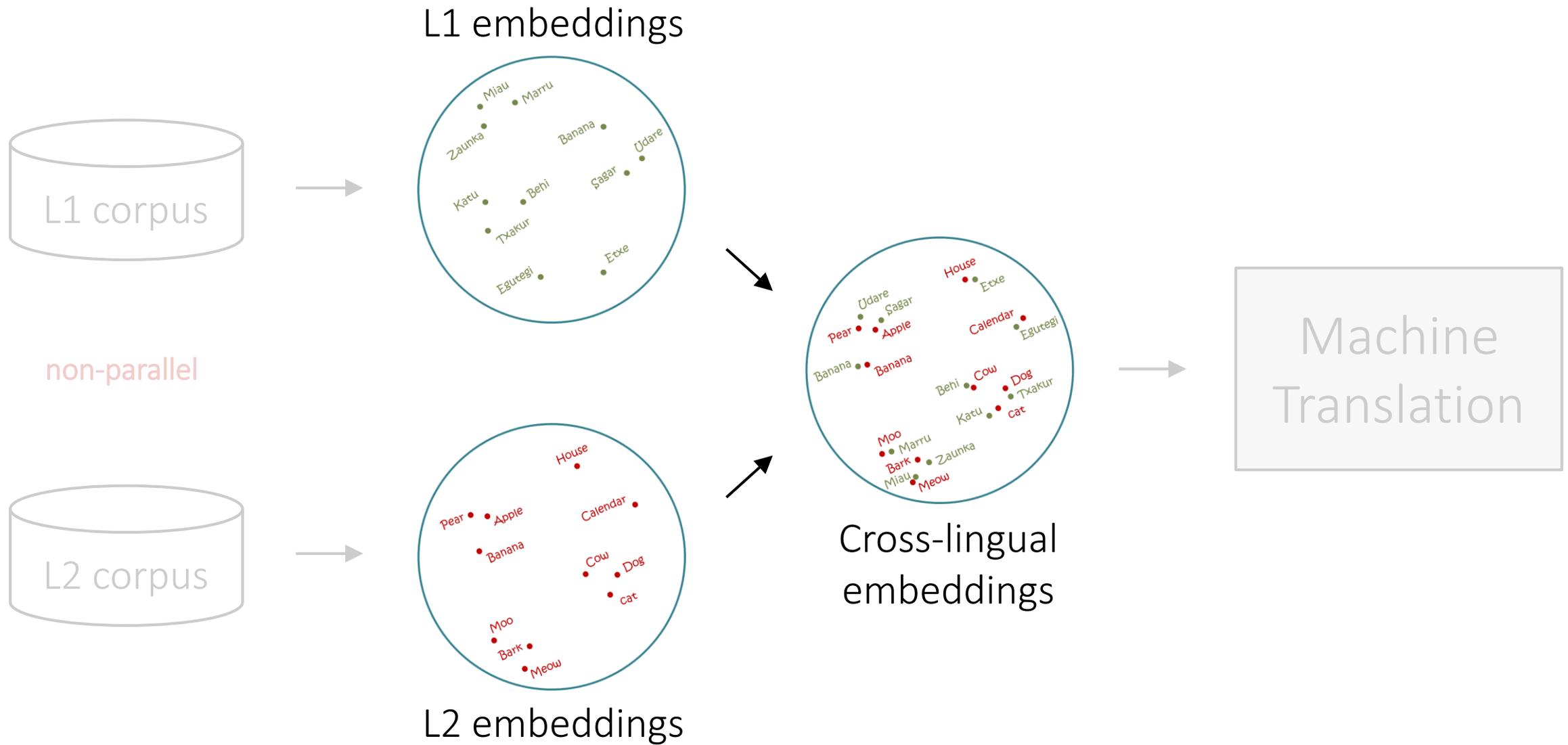
Cross-lingual embedding mappings

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	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 [†]
	Zhang et al. (2016)	36.73 [†]	40.80 [†]	28.16 [†]	31.07 [†]
	Artetxe et al. (2016)	39.27	41.87 [*]	30.62 [*]	31.40 [*]
	Smith et al. (2017)	43.1	43.33 [†]	29.42 [†]	35.13 [†]
	Artetxe et al. (2018a)	45.27	44.13	32.94	36.60
25 dict.	Artetxe et al. (2017)	37.27	39.60	28.16	-
Init.	Smith et al. (2017), cognates	39.9	-	-	-
heurist.	Artetxe et al. (2017), num.	39.40	40.27	26.47	-
None	Zhang et al. (2017), $\lambda = 1$	0.00 [*]	0.00 [*]	0.00 [*]	0.00 [*]
	Zhang et al. (2017), $\lambda = 10$	0.00 [*]	0.00 [*]	0.01 [*]	0.01 [*]
	Conneau et al. (2018), code [‡]	45.15 [*]	46.83 [*]	0.38 [*]	35.38 [*]
	Conneau et al. (2018), paper [‡]	45.1	0.01 [*]	0.01 [*]	35.44 [*]
		Artetxe et al. (2018b)	48.13	48.19	32.63

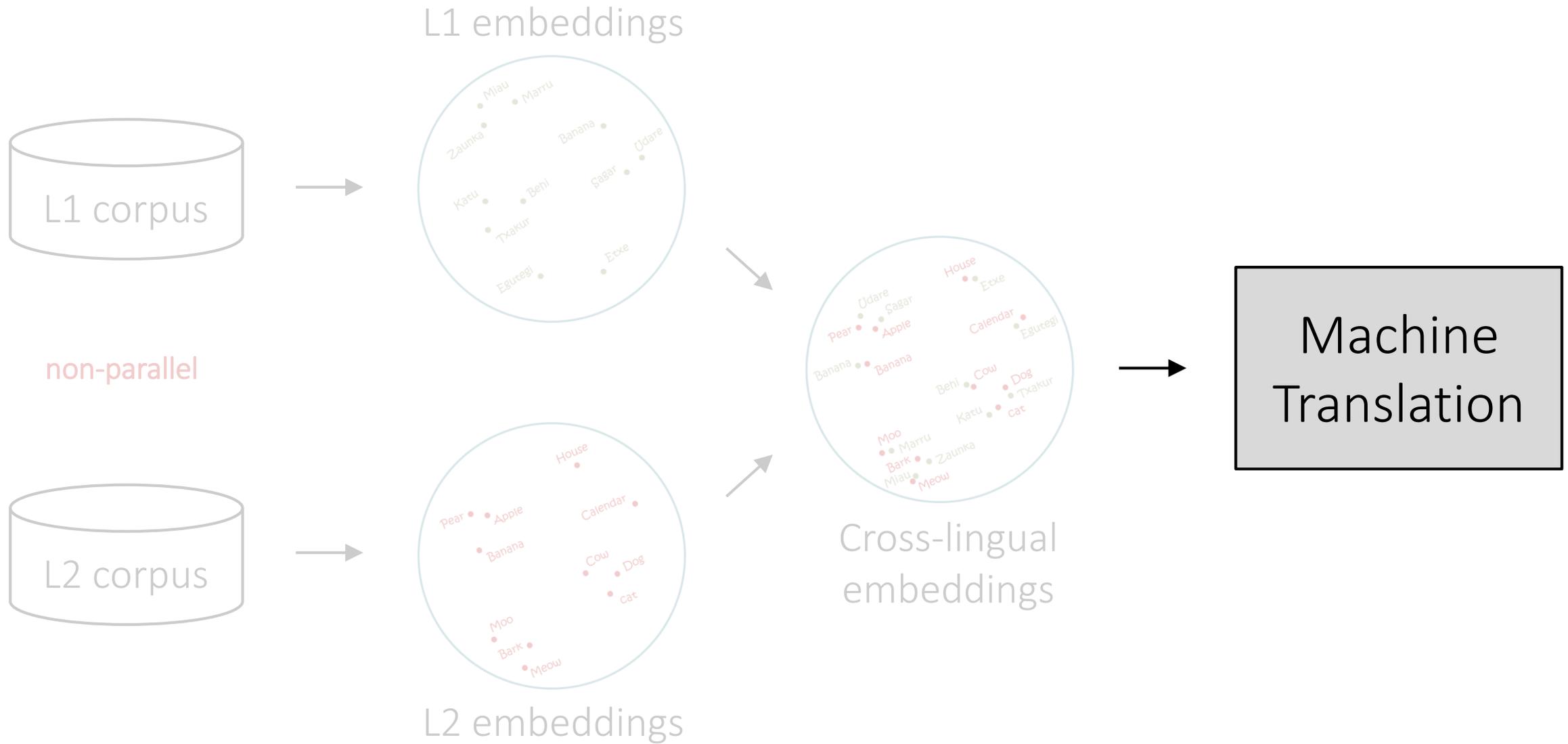
Cross-lingual embedding mappings

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	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 [†]
	Zhang et al. (2016)	36.73 [†]	40.80 [†]	28.16 [†]	31.07 [†]
	Artetxe et al. (2016)	39.27	41.87 [*]	30.62 [*]	31.40 [*]
	Smith et al. (2017)	43.1	43.33 [†]	29.42 [†]	35.13 [†]
	Artetxe et al. (2018a)	45.27	44.13	32.94	36.60
25 dict.	Artetxe et al. (2017)	37.27	39.60	28.16	-
Init.	Smith et al. (2017), cognates	39.9	-	-	-
heurist.	Artetxe et al. (2017), num.	39.40	40.27	26.47	-
None	Zhang et al. (2017), $\lambda = 1$	0.00 [*]	0.00 [*]	0.00 [*]	0.00 [*]
	Zhang et al. (2017), $\lambda = 10$	0.00 [*]	0.00 [*]	0.01 [*]	0.01 [*]
	Conneau et al. (2018), code [‡]	45.15 [*]	46.83 [*]	0.38 [*]	35.38 [*]
	Conneau et al. (2018), paper [‡]	45.1	0.01 [*]	0.01 [*]	35.44 [*]
		Artetxe et al. (2018b)	48.13	48.19	32.63

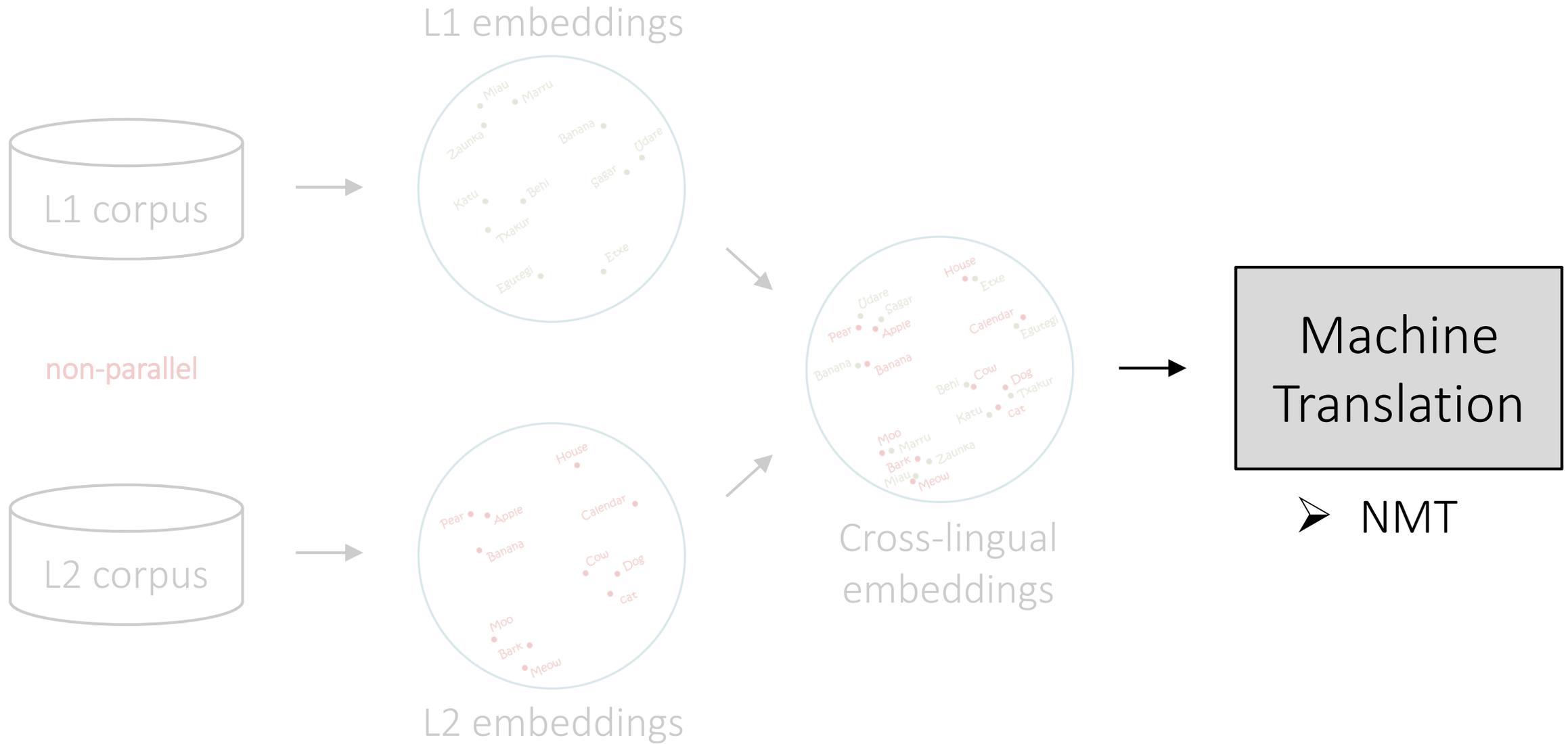
Overview



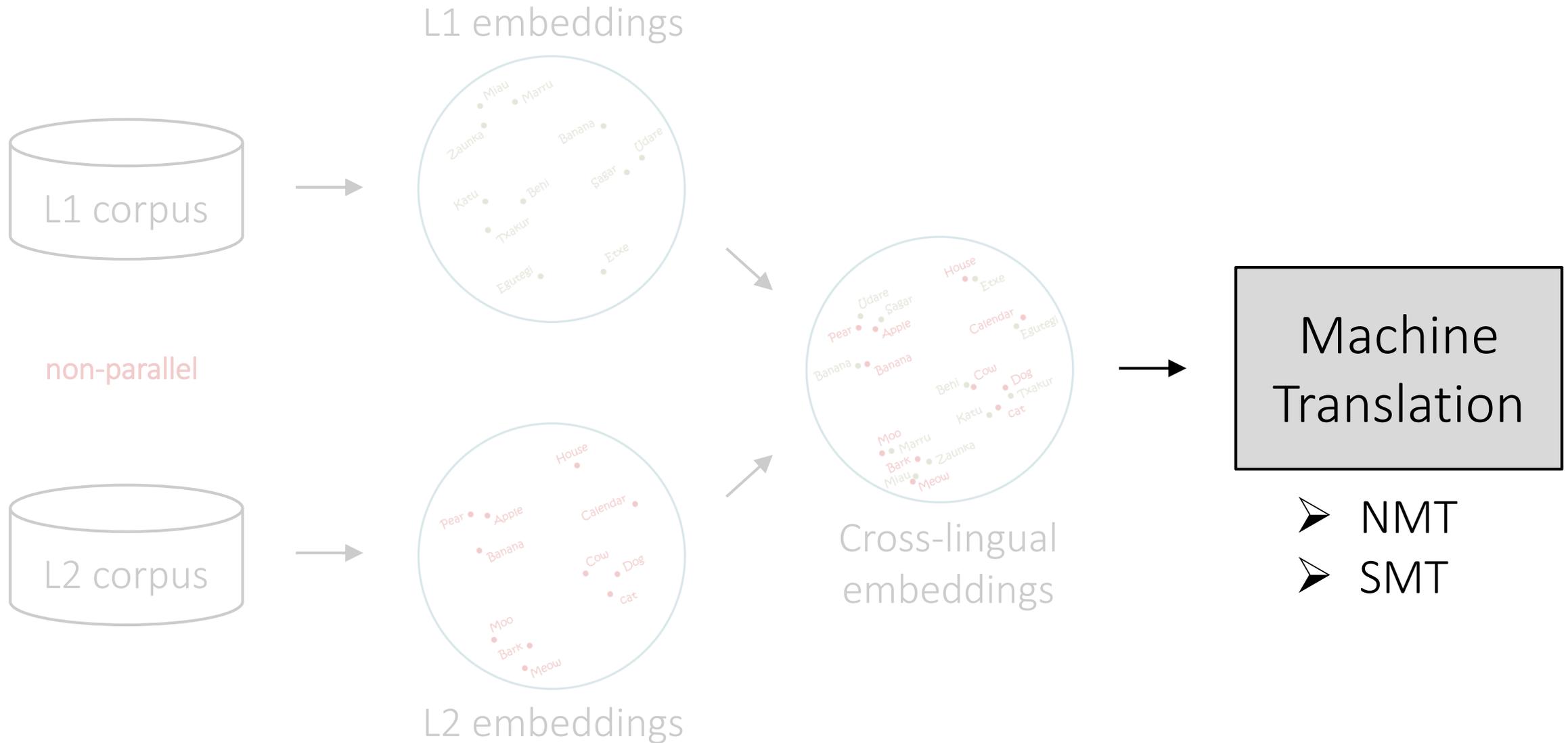
Overview



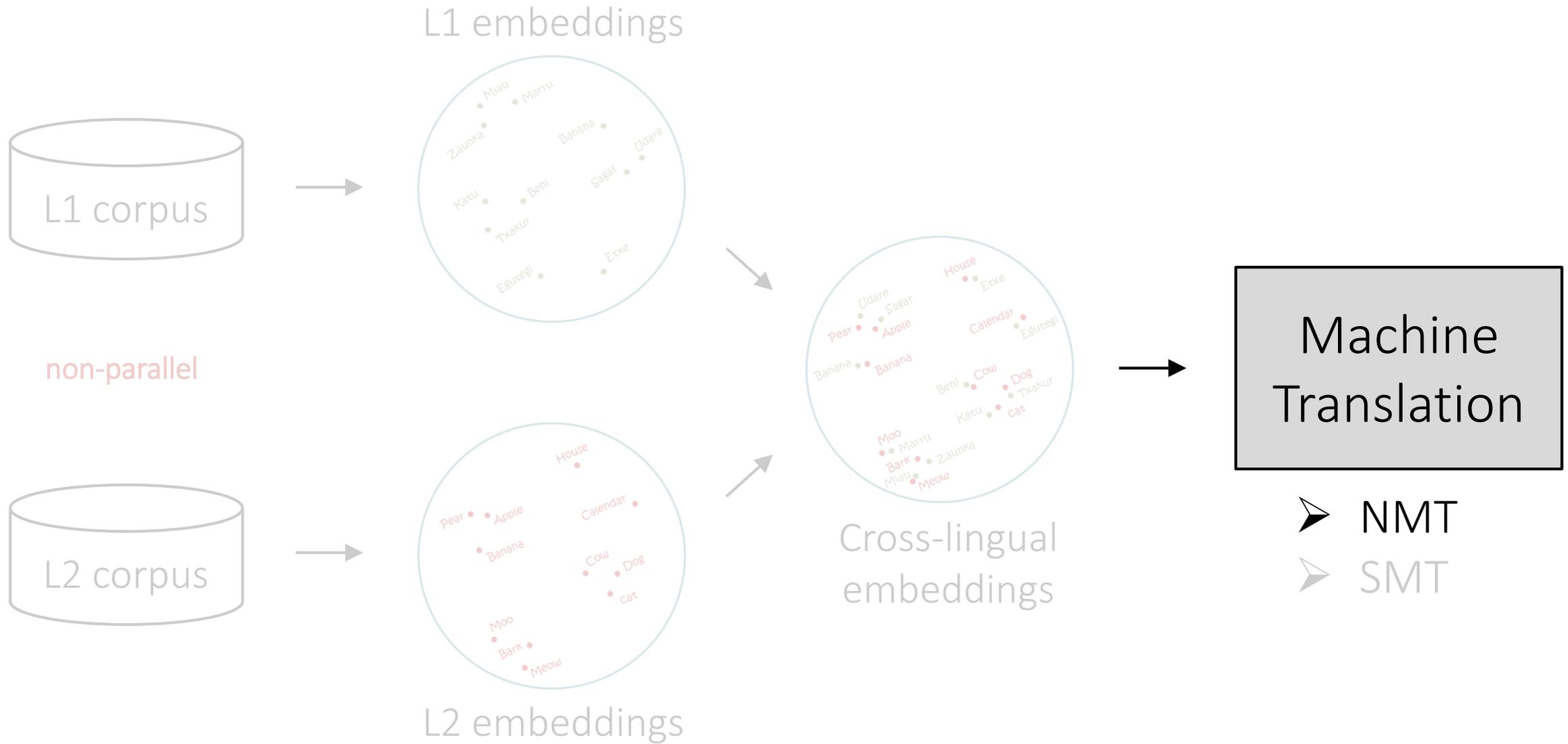
Overview



Overview

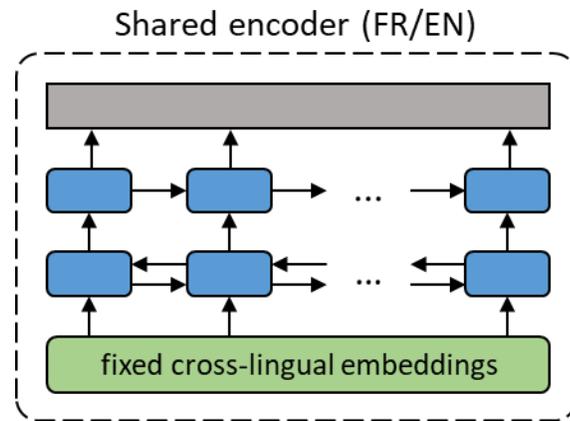


Overview

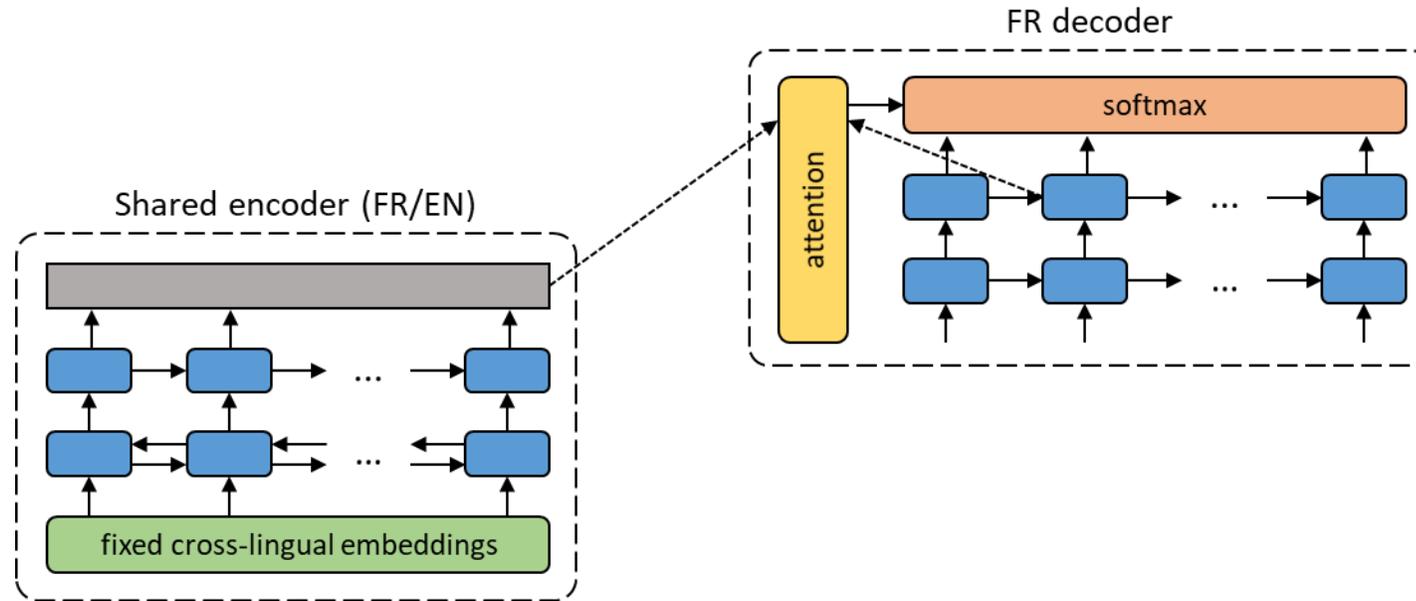


Unsupervised neural machine translation

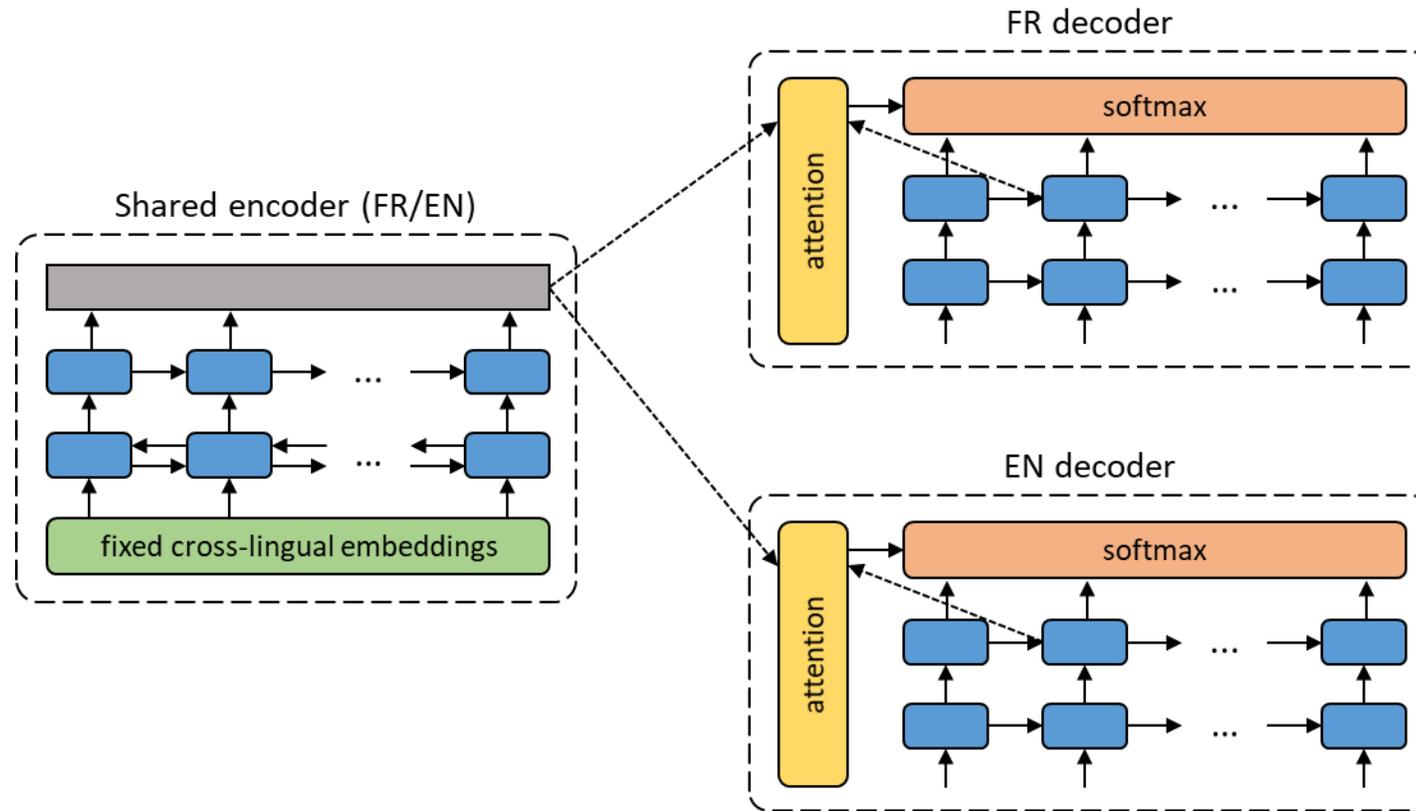
Unsupervised neural machine translation



Unsupervised neural machine translation

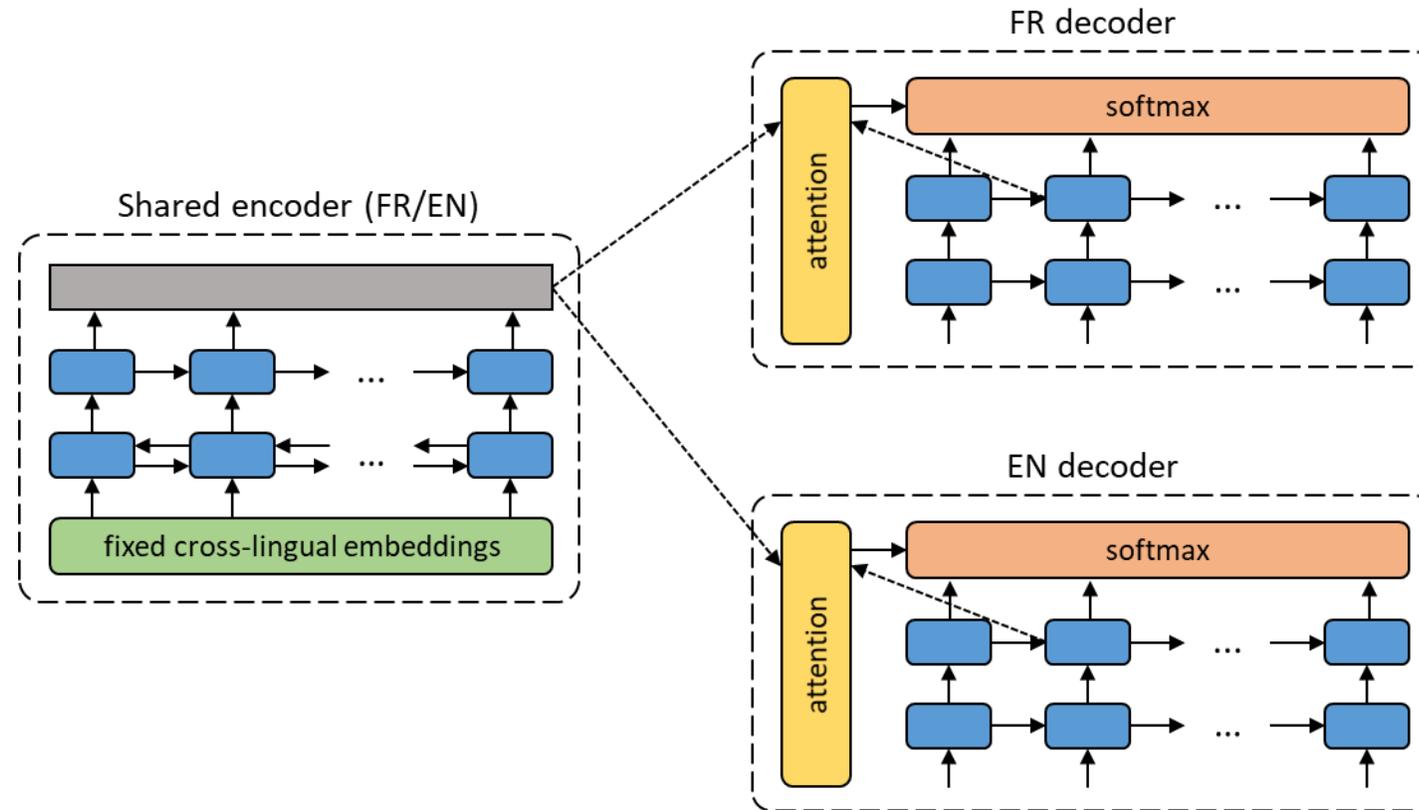


Unsupervised neural machine translation



Unsupervised neural machine translation

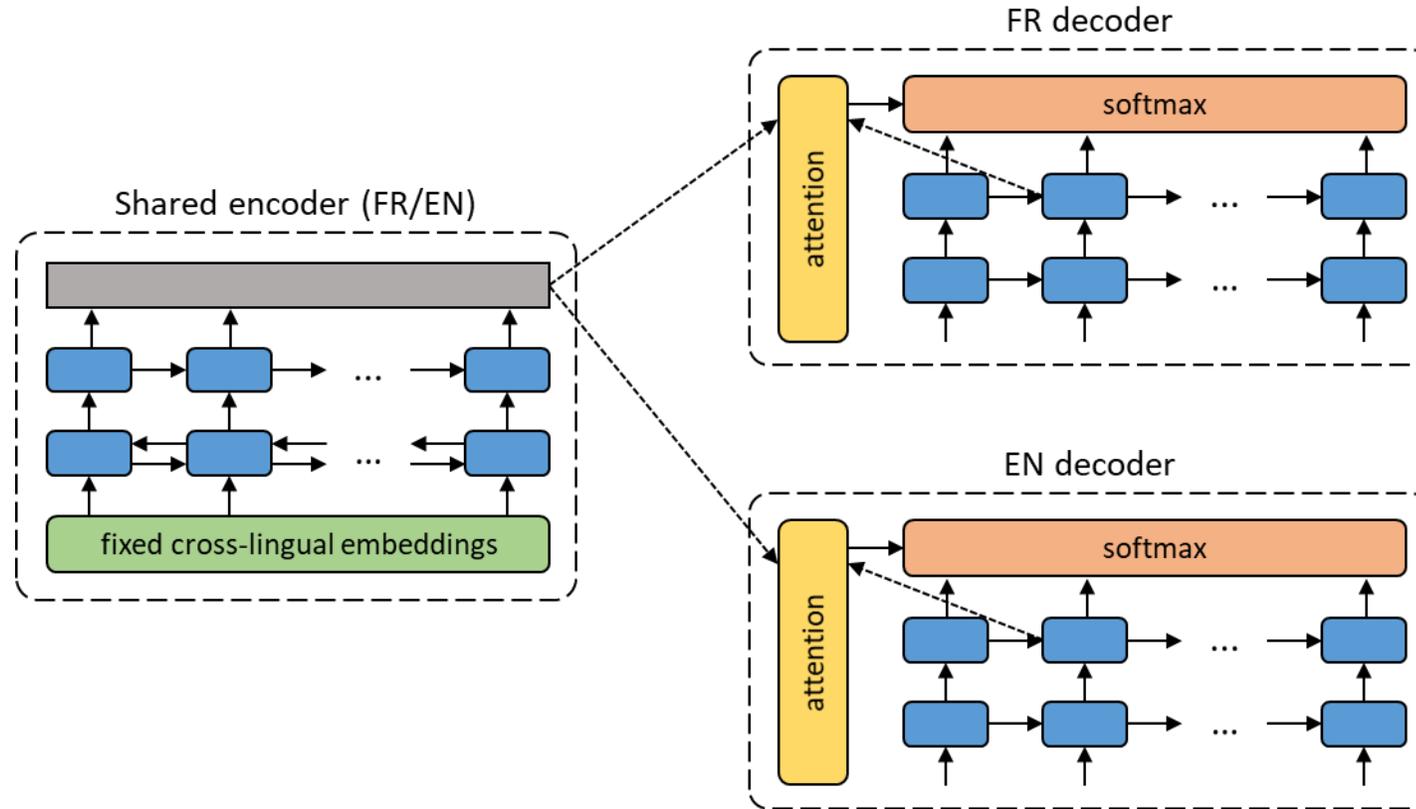
Training



Unsupervised neural machine translation

Training

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

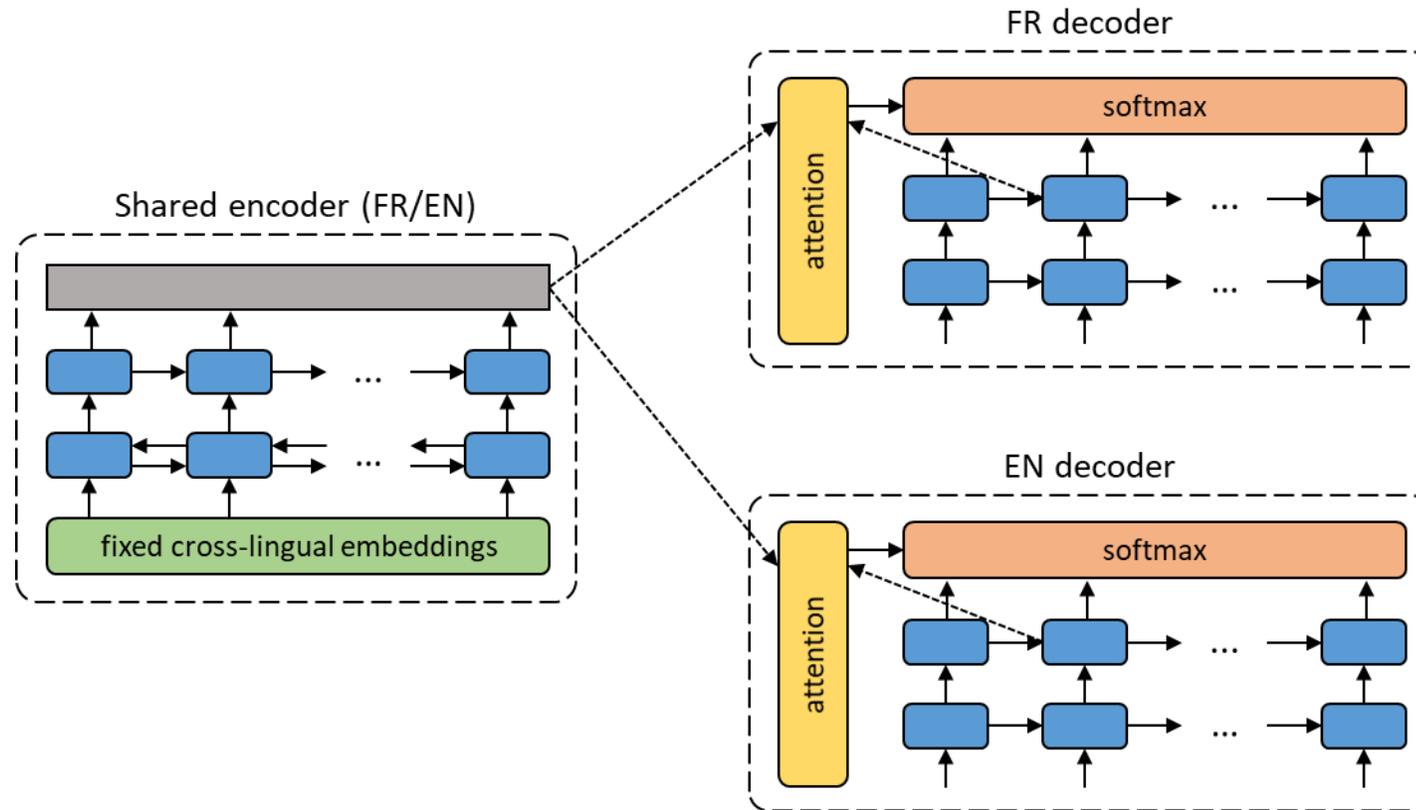


Unsupervised neural machine translation

Training

- Supervised

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

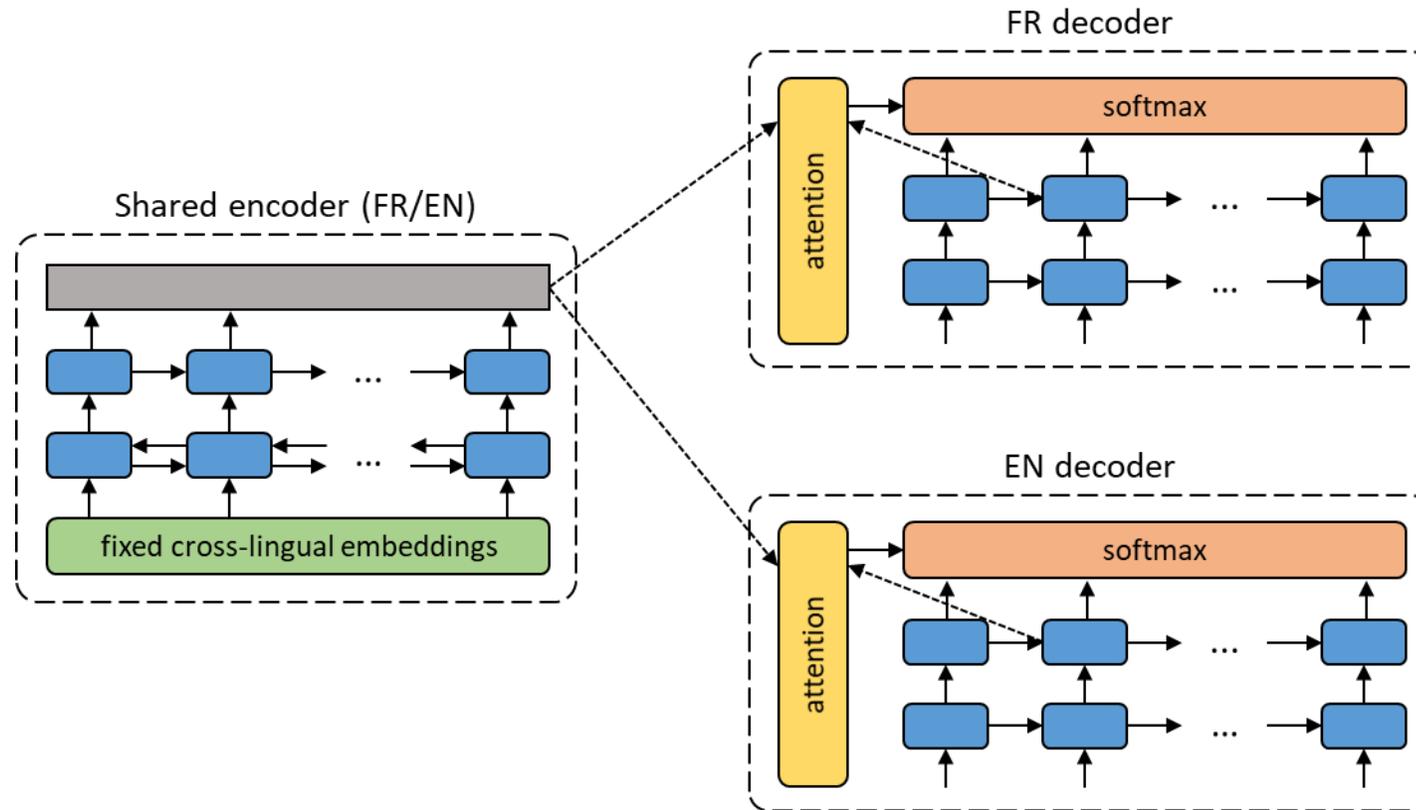


Unsupervised neural machine translation

Training

- Supervised

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



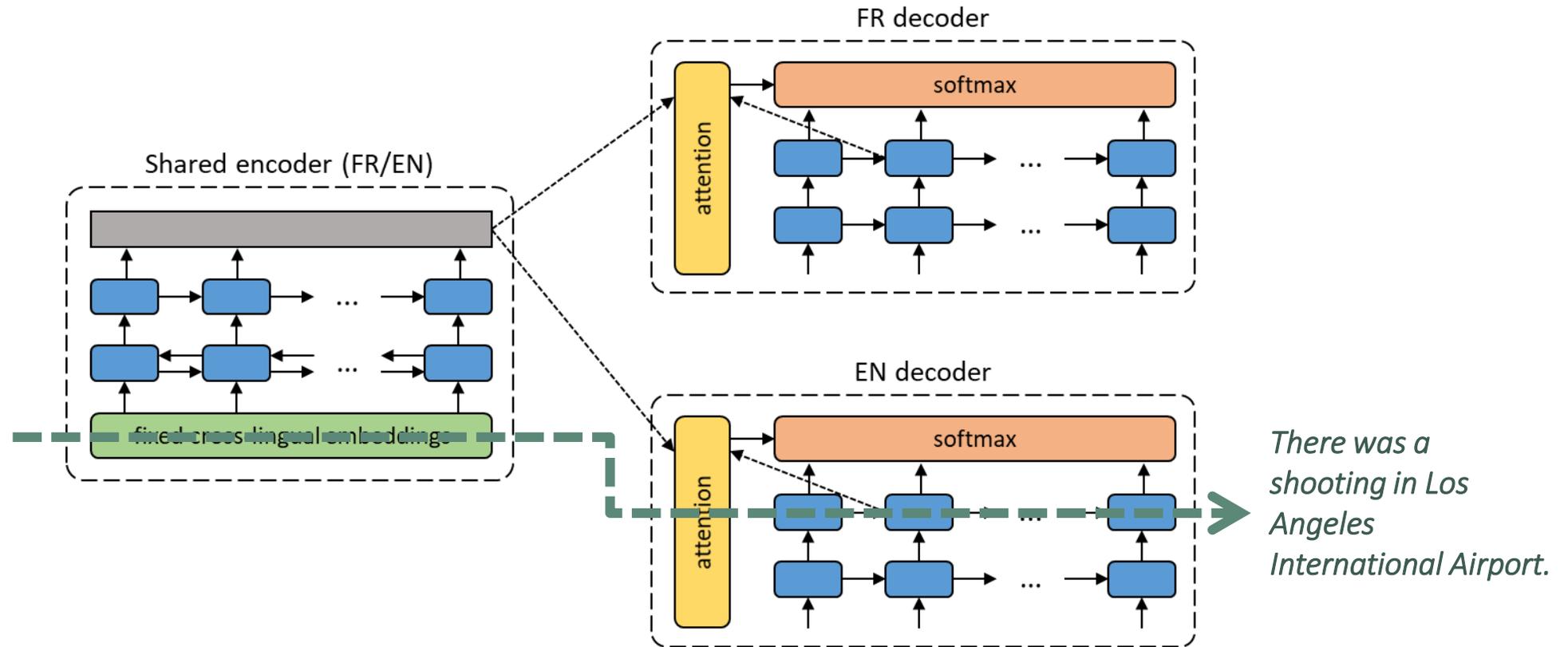
There was a shooting in Los Angeles International Airport.

Unsupervised neural machine translation

Training

- Supervised

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



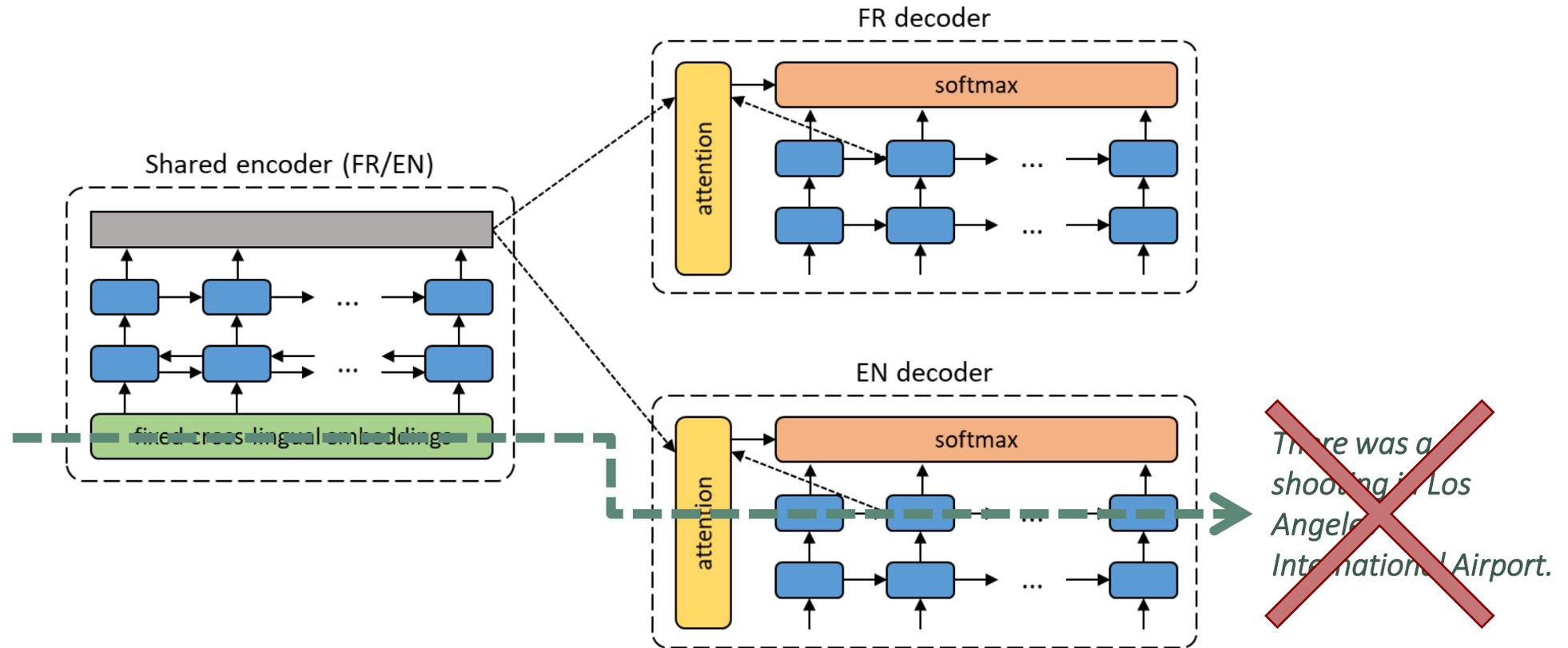
There was a shooting in Los Angeles International Airport.

Unsupervised neural machine translation

Training

- Supervised

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

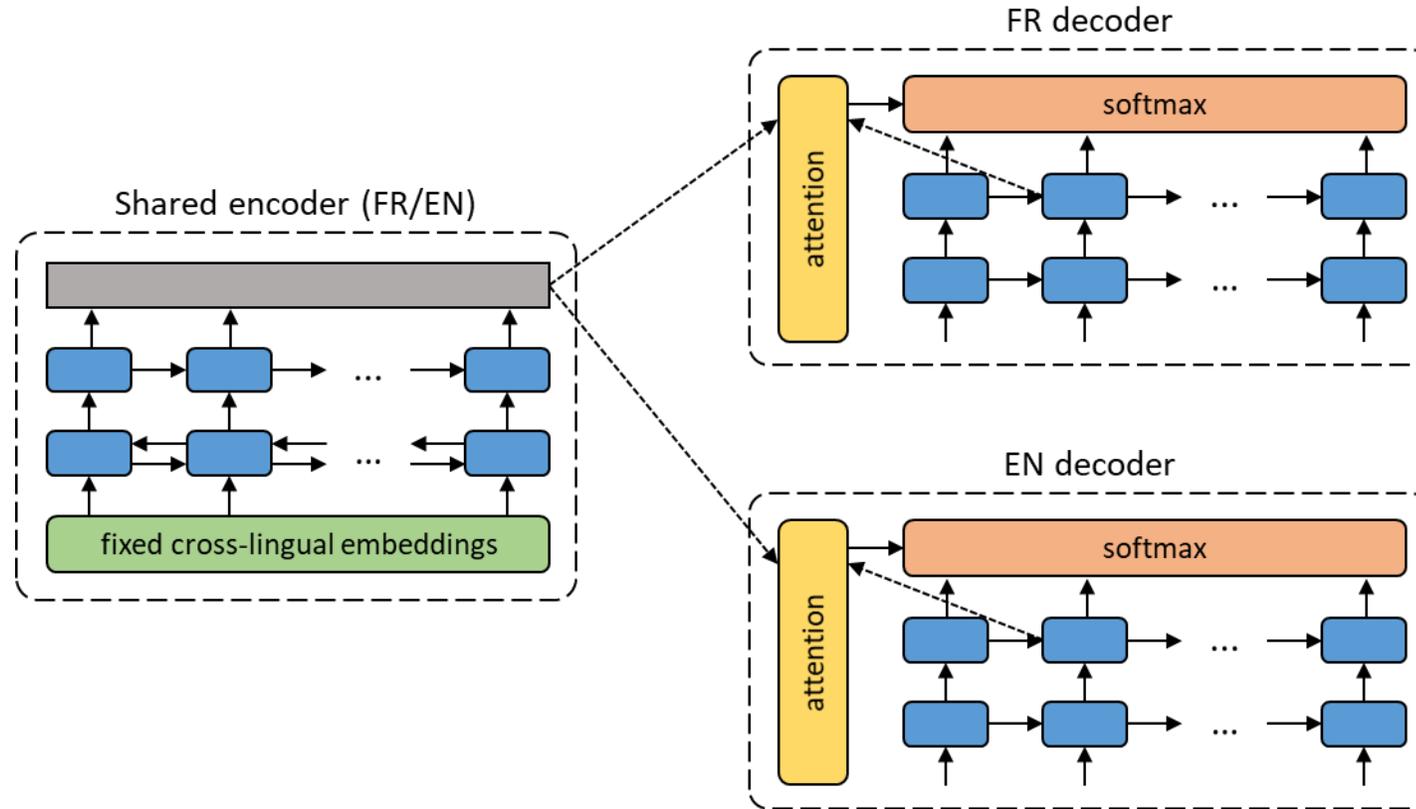


Unsupervised neural machine translation

Training

- Supervised

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

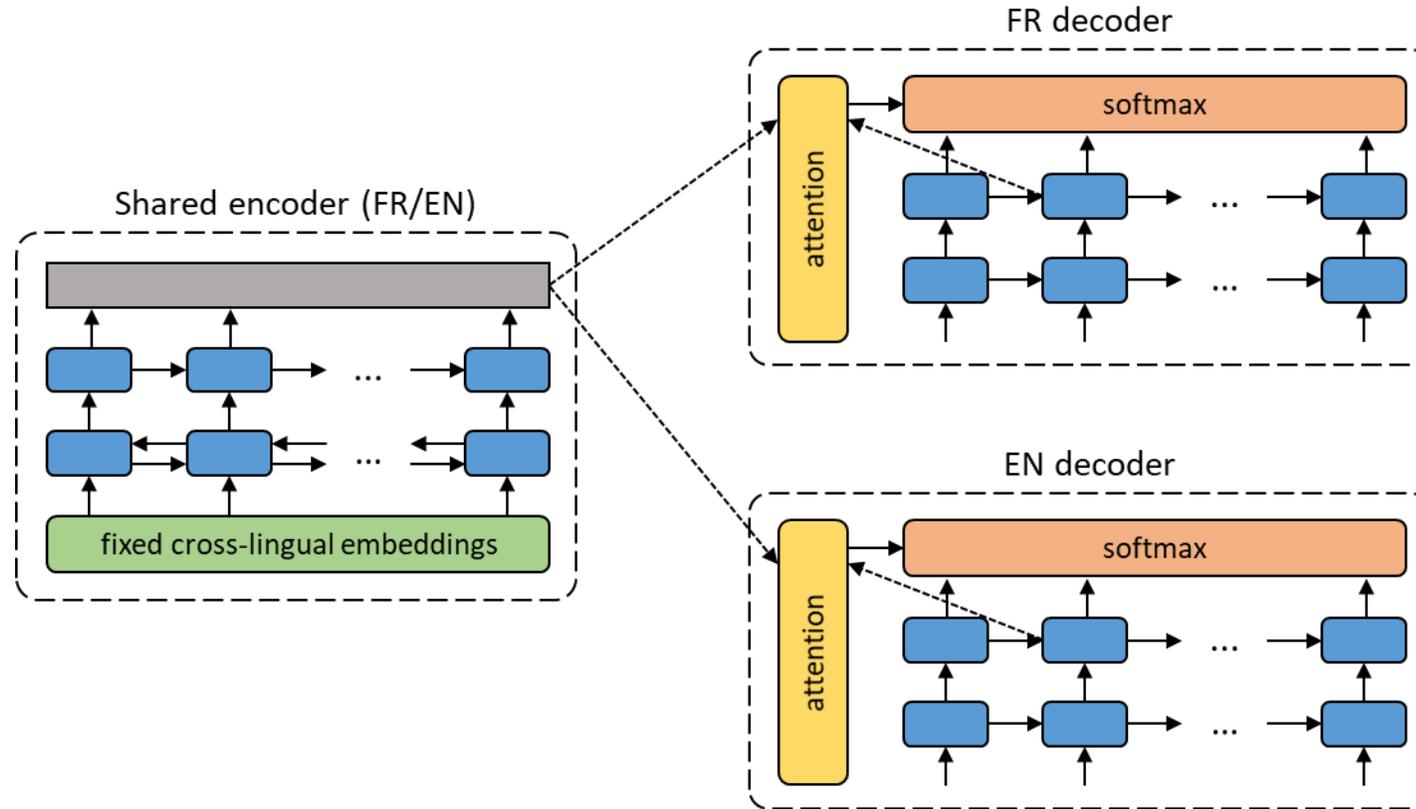


Unsupervised neural machine translation

Training

— Supervised

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

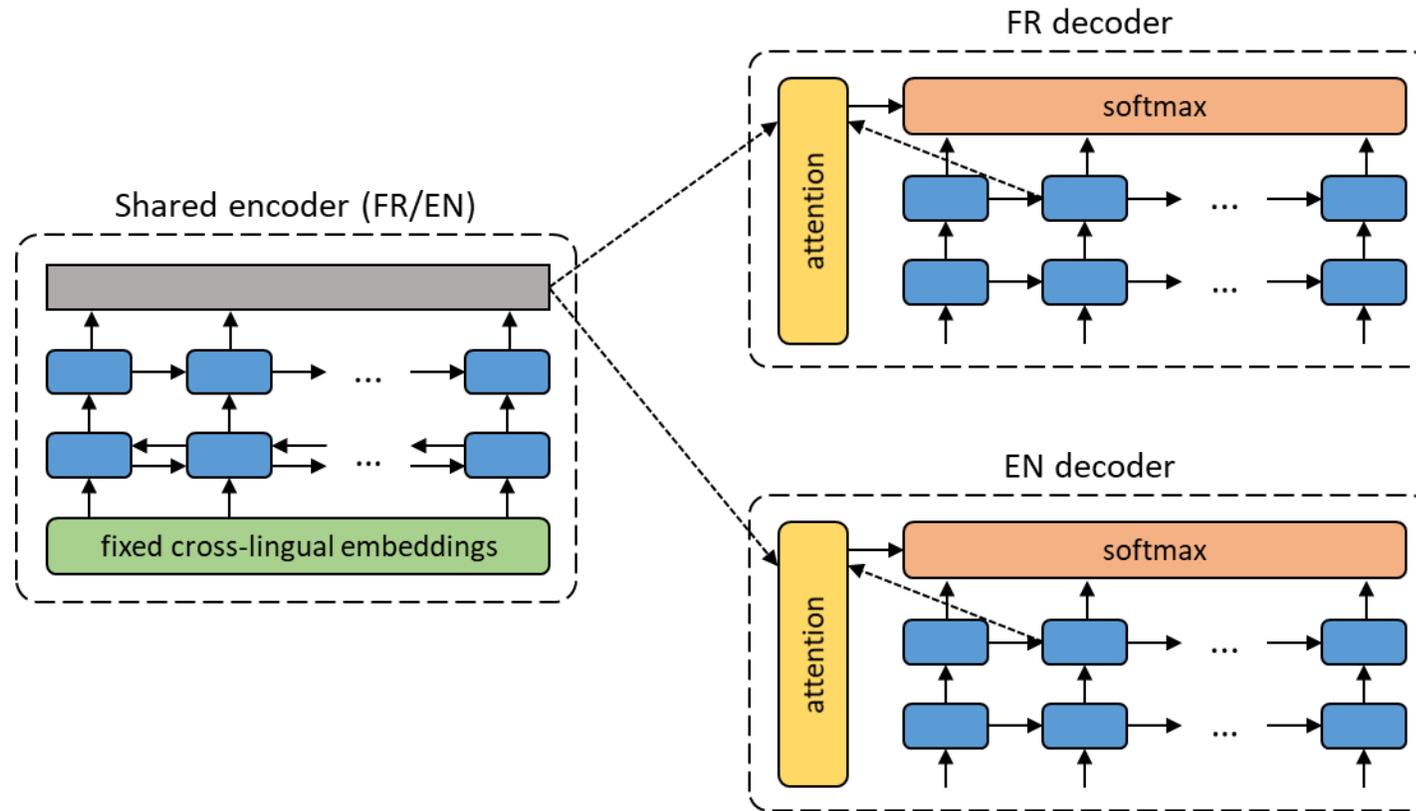


Unsupervised neural machine translation

Training

— Supervised

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



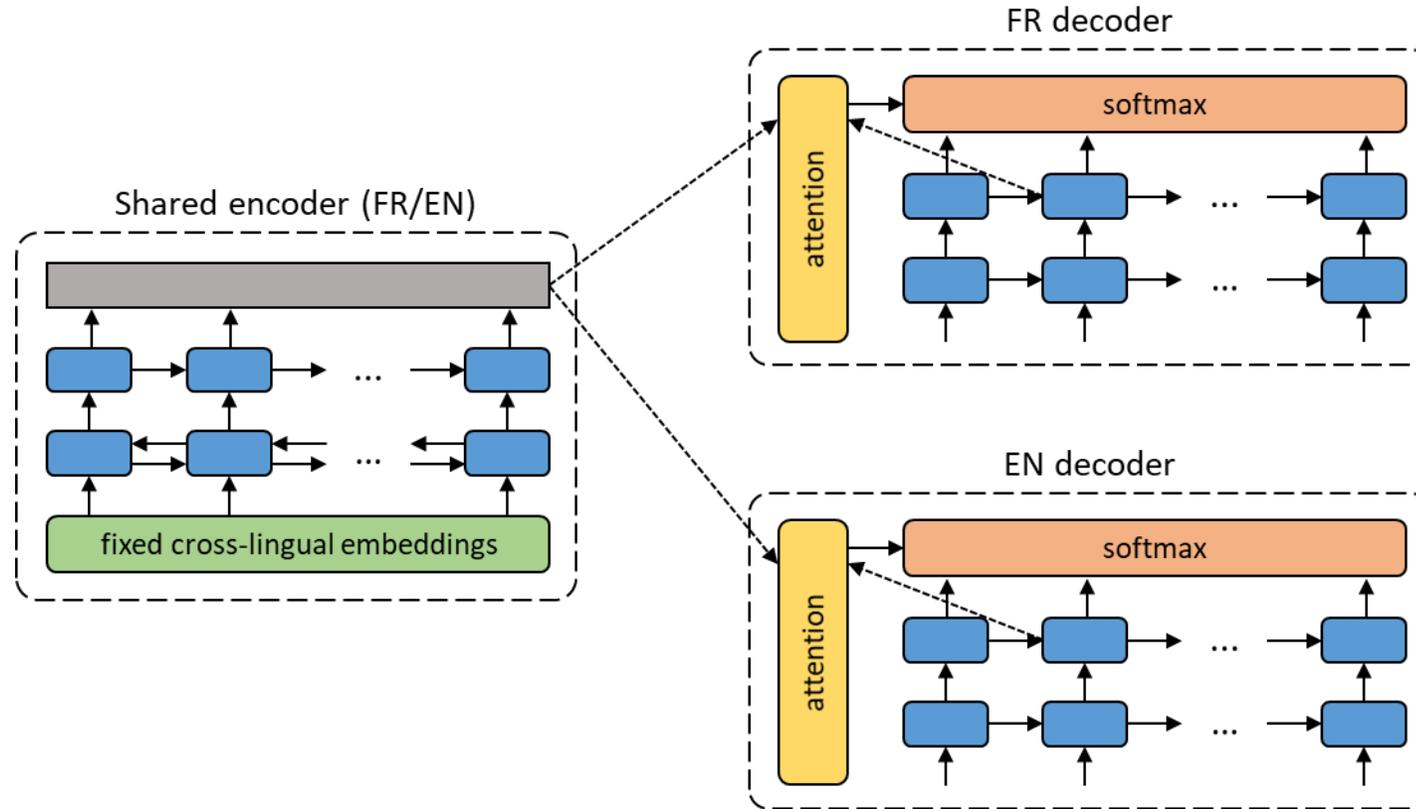
Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

- ~~— Supervised~~
- Denoising

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

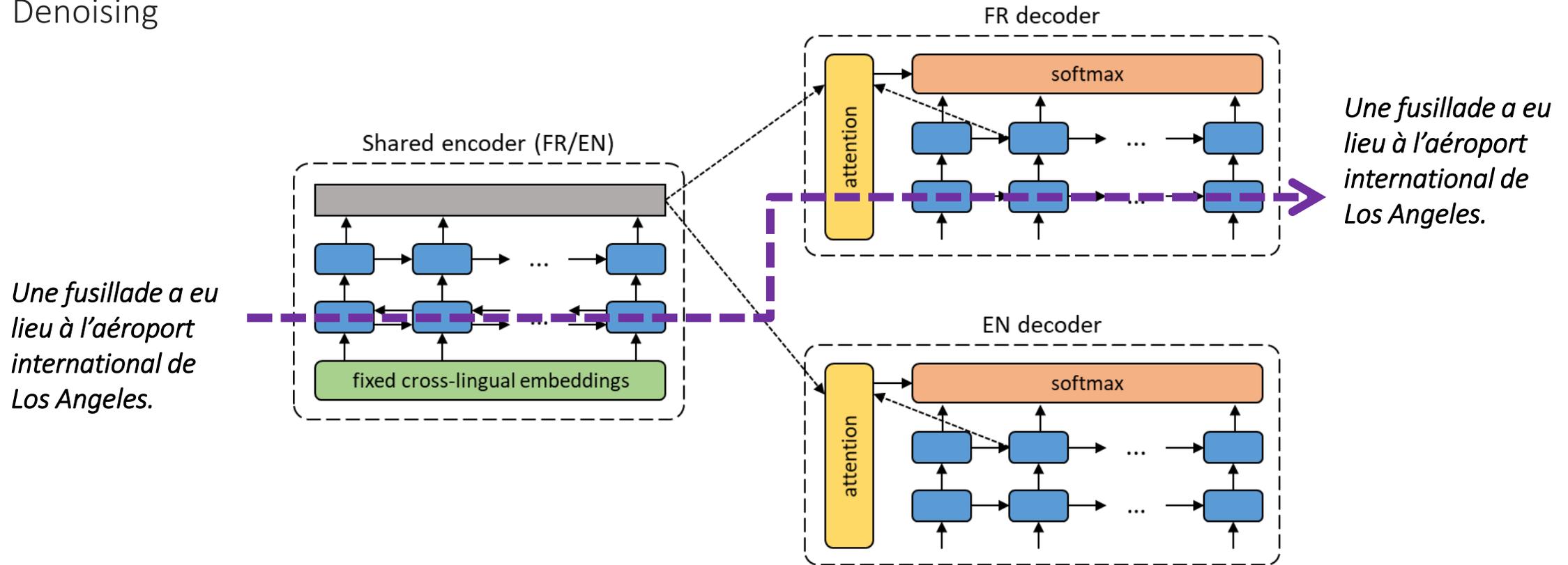


Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

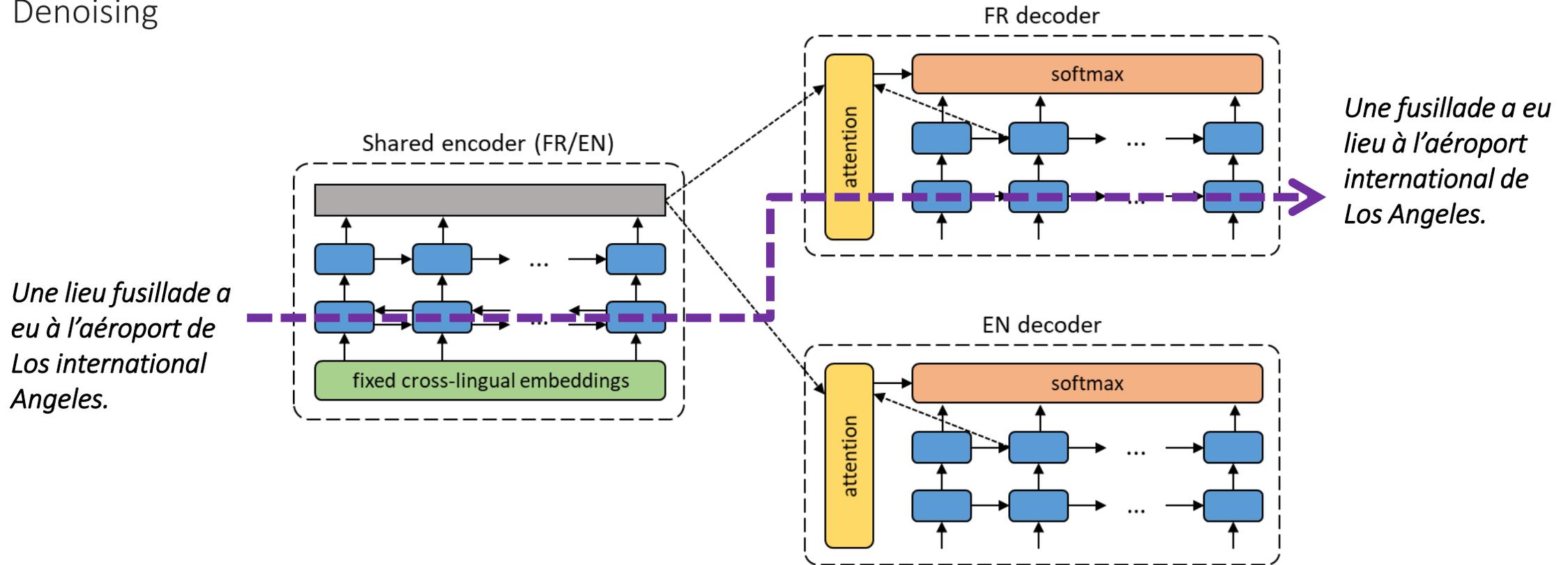
- ~~— Supervised~~
- Denoising



Unsupervised neural machine translation

Training

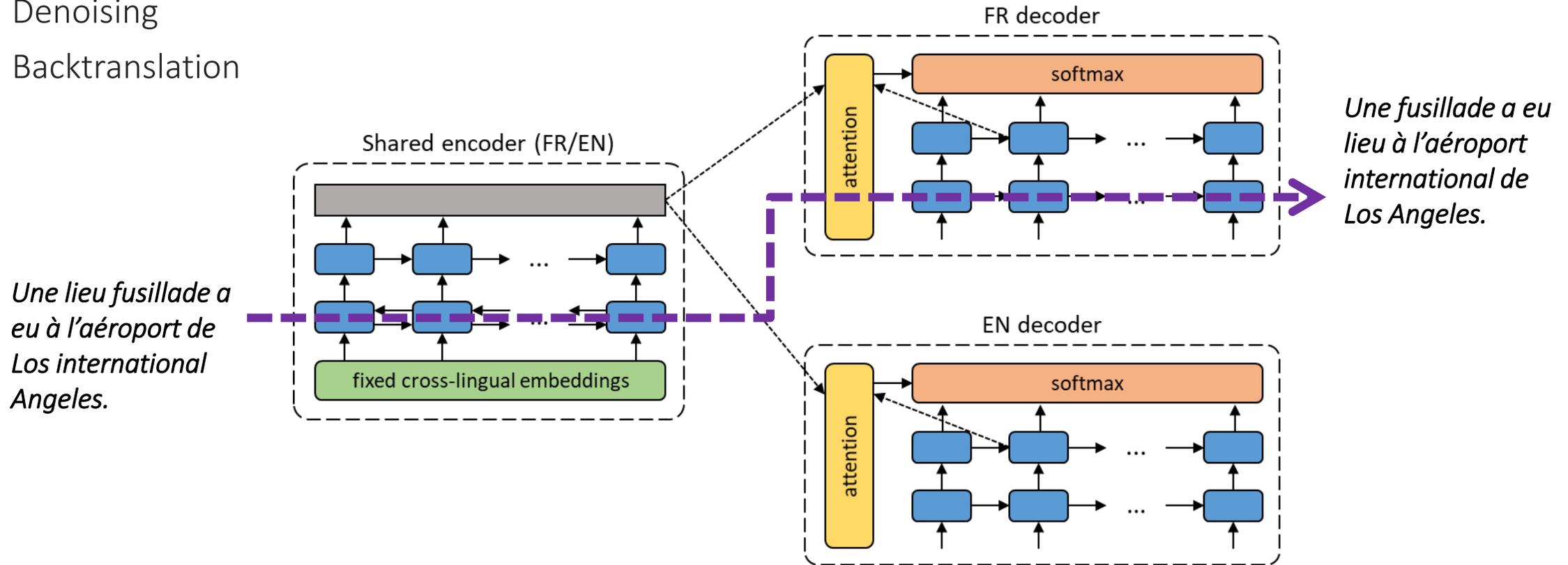
- ~~— Supervised~~
- Denoising



Unsupervised neural machine translation

Training

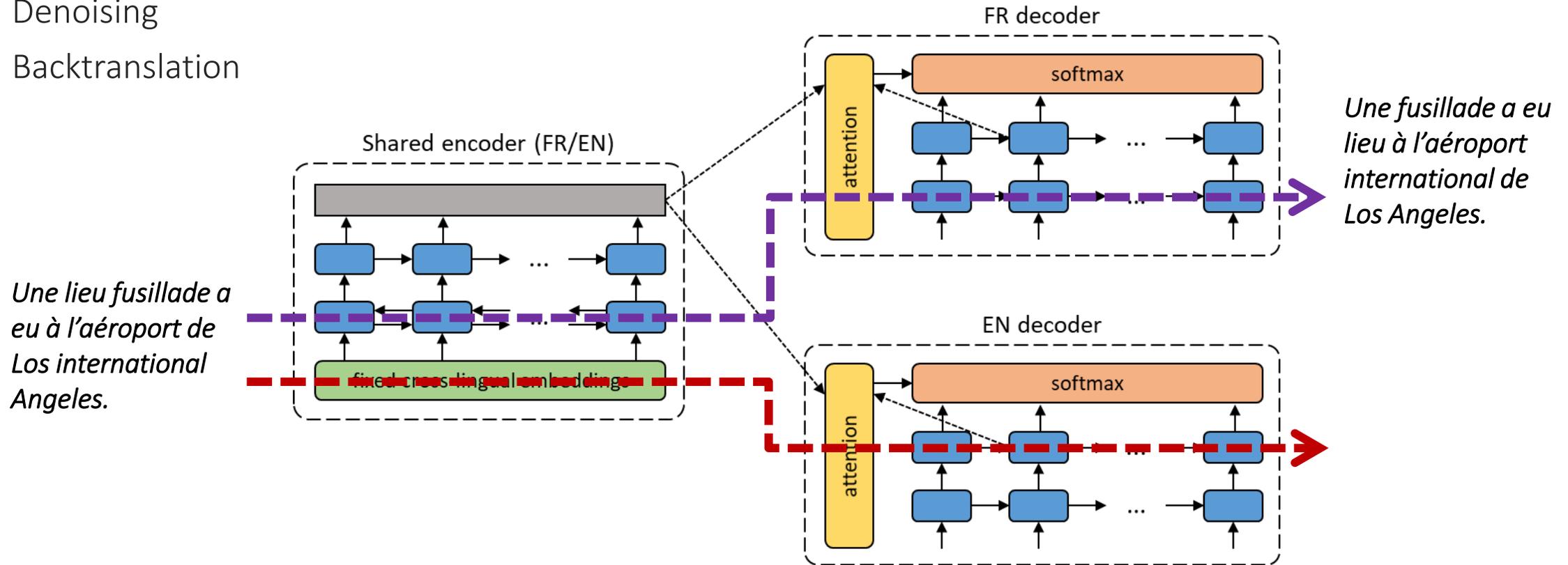
- ~~— Supervised~~
- Denoising
- Backtranslation



Unsupervised neural machine translation

Training

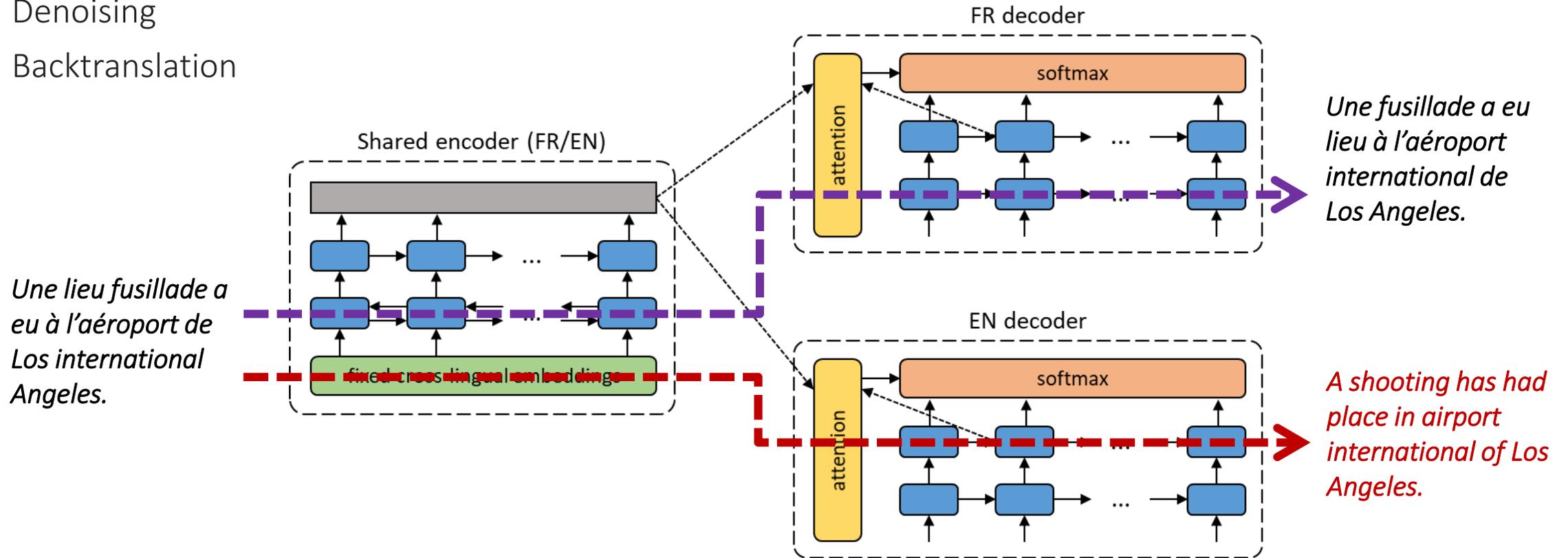
- ~~— Supervised~~
- Denoising
- Backtranslation



Unsupervised neural machine translation

Training

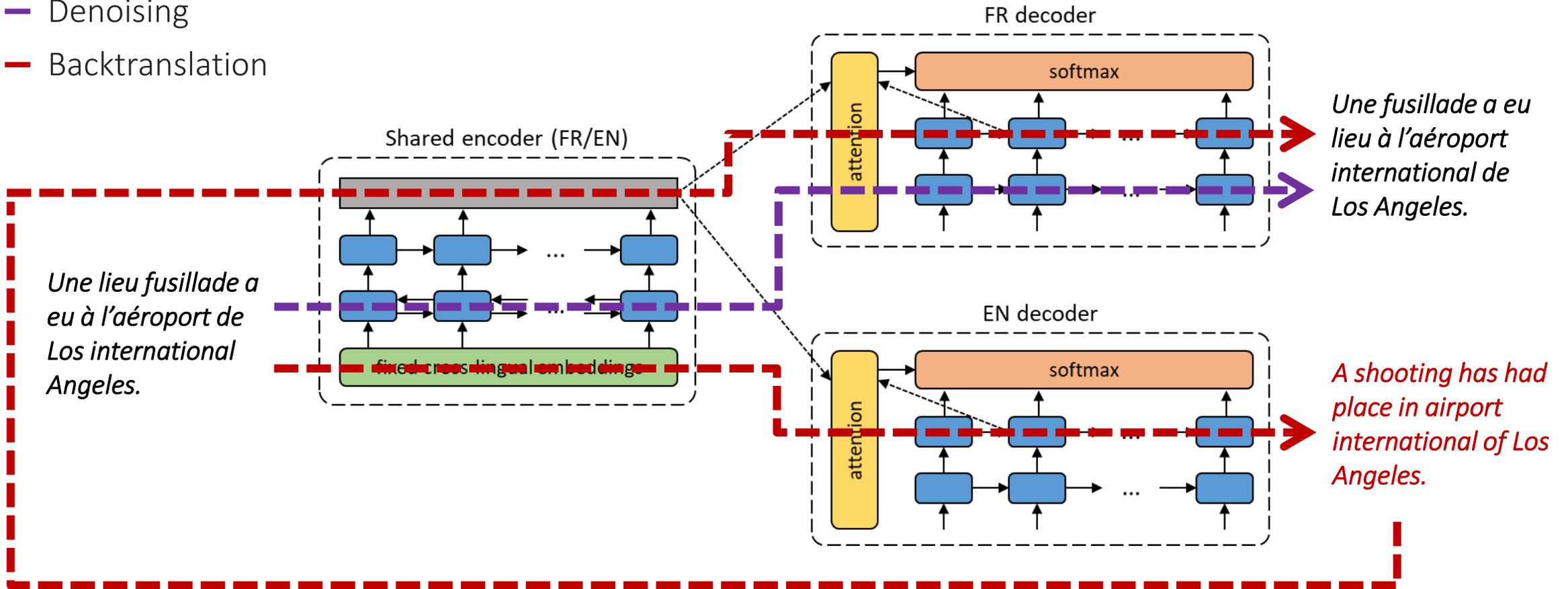
- ~~— Supervised~~
- Denoising
- Backtranslation



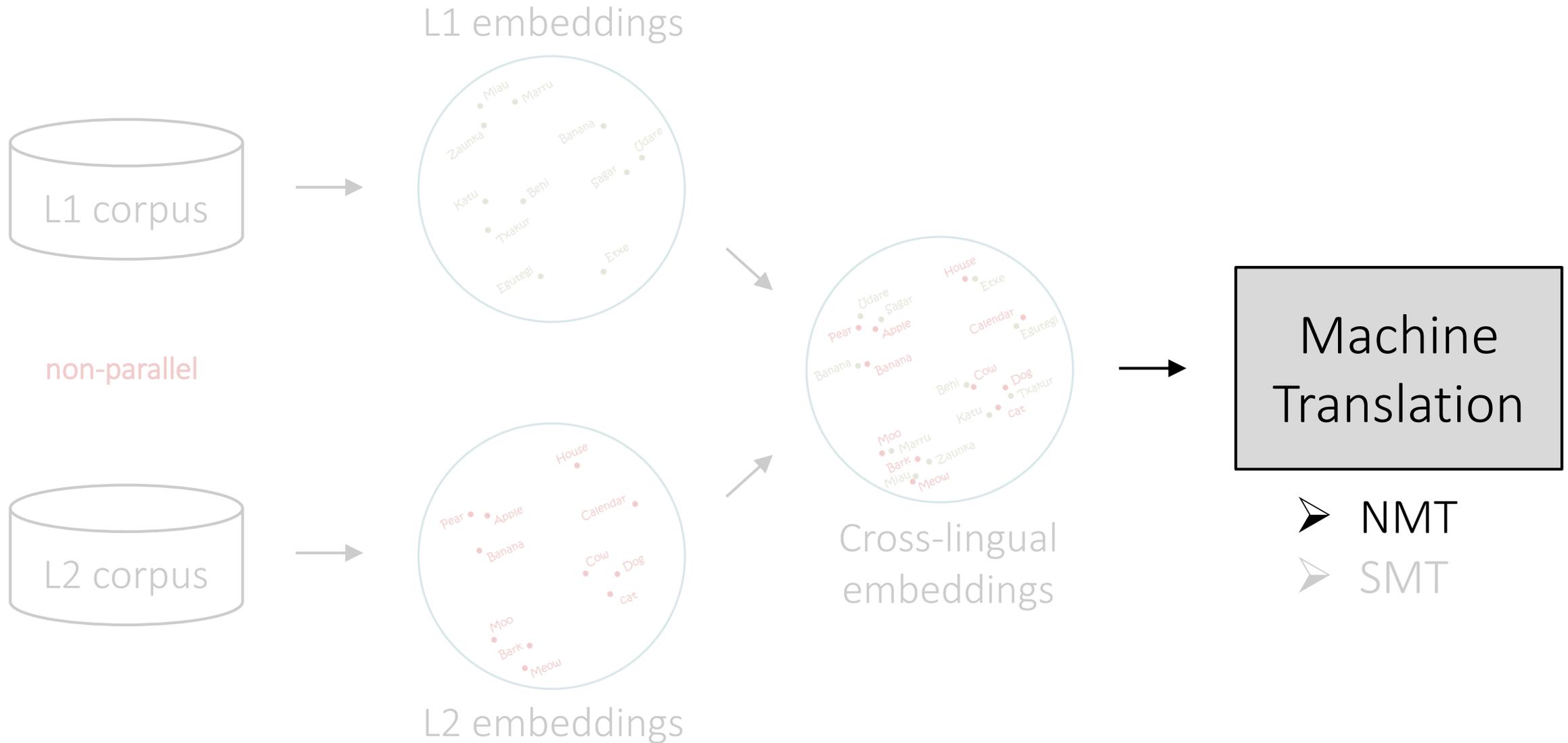
Unsupervised neural machine translation

Training

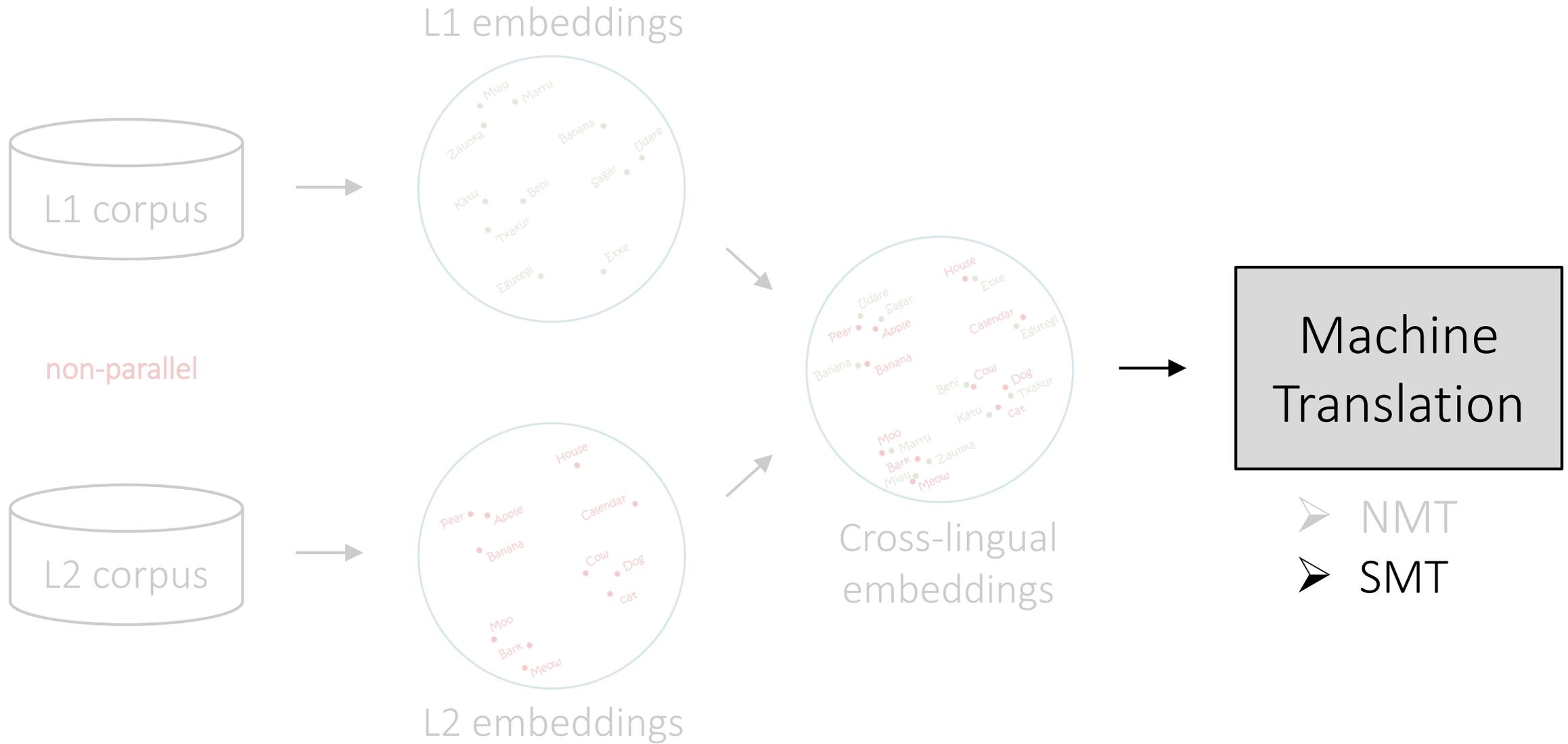
- ~~Supervised~~
- Denoising
- Backtranslation

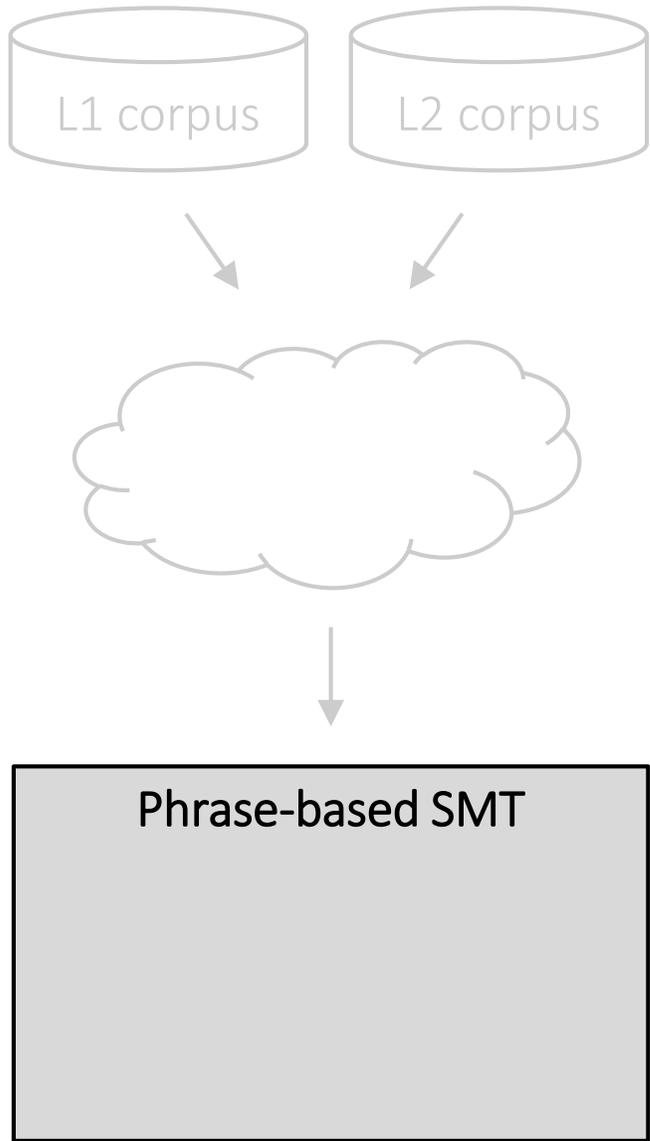


Overview

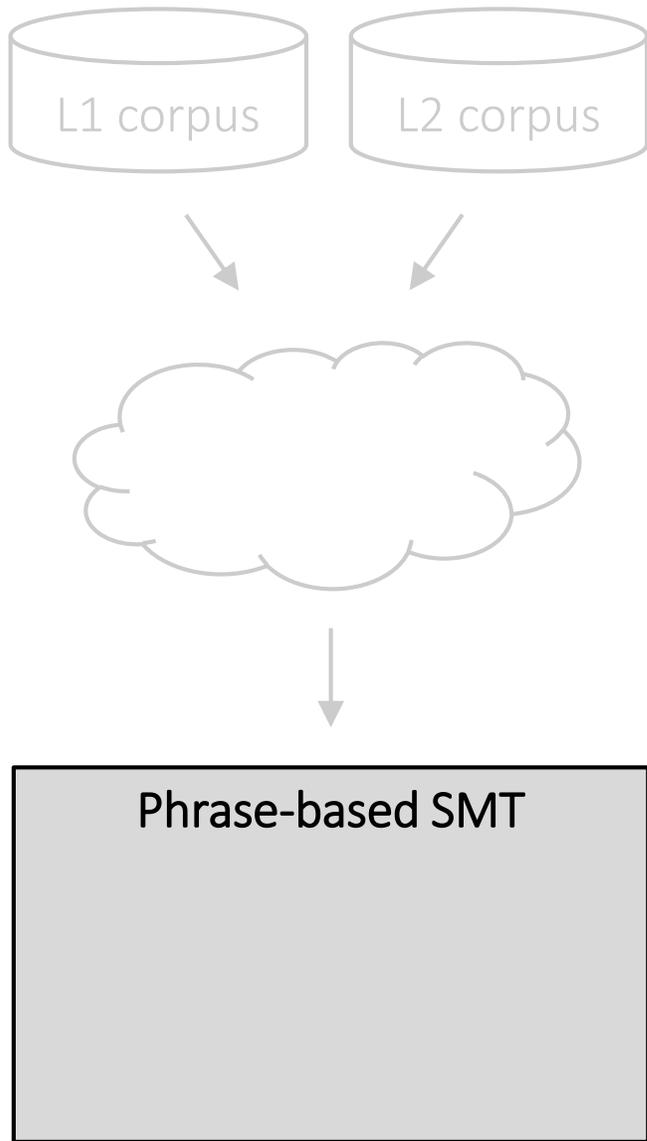


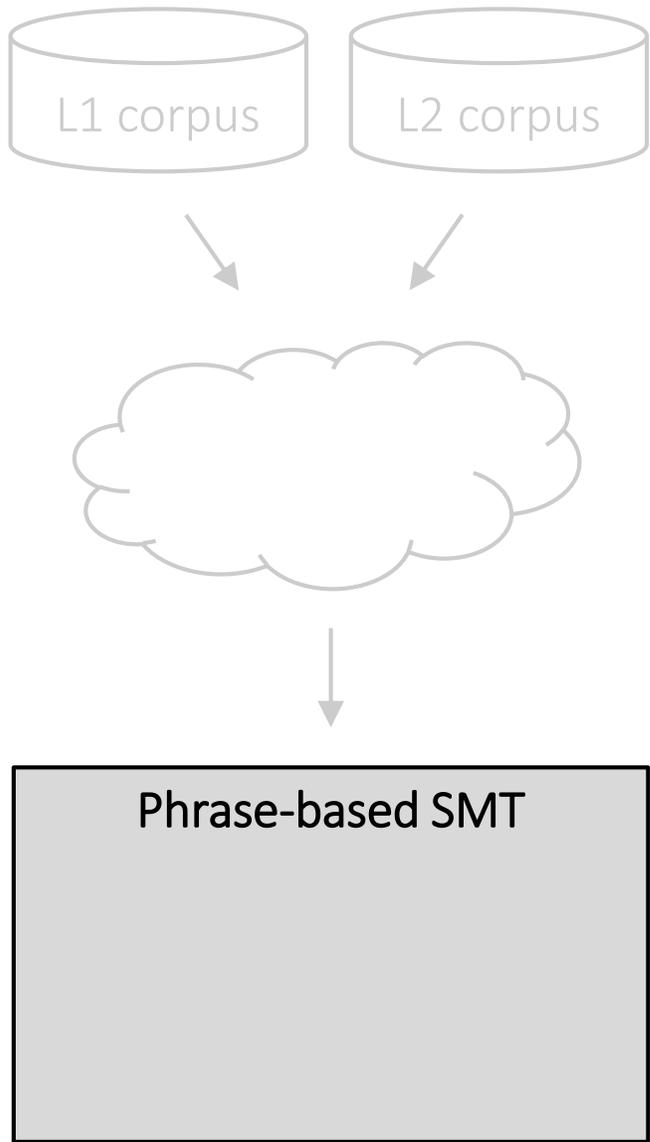
Overview





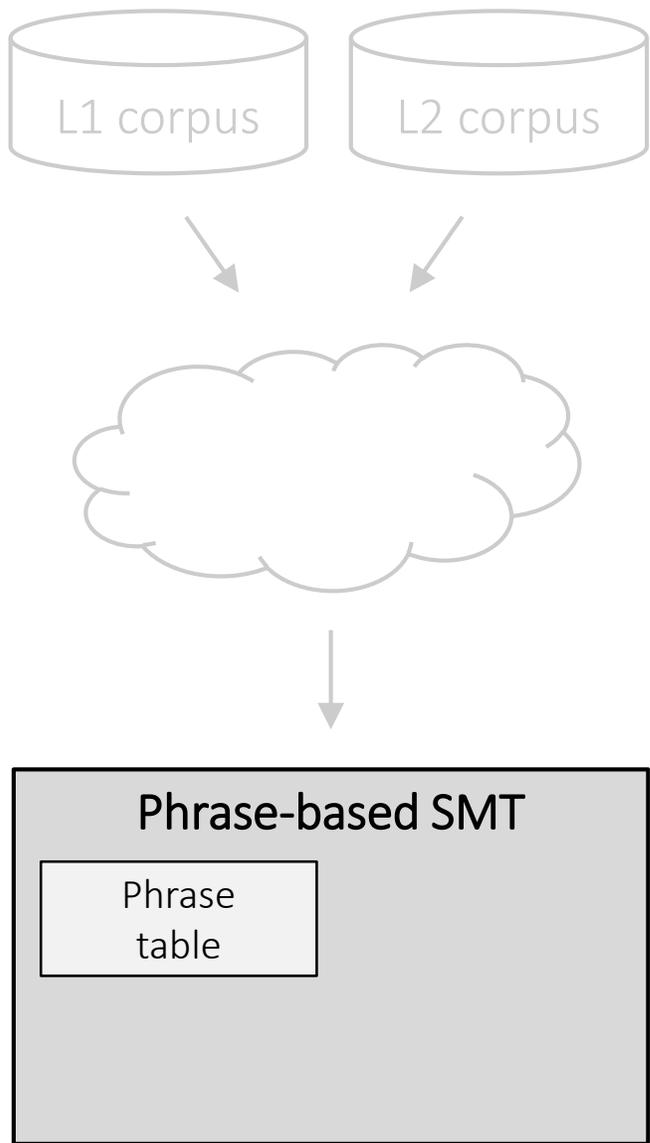
Phrase-based SMT





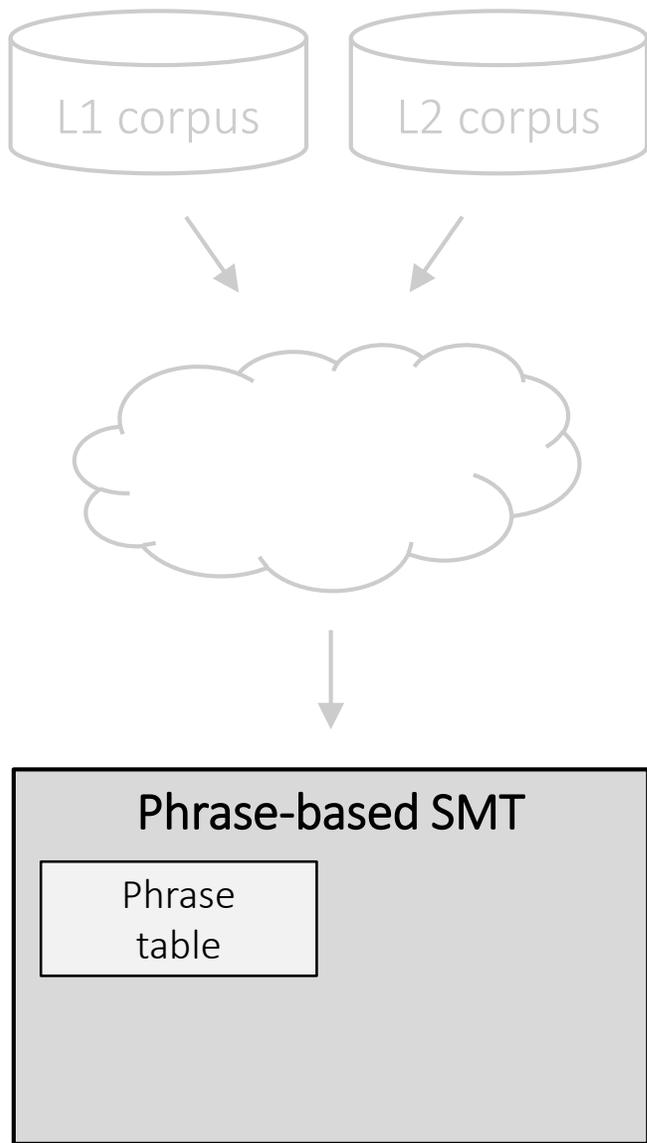
Phrase-based SMT

Log-linear model combining



Phrase-based SMT

Log-linear model combining
- Phrase table

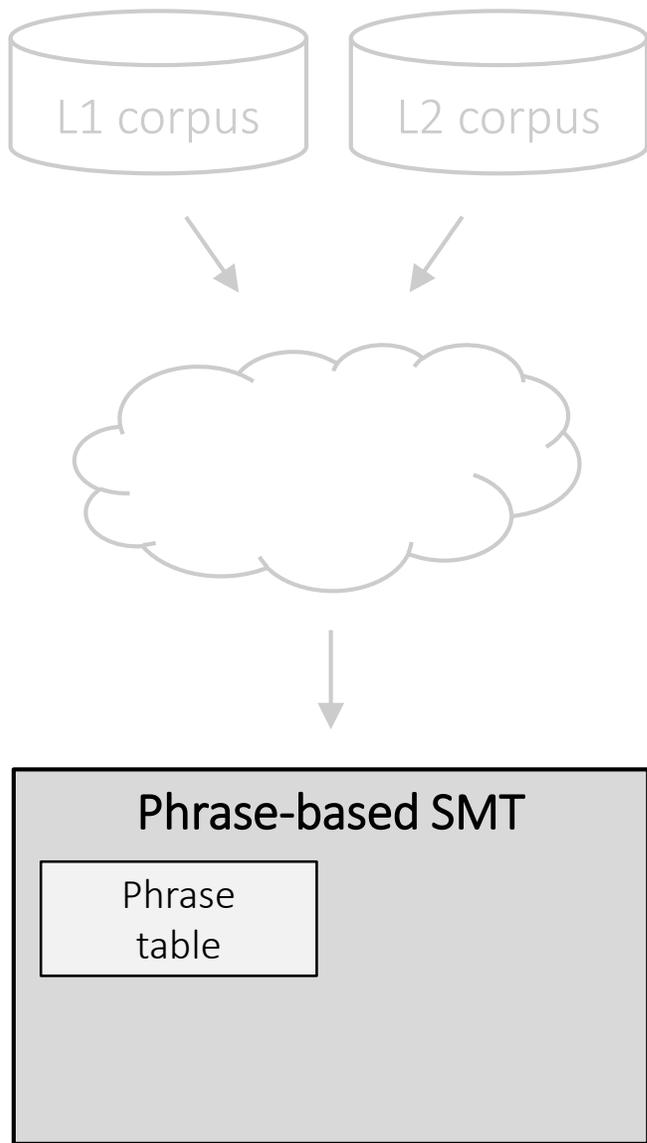


Phrase-based SMT

Log-linear model combining
- Phrase table

nire iritziz	in my opinion
nire iritziz	in my view
nire iritziz	I think
opari bat	a present
opari bat	one present
opari bat	a gift

⋮



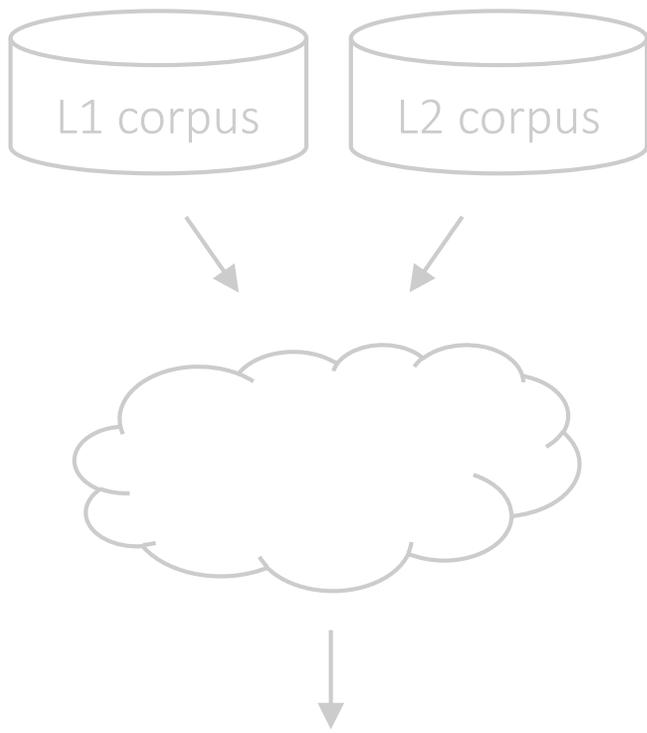
Phrase-based SMT

Log-linear model combining

- Phrase table
- Direct/inverse translation probabilities

		$\phi(\bar{f} \bar{e})$	$\phi(\bar{e} \bar{f})$
nire iritziz	in my opinion	0.54	0.63
nire iritziz	in my view	0.32	0.68
nire iritziz	I think	0.11	0.09
opari bat	a present	0.32	0.56
opari bat	one present	0.14	0.73
opari bat	a gift	0.11	0.49

⋮



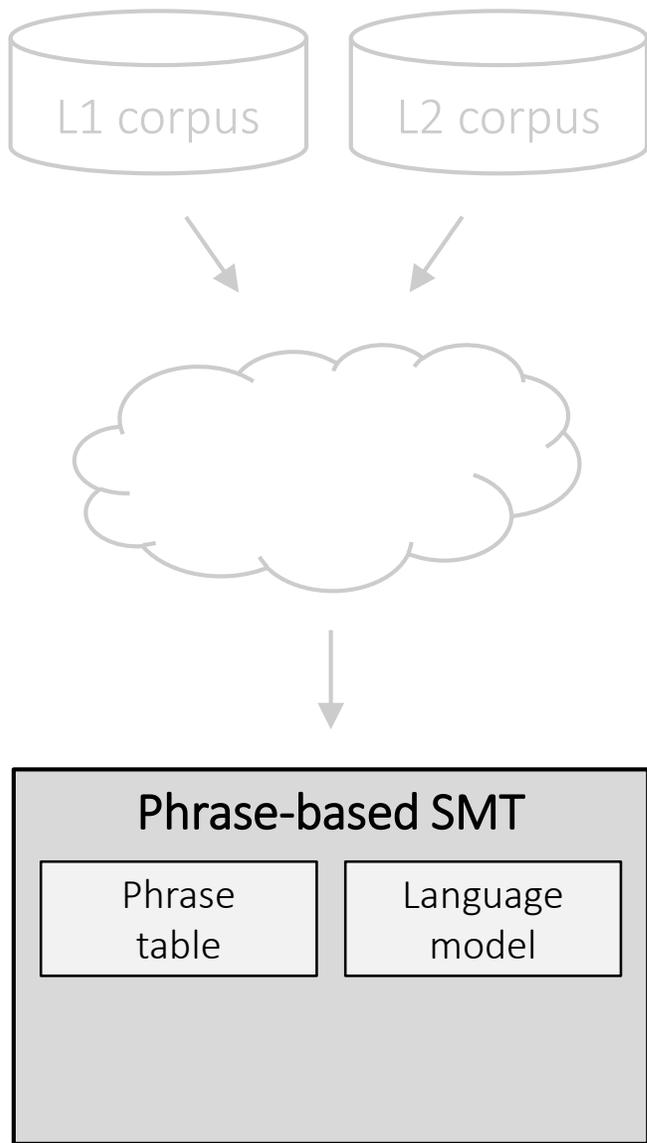
Phrase-based SMT

Log-linear model combining

- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings

		$\phi(\bar{f} \bar{e})$	$\phi(\bar{e} \bar{f})$	$\text{lex}(\bar{f} \bar{e})$	$\text{lex}(\bar{e} \bar{f})$
nire iritziz	in my opinion	0.54	0.63	0.12	0.15
nire iritziz	in my view	0.32	0.68	0.09	0.16
nire iritziz	I think	0.11	0.09	0.04	0.02
opari bat	a present	0.32	0.56	0.21	0.22
opari bat	one present	0.14	0.73	0.18	0.32
opari bat	a gift	0.11	0.49	0.11	0.13

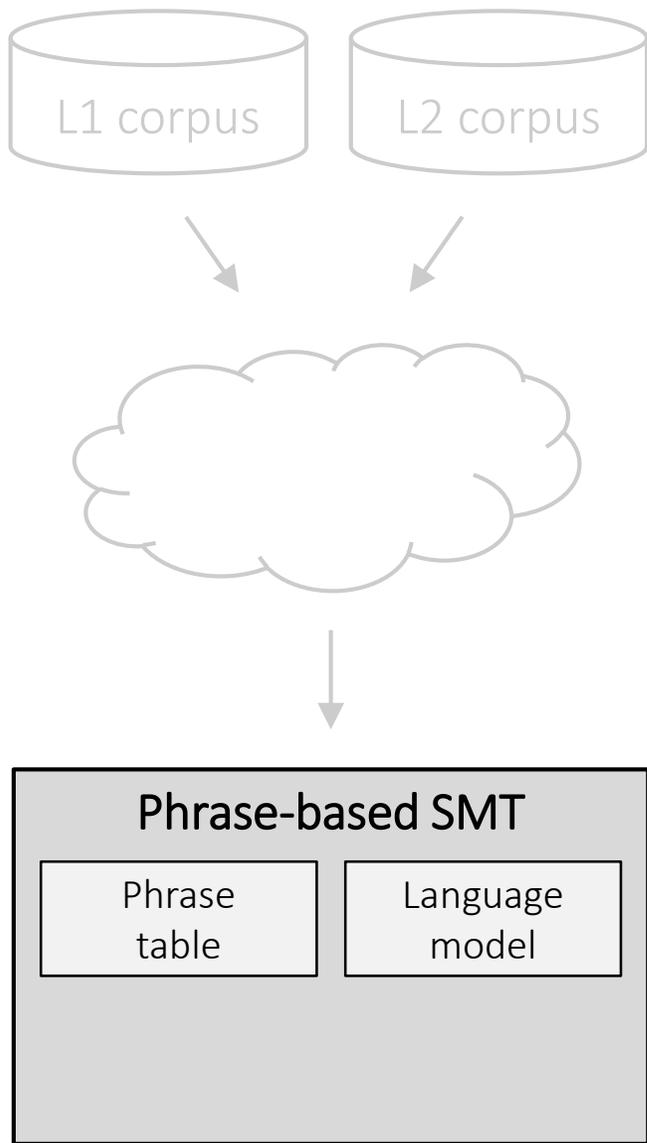
⋮



Phrase-based SMT

Log-linear model combining

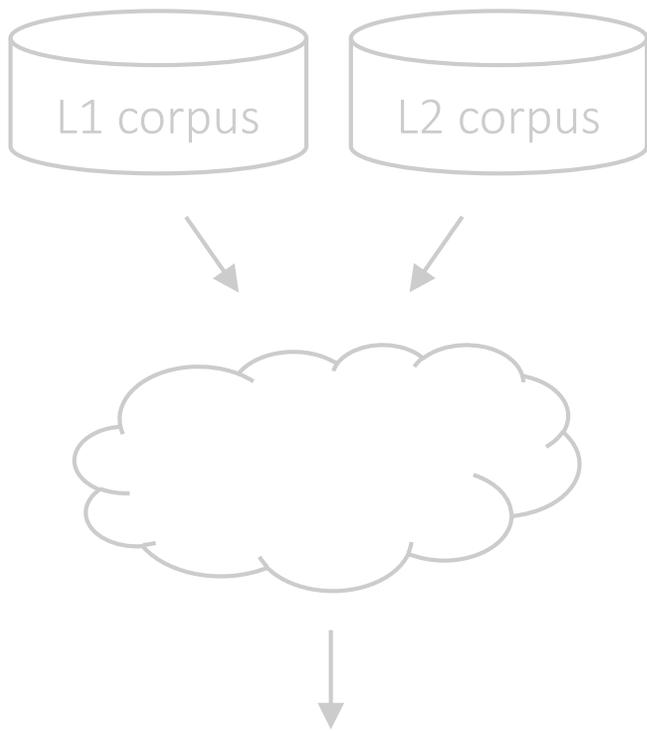
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model



Phrase-based SMT

Log-linear model combining

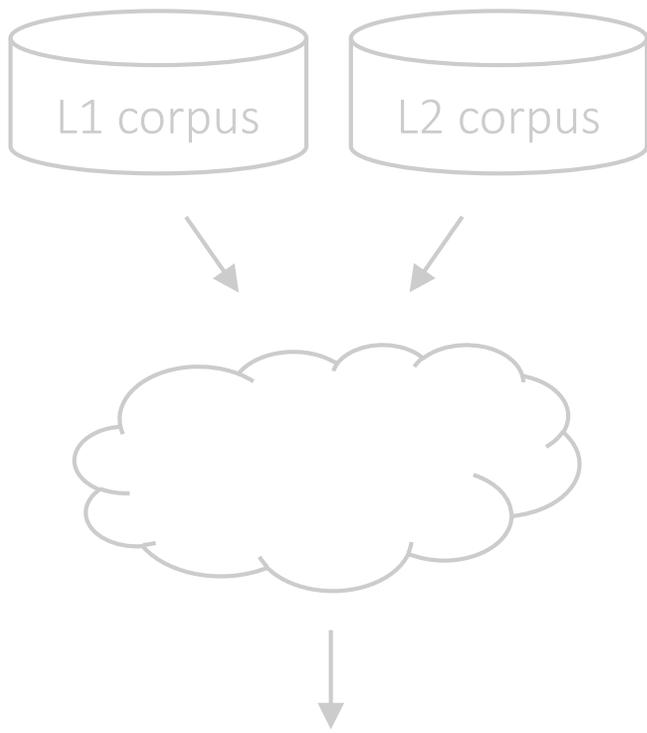
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing



Phrase-based SMT

Log-linear model combining

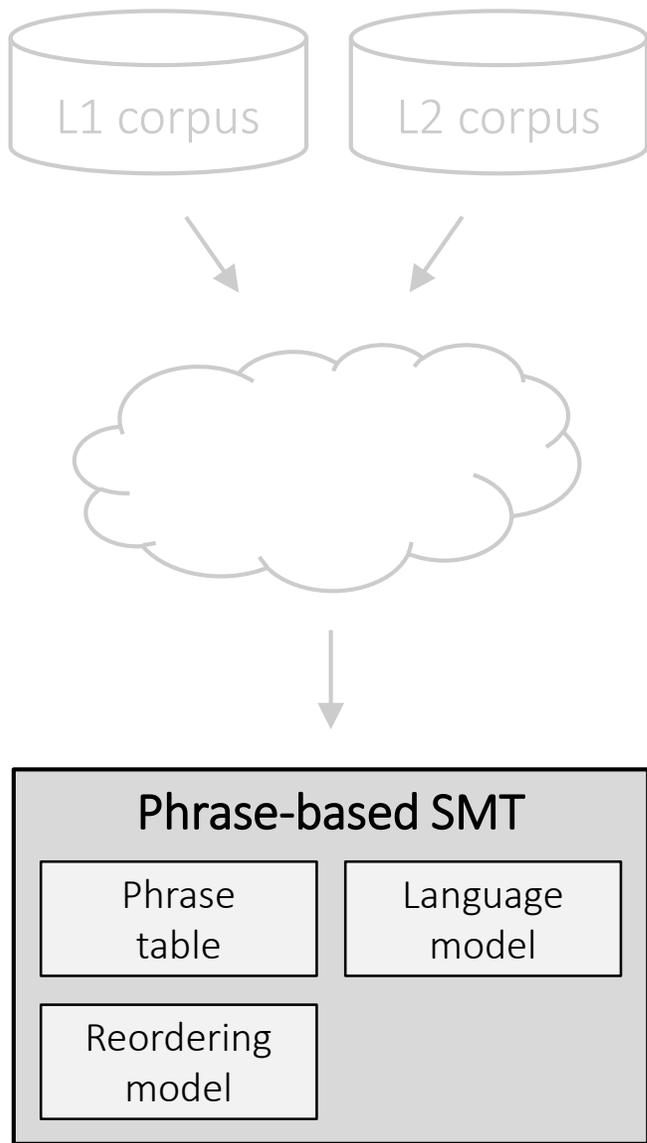
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model



Phrase-based SMT

Log-linear model combining

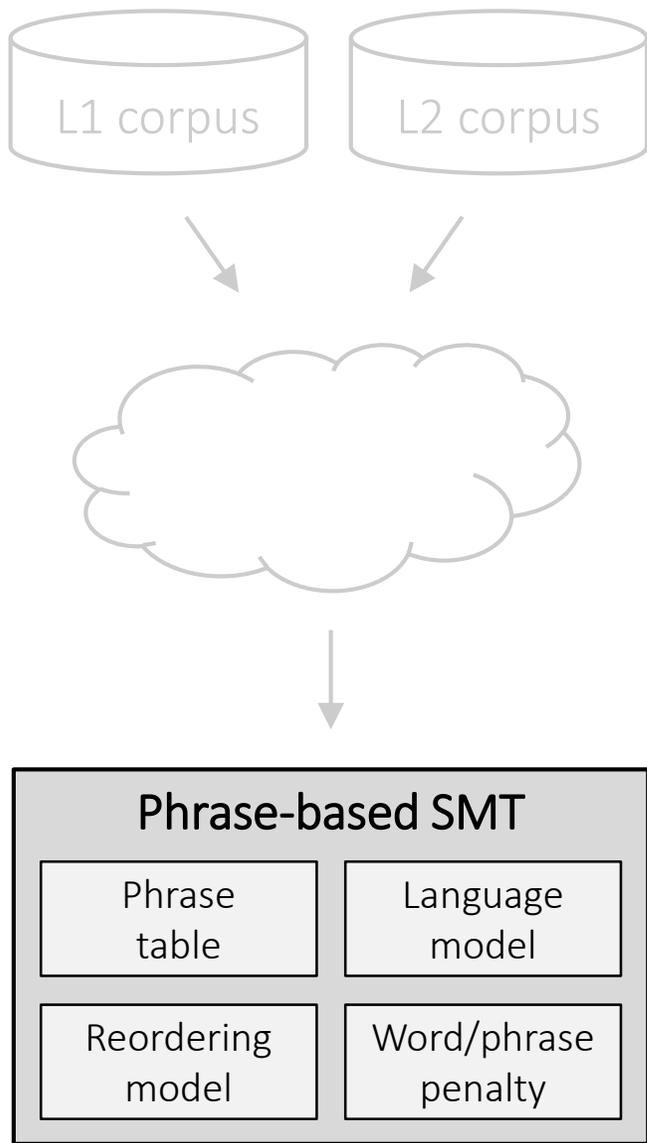
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)



Phrase-based SMT

Log-linear model combining

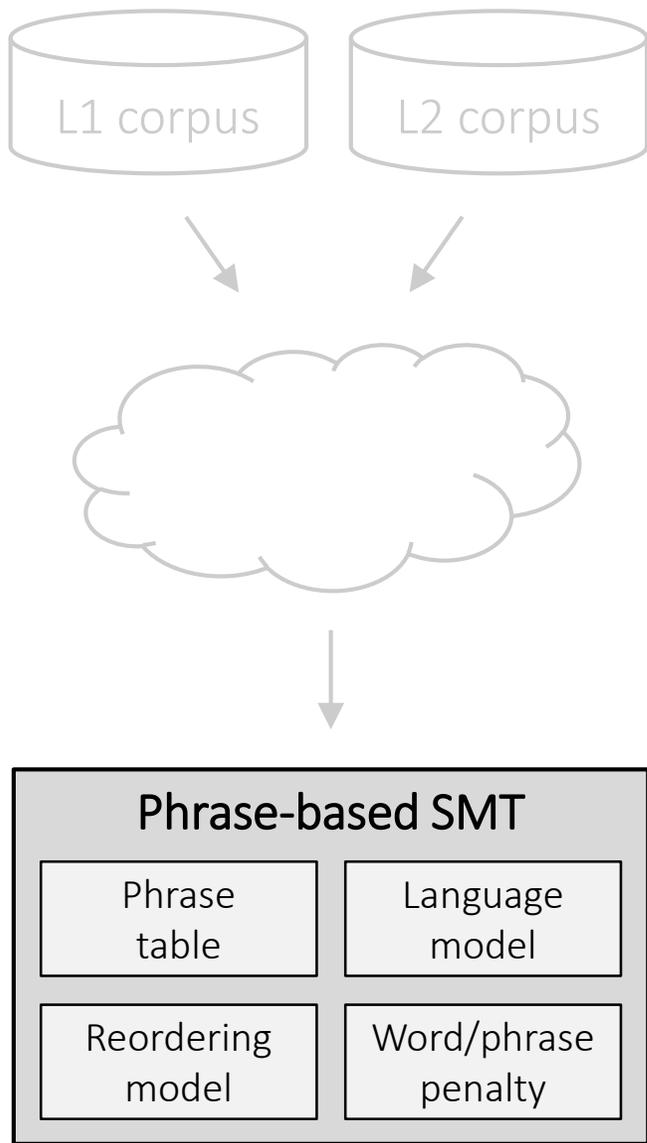
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model



Phrase-based SMT

Log-linear model combining

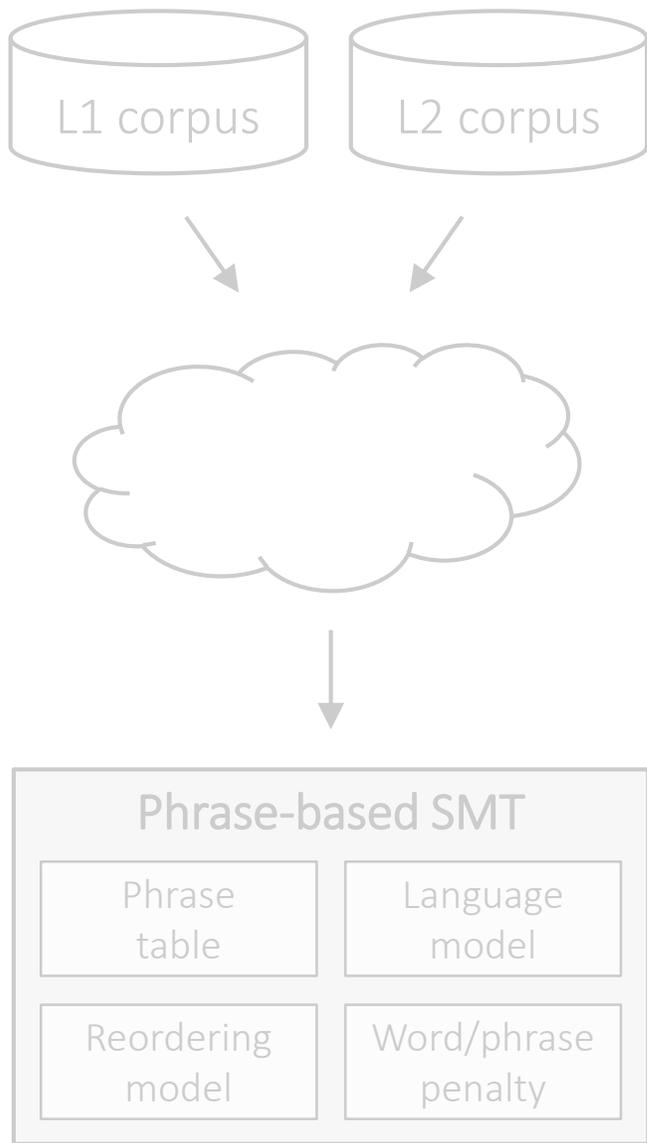
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty



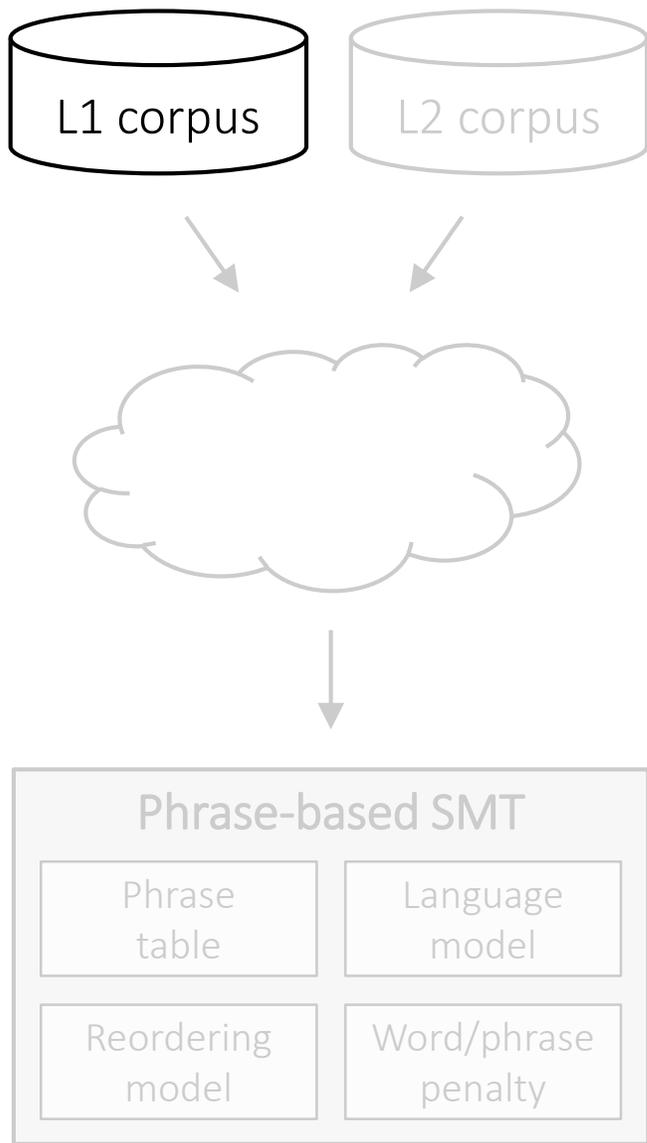
Phrase-based SMT

Log-linear model combining

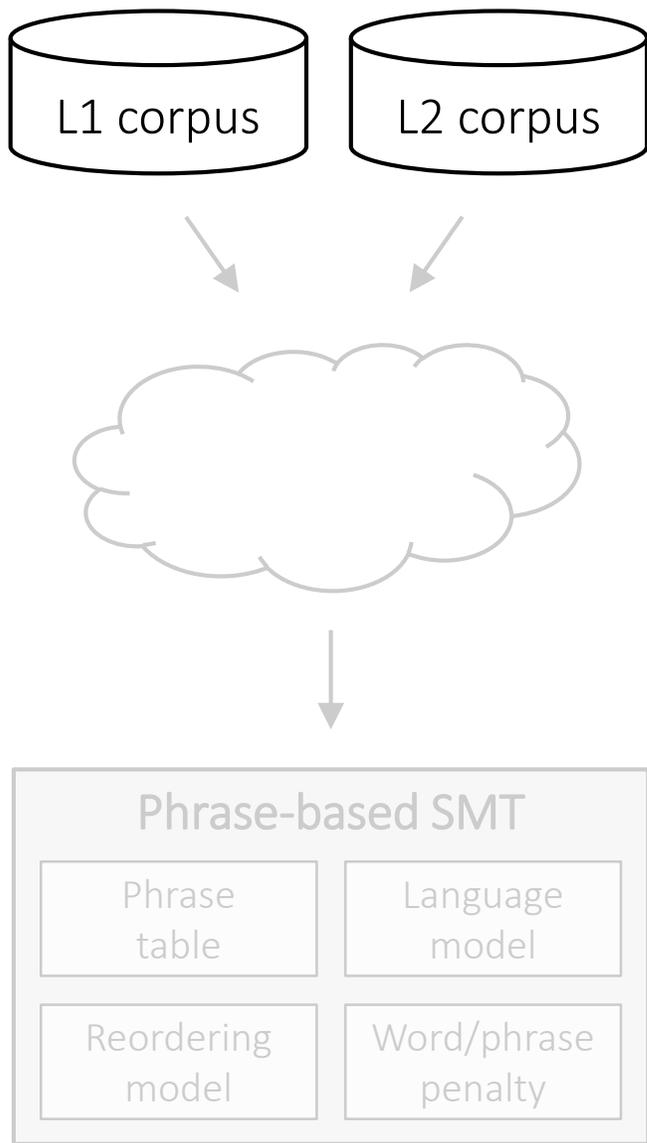
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output



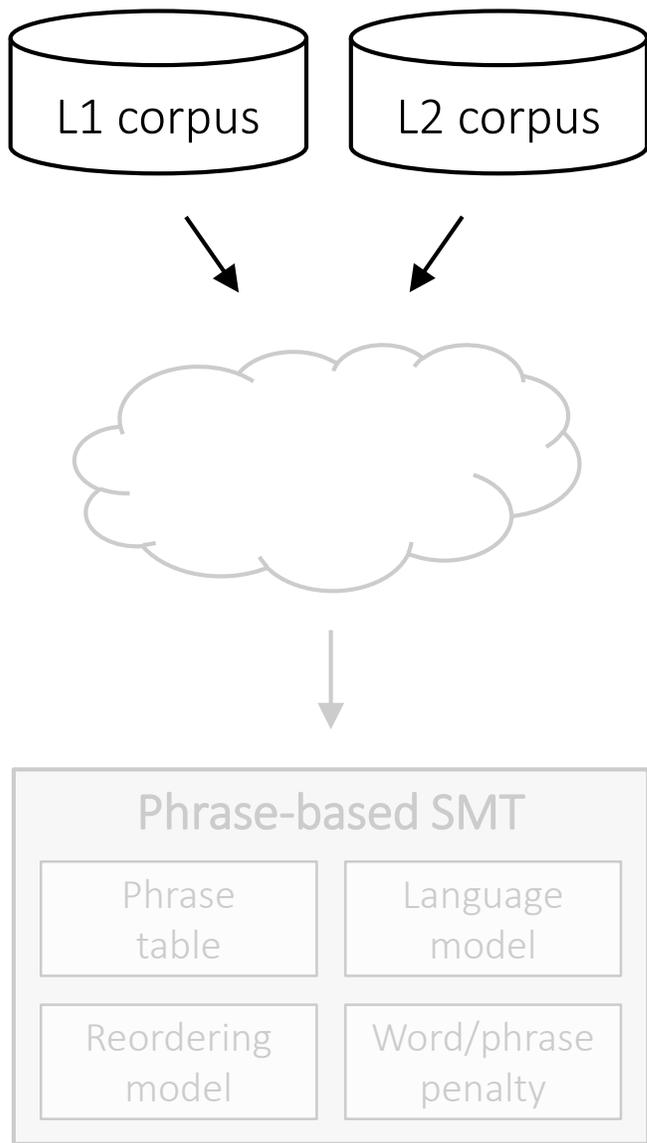
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output



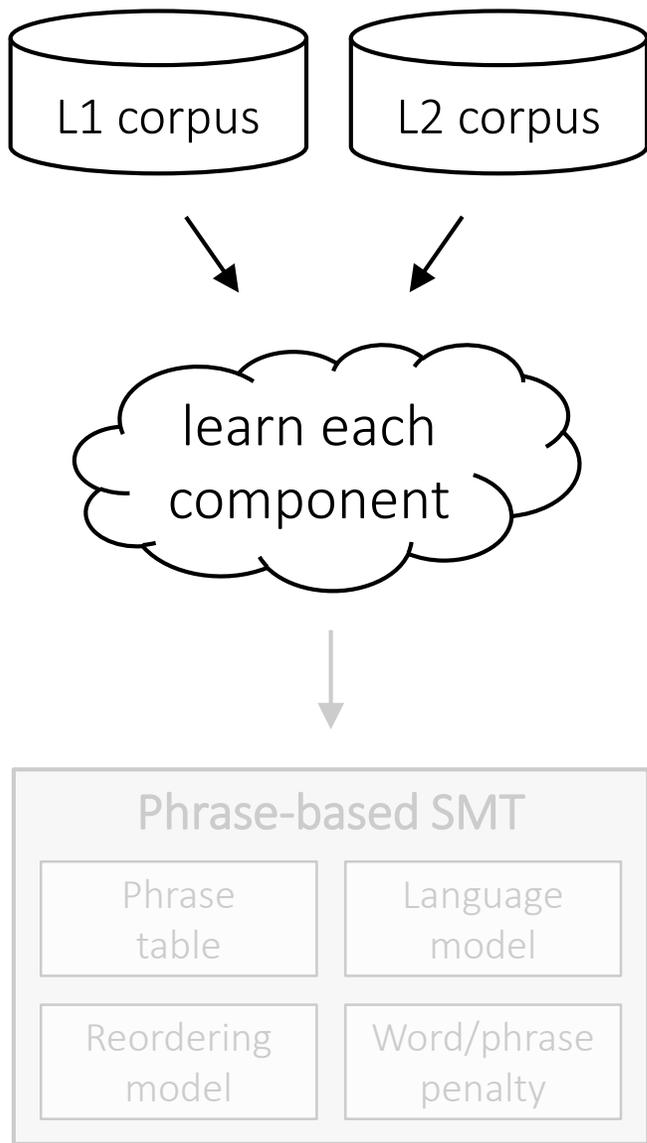
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output



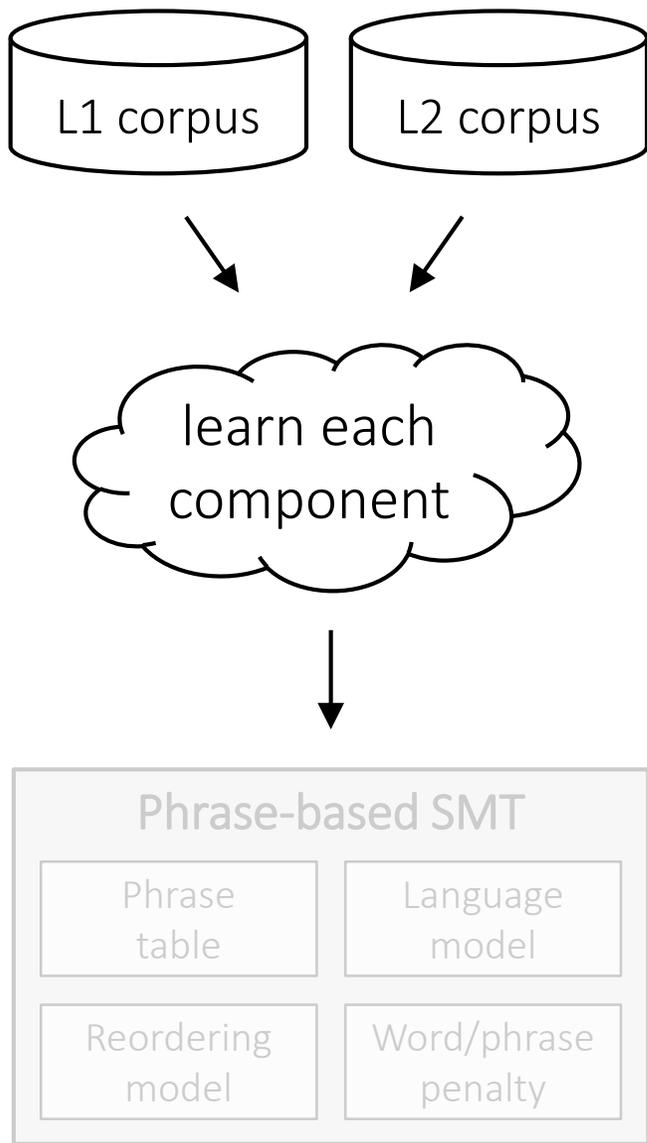
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output



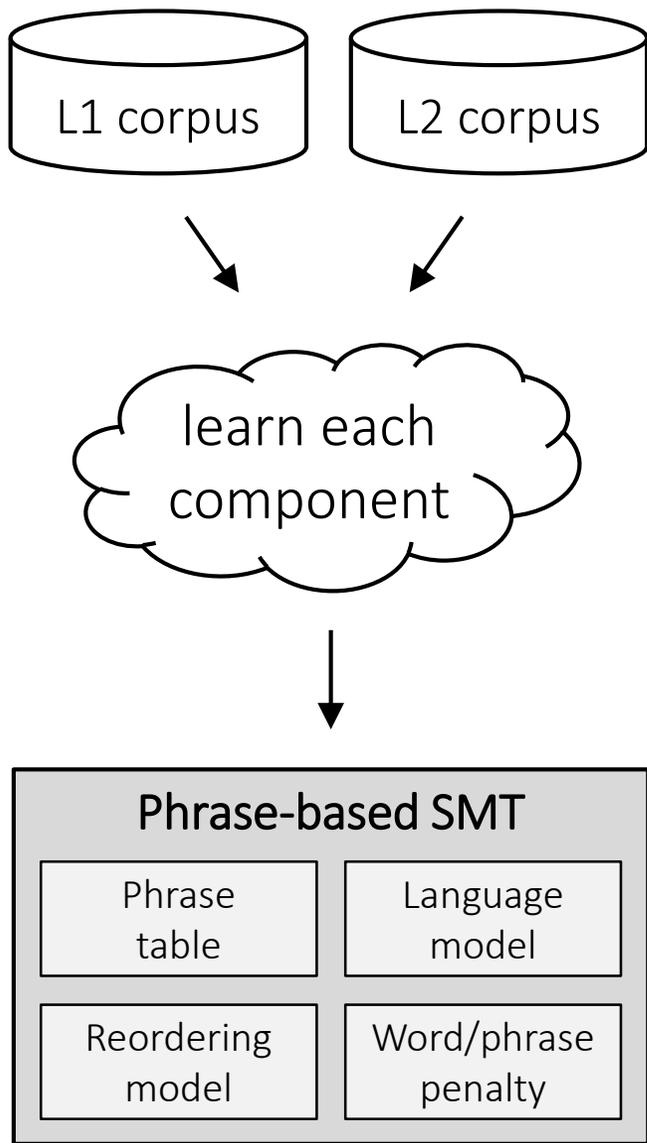
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output



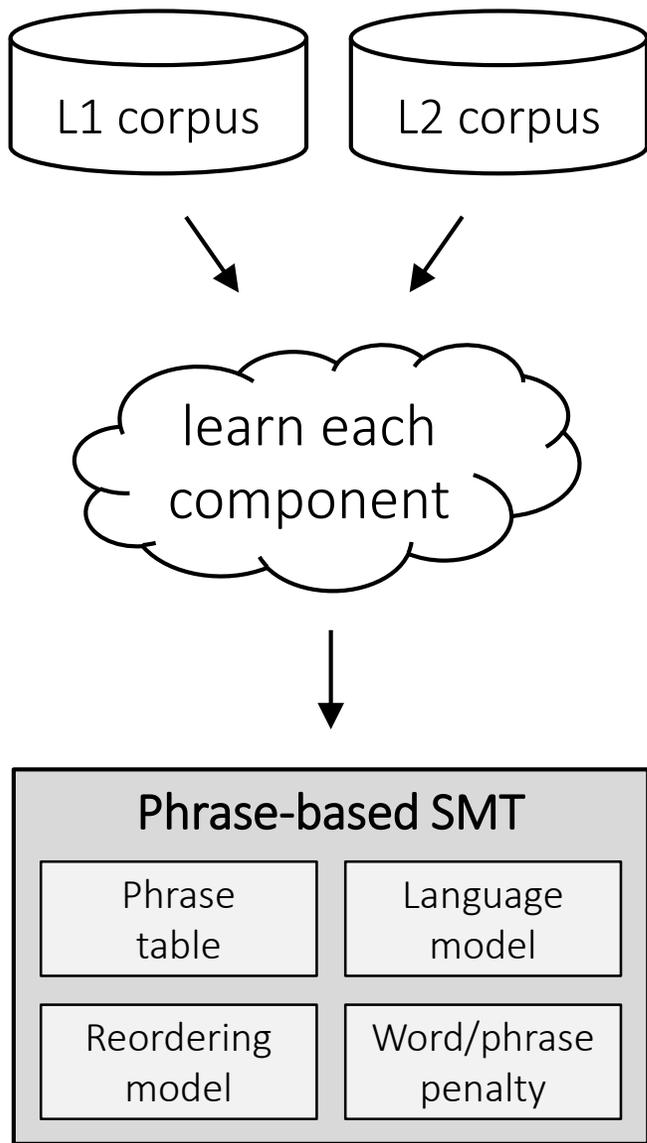
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output



- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output

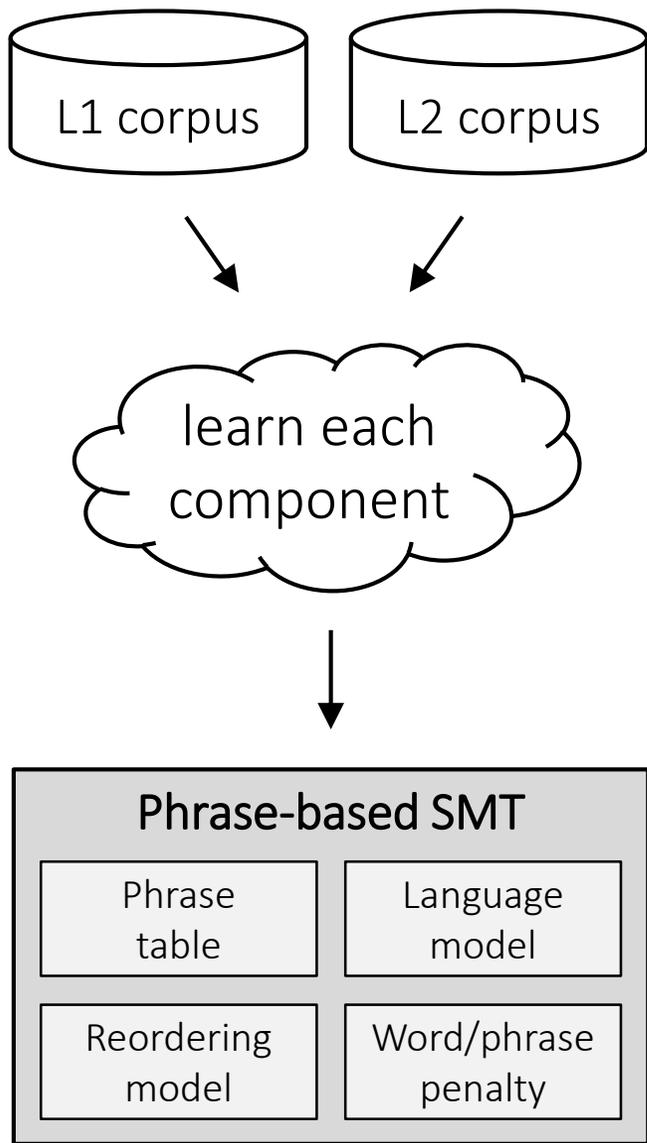


- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

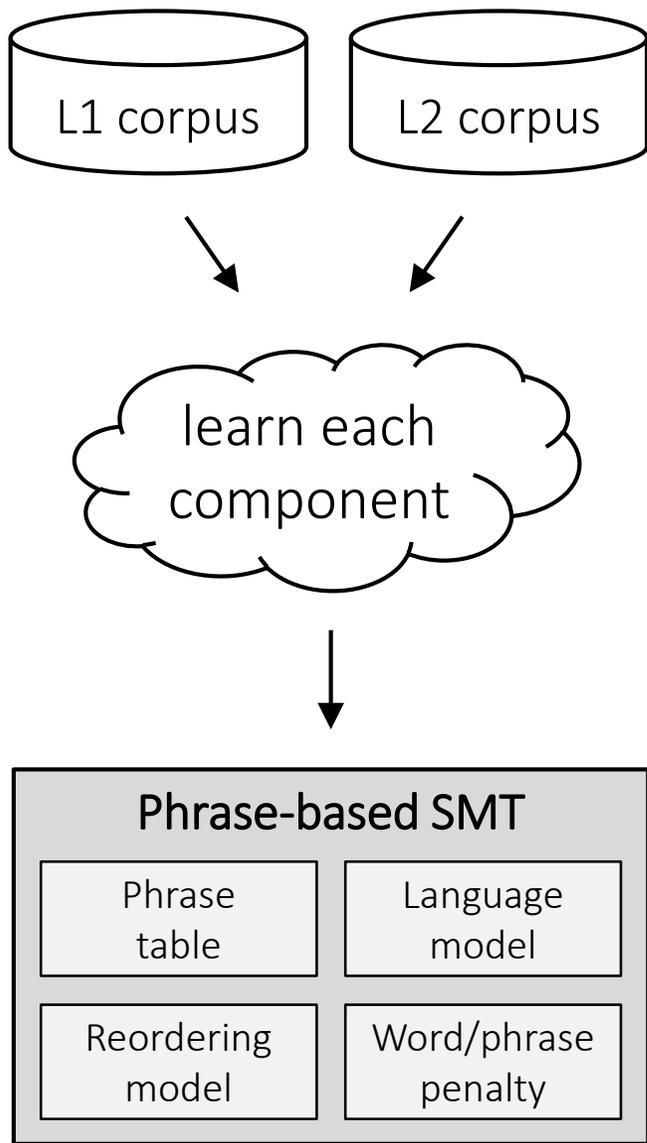
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

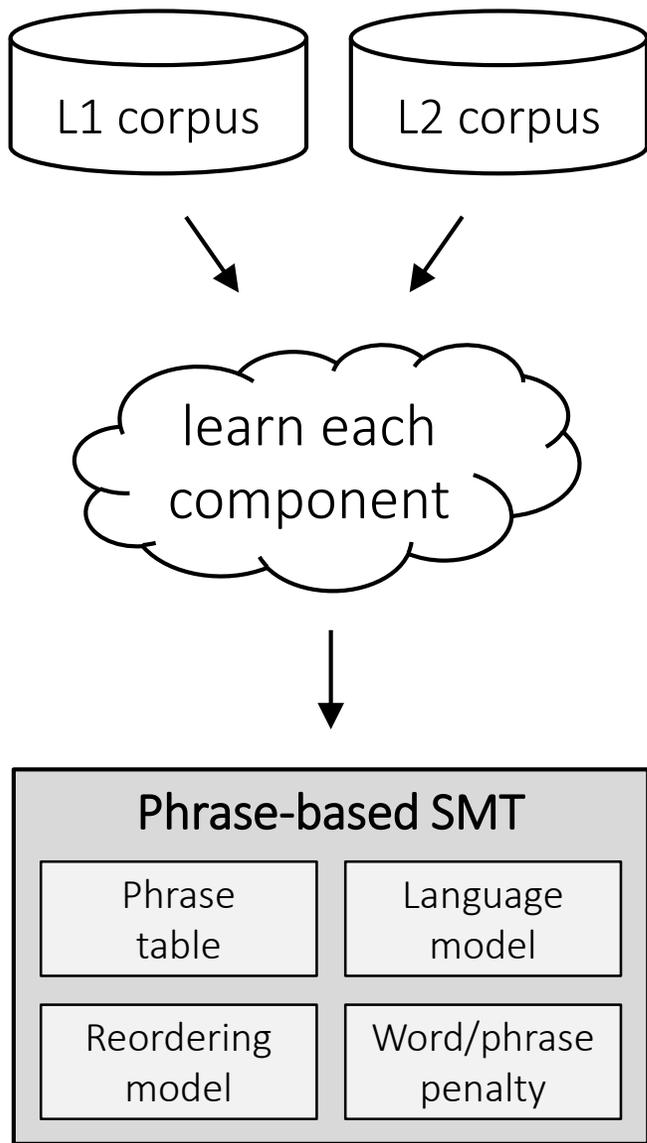
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

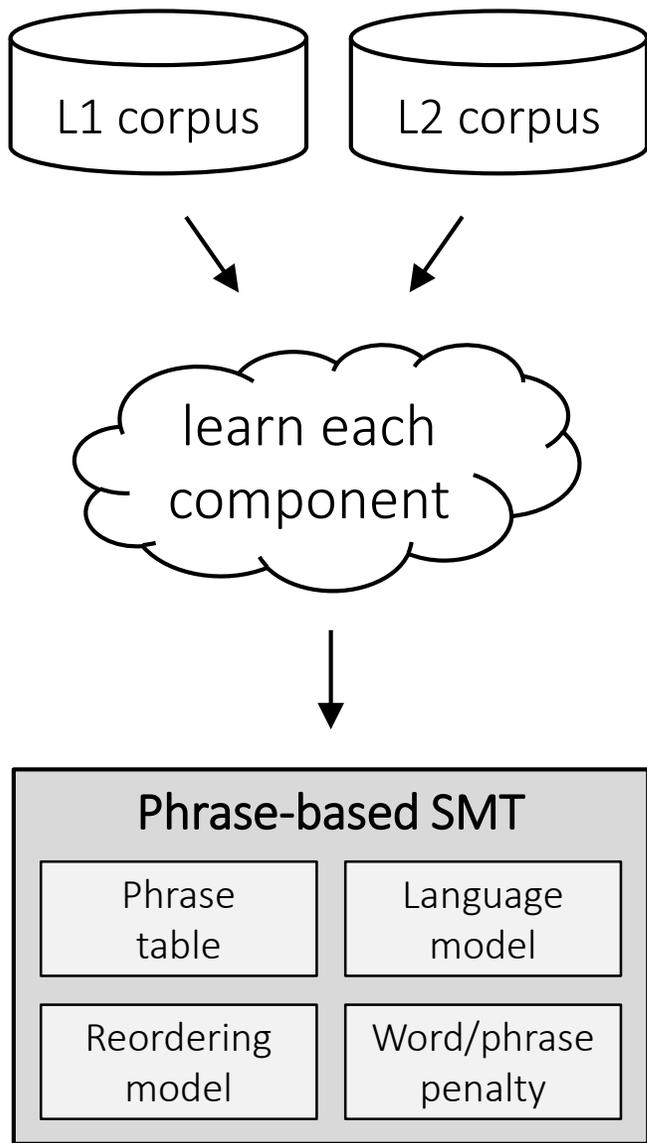
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

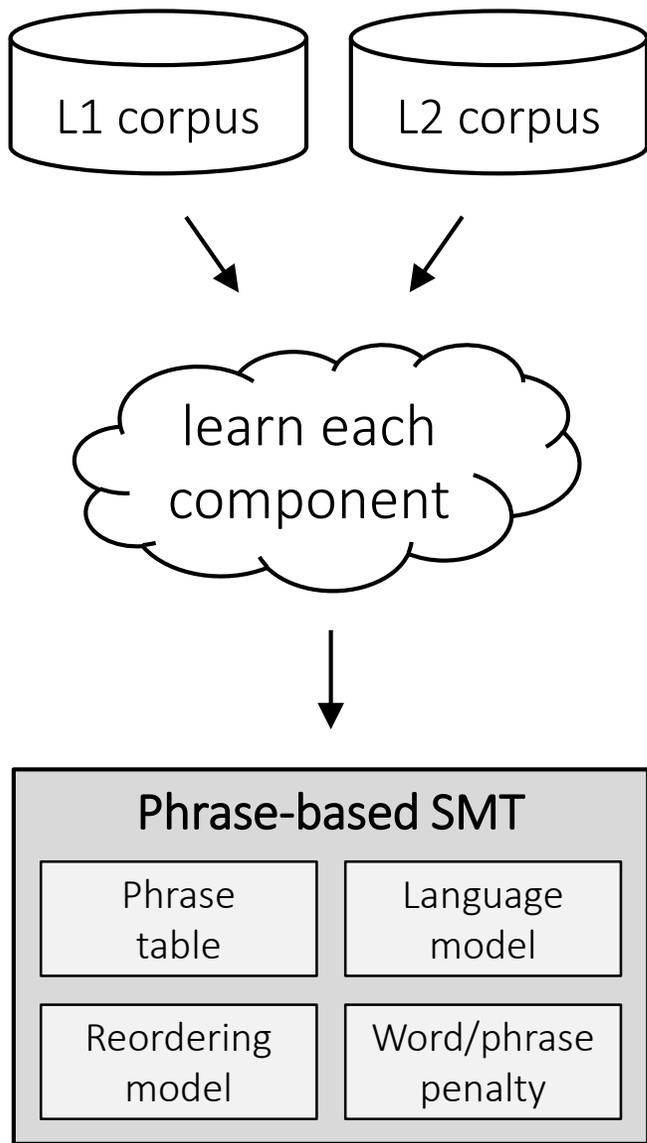
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

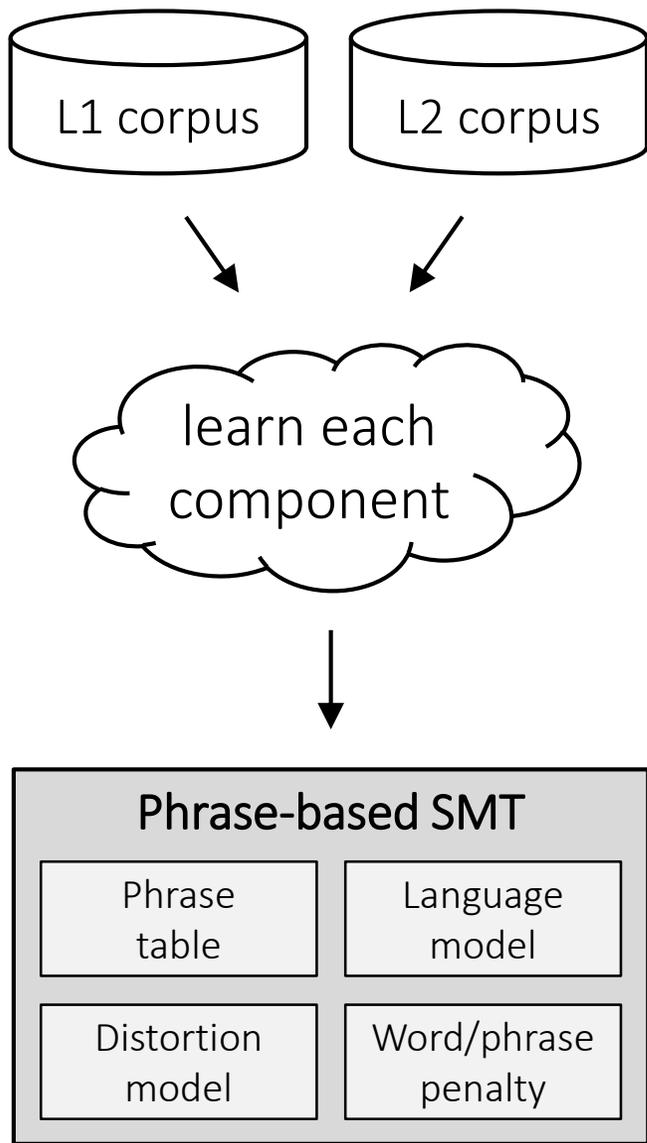
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - Lexical reordering model
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

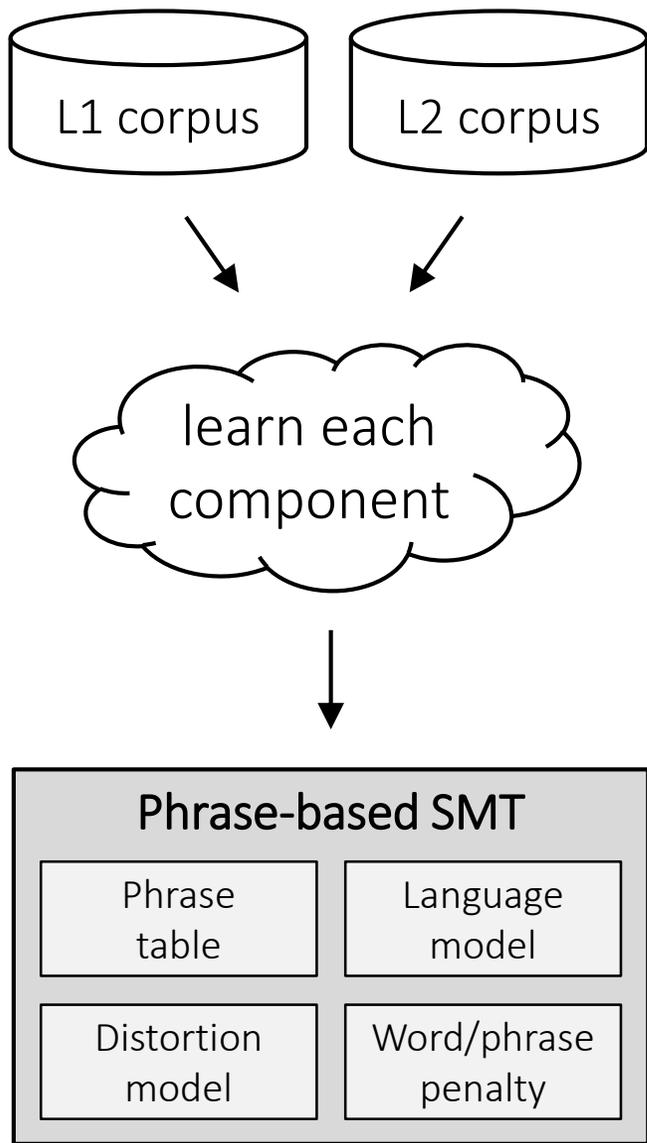
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

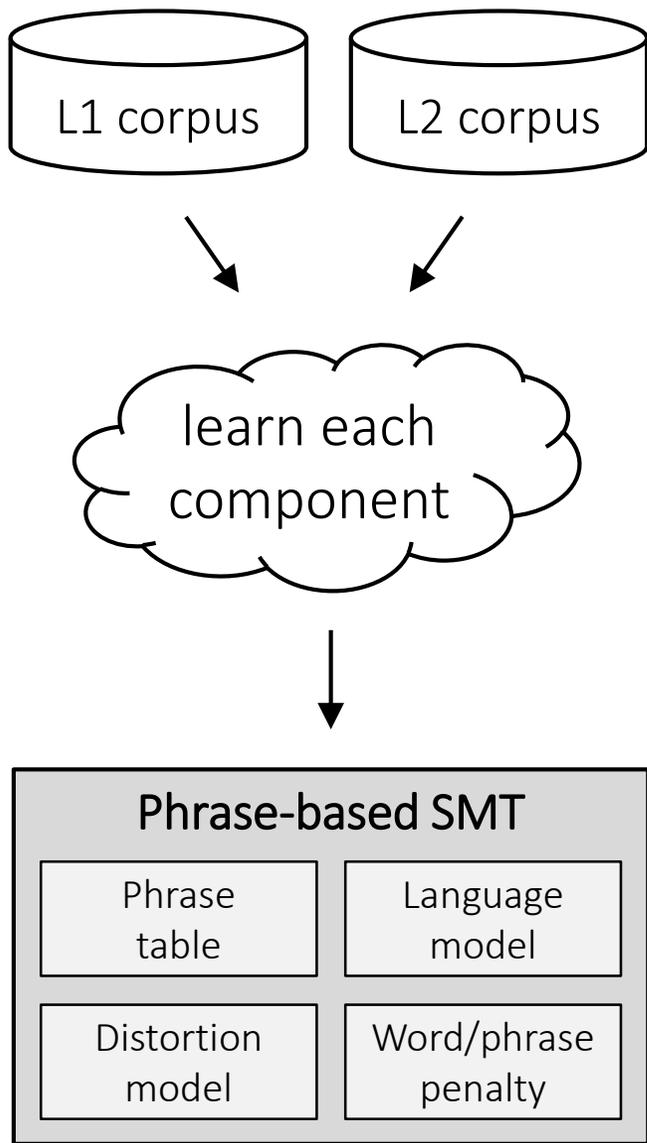
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

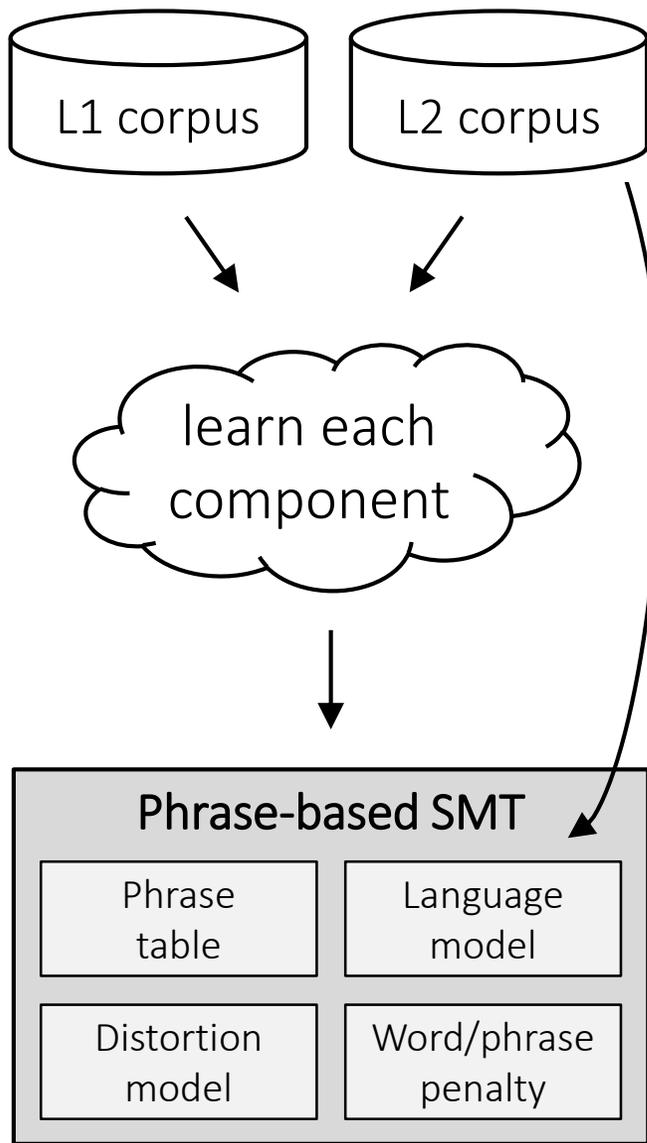
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

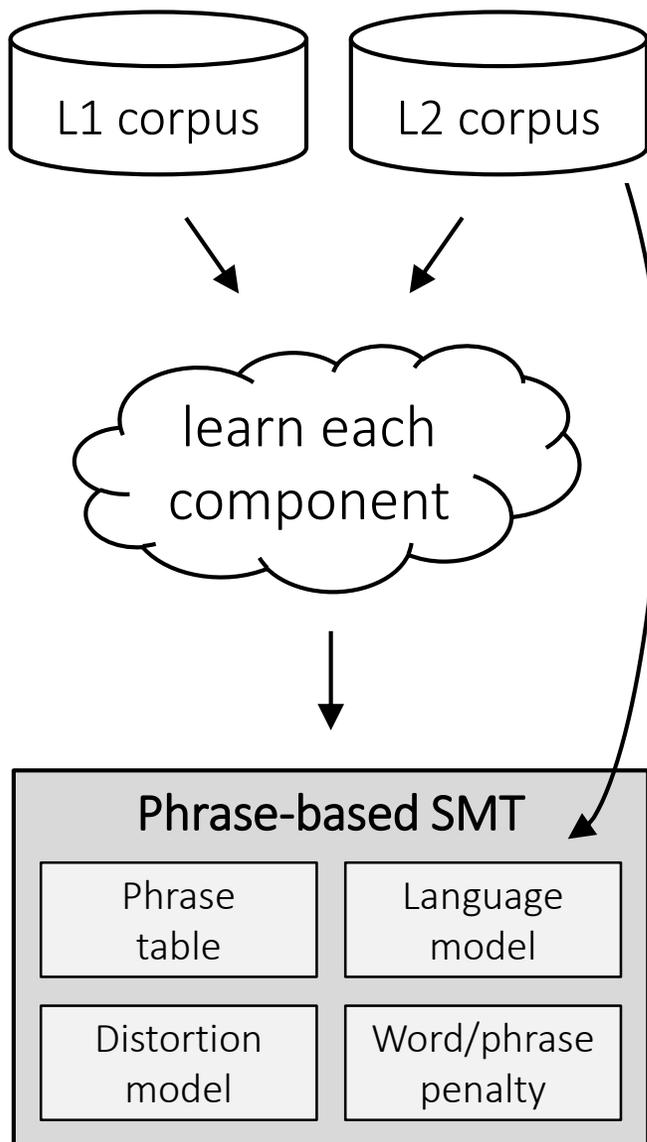
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

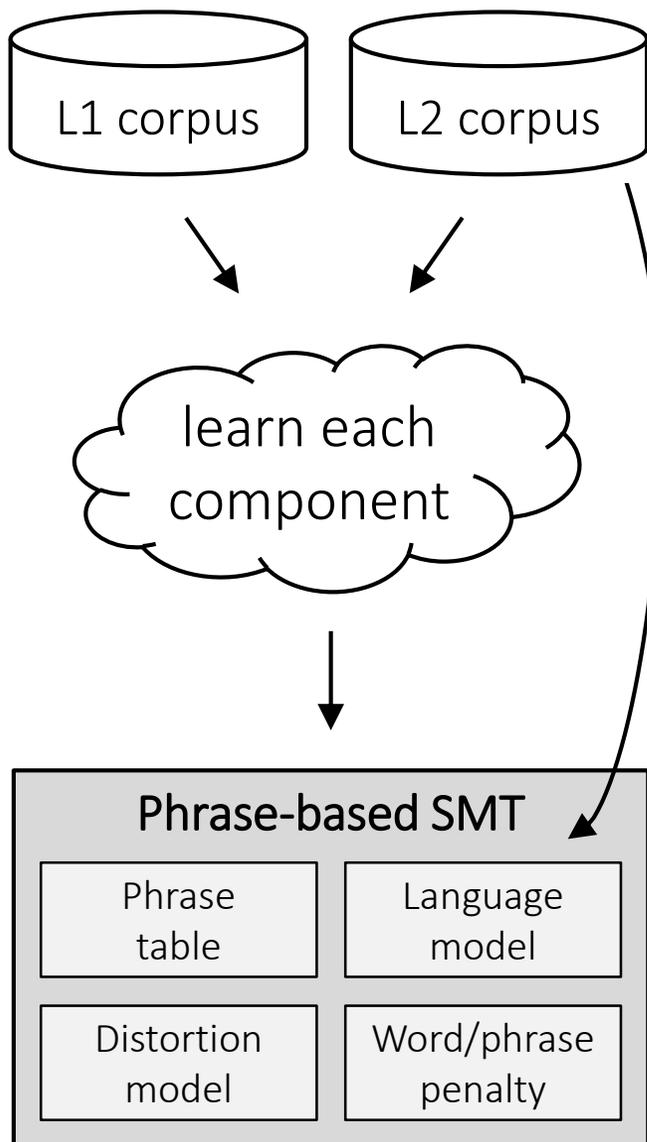
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

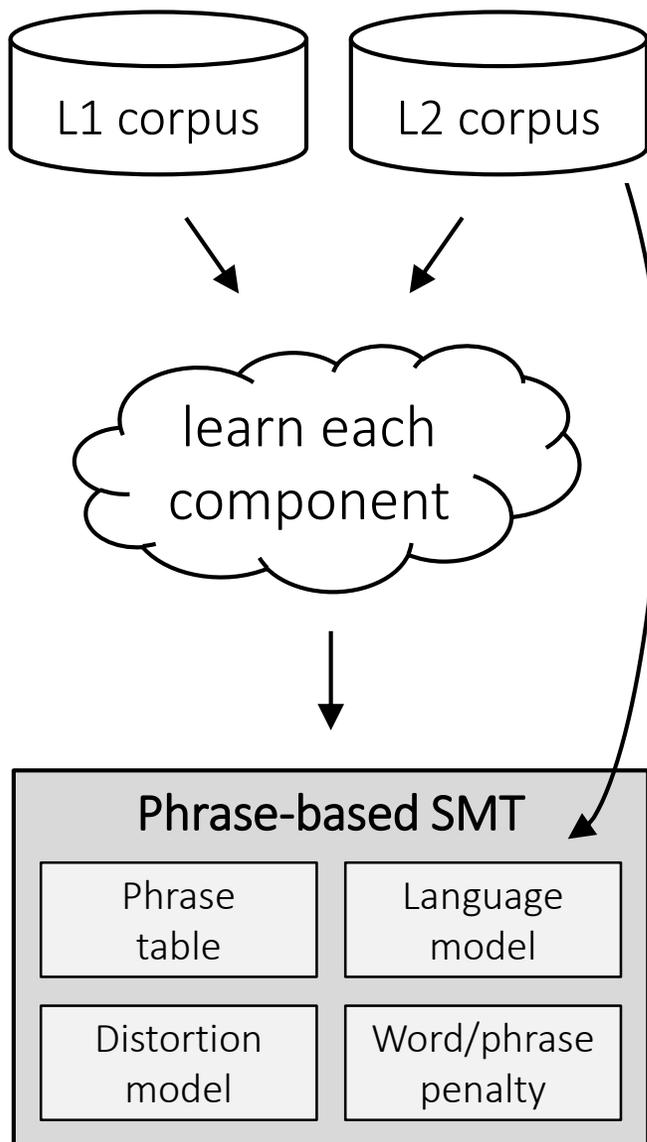
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

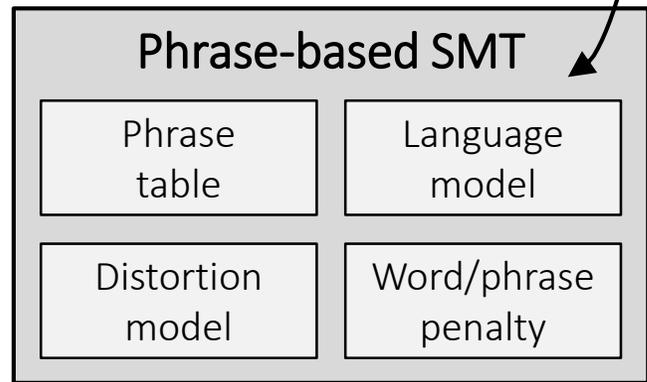
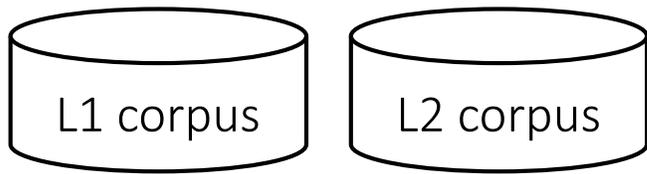
- Phrase table
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

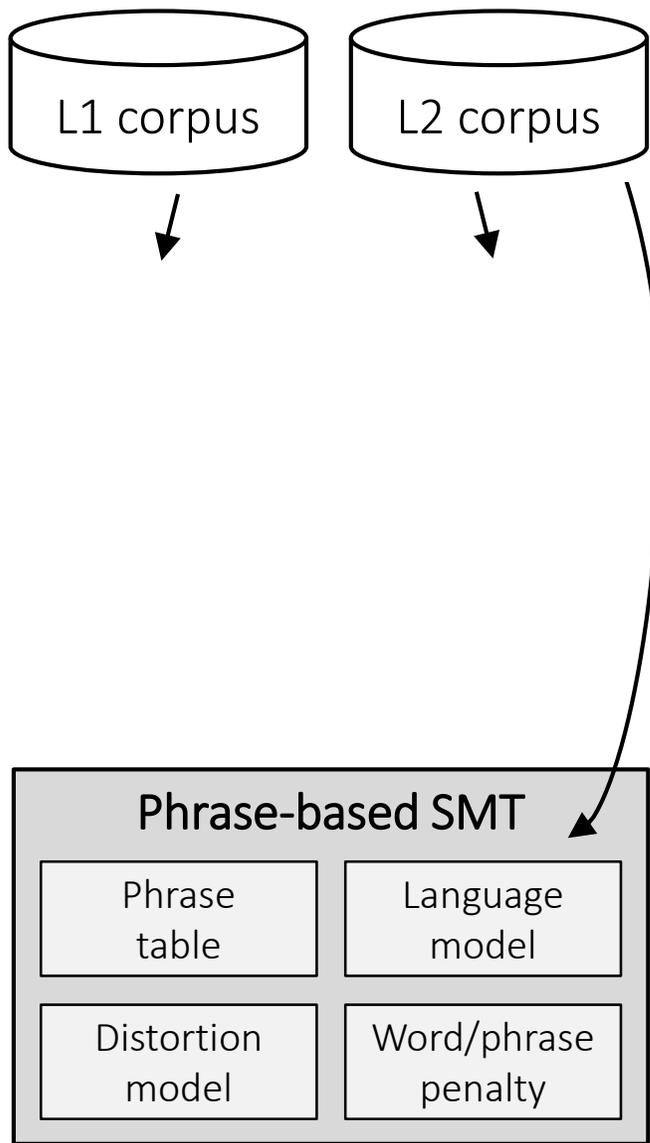
- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

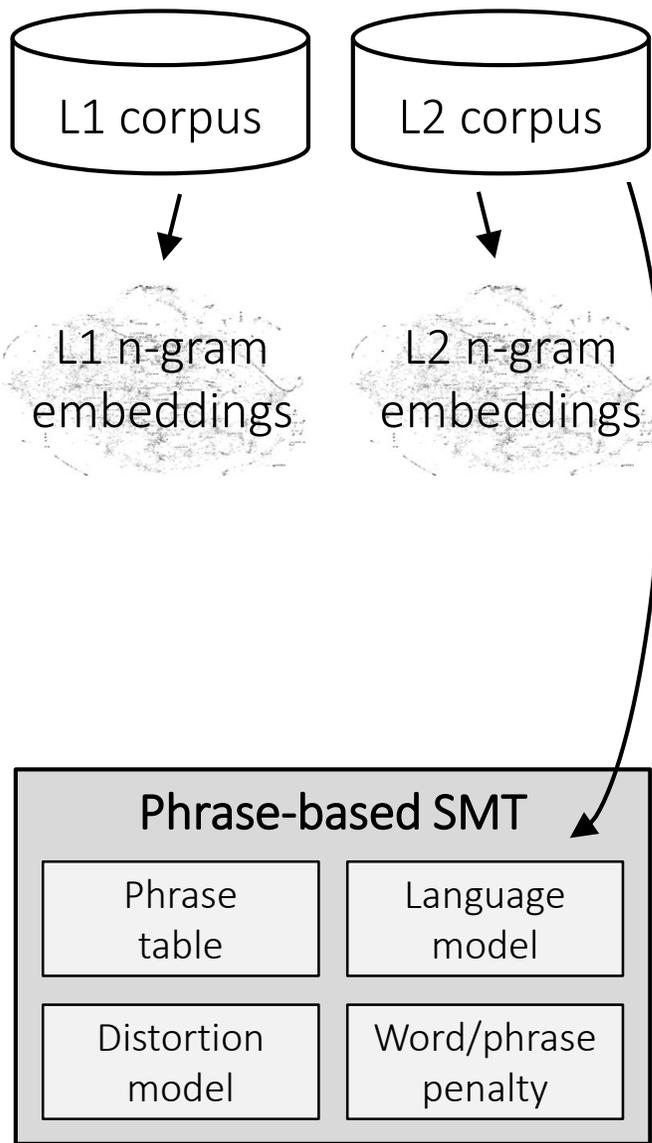
- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

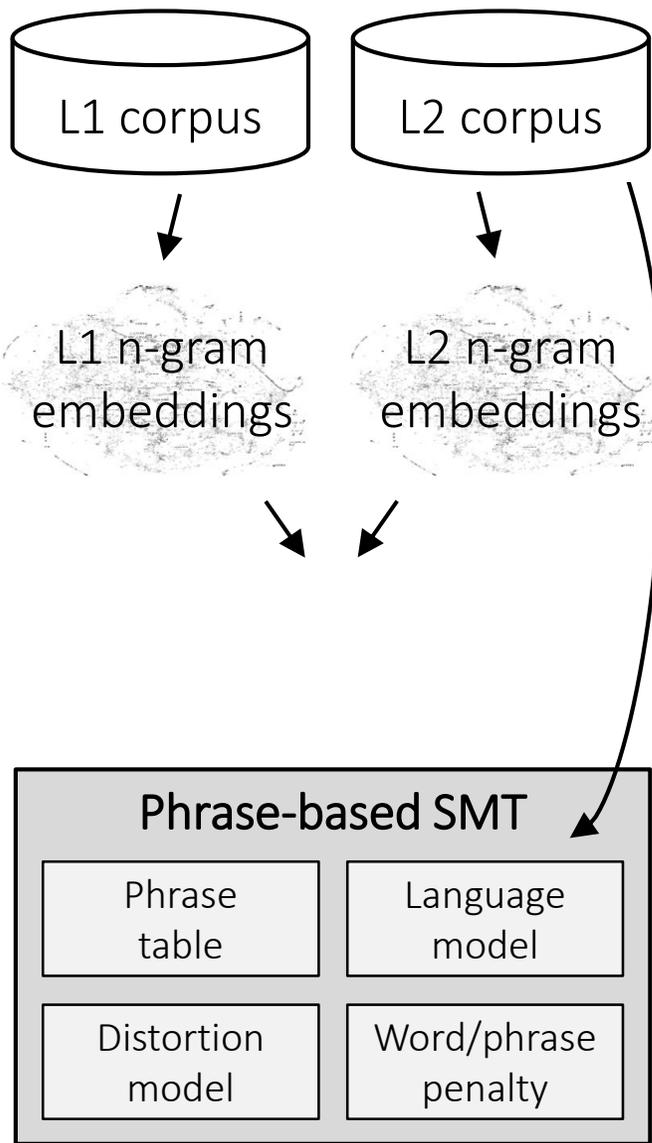
- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

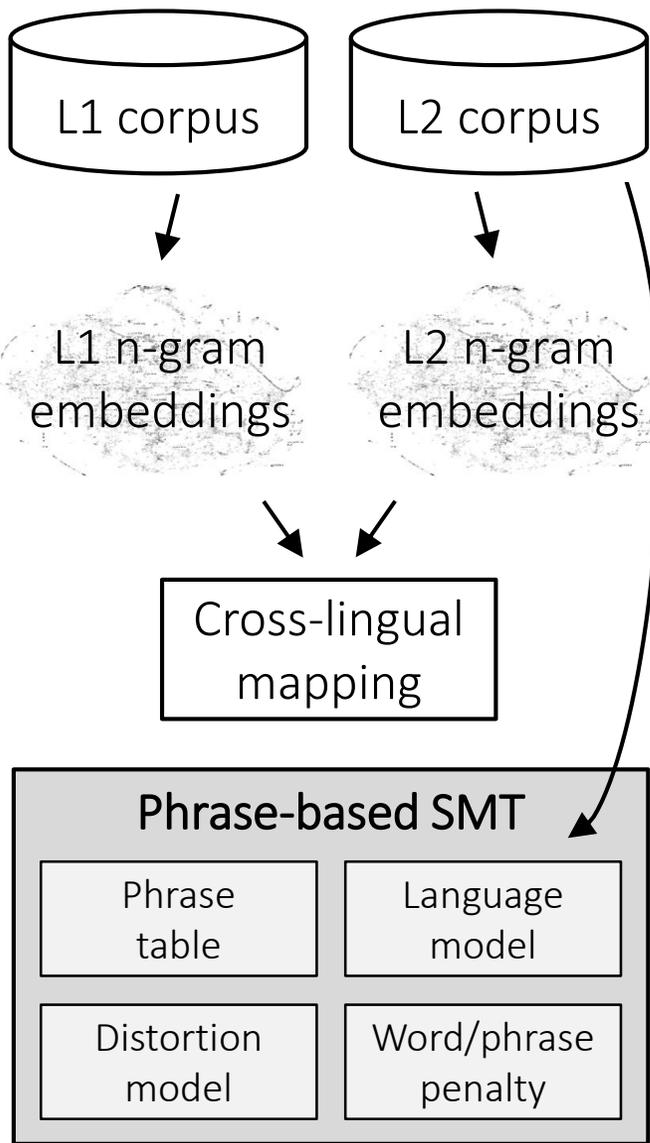
- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

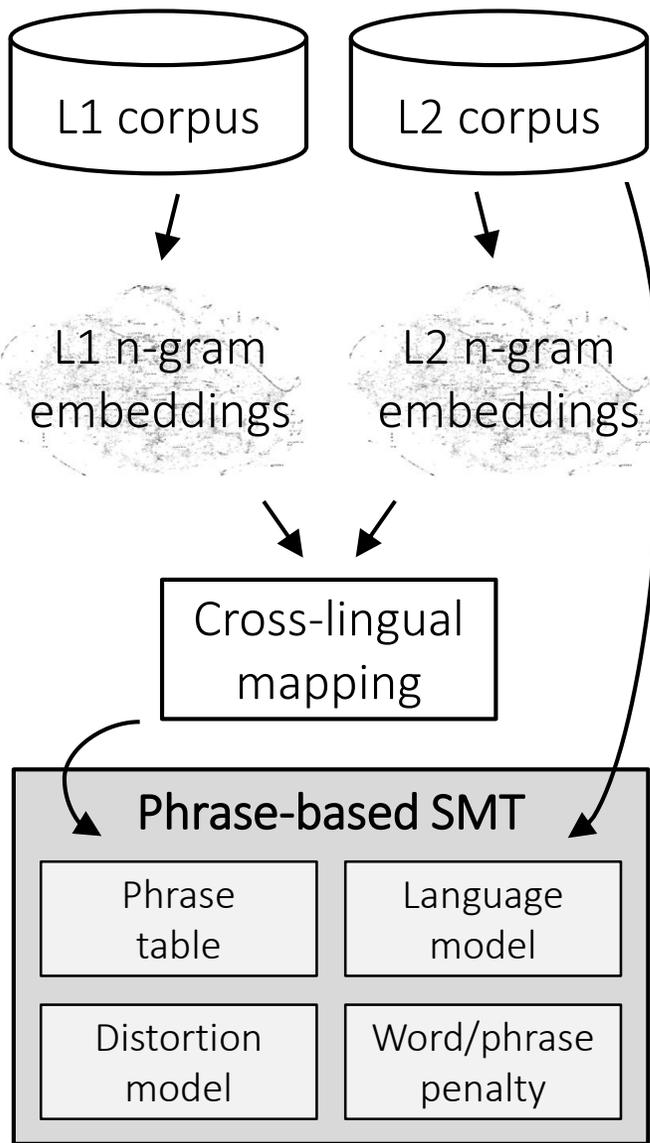
- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

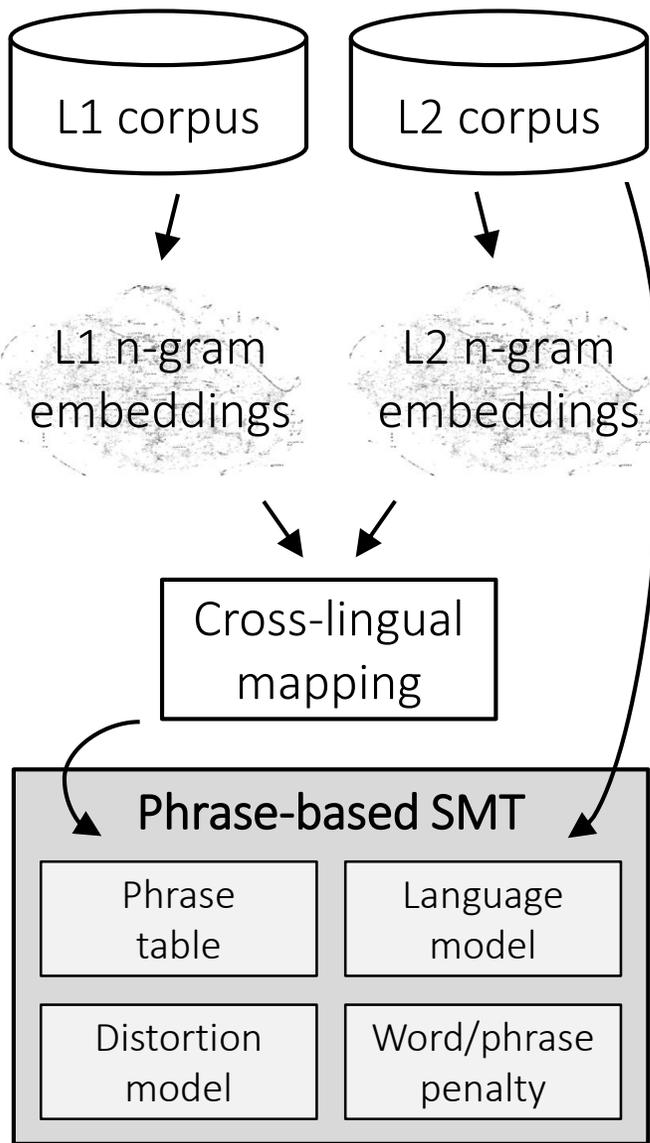
- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

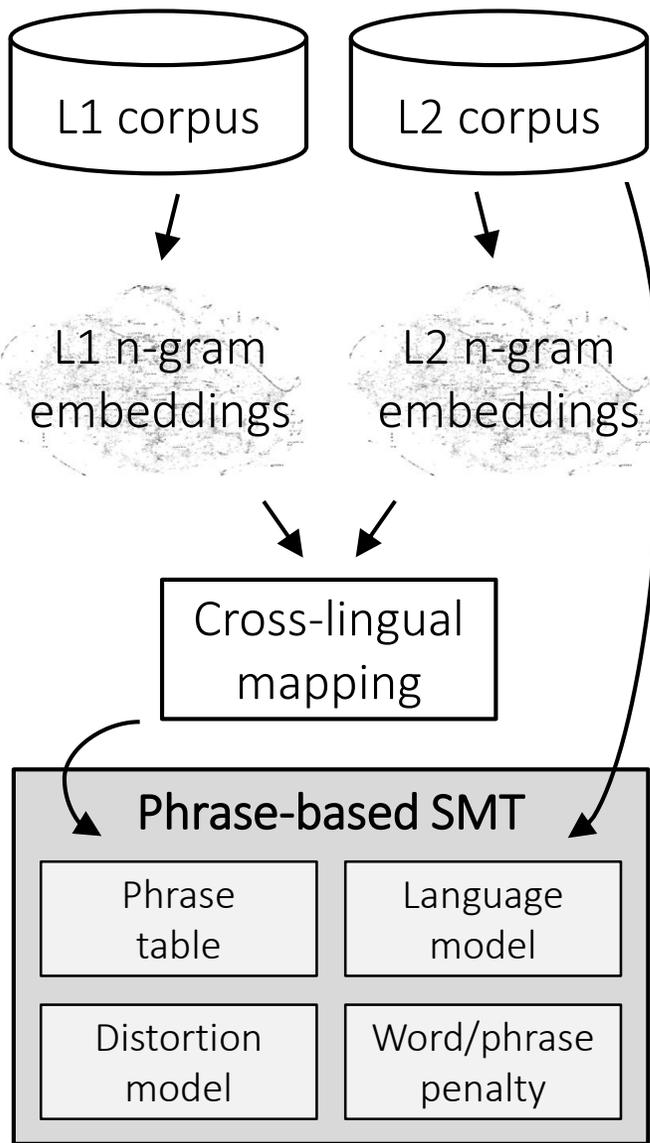
- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output

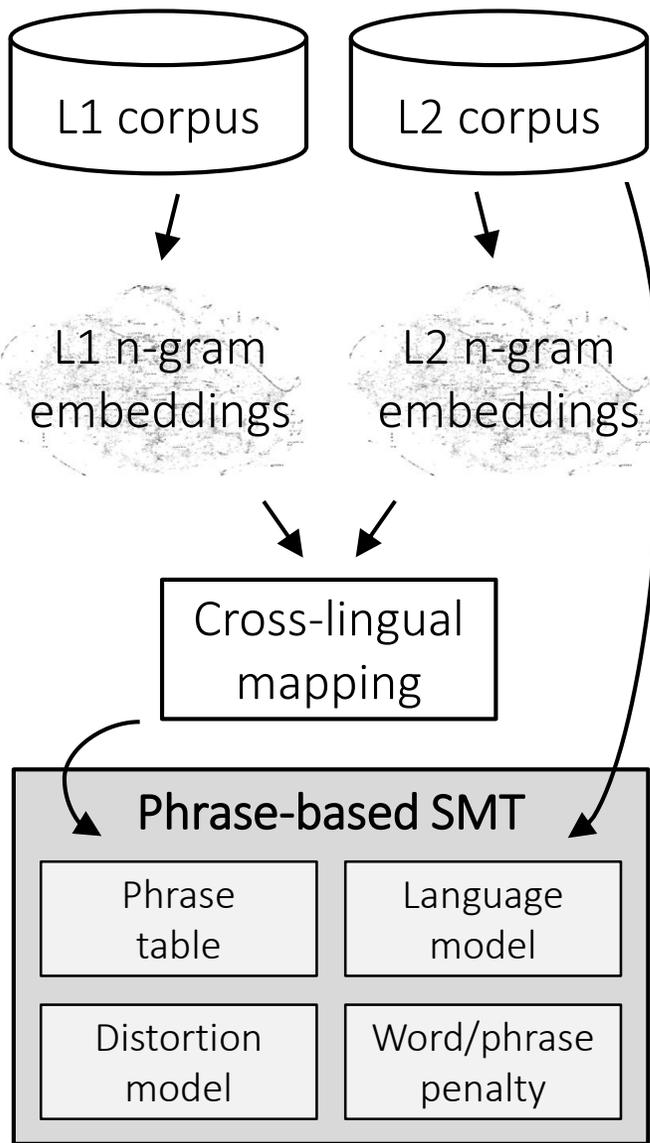


Unsupervised phrase-based SMT

Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output

I will go to New York by plane .

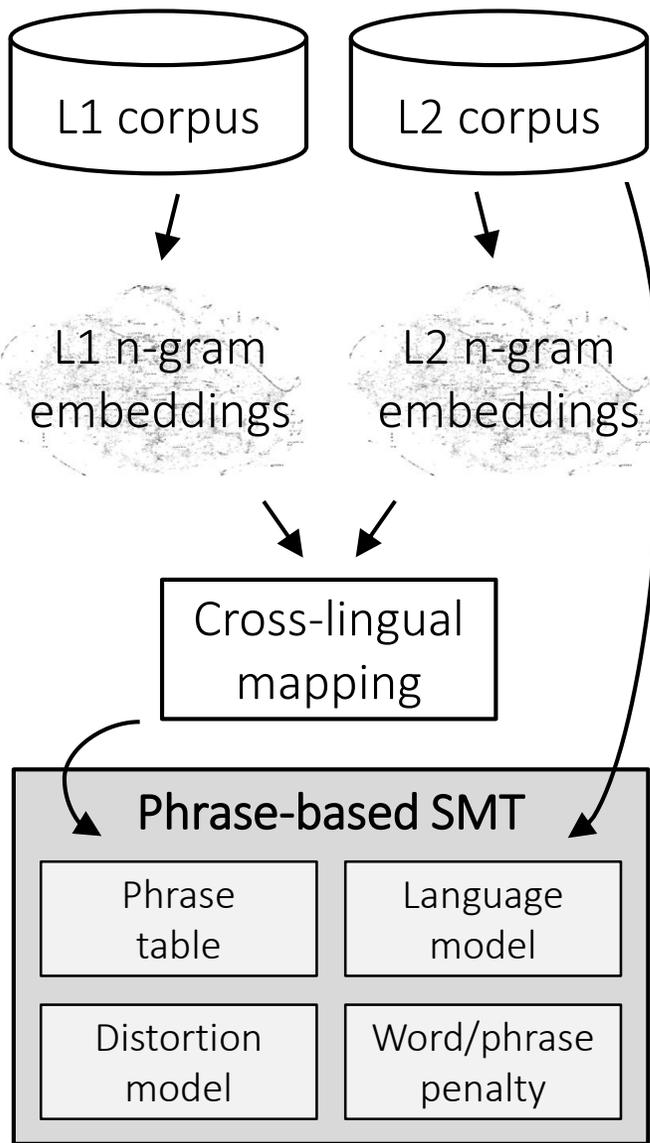


Unsupervised phrase-based SMT

Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output

I will go to New York by plane .
 w



Unsupervised phrase-based SMT

Learn components from monolingual corpora

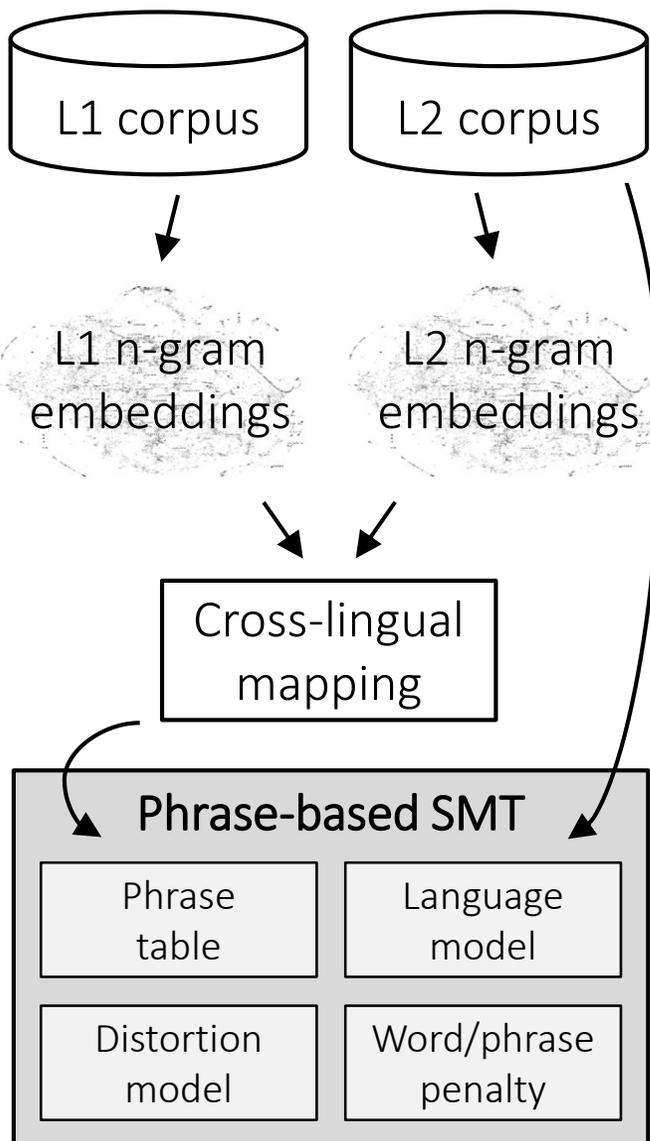
- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output

I will go to New York by plane .

w

↪

Unsupervised phrase-based SMT



Learn components from monolingual corpora

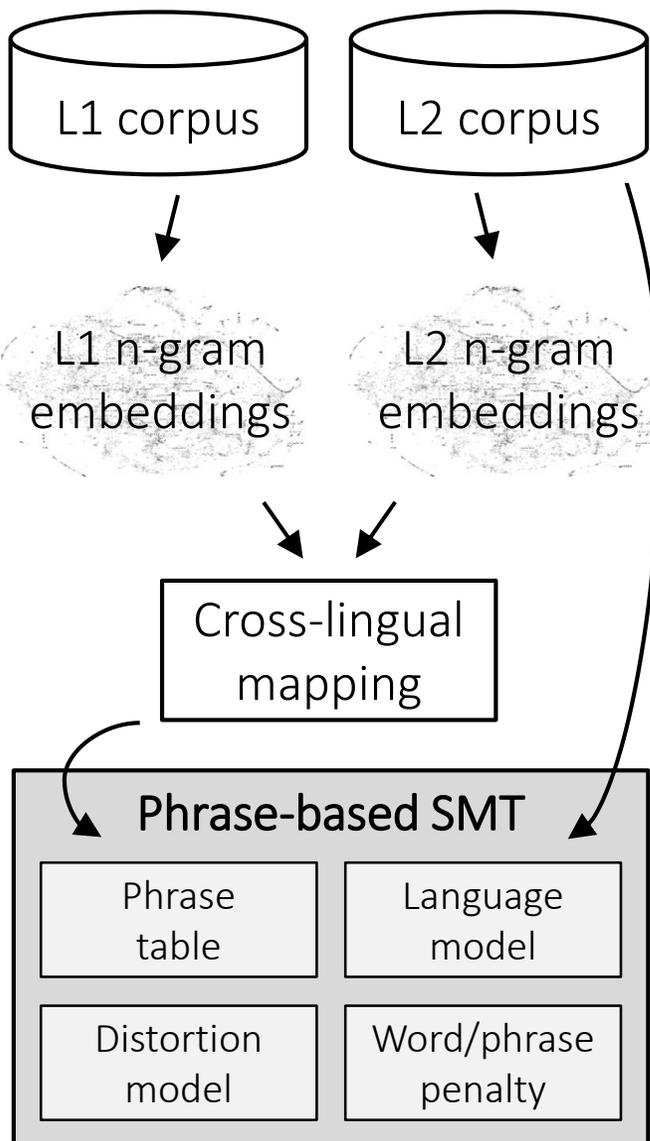
- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output

I will go to New York by plane .

w c

↪

Unsupervised phrase-based SMT



Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output

I will go to New York by plane .

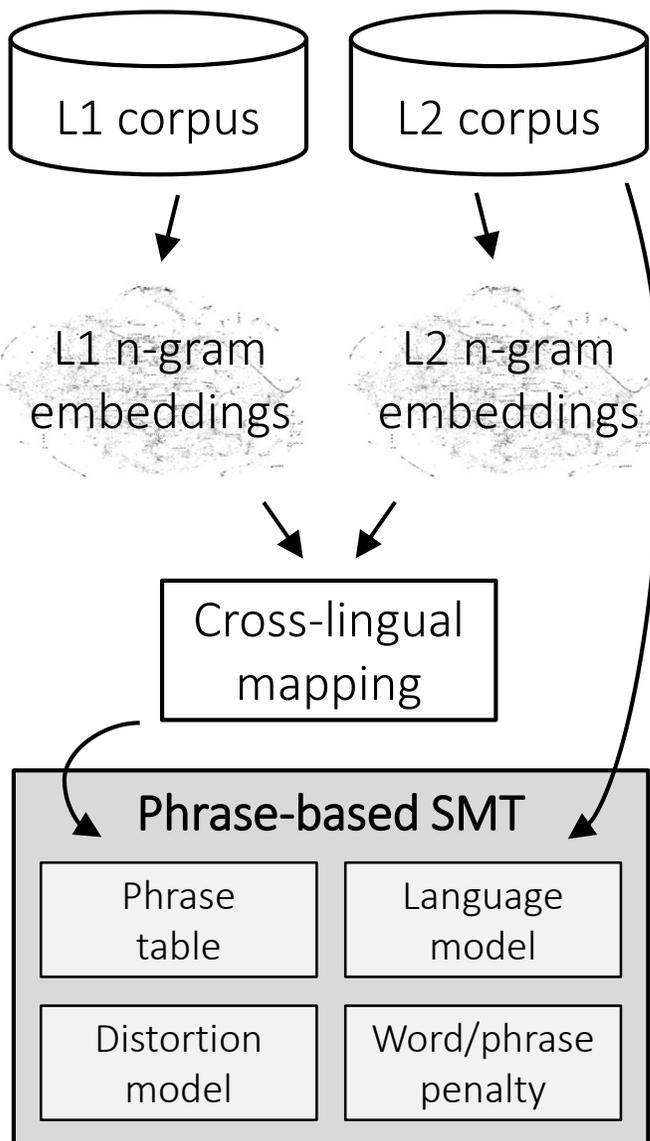
w *c*

↪

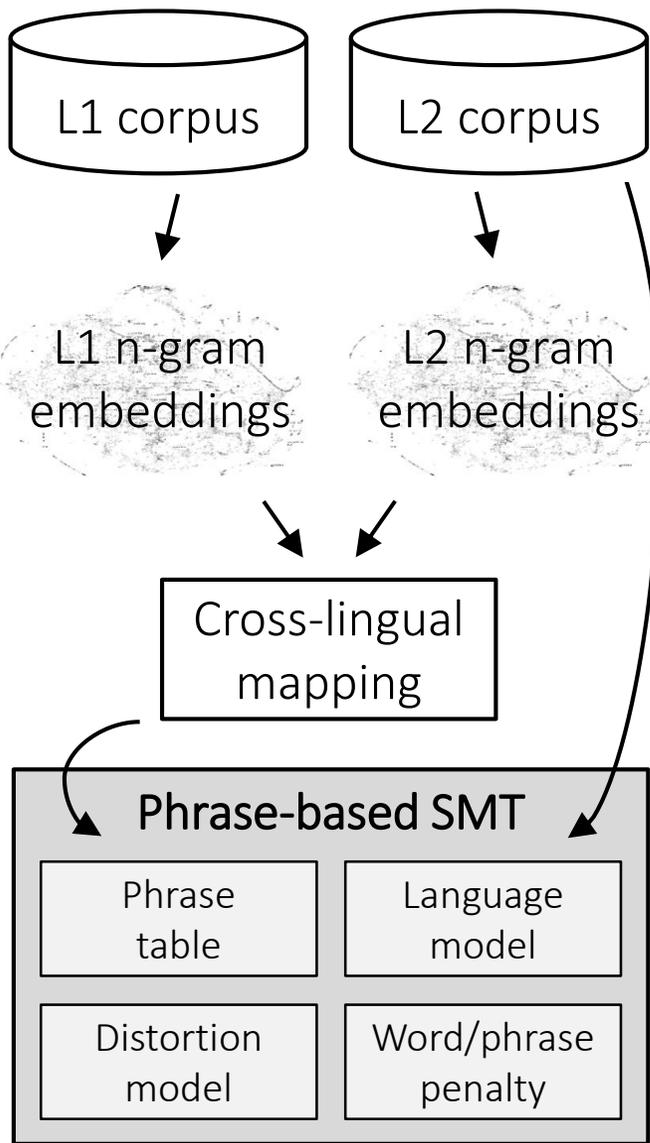
Unsupervised phrase-based SMT

Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



I will go to New York by plane .
 w c



Unsupervised phrase-based SMT

Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output

I will go to New York by plane .

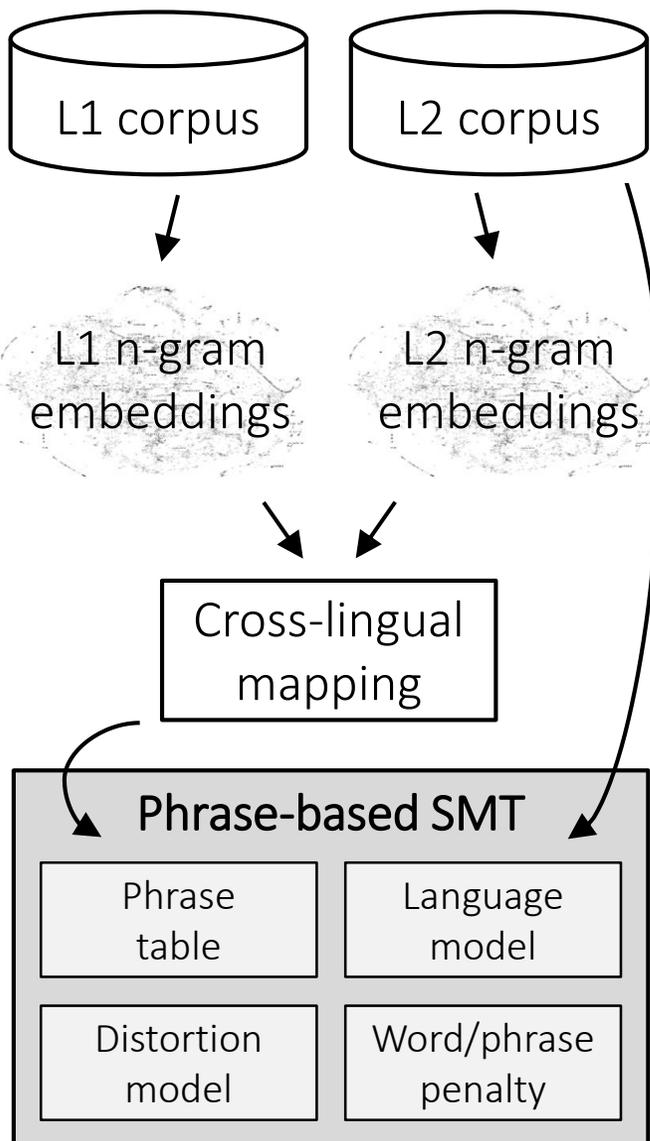
p *c*

↪

Unsupervised phrase-based SMT

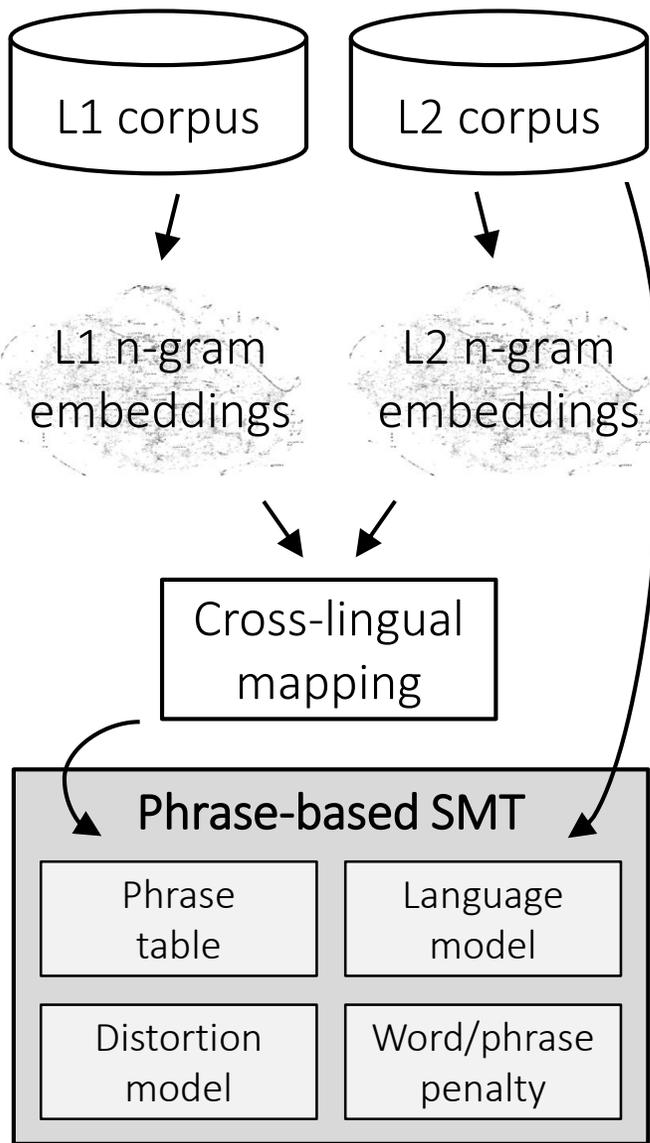
Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



I will go to New York by plane .

p *c*



Unsupervised phrase-based SMT

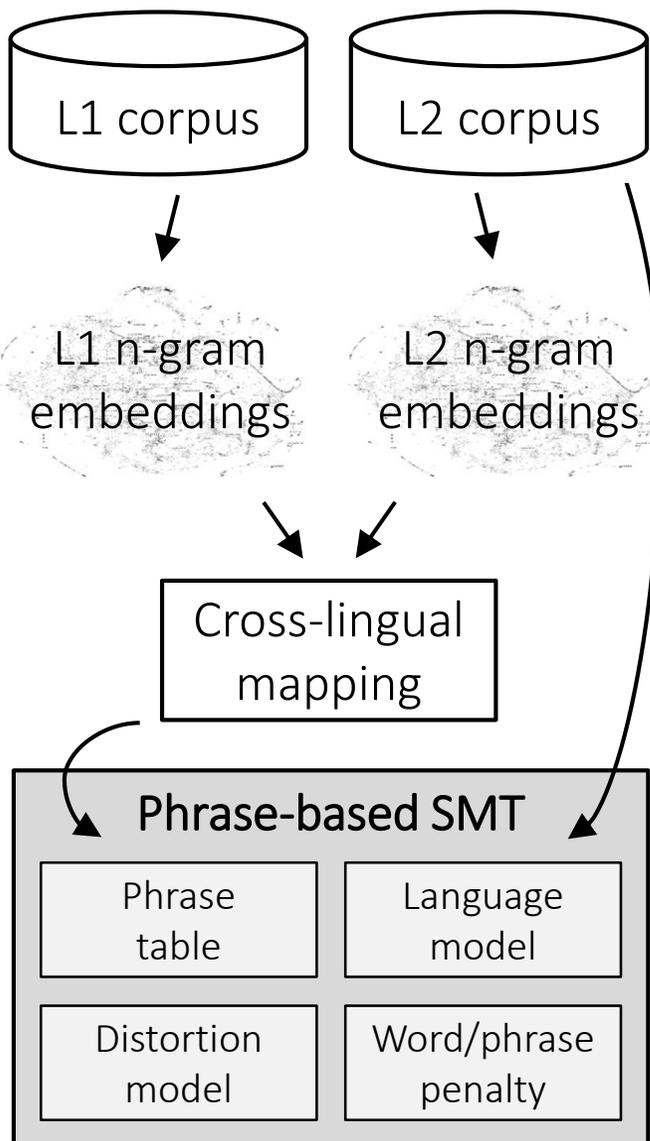
Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output

Unsupervised phrase-based SMT

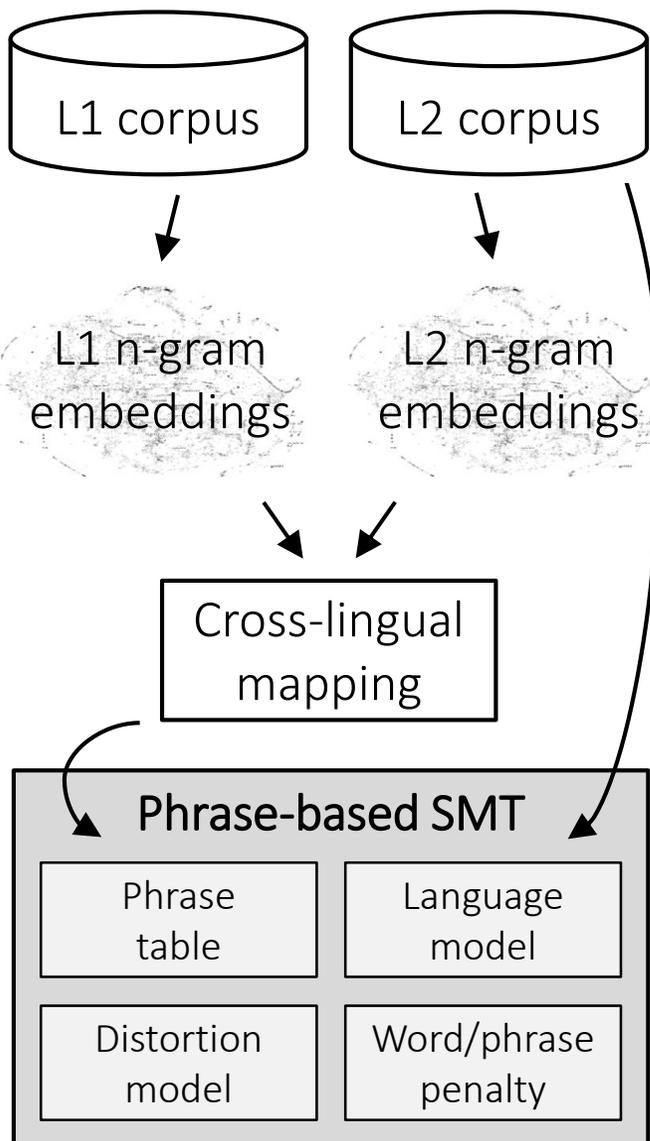
Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



For each \bar{e} , estimate $\phi(\bar{f}|\bar{e})$ for 100-NN:

Unsupervised phrase-based SMT



Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output

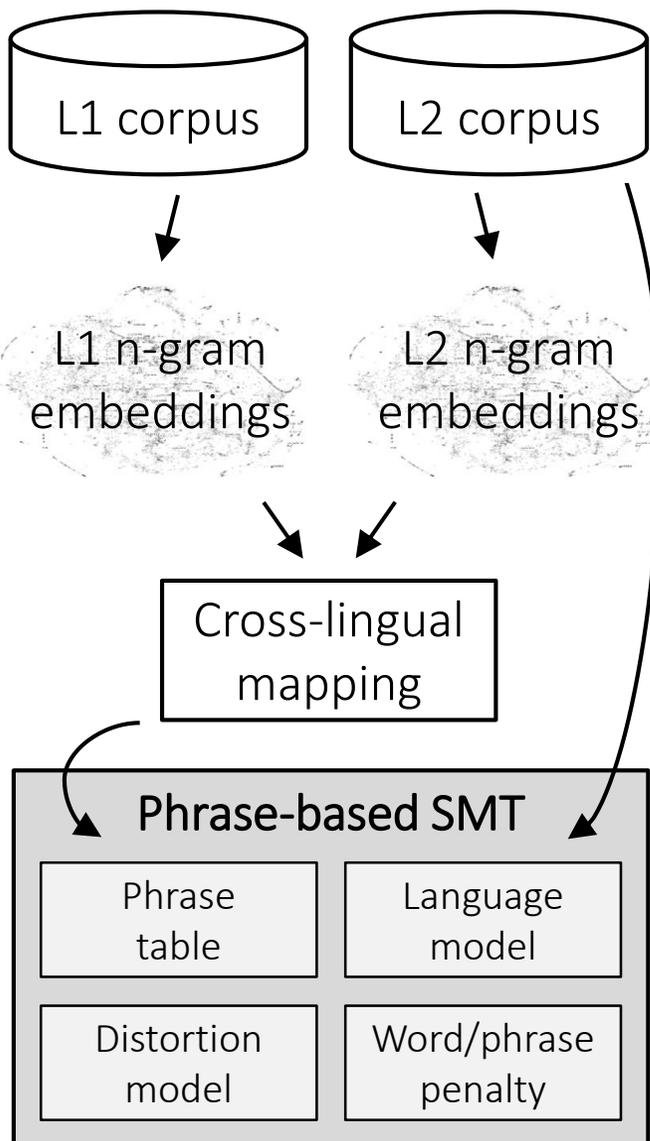
For each \bar{e} , estimate $\phi(\bar{f}|\bar{e})$ for 100-NN:

$$\phi(\bar{f}|\bar{e}) = \frac{e^{\cos(\bar{e}, \bar{f})/\tau}}{\sum_{\bar{f}'} e^{\cos(\bar{e}, \bar{f}')/\tau}}$$

Unsupervised phrase-based SMT

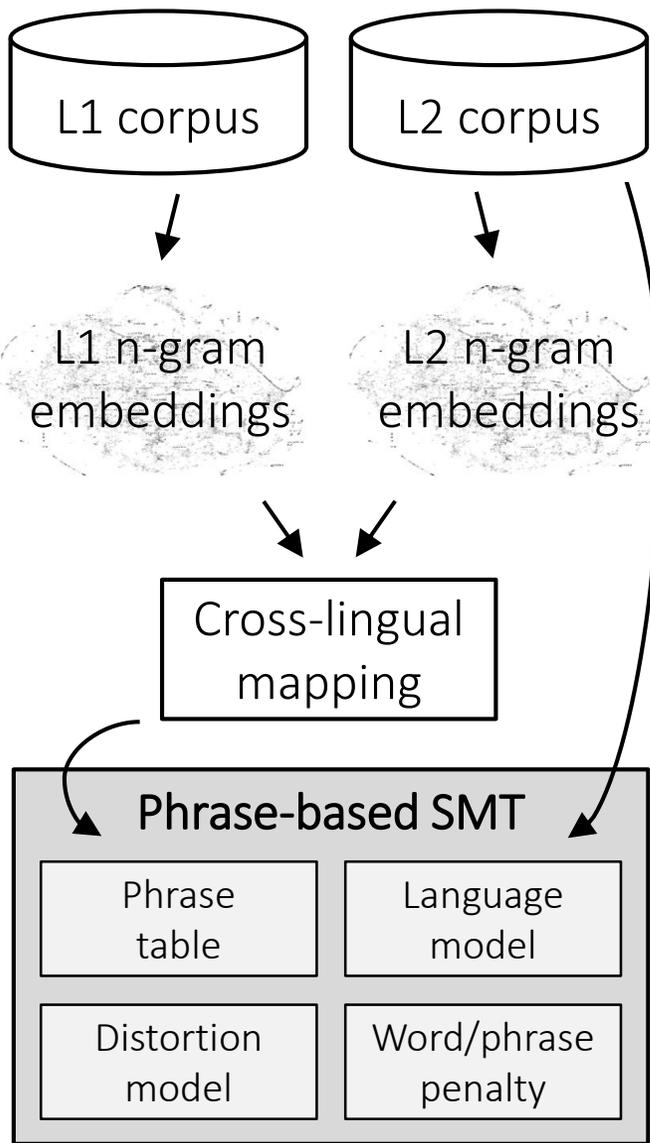
Learn components from monolingual corpora

- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



For each \bar{e} , estimate $\phi(\bar{f}|\bar{e})$ for 100-NN:

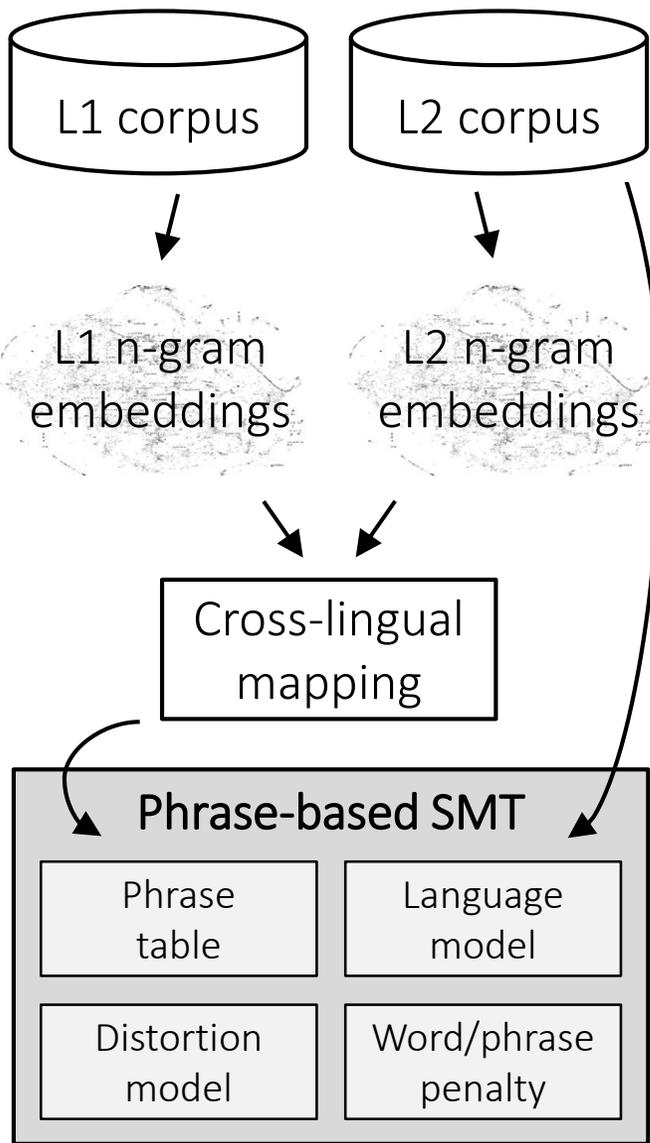
$$\phi(\bar{f}|\bar{e}) = \frac{e^{\cos(\bar{e}, \bar{f})/\tau}}{\sum_{\bar{f}'} e^{\cos(\bar{e}, \bar{f}')/\tau}} \quad \min_{\tau} \sum_{\bar{f}} \log \phi(\bar{f}|\text{NN}_{\bar{e}}(\bar{f})) + \sum_{\bar{e}} \log \phi(\bar{e}|\text{NN}_{\bar{f}}(\bar{e}))$$



Unsupervised phrase-based SMT

Learn components from monolingual corpora

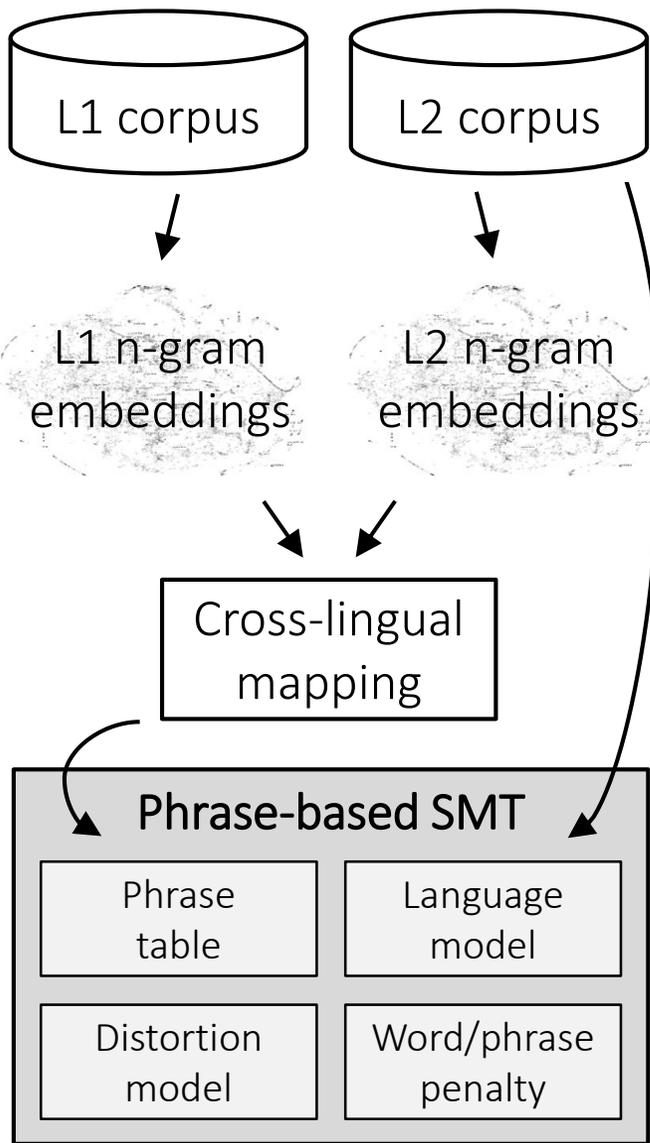
- Phrase table **TRICKY...**
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output



Unsupervised phrase-based SMT

Learn components from monolingual corpora

- Phrase table ~~TRICKY...~~
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model **EASY!!!**
 - N-gram frequency counts with back-off and smoothing
- Reordering model **EASY!!!**
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty **EASY!!!**
 - Fixed score to control the length of the output

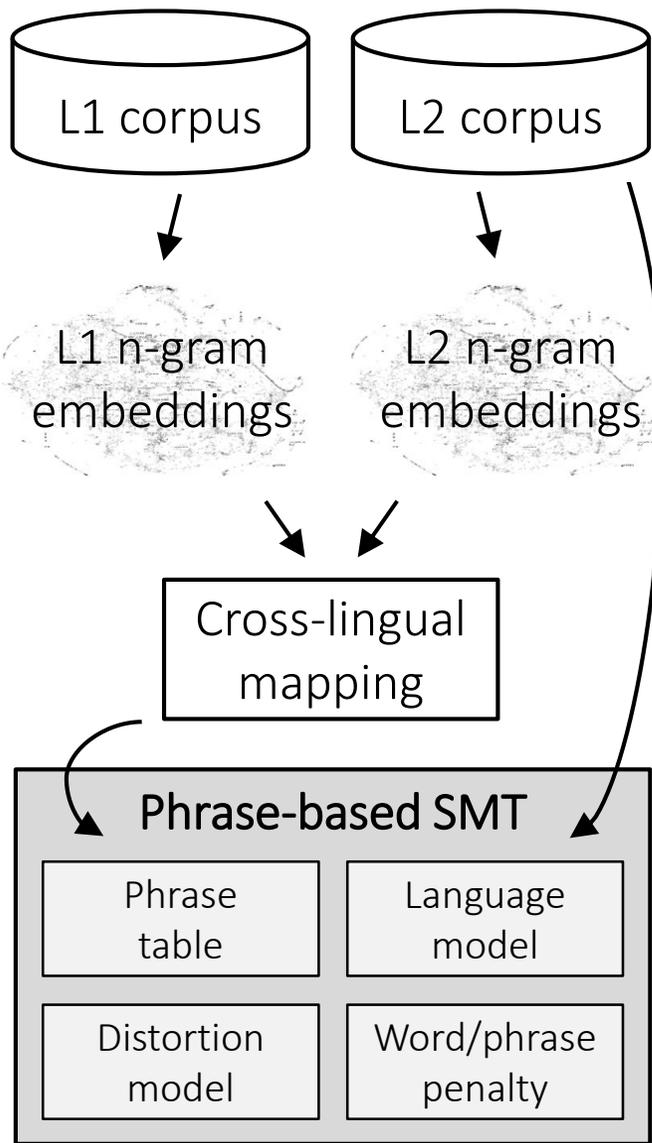


Unsupervised phrase-based SMT

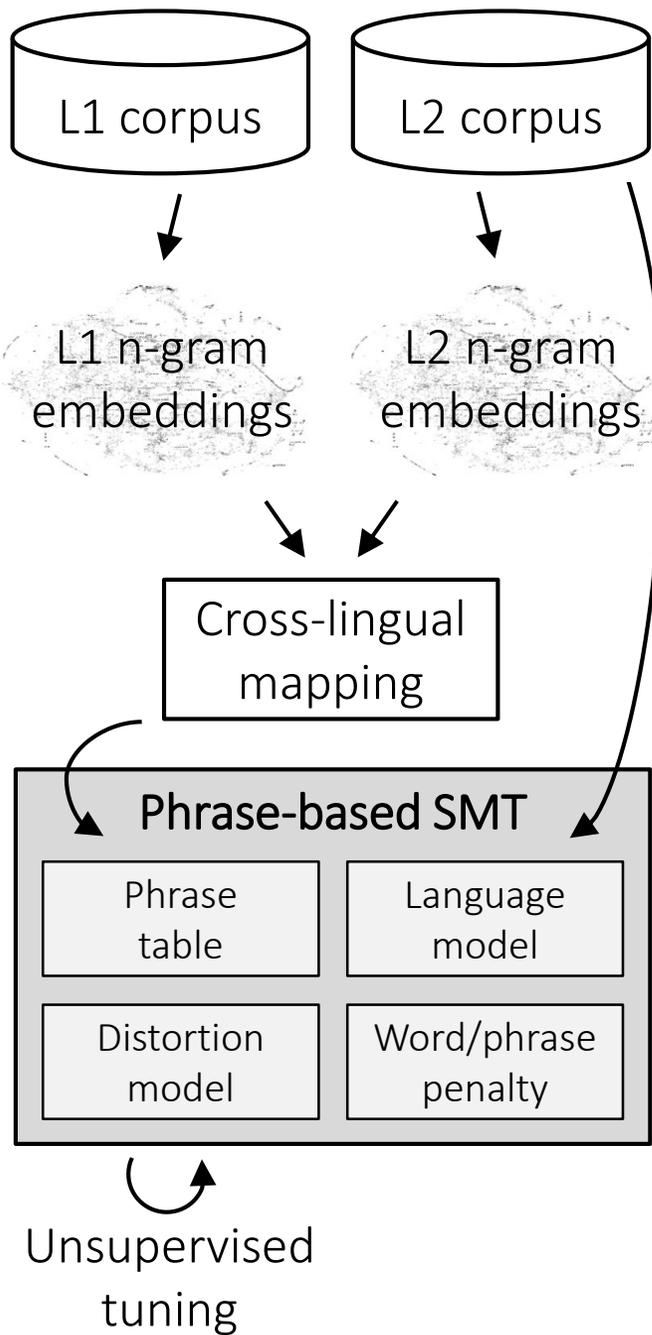
Learn components from monolingual corpora

- Phrase table ~~TRICKY...~~ EASY!!!
 - Direct/inverse translation probabilities
 - Direct/inverse lexical weightings
- Language model EASY!!!
 - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
 - Distortion model (distance based)
 - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!
 - Fixed score to control the length of the output

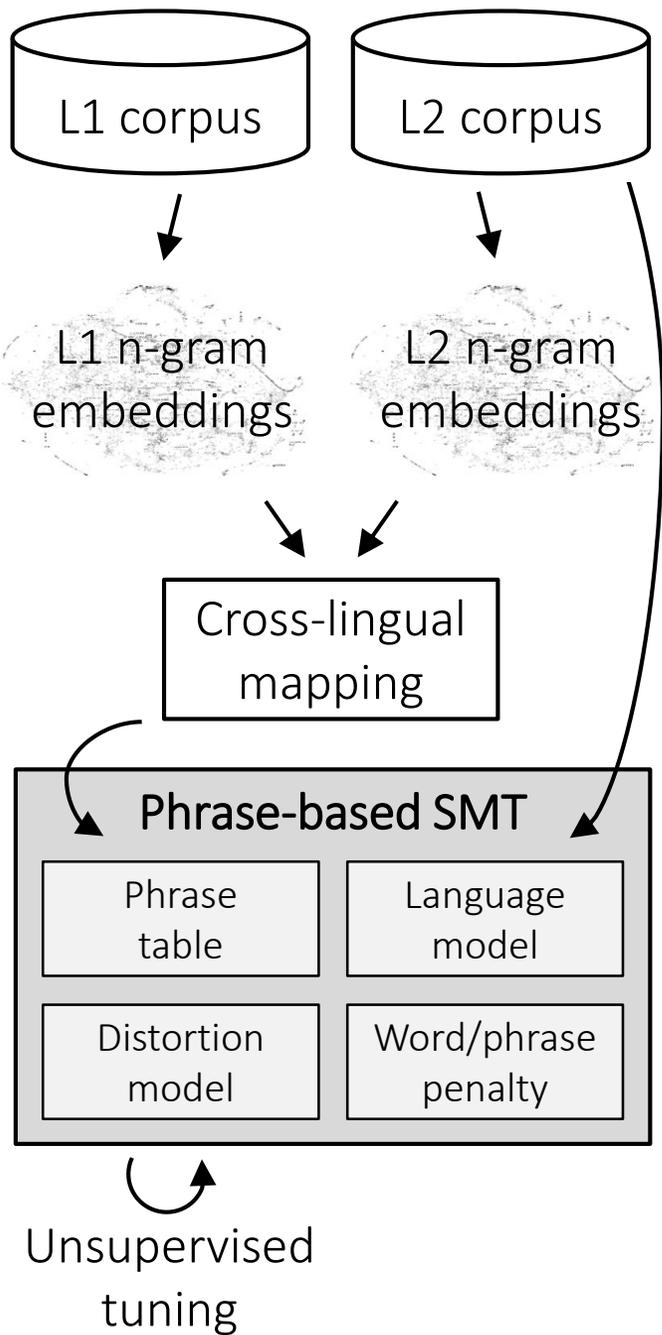
Unsupervised phrase-based SMT



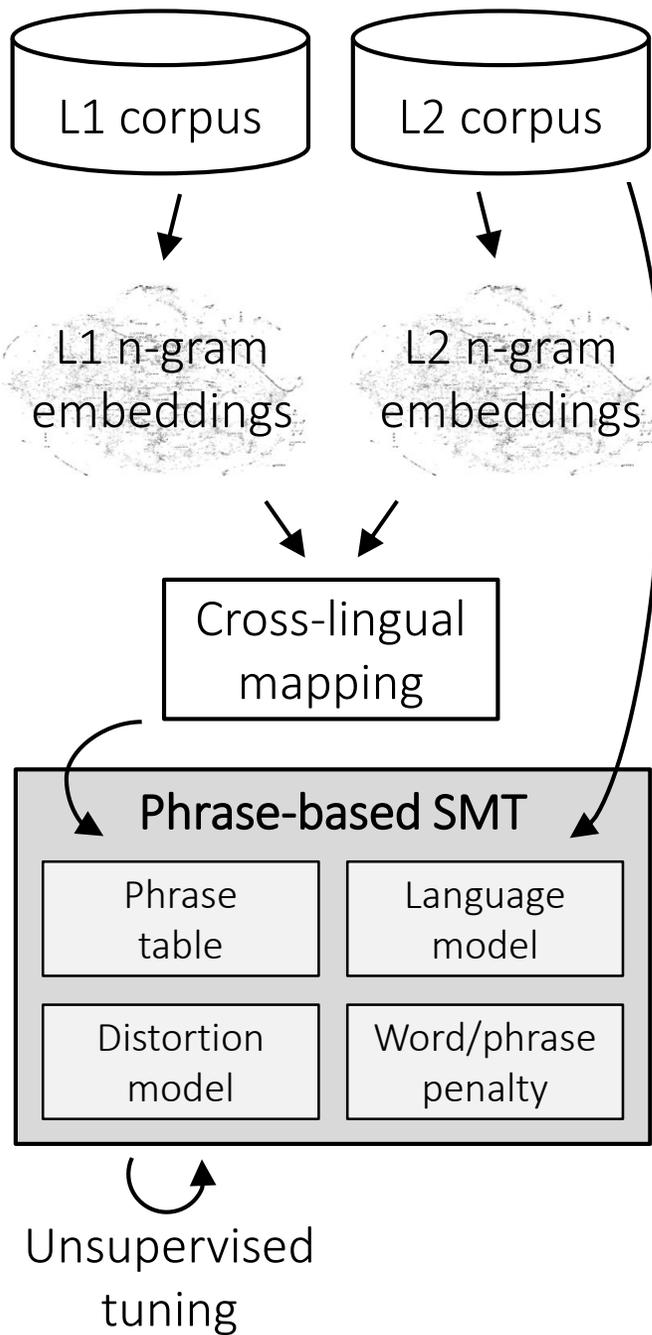
Unsupervised phrase-based SMT



Unsupervised phrase-based SMT



MERT

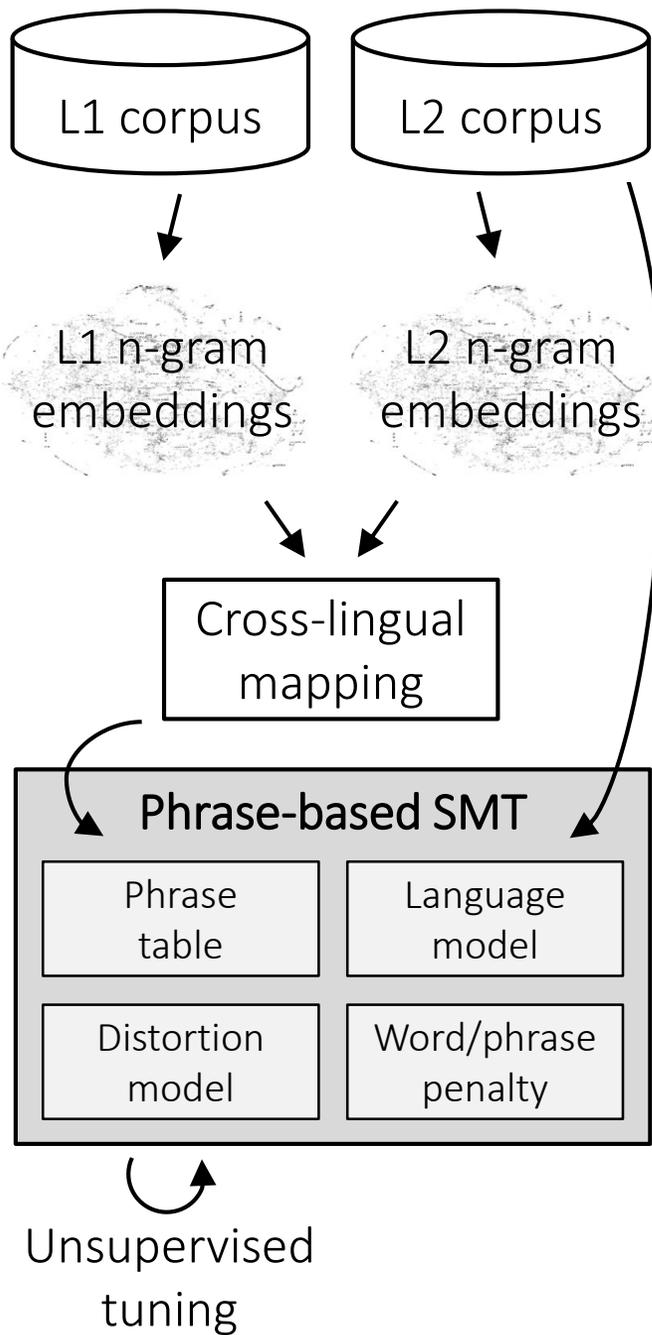


Unsupervised phrase-based SMT

L1 dev

L2 dev

MERT



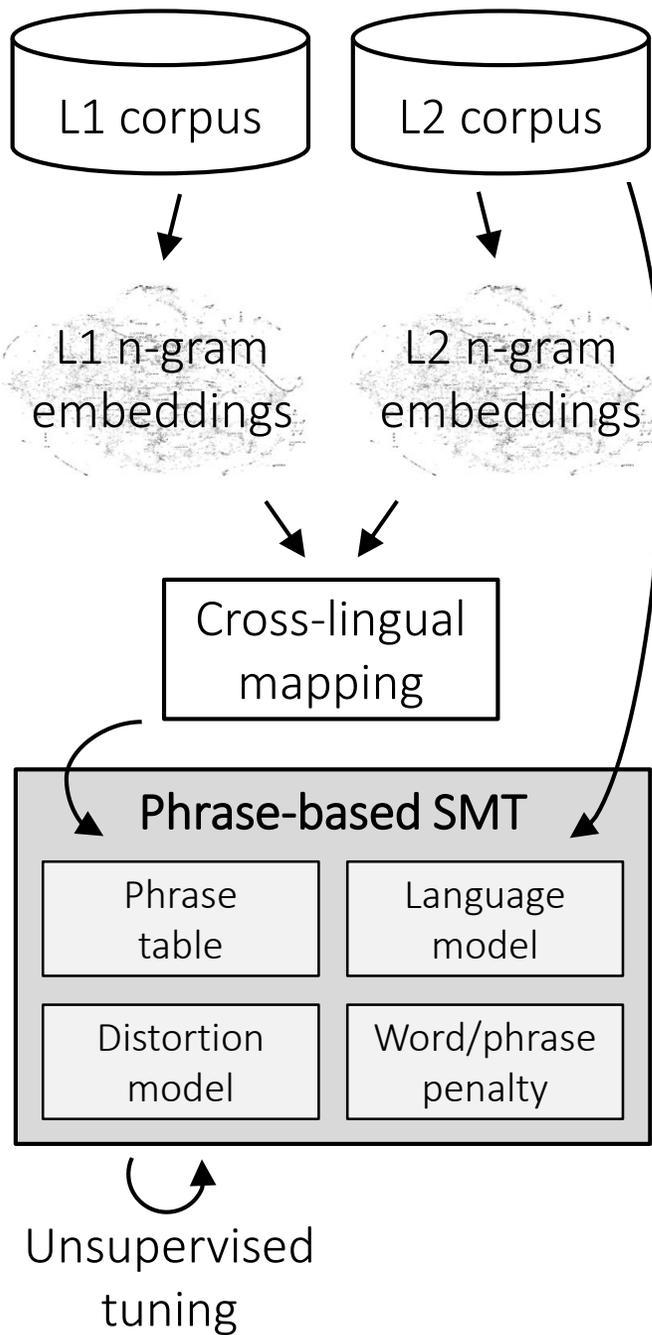
Unsupervised phrase-based SMT

L1 dev

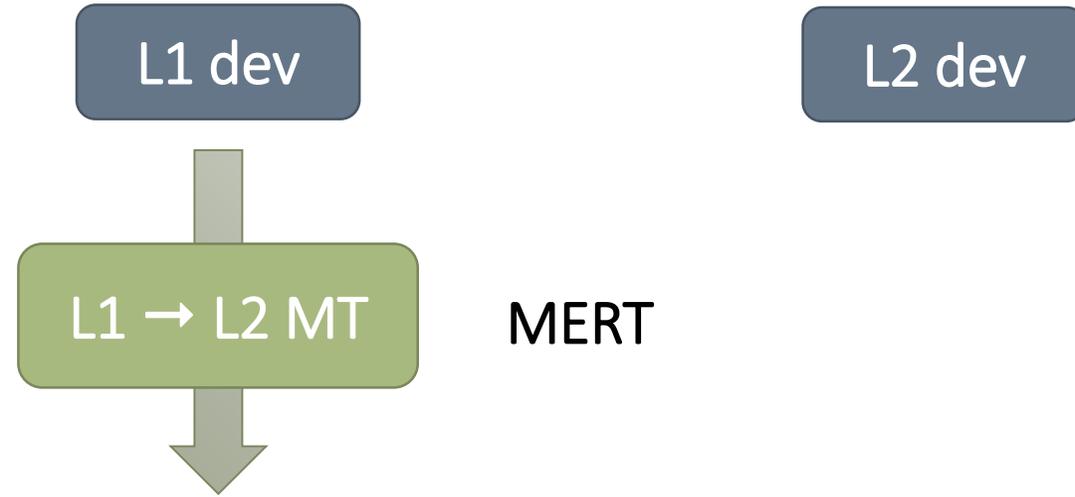
L2 dev

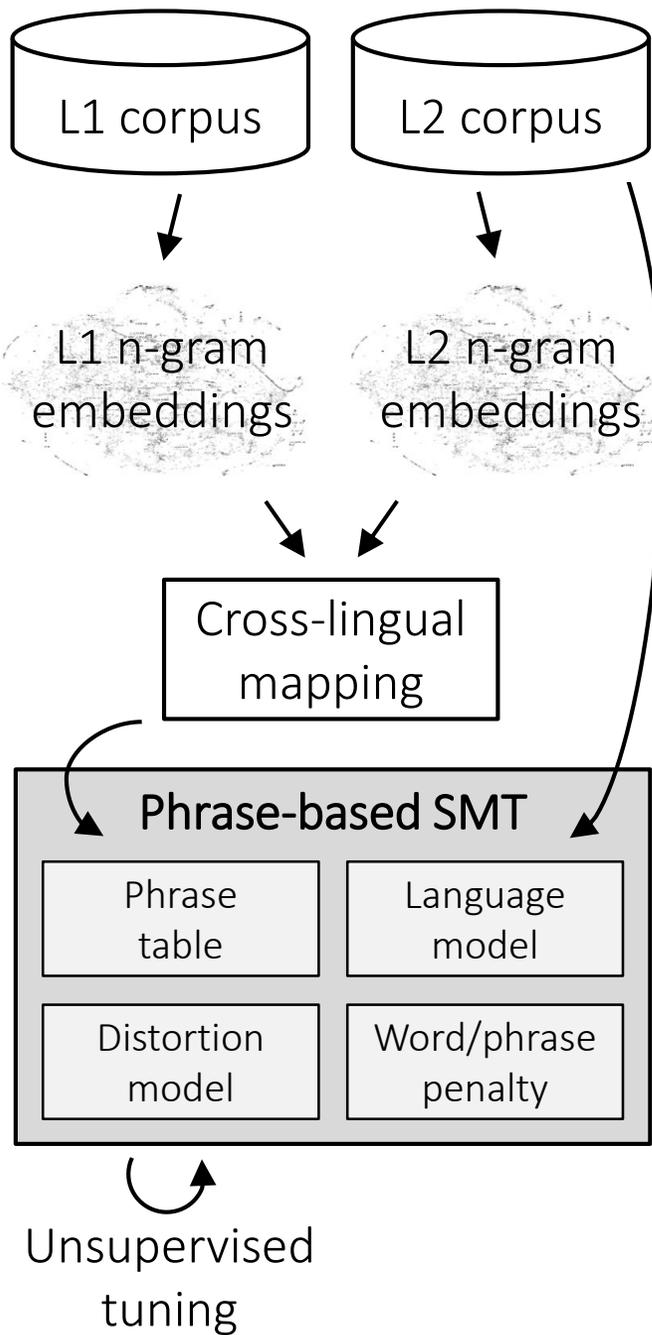
L1 → L2 MT

MERT

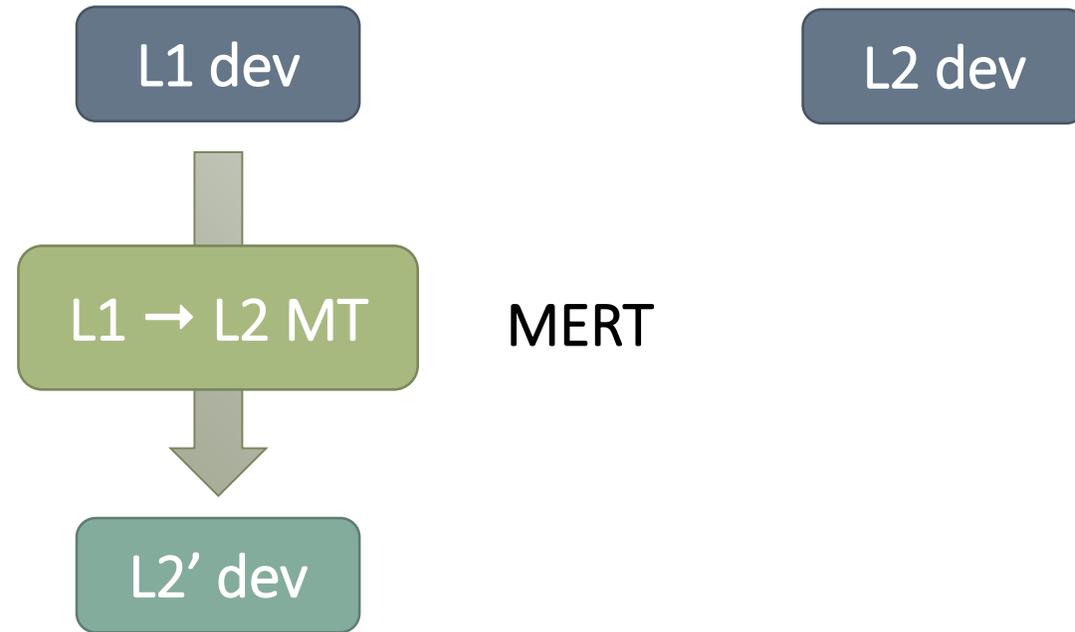


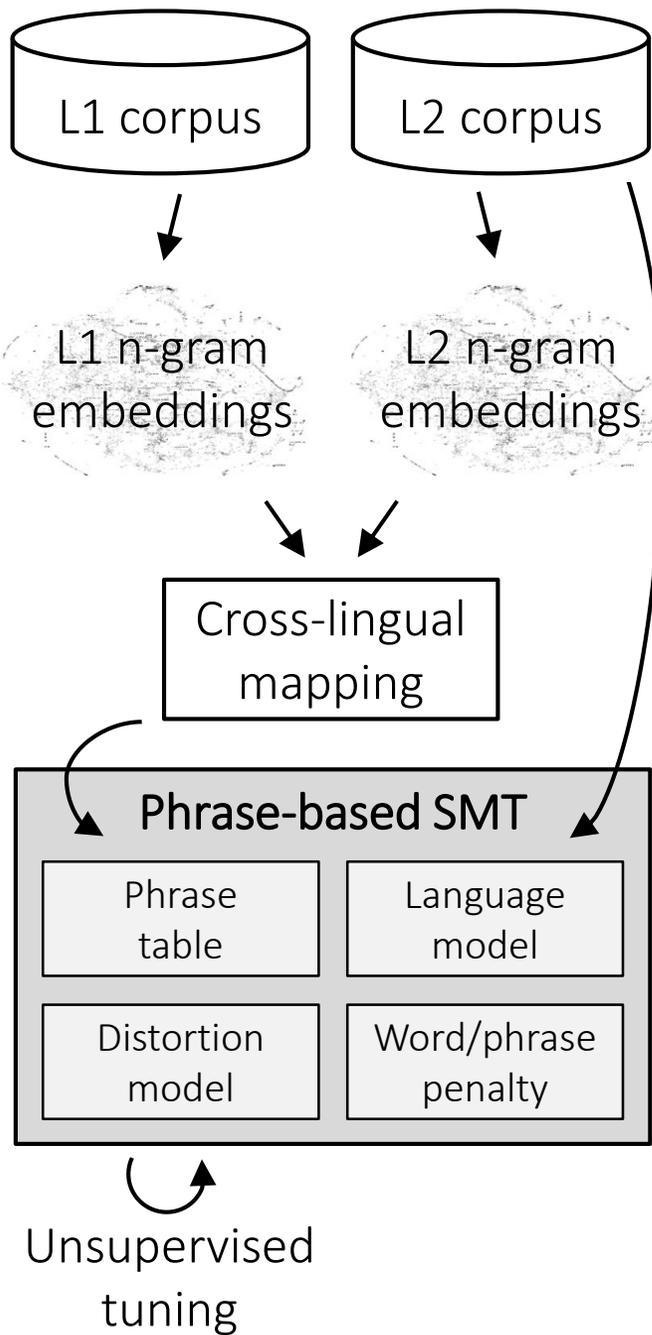
Unsupervised phrase-based SMT



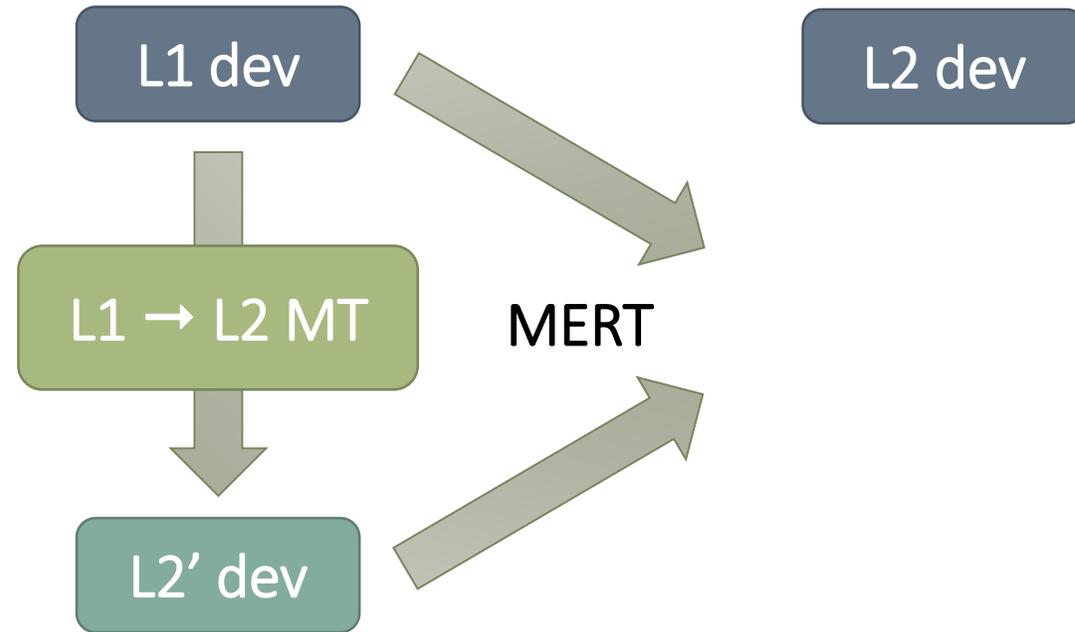


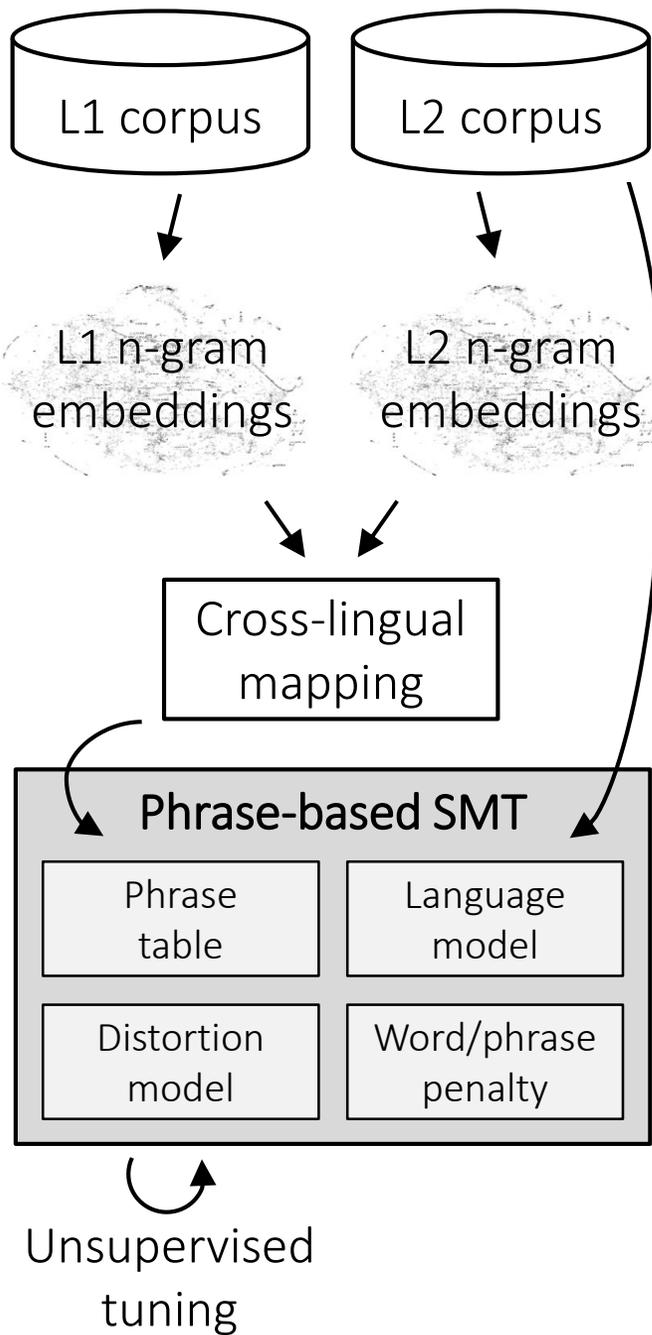
Unsupervised phrase-based SMT



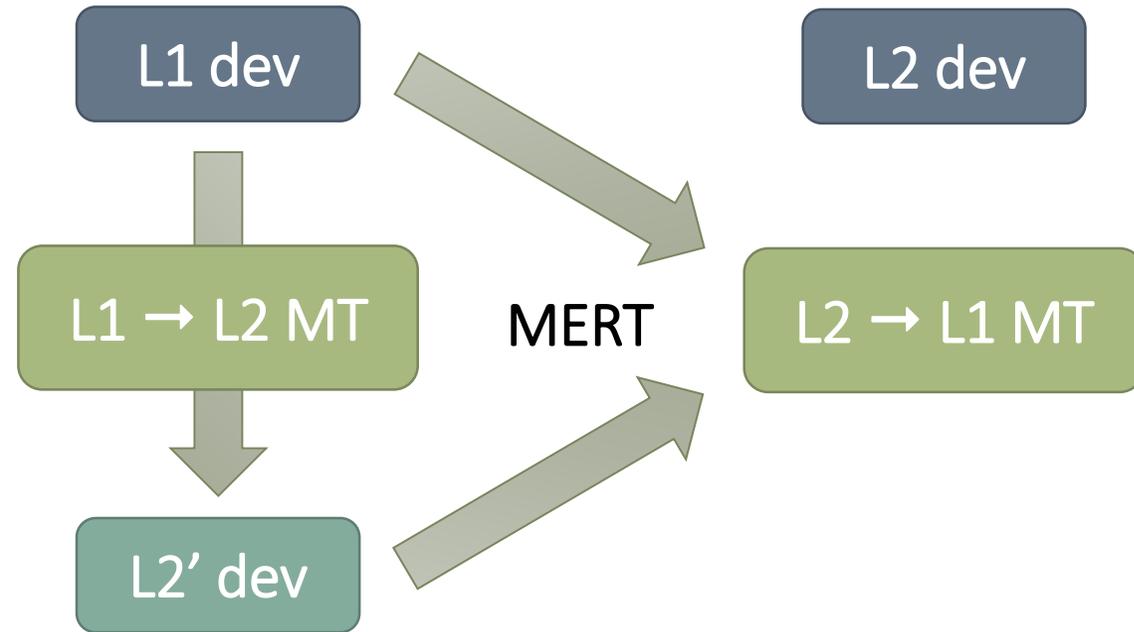


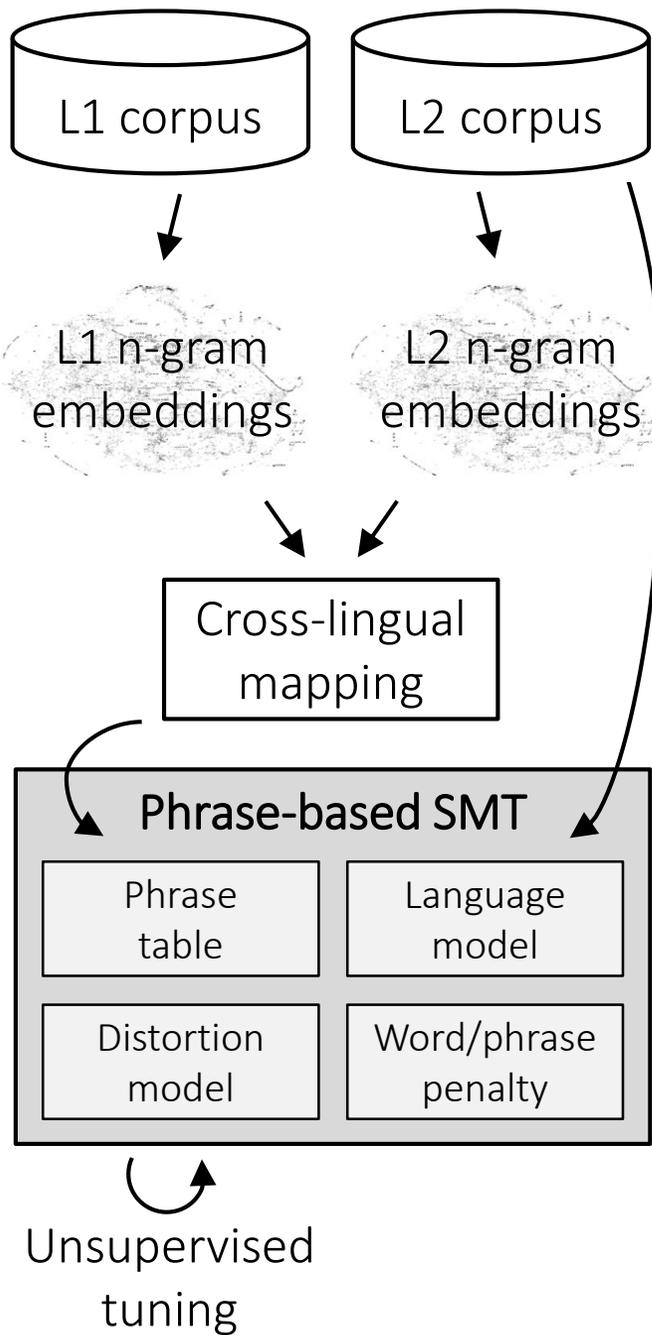
Unsupervised phrase-based SMT



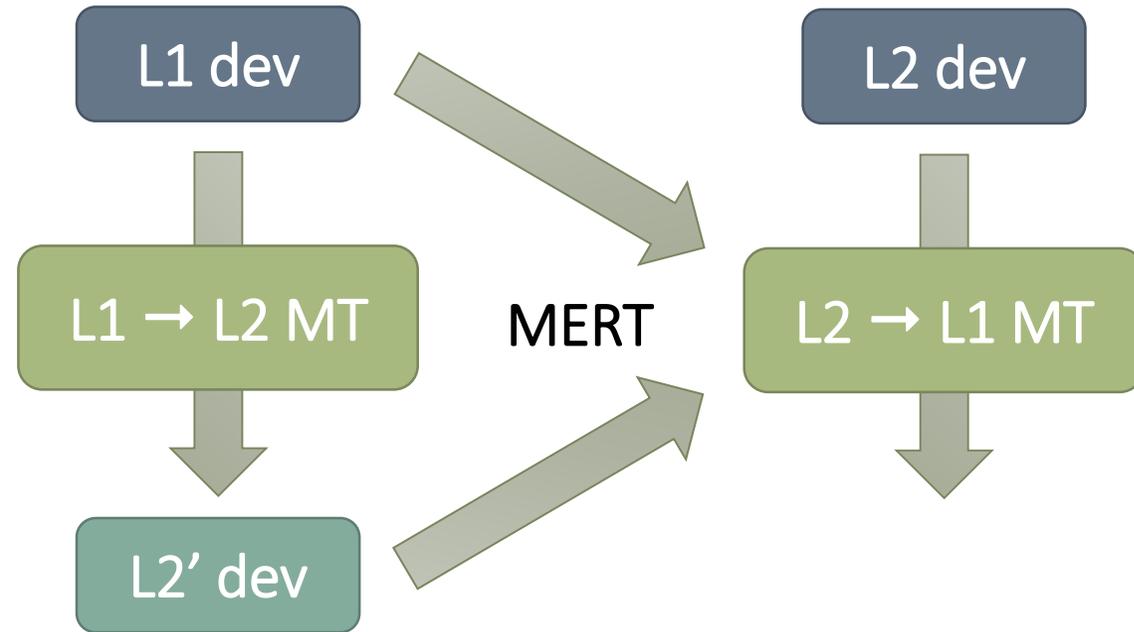


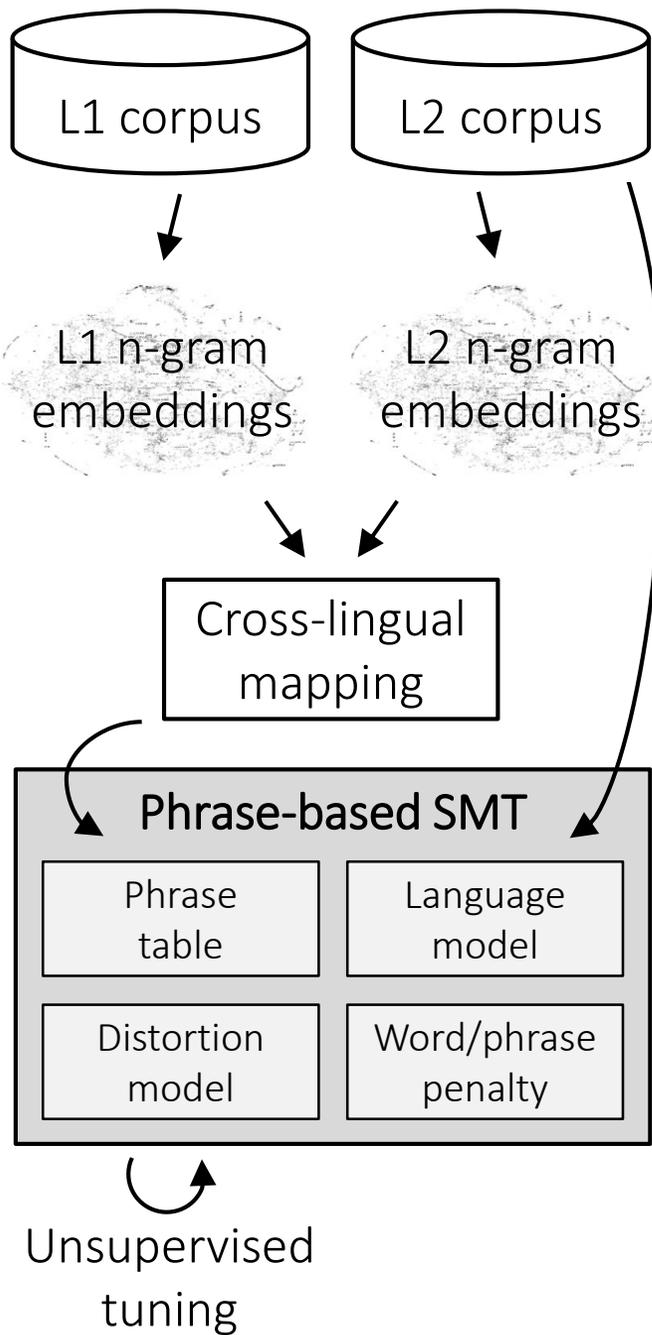
Unsupervised phrase-based SMT



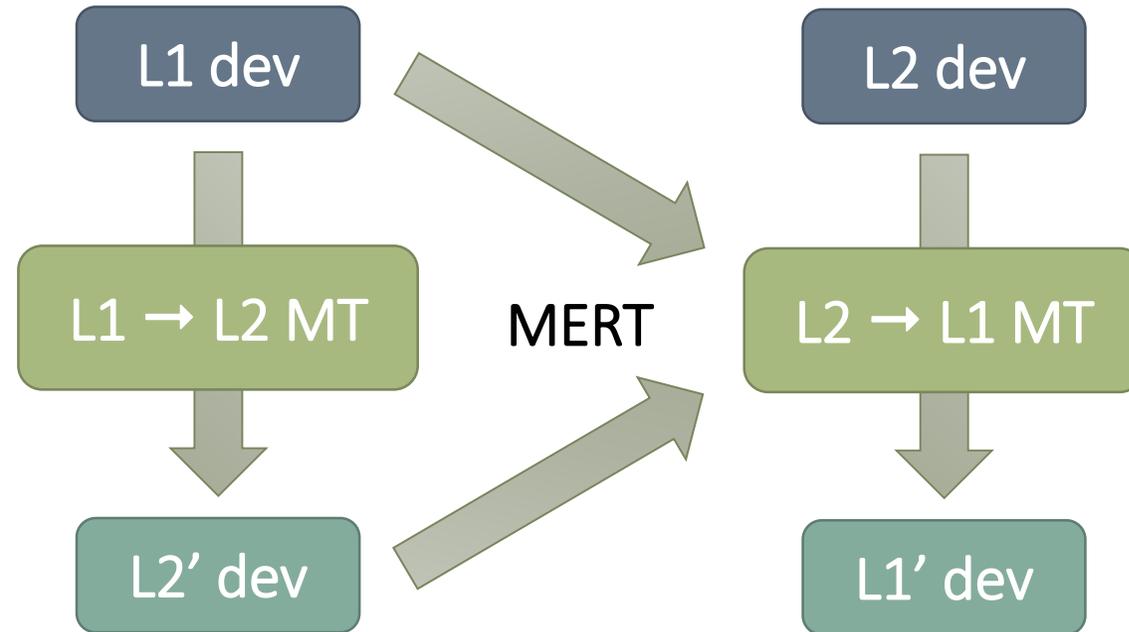


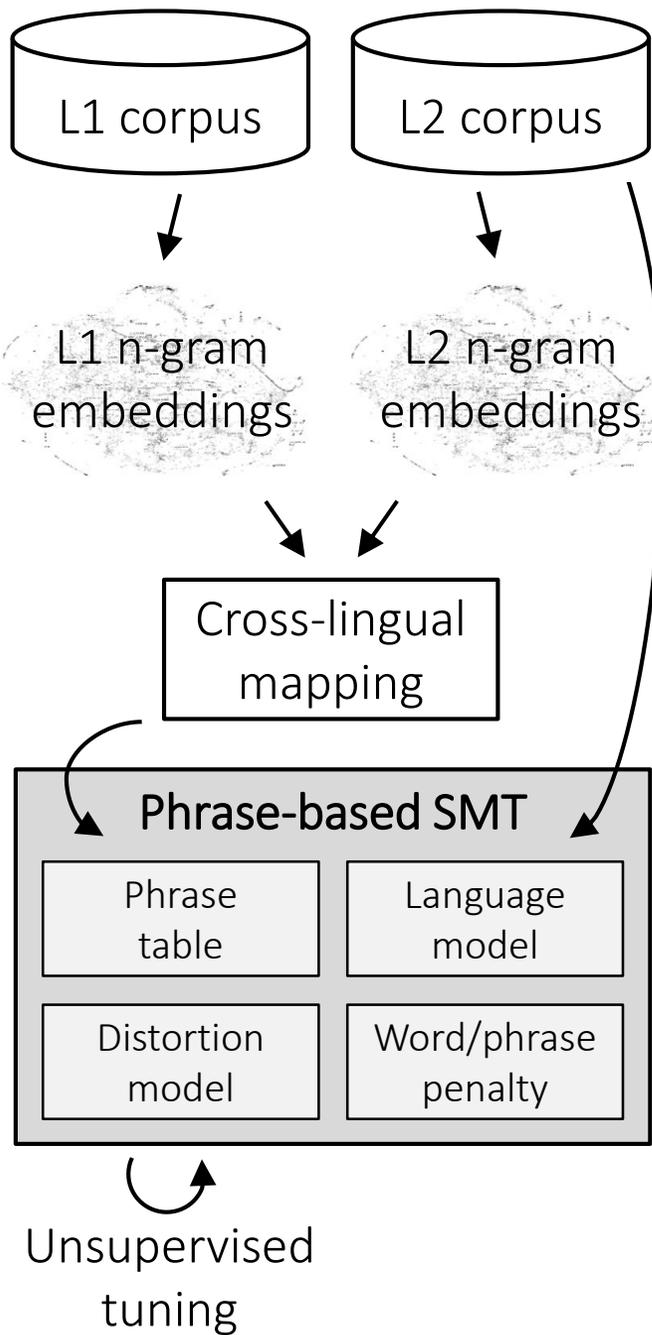
Unsupervised phrase-based SMT



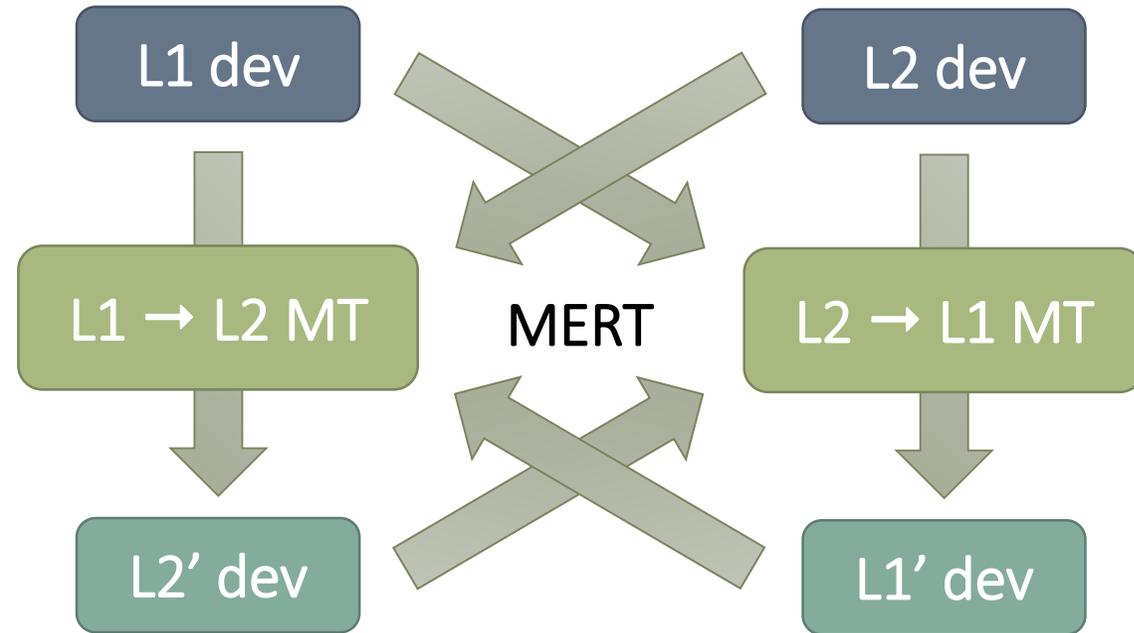


Unsupervised phrase-based SMT

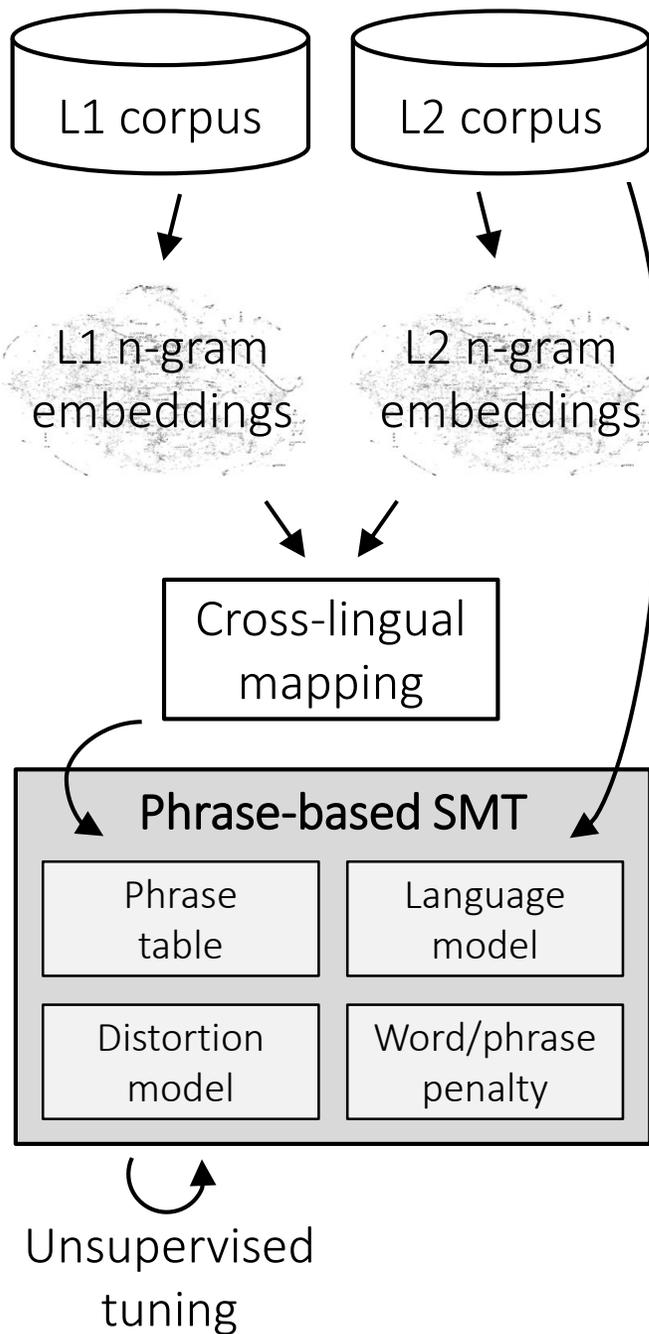




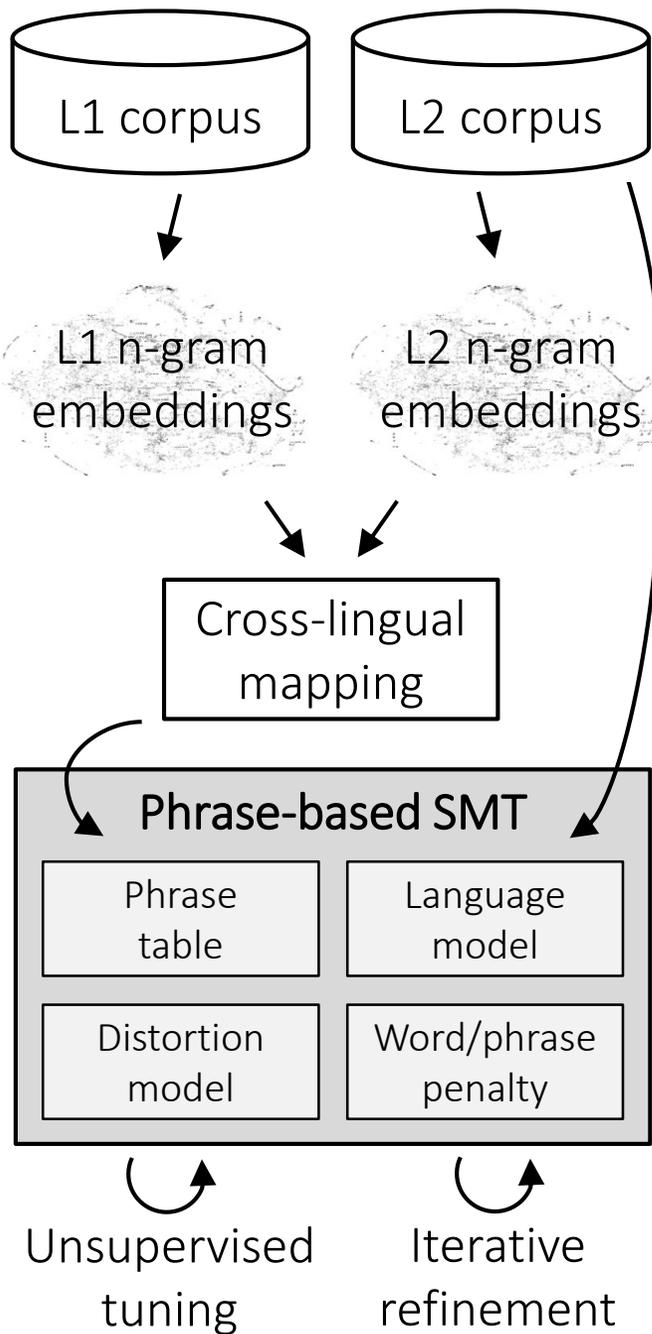
Unsupervised phrase-based SMT



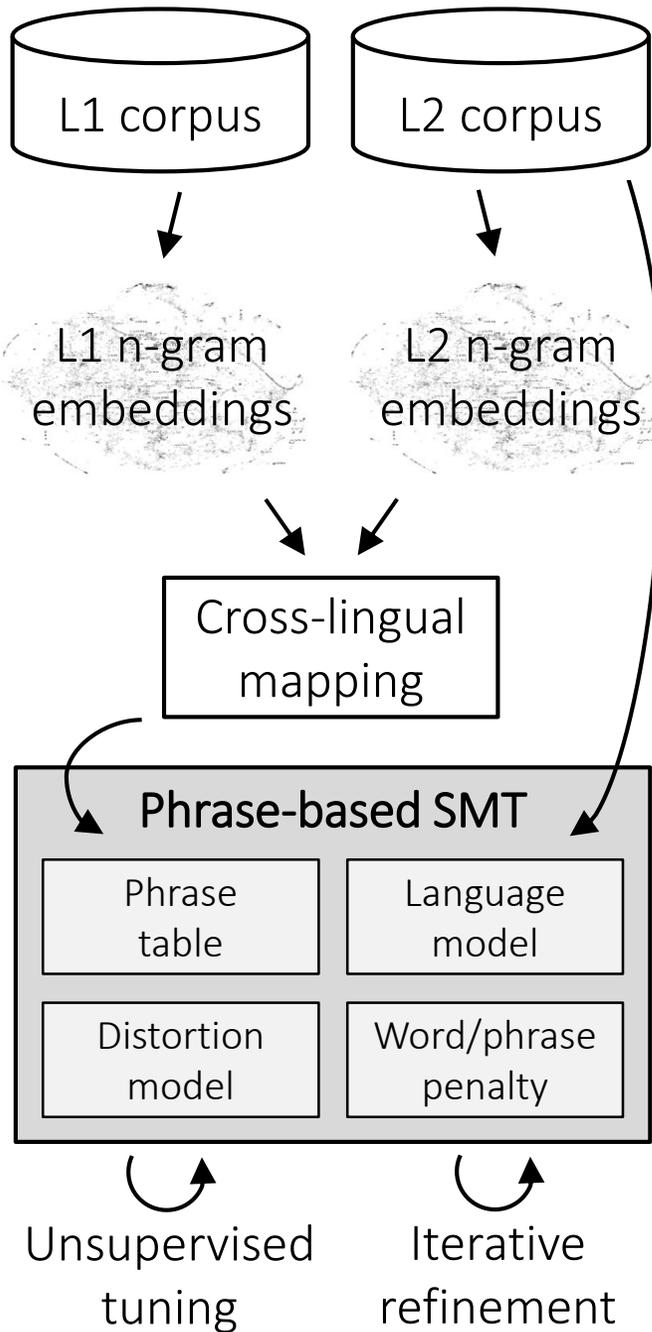
Unsupervised phrase-based SMT



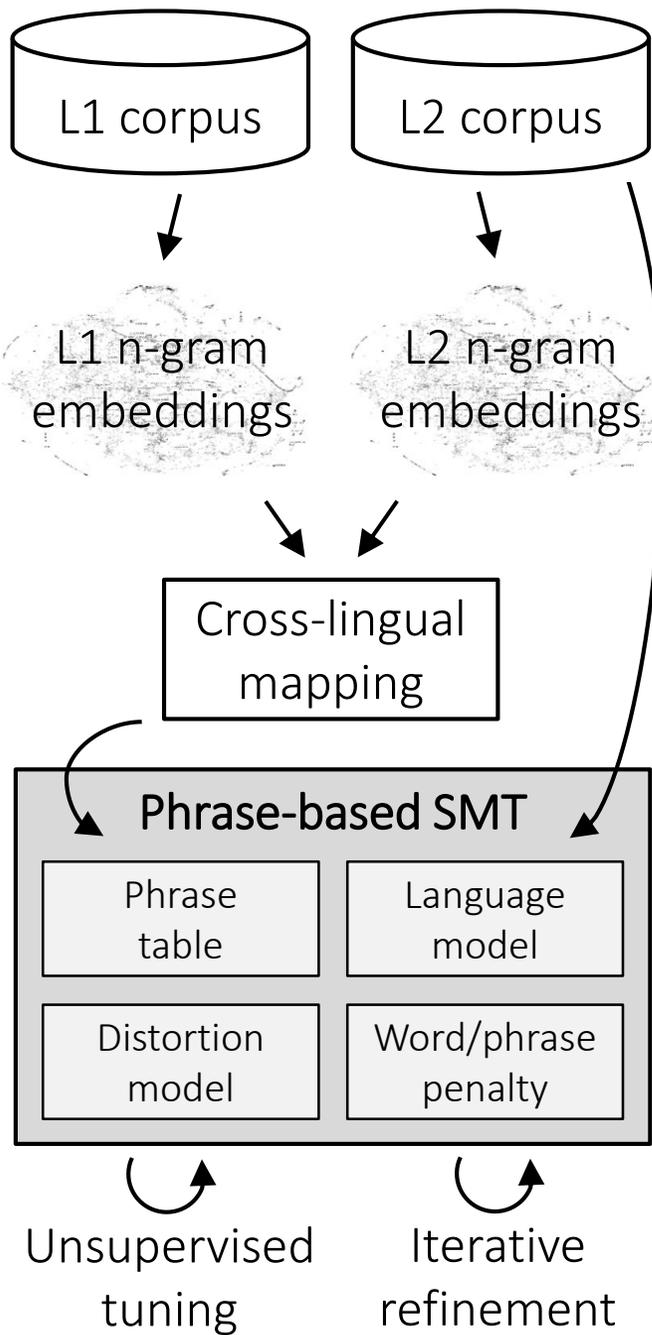
Unsupervised phrase-based SMT



Unsupervised phrase-based SMT

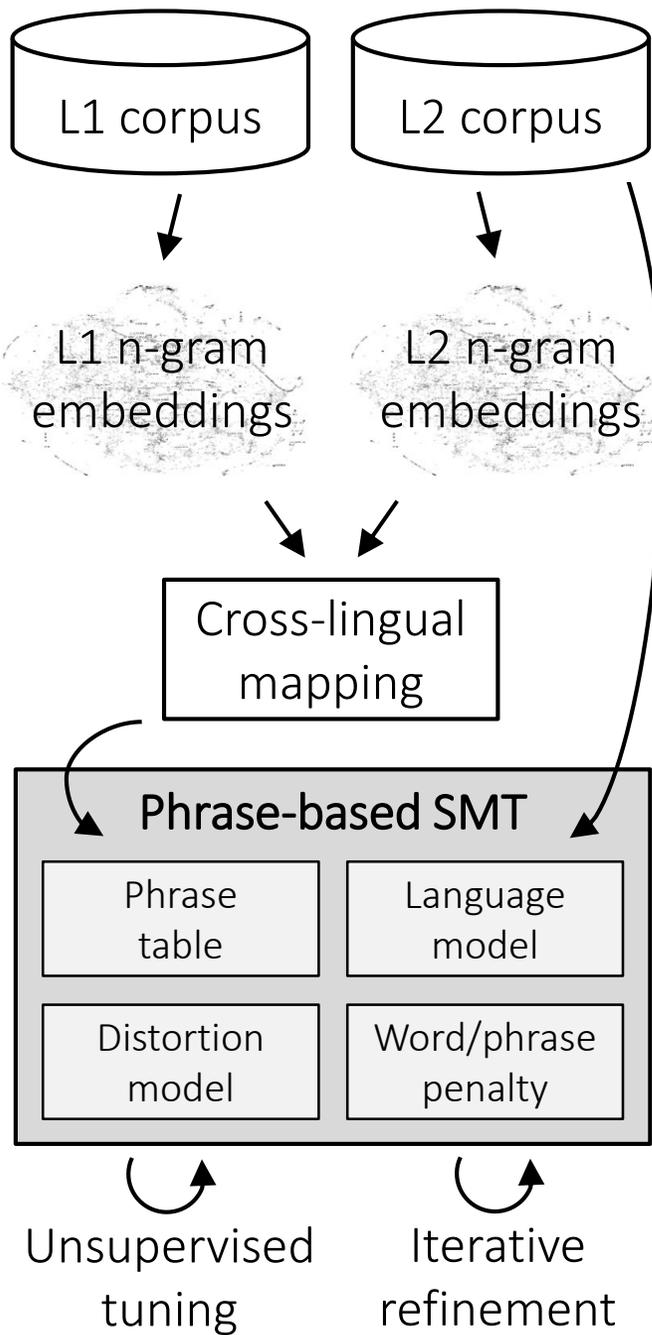


TRAIN

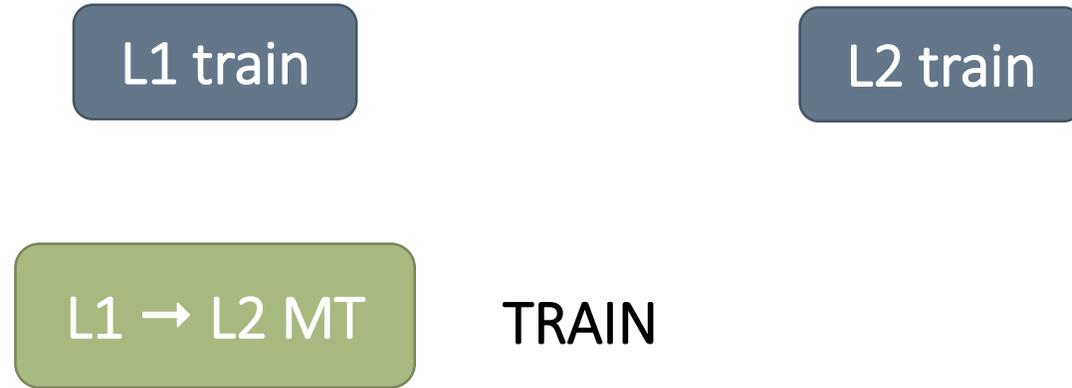


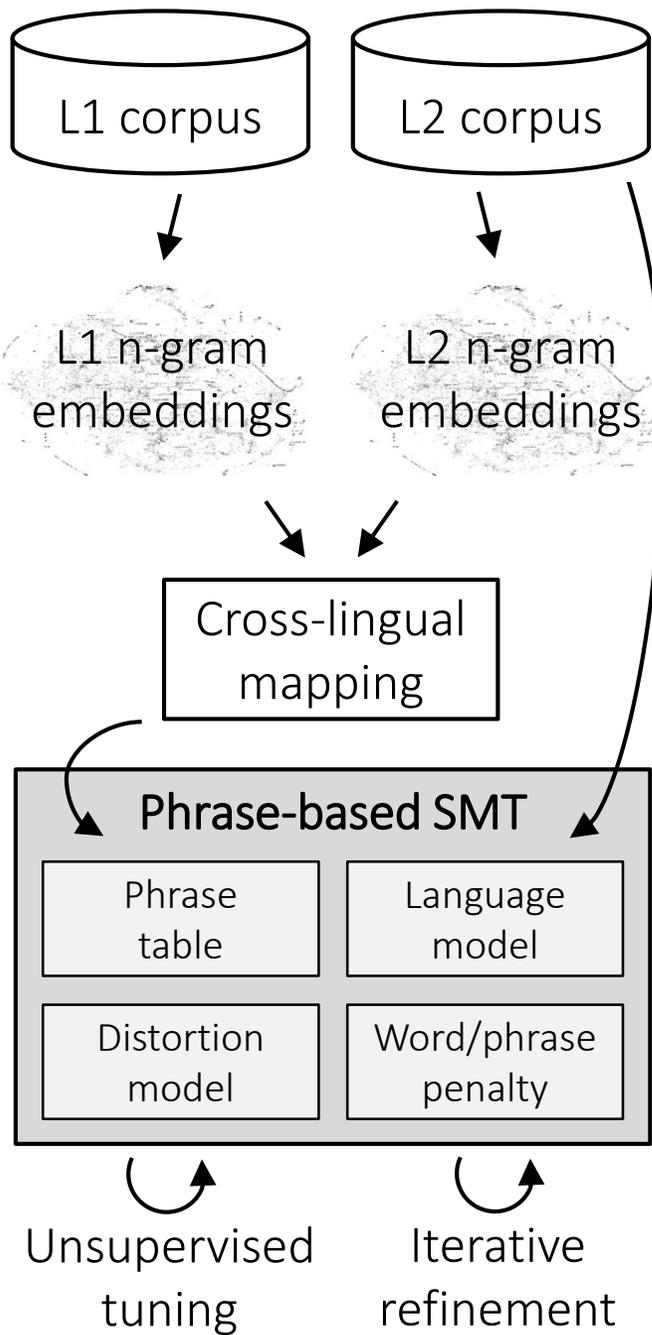
Unsupervised phrase-based SMT



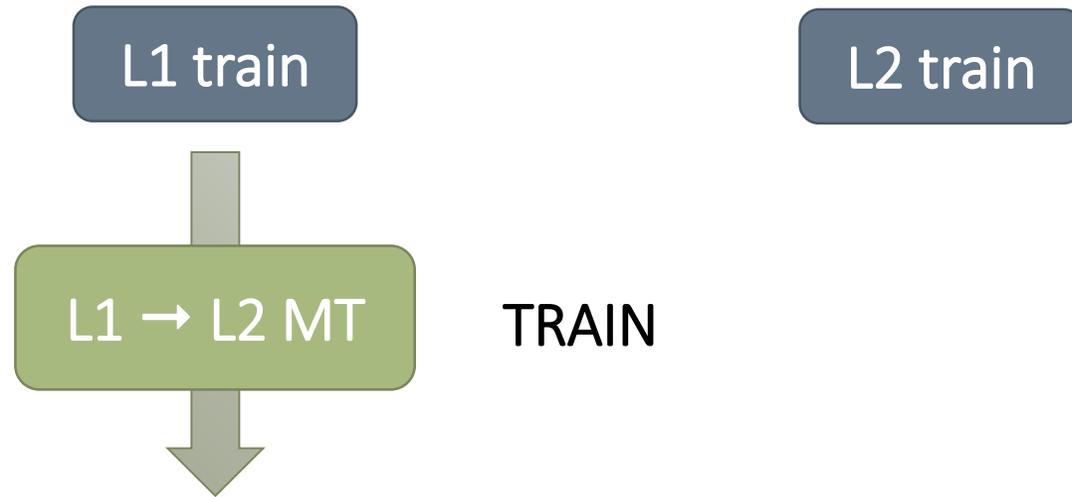


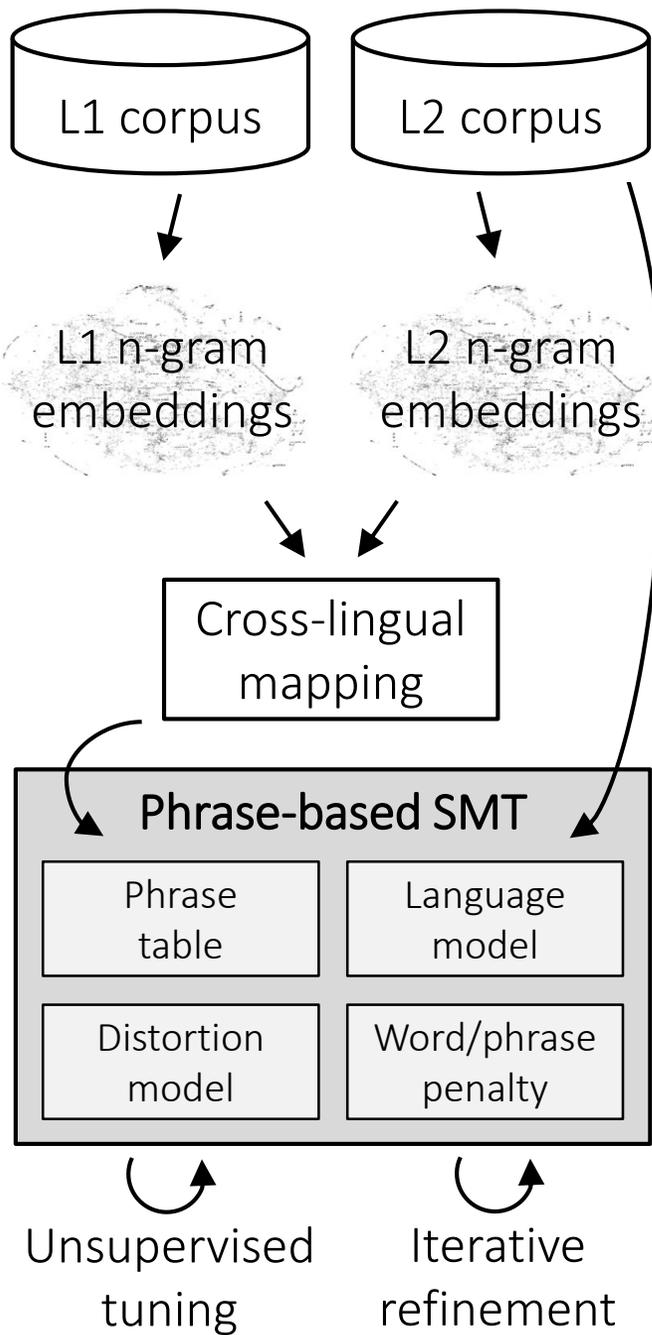
Unsupervised phrase-based SMT



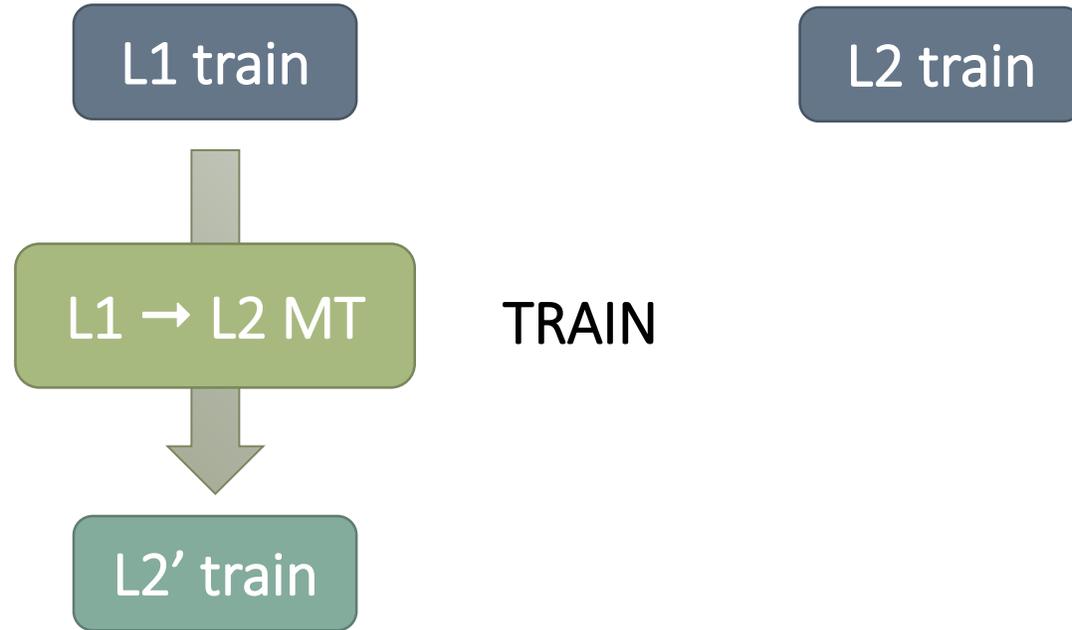


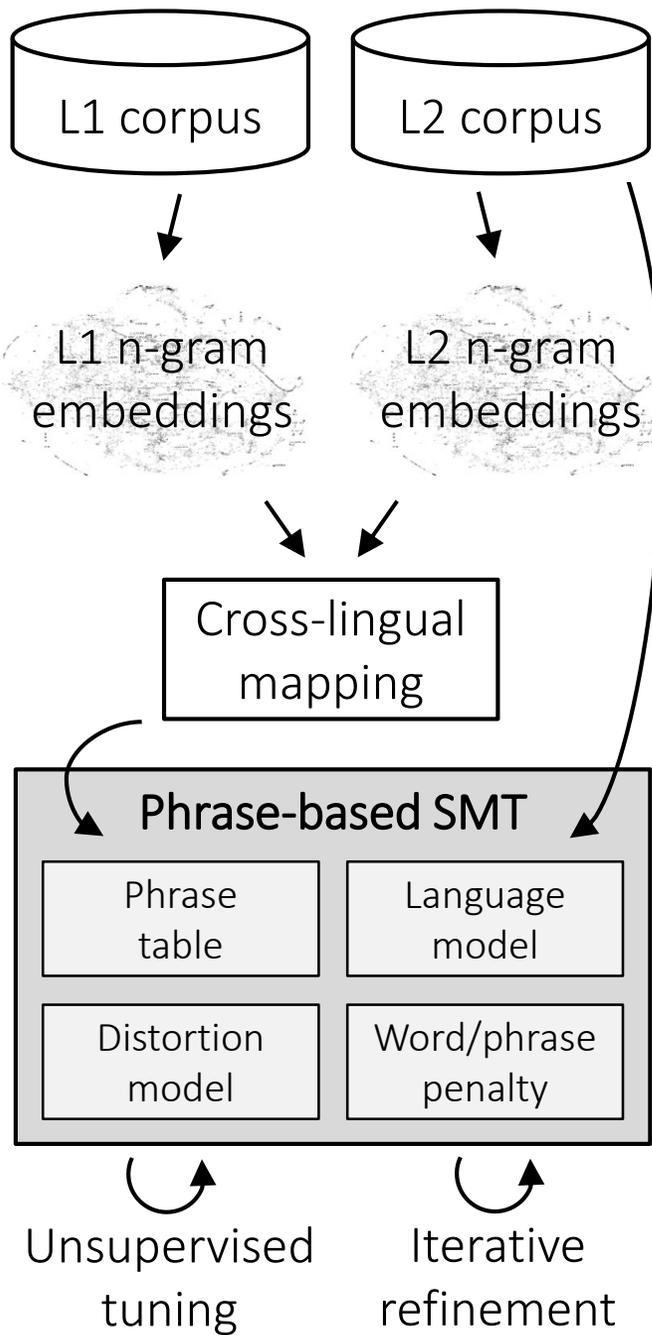
Unsupervised phrase-based SMT



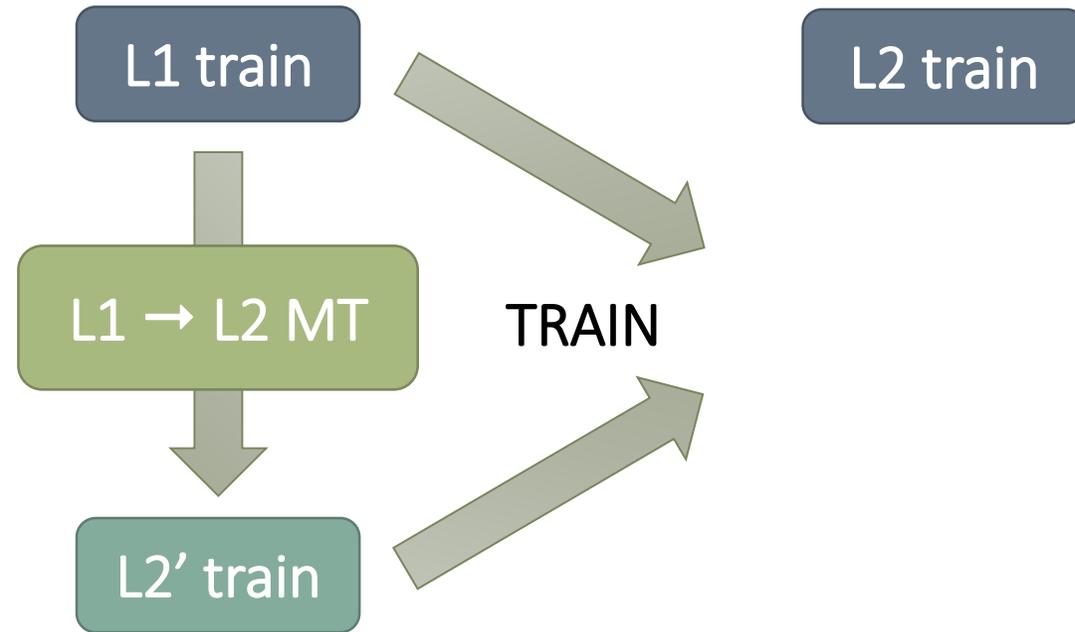


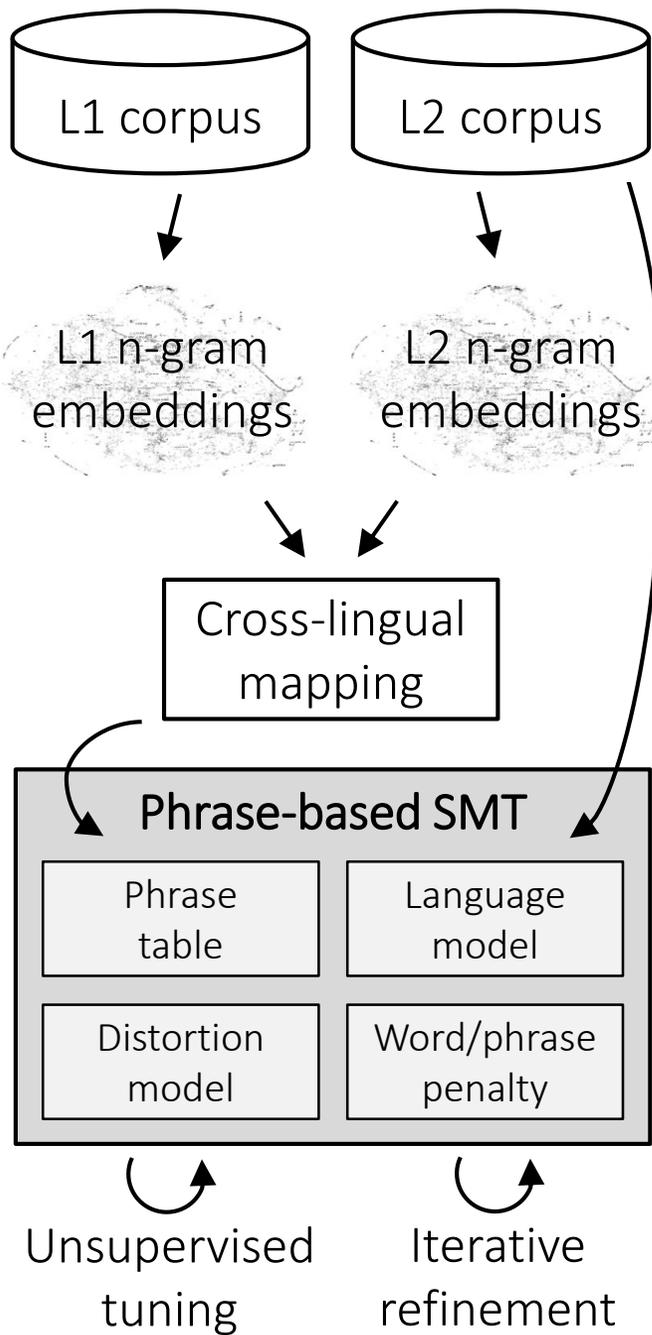
Unsupervised phrase-based SMT



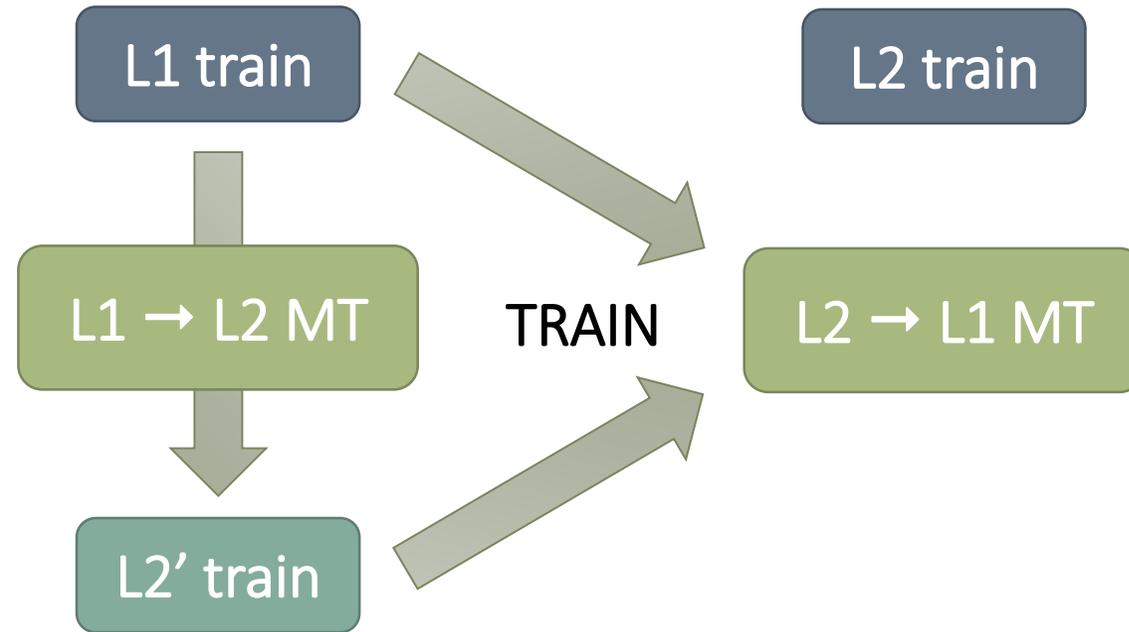


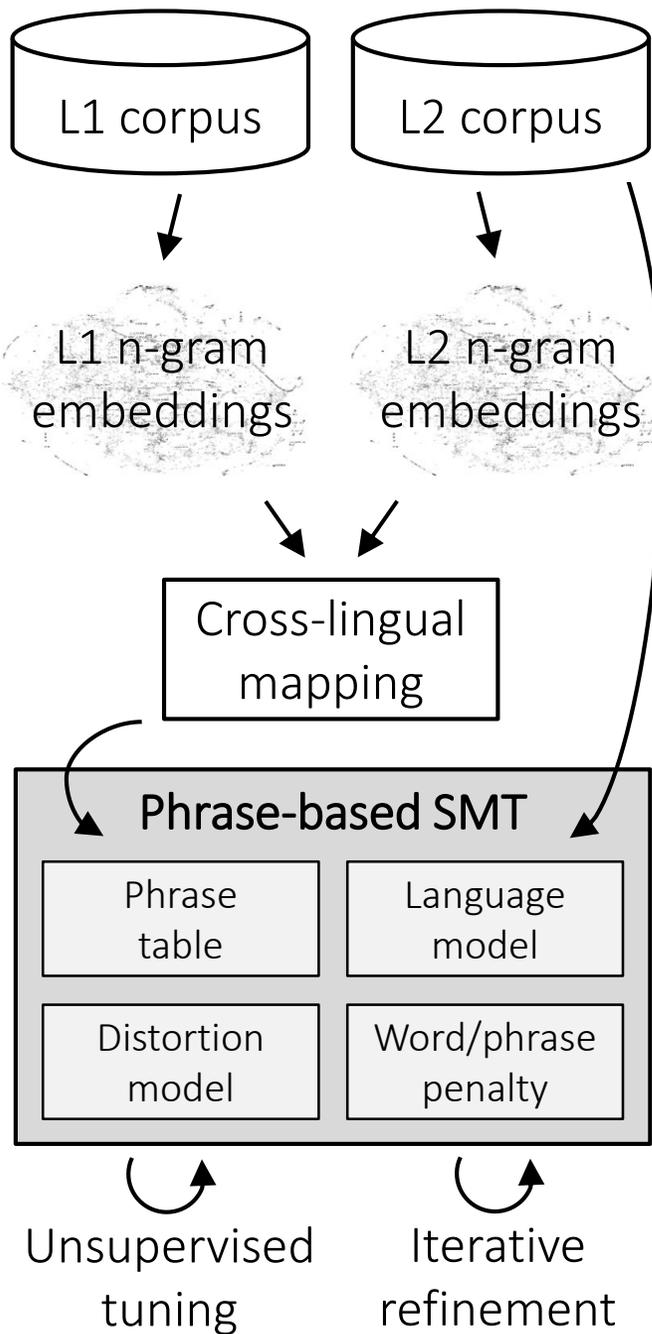
Unsupervised phrase-based SMT



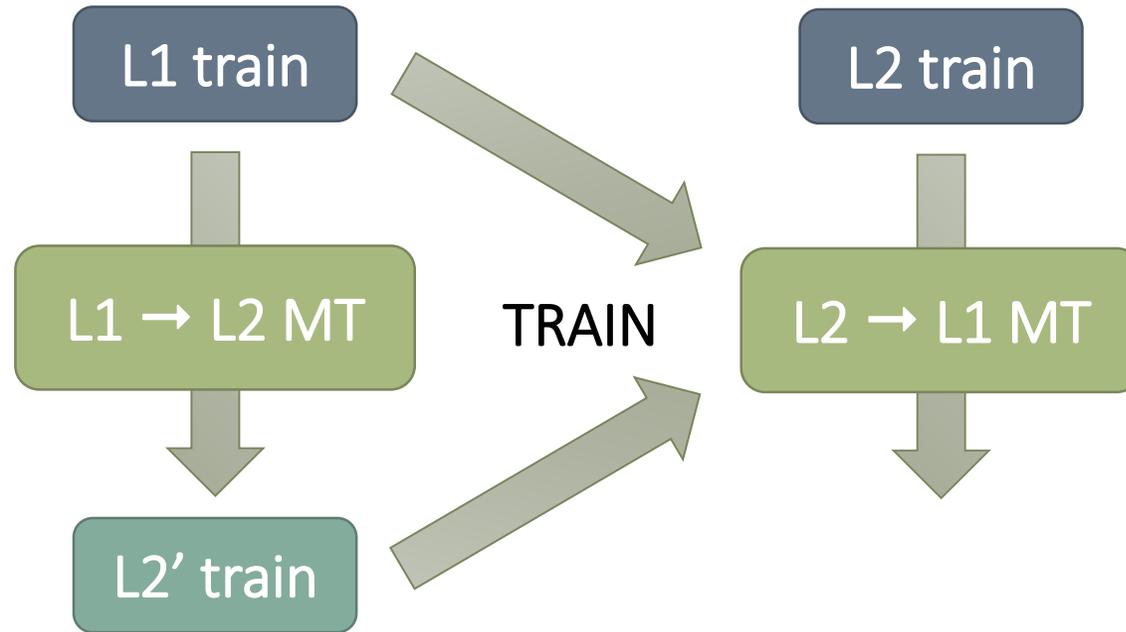


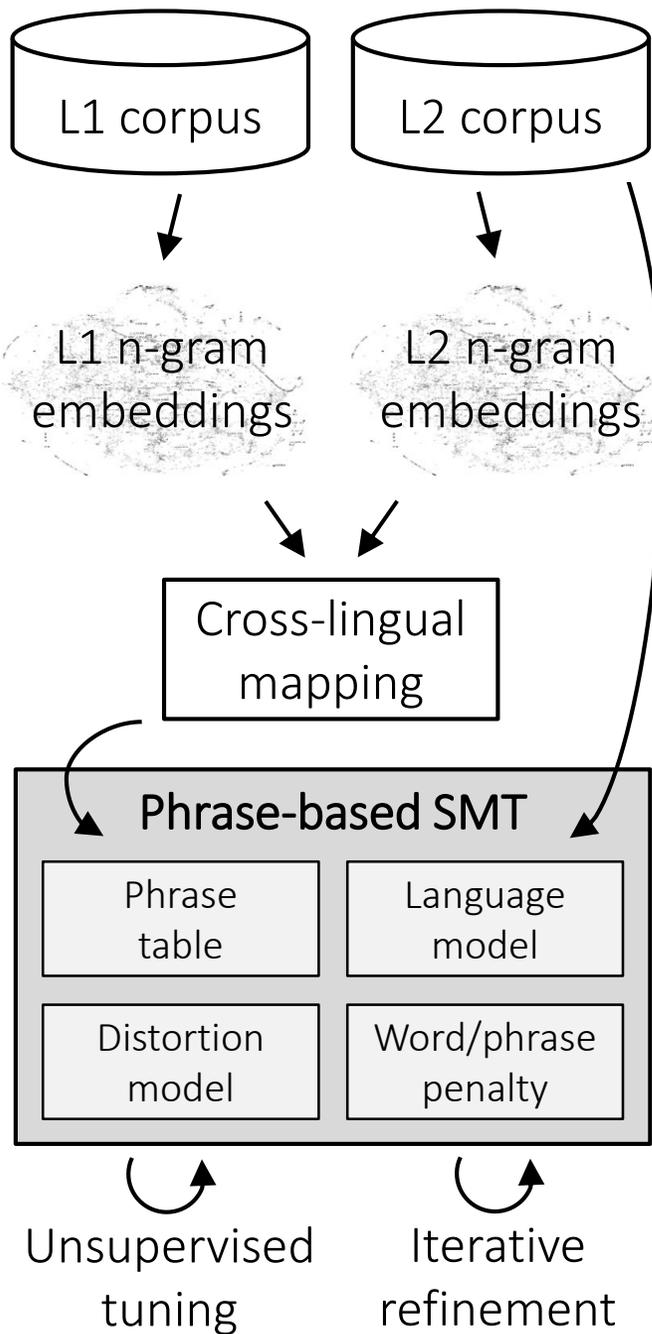
Unsupervised phrase-based SMT



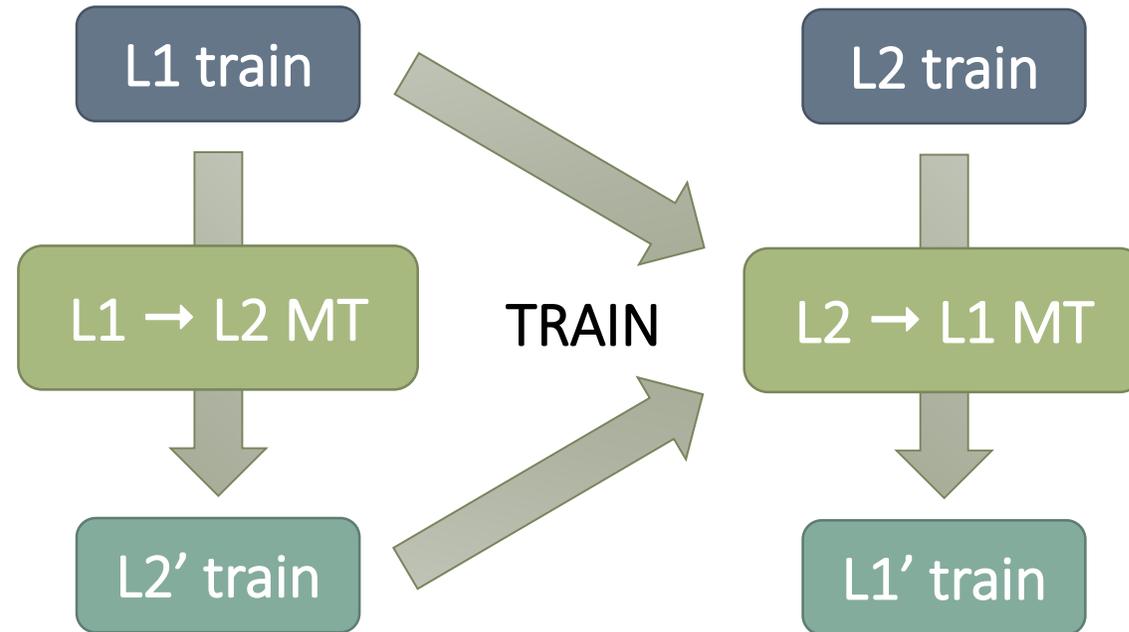


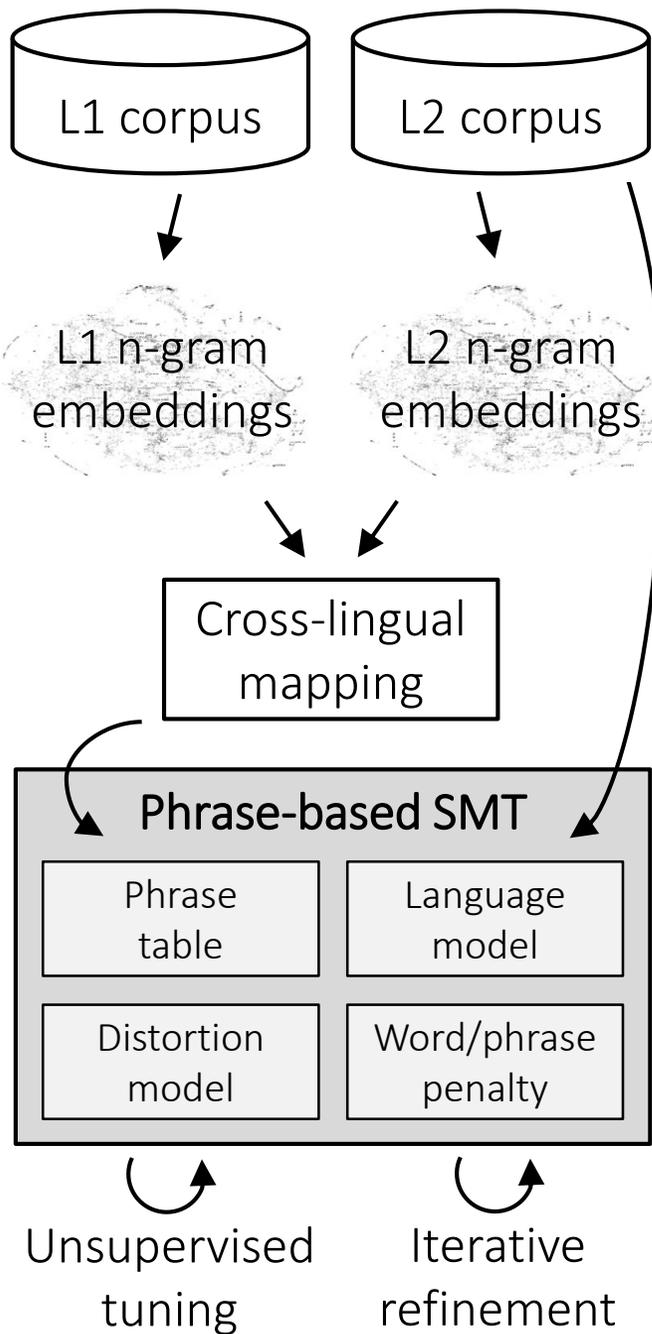
Unsupervised phrase-based SMT



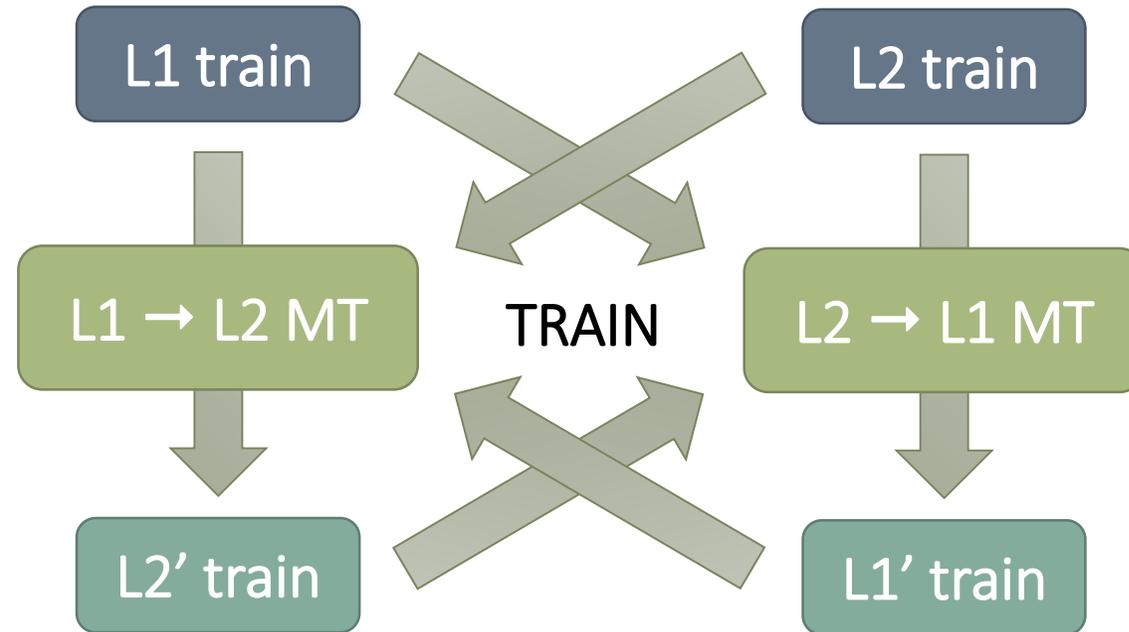


Unsupervised phrase-based SMT

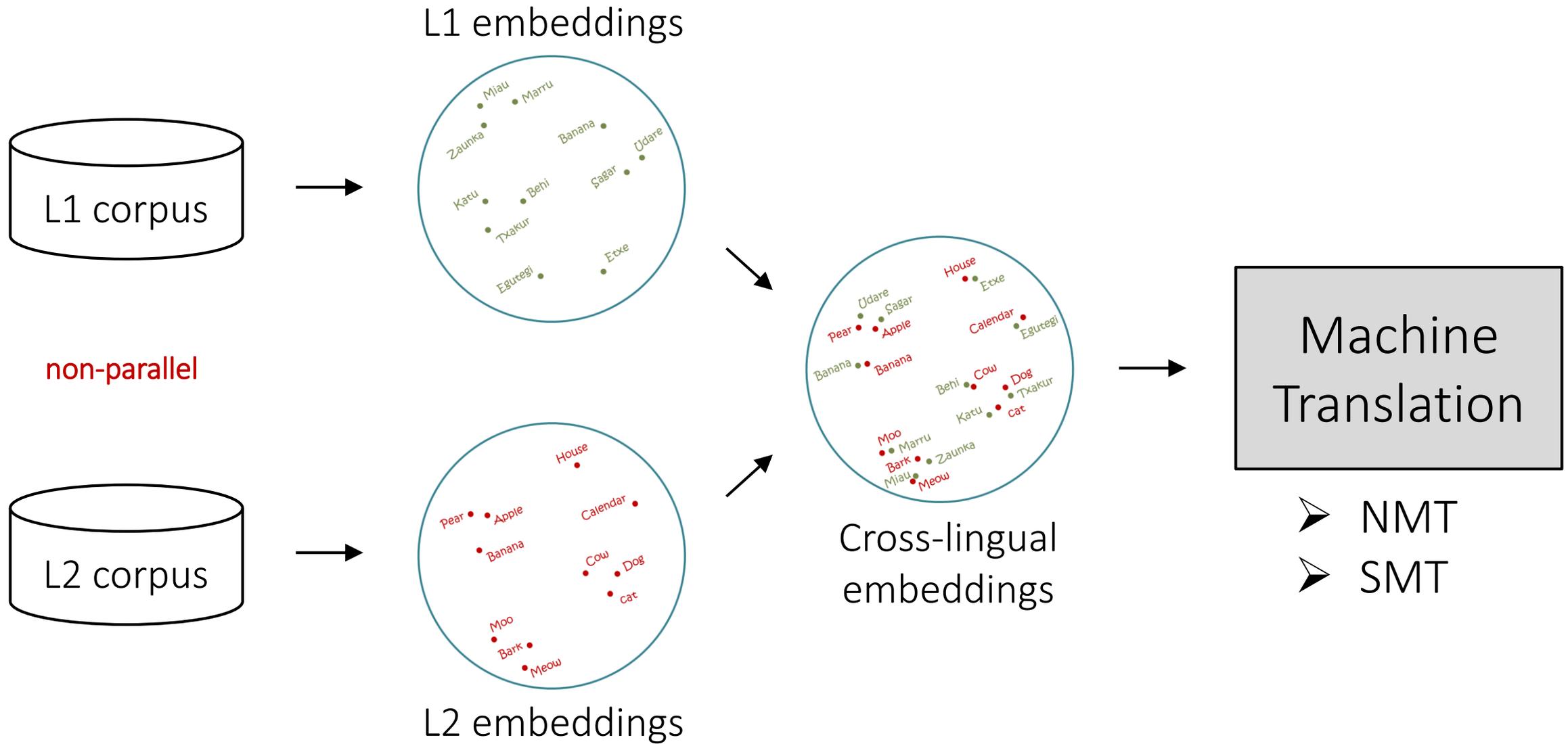




Unsupervised phrase-based SMT



Overview



Experiments

Experiments

- Languages: French-English, German-English

Experiments

- Languages: French-English, German-English
- Training: WMT-14 News Crawl

Experiments

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Evaluation: WMT-14 newstest (tok. BLEU)

Experiments

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Evaluation: WMT-14 newstest (tok. BLEU)

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

Experiments

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Evaluation: WMT-14 newstest (tok. BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Unsup. NMT	15.56	15.13	10.21	6.55

Experiments

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Evaluation: WMT-14 newstest (tok. BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Unsup. NMT	15.56	15.13	10.21	6.55
Unsup. SMT (initial)	21.16	20.13	13.86	10.59

Experiments

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Evaluation: WMT-14 newstest (tok. BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Unsup. NMT	15.56	15.13	10.21	6.55
Unsup. SMT (initial)	21.16	20.13	13.86	10.59
+ Unsupervised tuning	22.17	22.22	14.73	10.64

Experiments

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Evaluation: WMT-14 newstest (tok. BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Unsup. NMT	15.56	15.13	10.21	6.55
Unsup. SMT (initial)	21.16	20.13	13.86	10.59
+ Unsupervised tuning	22.17	22.22	14.73	10.64
+ Iterative refinement	25.87	26.22	17.43	14.08

Experiments

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Evaluation: WMT-14 newstest (tok. BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Unsup. NMT	15.56	15.13	10.21	6.55
Unsup. SMT (initial)	21.16	20.13	13.86	10.59
+ Unsupervised tuning	22.17	22.22	14.73	10.64
+ Iterative refinement	25.87	26.22	17.43	14.08
Sup. SMT (Europarl-Moses)	30.61	30.82	20.83	16.60

Experiments

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Evaluation: WMT-14 newstest (tok. BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Unsup. NMT	15.56	15.13	10.21	6.55
Unsup. SMT (initial)	21.16	20.13	13.86	10.59
+ Unsupervised tuning	22.17	22.22	14.73	10.64
+ Iterative refinement	25.87	26.22	17.43	14.08
Sup. SMT (Europarl-Moses)	30.61	30.82	20.83	16.60
Sup. NMT (Transformer)	-	41.8	-	28.4

Conclusions

Conclusions

Unsupervised machine translation

Conclusions

Unsupervised machine translation

- Map independently learned embeddings to a cross-lingual space

Conclusions

Unsupervised machine translation

- Map independently learned embeddings to a cross-lingual space
- Generalize from word level to sentence level translation (NMT/SMT)

Conclusions

Unsupervised machine translation

- Map independently learned embeddings to a cross-lingual space
- Generalize from word level to sentence level translation (NMT/SMT)

Strong results

Conclusions

Unsupervised machine translation

- Map independently learned embeddings to a cross-lingual space
- Generalize from word level to sentence level translation (NMT/SMT)

Strong results

- Gap with supervised SMT (Europarl Moses) down to 2-5 BLEU

Conclusions

Unsupervised machine translation

- Map independently learned embeddings to a cross-lingual space
- Generalize from word level to sentence level translation (NMT/SMT)

Strong results

- Gap with supervised SMT (Europarl Moses) down to 2-5 BLEU
- See also Lample et al. (2018a,b)

Conclusions

Unsupervised machine translation

- Map independently learned embeddings to a cross-lingual space
- Generalize from word level to sentence level translation (NMT/SMT)

Strong results

- Gap with supervised SMT (Europarl Moses) down to 2-5 BLEU
- See also Lample et al. (2018a,b)

Future work

Conclusions

Unsupervised machine translation

- Map independently learned embeddings to a cross-lingual space
- Generalize from word level to sentence level translation (NMT/SMT)

Strong results

- Gap with supervised SMT (Europarl Moses) down to 2-5 BLEU
- See also Lample et al. (2018a,b)

Future work

- Semi-supervised with small parallel corpora

Conclusions

Unsupervised machine translation

- Map independently learned embeddings to a cross-lingual space
- Generalize from word level to sentence level translation (NMT/SMT)

Strong results

- Gap with supervised SMT (Europarl Moses) down to 2-5 BLEU
- See also Lample et al. (2018a,b)

Future work

- Semi-supervised with small parallel corpora
- SMT/NMT hybridization

Thank you!