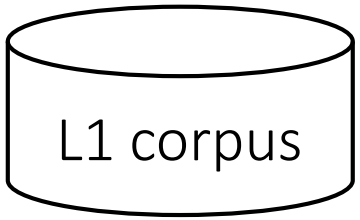
Unsupervised Machine Translation

Mikel Artetxe

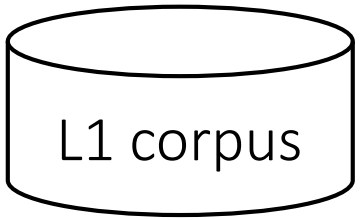
IXA NLP group – University of the Basque Country (UPV/EHU)

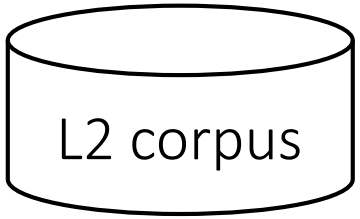
Introduction

Introduction

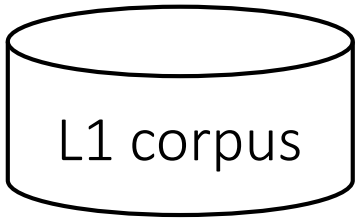


Introduction

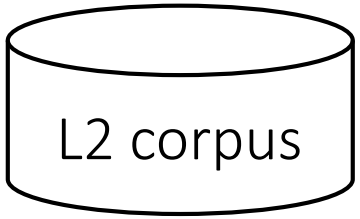




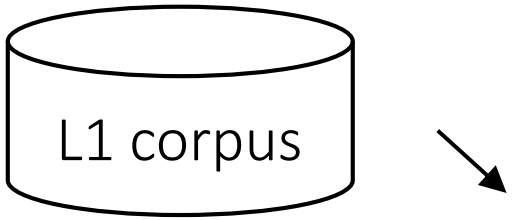
Introduction



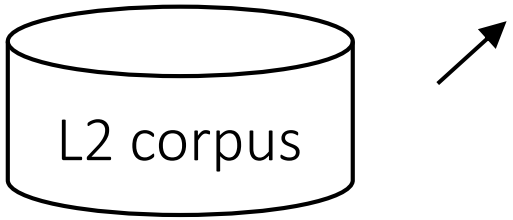
non-parallel



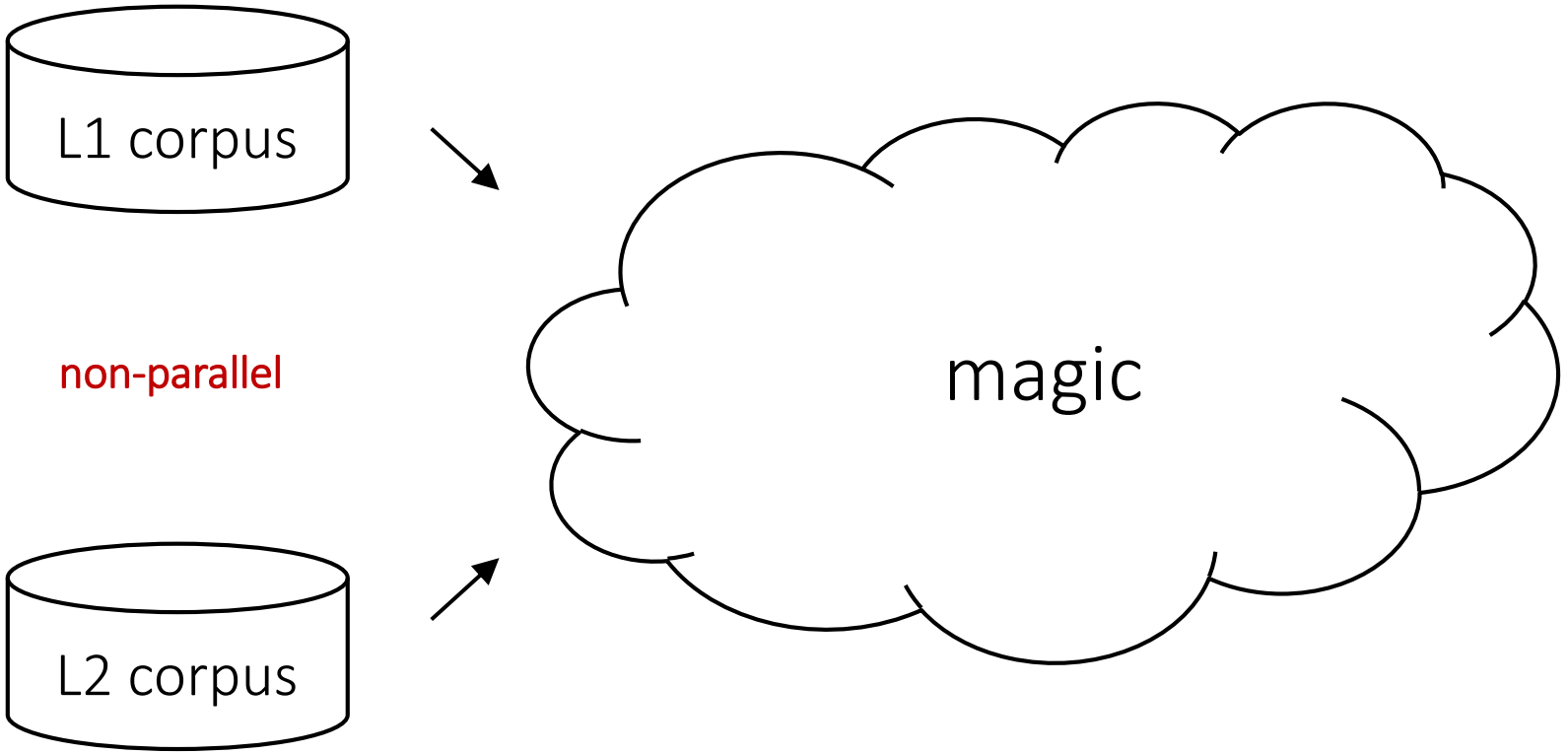
Introduction



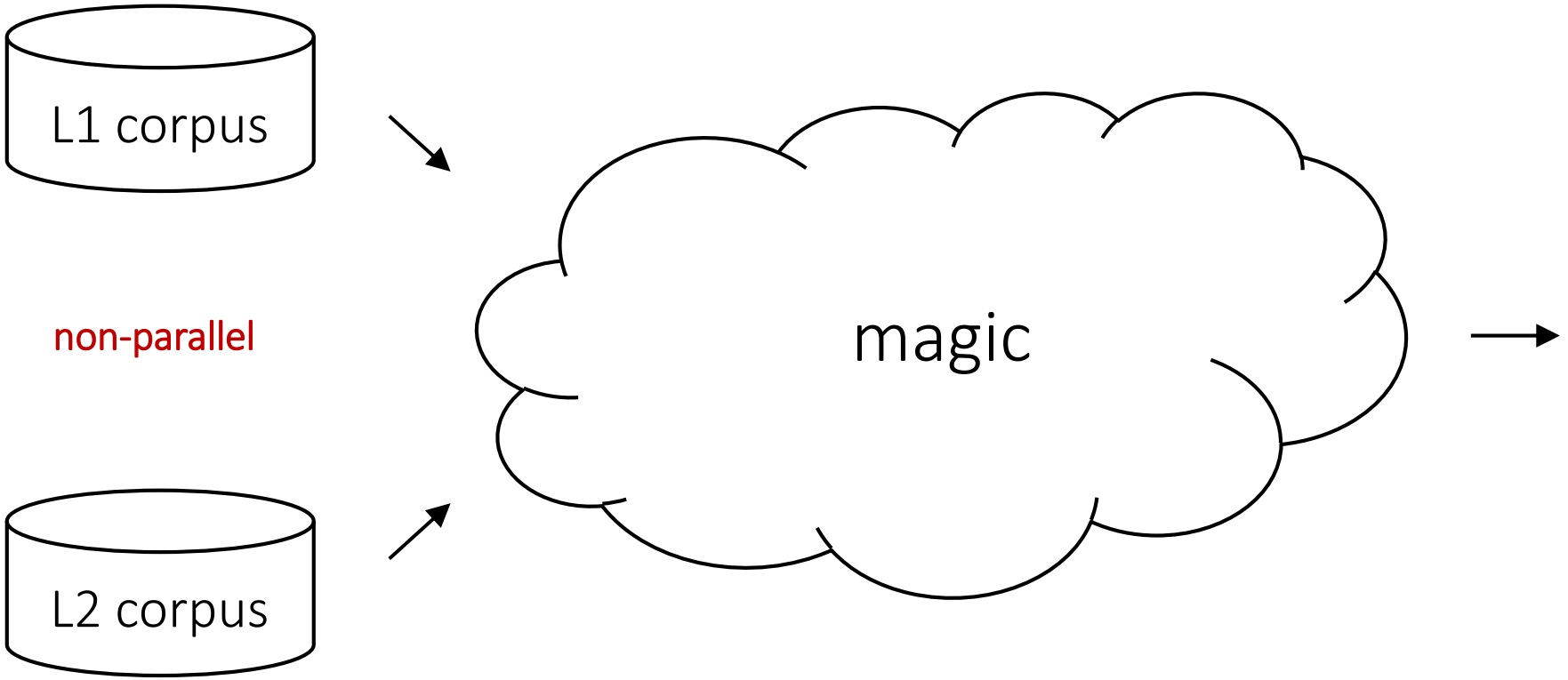
non-parallel



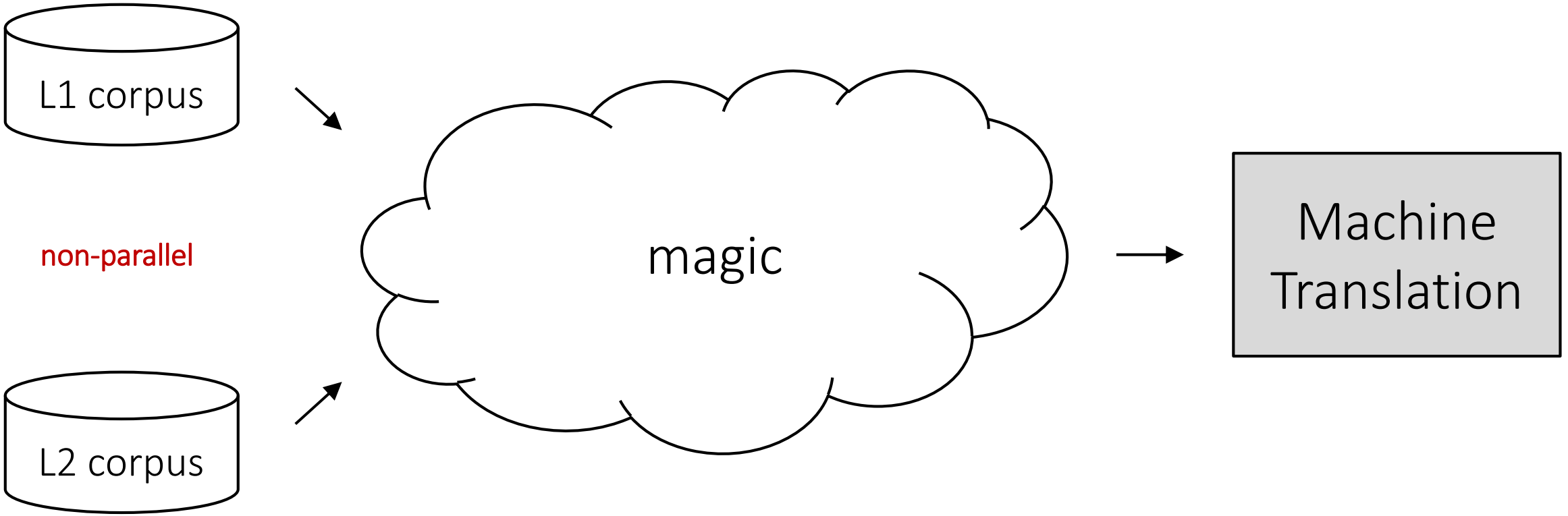
Introduction



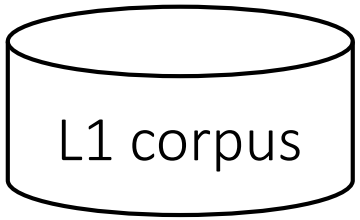
Introduction



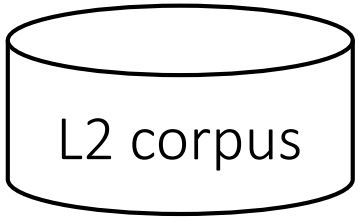
Introduction



Introduction

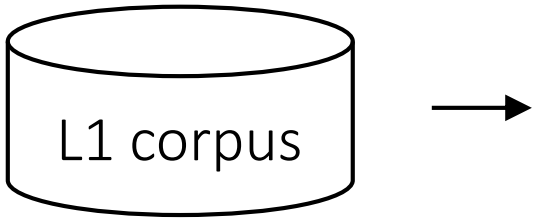


non-parallel

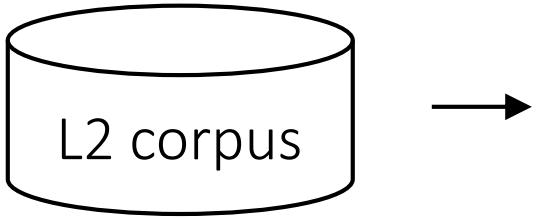


Machine Translation

Introduction



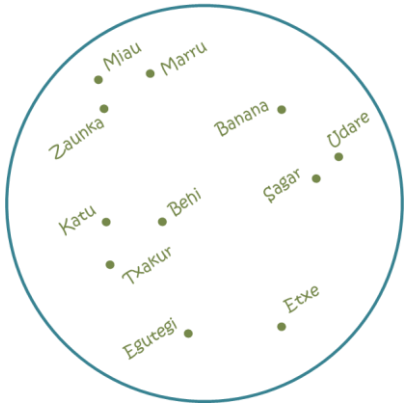
non-parallel



Machine Translation

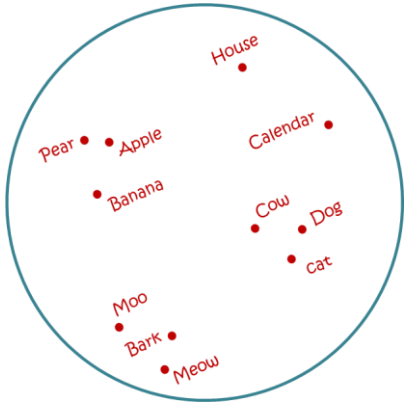
Introduction

L1 corpus



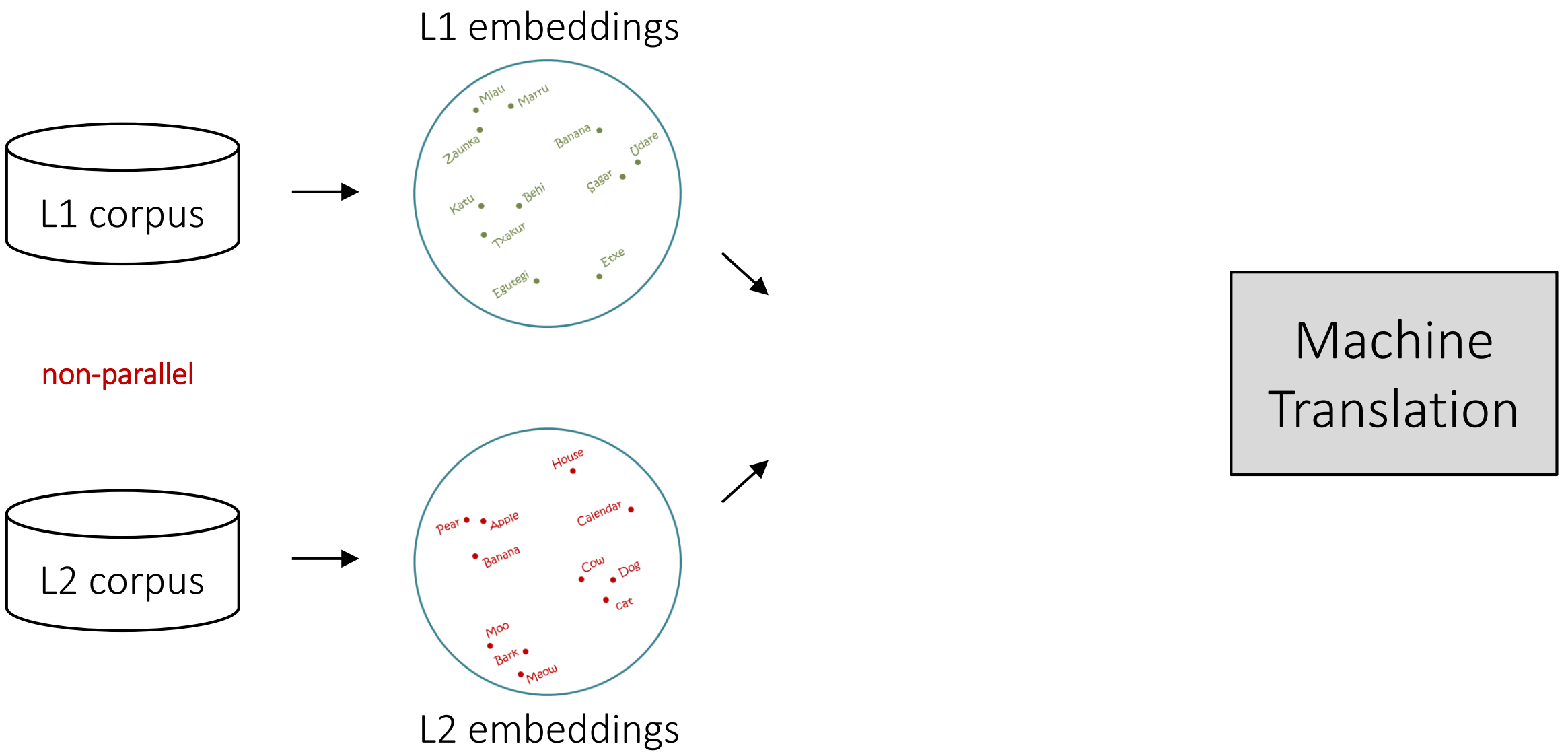
non-parallel

L2 corpus

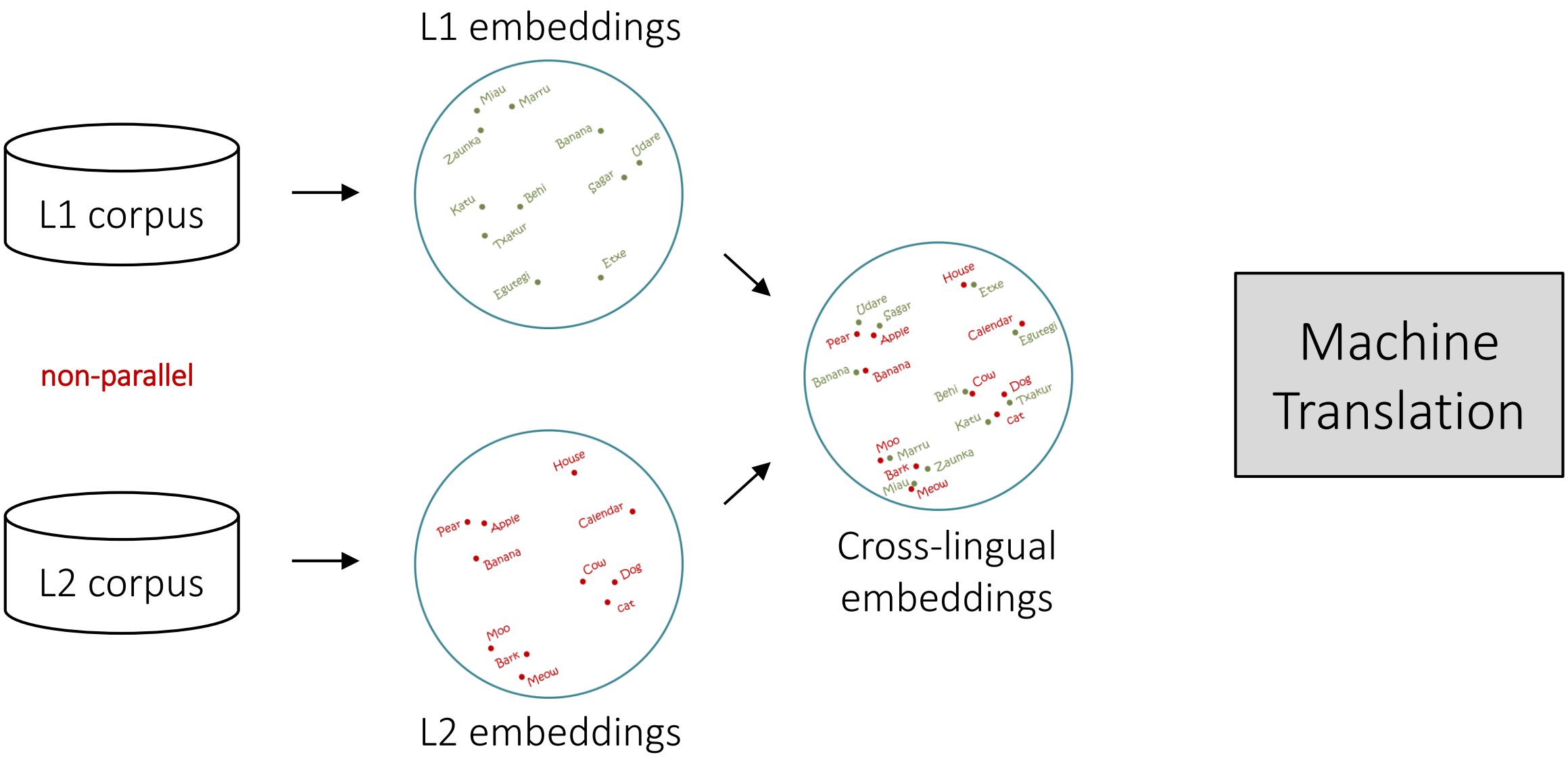


Machine Translation

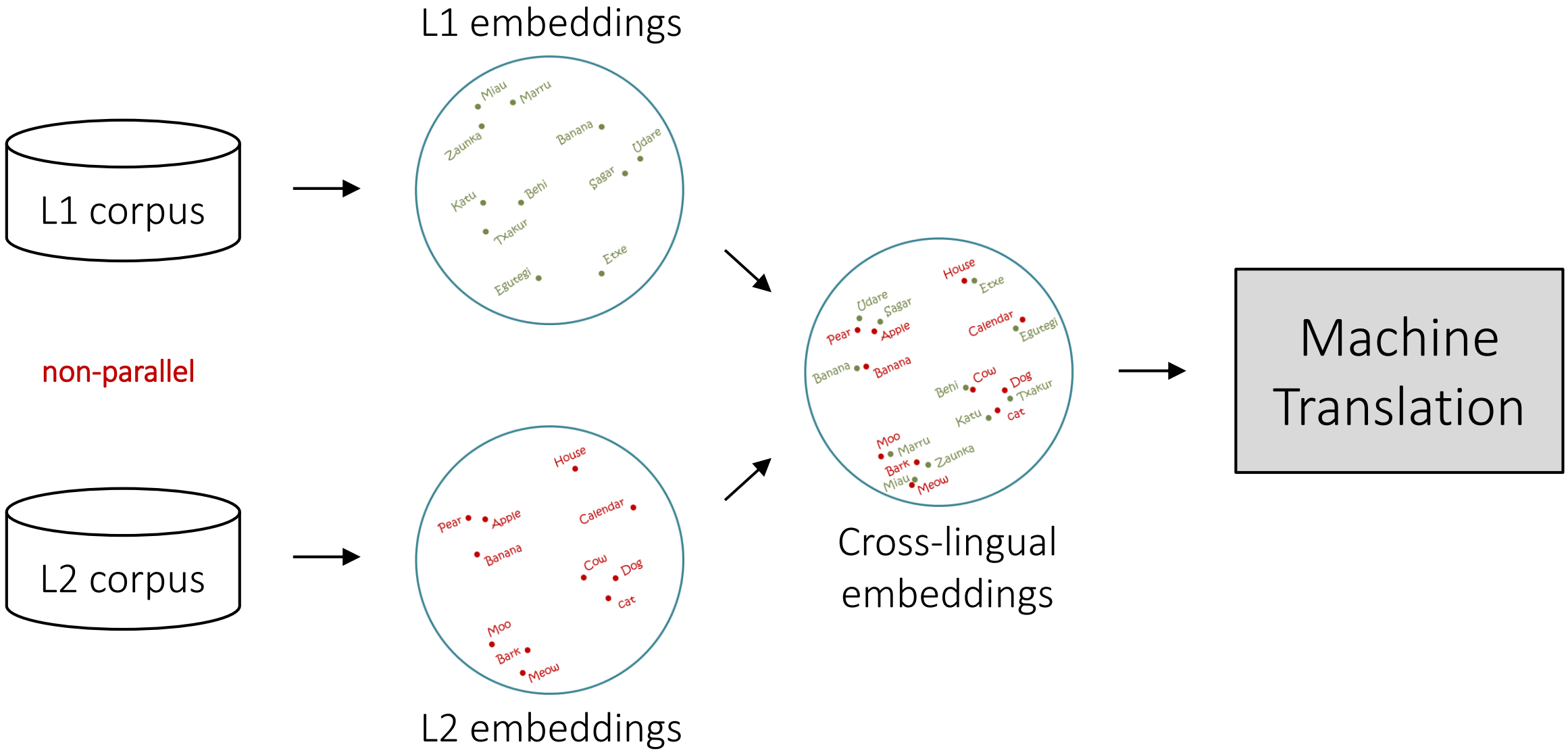
Introduction



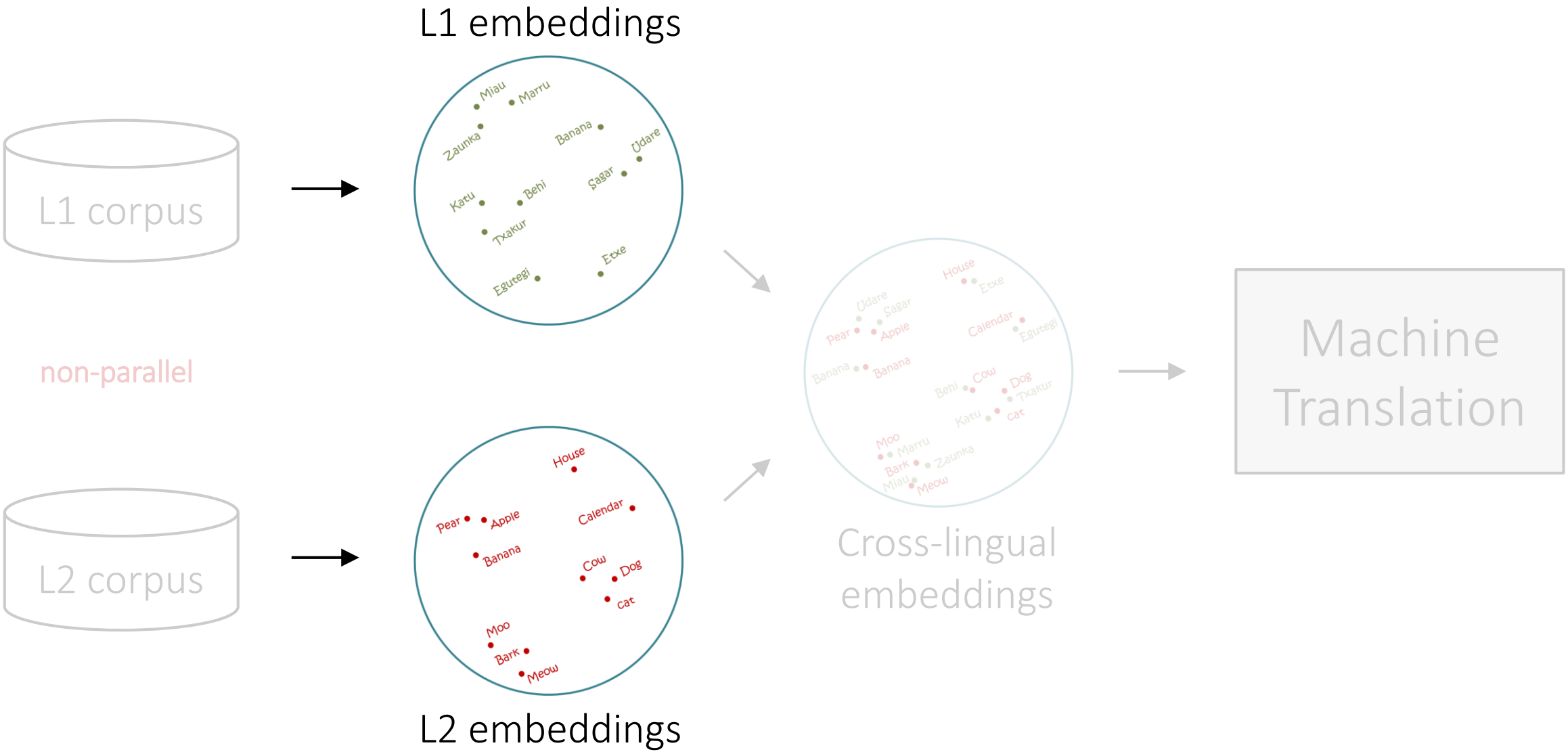
Introduction



Introduction

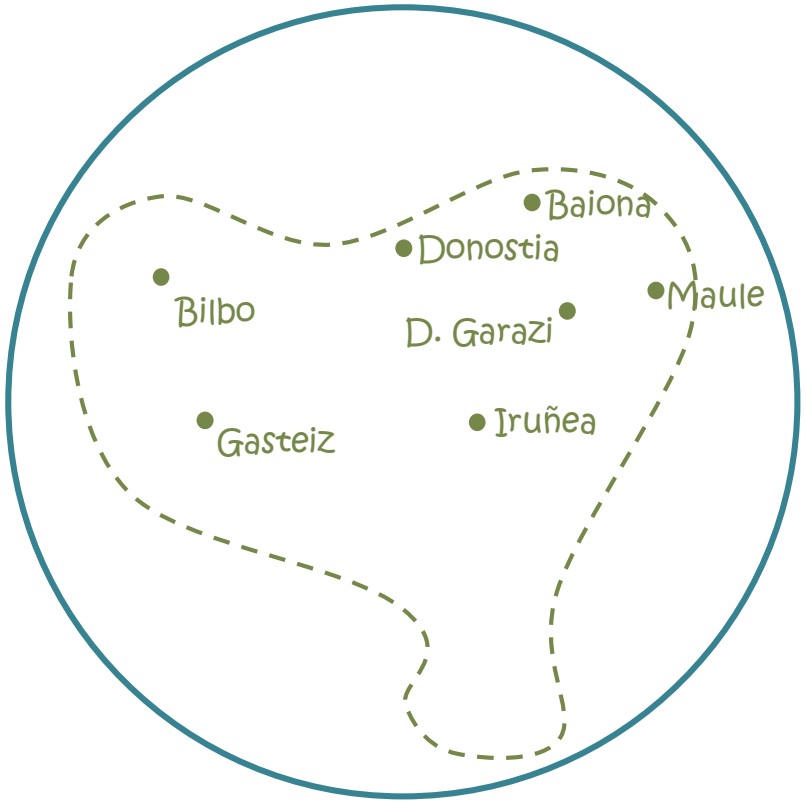


Introduction

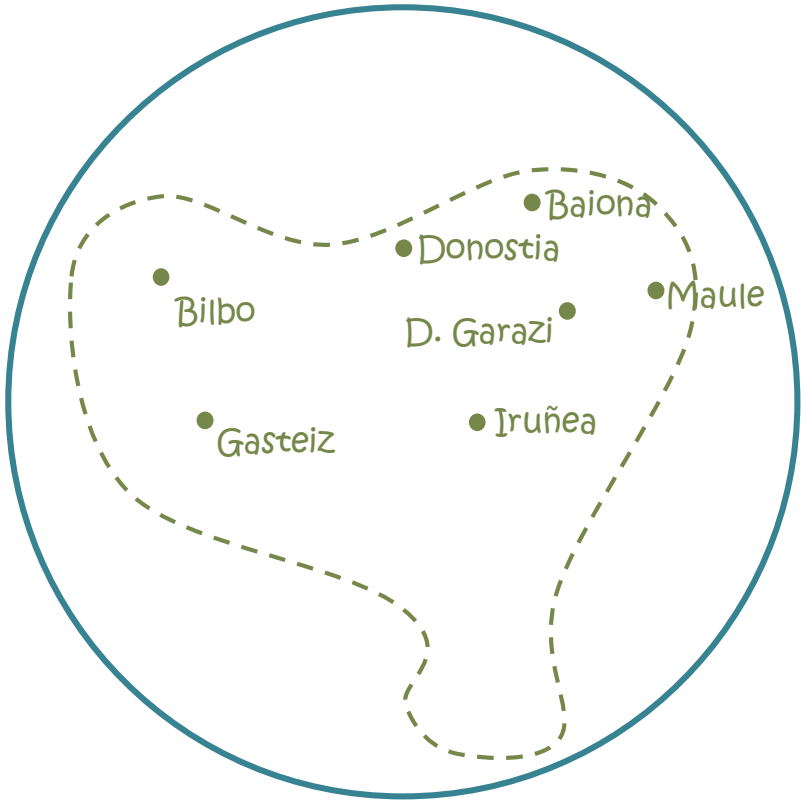


Embeddings

Embeddings

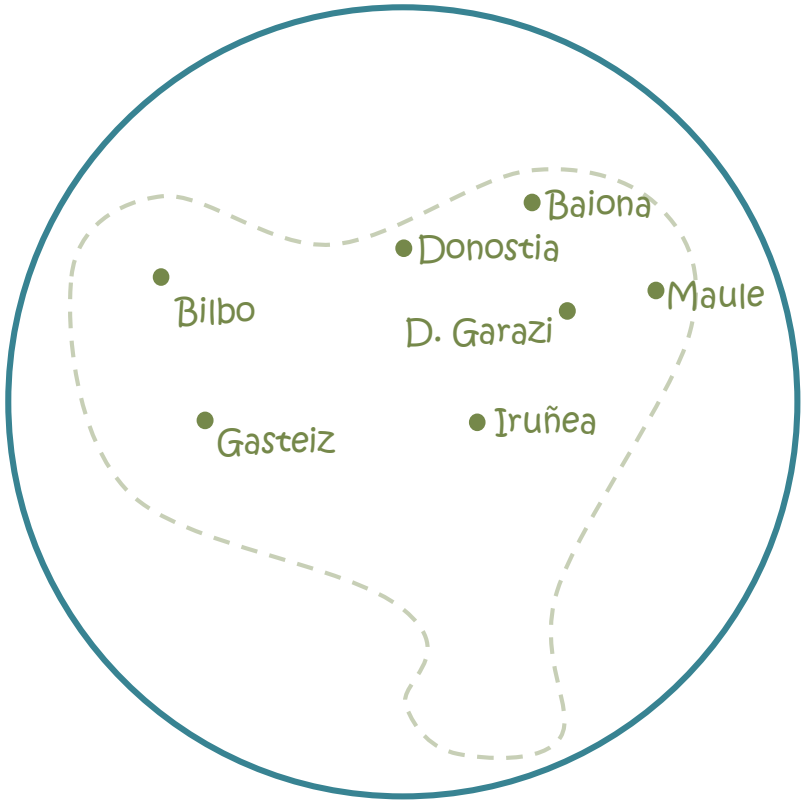


Embeddings



Geographical space

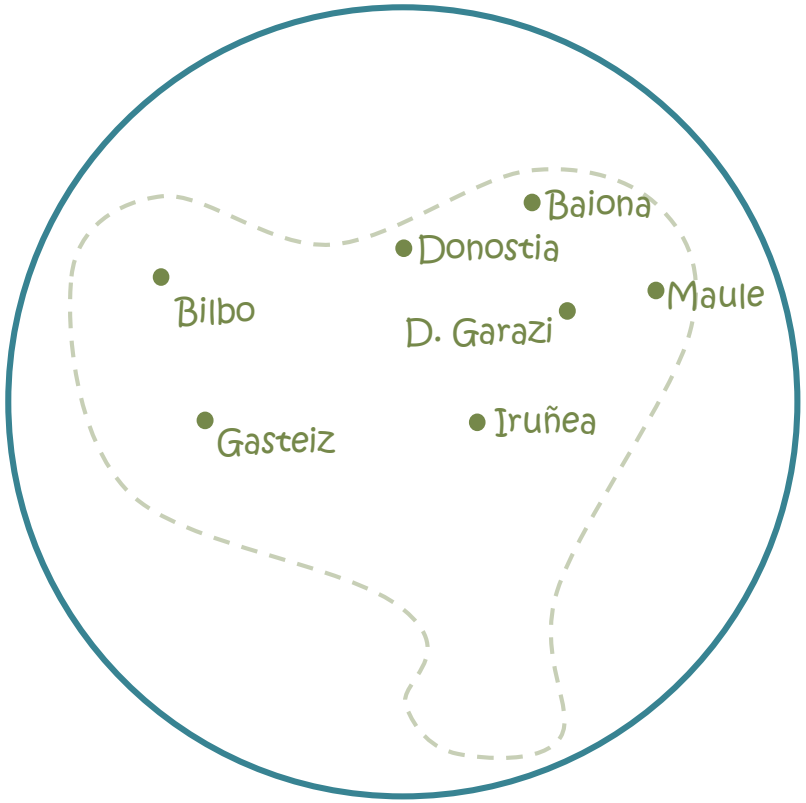
Embeddings



Geographical space

- Cities

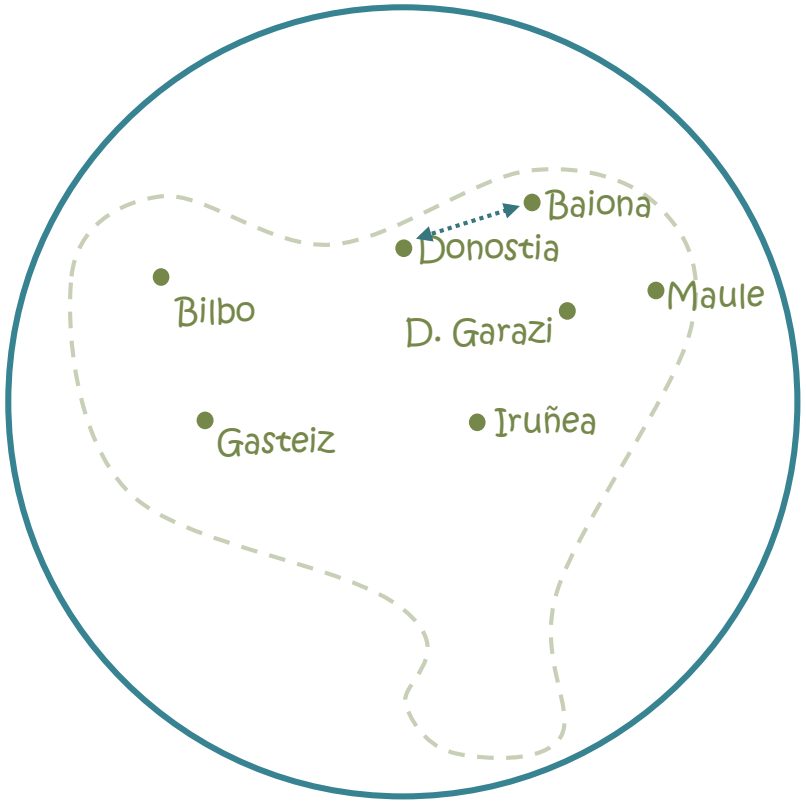
Embeddings



Geographical space

- Cities
- Meaningful distances

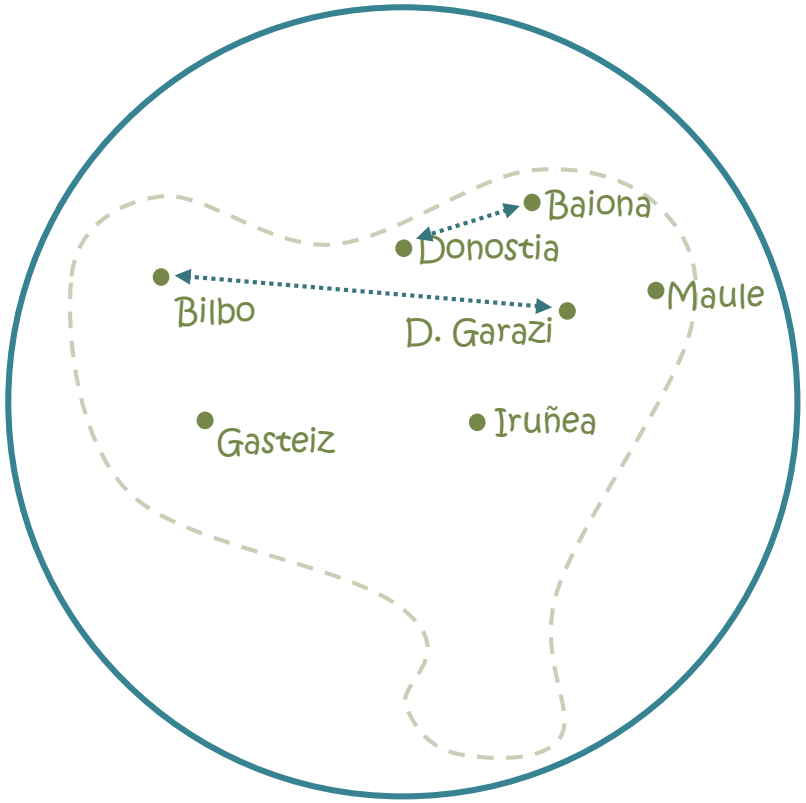
Embeddings



Geographical space

- Cities
- Meaningful distances

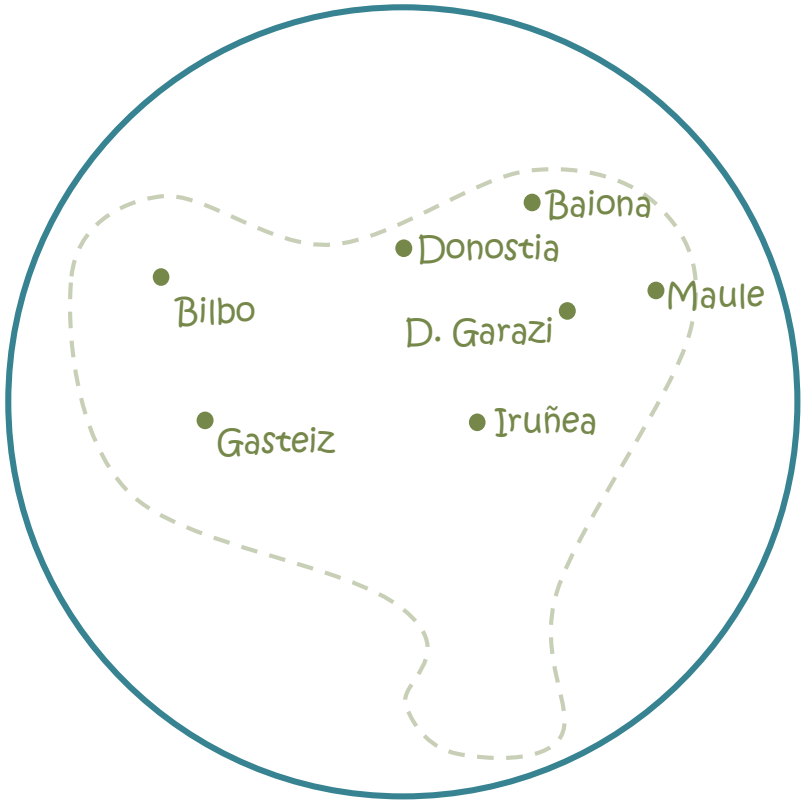
Embeddings



Geographical space

- Cities
- Meaningful distances

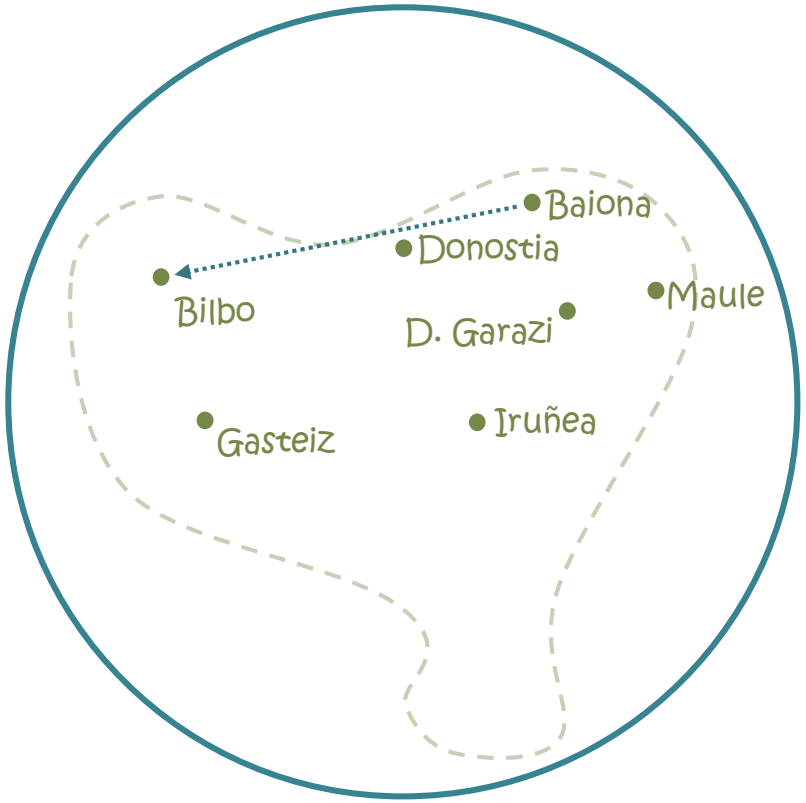
Embeddings



Geographical space

- Cities
- Meaningful distances
- Meaningful relations

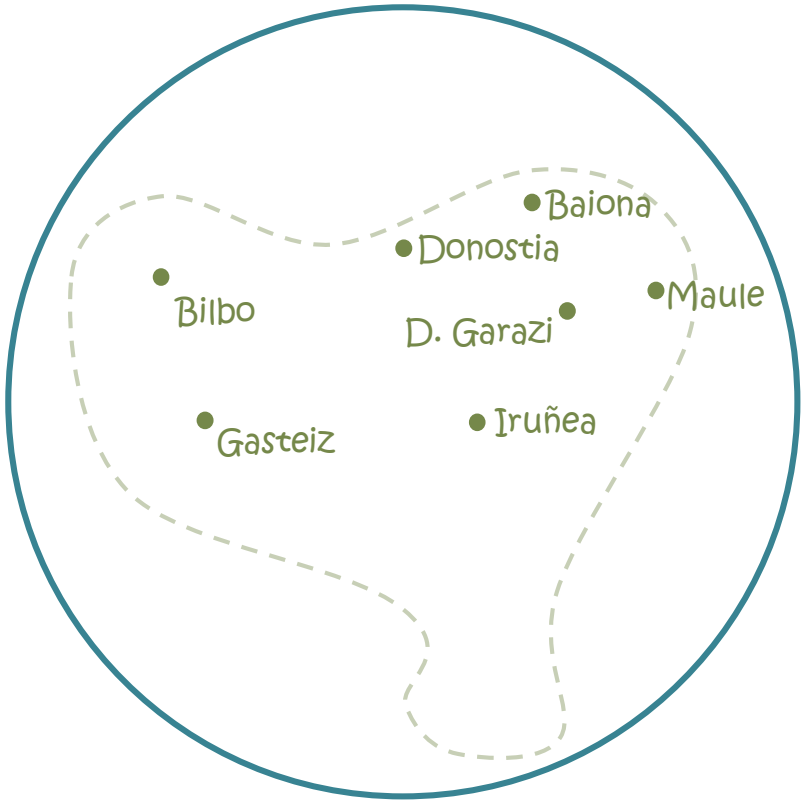
Embeddings



Geographical space

- Cities
- Meaningful distances
- Meaningful relations

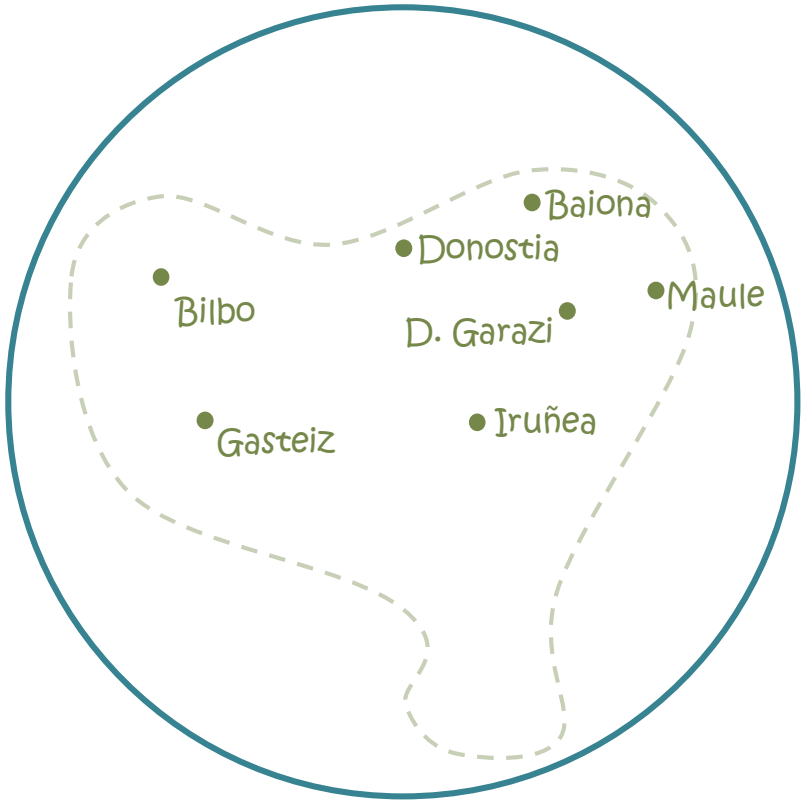
Embeddings



Geographical space

- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions

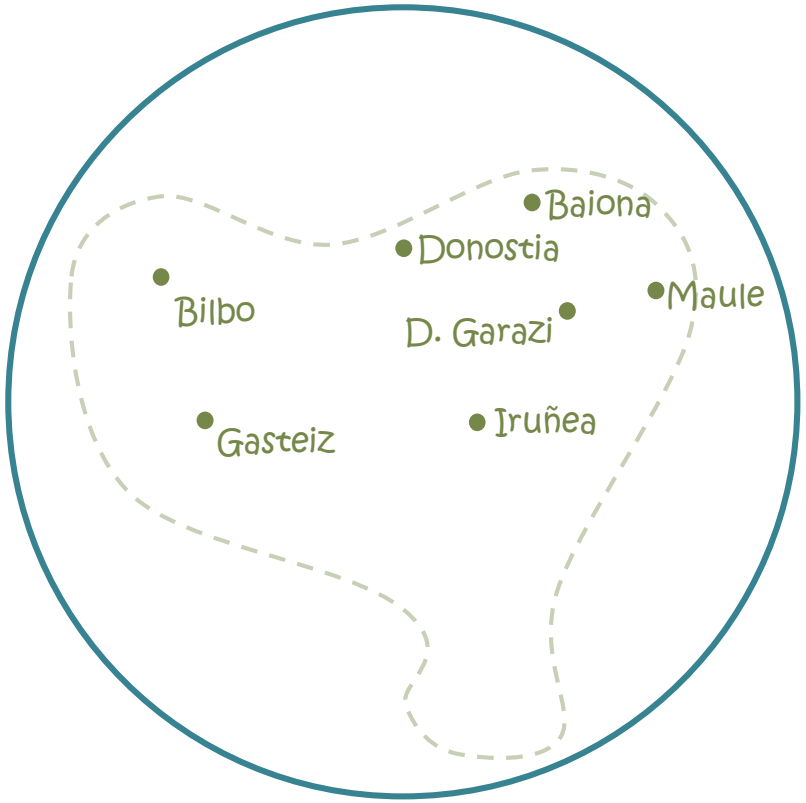
Embeddings



Geographical space

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

Embeddings

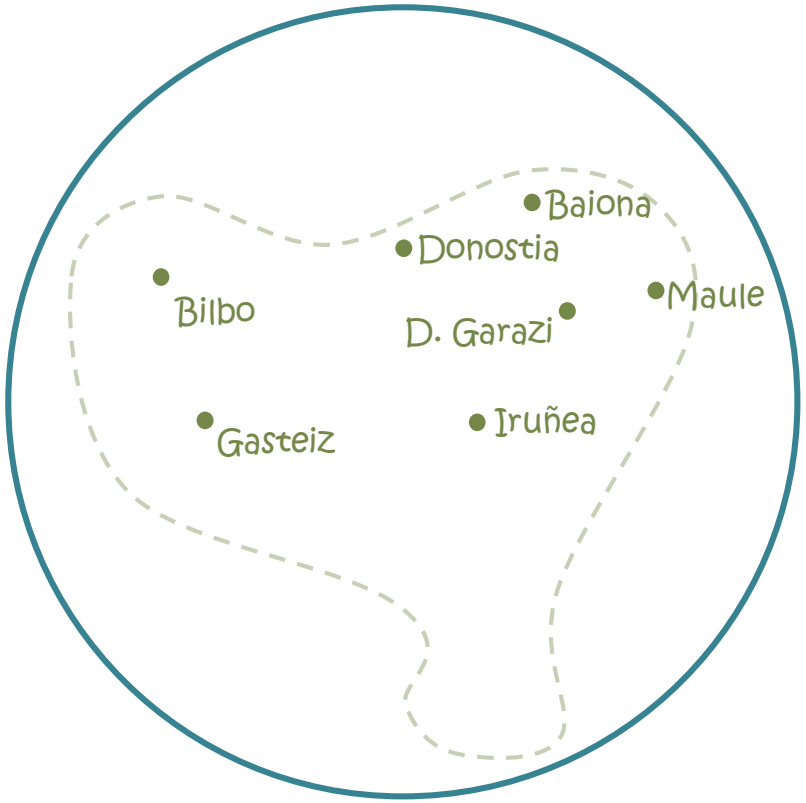


Geographical space

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

Semantic space

Embeddings



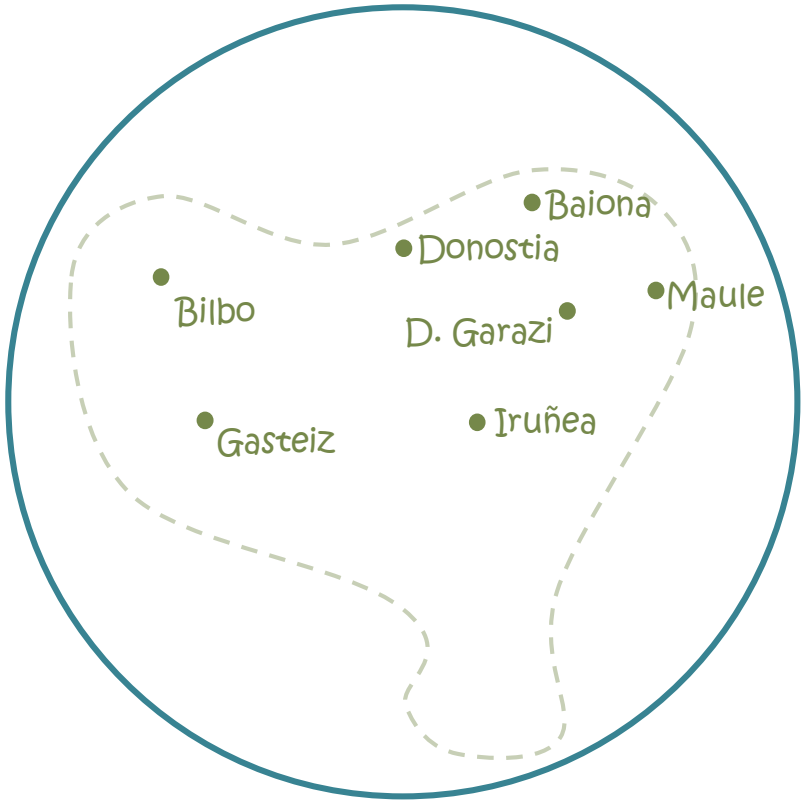
Geographical space

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

Semantic space

- Words

Embeddings

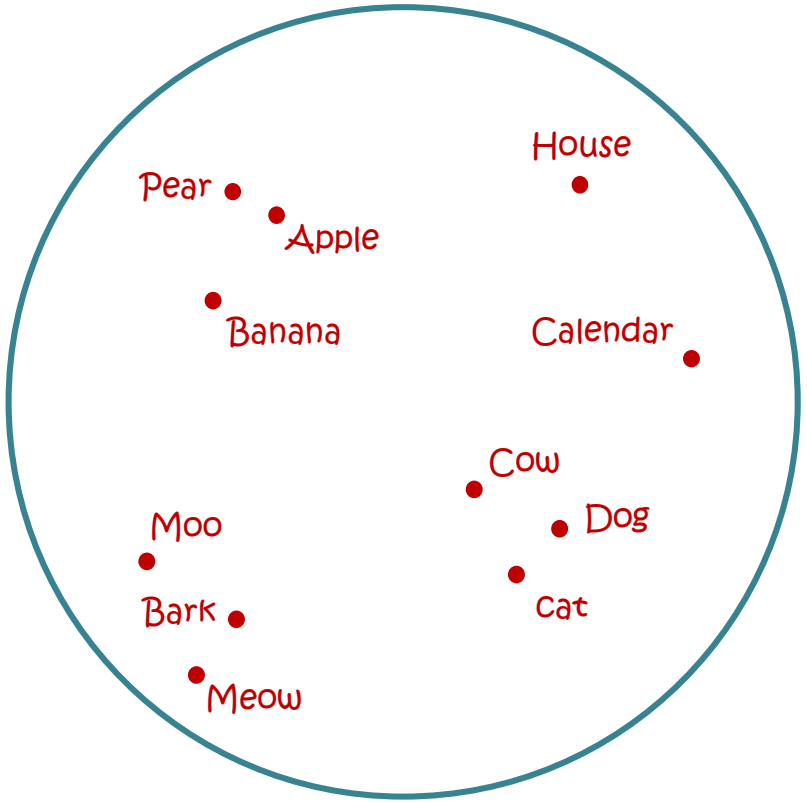


Geographical space

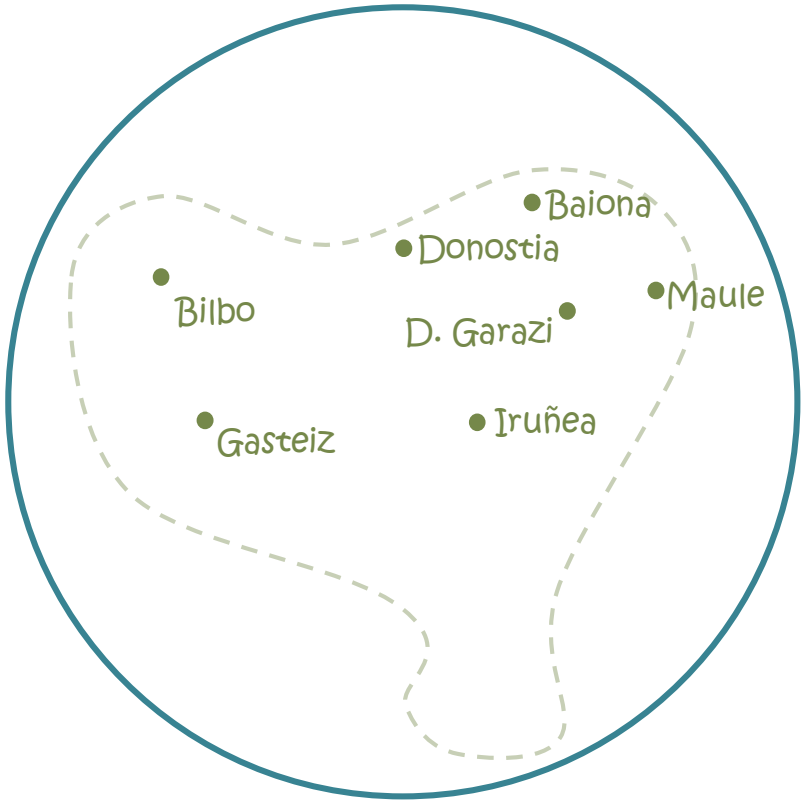
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words



Embeddings

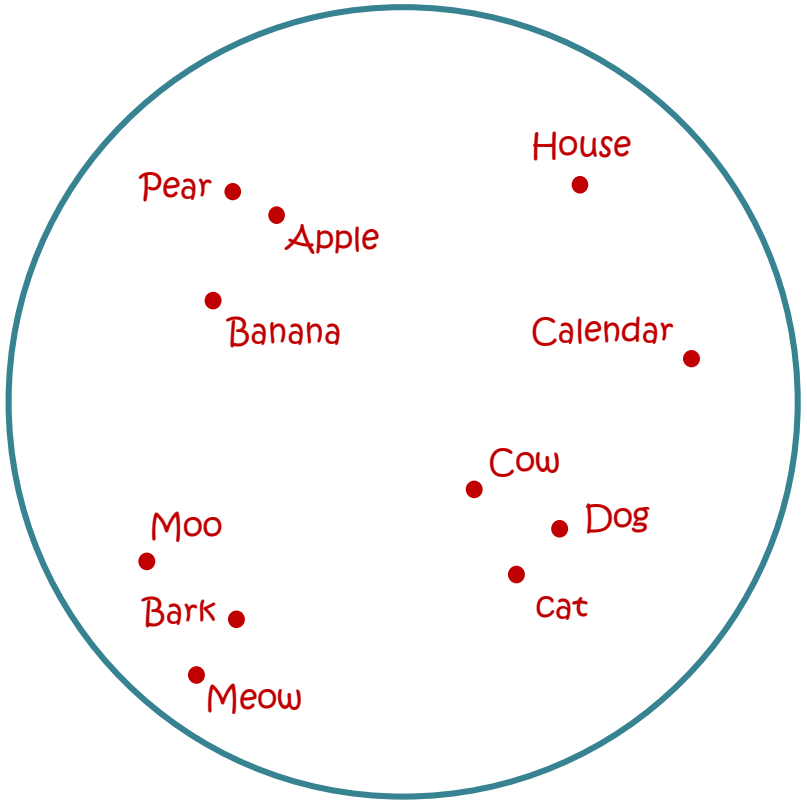


Geographical space

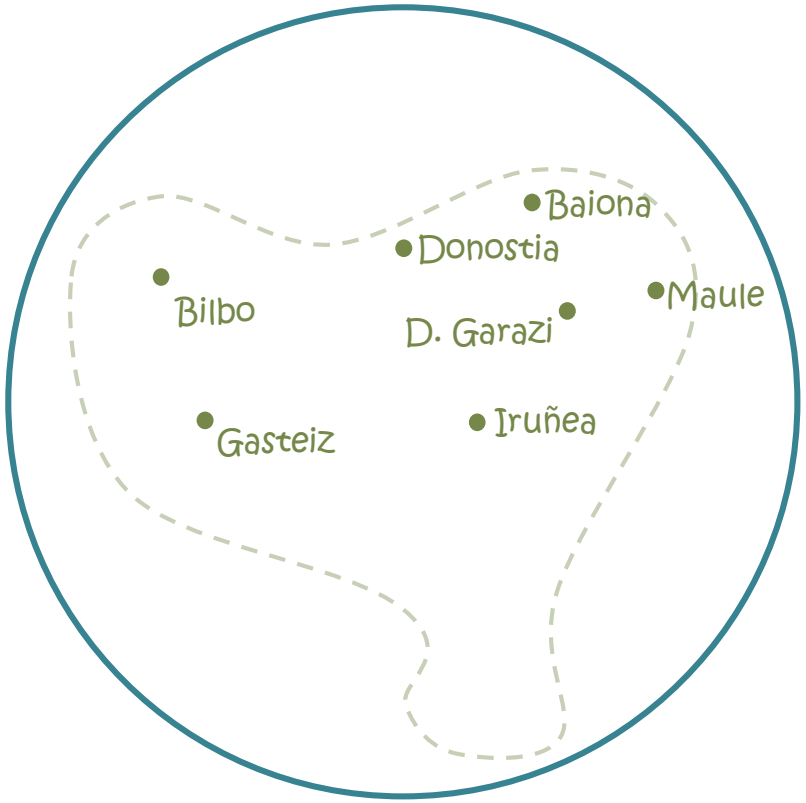
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances



Embeddings

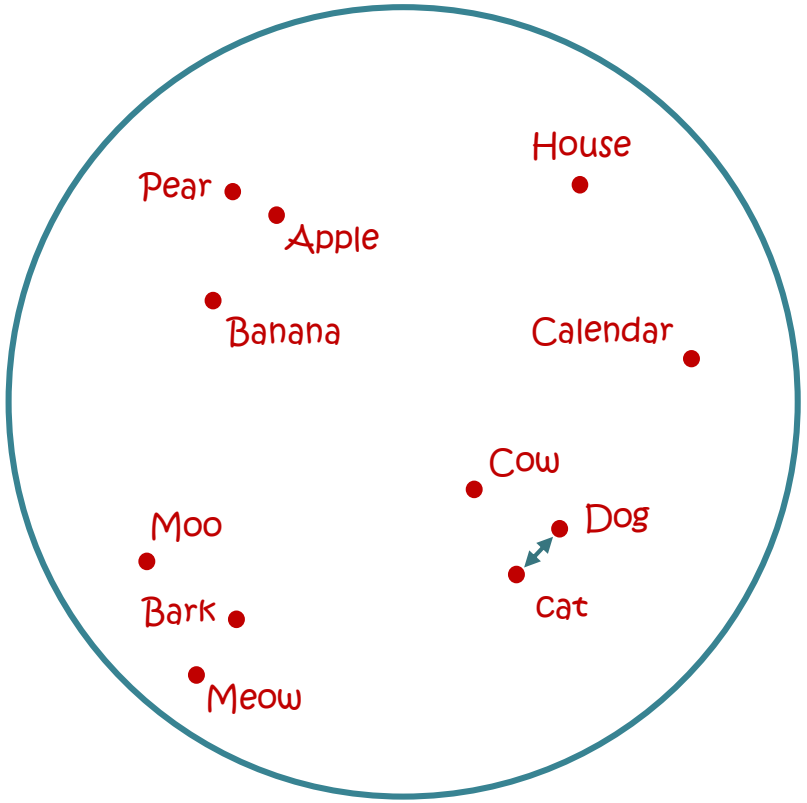


Geographical space

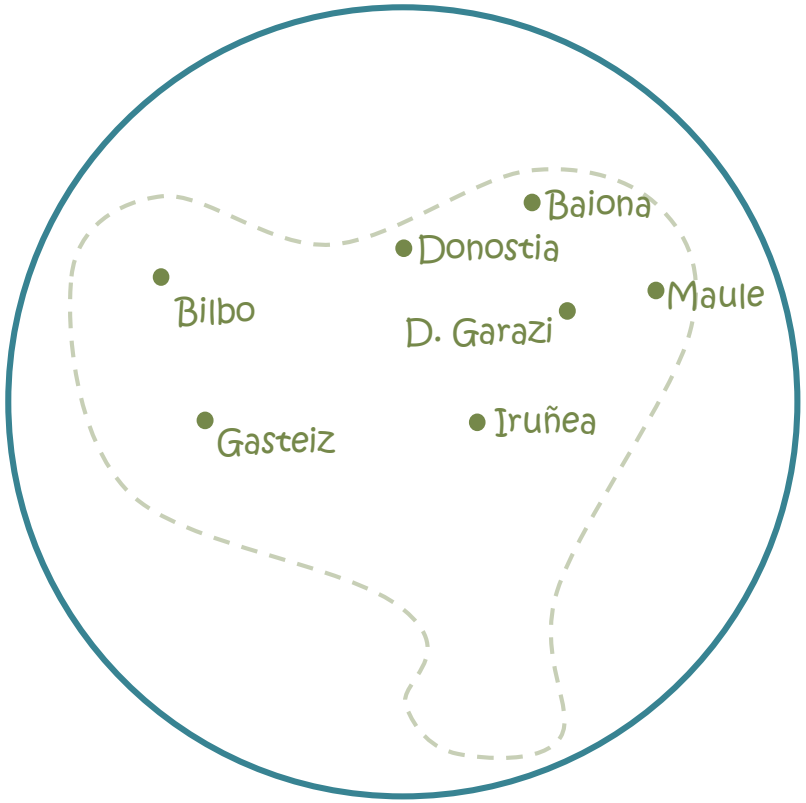
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances



Embeddings

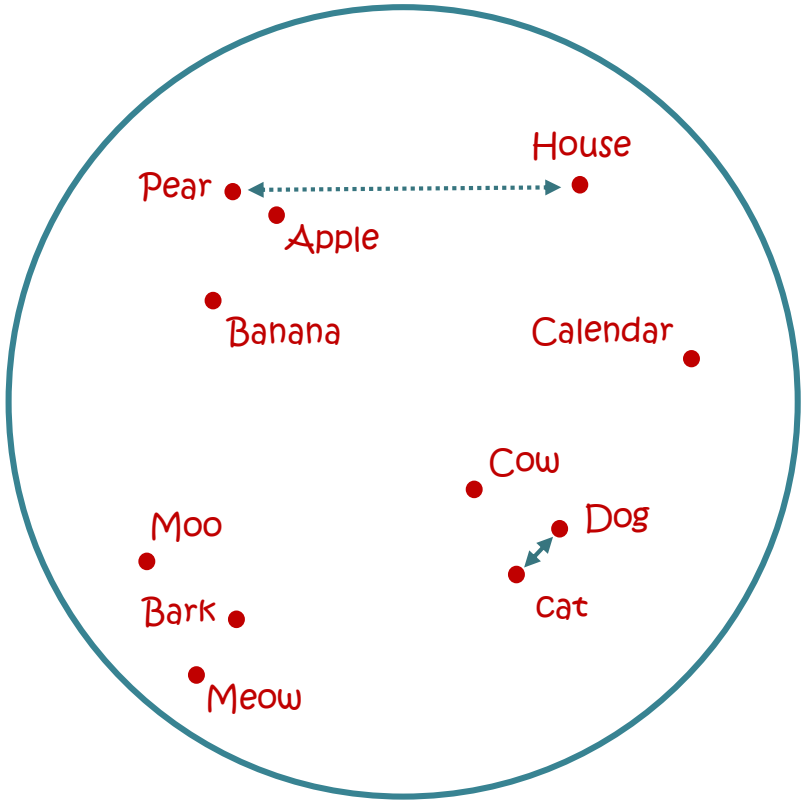


Geographical space

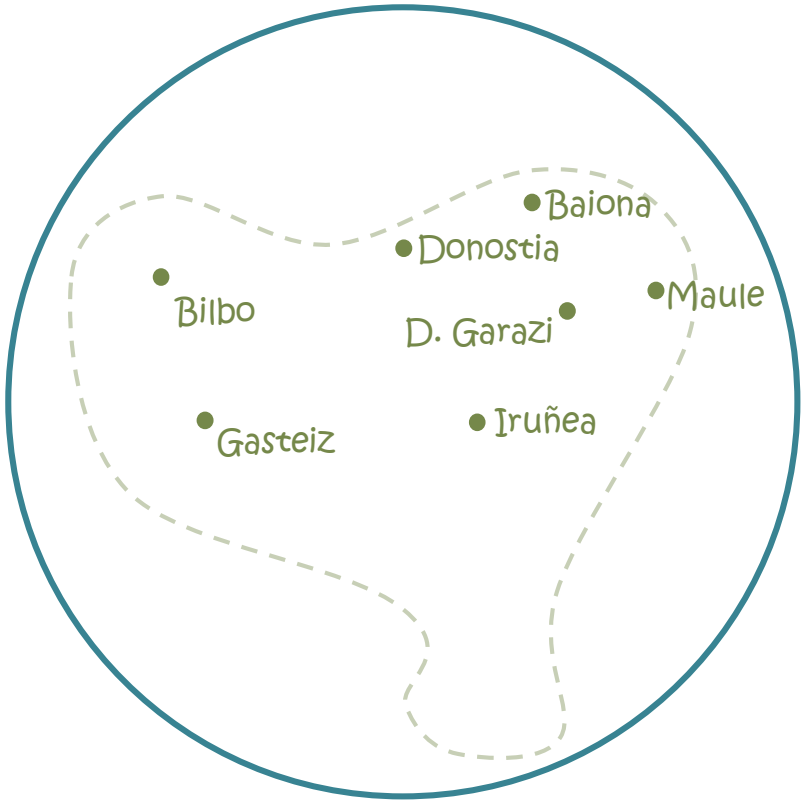
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances



Embeddings

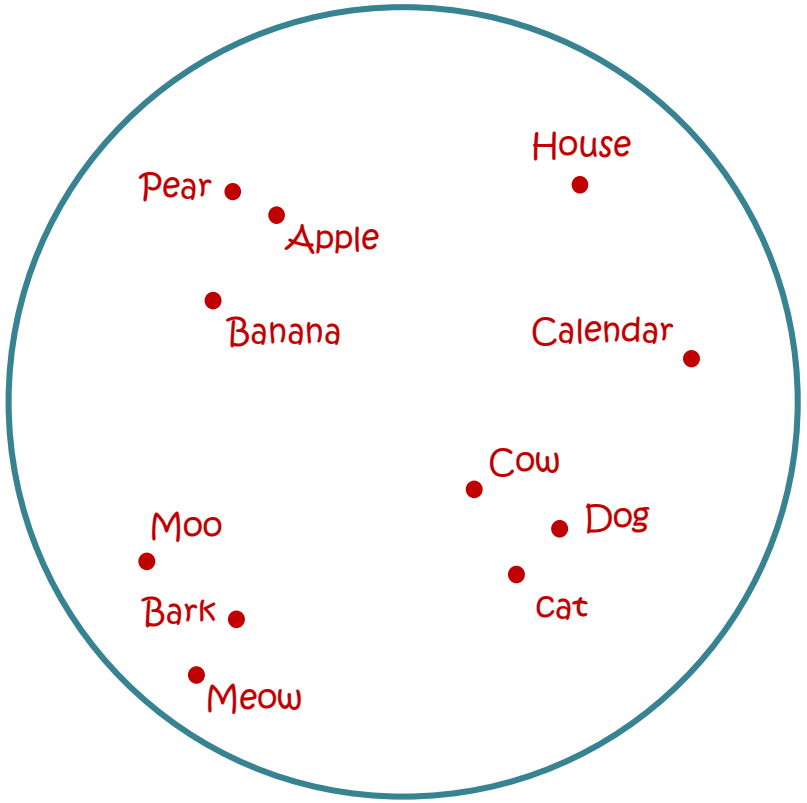


Geographical space

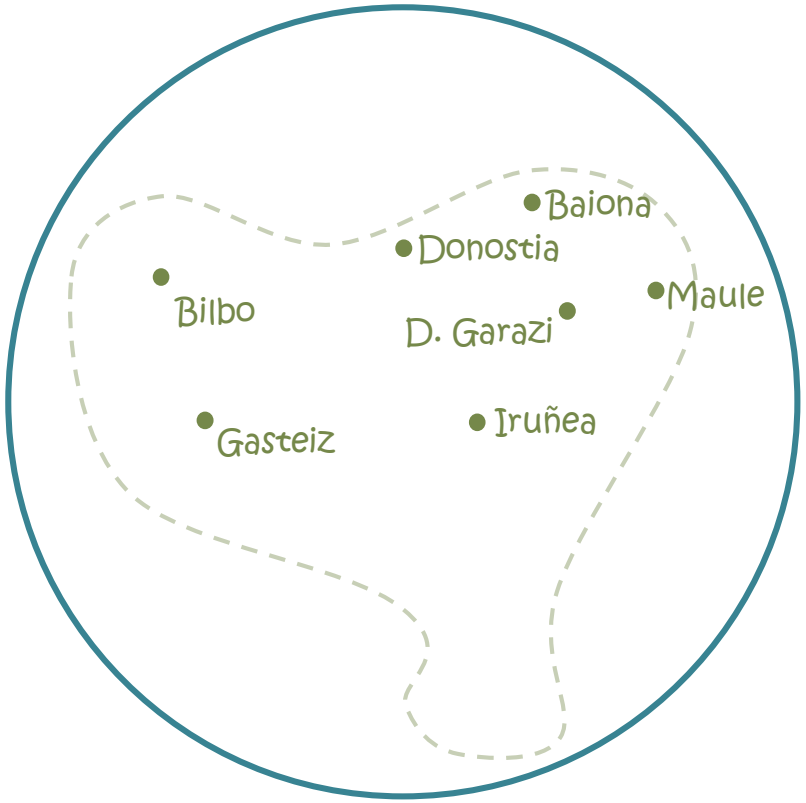
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Embeddings

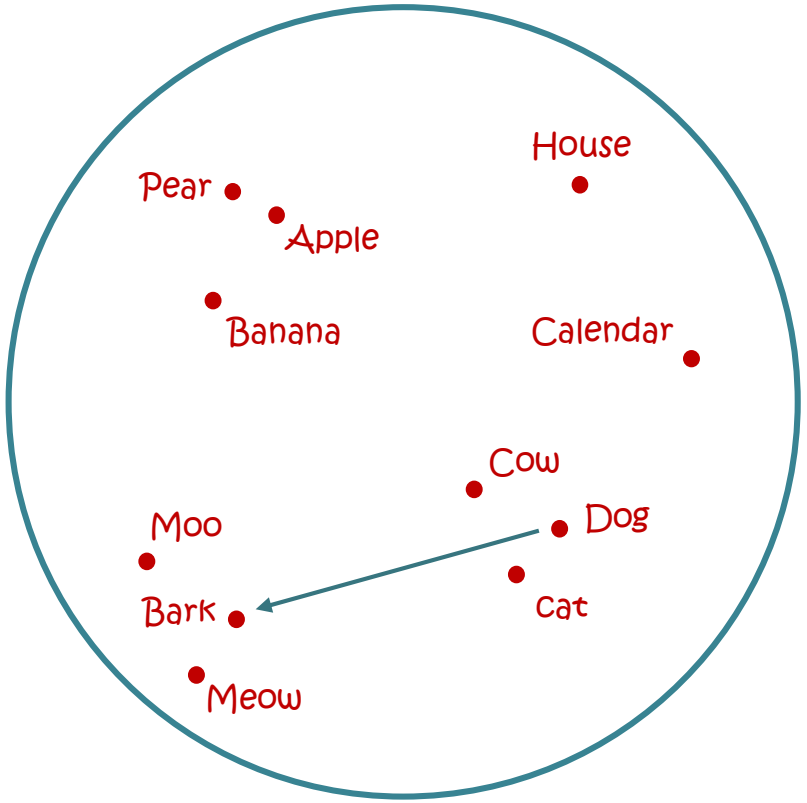


Geographical space

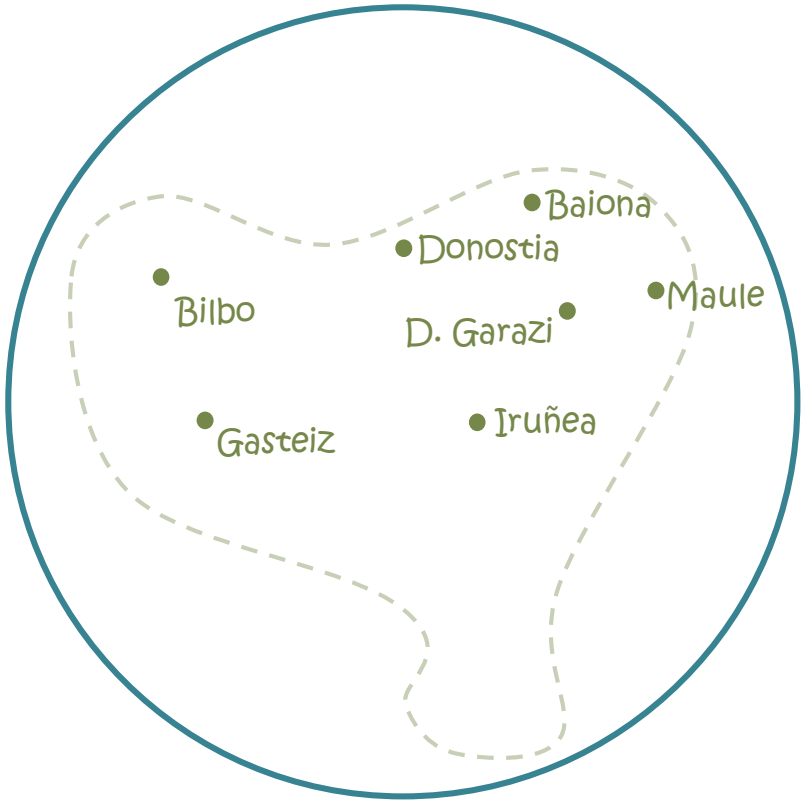
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Embeddings

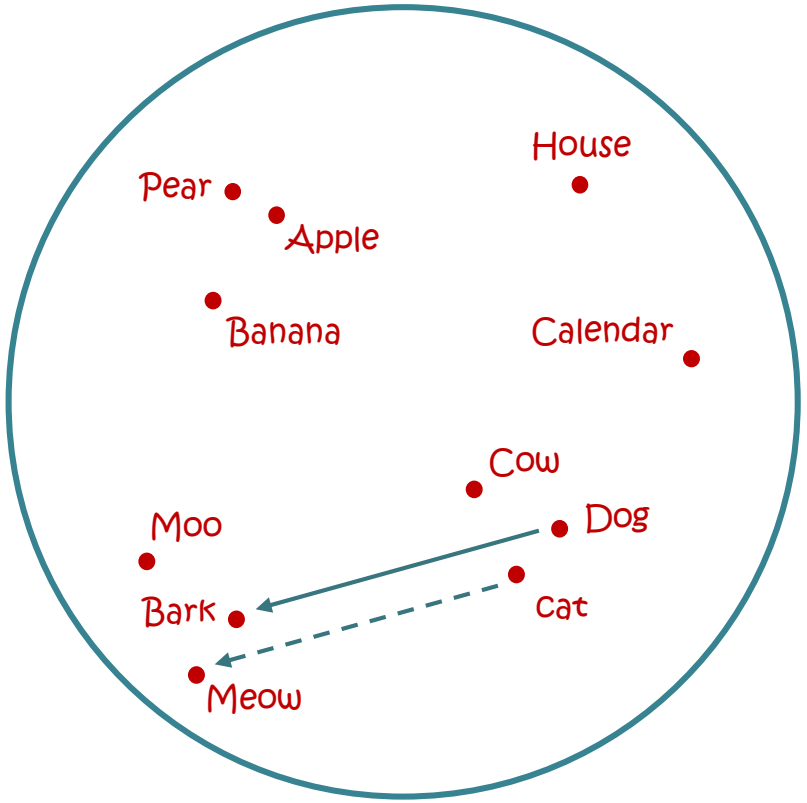


Geographical space

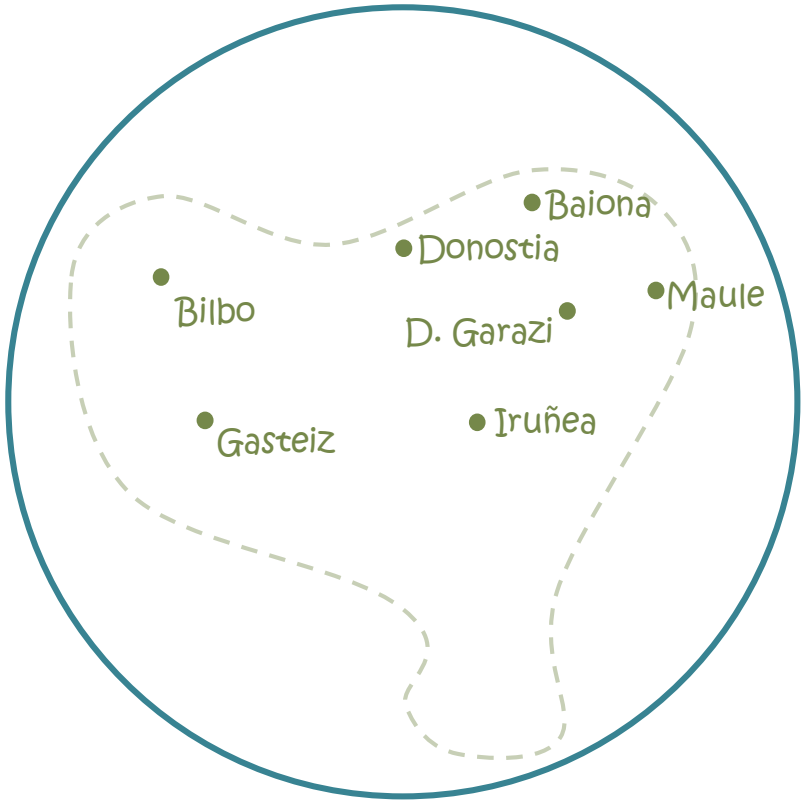
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Embeddings

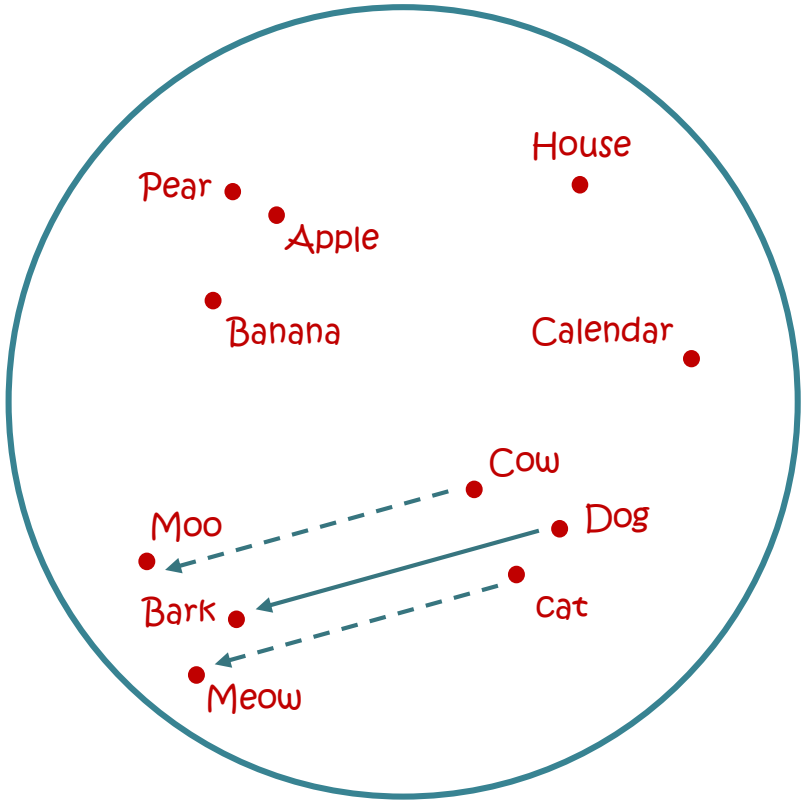


Geographical space

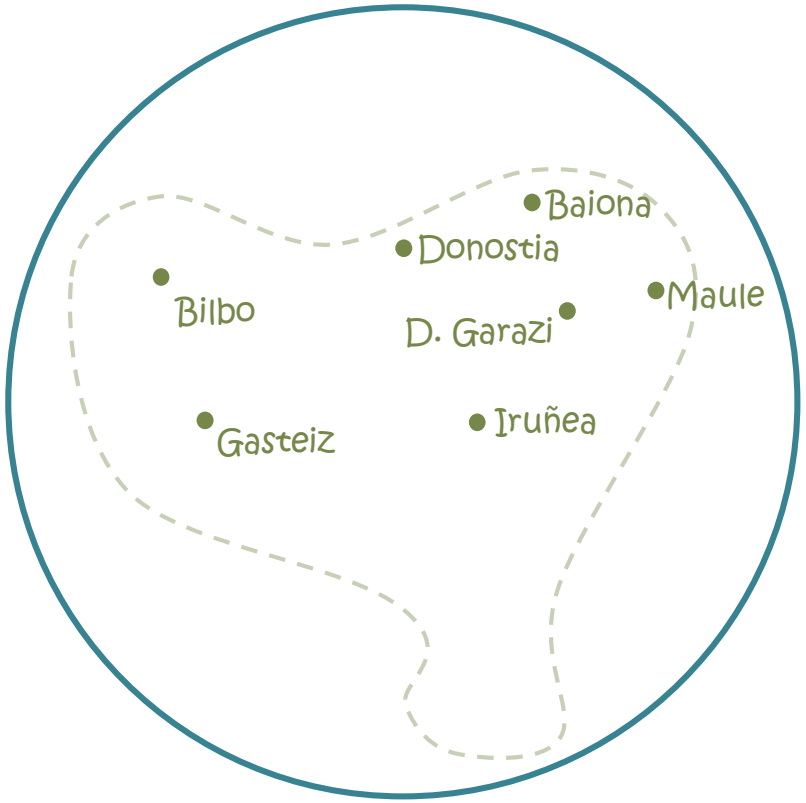
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations



Embeddings

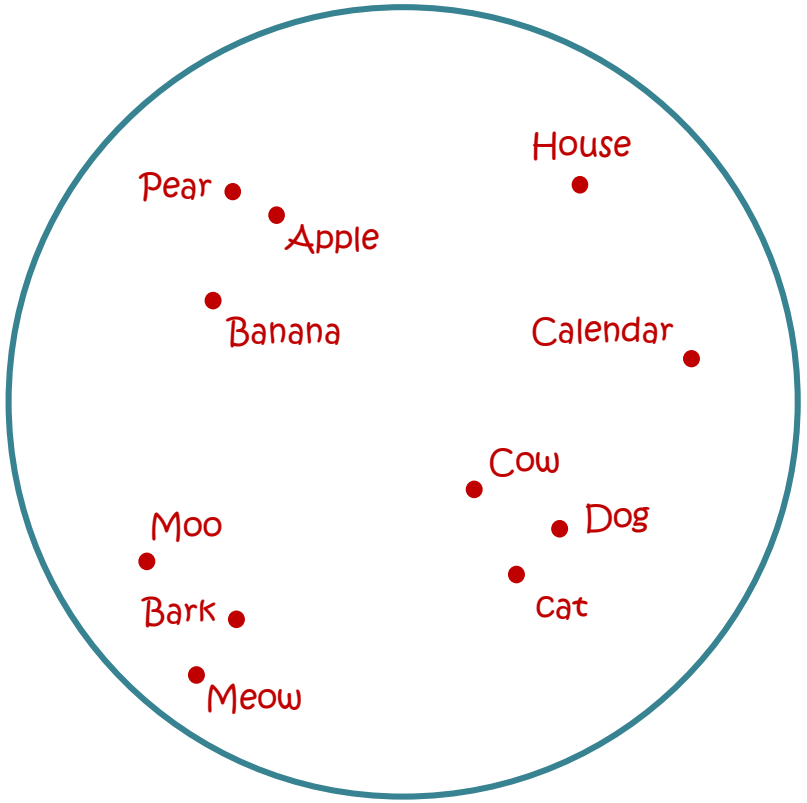


Geographical space

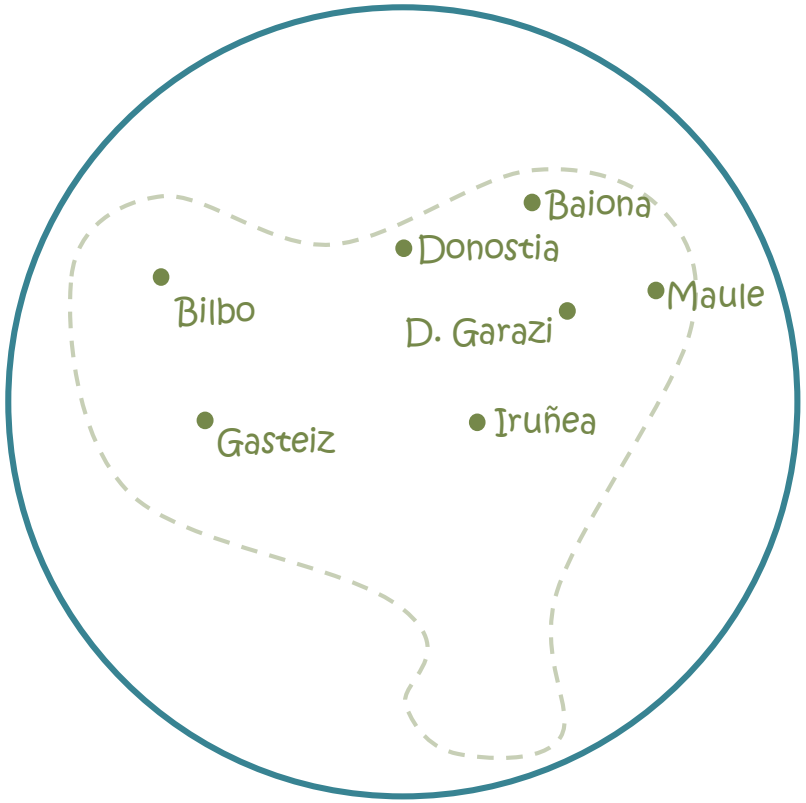
- Cities
- Meaningful distances
- Meaningful relations
- 2 dimensions
- Cartographers from 3D world

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions



Embeddings

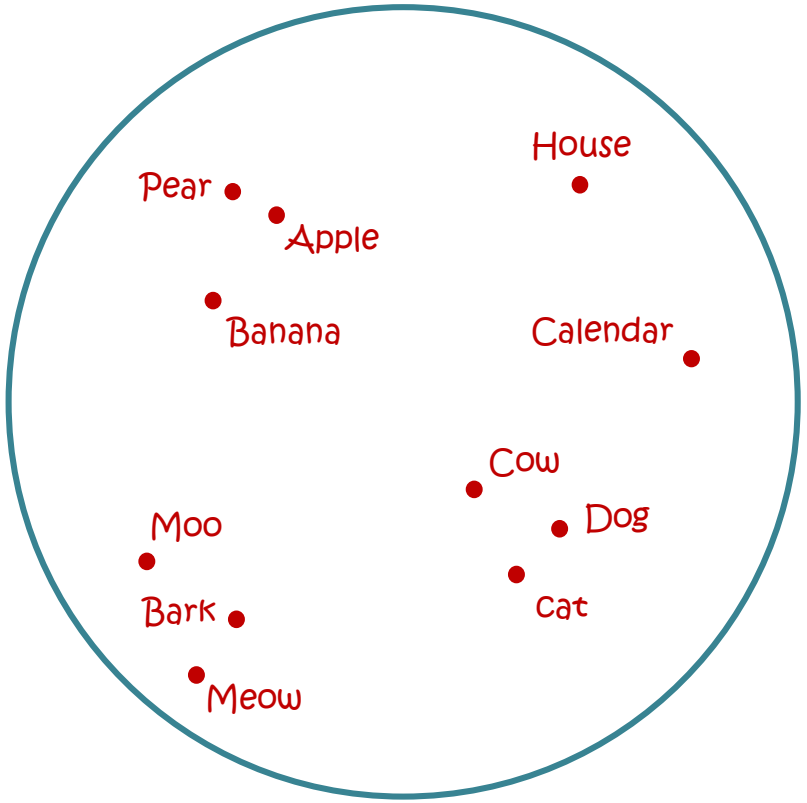


Geographical space

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

Semantic space

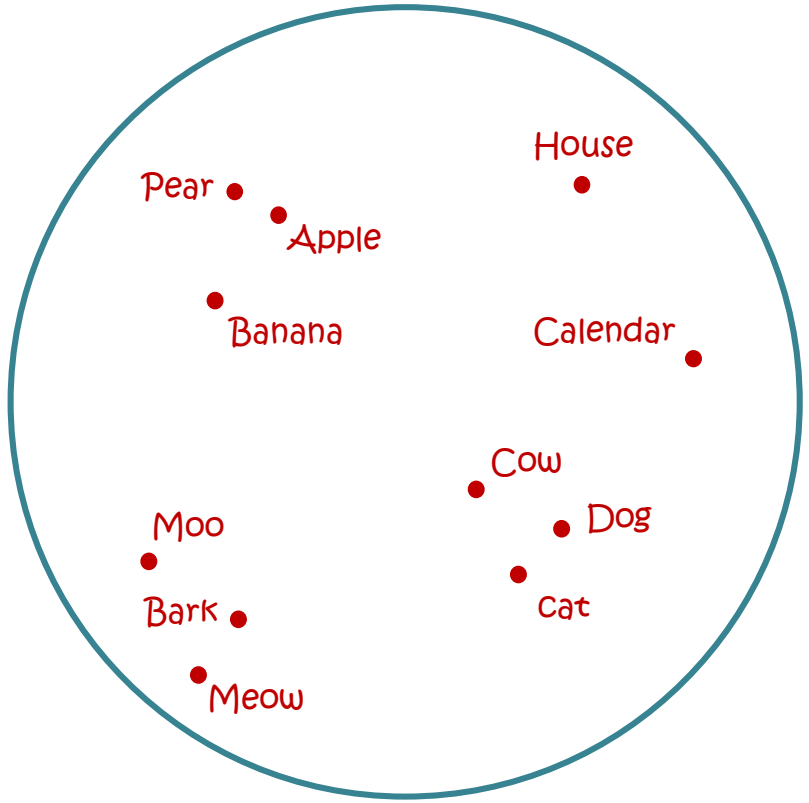
- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts

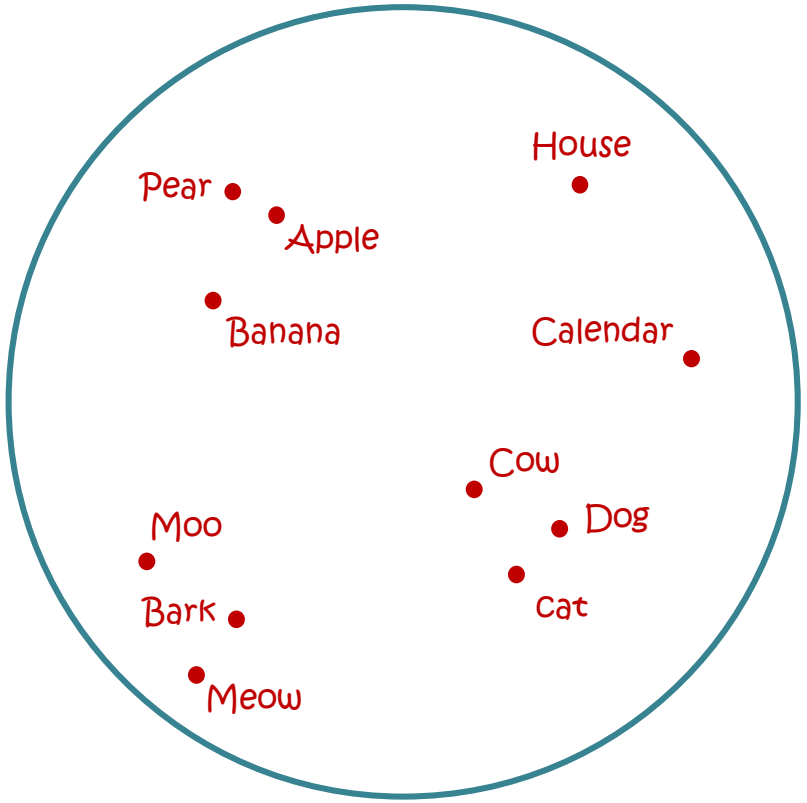


Embeddings

Traditional vector space models:

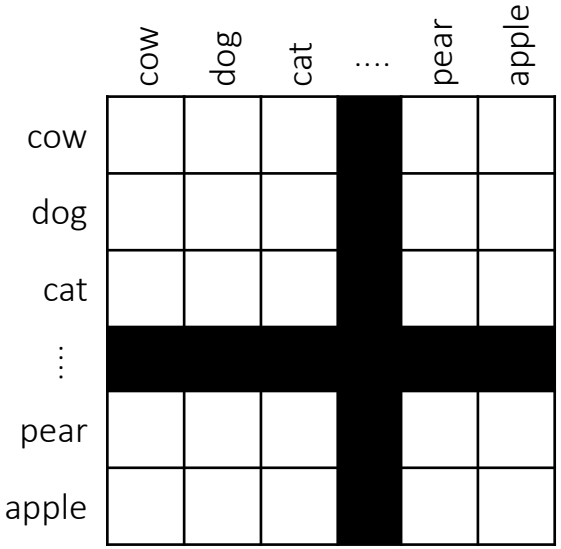
Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



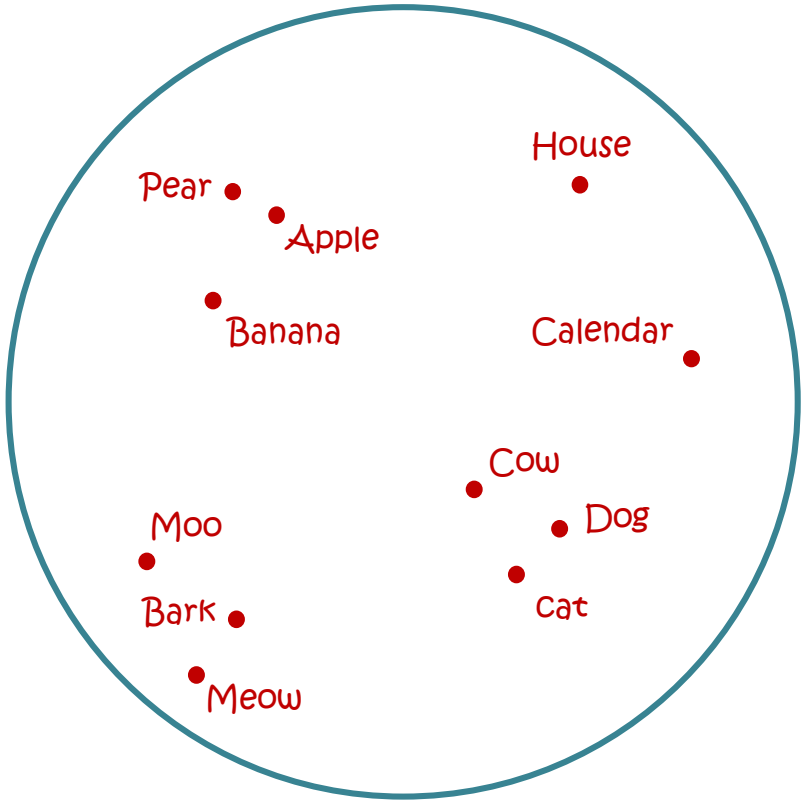
Embeddings

Traditional vector space models:



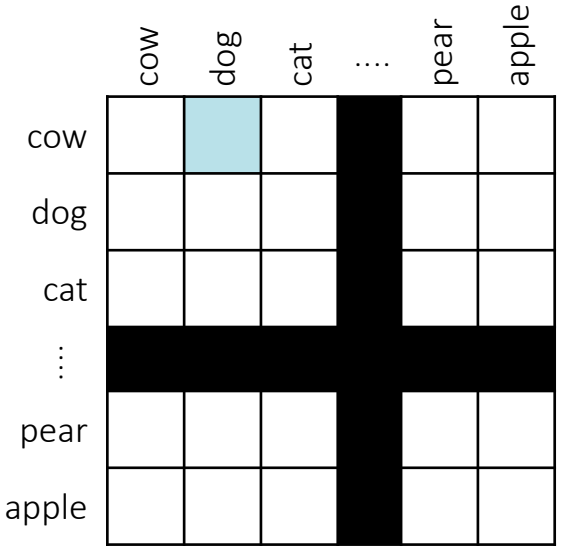
Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



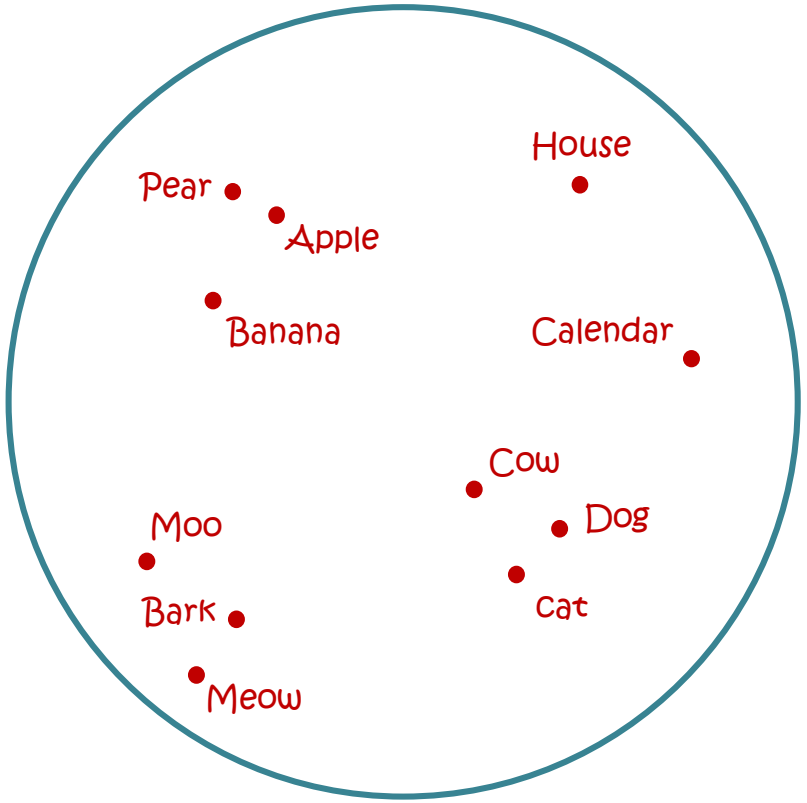
Embeddings

Traditional vector space models:



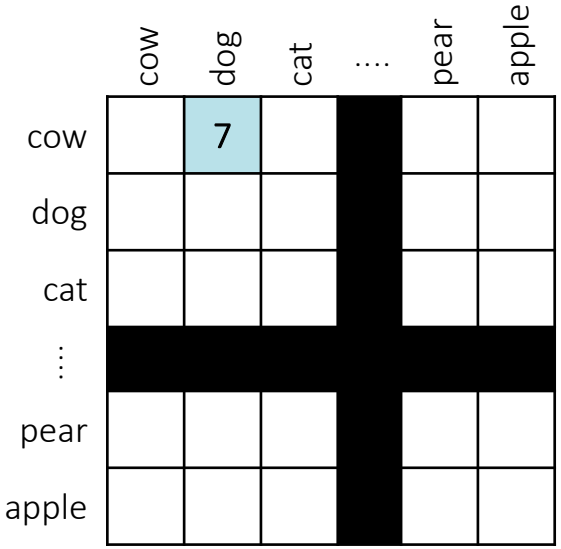
Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



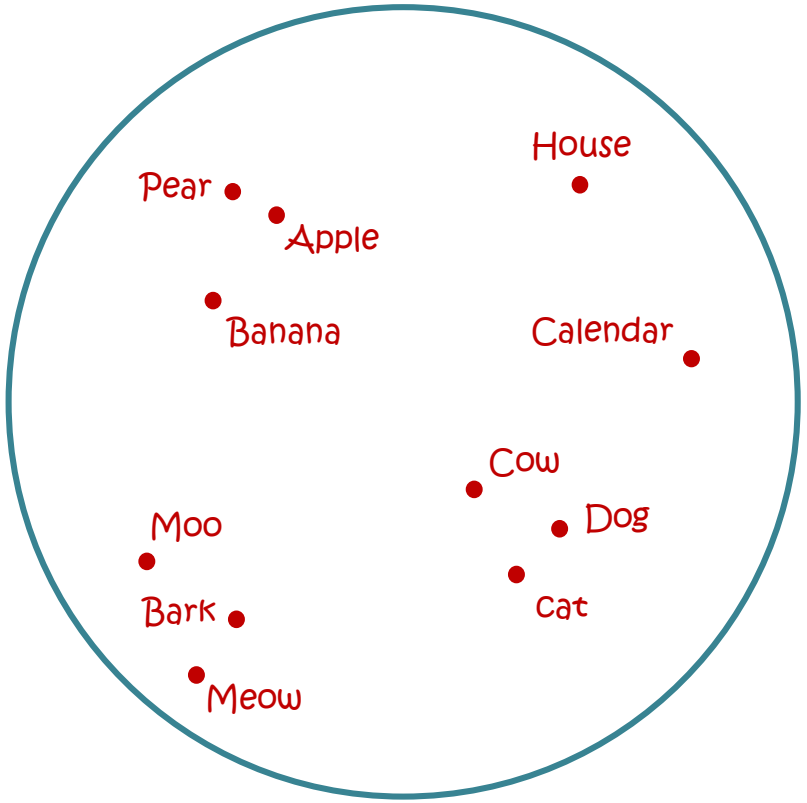
Embeddings

Traditional vector space models:



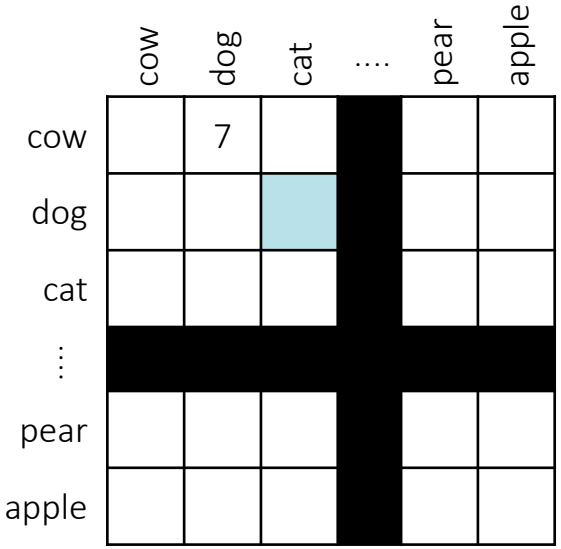
Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



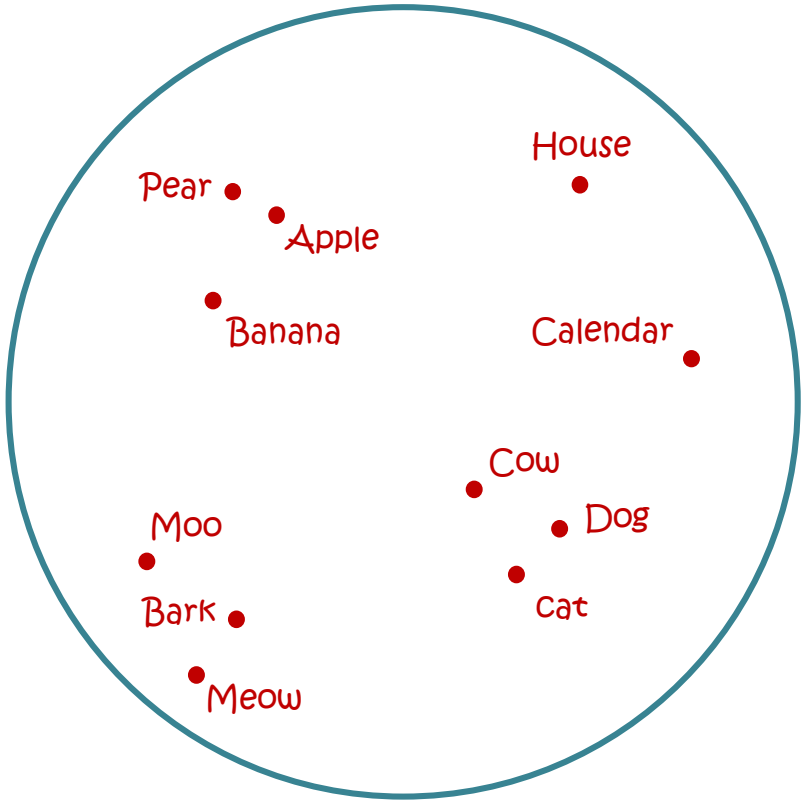
Embeddings

Traditional vector space models:



Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



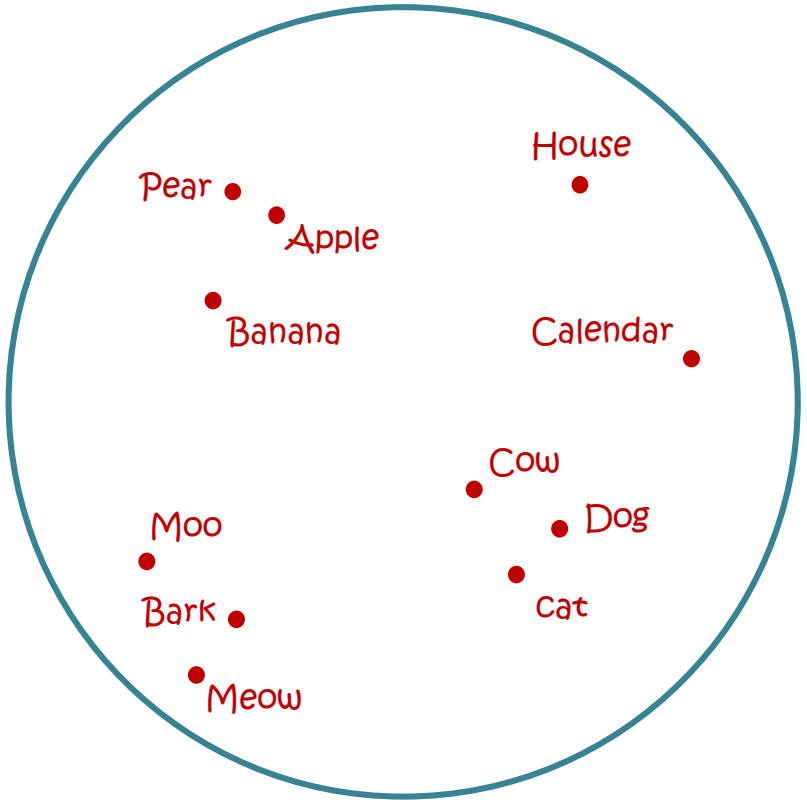
Embeddings

Traditional vector space models:

	cow	dog	cat	....	pear	apple
cow		7				
dog			24			
cat						
⋮						
pear						
apple						

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



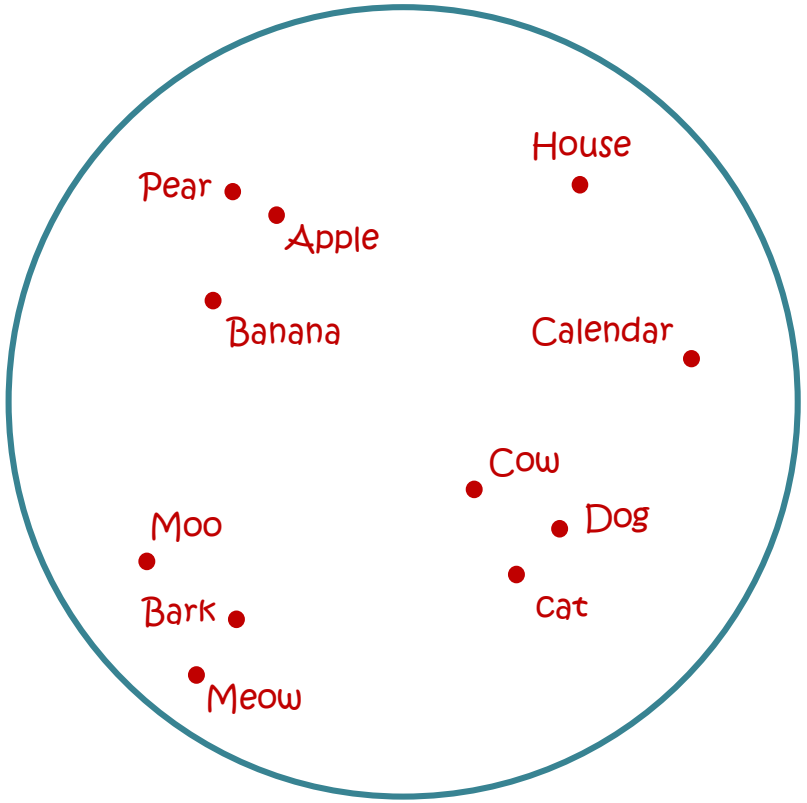
Embeddings

Traditional vector space models:

	cow	dog	cat	....	pear	apple
cow	23	7	4		1	2
dog	7	34	24		2	3
cat	4	24	27		2	1
⋮						
pear	1	2	2		19	21
apple	2	3	1		21	28

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



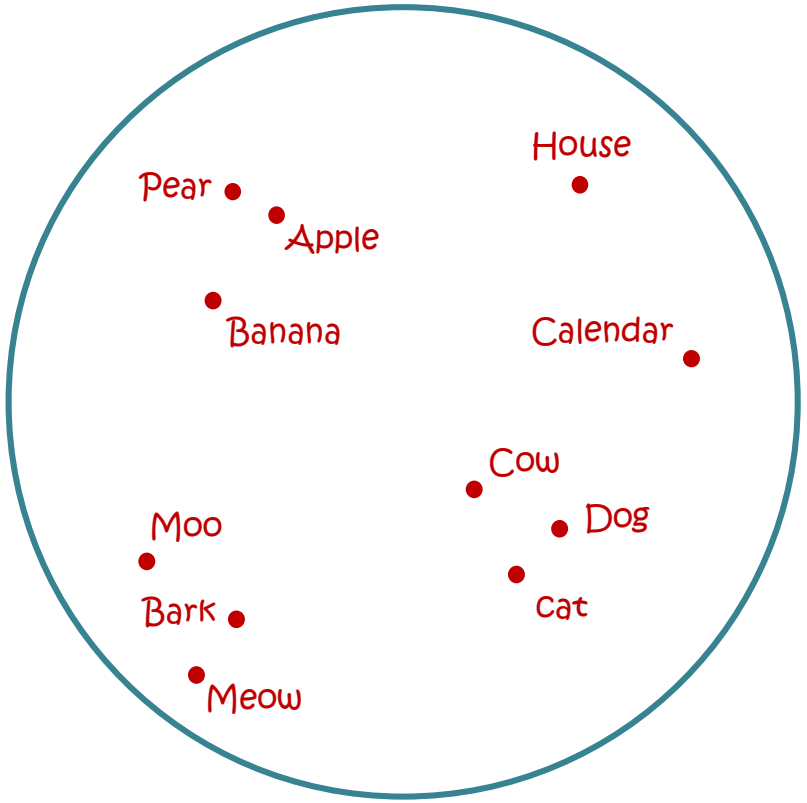
Embeddings

Traditional vector space models:

cow	23	7	4		1	2
dog	7	34	24		2	3
cat	4	24	27		2	1
⋮						
pear	1	2	2		19	21
apple	2	3	1		21	28

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



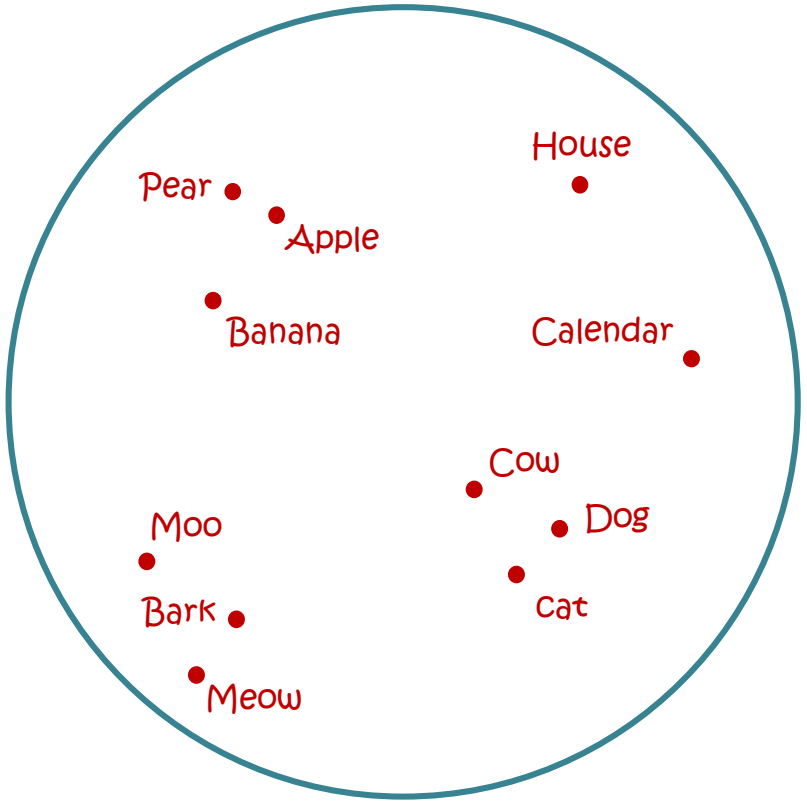
Embeddings

Traditional vector space models:

cow	23	7	4		1	2
dog	7	34	24		2	3
cat	4	24	27		2	1
⋮						
pear	1	2	2		19	21
apple	2	3	1		21	28

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

Traditional vector space models:

cow	23	7	4		1	2
dog	7	34	24		2	3
cat	4	24	27		2	1
⋮						
pear	1	2	2		19	21
apple	2	3	1		21	28

cow: (23, 7, 4, …, 1, 2)

dog: (7, 34, 24, …, 2, 3)

cat: (4, 24, 27, …, 2, 1)

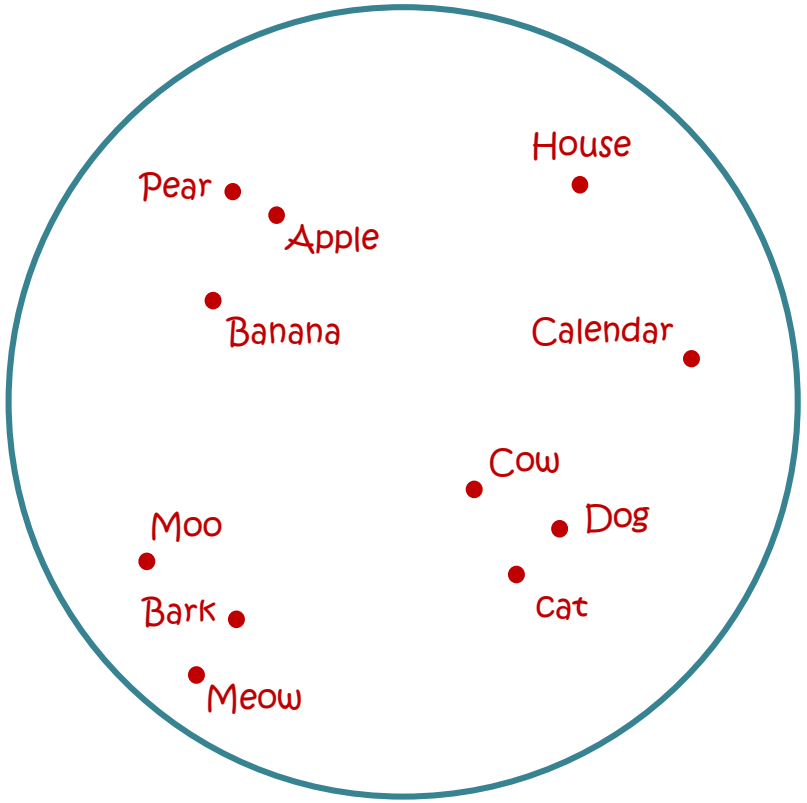
⋮

pear: (1, 2, 2, …, 19, 21)

apple: (2, 3, 1, …, 21, 28)

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

Traditional vector space models:

cow	23	7	4		1	2
dog	7	34	24		2	3
cat	4	24	27		2	1
⋮						
pear	1	2	2		19	21
apple	2	3	1		21	28

cow: (23, 7, 4, …, 1, 2)

dog: (7, 34, 24, …, 2, 3)

cat: (4, 24, 27, …, 2, 1)

⋮

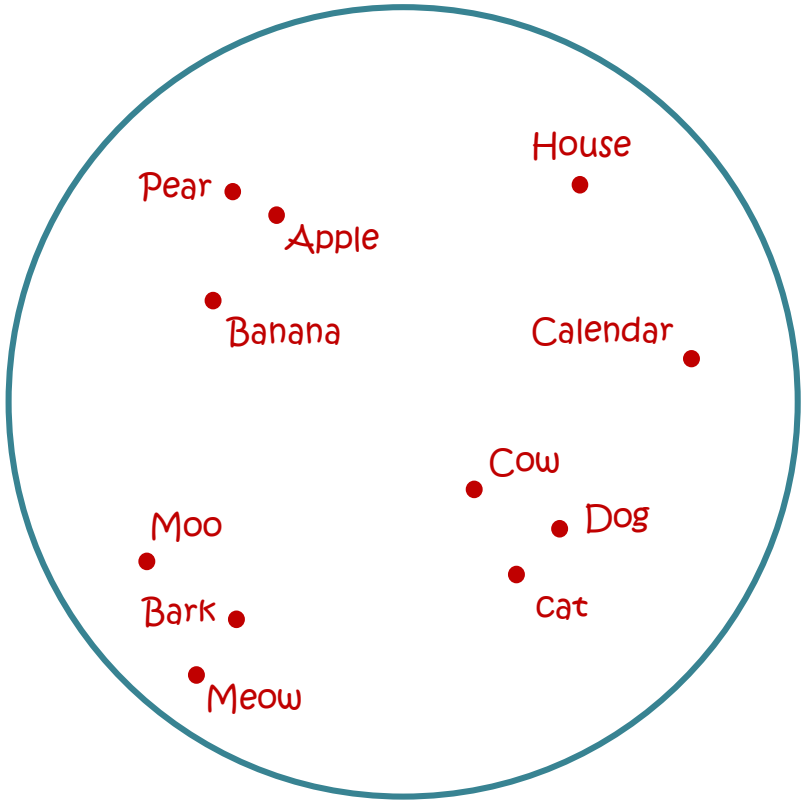
pear: (1, 2, 2, …, 19, 21)

apple: (2, 3, 1, …, 21, 28)

PCA

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

Traditional vector space models:

cow	23	7	4		1	2
dog	7	34	24		2	3
cat	4	24	27		2	1
⋮						
pear	1	2	2		19	21
apple	2	3	1		21	28

cow: (23, 7, 4, …, 1, 2)

dog: (7, 34, 24, …, 2, 3)

cat: (4, 24, 27, …, 2, 1)

⋮

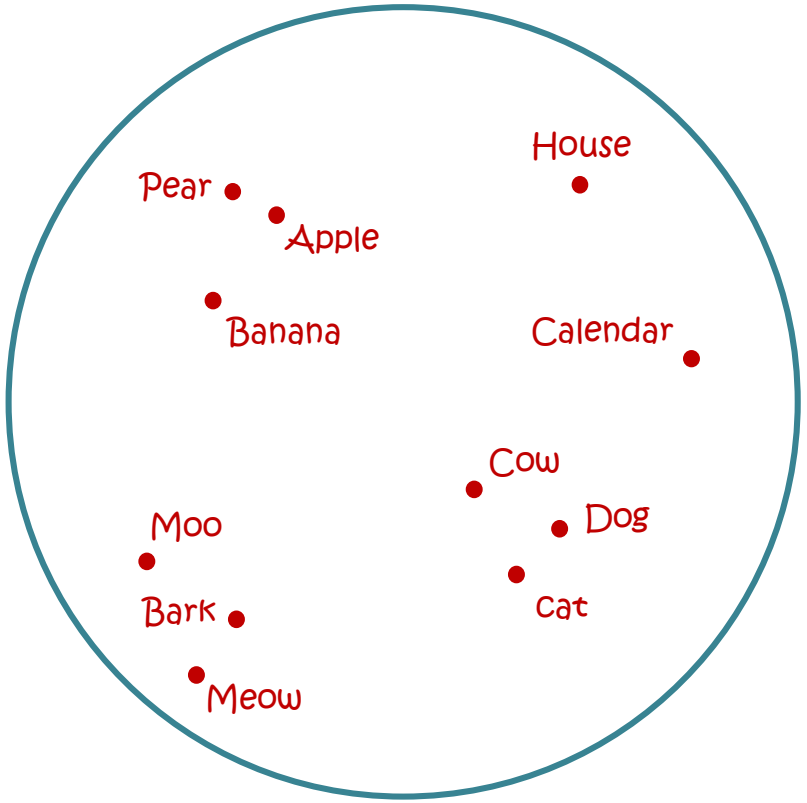
pear: (1, 2, 2, …, 19, 21)

apple: (2, 3, 1, …, 21, 28)

PCA

Semantic space

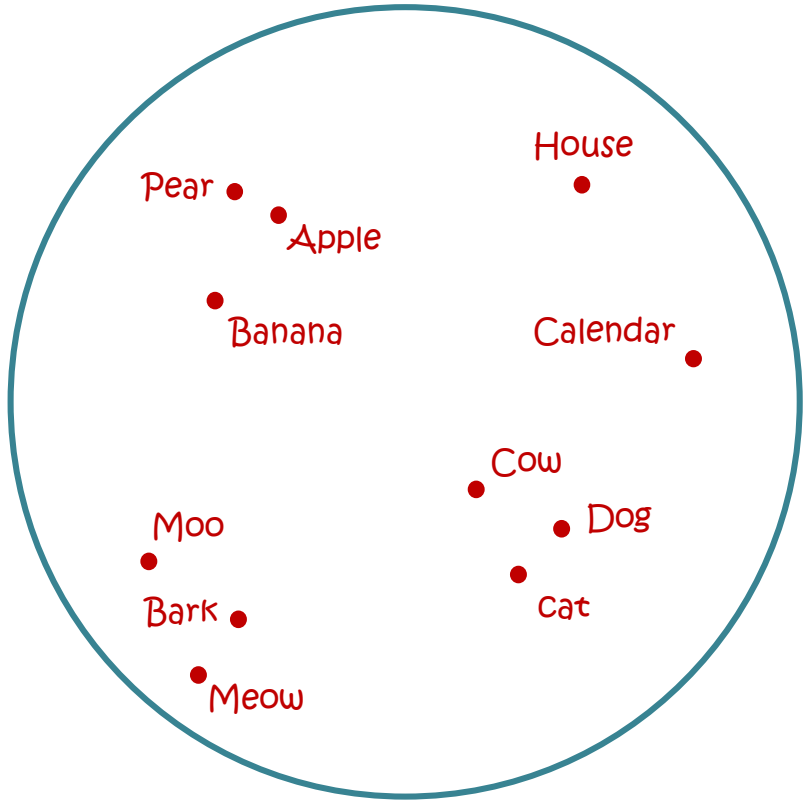
- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts

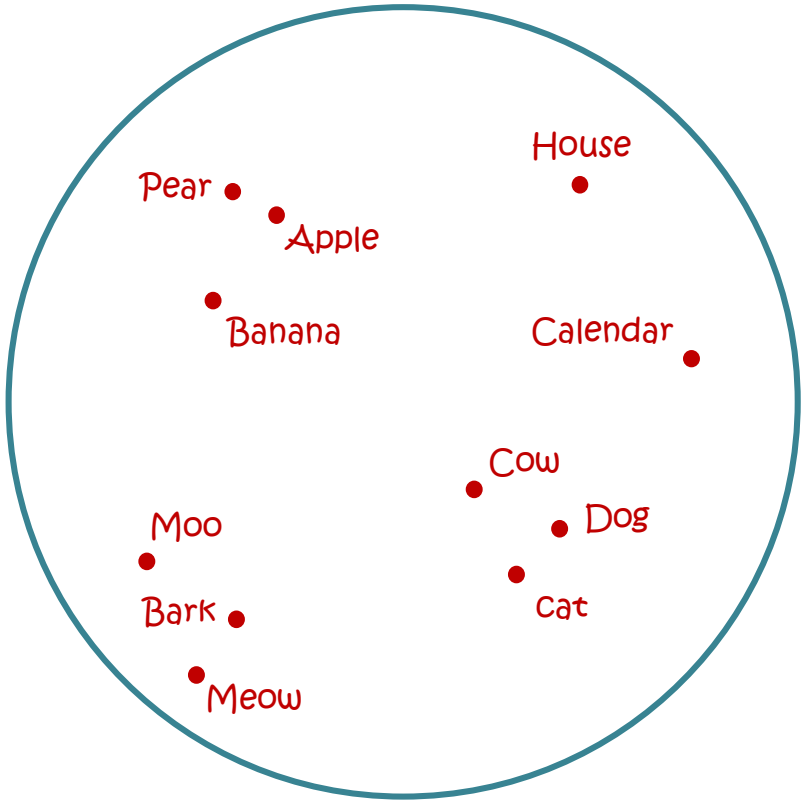


Embeddings

Skip-gram with negative sampling:

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



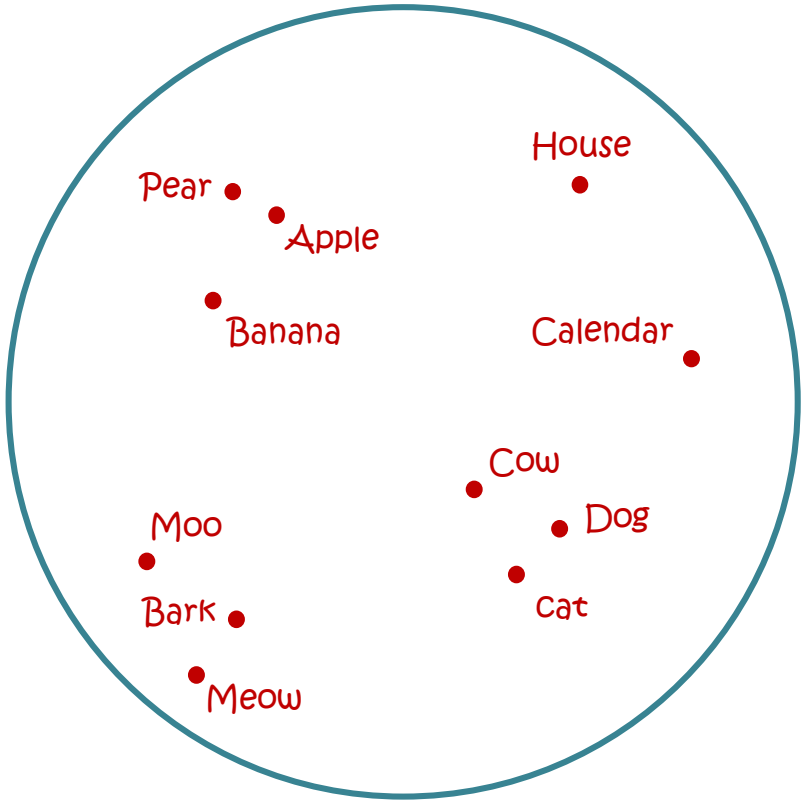
Embeddings

Skip-gram with negative sampling:

$$\log \sigma(w \cdot c) + \sum_{i=1}^{k} \mathbb{E}_{C_N \sim P_D}[\log \sigma(-w \cdot c_N)]$$

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

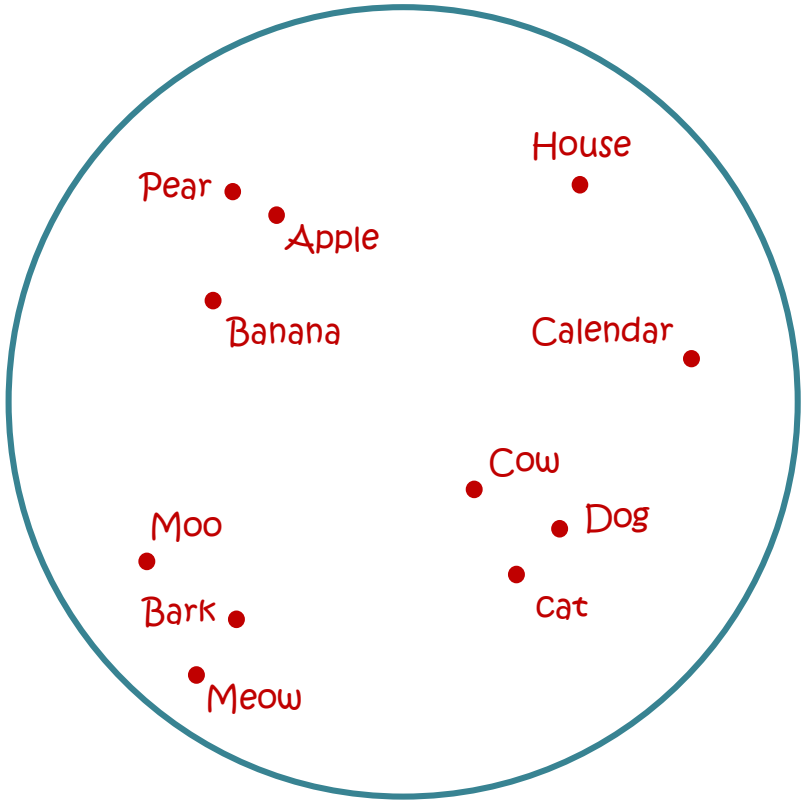
Skip-gram with negative sampling:

$$\log \sigma(w \cdot c) + \sum_{i=1}^{k} \mathbb{E}_{C_N \sim P_D}[\log \sigma(-w \cdot c_N)]$$

I will go to New York by plane .

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

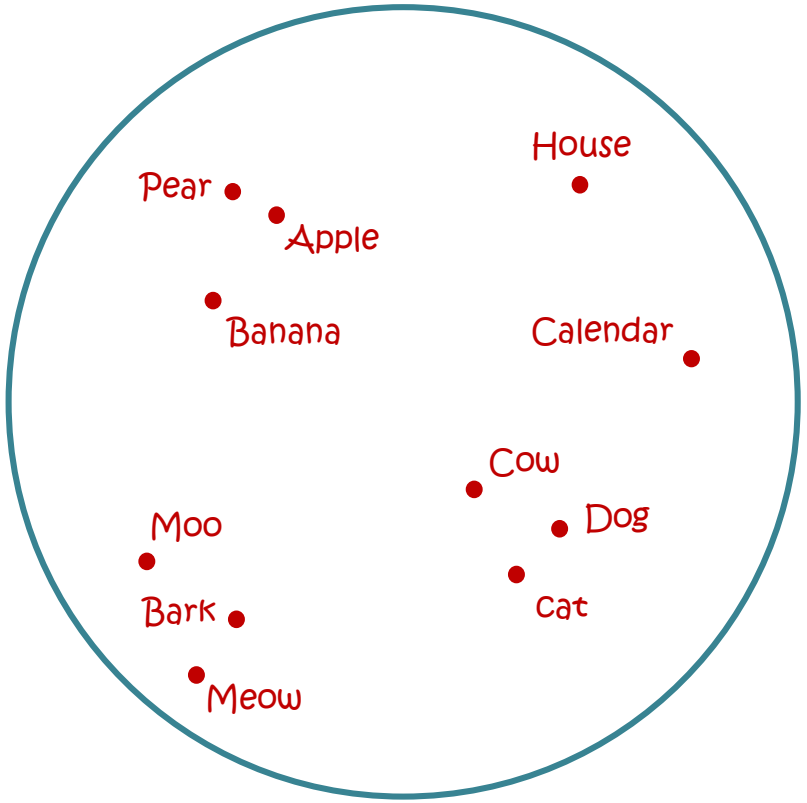
Skip-gram with negative sampling:

$$\log \sigma(w \cdot c) + \sum_{i=1}^{k} \mathbb{E}_{C_N \sim P_D}[\log \sigma(-w \cdot c_N)]$$

I will $\underset{w}{\underline{go}}$ *to New York by plane .*

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

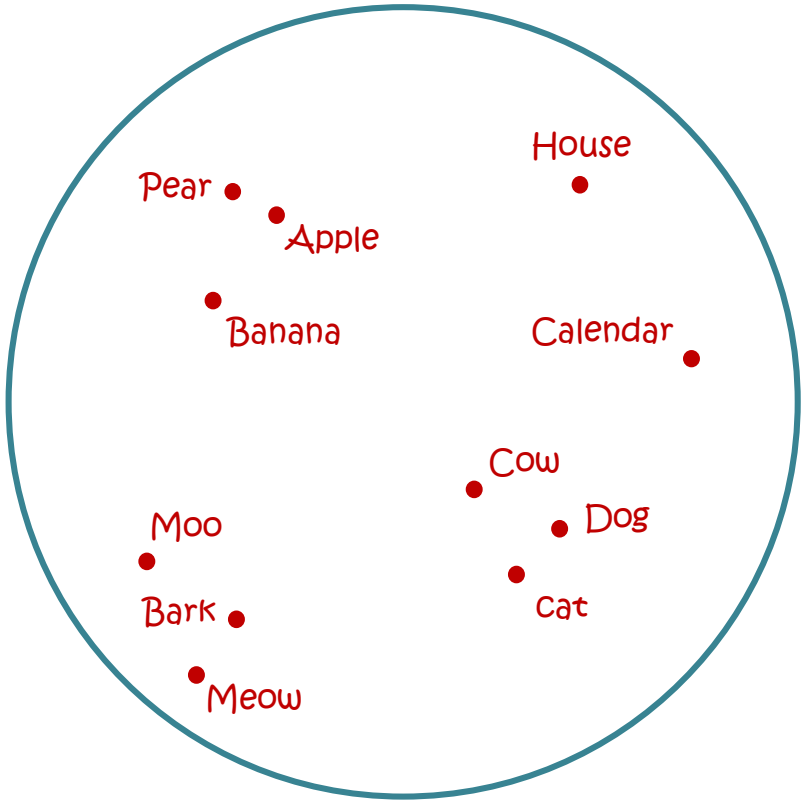
Skip-gram with negative sampling:

$$\log \sigma(w \cdot c) + \sum_{i=1}^{k} \mathbb{E}_{C_N \sim P_D}[\log \sigma(-w \cdot c_N)]$$

I will $\underset{w}{\underline{go}}$ *to New York by plane .*

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

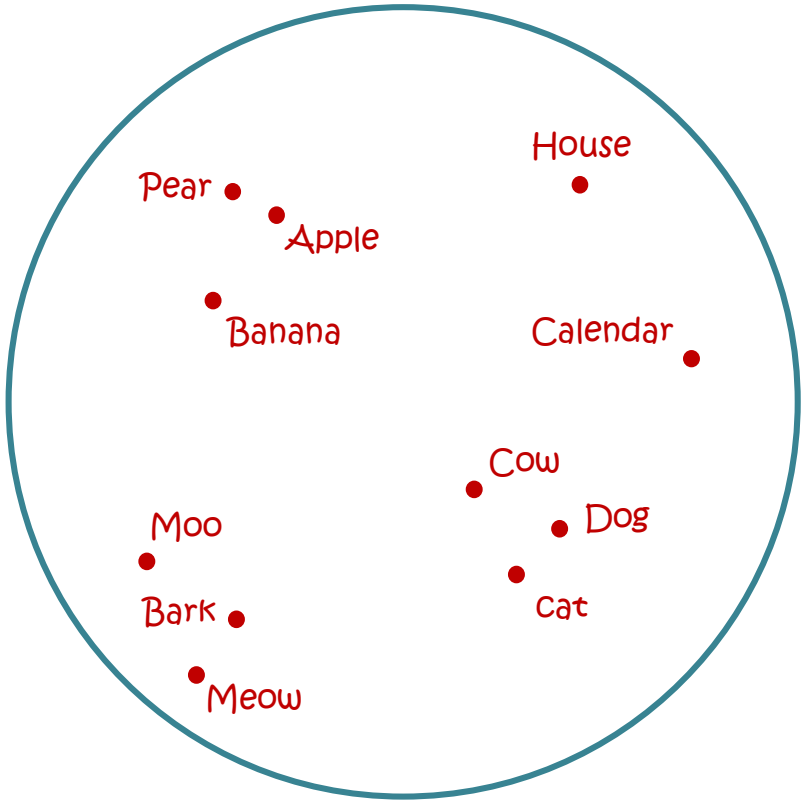
Skip-gram with negative sampling:

$$\log \sigma(w \cdot c) + \sum_{i=1}^{k} \mathbb{E}_{C_N \sim P_D}[\log \sigma(-w \cdot c_N)]$$

I will $\underset{w}{go}$ $\underset{c}{to}$ *New York by plane .*

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

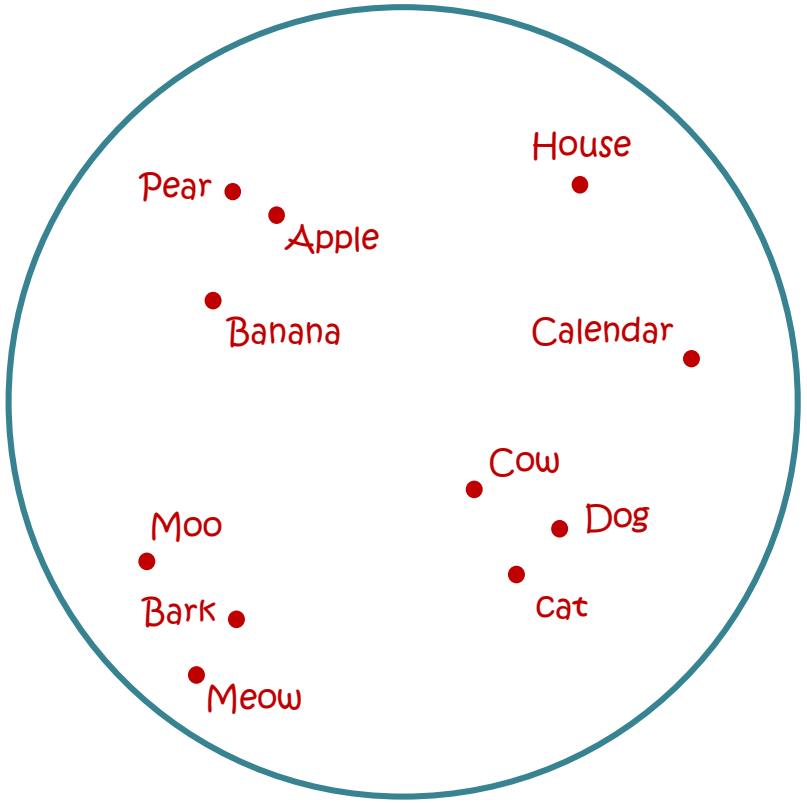
Skip-gram with negative sampling:

$$\log \sigma(w \cdot c) + \sum_{i=1}^{k} \mathbb{E}_{C_N \sim P_D}[\log \sigma(-w \cdot c_N)]$$

I will $\underset{w}{go}$ *to* $\underset{c}{New}$ *York by plane .*

Semantic space

- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Embeddings

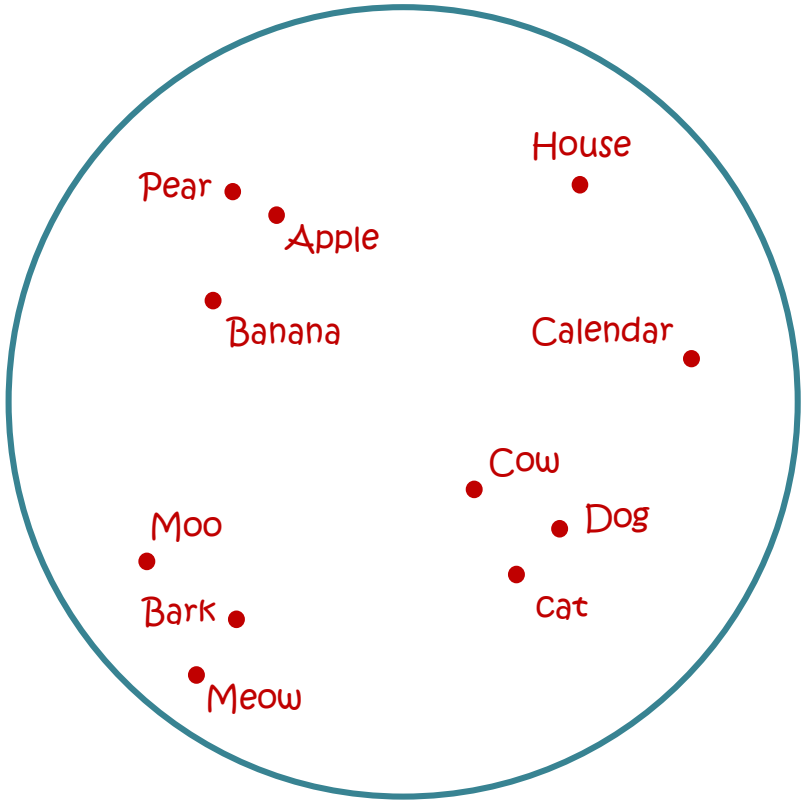
Skip-gram with negative sampling:

$$\log \sigma(w \cdot c) + \sum_{i=1}^{k} \mathbb{E}_{C_N \sim P_D}[\log \sigma(-w \cdot c_N)]$$

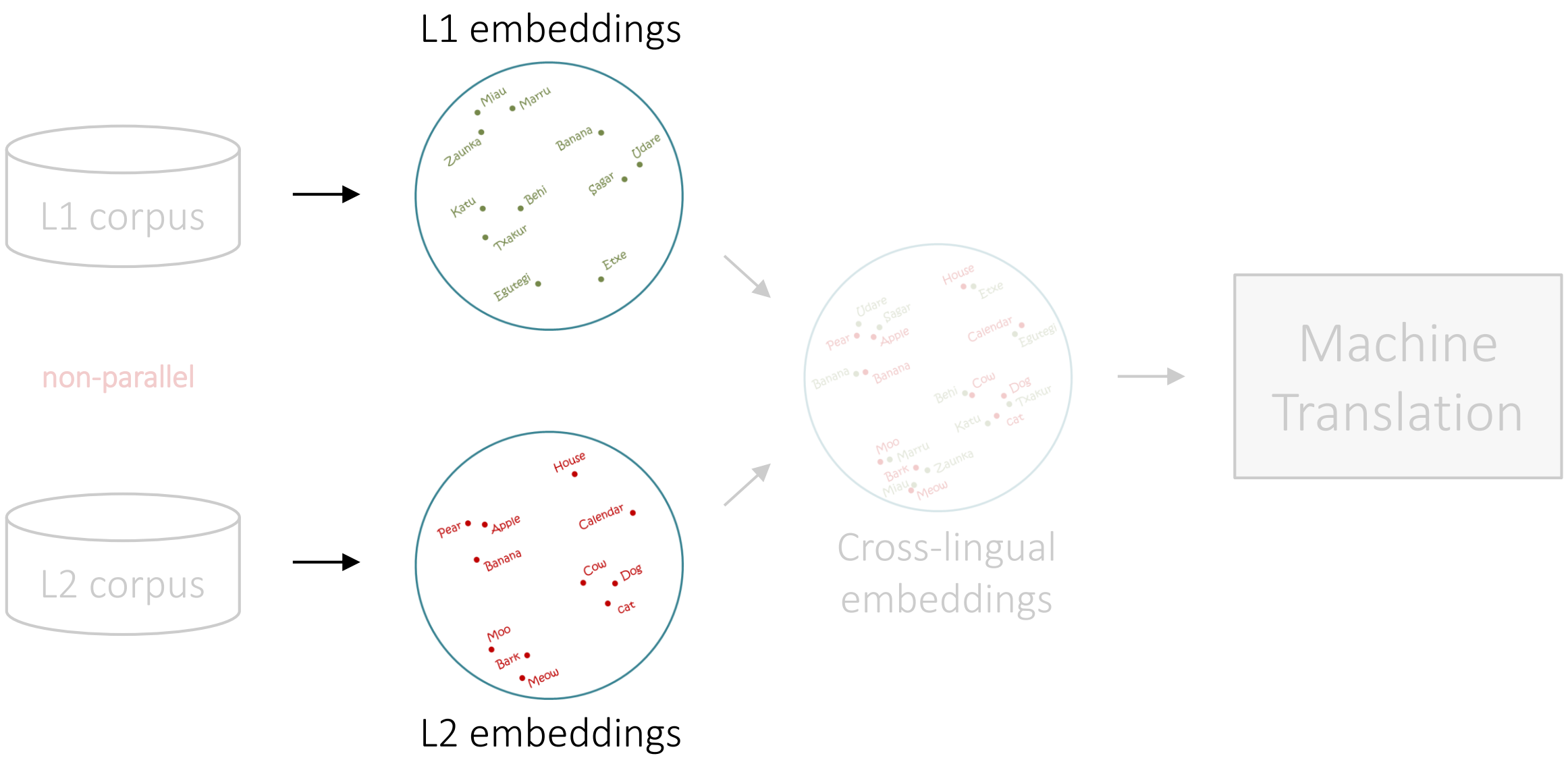
I will $\underset{w}{go}$ *to New* $\underset{c}{York}$ *by plane .*

Semantic space

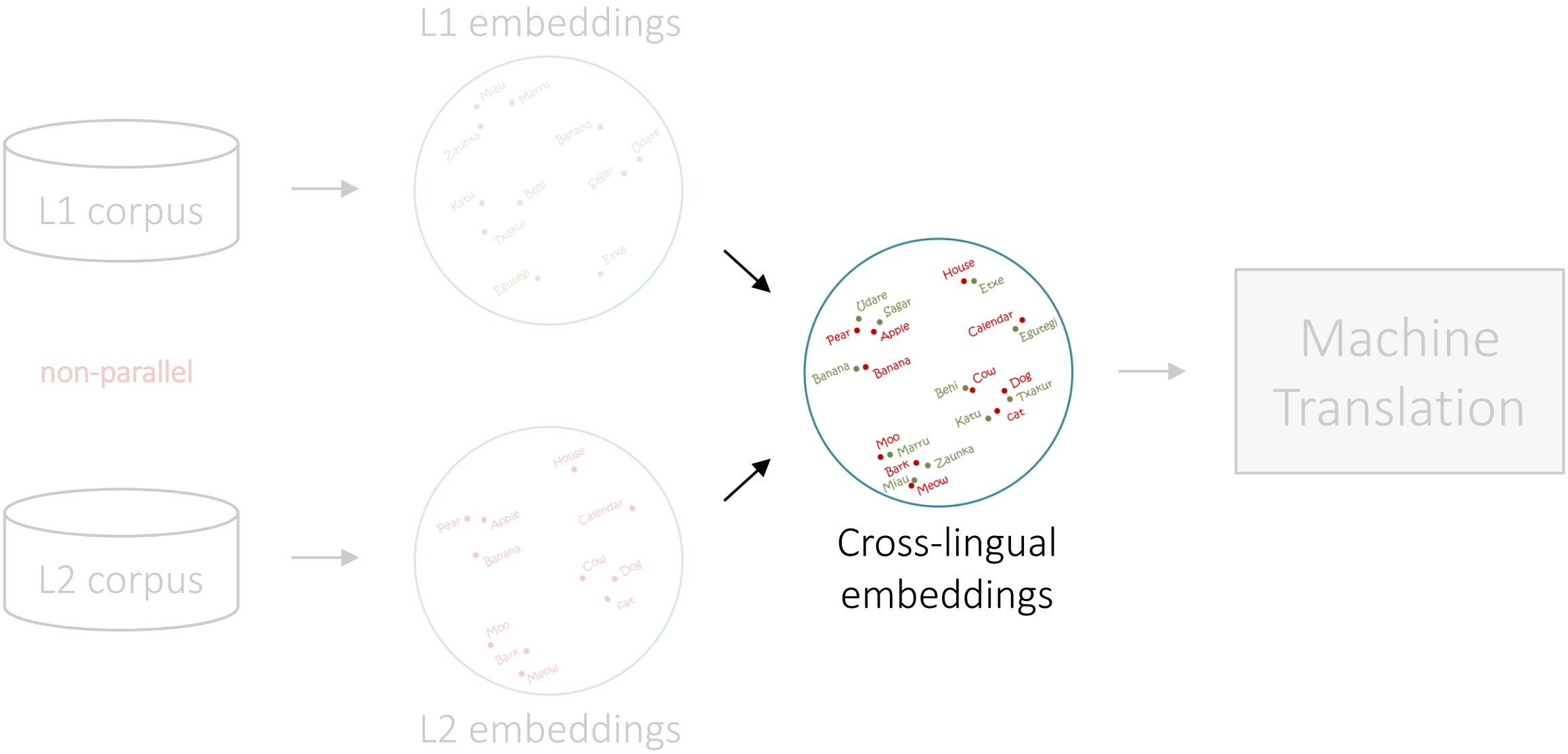
- Words
- Meaningful distances
- Meaningful relations
- 300 dimensions
- Machine learning / linear algebra from co-occurrence counts



Outline

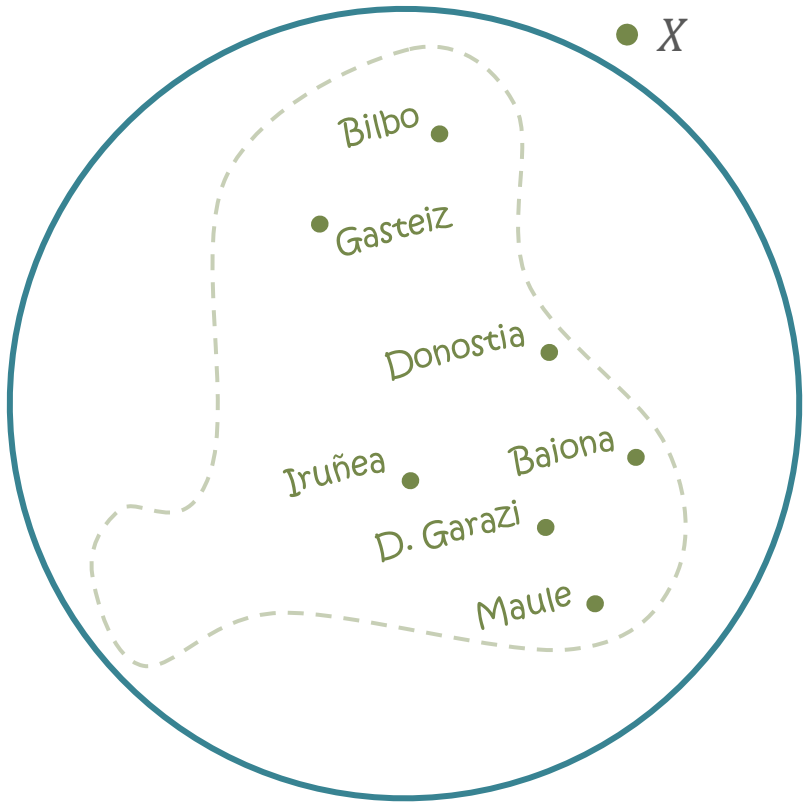

L1 embeddings
L1 corpus
non-parallel
L2 corpus
L2 embeddings
Cross-lingual embeddings
Machine Translation

Outline

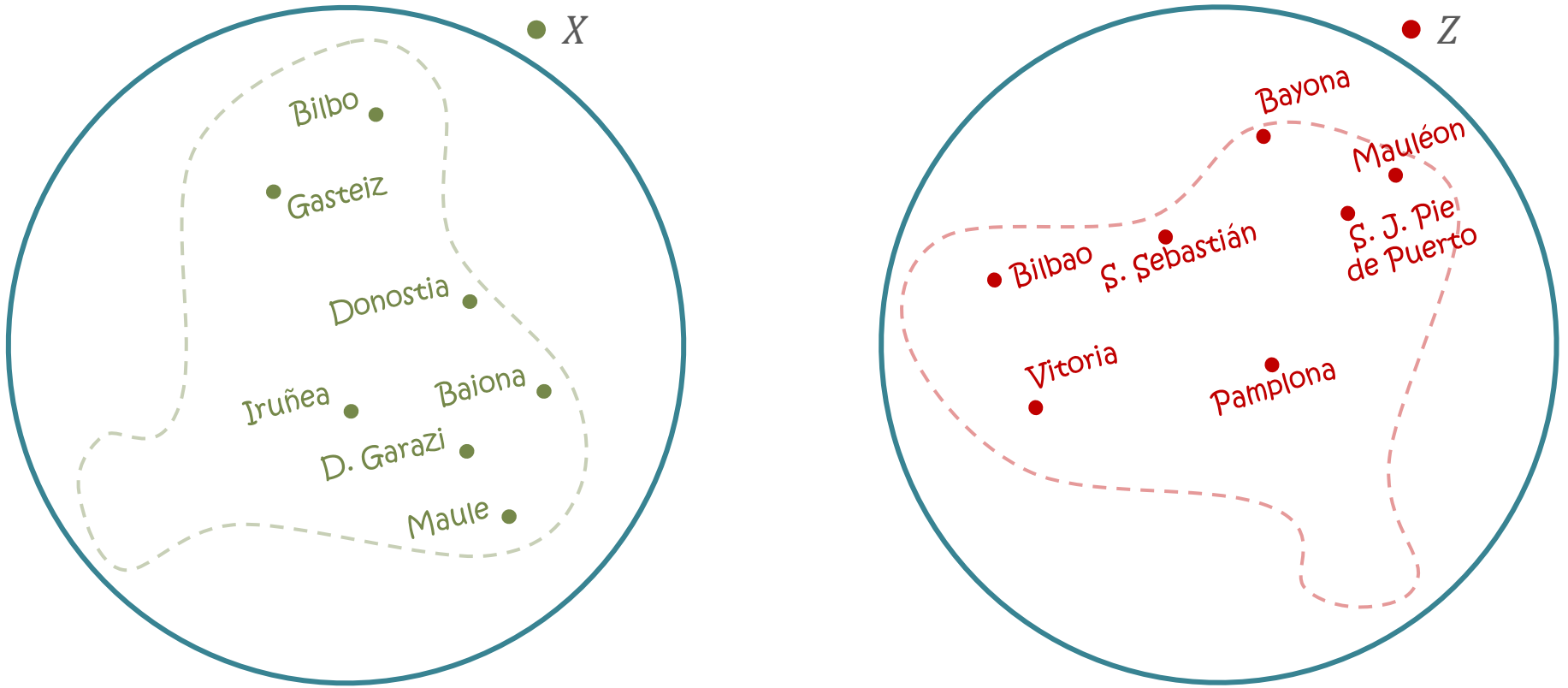

L1 embeddings
L1 corpus
non-parallel
L2 corpus
L2 embeddings
Cross-lingual embeddings
Machine Translation

Cross-lingual embedding mappings

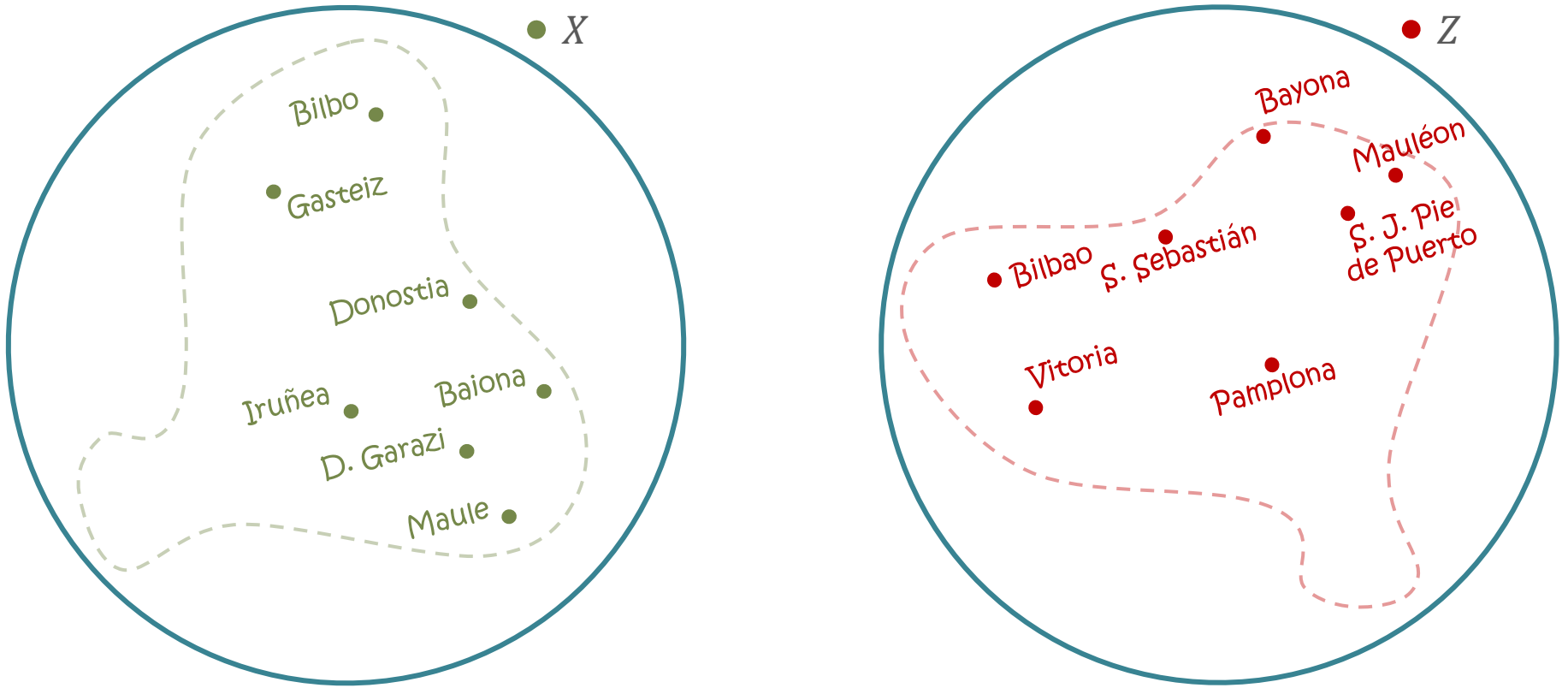
Cross-lingual embedding mappings



Cross-lingual embedding mappings

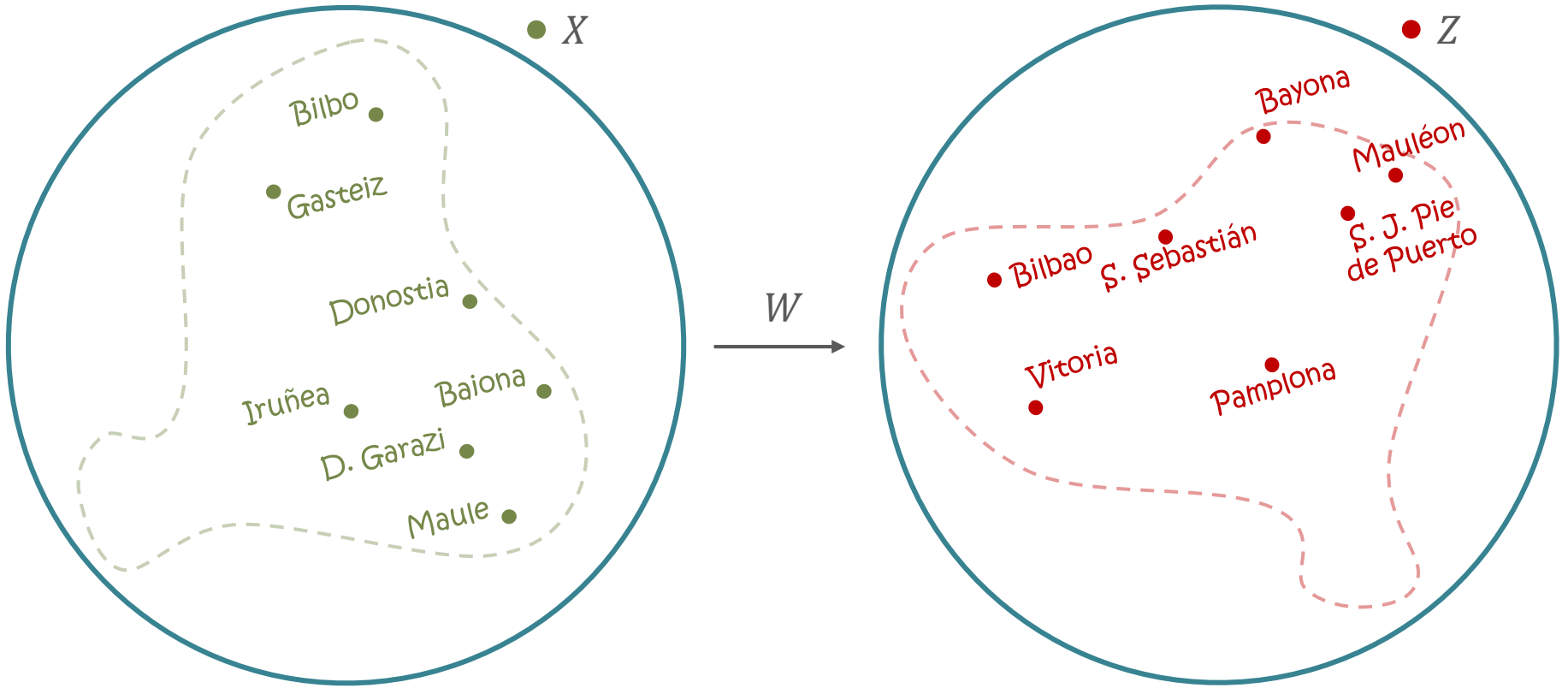


Cross-lingual embedding mappings



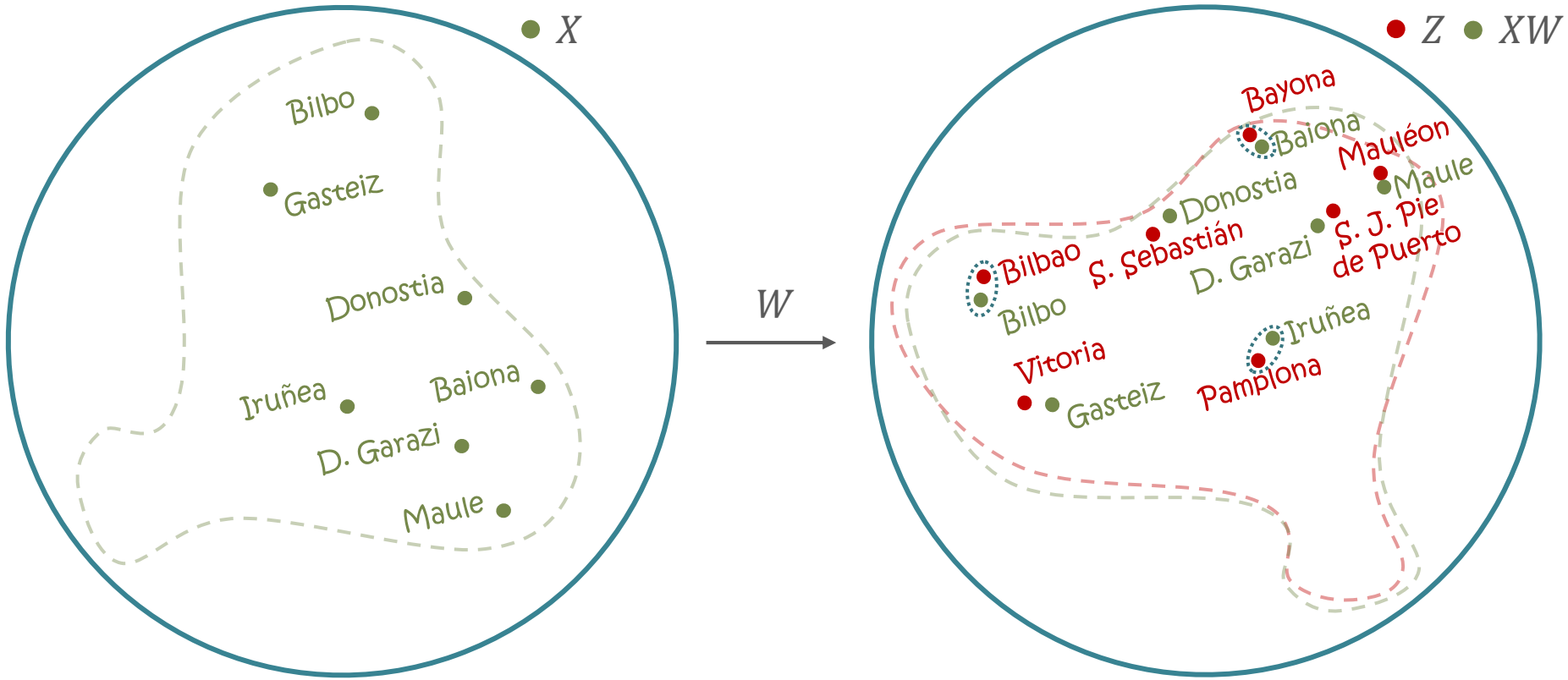
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Cross-lingual embedding mappings



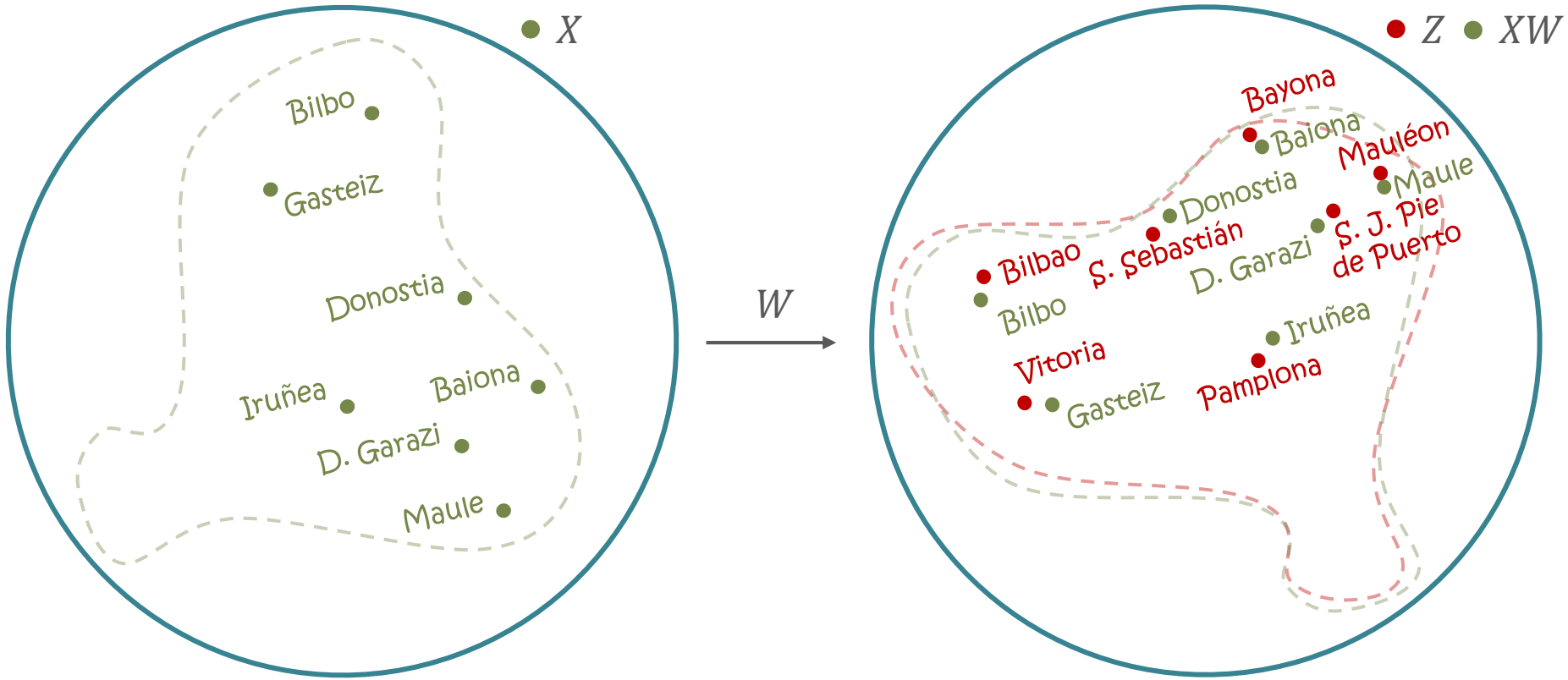
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Cross-lingual embedding mappings



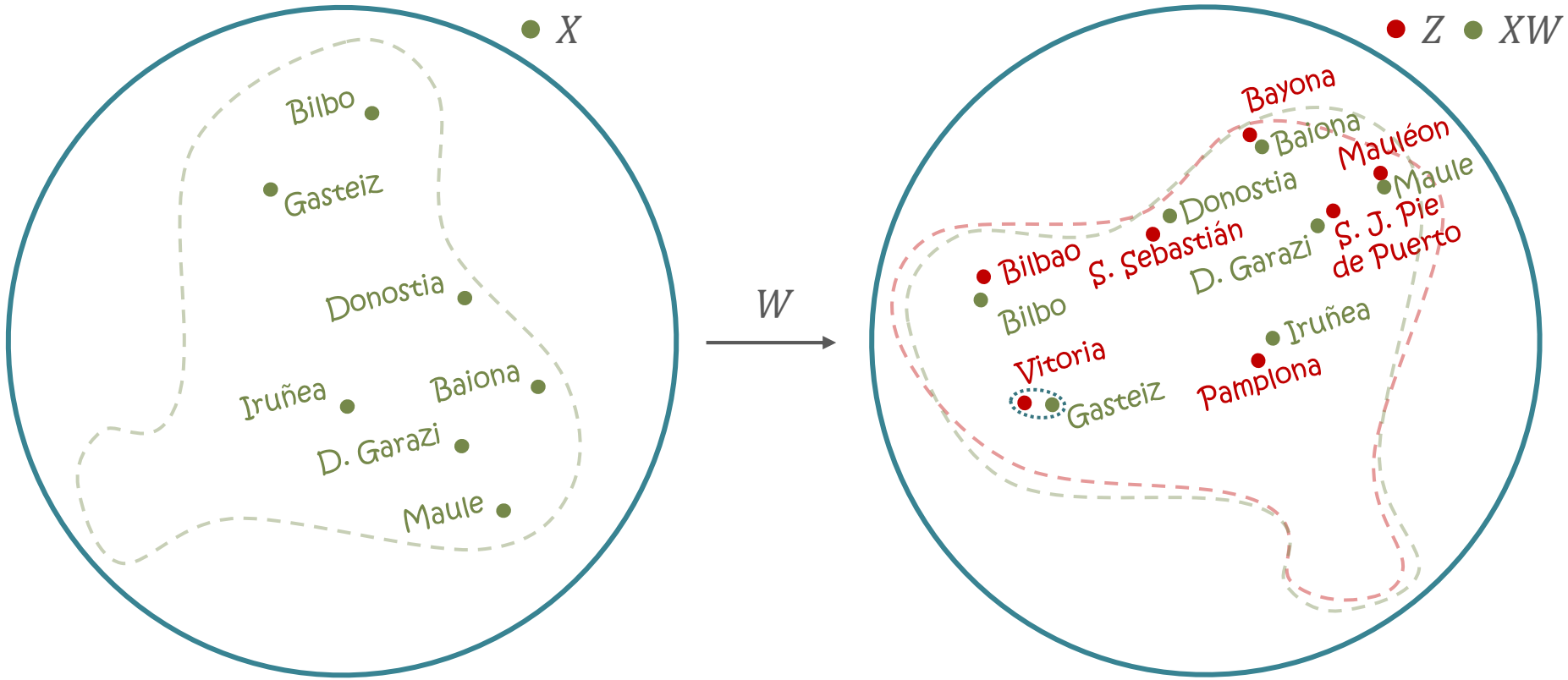
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Cross-lingual embedding mappings



Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

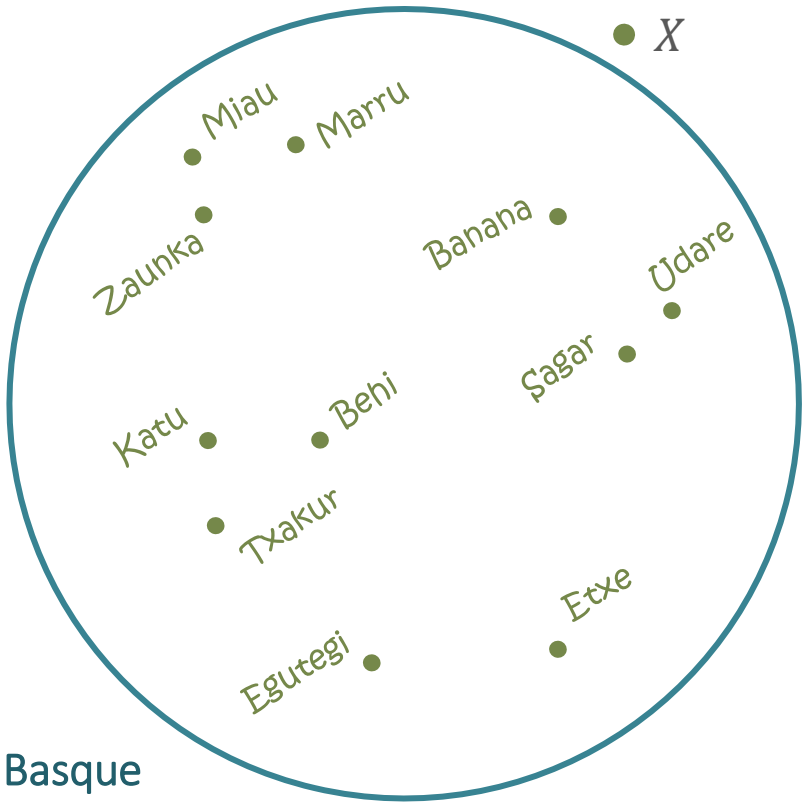
Cross-lingual embedding mappings



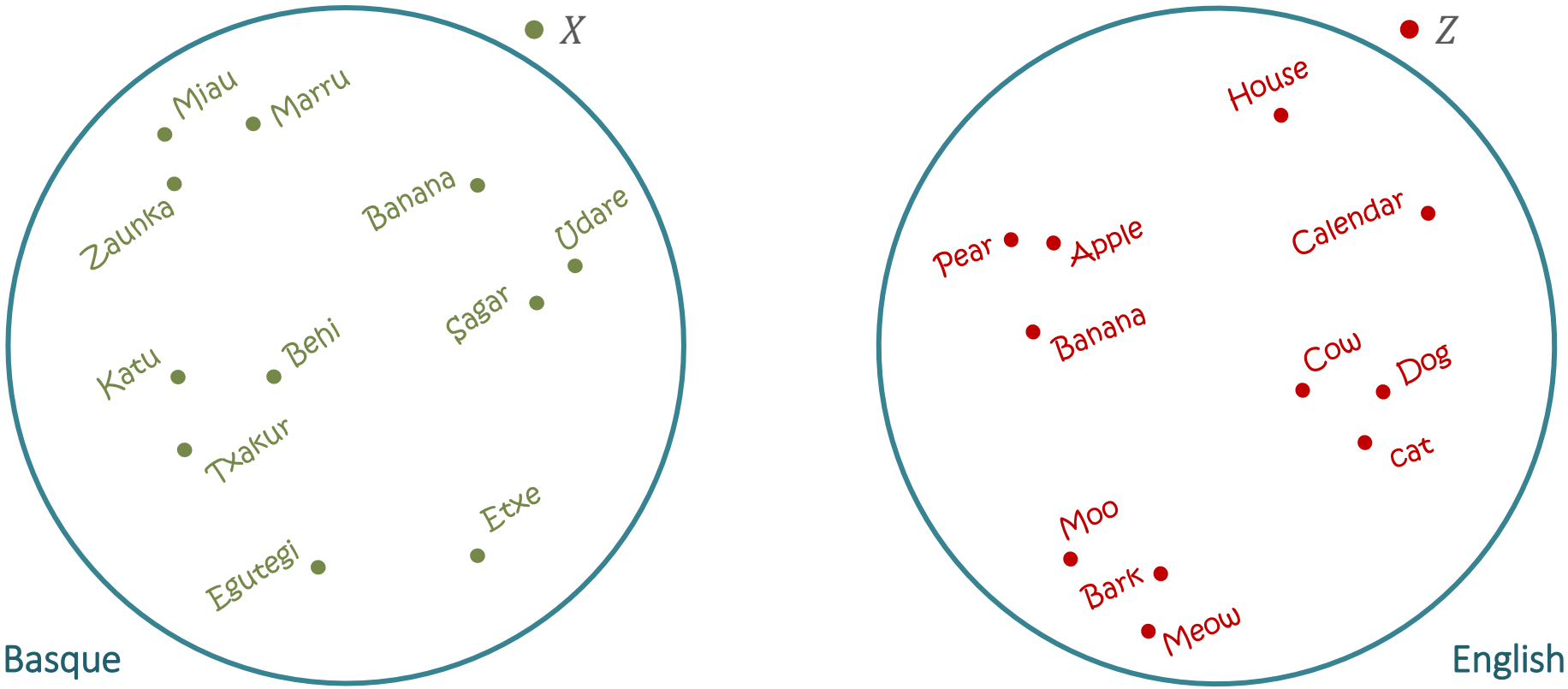
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Cross-lingual embedding mappings

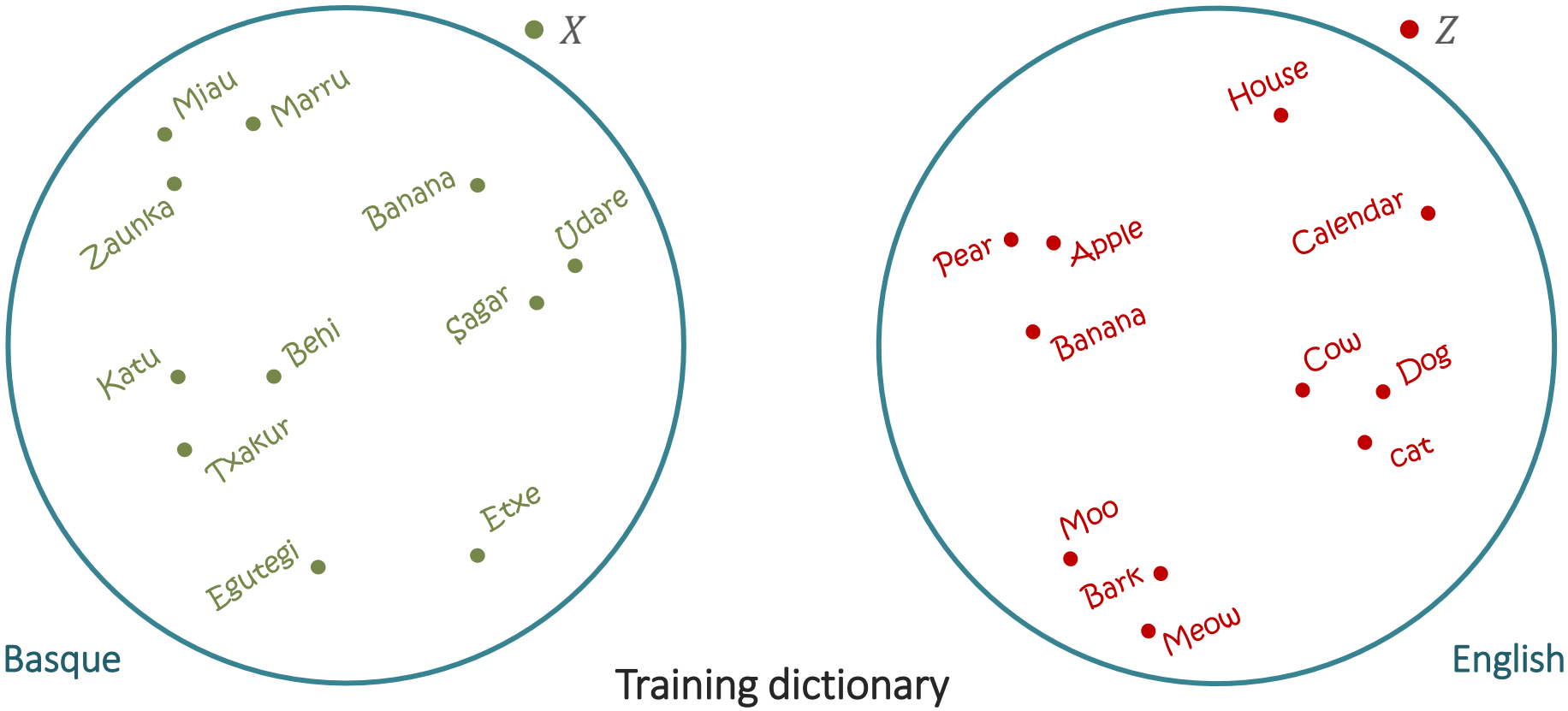
Cross-lingual embedding mappings



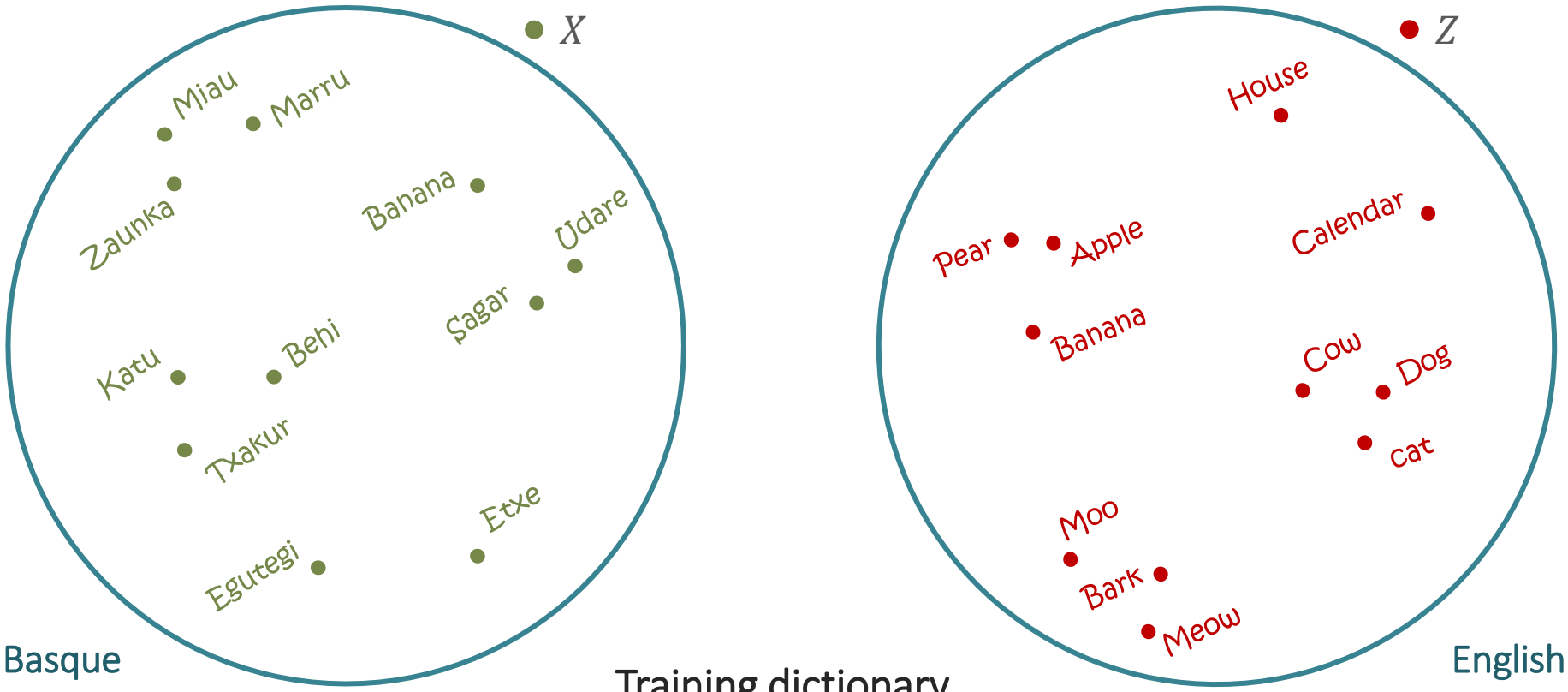
Cross-lingual embedding mappings



Cross-lingual embedding mappings



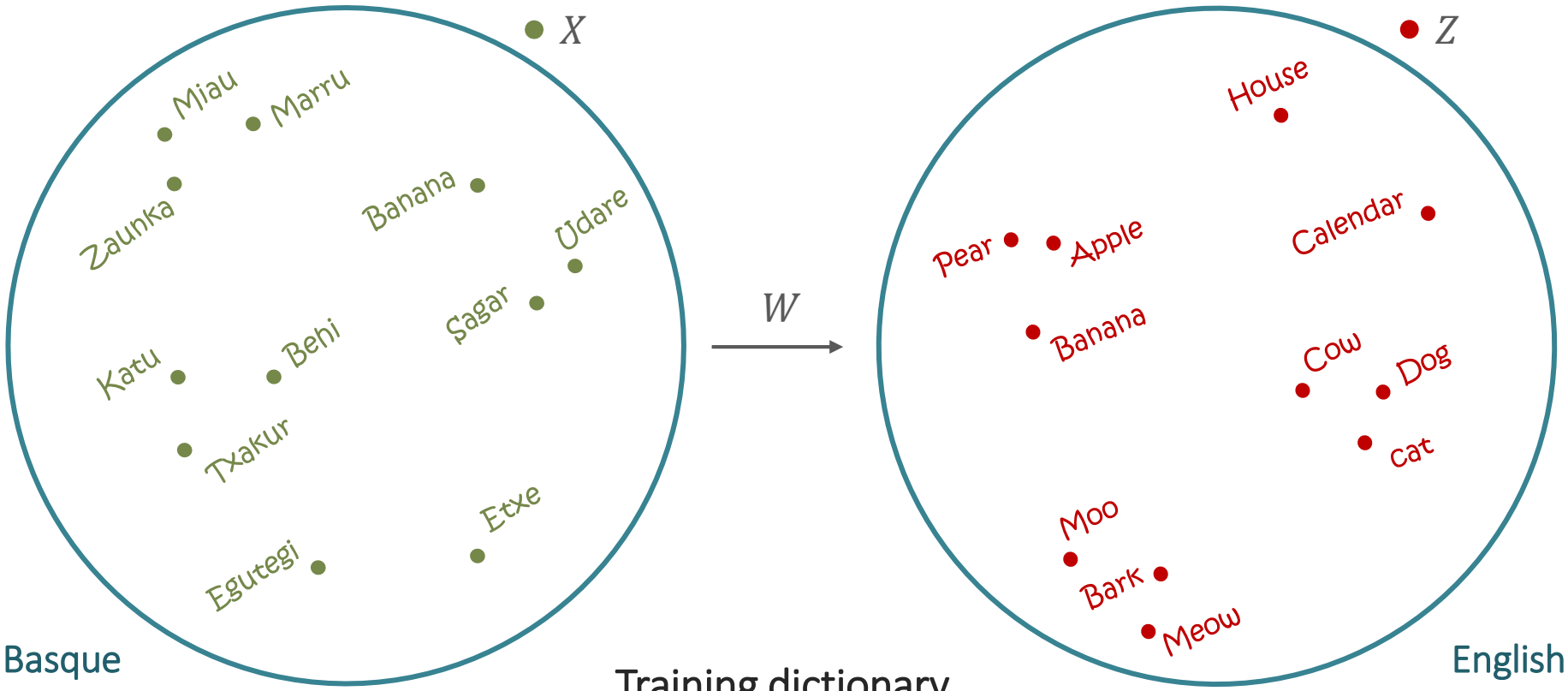
Cross-lingual embedding mappings



Training dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

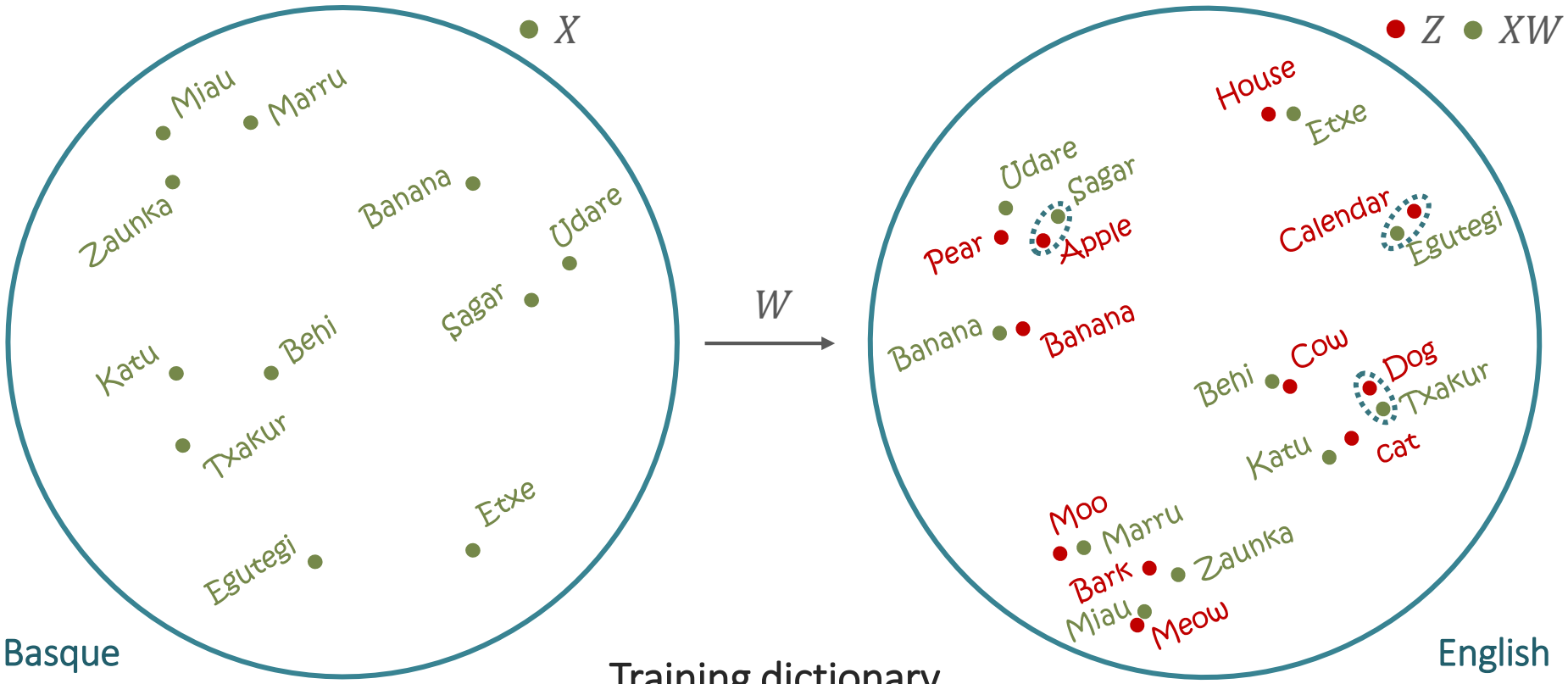
Cross-lingual embedding mappings



Training dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

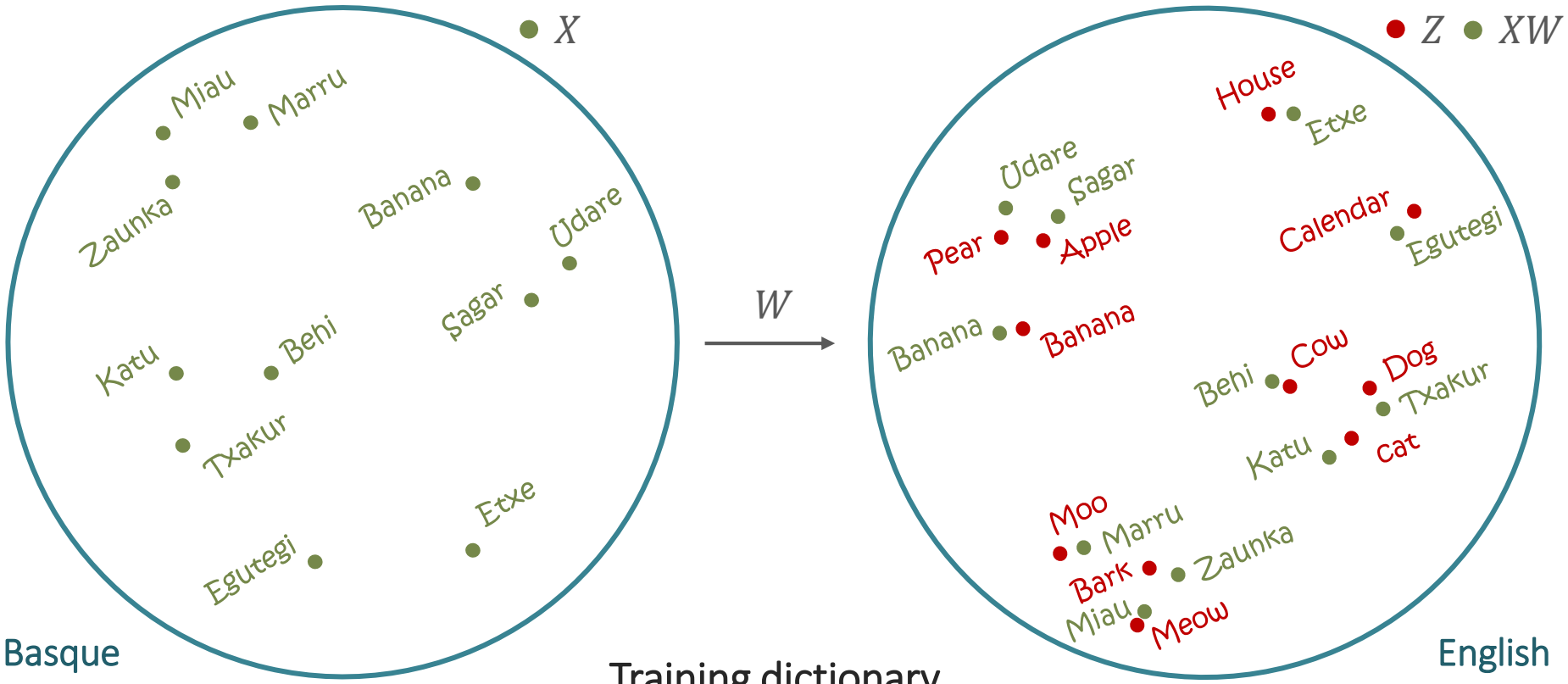
Cross-lingual embedding mappings



Training dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

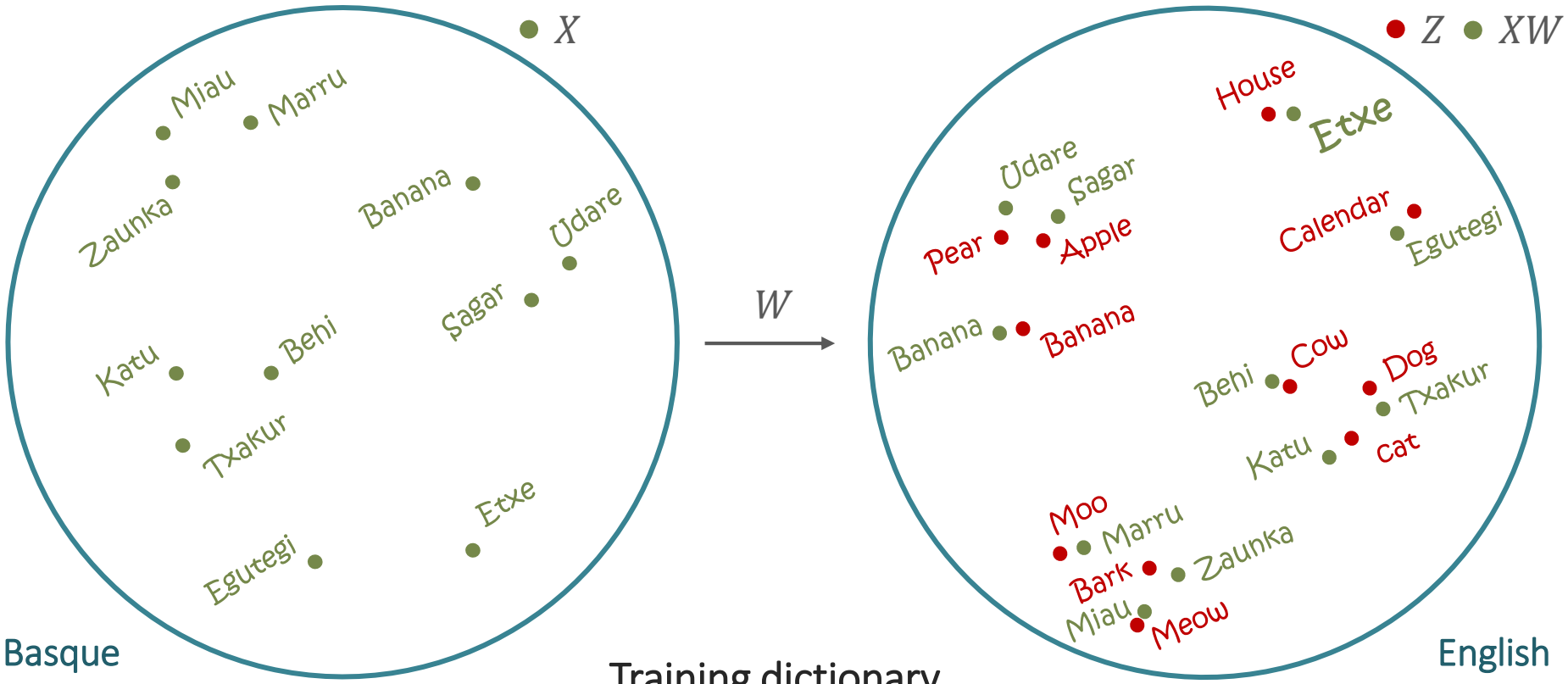
Cross-lingual embedding mappings



Training dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

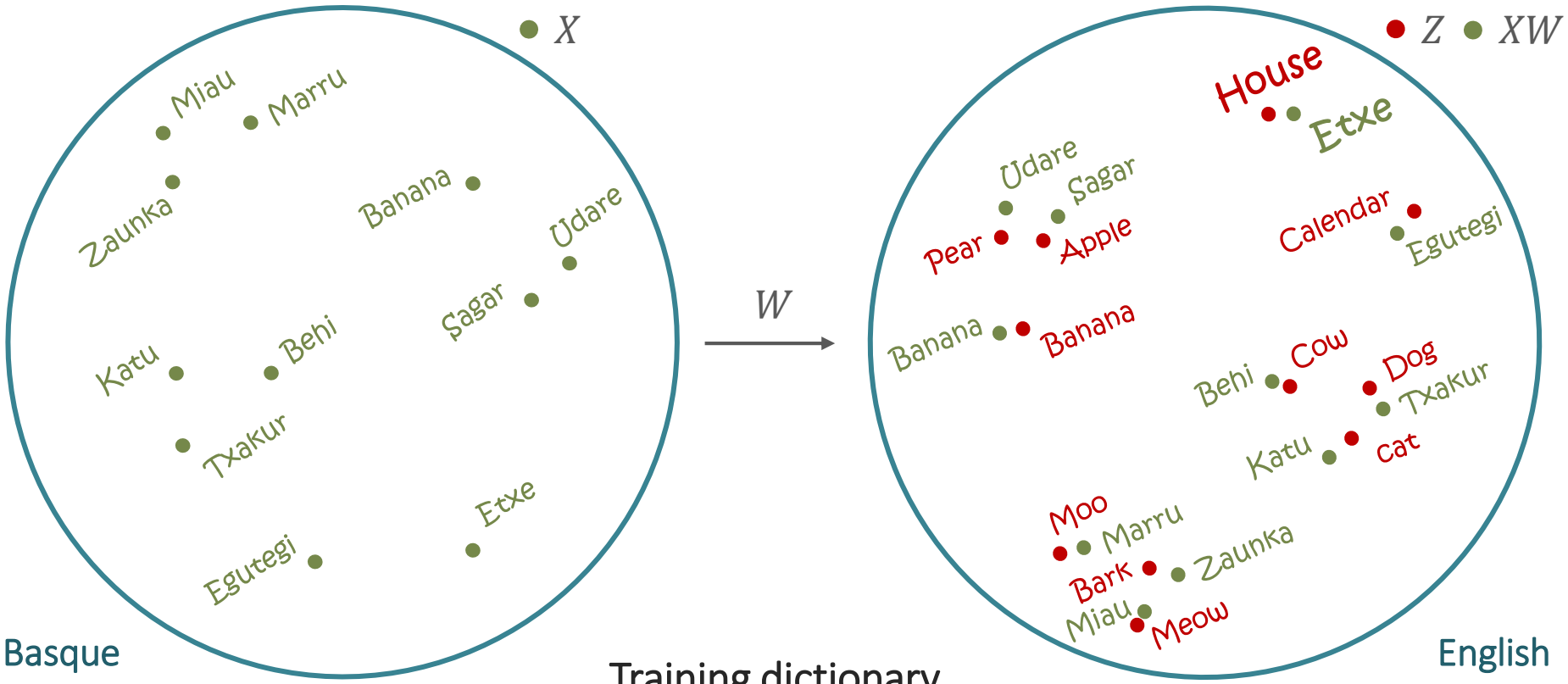
Cross-lingual embedding mappings



Training dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

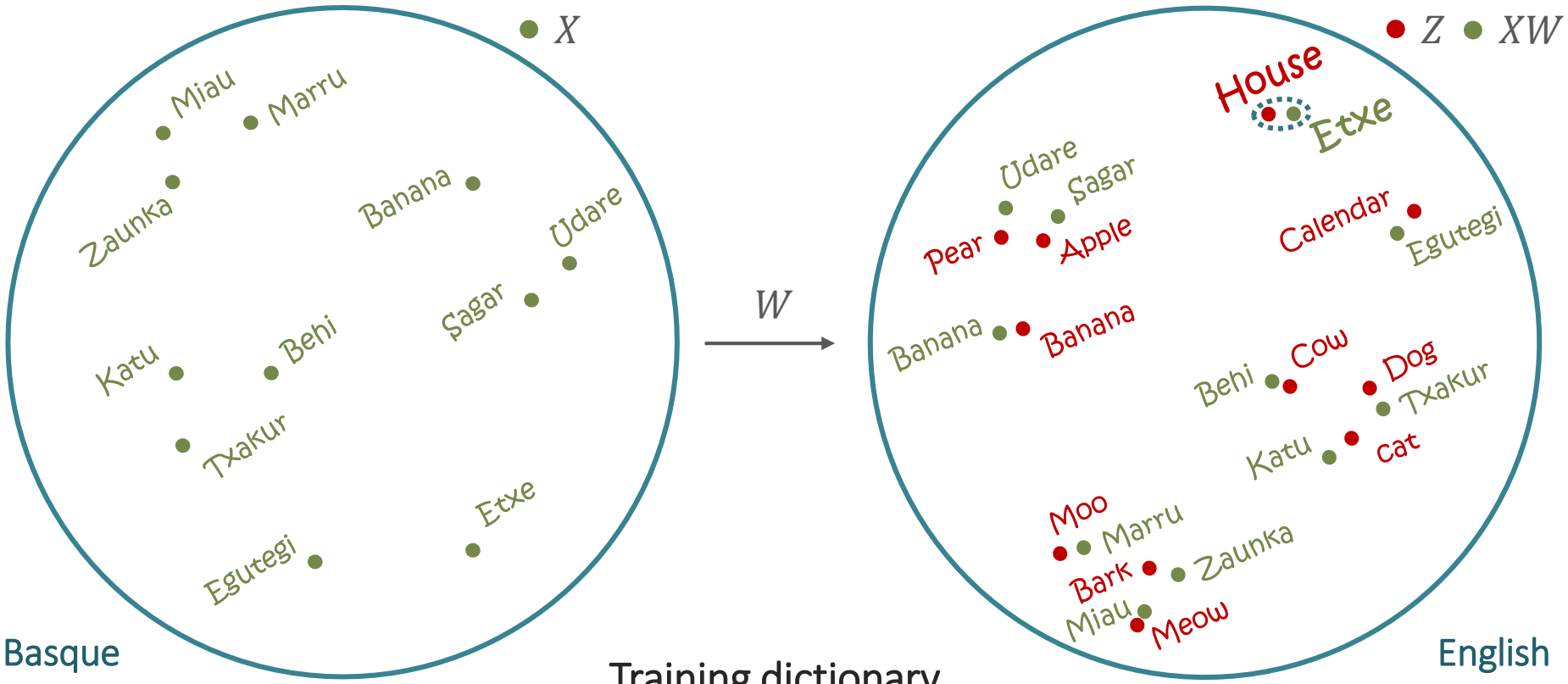
Cross-lingual embedding mappings



Training dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

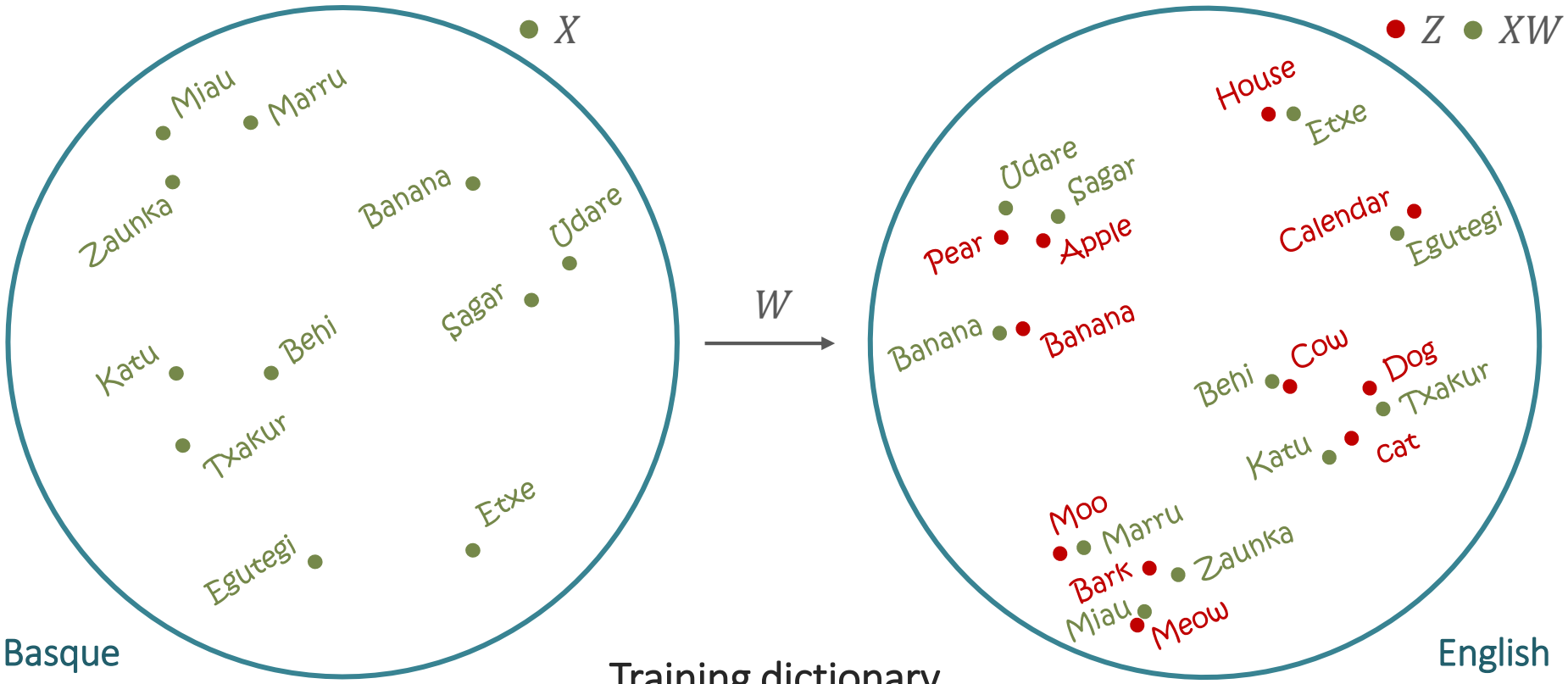
Cross-lingual embedding mappings



Training dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

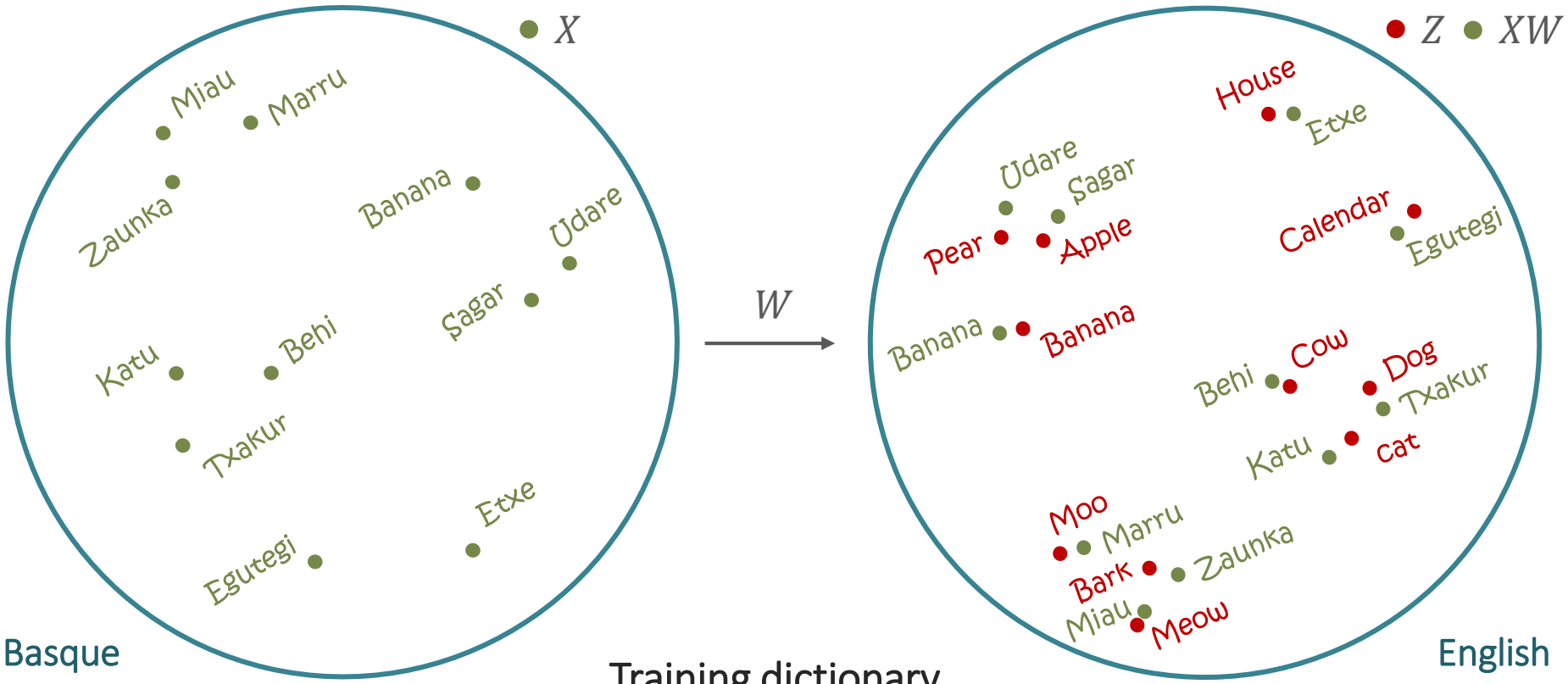
Cross-lingual embedding mappings



Training dictionary

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

Cross-lingual embedding mappings



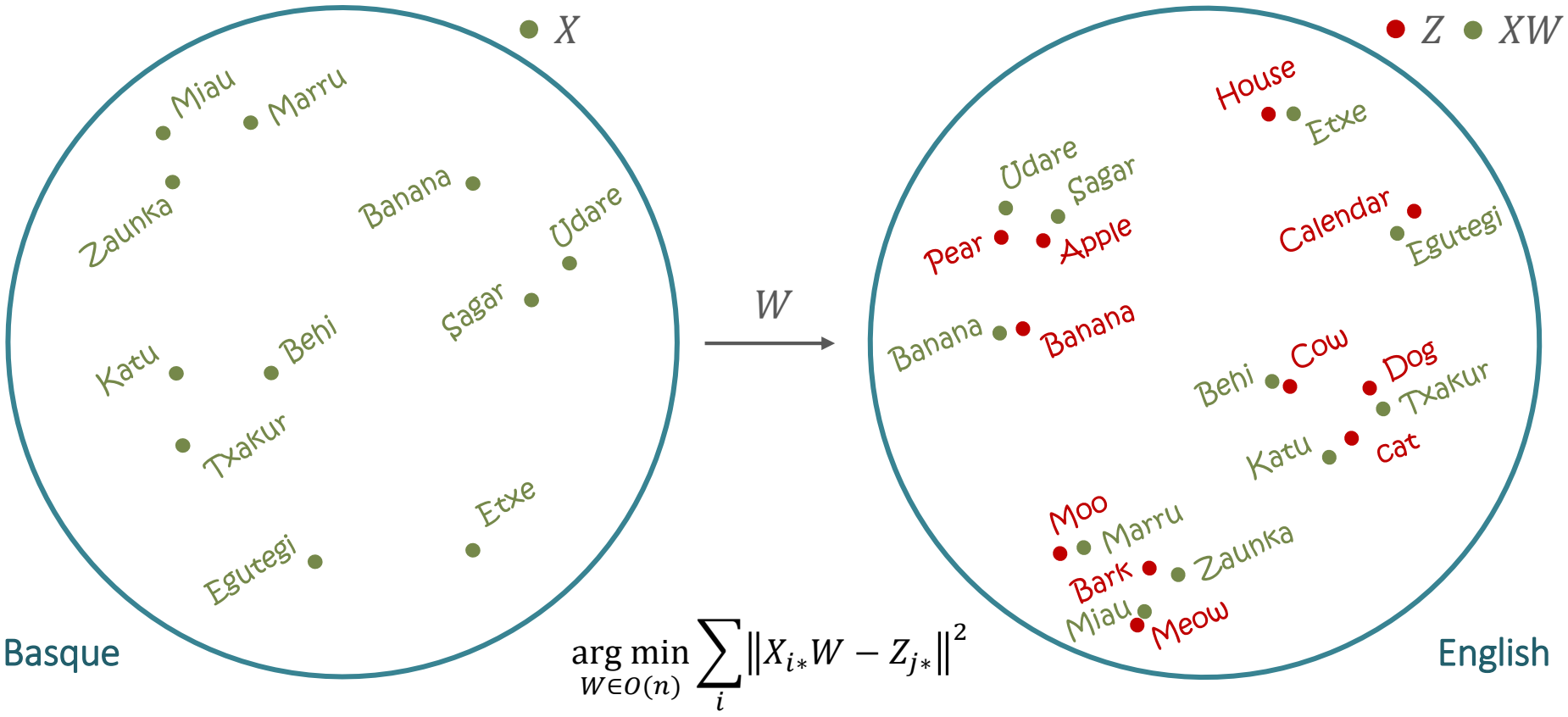
Training dictionary

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

Cross-lingual embedding mappings

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

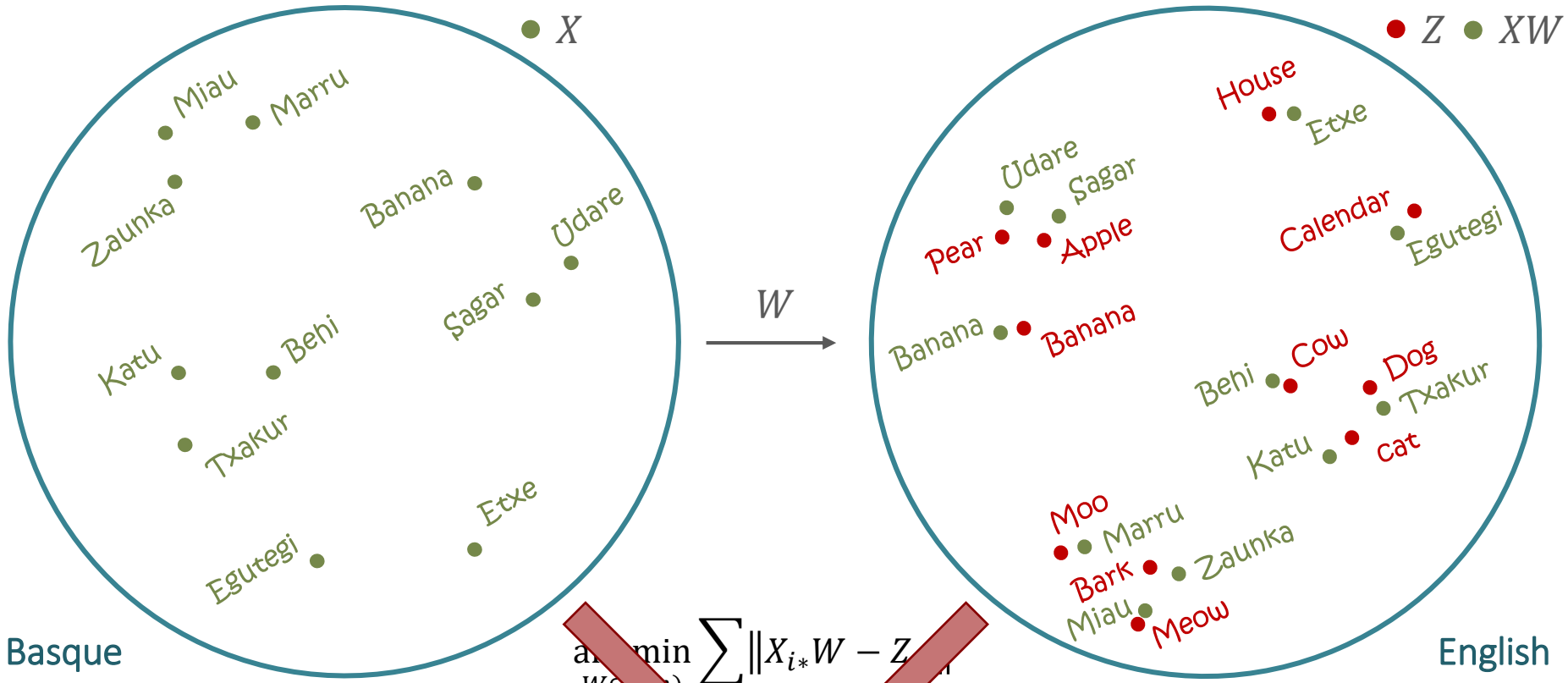
Cross-lingual embedding mappings



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

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

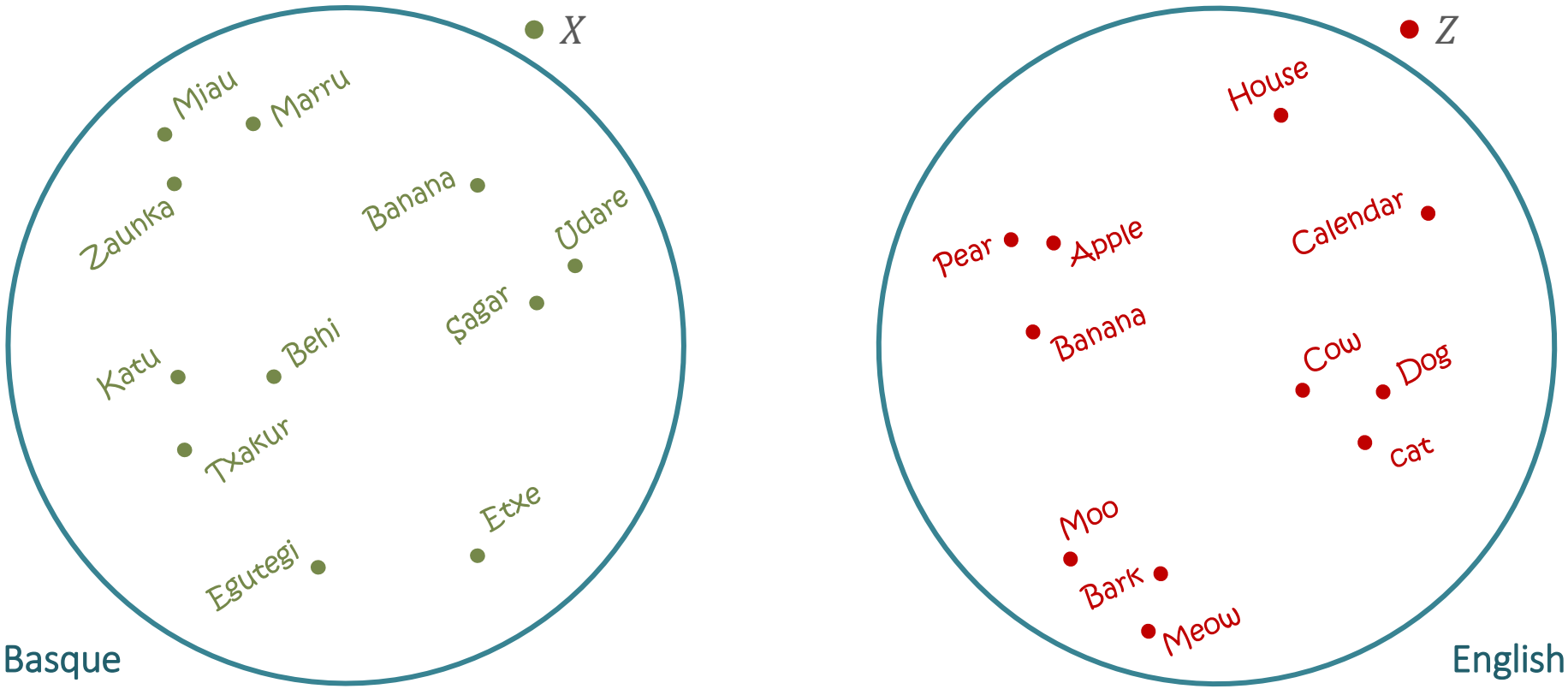
Cross-lingual embedding mappings



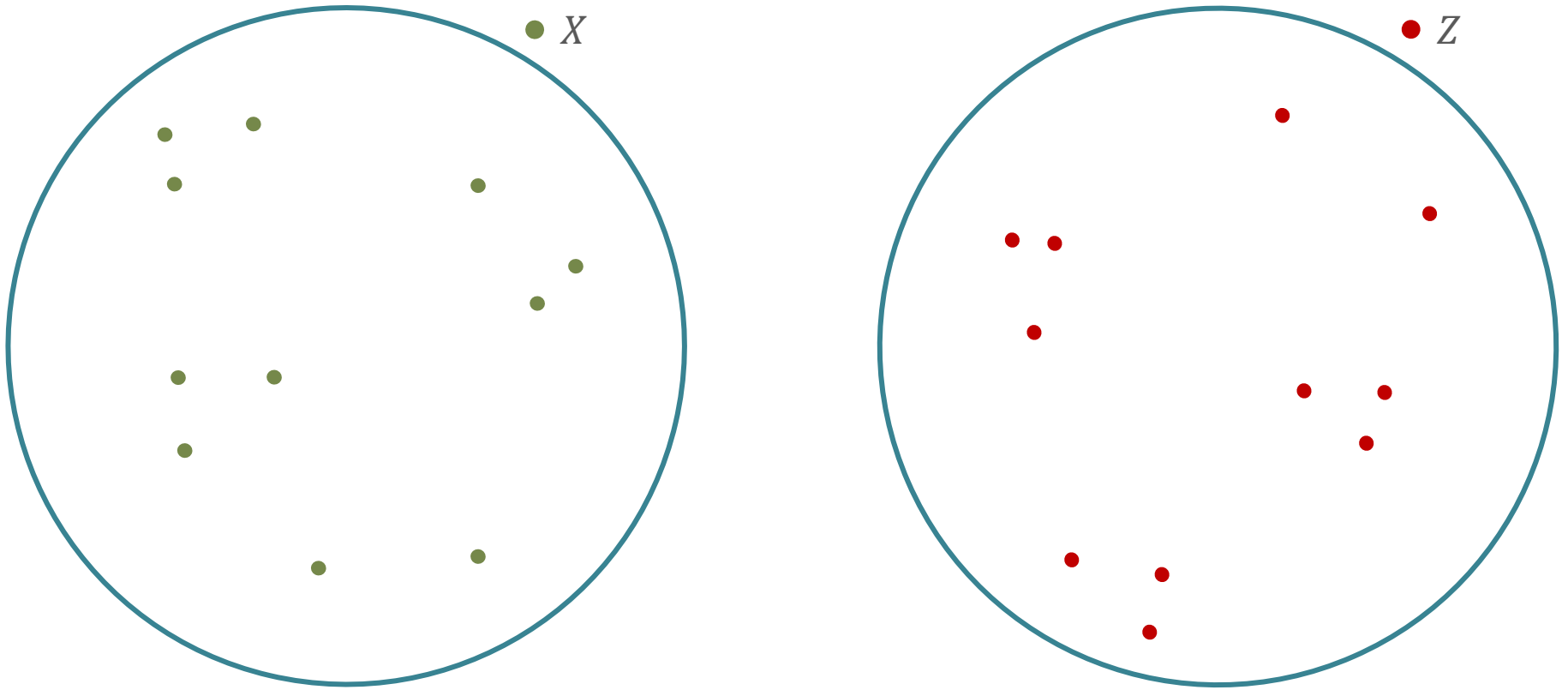
$$\min_{W} \sum_{i} \|X_{i*}W - Z_{i*}\|$$

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

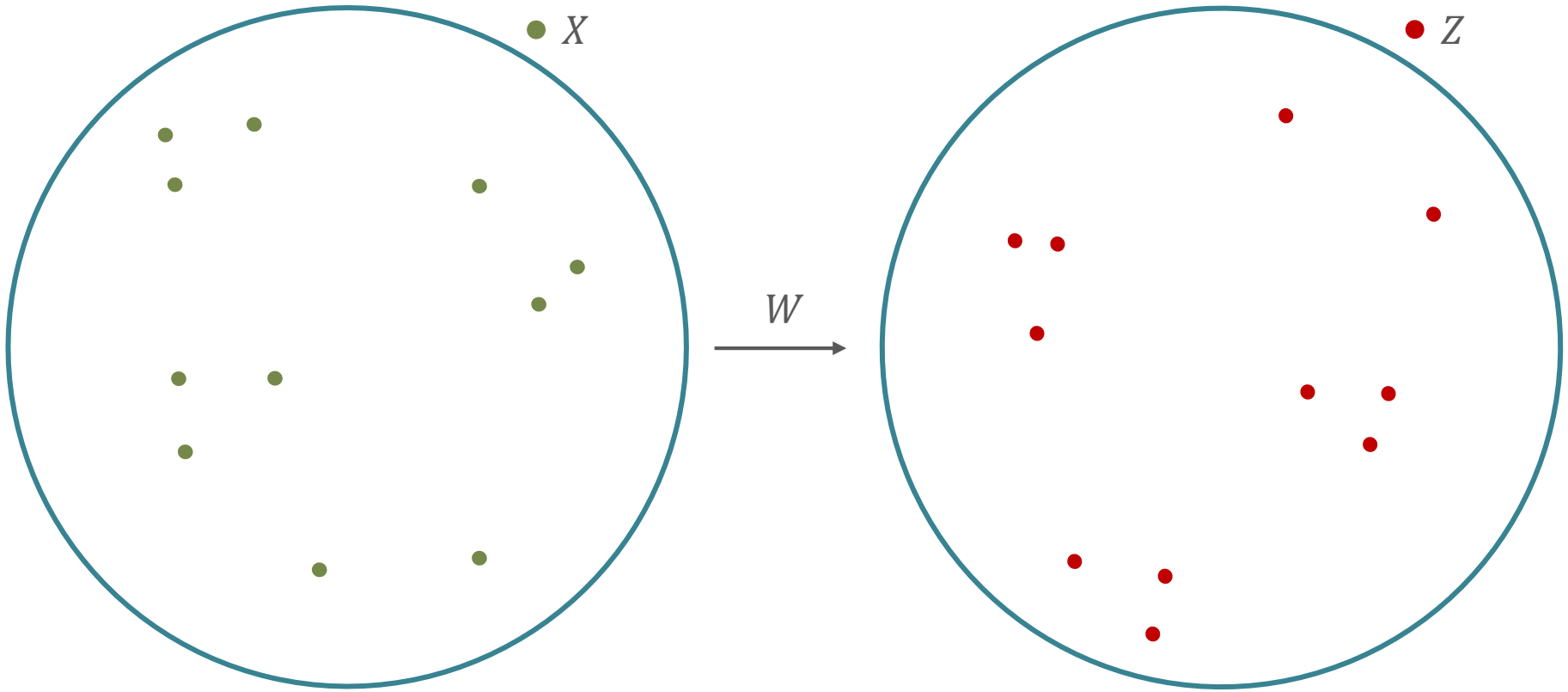
Cross-lingual embedding mappings



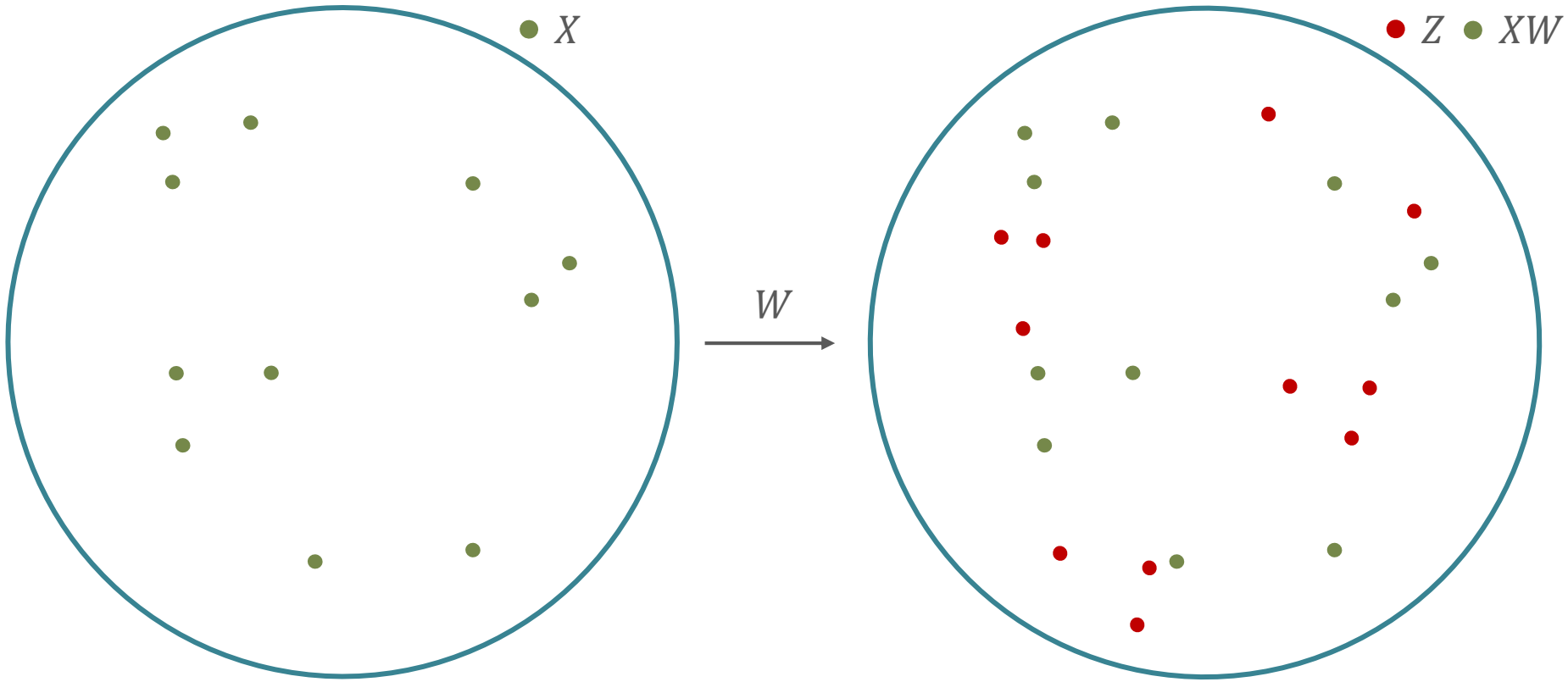
Cross-lingual embedding mappings



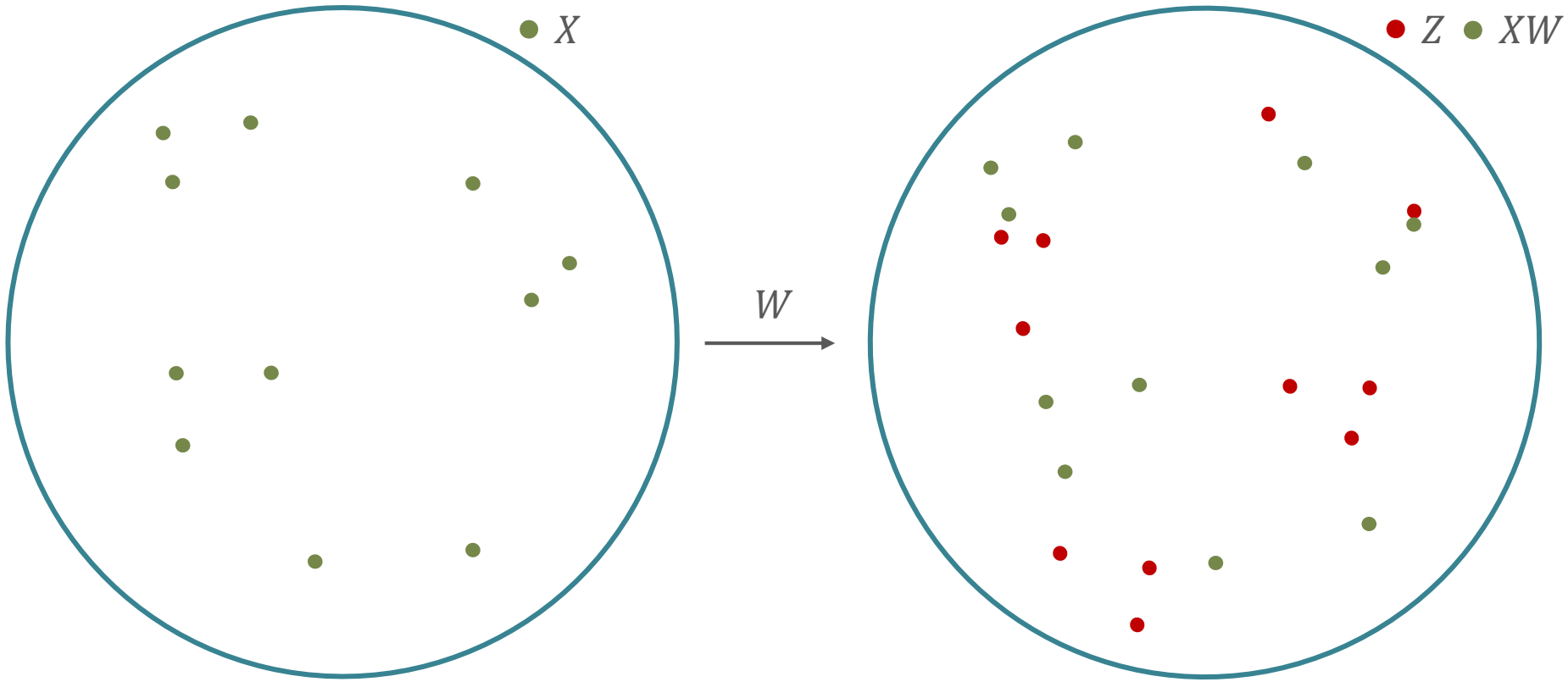
Cross-lingual embedding mappings



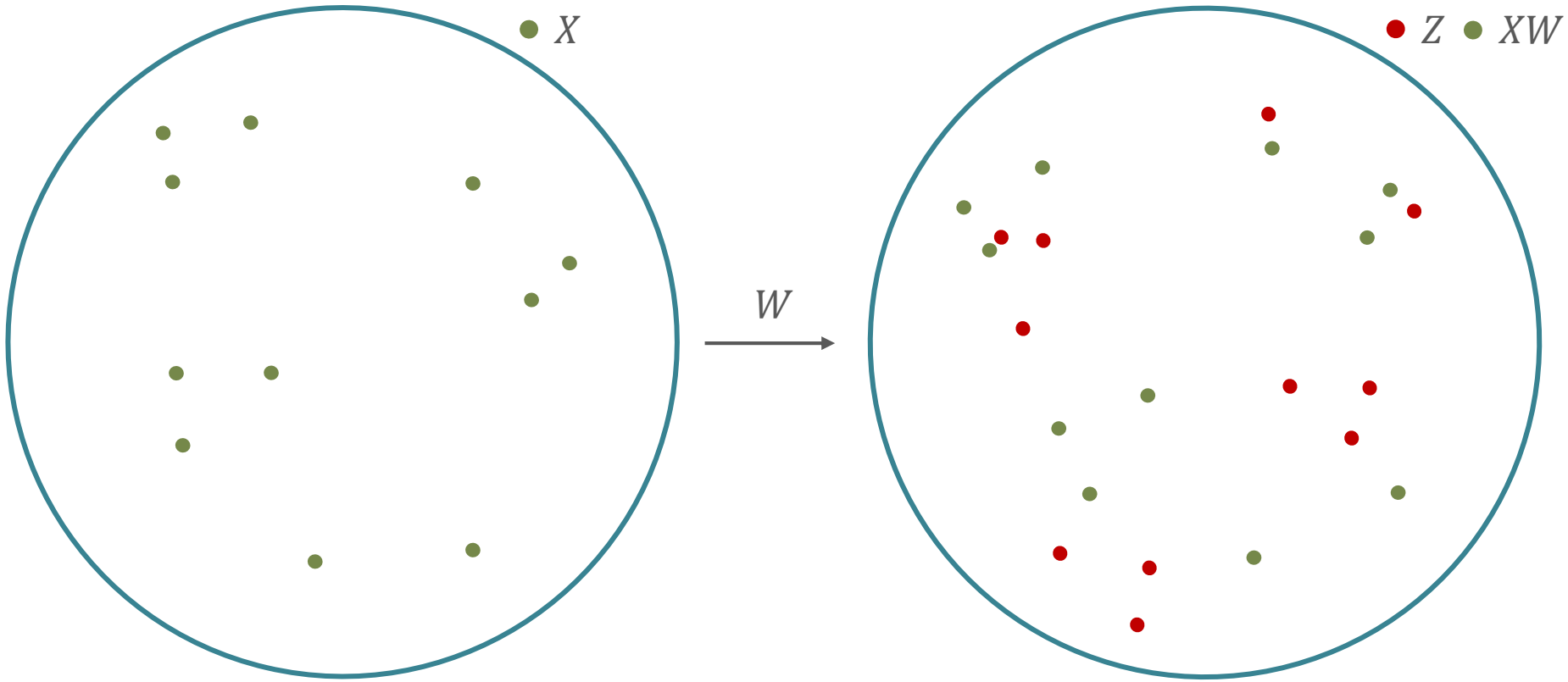
Cross-lingual embedding mappings



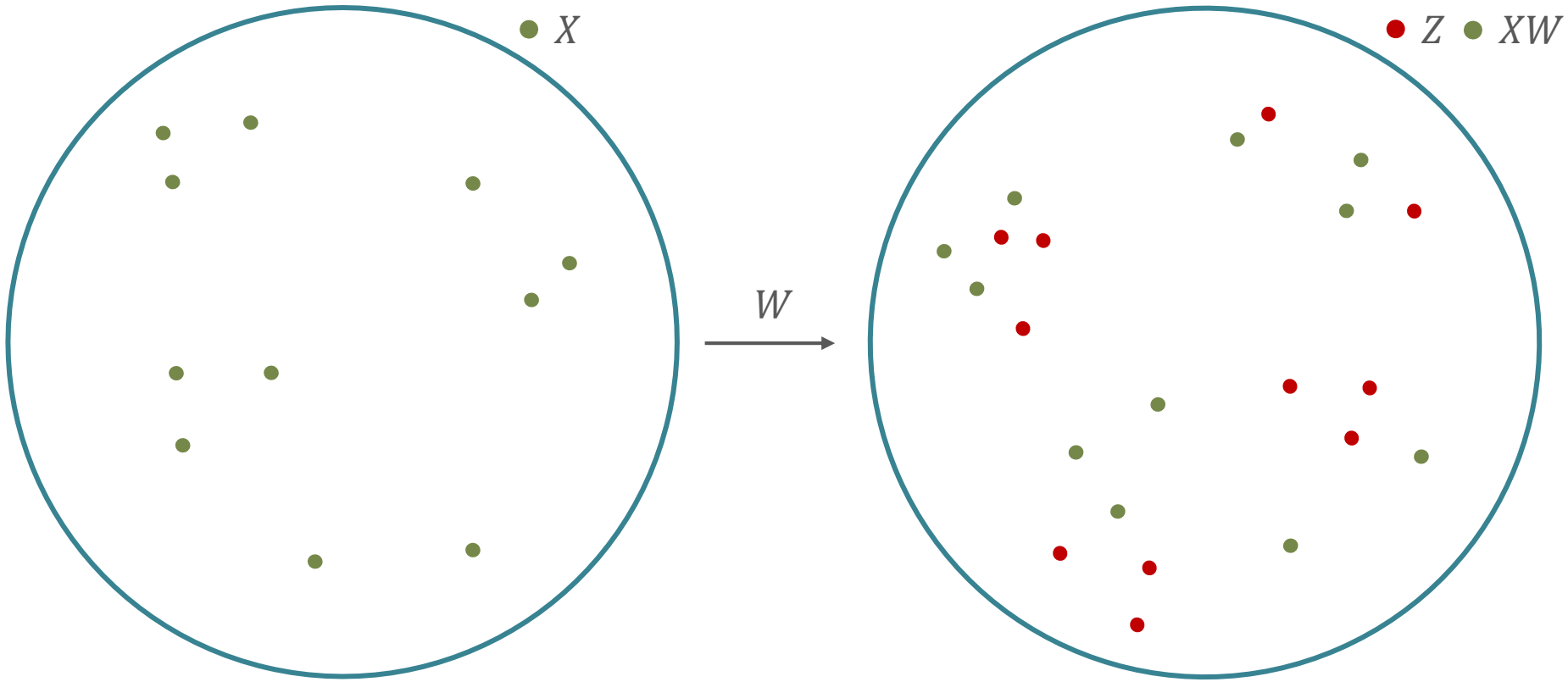
Cross-lingual embedding mappings



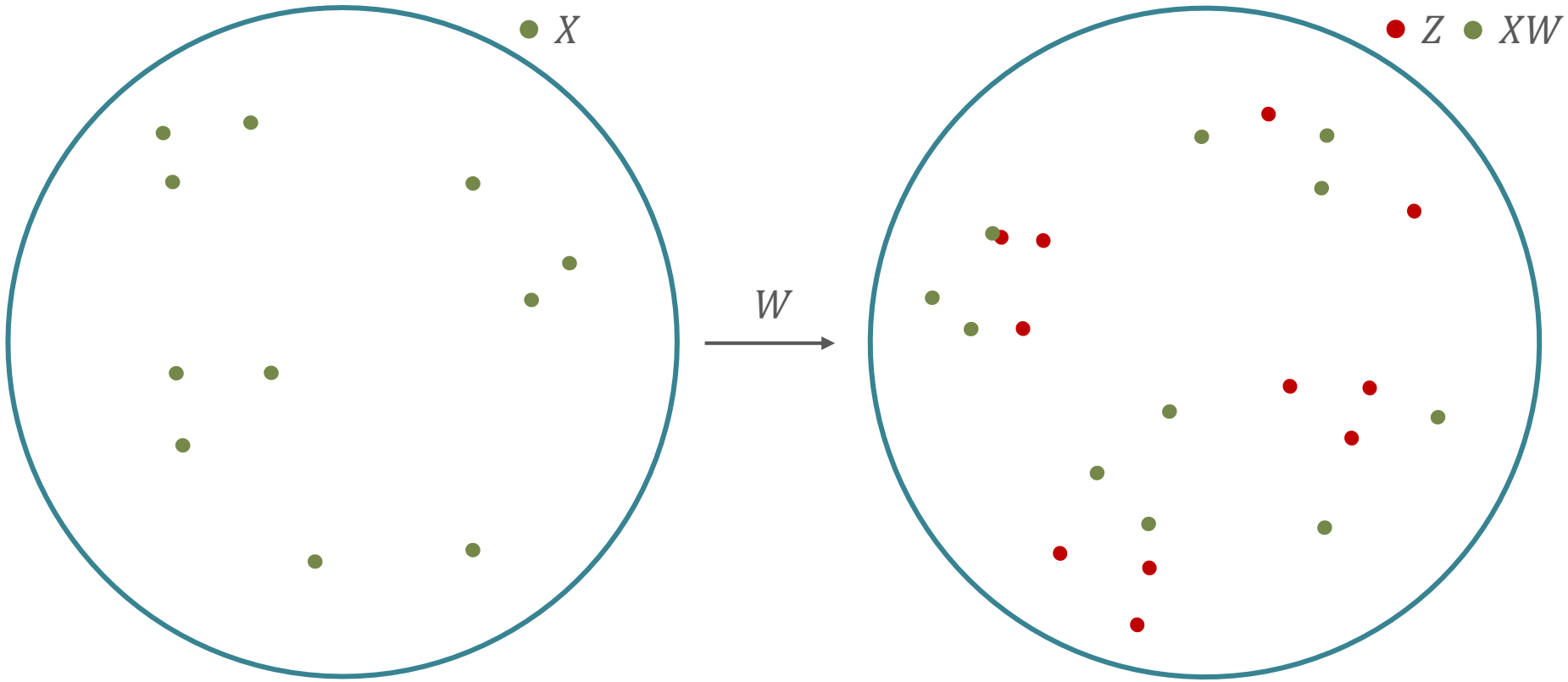
Cross-lingual embedding mappings



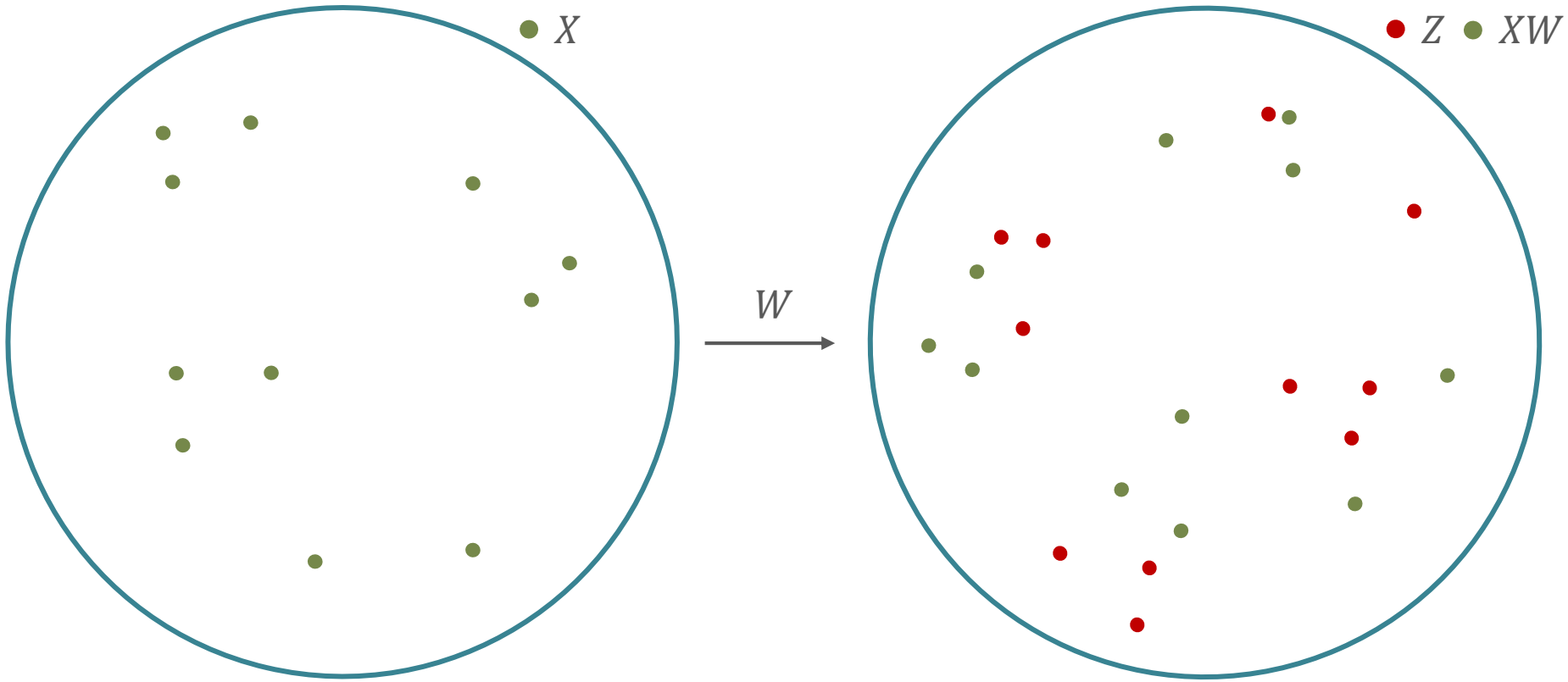
Cross-lingual embedding mappings



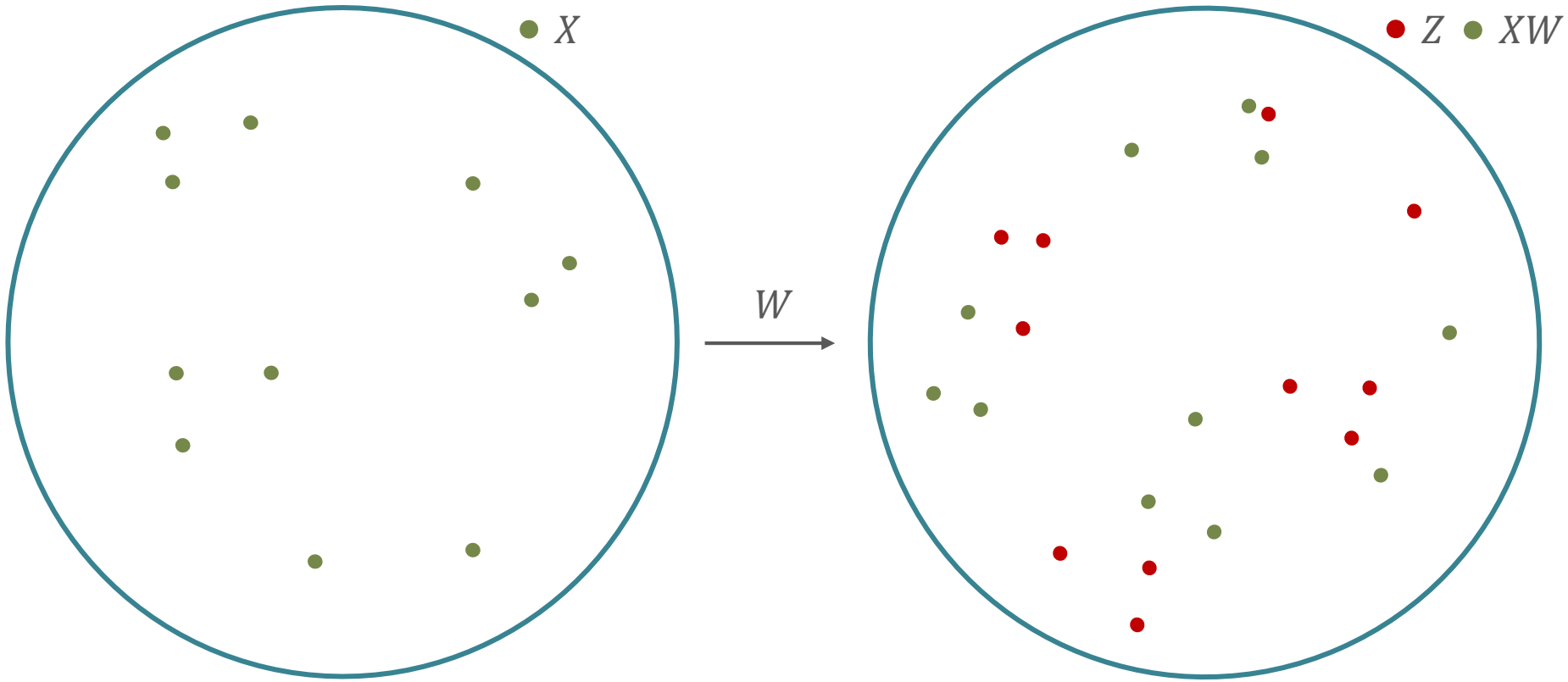
Cross-lingual embedding mappings



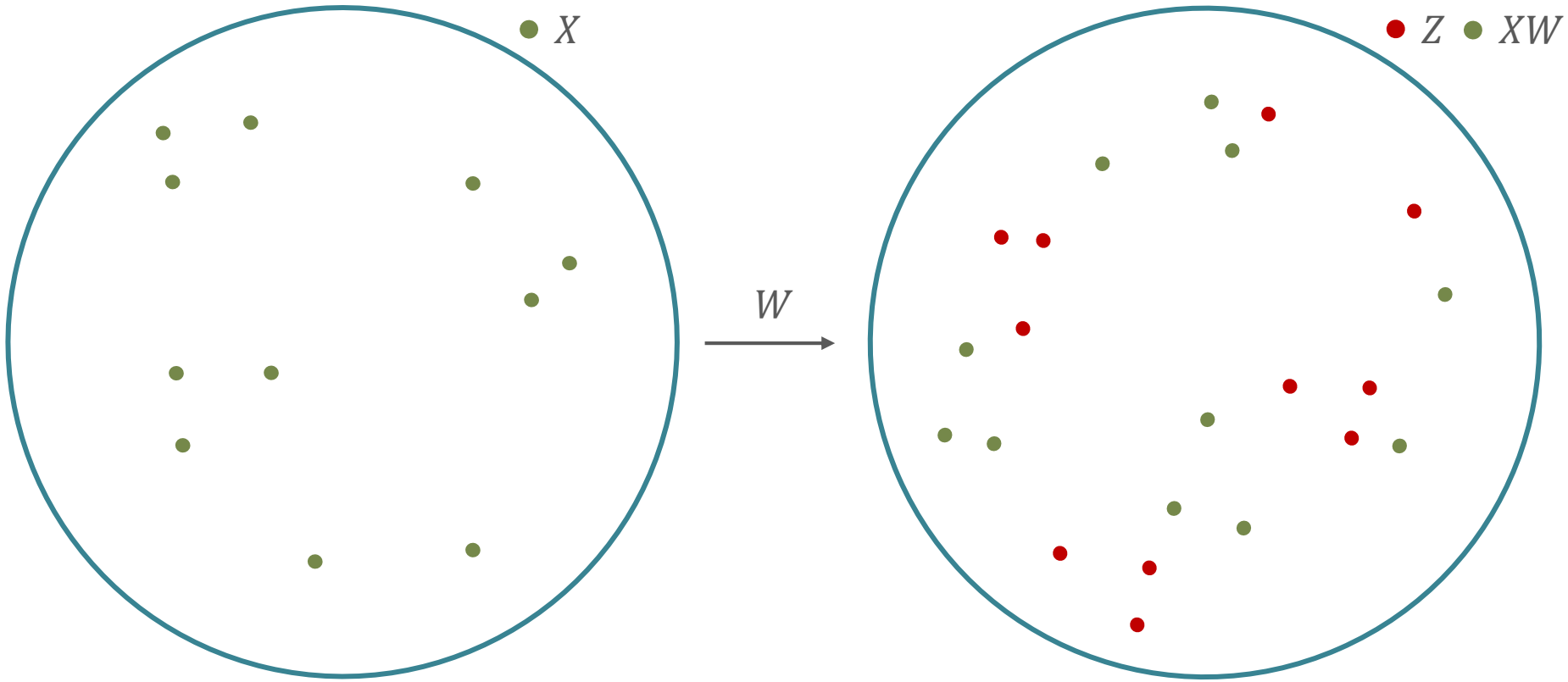
Cross-lingual embedding mappings



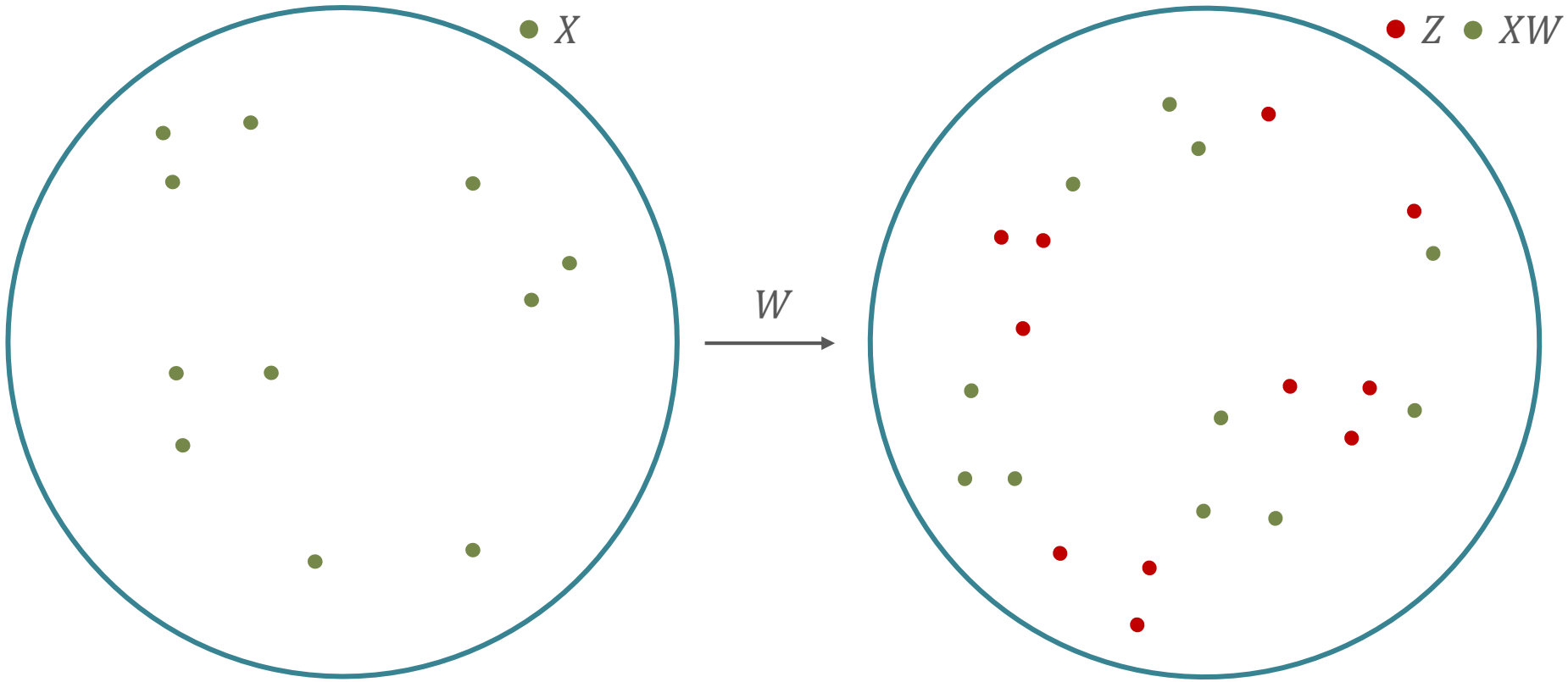
Cross-lingual embedding mappings



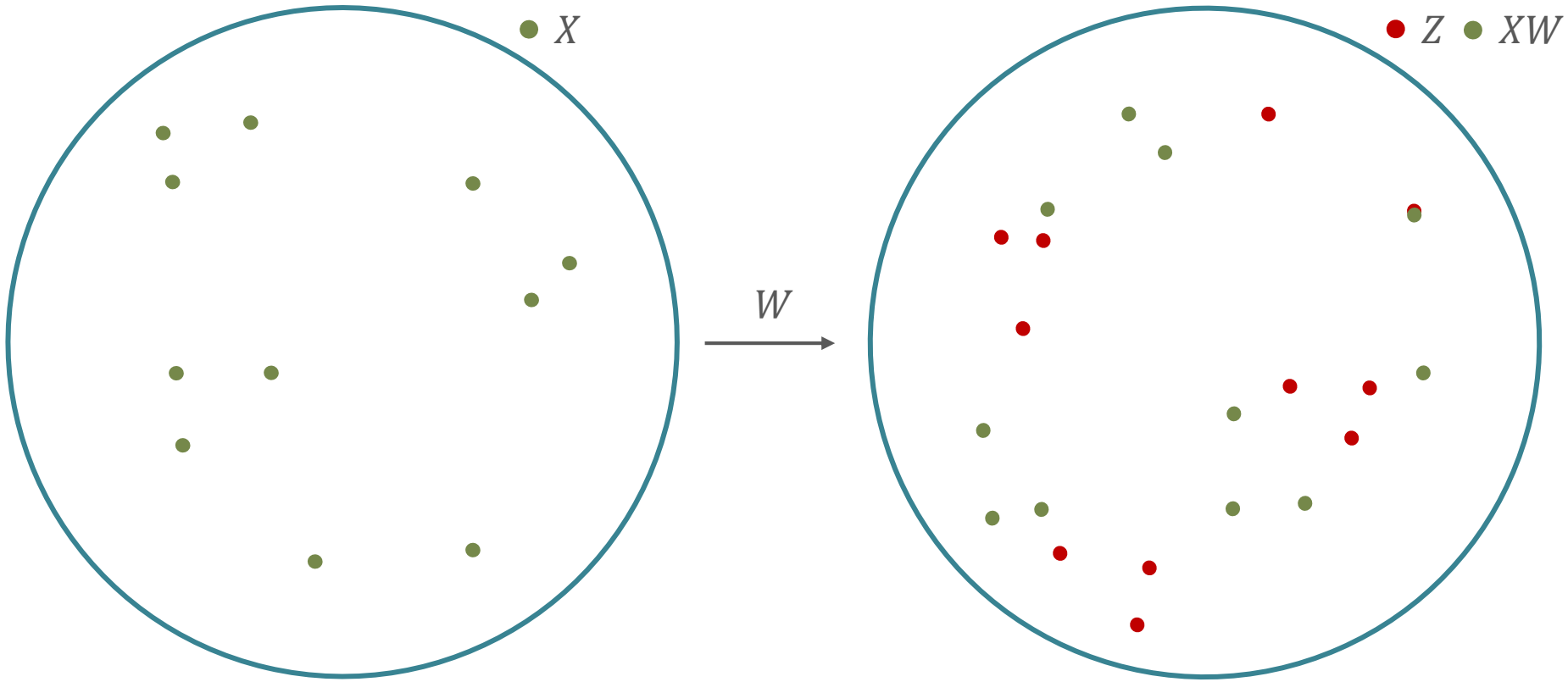
Cross-lingual embedding mappings



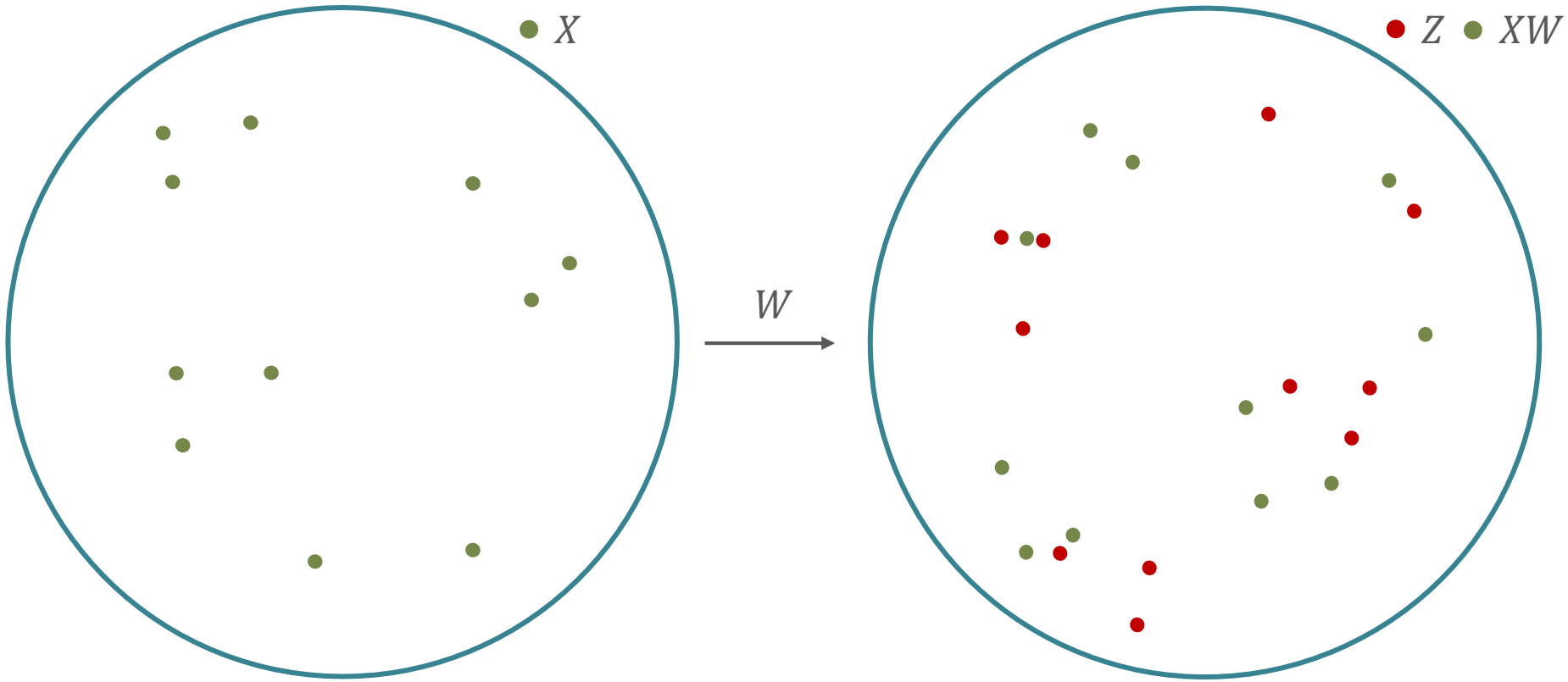
Cross-lingual embedding mappings



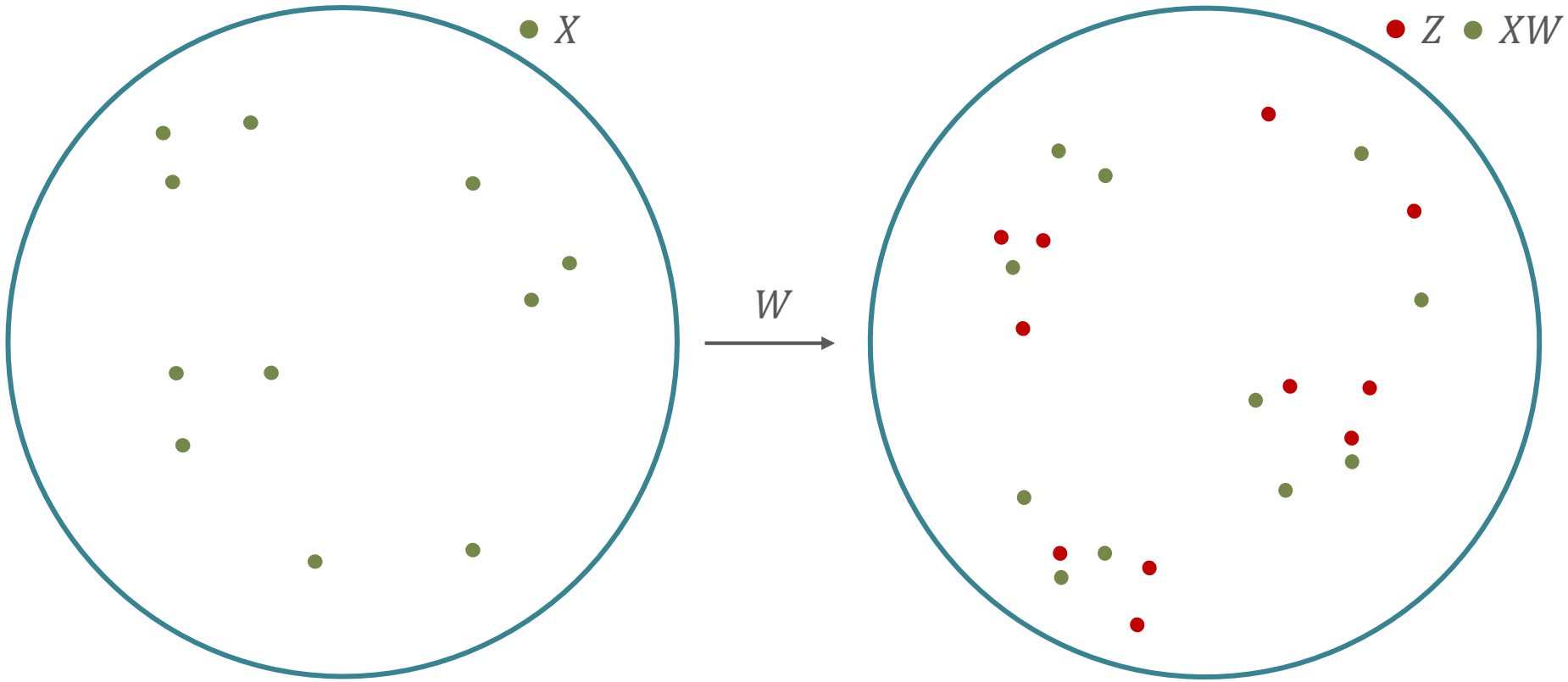
Cross-lingual embedding mappings



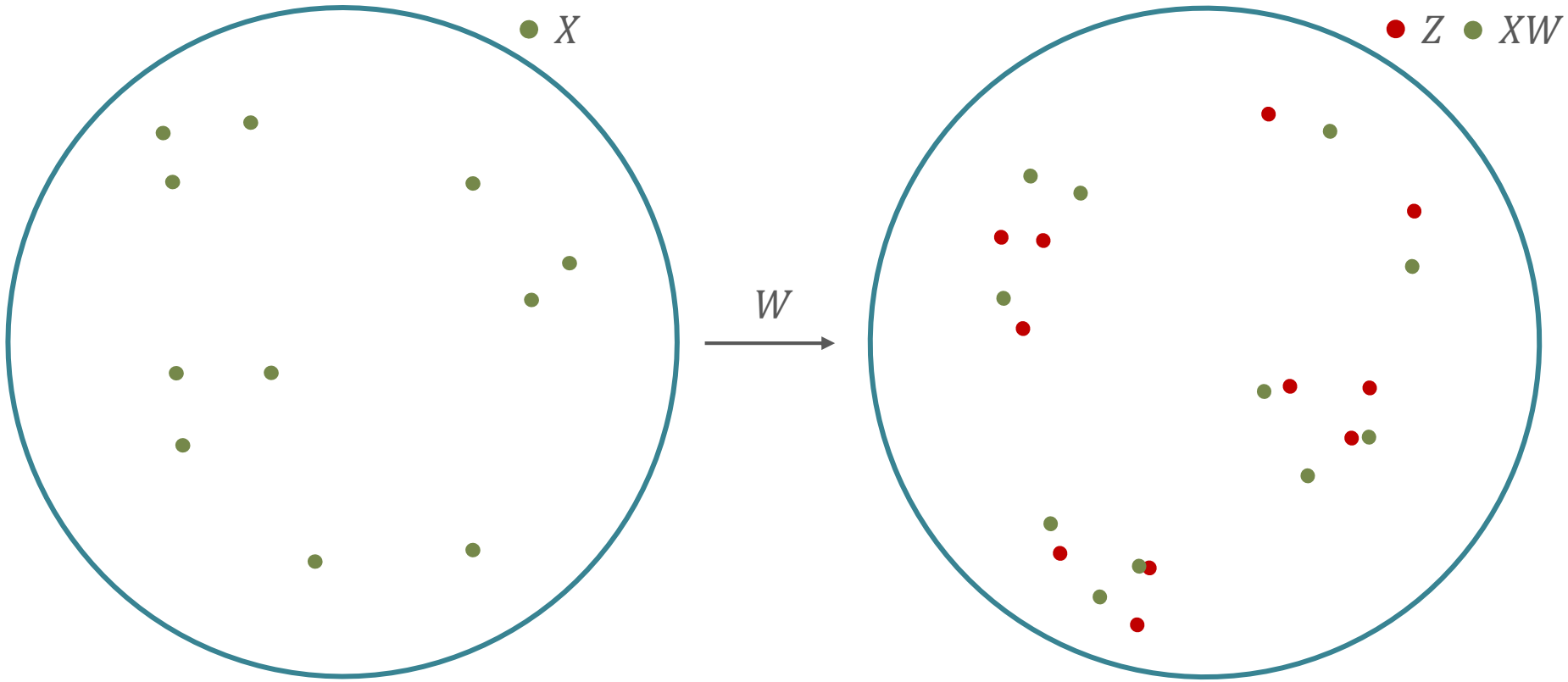
Cross-lingual embedding mappings



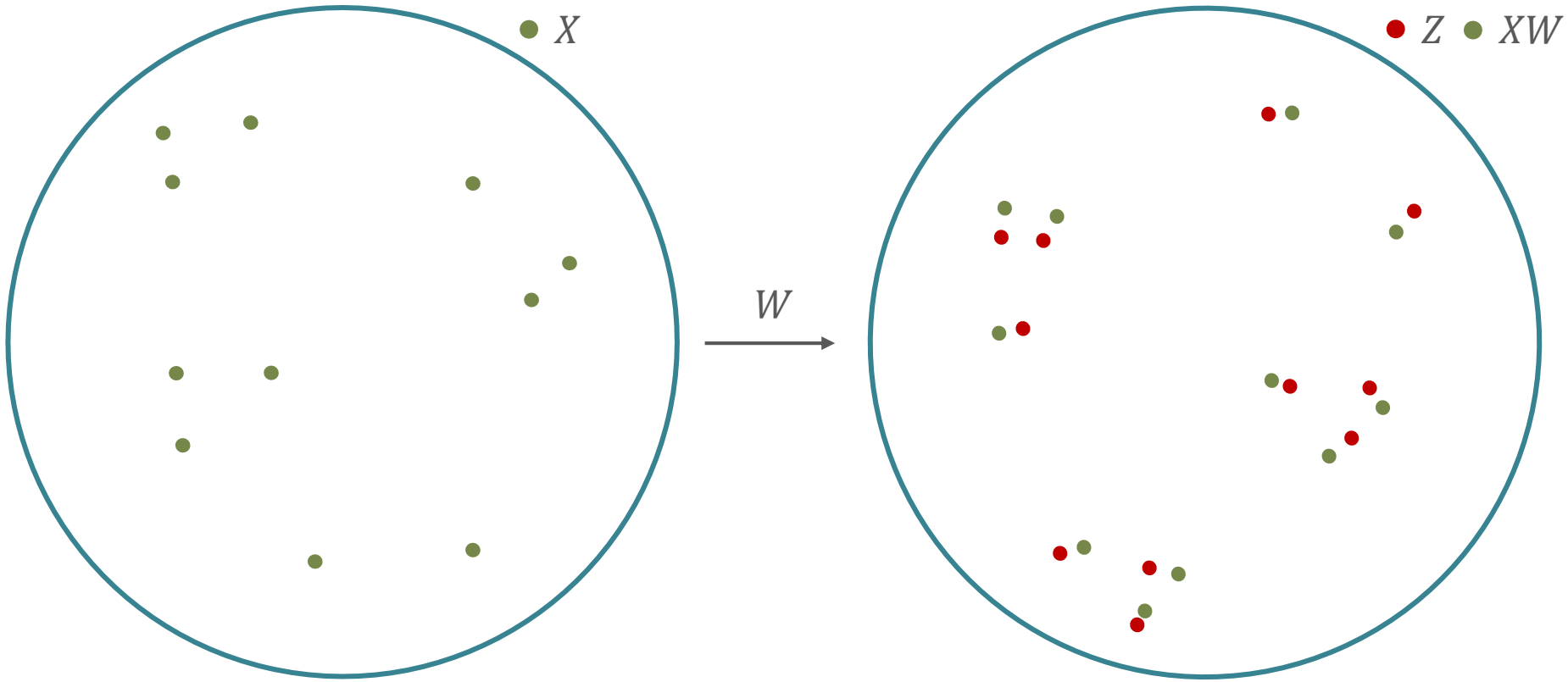
Cross-lingual embedding mappings



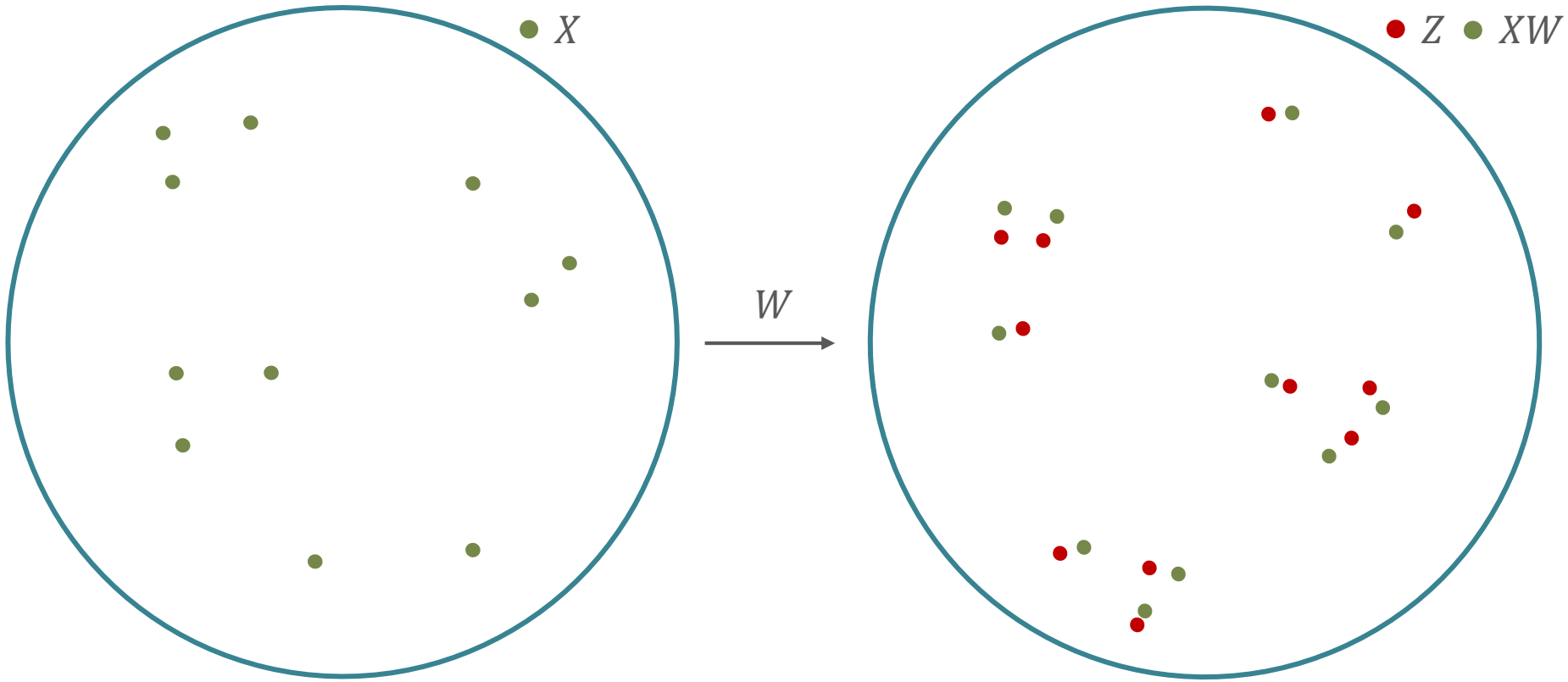
Cross-lingual embedding mappings



Cross-lingual embedding mappings



Cross-lingual embedding mappings



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

Cross-lingual embedding mappings

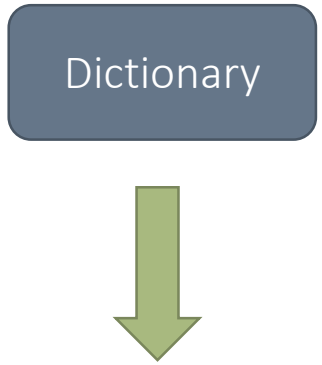
$$W^* = \arg\min_{W \in O(n)} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual embedding mappings

Dictionary

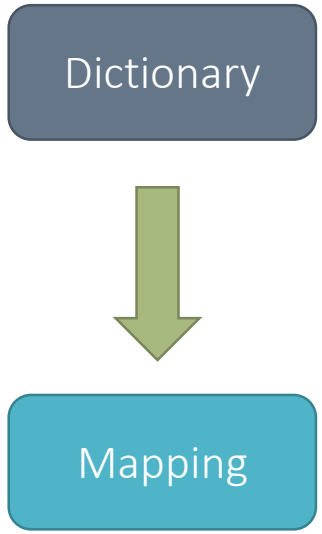
$$W^* = \arg\min_{W \in O(n)} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual embedding mappings



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

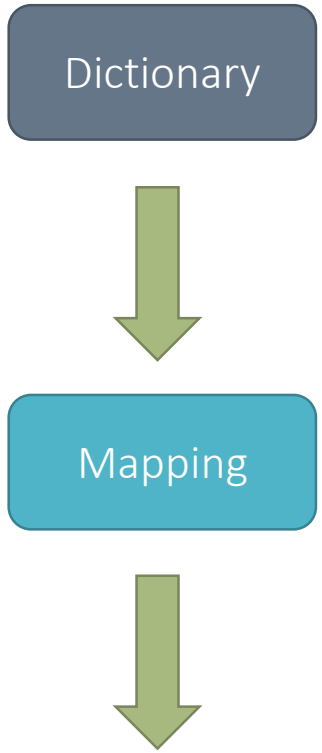
Cross-lingual embedding mappings



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

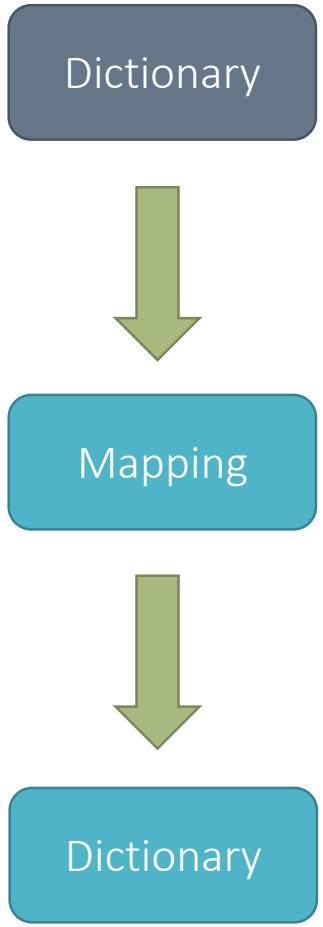
Cross-lingual embedding mappings

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



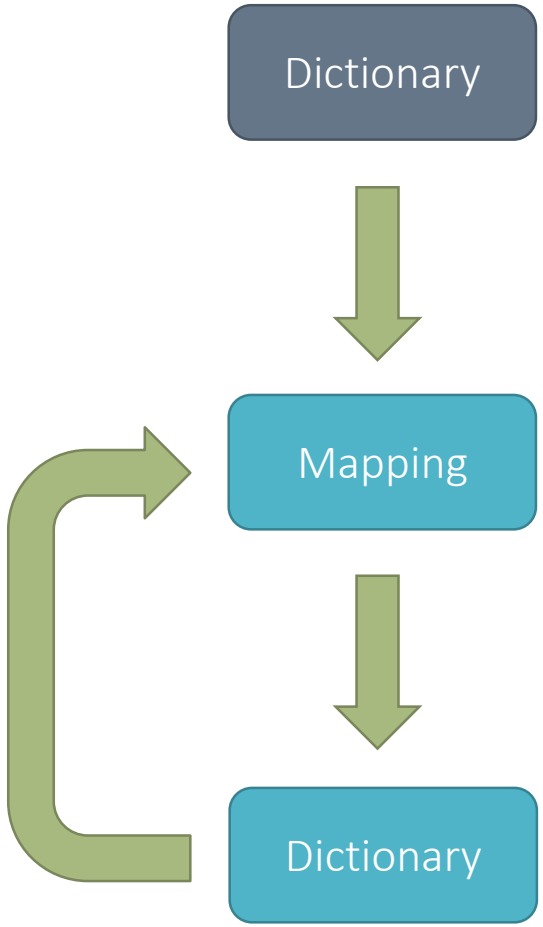
Cross-lingual embedding mappings

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



Cross-lingual embedding mappings

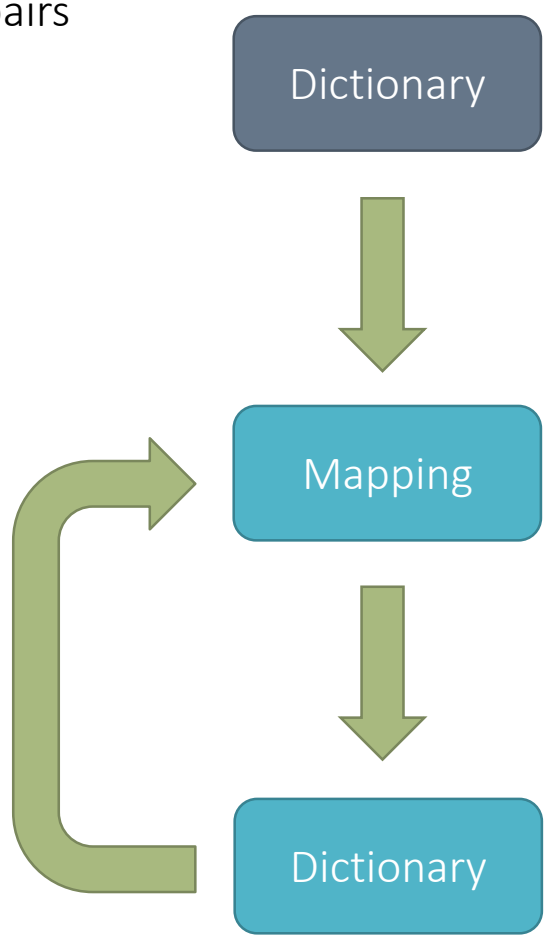
$$W^* = \arg\min_{W \in O(n)} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$



Cross-lingual embedding mappings

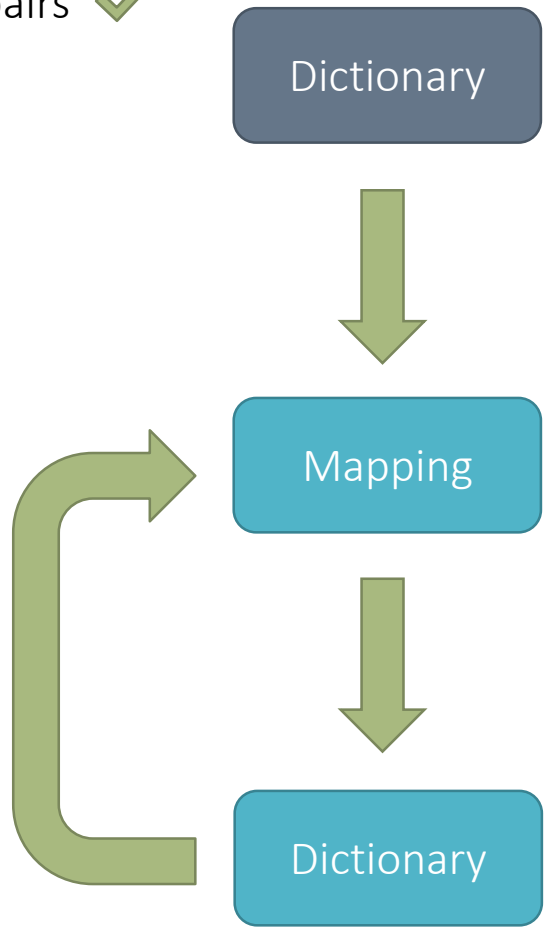
- 25 word pairs

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



Cross-lingual embedding mappings

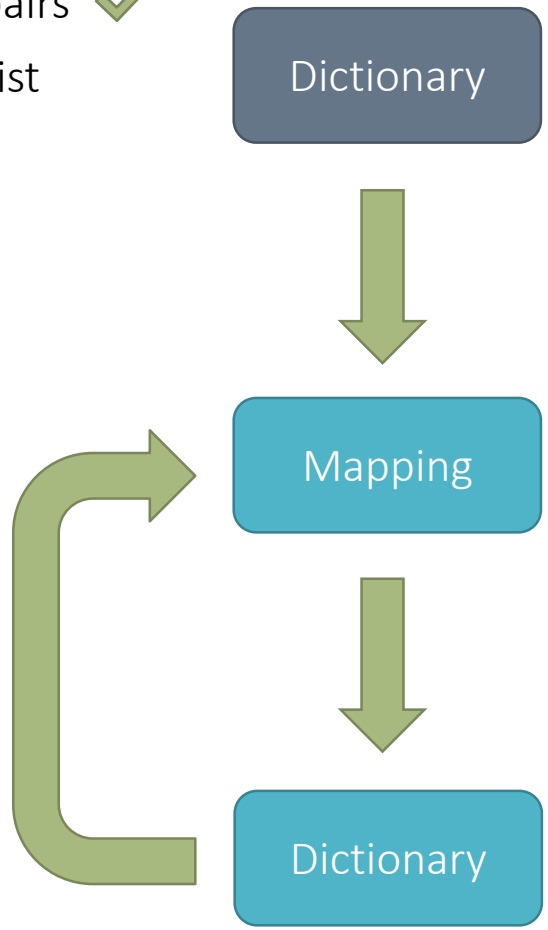
- 25 word pairs ✓



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

Cross-lingual embedding mappings

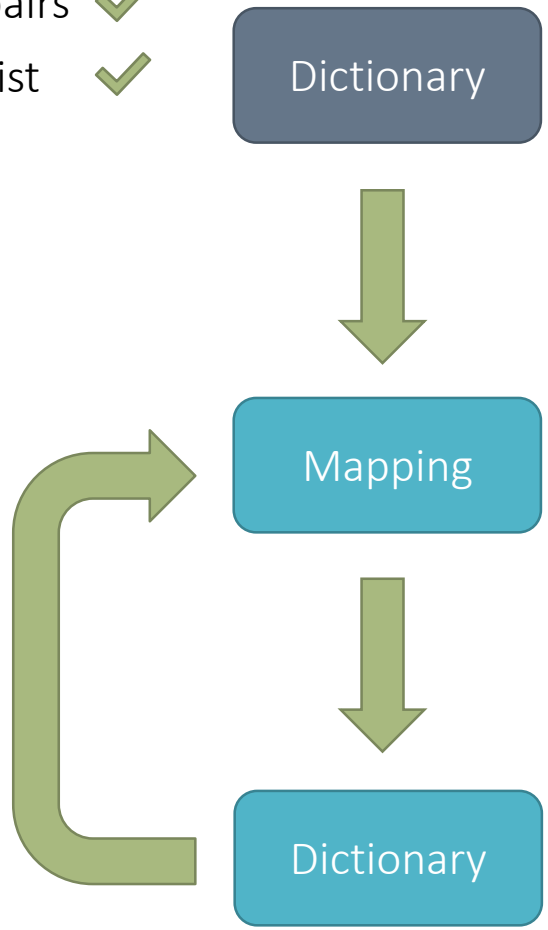
- 25 word pairs ✓
- Numeral list



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

Cross-lingual embedding mappings

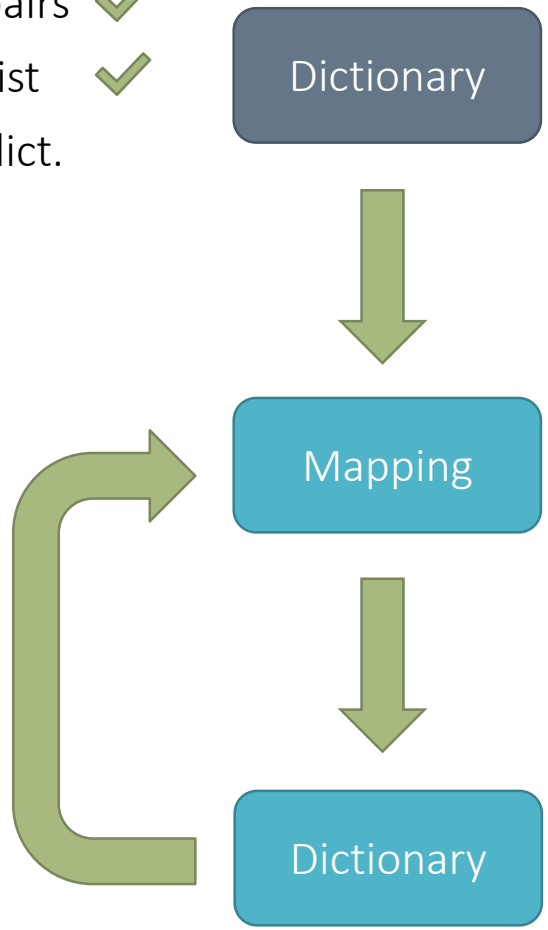
- 25 word pairs ✓
- Numeral list ✓



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

Cross-lingual embedding mappings

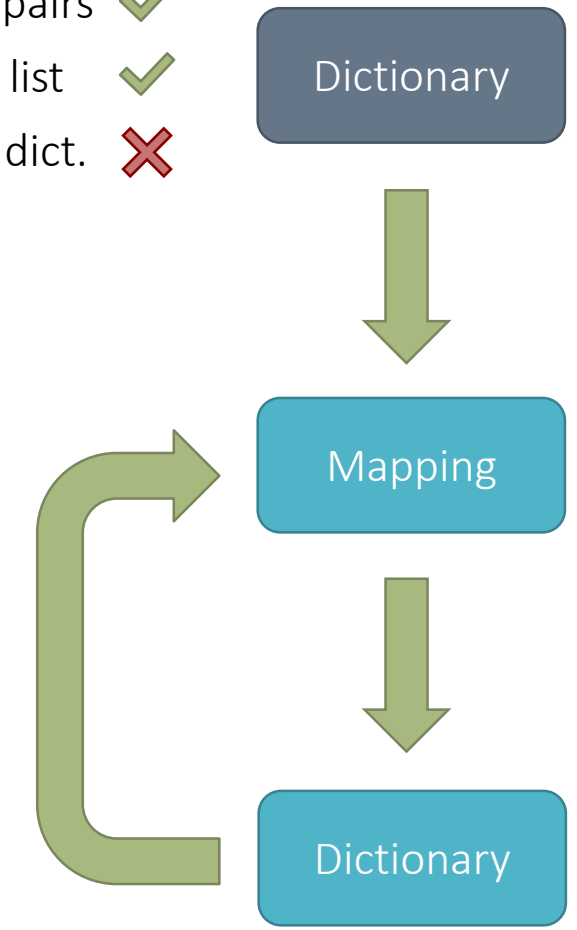
- 25 word pairs ✓
- Numeral list ✓
- Random dict.



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

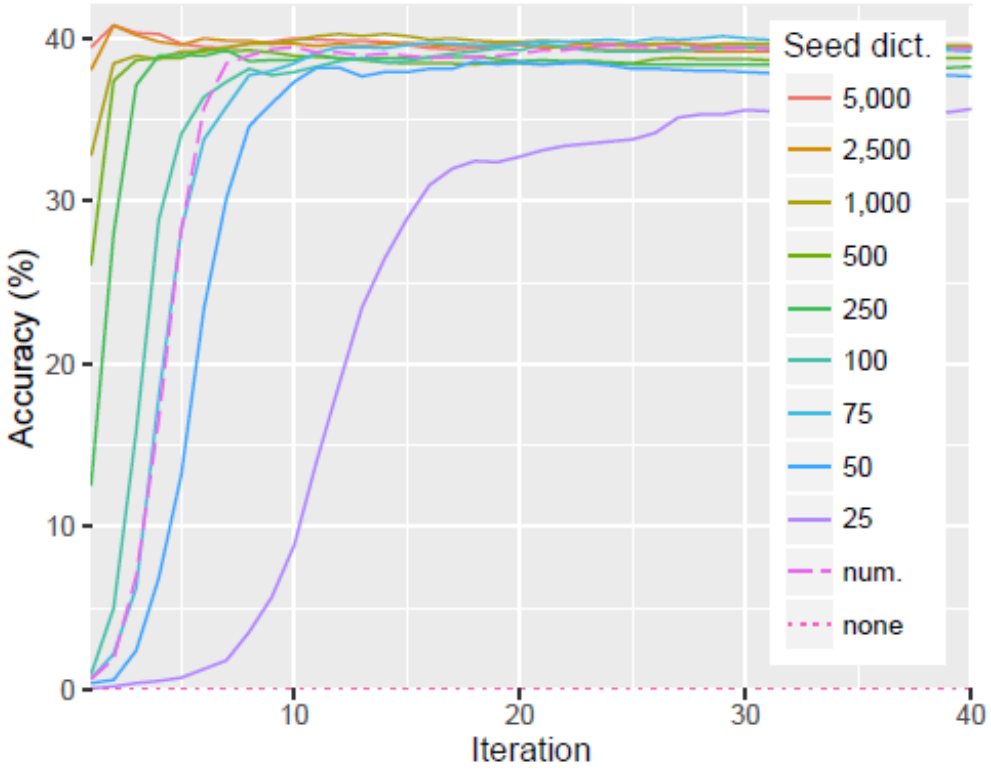
Cross-lingual embedding mappings

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

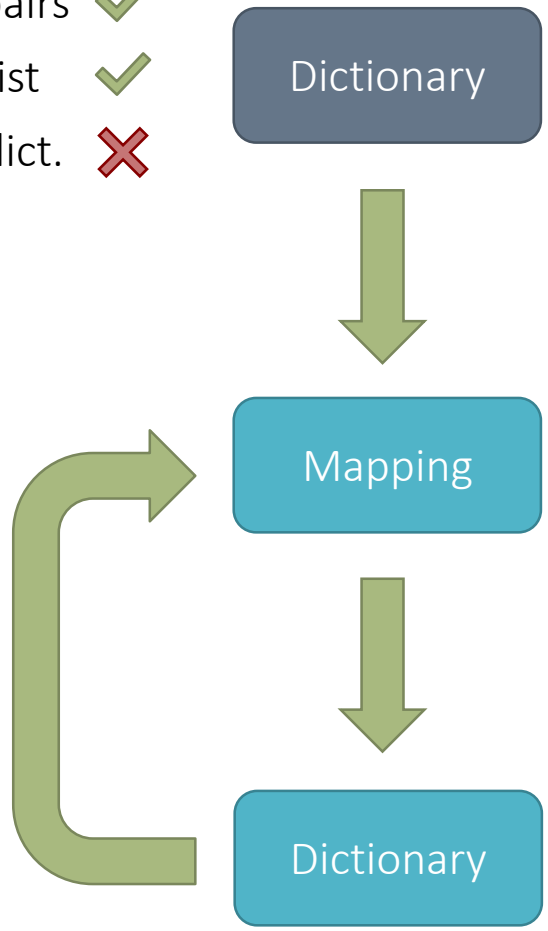


$$W^* = \arg\min_{W \in O(n)} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual embedding mappings

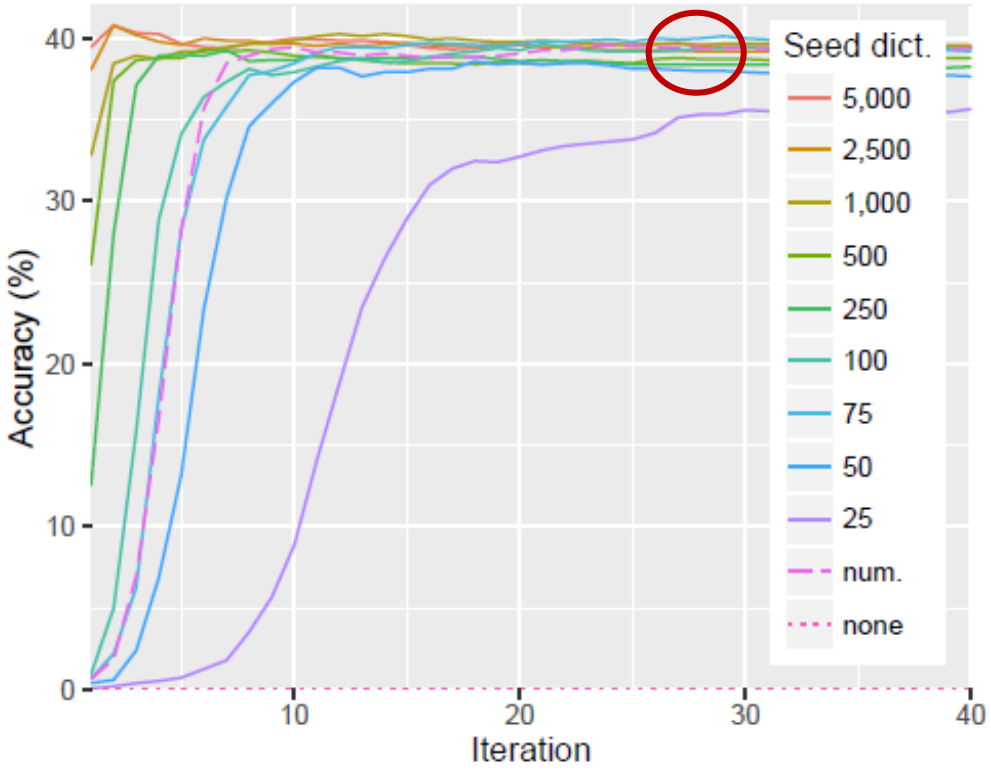


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

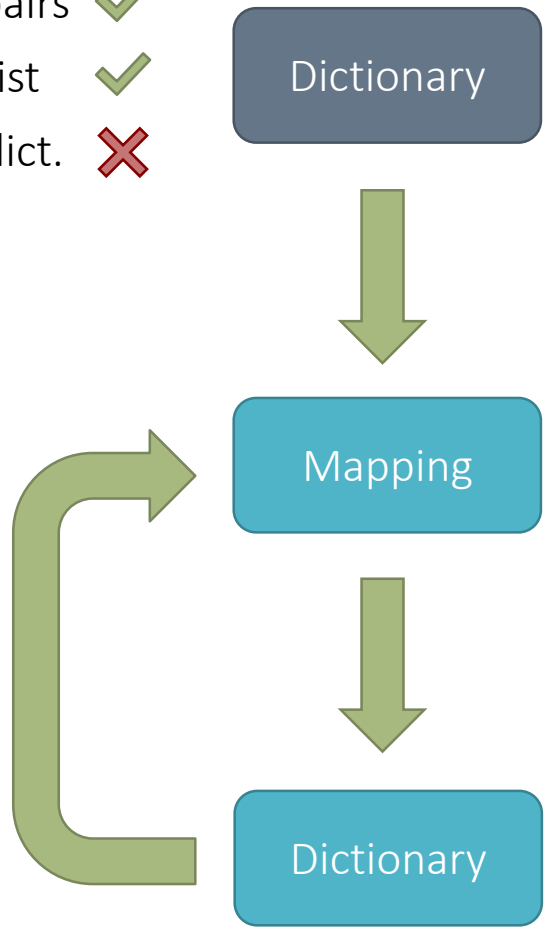


$$W^* = \arg\min_{W \in O(n)} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual embedding mappings

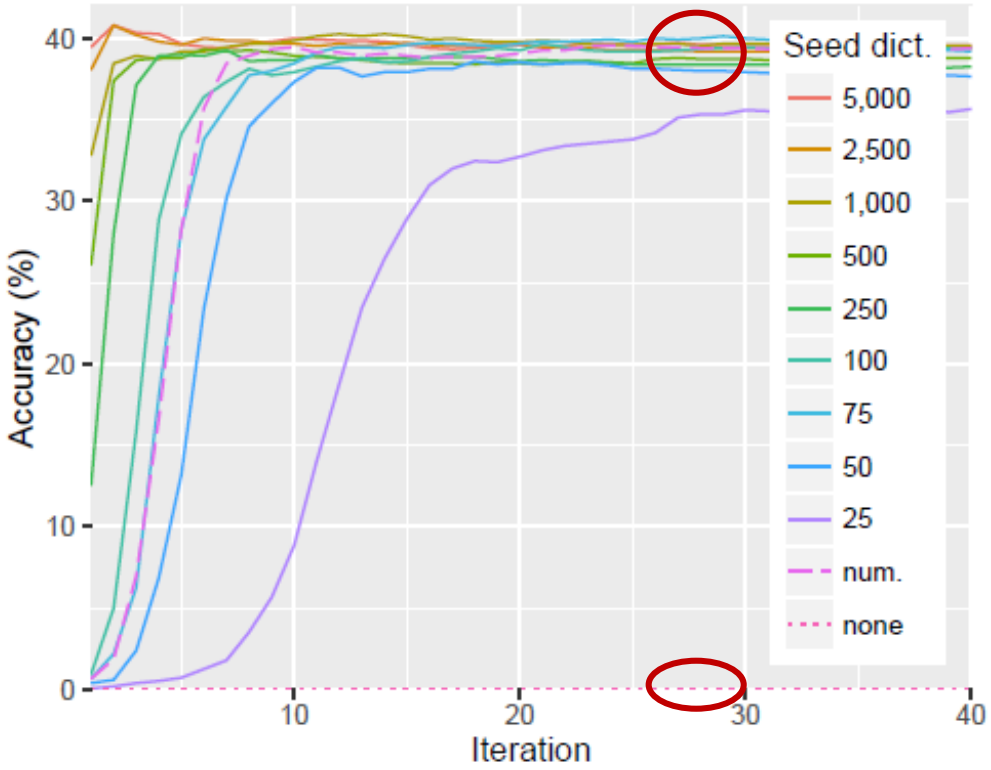


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

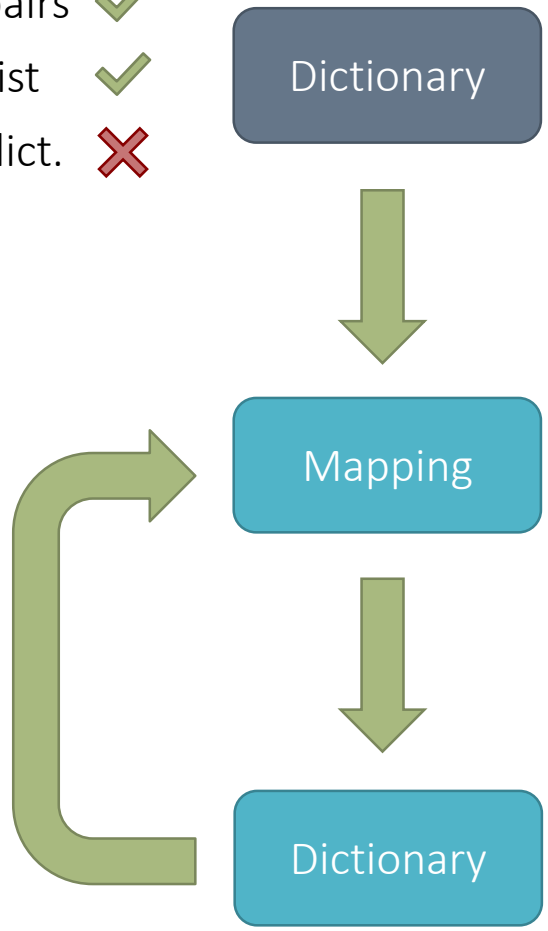


$$W^* = \arg\min_{W \in O(n)} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual embedding mappings

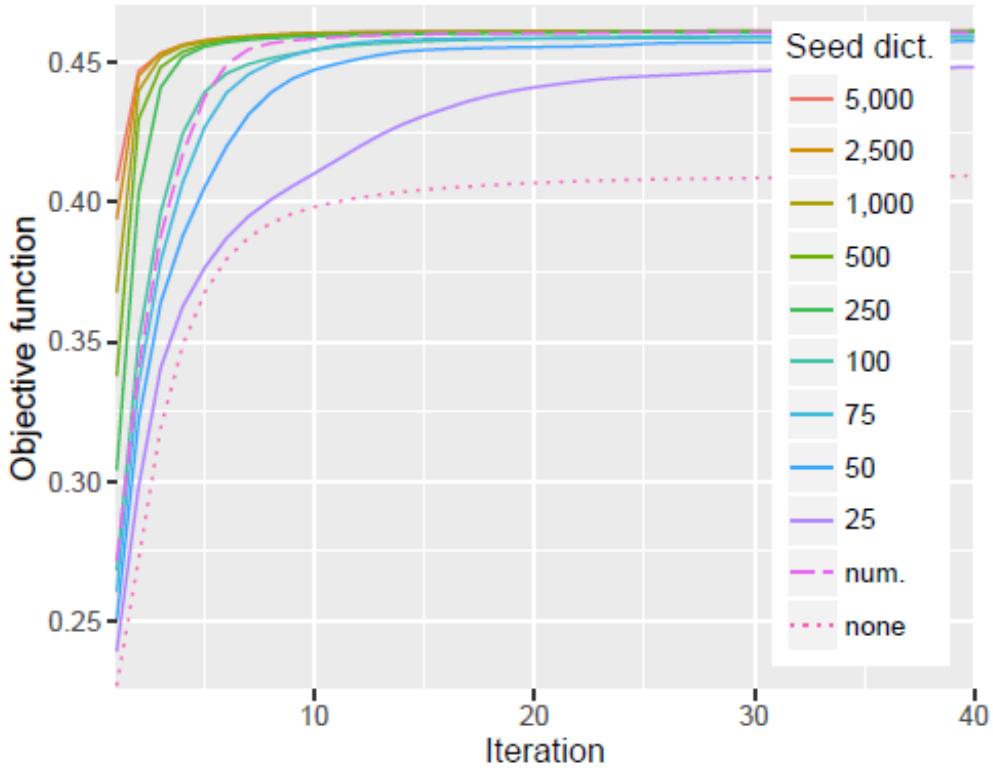


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

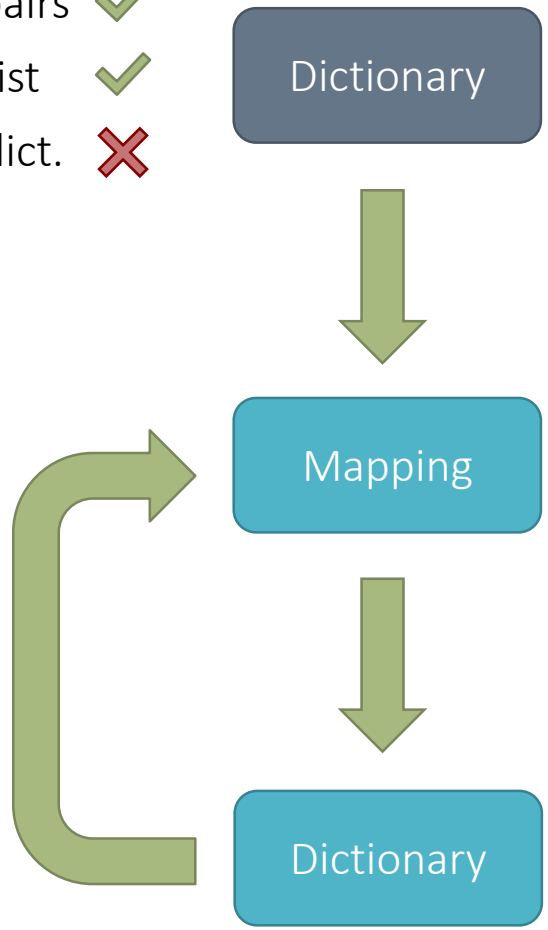


$$W^* = \arg\min_{W \in O(n)} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual embedding mappings

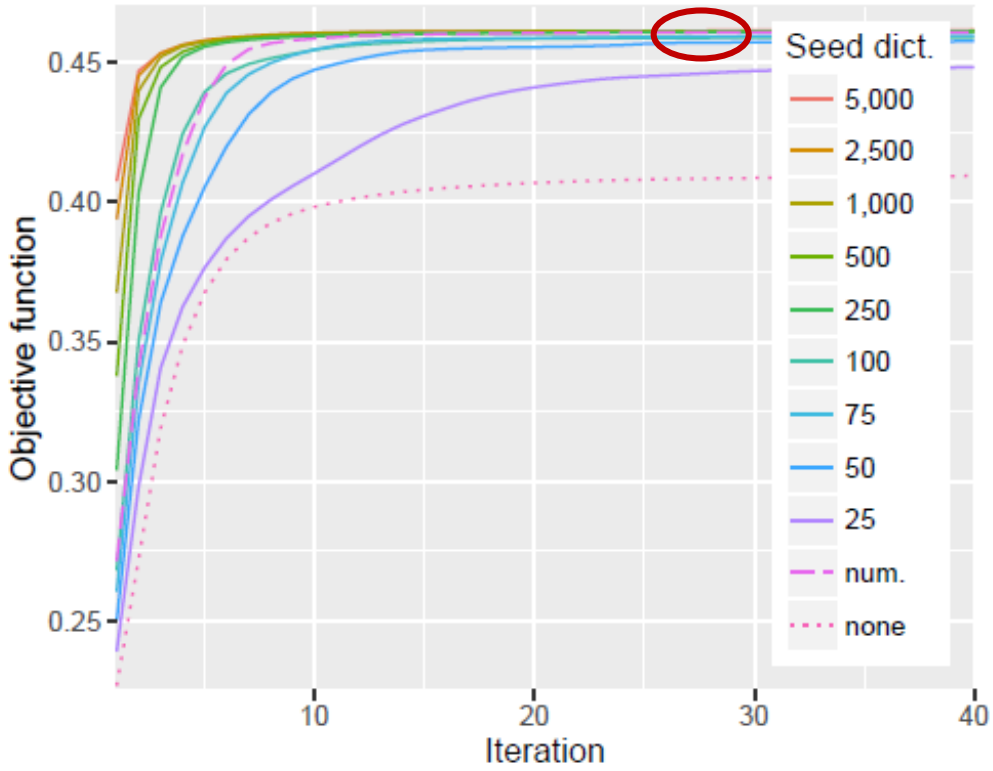


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

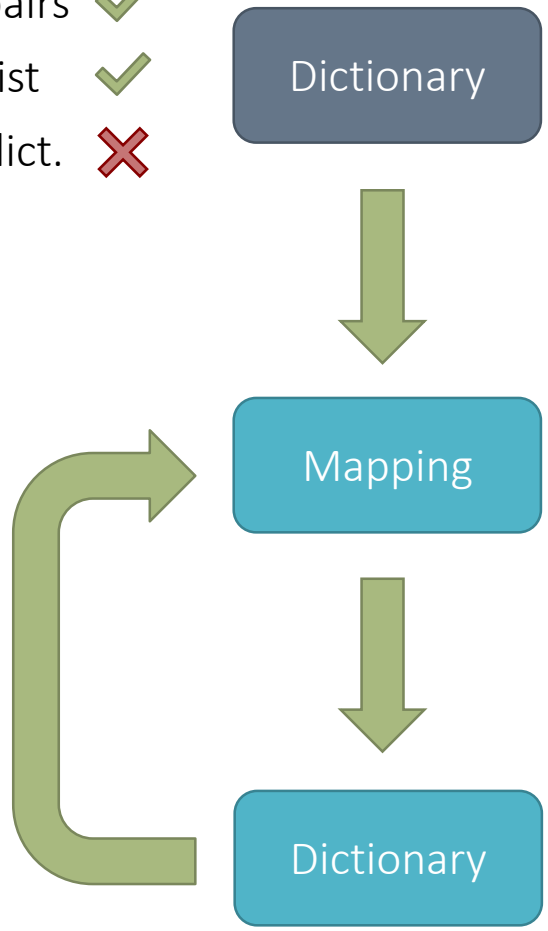


$$W^* = \arg\min_{W \in O(n)} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual embedding mappings

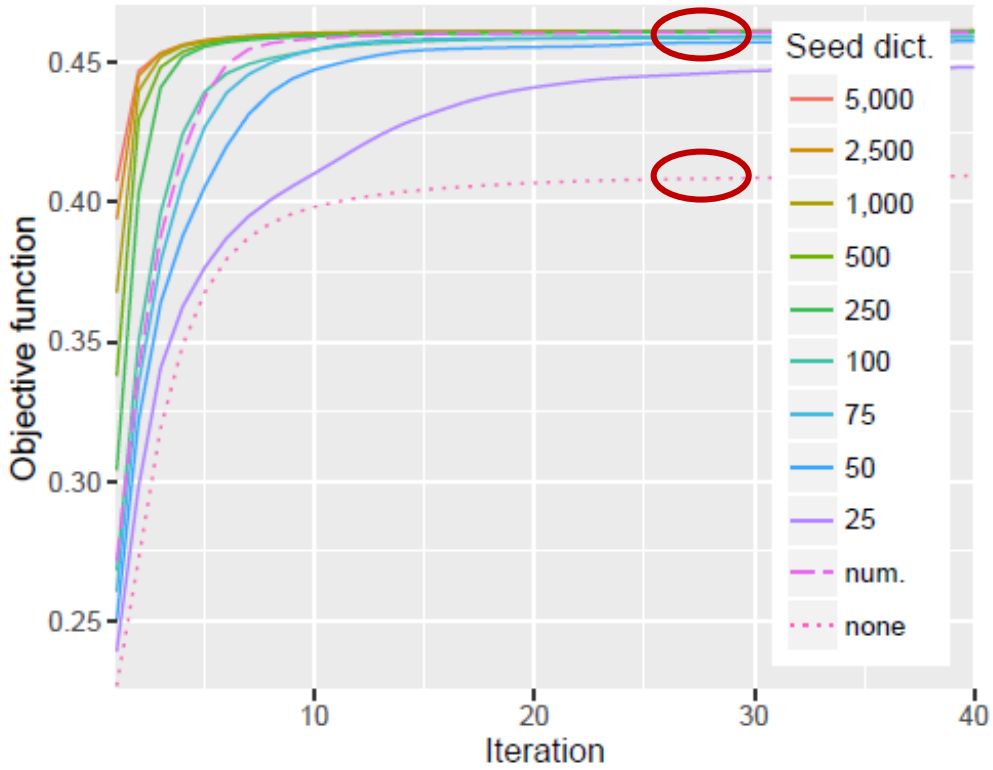


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

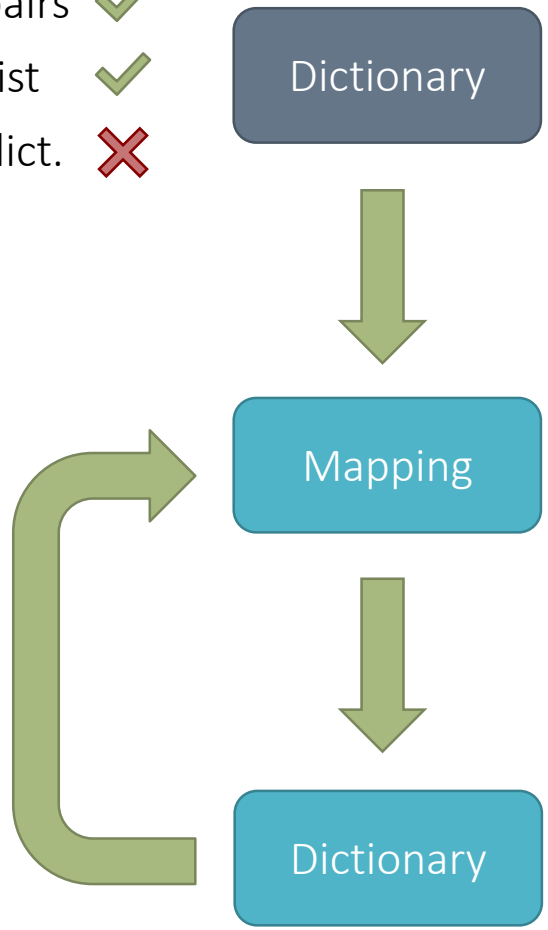


$$W^* = \arg\min_{W \in O(n)} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual embedding mappings



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

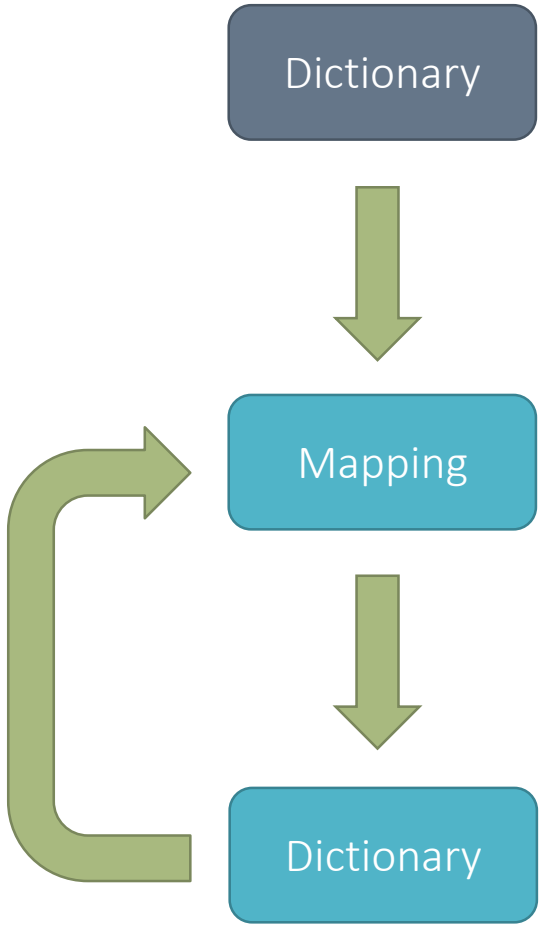


$$W^* = \arg\min_{W \in O(n)} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual embedding mappings

Intra-lingual similarity distribution

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

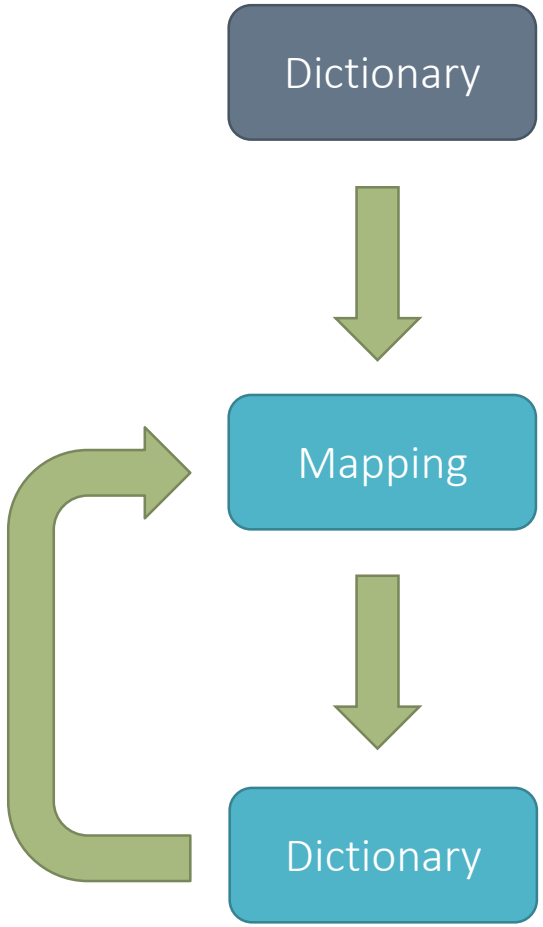


Cross-lingual embedding mappings

English

Intra-lingual similarity distribution

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



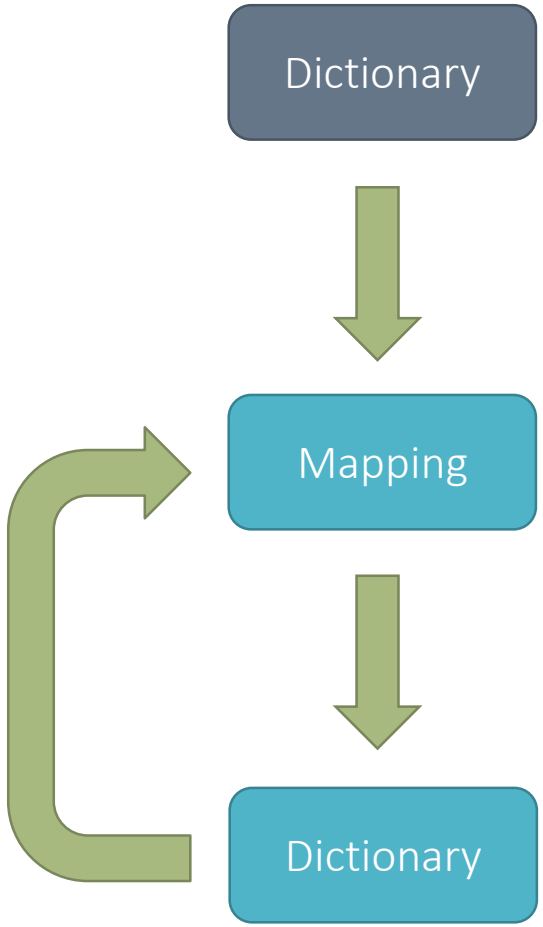
Cross-lingual embedding mappings

English

two

Intra-lingual similarity distribution

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



Cross-lingual embedding mappings

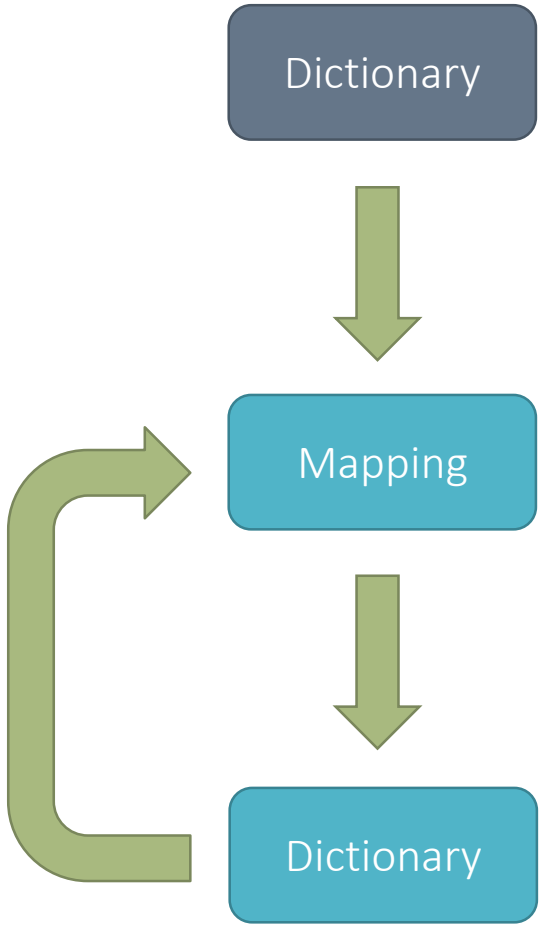
English

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

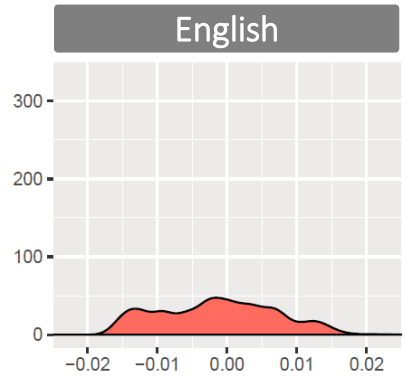
two

Intra-lingual similarity distribution

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



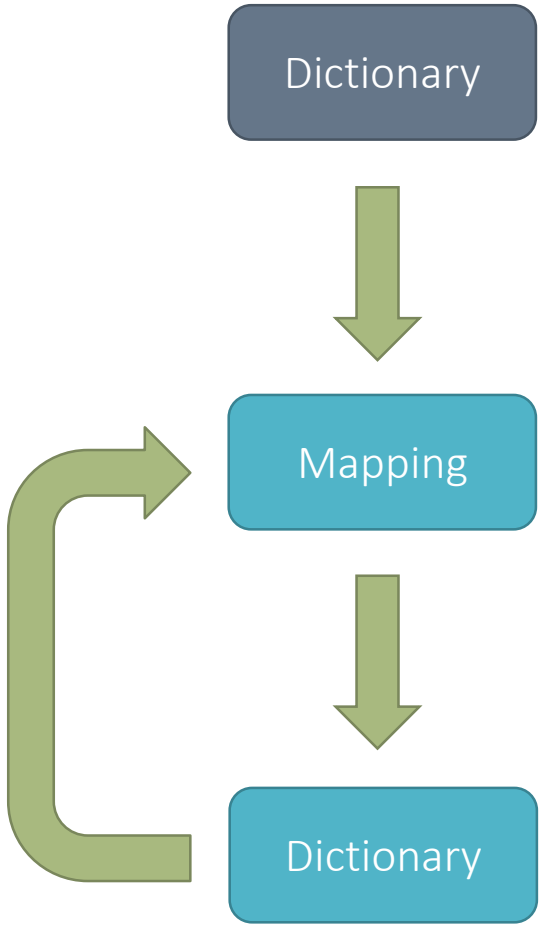
Cross-lingual embedding mappings



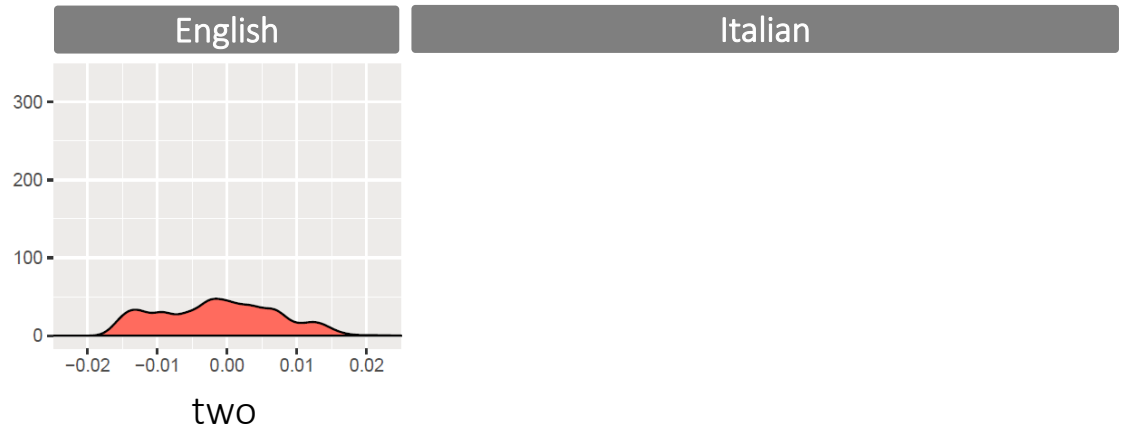
two

Intra-lingual similarity distribution

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

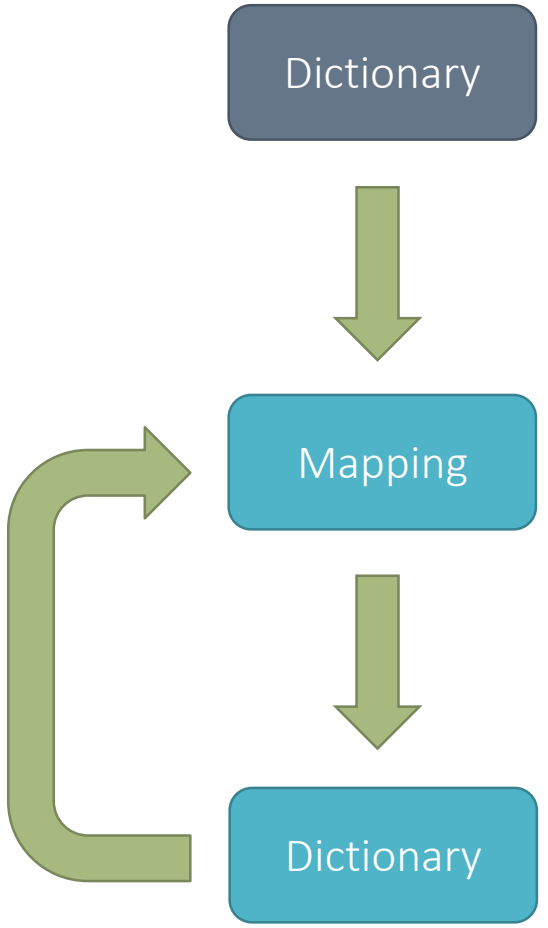


Cross-lingual embedding mappings

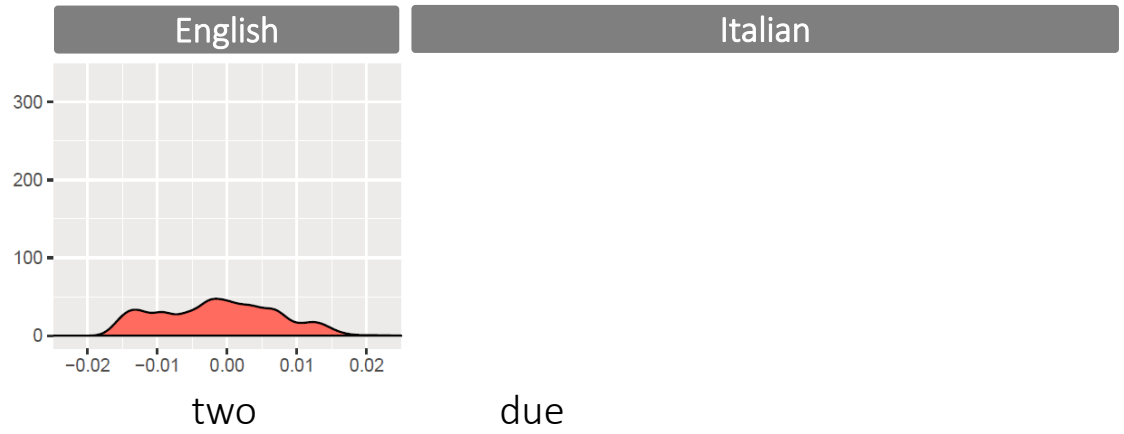


Intra-lingual similarity distribution

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

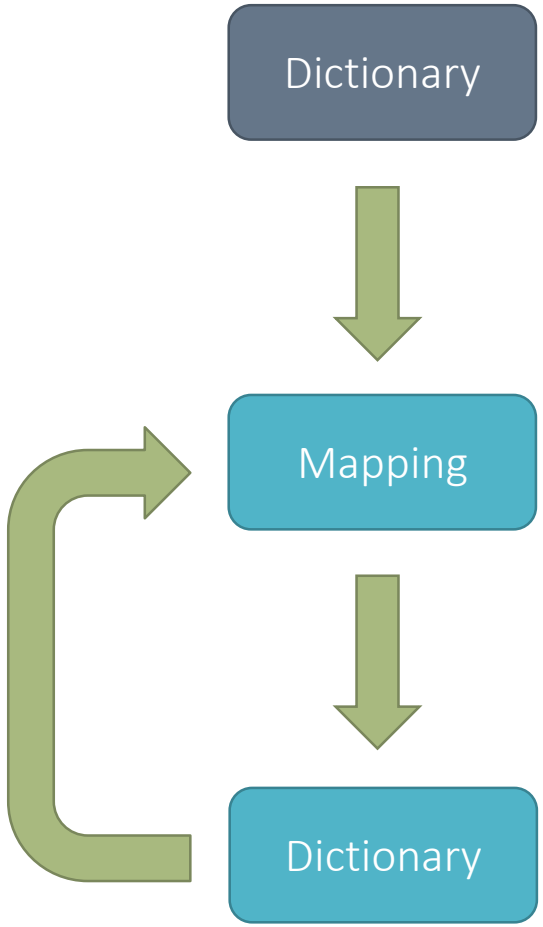


Cross-lingual embedding mappings

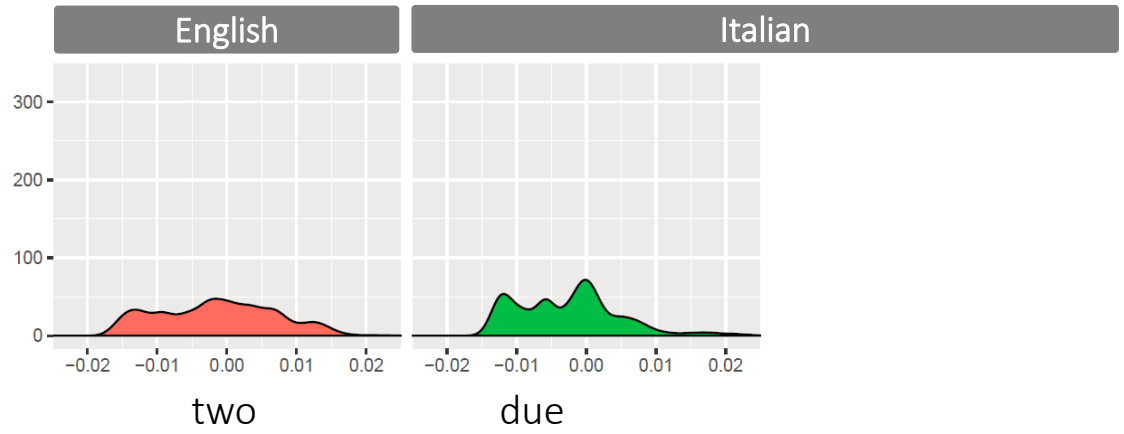


Intra-lingual similarity distribution

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

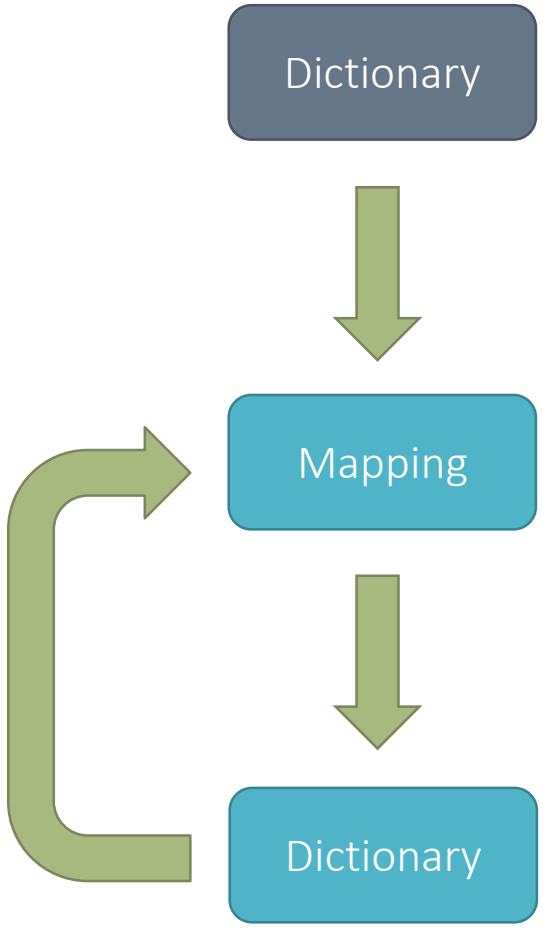


Cross-lingual embedding mappings

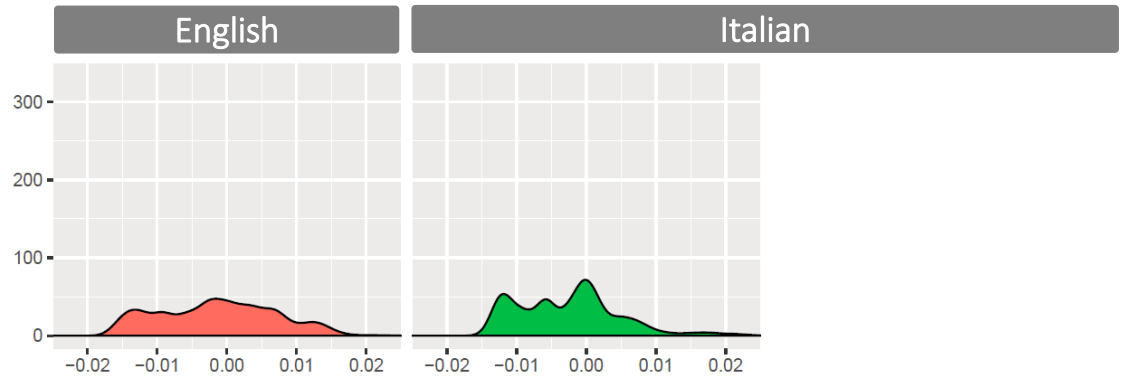


Intra-lingual similarity distribution

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

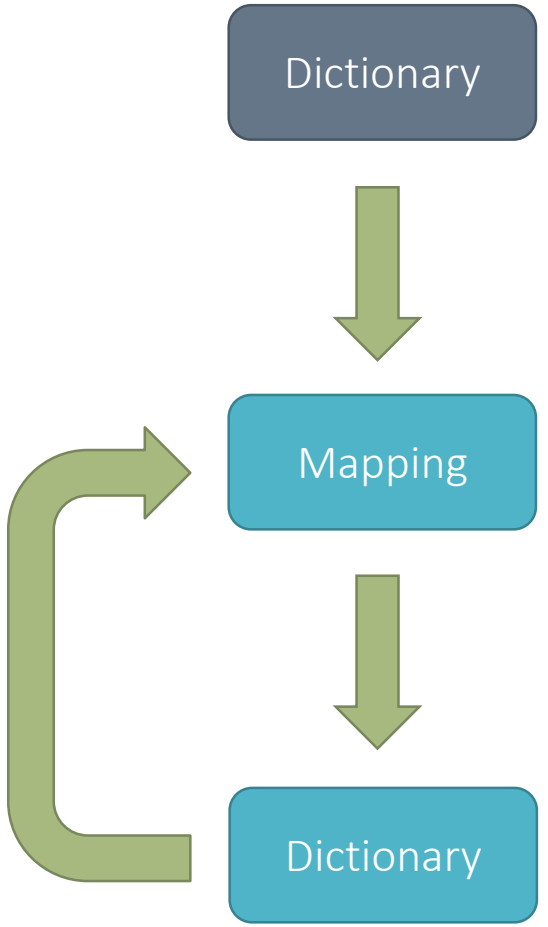


Cross-lingual embedding mappings

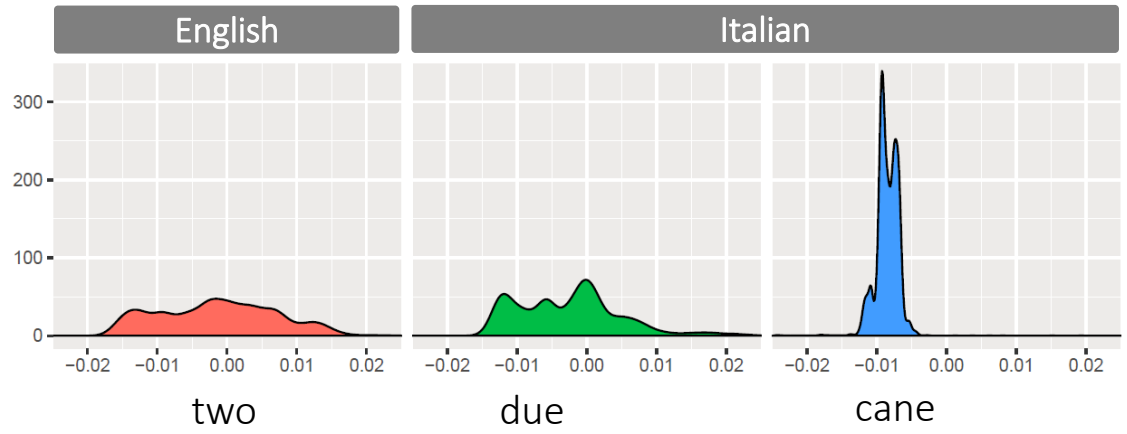


Intra-lingual similarity distribution

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

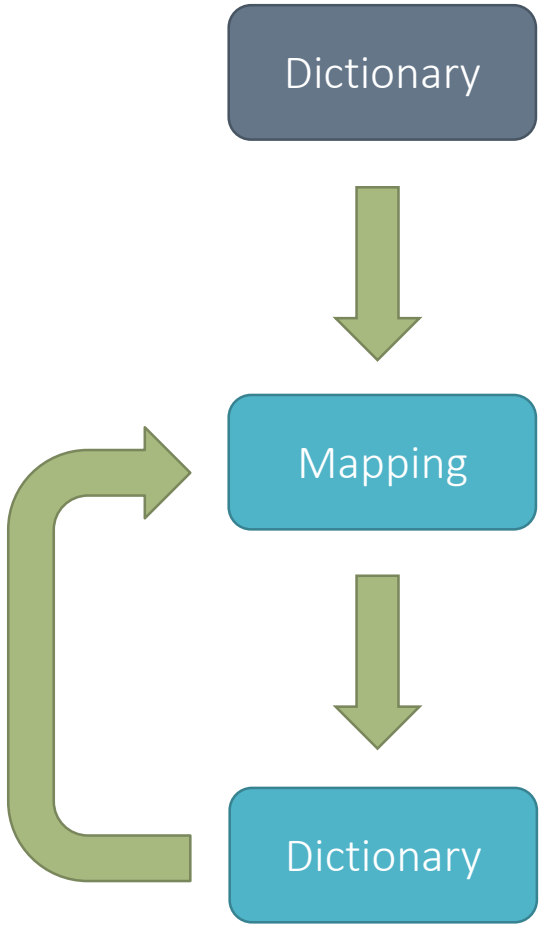


Cross-lingual embedding mappings

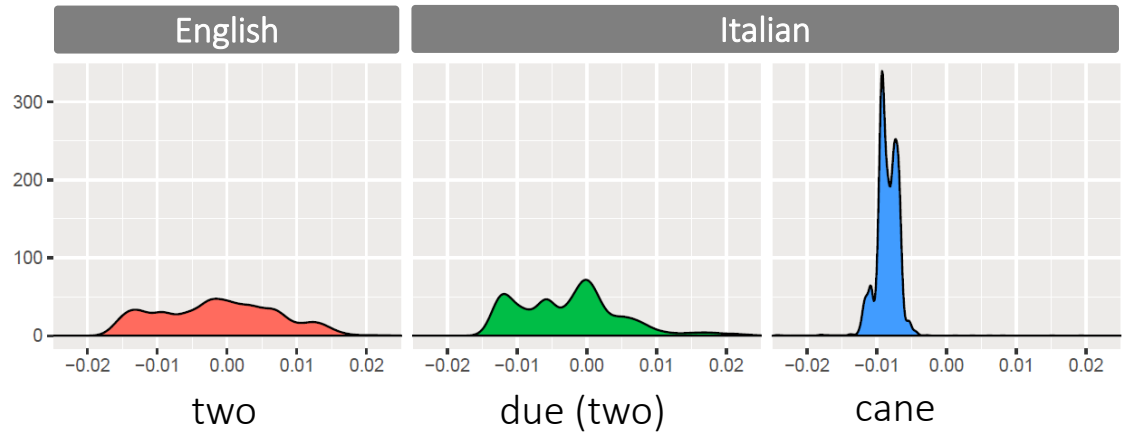


Intra-lingual similarity distribution

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

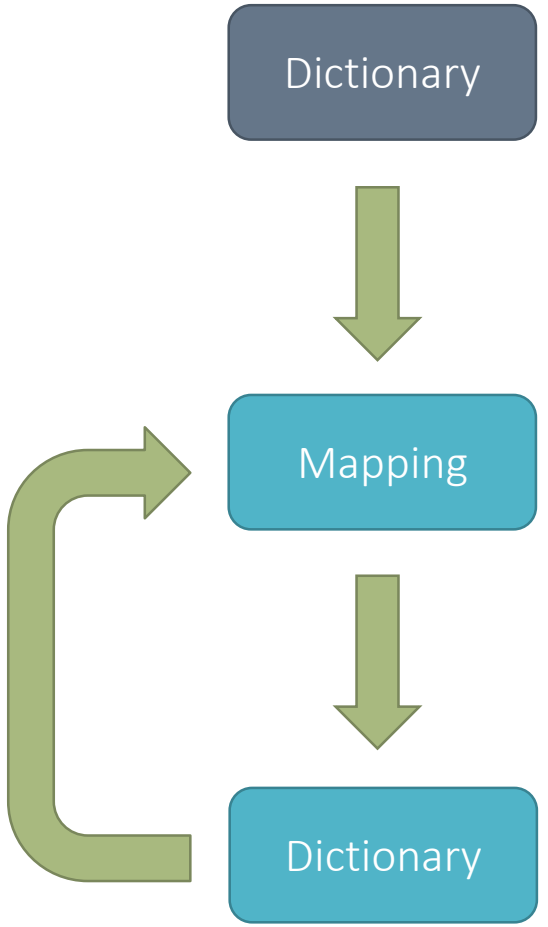


Cross-lingual embedding mappings

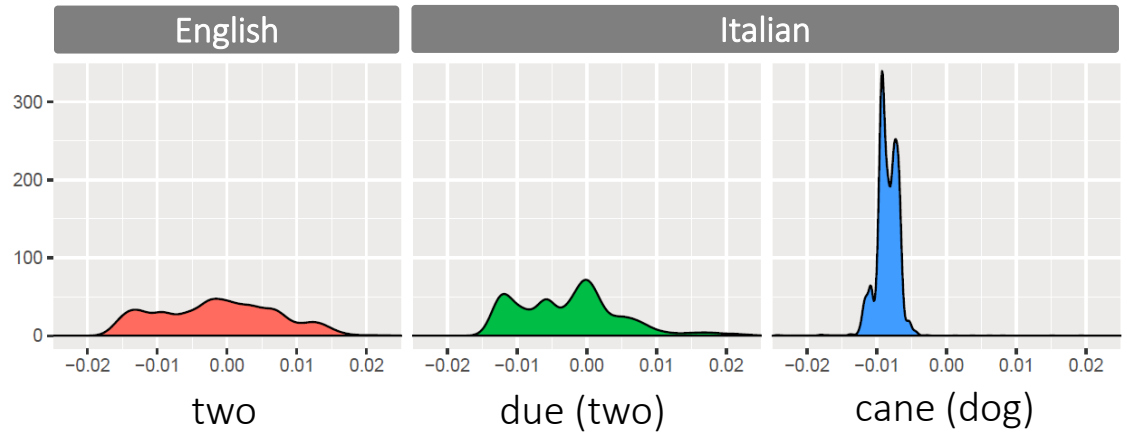


Intra-lingual similarity distribution

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

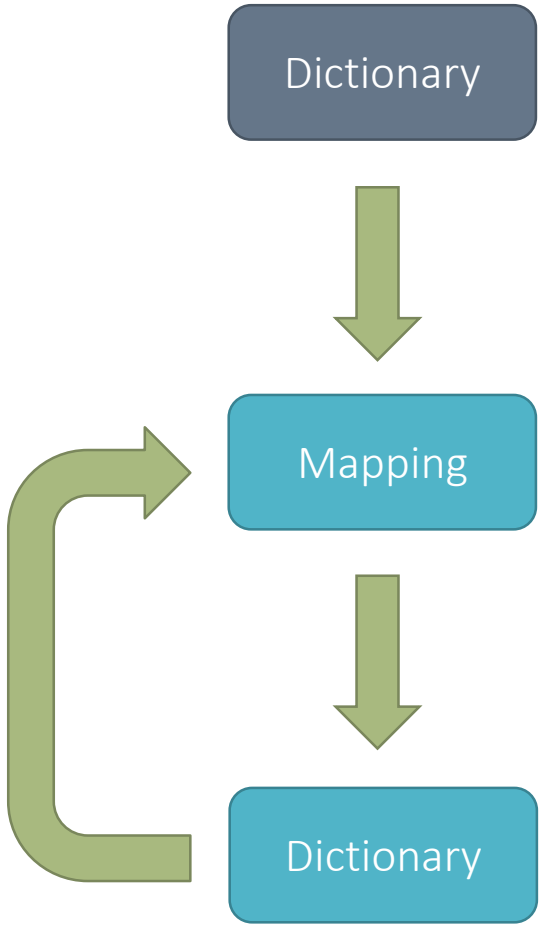


Cross-lingual embedding mappings

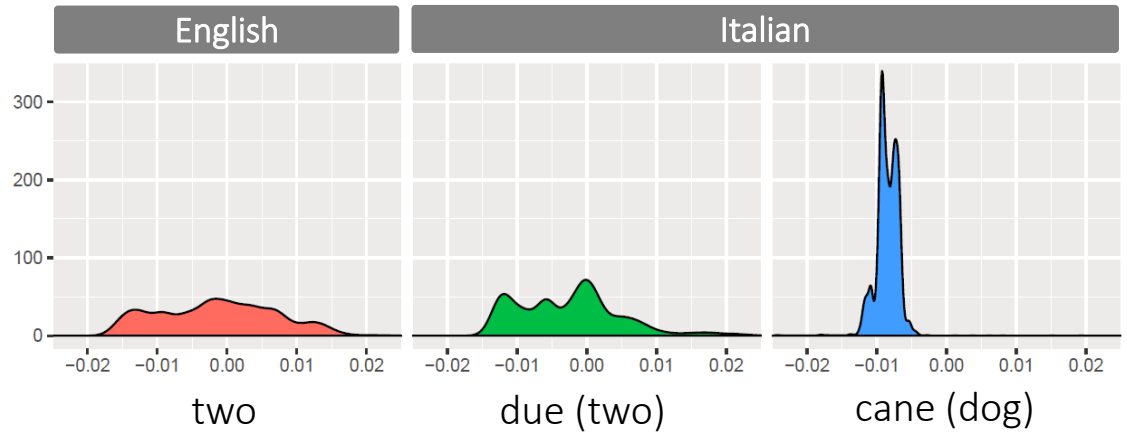


Intra-lingual similarity distribution

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

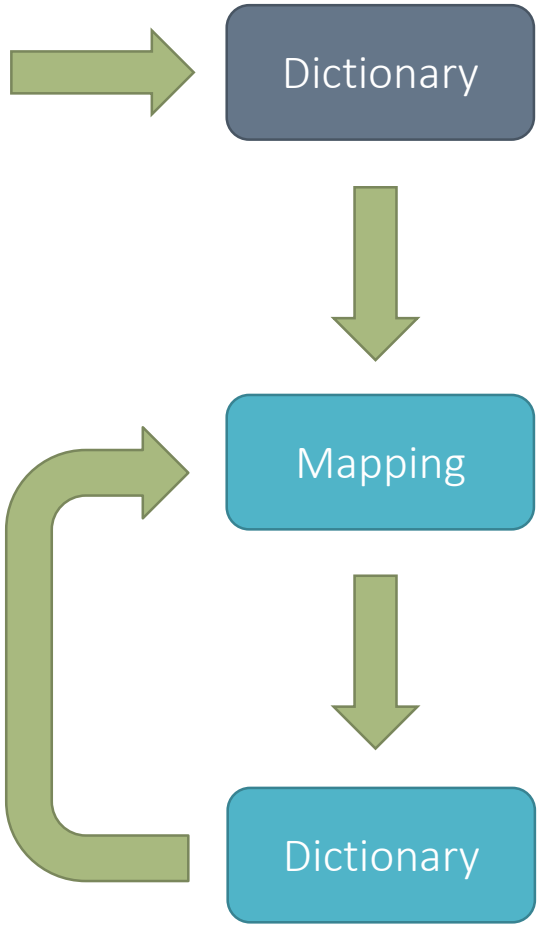


Cross-lingual embedding mappings

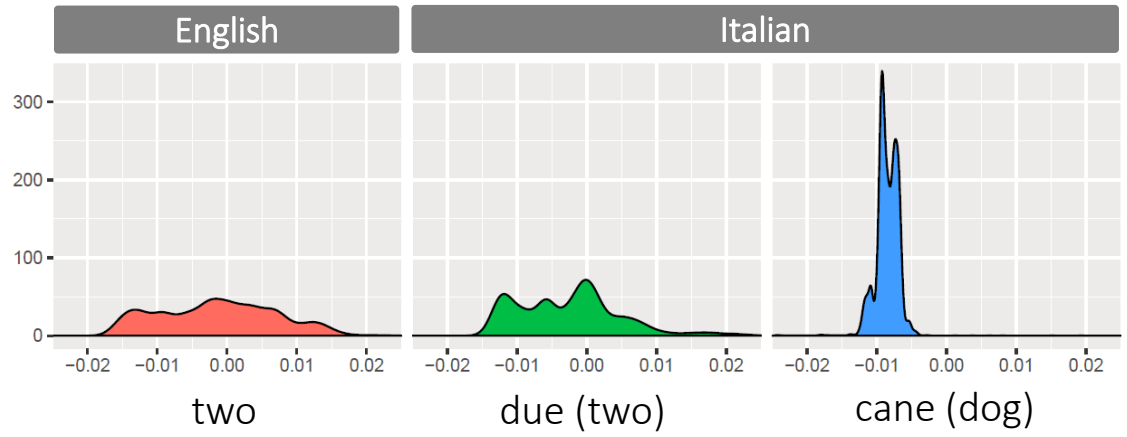


Intra-lingual similarity distribution

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



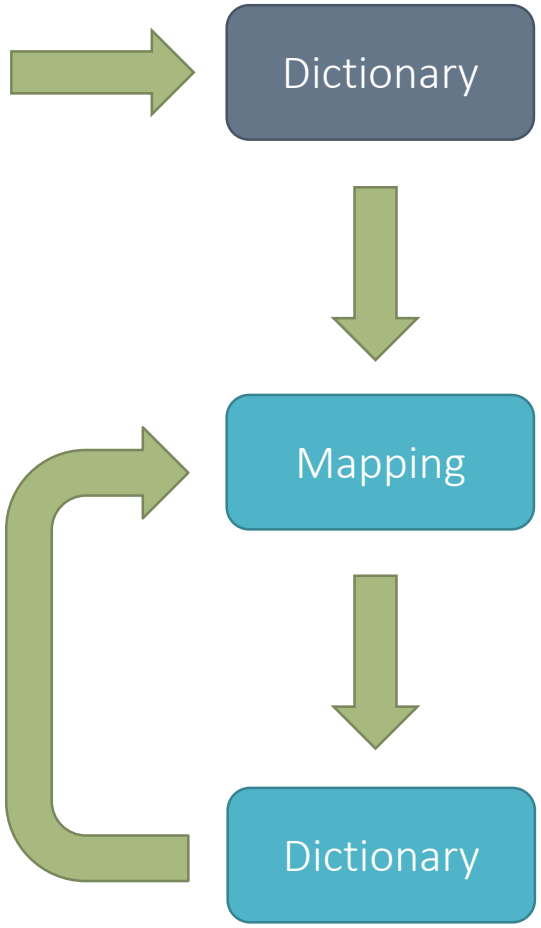
Cross-lingual embedding mappings



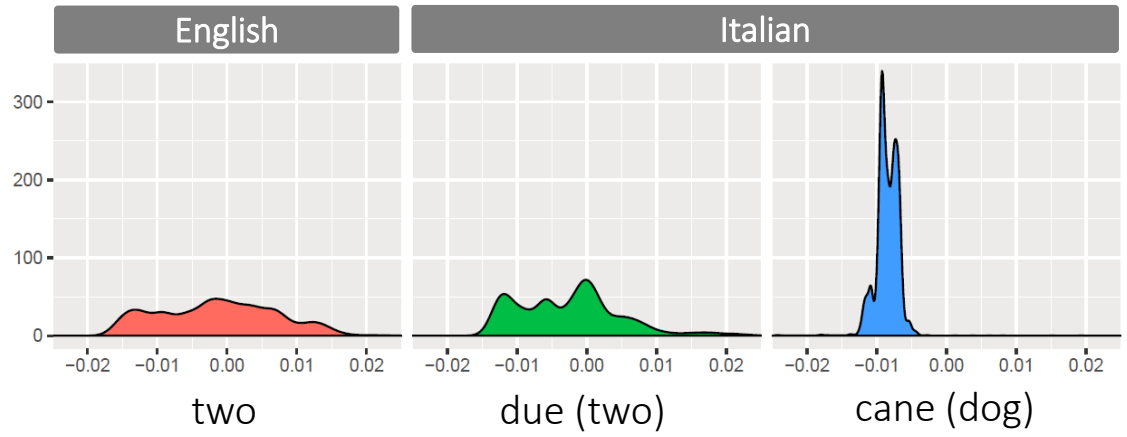
Intra-lingual similarity distribution

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

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



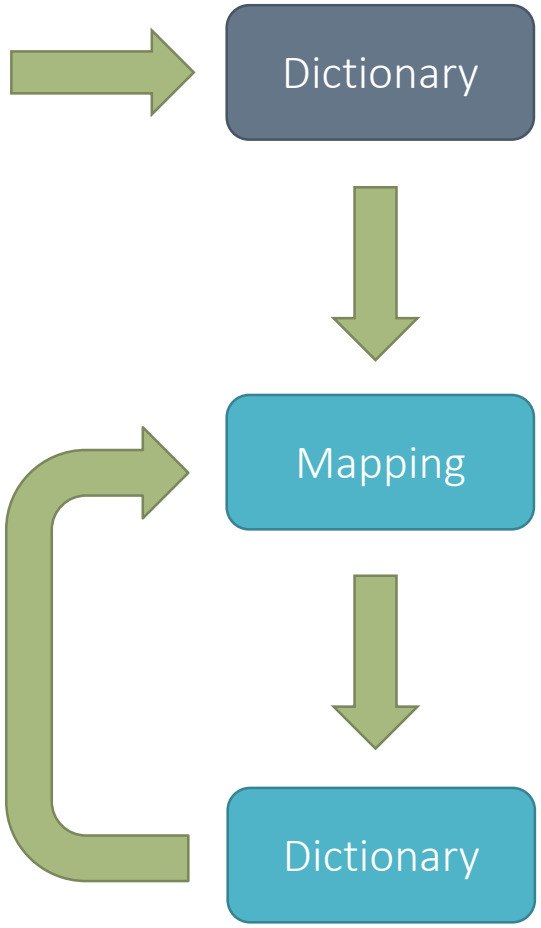
Cross-lingual embedding mappings



Intra-lingual similarity distribution

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

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



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. heurist.	Smith et al. (2017), cognates
	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. heurist.	Smith et al. (2017), cognates
	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. heurist.	Smith et al. (2017), cognates				
	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. heurist.	Smith et al. (2017), cognates	39.9	-	-	-
	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	**37.33**

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. heurist.	Smith et al. (2017), cognates	39.9	-	-	-
	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	**37.33**

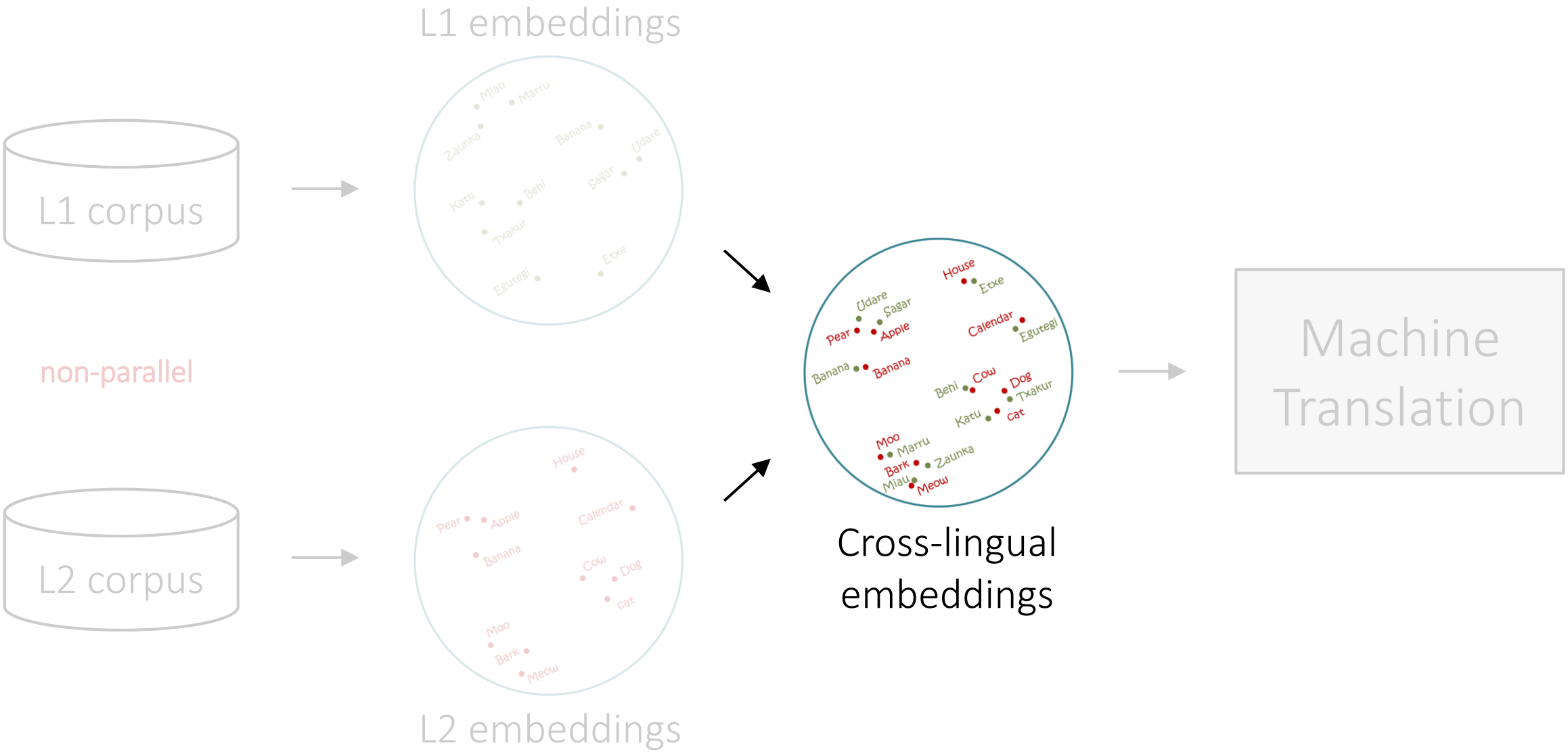
Cross-lingual embedding mappings

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. heurist.	Smith et al. (2017), cognates	39.9	-	-	-
	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	**37.33**

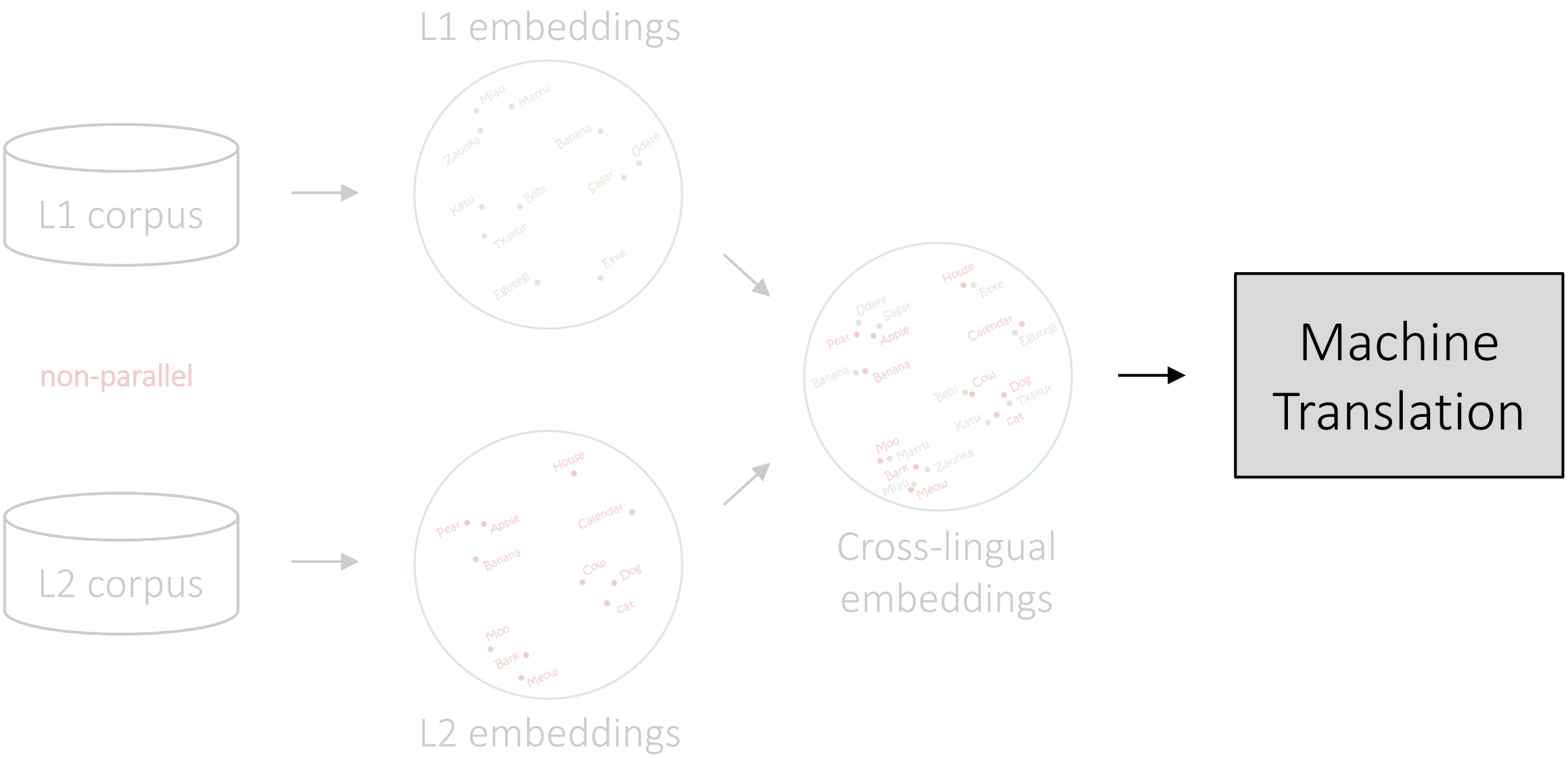
Cross-lingual embedding mappings

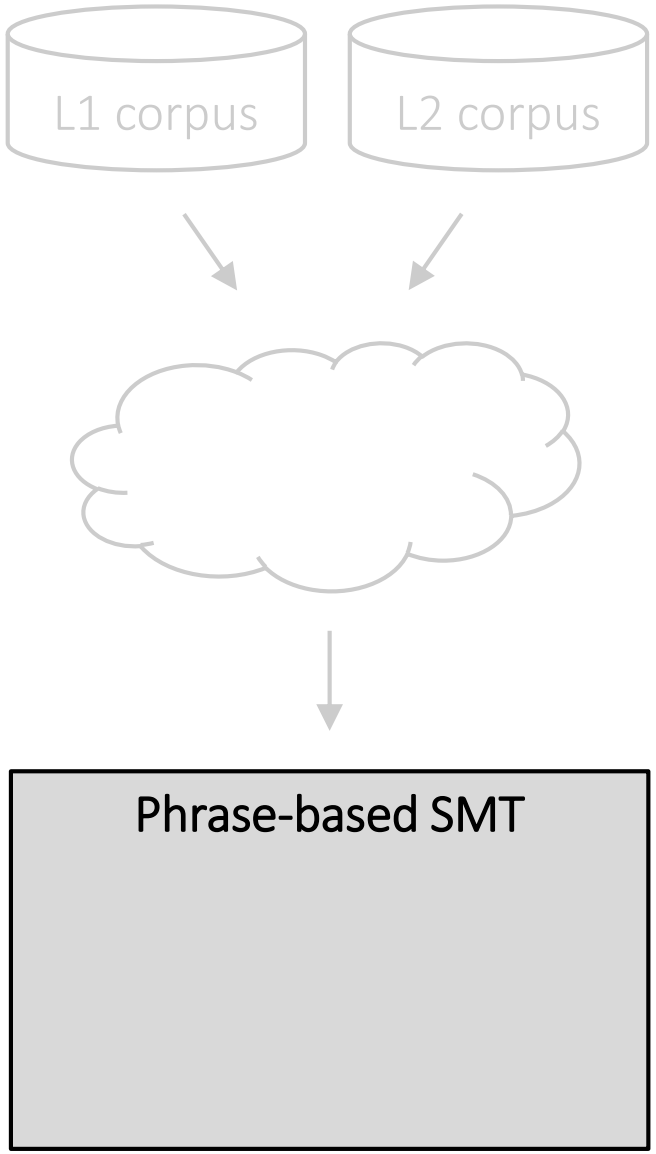
Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	34.93†	35.00†	25.91†	27.73†
	Faruqui and Dyer (2014)	38.40*	37.13*	27.60*	26.80*
	Shigeto et al. (2015)	41.53†	43.07†	31.04†	33.73†
	Dinu et al. (2015)	37.7	38.93*	29.14*	30.40*
	Lazaridou et al. (2015)	40.2	-	-	-
	Xing et al. (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. heurist.	Smith et al. (2017), cognates	39.9	-	-	-
	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	**37.33**

Outline

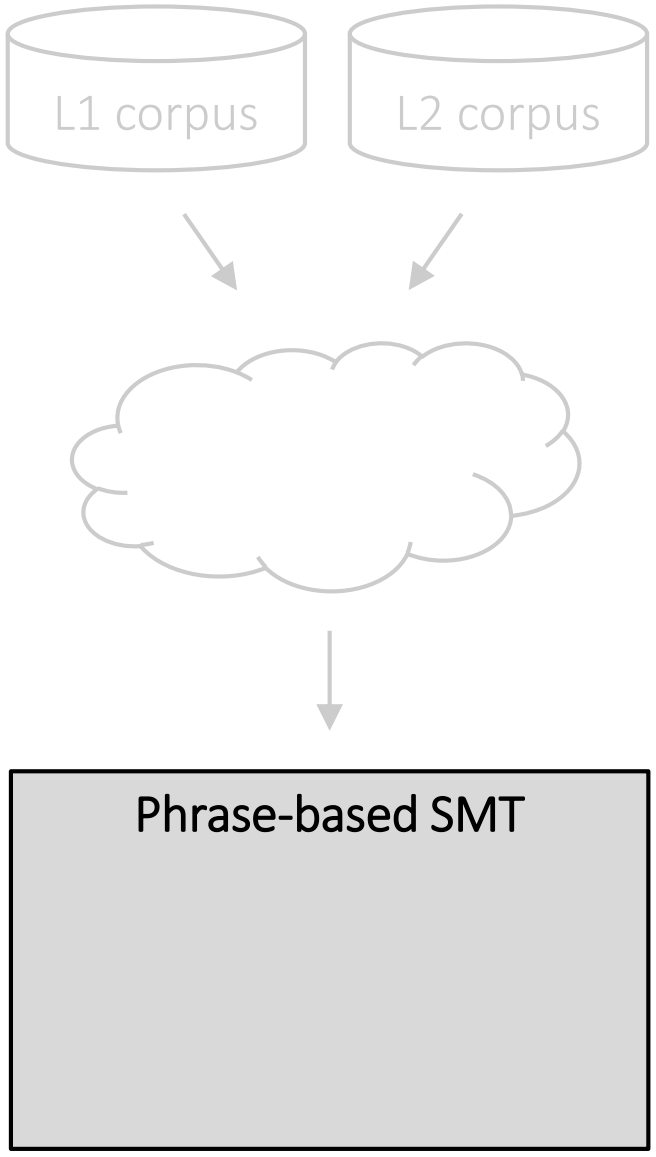

L1 embeddings
L1 corpus
non-parallel
L2 corpus
L2 embeddings
Cross-lingual embeddings
Machine Translation

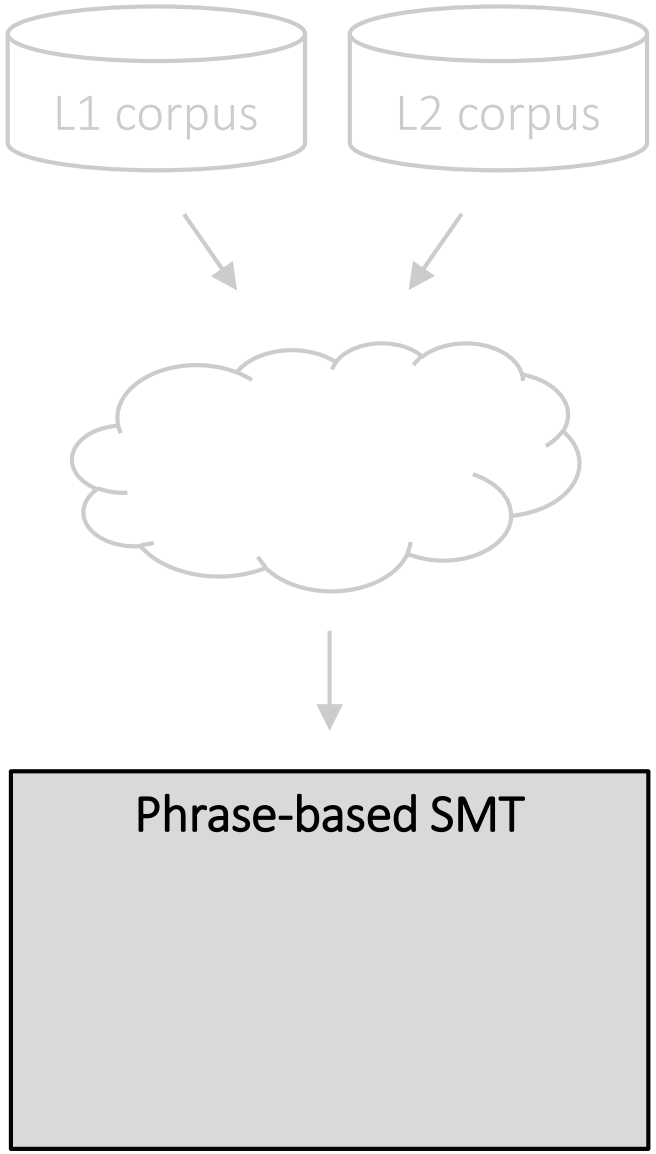
Outline


L1 embeddings
L1 corpus
non-parallel
L2 corpus
L2 embeddings
Cross-lingual embeddings
Machine Translation


L1 corpus
L2 corpus
Phrase-based SMT

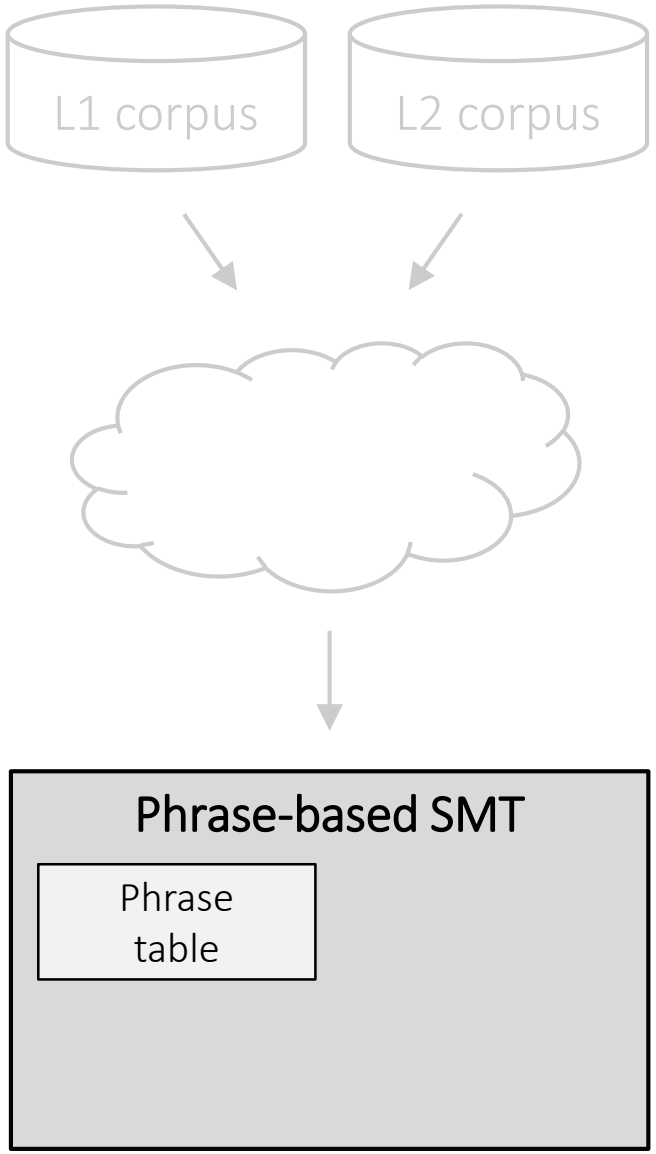
Phrase-based SMT





Phrase-based SMT

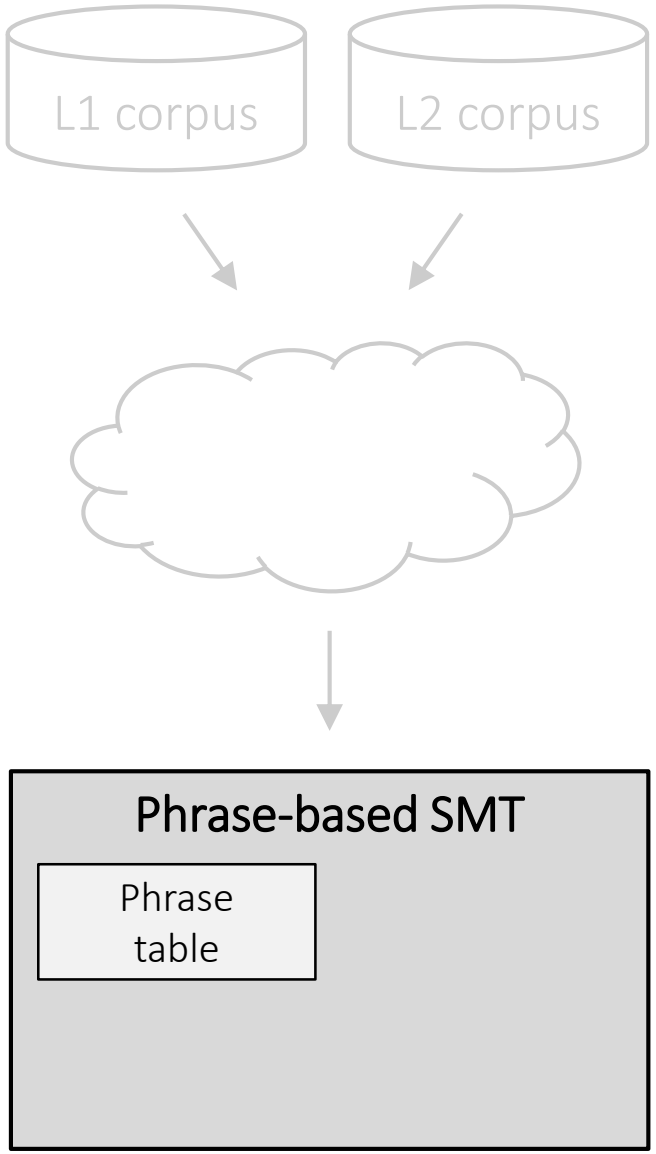
Log-linear model combining



Phrase-based SMT

Log-linear model combining

- Phrase table



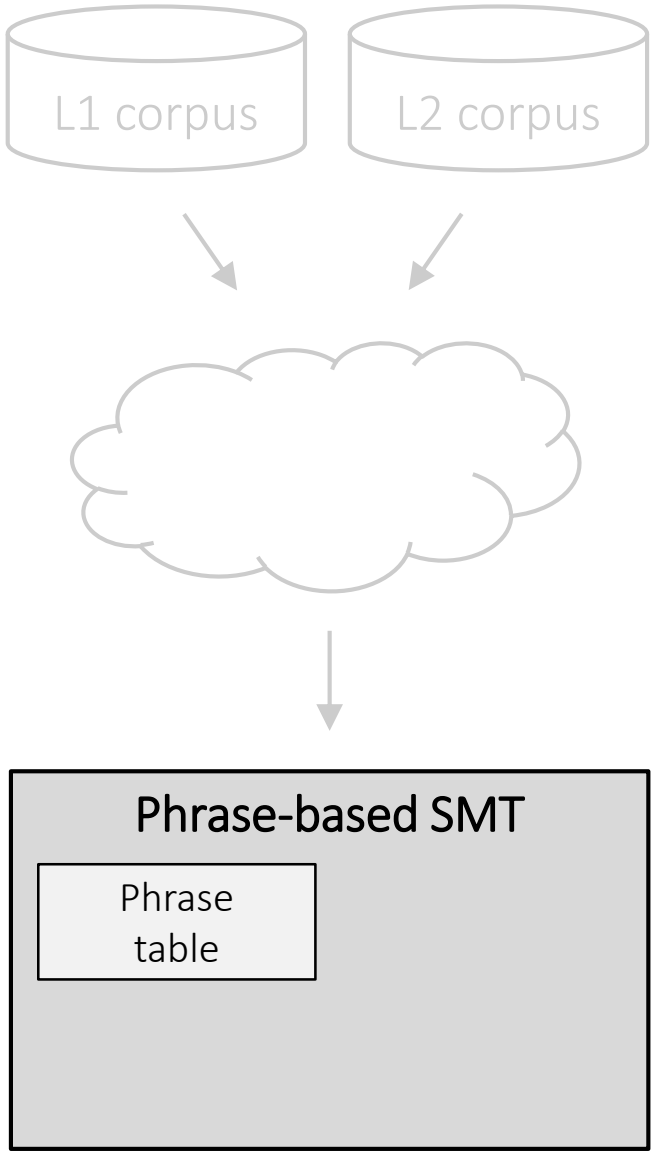
Phrase-based SMT

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

⋮



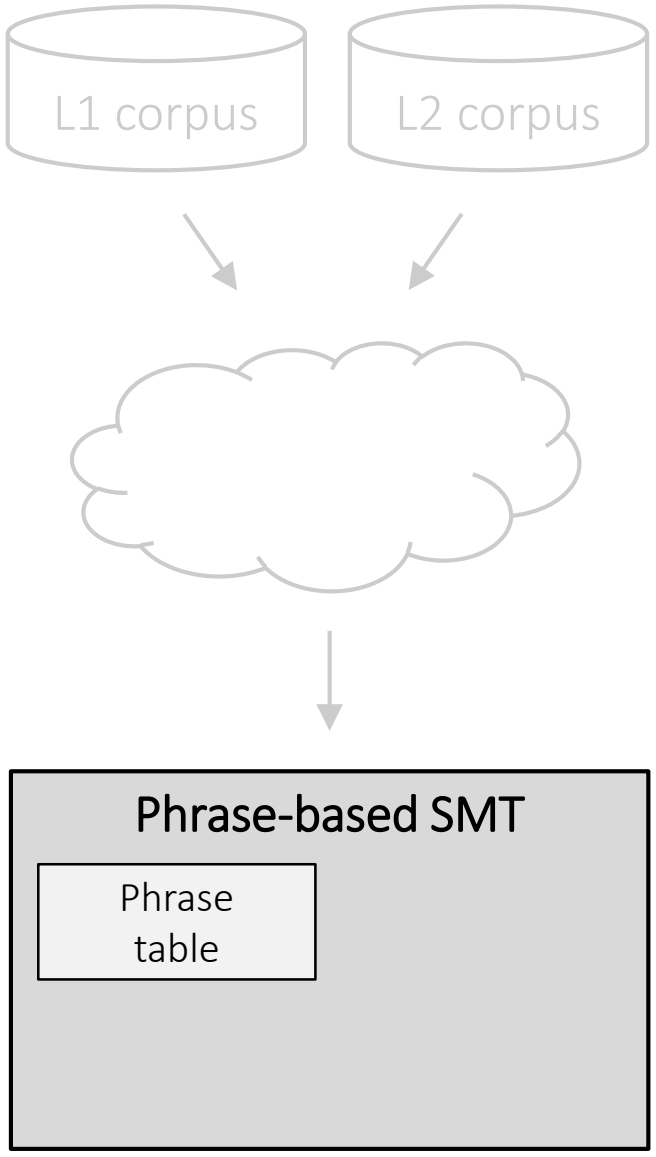
Phrase-based SMT

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

⋮



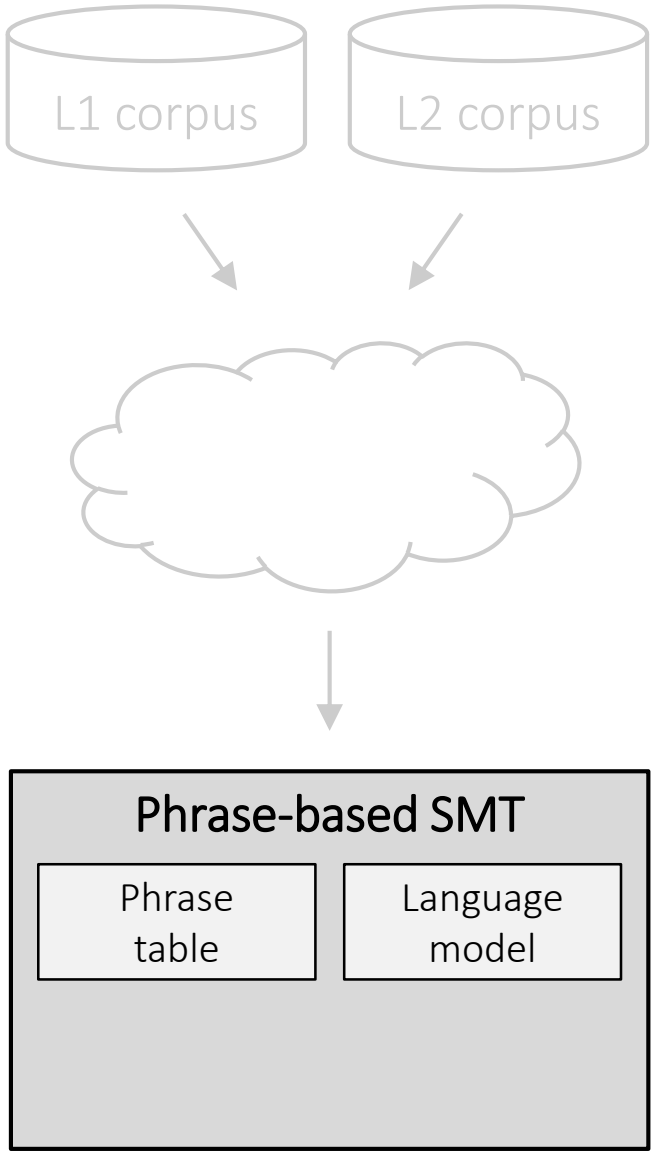
Phrase-based SMT

Log-linear model combining

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

		$\phi(\bar{f}\|\bar{e})$	$\phi(\bar{e}\|\bar{f})$	$\mathbf{lex}(\bar{f}\|\bar{e})$	$\mathbf{lex}(\bar{e}\|\bar{f})$
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

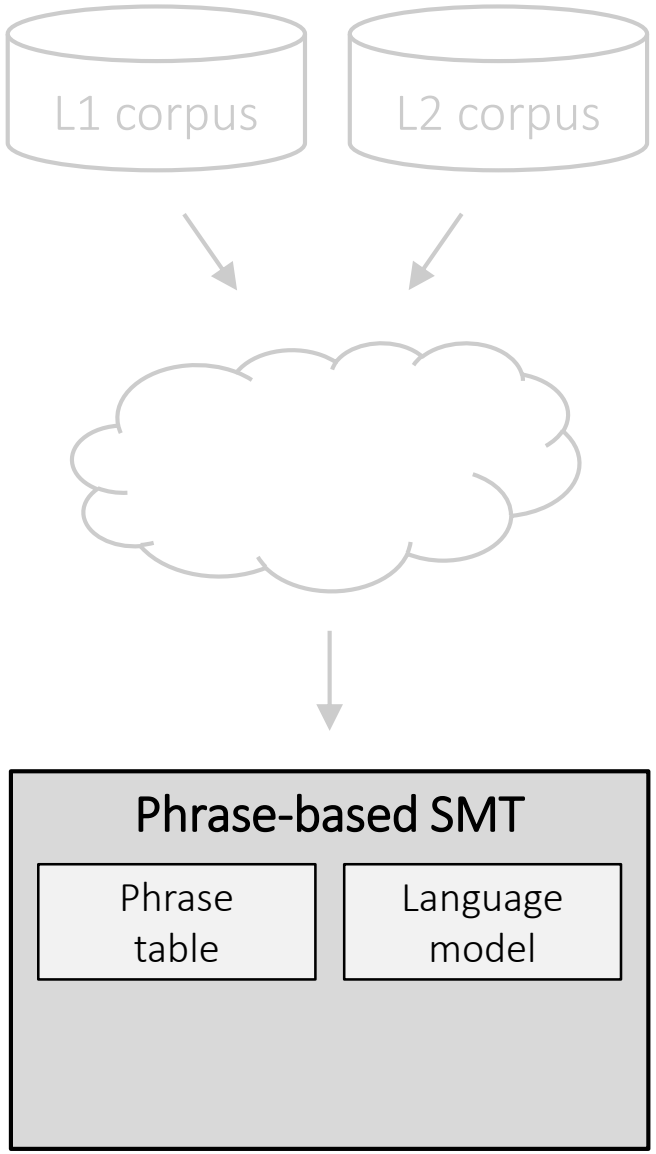
⋮



Phrase-based SMT

Log-linear model combining

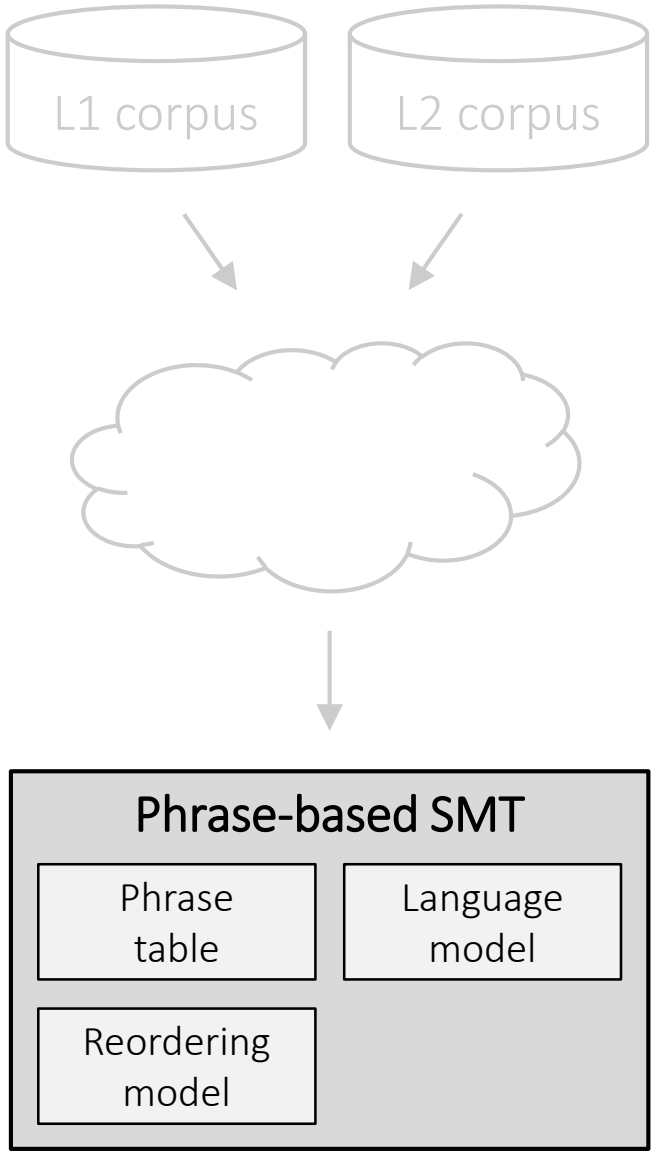
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model



Phrase-based SMT

Log-linear model combining

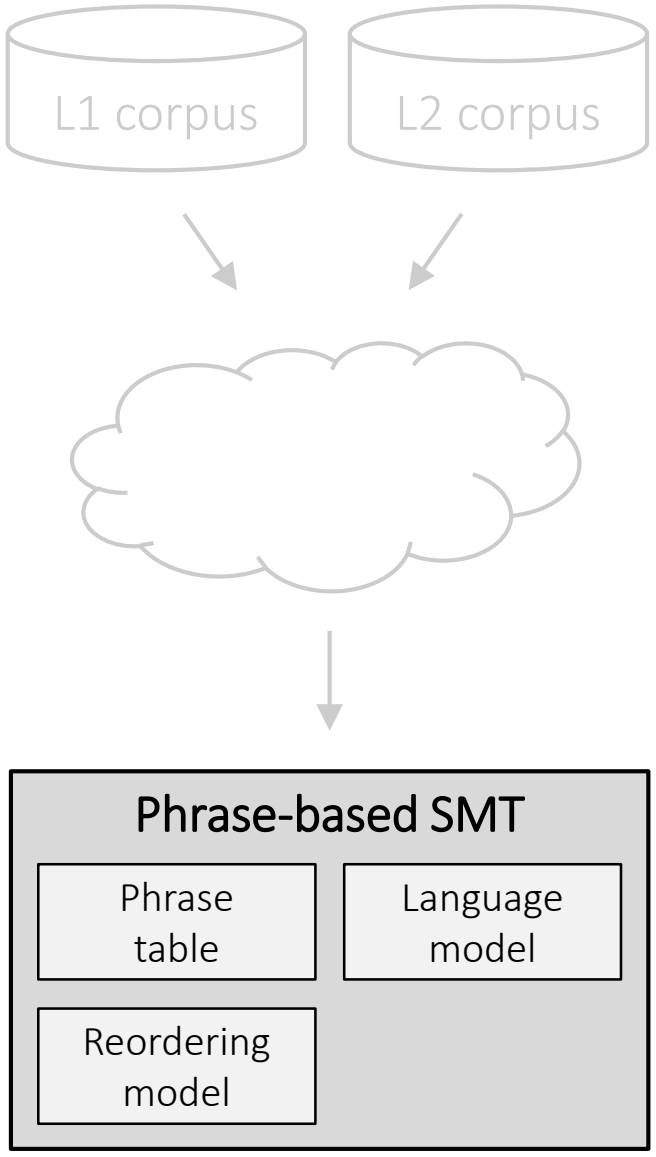
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing



Phrase-based SMT

Log-linear model combining

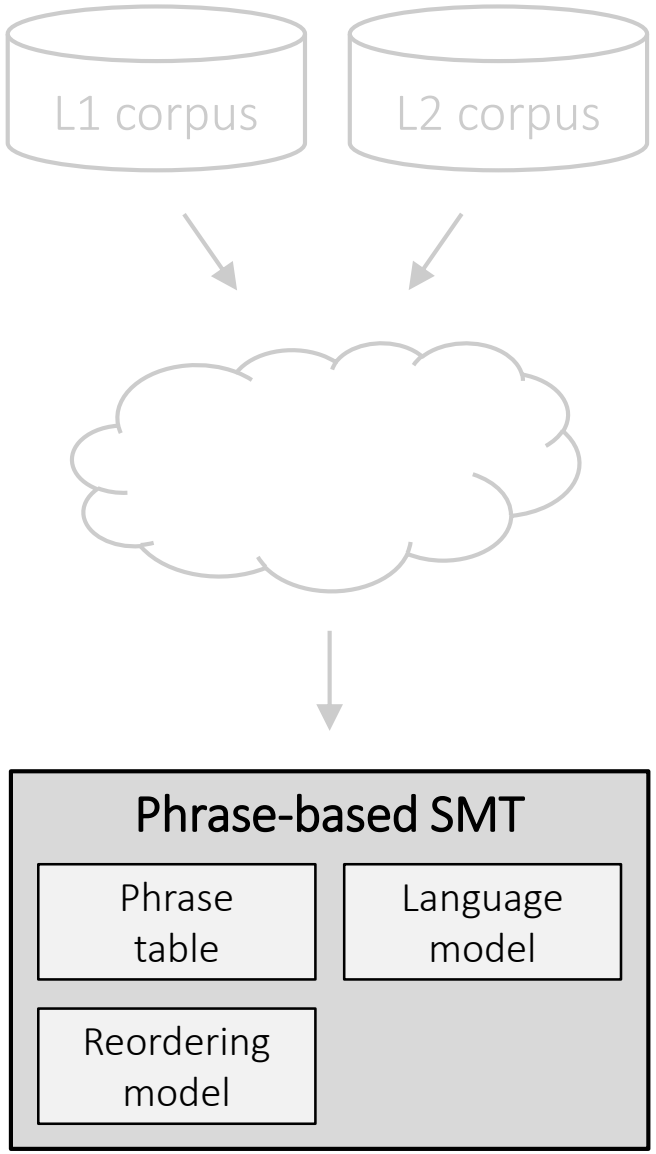
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model



Phrase-based SMT

Log-linear model combining

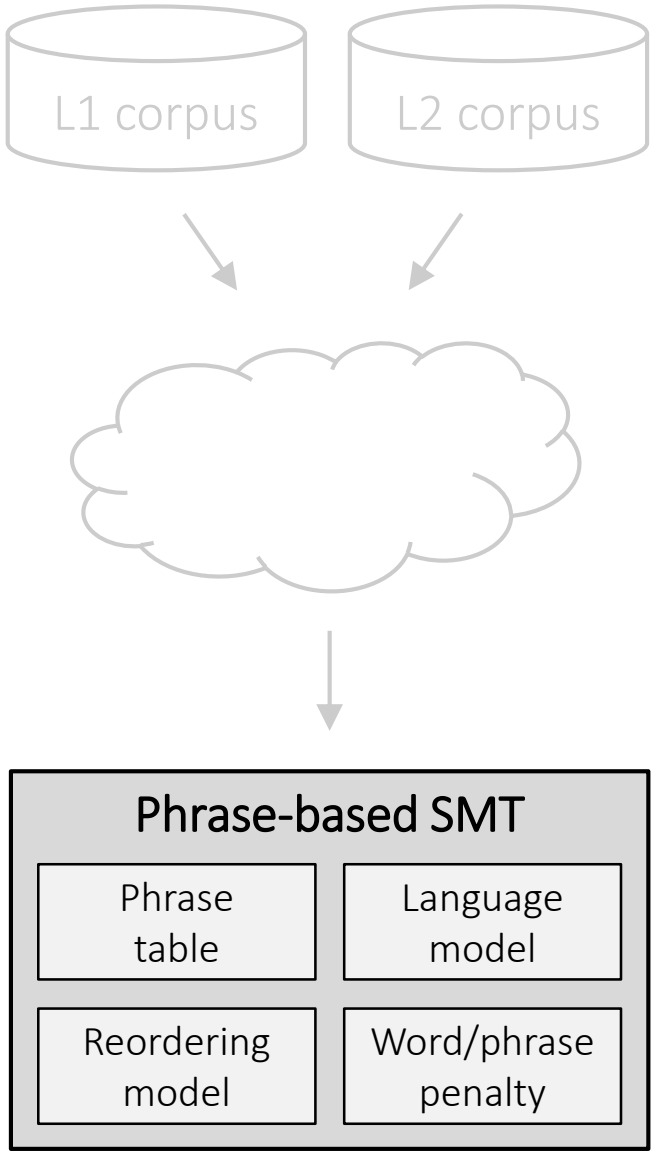
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)



Phrase-based SMT

Log-linear model combining

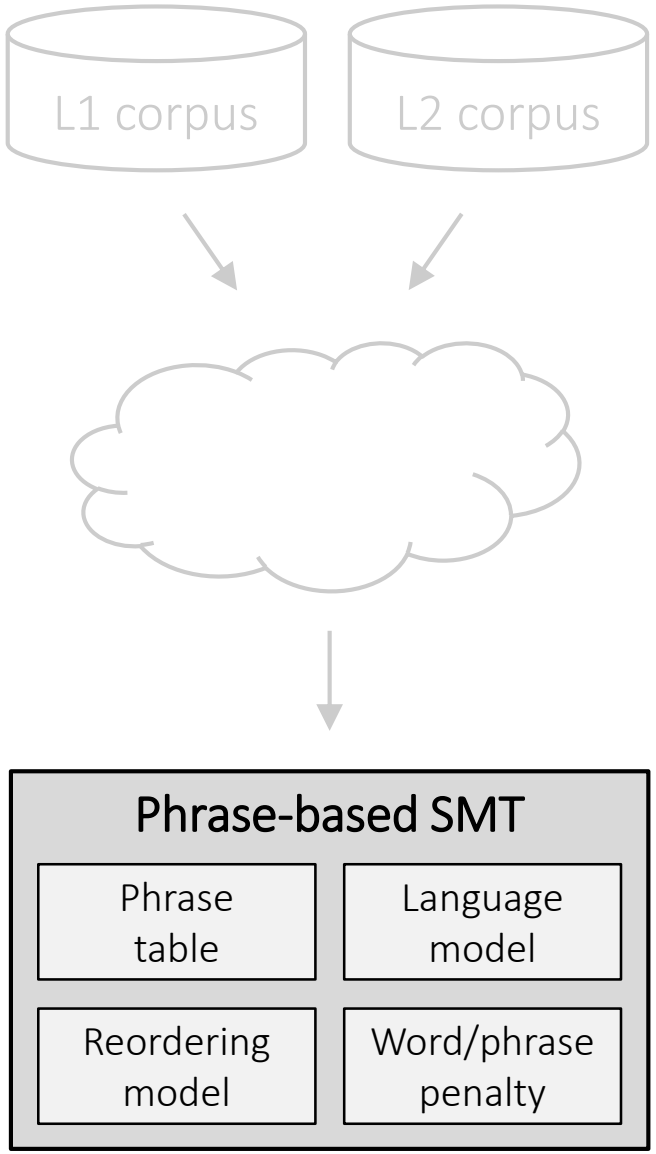
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model



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

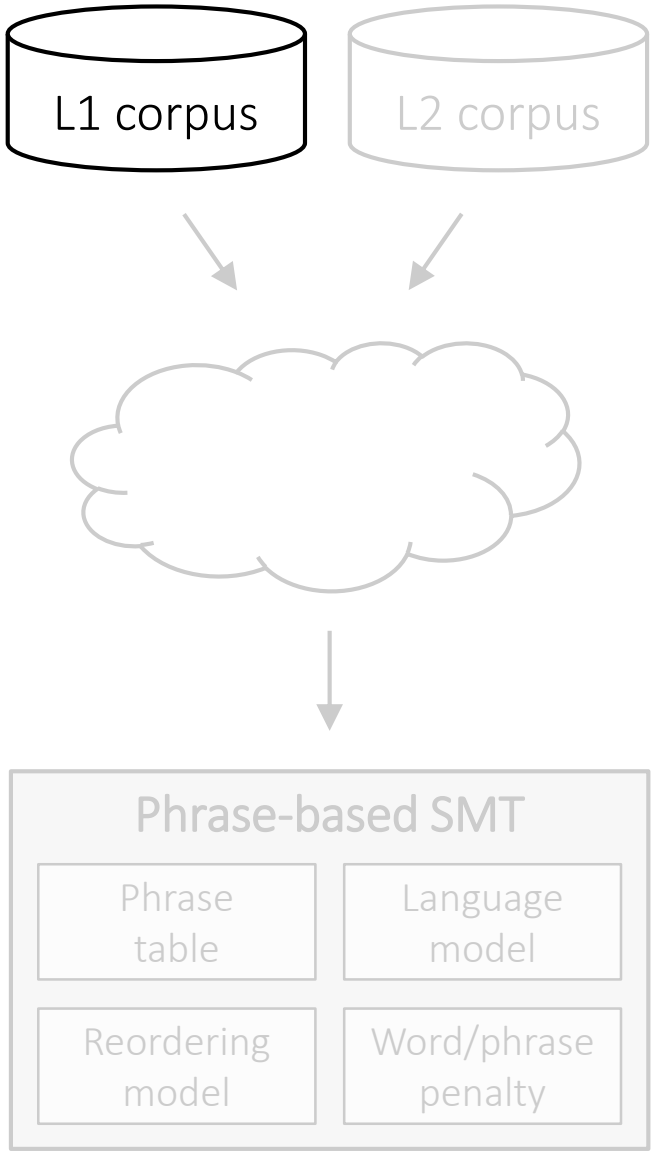


Phrase-based SMT

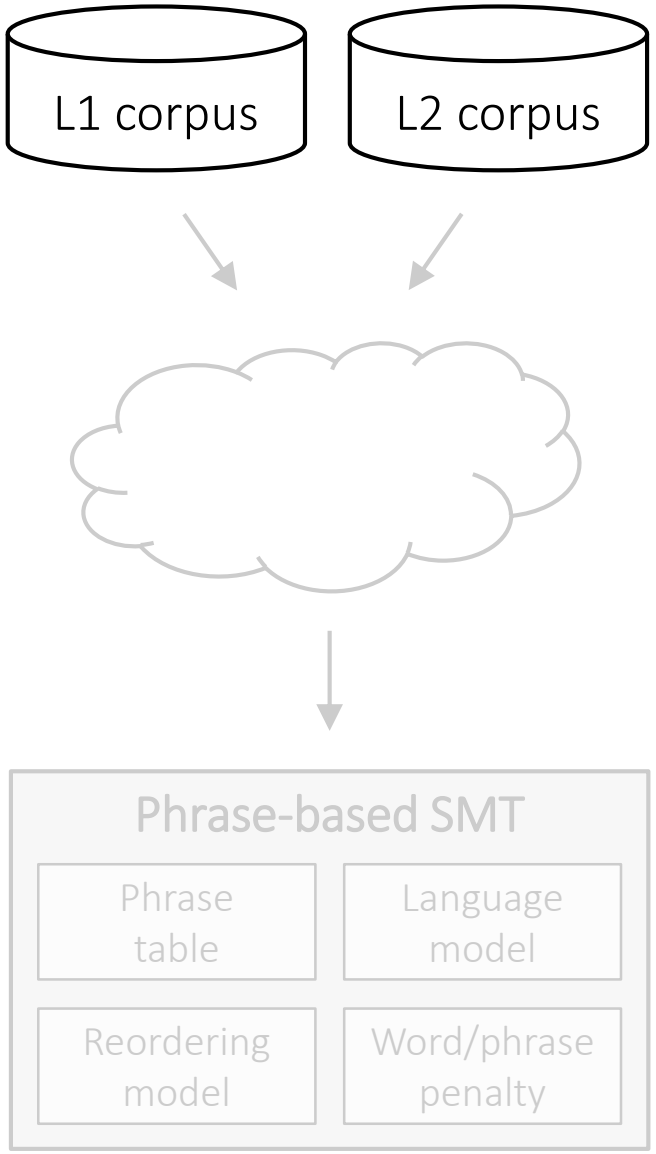
Log-linear model combining

- 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

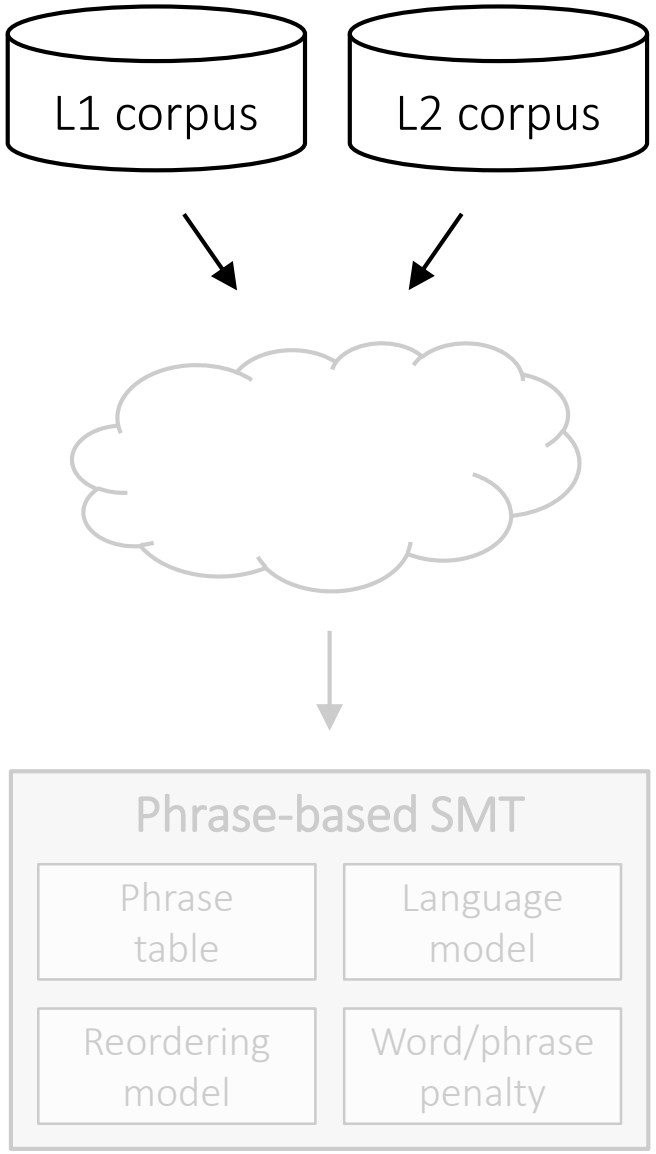
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



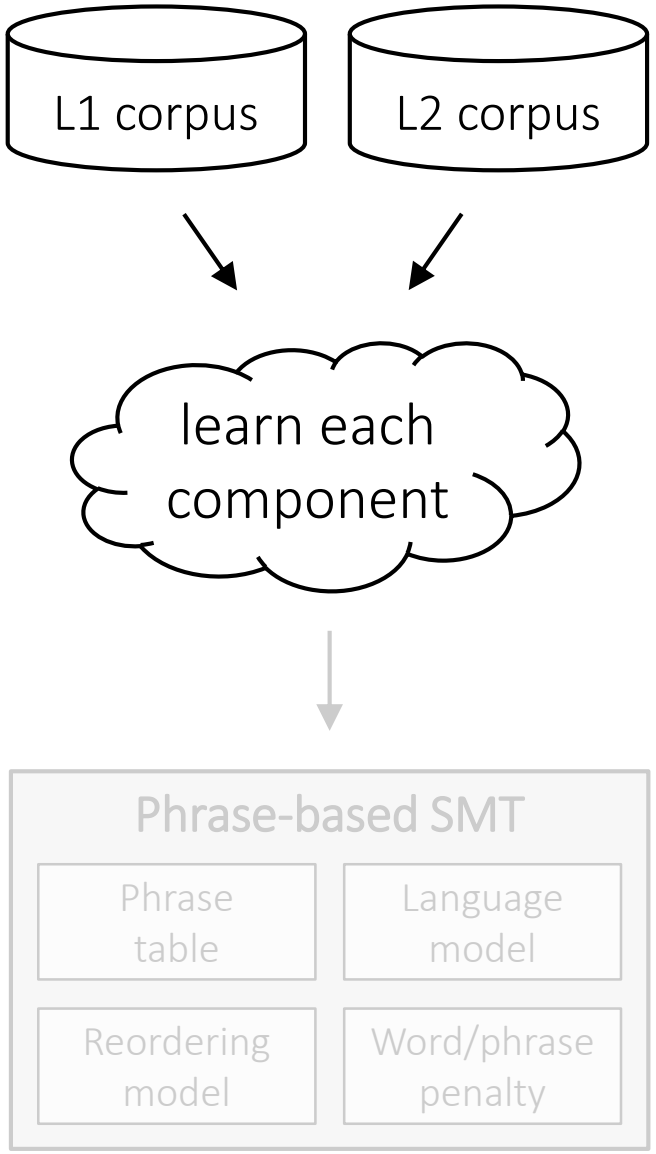
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



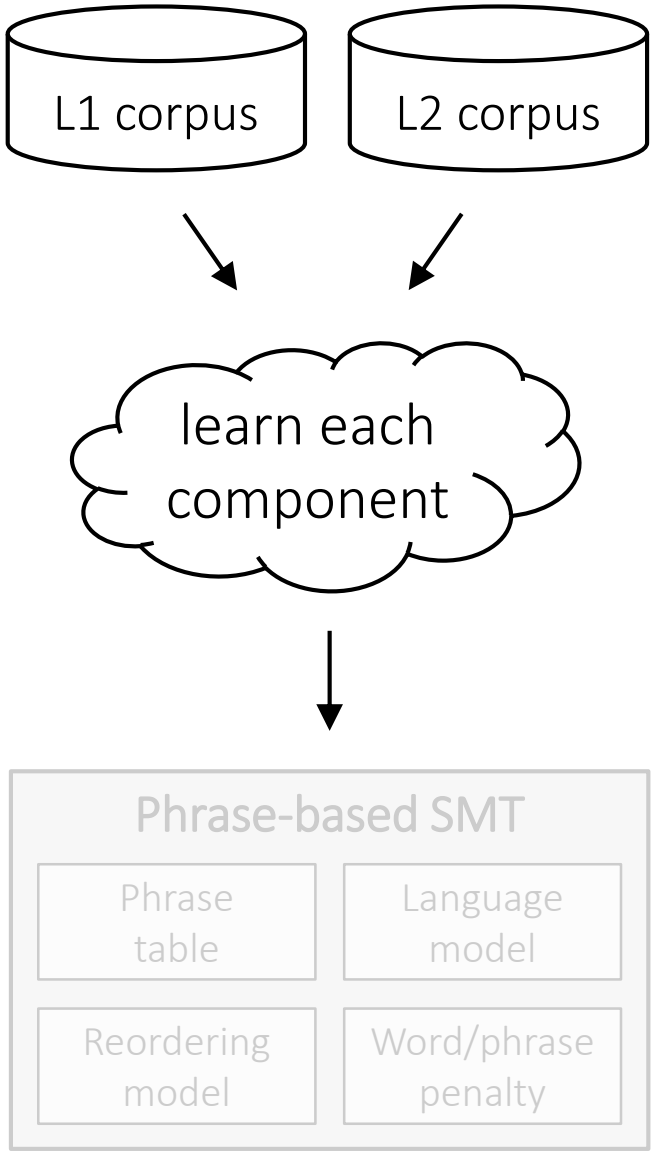
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



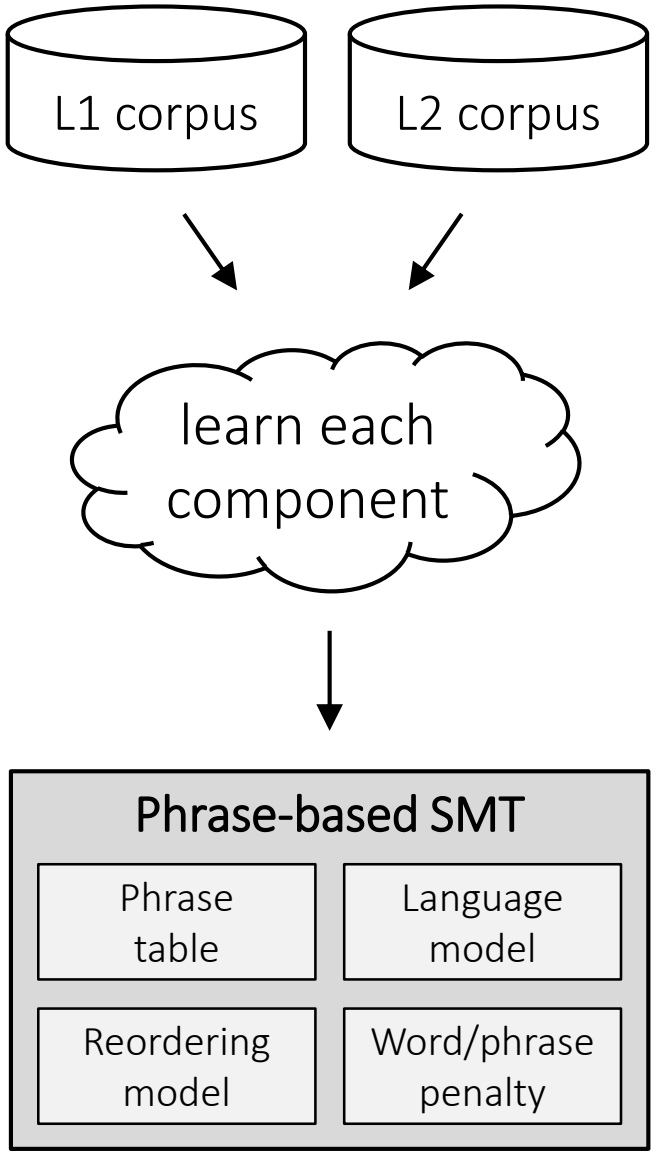
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



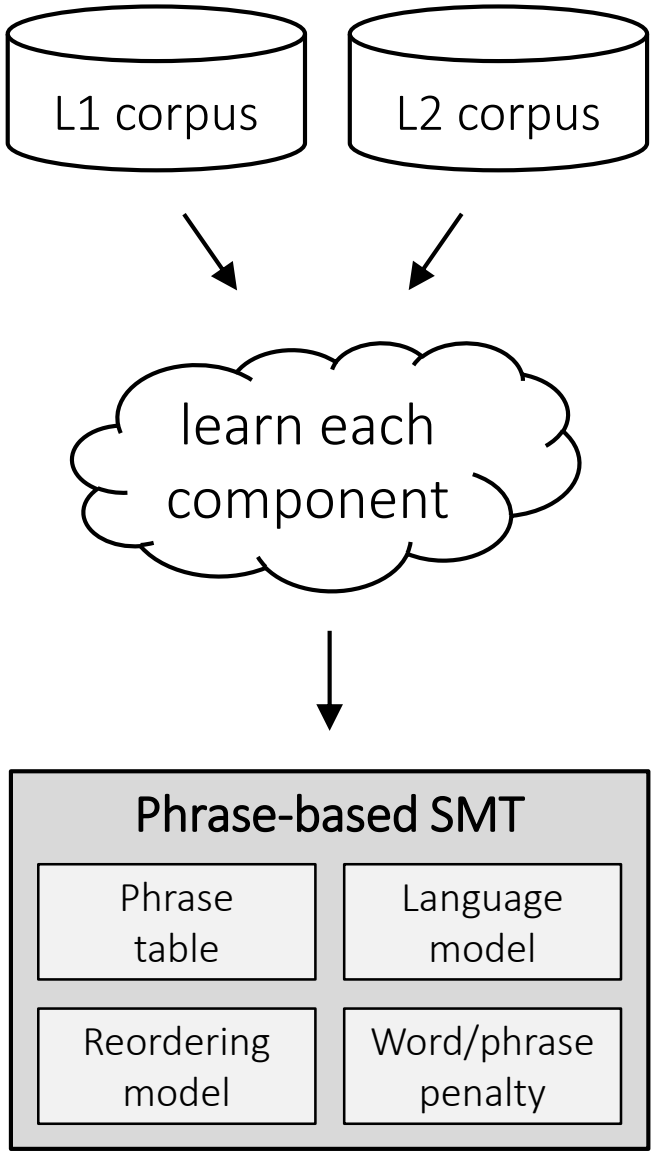
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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

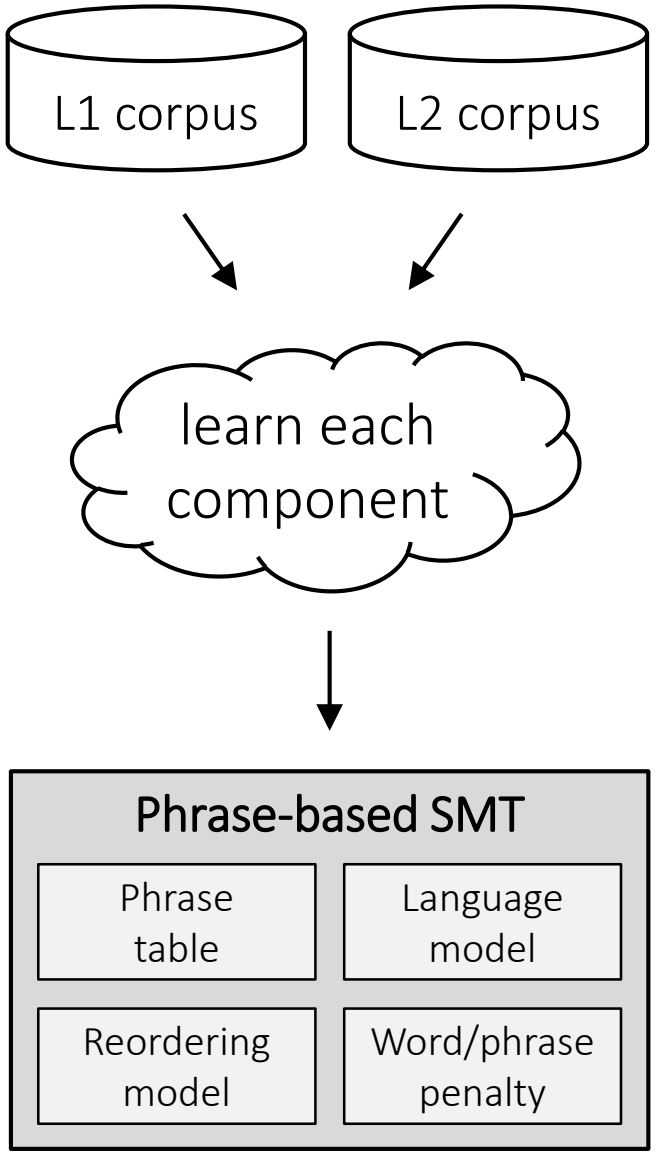


- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



Unsupervised phrase-based SMT

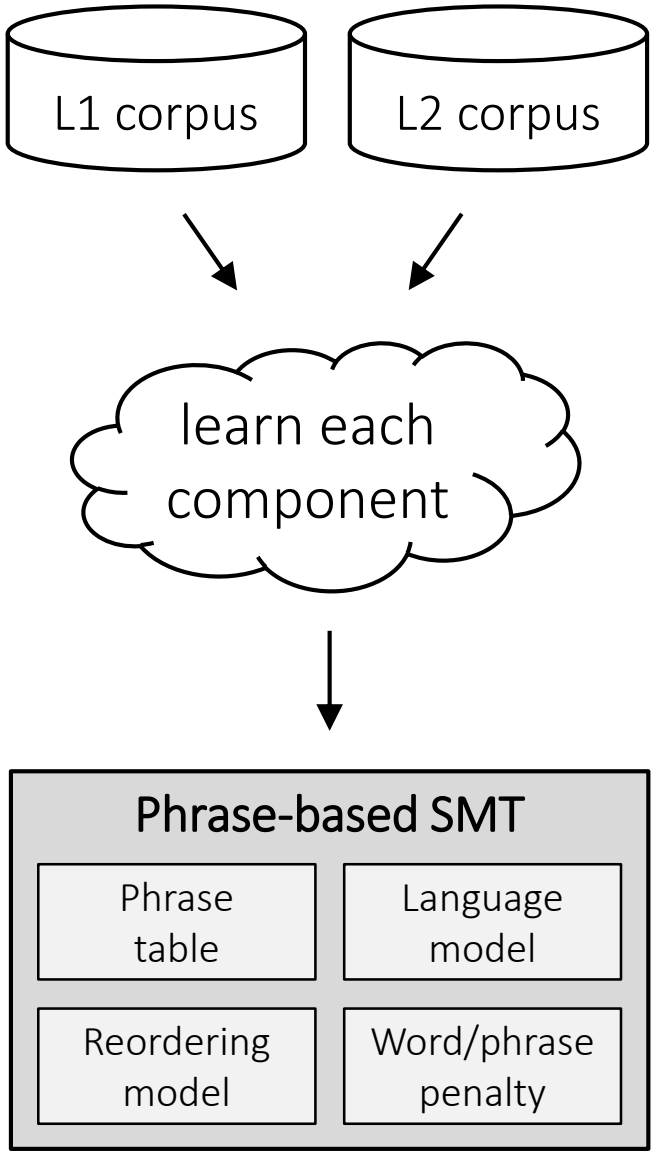
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

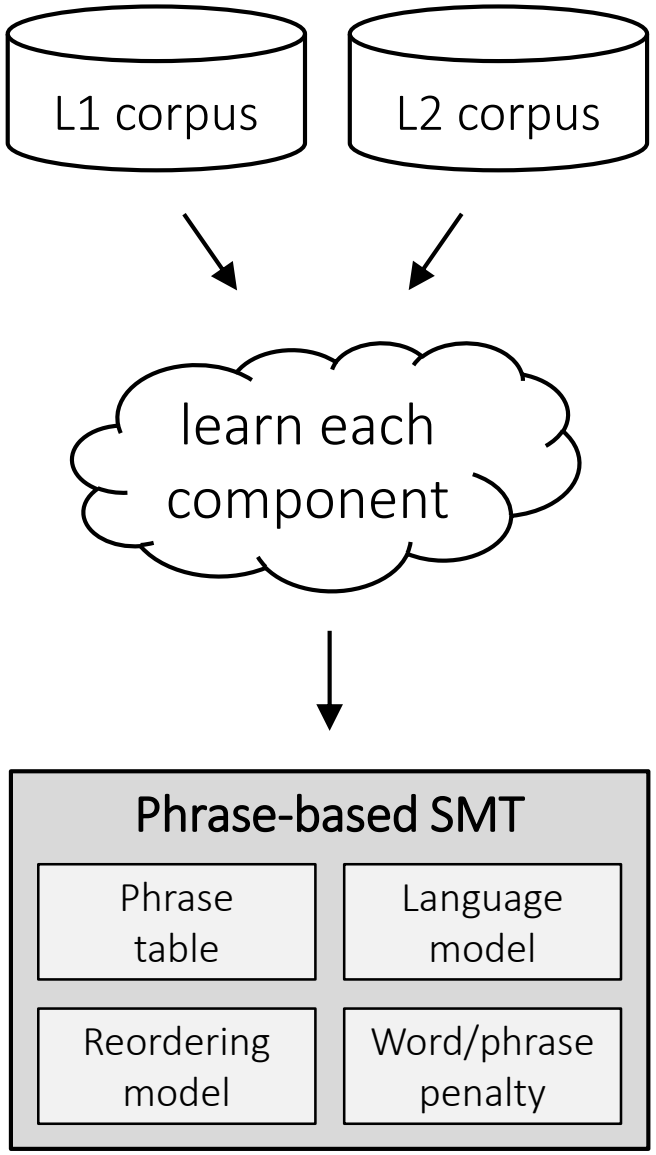
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

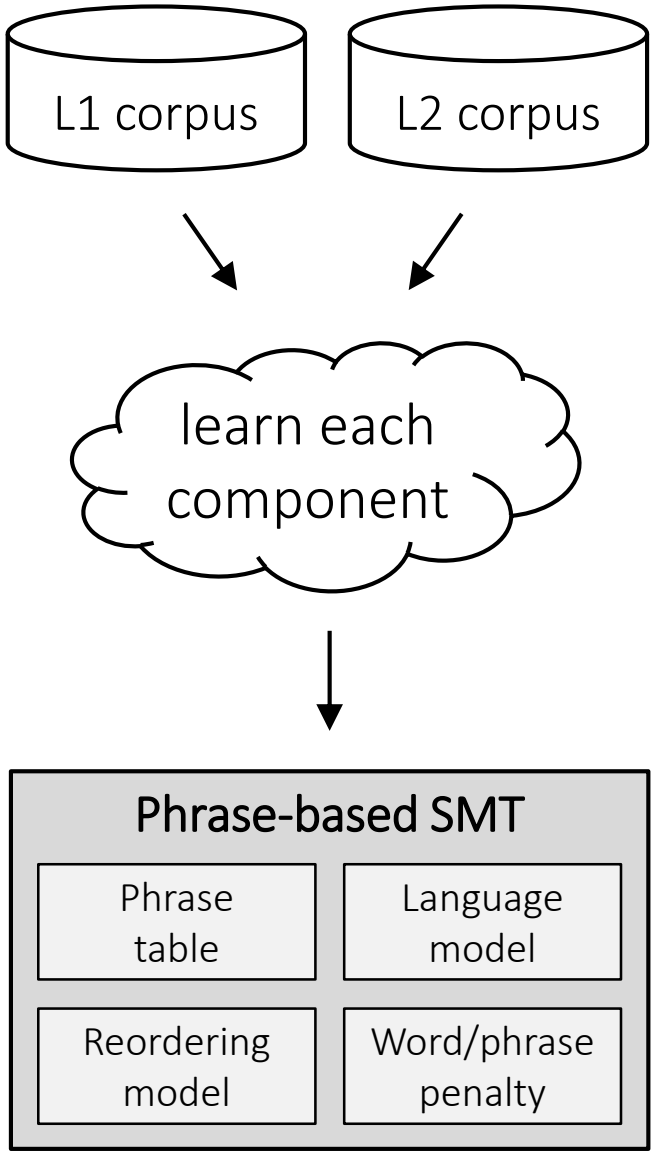
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

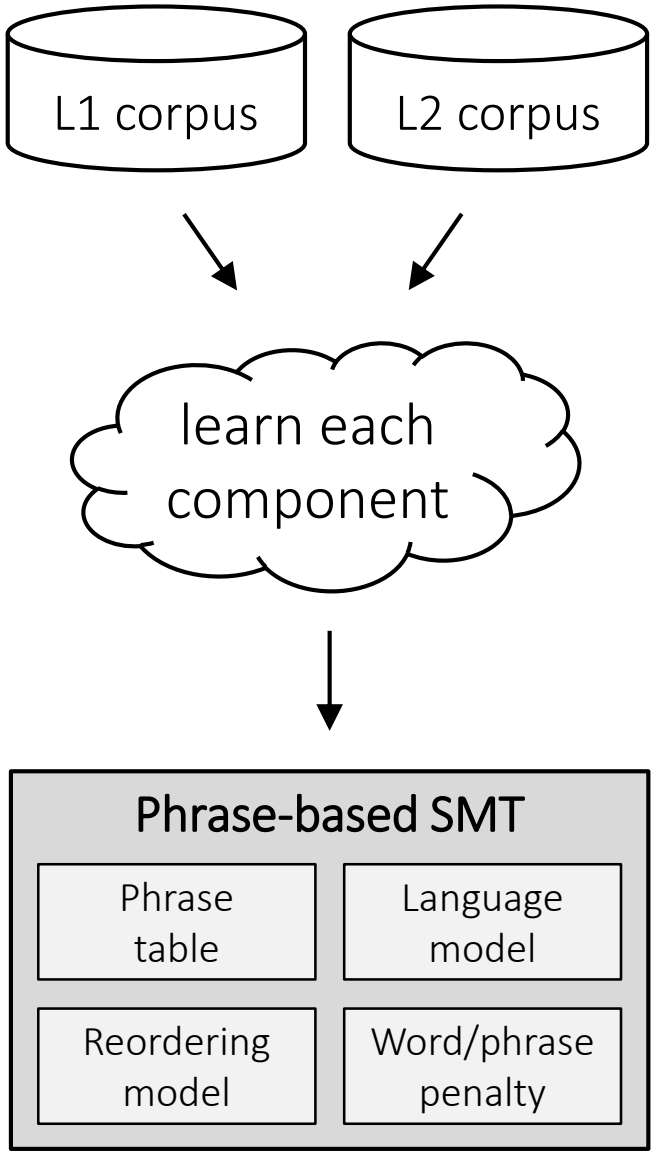
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty EASY!!!
  - Fixed score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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

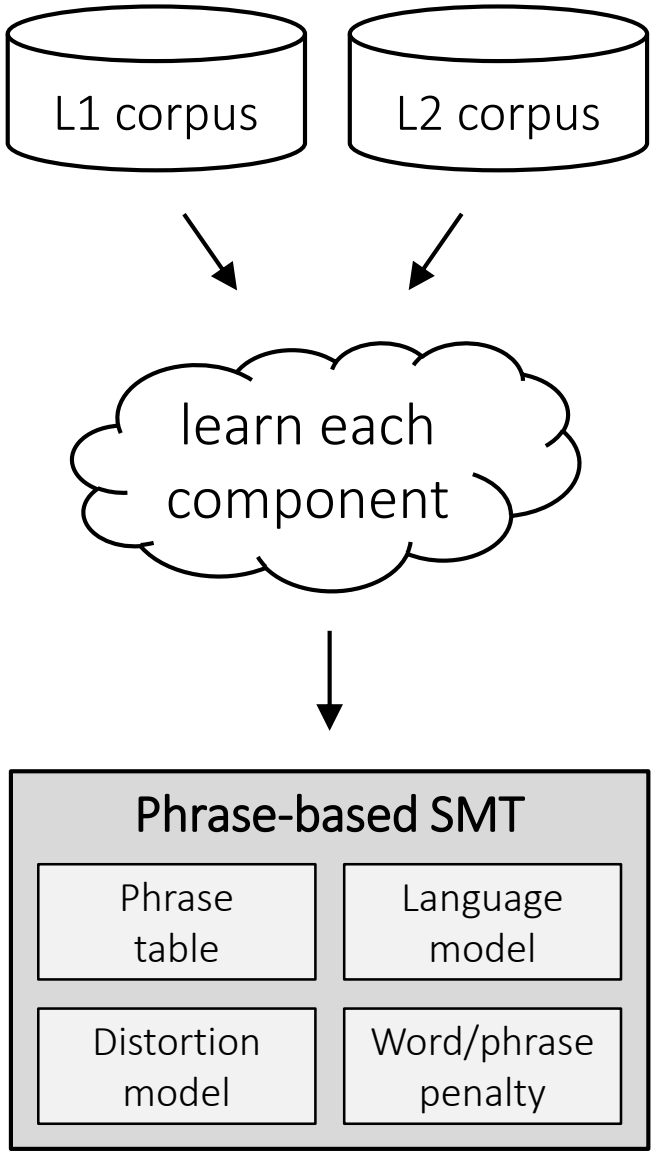
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty EASY!!!
  - Fixed score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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

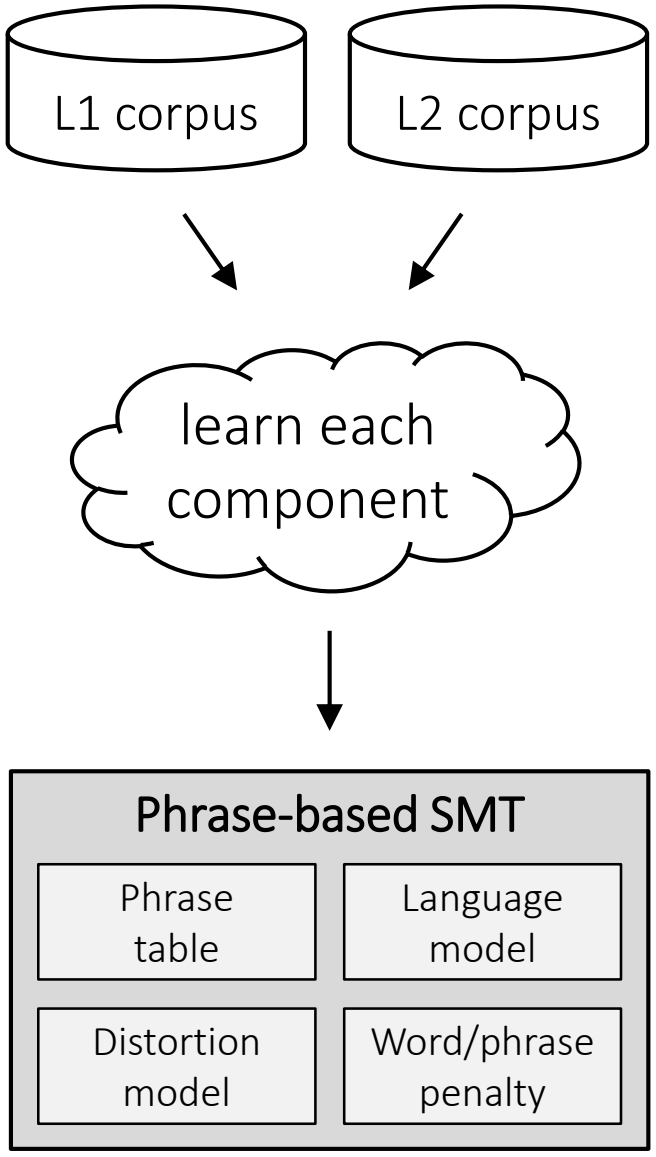
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!
  - Fixed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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

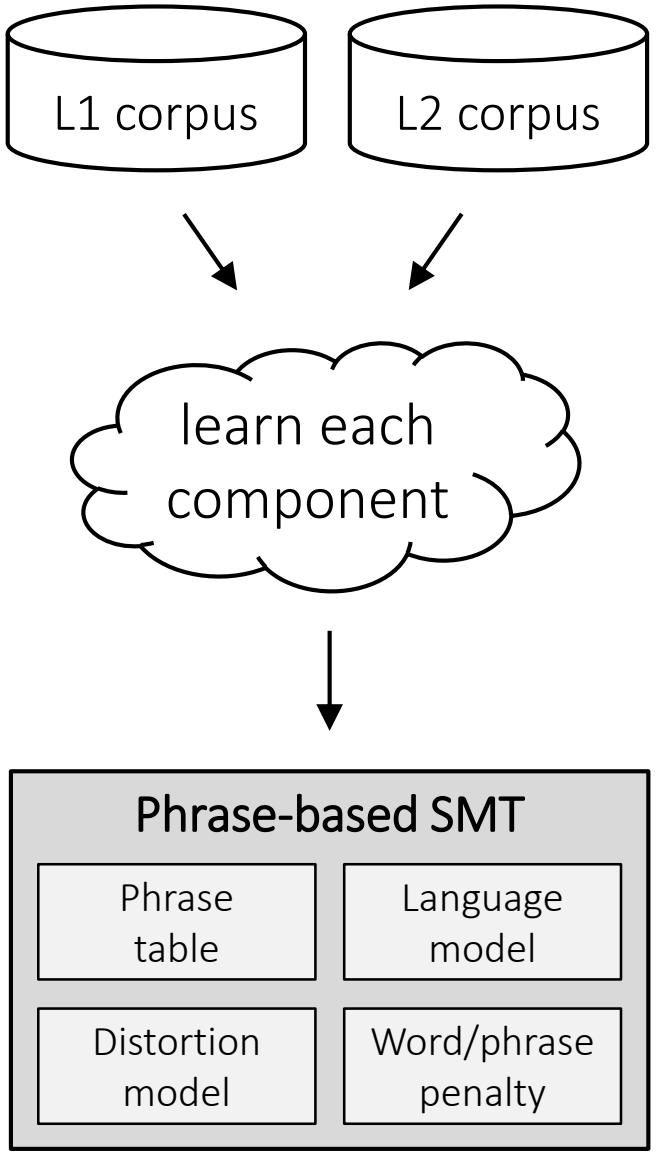
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!
  - Fixed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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

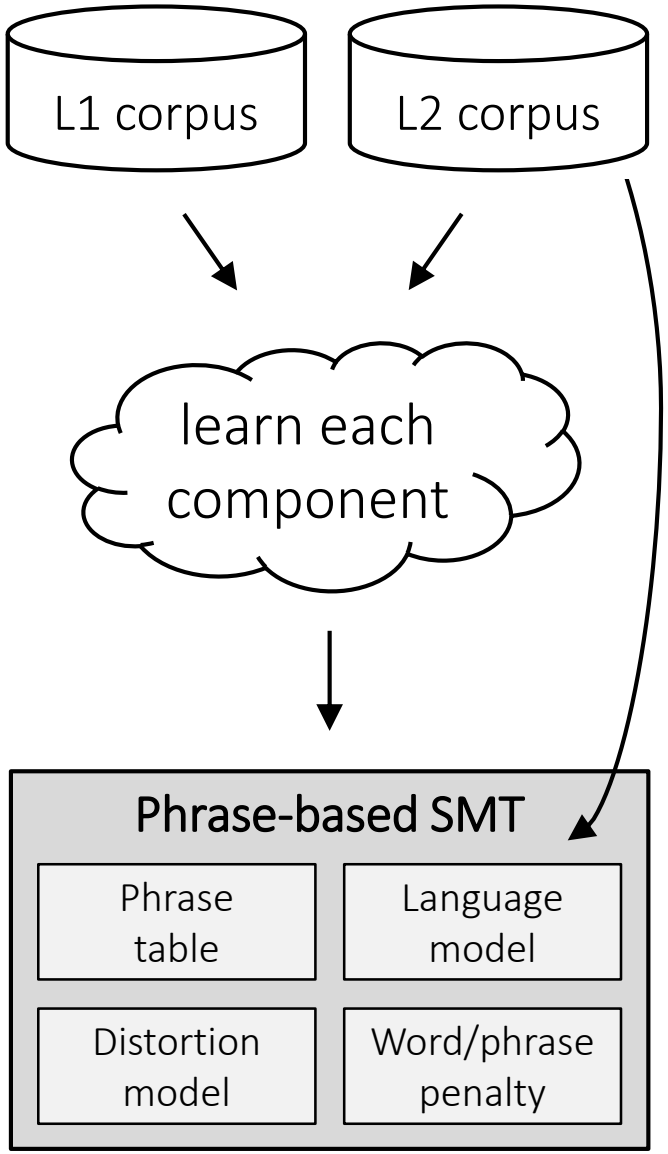
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!
 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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

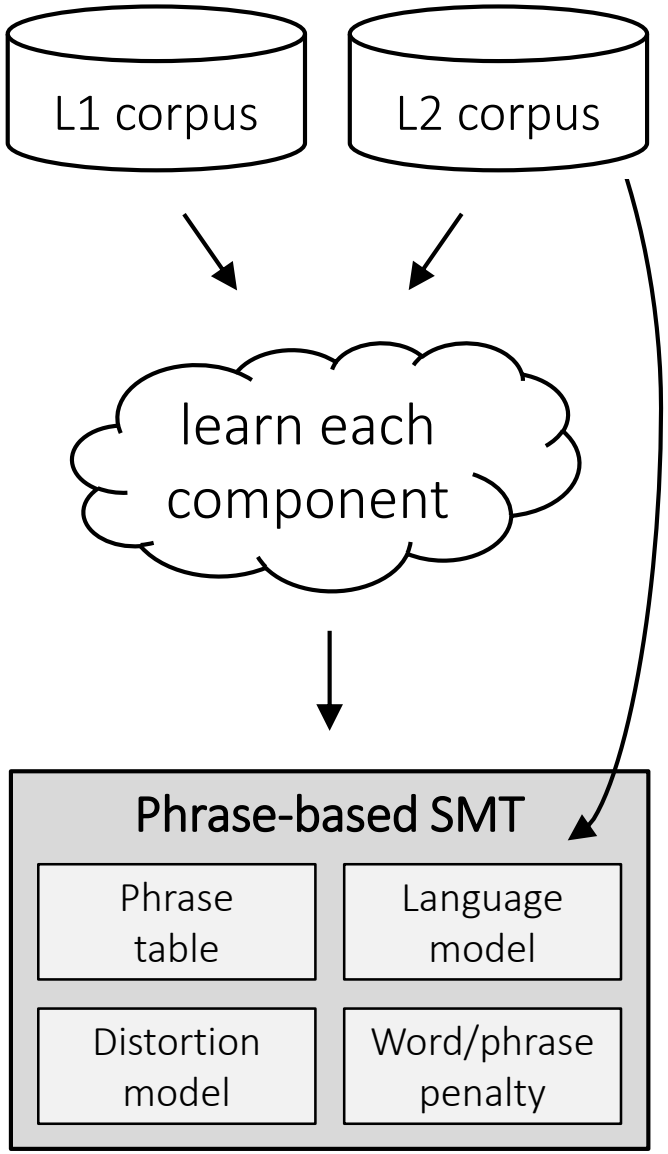
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!
 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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

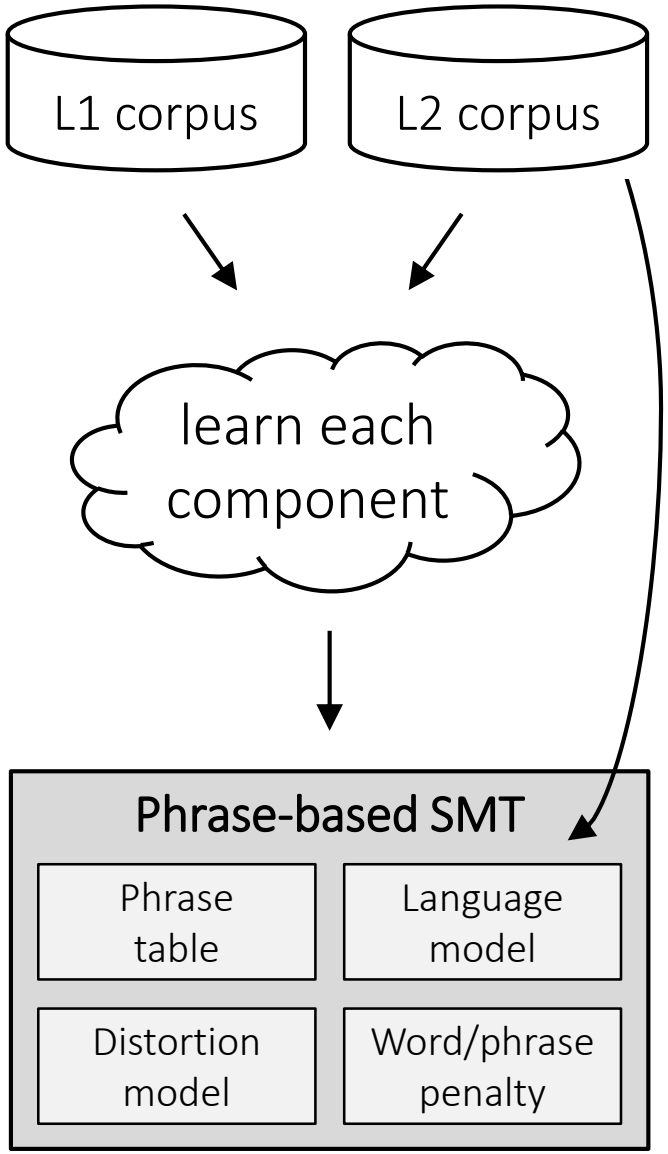
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!
 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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

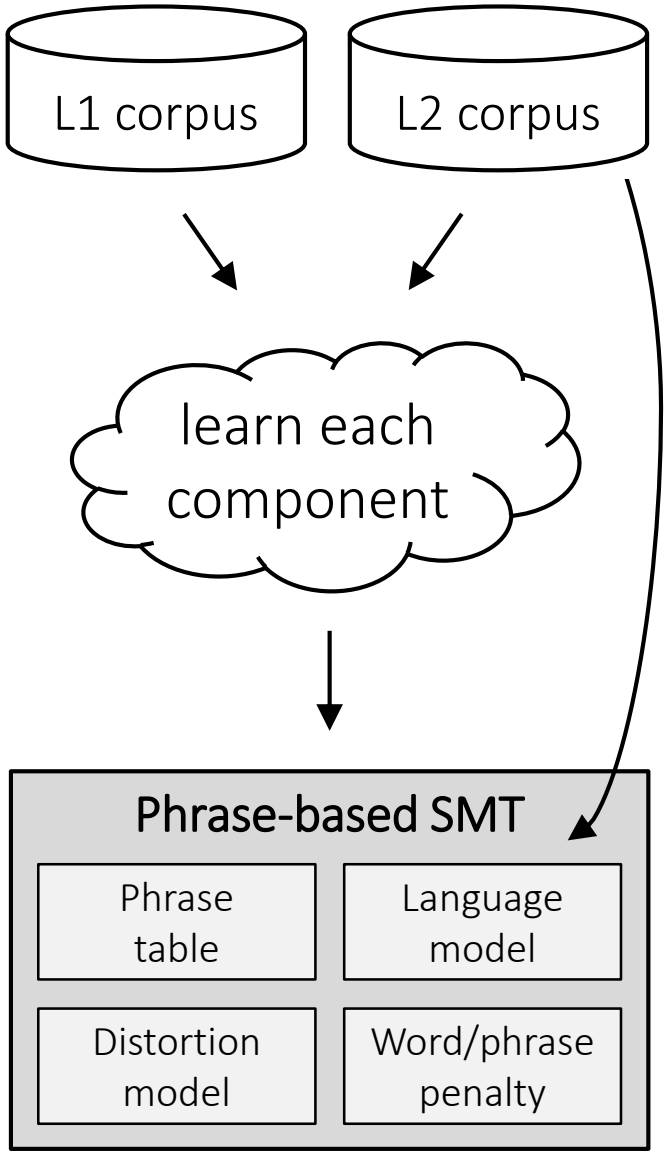
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

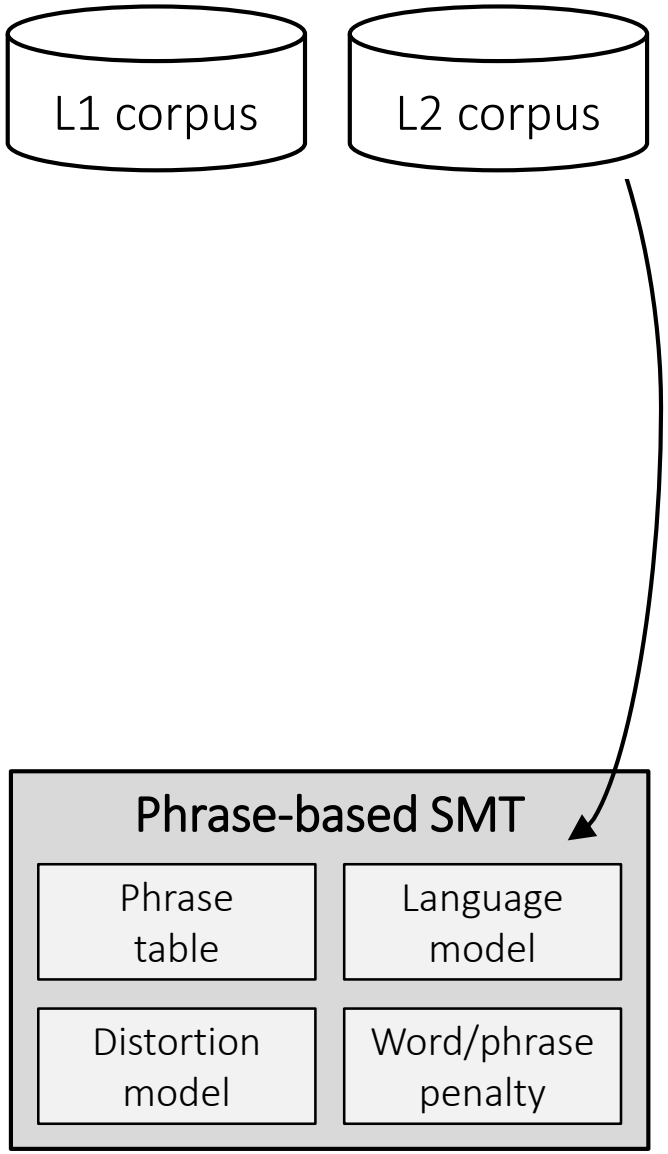
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

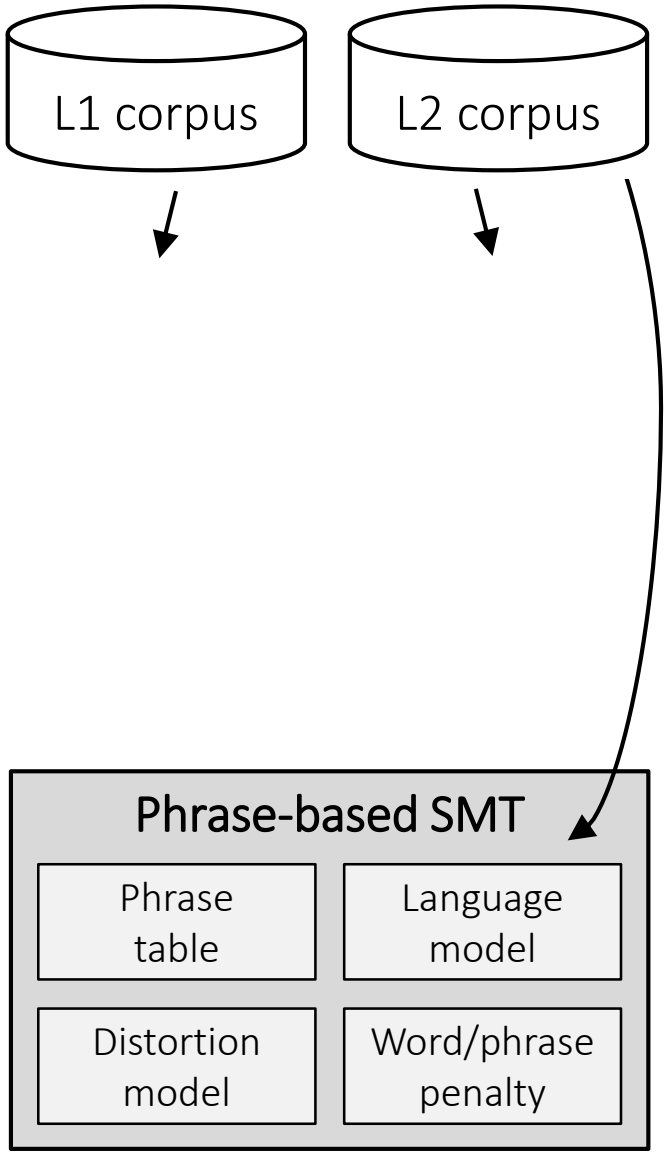
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

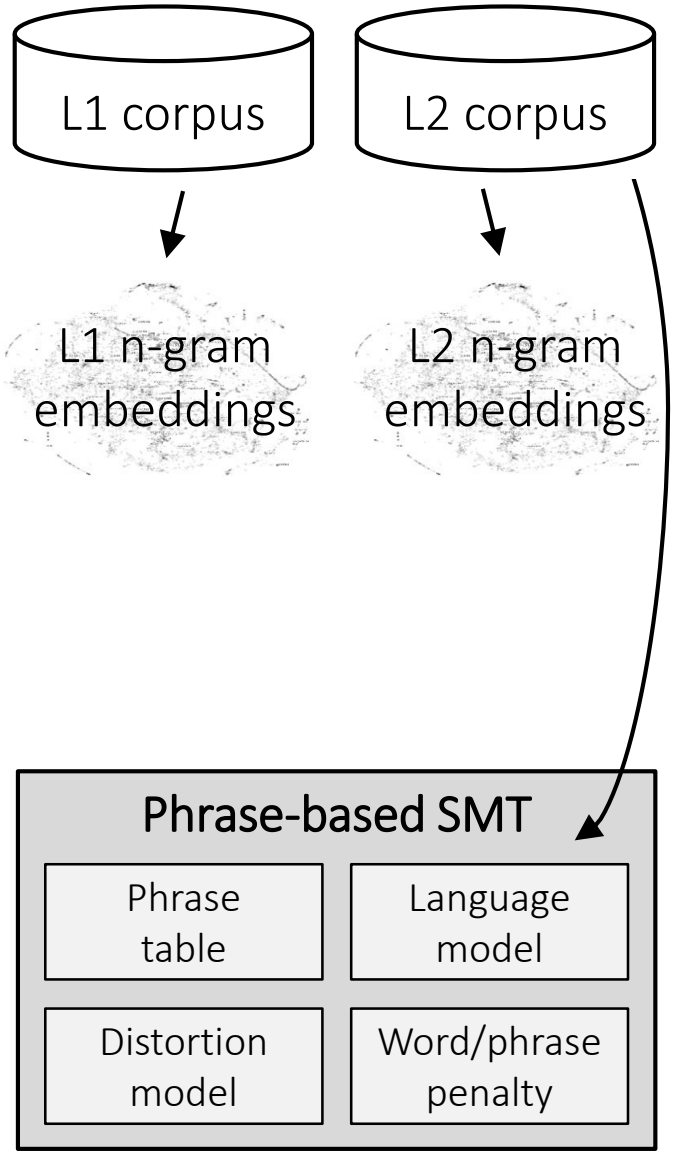
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

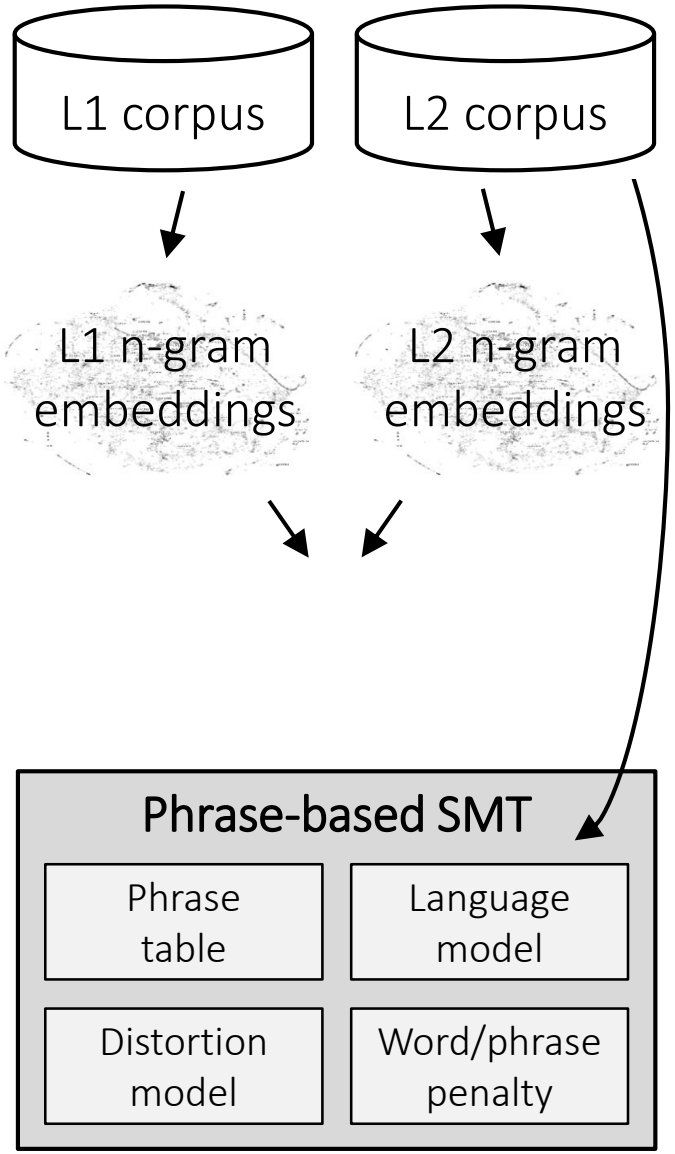
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

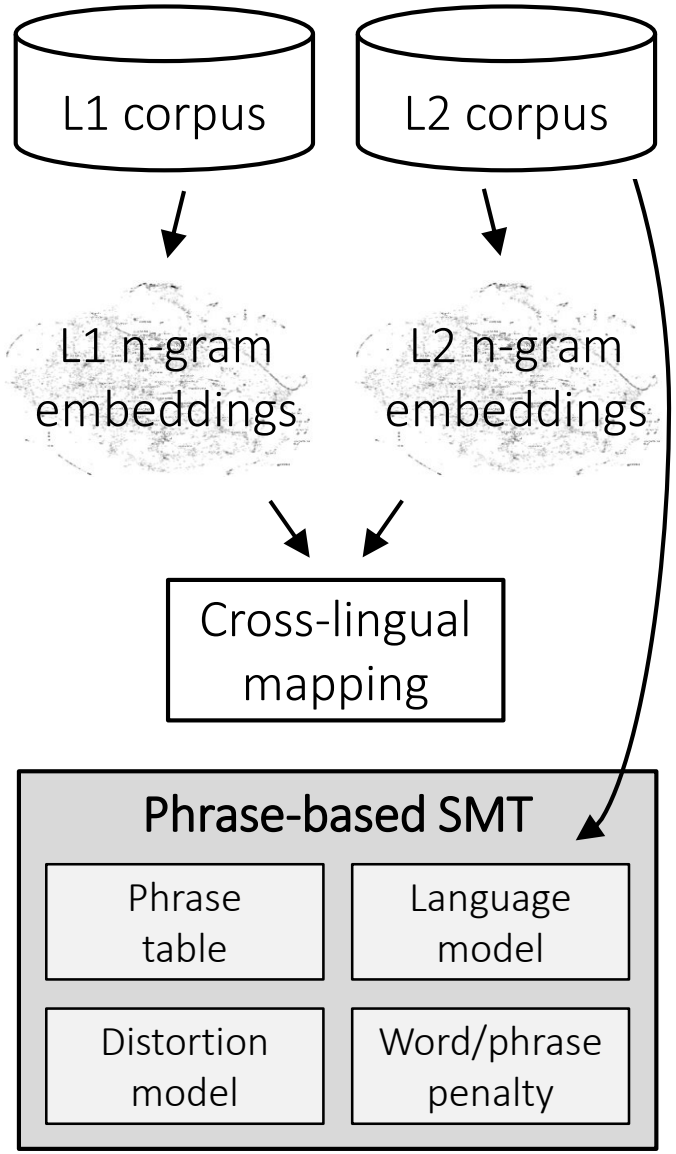
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

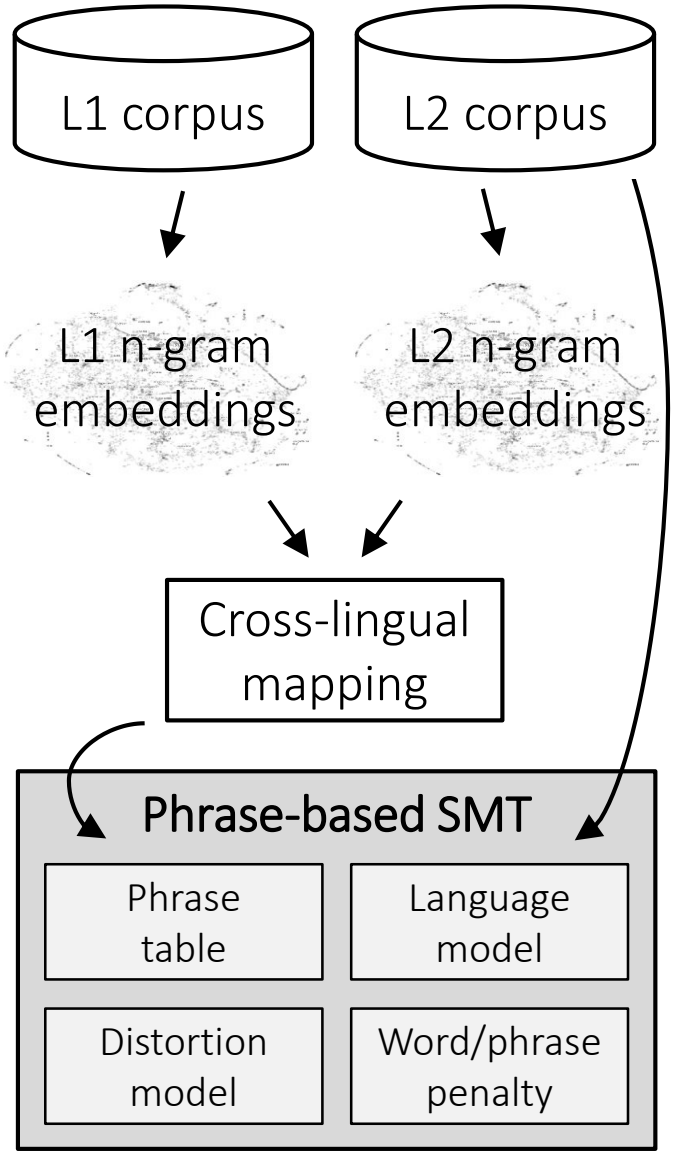
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

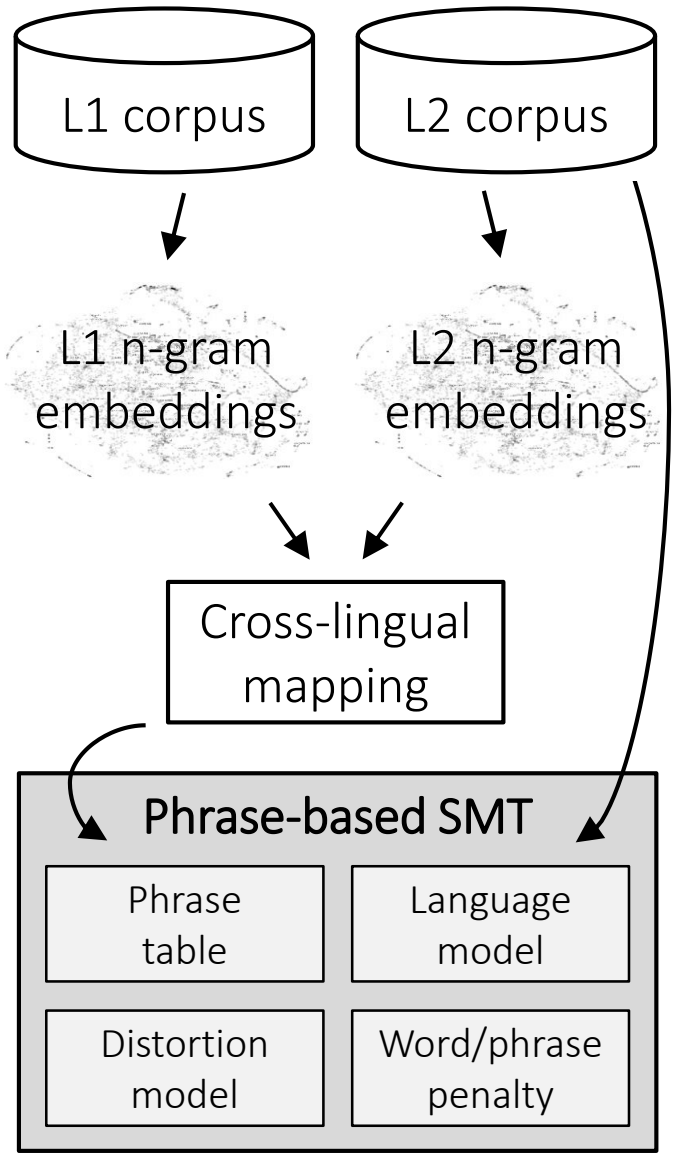
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

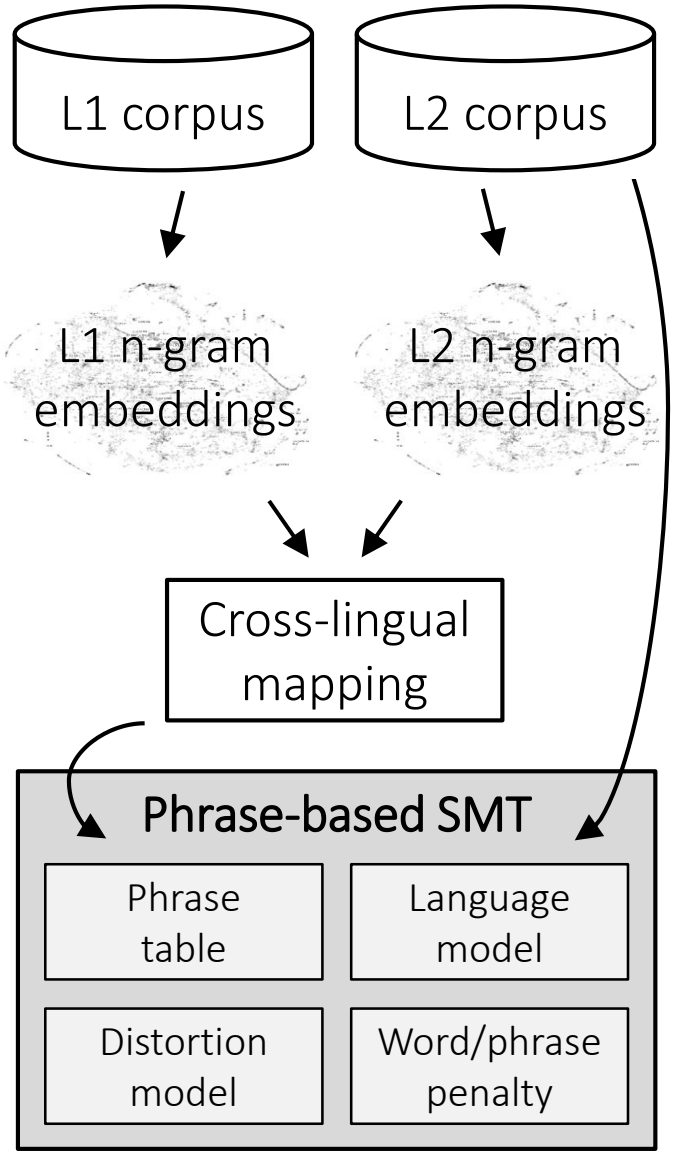
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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

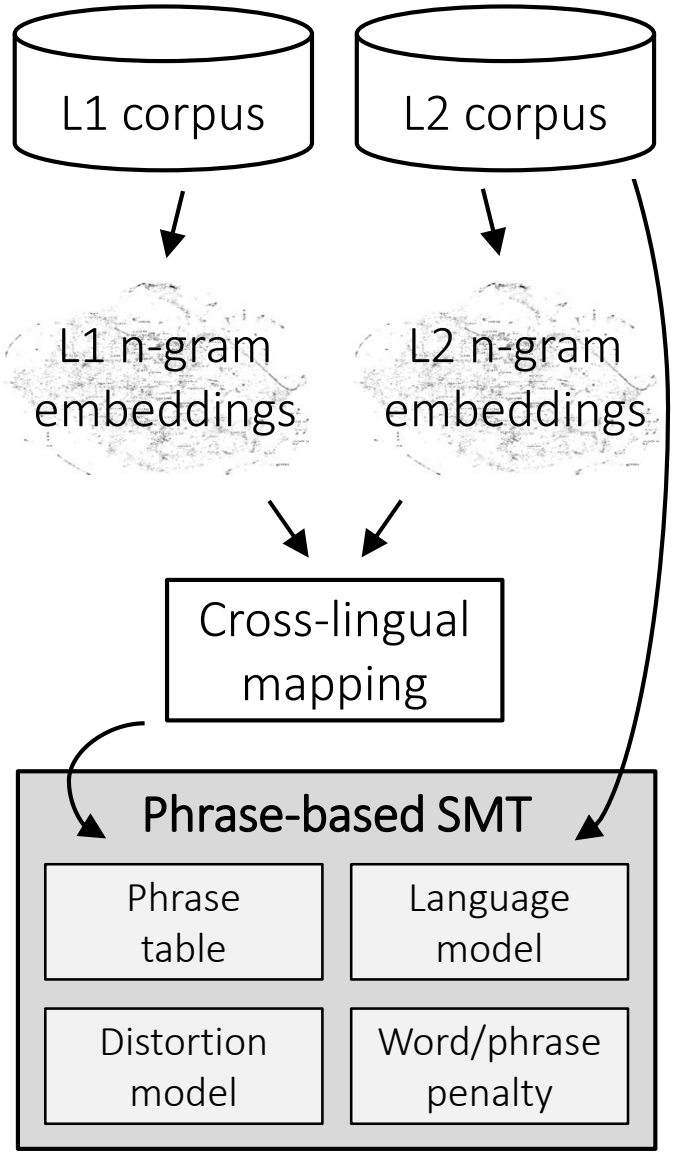


Unsupervised phrase-based SMT

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 .

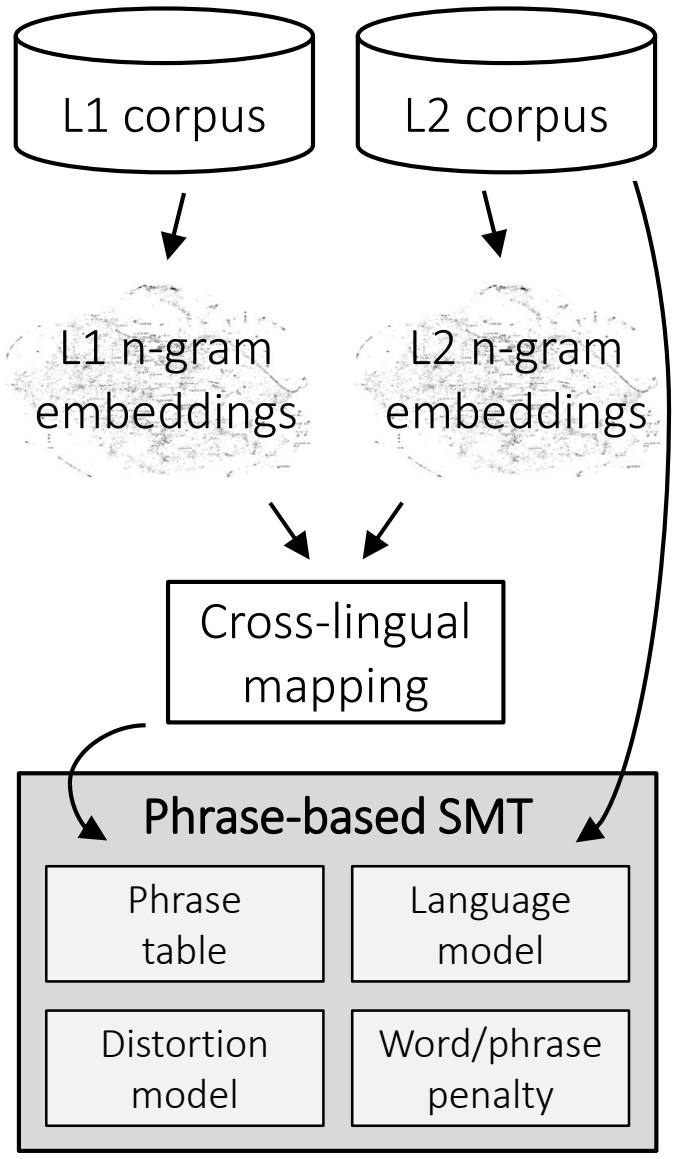


Unsupervised phrase-based SMT

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

$$I \ will \ \underset{w}{\underline{go}} \ to \ New \ York \ by \ plane \ .$$



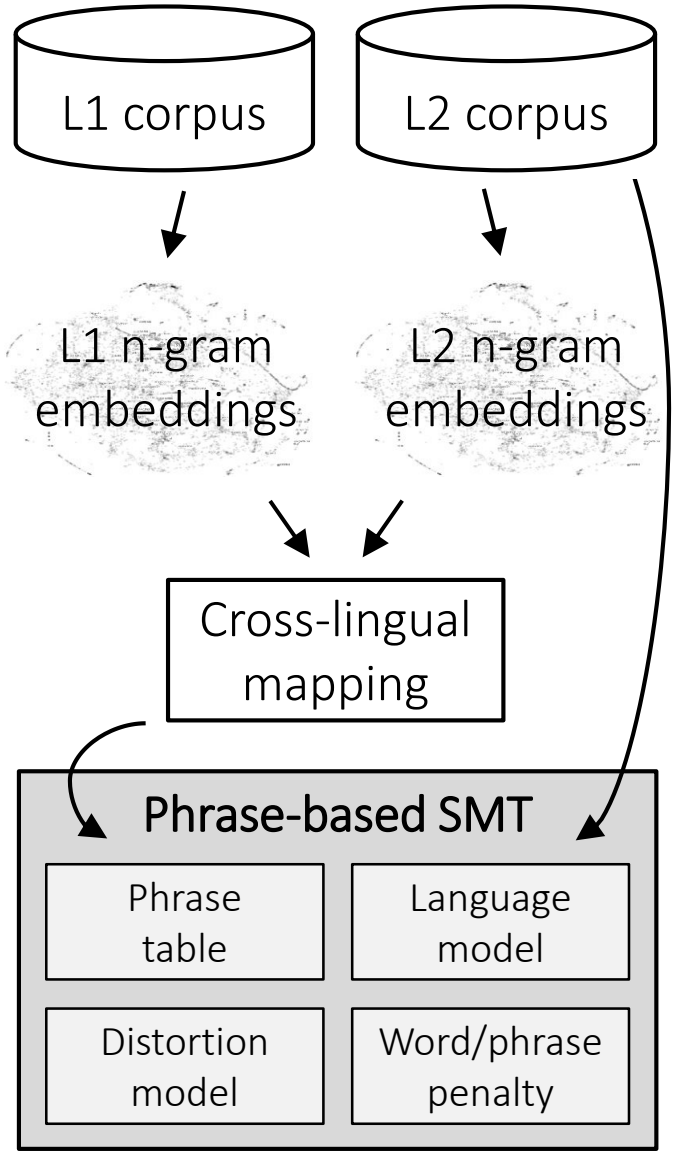
Unsupervised phrase-based SMT

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

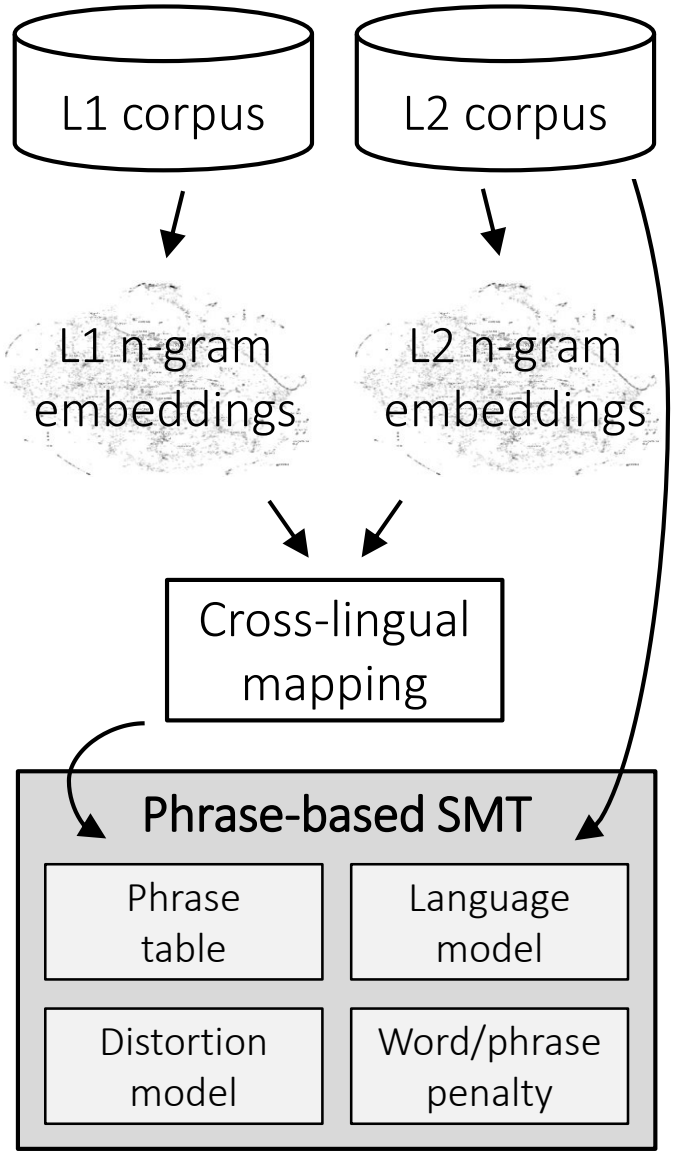


Unsupervised phrase-based SMT

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 $\underbrace{go}_{w}$ $\underbrace{to}_{c}$ *New York by plane .*



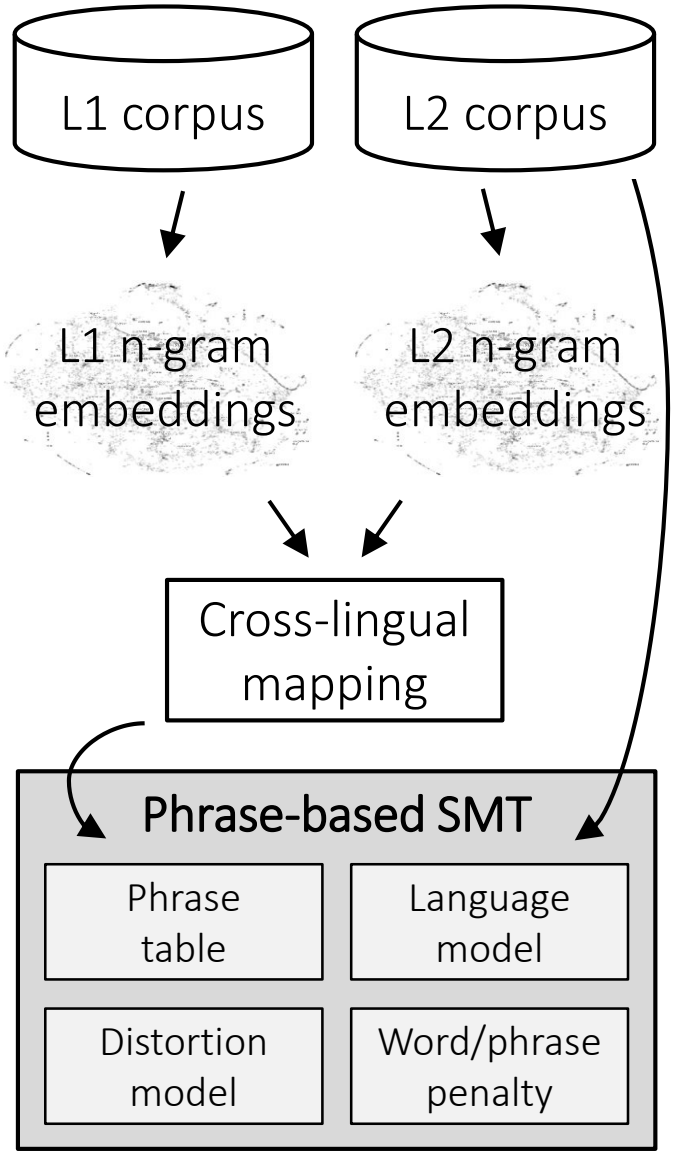
Unsupervised phrase-based SMT

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 c

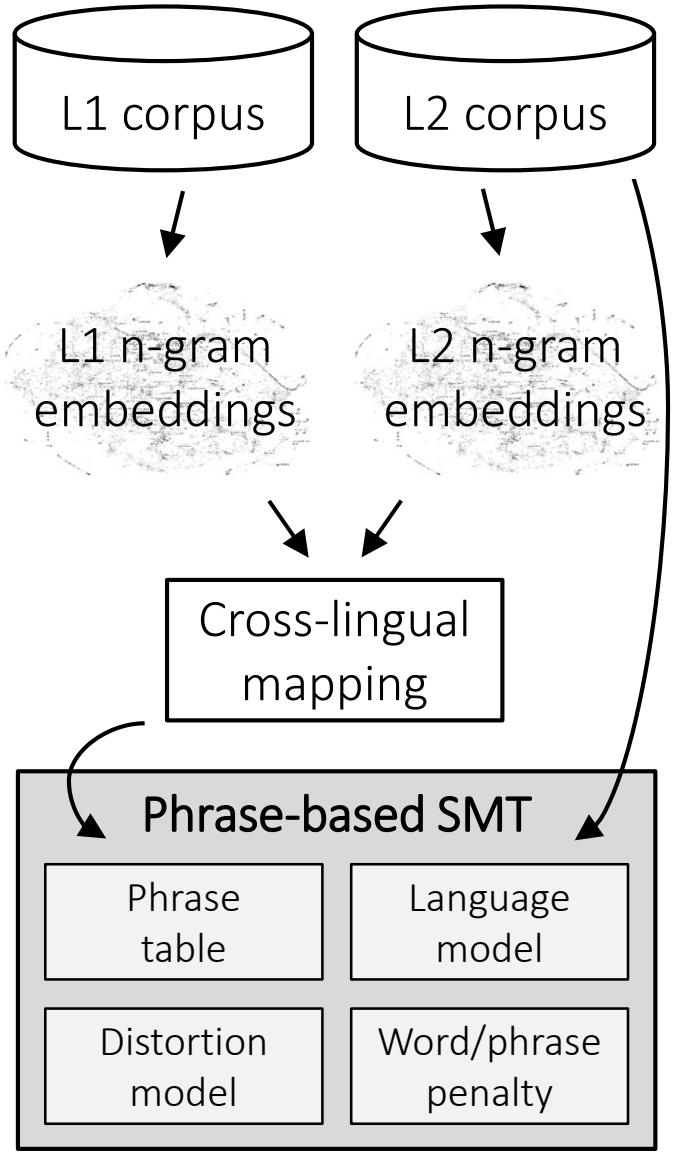


Unsupervised phrase-based SMT

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 $\underbrace{\textit{go}}_{w}$ *to New* $\underbrace{\textit{York}}_{c}$ *by plane .*



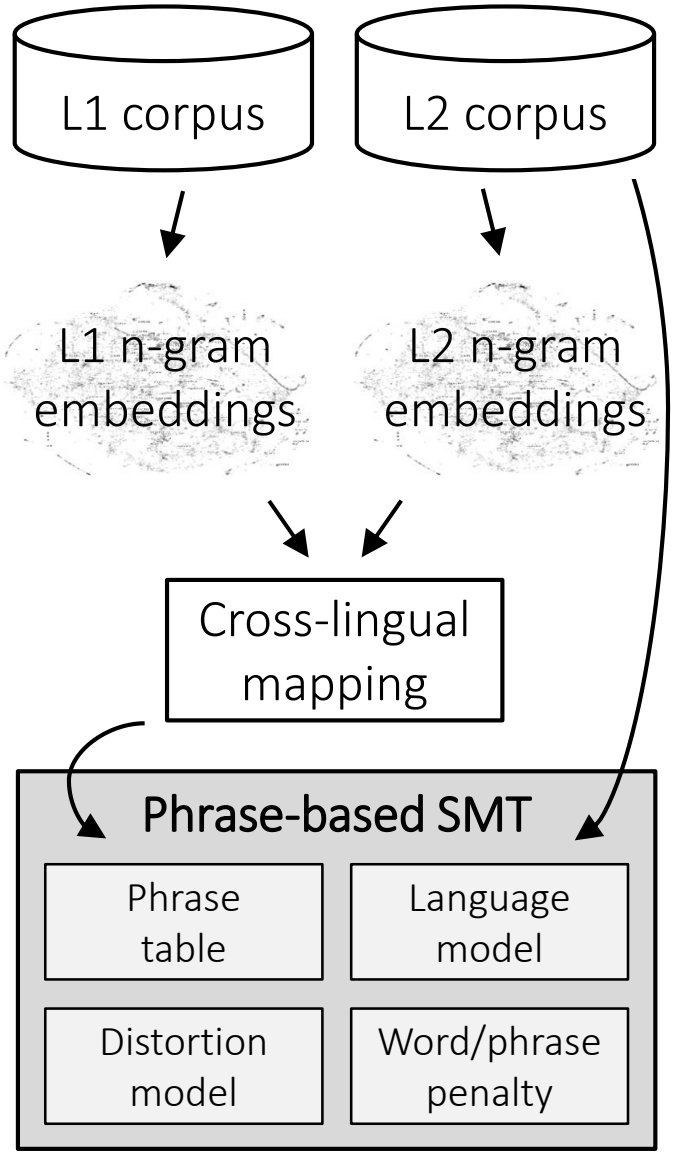
Unsupervised phrase-based SMT

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 .

p c



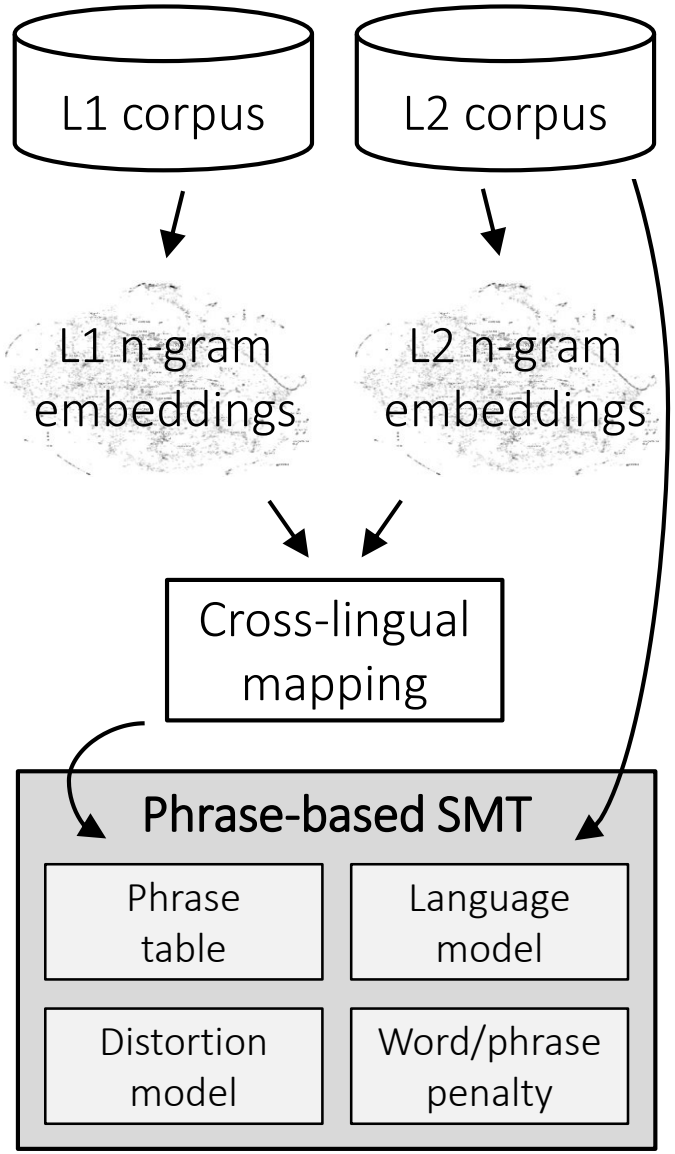
Unsupervised phrase-based SMT

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



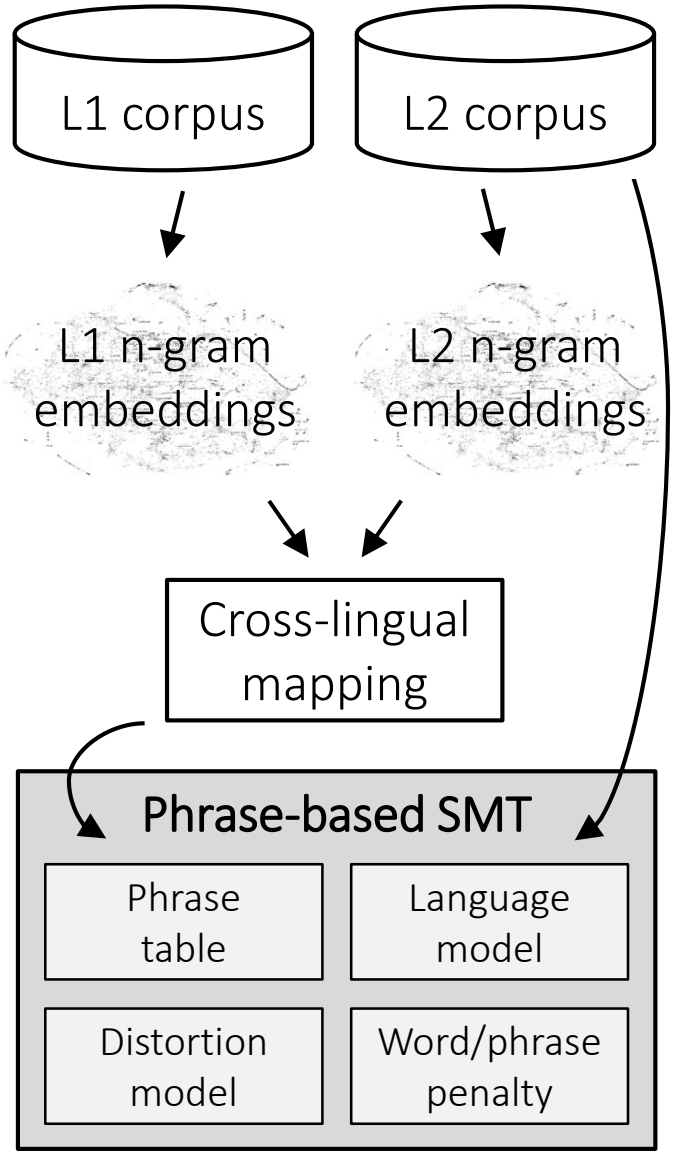
Unsupervised phrase-based SMT

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

I will go to New York by plane .

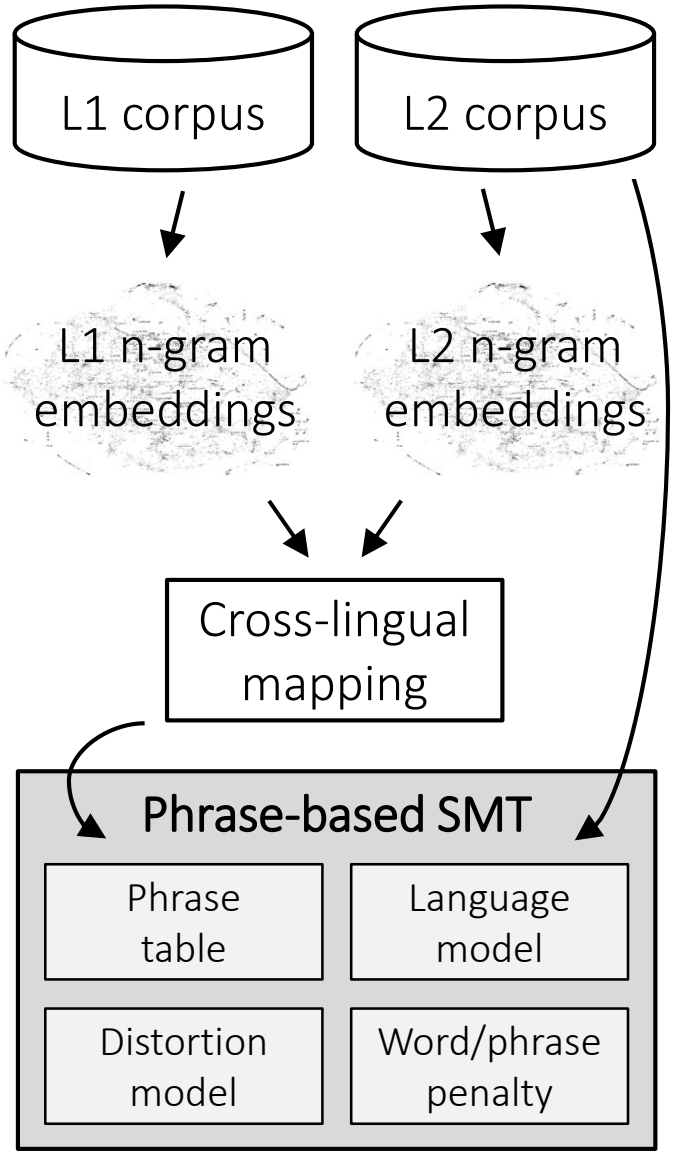
p c



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

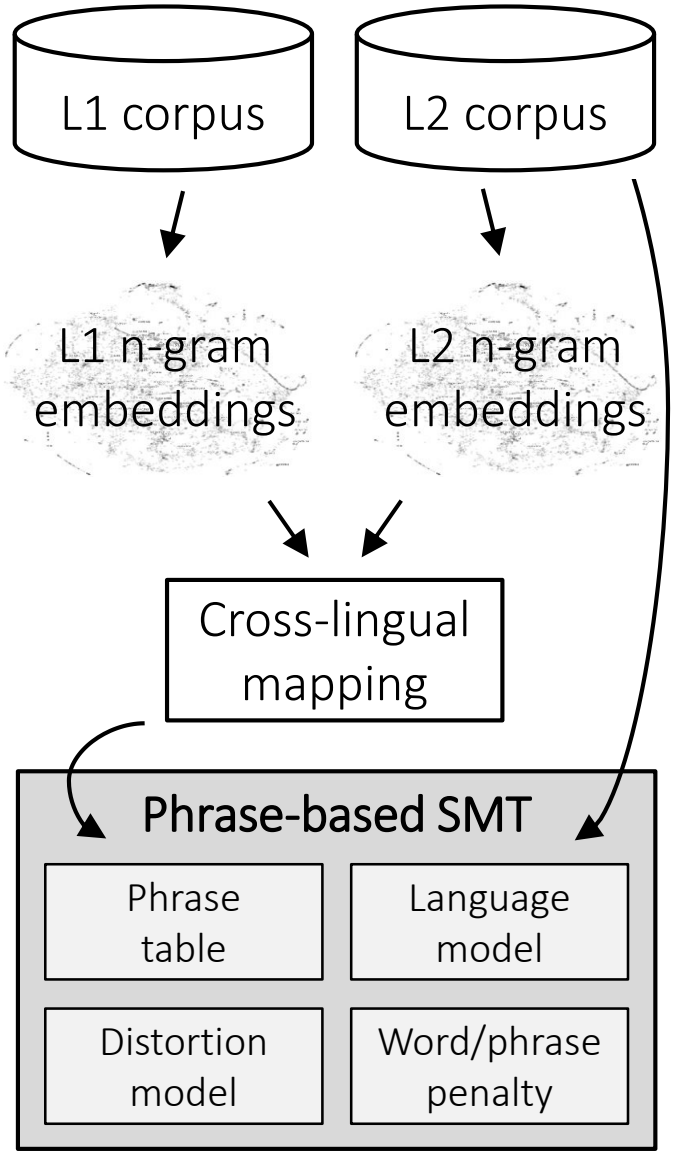


Unsupervised phrase-based SMT

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:



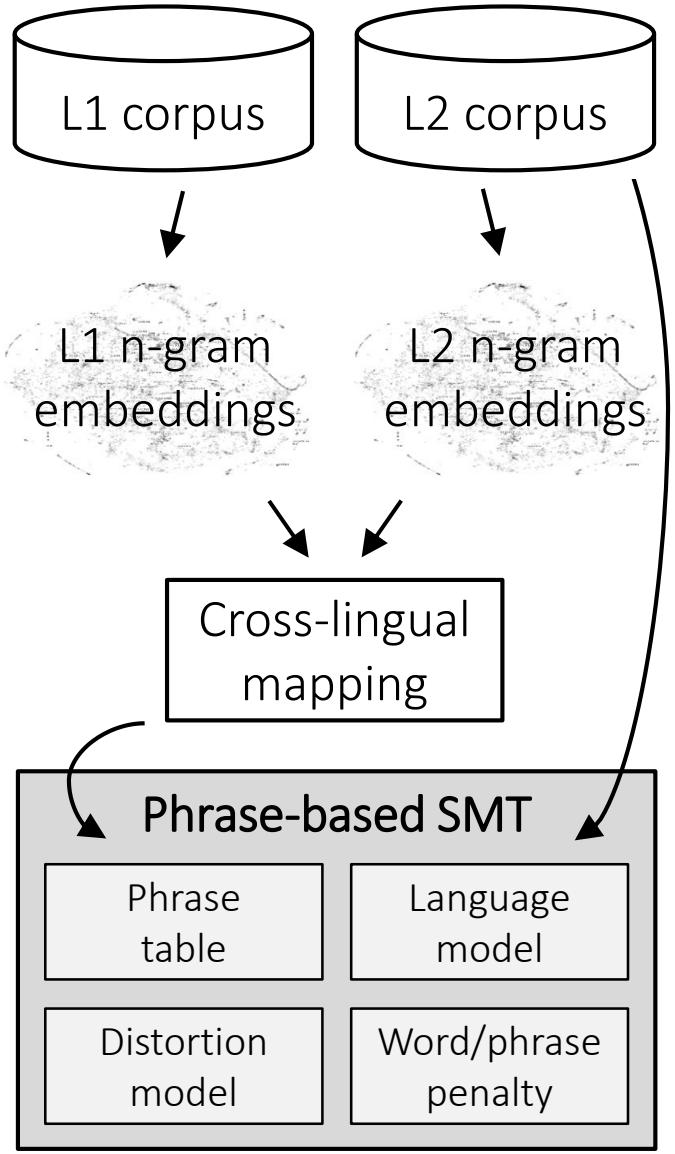
Unsupervised phrase-based SMT

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:

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



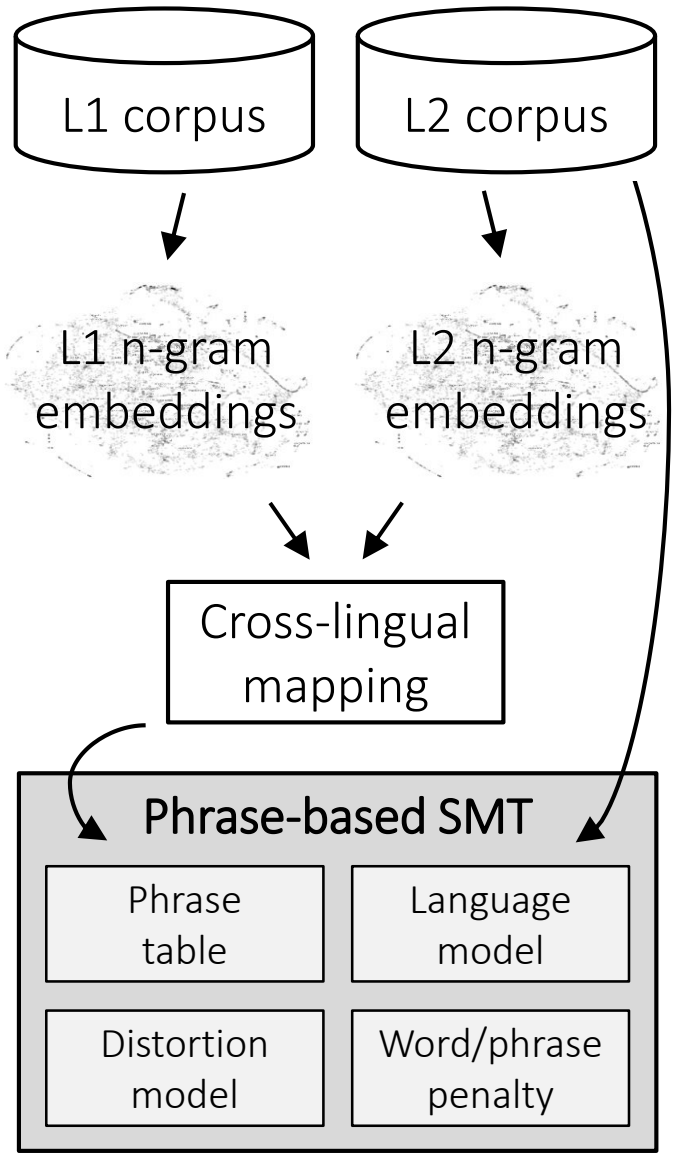
Unsupervised phrase-based SMT

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:

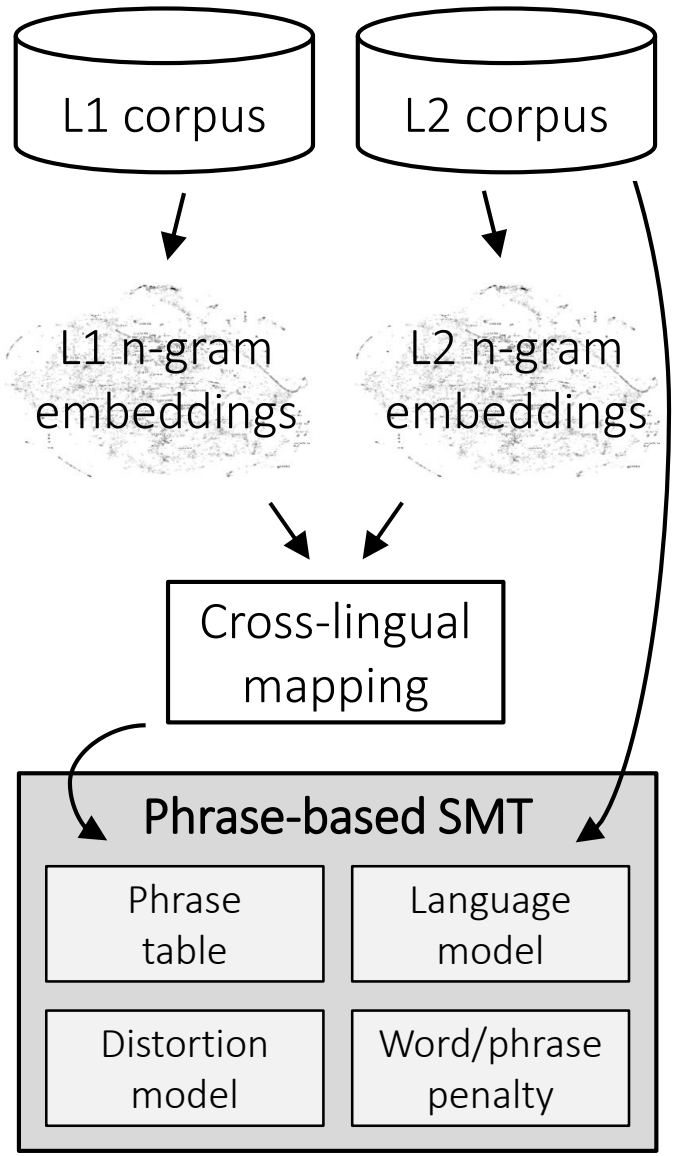
$$\phi(\bar{f}|\bar{e}) = \frac{e^{\cos(\bar{e},\bar{f})/\tau}}{\sum_{\bar{f}'} e^{\cos(\bar{e},\bar{f}')/\tau}} \qquad \min_{\tau} \sum_{\bar{f}} \log \phi\left(\bar{f}|\mathrm{NN}_{\bar{e}}(\bar{f})\right) + \sum_{\bar{e}} \log \phi\left(\bar{e}|\mathrm{NN}_{\bar{f}}(\bar{e})\right)$$



Unsupervised phrase-based SMT

Learn components from monolingual corpora

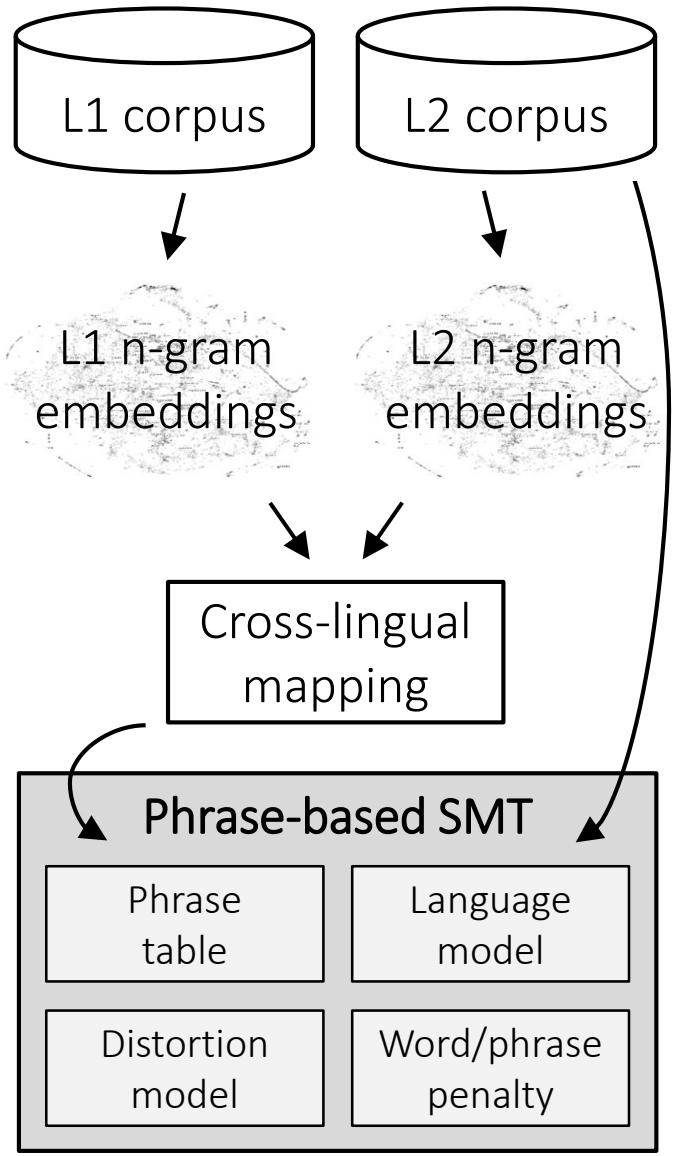
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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

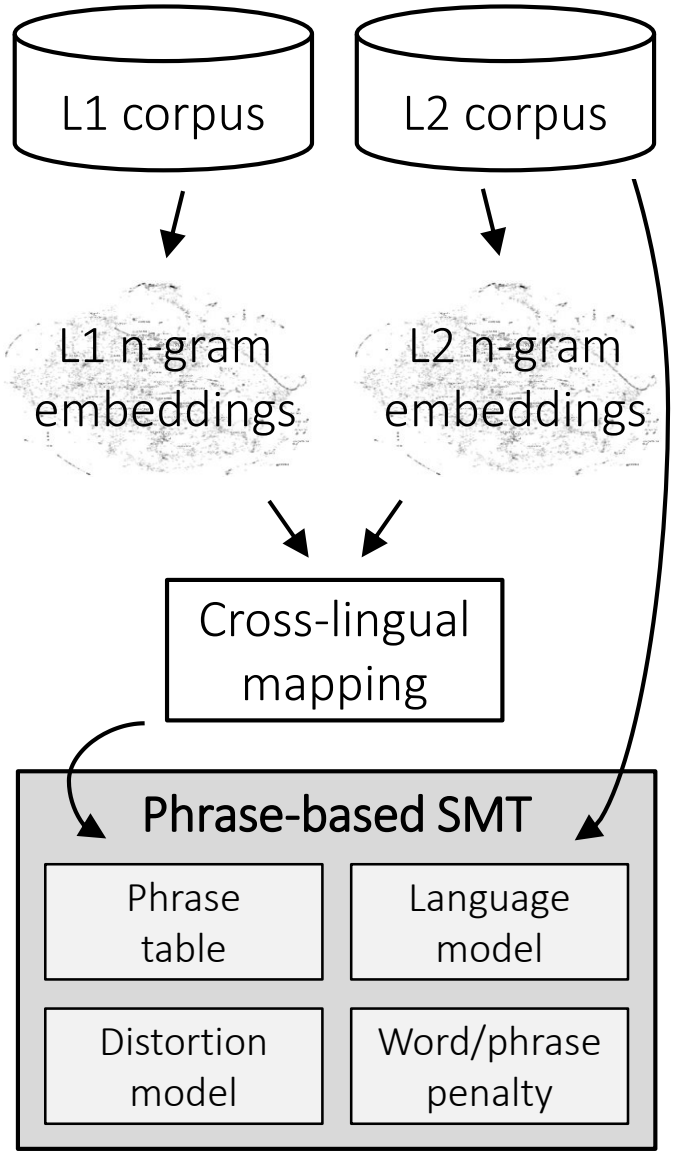


Unsupervised phrase-based SMT

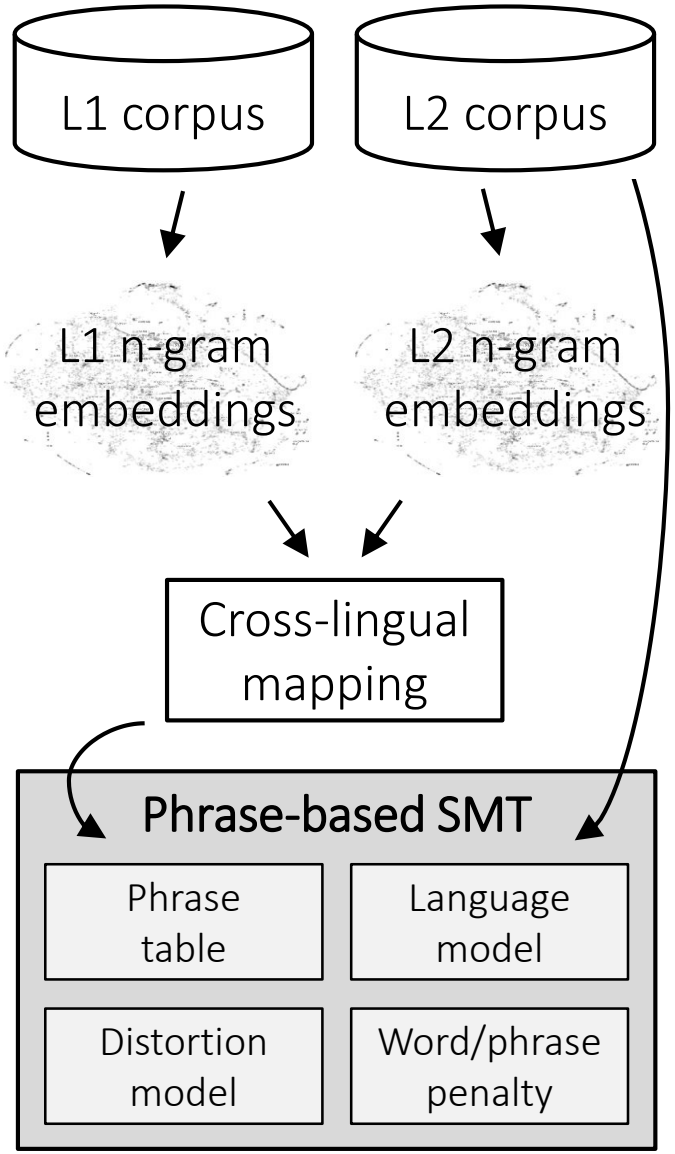
Learn components from monolingual corpora

- 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

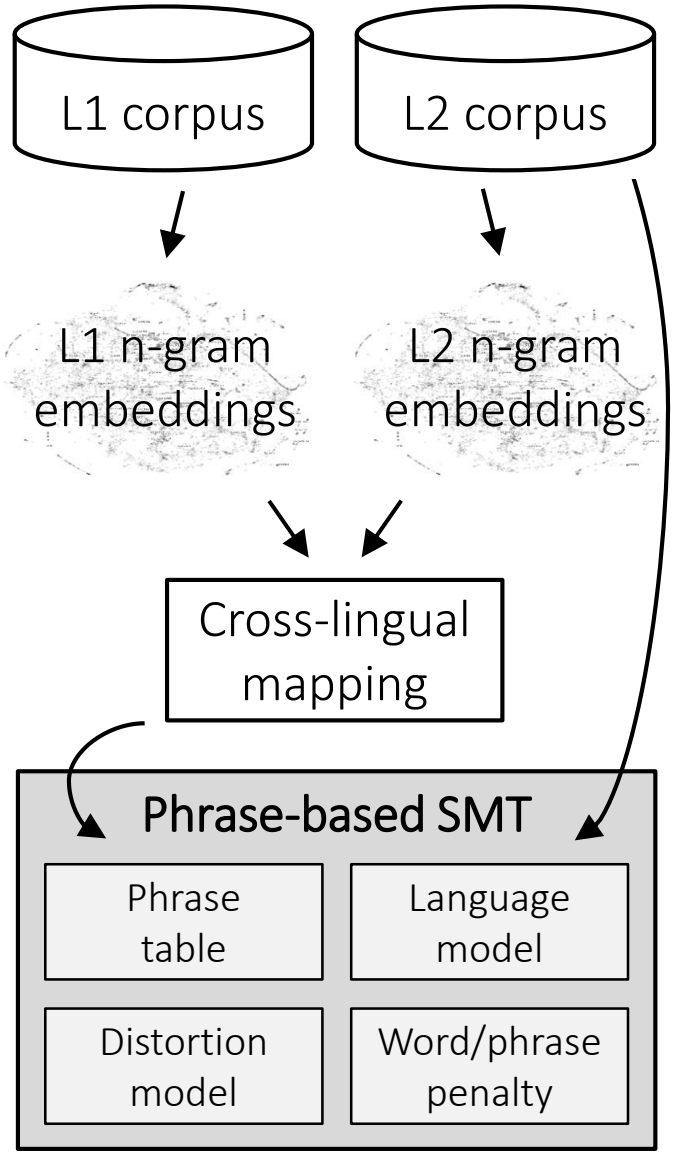
Unsupervised phrase-based SMT



Unsupervised phrase-based SMT



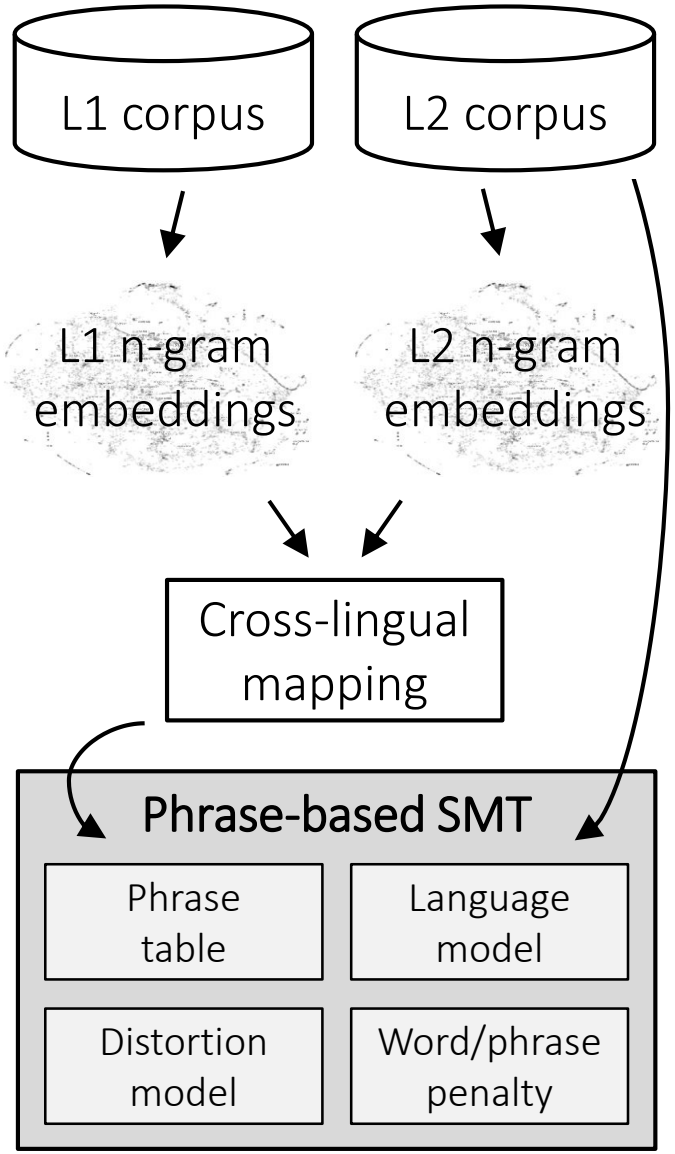
EXPERIMENTS



Unsupervised phrase-based SMT

EXPERIMENTS

- Languages: French-English, German-English



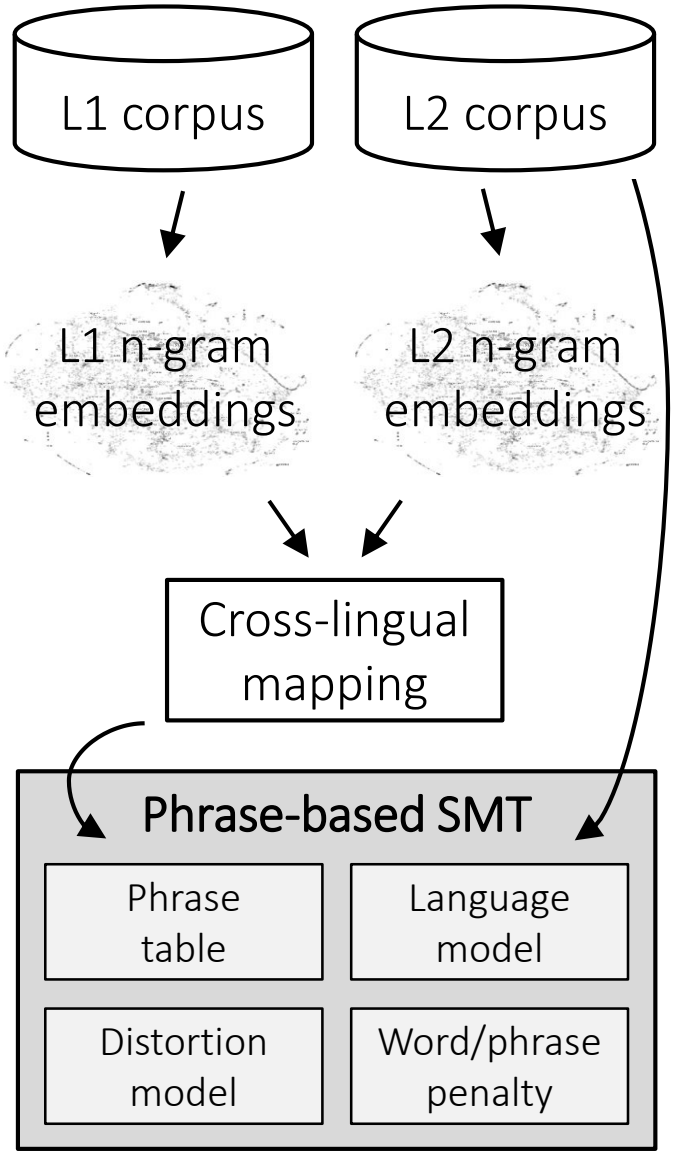
Unsupervised phrase-based SMT

EXPERIMENTS

- Languages: French-English, German-English

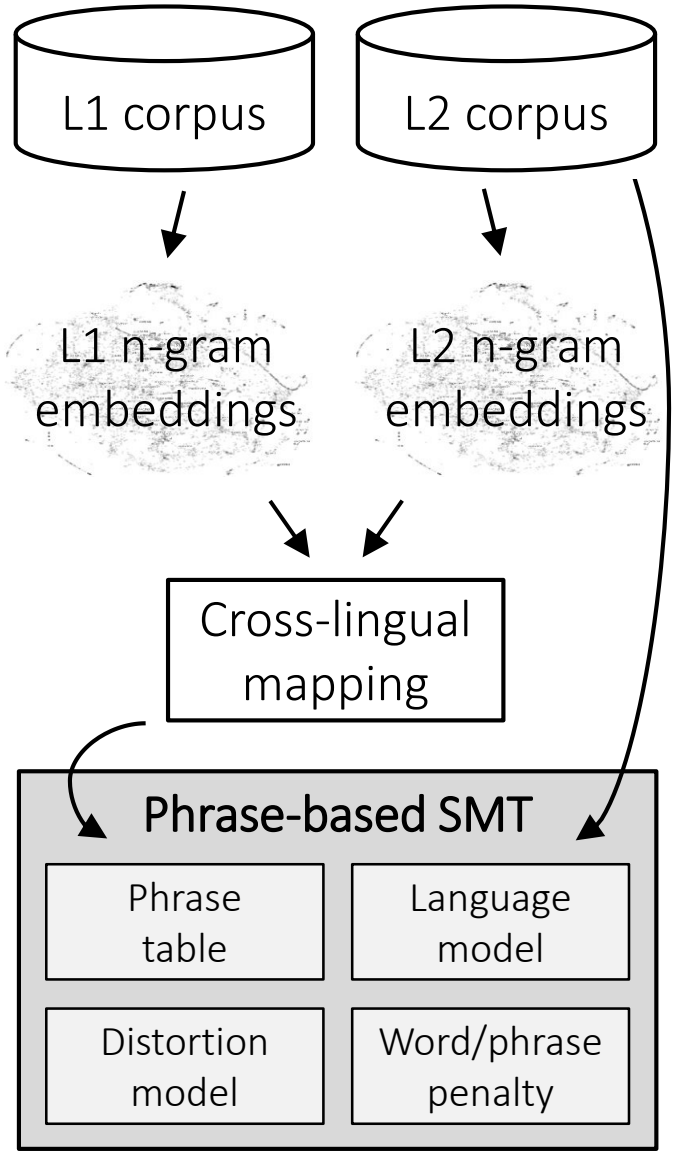
- Training: WMT-14 News Crawl

Unsupervised phrase-based SMT



EXPERIMENTS

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

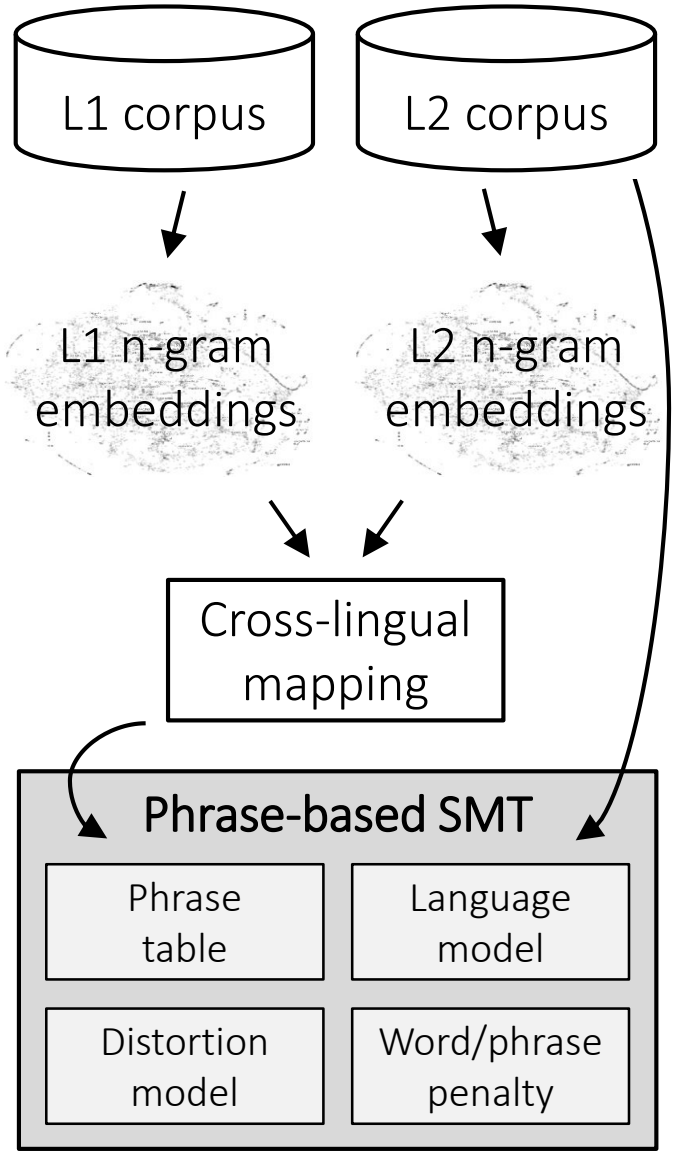


Unsupervised phrase-based SMT

EXPERIMENTS

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

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



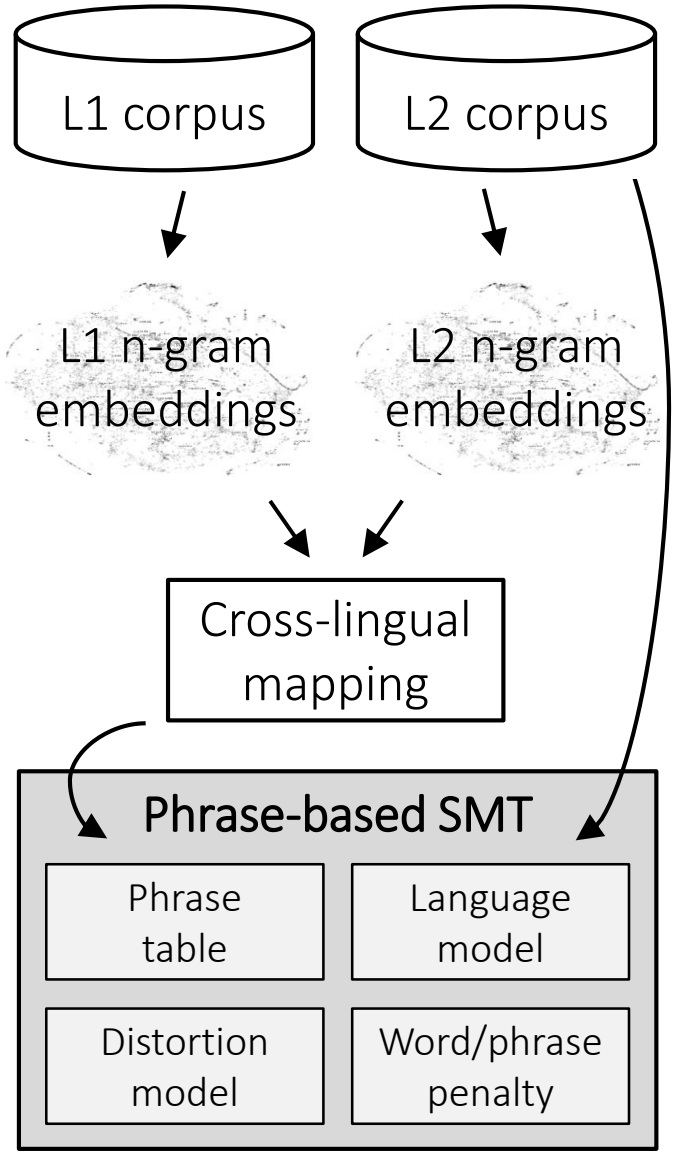
Unsupervised phrase-based SMT

EXPERIMENTS

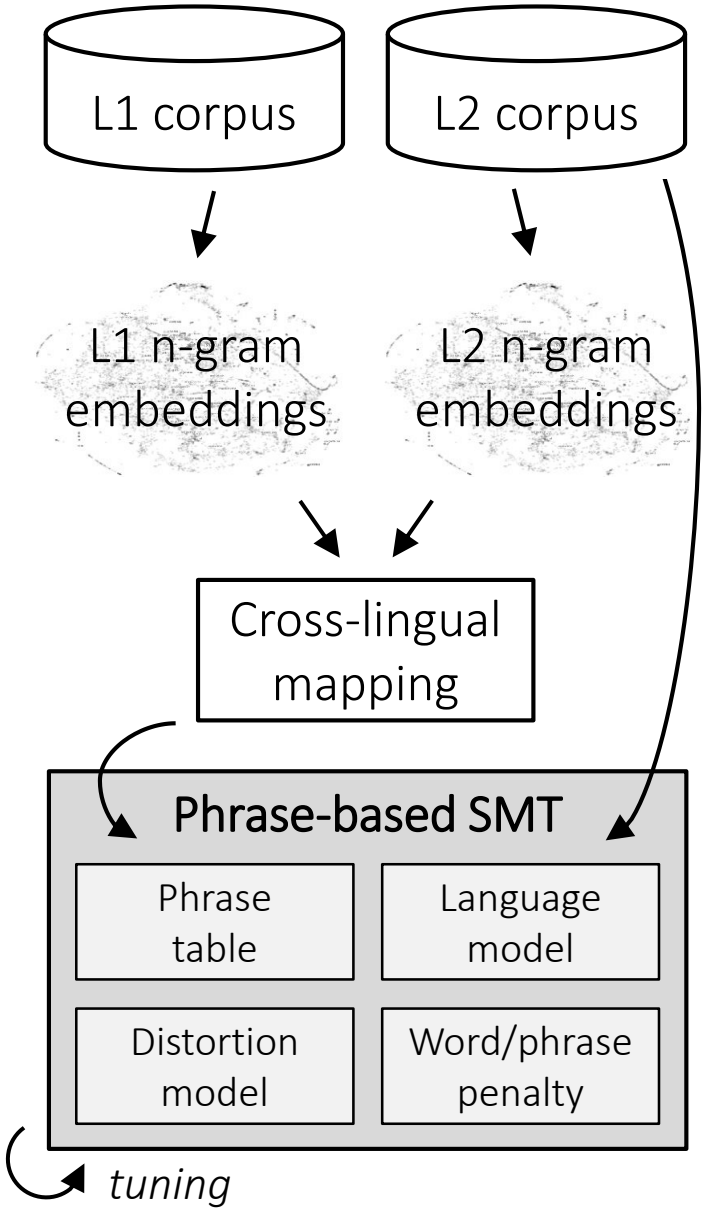
- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

FR-EN	EN-FR	DE-EN	EN-DE
22.4	19.6	15.3	11.0

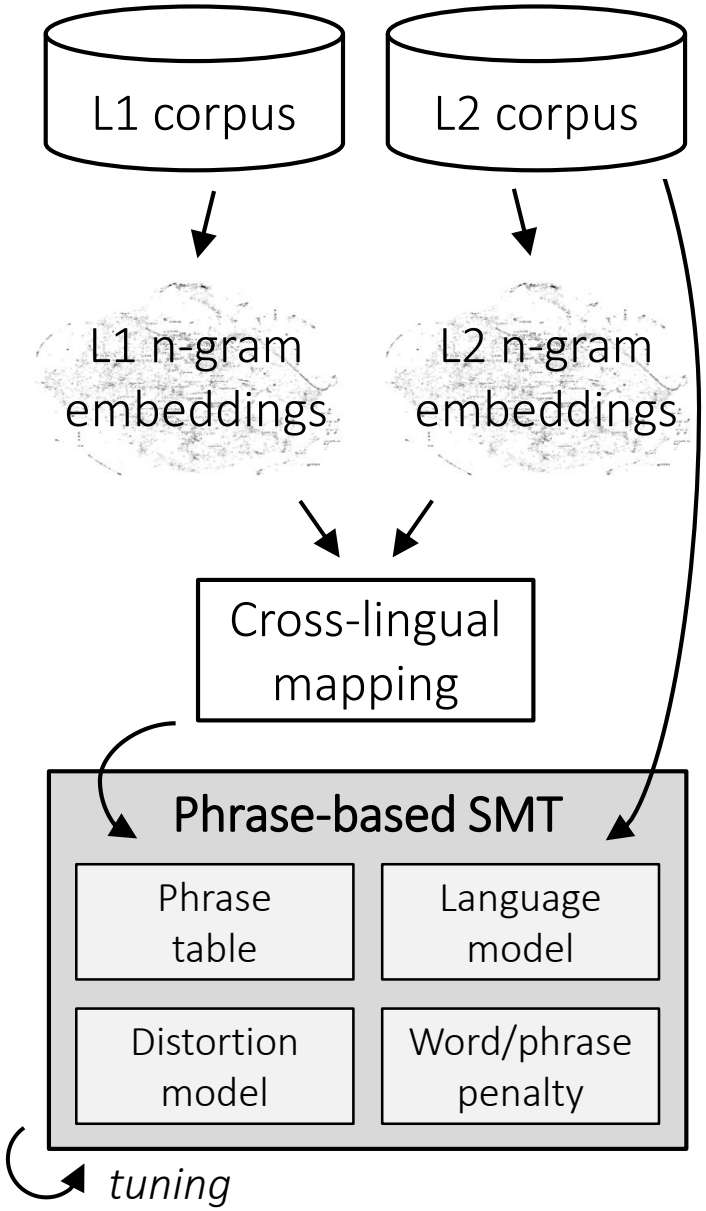
Unsupervised phrase-based SMT



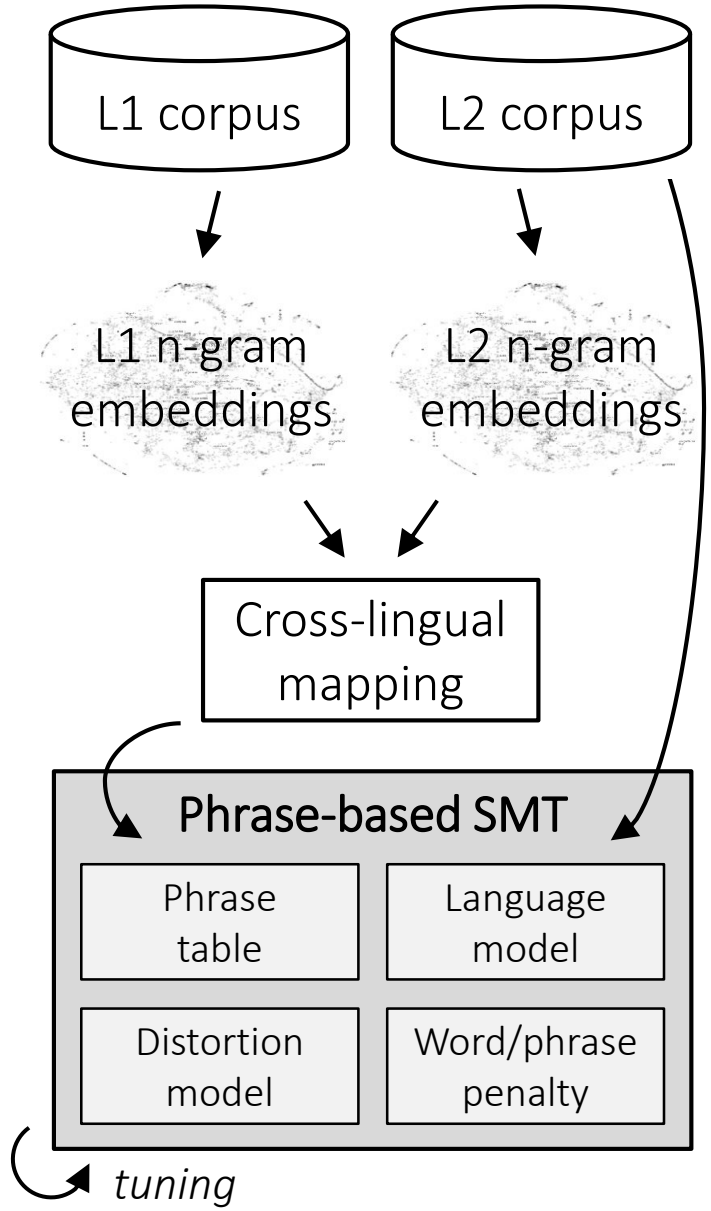
Unsupervised phrase-based SMT



Unsupervised phrase-based SMT



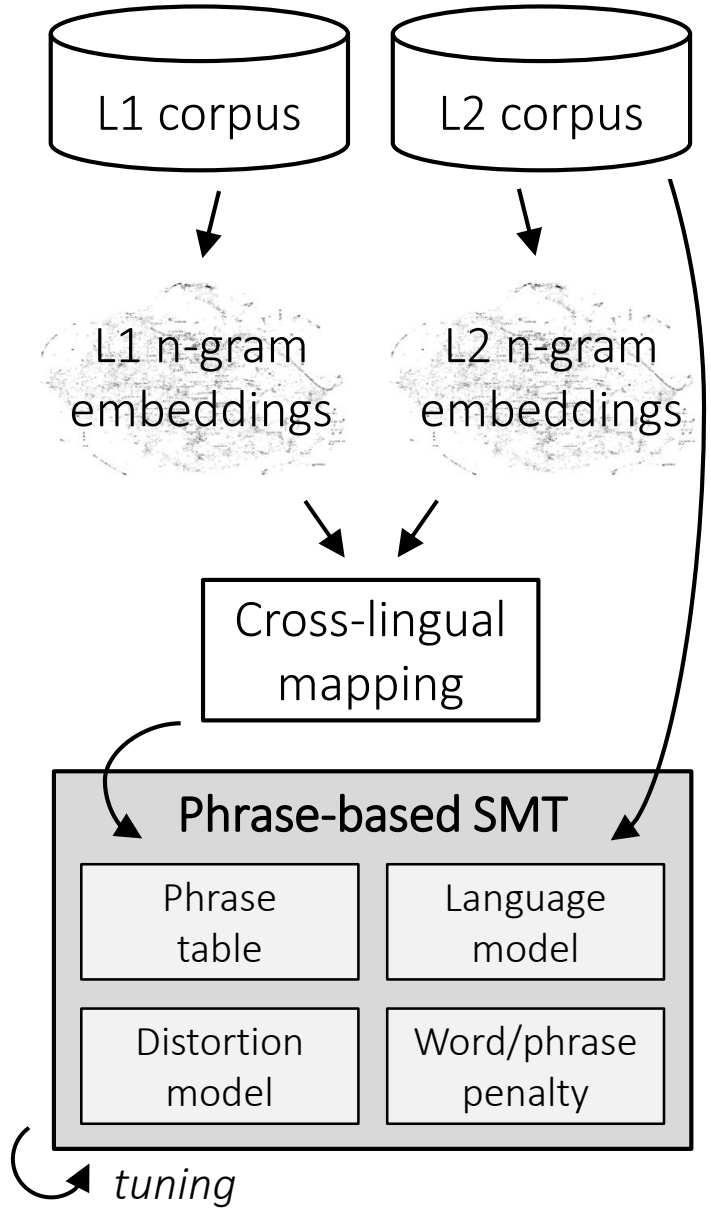
$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$



Unsupervised phrase-based SMT

$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

- $L_{cycle}(E) = 1 - \mathrm{BLEU}\left(\mathrm{T}_{F \to E}\left(\mathrm{T}_{E \to F}(E)\right), E\right)$

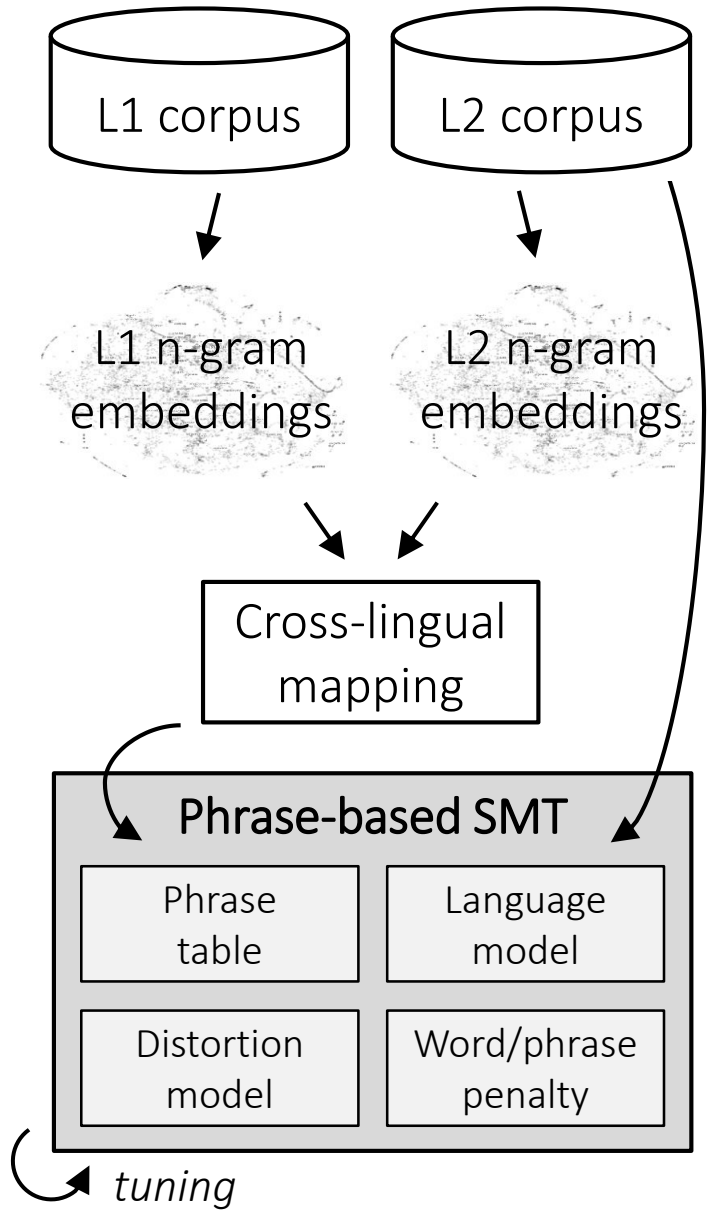


Unsupervised phrase-based SMT

$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

- $L_{cycle}(E) = 1 - \mathrm{BLEU}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right), E\right)$
- $L_{lm}(E) = \mathrm{LP} \cdot \max\left(0, \mathrm{H}(F) - \mathrm{H}\left(\mathrm{T}_{E\to F}(E)\right)\right)^2$

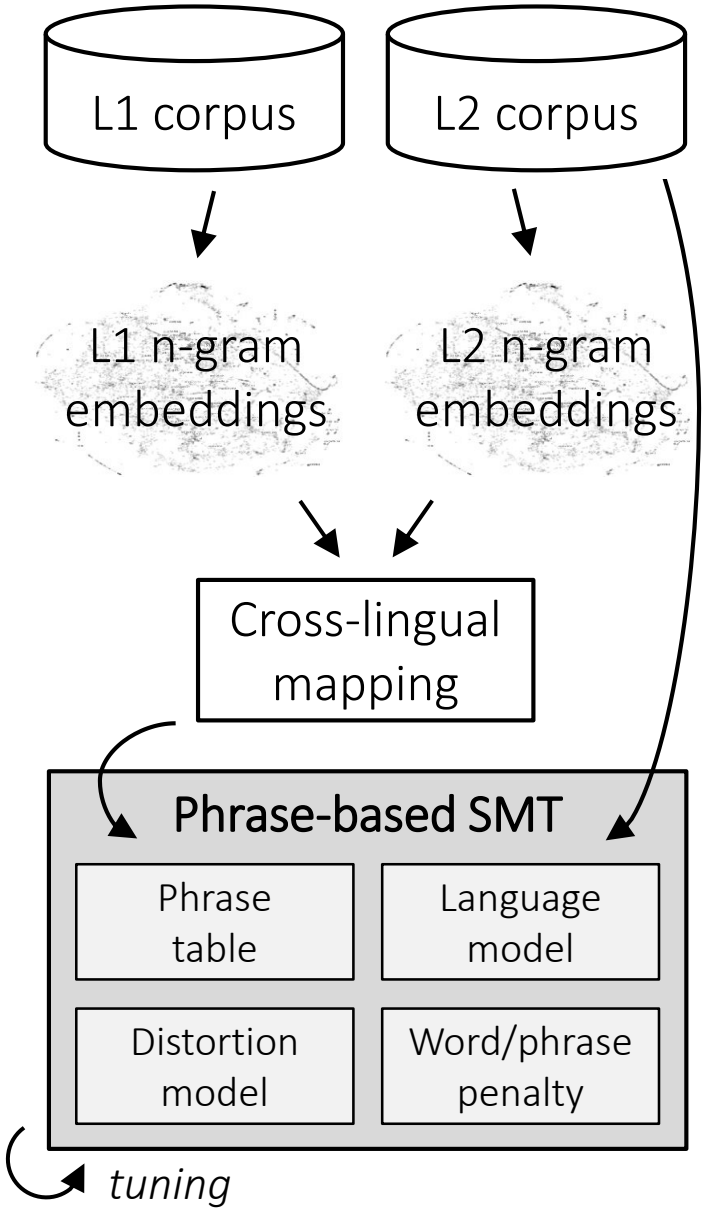
Unsupervised phrase-based SMT



$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

- $L_{cycle}(E) = 1 - \mathrm{BLEU}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right), E\right)$
- $L_{lm}(E) = \mathrm{LP}\cdot\max\left(0, \mathrm{H}(F) - \mathrm{H}\left(\mathrm{T}_{E\to F}(E)\right)\right)^2$

$$\mathrm{LP} = \mathrm{LP}(E)\cdot\mathrm{LP}(F)\,, \quad \mathrm{LP}(E) = \max\left(1, \frac{\mathrm{len}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right)\right)}{\mathrm{len}(E)}\right)$$



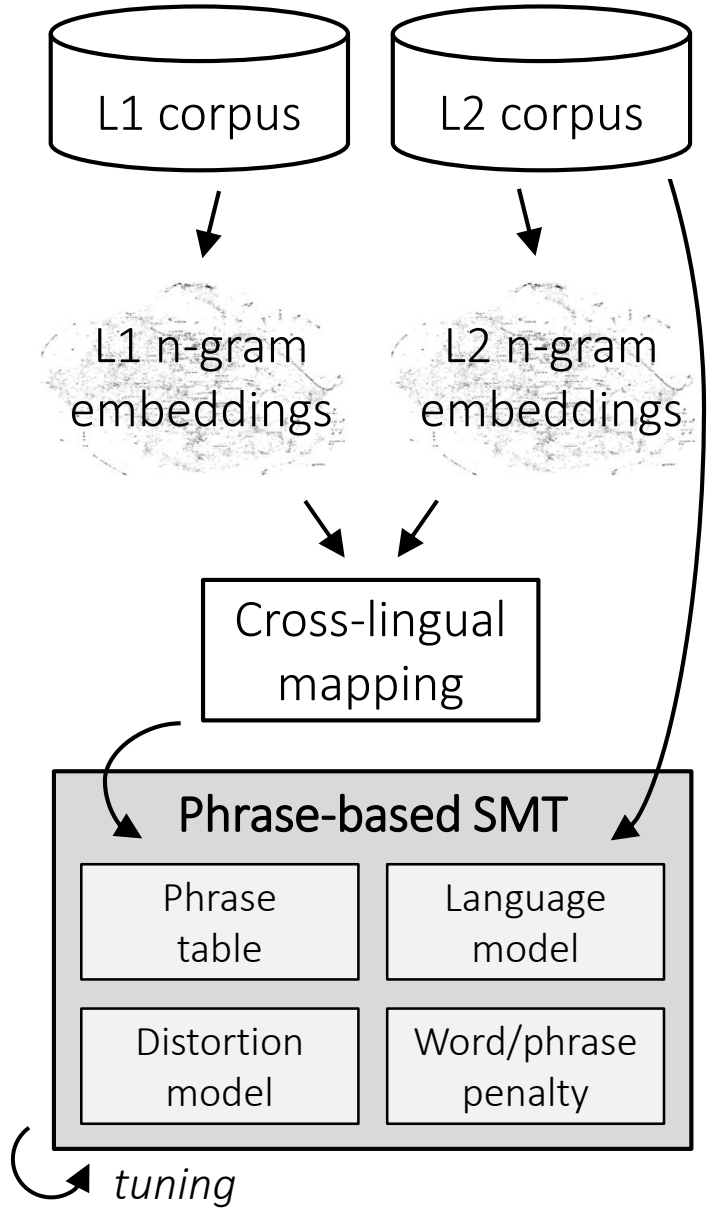
Unsupervised phrase-based SMT

$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

- $L_{cycle}(E) = 1 - \mathrm{BLEU}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right), E\right)$
- $L_{lm}(E) = \mathrm{LP}\cdot\max\left(0, \mathrm{H}(F) - \mathrm{H}\left(\mathrm{T}_{E\to F}(E)\right)\right)^2$

$$\mathrm{LP} = \mathrm{LP}(E)\cdot\mathrm{LP}(F)\,, \quad \mathrm{LP}(E) = \max\left(1, \frac{\mathrm{len}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right)\right)}{\mathrm{len}(E)}\right)$$

Alternating optimization



Unsupervised phrase-based SMT

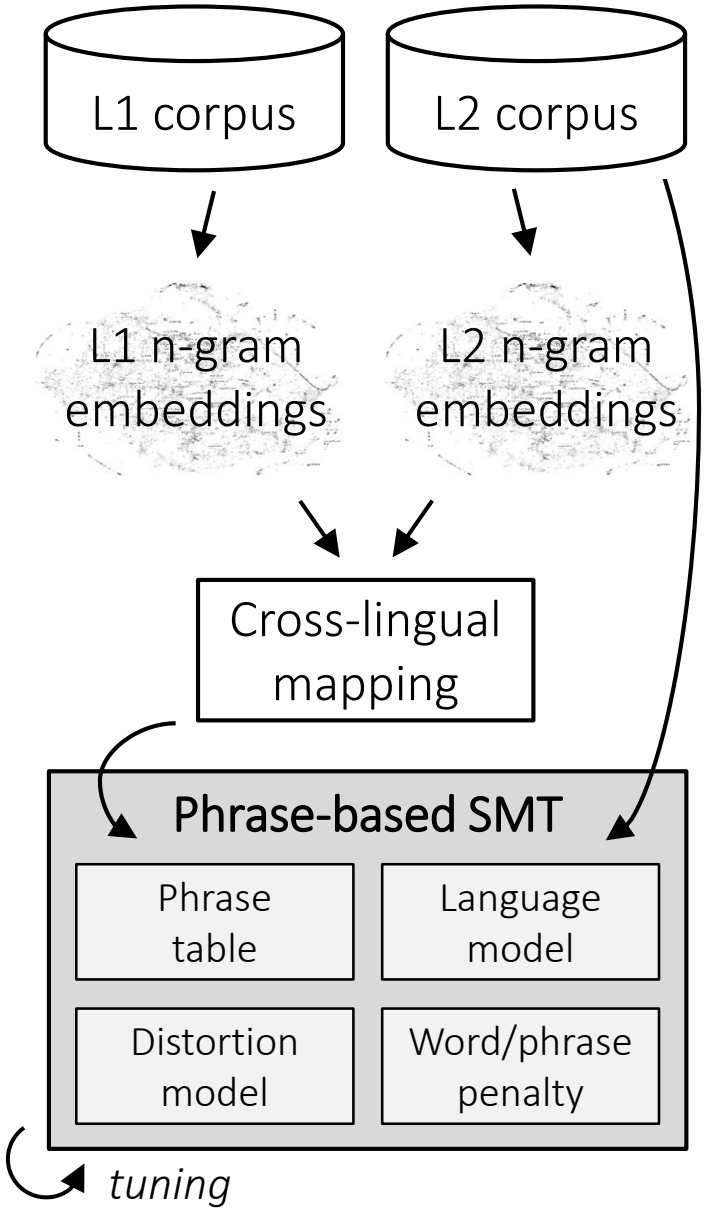
$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

- $L_{cycle}(E) = 1 - \mathrm{BLEU}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right), E\right)$
- $L_{lm}(E) = \mathrm{LP}\cdot\max\left(0, \mathrm{H}(F) - \mathrm{H}\left(\mathrm{T}_{E\to F}(E)\right)\right)^2$

$$\mathrm{LP} = \mathrm{LP}(E)\cdot\mathrm{LP}(F)\,, \quad \mathrm{LP}(E) = \max\left(1, \frac{\mathrm{len}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right)\right)}{\mathrm{len}(E)}\right)$$

Alternating optimization

- Fix $\mathrm{T}_{F\to E}$ and optimize $\mathrm{T}_{E\to F}$ using MERT



Unsupervised phrase-based SMT

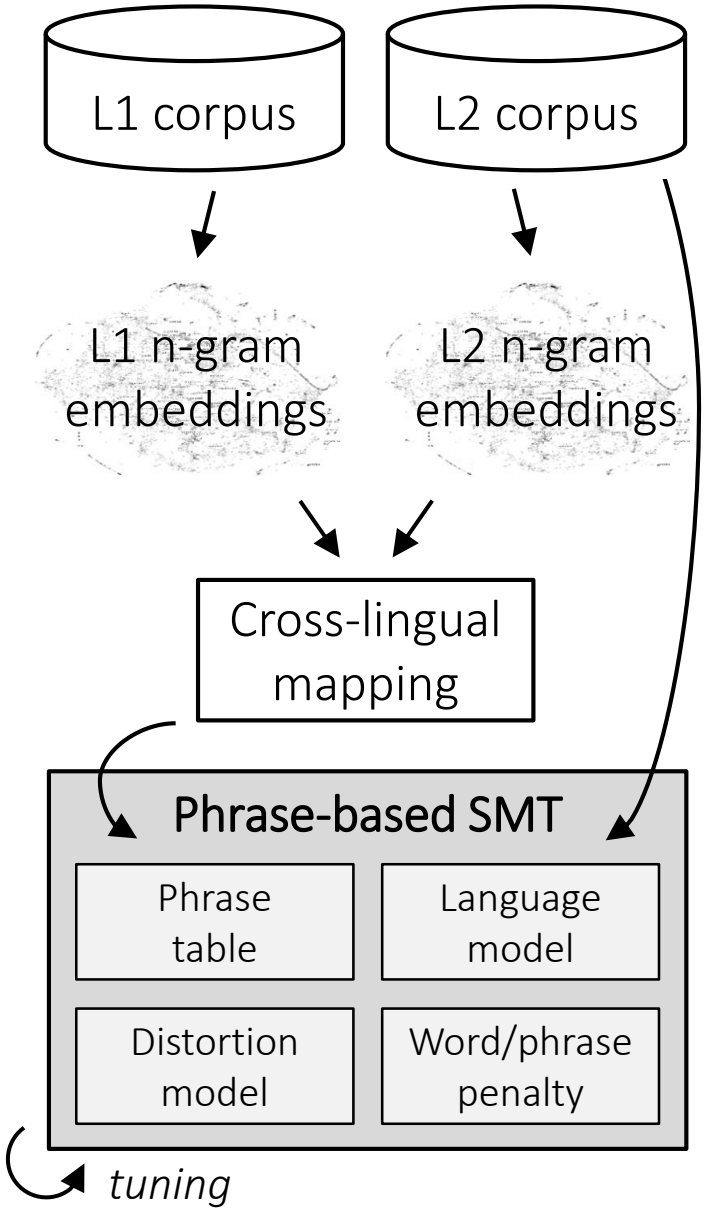
$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

- $L_{cycle}(E) = 1 - \mathrm{BLEU}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right), E\right)$
- $L_{lm}(E) = \mathrm{LP}\cdot\max\left(0, \mathrm{H}(F) - \mathrm{H}\left(\mathrm{T}_{E\to F}(E)\right)\right)^2$

$$\mathrm{LP} = \mathrm{LP}(E)\cdot\mathrm{LP}(F)\,, \quad \mathrm{LP}(E) = \max\left(1, \frac{\mathrm{len}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right)\right)}{\mathrm{len}(E)}\right)$$

Alternating optimization

- Fix $\mathrm{T}_{F\to E}$ and optimize $\mathrm{T}_{E\to F}$ using MERT
- Fix $\mathrm{T}_{E\to F}$ and optimize $\mathrm{T}_{F\to E}$ using MERT



Unsupervised phrase-based SMT

$$L = L_{cycle}(E) + L_{cycle}(F) + L_{lm}(E) + L_{lm}(F)$$

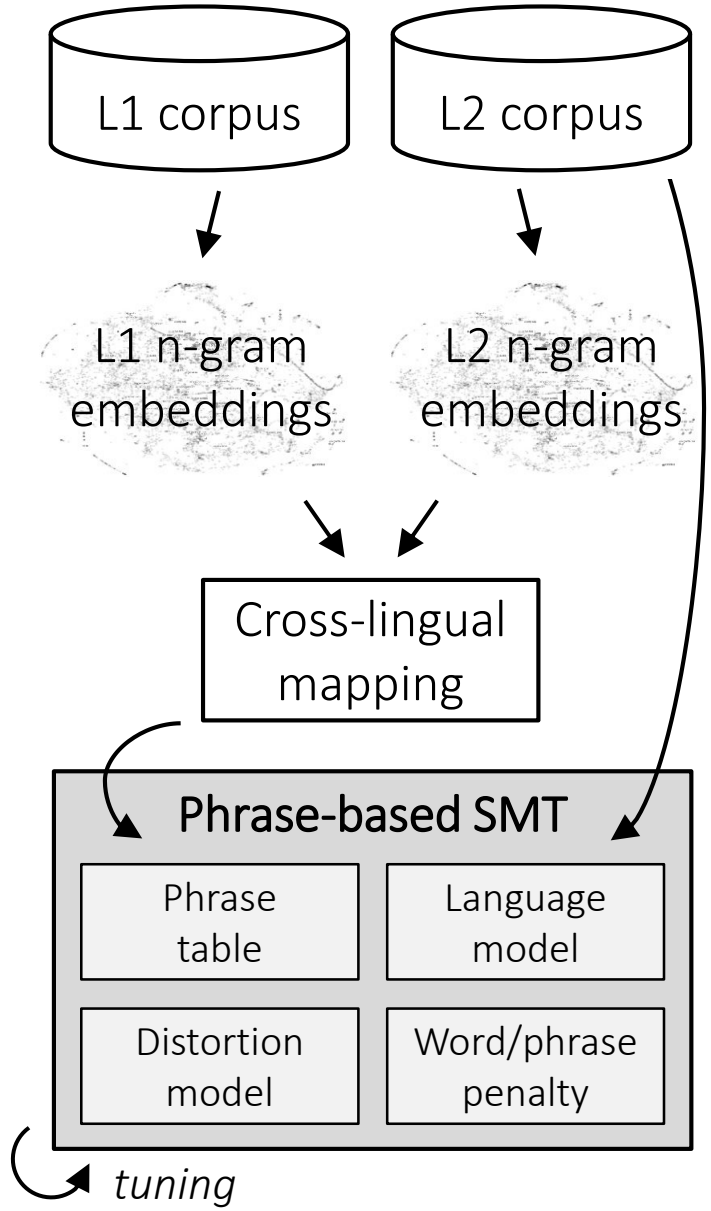
- $L_{cycle}(E) = 1 - \mathrm{BLEU}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right), E\right)$
- $L_{lm}(E) = \mathrm{LP}\cdot\max\left(0, \mathrm{H}(F) - \mathrm{H}\left(\mathrm{T}_{E\to F}(E)\right)\right)^2$

$$\mathrm{LP} = \mathrm{LP}(E)\cdot\mathrm{LP}(F)\,, \quad \mathrm{LP}(E) = \max\left(1, \frac{\mathrm{len}\left(\mathrm{T}_{F\to E}\left(\mathrm{T}_{E\to F}(E)\right)\right)}{\mathrm{len}(E)}\right)$$

Alternating optimization

- Fix $\mathrm{T}_{F\to E}$ and optimize $\mathrm{T}_{E\to F}$ using MERT
- Fix $\mathrm{T}_{E\to F}$ and optimize $\mathrm{T}_{F\to E}$ using MERT
- Iterate until convergence

Unsupervised phrase-based SMT

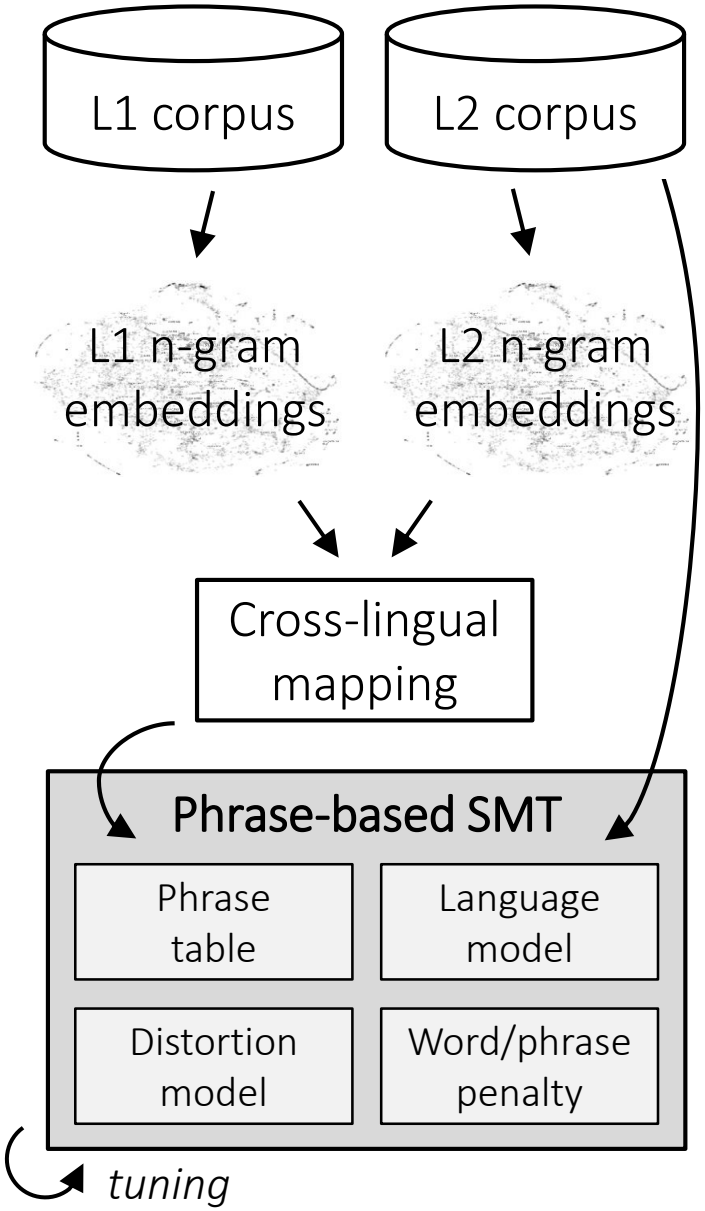


EXPERIMENTS

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Initial SMT	22.4	19.6	15.3	11.0

Unsupervised phrase-based SMT

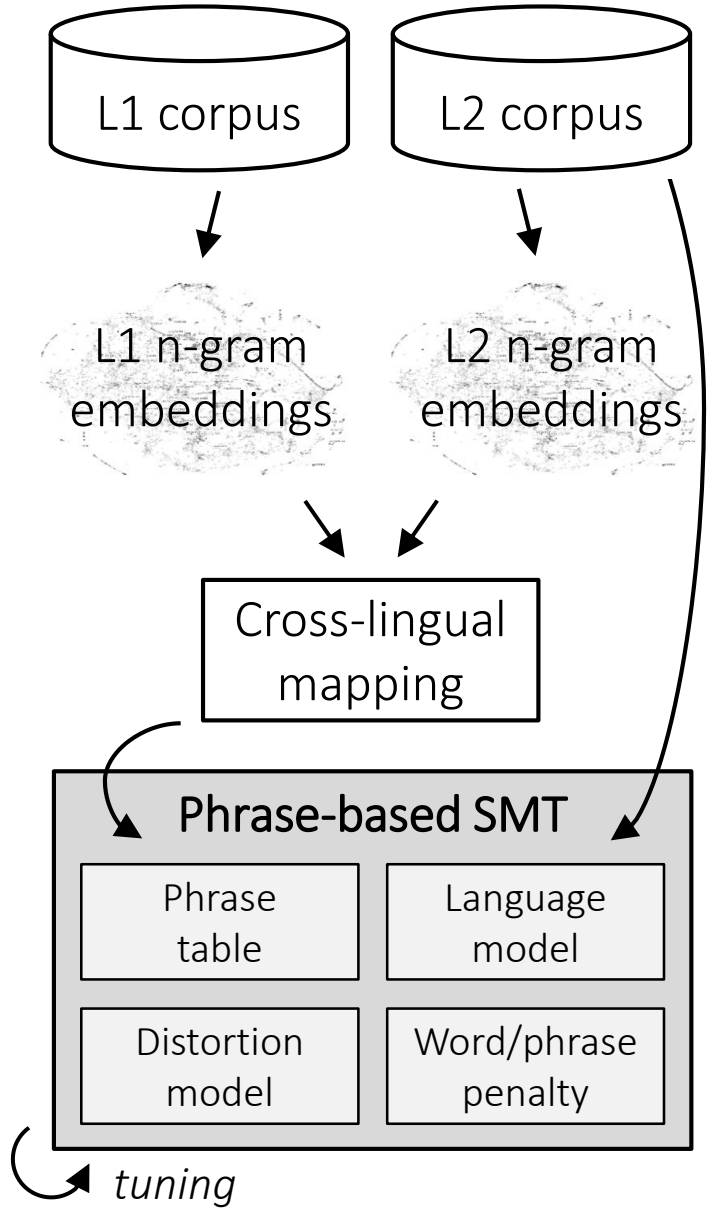


EXPERIMENTS

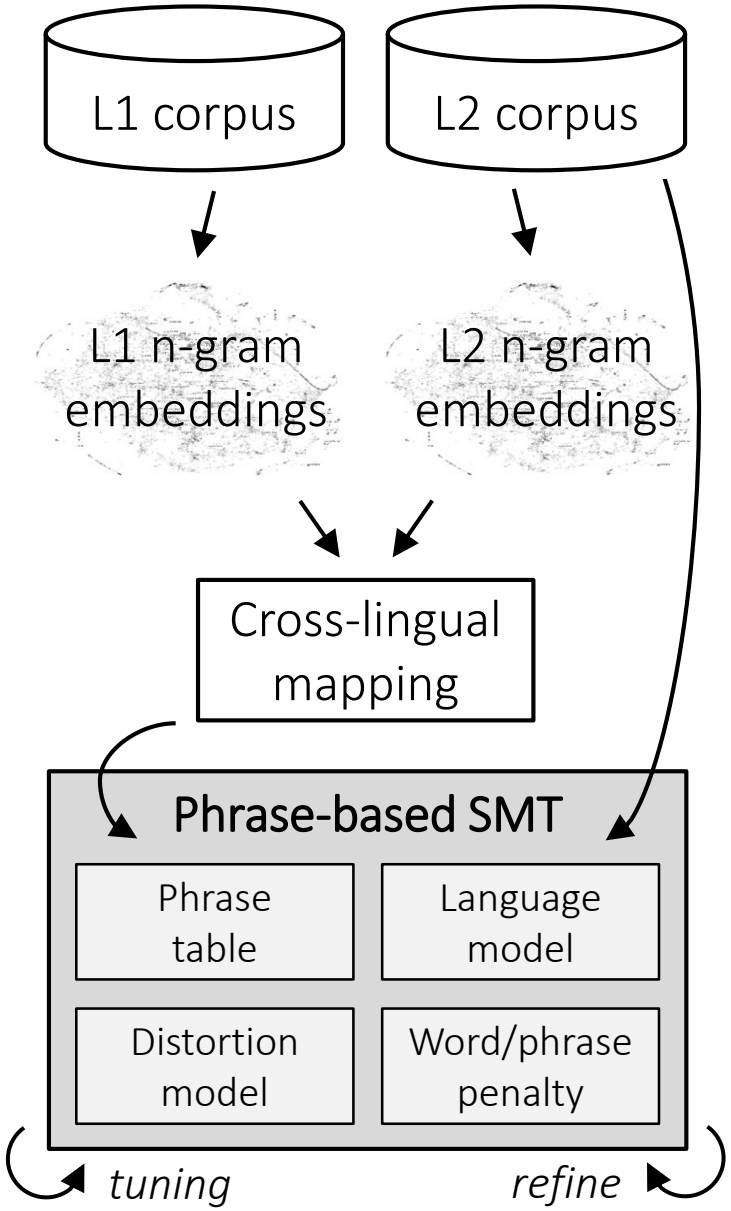
- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Initial SMT	22.4	19.6	15.3	11.0
+ Tuning	23.4	21.9	15.4	11.2

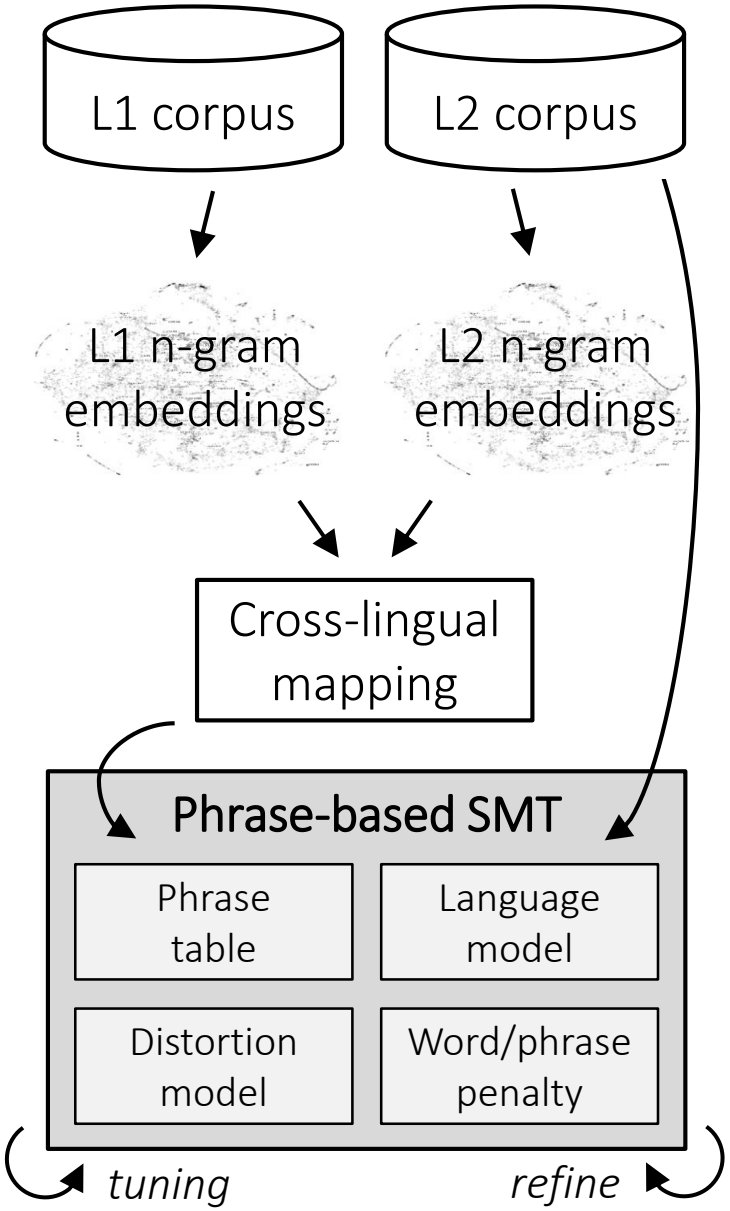
Unsupervised phrase-based SMT



Unsupervised phrase-based SMT



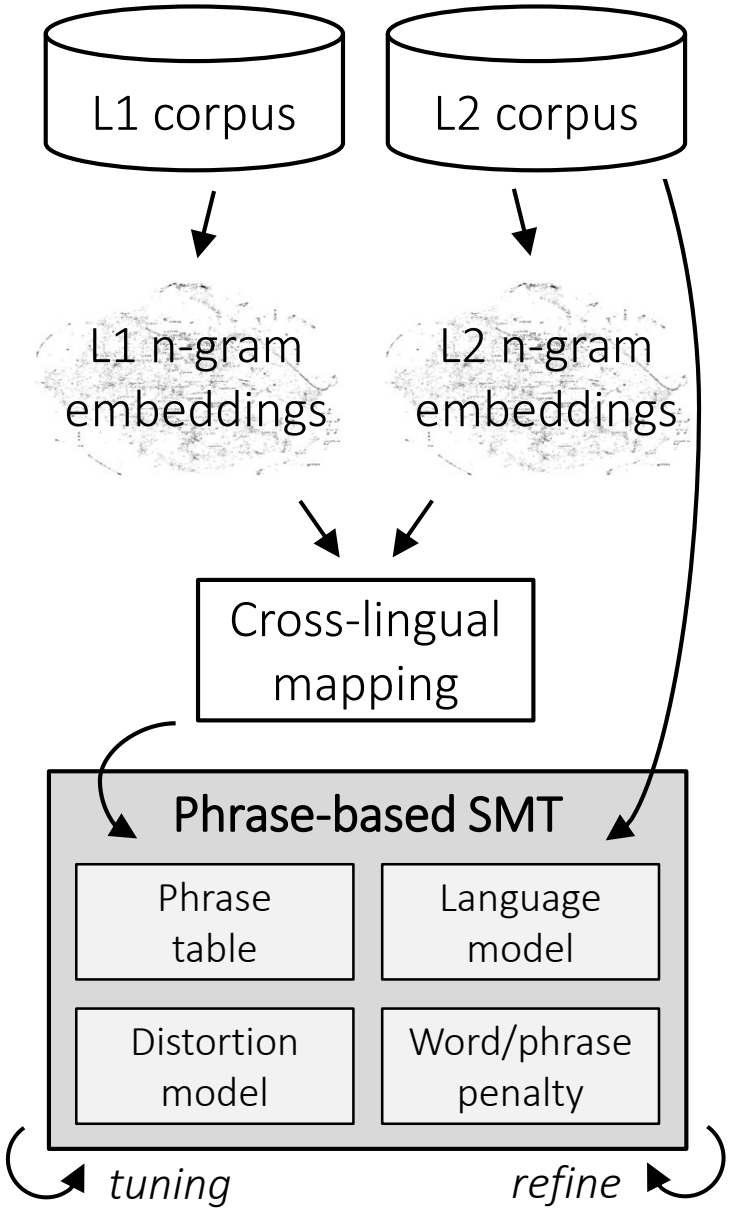
Unsupervised phrase-based SMT



L1 train

L2 train

Unsupervised phrase-based SMT

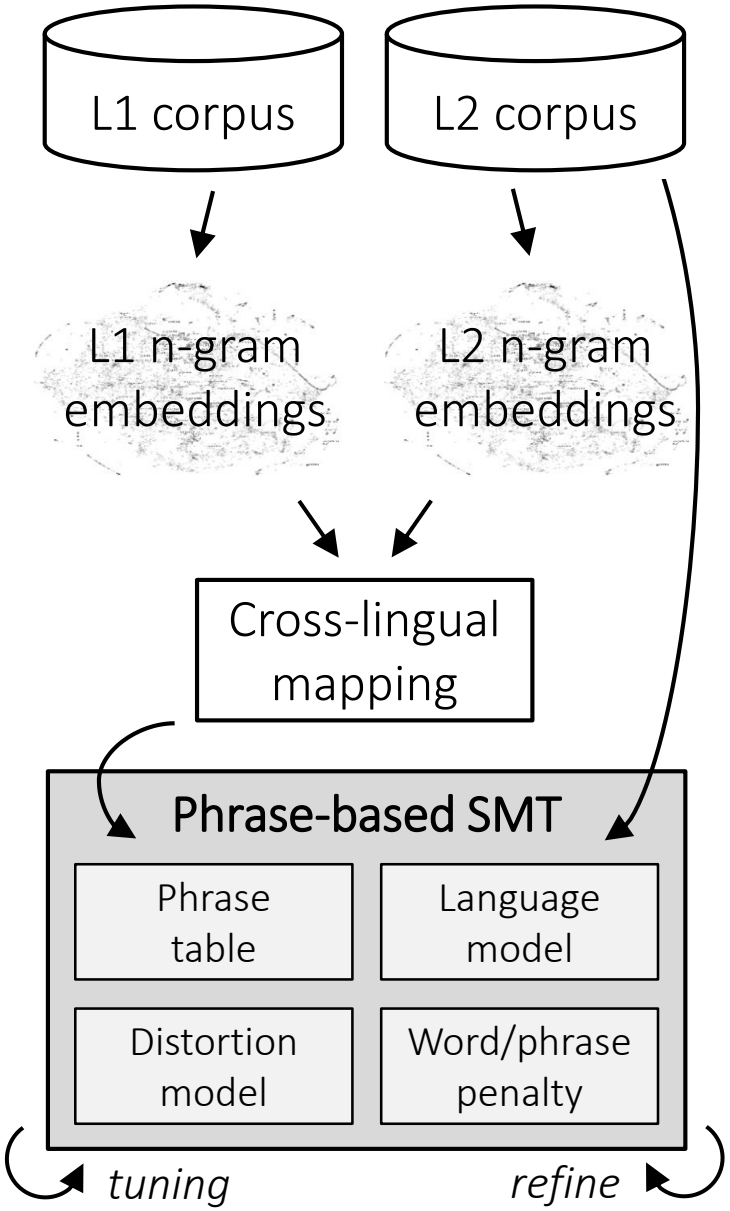


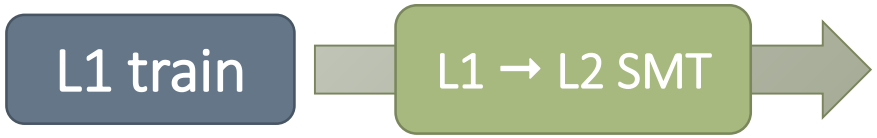
L1 train

L1 → L2 SMT

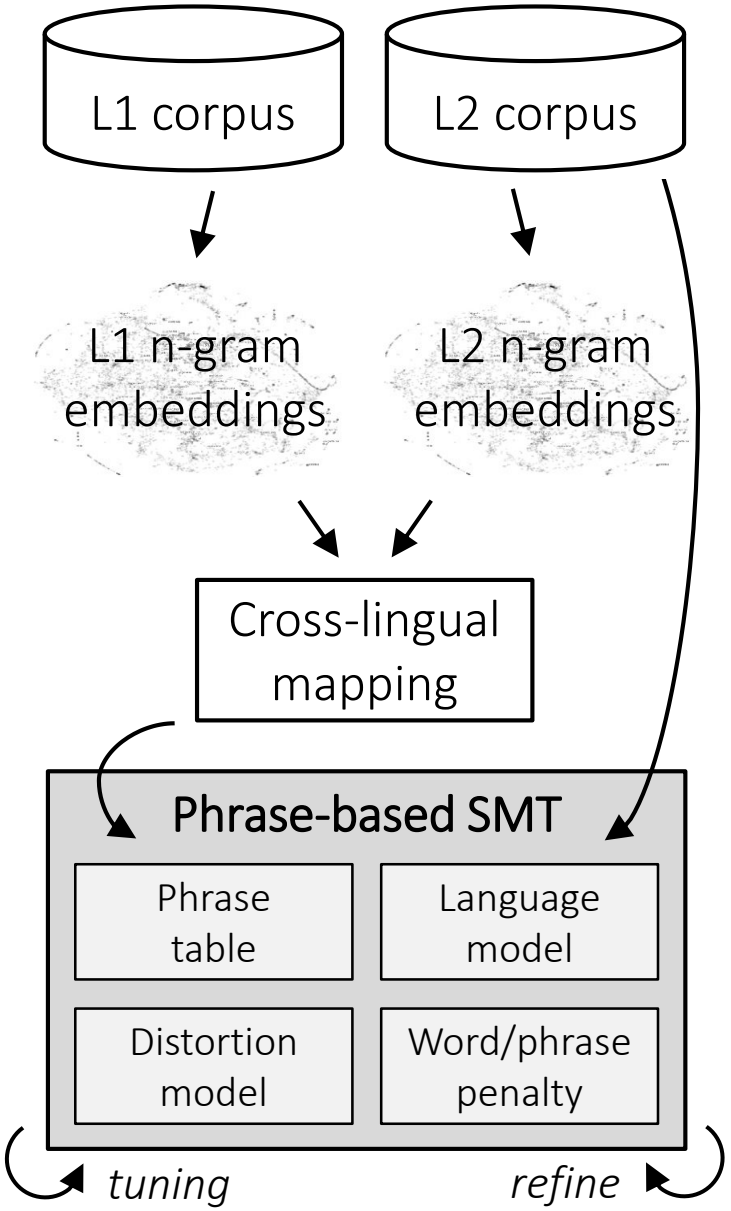
L2 train

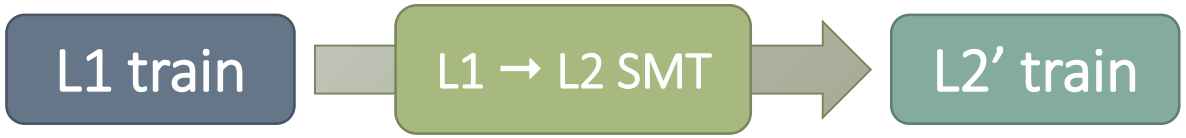
Unsupervised phrase-based SMT



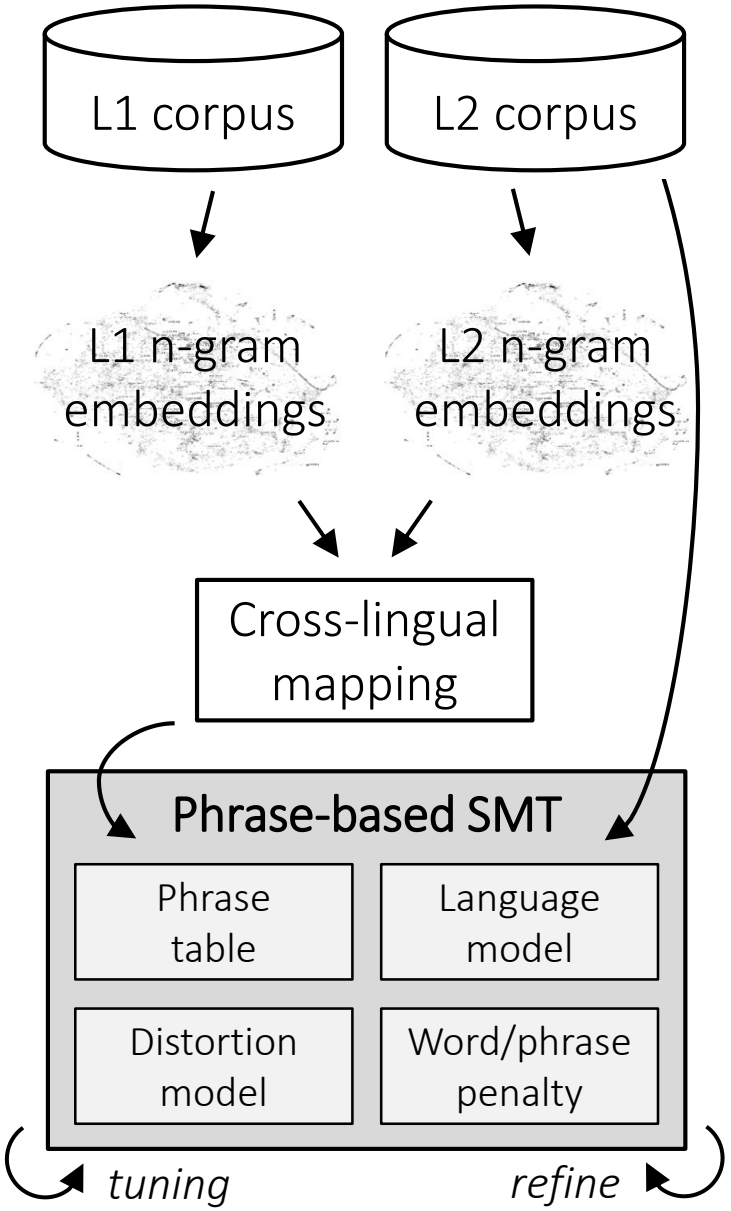


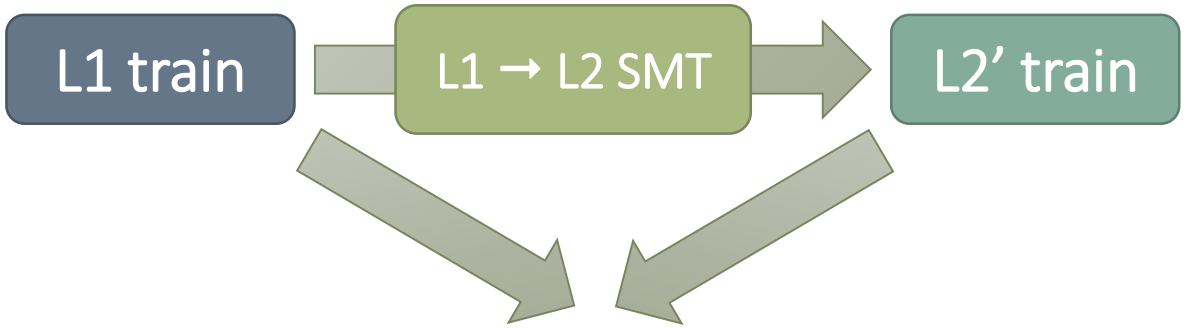
Unsupervised phrase-based SMT





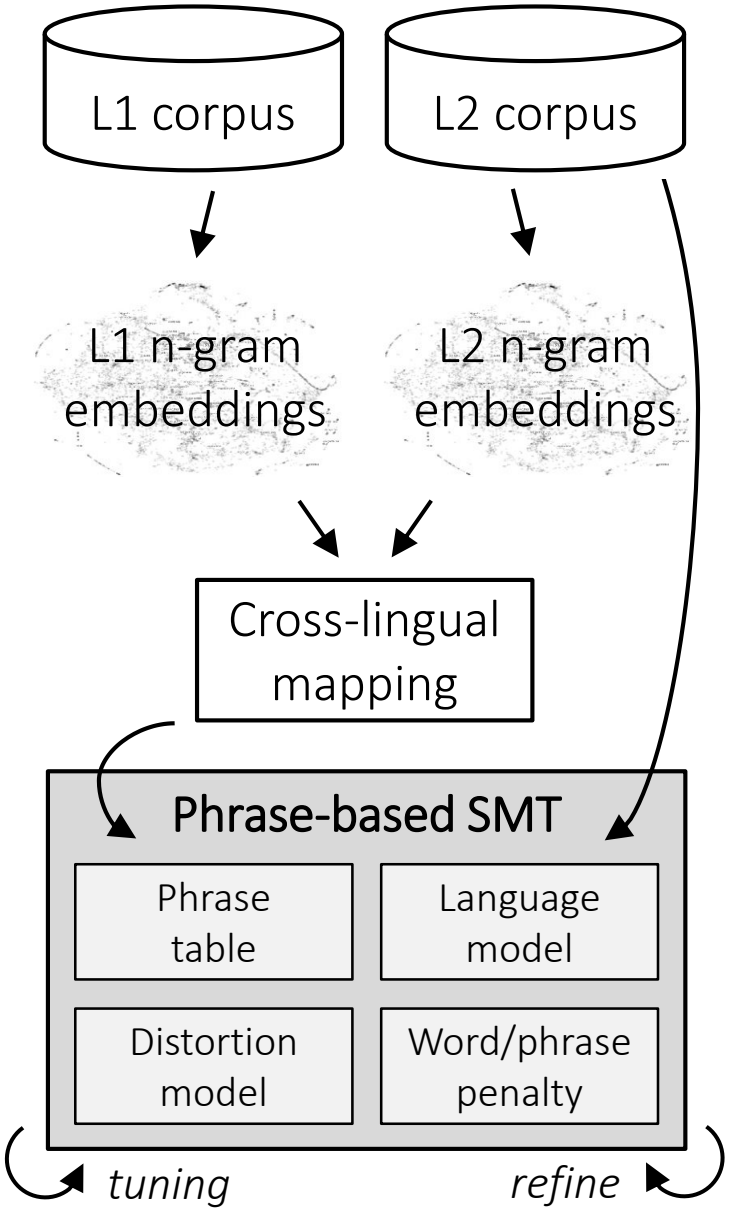
Unsupervised phrase-based SMT

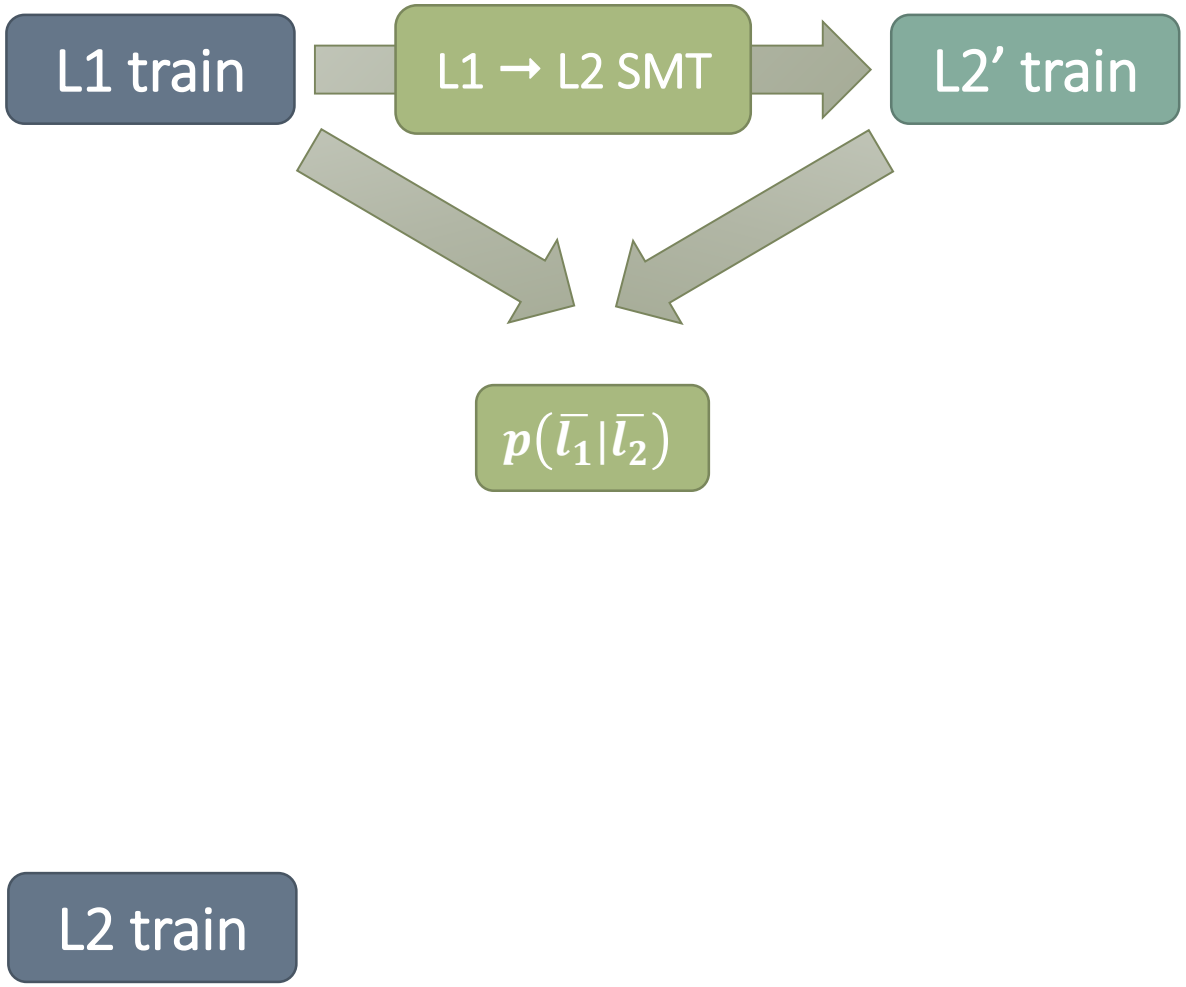




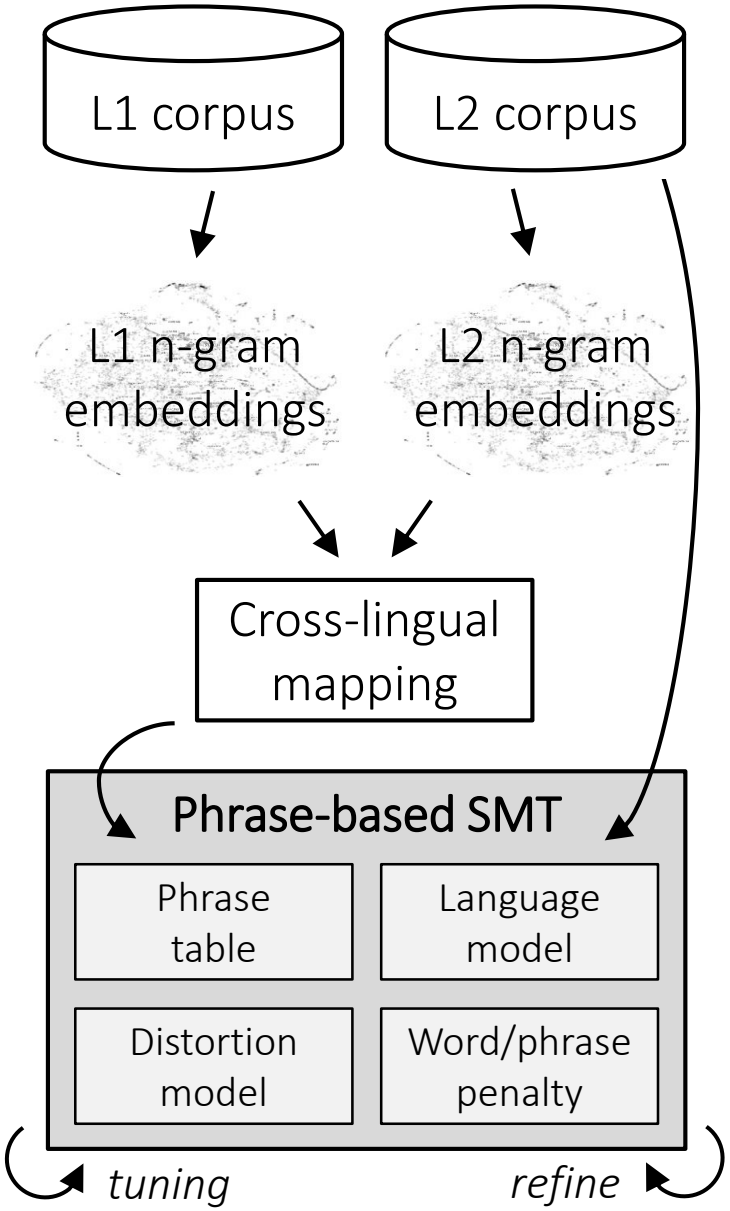
L2 train

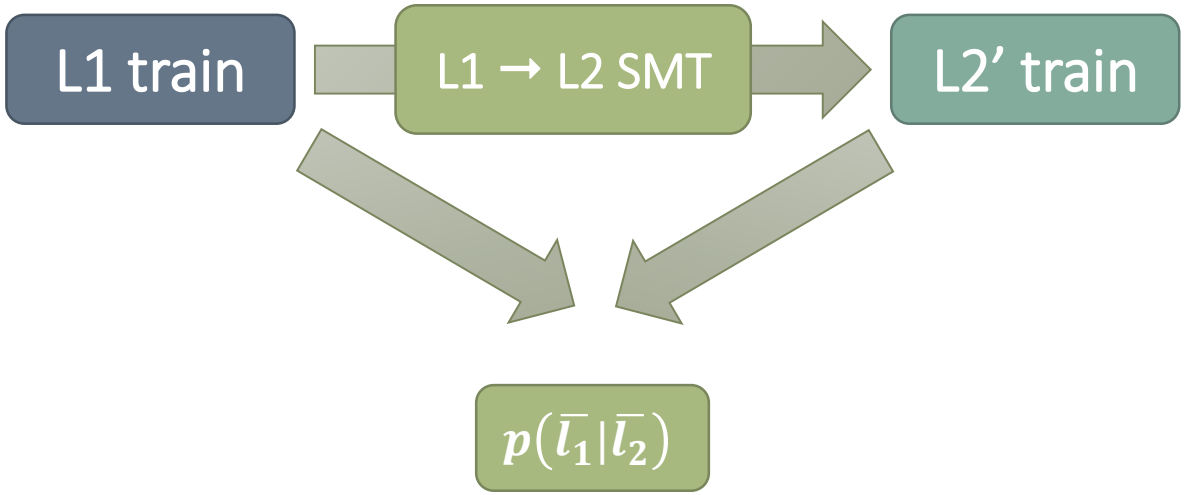
Unsupervised phrase-based SMT



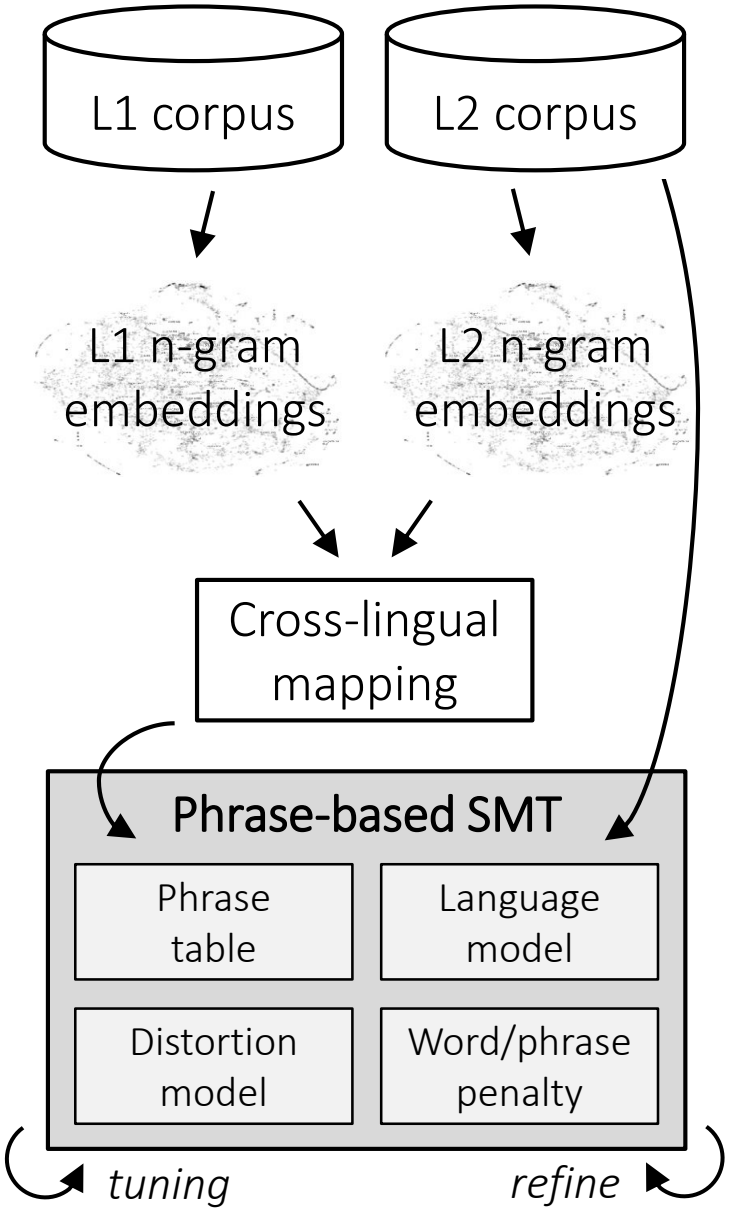


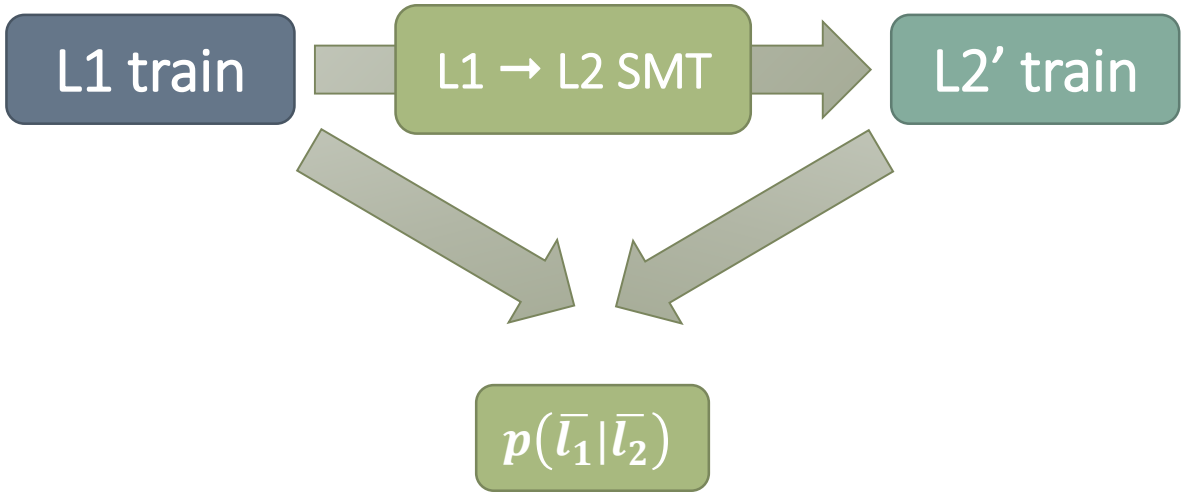
Unsupervised phrase-based SMT

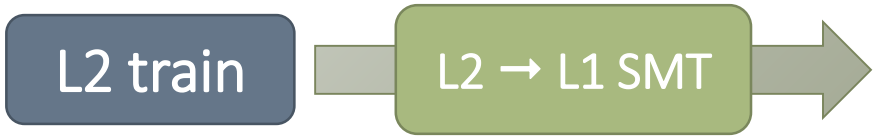




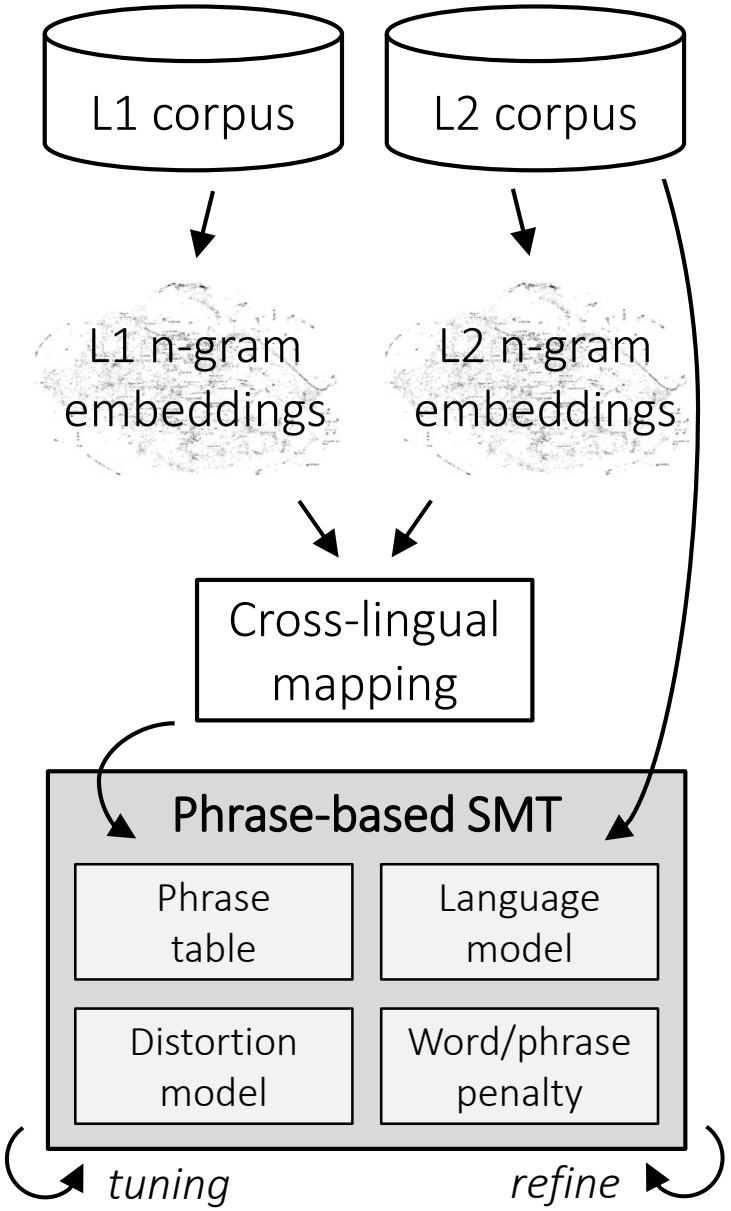
Unsupervised phrase-based SMT

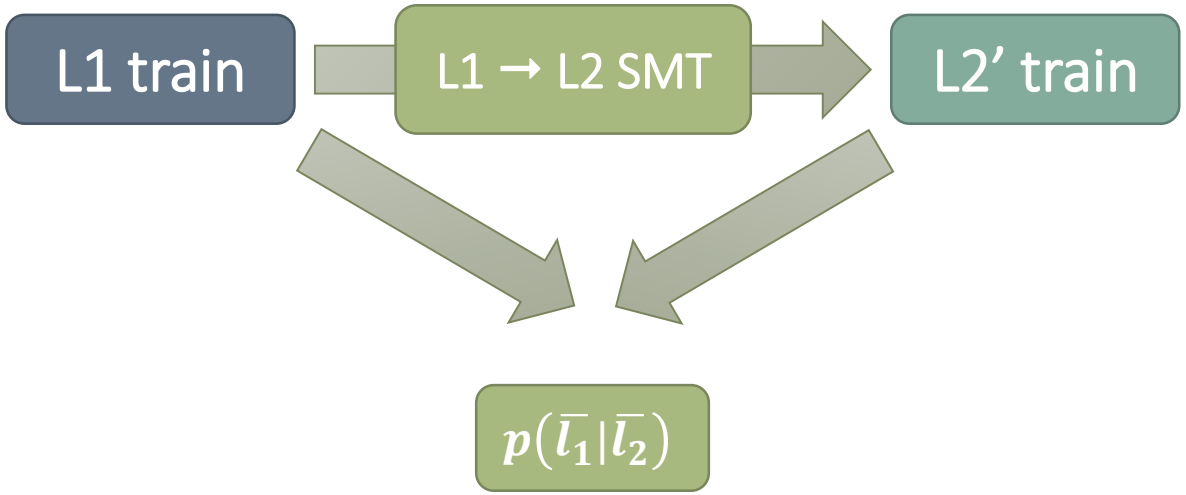


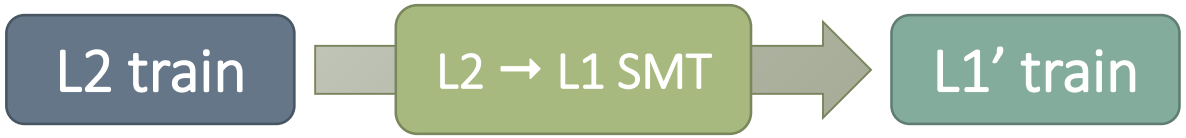




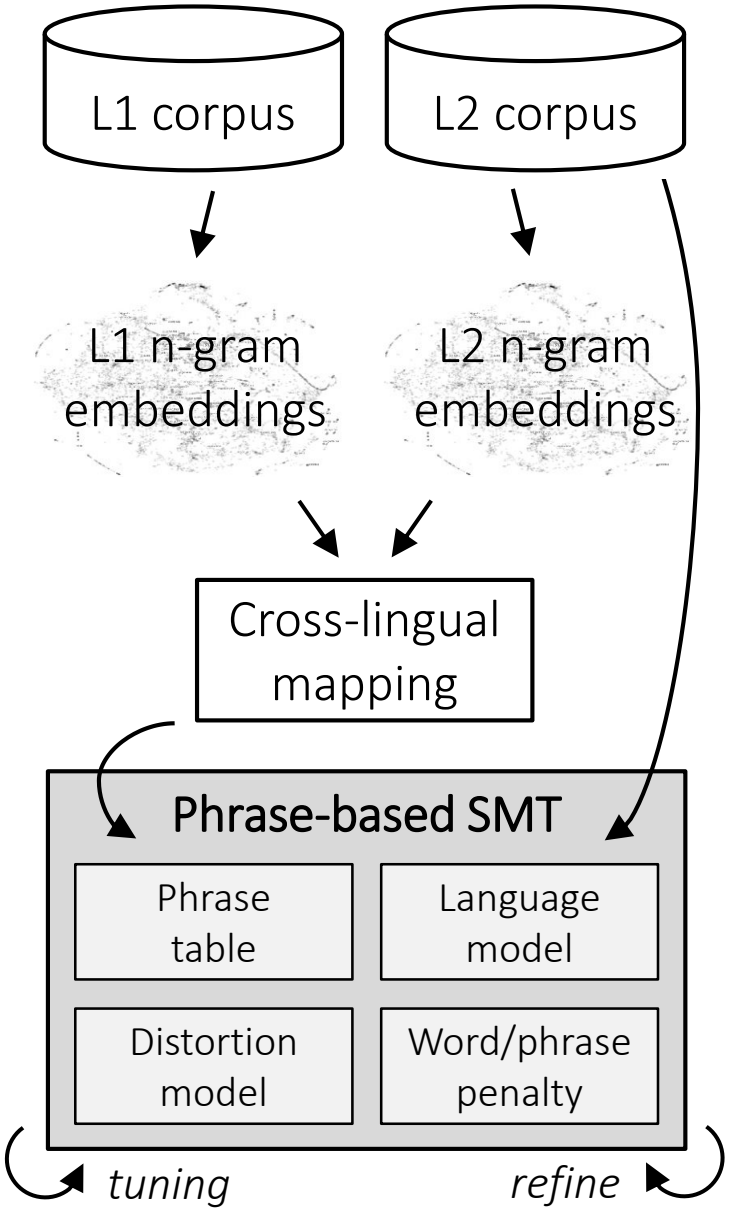
Unsupervised phrase-based SMT

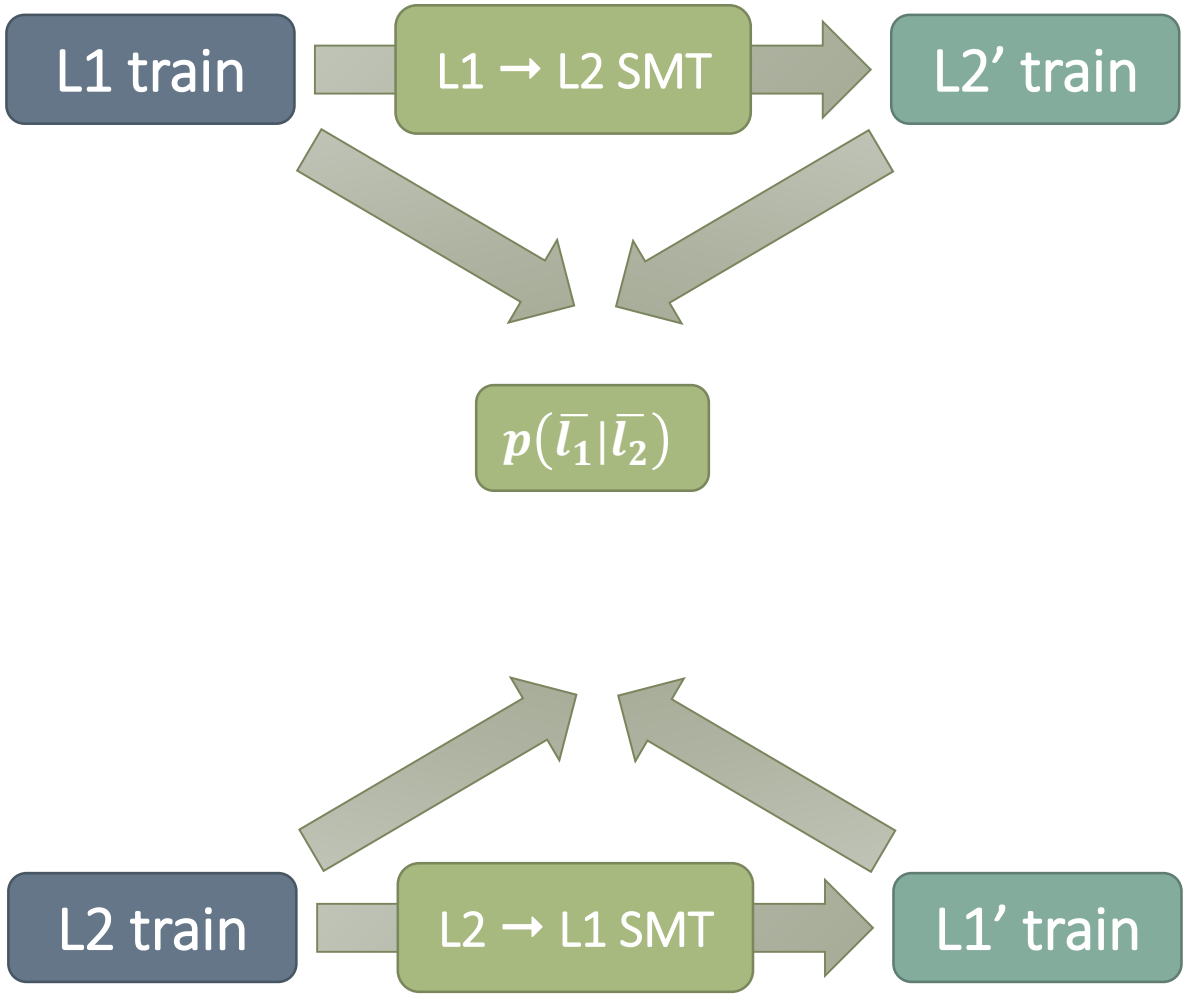




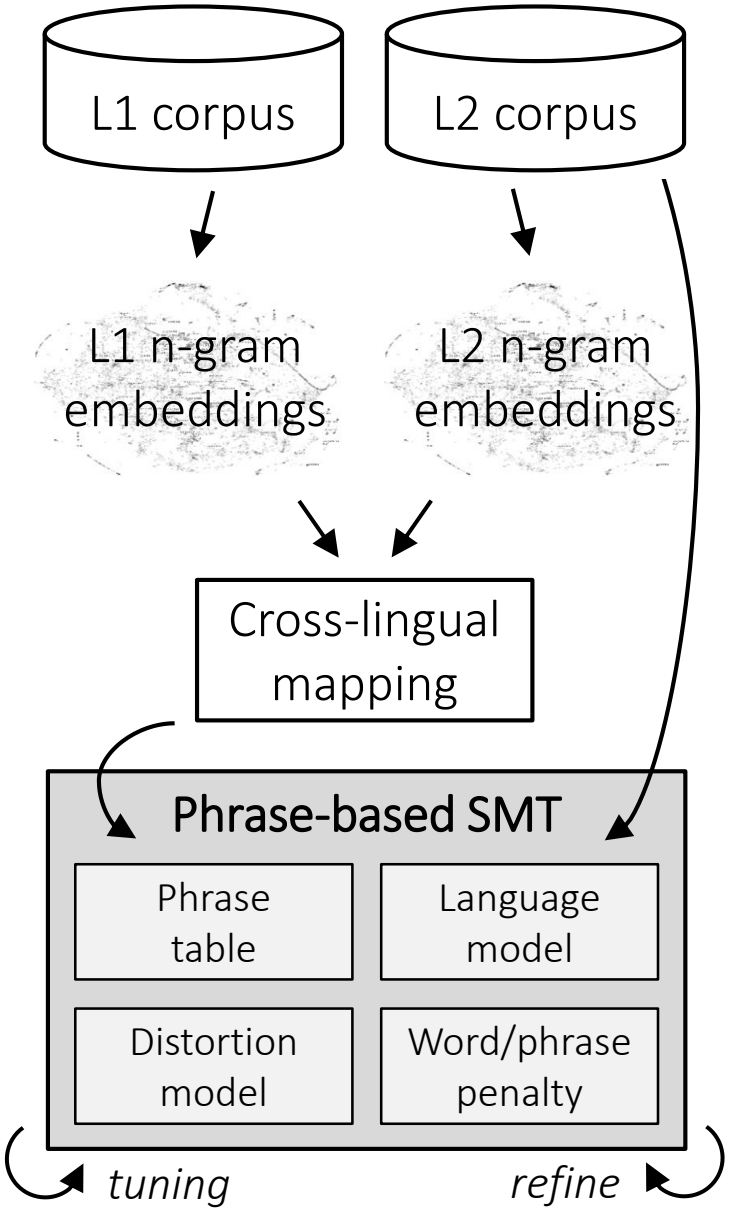


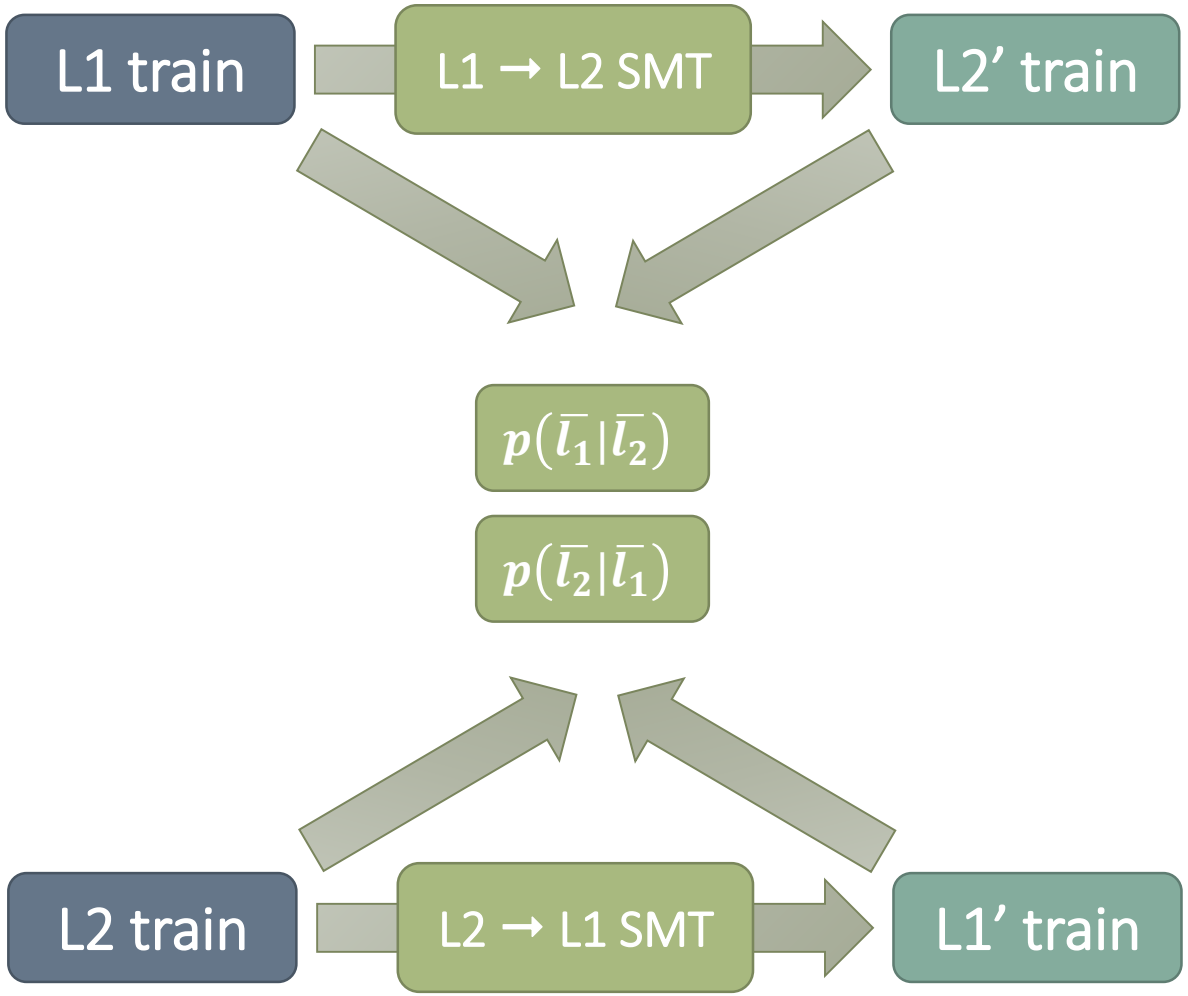
Unsupervised phrase-based SMT



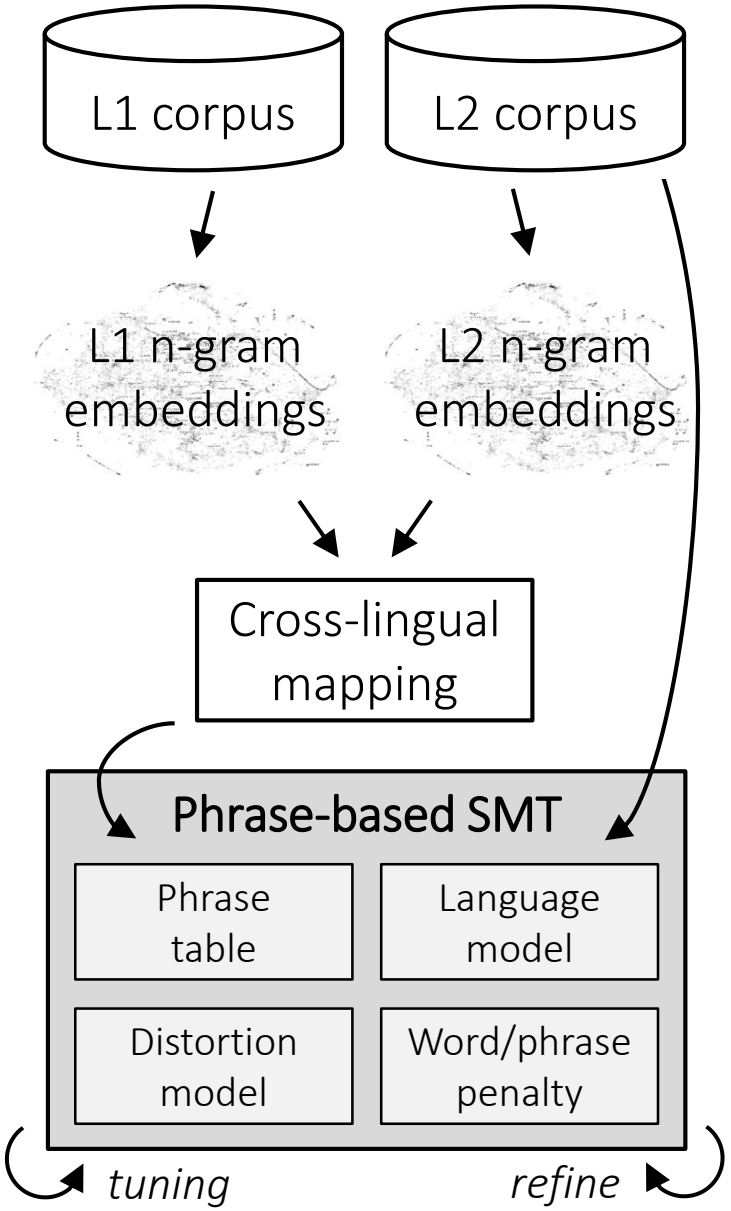


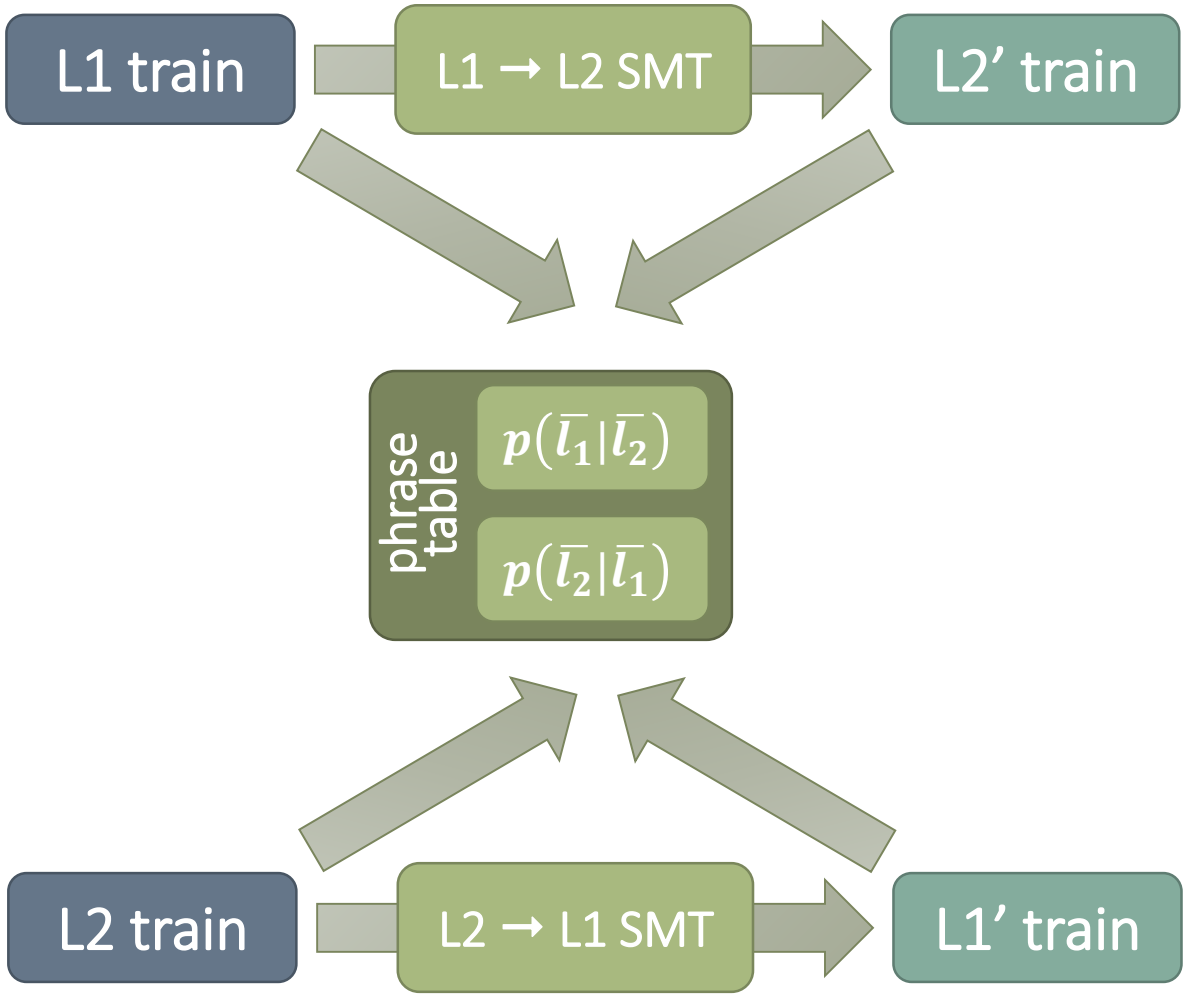
Unsupervised phrase-based SMT



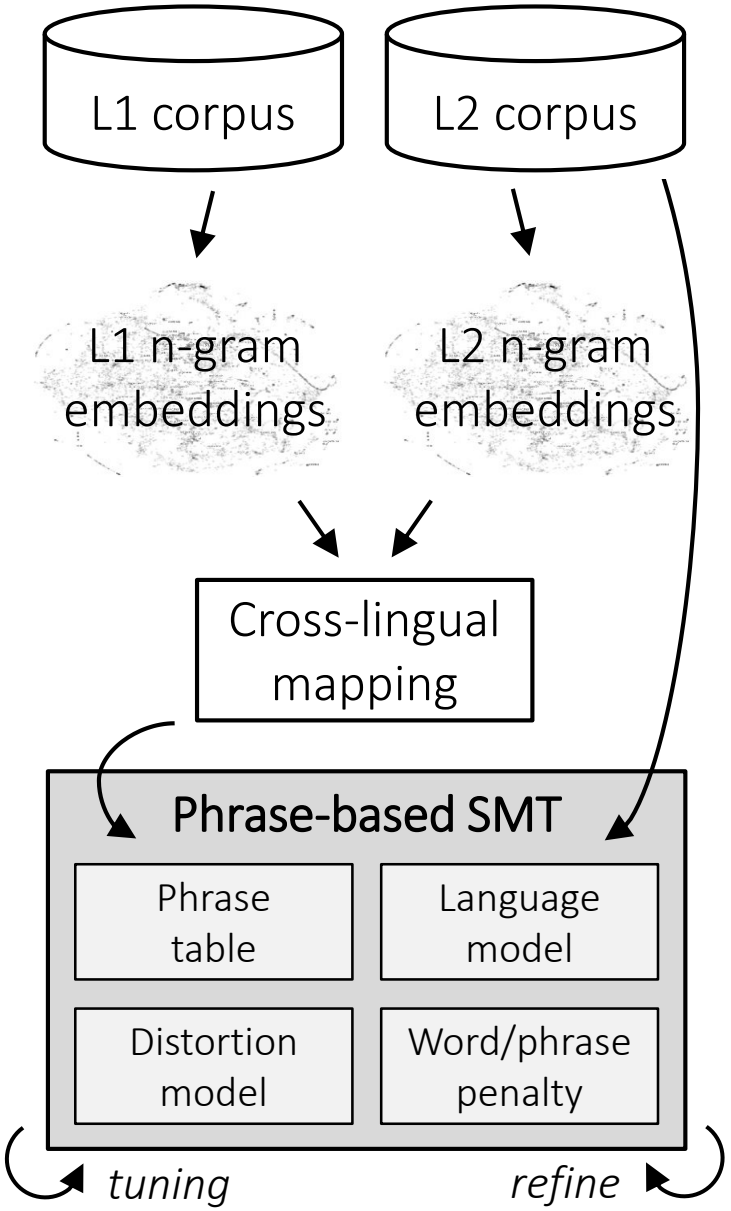


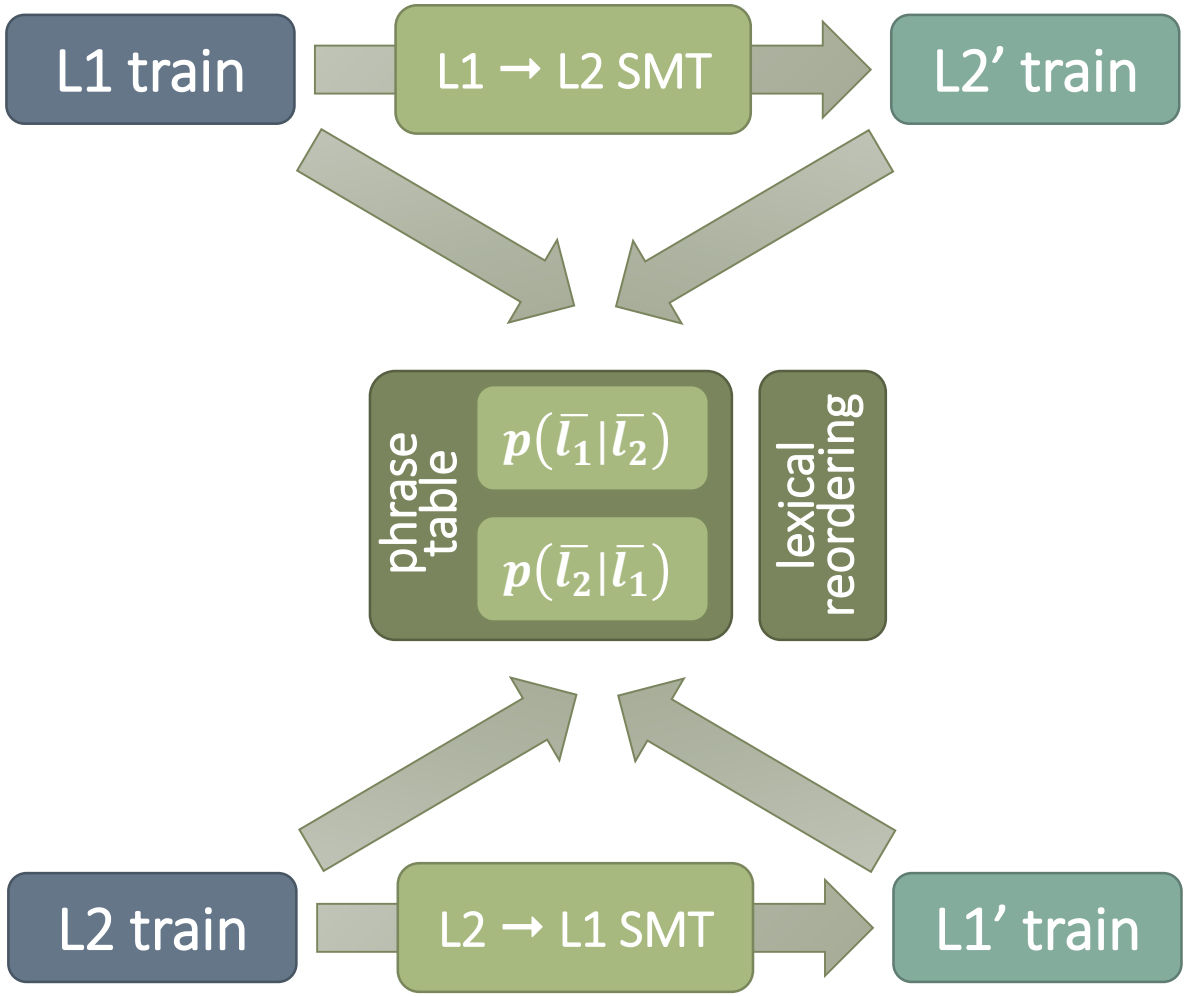
Unsupervised phrase-based SMT



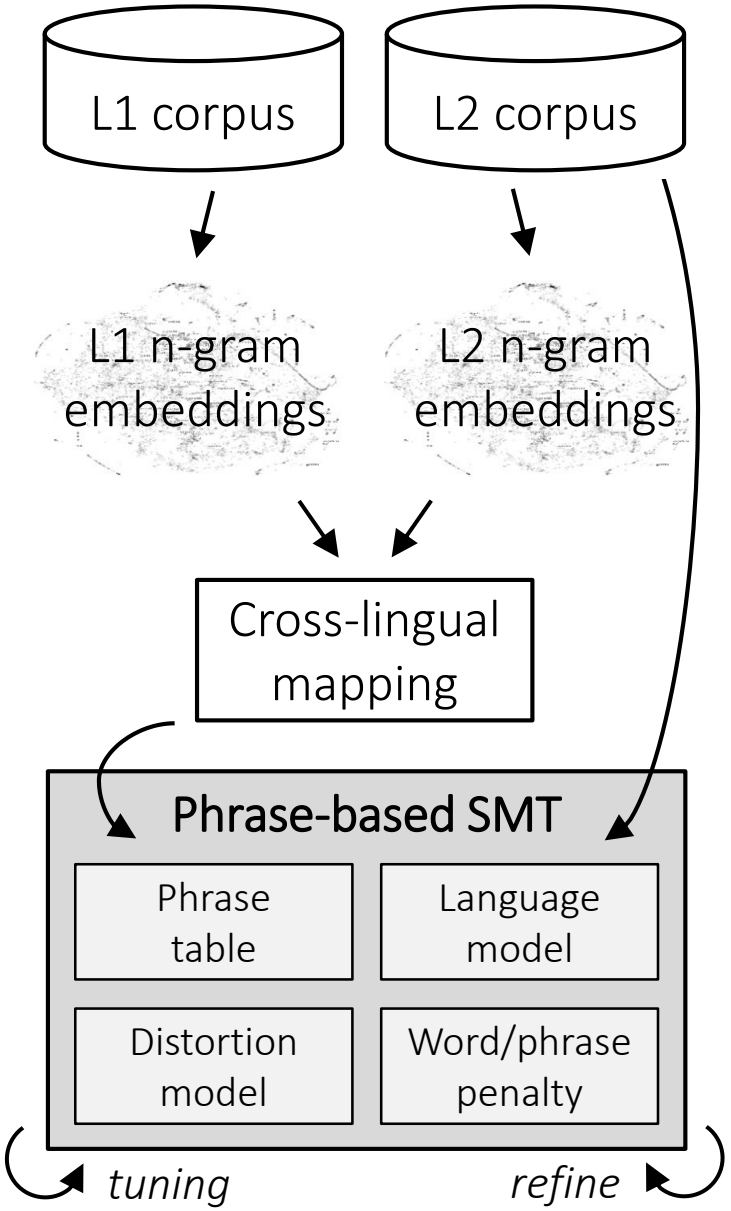


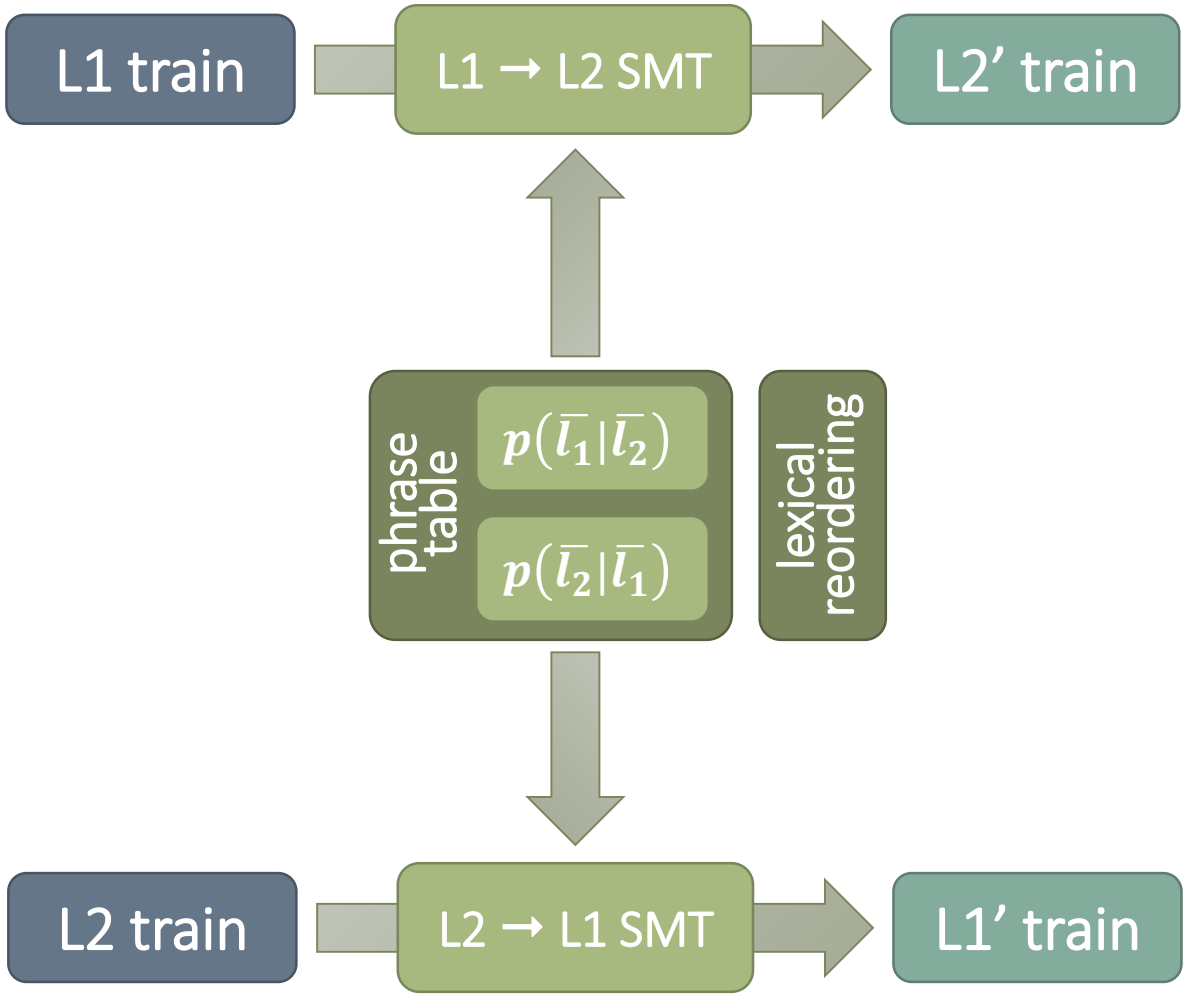
Unsupervised phrase-based SMT



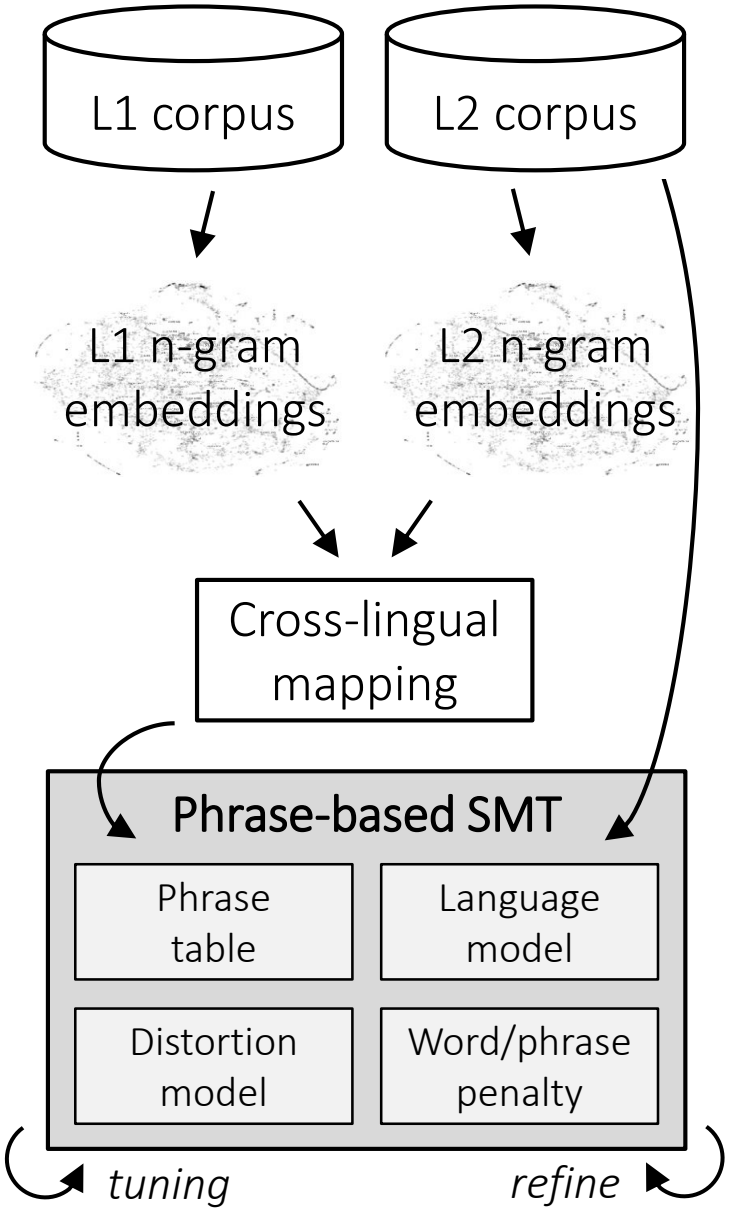


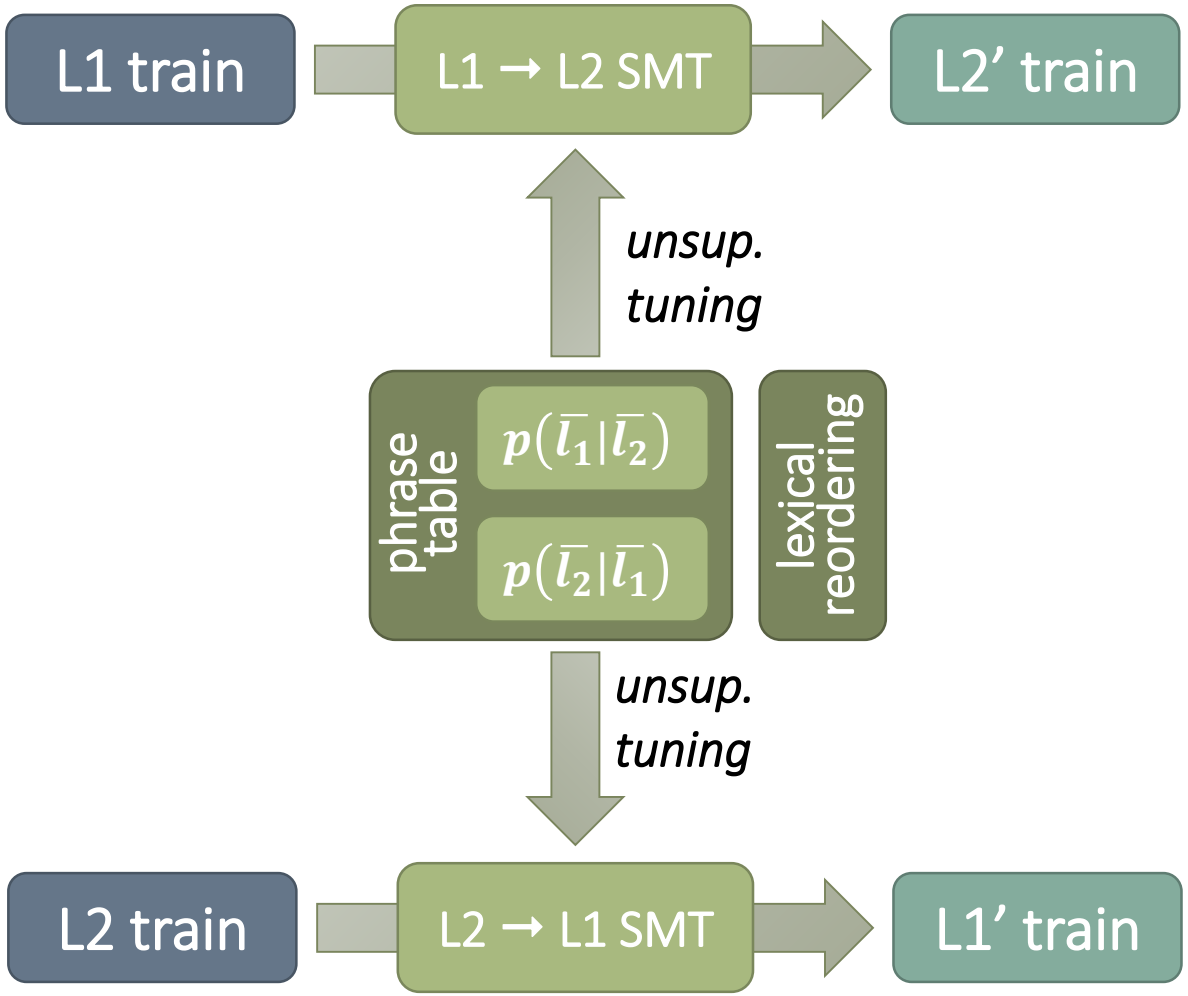
Unsupervised phrase-based SMT

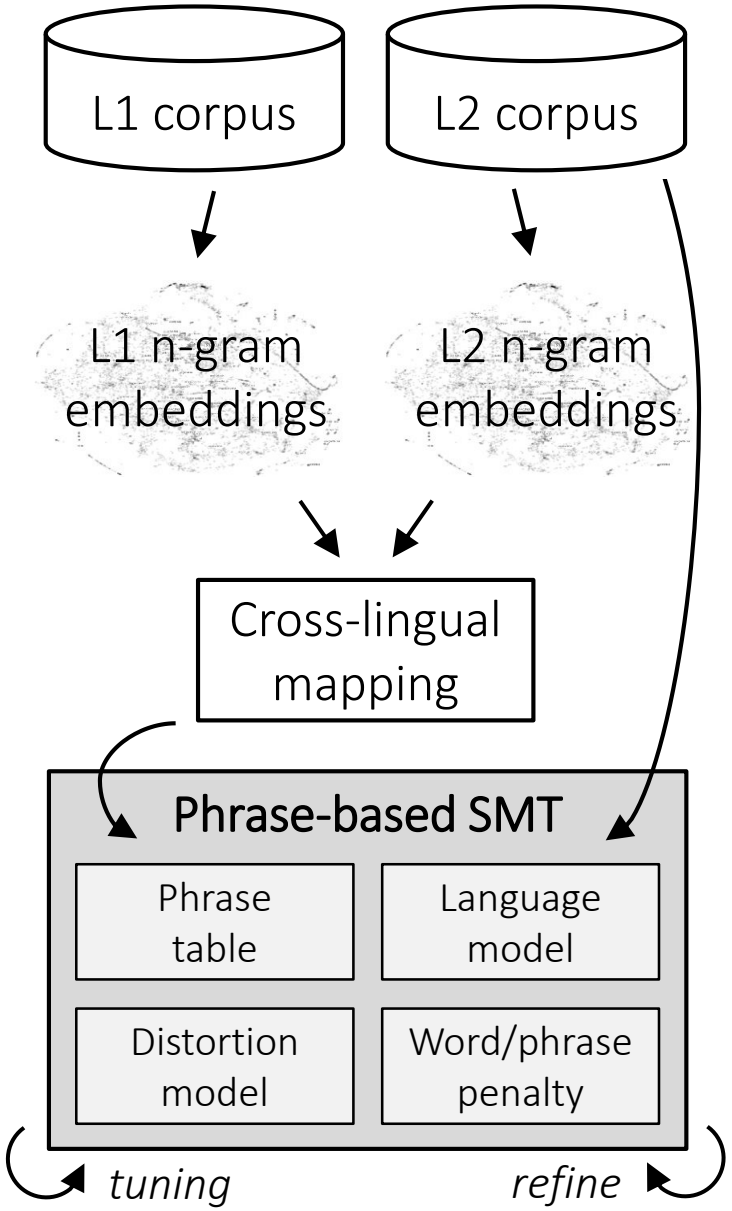




Unsupervised phrase-based SMT







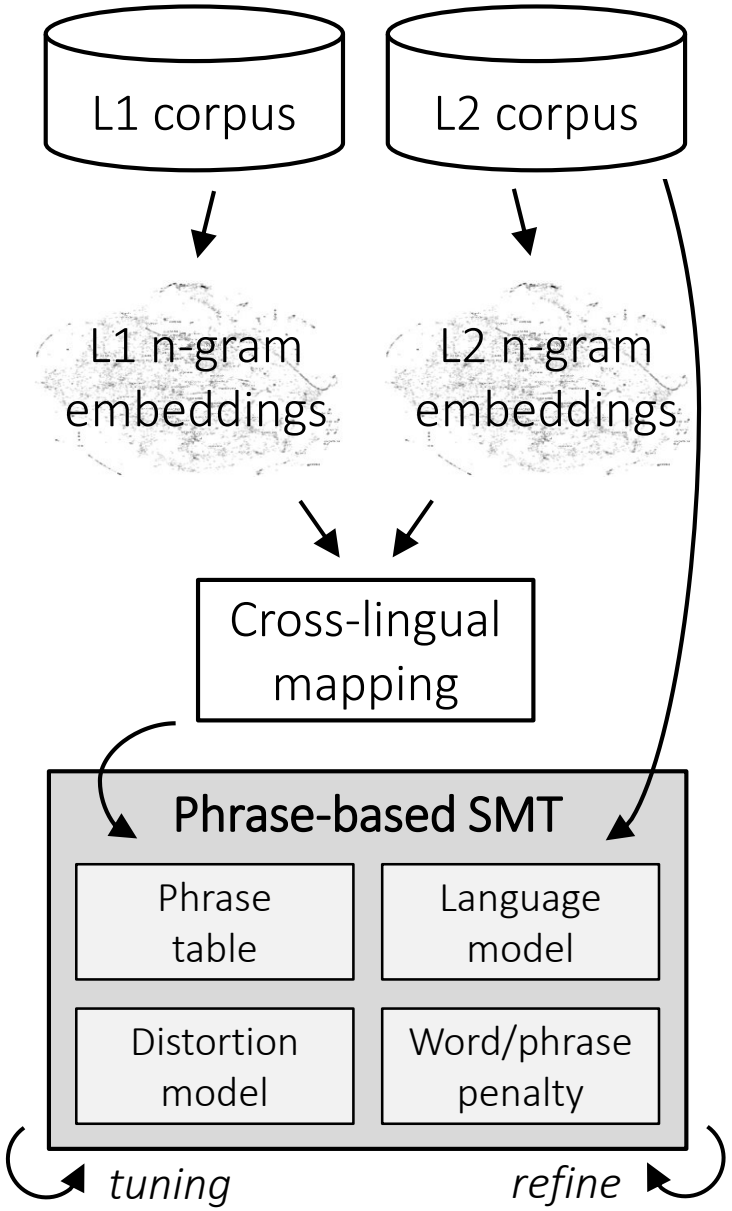
Unsupervised phrase-based SMT

EXPERIMENTS

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Initial SMT	22.4	19.6	15.3	11.0
+ Tuning	23.4	21.9	15.4	11.2

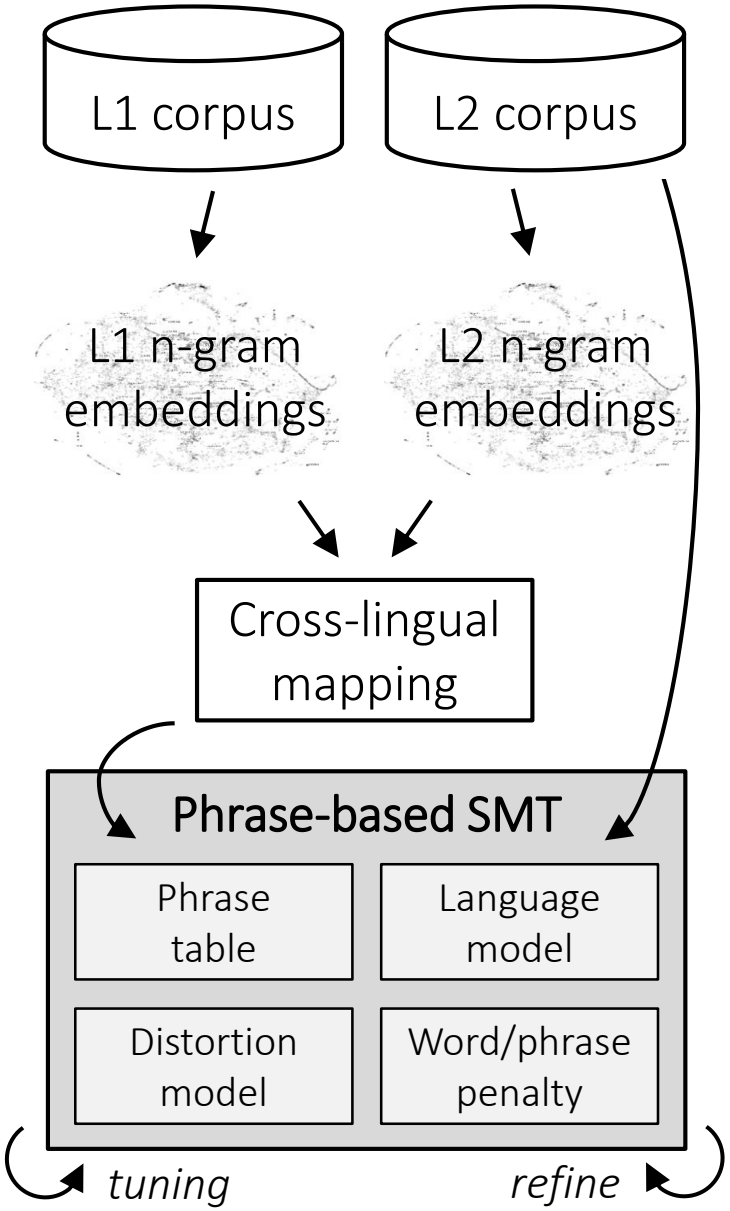
Unsupervised phrase-based SMT

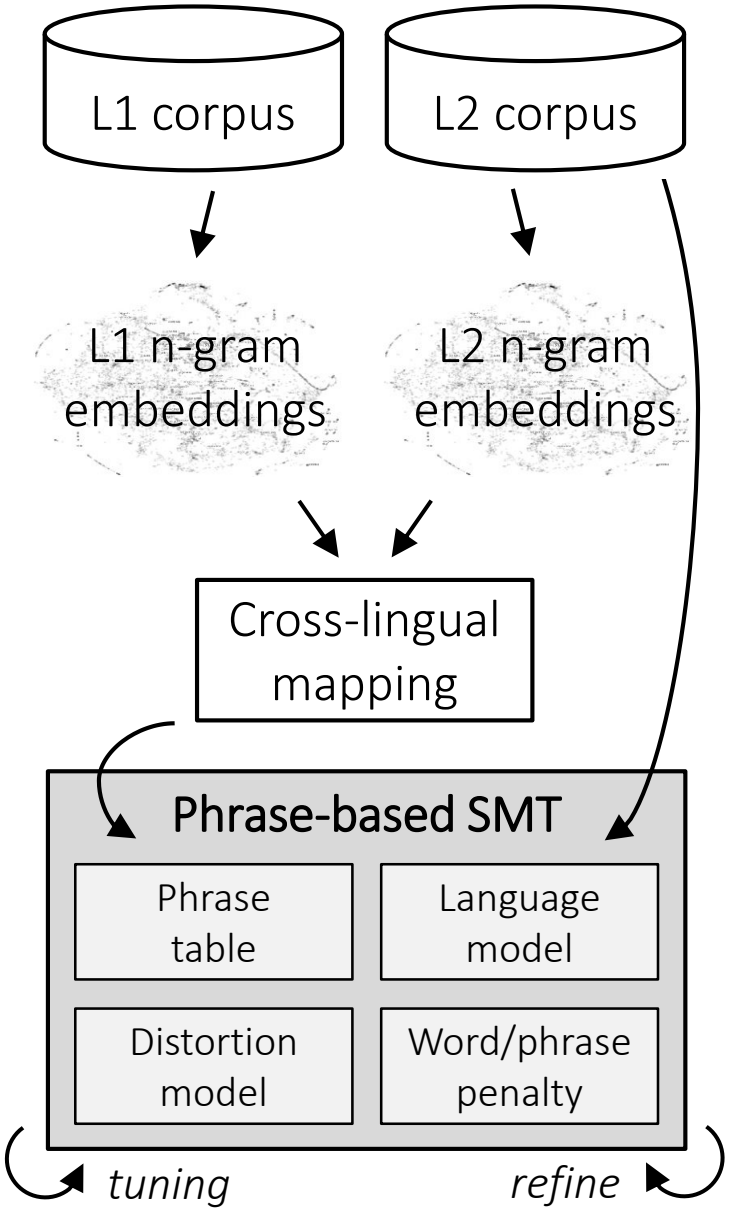


EXPERIMENTS

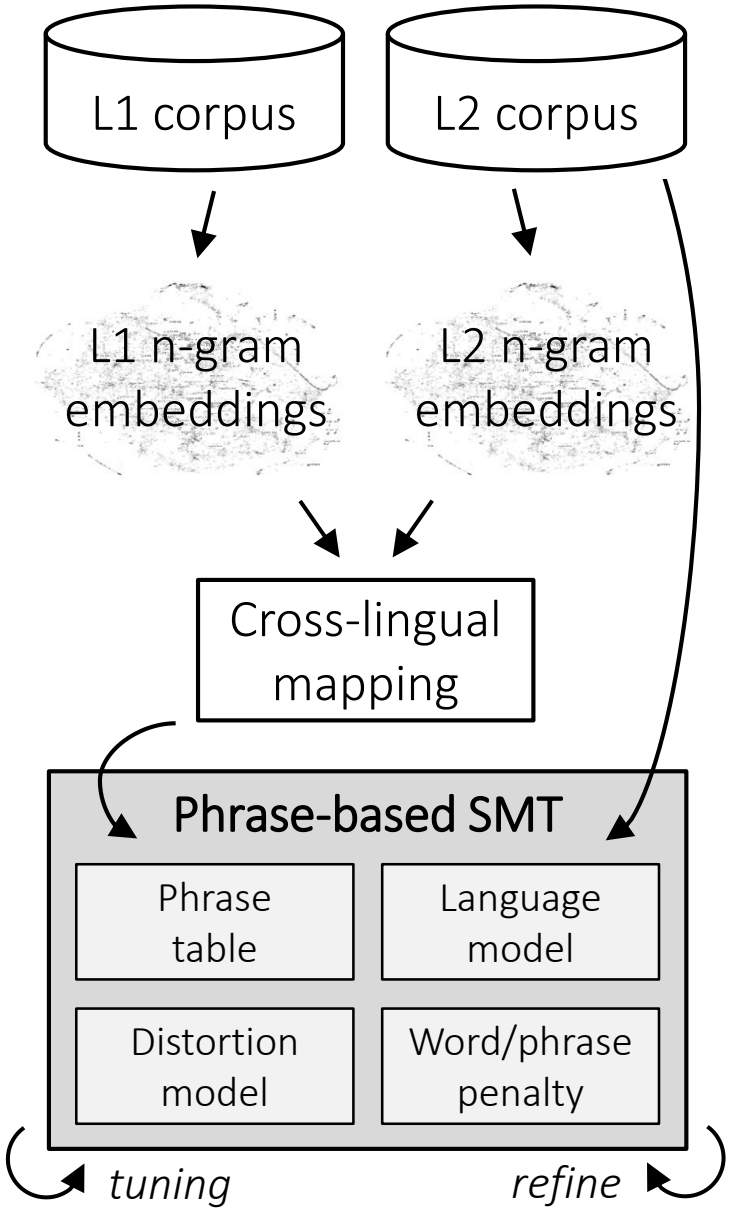
- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Initial SMT	22.4	19.6	15.3	11.0
+ Tuning	23.4	21.9	15.4	11.2
+ Refinement	27.9	27.8	19.7	14.7


L1 corpus
L2 corpus
L1 n-gram embeddings
L2 n-gram embeddings
Cross-lingual mapping
Phrase-based SMT
Phrase table
Language model
Distortion model
Word/phrase penalty
tuning
refine

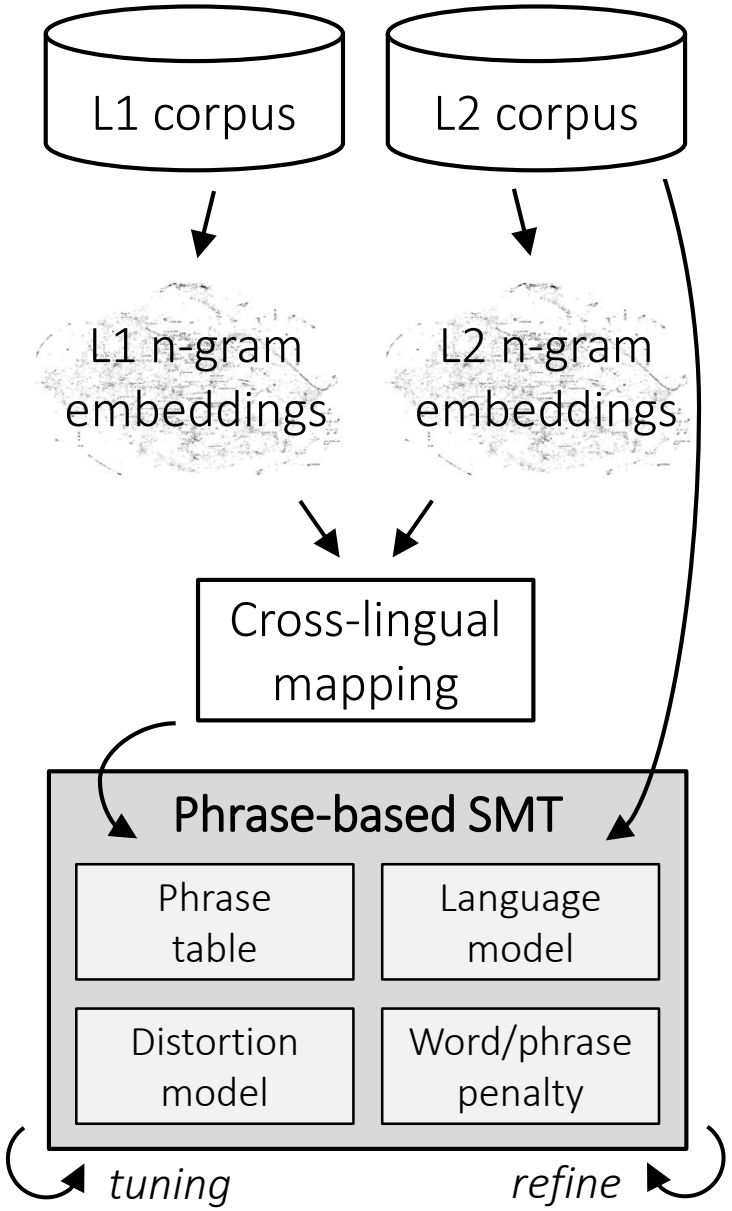

L1 corpus
L2 corpus
L1 n-gram embeddings
L2 n-gram embeddings
Cross-lingual mapping
Phrase-based SMT
Phrase table
Language model
Distortion model
Word/phrase penalty
tuning
refine

but…


L1 corpus
L2 corpus
L1 n-gram embeddings
L2 n-gram embeddings
Cross-lingual mapping
Phrase-based SMT
Phrase table
Language model
Distortion model
Word/phrase penalty
tuning
refine

but…

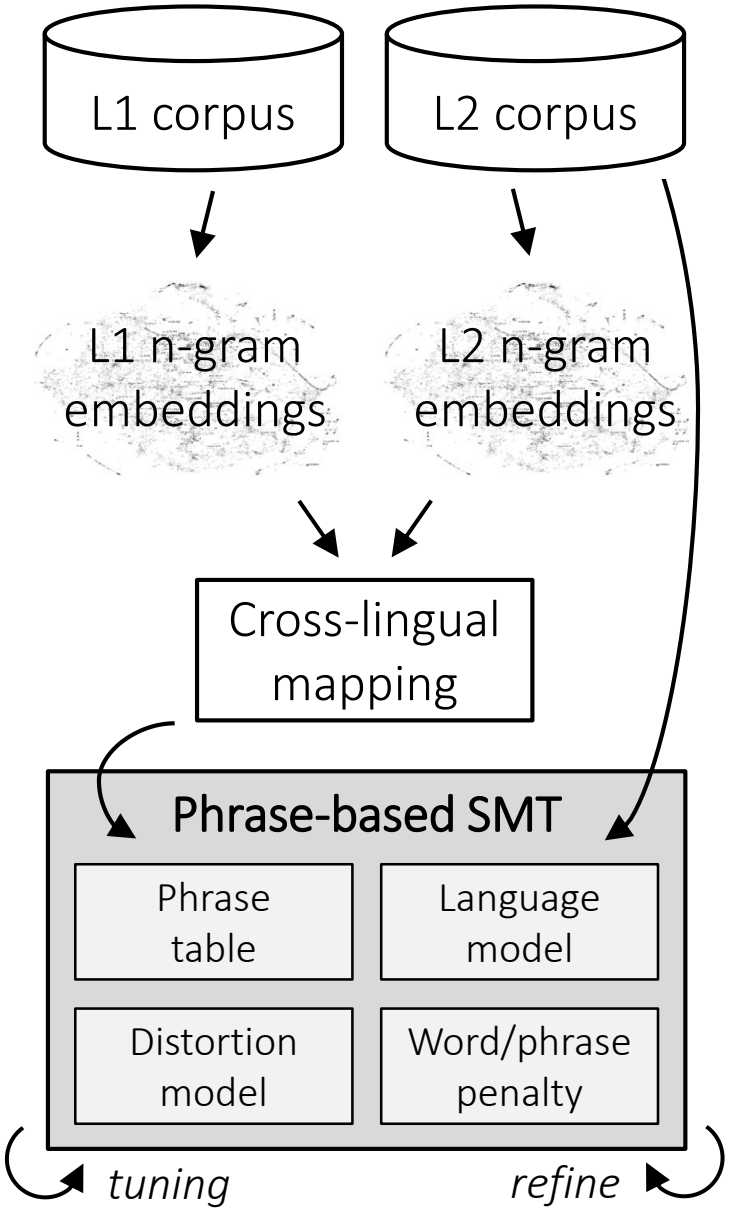
NMT >> SMT

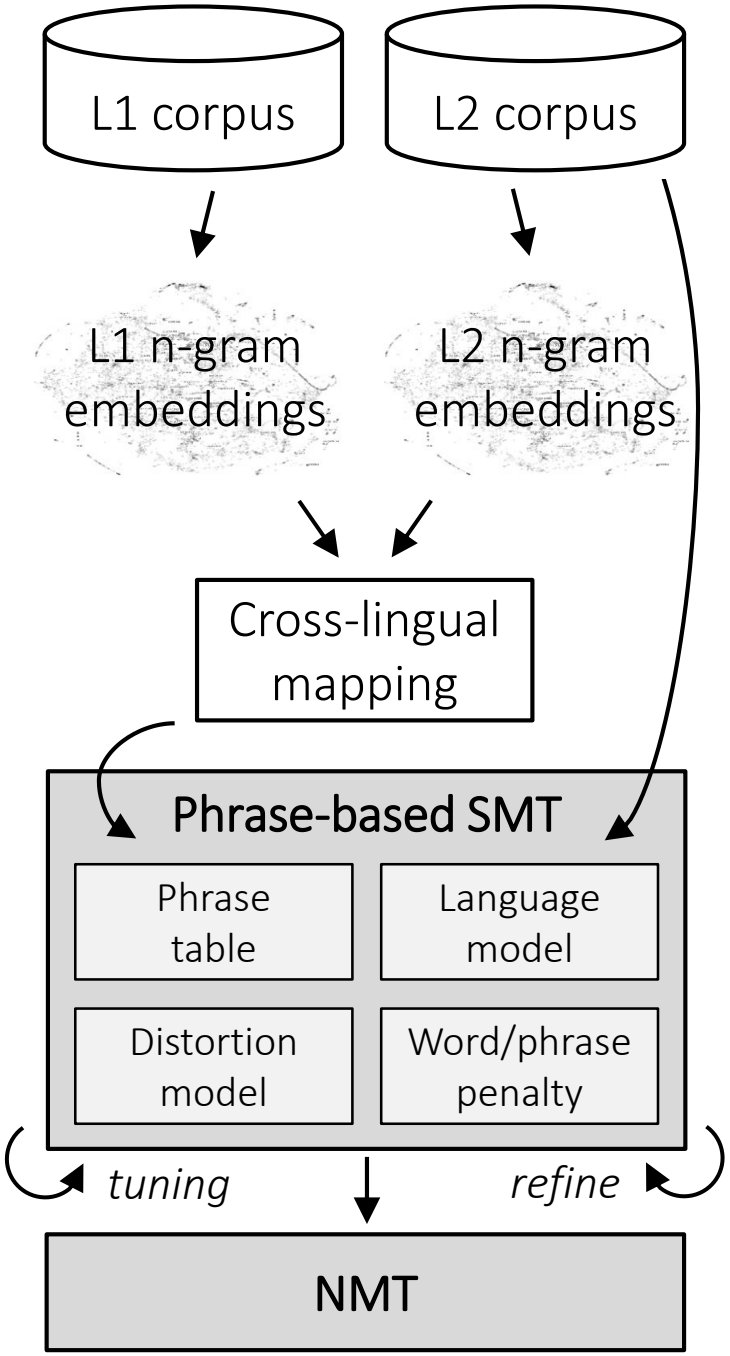


but…

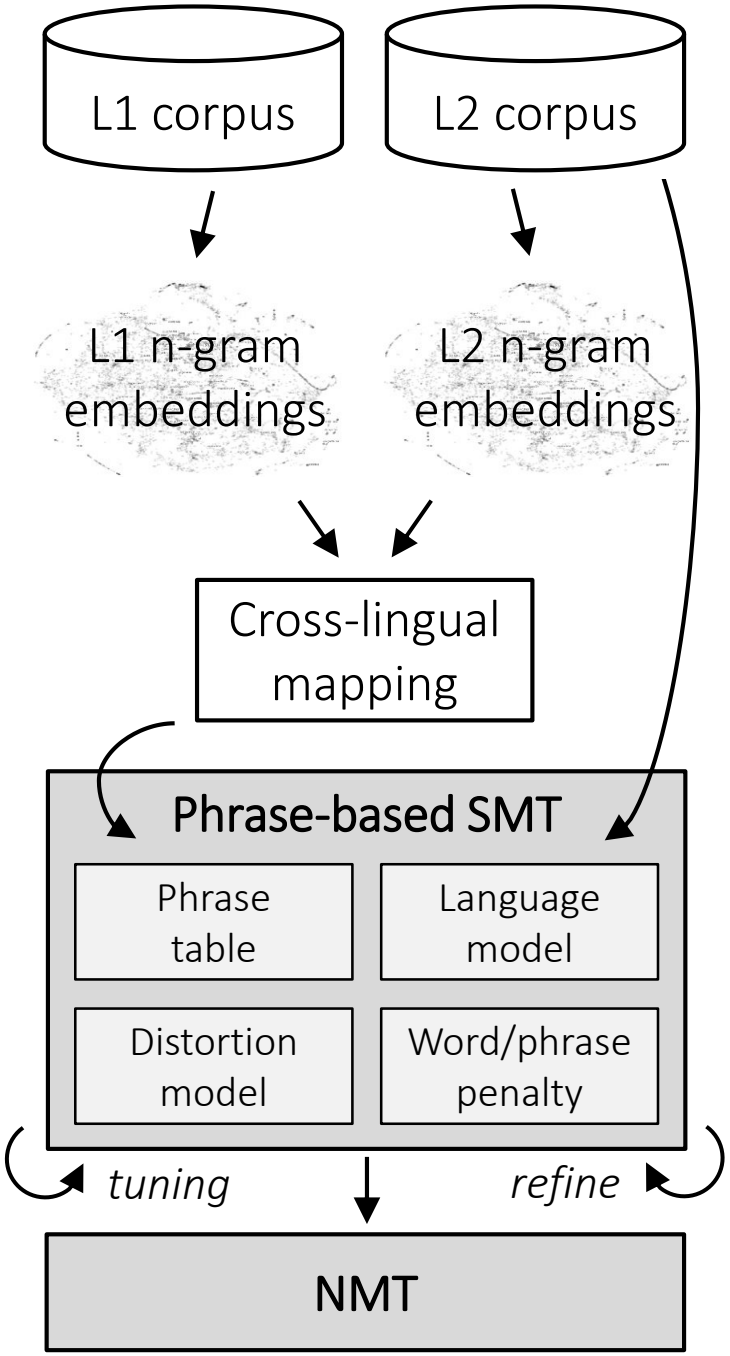
NMT >> SMT

(unsupervised) SMT has a hard ceiling!

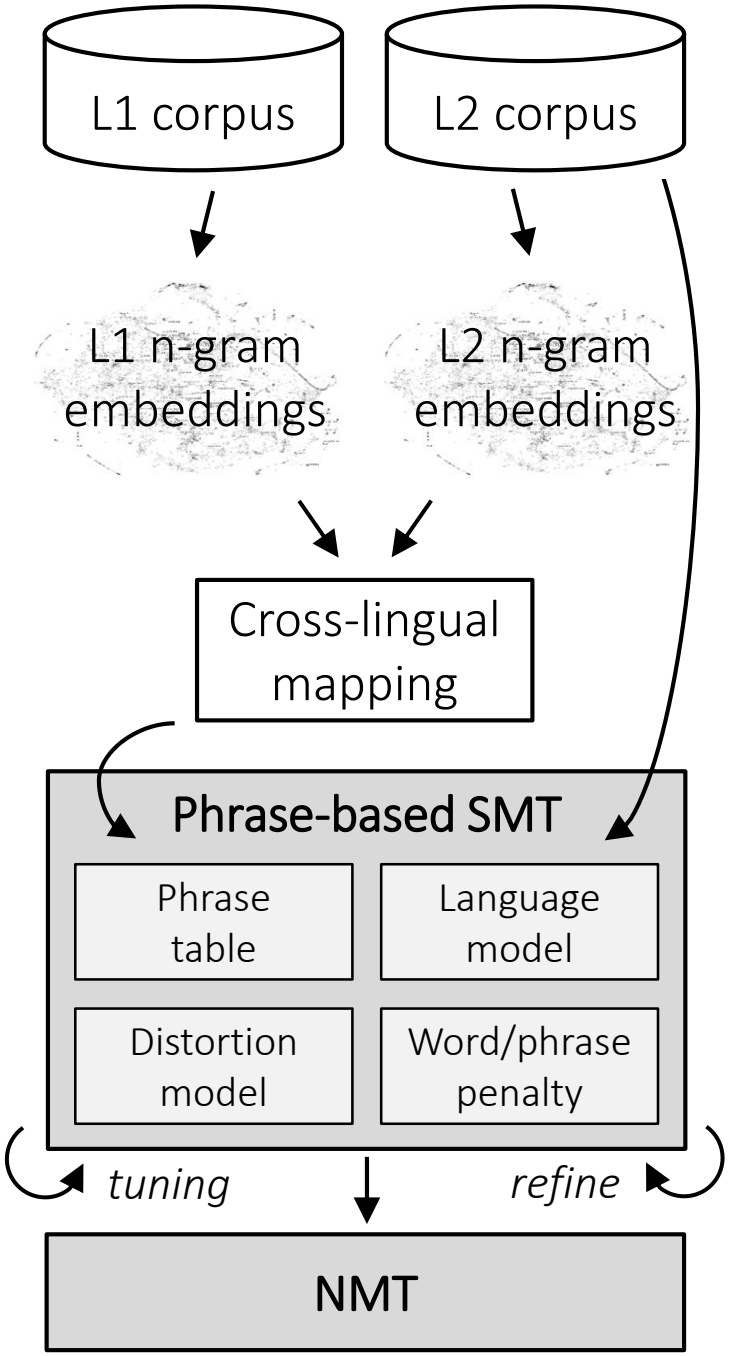

L1 corpus
L2 corpus
L1 n-gram embeddings
L2 n-gram embeddings
Cross-lingual mapping
Phrase-based SMT
Phrase table
Language model
Distortion model
Word/phrase penalty
tuning
refine


L1 corpus
L2 corpus
L1 n-gram embeddings
L2 n-gram embeddings
Cross-lingual mapping
Phrase-based SMT
Phrase table
Language model
Distortion model
Word/phrase penalty
tuning
refine
NMT

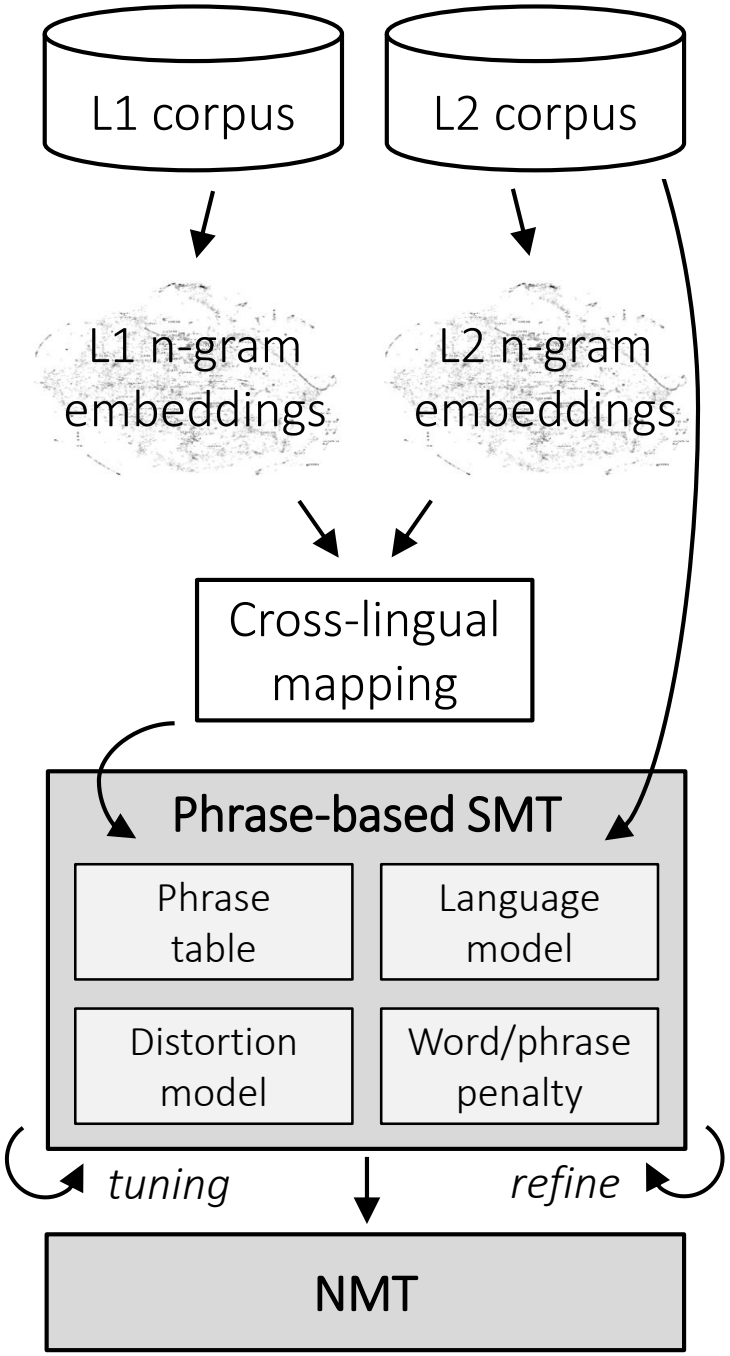
NMT hybridization



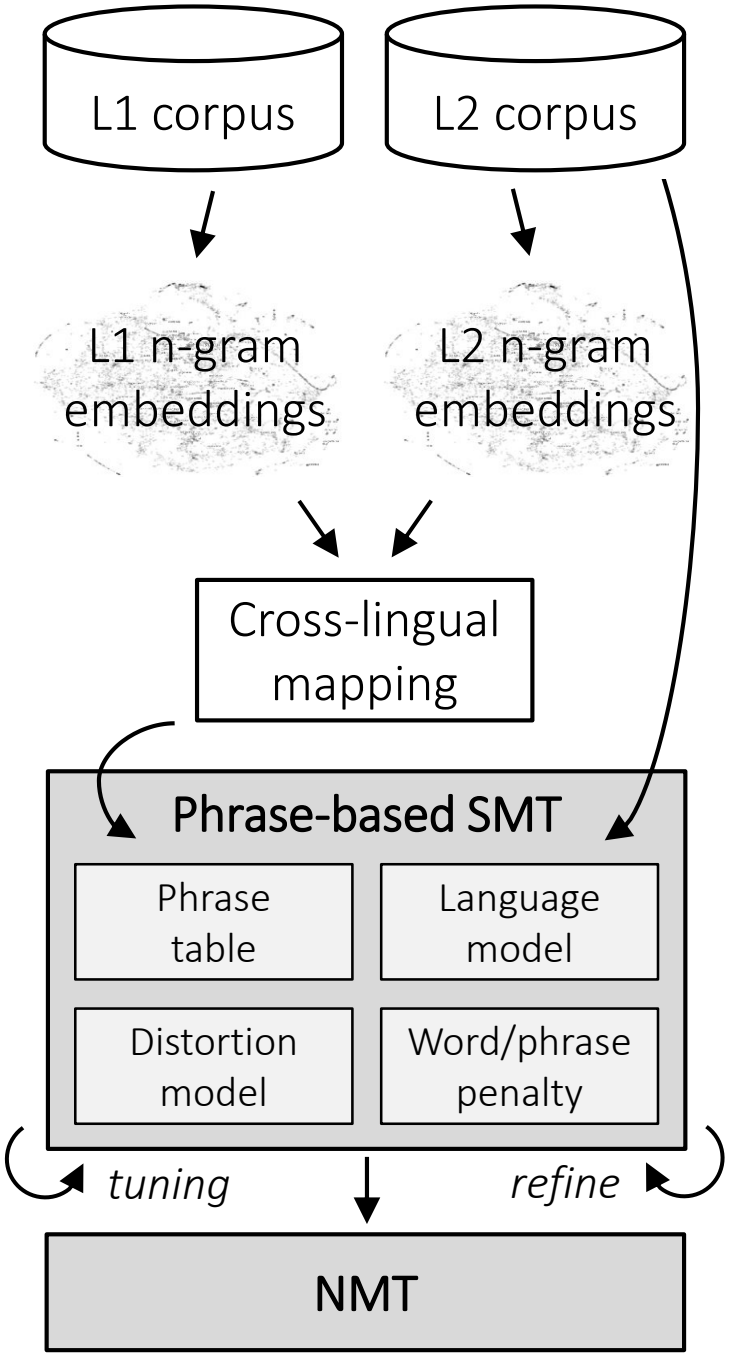
NMT hybridization

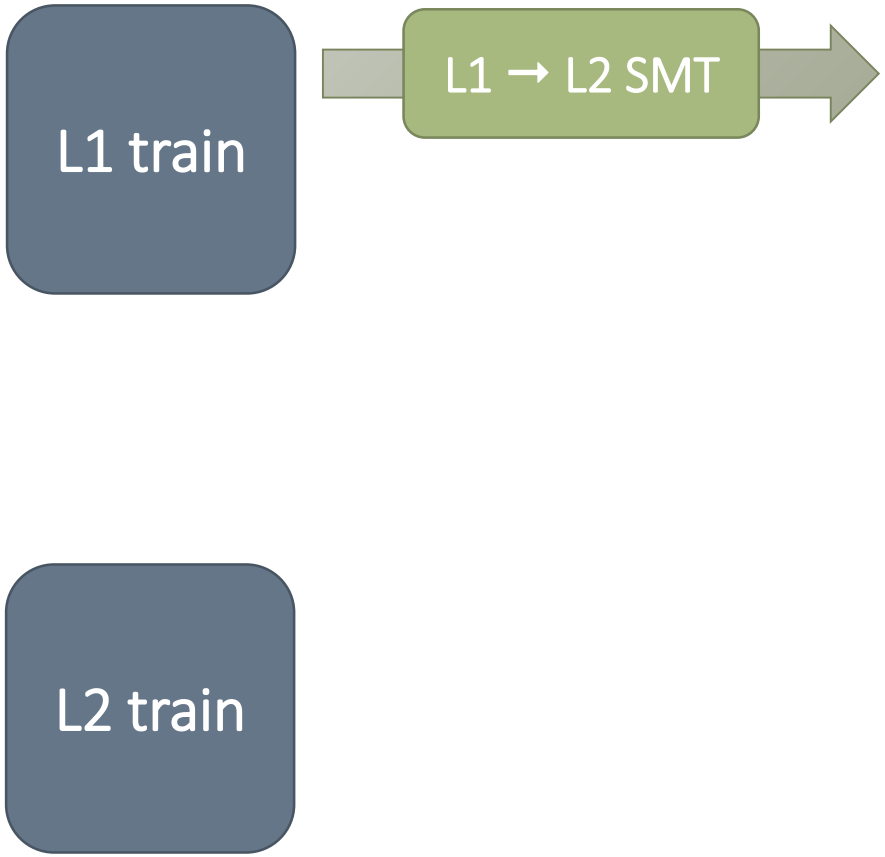


NMT hybridization

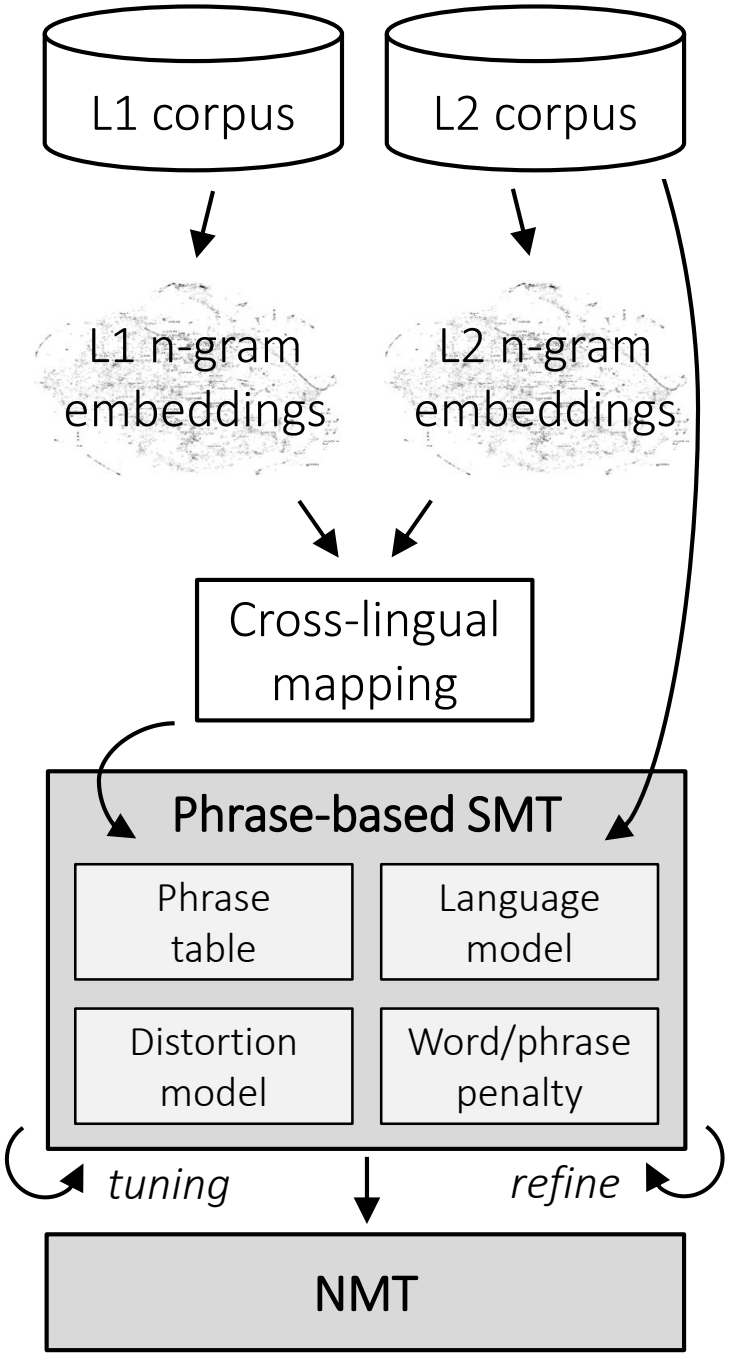


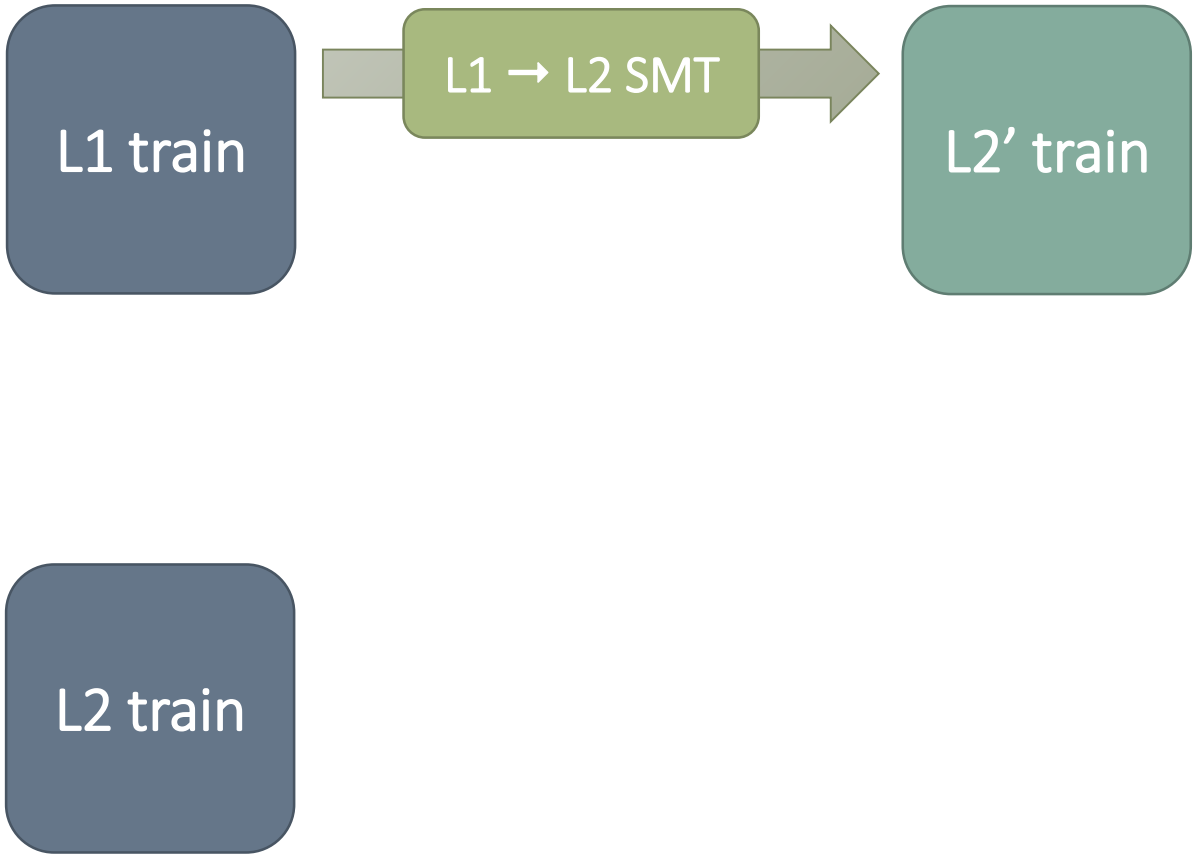
NMT hybridization



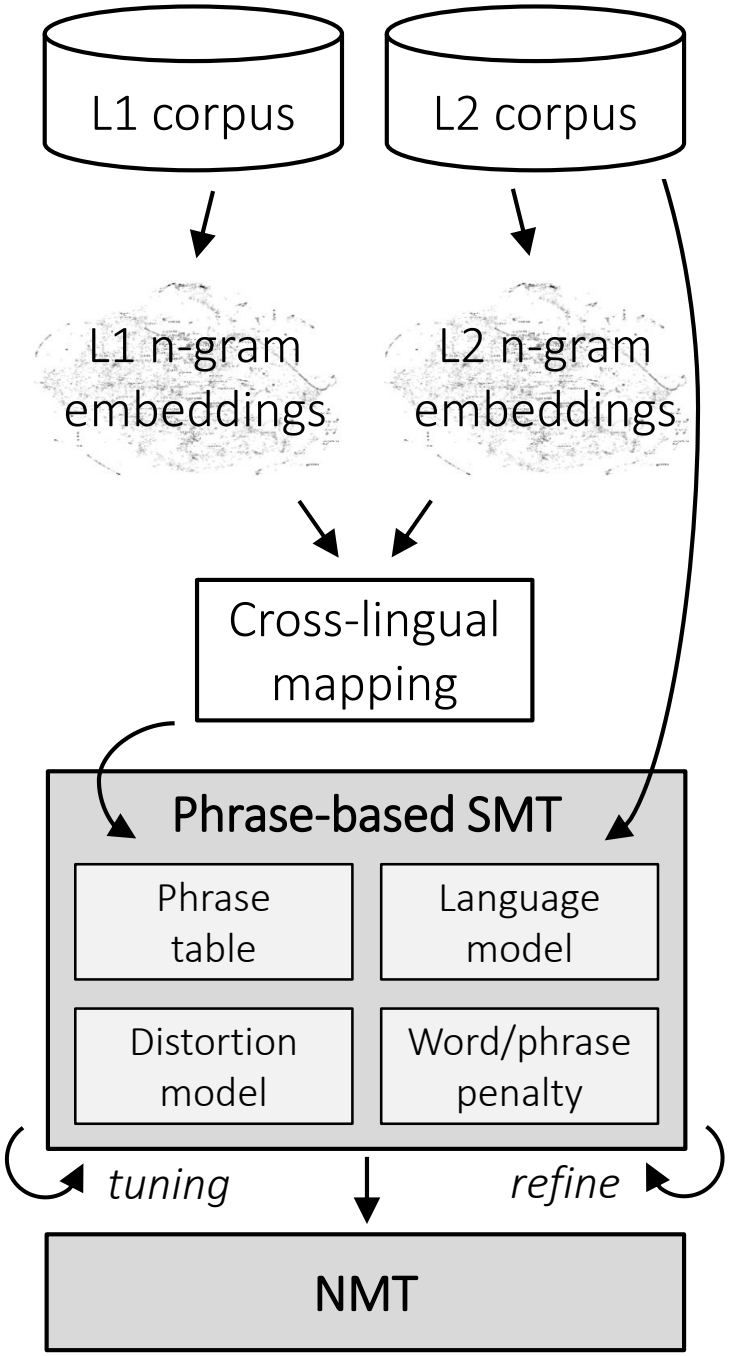


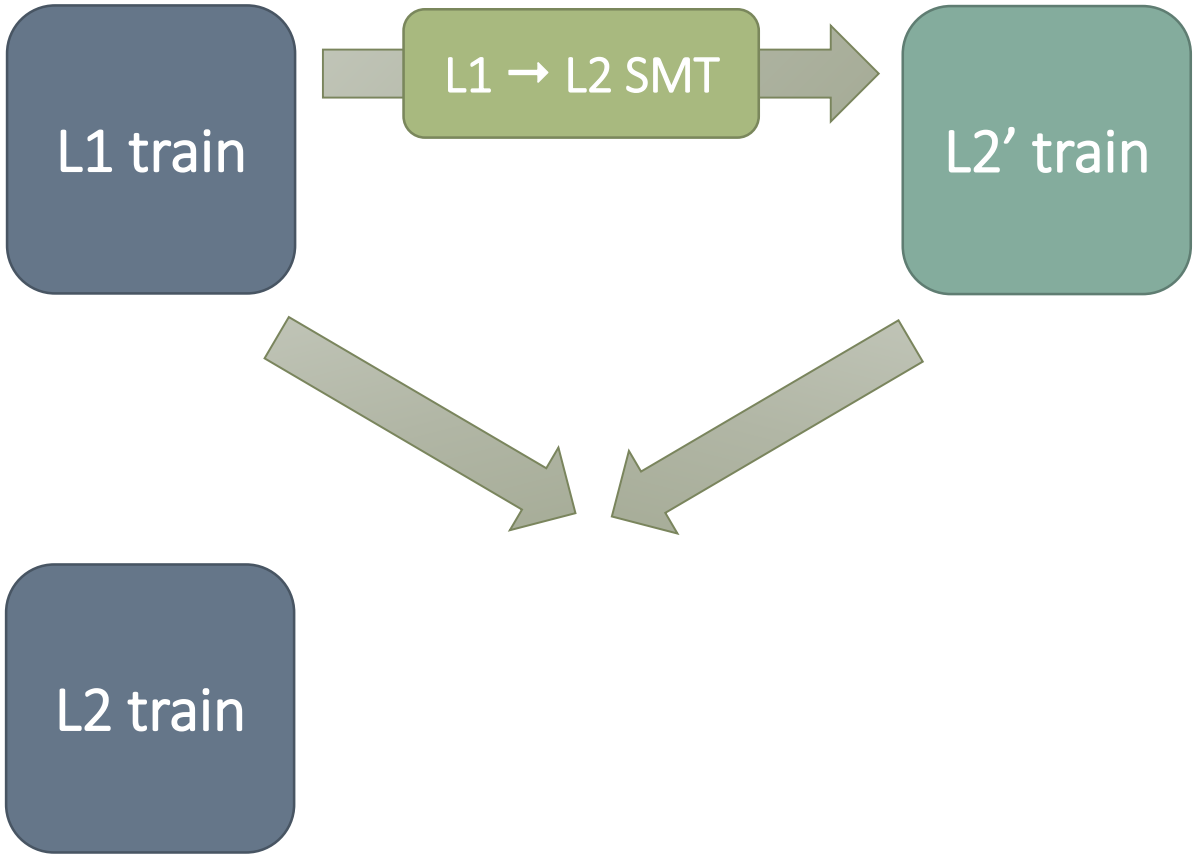
NMT hybridization



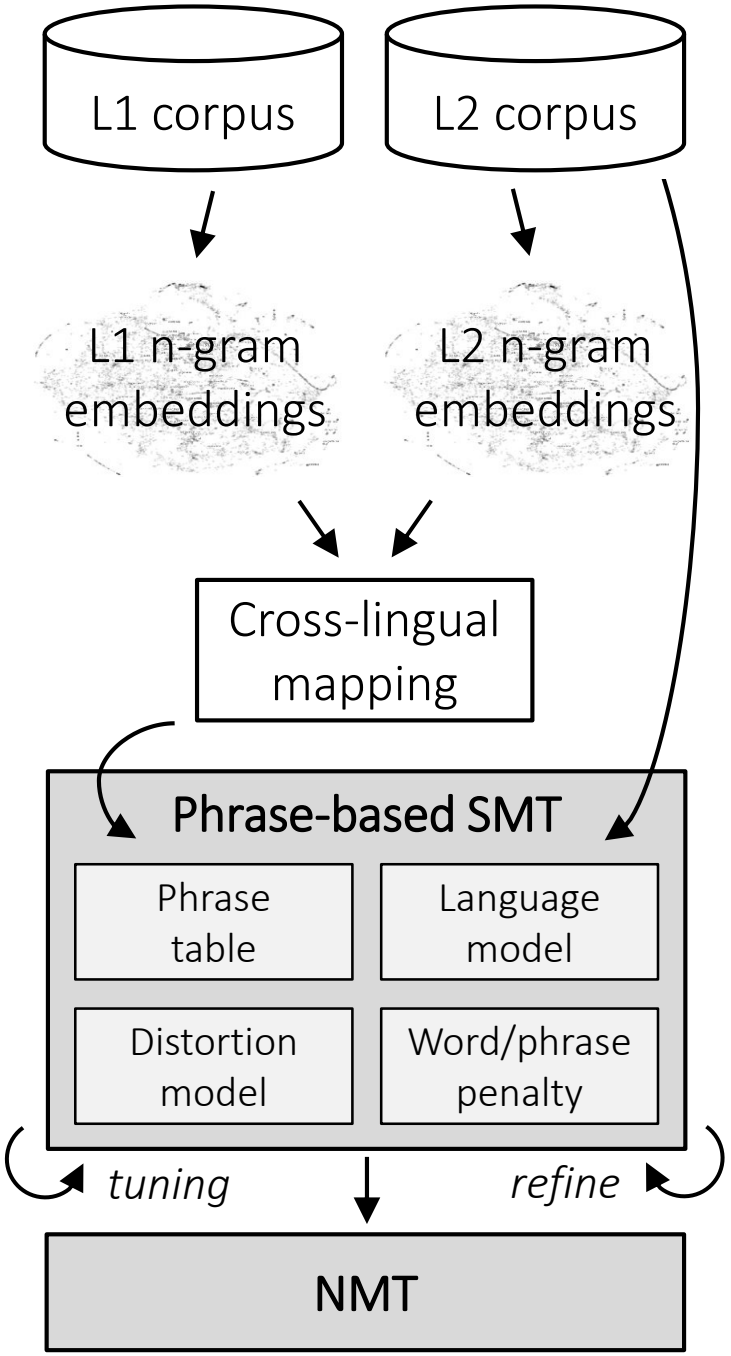


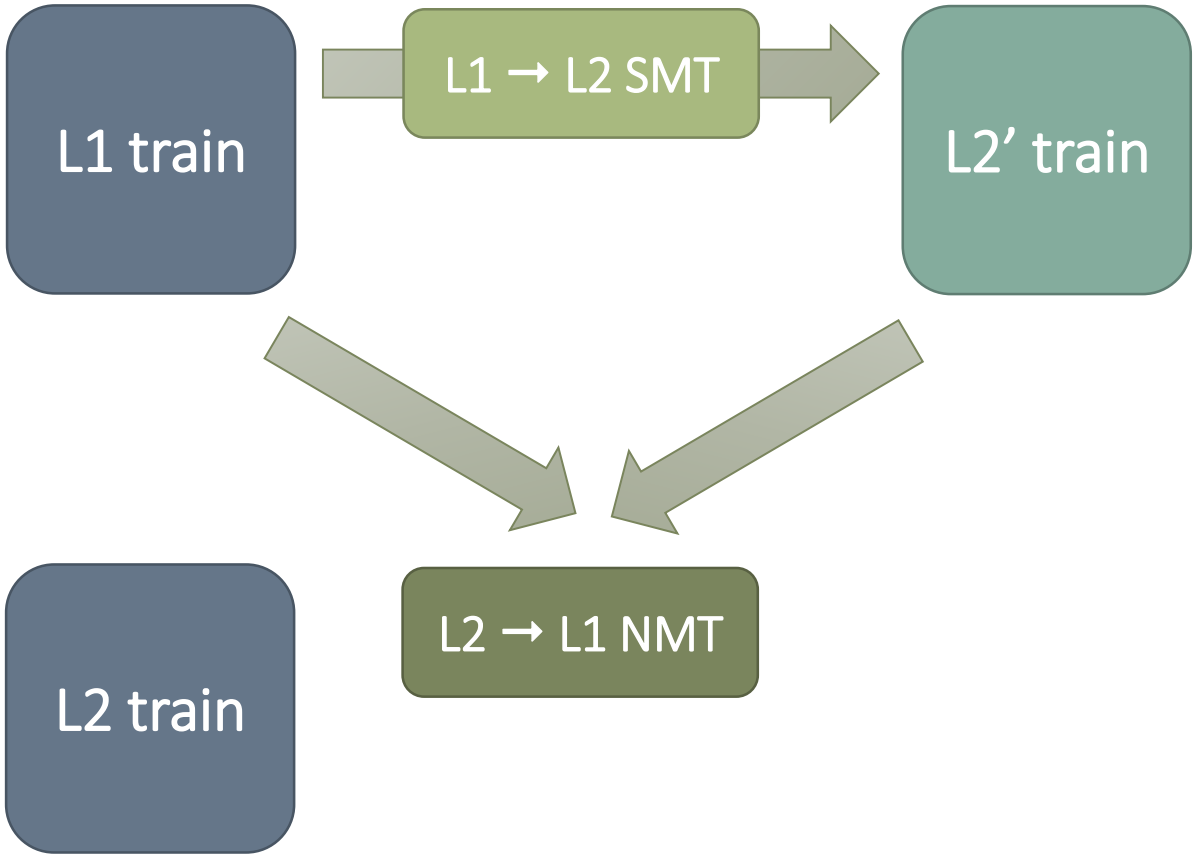
NMT hybridization



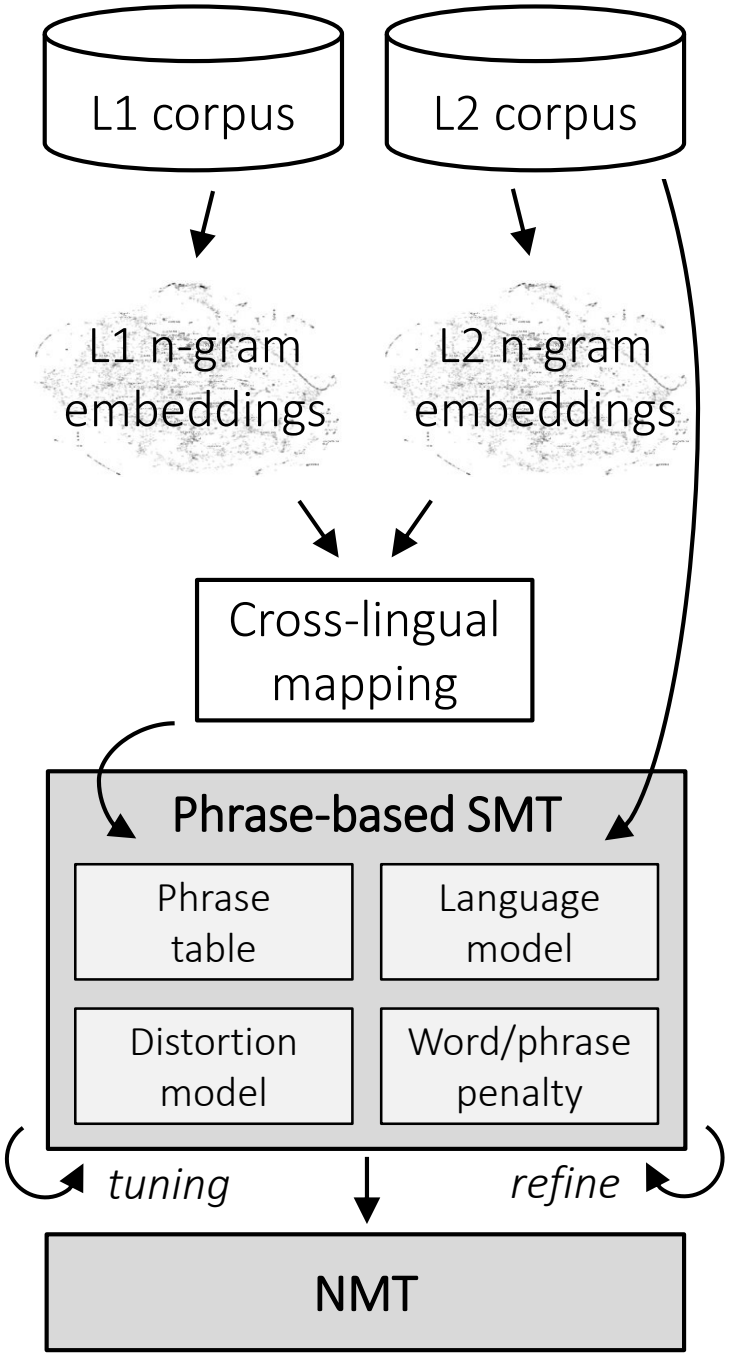


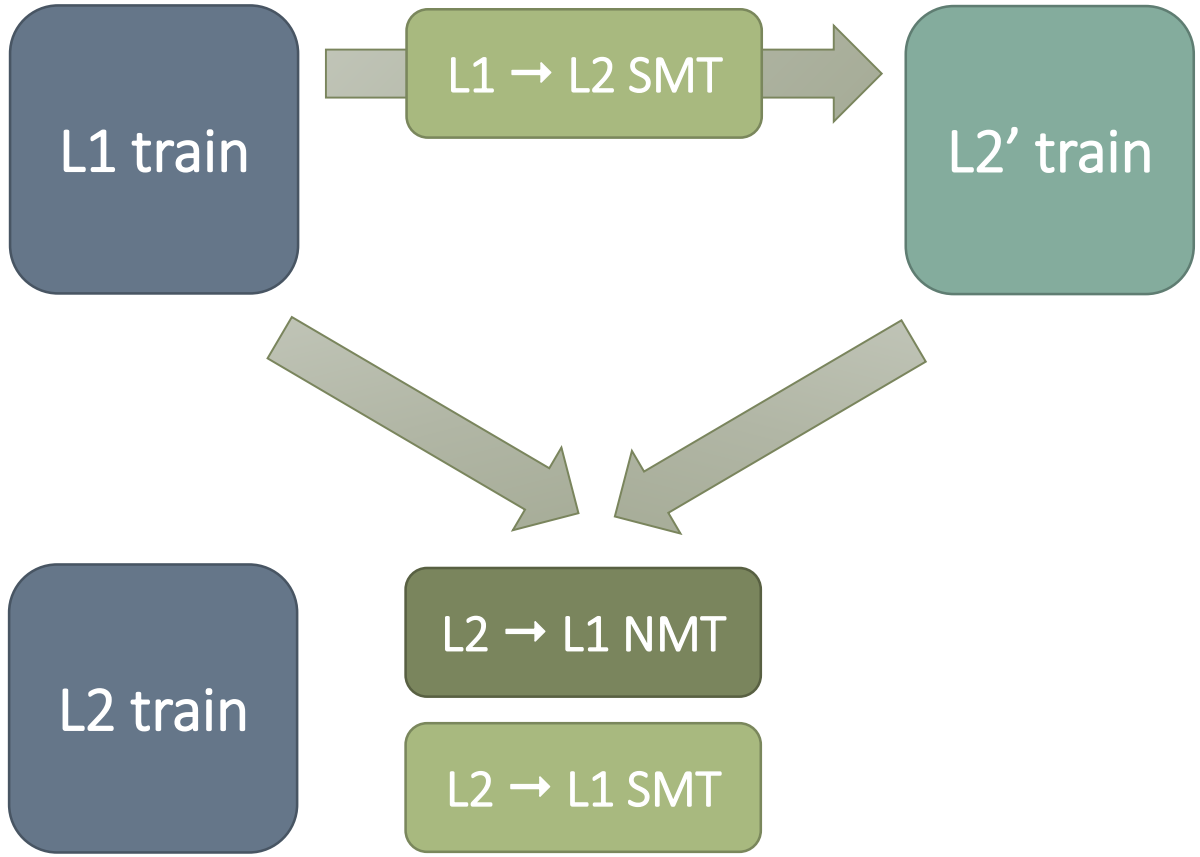
NMT hybridization



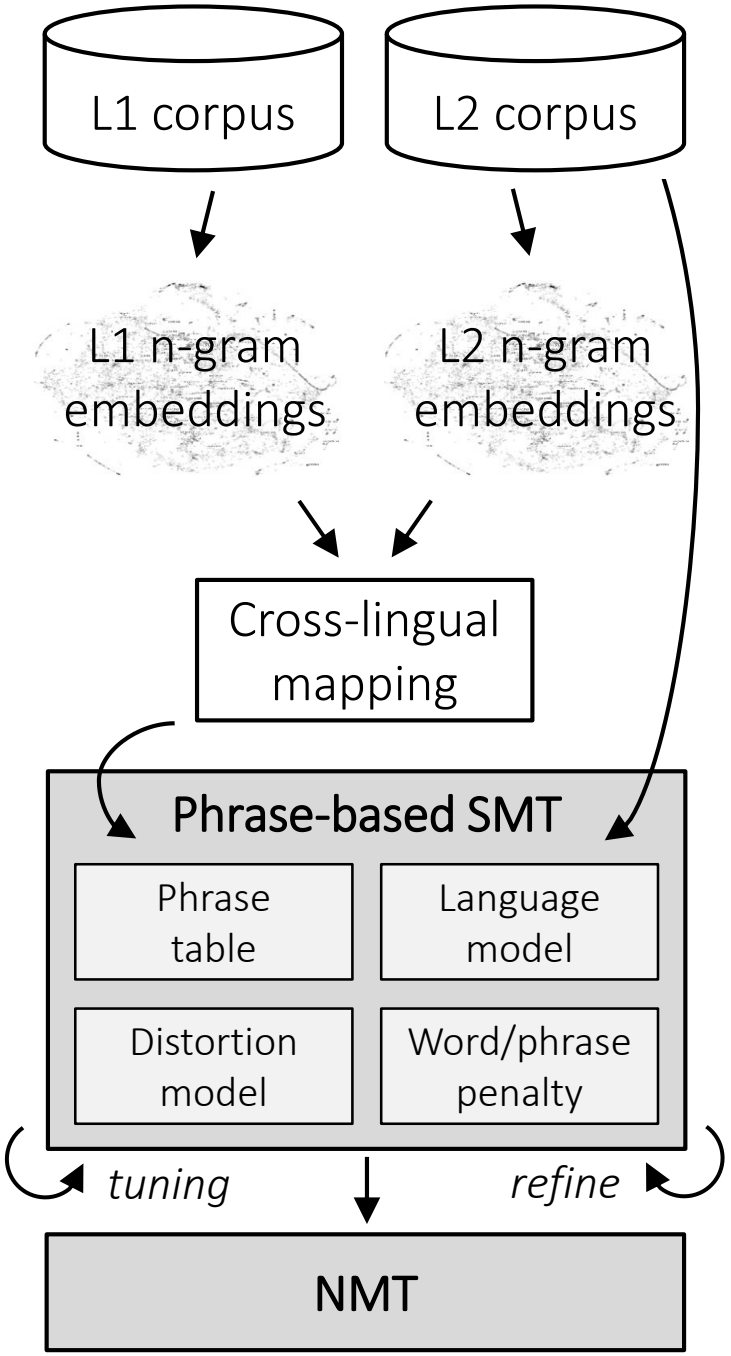


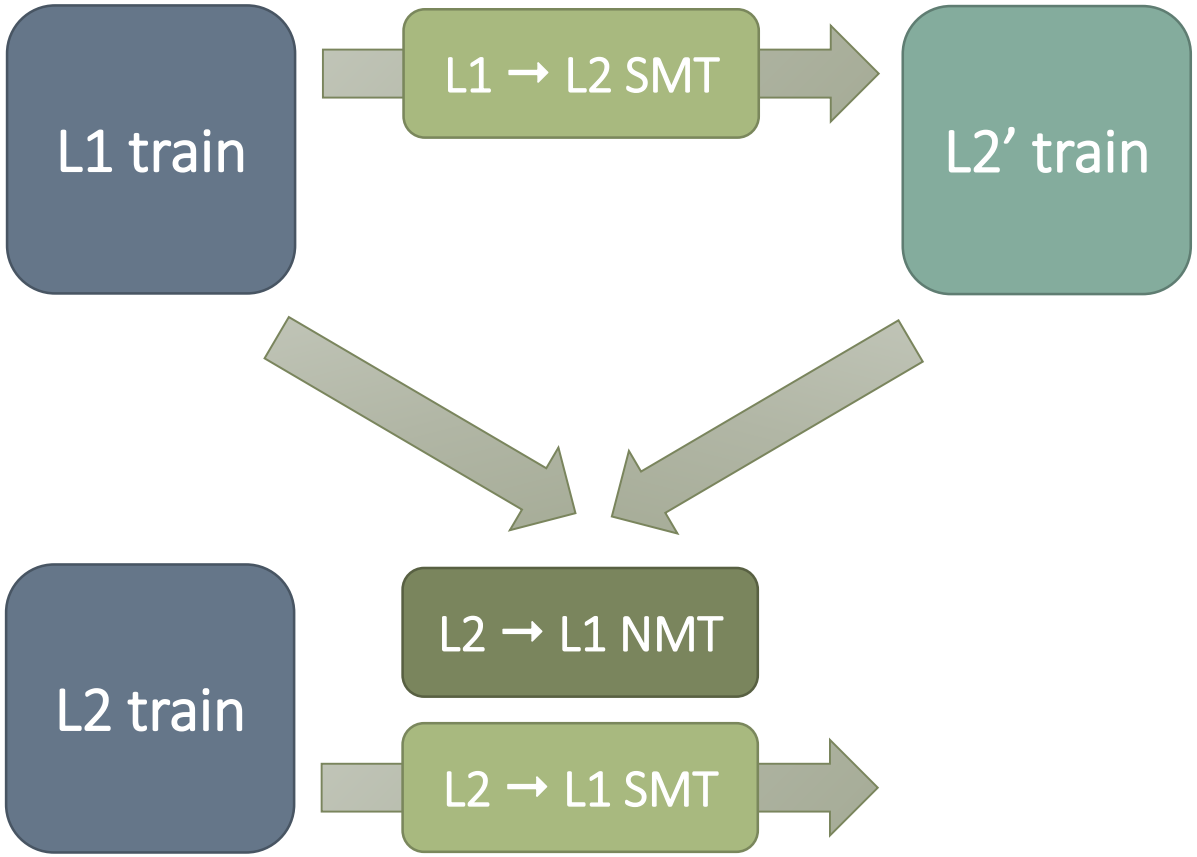
NMT hybridization



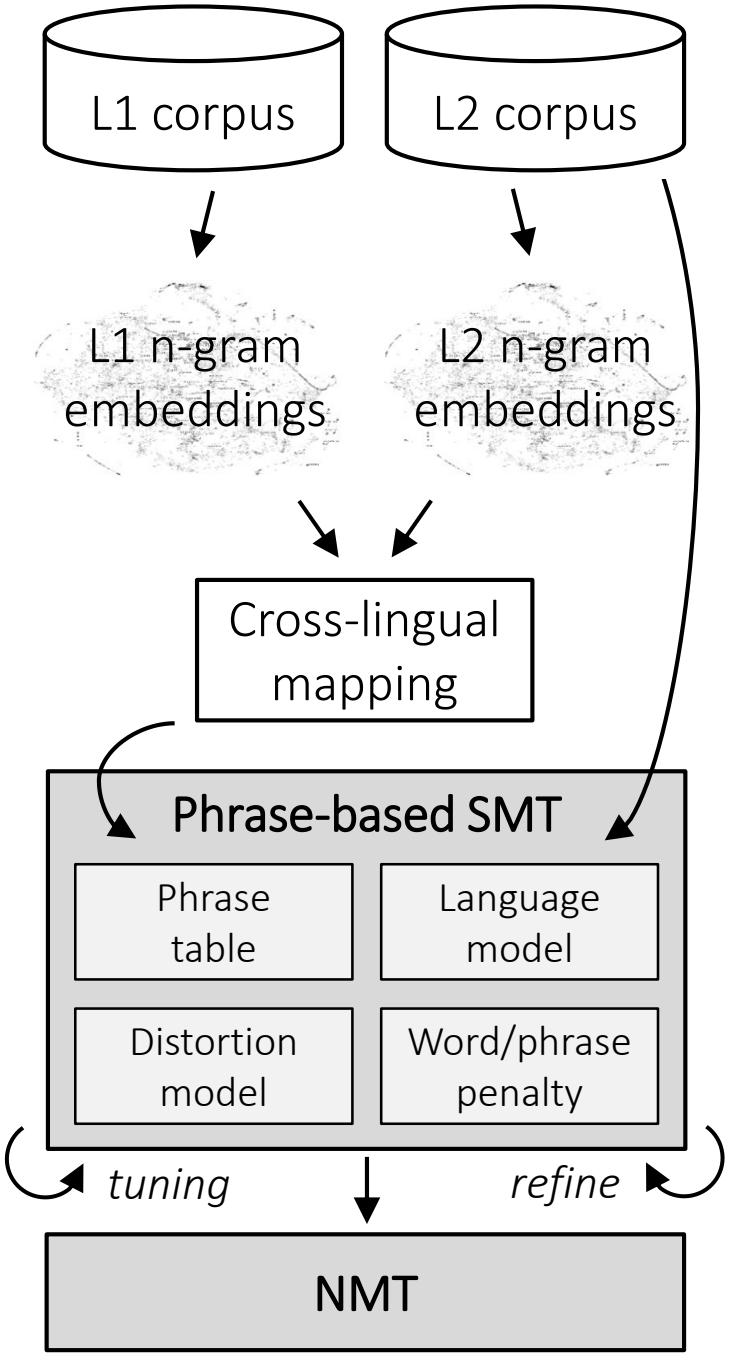


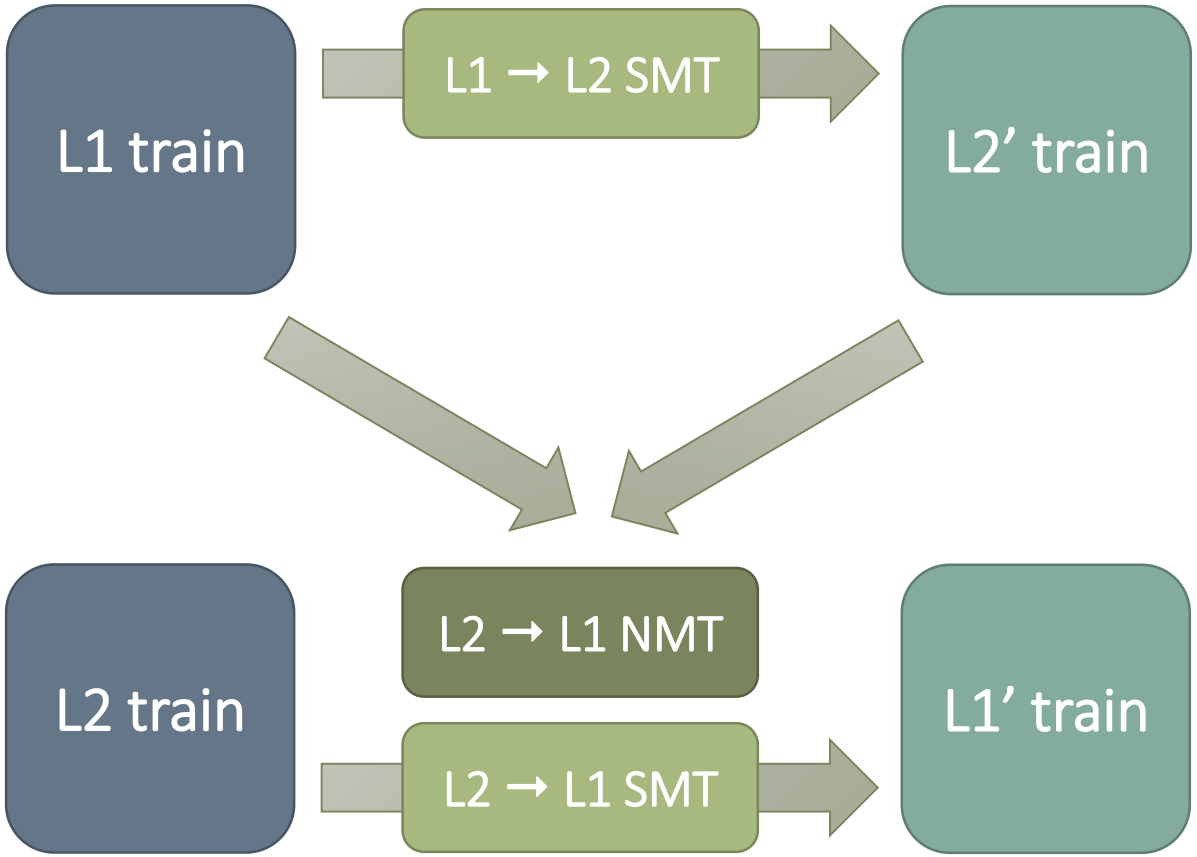
NMT hybridization



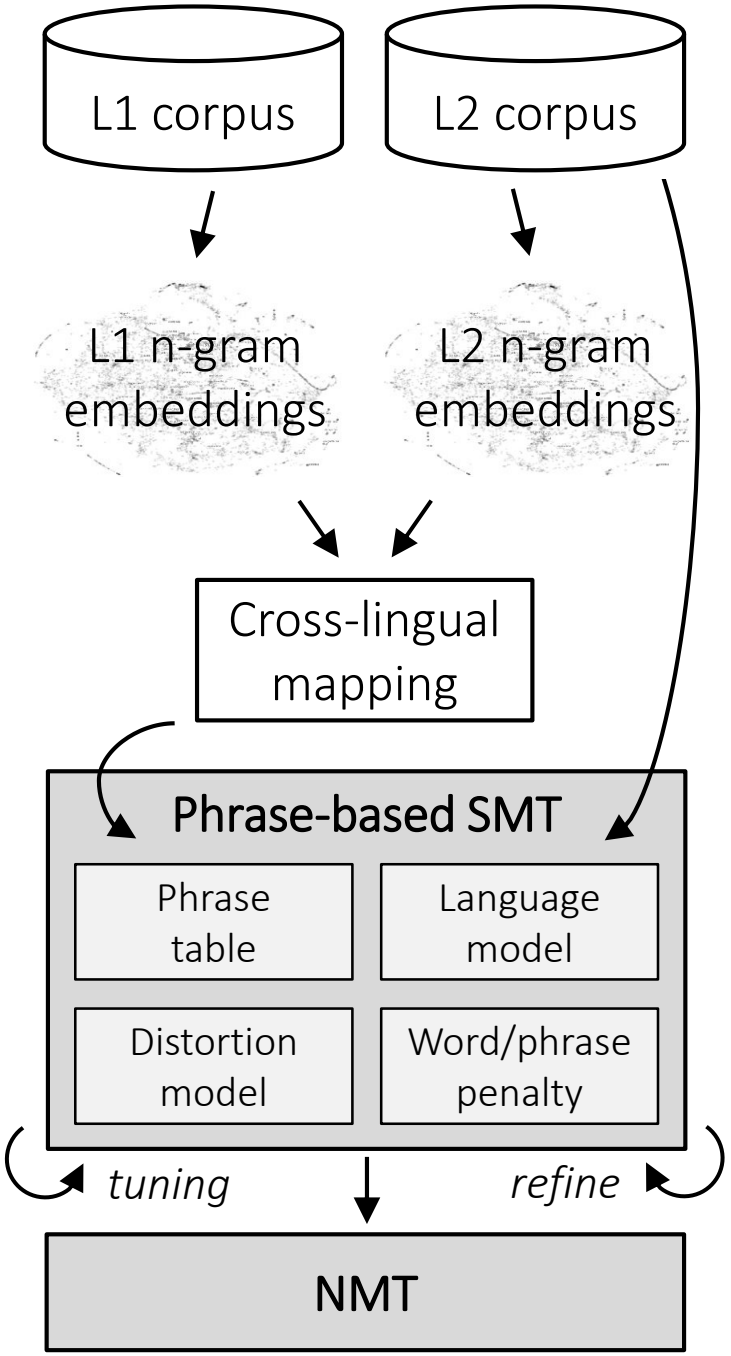


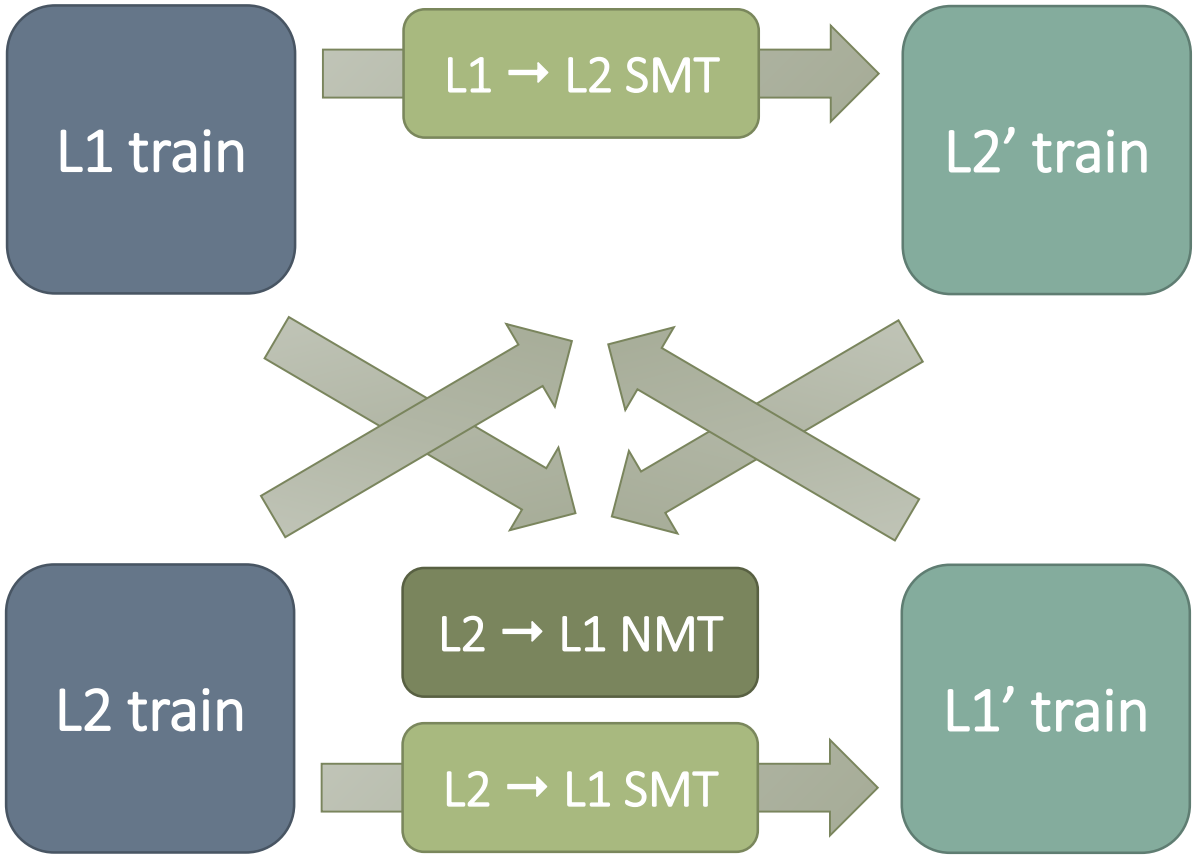
NMT hybridization



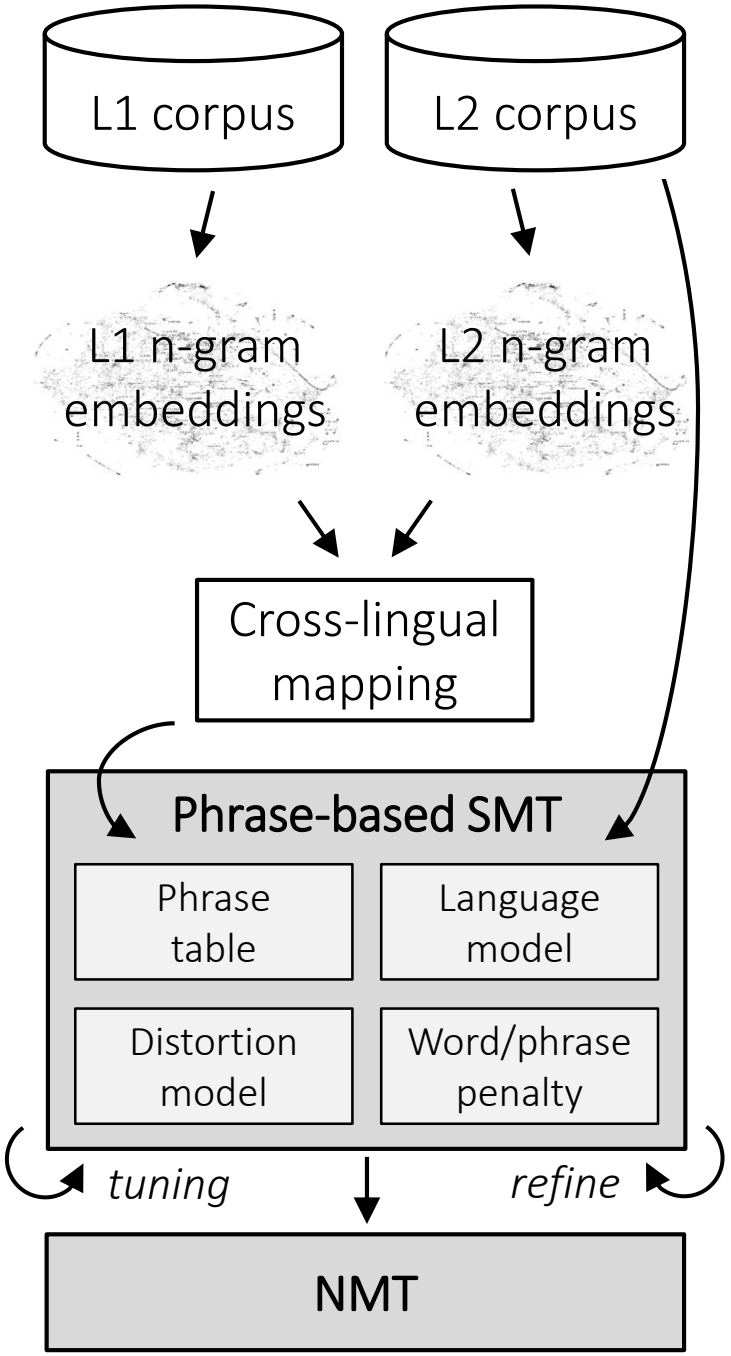


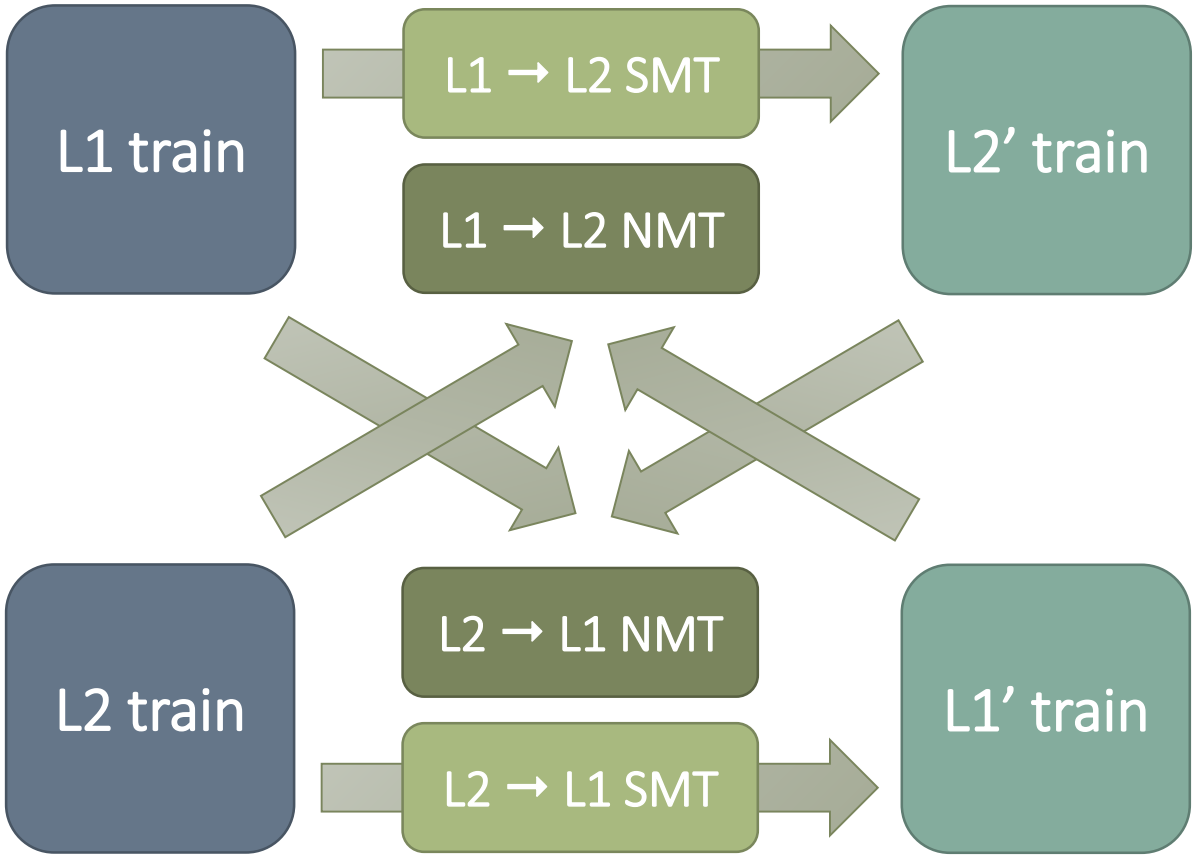
NMT hybridization



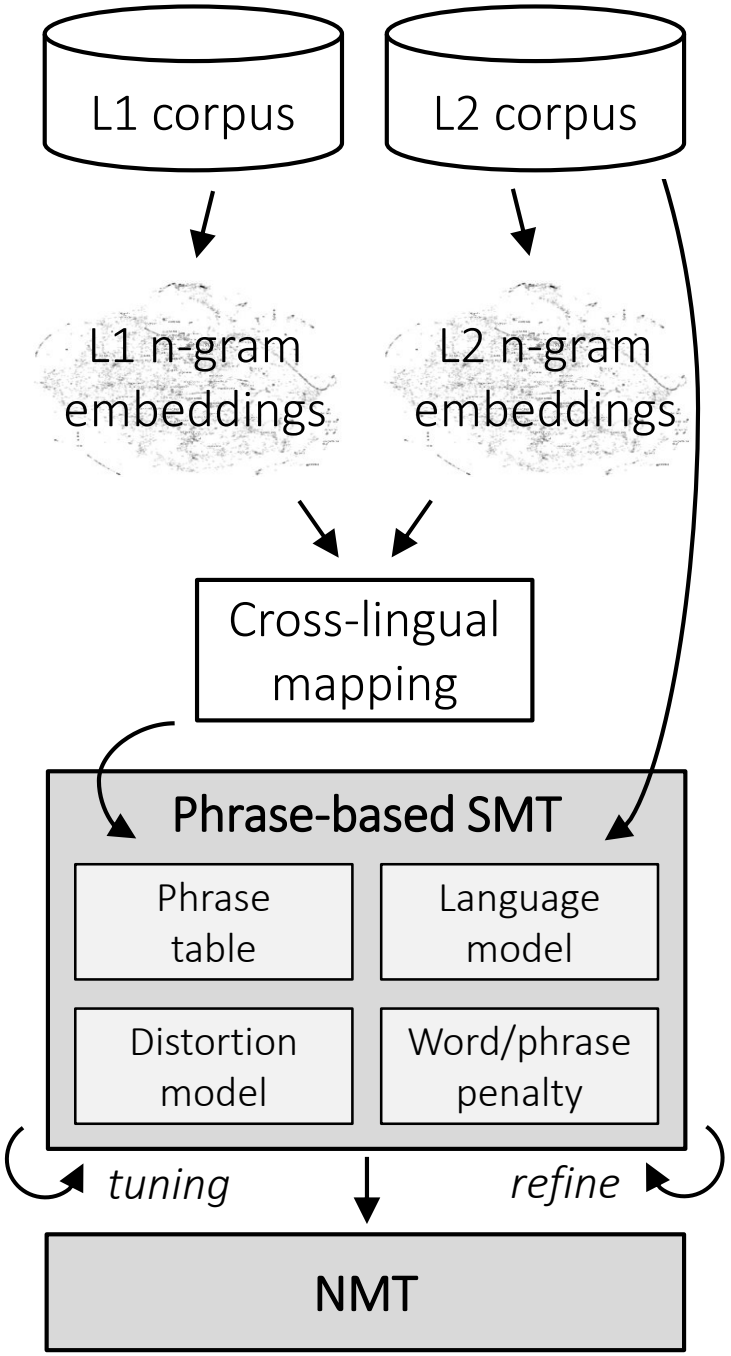


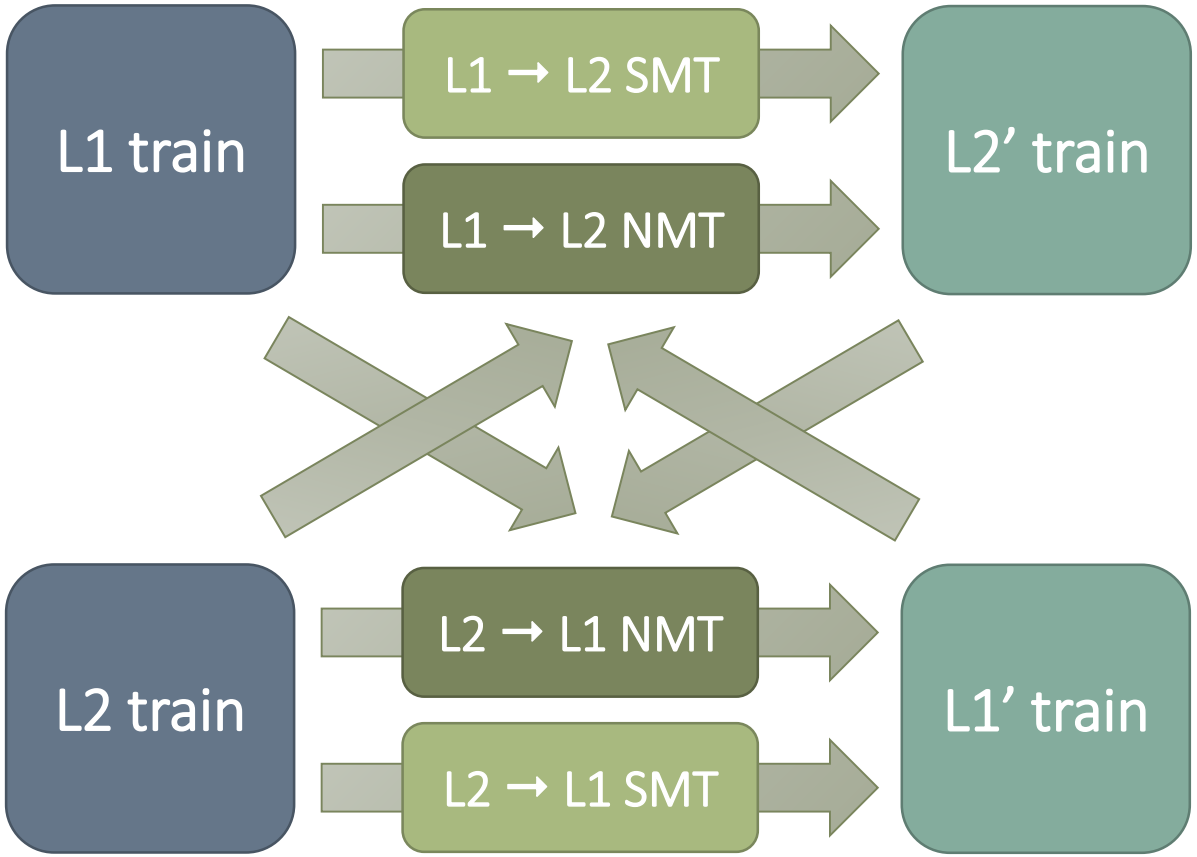
NMT hybridization



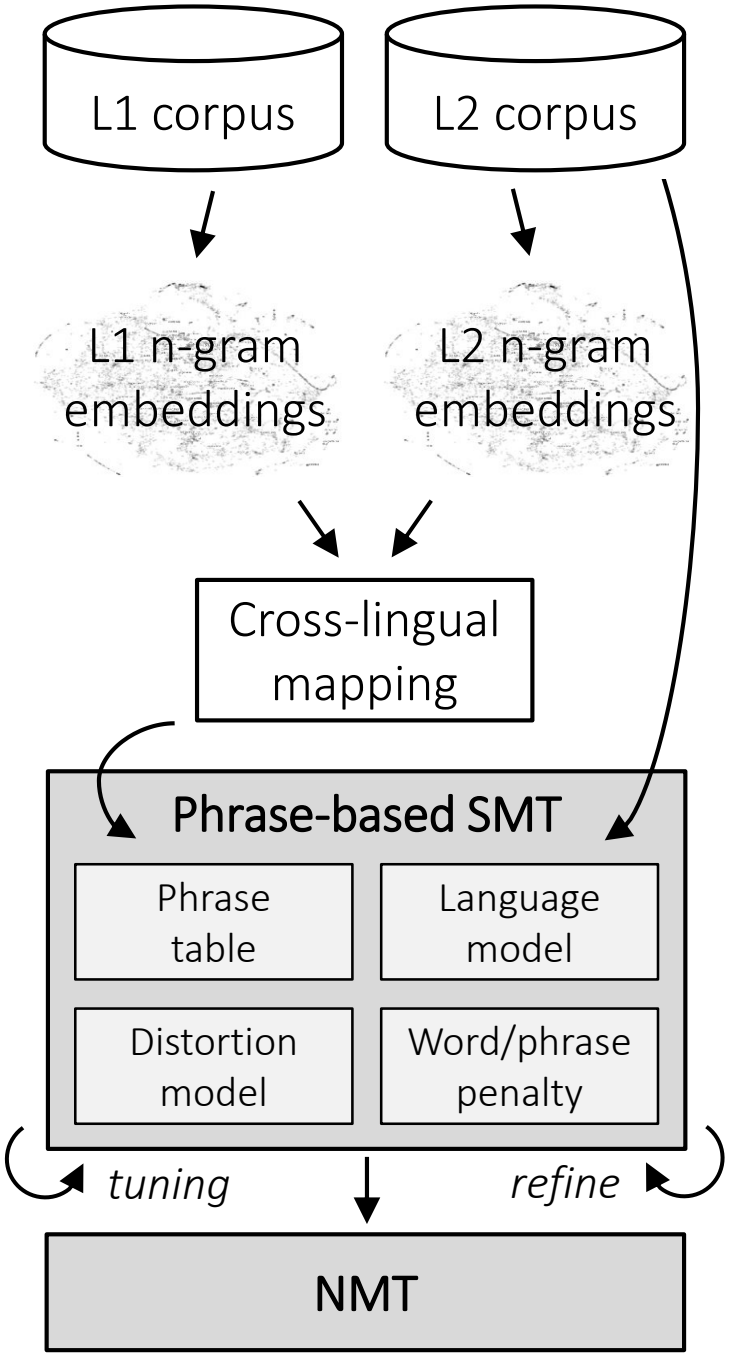


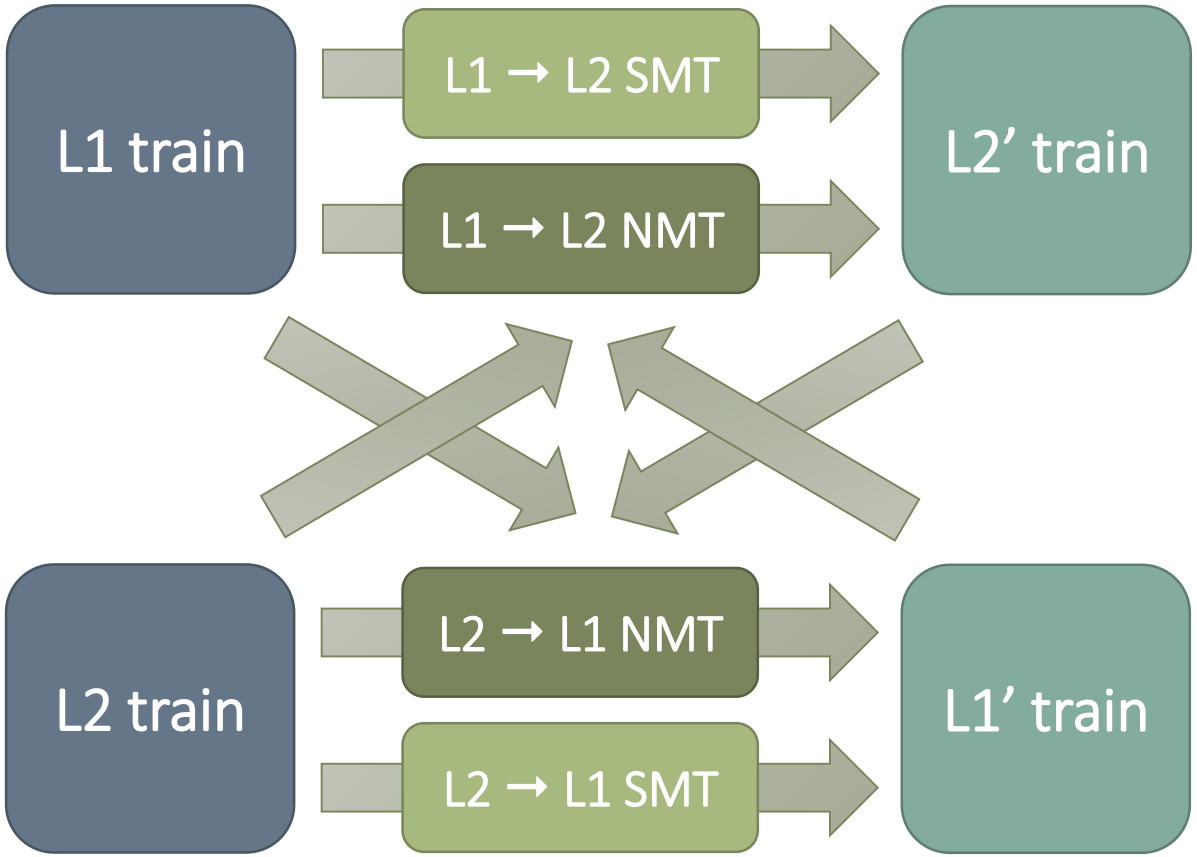
NMT hybridization





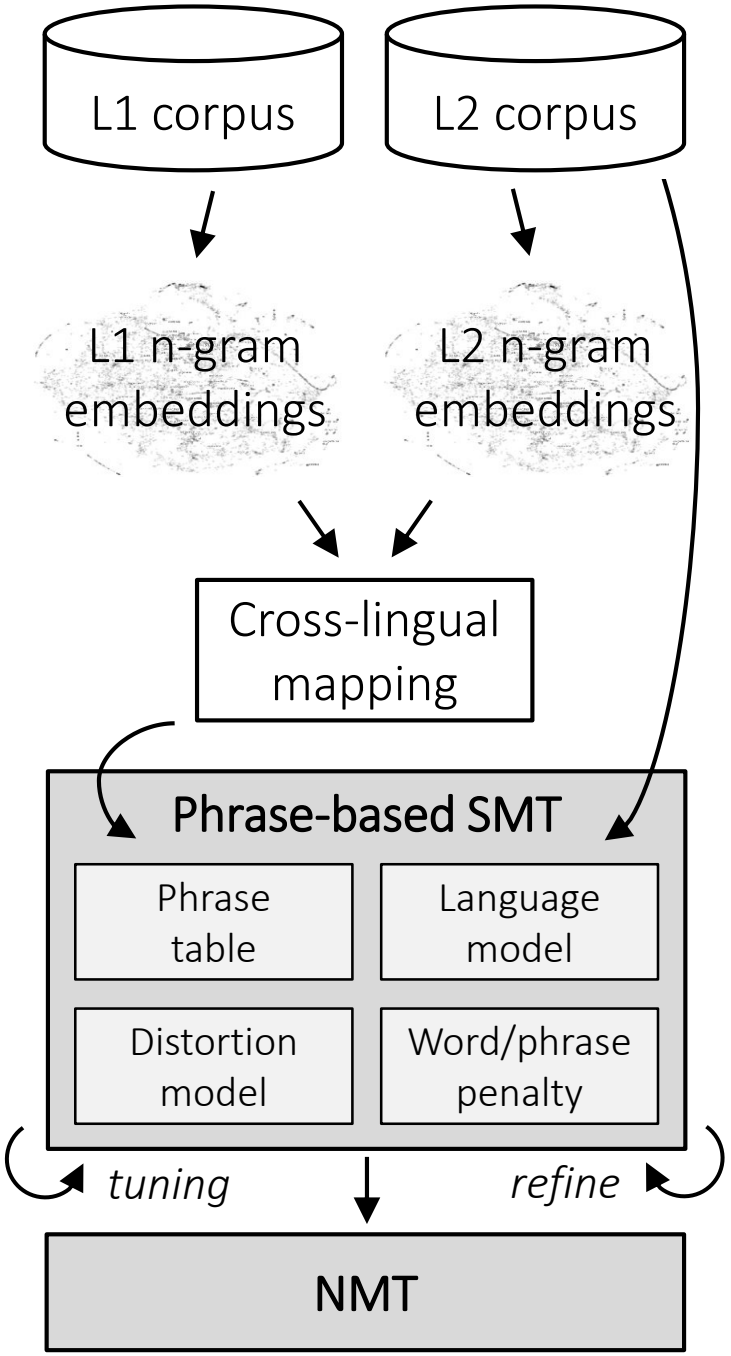
NMT hybridization

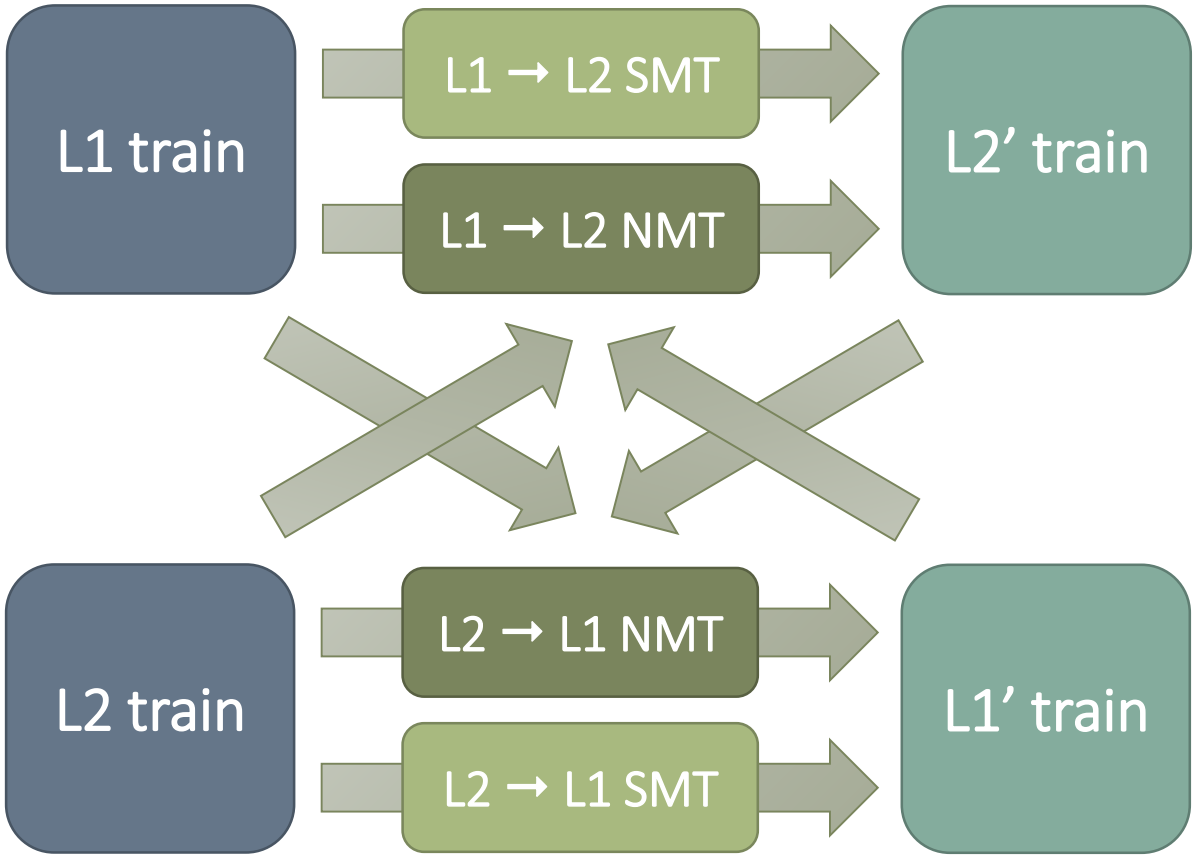




$$N_{SMT} = N \cdot \max(0, 1 - t/a)$$

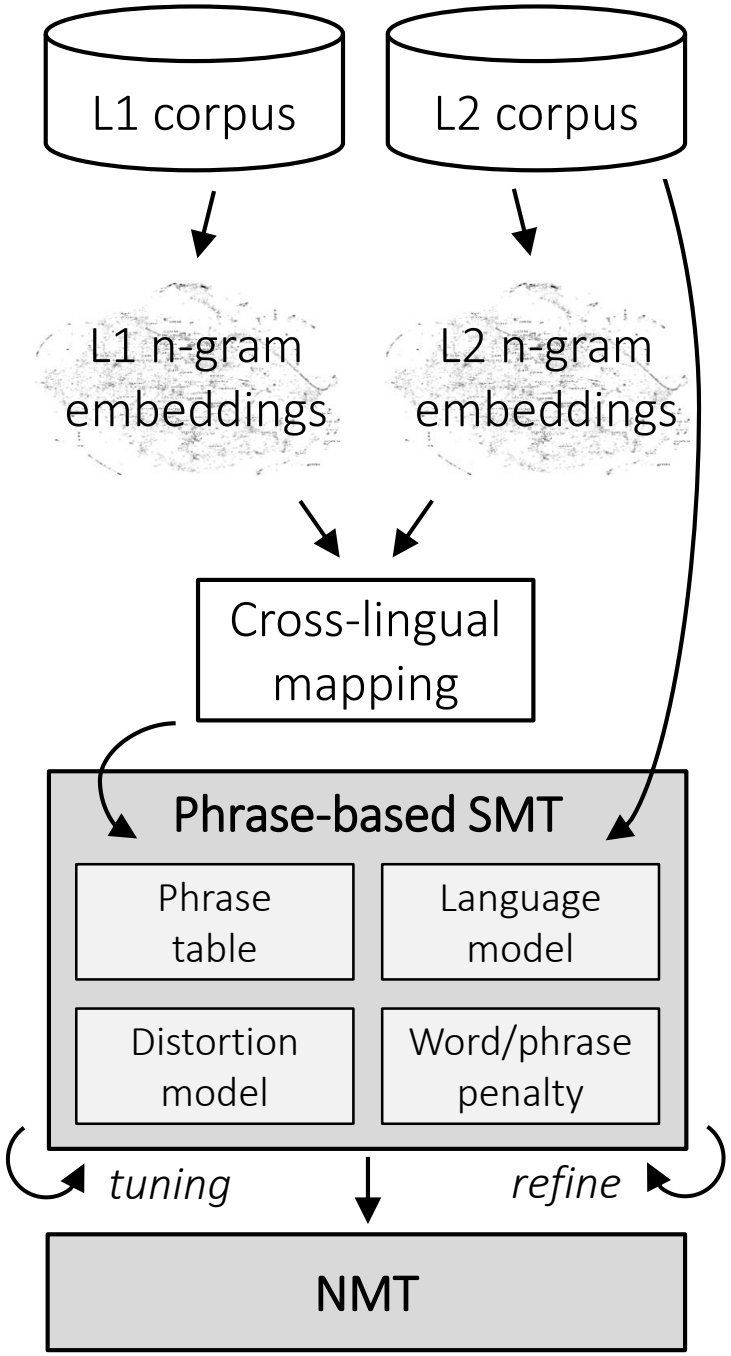
NMT hybridization





$$N_{SMT} = N \cdot \max(0, 1 - t/a)$$
$$N_{NMT} = N - N_{SMT}$$

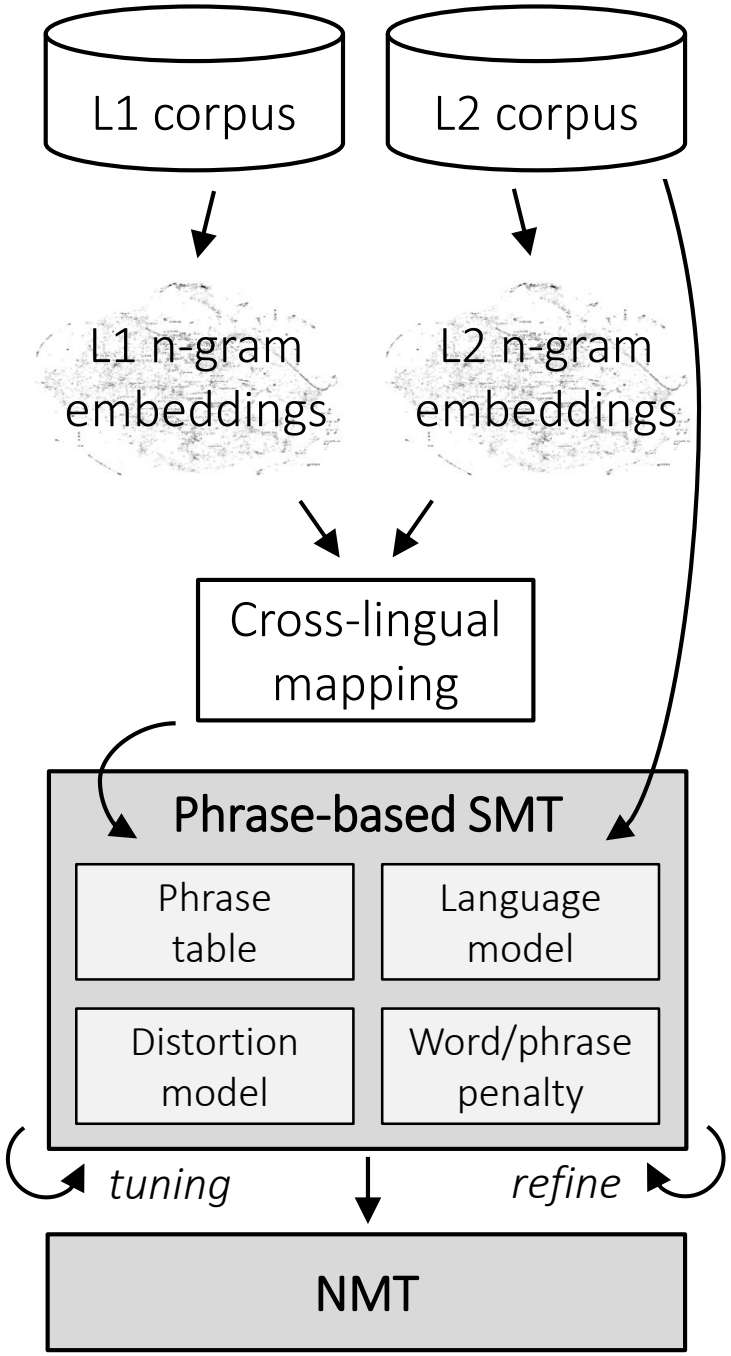
NMT hybridization



EXPERIMENTS

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Initial SMT	22.4	19.6	15.3	11.0
+ Tuning	23.4	21.9	15.4	11.2
+ Refinement	27.9	27.8	19.7	14.7



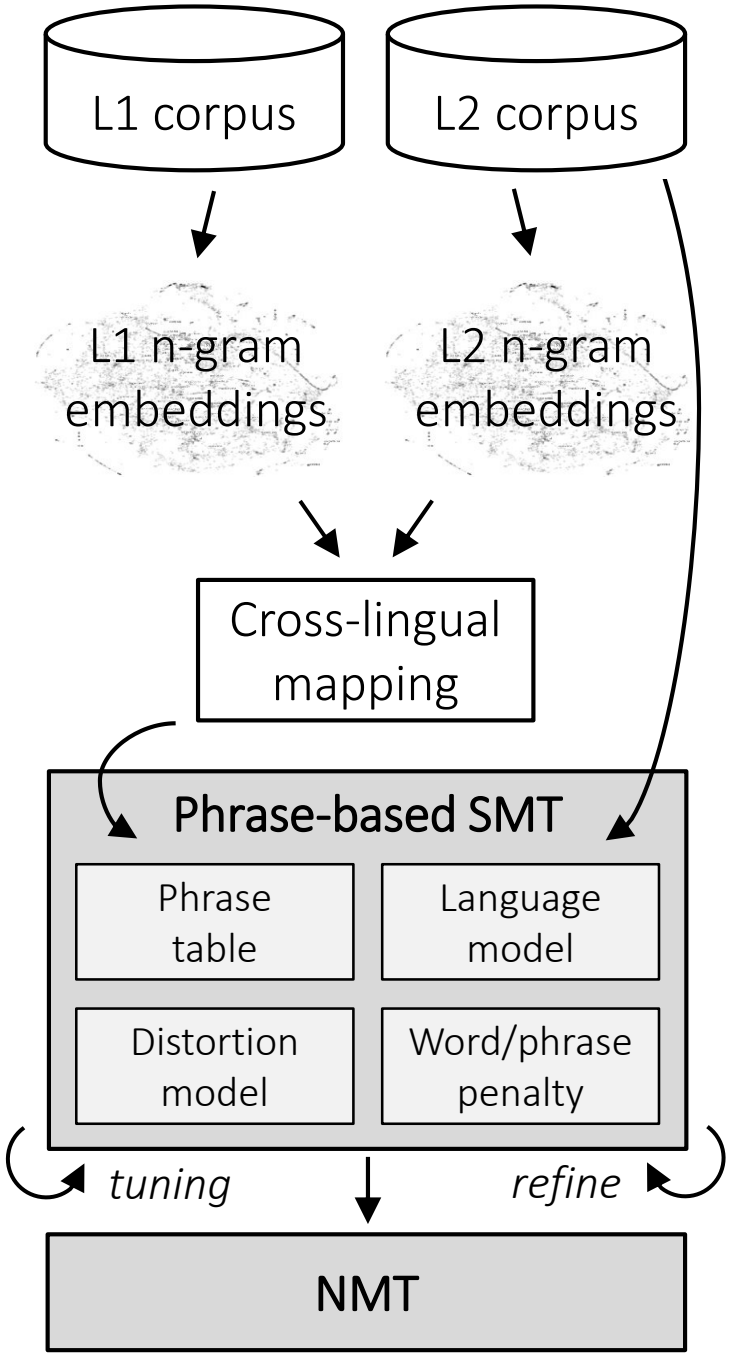
NMT hybridization

EXPERIMENTS

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Initial SMT	22.4	19.6	15.3	11.0
+ Tuning	23.4	21.9	15.4	11.2
+ Refinement	27.9	27.8	19.7	14.7
+ NMT hybrid	33.2	33.6	26.4	21.2

NMT hybridization

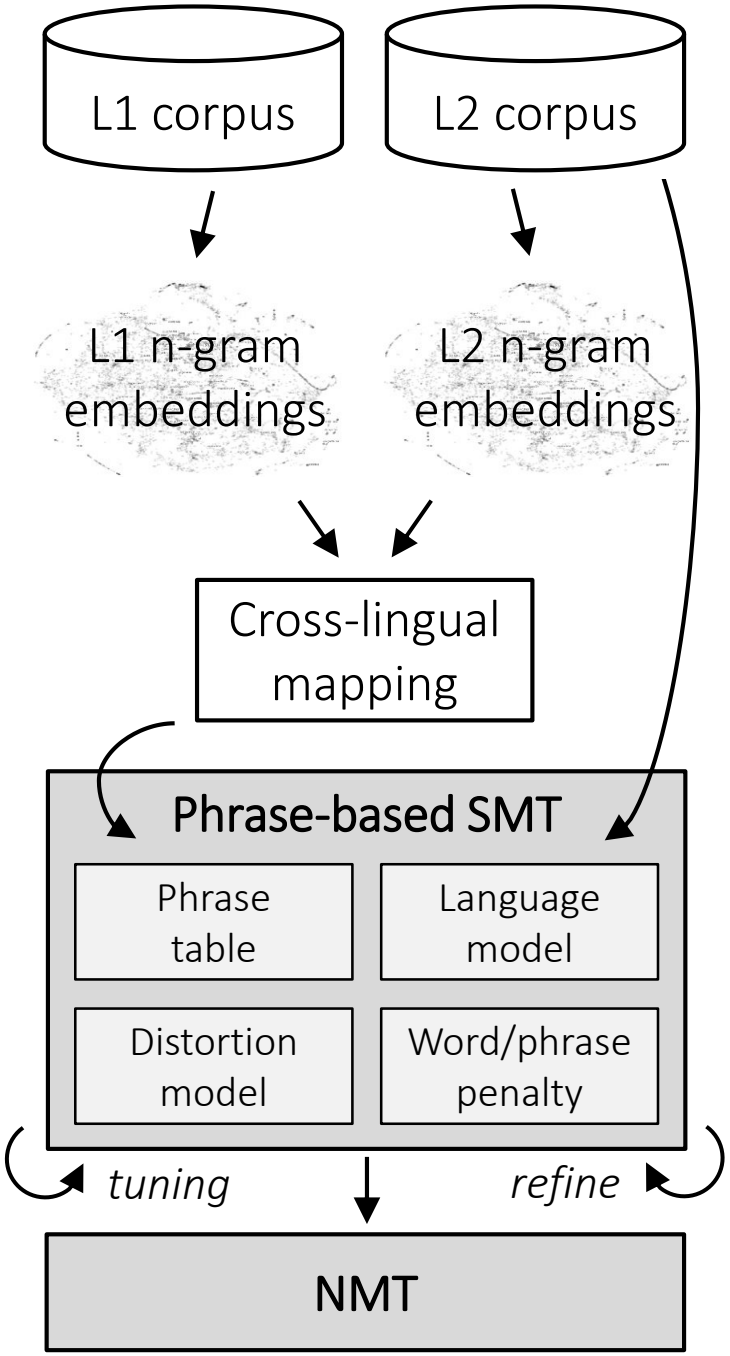


EXPERIMENTS

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Initial SMT	22.4	19.6	15.3	11.0
+ Tuning	23.4	21.9	15.4	11.2
+ Refinement	27.9	27.8	19.7	14.7
+ NMT hybrid	33.2	33.6	26.4	21.2

NMT hybridization

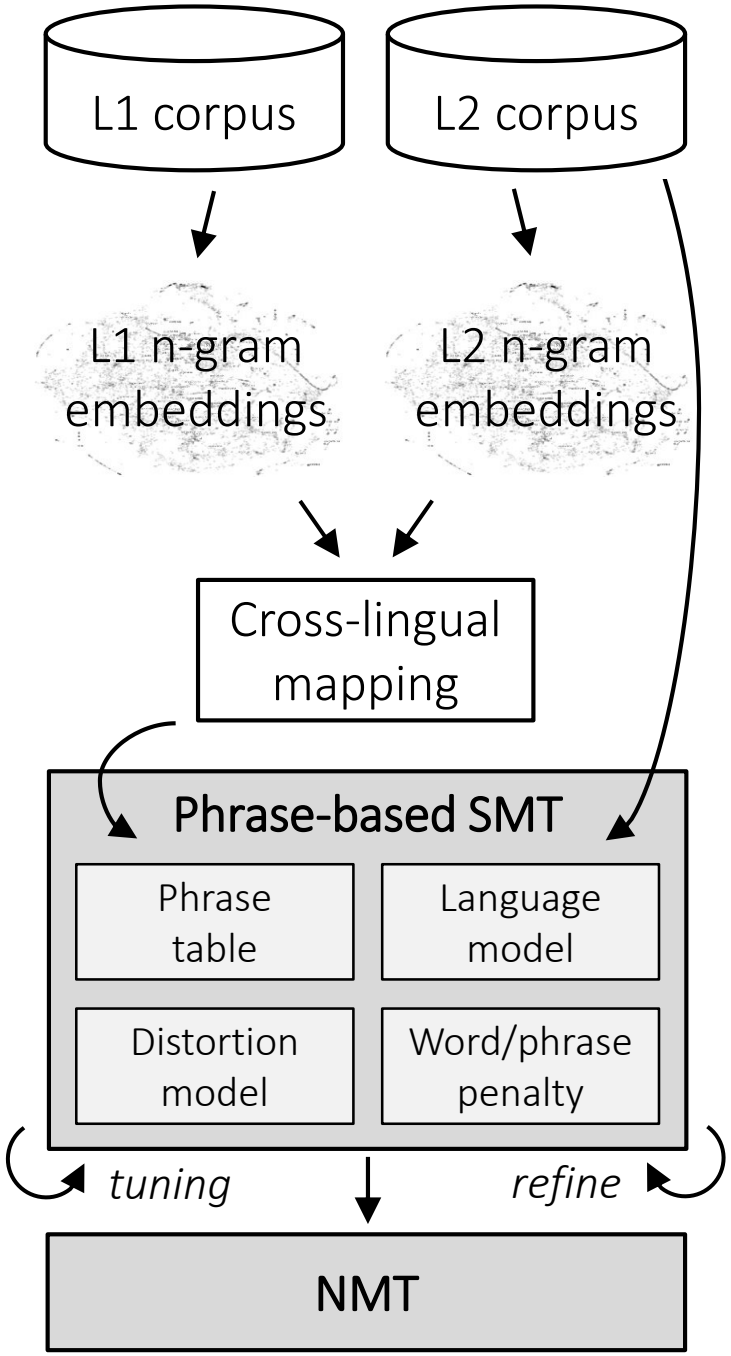


EXPERIMENTS

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsup.	Initial SMT	22.4	19.6	15.3	11.0
	+ Tuning	23.4	21.9	15.4	11.2
	+ Refinement	27.9	27.8	19.7	14.7
	+ NMT hybrid	33.2	33.6	26.4	21.2
Supervised					

NMT hybridization

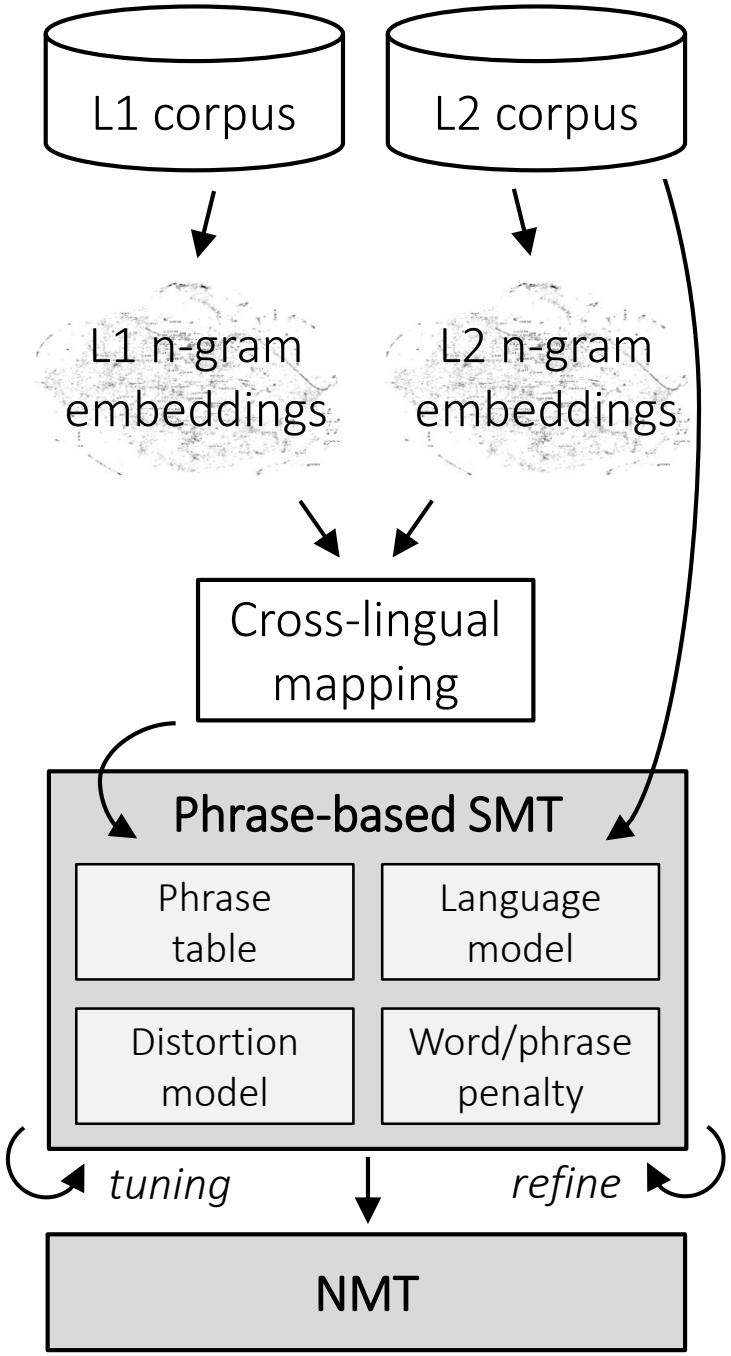


EXPERIMENTS

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsup.	Initial SMT	22.4	19.6	15.3	11.0
	+ Tuning	23.4	21.9	15.4	11.2
	+ Refinement	27.9	27.8	19.7	14.7
	+ NMT hybrid	33.2	33.6	26.4	21.2
Supervised	WMT winner	35.0	35.8	29.0	20.6

NMT hybridization

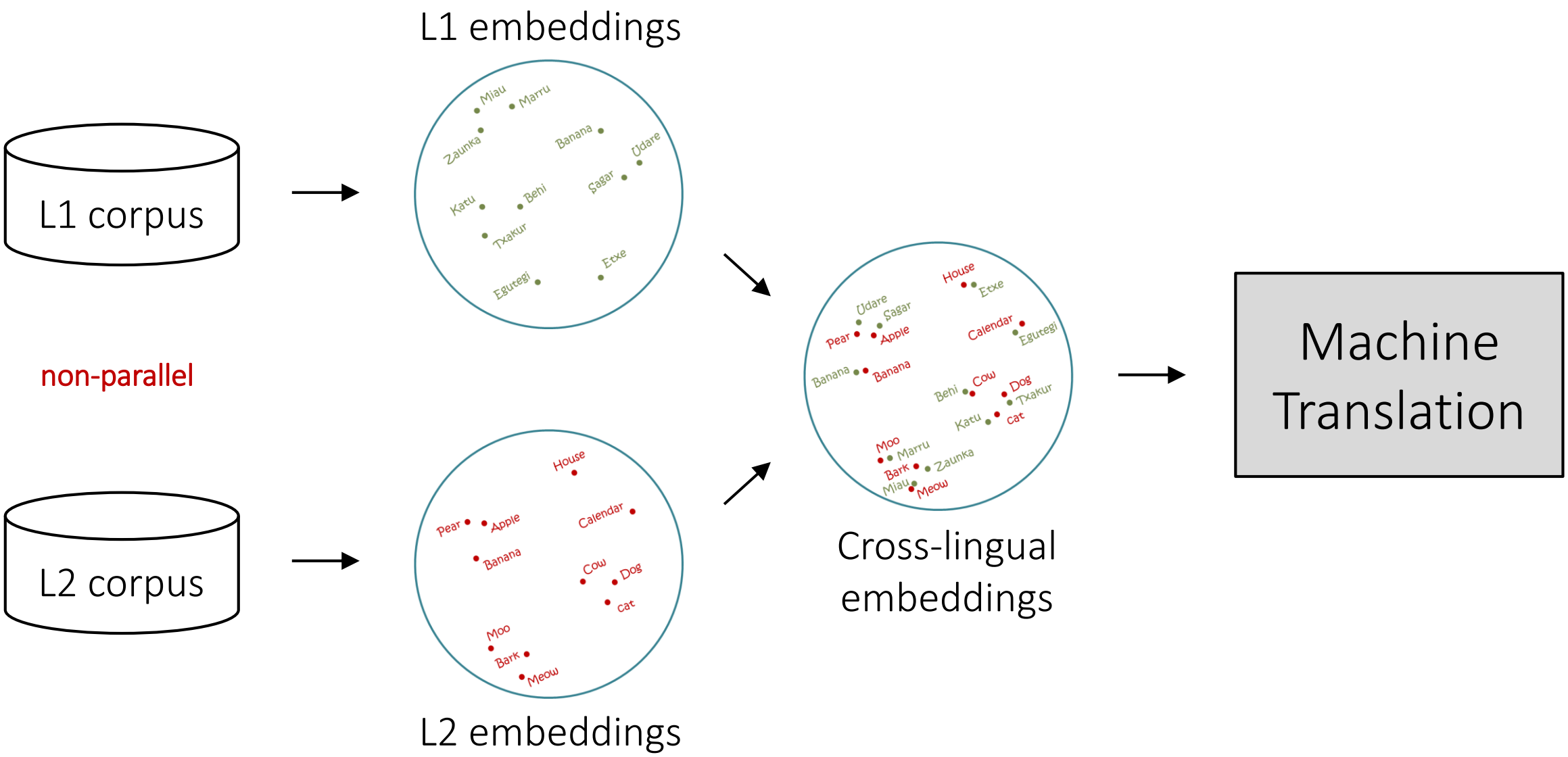


EXPERIMENTS

- Languages: French-English, German-English
- Training: WMT-14 News Crawl
- Test set: WMT-14 newstest (BLEU)

		FR-EN	EN-FR	DE-EN	EN-DE
Unsup.	Initial SMT	22.4	19.6	15.3	11.0
	+ Tuning	23.4	21.9	15.4	11.2
	+ Refinement	27.9	27.8	19.7	14.7
	+ NMT hybrid	33.2	33.6	26.4	21.2
Supervised	WMT winner	35.0	35.8	29.0	20.6
	Current SOTA	-	43.8	-	33.8

Outline


L1 embeddings
L1 corpus
non-parallel
L2 corpus
L2 embeddings
Cross-lingual embeddings
Machine Translation
Miau
Marru
Zaunka
Banana
Udare
Katu
Behi
Sagar
Txakur
Egutegi
Etxe
House
Pear
Apple
Calendar
Cow
Dog
cat
Moo
Bark
Meow

Publications

Publications

- Cross-lingual embedding mappings

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP’16)

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP’16)
    - Multi-step framework (AAAI’18)

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*
    - Self-learning (ACL'17)

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*
    - Self-learning (ACL'17)
    - Fully unsupervised method (ACL'18)

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*
    - Self-learning (ACL'17)
    - Fully unsupervised method (ACL'18)
  - *Calling for new paradigms*

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*
    - Self-learning (ACL'17)
    - Fully unsupervised method (ACL'18)
  - *Calling for new paradigms*
    - Joint learning vs offline mapping on parallel corpora (under review)

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*
    - Self-learning (ACL'17)
    - Fully unsupervised method (ACL'18)
  - *Calling for new paradigms*
    - Joint learning vs offline mapping on parallel corpora (under review)
    - Bilingual lexicon induction through unsupervised machine translation (under review)

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*
    - Self-learning (ACL'17)
    - Fully unsupervised method (ACL'18)
  - *Calling for new paradigms*
    - Joint learning vs offline mapping on parallel corpora (under review)
    - Bilingual lexicon induction through unsupervised machine translation (under review)
- Unsupervised machine translation

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*
    - Self-learning (ACL'17)
    - Fully unsupervised method (ACL'18)
  - *Calling for new paradigms*
    - Joint learning vs offline mapping on parallel corpora (under review)
    - Bilingual lexicon induction through unsupervised machine translation (under review)
- Unsupervised machine translation
    - NMT based (ICLR'18)

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*
    - Self-learning (ACL'17)
    - Fully unsupervised method (ACL'18)
  - *Calling for new paradigms*
    - Joint learning vs offline mapping on parallel corpora (under review)
    - Bilingual lexicon induction through unsupervised machine translation (under review)
- Unsupervised machine translation
    - NMT based (ICLR'18)
    - SMT based (EMNLP'18)

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*
    - Self-learning (ACL'17)
    - Fully unsupervised method (ACL'18)
  - *Calling for new paradigms*
    - Joint learning vs offline mapping on parallel corpora (under review)
    - Bilingual lexicon induction through unsupervised machine translation (under review)
- Unsupervised machine translation
  - NMT based (ICLR'18)
  - SMT based (EMNLP'18)
  - Improve SMT based + NMT (arXiv'19, under review)

Publications

- Cross-lingual embedding mappings
  - *Generalize & Improve*
    - Normalization & Orthogonality (EMNLP'16)
    - Multi-step framework (AAAI'18)
  - *Reduce supervision*
    - Self-learning (ACL'17)
    - Fully unsupervised method (ACL'18)
  - *Calling for new paradigms*
    - Joint learning vs offline mapping on parallel corpora (under review)
    - Bilingual lexicon induction through unsupervised machine translation (under review)
- Unsupervised machine translation
  - NMT based (ICLR'18)
  - SMT based (EMNLP'18)
  - Improve SMT based + NMT (arXiv'19, under review)
  - Understanding iterative back-translation (under review)

Publications

- Tailoring word embeddings (CoNLL'18)

Publications

- Tailoring word embeddings (CoNLL'18)
- Multilingual sentence embeddings (Facebook internship)

Publications

- Tailoring word embeddings (CoNLL'18)
- Multilingual sentence embeddings (Facebook internship)
  - Parallel corpus mining (arXiv'18, under review)

Publications

- Tailoring word embeddings (CoNLL'18)
- Multilingual sentence embeddings (Facebook internship)
  - Parallel corpus mining (arXiv'18, under review)
  - Cross-lingual transfer learning (arXiv'18, under review)