Unsupervised Machine Translation

Mikel Artetxe

Facebook AI Research

Introduction

Introduction

WMT 2019 English-German

Introduction

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

Introduction

super-human performance

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

Introduction

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

super-human performance

human evaluators preferred MT to human translation!

Introduction

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

NMT trained on 27.7M parallel sentence pairs

super-human performance

human evaluators preferred MT to human translation!

Introduction

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

super-human performance

human evaluators preferred MT to human translation!

NMT trained on 27.7M parallel sentence pairs

In my opinion, this exposes a serious inconsistency in industrial…
En mi opinión, esto representa una grave inconsistencia entre…

Introduction

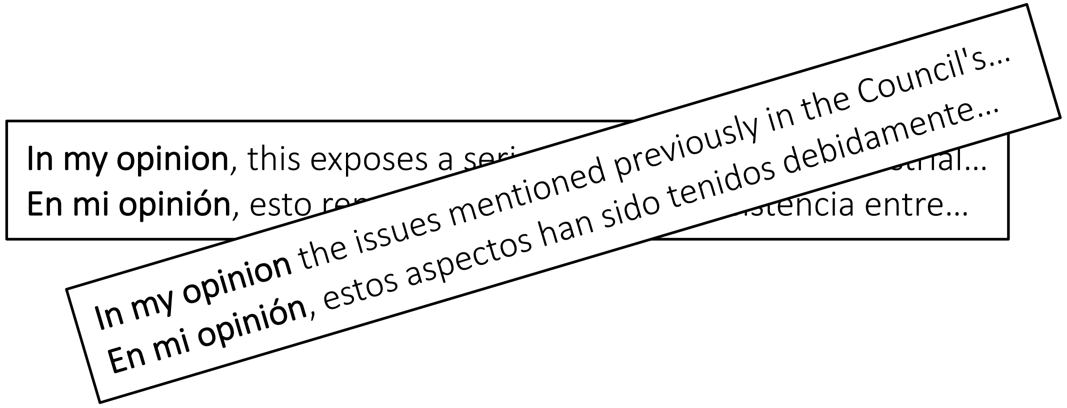
super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

NMT trained on 27.7M parallel sentence pairs



Introduction

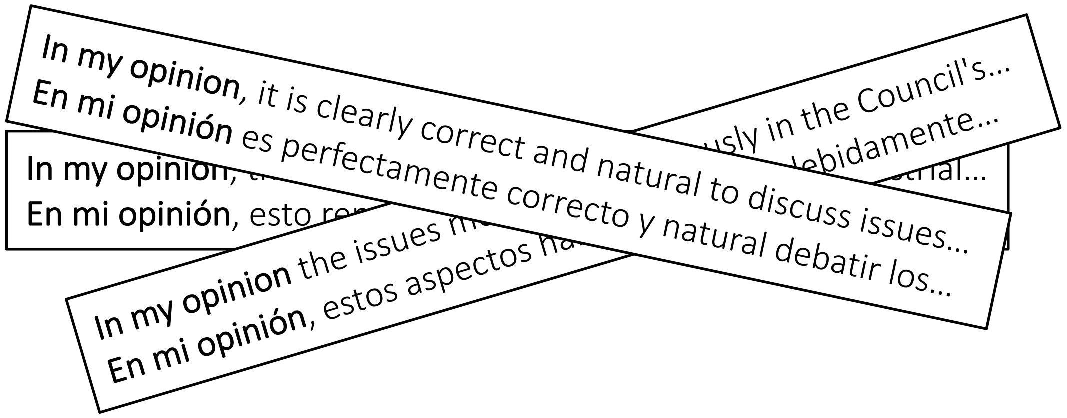
super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

← NMT trained on 27.7M parallel sentence pairs



Introduction

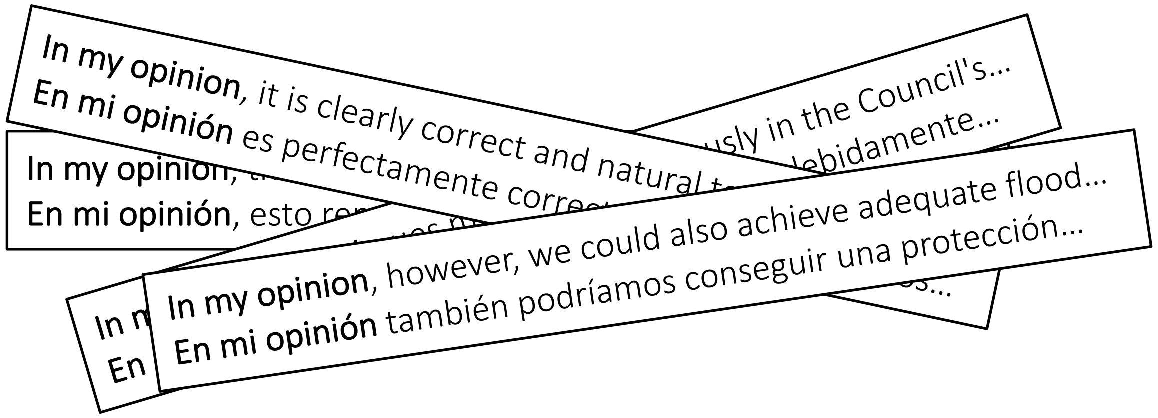
super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

NMT trained on 27.7M parallel sentence pairs



Introduction

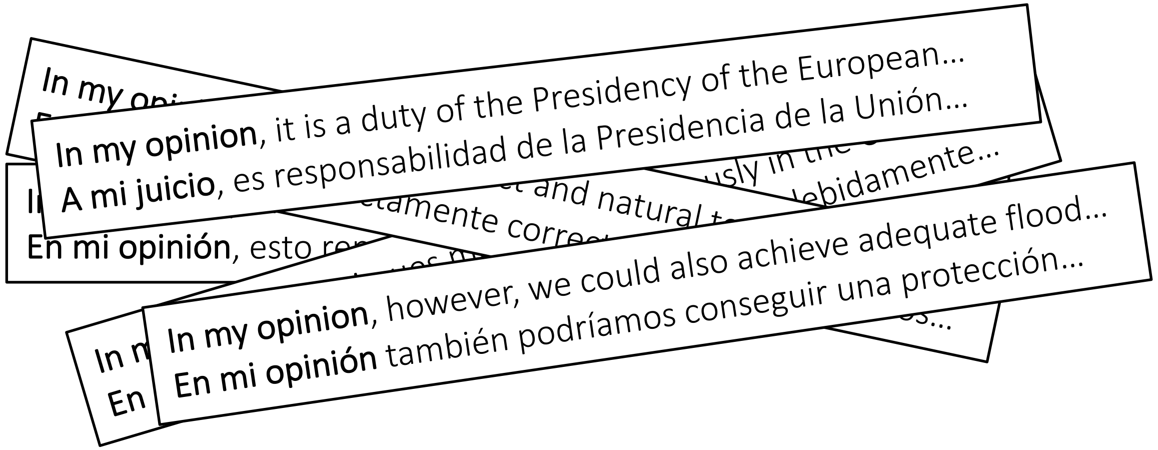
WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

super-human performance

human evaluators preferred MT to human translation!

NMT trained on 27.7M parallel sentence pairs



Introduction

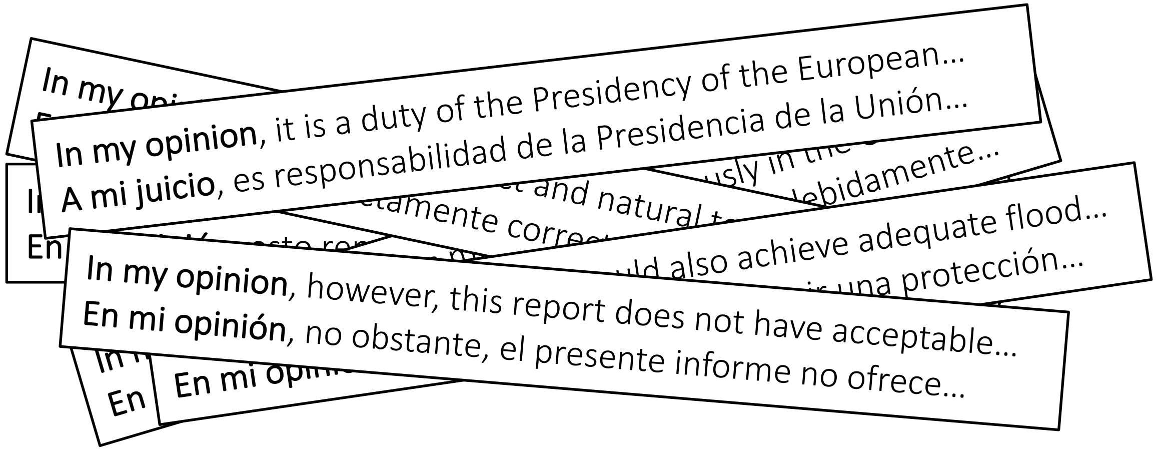
super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

← NMT trained on 27.7M parallel sentence pairs



Introduction

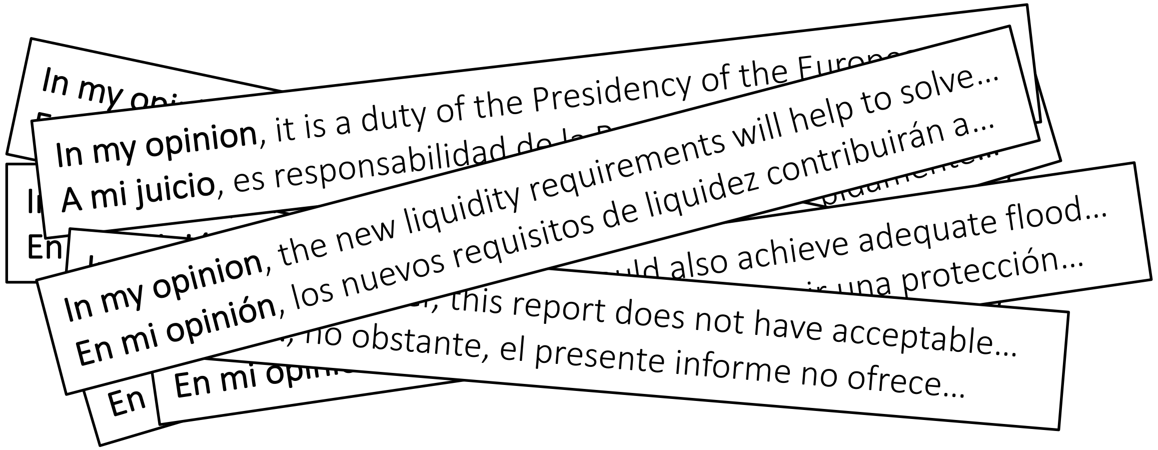
WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

super-human performance

human evaluators preferred MT to human translation!

NMT trained on 27.7M parallel sentence pairs



Introduction

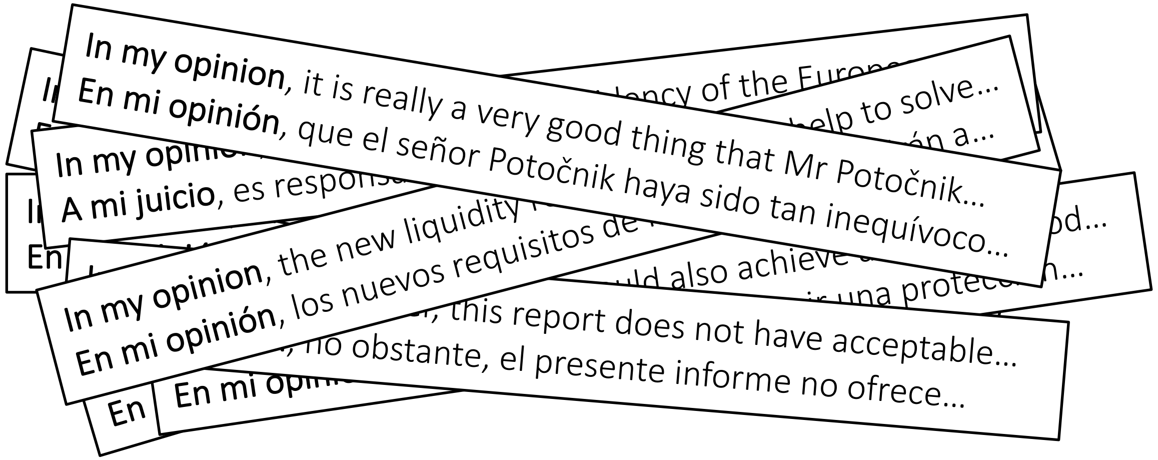
super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

← NMT trained on 27.7M parallel sentence pairs



Introduction

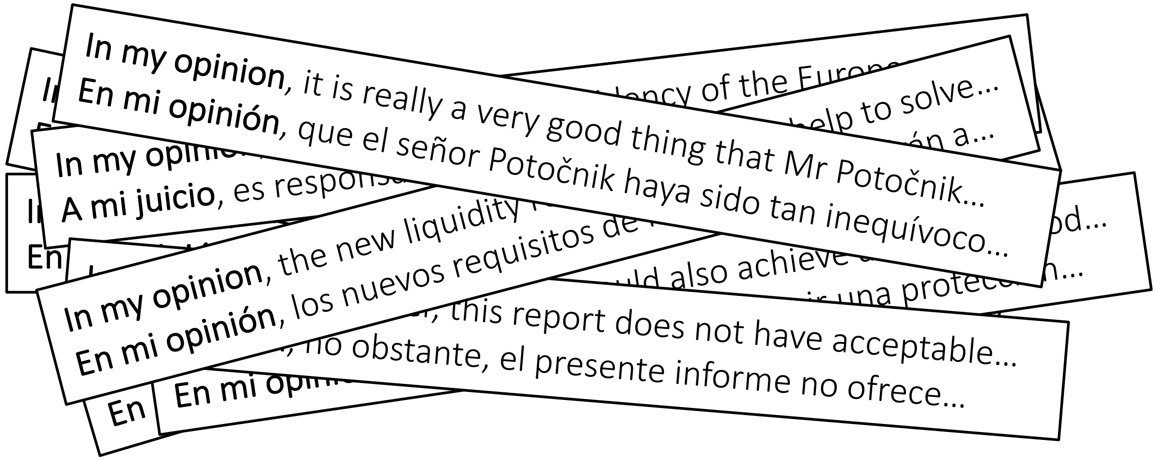
super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

← NMT trained on 27.7M parallel sentence pairs



In my opinion, this social dimension should, in fact, be brought into...

Introduction

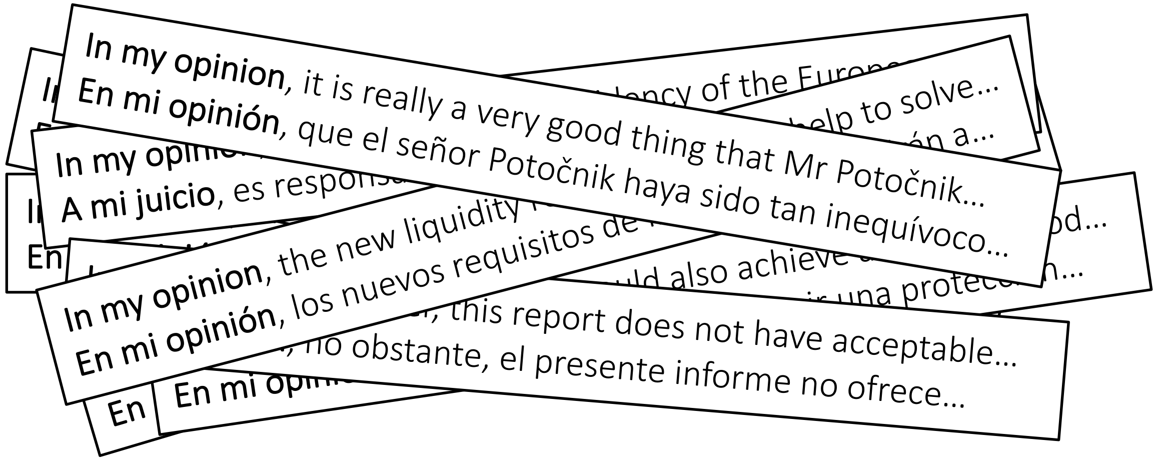
super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

NMT trained on 27.7M parallel sentence pairs



In my opinion, this social dimension should, in fact, be brought into...

???

Introduction

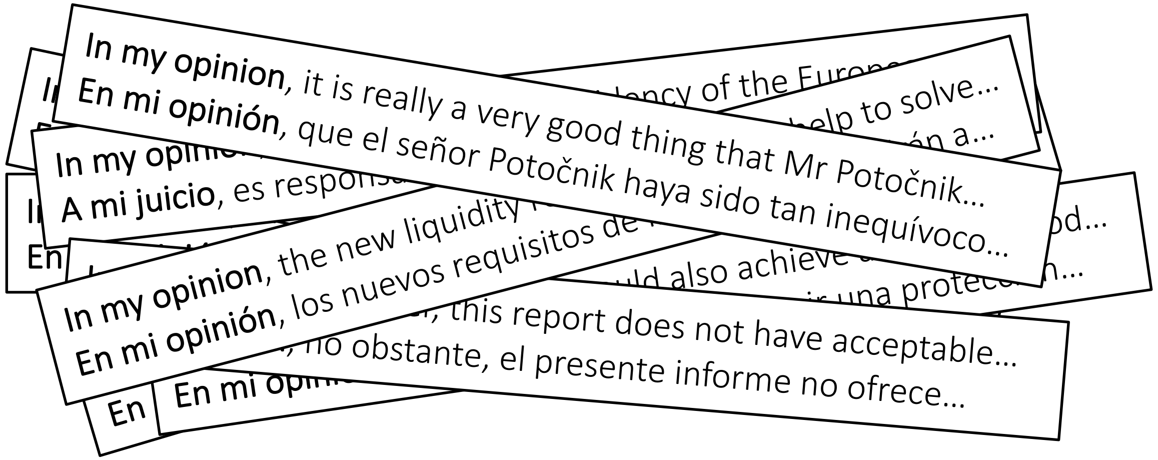
super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

← NMT trained on 27.7M parallel sentence pairs



In my opinion, this social dimension should, in fact, be brought into...

En mi opinion, ...

Introduction

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

← NMT trained on 27.7M parallel sentence pairs

super-human performance

human evaluators preferred MT to human translation!

Introduction

super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

NMT trained on 27.7M parallel sentence pairs

Introduction

super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

NMT trained on $\underbrace{\text{27.7M parallel sentence pairs}}_{\text{We would need 48 years to read that!}}$

Introduction

super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

NMT trained on $\underbrace{\text{27.7M parallel sentence pairs}}$

We would need 48 years to read that!

- Not sample efficient

Introduction

super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

NMT trained on $\underbrace{\text{27.7M parallel sentence pairs}}_{\text{We would need 48 years to read that!}}$

- Not sample efficient
- Not available for most language pairs

Introduction

super-human performance

human evaluators preferred MT to human translation!

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

NMT trained on $\underbrace{\text{27.7M parallel sentence pairs}}$

We would need 48 years to read that!

- Not sample efficient
- Not available for most language pairs

Introduction

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

super-human performance

human evaluators preferred MT to human translation!

NMT trained on $\underbrace{\text{27.7M parallel sentence pairs}}_{}$

We would need 48 years to read that!

- Not sample efficient
- Not available for most language pairs

Introduction

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

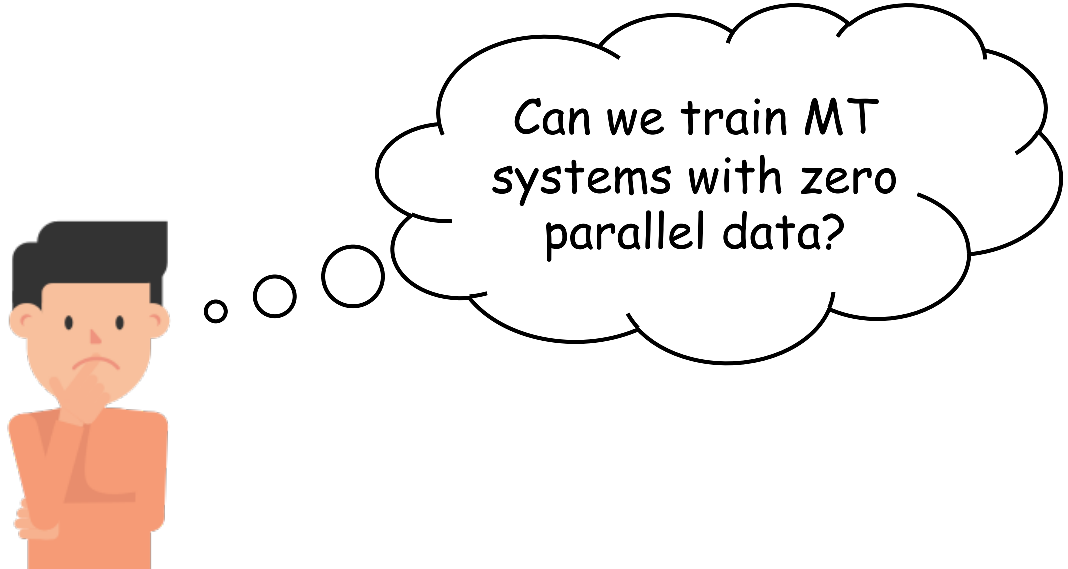
super-human performance

human evaluators preferred MT to human translation!

NMT trained on $\underbrace{\text{27.7M parallel sentence pairs}}$

We would need 48 years to read that!

- Not sample efficient
- Not available for most language pairs



Introduction

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

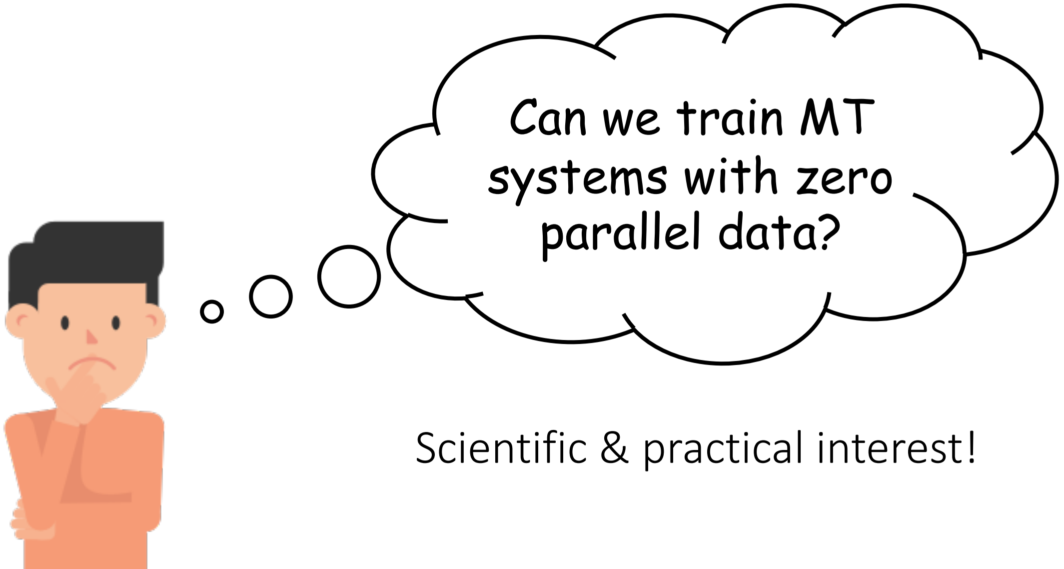
super-human performance

human evaluators preferred MT to human translation!

NMT trained on $\underbrace{\text{27.7M parallel sentence pairs}}$

We would need 48 years to read that!

- Not sample efficient
- Not available for most language pairs



Scientific & practical interest!

Introduction

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

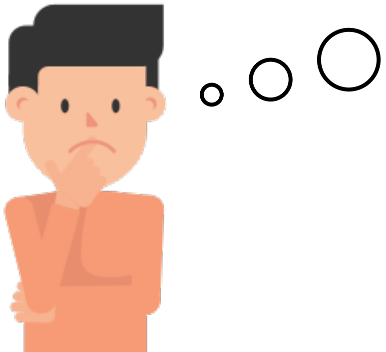
super-human performance

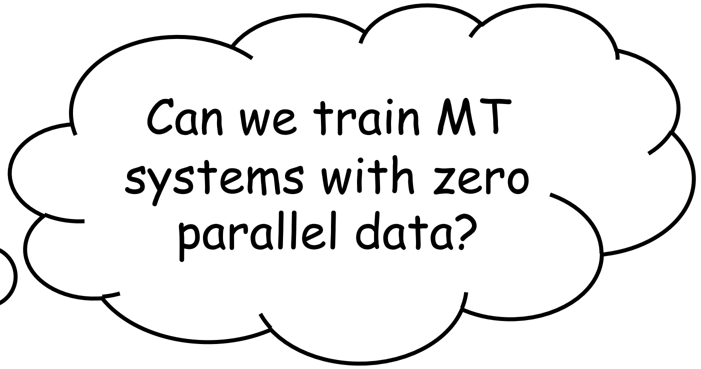
human evaluators preferred MT to human translation!

NMT trained on $\underbrace{\text{27.7M parallel sentence pairs}}$

We would need 48 years to read that!

- Not sample efficient
- Not available for most language pairs





Scientific & practical interest!

Previously explored in statistical decipherment

(Knight et al., ACL'06; Ravi & Knight, ACL'11; Dou et al., ACL'15; *inter alia*)

Introduction

WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

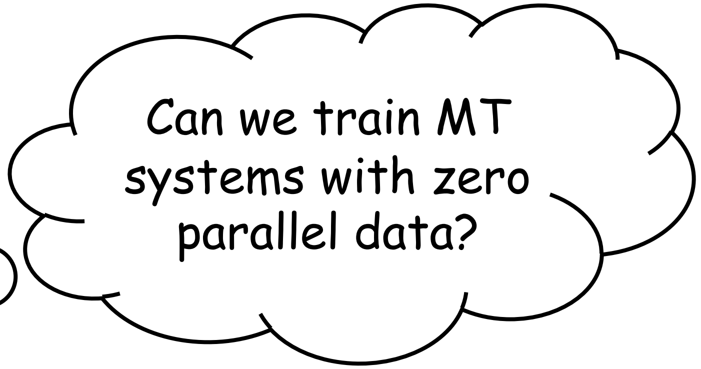
super-human performance

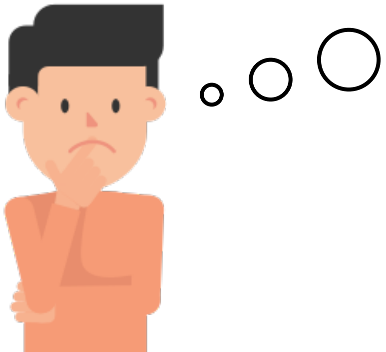
human evaluators preferred MT to human translation!

NMT trained on $\underbrace{\text{27.7M parallel sentence pairs}}$

We would need 48 years to read that!

- Not sample efficient
- Not available for most language pairs





Scientific & practical interest!

Previously explored in statistical decipherment

(Knight et al., ACL'06; Ravi & Knight, ACL'11; Dou et al., ACL'15; *inter alia*)

…but strong limitations…

Introduction

super-human performance

human evaluators preferred MT to human translation!

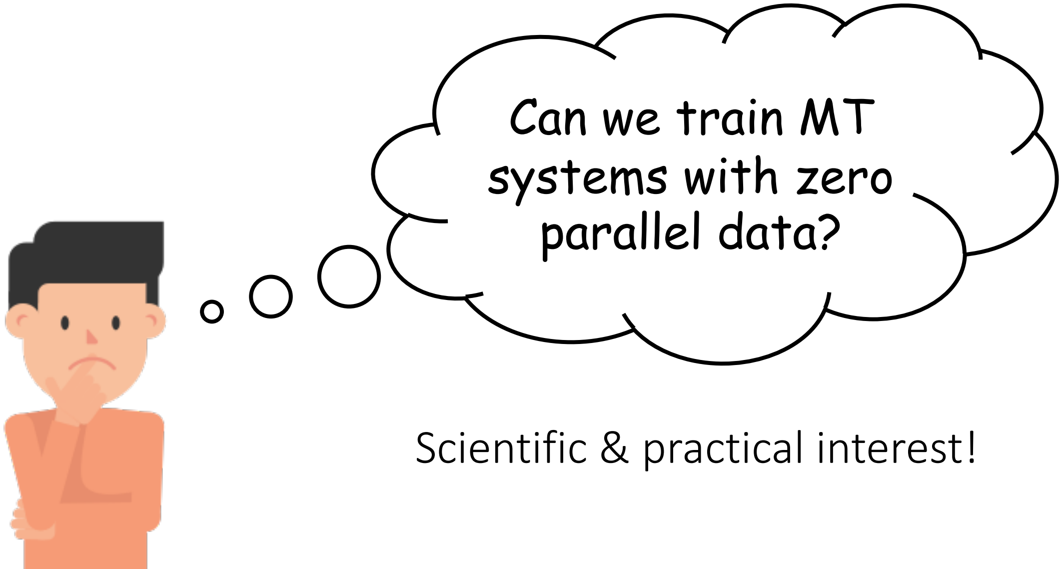
WMT 2019 English-German

System	Score
Facebook FAIR	0.347
Microsoft sent-doc	0.311
Microsoft doc-level	0.296
HUMAN	0.240
MSRA-MADL	0.214
UCAM	0.213
⋮	⋮

NMT trained on $\underbrace{\text{27.7M parallel sentence pairs}}$

We would need 48 years to read that!

- Not sample efficient
- Not available for most language pairs



Scientific & practical interest!

Previously explored in statistical decipherment

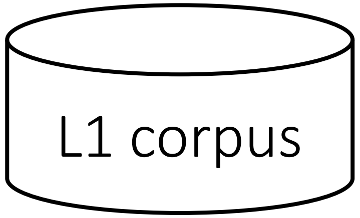
(Knight et al., ACL'06; Ravi & Knight, ACL'11; Dou et al., ACL'15; *inter alia*)

...but strong limitations...

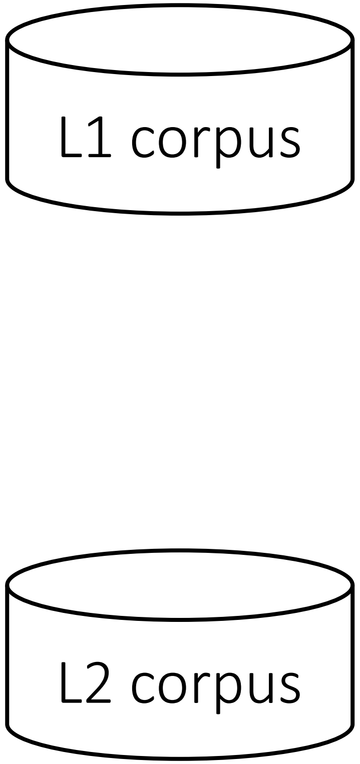
IMPRESSIVE PROGRESS IN THE LAST 2-3 YEARS!

Outline

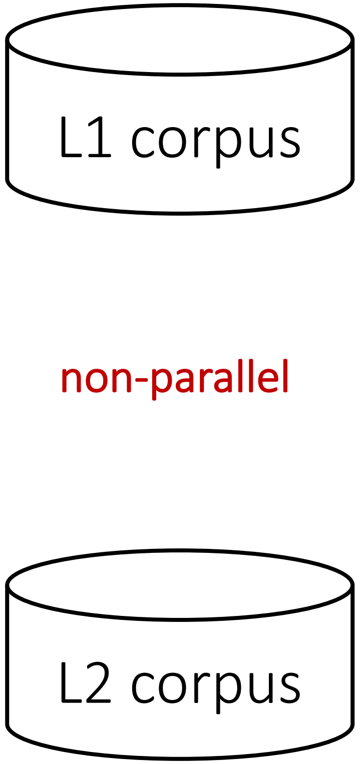
Outline



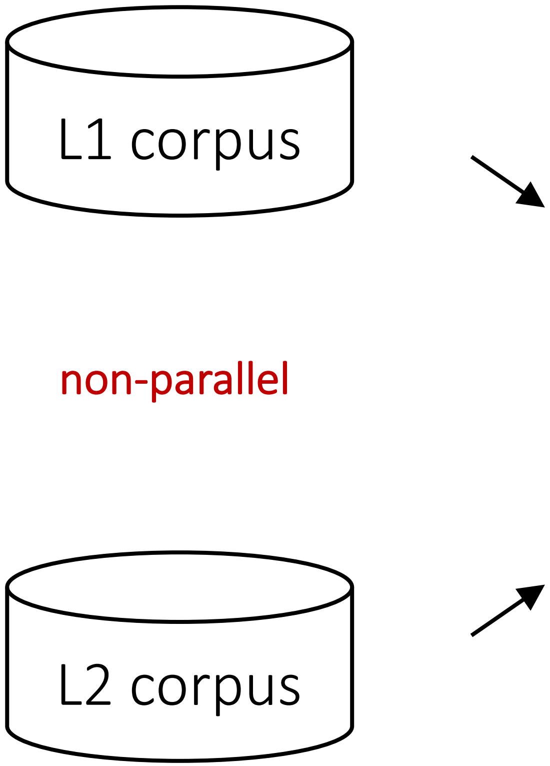
Outline



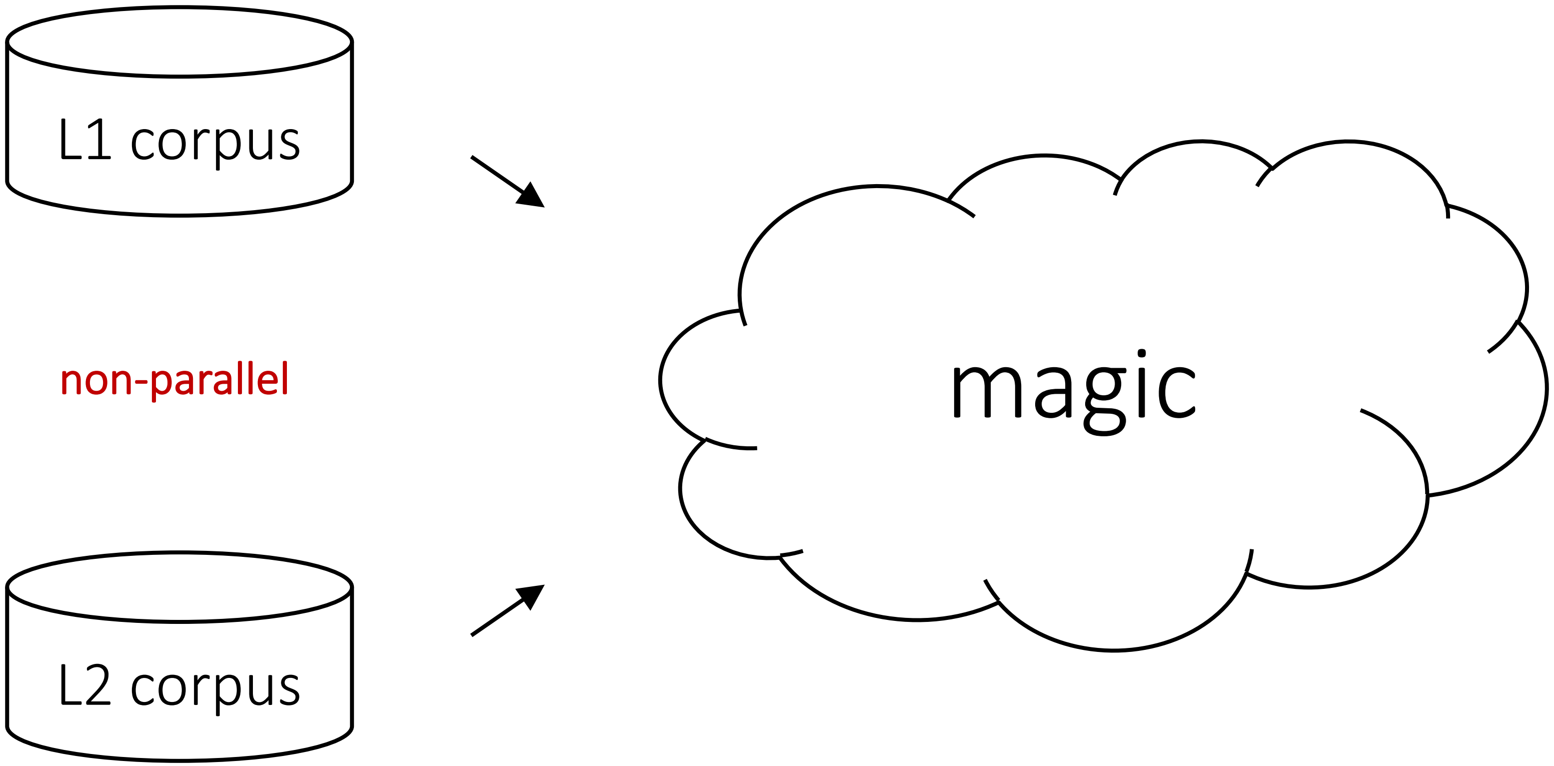
Outline



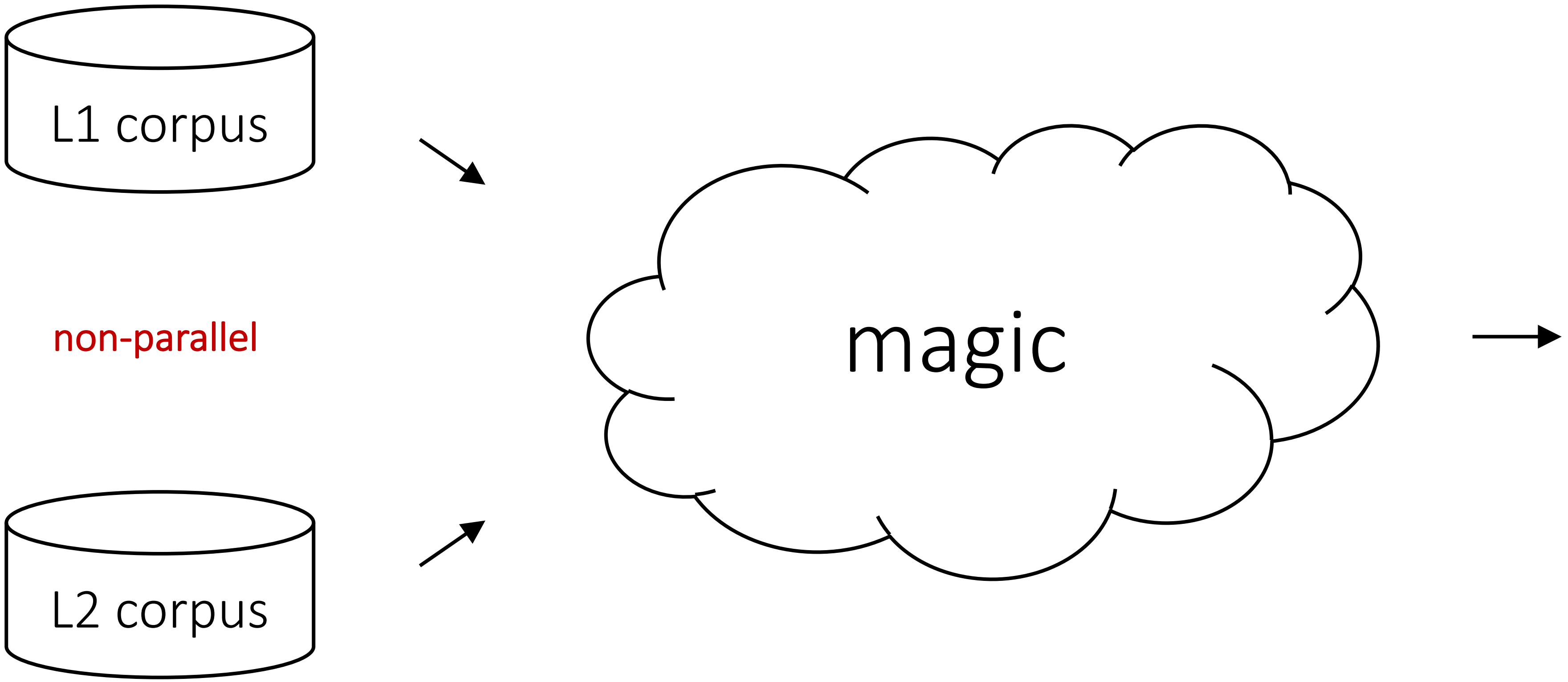
Outline



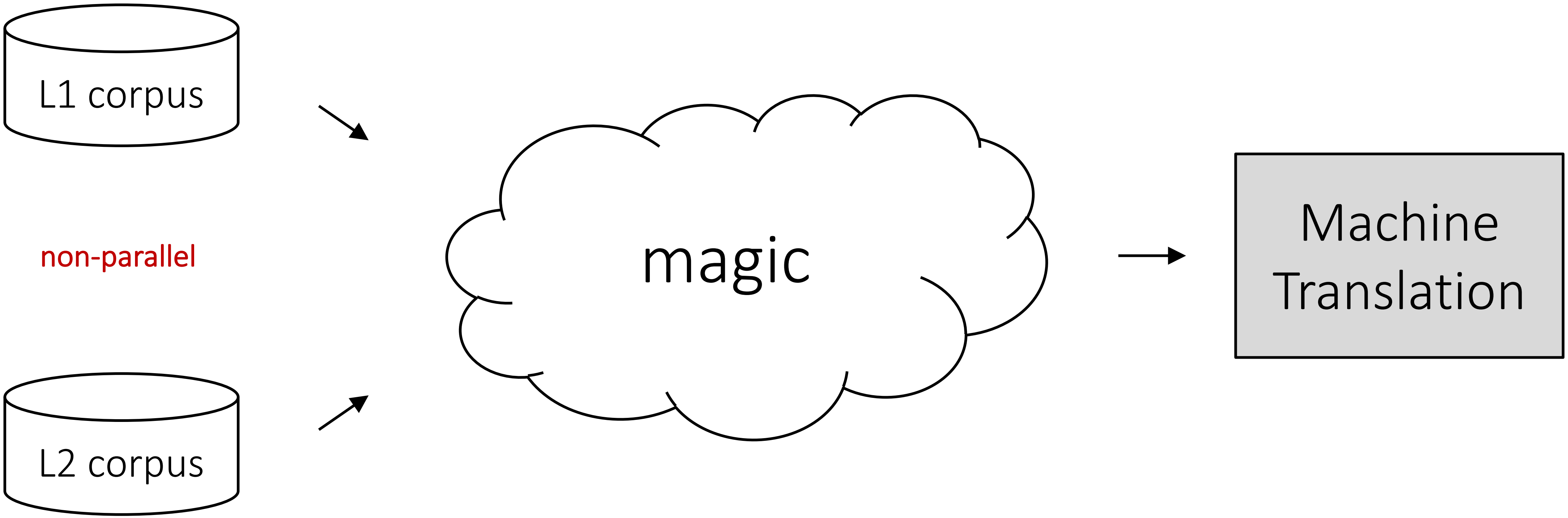
Outline



Outline



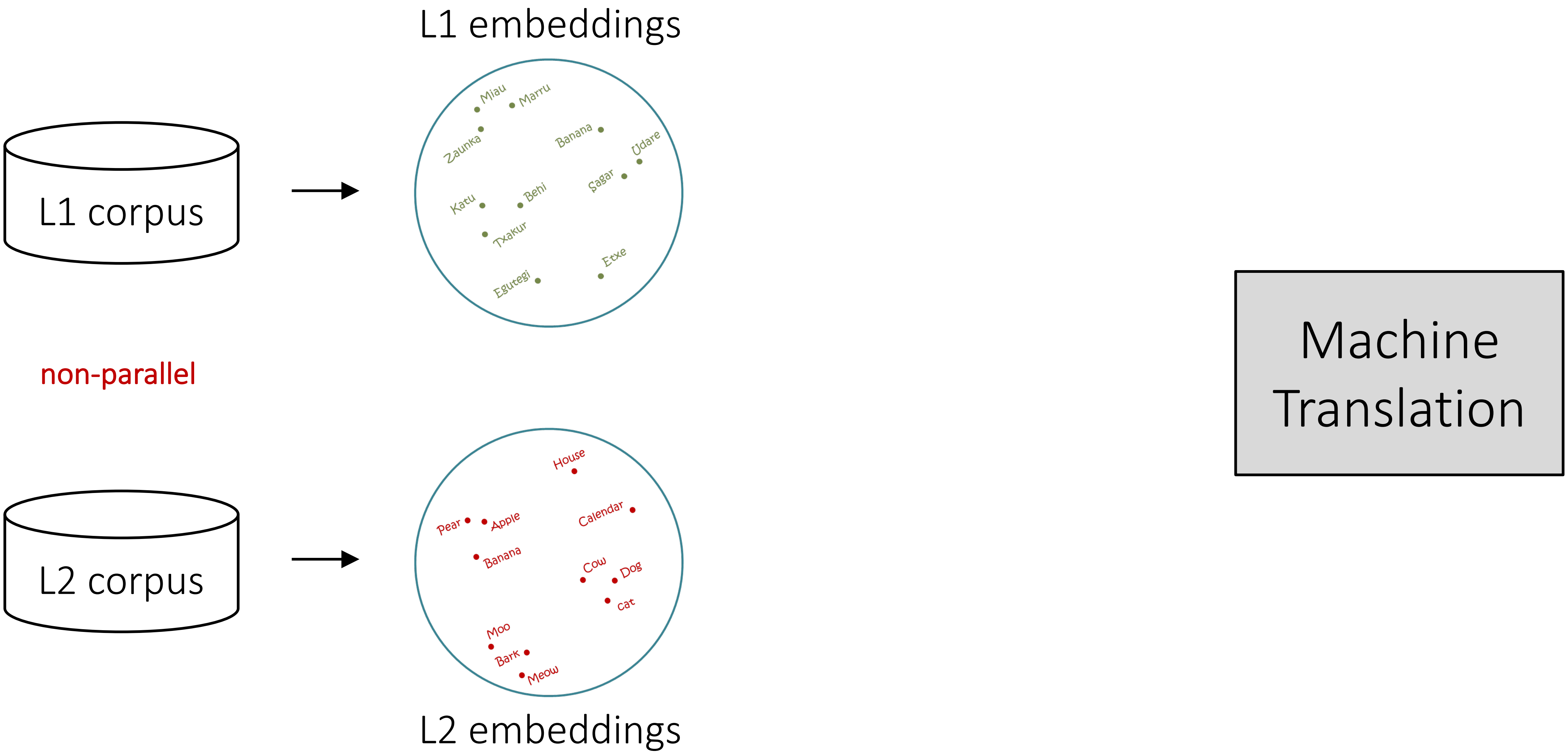
Outline


L1 corpus
non-parallel
L2 corpus
magic
Machine Translation

Outline

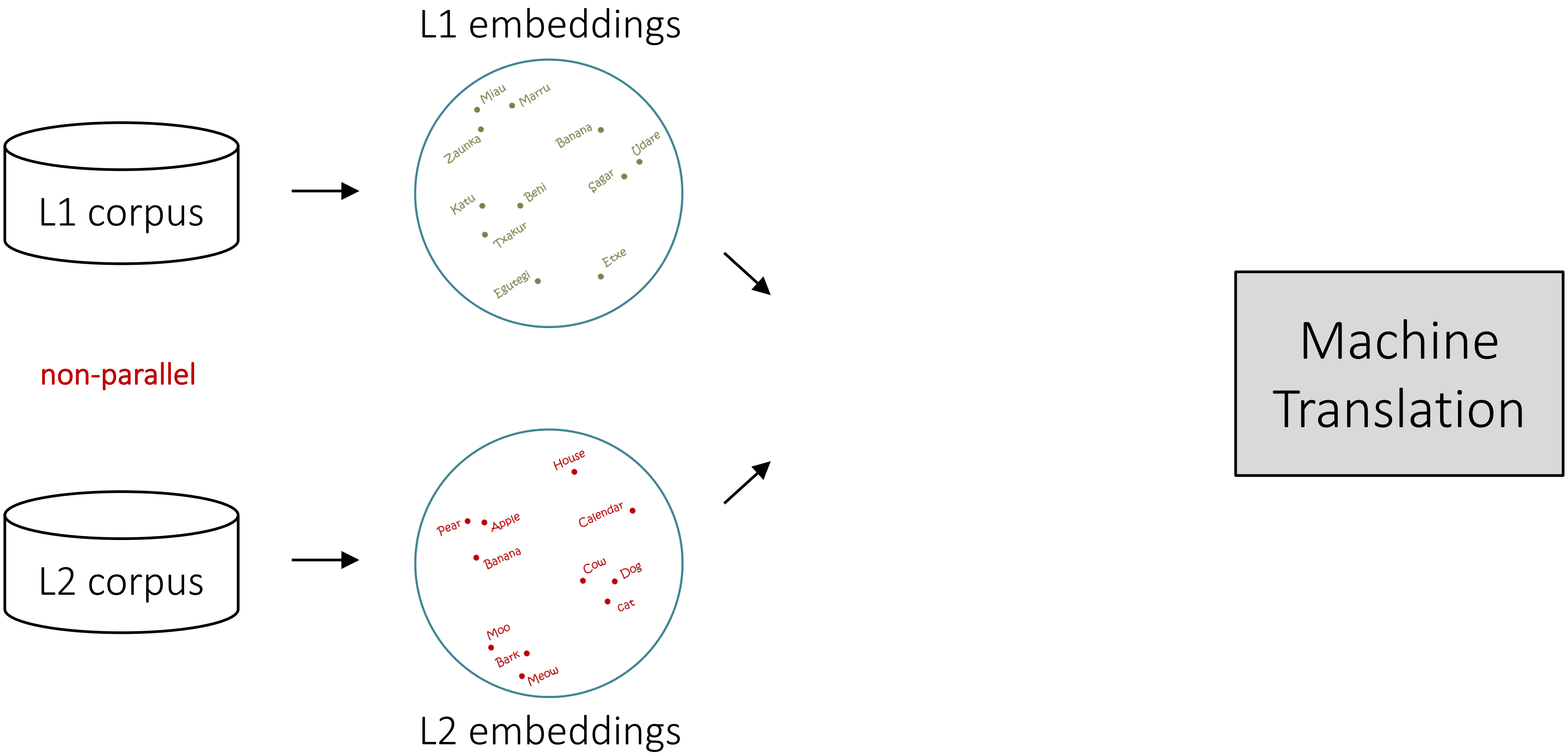
Outline

Outline

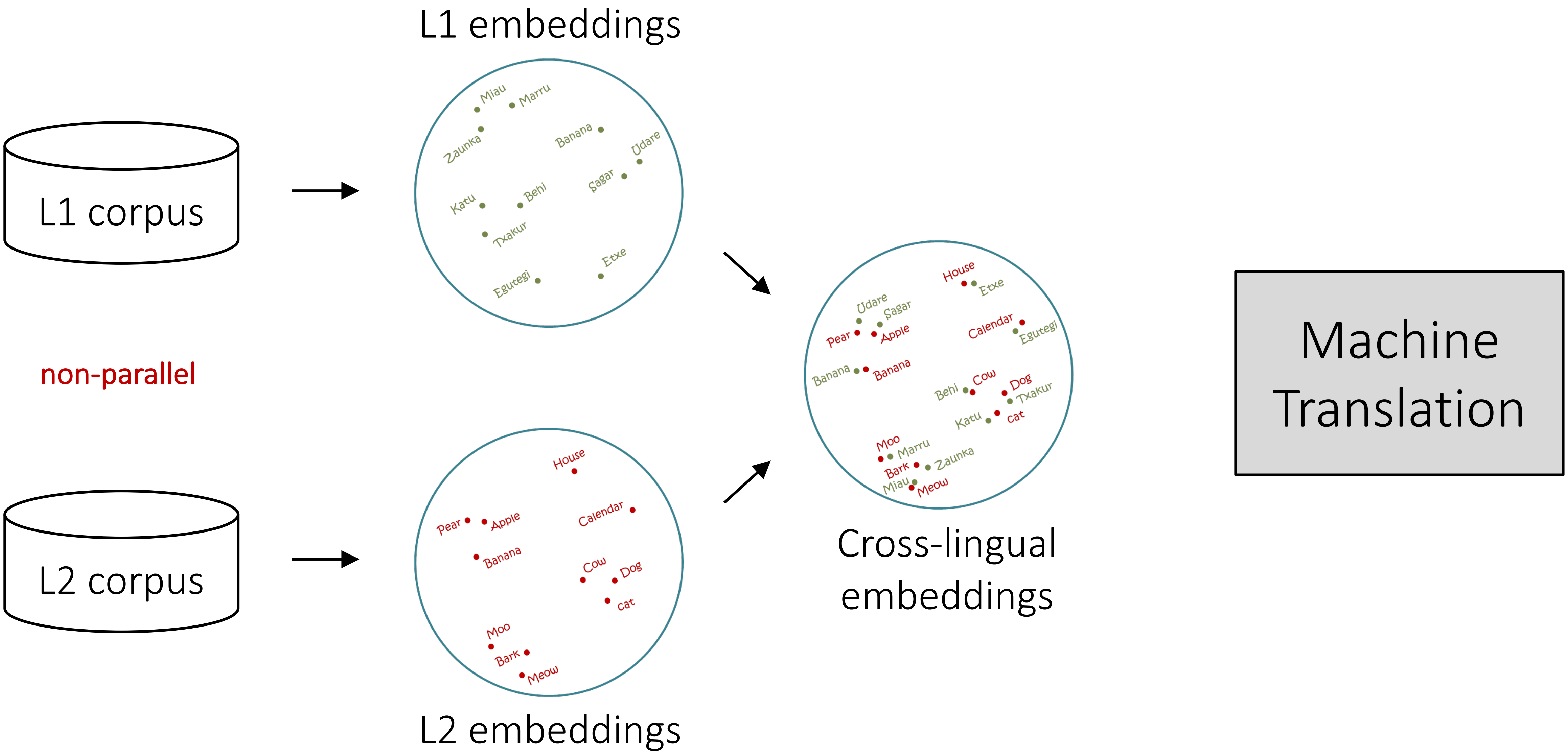


L1 embeddings
L1 corpus
Miau
Marru
Zaunka
Banana
Udare
Sagar
Katu
Behi
Txakur
Egutegi
Etxe
non-parallel
L2 corpus
House
Pear
Apple
Calendar
Banana
Cow
Dog
cat
Moo
Bark
Meow
L2 embeddings
Machine Translation

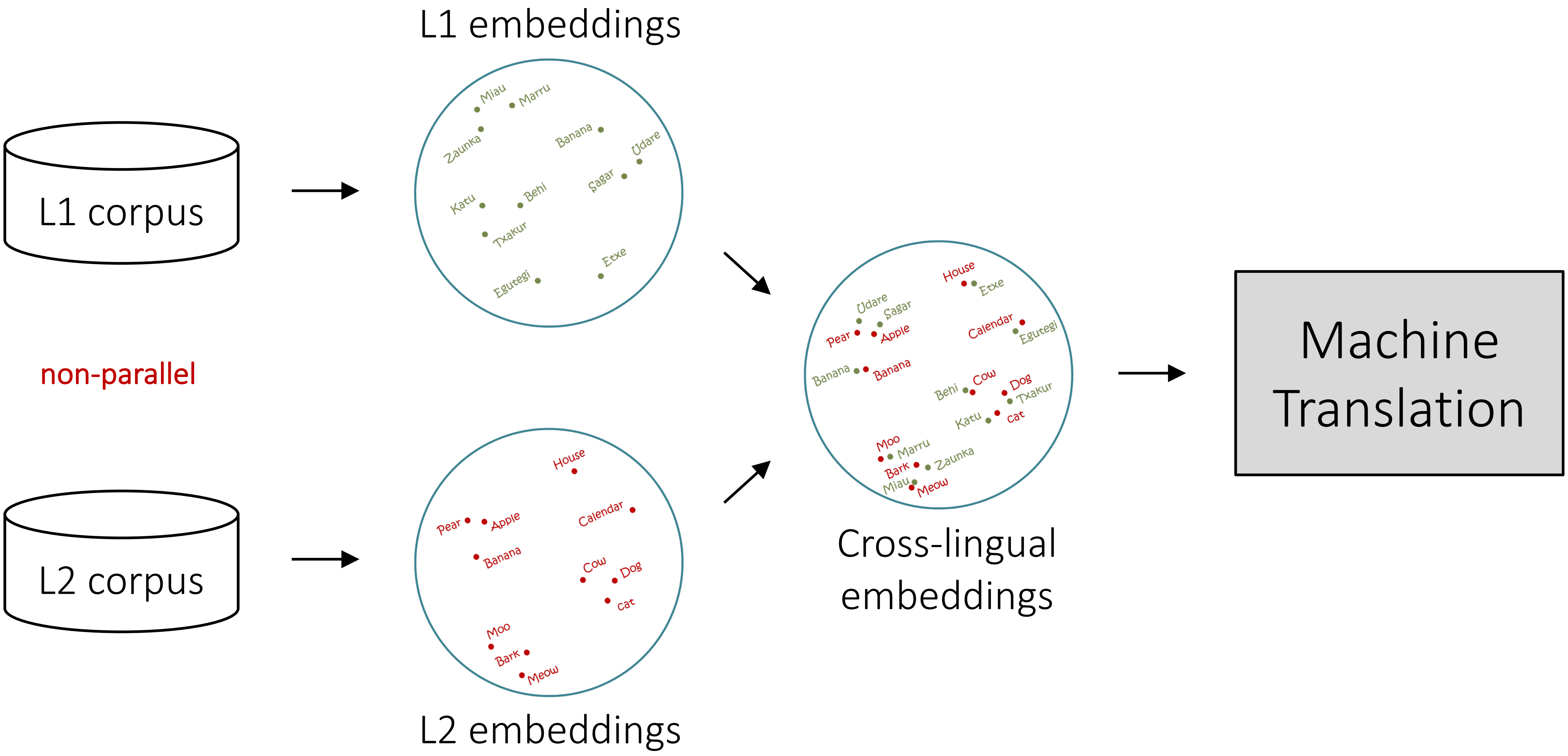
Outline


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

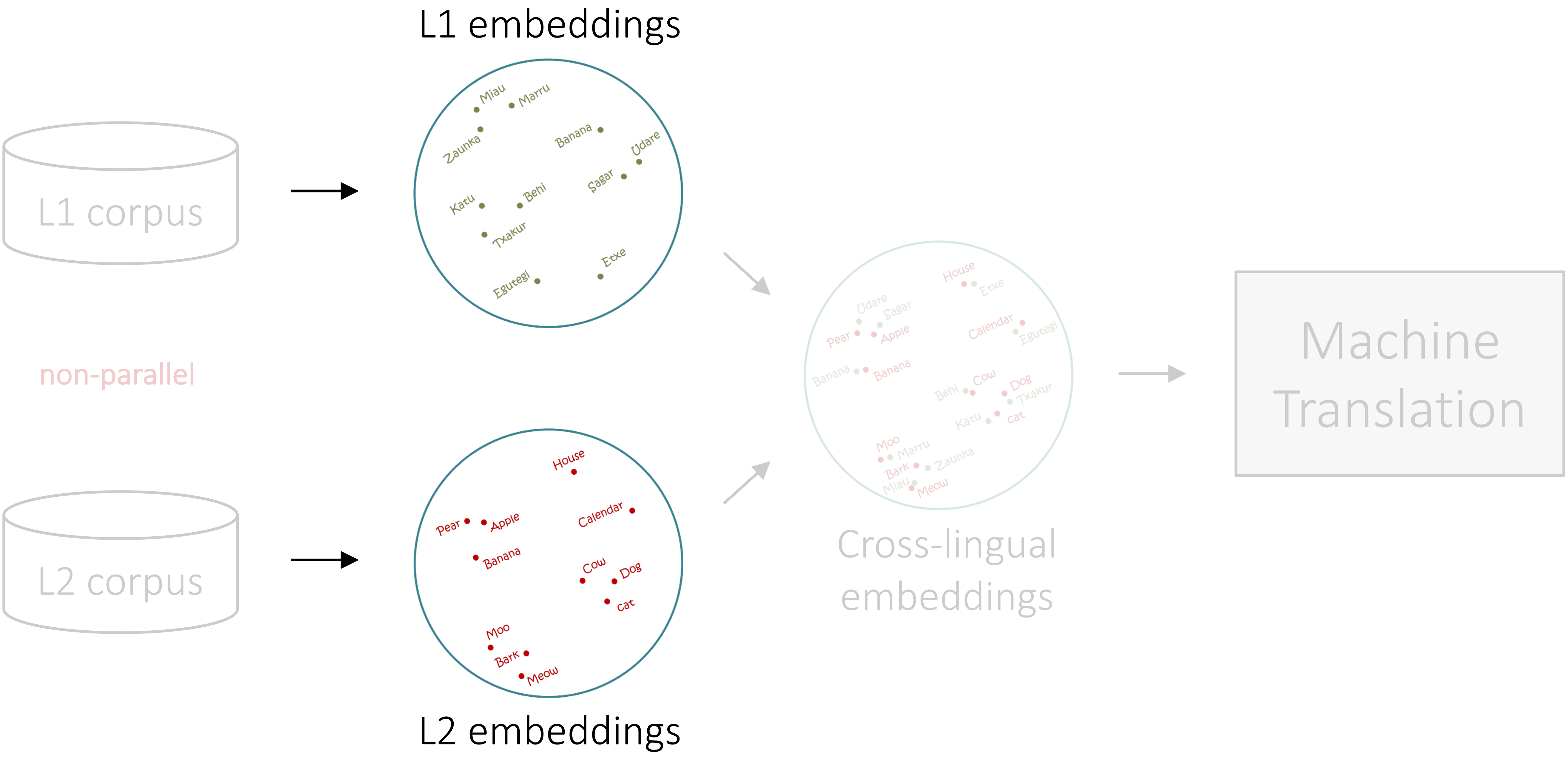
Outline


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

Outline


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

Outline



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

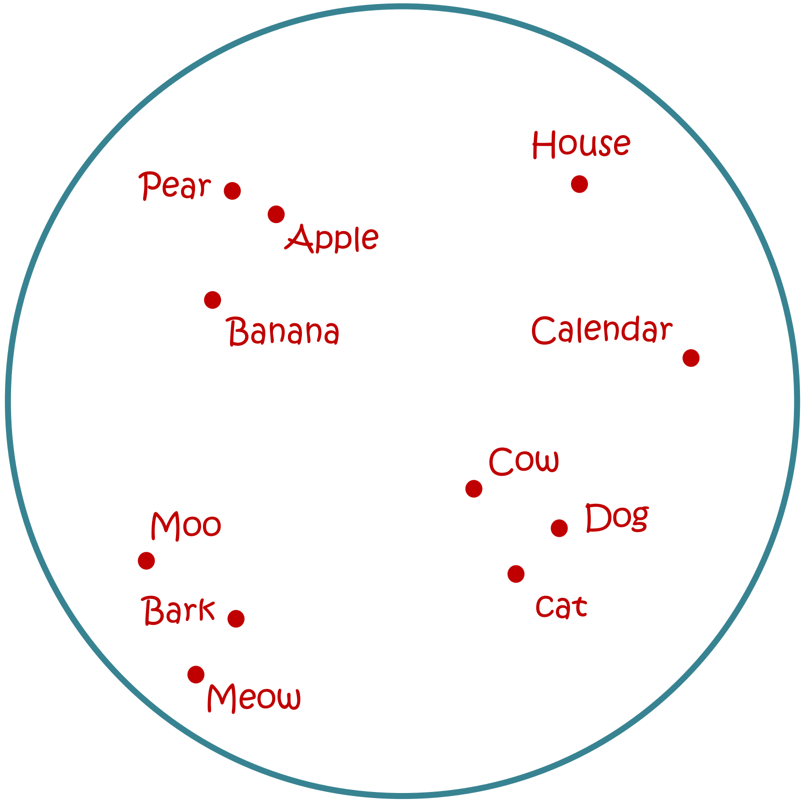
Word embeddings

Word embeddings

Distributed representations of words

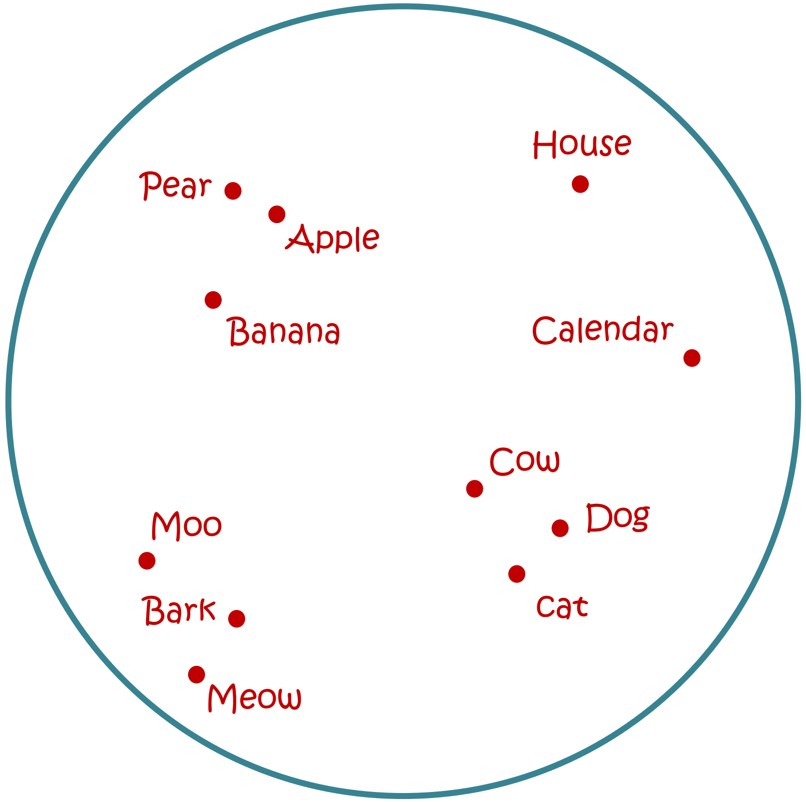
Word embeddings

Distributed representations of words



Word embeddings

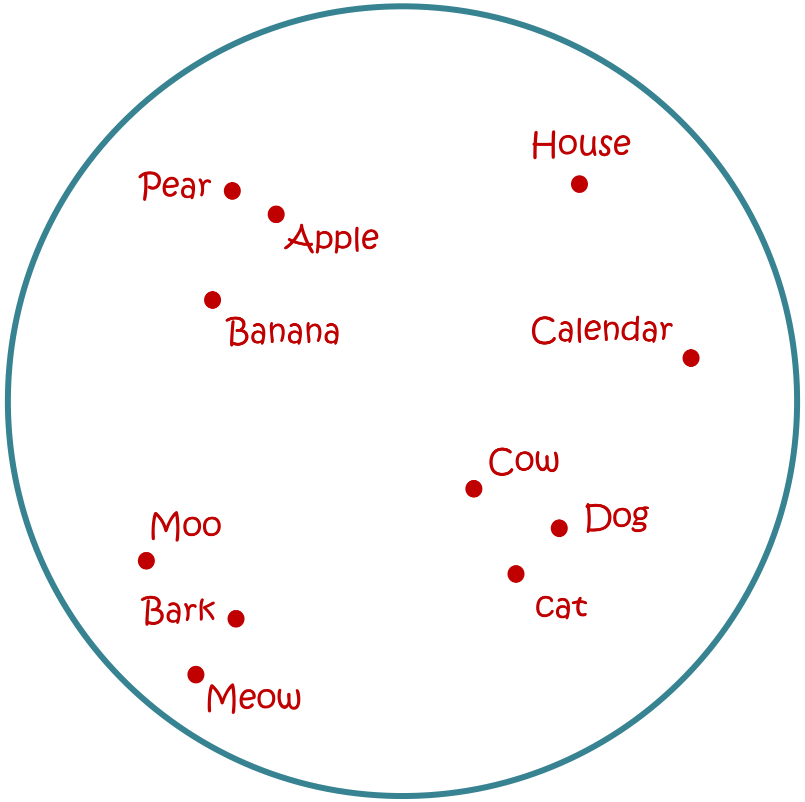
Distributed representations of words



$$\mathrm{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\| \|w_{cat}\|}$$

Word embeddings

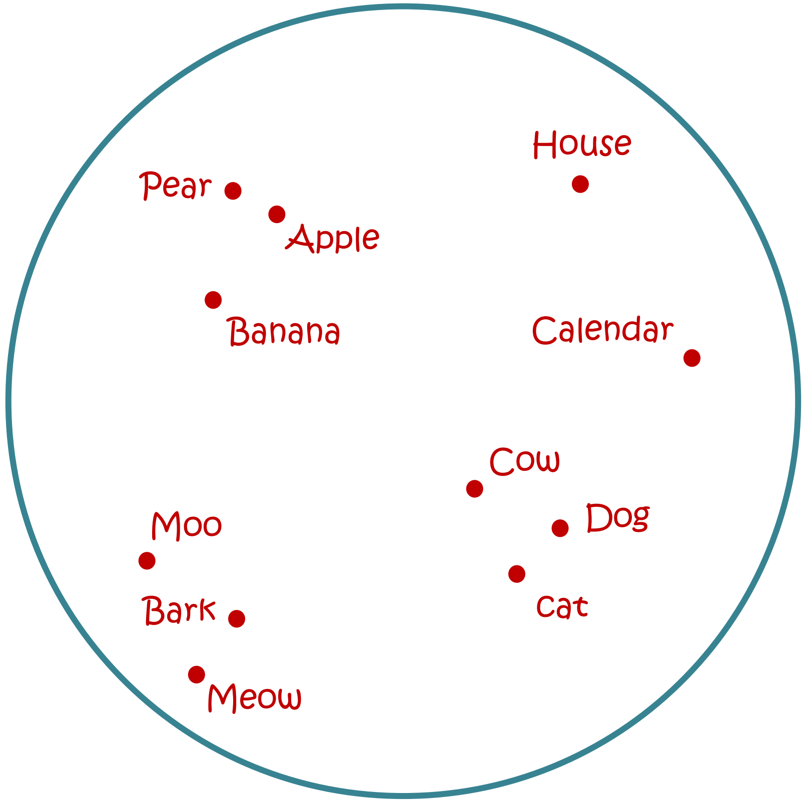
Distributed representations of words



$$\mathrm{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\| \|w_{cat}\|}$$

Word embeddings

Distributed representations of words

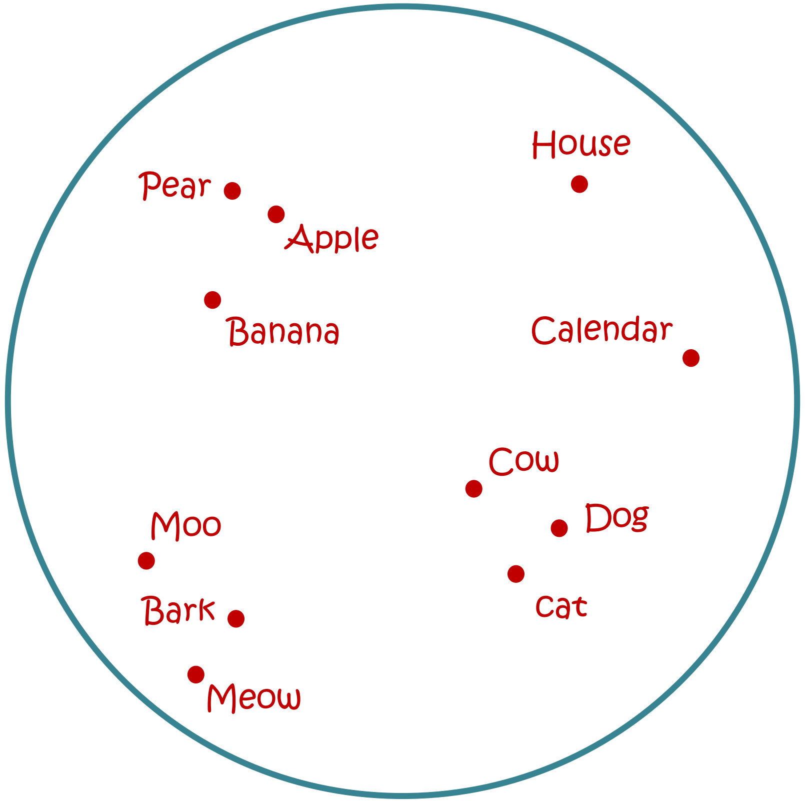


$$\mathrm{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\| \|w_{cat}\|}$$

Learned from co-occurrence patterns in a monolingual corpus

Word embeddings

Distributed representations of words



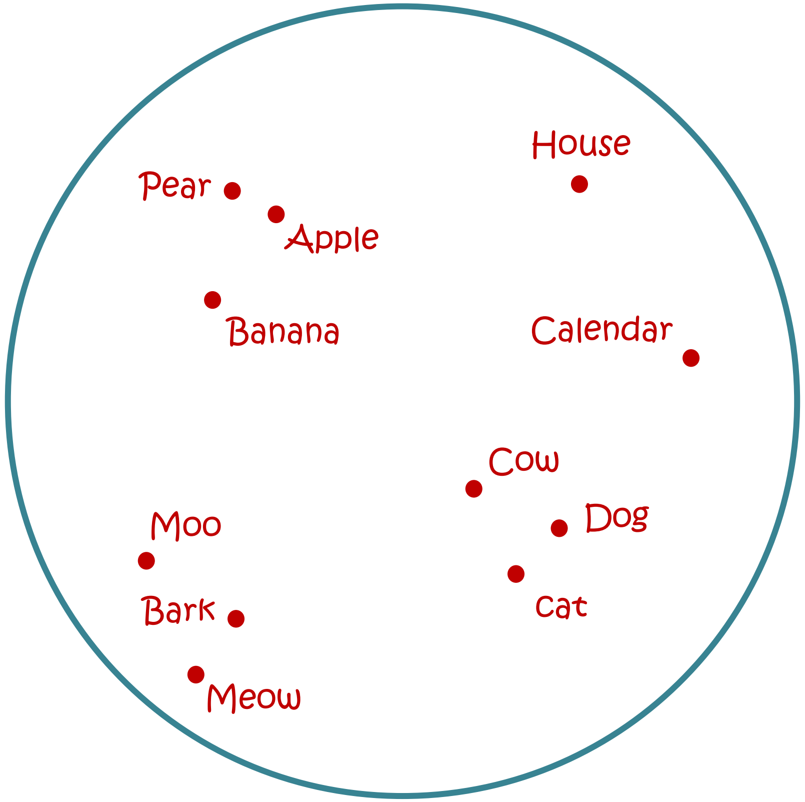
$$\text{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\| \|w_{cat}\|}$$

Learned from co-occurrence patterns in a monolingual corpus

e.g. skip-gram with negative sampling
(Mikolov et al., NIPS'13)

Word embeddings

Distributed representations of words



$$\mathrm{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\|\|w_{cat}\|}$$

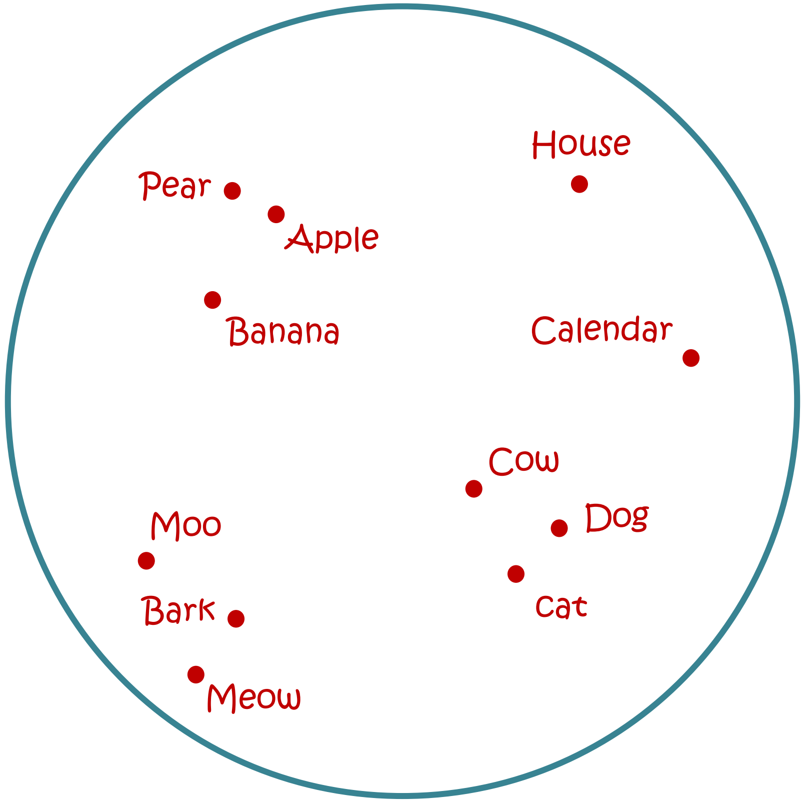
Learned from co-occurrence patterns in a monolingual corpus

e.g. skip-gram with negative sampling
(Mikolov et al., NIPS'13)

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

Word embeddings

Distributed representations of words



$$\mathrm{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\| \|w_{cat}\|}$$

Learned from co-occurrence patterns
in a monolingual corpus

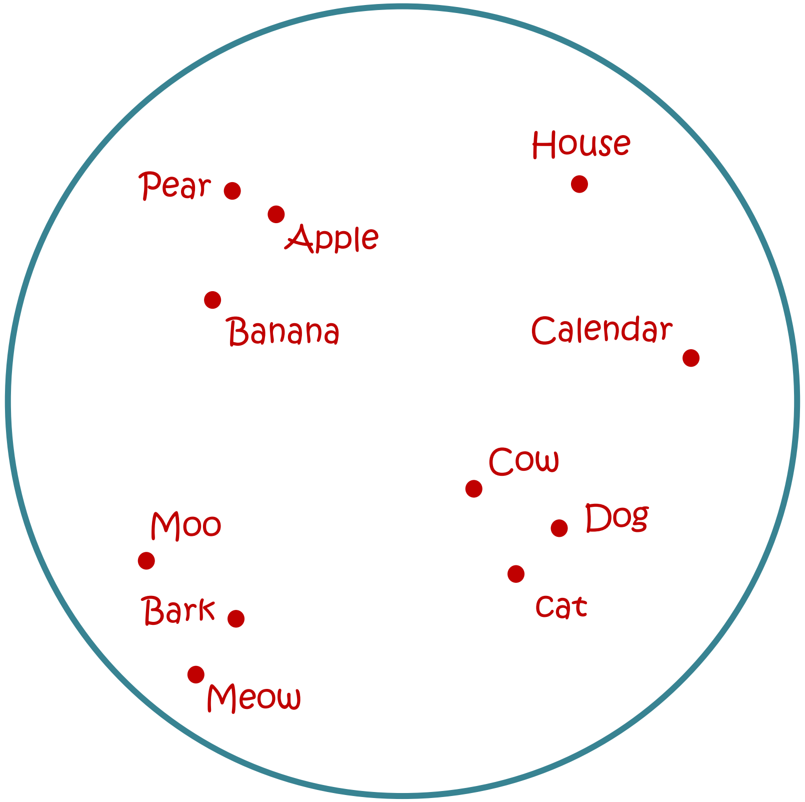
e.g. skip-gram with negative sampling
(Mikolov et al., NIPS'13)

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

I will go to New York by plane .

Word embeddings

Distributed representations of words



$$\mathrm{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\| \|w_{cat}\|}$$

Learned from co-occurrence patterns in a monolingual corpus

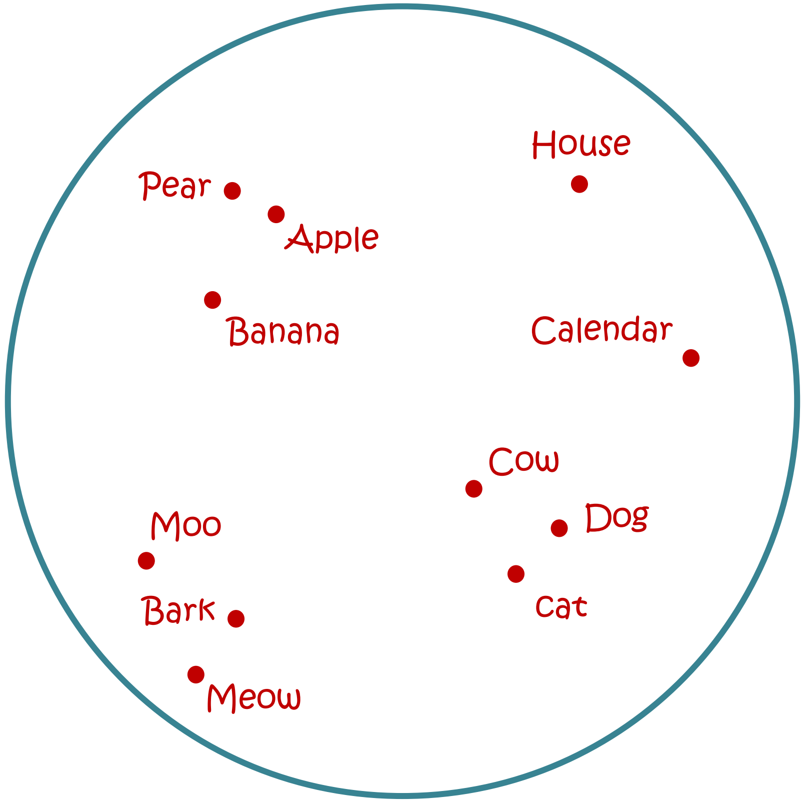
e.g. skip-gram with negative sampling
(Mikolov et al., NIPS'13)

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

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

Word embeddings

Distributed representations of words



$$\mathrm{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\| \|w_{cat}\|}$$

Learned from co-occurrence patterns
in a monolingual corpus

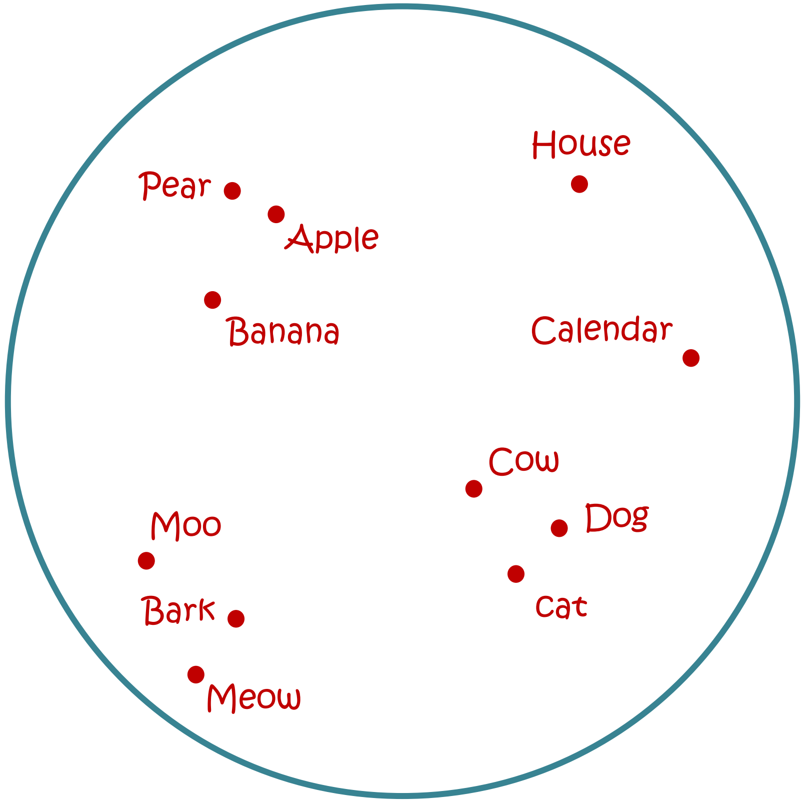
e.g. skip-gram with negative sampling
(Mikolov et al., NIPS'13)

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

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

Word embeddings

Distributed representations of words



$$\mathrm{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\| \|w_{cat}\|}$$

Learned from co-occurrence patterns in a monolingual corpus

e.g. skip-gram with negative sampling

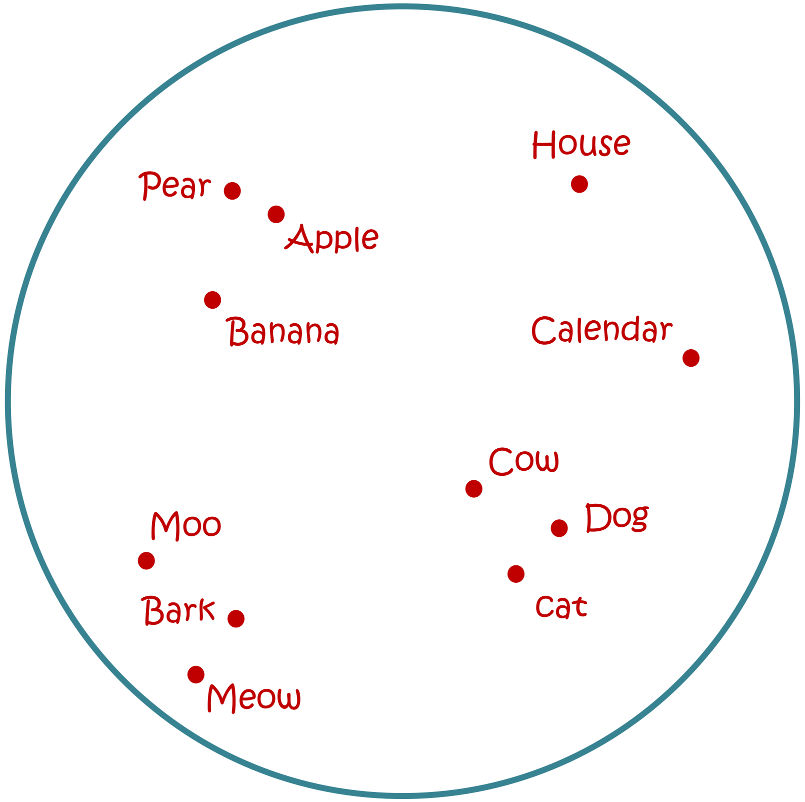
(Mikolov et al., NIPS'13)

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

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

Word embeddings

Distributed representations of words



$$\mathrm{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\| \|w_{cat}\|}$$

Learned from co-occurrence patterns in a monolingual corpus

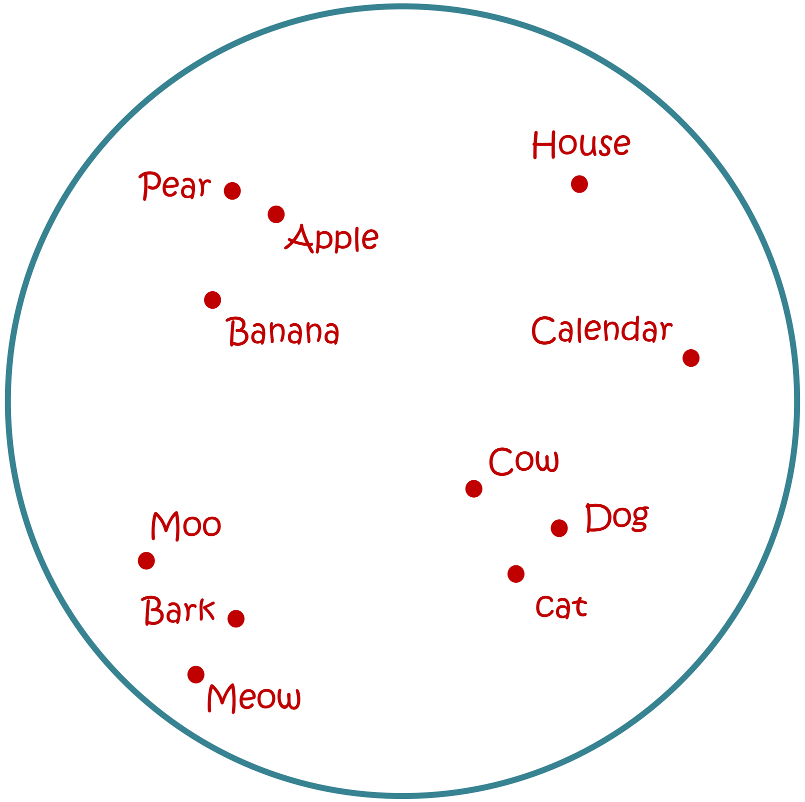
e.g. skip-gram with negative sampling
(Mikolov et al., NIPS'13)

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

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

Word embeddings

Distributed representations of words



$$\mathrm{sim}(cow, cat) \approx \cos(w_{cow}, w_{cat}) = \frac{w_{cow} \cdot w_{cat}}{\|w_{cow}\| \|w_{cat}\|}$$

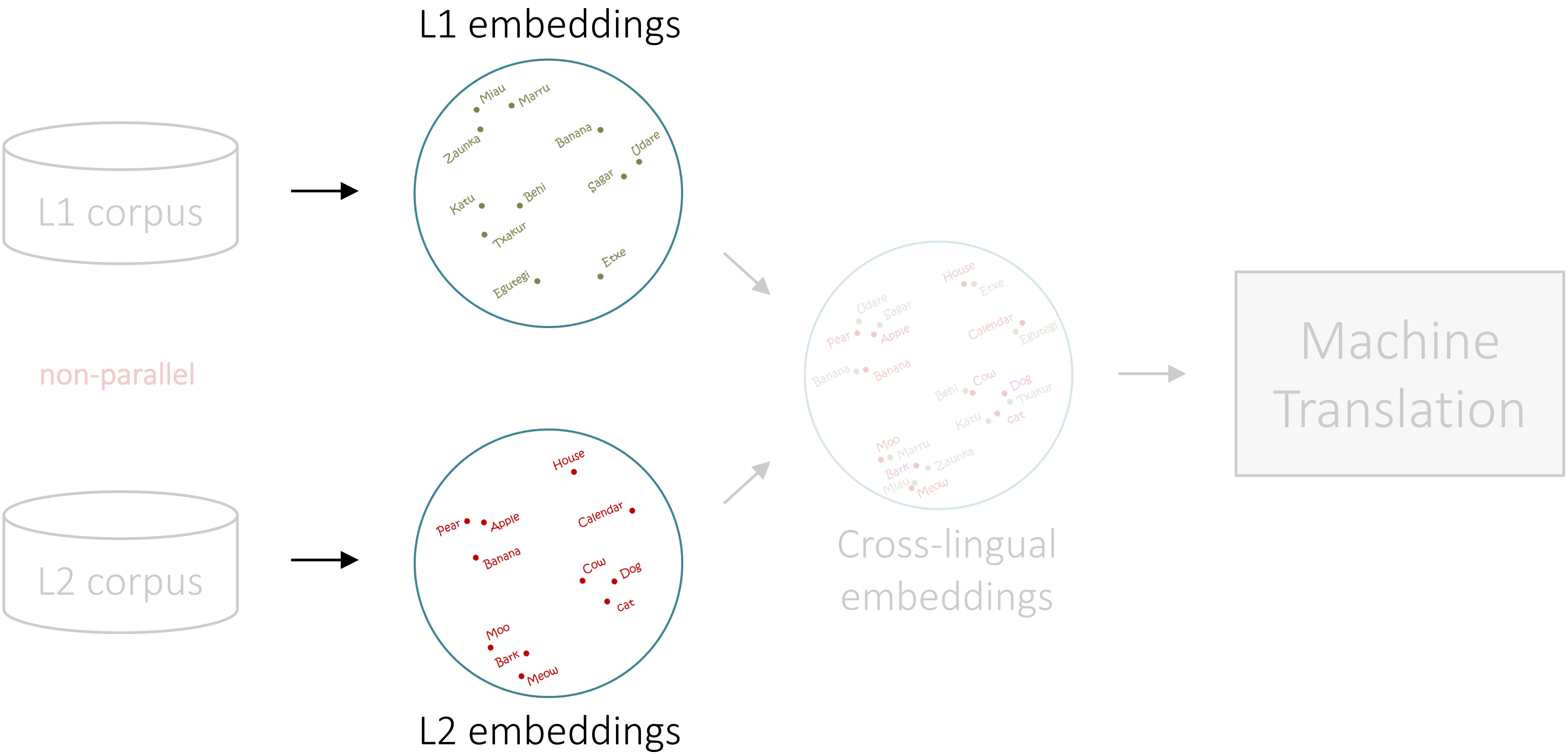
Learned from co-occurrence patterns in a monolingual corpus

e.g. skip-gram with negative sampling
(Mikolov et al., NIPS'13)

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

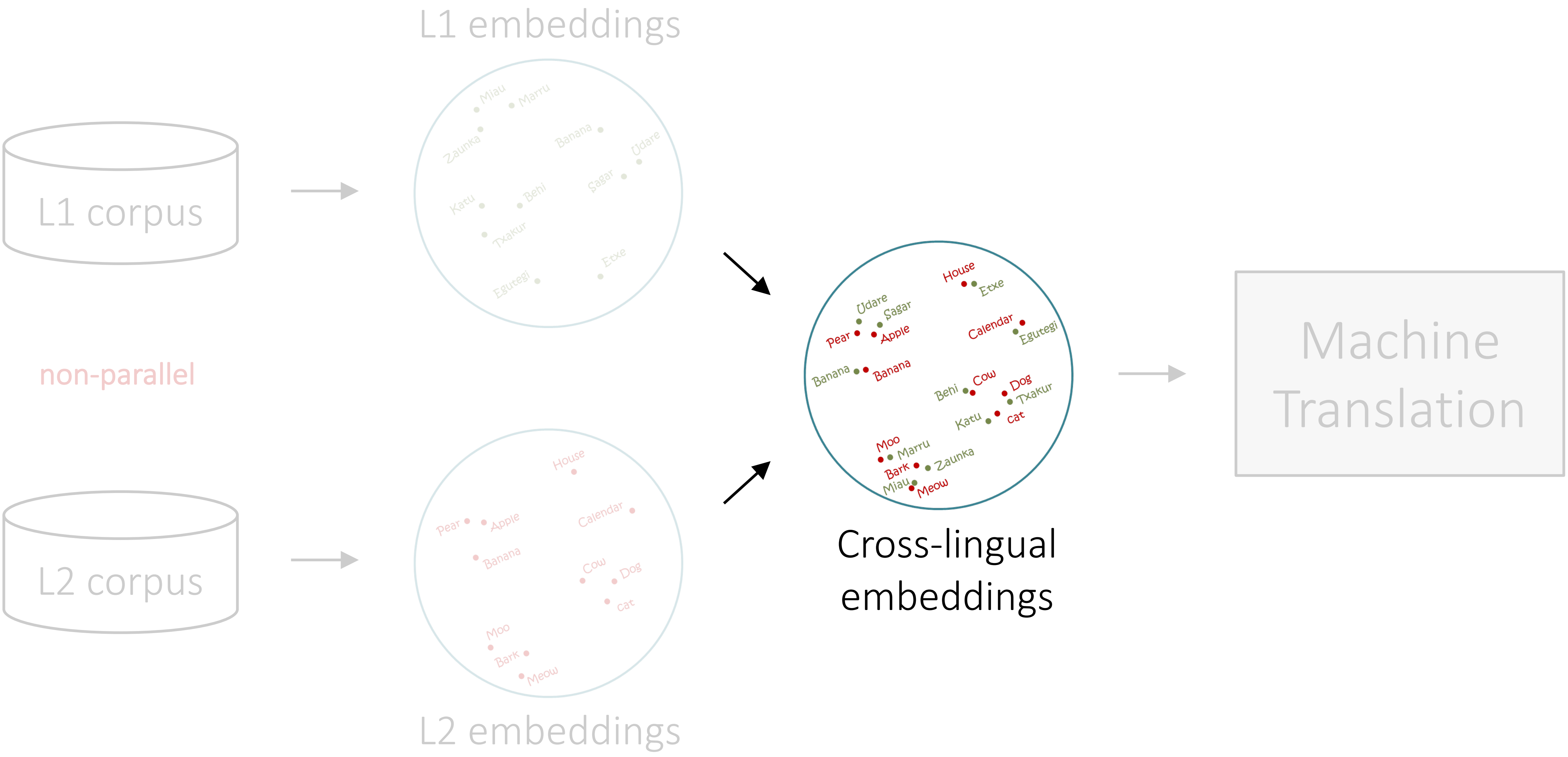
I will $\underset{w}{\underline{\textit{go}}}$ *to New* $\underset{c}{\underline{\textit{York}}}$ *by plane .*

Outline



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

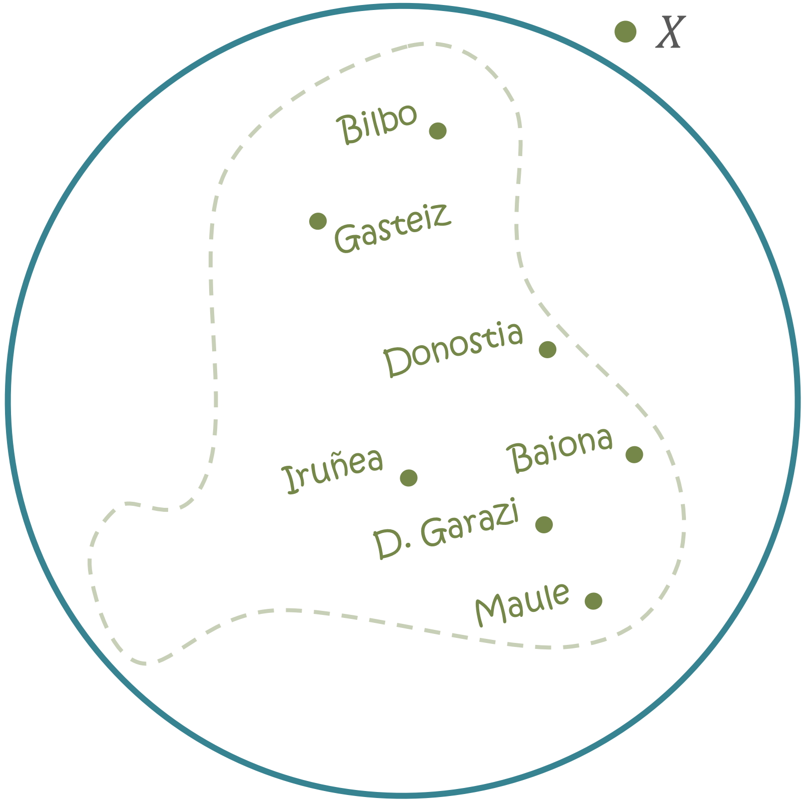
Outline



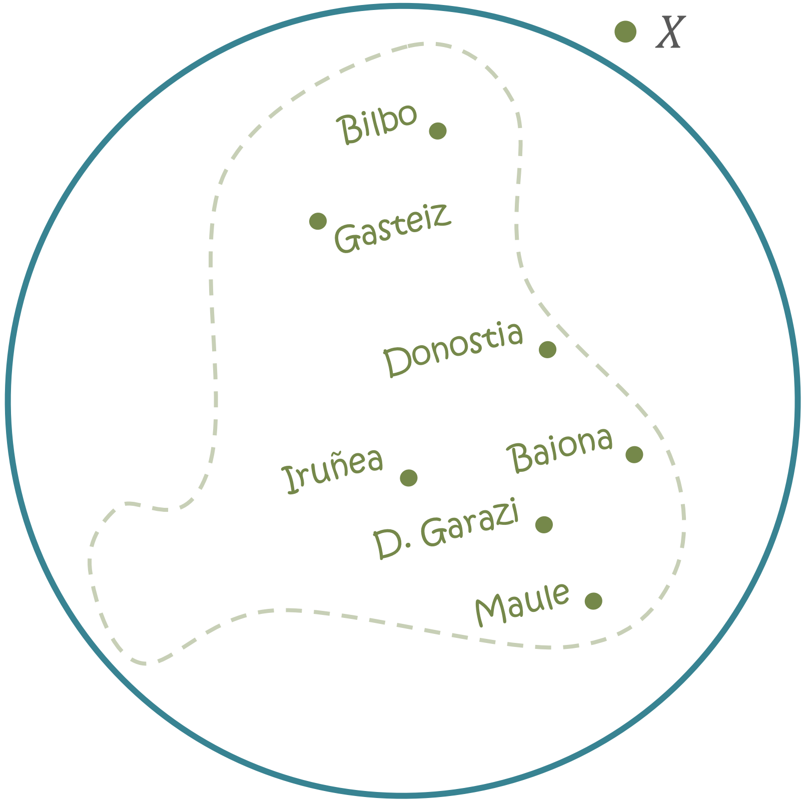
L1 embeddings
L1 corpus
Miau
Marru
Zaunka
Banana
Udare
Katu
Behi
Sagar
Txakur
Etxe
Egutegi
non-parallel
L2 corpus
House
Pear
Apple
Calendar
Banana
Cow
Dog
cat
Moo
Bark
Meow
L2 embeddings
House
Etxe
Udare
Sagar
Pear
Apple
Calendar
Egutegi
Banana
Banana
Behi
Cow
Dog
Txakur
Katu
cat
Moo
Marru
Bark
Zaunka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

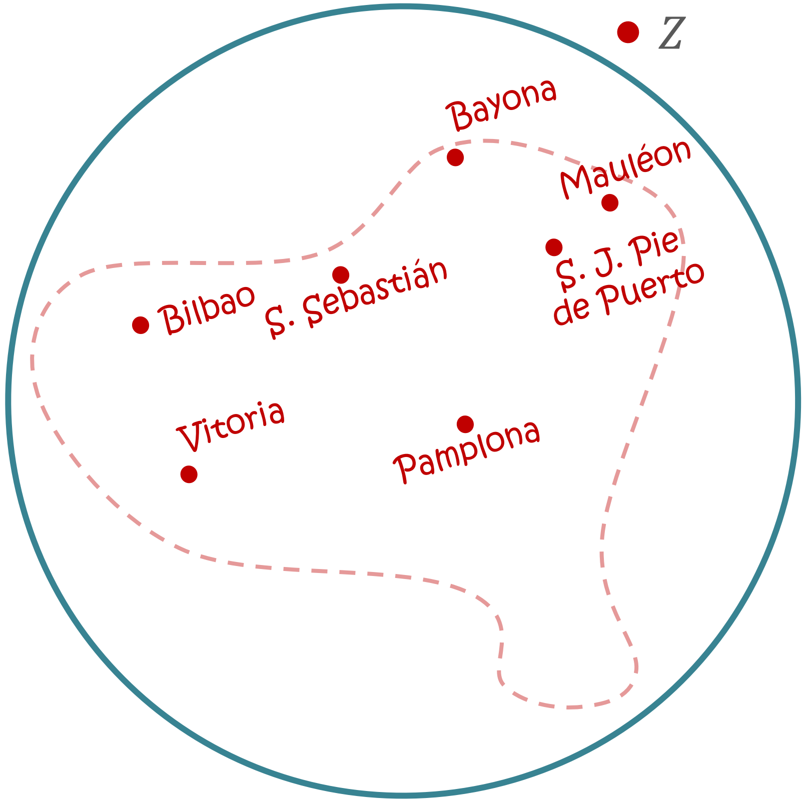
Cross-lingual word embedding alignment

Cross-lingual word embedding alignment

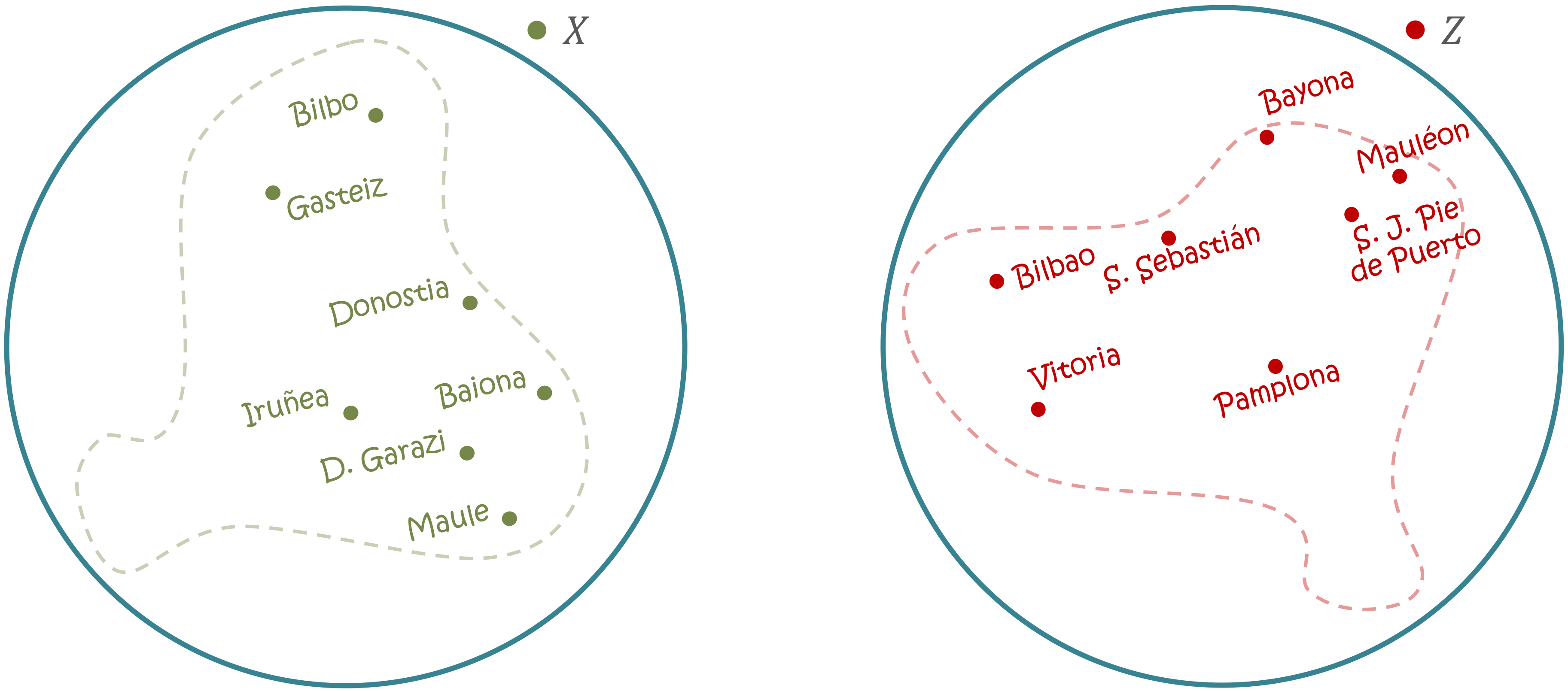


Cross-lingual word embedding alignment



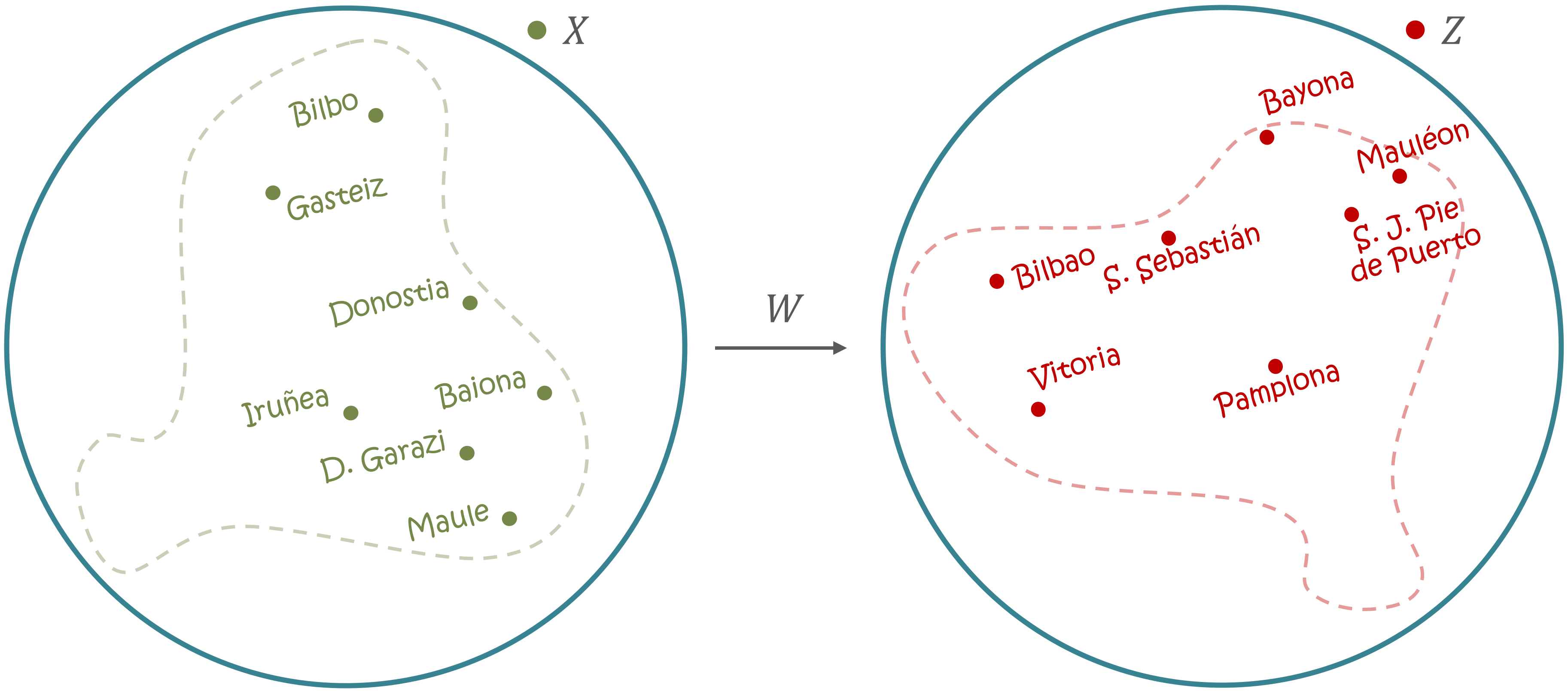


Cross-lingual word embedding alignment



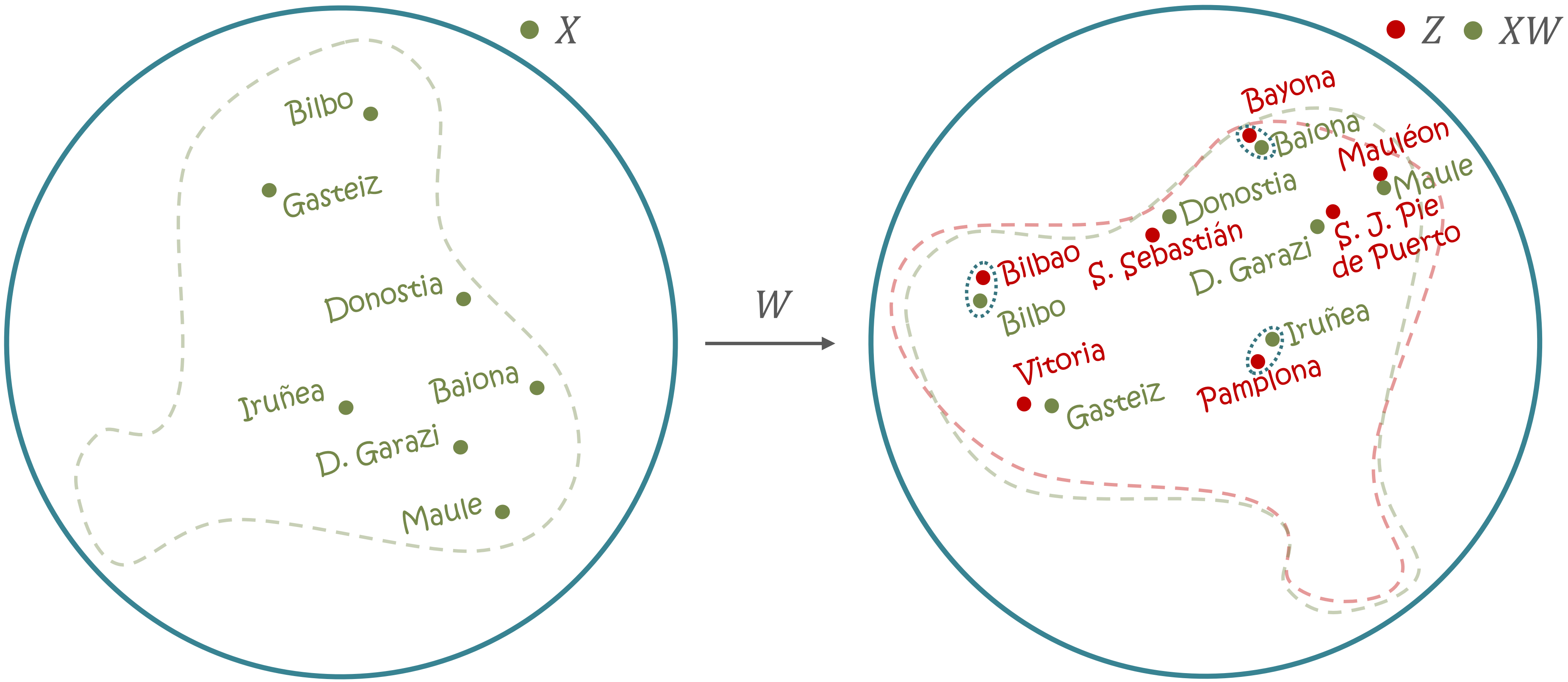
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Cross-lingual word embedding alignment



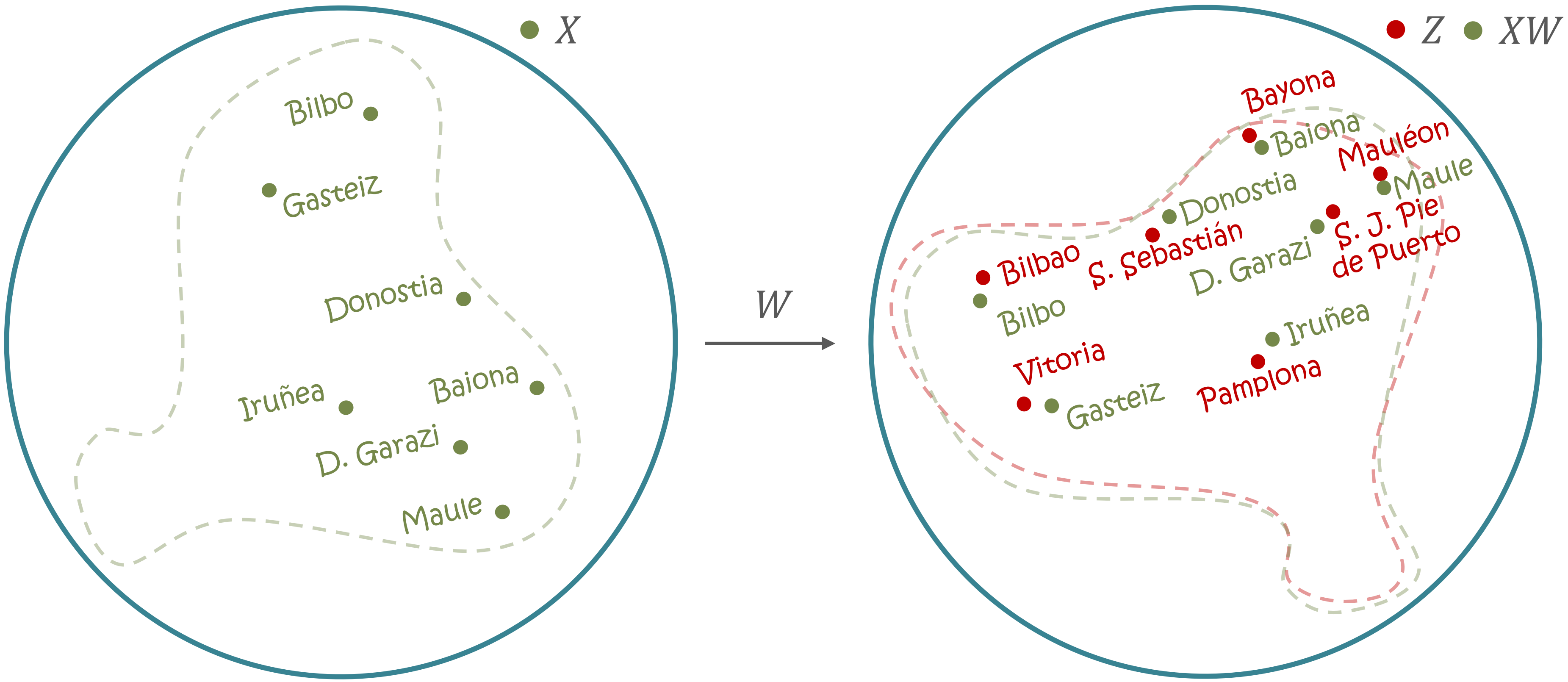
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Cross-lingual word embedding alignment



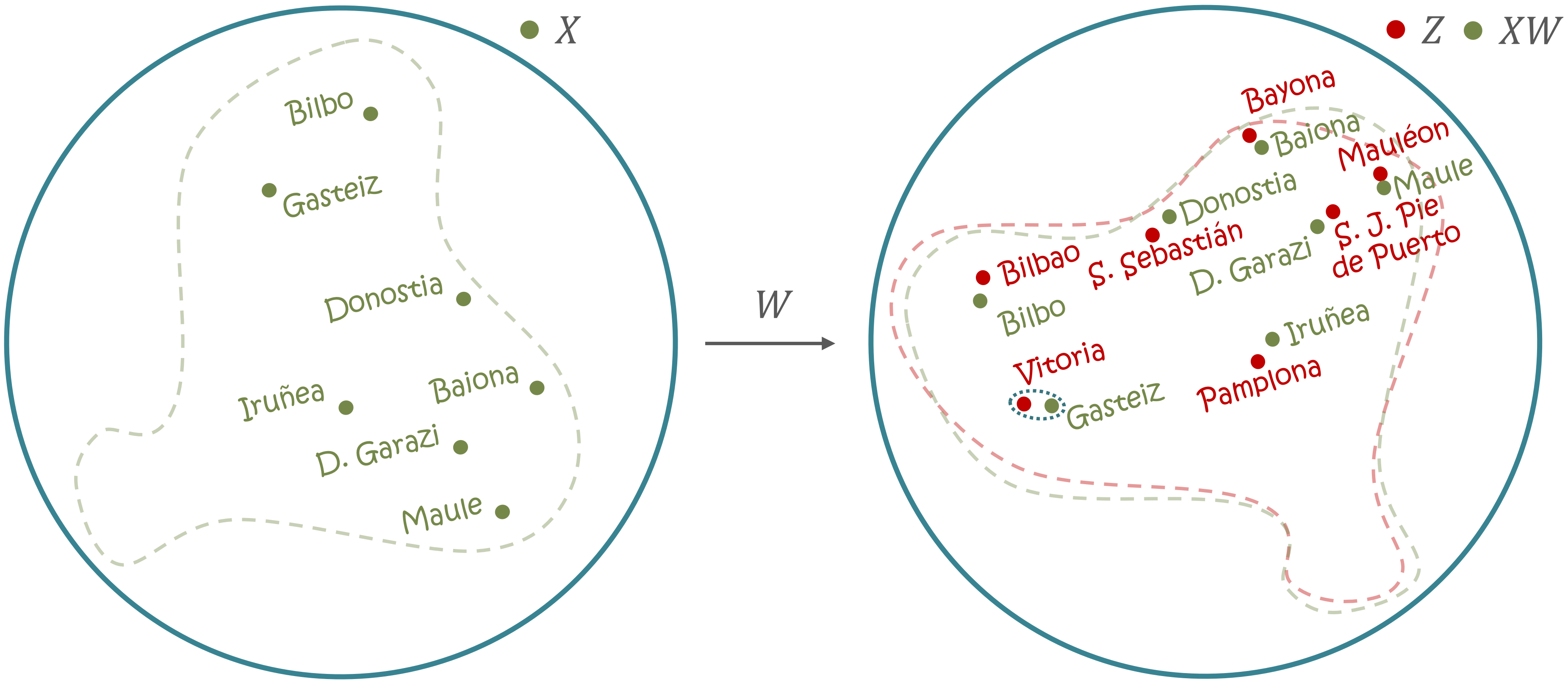
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Cross-lingual word embedding alignment



Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Cross-lingual word embedding alignment



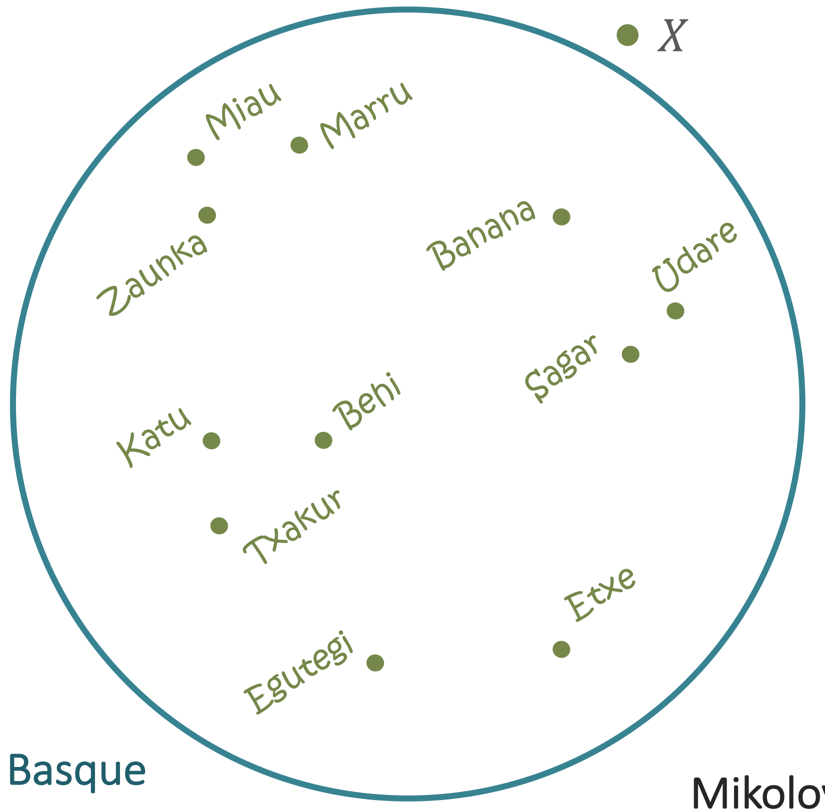
Bilbo	Bilbao
Baiona	Bayona
Iruñea	Pamplona

Cross-lingual word embedding alignment

Cross-lingual word embedding alignment

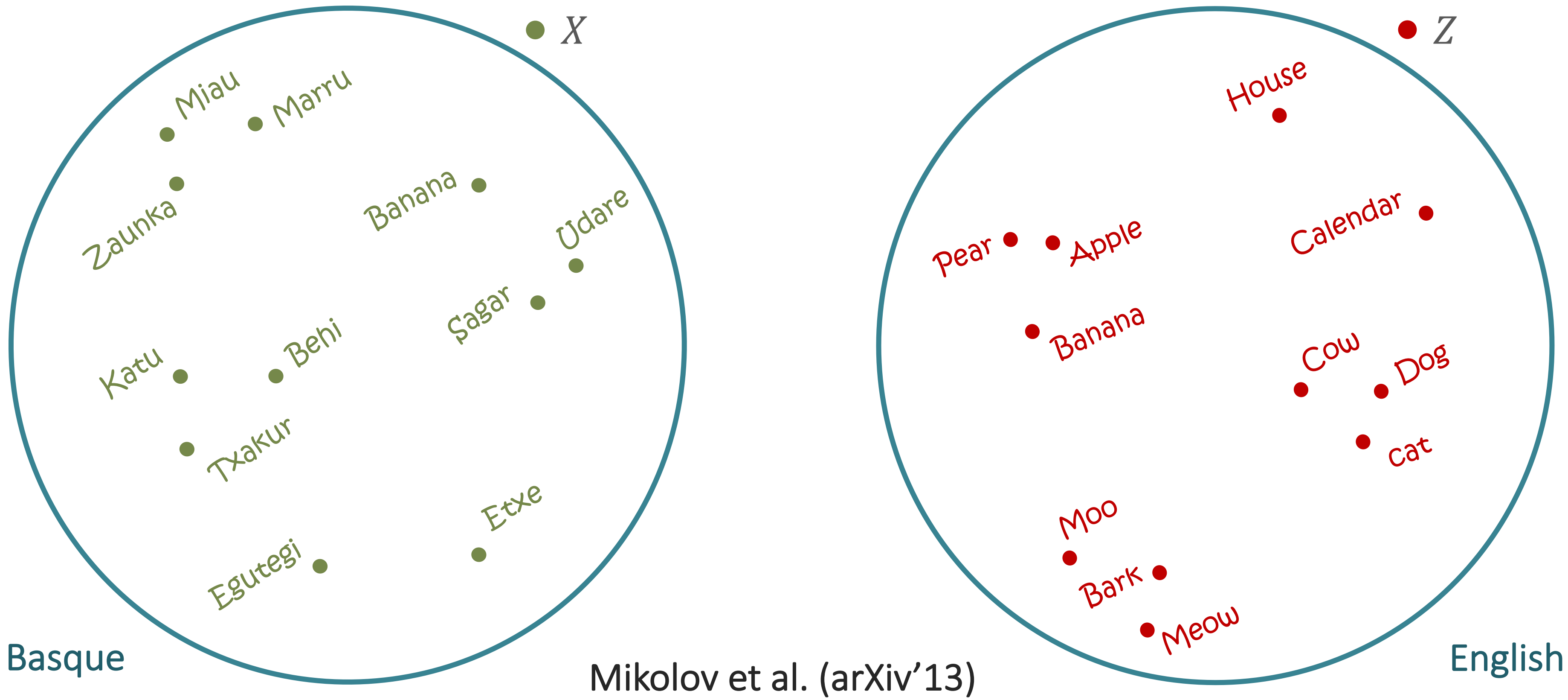
Mikolov et al. (arXiv'13)

Cross-lingual word embedding alignment



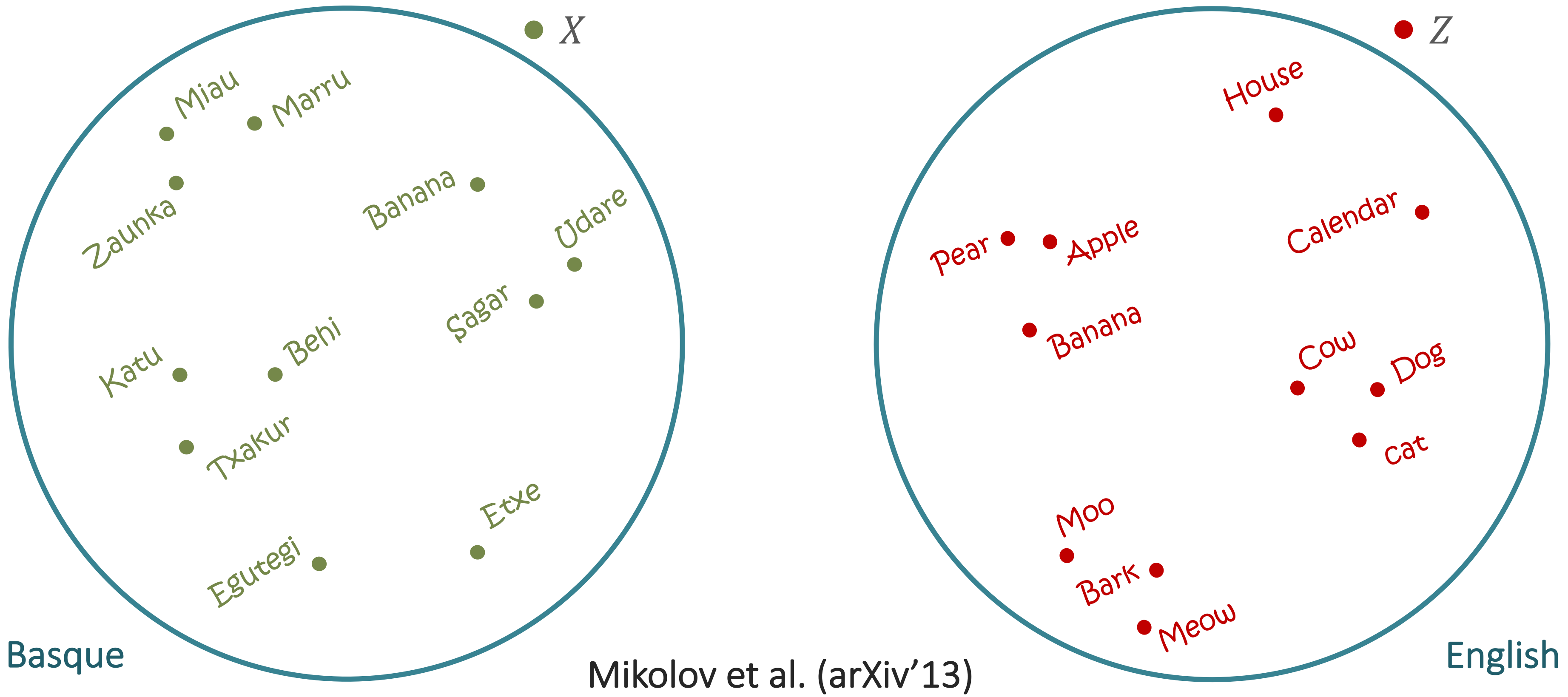
Mikolov et al. (arXiv’13)

Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

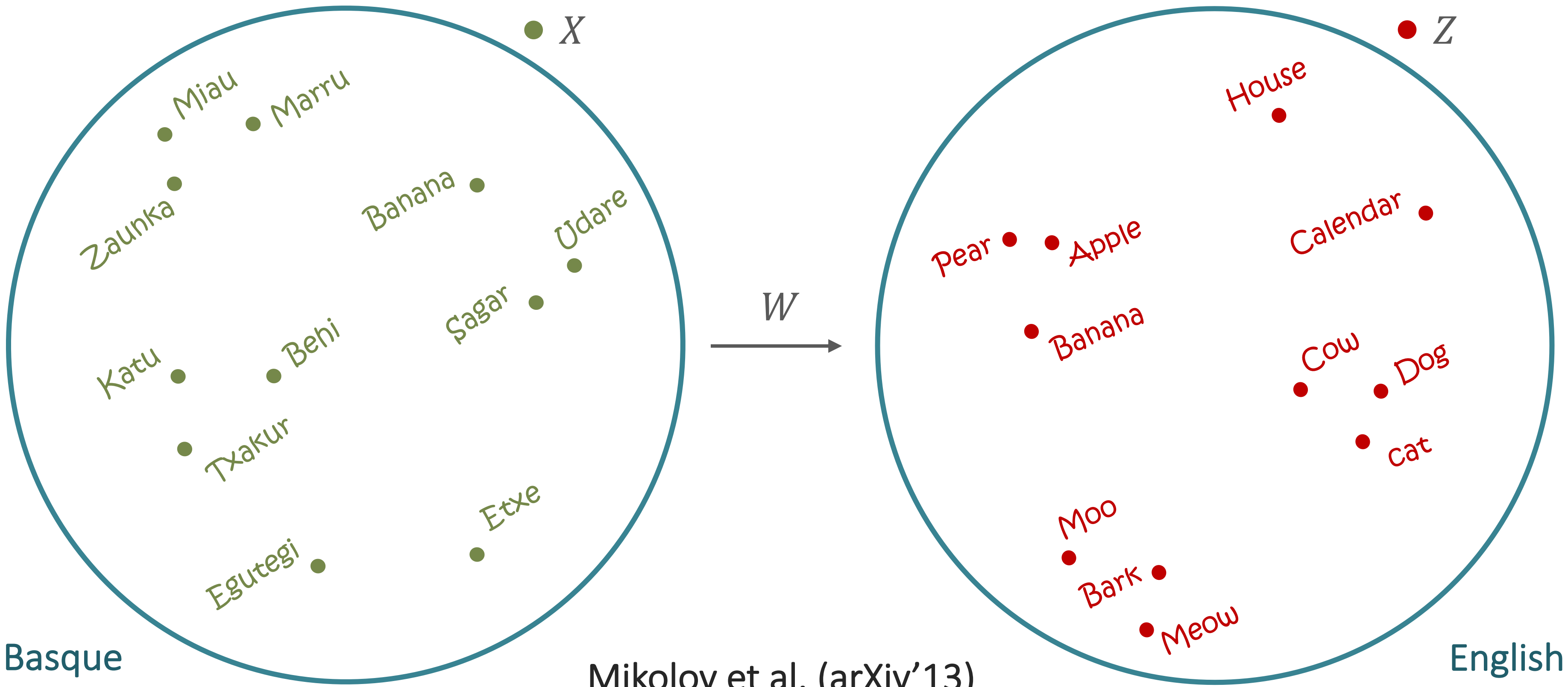
Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

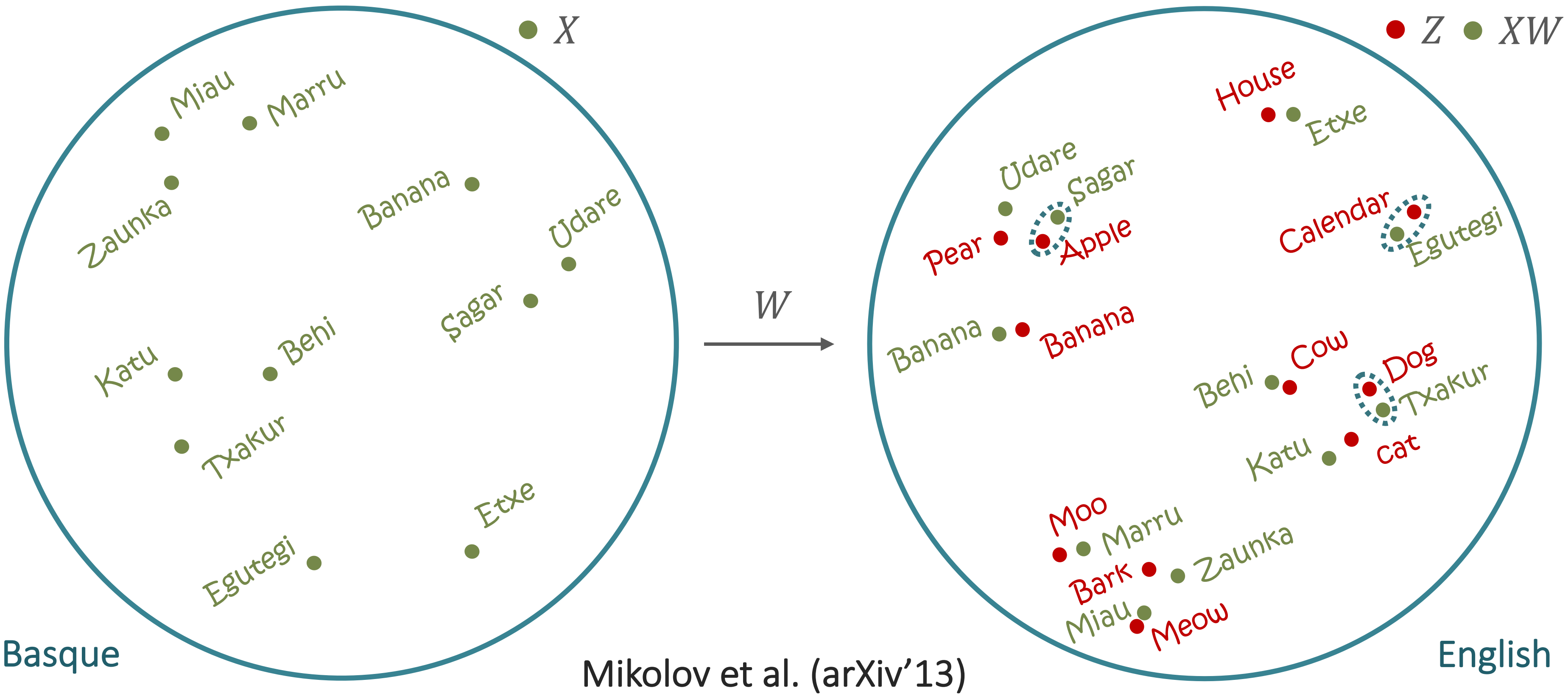
Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

Basque	English
Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

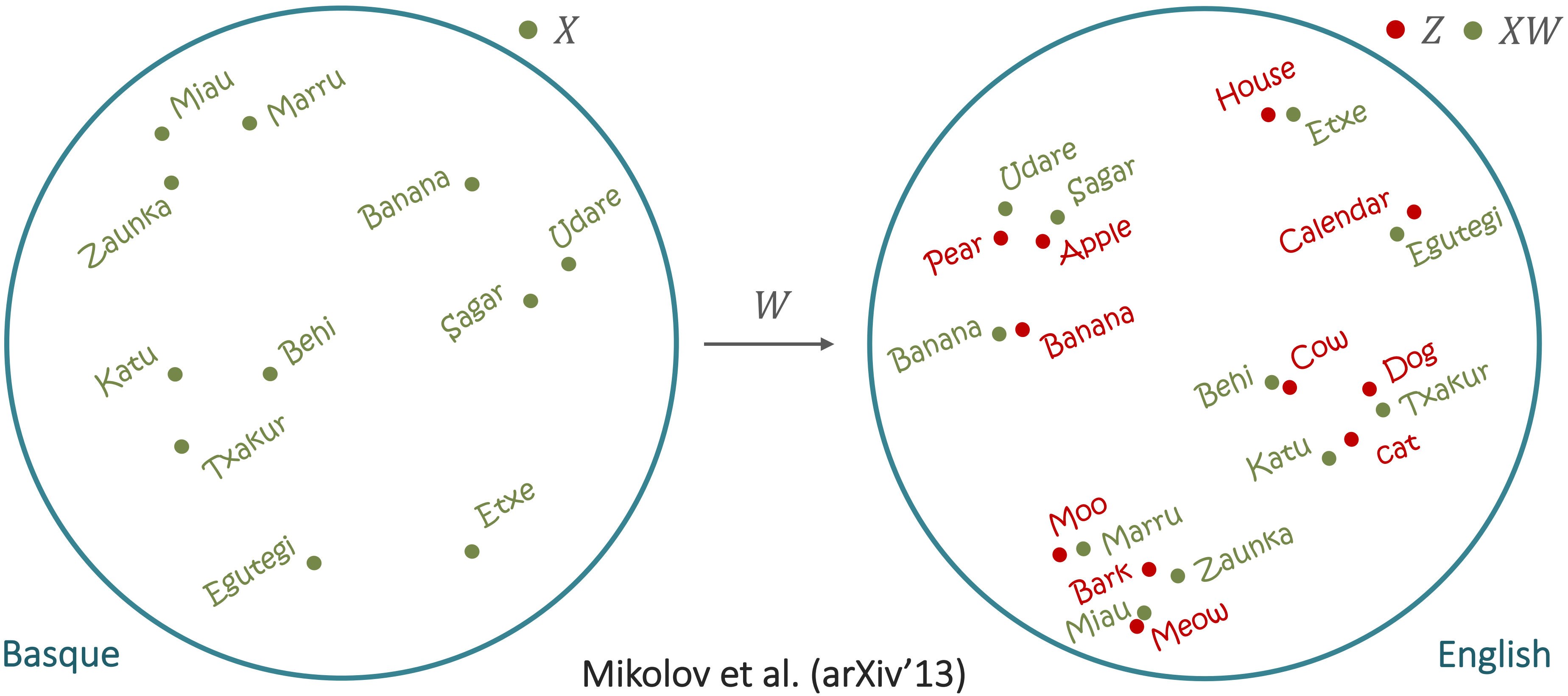
Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

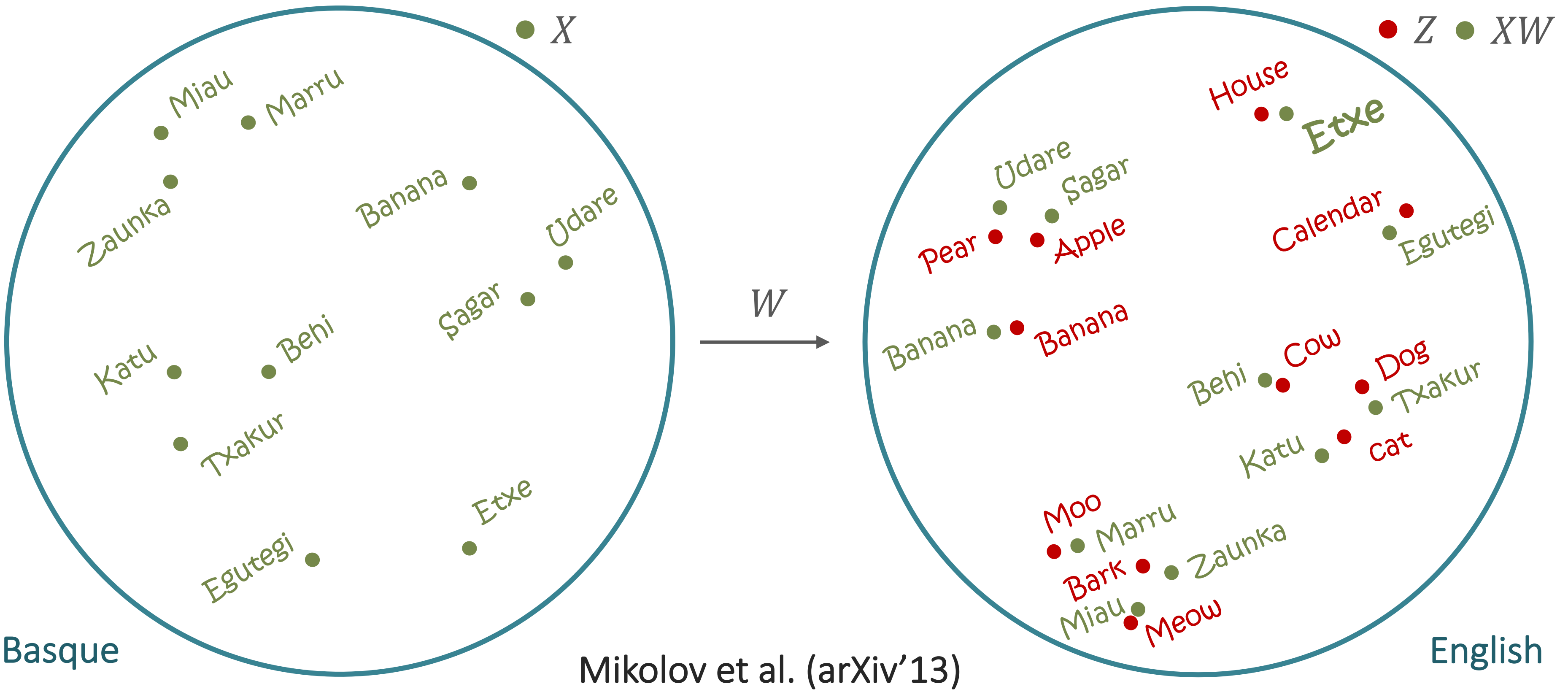
Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

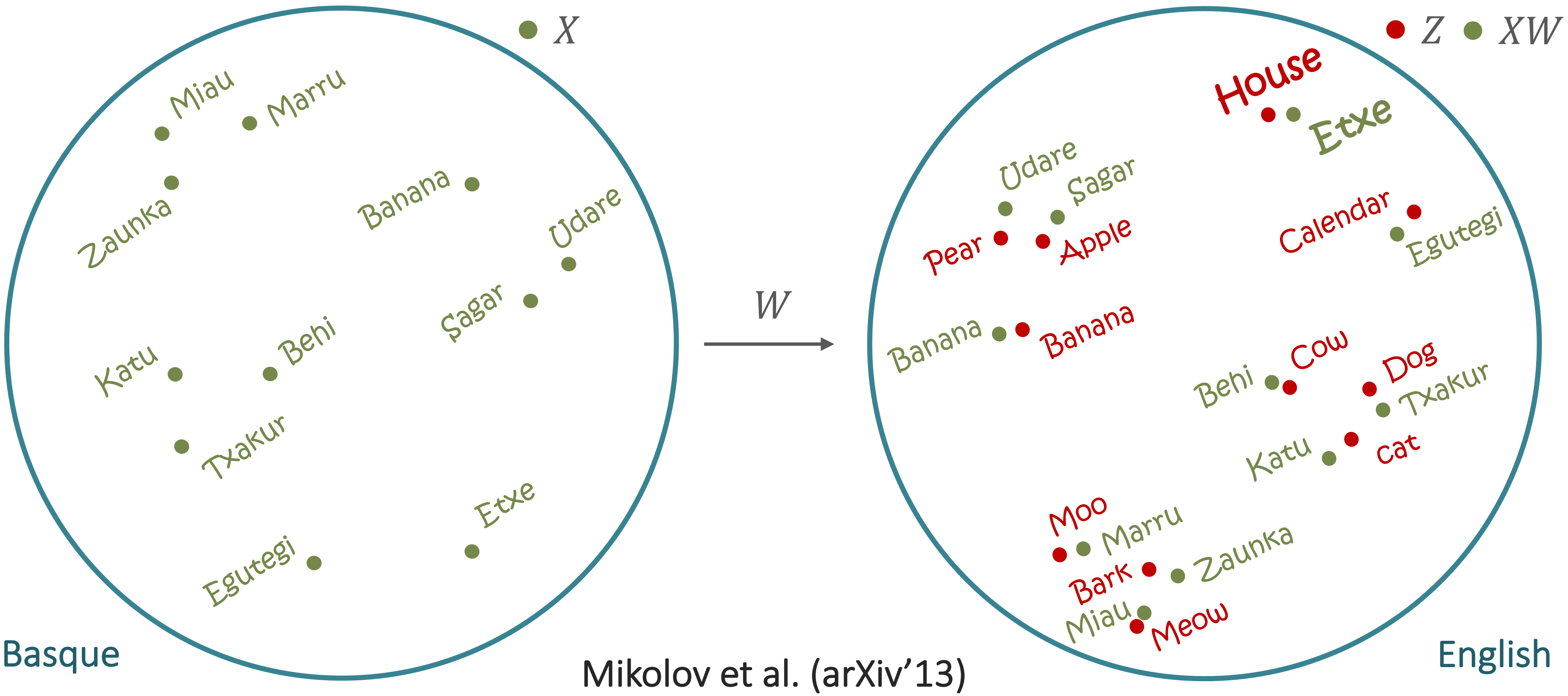
Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

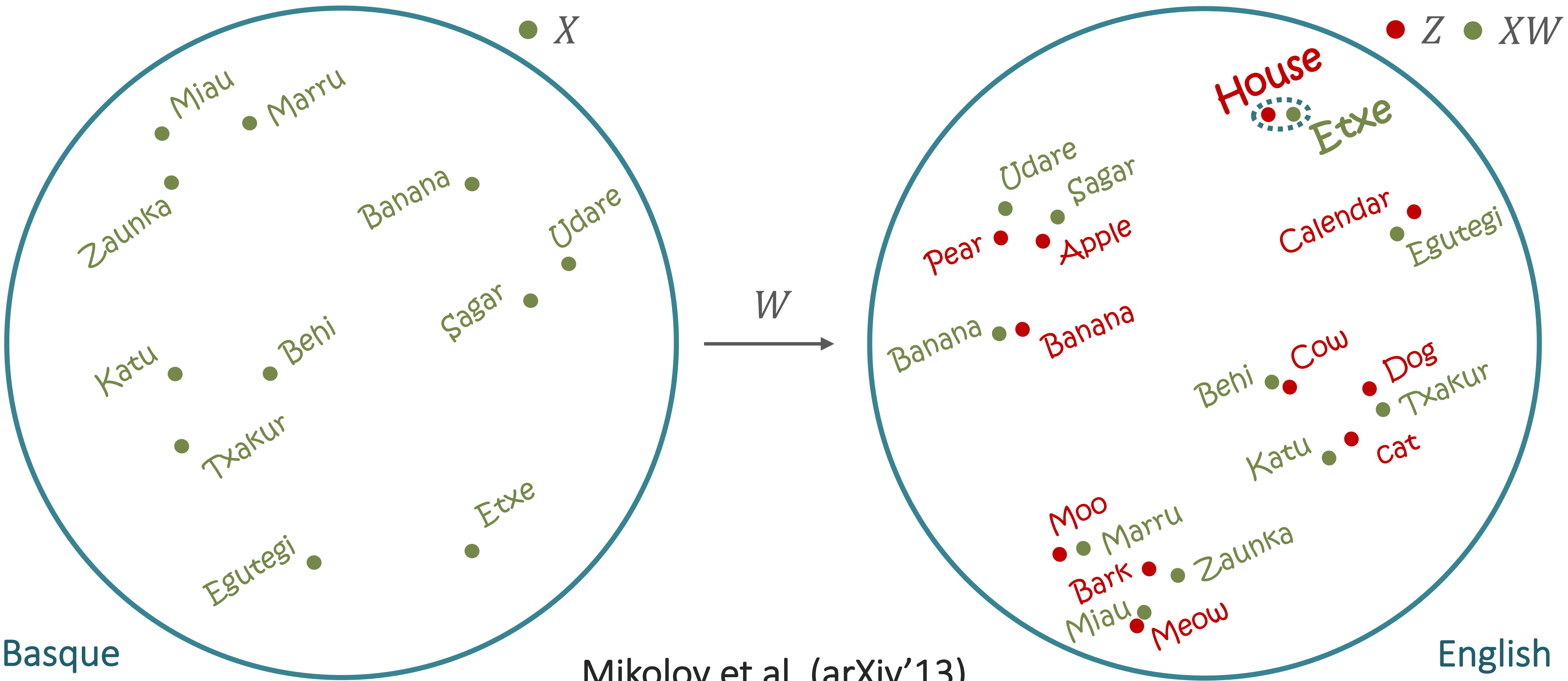
Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

Basque	English
Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

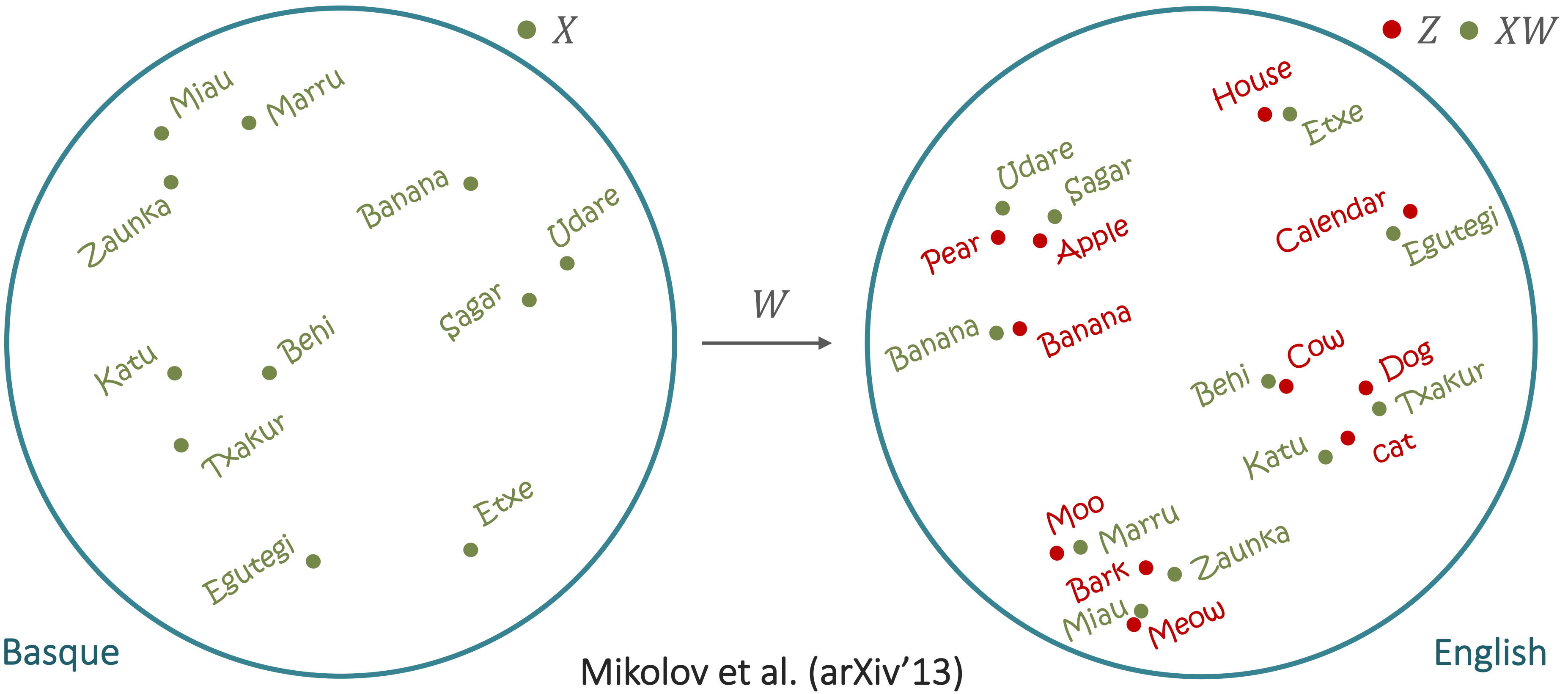
Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

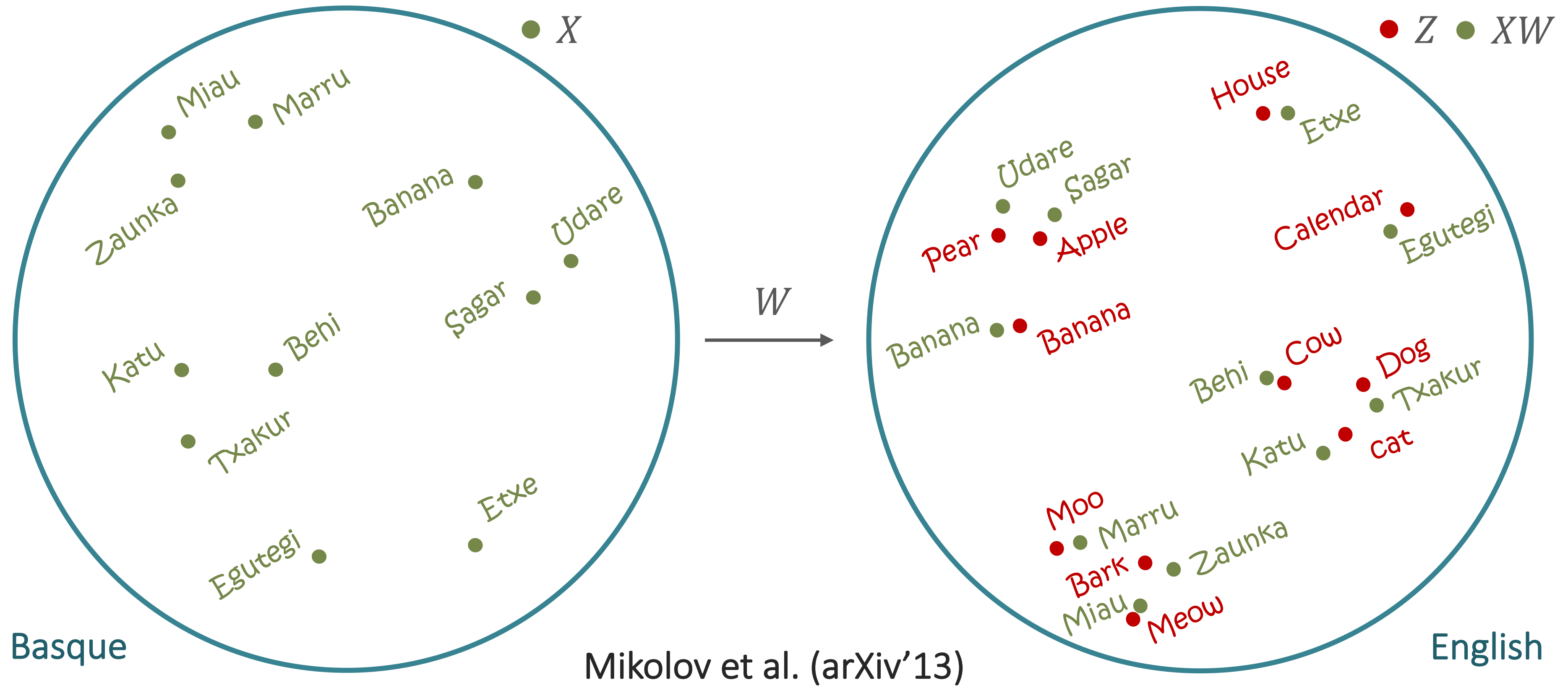
Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

Txakur	Dog
Sagar	Apple
⋮	⋮
Egutegi	Calendar

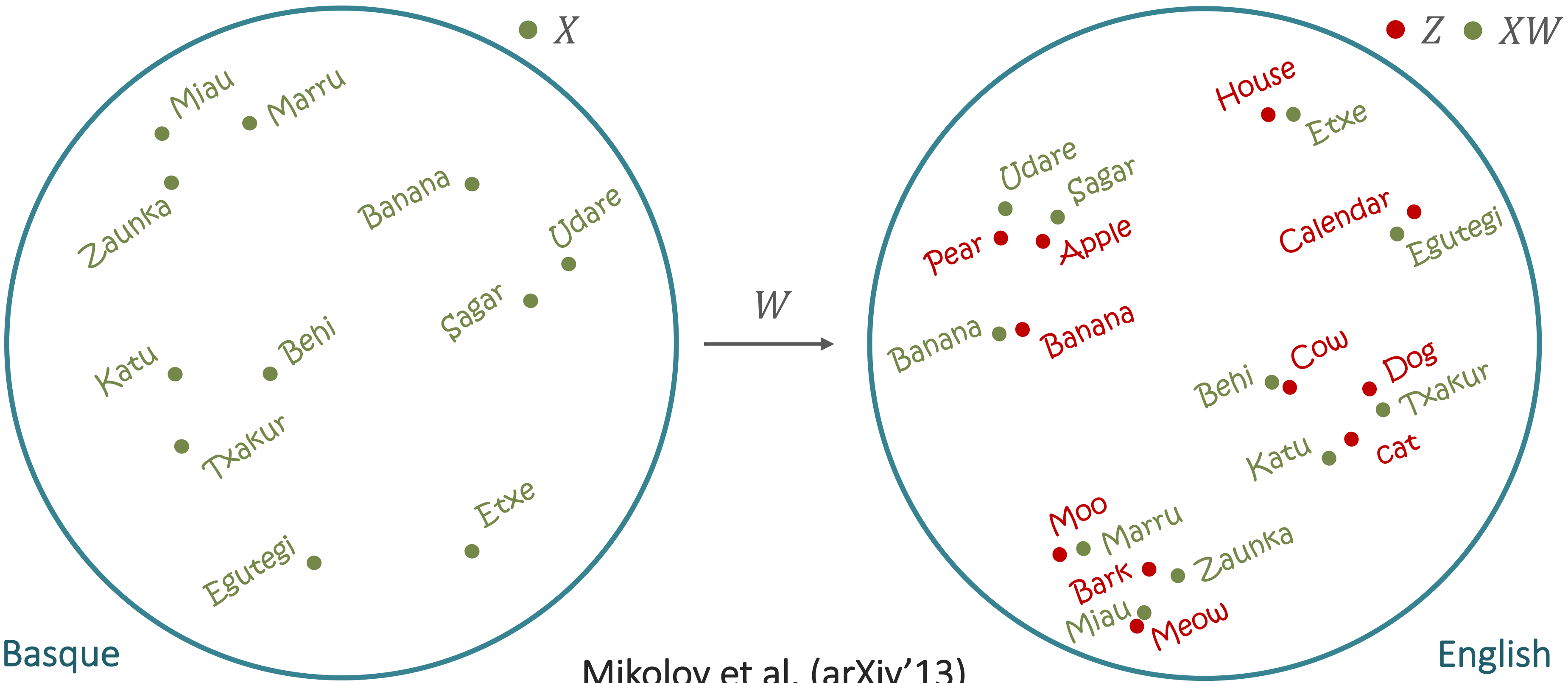
Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

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

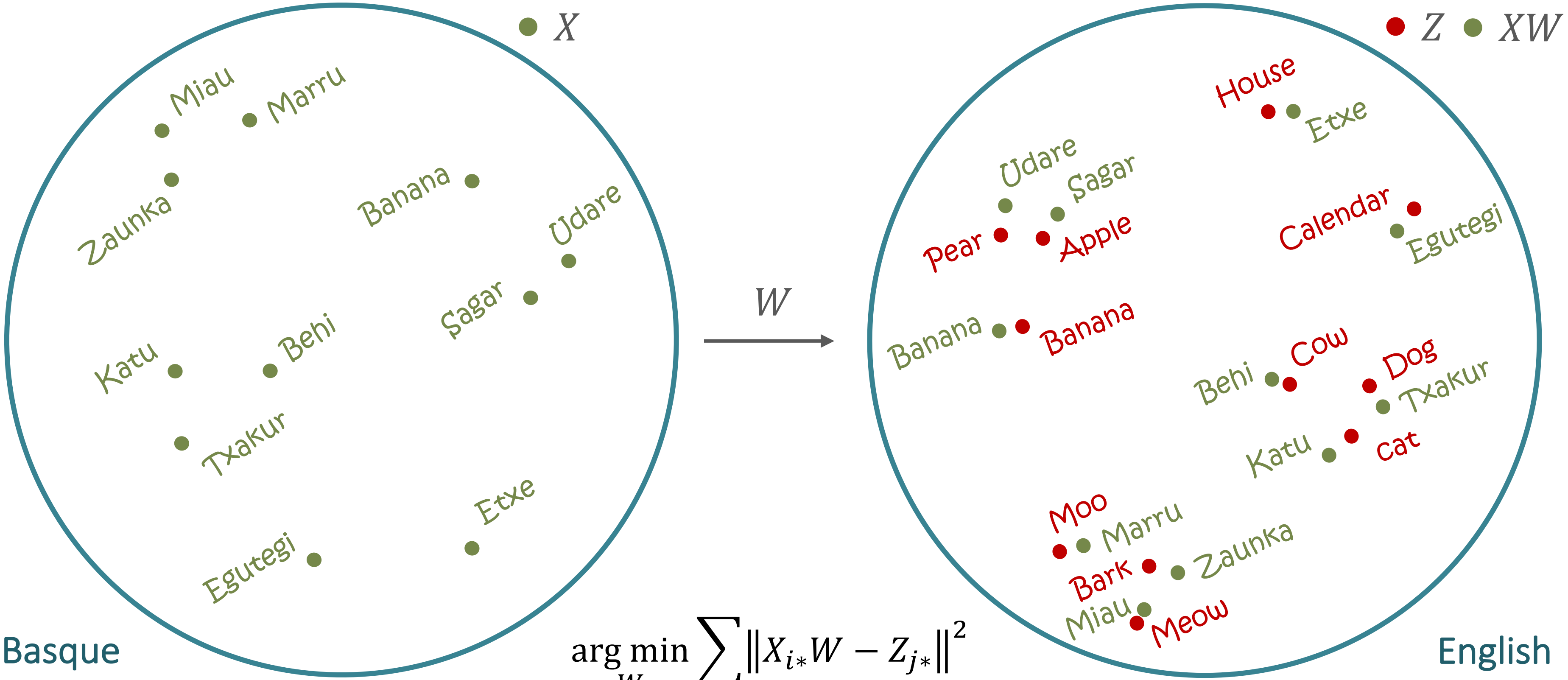
Cross-lingual word embedding alignment



Mikolov et al. (arXiv'13)

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

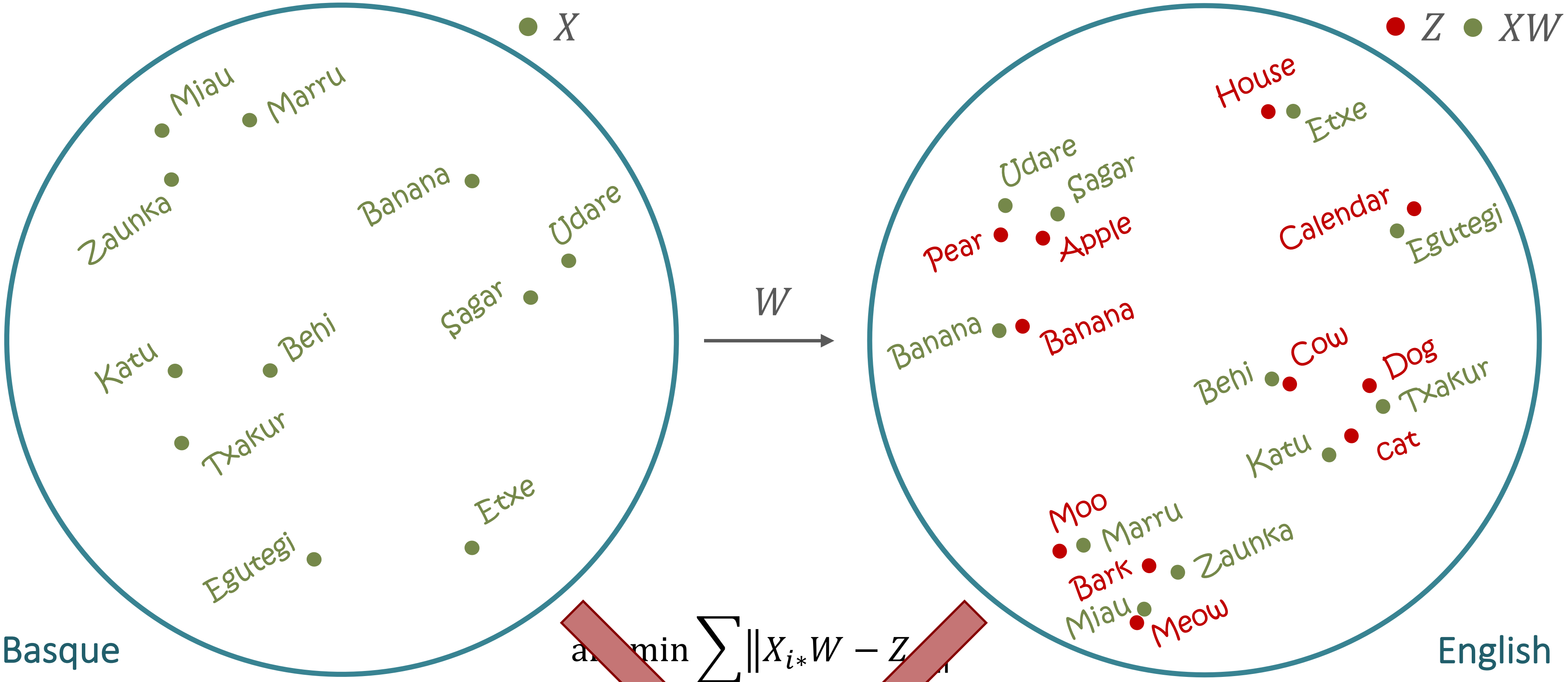
Cross-lingual word embedding alignment



$$\arg\min_W \sum_i \left\| X_{i*} W - Z_{j*} \right\|^2$$

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

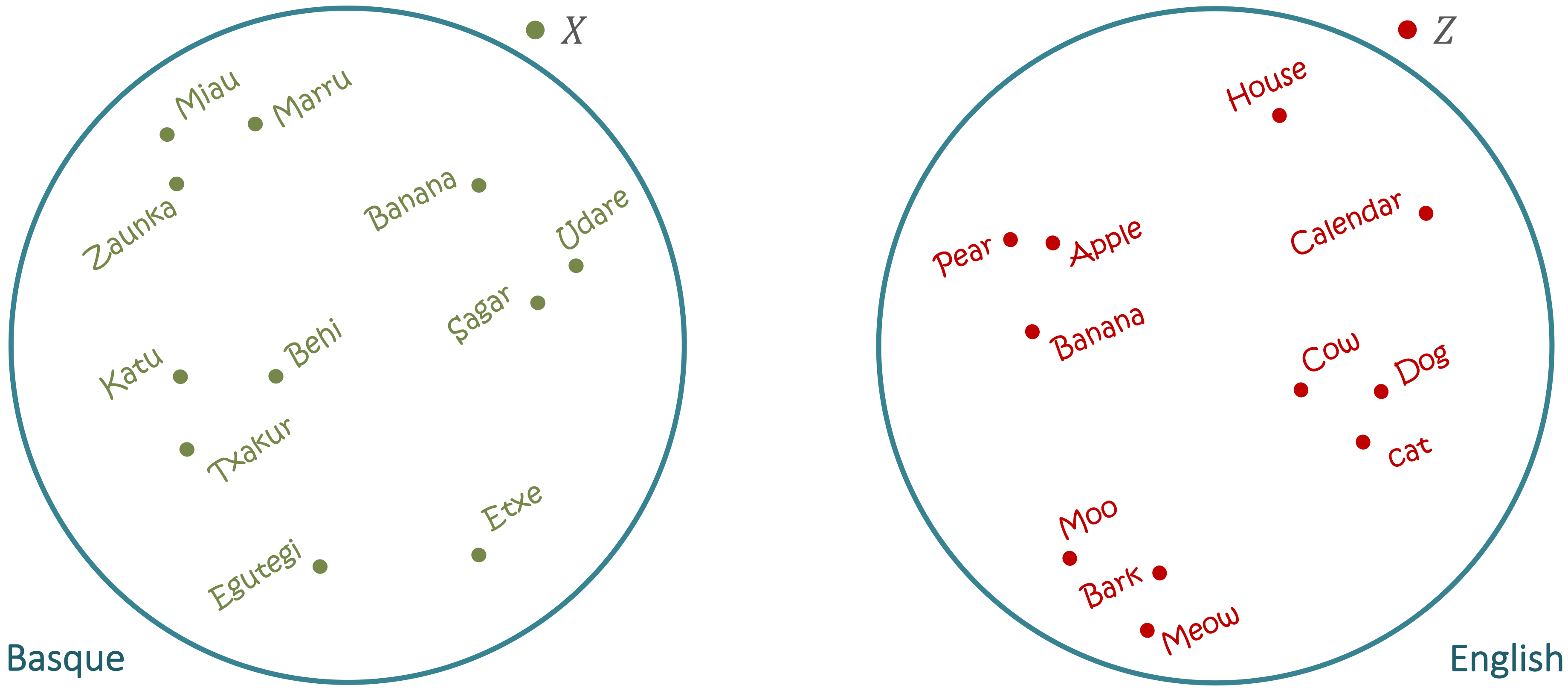
Cross-lingual word embedding alignment



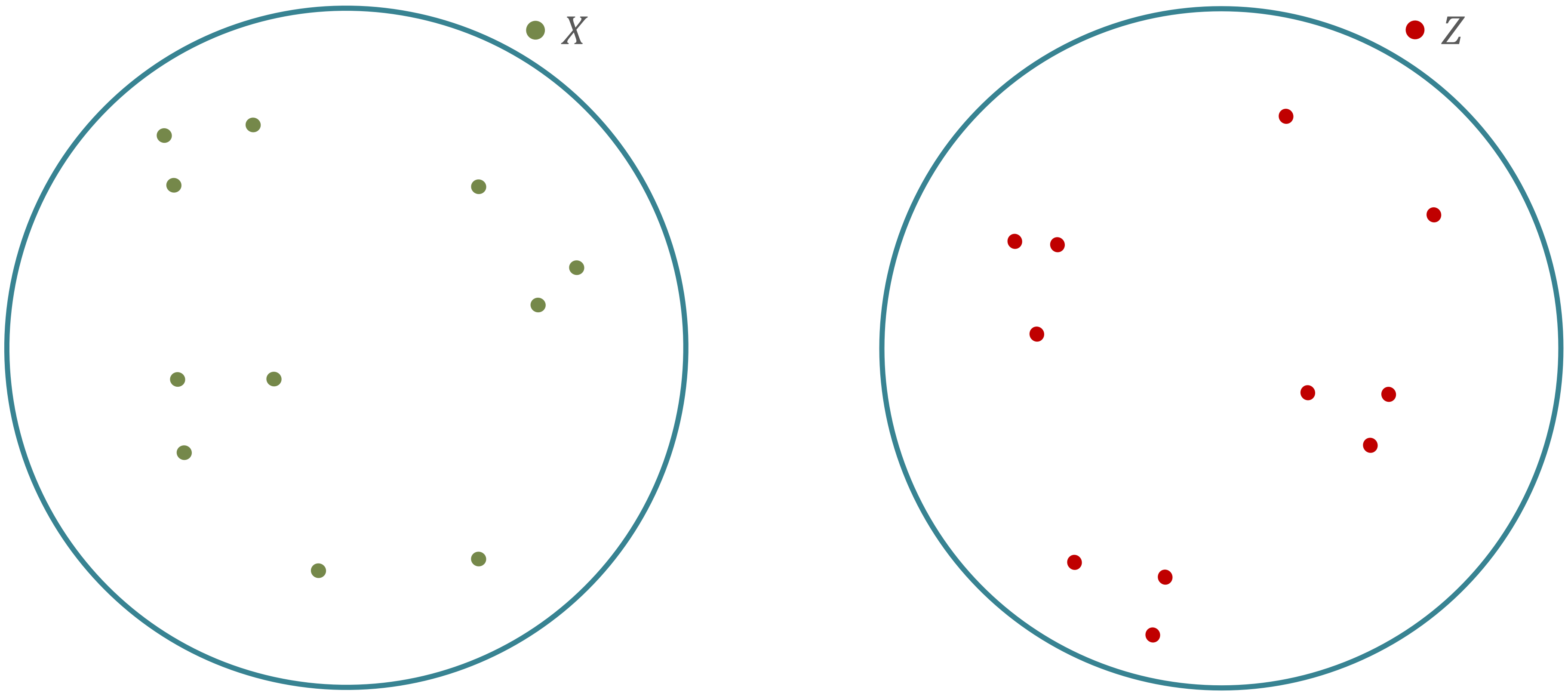
$$\arg\min_{W} \sum_i \|X_{i*}W - Z_{i*}\|$$

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

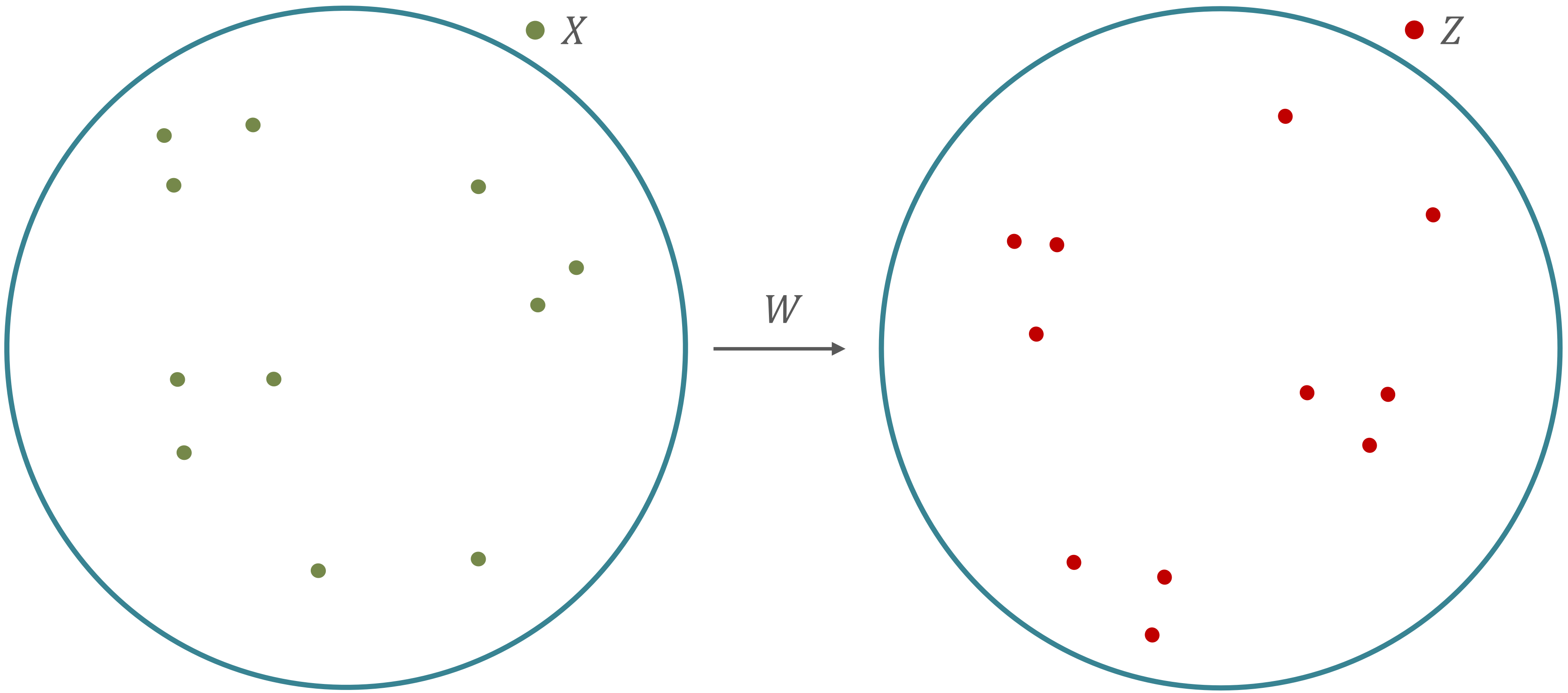
Cross-lingual word embedding alignment



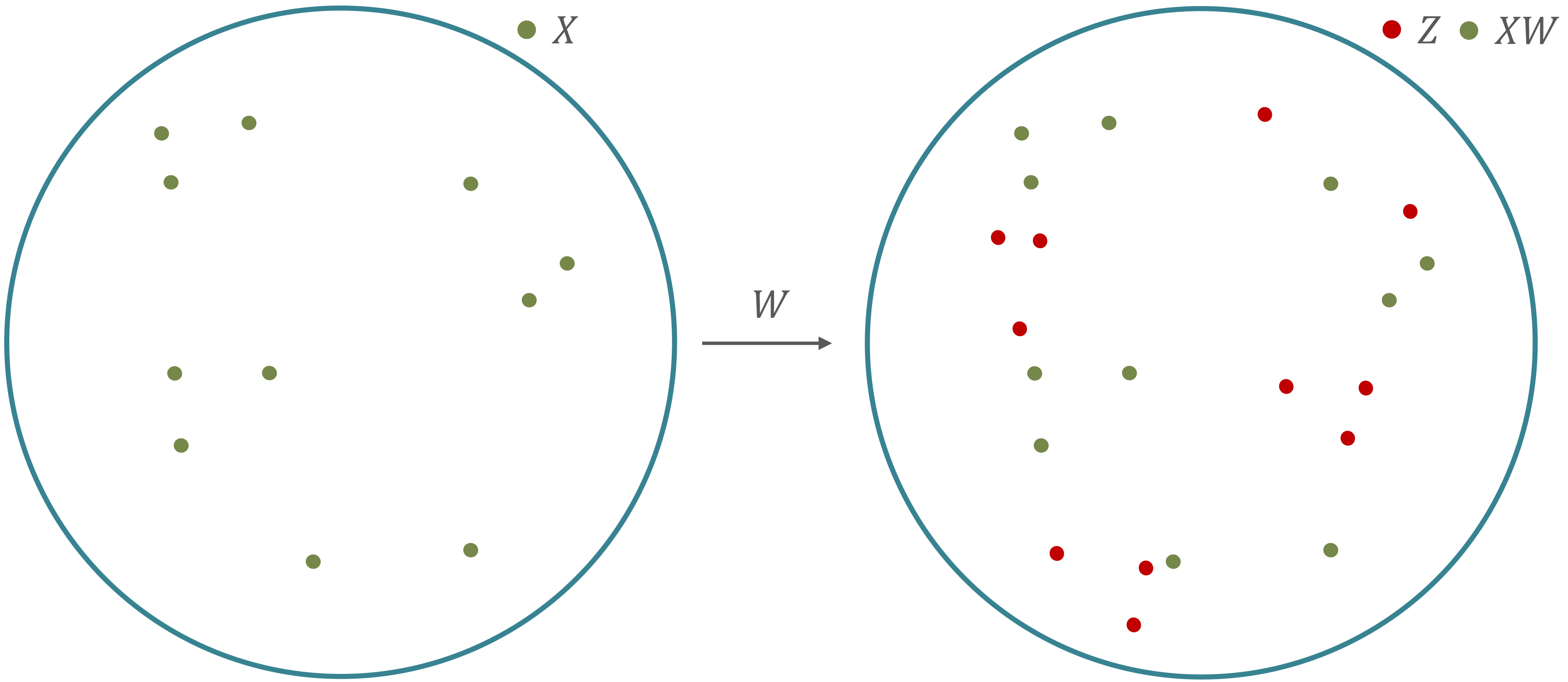
Cross-lingual word embedding alignment



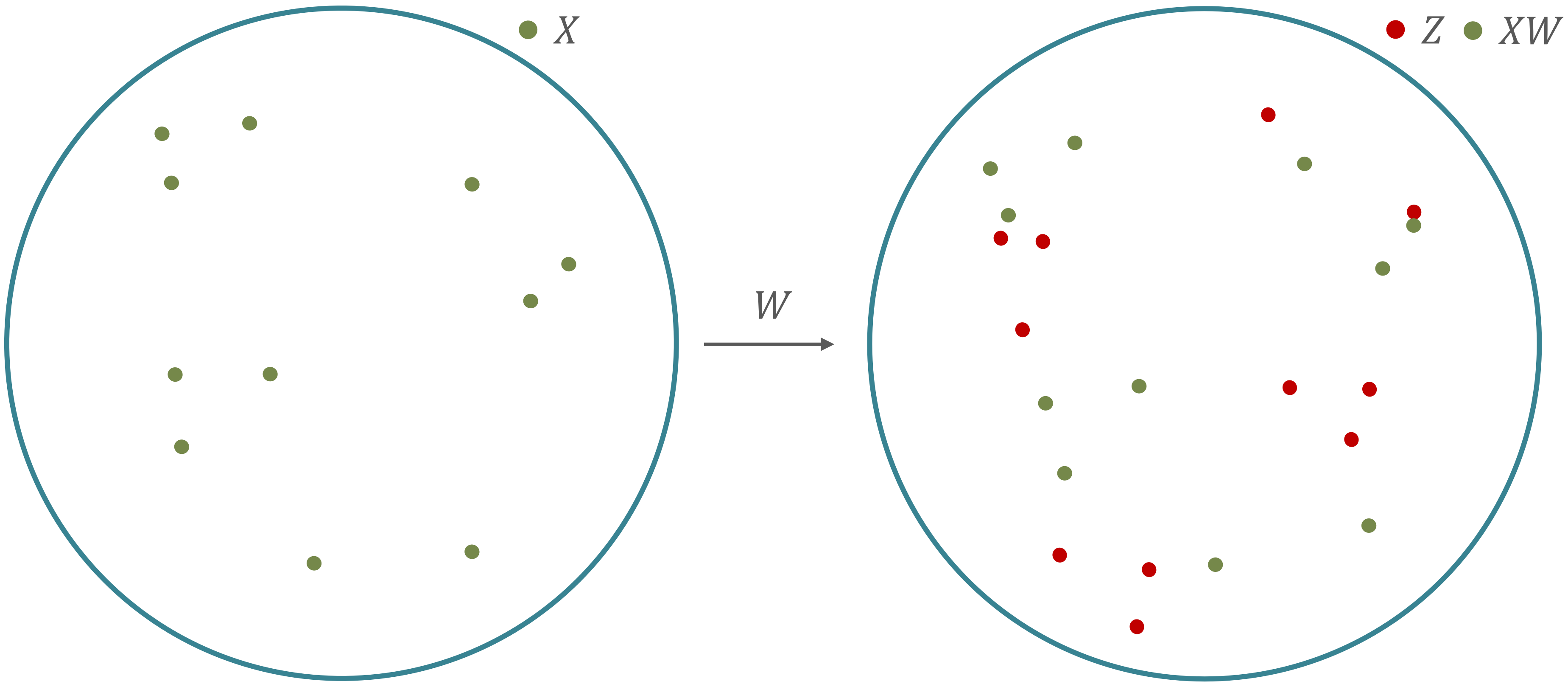
Cross-lingual word embedding alignment



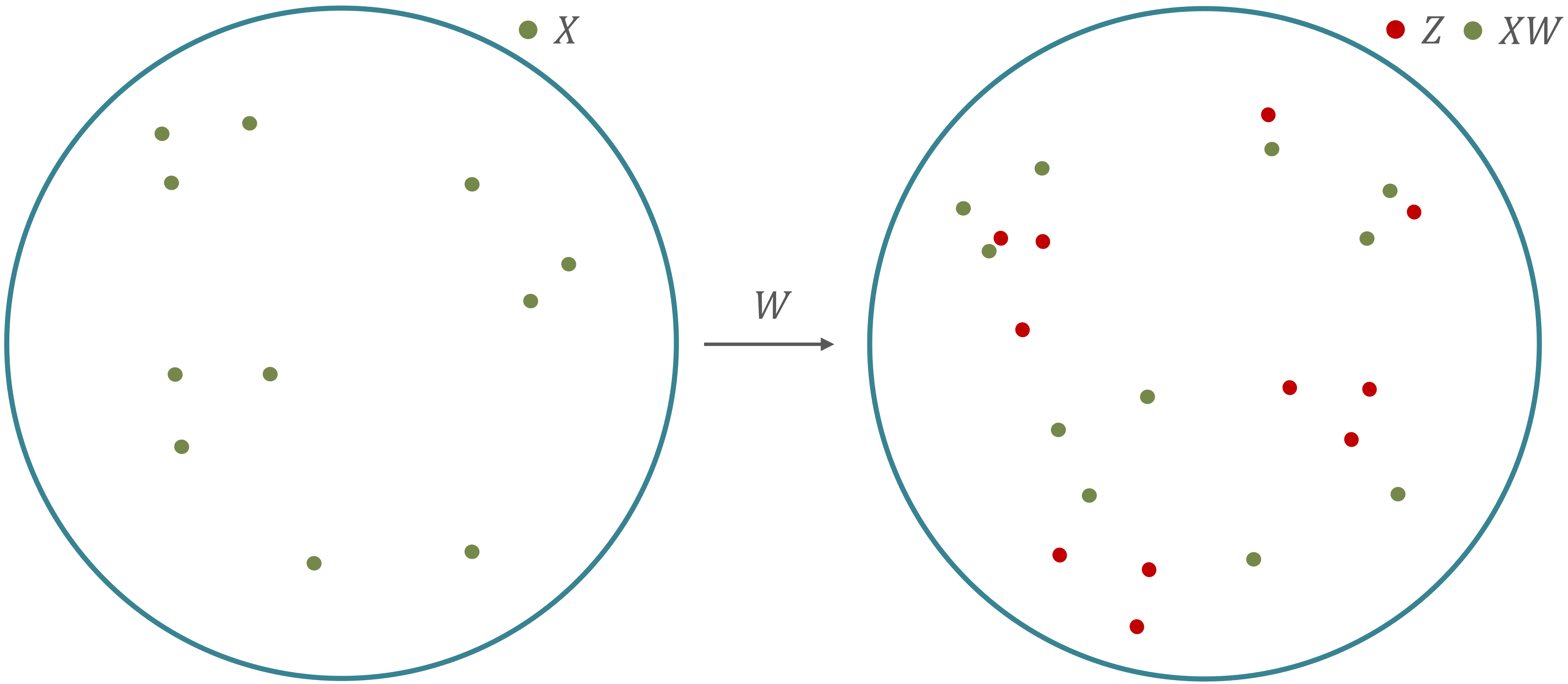
Cross-lingual word embedding alignment



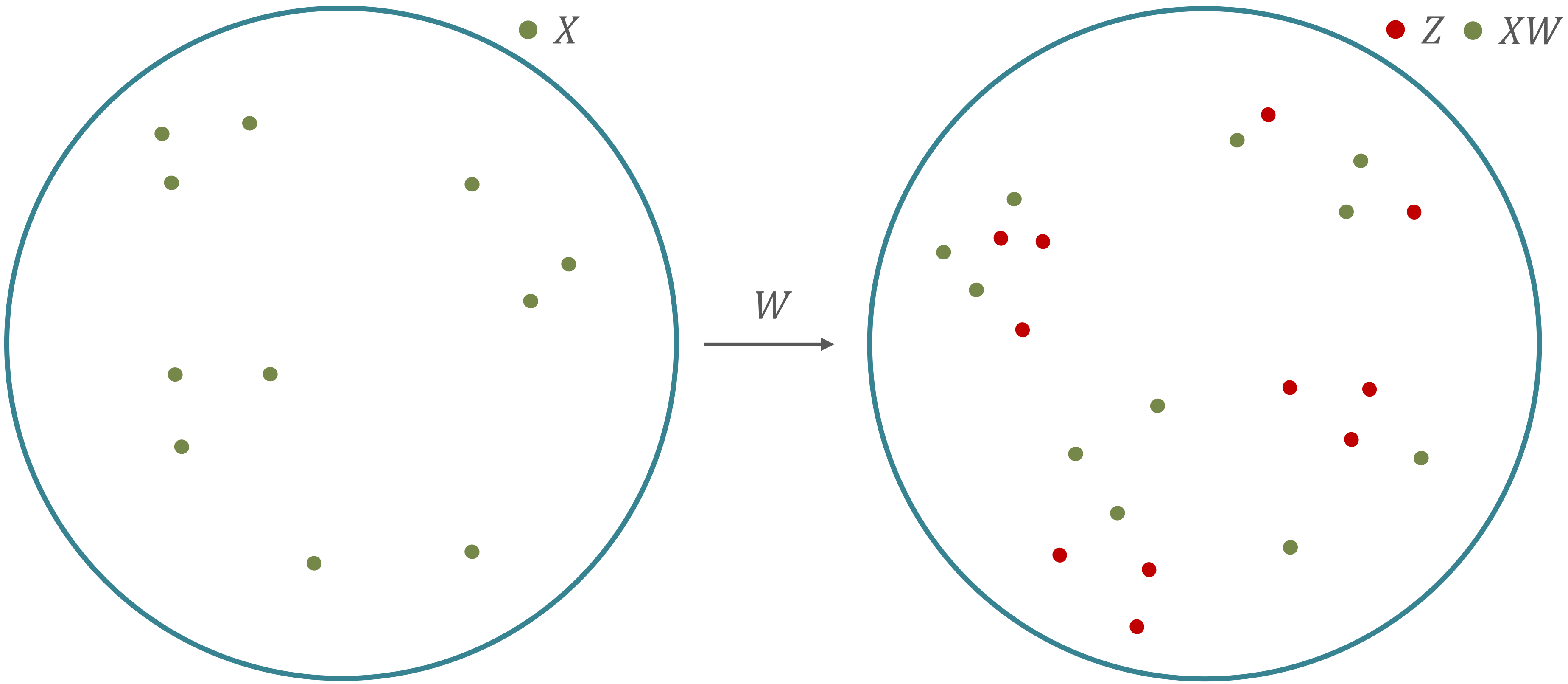
Cross-lingual word embedding alignment



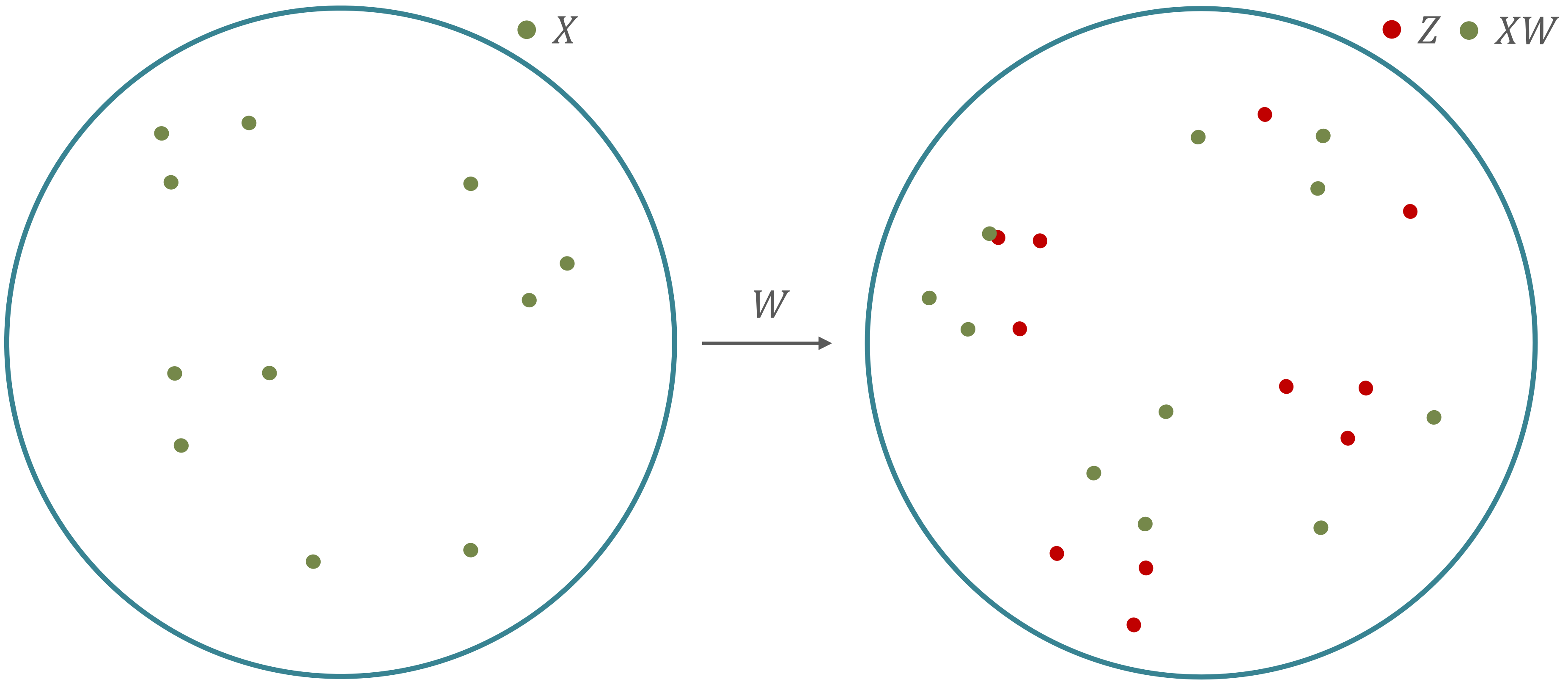
Cross-lingual word embedding alignment



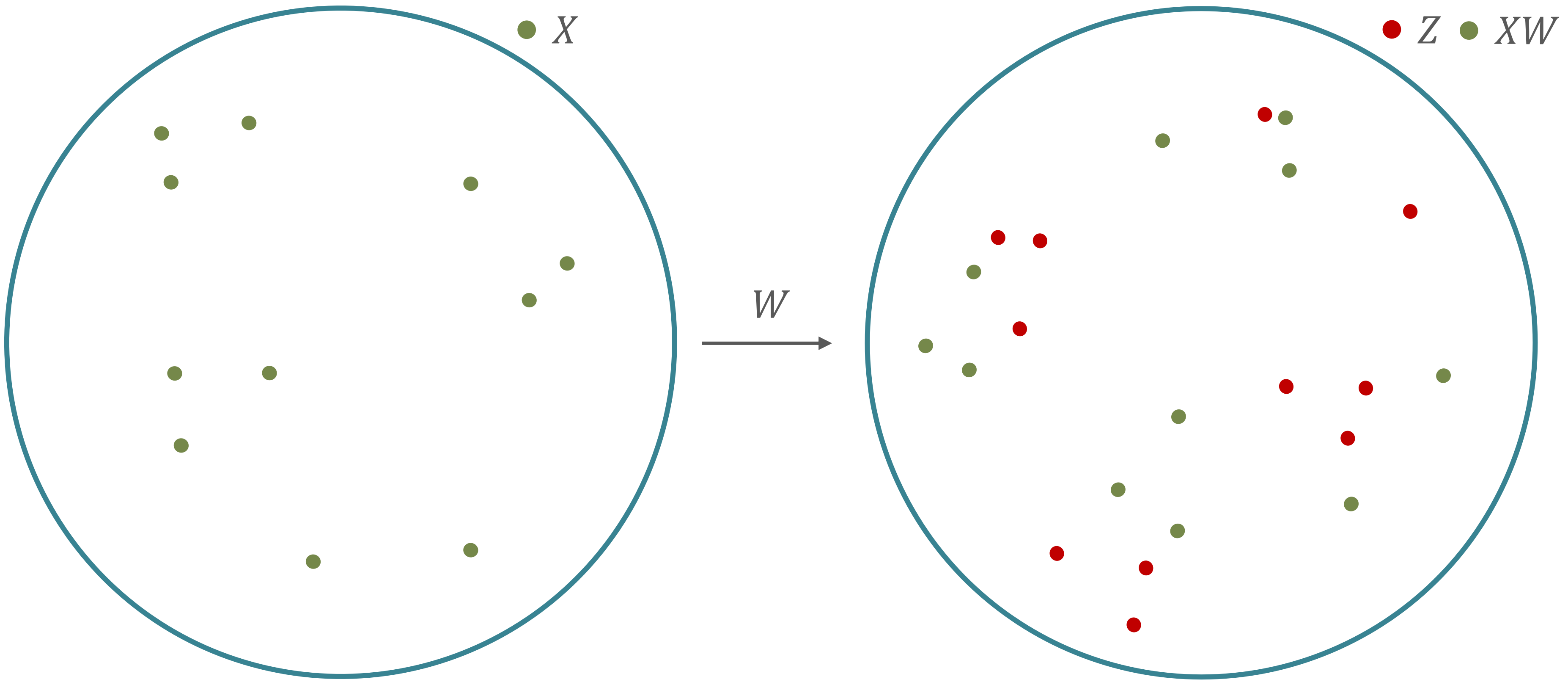
Cross-lingual word embedding alignment



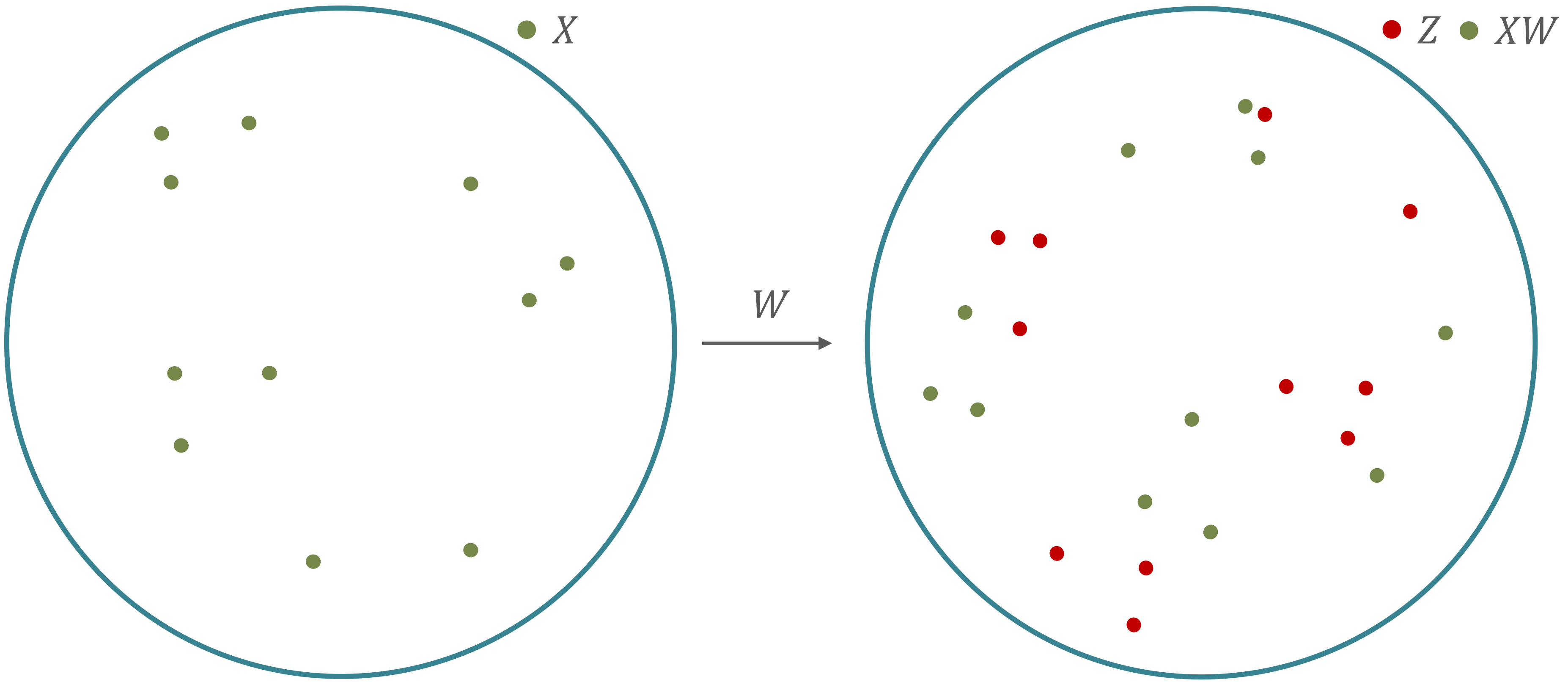
Cross-lingual word embedding alignment



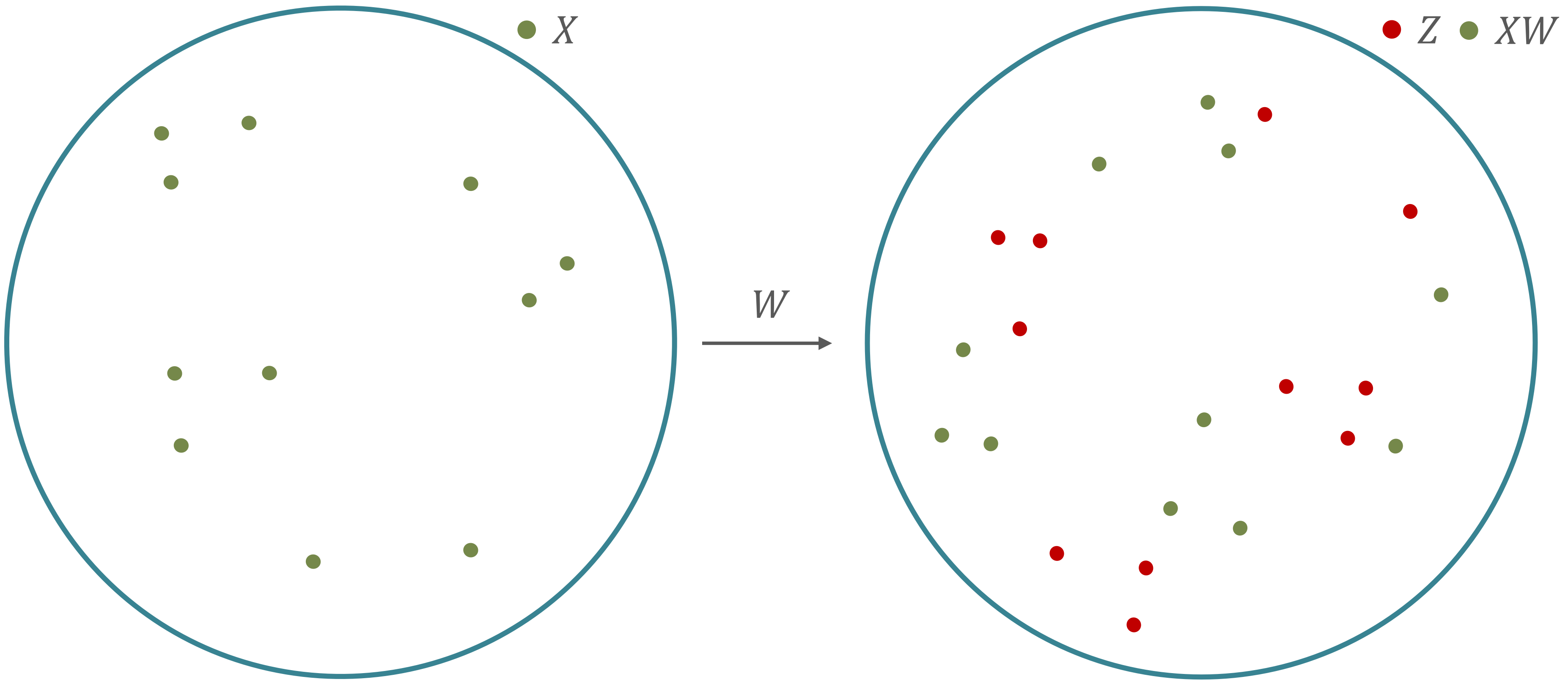
Cross-lingual word embedding alignment



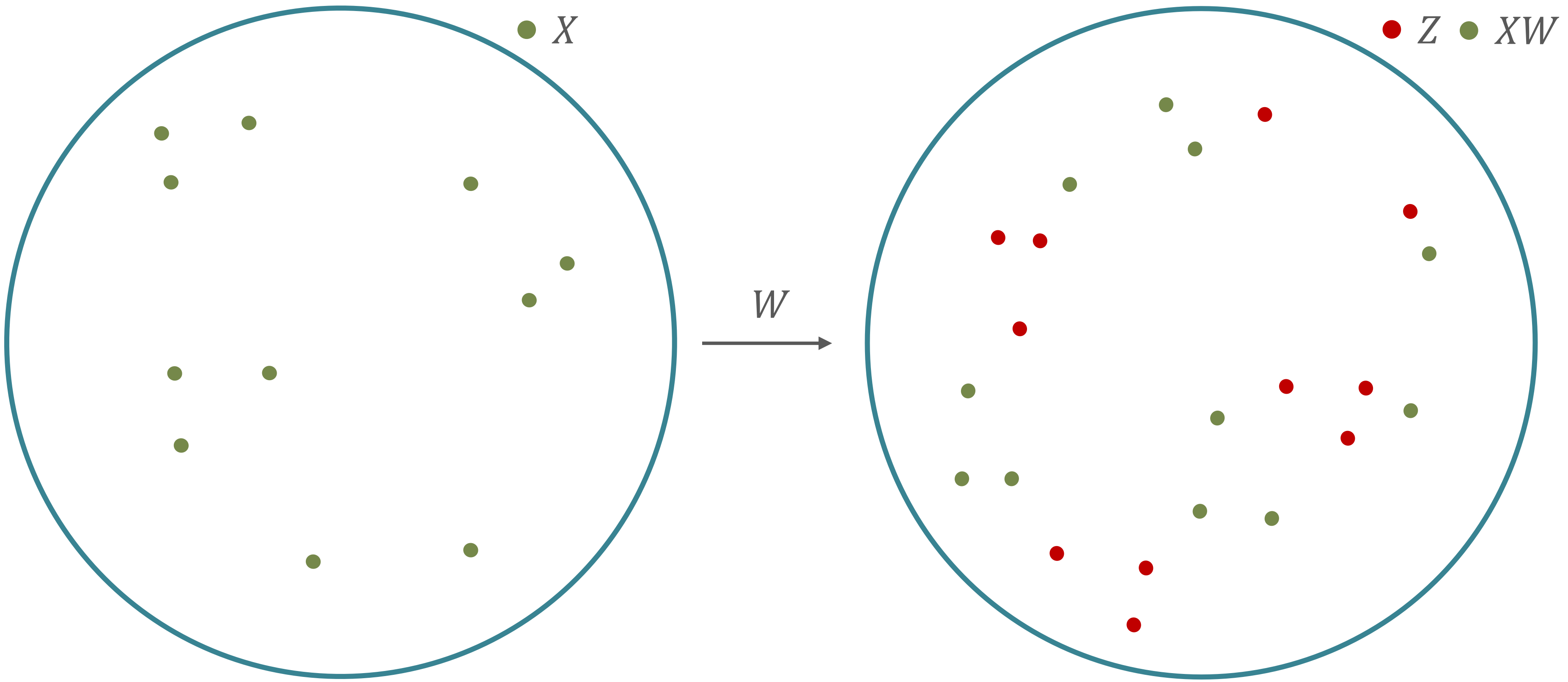
Cross-lingual word embedding alignment



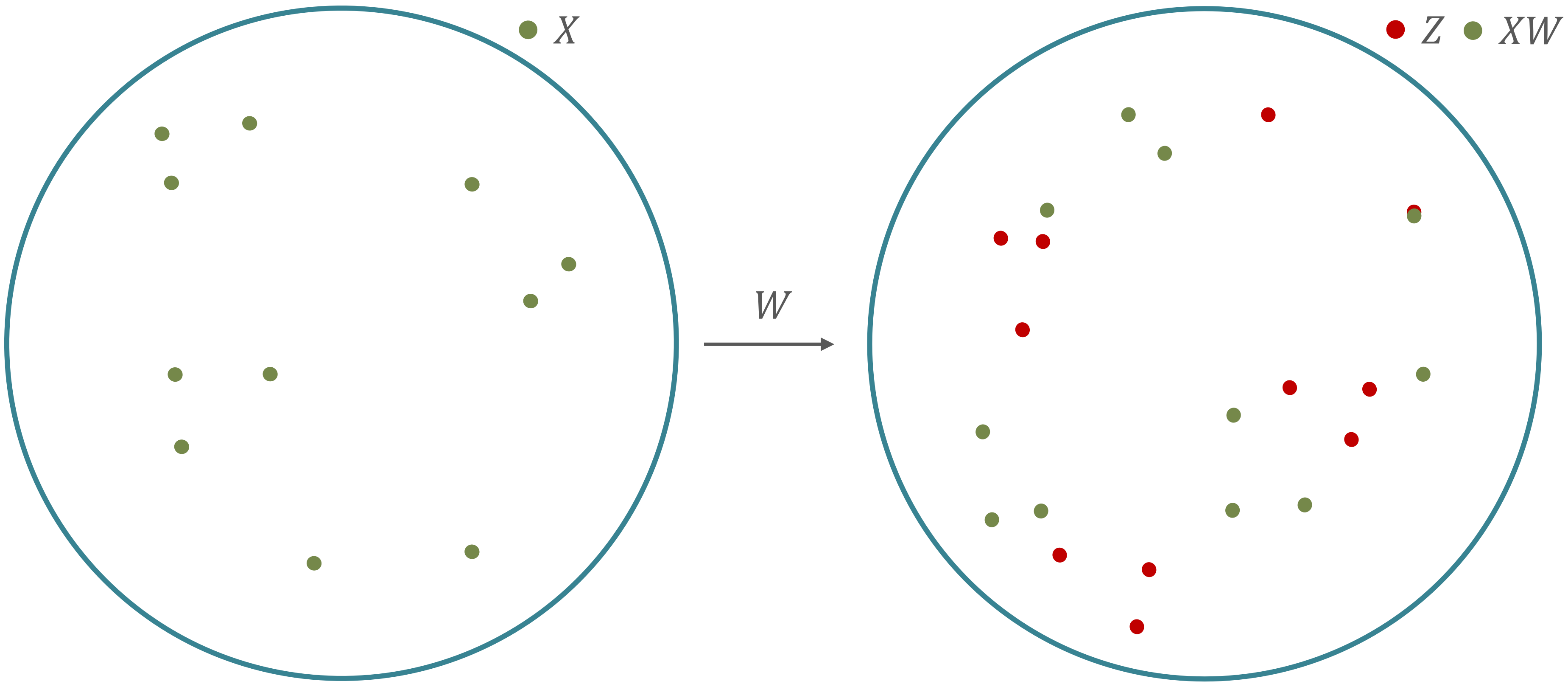
Cross-lingual word embedding alignment



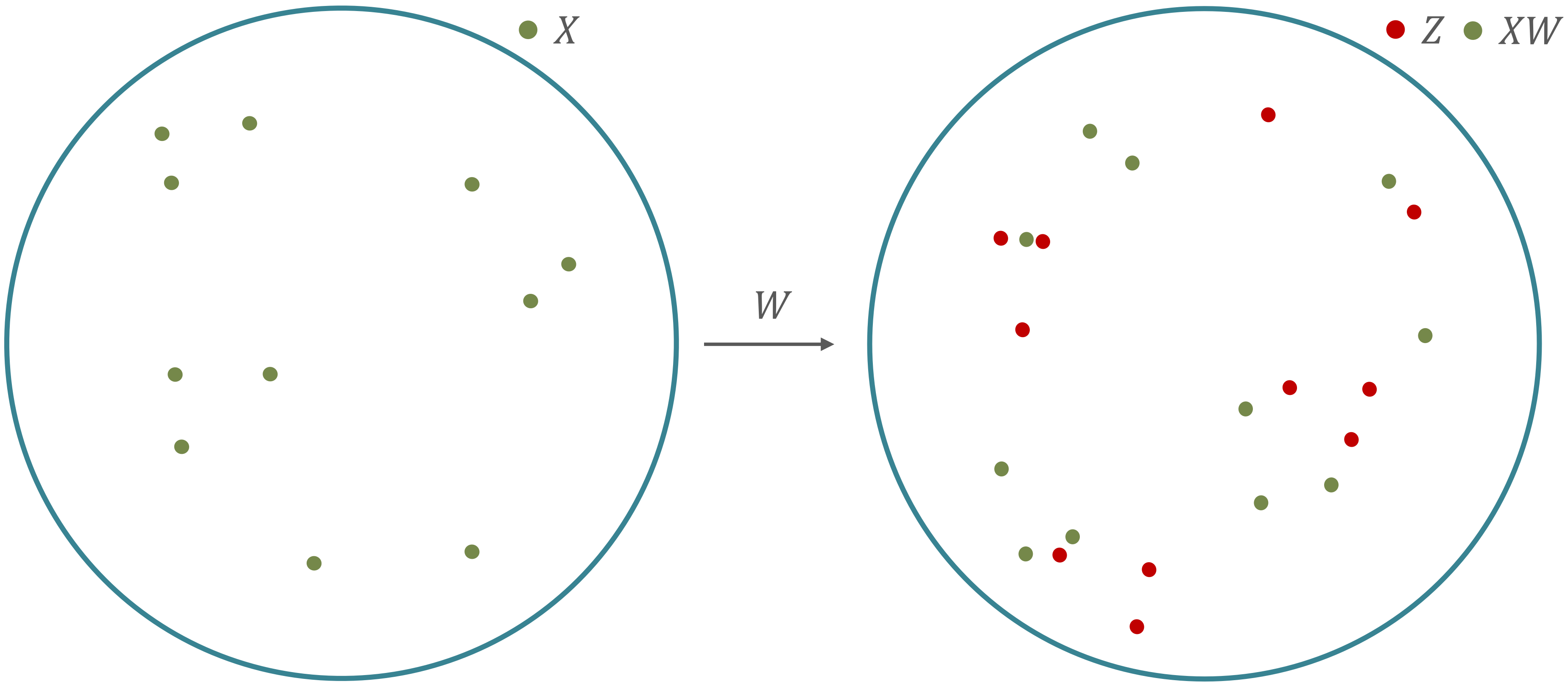
Cross-lingual word embedding alignment



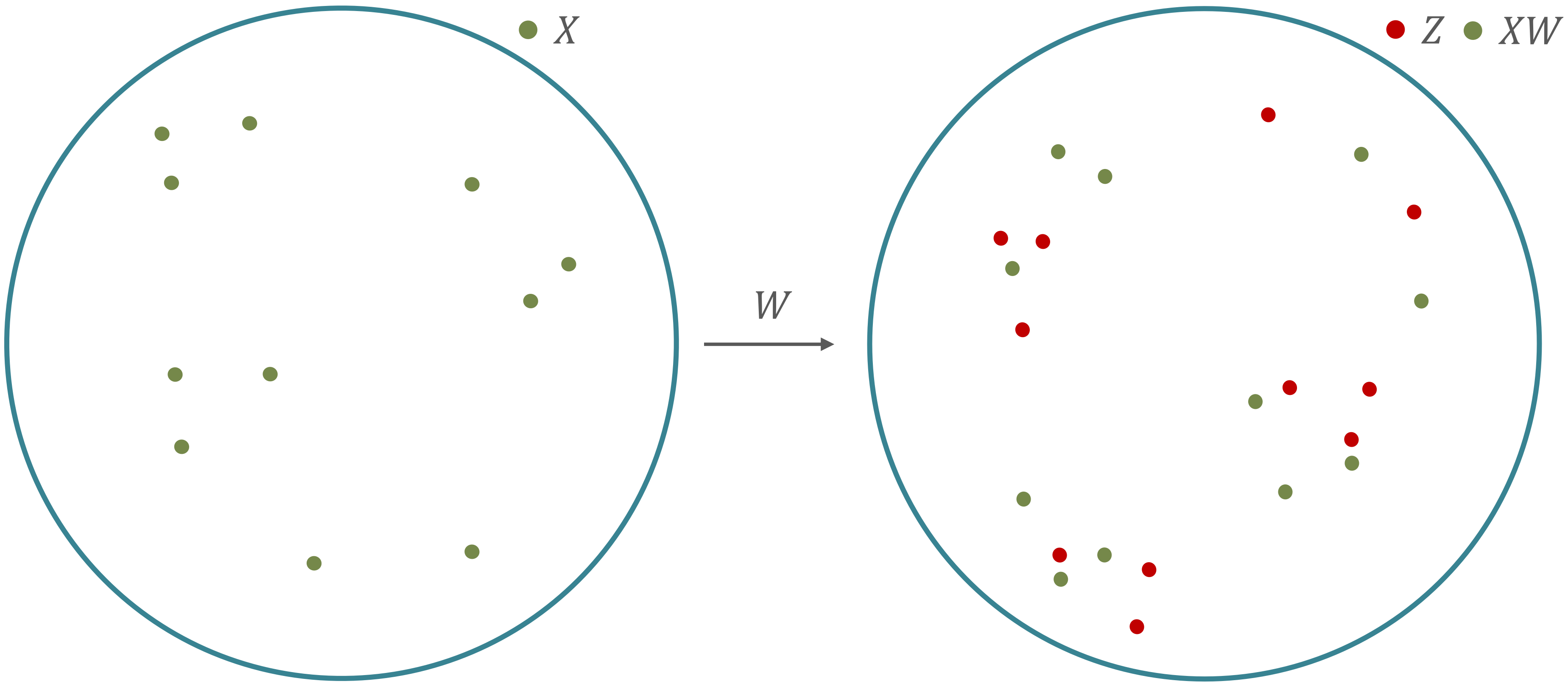
Cross-lingual word embedding alignment



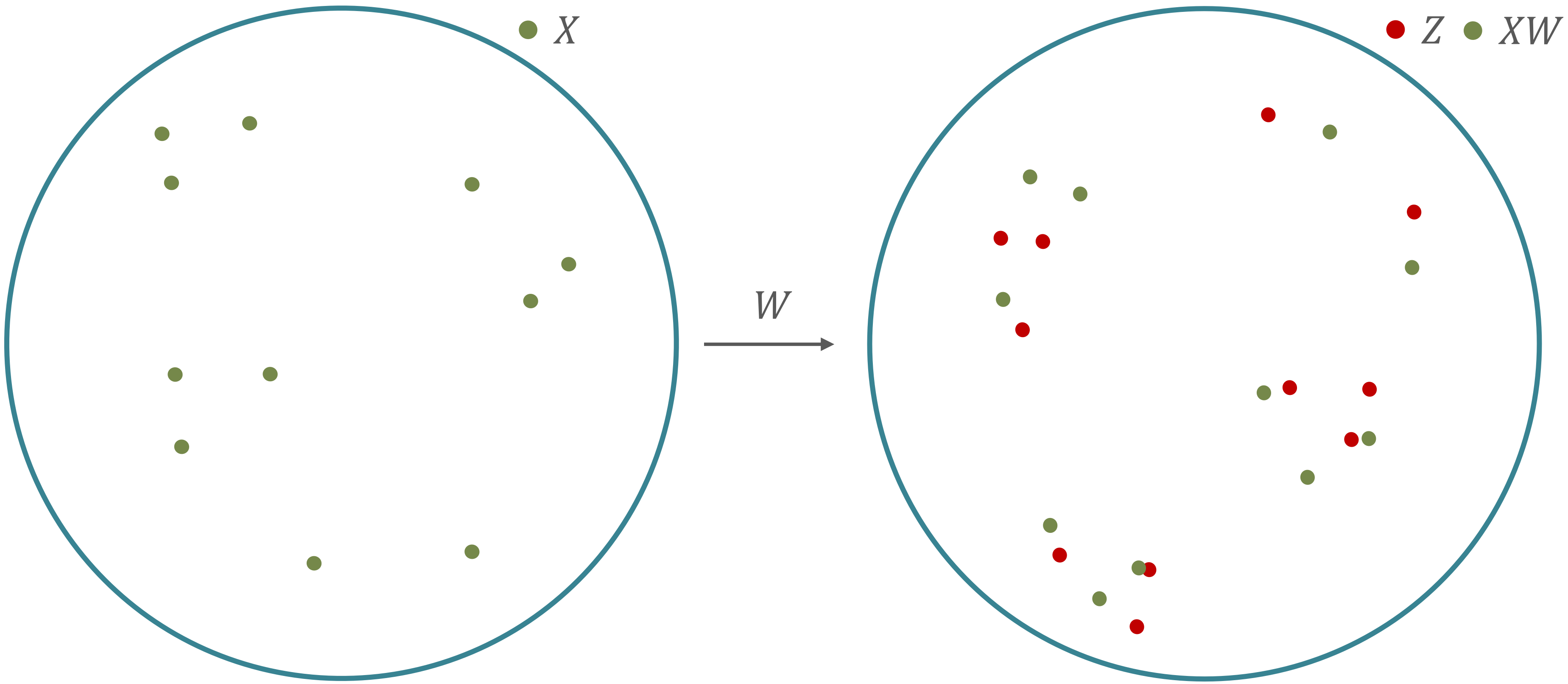
Cross-lingual word embedding alignment



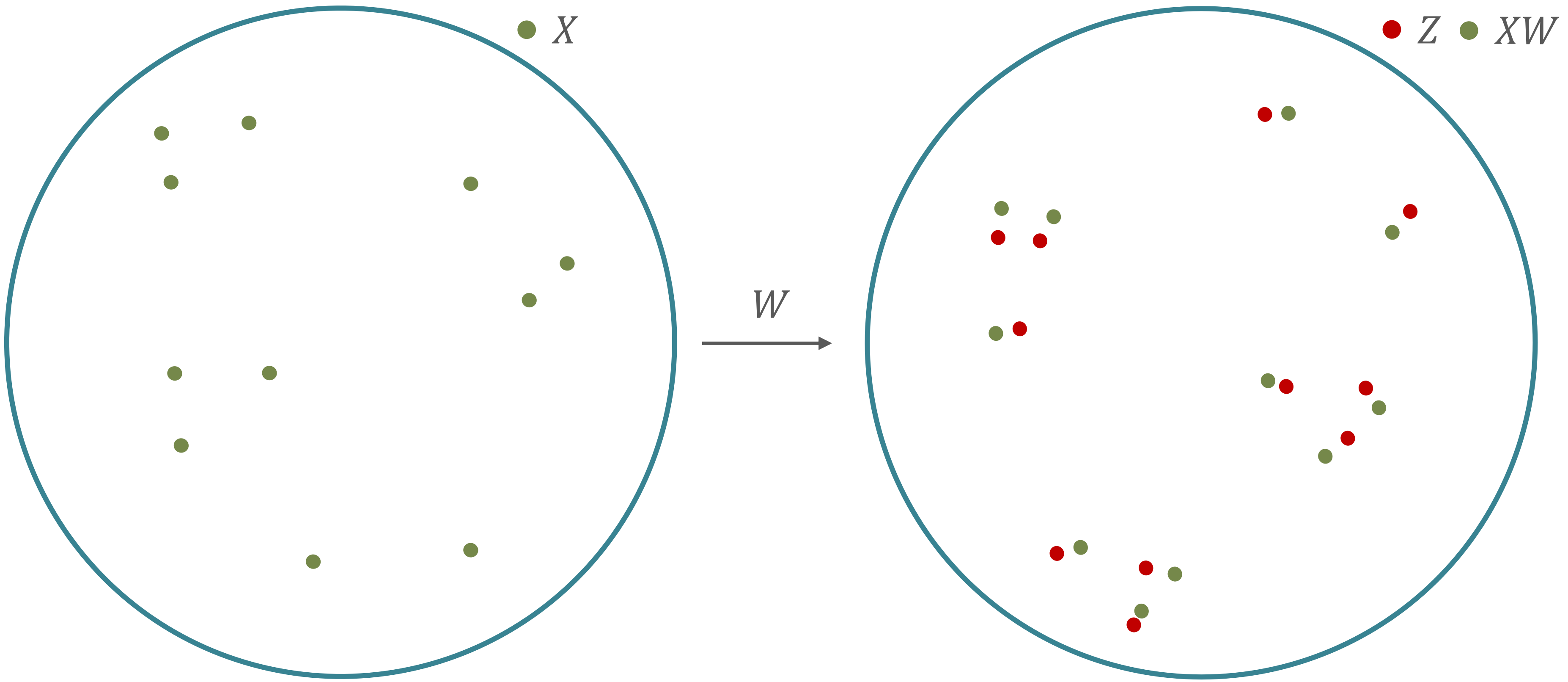
Cross-lingual word embedding alignment



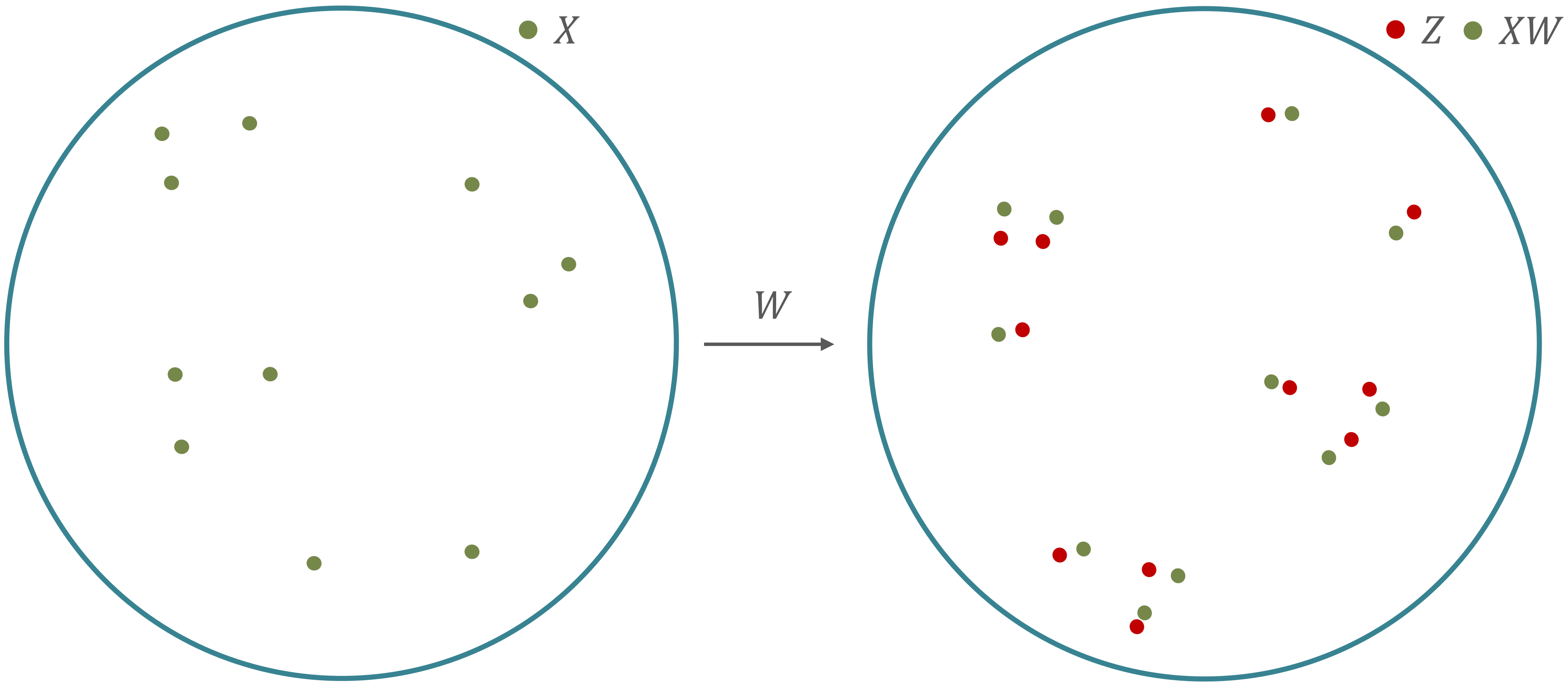
Cross-lingual word embedding alignment



Cross-lingual word embedding alignment



Cross-lingual word embedding alignment



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

Cross-lingual word embedding alignment

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

Cross-lingual word embedding alignment

Self-learning

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

Cross-lingual word embedding alignment

Self-learning

Dictionary

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

Cross-lingual word embedding alignment

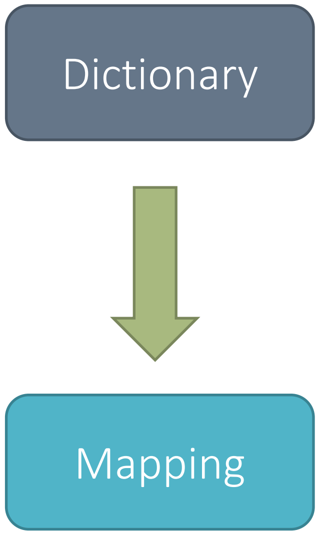
Self-learning

Dictionary

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

Cross-lingual word embedding alignment

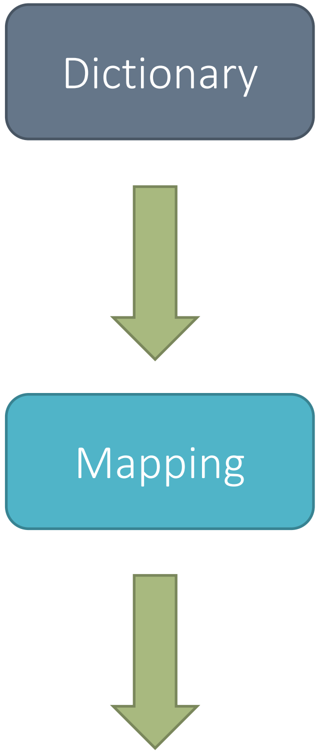
Self-learning



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

Cross-lingual word embedding alignment

Self-learning

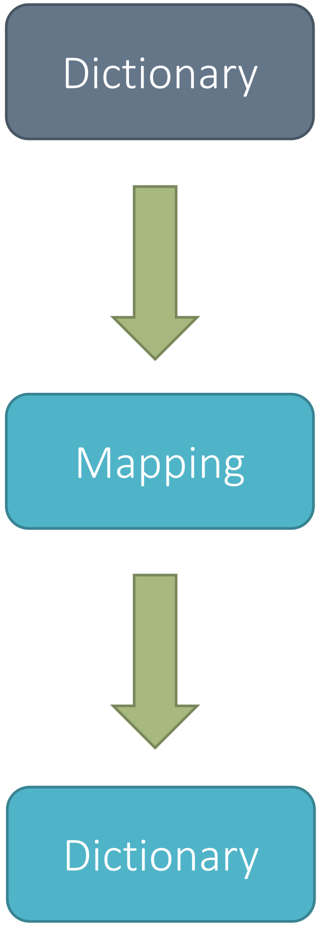


$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual word embedding alignment

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

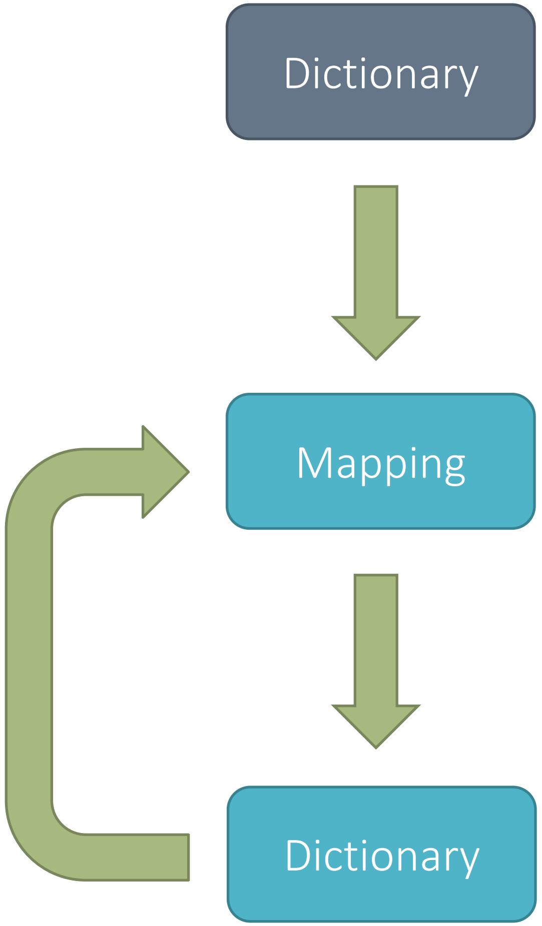
Self-learning



Cross-lingual word embedding alignment

<u>Self-learning</u>

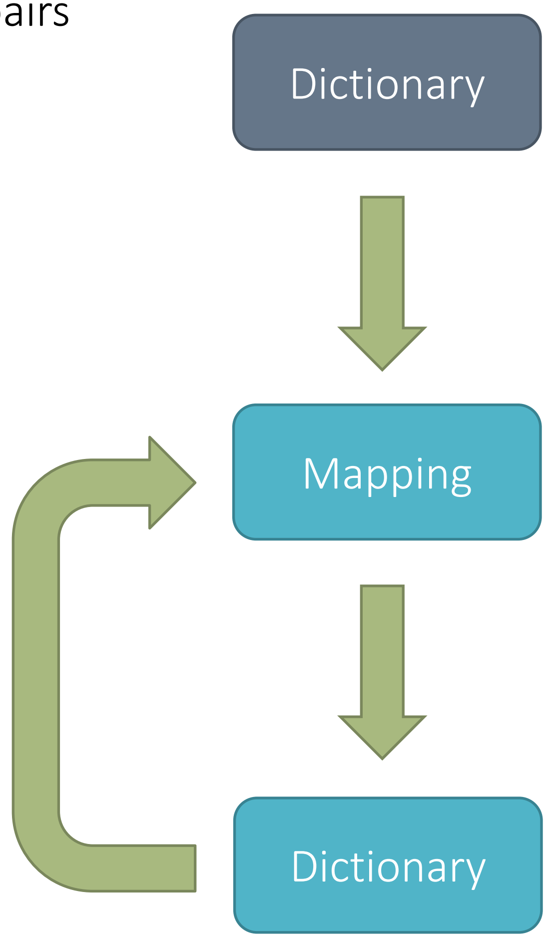
$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \|X_{i*}W - Z_{j*}\|^2$$



Cross-lingual word embedding alignment

Self-learning

- 25 word pairs

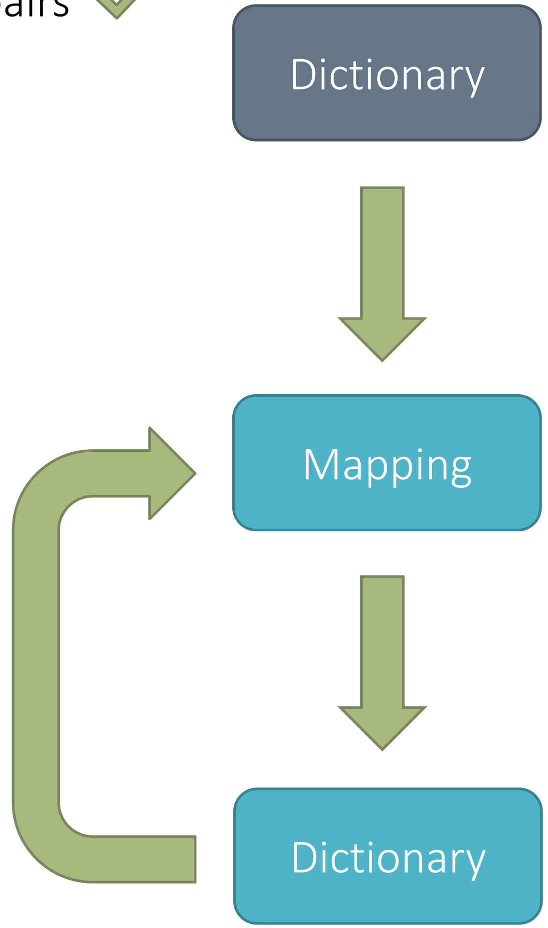


$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \left\| X_{i*} W - Z_{j*} \right\|^2$$

Cross-lingual word embedding alignment

Self-learning

- 25 word pairs ✓

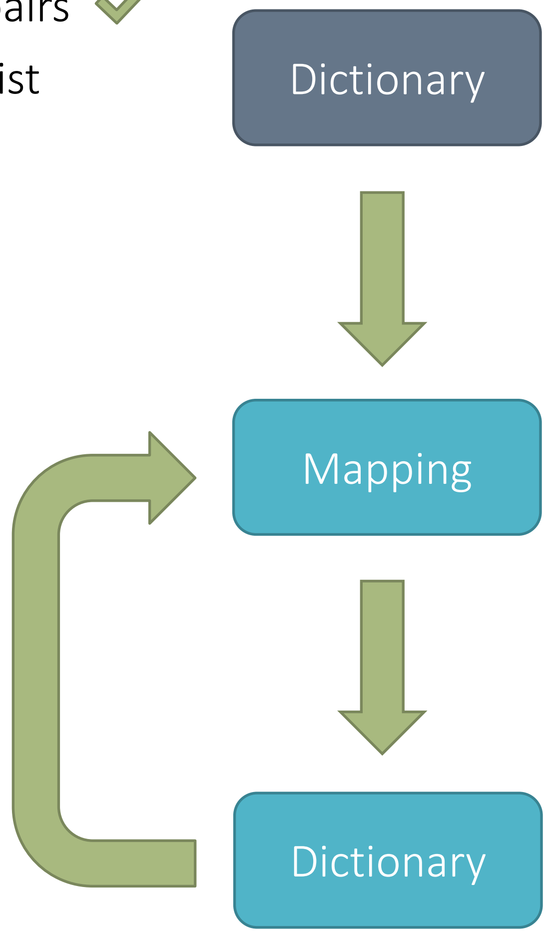


$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \|X_{i*}W - Z_{j*}\|^2$$

Cross-lingual word embedding alignment

Self-learning

- 25 word pairs
- Numeral list

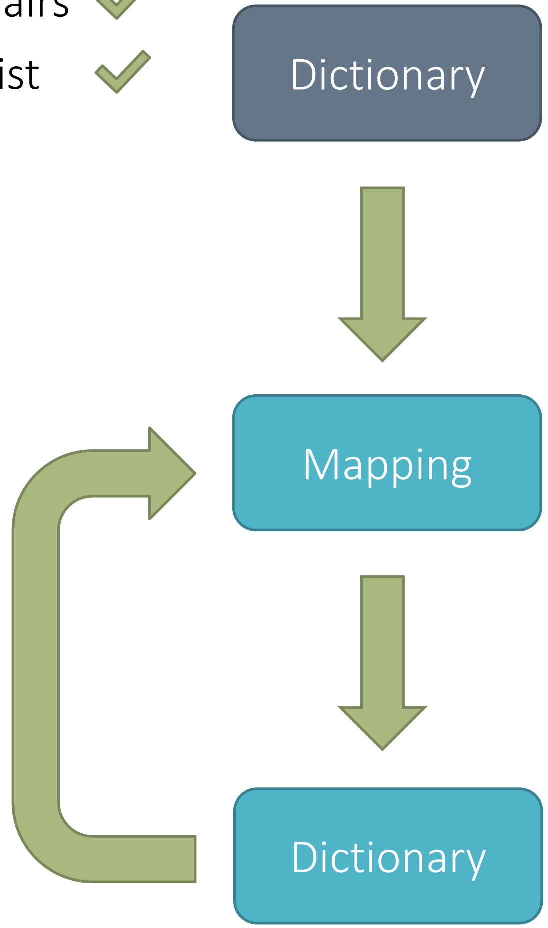


$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \|X_{i*}W - Z_{j*}\|^2$$

Cross-lingual word embedding alignment

Self-learning

- 25 word pairs
- Numeral list

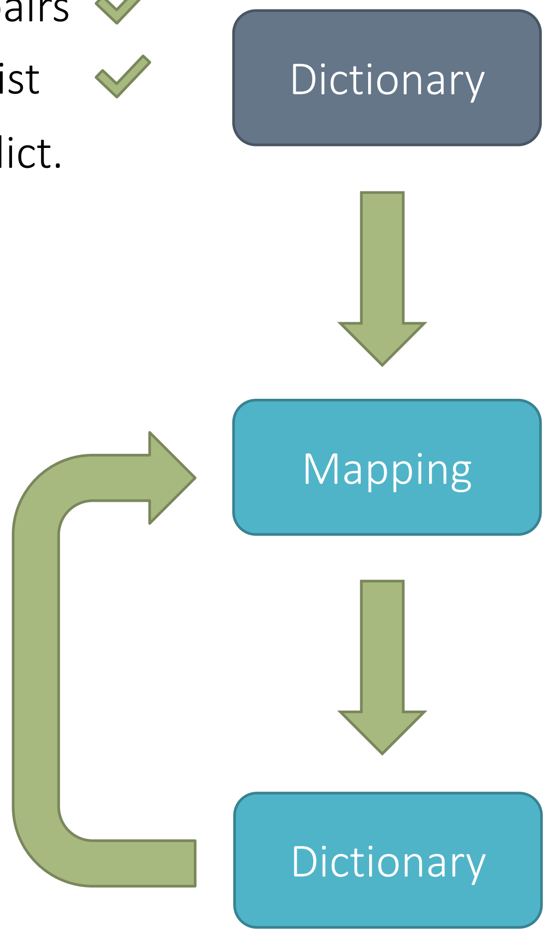


$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \|X_{i*}W - Z_{j*}\|^2$$

Cross-lingual word embedding alignment

Self-learning

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

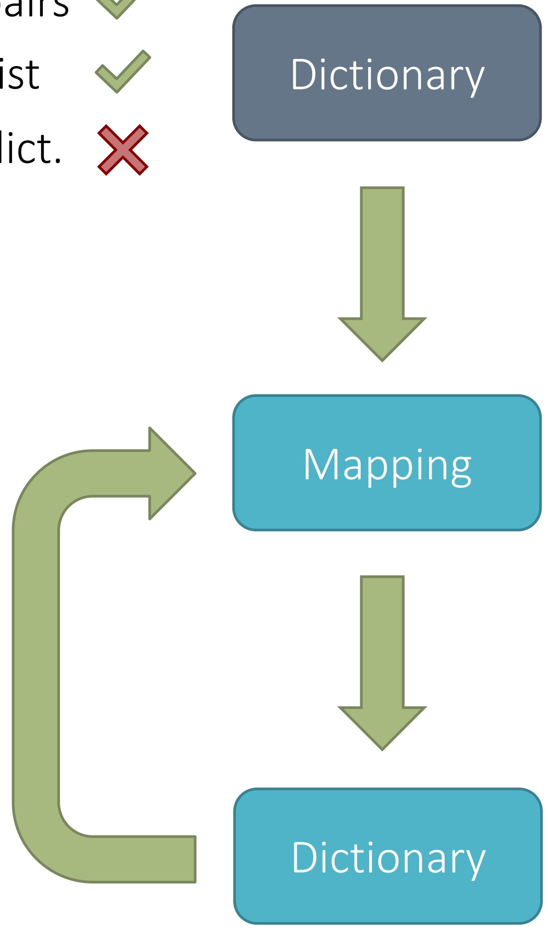


$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \lVert X_{i*}W - Z_{j*} \rVert^2$$

Cross-lingual word embedding alignment

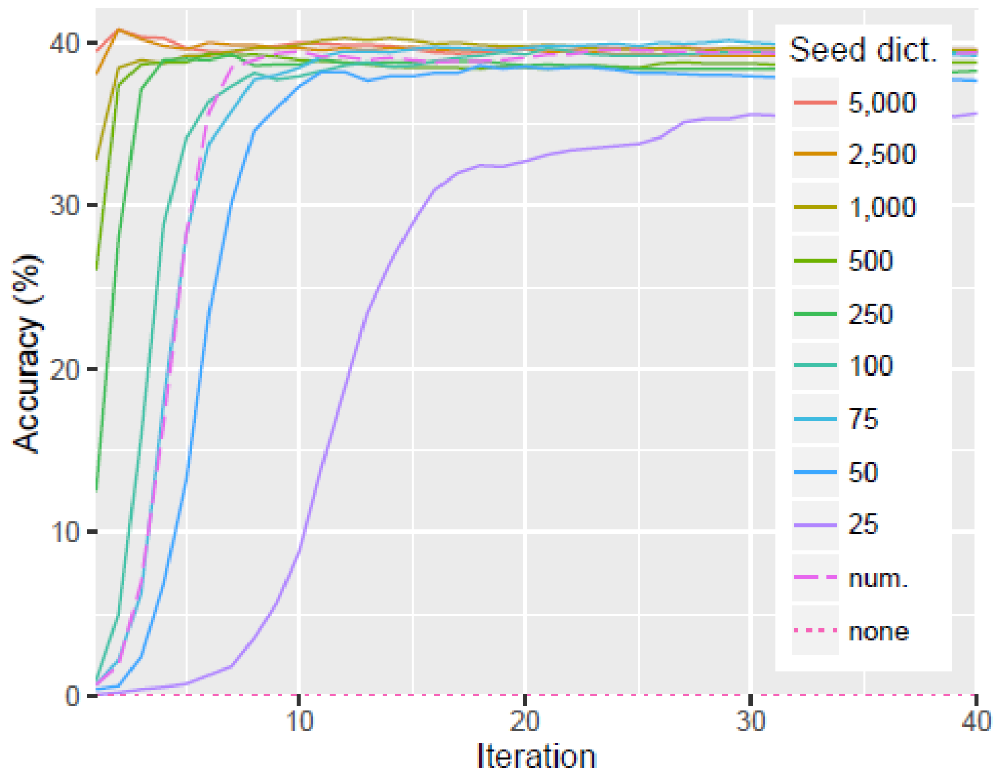
Self-learning

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



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

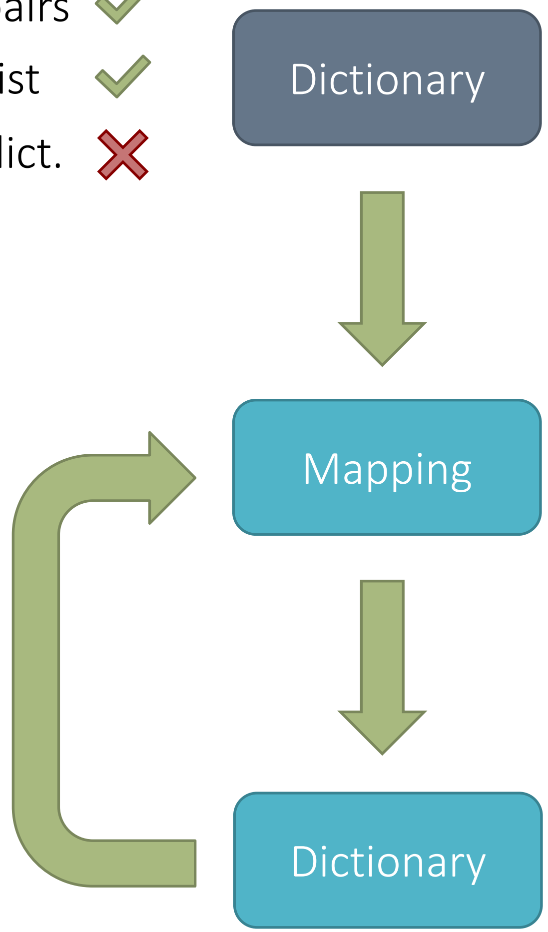
Cross-lingual word embedding alignment



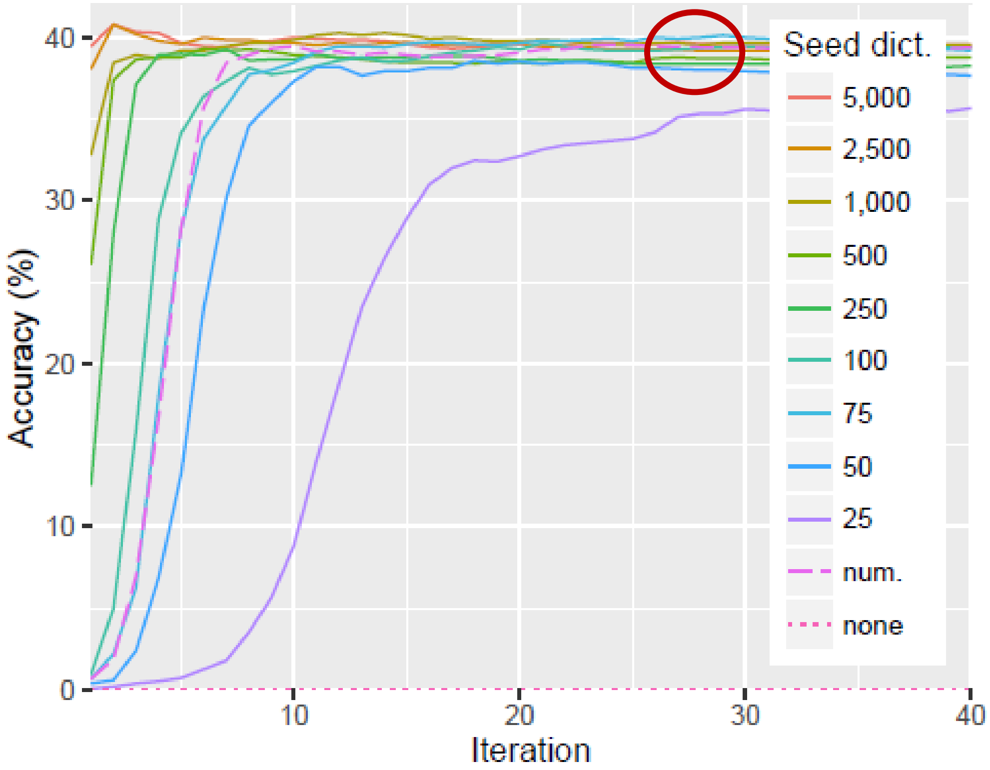
- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

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

Self-learning

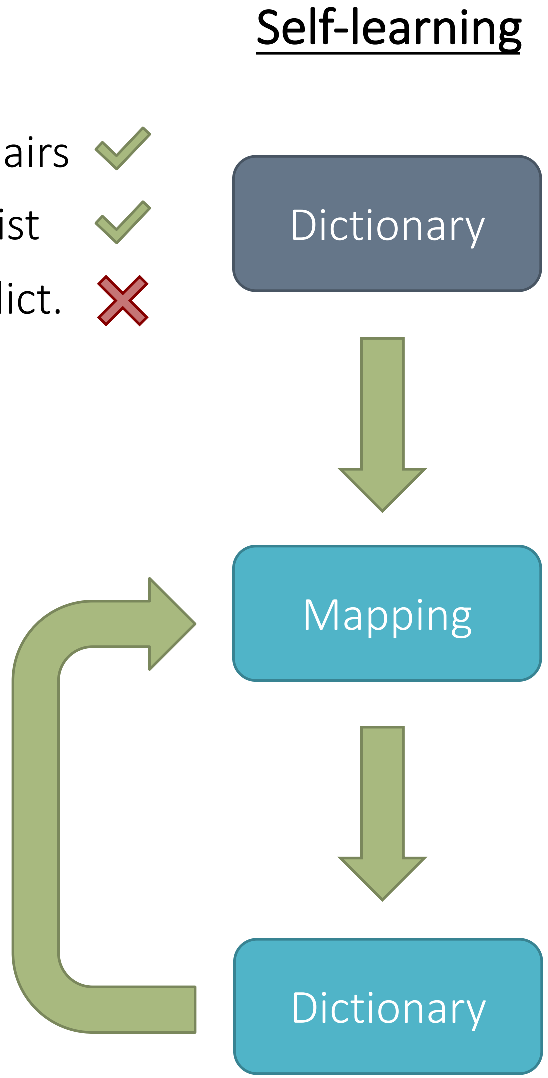


Cross-lingual word embedding alignment

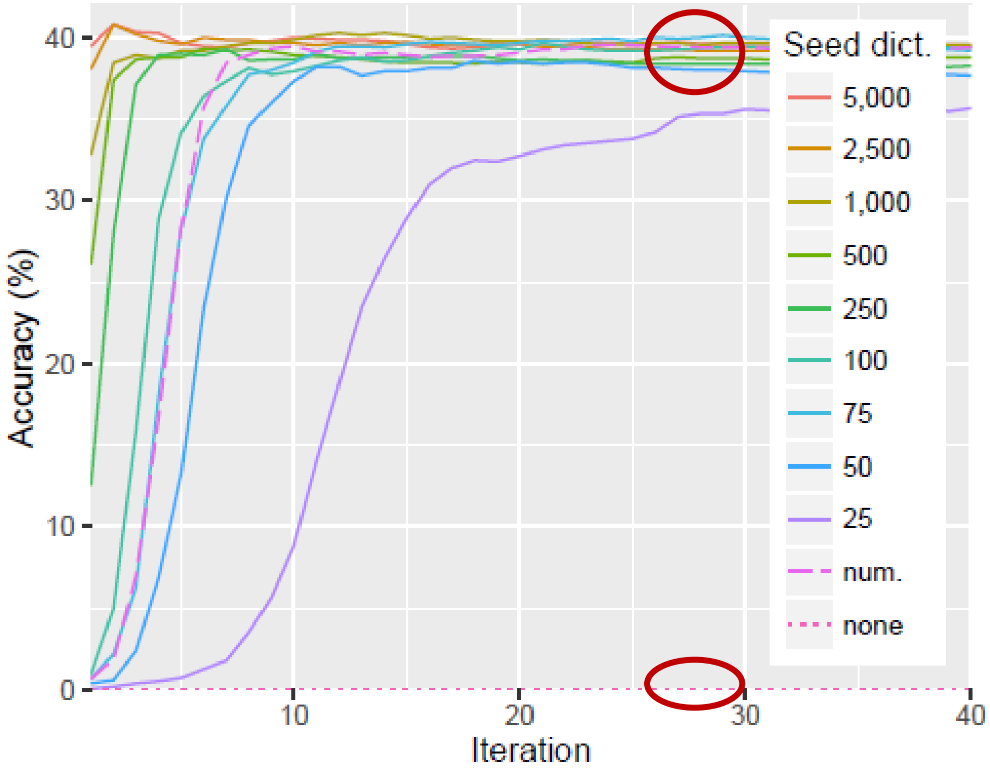


- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

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



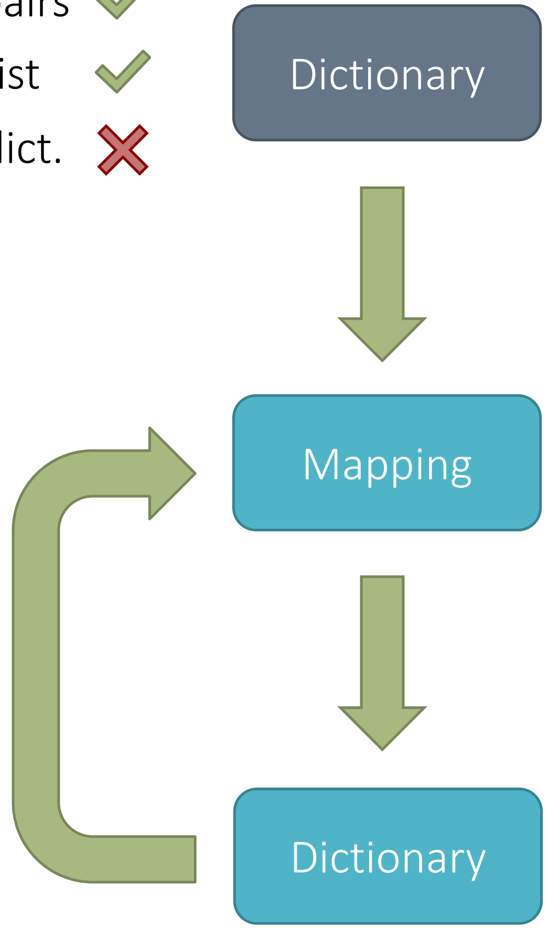
Cross-lingual word embedding alignment



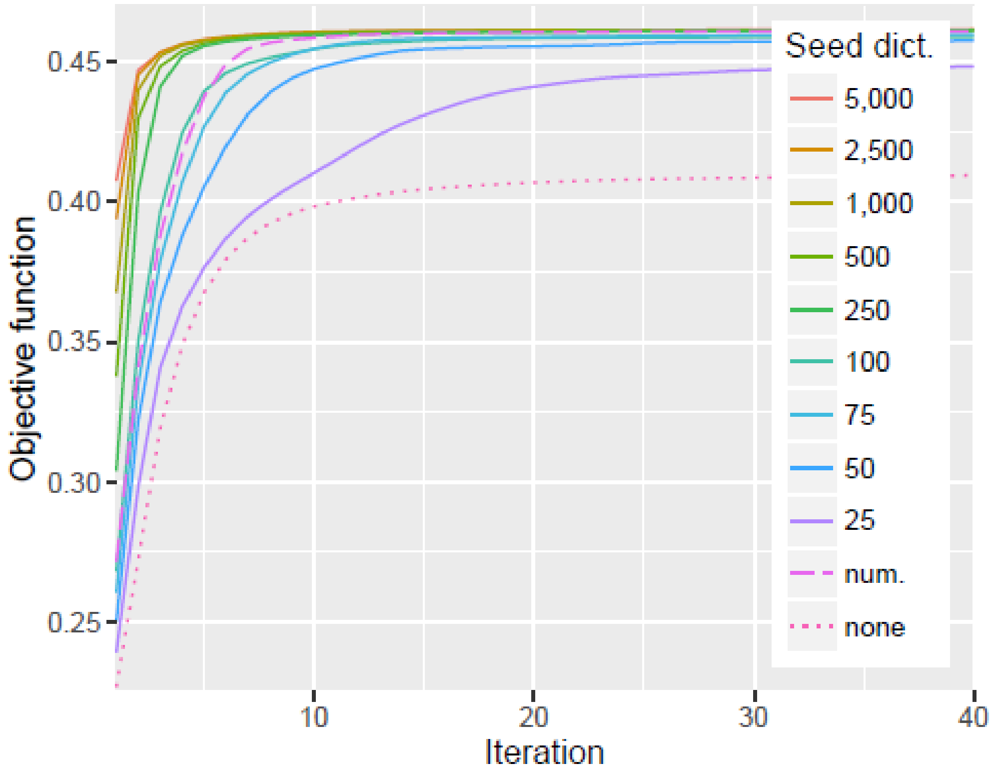
- 25 word pairs ✓
- Numeral list ✓
- Random dict. ✗

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

Self-learning

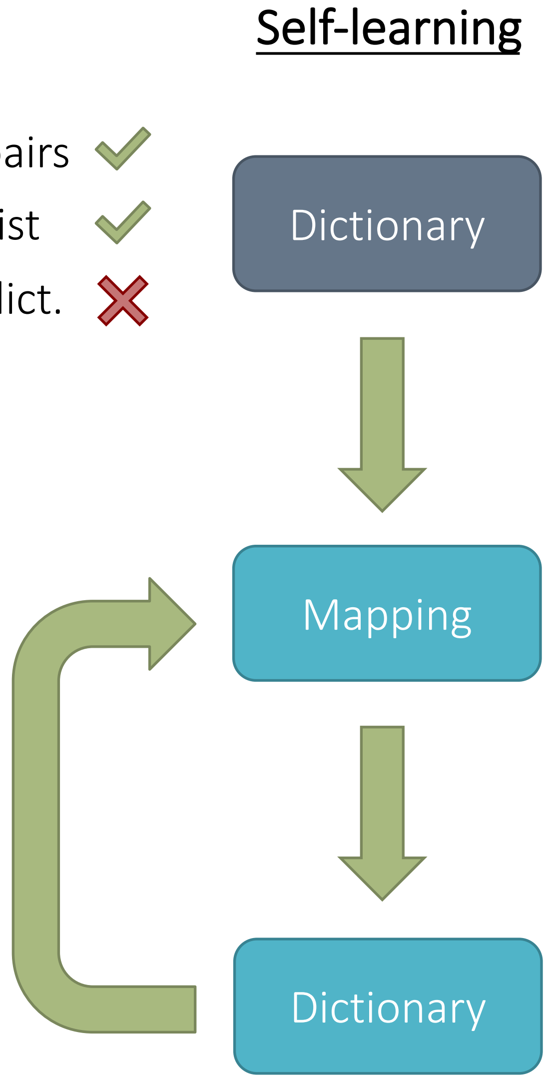


Cross-lingual word embedding alignment

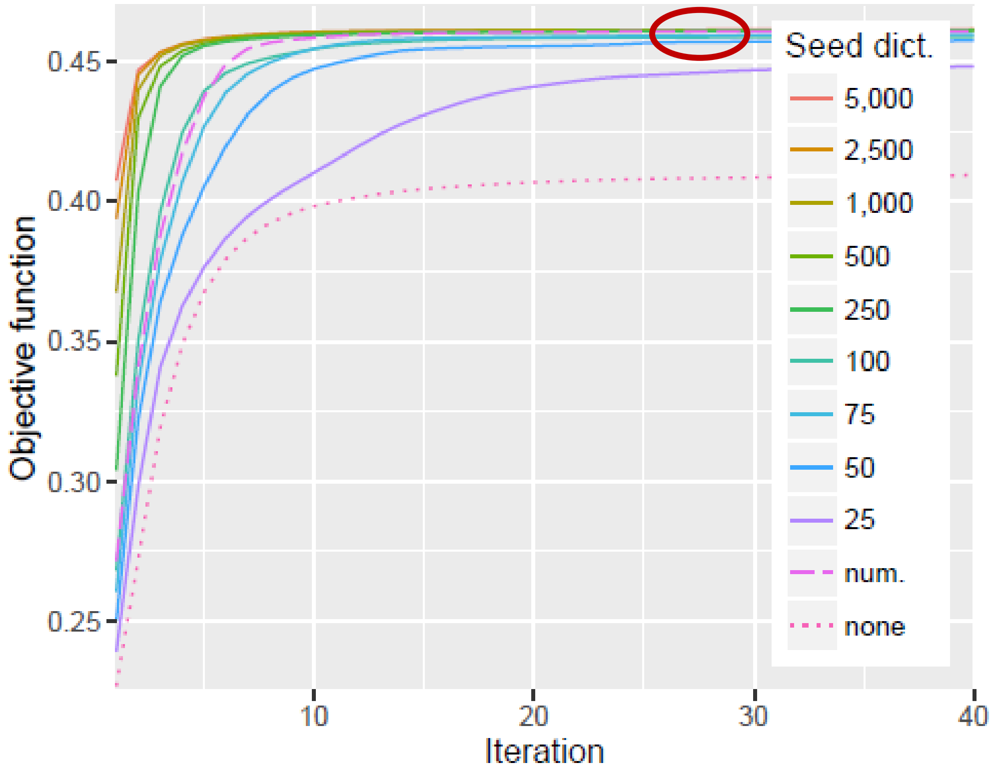


$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \|X_{i*}W - Z_{j*}\|^2$$

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



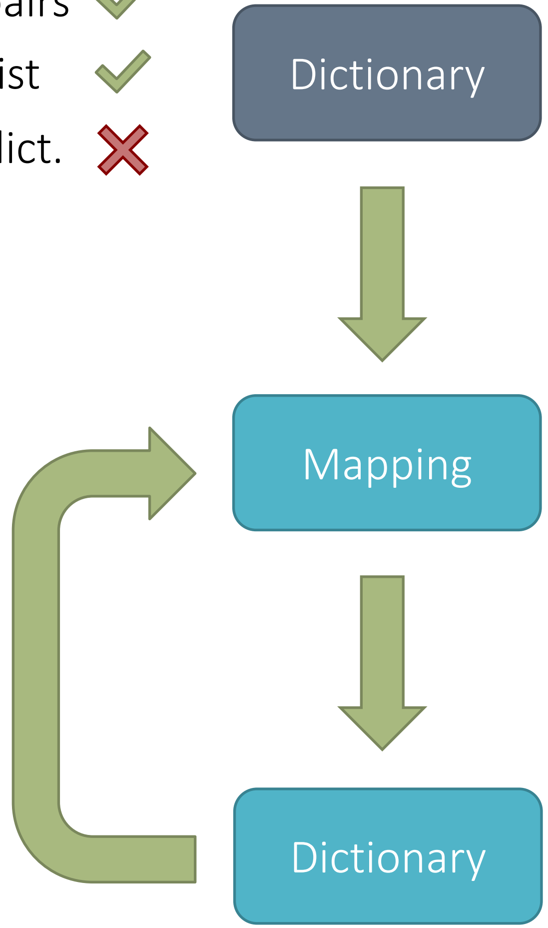
Cross-lingual word embedding alignment



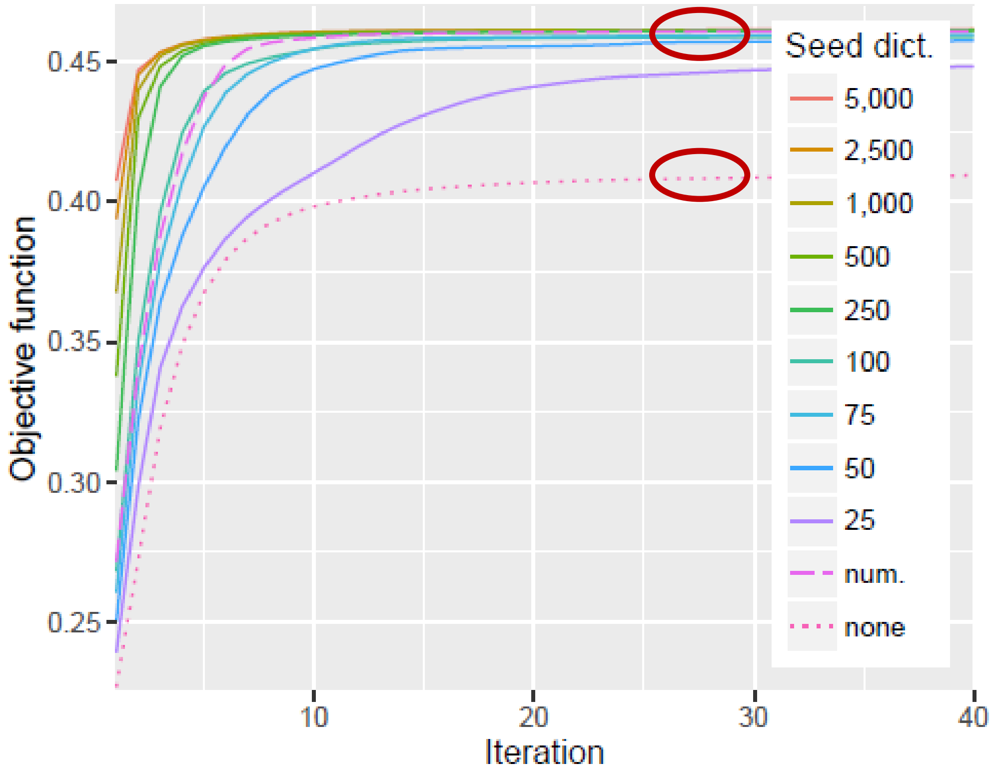
$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \|X_{i*}W - Z_{j*}\|^2$$

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

Self-learning

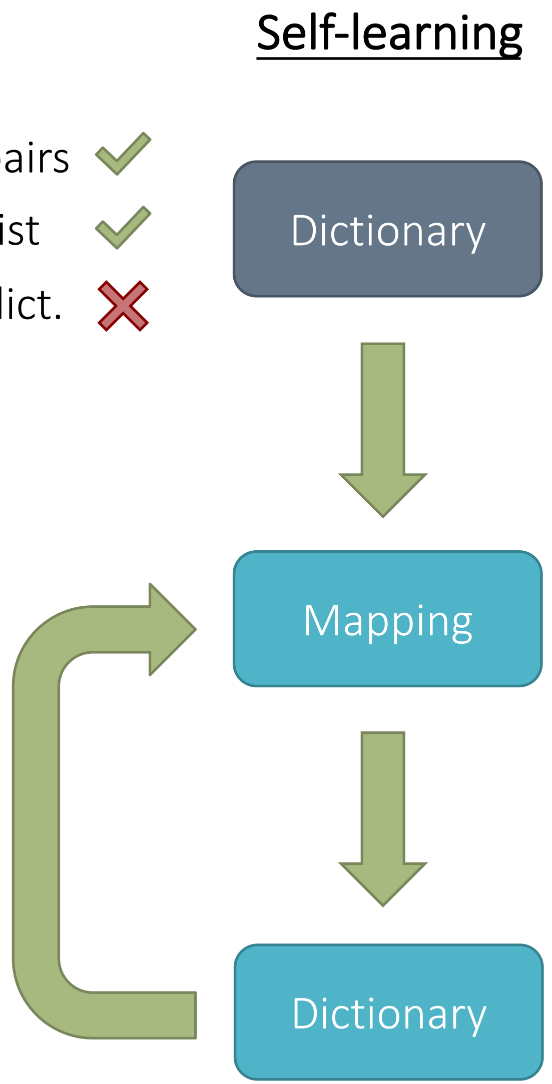


Cross-lingual word embedding alignment



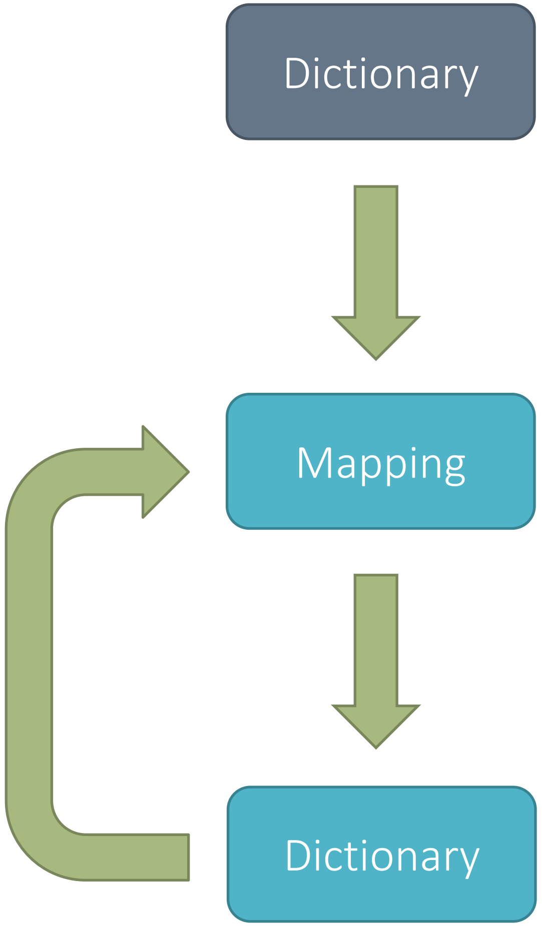
$$W^* = \underset{W \in O(n)}{\arg\min} \sum_i \min_j \|X_{i*}W - Z_{j*}\|^2$$

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



Cross-lingual word embedding alignment

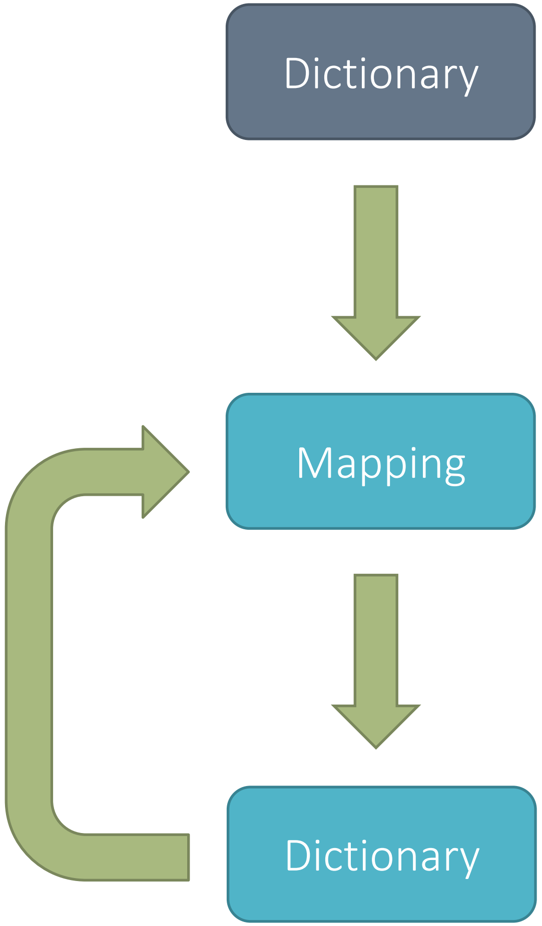
Intra-lingual similarity distribution



Cross-lingual word embedding alignment

English

<u>Intra-lingual similarity distribution</u>

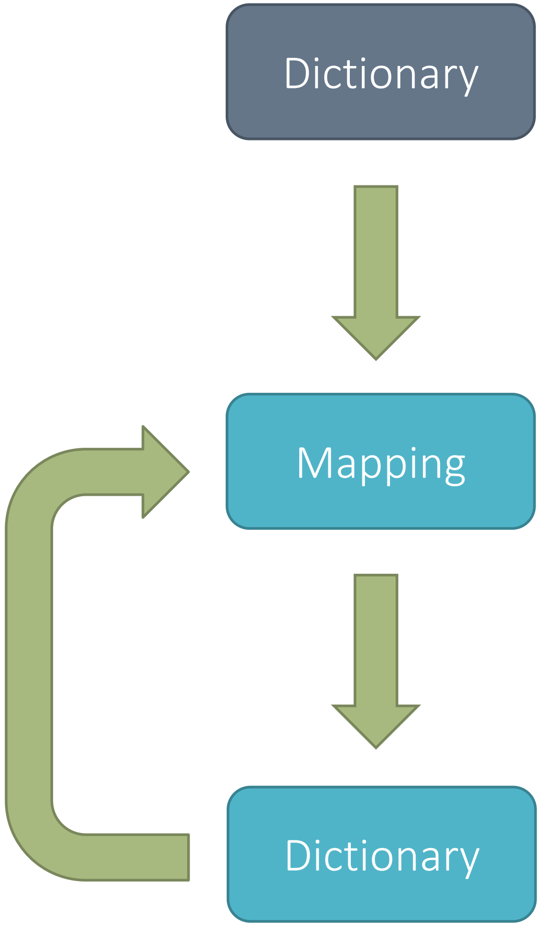


Cross-lingual word embedding alignment

English

two

Intra-lingual similarity distribution



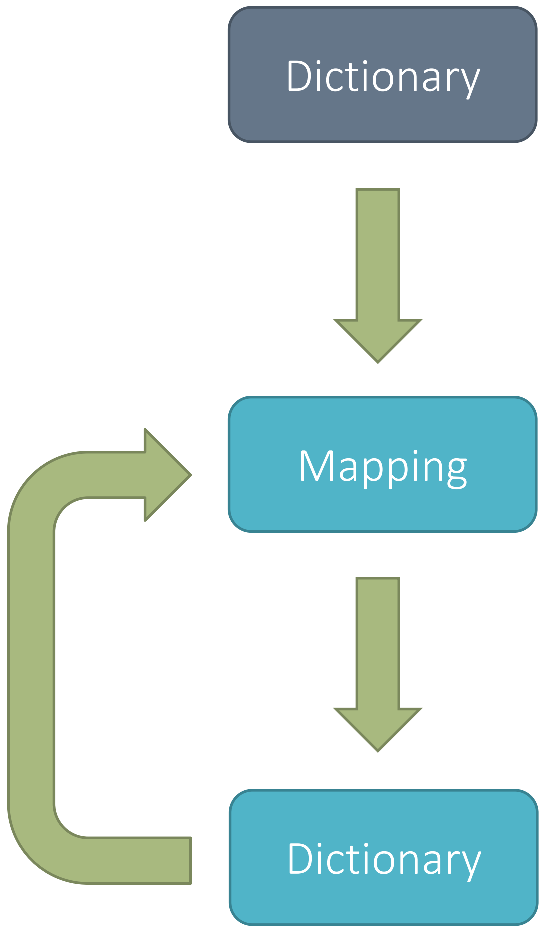
Cross-lingual word embedding alignment

English

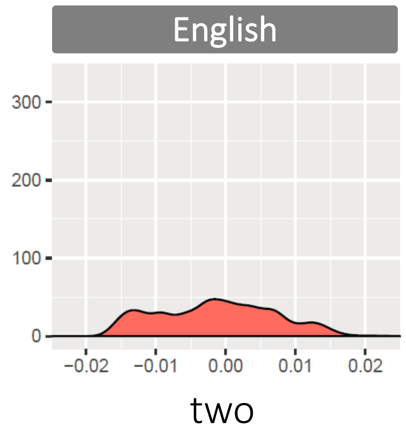
```
for x in vocab:
    sim("two", x)
```

two

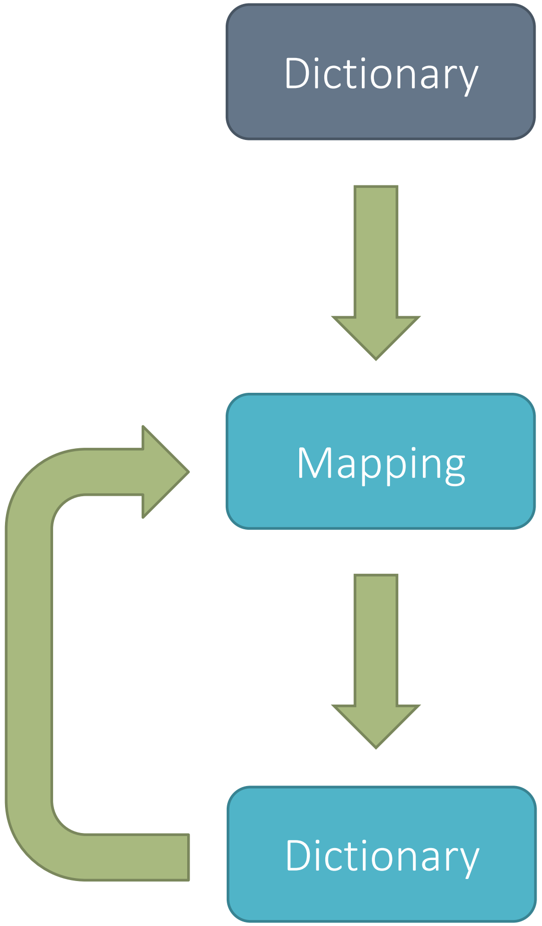
<u>Intra-lingual similarity distribution</u>



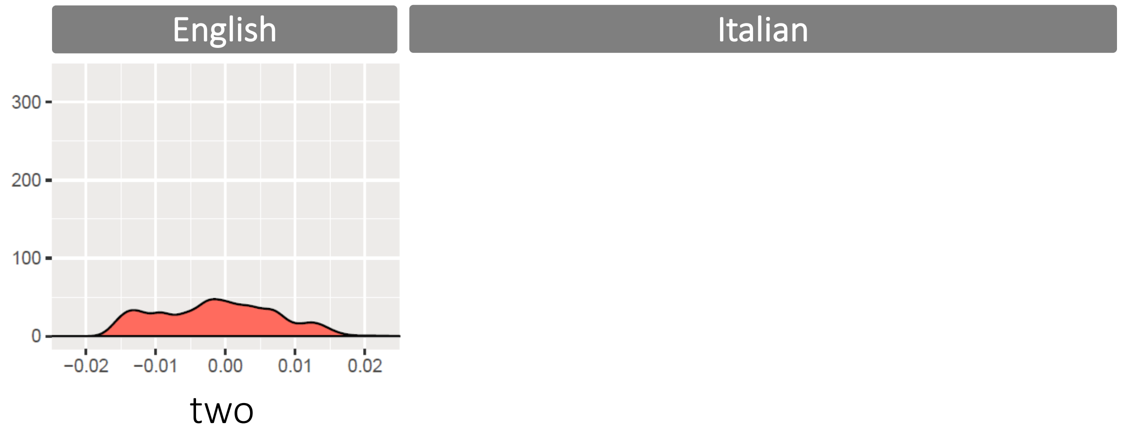
Cross-lingual word embedding alignment



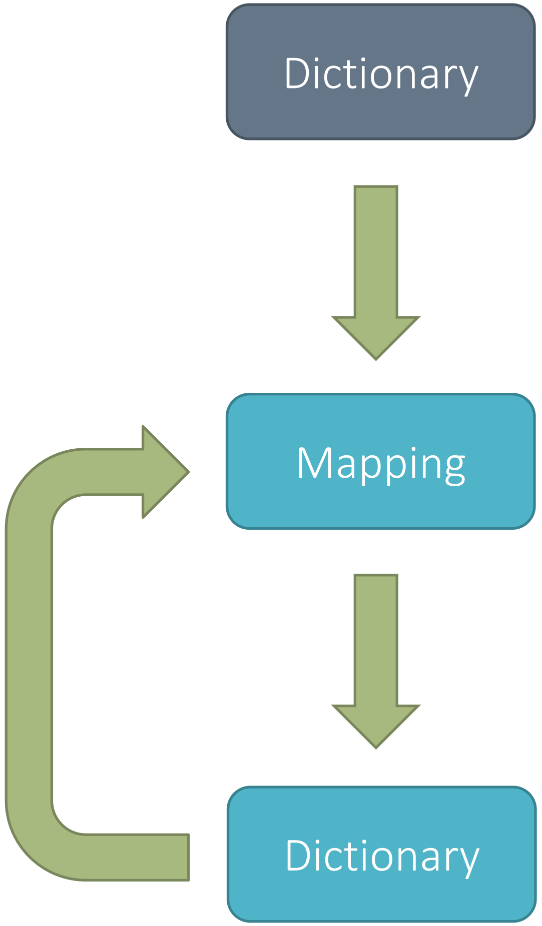
Intra-lingual similarity distribution



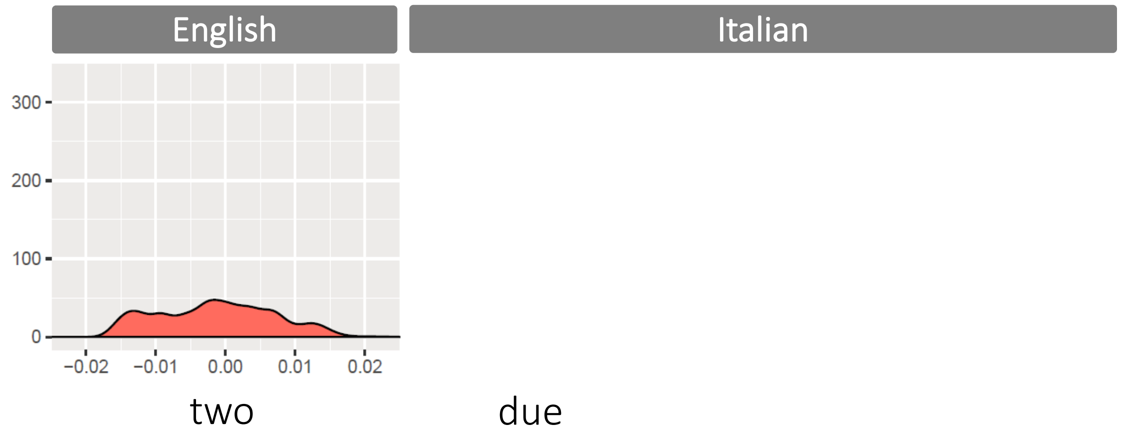
Cross-lingual word embedding alignment



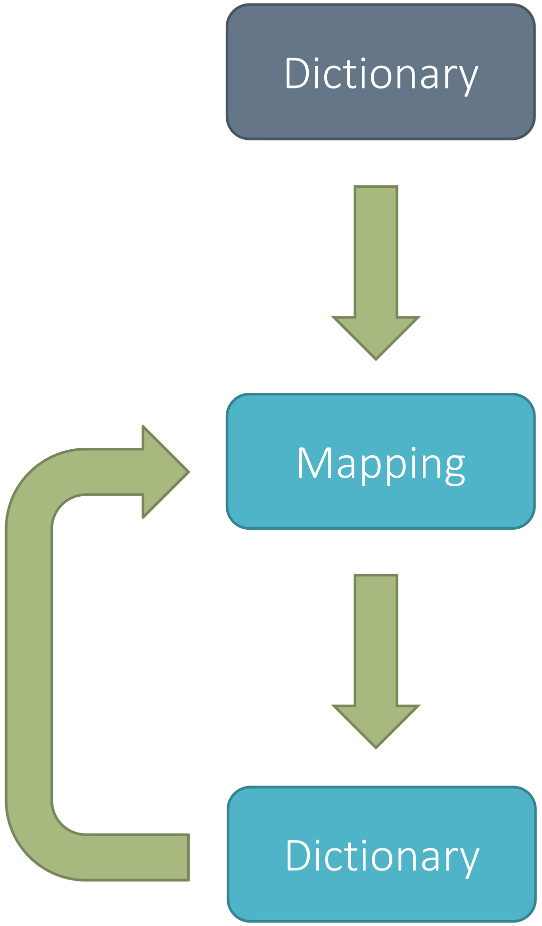
<u>Intra-lingual similarity distribution</u>



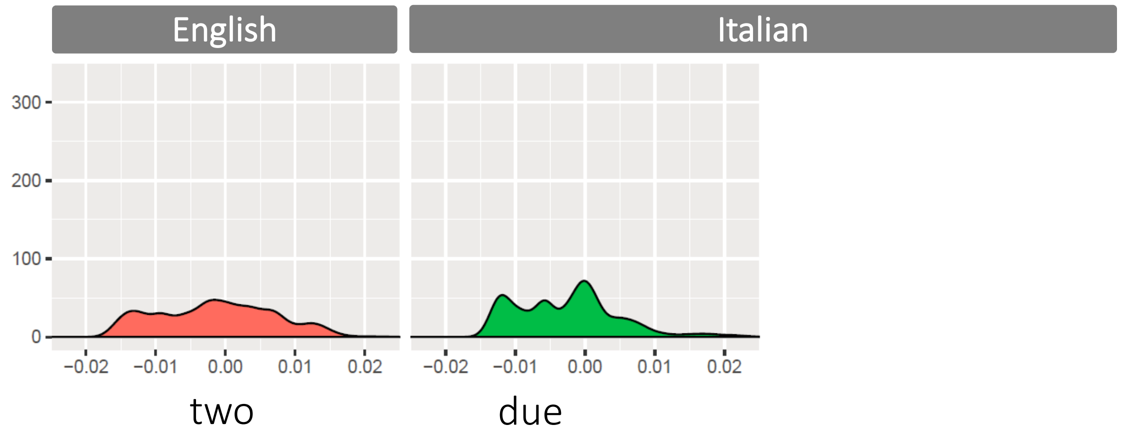
Cross-lingual word embedding alignment



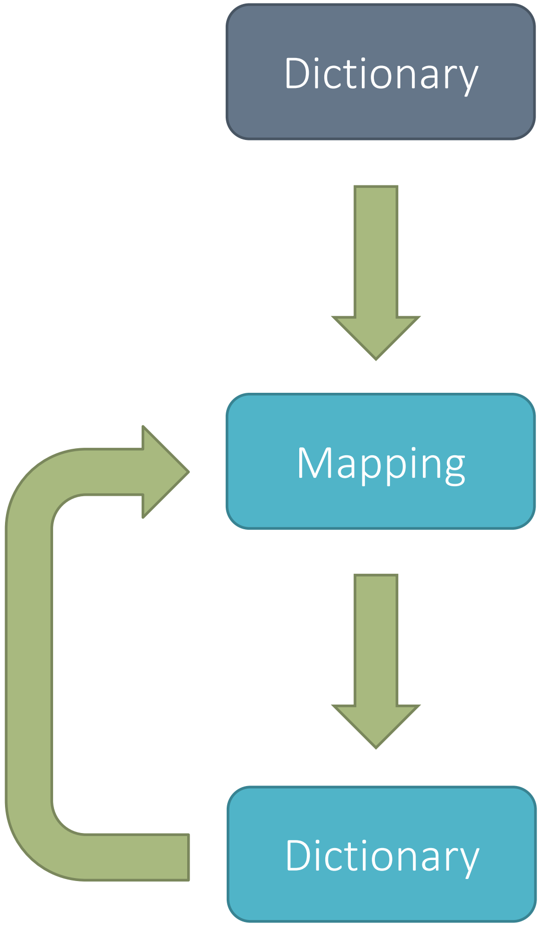
Intra-lingual similarity distribution



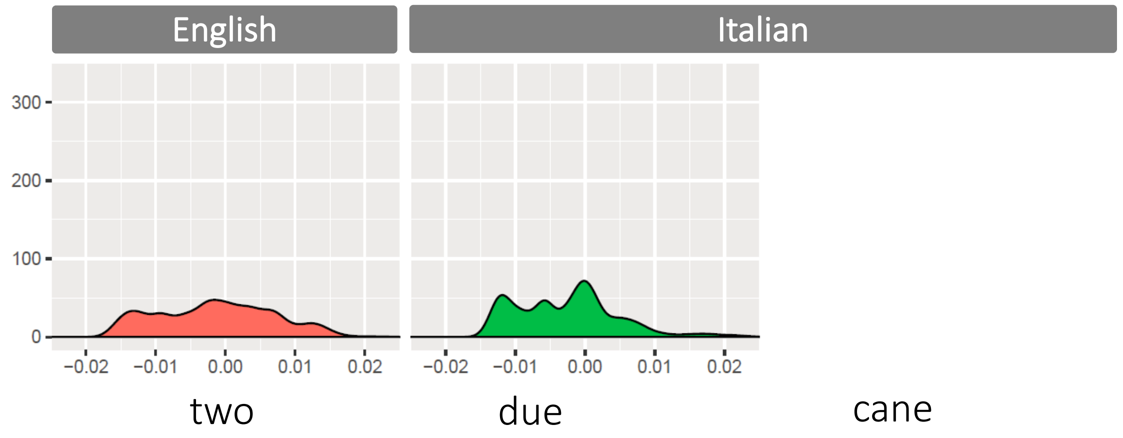
Cross-lingual word embedding alignment



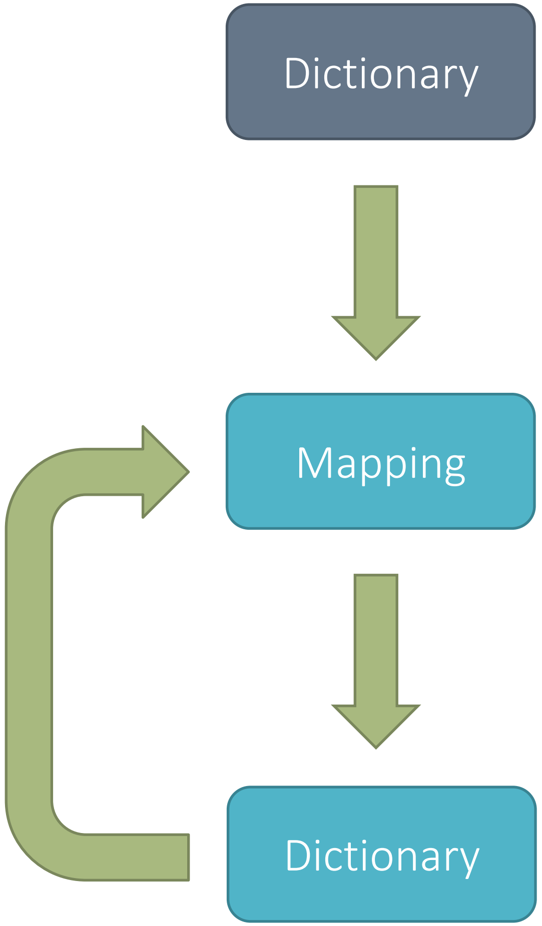
Intra-lingual similarity distribution



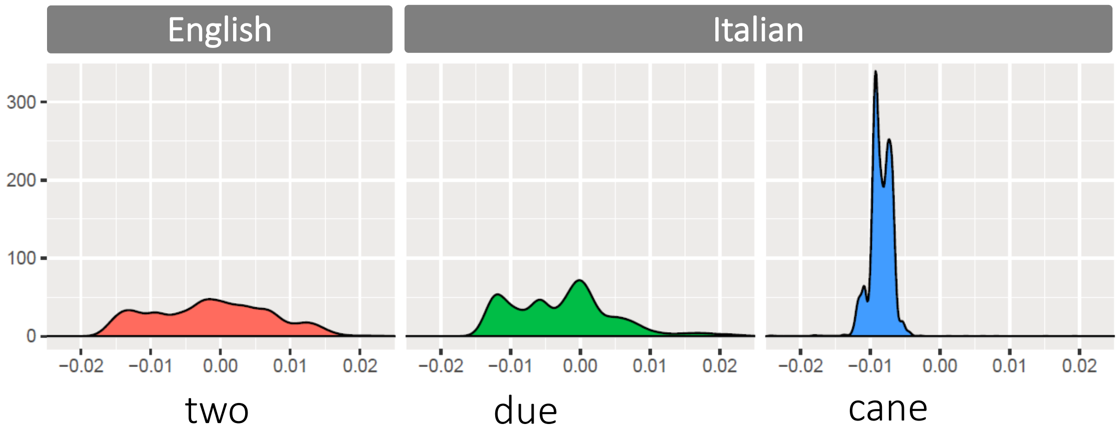
Cross-lingual word embedding alignment



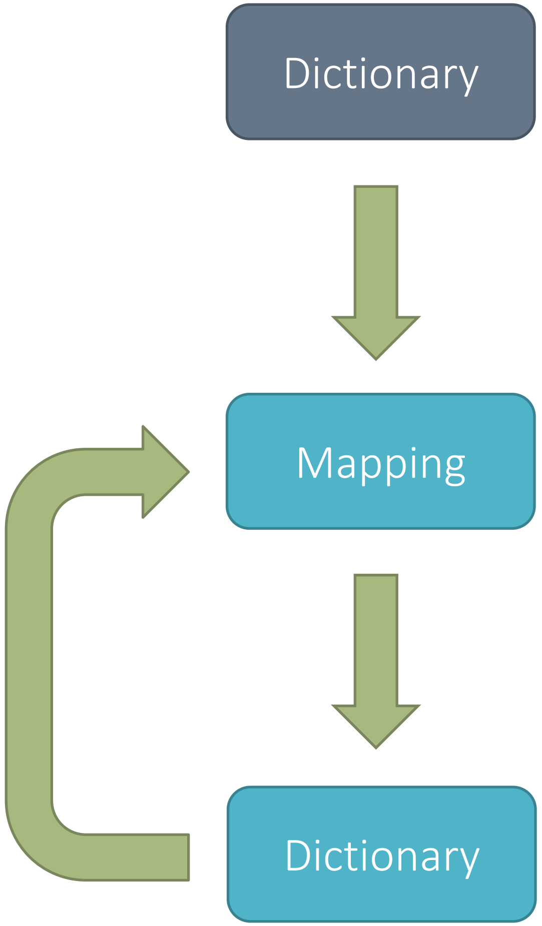
Intra-lingual similarity distribution



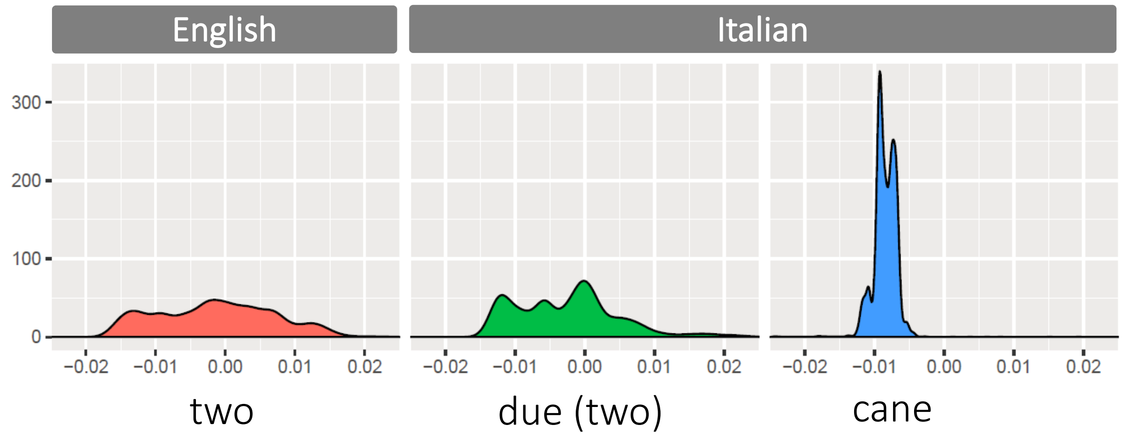
Cross-lingual word embedding alignment



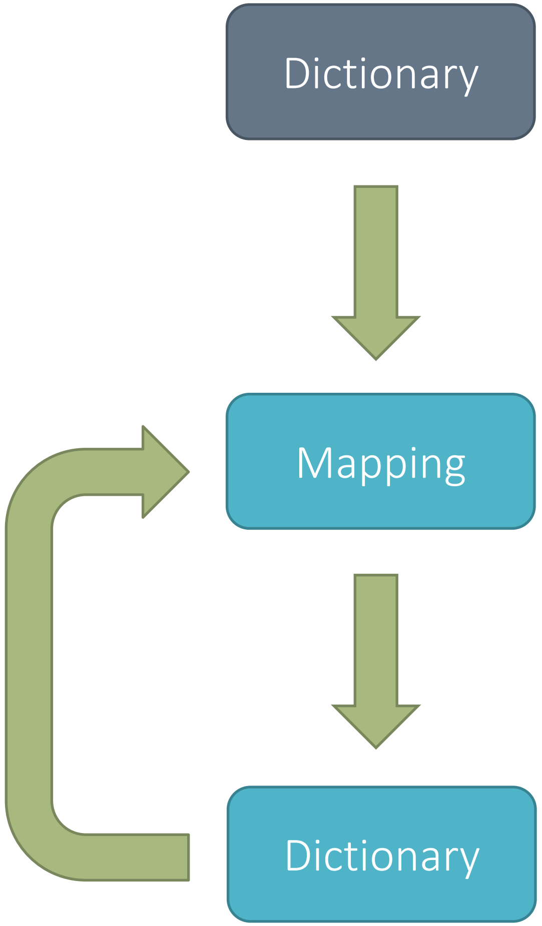
Intra-lingual similarity distribution



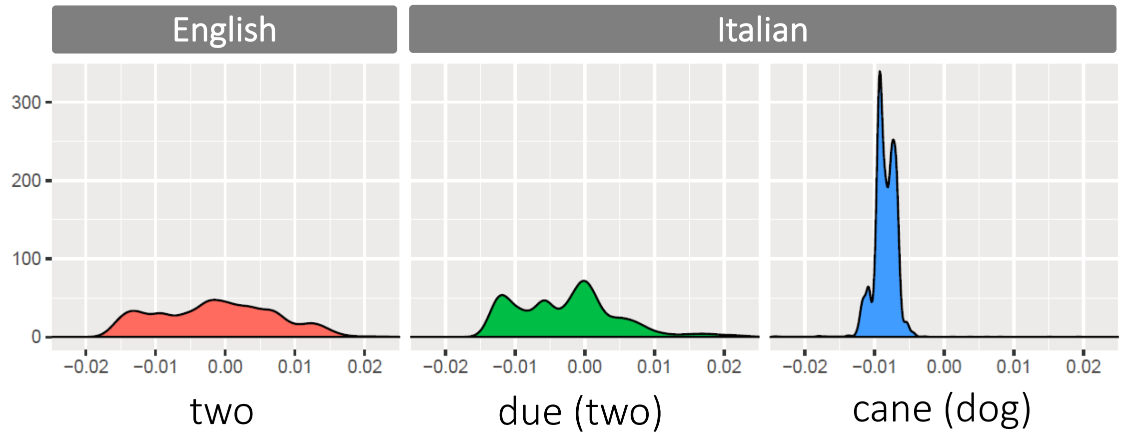
Cross-lingual word embedding alignment



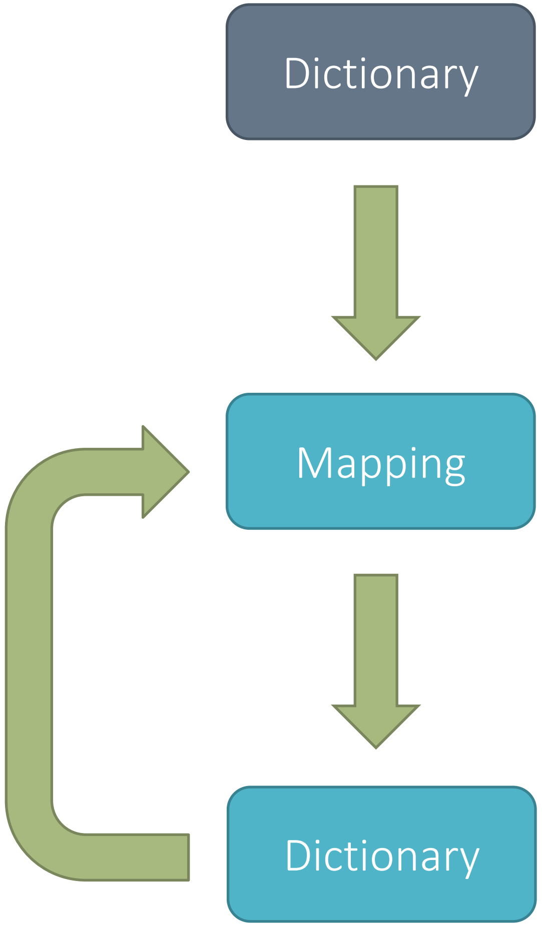
Intra-lingual similarity distribution



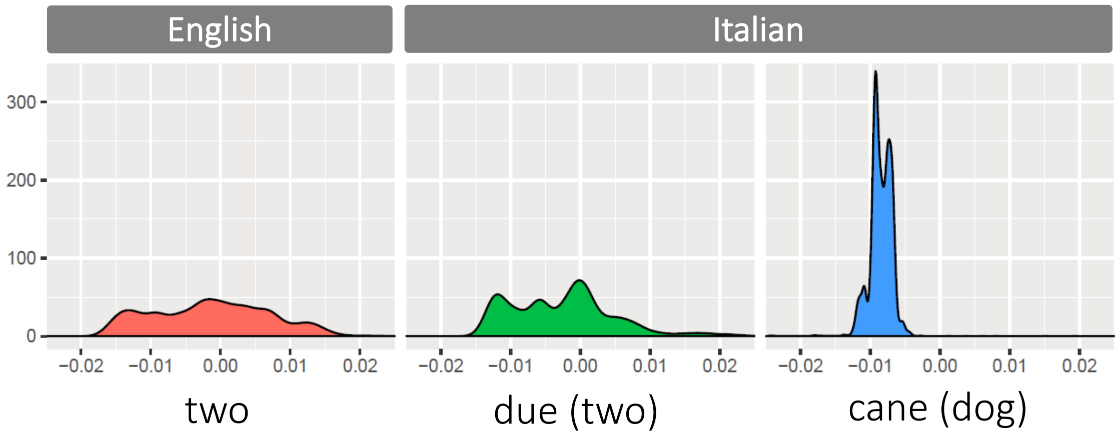
Cross-lingual word embedding alignment



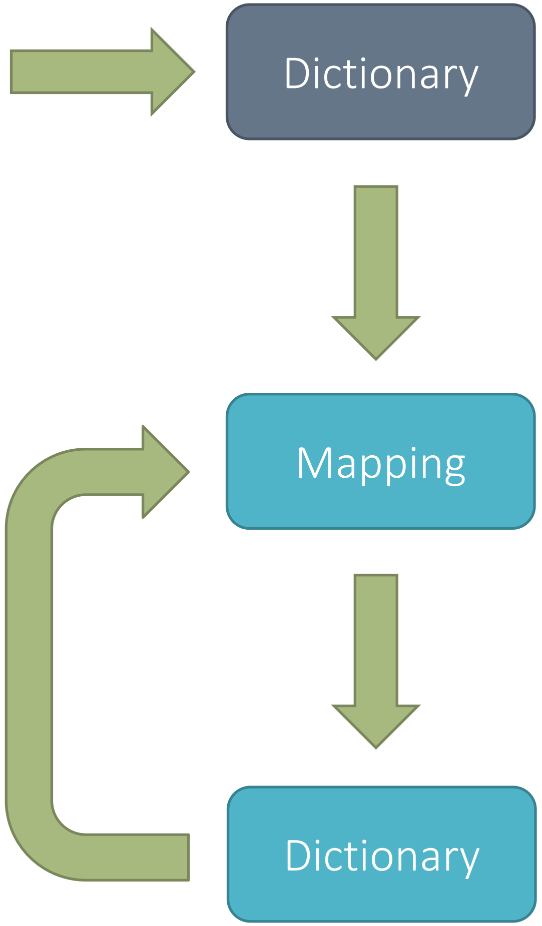
Intra-lingual similarity distribution



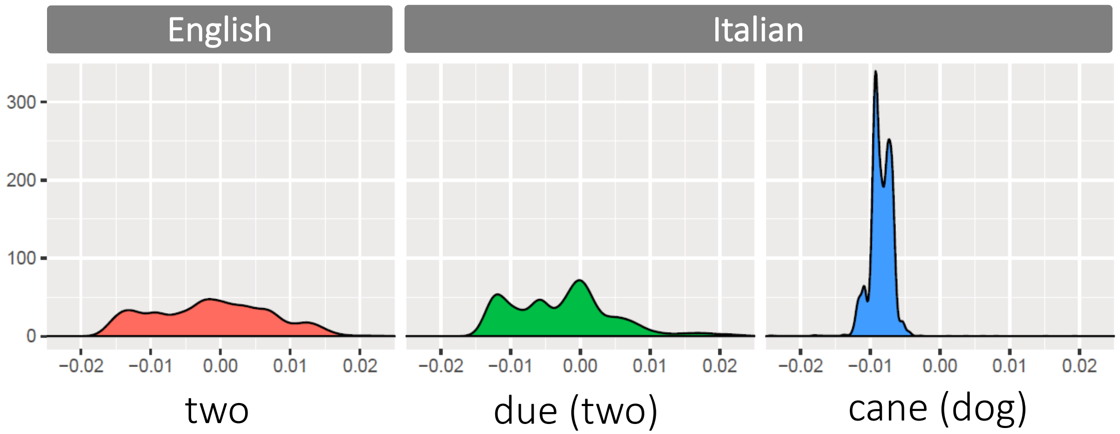
Cross-lingual word embedding alignment



Intra-lingual similarity distribution

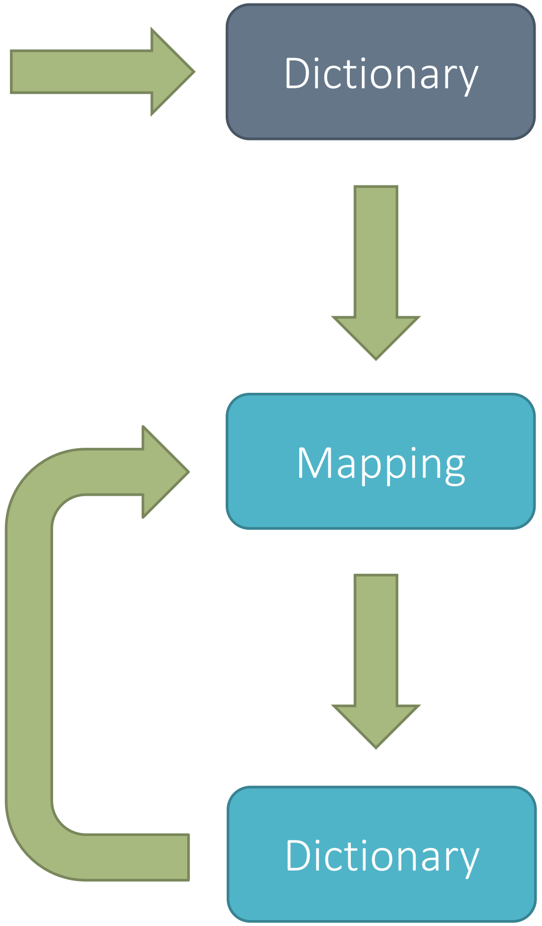


Cross-lingual word embedding alignment

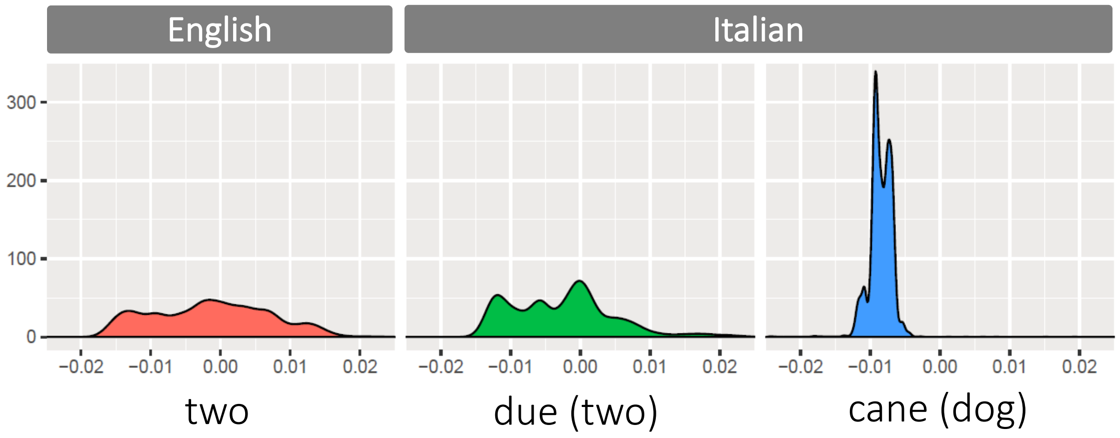


Intra-lingual similarity distribution

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

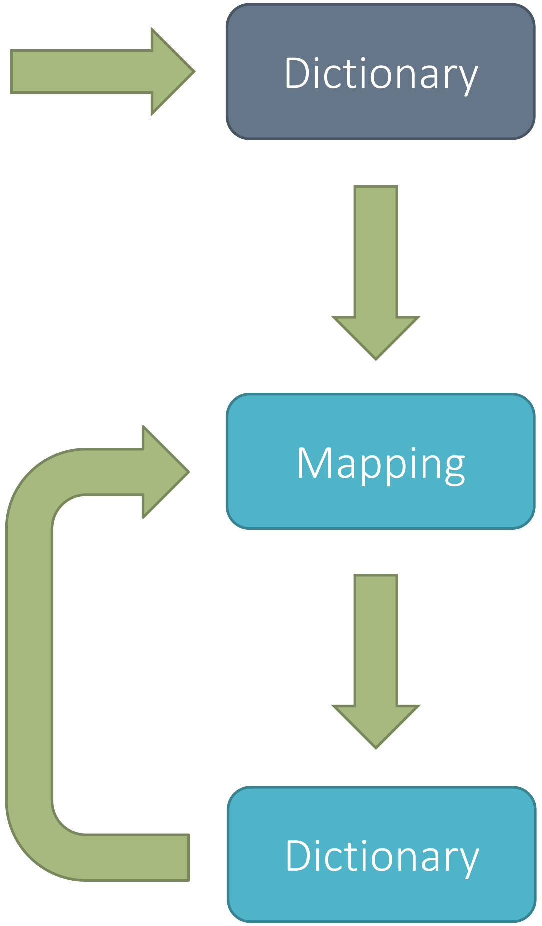


Cross-lingual word embedding alignment

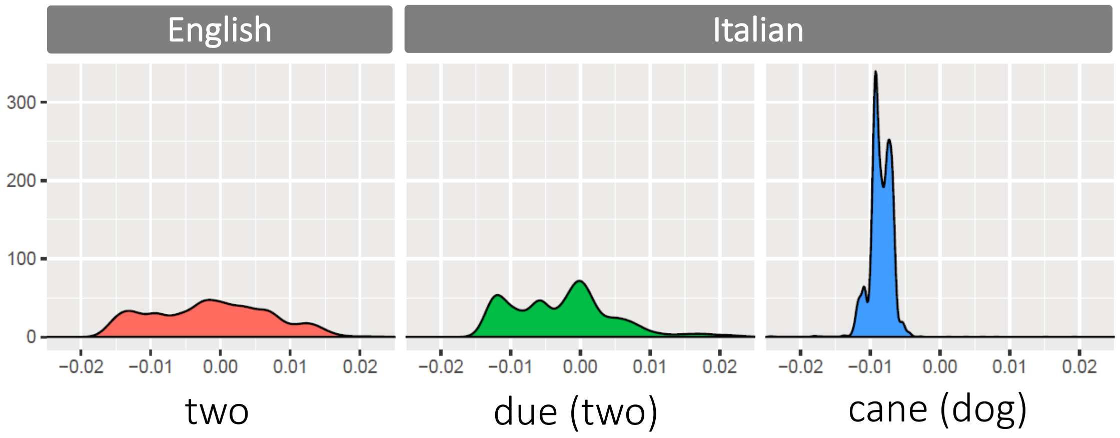


Intra-lingual similarity distribution

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



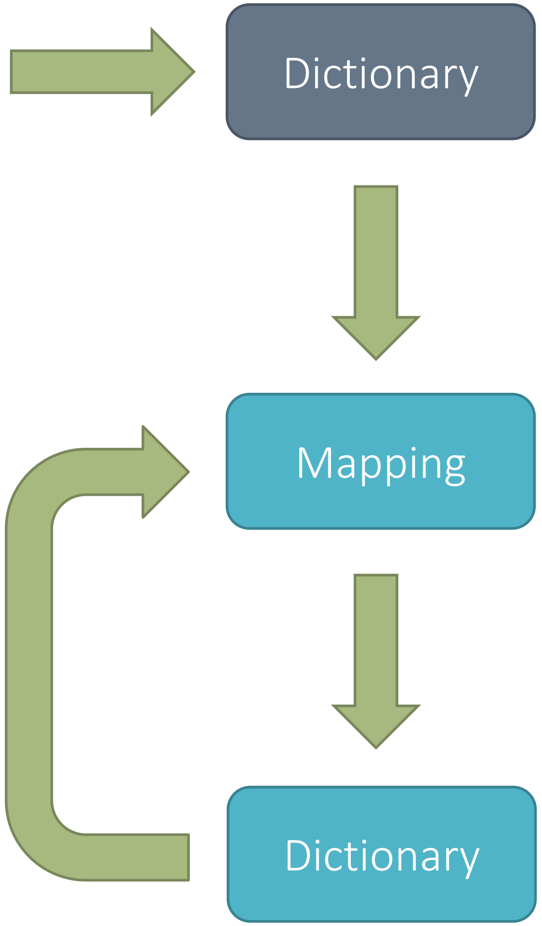
Cross-lingual word embedding alignment



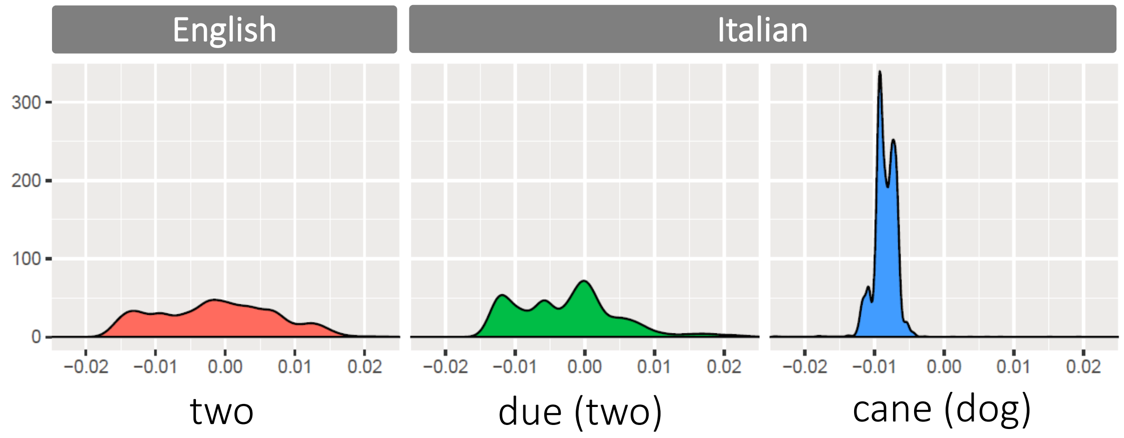
Intra-lingual similarity distribution

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

+



Cross-lingual word embedding alignment

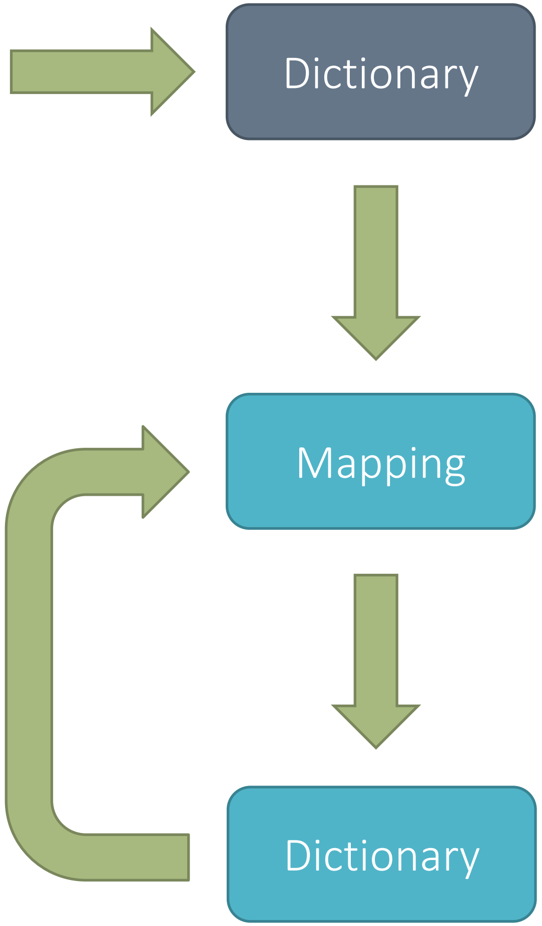


Intra-lingual similarity distribution

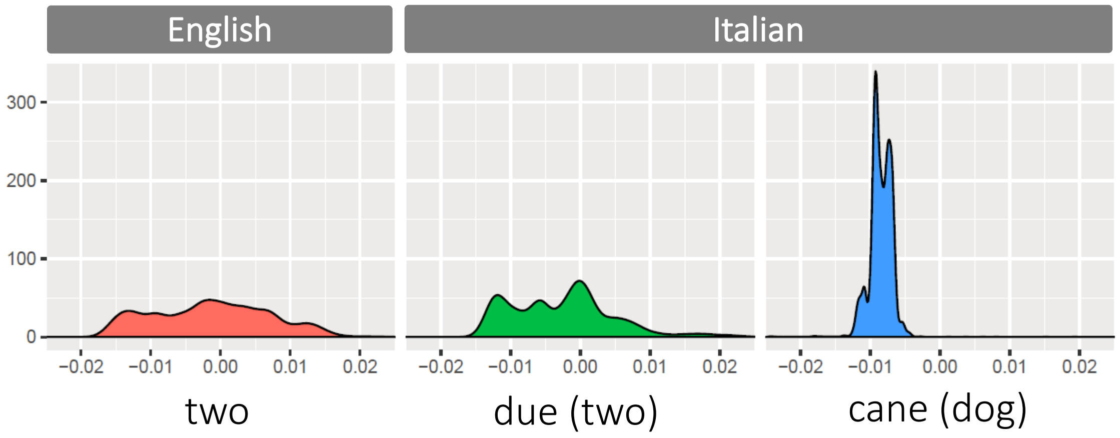
$$X' = \text{sorted}\left(\sqrt{XX^T}\right) \quad Z' = \text{sorted}\left(\sqrt{ZZ^T}\right)$$

+

Robust self-learning



Cross-lingual word embedding alignment



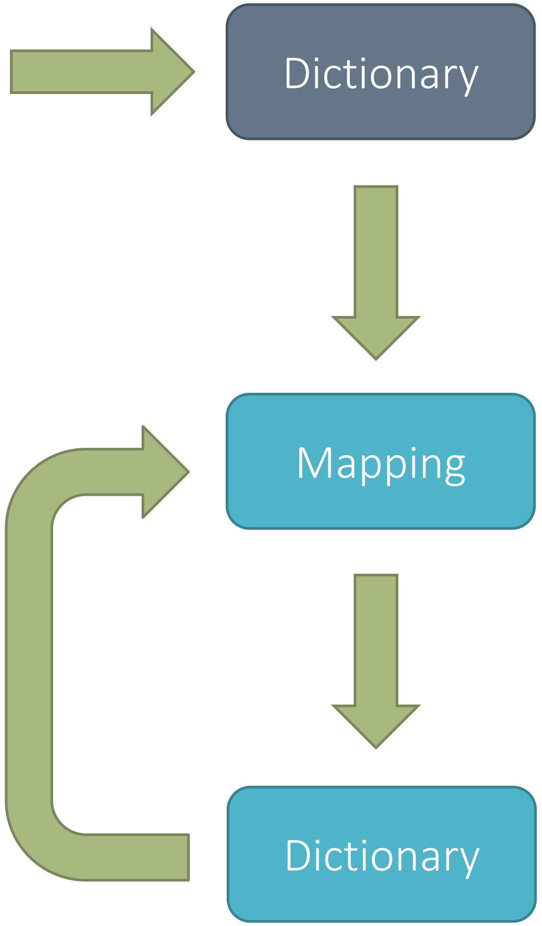
<u>Intra-lingual similarity distribution</u>

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

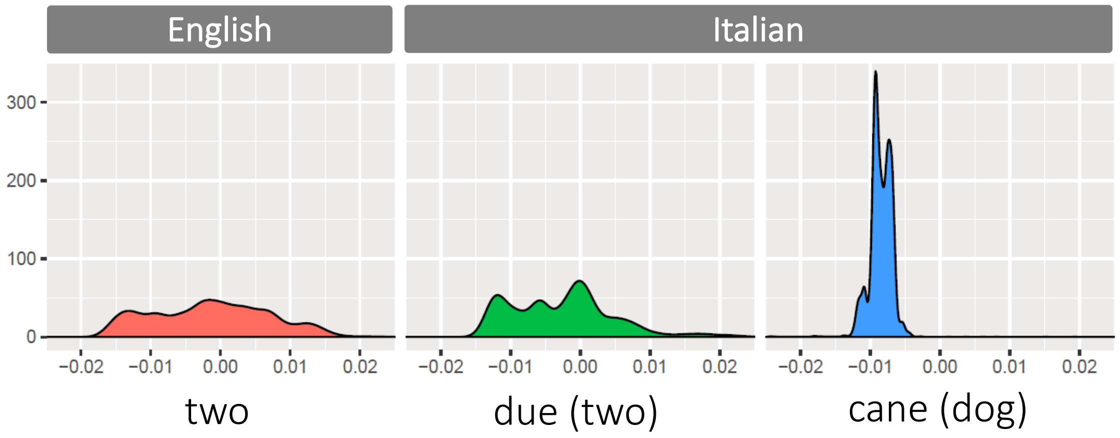
+

<u>Robust self-learning</u>

- Stochastic dictionary induction



Cross-lingual word embedding alignment



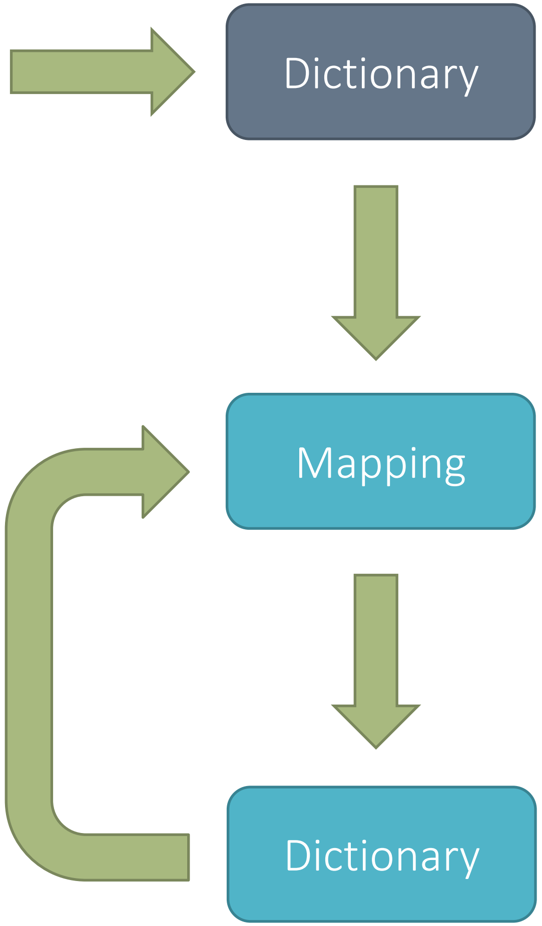
<u>Intra-lingual similarity distribution</u>

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

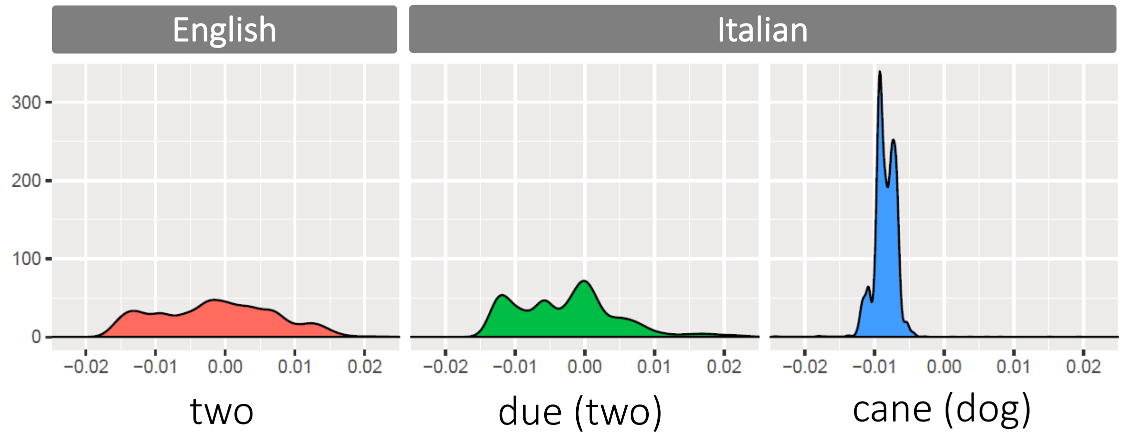
+

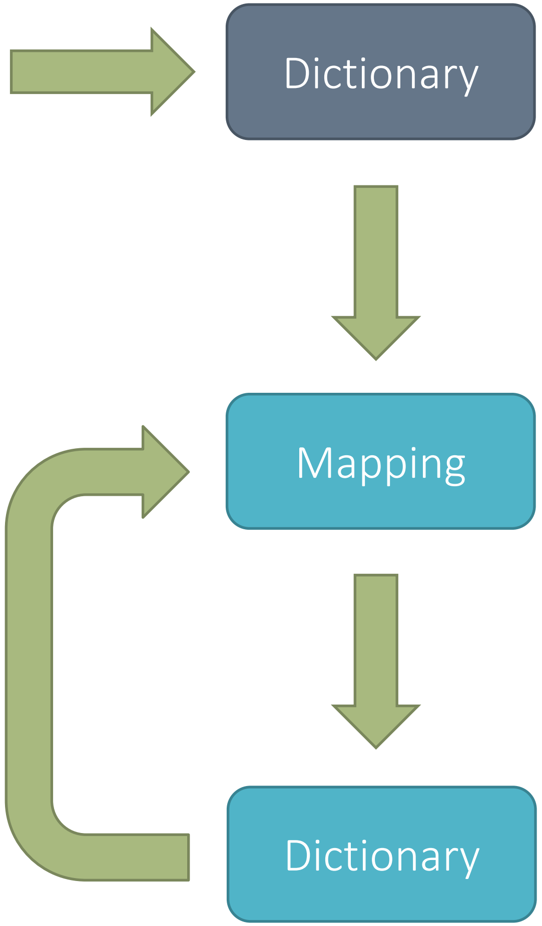
<u>Robust self-learning</u>

- Stochastic dictionary induction
- Frequency-based vocabulary cutoff



Cross-lingual word embedding alignment





Intra-lingual similarity distribution

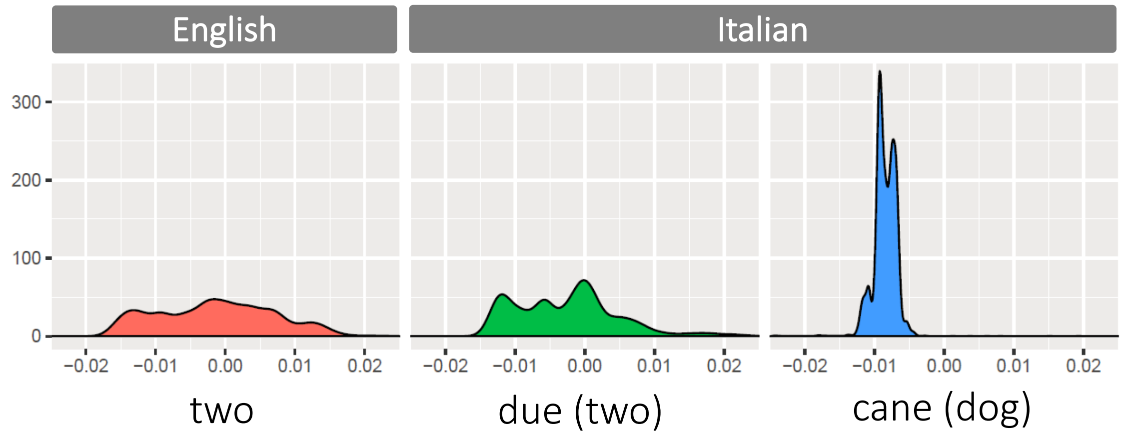
$$X' = \text{sorted}\left(\sqrt{XX^T}\right) \quad Z' = \text{sorted}\left(\sqrt{ZZ^T}\right)$$

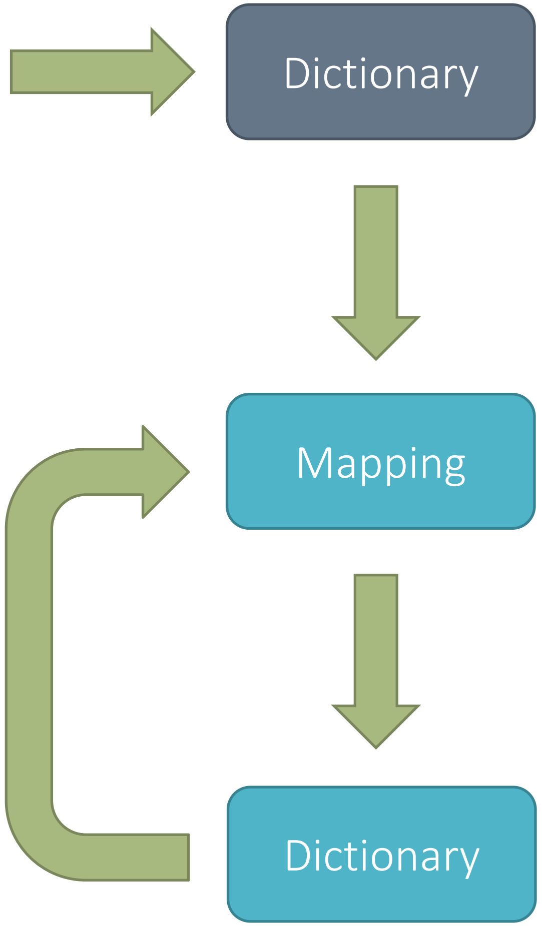
+

Robust self-learning

- Stochastic dictionary induction
- Frequency-based vocabulary cutoff
- CSLS retrieval (Conneau et al., ICLR’18)

Cross-lingual word embedding alignment





Intra-lingual similarity distribution

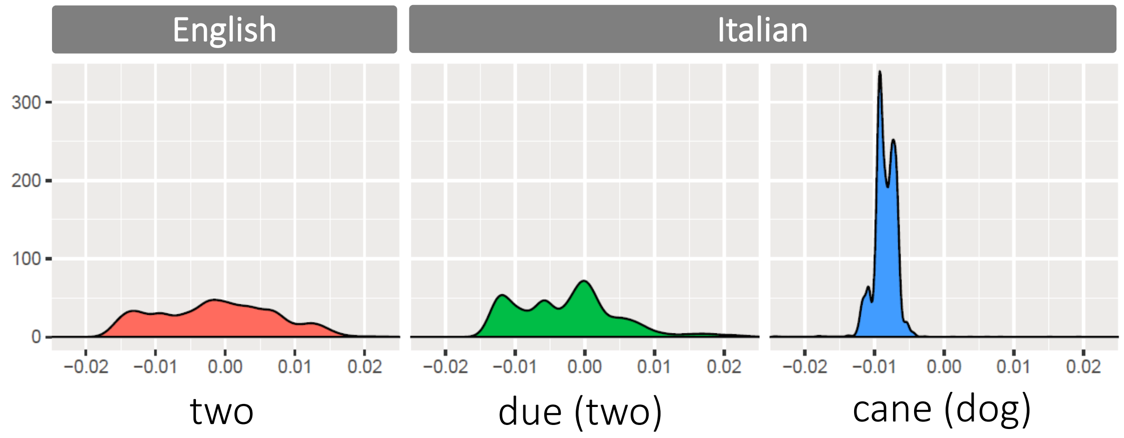
$$X' = \text{sorted}\left(\sqrt{XX^T}\right) \quad Z' = \text{sorted}\left(\sqrt{ZZ^T}\right)$$

+

Robust self-learning

- Stochastic dictionary induction
- Frequency-based vocabulary cutoff
- CSLS retrieval (Conneau et al., ICLR'18)
- Bidirectional dictionary induction

Cross-lingual word embedding alignment



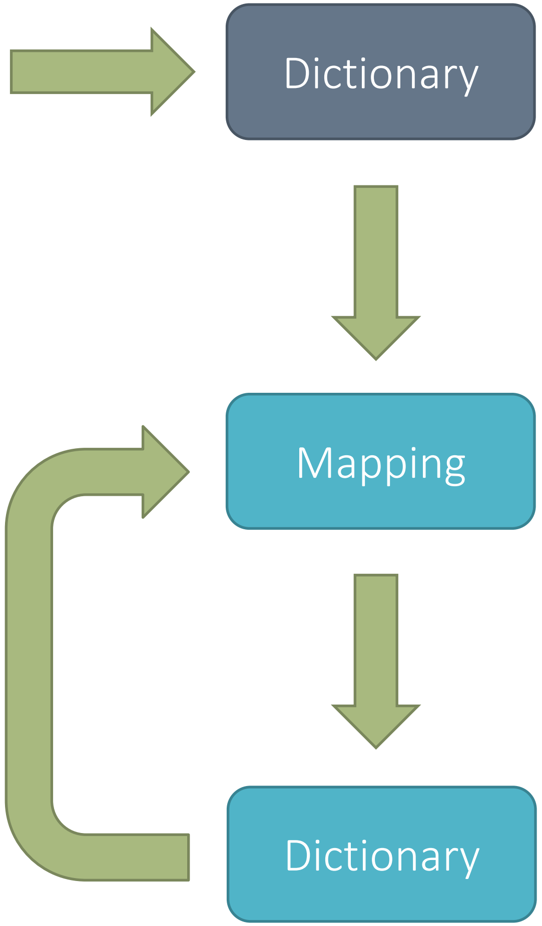
Intra-lingual similarity distribution

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

+

Robust self-learning

- Stochastic dictionary induction
- Frequency-based vocabulary cutoff
- CSLS retrieval (Conneau et al., ICLR'18)
- Bidirectional dictionary induction
- Final symmetric re-weighting



Cross-lingual word embedding alignment

Cross-lingual word embedding alignment

Supervision	Method

Cross-lingual word embedding alignment

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 word embedding alignment

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 word embedding alignment

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 word embedding alignment

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 word embedding alignment

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 word embedding alignment

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict.	Mikolov et al. (2013)	$34.93^{\dagger}$	$35.00^{\dagger}$	$25.91^{\dagger}$	$27.73^{\dagger}$
	Faruqui and Dyer (2014)	38.40^{*}	37.13^{*}	27.60^{*}	26.80^{*}
	Shigeto et al. (2015)	$41.53^{\dagger}$	$43.07^{\dagger}$	$31.04^{\dagger}$	$33.73^{\dagger}$
	Dinu et al. (2015)	37.7	38.93^{*}	29.14^{*}	30.40^{*}
	Lazaridou et al. (2015)	40.2	-	-	-
	Xing et al. (2015)	$36.87^{\dagger}$	$41.27^{\dagger}$	$28.23^{\dagger}$	$31.20^{\dagger}$
	Zhang et al. (2016)	$36.73^{\dagger}$	$40.80^{\dagger}$	$28.16^{\dagger}$	$31.07^{\dagger}$
	Artetxe et al. (2016)	39.27	41.87^{*}	30.62^{*}	31.40^{*}
	Smith et al. (2017)	43.1	$43.33^{\dagger}$	$29.42^{\dagger}$	$35.13^{\dagger}$
	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 word embedding alignment

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**

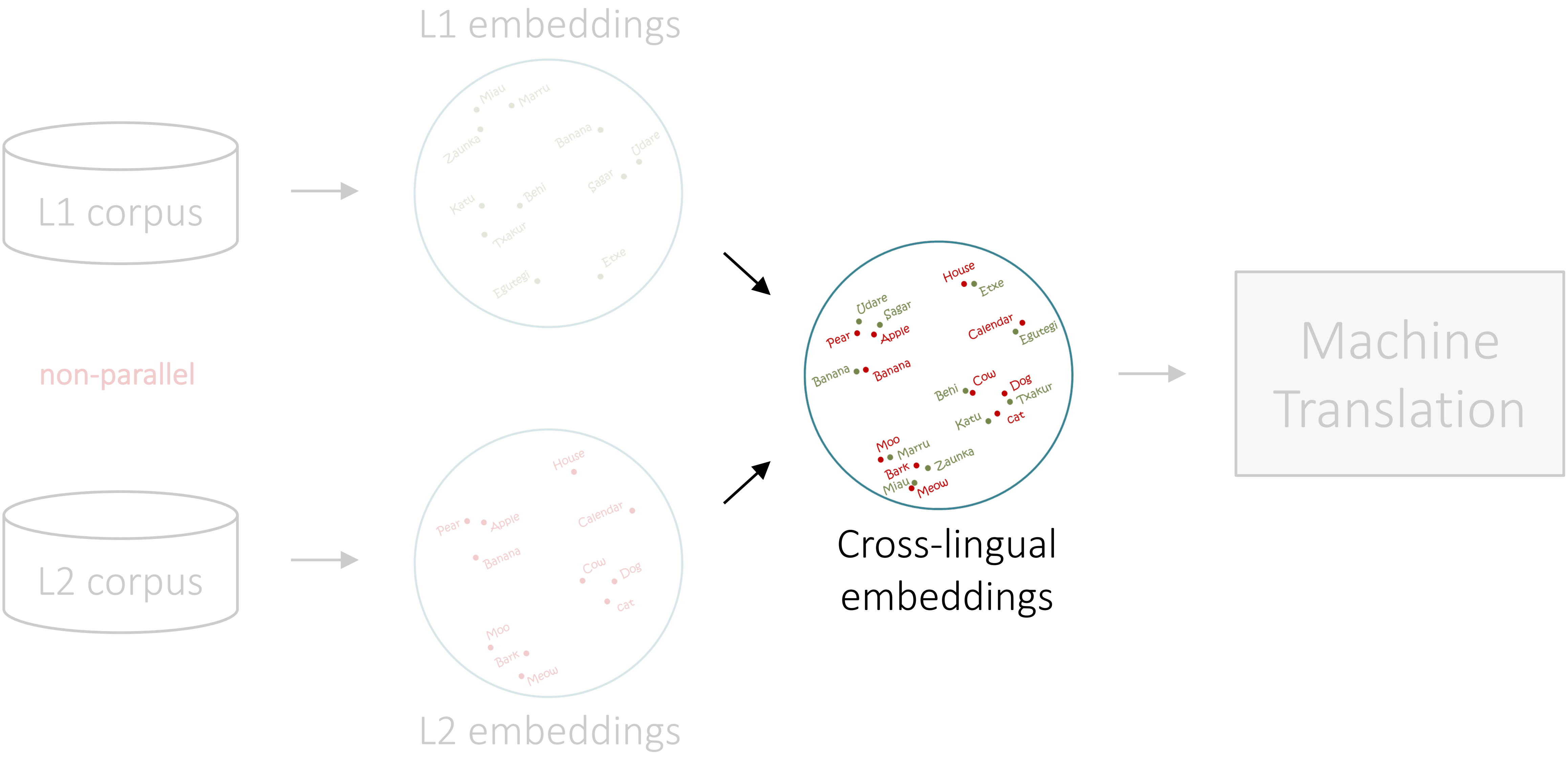
Cross-lingual word embedding alignment

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

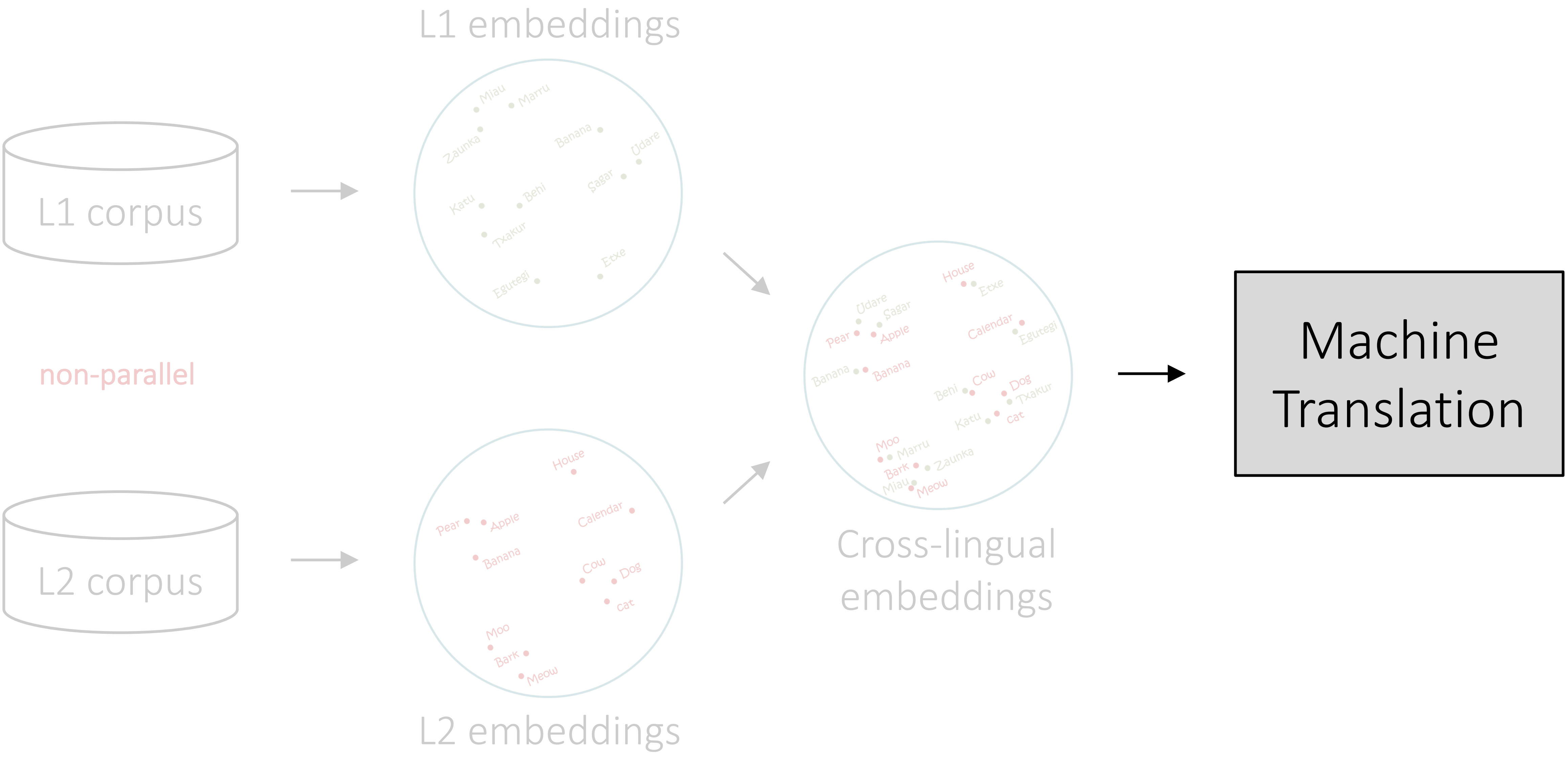
Cross-lingual word embedding alignment

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**

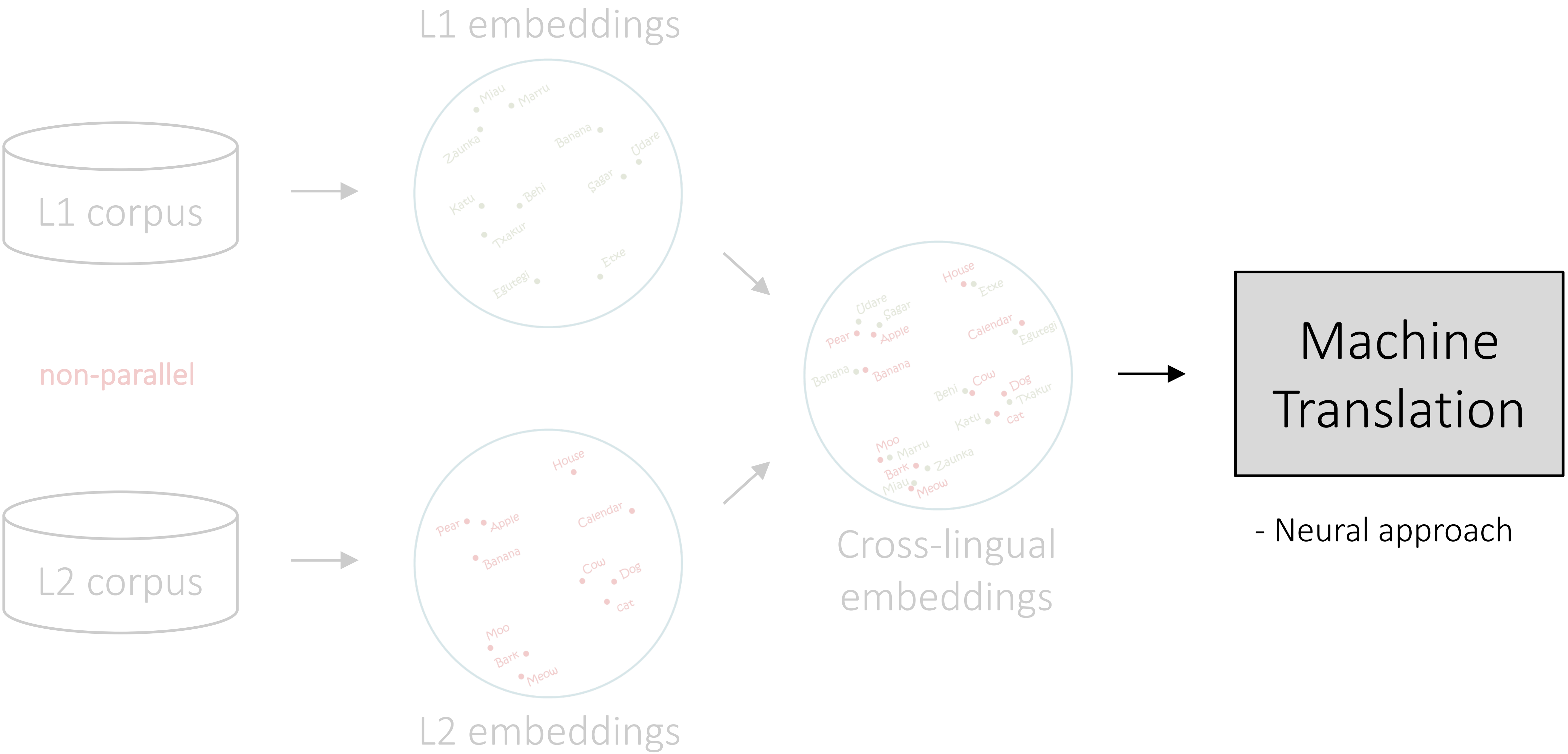
Outline


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

Outline

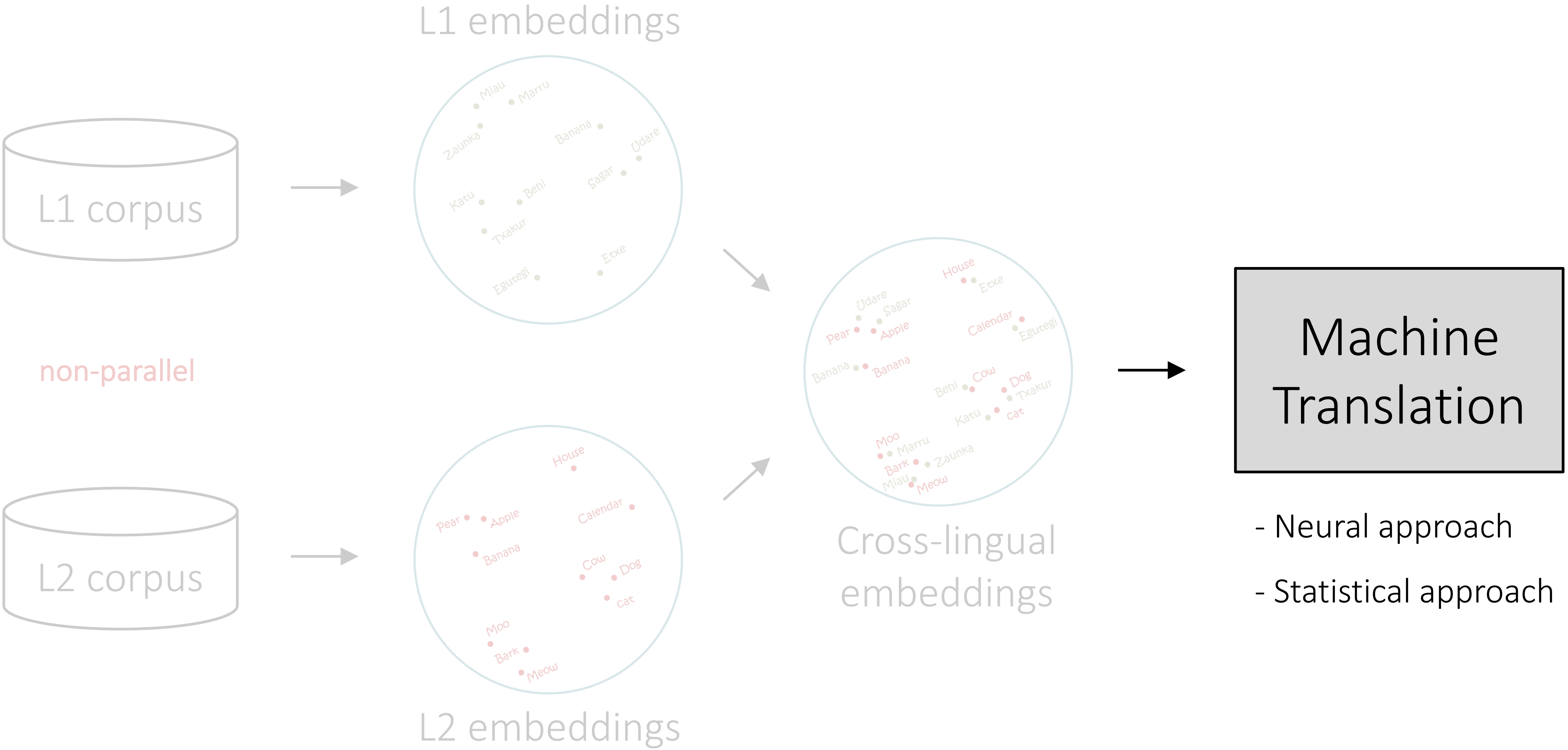

L1 embeddings
L1 corpus
non-parallel
L2 corpus
L2 embeddings
Cross-lingual embeddings
Machine Translation

Outline

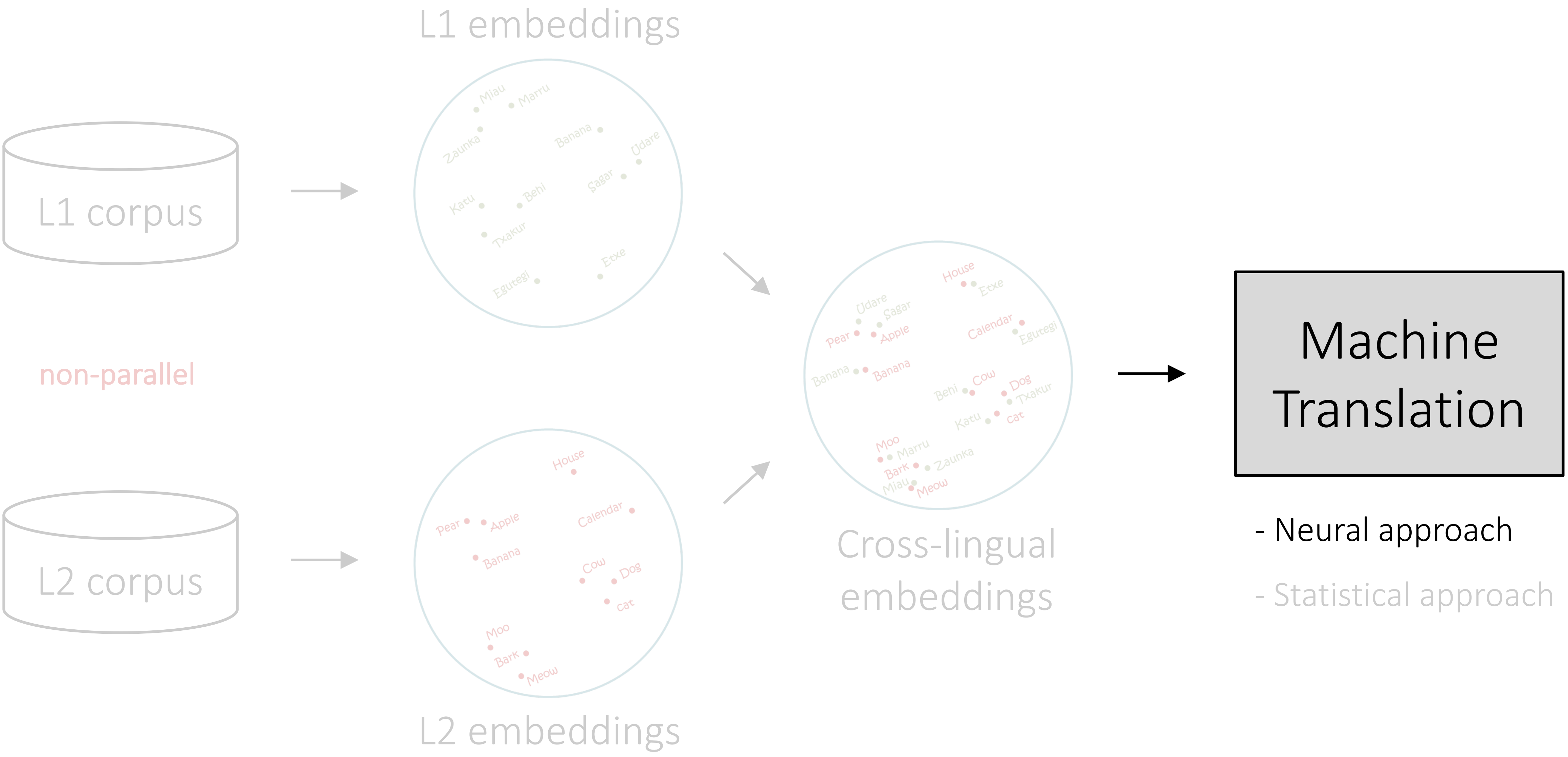


L1 embeddings
L1 corpus
Miau
Marru
Zaunka
Banana
Udare
Katu
Beni
Sagar
Txakur
Etxe
Egutegi
non-parallel
L2 corpus
House
Pear
Apple
Calendar
Banana
Cow
Dog
Cat
Moo
Bark
Meow
L2 embeddings
Cross-lingual
embeddings
Machine
Translation
- Neural approach

Outline

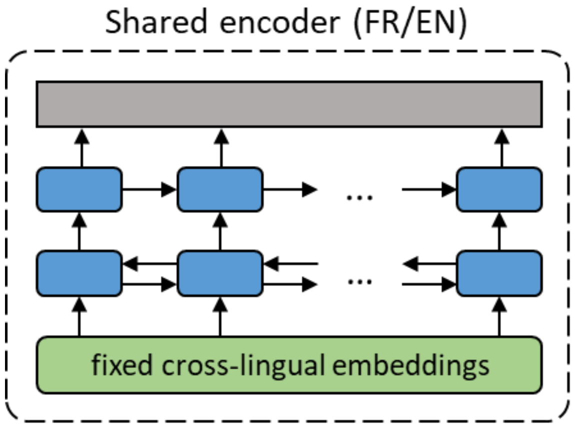

L1 embeddings
L1 corpus
non-parallel
L2 corpus
L2 embeddings
Cross-lingual embeddings
Machine Translation
- Neural approach
- Statistical approach

Outline

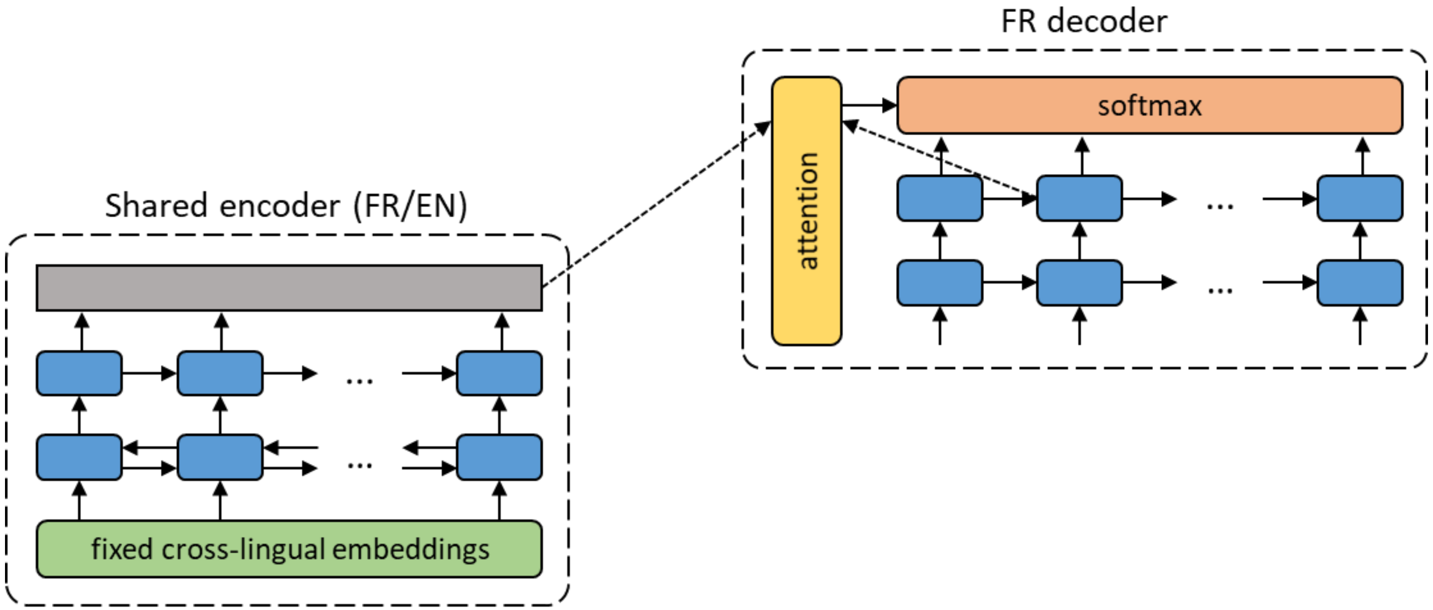

L1 embeddings
L1 corpus
Miau
Marru
Zaunka
Banana
Udare
Katu
Beni
Sagar
Txakur
Etxe
Egutegi
non-parallel
L2 corpus
House
Pear
Apple
Calendar
Banana
Cow
Dog
cat
Moo
Bark
Meow
L2 embeddings
Cross-lingual embeddings
Machine Translation
- Neural approach
- Statistical approach

Unsupervised neural machine translation

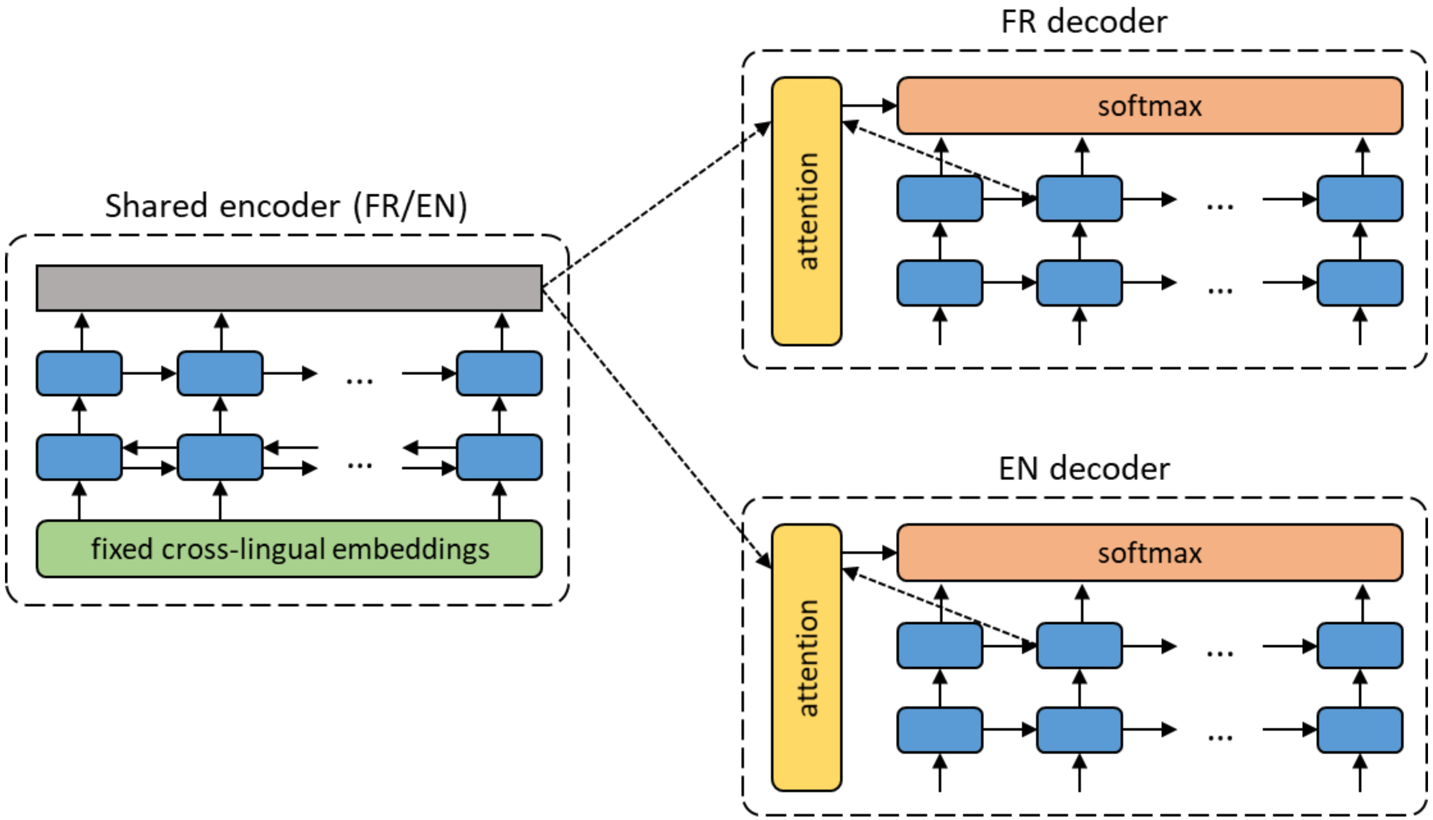
Unsupervised neural machine translation



Unsupervised neural machine translation

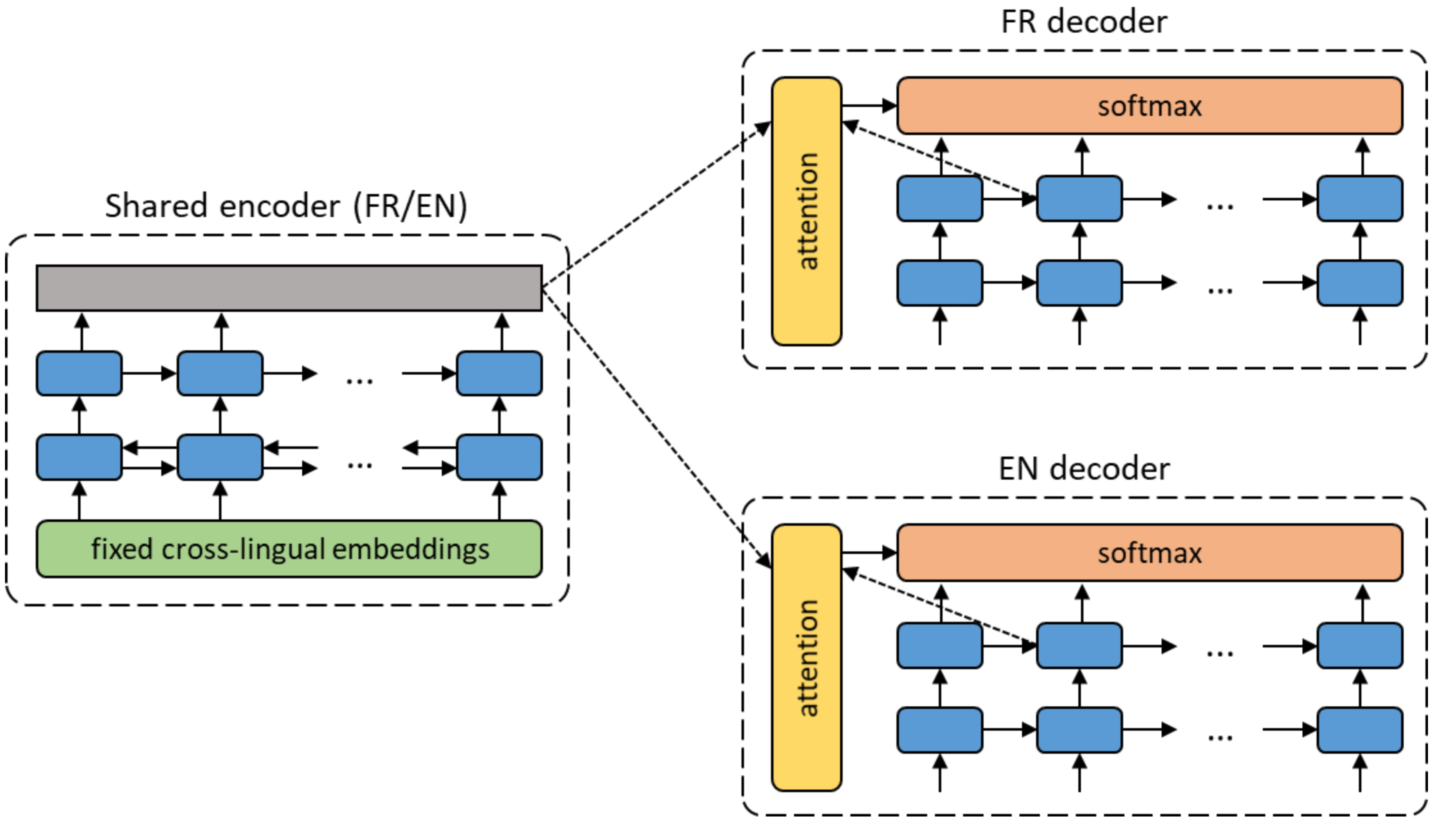


Unsupervised neural machine translation



Unsupervised neural machine translation

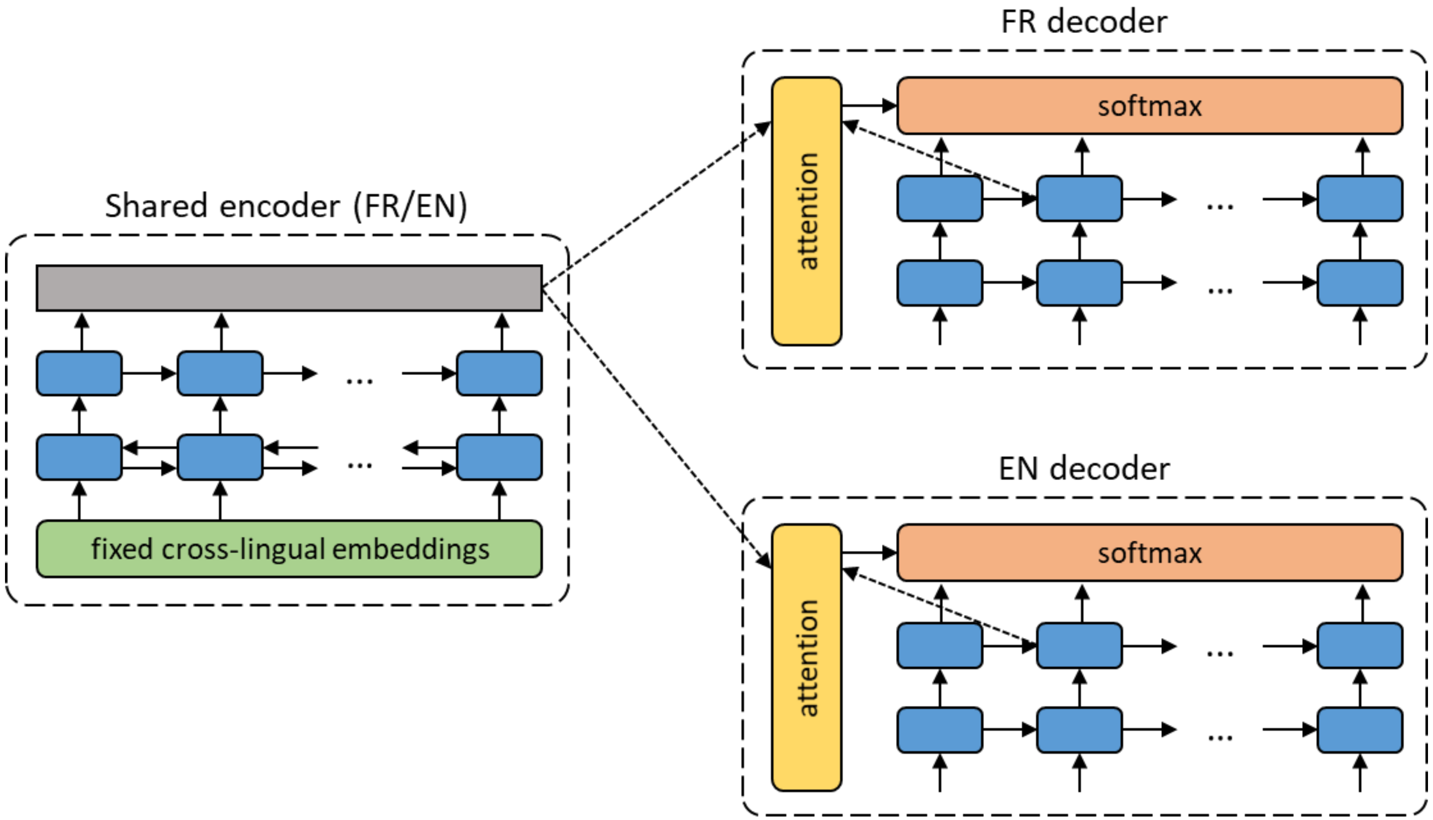
Training



Unsupervised neural machine translation

Training

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

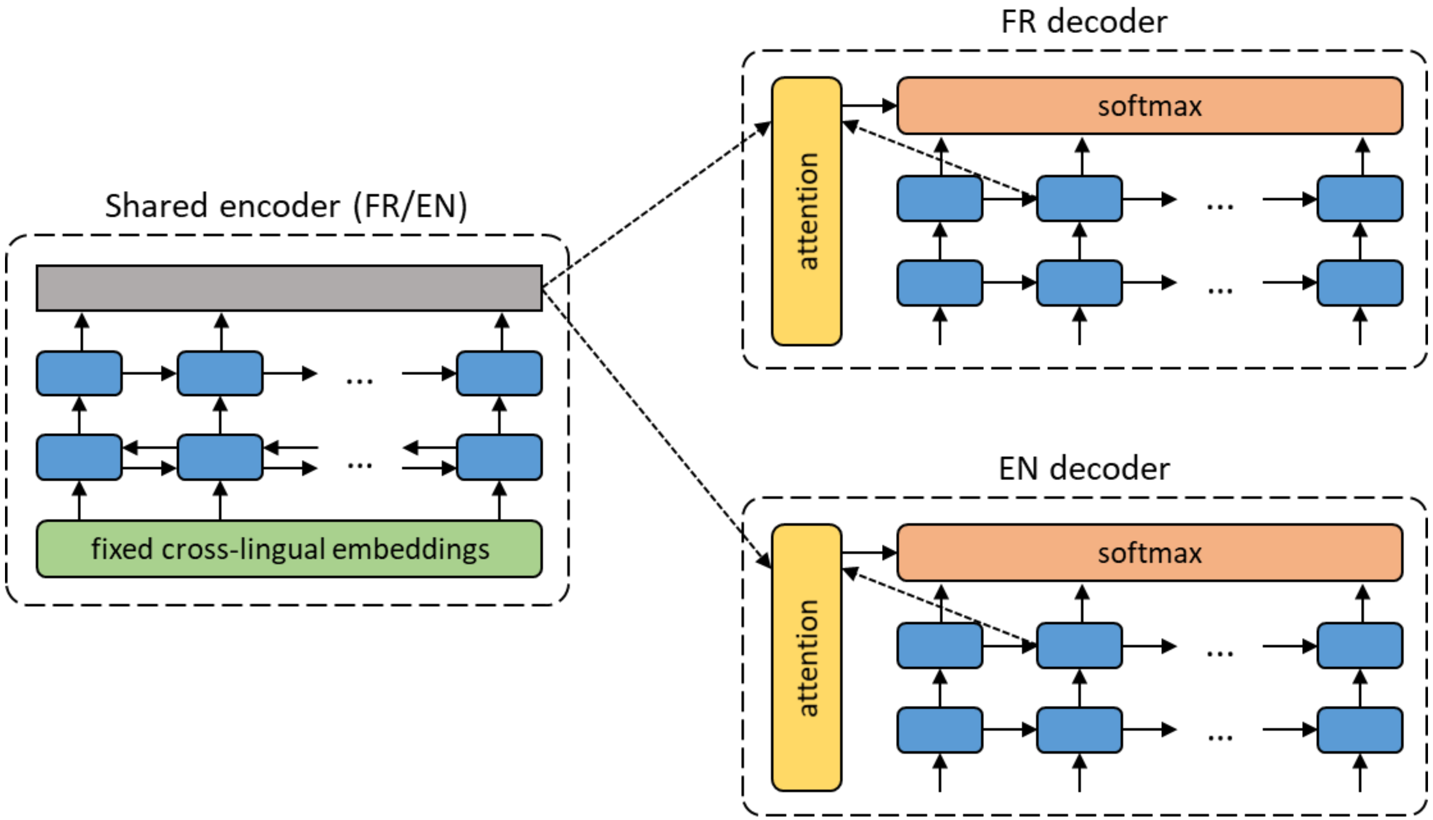


Unsupervised neural machine translation

Training

- Supervised

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

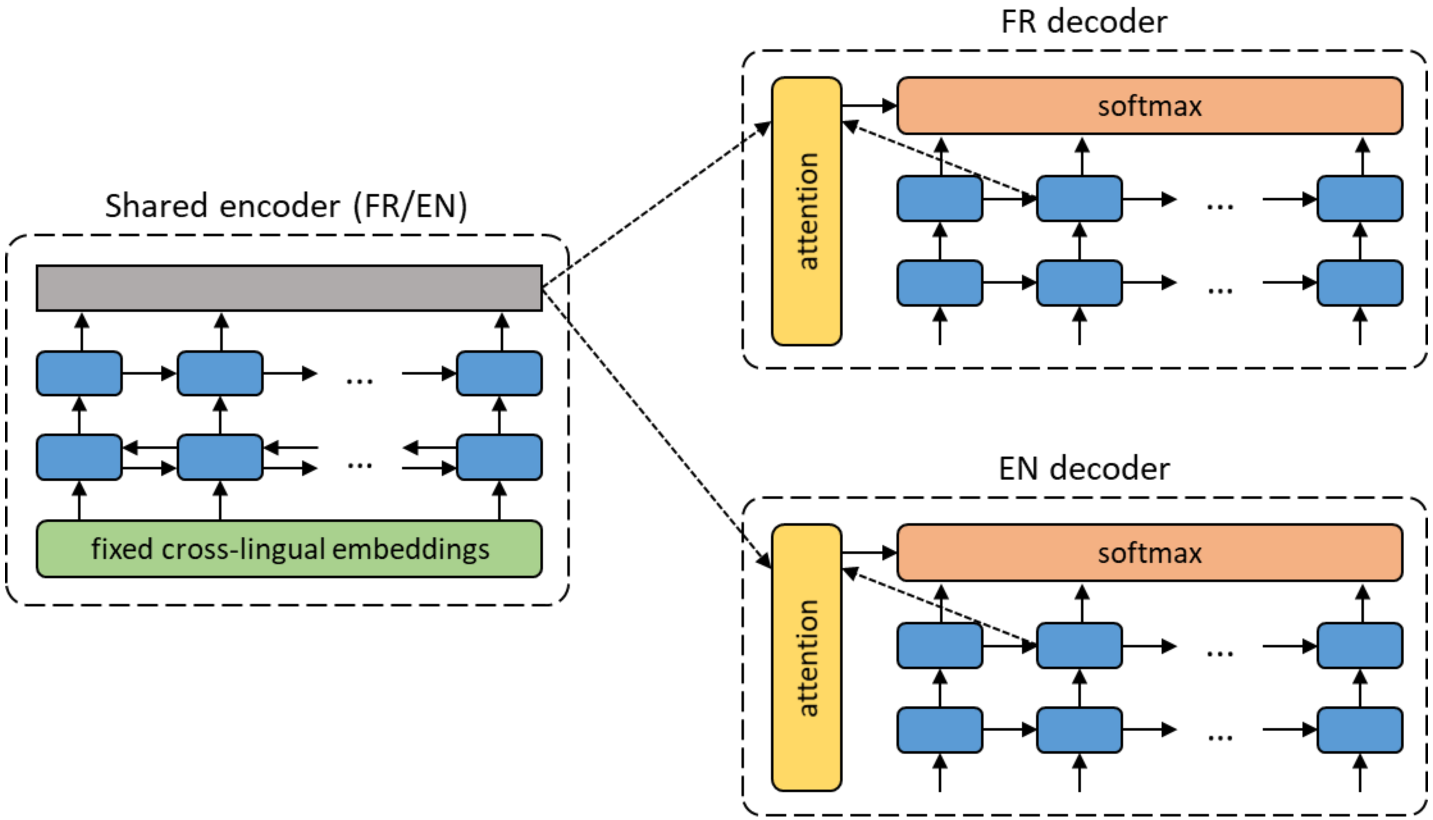


Unsupervised neural machine translation

Training

- Supervised

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



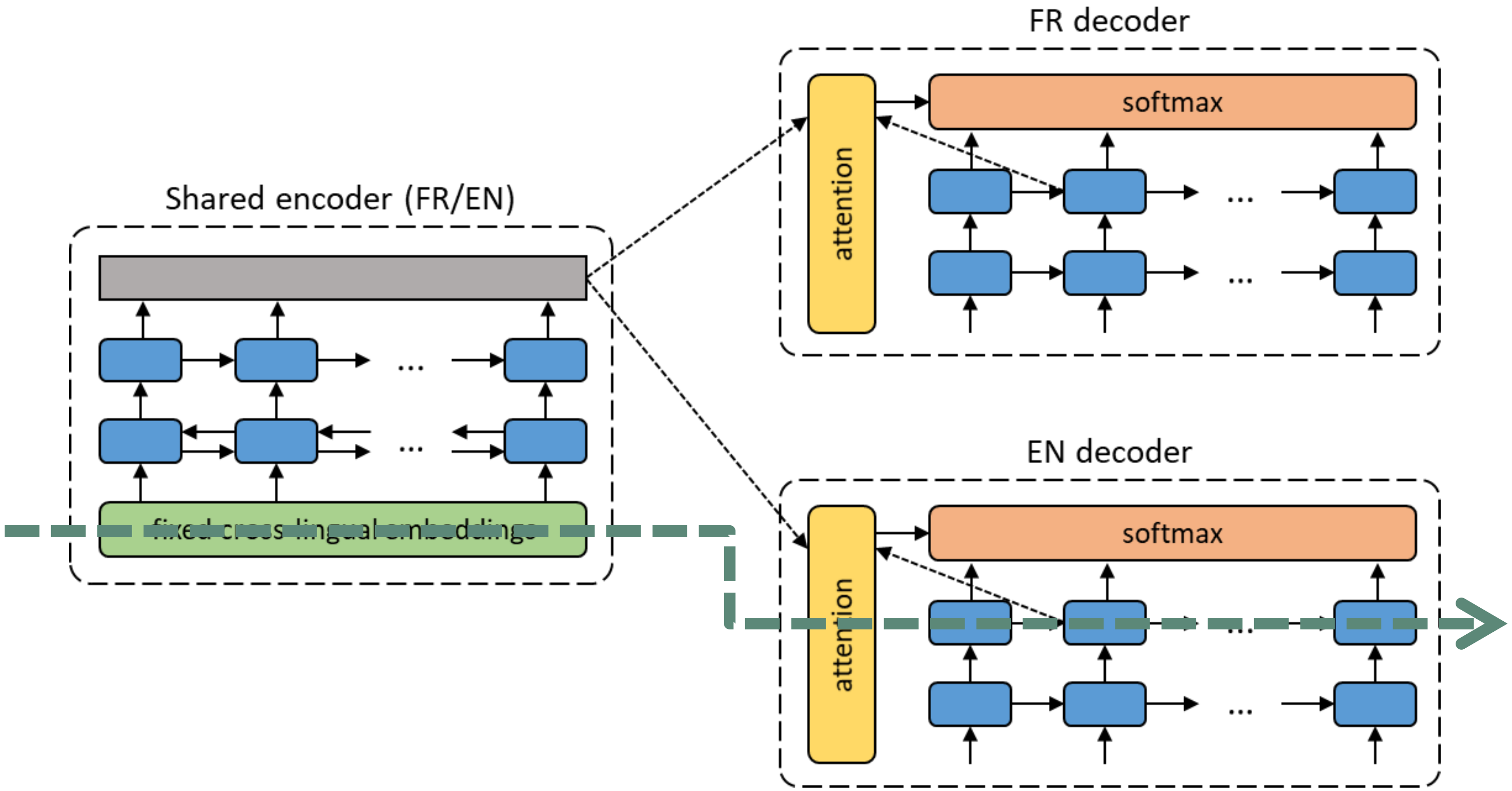
There was a shooting in Los Angeles International Airport.

Unsupervised neural machine translation

Training

- Supervised

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



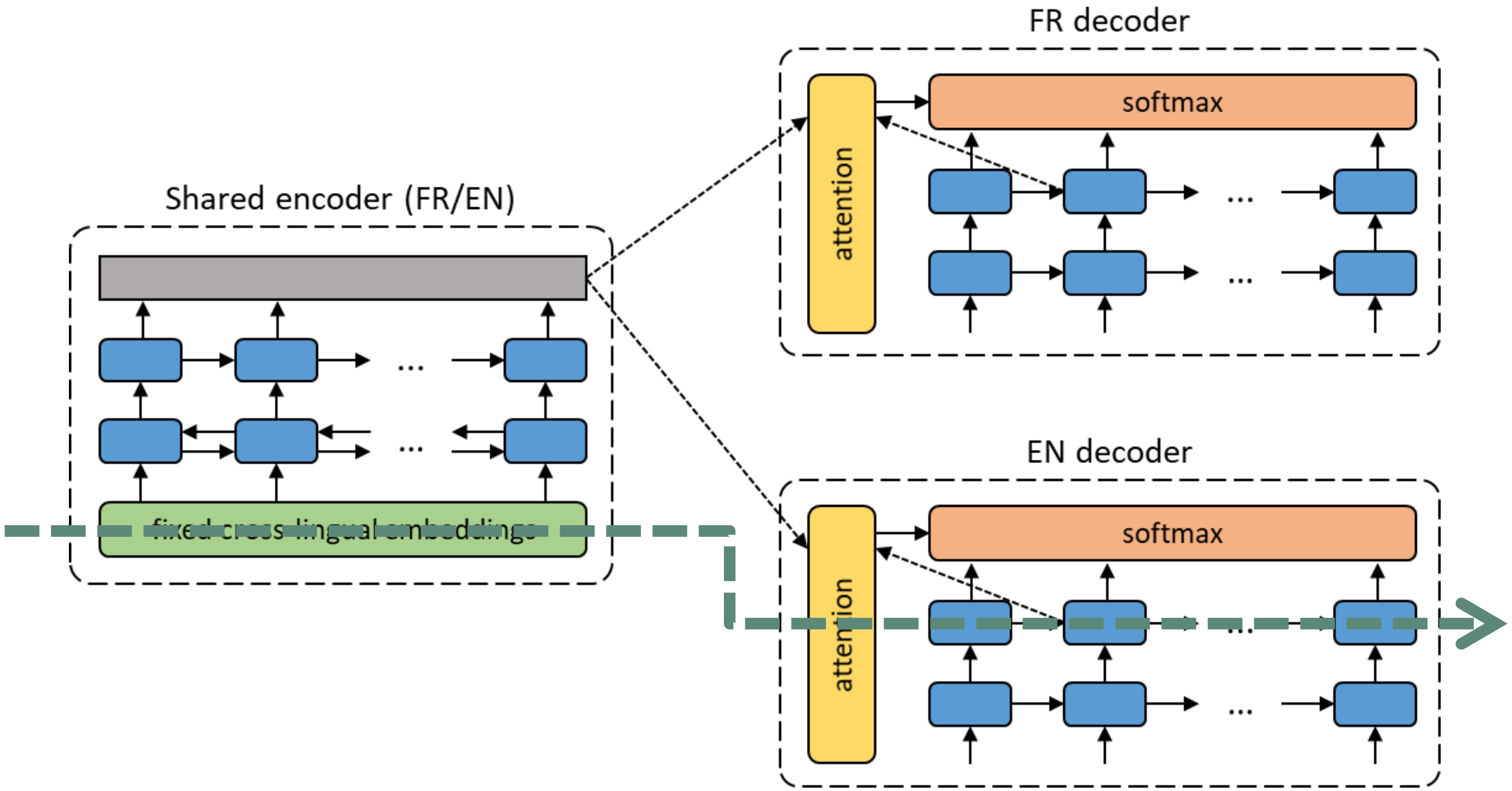
There was a shooting in Los Angeles International Airport.

Unsupervised neural machine translation

Training

- Supervised

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



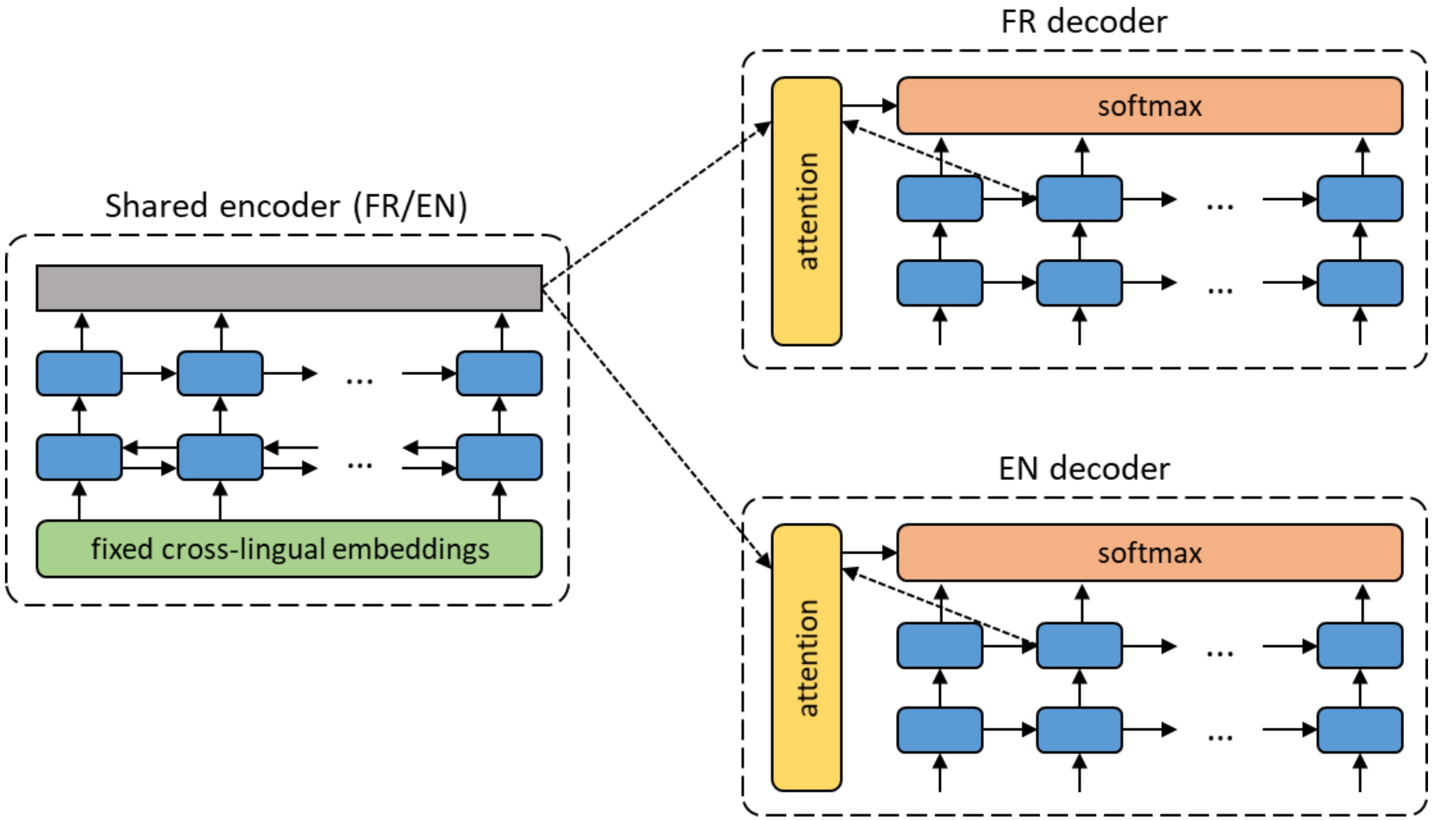
~~There was a shooting in Los Angeles International Airport.~~

Unsupervised neural machine translation

Training

- Supervised

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

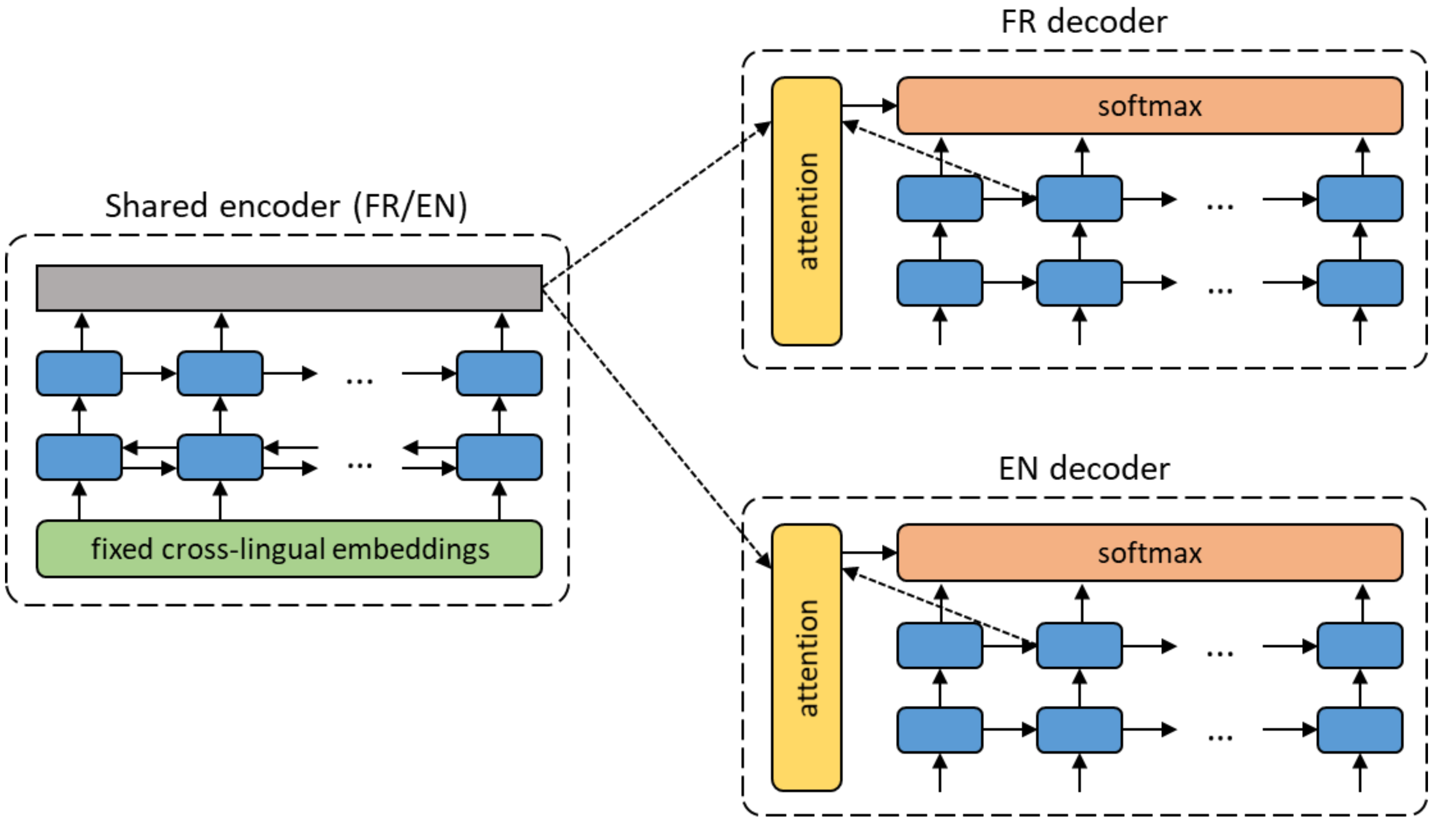


Unsupervised neural machine translation

Training

- ~~Supervised~~

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

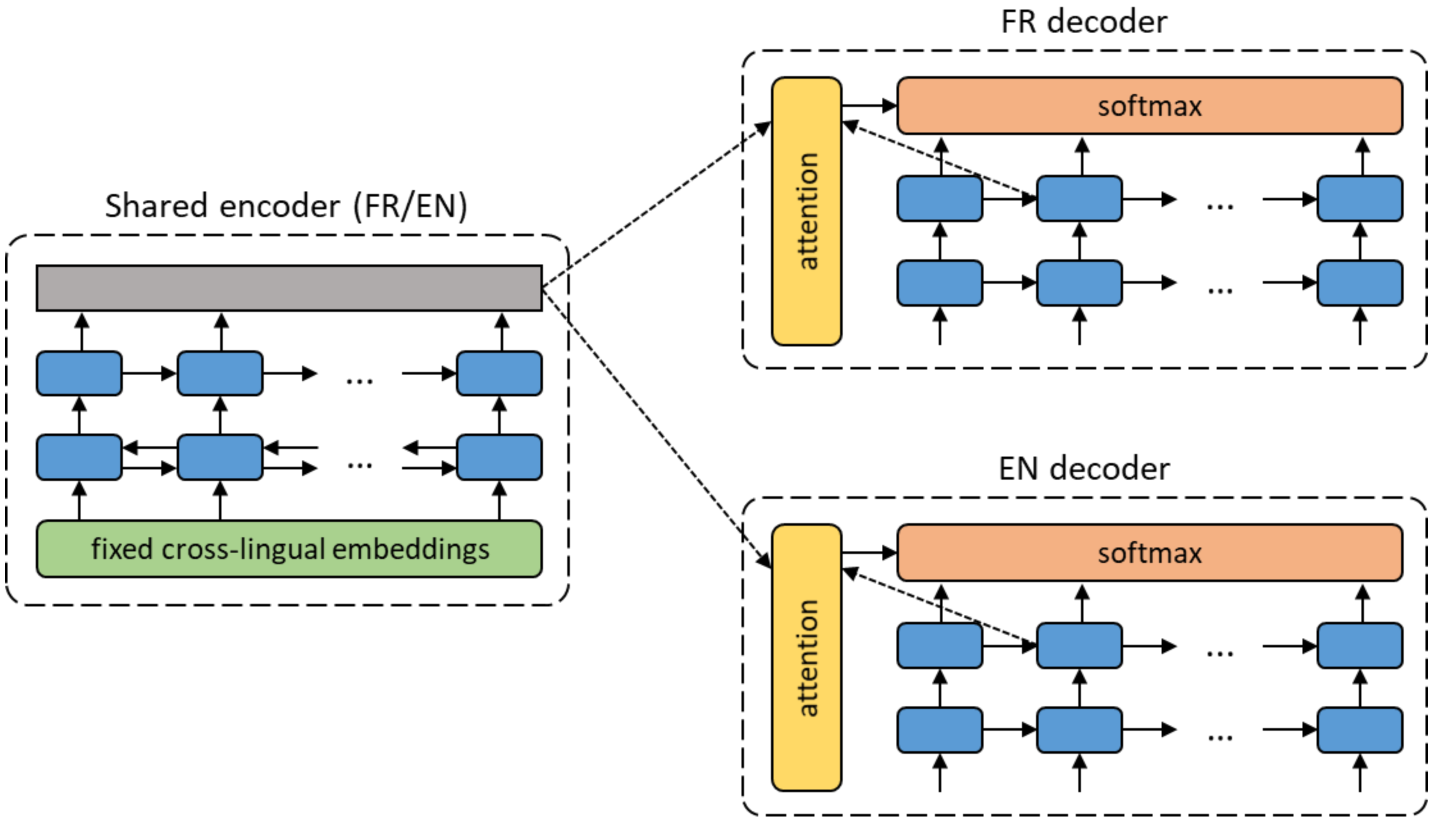


Unsupervised neural machine translation

Training

- ~~Supervised~~

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



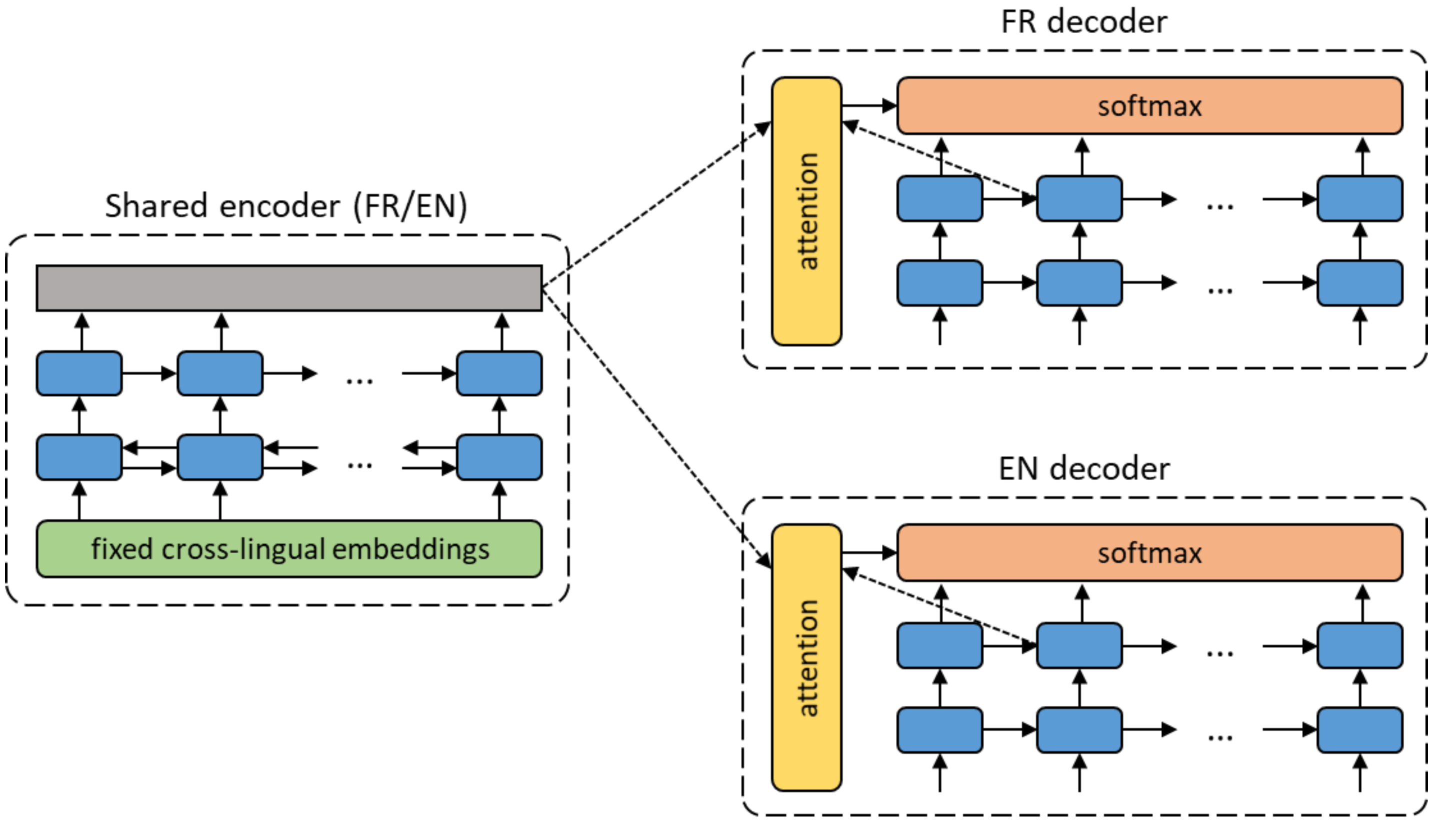
Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising

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



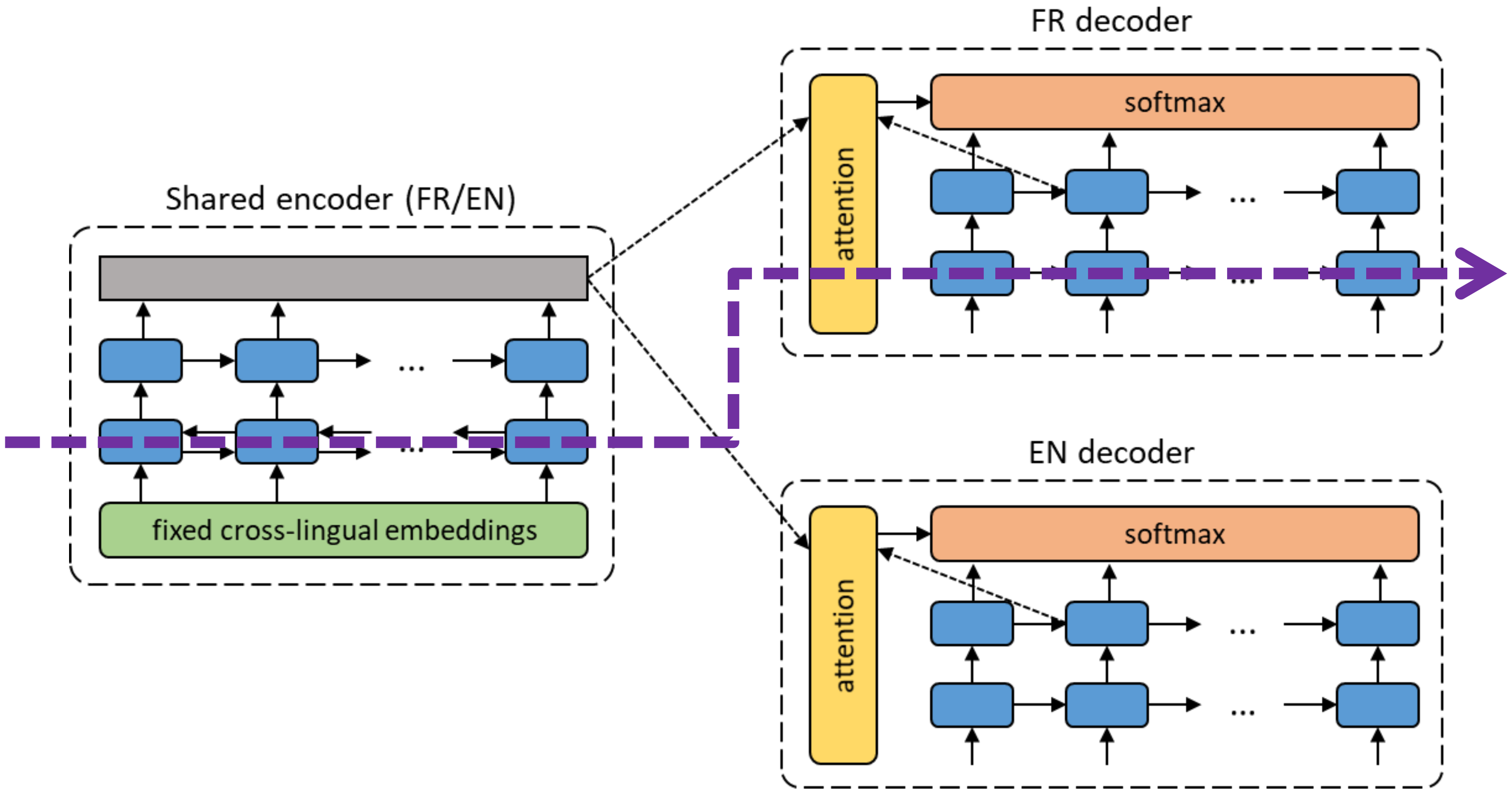
Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising

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



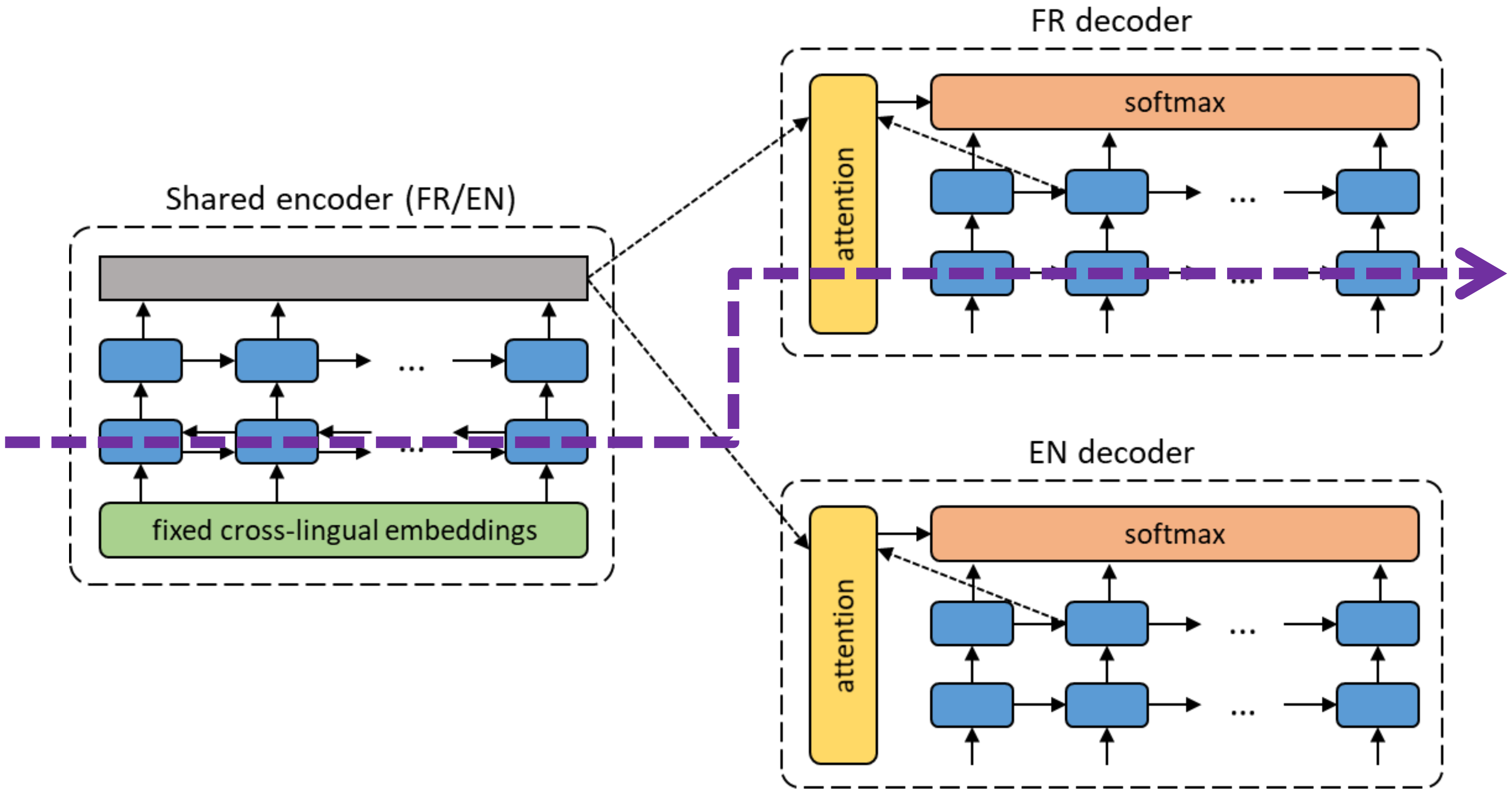
Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising

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



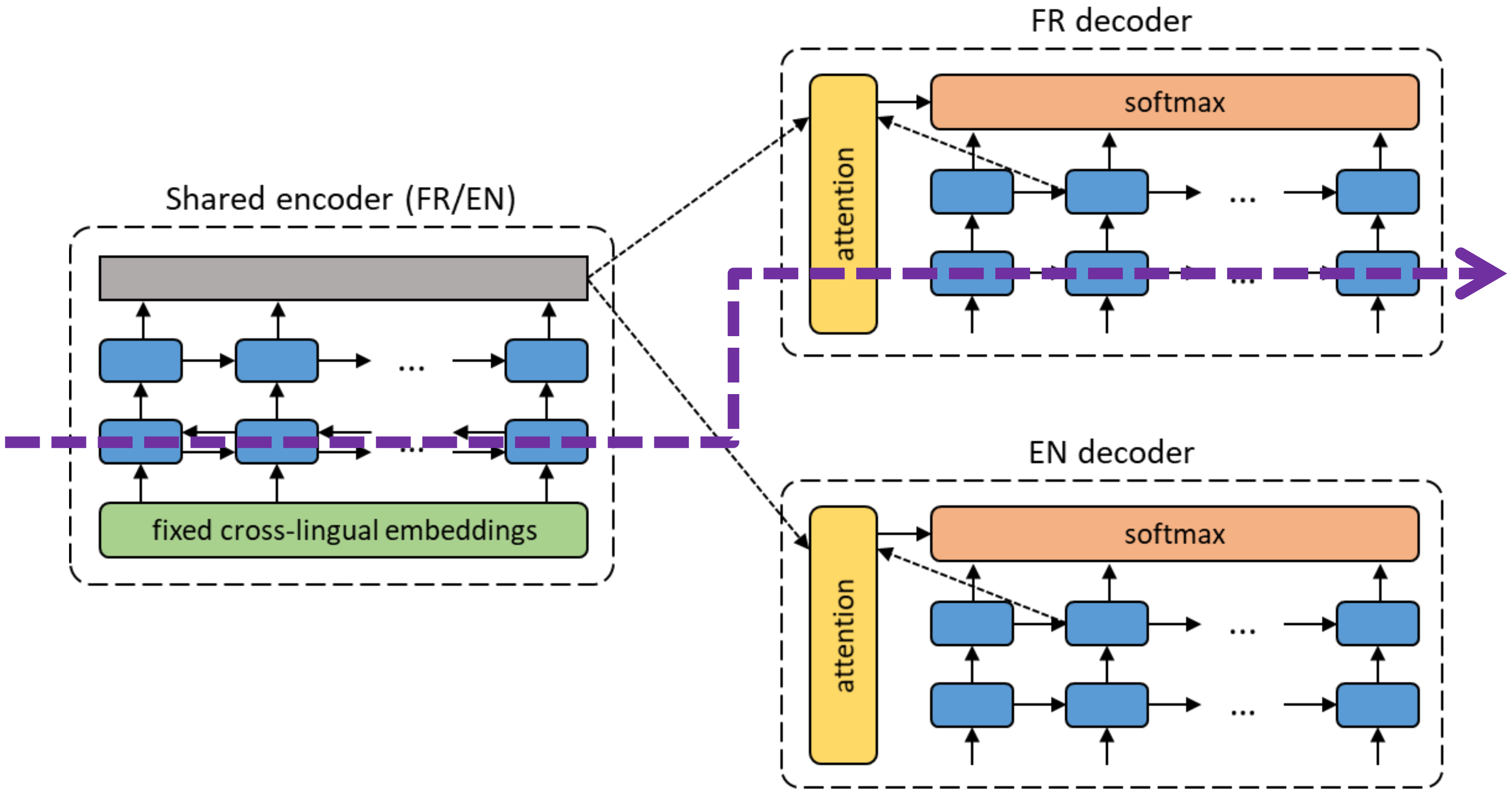
Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising
- Backtranslation

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

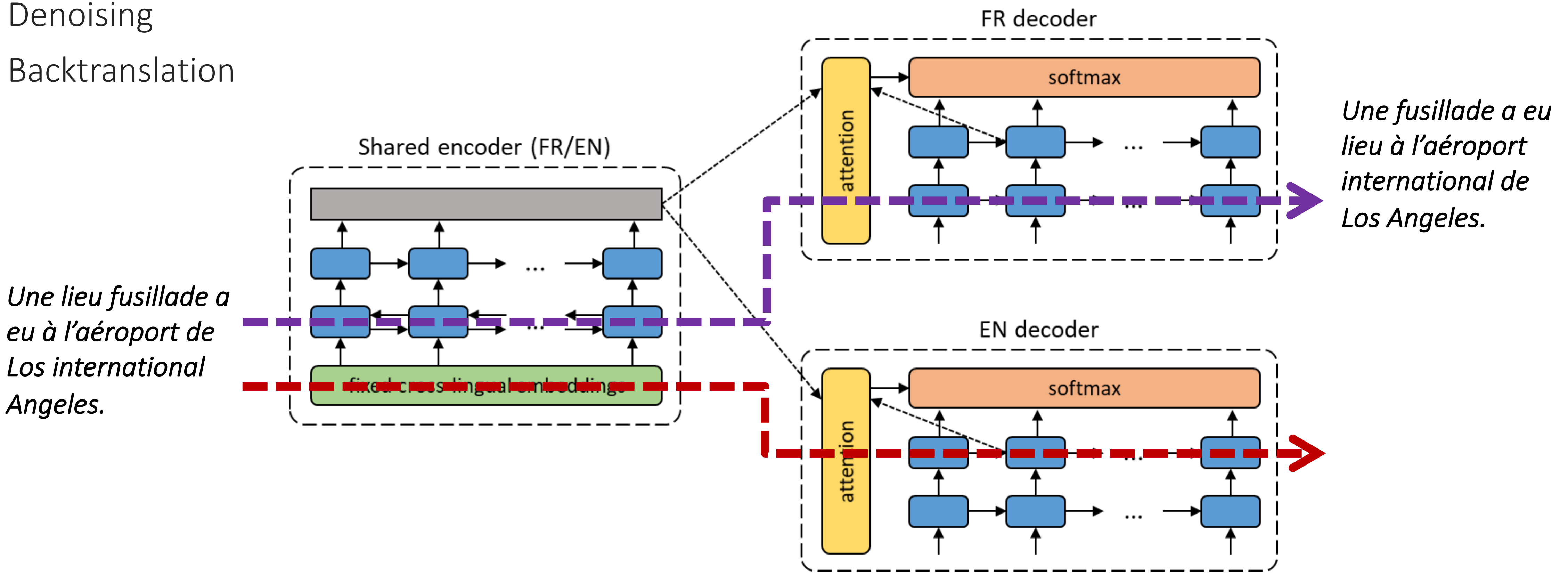


Une fusillade a eu lieu à l'aéroport international de Los Angeles.

Unsupervised neural machine translation

Training

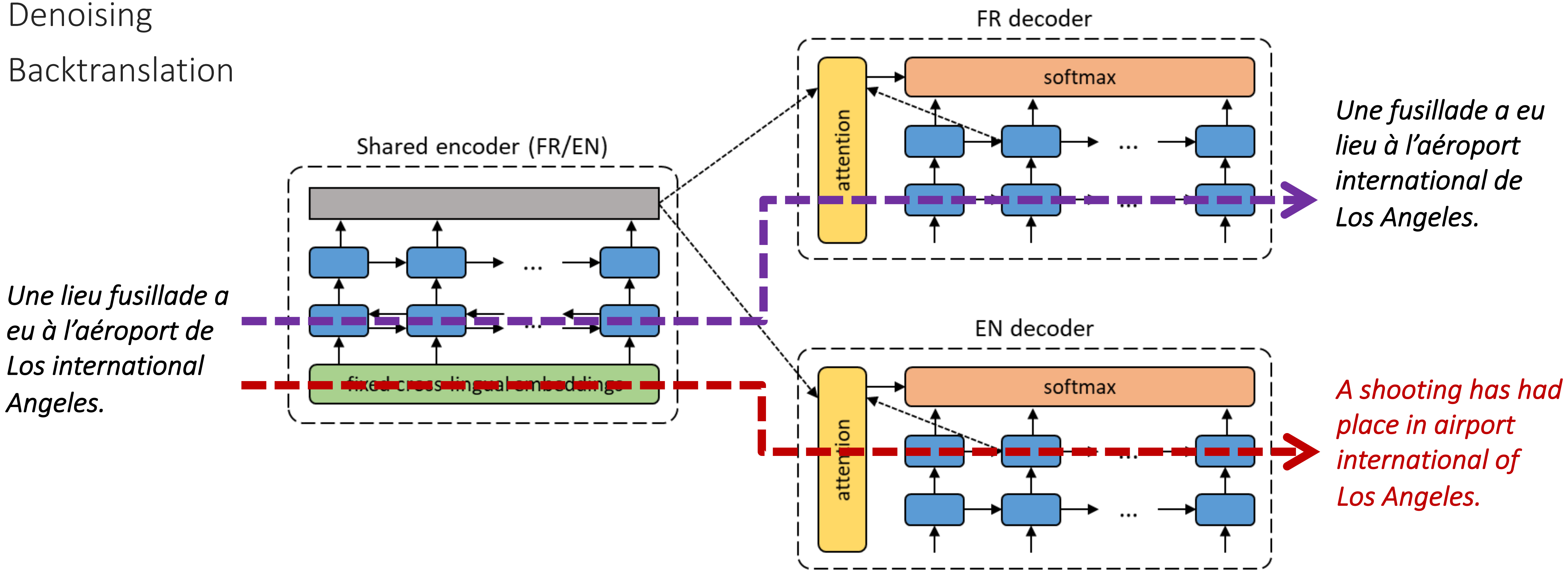
- ~~Supervised~~
- Denoising
- Backtranslation



Unsupervised neural machine translation

Training

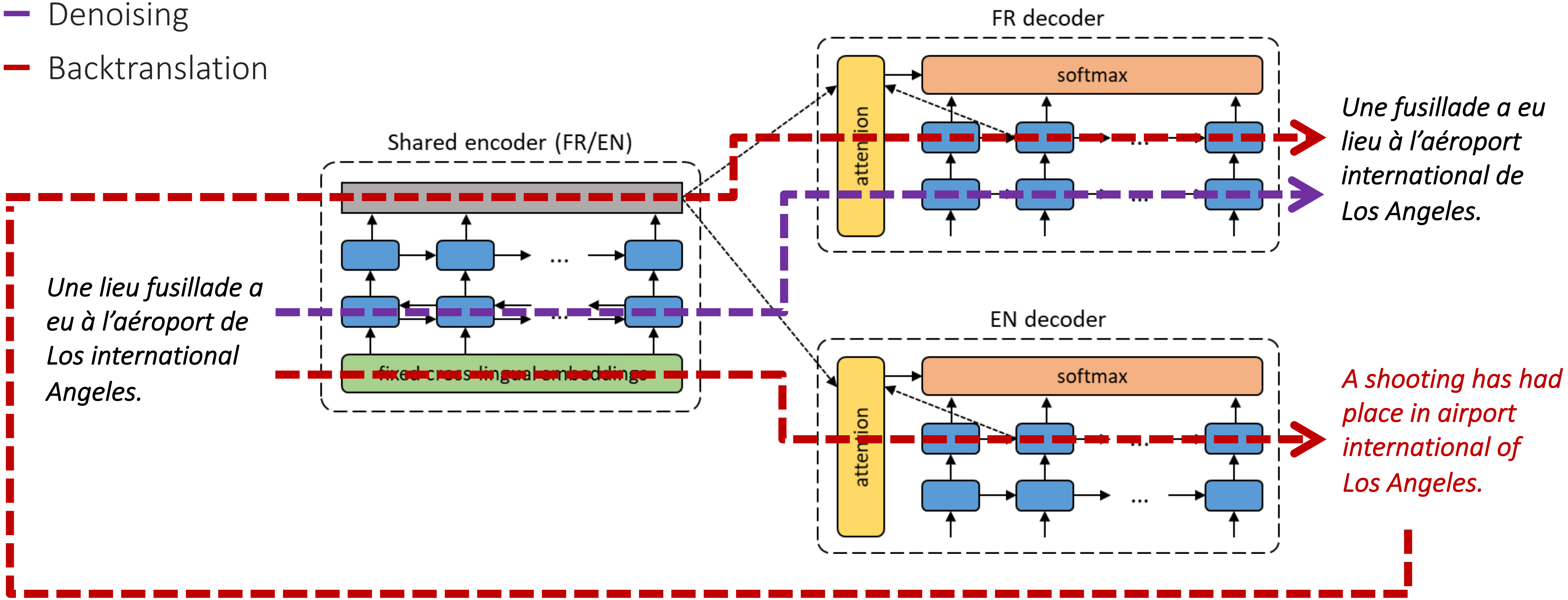
- ~~Supervised~~
- Denoising
- Backtranslation



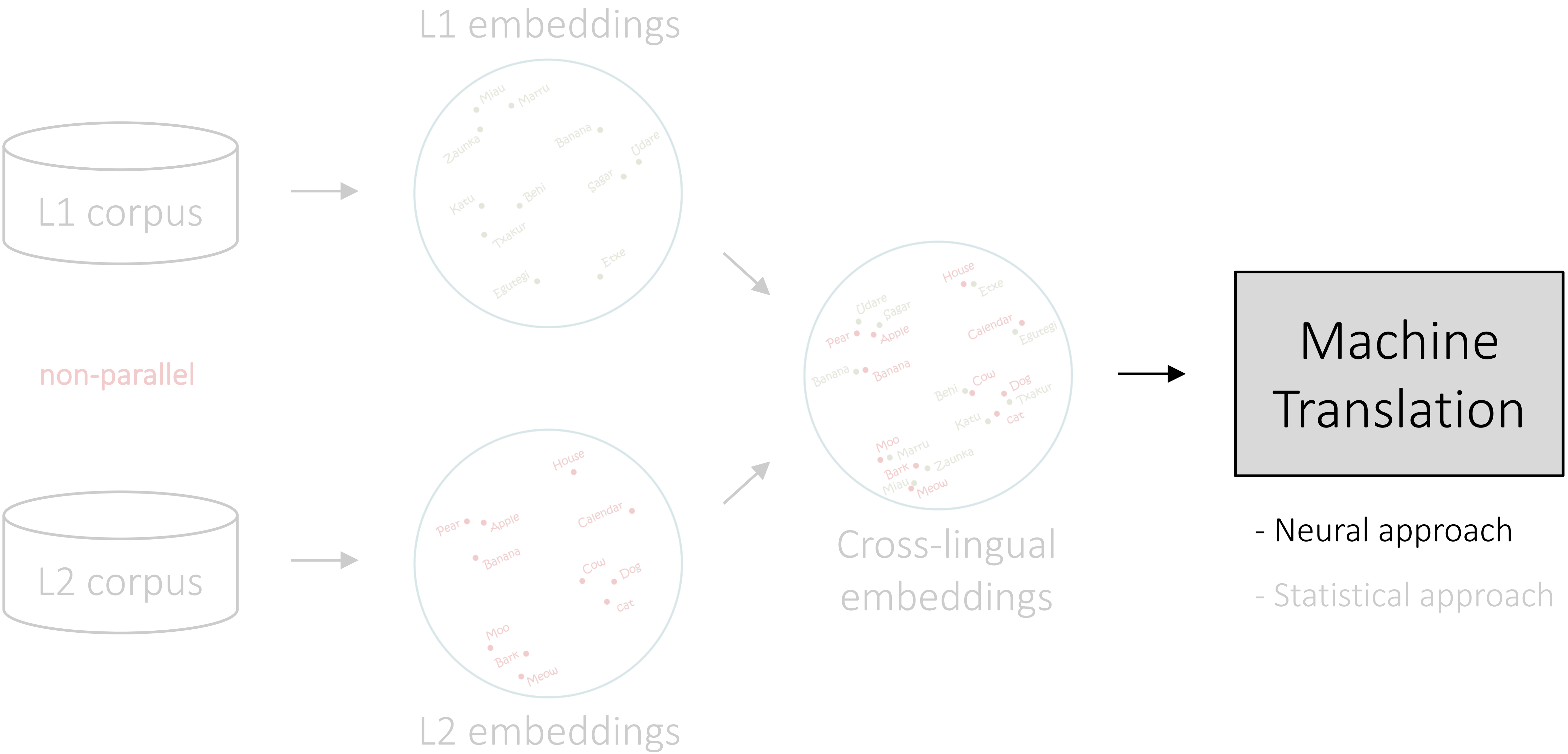
Unsupervised neural machine translation

Training

- ~~Supervised~~
- Denoising
- Backtranslation

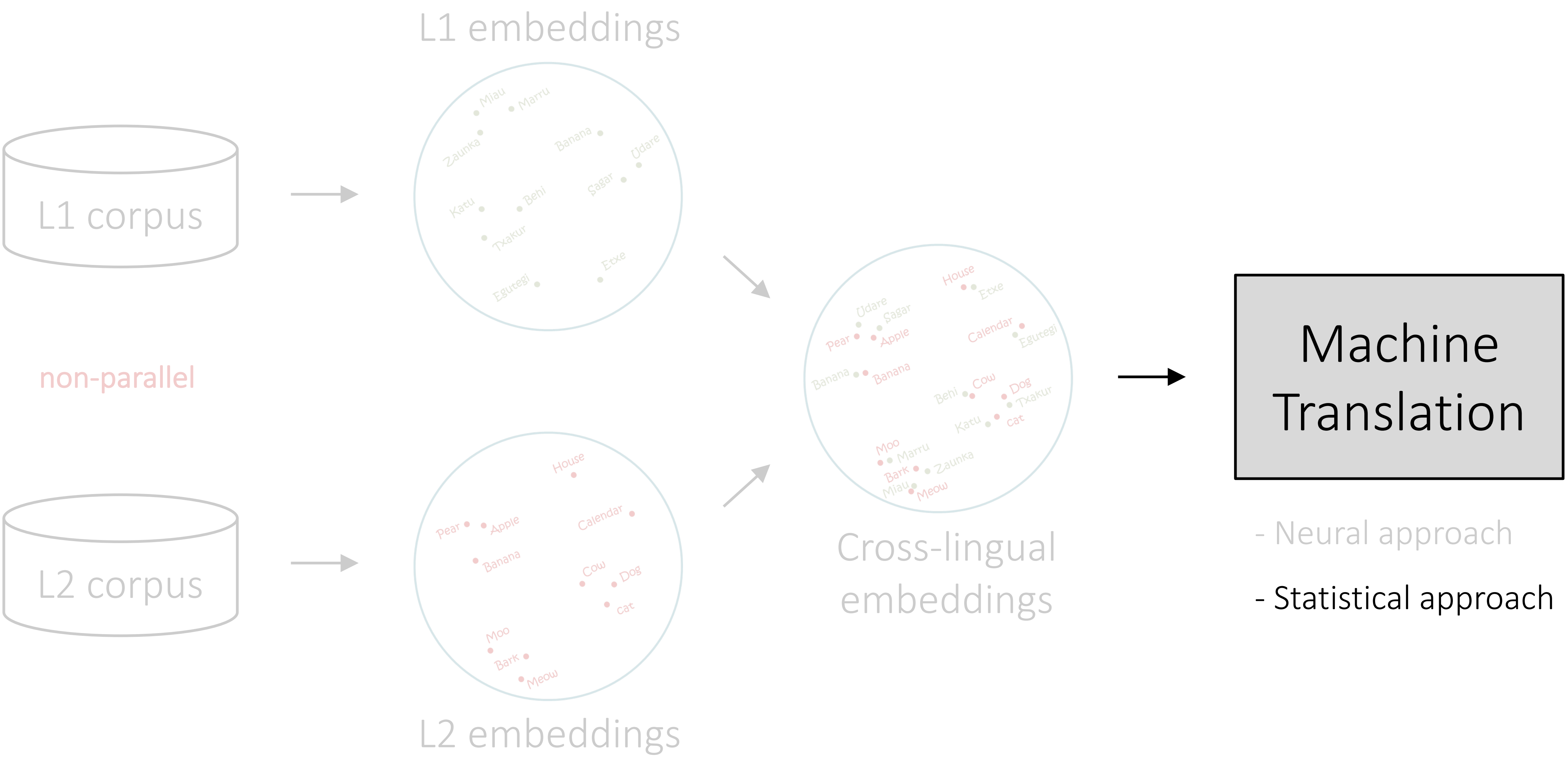


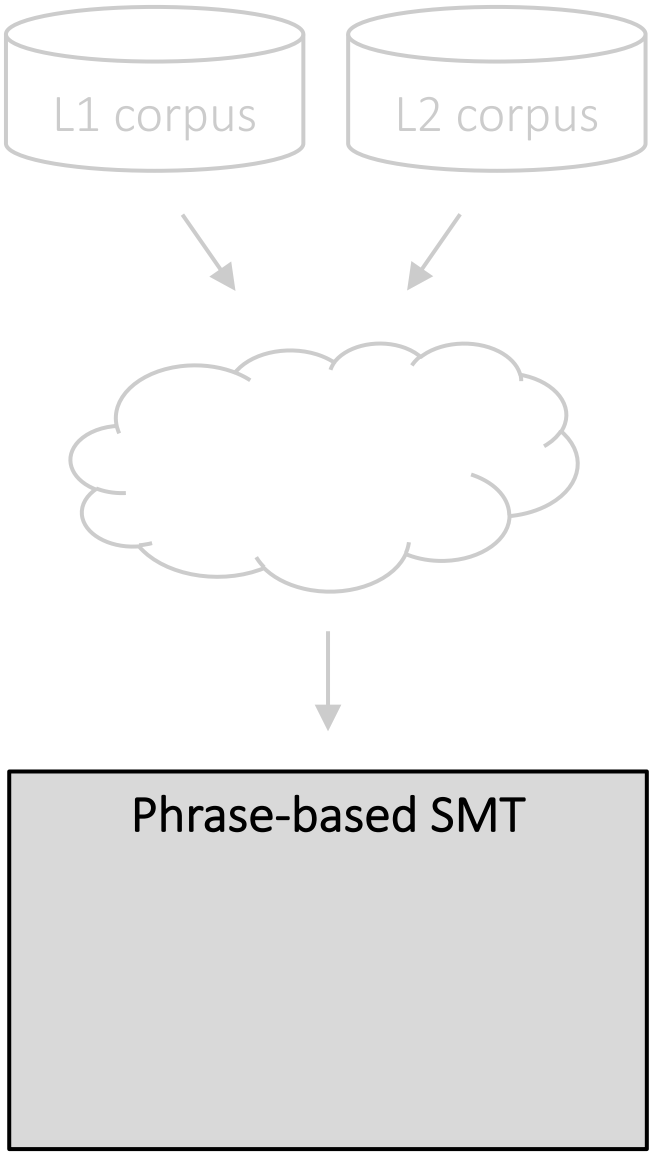
Outline



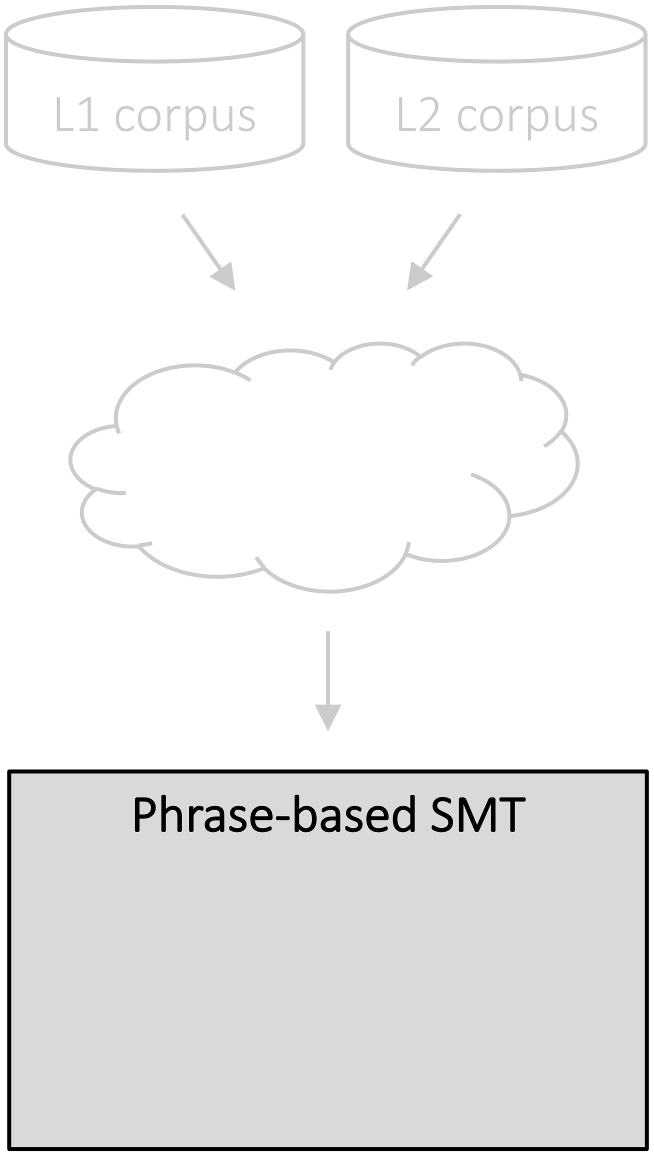
L1 embeddings
L1 corpus
non-parallel
L2 corpus
L2 embeddings
Cross-lingual embeddings
Machine Translation
- Neural approach
- Statistical approach

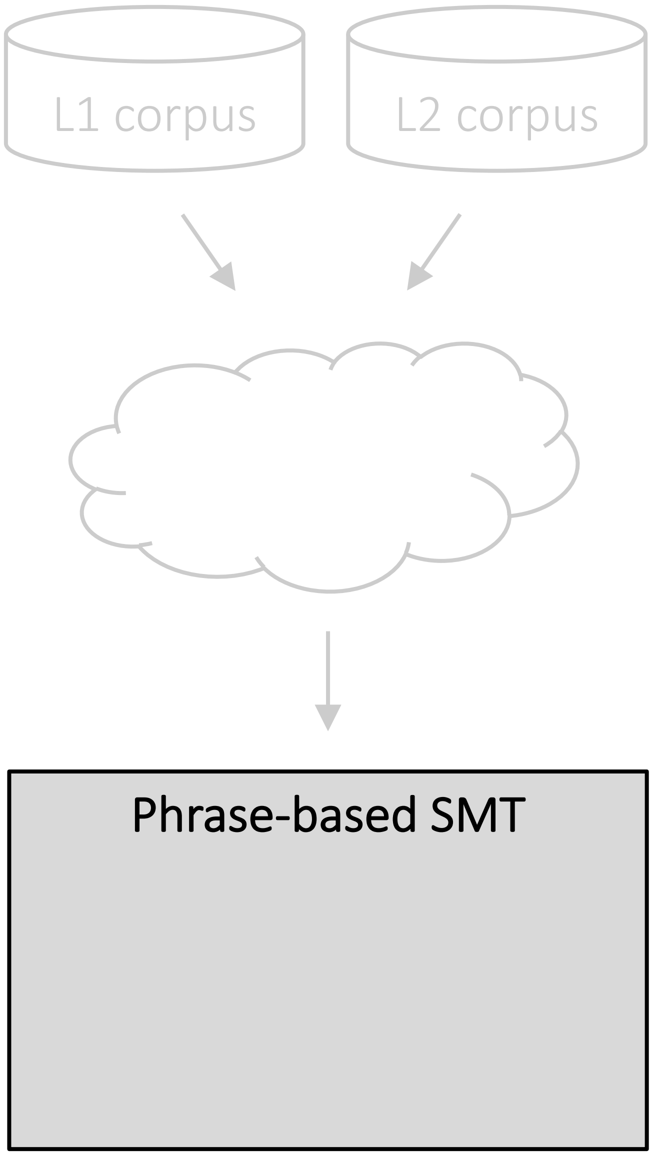
Outline


L1 embeddings
L1 corpus
non-parallel
L2 corpus
L2 embeddings
Cross-lingual embeddings
Machine Translation
- Neural approach
- Statistical approach


L1 corpus
L2 corpus
Phrase-based SMT

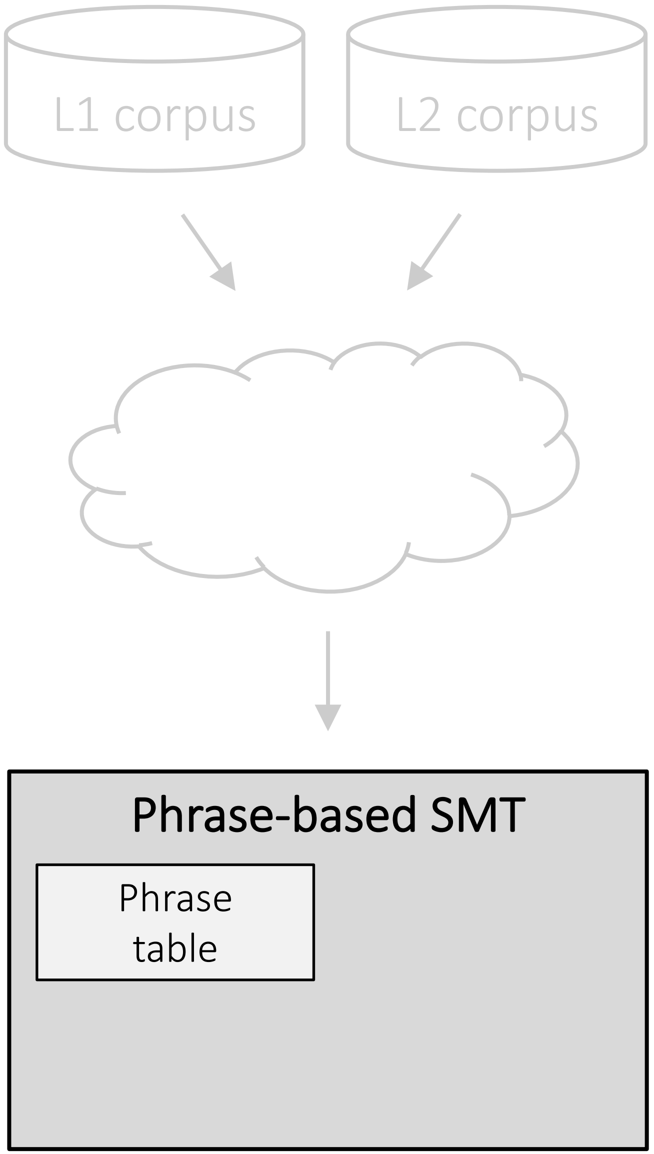
Phrase-based SMT





Phrase-based SMT

Log-linear model combining

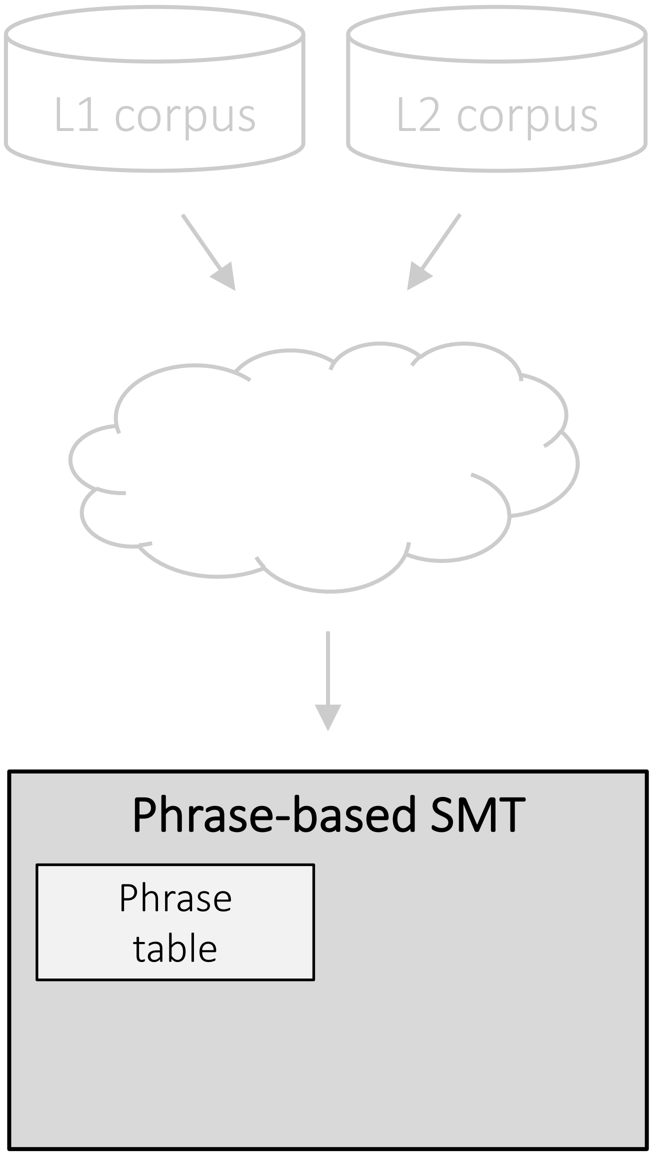


Phrase-based SMT

Log-linear model combining

- Phrase table

Phrase-based SMT

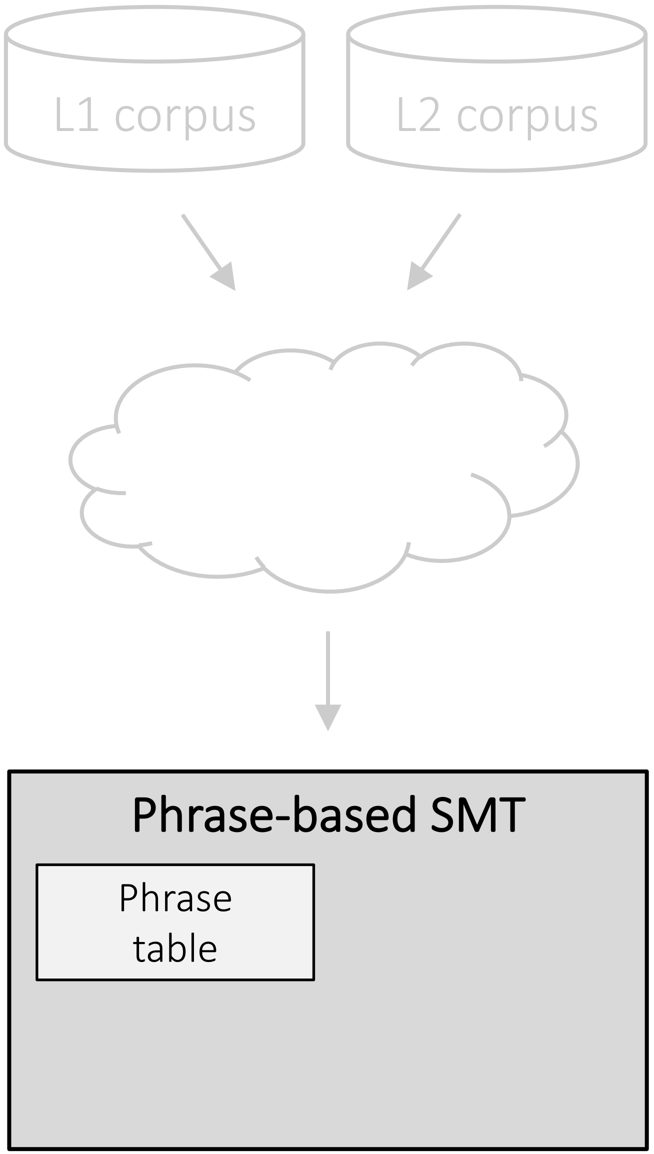


Log-linear model combining

- 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

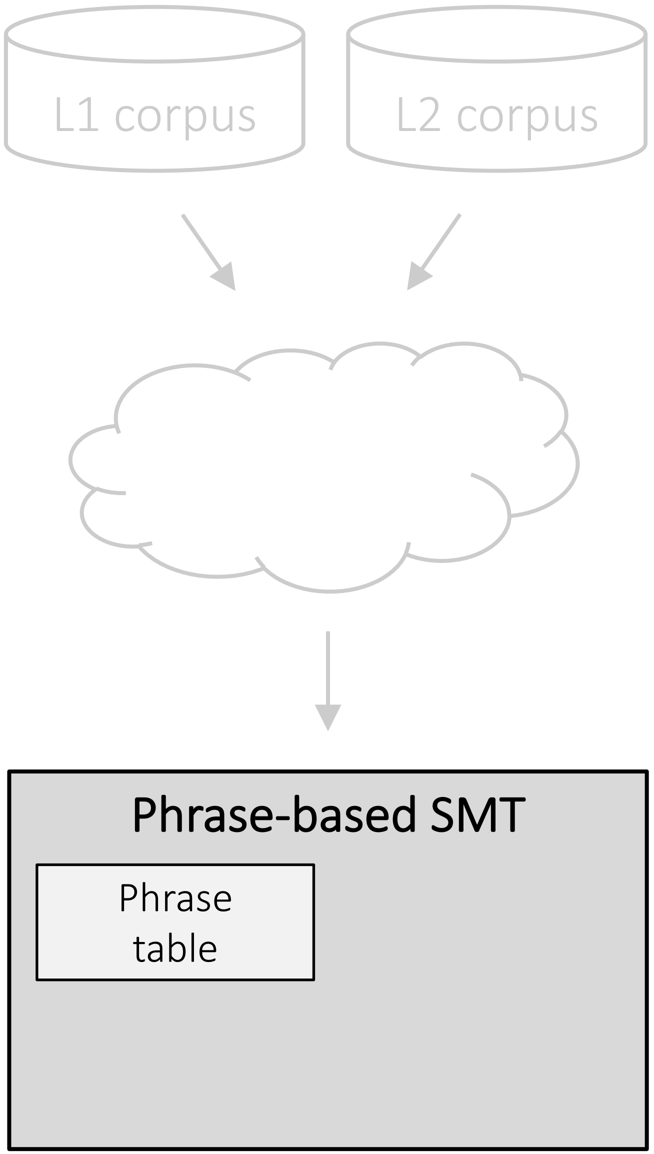
Log-linear model combining

- Phrase table
  - Direct/inverse translation probabilities

		$\phi(\bar{f}\vert\bar{e})$	$\phi(\bar{e}\vert\bar{f})$
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

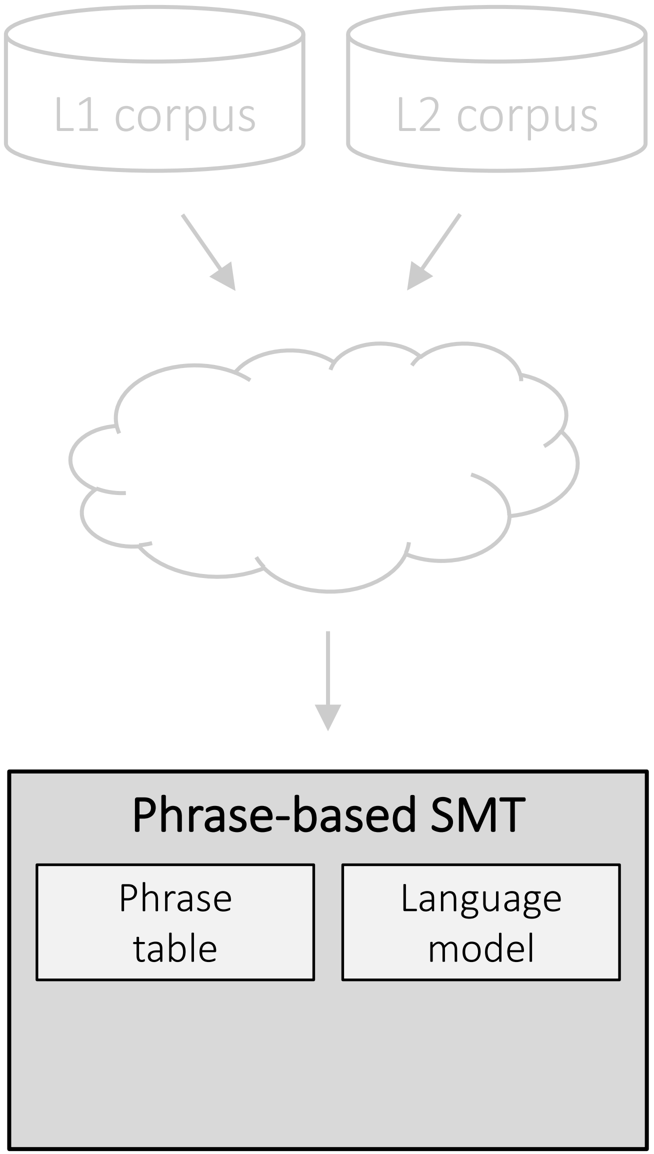


Log-linear model combining

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

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

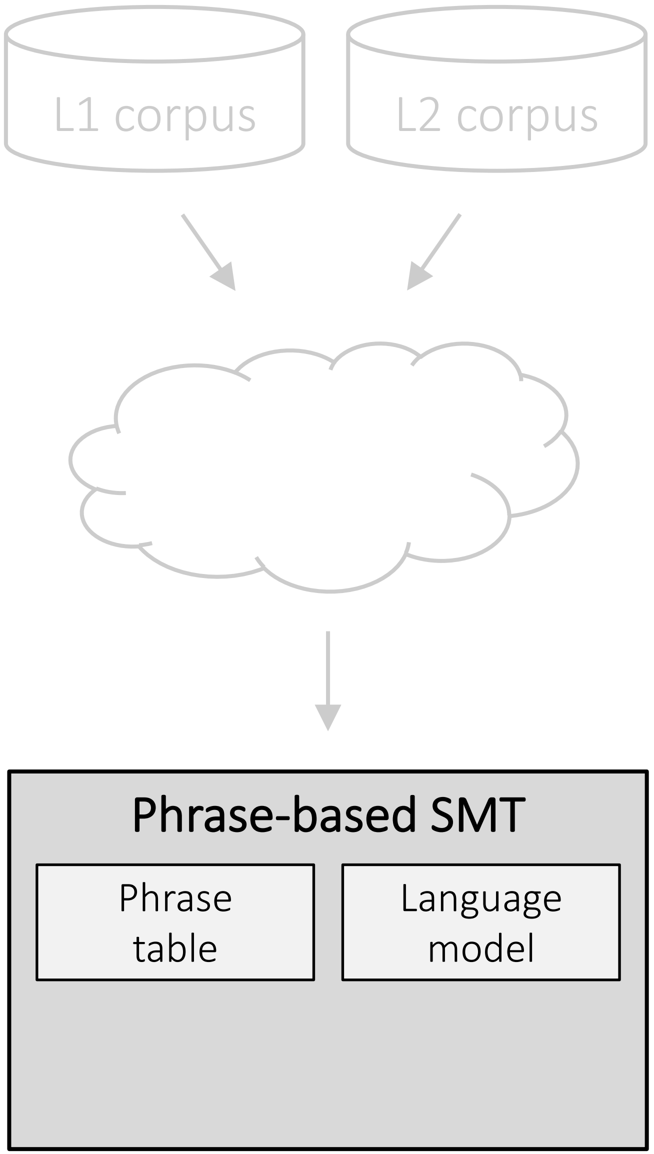
⋮



Phrase-based SMT

Log-linear model combining

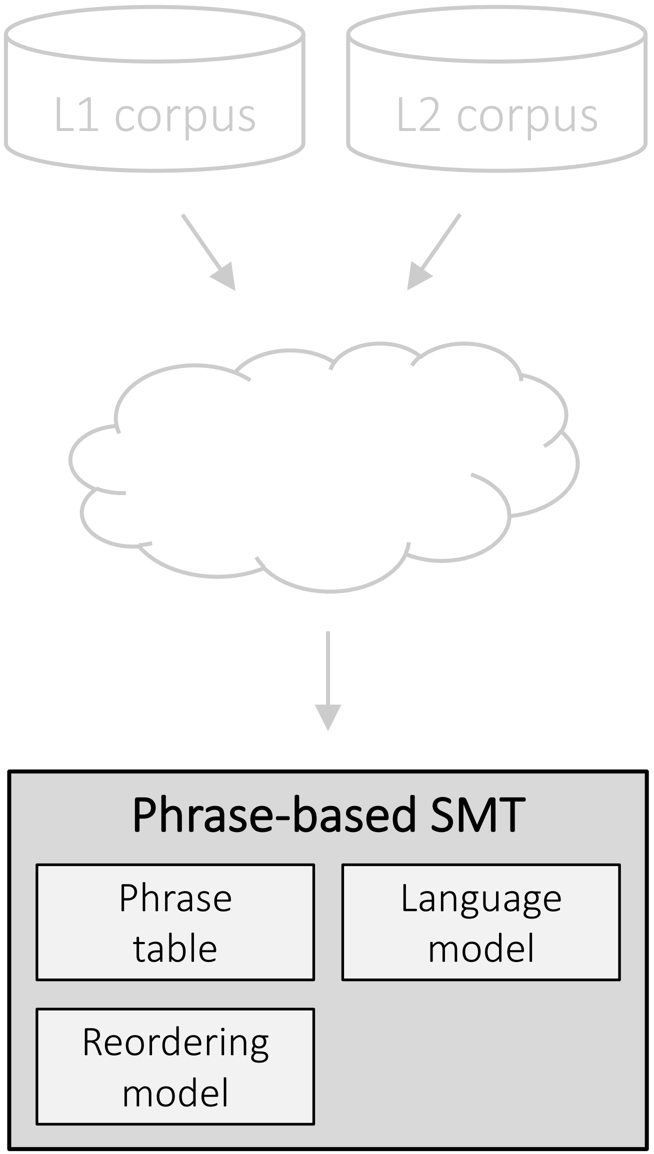
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model



Phrase-based SMT

Log-linear model combining

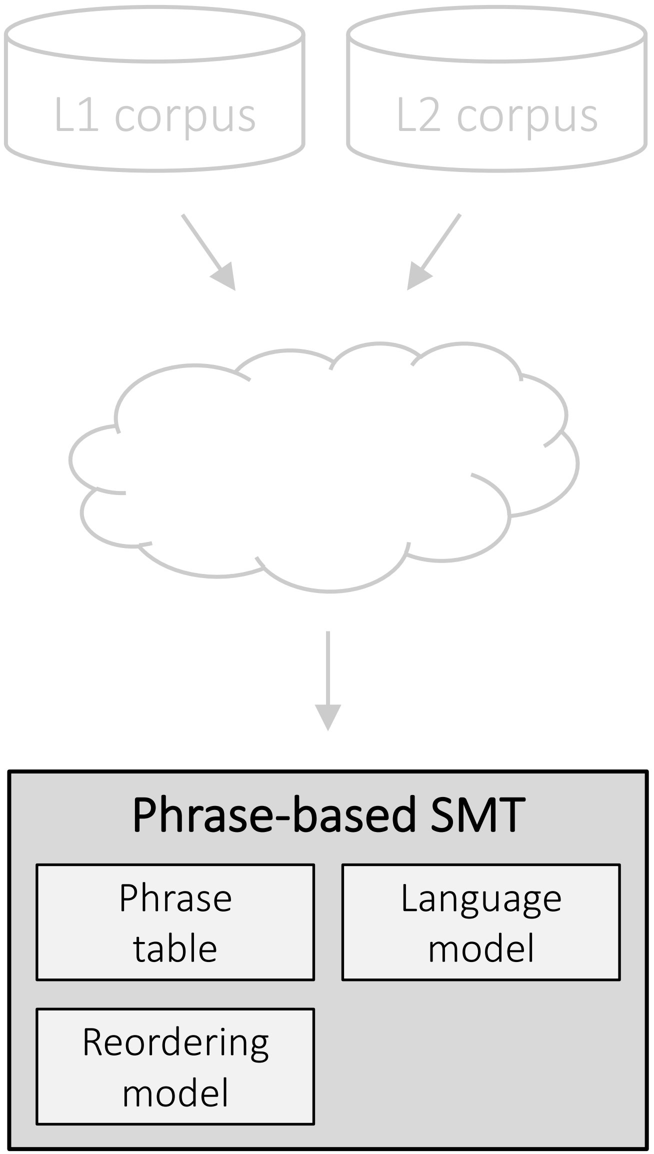
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing



Phrase-based SMT

Log-linear model combining

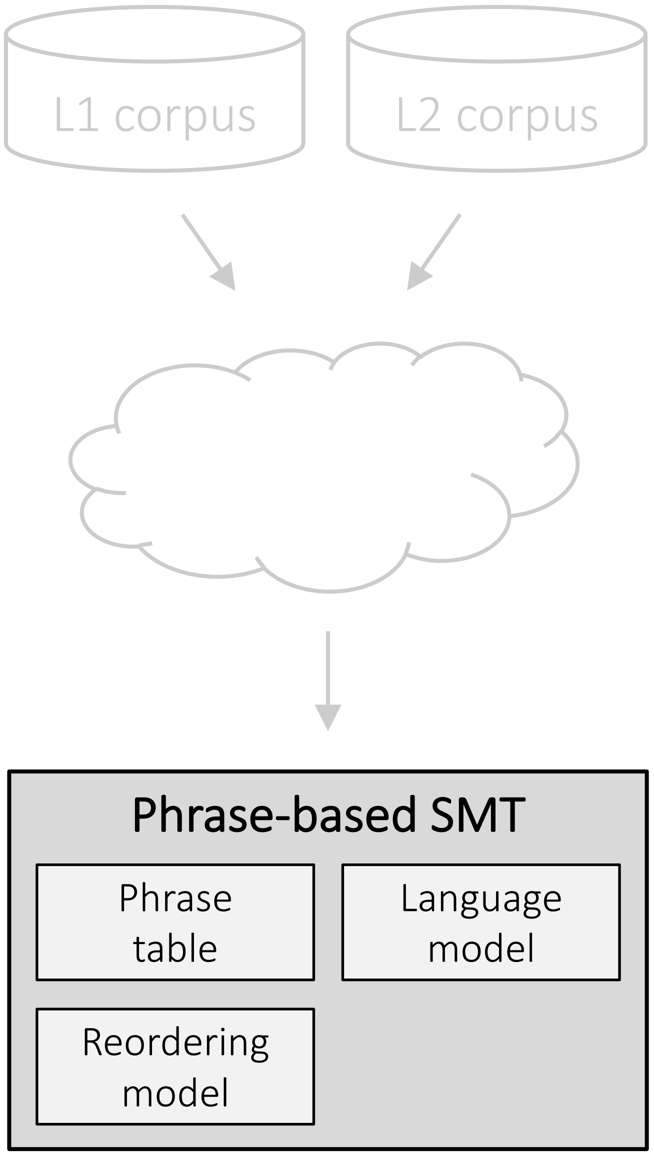
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model



Phrase-based SMT

Log-linear model combining

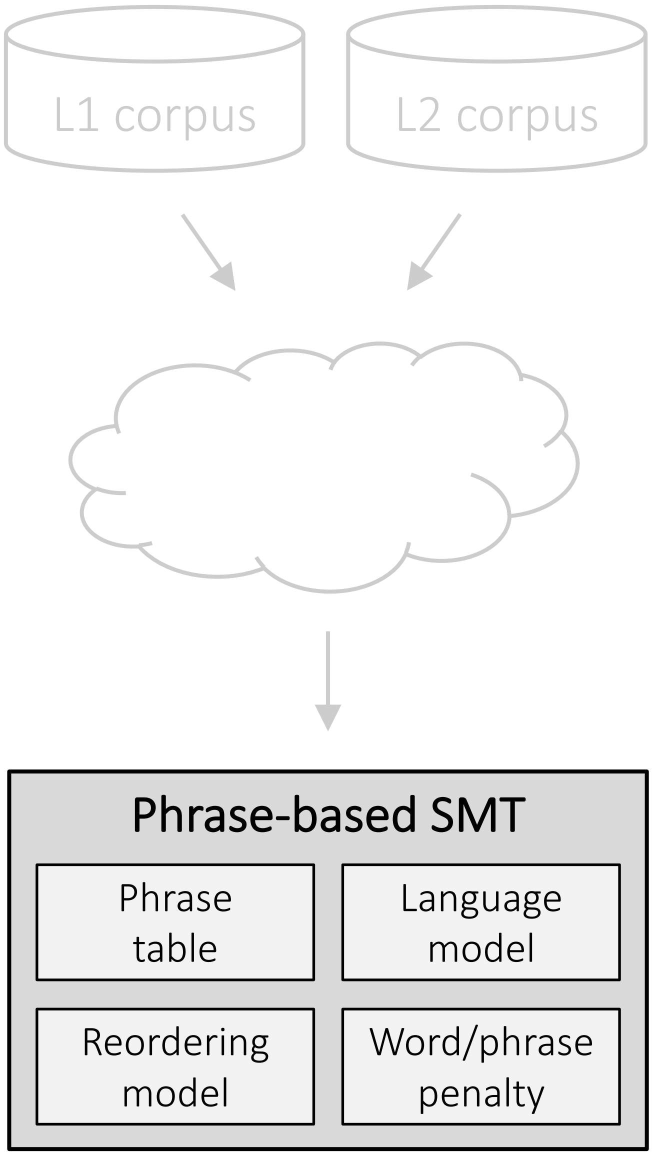
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)



Phrase-based SMT

Log-linear model combining

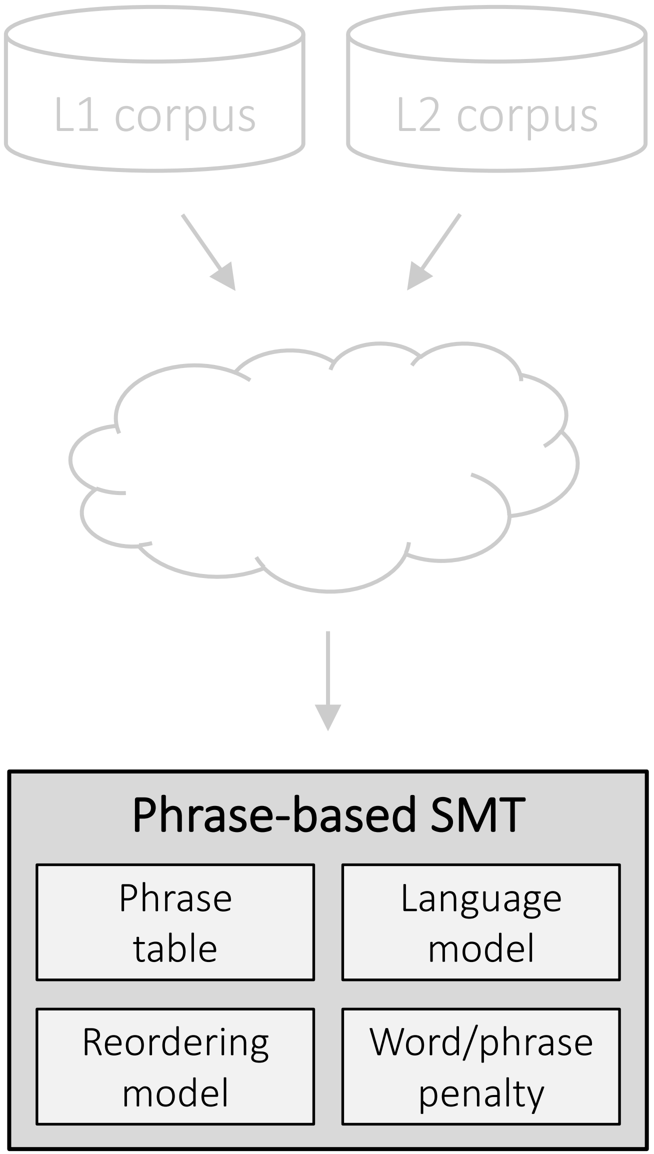
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model



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

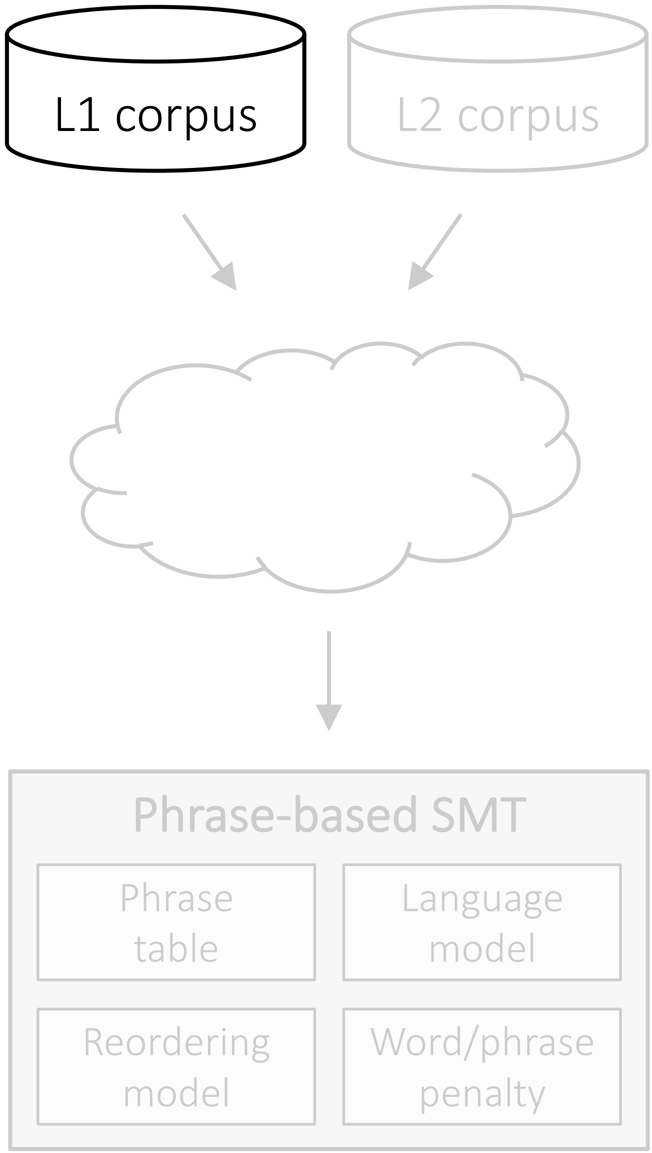
Phrase-based SMT



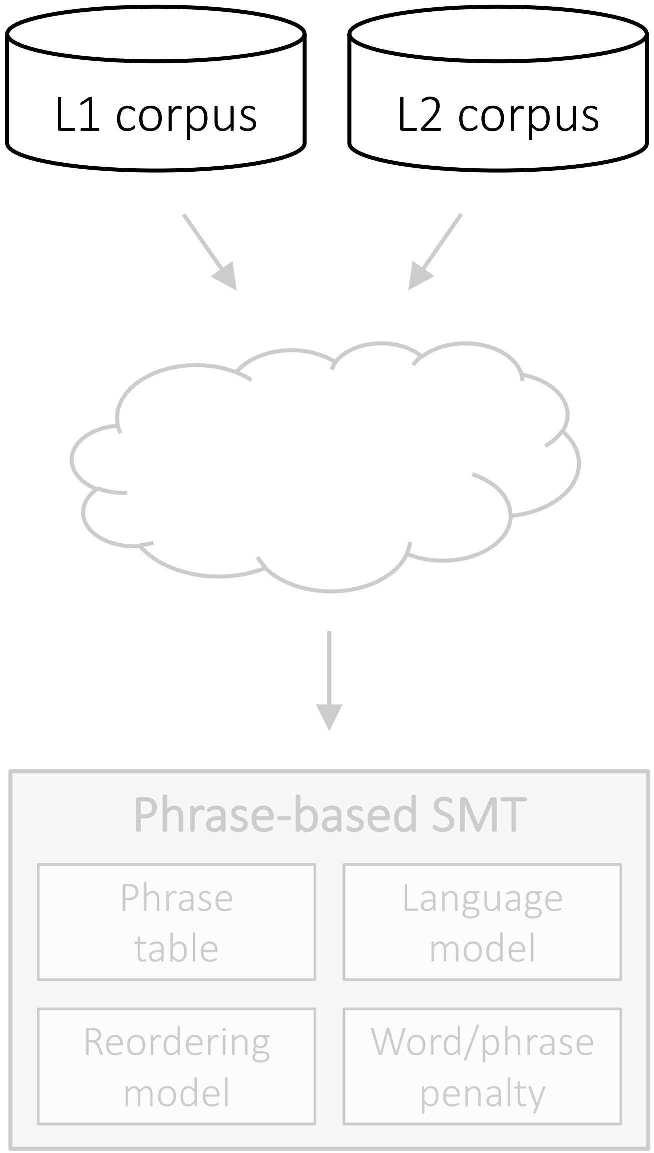
Log-linear model combining

- 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

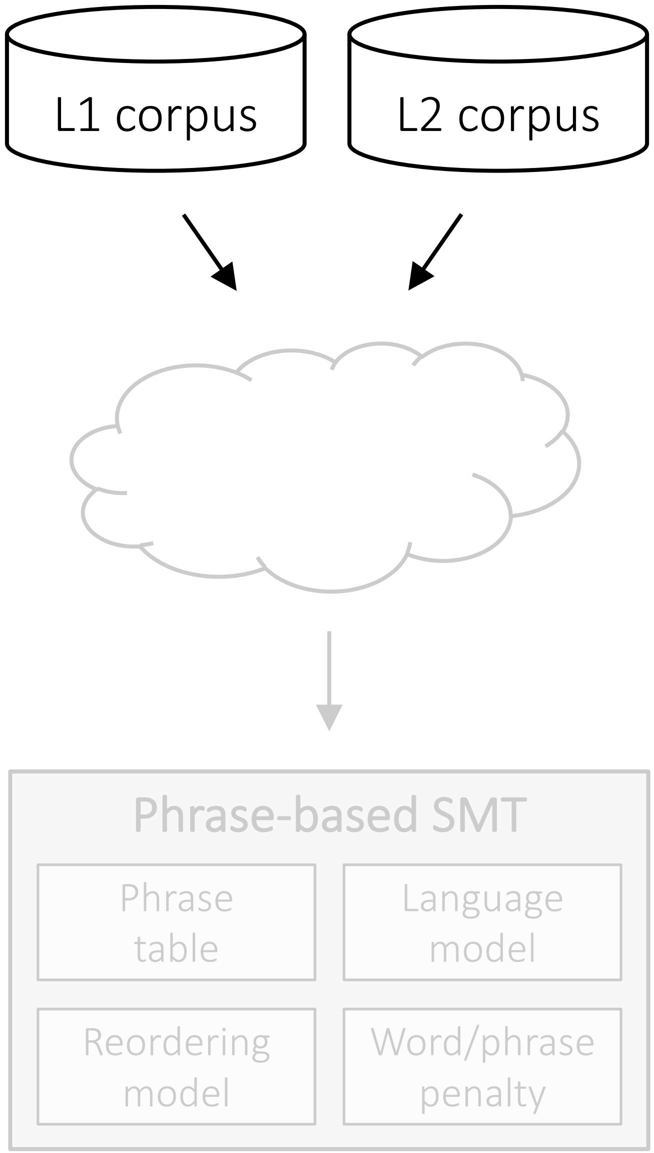
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



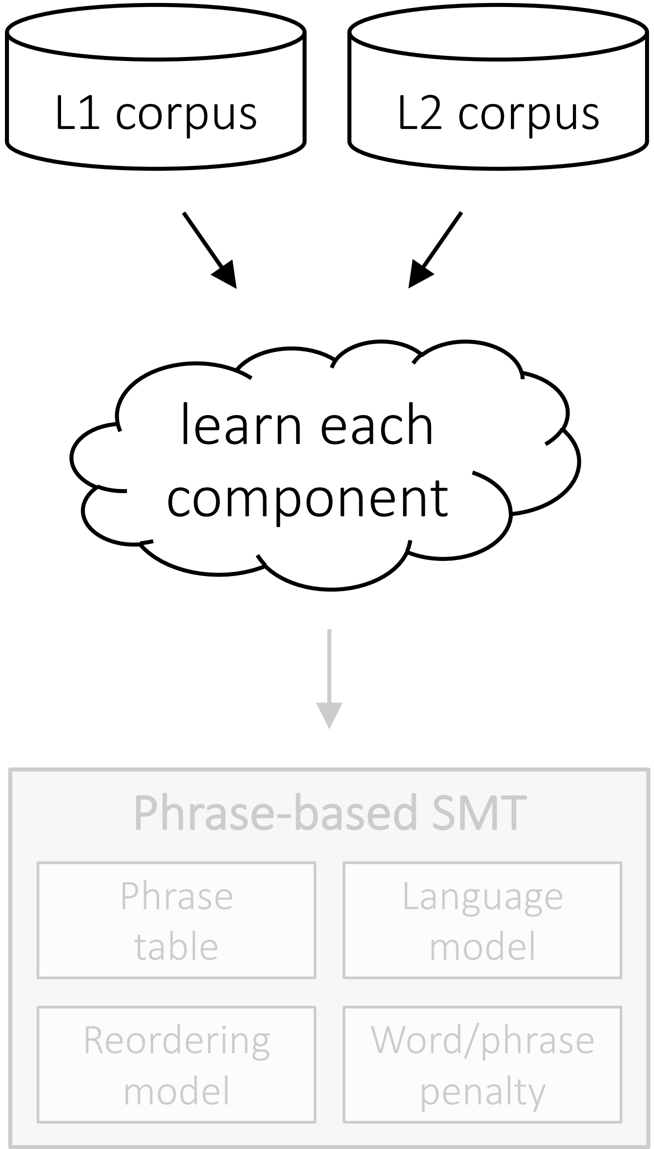
- Phrase table
    - Direct/inverse translation probabilities
    - Direct/inverse lexical weightings
- Language model
    - N-gram frequency counts with back-off and smoothing
- Reordering model
    - Distortion model (distance based)
    - Lexical reordering model
- Word/phrase penalty
    - Fixed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



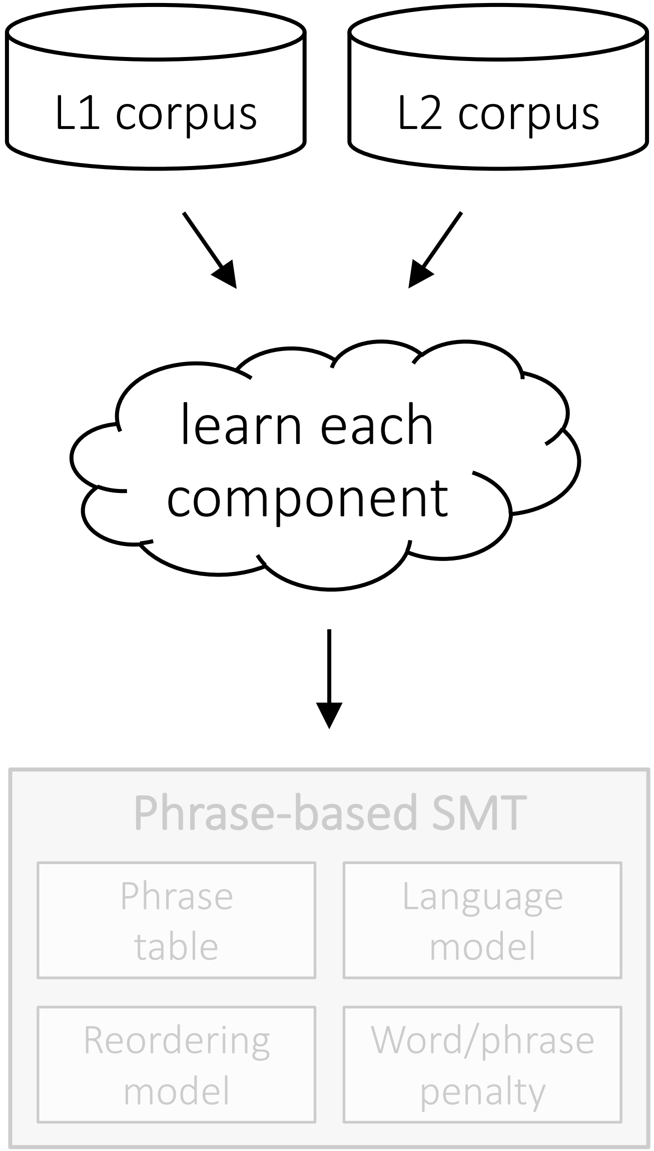
- Phrase table
    - Direct/inverse translation probabilities
    - Direct/inverse lexical weightings
- Language model
    - N-gram frequency counts with back-off and smoothing
- Reordering model
    - Distortion model (distance based)
    - Lexical reordering model
- Word/phrase penalty
    - Fixed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



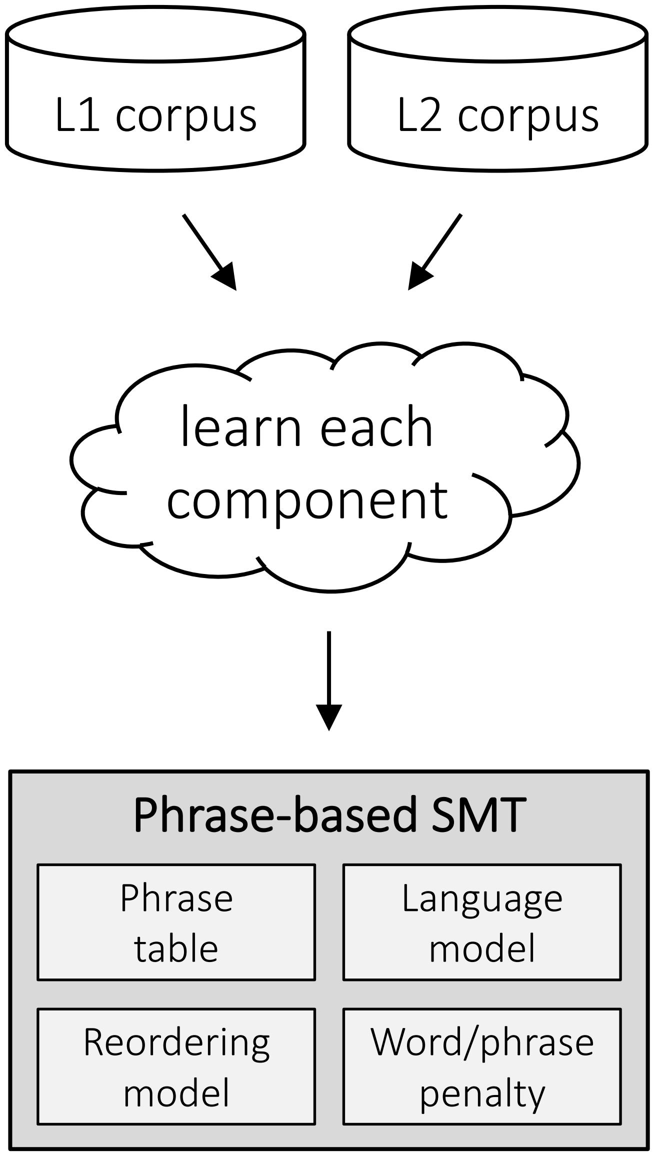
- Phrase table
    - Direct/inverse translation probabilities
    - Direct/inverse lexical weightings
- Language model
    - N-gram frequency counts with back-off and smoothing
- Reordering model
    - Distortion model (distance based)
    - Lexical reordering model
- Word/phrase penalty
    - Fixed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

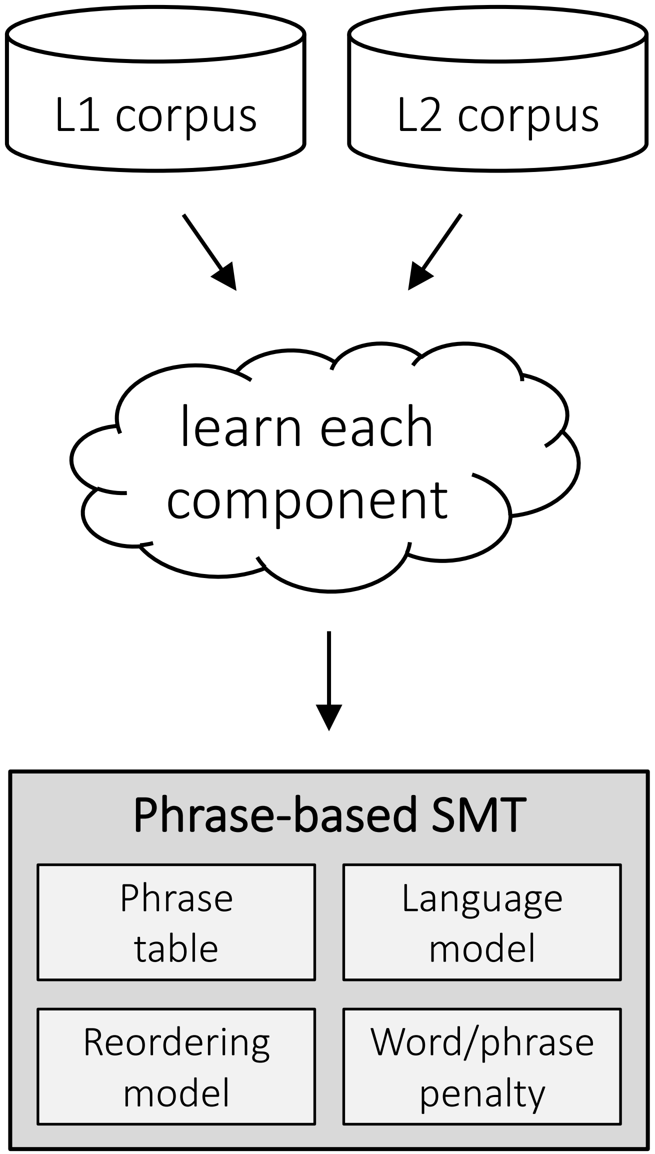


- Phrase table
    - Direct/inverse translation probabilities
    - Direct/inverse lexical weightings
- Language model
    - N-gram frequency counts with back-off and smoothing
- Reordering model
    - Distortion model (distance based)
    - Lexical reordering model
- Word/phrase penalty
    - Fixed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

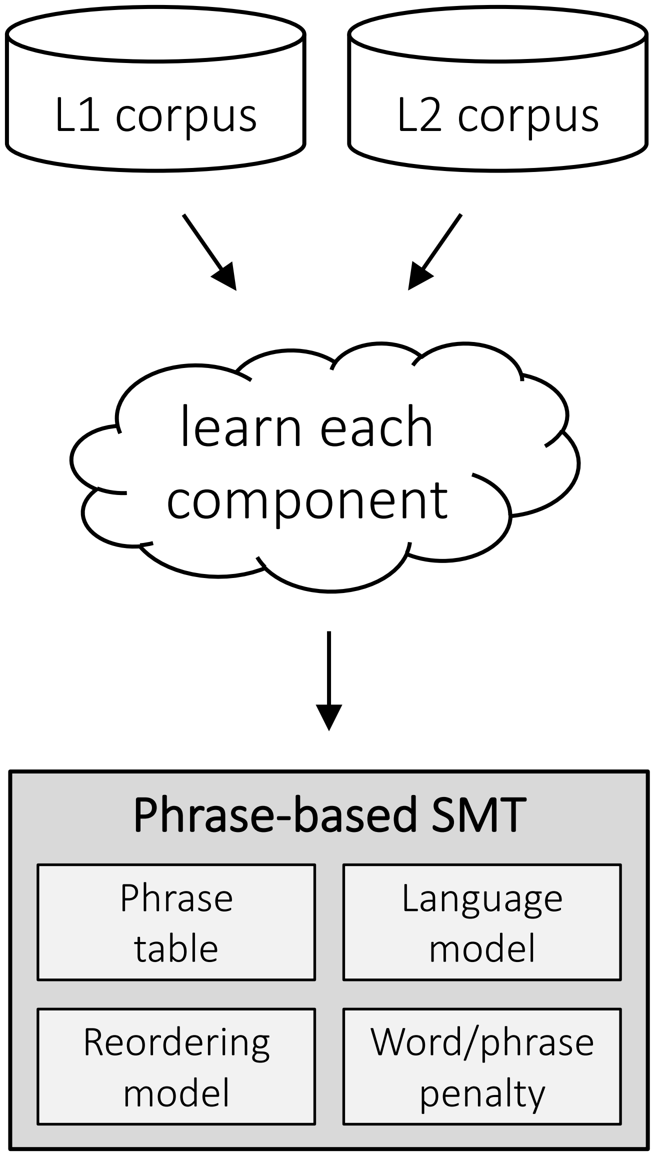


- Phrase table
    - Direct/inverse translation probabilities
    - Direct/inverse lexical weightings
- Language model
    - N-gram frequency counts with back-off and smoothing
- Reordering model
    - Distortion model (distance based)
    - Lexical reordering model
- Word/phrase penalty
    - Fixed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

Unsupervised phrase-based SMT



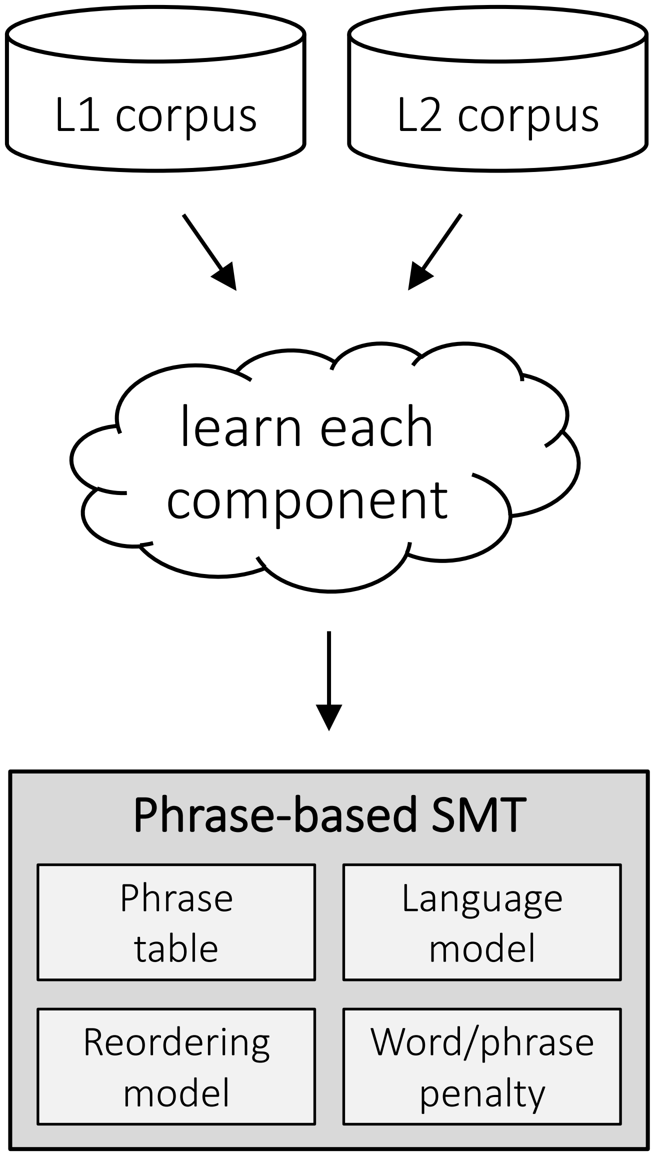
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

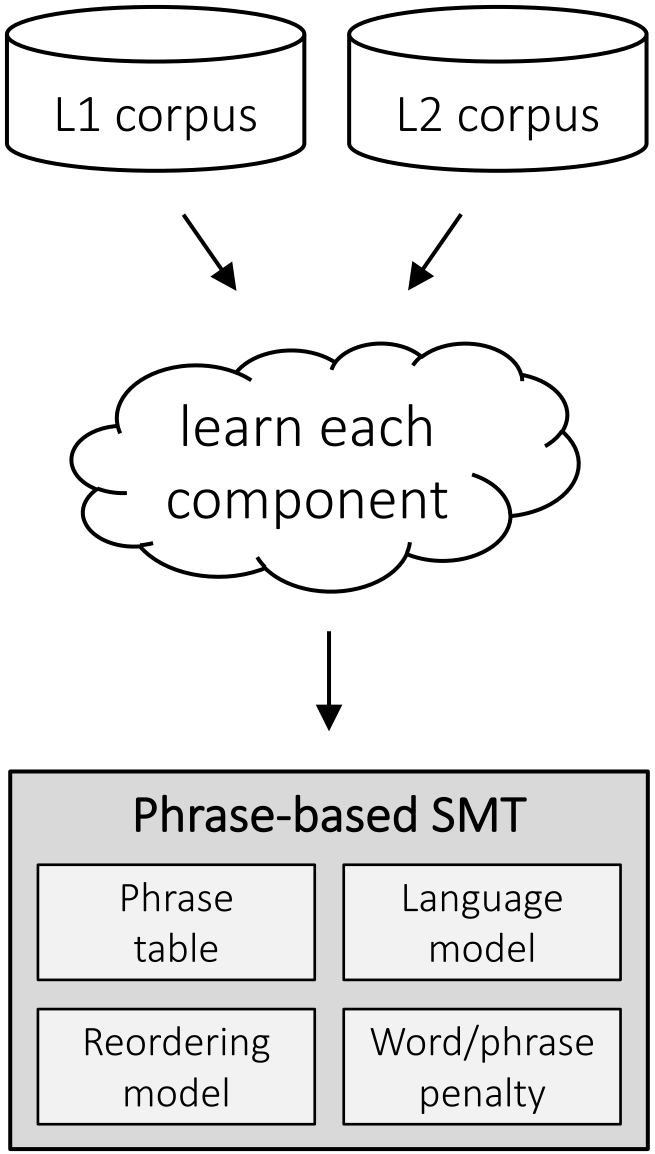
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty
  - Fixed score to control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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

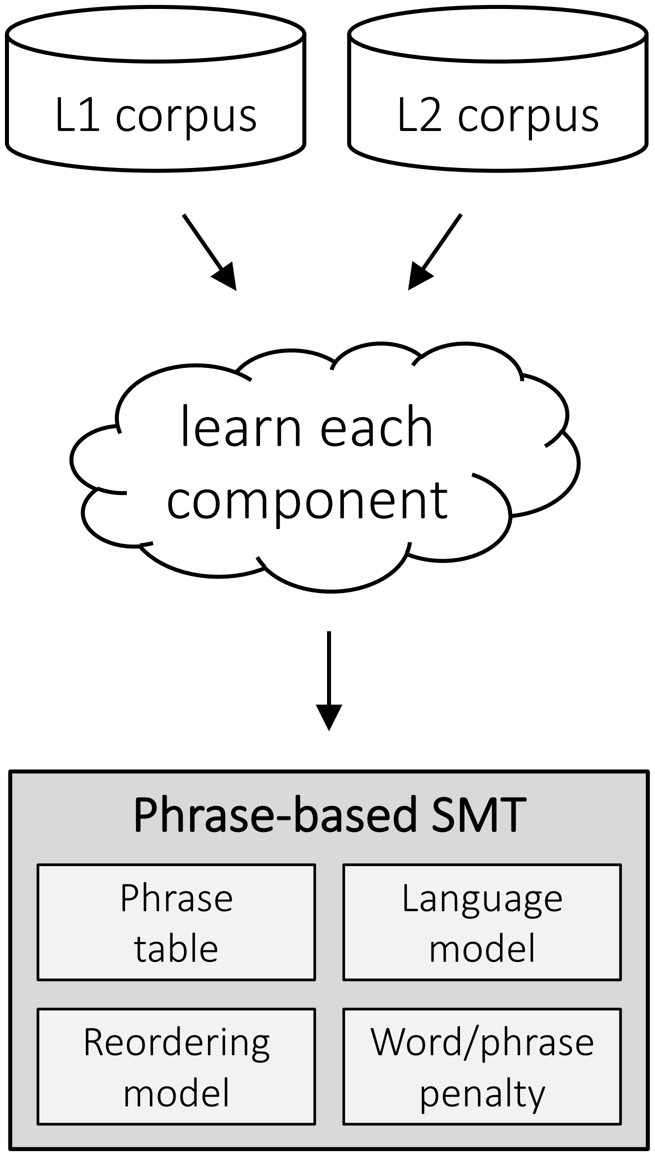
- Phrase table
    - Direct/inverse translation probabilities
    - Direct/inverse lexical weightings
- Language model
    - N-gram frequency counts with back-off and smoothing
- Reordering model
    - Distortion model (distance based)
    - Lexical reordering model
- Word/phrase penalty
    - Fixed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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

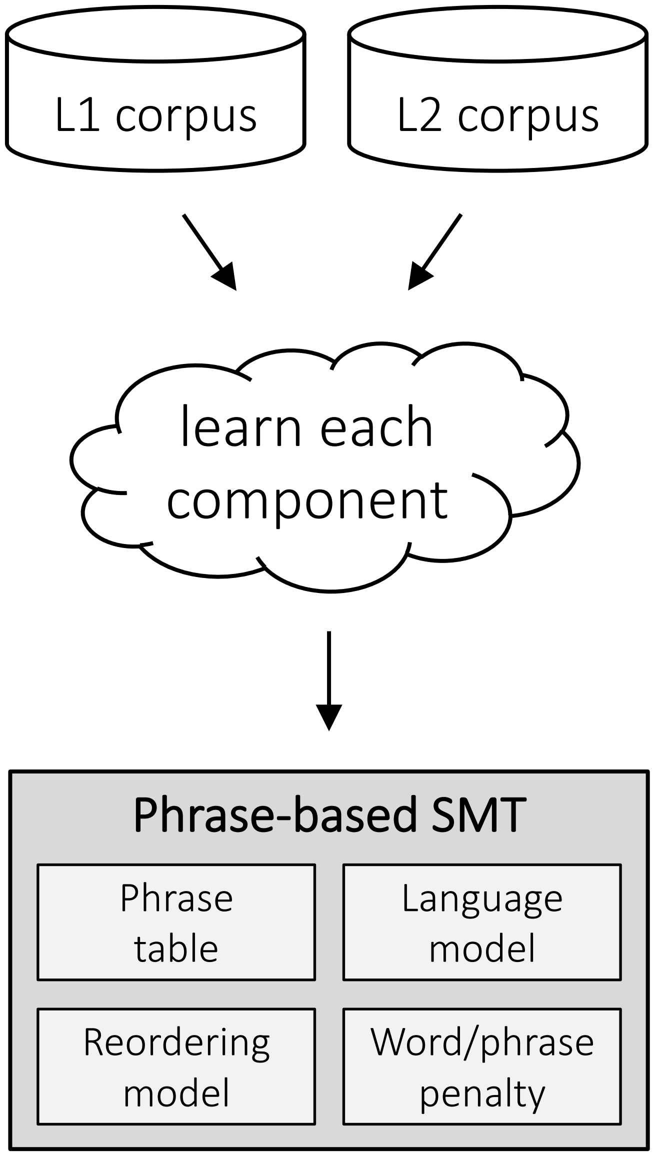
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model
  - N-gram frequency counts with back-off and smoothing
- Reordering model
  - Distortion model (distance based)
  - Lexical reordering model
- Word/phrase penalty EASY!!!
  - Fixed score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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

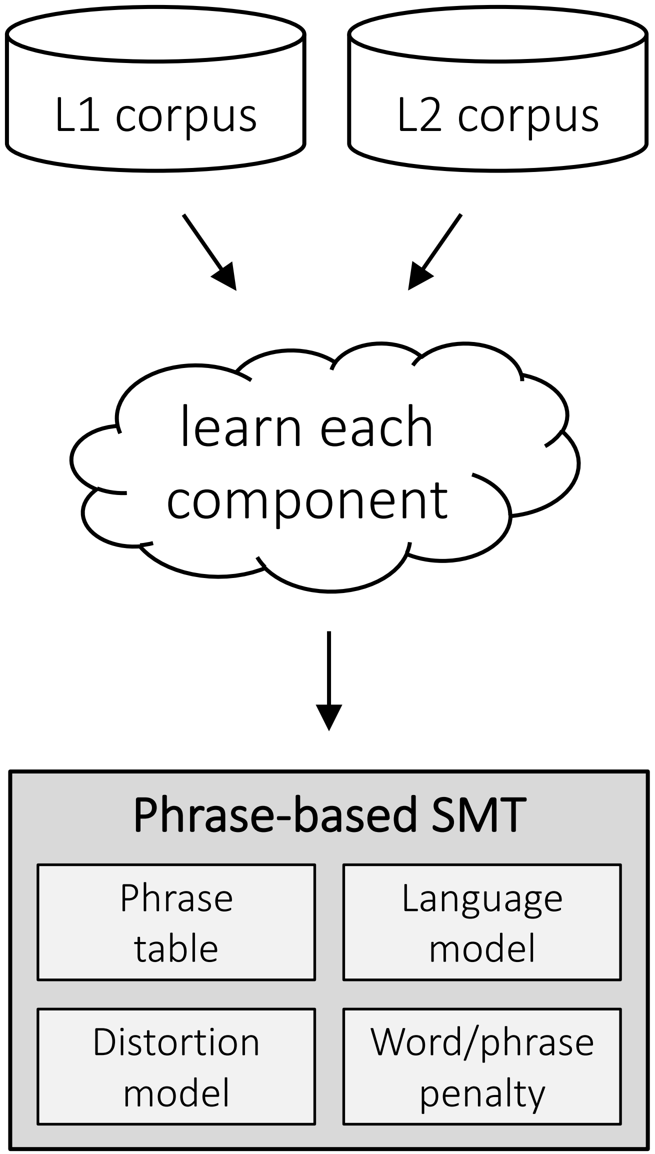
- Phrase table
    - Direct/inverse translation probabilities
    - Direct/inverse lexical weightings
- Language model
    - N-gram frequency counts with back-off and smoothing
- Reordering model
    - Distortion model (distance based)
    - Lexical reordering model
- Word/phrase penalty EASY!!!
   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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

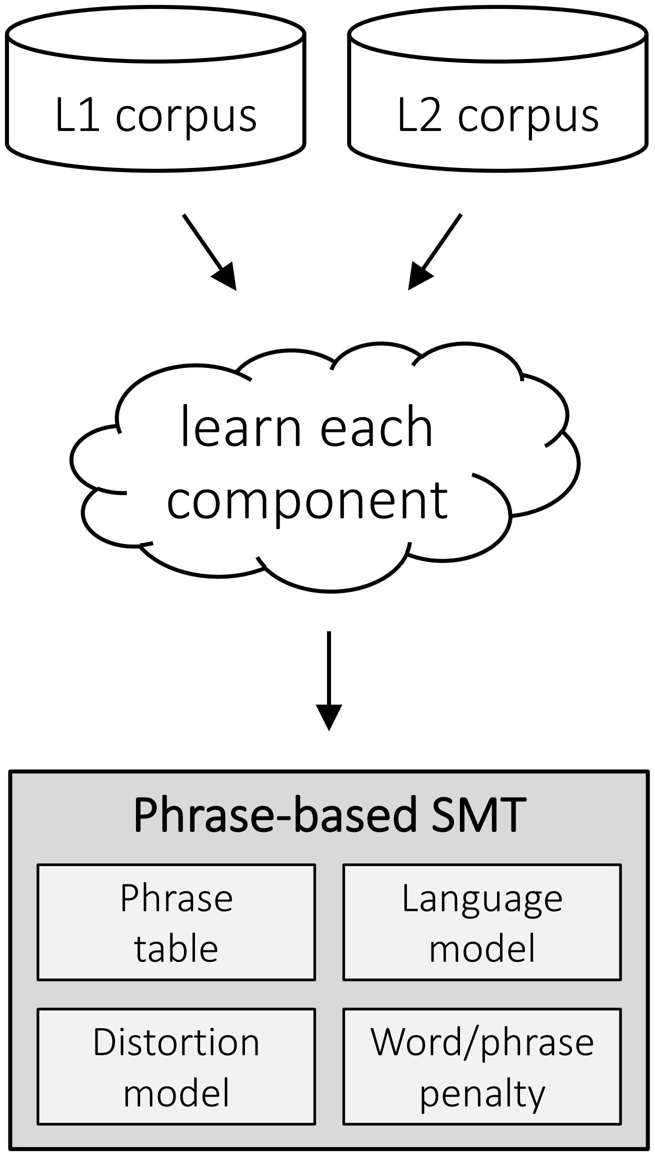
- Phrase table
    - Direct/inverse translation probabilities
    - Direct/inverse lexical weightings
- Language model
    - N-gram frequency counts with back-off and smoothing
- Reordering model
    - Distortion model (distance based)
    - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!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

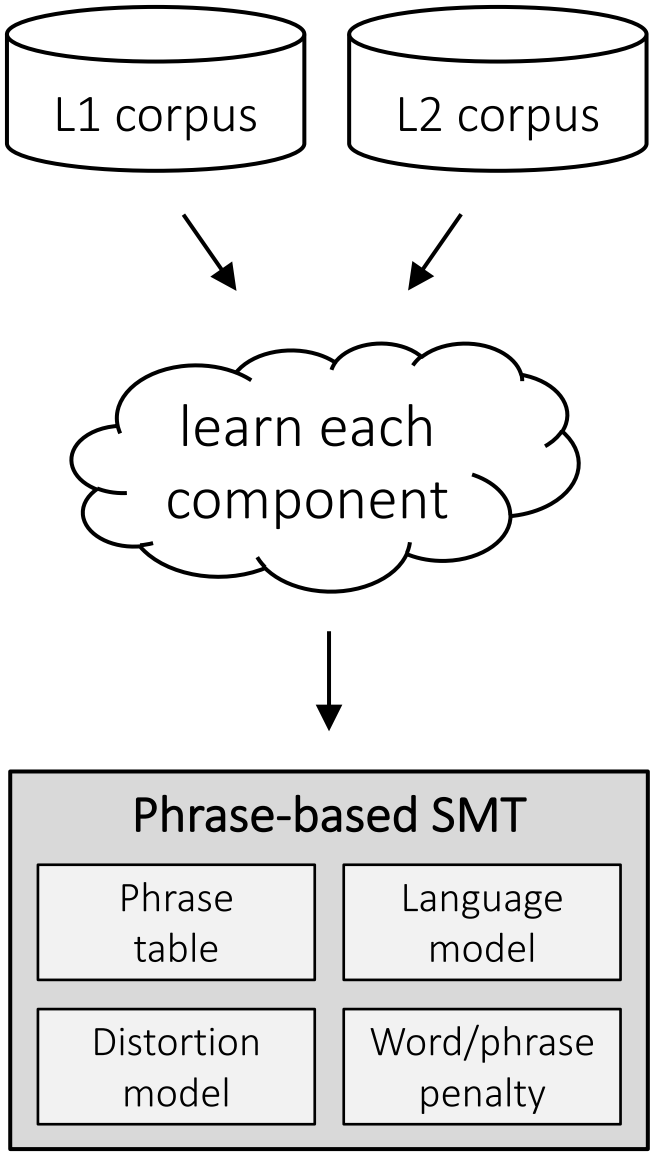


Unsupervised phrase-based SMT

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

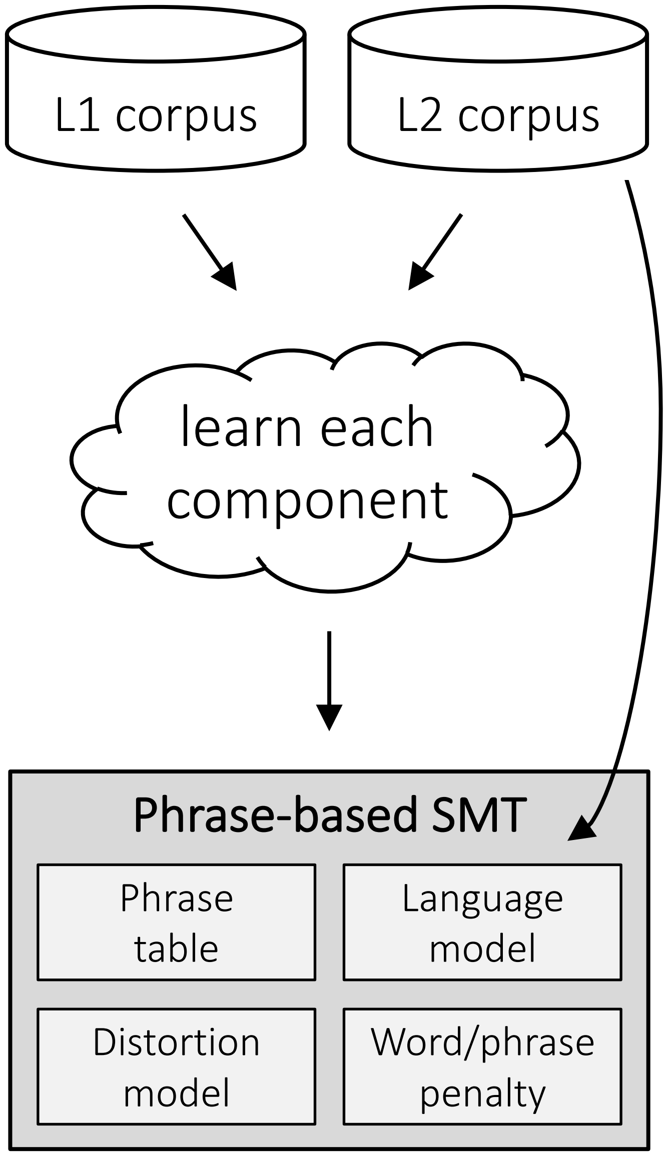
Unsupervised phrase-based SMT



Learn components from monolingual corpora

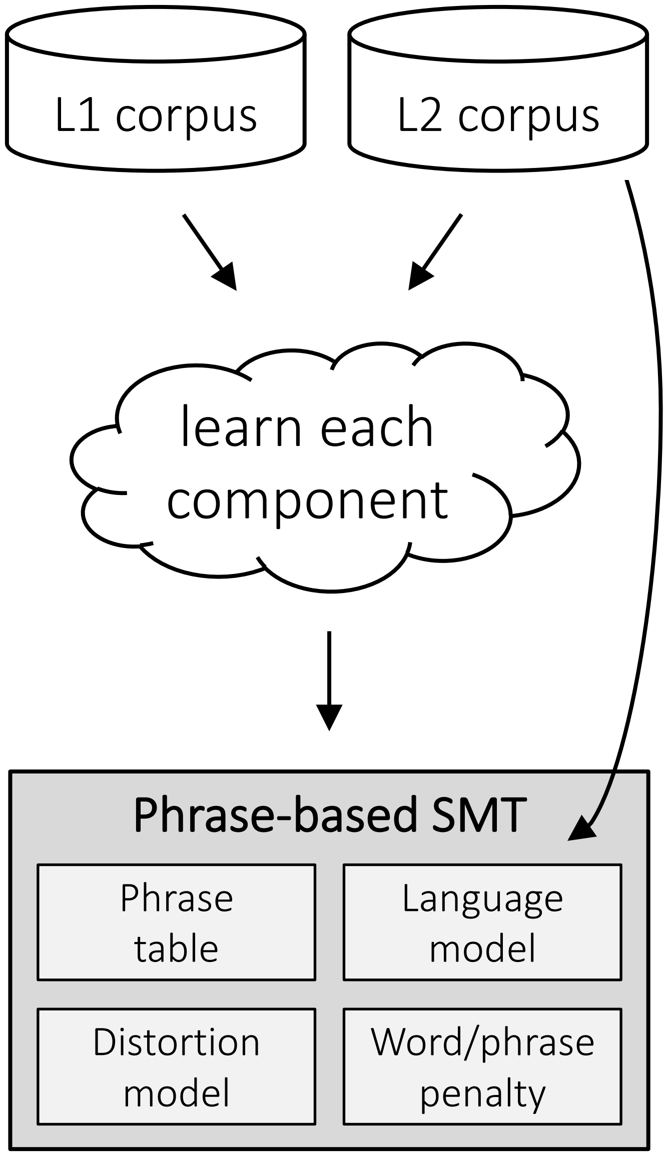
- Phrase table
    - Direct/inverse translation probabilities
    - Direct/inverse lexical weightings
- Language model
    - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
    - Distortion model (distance based)
    - ~~Lexical reordering model~~
- Word/phrase penalty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

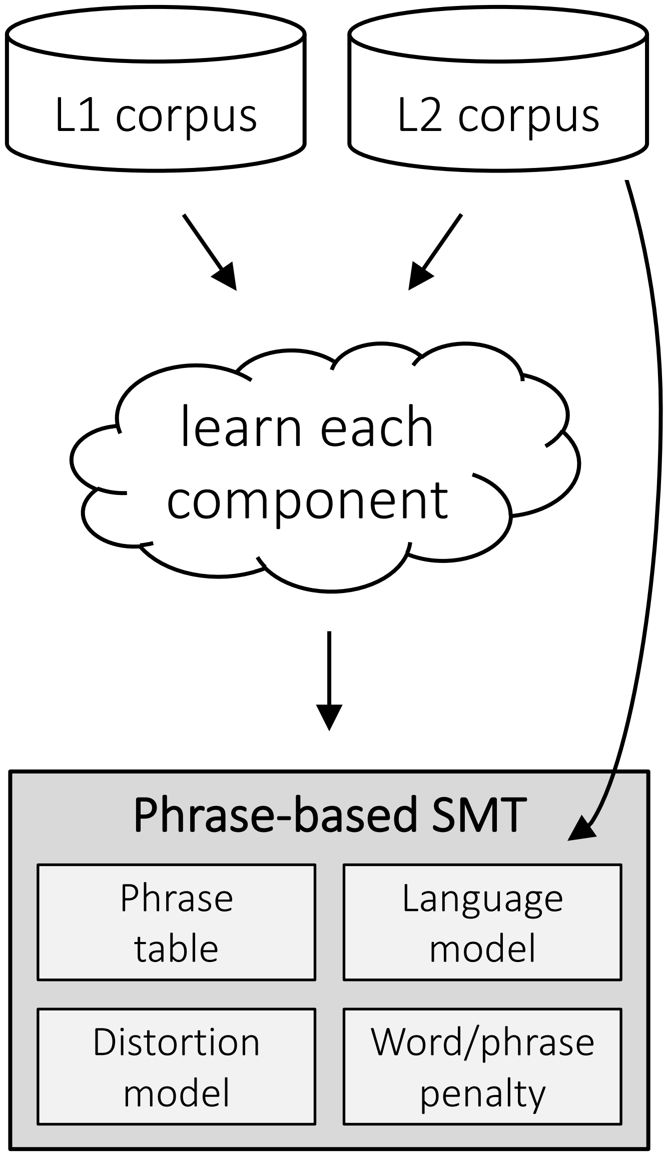


Unsupervised phrase-based SMT

Learn components from monolingual corpora

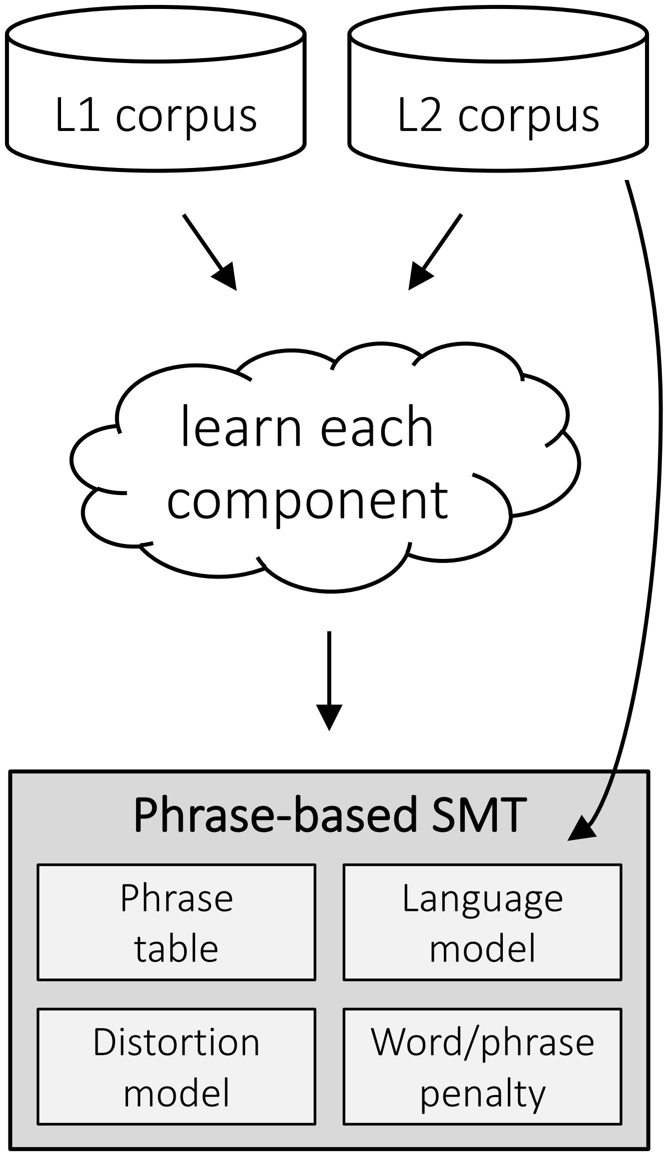
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!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

Unsupervised phrase-based SMT



Learn components from monolingual corpora

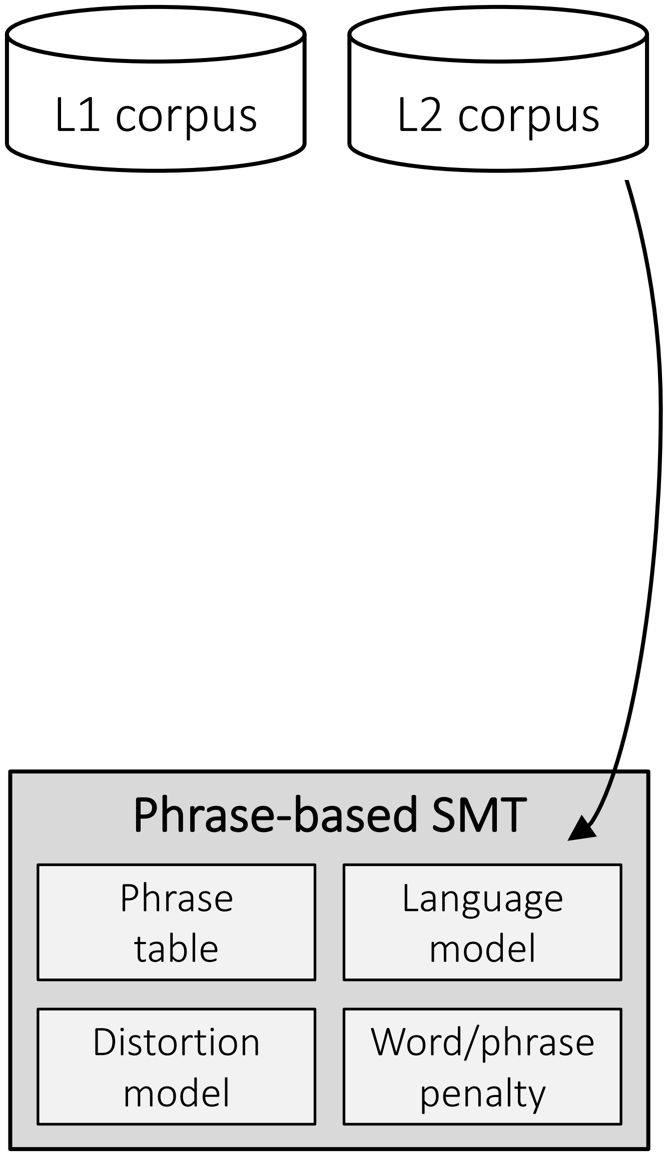
- Phrase table
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty EASY!!!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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

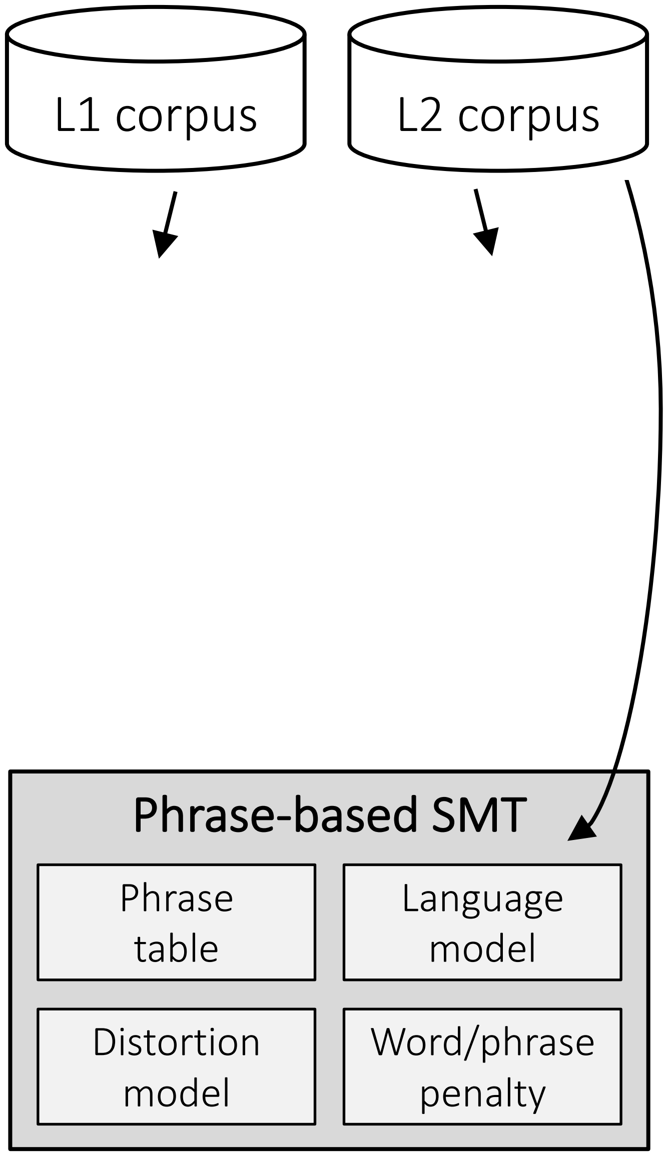
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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

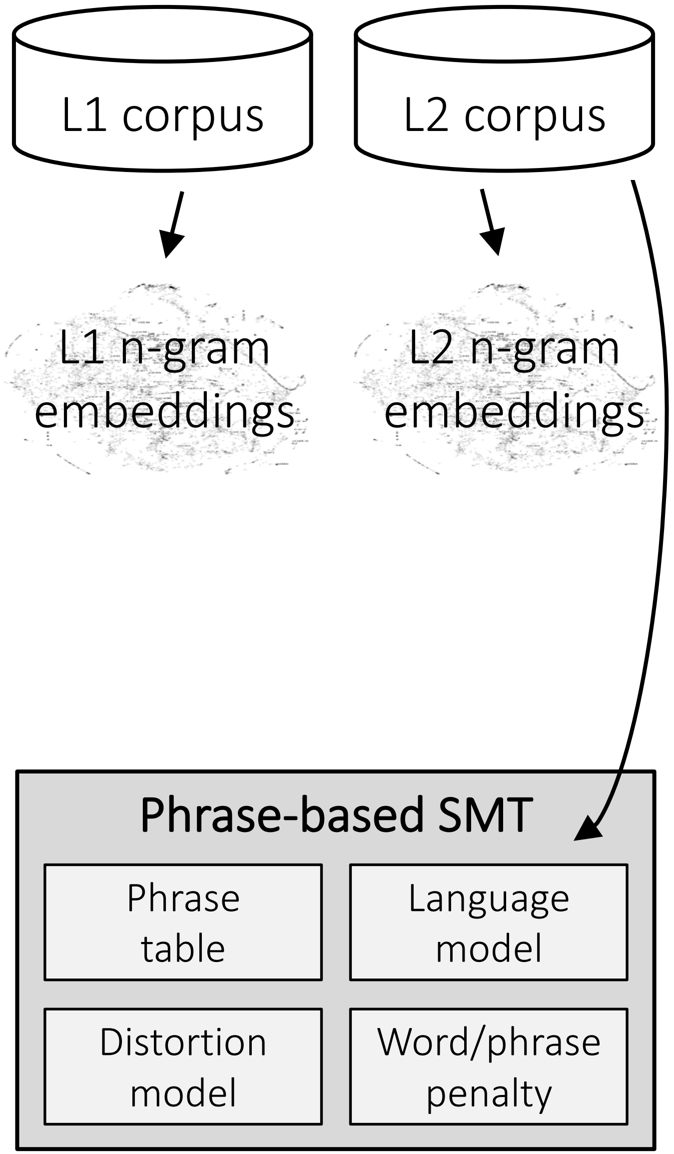


Unsupervised phrase-based SMT

Learn components from monolingual corpora

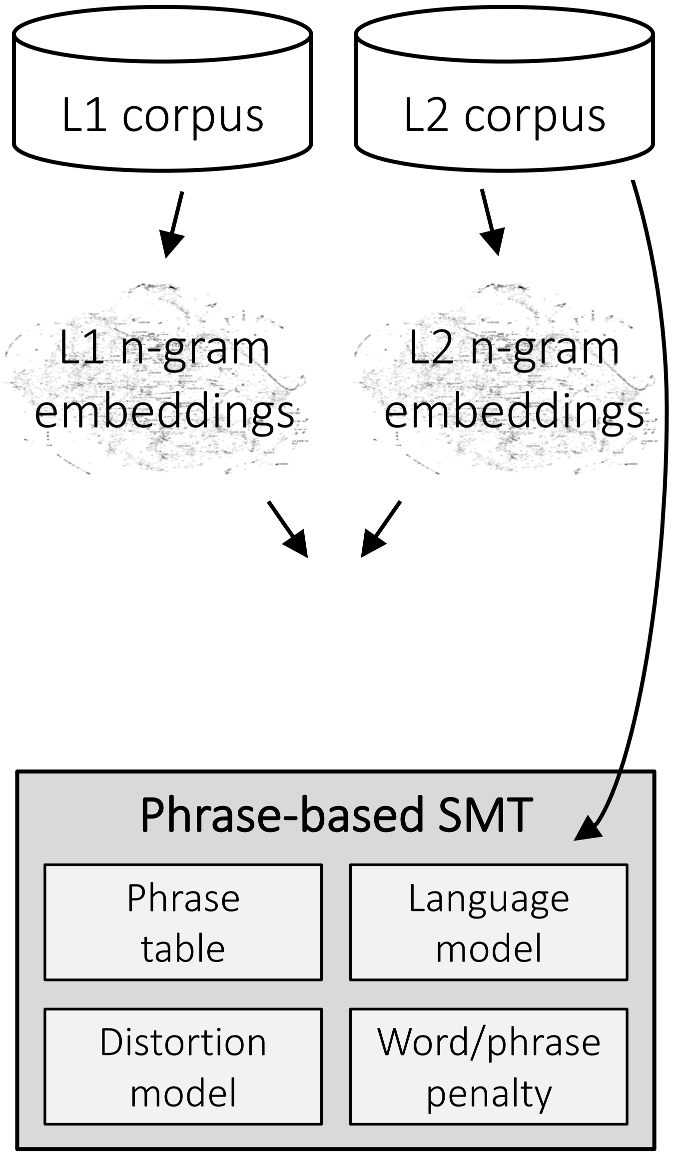
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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

Unsupervised phrase-based SMT



Learn components from monolingual corpora

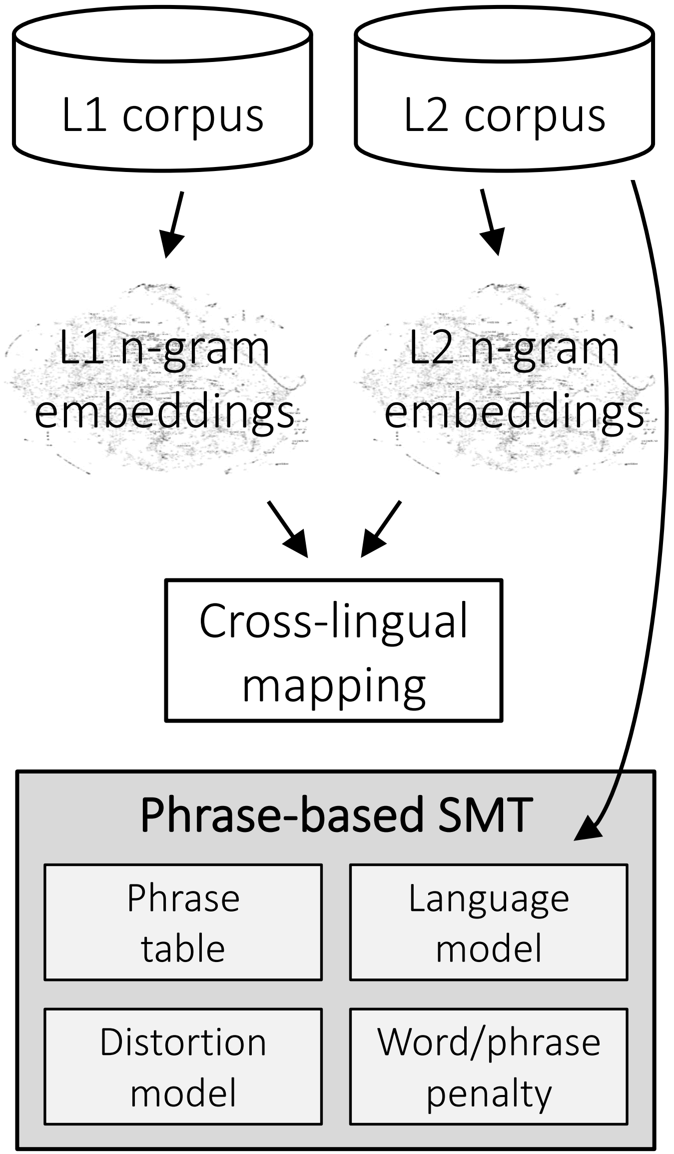
- Phrase table TRICKY...
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

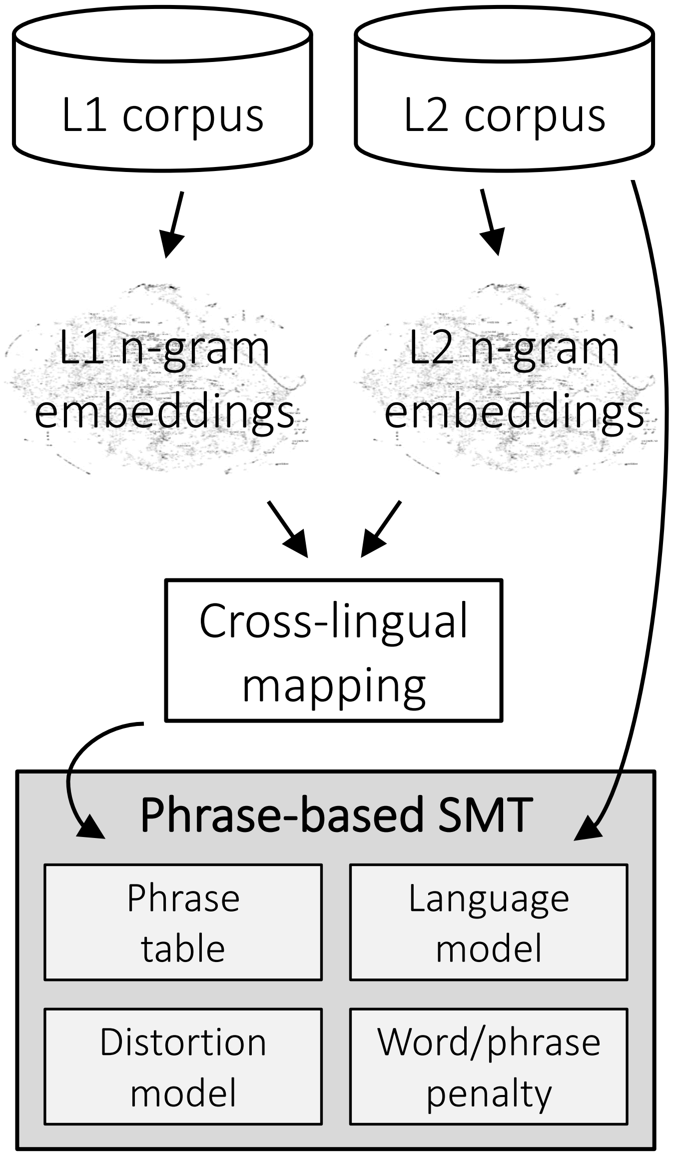
- Phrase table TRICKY...
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

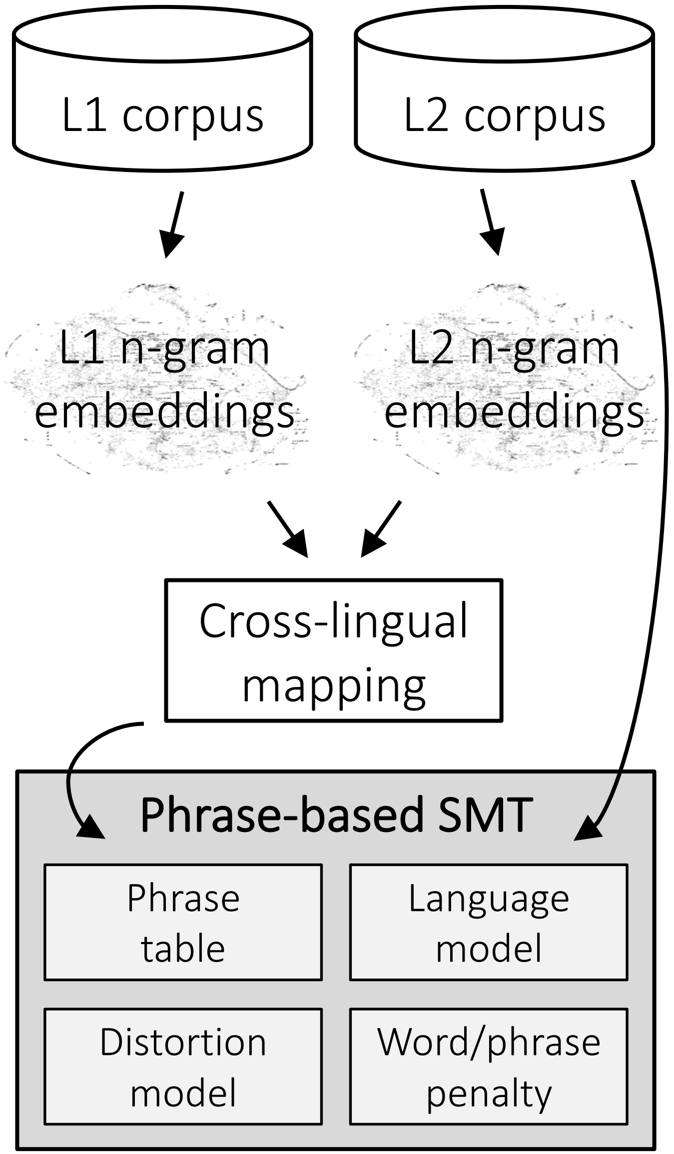
- Phrase table TRICKY...
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase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



Unsupervised phrase-based SMT

Learn components from monolingual corpora

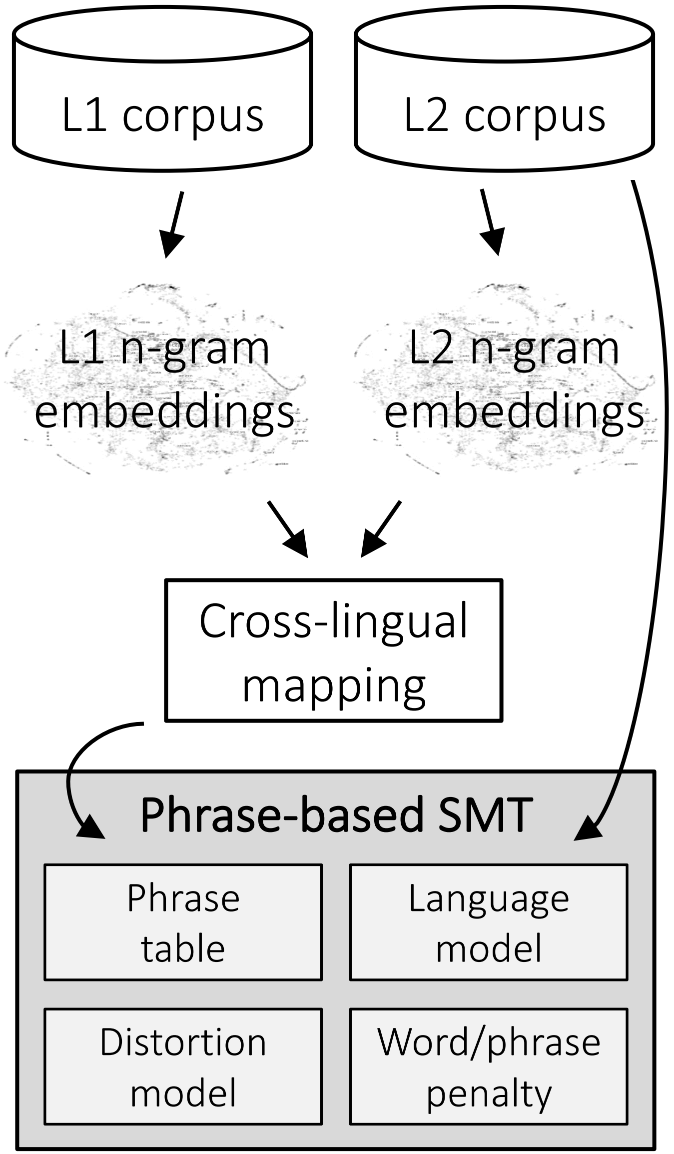
- Phrase table TRICKY…
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase penalty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

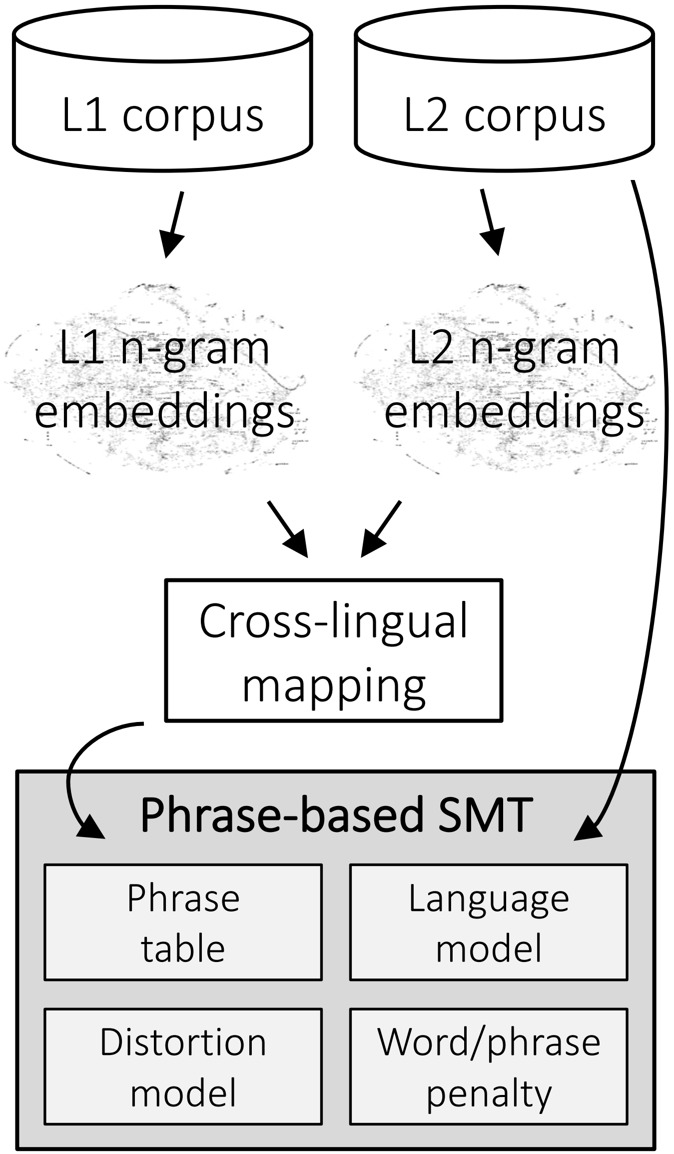


Unsupervised phrase-based SMT

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 .

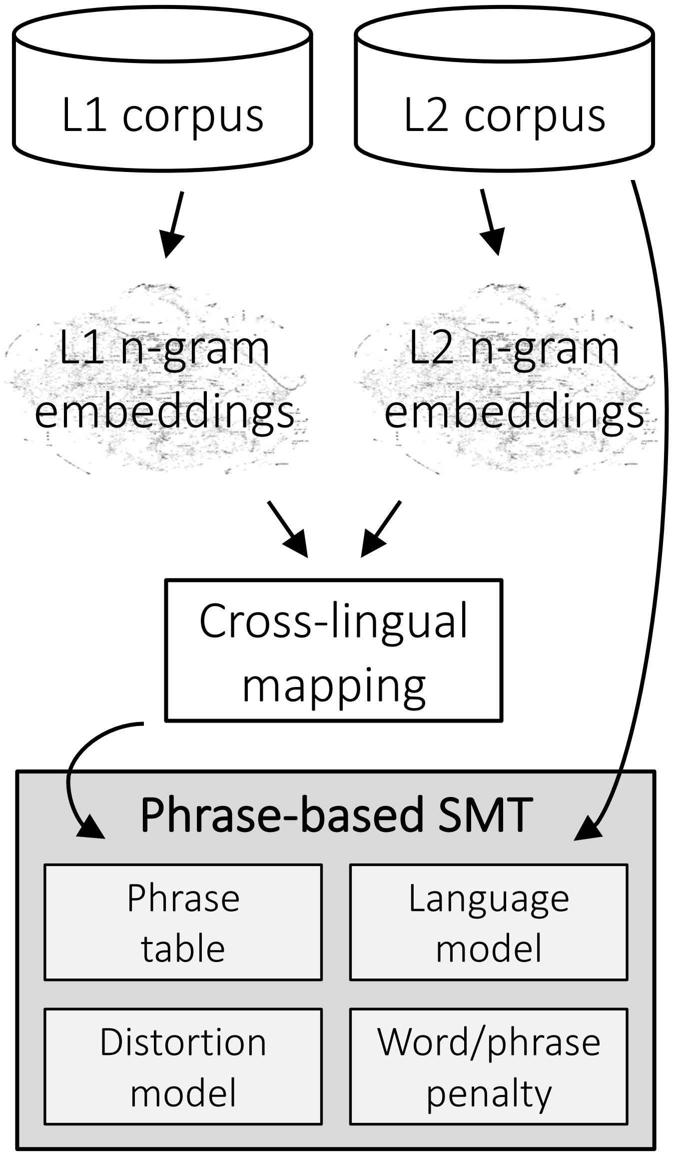


Unsupervised phrase-based SMT

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



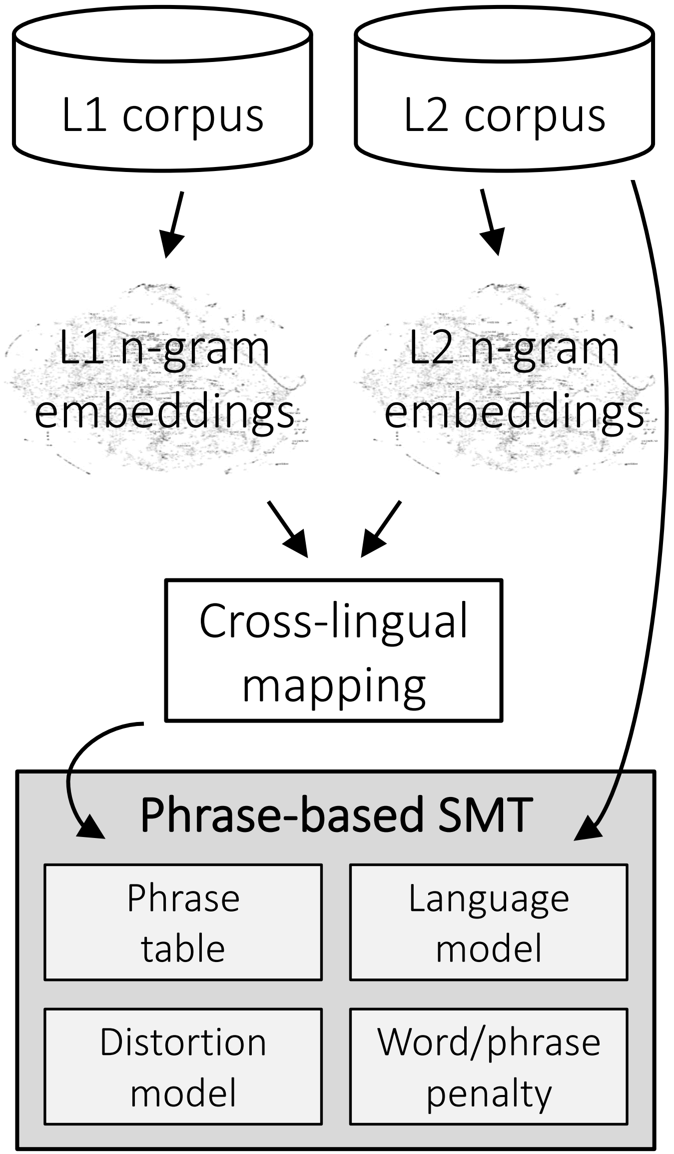
Unsupervised phrase-based SMT

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

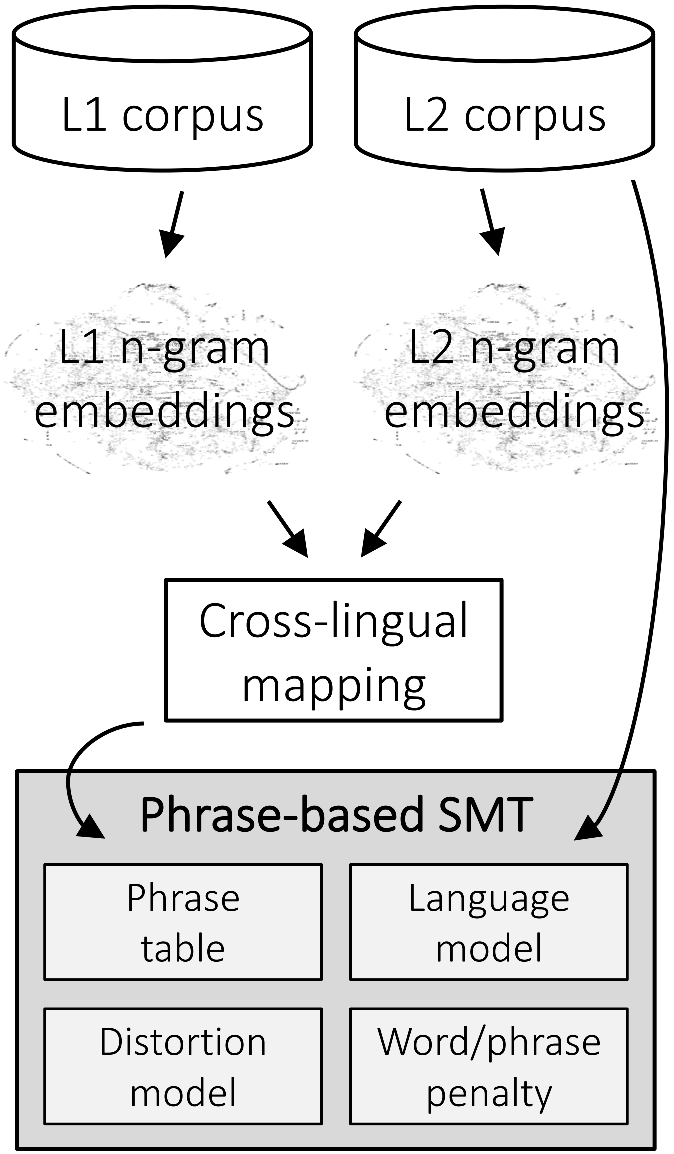


Unsupervised phrase-based SMT

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

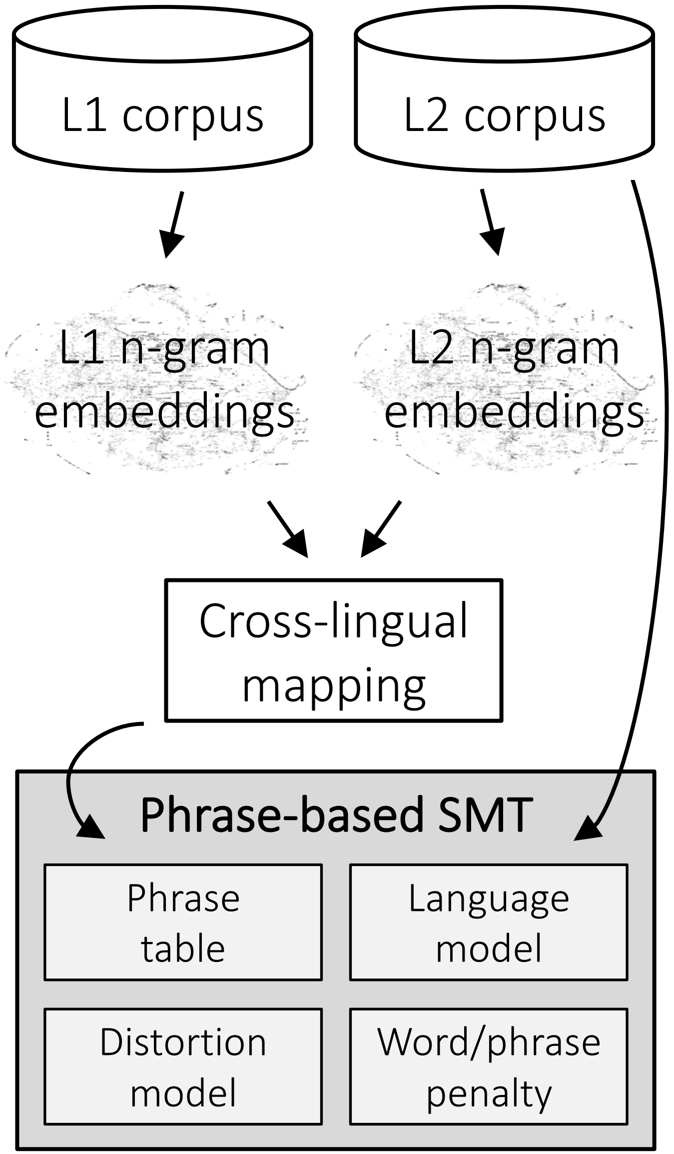


Unsupervised phrase-based SMT

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

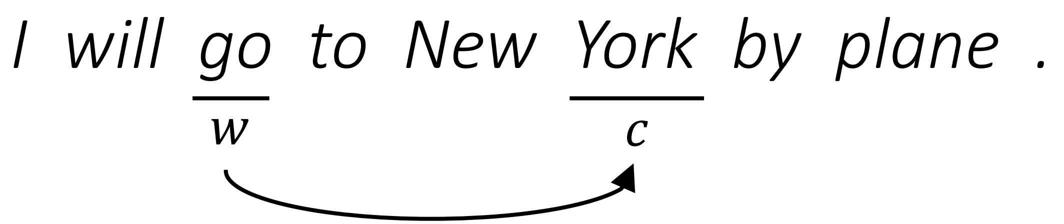
I will $\underbrace{\textit{go}}_{w}$ *to* $\underbrace{\textit{New}}_{c}$ *York by plane .*

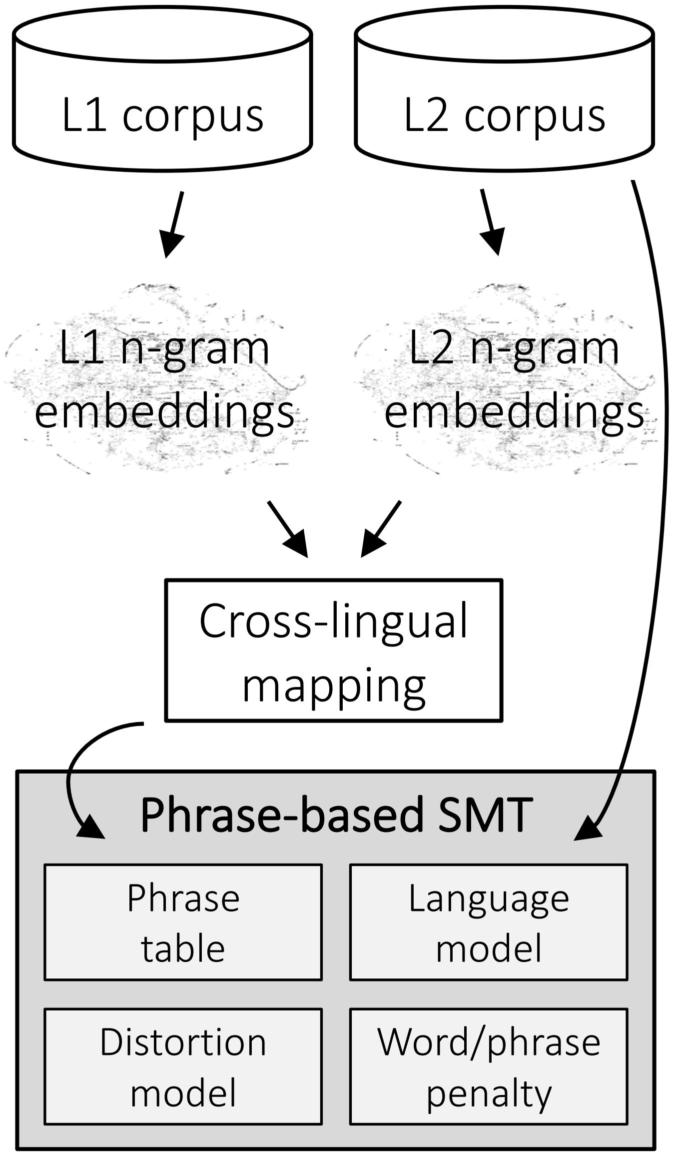


Unsupervised phrase-based SMT

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

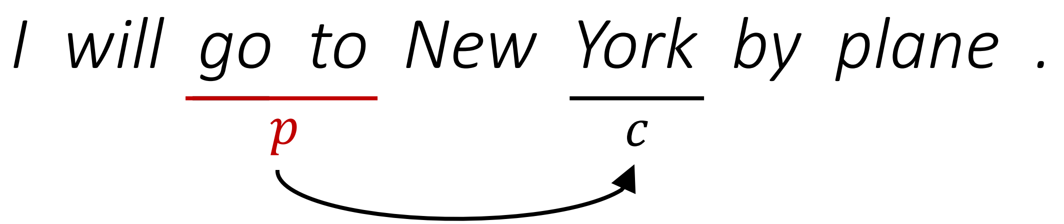


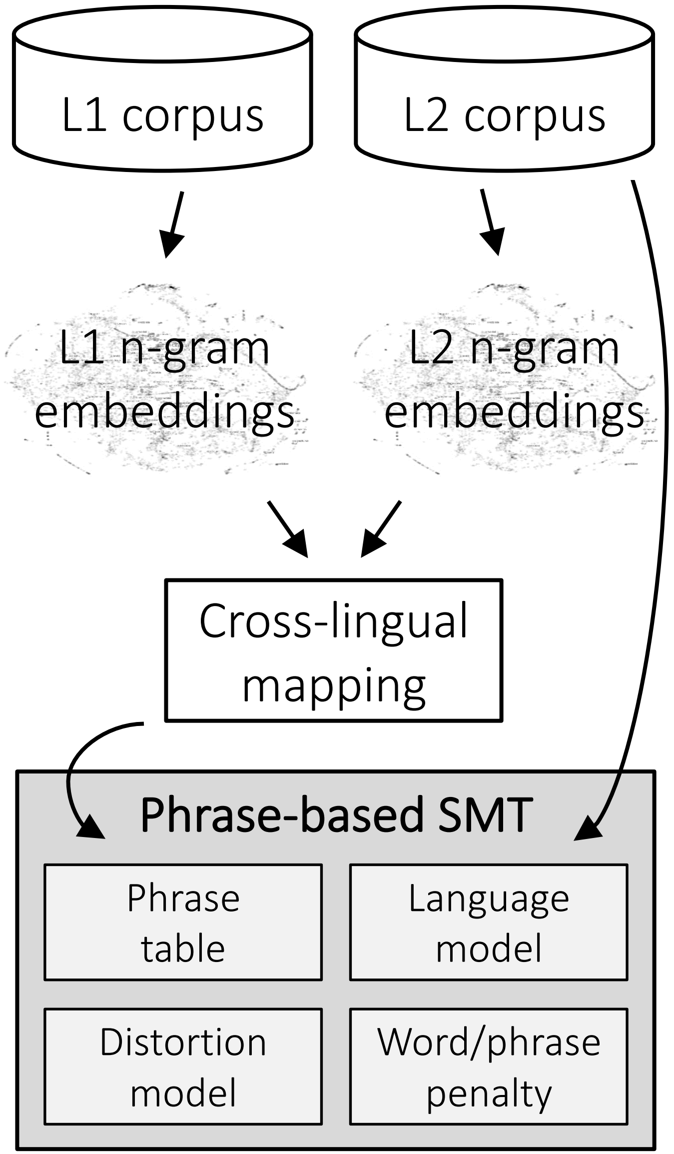


Unsupervised phrase-based SMT

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





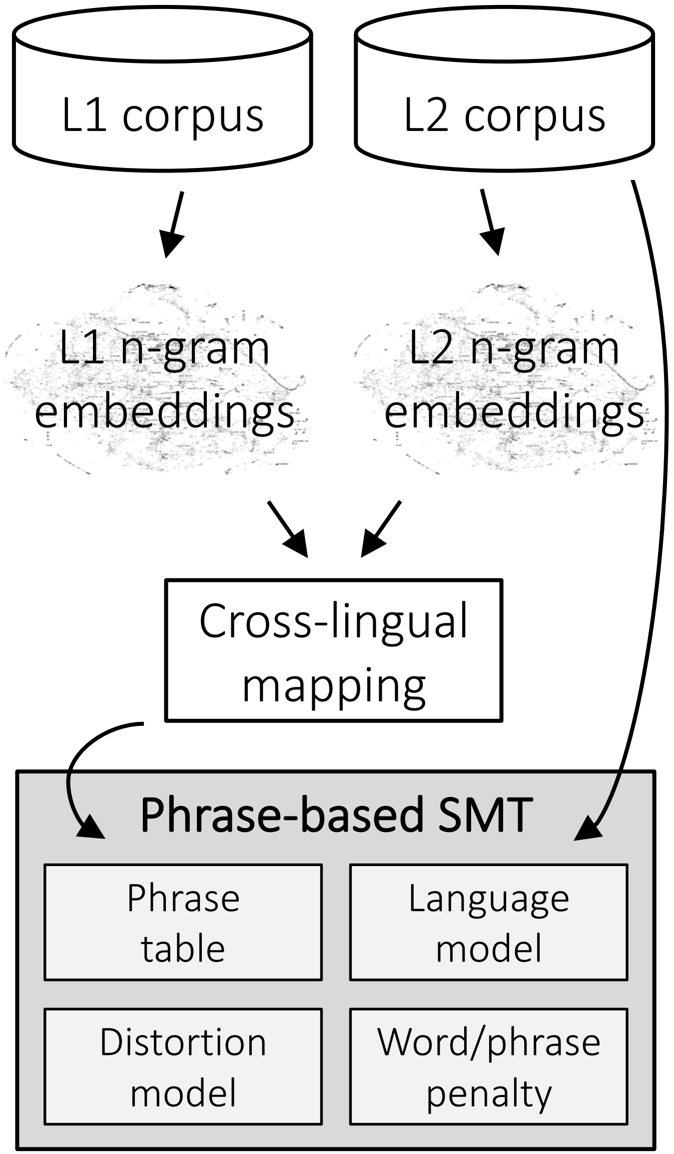
Unsupervised phrase-based SMT

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

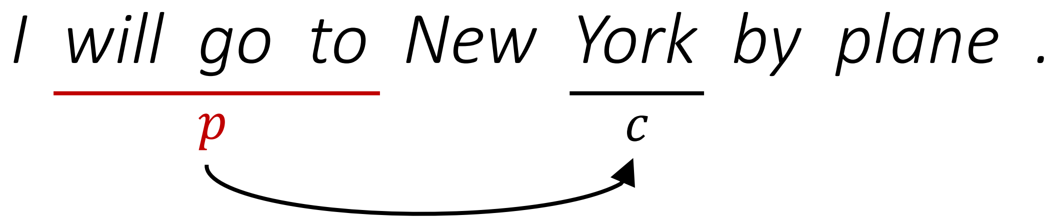
I $\underbrace{\textit{will go}}_{p}$ *to New* $\underbrace{\textit{York}}_{c}$ *by plane .*

Unsupervised phrase-based SMT

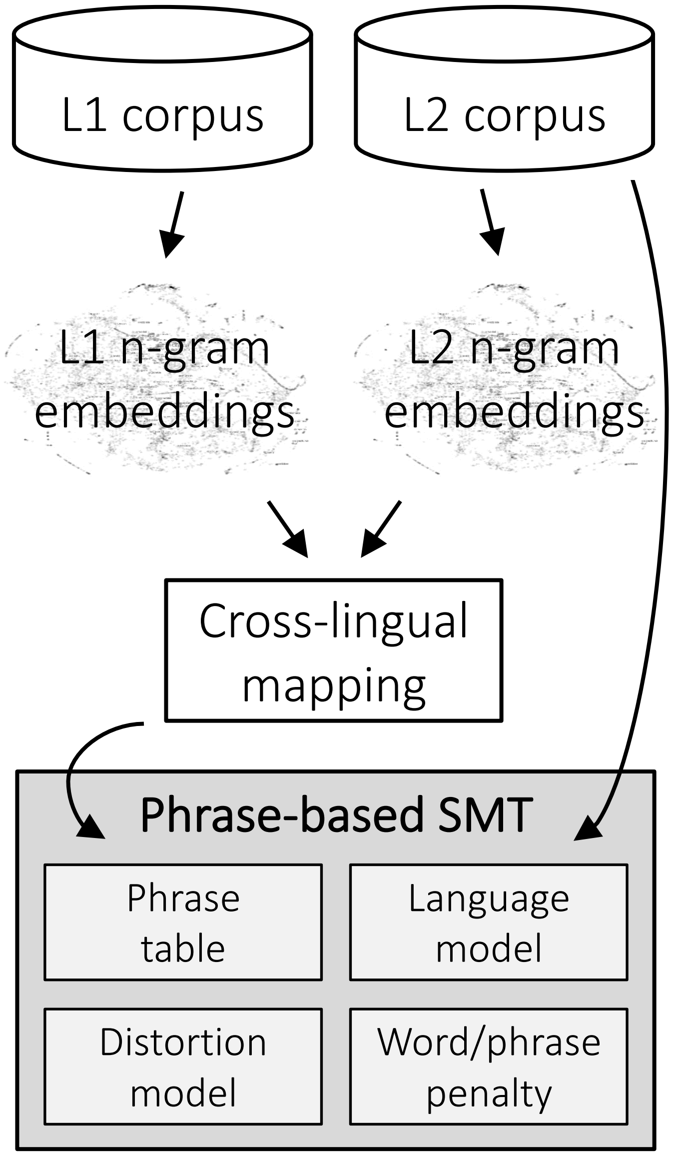


Learn components from monolingual corpora

- 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

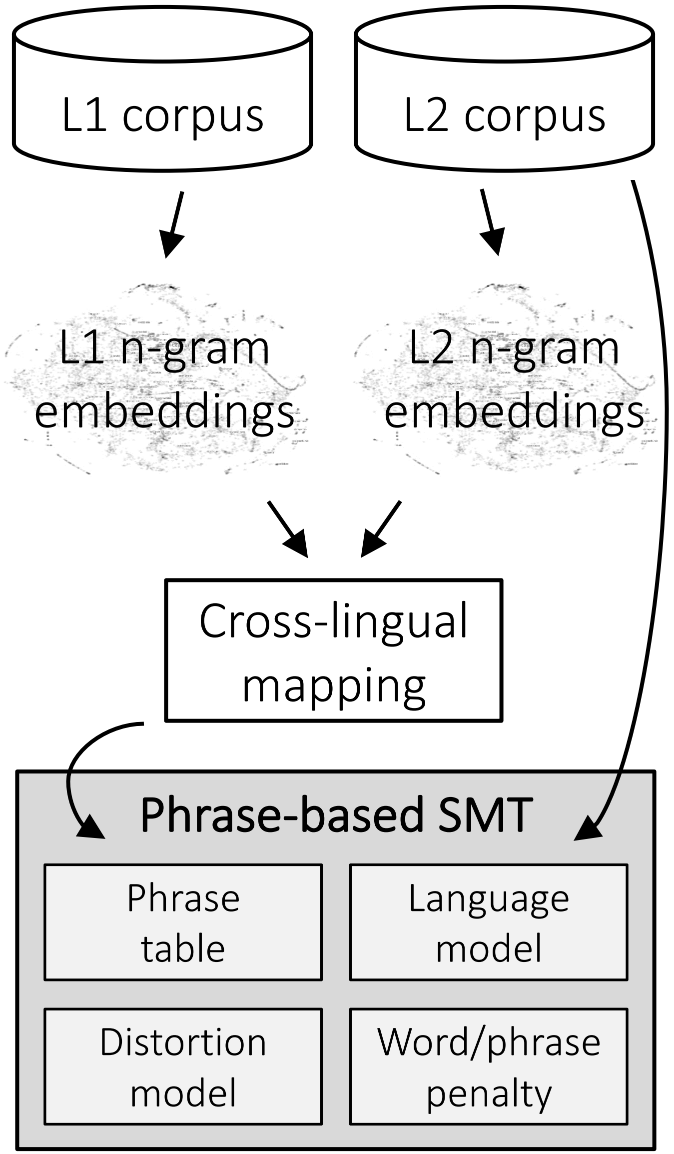


Unsupervised phrase-based SMT



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

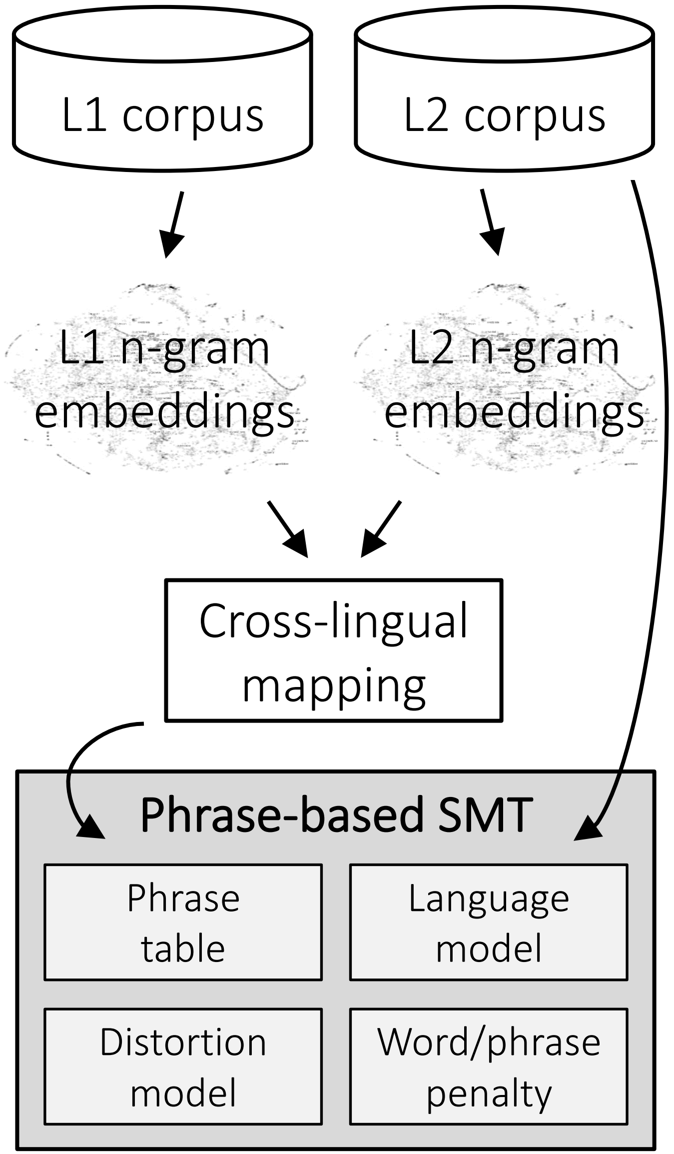


Unsupervised phrase-based SMT

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:



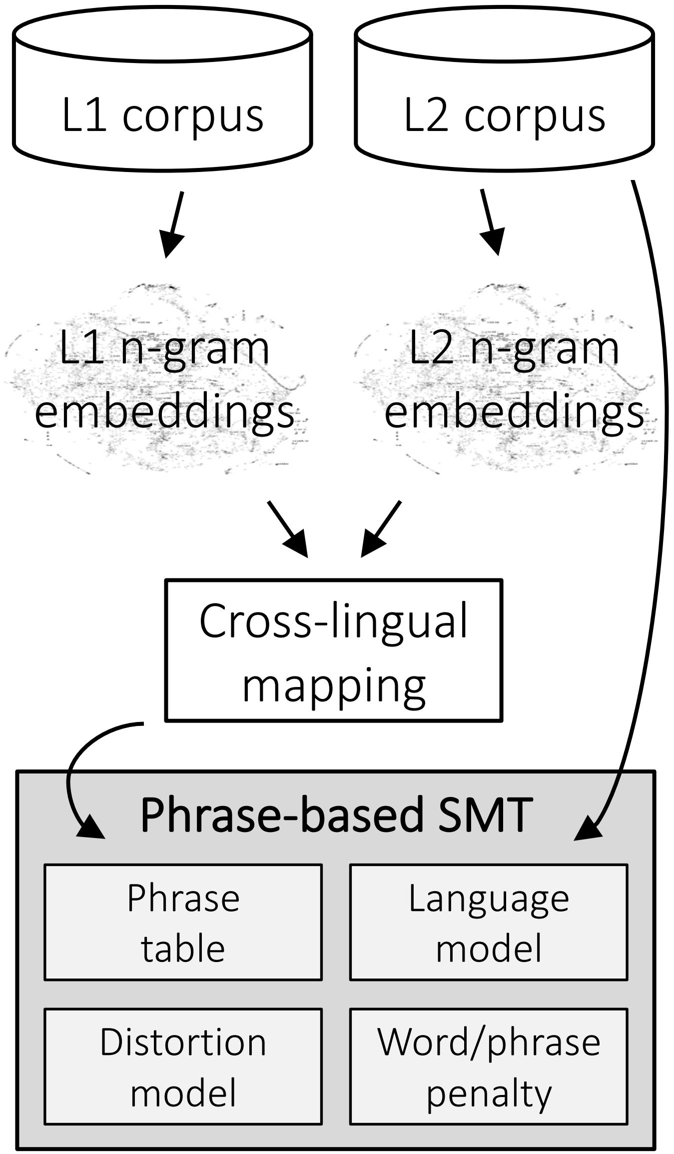
Unsupervised phrase-based SMT

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:

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



Unsupervised phrase-based SMT

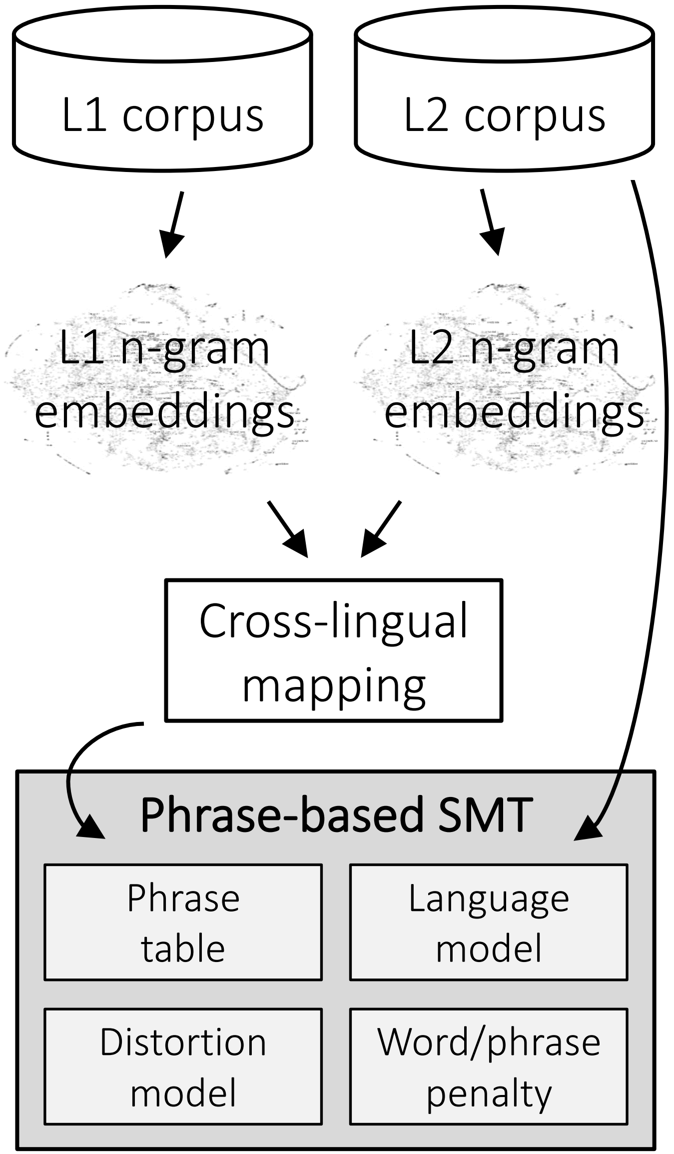
Learn components from monolingual corpora

- 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:

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

$$\min_{\tau} \sum_{\bar{f}} \log \phi\left(\bar{f}|\mathrm{NN}_{\bar{e}}(\bar{f})\right) + \sum_{\bar{e}} \log \phi\left(\bar{e}|\mathrm{NN}_{\bar{f}}(\bar{e})\right)$$

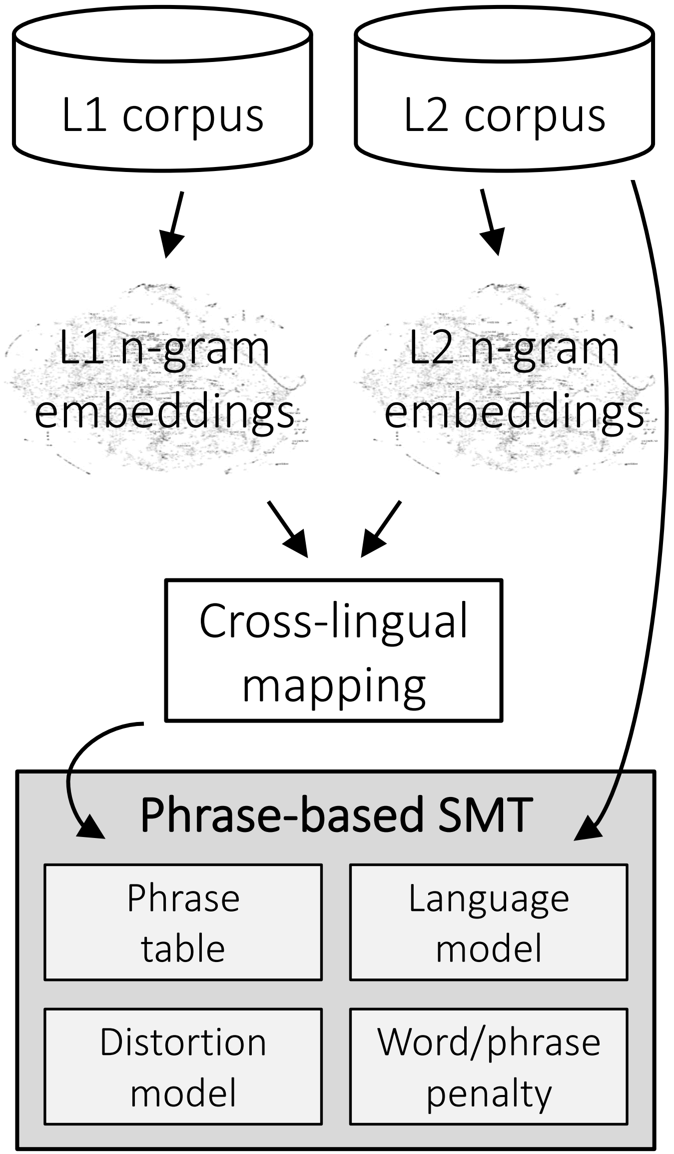


Unsupervised phrase-based SMT

Learn components from monolingual corpora

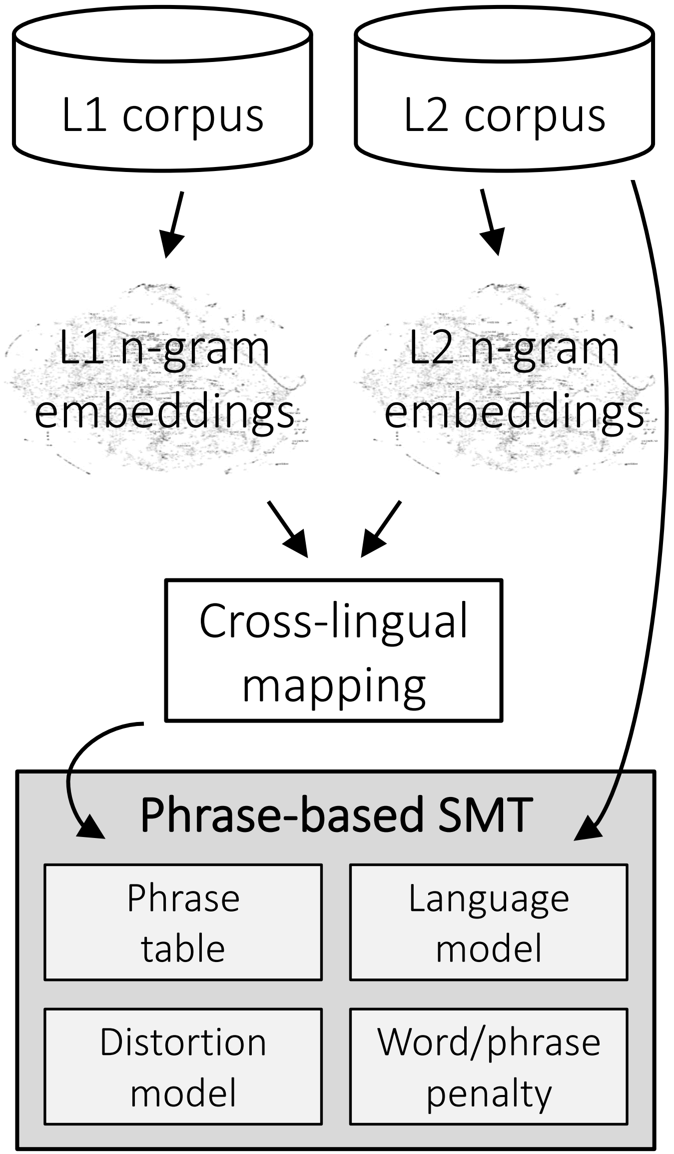
- Phrase table TRICKY...
  - Direct/inverse translation probabilities
  - Direct/inverse lexical weightings
- Language model EASY!!!
  - N-gram frequency counts with back-off and smoothing
- Reordering model EASY!!!
  - Distortion model (distance based)
  - ~~Lexical reordering model~~
- Word/phrase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

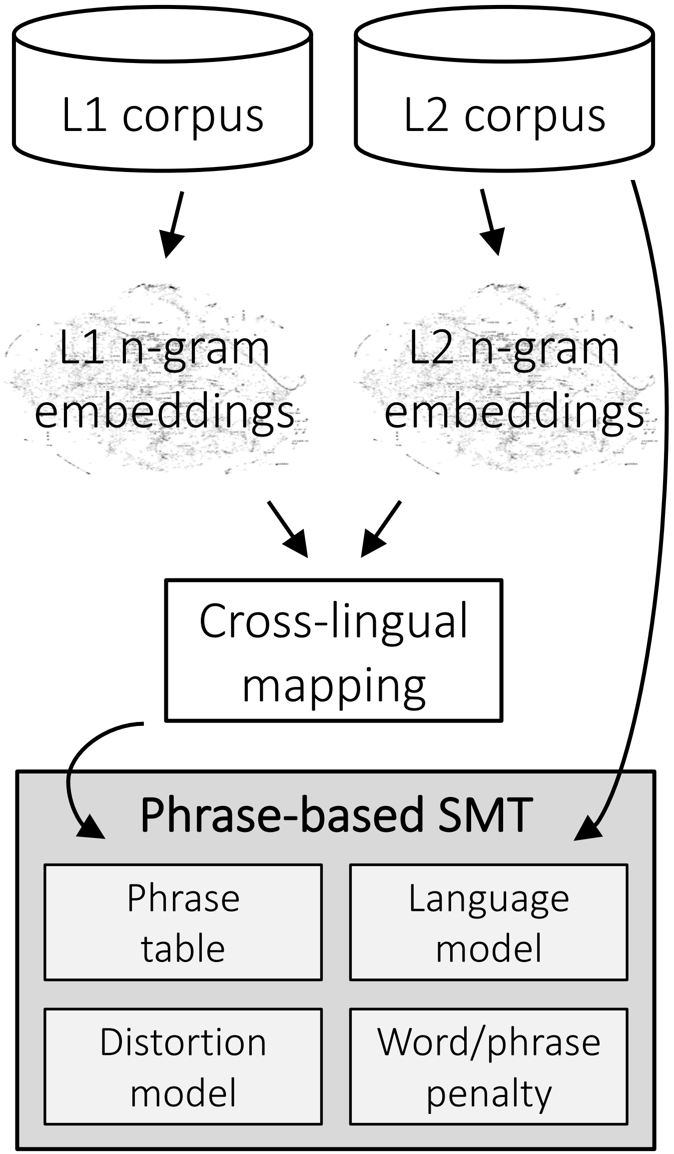


Unsupervised phrase-based SMT

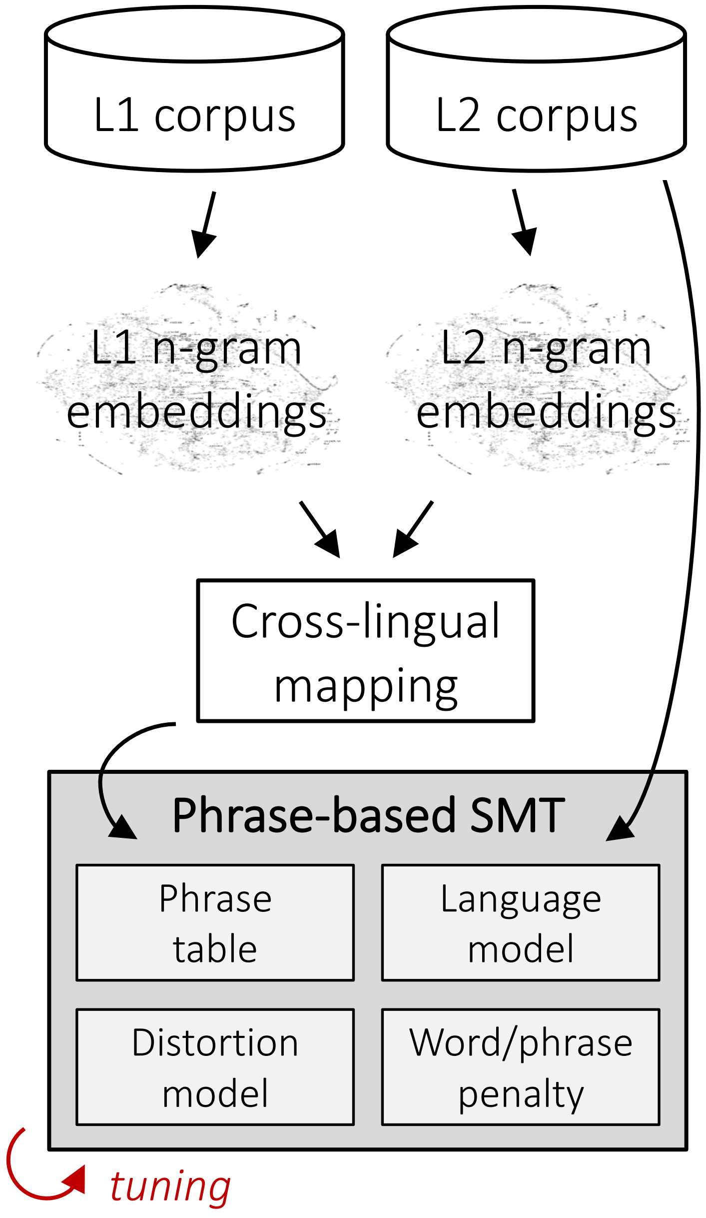
Learn components from monolingual corpora

- 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

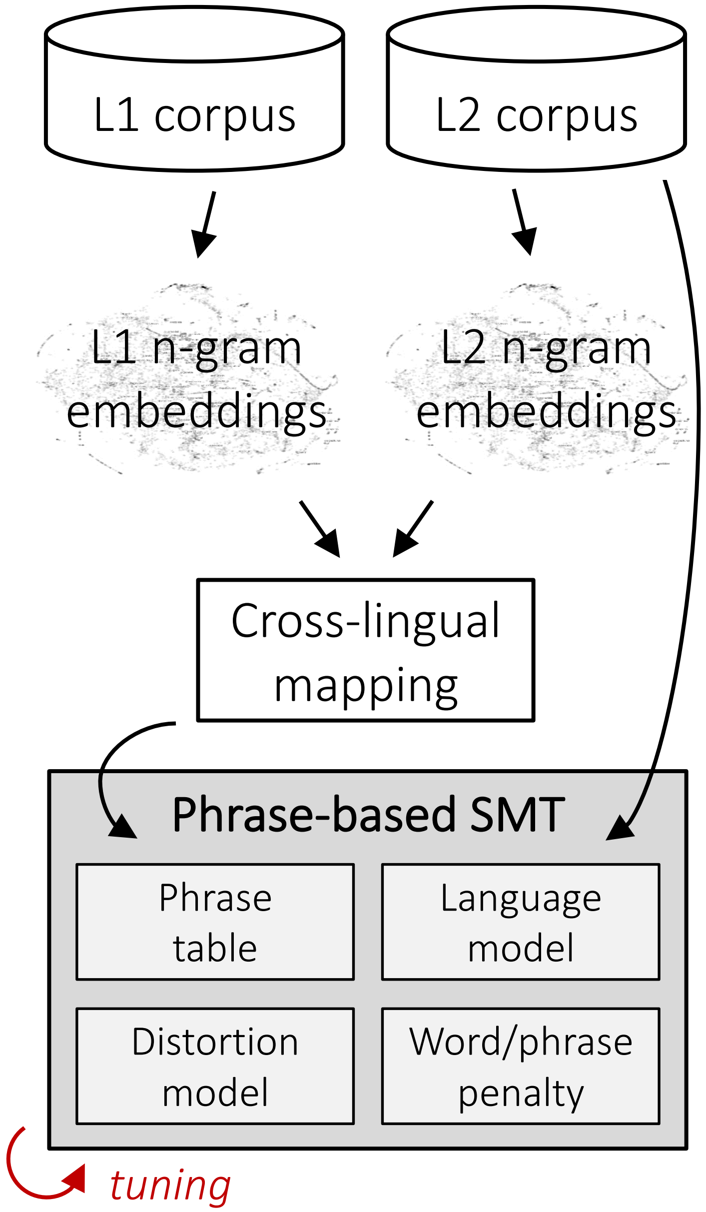
Unsupervised phrase-based SMT



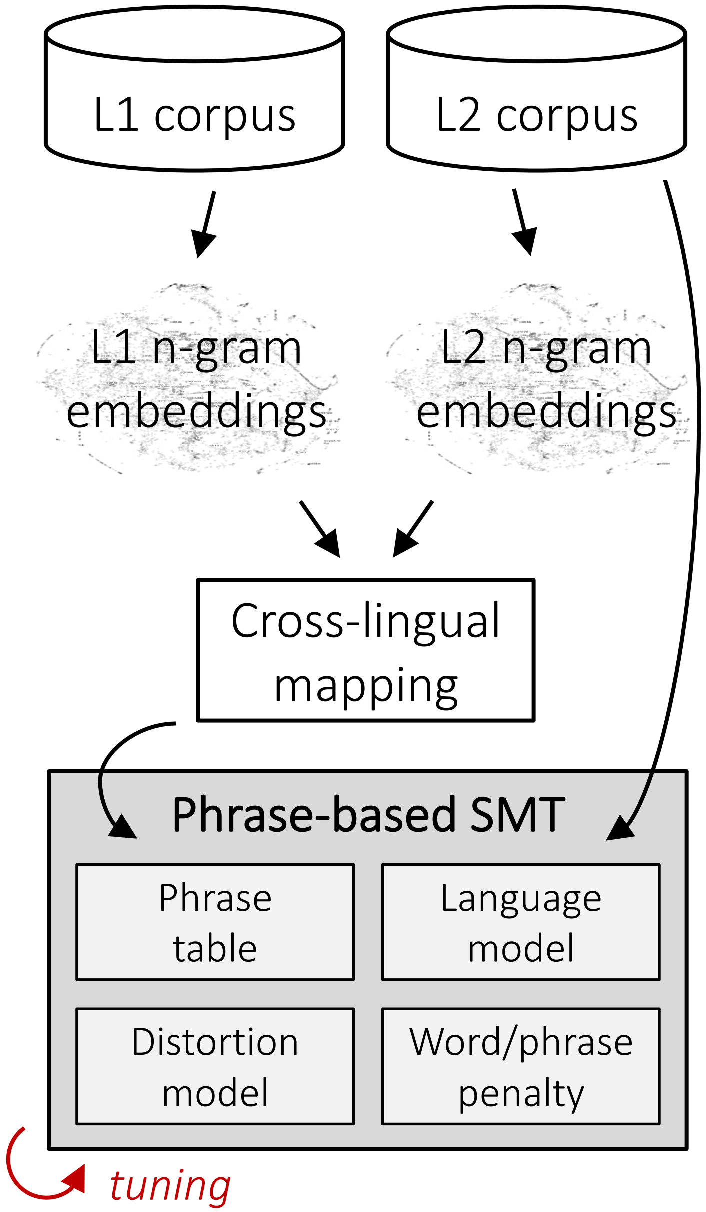
Unsupervised phrase-based SMT



Unsupervised phrase-based SMT



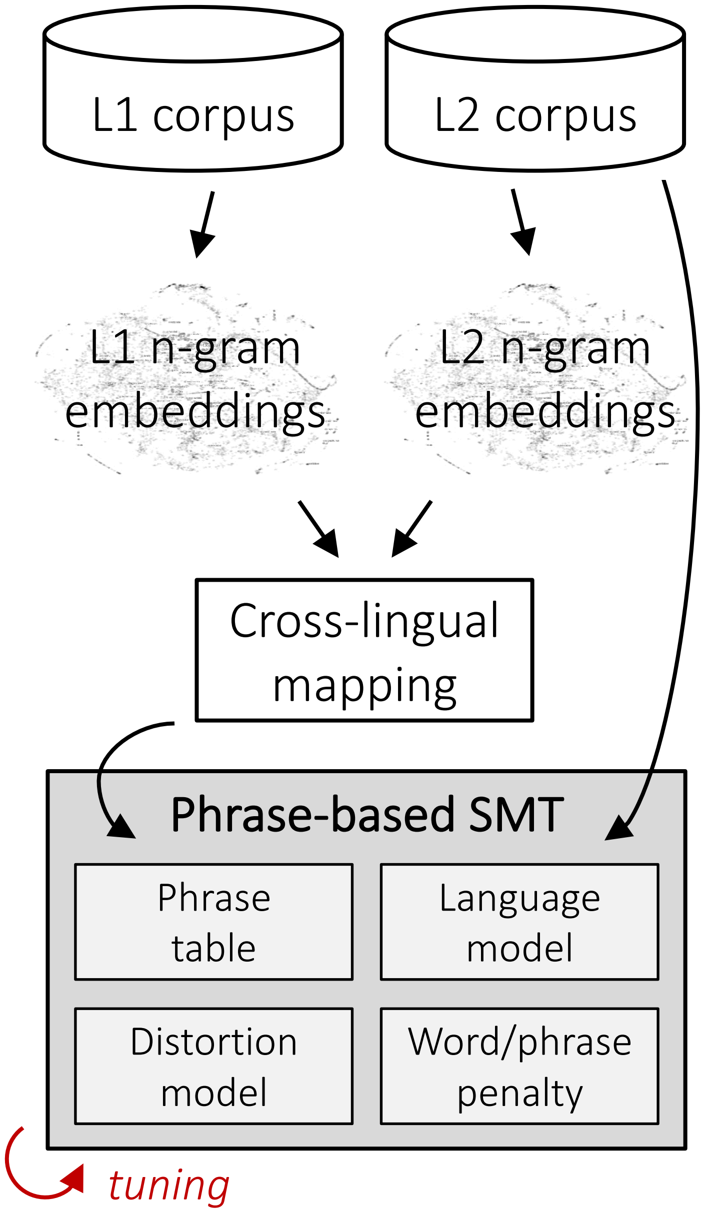
Goal: Adjust the weights of the resulting log-linear model to optimize some evaluation metric (e.g. BLEU)



Unsupervised phrase-based SMT

Goal: Adjust the weights of the resulting log-linear model to optimize some evaluation metric (e.g. BLEU)

... but we don't have a parallel development set!

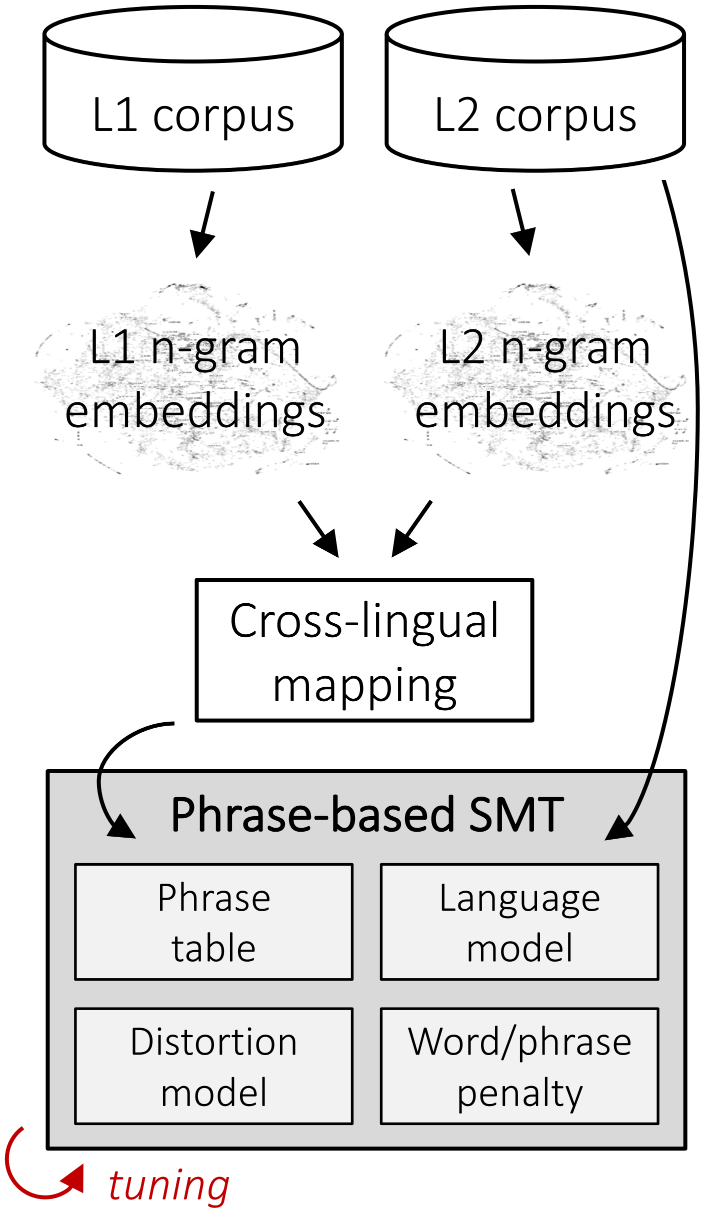


Unsupervised phrase-based SMT

Goal: Adjust the weights of the resulting log-linear model to optimize some evaluation metric (e.g. BLEU)

... but we don't have a parallel development set!

Solutions:



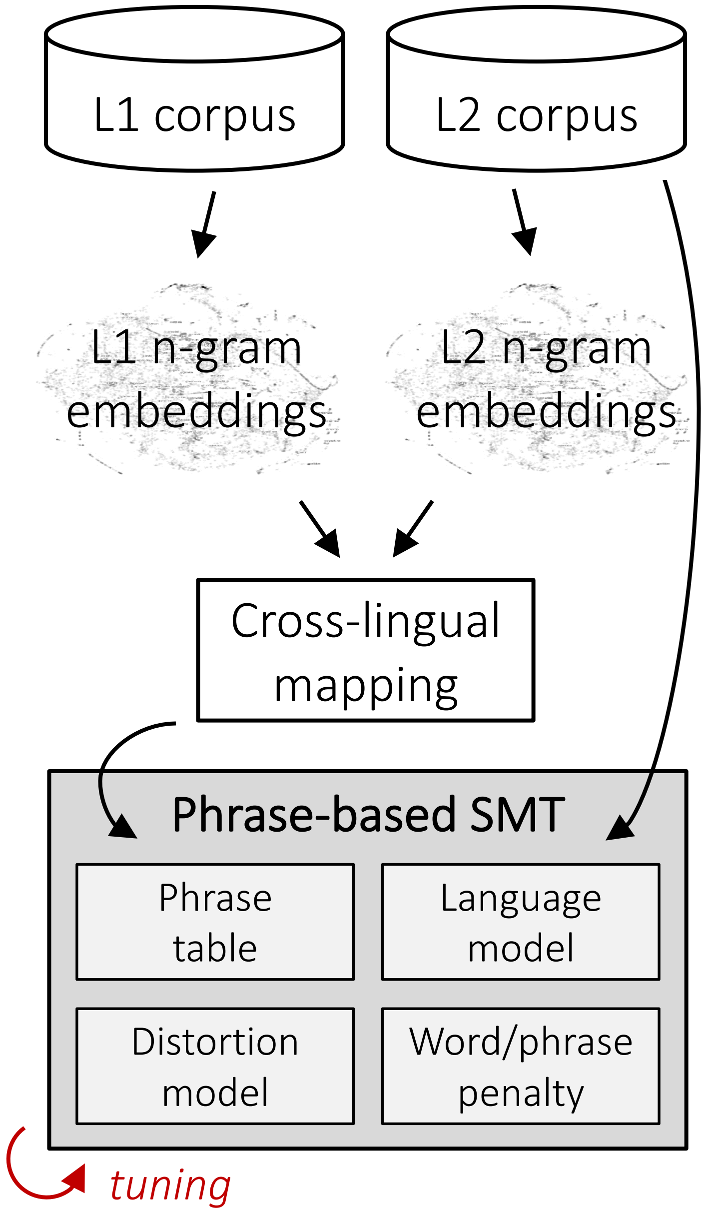
Unsupervised phrase-based SMT

Goal: Adjust the weights of the resulting log-linear model to optimize some evaluation metric (e.g. BLEU)

… but we don't have a parallel development set!

Solutions:

- Default weights without tuning (Lample et al., EMNLP'18)



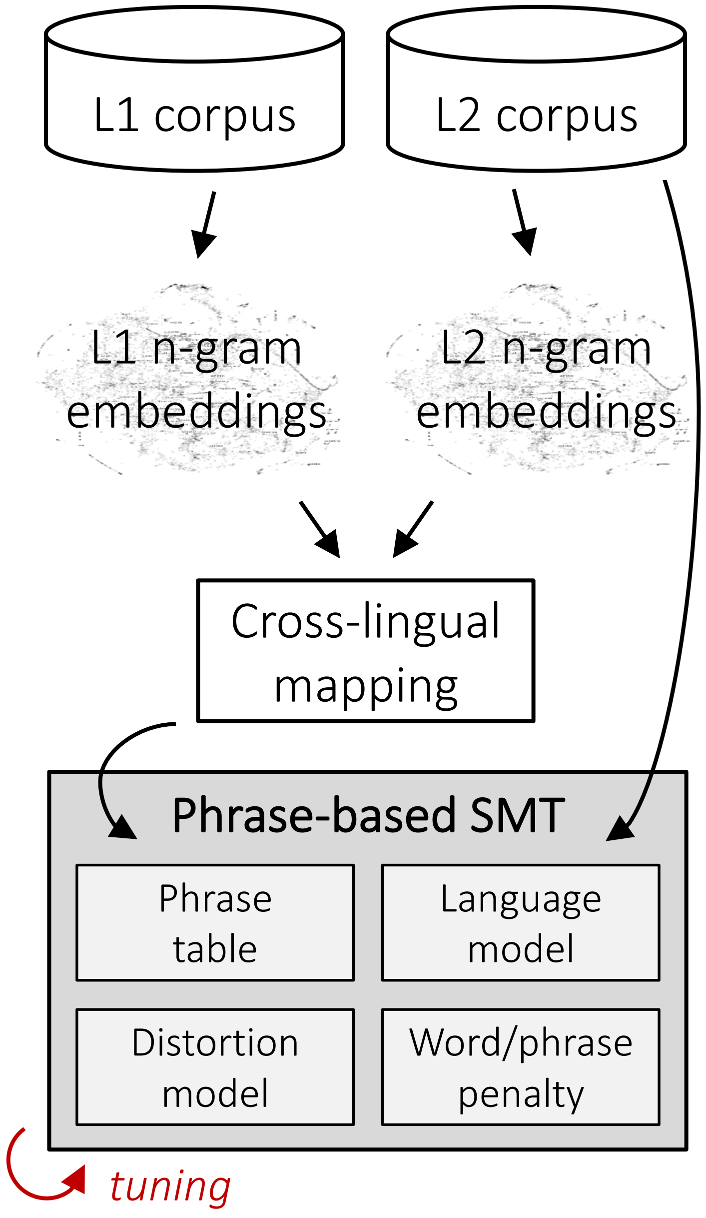
Unsupervised phrase-based SMT

Goal: Adjust the weights of the resulting log-linear model to optimize some evaluation metric (e.g. BLEU)

… but we don't have a parallel development set!

Solutions:

- Default weights without tuning (Lample et al., EMNLP'18)
- Alternating optimization with back-translation (Artetxe et al., EMNLP'18)



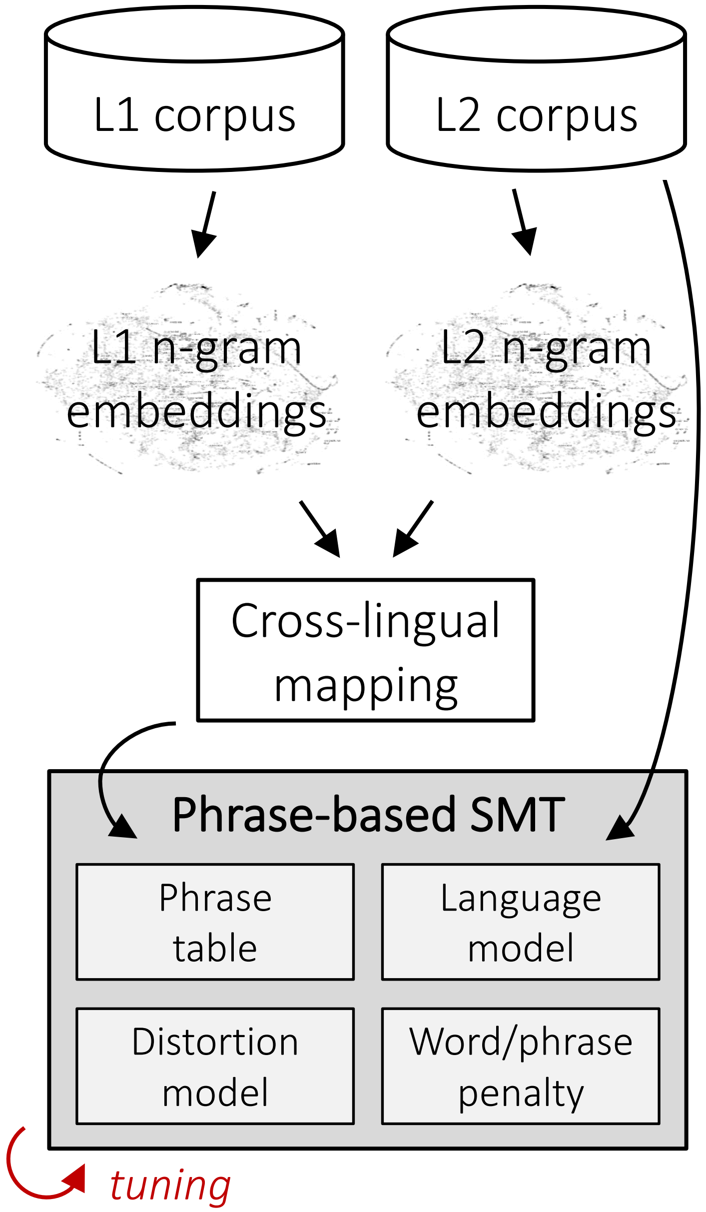
Unsupervised phrase-based SMT

Goal: Adjust the weights of the resulting log-linear model to optimize some evaluation metric (e.g. BLEU)

… but we don't have a parallel development set!

Solutions:

- Default weights without tuning (Lample et al., EMNLP'18)
- Alternating optimization with back-translation (Artetxe et al., EMNLP'18)
- Unsupervised optimization criterion (Artetxe et al., ACL'19)



Unsupervised phrase-based SMT

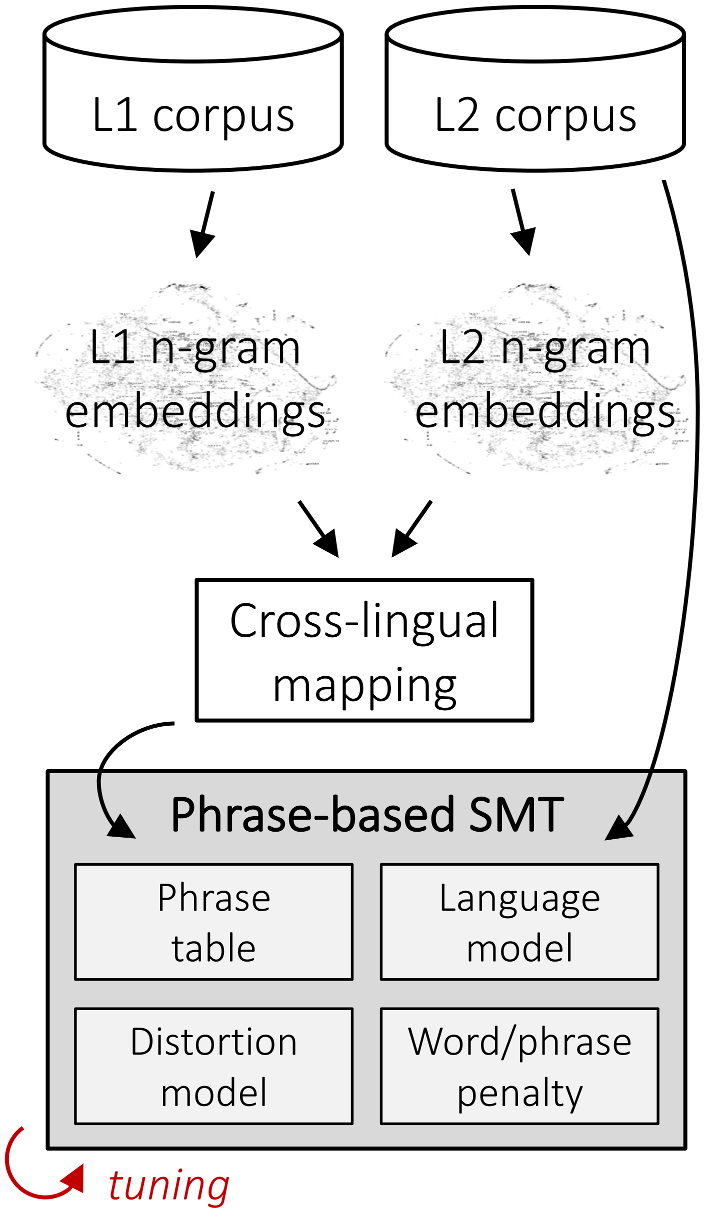
Goal: Adjust the weights of the resulting log-linear model to optimize some evaluation metric (e.g. BLEU)

… but we don't have a parallel development set!

Solutions:

- Default weights without tuning (Lample et al., EMNLP'18)
- Alternating optimization with back-translation (Artetxe et al., EMNLP'18)
- Unsupervised optimization criterion (Artetxe et al., ACL'19)

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



Unsupervised phrase-based SMT

Goal: Adjust the weights of the resulting log-linear model to optimize some evaluation metric (e.g. BLEU)

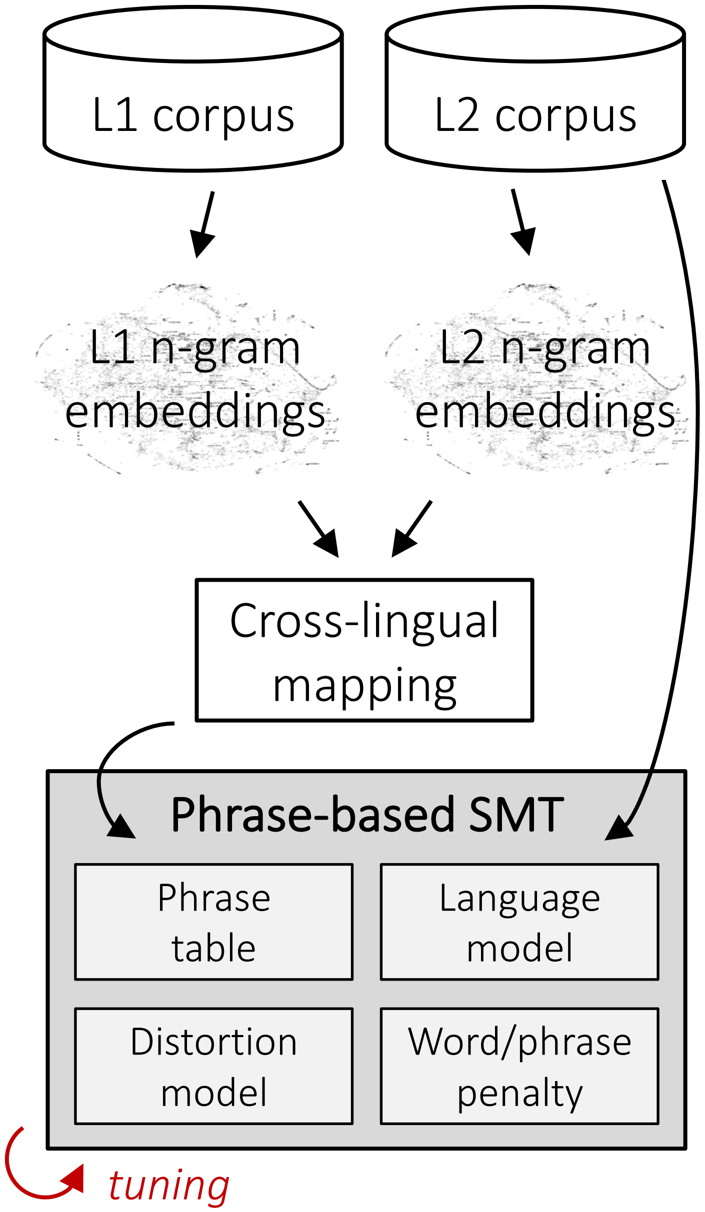
… but we don't have a parallel development set!

Solutions:

- Default weights without tuning (Lample et al., EMNLP'18)
- Alternating optimization with back-translation (Artetxe et al., EMNLP'18)
- Unsupervised optimization criterion (Artetxe et al., ACL'19)

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

- $L_{cycle}(E) = 1 - \mathrm{BLEU}\big(\mathrm{T}_{F\rightarrow E}\big(\mathrm{T}_{E\rightarrow F}(E)\big), E\big)$



Unsupervised phrase-based SMT

Goal: Adjust the weights of the resulting log-linear model to optimize some evaluation metric (e.g. BLEU)

... but we don't have a parallel development set!

Solutions:

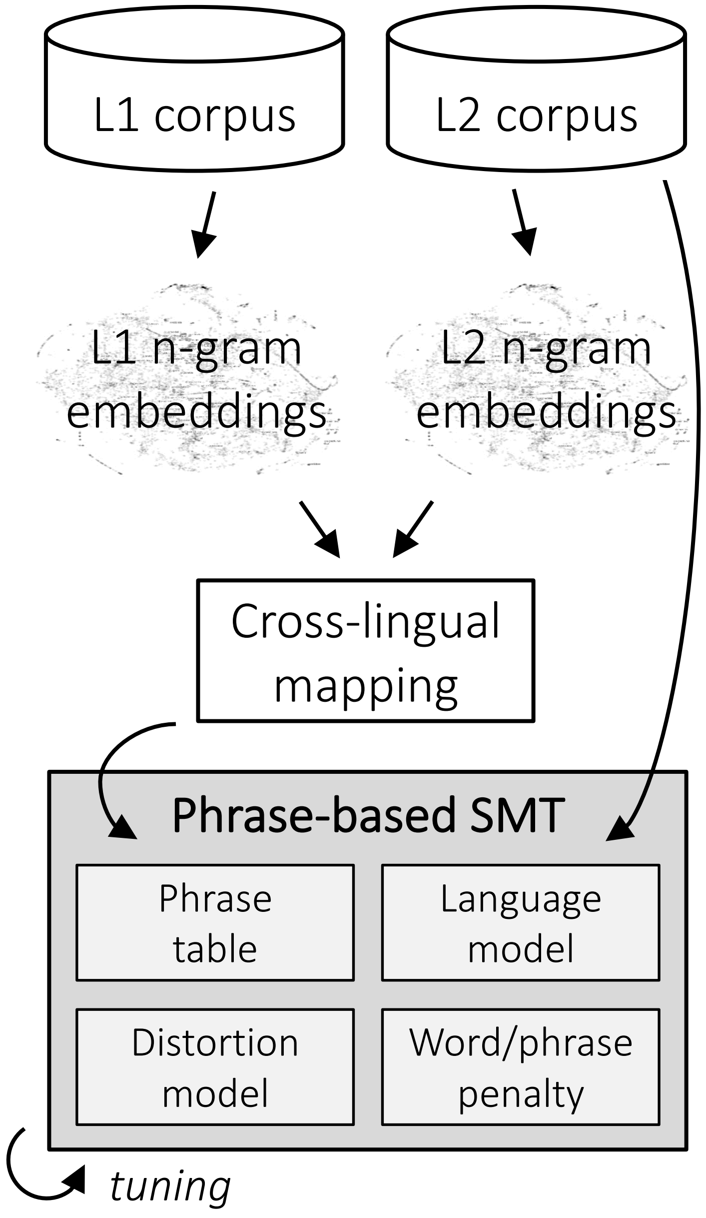
- Default weights without tuning (Lample et al., EMNLP'18)
- Alternating optimization with back-translation (Artetxe et al., EMNLP'18)
- Unsupervised optimization criterion (Artetxe et al., ACL'19)

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

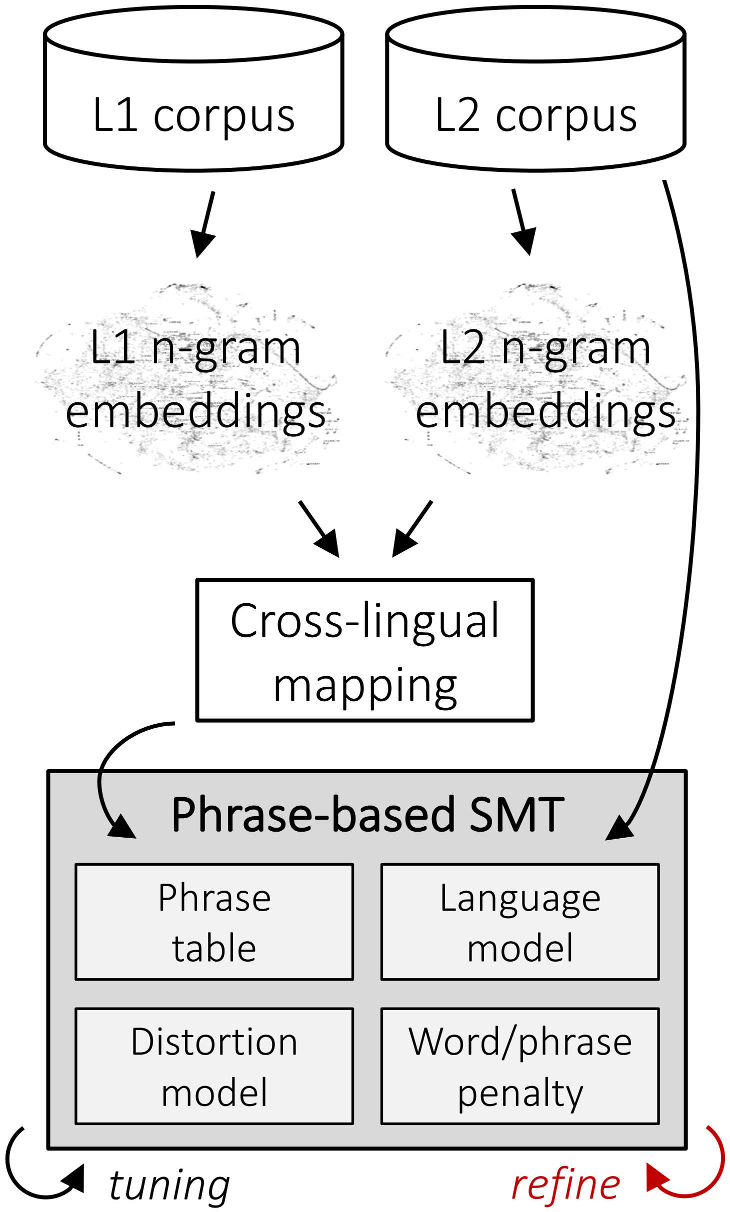
- $L_{cycle}(E) = 1 - \mathrm{BLEU}\big(\mathrm{T}_{F\rightarrow E}\big(\mathrm{T}_{E\rightarrow F}(E)\big), E\big)$
- $L_{lm}(E) = \max\Big(0, \mathrm{H}(F) - \mathrm{H}\big(\mathrm{T}_{E\rightarrow F}(E)\big)\Big)^2 \cdot \mathrm{LP}$

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

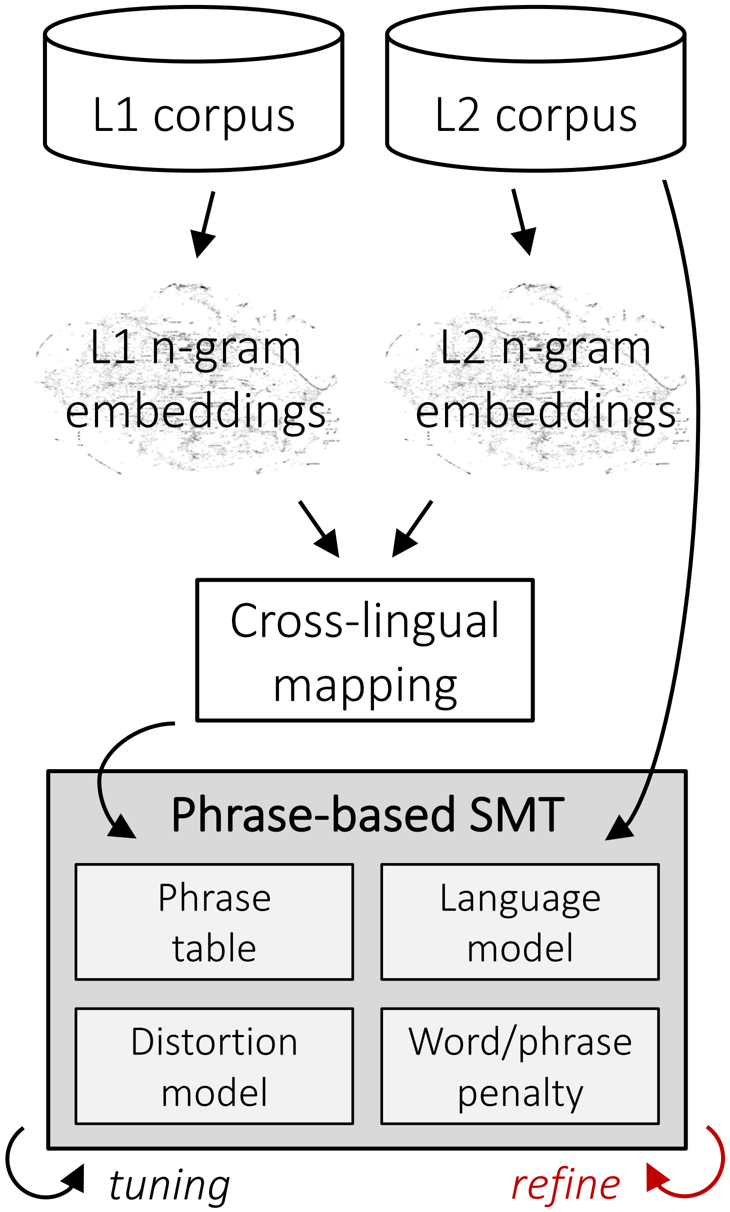
Unsupervised phrase-based SMT



Unsupervised phrase-based SMT

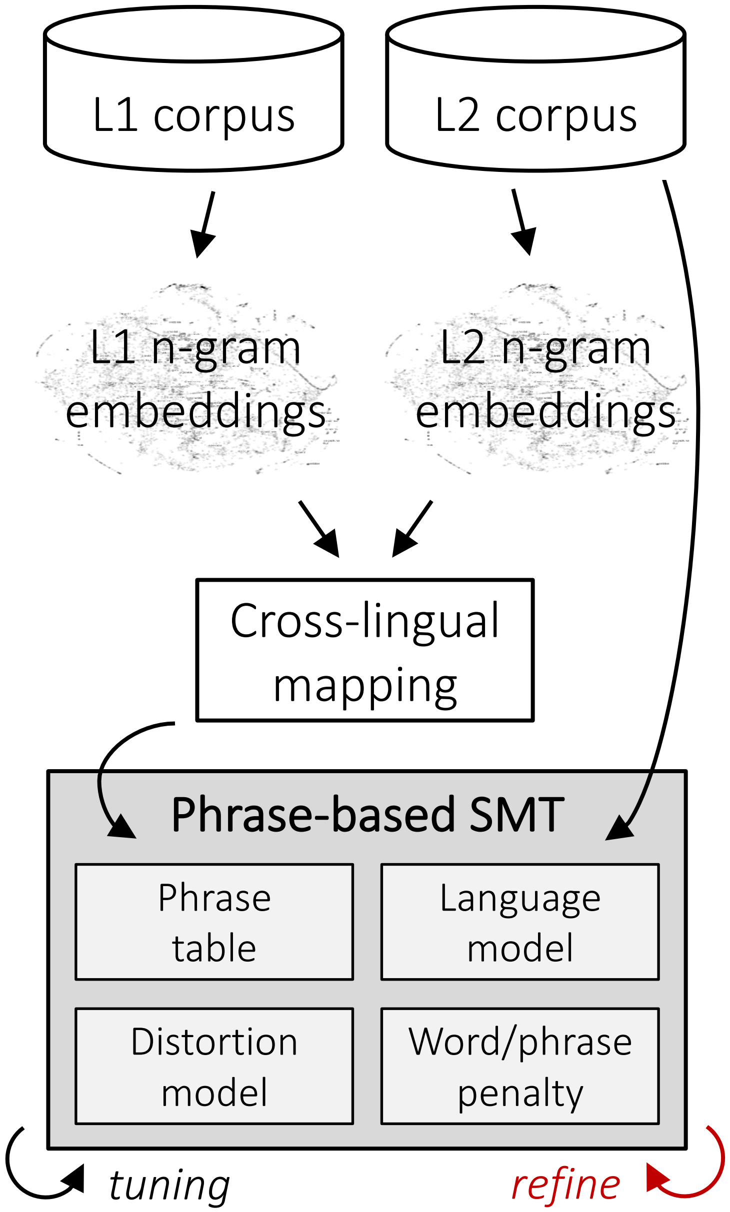


Unsupervised phrase-based SMT

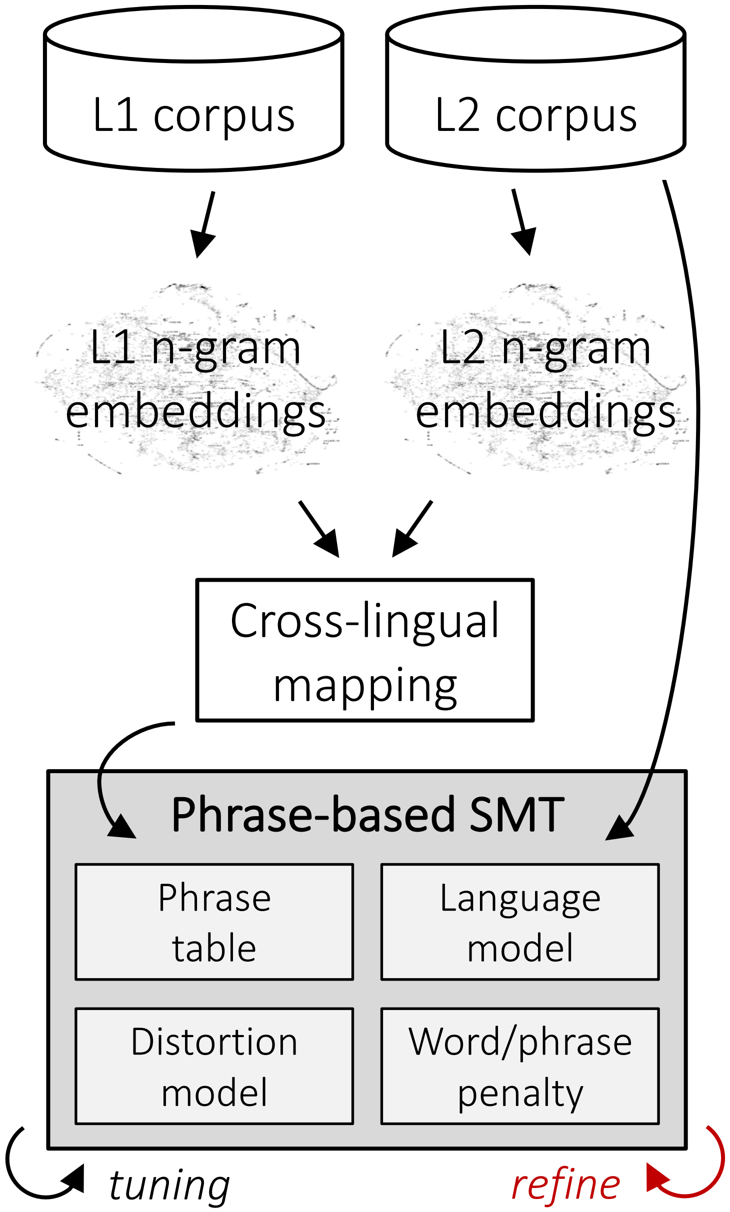


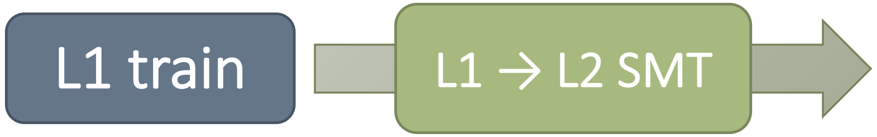
L1 train

Unsupervised phrase-based SMT

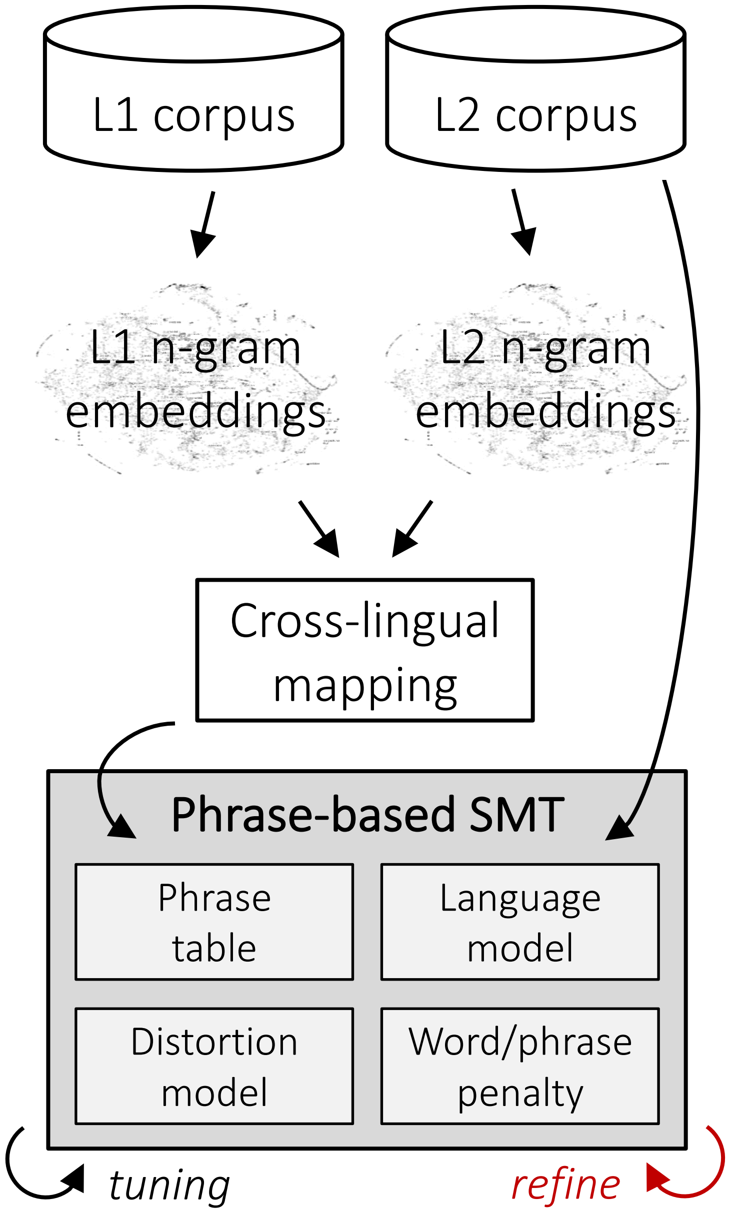


Unsupervised phrase-based SMT

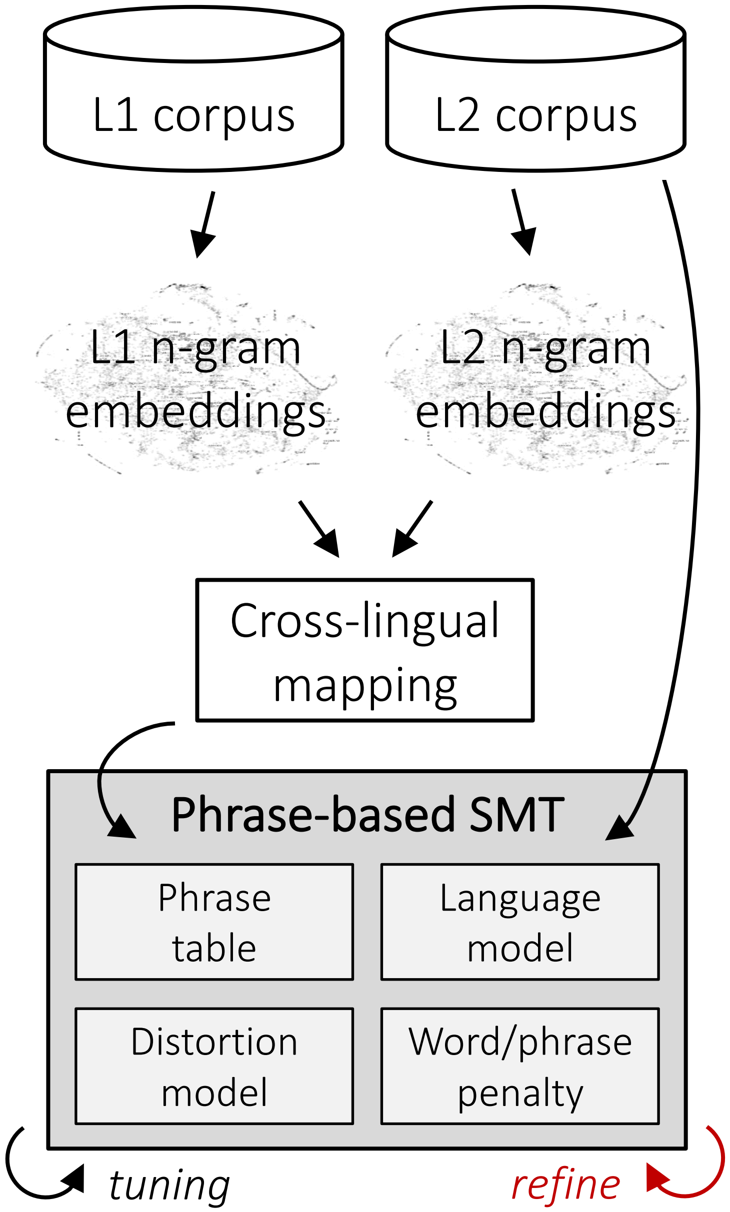


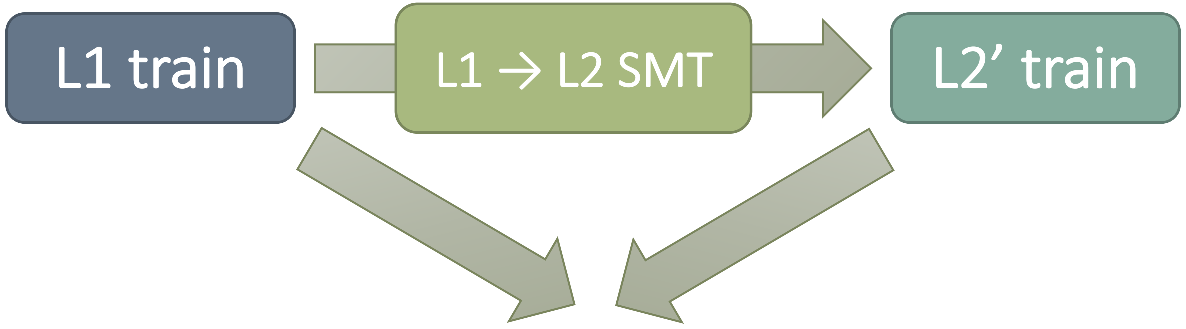


Unsupervised phrase-based SMT

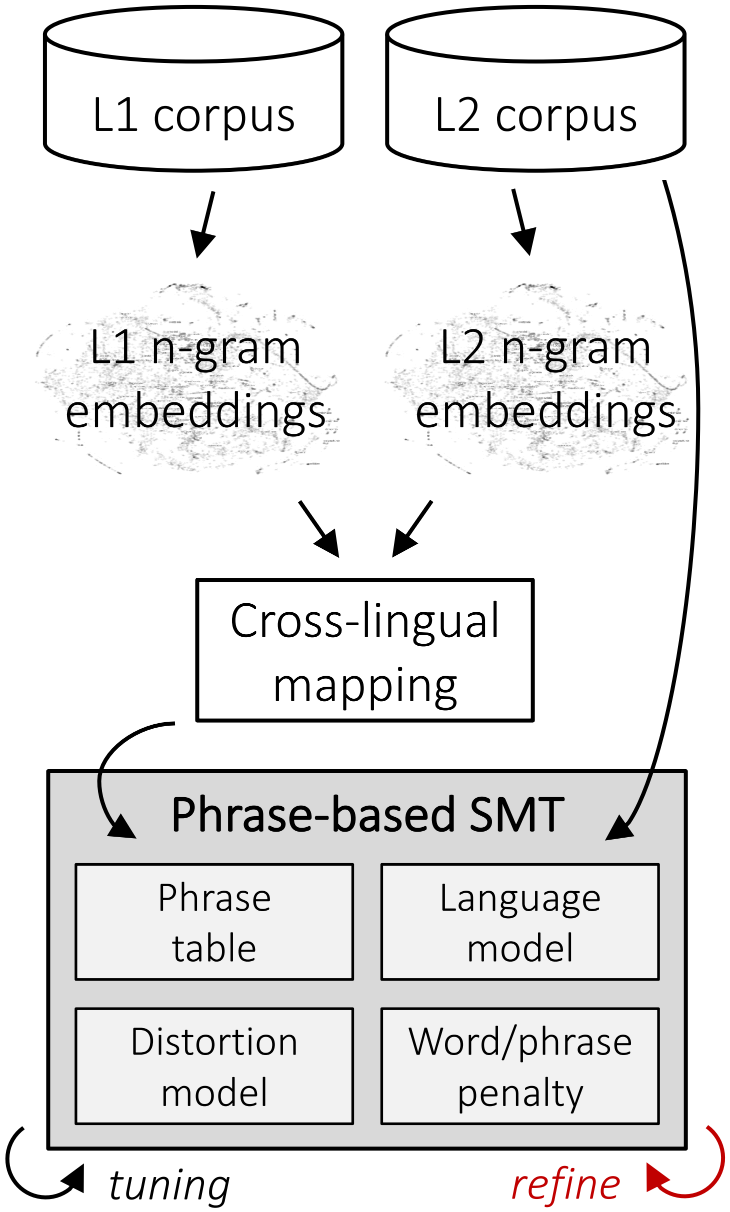


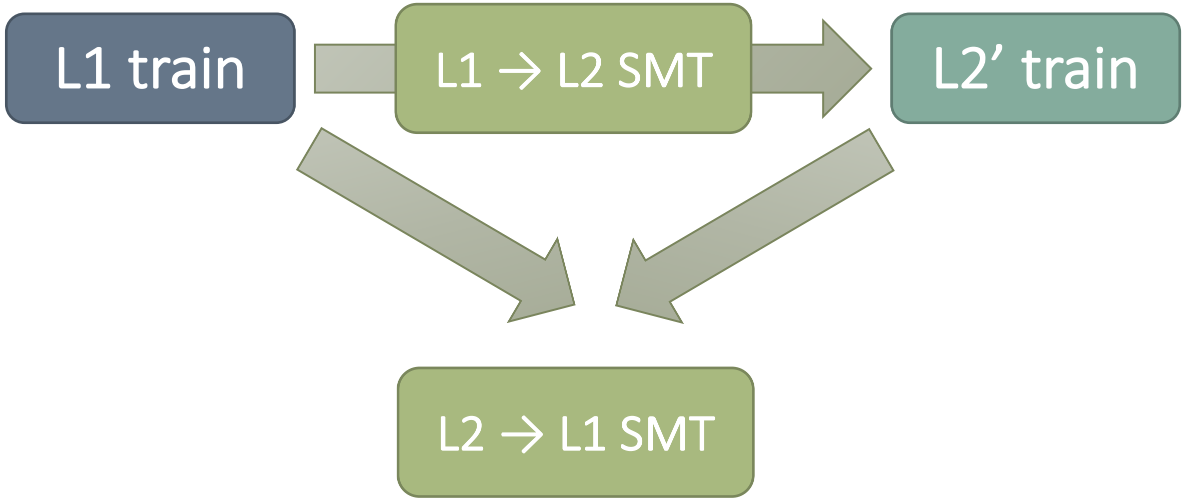
Unsupervised phrase-based SMT



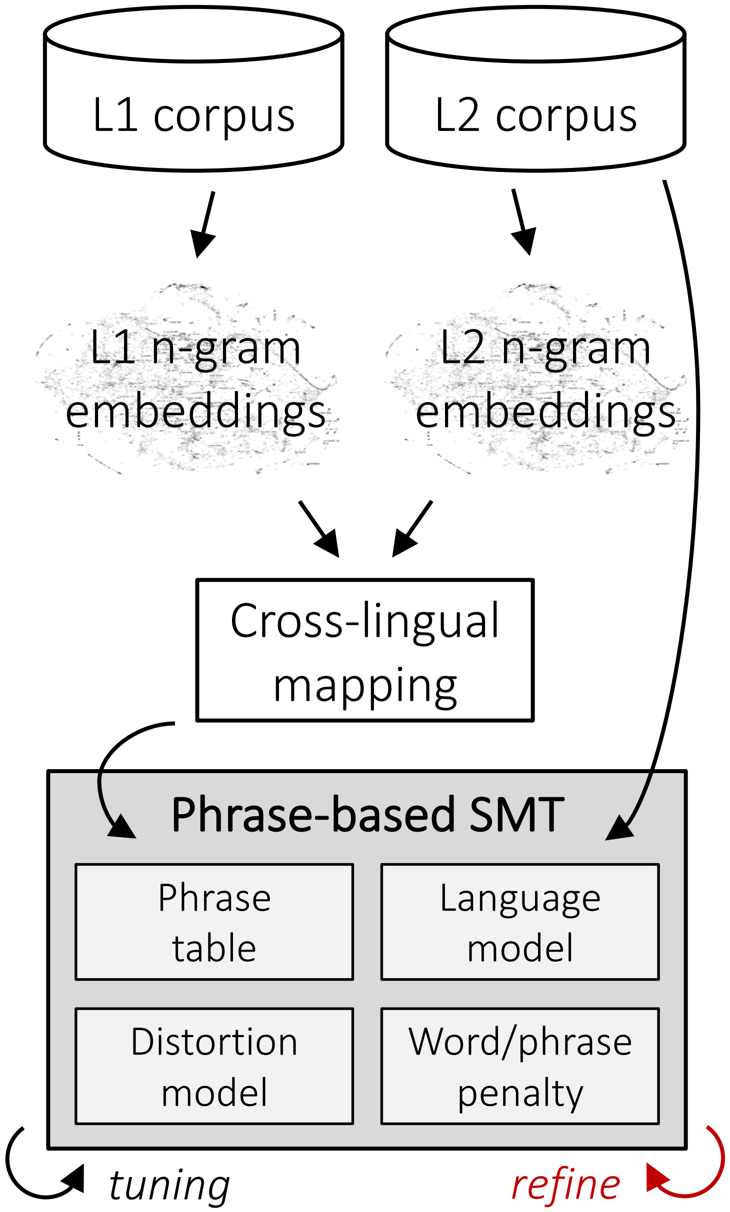


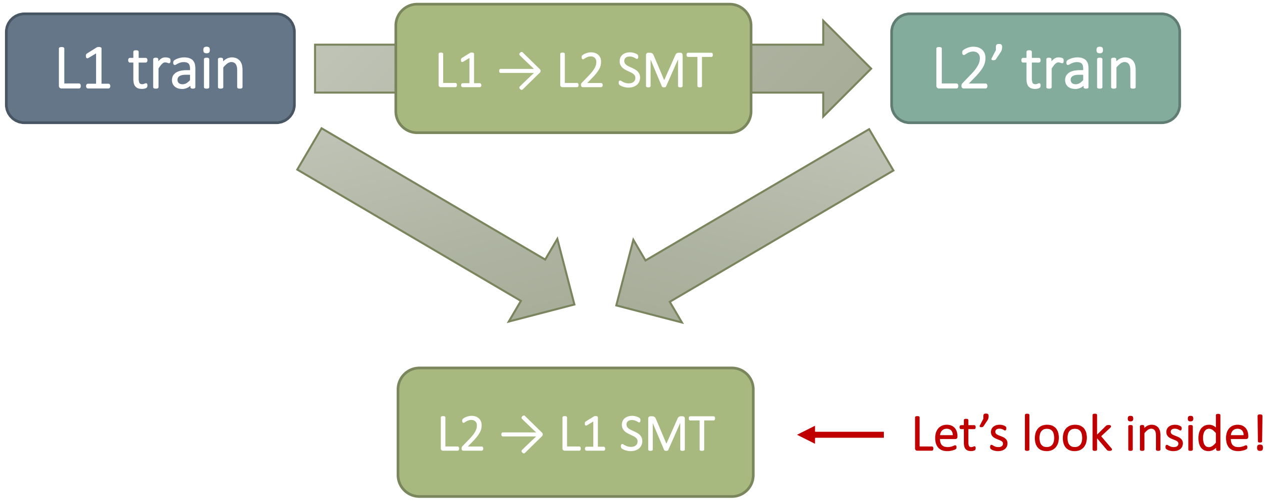
Unsupervised phrase-based SMT



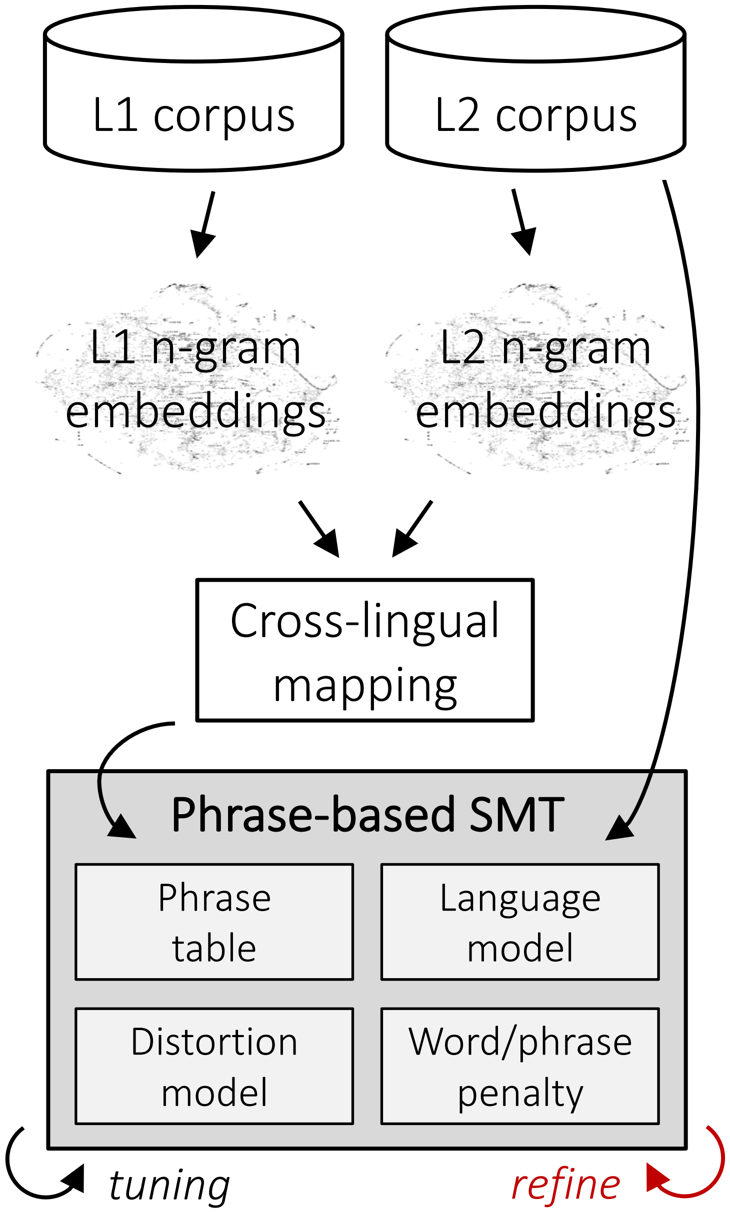


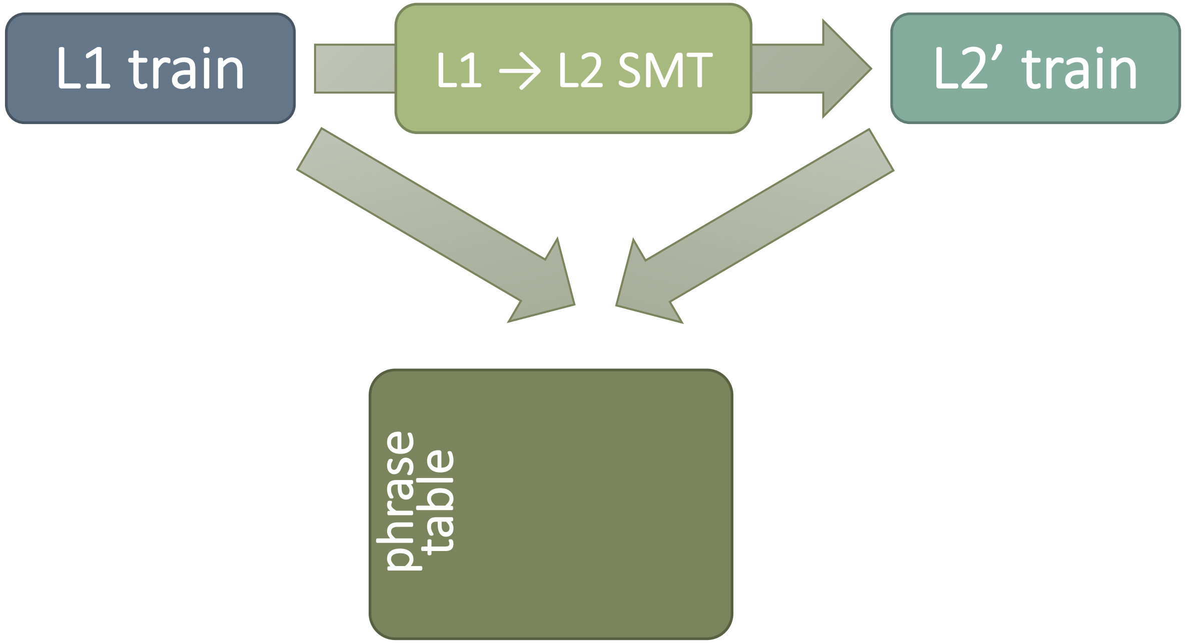
Unsupervised phrase-based SMT



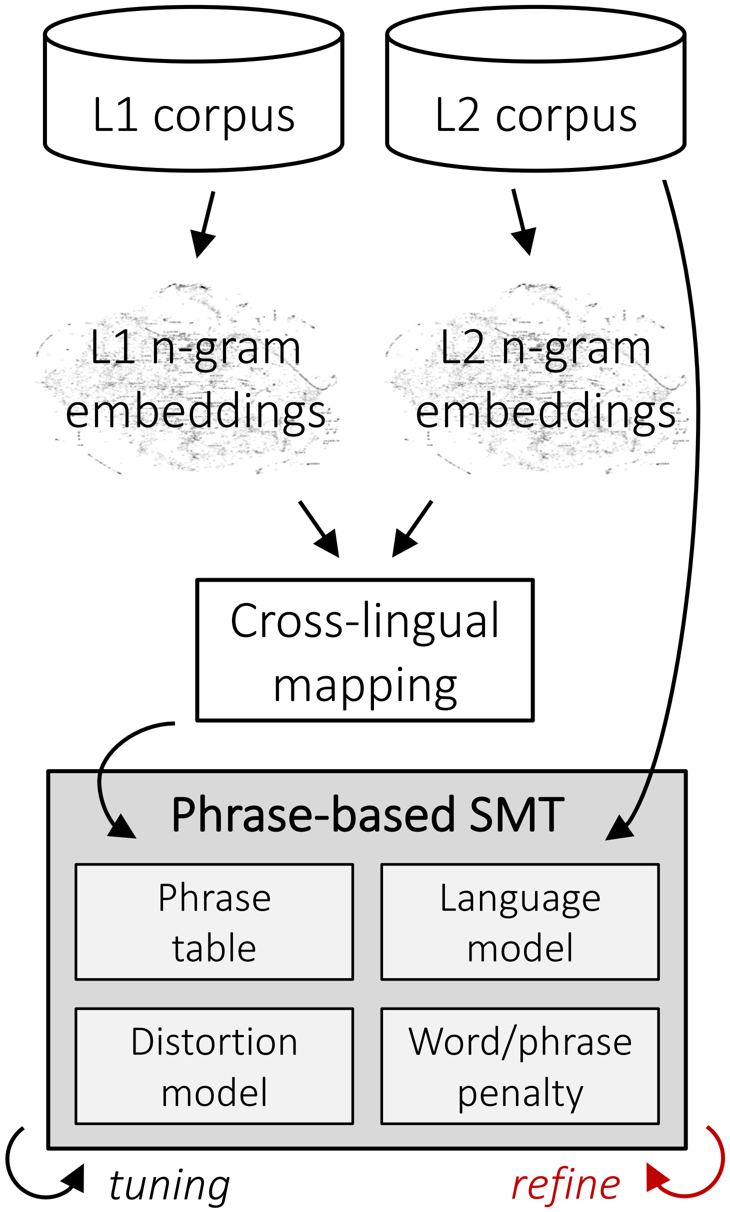


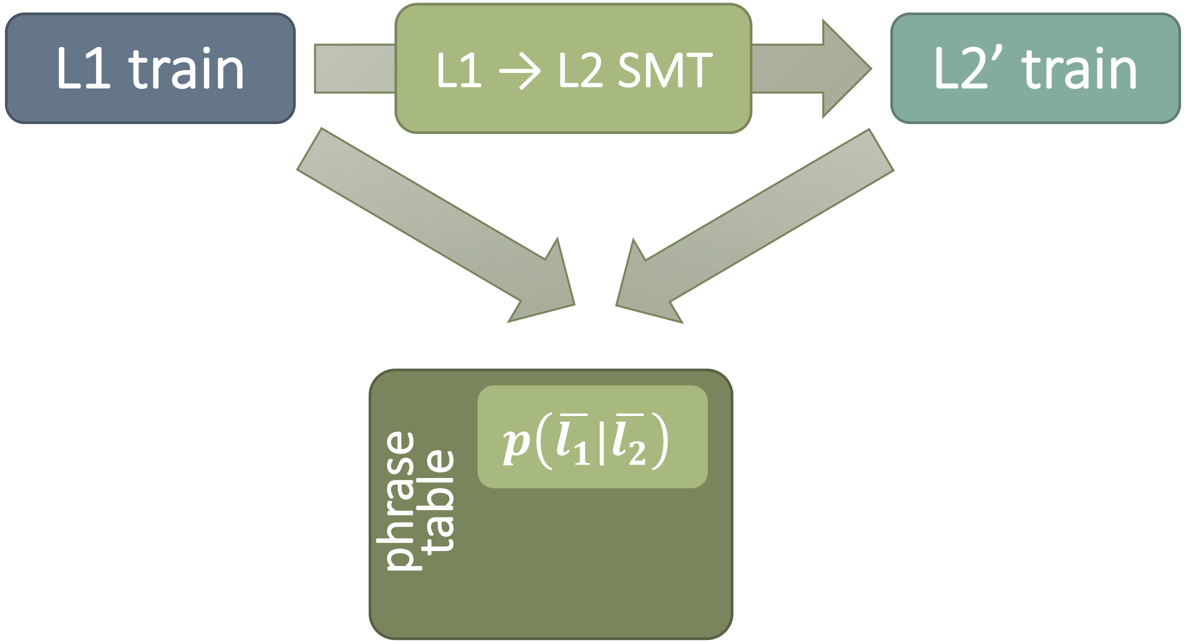
Unsupervised phrase-based SMT



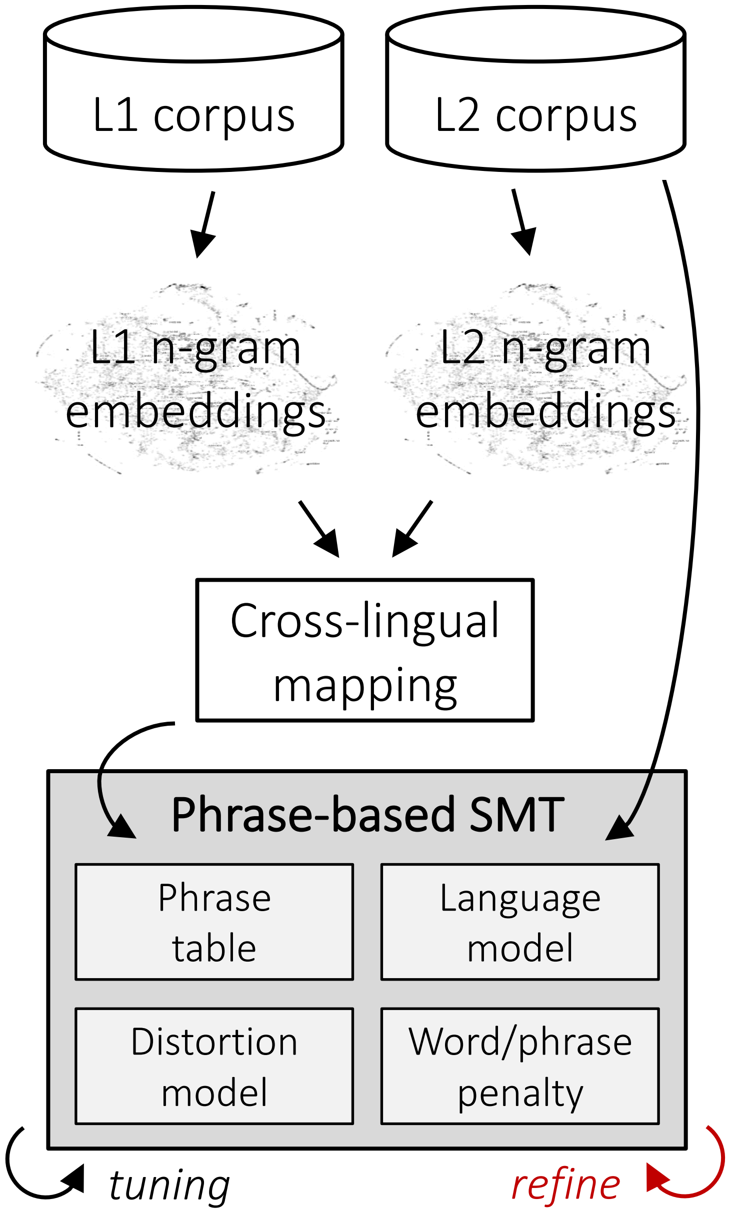


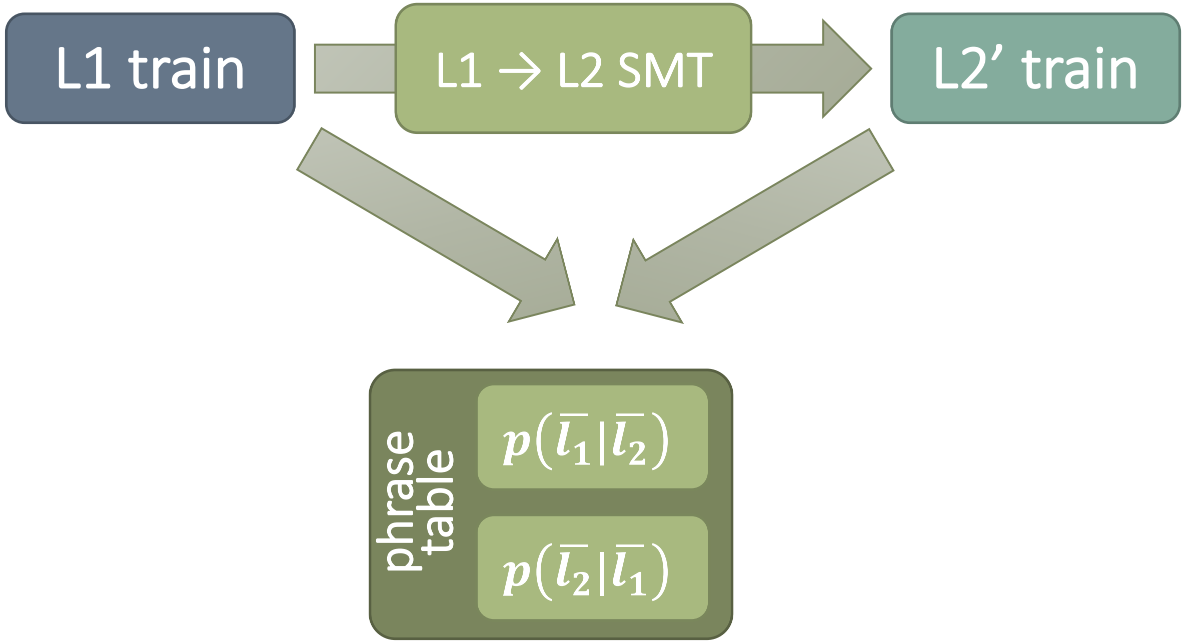
Unsupervised phrase-based SMT



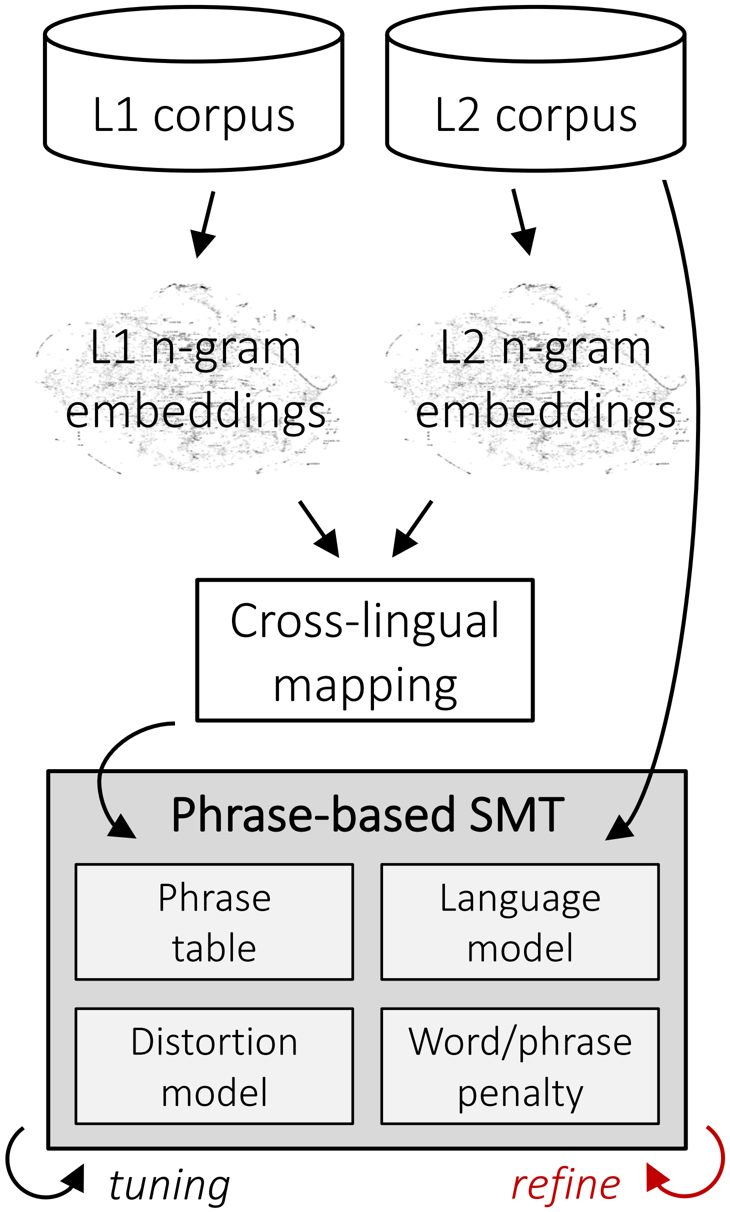


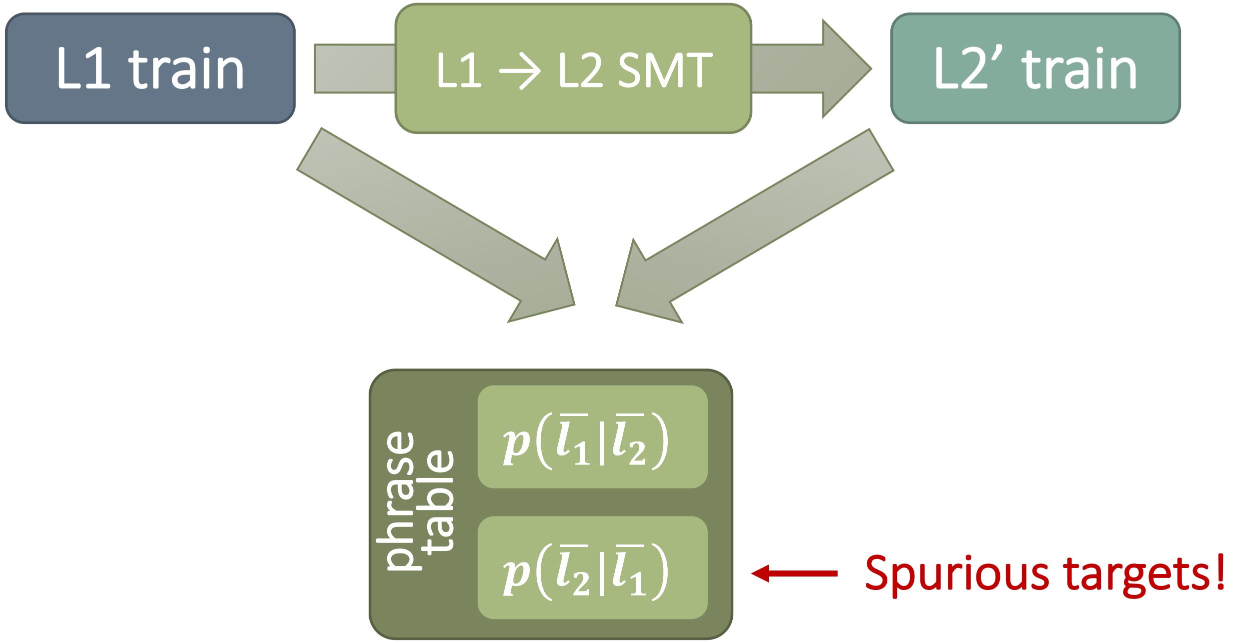
Unsupervised phrase-based SMT

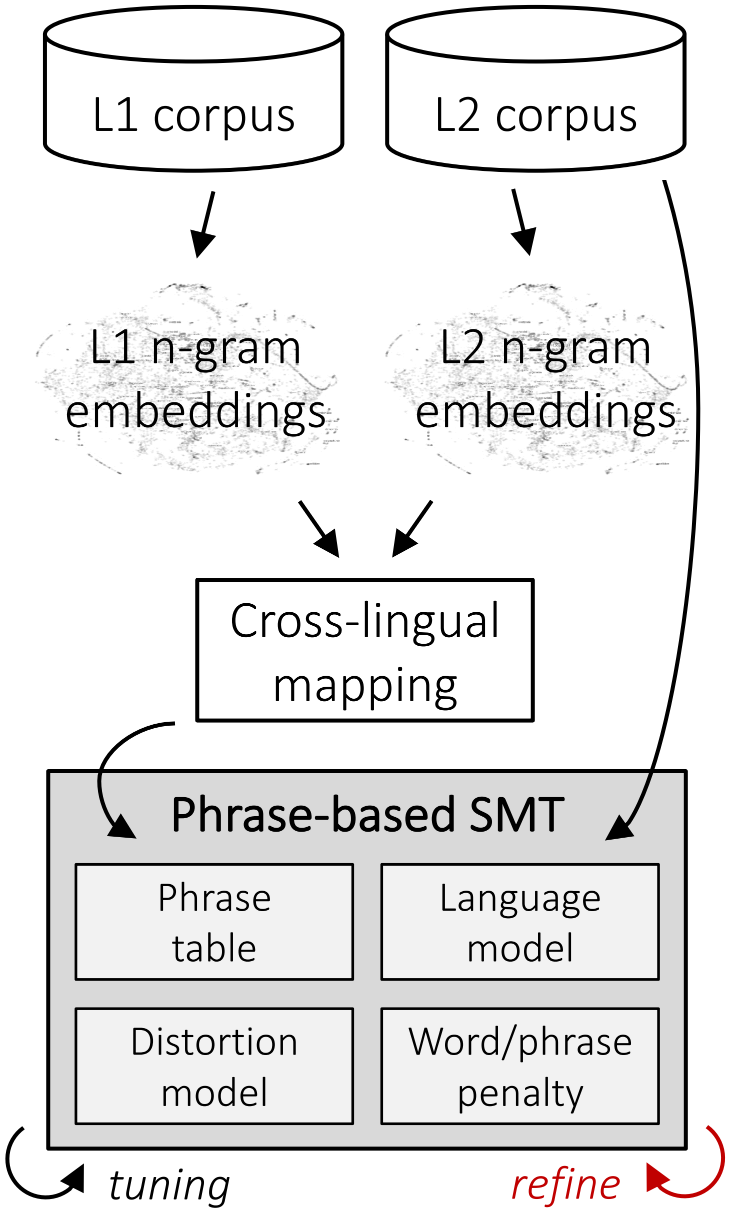




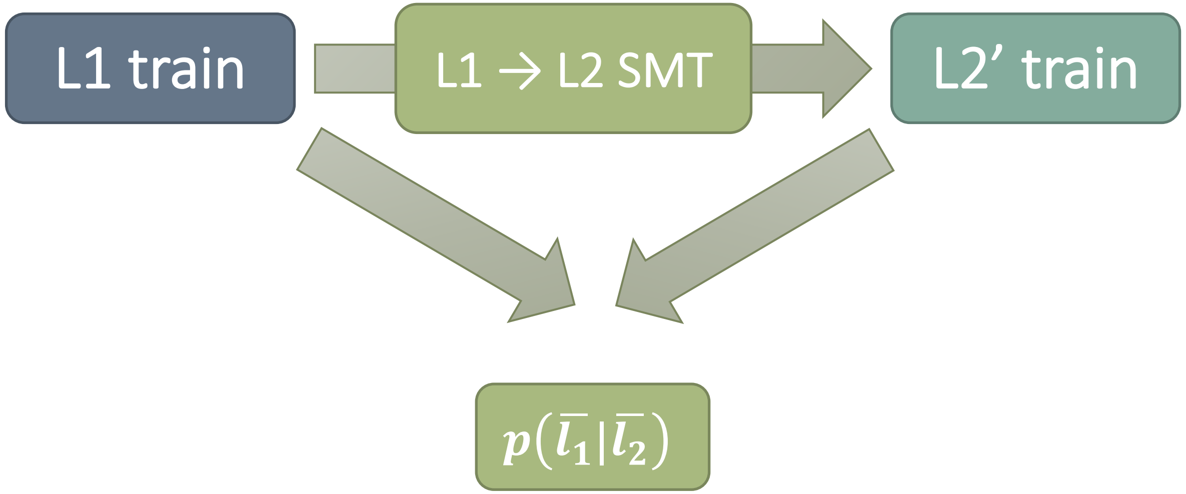
Unsupervised phrase-based SMT



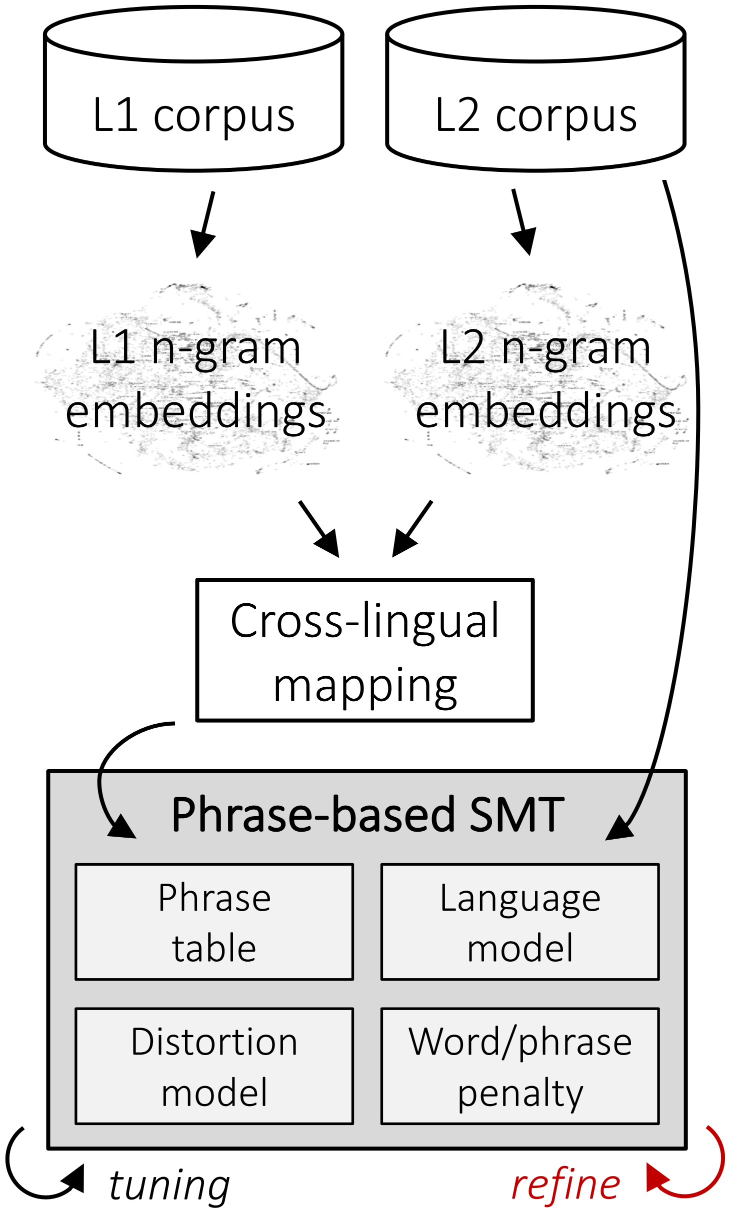


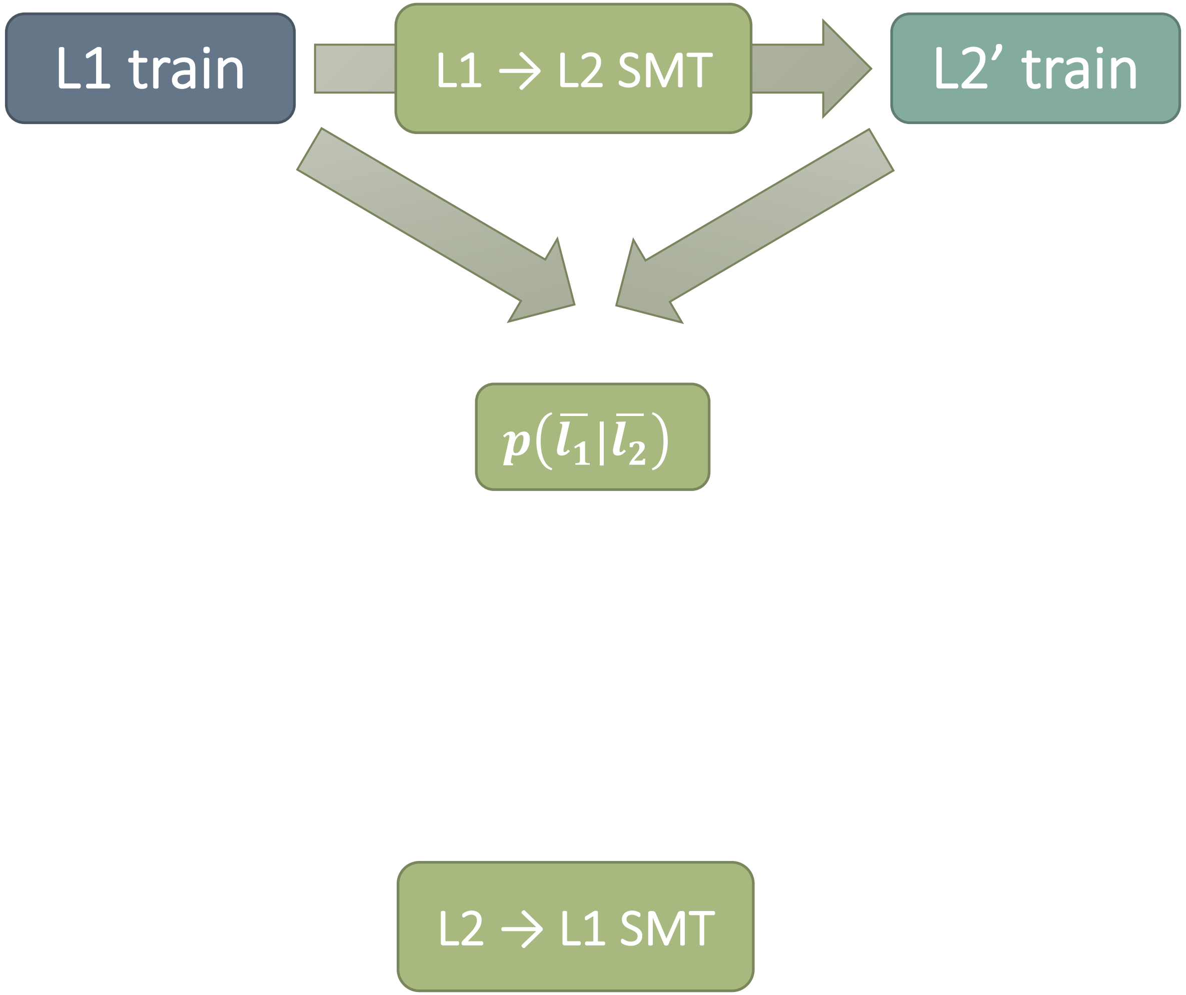


Unsupervised phrase-based SMT

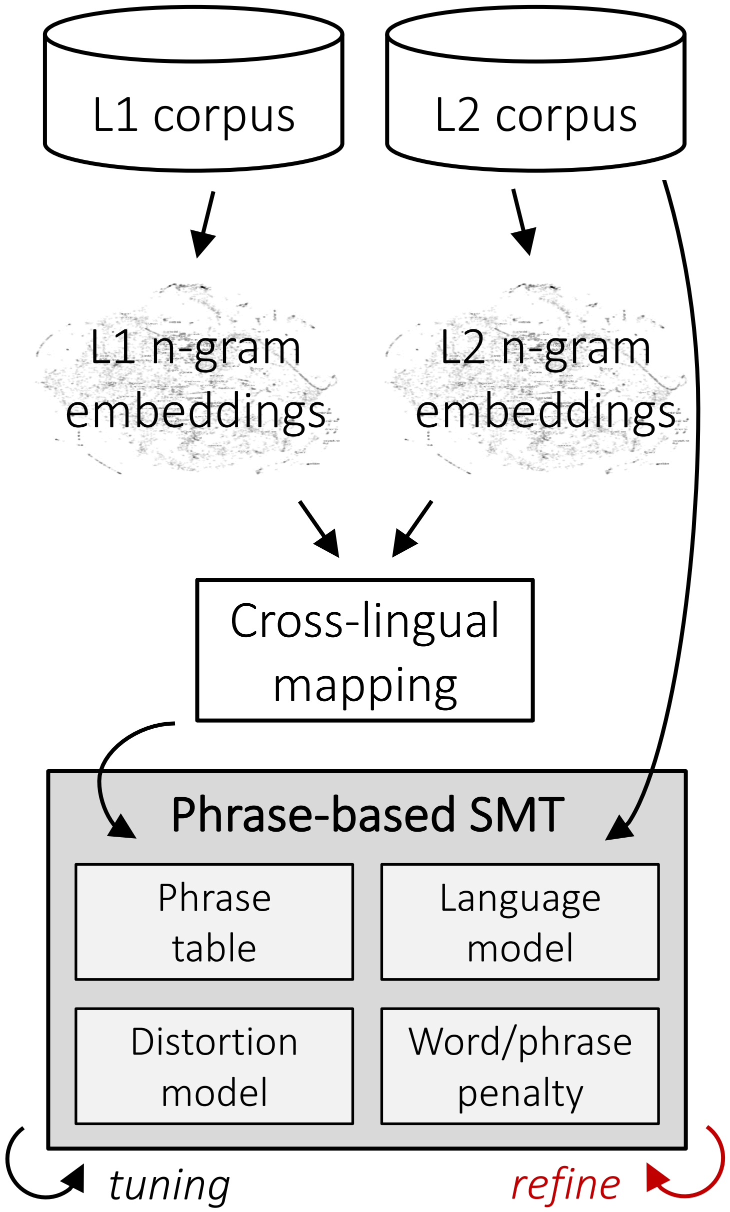


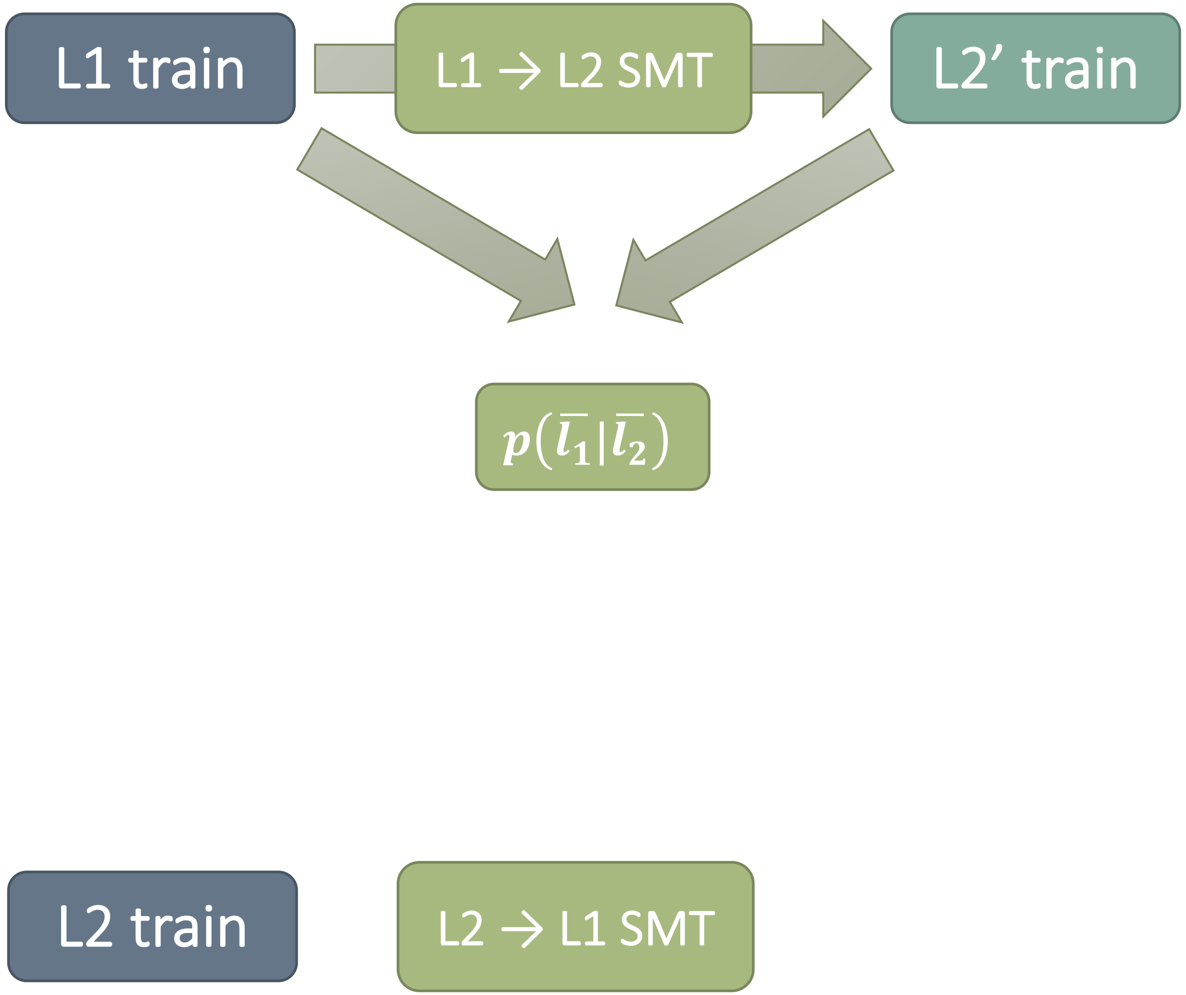
Unsupervised phrase-based SMT

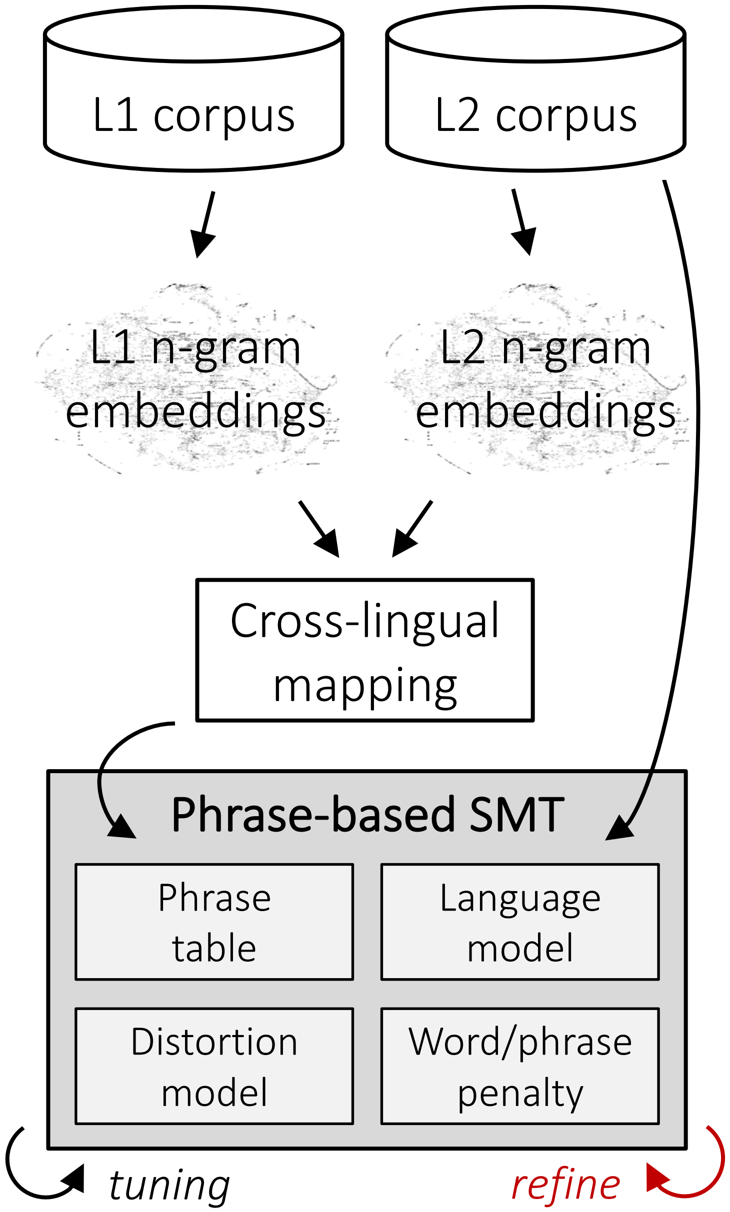




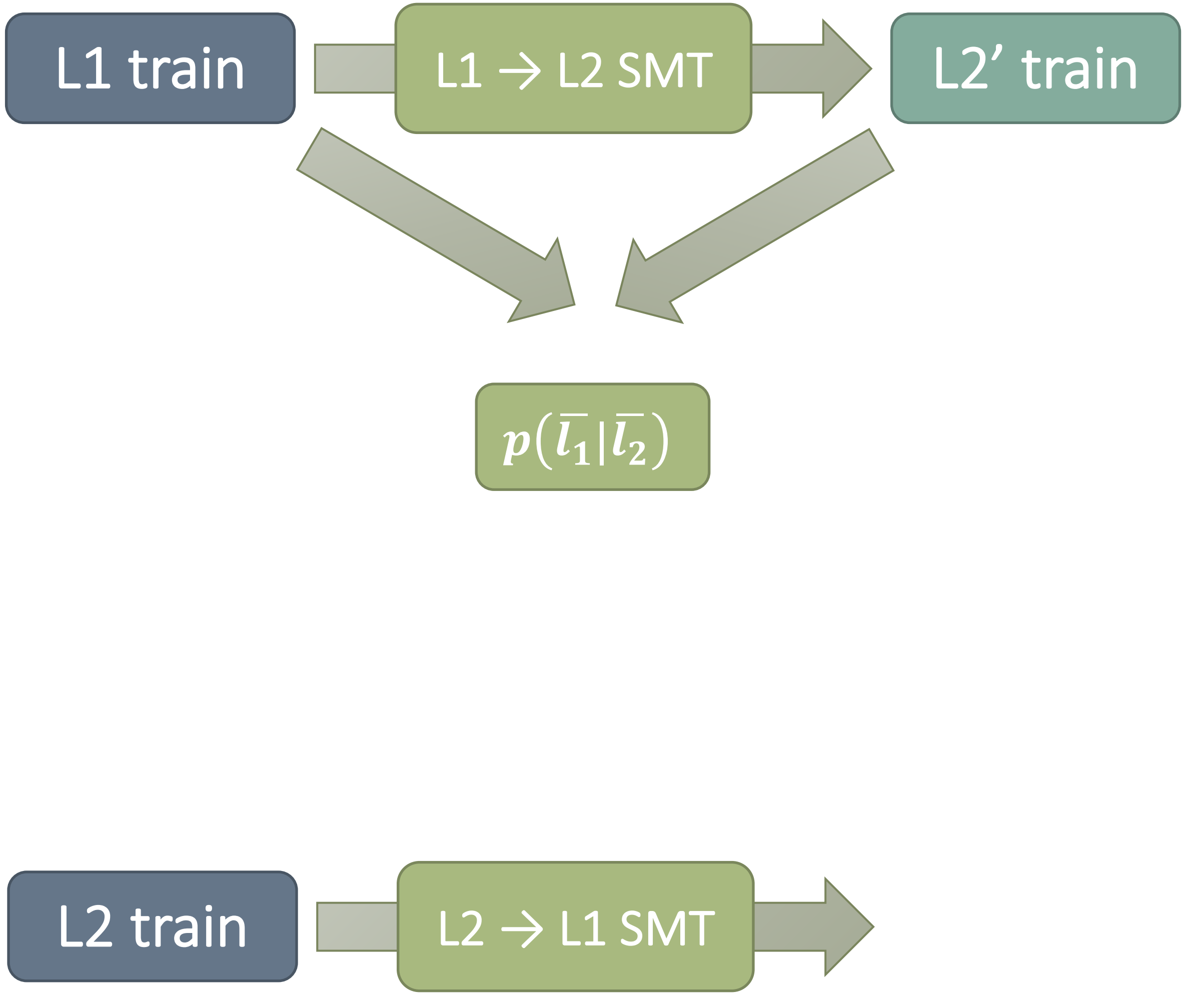
Unsupervised phrase-based SMT



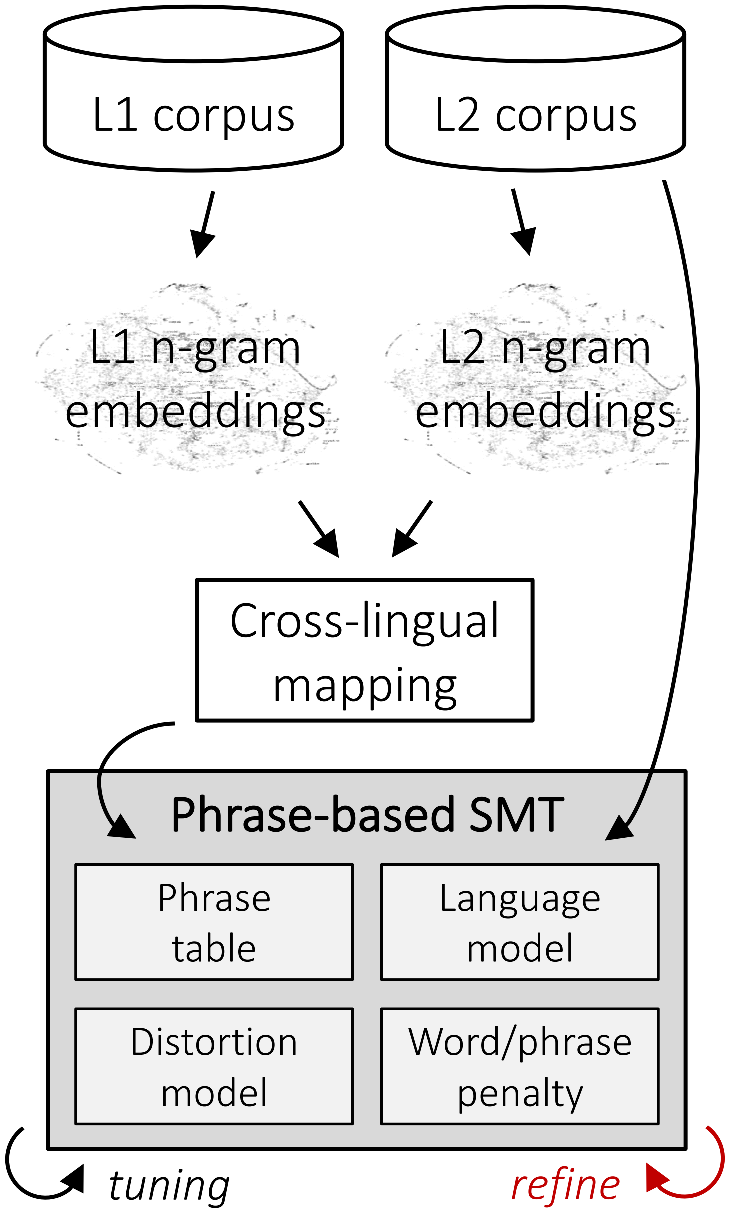


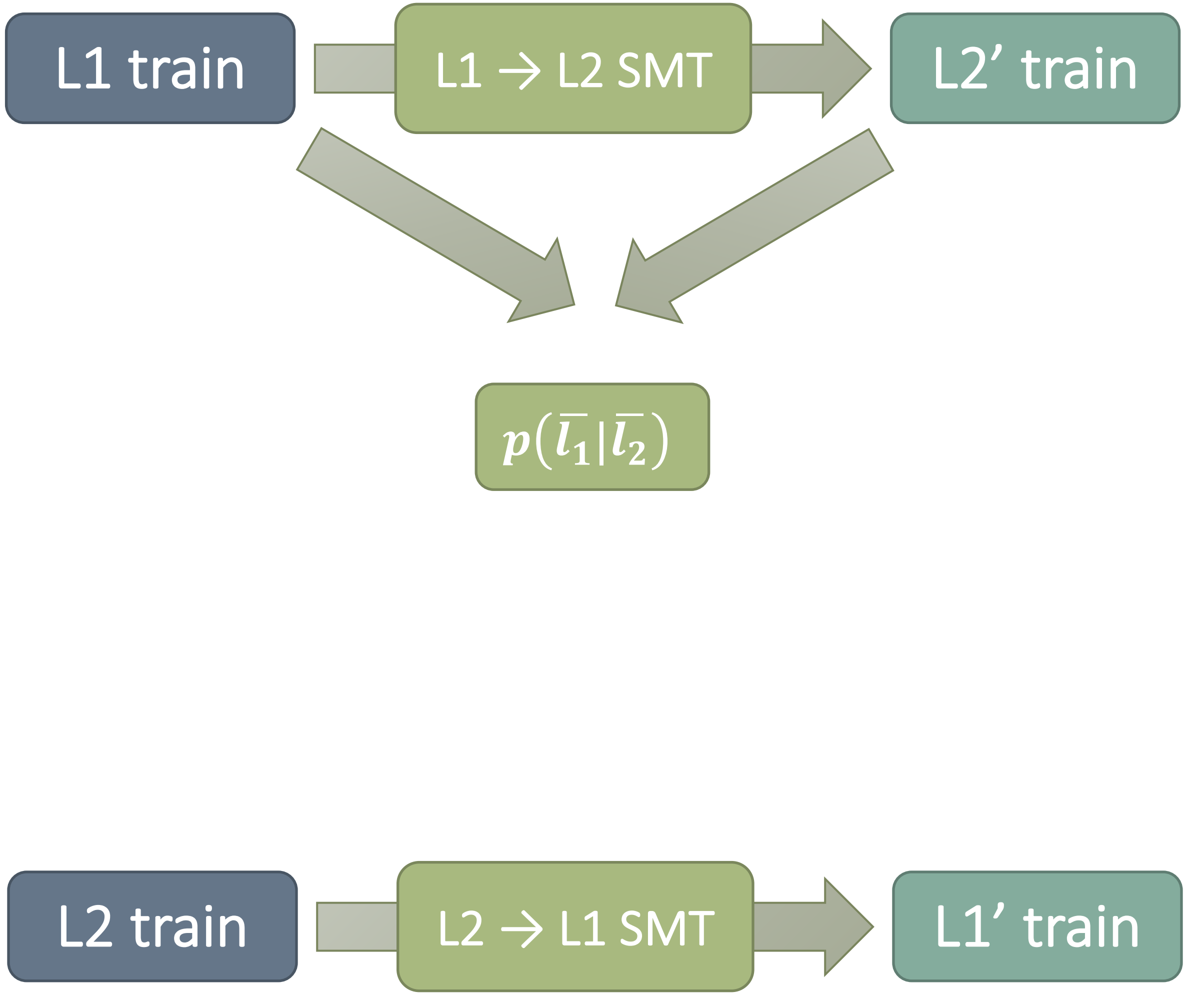


Unsupervised phrase-based SMT

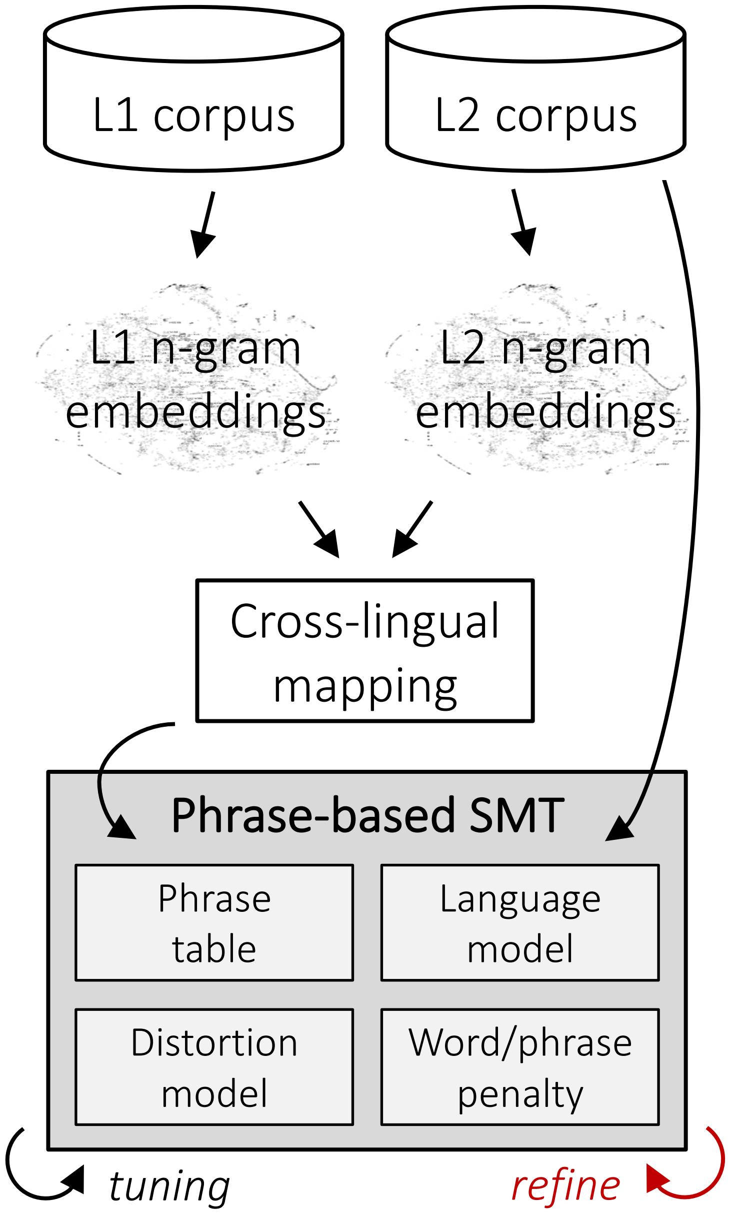


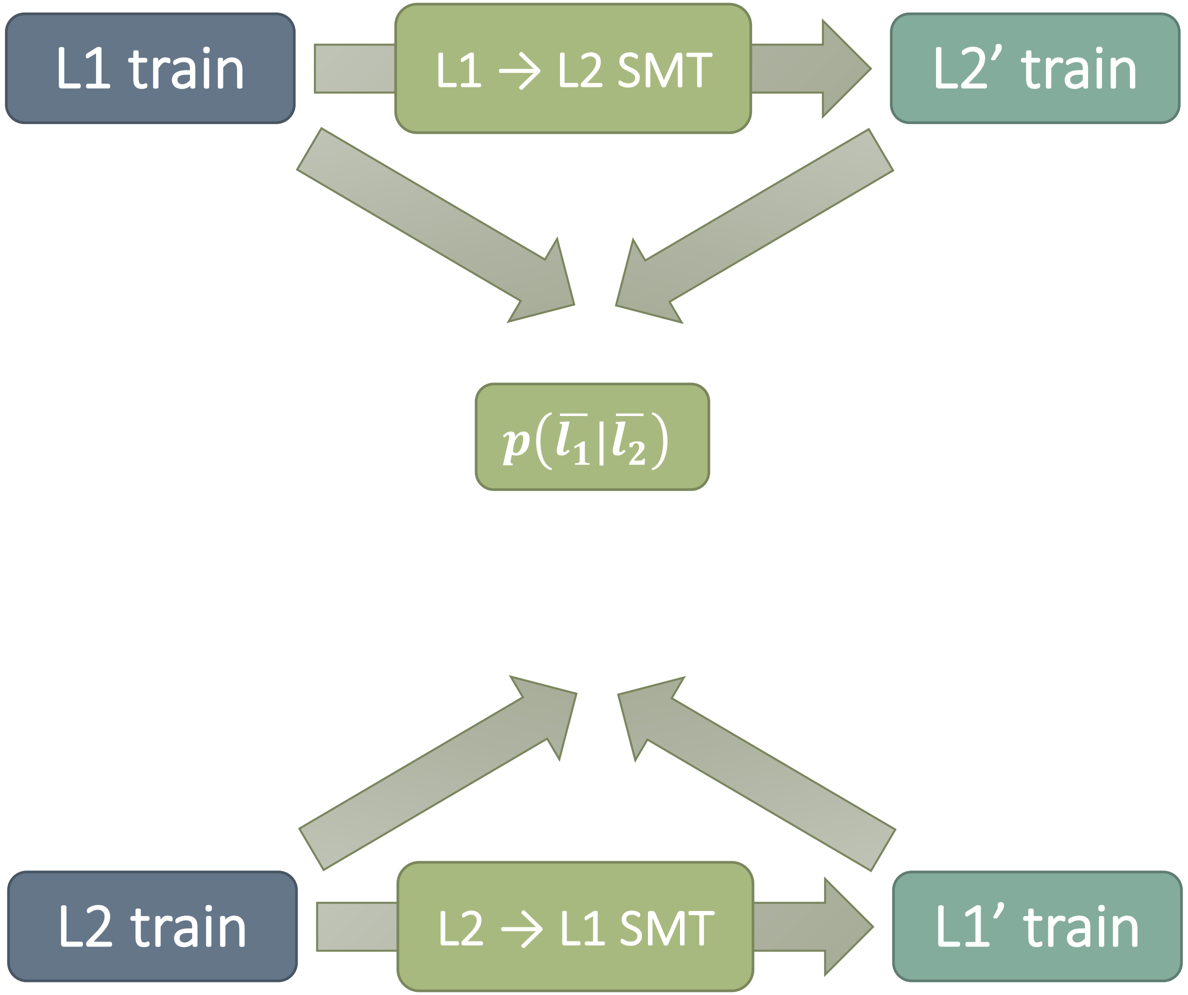
Unsupervised phrase-based SMT

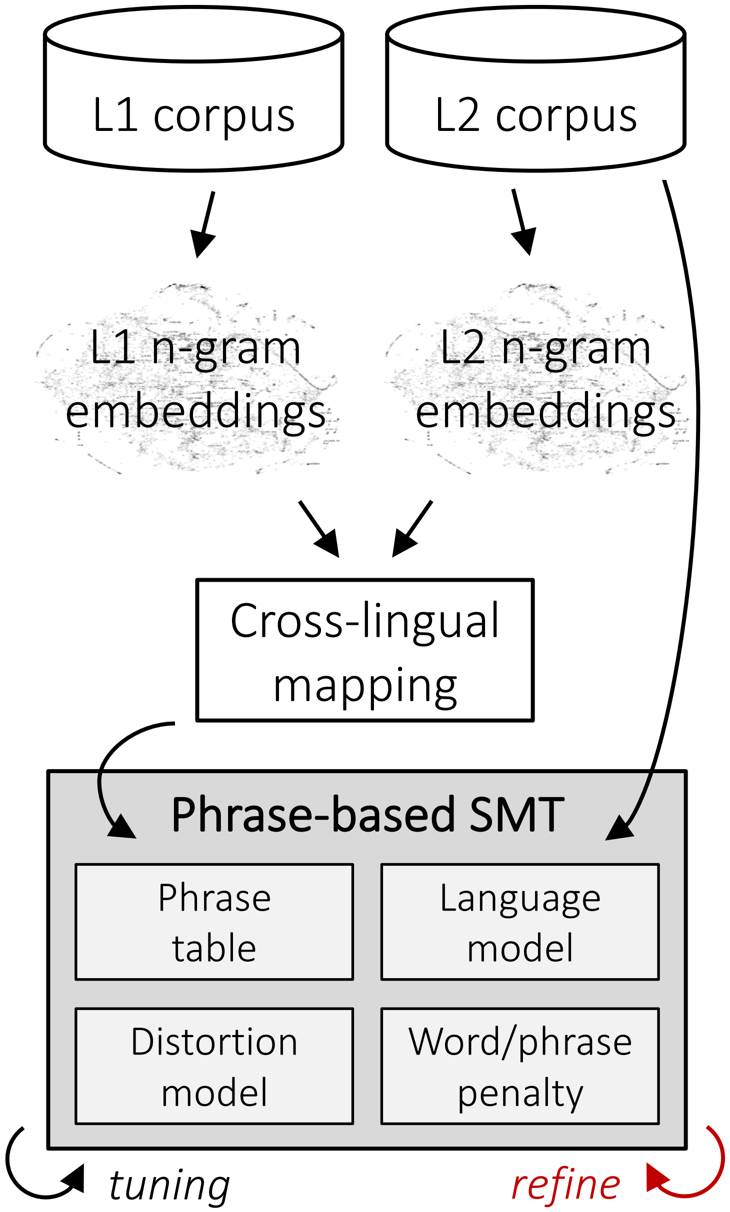




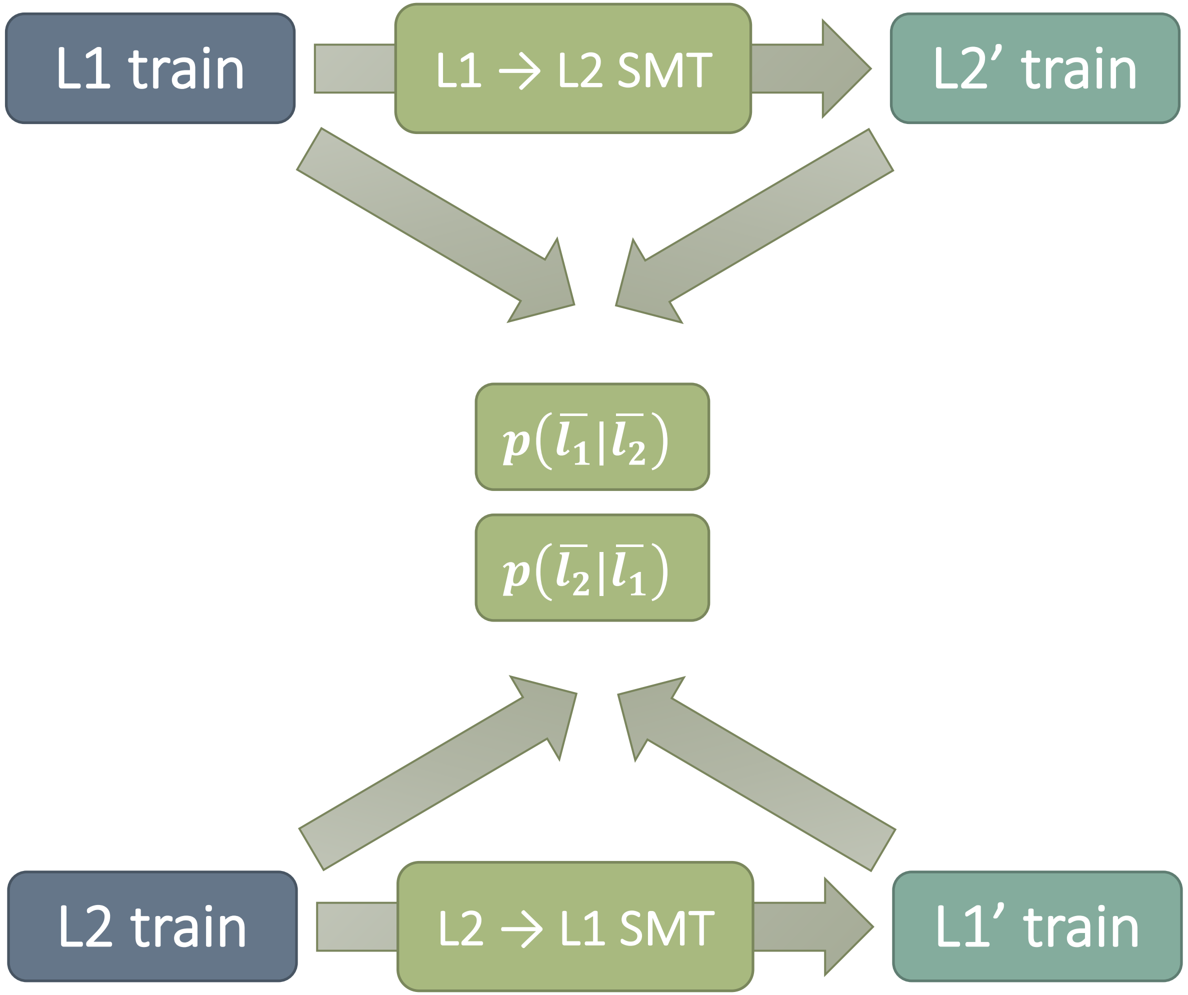
Unsupervised phrase-based SMT



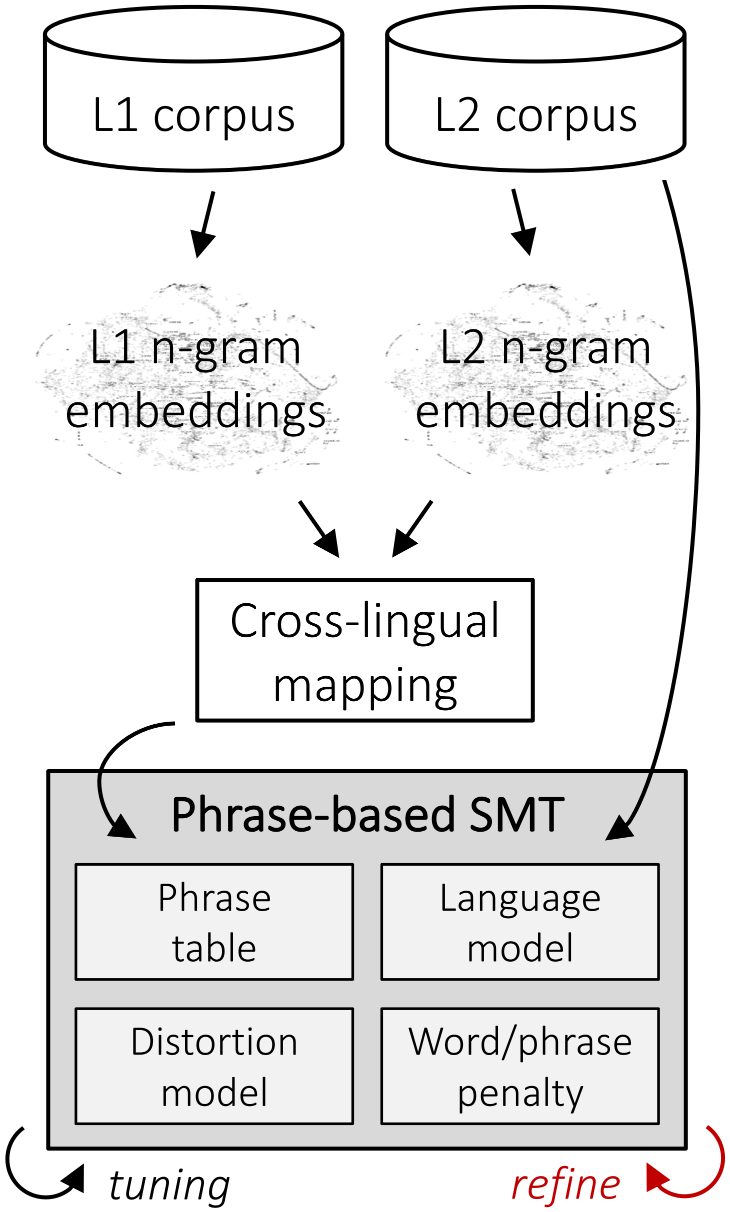


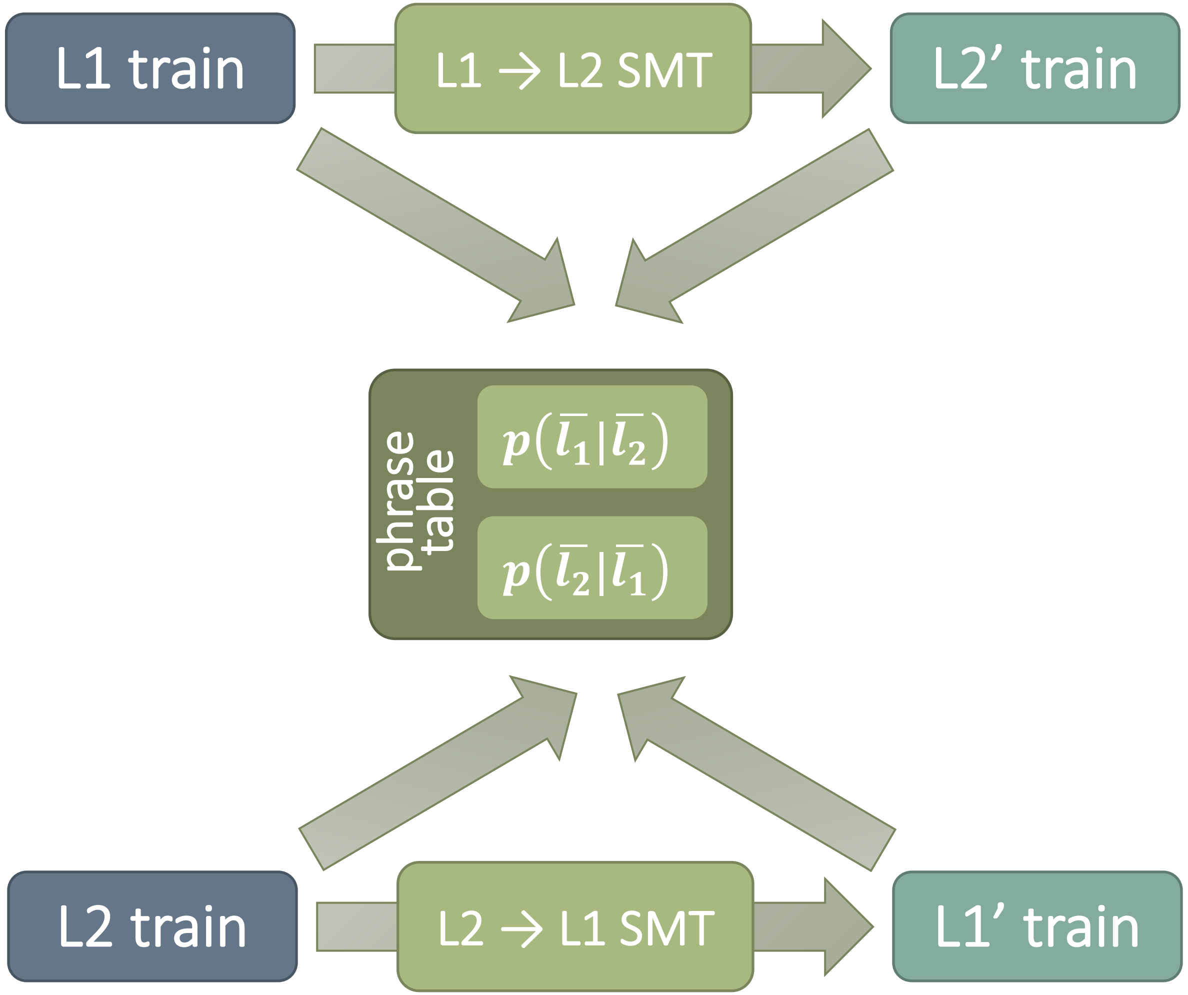


Unsupervised phrase-based SMT

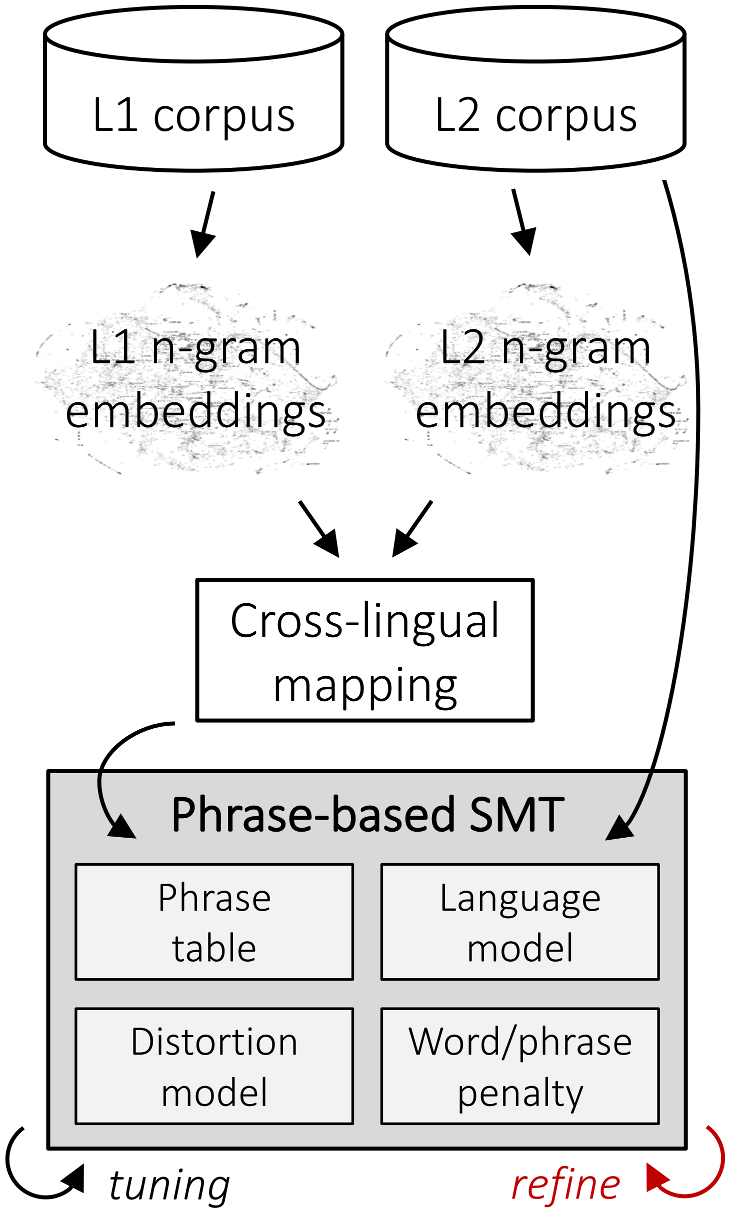


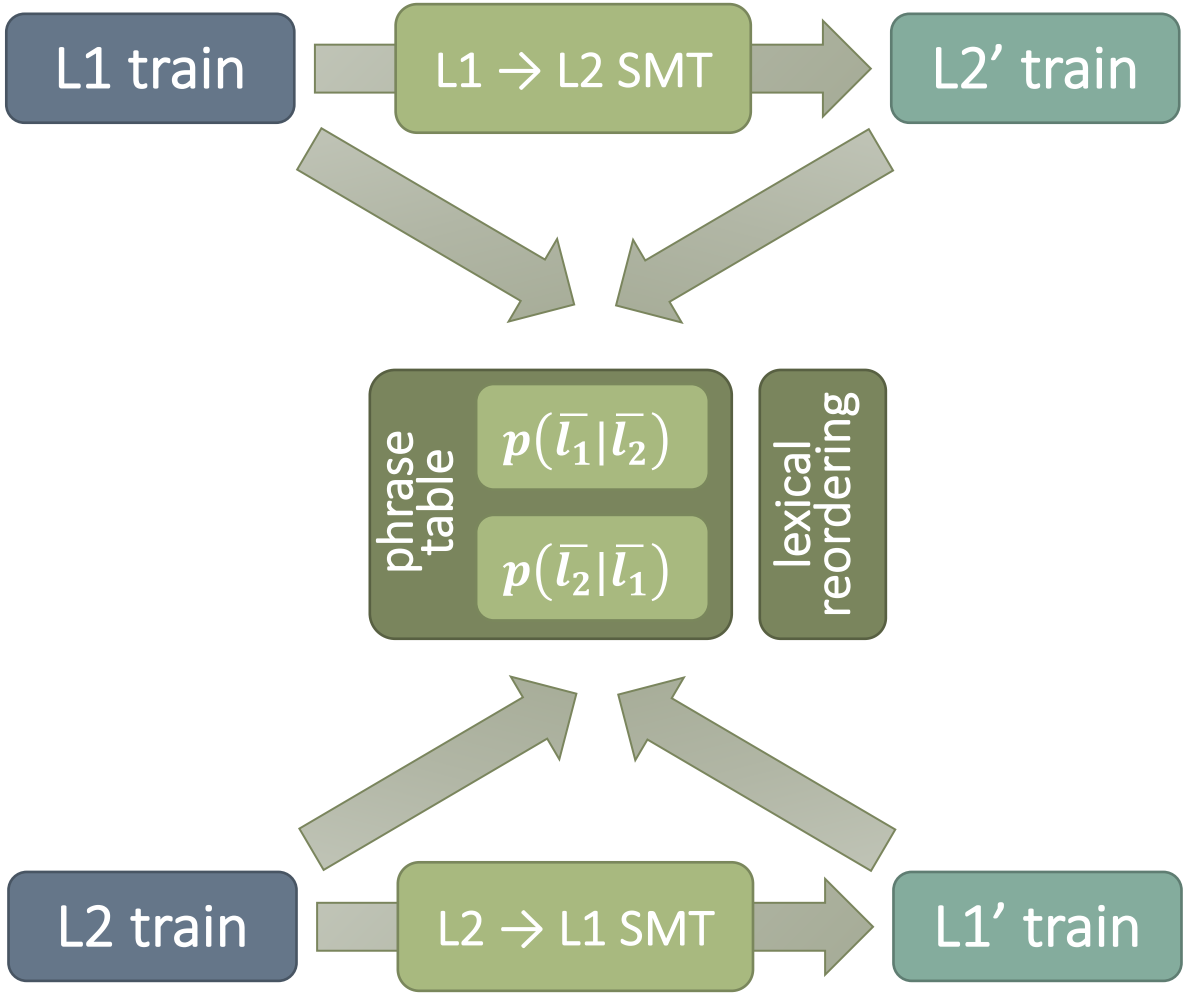
Unsupervised phrase-based SMT



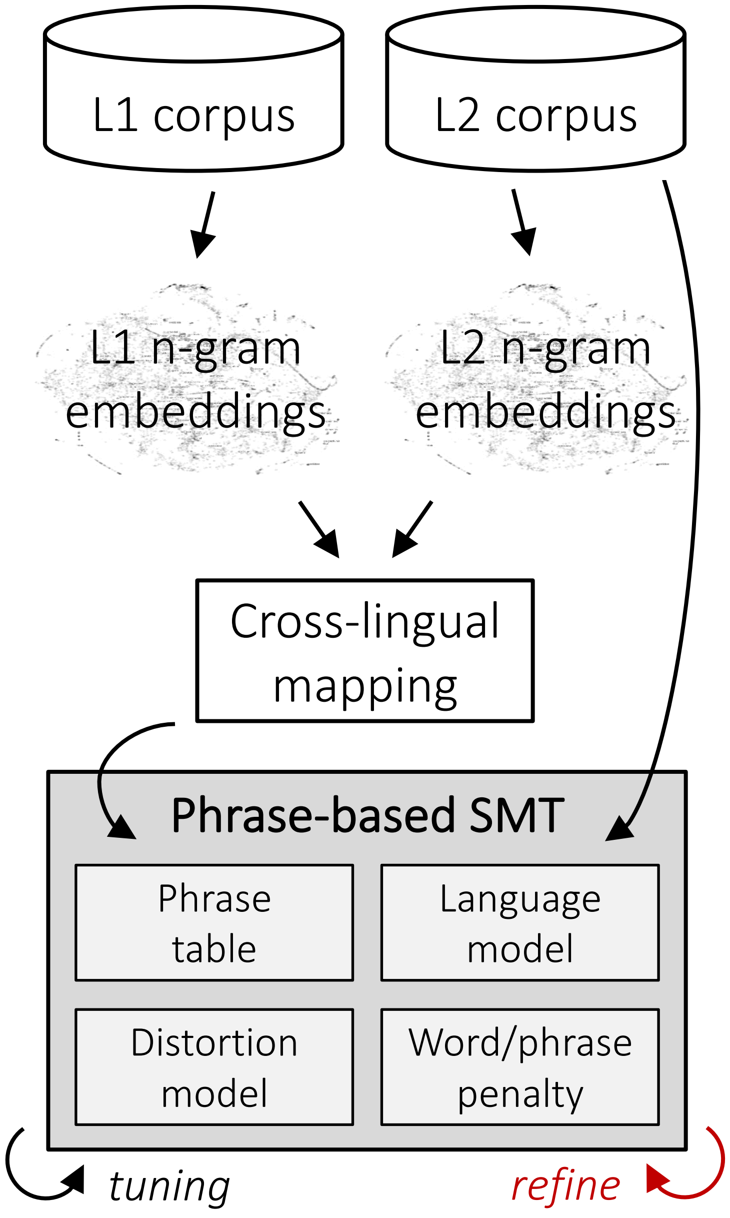


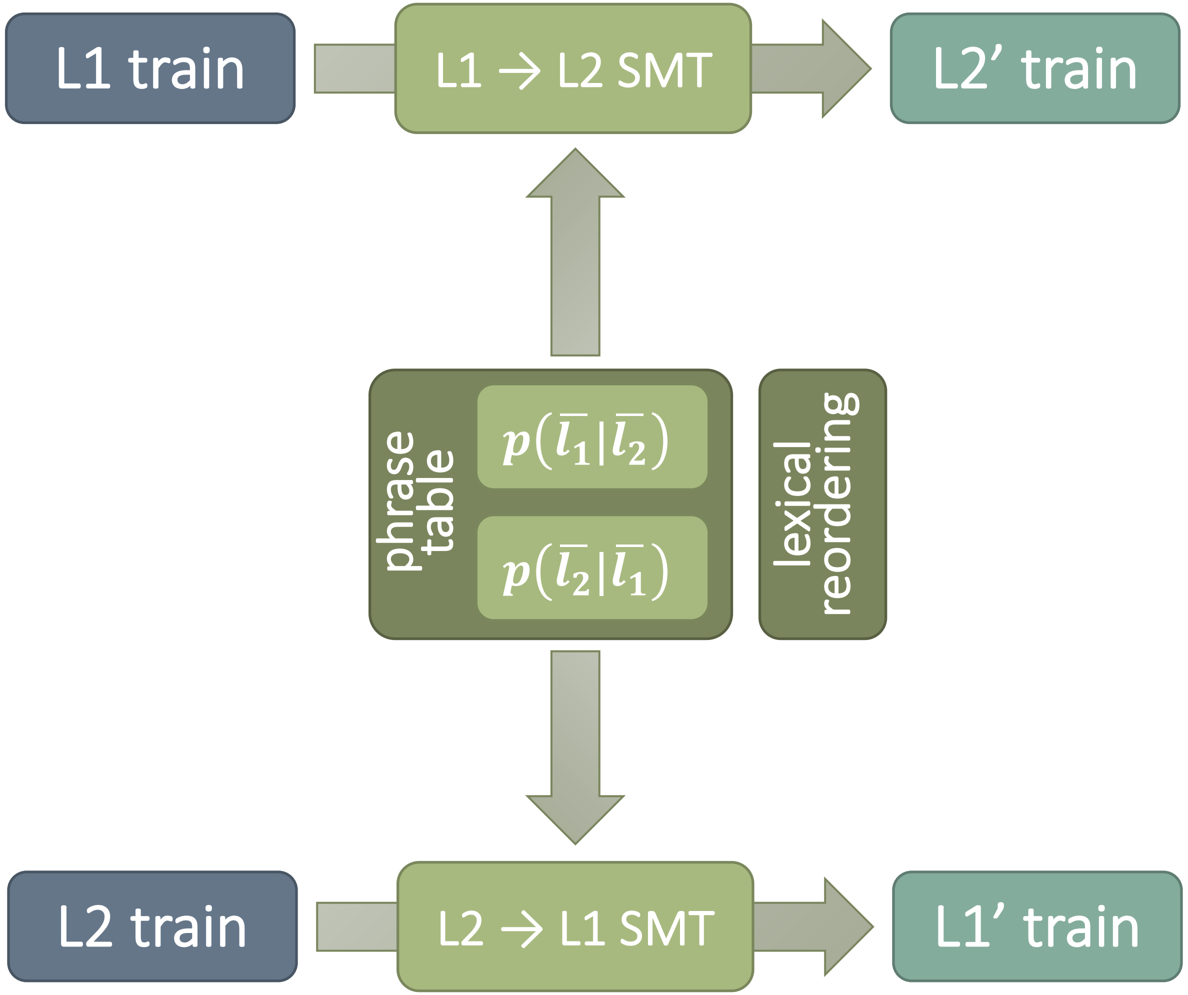
Unsupervised phrase-based SMT



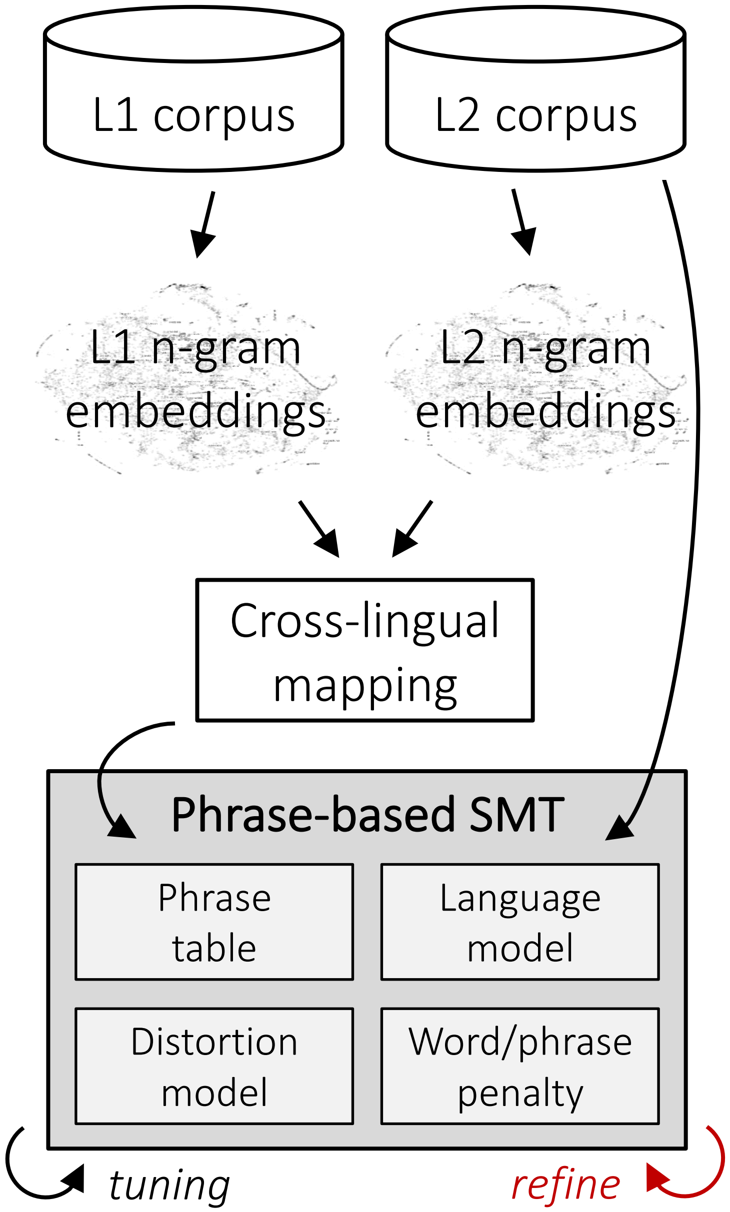


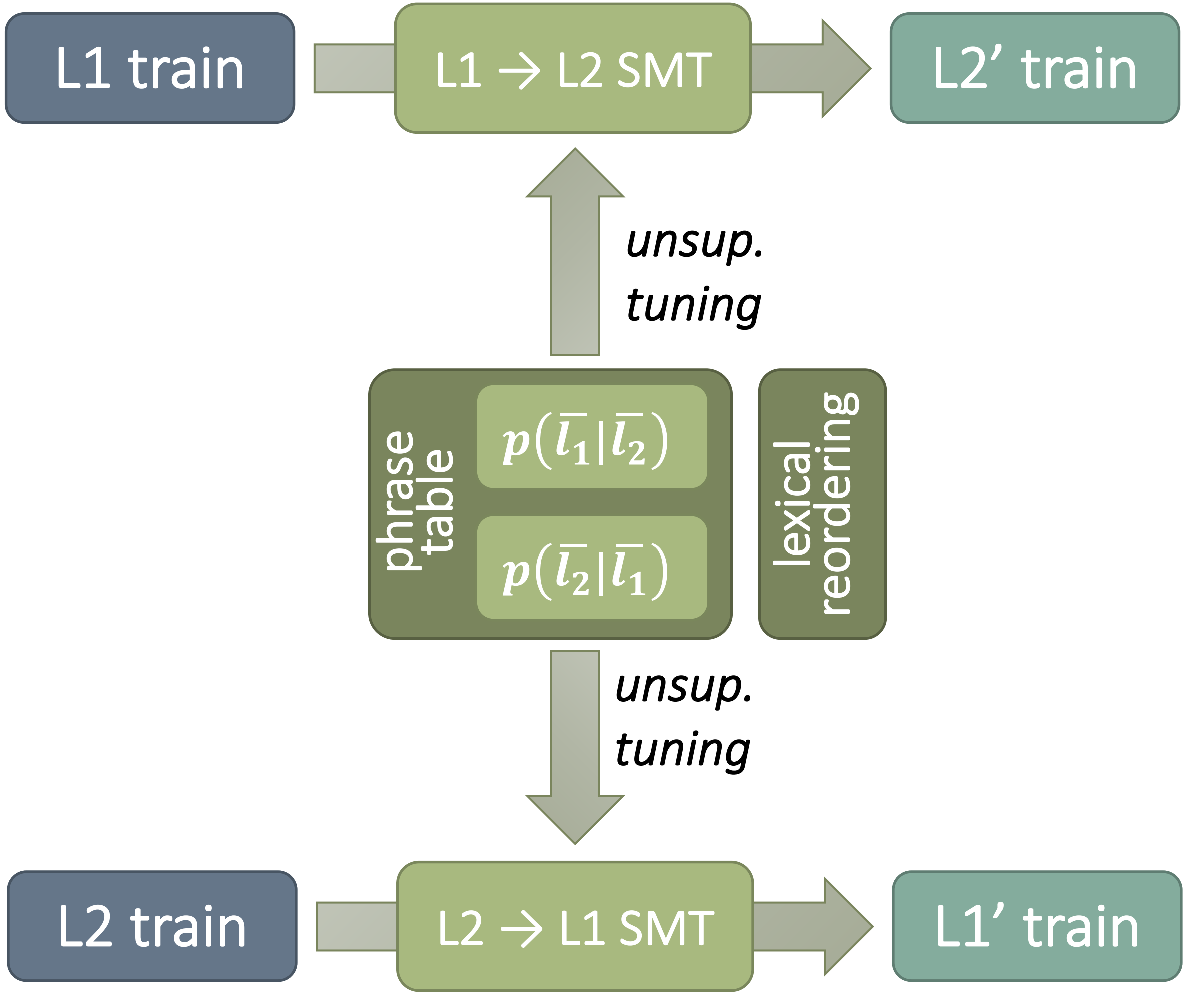
Unsupervised phrase-based SMT

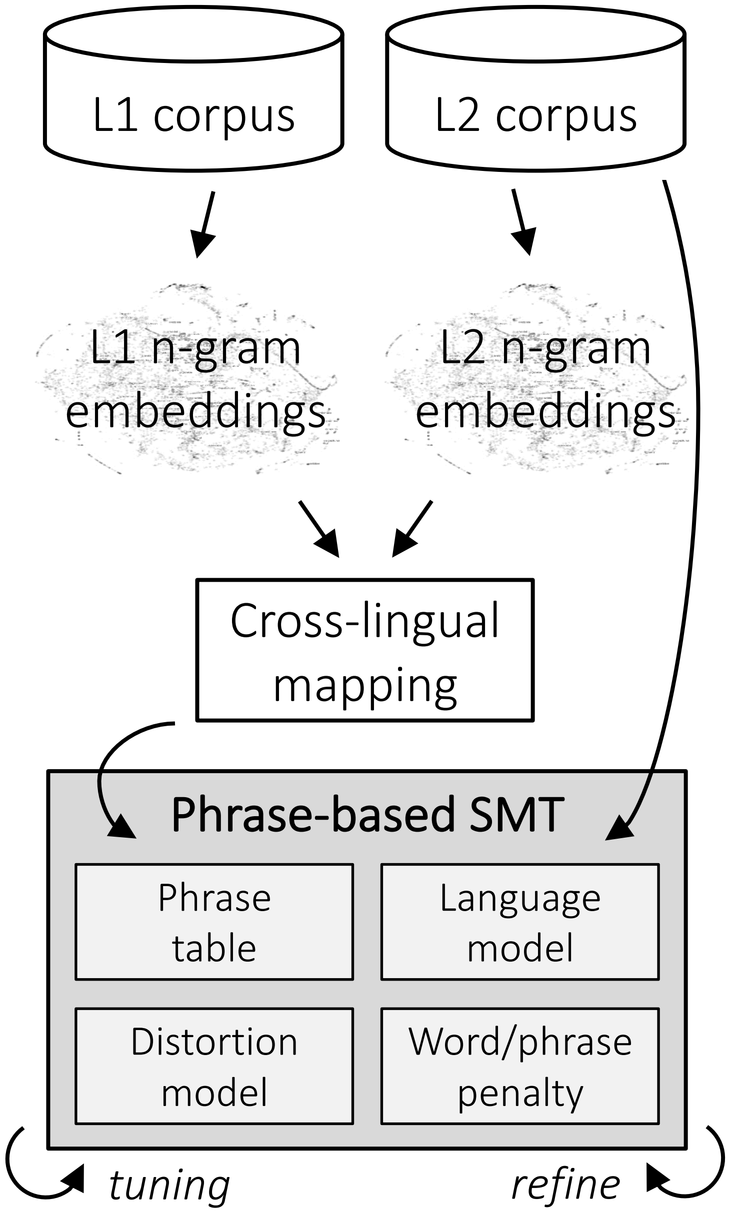


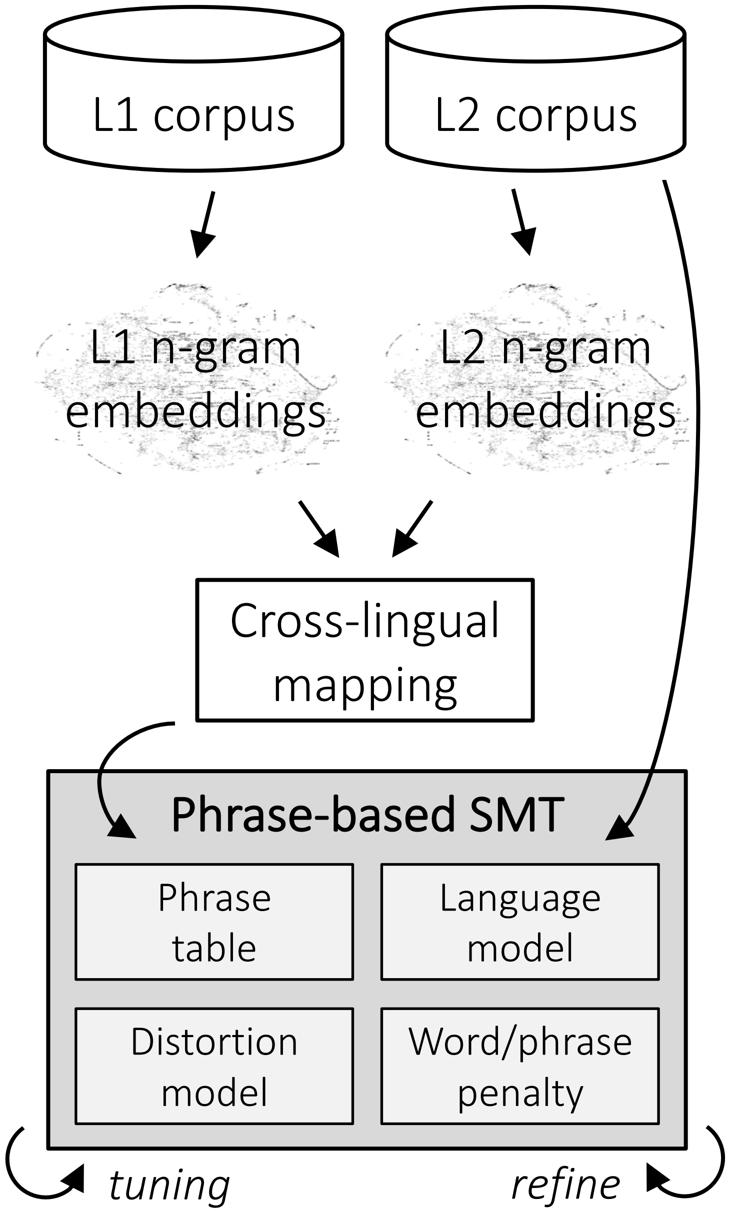


Unsupervised phrase-based SMT

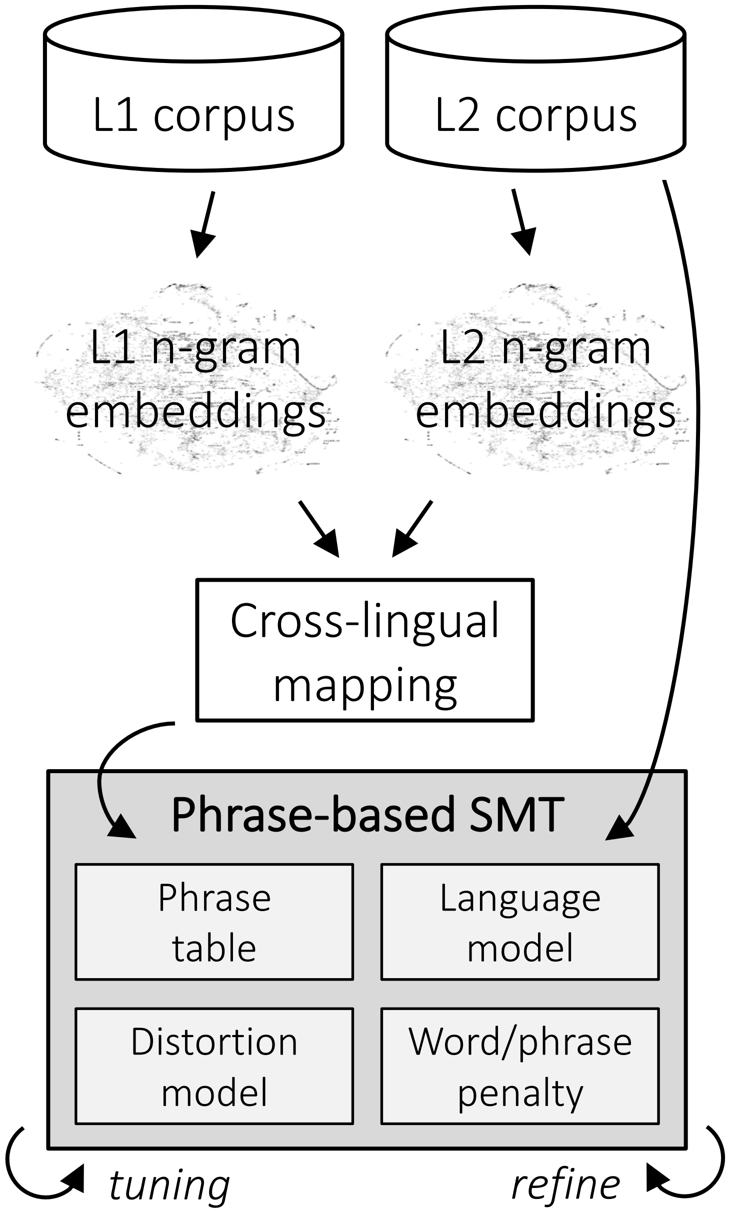





L1 corpus
L2 corpus
L1 n-gram embeddings
L2 n-gram embeddings
Cross-lingual mapping
Phrase-based SMT
Phrase table
Language model
Distortion model
Word/phrase penalty
tuning
refine

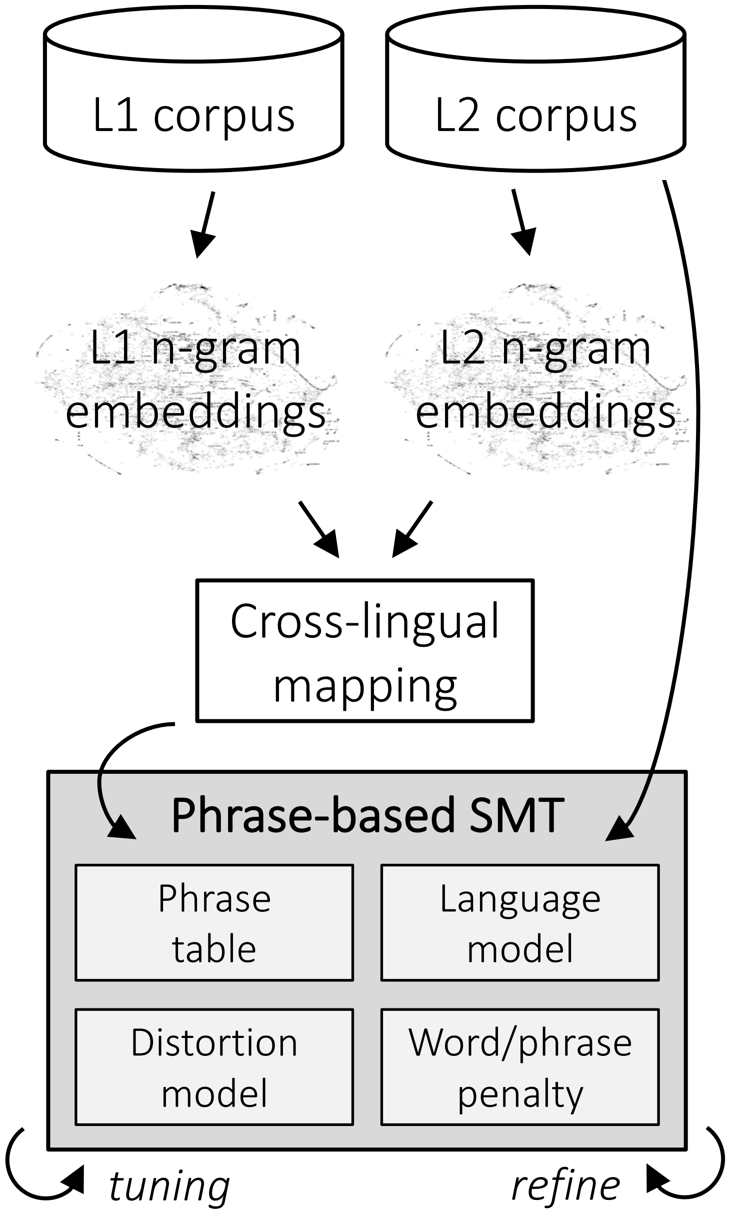

L1 corpus
L2 corpus
L1 n-gram embeddings
L2 n-gram embeddings
Cross-lingual mapping
Phrase-based SMT
Phrase table
Language model
Distortion model
Word/phrase penalty
tuning
refine

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…

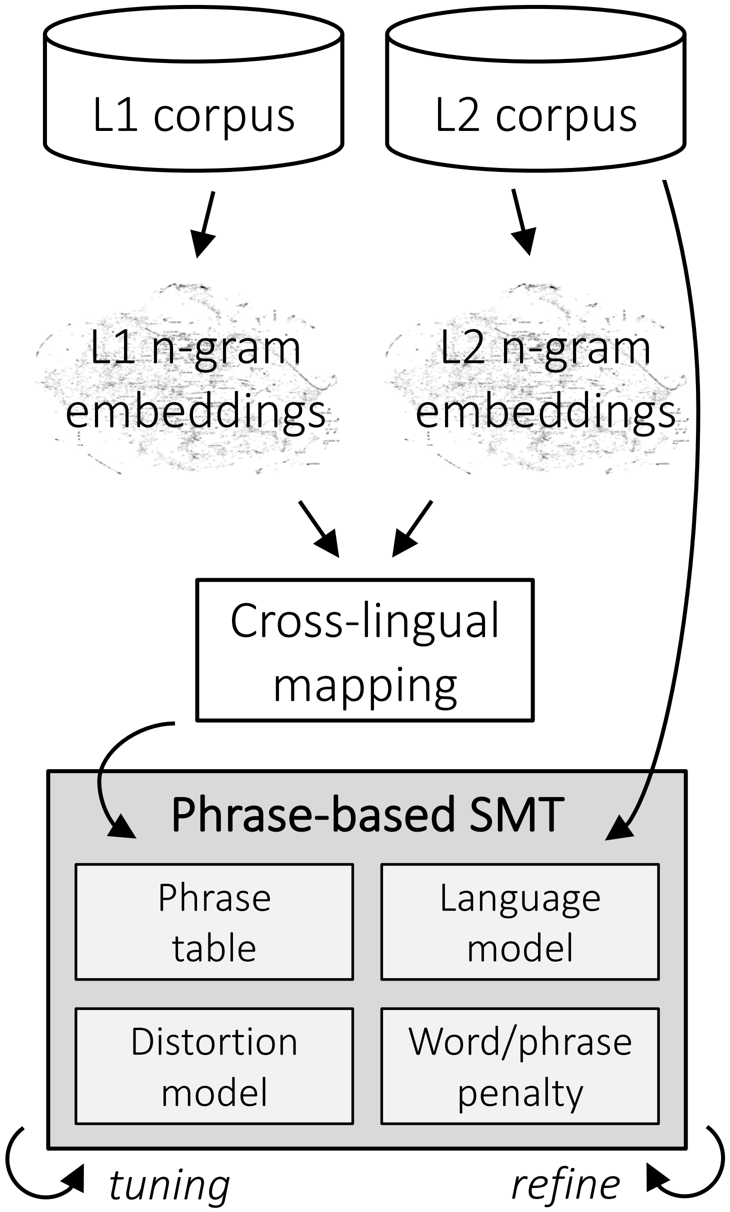
NMT >> SMT

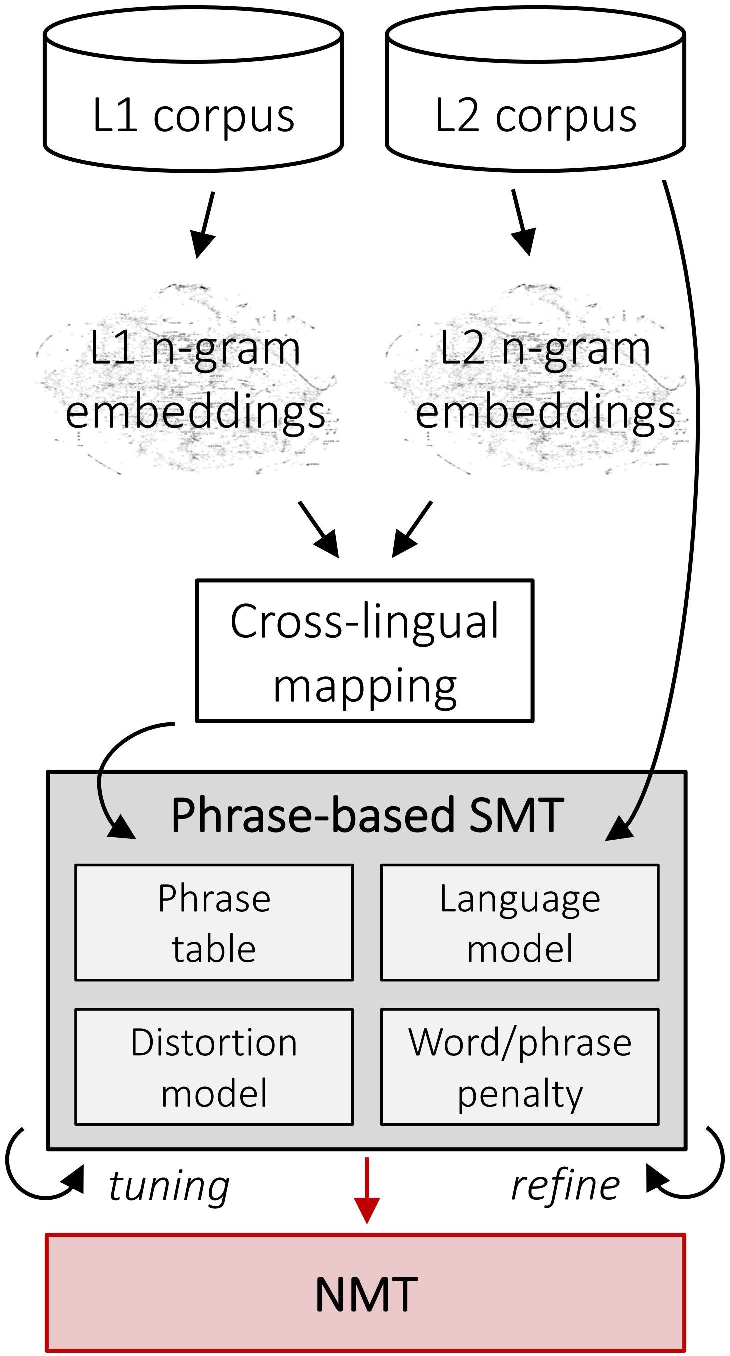


but…

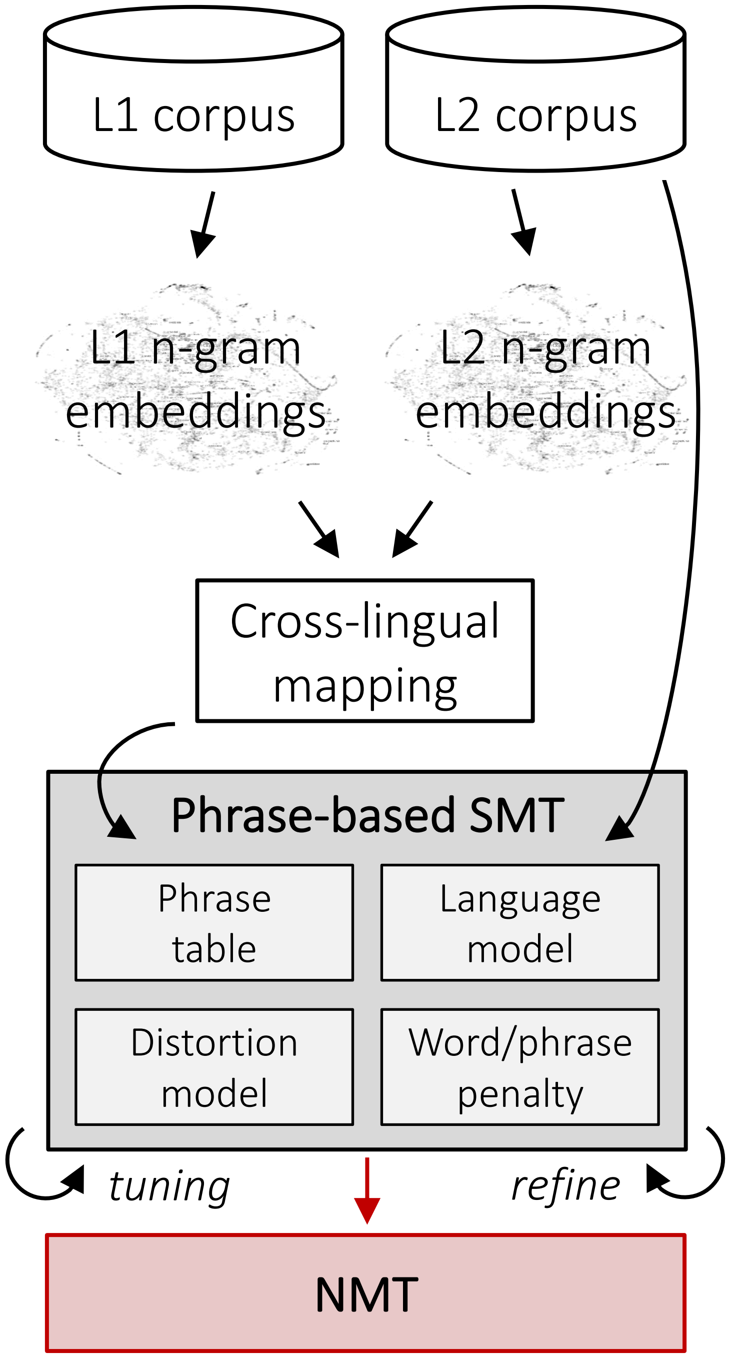
NMT >> SMT

(unsupervised) SMT has a hard ceiling!

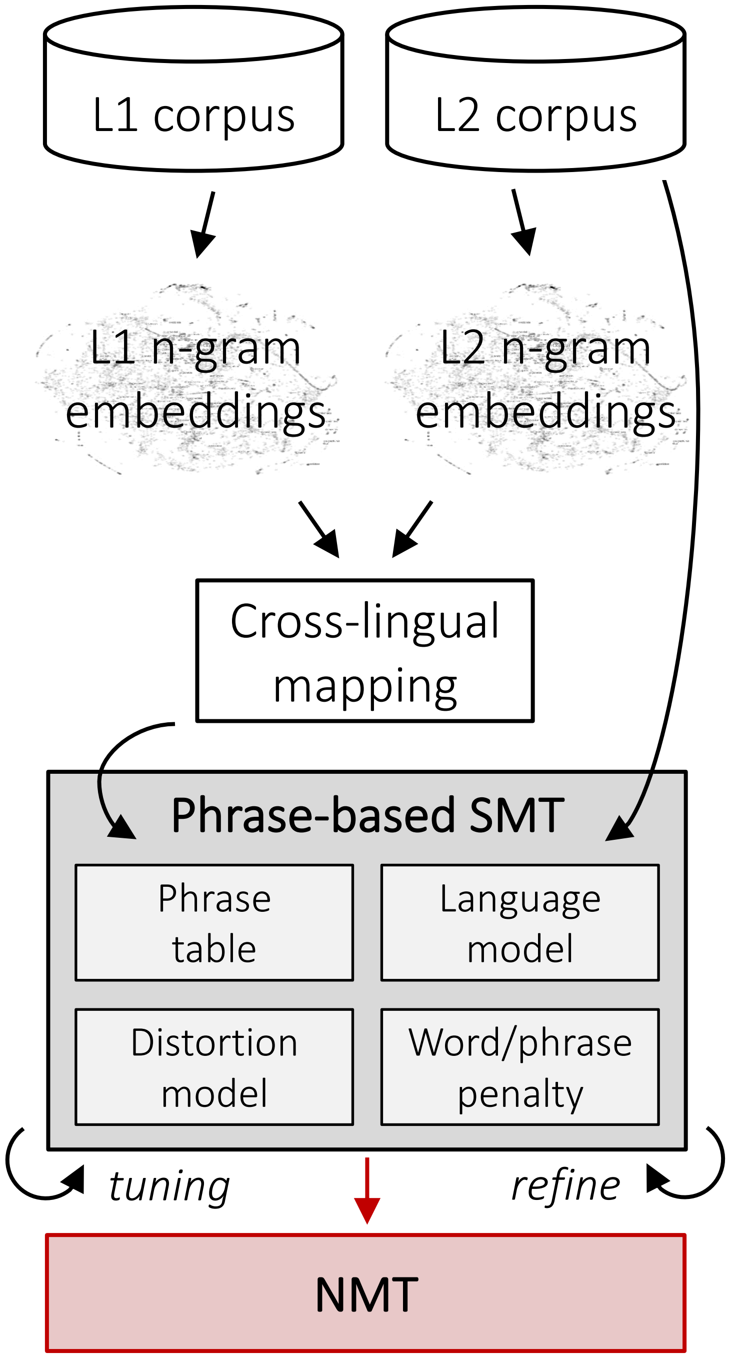

L1 corpus
L2 corpus
L1 n-gram embeddings
L2 n-gram embeddings
Cross-lingual mapping
Phrase-based SMT
Phrase table
Language model
Distortion model
Word/phrase penalty
tuning
refine

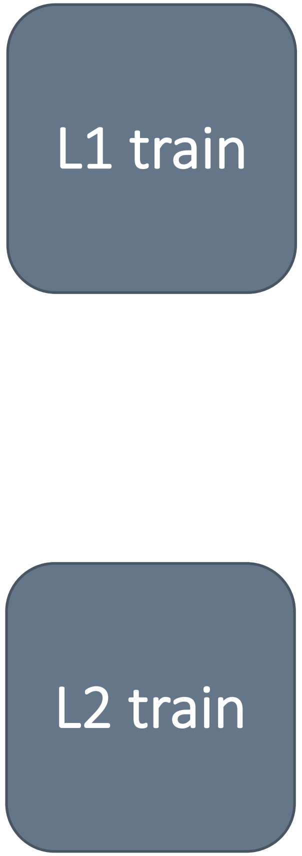

L1 corpus
L2 corpus
L1 n-gram embeddings
L2 n-gram embeddings
Cross-lingual mapping
Phrase-based SMT
Phrase table
Language model
Distortion model
Word/phrase penalty
tuning
refine
NMT

NMT hybridization

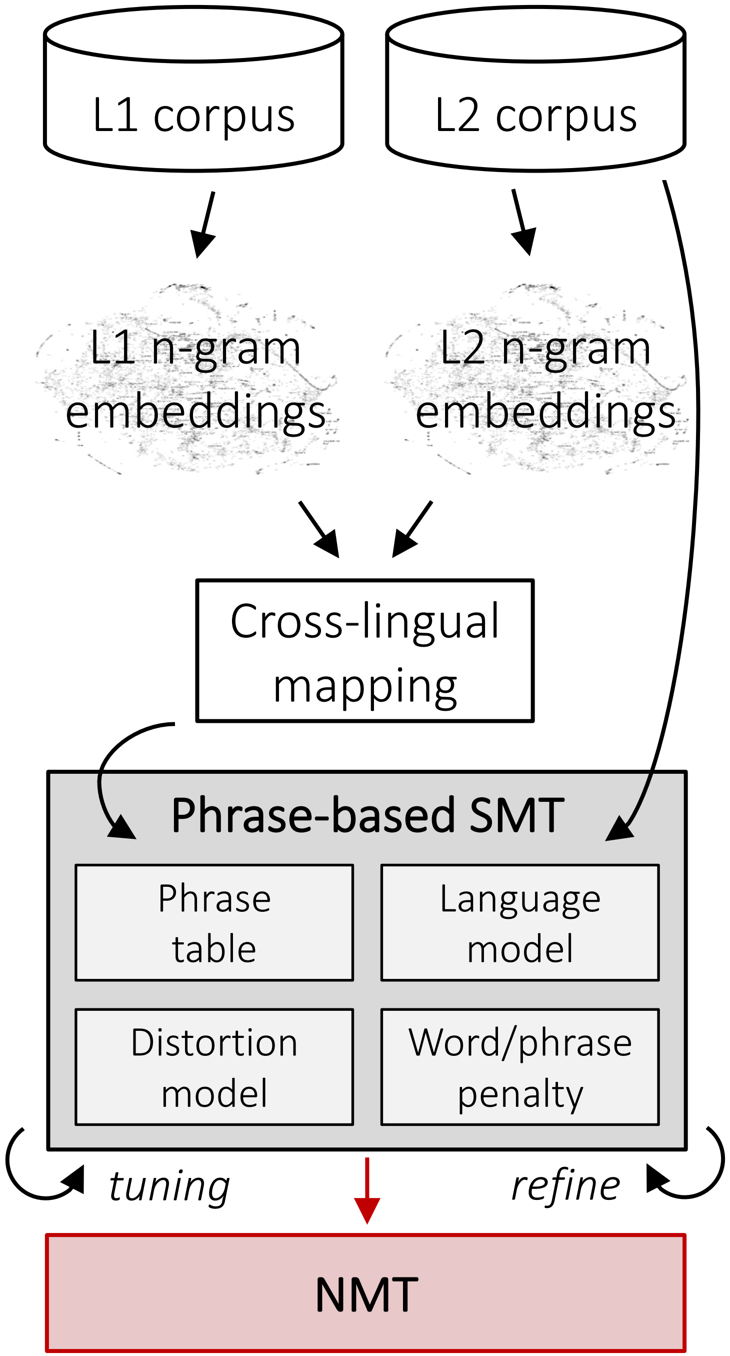


NMT hybridization

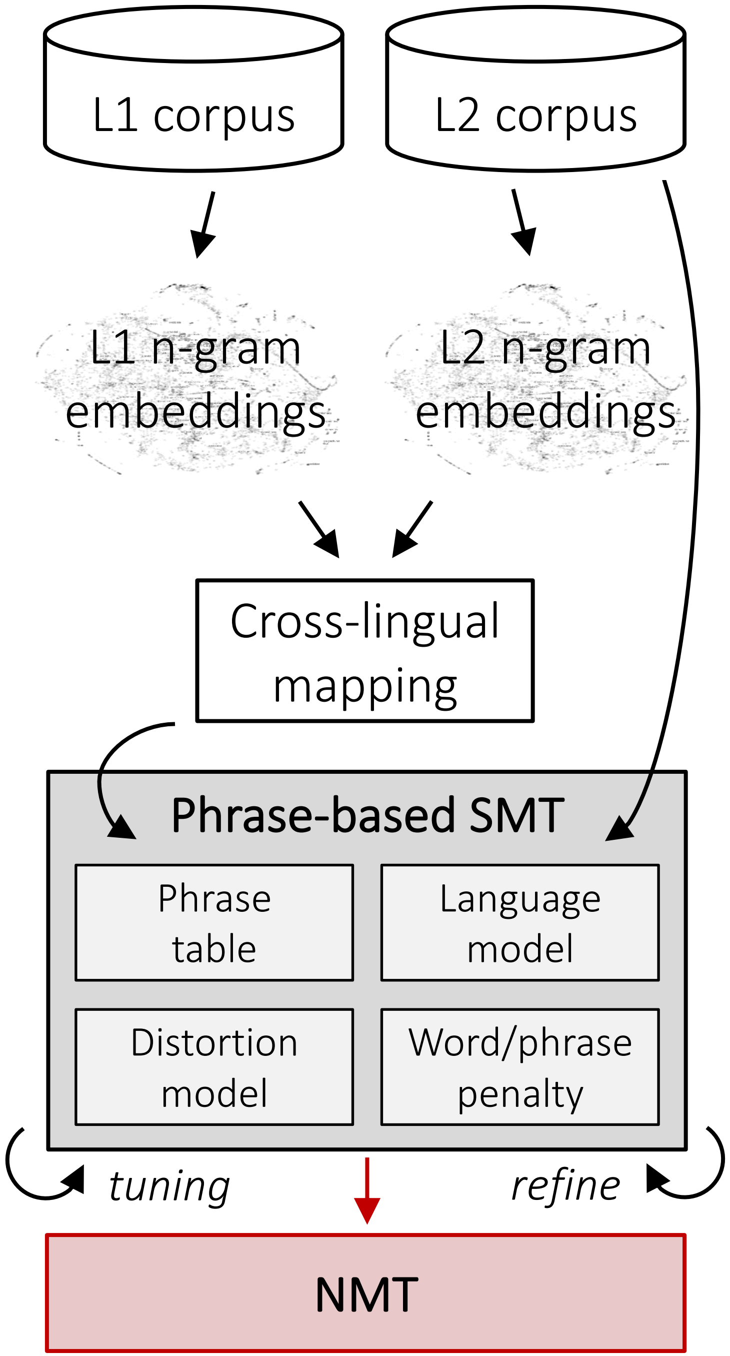


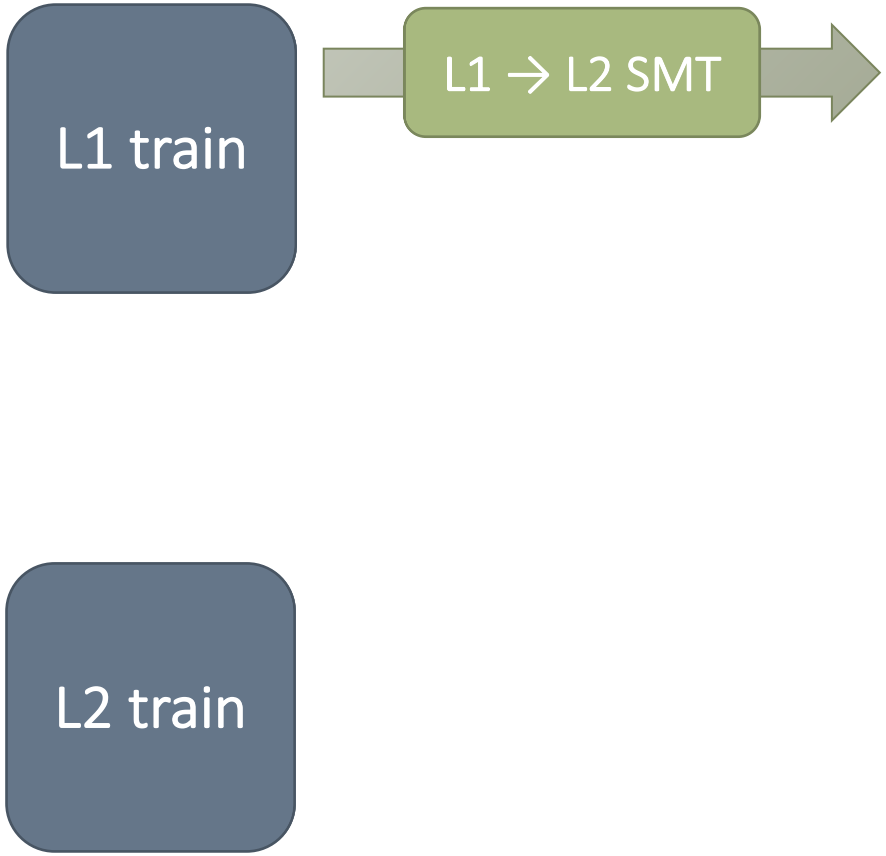


NMT hybridization

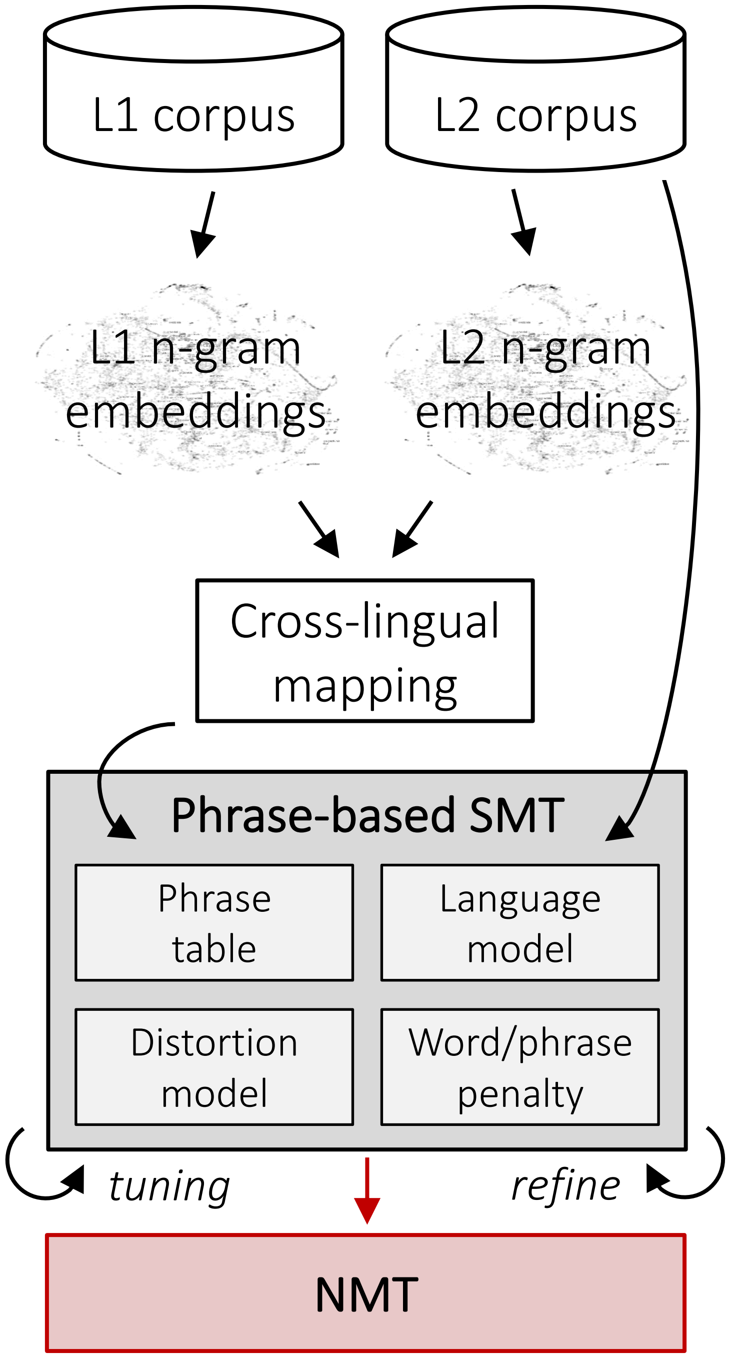


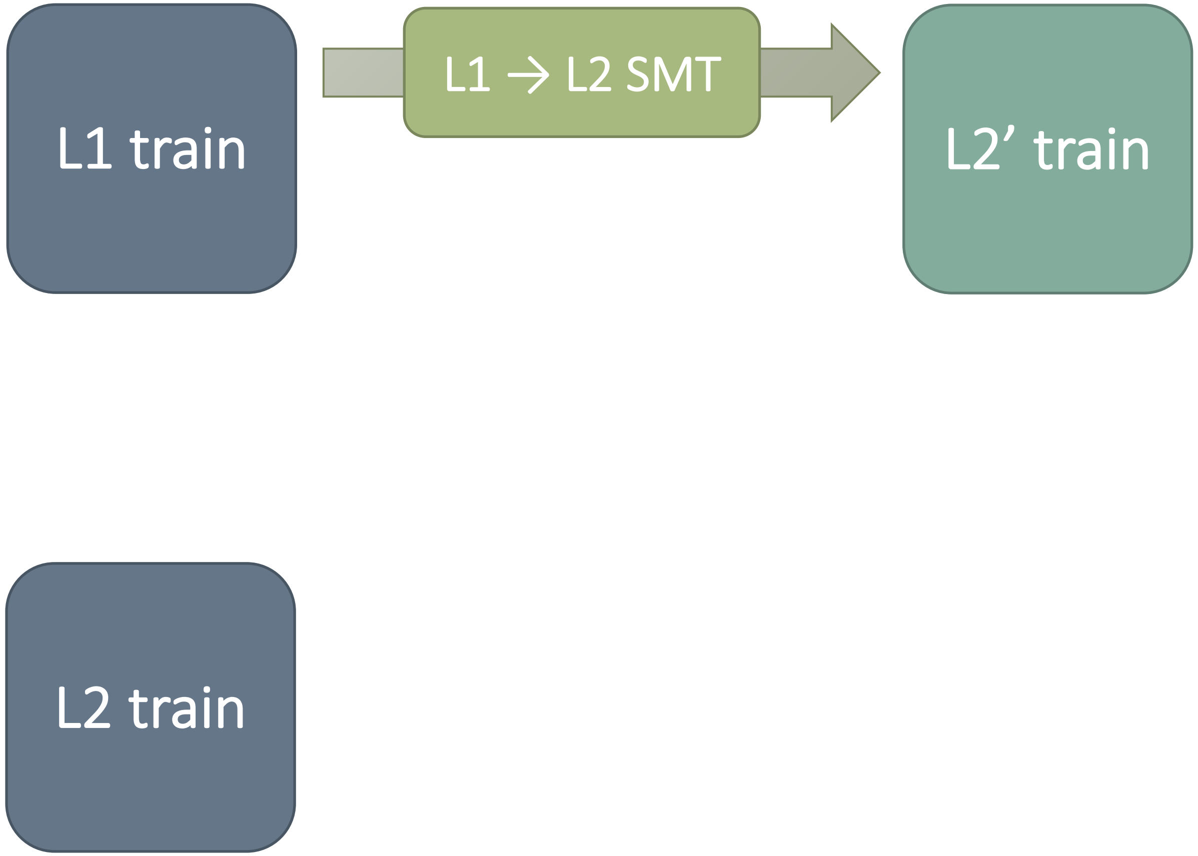
NMT hybridization



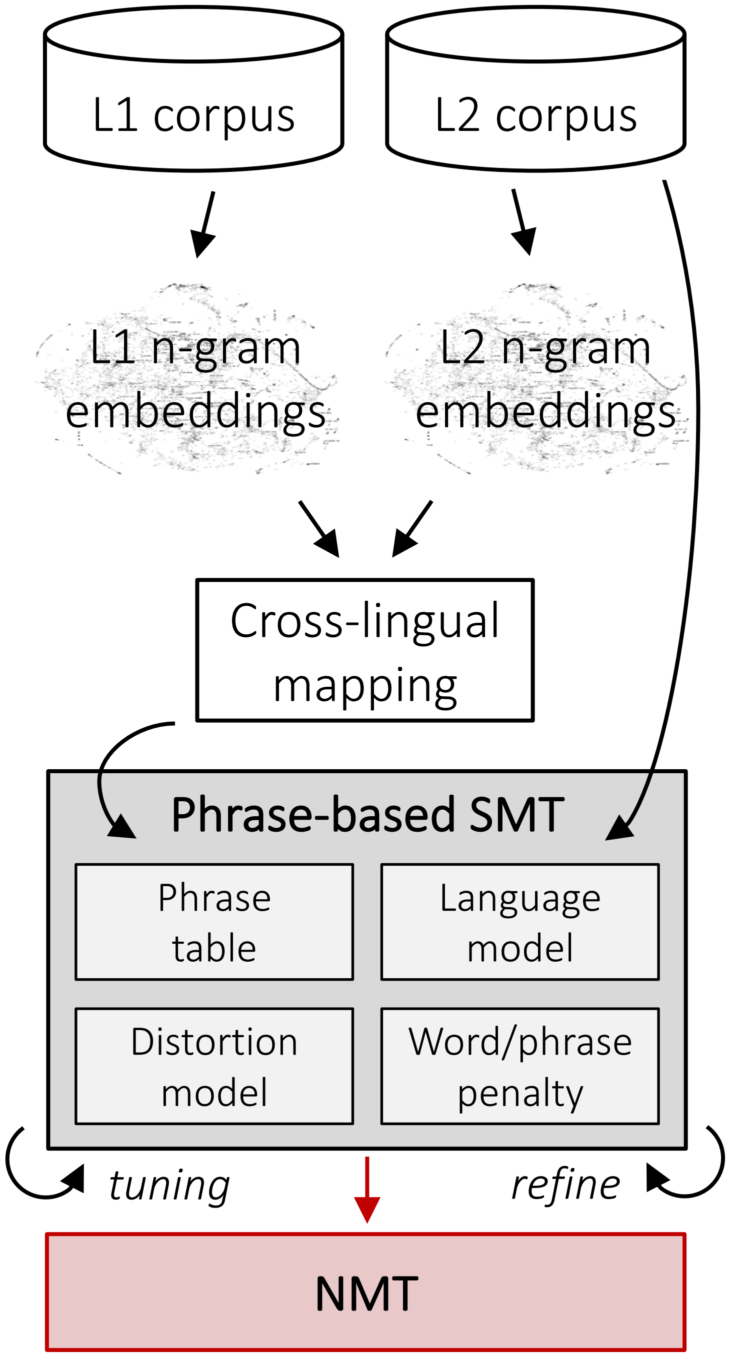


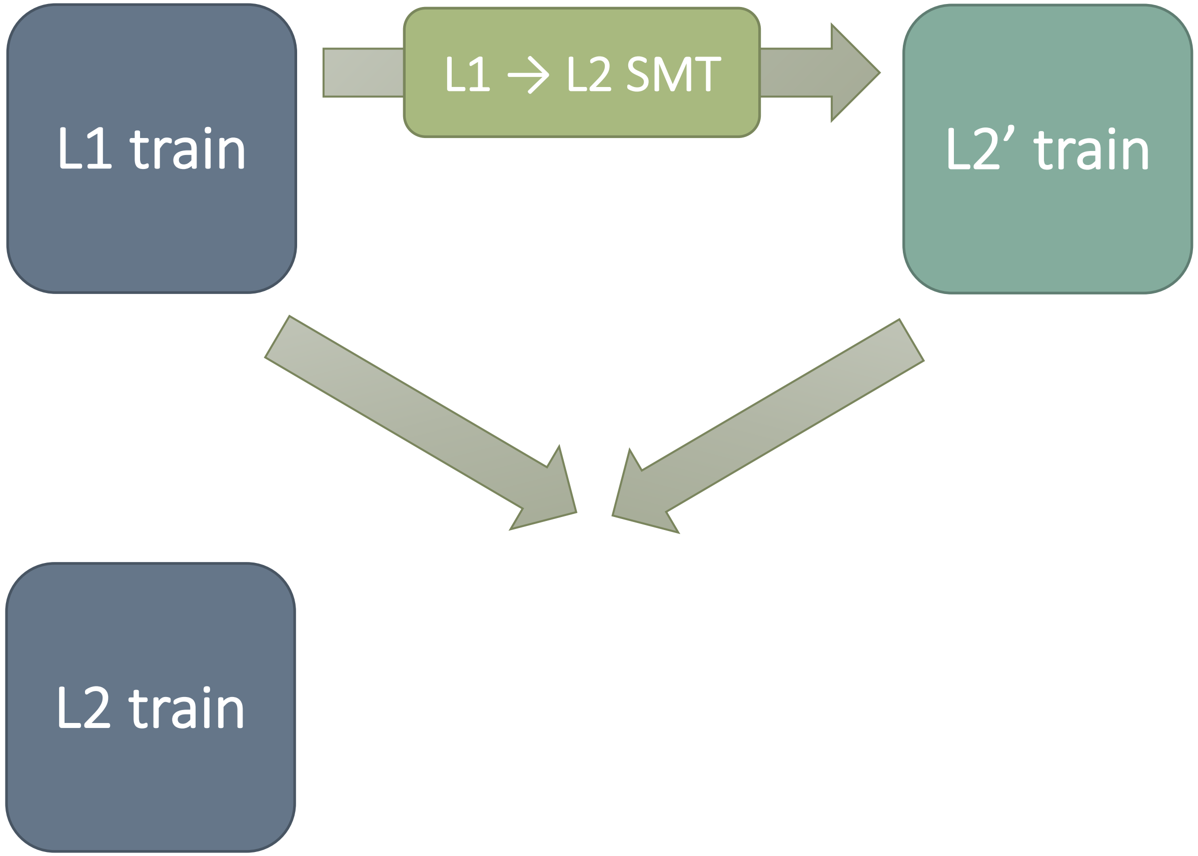
NMT hybridization



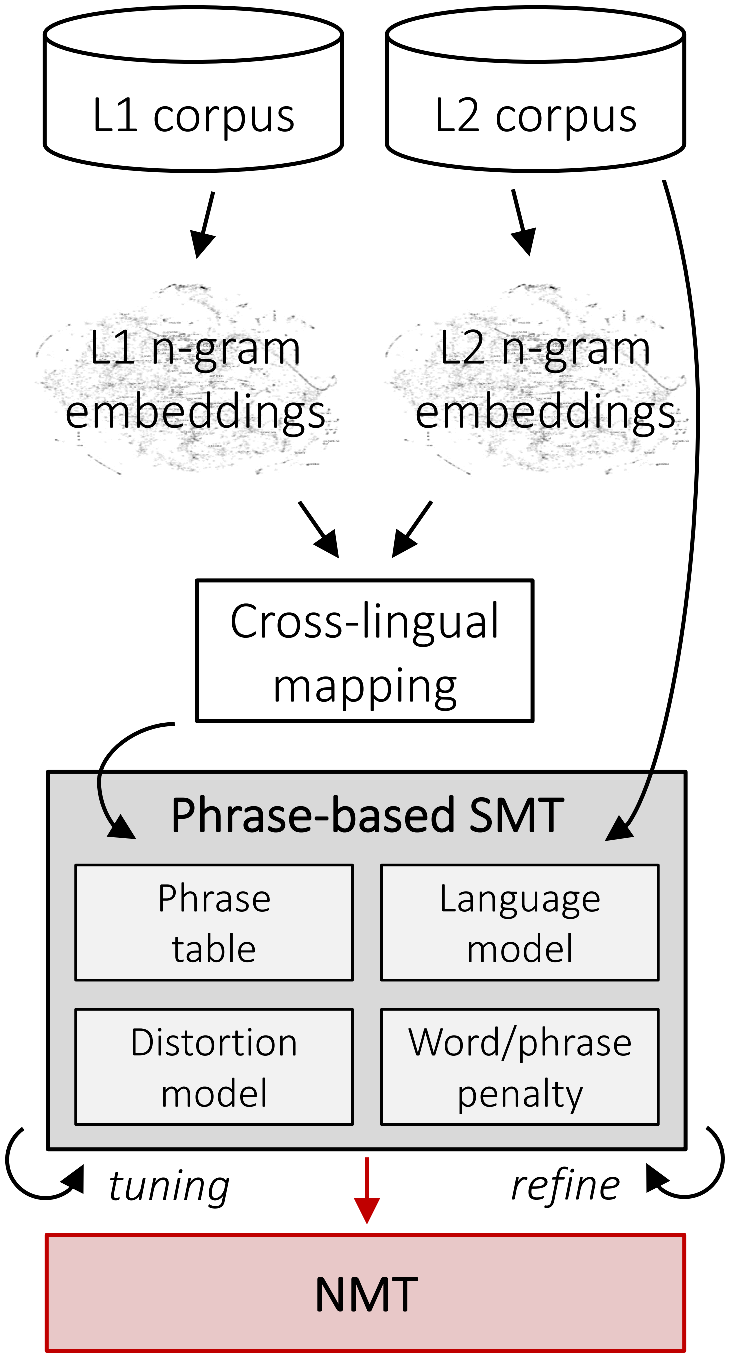


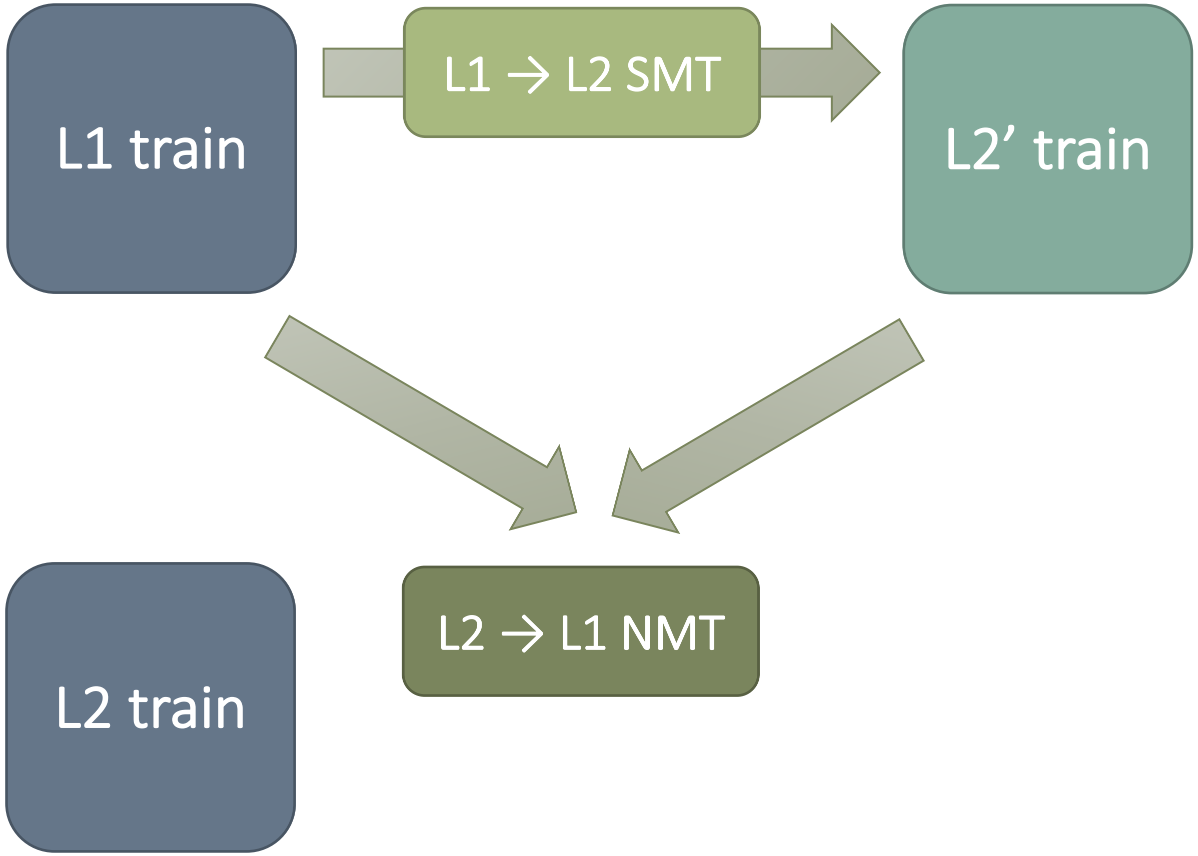
NMT hybridization



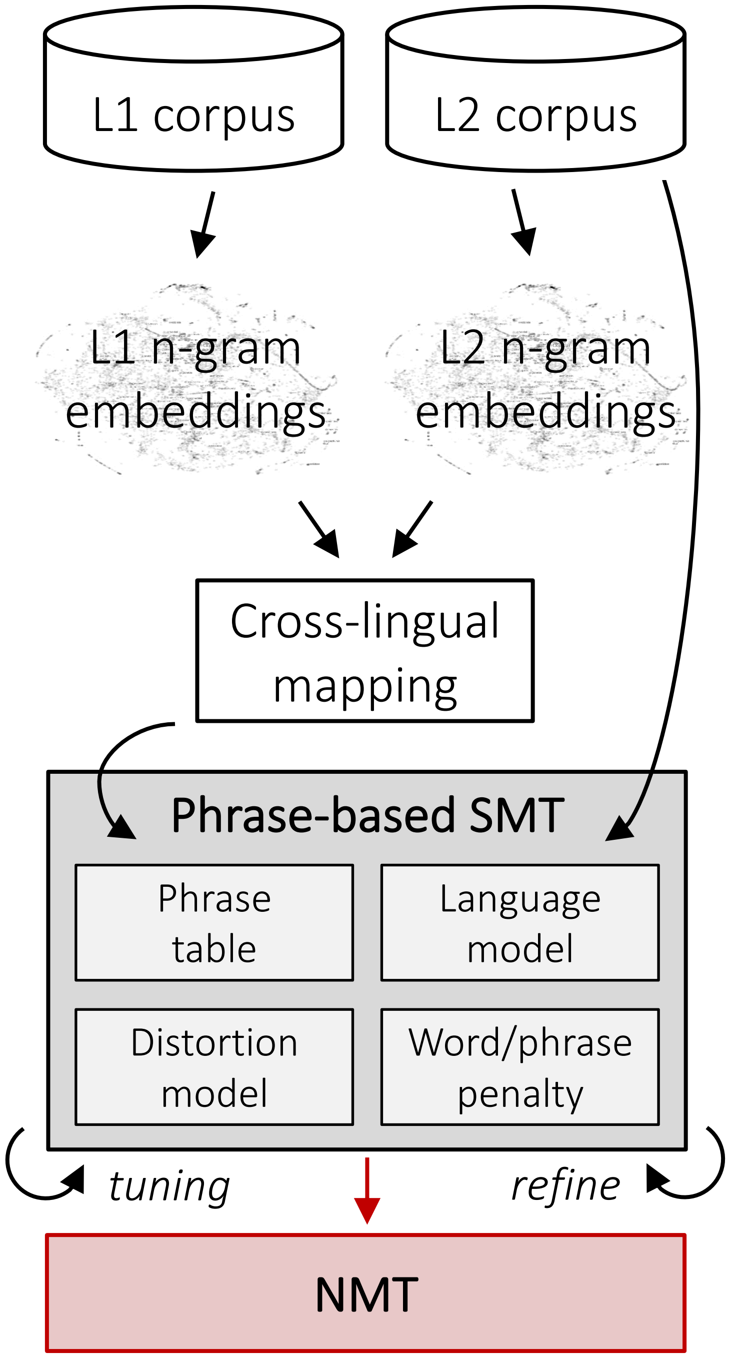


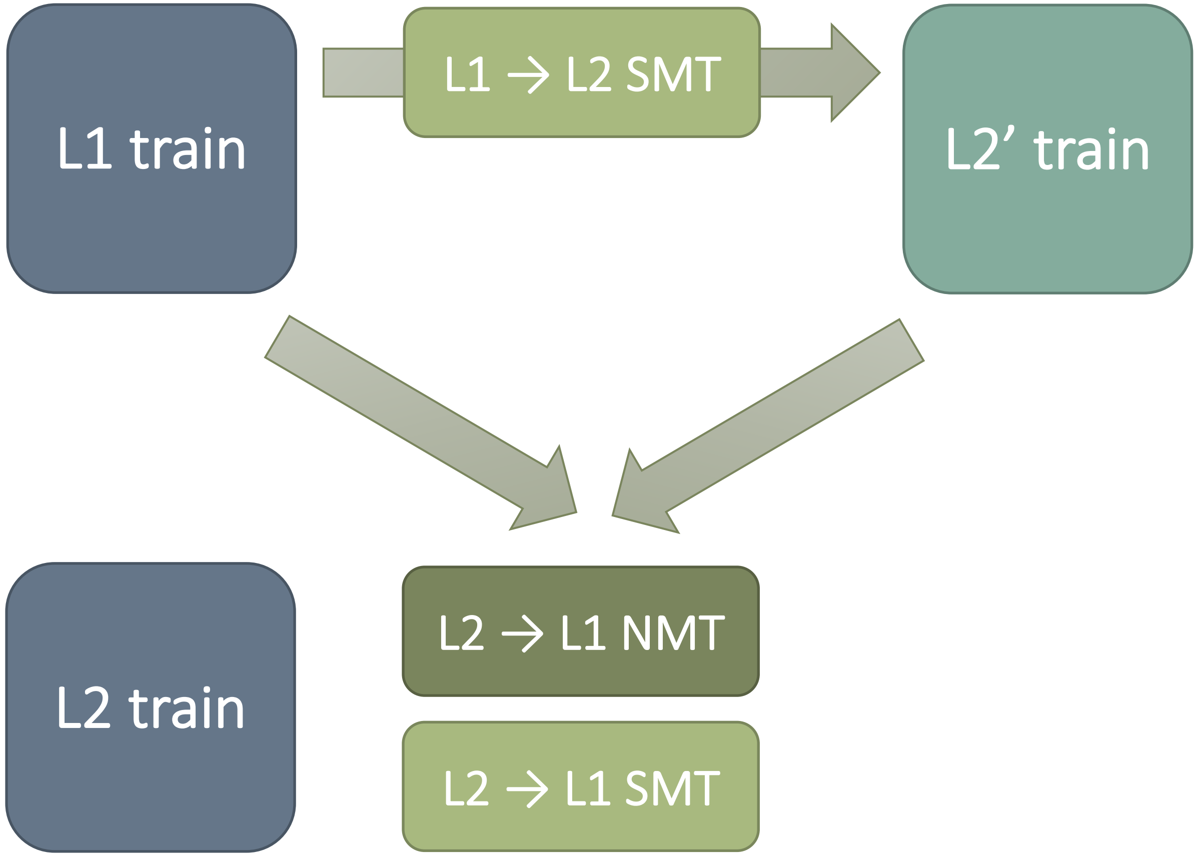
NMT hybridization



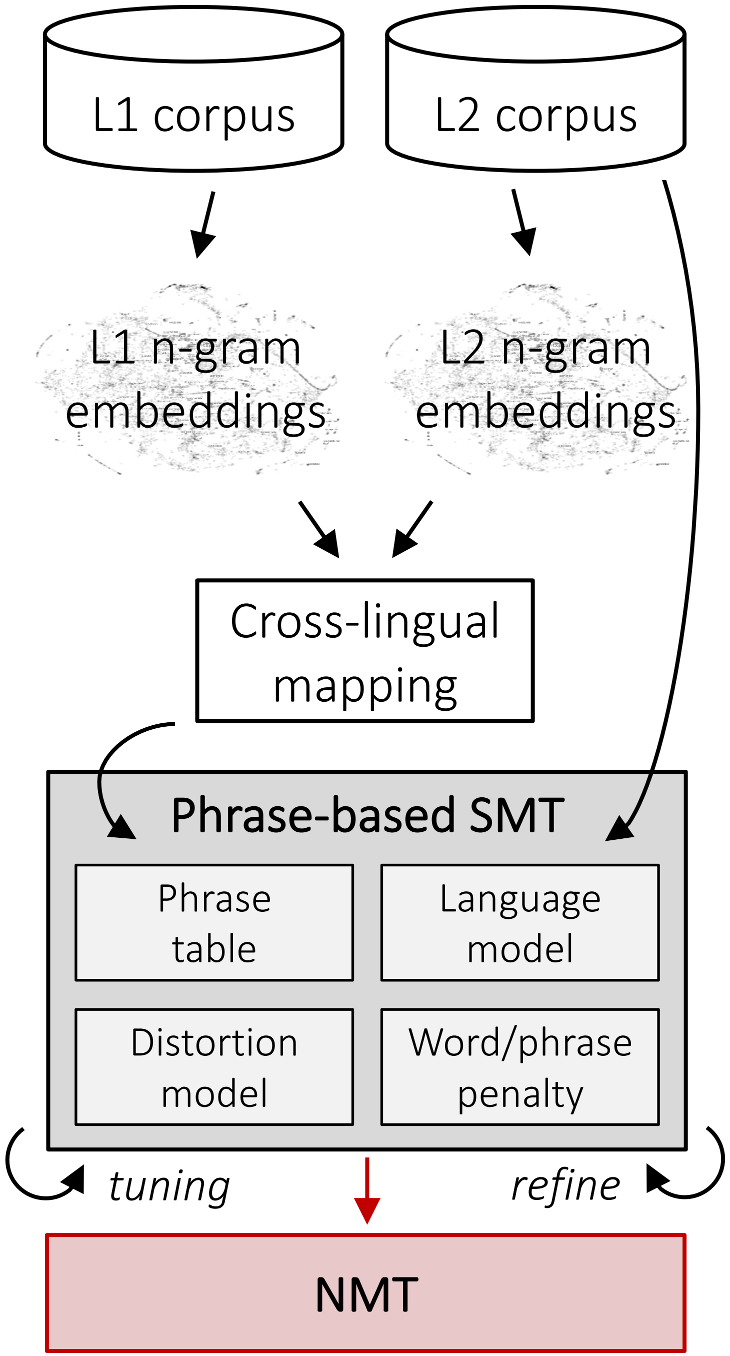


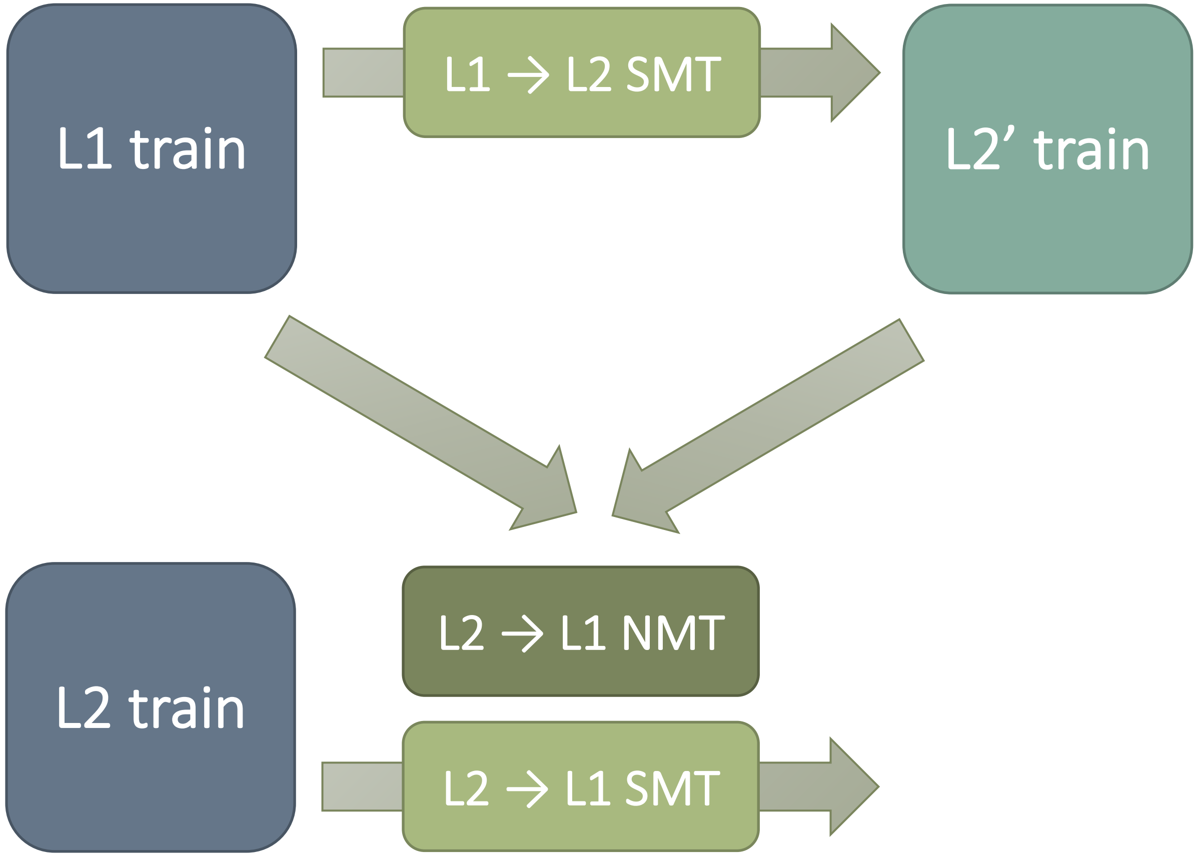
NMT hybridization



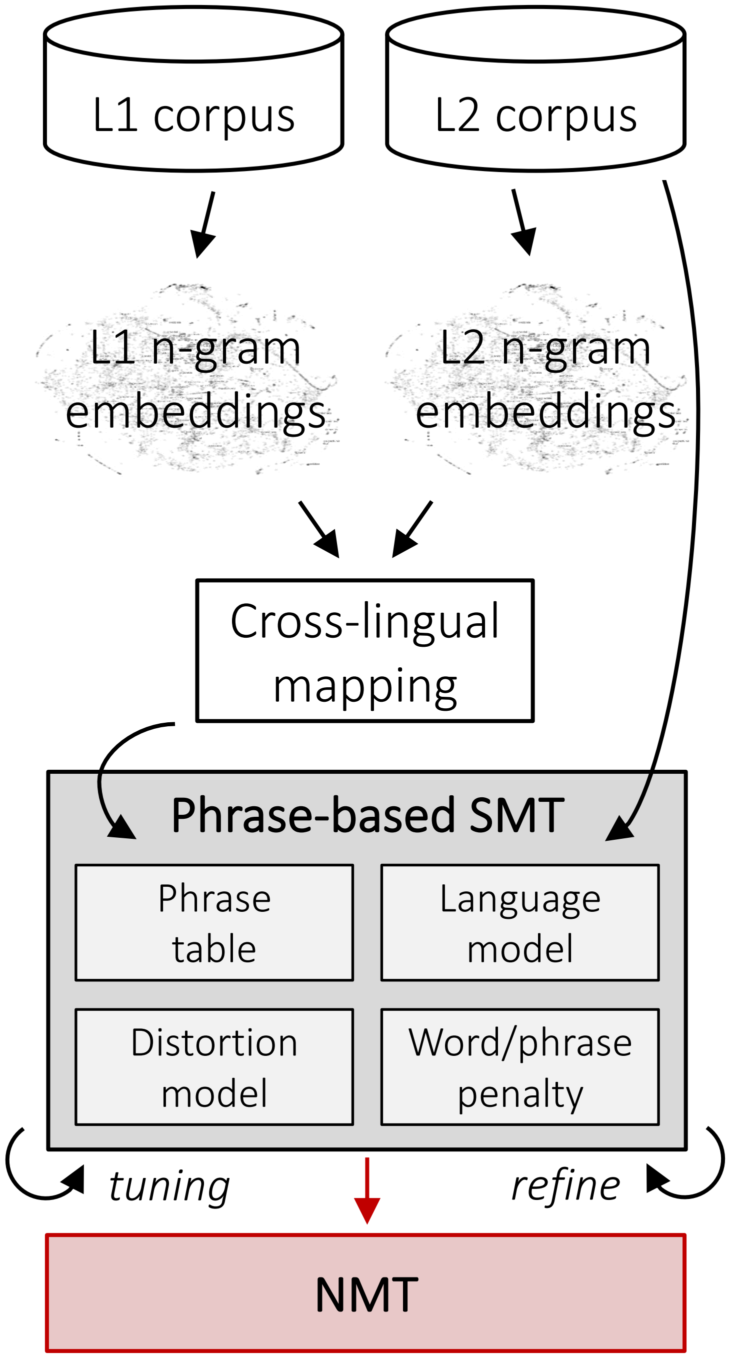


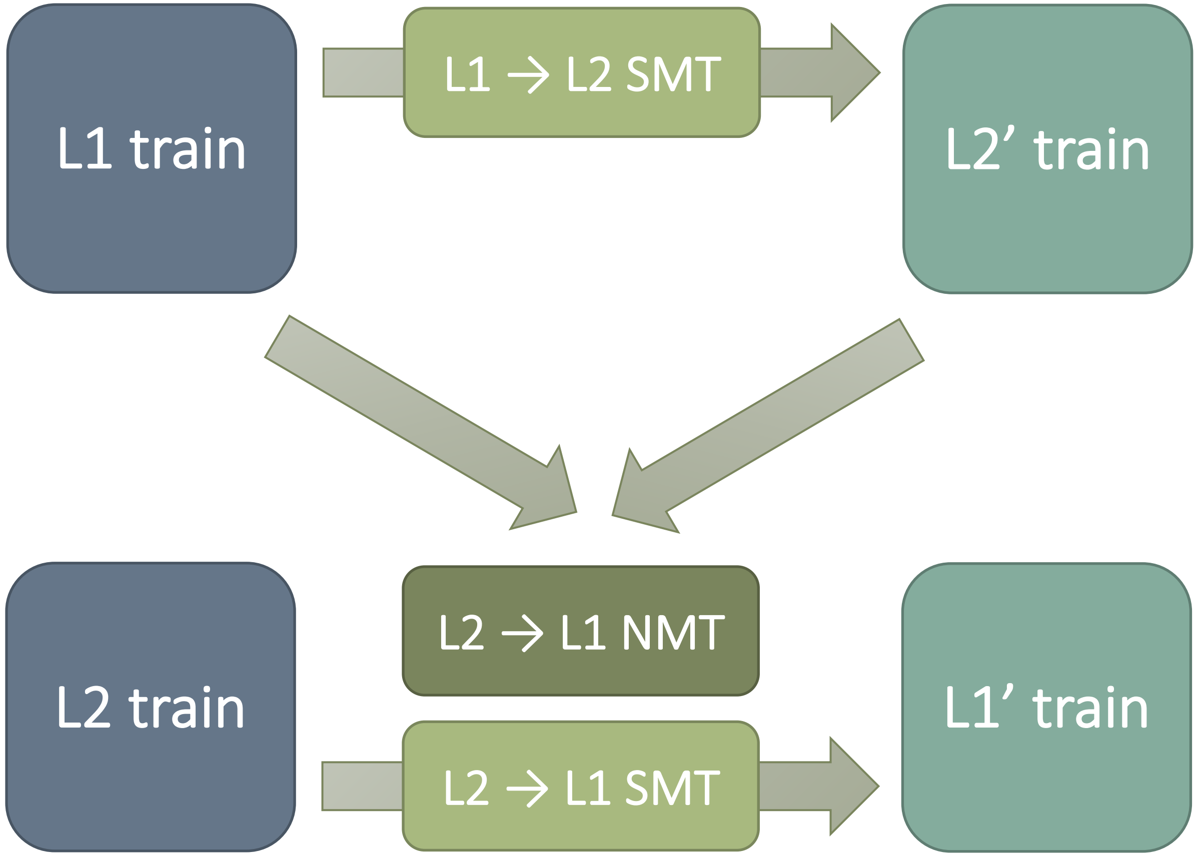
NMT hybridization



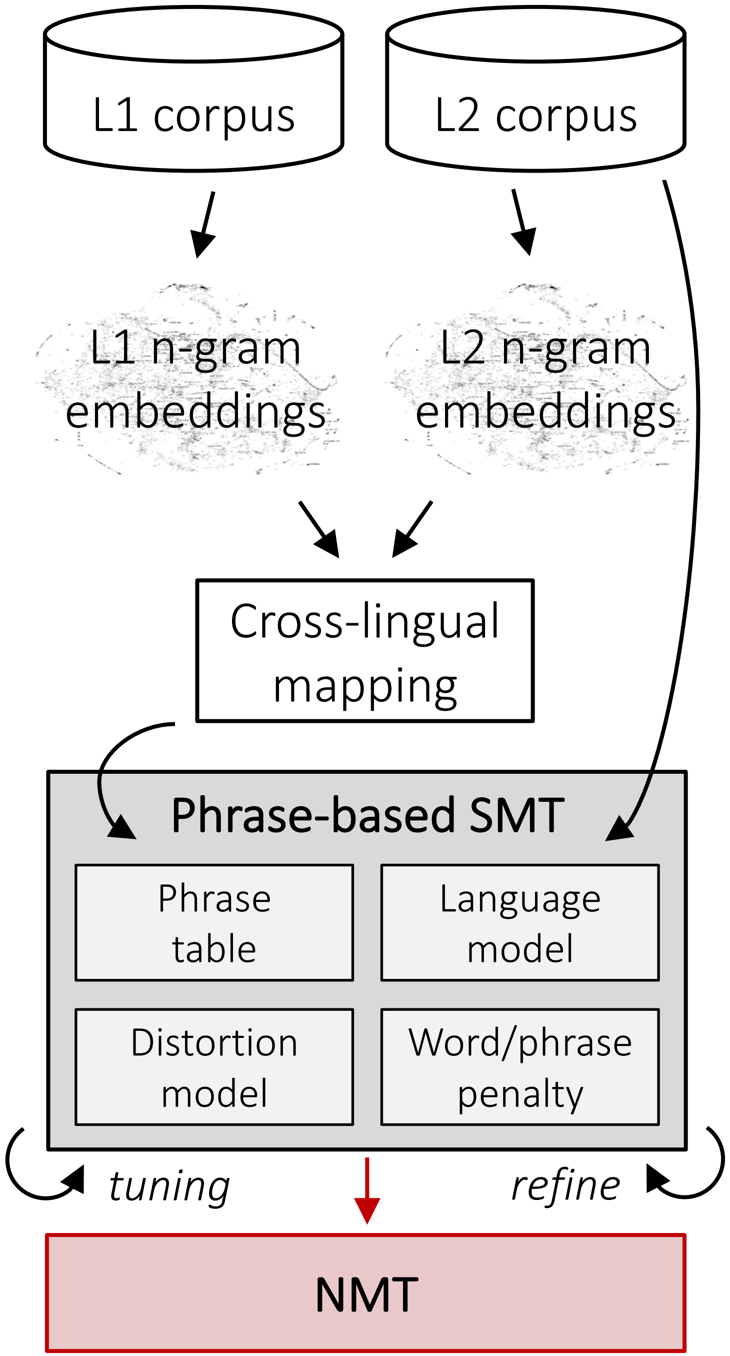


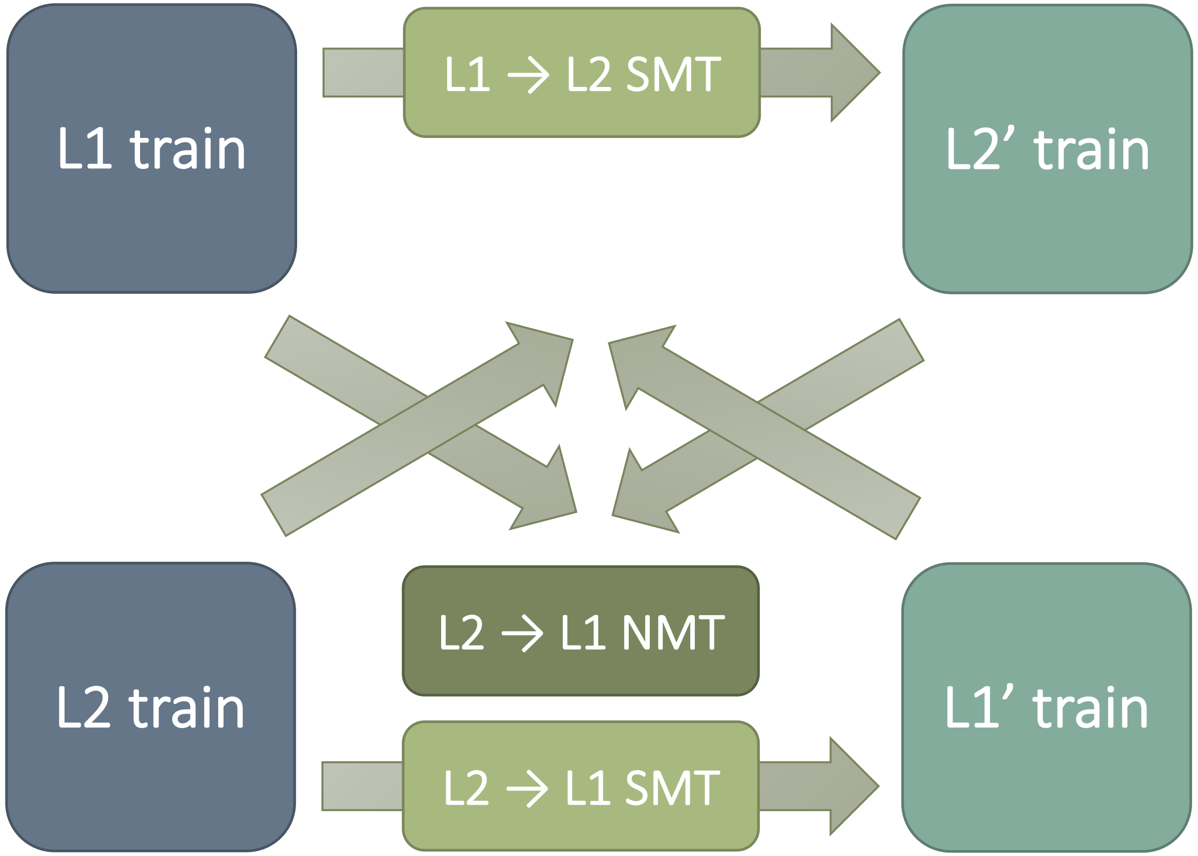
NMT hybridization



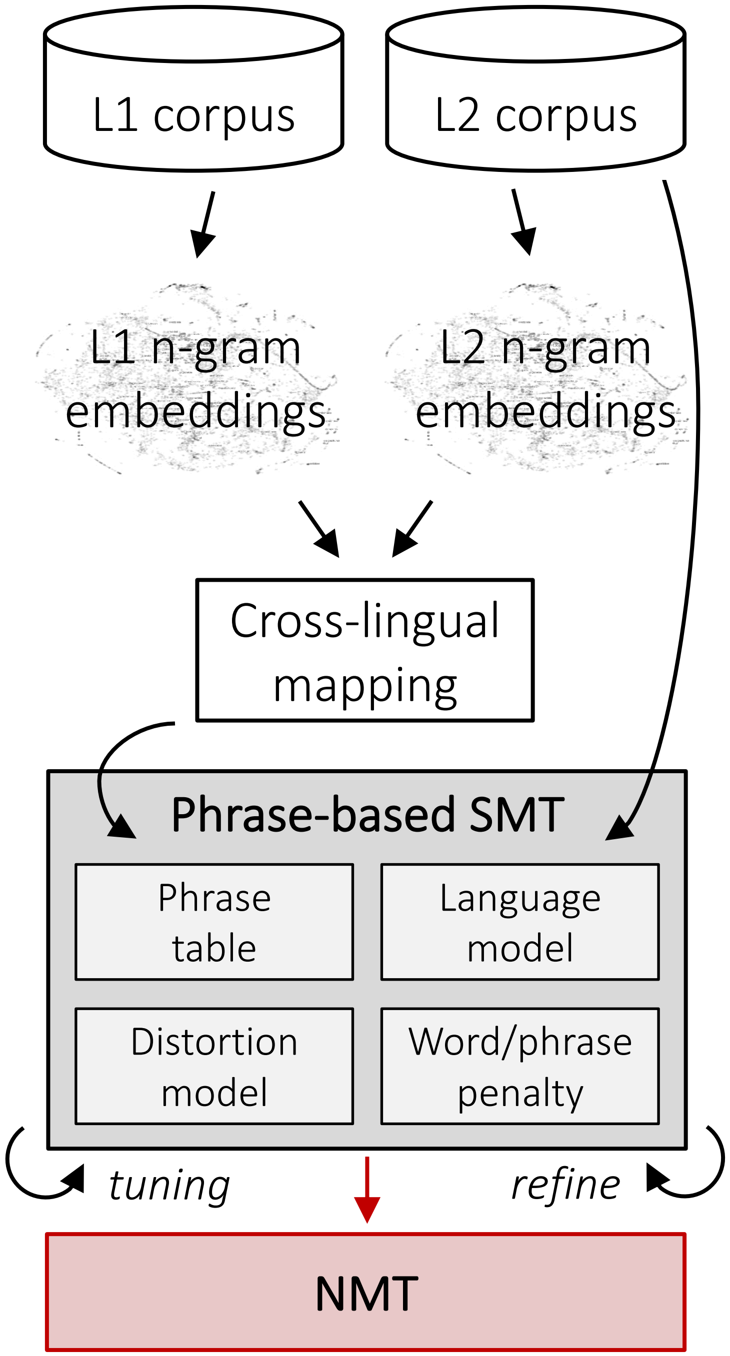


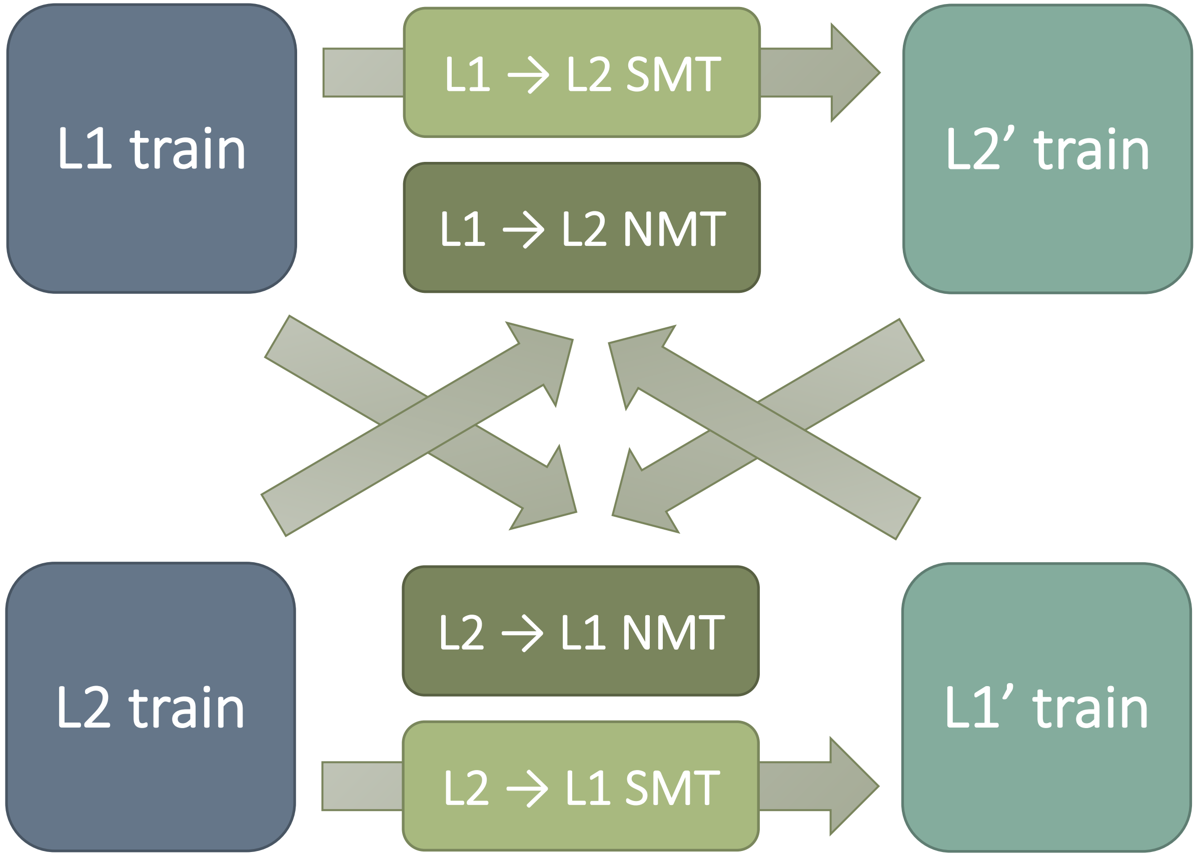
NMT hybridization



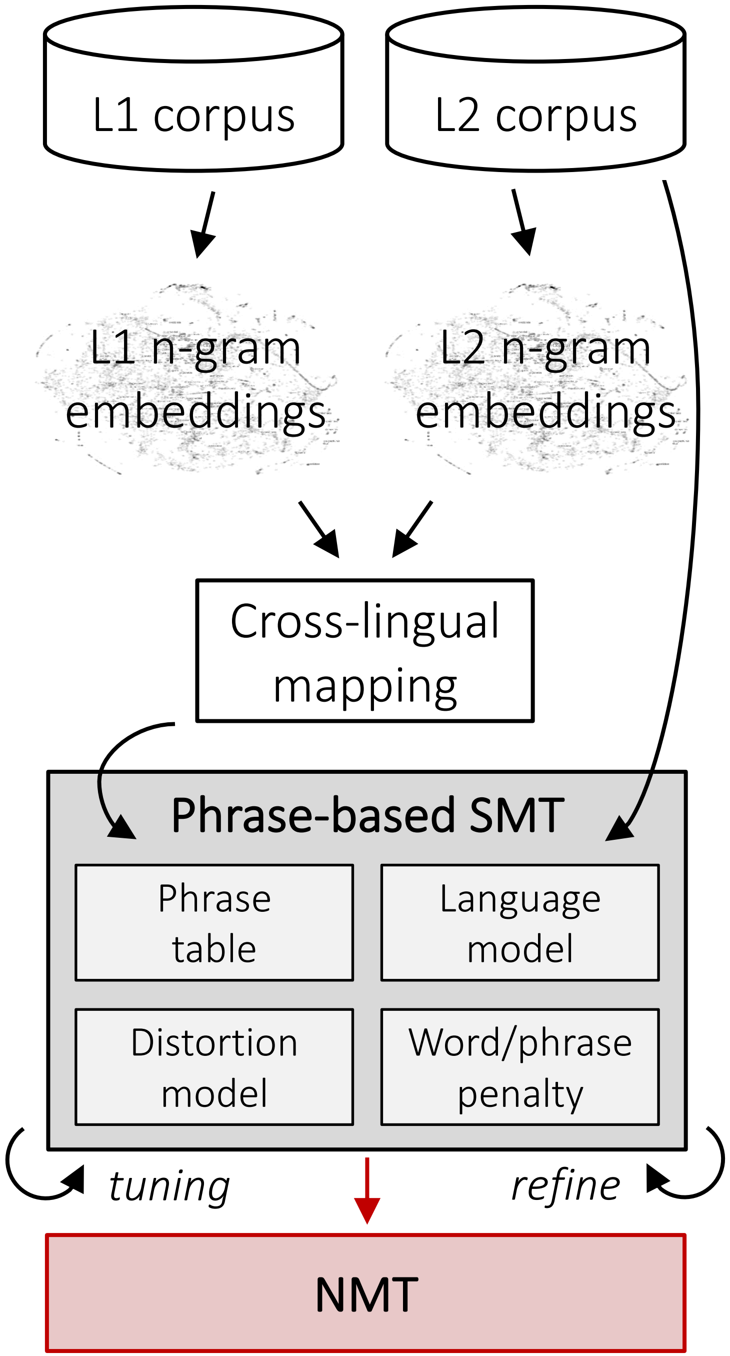


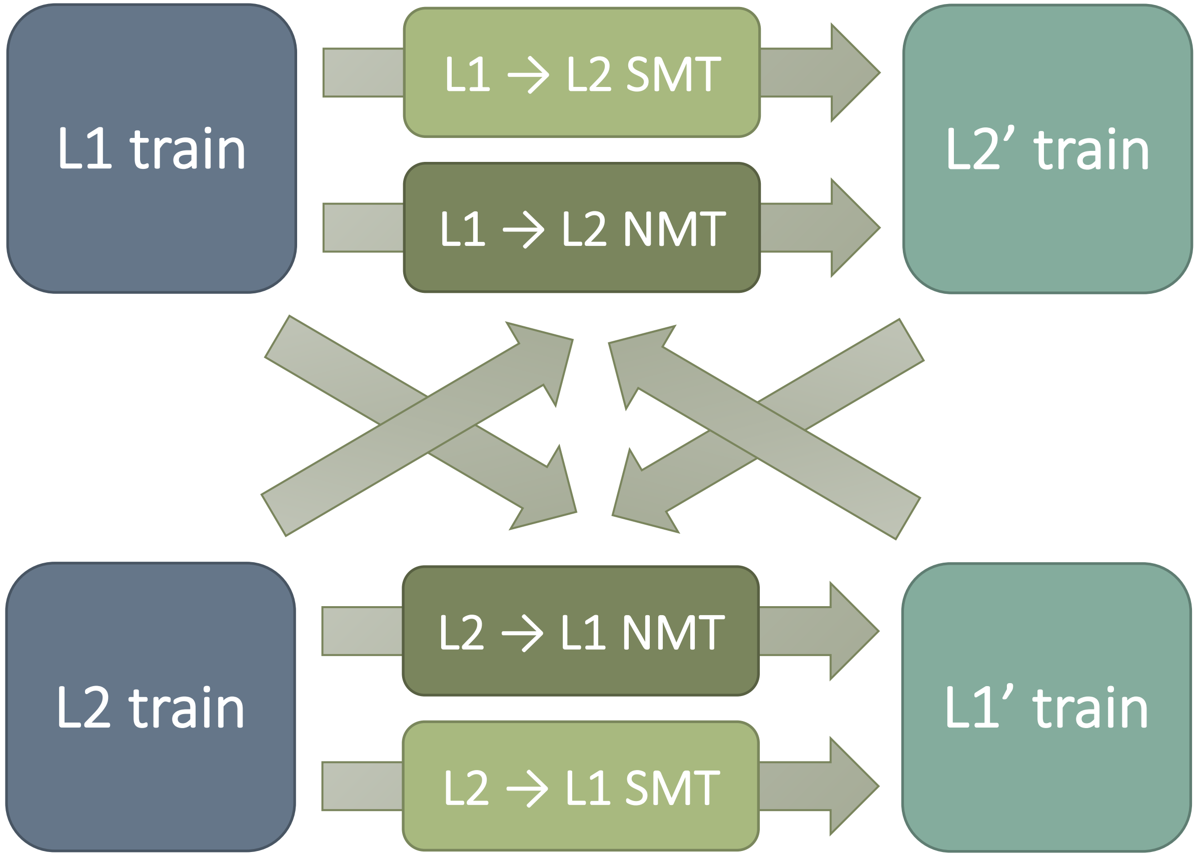
NMT hybridization



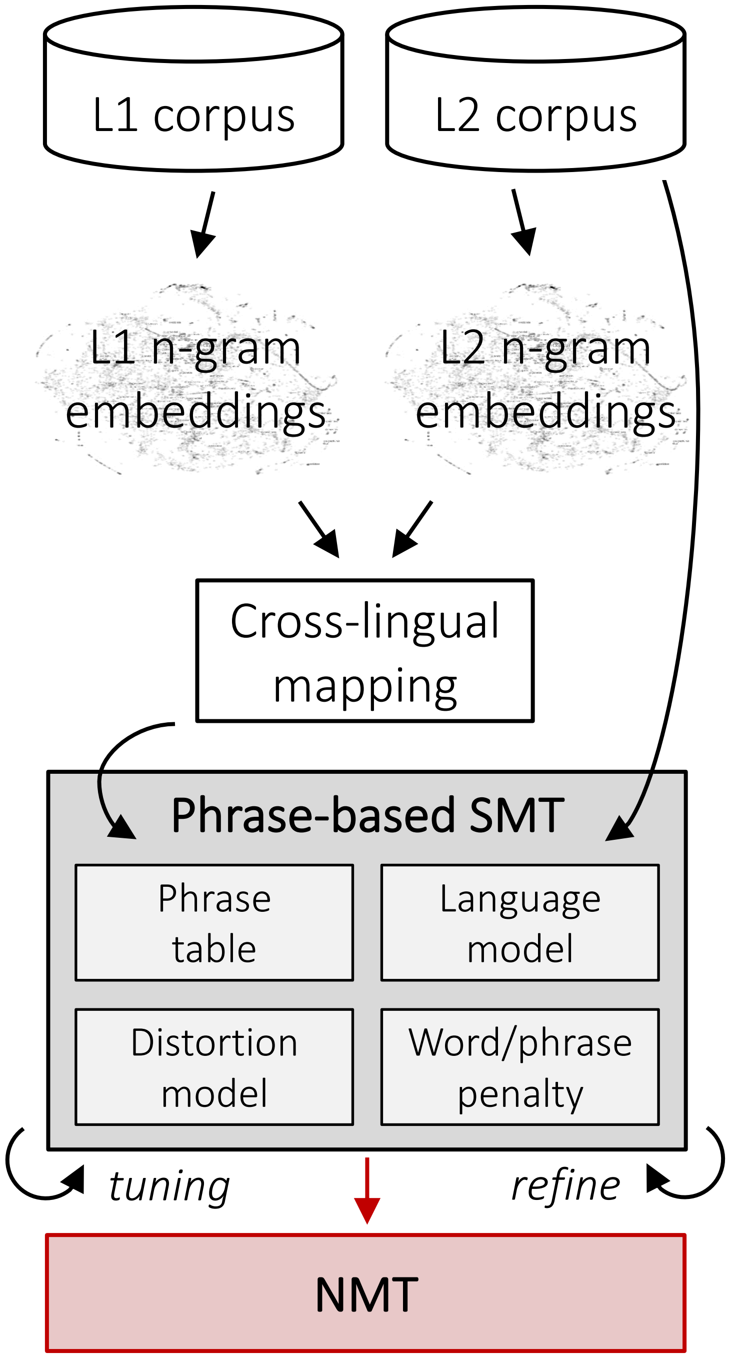


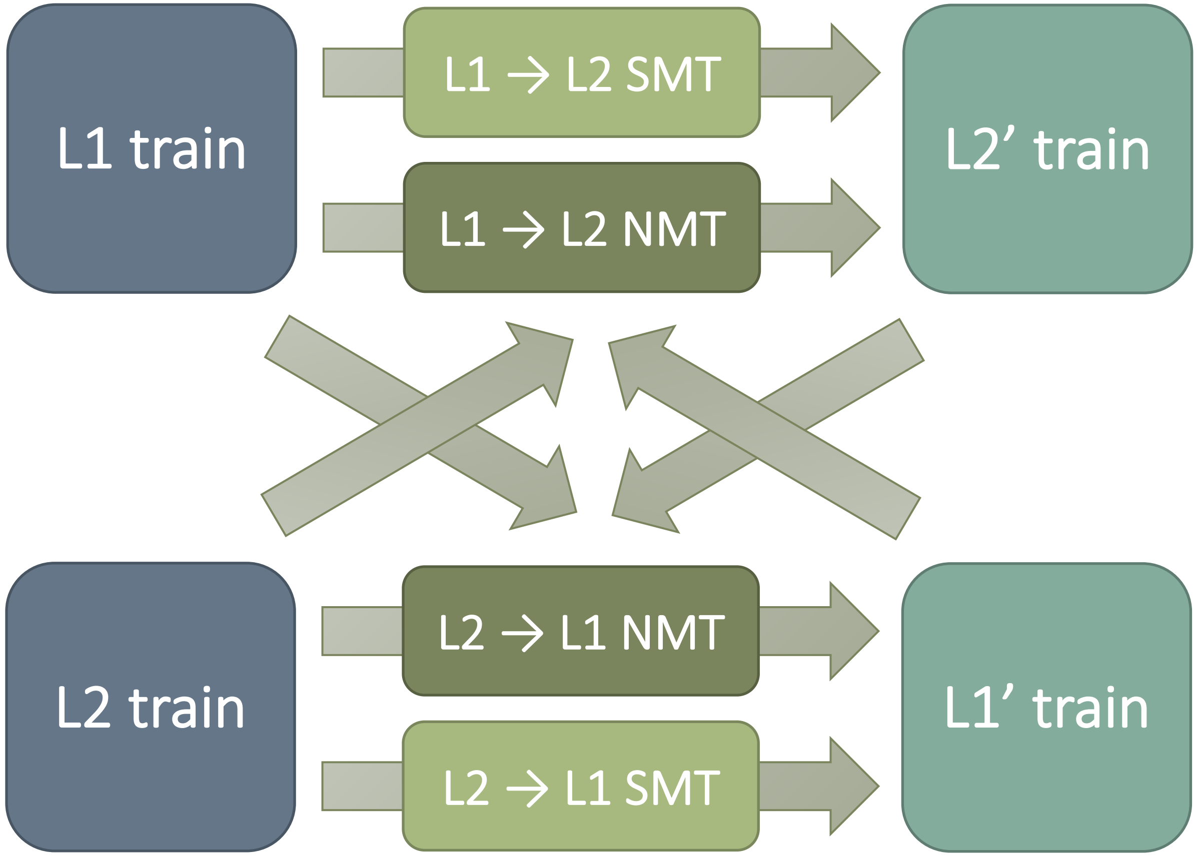
NMT hybridization





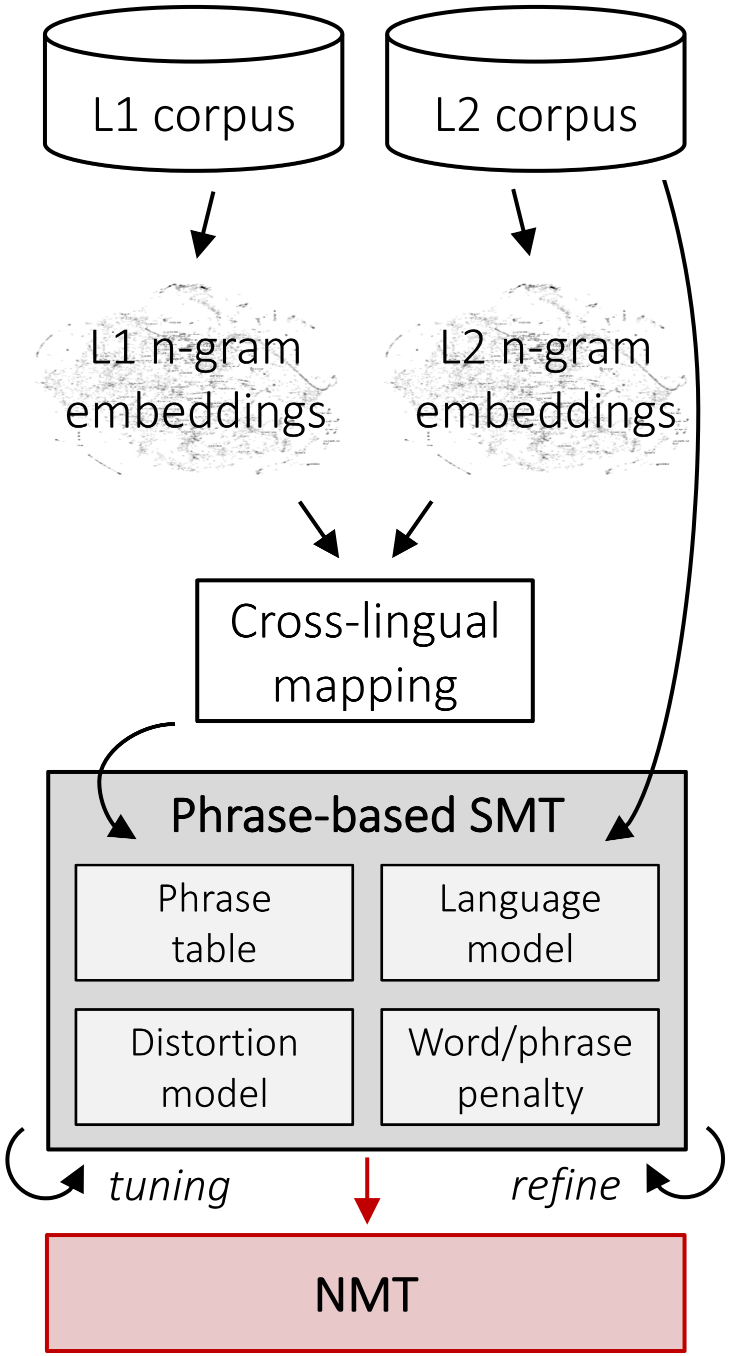
NMT hybridization

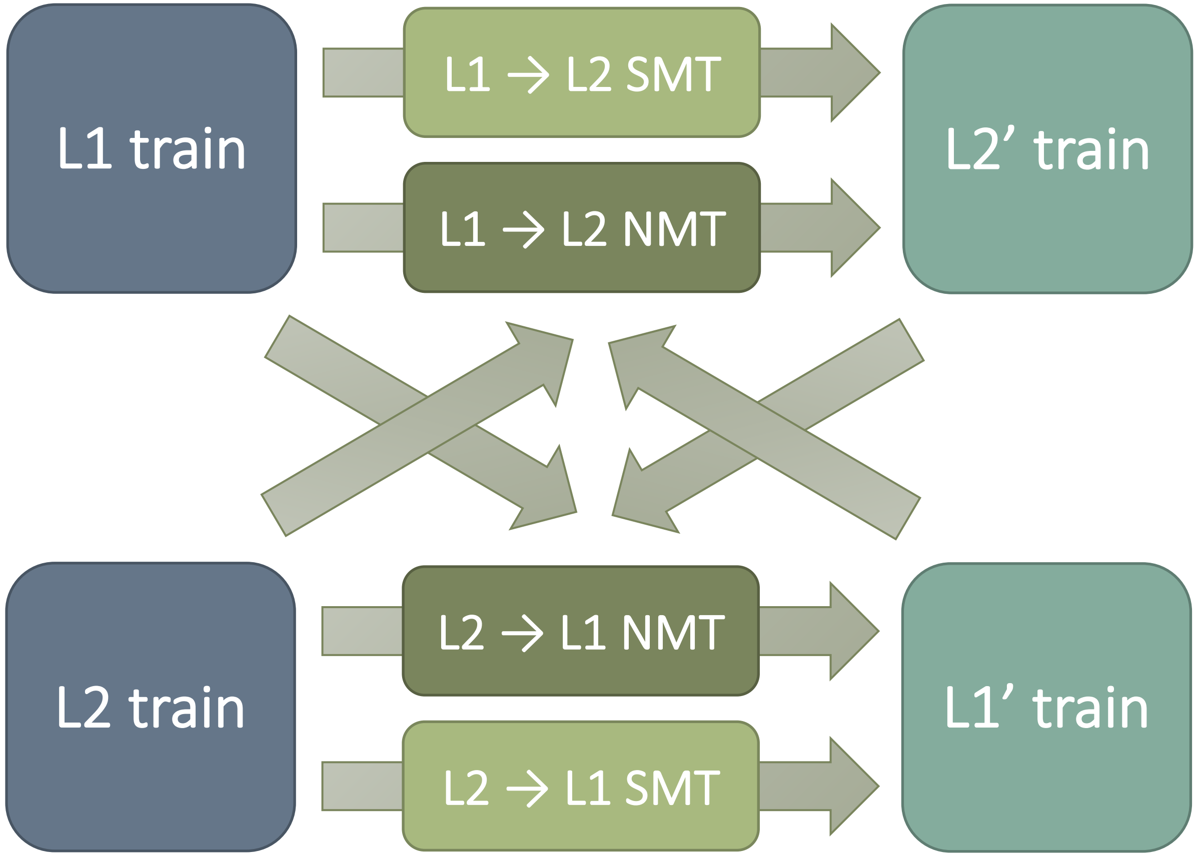




$$N_{SMT} = N \cdot \max(0, 1 - t/a)$$

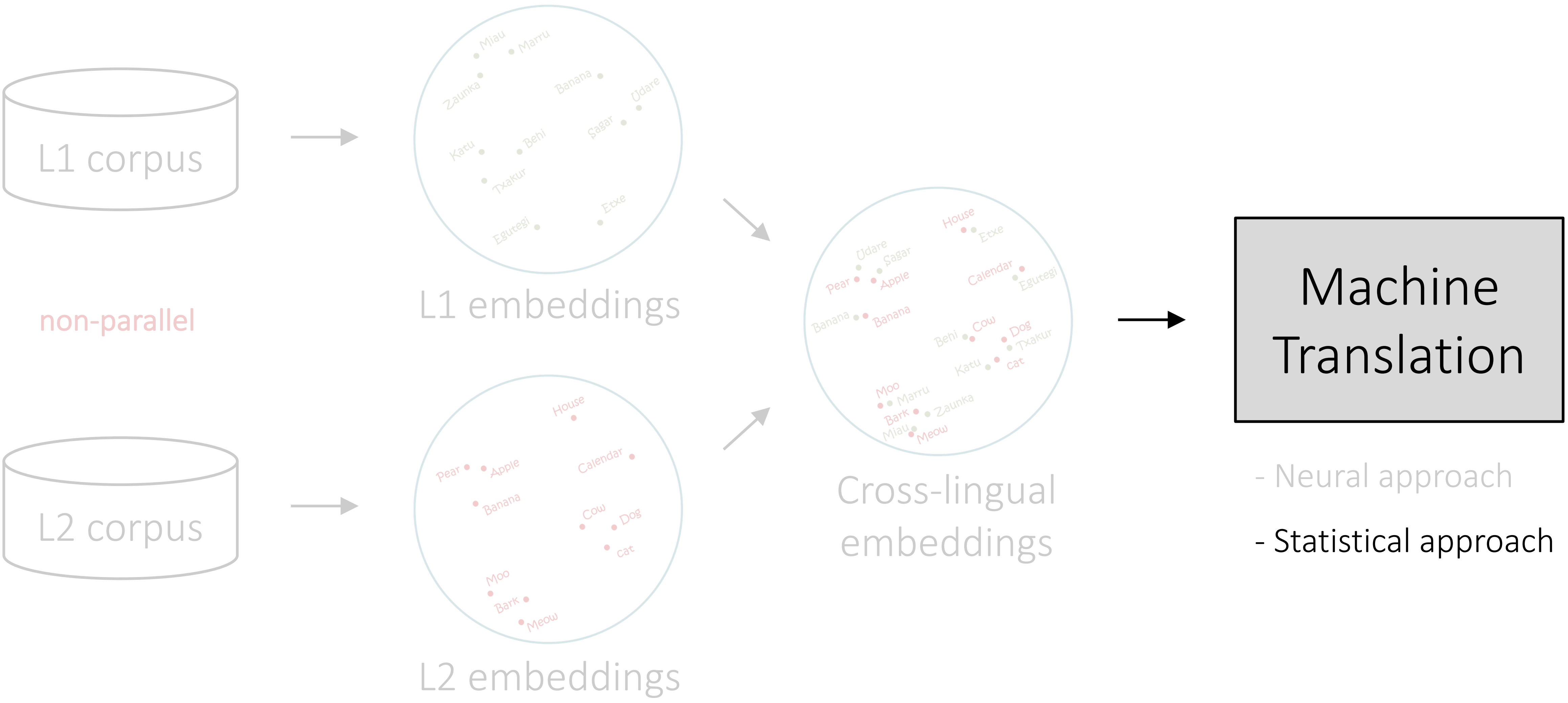
NMT hybridization



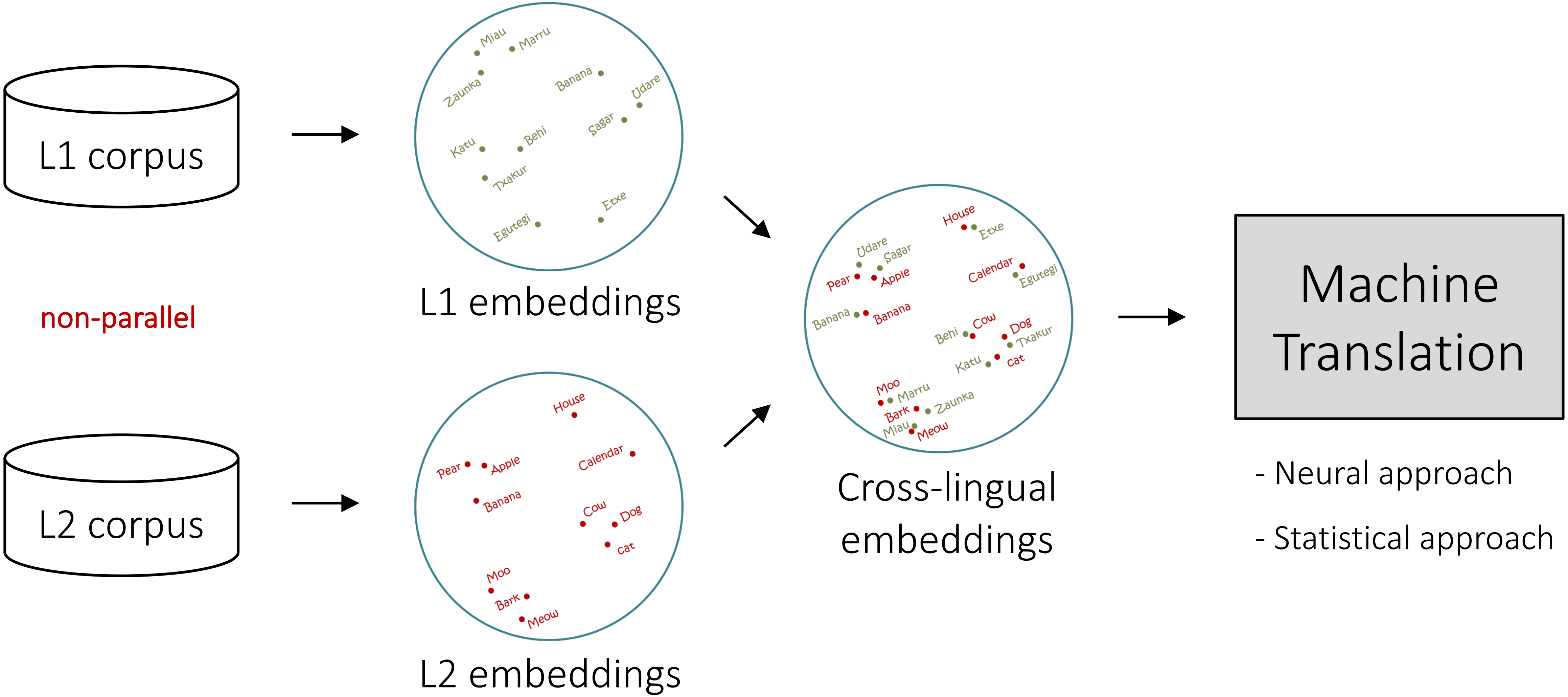


$$N_{SMT} = N \cdot \max(0, 1 - t/a)$$
$$N_{NMT} = N - N_{SMT}$$

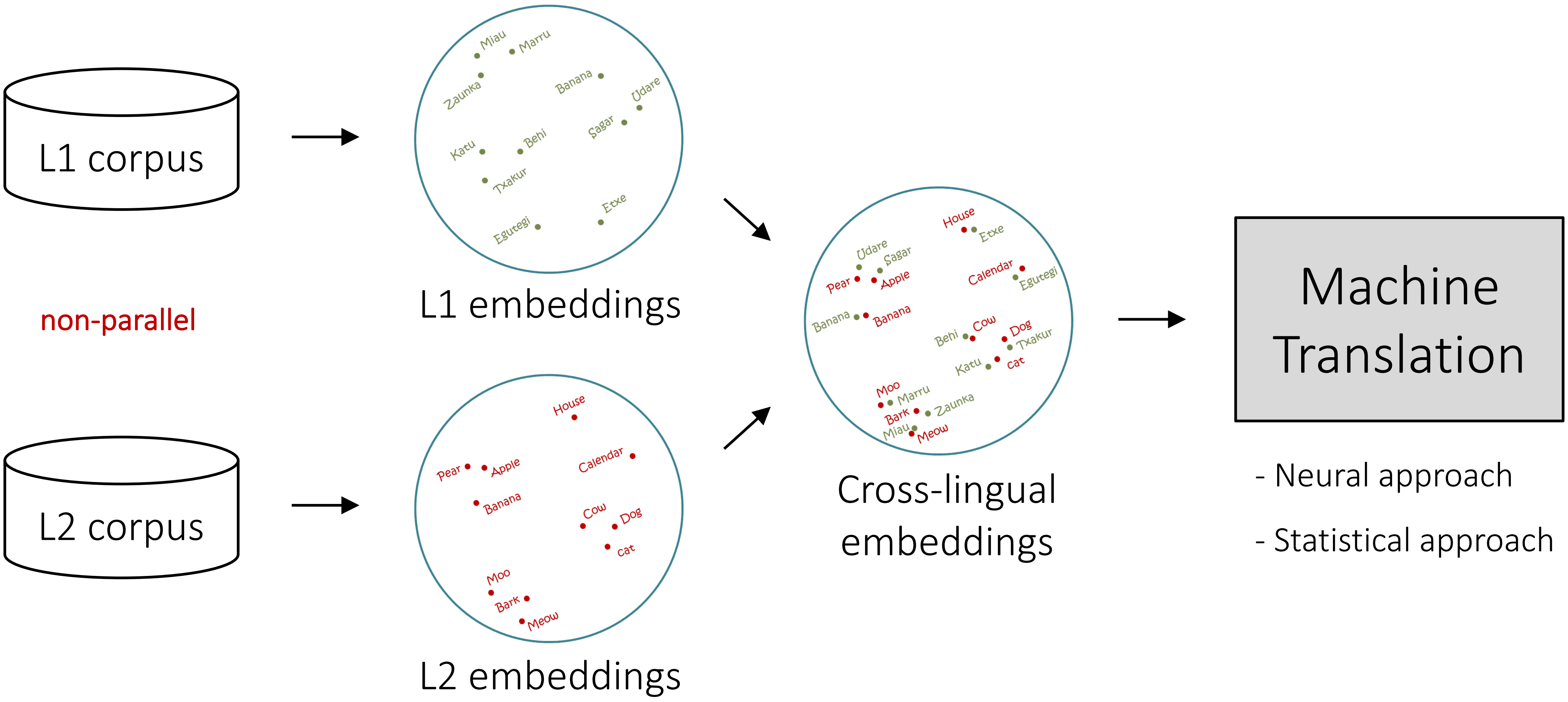
Outline


L1 corpus
L2 corpus
non-parallel
L1 embeddings
L2 embeddings
Cross-lingual embeddings
Machine Translation
- Neural approach
- Statistical approach

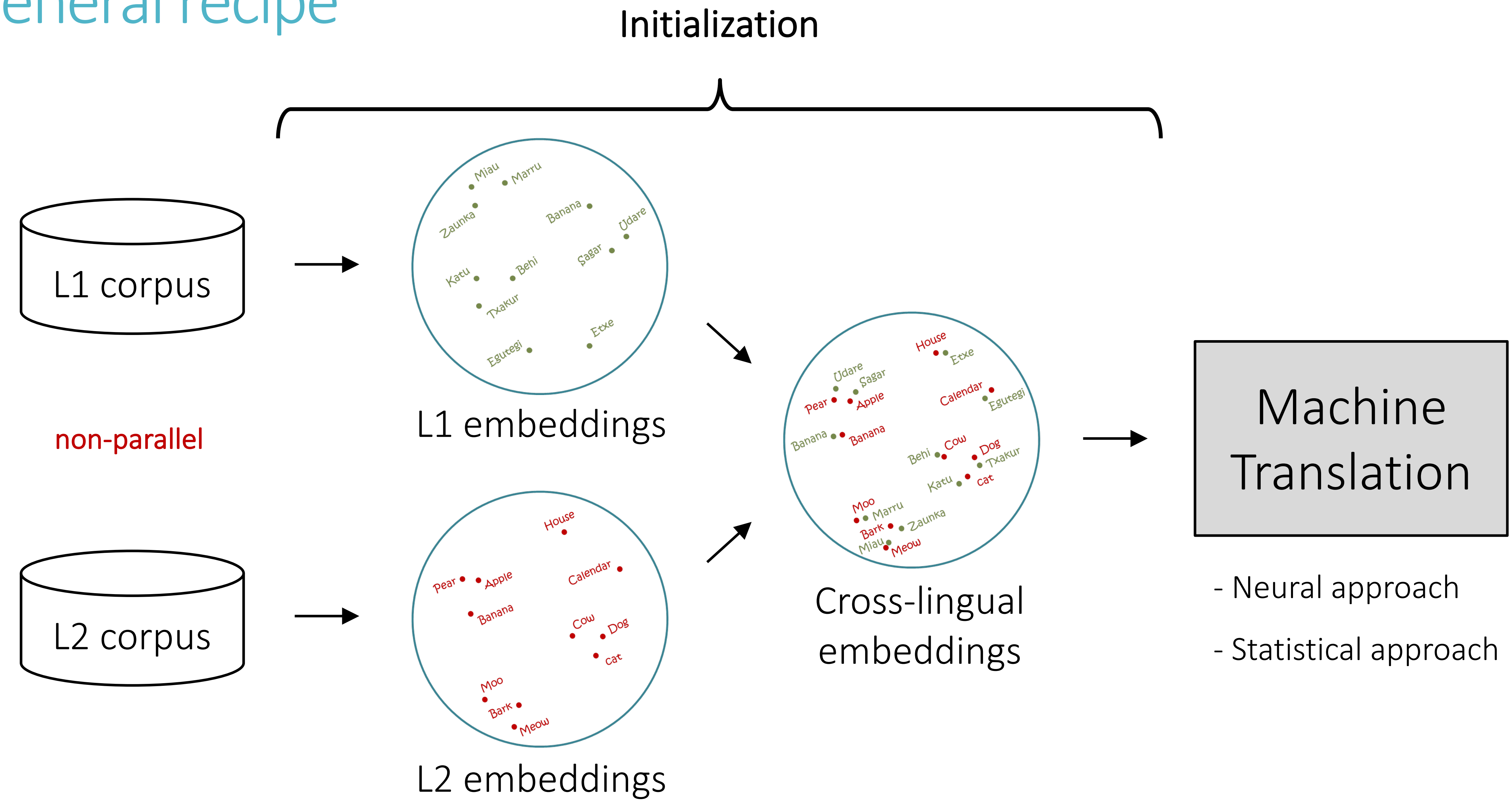
Outline


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

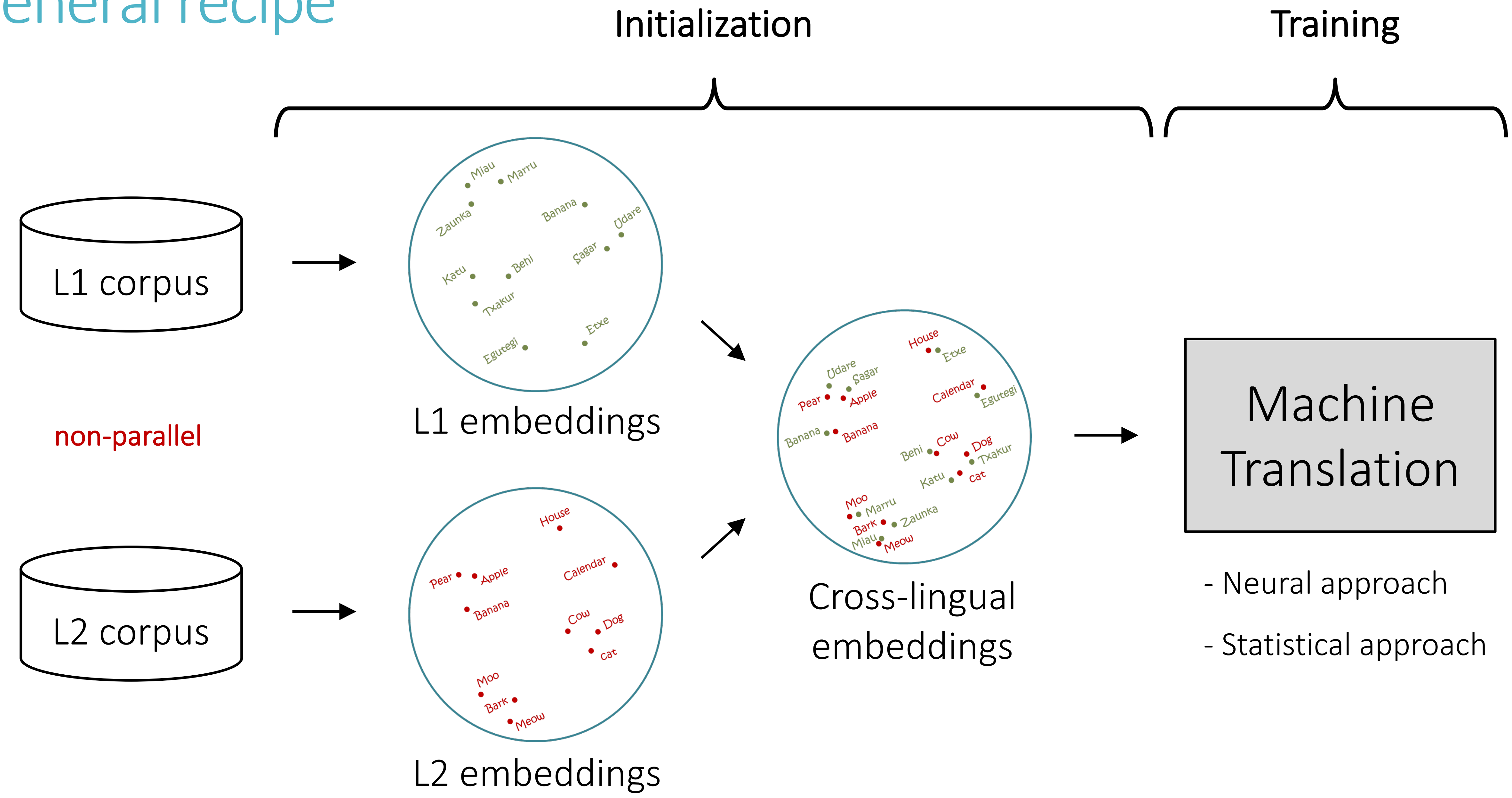
General recipe



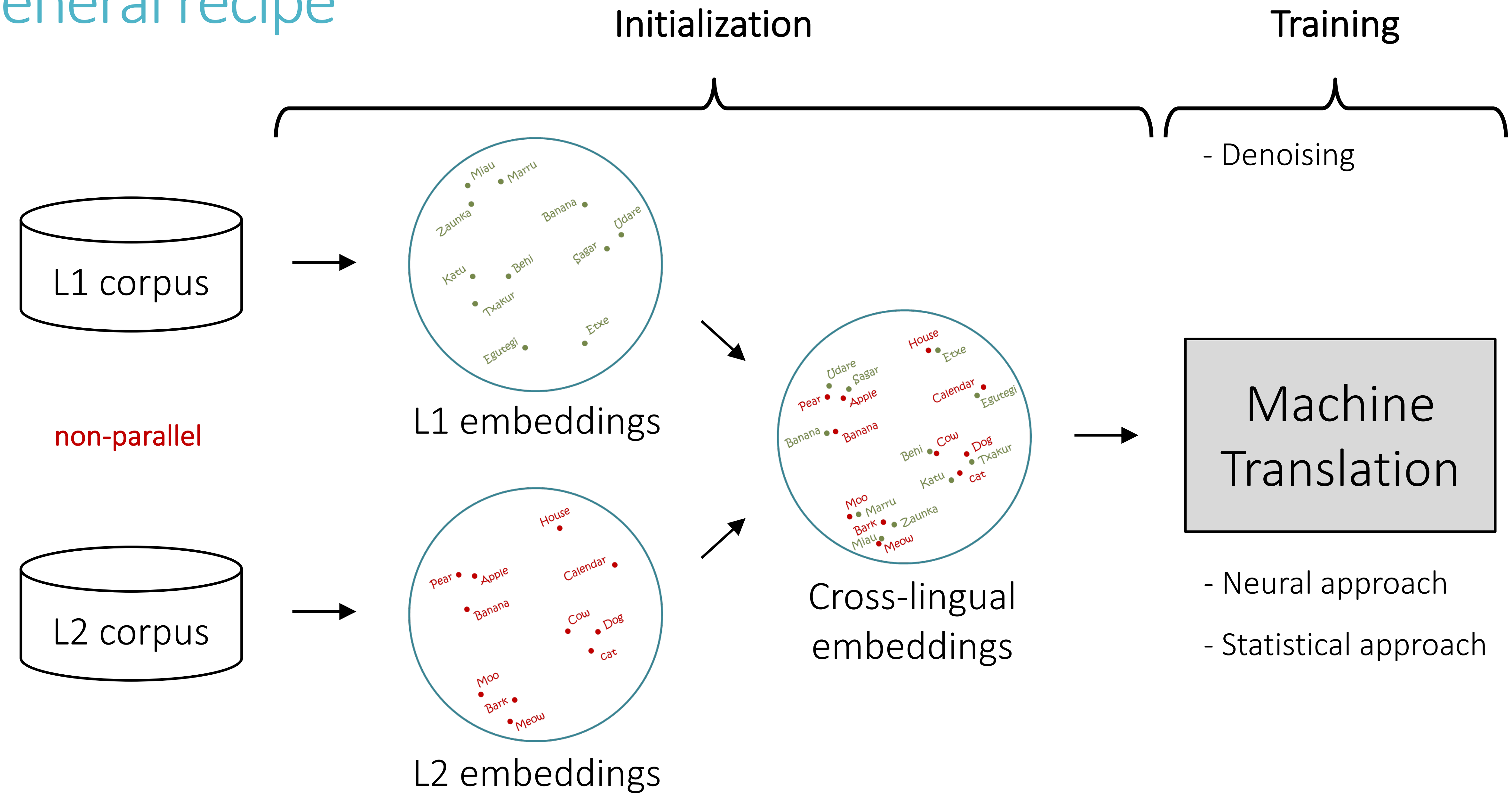
General recipe



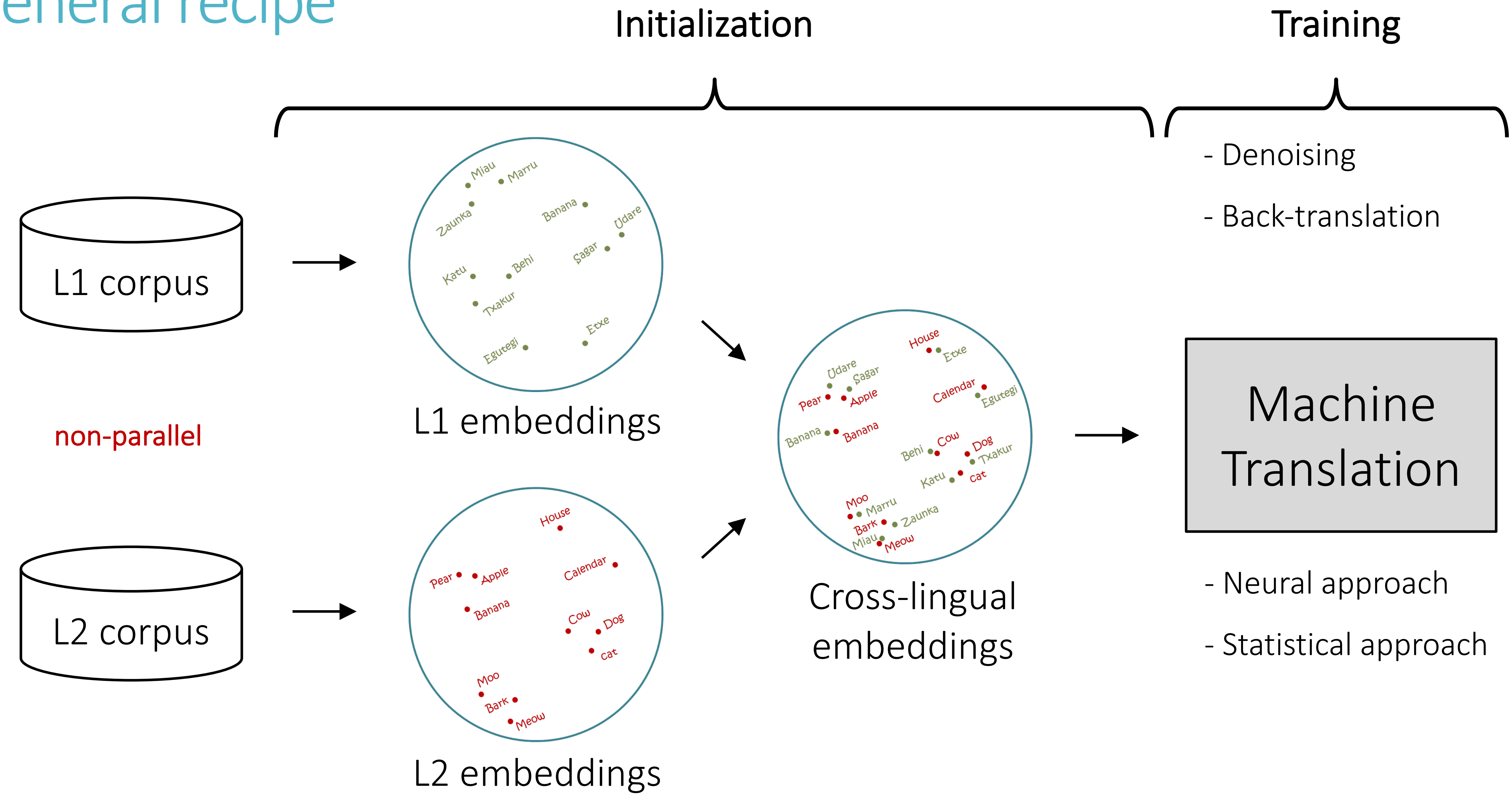
General recipe



General recipe



General recipe



General recipe

Initialization

Training

- Denoising
- Back-translation

L1 corpus

L1 embeddings

non-parallel

L2 corpus

L2 embeddings

Cross-lingual embeddings

Machine Translation

- Neural approach
- Statistical approach

General recipe

Initialization

- Cross-lingual word embeddings *(Artetxe et al., ICLR'18; Lample et al., ICLR'18)*

Training

- Denoising
- Back-translation

L1 corpus

non-parallel

L2 corpus

Machine Translation

- Neural approach
- Statistical approach

General recipe

Initialization

- Cross-lingual word embeddings *(Artetxe et al., ICLR'18; Lample et al., ICLR'18)*
- Deep pretraining *(Conneau & Lample, NeurIPS'19; Song et al., ICML'19; Liu et al., arXiv'20)*

Training

- Denoising
- Back-translation

L1 corpus

non-parallel

L2 corpus

Machine Translation

- Neural approach
- Statistical approach

General recipe

L1 corpus

non-parallel

L2 corpus

Initialization

- Cross-lingual word embeddings *(Artetxe et al., ICLR'18; Lample et al., ICLR'18)*
- Deep pretraining *(Conneau & Lample, NeurIPS'19; Song et al., ICML'19; Liu et al., arXiv'20)*
  - Pretrain the full network through masked language modeling (e.g., multilingual BERT, XLM, MASS, mBART)

Training

- Denoising
- Back-translation

Machine Translation

- Neural approach
- Statistical approach

General recipe

Initialization

- Cross-lingual word embeddings *(Artetxe et al., ICLR'18; Lample et al., ICLR'18)*
- Deep pretraining *(Conneau & Lample, NeurIPS'19; Song et al., ICML'19; Liu et al., arXiv'20)*
  - Pretrain the full network through masked language modeling (e.g., multilingual BERT, XLM, MASS, mBART)
  - Works well, but very expensive to train

Training

- Denoising
- Back-translation

L1 corpus

non-parallel

L2 corpus

Machine Translation

- Neural approach
- Statistical approach

General recipe

L1 corpus

non-parallel

L2 corpus

Initialization

- Cross-lingual word embeddings *(Artetxe et al., ICLR'18; Lample et al., ICLR'18)*
- Deep pretraining *(Conneau & Lample, NeurIPS'19; Song et al., ICML'19; Liu et al., arXiv'20)*
  - Pretrain the full network through masked language modeling (e.g., multilingual BERT, XLM, MASS, mBART)
  - Works well, but very expensive to train
  - Masking can be seen as a form of noise!

Training

- Denoising
- Back-translation

Machine Translation

- Neural approach
- Statistical approach

General recipe

Initialization

- Cross-lingual word embeddings *(Artetxe et al., ICLR'18; Lample et al., ICLR'18)*
- Deep pretraining *(Conneau & Lample, NeurIPS'19; Song et al., ICML'19; Liu et al., arXiv'20)*
  - Pretrain the full network through masked language modeling (e.g., multilingual BERT, XLM, MASS, mBART)
  - Works well, but very expensive to train
  - Masking can be seen as a form of noise!
- Synthetic parallel corpus *(Marie & Fujita, arXiv'18; Artetxe et al., ACL'19)*

Training

- Denoising
- Back-translation

L1 corpus

non-parallel

L2 corpus

Machine Translation

- Neural approach
- Statistical approach

General recipe

L1 corpus

non-parallel

L2 corpus

Initialization

- Cross-lingual word embeddings *(Artetxe et al., ICLR'18; Lample et al., ICLR'18)*
- Deep pretraining *(Conneau & Lample, NeurIPS'19; Song et al., ICML'19; Liu et al., arXiv'20)*
  - Pretrain the full network through masked language modeling (e.g., multilingual BERT, XLM, MASS, mBART)
  - Works well, but very expensive to train
  - Masking can be seen as a form of noise!
- Synthetic parallel corpus *(Marie & Fujita, arXiv'18; Artetxe et al., ACL'19)*
  - <u>Basic</u>: word-for-word translation using cross-lingual word embeddings

Training

- Denoising
- Back-translation

Machine Translation

- Neural approach
- Statistical approach

General recipe

L1 corpus

non-parallel

L2 corpus

Initialization

- Cross-lingual word embeddings *(Artetxe et al., ICLR'18; Lample et al., ICLR'18)*
- Deep pretraining *(Conneau & Lample, NeurIPS'19; Song et al., ICML'19; Liu et al., arXiv'20)*
  - Pretrain the full network through masked language modeling (e.g., multilingual BERT, XLM, MASS, mBART)
  - Works well, but very expensive to train
  - Masking can be seen as a form of noise!
- Synthetic parallel corpus *(Marie & Fujita, arXiv'18; Artetxe et al., ACL'19)*
  - Basic: word-for-word translation using cross-lingual word embeddings
  - Better: Unsupervised statistical machine translation

Training

- Denoising
- Back-translation

Machine Translation

- Neural approach
- Statistical approach

Results

Results

- <u>Languages</u>: French-English, German-English

Results

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

Results

- Languages: French-English, German-English

- Training: WMT-14 News Crawl

- Test set: WMT-14 newstest (BLEU)

Results

- Languages: French-English, German-English

- Training: WMT-14 News Crawl

- Test set: WMT-14 newstest (BLEU)

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

Results

- Languages: French-English, German-English

- Training: WMT-14 News Crawl

- Test set: WMT-14 newstest (BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Cross-lingual embs. (Artetxe et al., ICLR'18)*	15.6	15.1	10.2	6.6

*Tokenized BLEU (about 1-2 points higher)

Results

- Languages: French-English, German-English

- Training: WMT-14 News Crawl

- Test set: WMT-14 newstest (BLEU)

	FR-EN	EN-FR	DE-EN	EN-DE
Cross-lingual embs. (Artetxe et al., ICLR'18)*	15.6	15.1	10.2	6.6
+ scaling up (Conneau & Lample, NeurIPS'19)*	29.4	29.4	-	-

*Tokenized BLEU (about 1-2 points higher)

Results

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

	FR-EN	EN-FR	DE-EN	EN-DE
Cross-lingual embs. (Artetxe et al., ICLR'18)*	15.6	15.1	10.2	6.6
+ scaling up (Conneau & Lample, NeurIPS'19)*	29.4	29.4	-	-
Deep pre-training (Conneau & Lample, NeurIPS'19)*	33.3	33.4	-	-

*Tokenized BLEU (about 1-2 points higher)

Results

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

	FR-EN	EN-FR	DE-EN	EN-DE
Cross-lingual embs. (Artetxe et al., ICLR'18)*	15.6	15.1	10.2	6.6
+ scaling up (Conneau & Lample, NeurIPS'19)*	29.4	29.4	-	-
Deep pre-training (Conneau & Lample, NeurIPS'19)*	33.3	33.4	-	-
Unsup SMT + NMT (Artetxe et al., ACL'19)*	33.5	36.2	27.0	22.5

*Tokenized BLEU (about 1-2 points higher)

Results

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

	FR-EN	EN-FR	DE-EN	EN-DE
Cross-lingual embs. (Artetxe et al., ICLR'18)*	15.6	15.1	10.2	6.6
+ scaling up (Conneau & Lample, NeurIPS'19)*	29.4	29.4	-	-
Deep pre-training (Conneau & Lample, NeurIPS'19)*	33.3	33.4	-	-
Unsup SMT + NMT (Artetxe et al., ACL'19)*	33.5	36.2	27.0	22.5
detok. SacreBLEU	33.2	33.6	26.4	21.2

*Tokenized BLEU (about 1-2 points higher)

Results

- 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.	Cross-lingual embs. (Artetxe et al., ICLR'18)*	15.6	15.1	10.2	6.6
	+ scaling up (Conneau & Lample, NeurIPS'19)*	29.4	29.4	-	-
	Deep pre-training (Conneau & Lample, NeurIPS'19)*	33.3	33.4	-	-
	Unsup SMT + NMT (Artetxe et al., ACL'19)*	33.5	36.2	27.0	22.5
	detok. SacreBLEU	33.2	33.6	26.4	21.2
Supervised					

*Tokenized BLEU (about 1-2 points higher)

Results

- 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.	Cross-lingual embs. (Artetxe et al., ICLR'18)*	15.6	15.1	10.2	6.6
	+ scaling up (Conneau & Lample, NeurIPS'19)*	29.4	29.4	-	-
	Deep pre-training (Conneau & Lample, NeurIPS'19)*	33.3	33.4	-	-
	Unsup SMT + NMT (Artetxe et al., ACL'19)*	33.5	36.2	27.0	22.5
	detok. SacreBLEU	33.2	33.6	26.4	21.2
Supervised	WMT winner	35.0	35.8	29.0	20.6

*Tokenized BLEU (about 1-2 points higher)

Results

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

		FR-EN	EN-FR	DE-EN	EN-DE
Unsup.	Cross-lingual embs. (Artetxe et al., ICLR'18)*	15.6	15.1	10.2	6.6
	+ scaling up (Conneau & Lample, NeurIPS'19)*	29.4	29.4	-	-
	Deep pre-training (Conneau & Lample, NeurIPS'19)*	33.3	33.4	-	-
	Unsup SMT + NMT (Artetxe et al., ACL'19)*	33.5	36.2	27.0	22.5
	detok. SacreBLEU	33.2	33.6	26.4	21.2
Supervised	WMT winner	35.0	35.8	29.0	20.6
	Original transformer (Vaswani et al., NIPS'17)*	-	41.0	-	28.4

*Tokenized BLEU (about 1-2 points higher)

Results

- 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.	Cross-lingual embs. (Artetxe et al., ICLR'18)*	15.6	15.1	10.2	6.6
	+ scaling up (Conneau & Lample, NeurIPS'19)*	29.4	29.4	-	-
	Deep pre-training (Conneau & Lample, NeurIPS'19)*	33.3	33.4	-	-
	Unsup SMT + NMT (Artetxe et al., ACL'19)*	33.5	36.2	27.0	22.5
	detok. SacreBLEU	33.2	33.6	26.4	21.2
Supervised	WMT winner	35.0	35.8	29.0	20.6
	Original transformer (Vaswani et al., NIPS'17)*	-	41.0	-	28.4
	Large scale back-translation (Edunov et al., EMNLP'18)	-	43.8	-	33.8

*Tokenized BLEU (about 1-2 points higher)

Conclusions

Conclusions

Unsupervised machine translation is a reality!

Conclusions

Unsupervised machine translation is a reality!

- General recipe: back-translation + smart initialization

Conclusions

Unsupervised machine translation is a reality!

- General recipe: back-translation + smart initialization
    - Align monolingual representations

Conclusions

Unsupervised machine translation is a reality!

- General recipe: back-translation + smart initialization
    - Align monolingual representations
    - Generalize from word level to sentence level translation

Conclusions

Unsupervised machine translation is a reality!

- General recipe: back-translation + smart initialization
  - Align monolingual representations
  - Generalize from word level to sentence level translation

Strong results

Conclusions

Unsupervised machine translation is a reality!

- General recipe: back-translation + smart initialization
  - Align monolingual representations
  - Generalize from word level to sentence level translation

Strong results

- Competitive with supervised SOTA from 6 years ago

Conclusions

Unsupervised machine translation is a reality!

- General recipe: back-translation + smart initialization
  - Align monolingual representations
  - Generalize from word level to sentence level translation

Strong results

- Competitive with supervised SOTA from 6 years ago
- …in just 2-3 years!

Conclusions

Unsupervised machine translation is a reality!

- General recipe: back-translation + smart initialization
  - Align monolingual representations
  - Generalize from word level to sentence level translation

Strong results

- Competitive with supervised SOTA from 6 years ago
- …in just 2-3 years!

What's next?

Thank you!

Twitter: @artetxem
Email: artetxe@fb.com