Multilingual NLP?

Speaking more languages means communicating with more people…
...and reaching more users and customers...

https://tinyurl.com/xlingual

Why Cross-Lingual NLP?

...but there are **more profound** and **democratic** reasons to work in this area:

- decreasing **the digital divide**
- dealing with **inequality of information**
- mitigating **cross-cultural biases**
- deploying language technology for **underrepresented** languages, dialects, minorities; societal impact
- understanding cross-linguistic differences

"95% of all languages in use today will never gain traction online" (Andras Kornai)

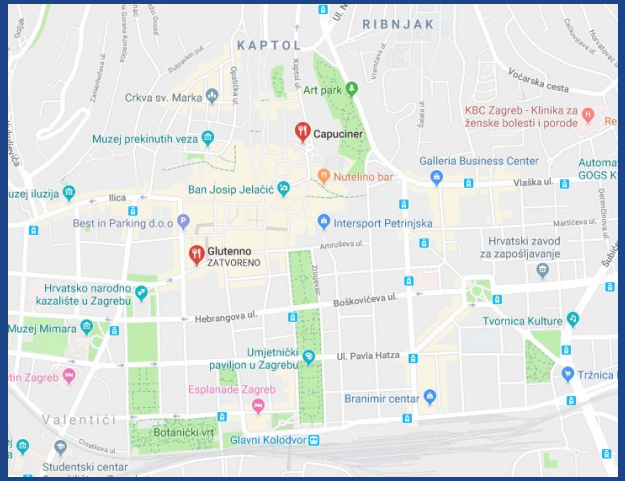
"The limits of my language *online* mean the limits of my world?"

Source: http://labs.theguardian.com/digital-language-divide/

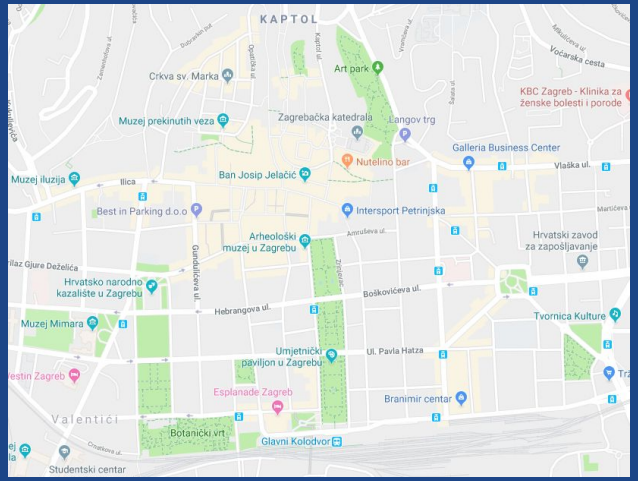
Why Cross-Lingual NLP?

Inequality of information and representation can also affect how we understand places, events, processes...

We're in Zagreb searching for...



...éttermek (HU)



...jatetxe (EU)

...restaurants (EN)

Motivation: Cross-Lingual Representations are Everywhere

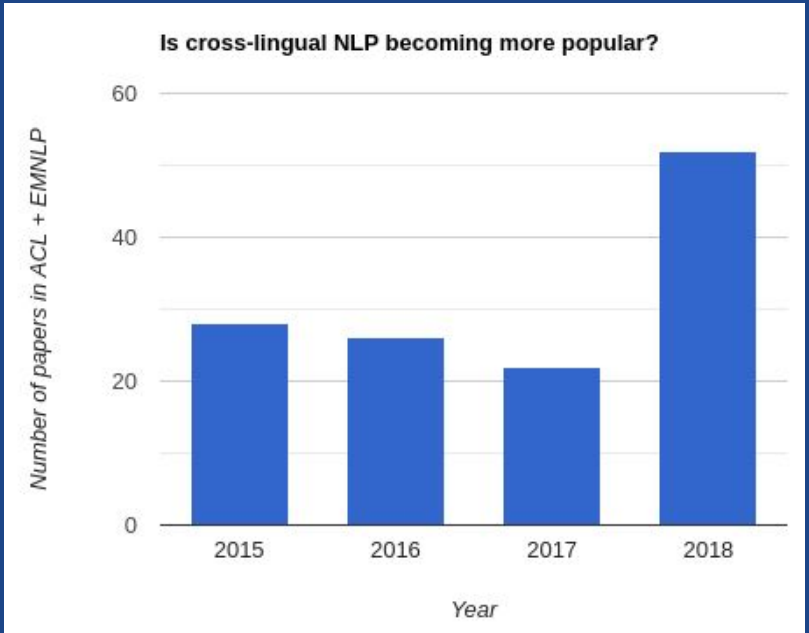
- Cross-lingual and multilingual semantic similarity
- Bilingual lexicon induction, multi-modal representations

- Cross-lingual SRL, POS tagging, NER
- Cross-lingual dependency parsing, sentiment analysis
- Cross-lingual natural language understanding for dialogue
- Cross-lingual lexical entailment
- Cross-lingual annotation and model transfer
- Cross-lingual *you-name-it-task*

- Statistical and neural MT
- Cross-lingual IR and QA

Motivation: Cross-Lingual Representations are Everywhere

Searching for “multilingual”, “cross-lingual” and “bilingual” in the ACL anthology (ACL+EMNLP)



- 10+ papers on unsupervised cross-lingual word embeddings at EMNLP 2018

- The trend continues:
  - 20+ papers on cross-lingual learning and applications at NAACL 2019.

Motivation (Very High-Level)

We want to understand and model the meaning of...

Source: dreamstime.com

...without manual/human input and without perfect MT

So, Why (Unsupervised) Cross-Lingual Embeddings Exactly?

Cross-lingual word embeddings (CLWE-s)

- Simple: quick and efficient to train
- (Still) state-of-the-art in cross-lingual NLP; omnipresent
- Lightweight and inexpensive
- Multilingual modeling of meaning and support for cross-lingual NLP

Unsupervised CLWE-s:

- Wide portability without bilingual resources?
- Deploying language technology for virtually any language?
- Increasing the ability of cross-lingual transfer?
- An interesting scientific problem still at its infancy:
  Potential for transforming cross-lingual and cross-domain NLP

- **What is the current state-of-the-art in unsupervised cross-lingual representation learning?**

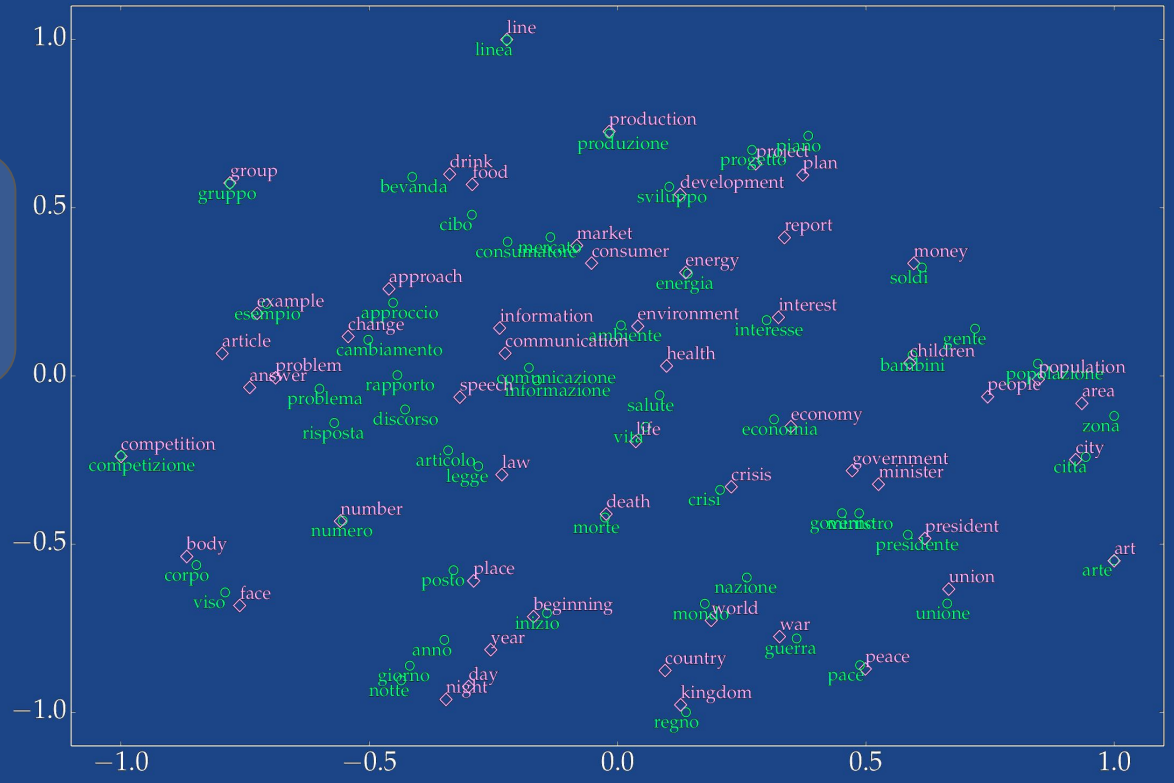
Motivation: Crossing the Lexical Chasm

Multilingual / Cross-lingual representation of meaning

- Word-level
  - Cross-lingual word embeddings
  - Words with similar meanings across languages have similar vectors

- Sentence-/paragraph-level
  - Most recent developments
  - Multilingual unsupervised pretraining [Conneau and Lample, arXiv-19]

- (Unsupervised) NMT?



Cross-Lingual Word Embeddings

Representation of a word $w_1^S \in V^S$:

$$vec(w_1^S) = [f_1^1, f_2^1, \ldots, f_{dim}^1]$$

Exactly the same representation for $w_2^T \in V^T$:

$$vec(w_2^T) = [f_1^2, f_2^2, \ldots, f_{dim}^2]$$

Language-independent word representations in the same shared semantic (or *embedding*) space!

Why Cross-Lingual Word Representations?

Capturing meaning across languages: a standard task of **bilingual lexicon induction (BLI)**

en_morning			en_carpet		
Slavic+EN	**Germanic**	**Romance+EN**	**Slavic+EN**	**Germanic**	**Romance+EN**
en_daybreak	de_vormittag	pt_madrugada	en_rug	de_teppichboden	en_rug
en_morn	nl_krieken	it_mattina	bg_килим	nl_tapijten	it_moquette
bg_разсъмване	en_dawn	en_dawn	ru_ковролин	en_rug	it_tappeti
hr_svitanje	nl_zonsopkomst	pt_madrugadas	bg_килими	de_teppich	pt_tapete
hr_zore	sv_morgonen	es_madrugada	pl_dywany	en_carpeting	es_moqueta
bg_изгрев	de_tagesanbruch	it_nascente	bg_мокет	de_teppiche	it_tappetino
en_dawn	en_sunrise	en_morn	pl_dywanów	sv_mattor	en_carpeting
ru_утро	nl_opgang	es_aurora	hr_tepih	sv_matta	pt_carpete
bg_аврора	de_sonnenaufgang	fr_matin	pl_wykładziny	en_carpets	pt_tapetes
hr_jutro	nl_dageraad	fr_aurora	ru_ковер	nl_tapijt	fr_moquette
ru_рассвет	de_anbruch	es_amaneceres	ru_коврик	nl_kleedje	en_carpets
hr_zora	sv_morgon	en_sunrises	hr_ćilim	nl_vloerbedekking	es_alfombra
hr_zoru	en_daybreak	es_mañanero	en_carpeting	de_brücke	es_alfombras
pl_poranek	de_morgengrauen	fr_matinée	pl_dywan	de_matta	fr_tapis
en_sunrise	nl_zonsopgang	it_mattinata	ru_ковров	nl_matta	pt_tapeçaria
bg_зазоряване	nl_goedemorgen	pt_amanhecer	en_carpets	en_mat	it_zerbino

Retrieving nearest neighbours from a shared cross-lingual embedding space (P@1, MRR, MAP)

Unsupervised MT

- Recently: **unsupervised neural and statistical machine translation**

[Artetxe et al., ICLR-18, EMNLP-18, ACL-19; Lample et al., ICLR-18, EMNLP-18; Wu et al., NAACL-19;...]

Key component: initialization via unsupervised cross-lingual word embeddings

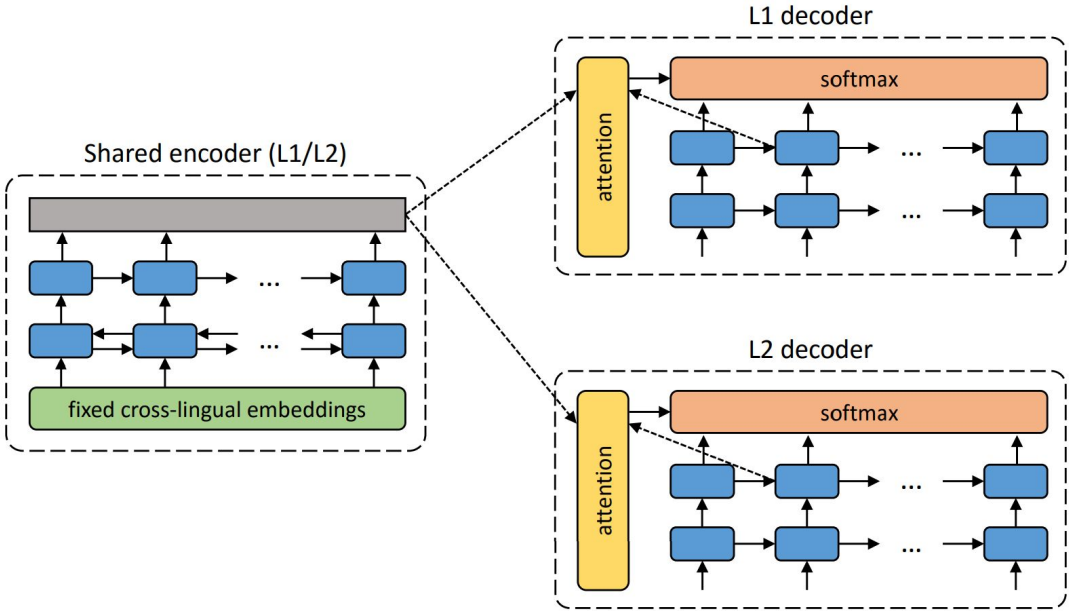
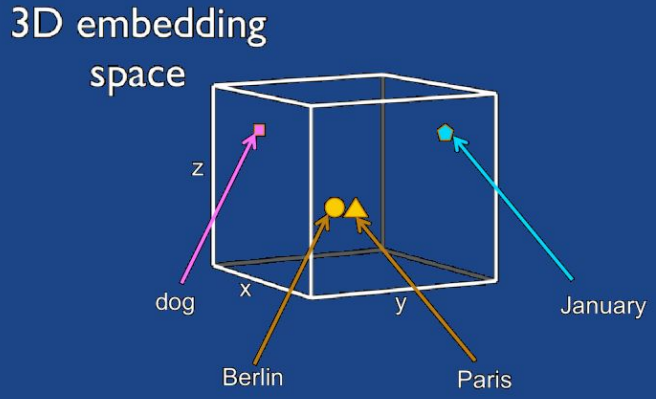
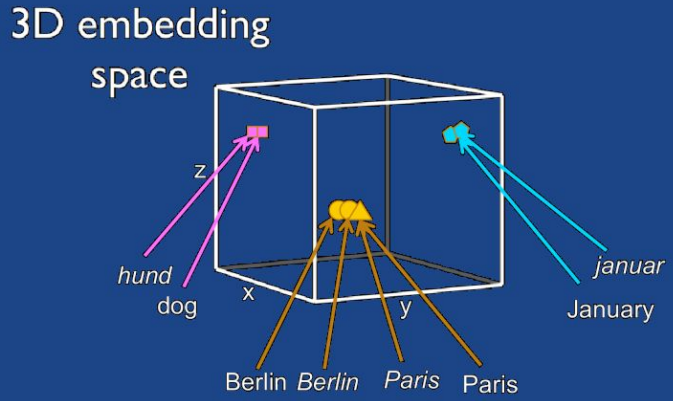


Image from [Artetxe et al., ICLR-18]

Cross-Lingual Word Embeddings





Monolingual vs. **Cross-lingual**

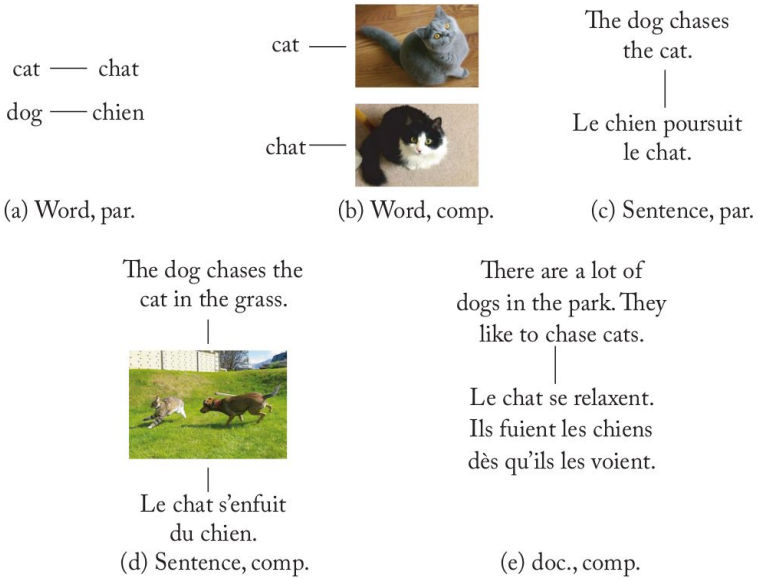
Q1 → Algorithm Design: How to align semantic spaces in two different languages?

Q2 → Data Requirements: Which **bilingual signals** are used for the alignment?

Cross-Lingual Word Embeddings

A large number of different methods, but **the same end goal**:

Induce a shared semantic vector space in which words with similar meaning end up with similar vectors, regardless of their actual language.



We need some bilingual supervision to learn CLWE-s.

Fully unsupervised CLWE-s: they rely only on monolingual data

Projection-based CLWE Learning

Most models learn a single projection matrix $\mathbf{W_{L1}}$ (i..e., $\mathbf{W_{L2} = I}$), but **bidirectional learning** is also common.

$$\begin{array}{r} \\ \text{bird} \\ \text{pretty} \\ \dots \\ \text{eat} \end{array} \begin{bmatrix} -1.18 & 0.21 & \dots & 0.11 \\ 0.23 & -0.53 & \dots & 0.34 \\ \dots & \dots & \dots & \dots \\ 0.78 & 1.33 & \dots & -0.47 \end{bmatrix} \mathbf{X_S} \quad \mathbf{W_{L1}} = \quad \mathbf{X_T} \begin{bmatrix} 0.59 & 1.01 & \dots & 0.37 \\ -0.34 & -0.27 & \dots & 0.41 \\ \dots & \dots & \dots & \dots \\ 0.81 & -0.31 & \dots & 0.29 \end{bmatrix} \begin{array}{l} \\ \text{Vogel} \\ \text{schön} \\ \dots \\ \text{essen} \end{array}$$

How do we find the "optimal" projection matrix $\mathbf{W_{L1}}$?

- **Mean square error:** [Mikolov et al., arXiv-13] and most follow-up work
  ...except…
- **Canonical methods** [Faruqui et al., EACL-14; Lu et al., NAACL-15; Rotman et al., ACL-18]
- **Max-margin framework:** [Lazaridou et al., ACL-15; Mrkšić et al., TACL-17]
- **Relaxed Cross-Domain Similarity Local Scaling**: [Joulin et al., EMNLP-18]

Minimising Euclidean Distance

[Mikolov et al., arXiv-13] minimize the Euclidean distances for translation pairs after projection

$$\mathbf{W}_{L1} = \arg\min_{\mathbf{W}} \| \mathbf{X}_S \mathbf{W} - \mathbf{X}_T \|_2$$

The optimisation problem has no closed-form solution

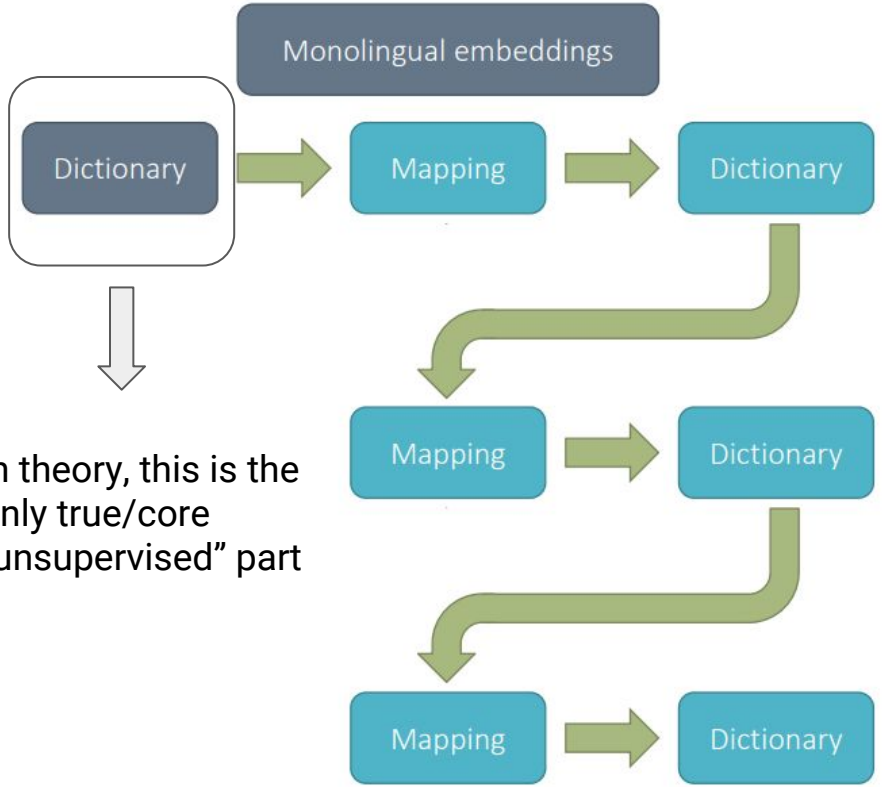
- Iterative SGD-based optimisation was used initially

More complex mappings: e.g., non-linear DFFNs instead of linear projection matrix yield worse performance

Better (word translation) results when $\mathbf{W}_{L1}$ is constrained to be **orthogonal**

- This preserves monolingual vector space topology

Self-Learning in a Nutshell



In theory, this is the only true/core “unsupervised” part

- The seed dictionary improves over time, but…
- How do we start?
- How do we choose new candidates?
- How do we guarantee that we do not introduce noise?
- How does the method converge?

Image courtesy of Eneko Agirre

Self-Learning in a Nutshell

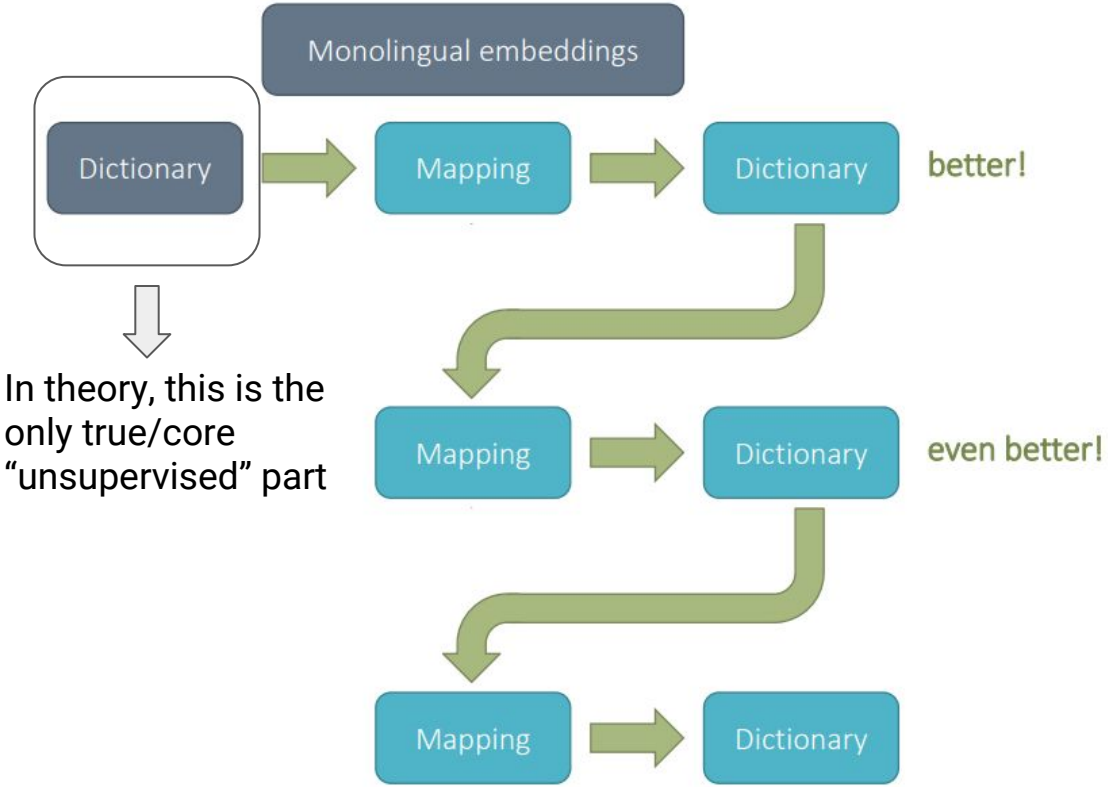
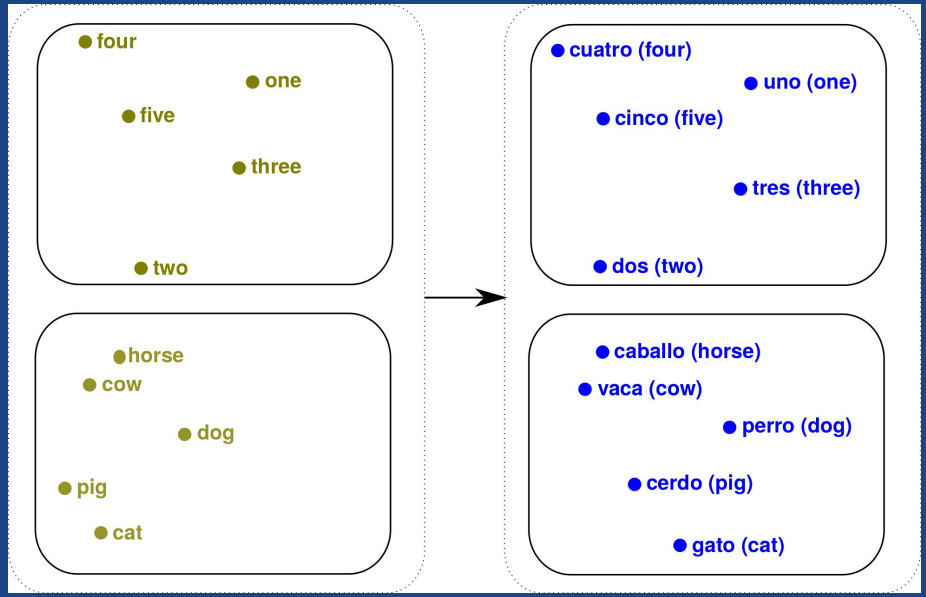


Image courtesy of Eneko Agirre

(Approximate) Isomorphism

“. . . we hypothesize that, if languages are used to convey thematically similar information in similar contexts, these random processes should be approximately isomorphic between languages, and that this isomorphism can be learned from the statistics of the realizations of these processes, the monolingual corpora, in principle without any form of explicit alignment.”

[Miceli Barone, RepL4NLP-16]



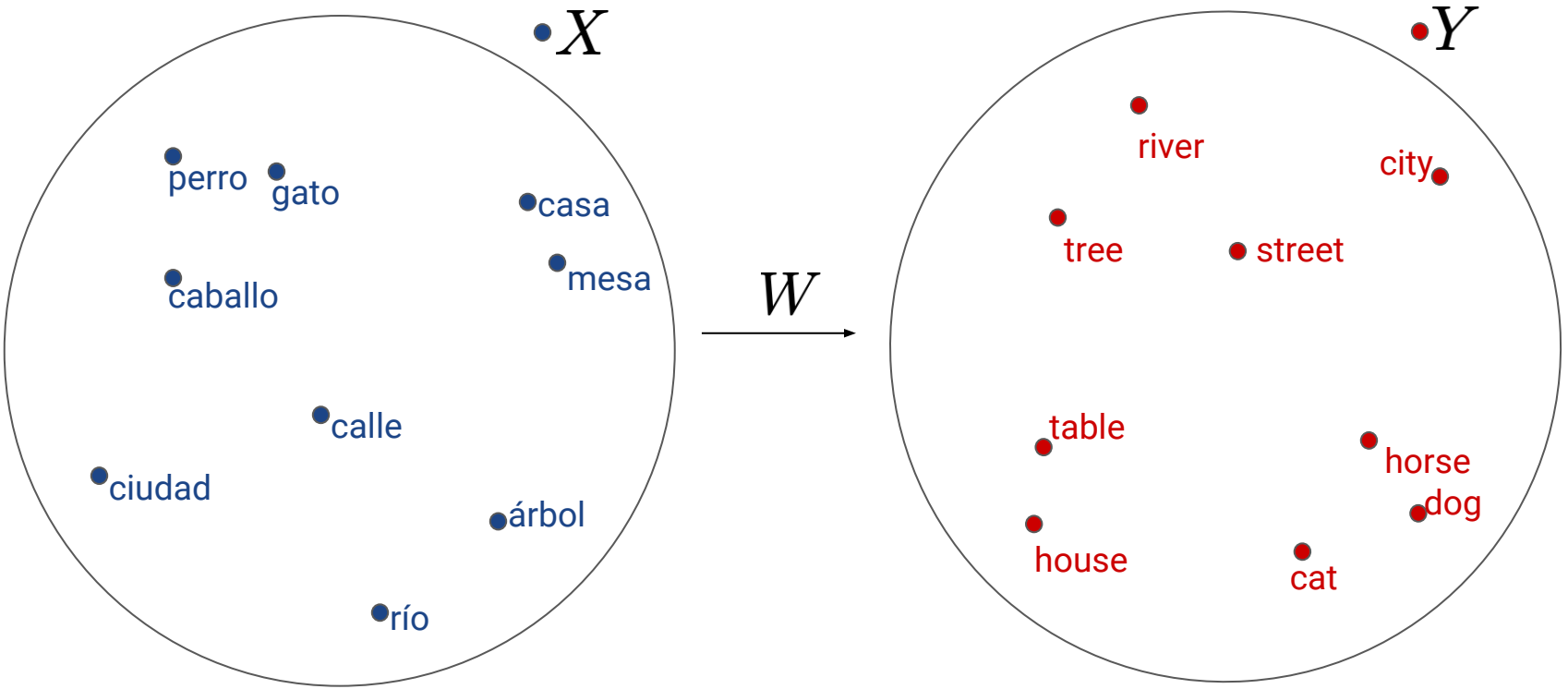
[Mikolov et al., arXiv-13]

Why should this work at all?

Main assumption:
Embedding spaces in different languages have similar structure so that the same transformation can align source language words with target language words.

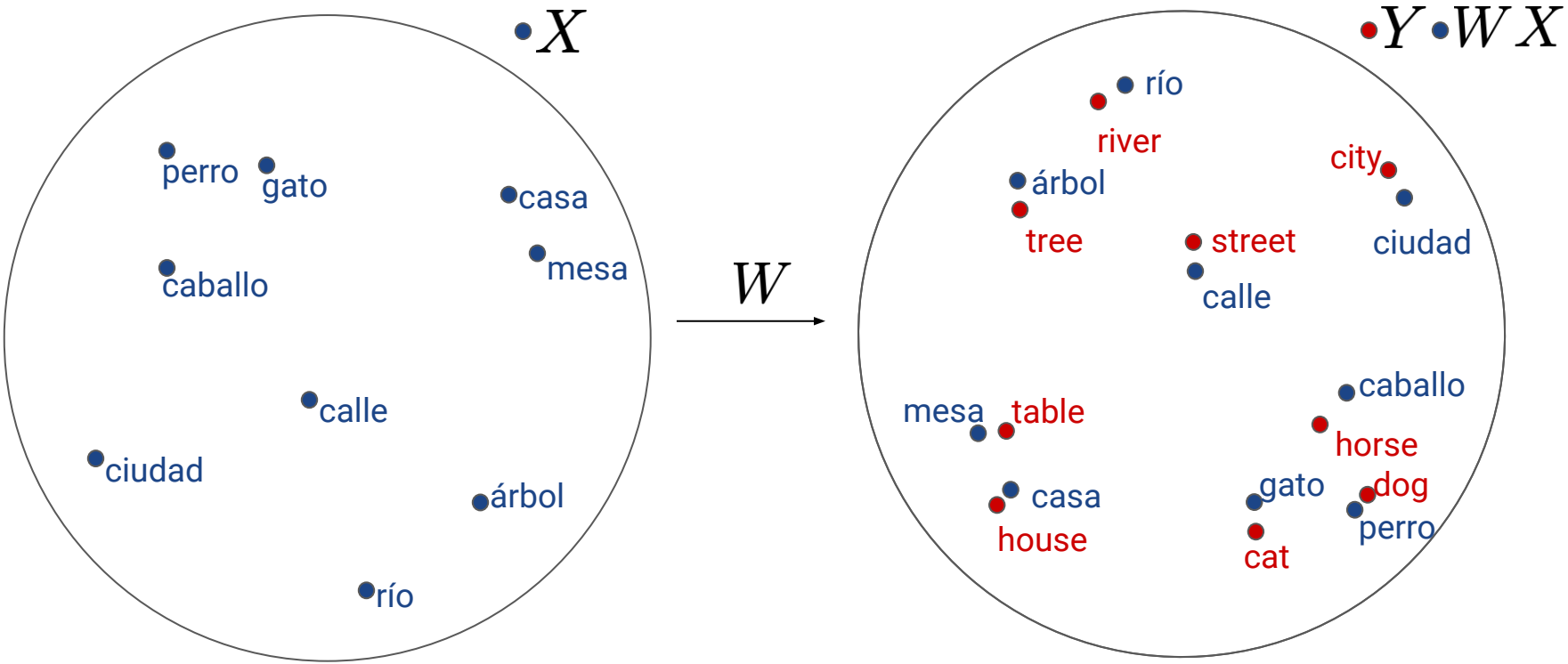
More specifically:
Embeddings spaces should be approximately isomorphic.

Supervised alignment



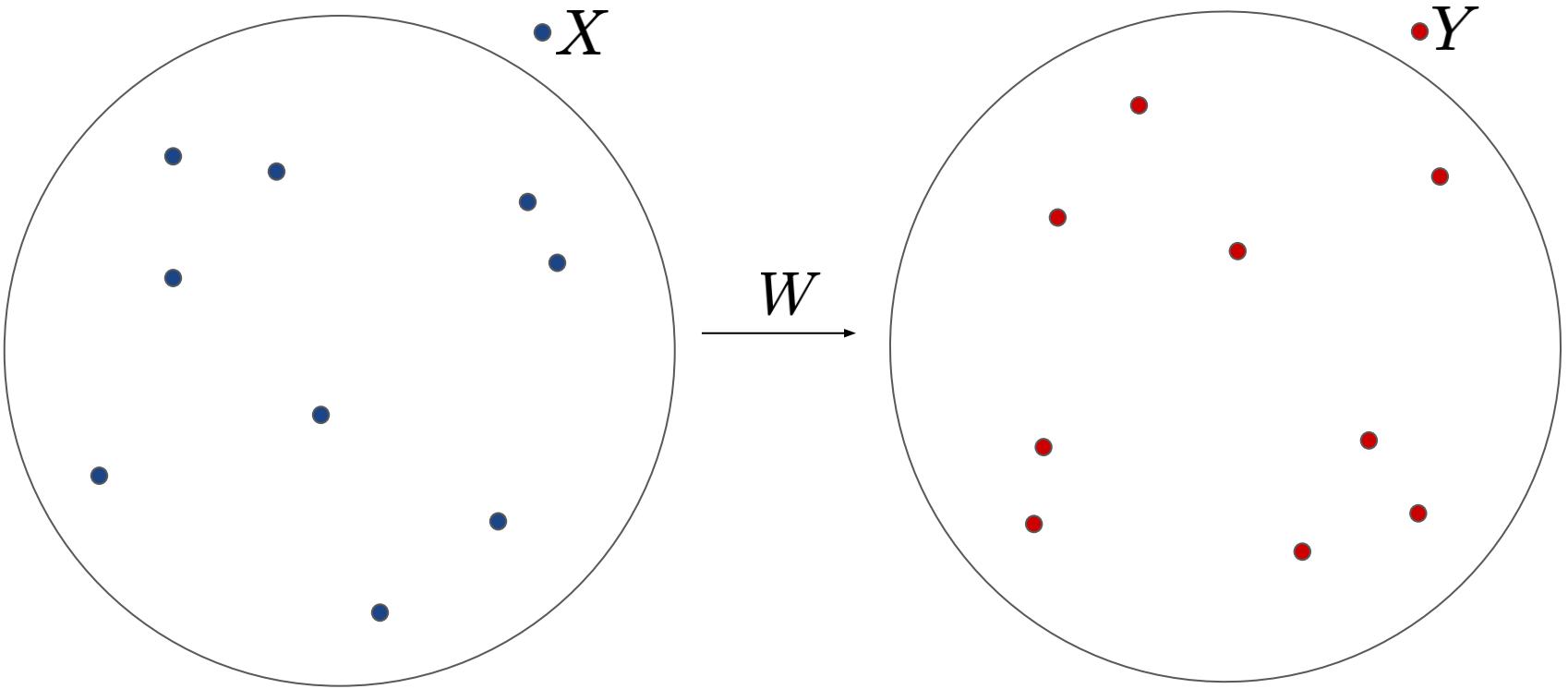
Inspired by: Slides from Eneko Agirre, Mikel Artetxe

Supervised alignment



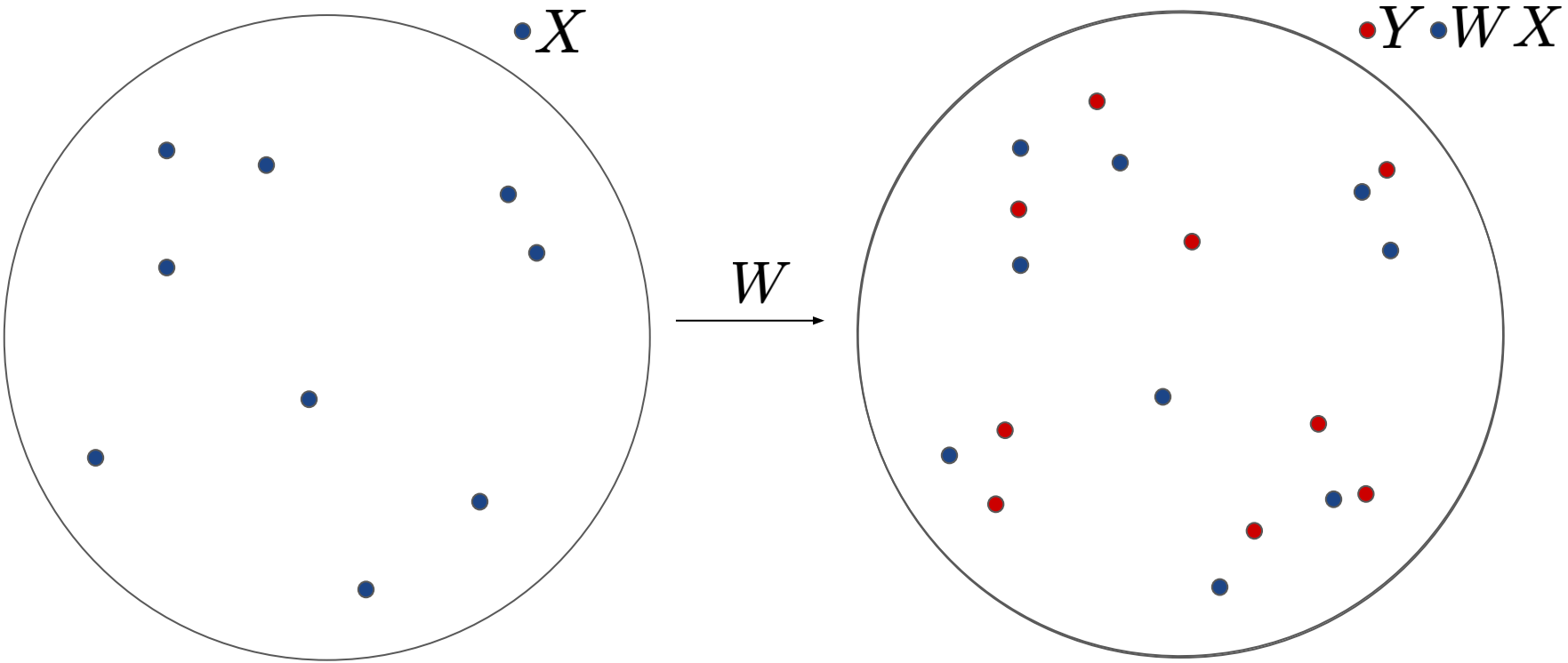
Inspired by: Slides from Eneko Agirre, Mikel Artetxe

Unsupervised alignment



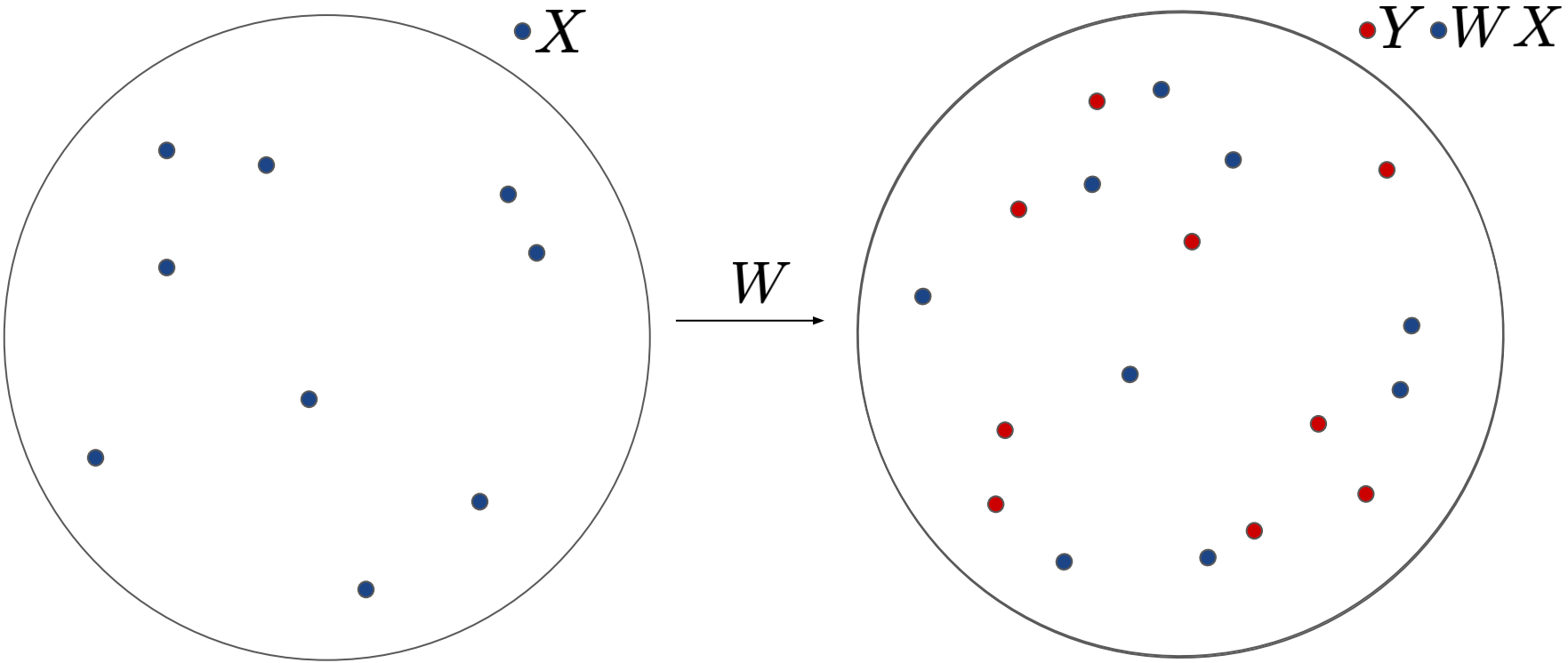
Inspired by: Slides from Eneko Agirre, Mikel Artetxe

Unsupervised alignment



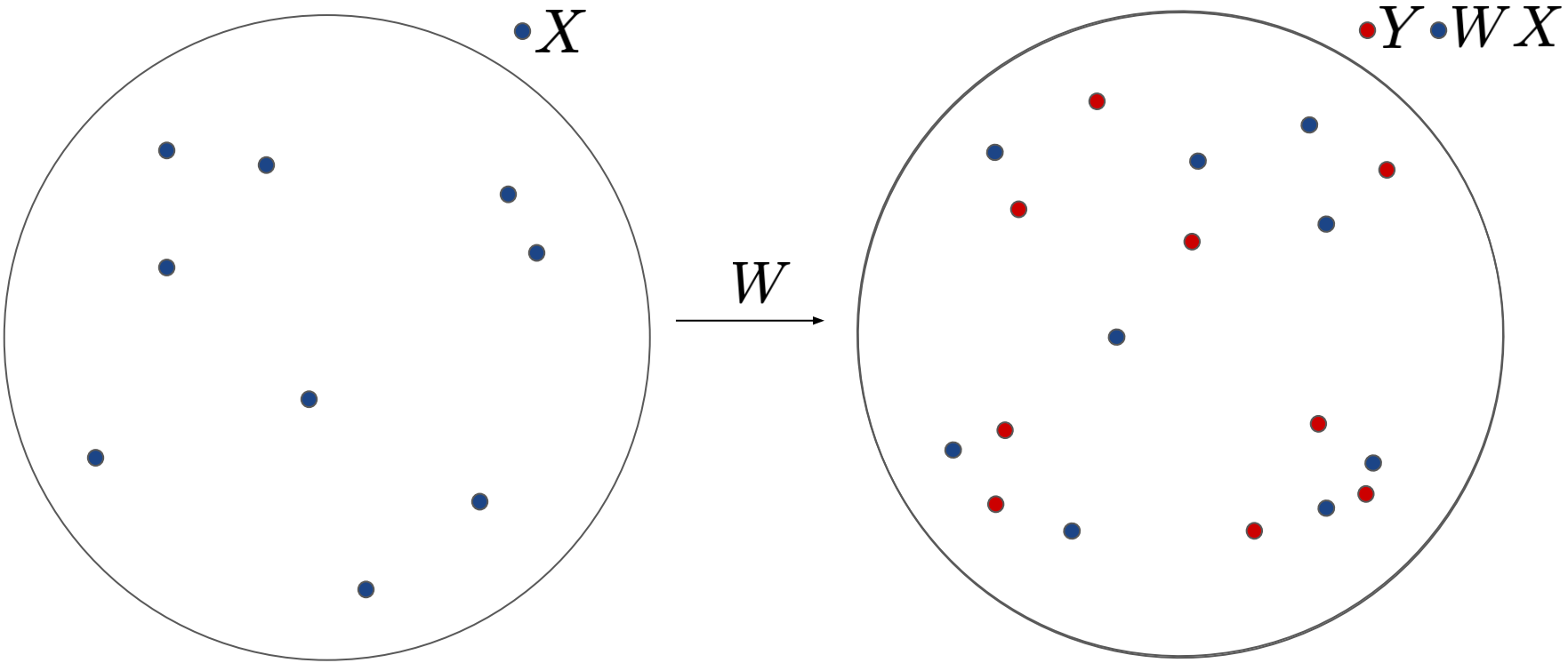
Inspired by: Slides from Eneko Agirre, Mikel Artetxe

Unsupervised alignment



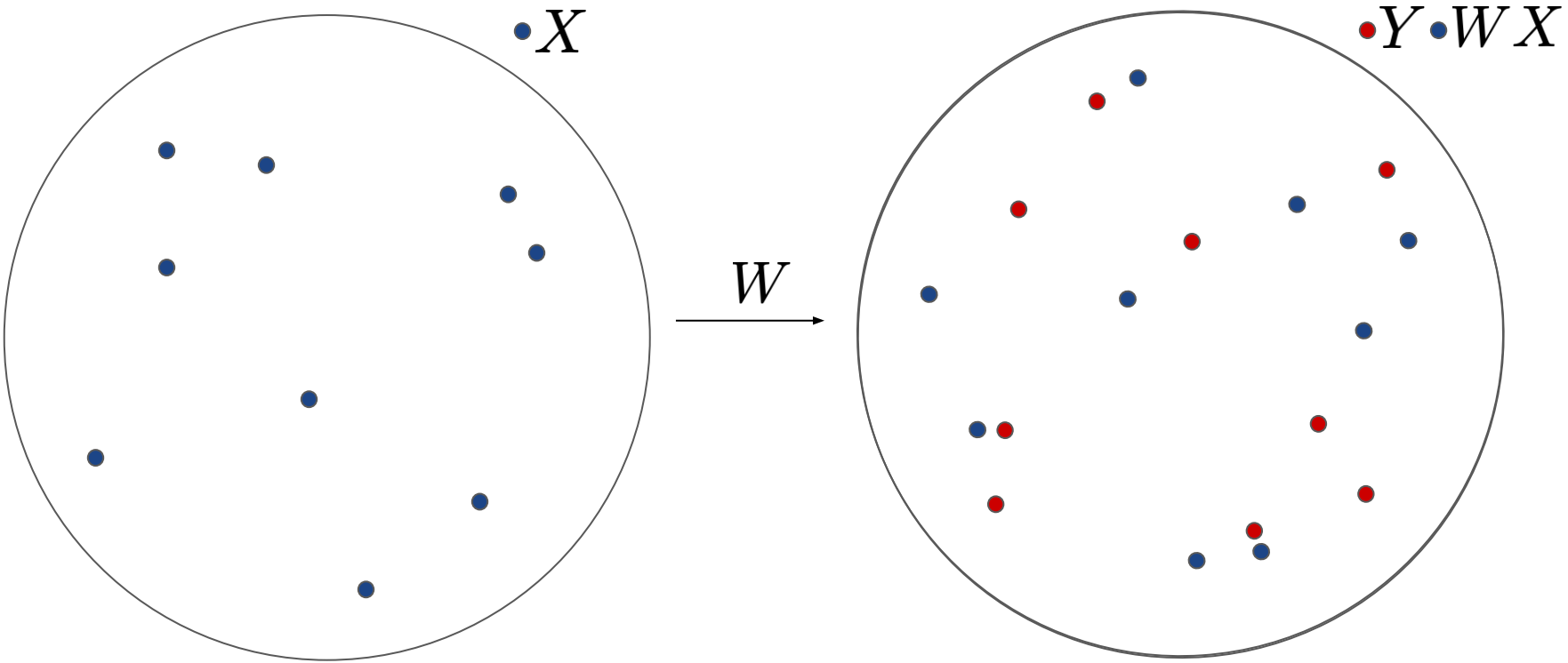
Inspired by: Slides from Eneko Agirre, Mikel Artetxe

Unsupervised alignment



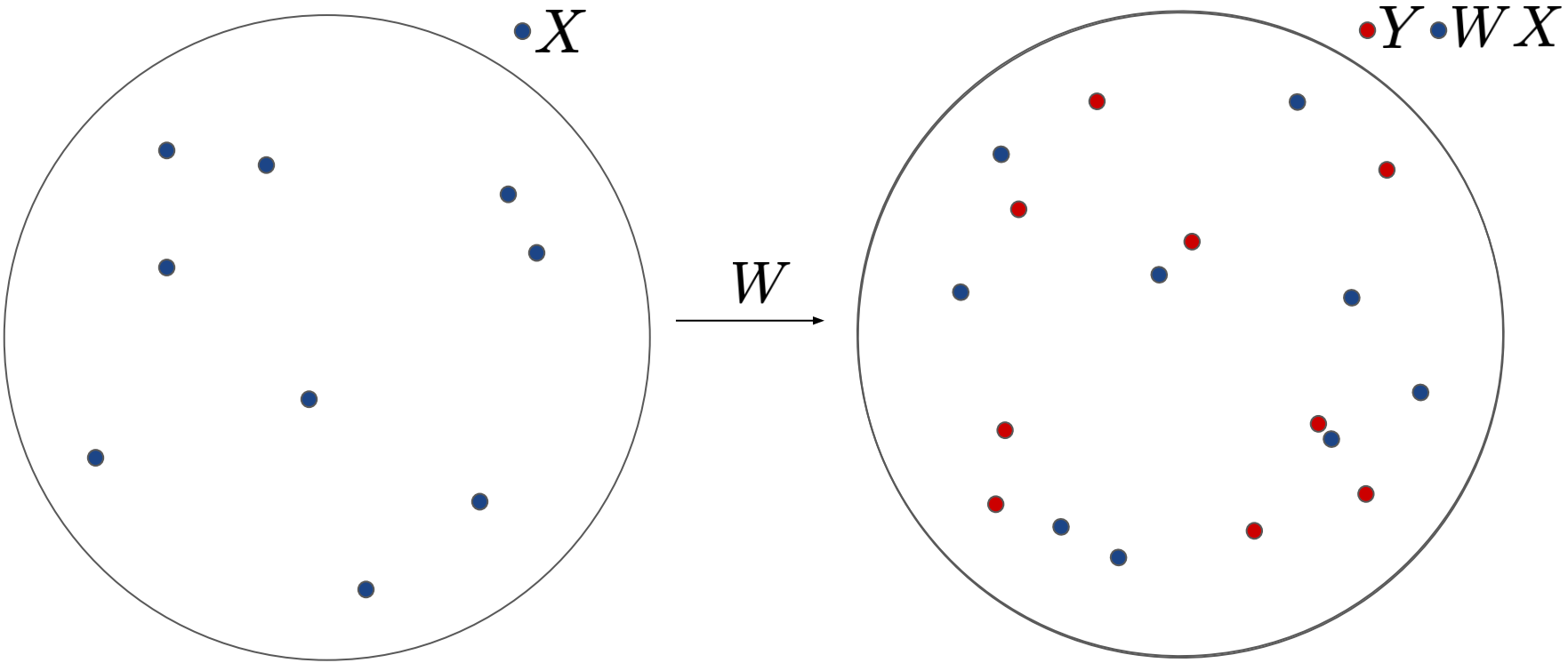
Inspired by: Slides from Eneko Agirre, Mikel Artetxe

Unsupervised alignment



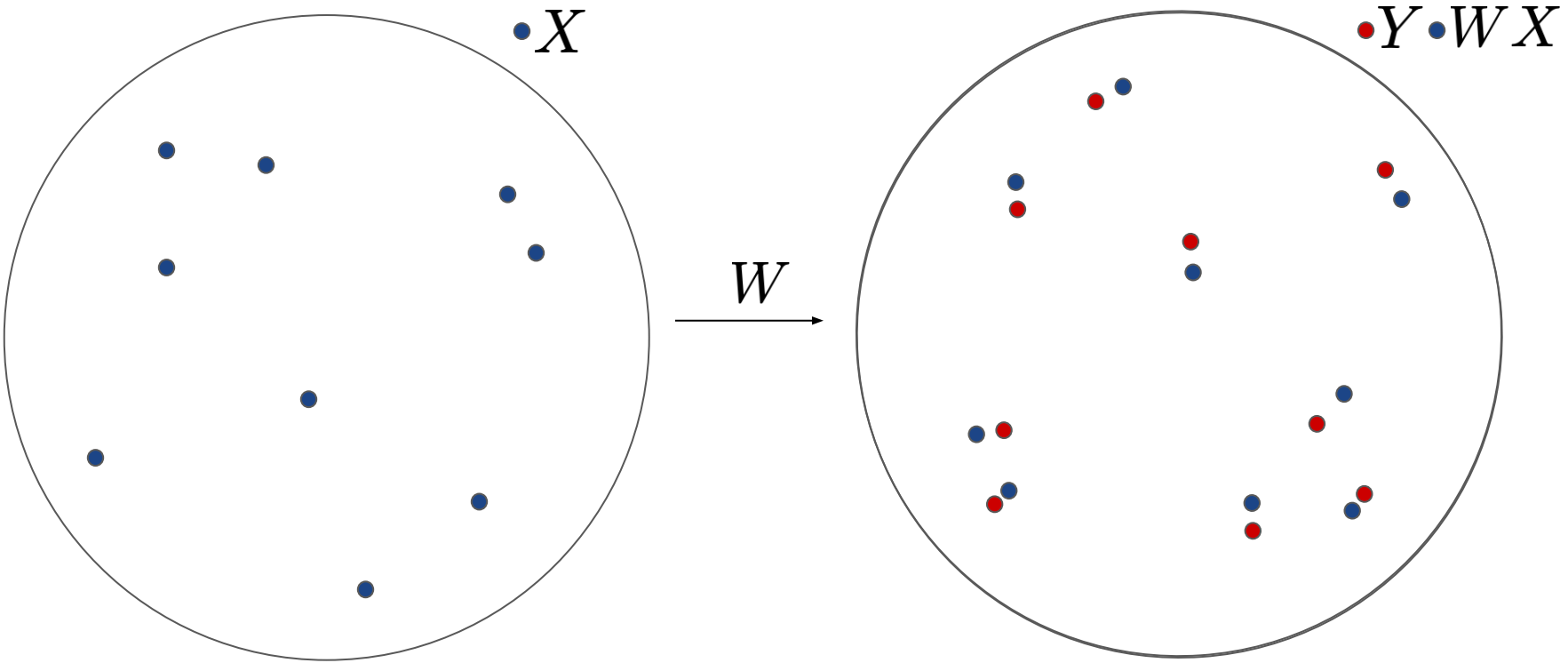
Inspired by: Slides from Eneko Agirre, Mikel Artetxe

Unsupervised alignment



Inspired by: Slides from Eneko Agirre, Mikel Artetxe

Unsupervised alignment

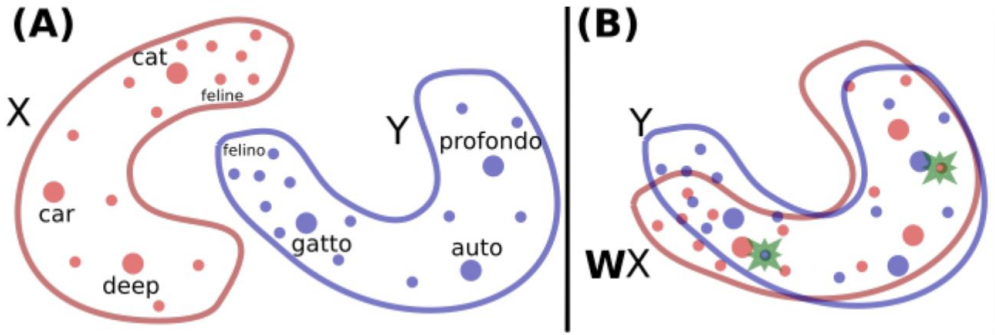


Inspired by: Slides from Eneko Agirre, Mikel Artetxe

Conneau et al. (2018)

1. **Monolingual word embeddings:**
   Learn monolingual vector spaces X and Y.

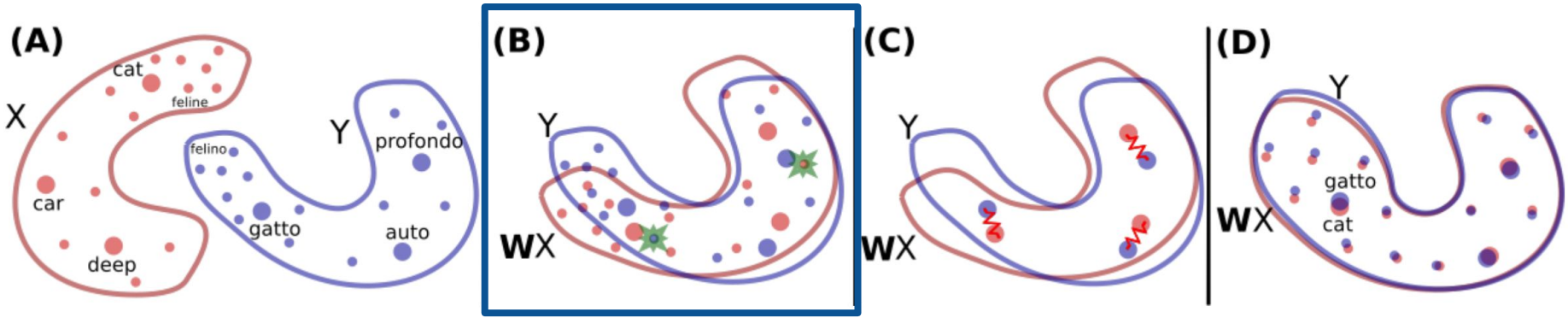
2. **Adversarial mapping:**
   Learn a translation matrix W. Train discriminator to discriminate samples from WX and Y.



Conneau et al. (2018)

3. **Refinement (Procrustes analysis):**
   Build bilingual dictionary of frequent words using W. Learn a new W based on frequent word pairs.

4. **Cross-domain similarity local scaling (CSLS):**
   Use similarity measure that increases similarity of isolated word vectors, decreases similarity of vectors in dense areas.



Systematic comparisons of unsupervised vs. supervised

Unsupervised vs. supervised

[Conneau et al.; ICLR 2018]: *"Without using any character information, our model even outperforms existing supervised methods on cross-lingual tasks for some language pairs"*

[Artetxe et al.; ACL 2018]: *"Our method succeeds in all tested scenarios and obtains the best published results in standard datasets, even surpassing previous supervised systems"*

[Hoshen and Wolf; EMNLP 2018]: *"...our method achieves better performance than recent state-of-the-art deep adversarial approaches and is competitive with the supervised baseline"*

[Xu et al.; EMNLP 2018]: *"Our evaluation (...) shows stronger or competitive performance of the proposed method compared to other state-of-the-art supervised and unsupervised methods..."*

[Chen and Cardie; EMNLP 2018]: *"In addition, our model even beats supervised approaches trained with cross-lingual resources."*

Unsupervised vs. supervised

- How come **unsupervised** is reportedly better than **supervised**?

Unsupervised vs. supervised

- How come **unsupervised** is reportedly better than **supervised**?
- *Argument 1:* Supervision is poor quality.

Unsupervised vs. supervised

- How come **unsupervised** is reportedly better than **supervised**?
- *Argument 1:* Supervision is poor quality. *Counter-argument:* We evaluate on the same data. *Possible counter-counter-argument:* Maybe the train splits are particularly poor?

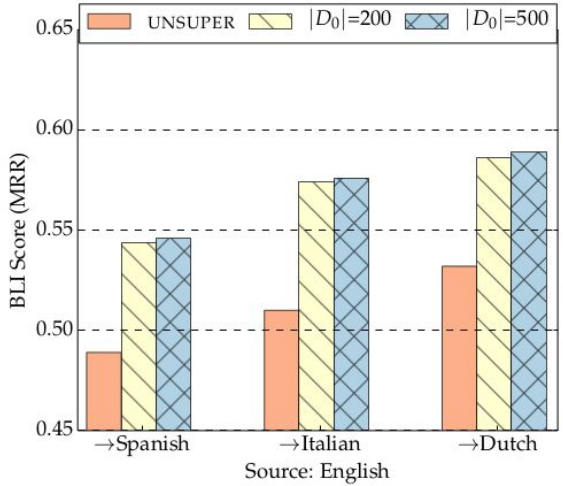
Unsupervised vs. supervised

{Bulgarian, Catalan, Esperanto, Estonian, Basque, Finnish, Hebrew, Hungarian, Indonesian, Georgian, Korean, Lithuanian, Bokmål, Thai, Turkish}

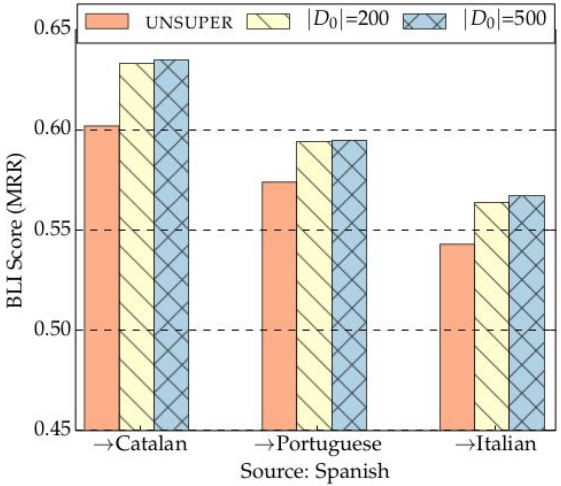
x

{Bulgarian, Catalan, Esperanto, Estonian, Basque, Finnish, Hebrew, Hungarian, Indonesian, Georgian, Korean, Lithuanian, Bokmål, Thai, Turkish}

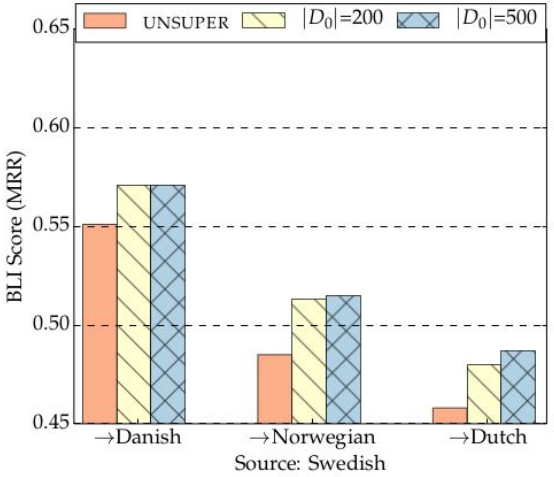
Unsupervised vs. supervised: similar languages



(a) English→ L_2



(b) Spanish→ L_2

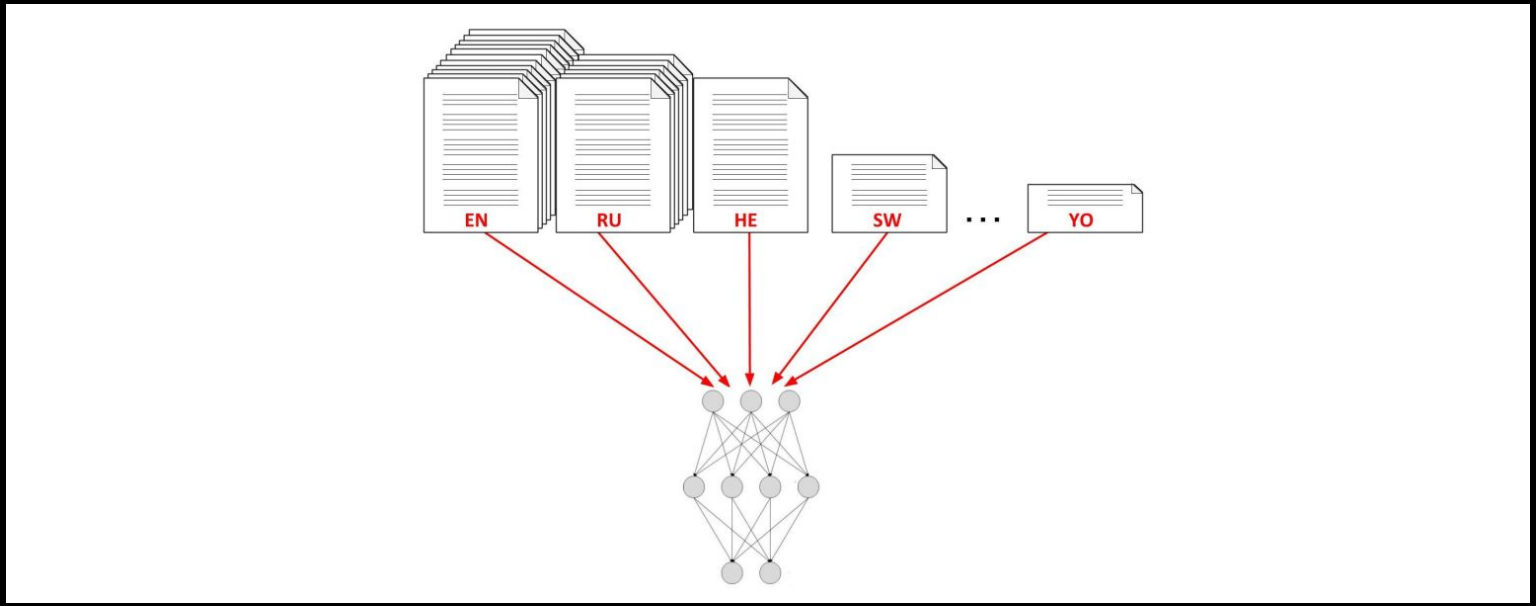


(c) Swedish→ L_2

While fully unsupervised CLWEs really show impressive performance for similar language pairs, they are still worse than weakly supervised methods…

- Furthermore, we don't really need them for these scenarios...

Joint Multilingual Learning in a Nutshell

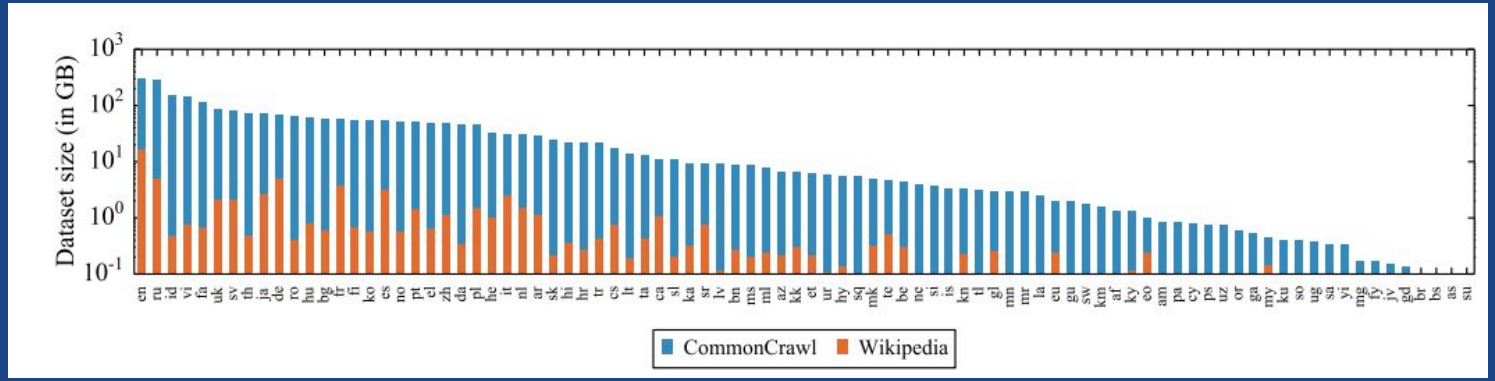


Joint multilingual learning – train a single model on a mix of datasets in all languages, to enable **data and parameter sharing** where possible

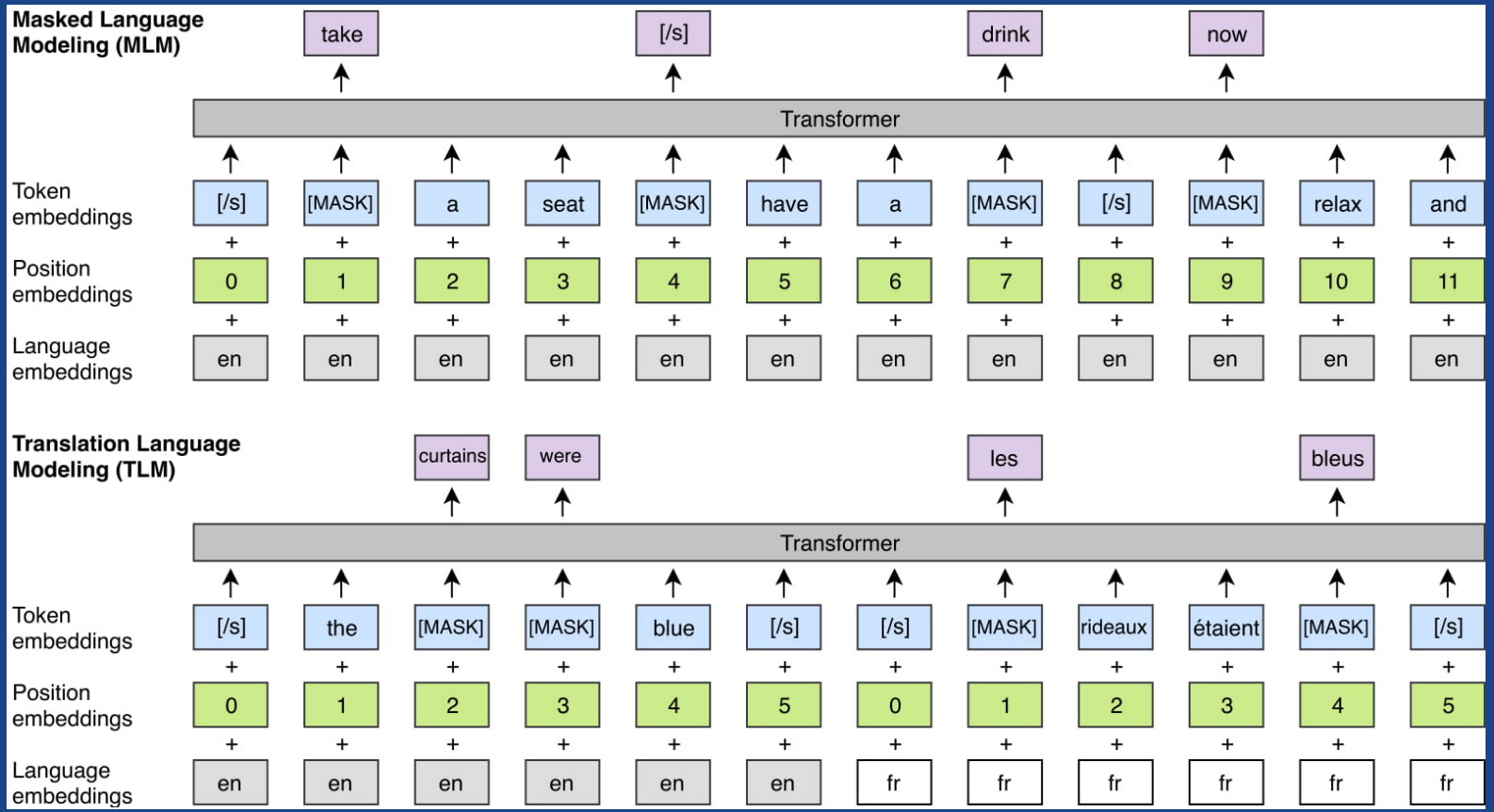
Baseline models

- Pretrained multilingual encoders are used as a baseline
  - mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020)
  - **Zero-shot cross-lingual transfer**

- The encoders are pre-trained on 100+ languages on the texts from Wikipedia and CC-100



Distribution of languages in CC-100 (Source: Conneau et al., 2020)


Masked Language Modeling (MLM)
take
[/s]
drink
now
Transformer
Token embeddings
[/s] [MASK] a seat [MASK] have a [MASK] [/s] [MASK] relax and
Position embeddings
0 1 2 3 4 5 6 7 8 9 10 11
Language embeddings
en en en en en en en en en en en en
Translation Language Modeling (TLM)
curtains
were
les
bleus
Transformer
Token embeddings
[/s] the [MASK] [MASK] blue [/s] [/s] [MASK] rideaux étaient [MASK] [/s]
Position embeddings
0 1 2 3 4 5 0 1 2 3 4 5
Language embeddings
en en en en en en fr fr fr fr fr fr