

Multilingual NLP

5. Cross-Lingual Word Embeddings

(+ Multilingual Resources)

Prof. Dr. Goran Glavaš
Center for AI and Data Science (CAIDAS), Uni Würzburg

After this lecture, you'll...

- Know what cross-lingual word embeddings (CLWEs) are
- Understand methods for inducing CLWEs from scratch
- Understand how to induce CLWEs from monolingual embeddings
- Know the limitations of unsupervised induction of CLWEs
- Be able to evaluate the quality of CLWEs
- Be aware of resources with word/sentence translations

Content

- **Cross-Lingual Word Embeddings**
 - Joint Training (from scratch)
 - Projection-Based CLWEs
 - Unsupervised Induction of CLWEs
- Evaluation of CLWEs

Cross-Lingual Word Embeddings

- A **semantic vector space** in which words with similar meaning have similar vectors
- Whether they come from the same language or from different languages.

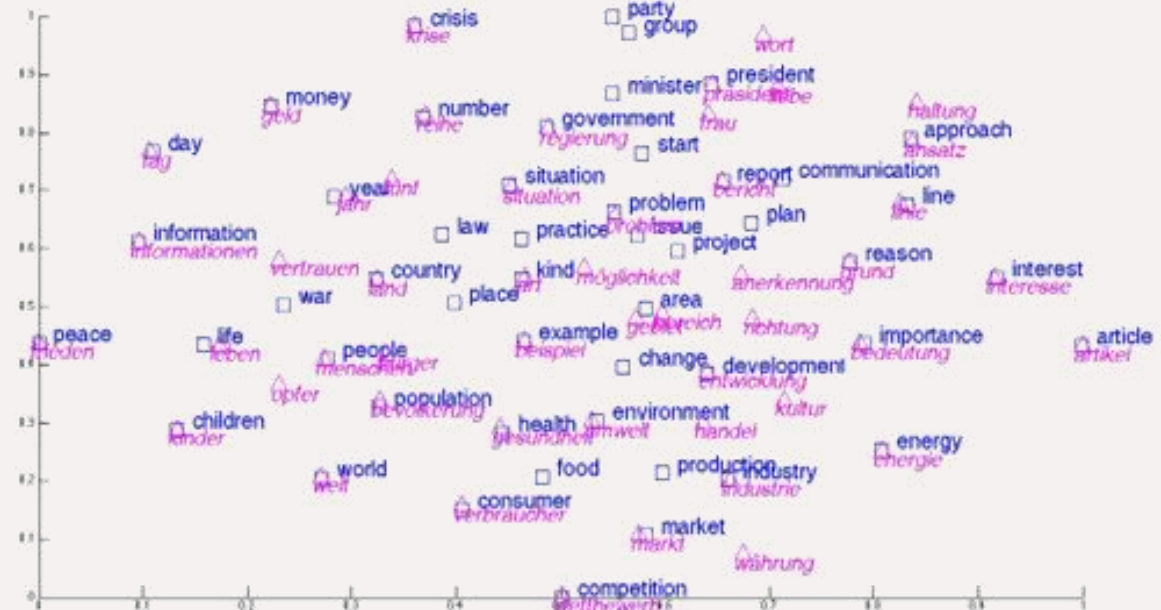


Image from: Luong, M. T., Pham, H., & Manning, C. D. (2015). [Bilingual word representations with monolingual quality in mind](#). *Proc. 1st Workshop on vector space modeling for natural language processing* (pp. 151-159).

Cross-Lingual Word Embeddings



Ruder, S., Vulić, I., & Søgaard, A. (2019). [A Survey of Cross-Lingual Word Embedding Models](#). Journal of Artificial Intelligence Research, 65, 569-631.

- Typology of methods for inducing Cross-Lingual Word Embeddings
 - **Type of bilingual / multilingual signal**
Document-level, sentence-level, word-level, no signal (i.e., unsupervised)
 - **Comparability**
Parallel texts, comparable texts, not comparable (i.e., randomly aligned)
 - **Point (time) of alignment**
Joint embedding models vs. Post-hoc alignment
 - **Modality**
Text only vs. using images for alignment

Content

- **Cross-Lingual Word Embeddings**
 - **Joint Training (from Scratch)**
 - Projection-Based CLWEs
 - Unsupervised Induction of CLWEs
- Evaluation of CLWEs



Joint CLWE Models

- Joint Cross-Lingual/Multilingual Word Embedding approaches induce embeddings of words from both/all languages simultaneously
- Using different types of (gold) bilingual signal:
 - Word translations
 - Easier/cheaper to obtain (+)
 - Less reliable signal, words out of context (-)
 - Sentence translations
 - More difficult/expensive to obtain (-)
 - Richer signal for aligning representations between languages (+)



Joint CLWEs with Word Translations

- Input
 - Dictionary of word translations $D = \{(w_s^k, w_t^k)\}_k$
 - Source language corpus C_s and vocabulary V_s
 - Target language corpus C_t and vocabulary V_t
- Q: Where to get D from?
 - Massively multilingual lexico-semantic resources!
 - **BabelNet**, [PanLex](#), ...
 - BabelNet covers over 500 languages
 - Caveat: not all languages have same coverage
 - PanLex covers 5,700 languages
 - Caveat: very low coverage for most languages



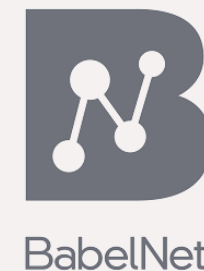
BabelNet



PANLEX

BabelNet

- Massively multilingual lexico-semantic network
 - Effectively, a **graph**
 - Nodes are so-called **synonym sets** (**synsets**)



song

bn:00072794n Noun Concept Categories: Articles with short description, Ritual, Wikipedia arti...

EN song /ə/ • vocal /ə/

Synset ID

TRANSLATIONS	DEFINITIONS	EXAMPLES
<div>English > Arabic x Ukrainian x Quechua x More languages...</div>		
<div><div>EN</div><div><p>A short musical composition with words WordNet 3.0 & Open English WordNet</p><p>A song is a musical composition intended to be performed by the human voice. Wikipedia</p><p>Musical composition for voice or voices. Wikipedia Disambiguation</p><p>Musical composition for voice Wikidata</p><p>A musical piece with lyrics (or "words to sing"); prose that one can sing. OmegaWiki</p><p>A musical composition with lyrics for voice or voices, performed by singing. Wiktionary</p><p>Musical composition. Wiktionary (translation)</p></div></div>		
<div>AR</div>		
<div>UK</div> <div><p>Пісня, співанка — словесно-музичний твір, призначений для співу. Wikipedia</p></div>		
<div>QU</div> <div><p>Rimay taki nisqaqa takisqa harawim, wachuchikunapi rurasqa. Wikipedia</p></div>		

BabelNet



BabelNet

- Massively multilingual lexico-semantic network
 - Effectively, a **graph**
 - Nodes are so-called **synonym sets** (**synsets**)
 - Multilingual **glosses** (definitions) available

TRANSLATIONS

DEFINITIONS

EXAMPLES

English > Arabic x Ukrainian x Quechua x

EN

A short musical composition with words [WordNet 3.0 & Open English WordNet](#)

A song is a musical composition intended to be performed by the human voice. [Wikipedia](#)

Musical composition for voice or voices. [Wikipedia Disambiguation](#)

Musical composition for voice [Wikidata](#)

A musical piece with lyrics (or "words to sing"); prose that one can sing. [OmegaWiki](#)

A musical composition with lyrics for voice or voices, performed by singing. [Wiktionary](#)

Musical composition. [Wiktionary \(translation\)](#)

AR

UK

Пісня, співанка — словесно-музичний твір, призначений для співу. [Wikipedia](#)

QU

Rimay taki nisqaqa takisqa harawim, wachuchikunapi rurasqa. [Wikipedia](#)

BabelNet

- Massively multilingual lexico-semantic network
 - Effectively, a **graph** with typed edges
 - Nodes are so-called **synonym sets** (**synsets**)
 - Edges are **lexico-semantic relations** between synsets, e.g.:
 - Hypernymy (**is-a**)
 - Meronymy (**part-of**)
 - ...



BabelNet

English > Arabic x Ukrainian x Quechua x

IS A	musical composition • literary form • literary genre • music • vocal music
HAS PART	refrain • lyrics • song verse • DE Reprise • couplet +1 relations
PART OF	songbook • Breton song
HAS KIND	anthem • aria • ballad • scion • barcarole +128 relations
HAS INSTANCE	Magnificat • I'm Free • Wishin' and Hopin' • Dame • Flying the Flag +9K relations
DERIVATION	songwriter • songster • sing • sing
DESCRIBED BY SOURCE	Brockhaus and Efron Encyclopedic Dictionary • Otto's encyclopedia • Gujin Tushu Jicheng
DIFFERENT FROM	canzona • song form • musical work
INSTRUMENTATION	voice
MODEL ITEM	Poovukkul • Wuthering Heights
ON FOCUS LIST OF WIK...	HY Հիբիպեդիա: Կարևորագույն հոդվածներ
PARTIALLY COINCIDEN...	piesn
SAID TO BE THE SAME ...	Song



Joint CLWEs with Word Translations

- Word-level alignments: $D = \{(w_s^k, w_t^k)\}_i$
- Source language corpus C_s and vocabulary V_s
- Target language corpus C_t and vocabulary V_t
- **Idea:** modify the word embedding model (e.g., Skip-Gram) so that words that are mutual translations share the embedding vector
 - I.e., for each pair (w_s^i, w_t^i) from D , enforce $\mathbf{x}_s^k = \mathbf{x}_t^k$
- Joint vocabulary $V = V_s \cup V_t$
 - Corresponding joint embedding matrices: $\mathbf{W}_1 \in \mathbb{R}^{|V| \times d}$ and $\mathbf{W}_2 \in \mathbb{R}^{d \times |V|}$
 - Shared embeddings \mathbf{x}_1^k and \mathbf{x}_2^k for mutual translations w_s^k and w_t^k



Joint CLWEs with Word Translations

- Training data: simple concatenation of the corpora in both languages

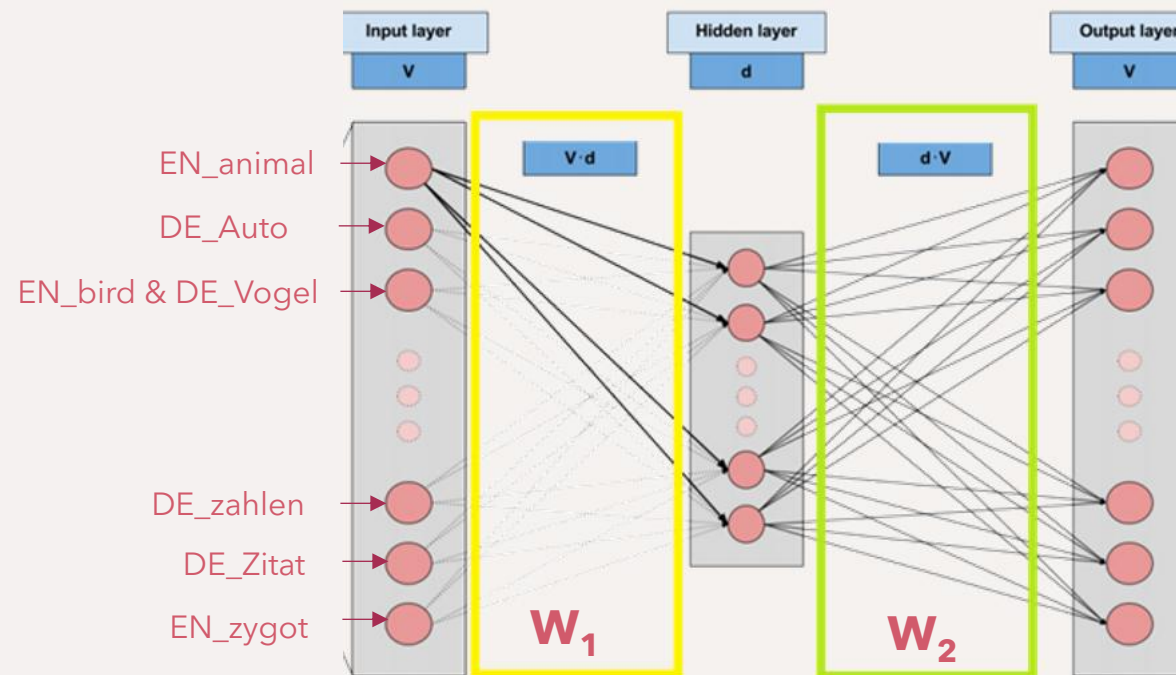
- Example: EN source, DE target

- $D = \{..., (bird, Vogel), ...\}$

Context (EN): blue **bird** flies over the nest...

Context (DE): Gesang des roten schönen **Vogels** ...

- Tied vectors of word translations drive the representational alignment between languages



Joint CLWEs with Sentence Translations



Luong, M. T., Pham, H., & Manning, C. D. (2015, June). [Bilingual word representations with monolingual quality in mind](#). In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing* (pp. 151-159).

- Example: Bilingual Skip-Gram (Bi-Skip-Gram) model of Luong et al.
- Parallel sentences required
 - A model for word alignment also needed
 - We'll cover word alignment in Lecture 8

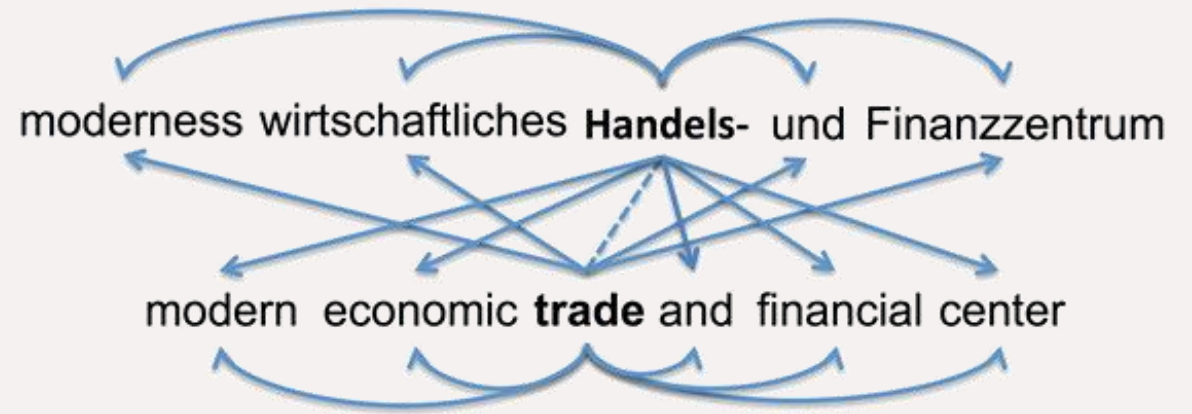


Image from: Luong et al.

Joint CLWEs with Sentence Translations



Luong, M. T., Pham, H., & Manning, C. D. (2015, June). [Bilingual word representations with monolingual quality in mind](#). In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing* (pp. 151-159).

- Example: Bilingual Skip-Gram (Bi-Skip-Gram) model of Luong et al.
- Parallel sentences required
- Monolingual (both languages):
 - *Handels-* → *moderness*
 - *Handels-* → *wirtschaftliches*
 - ...
 - *trade* → *modern*
 - *trade* → *economic*
 - ...

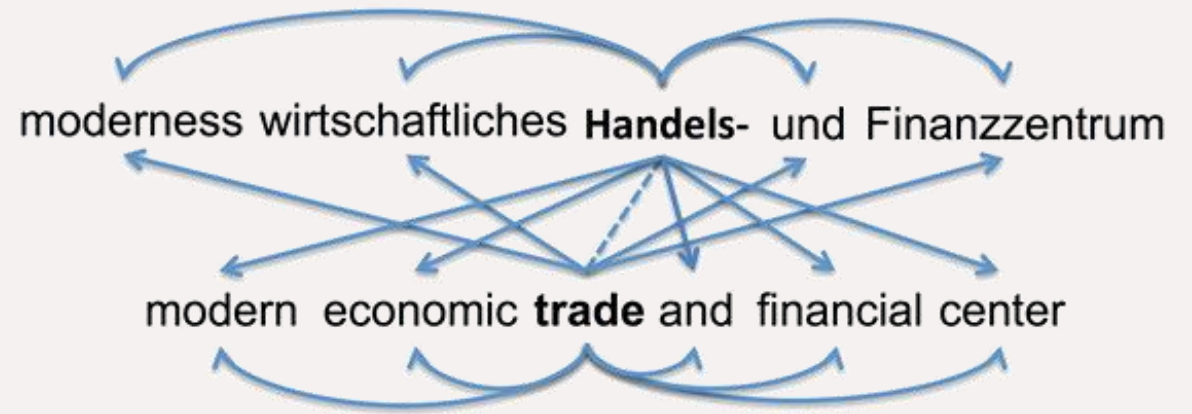


Image from: Luong et al.

Joint CLWEs with Sentence Translations



Luong, M. T., Pham, H., & Manning, C. D. (2015, June). [Bilingual word representations with monolingual quality in mind](#). In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing* (pp. 151-159).

- Example: Bilingual Skip-Gram (Bi-Skip-Gram) model of Luong et al.
- Parallel sentences required
- Cross-lingual (both languages):
 - *Handels-* → *modern*
 - *Handels-* → *economic*
 - ...
 - *trade* → *moderness*
 - *trade* → *wirtschaftliches*
 - ...

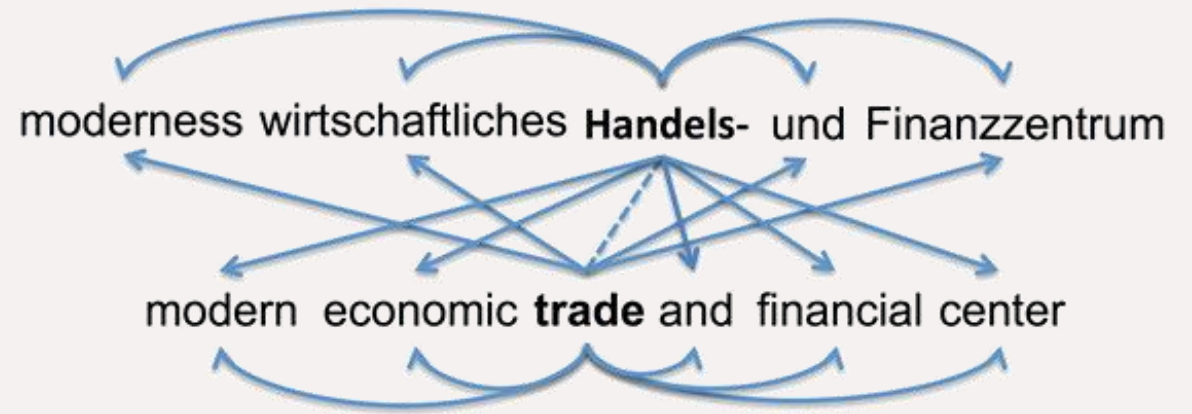


Image from: Luong et al.

Sentence Translations

- **Q:** Where to get parallel sentences from?
- Parallel corpora is the main training data for **machine translation**
 - Collecting it (manually, automatically, semi-automatically) has therefore been a major focus in MT
 - We will discuss approaches for creating parallel data in **Lecture 9**
- Some prominent sources of parallel data
 - Opus: Aggregator of all Open-Source parallel corpora
 - WikiMatrix: automatically created from Wikipedia
 - Based on **multilingual sentence encoders** (**Lecture 10**)
 - „Quasi-parallel“ – not manually curated
 - 85 languages and 1620 language pairs
 - Multi-Bible: Manual Bible translations exist in 1500+ languages
 - **Multi-parallel**: sentences aligned across many (all) languages

Content

- **Cross-Lingual Word Embeddings**
 - Joint Training (from Scratch)
 - **Projection-Based CLWEs**
 - Unsupervised Induction of CLWEs
- Evaluation of CLWEs

Projection-Based CLWEs

- Q: What could be the main shortcoming of joint CLWE models?
 - Let's say we have N languages
 - And we need words from all N in a joint embedding space
- For each language pair: train a bilingual model from scratch
- For a multilingual space:
 - Let's say we have a pivot language (commonly English)
 - We induce $N-1$ bilingual spaces E_N-L_2
 - Q: how to align these $N-1$ spaces?
- Q: Multilingual Skip-Gram?
 - We'd need multi-parallel corpora – usually very limited in size



Projection-Based CLWEs

- On the other hand, pretrained **monolingual word embeddings** exist for very many languages
- **Idea:** can we (cheaply) align monolingual embedding spaces **post-hoc**?
- To get a multilingual word embedding space for **N** languages :
 1. Train **N** monolingual spaces
 2. Learn **N-1** (cheap) alignments (**N-1** languages to **EN** as pivot)
- Let $\mathbf{X}_{L_1} \in \mathbb{R}^{|V_s| \times d}$ and $\mathbf{X}_{L_2} \in \mathbb{R}^{|V_t| \times d}$ be the independently trained monolingual embeddings of two languages **L1** and **L2**
- **Projection-based CLWEs:** find an „alignment“ between \mathbf{X}_{L_1} and \mathbf{X}_{L_2} such that **words with similar meaning (across langs) get similar vectors**



Projection-Based CLWEs

- **Post-hoc alignment** of monolingual word embedding spaces

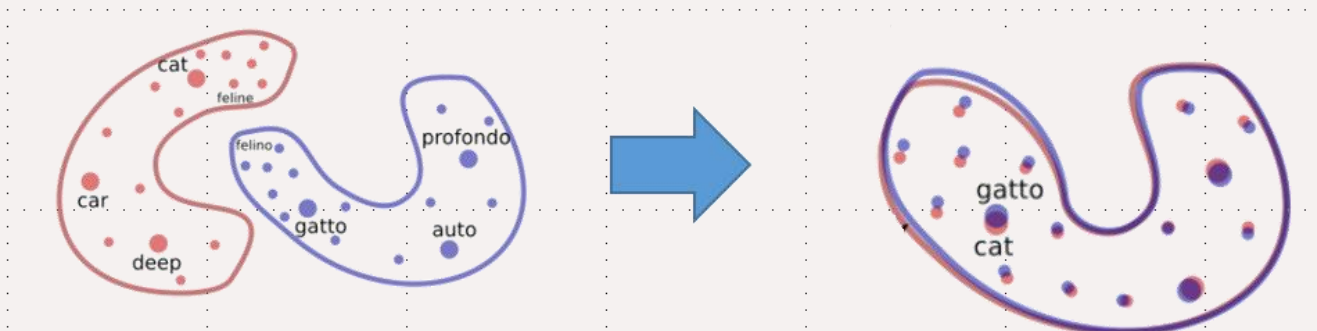
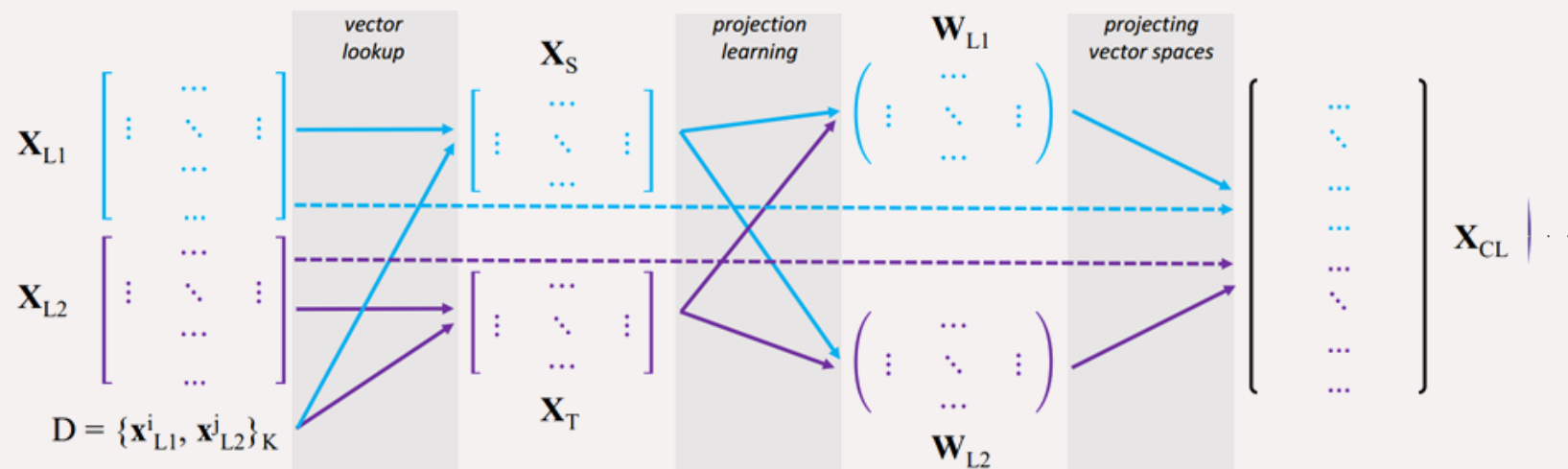


Image from: Lample, G., Conneau, A., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018) [Word translation without parallel data](#). In *International Conference on Learning Representations*.

- In general, we are looking for functions f and g that produce a meaningful bilingual embedding space $f(\mathbf{X}_{L1}|\boldsymbol{\theta}_{L1}) \cup g(\mathbf{X}_{L2}|\boldsymbol{\theta}_{L2})$

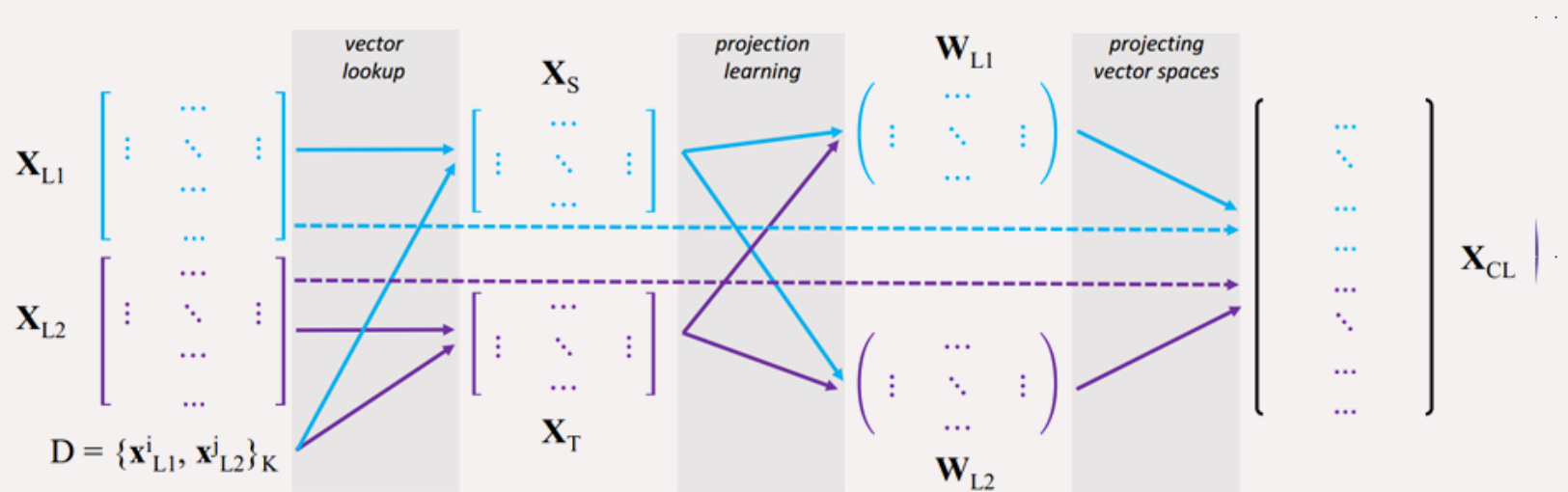
Projection-Based CLWEs

- **Post-hoc alignment** of independently trained monolingual word embedding spaces
 - Alignment based on word translation pairs, $\mathbf{D} = \{(\mathbf{x}_{L1}^k, \mathbf{x}_{L2}^k)\}_k$ is the set of word embedding pairs between the languages corresponding to pairs of mutual translations



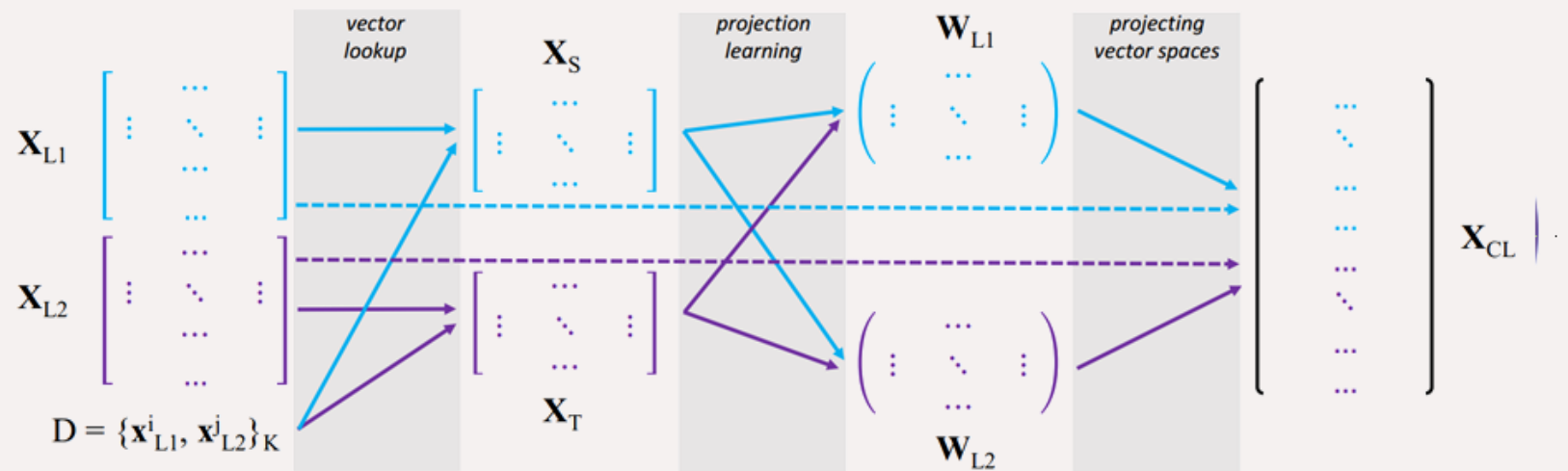
Projection-Based CLWEs

- **Post-hoc alignment** of independently trained monolingual word embedding spaces
 - Alignment based on word translation pairs, $\mathbf{D} = \{(\mathbf{x}_{L1}^k, \mathbf{x}_{L2}^k)\}_k$ is the set of word embedding pairs between the languages corresponding to pairs of mutual translations
 - We stack $\{\mathbf{x}_{L1}^k\}_k$ into matrix $\mathbf{X}_S \in \mathbb{R}^{k \times d_1}$ and $\{\mathbf{x}_{L2}^k\}_k$ into the matrix $\mathbf{X}_T \in \mathbb{R}^{k \times d_2}$



Projection-Based CLWEs

- **Post-hoc alignment** of independently trained monolingual word embedding spaces



- In the general case, we want to find **projection matrices** $\mathbf{W}_{L1} \in \mathbb{R}^{d1 \times d}$ and $\mathbf{W}_{L2} \in \mathbb{R}^{d2 \times d}$ such that $\mathbf{X}_S \mathbf{W}_{L1} = \mathbf{X}_T \mathbf{W}_{L2}$
 - This is a model, in which \mathbf{W}_{L1} and \mathbf{W}_{L2} are parameters
 - Q: What objective function to use?

Projection-Based CLWEs

- Find **projection matrices**
 - $\mathbf{W}_{L1} \in \mathbb{R}^{d1 \times d}$ and $\mathbf{W}_{L2} \in \mathbb{R}^{d2 \times d}$ such that $\mathbf{X}_S \mathbf{W}_{L1} = \mathbf{X}_T \mathbf{W}_{L2}$
 - In practice, the problem is equivalent to learning one parameter matrix \mathbf{W} , i.e., $\mathbf{X}_S \mathbf{W} = \mathbf{X}_T$

	\mathbf{X}_S		\mathbf{X}_T	
bird	$\begin{bmatrix} -1.18 & 0.21 & \dots & 0.11 \end{bmatrix}$	$\mathbf{W} =$	$\begin{bmatrix} 0.59 & 1.01 & \dots & 0.37 \end{bmatrix}$	Vogel
pretty	$\begin{bmatrix} 0.23 & -0.53 & \dots & 0.34 \end{bmatrix}$		$\begin{bmatrix} -0.34 & -0.27 & \dots & 0.41 \end{bmatrix}$	schön
...	$\begin{bmatrix} \dots & \dots & \dots & \dots \end{bmatrix}$		$\begin{bmatrix} \dots & \dots & \dots & \dots \end{bmatrix}$...
eat	$\begin{bmatrix} 0.78 & 1.33 & \dots & -0.47 \end{bmatrix}$		$\begin{bmatrix} 0.81 & -0.31 & \dots & 0.29 \end{bmatrix}$	essen

Projection-Based CLWEs

	\mathbf{X}_S		\mathbf{X}_T
bird	$\begin{bmatrix} -1.18 & 0.21 & \dots & 0.11 \end{bmatrix}$	\mathbf{W}	$\begin{bmatrix} 0.59 & 1.01 & \dots & 0.37 \end{bmatrix}$ Vogel
pretty	$\begin{bmatrix} 0.23 & -0.53 & \dots & 0.34 \end{bmatrix}$	$=$	$\begin{bmatrix} -0.34 & -0.27 & \dots & 0.41 \end{bmatrix}$ schön
...	$\begin{bmatrix} \dots & \dots & \dots & \dots \end{bmatrix}$		$\begin{bmatrix} \dots & \dots & \dots & \dots \end{bmatrix}$...
eat	$\begin{bmatrix} 0.78 & 1.33 & \dots & -0.47 \end{bmatrix}$		$\begin{bmatrix} 0.81 & -0.31 & \dots & 0.29 \end{bmatrix}$ essen

- The corresponding objective is „least squares“:

$$\operatorname{argmin}_{\mathbf{W}} \|\mathbf{X}_S \mathbf{W} - \mathbf{X}_T\|$$

- Minimize the Euclidean distance between source language projections and corresponding target language vectors
- If \mathbf{W} is unconstrained, no unique closed form solution
 - Numeric optimization \rightarrow minimization with GD

Projection-Based CLWEs



Mikolov, T., Le, Q. V., & Sutskever, I. (2013). [Exploiting similarities among languages for machine translation](#). *arXiv preprint arXiv:1309.4168*.

- The corresponding objective is **least squares**:

$$\operatorname{argmin}_{\mathbf{W}} \|\mathbf{X}_S \mathbf{W} - \mathbf{X}_T\|$$

- Mikolov et al. find **W** via numeric optimization
- Trains in mini-batches of **k** word pairs
- With mini-batch gradient descent

Projection-Based CLWEs



Smith, S. L., Turban, D. H., Hamblin, S., & Hammerla, N. Y. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations*.

- Turns out that we learn better projections if we constraint \mathbf{W} to be an orthogonal matrix, i.e., such that its rows and columns are orthonormal

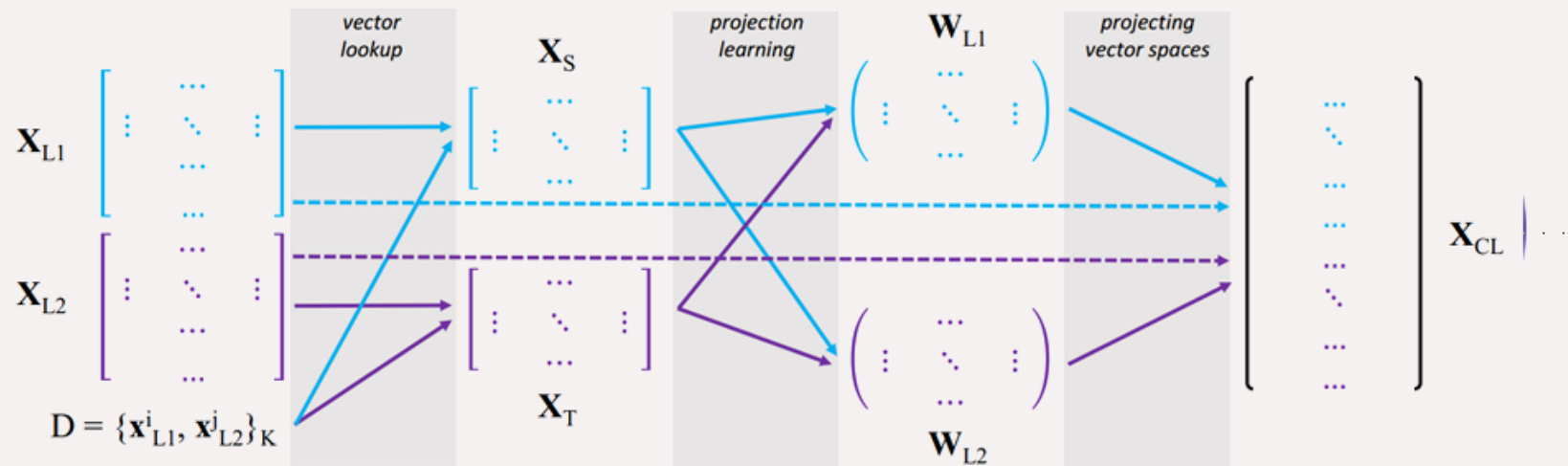
$$\operatorname{argmin}_{\mathbf{W}} \|\mathbf{X}_S \mathbf{W} - \mathbf{X}_T\|, \text{ s.t. } \mathbf{W}^T \mathbf{W} = \mathbf{I}$$

- This optimization problem is known as the Procrustes problem and has a closed-form solution:

$$\mathbf{W} = \mathbf{U}\mathbf{V}^T \text{ where } \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \text{SVD}(\mathbf{X}_S^T \mathbf{X}_T)$$

- SVD = a matrix factorization method called Singular Value Decomposition

Projection-Based CLWEs



- So, in practice, $\mathbf{W}_{L2} = \mathbf{I}$ and we obtain $\mathbf{W} = \mathbf{W}_{L1}$ by solving the Procrustes problem on \mathbf{X}_S and \mathbf{X}_T
- Having „learned“ the projection \mathbf{W} , we project the whole embedding space of L1 (source) into the embedding space of L2 (target)

$$\mathbf{X}_{\text{biling}} = \mathbf{X}_{L1} \mathbf{W} \cup \mathbf{X}_{L2}$$

Projection-Based CLWEs

- Advantage of projection-based CLWE methods over joint induction:
 - **Compute**: learning an orthogonal projection (i.e., solving Procrustes) is very computationally cheap
 - **Flexibility**: works regardless of how the monolingual embedding spaces \mathbf{X}_{L1} and \mathbf{X}_{L2} were obtained
 - Even if \mathbf{X}_{L1} and \mathbf{X}_{L2} trained with different methods
 - **Performance**: the quality of CLWEs induced via projection matches or surpasses that of jointly induced CLWEs
- **Q**: Where do we get word translations for training the projection \mathbf{W} ?
- **Q**: How many word translation pairs do we need to learn a **good** projection?
 - I.e., what value should we set k in $D = \{(w_s^k, w_t^k)\}_k$ to?

Projection-Based CLWEs



Glavaš, G., Litschko, R., Ruder, S., & Vulić, I. (2019, July). [How to \(Properly\) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions](#). In Proceedings of ACL (pp. 710-721).

- Q: How many word translation pairs do we need to learn a **good** projection?
- Depends on several factors, primarily
 - (1) **Lexical proximity** of languages,
 - (2) Quality of monolingual word embeddings (**size of pretraining corpora**)
- In general, performance saturates with ca. **5K** translation pairs
 - Marginal gains with more translation pairs
- Q: why do we stick to a linear model? Why not learn a **non-linear model** (with more parameters than a single projection matrix)?

Content

- **Cross-Lingual Word Embeddings**
 - Joint Training (from Scratch)
 - Projection-Based CLWEs
 - **Unsupervised Induction of CLWEs**
- Evaluation of CLWEs

Unsupervised Projection-Based CLWEs

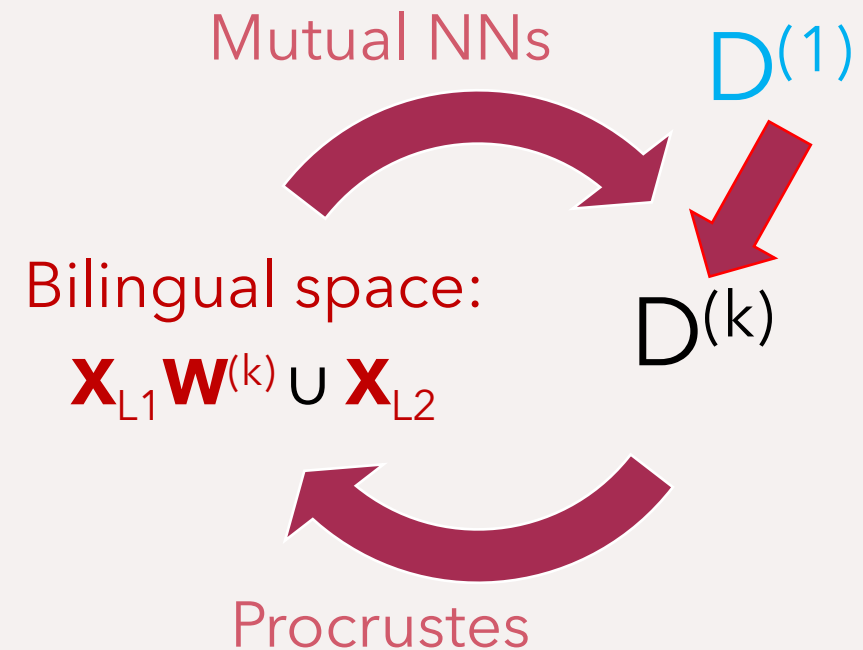
- **Unsupervised CLWEs:** In 2018, a flood of work introducing projection-based CLWE methods that do not require any word translations

- The **same general framework** for all unsupervised CLE models

1. Induce (automatically) initial word alignment dictionary $\mathbf{D}^{(1)}$

Repeat:

2. Learn the projection $\mathbf{W}^{(k)}$ using $\mathbf{D}^{(k)}$
3. Induce new dictionary $\mathbf{D}^{(k+1)}$ from $\mathbf{X}_{L1} \mathbf{W}^{(k)} \cup \mathbf{X}_{L2}$

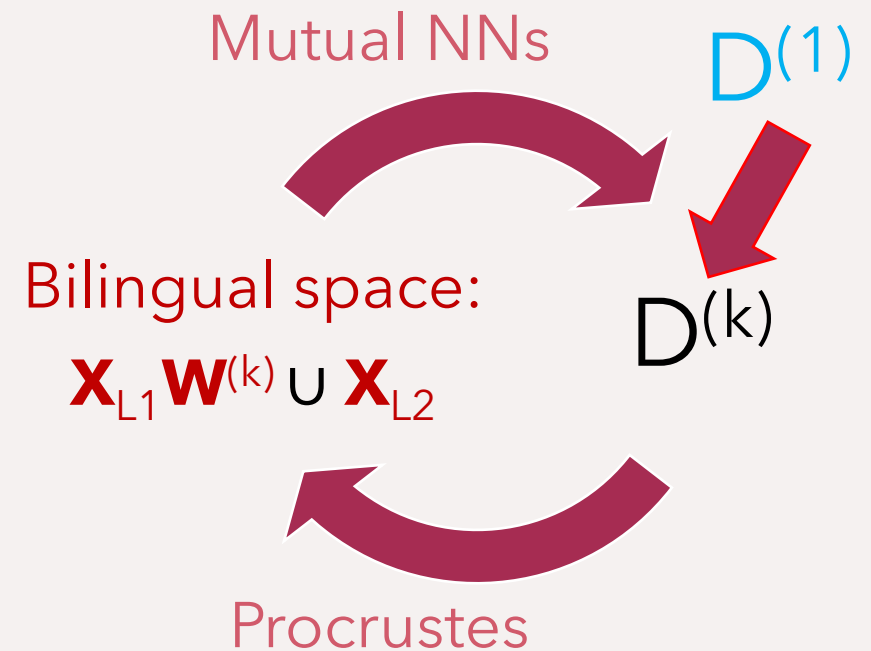


Unsupervised Projection-Based CLWEs



Lample, G., Conneau, A., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018) [Word translation without parallel data](#). In International Conference on Learning Representations.

- **Generative adversarial network** for initial alignment dictionary $D^{(1)}$
 - **Generator**: the projection matrix \mathbf{W}
 - **Discriminator**: classifier that distinguishes between $\mathbf{x}_{L1}\mathbf{W}$ and \mathbf{x}_{L2} , i.e., predicts whether a vector has been obtained by:
 1. Transforming source language vector \mathbf{x}_{L1} with the projection matrix \mathbf{W} (i.e., $\mathbf{x}_{L1}\mathbf{W}$) or
 2. if its an original target language vector \mathbf{x}_{L2}

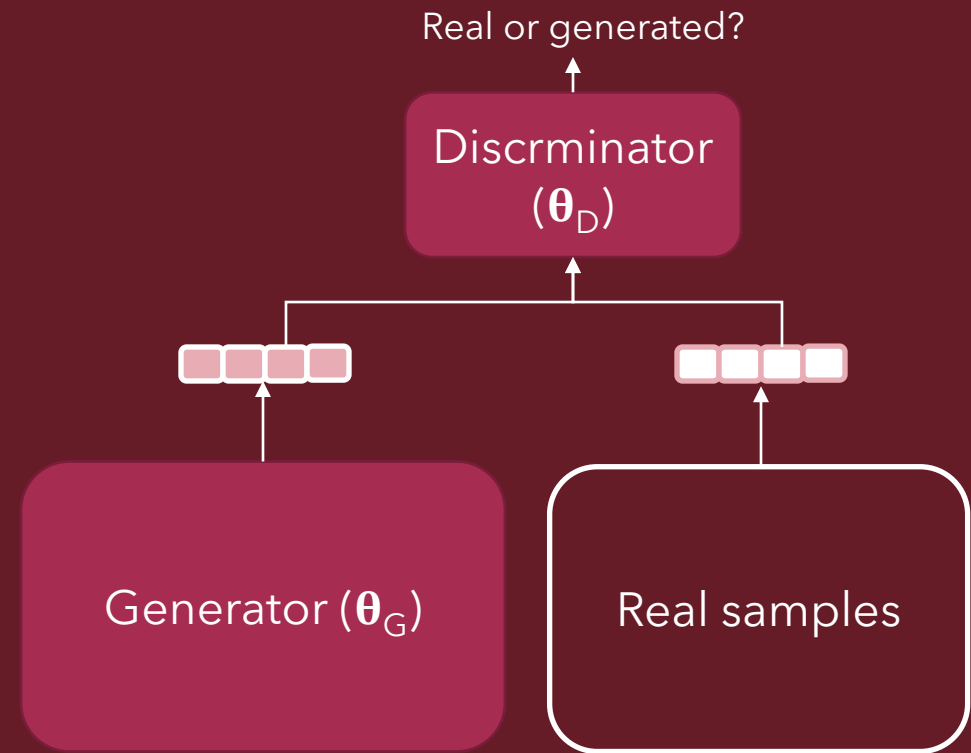


Generative Adversarial Networks



Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014, December). [Generative Adversarial Nets](#). In Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2 (pp. 2672-2680).

- **Generator:** our core **neural model** that **generates** vectors in continuous space
 - Images, word embeddings, ...
 - Parameters: θ_G
- **Discriminator:** a binary classifier that predicts whether a vector was
 - (1) generated by the generator or
 - (2) it is a real/original vector
 - Parameters: θ_D

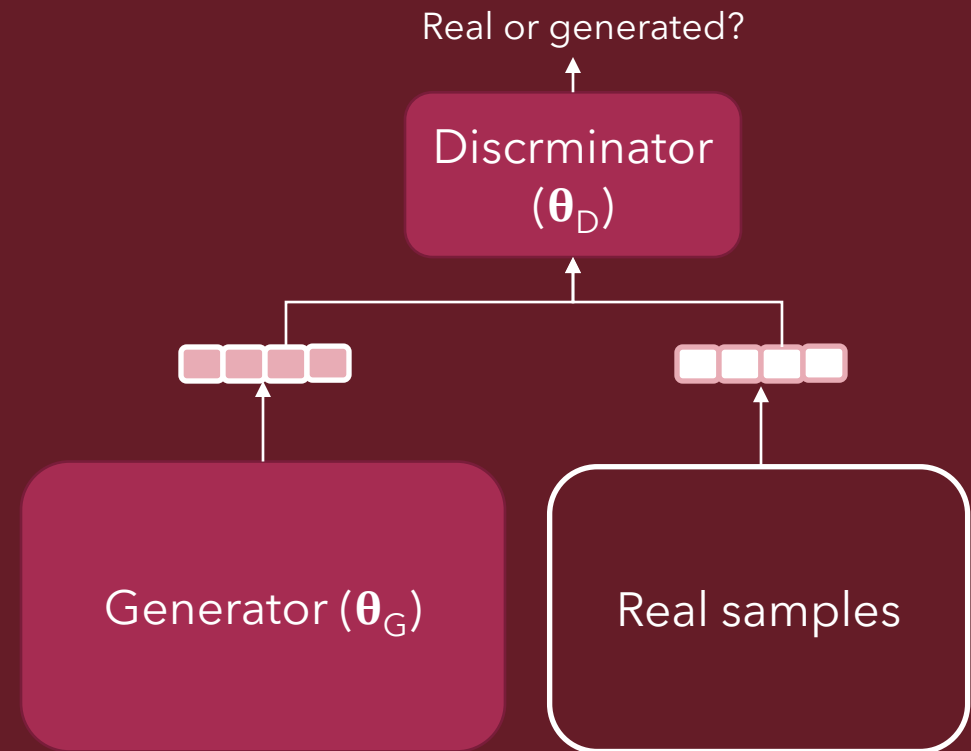


Generative Adversarial Networks



Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014, December). [Generative Adversarial Nets](#). In Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2 (pp. 2672-2680).

- **Generator:** $\text{Gen}(\mathbf{x}|\theta_G)$
- **Discriminator:** $\text{Disc}(\mathbf{x}|\theta_D)$
- Discriminator's job is to minimize its binary classification loss
- Generator's job is to **fool** the discriminator
 - I.e., maximize the discriminator's loss

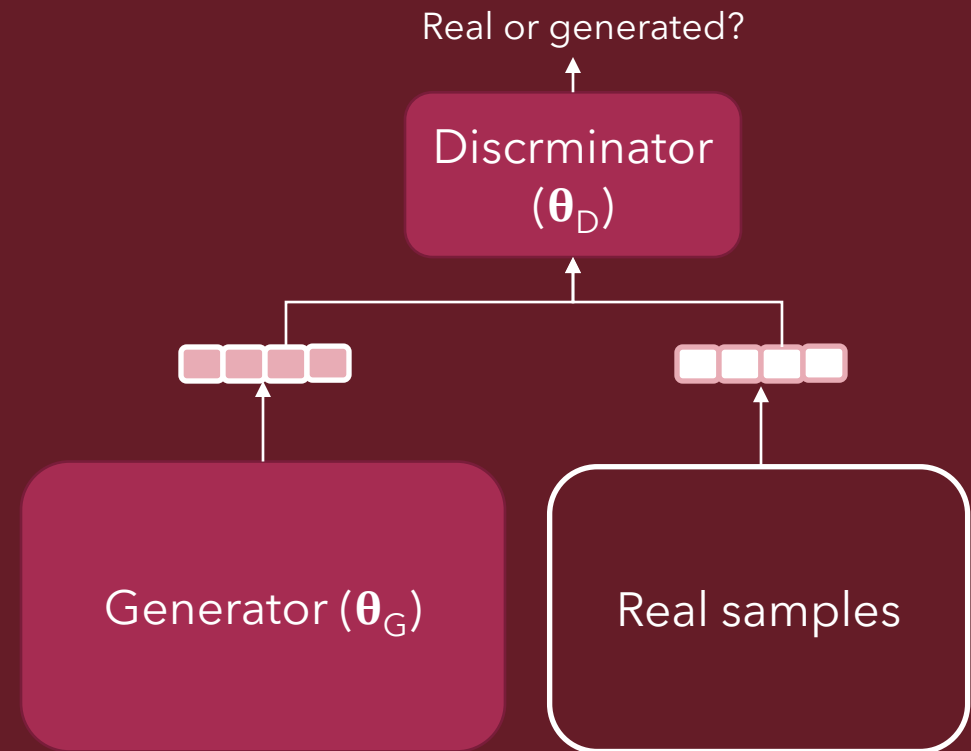


Generative Adversarial Networks



Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014, December). [Generative Adversarial Nets](#). In Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2 (pp. 2672-2680).

- **Generator:** $\text{Gen}(\mathbf{x}|\theta_G)$
- **Discriminator:** $\text{Disc}(\mathbf{x}|\theta_D)$
- Generator's job is to **fool** the discriminator
 - Generations are better the more they resemble the real examples
 - I.e., generations fit well into the „distribution“ of real examples

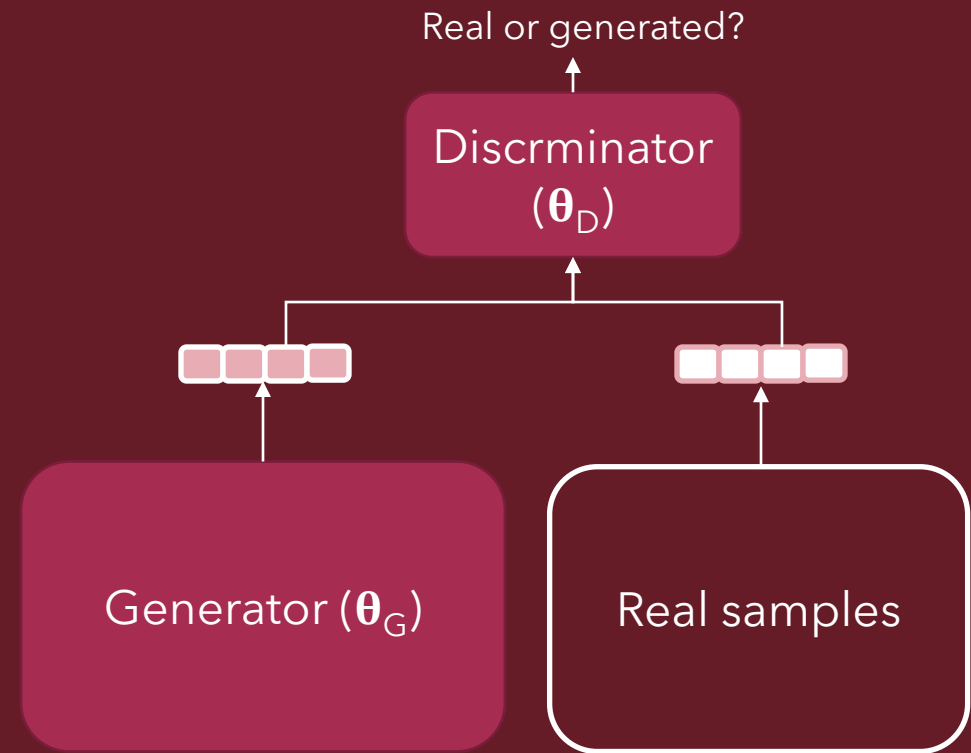


Generative Adversarial Networks



Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014, December). [Generative Adversarial Nets](#). In Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2 (pp. 2672-2680).

- A **competition** that iteratively makes both become better
- Iteratively:
 1. Feed into discriminator either (1) $\mathbf{x} = \text{Gen}(\text{input}|\boldsymbol{\theta}_G)$ or a real sample \mathbf{x}
 2. Compute the discriminator's loss $L_D(\text{Disc}(\mathbf{x}|\boldsymbol{\theta}_D))$
 3. Minimize discriminator's parameters with GD: $\boldsymbol{\theta}_D^{(k+1)} = \boldsymbol{\theta}_D^{(k)} - \eta \nabla_{\boldsymbol{\theta}} L_D$



Generative Adversarial Networks



Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014, December). [Generative Adversarial Nets](#). In Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2 (pp. 2672-2680).

- A **competition** that iteratively makes both become better
- Iteratively:

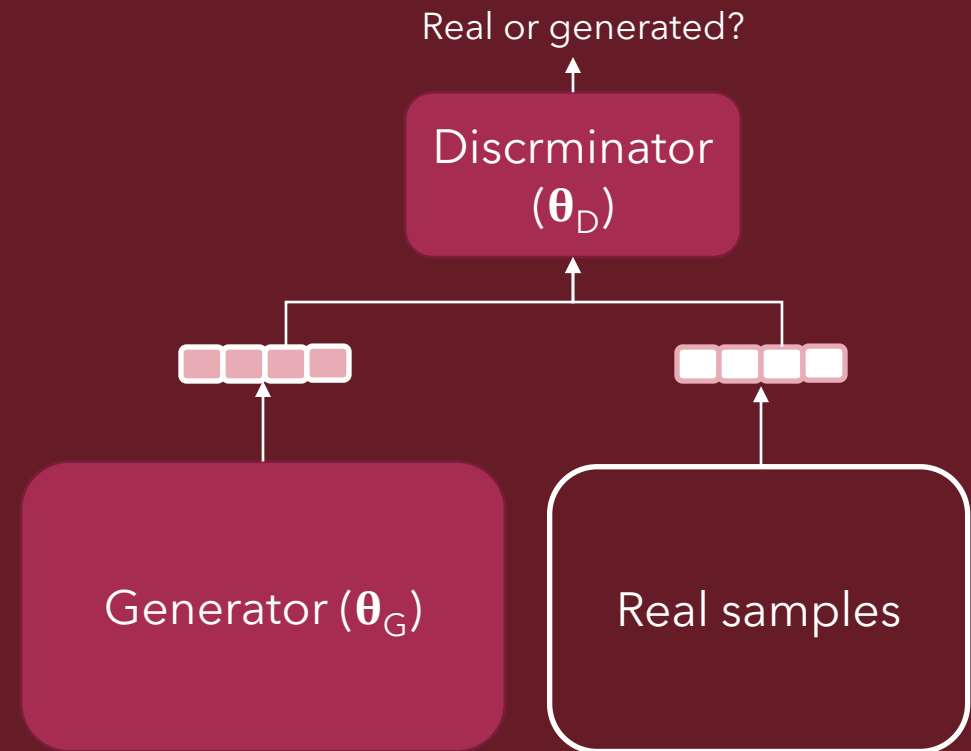
...

3. Minimize discriminator's parameters (GD):

$$\theta_D^{(k+1)} = \theta_D^{(k)} - \eta \nabla_{\theta_D} L_D$$

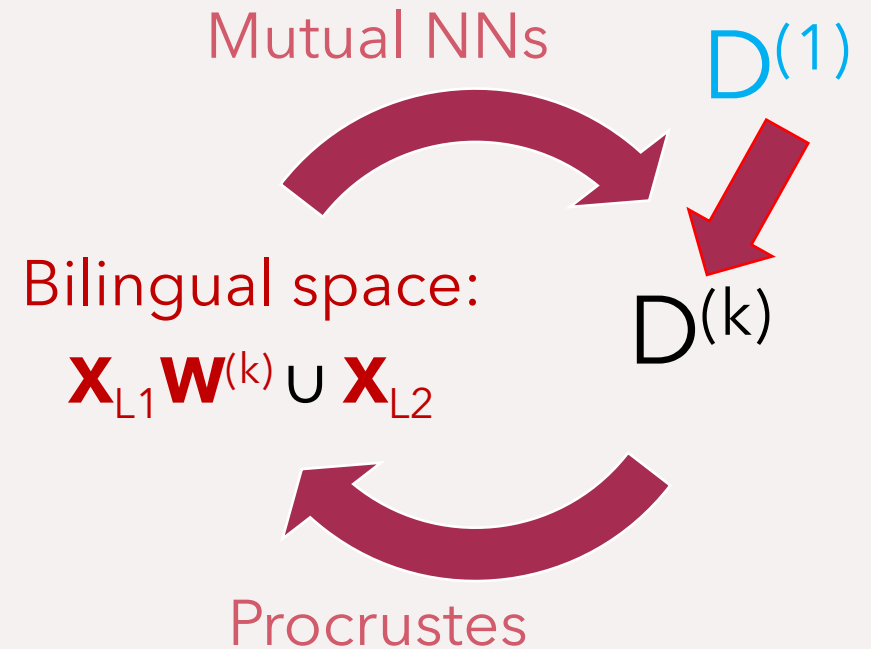
4. If \mathbf{x} is a generated sample, $\mathbf{x} = \text{Gen}(\text{input}|\theta_G)$ then update θ_G to maximize L_D :

$$\theta_G^{(k+1)} = \theta_G^{(k)} + \eta \nabla_{\theta_G} L_D$$



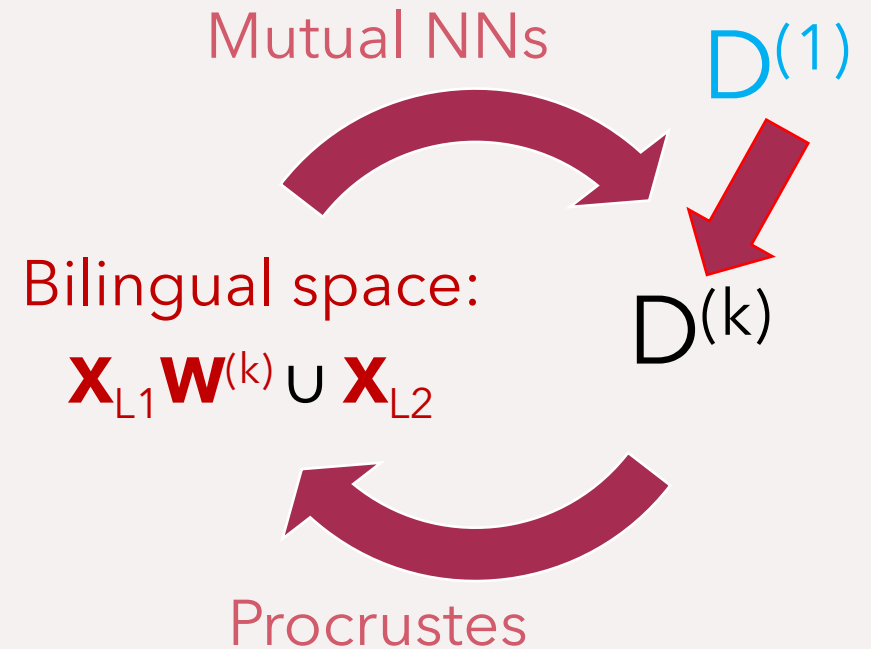
Unsupervised Projection-Based CLWEs

- The dictionary $D^{(k+1)}$ (next iteration):
 - Mutual nearest neighbours in $\mathbf{X}_{L1} \mathbf{W}^{(k)} \cup \mathbf{X}_{L2}$
 - $\mathbf{W}^{(k)}$ induced using dictionary $D^{(k)}$ from the current iteration
- Q: how do we find mutual NNs?
 1. For each \mathbf{x}_{L1}^i in $\mathbf{X}_{L1} \mathbf{W}^{(k)}$ rank all vectors from \mathbf{x}_{L2}^j in \mathbf{X}_{L2}
 2. For each \mathbf{x}_{L2}^j in \mathbf{X}_{L2} rank all vectors from \mathbf{x}_{L1}^i in $\mathbf{X}_{L1} \mathbf{W}^{(k)}$
 - Some measure of vector similarity
 - NNs are \mathbf{x}_{L1}^i and \mathbf{x}_{L2}^j that are on top of each other's ranking



Unsupervised Projection-Based CLWEs

- Q: how do we find mutual NNs?
 - Some measure of vector similarity
 - NNs are \mathbf{x}_{L1}^i and \mathbf{x}_{L2}^j that are on top of each other's ranking
- Similarity measure: cosine similarity
- **Hubness** problem:
 - Vector space: $\mathbf{X} \in \mathbb{R}^{d \times |V|}$
 - If $|V| \gg d$, there will be (by chance) vectors in $\mathbf{x} \in \mathbf{X}$ that have high-similarity with many/most other vectors
 - Skewes similarity measures like cosine



Unsupervised Projection-Based CLWEs



Lample, G., Conneau, A., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018) [Word translation without parallel data](#). In International Conference on Learning Representations.

- **Quality of CLWE**: accuracy of retrieving translation pair for a given word
 - When w_{L1}^i with vector \mathbf{x}_{L1}^i as „query“, we rank all $\mathbf{x} \in \mathbf{X}_{L2}$ based on similarity with \mathbf{x}_{L1}^i : where in the ranking is the vector \mathbf{x}_{L2}^j of the actual word translation w_{L2}^j
- **Hubness** problem in CLWEs:
 - A **hub** vector $\mathbf{x}_{L1}^i \in \mathbf{X}_{L1} \mathbf{W}$: high similarity with many vectors in \mathbf{X}_{L2} (and vice versa)
- **Cross-Domain Similarity Local Scaling**
 - Cosine similarity adjusted for the hubness of both vectors

$$\text{CSLS}(\mathbf{x}_{L1} \in \mathbf{X}_{L1} \mathbf{W}, \mathbf{x}_{L2} \in \mathbf{X}_{L2}) = 2 * \cos(\mathbf{x}_{L1}, \mathbf{x}_{L2}) - r_{L2}(\mathbf{x}_{L1}) - r_{L1}(\mathbf{x}_{L2})$$

Unsupervised Projection-Based CLWEs



Lample, G., Conneau, A., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018) [Word translation without parallel data](#). In International Conference on Learning Representations.

- Cross-Domain Similarity Local Scaling

- Cosine similarity adjusted for the hubness of both vectors

$$\text{CSLS}(\mathbf{x}_{L1} \in \mathbf{X}_{L1} \mathbf{W}, \mathbf{x}_{L2} \in \mathbf{X}_{L2}) = 2 * \cos(\mathbf{x}_{L1}, \mathbf{x}_{L2}) - r_{L2}(\mathbf{x}_{L1}) - r_{L1}(\mathbf{x}_{L2})$$

- $r_{L2}(\mathbf{x}_{L1})$ is the average cosine similarity that \mathbf{x}_{L1} has with K most similar vectors $\mathbf{x}_{L2} \in \mathbf{X}_{L2}$
- $r_{L1}(\mathbf{x}_{L2})$ is the average cosine similarity that \mathbf{x}_{L2} has with K most similar vectors $\mathbf{x}_{L1} \in \mathbf{X}_{L1} \mathbf{W}$

Unsupervised CLWEs: Criticism



Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019). [Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?](#) In Proceedings of the EMNLP (pp. 4407-4418).

- **Motivation**

- „No bilingual signal required“
- Thus applicable to „under-resourced languages“

- But: Supervised models don't need many word pairs (e.g., 1-5K)

- Trivial to obtain for any language pair from resources like: BabelNet, PanLex
- If a few thousand word translation pairs cannot be obtained
 - Then a language is **so low-resource** that we likely don't have **reliable monolingual embeddings** due to too small corpora in that language

Unsupervised CLWEs: Criticism



Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019). Do We Really Need Fully Unsupervised Cross-Lingual Embeddings? In Proceedings of the EMNLP (pp. 4407-4418).

- **Performance:** „Unsupervised CLE outperforms supervised CLE“
 - *„Without using any character information, our model even outperforms existing supervised methods on cross-lingual tasks for some language pairs“*
 - *„Our method succeeds in all tested scenarios and obtains the best published results in standard datasets, even surpassing previous supervised systems“*
 - *„...our method achieves better performance than recent state-of-the-art deep adversarial approaches and is competitive with the supervised baseline“*
- **Unintuitive:** unsupervised CLE models all solve Procrustes problem in the final step, only on the less reliable (automatically induced) **D**

Content

- **Cross-Lingual Word Embeddings**
 - Joint Training (from Scratch)
 - Projection-Based CLWEs
 - Unsupervised Induction of CLWEs
- **Evaluation of CLWEs**

Evaluation of CLWEs



Glavaš, G., Litschko, R., Ruder, S., & Vulić, I. (2019, July). [How to \(Properly\) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions](#). In Proceedings of ACL (pp. 710-721).

- **Intrinsic evaluation**

- Bilingual Lexicon Induction (BLI)
- Cross-Lingual Word Similarity (XL-SIM)

- **Extrinsic evaluation:**

- Cross-lingual transfer in downstream NLP tasks (e.g., text classification)
- More in [Lecture 6](#) 😊

Evaluation of CLWEs

- **Bilingual Lexicon Induction**

- Essentially the same task as in „training”: word translation
- Given a test dictionary $D_{\text{test}} = \{(w_{L1}^k, w_{L2}^k)\}_k$ and a bilingual embedding space $\mathbf{X}_{L1,L2}$ (for projection-based CLWEs $\mathbf{X}_{L1,L2} = \mathbf{X}_{L1}\mathbf{W} \cup \mathbf{X}_{L2}$)
- For w_{L1}^k with vector \mathbf{x}_{L1} as „query”, we rank all $\mathbf{x} \in \mathbf{X}_{L2}$ based on similarity with \mathbf{x}_{L1} : let r be the rank at which we find the vector \mathbf{x}_{L2}^j of the translation w_{L2}^j
- Two common performance measures:
 - Precision@1 (P@1): percentage of pairs (out of k) for which $r = 1$
 - Mean reciprocal rank (MRR): average of $1/r$ (across all k pairs)

Evaluation of CLWEs



Vulić, I., Baker, S., Ponti, E. M., Petti, U., Leviant, I., Wing, K., ... & Korhonen, A. (2020). [Multi-simlex: A large-scale evaluation of multilingual and cross-lingual lexical semantic similarity](#). Computational Linguistics, 46(4), 847-897.

- **Cross-Lingual Word Similarity**

- Evaluate CLWEs the same way we evaluate monolingual word embeddings
- Given two words, w_{L1} , w_{L2} measure the similarity of their vectors
 - E.g., $CSLS(\mathbf{x}_{L1}, \mathbf{x}_{L2})$
- Compare embedding similarities against human judgments of semantic similarity for pairs of words
 - Performance measure: Spearman correlation (of two sets of scores)
- XL-SIM: pairs of words from different languages
 - Need bilingual human annotators
 - Subjective task: need multiple annotators (average their scores)

Unsupervised CLWEs: Revisited



Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019). [Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?](#) In Proceedings of the EMNLP (pp. 4407-4418).

- **Performance:** „Unsupervised CLE outperforms supervised CLE“
 - *„Without using any character information, our model even outperforms existing supervised methods on cross-lingual tasks for some language pairs“*
 - *„Our method succeeds in all tested scenarios and obtains the best published results in standard datasets, even surpassing previous supervised systems“*
 - *„...our method achieves better performance than recent state-of-the-art deep adversarial approaches and is competitive with the supervised baseline“*
- **Unintuitive:** unsupervised CLE models all solve Procrustes problem in the final step, only on the less reliable (automatically induced) **D**

Unsupervised CLWEs: Revisited



Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019). [Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?](#) In Proceedings of the EMNLP (pp. 4407-4418).

- **Unintuitive:** unsupervised CLWE models all solve Procrustes problem in the final step, only on the less reliable (automatically induced) **D**
- Performance of **unsupervised CLWE models*** depends on the extent to which the monolingual embedding spaces **X_{L1}** and **X_{L2}** have the „same shape“ (isomorphism)
 - Good between close and high-resource languages
 - E.g., **EN-DE**, **EN-ES**, **EN-IT**, ...
 - **Q:** What about low-resource and distant languages?

Unsupervised CLWEs: Revisited



Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019). [Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?](#) In Proceedings of the EMNLP (pp. 4407-4418).

- Wider evaluation:
 - 15 languages
(210 BLI evaluations)

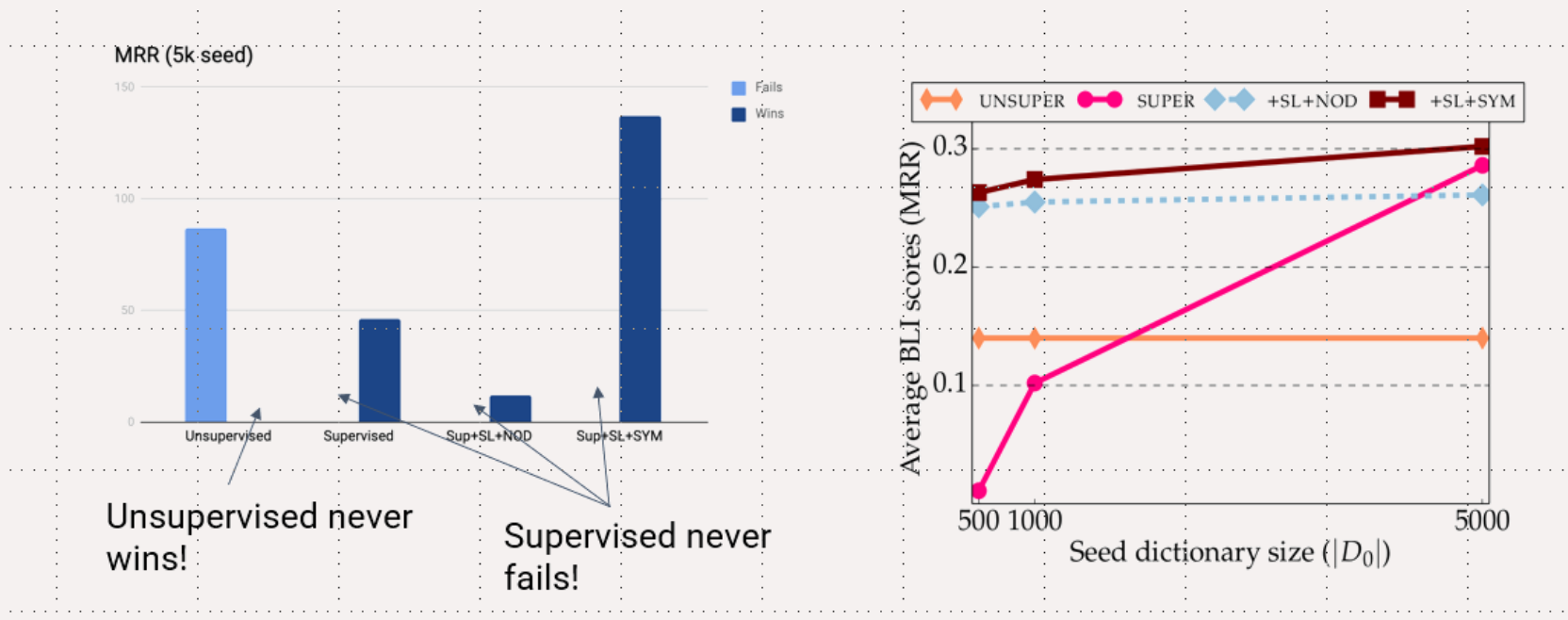
Language	Family	Type	ISO 639-1
Bulgarian	IE: Slavic	fusional	BG
Catalan	IE: Romance	fusional	CA
Esperanto	– (constructed)	agglutinative	EO
Estonian	Uralic	agglutinative	ET
Basque	– (isolate)	agglutinative	EU
Finnish	Uralic	agglutinative	FI
Hebrew	Afro-Asiatic	introflexive	HE
Hungarian	Uralic	agglutinative	HU
Indonesian	Austronesian	isolating	ID
Georgian	Kartvelian	agglutinative	KA
Korean	Koreanic	agglutinative	KO
Lithuanian	IE: Baltic	fusional	LT
Bokmål	IE: Germanic	fusional	NO
Thai	Kra-Dai	isolating	TH
Turkish	Turkic	agglutinative	TR

Unsupervised CLWEs: Revisited



Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019). [Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?](#) In Proceedings of the EMNLP (pp. 4407-4418).

- Wider evaluation: 15 language (210 BLI evaluations)





The End

Image: Alexander Mikhailchik