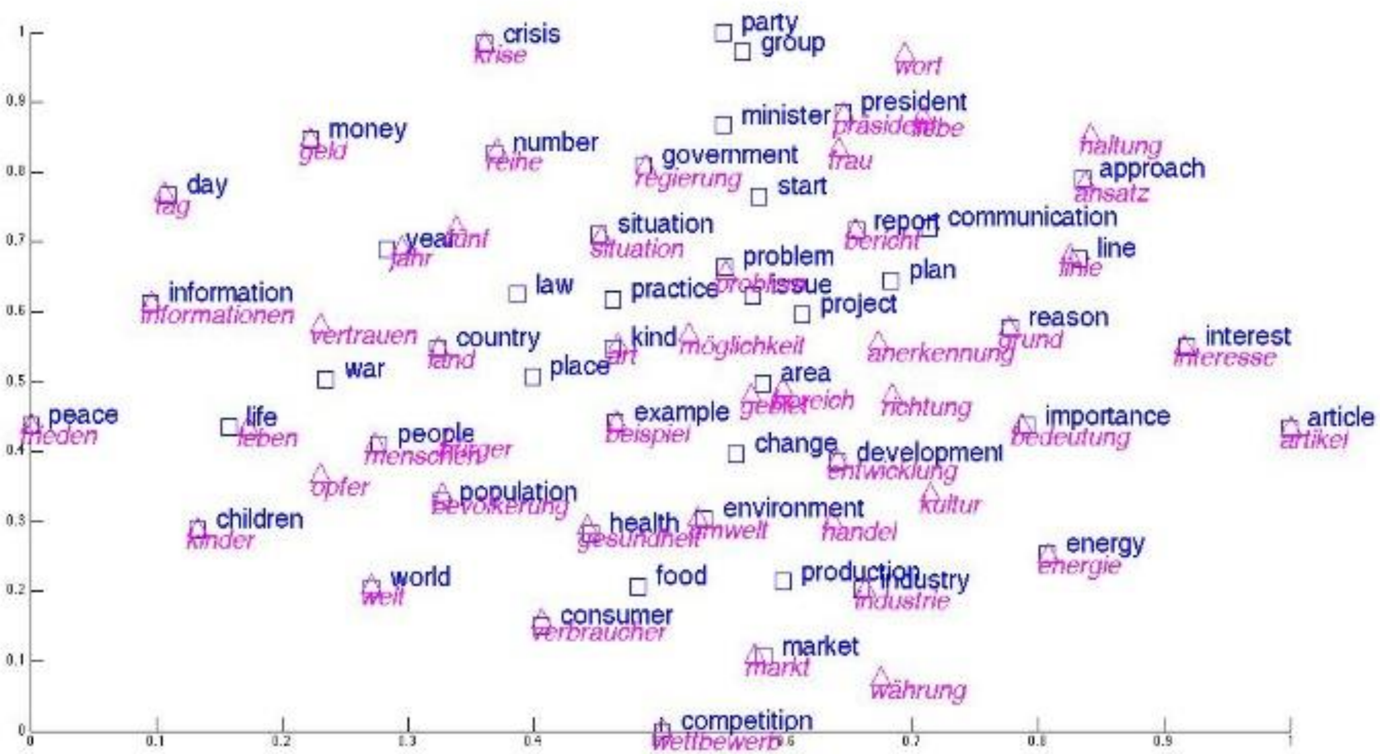
Cross-Lingual (Word) Embeddings (CLWE)

Different methodologies but the same **end goal**:

Induce a semantic vector space in which words with similar meaning end up with similar vectors, regardless of whether they come from the same language or from different languages.



Cross-Lingual (Word) Embeddings (CLWE)

Typology of methods for inducing CLWEs

1. **Type of bilingual / multilingual signal**
   - Document-level, sentence-level, word-level, no signal (i.e., unsupervised)

2. **Comparability**
   - Parallel texts, comparable texts, not comparable (i.e., randomly aligned)

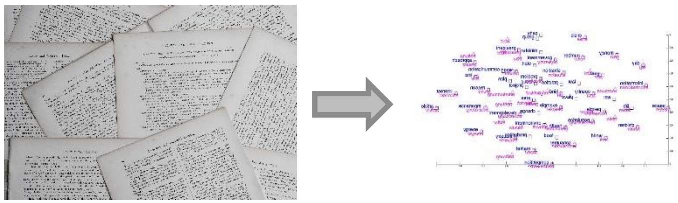
3. **Point (time) of alignment**
   - *Joint embedding models* vs. *Post-hoc alignment*

4. **Modality**
   - Text only vs. using images for alignment (e.g., [Kiela et al., ‘15])

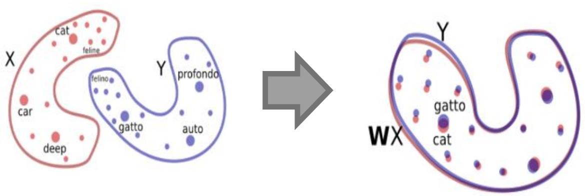
Joint Models vs. Post-hoc alignment

Regardless of the source of supervision, there are two main strategies for inducing a bilingual/multilingual word embedding space:

1. **Joint embedding models**
   - Start from raw texts in two (or more) languages
   - Induce a bilingual (multilingual) space from scratch



2. **Post-hoc alignment models** (aka *projection* models)
   - Start from two independently pretrained monolingual embedding spaces
     - E.g., We apply word2vec on EN Wikipedia; then (independently) on ES Wikipedia
   - Learn the alignment/projection between the two monolingual spaces



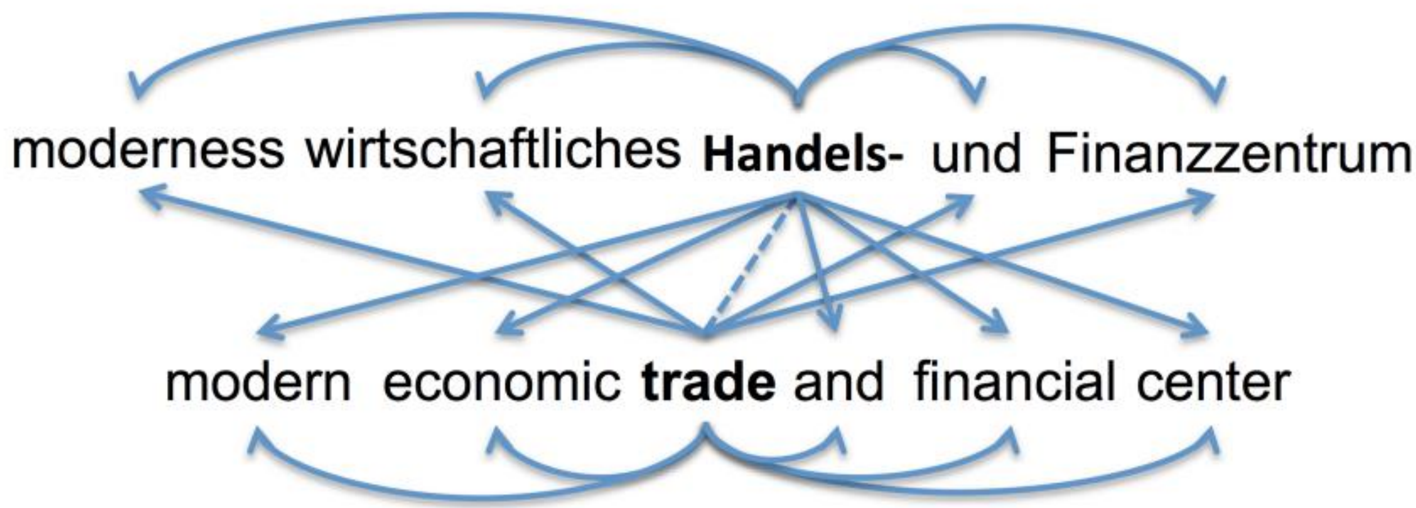
Joint CLWE induction

– A number of models

– Example: **Bilingual Skip-Gram**

  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).

– Skip-Gram extended with cross-lingual context prediction
  - Parallel data (mutual sentence translations) needed!
    - Automatic word alignment



Joint CLWE alignment

Some shortcomings

- Expensive model training for every pair of languages
- Bilingual models – not multilingual models
- Bilingual models not comparable, e.g., EN-DE vs. EN-ES
- Parallel sentences not so easily obtainable for all language pairs
  - Although there are extensions of Bilingual Skip-Gram that require only word-level supervision (i.e., word translations)

More elegant and less-resource demanding solution:

- Train monolingual vectors independently
- Light-weight post-hoc alignment between those spaces?
- Easy to induce truly multilingual spaces through post-hoc projections
- **Projection-based CLWE models**

Post-hoc embedding alignment

- Monolingual embeddings **independently** trained
  - Can be trained even with different embedding algorithms, e.g., GloVe vs. SkipGram
- Post-hoc aligning monolingual spaces

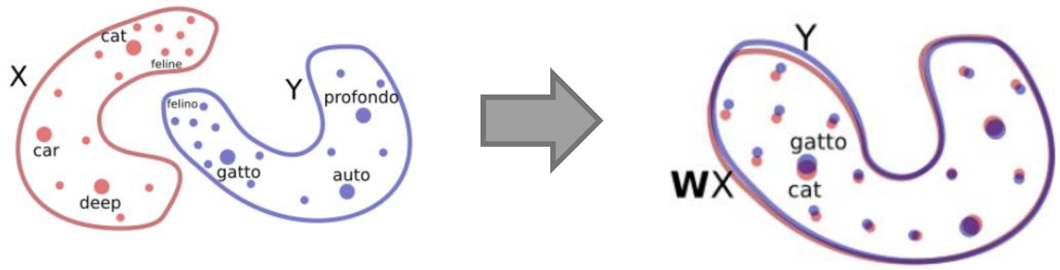


Image from [Conneau et al., '18]

- **X** is dist. space of L1, **Y** of L2
  - In general, we are looking for functions *f* and *g* that produce a meaningful bilingual embedding space $f(\mathbf{X}) \cup g(\mathbf{Y})$

Projection-Based CLWE

– **Post-hoc** alignment of independently trained monolingual distributional word vector spaces
  - Alignment based on **word translation pairs** (dictionary **D**)
  - Supervised models use pre-obtained **D**, unsupervised automatically induce **D**

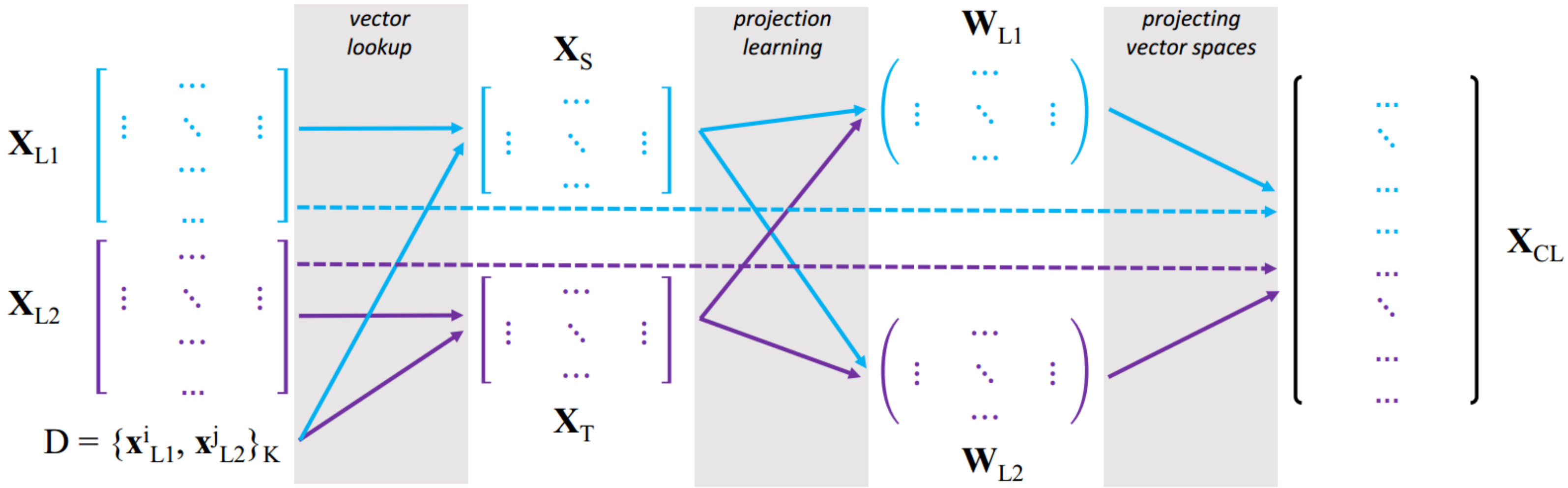


Image from [Glavaš et al., ACL '20]

Projection-Based CLWE

– Most models learn a single projection matrix $\mathbf{W}_{L1}$ (i.e., $\mathbf{W}_{L2} = \mathbf{I}$)

$$\begin{array}{c} \\ \text{bird} \\ \text{pretty} \\ \text{...} \\ \text{eat} \end{array} \overset{\mathbf{X}_S}{\begin{bmatrix} -1.18 & 0.21 & \dots & 0.11 \\ 0.23 & -0.53 & \dots & 0.34 \\ \dots & \dots & \dots & \dots \\ 0.78 & 1.33 & \dots & -0.47 \end{bmatrix}} \mathbf{W}_{L1} = \overset{\mathbf{X}_T}{\begin{bmatrix} 0.59 & 1.01 & \dots & 0.37 \\ -0.34 & -0.27 & \dots & 0.41 \\ \dots & \dots & \dots & \dots \\ 0.81 & -0.31 & \dots & 0.29 \end{bmatrix}} \begin{array}{c} \\ \text{Vogel} \\ \text{schön} \\ \text{...} \\ \text{essen} \end{array}$$

– How do we find the „optimal" projection matrix $\mathbf{W}_{L1}$?
  - We minimize the **mean square distance**

Minimizing Euclidean Distance

- Minimize the Euclidean distances for translation pairs after projection

$$\mathbf{W}_{L1} = \arg\min_{\mathbf{W}} \| \mathbf{X}_{\mathrm{S}} \mathbf{W} - \mathbf{X}_{\mathrm{T}} \|_2$$

- The optimization problem has no closed-form solution
  - SGD-based iterative optimization
- More complex mapping – DFFN instead of linear projection matrix yields worse performance
- Better (word translation) results when $\mathbf{W}_{L1}$ is constrained to be orthogonal

Solving the Procrustes Problem

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

– If **W** is orthogonal, the above optimization problem is the so-called **Procrustes problem** with a closed-form solution

$$\mathbf{W}_{L1} = \mathbf{U}\mathbf{V}^\top, \text{ with}$$
$$\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = SVD(\mathbf{X}_T {\mathbf{X}_S}^\top)$$

– All projection-based CLWE models, *supervised* and *unsupervised*, solve the Procrustes problem in the final step
- **Supervised**: clean, prepared word-translation dictionary (e.g., 5K entries)
- **Unsupervised**: initial translation dictionary automatically induced

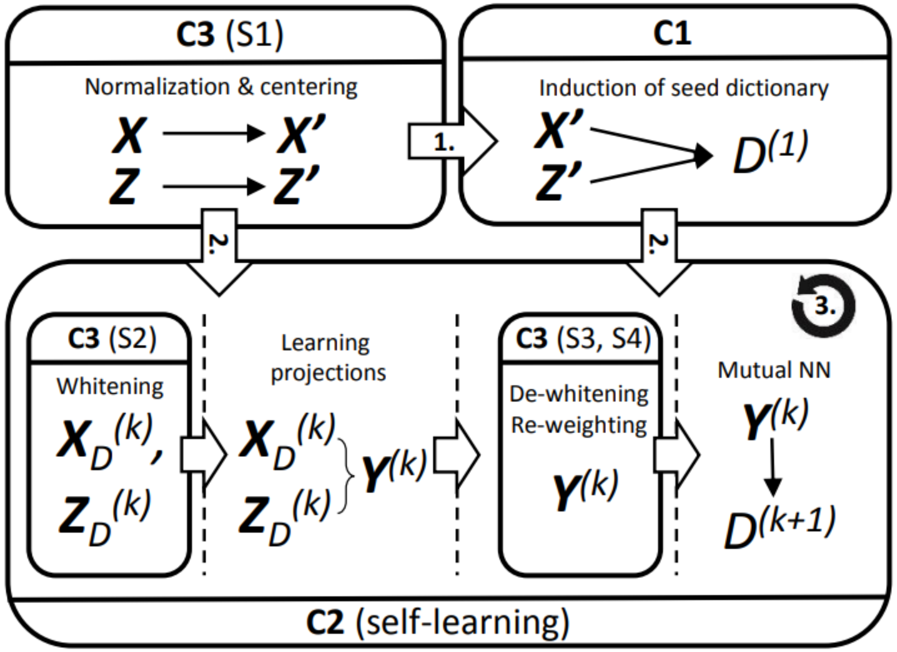
Unsupervised CLWE induction framework

The **same general framework** for all unsupervised CLWE models

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

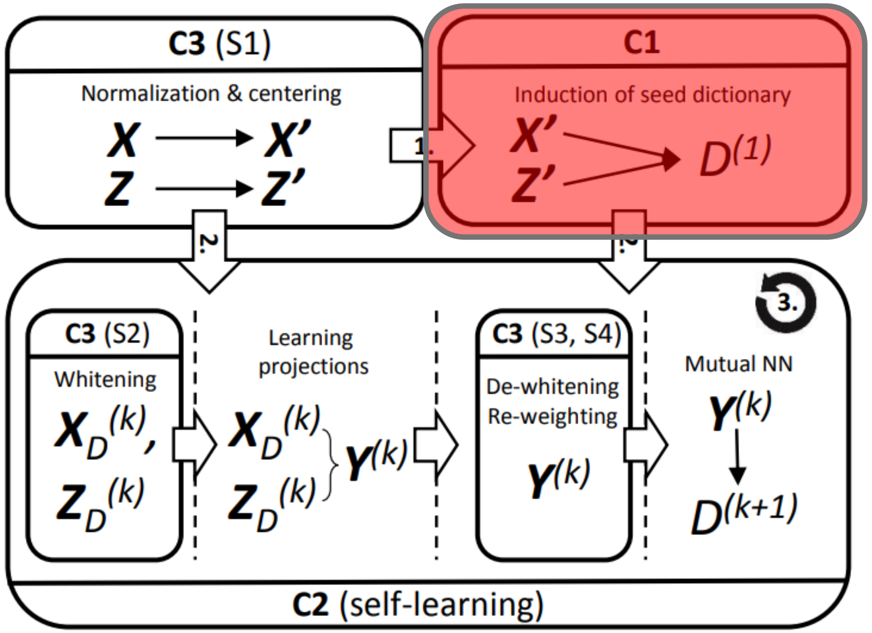
Repeat:

2. Learn the projection(s) using $\mathbf{D}^{(k)}$
3. Induce new dictionary $\mathbf{D}^{(k+1)}$ from the shared space $\mathbf{Y}^{(k)}$



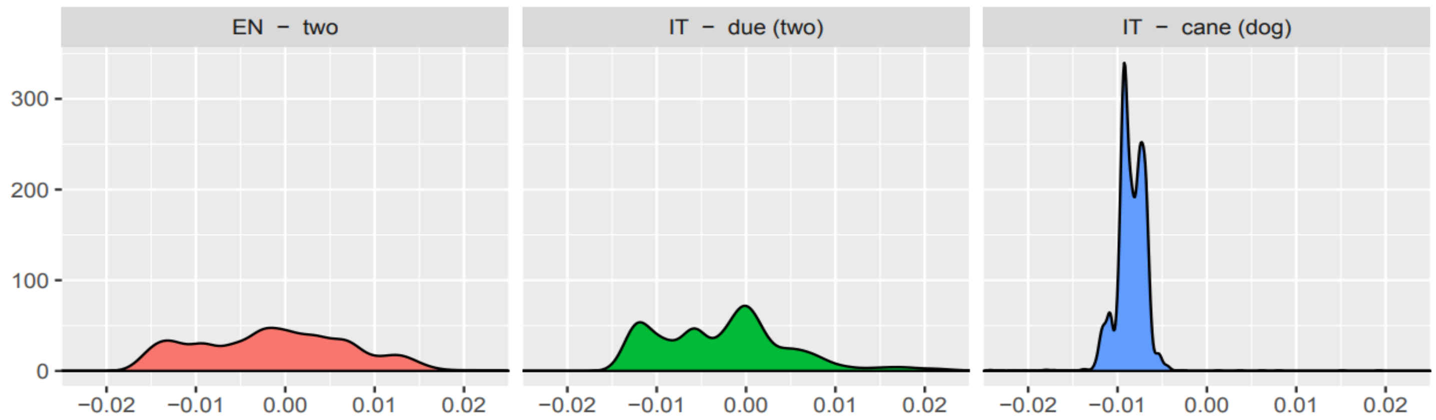
Unsupervised CLWE induction

- The **same general framework** for all unsupervised CLWE models
- Different approaches for step **C1**, i.e., inducing the initial dictionary $\mathbf{D}^{(1)}$:
  - Adversarial learning [Conneau et al., '18]
  - Similarities of similarity distributions [Artetxe et al., 2018]
  - PCA [Hoshen & Wolf, '18]
  - Solving optimal transport problem [Alvarez-Melis & Jaakkola, '18]
  - ...
- All assume (approximate) isomorphism of monolingual spaces!



Unsupervised CLWE: Example

– **VecMap** [Artetxe et al., 2018]
  - **Heuristic induction** of the initial word translation dictionary $\mathbf{D}^{(1)}$
    - Word with similar meanings will have similar monolingual similarity distributions (i.e., distributions of similarity across all words of the same lang.)



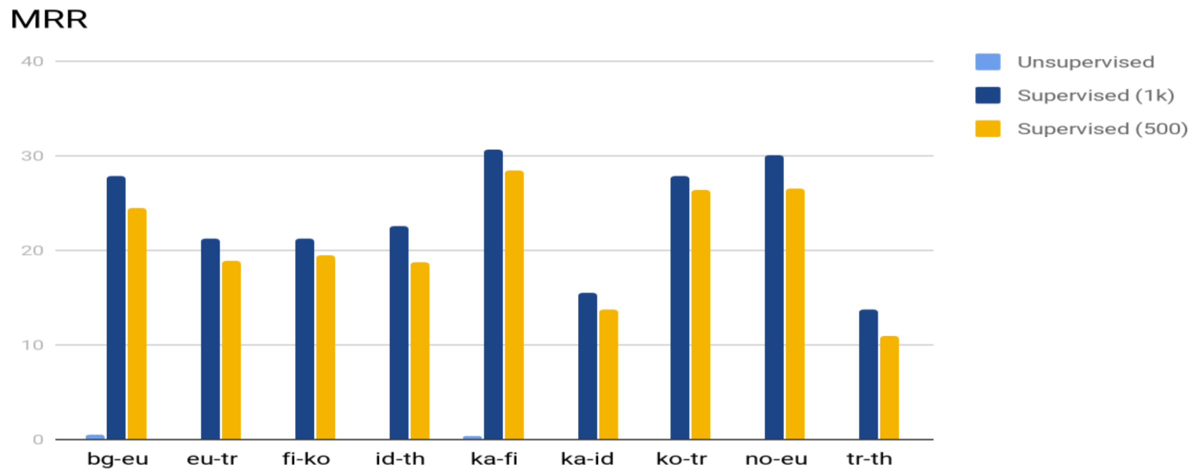
Why Unsupervised CLWE induction?

- **Original motivation**:
  - Does not require any bilingual/multilingual supervision, thus suitable for under-resourced languages
- **However...**
  1. Assumptions on which the automatic induction of an initial dictionary is based (approximate isomorphism of monolingual spaces) do not hold for
     - Pairs of etymologically and typologically distant languages
  2. Assumption that we „cannot find” clean word translations for low-resource languages is simply false
     - **PanLex –** a lexico-semantic resource covering 9000+ languages and dialects
     - For all languages some lexical alignment with other langs (for most with EN)
  3. Language „**X**” – no word translations to any other language
     - Then you probably don't have enough digital texts in X to induce reliable monolingual **X** embeddings in the first place

CLWEs – Evaluation

– Common evaluation: **Bilingual Lexicon Induction (BLI)**
  - Word translation task
  - Given a translation pair (w_s, w_t), rank all the words in the target language according to vector similarity with w_s and find where w_t is in the ranking

– Supervised vs. unsupervised CLWEs for low-resource setups
  - Vulić, I., Glavaš, G., Reichart, R., & Korhonen, A. (2019, November). *Do We Really Need Fully Unsupervised Cross-Lingual Embeddings?* In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4398-4409).



Cross-Lingual Transfer with CLWEs

Use CLWEs for cross-lingual transfer of supervised NLP tasks?

Assumption: **zero-shot transfer**

- Only task- annotated data for the source language L_S, no annotated data in target language L_T

Steps:

1. Induce the bilingual shared word embedding space $\mathbf{X}_{TS}$
   - E.g., by projecting the target lang. space $\mathbf{X}_S$ to the source lang. Space $\mathbf{X}_T$
2. Train the (neural) model using the task-specific data in Ls
   - E.g., for *Named Entity Recognition*, train a *Bi-LSTM+classifier* using embeddings of source language words from the shared space $\mathbf{X}_{TS}$ as input
3. At prediction time, for texts in target language L_T, feed as input the embeddings of target language words from the same shared space $\mathbf{X}_{TS}$