Prof. Dr. Goran Glavaš

Lecture 12: Multilinguality and Cross-Lingual Transfer

Learning outcomes

- After today's lecture, you'll...
  1. Understand what is multilingual NLP and why we need it
  2. Know the mechanisms for inducing multilingual representations spaces
     - Cross-lingual word embeddings (CLWEs)
     - Massively multilingual transformers (MMTs)
  3. Understand how to use multilingual representations spaces for CL transfer

Outline

1. Why Multilingual NLP?
2. Cross-lingual word embeddings
3. Multilingual transformers

Why Multilingual NLP?

– Because we want to **understand** and **model** the meaning of texts in...

[Image from: epthinktank.eu]

– ...without manual (i.e., human) input and without perfect MT!

– **How many different languages are there in the world?**

- How many have more than 10M speakers?

4

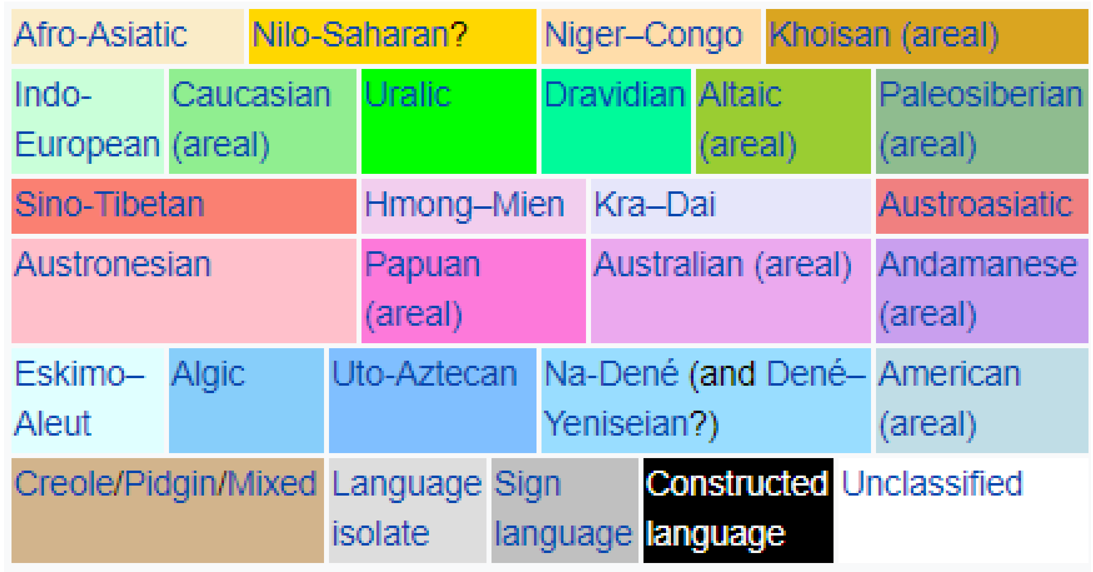
Why Multilingual NLP?

– According to Ethnologue (2020) there are 7,117 **living languages**

70	Hejazi Arabic	14.5	0.188%	Afroasiatic	Semitic
71	Nigerian Fulfulde	14.5	0.188%	Niger–Congo	Senegambian
72	Bavarian	14.1	0.183%	Indo-European	Germanic
73	South Azerbaijani	13.8	0.179%	Turkic	Oghuz
74	Greek	13.1	0.170%	Indo-European	Hellenic
75	Chittagonian	13.0	0.169%	Indo-European	Indo-Aryan
76	Kazakh	12.9	0.168%	Turkic	Kipchak
77	Deccan	12.8	0.166%	Indo-European	Indo-Aryan
78	Hungarian	12.6	0.164%	Uralic	Ugric
79	Kinyarwanda	12.1	0.157%	Niger–Congo	Bantu
80	Zulu	12.1	0.157%	Niger–Congo	Bantu
81	South Levantine Arabic	11.6	0.151%	Afroasiatic	Semitic
82	Tunisian Arabic	11.6	0.151%	Afroasiatic	Semitic
83	Sanaani Spoken Arabic	11.4	0.148%	Afroasiatic	Semitic
84	Min Bei Chinese	11.0	0.143%	Sino-Tibetan	Sinitic
85	Southern Pashto	10.9	0.142%	Indo-European	Iranian
86	Rundi	10.8	0.140%	Niger–Congo	Bantu
87	Czech	10.7	0.139%	Indo-European	Balto-Slavic
88	Ta'izzi-Adeni Arabic	10.5	0.136%	Afroasiatic	Semitic
89	Uyghur	10.4	0.135%	Turkic	Karluk
90	Min Dong Chinese	10.3	0.134%	Sino-Tibetan	Sinitic
91	Sylheti	10.3	0.134%	Indo-European	Indo-Aryan

Language variety

– **Language family**: group of languages that originate from the same *ancestral/parental* language (proto-language)



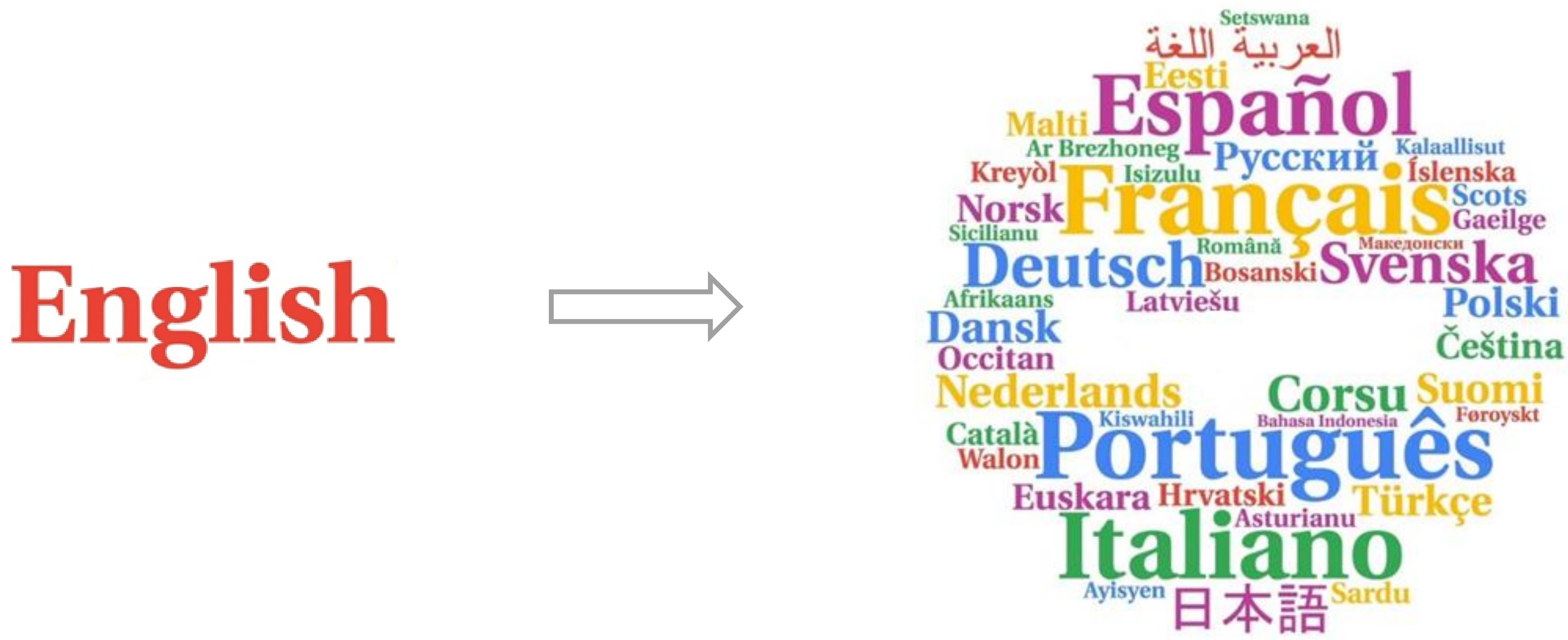
[Image from: Wikipedia]

– **Language isolates**: no known/demonstrable genealogical relationship with any other language:

- *Basque, Korean*
- Indo-European language isolates: *Albanian*, *Armenian*, *Greek*

Why Cross-Lingual NLP?

– Because we want to transfer supervised models for NLP tasks...
  - Trained on **annotated datasets** we have in **resource-rich languages**
  - Make predictions in resource-lean target languages



Language Transfer

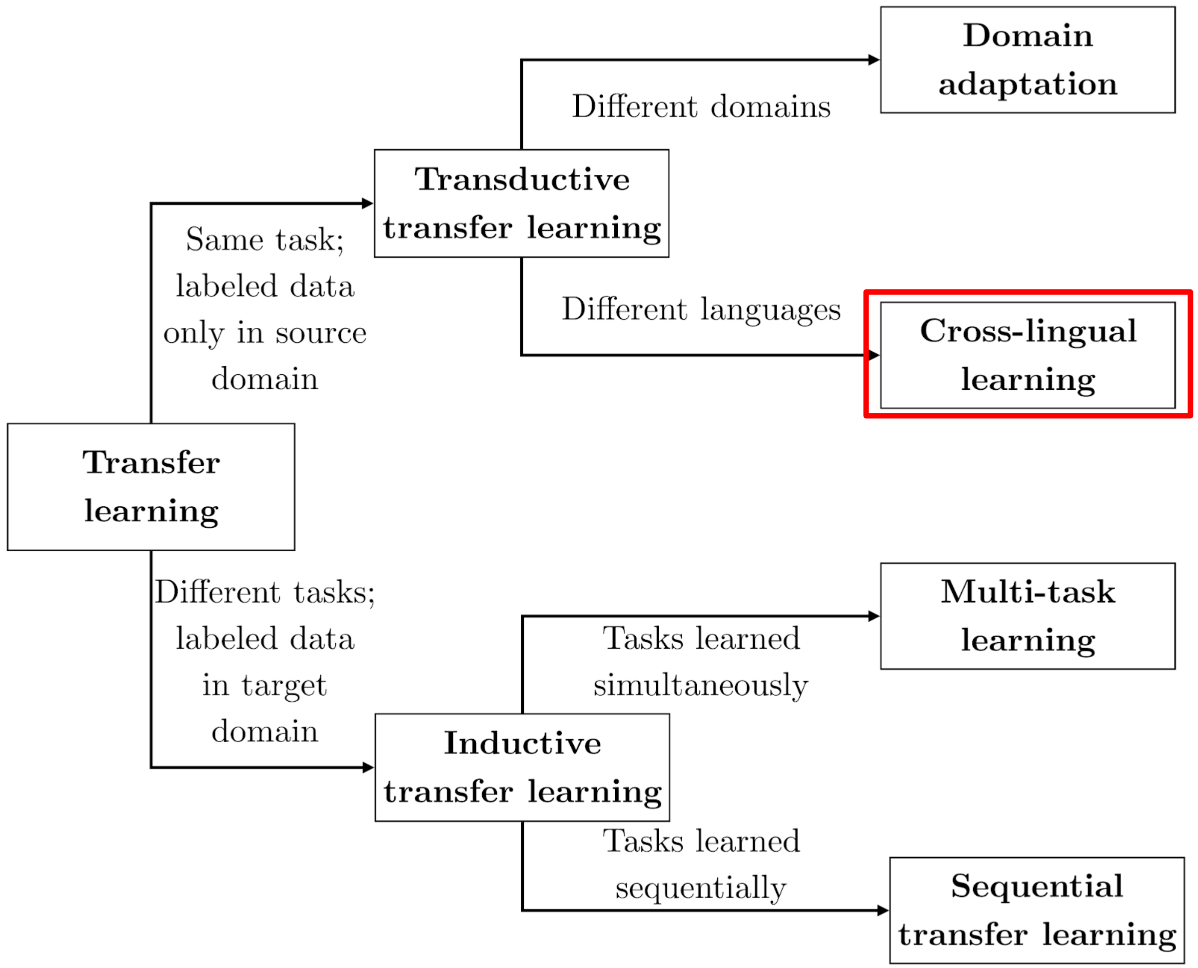
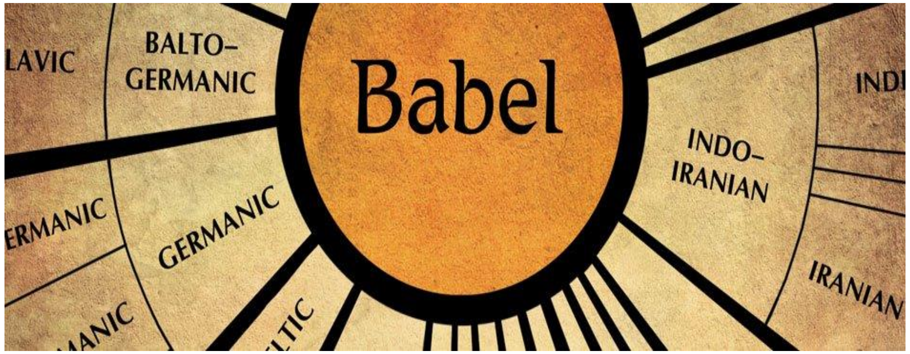


Image from [Ruder, 2019]

Crossing the Language Chasm

– **Old paradigm**:
  - Language-specific NLP models
  - Language-specific feature computation (i.e., preprocessing)

– **New paradigm(s)**:
  - Representation learning: semantic vectors (embeddings)
  - Multilingual / cross-lingual representation learning



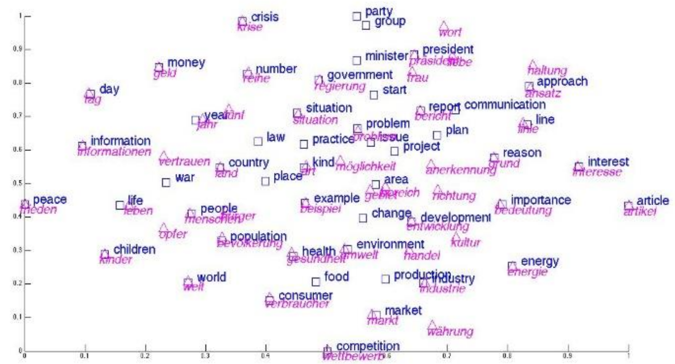
Crossing the Language Chasm: symbolic approaches

1. Full-Blown MT (SMT or NMT)
    - Parallel data needed, critical for under-resourced languages
    - Translate everything from the target language to the source language
2. Multilingual KBs
    - Texts represented using entities from a multilingual KB
    - Same entity ID for same concepts across languages
    - Issues: coverage, entity linking

BabelNet 2.0
A very large multilingual **encyclopedic dictionary** and **ontology**

Crossing the Language Chasm: representation learning

3. Multilingual / Cross-lingual representations of meaning



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

- **Text encoding**
  - Multilingual unsupervised pretraining
    - Multilingual BERT [Devlin et al., ‘19]
    - XLM(-R) [Conneau & Lample, ’19, Conneau et al., 2020]
    - mT5 [Xue et al., 2020]

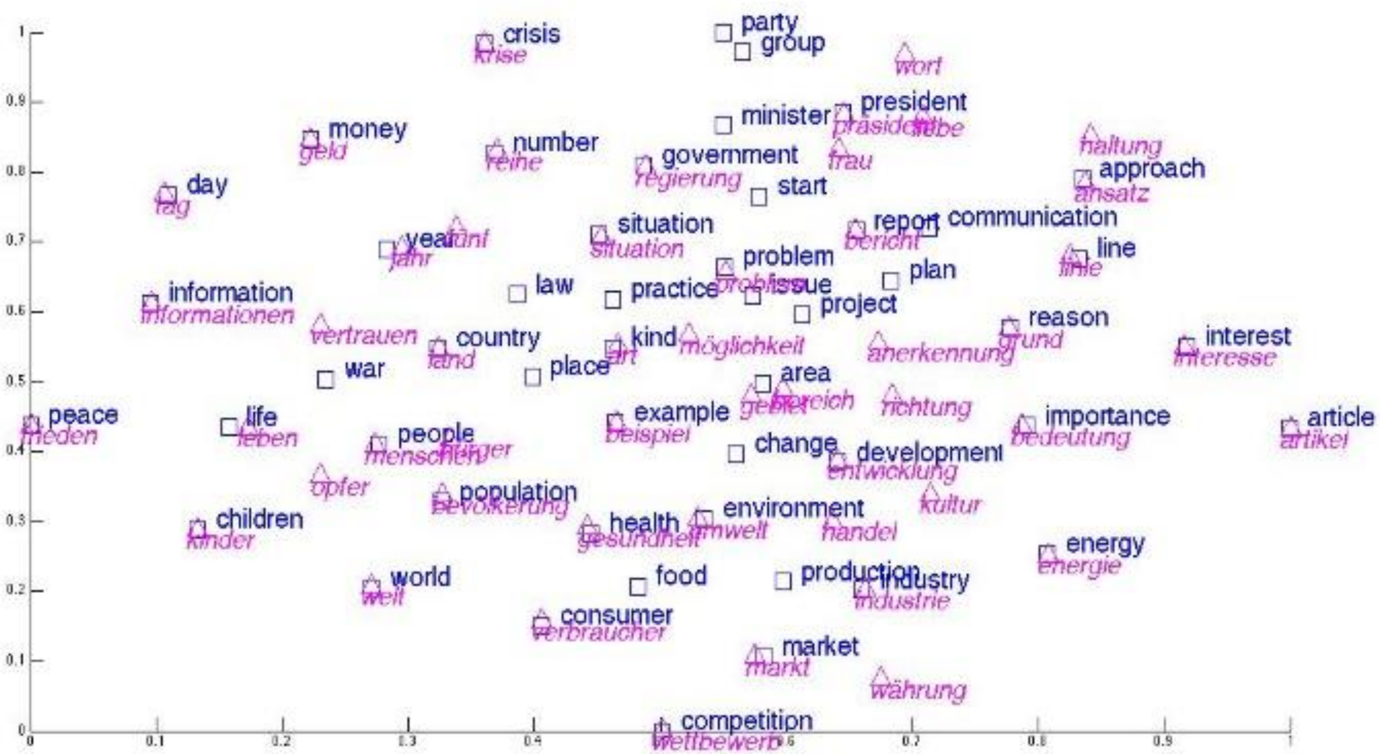
Outline

1. Why Multilingual NLP?
2. Cross-lingual word embeddings
3. Multilingual transformers

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*

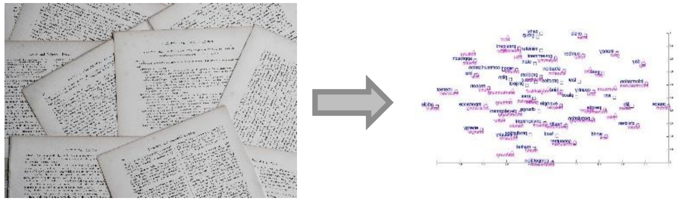
4. **Modality**
   - Text only vs. using images for alignment (e.g., [Kiela et al., ‘15])

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:

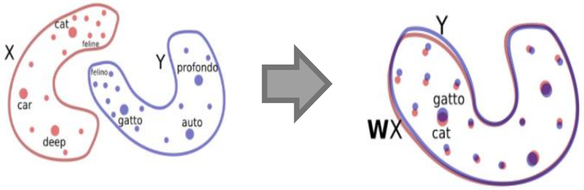
1. Joint embedding models

- 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



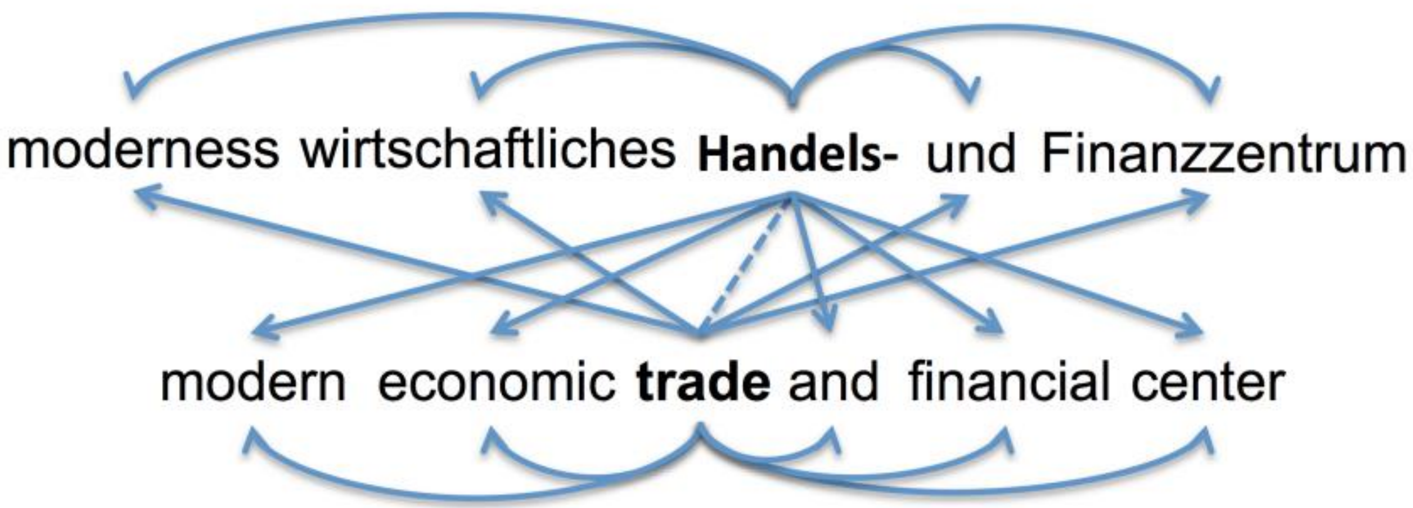
Joint CLWE induction

– 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

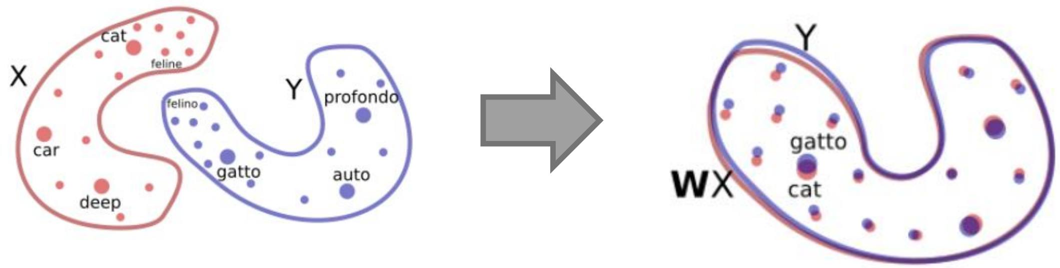


Image from [Conneau et al., '18]

– **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**

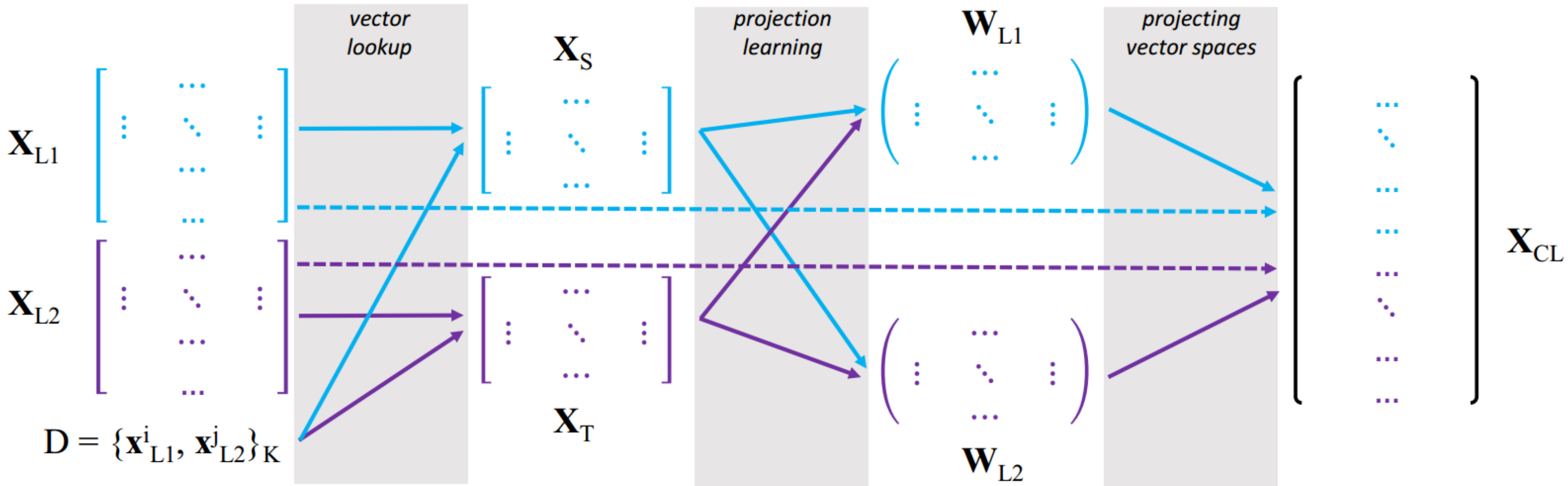


Image from [Glavaš et al., ACL '20]

Projection-Based CLWE

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

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

– 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

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

– 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

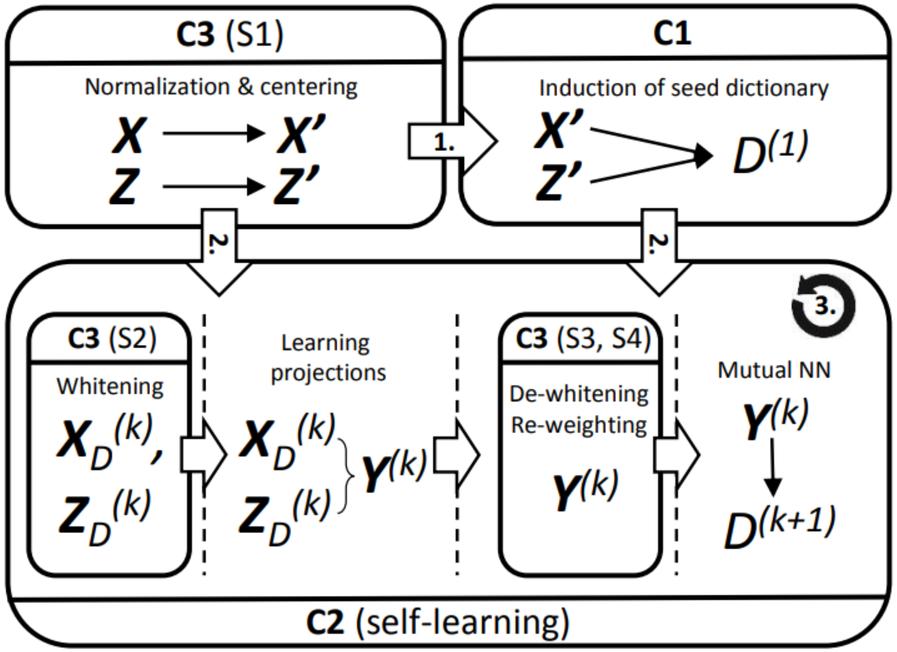
Unsupervised CLWE induction framework

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

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

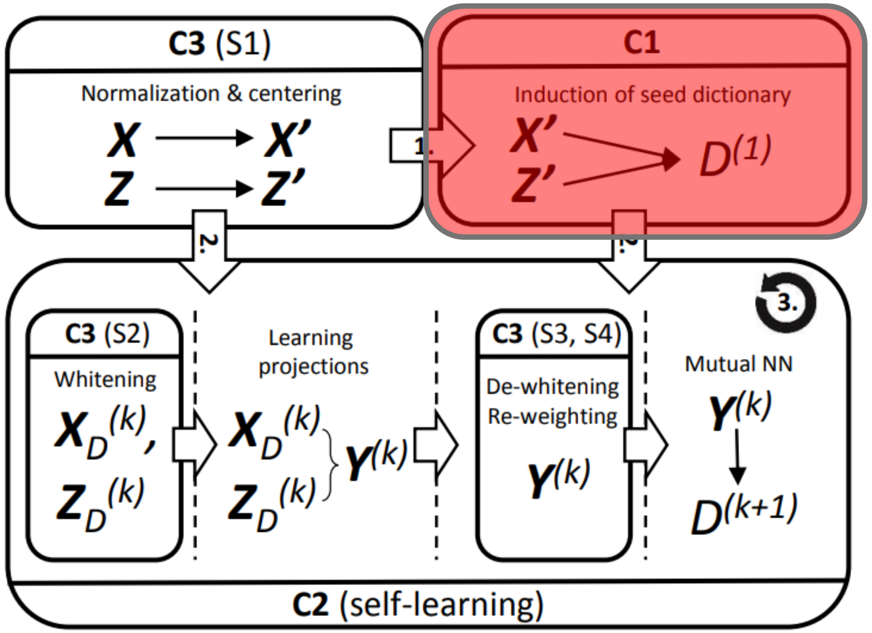
Repeat:

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



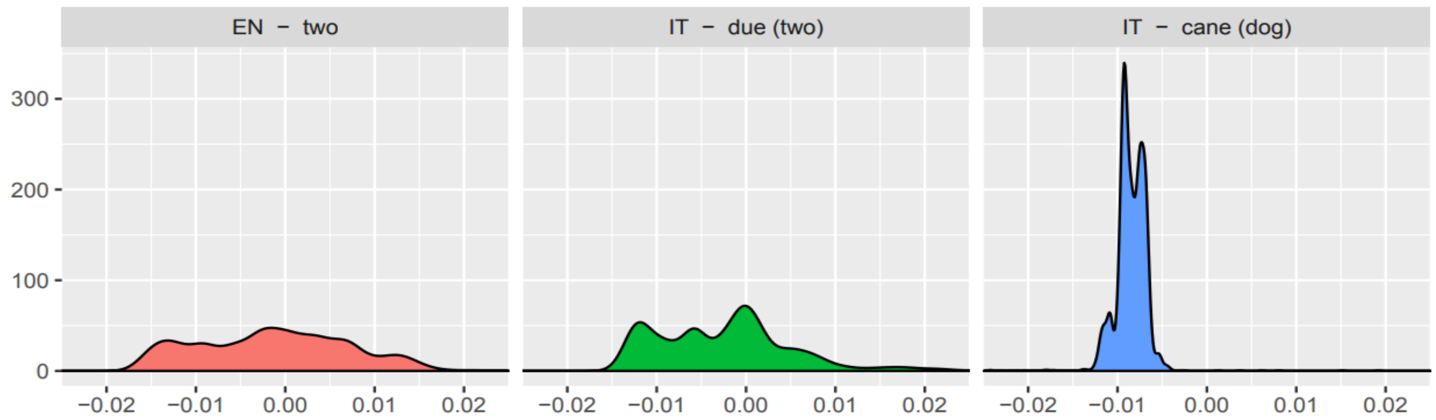
Unsupervised CLWE induction

- 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

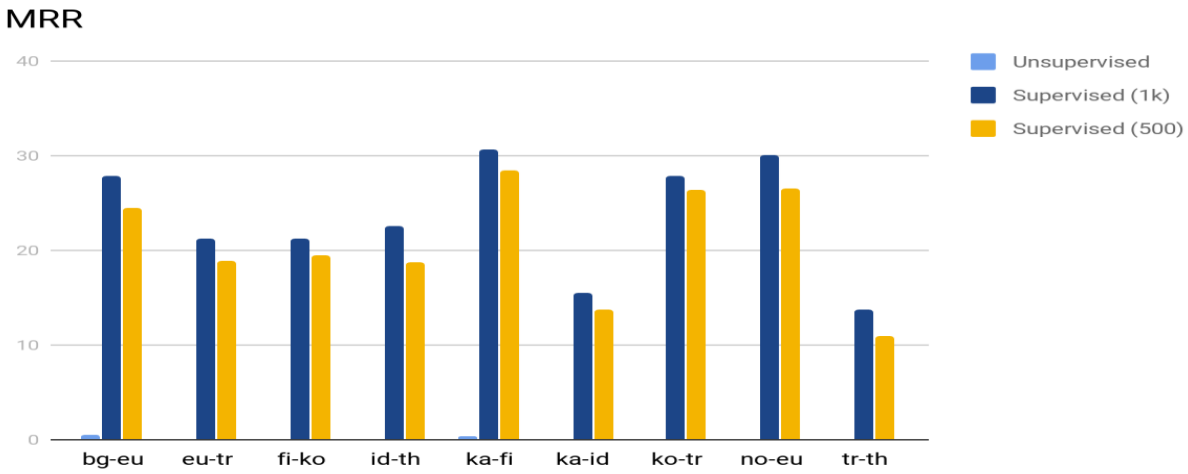
2. Assumption that we „cannot find” clean word translations for low-resource languages is simply false
   - **PanLex –** a lexico-semantic resource covering 9000+ languages and dialects
   - For all languages some lexical alignment with other langs (for most with EN)

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 $\mathbf{X}_{TS}$

Outline

1. Why Multilingual NLP?
2. Cross-lingual word embeddings
3. (Massively) Multilingual transformers

Massively Multilingual Transformers

- Deep Transformer nets pretrained on large **multilingual corpora** via (masked) **language modeling** objectives
  - Multilingual BERT, XLM-R, mT5

- Automatically induces shared (subword) vocabulary across all languages

- **Unsupervised** from the perspective of explicit cross-lingual signal
  - Deemed very effective for zero-shot CL transfer

„Suprising cross-lingual effectiveness of BERT"

„mBERT surprisingly good at zero-shot CL model transfer"

Massively Multilingual Transformers

- **Assumption**: after multilingual MLM pretraining mBERT can encode text from any of the languages seen in pretraining
- Automatically lends itself to **zero-shot language transfer** for downstream NLP tasks

- **mBERT** has its own *tokenizer* that can tokenize input texts from all languages seen in pre-training
  - Caveat: words from larger languages mostly have their own tokens
    - Words from smaller languages broken down into subwords which can be found across languages
    - Worst case scenario: input broken into characters

Cross-Lingual Transfer with MMTs

Zero-shot language transfer for downstream NLP tasks with mBERT:

1. Couple the mBERT Transformer with the task-specific classifier („head”)
2. Train the mBERT+classifier model jointly on source language data
   - Classifier parameters trained from scratch
   - mBERT’s Transformer parameters fine-tuned
3. Predict by feeding the target language text (tokenized with mBERT’s tokenizer) into the fine-tuned mBERT+classifier model

So...has mBERT solved zero-shot CL transfer?

– **No!** Settings in which they were evaluated were simply too favorable

„How multilingual is Multilingual BERT?" [Pires et al., ACL 19]

- **Tasks**: NER, POS; **Target languages**: DE, NL, ES

„Cross-lingual Ability of mBert: Empirical Study" [Karthikeyan et al., ICLR 20]

- **Tasks**: NER, NLI; **Target languages**: ES, HI, RU

– In most studies, the selected target languages were:

- (1) from the **same language family**,
- (2) with **large corpora in pretraining**

Zero-shot transfer performance drops

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). *From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4483-4499).

Task	Model	EN	ZH Δ	TR Δ	RU Δ	AR Δ	HI Δ	EU Δ	FI Δ	HE Δ	IT Δ	JA Δ	KO Δ	SV Δ	VI Δ	TH Δ	ES Δ	EL Δ	DE Δ	FR Δ	BG Δ	SW Δ	UR Δ
DEP	B	91.2	-43.9	-46.0	-28.1	-56.4	-36.1	-50.2	-30.7	-36.1	-17.1	**-60.1**	-56.1	-14.3	-	-	-	-	-	-	-	-	-
	X	92.0	**-85.4**	-44.2	-29.7	-54.6	-39	-49.5	-26.7	-39	-23.5	-80.5	-56.0	-16.3	-	-	-	-	-	-	-	-	-
POS	B	95.8	-38.0	-35.9	-16.0	-40.1	-33.4	-34.6	-21.9	-33.4	-19.8	**-46.1**	-42.0	-9.6	-	-	-	-	-	-	-	-	-
	X	96.3	-69.2	-27.7	-14.3	-37.1	-27.3	-31.9	-17.9	-27.3	-19.0	**-77.0**	-37.3	-10.7	-	-	-	-	-	-	-	-	-
NER	B	92.4	-23.3	-11.6	-10.7	**-31.7**	-11.1	-12.8	-3.8	-11.1	-2.6	-25.7	-13.8	-6.7	-	-	-	-	-	-	-	-	-
	X	91.6	**-34.8**	-6.2	-13.7	-24.6	-16.5	-8.0	-0.9	-16.5	-2.4	-30.1	-15.6	-2.2	-	-	-	-	-	-	-	-	-
XNLI	B	82.8	-13.6	-20.6	-13.5	-17.3	-21.3	-	-	-	-	-	-	-	-11.9	-28.1	-8.1	-14.1	-10.5	-7.8	-13.3	**-33.0**	-23.4
	X	84.3	-11.0	-11.3	-9.0	-13.0	-14.2	-	-	-	-	-	-	-	-9.7	-12.3	-5.8	-8.9	-7.8	-6.1	-6.6	**-20.2**	-17.3
XQuAD	B	71.1	-22.9	-34.2	-19.2	-24.7	-28.6	-	-	-	-	-	-	-	-22.1	**-43.2**	-16.6	-28.2	-14.8	-	-	-	-
	X	72.5	**-26.2**	-18.7	-15.4	-24.1	-22.8	-	-	-	-	-	-	-	-19.7	-14.8	-14.5	-15.7	-16.2	-	-	-	-

- B = mBERT (Base), X = XLM-R (Base)
- Drops **huge** for:
  1. Distant target languages and
  2. Target languages with small pretraining corpora

Language-Specific Representation Subspaces

– In representation spaces produced by MMTs, one can still relatively easy discern language-specific subspaces

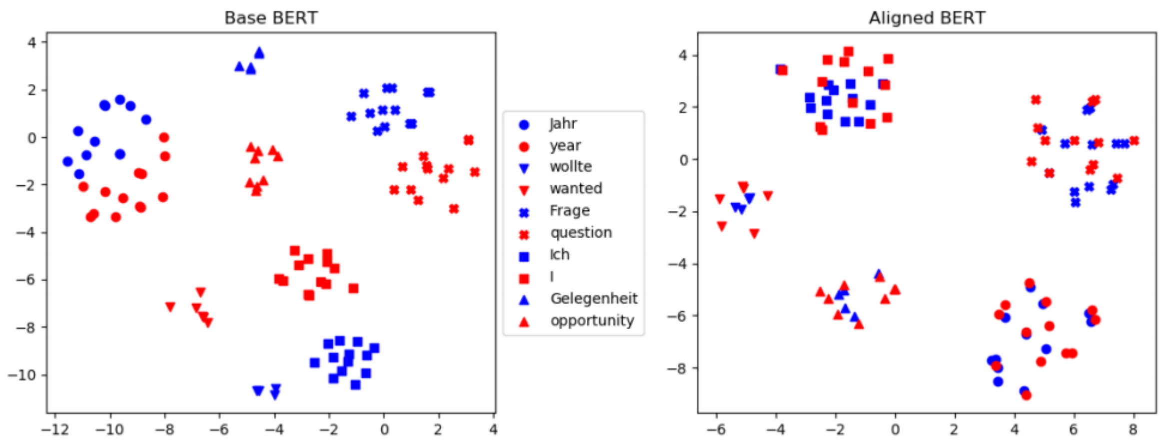


Image from [Cao et al., ‘20]

Better alignment between language subspaces...

– ...can be achieved with **bilingual supervision** (word translations of parallel data) [Wu & Conneau, ACL 20; Cao et al., ICLR 20; Hu et al., 2020]

– As with CLWEs: some bilingual/multilingual supervision ➔ better bilingual/multilingual representation space

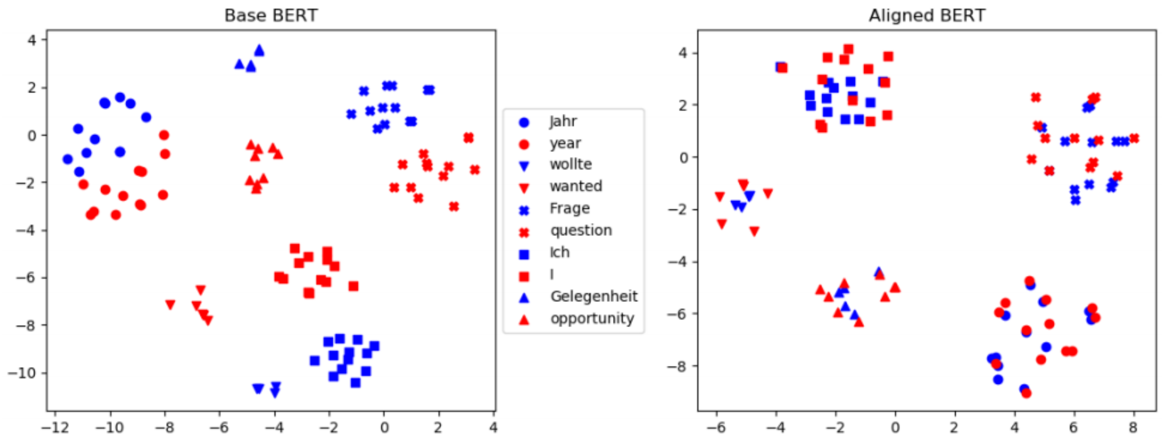


Image from [Cao et al., '20]

Choosing a Language Sample for CL Transfer Experiments

- Multilingual evaluation benchmarks should assess the expected performance of a model **across languages**
  - Sample of languages should be representative – but of what exactly?

- Findings can critically depend on the selection of languages
  - Most studies sample languages with the **largest digital footprint**
  - Such languages tend to belong to the same families (e.g., Indo-European)
  - Expected transfer performance is overestimated!

Variety sampling of languages

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., & Korhonen, A. (2020). *XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing,* pp. 2362-2376.

Idea: selection according to the distribution of linguistic properties

- **Variety sampling** favors the inclusion of outlier languages

1. **Typological diversity**: entropy of distribution of linguistic properties
2. **Family index**: number of different families / sample size
3. **Geography index**: entropy of lang. distr. over 6 geographic macro-areas

	Range	XCOPA	TyDiQA	XNLI	XQUAD	MLQA	PAWS-X
Typology	[0, 1]	0.41	0.41	0.39	0.36	0.32	0.31
Family	[0, 1]	1	0.9	0.5	0.6	0.66	0.66
Geography	[0, ln 6]	1.67	0.92	0.37	0	0	0

Learning outcomes

– Now you...

1. Understand what multilingual NLP is and why we need it
2. Know the mechanisms for inducing multilingual representations spaces
   - Cross-lingual word embeddings (CLWEs)
   - Massively multilingual transformers (MMTs)
3. Understand how to use multilingual representations spaces for CL transfer