UNIVERSITÄT MANNHEIM
Fakultät für Wirtschaftsinformatik und Wirtschaftsmathematik

Pitfalls in Evaluation of of Multilingual Text Representations

Goran Glavaš

Eval4NLP, November 20, 2020

Why Cross-Lingual 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!

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

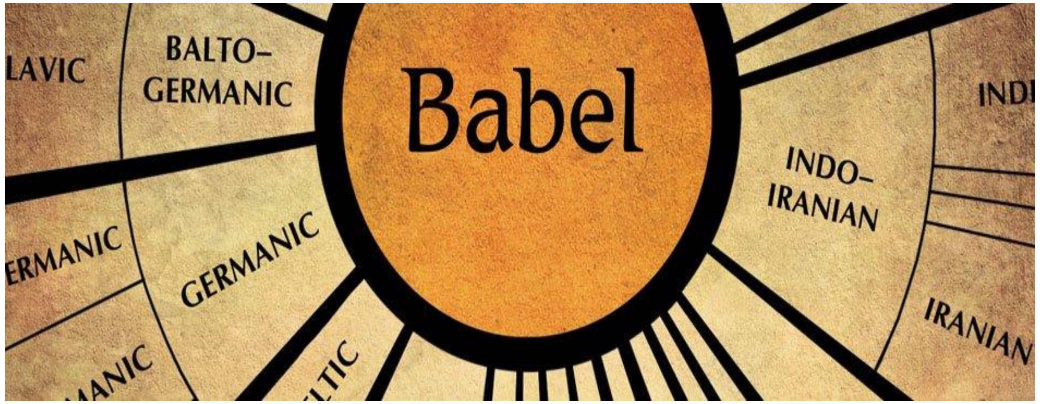
English

What this talk is about

- Crossing the Language Chasm
  - Cross-Lingual Word Embeddings (CLWEs)
  - Massively Multilingual Transformers (MMTs)
- Evaluation Pitfalls and Misleading Conclusions
  - Languages, Domains, and Corpora
  - Supervision
  - Tasks
  - Fair Comparisons

Crossing the Language Chasm

- **Old paradigm**:
  - Language-specific NLP models
  - Language-specific feature computation (i.e., preprocessing)
- **New paradigms**:
  - Representation learning: semantic vectors (embeddings)
  - Multilingual / cross-lingual representation learning



Crossing the Language Chasm

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
   - But...Unsupervised NMT?

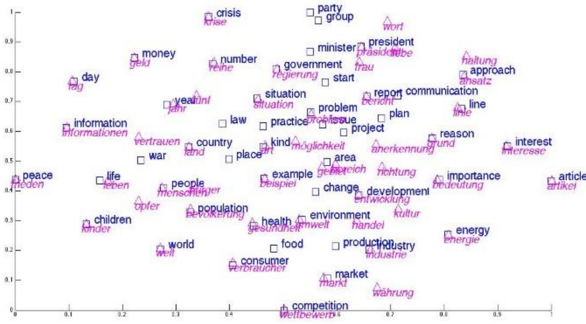
2. Multilingual KBs
   - Texts represented using entities from a multilingual KB
   - Same entity ID for same concepts across languages
   - Issues: coverage, entity linking

BabelNet 2.0

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

Crossing the Language Chasm

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]

Cross-lingual word embeddings

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.

- Typology of methods for inducing CLWEs [Ruder et al., ‘18]
  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 CLE models (*selection*)

- Jointly learning embeddings of two or more languages from scratch

1. Using word translations
   - Shared vectors for words in translation pairs [Guo et al., ‘14]
     - Feeding contexts from both languages to a standard embedding model (e.g., Skip-Gram)
   - Creation of pseudo-bilingual corpus [Gouws & Søgaard, ’15; Ammar et al., ’15; Duong et al., ’16; Adams et al., ‘17]
     - Pseudo-bilingual corpus by replacing words in a monolingual corpus with their translations

2. Using sentence translations (i.e., parallel data)
   - Compositional sentence model [Hermann & Blunsom, ‘13]
   - Bilingual Skip-Gram [Gouws et al., ‘15; Luong et al., ‘15]

Post-hoc embedding alignment

- Monolingual embeddings independently constructed
- Post-hoc aligning monolingual spaces

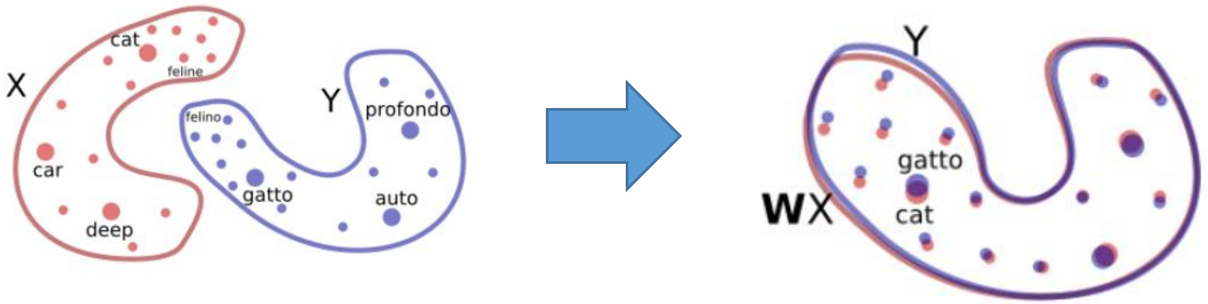
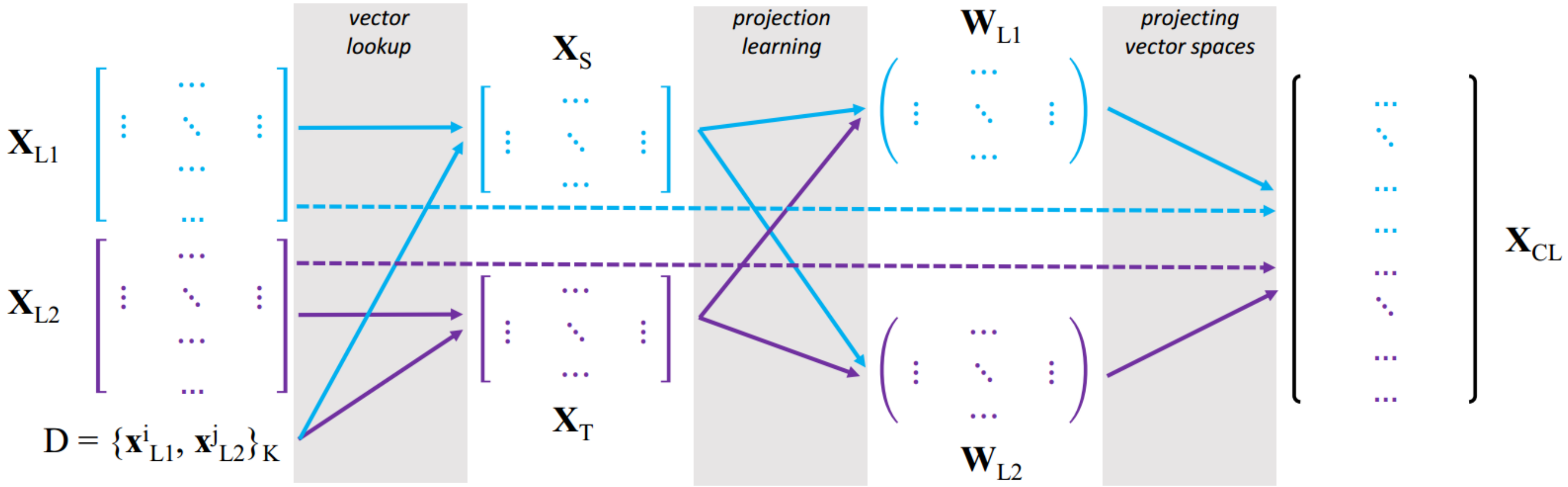


Image from [Conneau et al., '18]

- **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*(**X**) ∪ *g*(**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**



Projection-Based CLWE

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

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

- How do we find the „optimal" projection matrix $\mathbf{W}_{L1}$?
  - Mean square distance [Mikolov et al., '13] (and all subsequent work), except
  - (Relaxed) Cross-Domain Similarity Local Scaling [Joulin et al., '18]

Solving the Procrustes Problem

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

- If **W** is orthogonal, the above optimization problem is the so-called Procrustes problem with a closed-form solution [Schönemann, 1966]:

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

- Almost all projection-based CLWE models, supervised and unsupervised, solve the Procrustes problem in the final step

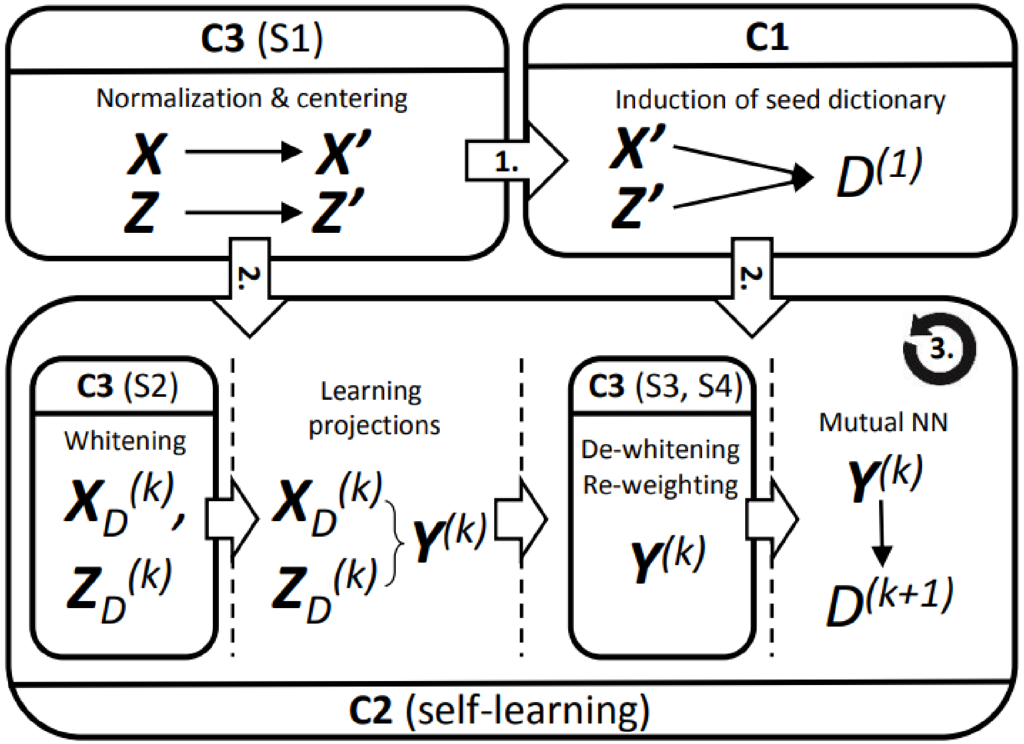
Unsupervised CLWE induction framework

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

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

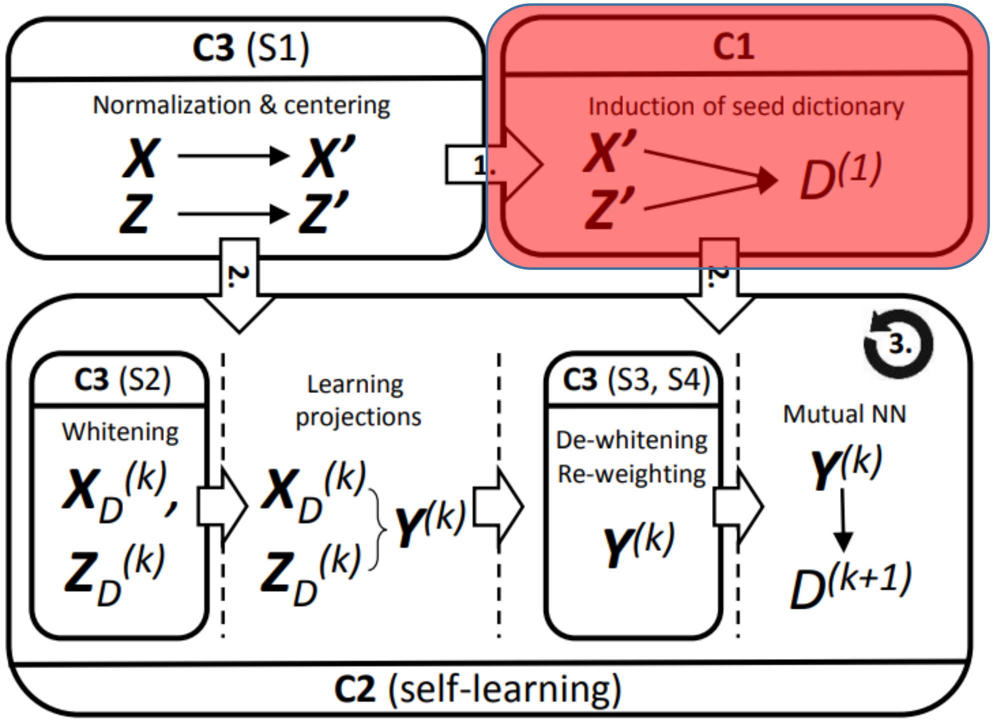
Repeat:

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

- 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!



CLWE Evaluation

Tasks: Why Do We Need CLWEs Exactly?

- Motivation for CLWEs in general
  - **Simple**: projection-based CLWEs can be obtained quickly (efficient training)
  - **Light-weight** and **inexpensive**

1. **Multilingual modeling of meaning** and
2. **Supporting cross-lingual transfer** for downstream NLP tasks

Tasks: Why Do We Need CLWEs Exactly?

- Most evaluations only on **Bilingual Lexicon Induction**
  - Effectively, **word translation**

- **BLI** is not (the only reason) why we induce CLWEs
  - To some extent tests multilingual modeling of meaning (at the word level)
  - Does it in reflect language transfer performance in downstream tasks?

- Even BLI results not comparable between models
  - Different language pairs, different training and testing dictionaries
  - No significance testing
    - Small numerical improvements (e.g., 0.5%) declared as „better performance"

Towards Better CLWE Evaluation [Glavaš et al., ACL 19]

- **Improved BLI evaluation**
  - Wide range of language pairs (pairs of languages not involving English)
    - Germanic (**DE**), Romance (**IT**, **FR**), Slavic (**RU**, **HR**), non Indo-European (**TR**, **FI**)
  - Same training / evaluation dictionaries
  - Testing differences in performance for statistical significance

- **Downstream evaluations**
  - BLI is not enough
  - **RQ**: do BLI results correlate with downstream performance?
  - **Three downstream tasks:**
    - Supervised: lang. transfer for (1) **Document classification** (TED-CLDC) and (2) **NLI**
    - Unsupervised: (3) ad-hoc cross-lingual document retrieval (CLIR)

Models in Evaluation

- Supervised models:
  - **CCA** [Faruqui & Dyer, '14]
  - **Procrustes (Proc)** [Smith et al., '17]
  - **Proc-B** [Glavaš et al., '19]
  - **RCSLS** [Joulin et al., '18]
- Unsupervised models:
  - **MUSE** [Conneau et al., '18]
  - **VecMap** [Artetxe et al., '18]
  - **ICP** [Hoshen et al., '18]
  - **GWA** [Alvarez-Melis & Jaakkola, '18]

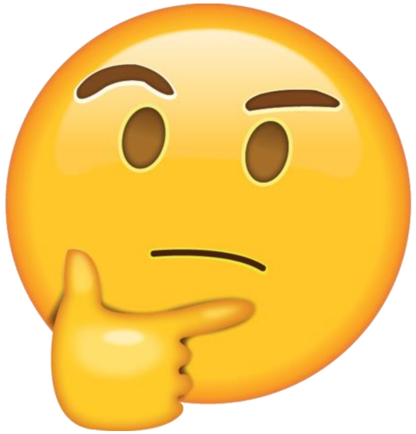
Results

		BLI	CLDC	XNLI	CLIR
SUP	Procrustes [Smith et al., 17]	.405 (2)	.267 (4)	.574 (3)	.196 (2)
	Proc-B [Glavaš et al., 19]	.398 (3)	.255 (5)	**.580 (1)**	**.216 (1)**
	RCSLS [Joulin et al., 18]	**.437 (1)**	**.510 (1)**	.385 (6)	.162 (4)
UNSUP	VecMap [Artetxe et al., 18]	.375 (4)	.405 (2)	**.581 (1)**	.155 (5)
	MUSE [Conneau et al., 18]	.183 (6)	.240 (6)	.467 (5)	.107 (6)
	ICP [Hoshen et al., 18]	.253 (5)	.348 (3)	.516 (4)	.182 (3)
	GWA [Alvarez-Melis & Jaakkola, 18]	.137 (7)	.184 (7)	.386 (6)	.072 (7)

- **BLI performance** (model ranking) poorly correlates with some of the downstream tasks
- BLI performance not enough to judge the quality of a CLWE space!

Do We Really Need Unsupervised CLWEs? [Vulić et al., EMNLP 19]

- **Motivation**
  - *„No bilingual/multilingual signal required"*
  - Thus suitable for / applicable to *„resource-lean languages"*
- Supervised models require **only a few thousand word pairs**
  - Almost trivial to obtain for any language pair
  - **PanLex** [Kamholz et al., '14] – aligned lexical entries for 9000+ language variants with the total of 1.1B translation pairs
- Unsupervised CLWE models thus not practically motivated
  - Are they *l'art pour l'art*?



Do We Really Need Unsupervised CLWEs? [Vulić et al., EMNLP 19]

- **Performance**: *„Unsupervised CLE outperforms supervised CLE”*
  - [Conneau et al., '18]: *„Without using any character information, our model even outperforms existing supervised methods on cross-lingual tasks for some language pairs”*
  - [Artetxe et al., '18]: *„Our method succeeds in all tested scenarios and obtains the best published results in standard datasets, even surpassing previous supervised systems”*
  - [Hoshen & Wolf, '18]: *„...our method achieves better performance than recent state-of-the-art deep adversarial approaches and is competitive with the supervised baseline”*

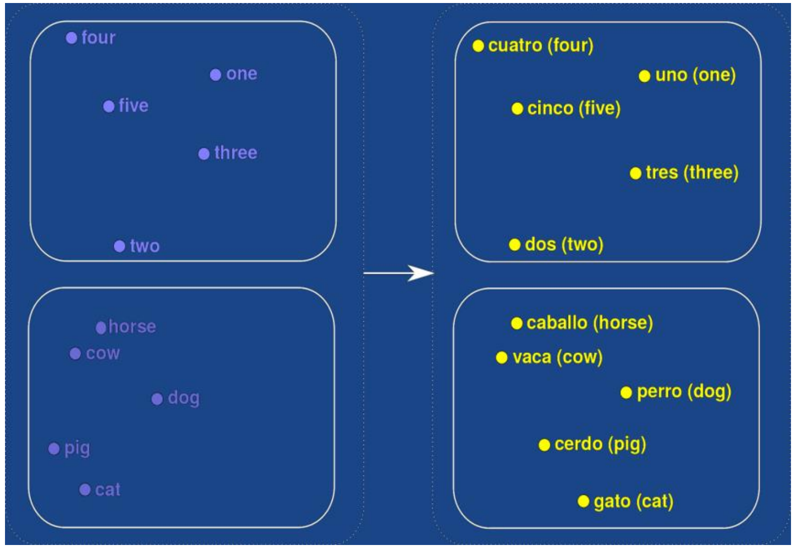
- **Unintuitive**: unsupervised CLE models all solve Procrustes problem in the final step, only using the less reliable (automatically induced) **D**
- Are unsupervised models compared fairly against supervised?

Unsupervised CLWEs and 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 & Baroni, '16]

- **Approximate isomorphism** of emb. spaces holds (loosely) only similar languages

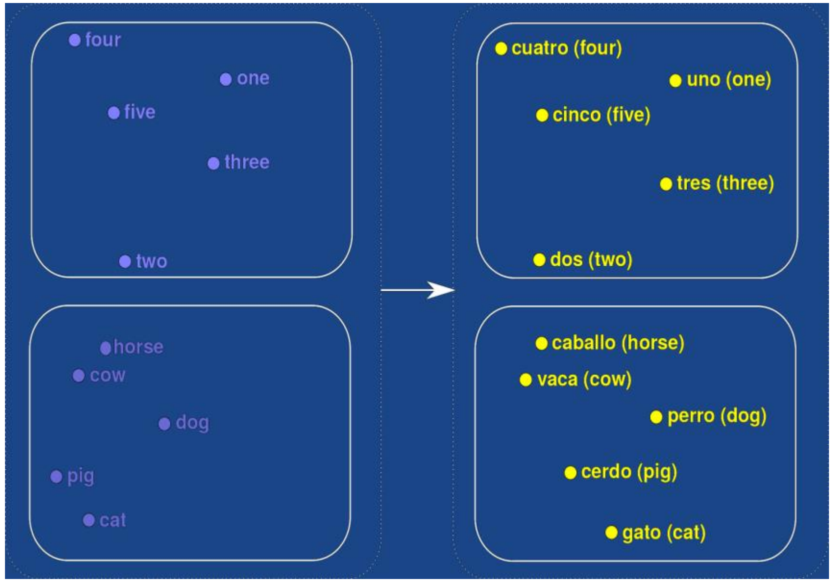
- It does not hold at all for distant languages and/or domains [Sogaard et al., '18; Vulić et al., EMNLP 20]



Unsupervised CLWEs and Approximate Isomorphism

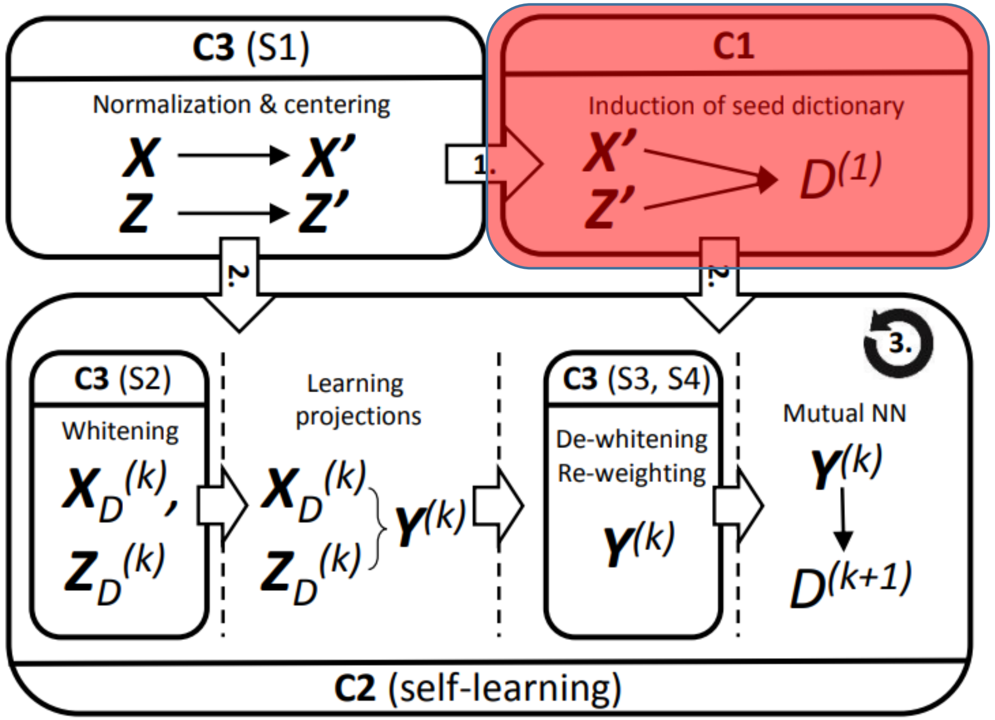
- "... we hypothesize that, if languages are used to convey thematically similar information in similar contexts, these random processes should be approximately isomorphic between languages, and that this isomorphism can be learned from the statistics of the realizations of these processes, the monolingual corpora, in principle without any form of explicit alignment." [Miceli & Baroni, '16]

- All (linear) projection-based CLWEs models rely on this assumption **once**
- But unsupervised CLWE models rely on it **twice** (additionally for inducing init. dict.)!



Do We Really Need Unsupervised CLWEs? [Vulić et al., EMNLP 19]

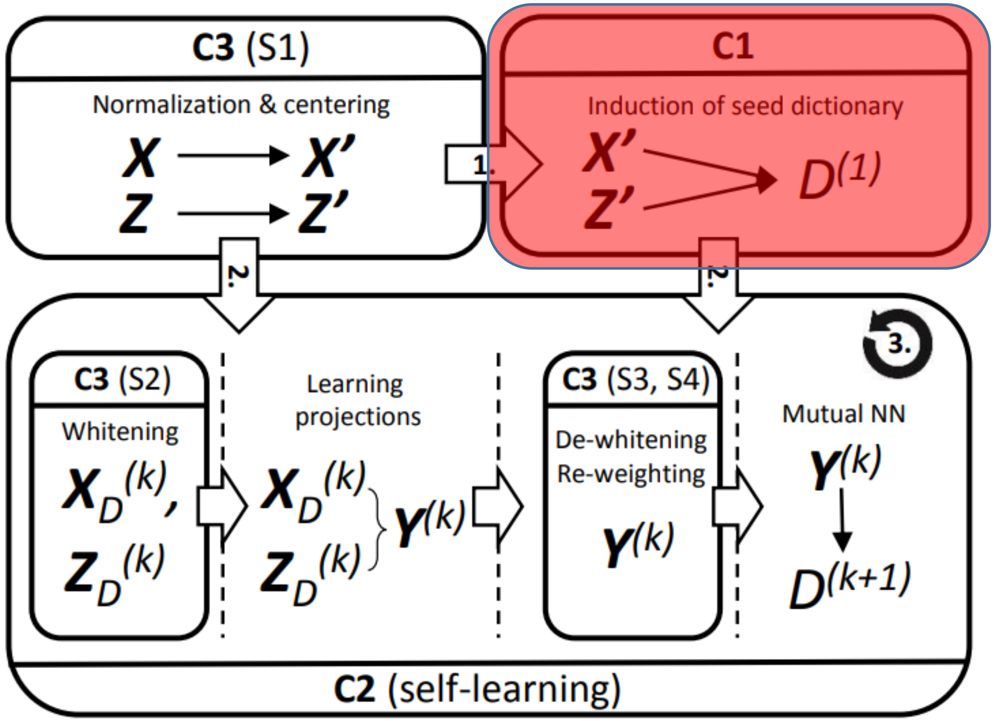
- Existing Supervised vs. Unsupervised CLWE evaluations are unfair
  - Evaluating the whole pipelines
  - **Unsup. + „bag of tricks”** vs. stripped down (basic) supervised models
  - Apples vs. oranges!



Supervised vs. Unsupervised CLWEs

- **Fair comparison**: vary only the component **C1** (dictionary induction)
  - Unsupervised induction (VecMap) vs. Supervised (clean initial dictionary)
- Keep all other useful „tricks"
  - Normalization
  - Centering
  - Whitening and de-whitening

  …

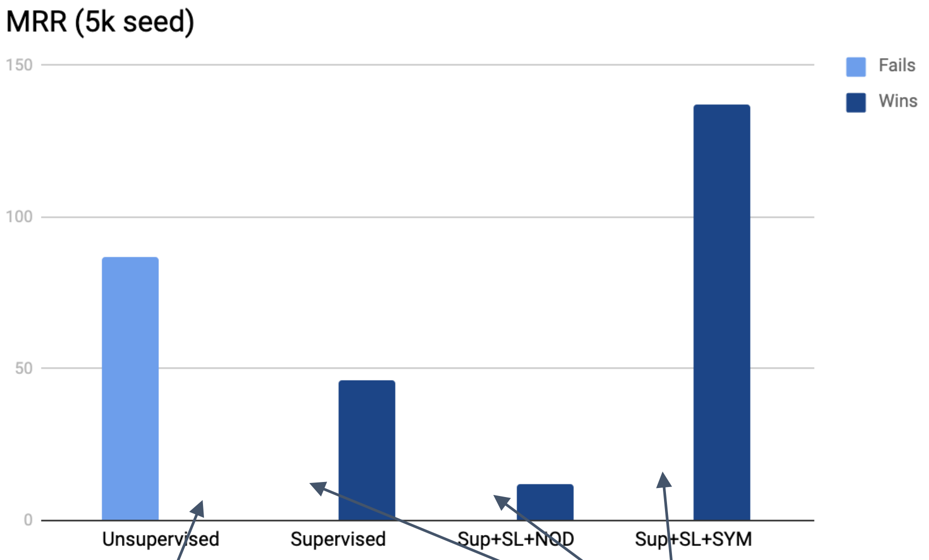


Do We Really Need Unsupervised CLWEs? [Vulić et al., EMNLP 19]

- **Wider evaluation**: **15** languages, **210** BLI setups

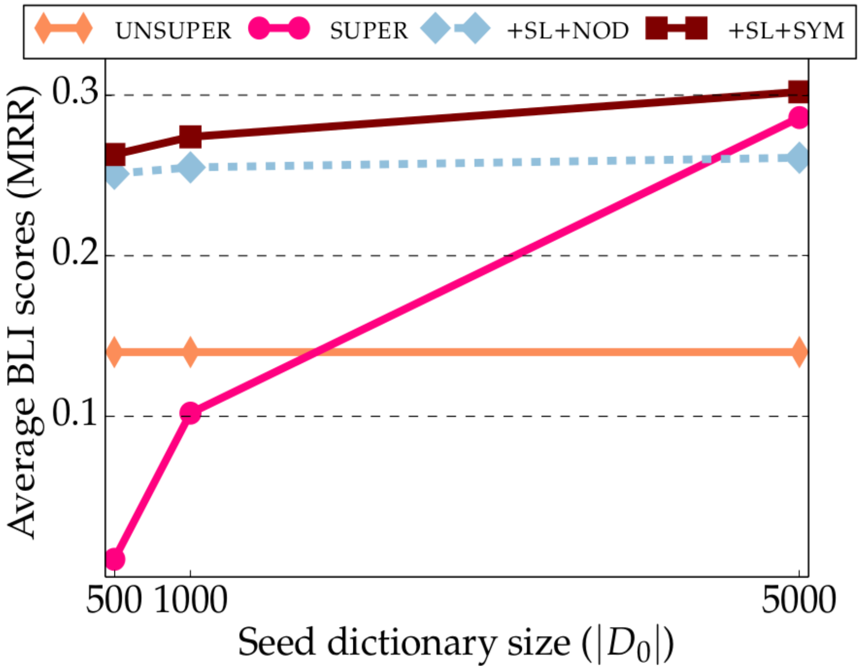
Language	Family	Type	ISO 639-1
Bulgarian	IE: Slavic	fusional	BG
Catalan	IE: Romance	fusional	CA
Esperanto	– (constructed)	agglutinative	EO
Estonian	Uralic	agglutinative	ET
Basque	– (isolate)	agglutinative	EU
Finnish	Uralic	agglutinative	FI
Hebrew	Afro-Asiatic	introflexive	HE
Hungarian	Uralic	agglutinative	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

Do We Really Need Unsupervised CLWEs? [Vulić et al., EMNLP 19]



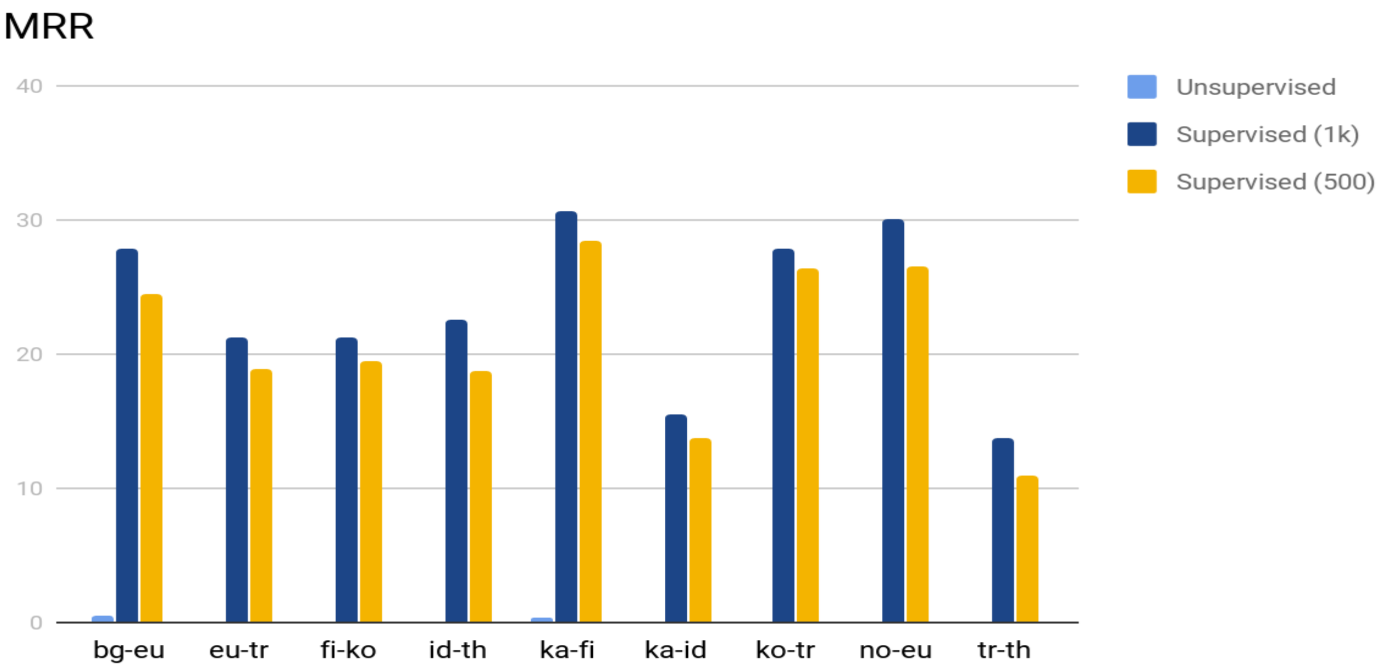
Unsupervised never wins!

Supervised never fails!



- Fully unsupervised VecMap completely fails for **87** lang. pairs

Do We Really Need Unsupervised CLWEs? [Vulić et al., EMNLP 19]



- Fully unsupervised VecMap completely fails for **87** language pairs
  - **All failed pairs include typologically and etymologically distant languages!**

(Massively) Multilingual Transformers

Massively Multilingual Transformers

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

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

  „Suprising cross-lingual effectiveness of BERT”
  [Wu & Dredze, 19]

  „mBERT surprisingly good at zero-shot CL model transfer”
  [Pires et al., 19]

So...has mBERT/XLM-R 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: An 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 et al., EMNLP 20]

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

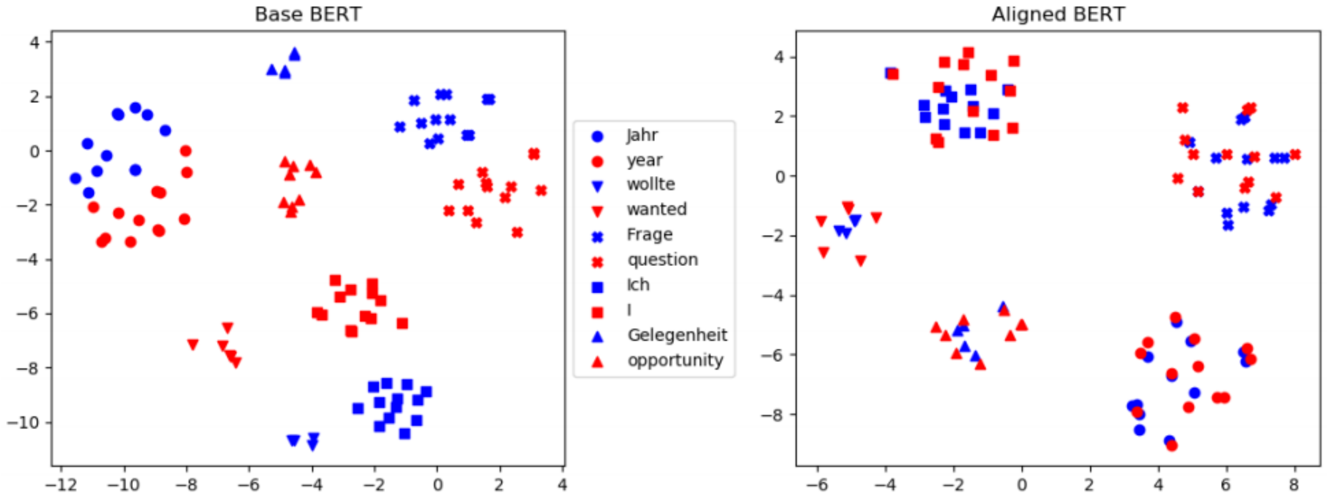


Image from [Cao et al., ‘20]

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

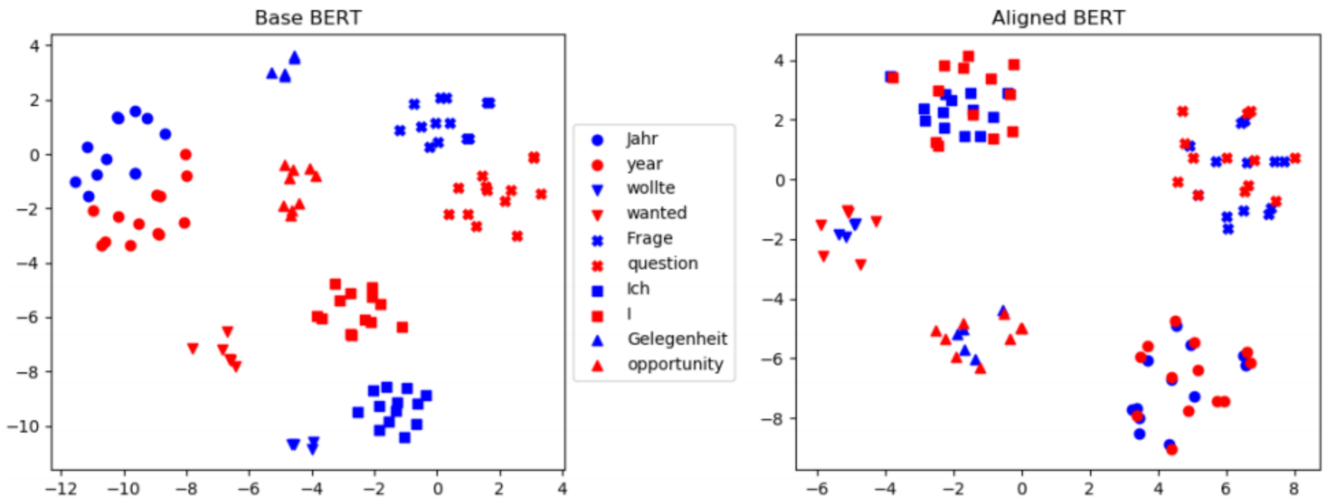


Image from [Cao et al., '20]

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 footpring**
  - Such languages tend to belong to the same families (e.g., Indo-European)
  - Expected transfer performance is overestimated!

Variety sampling of languages

Idea: selection according to the distribution of linguistic properties

- **Variety sampling** favors the inclusion of outliers

XCOPA (causal commonsense reasoning) [Ponti & Glavaš, et al., EMNLP 20]

1. **Typological diversity**: entropy of distribution of linguistic properties
   - E.g., from the URIEL database [Littel et al., 17]
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

Limitations uncovered by particular tasks

- Types of tasks also matter: **NLU tasks** dominate in CL benchmarks
  - QA, language inference, commonsense reasoning, etc.

- Limitations exposed by **reference-free MT** evaluation [Zhao et al., ACL 19]
  - **Adversarial** setup for MMTs
  - *„Translationese"* (bad literal „word-by-word" translations) receive representations similar to the source language sentences

source: *„Putin teilte aus und beschuldigte Ankara, Russland in den Rucken gefallen zu sein."*

system: „Putin lashed out and accused Ankara, Russia in the back fallen to be."

gold: „Putin lashed out, accusing Ankara of stabbing Moscow in the back."

Takeaways

Quick tought on **unsupervised** MT

- **Unsupervised MT** models are initialized either with...
  - A bilingual word embedding space [Artetxe et al., ‘18; Lample et al., ‘18]
  - A bilingual/multilingual pretrained transformer [Song et al., ’19; Liu et al., ‘20]

...and subjected to **denoising** and **back-translation** objectives

- All shortcomings/findings from unsupervised CLWEs and MMTs hold
  - UMT matches supervised MT performance only for close languages with large pretraining corpora – a setting where it’s not needed!
  - UMT fails for pairs of distant low-resource languages, a setting for which it is conceptually designed

Multilingual spaces induced without supervision

- **Absence of any explicit bilingual alignment**
  - Meaningful alignments between languages can only be obtained if there are prominent topological correspondences between language subspaces
  - Such topological alignments are inherently less likely to exist between typologically and etymologically distant languages
  - **Catch 22**: unsupervised multilingual representation learning unlikely to be work for intended use cases: distant and low-resource languages
- Be wary of any evaluation that renders a fully unsupervised MLRL method superior to supervised counterparts

Let’s not pretend we don’t have the resources we have

- Bilingual/multilingual signal is much more available than we think
  - **Parallel corpora**: JW300 [Agić & Vulić, ACL 19]
    Multilingual Bible Corpus [Mayer & Cysuow, LREC ‘14]
  - **Multilingual lexica**: PanLex [Kamholz et al., LREC ‘14]
- Language with no bilingual signal → most likely a language without a sufficiently large monolingual corpus

[Artetxe et al., ACL 20]: *„alleged scenario involving no parallel data and sufficient monolingual data is not met in the real world”*

On X* Benchmarks

- **Diversifying languages and tasks crucial**

- Diversifying languages easier: clearer criteria
  - Typological diversity,
  - etymological diversity,
  - geographic diversity

- Diversify the types of tasks
  - Language generation and LG **evaluation** tasks missing
  - MT, Cross-lingual summarization
  - **XGLUE** [Liang et al., EMNLP 20]**:** Question generation, new title generation

Thanks for your attention!

goran@informatik.uni-mannheim.de