How to (Properly) Evaluate Cross-Lingual Word Embeddings: On Strong Baselines, Comparative Analyses, and Some Misconceptions

Goran Glavaš[1], Robert Litschko[1], Sebastian Ruder[2], Ivan Vulić[3]

[1] Data & Web Science Group, University of Mannheim
[2] Insight Research Centre, National University of Ireland (now at DeepMind)
[3] PolyAI Ltd.

ACL 2019 Florence

MOTIVATION

Cross-Lingual Word Embeddings

- Cross-lingual word embeddings (CLWEs) conceptually allow for a **cheap language transfer** of NLP models
- Compared to full-blown MT, resource-light approach for bridging the language chasm in NLP applications
- While early CLWE models required parallel or comparable corpora, SOTA projection-based CLWE models require only limited **word-level supervision** or **no supervision**

CLWE Evaluation

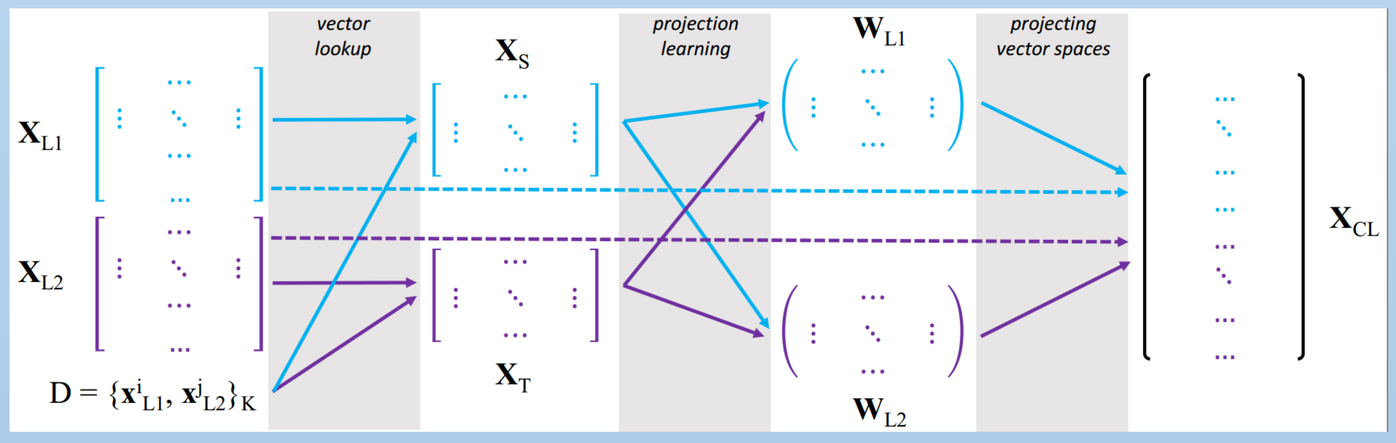
- Last few years have seen only **word-translation based** (bilingual lexicon induction, BLI) evaluation of induced CLWEs
- BLI is **NOT** why we induce CLWEs for: we use them primarily for CL transfer of NLP models; BLI evaluations vary greatly
- **Unsupervised CLWE induction models** are hot: no bilingual supervision needed, reported to (unintuitively) **outperform** supervised counterparts

RQs

1. Is BLI performance a good indicator of the quality of a CLWE space?

2. Are unsupervised CLWE induction models really better than supervised?

PROJECTIONS



Katze, er, ..., wählen $\begin{bmatrix} -1.18 & 0.21 & ... & 0.11 \\ 0.23 & -0.53 & ... & 0.34 \\ ... & ... & ... & ... \\ 0.78 & 1.33 & ... & -0.47 \end{bmatrix}$ $\times \mathbf{W}_{L1} =$ $\begin{bmatrix} 0.59 & 1.01 & ... & 0.37 \\ -0.34 & -0.27 & ... & 0.41 \\ ... & ... & ... & ... \\ 0.81 & -0.31 & ... & 0.29 \end{bmatrix}$ cat, he, ..., choose

- Initial seed dictionary D of word translations is given (**supervised**) or induced (**unsupervised**)
- $\mathbf{X}_S$ and $\mathbf{X}_T$: matrices with vectors of aligned words from the training dictionary D
- Solution by **solving the Procrustes problem**: $\mathbf{W}_{L1} = \mathbf{U}\mathbf{V}^T$ where $\mathbf{U}\mathbf{S}\mathbf{V}^T = \text{SVD}(\mathbf{X}_T\mathbf{X}_S^T)$

MODELS

- **Supervised models**:
  - CCA (Faruqui & Dyer, EACL 14); Simple Procrustes (Smith et al., ICLR 17); **Bootstrapped Procustres** (from smaller dicts; our contribution)
  - Discriminative Lat.-Var. Model (Ruder et al., EMNLP 18); **Relaxed CSLS** (Joulin et al., EMNLP 18) with **BLI-specialized objective/loss function**
- **Unsupervised models**:
  - MUSE (Conneau et al, ICLR 18); **VecMap** (Artetxe et al., ACL 18)
  - Gromov-Wasserstein Alignment (Alvarez.Melis & Jaakkola, EMNLP 18); Iterative Closest Point (Hoshen & Wolf, EMNLP 18)

RESULTS

- **Standardized BLI evaluation**
  - Same test sets (and same training sets for supervised models)
  - 8 languages, yielding 28 language pairs for evaluation: **EN** & **DE** (Germanic), **IT** & **FR** (Romance), **RU** & **HR** (Slavic), **FI** & **TR** (non-Indo-European, agglutinative)
- **Downstream evaluations**
  - 3 tasks: (1) CL document retrieval (CLIR) and CL transfer for (2) natural language inference (XNLI) and (3) document classification (CLDC)

Model	BLI (MAP / successful LPs)	XNLI (accuracy)	CLDC (micro avg. F_1)	CLIR (MAP)
Procrustes (train dict. 1K)	.299 (28)	.536	.190	.133
Procrustes (train dict. 5K)	.405 (28)	**.574**	.267	.196
Bootstrapped Proc. (train dict. 1K)	.379 (28)	**.579**	.251	**.216**
DLV (train dict. 5K)	.403 (28)	**.571**	.258	.197
RCSLS (train dict 5K)	**.437** (28)	.385	**.510**	.162
VecMap (unsup.)	.375 (28)	**.581**	**.405**	.155
MUSE (unsup.)	.183 (13)	.467	.240	.107
Iterative Closest Point (unsup.)	.253 (22)	.516	.348	**.182**
Gromov-Wasserstein (unsup.)	.137 (15)	.386	.184	.072

- **RCSLS**, tuned for BLI, is the most peculiar:
  - Best on BLI & CLDC, mediocre on CLIR and bad for NLI transfer
  - Overfitting for BLI may severely hurt downstream performance
- BLI results do not necessarily correlate with downstream results

Correlation w. BLI results	XNLI	CLDC	CLIR
All models	.269	.390	.764
All without RCSLS	**.951**	.266	**.910**

TAKEAWAYS

1. **BLI performance alone is not enough to judge the quality and usefulness of induced CLWE spaces**
2. **BLI must be coupled with a set of diverse downstream CL applications**
3. **Unsupervised CLWE induction methods do not outperform supervised counterparts**

Get in touch! @gg42554 @seb_ruder @licwu

Code & Data