We can obtain **linguistically-informed** and **dense** **<u>language representations</u>** by computing a **CCA** shared space from both **typological knowledge bases** and **task-learned** vectors

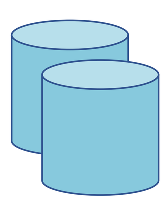
Towards a Multi-view Language Representation: a **shared space** of **discrete** and **continuous language features**

<u>Arturo</u> Oncevay, Barry Haddow, Alexandra Birch
School of Informatics, ILCC, University of Edinburgh, Scotland

0 WARNING
Language vector
!=
Word/Sentence vector

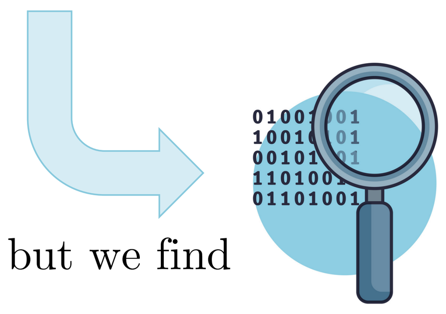
1 <u>Introduction</u>: where can we extract <u>language representations</u>?

Linguistically-informed language vectors from **typological Knowledge Bases** (KB)



e.g. WALS [1]
Word Order: Order of Subject, Object and Verb

but we find

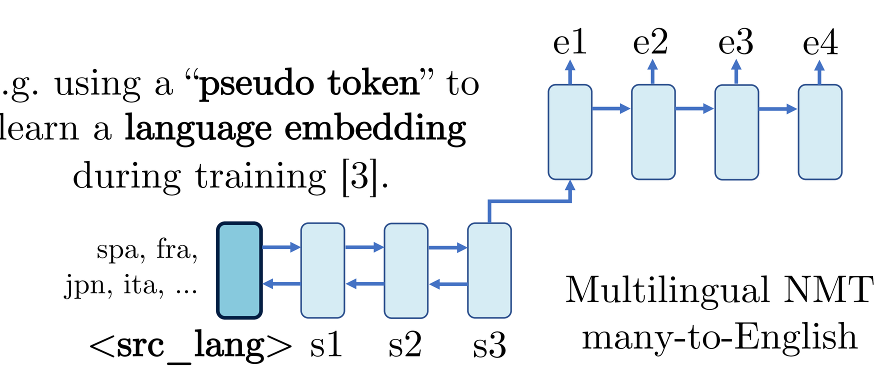


- x Categorical values
- x Sparse features
- x Missing entries
- x Redundant variables

How can we obtain the best of both worlds with minimal information loss?

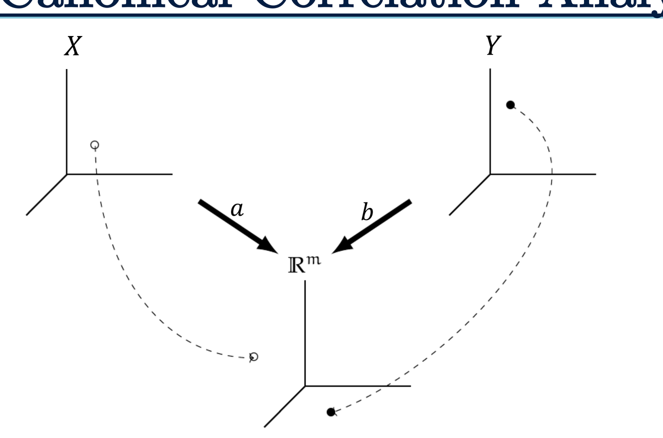
Dense and continuous task-learned language vectors, like from Language Modelling [2] or Neural Machine Translation (NMT) [3].

e.g. using a "**pseudo token**" to learn a **language embedding** during training [3].



2 <u>Multi-view Language Representations</u> with <u>Canonical Correlation Analysis (CCA)</u>:

Two views (X, Y) for a given set of data are **projected in a shared space** with m dimensions, by **maximising their correlation** in each coordinate and **retaining as little redundancy** as possible.

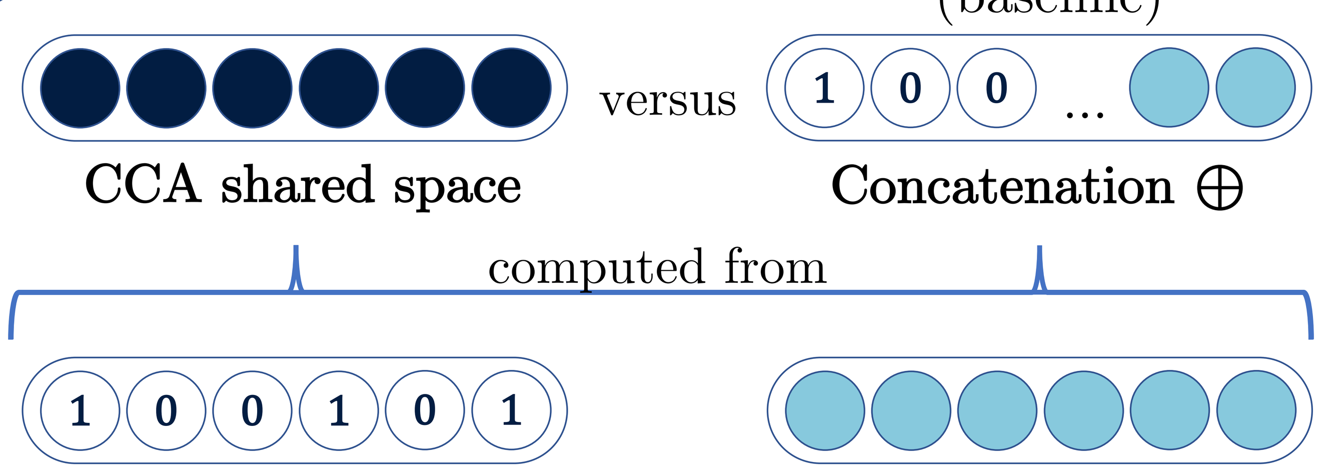


$$\underset{a_j, b_j}{\text{argmax}} \quad \text{corr}\left(a_j X^\top, b_j Y^\top\right) \qquad j \in \{1..m\}$$

such that
$$\text{corr}\left(a_j X^\top, a_k X^\top\right) = 0, \quad k < j$$
$$\text{corr}\left(b_j Y^\top, b_k Y^\top\right) = 0, \quad k < j$$

where: $X \in \mathbb{R}^{d}$ and $Y \in \mathbb{R}^{d'}$
$a_j \in \mathbb{R}^{1\times d}$ and $b_j \in \mathbb{R}^{1\times d'}$
corr function returns the Pearson correlation between two vectors (pairwise element)

3 <u>Experimental Setup:</u>



KB-vectors (lang2vec [4])
- Syntax (103 feats.)
- Phonology (25)
- Phonetic Inventory (158)

NMT-vectors [3] (512 dim.)

KB ∩ NMT = 729 lang.
KB − NMT = 2989 lang.
NMT − KB = 287 lang.

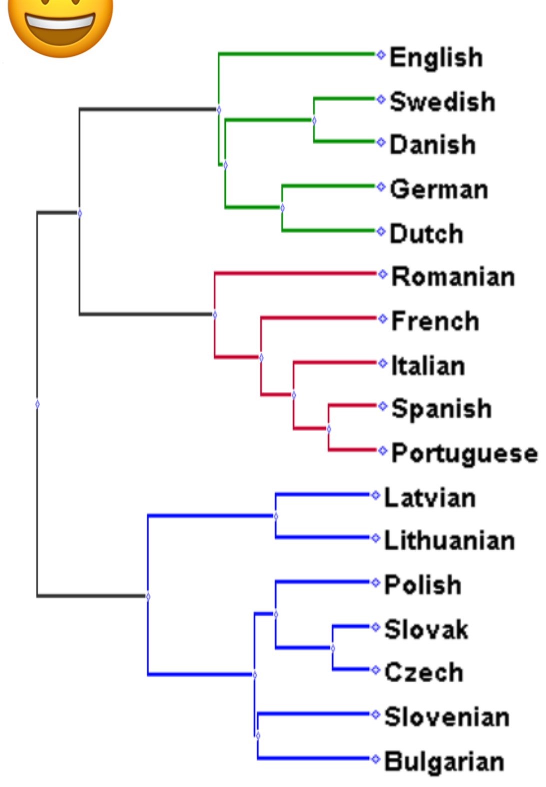
4 <u>Extrinsic Evaluation:</u>

A. Typological Feature Prediction: 😕

To predict one language feature given the others (with a one-leave-out setting). See <u>Table 1</u>.

B. Phylogenetic Tree Inference: 😃

To reconstruct a phylogenetic tree of 17 Indo-European languages and evaluate the distance to a Gold Standard language tree. See <u>Table 2</u>.



Gold Standard tree (Source: [5])

5 ~~Conclusion~~ Summary:

We projected multi-view language vectors with:
- ✓ Embedded information from both typological KBs and parallel corpora.
- ✓ Some retained genetic information.
- ✓ The potential to compute a language vector even if one of the two views is not available.

6 Ongoing Work:

- What kind of information we really retain?
- Can we take advantage of language representations in Neural Machine Translation?

Complementary information

4.A. Typological Feature Prediction:
Following [3], we performed a leave-one-out feature prediction with Logistic Regression classifiers, to identify the truth value when available. We use 10-fold cross-validation grouped by languages.

Feature class	# feats.	⊕	CCA
Syntax	97/103	**88.44**	85.29
Phonology	27/28	84.51	**89.62**
Phonetic Inventory	126/158	**91.66**	91.15

Table 1: Prediction in composed spaces (KB and NMT-learned) by concatenation ⊕ and CCA. Features are filtered out due to missing values and number of targets.

4.B. Phylogenetic Tree Inference:
We measured the distance [5] between a Gold Standard tree (τ) [6] and a reconstructed phylogenetic tree (g) of N languages or leaves (l) in different clustering settings.

$$\text{Dist}(\tau, g) = \sum_{i,j\in\{1..N\};i\neq j} \left(D_\tau\left(l_i, l_j\right) - D_g\left(l_i, l_j\right)\right)^2$$

linkage→	UPMGA		Ward	
#lang. (±eng)→	16	17	16	17
Random tree (avg.)	0.523	0.569	0.473	0.529
NMT-learned (L)	0.419	-	0.340	-
Syntax (S)	0.232	0.238	0.149	**0.160**
$S \oplus L$	0.291	-	0.159	-
CCA(S, L)	**0.205**	**0.216**	**0.140**	0.172
Phonology (P)	0.588	0.649	0.450	0.490
$P \oplus L$	0.466	-	0.422	-
CCA(P, L)	0.462	0.511	0.341	0.464
Phon. Inventory (I)	0.346	0.366	0.354	0.370
$I \oplus L$	0.440	-	0.547	-
CCA(I, L)	0.726	0.932	0.318	0.618

Table 2: Unweigthed distances to Gold Standard trees per metric (lower is better). English (eng) cannot be evaluated in all spaces without an NMT-learned vector.

References:
[1] Matthew S. Dryer and Martin Haspelmath (eds.) 2013. **The World Atlas of Language Structures Online**. *Leipzig: Max Planck Institute for Evolutionary Anthropology*. (Available online at <u>http://wals.info</u>).
[2] Robert Östling and Jörg Tiedemann. 2017. **Continuous multilinguality with language vectors**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 644–649, Valencia, Spain. Association for Computational Linguistics.
[3] Chaitanya Malaviya, Graham Neubig and Patrick Littell. 2017. **Learning language representations for typology prediction**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen, Denmark. Association for Computational Linguistics.
[4] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. **URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Vol. 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.
[5] Ella Rabinovich, Noam Ordan, and Shuly Wintner. 2017. **Found in translation: Reconstructing phylogenetic language trees from translations**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)*, p. 530–540, Vancouver, Canada. Association for Computational Linguistics.
[6] M. Serva and F. Petroni. 2008. **Indo-European languages tree by Levenshtein distance**. *EPL (Europhysics Letters)*, 81(6):68005.

THE UNIVERSITY of EDINBURGH

Take a picture to download the poster

Contact:
Arturo Oncevay
a.oncevay@ed.ac.uk
@a11byte