Towards a Multi-view Language Representation: A Shared Space of Discrete and Continuous Language Features

Arturo Oncevay, **Barry Haddow** and **Alexandra Birch**
School of Informatics, University of Edinburgh, UK
a.oncevay@ed.ac.uk

Abstract

Linguistic typology databases contain valuable knowledge of the distinguishing properties of different languages. Typically they contain sparse discrete features that are difficult to integrate into computational methods, and dense task-learned language vectors have emerged in response. To join both worlds, we compute a shared space between discrete (binary) and continuous features using canonical correlation analysis. We evaluate the new language representation against a concatenation baseline in typological feature prediction and in phylogenetic inference, obtaining promising results to explore further.

1 Introduction

Knowledge bases (KB) of linguistic typology are gaining attention for encoding languages mostly in the discrete space. URIEL and `lang2vec` (Littell et al., 2017) allow a straightforward extraction of typological binary features from different KBs, whereas Murawaki (2015, 2018) exploits them to build continuous and latent language representations. Furthermore, language embeddings have also been directly computed in continuous space from data-driven tasks such as neural machine translation (NMT, Malaviya et al., 2017) and language modelling (Tsvetkov et al., 2016; Östling and Tiedemann, 2017; Bjerva and Augenstein, 2018).

In this work, we hypothesise that an information-rich, dense representation of languages should include knowledge about the different properties and similarities between them, as this could be important for delivering more effective multi-lingual systems without the need of large corpora. Therefore, our research question is how can we obtain the best of both views (KB and task-learned, or discrete and continuous) with minimal information loss.

2 Multi-view Language Representation with Canonical Correlation Analysis

We take inspiration from work on mapping textual captions to abstract scenes (Papasarantopoulos et al., 2018), as Canonical Correlation Analysis (CCA) allows to find a projection of two views for a given set of data. Represented by two random vectors $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}^{d'}$, the views are projected in a shared space with m dimensions, by maximising their correlation in each coordinate and retaining as little redundancy as possible. CCA solves a sequence of optimisation problems for $j \in \{1..m\}$ where $a_j \in \mathbb{R}^{1\times d}$ and $b_j \in \mathbb{R}^{1\times d'}$:

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

where the corr function returns the Pearson correlation between two vectors (pairwise element).

CCA has been previously used to derive embeddings at word-level (Faruqui and Dyer, 2014; Dhillon et al., 2015; Osborne et al., 2016), but as far as we know, this is the first time it has been used to derive distributed language representations. A major advantage of CCA is that it enables us to project a sample only known in one of the two original views (e.g. a language described in a KB but without any parallel corpus).

3 Experimentation

We compare a CCA shared space of discrete and continuous features against a concatenation baseline ($\oplus$). We take the data from `lang2vec` (Littell et al., 2017), precisely, the kNN-based completed features of Syntax (103 feats.), Phonology (25) and Phonetic Inventory (158) classes, as well as the NMT-learned vectors (512 dim., Malaviya et al., 2017). We considered all 729 languages that appear in both the KB and learned sources.

Feature class	# feats.	$\oplus$	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 $\oplus$ and CCA. Features are filtered out due to missing values and number of targets.

3.1 Typological Feature Prediction

We perform a typological prediction task similar to Malaviya et al. (2017). For each feature class, we train Logistic Regressor classifiers to predict each feature targeting the truth value when available. We skip features with missing values or less than two targets. We compute the shared spaces ($\oplus$ and CCA) and train using a leave-one-out feature scheme with zero-mean normalised values. Then, we use 10-fold cross-validation grouped by languages. In Table 1, we observe that CCA cannot surpass the concatenation baseline in two of three feature classes. We plan to extend the analysis for language and feature groups, to examine why Phonology is the only benefited class.

3.2 Phylogenetic Inference

In our second experiment, we examine how well phylogenetic trees can be reconstructed from the language representations. Following previous work (Rabinovich et al., 2017; Bjerva et al., 2019b), we take a pruned tree of 17 Indo-European languages (Serva and Petroni, 2008) as a Gold Standard (GS). We tested two different clustering algorithms, since the GS was agglomerated with UPGMA, whereas Rabinovich et al. (2017) used variance minimisation (Ward, Ward Jr, 1963).

To measure the distance $\mathrm{Dist}(\tau, g)$ between the hypothesised tree τ and the GS tree g, we proceed as follows: Given a tree τ with N leaves, the weighted distance between two leaves, $D_\tau(l_i, l_j)$, is the sum of the weights of all edges on the shortest path that links each other, whereas the unweighted distance counts the edges involved (all weights equal one). Then, we obtain the desired distance between the trees by summing up the squared differences of all their leaf-pair distances.

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

We also compute a baseline by averaging distances of 10k random trees. For comparison purposes, we (zero-one) normalised the distances using the maximum one from the random samples. Table 2 presents the unweighted distances, where there are improvements with CCA against concatenated and single-view vectors in most of the cases. Phonetic Inventory is the exception, and we should analyse the latent genetic information lost in the blend.

cluster metric→	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.

A significant advantage is the representation of an "unknown" language, as we projected English from the KB view without a task-learned vector from the one-to-many NMT of Malaviya et al. (2017). Using our method, we can quickly address languages with single-view representations. For instance, `lang2vec` contains 2989 and 287 unique entries in the KB and NMT-learned sources, respectively.

4 Discussion and Ongoing Work

Multi-view language representations based on CCA have achieved positive results in phylogenetic inference, although they could not surpass all baselines for typological feature prediction. However, the former outcome is relevant due to its implications about language variations and clusters. Further correlation analysis should be performed for the discrete features in different KBs.

We plan to exploit the representations in multilingual NMT, where there are potential paths in parameter selection (Johnson et al., 2017) and generation (Platanios et al., 2018) by using language embeddings in neural architectures. Moreover, we expect to compose, in a multi-step scheme, different continuous learned representations with discrete features from linguistic KBs. Finally, we can also evaluate the CCA-based vectors in typological collaborative filtering (Bjerva et al., 2019a), as well as thoughtfully analyse what kind of information they really retain (Bjerva et al., 2019b).

References

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. *Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 907–916.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019a. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (In-press)*.

Johannes Bjerva, Robert Östling, Maria Han Veiga, Jörg Tiedemann, and Isabelle Augenstein. 2019b. What do language representations really represent? *Computational Linguistics (In-press)*.

Paramveer S Dhillon, Dean P Foster, and Lyle H Ungar. 2015. Eigenwords: Spectral word embeddings. *The Journal of Machine Learning Research*, 16(1):3035–3078.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling Zero-Shot translation. *Transactions of the Association of Computational Linguistics*, 5(1):339–351.

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: Volume 2, Short Papers*, volume 2, pages 8–14.

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.

Yugo Murawaki. 2015. Continuous space representations of linguistic typology and their application to phylogenetic inference. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 324–334.

Yugo Murawaki. 2018. Analyzing correlated evolution of multiple features using latent representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4371–4382.

Dominique Osborne, Shashi Narayan, and Shay B Cohen. 2016. Encoding prior knowledge with eigenword embeddings. *Transactions of the Association for Computational Linguistics*, 4:417–430.

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.

Nikos Papasarantopoulos, Helen Jiang, and Shay B Cohen. 2018. Canonical correlation inference for mapping abstract scenes to text. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435.

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 (Volume 1: Long Papers)*, pages 530–540, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maurizio Serva and Filippo Petroni. 2008. Indo-European languages tree by levenshtein distance. *EPL*, 81(6):68005.

Yulia Tsvetkov, Sunayana Sitaram, Manaal Faruqui, Guillaume Lample, Patrick Littell, David Mortensen, Alan W Black, Lori Levin, and Chris Dyer. 2016. Polyglot neural language models: A case study in Cross-Lingual phonetic representation learning. In *Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1366.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.