URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors

Patrick Littell[1], **David Mortensen**[1], **Ke Lin**[2],
Katherine Kairis[2], **Carlisle Turner**[3], **and Lori Levin**[1]

[1]Carnegie Mellon University, Language Technologies Institute
[2]University of Pittsburgh, Department of Linguistics
[3]University of Pittsburgh, Swanson School of Engineering
{plittell,dmortens,lsl}@cs.cmu.edu
{kel97,kak275,crt43}@pitt.edu

Abstract

We introduce the URIEL knowledge base for massively multilingual NLP and the lang2vec utility, which provides information-rich vector identifications of languages drawn from typological, geographical, and phylogenetic databases that are normalized to have straightforward and consistent formats, naming, and semantics. The goal of URIEL and lang2vec is to enable multilingual NLP, especially on less-resourced languages and make possible types of experiments (especially but not exclusively related to NLP tasks) that are otherwise difficult or impossible due to the sparsity and incommensurability of the data sources. lang2vec vectors have been shown to reduce perplexity in multilingual language modeling, when compared to one-hot language identification vectors.

1 Introduction

This article introduces lang2vec[1], a database and utility representing languages as information-rich typological, phylogenetic, and geographical vectors. lang2vec feature primarily represent binary language facts (e.g., that negation precedes the verb or is represented as a suffix, that the language is part of the Germanic family, etc.) and are sourced and predicted from a variety of linguistic resources including WALS (Dryer and Haspelmath, 2013), PHOIBLE (Moran et al., 2014), Ethnologue (Lewis et al., 2015), and Glottolog (Hammarström et al., 2015).

Despite the heterogeneity of its sources, lang2vec provides a simple interface with consistent formats, featuring naming, language codes, and feature semantics. lang2vec takes as its input a list of ISO 639-3 codes and outputs a matrix of [0.0, 1.0] feature values (like those in Table 1), allowing straightforward "plug and play" experimentation where different sources or types of information can easily be combined or contrasted.

lang2vec is a release of the URIEL project, a compendium of tools and resources to better enable multilingual NLP, especially in less-resourced languages where conventional NLP resources like parallel corpora are limited.

2 Motivation

The recent success of "polyglot" models (Hermann and Blunsom, 2014; Faruqui and Dyer, 2014; Ammar et al., 2016; Tsvetkov et al., 2016; Daiber et al., 2016), in which a language model is trained on multiple languages and shares representations across languages, represents a promising avenue for NLP, especially for less-resourced languages, as these models appear to be able to learn useful patterns from better-resourced languages even when training data in the target language is limited.

Just as neural NLP raises many questions about the best representations of words and sentences, these models raise the question of the representation of *languages*. Tsvetkov et al. (2016) shows that vectors that represent *information* about the language outperform a simple "one-hot" representation where each language is represented by a 1 in a single dimension. This result parallels the results of other recent work in sound/character representation, in which vectors of linguistically-aware features outperform one-hot character representations on some tasks (Bharadwaj et al., 2016;

[1]www.cs.cmu.edu/~dmortens/downloads/uriel_lang2vec_latest.tar.xz

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 8–14,
Valencia, Spain, April 3-7, 2017.

	S_SUBJECT-_BEFORE_VERB	S_SUBJECT-_AFTER_VERB	S_ADPOSITION-_BEFORE_NOUN	S_ADPOSITION-_AFTER_NOUN
eng	1	0	1	0
mlg	0	1	1	0
kaz	1	0	0	1

Table 1: Truncated `lang2vec` syntax vectors for English, Malagasy, and Kazakh, representing binary feature values converted from multi-class features in WALS (Dryer and Haspelmath, 2013) (§3.1), extracted by text-mining prose descriptions in Ethnologue (Lewis et al., 2015) (§3.1), and imputed by k-nearest-neighbors classification from related, nearby, and similar languages (§4).

Training set	baseline	`id`	`id+phonology+inventory`
Italian monolingual	4.36	—	—
Italian, French, Romanian	5.73	4.93	**4.24** (-26.0%)
Italian, French, Romanian, Hindi	5.88	4.98	**4.41** (-25.0%)
Hindi monolingual	3.70	—	—
Hindi, Tamil, Telegu	4.14	3.78	**3.35** (-19.1%)
Hindi, Tamil, Telegu, English	4.29	3.82	**3.42** (-20.3%)

Table 2: Perplexity of monoglot and polyglot language models in Italian and Hindi (Tsvetkov et al., 2016), when the languages are not identified to the model (baseline), when the languages are represented as one-hot vectors (`id`), and when languages are represented as `lang2vec` vectors (`id+phonology+inventory`).

Rama, 2016).

Sample results from Tsvetkov et al. (2016) are reproduced in Table 2, measuring the perplexity of monolingual and polyglot models, trained on pronunciation dictionaries in several languages and tested on Italian and Hindi. We can see that training on a set of three similar languages, and a set of four similar and dissimilar languages, raises perplexity above the baseline monolingual model, even when the language is identified to the model by a one-hot (`id`) vector. However, perplexity is lowered by the introduction of phonological feature vectors for each language (the `phonology` and `inventory` vector types described in §3.1), giving consistently lower perplexity than even the monolingual baseline.

Providing such vectors for many languages, however, is made difficult by the somewhat piecemeal digital representation of language information. There exist many information-rich sources of language data, but each source covers different sets of languages in different levels of detail, has different formats and semantics (ranging from binary features to trees to English prose descriptions), uses different identifiers for languages and different names for features, etc.

It does not take long in collecting a "polyglot" experiment like those in Ammar et al. (2016), Tsvetkov et al. (2016), or Daiber et al. (2016) before one adds a language for which an expected feature is missing, present only in another database or not present in any database; this problem is compounded when working on genuinely less-studied languages. The initial motivation for the URIEL knowledge base and the `lang2vec` utility is to make such research easier, allowing different sources of information to be easily used together or as different experimental conditions (e.g., is it better to provide this model information about the syntactic features of the language, or the phylogenetic relationships between the languages?). Standardizing the use of this kind of information also makes it easier to replicate and expand on previous work, without needing to know how the authors processed, for example, WALS feature classes or PHOIBLE inventories into model input.

While `lang2vec` was originally conceived as providing rich language representations to "polyglot" models, it can be utilized in a variety of kinds of research projects (O'Horan et al., 2016): helping to choose "bridge" or "pivot" languages for cross-lingual transfer (Deri and Knight, 2016), directly providing feature values to systems interested in those specific features, or acting as a dataset for the prediction of unknown or un-

recorded language facts (Daumé III and Campbell, 2007; Daumé III, 2009; Coke et al., 2016). By normalizing information from a variety of data sources, it can also allow the comparison of resources, due to format and semantic differences, that were difficult to compare directly before, and help to quantify knowledge gaps concerning world languages.

3 Vector types

`lang2vec` offers a variety of vector representations of languages, of different types and derived from different sources, but all reporting feature values between 0.0 (generally representing the absence of a phenomenon or non-membership in a class) and 1.0 (generally representing the presence of a phenomenon or membership in a class). This normalization makes vectors from different sources more easily interchangeable and more easily predictable for each other (§4).

As in SSWL (Collins and Kayne, 2011), different features are not held to be mutually exclusive; the features S_SVO and S_SOV can both be 1 if both orders are normally encountered in the language.

Phylogeny, geography, and identity vectors are complete—they have no missing values, due to the nature of how they are calculated. The typological features (syntax, phonology, and inventory), however, have missing values, reflecting the coverage of the original sources; missing values are represented in the output as "−−". Predicted typological vectors (§4) attempt to impute these values based on related, neighboring, and typologically similar languages.

All vectors within the `syntax`, `phonology`, and `inventory` categories have the same dimensionality as other types of vectors in the same category, even though the sources themselves may only represent a subset of these values, to allow straightforward element-wise comparison of values. (This way, when WALS happens not to contain a feature value that SSWL does, they can easily be combined by a vector operation, without needing to track down specific feature names or go back to the original sources. In general, users will probably want to use the union or average of relevant sources, or use the `knn` predictions.)

3.1 Typological vectors

The `syntax` features are adapted (after conversion to binary features) from the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), directly from Syntactic Structures of World Languages (Collins and Kayne, 2011) (whose features are already binary), and indirectly by text-mining the short prose descriptions on typological features in Ethnologue (Lewis et al., 2015).[2]

The `phonology` features are adapted in the same manner from WALS and Ethnologue.

The phonetic `inventory` features are adapted from the PHOIBLE database, itself a collection and normalization of seven phonological databases (Moran et al., 2014; Chanard, 2006; Crothers et al., 1979; Hartell, 1993; Michael et al., 2012; Maddieson and Precoda, 1990; Ramaswami, 1999). The PHOIBLE-based features in `lang2vec` primarily represent the presence or absence of natural classes of features (e.g., interdental fricatives, voiced uvulars, etc.), with 1 representing the presence of at least one sound of that class and 0 representing absence. They are derived from PHOIBLE's phonetic inventories by extracting each segment's articulatory features using the `PanPhon` feature extractor (Mortensen et al., 2016), and using these features to determine the presense or absence of the relevant natural classes.

3.2 Phylogeny vectors

The `fam` vectors express shared membership in language families, according to the world language family tree in Glottolog (Hammarström et al., 2015). Each dimension represents a language family or branch thereof (such as "Indo-European" or "West Germanic" in Table 4).

3.3 Geography vectors

Although another component of URIEL (to be described in a future publication) provides geographical distances *between* languages, `geo` vectors express geographical location with a fixed number of dimensions and each dimension representing the same feature even when different sets of languages are considered. Each dimension represents

[2]Descriptions of well-studied typological features are often expressed formulaically in prose ("SVO", "adjective before noun", "(C)(C)v(C)", etc.), and are relatively straightforward to extract given regular expressions and some Boolean logic (e.g., if "CV" and not "CCV" and ...).

Vector type	#Languages	#Features	#Data points	%Coverage
Syntax (from sources)				
syntax_wals	1808	98	78732	44%
syntax_sswl	230	33	6404	84%
syntax_ethnologue	1336	30	18105	45%
Syntax (averaged over sources)				
syntax_avg	2654	103	94227	34%
Syntax (predicted)				
syntax_knn	7970	103	820910	100%
Phonology (from sources)				
phonology_wals	832	27	14358	64%
phonology_ethnologue	543	8	1017	23%
Phonology (averaged over sources)				
phonology_avg	1296	28	15303	42%
Phonology (predicted)				
phonology_knn	7970	28	223160	100%
Inventory (from sources)				
inventory_phoible_aa	202	158	31916	100%
inventory_phoible_gm	428	158	67624	100%
inventory_phoible_ph	404	158	63832	100%
inventory_phoible_ra	100	158	15800	100%
inventory_phoible_saphon	334	158	52772	100%
inventory_phoible_spa	219	158	34602	100%
inventory_phoible_upsid	334	158	75050	100%
Inventory (averaged over sources)				
inventory_avg	1715	158	270970	100%
Inventory (predicted)				
inventory_knn	7970	158	1259260	100%

Table 3: Typological vectors available in lang2vec, along with the number of languages and features, the number of individual data points, and the percentage of those language/feature pairs for which that data point exists.

	Indo-European	Germanic	West Germanic	Romance	North Germanic
deu	1	1	1	0	0
eng	1	1	1	0	0
fra	1	0	0	1	0
swe	1	1	0	0	1
mlg	0	0	0	0	0

Table 4: Truncated lang2vec phylogeny vectors for German, English, French, Swedish, and Malagasy, where 1 represents membership in a particular language family or branch.

the orthodromic distance—that is, the "great circle" distance—from the language in question to a fixed point on the Earth's surface. These distances are expressed as a fraction of the Earth's antipodal distance, so that values will always be in between 0.0 (directly at the fixed point) and 1.0 (at the antipode of the fixed point).

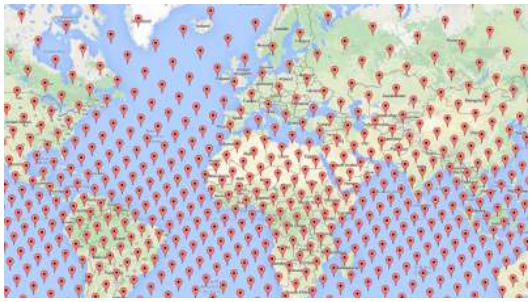

Figure 1: Example of a Fibonacci lattice overlaid on the Earth's surface, representing the "fixed points" of a `geo` vector (§3.3). (Map data: Google.)

The fixed points were derived by generating a spherical Fibonacci lattice (González, 2009; Keinert et al., 2015), a technique that approximates with high precision a uniform distribution of points on a sphere. Language points are derived from Glottolog, WALS, and SSWL's declarations of language location.[3]

3.4 Identity vectors

The `id` vector is simply a one-hot vector identifying each language. These vectors can serve as simple identifiers of languages to a system, serve as the control in an experiment in introducing (say) typological information to a system, as in Tsvetkov et al. (2016), or serve in combination with other vectors (such as `fam`) that do not always identify a language uniquely.

4 Feature prediction

One of the major difficulties in using typological features in multilingual processing is that many languages, and many features of individual languages, happen to be missing from the databases.

For example, no relevant syntactic features from Slovak were available in any of the source databases.[4] It is not a mystery, however, what sort of language Slovak is; it is probably very similar to Czech, somewhat similar to other West Slavic languages, etc. Likewise, it is probably more similar overall to nearby languages than far-away languages. [5]

The question of how we can best predict unknown typological features is a larger question (Daumé III and Campbell, 2007; Daumé III, 2009; Coke et al., 2016) than this article can capture in detail, but nonetheless we can offer a preliminary attempt at providing practically useful approximations of missing features by a k-nearest-neighbors approach. By taking an average of genetic, geographical, and feature distances between languages, and calculating a weighted 10-nearest-neighbors classification, we can predict feature missing values with an accuracy of 92.93% in 10-fold cross-validation. (We will describe these procedures, the exact notions of distance involved, alternative prediction methods that we also investigated, and their results in more detail in a future article.)

5 Conclusion

While there are many language-information resources available to NLP, their heterogeneity in format, semantics, language naming, and feature naming makes it difficult to combine them, compare them, and use them to predict missing values from each other. `lang2vec` aims to make cross-source and cross-information-type experiments straightforward by providing standardized, normalized vectors representing a variety of information types.

Acknowledgements

This project was sponsored by the Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O), program: Low Resource Languages for Emergent Incidents (LORELEI), issued by DARPA/I2O under Contract No. HR0011-15-C-0114.

[3] It should be emphasized that these points are abstractions rather than precise facts; there is no one point on Earth that best specifies "English", and no definition of the "center" of a language's area would have a known and an unambiguous answer for every language. About 2% of language codes had no corresponding geographical information in any database; we filled these in manually where possible.

[4] There are some features in WALS describing Slovak, but `lang2vec` does not index any of these.

[5] This principle cannot be trusted absolutely, of course—Slovak is in close geographic proximity to Hungarian, a very different language in many respects—but nonetheless there is almost always *some* information from which we can make a good guess at missing features.

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Akash Bharadwaj, David Mortensen, Chris Dyer, and Jaime Carbonell. 2016. Phonologically aware neural model for named entity recognition in low resource transfer settings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1472, Austin, Texas, November. Association for Computational Linguistics.

Christian Chanard. 2006. *Systmes Alphabtiques Des Langues Africaines*. UNESCO-SIL.

Reed Coke, Ben King, and Dragomir R. Radev. 2016. Classifying syntactic regularities for hundreds of languages. *Computing Research Repository*, abs/1603.08016.

Chris Collins and Richard Kayne. 2011. *Syntactic Structures of the World's Languages*. New York University, New York.

John H. Crothers, James P. Lorentz, Donald A. Sherman, and Marilyn M. Vihman. 1979. *Handbook of Phonological Data From a Sample of the World's Languages: A Report of the Stanford Phonology Archive*.

Joachim Daiber, Miloš Stanojević, and Khalil Sima'an. 2016. Universal reordering via linguistic typology. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3167–3176, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Hal Daumé III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 65–72, Prague, Czech Republic, June. Association for Computational Linguistics.

Hal Daumé III. 2009. Non-parametric Bayesian areal linguistics. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 593–601, Boulder, Colorado, June. Association for Computational Linguistics.

Aliya Deri and Kevin Knight. 2016. Grapheme-to-phoneme models for (almost) any language. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 399–408. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath. 2013. *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology.

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, Gothenburg, Sweden, April. Association for Computational Linguistics.

Álvaro González. 2009. Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices. *Mathematical Geosciences*, 42(1):49–64.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2015. *Glottolog 2.6.* Max Planck Institute for the Science of Human History, Jena.

Rhonda L. Hartell. 1993. *Alphabets des langues africaines*. UNESCO and Socit Internationale de Linguistique.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 58–68, Baltimore, Maryland, June. Association for Computational Linguistics.

Benjamin Keinert, Matthias Innmann, Michael Sänger, and Marc Stamminger. 2015. Spherical Fibonacci mapping. *ACM Transactions on Graphics*, 34(6):193:1–193:7, October.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2015. *Ethnologue: Languages of the World, Eighteenth edition*. SIL International, Dallas, Texas.

Ian Maddieson and Kristin Precoda. 1990. Updating UPSID. In *UCLA Working Papers in Phonetics*, pages 104–111. Department of Linguistics, UCLA.

Lev Michael, Tammy Stark, and Will Chang. 2012. *South American Phonological Inventory Database*. University of California, Berkeley.

Steven Moran, Daniel McCloy, and Richard Wright. 2014. *PHOIBLE Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Helen O'Horan, Yevgeni Berzak, Ivan Vulic, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027, Osaka, Japan, December. The COLING 2016 Organizing Committee.

N. Ramaswami. 1999. *Common Linguistic Features in Indian Languages: Phonetics*. Central Institute of Indian Languages.

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. pages 1357–1366, June.