What do phone embeddings learn about Phonology?

Sudheer Kolachina
sudheer.kpg08@gmail.com

Lilla Magyar
lillamagyar0929@gmail.com

Abstract

Recent work has looked at evaluation of phone embeddings using sound analogies and correlations between distinctive feature space and embedding space. It has not been clear what aspects of natural language phonology are learnt by neural network inspired distributed representational models such as word2vec. To study the kinds of phonological relationships learnt by phone embeddings, we present artificial phonology experiments that show that phone embeddings learn paradigmatic relationships such as phonemic and allophonic distribution quite well. They are also able to capture co-occurrence restrictions among vowels such as those observed in languages with vowel harmony. However, they are unable to learn co-occurrence restrictions among the class of consonants.

1 Introduction

Over the last few years, distributed representation models based on neural networks such as word2vec (Mikolov et al., 2013a) and GloVe (Pennington et al., 2014) have been of much importance in speech and natural language processing (NLP). The word2vec technique is a shallow neural network that takes a text corpus as input and outputs a vector space containing all unique words in the text. The dense vector representations of words induced using word2vec have been shown to capture multiple degrees of similarities between words. Mikolov et al. (2013a,b) show that word embeddings can solve word analogy questions and sentence completion tasks. Mikolov et al. (2013b) show that word embeddings represent words in continuous space, making it possible to perform algebraic operations, such as vector(King) − vector(Man) + vector(Woman) = vector(Queen). Considerable attention has been paid to evaluating these vector representations using human judgement datasets (Baroni et al., 2014; Levy et al., 2015). Asr and Jones (2017) use artificial language experiments to study the difference between similarity and relatedness in evaluating distributed semantic models. *Phone embeddings* induced from phonetic corpora have been used in tasks such as word inflection (Silfverberg et al., 2018) and sound sequence alignment (Sofroniev and Çöltekin, 2018). Silfverberg et al. (2018) show that dense vector representations of phones learnt using various techniques are able to solve analogies such as **p** is to **b** as **t** is to **X**, where **X** = **d**. They also show that there is a significant correlation between distinctive feature space and the phone embedding space.

Our goal in this paper is to understand better the evaluation of phone embeddings. We argue that significant correlation between distinctive feature space and phone embedding space cannot be automatically interpreted as the model's ability to capture facts about the phonology of natural language. Since many distinctive features tend to be phonetically based, natural classes denoted by these features capture *phonetic facts* as well as *phonological facts*. For example, the feature $[\pm long]$ denotes the distinction between long and short vowels, which is a language-independent phonetic fact. But, whether this distinction is a phonological fact varies from language to language. It is important to make this distinction between phonetic facts and phonological facts when evaluating phone embeddings for their learning of phonology. In this paper, we propose an alternative methodology to evaluate word2vec's ability to learn phonological facts. We define artificial languages with different kinds of phoneme-allophone distinctions and co-occurrence restrictions and study how well phone embeddings capture these relationships. Several interesting insights regarding the relationship between phonetics and phonol-

Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 160–169
Florence, Italy. August 2, 2019

ogy, the role of distinctive features and the task of distinctive feature/phoneme induction accrue from our experiments.

2 Background and Related work

One major difference between words and phones is that while words are meaningful units in language, phones have no meaning in themselves. However, as with words, there are clear patterns of organization of individual phones in a language. One well-known pattern in phonology is the distinction between **contrastive** and **complementary** distribution. Two phones are said to be in contrastive distribution if they occur in the same context and create a meaning contrast. For example, **b** and **k** occur in word-initial position and create a contrast in meaning, such as in **bæt** versus **kæt**. This is why they are considered distinct phonemes in the language. On the other hand, **p^{h}** and **p** never occur in the same context, which is referred to as being in complementary distribution. Since they are phonetically related, they are considered *allophones*, variants of the same underlying phoneme. The notions of contrastive and complementary distribution are purely based on context. They can be considered instances of paradigmatic similarity discussed in the distributed semantic literature. Allophony also involves the notion of phonetic similarity. Another pattern in natural language phonology is that of **co-occurrence restrictions**. A well-known example is homorganic consonant clusters. For example, in nasal plus stop clusters, the nasal must have identical place of articulation to the following stop. Yet another example of co-occurrence restriction in phonology is the phenomenon of vowel harmony. In some languages, a word can only have vowels which agree with respect to certain features, such as backness, rounding or height. Co-occurrence restrictions can be considered to be instances of syntagmatic similarity whereby words that frequently occur together form a syntagm (phrase). Again, most types of co-occurrence restrictions involve phonetic similarity.

The traditional method to describe phones in phonology is in terms of distinctive features (Jakobson et al., 1951). Distinctive features allow phones to be grouped into *natural classes*, which are established on the basis of participation in common phonological processes. They allow for generalizations about phonotactic contexts to be captured in an economical way. In addition to distinctive features in phonology, there are also phonetic features that describe the articulatory and acoustic properties of phones (Ladefoged and Johnson, 2010). However, in practice, there is considerable overlap between phonological distinctive features and phonetic features. This already poses an interesting question about the nature of the relationship between phonetics and phonology, which as we will see, is relevant to the evaluation of phone embeddings.

Next, let us examine the notion of correlation between distinctive feature space and phone embedding space to evaluate phone embeddings as proposed by Silfverberg et al. (2018). Pair-wise featural similarity is estimated using a metric such as Hamming distance or Jaccard index applied to feature representations of phones. Pair-wise contextual similarity is estimated as cosine similarity between phone embeddings induced using a technique like `word2vec`. The correlation between pairwise featural similarity and pairwise contextual similarity is estimated using Pearson's r or Spearman's ρ. The value of this correlation is shown for a number of languages in table 1. Data for Shona and Wargamay are taken from Hayes and Wilson (2008)[1]. Similar datasets were constructed for Telugu and the Vedic variety of Sanskrit[2]. For English, the CMU phonetic dictionary was used with a feature representation based on Parrish (2017) with some minor extensions. The `word2vec` implementation in the Gensim toolkit (Řehůřek and Sojka, 2010) was used to induce phone embeddings using the following parameters- CBOW, dimensionality of 30, window size of 4, negative sampling of 3, minimum count of 5, learning rate of 0.05. We use CBOW which predicts the most likely phone given a context of 4 phones in either direction as this is intuitively similar to the task of a phonologist. It would be interesting to compare CBOW and Skip-gram architectures and also, study the effect of different parameters on this correlation between distinctive feature space and phone embedding space. However, this is not the goal of our study. In this paper, we restrict our attention to the linguistic significance of this correlation.

All languages in Table 1 show a significant positive correlation between distinctive feature space

[1] https://linguistics.ucla.edu/people/hayes/Phonotactics/index.htm#simulations

[2] Datasets and code available at https://github.com/skolachi/sigmorphoncode

Language	Size	Pearson	Spearman
English	135091	0.589	0.612
Shona	4395	0.431	0.575
Telugu	19627	0.349	0.350
Wargamay	5910	0.411	0.428
Vedic	45334	0.351	0.285
English	4000	0.129	0.161
Shona	4000	0.507	0.533
Telugu	4000	0.202	0.206
Wargamay	4000	0.219	0.387
Vedic	4000	0.146	0.159

Table 1: Correlation between distinctive feature space and embedding space, all values significant ($p < 0.01$)

and embedding space. What is the physical interpretation of this correlation? Firstly, it is important to note the use of this correlation to evaluate phone embeddings presupposes that these handcrafted distinctive features are the gold standard descriptions of the phonology of these languages. Even if this were the case, the kind of distinctive features used to describe phones plays an important role in the interpretation of this correlation. If feature specifications of phones are based mostly on their phonetic properties, a positive correlation between featural space and embedding space indicates that phonetically similar phones tend to occur in similar contexts. In other words, the natural classes of phonology are tightly constrained by phonetics. To illustrate this point, we take the example of Wargamay natural classes derived from the distinctive features of Hayes and Wilson (2008) shown in Table 2. Examining the pairwise cosine similarities of phones based on embeddings induced by `word2vec` in the agglomerative clustering (WPGMA) dendrogram heatmap shown in Figure 1, `word2vec` CBOW embeddings identify the following natural classes- ii1, uu1, aa1 ($[+long, +main, +stress]$), i1, u1, a1 ($[-long, +main, +stress]$), i2, u2, a2 ($[-long, -main, +stress]$), i0, u0, a0 ($[-long, -stress]$) and $[-syllabic]$ which denotes the set of all consonants. Among the set of consonants, the velar consonants N, g ($[+dorsal]$) show up in the same cluster, as do the bilabials b and m. Sonorant consonants like R, l, n, w form one cluster and $[+approximant]$ r, y form another cluster. Notice that all these classes are based on place and manner of articulation. Therefore, it is not clear if the observed clustering is to interpreted as the model's learning of phonology or the fact phonetic features strictly constrain the contexts in which phones occur. Furthermore, as with word meaning, when embeddings of two phones show high similarity, it is not clear if it is an instance of paradigmatic similarity (phonemic relationship) or syntagmatic similarity (co-occurrence restriction).

Feature	Class
-high	a0,a1,a2,aa1
+high	i0,i1,i2,ii1,u0,u1,u2,uu1,w,y
+long	aa1,ii1,uu1
-long	a0,a1,a2,i0,i1,i2,u0,u1,u2
+back	a0,a1,a2,aa1,u0,u1,u2,uu1,w
-back	i0,i1,i2,ii1,y
-approximant	N,b,d,g,j,m,n,nj
+approximant	R,a0,a1,a2,aa1,i0,i1,i2,ii1,l,r,u0,u1,u2,uu1,w,y
-sonorant	b,d,g,j
+sonorant	N,R,a0,a1,a2,aa1,i0,i1,i2,ii1,l,m,n,nj,r,u0,u1,u2,uu1,w,y
+syllabic	a0,a1,a2,aa1,i0,i1,i2,ii1,u0,u1,u2,uu1
-syllabic	N,R,b,d,g,j,l,m,n,nj,r,w,y
+main	a1,aa1,i1,ii1,u1,uu1
-main	a0,a2,i0,i2,u0,u2
+stress	a1,a2,aa1,i1,i2,ii1,u1,u2,uu1
-stress	a0,i0,u0
-consonantal	a0,a1,a2,aa1,i0,i1,i2,ii1,u0,u1,u2,uu1,w,y
+consonantal	N,R,b,d,g,j,l,m,n,nj,r
+anterior	d,l,n,r
-anterior	R,j,nj,y
+lateral	l
-lateral	R,r
+coronal	R,d,j,l,n,nj,r,y
+dorsal	N,g
+labial	b,m

Table 2: Natural classes derived from distinctive features

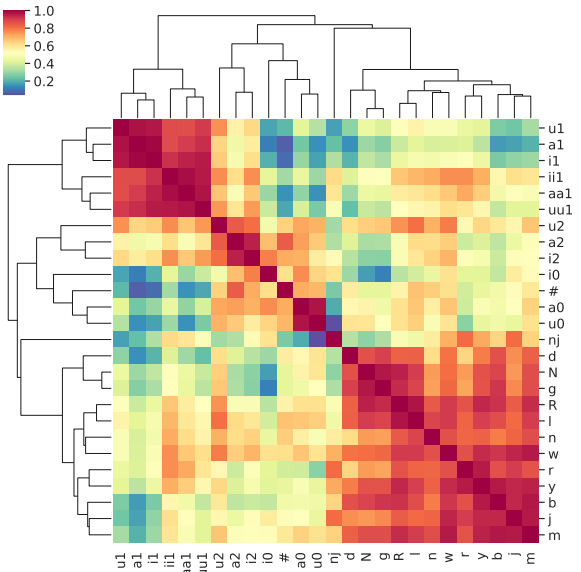


Figure 1: Phone clusters of Wargamay

Asr and Jones (2017) use an artificial language experiment to study the difference in performance of word embeddings between paradigmatic and syntagmatic tasks. In section 3, we propose a similar approach to study `word2vec`'s ability to learn

different kinds of phonological patterns. While natural language phonology can be complex with many interleaved phenomena, artificial language phonology makes it possible to test learning of each pattern independently. In addition, previous work on phonological learning such as Hayes and Wilson (2008) assumes that distinctive features exist *a priori*. In our experiments with artificial languages, we explore the possibility of deriving distinctive features from phone embeddings which capture contextual distributions of phones.

3 Learning artificial phonology with `word2vec`

In this section, we present experiments with `word2vec` on learning artificial languages with different kinds of phonological relationships. The languages studied in this experiment are described below. The minimal word is bimoraic CVC. The maximum word length is set at three syllables. Word boundary is indicated using #.

1. **Language 1** contains only open (CV) syllables in polysyllabic words. Monosyllabic words are all CVC. The set of possible consonants is **p t k** and the set of possible vowels is **a e i o u**.

2. **Language 2** is the same as Language 1 with the difference that intervocalic consonants are voiced- **b d g** instead of **p t k**. In other words, there is allophonic variation within the class of consonants.

3. **Language 3** is the same as Language 2 with the following differences: Final syllables in polysyllabic words are optionally closed, that is, codas are allowed. Word-initial consonants are aspirated, **P T K**. Word-final consonants are voiceless **p t k**. Thus, an additional degree of allophony for consonants is introduced.

4. **Language 4** is the same as Language 3 with the addition of nasal codas: **m n N (ŋ)** in all syllables. In the final syllable, the nasal and the voiceless stop form a coda cluster.

5. **Language 5** is the same as Language 4 with the difference that nasal codas are optional. This language is the union of Languages 3 and 4.

6. **Language 6** is the same as Language 5 with a restriction on nasal coda based on the place of articulation of the following voiced consonant. In other words, only **mb nd Ng** combinations are allowed.

7. **Language 7** is the same as Language 6 with the addition that **r** is optionally allowed following a voiced consonant. In other words, onset clusters **br dr gr** are permitted in medial syllables.

8. **Language 8** is the same as Language 7 with the addition that a sibilant **s** is optionally allowed in the coda position of the final syllable. This language allows a variety of contexts in the final syllable- voiceless stops, nasals and nasal+stop clusters, sibilant **s**, sibilant+stop clusters **sp st sk** and also nasal+sibilant+stop clusters.

9. **Language 9** is the same as Language 8 with the restriction that the nasal + sibilant + voiceless stop cluster in coda position must be homorganic- only **nst** is allowed.

10. **Language 10** is the same as Language 9 with the restriction that only high vowels **i u** can occur in initial syllables.

11. **Language 11** is the same as Language 10 with the difference that it has vowel harmony with respect to backness. Thus, words can only have either $[-back]$ (front) vowels **i e** or $[+back]$ vowels **u o**.

12. **Language 12** is the same as Language 11 with the difference that the transparent vowel **a** is permitted in non-initial syllables of polysyllabic words.

Phone embeddings were induced using the same parameters as in the previous section- CBOW, dimensionality 30, context window 4, negative sampling 3, minimum count 5 and learning rate 0.05. The number of words in each language is shown in table 3, alongside the correlations between distinctive feature space and embedding space. A set of distinctive features similar to those of Hayes and Wilson (2008) are used to estimate these correlations. Since the value of cosine similarity is bounded on $[-1, 1]$, we also use Euclidean distance to estimate correlation between contextual similarity based on phone embeddings

and featural similarity. We will return to the issue of the significance of these correlations shortly.

Language	size	Pearson's r	
		Cosine	Euclidean
Language 1	3645	0.873	0.882
Language 2	3645	0.632	0.408
Language 3	14445	0.573	0.396
Language 4	372780	0.477	0.362
Language 5	878625	0.470	0.354
Language 6	139635	0.503	0.343
Language 7	549135	0.500	0.305
Language 8	988455	0.394	0.263
Language 9	878625	0.421	0.254
Language 10	351450	0.481	0.286
Language 11	57690	0.476	0.277
Language 12	127962	0.430	0.209

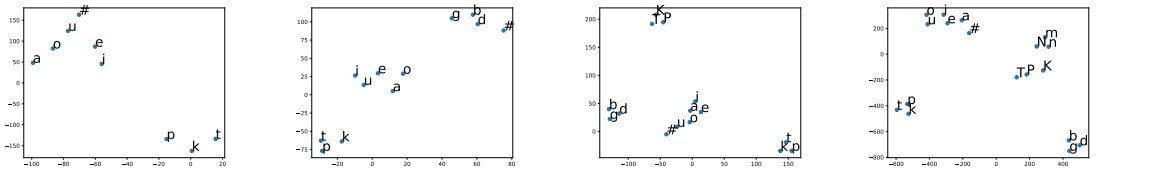
Table 3: Correlation between embedding and distinctive feature space, all values significant at $p < 0.01$

As can be noticed from the descriptions, each language defines different sets of equivalence relations among phones based on the contexts in which they occur. For example, in Language 3, aspirated stops occur word-initially, voiced stops occur inter-vocalically and voiceless stops occur word-finally. The task of phonology is to capture generalizations about these *natural classes*. Notice that although these natural classes are based on phonetic features such as aspiration and voicing, `word2vec` has no access to these features. The goal of our experiments is to investigate the extent to which these natural classes can be inferred solely based on phone embeddings. The embedding space for each language is visualized using T-distributed Stochastic Neighbor Embedding (t-SNE) plots. Multiple plots were generated for different values of perplexity and learning rate using the implementation in scikit-learn toolkit (Buitinck et al., 2013). The plots shown in Figure 2 correspond to perplexity 3 and learning rate 100. In addition, phone clusters derived using agglomerative clustering of cosine similarities between phone embeddings are also shown. Euclidean distance was used to plot the dendrogram heatmaps[3].

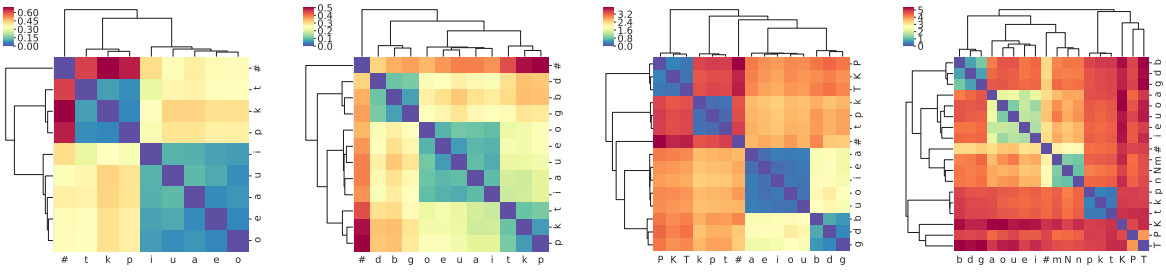
From the plots, we observe that phone embeddings capture the different context classes with varying degrees of success. Languages 1-3 were designed with unique contexts for each class of phones and the embeddings show clear separation between these classes. In Language 4-5, where nasal codas are allowed, the t-SNE plot shows less separation between nasal codas and word-initial aspirated voiceless stops. This is due to the fact that in monosyllabic words, aspirated stops and nasals co-occur within the same context (bimoraic) window. This is an unintended co-occurrence restriction learnt by `word2vec`. However, this pattern in monosyllabic words has no effect on the phone clusters in the dendrogram. Nasals and aspirated stops form separate clusters in the dendrogram. In Language 6, a co-occurrence constraint that nasal obstruent clusters be homorganic was introduced. Interestingly, the t-SNE plot for this language has nasals showing up with vowels. The syntagmatic relationship (co-occurrence restriction) between nasals and homorganic voiced obstruents introduced in this language is not seen in the t-SNE plot of the embedding space. But, the dendrogram heatmap for this language shows nasals and voiced obstruents forming a high-level cluster. It is plausible that with hyperparameter tuning, co-occurrence restrictions such as nasal-voiced obstruent clusters are captured even in the t-SNE plots of embedding space. Co-occurrence restrictions in phonology are much more rigid than word relatedness since the size of the phone inventory in a language is many degrees smaller than the size of the vocabulary.

A similar pattern is observed with languages 7, 8 and 9, where other kinds of co-occurrence relations between consonants are introduced. The t-SNE plot for Language 7 fails to capture the onset clusters **br dr gr** introduced in this language. The lateral **r** shows up with the word boundary. The dendrogram for this language fails to recover word-initial aspirated stops as a separate class. In Language 8, the introduction of the optional sibilant in the coda position of the final syllable has a same effect on the embedding space as visualized by the t-SNE plot. Nasals, aspirated stops, lateral, sibilant and word boundary are less separated in the t-SNE plot. In the dendrogram plot, the sibilant forms a cluster with the nasals and word boundary. Both the t-SNE and dendrogram plots for Language 9 are almost identical to those Language 8 indicating that the homorganic restriction on nasal sibilant voiceless stop clusters in the final syllable has no effect on the embedding space. In other words, phone embeddings are unable to learn these co-occurrence restrictions. Languages

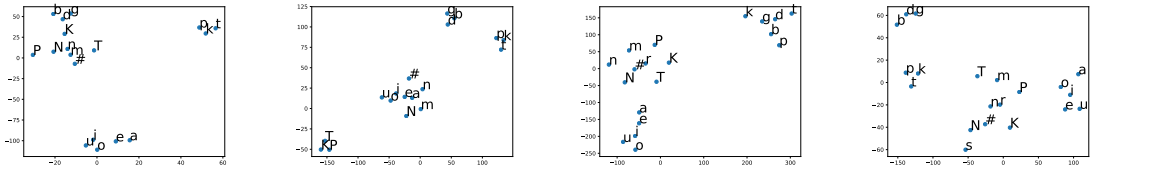
[3]The interpretation of these distance-based heatmaps differs from the cosine similarity-based heatmap of Wargamay presented in the previous section.



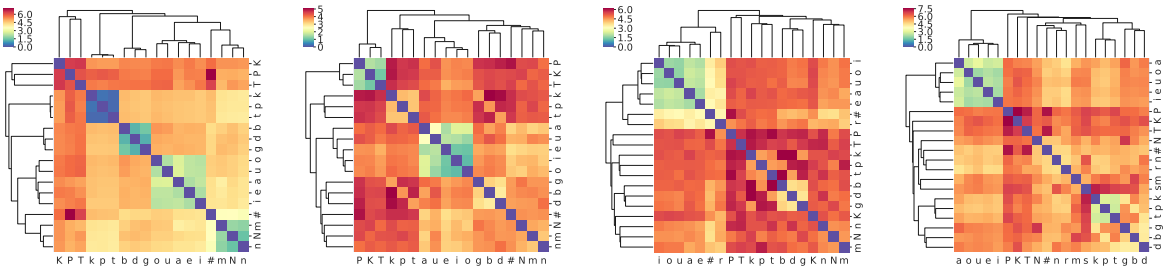
Language 1 Language 2 Language 3 Language 4



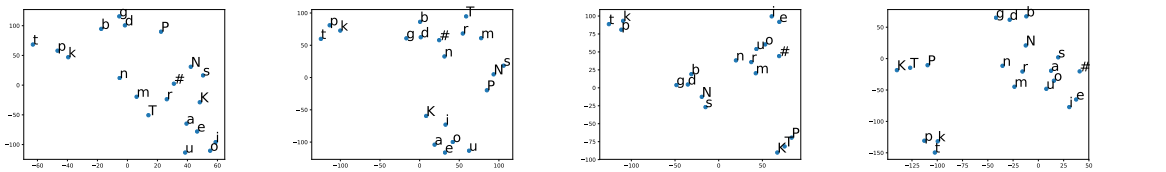
Language 1 Language 2 Language 3 Language 4



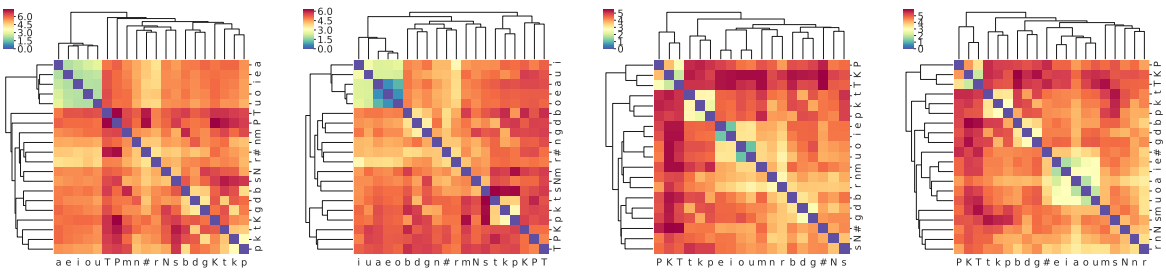
Language 5 Language 6 Language 7 Language 8



Language 5 Language 6 Language 7 Language 8



Language 9 Language 10 Language 11 Language 12



Language 9 Language 10 Language 11 Language 12

Figure 2: Embedding space of artificial languages

10-12 introduce contextual restrictions on vowels. In Language 10, only high vowels occur in the word-initial position and phone embeddings capture this distinct class of vowels as shown by the dendrogram heatmap. Languages 11 and 12 show a similar pattern with respect to a different feature, backness. Both of them are harmony languages, which still obey the constraint that vowels in initial syllables must be $[+high]$. Interestingly, vowels cluster with respect to $[\pm back]$ rather than $[\pm high]$ as can be seen from the plots. Evidence for agreement between vowels with respect to backness is three times more frequent than the evidence with respect to agreement between vowels in initial syllable with respect to height. Although vowel harmony is also an instance of co-occurrence restriction (syntagmatic relationship), `word2vec` infers these classes accurately. The number of vowels in a language tends to be much lower than the number of consonants. And therefore, it seems that a co-occurrence restriction between vowels is a relatively larger sample of the set of all possible vowel sequences ($5*5*5 = 125$ in this language) compared to a co-occurrence restriction between two or more consonants. The transparent vowel **a** has no effect on the distances between the other vowels in Language 12.

The ability of phone embeddings to learn phonology in our artificial language experiments can be summarized as follows-

1. Phone embeddings are able to capture paradigmatic relationships among phones very well. For example, word-initial aspirated stops, intervocalic voiced stops, word-final voiceless stops and vowels are recovered as separate classes in most languages.

2. Phone embeddings are also able to capture positional restrictions as well as co-occurrence restrictions on vowels as shown by Languages 10-12.

3. Phone embeddings are not able to capture co-occurrence restrictions among consonants such as homorganic nasal-voiced obstruent clusters, voiced obstruent-lateral cluster and homorganic nasal-sibilant-voiceless stop clusters. This observation is similar to one reported in the distributed semantic literature that word embeddings capture similarity better than relatedness (Asr et al., 2018). Based on insights from the word embedding literature, context embeddings denoted by the hidden to output layer weight matrix, are supposed to be able to capture better syntagmatic relationships like co-occurrence restrictions. In addition, it is plausible that these co-occurrence restrictions among consonants can be learnt using autosegmental tier-based representations. We leave this investigation to future work.

4 Distinctive Features and Phoneme Induction

The main argument of this paper is that phone embeddings should be evaluated in terms of their ability to capture phonological relationships. Applying this bottom-up approach to natural language phonology is not straightforward since the full set of phonological relationships is not known beforehand. Even the method of evaluating phone embeddings using the correlation between distinctive feature space and phone embedding space, as mentioned earlier, presupposes that the gold standard specification of distinctive features for that particular language is known. However, this is seldom the case. Natural languages are highly complex with processes such as borrowing, loanword adaptation and language changes such as drift. This is why experimenting with artificial phonology can be informative.

The artificial languages in our experiment had increasing levels of complexity, since the goal was to tease apart learnability of different phenomena. Recall that a fixed set of distinctive features along the lines of Hayes and Wilson (2008) was used to estimate the correlation between distinctive feature space and phone embedding space. Notice in table 3 that the value of this correlation goes down as we move from Language 1 to Language 12 regardless of the distance metric used to estimate distance between embeddings. Unlike the cross-linguistic comparison in section 2, the distinctive features are the same across languages. We observe that as the size of the phone inventory and the number of distinct context classes increase, the degree of correlation between feature space and embedding space decreases. How can this trend be accounted for? Examining the distances in the clustermaps, we observe that as the number of context classes goes up, intra-phone distances, especially among the class of consonants tend to increase. This can be noticed by comparing

the clusters corresponding to voiceless consonants and vowels between Language 1 and Language 12. Given the continuous space nature of phone embeddings and the dimensionality reduction property of `word2vec`, this is expected. When the weights of the neural network corresponding to a particular phone or phone-sequence are adjusted, the changes affect similar items (Mikolov et al., 2013b). This inverse "dispersion" effect is also relevant to the correlation between distinctive feature space and embedding space- the value of featural distance between phones is constant across languages when estimated using a fixed distinctive feature representation. But, as the number of context classes increases, distances between phone embeddings increase and the cumulative effect on the correlation between phonetic space and embedding space is downward. Thus, this correlation value clearly cannot be used as an evaluation metric for cross-linguistic comparison. Even within a language, a higher correlation value does not necessarily indicate better learning of phonology/phonetics. Rather it indicates a low inverse dispersion effect. One way to interpret the results of Silfverberg et al. (2018, pp.140) is that phone classes based on context are much less spread out in embedding space when learnt using supervised RNN compared to `word2vec`. At best, this can be interpreted as a difference in the dimensionality reduction properties of the two techniques.

This also raises an interesting question about the degree of specification of phones. Phonologists assume a language independent feature specification of phones. The results of our experiments suggest the following possibility- could the granularity of feature specification be dependent on how separable the different classes of phones are in embedding space? In other words, do learners infer distinctive features of phones based on the contexts in which they occur? If certain phone classes can be inferred purely based on context, phonetic features that distinguish these classes can be underspecified. For example, in Language 10, the difference between high and non-high vowels in a language could be inferred based on context. For such a language, is it necessary to include height ($[\pm high]$) as a distinctive feature? Intuitively, the task of distinctive feature induction is related to phoneme induction.

A quantitative approach to phoneme induction based on phone embeddings and phonetic features

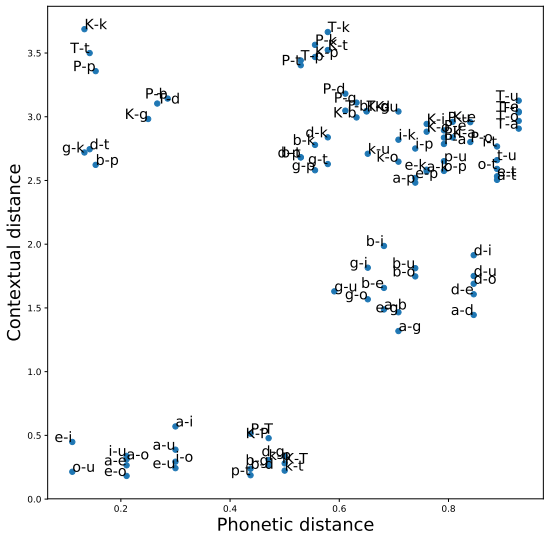


Figure 3: Contextual distance versus Phonetic distance

can be outlined as follows. If embeddings of two phones show low similarity (or high distance), their contexts are very different. If the phones show a high degree of phonetic similarity, then this is very likely to be a case of allophony. If embeddings of two phones show high degree of similarity (or low distance), then their contexts are very similar. If the phones show low degree of phonetic similarity, these are clearly two distinct phonemes in the language. If the phones also show a high degree of phonetic similarity, then this could be either an instance of a phonemic relationship or a co-occurrence restriction. The feature specifications of such phones can be compared to discover distinctive features of phonology. If no such feature is found, it means the default phonetic feature specification is too coarse-grained. If more than one distinctive feature is found, the feature specification is too fine-grained. The exact feature corresponding to the contrast between two phones can be discovered by iterating over the full set of features of the two phones and checking if leaving out a particular feature leads to a drop in the overall correlation between distinctive feature space and embedding space. These ideas are illustrated by the plots in Figures 3 and 4. Figure 3 shows a scatter plot of phone pairs along the phonetic distance-contextual distance axes for Language 3 in the artificial language experiment. Allophonic phone pairs such as **P**-**p**, **p**-**b**, **T**-**t**, **t**-**d**, **K**-**k**, **k**-**g**, etc. show up at the top left corner of the scatter

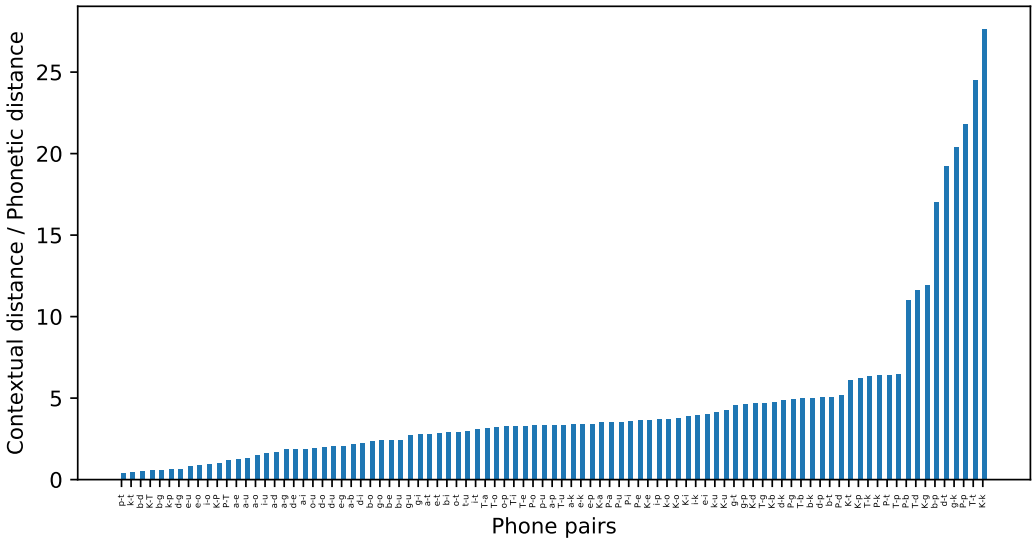


Figure 4: Allophonic index derived from embeddings

plot. The phonetic feature specifications of these pairs can be compared to discover that voicing and aspiration are not phonemic in this language. Similarly, phone pairs that show up at the bottom left corner of this plot such as the 10 pairs of vowels and **P-T**, **P-K**, **K-T**, **p-t**, **t-k**, **p-k**, **b-d**, **d-g** and **g-b** are all phonemic contrasts. The phonetic specifications of these phone pairs can be compared to discover that both height and backness are contrastive for vowels and place of articulation is contrastive for consonants. The remaining phone pairs in the top right corner of the scatter plot are all phonemic contrasts. However, they might not yield any new distinctive features. The bar plot in Figure 4 is another way of visualizing the usefulness of distances between phone embeddings to identify phonemic versus allophonic relationships. We define allophonic index as the ratio of contextual distance estimated using phone embeddings to phonetic distance. The higher the value of this index for a phone pair, the more likely the pair is to be allophonic. The sorted bar plot in Figure 4 corresponding to artificial Language 3 shows allophonic pairs at the right edge and phonemic pairs at the left edge. A precise formulation of a phoneme/distinctive feature induction algorithm based on these metrics is reserved for future work.

5 Conclusions and Future work

This paper presents a discussion of evaluation of phone embeddings. Artificial language experiments are used to study `word2vec`'s ability to learn different kinds of phonological relationships. The results show that phone embeddings are able to capture phonemic and allophonic relationships quite well. Phone embeddings are also able to capture co-occurrence restrictions among vowels found in harmony languages. Phone embeddings do not perform well on capturing co-occurrence restrictions among consonants. The experimental results also show an interesting correlation between size and complexity of phone inventory and magnitude of inter-phone distances based on phone embeddings. An analysis of the limitation of correlation between embedding space and distinctive feature space to evaluate phone embeddings for their learning of phonology is also provided. The analytical framework presented here and the proposal for distinctive feature induction will be developed in future work and can be applied to diverse problems ranging from bootstrapping pronunciations of OOV words in ASR to modeling historical phonology. A similar analysis of sound analogies is required to better understand their significance to phonology.

6 Acknowledgements

We thank Giorgio Magri and Mark Steedman for useful comments and discussion. Thanks are also due to the anonymous reviewers for their much useful feedback.

References

Fatemeh Torabi Asr and Michael Jones. 2017. An artificial language evaluation of distributional semantic models. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 134–142. Association for Computational Linguistics.

Fatemeh Torabi Asr, Robert Zinkov, and Michael Jones. 2018. Querying word embeddings for similarity and relatedness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 675–684. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247. Association for Computational Linguistics.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic inquiry*, 39(3):379–440.

Roman Jakobson, C Gunnar Fant, and Morris Halle. 1951. *Preliminaries to speech analysis: The distinctive features and their correlates*. MIT press.

Peter Ladefoged and Keith Johnson. 2010. *A course in Phonetics*. Thomson Wadsworth Boston.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751.

Allison Parrish. 2017. Poetic sound similarity vectors using phonetic features. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Miikka P Silfverberg, Lingshuang Mao, and Mans Hulden. 2018. Sound analogies with phoneme embeddings. *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 136–144.

Pavel Sofroniev and Çağrı Çöltekin. 2018. Phonetic vector representations for sound sequence alignment. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 111–116.