The Original Sound of Clustering

Comparison of language classification based on pronunciation and origin

Isa Rethans – 02-07-2021
Supervised by Marco Loog; Tom Viering; Arman Naseri Jahfari; Stavros Makrodimitris

1 Background

- Language similarity useful for data enrichment in Natural Language Processing
- IPA - international phonetic alphabet - generic for all languages, used for written pronunciation
- Combination of IPA and clustering could be efficient way of language classification

2 Problem

How does data-driven language classification using IPA and clustering compare to linguistic language classification based on origin?

3 Method

I – Dataset creation

ipa-dict - word to pronunciation mapping
wiktextract - dictionary including information for pronunciation for some words

word	ipa	Iso-code
hello	həlˈəʊ	en
world	wˈɜːld	en

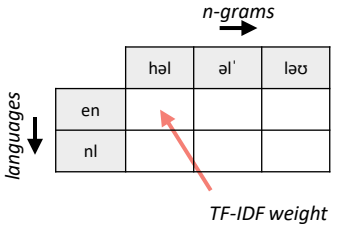
II – Data preparation

n-grams - Sequence of *n* IPA symbols representing sound combinations in language

həl	ələ	ləʊ

Pronunciation of hello in 3-grams

TFIDF – weight indicating the importance of an n-gram in a language compared to the other languages



III - Clustering

Singular Value Decomposition - dimensionality reduction method
k-means clustering - partition data into *k* clusters

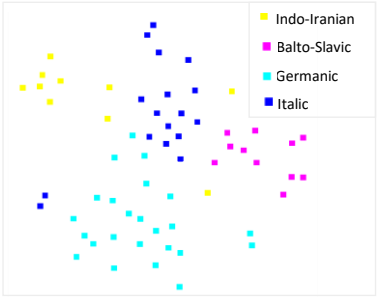
4 Approach

Compare k-means clustering with k=4 to four big language groups that linguists describe within the Indo-European language family

5 Results

t-SNE – dimensionality reduction used for visualisation

Figure showing the data distribution of languages with vectors based on 2-grams

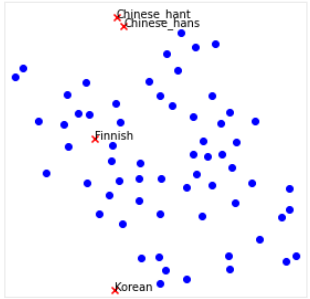


Adjusted Rand Index (ARI) - measure for partition similarity. Ranges from -1 to 1 with 0 as expected value for two random partitions.

		n for n-grams			
		1	*2*	*3*	*4*
Adjusted Rand Index	min	0.2298	0.1138	0.0588	0.0132
	max	0.6889	0.8053	0.5903	0.4513
	mean	0.4304	0.4491	0.2716	0.1170

ARI for comparison of k-means and linguistic partitions calculated over 100 iterations of k-means clustering

t-SNE plot showing the position of non – Indo European languages as compared to the position of Indo European languages



6 Conclusions

- Relative even spread of data over the feature space
- Significance in relationship between language classification based on pronunciation and origin
- Source for much more research

7 Future work

- More in depth analysis of k-means
- Another approach using hierarchical clustering
- Extension by involving other language families
- Use of dataset and vectorization method for other research related to language pronunciation