The role of similarity in non-local dependencies

Jason Riggle

Information-Theoretic Approaches to Linguistics
LSA Summer Institute, Boulder CO, July 16

Idea

(1) Non-local dependencies in phonology can be captured with a simple (albeit large) class of models.

(2) Similarity can serve as a guide that makes it feasible to work with this class.

These conjectures could both be somewhat wrong and yet still interesting.

Execution

- How do we measure similarity?
- How do we incorporate similarity into models of non-local dependencies?
- How do we evaluate such models?

These questions are interesting regardless of the status of conjectures 1 and 2.

Motivation

- Success (even for a few cases) provides a tool that allows unsupervised learning of non-local phonological patterns.
- But, the methods are equally important: the information-theoretic/algorithmic approach is:
  a. **axiomatically defined**,
  b. **replicable**, and
  c. **transparent**

Motivation

- Mathematical methods do not obscure the linguistic insights. On the contrary, they clarify the differences between proposed models and make the relationship between the models and insights more transparent.

Local and non-local dependencies

- A strictly local pattern is a pattern that can be described with n-grams (for reasonable n).
- Many observed phenomena of natural language (e.g., harmony) are not strictly local.
- This is unfortunate because locals are cool.

Local and non-local dependencies

Yes, there are are n-grams that are large enough to capture all actual utterances because there is a longest sentence ever uttered. But this misses the point in two crucial ways:

1. This bound is an accident of flesh, not a deep property of grammar (debate, debate, debate, ...)
2. More straightforwardly, such a bound doesn't help us at all in the task at hand because, for a set of s symbols, there are s^n strings of length n.

Relative non-locality

- Idea: phonemes can be members of several (potentially overlapping) classes, tiers, etc.
- Grammars refer to n-grams (i.e., strictly local) that are indexed to the classes (i.e., relative).
- The probability of a form is a combination of its probability on several tiers; on each tier the elements ignore intervening material not on the tier. (see Goldsmith & Riggle, to appear)

Success and peril

- Autosegmental models succeed at making the non-local interactions relatively-local.
- But, if there are n phonemes in the inventory then there are 2^n subsets of the segments; this is way too many to evaluate all of them.
- How do we tame this class of models?
- There are s-chose-n relatively local models for inventories of s segments that split n of them off onto their own tier.

Success and peril

How do we tame this class of models?

25 choose n	n
300	2
2,300	3
12,650	4
53,130	5
177,100	6
480,700	7
1,081,575	8
2,042,975	9
3,268,760	10

Guidance

- Many linguists have observed that non-local dependencies target "similar" segments.

  Pierrehumbert (1993), Frisch (1996), Frisch et al. (1997), MacEachern (1999), Zuraw (2002), Rose & Walker (2004) ... and others
- The critical question is how similarity is to be measured and whether the same or different metrics are needed for various phenomena.

Distribution, articulation, perception

- Classes can be derived via clustering based on:
  a. distributional similarity
     (e.g., KL divergence)
  b. articulatory similarity
     (e.g., shared features)
  c. perceptual similarity
     (e.g., confusion matrices)

Distribution, articulation, perception

- Basing similarity on features is tricky:

  "However, linguistics and psycholinguistics have seen a range of competing feature systems without any one having emerged as an outright winner, and some researchers have abandoned attempts to theoretically derive metrics of phoneme similarity in favor of the use of phoneme confusability as an indicator of similarity..." — Bailey & Hahn (2005)

Introduction Locality Similarity Distance Clustering Evaluation Conclusions

Distribution, articulation, perception

- Basing similarity on features is tricky:

  "Featural representations of speech sounds were developed originally to explain natural classes, that is, sets of phonemes that behave like members of a single category in phonological patterns"

  "Nevertheless, they are articulatory, not acoustic features, so it is possible that they are more relevant to short-term memory and speech production tasks than to word recognition tasks." — Bailey & Hahn (2005)

Introduction Locality Similarity Distance Clustering Evaluation Conclusions

Which is best for what (when and why)?

- Rose & Walker observe that [son], [cont] and place features are the most important in similarity for harmony, and that these are not the features which agree at a distance.
- Clearly there is room for many factors specific to speech, but distribution will often reflect these and is accessible immediately.
- The fact that distribution may be derivative of these factors makes it a good starting point.

Introduction Locality Similarity Distance Clustering Evaluation Conclusions

Conditional probability

- The probability of some event *y*, given that some other event *x* has occurred.
- In the case of bigrams, this is the probability of a particular phone, given the one that preceded it.

$$P(y|x) = \frac{P(xy)}{P(x)}$$

KL-divergence (relative entropy)

- Given two probability distributions P and Q, $D_{KL}(P||Q)$ is the expected number of extra bits required to encode samples from P with a code that is based on distribution Q.

$$KL(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

KL-divergence (relative entropy)

Consider the problem of modeling the distribution over a random variable X given a set of instances $x_i, i = 1...m$ where $\hat{P}(X)$ is the distribution corresponding to observed counts of the values of X.

If one were attempting to do this with a parameterized family of distributions, P_θ, the **maximum likelihood** value for θ is precisely the value of P_θ that has the **minimum KL-divergence** from $\hat{P}$.

$$\arg\max_\theta \sum_{i=1}^{m} \log P_\theta(x_i) = \arg\min_\theta KL(\hat{P}||P_\theta)$$

Conditional divergence

- Since conditional probability distributions are probability distributions, we can measure the KL-divergence between two of them.

For two conditional distributions $P(Y|X)$, $Q(Y|X)$:

$$KL(P||Q) = \sum_x P(x) \left(\sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)} \right)$$

Conditional divergence (as a difference of divergences)

- This can be rewritten as the difference between the KL-divergence of the bigrams and the KL-divergence of the unigrams.

For two conditional distributions $P(Y|X)$, $Q(Y|X)$:

$$KL(P||Q) = KL(P(X,Y)||Q(X,Y)) - KL(P(X)||P(Y))$$

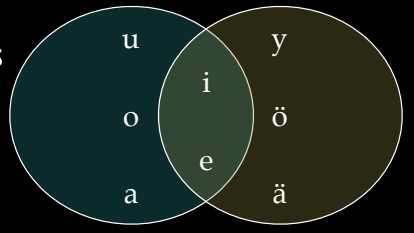
Swap Distance

- The **swap distance** for a pair of symbols in the inventory is the average cost (in bits) of systematically confusing the two.
- If P is the conditional distribution over the corpus. Each pair of symbols (a, b) yields a swap-distribution $Q_{(a,b)}$ by re-labeling each instance of 'b' as 'a' and vice versa.
- The swap distance for the pair (a, b) is then the KL-divergence of P and $Q_{(a,b)}$.

Illustration

- Assigning sounds (or other linguistic structures) to groups is a familiar part of linguistic analysis.

E.g., Finnish Vowels

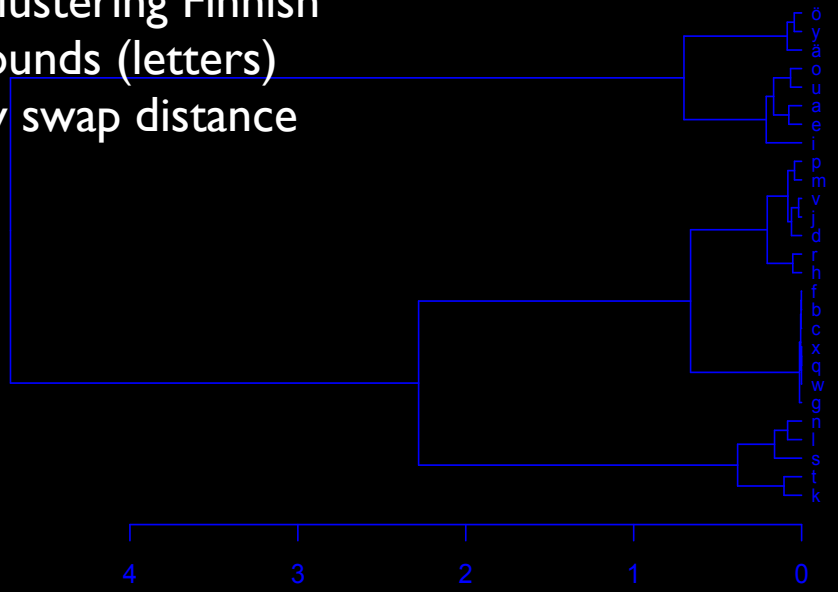


- Clustering algorithms derive the groups based on a measure of distance between the items.

Ward's strategy

- Ward's agglomerative hierarchical clustering strategy works by successively merging nodes (groups of elements) in a way that minimizes the loss in predictive power for the distances.
  1. start with each letter in its own cluster
  2. iteratively merge pairs of clusters in a way that minimizes the variance within each cluster (summed across all clusters)
  3. stop when only one cluster remains

Clustering Finnish sounds (letters) by swap distance



Linguistic insight

- Clustering on similarity yields categories, classes, features, tiers...
- Allowing these clusters to interact non-locally is where we add linguistics (i.e., domain specific) insights.
- But these should still be grounded if possible.
- see, e.g., Auditory Scene Analysis

Asymmetry?

- Some properties of interest are not symmetric.
- Confusability, KL-divergence, ...
- In general using asymmetric measures can make it NP-hard to do optimal clustering.
- Unfortunately, the operation of symmetrizing measures like confusability will hide some properties that seem likely to be relevant.

How far do we go?

- How many clusters do we evaluate?
- What threshold of difference is relevant?
- What constrains this process?

Maximizing the probability of the data

- With a proposed set of segments to separate onto their own tier, we can recompute the probability of the corpus.

  If it goes up (cost in bits goes down) then the tier is encoding information.
- If we assume (following Goldsmith & Riggle, to appear) that the conditional probabilities for bigrams are modified by mutual information on other tiers (rather than fitting MLE values), then some tiers make the model worse.

Amidoinitrite?

- How do we know that adding the non-local interactions is helping (in the right way)?
- Are we getting at something that's linguistically interesting or simply adding parameters?

saves 23,186	ä ö y _ e a o u	
saves 22,981	ä ö y _ _ a o u	(205 bits off best)
saves 20,540	ä ö y _ _ a _ u	
saves 20,426	ä ö y _ _ a o _	
saves 19,810	ä _ y _ e a o u	
saves 19,755	ä ö y i _ a o u	
saves 19,608	ä ö y i e a o u	(3,578 bits from best)

Idea recap

- Unsupervised learning of non-local patterns
  - conjecture 1: **relative locality**
  - conjecture 2: **similarity-based**
- It works for Finnish (and also for Turkish and probably Hungarian).
- Consonantal harmony and/or co-occurrence restrictions will allow easier comparisons of the role of different kinds of similarity.

Execution & benefit

- A metric of distributional similarity:
  - swap distance
- A strategy for discovering tiers:
  - clustering (Ward)
- A metric for model selection:
  - information theory
- This approach to non-local dependencies is axiomatically defined, replicable & transparent.

Thank You

Thank You

And many thanks to our hosts and organizers for putting this workshop together.