Unsupervised Consonant-Vowel Prediction over Hundreds of Languages

Young-Bum Kim and Benjamin Snyder
Wisconsin-Madison
[ACL 2010] Computational Decipherment of Ugaritic

⇒ Leverages related languages
Remaining Challenges

- Language family unknown
- No related living languages
  - Linear A
  - Linear Elamite
  - Byblos
  - Cypro-Minoan
  - Isthmian
- Syllabic writing systems

Undeciphered:
Linear A
Trigram HMM over character sequences
- Two hidden states
- Estimated using EM

Distinguishes Spanish consonants/vowels
Lingering Questions

1. Just Spanish?
   - Georgian: prtskvna (peeling) : CCCCCCCCV

2. Leverage known languages?

3. sibilant, nasal, liquid, etc?
Bible Data for 503 alphabetic languages

1. Cons vs Vowel
2. Nasal vs Cons vs Vowel
HMM Results

- 503 alphabetic languages
- 10 EM restarts
- Oracle tag mappings:

| C / V | N / C / V |
|-------|-----------|
| 93.4% | 74.6%     |

- [C / V] dominant natural pattern
- [N / C / V] less so...
Model Design Goals

1. Cluster letters:
   - HMM does soft clustering
   - letters typically in single phonetic class

2. Learn from hundreds of languages
   - Avoid oracle tag mappings

3. Model typologically coherent groupings
Bayesian HMM: Dirichlet Priors

\[ \alpha_1, \alpha_2, \alpha_3 > 0 \]
Hard Bayesian HMM

- One tag per character type
  (stochastic deterministic FSA)

- Dimensionality of tag emissions unknown:

  \[ \phi | N \sim \text{Dirichlet}(\beta_1 \ldots \beta_N) \]

  \[ N | \lambda \sim \text{Poisson}(\lambda) \]
Model 1: SYMM

Assume fixed, symmetric priors

transition dist

\[ \theta \]

\[ V \]

\[ t_1 \rightarrow t_2 \rightarrow \ldots \rightarrow t_n \]

\[ w_1 \rightarrow w_2 \rightarrow \ldots \rightarrow w_n \]

\[ \phi \]

\[ N \]

\[ \alpha \]

\[ \beta \]

\[ \lambda \]

Dirichlet priors

Poisson prior

Emission dist

\# symbols
Inference: Gibbs Sampling

- Integrate out parameters
- Iteratively sample tag assignment $t_w$ of symbol types $W$:

$$f(t_w|t_{-w}, w_1 \ldots w_n) \propto$$
Inference: Gibbs Sampling

- Integrate out parameters
- Iteratively sample tag assignment $t_w$ of symbol types $W$:

$$f(t_w|t_{-w}, w_1 \ldots w_n) \propto$$

$$\int f(N|\lambda) f(\lambda) \, d\lambda$$  \hspace{1cm} \text{(type counts)} \hspace{1cm} (1)$$

$$\int f(t_1 \ldots t_n|\theta) f(\theta) \, d\theta$$  \hspace{1cm} \text{(tags)} \hspace{1cm} (2)$$

$$\int f(w_1 \ldots w_n|t_1 \ldots t_n, \phi) f(\phi|N) \, d\phi$$  \hspace{1cm} \text{(symbols)} \hspace{1cm} (3)$$
Tag Predictive Distribution

$$\Pi_{t,t'} \left( \alpha_{t,t'} + n(t, t') \right)^{[\delta(t,t')]$$

Dirichlet hyperparameter

Tag bigram counts excluding symbol $w$

Tag bigram counts only with symbol $w$
Next Idea

- Trying to decipher an unknown language
- Learn from our 502 other languages where the tags are known

⇒ Assume shared, unknown priors across languages
transition dist

dirichlet priors

emission dist

Poisson prior

# symbols
Model 2: Infer universal language priors

MERGED

transition dist

emission dist

# symbols

Dirichlet priors

Poisson prior

Model 2: Infer universal language priors

MERGED

transition dist

emission dist

# symbols

Dirichlet priors

Poisson prior
Inference: Gibbs Sampling

- Slice sample hyperparameters

  \( f(\alpha_{t,t'} | \ldots) \propto \prod_{k} \frac{(\alpha_{t,t'} + k)^\#(k,t,t')}{\ldots \ldots} \)

  \( \sim \) product of predictive distributions with common prior

  \( \sim \) combine terms across languages for efficiency

  number of languages which have \( \geq k \) occurrences of bigram \((t, t')\)
MAP Transition Priors: 503 languages

context = START...

context = START, CONS...

context = START, VOWEL...

context = VOWEL, CONS...
MAP Transition Priors: 503 languages

case = CONS, CONS...
case = CONS, VOWEL...

context = CONS, CONS...
context = CONS, VOWEL...
Final Idea

- Languages vary typologically

  ⇒ Assume latent clustering of languages

  ⇒ Hyperparameters shared at the cluster level
Model 3: CLUST

Cluster assignment

Transition dist

Emission dist

# symbols

K clusters

Dirichlet priors

Poisson prior

Model 3

CLUST

$L$

$\pi$

$Z^L$

$\theta^L$

$V$

$t_1$

t_2

$\ldots$

$t_n$

$w_1$

$w_2$

$\ldots$

$w_n$

$\phi^L$

$N^L$

$K$

$\alpha_k$

$\beta_k$

$\lambda_k$
Evaluation

- Token-level accuracy
  - 503 languages
  - leave-one-out average

- Baseline:
  - HMM (EM) [Knight 2006]

- Hard Bayesian HMM
  - SYMM: symmetric fixed priors
  - MERGE: uses 502 observed languages
  - CLUST: 20 latent language clusters

Two Tasks

Cons / Vowel
Nasal / Cons / Vowel

← oracle mappings

← oracle mappings
Results

- 503 alphabetic languages
Results

- 30 language isolates

### Cons vs Vowel

- EM
- SYMM
- MERGED
- CLUST

### Nasal vs Cons vs Vowel

- EM
- SYMM
- MERGED
- CLUST
Confusion matrix

CLUST

true

predicted

C V N
C  
V  
N  

predicted

C V N
C  
V  
N  

baseline
## Confusion matrix

|       | C    | V    | N    |
|-------|------|------|------|
| **C** | ![C] | ![V] | ![N] |
| **V** | ![C] | ![V] | ![N] |
| **N** | ![C] | ![V] | ![N] |

**CLUST**

**Baseline**
Cluster Analysis

| Plurality Family | Proportion (%) | Plurality Family | Proportion (%) |
|------------------|----------------|------------------|----------------|
| Indo-European    | 38             |                  |                |
|                  | 24             |                  |                |
|                  | 21             |                  |                |
| Quechuan         | 89             |                  |                |
| Mayan            | 64             |                  |                |
| Oto-Manguean     | 55             |                  |                |
| Maipurean        | 25             |                  |                |
| Tucanoan         | 20             |                  |                |
| Uto-Aztecan      | 40             |                  |                |
| Altaic           | 44             |                  |                |
|                  |                | Austronesian     | 91             |
|                  |                |                  | 71             |
|                  |                |                  | 24             |
| Niger-Congo      | 100            |                  |                |
|                  | 68             |                  |                |
|                  | 67             |                  |                |
|                  | 50             |                  |                |
|                  | 24             |                  |                |
Future Work

- Finer-grained phonetic categories
- Syllable writing systems
  - Syllables as hidden variables in our data
- Apply to Linear A

Thank you!