Inferring phylogenetic graphs of Natural Languages using Minimum Message Length

Jane N. Ooi and David L. Dowe,
Monash University, Australia,
www.csse.monash.edu.au/~dld
Table of Contents

- Motivation and Background
- What is a phylogenetic model?
- Phylogenetic Trees and Graphs
- Types of evolution of languages
- Minimum Message Length (MML)
  - Multistate distribution – modelling of mutations
- Results/Discussion
- Conclusion and future work
Motivation

To study how languages have evolved (Phylogeny of languages).

- e.g. Artificial languages,
- European languages.

To refine natural language compression method.
Evolution of languages

- What is phylogeny?
  - Phylogeny means Evolution

- What is a phylogenetic model?
  - A phylogenetic tree/graph is a tree/graph showing the evolutionary interrelationships among various species or other entities that are believed to have a common ancestor.
Difference between a phylogenetic tree and a phylogenetic graph

**Phylogenetic trees**
- Each child node has exactly one parent node.

```
  X
 / \
/    \
Y    Z
```

**Phylogenetic graphs (new concept)**
- Each child node can descend from one or more parent node(s).

```
  X
 / \
/    \
Y    Z
```
Evolution of languages

3 types of evolution

- Evolution of phonology/pronunciation

<table>
<thead>
<tr>
<th>Words</th>
<th>US</th>
<th>UK</th>
</tr>
</thead>
<tbody>
<tr>
<td>schedule</td>
<td>skedule</td>
<td>shedule</td>
</tr>
<tr>
<td>leisure</td>
<td>leezhure</td>
<td>lezhure</td>
</tr>
</tbody>
</table>

- Evolution of written script/spelling

<table>
<thead>
<tr>
<th>English</th>
<th>Malay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mobile</td>
<td>Mobil</td>
</tr>
<tr>
<td>Television</td>
<td>Televisyen</td>
</tr>
</tbody>
</table>

- Evolution of grammatical structures
Minimum Message Length (MML)

What is MML?
- A measure of **goodness of classification** based on information theory (Wallace and Boulton, 1968; Wallace and Dow, 1999a; Wallace, 2005).

Data can be described using “models”

MML methods favour the “best” description of data where
- “best” = **shortest** overall two-part message length

Two part message
- \(\text{Msg length} = \text{Msg length (model)} + \text{msg length (data | model)}\)
Degree of similarity between languages can be measured by compressing them in terms of one another.

Example:

- Language A Language B
  - 3 possibilities —
    - Unrelated — shortest message length when compressed separately.
    - A descended from B — shortest message length when B compressed and then A compressed in terms of B.
    - B descended from A — shortest message length when A compressed and then B compressed in terms of A.
Minimum Message Length (MML)

The best phylogenetic model is the tree/graph that achieves the shortest overall two-part message length.
Modelling mutation between words

Root language

- Equal frequencies for all characters.
  - \( \log(\text{size of alphabet}) \times \text{no. of chars} \).
- Some characters occur more frequently than others.
  - e.g.: English “x” compared with “a”.
  - Multi-state (multinomial) distribution of characters.
Modelling mutation between words

Child languages

- Muti-state distribution
  - 4 states.
    - Insert
    - Delete
    - Copy
    - Change

- Use string alignment techniques to find the best alignment between words.

- Dynamic Programming Algorithm to find alignment between strings.

- MML favors the alignment between words that produces the shortest overall message length.
Example:

recommend

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>r e c o m m e n d</td>
<td></td>
</tr>
</tbody>
</table>

---
Work to date

**Preliminary model**
- Only *copy* and *change* mutations
- Words of the same length
- artificial and some European languages.

**Expanded model**
- *Copy*, *change*, *insert* and *delete* mutations
- Words of different length
- artificial and some European languages.
Results – Preliminary model

- Artificial languages
  - A – random
  - B – 5% mutation from A
  - C – 5% mutation from B

Full stop “.” marks the end of string.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>asdfge.</td>
<td>assfge.</td>
<td>assfge.</td>
</tr>
<tr>
<td>2</td>
<td>zlsdrya.</td>
<td>zlsdrya.</td>
<td>zlsdrya.</td>
</tr>
<tr>
<td>3</td>
<td>wet.</td>
<td>wet.</td>
<td>wbt.</td>
</tr>
<tr>
<td>4</td>
<td>vsert.</td>
<td>vsegt.</td>
<td>vsagt.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Results – Preliminary model

Possible tree topologies for 3 languages:

- Null hypothesis: totally unrelated
- Partially related
- Fully related

Expected topology:

Expanded model only
Results – Preliminary model

Possible graph topologies for 3 languages:

Non-related parents

Related parents
Results – Preliminary model

Results:
- Best tree =

  B
   / \       /
  /   \     /   
 A     C

- \(P_{mut}(B,A) \sim 0.051648, P_{mut}(B,C) \sim 0.049451\)

- Overall Message Length = 2933.26 bits
  - Cost of topology = \(\log(5)\)
  - Cost of fixing root language (B) = \(\log(3)\)
  - Cost of root language = 2158.7186 bits
  - Branch 1
    - Cost of child language (Lang. A) binomial distribution = 392.069784 bits
  - Branch 2
    - Cost of child language (Lang. C) binomial distribution = 378.562159 bits
## Results – Preliminary model

- European Languages (with accents removed)
  - French
  - English
  - Spanish

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>French</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>baby.</td>
<td>bebe.</td>
<td>nene.</td>
</tr>
<tr>
<td>2</td>
<td>beach.</td>
<td>plage.</td>
<td>playa.</td>
</tr>
<tr>
<td>3</td>
<td>biscuits.</td>
<td>biscuits.</td>
<td>bizcocho.</td>
</tr>
<tr>
<td>4</td>
<td>cream.</td>
<td>creme.</td>
<td>crema.</td>
</tr>
<tr>
<td>...</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
<tr>
<td>30</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>
Results – Preliminary model

- $P(\text{from French}) \sim 0.834297$
- $P(\text{from Spanish not French}) \sim 0.090559$
- $P(\text{from neither}) \sim 0.075145$
- $P_{\text{mut}}(\text{French, Spanish}) \sim 0.245174$

Cost of “parent” language (French) = 1226.76 bits
Cost of language (Spanish) binomial distribution = 734.59 bits
Cost of child language (English) trinomial distribution = 537.70 bits
Total tree cost = $\log(5) + \log(3) + \log(2) + 1226.76 + 734.59 + 537.70$
= 2503.95 bits
Results – Expanded model

- 16 sets of 4 languages

- Different length vocabularies
  - A – randomly generated
  - B – mutated from A
  - C – mutated from A
  - D – mutated from B

- Mutation probabilities
  - Copy – 0.65
  - Change – 0.20
  - Insert – 0.05
  - Delete – 0.10
# Results – Expanded model

<table>
<thead>
<tr>
<th></th>
<th>Language A</th>
<th>Language B</th>
<th>Language C</th>
<th>Language D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>awjmv.</td>
<td>afjmv.</td>
<td>wqmv.</td>
<td>afjnv.</td>
</tr>
<tr>
<td>2</td>
<td>bauke.</td>
<td>baxke.</td>
<td>auke.</td>
<td>bave.</td>
</tr>
<tr>
<td>3</td>
<td>doinett.</td>
<td>domnit</td>
<td>deoinet.</td>
<td>domnit.</td>
</tr>
<tr>
<td>4</td>
<td>eni.</td>
<td>eol.</td>
<td>enc.</td>
<td>eol.</td>
</tr>
<tr>
<td>5</td>
<td>foijgnw.</td>
<td>fiogw.</td>
<td>foijn.</td>
<td>fidgw.</td>
</tr>
<tr>
<td>50</td>
<td>........</td>
<td>........</td>
<td>........</td>
<td>........</td>
</tr>
</tbody>
</table>

Examples of a set of 4 vocabularies used
Results – Expanded model

Possible tree structures for 4 languages:

Null hypothesis: totally unrelated

Partially related
Results – Expanded model

Expected topology

Fully related
Results – Expanded model

- Correct tree structure 100% of the time.
- Sample of inferred tree and cost:

```
  A
 / \
B   C
  \
   D
```

- Language A: size = 383 chars, cost = 1821.121913 bits
Results – Expanded model

- \( \Pr(\text{Delete}) = 0.076250 \)
- \( \Pr(\text{Insert}) = 0.038750 \)
- \( \Pr(\text{Mismatch}) = 0.186250 \)
- \( \Pr(\text{Match}) = 0.698750 \)
- 4 state Multinomial cost = 930.108894 bits

- \( \Pr(\text{Delete}) = 0.071250 \)
- \( \Pr(\text{Insert}) = 0.038750 \)
- \( \Pr(\text{Mismatch}) = 0.183750 \)
- \( \Pr(\text{Match}) = 0.706250 \)
- 4 state Multinomial cost = 916.979371 bits

*Note that all multinomial cost includes an extra cost of \( \log(26) \) to state the new character for mismatch and insert*
Results – Expanded model

- $\Pr(\text{Delete}) = 0.066580$
- $\Pr(\text{Insert}) = 0.035248$
- $\Pr(\text{Mismatch}) = 0.189295$
- $\Pr(\text{Match}) = 0.708877$
- 4 state Multinomial cost = 873.869382 bits

Cost of fixing topology = $\log(7) = 2.81$ bits

Total tree cost = $930.11 + 916.98 + 873.87 + 1821.11 + \log(7) + \log(4) + \log(3) + \log(2)$
= 4549.46 bits
## Results – Expanded model

### European Languages
- French
- English
- German

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>French</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>even.</td>
<td>meme.</td>
<td>sogar.</td>
</tr>
<tr>
<td>2</td>
<td>eyes.</td>
<td>oeil.</td>
<td>auge.</td>
</tr>
<tr>
<td>3</td>
<td>false.</td>
<td>faux.</td>
<td>falsch.</td>
</tr>
<tr>
<td>4</td>
<td>fear.</td>
<td>peur.</td>
<td>angst.</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>601</td>
<td>....</td>
<td>....</td>
<td>....</td>
</tr>
</tbody>
</table>
Results – Expanded model

Total cost of this tree = 56807.155 bits

Cost of fixing topology = \log(4) = 2 \text{ bits}

Cost of fixing root language (French) = \log(3) = 1.585 \text{ bits}

Cost of French = \text{no. of chars} \times \log(27) = 21054.64 \text{ bits}
Results — Expanded model

- Cost of fixing parent/child language (English) = \( \log(2) = 1 \) bit
- Cost of multistate distribution (French -> English) = 15567.98 bits
- MML inferred probabilities:
  - \( \Pr(\text{Delete}) = 0.164322 \)
  - \( \Pr(\text{Insert}) = 0.071429 \)
  - \( \Pr(\text{Mismatch}) = 0.357143 \)
  - \( \Pr(\text{Match}) = 0.407106 \)

- Cost of multistate distribution (English -> German) = 20179.95 bits
- MML inferred probabilities:
  - \( \Pr(\text{Delete}) = 0.069480 \)
  - \( \Pr(\text{Insert}) = 0.189866 \)
  - \( \Pr(\text{Mismatch}) = 0.442394 \)
  - \( \Pr(\text{Match}) = 0.298260 \)

- Note that an extra cost of \( \log(26) \) is needed for each mismatch and \( \log(27) \) for each insert to state the new character.
MML methods have managed to

- infer the **correct** phylogenetic tree/graphs for artificial languages.
- infer phylogenetic trees/graphs for languages by encoding them in terms of one another.

We can not (or can we?) conclude that one language really descends from another language. We can only conclude that they are related.
Future work:

- Compression – grammar and vocabulary.
- Compression – phonemes of languages.
- Endangered languages – Indigenous languages.
- Refine coding scheme.
  - Some characters occur more frequently than others.
    E.g.: English - “x” compared with “a”.
  - Some characters are more likely to mutate from one language to another language.
Questions?
Some further reading on MML
