CPSC 503 – Final project presentation: Inferring history of Austronesian languages using language models

Yongliang (Vincent) Zhai

April 15, 2013

Abstract:

- In this project, we use language models, including edit distances and N-gram methods, to measure the dissimilarity of languages, in order to reconstruct a phylogenetic tree for the history of languages.
- We propose a new dissimilarity measure for the sound of words describing the same content in different languages based on N-grams.
- We show that this dissimilarity measure performs best in identifying the correct history of languages in our data analysis compared with other dissimilarity measures derived from edit distances.
- This work can be applied to larger number of languages to assist human annotators to classify word cognates and languages.

Method: Edit Distances for two words A and B

- Minimum Edit Distance:
  insertion (1), deletion (1), and replacement (2).

$$d_{MED} < l_A + l_B.$$

- Normalized Minimum Edit Distance:

$$d_{NMED} = \frac{d_{MED}}{l_A + l_B}, \quad 0 \leq d_{NMED} \leq 1.$$

- Levenshtein Edit Distance:
  insertion (1), deletion (1), and replacement (1).

$$d_{LED} \leq \max\{l_A, l_B\}.$$

- Normalized Levenshtein Edit Distance:

$$d_{NLED} = \frac{d_{LED}}{\max\{l_A, l_B\}}, \quad 0 \leq d_{NMED} \leq 1.$$

Method: N-grams model for two words A and B

- grams(A): all N-grams of characters in A.
  For example, $A = add$.

$$grams(A) = \{\{a\}, \{d\}, \{a, d\}, \{d, d\}, \{a, d, d\}\}$$

- grams(A, B): common grams of $grams(A)$ and $grams(B)$.

$$grams(A, B) = \{x | x \in grams(A) \text{ and } x \in grams(B)\}.$$

- d_{grams}: dissimilarity of A and B.

$$d_{grams} = 1 - \frac{2 \times grams(A, B)}{grams(A) + grams(B)}.$$

  Therefore,

$$0 \leq d_{grams} \leq 1.$$

Dissimilarity Matrix for Languages

- The dissimilarity of two languages L_1 and L_2 is defined as the average of dissimilarities of N words, i.e.,

$$d_L(L_1, L_2) = \frac{\sum_{i=1}^{N} d(W_{i,1}, W_{i,2})}{N}.$$

- The dissimilarity of M languages $L_1, L_2, \ldots, L_M$ forms a dissimilarity matrix $\mathbf{D}$, which is $M \times M$, and the (i, j) element of $\mathbf{D}$ is

$$d_{ij} = d_L(L_i, L_j).$$

  The diagonal elements of $\mathbf{D}$ are set to 0.

Constructing a Tree Using **D**

- Neighbour-Joining algorithm.
- Input: dissimilarity matrix **D**.
- Output: an unrooted bifurcating tree T.
- Guaranteed to return the "correct tree" under some conditions.

Data: Summary

- 14 languages are chosen.
- 87 items are annotated for all 14 languages.
- Some items are annotated with more than one sounds.
- Choose the last one if more than one annotations exist.

Table: Three different annotations for the item "to walk" in Fagani language.

ID	*Item*	*annotation*	*notes*	*cognacy*
137031	to walk	pwapwahe	walk	81
137032	to walk	sio	go down	38
137034	to walk	akau	go up	1

Data: Relations of the 14 Languages.

Table: Classification of 14 languages chosen.

Language	*Classification*
Rukai Tona	A:F:Rukai
Rukai Budai	A:F:Rukai
Ivasay	A:M:P:B:Ivatan
Isamorong	A:M:P:B:Ivatan
Babuyan	A:M:P:B:Ivatan
Muna	A:M:C:E:S:M:N:M:M:Western
Wuna	A:M:C:E:S:M:N:M:M:Western
Bonerate	A:M:C:E:S:M:Tukangbesi-Bonerate
Popalia	A:M:C:E:S:M:Tukangbesi-Bonerate
Mouk	A:M:C:E:O:W:N:N:V:S:Bibling
Aria	A:M:C:E:O:W:N:N:V:S:Bibling
Megiar	A:M:C:E:O:W:N:N:V:B:N:Northern
Matukar	A:M:C:E:O:W:N:N:V:B:N:Northern
Fagani	A:M:C:E:O:C:S:M:San Cristobal

Results:

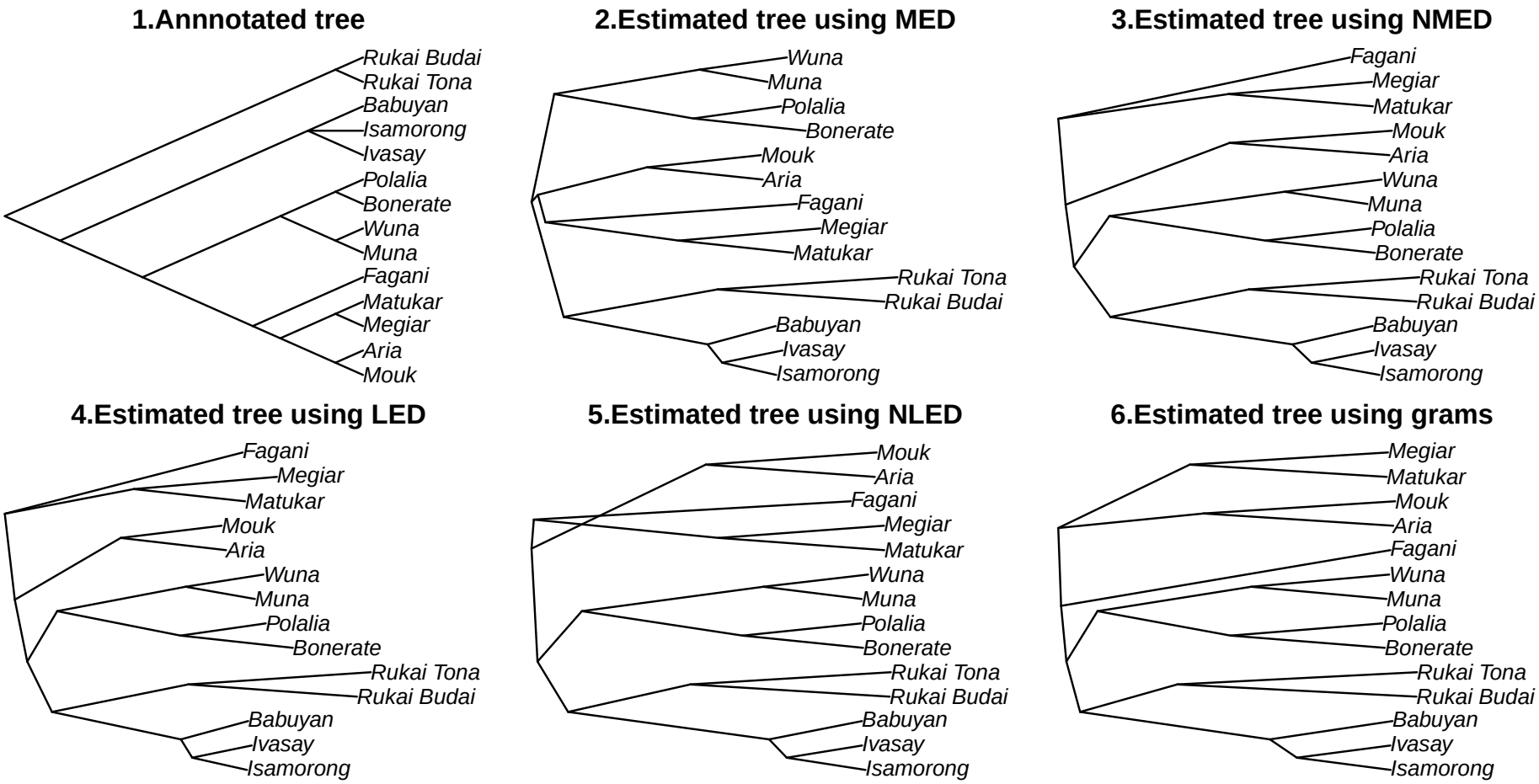

Figure: The annotated tree and estimated trees using different language models.

Conclusions and Limitations

Conclusions:

- Edit distances can be used to infer history of languages, although not very accurate at some details.
- Our proposed N-grams model performs best in this analysis (but it may be too early to claim this is true in general).

Limitations and Future Work:

- Create a larger experiment to check the accuracy of the results automatically.
- Repeat the analysis on other languages
- Repeat the analysis with more languages.

Thank you!