Clustering Affixes:

Applying ML Techniques to Morphological Analysis

Olga Zamaraeva

University of Washington

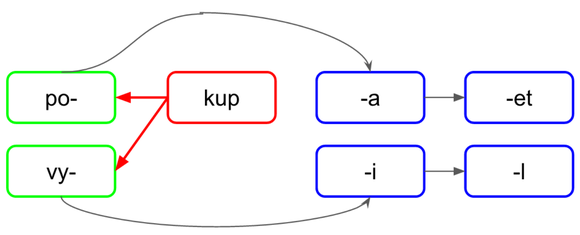
olzama@uw.edu

1. Morphological Analysis

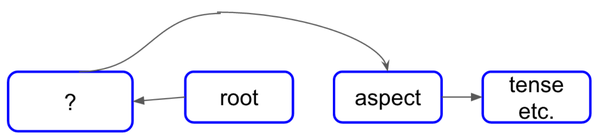
- Group morphemes into Position Classes
- Infer linguistic features associated with the affixes
- Include morphological rules in the grammar

1) Ivan po-kup-a-et sobak-u
   Ivan.NOM PREF1-buy-IMPF-3SG.PRES dog-ACC [rus]
   Ivan is buying a dog.
2) Ivan vy-kup-i-l sobak-u
   Ivan.NOM PREF2-buy-PFV-3SG.M.PST dog.ACC [rus]
   Ivan bought the dog out.

Position classes and Interlinear Glossed Text



Affix graph



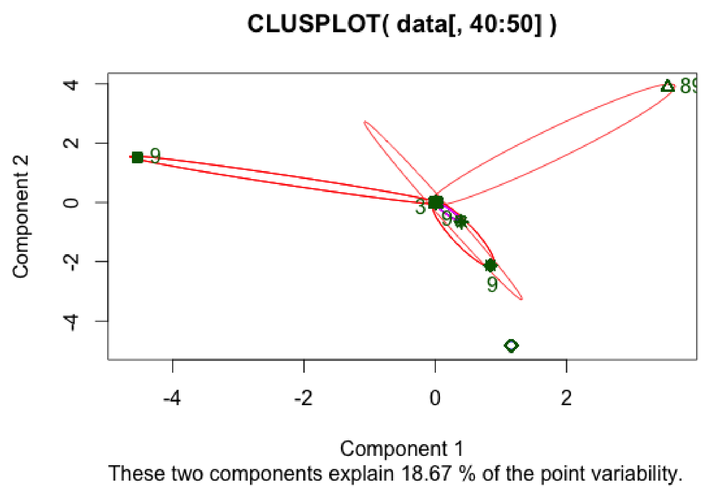
Position classes graph

2. Dataset

- Chintang (ISO-639: ctn) IGT collection, shared by Bickel et al. (2013)
- 9K IGT
- Gold standard exists for morphological rules

3. Clustering affixes

- Use simple classification technique instead of relying on and updating a big graph
  k-means
- Store input information and build the result graph once



These two components explain 18.67 % of the point variability.

2D plot using a subset of features

4. Feature selection

- Orthography
- Gloss
- Right and left context (1 position)
- Linguistic features of the affix
- Prefix or suffix

5. Evaluation

- Precision: how many mistakes in each cluster
  - Pick a cluster (all the assigned labels are the same)
  - Determine which label is the majority label in the corresponding gold labels portion
  - Count how many labels are not the majority label in the corr. gold labels portion
    "false positives"
- Recall: how many things that should be in one cluster are in different clusters
  - Pick a true cluster (all true labels are the same)
  - Determine which label is the majority label in the corresponding assigned labels portion
  - Count how many labels are not the majority label in the corr. assigned labels portion
    "false negatives"
- Excluded from evaluation:
  - Unknown affixes
    not found in the gold file
    constitute a big portion of the data
- Gold labels assigned automatically to datapoints
  mostly based on gloss / orthography

6. Results

Baseline: Graph-based system Wax (2014)

k	24	26
Precision Baseline	75.0	---
Precision k-means	**79.0**	81.0
Recall Baseline	**80.8**	---
Recall k-means	70.6	68.7

(average over 100 k-means runs)

F1 score:

Baseline: 77.4

k-means: 74.2

k-means 26: 73.9

7. Error Analysis

- Recall may be less important.
- There may be errors in the gold standard.
- Need better way to assign gold labels to observations automatically.
- Should use evaluation techniques appropriate for clustering.
  V-measure
- Better feature selection always possible.
- Sparse vectors.
- Recall problems could be solved with the users help
  Active learning

8. Future Work

- Agglomerative clustering and Active learning
  Field linguists have expressed interest in such a system
- Other algorithms may work better on sparse, binary vectors
- Dimensionality reduction
- Evaluation methods for unknown affixes
  Evaluation by parsing

9. Acknowledgments

Thanks to Gina Levow and Rik Koncel-Kedziorsky for suggestions on evaluation. Thanks to Emily M. Bender for the poster LaTeX source.

My work on this project is partially supported by Microsoft Graduate Women Fellowship.

References

Bickel, B., Gaenszle, M., Rai, N. K., Rai, V. S., Lieven, E., Stoll, S., ... Rai, I. P. (2013). *Tale of a poor guy.* (Accessed online on 15-January-2013)

Wax, D. (2014). *Automated grammar engineering for verbal morphology*. Unpublished master's thesis, University of Washington.