Automatic Detection of Morphological Processes in the Yorùbá Language

Tunde Adegbola

African Languages Technology Initiative (Alt-i)

Unsupervised Morphology Induction

- Unsupervised morphology induction of primarily concatenative morphology is quite well understood
- Methods generally based on the works of Harris (1955); Déjean (1998); Goldsmith (2000); Creutz and Lagus (2002); Creutz (2003); Creutz and Lagus (2004); Monson et al. (2007); Hammarström (2009) and a host of others
- But many productive morphological processes depend on **stem-derived** morphemes that do not recure due to derivation from stems
- 7000 languages in the world and probably less than 100 subjected to NLP at a significant level
- Algorithms for totally automated, unsupervised morphology induction is of prime importance in NLP, particularly for low resourced languages

Stem-derived Morphemes

- An affix that depends on and therefore reflects the stem
- For example, partial reduplication in Yorùbá: $CV \rightarrow CíCV$

Stem	Gloss	Derived Word	Gloss	
Lọ	Go	Lílọ	Going (N)	
Wá	Come	Wíwá	Coming (N)	
Ṣe	Do	Ṣíṣe	Doing (N)	
Kọ	Write	Kíkọ	Writing (N)	
Ké	Cry	Kíké	Crying (N)	
Sè	Cook	Sísè	Cooking (N)	

- Other morphological process employ stem-derived (templatised) morphemes

Recurrent Patterns and Word-labels

- **Stem-derived** morphemes **do not reoccur** but manifest recurrent patterns
- **Word-labels** serve as a textual proxy of the recurrent patterns
- Words formed by common morphological process tend to produce the same word-label
- A Word-label naturally clusters words of common morphological processes, leading to automatic and unsupervised induction of morphology
- Patterns in word-labels indicate the morphological process that motivated the formation of the words in their clusters
- Significant substrings of word-labels suggest morphemic boundaries

The Word Label

A sequence of symbols $A_i X_i,\ \text{where}\ A_i \in \{C,\ V\},\ and\ X_i \in \{0, 1, 2, \ldots n\},$ in which each symbol represents a consonant or vowel uniquely

Word	Word Label
Deal	C0V0V1V1
Said	C0V0V1C1
Deed	C0V0V0C0
Seek	C0V0V0C1
Razzle dazzle	C0V0C1C1C2V1 C3V0C1C1C2V1
Dilly dally	C0V0C1C1C2 C0V1C1C1C2
Willy nilly	C0V0C1C1C2 C3V0C1C1C2

Metrization of word-labels orders them based on morphological processes that produced the words they cluster

Predicted Vs Observed Probabilities of Word-labels

- The predicted probability of a word-label assumes equiprobability and independence of its symbols
- The observed probability of a word-label manifests morphological influences as well as resource scarcity effects
- A significant difference in the predicted and observed probabilities of a word-label is indicative of stem-derived morphology
- The nature of the differences between the predicted and observed probabilities is also indicative of the level of coverage of the corpus from which the observed probability was obtained

Computing the Predicted Probability

- A word-label is a sequence of symbols $A_i X_i$, where:
  $A_i \in \{C, V\}$ *and* $X_i \in \{0, 1, 2, \ldots . n\}$
- For example,
  - C0V0C1V1 is made up of symbols $A_1 X_1 A_2 X_2 A_3 X_3 A_4 X_4$, where
    $A_1 = C,\ X_1 = 0,\ A_2 = V,\ X_2 = 0,\ A_3 = C,\ X_3 = 1,\ A_4 = V\ and\ X_4 = 1.$
- Given c consonants and v vowels, the probability of the first symbol being C0 or V0 is $\frac{c}{c}$ or $\frac{v}{v}$. respectively
- The probability of any symbol *CX* or *VX* is $\frac{(c - X)}{c}$ or $\frac{(v - X)}{v}$ respectively
- The probability of incidence of an already used symbol is $\frac{1}{c}$ or $\frac{1}{v}$

From Likelihood to Probability

- The probability of each symbol in a word-label assumes equiprobability of each and independence between any two incident consonant or vowel
- The likelihood of a word-label is the product of the probabilities of all symbols in it as shown in eqn. (1) below
- $$L\left(A_1X_1A_2X_2\ldots\ A_nX_n\right)=\prod_{i=1}^{n}P(A_iX_i) \qquad (1)$$
- n is the number of symbols in a word-label
- $$S=\sum_{j=1}^{m}L_j\left(A_1X_1A_2X_2\ldots\ A_nX_n\right) \qquad (2)$$
- m is the number of word-labels in a group of word-labels of same length and identical CV structure
- $$P\left(A_1X_1A_2X_2\ldots\ A_nX_n\right)=\frac{1}{S}\prod_{i=1}^{n}P(A_iX_i) \qquad (3)$$
- Each likelihood is normalised with S, the cumulative likelihoods of all word-labels in a group to turn them into probabilities cumulating to unity

Computing the Observed Probabilities

- Given a group of word-labels of identical length and common sequence of consonants and vowels, clustering a total of n word-tokens around all word-labels in the group
- The observed probability $P(i)$ of a word-label i with a cluster of n_i word-tokens is given by: $P(i) = \frac{n_i}{n}$

Test

- A lexicon of 14,670 word-tokens was extracted from a Yorùbá corpus
- It produces 1,282 distinct word-labels
- Word-labels were grouped according to their lengths and CV structures (sequence of consonants and vowels)
- Predicted and observed probabilities of the word-tokens were computed based on 18 consonants and 12 vowels
- Comparison between the predicted and observed probabilities of word-labels was undertaken to determine the incidence of stem-derived morphemes

Results

- The word-label with the highest disparity between the predicted and observed probabilities is C0V0C1V0C0V0C1V0
- It featured predicted and observed probabilities of 1.04697E-06 and 0.133116883 respectively, yielding a ratio of 127145.48
- The symmetry in the word-label by virtue of the duplication of the substring C0V0C1V0 suggest the morphological process of full reduplication and C0V0C1V0 as the morphemic boundary
- Samples words clustered around th word-label include *biribiri, bọ̀lọ̀bọ̀lọ̀, fírífírí,* and *gbẹ̀jẹ̀gbẹ̀jẹ̀,* all of them being words formed through the morphological process of full reduplication

Results (cont.)

Word-Label	Ratio	Morphological Process	Sample words
C0V0C1V0C0V0C1V0	127145.48	Full Reduplication	biribiri, bọ̀lọ̀bọ̀lọ̀ , fírífírí, gbẹ̀jẹ̀gbẹ̀jẹ
C0V0C1V1C0V0C1V1	19452.41	Full Reduplication	bojúbojú, bàmùbàmù, fọ́rífọ́rí, jayéjayé
C0V0V0C1V0C0V0	6174.55	Partial Reduplication	fẹ́ẹ́rẹ́fẹ́ , gbuurugbu, tààràtà, pẹẹrẹpẹ
C0V0C0V0C0V0	1074.26	Full Reduplication	dandandan, gangangan, jẹ́ jẹ́ jẹ́ , tantantan
C0V0C1V1C0V2C1V1	352.40	Full Reduplication	fálafàla, jágbajàgba, kóbokòbo, pálapàla
V0C0V0C1V1C0V0	20.58	Interfixation	àgbàlágbà, ọmọkọ́mọ, ọ̀pọ̀ lọpọ
V0C0V1C1V2C2V2	1.88	Prefixation	alágídí, alákàrà, ọlọ́gẹ̀dẹ̀ , ónígbèsè
V0V1C0V2C1V3	1.30	Prefix+Compounding	àìdúpẹ́ , àìlera, àìmọ̀kan, àìrójú, àìgbọràn
C0V0C0V1C1V2	1.30	Partial Reduplication	dídọ́gba, jíjóná, kíkorò, lílépa, pípadà
C0V0C1V1C0V2	0.85	Compounding	jogóji, kàgbákò, láyọ̀ lé, pawọ́pọ̀ , ṣojúṣe
C0V0C1V1C2V1C3V2	0.53	Desentencialisation	kòbọmọjẹ́ , mójúkúrò, yírapadà, ṣàfarawé

Conclusions

- The predicted and observed probabilities ratio of word-labels is a valuable metric for the identification of word-labels that incorporate stem-derived morphemes
- Word-labels with stem-derived morphemes but low cardinality are:
  - indicative of the paucity of the employed corpus
  - can be used to project and thereby validate or even generate out-of-vocabulary words in relevant NLP situations

Future Studies

- Application of word-labels in
  - the analysis of the morphology of other similar languages
  - modelling the coverage of corpora
  - projecting, validating or generating out-of-vocabulary words in ASR and other NLP endeavours
- Development of an Automatic Morphological Analyzer that addresses both recurrent partials and stem-derived morphemes to supplement the results offered by Linguistica, Morfessor, Paramor and others
- Possible extension of word-labels to address other radically different non-concatenative morphologies such as Root-and-Pattern morphology of semitic languages

Thank You