Improving automatically induced lexicons for highly agglutinating languages using data-driven morphological segmentation

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Wiehan Agenbag · Thomas Niesler

Department of Electrical and Electronic Engineering, University of Stellenbosch, South Africa
wagenbag@sun.ac.za · trn@sun.ac.za

Motivation

Automatic lexicon induction enables ASR for under-resourced languages by requiring only recorded speech and orthographic transcriptions.

Agglutination is a challenge

Agglutinating languages consist of large vocabularies of long words that occur infrequently, which makes automatic lexicon induction particularly difficult.

Proposed solution

Data-driven morphological segmentation to create shorter, more frequently occurring types, prior to lexicon induction. Then recombine pronunciations of morphs.

The challenge: Luganda

Table 1. Vocabulary distributions for three under-resourced languages.

Dataset	Hours	Words	$n > 3$		$n > 9$	
			Types	Tokens	Types	Tokens
Luganda	9.59	**18305**	14.4%	75.0%	5.6%	63.8%
Acholi	9.19	8719	27.9%	91.9%	13.3%	85.3%
Ugandan English	5.75	6737	26.9%	88.2%	11.1%	78.3%

- Luganda is a highly agglutinating, under-resourced language spoken in Uganda.
- Large number of words compared to more isolating languages: $>$ 2x compared to Acholi for the same corpus size. Many tokens (25%) are seen 3 times or fewer in the corpus.
- This causes a large performance deficit for automatically induced lexicons compared with hand-designed lexicons.

Solution: morphological segmentation

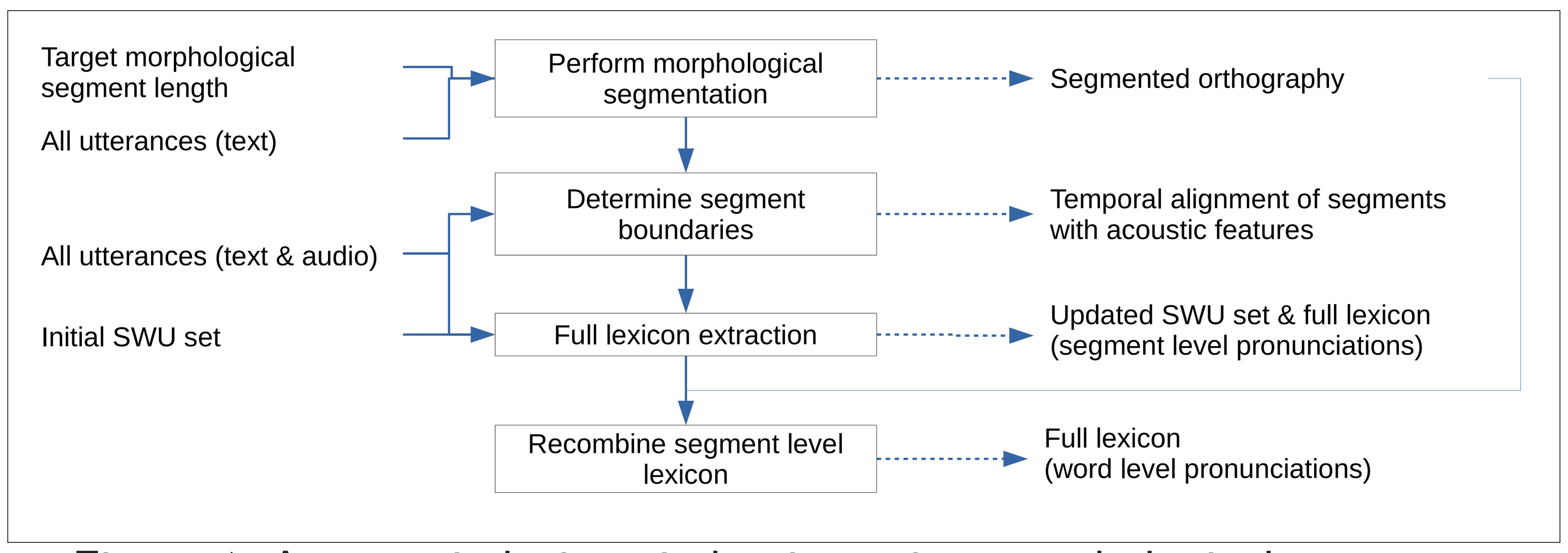

Figure 1. Automatic lexicon induction using morphological segments.

- Break down longer agglomerations into shorter morphs: e.g. *EKIGAMBO* $\rightarrow$ *EKI* + *GAM* + *BO*.
- In the absence of expert knowledge we used data-driven morphological segmentation (Morfessor 2.0), which yields a more favorable distribution of types.
- Automatic lexicon induction is then performed using the segmented orthography. Word-level pronunciations are obtained by concatenating segment pronunciations.

Results

Table 2. Vocabulary distributions for various degrees of segmentation, and the ASR performance of the associated lexicons.

Lexicon	Average segment length	# Types	$n > 3$		$n > 9$		% WER
			Types	Tokens	Types	Tokens	
Phoneme							54.94%
Grapheme							55.14%
Auto (unsegmented)	5.7	18305	14.4%	75.0%	5.6%	63.8%	60.91%
Auto + segmentation	3.0	2135	70.12%	99.31%	49.51%	97.72%	56.35%
	2.5	836	93.42%	99.94%	82.30%	99.63%	56.35%
	2.0	230	99.57%	100.00%	99.13%	100.00%	54.95%
	1.5	70	98.57%	100.00%	97.14%	100.00%	**54.90**%
	1.0	33	96.97%	100.00%	93.94%	100.00%	55.14%

Adding context

- Shorter segments appear to yield better lexicons, but are also associated with a rapidly diminishing number of types.
- This may not reflect all the acoustically distinct types useful for ASR.
- Solution: increase number of types by adding context-dependence to segments prior to inducing pronunciations; e.g. EKIGAMBO $\rightarrow$ *|EKI|GAM + EKI|GAM|BO + GAM|BO|*
- The use of a threshold ensures that only contexts with an adequate occurrence count are considered. For the rest, context is discarded.

Table 3. The ASR performance of lexicons induced using context-dependant morphological segmentation for various pooling thresholds. A threshold of ∞ indicates context independent segments.

Average segment length	Threshold	# Types	% WER
2	∞	230	54.95%
	250	434	**54.33**%
	125	720	54.95%
	62	1358	55.05%
1.5	∞	70	54.90%
	250	462	55.53%
	125	856	**54.81**%
	62	1573	56.83%
1	∞	33	55.15%
	250	655	54.52%
	125	992	**54.28**%
	62	1347	55.87%

Summary and conclusion

Table 4. Best ASR performance for various lexicons.

System	% WER
Phoneme	54.94%
Grapheme	55.14%
Auto	60.91%
Auto + segmentation	54.90%
Auto + segmentation + context	54.28%

Automatically induced pronunciation lexicon that exceeds an expert baseline even for a highly agglutinating language (Luganda).