A Systematic Approach to Derive a Refined Speech Corpus for Sinhala

Disura Warusawithana, Nilmani Kulaweera, Lakshan Weerasinghe, Buddhika Karunarathne

University of Moratuwa, Sri Lanka

PROBLEM

Limited availability of publicly available *resources* for Sinhala ASR

Room for *improvement of quality* in existing resources

MOTIVATION

Providing a resource of good quality

- to *train* Sinhala ASR models for high accuracy
- to *evaluate* existing Sinhala ASR models

DATASET

OpenSLR-52 Speech Corpus

Publicly available, crowdsourced corpus for Sinhala Language

Corpus Statistics

185,293 utterances
~224 hours of speech data
478 speakers

Issues in the Corpus

Unavailable Metadata

No information about the genders of speakers

Issues Related to Textual Characters

Punctuation marks සිංහල උපසිරසියට! ජය වේවා!.
English utterances winners of miss world
Numeric characters 2004 දෙසැම්බර් 26 වැනි දා
Unnecessarily applied non-printable characters (e.g. ZWJ)

Issues Related to Linguistics

Transcriptions contain obviously misspelled words

පුලුවන් ⟶ පුළුවන් (possible)

Transcriptions contain words having contextually incorrect spellings

(of)
එකේ සුන්දරත්වය සංචාරක සටහනකින් විස්තර කරන්න බැහැ
✖ ✔
ඒකේ (of that)

Transcriptions have obviously incorrect omission of spaces

ඔබවටා ⟶ ඔබ වටා (around you)

Transcriptions have obviously incorrect inclusion of spaces

අපිව ත් ⟶ අපිවත් (also us)

Transcriptions have contextually incorrect omission of spaces

(time)
චිත්‍රපටියක් බලන්නට යාමය
✖ ✔
යාම ය (is going)

METHODOLOGY

Completing Required Metadata

Tagged the gender of each speaker by listening to utterances

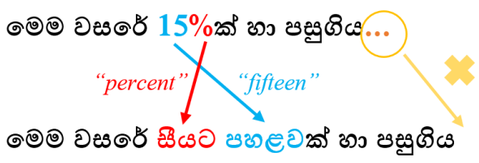
b1a64	f
d6ccd	m
0a2fe	m
6054f	f
da449	f

Treating Character-wise Errors

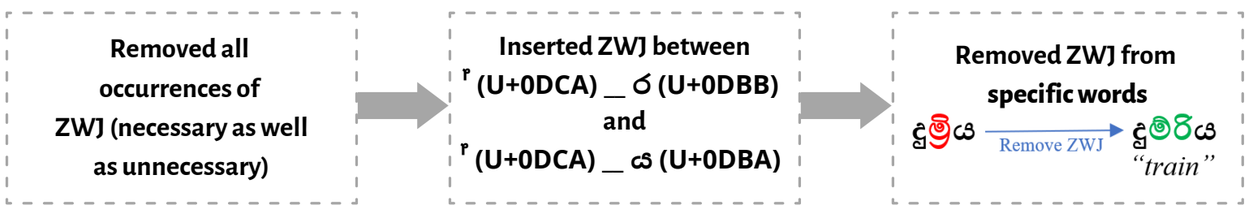
Removed Punctuation marks (inserted "සියට" for "%")

Excluded transcriptions of English utterances

Replaced numbers with their textual format



Removed unnecessary non-printable characters



Applying Linguistic Corrections

Prepared a list of 61 rules for grammatically-correct spacing

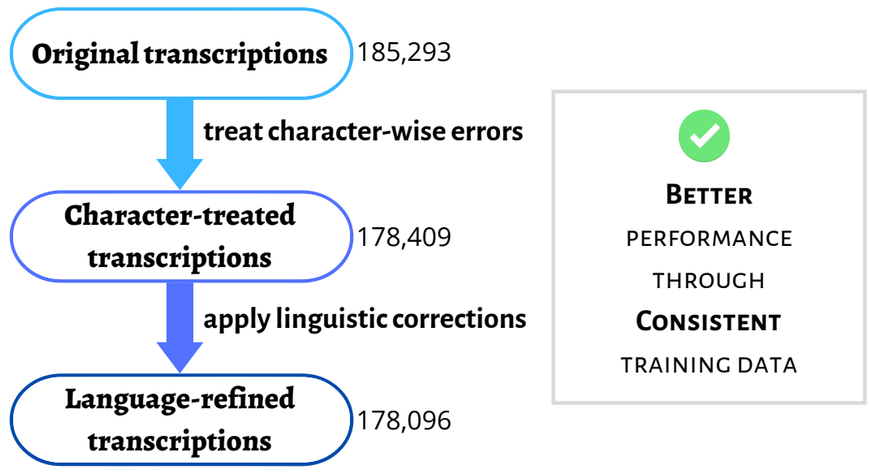
Independent of the order of applying them on text

Defined corrections as 'dictionaries' of key-value pairs

incorrect,	correct
ගියේය,	ගියේ ය
කාල වලට,	කාලවලට

Developed scripts to find all occurrences of keys and replace with values

pandas



EVALUATION

Experiments

Experiment 1 : Using Character-treated transcriptions

Experiment 2 : Using Language-refined transcriptions

Train-Test split: 80%-20%
GMM-HMM based ASR model

KALDI

Experimental Results (WER)

Training Pass	Experiment 1	Experiment 2
Monophone	64.17	**59.25**
Triphone pass 1	49.20	**42.72**
Triphone pass 2	47.24	**40.67**
Triphone pass 3	43.21	**36.34**

Comparison of Decoded Texts

Original transcription	Decoded text	
	Experiment 1	Experiment 2
Effect of proper removal of spaces		
මේ තොරතුරුවලට (For these details)	මේ තොරතුරු වලට	මේ තොරතුරුවලට
Effect of proper inclusion of spaces		
නිදසුනක් වශයෙන් දැක්විය හැකි ය (Can give as examples)	නිදසුනක් වශයෙන් දැක්විය හැකිය	නිදසුනක් වශයෙන් දැක්විය හැකි ය
Effect of spelling corrections		
එකක් නමයි යම් දියුණුවක් ලබමින් තිබුණ අපේ පොඩි පොඩි කර්මාන්තවලට (One thing is for our small-scale industries which were growing)	එකක් නමයි යම් දියුණුවක් ලබමින් තිබුන අපේ පොඩි පොඩි කර්මාන්තවලට	එකක් නමයි යම් දියුණුවක් ලබමින් තිබුණ අපේ පොඩි පොඩි කර්මාන්තවලට