Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban

Sarah Samson, Laurent Besacier, Benjamin Lecouteux, Mohamed Dyab

Using Resources from a Closely-related Language to Develop ASR for a Very Under-resourced Language: A Case Study for Iban

Sarah Samson Juan[1,2], Laurent Besacier[2], Benjamin Lecouteux[2], Mohamed Dyab[2]

[1]Faculty of Computer Science and Information Technology, UNIMAS, Malaysia
[2]Grenoble Informatics Laboratory (LIG), Univ. Grenoble Alpes, Grenoble, France
contact: laurent.besacier@imag.fr

Introduction

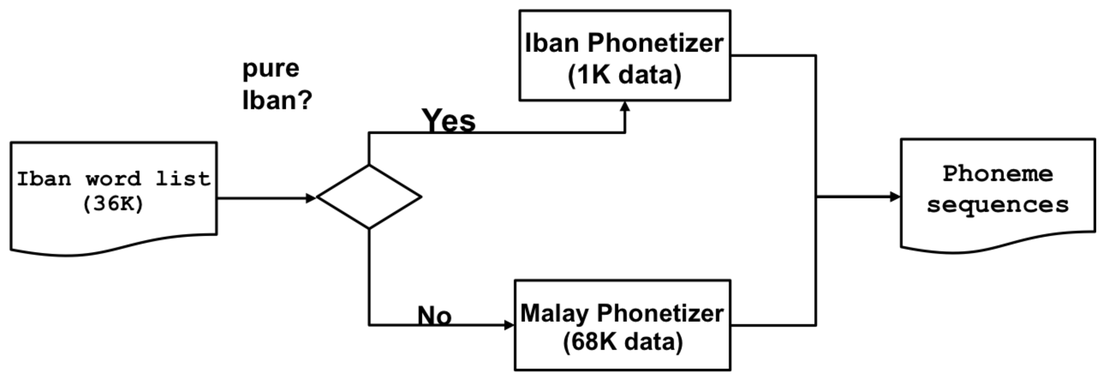
- **Context**
  - Exploit data from **closely-related language** for developing ASR for very under-resourced language
  - Case study on Iban, a language spoken in Sarawak, Borneo Island. The language is close to Malay, which is largely spoken in Malaysia.
- **Objectives**
  - Fast-bootstrapping approach for building Iban pronunciation dictionary
  - Improve performance of (low-resourced) Iban acoustic models
- **Problems**
  - Building an Iban pronunciation dictionary from scratch
  - Very limited training data for acoustic model training

Methodology

- **Semi-supervised approach for Iban pronunciation dictionary**
  - Semi-supervised lexicon design using Malay data
- **Cross-lingual acoustic modelling with limited training data**
  - Cross-lingual SGMMs - porting Universal Background Model (UBM) across languages
  - Cross-lingual DNNs - language-specific top layer for DNN

Using a closely related language (Malay) for Iban pronunciation dictionary

Hybrid G2P:



Pronunciation dictionary evaluation on Iban ASR

- **Acoustic model :**
  - Context dependent
  - HMM triphone state
  - 39 MFCCs
  - Train on 7hr news data
- **Language model :**
  - Trigram
  - 2M Iban words - news articles
- **ASR tool :** Kaldi

Pronunciation Dictionary	HMM/GMM ASR (WER %)	
	no spkr adapt	spkr adapt
Grapheme	32.9	20.5
English G2P	39.2	22.9
Malay G2P	35.7	19.8
Iban G2P	36.2	20.9
Hybrid G2P	36.0	**19.7**

Baseline ASR using Hybrid G2P pronunciation dictionary

- **No speaker adaptation :**

Training approach	Amount of training data	
	1h	7h
GMM	40.3	36.0
SGMM	37.8	18.9
DNN	26.9	18.4
# of states	661	2998

Cross-lingual acoustic modelling for low-resourced Iban ASR

- **Training data :**
  - Malay - 120h MASS corpus
  - English - 118h TED corpus
  - Iban - 1h condition ; 7h condition
- **Using SGMM :**
  - No speaker adaptation
  - UBM Gaussians : 600
  - No. of substates : 805 (1h) and 10K (7h)
  - Approach: **initialize Iban SGMM using UBM trained on source language data** (monolingual or multilingual)
- **Using DNN :**
  - 6 hidden layers, each with 1024 units
  - LDA, MLLT, SAT-fMLLR (speaker adaptation)
  - Approach: **fine-tune hidden layers trained on source language data to Iban training data**

Evaluation of cross-lingual/multilingual SGMM on Iban ASR

Cross-lingual SGMM	Amount of training data	
	1h	7h
Using monolingual UBM:		
a. Malay	**28.3**	19.4
b. English	30.8	19.2
Using multilingual UBM:		
a. Iban + Malay	**27.2**	19.6
b. Iban + English	29.8	19.2
c. English + Malay	29.4	19.1
d. Iban + Malay + English	28.3	19.2
# of substates	805	10K

Evaluation of language specific DNN on Iban ASR

DNN with lang. specific top layer	Amount of train data	
	1h	7h
a. Hidden layers from English	19.1	15.2
b. Hidden layers from Malay	18.9	15.2

Towards a zero-shot ASR

- **Approach and setup :**
  - Train Iban ASR on automatic transcripts - obtained from decoding Iban training data with Malay acoustic models
  - Malay acoustic models - 120h training data; SGMMs
  - Iban ASR (from automatic transcripts) - 7h training data; train GMM, SGMM and DNN models
- **Results :**

ASR system (7h)	GMM	SGMM	DNN
Supervised (no spkr adapt.)	36.0	18.9	18.4
Supervised (with spkr adapt.)	19.7	16.6	15.8
Unsupervised (with spkr adapt.)	21.4	18.6	18.9

Conclusions

- Built first ASR system for Iban - corpus and Kaldi scripts available on github : **https://github.com/sarahjuan/iban**
- **Using Malay (closely-related) data in the lexicon design for Iban is better than using English (not a close language)**
- Cross-lingual effect on acoustic model is more evident on SGMM experiment for 1h training data (very limited condition)
- Language specific top layer for DNN (**English and Malay source languages do not make a difference for Iban DNN**)
- Improving **Zero-shot ASR**: conf measures to select the best transcripts