1-2-2-MLP

Code-switched language modelling using a code predictive LSTM in under-resourced South African languages

Joshua Jansen van Vüren Thomas Niesler
Department of Electrical and Electronic Engineering, University of Stellenbosch, South Africa
jjvanvueren@sun.ac.za trn@sun.ac.za

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT

Motivation

- Code-switching is the use of more than one language within or between sentences.
- The phenomenon occurs predominantly in spontaneous speech, but is sparse compared to monolingual text.
- It is extremely costly to transcribe corpora of code-switched speech because specialised multilingual transcribers are required.
- Higher language model perplexities and word error rates are observed across language switches.
- Previous research has investigated using the current language as a selection signal for a monolingual language model to model the current input token (Garg et al., 2018).
- Our work extends this idea by incorporating a mechanism to directly model a language switch.

Dataset

- Our dataset contains five languages arranged as four bilingual corpora. It is highly under-resourced.

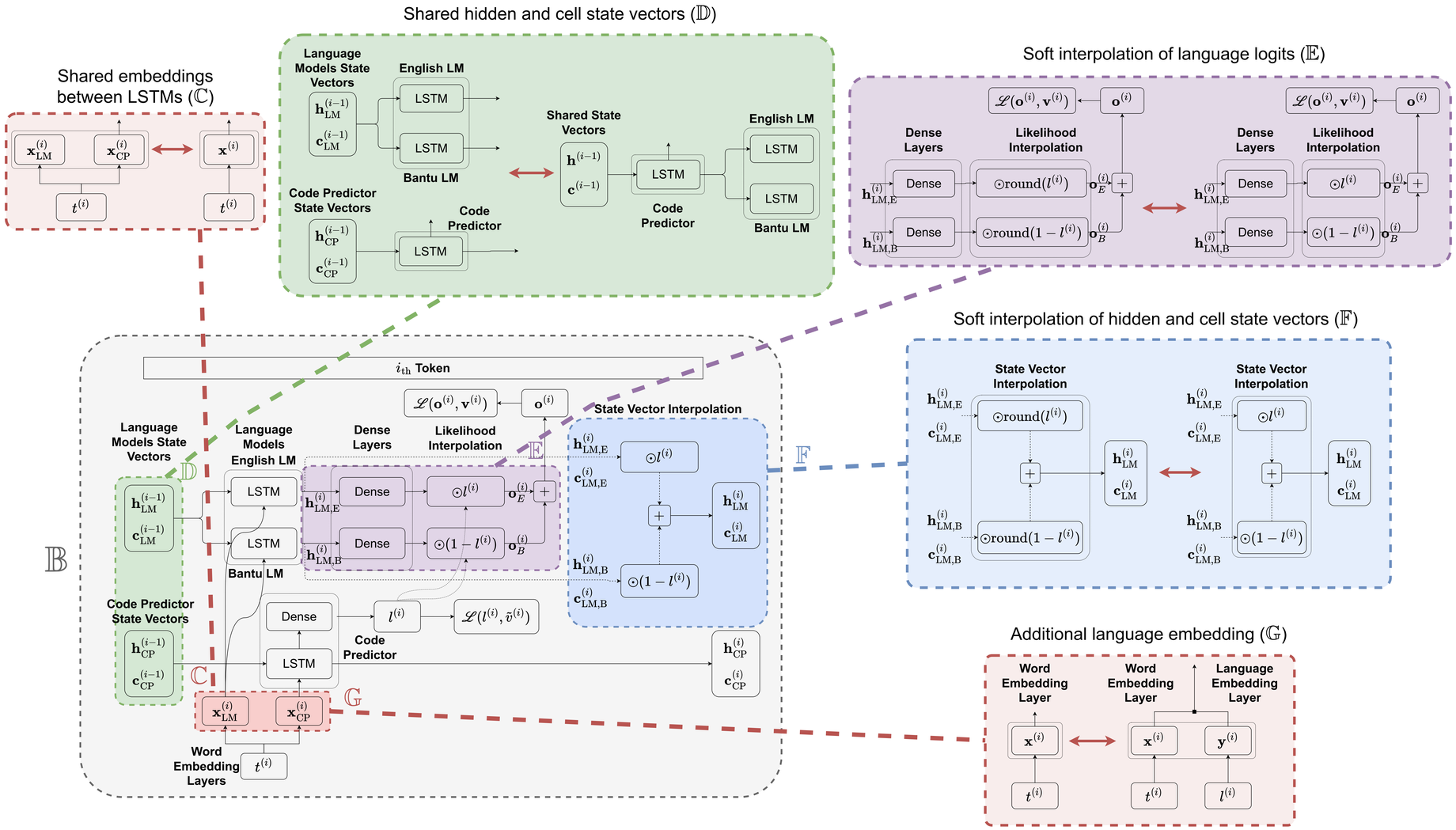
Pair	Partition	Tok	Typ	CS_{EB}	CS_{BE}	Dur
English-isiZulu (EZ)	Train	52383	10396	2236	2743	4.81h
	Dev	1566	866	175	198	8.00m
	Test	5656	2305	688	776	30.4m
English-isiXhosa (EX)	Train	32539	7716	776	1003	2.68h
	Dev	2300	1246	91	113	13.7m
	Test	2651	1387	328	363	14.3m
English-Sesotho (ES)	Train	35197	4339	1565	1719	2.36h
	Dev	3067	1050	156	166	12.8m
	Test	4054	1193	403	396	15.5m
English-Setswana (ET)	Train	35725	3808	1885	1951	2.33h
	Dev	3707	1052	224	251	13.8m
	Test	4939	1254	505	526	17.8m

Experiments

- We conducted an ablative investigation of six configurations of the proposed code-switched language model.
- The language models were also used to generate corpora of synthetic code-switched text and employed to rescore N-best lists.

Model Configurations

Alias	Configuration
$\mathbb{A}_\phi$	Baseline Kaldi system (CNN-TDNN-F acoustic model)
$\mathbb{A}$	Baseline LSTM (Single layer LSTM)
$\mathbb{B}$	Separate embeddings per LSTM (Default)
$\mathbb{C}$	$\mathbb{B}$ with shared embeddings between LSTMs
$\mathbb{D}$	$\mathbb{C}$ with shared hidden and cell state vectors
$\mathbb{E}$	$\mathbb{D}$ with soft interpolation of language logits
$\mathbb{F}$	$\mathbb{E}$ with soft interpolation of hidden and cell state vectors
$\mathbb{G}$	$\mathbb{E}$ with additional language embedding



Results

- The final model $\mathbb{G}$ consistently outperforms the baseline LSTM ($\mathbb{A}$) in terms of perplexity and code-switched perplexity.
- The same model $\mathbb{G}$ also outperforms the baseline speech recognition system ($\mathbb{A}_\phi$) in word error rate and word error rate across over language switches.

Development set perpelxity (PP), code-switched perplexity (CPP), test set word error rate (WER), code-switched bigram error (CSBG).

	English-isiZulu				English-isiXhosa				English-Sesotho				English-Setswana			
Alias	PP	CPP	WER	CSBG	PP	CPP	WER	CSBG	PP	CPP	WER	CSBG	PP	CPP	WER	CSBG
$\mathbb{A}_\phi$	-	-	41.8	63.7	-	-	42.9	69.5	-	-	50.6	67.2	-	-	41.9	57.4
$\mathbb{A}$	976.6	5228.8	40.3	60.3	888.1	8048.9	42.6	69.6	350.3	3209.7	49.7	**66.2**	204.9	1476.1	40.1	53.2
$\mathbb{D}$	956.1	5498.4	40.7	61.7	738.6	12284.8	42.8	70.9	327.8	2958.4	50.6	68.7	198.7	1567.9	39.6	53.7
$\mathbb{E}$	904.8	5147.1	40.3	61.1	744.2	12077.1	**42.0**	**68.3**	322.5	2449.7	49.0	66.5	199.3	1260.6	39.8	**52.7**
$\mathbb{G}$	**832.1**	4841.4	**39.9**	**60.0**	**725.4**	13198.6	42.1	69.3	**319.1**	2728.6	**48.9**	66.3	**194.2**	1358.3	**39.3**	53.1

Supported by the Department of Sport, Arts and Culture of the Republic of South Africa