Human Language Technologies Assignment 2 Report

Daniel Conroy

The output of the language model builder LMTool includes an .lm language model file and a .dic pronouncing dictionary file. The .lm file is produced by the QuickLM statistical language modeler. QuickLM is open-source, so it's source code[1] can be read to understand it's outputs in the .lm file. I discuss examples of it's outputs later. The pronunciations LMTool generates and outputs in the .dic file are taken from the pronunciation dictionary cmudict where possible, but this dictionary may contain errors[2] or pronunciations specific to American English.

After using LMTool to build a language model for the text 100west.txt, I chose the following ten words from the .dic file to model the syllables in them. Alongside them are the phoneme representations of the words in the .dic file and slightly modified versions of these phoneme representations (changes are in red) that I will use to model syllables in the words. (I modify the generated phoneme representation for camaraderie to make it sound closer to the French pronunciation.) I also place syllable boundaries (denoted by $) in the latter phoneme representations of the words.

Words	.dic Phoneme Representations	Personal Phoneme Representations
Cabins	K AE B AH N Z	K AE B $ IH N Z
Cable	K EY B AH L	K EY $ B AH L
Camaraderie	K AA M ER AA D ER IY	K AE $ M AE $ R AE $ D ER $ IY
Campgrounds	K AE M P G R AW N D Z	K AE M P $ G R AW N D Z
Capable	K EY P AH B AH L	K EY $ P AX $ B AH L
Civilization	S IH V AH L IH Z EY SH AH N	S IH $ V AH L $ AY Z $ EY $ SH AH N
Conservationists	K AA N S ER V EY SH AH N IH S T S	K AA N $ S ER $ V EY $ SH AH N $ IH S T S
Coyotes	K AY OW T IY S	K AY $ OW $ T IY S
Cry	K R AY	K R AY
Crystals	K R IH S T AH L Z	K R IH $ S T AH L Z

Here is a list of the phoneme representations of the syllables in these words.

Onset		Peak	Coda		
	K	AE	B		
	K	AE			
	K	AE	M	P	
	M	AE			
	R	AE			
		IH	N	Z	
	S	IH			
		IH	S	T	S
K	R	IH			
	K	EY			
		EY			
	V	EY			
	B	AH	L		
	V	AH	L		
	SH	AH	N		
S	T	AH	L	Z	
	D	ER			
	S	ER			

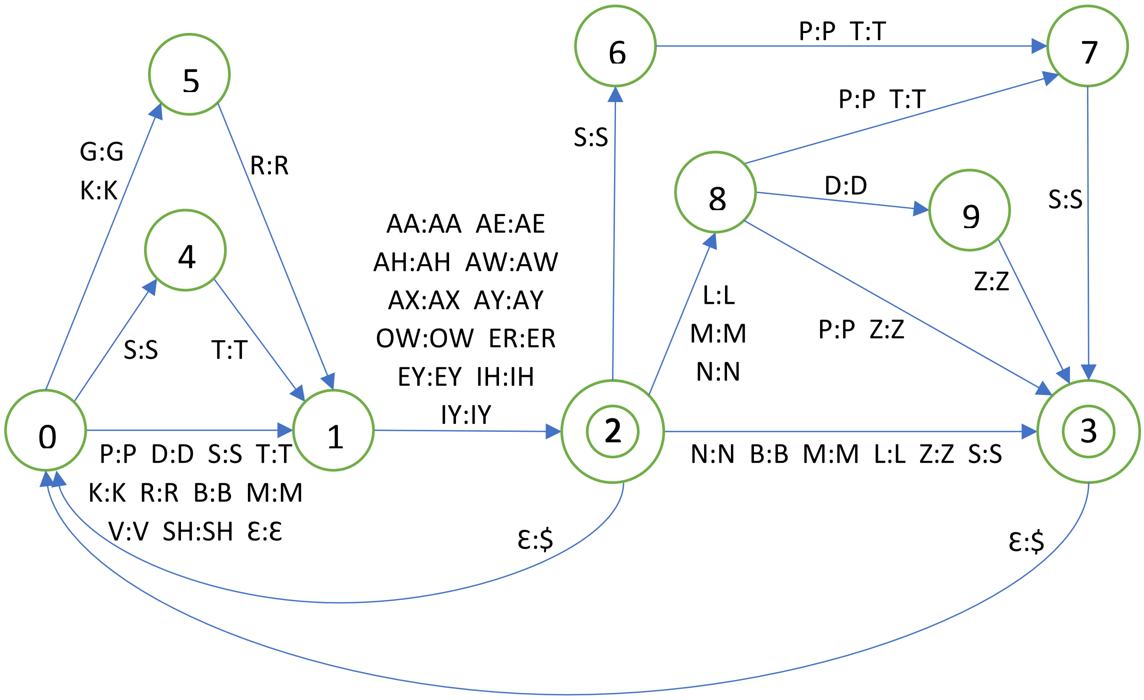
[1] https://www.speech.cs.cmu.edu/tools/download/quick_lm.pl
[2] www.speech.cs.cmu.edu/tools/FAQ.html

Onset		Peak	Coda		
		IY			
	T	IY	S		
G	R	AW	N	D	Z
	P	AX			
		AY	Z		
	K	AY			
K	R	AY			
	K	AA	N		
		OW			

Here is a phonotactic transducer that models these syllables. The transition sequence (1, 2) models the peak.

The transition sequences (0, 1), (0, 4, 1) and (0, 5, 1) model the onset. Since the onset is an optional part of a syllable, there is an epsilon-transition from state 0 to state 1. It's necessary to model onsets with two consonants using both states 4 and 5 since the phoneme combinations of G T and K T are impossible in English.

The transition sequences (2, 3), (2, 8, 3), (2, 6, 7, 3), (2, 8, 7, 3) and (2, 8, 9, 3) model the coda. This transducer includes extra transitions that model well-formed syllable codas in English that aren't in the above syllables. These are: M, L P, M Z, N P, M D Z, L D Z, S P S (as in crisps), L P S, L T S, M P S, M T S, N P S, N T S. Some of these may not actually occur in English.



The .lm file output LMTool produced for just the syllables in the words I chose is included below.

This language model provides information about the structure of the syllables, specifically which phonemes follow which phonemes in the syllables (indicated by bigrams) and what sequences of three phonemes (trigrams) are there in the syllables.

However, unlike the language model, the phonotactic transducer may be used to efficiently check (either programmatically or with the aid of a visualization of the transducer) whether a sequence of phonemes of any length is one of the syllables (or a part of one) or has a similar structure to one of the syllables (or a part of one).

Unlike the phonotactic transducer, the language model provides statistical information about the phonemes in the syllables.

The most common vowel in the syllables is AH. This is indicated by AH having the lowest magnitude log probability for vowels, -1.6659.

The most common consonant in the syllables is K, which has the lowest magnitude log probability for consonants, -1.4898. This is unsurprising since all the words start with the letter C, which is pronounced as K in all but one case.

The information provided for the bigram <s> K indicates that K is the most common first phoneme in the syllables, since the bigram has the lowest magnitude log “biprobability” for bigrams starting with <s>, -0.8239. The biprobability of a bigram indicates the proportion of times the second “word” of the bigram follows the first “word” of the bigram.

The most common consonant following the most common vowel in the syllables, AH, is L since the biprobability (log is -0.4771) for the bigram AH L is higher than the biprobability (log is -0.7782) for the only other bigram beginning with AH, AH N.

The most common vowel following the most common consonant in the syllables, K, is AE since the biprobability (log is -0.7782) for the bigram K AE is higher than the biprobabilities for all other bigrams beginning with K.

.lm file output:

```
\data\
ngram 1=27
ngram 2=68
ngram 3=67

\1-grams:
-0.9669 </s> -0.3010
-0.9669 <s> -0.2149
-2.4440 AA -0.2931
-1.7451 AE -0.2426
-1.6659 AH -0.2867
-2.4440 AW -0.2931
-2.4440 AX -0.2514
-1.9669 AY -0.2444
-1.9669 B -0.2408
-2.1430 D -0.2916
-2.1430 ER -0.2514
-1.8420 EY -0.2514
-2.4440 G -0.2947
-1.8420 IH -0.2317
-2.1430 IY -0.2408
-1.4898 K -0.2736
-1.8420 L -0.2444
-2.1430 M -0.2900
-1.7451 N -0.2408
-2.4440 OW -0.2514
```

```
-2.1430 P -0.2497
-1.8420 R -0.2802
-1.6659 S -0.2354
-2.1430 SH -0.2916
-1.9669 T -0.2786
-2.1430 V -0.2851
-1.8420 Z -0.2514

\2-grams:
-1.7782 <s> AY -0.2218
-1.4771 <s> B -0.1249
-1.7782 <s> D -0.1761
-1.7782 <s> EY 0.0000
-1.7782 <s> G 0.0000
-1.4771 <s> IH -0.1761
-1.7782 <s> IY -0.1761
-0.8239 <s> K 0.0000
-1.7782 <s> M -0.1761
-1.7782 <s> OW 0.0000
-1.7782 <s> P -0.1761
-1.7782 <s> R -0.2430
-1.3010 <s> S -0.1249
-1.4771 <s> SH 0.0000
-1.7782 <s> T -0.2218
-1.4771 <s> V 0.0000
-0.3010 AA N -0.1461
-0.5229 AE </s> -0.3010
-1.0000 AE B -0.2218
-1.0000 AE M -0.1761
-0.4771 AH L 0.0000
-0.7782 AH N -0.1461
-0.3010 AW N -0.2553
-0.3010 AX </s> -0.3010
-0.4771 AY </s> -0.3010
-0.7782 AY Z 0.0000
-0.7782 B </s> -0.3010
-0.4771 B AH -0.1249
-0.6021 D ER 0.0000
-0.6021 D Z 0.0000
-0.3010 ER </s> -0.3010
-0.3010 EY </s> -0.3010
-0.3010 G R -0.2430
-0.6021 IH </s> -0.3010
-0.9031 IH N -0.2553
-0.9031 IH S -0.2218
-0.6021 IY </s> -0.3010
-0.6021 IY S -0.2218
-1.2553 K AA 0.0000
-0.7782 K AE 0.0000
-1.2553 K AY -0.1249
-0.9542 K EY 0.0000
-0.9542 K R -0.1761
-0.4260 L </s> -0.3010
-0.9031 L Z 0.0000
-0.6021 M AE -0.1461
-0.6021 M P -0.1761
-0.5229 N </s> -0.3010
-1.0000 N D -0.1761
```

```
-1.0000 N Z 0.0000
-0.3010 OW </s> -0.3010
-0.6021 P </s> -0.3010
-0.6021 P AX 0.0000
-0.9031 R AE -0.1461
-0.9031 R AW 0.0000
-0.9031 R AY -0.1249
-0.9031 R IH -0.1761
-0.7782 S </s> -0.3010
-1.0792 S ER 0.0000
-1.0792 S IH -0.1761
-0.7782 S T -0.1249
-0.3010 SH AH -0.2218
-0.7782 T AH -0.1249
-0.7782 T IY -0.1761
-0.7782 T S -0.2218
-0.6021 V AH -0.1249
-0.6021 V EY 0.0000
-0.3010 Z </s> -0.3010

\3-grams:
-0.3010 <s> AY Z
-0.3010 <s> B AH
-0.3010 <s> D ER
-0.3010 <s> EY </s>
-0.3010 <s> G R
-0.6021 <s> IH N
-0.6021 <s> IH S
-0.3010 <s> IY </s>
-1.2553 <s> K AA
-0.7782 <s> K AE
-1.2553 <s> K AY
-0.9542 <s> K EY
-0.9542 <s> K R
-0.3010 <s> M AE
-0.3010 <s> OW </s>
-0.3010 <s> P AX
-0.3010 <s> R AE
-0.7782 <s> S ER
-0.7782 <s> S IH
-0.7782 <s> S T
-0.3010 <s> SH AH
-0.3010 <s> T IY
-0.6021 <s> V AH
-0.6021 <s> V EY
-0.3010 AA N </s>
-0.3010 AE B </s>
-0.3010 AE M P
-0.4260 AH L </s>
-0.9031 AH L Z
-0.3010 AH N </s>
-0.3010 AW N D
-0.3010 AY Z </s>
-0.3010 B AH L
-0.3010 D ER </s>
-0.3010 D Z </s>
-0.3010 G R AW
-0.3010 IH N Z
```

```
-0.3010 IH S T
-0.3010 IY S </s>
-0.3010 K AA N
-0.7782 K AE </s>
-0.7782 K AE B
-0.7782 K AE M
-0.3010 K AY </s>
-0.3010 K EY </s>
-0.6021 K R AY
-0.6021 K R IH
-0.3010 L Z </s>
-0.3010 M AE </s>
-0.3010 M P </s>
-0.3010 N D Z
-0.3010 N Z </s>
-0.3010 P AX </s>
-0.3010 R AE </s>
-0.3010 R AW N
-0.3010 R AY </s>
-0.3010 R IH </s>
-0.3010 S ER </s>
-0.3010 S IH </s>
-0.6021 S T AH
-0.6021 S T S
-0.3010 SH AH N
-0.3010 T AH L
-0.3010 T IY S
-0.3010 T S </s>
-0.3010 V AH L
-0.3010 V EY </s>

\end\
```