Homework 2: Solutions

Junkun's Notes Revised from a Previous Rubric
CS 539

1 Shannon Game and Entropy of English

2. To compute the entropy, suppose the probability of taking k guesses until the correct letter is guessed is p_k, then we have

$$H = -\sum_{k=1}^{27} p_k log(p_k)$$

2 Part-of-Speech Tagging as WFST Composition

2. Make sure to write `lexicon.wfst` in terms of $p(w_1, w_2, ..., w_n|t_1, t_2, ..., t_n)$ instead of $p(t_1, t_2, ..., t_n|w_1, w_2, ..., w_n)$.

   So your wfst file should look similar to:

```
(NOUN (FISH NOUN FISH 0.2))
(NOUN (PANDA NOUN PANDA 0.3))
(NOUN (CAN NOUN CAN 0.1))
(NOUN (ARROW NOUN ARROW 0.2))
...

(VERB (EATS VERB EATS 0.3))
...
```

   etc

4. If `bigram.wfst` is our model of $p(t_1, t_2, ..., t_n)$, and if `lexicon.wfst` is a model of $p(w_1, w_2, ..., w_n|t_1, t_2, ..., t_n)$, then composing them together we have the joint distribution:

$$p(w_1, ..., w_n|t_1, ..., t_n)p(t_1, ..., t_n) = p(w_1, ..., w_n, t_1, ..., t_n)$$

   By the chain rule. Or you can show $p(t_1, t_2, ..., t_n|w_1, w_2, ..., w_n)$ in Beyas' Rule.

3 Pronouncing and Spelling English

5. I used a prefix trie to carry this out, but I don't think you even need to do that. You could go through the data file line by line and generate your wfst with extra states this way.

   Here is how I designed my `espell-eword.wfst`, generated from `gen_espell_eword.py`:

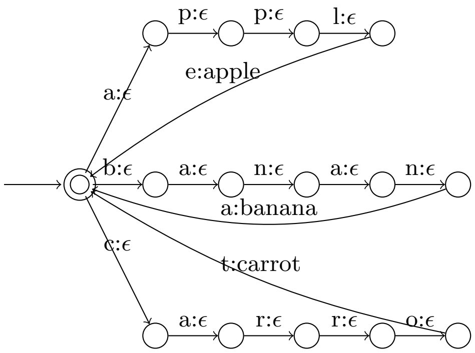


Figure 1: `espell-eword.wfst` with apple, banana, and carrot

7. `$ echo "W H A L E B O N E S" | carmel -sriIEQk 5 epron.wfsa epron-espell.wfst`

```
Input line 1: W H A L E B O N E S
(767405 states / 768107 arcs reduce-> 1676/2378)
(21019 states / 233760 arcs reduce-> 18074/214247)
W EY L B OW N Z 5.54489826143453e-12
W AH L IY B OW N Z 5.40084628191659e-12
W EY L B OW N Z 3.51523275430919e-12
W EY L B OW N Z 3.08999996423283e-12
W EY L AH B AH N Z 2.85840793768298e-12
Derivations found for all 1 inputs
Viterbi (best path) product of probs=5.54489826143453e-12, probability=2^-37.392 per-input-symbol-p
```

10. `echo "BEAR" | carmel -sliOEQk 10 eword-epron.wfst epron-eword.wfst eword.wfsa`

```
Input line 1: BEAR
(209214 states / 209213 arcs reduce-> 7/6)
(27 states / 38 arcs reduce-> 11/14)
(11 states / 14 arcs)
BEAR 0.00010622814
BARE 1.4904459e-05
BAER 2.167921e-06
BEHR 1.3549508e-06
BAHR 5.4198027e-07
```

This takes the word “BEAR”, turns it into its pronounciation phoneme components, and then tries to map this back to the original word, looking for the most likely word that the phonemes sound like.

4 Decoding English Words from Japanese Katakana

1. In my `estimate.py` I used a dictionary of dictionaries to estimate the probability of a Japanese phoneme sequence given an English phoneme $p(j|e)$.

   I used this to generate both `epron-jpron.probs` and `epron-jpron.wfst` at the same time.

   Lets say I had a dictionary like this:

   ```
   en_jp = {}
   en_jp["AH"] = {}
   ```

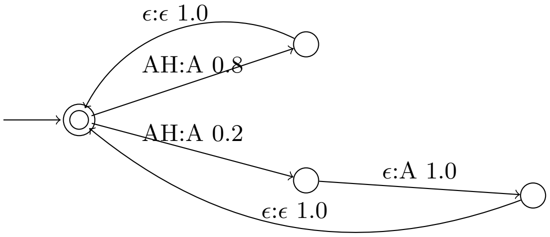
```
en_jp["AH"]["A"] = 80
en_jp["AH"]["A A"] = 20
en_jp["AH"]["A A A A"] = 1
```

Then maybe we could get rid of that one singleton mapping (that is longer than 3 phonemes) because it might be noise. I've also tried a version where I've kept everything, and it seems to work about the same. With the remaining elements, we can compute the probabilities:

$$p(jp = A|en = AH) = 0.8$$

$$p(jp = AA|en = AH) = 0.2$$

2. From those probabilities we can generate the wfst:



5. Here are the first two results from Japanese Katakana to English phonemes

```
$ cat jprons.txt | carmel -sribIEQk 5 epron-jpron.wfst

Input line 1: H I R A R I K U R I N T O N
(2499 states / 2642 arcs reduce-> 25/168)
HH IH L AE L IH K UH L IH N T AO N 0.0193183156429586
HH IH L AE L IH K UH R IH N T AO N 0.0186962177498504
HH IH L AE R IH K UH L IH N T AO N 0.0186962177498504
HH IH R AE L IH K UH L IH N T AO N 0.0186962177498504
HH IH L AE R IH K UH R IH N T AO N 0.0180941529587871

Input line 2: D O N A R U D O T O R A N P U
(2705 states / 2882 arcs reduce-> 51/228)
D AO N AE L UH D AO T AO L AE N P UH 0.00445507397728098
D AO N AE L UH D AO T AO R AE N P UH 0.00431160949589813
D AO N AE R UH D AO T AO L AE N P UH 0.00431160949589813
D AO N AE R UH D AO T AO R AE N P UH 0.00417276492824138
D AO N AE L UH D OW T AO L AE N P UH 0.00388852153589646
```

6. Here are all the results from Japanese Katakana to English words:

```
$ cat jprons.txt | carmel -sribIEQk 5 eword.wfsa eword-epron.wfst epron-jpron.wfst

Input line 1: H I R A R I K U R I N T O N
(1277 states / 1349 arcs reduce-> 21/93)
(1762 states / 2477 arcs reduce-> 1101/1662)
(962 states / 1384 arcs reduce-> 901/1323)
HILARY CLINTON 9.70302583148299e-15
HILLARY CLINTON 2.87497052832624e-15
```

```
HERE EARLY CLINTON 7.57797007325967e-16
HILL EARLY CLINTON 4.10739828626111e-16
HIGH RALLY CLINTON 3.37406355927389e-16

Input line 2: D O N A R U D O T O R A N P U
(1393 states / 1484 arcs reduce-> 35/126)
(2769 states / 3917 arcs reduce-> 1501/2266)
(1279 states / 1822 arcs reduce-> 1161/1703)
DONALD TRUMP 4.35362445143923e-14
DONALD TRAMP 2.42681870525719e-15
DONALD TRUMPS 1.40959006671416e-16
DONALD TRUMPED 4.83861140066993e-17
DONALD AUTO LUMP 6.6816683414895e-18

Input line 3: B I D E O T E E P U
(939 states / 986 arcs reduce-> 17/64)
(621 states / 870 arcs reduce-> 401/601)
(358 states / 515 arcs reduce-> 343/500)
VIDEOTAPE 6.73821209011932e-09
VIDEOTAPES 8.21757648780644e-10
VIDEOTAPED 8.97525972363738e-11
VIDEO TAPE 4.09038126127251e-12
VIDEO TAPES 4.19757140444325e-13

Input line 4: H O M A A SH I N P U S O N
(1214 states / 1291 arcs reduce-> 22/99)
(1331 states / 1836 arcs reduce-> 762/1137)
(667 states / 947 arcs reduce-> 623/902)
HOMER SIMPSON 5.2963416765271e-16
HOME RUSSIAN PUSAN 4.19732401380608e-18
HOME RE SIMPSON 3.03267977341654e-18
HOME RUSHING PUSAN 2.60451390983126e-18
HOME AIR SIMPSON 6.63666606257912e-19

Input line 5: R A PP U T O PP U
(777 states / 821 arcs reduce-> 18/62)
(583 states / 803 arcs reduce-> 320/470)
(308 states / 446 arcs reduce-> 302/440)
LAPTOP 1.71622503237902e-08
LAPTOPS 1.42493411813258e-10
WRAPPED UP 1.4701466646139e-11
WRAP TOP 1.43145429769306e-11
RAP TOP 1.32920756214356e-11

Input line 6: SH E E B I N G U K U R I I M U
(1355 states / 1428 arcs reduce-> 33/106)
(1337 states / 1808 arcs reduce-> 707/1028)
(632 states / 878 arcs reduce-> 570/815)
SAVING CREAM 8.66013834286229e-14
SAVINGS CREAM 5.18805370773068e-14
SHAVING CREAM 1.48421614965075e-14
```

```
SAVING SCREAM 3.81048598039438e-16
SAVING SCREAMS 8.47856736298164e-17

Input line 7: CH A I R U D O SH I I T O
(1117 states / 1196 arcs reduce-> 29/108)
(1288 states / 1794 arcs reduce-> 685/1046)
(622 states / 920 arcs reduce-> 596/894)
CHILD SHEET 6.37034664082233e-12
CHILD SEAT 2.08530875762869e-12
CHILD SHEET 3.67104111932581e-13
CHILD SHUT 9.25588890050669e-14
CHILD SHIITE 6.13002865992803e-14

Input line 8: SH I I T O B E R U T O
(1025 states / 1088 arcs reduce-> 29/92)
(1339 states / 1857 arcs reduce-> 656/1003)
(588 states / 867 arcs reduce-> 542/821)
SHEET BELT 3.53144709237838e-12
SHEET BUILT 1.19311434465159e-12
SEAT BELT 1.15600578179658e-12
SEAT BUILT 3.90561445402474e-13
SHEET BELT 2.03506782562953e-13

Input line 9: SH I N G U R U R U U M U
(1105 states / 1168 arcs reduce-> 38/101)
(1182 states / 1611 arcs reduce-> 578/862)
(502 states / 710 arcs reduce-> 468/676)
SINGLE ROOM 3.50896454618416e-12
SINGLE ROOMS 4.27377729298878e-13
SINGLE LOOM 2.11125023995056e-13
SINGLE LOOMS 5.11523247290928e-14
SINGLE ROOM 4.67098858356863e-14

Input line 10: G A A R U H U R E N D O
(1128 states / 1201 arcs reduce-> 29/102)
(1717 states / 2446 arcs reduce-> 953/1442)
(829 states / 1193 arcs reduce-> 776/1139)
GIRLFRIEND 5.65079271302624e-08
GIRLFRIENDS 5.83019569972329e-10
GIRLFRIEND 9.04175151631849e-11
GIRL FRIEND 1.9672218102104e-11
GIRLS FRIEND 4.86816267874451e-12

Input line 11: T O R A B E R A A Z U TCH E KK U
(1393 states / 1488 arcs reduce-> 34/129)
(3039 states / 4156 arcs reduce-> 1770/2646)
(1598 states / 2302 arcs reduce-> 1469/2169)
TRAVELERS CHECK 2.58540012505422e-14
TRAVELERS CZECH 2.34023292995593e-15
TRAVELLERS CHECK 6.80368525317696e-16
TRAVELERS CHUCK 3.40239801299651e-16
TRAVELERS CHECKS 2.76381590328967e-16

Input line 12: B E B I I SH I TT A A
(943 states / 1008 arcs reduce-> 19/84)
(983 states / 1366 arcs reduce-> 616/941)
(533 states / 775 arcs reduce-> 498/740)
BABY SITTER 5.60486230580902e-15
BABY SHUTTER 4.43615834526301e-15
BABY SEATER 3.56457475869954e-15
BABY SCHUTTER 1.66355955340684e-15
BABY SEITER 5.16614673368459e-16

Input line 13: S U K O TT O R A N D O
(1044 states / 1108 arcs reduce-> 24/88)
(1644 states / 2330 arcs reduce-> 935/1400)
(825 states / 1180 arcs reduce-> 756/1110)
SCOTLAND 3.15387954815166e-08
SCOTT LAND 4.29436353952433e-11
SCOTT RAND 4.04752498514282e-12
SCOTT LEARNED 1.55545515333612e-12
SCOTT ROUND 9.45535101302506e-13

Input line 14: B A I A R I N K O N TCH E R U T O
(1471 states / 1568 arcs reduce-> 34/131)
(1792 states / 2509 arcs reduce-> 1221/1850)
(1044 states / 1496 arcs reduce-> 944/1396)
VIOLIN CONCERTO 1.03424424658838e-17
VIA RIM CONCERTO 4.77579923845313e-20
VIA LINE CONCERTO 3.04493864934835e-20
BUY ALAN CONCERTO 2.33219219343927e-20
VIA RUN CONCERTO 1.87310820909558e-20

Input line 15: A PP U R U M A KK U B U KK U P U R O
(1553 states / 1645 arcs reduce-> 34/126)
(1312 states / 1724 arcs reduce-> 662/981)
(602 states / 861 arcs reduce-> 582/841)
APPLE MAKE BOOK PRO 2.10820620368463e-22
APPLE MAC BOOK PRO 2.07250509727058e-22
OPERA MAKE BOOK PRO 1.06235432441317e-22
OPERA MAC BOOK PRO 1.04436404209685e-22
APPLE MAKE BOOK PLOUGH 2.28684958197915e-23

Input line 16: K O N P I U U T A S A I E N S U
(1470 states / 1570 arcs reduce-> 33/133)
(1897 states / 2554 arcs reduce-> 1020/1522)
(918 states / 1318 arcs reduce-> 857/1256)
COMPUTER SCIENCE 2.67454347672193e-14
COMPUTER SAYINGS 2.12261588573531e-19
COMPUTER THREE INS 1.61631262585561e-19
COMPUTER SAY ONCE 1.6101316492732e-19
COMPUTER SAY INS 8.56862905185082e-20
```

```
Input line 17: H I J I K A R U T O R E E N I N G U
(1626 states / 1728 arcs reduce-> 47/149)
(3479 states / 4845 arcs reduce-> 1749/2620)
(1546 states / 2214 arcs reduce-> 1457/2125)
PHYSICAL TRAINING 2.1025911813503e-13
PHYSICAL STRAINING 5.34734836388998e-16
HIS EAKLE TRAINING 4.12762975614606e-17
HIS OCCURS TRAINING 1.00568743709958e-18
PHYSICAL TRAIN ING 7.14914486839616e-19

Input line 18: H I J I K A R U E K U S A S A I S U
(1641 states / 1758 arcs reduce-> 41/158)
(3744 states / 5168 arcs reduce-> 2173/3249)
(1944 states / 2791 arcs reduce-> 1834/2681)
PHYSICAL EXERCISE 1.3415837860622e-15
PHYSICAL TEXAS ICE 5.69500819575834e-19
HIS EAKLE EXERCISE 2.63368419159696e-19
PHYSICAL EXCESS ICE 1.09342764704599e-19
PHYSICAL TEXAS EYES 8.03699531645734e-20

Input line 19: A I S U K U R I I M U
(1024 states / 1095 arcs reduce-> 28/99)
(1463 states / 1979 arcs reduce-> 755/1127)
(669 states / 955 arcs reduce-> 614/899)
ICE CREAM 4.9669803665802e-12
RE SCREAM 1.41322936061037e-12
EYE SCREAM 1.20259812270587e-12
RISK THEME 7.35401617300449e-13
EYES CREAM 7.00957690857633e-13

Input line 20: H O TT O M I R U K U
(940 states / 989 arcs reduce-> 19/68)
(744 states / 1020 arcs reduce-> 407/614)
(350 states / 500 arcs reduce-> 321/471)
HOT MILK 8.64424385294561e-12
FOUGHT MILK 2.60206336823283e-12
HUT MILK 4.03271704329609e-13
PHOTO MILK 2.15217920399196e-13
FOOT MILK 1.96123576495779e-13

Input line 21: T O R I P U R U R U U M U
(1193 states / 1263 arcs reduce-> 36/106)
(1345 states / 1831 arcs reduce-> 680/1025)
(586 states / 837 arcs reduce-> 543/794)
TRIPLE ROOM 2.83702641921718e-12
TRIPLE ROOMS 3.45538375508666e-13
TRIPLE LOOM 1.70696301700509e-13
TRIPLE LOOMS 4.13570712245085e-14
TRIPLE ROOM 3.7765323191586e-14
```

```
Input line 22: K U R A U N P U R A Z A H O T E R U
(1650 states / 1758 arcs reduce-> 35/143)
(2323 states / 3174 arcs reduce-> 1160/1730)
(1048 states / 1506 arcs reduce-> 983/1440)
CROWN PLAZA HOTEL 3.15545205414176e-18
CROWN PLAZA HOTEL 1.54171826136237e-18
CROWN PLAZA HOTELS 5.28947454483843e-19
CROWN PLAZA HOTELS 2.58437756583399e-19
CROWN POOR OTHER HOTEL 1.94355947620941e-19

Input line 23: H E E S U B U KK U R I S A A TCH I S A I E N T I S U T O
(2411 states / 2575 arcs reduce-> 65/229)
(3099 states / 4092 arcs reduce-> 1589/2367)
(1435 states / 2059 arcs reduce-> 1342/1965)
FACE BOOK RESEARCH SCIENTIST 3.89870846090939e-22
FAITH BOOK RESEARCH SCIENTIST 5.13598432306903e-23
PHASE BOOK RESEARCH SCIENTIST 1.06166545584978e-23
HAYES BOOK RESEARCH SCIENTIST 4.96597768133716e-24
FACE BOOKS RESEARCH SCIENTIST 3.54148270411558e-24

Input line 24: W O R U H U G A N G U M O TS U A R U T O
(1820 states / 1938 arcs reduce-> 53/171)
(3379 states / 4701 arcs reduce-> 1617/2465)
(1434 states / 2099 arcs reduce-> 1335/1998)
WOLFGANG MOZART 6.09837389646259e-21
WOLFGANG MOTOR ROUTE 6.04492840795533e-23
WOLFGANG MOTOR ROOT 1.8987273459307e-23
WOLFGANG MOTE ALTO 7.12715237408978e-24
WALL WHO GANG MOZART 6.74597343995131e-24
Derivations found for all 24 inputs
Viterbi (best path) product of probs=e^-739.505694470539, probability=2^-1066.88 per-input-symbol-p
```

9. One possibility could be to go from English pronunciations to English spellings. And then English spellings to English words. This would let us spell any word from any Japanese Katakana sequence, making us robust against words not found in our training set. At the same time, this would make our composed machine a lot larger slowing us down.

   Another possibility would be to take the intersection of the `epron-jpron.data` and `eword-epron.data` datasets. And then we could build a machine that directly goes from Japanese Katakana to English words. This would give us a small lightweight machine that would run quickly but it would be very brittle, giving us no output for anything not contained in our combined dataset.