Independent Project 2: Sequence Labeling

CS 6501-005 Natural Language Processing

Sri Vaishnavi Vemulapalli (sv2fr)

1 Hidden Markov Models

1.1 A Toy Problem

Note: The Viterbi algorithm was executed after converting given probabilities to log space.

1. Trellis tables

   Viterbi Variables

	G	C	A	C	T	G
H	-1.714798	-3.835062	-6.360791	-8.481054	-11.006783	-12.968441
L	-2.525729	-4.240527	-5.955326	-8.257911	-10.155031	-12.457616

   Back Pointers

	G	C	A	C	T	G
H	Start	H	H	H	H	L
L	Start	H	H	L	L	L

2. Hidden states of DNA sequence

   ['H', 'H', 'L', 'L', 'L', 'L']

1.2 POS Tagging

1. K = 6

   The cutoff value was chosen by examining the count of words per frequency.

```
(1, 20922) => 20922 words have frequency 1
[(1, 20922), (2, 6645), (3, 3290), (4, 2154), (5, 1493),
 (6, 1188), (7, 849), (8, 670), (9, 587), (10, 499),
 (11, 476), (12, 405), (13, 342), (14, 282), (16, 258),
 (15, 256), (17, 217), (18, 195), (19, 187), (20, 165)]
```

   Initially a cutoff value of 10 was chosen as it seemed ideal. However, while tuning the alpha-beta parameters, there was no significant improvement in the accuracy on dev set.

   Hence, by re-examining the frequency counts, a cutoff value of 6 was chosen. This gave a slightly better accuracy on the dev set.

   V = 9739

   The size of vocabulary was calculated after replacing all the words with frequency less than or equal to 6 with Unk. (for cutoff 10 the vocab size was 7134.)

2. The transition probability table:

	A	C	D	M	N	O	P	R	V	W	End
Start	0.037987	0.141893	0.240420	0.000536	0.236926	0.133002	0.076899	0.047988	0.021388	0.062961	0.000000
A	0.075417	0.088937	0.003324	0.000465	0.702922	0.083980	0.001310	0.030772	0.009380	0.000972	0.002521
C	0.104518	0.028970	0.286880	0.002168	0.324483	0.097026	0.063766	0.026628	0.057393	0.007575	0.000593
D	0.231103	0.010097	0.004815	0.002060	0.660307	0.042826	0.001138	0.014867	0.027788	0.000023	0.004975
M	0.000574	0.002868	0.004110	0.000191	0.001816	0.011566	0.005353	0.173389	0.799656	0.000000	0.000478
N	0.011204	0.221785	0.000563	0.017756	0.259672	0.252889	0.030655	0.049085	0.133388	0.001994	0.021009
O	0.031756	0.098601	0.056478	0.005487	0.179843	0.184832	0.031083	0.036685	0.078881	0.021238	0.275118
P	0.127741	0.024720	0.000055	0.062486	0.334603	0.054319	0.000164	0.040863	0.340496	0.000658	0.013895
R	0.092873	0.090047	0.080373	0.005946	0.056341	0.176629	0.014783	0.069334	0.408811	0.003811	0.001051
V	0.077035	0.163538	0.160559	0.000756	0.134886	0.103973	0.057537	0.165761	0.129571	0.005067	0.001317
W	0.027626	0.011660	0.062290	0.086870	0.090756	0.015126	0.058613	0.051155	0.504727	0.000210	0.090966

The sum of transition probabilities was verified:
Note: Because floating point numbers are not accurately represented in computers, the sum of probabilities does not add up to exactly 1.0. However, the difference is very minute and this should be considered as a limitation in computers and not as a problem in the algorithm.

```
Start 1.0
A 1.0000000000000002
C 0.9999999999999999
D 1.0
M 0.9999999999999999
N 0.9999999999999999
O 1.0000000000000002
P 0.9999999999999999
R 1.0
V 0.9999999999999999
W 1.0
```

3. The emission probability table:

	In	the	aftermath	of	stock	market	's	Unk	190-point	drop	...
A	0.000000	0.000028	0.000000	0.000000	0.000014	0.000000	0.000000	0.152215	0.000268	0.000000	...
C	0.014344	0.000000	0.000000	0.185907	0.000000	0.000000	0.000000	0.001034	0.000000	0.000000	...
D	0.000000	0.501400	0.000000	0.000000	0.000000	0.000000	0.000000	0.000171	0.000000	0.000000	...
M	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.001338	0.000000	0.000000	...
N	0.000010	0.000020	0.000072	0.000000	0.004242	0.006383	0.000007	0.120463	0.000003	0.000569	...
O	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.045006	0.000000	0.000000	...
P	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.237695	0.000822	0.000000	0.000000	...
R	0.000033	0.000000	0.000000	0.000066	0.000000	0.000000	0.000000	0.024853	0.000000	0.000000	...
V	0.000000	0.000007	0.000000	0.000000	0.000082	0.000180	0.009319	0.077050	0.000000	0.000359	...
W	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.003887	0.000000	0.000000	...

10 rows × 9739 columns

The sum of emission probabilities was verified:
Note: Because floating point numbers are not accurately represented in computers, the sum of probabilities does not add up to exactly 1.0. However, the difference is very minute and this should be considered as a limitation in computers and not as a problem in the algorithm.

```
A 1.0000000000000082
C 1.0
D 1.0000000000000002
M 1.0
N 1.0000000000000049
O 1.0000000000000018
P 1.0000000000000004
R 0.9999999999999992
V 1.000000000000012
W 0.9999999999999998
```

4. The emission probabilities were smoothed using smoothing parameter: $\alpha = 1$

   Comparing the emission table from before, it can be observed that the probabilities are no longer zero and have a small number.

	In	the	aftermath	of	stock	market	's	Unk	190-point	drop	...
A	0.000012	0.000037	0.000012	0.000012	0.000025	0.000012	0.000012	0.133868	0.000248	0.000012	...
C	0.013362	0.000007	0.000007	0.173093	0.000007	0.000007	0.000007	0.000970	0.000007	0.000007	...
D	0.000010	0.451370	0.000010	0.000010	0.000010	0.000010	0.000010	0.000164	0.000010	0.000010	...
M	0.000050	0.000050	0.000050	0.000050	0.000050	0.000050	0.000050	0.000743	0.000050	0.000050	...
N	0.000013	0.000022	0.000073	0.000003	0.004115	0.006190	0.000009	0.116767	0.000006	0.000555	...
O	0.000006	0.000006	0.000006	0.000006	0.000006	0.000006	0.000006	0.042500	0.000006	0.000006	...
P	0.000022	0.000022	0.000022	0.000022	0.000022	0.000022	0.187639	0.000671	0.000022	0.000022	...
R	0.000042	0.000014	0.000014	0.000071	0.000014	0.000014	0.000014	0.021439	0.000014	0.000014	...
V	0.000007	0.000014	0.000007	0.000007	0.000084	0.000174	0.008693	0.071822	0.000007	0.000342	...
W	0.000052	0.000052	0.000052	0.000052	0.000052	0.000052	0.000052	0.001973	0.000052	0.000052	...

10 rows × 9739 columns

The transition probabilities were smoothed using smoothing parameter: $\beta = 1$

Comparing the transition table from before, it can be observed that the probabilities are no longer zero and have a small number.

	A	C	D	M	N	O	P	R	V	W	End
Start	0.037999	0.141886	0.240394	0.000554	0.236901	0.132996	0.076903	0.047997	0.021402	0.062968	0.000018
A	0.075421	0.088939	0.003337	0.000479	0.702837	0.083982	0.001324	0.030782	0.009392	0.000986	0.002535
C	0.104517	0.028976	0.286866	0.002175	0.324466	0.097026	0.063769	0.026633	0.057396	0.007582	0.000601
D	0.231088	0.010108	0.004826	0.002072	0.660243	0.042832	0.001150	0.014877	0.027796	0.000034	0.004986
M	0.000668	0.002960	0.004202	0.000286	0.001910	0.011650	0.005443	0.173319	0.798988	0.000095	0.000573
N	0.011207	0.221781	0.000566	0.017759	0.259667	0.252884	0.030657	0.049086	0.133387	0.001997	0.021012
O	0.031760	0.098601	0.056480	0.005493	0.179838	0.184827	0.031087	0.036689	0.078882	0.021242	0.275107
P	0.127733	0.024741	0.000082	0.062497	0.334539	0.054332	0.000192	0.040879	0.340430	0.000685	0.013919
R	0.092874	0.090049	0.080376	0.005962	0.056348	0.176616	0.014797	0.069339	0.408760	0.003827	0.001068
V	0.077036	0.163533	0.160554	0.000763	0.134883	0.103973	0.057540	0.165756	0.129569	0.005074	0.001325
W	0.027702	0.011752	0.062329	0.086884	0.090766	0.015215	0.058657	0.051207	0.504302	0.000315	0.090976

5. The smoothed transition probabilities were converted to log-space:

	A	C	D	M	N	O	P	R	V	W	End
Start	-3.270201	-1.952735	-1.425476	-7.497521	-1.440113	-2.017438	-2.565208	-3.036607	-3.844269	-2.765131	-10.898719
A	-2.584674	-2.419805	-5.702586	-7.644286	-0.352630	-2.477150	-6.627352	-3.480818	-4.667856	-6.922151	-5.977690
C	-2.258401	-3.541302	-1.248740	-6.130694	-1.125575	-2.332772	-2.752486	-3.625597	-2.857781	-4.881935	-7.417238
D	-1.464957	-4.594460	-5.333698	-6.179425	-0.415147	-3.150460	-6.768311	-4.207942	-3.582859	-10.284819	-5.301213
M	-7.310550	-5.822473	-5.472271	-8.157848	-6.260728	-4.452439	-5.213409	-1.752620	-0.224410	-9.256460	-7.464701
N	-4.491219	-1.506065	-7.476843	-4.030878	-1.348357	-1.374825	-3.484884	-3.014177	-2.014498	-6.215904	-3.862669
O	-3.449535	-2.316673	-2.873861	-5.204341	-1.715699	-1.688336	-3.470975	-3.305289	-2.539802	-3.851757	-1.290595
P	-2.057813	-3.699290	-9.406400	-2.772644	-1.095002	-2.912647	-8.559103	-3.197140	-1.077547	-7.286137	-4.274531
R	-2.376512	-2.407400	-2.521043	-5.122405	-2.876201	-1.733775	-4.213303	-2.668744	-0.894627	-5.565769	-6.842421
V	-2.563476	-1.810739	-1.829122	-7.177723	-2.003347	-2.263627	-2.855279	-1.797238	-2.043540	-5.283548	-6.626546
W	-3.586251	-4.443701	-2.775321	-2.443187	-2.399470	-4.185466	-2.836051	-2.971885	-0.684580	-8.063588	-2.397161

The smoothed emission probabilities were converted to log-space:

	In	the	aftermath	of	stock	market	's	Unk	190-point	drop	...
A	-11.299039	-10.200427	-11.299039	-11.299039	-10.605892	-11.299039	-11.299039	-2.010905	-8.303307	-11.299039	...
C	-4.315324	-11.858067	-11.858067	-1.753927	-11.858067	-11.858067	-11.858067	-6.938086	-11.858067	-11.858067	...
D	-11.488459	-0.795469	-11.488459	-11.488459	-11.488459	-11.488459	-11.488459	-8.715870	-11.488459	-11.488459	...
M	-9.913487	-9.913487	-9.913487	-9.913487	-9.913487	-9.913487	-9.913487	-7.205437	-9.913487	-9.913487	...
N	-11.280763	-10.721147	-9.531563	-12.667058	-5.493099	-5.084828	-11.568445	-2.147574	-11.973910	-7.496574	...
O	-12.069646	-12.069646	-12.069646	-12.069646	-12.069646	-12.069646	-12.069646	-3.158251	-12.069646	-12.069646	...
P	-10.741319	-10.741319	-10.741319	-10.741319	-10.741319	-10.741319	-1.673234	-7.307332	-10.741319	-10.741319	...
R	-10.066428	-11.165040	-11.165040	-9.555602	-11.165040	-11.165040	-11.165040	-3.842530	-11.165040	-11.165040	...
V	-11.872982	-11.179835	-11.872982	-11.872982	-9.388075	-8.654106	-4.745288	-2.633568	-11.872982	-7.981161	...
W	-9.865734	-9.865734	-9.865734	-9.865734	-9.865734	-9.865734	-9.865734	-6.228148	-9.865734	-9.865734	...

10 rows × 9739 columns

The Viterbi algorithm was implemented using the above transition and emission probability tables.

Decoding obtained for the first five sentences is shown below.

	predicted	actual
0	[D, N, O, D, N, A, N, O, M, V, D, N, M, V, M, V, D, N, O, A, N, M, V, D, N, O]	[N, N, O, D, A, N, N, O, V, V, A, N, C, R, C, V, D, A, O, A, N, C, A, N, V, O]
1	[D, N, M, V, D, N, M, V, D, N, C, D, N, C, D, N, C, D, N]	[R, W, V, N, N, V, C, R, O, N, C, N, N, C, O, N, C, D, A]
2	[D, N]	[N, O]
3	[D, N, C, D, N, O, D, N, M, V, D, N, M, V, D, N, O]	[R, V, N, V, P, O, D, N, R, V, R, V, R, C, D, N, O]
4	[D, N, M, V, D, N, C, D, N, C, D, N, C, D, N, M, V, M, V, R, A, N, M, V, D, N, M, V, R, A, N, M, V, M, V, A, N, O]	[C, V, N, C, D, N, C, A, N, C, D, N, C, D, N, N, O, P, V, R, V, A, R, V, P, N, C, D, N, N, V, R, V, R, C, N, C, A, N, N, O]

Accuracy on the dev set: 0.94259
(The accuracy was calculated at the tag level.)

Average accuracy per sentence: 17.5554

6. The algorithm was run on the test set. The decoding for first five sentences:

```
Newspaper/N publishers/N are/V reporting/V mixed/A third-quarter/A results/N ,/O aided/V by/C favorable/A newsprint/N prices/N
and/C hampered/V by/C flat/A or/C declining/V advertising/N linage/N ,/O especially/R in/C the/D Northeast/N ./O

Adding/N to/R unsteadiness/V in/C the/D industry/N ,/O seasonal/A retail/A ad/N spending/N patterns/N in/C newspapers/N have/V
been/V upset/V by/C shifts/N in/C ownership/N and/C general/A hardships/N within/C the/D retail/A industry/N ./O

In/C New/N York/N ,/O the/D Bonwit/A Teller/N and/C B./N Altman/N &/C Co./N department/N stores/N have/V filed/V for/C protect
ion/N from/C creditors/N under/C Chapter/N 11/O of/C the/D federal/A Bankruptcy/N Code/N ,/O while/C the/D R.H./A Macy/N &/C C
o./N ,/O Bloomingdale/N 's/V and/C Saks/N Fifth/N Avenue/N department-store/N chains/N are/V for/C sale/N ./O
Many/A papers/N throughout/C the/D country/N are/V also/R faced/V with/C a/D slowdown/N in/C classified-ad/N spending/N ,/O a/
D booming/A category/N for/C newspapers/N in/C recent/A years/N ./O

Until/C recently/R ,/O industry/N analysts/N believed/V decreases/N in/C retail/A ad/N spending/N had/V bottomed/V out/R and/C
would/M in/C fact/N increase/N in/C this/D year/N 's/P third/A and/C fourth/A quarters/N ./O
```

7. Initially the cutoff frequency was chosen as 10. The alpha-beta parameters were tuned and tag-level accuracy on dev set was obtained.

cutoff	alpha	beta	accuracy at tag level
10	0.001	0.001	0.9390
	0.05	0.05	0.9389
	0.1	0.1	0.9389
	0.5	0.5	0.9383
	1	1	0.9388
	2	1	0.9341
	3	1	0.9305
	3	2	0.9305
	3	3	0.9388

Even though there is a very minute difference in accuracy values, it can be observed that when alpha and beta values are equal, the accuracy is slightly better.

Since there was not much change in accuracy, the cutoff was changed to 6 to investigate if that resulted in better performance.

cutoff	alpha	beta	accuracy at tag level
6	1	1	0.9426
	0.001	0.001	0.9446

Decreasing the cutoff definitely increased the accuracy, but not significantly.

Since no significant difference was observed in accuracy levels with different values of parameters for the previous cutoff, the values of alpha-beta for which the dev accuracy was the best (0.001) were chosen to run as the final model for cutoff 6.

8. The algorithm was run on the test set using alpha and beta values as 0.001. The decoding for first five sentences:

```
Newspaper/N publishers/N are/V reporting/V mixed/A third-quarter/A results/N ,/O aided/V by/C favorable/A newsprint/N prices/N
and/C hampered/V by/C flat/A or/C declining/V advertising/N linage/N ,/O especially/R in/C the/D Northeast/N ./O

Adding/N to/R unsteadiness/V in/C the/D industry/N ,/O seasonal/A retail/A ad/N spending/N patterns/N in/C newspapers/N have/V
been/V upset/V by/C shifts/N in/C ownership/N and/C general/A hardships/N within/C the/D retail/A industry/N ./O

In/C New/N York/N ,/O the/D Bonwit/A Teller/N and/C B./N Altman/N &/C Co./N department/N stores/N have/V filed/V for/C protect
ion/N from/C creditors/N under/C Chapter/N 11/O of/C the/D federal/A Bankruptcy/N Code/N ,/O while/C the/D R.H./A Macy/N &/C C
o./N ,/O Bloomingdale/N 's/V and/C Saks/N Fifth/N Avenue/N department-store/N chains/N are/V for/C sale/N ./O
Many/A papers/N throughout/C the/D country/N are/V also/R faced/V with/C a/D slowdown/N in/C classified-ad/A spending/N ,/O a/
D booming/A category/N for/C newspapers/N in/C recent/A years/N ./O

Until/C recently/R ,/O industry/N analysts/N believed/V decreases/N in/C retail/A ad/N spending/N had/V bottomed/V out/R and/C
would/M in/C fact/N increase/N in/C this/D year/N 's/P third/A and/C fourth/A quarters/N ./O
```

2 Conditional Random Fields

1. Performance on dev set after adding the features:

feature	lbfgs	ap
tokens only	0.4569	0.7971
first letter (fl)	0.5953	0.8511
last letter (ll)	0.5373	0.8144
first letter and last letter	0.6321	0.8567
first two letters (f2l)	0.5712	0.8484
last two letters (l2l)	0.4748	0.8125
first two letters and last two letters	0.5758	0.8516
first three letters (f3l)	0.5613	0.8330
last three letters (l3l)	0.4545	0.8148
first three letters and last three letters	0.5410	0.8393
previous word (prev)	0.4659	0.7944
next word (next)	0.4850	0.7877
previous and next word	0.5031	0.7942
all of the above (fl ll f2l l2l f3l l3l prev next)	0.7235	0.8753

2. CRF consists of two parts, the decoding problem and the parameter estimation problem. For the later, we can use techniques such as logistic regression or perceptron.

   In case of logistic regression, the fitted output labels or $\hat{y}$ are defined as: $\hat{y} \leftarrow \text{argmax}_y \, P(y|x;\theta)$

   i.e., the class that has the maximum probability is chosen as the fitted label for the given input.

   Whereas for perceptron algorithm, $\hat{y}$ is defined as: $\hat{y} \leftarrow \text{argmax}_y \, w^T f(x,y)$

   Here, $f(x,y)$ is a feature function and w is the corresponding weight. The feature function is used to model contextual information. Hence, it contains information about either the previous or next labels in the sequence for the current label. (In the previous question, we constructed different features by passing the previous, next, first two letters etc.)

   Hence, the perceptron algorithm is trying to learn using contextual information from its surroundings whereas, logistic regression is trying to maximize the probability over all.

3 Appendix

List of files attached and description:

- **sv2fr-tprob.txt** - file containing transition probabilities for K = 6
- **sv2fr-eprob.txt** - file containing emission probabilities for K = 6
- **sv2fr-tprob-smoothed.txt** - file containing the smoothed transition probabilities with $\beta = 1$
- **sv2fr-eprob-smoothed.txt** - file containing the smoothed emisison probabilities with $\alpha = 1$
- **sv2fr-viterbi.txt** - file containing the decoding on test set (performed using log-transformed smoothed probabilities with $\alpha = 1$ and $\beta = 1$) in the same format as trn.pos
- **sv2fr-viterbi-tuned.txt** - file containing decoding on test set (performed using log-transformed smoothed probabilities with $\alpha = 0.001$ and $\beta = 0.001$) in the same format as trn.pos
- **sv2fr_viterbi.py** - implementation of Viterbi algorithm
- **sv2fr_utils.py** - utility functions used during implementation
- **sv2fr_1.1.py** - implementation for problem 1.1

- **sv2fr_1.2.py** - implementation for problem 1.2
- **util.py** - the util file for crf modified to suit python3
- **crf.py** - the modified implementation of crf for problem 2