22417.202210 Homework 6 - Expectation Maximization

Pankaj Kumar Jatav

TOTAL POINTS

19 / 19

QUESTION 1

1 q1 **2 / 2**

✓ **- 0 pts** *Correct*

- 1.5 pts Off by two orders of magnitude.

QUESTION 2

2 q2 **2 / 2**

✓ **- 0 pts** *Correct*

- 0.5 pts Off but still same order of magnitude.

QUESTION 3

3 q3 **10 / 10**

✓ **- 0 pts** *Correct*

- 1 pts Similar to correct log-likelihood

- 2 pts Seems correct, but miss some parts of answers. Init log-likelihood starts from ~ -357642

- 5 pts Miss most part of answers. Init log-likelihood starts from ~ -357642

QUESTION 4

4 q4 **3 / 3**

✓ **- 0 pts** *Correct*

- 0.5 pts Similar to correct answers

- 1 pts A bit off from correct answers

QUESTION 5

5 q5 **1 / 1**

✓ **- 0 pts** *Correct*

- 0.5 pts Incorrect

QUESTION 6

6 q6 **1 / 1**

✓ **- 0 pts** *Correct*

- 0.5 pts Far from correct answers. BLEU Yours: 0.0615

QUESTION 7

7 Collaboration Question **0 / 0**

✓ **- 0 pts** *Correct*

HOMEWORK 6
UNSUPERVISED LEARNING: THE EM ALGORITHM[1]

CS 688 MACHINE LEARNING (SPRING 2022)
https://nlp.cs.gmu.edu/course/cs688-spring22/

OUT: April 28, 2022
DUE: May 6, 2022

Your name: Pankaj Kumar Jatav

Your GID: G01338769

[1] Compiled on Saturday 7th May, 2022 at 03:59

1 Written Questions [20 pts]

Make sure to download the assignment materials from the class page.

Intro

In 2005, a blog post[2] went viral that showed a bootleg copy of Revenge of the Sith with its Chinese version translated (apparently by low-quality machine translation) back into an English movie called Backstroke of the West. Can you do better?

Setup Download the assignment materials from the class page. It contains the following files:

- `train.zh-en`: training data (Chinese-English)
- `train.zh`: training data (Chinese side)
- `train.en`: training data (English side)
- `test.zh`: test data (Chinese side): don't peak!
- `test.en`: test data (English side): don't peak!
- `backstroke.en`: *Backstroke of the West*
- `bleu.py`: evaluation script for translation
- `translate.py`: IBM Model 1 decoder
- `lm.py`: Language Model used by translate.py

The training data is Star Wars Episodes 1, 2, and 4–6. The test data is Episode 3.

You may write code in any language you choose. You may reuse any code you've used in previous homework assignments, or even the solution or another student's code as long as you cite properly.

2 These are your first steps (4 credits total)

Write code to read in `train.zh-en`. It contains one sentence pair per line, with the Chinese and English versions separated by a tab. Both sides have been tokenized (i.e. splitting "`hello!`" to two words: `hello` and `!`. This means you can easily get the tokens of a sentence by calling the `str.split()` python function. Below, we assume that a fake word NULL is prepended to every English sentence, so that $e_0 =$ NULL (so you'll need to write code to do exactly that).

Next, write code to create the data structure(s) to store the model parameters $t(f \mid e)$ and initialize them all to uniform. **Important: You only need a $t(f \mid e)$ for every Chinese word f and English word e that occur in the same line. If f and e never occur in the same line, don't create $t(f \mid e)$.** This will save you a lot of memory and time.

1. (2 points) What is the size of the model? (as in, how many model parameters are there?)

Model Size
1,75,513

2. (2 points) What would the size of the model be if we took all possible pairs of English-Chinese words into account?

[2] https://web.archive.org/web/20170115091456/http://winterson.com/2009/01/episode-iii-backstroke-of-west-redux.html

Model Size

2,13,10,282

3 Join me, and I will complete your training (13 credits total)

E-step Write code to perform the E step. As described in the slides, for each sentence pair and for each Chinese position $j = 1, \ldots, m$ and English position $i = 0, \ldots, l$, do:

$$c(f_j, e_i) \leftarrow c(f_j, e_i) + \frac{t(f_j \mid e_i)}{\Sigma_{i'=0}^{l} t(f_j \mid e_{i'})}.$$

M-step Write code to perform the M step. As described in the slides, for each Chinese word type f and English word type e (including NULL), do:

$$t(f \mid e) \leftarrow \frac{c(f, e)}{\Sigma_{f'} c(f', e)}.$$

3. (10 points) **Training** Train the model on the training data. Report the total (natural) log-likelihood of the data after each pass through the data.

$$\text{log-likelihood} = \sum_{(\mathbf{f},\mathbf{e})\text{in data}} \log P(\mathbf{f} \mid \mathbf{e})$$

$$P(\mathbf{f} \mid \mathbf{e}) = \frac{1}{100} \times \prod_{j=1}^{m} \frac{1}{l+1} \left(\sum_{i=0}^{l} t(f_j \mid e_i) \right).$$

It should increase every time and eventually get better than -152000 (you can stop when that happens, and just leave the cells blank below).[3]

[3] It happens within 10 iterations in my implementation.

Model Size

Iteration	Log-Likelihood
Init Log-Likelihood:	-308242.87179902365
Iteration 0	-160652.4129366121
Iteration 1	-146405.10858173514
Iteration 2	
Iteration 3	
Iteration 4	
Iteration 5	
Iteration 6	
Iteration 7	
Iteration 8	
Iteration 9	
Iteration 10	
Iteration 11	
Iteration 12	
Iteration 13	
Iteration 14	

4. (3 points) After training, for each English word e in {jedi, force, droid, sith, lightsabre}, for each of the five Chinese words f with the highest $t(fe)$, report both f and $t(f \mid e)$. The top translations should be: , , , , and .

Probabilities

Parameter	Probability
t(—jedi)	0.4236256003066005
t(—force)	0.38666328948131196
t(—droid)	0.47579858368794126
t(—sith)	0.3562116821228999
t(—lightsabre)	0.27249817570312096

4 Now witness the power of this fully operational translation system (2 credits total)

In this part, you'll use the provided Model 1 decoder to try to translate Episode 3 better than Backstroke of the West.

5. (1 point) Write code to dump the word-translation probabilities in a text file, named `ttable.txt` (translation table), in the following format:

```
captain         0.7648140965086824
captain     2.1415989262858237e-11
captain        2.5532871116116466e-09
captain        2.1644056596004747e-09
captain     1.1592984773140098e-07
NULL        3.7303483172036556e-20
NULL     0.04330894875722255
NULL        7.859239306612376e-14
NULL ? 3.5071699071940116e-05
NULL        3.890691080886856e-19
```

The order of the lines does not matter.

Translate Episode 3 (`test.zh`). The decoder should be run (in your terminal/command prompt) like this:

```
translate.py ttable.txt train.en test.zh
```

Translations are written to stdout. Here, show the output on lines 475–492.

Output

```
i told you to the power of the sith knows of a tragedy darth
no
i ’m trying to the jedi is not to tell you
it’s a sith the swamps ...
he is the force ... as powerful as a matter of the sith maul , two ... and the controls to hate
life
he is the right side of the force will be able to know the common
he is to be true
what ?
the dark side of the force , i think i had a lot of special flutter
what happened to him
he is a dangerous situation , yes , sir , the one you wish
he ’s the power
but he is the power
he ’s all the difference apprentice
he ’s apprentice ... and that ?
wonderful
but he could to rescue anybody
```

6. (1 point) To evaluate translations accuracy using the BLEU metric, store the translations in a text file and run (in your terminal/command prompt):

```
bleu.py your_translations test.en
```

The score is between 0 and 1 (higher is better). What BLEU score does Backstroke of the West get? What does your system get? Your system should do better than Backstroke of the West.

BLEU Blackstroke

BLEU: 0.04346211595059747

BLEU Yours

BLEU: 0.05603082283491989

e

5 Collaboration Questions

After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found in the syllabus.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details.
3. Did you find or come across code that implements any part of this assignment? If so, include full details.

Your Answer