Report **Jay Priyadarshi**
Date: 05/10/2016

Goal

Given a corrupted text, reconstruct the original English Text. The cipher text is a columnar transposition cipher with 10 rows each containing 50 characters.

Example

Corrupted text:

heeemirhcletshlohttwhsnpuesecetottipoenadeamtoaexd
tnreaaahninsnootottnhiaihkvnnhimttwmtieadnrisegena
raomnnaytihhgchawbsindcasveedreeipeeenngcethcltswi
rrginhnhrllteantfttpaiashneshoeiehrbettshofodiroho
eaerigtolelyeaarufakoglceesoifwsolsbshsdgoosnkbata
nebtrdpueroribgffbhmsmcaeutaoseprrerprttoocioonbah
aawituendgiriygswarlaernplnotseghydtssttwathgenusu
eotsruilahtnridseaicennoasadnefwhteeenfltcostyslhs
aeoaaebhabmtrnoniwtarwetgnfamthsoryhoedemhmnebspti
mgreironluhrdleraowlniykwafluisreaeecdtehcenpolstn

Original English text:

thepanelisalsoexpectedtorecommendthatthewhitehouse
theiraniangovernmenthasmaintainedthatitknowsnothin
bymatchingthelowestpriceandenhancingservicehewasde
thisfallherneighborhoodinthenortheasternpartofthis
forcollegebasketballfansthisisasgoodaweekasyourego
butaccordingtobarbershopproprietorsthenumberoffema
anintriguingnewstudysuggeststhatwhatreallydrawspeo
alsoonthursdaythelatestwinnersofthelifesciencesand
whatmembersofbothpartiesbemoanedmorethananythingwa
onekeyhurdleforteslainproducingthenewsmallercarwil

Data

The Open American National Corpus(OANC) was used to calculate bi-gram and tri-gram statistics for English Language. OANC is a roughly 15 million word corpus. The OANC corpus is available here: http://www.anc.org/data/oanc/download/

Evaluation

Bi-gram precision can quantify how close the produced result is from the actual answer, which in this case is the actual English text. The approach is similar to how BLEU evaluates the candidate and reference solutions. There are 49 bi-grams in each row and as there are 10 rows, we have a total of 490 bi-grams. For evaluating the result, consider the text row-by-row and whenever a bi-gram from produced result row matches a bi-gram from the actual result row, there is a match. Total occurrence of all bi-grams are calculated. This count for a bi-gram is then modified by considering the minimum value of the bi-gram count in produced result row and the bi-gram count in actual result row.

$$Bigram\ precision = \frac{x}{49 * 10}$$

where x is total bi-grams in the produced output which are also present in the actual answer.

Method

The following objective function was defined and the aim of the search algorithm was to minimize this function.

$$Objective\ funciton_{trigram} = \sum_{i=0}^{10} \sum_{j=0}^{48} -\log P(text_{ij+2} | text_{ij}\ text_{ij+1})$$

$$Objective\ funciton_{bigram} = \sum_{i=0}^{10} \sum_{j=0}^{49} -\log P(text_{ij+1} | text_{ij})$$

where $text_{ij}$ is the character present at row i and column j in the text.

While calculating bi-gram and tri-gram probabilities, add-one smoothing was used to handle the cases when a bi-gram/tri-gram was never seen in the training. While calculating bi-gram/tri-gram probabilities, the spaces between the words in the corpus were not considered as the given input cipher text and the plain text does not have any spaces between the words.

Tri-gram probability was used because it turned out to be a better predictor of the text. The following table shows the value of objective function when bi-grams and tri-grams were used on the example text given in the Example section. Here the **gold** solution is the actual answer, **random** solution is a random permutation of the numbers to show that it does worse on the objective function. 10x50 case denotes that all 10 rows, each with 50 characters were used. 10x10 case denotes that all 10 rows were used but only the first 10 numbers in the key were found.

10x50	Bi-gram	Tri-gram
Gold	1271.8	1113.7
Greedy	1334.3	1285.1
Random	1532.1	1778.6

10x10	Bi-gram	Tri-gram
Gold	234.1	187.6
Exhaustive	234.1	187.6
Random	292.2	291.1

The following approaches were used:

i) **Greedy Search**: choosing the next best column given the current column did not yield the required result

ii) **Simulated Annealing** with temperature reduction factor $\alpha = 0.9$ and a new solution was accepted if acceptance probability was more than 0.5. The new solution at each iteration was produced by randomly swapping two columns. This approach did not yield the deciphered text

iii) **Tabu search**: this approach is similar to the Simulated Annealing algorithm, with an addition of tabu list. At each iteration the algorithm tries to replaces the worst key from the tabu list with the best key found in that iteration. The new solutions are produced by randomly swapping two columns of the key. Tabu search did not produced the actual text as well

iv) **Beam Search** with 20000 beam size generated a partial key of size 32. Another beam search with 100 beam size was used to find the remaining 18 column orderings. The aim of the second beam search was to find the remaining 18 column numbers which together with the already found 32 column ordering reduces the objective function's value. This approach worked and generated the actual English text from the given corrupted text

Result

The **Beam Search** algorithm was used on the following cipher text:

dtjmeoftumhbhstehresweseeoatearthkteoouietohfyetri
ehsinaeinctnecfdbouunuetomltoimtonsrsihyeognrcfesi
heldoehwaenwkleomnaegnonwefrimimynhenopngotiwymltt
ikthgatysiowiitdeedrofeeyexonroaneitthtntsnhleofhc
oslmnemjrtoebeftegernesioiweayksphluethuaraeahafth
nfiehtuohugiandettheeoartttrbeybuiohvsatsdousnwona
sltaeytgnnhaatgoehtkssulttroleulenaofiolgtwanlleli
wmlptepmsttoehwchhieiavoehikrnwthaterecnltewacwele
icnoovslantrebdweoifseundstntewanhioottaeataairetds
cggoarrhsdewarhtipntiiretnowrddcahipwtgneoeboeldwa

After the running both beam searches the following text was produced:

theremustbesomewayoutofheresaidthejokertothethieft
toomuchconfusionicantgetnoreliefbusinessmentheydri
mywineplowmendigmyearthnoneofthemalongthelineknoww
anyofitisworthnoreasontogetexcitedthethiefhekindly
spoketherearemanyhereamonguswhofeelthatlifeisbutaj
butyouandiwevebeenthroughthatandthisisnotourfateso
letusnottalkfalselynowthehourisgettinglateallalong
thewatchtowerprinceskepttheviewwhileallthewomencam
andwentbarefootservantstoooutsideinthedistanceawil
catdidgrowltworiderswereapproachingthewindbegantoh

Analysis

The Bi-gram Precision values for the beam search are given below:

Search	Bi-gram Precision
1^{st} beam search	0.92
2^{nd} beam search	1

The search problem is similar to the Traveling Salesman problem (TSP) and the task of finding a tour of minimum overall cost can be thought in terms of finding a permutation of columns which results in minimum objective function value. TSP belongs to the class of NP-Complete problems. Hence, all the algorithms trying to solve this problem usually try to find the best approximate solution. In the above beam search approach, the first beam search was able to find this approximate solution. It was able to recover 451 bi-grams out of 490 bi-grams present in the actual solution. The second beam search takes the part of the approximate solution (the part which makes sense) and tries to build the whole key on top of that.