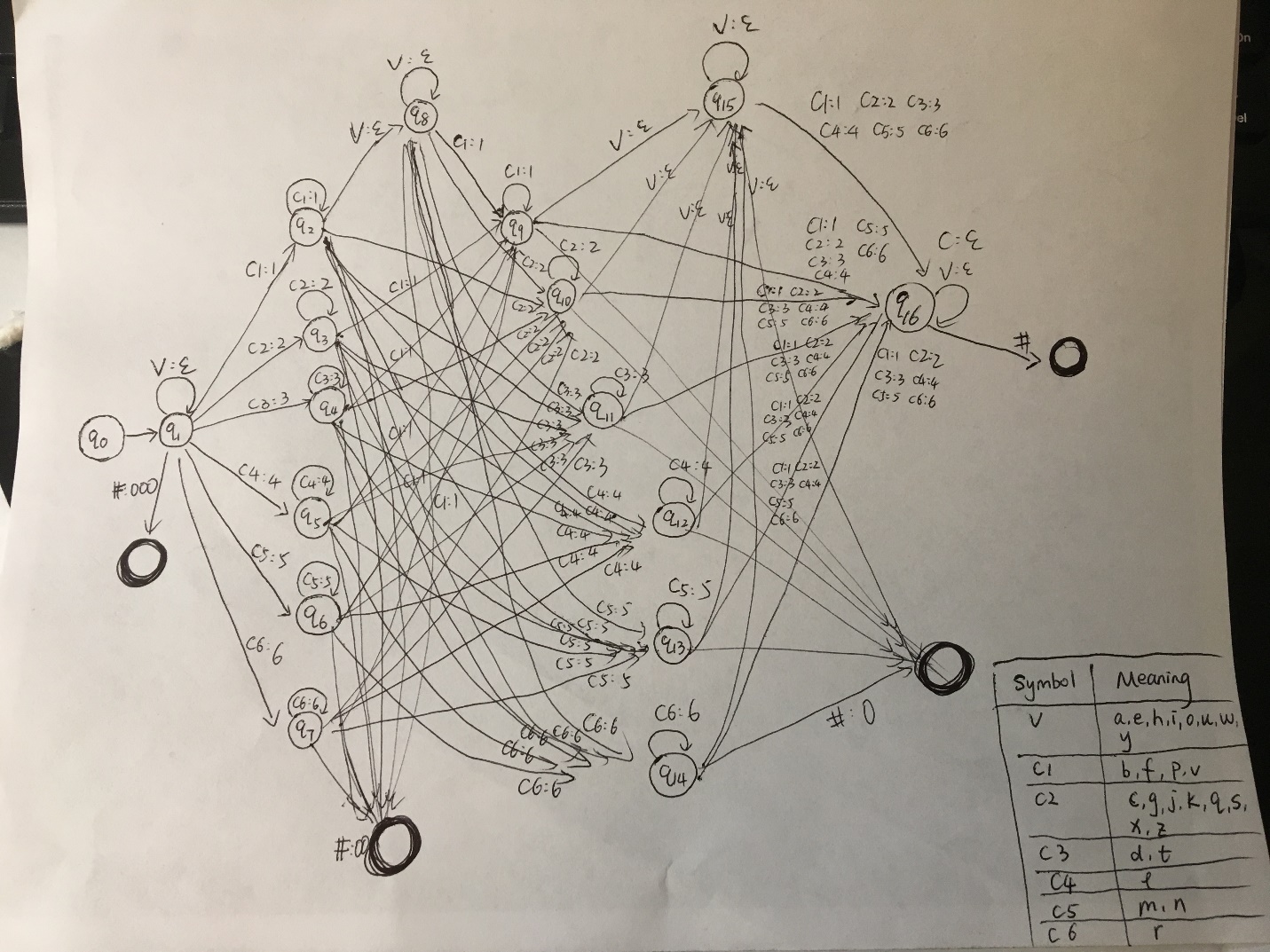
**Homework2\_ReadMe**

Mengxi Nie eva\_happylife@gwu.edu

Q1:

Here are some symbols.

|  |  |
| --- | --- |
| Symbol | Meaning |
| V | a, e, h, i, o, u, w, y |
| C1 | b, f, p, v |
| C2 | c, g, j, k, q, s, x, z |
| C3 | d, t |
| C4 | l |
| C5 | m, n |
| C6 | r |



Q2:

The code and result are in the fold named “Q2”

I use perl5 and the operation system is ubuntu14(64 bit)

Running command: chmod 755 letterLangId.pl

./letterLangId.pl

The output is in the file named “BigramLetterLangId.out”

Q3:

The code and result are in the fold named “Q3”

I use perl5 and the operation system is ubuntu14(64 bit)

Running command: chmod 755 BigramWordLangId-AO.pl

./BigramWordLangId-AO.pl

The output is in the file named “BigramWordLangId-AO.out”

Q4:

The code and result are in the fold named “Q4”

I use perl5 and the operation system is ubuntu14(64 bit)

Running command: chmod 755 BigramWordLangId-GT.pl

./BigramWordLangId-GT.pl

The output is in the file named “BigramWordLangId-GT.out”

Q5:

The code and result are in the fold named “Q5”

I use perl5 and the operation system is ubuntu14(64 bit)

Running command: chmod 755 TrigramWordLangId-KBO.pl

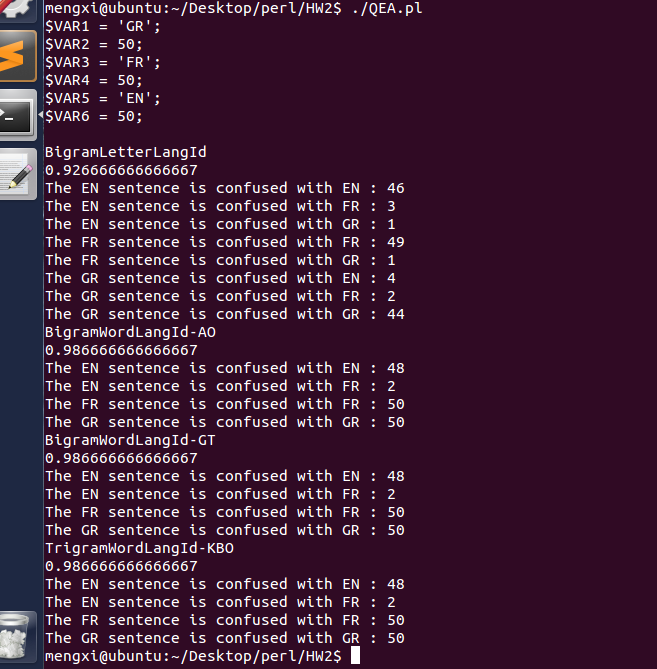
./TrigramWordLangId-KBO.pl

The output is in the file named “TrigramWordLangId-KBO.out”

Q6:

I use the perl5 to do the error analysis and the code in the file named “QEA.pl”. Besides, the operation system is ubuntu14(64 bit).

The result is as follows:



* **Accuracy**

For “BigramLetterLangID”:

The sum of wrong choices is 3 + 1 + 1 + 4 + 2 = 11.

The sum of sentences is 150.

The accuracy is (150 – 11) / 150 = 92.67%, as the result shown in the picture.

For “BigramLetterLangID-AO”:

The sum of wrong choices is 2.

The sum of sentences is 150.

The accuracy is (150 – 11) / 150 = 98.67%, as the result shown in the picture.

For “BigramLetterLangID-GT”:

The sum of wrong choices is 2.

The sum of sentences is 150.

The accuracy is (150 – 11) / 150 = 98.67%, as the result shown in the picture.

For “BigramLetterLangID-KBO”:

The sum of wrong choices is 2.

The sum of sentences is 150.

The accuracy is (150 – 11) / 150 = 98.67%, as the result shown in the picture.

|  |  |
| --- | --- |
| **Experimental Condition** | **Overall Accuracy %** |
| BigramLetterLangID | 92.67% |
| BigramLetterLangID-AO | 98.67% |
| BigramLetterLangID-GT | 98.67% |
| BigramLetterLangID-KBO | 98.67% |

* **Confusion Matrix**

For “BigramLetterLangID”:

In English, the sum of wrong choices is 4.

In French, the sum of wrong choices is 1.

In German, the sum of wrong choices is 6.

There are 3 EN sentences are confused with FR and the percentage is 3 / 4 = 75%.

There is 1 EN sentences are confused with GR and the percentage is 1 / 4 = 25%.

There is 1 FR sentence is confused with GR and the percentage is 1 / 1 = 100%.

There are 4 GR sentences are confused with EN and the percentage is 4 / 6 = 66.7%.

There are 2 GR sentences are confused with FR and the percentage is 2 / 6 = 33.3%.

|  |  |  |  |
| --- | --- | --- | --- |
| **BigramLetterLangID** | **EN** | **FR** | **GR** |
| **EN** |  |  | 66.7% |
| **FR** | 75% |  | 33.3% |
| **GR** | 25% | 100% |  |

For “BigramLetterLangID-AO”:

In English, the sum of wrong choices is 2.

There are 2 EN sentences are confused with FR and the percentage is 2 / 2 = 100%.

|  |  |  |  |
| --- | --- | --- | --- |
| **BigramLetterLangID-AO** | **EN** | **FR** | **GR** |
| **EN** |  |  |  |
| **FR** | 100% |  |  |
| **GR** |  |  |  |

For “BigramLetterLangID-GT”:

In English, the sum of wrong choices is 2.

There are 2 EN sentences are confused with FR and the percentage is 2 / 2 = 100%.

|  |  |  |  |
| --- | --- | --- | --- |
| **BigramLetterLangID-GT** | **EN** | **FR** | **GR** |
| **EN** |  |  |  |
| **FR** | 100% |  |  |
| **GR** |  |  |  |

For “BigramLetterLangID-KBO”:

In English, the sum of wrong choices is 2.

There are 2 EN sentences are confused with FR and the percentage is 2 / 2 = 100%.

|  |  |  |  |
| --- | --- | --- | --- |
| **BigramLetterLangID-KBO** | **EN** | **FR** | **GR** |
| **EN** |  |  |  |
| **FR** | 100% |  |  |
| **GR** |  |  |  |

* **Perplexity of Test Set**

|  |  |  |  |
| --- | --- | --- | --- |
| **Experimental Condition** | **EN** | **FR** | **GR** |
| **BigramLetterLangID** | 3.54186570274e+13 | 3.29804661082187e+13 | 4.82475250112298e+13 |
| **BigramLetterLangID-AO** | 3.43170673211e+12 | 3.99804661082187e+11 | 3.61999613720754e+11 |
| **BigramLetterLangID-GT** | 3.29600089683e+12 | 3.86153675671303e+11 | 3.33842836130829e+11 |
| **BigramLetterLangID-KBO** | 3.23552399825e+12 | 3.64393851958562e+11 | 3.28391846554303e+11 |

Because I use the *log* probability and when comes to a new word in the test, I use add (-999999999) to deal with this situation. Therefore, the perplexity of each experimental condition is very large. These models have a good performance at judging a German sentence. BigramLetterLangID-AO, BigramLetterLangID-GT and BigramLetterLangID-KBO have a bad performance when comes to judging a sentence is an English sentence.