DIALOUGE SYSTEMS AND NATURAL LANGUAGE INTERFACES :HW2

Ajantha Ramineni, Poorna Satya Sainath Kanamathareddy, Sreeharsha Poluru

Introduction

The HW2 takes the one act play titled “St. Valentine’s Day” by Annie Eliot,extracted from one act plays’ website

`(http://www.one-act-plays.com/comedies/st_valentines_day.html)`

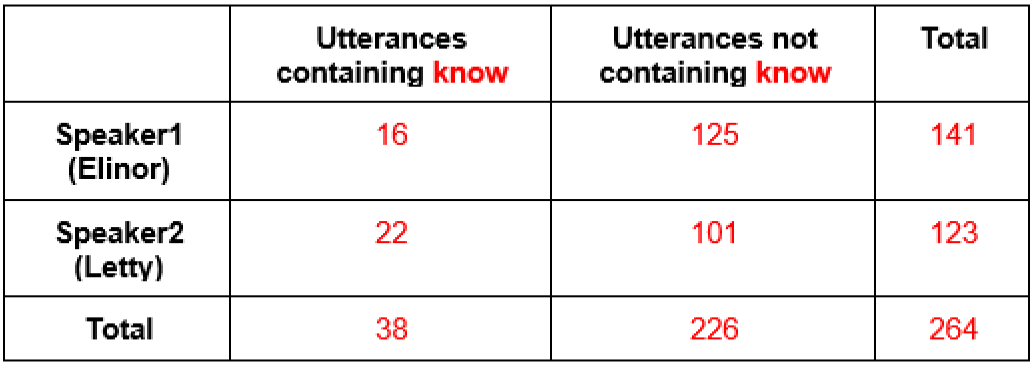
as the input. The play is a two actor play with the characters Elinor and Letty who are Speaker 1 and Speaker 2 respectively. We divided the corpus into utterances based on the turns. Each sentence is considered as a single utterance, this includes all the short poems (which are a part of the corpus).

Note: Throughout the document ***Speaker1 is Elinor*** *and* ***Speaker2 is Letty***

1 Contingency Tables

(a) The contingency table has been constructed by selecting the utterances containing the word *’know’*.The word *’know’* has been chosen in particular because of the frequency with which it occurs in the corpus and this may help us analyze the speakers’ dialogues.The values in the cells of the table i.e, number of utterances containing *’know’* in case of Speaker1 and Speaker2 etc,. have been calculated by using a Python program.

The following is the contingency table thus obtained.



	Utterances containing know	Utterances not containing know	Total
Speaker1 (Elinor)	16	125	141
Speaker2 (Letty)	22	101	123
Total	38	226	264

Contingency Table for Speaker1 and Speaker2

(b) The probabilities were calculated with the help of the values obtained in the contingency table above. The probabilities were calculated by using the conditional probability.

The calculated probabilities (rounded off to 2 decimals) are as follows:

(a) P(u is from Speaker1) = 0.53

(b) P(u is from Speaker2) = 0.47

(c) P(u contains know | u is from Speaker1) = 0.11

(d) P(u contains know | u is form Speaker2) = 0.18

(e) P(u being from Speaker1 | u contains know) = 0.42

(f) P(u being from Speaker1 | u does not contain know) = 0.55

(g) P(u being from Speaker2 | u contains know) = 0.58

2 Sequence Analysis

(a) The most common token in the corpus(including the stop words) is ***I*** , **which occurred 248 times in the corpus.**

The most common token(including the stop words) for Speaker1 is ***I***, **which occurred 127 times,** followed by ***of*****, which occurred 76 times.**

The most common token(including the stop words) for Speaker2 is ***I***, **which occurred 121 times,** followed by ***to*****, which occurred 77 times.**

(b) The most common bi-gram in the corpus is **(’I’, ”don’t”), which occurred 22 times.**

(c) The most common trigram in the corpus is **(’I’, ”don’t”, ’know’), which has a count of 11.**

(d) The probability that the next dialogue turn will be Speaker1, given that Speaker1 just finished a dialogue turn is **0.73**.

i.e, P(Speaker2’s turn | Speaker1 just finished dialogue turn) = 0.73

3 Descriptive Statistics

(a) The median number of words per utterance for Speaker1 in the data is **11**.

The median number of words per utterance for Speaker2 in the data is **16**.

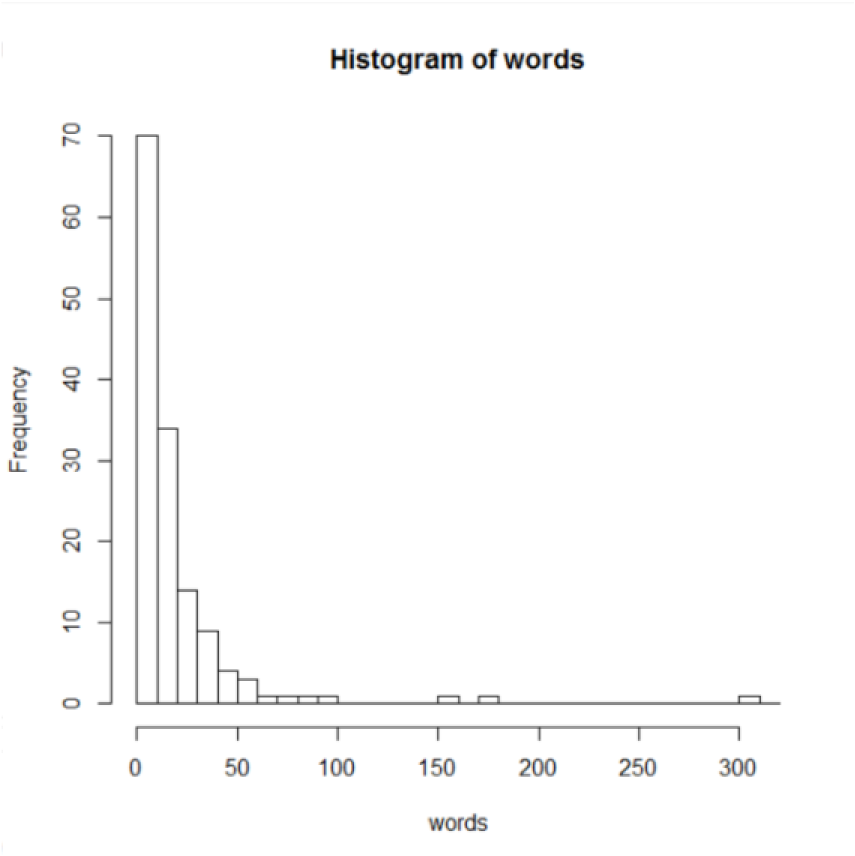
(b) The mode of words per utterance for Speaker1 is **7**.

The mode of words per utterance for Speaker2 is **10,11**.

(c) The standard deviation for number of modes per utterance for Speaker1 is **34.4362**.

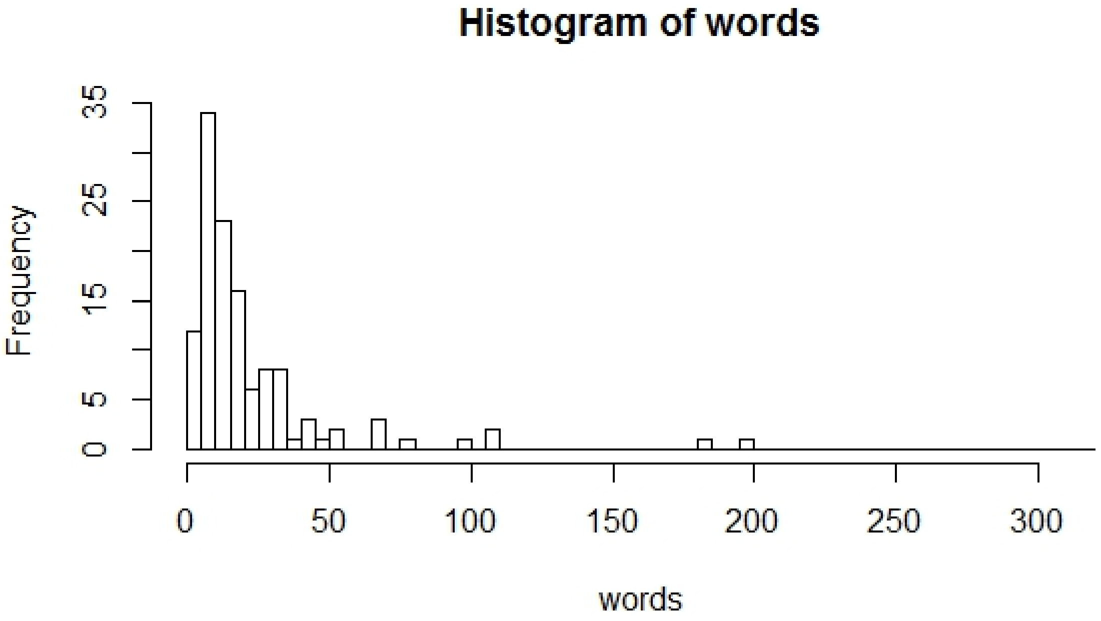
The standard deviation for number of words per utterance for Speaker2 is **31.47944**.

(d) **The histogram for number of words per utterance for Speaker1 is as follows:**



By observing the histogram for Speaker1, we can infer the that the Speaker1 uttered maximum number of dialogues which had a word count in the range (0-10). Further, the utterances with the word count in the range (10-20) occur with a frequency of 35, which in fact is the second highest.Also, the highest number of words uttered in a single turn by Speaker1 are in the range(300-305), which occurs for a single turn. Apart, form these, it may be observed that most of the utterances contained about 15-20 in case of Speaker1.

The histogram for number of words per utterance for Speaker2 is as follows:



From the histogram of Speaker2, it can be observed that each utterance of contained an average of 15 to 20 words. Highest number of utterances have a word count in the range(10-20), which is followed by utterances which have the word count in the range (20-30). Very few Speaker2 dialogues have a word count greater than 100, in fact, there are 2 such utterances(which can be seen in the histogram).

Also, by observing the two histograms, it can be concluded that the number of utterances of Speaker1 are more in number when compared with Speaker2.Among the two, Speaker1 has the utterance with the highest number of words.In general, per turn or utterance, Speaker2 speaks more words than Speaker1.

4 Hypothesis Testing

(a) The statistical test most suitable to determine whether presence of know depends on Speaker, is the Chi - square test.Though,both the Fischer and Chi-square tests are used to test the hypothesis for categorical data, Fischer is used for those categorical values whose sample size is small or whose contingency cells' value is less than 5.

The sample size of the input corpus can be without doubt categorized under the large sample size. Hence, the Chi-square test is the more appropriate between the two.

The null hypothesis and the alternate hypothesis is considered as follows:

H_0 : "The probability of know in an utterance is not significantly different conditioned upon who the speaker was."

H_A : "The probability of know occurring in an utterance is different conditioned upon who the speaker was."

The values from the contingency table above is to be given as the input to the function, which performs the Chi-square test.The values must be first converted to the matrix form before they are passed as the input.

Now, the matrix form containing the contingency table values is given as the input to the function in R.The whole sequence of steps is stored as a R script file.This script file is run to perform the test.

The script is given below:

```
myfunction<-function()
{
  data_m<-matrix(c(16,22,125,101), nrow=2)
  chisq.test(data_m,correct=FALSE)
}
```

Once the script is run, the output values along with the p-value and Chi-square values are displayed. The output is as follows:

```
Pearson's Chi-squared test

data:  data_m
X-squared = 2.2794, df = 1, p-value = 0.1311
```

The threshold value for p is taken to be 0.05. From the result, we observe that the value of p is 0.1311, which, is greater than the threshold value of p.

Hence we may accept the null hypothesis H_0.

(b) The given alternate hypothesis is as follows: H_A :Speaker 1 said more words in this corpus than Speaker 2.

But the total word count of the Speaker1 and total word count of Speaker2 come under the descriptive statistics. The difference between the descriptive statistics and inferential statistics is that the differential statics provides simple summaries of data, while the inferential statistics aims to learn about the population.

Hence no test is required to perform the comparison between total words of Speaker1 and Speaker2.

To test this, we may use the total word count which we found previously by using the Python program in the HW1.

For this corpus the total word count for Speaker1 is 2938 and for Speaker2 it is 2793. From this data we may infer that the Speaker1 spoke more words than Speaker2 in this corpus.

(c) The null hypothesis(H_0) and the alternative hypothesis(H_A) is as given below:

H_0 : The difference of mean utterances of Speaker1 and Speaker2 is zero.

H_A : The difference of mean utterances of Speaker1 and Speaker2 is greater than zero.

We use the Welch t-test to test the hypothesis.This test was chosen in particular because of the following reasons:

- the test deals with comparing the means of two data.
- the test is appropriate for the data which have different sizes.The input data has 2 different sizes - Speaker1 has 141 utterances while the Speaker2 has 123 utterances.

A Python program calculates and stores the count of words for each utterance for Speaker1 and Speaker2 in the text files $output_speaker1.txt$ and $output_speaker.txt$ respectively. This list of values is then passed as the input to the function to perform the Welch's t-test. The script to perform the test is as follows:

```
t_function<-function()
{
  obj1<-read.table("output_speaker1.txt")
  obj2<-read.table("output_speaker2.txt")
  t.test(obj1, obj2)
}
```

The output of the test is as follows:

```
Welch Two Sample t-test

data:  obj1 and obj2
t = -0.475, df = 261.93, p-value = 0.6352
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.624099  5.883223
sample estimates:
mean of x mean of y
 20.83688  22.70732
```

The threshold value for the p-value is taken as 0.05.As observed in the output, the value of p is 0.2995, which is greater than the threshold value.

Hence we may accept the null hypothesis.This indicates that the average words per utterance of Speaker1 and Speaker2 (i.e means of the two data).

5 Practice with Definitions

(a) **There are 19 tokens in the given quote**, without including the punctuation.The following is the list of tokens:

- Good
- Night
- good
- night
- Parting
- is
- such
- sweet
- sorrow
- that
- I
- shall
- say
- good
- night
- till
- it
- be

- morrow

(b) There are **15 distinct word types in the quote, when punctuation are not considered**. **If the punctuation be considered, the total distinct word types would be 18**

(c) ***'Say'*** is the verb, which remains unchanged when stemmed.

Explanation:

Original form = say

Stem of the word 'say' is say

Hence no change.

(d) The verb ***'parting'*** can be taken as the example of the inflectional morphology.

Explanation:

Verb = Parting

Stem = Part

By adding -ing to the stem(i.e,part) it is inflicted, thus, making the stem a *-ing participle(i.e,parting).*

(e) The verb ***be*** is the example of the derivational morphology in the given quotation.The derived verb is ***is***.

That is, the verb ***be***is morphed into the verb ***is***, which is a 3rd person singular present indicative of ***be*** (definition source: $http://www.dictionary.com/browse/is?s = t$).