Faculty of Engineering &Technology
Electrical &Computer Engineering Department

NATURAL LANGUAGE PROCESSING (NLP) AND INFORMATION RETRIEVAL

Major Assignment #1

Arabic Resources Collection

Arabic Ambiguous Entities

Prepared For

Dr. Adnan Yahya

By

Name: Rand Sandouka ID No.:1141462

Name: Ruba Hamad ID No.: 1140251

Date: 19/4/2019

Abstract:

Machines too like humans are capable of learning once they see relevant data. But where they vary from humans is the amount of data they need to learn from. You need to feed your machines with enough data in order for them to do anything useful for you. The data-set in some cases must be very large to enable sufficient learning for the model to be generated; therefore, the experiment will be transformed to a data collection task.

Introduction:

- **Data Set and its importance:**

A data set is a collection of data which corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.

Several characteristics define a data set's structure and properties. These include the number and types of the attributes or variables, and various statistical measures applicable to them. [1]

- **Kappa measure:**

A common measure for agreement between judges is the kappa statistic. It is designed for categorical judgments and corrects a simple agreement rate for the rate of chance agreement.

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Where:

P(A): is the proportion of the times the judges agreed.

P(E): is the proportion of the times they would be expected to agree by chance.[2]

Procedures:

1. First of all, we have chosen 20 Arabic words that have at least two meanings, which are called ambiguous entities, e.g.,عشاء، سهم، لحام .
2. We have collected 50 document links for each word with its different meanings using Bing search engine.

3. We have chosen 8 judges, with 4 judges for the first ten words and 4 judges for the rest of the words.

4. Then we distributed forms (shown in A1 & A2) among judges to fill them.

5. Lastly, we calculated Kappa.

- **Results and Conclusion:**

In this part, we calculated the Pairwise Kappa for each word with its different meanings. We filled 20 excel sheets; Since we have 20 different words, after that we applied a python code (as shown in A3) to calculate the pair wise and the average kappa as shown in appendix and got results as shown in Figure 1.

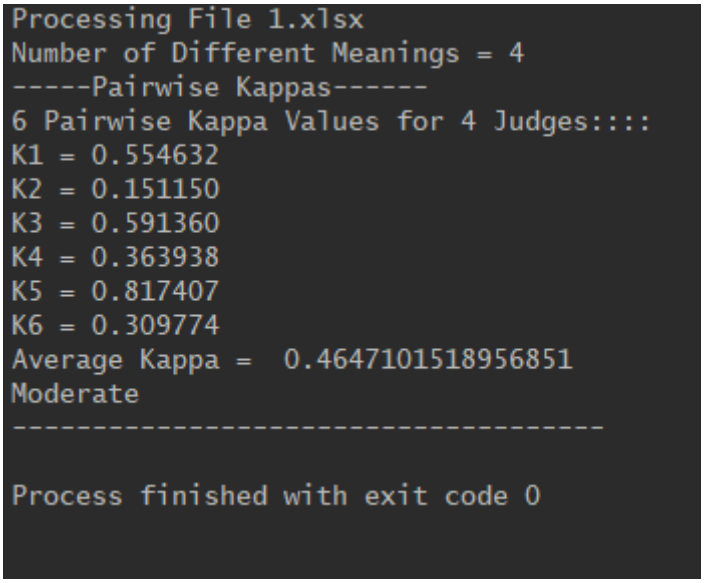

Figure 1: Kappa measure for the first word.

The table below illustrates the kappas and average kappas for four words with these pair wise of judges.

Judge 1 + Judge 2	*Judge 1 + Judge 3*
Judge 1 + Judge 4	*Judge 2 + Judge 3*
Judge2 + Judge4	*Judge3 + Judge4*

Table 1: Kappa Calculation

Pairwise Kappa	W1	W2	W3	W6
K1	0.55	0.2	0.449	0.134
K2	0.15	0.054	0.26	0.1002
K3	0.59	0.8	0.671	0.076
K4	0.36	.0.02	0.181	0.599
K5	0.817	0.303	0.638	0.088
K6	0.39	0.099	0.205	0.149
Average Kappa	**0.46 Moderate**	**0.247 Fair**	**0.4011 Moderate**	**0.191 Poor**

References:

[1] https://en.wikipedia.org/wiki/Data_set

[2] https://nlp.stanford.edu/IR-book/pdf/08eval.pdf

Appendix:

A1.

https://docs.google.com/document/d/12-wgyPIBeyn5bOvHBkCUULjE2RBoK3rKMatDkZ_Q-os/edit?usp=sharing

A2.

https://docs.google.com/document/d/1MZuwsAGOe8HKYo0QCHkToPocVLvGpwJs1aBC9sZxuDY/edit?usp=sharing

A3.

https://docs.google.com/document/d/17Lb_w0RtACVXDHOloGYkQG5a8C0i7fNOC9VPSfVwgYo/edit?usp=sharing