Faculty of Engineering & Technology
Electrical & Computer Engineering Department

Natural Language Processing – ENCS539

Resources Collecting Project

Prepared by:

Aminah Ankoush 1152161

Khulood Sabri 1150697

Supervised by:

Dr. Adnan Yahya

April 19, 2019

Introduction

In any Information retrieval system, it is required to have a collection of data to test the performance of the system. In this project, a hundred of articles were collected from different resources involving political and social topics. Then, the collected data were preprocessed for the purpose of evaluation. After that, they were distributed over a 150 judge for the evaluation process. Finally, the results of the evaluation were analyzed and the degree of agreement (Kappa Measure) were calculated.
This report discusses in details the steps mentioned earlier.

Resources Collection

A collection of 100 Arabic articles were gathered from more than 50 resources including, websites, social media and newspapers. Appendix A attached with the report includes the sites and the links of the selected articles. The diversity of resources ensures that the quality and the writing styles of articles are totally different. Moreover, the selected articles involved political and social topics with an average length of two pages in Word. It was considered to choose recent articles to catch the attention of the readers.

Data Normalization

Some factors that are unrelated to the evaluation process may affect the judges' decision. These factors include photos attached in the article, font size, color and style, the article's author, and the web or Facebook page which has published it. These factors were eliminated by extracting only the text of the articles, copying it to Word documents and using one font style to all of them. Each document was given a unique number from one to hundred for making the process of distributing and evaluating easier. And finally, they were converted to pdf format before distributing them to judges. Appendix B attached with the report contains the distributed articles.

Articles Distribution

The big challenge was neither collecting the data nor normalizing it, but getting five evaluations to each article, a total of 500 evaluations. For this purpose, about 150 judges were involved in the process of evaluation. Actually, the judges were normal people with ages range from 20 to 60 including, students, graduates, teachers and employees. The communication with the most of them was using Facebook groups (Birzeit University Students and Tanween). The judges were given about five days to end evaluation. Excel sheets attached in Appendix C were used to organize the exhausting process of distributing articles.

Evaluation Criteria

Each judge read and evaluated an average of 5 articles. The number actually is according to their desire. The evaluation for each article is done by filling a google form that contains five questions. First one for the name of the judge, second for the number of the article and the three others questions involve evaluation. The main criteria considered in the evaluation are the quality of the language, the style of writing both with classes (Excellent, good, acceptable, bad), and the ability of the writer to convey the goal of his writing with classes (yes, partially, no) . The form is attached with the report in Appendix D.

Summary of Results

Google Forms makes the process of evaluation clearer by providing a summary of evaluations and the proportion of each class for each question as shown in Figures 1,2 and 3.

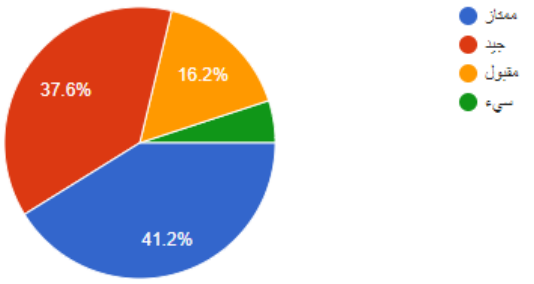

Figure 1: Language Evaluation

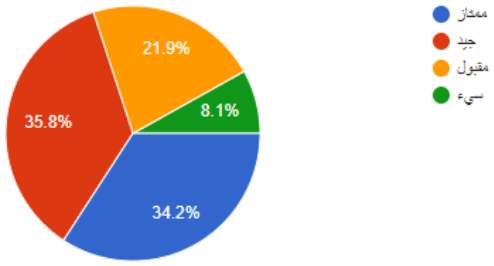

Figure 2: Style of Writing Evaluation

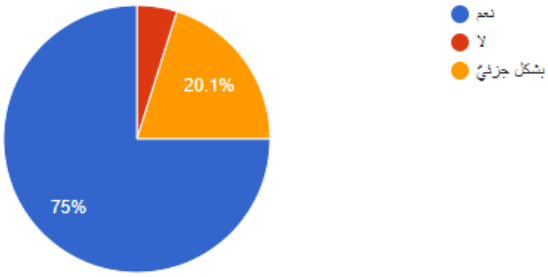

Figure 3: Goal Achievement Evaluation

Agreement Evaluation

Evaluations were retrieved as an excel file from Google Forms (Appendix E). The python code attached in Appendix F was written to extract the data and get the agreement measure. It should be noticed that the number of categories for each evaluation element was reduced. ‘Excellent’ and ‘good’ were treated as one class, and ‘acceptable’ and ‘bad’ were collapsed into one class for the language and style evaluations. For the goal conveyance evaluation, ‘No’ and ‘Partially’ categories were merged.

Fleiss measure was used for agreement evaluation. Fleiss measure is an extension of kappa with more than two raters. In this measure there is a population of raters which is large relative to the number of subjects. Each subject is evaluated by randomly chosen raters from this population. It should be mentioned that Fleiss does not care who is the rater; It is only concerned in the rates themselves.

Fleiss measure has two variations, fixed marginal and free-marginal. Marginal distributions are considered to be free when raters do not know a priori the quantities of cases that should be distributed into each category. This is the case when a rater is free to assign cases to categories with no limits on how many cases must go into each category. In our case free-marginal Fleiss was used.

In python a built-in function can be used to get free-marginal Fleiss measure:

$$statsmodels.stats.inter\_rater.fleiss\_kappa\ (table,\ method =' rand')$$

Where *table* is a two-dimensional array, with rows representing the subjects and columns representing the categories. Table[i][j] is the number of times the subject i is classified in category j.

The results of free-marginal Fleiss for the three evaluation aspects mentioned before were as follows:

- For language evaluation: 0.57
- For style evaluation: 0.47
- For goal conveyance evaluation: 0.53

The table below shows how kappa measures are usually interpreted. It can be observed that our values indicate a moderate agreement.

Table 1: Kappa Interpretation

κ	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

References

[1] https://www.statsmodels.org/dev/generated/statsmodels.stats.inter_rater.fleiss_kappa.html

[2] https://towardsdatascience.com/inter-rater-agreement-kappas-69cd8b91ff75

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/?fbclid=IwAR2xoDlYDvF85csFj113LHUbu9951Z6kyMGc5D2se97a-XEN6h086eLx2G4

[4] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.8776&rep=rep1&type=pdf