CISC839 G-# Final Report: Exploring question answering data for a specific domain

Marim Amer[1], **Mawada Amer**[2], **and Youmna Alsayed**[3]

[1]**21maem@queensu.ca**
[2]**21maem1@queensu.ca**
[3]**21yaaa@queensu.ca**

1 BACKGROUND AND OBJECTIVE

A question-answering system for the medical field is used to predict the judgment score for question pairs and predict the question type. It is a very important system that makes it easier for the patients, students, and medical staff... to know the validity of the questions asked and the type of questions.

In this project, we will use two data-sets:

- The first one is the MedQuAD Ben Abacha and Demner-Fushman (2019) data set that contains XML files and 47,457 medical question-answers, 10 columns (e.g., source, URL, Focus, Synonym, Cui, Answer, type, SemanticType, SemanticGroup, and question) and covers 37 question types (e.g., treatment, diagnosis, side effects). We will use it to answer predictive and Hypothesis test questions.
- The second one is the paired question data set, which contains four features and 2479 rows. We will use it in our model and answer a regression question.

The project covers the most common questions in the medical field, where it compares the test questions and sees what the score is for the similarity of questions.

1. What are the entailment levels for each question pair?
    - In a question entailment system, a user asks his/her question, and the system tries to find the most similar question(s) in the data-set and return their answers. So, the most important part is finding the question's equivalent in the data-set. In this task, our goal is to provide a regression system that gets two questions as input (one is the user question, and the other one i3s a question in the data-set) and predict their entailment score. The entailment score has four levels: 1: Incorrect, 2: Related, 3: Incomplete, 4: Excellent.
2. What is the qtype of questions?
    - When we want to add a new question to the main QA data set, we need to label its qtype. Also, if the system cannot provide a satisfactory answer, according to the qtype, the system can send the question to the appropriate person to answer.
3. Is the ratio of questions with qtype-coarse='Drugs' from the 'GHR' source significantly higher than this ratio for all other sources?
    - When we calculated the Drugs ratio for each source, we observed that the drugs in the 'GHR' source have a higher ratio than the ratio for all other sources. This question is important for patients, doctors, and the general public to know where is the source where we can find the most drugs? or places where the drug is available.

2 DATASET

- According to the MedQuAD dataset:
  - This data was provided by Asma Ben Abacha and Dina Demner-Fushman.
  - Created from 12 NIH websites.
  - The data collected from the medical field
- According to the paired questions data-set:
  - We collected this data-set from MedQuAD and QA-TestSet-LiveQA-Med-Qrels-2479-Answers Ben Abacha (2017).
  - The data was collected from the medical field.

2.1 Data Preprocessing

- According to the MedQuAD data-set
  1. Cleaning Qtype-Coarse, source columns
     (a) Keep ASCII + European Chars, white-space, and no digits, and remove single letter chars
     (b) Convert all white-spaces (tabs, etc.) to a single white space, and convert all words into lowercase.
     (c) Lemmatization, and TF-IDF.
- According to paired questions
  1. Clean user question and database question columns.
     (a) Keep only ASCII + European Chars and white-space, no digits, and remove single letter chars.
     (b) Convert all white-spaces (tabs, etc.) to a single white-space, and convert all words into lowercase.
     (c) Remove stop-words, punctuation, and stemming, Lemmatization.
  2. Clean column "judge score"
     (a) Remove all words in the column just we take the number'1':'Incorrect', '2': 'Related', '3': 'Incomplete', '4': 'Excellent', and removing '-'

2.2 Basic Statistics of the Dataset

1. Paired questions data-set
   (a) 4 columns (question-id, database-question, user-question, and judge-score) and 2479 rows.
   (b) It has 163 duplicated data.
   (c) We used question-id and judge-score columns, and their are integers and user-question and database-question are text.
   (d) There is unbalanced data in the judge-score column.As shown in Figure 1
2. MeQuAD data-set
   (a) 10 columns (source, URL, Focus, Synonym, Cui, Answer, type, SemanticType, Semantic-Group, and question) and 16377 rows.
   (b) 37 question types (treatment, diagnosis, side effects).
   (c) We used only Qtype-Coarse and source columns.
   (d) It has 19 null values in its focus column. As shown in Figure 1
   (e) It has 30 duplicated data.

	Source	url	Answer	qtype	Focus	Question	Qtype-Coarse
Count	16377	16377	16377	16377	16377	16377	16377
Unique	9	5476	15817	16	5125	14979	4
top	GHR	http://nihsenioriohealth.gov/breastcancer/toc.html	This condition is inherited in an autosomal re...	information	Breast Cancer	What is (are) High Blood Cholesterol ?	Undefined
freq	5430	28	348	4523	53	19	12584

Table 1. Statistics of the Dataset.

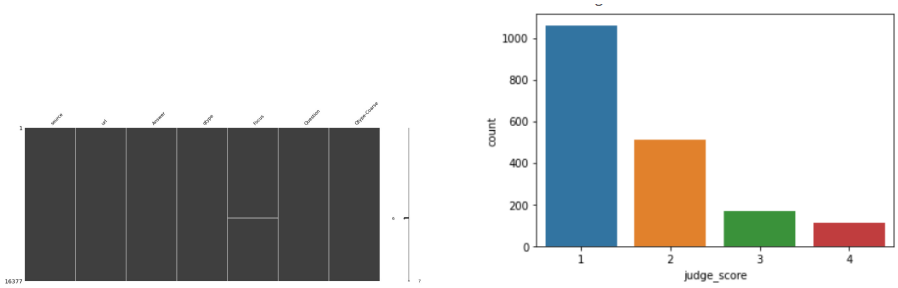

Figure 1. Left image (Null Values), and Right image (Imbalanced data)

3 ANSWERS TO THE RESEARCH QUESTIONS

3.1 Answer to question #1: What are the entailment levels for each question pair?

We split the data into 0.8 for training and 0.2 for testing, and then we solved imbalanced data by doing random over sampling. We did not encode the text because the similarity functions accept text words.

It is a **regression question**. We used six models. The best two models with the minimum error are the Decision Tree Regressor with hyper-parameters (criterion = 'gini', min samples split = 2, min samples leaf =1) with an error of 0.64, mean absolute percentage error is 0.64, and MSE = 0.98. The second model is KNN with hyper-parameters (n neighbors = 4) with an error of 0.66. The mean absolute percentage error is 0.42, and the MSE is 0.88.

3.2 Answer to question #2: What is the qtype of questions?

We solved null values, dropped duplicated rows, and cleaned the text from any stop words, punctuation, lemmatization, and... Then we split the data into 0.8 for training and 0.2 for testing data. Then do Tf-idf Vectorizer.

It is a **predictive question**. We used 4 classification models like the Decision Tree Classifier with hyper-parameters (max-depth = 9) with an accuracy of 0.94 and f1-score = 0.94 and the KNN with hyper-parameters (n-neighbors = 80) classifier with an accuracy of 0.81 and f1-score = 0.81.

The best model is the Decision Tree Classifier, and the results are accuracy = 0.94 and f1-score = 0.94. As shown in Figure 2

3.3 Answer to question #3: Is the ratio of questions with qtype-coarse='Drugs' from the 'GHR' source significantly higher than this ratio for all other sources?

It is a **hypothesis test question**. We calculate the drug ratio for each source that equals (number of questions with qtype-coarse = 'drugs' in that source) divided by the total number of questions in that source. The source that has the higher ratio is GHR. As shown in the Table 2

The p-value = 1.6634689505259636e-16 that means when the null hypothesis is true, the statistical summary would be equal to or more extreme than the actual observed results.

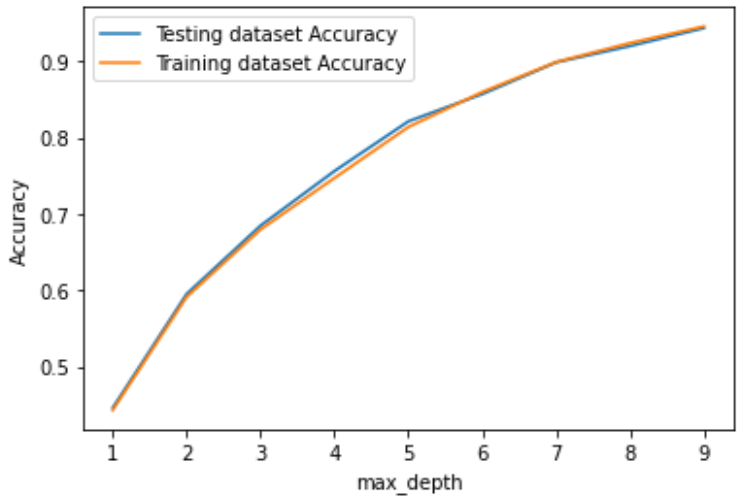

Figure 2. Decision Tree Classifier accuracy.

NIHSenior Health	GARD	NINDS	NHLBI	MPlusHealth Topics	CancerGov	CDC	GHR	NIDDK
1.2029	3.4072	1.6547	0.8731	0.01221	0.6533	0.5251	6.6312	1.1662

Table 2. Sources Ration

4 LIMITATIONS

1. We have fewer data points, so the error is high.
2. Lack of knowledge in clinical field.
3. There is under-fitting in the MedQuAD dataset and we solved it.

5 TAKE-AWAY MESSAGES

1. We need more data to train and test our algorithms.
2. Predict more information about the medical questions like the focus column and so on.
3. Make Chat-bot and student, patients, and medical staff.

6 REPLICATION PACKAGE

The first data set $(\text{paired}_{q}uestions)andanswerstothefirstnon-trivialquestionlink$:
https://drive.google.com/file/d/1B5QiIc9z93NNNQdjI3k7gU2ws9Q2Ip0F/view?usp=sharing
The second data set (MedQuAD) answered the second and third non-trivial questions linked:
https://drive.google.com/file/d/1AYjgrvm9YEElEXYBjZvrwSC6O2U0JWq3/view?usp=sharing

7 DISTRIBUTION OF WORKLOAD

1. Marim Amer
   (a) Reprocessing (user question, database question) columns from paired questions CSV file.
   (b) Hypothesis test question, and part of regression question.
   (c) Imbalanced data.
   (d) Jaccard, Levenshtein, and Dice Coefficient Similarity.
   (e) Neural Network, Passive Aggressive Regressor, and Linear Regression models.

2. Mawada Amer
    (a) Reprocessing columns from MedQuAD CSV file.
    (b) Hypothesis test question, and predictive question.
    (c) Decision Tree, KNN, SVM model, and Logistic Regression models.
3. Youmna Abd Alatty
    (a) Reprocessing(judge score) columns from paired questions CSV file.
    (b) Part of regression question.
    (c) Overlap Coefficient Similarity, and Cosine Similarity.
    (d) KNN Model, Decision Tree, and SVM models.

REFERENCES

Ben Abacha, A. (2017). Qa-testset-liveqa-med-qrels-2479-answers.

Ben Abacha, A. and Demner-Fushman, D. (2019). A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

(Ben Abacha and Demner-Fushman, 2019).

(Ben Abacha, 2017).