CISC839 Project-20: A QUESTION ENTAILMENT APPROACH TO QUESTION ANSWERING

Nourhan Abdelkerim[1], **Sondos Mahmoud**[2], **and Mostafa Awad**[3]

[1]**21nmma1@queensu.ca**
[2]**21smam1@queensu.ca**
[3]**21mmaa1@queensu.ca**

1 BACKGROUND AND OBJECTIVE

One of the challenges in large-scale information retrieval (IR) is to develop domain-specific methods to answer natural language questions.That is when users tries to search online for an answer to their question they struggle to navigate through many links and web pages to find a complete sufficient answer.

Despite the availability of numerous sources and datasets for answer retrieval, Question Answering (QA) remains a challenging problem due to the difficulty of the question understanding and answer extraction tasks when tackling domain-specific searches. This is due to several factors, such as the lexical and semantic challenges of domain-specific data that often include advanced argumentation and complex contextual information, the higher sparseness of relevant information sources, and the more pronounced lack of similarities between users' searches.

The objective of the project is to help in building a system to answer questions based on Recognizing Question Entailment (RQE). This system retrieve answers to a premise question by retrieving entailed questions that already have associated answers. Therefore, the entailment relation between two questions is defined as: question A entails question B if every answer to B is also a correct answer to A.

Our main task is to provide solutions to improve the performance of the Question Answering task. By build a model to measure the entailment score between two questions, this will improve the performance by retrieving questions with the highest similarity to the proposed one. The second contribution is a model that can predict the type of the question, It can improve the performance by comparing the proposed question only with questions with the same type, this will reduce the run time and make the Question Answering system faster and provide better results. To accomplish this task we will use two datasets the first one is MedQuAD dataset and the second one is dataset retrieved from the evaluation of LiveQAMed2017-TestQuestions Dataset.

1.1 Hypothesis Question

Is the ratio of treatment questions in the dataset significantly different from this ratio in the test dataset? When we want to test the question answering system, to achieve more accurate results so we need to know if the distribution of question types in train data will cover all users need by checking test data.

1.2 Regression Question

What is the entailment score between two question? This will help question answering system to decide which question is similar to the proposed one and retrieve its answer.

1.3 Predictive Question

Can we predict any characteristics of the user question like question type? this can help the question answering system to provide a better answer which is related to the type of the question by filtering the dataset so that the model find the entailment within smaller size of dataset.

2 DATASET

2.1 MedQuAD Dataset

2.1.1 Train Data

The first dataset is MedQuAD. MedQuAD includes 47,457 medical question-answer pairs created from 12 NIH (National Institutes of Health) websites. The collection covers 37 question types (e.g. Treatment, Diagnosis, Side Effects). Data was crawled from websites from the National Institutes of Health7. Each web page describes a specific topic (e.g. name of a disease or a drug). Hand-crafted patterns were constructed for each website to automatically generate the question-answer pairs based on the document structure and the section titles. To provide additional information about the questions that could be used for diverse IR and NLP tasks, the questions were automatically annotated with Semantic Type.

2.1.2 Test Data

Testset is generated from the same dataset (MedQuAD Dataset), we may split the dataset into train and test and use the trainset to learn the model and the testset for evaluation.

2.2 LiveQAMed2017-TestQuestions Dataset

2.2.1 Train Data

The second dataset is extraced from the manual evaluation of the results for the testset (TREC-2017-LiveQA-Medical-Test). To create this dataset, the test questions of the medical task at TREC-2017 LiveQA were passed to a model that can measure the entailment degree between two questions (questions from this testset and questions from MedQuAD Dataset) then retrieve the answers of the top 10 candidates. Then an interface to perform the manual evaluation of the retrieved answers was developed.

The evaluation was done by 3 assessors: a medical doctor, a medical librarian and a researcher in medical informatics. They gave a score for each answer from 1 to 4 where Excellent Answer takes rank 4, Correct but Incomplete takes rank 3, Incorrect but Related takes rank 2, Incorrect takes rank 1. Then these scores are used to tell the entailment degree between the questions.

Medical questions at TREC-2017 LiveQA are randomly selected from the consumer health questions that the NLM receives daily from all over the world. The test questions cover different medical entities and have a wide list of question types such as Comparison, Diagnosis, Ingredient, Side effects and Tapering.

2.2.2 Test Data

To generate a testset LiveQAMed2017-TestQuestions Dataset is splitted into two parts: 70% for train and 30% for test.

2.3 Data Preprocessing

Data preprocessing is an essential step in building a Machine Learning model and depending on how well the data has been preprocessed, the results are seen. At the first we removed punctuation like (.,!*"@') by selecting pattern of the text that we want and neglect others then convert all characters to lowercase. Second step we removed stopwords from the text by NLTK library as they don't add any value to the analysis like ['am', 'is', 'are', 'does', 'has', 'not', 'or' and etc.]. Based on the medical dataset we selected some stopwords and did not remove them like ['who', 'what', 'when', 'why', 'how', 'which', 'where', 'whom'], because they make a different in the meaning of the medical questions. We applied lemmatizing or diminished text to the root form and makes sure that it doesn't lose its meaning by WordNetLemmatizer. After all the text processing steps are performed, the final acquired data is converted into the numeric form using TF-IDF transformer, we applied it during pipeline of the model.

2.4 Basic Statistics of the Dataset

Both datasets have no missing values.

Dataset	MedQuAD	pair questions
	4119	234

Table 1. Number of duplicated.

	Question	qtype
count	47441	47441
unique	44603	39
top	What causes Causes of Diabetes ?	information
freq	20	9214

Figure 1. Statistics summary of MedQuAD dataset.

	Question_1	normal_Ques	Judge_score
count	2479	2479	2479
unique	1805	103	4
top	What is (are) Shingles ?	My question to you is: what is the reason that...	1-Incorrect
freq	15	45	1436

Figure 2. Statistics summary of Pair questions categorical dataset.

3 ANSWERS TO THE RESEARCH QUESTIONS

3.1 Answer to question #1: Hypothesis Question

By using MedQuAD dataset with question and question type columns as inputs. I split the dataset into 70 percent for the train data and 30 percent for test data. After that, i apply Chi-square Test for comparing differences between two independent groups. It tests the hypothesis that if the two groups come from same population or have the same medians.Chi-squared tests are based on chi-squared statistic.I calculate the chi-squared statistic with the following formula:

$$\sum \frac{(observedexpected)^2}{expected}.$$

In the formula, observed is the actual observed count for each category and expected is the expected count based on the distribution of the population for the corresponding category.In the chi-square test we compare the chi-square test statistic to a critical value based on the chi-square distribution.
Null hypothesis: Two groups have equal median.
Alternative hypothesis: Two groups does not have equal median.
SO, i found p value=1.0 which means that i failed to reject the null hypothesis and the two datasets have equal median.

3.2 Answer to question #2: Regression Question

To find the entailment score between two questions we used a feature-based approach in which we extract features from the input questions and feed them to a liner regression model. The first step is to compute six similarity measures between the pre-processed questions and use their values as features. We use Word Overlap, the Dice coefficient based on the number of common bigrams, Cosine, Levenshtein, Jaro Winkler, and the Jaccard similarities. Our feature list also includes the maximum and average values obtained with these measures and the question length ratio (length(PQ)/length(HQ)). Using LiveQAMed2017-TestQuestions Dataset after splitting it to train and test set with 30% for the testset we got the results in table(2). To improve the model performance we tried different approaches: we tried to use polynomial

	MSE	MAR	MAPE
Linear Regression	0.81	0.63	0.45

Table 2. Liner regression score.

regression and replacing the proposed question with the scientific paraphrasing of it. To tune the degree hyper-parameter for the polynomial regression we using grid search and the best model was at degree = 2.

Then using the polynomial regression model with scientific question we got the best result for far. you can find the full compression for all trial in table(3). Using the polynomial features the Mean square error

	MSE	MAR	MAPE
Linear Regression	0.81	0.63	0.45
polynomial regression	0.63	0.62	0.43
polynomial regression with scientific questions	0.49	0.52	0.63

Table 3. Regression Evaluation scores.

decreased significantly which means more error are now less than one which mean better performance. Regarding using scientific questions the model shows great performance with means if user ask a question using clinical words it's more likely to get better answers.

3.3 Answer to question #3: Predictive Question

I use MedQuAD dataset with Questions and question type columns as inputs. Firstly, I apply Label Encoder preprocessing to qtype column .For Questions column ,after preprocessing, i apply TfidfTransformer and CountVectorizer methods. All of these methods are considered in pipeline. Then,i split dataset by using Train Test Split method with 70 percent for training dataset and 30 percent for testing dataset. Then, i create support vector machine and logistic regression models with OneVsRest classifier. This classifier can be used to use a binary classifier like Logistic Regression for multi-class classification and. I set C=1, penalty= l1 and solver= liblinear for logistic regression's hyperparameters and C= 1, gamma= 1, kernel=sigmoid for support vector machine model.

3.3.1 Evaluate LogisticRegression model

1-The area under the curve score = 0.99, which mean that the model has the best performance for classification.

	F1 Score	Accuracy	recall	precision
		0.98		
macro avg	0.98		0.97	0.97
weighted avg	0.98		0.98	0.98

Table 4. Evaluation scores.

3.3.2 Evaluate SVC model

Also AUC = 0.99, so the two models have the same accuracy which mean that the dataset Very suitable to predict the question type. So any of these models can help the question system to provide a better answer easily which is related to question type.

	F1 Score	Accuracy	recall	precision
		0.98		
macro avg	0.97		0.97	0.97
weighted avg	0.98		0.98	0.98

Table 5. Evaluation scores.

4 LIMITATIONS

For paired questions dataset, during exploration we found data is imbalanced, so we tried to oversampling it by resample library. After oversampling we found the mse increased. We found that the resample is not a good idea with regression tasks.

The second thing we found is the data contains duplicate pair of question but have different rank, so we tried to replace this ranks with another value of rank indicates to both of them then drop duplicates. After replace ranks we found the model's performance is being better than before.

Our dataset contains three columns of questions, first one is hypothesis question, second is normal question asked by people has no medical background but the third is scientific question. So we used the third column to measure what will happen to the performance if the asked questions was scientific not normal. After that we found the performance is being great.

5 TAKE-AWAY MESSAGES

1. The ratio of treatment questions in the train dataset is the same as the ratio in the testset. 2. Logistic regression achieved high and very sufficient accuracy. 3. Question entailment model achieved moderate results and needs more work. 4. If users ask question using clinical words it's more likely to find a better answer. 5. This project needs large scale of datasets instead of the available size of paired question data. 6. Need some medical knowledge. 7. According to the time limitation we couldn't apply huge change in deep learning model hyperparameters on paired question problem, as we searched and according to the papers that we read the Siamese Network Text Similarity model is perfect to fit this problem, and according to the results that we achieved deep learning model achieved higher performance than machine learning models, so we suggest to dive more inside it in the future. We couldn't have a chance to talk about it in this report cause of slides limit.

6 REPLICATION PACKAGE

You can check the notebook of this project from the link:

https://colab.research.google.com/drive/1MwLDDooXA9OCH4Nsc3jxZ3F6Zy3S8BQM?usp=sharing
We added the link of datasets inside the notebook.

7 DISTRIBUTION OF WORKLOAD

7.1 MedQuAD Data

7.1.1 First Member worked on:

1. Data cleaning and preprocessing. 2. Build SVC model.

7.1.2 Second Member worked on:

1. Solve hypothesis question. 2. Build Logistic Regression model.

7.2 QA-TestSet-LiveQA-Med-Qrels-2479-Answers Data

7.2.1 First Member worked on:

1. Prepare data and extract features from files. 2. Data Cleaning and preprocessing. 3. Data expolration. 4. Build Deep Learning model.

7.2.2 Third Member worked on:

1. Prepare data and extract features from files. 2. Find similarities with similarity functions. 3. Build Linear Regression and Polynomial models.

REFERENCES

@articleBenAbacha-BMC-2019, title=A Question-Entailment Approach to Question Answering, author=Asma Ben Abacha and Dina Demner-Fushman, journal==BMC Bioinform., number=1, pages=511:1–511:23, year=2019, url = https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3119-4