KANTIPUR ENGINEERING COLLEGE

(Affiliated to Tribhuvan University)

Dhapakhel, Lalitpur

[Subject Code: CT755]

A MAJOR PROJECT FINAL REPORT ON

SEMANTIC QUESTION MATCHING WITH DEEP LEARNING

Submitted by:

Ankit Kharel [2072/BCT/95]

Anshul Bikram Rana [2072/BCT/96]

Kritish Dhaubanjar [2072/BCT/102]

Pradyumna Subedi [2072/BCT/110]

A MAJOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

December, 2018

SEMANTIC QUESTION MATCHING WITH DEEP LEARNING

Submitted by:

Ankit Kharel [2072/BCT/95]

Anshul Bikram Rana [2072/BCT/96]

Kritish Dhaubanjar [2072/BCT/102]

Pradyumna Subedi [2072/BCT/110]

A MAJOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR IN COMPUTER ENGINEERING

Submitted to:

Department of Computer and Electronics Engineering

Kantipur Engineering College

Dhapakhel, Lalitpur

December, 2018

KANTIPUR ENGINEERING COLLEGE
DEPARTMENT OF COMPUTER AND ELECTRONICS ENGINEERING

APPROVAL LETTER

The undersigned certify that they have read and recommended to the Institute of Engineering for acceptance, a project report entitled ”Semantic Question Matching with Deep Learning” submitted by

Ankit Kharel [2072/BCT/95]

Anshul Bikram Rana [2072/BCT/96]

Kritish Dhaubanjar [2072/BCT/102]

Pradyumna Subedi [2072/BCT/110]

in partial fulfillment for the degree of Bachelor in Computer Engineering.

.....
Supervisor
Supervisor’s Name
Supervisor’s Designation
Second Line of Designation (if required)

.....
External Examiner
External’s Name
External’s Designation
Second Line of Designation (if required)

.....
Er. Rabindra Khati
Head of Department
Department of Computer and Electronics Engineering

Date: December 26, 2018

ABSTRACT

In order to build a high-quality knowledge base, it's important to ensure each unique question exists on Q&A forums like Quora, Reddit and StackOverflow, only once. Writers shouldn't have to write the same answer to multiple versions of the same question, and readers should be able to find a single canonical section with the question they're looking for.

To prevent duplicate questions from existing on Q&A forums, we propose a machine learning system to identify if two questions have the same intent via analysis of exact semantic coincidence between questions in the forums and learn to match questions with answers by considering their semantic encoding.

Keywords – Keyword 1, Keyword 2...

ACKNOWLEDGMENT

This project is prepared in partial fulfillment of the requirement for the bachelor's degree in Computer engineering. First and foremost, we would like to express our sincere gratitude towards **Er. Rabindra Khati**, HOD, for his precious encouragement. We would like to thank **Er. Dipesh Shrestha**, Deputy HOD, for his constant guidance and inspiring lectures. Without his invaluable supervision and suggestions, it would have been very difficult for us. We would like to thank **Mr. Tek Narayan Adhikari** and **Er. Shiva Ghimire** for helping us with the Machine Learning. We are also highly indebted to **Er. Bishal Thapa** for co-operating with us and letting us carry out this project smoothly.

Last but not the least, we are thankful to all our teachers and all other people who have contributed in preparation of this report. We have tried to make this report error free. Any error in this report is our own responsibility. Suggestions and comments from readers are highly appreciable.

Ankit Kharel	[2072/BCT/95]
Anshul Bikram Rana	[2072/BCT/96]
Kritish Dhaubanjar	[2072/BCT/102]
Pradyumna Subedi	[2072/BCT/110]

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABC Annapurna Base Camp

UN United Nations

LIST OF SYMBOLS

a Area of Triangle

α Transparency Factor

CHAPTER 1
INTRODUCTION

Background

Over 300 million people visit Quora every month and more than 38 million questions have been asked till date. It comes in no surprise that many users ask similarly intended questions. Multiple questions with the same intent can cause seekers to spend more time finding the best answer to their question and make writers feel they need to answer multiple versions of the same question. Quora values canonical questions because they provide a better experience to active seekers and writers, and offer more value to both of these groups in the long term.

It was a severe issue and hence Quora published its dataset for the first time in 27 Feb, 2017, which could be used for the machine learning for data analysis. Quora has given an (almost) real-world dataset of question pairs, with the label of is_duplicate along with every question pair. The objective was to minimize the log loss of predictions on duplicity in the testing dataset.

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

Figure 1.1: Few sample lines of the quora dataset

Source: https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

In figure 1.1 though the questions with qid1 and qid2 of 6542 and 6543 respectively mean the same, the phrase are completely different and hence are the duplicate questions. The existence of duplicate questions can be analyzed using the machine learning and training them using the datasets.

Unlike the first one, the questions with qui1 and qid2 of 895 and 896 respecively shown in figure 1.1 look similar in terms of phrases but are not the duplicates. We need to test

for exact semantic coincidence between questions and analyze if it is the question with same intent. Hence, this proposal is all about tackling this natural language processing problem by applying a noval approach to classify whether question pairs are duplicates or not. Doing so will make it easier to find high quality answers to questions resulting in an improved experience for Q&A forum writers, seekers and readers.

Problem Statement

All the forums are based on the principle that ensures, each unique question exists on the forum only once and that there should be a single question section for each logically distinct question. For example, we'd consider questions like "What are the best ways to lose weight?", "How can a person reduce weight?" and "What are effective weight loss plans?" to be duplicate questions because they all have the same intent and should not exist separately because the intent behind both is identical.

The problems resulted due to question duplication may be enlisted as:

1. Lack of a sensible search.
2. Dull repetition and paucity of response to questioners.
3. Answer fatigue
4. Segregation of information
5. Distribution of user based on varying level of activity and intellect of answers.
6. Redundancy in answers.

Objectives

1. To analyze and merge the questions having same semantic meaning so as to prevent the duplication.
2. To mitigate the redundancy in answers and increase the reliability.

CHAPTER 2
FIGURES AND TABLES

Section for Sample Figure

Partem graece mucius est ei, usu ex facilisis dissentiet, probo prodesset scriptorem ius ea. Ne mei essent splendide, case tritani duo ei. Eam ne stet feugiat albucius. Ne diam dicant viderer eam, at pri minim vivendo periculis, in vix commodo invidunt. In eum omnis vocent aperiam. Sit prima atqui splendide ad, singulis recusabo quaerendum at his, enim eligendi forensibus ea eam.

Sed veri aeque persecuti ut. Ut accusam mediocrem accusamus eos, quis clita probatus ex his. Mea oratio deleniti interesset at. Odio melius antiopam eu his, ut nec tantas maiestatis ullamcorper. Ut numquam reprimique cum.

Figure 2.1: The Sample Image One

Referring Figure

This is an example where Figure 2.1 of page 3 is referenced. Detracto suavitate id per, no est putent accusata quaestio, purto quaeque oporteat ei sea. Id eam erat affert, ex has summo inimicus partiendo. Option aliquam imperdiet ius ex. Efficiendi omittantur in mea, id usu tacimates rationibus. Ei accusamus dissentias vix, eos aperiam percipit id.

Mea cu vitae noluisse. Tation eirmod iracundia sea no, duo no aliquando elaboraret.

Qui ut legere mucius, dolore efficiendi definitionem quo ex. Usu te falli similique posidonium, eum eu dicat aeterno phaedrum, te paulo deleniti ius. Pro te aliquam platonem, eos ea dolore phaedrum. Graece honestatis sit at, nec id ubique legendos.

Section for Sample Table

Prompta albucius ne vel. Has te intellegat definitionem, vim duis animal adipiscing ei. At qui iisque accusata, id his possim nominati. Ne agam dissentias eos, dico vivendo percipit ut mea.

Table 2.1: Table Example

x	3	4	5	5	5
f(x)	111	22	30	40	500

Referring Table

This is an example where Table 2.1 of page 4 is referenced. Detracto suavitate id per, no est putent accusata quaestio, purto quaeque oporteat ei sea. Id eam erat affert, ex has summo inimicus partiendo. Option aliquam imperdiet ius ex. Efficiendi omittantur in mea, id usu tacimates rationibus. Ei accusamus dissentias vix, eos aperiam percipit id.

Mea cu vitae noluisse. Tation eirmod iracundia sea no, duo no aliquando elaboraret. Qui ut legere mucius, dolore efficiendi definitionem quo ex. Usu te falli similique posidonium, eum eu dicat aeterno phaedrum, te paulo deleniti ius. Pro te aliquam platonem, eos ea dolore phaedrum. Graece honestatis sit at, nec id ubique legendos. Detracto suavitate id per, no est putent accusata quaestio, purto quaeque oporteat ei sea. Id eam erat affert, ex has summo inimicus partiendo. Option aliquam imperdiet ius ex. Efficiendi omittantur in mea, id usu tacimates rationibus. Ei accusamus dissentias vix, eos aperiam percipit id.

Mea cu vitae noluisse. Tation eirmod iracundia sea no, duo no aliquando elaboraret. Qui ut legere mucius, dolore efficiendi definitionem quo ex. Usu te falli similique posidonium, eum eu dicat aeterno phaedrum, te paulo deleniti ius. Pro te aliquam platonem, eos ea dolore phaedrum. Graece honestatis sit at, nec id ubique legendos. Detracto

suavitate id per, no est putent accusata quaestio, purto quaeque oporteat ei sea. Id eam erat affert, ex has summo inimicus partiendo. Option aliquam imperdiet ius ex. Efficiendi omittantur in mea, id usu tacimates rationibus. Ei accusamus dissentias vix, eos aperiam percipit id.

Mea cu vitae noluisse. Tation eirmod iracundia sea no, duo no aliquando elaboraret. Qui ut legere mucius, dolore efficiendi definitionem quo ex. Usu te falli similique posidonium, eum eu dicat aeterno phaedrum, te paulo deleniti ius. Pro te aliquam platonem, eos ea dolore phaedrum. Graece honestatis sit at, nec id ubique legendos. Detracto suavitate id per, no est putent accusata quaestio, purto quaeque oporteat ei sea. Id eam erat affert, ex has summo inimicus partiendo. Option aliquam imperdiet ius ex. Efficiendi omittantur in mea, id usu tacimates rationibus. Ei accusamus dissentias vix, eos aperiam percipit id.

Mea cu vitae noluisse. Tation eirmod iracundia sea no, duo no aliquando elaboraret. Qui ut legere mucius, dolore efficiendi definitionem quo ex. Usu te falli similique posidonium, eum eu dicat aeterno phaedrum, te paulo deleniti ius. Pro te aliquam platonem, eos ea dolore phaedrum. Graece honestatis sit at, nec id ubique legendos.

Complex Table from Excel

Complex tables can be created in MS-Excel and latex code for the table can be generated using ”excel2latex” add-in. An example of a complex table is shown in Table 2.2 in page 6
you can download excel2latex add-in from
https://www.ctan.org/tex-archive/support/excel2latex
Some extra packages are required. like bigstrut, multirow etc. Mea cu vitae noluisse. Tation eirmod iracundia sea no, duo no aliquando elaboraret. Qui ut legere mucius, dolore efficiendi definitionem quo ex. Usu te falli similique posidonium, eum eu dicat aeterno phaedrum, te paulo deleniti ius. Pro te aliquam platonem, eos ea dolore phaedrum. Graece honestatis sit at, nec id ubique legendos. Detracto suavitate id per, no est putent accusata quaestio, purto quaeque oporteat ei sea. Id eam erat affert, ex has summo inimicus partiendo. Option aliquam imperdiet ius ex. Efficiendi omittantur in

Table 2.2: Complex table converted from Excel using excel2latex

<table>
<tr><th>SN</th><th>Col 1</th><th>Col 2</th><th colspan="2">Col 3</th><th>Col 4</th></tr>
<tr><td>1</td><td colspan="2" rowspan="2">Merged Cells 1</td><td>a</td><td>b</td><td>e</td></tr>
<tr><td>2</td><td>c</td><td>d</td><td rowspan="3">f</td></tr>
<tr><td>3</td><td colspan="3">Merged Cells 2</td></tr>
<tr><td rowspan="3">4</td><td rowspan="3">abc</td><td rowspan="4">Merged Cells 3</td><td>a1</td><td>a2</td></tr>
<tr><td>a3</td><td>q</td><td>ww</td></tr>
<tr><td colspan="2" rowspan="2">a4</td><td>as</td></tr>
<tr><td rowspan="2">5</td><td>a11</td><td>qq</td></tr>
<tr><td></td><td colspan="2">Merged</td></tr>
</table>

mea, id usu tacimates rationibus. Ei accusamus dissentias vix, eos aperiam percipit id.

Mea cu vitae noluisse. Tation eirmod iracundia sea no, duo no aliquando elaboraret. Qui ut legere mucius, dolore efficiendi definitionem quo ex. Usu te falli similique posidonium, eum eu dicat aeterno phaedrum, te paulo deleniti ius. Pro te aliquam platonem, eos ea dolore phaedrum. Graece honestatis sit at, nec id ubique legendos. Detracto suavitate id per, no est putent accusata quaestio, purto quaeque oporteat ei sea. Id eam erat affert, ex has summo inimicus partiendo. Option aliquam imperdiet ius ex. Efficiendi omittantur in mea, id usu tacimates rationibus. Ei accusamus dissentias vix, eos aperiam percipit id.

Mea cu vitae noluisse. Tation eirmod iracundia sea no, duo no aliquando elaboraret. Qui ut legere mucius, dolore efficiendi definitionem quo ex. Usu te falli similique posidonium, eum eu dicat aeterno phaedrum, te paulo deleniti ius. Pro te aliquam platonem, eos ea dolore phaedrum. Graece honestatis sit at, nec id ubique legendos.

CHAPTER 3
EQUATIONS

Basics of Equations

Mathematical expression within text can be written as $y = mx + c$. In separate line as

$$ax + by + c = 0$$

Some latex mathmatics examples:

Superscript:

x^3

$$x^3$$

$$x^{3x+4}$$

$$x^{3x+4^4+5}$$

Subscript:

$$x_{13}$$

x_{12}

$$x_{123}$$

Greek letters

$$\pi$$

$$\alpha$$

αA β βB δ γ ϑ Θ ϕ φ Φ trigonometric: $y = \sin(\pi)$ Log: $y = \log(\pi)$ $y = \ln(\pi)$ $y = \log_{10}$ Square root: $\sqrt{2}$ $\sqrt[3]{4}$ $\sqrt{x^2 + y^2}$ $\sqrt[3]{x^2 + y^2}$ $\sqrt{\sqrt[3]{x^2 + y^2}}$ Fraction: About 2/3 of the class is full. About $\frac{2}{3}$ of the class is full. About $\dfrac{2}{3}$ of the class is full. About $\dfrac{\sqrt{\sqrt[3]{x^2 + y^2}}}{\sqrt{x^2 + x + 1}}$ of About $\dfrac{2}{1 + \sqrt{\sqrt[3]{x^2 + y^2}}}$ of the class is full.

Reserved characters: $\{a, b, c\}$ \$20 $10\%of100is100$ $10 \ \% \quad of \ \ 100 \qquad is \ \ 100$

Braket Style: $3(\frac{2}{3})\ 3\left(Hello(\frac{2}{3})\right)\ 3\left\{Hello(\frac{2}{3})\right\}\ 3\left\{Hello(\frac{2}{3})\ 3\ Hello(\frac{2}{3})\ \frac{dy}{dx}\big|_{x=1}\ (\ (\ \Big(\ \Big(\ 3(\frac{2}{3})\ 3\Big(\frac{2}{3}\Big)$

Equation:

$$E = mc^2 \tag{3.1}$$

$$E = mc^2$$

$$E = mc^2$$

$$E = mc^2 \tag{3.2}$$

$$E = mc^4 \tag{3.3}$$

$$E = mc^7 \tag{3.4}$$

$$E = mc^2$$

$$E = mc^4$$

$$E \approx \pm(mc^7 + 3)$$

$$E \approx \pm(mc^7 + 3)$$

$$E = mc^2 \tag{3.5}$$

$$E = mc^4$$

$$E \approx \pm(mc^7 + 3) \tag{3.6}$$

Limit: $\lim_{x \to a} f(x)$
$\lim_{x \to a} \frac{f(x)-f(a)}{x-a} = f'(a)$ Integration: $\int$
$\int(\quad sin \quad x \quad dx =)$
$\int(\quad sin \quad x \quad dx =) \int_a^b (\quad sin \quad x \quad dx =) \int_a^b x^2 \, dx \; = \left[\frac{x^3}{3} \right]_a^b$

Summation: $\sum_{n=1}^{10} \int_a^b f(x) \, dx \; = \lim_{x \to \infty} \sum_{K=1}^{10} f(x_k).\delta x$

Referencing the Equation

The Equation can be referenced using labels. example Equation 3.2 of page 8 is referenced here. LaTeX is the de facto standard for the communication and publication of scientific documents. LaTeX is available as free software.

CHAPTER 4
LISTING EXAMPLES

Here are some examples of listing

Ordered List Technique:

Listing Technique 1:

1. Pencil
2. Paper
3. Calculator
4. Notebook
    (a) Assignment
        i. Test
            A. Test 1
            B. Test 2
        ii. Quiz
    (b) Classwork

Listing Technique 2:

- Pencil
- Paper
- Calculator
- Notebook
    - Assignment
        1. Test
            * Test 1
            * Test 2
        2. Quiz
    - Classwork

Listing Technique 3:

1. Pencil
2. Paper

3. Calculator
4. Notebook
    A Assignment
        I Test
            i Test 1
            ii Test 2
        II Quiz
    B Classwork

The Subsections

LaTeX is a document preparation system for the communication and publication of scientific documents.

It is most often used for medium-to-large technical or scientific documents but it can be used for almost any form of publishing. as said in 1 of page no. 1 LaTeX is the de facto standard for the communication and publication of scientific documents. LaTeX is available as free software.

The subsubsection

LaTeX is a document preparation system for the communication and publication of scientific documents.

It is most often used for medium-to-large technical or scientific documents but it can be used for almost any form of publishing.

LaTeX is the de facto standard for the communication and publication of scientific documents. LaTeX is available as free software.

The paragraph LaTeX is a document preparation system for the communication and publication of scientific documents.

It is most often used for medium-to-large technical or scientific documents but it can be used for almost any form of publishing.

The subparagraph LaTeX is a document preparation system for the communication and publication of scientific documents.

It is most often used for medium-to-large technical or scientific documents but it can be used for almost any form of publishing.

CHAPTER 5
EXAMPLE ABBR AND SYMBOLS

Abbr

Here is an example of writing abbreviations ABC is abbr of Annapurna Base Camp UN is abbr of United Nations

Symbols

Now the symbol α is the Transparency Factor a is abbr of Area of Triangle Note that only those abbrs and bymbols that are included in the text will be listed in List of Abbr/Symbols.

CHAPTER 6
CITATION EXAMPLE

Citation and compiling bib file

This is an example of citing texts [**?**]. This is second citation [**?**] The first cited reference will be numbered ”1”, second ”2” and so on. Only those cited in the document will be listed in the reference section.

Note that to compile documents with reference correctly, you need to follow following steps:

1. Run pdfLatex (or Quick build)
2. Run Bibtex
3. Run pdflatex 2 times

Tation eirmod iracundia sea no, duo no aliquando elaboraret. Qui ut legere mucius, dolore efficiendi definitionem quo ex. Usu te falli similique posidonium, eum eu dicat aeterno phaedrum, te paulo deleniti ius. Pro te aliquam platonem, eos ea dolore phaedrum. Graece honestatis sit at, nec id ubique legendos. Detracto suavitate id per, no est putent accusata quaestio, purto quaeque oporteat ei sea. Id eam erat affert, ex has summo inimicus partiendo. Option aliquam imperdiet ius ex. Efficiendi omittantur in mea, id usu tacimates rationibus. Ei accusamus dissentias vix, eos aperiam percipit id.

Mea cu vitae noluisse. Tation eirmod iracundia sea no, duo no aliquando elaboraret. Qui ut legere mucius, dolore efficiendi definitionem quo ex. Usu te falli similique posidonium, eum eu dicat aeterno phaedrum, te paulo deleniti ius. Pro te aliquam platonem, eos ea dolore phaedrum. Graece honestatis sit at, nec id ubique legendos. Detracto suavitate id per, no est putent accusata quaestio, purto quaeque oporteat ei sea. Id eam erat affert, ex has summo inimicus partiendo. Option aliquam imperdiet ius ex. Efficiendi omittantur in mea, id usu tacimates rationibus. Ei accusamus dissentias vix, eos aperiam percipit id.

APPENDIX

Appendix Text Comes Here