Applying natural language inference or question entailment for crowdsourcing more data

Jingjing Shi u5792801
Course code: 8755
Supervisor: Zhenchang Xing
Motivation

- **35%** of US adults had gone online to figure out what medical conditions they might have. (Ben Abacha & Zweigenbaum, 2015)
- There are **2,000 to 100,000** health related Web sites to get large amount of medical information. (Diaz et al., 2002)
- However,

  Questions >> Answers
So…

How to expand the current medical question-answer pairs and find the best answer from the candidates?

e.g. I am diabetic, 56 years old and my legs are burning from my knees to my toes. Can you help please

Which is the best answer?

1. You are experiencing neuropathic pain due to diabetes. Consult your doctor for blood tests to determine if your diabetes is well controlled.
2. You definitely needs medical attention. You diabetes will have to be controlled in order for the sores to heal. The sores on your legs will progressively get worse and could even cause septicemia.
Recognizing Question Entailment (RQE)

Question A entails question B if every answer to B is also a correct answer to A (X and Y are similar questions)

e.g.

Question A: Hi I have retinitis pigmentosa for 3 years. I’m suffering from this disease. Please introduce me any way to treat my eyes.

Question B: Are there treatments for RP?

A entails B since every answer to B is also an answer to A.
Natural Language Inference (NLI)

Whether a given hypothesis (H) can be inferred from a given premise (P) (Romanov & Shivade, 2018)

Entailment: ( A and B are semantically similar)

e.g. mother developed separation of symphysis pubis and was put in traction (P), She has orthopaedic injuries (H)

Natural (A and B are unrelated)
e.g. No known sick contacts (P), No recent travel (H)

Contradiction (A and B have opposite semantic meaning)
e.g. The infant emerged with spontaneous cry (P), The infant was till born (H)
Process

Using existing medical question-answer pairs to train NLP models to achieve RQE and NLI

Crawling medical question-answer pairs from medical website (including the questions were not answered)

Preprocessing the dataset

Using trained models to find the best answers and expand the current dataset
Reference

