Applying Natural Language Inference or Question Entailment for Crowdsourcing More Data

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

• Motivation
• Data Description
• Methodology
• Results
• Conclusion & Future Work
• Reference
Motivation

Question answering (QA) is widely investigated in the open domain. However, for the specific domain, such as patient question answering, it is less studied.

Challenges:
- less annotated data
- Quality of data set
- Experts require knowledge to answer
- Take long time to answer each question (up to 4 hours) [Russell-Rose and Chamberlain, 2017]

Picture is from: https://towardsdatascience.com/automatic-question-answering-ac7593432842
Motivation

If the data size is small, it is easier to cause overfitting, and the model cannot generalize well:

• Cause out-of-vocabulary (OOV) problem
  – an unseen word encountered in the fixed vocabulary language model, but the model cannot handle it appropriately
Motivation

To address these problems, our aim is to use recognizing question entailment (RQE) and natural language inference (NLI) to crowd source data for this particular domain.
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) [Abacha and Demner-Fushman, 2016] (Examples are from CHQs)

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] (Examples are from MedNLI)

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)
Data Description

Q. Help with symptoms

My friends 18-year-old son who is having some serious issues. He spent the last week in the hospital and they can’t figure out what it is. I thought I would post here in hopes that someone might have an idea. His symptoms are as follows; rapid weight loss, excruciating muscle pain throughout the entire body (this occurs in flares with no predictability), extreme night sweats, fever, rash/skin peeling on the fingers and toes. The hospital instructed them to take him to the rheumatologist. Any thoughts on what this might be?

More: ALL EXPERTS CANCER EXPERT

A. Cancer Expert - 2019/09/16

Dear Anonymous, I can quite understand your concerns. It is not possible to make any diagnosis from the signs and symptoms listed. It is suggested that he be seen by a specialist physician who will conduct a battery of tests to determine exactly what the cause could be. (MCH).

The information provided does not constitute a diagnosis of your condition. You should consult a medical practitioner or other appropriate health care professional for a physical examination, diagnosis and formal advice. Health24 and the expert accept no responsibility or liability for any damage or personal harm you may suffer resulting from making use of this content.
Data Description

22 categories with 25,705 questions.

There are 23,755 questions have answers, and 1,950 questions have not been answered yet.
Methodology

Pre-processing:
Use ScispaCy to expand abbreviation, such as from bp to blood pressure

Top-performing system for RQE and NLI is transform learning since pre-training can provide better generalization [Erhan et al., 2010]:
• Do pre-training on the generic task by using supervised or unsupervised methods
• Fine-tune on the specific downstream tasks

Pre-trained BERT (Bidirectional Encoder Representations from Transformer) [Devlin et al., 2018] used here for RQE and NLI tasks
Methodology

BERT (Bidirectional Encoder Representations from Transformer)

The pre-trained model we use is SciBERT since it is pre-trained in the scientific text.

Methodology

CHQs: received by U.S National Library of medicine (NLM) and Frequently Asked Questions (FAQs) from NIH institutes

MedNLI: from the MIMIC-III database

Use fine-tuned model in RQE task to locate similar question (answered question) to the unanswered questions.

Use fine-tuned model in NLI task to validate if the similar answered questions’ answer can be inferred from its corresponding unanswered question.
Methodology

Use fine-tuned model in RQE task to locate similar question (answered question) to the unanswered questions.

Use fine-tuned model in NLI task to validate if the similar answered questions’ answer can be inferred from its corresponding unanswered question.

Set threshold in the output of testing in order to deal with ambiguous case.

Use the idea of pseudo-relevance feedback to evaluate results.
Methodology

- Filter some dissimilar question pairs:
  - TF-IDF
  - BM25
  - Word2Vec
Results

For RQE task:
Use p-value to determine the significance of the result.
Null hypothesis: the generated data do not have high accuracy.
Alternative hypothesis: the generated data have high accuracy.

<table>
<thead>
<tr>
<th></th>
<th>CHQs</th>
<th>CHQs +TF-IDF</th>
<th>CHQs + BM25</th>
<th>CHQs + Word2Vec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iteration 1</td>
<td>0.9678</td>
<td>0.9715</td>
<td><strong>0.9770</strong></td>
<td>0.9709</td>
</tr>
<tr>
<td>Iteration 2</td>
<td>0.9682</td>
<td>0.9737</td>
<td>0.9720</td>
<td>0.9666</td>
</tr>
<tr>
<td>Iteration 3</td>
<td>0.9727</td>
<td>0.9720</td>
<td>0.9753</td>
<td><strong>0.9748</strong></td>
</tr>
<tr>
<td>Iteration 4</td>
<td>0.9698</td>
<td>0.9742</td>
<td>0.9748</td>
<td>0.9676</td>
</tr>
<tr>
<td>Iteration 5</td>
<td><strong>0.9732</strong></td>
<td><strong>0.9753</strong></td>
<td>0.9737</td>
<td>0.9709</td>
</tr>
<tr>
<td>Average</td>
<td>0.9703</td>
<td>0.9733</td>
<td>0.9746</td>
<td>0.9702</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>added data number</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td><strong>7,412</strong></td>
</tr>
<tr>
<td>BM25</td>
<td>7,104</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>1,936</td>
</tr>
</tbody>
</table>
Results

For NLI task:

Null hypothesis: the generated data do not have high accuracy.
Alternative hypothesis: the generated data have high accuracy.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Iteration 1</td>
<td>0.7256</td>
<td>0.7313</td>
<td>0.7409</td>
<td>0.7412</td>
</tr>
<tr>
<td>Iteration 2</td>
<td>0.7255</td>
<td>0.7476</td>
<td>0.7266</td>
<td>0.7295</td>
</tr>
<tr>
<td>Iteration 3</td>
<td>0.7262</td>
<td>0.7334</td>
<td>0.7366</td>
<td>0.7287</td>
</tr>
<tr>
<td>Iteration 4</td>
<td>0.7319</td>
<td>0.7357</td>
<td>0.7369</td>
<td>0.7348</td>
</tr>
<tr>
<td>Iteration 5</td>
<td>0.7504</td>
<td>0.7497</td>
<td>0.7469</td>
<td>0.7223</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>0.7320</td>
<td>0.7395</td>
<td>0.7376</td>
<td>0.7313</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>added data number</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>1,872</td>
<td>0.1406</td>
</tr>
<tr>
<td>BM25</td>
<td>1,707</td>
<td>0.1940</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>444</td>
<td>0.5420</td>
</tr>
</tbody>
</table>

Standard derivation:
MedNLI: 0.009543  
TF-IDF: 0.007596  
BM25: 0.006631  
Word2Vec: 0.006344
Conclusion & Future Work

• The data used in NLI task is small, they cause the result is not obvious. For the next step, we can expand our data.

• Fine-tuning and label generating are done in CoLab. Due to the limited hardware, we set small k for k-fold cross-validation. So, the samples for p-value are small. After finding a more powerful hardware, we can set k to 10 and see the result. Besides, we can do hyperparameter optimization and find the best hyperparameters for these models.

• Optimizing the way of using Word2Vec
  – Deal with typos and the words that do not occur in embeddings
  – Or find a more suitable pre-trained model to do embedding.
Reference


