Recognizing Question Entailment using Sci-BERT stacked with a Gradient Boosting Classifier

Prakhar Sharma, Sumegh Roychowdhury

Indian Institute of Technology Kharagpur, West Bengal – 721302, India

1. Motivation

- The number of people turning to the Internet to search for a diverse range of health-related subjects continues to grow and with this multitude of information available, duplicate questions become more frequent and finding the most appropriate answers becomes problematic.
- In all, 80 percent of Internet users, or about 93 million Americans, have searched for a health-related topic online, according to a study released on 16th July 2018 by the Pew Internet American Life Project [8].
- So we aim to retrieve similar questions that are already answered by human experts (like retrieving FAQs) to reduce the burden of manually searching through the entire database available, because manually checking entailment is a very tedious task.

2. Question Entailment

- We use the following definition of question entailment: *Question A entails a Question B if every answer to B is also a complete or partial answer to A*.

For example:
Q1: "Can you mail me patient information about Glaucoma, I was recently diagnosed and want to learn all I can about the disease."
Q2: "What is Glaucoma?"
In the above example, the answer of **Q1** implies the answer of **Q2** (Entailment).

3. Dataset

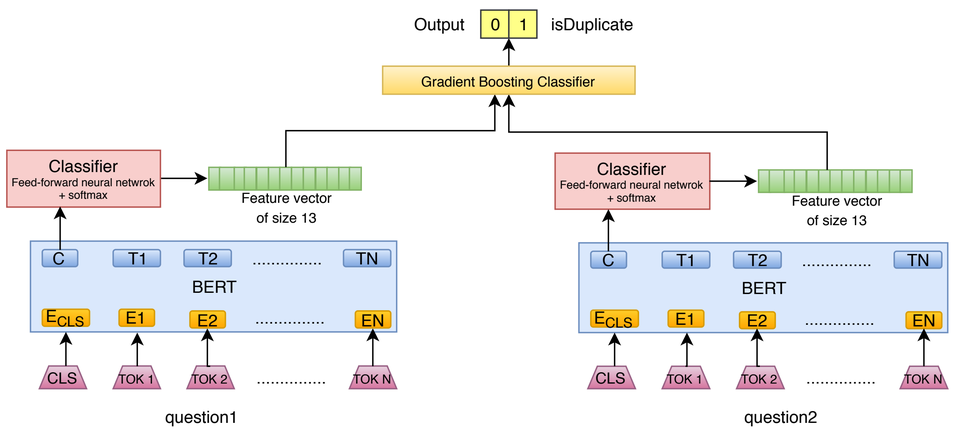
- We use the dataset provided by the **MEDIQA 2019 RQE Shared Task** [2].
- The training corpus contains 8,588 training pairs, containing 54.2% positive pairs, and the remaining 3,933 pairs are negative examples collected by associating a random short form of NLM dataset question [4] having at least one common keyword and one different keyword for each original question.
- The validation corpus contains 302 pairs of questions consisting of 173 negative and 129 positive pairs.
- The hidden test set had 230 pairs of which exactly 115 pairs (50%) were positive and rest 115 negative.
- We additionally used an annotated corpus of consumer health questions [6] to build our question type prediction classifier. The corpus consists of 1,467 consumer-generated requests for disease information, containing a total of 2,937 questions. The dataset has these requests classified into 13 question types/classes namely: *Anatomy, Cause, Complication, Diagnosis, Information, Management, Manifestation, Other Effects, Person-Org, Prognosis, Susceptibility, Other, Not Disease* .
- Table 1. shows an elementary data statistics, Positive means Entailment & Negative means Not Entailment.

Validation Set	Positive	Negative
Same Medical Entity	112	54
Different Medical Entity	17	119

Test Set	Positive	Negative
Same Medical Entity	87	101
Different Medical Entity	28	14

4. Proposed Approach - QSpider

- **QSpider** is a staged system consisting of Sci-BERT [1] stacked with a Gradient Boosting Classifier.
- Our model aims at capturing *question types* and use them as features to detect question entailment.
- **Embedding Layer:**
  Each Sub-network has an embedding layer which takes input, tokenizes them into words utilizing *BertTokenizer* and converts each word into vector representation. The positional embeddings, segment embeddings and token embeddings are all taken into account in BertEncoder. We also use Scibert-scivocab-uncased as the vocabulary for our model.
- **Classification Layer:**
  We use the *[CLS]* token for classification which is passed through a feed-forward neural network + softmax and mapped into a feature vector of size 13 containing probabilities for occurence of each of the 13 **question types** mentioned in Section 3.
- **Output Layer:**
  Both the feature vectors obtained are concatenated and passed into a Gradient Boosting Classifier which predicts whether the two questions are an entailment or not.



5. Baselines

- **Dependency Tree-LSTM:**
  We refer to Child-Sum Tree-LSTM [7] applied to a dependency tree as Dependency Tree-LSTM. We produced dependency parses using the *Stanford Neural Network Dependency Parser*. We first produce sentence representations h_L and h_R for **question1** and **question2** respectively in the pair using a Tree-LSTM model over question's parse tree. Given these sentence representations, we calculate the entailment probability θ using a neural network that considers both the distance and angle between the pair (h_L, h_R):

$$\mathrm{h}_\times = h_L \odot h_R,$$
$$\mathrm{h}_+ = |h_L - h_R|,$$
$$\mathrm{h}_s = \sigma\left(W^{(\times)}h_\times + W^{(+)}h_+ + b^{(h)}\right),$$
$$\hat{p}_\theta = \mathrm{softmax}\left(W^{(p)}h_s + b^{(p)}\right),$$

  We want $\hat{p}_\theta$ given model parameters θ to be close to the p. Here y denotes whether it is an entailment. Hence we decide the cost function as the regularized KL-divergence between p and $\hat{p}_\theta$:

$$\mathrm{p}_i = || \; i - y \; | - 1 \; | \quad i = \{0, 1\}$$
$$\mathrm{J}(\theta) = \frac{1}{m}\sum_{k=1}^{m} \mathrm{KL}\left(p^{(k)} \,\middle\|\, \hat{p}_\theta^{(k)}\right) + \frac{\lambda}{2}\|\theta\|_2^2$$

- **$\mathrm{BERT}_{\mathrm{Large,uncased}}$** :
  We chose $\mathrm{BERT}_{\mathrm{large,\ uncased}}$ [3] as our underlying BERT model. It consists of 24-layers, 1024-hidden, 16-heads and 340M parameters. The maximum sequence length was kept to be 128. The loss function was defined as the Binary Cross-Entropy Loss.

$$L_{entropy}(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right)$$

- **BERT variants + Hinge Loss:**
  We then tried using domain specific variants of BERT such as Bio-BERT [5] and Sci-BERT [1].Sci-BERT outperformed Bio-BERT in this task. Since for binary classification tasks, both Hinge Loss and Cross-Entropy Loss are widely used, we tried incorporating both of these losses in our model. In this task, Hinge Loss did give a better accuracy as reported below.

$$L_{hinge}(f) = \sum_{j \neq y_i} max(0, s_j - s_{y_i} + 1)$$

6. Results

Model	Valid	Test
Tree-LSTM	64.0	60.2
$\mathrm{BERT}_{\mathrm{large,\ uncased}}$	76.2	48.1
Bio-BERT	77.6	49.6
Sci-BERT + Hinge Loss	**80.5**	51.3
QSpider	62.0	**68.4**

- Our model was ranked 1st among all other competitors in the Development Phase, with *Sci-BERT + Hinge Loss* giving the highest accuracy of *80.5%*.
- We were ranked 3rd in the Testing Phase with our final model *QSpider* giving *68.4%* accuracy on the hidden test.

7. Error Analysis

- The dev set had 112 out of 302 examples with same medical entity that entail each other and 119 examples with different entities which don't entail each other. This was the reason why attention models like BERT performed well in the dev set by focusing more on the entity name rather than the actual meaning.
- The hidden test set had more than 80% pairs where they both have same medical entities but almost half of them don't entail each other. This caused a **huge drop in the performance** of *BERT*. But *Qspider* solves this problem by not only focusing on the entity name but also the semantics by capturing question type.

Validation Set	Correct	Wrong
Same Medical Entity (Positive)	112	0
Same Medical Entity (Negative)	1	53
Different Medical Entity (Positive)	13	4
Different Medical Entity (Negative)	117	2

Test Set	Correct	Wrong
Same Medical Entity (Positive)	87	0
Same Medical Entity (Negative)	1	100
Different Medical Entity (Positive)	24	4
Different Medical Entity (Negative)	6	8

References

[1] Iz Beltagy, Arman Cohan, and Kyle Lo. Scibert: Pretrained contextualized embeddings for scientific text. 2019.

[2] Asma Ben Abacha, Chaitanya Shivade, and Dima Demner-Fushman. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering, acl-bionlp 2019. In *Proceedings of the BioNLP 2019 workshop, Florence, Italy, August 1, 2019*. Association for Computational Linguistics, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[4] Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, and Pifer EA. A taxonomy of generic clinical questions: classification study. *British Medical Journal*, 321:429–432, 2000.

[5] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical languagerepresentation model for biomedical text mining. 2019.

[6] Kirk Roberts, Kate Masterton, Marcelo Fiszman, Halil Kilicoglu, and Dina Demner-Fushman. Annotating question types for consumer health questions. 2014.

[7] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. 2015.

[8] Jane Weaver. More people search for health online. 2016.

Contact

Sumegh Roychowdhury (sumegh01@iitkgp.ac.in)
Prakhar Sharma (prakharsharma@iitkgp.ac.in)