Towards More Robust Natural Language Understanding

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Bachelor of Science in Computer Science and Engineering with Honors Undergraduate Research Distinction in the College of Engineering of The Ohio State University

By

Xinliang (Frederick) Zhang,

Undergraduate Program in Computer Science and Engineering

The Ohio State University

2021

Thesis Committee:

Dr. Huan Sun, Advisor

Dr. Marie-Catherine de Marneffe

Abstract

Natural Language Understanding (NLU) is a branch of Natural Language Processing (NLP) that uses intelligent computer software to understand texts that encode human knowledge. Recent years have witnessed notable progress across various NLU tasks with deep learning techniques, especially with pretrained language models. Besides proposing more advanced model architectures, constructing more reliable and trustworthy datasets also plays a huge role in improving NLU systems, without which it would be impossible to train a decent NLU model. It's worth noting that the human ability of understanding natural language is flexible and robust. On the contrary, most of existing NLU systems fail to achieve desirable performance on out-of-domain data or struggle on handling challenging items (e.g., inherently ambiguous items, adversarial items) in the real world. Therefore, in order to have NLU models understand human language more effectively, it is expected to prioritize the study on robust natural language understanding.

In this thesis, we deem that NLU systems are consisting of two components: NLU models and NLU datasets. As such, we argue that, to achieve robust NLU, the model architecture/training and the dataset are equally important. Specifically, we will focus on three NLU tasks to illustrate the robustness problem in different NLU tasks and our contributions (i.e., novel models and new datasets) to help achieve more robust natural language understanding. The major technical contributions of this thesis are:

1. We study how to utilize diversity boosters (e.g., beam search & QPP) to help neural question generator synthesize diverse QA pairs, upon which a Question Answering (QA) system is trained to improve the generalization on the unseen target domain. It's worth mentioning that our proposed QPP (question phrase prediction) module, which predicts a set of valid question phrases given an answer evidence, plays an important role in improving the cross-domain generalizability for QA systems. Besides, a target-domain test set is constructed and approved by the community to help evaluate the model robustness under the cross-domain generalization setting.
2. We investigate inherently ambiguous items in Natural Language Inference, for which annotators don't agree on the label. Ambiguous items are overlooked in the literature but often occurring in the real world. We build an ensemble model, AAs (Artificial Annotators), that simulates underlying annotation distribution to effectively identify such inherently ambiguous items. Our AAs are better at handling inherently ambiguous items since the model design captures the essence of the problem better than vanilla model architectures.
3. We follow a standard practice to build a robust dataset for FAQ retrieval task, `COUGH`. In our dataset analysis, we show how `COUGH` better reflects the challenge of FAQ retrieval in the real situation than its counterparts. The imposed challenge will push forward the boundary of research on FAQ retrieval in real scenarios.

Moving forward, the ultimate goal for robust natural language understanding is to build NLU models which can behave humanly. That is, it's expected that robust NLU systems are capable to transfer the knowledge from training corpus to unseen documents more reliably and survive when encountering challenging items even if the system doesn't know a priori of users' inputs.

Dedicated to my parents.

Acknowledgments

I feel incredibly fortunate to have Dr. Huan Sun as my advisor, without whom nothing in this thesis is possible. I would like to express my sincere gratitude to her for critiquing my work and my ideas in a constructive way. Her vision and rigorous research attitudes have shaped my thoughts. I am always indebted to her for guiding me all the way, for her contagious energy, for being so supportive and caring about students.

I owe a great debt of gratitude to Dr. Marie-Catherine de Marneffe for her countless help. I am super grateful for her many invaluable insights and suggestions on my work. She has been so generous with her time, reading, reviewing and commenting on many of my writings. I am always thankful for her positive encouragement and praise.

It's also my great privilege to collaborate with my friends, lab-mates at SunLab and my past mentor: Xiang Yue, Ziyu Yao, Heming Sun, Emmett Jesrani and Dr. Chen Chen.

I appreciate the help from anyone who helped me along my education journey: Hangzhou Jindu Tianchang Elementary School, Hangzhou Caihe Experimental School, Hangzhou Xuejun High School, Sichuan University and Ohio State University. I am especially thankful to my Chinese friends, who always compliment me and cheer me up no matter what.

Last but not the least, my deepest gratitude, without any doubt, goes to my parents, Rongchang and Hangjuan. They gave birth to me, raised me up, set good examples for me, and taught me tremendously many invaluable lessons. I wouldn't become who I am without their trust and support. All in all, thanks for their unconditional love and upbringing.

Vita

2021-	Ph.D. in Computer Science and Engineering, University of Michigan.
2018-2021	B.S. in Computer Science and Engineering & Industrial and Systems Engineering, The Ohio State University.
2016-2018	B.E. in Industrial Engineering, Sichuan University.

Publications

Research Publications

X. F. Zhang and M. de Marneffe. Identifying inherent disagreement in natural language inference. *In NAACL 2021.* 2021.

X. F. Zhang, H. Sun, X. Yue, E. Jesrani, S. Lin, and H. Sun. COUGH: A challenge dataset and models for COVID-19 FAQ retrieval. *arXiv preprint.* 2020.

X. Yue*, X. F. Zhang*, Z. Yao, S. Lin, and H. Sun. CliniQG4QA: Generating diverse questions for domain adaptation of clinical question answering. *In ML4H workshop at NeurIPS 2020.* 2020. (*equal contributions)

Fields of Study

Major Field: Computer Science and Engineering

Table of Contents

Page

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Natural Language Understanding (NLU)

Have you ever asked: “Siri, how is the weather today?”, “Cortana, what is the best spot for hiking in Columbus?” or ”Xiaoice, could you tell me how’s traffic outside?”. If so, you have experienced receiving a data-supported answer from your personalized AI assistant. A natural question that people would ask is how can the agent understand an utterance and intents and generate a relevant response. The answer is Natural Language Understanding.

Natural Language Understanding (NLU) is a branch of Natural Language Processing (NLP) in the area of Artificial Intelligence (AI) that uses intelligent computer software to understand texts that encode human knowledge. Some representative NLU applications (and there are way more) are: Automated Reasoning, Question Answering, Text Categorization, Large-scale Content Analysis, Information Retrieval and Textual Entailment. NLU is generally considered an AI-hard problem (i.e., a problem that is hard to be solved by AI systems) [Yampolskiy, 2013]. NLU is an AI-hard problem mainly because the nature of human language (e.g., ambiguity) makes NLU difficult. For example, given the following sentence “when the hammer hit the glass table, it shattered”,[1] humans know that it is the glass table that shattered but not the hammer. This is because our prior knowledge let us

[1] https://www.colorado.edu/earthlab/2020/02/07/what-natural-language-processing-and-why-it-hard.

know what glass is and that glass can shatter easily. However, coreference resolution is still a challenging task for NLU models, and thus, NLU systems still have difficulties figuring out which one of these two objects shatters.

Recent years have witnessed notable progress across various Natural Language Understanding tasks, especially after entering the deep learning era in 2012. Deep learning approaches quickly outperformed statistical learning methods by a large margin on many NLU tasks. As today, neural network-based NLP models have reached many new milestones (e.g., model performance comes close to or surpasses the level of non-expert humans) and have become the dominating approach for NLP tasks. Typical neural network-based NLP models/algorithms are RNN [Elman, 1990], LSTM [Hochreiter and Schmidhuber, 1997], GRU [Cho et al., 2014], Seq2Seq [Sutskever et al., 2014], attention mechanism [Luong et al., 2015] and Transformer [Vaswani et al., 2017]. Recently, pretrained language models, such as GPT [Radford et al., 2018] and BERT [Devlin et al., 2019b], have dramatically altered the NLP landscape and marked new records on the majority of NLU tasks. However, the neural NLP models work well for supervised tasks in which there is abundant labeled data for learning, but still perform poorly for low-resource and cross-domain tasks where the training data is insufficient and the test data is from different domains, respectively.

Besides more advanced model architectures, reliable and trustworthy datasets also play a huge role in improving NLU systems. Without a decent dataset, it would be challenging to train a machine learning model, not to mention carrying out a valid evaluation. As such, comprehensive evaluation benchmarks, aggregating datasets of multiple NLU tasks, emerged in the past few years such as GLUE [Wang et al., 2019b] and SuperGLUE [Wang et al., 2019a]. They are diagnostic datasets designed to evaluate and analyze model performance with respect to a wide range of linguistic phenomena found in human language.

The human ability of understanding natural language is flexible and robust. Therefore, human capability of understanding multiple language tasks simultaneously and transferring the knowledge to unseen documents is mostly reliable. On the contrary, most of existing NLU models built on word/character levels are exclusively trained on a restricted dataset. These restricted datasets normally only characterize one particular domain or only include simple examples which might not well reflect the task difficulties in reality. Consequently, such models usually fail to achieve desirable performance on out-of-domain data or struggle on handling challenging items (e.g., inherently ambiguous items, adversarial items) in the real world. Moreover, machine learning algorithms are usually data-hungry and can easily malfunction when there is insufficient amount of training data. Therefore, in order to have NLU models understand human language more effectively, it is expected to prioritize the study on robust natural language understanding.

1.2 Robustness Problem in NLU

In this thesis, we deem that NLU systems are consisting of two components: NLU models and NLU datasets. As such, we argue that, to achieve robust NLU, the model architecture/training and the dataset are equally important. If either component is weak, it would be hard to achieve full robustness. Therefore, in order to achieve full robustness in NLU, researchers are expected to implement robust models which then are trained on constructed robust datasets. In this thesis, we define robust models and robust datasets as follow:

1. **Robust models** are expected to be resistant to domain changes and resilient to challenging items (e.g., inherently ambiguous items, adversarial items).

2. **Robust datasets** are expected to reflect real-world challenges and encode knowledge that is difficult to be unraveled simply by surface-level[2] understanding.

In short, a truly robust NLU system is expected to be a robust model trained on robust datasets.

Three NLU tasks for NLU robustness problem

In the context of NLP, robustness is an umbrella term which could be interpreted differently from different angles. In this thesis, we will focus on three NLU tasks to illustrate the robustness problem in different NLU tasks and our contributions (i.e., novel models and new datasets) to help achieve more robust natural language understanding.

The first robustness problem that will be studied in this thesis is the cross-domain generalization.In Question Answering, most past work on open-domain were only testing models on in-domain data (*source* domain), despite outperforming human performance. However, these well-performing models have a relatively weak generalizability, which is the crux of this robustness problem. That is, when such models are deployed on out-of-domain data (*target* domain), their performances go down drastically, which is way behind human performance. Similar trend is also observed under the clinical setting where a model trained on one corpus may not generalize well to new clinical texts collected from different medical institutions [Yue et al., 2020a,b]. In Chapter 2, we will study how to utilize diversity boosters to help Question Generator (QG) synthesize diverse[3] QA pairs, upon which a Question Answering system is trained to improve the generalization to the unseen target domain. We also construct a target-domain test set to help evaluate models' generalizability.

[2]For example, the presence of "not" or "bad" doesn't always indicate a negative sentiment.

[3]"Diverse" here means questions with different syntactic structures or different topics.

The second robustness problem that will be studied in this thesis is how to better handle inherently ambiguous items, one type of challenging items in reality. In sentiment analysis and textual entailment tasks, it has been observed that there are inherently ambiguous/disagreement[4] items for which annotators have different annotations [Kenyon-Dean et al., 2018, Pavlick and Kwiatkowski, 2019, Zhang and de Marneffe, 2021]. These items were usually treated as noise and removed in the dataset construction phase, which is problematic. In Chapter 3, we will investigate inherently ambiguous items, which are overlooked in the literature but often occurring in the real world, in the NLI (Natural Language Inference) task. To this end, we build an ensemble model, AAs (Artificial Annotators), which simulates underlying annotation distribution by capturing the modes in annotations to effectively identify such inherently ambiguous items.

The third robustness problem that will be studied in this thesis is how to construct a reliable and challenging dataset (i.e., robust dataset). In textual entailment and FAQ retrieval tasks, common datasets (e.g., SNLI [Bowman et al., 2015] and MultiNLI [Williams et al., 2018] for textual entailment; FAQIR [Karan and Šnajder, 2016] and StackFAQ [Karan and Šnajder, 2018] for FAQ retrieval) used for training and testing might not well characterize the real difficulties of respective tasks. In the aforementioned datasets, sentence lengths and language complexities are generally low, styles are limited and the search space is small. In Chapter 4, we will follow a standard practice to build a robust dataset for the FAQ Retrieval task. In our dataset analysis, we will also show how this dataset better reflects the challenge of FAQ Retrieval in the real situation than its counterparts.

We will conclude with recommendations for future work about how to better approach robustness problem in NLU in Chapter 5.

[4]In this thesis, "ambiguous" and "disagreement" will be used interchangeably.

Chapter 2: Clinical Question Answering

2.1 Introduction

Clinical question answering (QA), which aims to automatically answer natural language questions based on clinical texts in Electronic Medical Records (EMR), has been identified as an important task to assist clinical practitioners [Patrick and Li, 2012, Raghavan et al., 2018, Pampari et al., 2018, Fan, 2019, Rawat et al., 2020]. Neural QA models in recent years [Chen et al., 2017, Devlin et al., 2019b] show promising results in this research. However, answering clinical questions still remains challenging in real-world scenarios because well-trained QA systems may not generalize well to new clinical contexts from a different institute or patient group. For example, Yue et al. [2020a] pointed out when a clinical QA model trained on the emrQA [Pampari et al., 2018] dataset is deployed to answer questions on MIMIC-III clinical texts [Johnson et al., 2016], its performance drops by around 30% even on questions that are similar to those in training.

Most of the existing clinical QA datasets and setups focus on in-domain testing while leaving the generalization challenge under-explored. In this chapter, we propose to evaluate *the performance of clinical QA models on target contexts and questions which may have different distributions from the training data*. Due to the lack of publicly-available clinical

QA pairs for our proposed evaluation setting, we ask clinical experts to annotate a new test set on the sampled MIMIC-III [Johnson et al., 2016] clinical texts.

Inspired by recent work on question generation (QG) for improving QA performance in the open domain [Golub et al., 2017, Wang et al., 2019c, Shakeri et al., 2020], we implement an answer evidence extractor and a seq2seq-based QG model to synthesize QA pairs on target contexts to train a QA model. However, we do not observe that such QA models achieve better performance on our curated MIMIC-III QA set, compared with that trained on emrQA. Our error analysis reveals that the automatic generation technique often falls short of generating questions that are *diverse* enough to serve as useful training data for clinical QA models.

To this end, we investigate two kinds of approaches to diversify the generation. Inspired by Ippolito et al. [2019] whio study various decoding-based methods, we pick the standard beam search as a representative of the decoding-based approach since it achieves satisfying performance in various generation tasks. On the other hand, another practice (topic-guided approach) is to have a diversification step followed by a conditional generation. In general, such techniques first decide question topics and then generate questions conditioned on selected topics [Kang et al., 2019, Cho et al., 2019, Liu et al., 2020]. Following the second approach, we propose a simple but effective question phrase prediction (QPP) module to diversify the generation. Specifically, QPP takes the extracted answer evidence as input and sequentially predicts potential question phrases (e.g., "What treatment", "How often") that signify what types of questions humans may ask about the answer evidence. Then, by directly forcing a QG model to produce specified question phrases at the beginning of the question generation process (both in training and inference), QPP enables diverse questions to be generated.

Through comprehensive experiments, we demonstrate that when using QA pairs automatically synthesized by diverse QG, especially by the QPP-enhanced QG, we are able to boost QA performance by 4.5%-9% in terms of Exact Match (EM), compared with their counterparts directly trained on the source QA dataset (i.e., emrQA).

2.2 Out-of-Domain Test Set

Unlike open domain, there are very few publicly available QA datasets in the clinical domain. EmrQA dataset [Pampari et al., 2018], which was generated based on medical expert-made question templates and existing annotations on n2c2 challenge datasets [n2c2, 2006], is a commonly adopted dataset for clinical reading comprehension.

However, all the QA pairs in emrQA are based on n2c2 clinical texts and thus not suitable for our generalization setting. Yue et al. [2020a] studied a similar problem and annotated a test set on MIMIC-III clinical texts [Johnson et al., 2016]. However, their test set is too small (only 50 QA pairs) and not publicly available. Given the lack of a reasonably large clinical QA test set for studying generalization, with the help of three clinical experts, we create 1287 QA pairs on a sampled set of MIMIC-III [Johnson et al., 2016] clinical notes, which have been reviewed and approved by PhysioNet.[5]

Annotation Process. We sample 36 MIMIC-III clinical notes[6] as contexts. For each context, clinical experts can ask any questions as long as an answer can be extracted from the context. To save annotation effort, QA pairs generated by 9 QG models (i.e., all base QG models and their diversity-enhanced variants; see Section 2.5.1) are provided as references, and (nearly)

[5]https://physionet.org/. PhysioNet is a resource center with missions to conduct and catalyze for biomedical research, which offers free access to large collections of physiological and clinical data, such as MIMIC-III [Johnson et al., 2016].

[6]When sampling MIMIC-III notes, we ensure that all the sampled clinical texts do not appear in emrQA, acknowledging that there is a small overlap between the two datasets.

(Question / Context)	emrQA	MIMIC-III
# Train	781,857 / 337	- / 337
# Dev	86,663 / 41	8,824 / 40
# Test	98,994 / 42	1,287 / 36
# Total	967,514 / 420	- / 413
for purpose of	QG & QA (*source* domain)	QA (*target* domain)

Table 2.1: Statistics of the datasets. We synthesize a machine-generated dev set and ask human experts to annotate a test set for MIMIC-III. Details of dev set construction can be found in Setion A.1.3.

duplicates are removed. Meanwhile, clinical experts are highly encouraged to create new questions based on the given clinical text (which are marked as *"human-generated"*). But if they do find that the machine-generated questions sound natural and match the provided answer, they can keep them (which are marked as *"human-verified"*). After obtaining the annotated questions, we ask another clinical expert to do a final pass of the questions in order to further ensure the quality of the test set. The final test set consists of 1287 questions (of which 975 are *"human-verified"* and 312 are *"human-generated"*).

In the following sections, we consider emrQA as the *source* dataset and our annotated MIMIC-III QA dataset as the *target* data. Detailed statistics of the two datasets are given in Table 2.1.

2.3 Framework

2.3.1 Overview of Our Framework

We first give an overview of our framework without including any diversity booster.

To solve the proposed generalization challenge of clinical QA, inspired by recent work on question generation (QG) for QA in the open domain [Golub et al., 2017, Wang et al.,

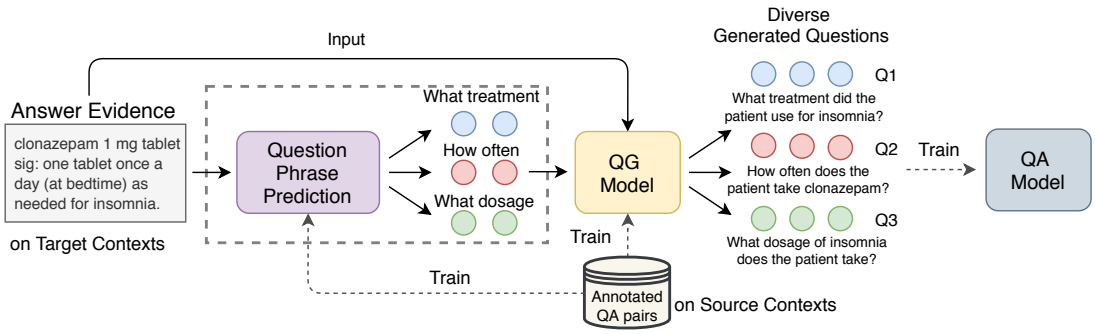

Figure 2.1: Illustration of our framework equipped with QPP: A key component is our question phrase prediction (QPP) module, which aims to generate diverse question phrases and can be "plugged-and-played" with most existing QG models to diversify their generation.

2019c, Shakeri et al., 2020], we implement an answer evidence extractor and a seq2seq-based neural QG model [Du et al., 2017, NQG] to synthesize QA pairs on target contexts. Specifically, given a document, we deploy a ClinicalBERT [Alsentzer et al., 2019] model to extract a long text[7] span as an answer evidence. We formulate such span prediction problem as a `BIO` tagging task. After prediction, we develop some heuristic rules (e.g., removing/merging very short extracted evidences) to further improve the quality of the extracted evidences; more details are listed in Appendix A.1.1. Based on the extracted answer evidences, a seq2seq-based QG model can be used to generate questions. Both answer evidence extractor and QG model are trained on the source data and then used to synthesize QA pairs on target contexts, based on which a QA model can be trained.

2.3.2 Preliminary Observation

To our surprise, training on the synthesized target-context QA pairs does not yield an improvement of QA on the constructed MIMIC-III QA set. Specifically, F1 is 79.43 for the QA model trained on corpus synthesized by NQG (neural question generation) model, which

[7]Following Pampari et al. [2018], Yue et al. [2020a], we focus on long text spans instead of short answers since the former often contain richer information, which is more useful to support clinical decision making.

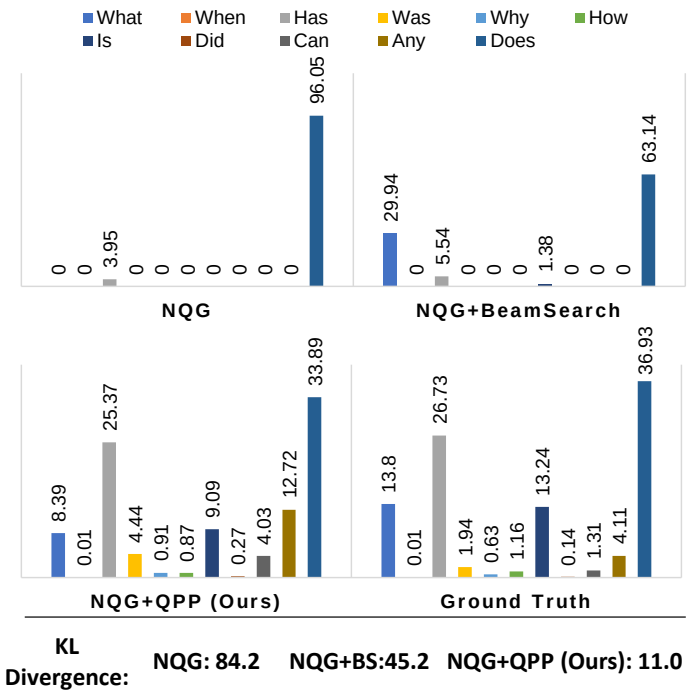

Figure 2.2: Distributions over types of questions generated by NQG models and the ground truth. BS: Beam Search; QPP: Question Phrase Prediction module.

is a little inferior to directly training the QA model on emrQA (79.99 F1). An outstanding characteristic we observe in the generated questions is the large bias of question types (e.g., most questions are "Does" while there is few "Why" and no "How" question). The distributions of question types are in Figure 2.2 (see top-left sub-plot).

2.4 Diverse Question Generation for QA

Given the observation above, we argue that the synthetic questions should be diverse so that they could serve as more useful training corpora.

2.4.1 Overview of Diverse Question Generation

We investigate two kinds of approaches to diversify the generation. In the first decoding-based approach, we select the standard beam search as the representative since it is well

Algorithm 1 Training procedure of our framework equipped with QPP.

Input: labeled *source* data $\{(P_S, A_S, Q_S)\}$, unlabeled *target* data $\{P_T\}$
Output: Generated QA pairs $\{(A'_T, Q'_T)\}$ on *target* contexts; An optimized QA model for answering questions on target contexts

Pretraining Stage

1: Train *Answer Evidence Extractor* based on the *source* data $\{(P_S, A_S)\}$
2: Obtain question phrase data Y_S from Q_S and train *Question Phrase Prediction* module on the *source* data $\{(A_S, Y_S)\}$
3: Train a *QPP-enhanced QG* model on the *source* data $\{(A_S, Y_S, Q_S)\}$

Training Stage

4: Use *AEE* to extract potential answer evidences $\{A'_T\}$ on the *target* contexts $\{P_S\}$
5: Use *QPP* to predict potential question phrases set $\{Y'_T\}$ on $\{A'_T\}$
6: Use *QPP-enhanced QG* to generate diverse questions $\{Q'_T\}$ based on $\{(A'_T, Y'_T)\}$
7: Train a *QA* model on synthetic *target* data $\{(P_T, A'_T, Q'_T)\}$

studied and shows competitive performance in diversifying generations [Ippolito et al., 2019]. For the other kind (topic-guided approach), we propose a *question phrase prediction (QPP)* module, which predicts a set of valid question phrases given an answer evidence (Figure 2.1). Then, conditioned on a question phrase sampled from the set predicted by the QPP, a QG model is utilized to complete the rest of the question.

2.4.2 Question Phrase Prediction (QPP)

We formulate the question phrase prediction task as a *sequence prediction* problem and adopt a commonly used seq2seq model [Luong et al., 2015]. More formally, given an answer evidence $\mathbf{a}$, QPP aims to predict a sequence of question phrases $\mathbf{s} = (s_1, ..., s_{|\mathbf{s}|})$(e.g., "What treatment" (s_1) $\rightarrow$ "How often" (s_2) $\rightarrow$ "What dosage" (s_3), with $|\mathbf{s}| = 3$).

During training, we assume that the set of question phrases is arranged in a pre-defined order. Such orderings can be obtained with some heuristic methods, e.g., using a descending order based on question phrase frequency in the corpus[8] (more details are in Appendix A.1.2). As such, we aim to minimize:

[8]In emrQA, each answer evidence is tied with multiple questions, which allows the training for QPP.

$$L_{QPP} = -\sum \log P(\mathbf{s}|\mathbf{a};\theta) \tag{2.1}$$

where $\mathbf{s}, \mathbf{a}, \theta$ denote question phrase sequence, input answer evidence and all the parameters in QPP, respectively. Algorithm 1 illustrates the pretraining and training procedure of our framework when equipped with our proposed QPP module.

In the inference stage, QPP can dynamically decide the number of question phrases for each answer evidence by predicting a special `[STOP]` type. By decomposing QG into two steps (diversification followed by generation), the proposed QPP can increase the diversity in a more controllable way compared with decoding-based approach.

2.5 Evaluation and Results

2.5.1 Experiment Setup

Base QG and QA models: In our experiments, we adopt three base QG models: NQG [Du et al., 2017], NQG++ [Zhou et al., 2017] and BERT-SQG [Chan and Fan, 2019]. For QA, we use two base models, DocReader [Chen et al., 2017] and ClinicalBERT [Alsentzer et al., 2019].

To investigate the effectiveness of diverse QG for QA, we consider the following variants of each base QG model: (1) Base Model: Inference with greedy search; (2) Base Model + Beam Search: Inference with Beam Search with the beam size at K and keep top K beams (we set $K = 3$) (3) Base Model + QPP: Inference with greedy search for both QPP module and Base model.

When training a QA model, we only use the synthetic data on the target contexts and do not combine the synthetic data with the source data since the combination does not help in our preliminary experiments.

QA Datasets	**DocReader** [Chen et al., 2017]						**ClinicalBERT** [Alsentzer et al., 2019]					
	Human Verified		**Human Generated**		**Overall Test**		**Human Verified**		**Human Generated**		**Overall Test**	
	EM	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**	**EM**	**F1**
emrQA [Pampari et al., 2018]	61.44	78.82	69.87	83.66	63.48	79.99	61.23	78.56	69.23	82.83	63.17	79.59
NQG [Du et al., 2017]	64.71	79.36	66.99	79.67	65.26	79.43	59.49	76.68	67.3	82.59	61.38	78.11
+ BeamSearch	67.07	81.21	71.15	83.07	68.07	81.66	63.17	79.17	68.91	84.26	64.56	80.4
+ QPP (Ours)	**68.82**	**82.89**	**74.68**	**85.18**	**70.09**	**83.44**	**63.79**	**79.56**	**69.23**	**84.33**	**65.11**	**80.72**
NQG++ [Zhou et al., 2017]	65.94	78.71	66.34	81.34	66.04	79.35	59.59	75.85	65.06	80.11	60.92	76.88
+ BeamSearch	68.10	80.09	72.11	84.56	69.07	81.17	64.61	80.30	68.26	83.70	65.50	81.12
+ QPP (Ours)	**70.05**	**83.47**	**74.36**	**85.92**	**71.10**	**84.06**	**65.33**	**80.64**	**70.83**	**85.76**	**66.67**	**81.88**
BERT-SQG [Chan et al., 2019]	66.05	79.64	70.19	81.47	67.05	80.08	59.59	78.04	65.06	82.20	60.92	79.05
+ BeamSearch	68.71	81.98	73.71	84.44	69.93	82.58	61.94	79.02	67.31	82.54	63.25	79.88
+ QPP (Ours)	**70.77**	**83.60**	**74.36**	**85.53**	**71.64**	**84.07**	**64.21**	**80.53**	**69.23**	**85.38**	**65.43**	**81.71**

Table 2.2: QA performance on MIMIC-III test set. emrQA is also included as a baseline dataset to illustrate that the generated diverse questions on MIMIC-III are useful to improve the QA model performance on new contexts.

Evaluation Metrics: For QG evaluation, we focus on evaluating both relevance and diversity. Following previous work [Du et al., 2017, Zhang et al., 2018], we use BLEU [Papineni et al., 2002], ROUGE-L [Lin, 2004] as well as METEOR [Lavie and Denkowski, 2009] for relevance evaluation. Since the Beam Search and our QPP module enable QG models to generate multiple questions given an evidence, we report the top-1 relevance among the generated questions following Cho et al. [2019]. For diversity, we report Distinct [Li et al., 2016] as well as Entropy [Zhang et al., 2018] scores. We calculate BLEU and the diversity measures based on 3- and 4-grams.

For QA evaluation, we report exact match (EM) (the percentage of predictions that match the ground truth answers exactly) and F1 (the average overlap between the predictions and ground truth answers) as in Rajpurkar et al. [2016].

2.5.2 Results

Table 2.2 summarizes the performance of two widely used QA models, DocReader [Chen et al., 2017] and ClinicalBERT [Alsentzer et al., 2019], on the MIMIC-III test set. The QA models are trained on different corpora, including the emrQA dataset as well as QA pairs generated by different models.

Models	Relevance				Diversity			
	BLEU3	**BLEU4**	**MR**	**RG**	**Dist3**	**Dist4**	**Ent3**	**Ent4**
NQG [Du et al., 2017]	91.45	90.11	60.70	94.62	0.233	0.282	4.473	4.738
+ BeamSearch	94.33	93.42	62.08	95.56	0.569	0.775	5.406	5.812
+ QPP (Ours)	**96.82**	**96.33**	**64.38**	**97.49**	**3.177**	**5.289**	**7.100**	**7.777**
NQG++ [Zhou et al., 2017]	97.11	96.65	71.57	97.86	0.229	0.275	4.419	4.648
+ BeamSearch	98.35	98.07	72.98	98.55	0.618	0.848	5.497	5.953
+ QPP (Ours)	**99.15**	**99.03**	**74.01**	**99.11**	**3.183**	**5.293**	**7.111**	**7.798**
BERT-SQG [Chan and Fan, 2019]	89.07	87.99	65.25	94.91	0.228	0.276	4.594	4.849
+ BeamSearch	95.45	94.84	66.39	96.22	0.510	0.713	5.522	6.015
+ QPP (Ours)	**96.54**	**96.19**	**67.51**	**97.42**	**3.344**	**5.332**	**7.173**	**7.816**

Table 2.3: Automatic evaluation of the generated questions on emrQA dataset. For each base model, the best performing variant is in **bold**. RG: ROUGE-L, MR: METEOR, Dist: Distinct, Ent: Entropy.

We also evaluate QG models on the emrQA dataset (i.e., train and test QG solely on source domain). As can be seen from Table 2.3, the three selected base models (NQG, NQG++ and BERT-SQG) all achieve very promising relevance scores; however, they do not perform well with diversity scores. The diversity of generated questions is boosted to some extent when the Beam Search is used since it can offer flexibility for QG models to explore more candidates when decoding. In comparison, the QPP module in our framework leads to the best results under both relevance and diversity evaluation. Particularly, it obtains 5% absolute improvement in terms of Dist4 for each base model.

2.6 Analysis

2.6.1 Quantitative Analysis

Analysis on QA Generalization: As expected, the corpora generated by diverse QG help the QA model perform consistently better than those generated by their respective base QG version as well as emrQA (Table 2.2). Between the two diversity-boosting approaches, we observe that the QA model trained on the corpora by QPP-enhanced QG achieves the best performance. Moreover, results on the human-generated portion are consistently better than those on human-verified. This is likely due to the fact that human-generated questions

QA Example from MIMIC-III	QG Example from MIMIC-III
Context: ... he was guaiac negative on admission. hematocrit remained stable overnight. **5. abd pain: suspect secondary to chronic pancreatitis.** amylase unchanged	**Context:** ... the patient was taking at home prior to admission were not restarted. **25. acetaminophen 325-650 mg po/ng q6h:prn pain** 26. dabigatran etexilate 150 mg po bid...
Question: Why did the patient get abd pain? **Answer by QA model trained on** -*emrQA:* 5. abd pain -*NQG Generated:* 5. abd pain: -*NQG+BeamSearch:* 5. abd pain: ***-NQG+QPP:*** 5. abd pain: suspect secondary to chronic pancreatitis**.**	**Questions generated by** -*NQG:* **Does** the patient have any pain? -*NQG+BeamSearch:* **Does** the patient have any pain history? **Does** the patient have pain? **Does** the patient have any pain? ***-NQG+QPP:*** **Why** did the patient have acetaminophen? **What** treatment has the patient had for his pain? **How** was pain treated? **Does** the patient have any pain? ...

Figure 2.3: QA and QG examples. The red parts in contexts are ground-truth answer evidences.

are more readable and sensible while human-verified ones are less natural (though the correctness is ensured). All these results indicate that improving the diversity of generated questions can help better train QA models on the new contexts and better address the generalization challenge.

Analysis on QG diversity: Figure 2.2 shows the distribution over types of questions generated by NQG-based models (i.e., base model, base + beam search and base + QPP) and the ground truth on emrQA dataset. We observe that the Kullback–Leibler (KL) divergence between the distributions of generated questions and the ground truth is smaller after enabling diversity booster. The gap reaches the minimum when our QPP module is plugged in. It's worth noting that even some of the least frequent types of questions (e.g., "How", "Why") can be generated when our QPP module is turned on. These observations demonstrate diversity booster, especially our QPP module, can help generate diverse questions.

2.6.2 Qualitative Analysis: Error Analysis

In Figure 2.3, we first present a QA example and a QG example from MIMIC-III. In the QA example, this "why" question can be correctly answered by the QA model (DocReader) trained on the "NQG+QPP" generated corpus while the QA models trained on other generated corpora fail. This is because the NQG model and "NQG+BeamSearch"

cannot generate any "why" questions as shown in Figure 2.2. Thus QA models trained on such corpora cannot answer questions of less frequent types. Though the emrQA dataset contains diverse questions (including "why" questions), its contexts might be different from MIMIC-III in terms of topic, note structures, writing styles, etc. So the model trained on emrQA struggles to answer some questions. In the QG example, the base model NQG can only generate one question. Though utilizing the Beam Search enables the model to explore multiple candidates, the generated questions are quite similar and are less likely to help improve QA. Enabling our QPP module helps generate diverse questions including "Why", "What", "How", etc.

2.7 Conclusion

In this chapter, we systematically investigate the generalization challenge of clinical reading comprehension and construct a new test set on MIMIC-III clinical texts. After observing simply using QG for QA does not work, we explore the importance of generating *diverse* questions. That is, we study two approaches for boosting question diversity, beam search and QPP. Particularly, our proposed QPP (question phrase prediction) module significantly improves the cross-domain generalizability of QA systems. Our comprehensive experiments allow for a better understanding of why diverse question generation can help QA on new clinical documents (i.e., target domain).

Chapter 3: Natural Language Inference

3.1 Introduction

Natural language inference (NLI)[9] is the problem of determining whether a natural language hypothesis *h* can be inferred (or entailed) from a natural language premise *p* [i.a., Dagan et al., 2005, MacCartney and Manning, 2009]. Conventionally, people only examine items that are suitable for systematic inferences (i.e., items for which people consistently agree on the NLI label).

However, Pavlick and Kwiatkowski [2019] observed inherent disagreements among annotators in several NLI datasets (e.g., SNLI [Bowman et al., 2015]), which cannot be smoothed out by hiring more people. They pointed out that to achieve robust NLU, we need to be able to tease apart systematic inferences (i.e., items for which most people agree on the annotations) from items inherently leading to disagreement. The last example in Table 3.1 is a typical disagreement item: some annotators consider it to be an entailment (3 or 2), while others view it as a contradiction (-3). Clearly, the annotators have two different interpretations on the complement clause "If she'd said Carolyn had borrowed a book from Clare and wanted to return it". Moreover, a common practice in the literature to generate an inference label from annotations is to take the average [i.a., Pavlick and Callison-Burch, 2016]. In this case, it would be "Neutral", but such label is not accurately capturing the

[9]In this thesis, we use "textual entailment" and "Natural Language Inference" or "NLI" interchangeably.

1	*Premise:* Some of them, like for instance the farm in Connecticut, are quite small. If I like a place I buy it. I guess you could say it's a hobby. *Hypothesis:* buying places is a hobby. `Entailment (Entailment)` [3, 3, 2, 2, 2, 2, 1, 1]
2	*Premise:* "I hope you are settling down and the cat is well." This was a lie. She did not hope the cat was well. *Hypothesis:* the cat was well. `Neutral (Neutral)` [0, 0, 0, 0, 0, 0, 0, 0, -3]
3	*Premise:* "All right, so it wasn't the bottle by the bed. What was it, then?" Cobalt shook his head which might have meant he didn't know or might have been admonishment for Oliver who was still holding the bottle of wine. *Hypothesis:* Cobalt didn't know. `Neutral (Disagreement)` [1, 0, 0, 0, 0, 0, 0, -2]
4	*Premise:* A: No, it doesn't. B: And, of course, your court system when you get into the appeals, I don't believe criminal is in a court by itself. *Hypothesis:* criminal is in a court by itself. `Contradiction (Contradiction)` [-1, -1, -2, -2, -2, -2, -2, -3]
5	*Premise:* A: The last one I saw was Dances With The Wolves. B: Yeah, we talked about that one too. And he said he didn't think it should have gotten all those awards. *Hypothesis:* Dances with the Wolves should have gotten all those awards. `Contradiction (Disagreement)` [0, 0, -1, -1, -2, -2, -2, -3]
6	*Premise:* Meg realized she'd been a complete fool. She could have said it differently. If she'd said Carolyn had borrowed a book from Clare and wanted to return it they 'd have given her the address. *Hypothesis:* Carolyn had borrowed a book from Clare. `Disagreement (Disagreement)` [3, 3, 3, 2, 0, -3, -3, -3]

Table 3.1: Examples from CommitmentBank, with finer-grained NLI labels. The labels in parentheses come from Jiang and de Marneffe [2019a]. Scores in brackets are the raw human annotations.

distribution. Alternatively, some work simply ignored the "Disagreement" portion but only studied systematic inferences items [Jiang and de Marneffe, 2019b,a, Raffel et al., 2019].

Kenyon-Dean et al. [2018] also pointed out in sentiment analysis task, when performing real-time sentiment classification, an automated system cannot know a priori whether the data sample is inherently non-ambiguous. Here, in line with what Kenyon-Dean et al. [2018] suggested for sentiment analysis, we propose a finer-grained labeling for NLI: teasing disagreement items, labeled "Disagreement", from systematic inferences, which can be "Contradiction", "Neutral" or "Entailment". As such, in order to achieve robust NLU in NLI task, the developed models should be able to identify inherent disagreement items when possible and carry out systematic inferences on non-disagreement items.

To this end, we propose Artificial Annotators (AAs), an ensemble of BERT models [Devlin et al., 2019a], which simulate the uncertainty in the annotation process by capturing

modes in annotations. That is, we expect to utilize simulated modes of annotations to enhance finer-grained NLI label prediction. Our results, on the CommitmentBank, show that AAs perform statistically significantly better than all baselines (including BERT baselines) by a large margin in terms of both F1 and accuracy. We also show that AAs manage to learn linguistic patterns and context-dependent reasoning.

3.2 Inherently Ambiguous Items in CB

We start with the introduction to the dataset used in this chapter, CommitmentBank [de Marneffe et al., 2019], and then move on to how we determine ambiguous items and systematic inference items.

The CommitmentBank (CB) is a corpus of 1,200 naturally occurring discourses originally collected from news articles, fiction and dialogues. Each discourse consists of up to 2 prior context sentences and 1 target sentence with a clause-embedding predicate under 4 embedding environments (negation, modal, question or antecedent of conditional). Annotators judged the extent to which the speaker/author of the sentences is committed to the truth of the content of the embedded clause (CC), responding on a Likert scale from +3 to -3, labeled at 3 points (+3/speaker is certain the CC is true, 0/speaker is not certain whether the CC is true or false, -3/speaker is certain the CC is false). Following Jiang and de Marneffe [2019a], we recast CB by taking the context and target as the premise and the embedded clause in the target as the hypothesis.

Common NLI benchmark datasets are SNLI [Bowman et al., 2015] and MultiNLI [Williams et al., 2018], but these datasets have only one annotation per item in the training set. CB has at least 8 annotations per item, which permits to identify items on which annotators disagree. Jiang and de Marneffe [2019a] discarded items if less than 80% of the

	Entailment	Neutral	Contradiction	Disagreement	Total
Train	177	57	196	410	840
Dev	23	9	22	66	120
Test	58	19	54	109	240
Total	258	85	272	585	1,200

Table 3.2: Number of items in each class in train/dev/test.

annotations are within one of the following three ranges: [1,3] Entailment, 0 Neutral, [-3,-1] Contradiction. The gold label for example 3 in Table 3.1 would thus be "Disagreement". However, this seems a bit too stringent, given that 70% of the annotators all agree on the 0 label and there is only one annotation towards the extreme. Likewise, for example 5, most annotators chose a negative score and the item might therefore be better labeled as "Contradiction" rather than "Disagreement". To decide on the **finer-grained NLI labels**, we therefore also took variance and mean into account, as follows:[10]

- **Entailment:** 80% of annotations fall in the range [1,3] OR the annotation variance $\leq$ 1 and the annotation mean $>$ 1.
- **Neutral:** 80% of annotations is 0 OR the annotation variance $\leq$ 1 and the absolute mean of annotations is bound within 0.5.
- **Contradiction:** 80% of annotations fall in the range [-3, -1] OR the annotation variance $\leq$ 1 and the annotation mean $<$ -1.
- **Disagreement:** Items which do not fall in any of the three categories above.

We randomly split CB into train/dev/test sets in a 7:1:2 ratio.[11] Table 3.2 gives splits' basic statistics.

[10]Compared with the labeling scheme in Jiang and de Marneffe [2019a], our labeling scheme results in 59 fewer Disagreement items, 48 of which are labeled as Neutral.

[11]We don't follow the SuperGLUE splits [Wang et al., 2019a] as they do not include disagreement items.

3.3 Linguistic Rules

Our developed linguistic rules are inspired by and adapted from Jiang and de Marneffe [2019a] to explicitly include the most discriminating expressions for disagreement items. We utilize three linguistic features which are provided in CB: entailment-canceling environment (negation, modal, question, antecedent of conditional), matrix verb and its subject person.

1. Items under conditional are disagreement.
2. Items under question and with second person are neutral.
3. Items under question and with non-second person are disagreement.
4. Items of the form "I don't know/think/believe" are contradiction (i.e., negRaising structure).
5. Items with factive verbs are entailment.
6. Items under negation and with non-factive verbs are disagreement.
7. Items under modal and with non-third person are entailment.

When this policy is executed, there are two additional auxiliary rules: Items not falling in any group above are assigned a disagreement label as it is the dominant class in CB; For items satisfying more than one rule, the label will be determined by the higher-ranked rule (a smaller number indicates a higher rank). Note that the rules above also reveal the most discriminating expressions for each class.

3.4 Artificial Annotators

We aim at finding an effective way to tease items leading to systematic inferences apart from items leading to disagreement. As pointed out by Calma and Sick [2017], annotated labels are subject to uncertainty. Annotations are indeed influenced by several factors: workers' past experience and concentration level, cognition complexities of items, etc. They

proposed to simulate the annotation process in an active learning paradigm to make use of the annotations that contribute to uncertainty. Likewise, for NLI, Gantt et al. [2020] observed that directly training on raw annotations using annotator identifier improves performance. Essentially, Gantt et al. [2020] used a mixed-effect model to learn a mapping from an item and the associated annotator identifier to a NLI label. However, annotator identifiers are not always accessible, especially in many datasets that have been there for a while. Thus, we decide to simulate the annotation process instead of learning from real identifiers.

As shown by Pavlick and Kwiatkowski [2019], if annotations of an item follow unimodal distributions, then it is suitable to use aggregation (i.e., take an average) to obtain a inference label; but such an aggregation is not appropriate when annotations follow multi-modal distributions. Without loss of generality, we assume that items are associated with n-modal distributions, where n $\geq$ 1. Usually, systematic inference items are tied to unimodal annotations while disagreement items are tied to multi-modal annotations. We, thus, introduce the notion of Artificial Annotators (AAs), where each individual "annotator" learns to model one mode.

3.4.1 Architecture

AAs is an ensemble of n BERT models [Devlin et al., 2019a] with a primary goal of finer-grained NLI label prediction. n is determined to be 3 as there are up to 3 relationships between premise and hypothesis, excluding the disagreement class. Within AAs, each BERT is trained for an auxiliary systematic inference task which is to predict entailment/neutral/contradiction based on a respective subset of annotations. The subsets of annotations for the three BERT are mutually exclusive.

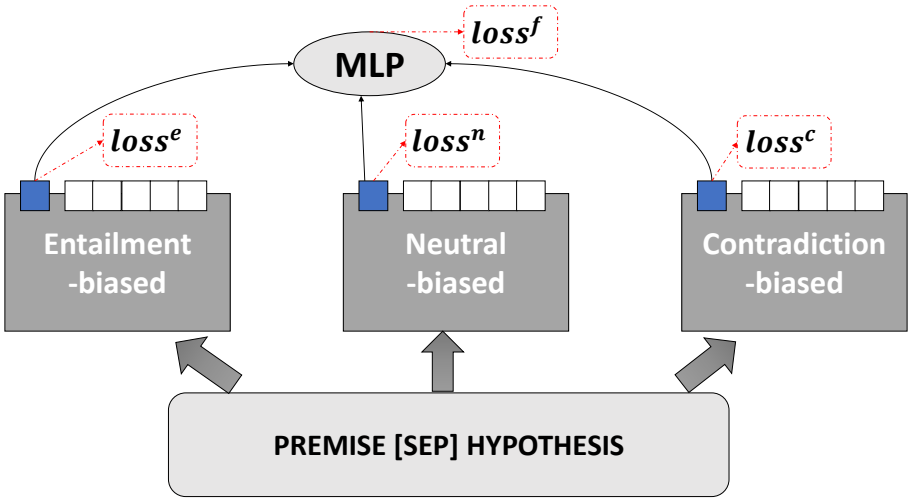

Figure 3.1: Artificial Annotators (AAs) setup.

A high-level overview of AAs is shown in Figure 3.1. Intuitively, each BERT separately predicts a systematic inference label, each of which represents a mode[12] of the annotations. The representations of these three labels are further aggregated as augmented information to enhance final fine-grained NLI label prediction (see Eq. 3.1).

If we view the AAs as a committee of three members, our architecture is reminiscent of the Query by Committee (QBC) [Seung et al., 1992], an effective approach for active learning paradigm. The essence of QBC is to select unlabeled data for labeling on which disagreement among committee members (i.e., learners pre-trained on the same labeled data) occurs. The selected data will be labeled by an oracle (e.g., domain experts) and then used to further train the learners. Likewise, in our approach, each AA votes for an item independently. However, the purpose is to detect disagreements instead of using disagreements as a measure to select items for further annotations. Moreover, in our AAs, the three members are trained on three disjoint annotation partitions for each item (see Section 3.4.2).

[12]It's possible that three modes collapse to (almost) a point.

3.4.2 Training

We first sort the annotations in descending order for each item and divide them into three partitions.[13] For each partition, we generate an auxiliary label derived from the annotation mean. If the mean is greater/smaller than +0.5/-0.5, then it's entailment/contradiction; otherwise, it's neutral. The first BERT model is always enforced to predict the auxiliary label of the first partition to simulate an entailment-biased annotator. Likewise, the second and third BERT models are trained to simulate neutral-biased and contradiction-biased annotators.

Each BERT produces a pooled representation for the `[CLS]` token. The three representations are passed through a multi-layer perceptron (MLP) to obtain the finer-grained NLI label:

$$P(y|\mathbf{x}) = \text{softmax}(\mathbf{W_s} \tanh(\mathbf{W_t}[\mathbf{e}; \mathbf{n}; \mathbf{c}])) \tag{3.1}$$

with $[\mathbf{e}; \mathbf{n}; \mathbf{c}]$ being the concatenation of three learned representations out of **e**ntailment-biased, **n**eutral-biased and **c**ontradiction-biased BERT models. $\mathbf{W_s}$ and $\mathbf{W_t}$ are parameters to be learned.

The overall loss is defined as the weighted sums of four cross-entropy losses:

$$loss = r * loss^f + \frac{1-r}{3}(loss^e + loss^n + loss^c) \tag{3.2}$$

where $r \in [0, 1]$ controls the primary finer-grained NLI label prediction task loss ratio.

3.5 Evaluation and Results

Evaluation Setting: We include five baselines to compare with:

- **"Always 0"**: Always predict Disagreement.

[13]For example, if there are 8 annotations for a given item, the annotations are divided into partitions of size 3, 2 and 3.

	Dev		Test					
	Acc.	F1	Acc.	F1	Entail	Neutral	Contradict	Disagree
Always 0	55.00	39.03	45.42	28.37	0.00	0.00	0.00	62.46
CBOW	55.25	40.54	45.09	28.37	0.00	0.00	0.69	62.17
Heuristic	65.00	62.08	54.17	50.60	22.54	**52.94**	64.46	58.20
Vanilla BERT	63.71	63.54	62.50	61.93	59.26	49.64	69.09	61.93
Joint BERT	64.47	64.28	62.61	62.07	59.77	47.27	67.36	63.21
AAs (ours)	**65.15**	**64.41**	**65.60***	**64.97***	**61.07**	51.27	**70.89**	**66.49***

Table 3.3: Baselines and AAs overall performance on CB dev and test sets, and F1 scores of each class on the test set (average of 10 runs). * indicates statistically significant difference (t-test, $p \leq 0.01$).

- **CBOW** (Continuous Bags of Words): Each item is represented as the average of its tokens' GLOVE vectors [Pennington et al., 2014].
- **Heuristic baseline**: Linguistics-driven rules (detailed out in chapter 3.3), adapted from Jiang and de Marneffe [2019a]; e.g., conditional environment discriminates for disagreement items.
- **Vanilla BERT**: [Devlin et al., 2019a] Straightforwardly predict among 4 finer-grained NLI labels.
- **Joint BERT**: Two BERT models are jointly trained, each of which has a different speciality. The first one (2-way) identifies whether a sentence pair is a disagreement item. If not, this item is fed into the second BERT (3-way) which carries out systematic inference.

For all baselines involving BERT, we follow the standard practice of concatenating the premise and the hypothesis with `[SEP]`.

Results: Table 3.3 gives the accuracy and F1 for each baseline and AAs, on the CB dev and test sets. We run each model 10 times, and report the average. Also, Our AAs achieve the lowest standard deviations on test set items compared to BERT-based models, indicating that it is more stable and potentially more robust to wild environments.

3.6 Analysis

3.6.1 Empirical Results Analysis

CBOW is essentially the same as the "Always 0" baseline as it keeps predicting Disagreement regardless of the input. The Heuristic baseline achieves competitive performance on the dev set, though it has a significantly worse result on the test set. Not surprisingly, both BERT-based baselines outperform the Heuristic on the test set: fine-tuning BERT often lead to better performance, including for NLI [Peters et al., 2019, McCoy et al., 2019]. These observations are consistent with Jiang and de Marneffe [2019a] who observed a similar trend, though only on systematic inferences. Our proposed AAs perform consistently better than all baselines, and statistically significantly better on the test set (t-test, $p \leq 0.01$).

Table 3.3 also gives F1 for each class on the test set. AAs outperform all BERT-based models under all classes. However, compared with the Heuristic, AAs show an inferior result on "Neutral" items mainly due to the lack of "Neutral" training data. The first 4 examples in Table 3.4 show examples for which AAs make the correct prediction while other baselines might not. The confusion matrix in Table 3.5 shows that the majority ($\sim$60%) of errors come from wrongly predicting a systematic inference item as a disagreement item. In 91% of such errors, AAs predict that there is more than one mode for the annotation (i.e., the three labels predicted by individual "annotators" in AAs are not unanimous), as in example 5 in Table 3.4. AAs are thus predicting more modes than necessary when the annotation is actually following a uni-modal distribution. On the contrary, when the item is supposed to be a disagreement item but is missed by AAs (as in example 6 and 7 in Table 3.4), AAs mistakenly predict that there is only one mode in the annotations 78% of the time. It thus seems that a method which captures accurately the number of modes in the annotation distribution would lead to a better model.

1	*Premise:* B: Yeah, it is. A: For instance, B: I'm a historian, and my father had kept them, I think, since nineteen twenty-seven uh, but he burned the ones from twenty-seven to fi-, A: My goodness. B: I could not believe he did that, *Hypothesis:* his father burned the ones from twenty-seven Heuristics: C V. BERT: D J. BERT: E AAs: E {E, E, E} Gold: E [3, 3, 3, 3, 3, 2, 2, -1]
2	*Premise:* 'She was about to tell him that was his own stupid fault and that she wasn't here to wait on him - particularly since he had proved to be so inhospitable. But she bit back the words. Perhaps if she made herself useful he might decide she could stay - for a while at least just until she got something else sorted out. *Hypothesis:* she could stay Heuristics: D V. BERT: D J. BERT: D AAs: N {N, N, N} Gold: N [3, 0, 0, 0, 0, 0, 0, 0, 0, 0]
3	*Premise:* A: but that is one of my solutions. Uh... B: I know here in Dallas that they have just instituted in the last couple of years, uh, a real long period of time that you can absentee vote before the elections. And I do not think they have seen a really high improvement. *Hypothesis:* they have seen a really high improvement. Heuristics: C V. BERT: C J. BERT: C AAs: C {C, C, C} Gold: C [-1, -2, -2, -2, -2, -2, -2, -2, -3, -3]
4	*Premise:*B: So did you commute everyday then or, A: No. B: Oh, okay. A: No, no, it was a six hour drive. B: Oh, okay, when you said it was quite a way away, I did not know that meant you had to drive like an hour *Hypothesis:* speaker A had to drive like an hour Heuristics: C V. BERT: D J. BERT: E AAs: D {E, C, C} Gold: D [3, 2, 2, 1, 0, 0, -1, -1, -1, -3]
5	*Premise:* The assassin's tone and bearing were completely confident. If he noticed that Zukov was now edging further to the side widening the arc of fire he did not appear to be troubled. *Hypothesis:* Zukov was edging further to the side Heuristics: D V. BERT: D J. BERT: D AAs: D {E, E, N} Gold: E [3, 3, 3, 3, 2, 2, 1, 1]
6	*Premise:* B: Yeah, and EDS is very particular about this, hair cuts, A: Wow. B: I mean it was like you can't have, you know, such and such facial hair, no beards, you know, and just really detailed. A: A: I don't know that that would be a good environment to work in. *Hypothesis:* that would be a good environment to work in Heuristics: C V. BERT: C J. BERT: D AAs: C {C, C, C} Gold: D [2, 0, 0, 0, 0, -1, -2, -3]
7	*Premise:* "Willy did mention it. I was puzzled, I 'll admit, but now I understand." How did you know Heather had been there? *Hypothesis:* Heather had been there Heuristics: N V. BERT: E J. BERT: E AAs: E {E, E, E} Gold: D [3, 3, 3, 2, 1, 1, 0, 0, 0]

Table 3.4: Models' predictions for CB test items. Labels in [] are predictions by individual AAs.

Predict \ Gold	E	N	C	D	Total
E	37	2	0	13	52
N	1	10	0	3	14
C	0	0	34	13	47
D	20	7	20	80	127
Total	58	19	54	109	240

Table 3.5: Confusion matrix for the test set. E: entailment, N: neutral, C: contradiction, D: disagreement.

	negation	modal	conditional	question	negR
Heuristic	51.29	48.02	37.69	44.64	54.16
V. BERT	60.91	73.98	44.84	53.02	61.91
J. BERT	60.94	73.95	46.02	51.68	63.67
AAs	**65.96**	**80.18**	**48.05**	**54.95**	**68.00**

Table 3.6: F1 for CB test set under the embedding environments and "I don't know/believe/think" ("negR").

3.6.2 Linguistic Construction Analysis

We also examine the model performance for different linguistic constructions to investigate whether the model learns some of the linguistic patterns present in the Heuristic baseline. The Heuristic rules are strongly tied to the embedding environments. Another construction used is one which can lead to "neg-raising" reading, where a negation in the matrix clause is interpreted as negating the content of the complement, as in example 3 (Table 3.4) where *I **do not think** they have seen a really high improvement* is interpreted as *I think they **did not see** a really high improvement*. "Neg-raising" readings often occur with *know*, *believe* or *think* in the first person under negation. There are 85 such items in the test set: 41 contradictions (thus neg-raising items), 39 disagreements and 5 entailments. Context determines whether a neg-raising inference is triggered [An and White, 2019].

Table 3.6 gives F1 scores for the Heuristic, BERT models and AAs for items under the different embedding environments and potential neg-raising items in the test set. Though AAs achieve the best overall results, it suffers under conditional and question environments, as the corresponding training data is scarce (9.04% and 14.17%, respectively). The Heuristic baseline always assigns contradiction to the "I don't know/believe/think" items, thus capturing all 41 neg-raising items but missing disagreements and entailments. BERT, a SOTA NLP model, is not great at capturing such items either: 71.64 F1 on contradiction vs. 52.84 on the others (Vanilla BERT); 71.69 F1 vs. 56.16 (Joint BERT). Our AAs capture neg-raising items

Correct inference by Heuristic?	Yes (130)		No (110)	
	Acc.	F1	Acc.	F1
V. BERT	80.00	80.45	41.51	42.48
J. BERT	79.74	80.04	42.73	44.15
AAs	**84.37**	**84.85**	**46.97**	**48.75**

Table 3.7: BERT-based models performance on test items correctly predicted by vs. items missed by linguistic rules. Numbers next to Yes/No denote the size.

better with 77.26 F1 vs. 59.38, showing an ability to carry out context-dependent inference on top of the learned linguistic patterns. Table 3.7, comparing performance on test items correctly predicted by the linguistic rules vs. items for which context-dependent reasoning is necessary, confirms this: AAs outperform the BERT baselines in both categories.

3.7 Conclusion

In this chapter, we introduced finer-grained natural language inference. This task aims at teasing systematic inferences from inherent disagreements. The inherent disagreement items are challenging for NLU models to handle, rarely studied in past NLI work. We show that our proposed AAs, which simulate the uncertainty in annotation process by capturing the modes in annotations, perform statistically significantly better than all baselines. However the performance obtained ($\sim$66%) is still far from achieving truly robust NLU, leaving room for improvement.

Chapter 4: FAQ Retrieval

4.1 Introduction

FAQ, short for frequently asked questions, is designed for the purpose of providing information on frequent questions or concerns. The FAQ retrieval task is defined as ranking FAQ items $\{(q_i, a_i)\}$ from an FAQ Bank given a user query Q. In the FAQ retrieval literature [Karan and Šnajder, 2016, 2018, Sakata et al., 2019], a user query Q can be learned to match with the question field q_i, the answer field a_i or their concatenation (i.e., FAQ tuple) $q_i + a_i$.

To advance the COVID-19 information search, we present an FAQ dataset, COUGH, consisting of FAQ Bank, Query Bank, and Relevance Set, following the standard of constructing a robust FAQ dataset [Manning et al., 2008]. The FAQ Bank contains 15919 FAQ items scraped from 55 authoritative institutional websites. COUGH covers a wide range of perspectives on COVID-19, spanning from general information about the virus to specific COVID-19-related instructions for a healthy diet. For evaluation, we further construct Query Bank and Relevance Set, including 1201 crowd-sourced queries and their relevance to a set of FAQ items judged by annotators. Examples from COUGH are shown in Figure 4.1.

Our dataset poses several new challenges (e.g., the answers being long and noisy, and hard to match due to larger search space) to existing FAQ retrieval models. The diversity of FAQ items, which is reflected in their varying query forms and lengths as well as in narrative styles, also contributes to these challenges. Furthermore, these challenges can

COUGH: The COVID-19 FAQ Dataset

FAQ Bank	User Query Bank
Question1: Should children wear masks? **Answer1**: *In general, children 2 years and older should wear a mask...Appropriate and consistent use of masks...* **Question2:** Coping with Self-Quarantine **Answer2**: *Remind yourself that difficult emotions are normal during self-quarantine...* **Question3:** COVID-19是如何在人与人之间传播的? (How does COVID-19 spread between people?) **Answer3**: ***...该病毒的人际传播主要通过感染者与他人密切接触...(...mainly when an infected person is in close contact with another person...)***	**Query1:** Is it possible for human beings to get sick with COVID-19 transmitted to them from animals? **Query2:** Is it possible to get infected by COVID 19 if I touch food surface packaging?

Annotated Relevance Set

Query	Relevant FAQ in FAQ Bank	Score
Query1	Q: Can wild animals spread the virus that causes COVID-19 to people or pets? A: Currently, there is no evidence to suggest...	3.67
Query1	Q: How is COVID-19 transmitted? A: COVID-19 illness is spread mainly from person to person through respiratory...	2.67
Query2	Q: What are the lab protocols for identifying the virus in food? On surfaces?A: As food hasn't been implicated in transmission	3.67

Figure 4.1: Examples from the COUGH dataset.

reflect the characteristics and difficulties of FAQ retrieval in real scenarios better than counterparts like FAQIR [Karan and Šnajder, 2016] and StackFAQ [Karan and Šnajder, 2018] (Table 4.1). Moreover, in contrast to all prior datasets, COUGH covers multiple query forms (e.g., question and query string forms) and has many annotated FAQs for each user query, whereas queries in existing FAQ datasets are limited to the question form and have much fewer annotations. As such, our COUGH is deemed as a robust dataset, upon which a robust FAQ retriever could be developed to handle some real challenges (e.g., lengthy answer, enormous search space) better.

The contribution in this chapter is two-fold. First, we construct a challenging dataset COUGH to aid the development of COVID-19 FAQ retrieval models. Second, we conduct extensive experiments using various SOTA models across different settings, explore limitations of current FAQ retrieval models, and discuss future work along this line.

4.2 Standard FAQ Dataset Construction: COUGH

Since the outbreak of COVID-19, the community has witnessed many datasets released to advance the research of COVID-19.The most related work to ours are Sun and Sedoc [2020] and Poliak et al. [2020], both of which constructed a collection of COVID-19 FAQs

	FAQIR (Karan and Šnajder)	**StackFAQ** (Karan and Šnajder)	**LocalGov** (Sakata et al.)	**Sun and Sedoc**	**Poliak et al.**	**COUGH (ours)**
Domain	Yahoo!	StackExchange	Government	COVID-19	COVID-19	COVID-19
# of FAQs	4,313	719	1,786	690	2,115	15,919
# of Queries (Q)	1,233	1,249	784	6,495*	24,240*	1,201
# of annotations per Q	8.22	Not Applicable	<10	5	5	32.17
Query Length	7.30	13.84	**	**	**	12.97
FAQ-query Length	12.30	10.39	**	**	**	13.00
FAQ-answer Length	33.00	76.54	**	**	**	113.58
Language	English	English	Japanese	English	Multi-lingual	Multi-lingual
# of sources	1	1	1	12	34	55

Table 4.1: Comparison of COUGH with representative counterparts. *: Extracted from existing resources (e.g., COVID-19 Twitter dataset [Chen et al., 2020]). **: Not Applicable, either not in English or not publicly available.

by scraping authoritative websites (e.g., CDC and WHO). However, the dataset in the former work is not available yet and the latter work does not evaluate models on their dataset, and there is still a great need to understand how existing models would perform on the COVID-19 FAQ retrieval task. Moreover, the numbers of FAQs[14] in the 5 existing FAQ datasets (Table 4.1) are generally lower than 2000, which renders a small search space and thus the ease for FAQ retrievers to find the most relevant FAQ given a query.

A typical research-oriented FAQ dataset [Manning et al., 2008] consists of three parts: FAQ Bank, User Query Bank and Annotated Relevance Set. In this section, we will describe how we construct each of the three in detail.

4.2.1 FAQ Bank Construction

We developed scrapers based on JHU-COVID-QA library Poliak et al. [2020] with modifications to enable special features for our COUGH dataset.

Web scraping: We collect FAQ items from authoritative international organizations, state governments and some other credible websites including reliable encyclopedias and medical forums. Moreover, we scrape three types of FAQs: question form (i.e., an interrogative statement), query string (i.e., a string of words to elicit information) form and forum form

[14]In the literature, only 789 FAQ items are used for evaluation on FAQIR [Karan and Šnajder, 2018, Mass et al., 2020].

(FAQs scrapped from medical forums). Inspired by Manning et al. [2008], we loosen the constraint that queries must be in question form since we want to study a more generic and challenging problem. We also scrape 6,768 non-English FAQs to increase language diversity. Overall, we scraped a total of 15,919 FAQ items covering all three types and 19 languages. All FAQ items were collected and finalized on Aug. 30^{th}, 2020.

4.2.2 User Query Bank Construction

Following Karan and Šnajder [2016], Manning et al. [2008], we do not crowdsource queries from scratch, but instead ask annotators to paraphrase our provided query templates (See phase 1 below for details). In this way, we can ensure that 1) the collected queries are pertinent to COVID-19; 2) the collected queries are not too simple; 3) the chance of getting (nearly) duplicate user queries is reduced.

Phase 1: Query Template Creation: We sample 5% of FAQ items from each English non-forum source and use the question part as the template.

Phase 2: Paraphrasing for Queries: In this phase, each annotator is expected to give three paraphrases for each query template. Annotators are encouraged to give deep paraphrases (i.e., grammatically different but semantically similar/same) to simulate the noisy and diverse environment in real scenarios. In the end, we obtain 1236 user queries.

4.2.3 Annotated Relevance Set Construction

Phase 1: Initial Candidate Pool Construction: For each user query, as suggested by previous work [Manning et al., 2008, Karan and Šnajder, 2016, Sakata et al., 2019], we run 4 models[15], BM25 (Q-q), BM25 (Q-q+a), BERT (Q-q) and BERT (Q-a) fine-tuned on `COUGH`, to instantiate a candidate FAQ pool. Each model complements the others and contributes its

[15]Explanations of these models are in chapter 4.4.2.

	Type	Number	Q-Length	A-length
# English	Question	4978	14.64	123.89
	Query String	2139	9.18	89.60
	Forum	2034	147.46	90.49
# Non-English	Question	3396	-	-
	Query String	3372	-	-
# Total	-	15919	-	-

Table 4.2: Basic statistics of FAQ bank in COUGH.

top-10 relevant FAQ items. We then take the union to remove duplicates, giving an average pool size of 32.2.

Phase 2: Human Annotation: Each annotator gives each ⟨Query, FAQ item⟩ tuple a score based on the annotation scheme (i.e., Matched (4), Useful (3), Useless (2) and Non-relevant (1)) which is adapted from Karan and Šnajder [2016], Sakata et al. [2019]. In order to reduce the variance and bias in annotation, each tuple has at least 3 annotation scores. In our finalized Annotated Relevance Set, we keep all raw scores and include two additional labels: 1) mean of raw annotation scores; 2) binary label (positive/negative). We identify all tuples with mean score greater than 3 as positive examples.

Among 1236 user queries, we find that there are 35 "unanswerable" queries that have no associated positive FAQ item. In the end, there are 1201 user queries involved for evaluation after removing "unanswerable" queries.

4.3 COUGH Dataset Analysis

Besides the generic goal of large size, diversity, and low noise, COUGH features 4 additional aspects:

Varying Query Forms: As indicated in Table 4.2, there are multiple query forms. In evaluation, we include both question and query string forms. These two distinct forms are different in terms of query format (interrogative vs. declarative), average answer length

(123.89 vs. 89.60) and topics. Question form is usually related to general information about the virus while query string form is often searching for more specific instructions concerning COVID-19 (e.g., healthy diet during pandemic). In Figure 4.1, the first FAQ item is in question form while the second one is in query string form.

Answer Nature: Table 4.1 shows the answer fields in COUGH are much longer than those in any prior dataset. We also observe that answers might contain some contents which are not directly pertinent to the query, partially resulting in the long length nature. For example, in COUGH, the answer to a query "What is novel coronavirus" contains extra information about comparisons with other viruses. Such lengthy and noisy nature of answers shows the difficulty of FAQ retrieval in real scenarios.

Large-scale Relevance Annotation: Many existing FAQ datasets overlooked the scale of annotations (Table 4.1); yet, that would hurt the evaluation reliability since many true positive ⟨Query, FAQ item⟩ tuples were omitted. Following Manning et al. [2008], for each user query, we constructed a large-scale candidate pool to reduce the chance of missing true positive tuples. The annotation procedure yielded 39760 annotated ⟨Query, FAQ item⟩ tuples, each of which is annotated by at least 3 people to reduce annotation bias. Furthermore, we find that there are 7856 (19.76%) positive tuples (i.e., mean score > 3). Besides, from the perspective of FAQ Bank, 6648 of 7117 English non-forum items appear at least once in Initial Candidate Pool, and 3790 of them have at least one "matched" user query.

Multilinguality: COUGH includes 6768 FAQ items covering 18 non-English languages. In this thesis, we do not include FAQ items in languages other than English in the evaluation.[16] However, we do encourage investigators who use COUGH to better utilize non-English FAQ

[16]No annotation is done on non-Engligh items.

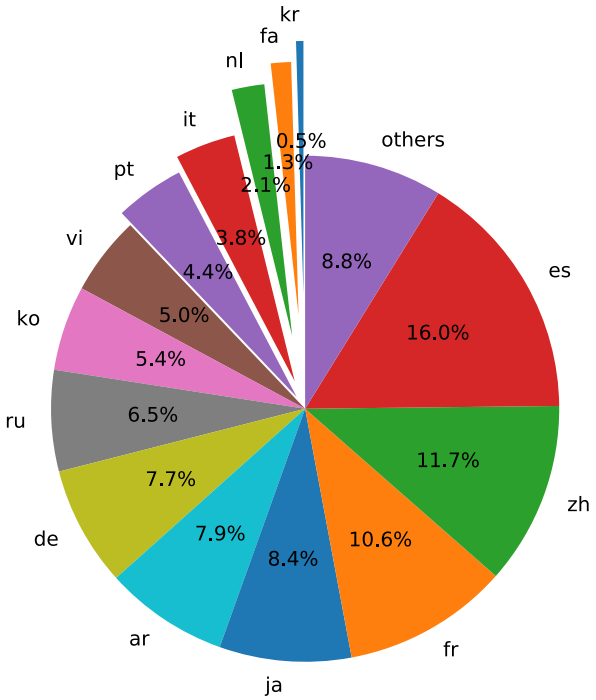

Figure 4.2: Language distribution for non-English FAQ items.

items for other potential tasks, such as multi-lingual FAQ retrieval and transfer learning from English FAQ items to low-resource non-English FAQ items.

Figure 4.2 shows the language distribution (excluding English) of FAQ items in COUGH dataset. Like English FAQ items, non-English FAQ items are also presented in both question and query string forms. Statistics of non-English items can be found in Table 4.2.

4.4 FAQ Retrieval Methods

4.4.1 FAQ Retrieval Methods Overview

The standard practice in FAQ retrieval focuses on retrieving the most-matched FAQ items given a user query [Karan and Šnajder, 2018]. Many earlier work, such as FAQ FINDER [Burke et al., 1997], query expansion [Kim and Seo, 2006] and BM25 [Robertson and Zaragoza, 2009], resorted to traditional IR techniques by leveraging lexical mapping and/or semantic similarity. In the deep learning era, many studies show that Neural Networks

are useful for FAQ retrieval as they are good at learning the semantic relevance between queries and FAQ items. Along this line, [Karan and Šnajder, 2016] adopted Convolution Neural Networks, [Gupta and Carvalho, 2019] utilized LSTM, and [Sakata et al., 2019] leveraged an ensemble of TSUBAKI [Shinzato et al., 2012] and BERT [Devlin et al., 2019b]. Recently, Mass et al. [2020] employed CombSum and PoolRank, ensembles of BM25 and BERT models, to learn ranking without requiring manual annotations.

4.4.2 Unsupervised FAQ Retrieval

In this chapter, we only focus on the unsupervised models since the size of User Query Bank (1201 items) is not large enough for supervised learning, especially for fine-tuning complex language models like BERT. We experiment with three commonly-used and SOTA unsupervised models to understand their limitations and figure out the challenge present in real scenarios for FAQ retrieval. Besides, each model has three configurable modes, Q-q, Q-a and Q-q+a, where we match user queries (Q) to the question (q) and answer (a) of an FAQ item as well as their concatenation (q+a)[17], respectively.

Baseline Models

(1) **BM25** [Robertson and Zaragoza, 2009], a commonly adopted IR baseline, is a nonlinear combination of term frequency, document frequency and document length.

(2) **BERT** [Devlin et al., 2019b] is a pretrained language model. We experiment with Sentence-BERT [Reimers and Gurevych, 2019], a Siamese network built for comparison between sentence-pair embeddings, which specializes in generating meaningful sentence representations.

[17]Q-q+a mode is only used for BM25 and BM25 in CombSum.

Fine-tuning: We use Multiple Negatives Ranking (MNR) loss[18] [Henderson et al., 2017] to fine-tune Sentence-BERT on FAQ bank. For the Q-q mode, similar to Mass et al. [2020], we use GPT2 [Radford et al., 2019] to generate synthetic questions as positive q's to match with Q and filter out low-quality ones via Elasticsearch. For the Q-a mode, an FAQ item itself is a positive pair. For both modes, negative q's or a's are randomly sampled.

(3) **CombSum** [Mass et al., 2020] first computes three matching scores between the user query and FAQ items via BM25 (Q-q+a), BERT (Q-q) and BERT (Q-a) models, respectively. Then, the three scores are normalized and combined by averaging.

4.5 Evaluation and Results

Evaluation Metric: We adopt our binary label (positive/negative) as ground truth labels. Following previous work [Karan and Šnajder, 2016, 2018, Sakata et al., 2019], we adopt widely-used MAP (Mean Average Precision)[19], MRR (Mean Reciprocal Rank) and P@5 (Precision at top 5) metrics.

Evaluation Settings: For the scope of this chapter, we only evaluate on English non-forum FAQ items, and leave the non-English and forum ones for future research as great challenges have already been observed under the current setting. However, we do encourage investigators who use `COUGH` to utilize these two categories for other potential applications (e.g., multi-lingual IR, transfer learning in IR).

Evaluation Results: Models' results are listed in Table 4.3. The current best results (P@5: 0.31; MAP: 0.42; MRR: 0.64) are not satisfying, showing a large room for improvement.

[18]For efficiency, MNR loss is computed using answers of other FAQs in the same training batch as negative answers.

[19]Evaluated on top-100 retrieved FAQ items.

	P@5	MAP	MRR
BM25 (Q-q)	0.27	0.38	0.56
BM25 (Q-a)	0.16	0.23	0.34
BM25 (Q-q+a)	0.25	0.34	0.52
BERT (Q-q) w/o finetune	0.29	**0.42**	0.59
+ finetune on pesudo Q-q	0.26	0.36	0.60
BERT (Q-a) w/o finetune	0.06	0.12	0.17
+ finetune on FAQ Bank	0.23	0.30	0.50
CombSum w/o finetune	0.21	0.31	0.49
+ fintune on pesudo Q-q	0.23	0.31	0.53
+ fintune on FAQ Bank	0.31	0.39	0.63
+ fintune on pesudo Q-q and FAQ Bank	**0.31**	0.39	**0.64**

Table 4.3: Evaluation on COUGH. BERT refers to Sentence-BERT [Reimers and Gurevych, 2019].

These results not only confirm that COUGH is challenging but also signify more robust methods and models are needed to handle challenges imposed by COUGH more effectively.

4.6 Analysis

Quantitative Analysis: It is not surprising to see that the Q-q mode consistently performs better than the Q-a mode regardless of underlying models. This is mainly caused by the fact that question fields are more similar to user queries than answer fields, in terms of syntactic structures and semantic meanings. As discussed in Section 4.3, the answer nature (lengthy and noisy) and large search space, albeit well characterize the FAQ retrieval task in real scenarios, do bring a great challenge to current FAQ retrieval models.

We observe that fine-tuning in the way we experimented with can only help improve the performance of the Q-a mode by a small margin, but might slightly hurt the Q-q mode due to the noise introduced in generating synthetic queries. Moreover, ensemble models don't perform as well as expected, since the particular Q-a model involved is weak (even after fine-tuning), which negatively impacts performance. In consequence, doing straightforward fine-tuning or ensemble simply by stacking models wouldn't improve the performance

Query: What research is being done on antibody tests and their accuracy? **FAQ item:** Q: What is antibody testing? How do I get a COVID-19 antibody test? A: CDC and partners are investigating to determine if you can get sick with COVID-19 more than once ... **Gold label:** `Negative [useful, useless, useless]` **Predicted rank:** `3`
Query: Are COVID-19 antibody tests accurate? **FAQ item:** Q: Should I be tested with an antibody (serology) test for COVID-19? A: ... Antibody tests have limited ability to diagnose COVID-19 and should not be used alone to diagnose COVID-19 ... **Gold label:** `Positive [useful, useful, matched]` **Predicted rank:** `26`

Table 4.4: Error analysis with fine-tuned BERT (Q-q). Human annotations are inside `[]`.

significantly, which confirms that COUGH is a challenging dataset. Interesting future work includes developing more advanced techniques to handle long and noisy answer fields.

Qualitative Analysis: To understand finetuned BERT (Q-q) better, we conduct error analysis as shown in Table 4.4 to show its major types of errors, hoping to further improve it in the future. Currently, finetuned BERT (Q-q) suffers from the following issues: 1) biased towards responses with similar texts (e.g., "antibody tests" and "antibody testing"); 2) fails to capture the semantic similarities under complex environments (e.g., pragmatic reasoning is required to understand that "limited ability" indicates results are not accurate for diagnosing COVID-19).

4.7 Conclusion

In this chapter, we introduce COUGH, a large challenging dataset for COVID-19 FAQ retrieval. COUGH features varying query forms, long and noisy answers, larger search space and multilinguality. COUGH also serves as a better evaluation benchmark since it has large-scale relevance annotations. Albeit results show the limitations of current FAQ retrieval models, COUGH is a more robust dataset than its counterparts since it better characterizes the challenges present in real scenarios for FAQ retrieval.

Chapter 5: Conclusion

In this thesis, I have embarked on building models and constructing datasets towards more robust natural language understanding. We start with a discussion on what robustness problem is in natural language understanding. That is, fully-trained NLU models are usually lacking generalizability and flexibility. In this thesis, we argue that, in order to achieve truly robust natural language understanding, implementing robust models and curating robust datasets are equally important. In this thesis, we investigate the NLU robustness problem in three NLU tasks (i.e., Question Answering, Natural Language Inference and Information Retrieval). We then propose novel methods and construct new datasets to advance research on improving the robustness of NLU systems.

In Chapter 2, we study how to utilize diversity boosters (e.g., beam search & QPP) to help Question Generator synthesize diverse QA pairs, upon which a Question Answering (QA) system is trained to improve the generalization onto unseen target domain. It's worth mentioning that our proposed QPP (question phrase prediction) module, which predicts a set of valid question phrases given an answer evidence, plays an important role in improving the cross-domain generalizability for QA systems. Besides, a target-domain test set is constructed and approved by the community to help evaluate the model robustness under cross-domain generalization setting. In Chapter 3, we investigate inherently ambiguous items in the NLI (Natural Language Inference) task, which are overlooked in the literature

but often occurring in the real world, for which annotators don't agree with the gold label. We build an ensemble model, AAs (Artificial Annotators), which simulates underlying annotation distribution to effectively identify such inherently ambiguous items. Our AAs, motivated by the nature of inherently ambiguous items, are better than vanilla models since our model design captures the essence of the problem better. In Chapter 4, we follow a standard practice to build a robust dataset for FAQ retrieval task. In our dataset analysis, we show how COUGH better reflects the challenge of FAQ retrieval in the real situation than its counterparts. The imposed challenge (e.g., long and noisy answer, large search space) will push forward the boundary of research on FAQ retrieval in real scenarios.

Overall, the technical contributions of this thesis are as follows:

1. We investigate the robustness problem in depth, and identify the equal importance of models implementation and datasets construction towards improving the robustness of NLU systems. In this thesis, we specifically study three concrete NLU tasks.
2. We propose two novel methods to help improve NLU model robustness. Specifically, we evaluate the effect of diverse question generation (QG) for clinical QA under the cross-domain evaluation setting, and propose QPP (Question Phrase Prediction) module as an effective diversity booster for QG [Yue et al., 2020b]. Moreover, we propose AAs (Artificial Annotators) to simulate underlying annotation distribution to handle a previously-overlooked NLI class better, inherent disagreement items [Zhang and de Marneffe, 2021].
3. We construct two robust datasets, QA test set on MIMIC-III Database [Yue et al., 2020b] and COUGH [Zhang et al., 2020]. They will serve as better evaluation benchmarks to examine designed models' generalization capabilities and abilities to handle real-scenario challenges (e.g., longer FAQ and larger search space).

Future Research: Moving forward, the ultimate goal for robust natural language understanding is to build NLU models which can behave humanly. That is, it's expected that robust NLU models are capable to transfer the knowledge from training corpus to unseen documents more reliably and survive when encountering challenging items even if the model doesn't know a priori of users' inputs. Two suggested important research frontiers are:

1) Improve model generalization under cross-domain setting: In Chapter 2, we discussed how we utilized QG model to help alleviate the generalization challenge encountered by QA systems. However, the question whether a better QA system could further improve the QG is yet known, which is, however, worth deeper investigation. Ideally, when introducing an auxiliary module to help the main model, we also expect to see that the auxiliary module could be benefited by the joint training with the main model. Besides, in Chapter 2, the reason we decided to utilize QG that way is that we observed that the QG system didn't suffer from severe generalization issues under the clinical setting. However, in open-domain, the aforementioned observation might not hold. In that case, it might be better to enforce the model to learn text representations that are invariant to domain changes. Recent work on cross-domain NER (Named Entity Recognition) have shown some progress along this path [Jia et al., 2019]. I also have a great interest in text generation. Though the majority of work that utilize domain adaptation techniques to tackle the generalization challenge focuses on classification tasks [Ganin et al., 2016, Chen and Cardie, 2018, Chen et al., 2018], could we effectively extend the success of domain adaptation to text generation? This might be a promising research direction since the text generation can be formulated as a sequence of classifications.

2) Embrace more challenges in NLU: In Chapter 3 and 4, we discussed two datasets, CommitmentBank & `COUGH`, on which we could develop methods that target at solving

NLU challenges under more realistic scenarios. SQuAD 2.0 [Rajpurkar et al., 2018] is a another great role model for datasets that aim at this goal. To do well on SQuAD 2.0, models must not only answer questions when possible, but also determine when no answer is supported by the paragraph and then say "no". This is a real challenge for QA system as it's not always the case that an answer could be found in a seemingly relevant document for a question. Another typical real challenge in NLU is how to solve mathematical problems. Hendrycks et al. [2021] presents a new math dataset on which a standard CS PhD student who doesn't especially like Math gets 40% accuracy while a fully-trained GPT-3 [Brown et al., 2020] models only gets 5%. Pretrained language models like GPT-3 or BERT is believed to heavily rely on the context to reason about the given prompt. However, mathematical language isn't necessarily constrained by contexts,[20] which imposes a great challenge to NLU systems. Additionally, in order to get full credits for a problem, the deployed system is also required to give correct reasoning steps, which is way more difficult than simply generating an answer. The following is an example from MATH dataset:[21]

Problem: *If* $\Sigma_{n=0}^{\infty} cos^{2n}\theta = 5$*, what is* $cos2\theta$*?*

Solution: *The geometric series is* $1 + cos^2\theta + cos^4\theta + ... = \frac{1}{1-cos^2\theta} = 5$*. Hence,* $cos^2\theta = \frac{4}{5}$*. Then,* $cos2\theta = 2cos^2\theta - 1 = \frac{3}{5}$

Moreover, linguistic rules or features, without any doubt, deserve more attention even if we are living in the realm of neural computing world. This is because linguistic rules or features exhibit great power when tackling challenging NLU problems. In Chapter 3, we find that SOTA NLU models, BERT, obtain inferior results to our linguistics-driven heuristic rules on dev set. This shows that giant neural models still fail to capture some necessary

[20]Math question could be context-free such as "let a equal one plus two minus three times four, is a congruent to zero when the modulo is five?"

[21]This example corresponds to the second example in their Figure 1.

linguistic phenomena. As such, it's essential to discover how to effectively incorporate linguistic information into neural models to compensate for what the neural network-model is weak at. A simple practice is to embed linguistic features such as NER and POS tags into original texts. Particularly, I observed that a vanilla attention-based Seq2Seq model, when being equipped with linguistic features, could achieve better performance than BART [Lewis et al., 2020],[22] a variant of BERT specializing in text generation, on both in-domain and cross-domain question generation tasks.

[22]In general, BART ($\sim$139M) has 8 times more parameters than vanilla attention-based Seq2Seq ($\sim$17M).

Appendix A: Supplementary Materials

A.1 Clinical Question Answering

A.1.1 Answer Evidence Extractor

Formulation and Implementation Formally, given a document (context) $\mathbf{p} = \{p_1, p_2, ..., p_m\}$, where p_i is the i-th token of the document and m is the total number of tokens, we aim to extract potential evidence sequences. Firstly, we adopt the ClinicalBERT model [Alsentzer et al., 2019] to encode the document:

$$\mathbf{U} = \texttt{ClinicalBERT}\{p_1, ..., p_m\}. \tag{A.1}$$

where $\mathbf{U} \in \mathbb{R}^{m \times d}$, and d is size of the dimension.

Following the same paradigm of the BERT model for the sequence labeling task [Devlin et al., 2019b], we predict the `BIO` tag for each a_j as follows:

$$\Pr(a_j|p_i) = \text{softmax}(\mathbf{U} \cdot \mathbf{W} + \mathbf{b}),\ \forall p_i \in \mathbf{p} \tag{A.2}$$

We train model on source contexts by minimizing the negative log-likelihood loss.

Post-processing Heuristic Rules We observe that when we directly apply the ClinicalBERT [Alsentzer et al., 2019] system described in Section 2.3.1 on clinical texts, the extracted answer evidences sometimes are broken sentences due to the noisy nature and uninformative language (e.g., acronyms) of clinical texts. To make sure the extracted evidences are meaningful, we designed a *"merge-and-drop"* heuristic rule to further improve the extractor's

accuracy. Specifically, for each extracted evidence candidate, we first examine the *length* (number of tokens) of the extracted evidence. If the length is larger than the threshold η, we keep this evidence; otherwise, we compute the *distance*, i.e., the number of tokens between the current candidate span and the closest span. If the *distance* is smaller than the threshold γ, we merge these two "close-sitting" spans; otherwise, we drop this over-short evidence span. In our experiments, we set η and γ to be 3 and 3, respectively, since they help the QA system achieve the best performance on the dev set.

A.1.2 Question Phrases Identification

In order to utilize the Question Phrase Prediction (QPP) module and make the QPP module generic enough without loss of generality, we identify valid n-gram Question Phrases in an automatic way.

To prepare an exhaustive list of valid n-gram Question Phrases, we first collect all of the first *n* words appearing in questions in emrQA, forming three (i.e., *n*=1, 2, 3) raw Question Phrases set.

We observe that all uni-grams are valid question phrases (e.g., "How", "When", "What"), so we don't do any pruning and keep the uni-gram question phrases set as it is.

As for n-gram (n $\geq$ 2) Question Phrases set, we conduct fine-grained filtering. We only consider n-grams with occurrence frequency greater than the threshold ζ as valid n-gram Question Phrases. In our experiment, we set ζ as 0.02%. Less frequent n-gram words (i.e., frequency $< 0.02\%$) will degrade to unigram Question Phrases in accordance with corresponding question types (e.g., "Has lasix" $\rightarrow$ "Has"*) so as to maintain lossless. In the end, n-gram (n $\geq$ 2) Question Phrases sets, without any information loss, consist of both n-gram Question Phrases and degraded unigram Question Phrases.

A.1.3 Dev Set Construction

The dev set on MIMIC-III is constructed by sampling generated questions from 9 QG models and is used to tune the hyper-parameters only. Instead of uniformly sampling from 9 QG models, we followed the sampling ratio of 1:3:6 (Base model, Base+BeamSearch, Base+QPP) for each QG method, which made the dev set cover as many diverse questions as possible.

Bibliography

E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. *NAACL Clinical NLP Workshop 2019*, 2019.

H. Y. An and A. S. White. The lexical and grammatical sources of neg-raising inferences. In *Proceedings of the Society for Computation in Linguistics (SCiL 2020)*, pages 220–233, 2019.

S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP'15*, pages 632–642, 2015.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

R. Burke, K. Hammond, V. Kulyukin, S. Lytinen, N. Tomuro, and S. Schoenberg. Question answering from frequently asked question files: Experiences with the FAQ FINDER system. *AI Magazine*, pages 57–66, 1997.

A. Calma and B. Sick. Simulation of annotators for active learning: Uncertain Oracles. In *Proceedings of the Workshop and Tutorial on Interactive Adaptive Learning co-located with European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2017)*, pages 49–58, 2017.

Y.-H. Chan and Y.-C. Fan. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, 2019.

D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading wikipedia to answer open-domain questions. In *ACL'17*, pages 1870–1879, 2017.

E. Chen, K. Lerman, and E. Ferrara. Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 2020.

X. Chen and C. Cardie. Multinomial adversarial networks for multi-domain text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1226–1240, 2018.

X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Q. Weinberger. Adversarial deep averaging networks for cross-lingual sentiment classification. *Trans. Assoc. Comput. Linguistics*, 6:557–570, 2018.

J. Cho, M. Seo, and H. Hajishirzi. Mixture content selection for diverse sequence generation. In *EMNLP-IJCNLP'19*, pages 3112–3122, 2019.

K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In A. Moschitti, B. Pang, and W. Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734, 2014.

I. Dagan, O. Glickman, and B. Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005*, pages 177–190, 2005.

M. de Marneffe, M. Simons, and J. Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. In *Sinn und Bedeutung 23*, 2019.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186, 2019a.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT'19*, pages 4171–4186, 2019b.

X. Du, J. Shao, and C. Cardie. Learning to ask: Neural question generation for reading comprehension. In *ACL'17*, pages 1342–1352, 2017.

J. L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

J. Fan. Annotating and characterizing clinical sentences with explicit why-qa cues. In *NAACL Clinical NLP Workshop*, pages 101–106, 2019.

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. S. Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17:59:1–59:35, 2016.

W. Gantt, B. Kane, and A. S. White. Natural language inference with mixed effects. In *The Ninth Joint Conference on Lexical and Computational Semantics (*SEM 2020)*, 2020.

D. Golub, P.-S. Huang, X. He, and L. Deng. Two-stage synthesis networks for transfer learning in machine comprehension. In *EMNLP'17*, pages 835–844, 2017.

S. Gupta and V. R. Carvalho. FAQ retrieval using attentive matching. In *SIGIR'19*, page 929–932, 2019.

M. Henderson, R. Al-Rfou, B. Strope, Y. hsuan Sung, L. Lukacs, R. Guo, S. Kumar, B. Miklos, and R. Kurzweil. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652, 2017.

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset. *CoRR*, abs/2103.03874, 2021.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, pages 1735–1780, 1997.

D. Ippolito, R. Kriz, J. Sedoc, M. Kustikova, and C. Callison-Burch. Comparison of diverse decoding methods from conditional language models. In *ACL '19*, pages 3752–3762, 2019.

C. Jia, L. Xiao, and Y. Zhang. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2464–2474, 2019.

N. Jiang and M. de Marneffe. Evaluating BERT for natural language inference: A case study on the Commitmentbank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP 2019*, pages 6085–6090, 2019a.

N. Jiang and M. de Marneffe. Do you know that Florence is packed with visitors? Evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 4208–4213, 2019b.

A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

J. Kang, H. P. San Roman, et al. Let me know what to ask: Interrogative-word-aware question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 163–171, 2019.

M. Karan and J. Šnajder. FAQIR - a frequently asked questions retrieval test collection. In *Text, Speech, and Dialogue - 19th International Conference, TSD 2016*, pages 74–81, 2016.

M. Karan and J. Šnajder. Paraphrase-focused learning to rank for domain-specific frequently asked questions retrieval. In *Expert Systems with Applications*, pages 418–433, 2018.

K. Kenyon-Dean, E. Ahmed, S. Fujimoto, J. Georges-Filteau, C. Glasz, B. Kaur, A. Lalande, S. Bhanderi, R. Belfer, N. Kanagasabai, R. Sarrazingendron, R. Verma, and D. Ruths. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018*, pages 1886–1895, 2018.

H. Kim and J. Seo. High-performance FAQ retrieval using an automatic clustering method of query logs. *Information Processing & Management*, pages 650 – 661, 2006.

A. Lavie and M. J. Denkowski. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2-3):105–115, 2009.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880, 2020.

J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT'16*, pages 110–119, 2016.

C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004.

B. Liu, H. Wei, D. Niu, H. Chen, and Y. He. Asking questions the human way: Scalable question-answer generation from text corpus. In *WWW'20*, pages 2032–2043, 2020.

M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP'15*, pages 1412–1421, 2015.

B. MacCartney and C. D. Manning. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics, IWCS 2009*, pages 140–156, 2009.

C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Y. Mass, B. Carmeli, H. Roitman, and D. Konopnicki. Unsupervised FAQ retrieval with question generation and BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 807–812, 2020.

T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, pages 3428–3448, 2019.

n2c2. n2c2 nlp research data sets, 2006. URL `https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/`.

A. Pampari, P. Raghavan, J. Liang, and J. Peng. emrqa: A large corpus for question answering on electronic medical records. In *EMNLP'18*, pages 2357–2368, 2018.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *ACL'02*, pages 311–318, 2002.

J. Patrick and M. Li. An ontology for clinical questions about the contents of patient notes. *Journal of Biomedical Informatics*, 45(2):292–306, 2012.

E. Pavlick and C. Callison-Burch. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, 2016.

E. Pavlick and T. Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, pages 677–694, 2019.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543, 2014.

M. E. Peters, S. Ruder, and N. A. Smith. To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proceedings of the 4th Workshop on Representation Learning for NLP, RepL4NLP@ACL 2019*, pages 7–14, 2019.

A. Poliak, M. Fleming, C. Costello, K. W. Murray, M. Yarmohammadi, S. Pandya, D. Irani, M. Agarwal, U. Sharma, S. Sun, N. Ivanov, L. Shang, K. Srinivasan, S. Lee, X. Han, S. Agarwal, and J. Sedoc. Collecting verified COVID-19 question answer pairs, 2020.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. *Technical report, OpenAI*, 2018.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners, 2019.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019.

P. Raghavan, S. Patwardhan, J. J. Liang, and M. V. Devarakonda. Annotating electronic medical records for question answering. *arXiv preprint arXiv:1805.06816*, 2018.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP'16*, pages 2383–2392, 2016.

P. Rajpurkar, R. Jia, and P. Liang. Know what you don't know: Unanswerable questions for squad. In *ACL'18*, pages 784–789, 2018.

B. P. S. Rawat, W.-H. Weng, P. Raghavan, and P. Szolovits. Entity-enriched neural models for clinical question answering. *arXiv preprint arXiv:2005.06587*, 2020.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 3980–3990, 2019.

S. Robertson and H. Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, pages 333–389, 2009.

W. Sakata, T. Shibata, R. Tanaka, and S. Kurohashi. FAQ retrieval using query question similarity and BERT-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019*, pages 1113–1116, 2019.

H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992, Pittsburgh, PA, USA, July 27-29, 1992*, pages 287–294, 1992.

S. Shakeri, C. N. dos Santos, H. Zhu, P. Ng, F. Nan, Z. Wang, R. Nallapati, and B. Xiang. End-to-end synthetic data generation for domain adaptation of question answering systems. In *EMNLP'20*, pages 5445–5460, 2020.

K. Shinzato, T. Shibata, D. Kawahara, and S. Kurohashi. Tsubaki: An open search engine infrastructure for developing information access methodology. *Journal of Information Processing*, pages 216–227, 2012.

S. Sun and J. Sedoc. An analysis of BERT FAQ retrieval models for COVID-19 infobot, 2020.

I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS '14*, pages 3104–3112, 2014.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 3261–3275, 2019a.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019b.

H. Wang, Z. Gan, X. Liu, J. Liu, J. Gao, and H. Wang. Adversarial domain adaptation for machine reading comprehension. In *EMNLP-IJCNLP'19*, pages 2510–2520, 2019c.

A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL'18*, pages 1112–1122, 2018.

R. V. Yampolskiy. Turing test as a defining feature of ai-completeness. In *Artificial Intelligence, Evolutionary Computing and Metaheuristics - In the Footsteps of Alan Turing*, volume 427, pages 3–17. Springer, 2013.

X. Yue, B. J. Gutierrez, and H. Sun. Clinical reading comprehension: A thorough analysis of the emrQA dataset. In *ACL'20*, 2020a.

X. Yue, X. F. Zhang, Z. Yao, S. Lin, and H. Sun. CliniQG4QA: Generating diverse questions for domain adaptation of clinical question answering. *Machine Learning for Health Workshop (ML4H) at NeurIPS 2020*, 2020b.

X. F. Zhang and M. de Marneffe. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*, 2021.

X. F. Zhang, H. Sun, X. Yue, E. Jesrani, S. Lin, and H. Sun. COUGH: A challenge dataset and models for COVID-19 FAQ retrieval. *arXiv preprint arXiv:2010.12800*, 2020.

Y. Zhang, M. Galley, J. Gao, Z. Gan, X. Li, C. Brockett, and B. Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *NeurIPS'18*, pages 1810–1820, 2018.

Q. Zhou, N. Yang, F. Wei, C. Tan, H. Bao, and M. Zhou. Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 662–671. Springer, 2017.