A Zero-Shot Claim Detection Framework using Question Answering

Revanth Gangi Reddy[1], **Sai Chetan**[1], **Yi R. Fung**[1], **Kevin Small**[2], **Heng Ji**[1]
[1]University of Illinois at Urbana-Champaign [2]Amazon Alexa AI
{revanth3, scc8, yifung2, hengji}@illinois.edu
smakevin@amazon.com

Abstract

In recent years, there has been an increasing interest in claim detection as an important building block for misinformation detection. This involves detecting more fine-grained attributes relating to the claim, such as the claimer, claim topic, claim object pertaining to the topic, etc. Yet, a notable bottleneck of existing claim detection approaches is their portability to emerging events and low-resource training data settings. In this regard, we propose a fine-grained claim detection framework that leverages zero-shot Question Answering (QA) using directed questions to solve a diverse set of sub-tasks such as topic filtering, claim object detection, and claimer detection. We show that our approach[1] significantly outperforms various zero-shot, few-shot and task-specific baselines on the NEWSCLAIMS benchmark (Reddy et al., 2021).

1 Introduction

Claim detection over news involves identifying claims related to various topics in news articles. Identifying such claims is a crucial first step for fighting misinformation and disinformation online. However, such harmful content can evolve rapidly, triggered by relatively new events which can gain extensive media coverage within a short time span. Hence, claim detection in such scenarios requires systems that are able to adapt quickly, by working well under zero-shot or few-shot settings.

Towards this goal, Reddy et al. (2021) propose a new benchmark, NEWSCLAIMS, that evaluates claim detection for previously unseen topics without access to any training data. Given a collection of news articles, the task involves identifying claims that are related to a pre-defined set of topics, along with extracting attributes for each claim, such as claim span, claim object, claimer and stance. Essentially, this benchmark extends the claim detection task to involve extracting additional background attributes relating to the claim, such as the claim object and claimer. Furthermore, the claimer detection sub-task within NEWSCLAIMS requires considerable document-level reasoning, making it harder than existing attribution tasks (Pareti, 2016; Newell et al., 2018), which mainly involve sentence-level reasoning. Reddy et al. (2021) propose various baselines for each sub-task in the NEWSCLAIMS benchmark, which involve zero-shot and few-shot approaches in addition to baselines trained using task-specific data.

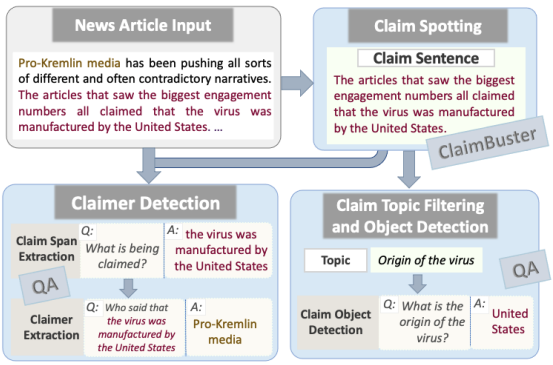

Figure 1: An overview of our claim detection framework. We leverage zero-shot QA for claim topic filtering, claim object detection and claimer detection for the claims spotted by ClaimBuster (Hassan et al., 2017).

To handle this scenario in a low-resource setting, we hypothesize that identifying claim topics and extracting corresponding claim attributes can be formulated as a question answering task. Hence, we propose to leverage a Question Answering (QA) system for the claim detection task and show that the *same QA model* can be used to solve multiple sub-tasks within claim detection, without the need of any task-specific training data. This involves filtering claims relating to specific topics, identifying claim objects associated with such topics and attribution for identifying the claimers making these

[1]Code is available here: https://github.com/blender-nlp/NewsClaims/tree/main/zero-shot-qa

Proceedings of the 29th International Conference on Computational Linguistics, pages 6927–6933
October 12–17, 2022.

claims. We realise this by using directed questions to help solve connected sub-tasks such as topic filtering, claim object detection and claimer detection within the claim detection framework. An overview of this framework is shown in Figure 1.

Leveraging pre-trained language models for directly solving end-tasks has been explored in prompting (Liu et al., 2021), with promising performance in both zero-shot (Zhong et al., 2021) and few-shot (Brown et al., 2020) settings. Prior work uses prompts that are in the form of prefix (Li and Liang, 2021) or cloze-style tasks (Schick and Schütze, 2021). In this work, we solve end-tasks by formulating them as directed questions, instead of prompts, that are fed into a QA model (see Figure 1 1 for examples of directed questions). We will first briefly introduce the NEWSCLAIMS benchmark (Section 2) and then describe our zero-shot claim detection framework (Section 3).

Our main contributions are: (1) we propose to use a single pre-trained question answering system in a zero-shot setting for various sub-tasks in claim detection, such as topic-filtering, claim object detection and claimer detection; (2) we show that, using directed questions, a QA model is able to outperform other attribution methods for claimer detection, which requires document-level reasoning; and (3) our proposed claim detection framework achieves state-of-the-art performance on multiple sub-tasks in the NEWSCLAIMS benchmark, outperforming various zero-shot, few-shot and task-specific baselines.

2 NEWSCLAIMS Background

NEWSCLAIMS (Reddy et al., 2021) extends claim detection to extract additional background attributes relating to the claim, such as claim objects and claimers. NEWSCLAIMS evaluates claim detection in the context of an emerging real-world scenario, by considering claims relating to various aspects of COVID-19. Specifically, the topics involved in the benchmark are about the *origin* of the virus, *transmission* of the virus, *cure* for the virus and *protection* from the virus. We refer the reader to Reddy et al. (2021) for detailed definitions and sample claims. Below, we briefly describe the different sub-tasks that we consider within this benchmark and their corresponding baselines, that were introduced in Reddy et al. (2021).

Claim Sentence Detection with Topic-Filtering: This sub-task involves identifying sentences that contain claims relating to COVID-19. For this, Reddy et al. (2021) begin with ClaimBuster (Hassan et al., 2017), which has been trained to identify check-worthy claims (Arslan et al., 2020). In order to then select claims relating to specific topics, from those extracted by ClaimBuster, Reddy et al. (2021) use pre-trained NLI models for zero-shot topic-filtering (Yin et al., 2019). This is done by posing the claim sentence as the premise and constructing a hypothesis for each candidate topic.

Claim Object Detection: A *claim object* relates to what is being claimed in the claim sentence with respect to the topic. Reddy et al. (2021) use zero-shot and few-shot approaches for this sub-task, via zero-shot prompting and leveraging few-shot examples for in-context learning (Brown et al., 2020) and prompt-based fine-tuning (Gao et al., 2021). The prompts are hand-crafted using the topic of the claim sentence.

Claimer Detection: Claimer detection involves identifying the source of the claims made within news articles. These claims can be categorized as either *reported* or those that are made by the author of the news article, i.e., the *journalist*, themselves. Reported claims could originate from people, organizations or other sources in news. The claimer detection sub-task within NEWSCLAIMS involves identifying whether the claim was made by the journalist or who the claimer is, in case it is reported. Since the sub-task requires attribution, Reddy et al. (2021) consider baselines that leverage semantic role labeling (SRL) or are trained on existing attribution datasets (Newell et al., 2018).

3 Method

We first describe the QA model which we use as the pre-trained model to feed directed questions as input. Next, we outline our zero-shot claim detection framework which leverages the above model for sub-tasks such as topic filtering, claim object detection and claimer detection.

3.1 Question Answering Model

The model is a transformer-based extractive question answering system that takes the question and context as input. The QA model has an extractive answer span predictor that predicts the answer spans using the output representations $\mathbf{H}$ from pre-trained language models (LM) (Devlin et al., 2019).

Specifically, the model has a predictor for the beginning α and ending β of the answer span as follows:

$$\boldsymbol{\alpha} = \text{softmax}(\mathbf{W}_1\mathbf{H}) \quad (1)$$

$$\boldsymbol{\beta} = \text{softmax}(\mathbf{W}_2\mathbf{H}) \quad (2)$$

where $\mathbf{W}_1$, $\mathbf{W}_2 \in R^{1\times D}$, D is the dimension of language model's output $\mathbf{H} \in R^{T\times D}$ and T is length of input context. Given $\mathbf{b}$ and $\mathbf{e}$ as the one-hot vectors for the ground-truth start and end positions, the loss function during training is the averaged cross entropy on the two span predictors:

$$\mathcal{L} = -\frac{1}{2}\sum_{t=1}^{T}\{1(\mathbf{b}_t)\log\boldsymbol{\alpha}_t + 1(\mathbf{e}_t)\log\boldsymbol{\beta}_t\} \quad (3)$$

At inference, the answer score for a span (i, j) within the context is computed as $S(i, j) = \alpha_i + \beta_j$, with the highest scoring span taken as the final answer.

3.2 Claim Detection Framework

Given a news article as input, our claim detection framework outputs claims relating to specific topics, along with their corresponding claim objects and claimers. Following Reddy et al. (2021), we use ClaimBuster (Hassan et al., 2017) as the claim-spotting model to first identify sentences that contain claims. Next, we leverage the QA model described in Section 3.1 for topic-filtering, claim object detection and claimer detection. An outline of our claim detection framework can be seen in Figure 1, with each step described in detail below.

Claim Topic Filtering: We propose to do topic filtering by measuring topic relevance as the answer confidence from a QA model, when a question corresponding to the topic is passed as input. We achieve this by formulating a question for each topic, with the claim sentence as context, as shown in Figure 2. The answer score for each question is taken as the corresponding topic relevance. Claims are then filtered based on the highest topic score using a threshold. In comparison, NLI does filtering based on the corresponding highest entailment score. The motivation behind using QA for this sub-task is for the directed questions to be more relevant towards identifying these topics, compared to the implicit inference in NLI.

Claim Object Detection: We pose the *claim object* detection sub-task as an extractive QA task using the same directed questions shown in Figure 2. While the answer score for a question is used for topic filtering, the corresponding answer span is used as the claim object.

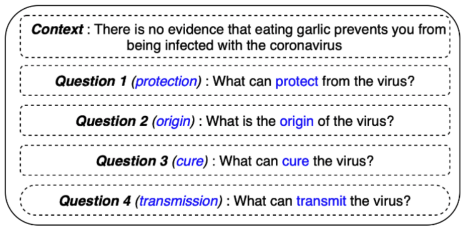

Figure 2: Questions corresponding to individual topics, with the claim sentence as context.

Claimer Detection: We formulate the claimer detection sub-task as a two-step process: first detect the exact claim span within the claim sentence and then identify the claimer. We leverage QA for both steps as follows. The claim span is obtained from the QA model's answer by using "*What is being claimed?*" as the question and the claim sentence as context. Next, for claimer identification, we use the entire news article as context, with the previously extracted claim span inserted into the question, "*Who said that <claim span>?*". We threshold on the answer score to determine if no claimer was identified, in which case, the claim sentence is attributed to the journalist.

4 Experiments

4.1 Setup

The QA model uses *bert-large-uncased* as the underlying language model. It is trained on SQuAD 2.0 (Rajpurkar et al., 2018) for four epochs with a learning rate of 3e-5 and Natural Questions (Kwiatkowski et al., 2019) for one epoch with a learning rate of 1e-5. Batch size is 16 in both cases.

We evaluate our approach on the NEWSCLAIMS benchmark[2] (Reddy et al., 2021). We refer the reader to Reddy et al. (2021) for a detailed description of each of the baselines. The development and test splits comprise 18 news articles with 103 claims and 125 news articles with 786 claims respectively. The thresholds for claim topic filtering and claimer detection were tuned on the development set. All numbers are reported on the test set.

4.2 Results and Analysis

In this section, we evaluate our proposed claim-detection framework for individual sub-tasks such

[2]https://github.com/uiucnlp/NewsClaims

as detecting claims relating to specific topics about COVID-19 (Section 4.2.1), extracting the claim object pertaining to the claim topic (Section 4.2.2) and identifying the claimer (Section 4.2.3).

4.2.1 Claim Sentence Detection

Here, we measure the performance for both zero-shot topic classification and the subsequent filtering. We first evaluate the performance of the QA system for classifying topics, given the claim sentence. Table 1 compares the performance of the NLI and QA systems for zero-shot classification over the four COVID-19 topics. We can see that our zero-shot QA approach considerably outperforms zero-shot NLI, demonstrating that QA can be better at measuring topic relevance. Further, we see that QA is able to overcome the NLI model's inability to distinguish between similar topics such as (*protection* and *cure*) or (*origin* and *transmission*). Some representative examples are in Table 2 with more detailed confusion matrices present in Section A.1 in the Appendix.

Model	Or.	Trans.	Prot.	Cure	All
NLI	56.9	45.1	54.5	3.3	46.6
QA	**85.9**	**64.7**	**63.9**	**66.5**	**72.3**

Table 1: Topic-wise F1 and overall accuracy (both in %) for topic classification given the claim sentence.

Claim Sentence	Topic
This novel coronavirus was believed to have started in a large seafood or wet market, suggesting animal-to-person spread.	**Gold:** Origin **NLI:** Trans. **QA:** Origin
One medication, an antiviral drug called Remdesivir, has been shown in certain studies to improve symptoms and shorten hospital stays.	**Gold:** Cure **NLI:** Protection **QA:** Cure

Table 2: Some examples of incorrect topic predictions from the NLI model which the QA model overcomes. We see that QA, which uses directed questions, is better at being able to distinguish between similar topics such as (*origin* and *transmission*) or (*protection* and *cure*), compared to NLI, which uses implicit inference.

Next, we measure the claim sentence detection performance to evaluate the QA model for topic filtering on the claims outputs by ClaimBuster. Table 3 compares our QA-based topic filtering approach against a pre-trained NLI model, as used in Reddy et al. (2021). We can see that using QA provides up to 5 point improvement in F1, suggesting that the answer confidence from the QA model can be a better estimate for filtering claims relating to these topics, compared to entailment score.

Model	P	R	F1
ClaimBuster	13.0	**86.5**	22.6
ClaimBuster + NLI	21.8	53.3	30.9
ClaimBuster + QA	**30.7**	43.4	**36.0**

Table 3: Performance (in %) of various systems for detecting sentences with claims relating to COVID-19.

4.2.2 Claim Object Detection

For the claim object detection sub-task, we compare our QA approach with various zero-shot and few-shot approaches used in Reddy et al. (2021). Table 4 shows the performance of the QA system along with different prompt-based approaches, that leverage generative language models to output the claim object. While GPT-3 and T5 show competitive performance in few-shot settings, our zero-shot QA approach outperforms by more than 5 points.

Approach	Model	Type	F1
Prompting	GPT-3	Zero-shot	15.2
Prompting	T5	Zero-shot	11.4
In-context learning	GPT-3	Few-Shot	51.9
Prompt-based fine-tuning	T5	Few-Shot	51.6
QA	BERT	Zero-shot	**57.0**

Table 4: F1 score (in %) of different zero-shot and few-shot approaches for the claim object detection sub-task.

4.2.3 Claimer Detection

The claimer detection sub-task is evaluated based on the classification F1 for predicting which claims are from journalists, along with a string-match F1 (Rajpurkar et al., 2018) for extracting the mention of the claimer in case of reported claims. Table 5 compares our zero-shot QA-based approach for claimer detection with the Semantic Role Labeling (SRL) and PolNeAR news-attribution (Newell et al., 2018) baselines from Reddy et al. (2021).

Model	Overall F1	Reported	Journalist
SRL	41.7	23.5	**67.2**
PolNeAR	42.3	25.5	65.9
QA	**50.1**	**39.8**	64.4

Table 5: F1 (in %) for identifying the claimer. Numbers for reported and journalist are shown separately.

To understand why sentence-level approaches, such as SRL and PolNeAR, can be very competitive at identifying claims that are from the journalist, we manually analyzed some examples. We observed that claims from the journalist are usually made in a first-person point of view, which can be identified

by sentence-level reasoning. Table 6 shows some examples for claims that directly come from the journalist and those that are reported from other sources. It can be seen that those that come from the journalist do not involve cue words and are usually presented in a first-person point of view. This explains the competitive performance of SRL and PolNeAR for predicting which claims come from the journalist.

Claim Sentence	Type
It is not yet known if remdesivir is safe for the treatment of COVID-19.	Journalist
Inhaling bleach fumes is dangerous and will not kill viruses that are already inside.	Journalist
An earlier version of this article claimed a laboratory near Wuhan could be to blame for the outbreak of coronavirus.	Reported
The South China Agricultural University in Guangzhou says that two of its researchers have identified the pangolin as the potential source of nCoV-2019.	Reported

Table 6: Some examples for when the claimer is journalist vs when it is a reported claim.

However, promising results in using QA for claimer detection can be seen for reported claims, which can require document-level reasoning skills for identifying the claimer. Table 7 breaks down the performance for reported claims based on where the claimer mention is present. We can see that QA outperforms other attribution approaches for both cases, with even larger gains for when the claimer is present outside the claim sentence (which necessitates cross-sentence attribution).

Model	In-sentence	Out-of-sentence
SRL	35.8	2.4
PolNeAR	38.9	2.7
QA	**46.2**	**29.0**

Table 7: F1 (in %) for claimer detection for when it is present within or outside the claim sentence.

4.2.4 Analysis of Question Templates

Instead of hand-crafting questions from topics, we experiment with using each topic directly as a question to be fed as input to the QA model. Table 8 shows the performance for the claim sentence detection, topic classification and claim object detection sub-tasks for the settings where a directed question is manually constructed from the topic, compared to where the topic is used as the question. We can see while claim sentence detection and topic classification performance are almost similar, the performance is considerably better for claim object detection in case of directed (hand-crafted) questions. This implies that the answer confidence from a weakly-defined question (by just using topic as the question) is still a reliable measure of topic relevance. However, directed questions are useful for getting the right answer spans, which is crucial in case of claim object detection. Note that hand-crafting the question does not need considerable effort, as it mainly involves converting the topic into an information-seeking format by prepending with a "what", for e.g.: "*protection from the virus*" –> "*What can protect from the virus?*", "*transmission of the virus*" –> "*What can transmit the virus?*"

Sub-Task	Hand-crafted	Topic
Claim Sentence Detection	36.0	35.8
Topic Classification	72.3	73.2
Claim Object Detection	57.0	47.0

Table 8: Comparison of performance (in %) for claim sentence detection (F1), topic classification (Acc.) and claim object detection (F1) sub-tasks when using questions that are *hand-crafted* from the topic vs using the *topic* directly as the question.

5 Conclusions and Future Work

We propose a new claim detection framework that leverages zero-shot QA with directed questions for various sub-tasks such as topic filtering, claim object detection and claimer detection. We show that these questions can be adept at identifying topic relevance for claims related to COVID-19. We demonstrate that QA can be leveraged for claimer detection with document-level attribution, while considerably outperforming attribution systems that can be limited by sentence-level reasoning. Future work involves building a unified model that can extract claims and corresponding attributes together, without the need for separate components for each individual attribute.

Acknowledgement

This research is based upon work supported by U.S. DARPA AIDA Program No. FA8750-18-2-0014. FA8750-19-2-1004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A Benchmark Dataset of Check-worthy Factual Claims. In *14th International AAAI Conference on Web and Social Media*. AAAI.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Association for Computational Linguistics (ACL)*.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, et al. 2017. Claimbuster: The first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Edward Newell, Drew Margolin, and Derek Ruths. 2018. An attribution relations corpus for political news. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Silvia Pareti. 2016. PARC 3.0: A corpus of attribution relations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3914–3920, Portorož, Slovenia. European Language Resources Association (ELRA).

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Revanth Gangi Reddy, Sai Chinthakindi, Zhenhailong Wang, Yi R Fung, Kathryn S Conger, Ahmed S Elsayed, Martha Palmer, and Heng Ji. 2021. Newsclaims: A new benchmark for claim detection from news with background knowledge. *arXiv preprint arXiv:2112.08544*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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)*, pages 3914–3923.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878.

A Appendix

A.1 Topic Classification performance

Figures 3 and 4 show the topic classification confusion matrices for the NLI and QA models respectively. As Reddy et al. (2021) point out, the NLI model predominantly suffers from classifying claims related to *cure* as *protection* and those related to *origin* as *transmission*. However, our QA model is able to overcome this, which explains the improved performance in topic classification and topic filtering for claim sentence detection.

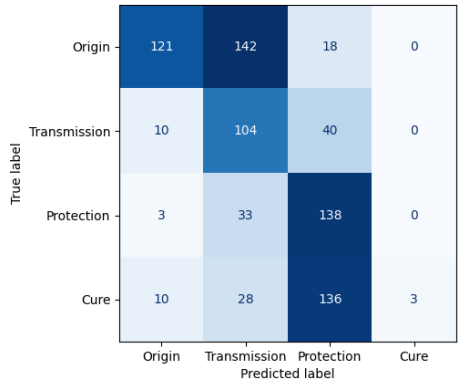

Figure 3: Topic classification confusion matrix for the the NLI model.

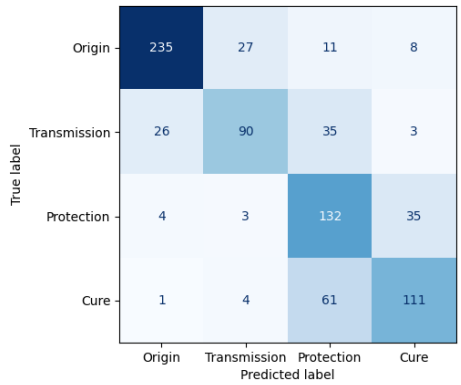

Figure 4: Topic classification confusion matrix for the the QA model.

A.2 QA Training Datasets

We give a brief overview of the datasets used to train the extractive QA model.

SQuAD: SQuAD1.1 (Rajpurkar et al., 2016) is an extractive machine reaching comprehension dataset containing questions posed by crowdworkers on a set of wikipedia articles. SQuAD2.0 (Rajpurkar et al., 2018) combines the 100,000+ questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones.

Natural Questions: NQ (Kwiatkowski et al., 2019) is an english machine reading comprehension benchmark which contains 300,000+ questions from Google users, and requires systems to read and comprehend entire Wikipedia articles to answer them.