ColumbiaNLP at SemEval-2019 Task 8: The Answer is Language Model Fine-tuning

Tuhin Chakrabarty
Columbia University
Department Of Computer Science
tc2896@columbia.edu

Smaranda Muresan
Columbia University
Data Science Institute
smara@columbia.edu

Abstract

Community Question Answering forums are very popular nowadays, as they represent effective means for communities to share information around particular topics. But the information shared on these forums is often not correct or misleading. This paper presents the ColumbiaNLP submission for the SemEval-2019 Task 8: Fact-Checking in Community Question Answering Forums. We show how fine-tuning a language model on a large unannotated corpus of old threads from Qatar Living forum helps us to classify question types (factual, opinion, socializing) and to judge the factuality of answers on the shared task labeled data from the same forum. Our system finished 4th and 2nd on Subtask A (question type classification) and B (answer factuality prediction), respectively, based on the official metric of accuracy.

1 Introduction

Community Question Answering (cQA) forums such as StackOverflow, Yahoo! Answers, and Quora are very popular nowadays, as they represent effective means for communities to share information and to collectively satisfy their information needs. Questions asked on these sites can be of different types, and the answers can often be false, misleading or irrelevant.

SemEval-2019 Task 8 is structured around two subtasks. Subtask A is a question classification task, where the questions types are:

- **Factual**: The question is asking for factual information, which can be answered by checking various information sources, and it is not ambiguous (e.g., "What is Ooredoo customer service number?").
- **Opinion**: The question asks for an opinion or an advice, not for a fact. (e.g., "Can anyone recommend a good Vet in Doha?"")
- **Socializing**: Not a real question, but intended for socializing or for chatting. This can also mean expressing an opinion or sharing some information, without really asking anything of general interest (e.g., "What was your first car?")

Subtask B is an answer classification task: are the answers to *factual questions* factual or not, and if they are factual are they true or false:

- **Factual - TRUE**: The answer is True and can be proven with an external resource. (Q: "I wanted to know if there were any specific shots and vaccinations I should get before coming over [to Doha]."; A: "Yes there are; though it varies depending on which country you come from. In the UK; the doctor has a list of all countries and the vaccinations needed for each.").
- **Factual - FALSE**: The answer gives a factual response, but it is False, it is partially false or the responder is unsure about (Q:"Can I bring my pitbulls to Qatar?"; A: "Yes you can bring it but be careful this kind of dog is very dangerous.").
- **Non-Factual**: When the answer does not provide factual information to the question; it can be an opinion or an advice that cannot be verified (e.g., "It's better to buy a new one.").

2 Related Work

Yu and Hatzivassiloglou (2003) separated opinions from fact, at both the document and sentence level.

(Mihaylova et al., 2018) were the first to propose a novel multi-faceted model for fact checking of answers on community question answering forums. Their proposed model captures information

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1144–1148
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

OPINION	FACTUAL	SOCIALIZING
586	311	254

Table 1: Size of Subtask A dataset (question types).

from the answer content (what is said and how), from the author profile (who says it), from the rest of the community forum (where it is said), and from external authoritative sources of information (external support). (Nakov et al., 2017) proposed models for credibility assessment in community question answering forums. However, credibility is different from veracity as it is a subjective perception about whether a statement is credible, rather than verifying whether it is true/false as a matter of fact.

Recently there has been a lot of attention on building models for fact checking. (Thorne et al., 2018) introduce a new publicly available dataset for fact extraction and verification (FEVER Shared Task). The dataset consists of 185,445 claims generated by altering sentences extracted from Wikipedia, and the task is to classify claims as SUPPORTED, REFUTED or NOTENOUGHINFO. However, the verification of the claims is limited to a particular database (namely Wikipedia) unlike Subtask B. Also, the claims are inherently less noisy as compared to answers in Community Question Answering forums.

Pre-trained language models have been recently used to achieve state-of-the-art results on a wide range of NLP tasks (e.g., sequence labeling and sentence classification). Some of the recent works that have employed pre-trained language models include (Howard and Ruder, 2018), (Peters et al., 2018), (Yang et al., 2018), and (Radford et al., 2018). In this paper, we show the effectiveness of the Universal Language Model Fine-tunig (ULMFiT) method (Howard and Ruder, 2018) for both question classification and answer fact checking.

3 Data

One of key challenges for both Subtask A and B is the limited amount of annotated data. This poses a challenge to apply state-of-the-art neural discrimination models without using additional data.

3.1 Labeled Data

Subtask A has a total of 1,118 questions divided into three types. Table 1 show the class distribution. Subtask B has a total of 495 answers divided into three types. Table 2 shows the class distribution.

TRUE	FALSE	NON-FACTUAL
166	135	194

Table 2: Size of Subtask B dataset (answer types).

QUESTIONS	ANSWERS
189,941	1,894,456

Table 3: External unannoted questions and answers.

3.2 Unlabeled Data

The task allows the use of external unnannoted data of 189,941 threads from Qatar Living Forum. Each of these threads have questions and answers just as our training data but without any labels. These threads may contain enough information to estimate the factuality of the answers in Subtask B as well as linguistic patterns in the questions asked for Subtask A. We refer to the resulting collection of comments as the **QL** dataset.

4 Model and Analysis

As the QL data is from the same distribution as our shared task data (Quatar Living), we need a method of incorporating this dataset into our models for both subtasks. We use a language model fine-tuning approach, which requires only unlabeled data similar to the task of interest.

The Universal Language Model Fine-Tuning method (ULMFiT) (Howard and Ruder, 2018) consists of the following stages: a) General-domain LM pre-training b) Task-specific LM fine-tuning and c) Task-specific classifier fine-tuning. In stage (a), the language model is trained on Wikitext-103 (Merity et al., 2017) consisting of 28,595 pre-processed Wikipedia articles and 103 million words capturing general properties of language. Stage (b) fine-tunes the LM on task-specific data, as no matter how diverse the general-domain data used for pre-training is, the data of the target task will likely come from a different distribution. In stage (c), a classifier is trained on the target task, fine-tuning the pre-trained LM but with an additional layer for class prediction. The models use a stacked Long Short Term Memory (LSTM) network to represent each sentence. For stages (a) and (b), the output of the LSTM is used to make a prediction of the next token and the parameters from stage (a) are used to initialize stage

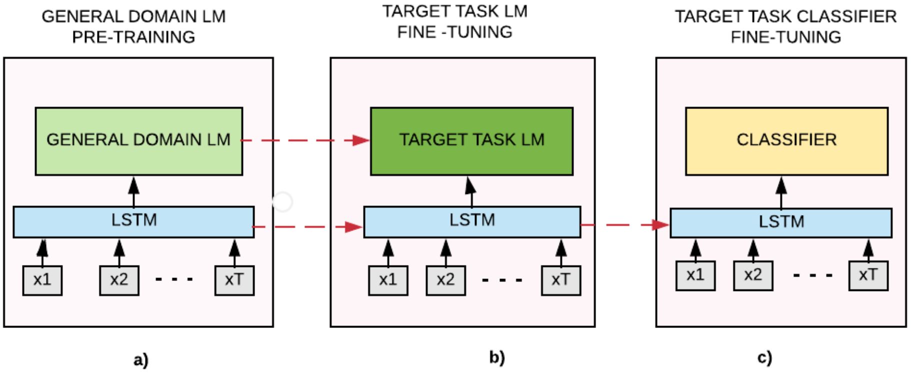

Figure 1: Schematic of ULMFiT showing the three stages. The dashed arrows indicate that the parameters from the previous step were used to initialize the next step.

Task Specific LM Fine-Tuning	QL LM Fine-Tuning
65	**81**

Table 4: Accuracy on the test splits while doing cross validation on training data for Subtask A

(b). For stage (c), the model is initialized with the same LSTM but with a new classifier layer given the output of the LSTM.

This process is illustrated in Figure 1. We refer the reader to Howard and Ruder (2018) for further details. In our work, we maintain stages (a) and (c) but modify stage (b) so that we fine-tune the language model on the unlabelled data rather than the task-specific data. The goal of ULMFiT is to allow training on small datasets of only a few hundred examples, but our experiments will show that fine-tuning the language model on the QL data improves over only task-specific LM fine-tuning.

4.1 Subtask A

For Subtask A we fine-tune a language model on the 189,941 questions from the QL dataset. Our initial experiments show that fine-tuning the LM on the QL dataset give large performance gains over fine-tuning on task specific data as demonstrated in Table 4.

Lets take the following question:

> *Ramadan Working Hours? For companies who are operating 5 days a week; what are your timings? Ours is 8:00am to 3:00pm.?*

This is a **Factual** question, but the task-specific LM fine-tuning labels it as **Socializing**, while fine-tuning on QL data allows the model to correctly classify it as **Factual**. To understand why this happens, we delve deeper into the unlabeled data set where we find multiple similar questions based on TF-IDF similarity, demonstrating that the LM Fine-Tuning on QL data learns representations of questions based on discriminatory phrases.

- ***Ramadan Working Hours?*** *Eid holidays announced*
- ***Ramadan Working Hours?*** *good morning; Did anybody knows what is the right time* ***timing*** *or* ***working hours*** *during* ***Ramadan****? Thanks and advance.*
- ***Ramadan Working Hours?*** *Ministry of Civil Service Affairs and Housing has issued a circular in this regard defining the restricted* ***working hours****. Can somebody help me to find the English translation of that. Thank you*
- *help pls! can somebody tell me the* ***Ramadan Working Hours?*** *of ministry of Foreign Affairs???*

On the official test data, the ULMFiT approach where the target task classifier is fine-tuned on the LM fine-tuned on questions from QL data gives us an accuracy of **83** placing us 4th on the leaderboard.

4.2 Subtask B

For Subtask B we followed a similar approach of LM fine-tuning. We obtained representations of answers by fine-tuning a LM on 1,894,456 answers from the QL dataset. Next, we obtained

ANSWER	AVG COSINE SIMILARITY
Medical Check is for everyone mate	0.81
The test is done for everyone; but is restricted to the above categories u mentioned.	0.44
Regardless what your job is...everybody gets tested for hepatitis B/C cheers Never say never	0.76
As I told u hepatitis B/C are checked for everyone applying for residence permit.If the result is positive u go back where u came from. And I know all of the above because my husband is a consultant pathologist. cheers	0.72

Table 5: Average Cosine Similarity scores of contextual representations of each answer to every other answer in the thread

representations by fine-tuning a LM on 1,894,456 question-answer pairs, in order to capture whether an answer is actually suited for the question asked or something irrelevant. An answer which is relevant to the question asked can then be easily discerned from an irrelevant one by a discriminative classifier.

Our model did not take into account external evidence from search engines as done by (Mihaylova et al., 2018), so we had to rely on intra-forum evidence for factuality features. Our hypothesis is that for factual questions, the answers which are factually true are similar to each other, while answers which are false or irrelevant are different from other answers. We incorporated this behaviour in our model: for every answer we computed the cosine similarity between the contextual representation of the answer obtained from last layer of the LSTM used to train the language model for answers. For each answer, we averaged the cosine similarity between that answer and the other answers in the same thread.

For example, take the question:

> *Hi all; are hepatitis B and C checked for in the medical test for non-medical professionals? Basically;I have been getting conflicting information on this. Some say that Hep B and C are tested for everyone applying for residence permit. Others say that only medical professionals; primary school teachers and food handlers are tested for Hep B and C. Please discuss!*

From Table 5 we see that the answer with the lowest cosine similarity (0.44) is the answer which is factually false, compared to the other answers which are factually true and have a higher cosine similarity.

We use these answer representations, question-answer pair representations and the average cosine similarity as features to train an XGB classifier to obtain our final results. For threads with only 1 answer we took the cosine similarity as 0.5.

FEATURES	ACCUARCY
LM on answers	61
LM on Q-A pairs	64
LM on answers and Q-A pairs	72
LM on answers and Q-A pairs + Avg Cosine Similarity	**79**

Table 6: Ablation scores for Subtask B on the final test set for Task 8.

Table 6 shows us the ablation scores for each approach. The best accuracy obtained is a combination of all the approaches together. We obtain an accuracy of **79** placing us 2nd on the leaderboard. The absence of true labels for both the dev and the test set prevents us from conducting an error analysis.

5 Implementation

The language model is fine tuned for 15 epochs as done in the ULMFIT original paper for both Subtasks. For classifier fine-tuning we use the same hyper-parameters as (Howard and Ruder, 2018) except the learning rate which is set to .0001 . We train our classifier for 5 epochs on both sub-tasks. Each model was run 10 times to account for variance and the results reported for both the tasks are the average of 10 runs. We did not use any special pre-processing technique and use the same approach as done in the ULMFIT paper, i.e clean up extra spaces, tab characters, new line characters and other characters and replace them with standard ones. We also use Spacy library to tokenize the data. The implementation can be found here [1]

6 Conclusion

We show that fine-tuning a language model on a large unsupervised corpus from the same community forum helps us achieve better accuracy for question classification. Most community

[1] https://github.com/fastai/fastai/blob/master/courses/dl2/imdb.ipynb

question-answering forums have such unlabeled data, which can be used in the absence of large labeled training data .

For answer classification we show how we can leverage information from previously answered questions on the thread through language model fine tuning. Our experiments also show that modeling an answer individually is not the best idea for fact-verification and results are improved when considering the context of the question.

Determining factuality of answers definitely requires modeling world knowledge or external evidence. The questions asked are often very noisy and require reformulation. As a future step we would want to incorporate external evidence from the internet in the factual answer classification problem.

References

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 328–339.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. In proceedings of the international conference on learning representations.

Tsvetomila Mihaylova, Preslav Nakov, Llu'is Marquez, Alberto Barron-Cede'no, Mitra Mohtarami, Georgi Karadzhov, and James Glass. 2018. Fact checking in community forums. In *Association for the Advancement of Artificial Intelligence*.

Preslav Nakov, Tsvetomila Mihaylova, Llu'is Marquez, Y Shiroya, and I Koychev. 2017. Do not trust the trolls: Predicting credibility in community question answering forums. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 551–560.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of NAACL-HLT 2018.*, pages 809–819.

Zhilin Yang, Jake Zhao, Bhuwan Dhingra, Kaiming He, William W. Cohen, Ruslan Salakhutdinov, and Yann LeCun. 2018. Glomo: Unsupervisedly learned relational graphs as transferable. In *arXiv:1806.05662*.

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*.