CodeForTheChange at SemEval-2019 Task 8: Skip-Thoughts for Fact Checking in Community Question Answering

Adithya Avvaru[1,3] and **Anupam Pandey**[2,3]
[1]Teradata India Pvt. Ltd, India
[2]Qubole, India
[3] International Institute of Information Technology, Hyderabad, India
{adithya.avvaru,anupam.pandey}@students.iiit.ac.in

Abstract

Community Question Answering (cQA) is one of the popular Natural Language Processing (NLP) problems being targeted by researchers across the globe. Couple of the unanswered questions in the domain of cQA are 'can we label the questions/answers as factual or not?' and 'Is the given answer by the user to a particular factual question is correct and if it is correct, can we measure the correctness and factuality of the given answer?'. We have participated in SemEval-2019 Task 8 which deals with these questions. In this paper, we present the features used, approaches followed for feature engineering, models experimented with and finally the results. Our primary submission with accuracy (official metric for SemEval Task 8) of 0.65 in Subtask B (Answer Classification) and 0.63 in Subtask A (Question Classification) stood at 6^{th} and 16^{th} places respectively.

1 Introduction

Community Question Answering (cQA) forums such as Quora, StackOverflow, Yahoo! Answers, Qatar Living etc., now-a-days are fast and effective means of getting answers for any question. But the answers may or may not be correct and factual always. The focus of cQA research, for the last few couple of years, is revolving around determining the model which predicts the best answer for the question, given a question and a number of answers (might be hundreds or even thousands in number).

cQA is one of the popular problems being constantly in focus of SemEval organizers since 2015. The subtasks that were targeted earlier include (i) classifying the answer to a particular question as good or potentially good or bad in 2015[1], (ii) three reranking subtasks i.e., Question-Comment Similarity, Question-Question Similarity and Question-External Comment Similarity in 2016[2] and (iii) Question Similarity (QS) to detect duplicate questions and Relevance Classification (RC) in 2017[3]. Contrary to earlier tasks of SemEval focusing mainly on classification and similarity of questions and/or answers and/or comments, SemEval-2019 targets the factuality of the questions (whether the question is factual or not) and the factuality of the answers (whether the answers provided to the factual questions are factual or not). The tasks become more challenging as data have noisy (like *!!!*), and unstructured (like *Oh..*) words.

SemEval-2019 Task 8 features the following two subtasks:

Subtask A (Question Classification) - determine whether a question asks for a factual information, an opinion/advice or is just socializing. Example from the "Qatar Living" forum given in competition page[4] for this subtask is as follows:
Q: I have heard its not possible to extend visit visa more than 6 months? Can U please answer me.. Thankzzz...
answer 1: Maximum period is 9 Months....
answer 2: 6 months maximum
answer 3: This has been answered in QL so many times. Please do search for information regarding this. BTW answer is 6 months.
This subtask aims at building models to detect true factual information in cQA forums.

Subtask B (Answer Classification) - determine whether an answer to a factual question is true, false, or does not constitute a proper answer.
This subtask aims at building models that classify the answers into the following three categories, given a factual question: ***a)*** *Fac-*

[1]http://alt.qcri.org/semeval2015/task3/
[2]http://alt.qcri.org/semeval2016/task3/
[3]http://alt.qcri.org/semeval2017/task3/
[4]https://competitions.codalab.org/competitions/20022

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1138–1143
Minneapolis, Minnesota, USA, June 6–7, 2019.

tual - True ***b)*** *Factual - False and* ***c)*** *Non-Factual.* The examples for each of them are as follows:

- **Factual - True:**
  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:**
  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:**
  Q: Which is suggested - buy a new car or an used one?
  A: Its better to buy a new one.

We participated in both the subtasks of SemEval-2019 Task 8. For detailed description of the task, different approaches used by other participants and results obtained by all the participants, please refer the task description paper (Mihaylova et al., 2019).

The rest of the paper is organized as follows: Section 2 describes the related work. Section 3 describes the data used for this SemEval task. Sections 4 and 5 elucidate the system architecture (feature extraction and model building) and experimentation details (along with the results) respectively. Section 6 concludes the paper with focus on future research on this task.

2 Related Work

Some of the earlier works on cQA include the use of classification models - Support Vector Machines(SVMs) (Šaina et al., 2017; Nandi et al., 2017; Xie et al., 2017; Mihaylova et al., 2016; Wang and Poupart, 2016; Balchev et al., 2016) for Similarity tasks; Convolutional Neural Networks (CNNs) for Similarity tasks (Šaina et al., 2017; Mohtarami et al., 2016) and for answer selection (Zhang et al., 2017); Long-Short Term Memory (LSTM) model for answer selection (Zhang et al., 2017; Feng et al., 2017; Mohtarami et al., 2016); Random Forests (Wang and Poupart, 2016); LDA topic language model to match the questions at both the term level and topic level (Zhang et al., 2014); translation based retrieval models (Jeon et al., 2005; Zhou et al., 2011); XgBoost (Feng et al., 2017) and Feedforward Neural Network (NN) (Wang and Poupart, 2016).

All of the above related works on cQA used the features such as Bag of Words (BoW) (Franco-Salvador et al., 2016), Bag of vectors (BoV) (Mohtarami et al., 2016), Lexical features (for example, Cosine Similarity, Word Overlap, Noun Overlap, N-gram Overlap, Longest Common Substring/Subsequence, Keyword and Named Entity features etc.) (Franco-Salvador et al., 2016; Mohtarami et al., 2016; Nandi et al., 2017); Semantic features (for eg, Distributed representations of text, Knowledge Graphs, Distributed word alignments, Word Cluster Similarity, etc.) (Franco-Salvador et al., 2016); Word Embedding Features (like Word2vec[5] (Mikolov et al., 2013), GloVe[6](Pennington et al., 2014) etc.) (Wang and Poupart, 2016; Mohtarami et al., 2016; Nandi et al., 2017); Metadata-based features (like user information, answer length, question length, question marks in answer, question to comment length etc.) (Mohtarami et al., 2016; Mihaylova et al., 2016; Xie et al., 2017).

Another related task to cQA is Fact Checking in Community Forums (Mihaylova et al., 2018). This work doesn't involve classification of questions/answers based on factuality but it determines the veracity of the answer given a particular question. This work is related to our task in a way that the data being used in our task is annotated and released to the research community by Tsvetomila Mihaylova and her team.

The fact that this research problem is relatively new, the strengths of the scalable gradient tree boosting algorithm, XGBoost (Chen and Guestrin, 2016) and distributed sentence encoder, Skip-Thought vectors (Kiros et al., 2015) are not explored yet. We tried to apply and combine these two effective methods for finding factual nature of the questions and answers.

3 Data Description

The data for both Question Classification - Subtask A and Answer Classification - Subtask B, is organized into train, dev and test sets. The number of samples in each of these datasets is shown in the Table 1.

[5]https://code.google.com/archive/p/word2vec/
[6]http://nlp.stanford.edu/projects/glove/

Subtask	Datasets		
	Train	Dev	Test
A	1118	239	935
B	495	112	310

Table 1: Dataset Description

The data, in Question Classification, has both subject and body for each question. Similarly, for Answer Classification, the data has question subject, question body and an answer (as a comment text). The data of both the subtasks also have other information related to meta-data like user information, date and time of the question and answer post. The detailed description of data can be seen in task description paper (Mihaylova et al., 2019).

4 System Description

4.1 Feature Extraction

4.1.1 Data pre-processing

We have applied some basic preprocessing tasks like removing URLs, converting text to lowercase along with removing stopwords.

4.1.2 Extract Skip-Thought vectors

We choose Skip-Thought Vectors as word embeddings for this task mainly because these are highly generic sentence representations unlike GloVe or Word2Vec which averages word embeddings of each individual word to calculate the word embedding for a complete sentence.

In subtask A, we have retrieved Skip-Thought vectors for question body and question subject. In subtask B, we extracted Skip-Thought vectors for question body, question subject and answer comment. For both the subtasks, we have used the code[7] written by the Skip-Thought vectors' authors.

4.2 Model Building

Once we have extracted Skip-Thought vectors, we used these vectors to train different models - AdaBoost Classifier (only in case of Subtask B), DecisionTree Classifier, RandomForest Classifier, ExtraTrees Classifier, XGBoost Classifier and Multi-layer Neural Network with dropout layers in between, Adam optimizer and softmax activation in the final layer. The hyper-parameters of all the models is determined by applying GridSearch with 10-fold cross-validation. The hyper-parameters are shown in the Table 2.

Classifier	Hyper-parameters
Decision Tree	min_samples_split = 2
Random Forest	n_estimators = 25 min_samples_leaf = 1 min_samples_split = 2
Extra Trees	n_estimators = 20 max_features = 37
XGBoost	learning_rate = 0.1 n_estimators = 100 max_depth = 5 objective = 'multi:softprob'
Adaboost	n_estimators = 45 learning_rate = 1.0

Table 2: Hyper-parameters used for models

5 Evaluation and Results

5.1 Subtask A (Question Classification)

For this subtask, we extract Skip-Thought vectors as described in section 4.1.2. Once we get these two vectors, we generated four different combinations of vectors - (i) question body only, (ii) question subject only, (iii) concatenation vector of both question body and question subject and (iv) average vector of both question body and question subject. We trained all the models mentioned in the section 4.2 with each one of these vectors. The evaluation scores for these models on test data are shown in the Table 3.

5.2 Subtask B (Answer Classification)

For this subtask, we extract Skip-Thought vectors as described in section 4.1.2. Once we get these three vectors, we generated two different combinations of vectors - (i) concatenation vector of question body, question subject & answer and (ii) average vector of question body, question subject & answer. We trained all the models mentioned in the section 4.2 using each one of these embedding vectors. The evaluation scores for these models (except MAP scores) on test data are shown in the Table 4.

In both the tables 3 and 4, the column **Vector** represents Skip-Thought vector combination type (whether it is body only (in case of Subtask A) or subject only (in case of Subtask A) or

[7] https://github.com/ryankiros/Skip-Thoughts

Model	Vector	Accuracy	F-score	Avgrec
Decision Tree	**Bodies**	0.5728	0.3550	0.3893
	Subjects	0.5567	0.3308	0.3626
	Avg	0.5966	0.3904	0.4277
	Concat	0.5691	0.3498	0.3909
Extra Trees	**Bodies**	0.5406	0.3015	0.4075
	Subjects	0.5329	0.2992	0.4002
	Avg	0.5315	0.2902	0.4015
	Concat	0.5509	0.3158	0.4180
Random Forest	**Bodies**	0.5476	0.3119	0.4161
	Subjects	0.5329	0.2971	0.3950
	Avg	0.5567	0.3275	0.4236
	Concat	0.5446	0.3153	0.4139
Neural Network	**Bodies**	0.6849	0.5118	0.5426
	Subjects	0.6338	0.4404	0.4677
	Avg	**0.6884**	**0.5228**	**0.5561**
	Concat	0.6740	0.5007	0.5405
XGBoost	**Bodies**	0.6268	0.4382	0.5194
	Subjects	0.5959	0.4032	0.4646
	Avg**	0.6366	0.4474	0.5195
	Concat*	0.6299	0.4416	0.5130

Table 3: Evaluation scores for Subtask A
$*$ - marks the scores of our primary submission
$**$ - marks the scores of our contrastive submission
Row in bold - post evaluation accuracy score (improved over actual submission)

concatenation of vectors of body, subject and answer/comment or average of vectors of body, subject and answer/comment). On dev data set, XGBoost Classifier with concatenated Skip-Thought vectors generated best scores for both subtasks. Hence, these are part of final submissions.

However, the rows which are marked in bold (in both subtasks) produced best accuracy score with Multi-layer Neural Network Classifier beating the best score of our CodaLab final submission. The Multi-layer Neural Network is designed to have an input layer, 2 hidden layers and an output layer with "relu" activations at input and hidden layers and "sigmoid" activation at output layer. All the layers are trained with 50 neurons except the output layer which has one neural node. This model counters overfitting problem by introduction of intermittent Dropout layers.

Another interesting observation that we found is the models, surprisingly, performed better when URLs are kept in the text compared to when URLs were removed.

Model	Vector	Accuracy	F-score	Avgrec
Decision Tree	**Avg**	0.5354	0.2755	0.3791
	Concat	0.5438	0.2843	0.3284
Extra Trees	**Avg**	0.5763	0.2845	0.3229
	Concat	0.6021	0.3150	0.3558
Random Forest	**Avg**	0.6215	0.3068	0.3285
	Concat	0.6172	0.2890	0.2943
Adaboost	**Avg**	0.5570	0.2607	0.2813
	Concat	0.5743	0.2612	0.2564
Neural Network	**Avg**	0.6129	0.3434	0.4036
	Concat	**0.6752**	**0.3420**	**0.3559**
XGBoost	**Avg****	0.6150	0.3225	0.3529
	Concat*	0.6537	0.3252	0.1555

Table 4: Evaluation scores for Subtask B
$*$ - marks the scores of our primary submission
$**$ - marks the scores of our contrastive submission
Row in bold - post evaluation accuracy score (improved over actual submission)

6 Conclusion

The earlier works on cQA didn't use Skip-Thought vectors, to the best of our knowledge. Hence, we used these vectors for both subtasks. We also have tried unique combinations of Skip-Thought vectors of question body, question subject and comments/answers (only in case of Subtask B) - either concatenation or average of vectors with different models. Out of all the models, concatenated Skip-Thought vectors with XGBoost Classifier generated best result out of all the combinations; as a result of which we stood 6^{th} in Subtask B and 16^{th} in Subtask A. However, post-evaluation submission which used concatenated Skip-Thought vectors with Neural Network classifier produced better accuracy score of 0.6752 compared to 0.6537 (which is official best result for Task B) and 0.6884 compared to 0.6299 (which is official best result for Task A). However, in future we would like to extend our work with other word embeddings like Word2vec, GloVe and BERT (Devlin et al., 2018) features and compare the results with current work using Skip-Thought vectors.

Acknowledgments

This research is supported by Teradata, India and Qubole, India in collaboration with Language Technologies Research Centre (LTRC) of International Institute of Information Technology, Hyderabad, India.

References

Daniel Balchev, Yasen Kiprov, Ivan Koychev, and Preslav Nakov. 2016. PMI-cool at SemEval-2016 Task 3: Experiments with PMI and Goodness Polarity Lexicons for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 844–850, San Diego, California. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 785–794. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Wenzheng Feng, Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2017. Beihang-MSRA at SemEval-2017 Task 3: A Ranking System with Neural Matching Features for Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 280–286, Vancouver, Canada. Association for Computational Linguistics.

Marc Franco-Salvador, Sudipta Kar, Thamar Solorio, and Paolo Rosso. 2016. UH-PRHLT at SemEval-2016 Task 3: Combining Lexical and Semantic-based Features for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 814–821, San Diego, California. Association for Computational Linguistics.

Jiwoon Jeon, W Bruce Croft, and Joon Ho Lee. 2005. Finding Similar Questions in Large Question and Answer Archives. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 84–90. ACM.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-Thought Vectors. *arXiv preprint arXiv:1506.06726*.

Tsvetomila Mihaylova, Pepa Gencheva, Martin Boyanov, Ivana Yovcheva, Todor Mihaylov, Momchil Hardalov, Yasen Kiprov, Daniel Balchev, Ivan Koychev, Preslav Nakov, Ivelina Nikolova, and Galia Angelova. 2016. SUper Team at SemEval-2016 Task 3: Building a Feature-Rich System for Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 836–843, San Diego, California. Association for Computational Linguistics.

Tsvetomila Mihaylova, Georgi Karadzhov, Atanasova Pepa, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact Checking in Community Question Answering Forums. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '19, Minneapolis, MN, USA.

Tsvetomila Mihaylova, Preslav Nakov, Lluis Marquez, Alberto Barron-Cedeno, Mitra Mohtarami, Georgi Karadzhov, and James Glass. 2018. Fact Checking in Community Forums. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mitra Mohtarami, Yonatan Belinkov, Wei-Ning Hsu, Yu Zhang, Tao Lei, Kfir Bar, Scott Cyphers, and Jim Glass. 2016. SLS at SemEval-2016 Task 3: Neural-based Approaches for Ranking in Community Question Answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 828–835, San Diego, California. Association for Computational Linguistics.

Titas Nandi, Chris Biemann, Seid Muhie Yimam, Deepak Gupta, Sarah Kohail, Asif Ekbal, and Pushpak Bhattacharyya. 2017. IIT-UHH at SemEval-2017 Task 3: Exploring Multiple Features for Community Question Answering and Implicit Dialogue Identification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 90–97, Vancouver, Canada. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Filip Šaina, Toni Kukurin, Lukrecija Puljić, Mladen Karan, and Jan Šnajder. 2017. TakeLab-QA at SemEval-2017 Task 3: Classification Experiments for Answer Retrieval in Community QA. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 339–343, Vancouver, Canada. Association for Computational Linguistics.

Hujie Wang and Pascal Poupart. 2016. Overfitting at SemEval-2016 Task 3: Detecting Semantically Similar Questions in Community Question Answering Forums with Word Embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 861–865, San Diego, California. Association for Computational Linguistics.

Yufei Xie, Maoquan Wang, Jing Ma, Jian Jiang, and Zhao Lu. 2017. EICA Team at SemEval-2017 Task 3: Semantic and Metadata-based Features for Community Question Answering. In *Proceedings of the*

11th International Workshop on Semantic Evaluation (SemEval-2017), pages 292–298, Vancouver, Canada. Association for Computational Linguistics.

Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, and Ming Zhou. 2014. Question Retrieval with High Quality Answers in Community Question Answering. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 371–380. ACM.

Sheng Zhang, Jiajun Cheng, Hui Wang, Xin Zhang, Pei Li, and Zhaoyun Ding. 2017. FuRongWang at SemEval-2017 Task 3: Deep Neural Networks for Selecting Relevant Answers in Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 320–325, Vancouver, Canada. Association for Computational Linguistics.

Guangyou Zhou, Li Cai, Jun Zhao, and Kang Liu. 2011. Phrase-Based Translation Model for Question Retrieval in Community Question Answer Archives. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 653–662. Association for Computational Linguistics.