TueFact at SemEval 2019 Task 8: Fact checking in community question answering forums: context matters

Réka Juhász
University of Tübingen
Wilhelmstr. 19-23
72074 Tübingen, Germany
reka.juhasz@student.uni-tuebingen.de

Franziska Barbara Linnenschmidt
University of Tübingen
Wilhelmstr. 19-23
72074 Tübingen, Germany
franziska-barbara.linnenschmidt@student.uni-tuebingen.de

Teslin Roys
University of Tübingen
Wilhelmstr. 19-23
72074 Tübingen, Germany
teslin.roys@student.uni-tuebingen.de

Abstract

The SemEval 2019 Task 8 on Fact-Checking in community question answering forums aimed to classify questions into categories and verify the correctness of answers given on the QatarLiving public forum. The task was divided into two subtasks: the first classifying the question, the second the answers. The TueFact system described in this paper used different approaches for the two subtasks. Subtask A makes use of word vectors based on a bag-of-word-ngram model using up to trigrams. Predictions are done using multi-class logistic regression. The official SemEval result lists an accuracy of 0.60. Subtask B uses vectorized character n-grams up to trigrams instead. Predictions are done using a LSTM model and achieved an accuracy of 0.53 on the final SemEval Task 8 evaluation set. In a comparison of contextual inputs to subtask B, it was determined that more contextual data improved results, but only up to a point.

1 Introduction

The SemEval 2019 Task 8 on Fact-Checking gave us the opportunity to develop a system that evaluates the factual content of questions and answers in the field of community question answering forums. This popular niche on the internet provides helpful information for specific interests, such as information on life in Qatar or elsewhere in the world, coding help on Stack Overflow, or answers to a wide range of questions on Quora, Reddit/r/Ask or Yahoo! Answers. Often other resources are not at hand or misleading, and it proves difficult to find what one is looking for amid non-relevant questions, let alone be sure the answers found are correct. A system that can, with some degree of accuracy, pick out the factual questions and then predict the correctness of the given answers, is a means to ensure quality in community question answering forums. There may also be many further applications in information retrieval – ordering search results based on estimated factuality in a web query or even identifying truthful answers in automatic question answering systems.

The task was divided into two subtasks. While Subtask A required a classification of the questions into three distinct categories "Factual", "Socializing", and "Opinion", Subtask B took the subset of the "Factual" questions to classify them according to either "True", "False" or "NonFactual" in terms of the actual answer. Similar tasks were already part of SemEval 2015 (Nakov et al., 2015) and SemEval 2016 (Nakov et al., 2016).

The Tuefact system follows this division, even going as far as using different pre-processing to accommodate for the different needs. While the question classification is done with a bag of word approach using word trigrams, the answer classification uses character trigrams. The models used to make predictions are logistic regression for Subtask A and a long short-term memory model (LSTM) (Hochreiter and Schmidhuber, 1997) for Subtask B.

The following section describes the data provided for the tasks. The next two sections describe the two components of our language independent system in detail together with a brief discussion of failed approaches and changes that lead to improvements. In the last two sections we discuss further work to be done on the TueFact system and our conclusion about the current version of it.

2 The data

The pre-annotated data from the QatarLiving forum was provided in XML format. It was split into two files: one for the question classification, the other for the answer classification.

The data for Subtask A comprised a total of

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 1176–1179
Minneapolis, Minnesota, USA, June 6–7, 2019. ©2019 Association for Computational Linguistics

168 questions. Each question contained a subject line, i.e. a summary of the question, the detailed question, and all answers given to the question. The meta information contained amongst others the date, user name, and id. The questions were annotated into the three categories "Factual", "Opinion", and "Socializing". 33 of the questions were labeled as "Factual", 73 as "Opinion", and 62 as "Socializing". The longest question was 98 words long, the shortest only 5, the average length of questions was 41.1 words.

The data provided for Subtask B contained a total of 95 questions labeled "Factual". The meta information was the same. Of the 356 given answers 128 were labeled as "True", 102 as "False", and 126 as "NonFactual". The longest answer was 195 words long, the shortest consisted of only one word, the average length of answers was 31.5 words.

For further information about the data please refer to the task description paper from Mihaylova et. al (2019). No special data pre-processing steps were used in preparation for either subtask – no noticeable performance gains were observed when stripping accents, punctuation symbols, or lowercasing.

3 Question classification

The goal of this subtask was to classify the questions posted on the QatarLiving forum,[1] a community question answering forum, as either: 1. "Factual", meaning that it asks about something specific, and can receive a correct or incorrect answer, 2. "Opinion", in which the answers cannot be right or wrong, as it does not ask for objective facts but the personal input of the answering people, and 3. "Socializing", where the goal of posting the question was not seeking information at all, but rather looking for online communication.

[1] https://www.qatarliving.com/forum

3.1 Approach

We approached this task as a multi-class learning problem, instead of first dividing the questions into "Factual" and "NonFactual", and then further dividing the "NonFactual" questions into "Opinion" and "Socializing". As baseline model we decided to use multi-class logistic regression based on character bigrams, and only the subject line of the questions as input. Our reasoning for the use of this baseline model was to see how such a basic model would perform, and where we could take it from there. It achieved an accuracy of just over 50% on our development set, a randomly split quarter from the data set provided.

The current model uses a multi-class logistic regression model with a limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Saputro and Widyaningsih, 2017) solver to predict the labels. The labels were encoded using the scikit-learn library (Pedregosa et al., 2011) and the questions vectorized from word uni-, bi- and trigrams. We used raw counts for all word-grams, and did not add weighting. We further decided not to use a language dictionary as external feature, in order to maintain language independence.

3.2 Results

For means of evaluating our models, and as the data set was small, we have randomly split the training data and used one quarter of it as a development set. The development set was then not only used to evaluate our model in terms of accurcay, but also to compare the predicted to the actual labels. Our system was best at correctly predicting "Opinion" in an average of 10 cases compared to an average of four misclassified cases with no clear tendency of misclassifying it as either "Factual" or "Socializing". False predictions were equally often made in the classification of the labels "Socializing" and "Factual", which suggests further improvement possibilities.

Best training results on the development set were achieved at an accuracy of 0.714. The official SemEval result lists an accuracy of 0.60.

	No CC	Limited CC	Subject Only + IDs	Full CC
Accuracy	61.5	67.12	79.94	79.92

Table 1: Averaged 5-fold cross-validated accuracy on the development set for the character-based embeddings variation.

4 Answer classification

In this subtask, answers to the questions from the first task were to be classified into "Factual" and "Non-Factual", and "Factual" items further classified into "True" or "False". Unlike in the first task, there is no distinction made between types of non-factual comments (e.g. socializing or opinion).

4.1 Approach

Our approach to this task was to treat the problem as a multiple-class learning problem with three categories: "Non-Factual", "True" and "False". Variations of the model which split it into two sequential learning problems – fact or non-fact, then true or false – showed no significant difference.

The basic task B model consists of a LSTM network architecture (Hochreiter and Schmidhuber, 1997) with an embedding layer, two hidden layers of 100 nodes followed by a softmax activation layer to permit multiple classification. Embeddings were trained using the top 400 uni-, bi- and tri-gram features by frequency in the data. In training, we used categorical cross-entropy as a loss function. In all model variations, we trained for 200 epochs with a batch size of 128 and used L2 weight regularization with a factor of 0.012.

4.2 Variations and results

We examined three main variations of input to the model for answer classification: no comment chain (CC), limited comment chain, subject and comment identifier only and full comment-chain (see table 1). In no CC, we included only the comment itself being classified. In limited CC, we included the comment and all previous comments in the chain. In the subject plus comment identifier variation, we included the question subject heading, the comment text and commenter identifier. In the full CC variation we included the subject heading and the entirety of all comments in the chain. In tables 1 and 2, we abbreviate the same way.

Next, we experimented with word-level versus character-level embeddings (see table 2). In the end, unfortunately, our best results for task B on the evaluation set (also with the subject and identifiers only variation) were less encouraging at 53 percent.

5 Future work

Due to time constraints, several improvements to the models for both tasks weren't completed in time for the final evaluation. For the question classification task, we aimed to experiment with word- or character-embeddings instead of only a bag-of-n-grams approach in order to enable work on a multi-channel convolutional neural network model, which is also being pursued further. In some NLP tasks, CNNs have been shown to outperform not only traditional machine learning models but recurrent neural networks as well, and this may also be the case here (Wu et al., 2016).

For the answer classification task, work is underway on a model variation which tags comments based on the proportion of their content that can be found on multiple websites from a web query consisting of the question subject line.

6 Conclusions and analysis

For Subtask A, we considered two approaches: simple logistic regression, and a convolutional neural network approach. The initial success of the logistic regression approach on the development data suggests a 'simplicity first' strategy is sensible in this case, but its mediocre performance on the evaluation data indicates it is not especially robust.

In Subtask B, we examined only a single approach using a LSTM trained with embeddings, but with a number of variations. Variations including more of the comment chain as input were more successful than using single-comment answers as input, however we found little difference between including the entire comment chain and including only the subject question and comment identifiers. We suspect this is due to the fact that all answers include each other at least once in the full-chain variation, so it provides less to distinguish the answers from one another. Further, all vari-

	No CC/WB	Limited CC/WB	Full CC/WB
Accuracy	59.87	66.97	79.14

Table 2: Averaged 5-fold cross-validated accuracy on the development set for the word-based embeddings variation.

ations using character-based embeddings slightly outperformed model variations employing word-based embeddings (1 to 5 percent). This isn't entirely unexpected – solely character embedding based models performed very well in the SemEval 2018 Irony Detection task (Van Hee et al., 2018) as well, which bore similarities. One hypothesis for this result is that with small input sizes (such as tweets or forum posts) word-based embeddings may distinguish fewer distinctions, but it may also simply be that there is no significant performance difference. In which case, character- embeddings should be preferred as they do not require a maintenance of a large dictionary.

Overall, the TueFact system is an acceptable baseline and a solid starting point for further work in the direction of fact checking in community question answering forums. But perhaps of greatest interest are our comparative results under different input. The significance of input choice on performance is highlighted: our results show that while inclusion of more context can certainly be useful, the intuition that more data will always improve performance in this task should not be taken as a given.

References

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

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.

Preslav Nakov, Lluís Màrquez, Walid Magdy, Alessandro Moschitti, Jim Glass, and Bilal Randeree. 2015. Semeval-2015 task 3: Answer selection in community question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 269–281.

Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. Semeval-2016 task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 525–545.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Dewi Retno Sari Saputro and Purnami Widyaningsih. 2017. Limited memory broyden-fletcher-goldfarb-shanno (l-bfgs) method for the parameter estimation on geographically weighted ordinal logistic regression model (gwolr). In *AIP Conference Proceedings*, volume 1868, page 040009. AIP Publishing.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.