COMP90042 Project:
Automated Fact Checking For Climate Science Claims

1068324

1 Introduction

As stated by Zeng et al., automated fact checking comprises of document retrieval, rationale selection and claim verification. For the system built in this project, based on the experimental analysis illustrated in the report, since document retrieval is out of scope, it firstly achieves rationale selection which aims to choose the relevant evidence at sentence-level by exploiting the last hidden layer as sentence embedding of the fine-tuned RoBERTa model for the comparison of the similarity between claims and evidences. After that, the task of claim verification is completed by the fine-tuned RoBERTa specified for the verdict classification of claims in the training set. However, the result of the system is less than expected and the potential reasons will be discussed in the Critical Analysis section.

2 Exploratory Analysis

By observing the train data, there are significant problems between the claims and their corresponding evidences, which can be mainly but not fully concluded in the following points:

1. Chemical Formula <-> Full Name (CO2 <-> Carbon Dioxide)
2. Synonyms (Man-made <-> human activity <-> human-caused)
3. Abbreviations (i.e. "PDO" exposed in the claim but not in the evidence)
4. Same information by different expressions (Global Warming <-> Catastrophic heating of the Earth's atmosphere)
5. 5. Wrong Format (CO 2 <-> CO2)

The above problems potentially cause severe influence while applying any count-based methods since the tokens from claims and evidences after being preprocessed are likely to be various and thus the sentence similarity may be lower and not suitable to be used for the evidence retrieval, which will be further experimented in the later sections.

3 Experiments for Sentence-similarity

Zeng et al. concludes that keywords matching and sentence similarity scoring are popular methods to select rationale. In this section, Distributional semantics is used to compare the sentence similarity.

3.1 Distributional semantics

3.1.1 Count based methods

The count based method chosen is BM25 which can be used for search engine. As illustrated by Trotman et al., the algorithms of BM25 utilize some kinds of sparse vector similarity between the weighted BOW of two sentences based on the concept of tf-idf. However, its shortcoming appears to be that matching exact words play a significant role in the retrieval of context. The results returend from BM25 with higher scores will contain more of the words in the claim sentence and less words that do not apear in it, which fails to address the problems of the discrepancy between the claim and its evidences mentioned in the former section. Therefore, it is trivial to only use BM25 for the retrieval of evidences. Nevertheless, BM25 can still be useful to handle the amount of possible evidences in the Evidence Retrieval section.

3.1.2 Neural methods

Word2Vec is applied to firstly compute the word embedding of each token, after which it averages the token embeddings to extract the sentence embeddings and compares them by cosine similarity. To train the CBOW model , the claims in the training set and the source of evidences which are preprocessed by lowering cases and removing stop words and punctuations are used as the corpus. Nevertheless, due to the existing problems

mentioned before, the model does not perform as expected. For instance, the similarity between the terms "co2" and "carbon dioxide" is around 0.14, which is considered to be low for sentence similarity while comparing each claim with almost 1.21 million evidences. Hence, Word2Vec is not desirable as there exists many issues in the training corpus as stated before.

4 Evidence Retrieval by Contextual Representation

At this section, to address the task of retrieving the evidences, as inspired by the results from Soleimani et al., the last hidden layer of the pre-trained BERT model is used for obtaining the sentence embeddings by averaging the word embeddings of tokens in the sentence. The choice of the pre-trained model is RoBERTa and the reason for the choice will be discussed in the next section as those two tasks can be processed separately and simultaneously. The first claim in the training set is taken for illustration. By empirical results, with utilizing the fine-tuned RoBERTa model for the sentence similarity, the disparity of similarity between the corresponding evidences and the same set of 20 randomly selected evidences has been increased even though the overall similarity has dropped to some extent, which is same as the expectation since it can better distinguish the true evidences for the claim. The fine-tuning process for the RoBERTa is divided as follows: firstly, the positive samples are generated by combining the claim with each of its corresponding evidences in the manner of inputs to the model. Secondly, the negative samples are produced by using the BM25 to search same amount of evidences as positive samples for the same claim from the source of evidences for those with the highest scores but not related to the claim which are considered to be hard negative from the outcomes of the experiments to implement DPR by Karpukhin et al., after which the negative evidences are combined with the claim as same as the first step. In such step, the identical quantity of positive and negative samples is designed to avoid imbalanced classes problem. At the same time, the development data set is generated in the same format except for the negative sampling, instead of BM25, the corresponding evidences for other claims are chosen as negative samples, which also yields good performance by Karpukhin et al.. Lastly, by adding a classification layer on top of the contextual representations of the RoBERTa model to classify whether the claim and the evidence in the combined sentence are corresponding, the fine-tuned RoBERTa model is fed with the new generated training data set to train and validated on the development data set with same format, where the loss function is Binary Cross Entropy with logistic loss, the optimizer is Adam and the maximum length of inputs is specified to be 100.

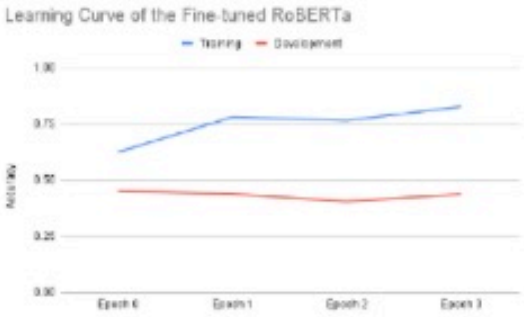

Figure 1: Learning Curve of Fine-tuned RoBERTa

The result of training is shown in the Figure 1. From the training and development accuracy, it is obvious that the model tends to overfit the training data, which can be potentially caused by lack of amount or generalization of data and the complexity of the model. This issue will be further analyzed in the Critical Analysis section. As the fine-tuned model is prepared well and the previous result does indicate that it can better deal with the similarity between claims and evidences, the next step is to decide the proper number for each claim in the test set.

Number of Evidences	1	2	3	4	5
Number of Claims	210	223	191	127	477

Table 1: The Distribution of Number of Evidences

As shown in Table 1, for the training data set, the majority of claims has 5 evidences. Since the number of evidences for a specific claim depends on its context and the source of evidences, it is nearly impossible to specify an exact number for each test claim due to the variation of sentence embeddings and the tremendous amount of evidences from the source. Therefore, from the perspective of frequentist hypothesis and the fact proposed by Zeng et al. that for rationale selection, it manually chooses the value k for the selection of top k sentences for retrieval, it is reasonable to assume that each test claim has 5 evidences. The next part is that for each test claim, compute the sentence similarity

between it and each evidence in the source and find the top 5 evidences. However, as proposed by Reimers and Gurevych, on a modern V100 GPU, 49995000 inference computations of BERT takes about 65 hours. Empirically, since it only needs the last hidden layer of the fine-tuned RoBERTa model, it roughly requires 1.2 million inference computations including the source of evidences and the test claims. Nevertheless, after trying to compute the sentence embeddings for evidences in the source with no trick on parallel processing, it did not finish the computation after 6 hours and thus the overall comparison of all evidences in the source is discarded as even though the sentence embeddings can be computed beforehand, it still takes a long period to compute the similarity and sort the results. To address the issue to avoid exhaustive comparison, intuitively, BM25 can be used for initial filtering of evidences from the source since that from the perspective of a human, when trying to find the evidences for the claim, the context with the tokens appearing in the claim will be more likely related to it. Thus, BM25 is used to firstly select top N evidences with highest scores based on tf-idf where N is the hyper parameter and then select the top 5 evidences with the highest cosine similarities with the claim according to the sentence embeddings extracted from the fine-tuned RoBERTa model.

Top N Evidences	N=20	N=40	N=60	N=100
Evidences Retrieval F-score	0.11235	0.11241	0.11737	0.10316

Table 2: The F-scores for different N

From the Table 2 of tuning the hyperparameter N on the development data set, it is possible that with larger N, the evidences of the claim will be included in the top N evidences. However, since it also depends on the sentence embeddings, more experiments of larger N should be tested. From the consideration of limited resources and time for the project, it could be improved in the future. And the reason for such low F-scores will be discussed in the Critical Analysis section.

5 Claim Classification

To classify the verdict for the claims, it exploits the same idea in the Evidence Retrieval section for fine-tuning the pre-trained model. Firstly, the choice of pre-trained model should be addressed. As suggested by Reimers and Gurevych, BERT and RoBERTa perform well for sentence embedding methods. Apart from those two models, DistilBERT is also included for comparison since it has fewer parameters and thus less time for inference computation. Before training, the training and development data set are preprocessed by combining the claim and each of its evidences in the form of inputs to the pre-trained models and labelled by their verdicts. Similarly, a classification layer is added on top of the last hidden layer, where the loss function is Cross Entropy Loss, the optimizer is Adam and the maximum length of inputs is 100.

Figure 2: Learning Curve of claim Classification

The result from Graph 2 of training the model on training set and validating it on development set shows that those models have similar limitations as the fine-tuned RoBERTa from the fact that with such high accuracy on train data and relatively low accuracy on development data, they overfit on the train data due to the same reasons mentioned before. From the result, it is observed that the RoBERTa model performs better on new unseen data and thus can be chosen for being used in Evidence Retrieval and Claim Classification. The next step is to predict the label of the top 5 evidences. As the fine-tuned RoBERTa model for claim classification can only achieve around 50% accuracy on development set, it is very likely that the top 5 evidences belong to different labels. To deal with this issue, the majority vote is applied and the tie is broken by choosing the label which has higher sum of the similarities with the claim for those evidences belonging to it.

Top N Evidences	N=20	N=40	N=60	N=100
Classification Accuracy	0.5649	0.5714	0.5649	0.5714

Table 3: The Classification Accuracy

As shown in the Table 3, the performance of the fine-tuned RoBERTa model on the retrieved evidences set from development claims is in accordance with its performance on the real evidences set under the variation of the hyperparameter N.

Thus, the underlying problems lie in how to improve the performance of the fine-tuned RoBERTa model, which will be explored in the next section.

6 Critical Analysis

6.1

The first one is about the performance of the fine-tuned RoBERTa models in Evidence Retrieval and Claim Classification. Both models have shown that they achieve high accuracy on training data set and relatively low accuracy on development data set, which means they overfit on the training data set. Two factors potentially contribute to this result: the data fed to the models and the complexity of the models. Firstly, as for the data, it should fulfill two requirements : the data is general and the amount should be large. Originally, there are around 1200 claims in the train set. Although the number of samples increases to approximately 8000 after being combined by the above methods, it is still possible to be regarded as not enough amount of training data as those claims have repeatedly occurred in the start of some samples and thus the models tend to recognize those patterns more frequently. Thus, by introducing more general training data, overfitting should be alleviated. Secondly, as for the complexity of the models, both models has only applied a single linear classification layer on their top and they have already overfitted, which means that the data amount is not big enough for such quantity of parameters in the models. Thus, it can be supposed that the model with higher complexity in the form of more layers on the top with regularization is able to better perform on both tasks with the reinforcement of larger number of general data set, before which it should firstly involve more data to train current models to observe any improvement of the performance.

6.2

The second one is about low F-scores in Evidence Retrieval. As stated in the outcomes of Evidence Retrieval, by the definition of the F-scores, the system does not work well in the retrieval component. Internally, the choice of the number of evidences for each claim, the choice of Top N evidences selected by BM25 and the performance of the fine-tuned RoBERTa to compare sentence similarity affect the result of F-scores. As for the choice of the number of evidences for each claim, if it is tuned as a hyperparameter, the high performance on the development data set does not guarantee the performance on the test data set since this number is very likely different for each claim. To avoid overfit on a specific data set, this number remains the same. However, the choice of N can be further increased to 1000 or more given more time and computing resources, as compared to fully search in 1.2 million evidences, the number and the time to generate such a list are still fairly small. It is supported by the fact that the tokens in the evidences do not necessary to be identical with the related tokens in the claims as discussed in the Exploratory Analysis section. However, the tuning of N also encounters with the same issue of k, that is, the optimal N for development set yields a lower Harmonic Mean of F and A in the "Final Evaluation" performance on the Codalab leader-board, which means N is overfitted on the development set. As for the fine-tuned RoBERTa model, with its better performance mentioned above, it is likely able to distinguish the similarity better. Externally, it should carefully consider what the meaning of evidences to a claim carefully. With the problems mentioned in the Exploratory Analysis section, some tokens in the claims may be explained with more tokens in the evidences. And also the data cleansing is important considering that the size of training set is small. Hence, by those factors, the trick of using BM25 which is count-based to reduce the number of candidates will be degraded.

7 Conclusion

The report firstly explores the problems of the discrepancy between claims and their evidences in the training data set and then addresses the unavailability of implementing distributional semantics methods for sentence similarity. The final automated fact checking system consists of evidence retrieval and claim classification based on the contextual representation in the form of applying the pre-trained RoBERTa models fine-tuned for specific tasks accordingly. Lastly, it discusses the performance of both tasks and proposes that the improvements should be based on introducing more training data and refining the internal mechanism to select the potential evidences to prevent exhaustive search for the source of evidences.

References

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense