Understanding the Impact of Evidence-Aware Sentence Selection for Fact Checking

Giannis Bekoulis[1,2] **Christina Papagiannopoulou**[3] **Nikos Deligiannis**[1,2]
[1]ETRO, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[2]imec, Kapeldreef 75, 3001 Leuven, Belgium
{gbekouli,ndeligia}@etrovub.be
[3]cppapagi@gmail.com

Abstract

Fact Extraction and VERification (FEVER) is a recently introduced task that consists of the following subtasks (i) document retrieval, (ii) sentence retrieval, and (iii) claim verification. In this work, we focus on the subtask of sentence retrieval. Specifically, we propose an evidence-aware transformer-based model that outperforms all other models in terms of FEVER score by using a subset of training instances. In addition, we conduct a large experimental study to get a better understanding of the problem, while we summarize our findings by presenting future research challenges[1]

1 Introduction

Recently a lot of research in the NLP community has been focused on the problem of automated fact checking (Liu et al., 2020; Zhong et al., 2020). In this work, we focus on the FEVER dataset that is the largest fact checking dataset (Thorne et al., 2018). The goal of the task is to identify the veracity of a given claim based on Wikipedia documents. The problem is traditionally approached as a series of three subtasks, namely (i) document retrieval (select the most relevant documents to the claim), (ii) sentence retrieval (select the most relevant sentences to the claim from the retrieved documents), and (iii) claim verification (validate the veracity of the claim based on the relevant sentences).

Several models have been proposed for the FEVER dataset (Hanselowski et al., 2018; Nie et al., 2019a; Soleimani et al., 2020). Most of the existing literature (Liu et al., 2020; Zhong et al., 2020) focuses on the task of claim verification, while little work has been done on the tasks of document retrieval and sentence retrieval. We suspect that this is because it is more straightforward for researchers to focus only on the improvement in terms of performance of the last component (i.e., claim verification) instead of experimenting with the whole pipeline of the three subtasks. In addition, the performance in the first two components is already quite high (i.e., $>$90% in terms of document accuracy for the document retrieval step and $>$87% in terms of sentence recall).

Unlike the aforementioned studies, in this work, we focus on the task of sentence retrieval on the FEVER dataset. Specifically, inspired by studies that investigate the impact of loss functions and sampling on other domains (e.g., computer vision (Wu et al., 2017; Wang et al., 2017), information retrieval (Pobrotyn et al., 2020)), this paper – to the best of our knowledge – is the first attempt to shed some light on the sentence retrieval task by performing the largest experimental study to date and investigating the performance of a model that is able to take into account the relations between all potential evidences in a given list of evidences. The contributions of our work are as follows: (i) we propose a simple yet effective evidence-aware transformer-based model that is able to outperform all other models in terms of the FEVER score (i.e., metric of the claim verification subtask) and improve a baseline model by 0.7% even by using a small subset of training instances; (ii) we conduct an extensive experimental study on various settings (i.e., loss functions, sampling instances) showcasing the effect in performance of each architectural choice on the sentence retrieval and the claim verification subtasks; (iii) the results of our study point researchers to certain directions in order to improve the overall performance of the task.

2 Models

We frame the sentence selection subtask, where the input is a claim sentence and a list of candidate evidence sentences (i.e., as retrieved from the document retrieval step, for that we used the same input as in the work of Liu et al. (2020)), as an NLI problem. Specifically, the claim is the "hypothesis"

[1]https://github.com/bekou/evidence_aware_nlp4if

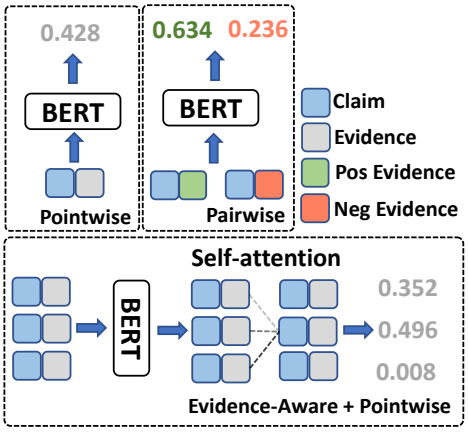

Figure 1: The architectures used for the sentence retrieval subtask. The pointwise loss considers each potential evidence independently. The pairwise loss considers the potential evidences in pairs (positive, negative). The proposed evidence-aware selection model uses self-attention to consider all the potential evidences in the evidence set simultaneously.

sentence and the potential evidence sentence is a "premise" sentence. In Fig. 1, we present the various architectures that we used in our experiments.

2.1 Baseline

Pointwise: Our model is similar to the one described in the work of Soleimani et al. (2020). We use a BERT-based model (Devlin et al., 2019) to obtain the representation of the input sentences. For training, we use the cross-entropy loss and the input to our model is the claim along with an evidence sentence. The goal of the sentence retrieval component paired with the pointwise loss is to predict whether a candidate evidence sentence is an evidence or not for a given claim. Thus, the problem of sentence retrieval is framed as a binary classification task.

2.2 Distance-based

Pairwise: In our work, we also exploit the pairwise loss, where the goal is to maximize the margin between the positive and the negative examples. Specifically, we use the pairwise loss that is similar to the margin based loss presented in the work of Wu et al. (2017). The pairwise loss is:

$$\mathcal{L}^{pairwise}(p,n) = [-y_{ij}(f(x_p) - f(x_n)) + m]_+ \quad (1)$$

In Eq. (1), $y_{ij} \in \{-1, 1\}$, $f(x)$ is the representation that we obtain from the BERT-based model, m is the margin and the indices p and n indicate a pair of a positive and a negative example. In order to obtain a claim aware representation of the (positive-negative) instances, we concatenate the claim with the corresponding evidence.

Triplet: Unlike the pairwise loss that considers only pairs of positive and negative examples, the triplet loss (Wu et al., 2017) uses triplets of training instances. Specifically, given an anchor sample a (i.e., claim), the goal is the distance $D_{ij} = \|f(x_i) - f(x_j)\|_2$ to be greater between the anchor and a negative example than the distance between the anchor and a positive example. The triplet loss is depicted in:

$$\mathcal{L}^{triplet}(a,p,n) = [D_{ap}^2 - D_{an}^2 + m]_+ \quad (2)$$

Similar to the previous equation, in Eq. (2), m is the margin and the indices a, p and n indicate the triplet of the anchor, a positive and a negative example. As anchor we use the claim, while similar to the pairwise loss, we concatenate the claim with the corresponding evidence for the positive and the negative examples.

Cosine: We have also experimented with the cosine loss. Specifically, we exploit positive and negative samples using the following formula:

$$\mathcal{L}^{cos}(p,n) = y_{ij}(1 - \cos(f(x_p), f(x_n))) + (1 - y_{ij})[(\cos(f(x_p), f(x_n)) - m)]_+ \quad (3)$$

In Eq. (3), $y_{ij} \in \{0, 1\}$ and cos indicates the cosine distance between the positive and the negative samples.

Angular: The angular loss (Wang et al., 2017) uses triplets of instances (i.e., similar to the triplet loss) while imposing angular constraints between the examples of the triplet. The formula is given by:

$$\mathcal{L}^{ang}(a,p,n) = [D_{ap}^2 - 4\tan^2 r D_{nc}^2]_+ \quad (4)$$

In Eq. (4), $f(x_c) = (f(x_a) - f(x_p))/2$ and r is a fixed margin (angle).

2.3 Evidence-Aware Selection

Unlike the aforementioned loss functions, the proposed model relies on a transformer-based model, similar to the retrieval model proposed in the work of Pobrotyn et al. (2020). This model exploits the use of self-attention over the potential evidence sentences in the evidence set. Unlike (i) the pointwise

Loss	# Negative Examples	# Max Instances	Dev					Test				
			P@5	R@5	F_1@5	LA	FEVER	P@5	R@5	F_1@5	LA	FEVER
Angular	✓	✓	26.90	93.93	41.82	77.22	74.81	24.36	86.14	37.98	72.30	68.30
Cosine	✓	✓	27.02	93.85	41.96	77.50	75.10	**24.83**	86.73	**38.61**	72.49	68.81
Triplet	✓	✓	26.99	94.24	41.96	77.51	75.32	24.74	**86.86**	38.51	72.76	69.31
Pairwise	✓	✓	26.88	93.90	41.79	78.05	75.61	24.44	86.17	38.08	72.92	69.34
	5	✓	26.76	93.23	41.58	77.21	74.74	24.53	85.90	38.17	72.05	68.22
	10	✓	26.77	92.99	41.57	77.58	75.04	24.62	86.15	38.29	72.65	68.93
	5	20	27.11	94.13	42.10	77.53	75.37	24.75	86.67	38.51	72.87	69.25
	10	20	27.09	94.40	42.10	78.05	75.79	24.74	86.84	38.51	**73.02**	69.38
Pointwise	✓	✓	25.77	91.96	40.26	77.94	75.12	22.28	82.61	35.01	71.63	67.63
	5	✓	27.74	95.93	43.04	78.43	76.71	23.99	85.67	37.48	72.54	68.71
	5	20	27.39	95.25	42.54	78.49	76.58	23.79	85.24	37.19	72.55	68.64
Evidence-Aware	5	20	**28.52**	**97.16**	**44.09**	**78.67**	**77.38**	24.70	86.81	38.46	72.93	**69.40**
	10	20	28.50	96.82	44.04	78.26	76.78	24.76	86.83	38.53	72.70	68.46

Table 1: Results of the (i) sentence retrieval task in terms of Precision (P), Recall (R), and F_1 scores and (ii) claim verification task in terms of the label accuracy (LA) and the FEVER score evaluation metrics in the dev and the test sets. The best performing models per column are highlighted in bold font. For more details, see Section 3.3.

loss that does not take into account the relations between the evidence sentences, and (ii) the distance-based losses (e.g., triplet) that considers only pairs of sentences, the transformer model considers subsets of evidence sentences simultaneously at the training phase. Specifically, the input to the transformer is a list of BERT-based representations of the evidence sentences. Despite its simplicity, the model is able to reason and rank the evidence sentences by taking into account all the other evidence sentences in the list. On top of the transformer, we exploit a binary cross-entropy loss similar to the one presented in the case of the pointwise loss.

3 Experimental Study

3.1 Setup

For the conducted experiments in the sentence retrieval task, in all the loss functions except for the evidence-aware one, we present results using all the potential evidence sentences (retrieved from document retrieval). For the evidence-aware model, we conduct experiments using either 5 or 10 negative examples per positive instance during training. In addition, the overall (positive and negative) maximum number of instances that are kept is 20. This is because unlike the other models that the evidences are considered individually or in pairs, in the evidence-aware model, we cannot consider all the evidences simultaneously. We experiment also with a limited number of instances in the other settings to have a fair comparison among the different setups. Note that for the distance-based losses, we conduct additional experiments only in the best performing model when all instances are included (i.e., pairwise). We also present results on the claim verification task with all of the examined architectures. For the claim verification step, we use the model of Liu et al. (2020). We evaluate the performance of our models using the official evaluation metrics for sentence retrieval (precision, recall and F_1 using the 5 highly ranked evidence sentences) and claim verification (label accuracy and FEVER score) in the dev and test sets.

3.2 Evaluation Metrics

We use the official evaluation metrics of the FEVER task for the sentence retrieval and the claim verification subtasks.

Sentence Retrieval: The organizers of the shared task suggested the precision to count the number of the correct evidences retrieved by the sentence retrieval component with respect to the number of the predicted evidences. The recall has also been exploited. Note that a claim is considered correct in the case that at least a complete evidence group is identified. Finally, the F_1 score is calculated based on the aforementioned metrics.

Claim Verification: The evaluation of the claim verification subtask is based on the *label accuracy* and the *FEVER score* metrics. The label accuracy measures the accuracy of the label predictions without taking the retrieved evidences into account. On the other hand, the FEVER score counts a claim as correct if a complete evidence group has been correctly identified as well as the corresponding label. Thus, the FEVER score is considered as a

strict evaluation metric and it was the primary metric for ranking the systems on the leaderboard of the shared task.

3.3 Results

In Table 1, we present our results on the sentence retrieval and claim verification tasks. The "# Negative Examples" column indicates the number of negative evidences that are randomly sampled for each positive instance during training, while the "# Max Instances" column indicates the maximum number of instances that we keep for each claim. The ✓ symbol denotes that we keep all the instances from this category (i.e., "# Negative Examples" or "# Max Instances"). Note that for the number of maximum instances, we keep as many as possible from the positive samples, and then we randomly sample from the negative instances.

Benefit of Evidence-Aware Model: The evidence-aware model (see the setting with 5 negative examples and 20 maximum instances denoted as (5, 20)) is the best performing one both in dev and test set in terms of FEVER score. The pairwise loss performs best in terms of label accuracy on the test set. However, the most important evaluation metric is the FEVER score, since it takes into account both the label accuracy and the predicted evidence sentences. The pointwise loss is the worst performing one when using all the evidence sentences. This is because in the case that we use all the potential evidences, the number of negative samples is too large and we have a highly imbalance problem leading to low recall and FEVER score in both the dev and test set. Note that the evidence-aware model relies on the pointwise loss (i.e., the worst performing one). However, a benefit of the evidence-aware model (0.7% in terms of FEVER score) is reported (see pointwise (5, 20)). This showcases the important effect of ranking potential evidences simultaneously using self-attention. From the distance-based loss functions (e.g., triplet) except for the pairwise, we observe that the angular and the cosine loss have worst performance compared to the pairwise and the triplet loss when using all the instances. We hypothesize that this is because the norm-based distance measures fit best for scoring pairs using the BERT-based representations.

Performance Gain: Most recent research works (e.g., Zhao et al. (2020); Liu et al. (2020)) focus on creating complex models for claim verification. We conducted a small scale experiment (that is not present in Table 1), where we replaced our model for claim verification (recall that we rely on the method of Liu et al. (2020)) with a BERT-based classifier. We observed that when using the model of the Liu et al. (2020) instead of the BERT-classifier (in our early experiments on the dev set), the benefit for the pointwise loss was 0.2 percentage points, a benefit of 0.1 percentage points for the triplet loss and a drop of 1 percentage point in the performance of the cosine loss. Therefore, the seemingly small performance increase in our model (i.e., a benefit of 0.7% in terms of FEVER score) is in line with the performance benefit of complex architectures for the claim verification task. In our paper, we do not claim state-of-the-art performance on the task, but rather showcase the benefit of our proposed methodology over a strong baseline model that relies on BERT_{base}.

Number of Samples Matters: The evidence-aware model is the best performing one (5, 20), while using only a small fraction of the overall training instances. This is because the evidence-aware model is able to take into account all possible combinations of the sampled evidences while computing attention weights. However, the same model in the (10, 20) setting showcases a reduced performance. This is due to the fact that the pointwise loss affects the model in a similar way as in the pointwise setting leading to a lower performance (due to class imbalance). For the pairwise loss, we observe that the performance of the model when sampling constrained evidence sentences (see (5, 20), (10, 20) settings) is similar to the performance of the model when we do not sample evidence sentences. In addition, it seems that when one constrains the number of negative samples should also constrain the overall number of instances in order to achieve the same performance as in the non-sampling setting. We hypothesize that this is due to that fact that when we have a limited number of instances it is better to have a more balanced version of the dataset.

Outcome: Therefore, we conclude that the evidence-aware model achieves high performance by using few examples, and thus it can be used even

in the case that we have a small amount of training instances. In the case of the pairwise loss is important to sample instances, otherwise it becomes computationally intensive when we take all the possible combinations between the positive and negative training instances into account. In addition, it is crucial to sample negative sentences to control: (i) the computational complexity in the case of the distance-based loss functions, (ii) the memory constraints in the case of the evidence-aware model and (iii) the imbalance issue in the case of the pointwise loss. However, more sophisticated techniques than random sampling should be investigated to select examples that are more informative. Finally, as indicated by our performance gain, we motivate future researchers to work also on the sentence retrieval subtask, as the improvement in this subtask leads to similar improvements with architectures proposed for the claim verification subtask.

4 Related Work

An extensive review on the task of fact extraction and verification can be found in Bekoulis et al. (2020). For the sentence retrieval task, several pipeline methods (Chernyavskiy and Ilvovsky, 2019; Portelli et al., 2020) rely on the sentence retrieval component of Thorne et al. (2018) that use TF-IDF representations. An important line of research (Hanselowski et al., 2018; Nie et al., 2019a; Zhou et al., 2019) includes the use of ESIM-based models (Chen et al. (2017). Those works formulate the sentence selection subtask as an NLI problem where the claim is the "premise" sentence and the potential evidence sentence is a "hypothesis" sentence. Similar to the ESIM-based methods, language model based methods (Nie et al., 2019b; Zhong et al., 2020; Soleimani et al., 2020; Liu et al., 2020; Zhao et al., 2020) transform the sentence retrieval task to an NLI problem using pre-trained language models. For the language model based sentence retrieval two types of losses have been exploited (i) pointwise loss, and (ii) pairwise loss, as presented also in Section 2. Unlike the aforementioned studies that rely only on losses of type (i) and (ii), we conduct the largest experimental study to date by using various functions on the sentence retrieval subtask of the FEVER task. In addition, we propose a new evidence-aware model that is able to outperform all other methods using a limited number of training instances.

5 Conclusion

In this paper, we focus on the subtask of sentence retrieval of the FEVER task. In particular, we propose a simple and effective evidence-aware model that outperforms all other models in which each potential evidence takes into account information about other potential evidences. The model uses only a few training instances and improves a simple pointwise loss by 0.7% percentage points in terms of FEVER score. In addition, we conduct a large experimental study, compare the pros and cons of the studied architectures and discuss the results in a comprehensive way, while pointing researchers to future research directions.

References

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2020. Fact extraction and verification–the fever case: An overview. *arXiv preprint arXiv:2010.03001*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Anton Chernyavskiy and Dmitry Ilvovsky. 2019. Extract and aggregate: A novel domain-independent approach to factual data verification. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 69–78, Hong Kong, China. Association for Computational Linguistics.

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.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. Association for Computational Linguistics.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for*

Computational Linguistics, pages 7342–7351, Online. Association for Computational Linguistics.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019a. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866, Honolulu, Hawai. AAAI Press.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019b. Revealing the importance of semantic retrieval for machine reading at scale. 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 2553–2566, Hong Kong, China. Association for Computational Linguistics.

Przemysław Pobrotyn, Tomasz Bartczak, Mikołaj Synowiec, Radosław Białobrzeski, and Jarosław Bojar. 2020. Context-aware learning to rank with self-attention. In *Proceedings of the ECOM'20: The SIGIR 2020 Workshop on eCommerce*, Online. Association for Computing Machinery.

Beatrice Portelli, Jason Zhao, Tal Schuster, Giuseppe Serra, and Enrico Santus. 2020. Distilling the evidence to augment fact verification models. In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 47–51, Online. Association for Computational Linguistics.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. 2017. Deep metric learning with angular loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2593–2601, Venice, Italy. IEEE.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. 2017. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2840–2848, Venice, Italy. IEEE.

Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Transformer-xh: Multi-evidence reasoning with extra hop attention. In *International Conference on Learning Representations*, Online.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.