Improving Evidence Retrieval with Claim-Evidence Entailment

Fan Yang*
Amazon Alexa
Cambridge, MA, USA
fyaamz@amazon.com

Eduard Dragut
Temple University
Philadelphia, PA, USA
edragut@temple.edu

Arjun Mukherjee
University of Houston
Houston, TX, USA
arjun4787@gmail.com

Abstract

Claim verification is challenging because it requires first to find textual evidence and then apply claim-evidence entailment to verify a claim. Previous works evaluate the entailment step based on the retrieved evidence, whereas we hypothesize that the entailment prediction can provide useful signals for evidence retrieval, in the sense that if a sentence supports or refutes a claim, the sentence must be relevant. We propose a novel model that uses the entailment score to express the relevancy. Our experiments verify that leveraging entailment prediction improves ranking multiple pieces of evidence.

1 Introduction

Claim verification verifies the credibility of a textual claim by inferring relevant and reliable textual evidence. An example in this space is FEVER, which regards Wiki pages as potential evidence and creates claims by crowdsourcing (Thorne et al., 2018). They propose a three-step pipeline: (i) document-level evidence retrieval; (ii) sentence-level evidence retrieval; (iii) claim-evidence entailment. Some works follow the pipeline and propose new models to improve claim verification (Yoneda et al., 2018; Nie et al., 2019; Hanselowski et al., 2018; Zhou et al., 2019), while other works combine the second and the third step and leverage all possible sentences for claim verification (Yin and Roth, 2018; Ma et al., 2019). We refer to the former as the pipeline framework and the latter as the multi-task framework.

The pipeline framework restricts a few sentences, and therefore it may not cover relevant evidence. The multi-task framework includes all possible sentences, where irrelevant ones may bring overwhelming noise and hurt claim verification. We argue that previous works focus on improving individual components but neglect to examine how those components connect. For example, will the entailment improve the retrieval?

Claim: Andy Roddick lost 5 Master Series between 2002 and 2010.
Candidate: Roddick was ranked in the top 10 for nine consecutive years between 2002 and 2010, and won five Masters Series in that period. **Label**: REFUTE
Candidate: Roddick's hard-court record in 2003 included his first Masters Series titles coming at Canada and Cincinnati and his only Grand Slam title. **Label**: NOT ENOUGH INFO
Candidate: Federer won his first Master Series event at the 2002 Hamburg Masters on clay, over Marat Safin. **Label**: NOT ENOUGH INFO

Table 1: The entailment result can imply whether the candidate is relevant or not.

We hypothesize that claim-evidence entailment can provide useful signals for evidence retrieval: if a sentence supports or refutes a claim, the sentence must be relevant. As in Table 1, the first candidate (actual evidence) shares more words and longer phrases than the other candidates. In contrast, the other two candidates may be relevant to the claim to some extent: the second sentence mentions *Roddick* and *masters series*, and the third sentence mentions *masters series* and *2002*. Thus, we propose a novel method to link the entailment prediction to the relevance score. Our method predicts the entailment for all retrieved candidates and utilizes the entailment score to express the relevance.

To our knowledge, this is the first work that

*The paper was done before the author joined Amazon.com Inc.

Proceedings of Recent Advances in Natural Language Processing, pages 1553–1558
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_174

uses the entailment prediction to measure relevancy. Our experiment demonstrates that a reliable entailment prediction improves evidence retrieval. This is beyond previous works that merely share low-level text encoder and train the two steps together.

2 Method

We adopt two base models, the Decomposable Attention (DA) model (Parikh et al., 2016) and the Enhanced Sequential Inference (ESI) model (Chen et al., 2017a), to encode claim-evidence pairs. Both models are designed for textual entailment (Dagan et al., 2005; Bowman et al., 2015; Parikh et al., 2016; Williams et al., 2018). Although there are more methods in the area of textual entailment (Sha et al., 2016; Chen et al., 2017b; Conneau et al., 2017; Nie et al., 2019; Munkhdalai and Yu, 2017; Tay et al., 2018; Ghaeini et al., 2018), we prefer the DA model and the ESI model because they have been widely applied for sentence retrieval and claim-evidence entailment (Thorne et al., 2018; Hanselowski et al., 2018; Nie et al., 2019; Yoneda et al., 2018).

2.1 Claim Verification Pipeline

We follow the three-step pipeline as proposed in (Thorne et al., 2018). We first apply the strategy proposed in (Hanselowski et al., 2018) to retrieve documents. It employs the constituency parser from AllenNLP (Gardner et al., 2018) to find entities. It uses MediaWiki API to obtain relevant articles by matching the title of the article with claim entities. Once we collect K^D document candidates, we treat each sentence of the article as potential evidence. **Evidence retrieval** considers the claim and all candidate sentences as the input and outputs evidence by selecting a subset of sentences. We use $\mathbf{h}_i^R = \text{Enc}^R(\mathbf{w}^c, \mathbf{w}_i^s)$ to denote the process that the relevance encoder encodes the claim $\mathbf{w}^c$ and the i-th evidence candidate $\mathbf{w}_i^s$ into the representation $\mathbf{h}_i^c$. We obtain the relevance score s_i by giving $\mathbf{h}_i$ to a fully connected network (FCN). After sorting the relevance score of evidence candidates, we collect the top K^S candidates as the evidence. **Claim-evidence entailment** predicts three probable outcomes: (i) the evidence supports the claim; (ii) the evidence refutes the claim; (iii) the evidence needs more information. The entailment encoder encodes the claim $\mathbf{w}^c$ and the retrieved evidence $\mathbf{w}^r$, and we denote the process as $\mathbf{h}^E = \text{Enc}^E(\mathbf{w}^c, \mathbf{w}^r)$. Then we feed $\mathbf{h}^E$ to another FCN for the entailment probability.

2.2 Sentence Pair Encoder

We design the relevance encoder and the entailment encoder to share the same architecture, because they both take two sentences as input and produce vector representation that captures claim, evidence, and the correlation of them. Although we consider the DA model and the ESI model in this work, we do not limit the choice of architectures. Let $\mathbf{a}$ and $\mathbf{b}$ be two sentences. The core idea of the two models is to obtain the attention weights $e_{j,k}$ of word a_j and word b_k as in equation 1, where $F(x)$ follows either DA or ESI to encode a single sentence. With the attention weights $\mathbf{e}$, we obtain $\bar{\mathbf{a}}$ and $\bar{\mathbf{b}}$:

$$\mathbf{e} = F(\mathbf{a})^\top F(\mathbf{b}),\ \mathbf{e} \in \mathcal{R}^{n^a \times n^b} \tag{1}$$

$$\bar{\mathbf{a}}_j = \sum_{k=0}^{K} \frac{\exp(e_{j,k})}{\sum_{m=0}^{K} \exp(e_{j,m})} F(\mathbf{b})_k \tag{2}$$

$$\bar{\mathbf{b}}_k = \sum_{j=0}^{K} \frac{\exp(e_{j,k})}{\sum_{m=0}^{K} \exp(e_{m,k})} F(\mathbf{a})_j \tag{3}$$

Then DA and ESI introduce $G(x_1, x_2)$ to update the representation by taking ($F(\mathbf{a})$, $\bar{\mathbf{a}}$), or ($F(\mathbf{b})$, $\bar{\mathbf{b}}$), as the input. Formally, $\mathbf{h}^a = G(F(\mathbf{a}), \bar{\mathbf{a}})$ and $\mathbf{h}^b = G(F(\mathbf{b}), \bar{\mathbf{b}})$. We recommend readers to refer the origin papers for the implementation of $F(x)$ and $G(x_1, x_2)$. We concatenate the two representations as the final output of the encoder, $\mathbf{h} = [\mathbf{h}^a; \mathbf{h}^b]$. We use the encoder for the retrieval step and the entailment step by varying the input pairs.

2.3 Entailment Score as Relevance Measure

A common design of $V(x)$ is to generate three values $[v^S, v^R, v^N]$, representing the evidence supports the claim, the evidence refutes the claim, and the evidence does not have enough information, respectively. The largest value decides the entailment: $v = \max([v^S, v^R, v^N])$. Intuitively, if one sentence supports or refutes the claim, the sentence must be relevant. Thus, we apply $V(x)$ on all candidate sentences and propose a new form of the relevance score in Equation 4.

Also, we can combine the new relevance score with the old one that intends to capture the relevance on a single sentence. We introduce the final relevance score r^{com} in Equation 5, where α and β can be learnable parameters or fixed hyper-parameters, and r^{rFCN} is the common way that ob-

tains the relevance score via a fully connected network.

$$r^{\text{vDiff}} = \max([v^S, v^R]) - v^N \quad (4)$$

$$r^{\text{com}} = \alpha \cdot r^{\text{rFCN}} + \beta \cdot r^{\text{vDiff}} \quad (5)$$

We optimize the retrieval objective as in Equation 6, by ranking the minimum score of evidence and the maximum score of irrelevant sentences. We use cross-entropy as in Equation 7 for the entailment objective. We sum the two as the joint training objective: $\mathcal{L} = \mathcal{L}^R + \mathcal{L}^V$.

$$\mathcal{L}^R = \frac{1}{N^c} \sum_{i}^{N^c} \max(0, \min([r^+_{i,0} \dots r^+_{i,N^+}]) - \max([r^-_{i,0} \dots r^-_{i,N^-}]) + m) \quad (6)$$

$$\mathcal{L}^V = -\frac{1}{N^c} \sum_{i}^{N^c} y^v_i \log(v_i) \quad (7)$$

One may wonder if a claim requires multiple sentences to form evidence. In that case, v may predict a single sentence *irrelevant*. We argue it is not a concern because the r^+ in our design is capable of taking a negative value while the r^- can take a positive value. As long as r^+ is greater than r^-, we can retrieve the right evidence.

3 Experiments

The focus of the experiments is to understand if the entailment score can benefit the retrieval. We conducted experiments on the FEVER dataset (Thorne et al., 2018)[1]. FEVER contained 80,035 *Support* claims, 29,775 *Refute* claims, and 35,639 *NotEnoughInfo* claims for training. The shared task of FEVER released 6,666 *Support* claims, 6,666 *Refute* claims, and 6,666 *NotEnoughInfo* claims for validation, and held another blind test set of 6,666 *Support* claims, 6,666 *Refute* claims, and 6,666 *NotEnoughInfo* claims. We considered two scenarios in our experiment, and we describe them as follows:

EC: short for **E**ntailment **C**omparison. We explored the claim-evidence entailment by augmenting gold evidence with irrelevant sentences. We varied the irrelevant sentences so that we maintained the recall of the evidence retrieval. This scenario emphasizes the importance of evidence retrieval.

[1]https://fever.ai/data.html

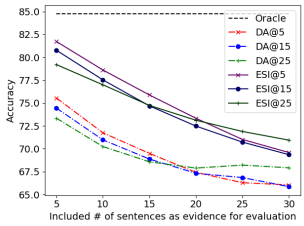

Figure 1: The entailment results on a different amount of sentences as evidence. The Oracle took gold evidence. *Model@number* means the model was trained on the evidence given that number of sentences.

RC: short for **R**etrieval **C**omparison. We followed the three-step pipeline and focused on sentence retrieval. At the document retrieval step, we adopted the strategy of (Zhou et al., 2019; Hanselowski et al., 2018). At the sentence retrieval step, we trained on gold evidence and applied retrieved documents for validation.

3.1 The Importance of Evidence Retrieval

We first experimented against the **EC** to understand the importance of evidence retrieval. The irrelevant sentences are sampled from the same document as the evidence. For claims that did not have gold evidence, we sampled sentences from high ranked documents. We evaluated cases where the evidence contained [5, 10, 15, 20, 25, 30] sentences, while we constructed evidence with [5, 15, 25] sentences for training. Besides, we included the oracle setting that claims were paired with only gold evidence.

We report the result in Figure 1. We notice a clear trend that having irrelevant sentences hurt claim verification, which strengthens the importance of evidence retrieval. We also see that the ESI model performs better than the DA model in all cases, possibly because the ESI model leveraged sequential orders.

3.2 Evidence Retrieval

We conducted experiments against the **RC** scenario to investigate if the claim-evidence entailment can enhance evidence retrieval. We considered three variants of the sentence retrieval step for comparison: **R** was the baseline that no entailment signal was used; **R+V-J** measured the relevance score as in Equation 5; **R-J** also leveraged the entailment task but only used r^{rFCN} for the relevance score.

We selected three previous works to compare against: **TwoW** (Yin and Roth, 2018) combined the retrieval step and the entailment step as **R-J**.

	TwoW	HAN	GEAR	DA			ESI		
				R	R-J	R+V-J	R	R-J	R+V-J
F-Recall	54.81	53.60	**86.72**	85.59	85.51	85.43	86.34	86.55	86.51
MRR@5	-	-	85.19	**82.72**	82.69	82.29	85.15	85.54***	**85.77***
MAP@5	-	-	84.10	81.68	**81.73**	81.29	84.03	84.34***	**84.62****
MRR@A	-	-	-	**83.16**	83.14	83.08	85.55	85.78***	**86.11****
MAP@A	-	-	-	80.04	**80.19**	80.07	82.48	82.53***	**83.16*****

Table 2: Evidence Retrieval Comparison. R is the retrieval baseline. R-J leverages the entailment signal via joint training. R+V-J measures the relevance score as in Equation 5. We use T-test, and ***, **, and * means the difference was significant under level $\alpha < 0.01$, $\alpha < 0.05$, and $\alpha < 0.1$, respectively.

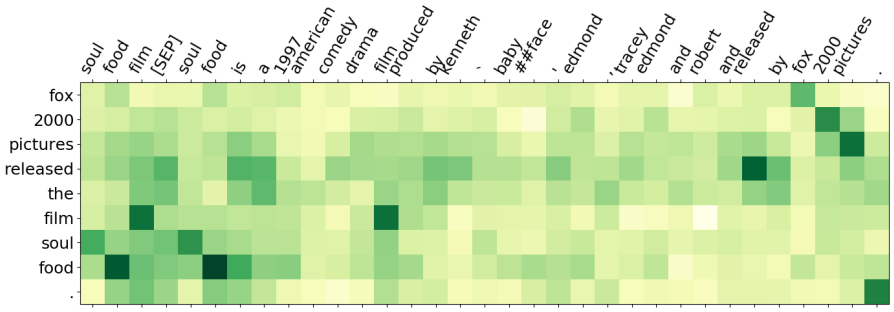

Figure 2: The visualization of the equation 1 using ESI. Each row corresponds to a token of the claim, and each column is a token of the evidence. A darker color means a higher attention score.

They filtered out low confident sentences so that the entailment step could focus on a relevant subset; **HAN** (Ma et al., 2019) also combined the retrieval step and the entailment step and filtered out irrelevant sentences. They leveraged coherence-based attention to model sentence candidates; **GEAR** (Zhou et al., 2019) adopted the three-step pipeline. They shared the same document retrieval step as ours and a similar sentence retrieval model.

We evaluated the retrieval step on three metrics: **F-recall** is the FEVER recall that measures if the top five sentences contained evidence. F-recall would count true positive if at least one evidence was found; **MRR** stands for mean reciprocal rank. Not only measuring if one evidence was selected, but it also considers the highest ranking position of the evidence. MRR favors to select one evidence as confident as possible; **MAP** stands for mean average precision. It cares for *all* evidence to be highly ranked and encourages the retrieval step to have all evidence confidently selected so that the retrieved candidates had less irrelevant sentences.

We report the result in Table 2. Since GEAR only reported results on the top five sentences, we calculated MRR and MAP on top five sentences (MRR@5 and MAP@5) and all sentences (MRR@A and MAP@A). We first observe that leveraging the entailment signal improves evidence retrieval on the ESI model, whereas it shows no improvement in the DA model. One possible reason is that the DA model did not perform well on claim verification compared to the ESI model. Therefore, the DA model could not provide a reliable entailment signal to enhance the retrieval. The ESI model, showing better accuracy to predict the entailment, improves MAP and MRR when we leveraged the entailment prediction (ESI-R+V-J v.s. ESI-R), which reinforced the thought that leveraging the entailment signal would require a good entailment predictor.

We also observe that TwoW and HAN could not efficiently retrieve relevant evidence as other methods. Although they show descent accuracy on claim-evidence entailment, a low F-recall means that filtering out low-rank candidates removed relevant evidence as well. Thus, these models show a disadvantage when people care about the evidence that leaded to a verification result.

Finally, we observe that leveraging the entailment signal did not offer an improvement in terms of F-recall. This might indicate that our method benefits ranking multiple pieces of evidence, as we see better performance on MAP and MRR. Besides, GEAR deployed an ensemble of ten models for retrieval, which could explain the difference.

3.3 Visualization

In Figure 2 we provide one visualization of the claim-evidence attention. We see that the claim and the evidence are attending on the same words and phrases. This explains why the entailment can benefit the retrieval: they reinforce each other to find similar lexicons.

4 Conclusion

In this work, we show that leveraging the entailment prediction can improve evidence retrieval when the entailment step produces a reliable result. In the future, we will adopt pre-trained models, e.g., BERT (Devlin et al., 2019), for our experiments. We expect improvement because BERT shows competitive results on the textual entailment tasks (Zhou et al., 2019).

Ethical consideration This work conducts experiments on benchmark datasets that have been extensively studied in the literature. Although the datasets used in the work was manually annotated, there is no identity characteristics. Also, we use RNN-based models with only a few layers, which are more eco-friendly compared to transformer based models.

Acknowledgments

Research was supported in part by grants NSF 1838147, NSF 1838145, ARO W911NF-20-1-0254. The views and conclusions contained in this document are those of the authors and not of the sponsors. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal.

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

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017b. Recurrent neural network-based sentence encoder with gated attention for natural language inference. In *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*, pages 36–40, Copenhagen, Denmark.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190, Southampton, United Kingdom. Springer.

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, pages 4171–4186, Florence, Italy.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia.

Reza Ghaeini, Sadid A Hasan, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Fern, and Oladimeji Farri. 2018. Dr-bilstm: Dependent reading bidirectional lstm for natural language inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1460–1469, New Orleans, Louisiana, USA.

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

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571, Florence, Italy.

Tsendsuren Munkhdalai and Hong Yu. 2017. Neural tree indexers for text understanding. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 11–21, Lyon, France.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. 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, Hawaii, USA.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, TX, USA.

Lei Sha, Baobao Chang, Zhifang Sui, and Sujian Li. 2016. Reading and thinking: Re-read lstm unit for textual entailment recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2870–2879, Osaka, Japan.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Compare, compress and propagate: Enhancing neural architectures with alignment factorization for natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1565–1575, Brussels, Belgium.

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, pages 809–819, New Orleans, LA, USA.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 1112–1122, New Orleans, Louisiana, USA.

Wenpeng Yin and Dan Roth. 2018. Twowingos: A two-wing optimization strategy for evidential claim verification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 105–114, Brussels, Belgium.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Ucl machine reading group: Four factor framework for fact finding (hexaf). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium.

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.