e-FEVER: Explanations and Summaries for Automated Fact Checking

Dominik Stammbach
Center for Law & Economics
ETH Zurich
dominik.stammbach@gess.ethz.ch

Elliott Ash
Center for Law & Economics
ETH Zurich
ashe@ethz.ch

Abstract

This paper demonstrates the capability of a large pre-trained language model (GPT-3) to automatically generate explanations for fact checks. Given a claim and the retrieved potential evidence, our system summarizes the evidence and how it supports the fact-check determination. The system does not require any additional parameter training; instead, we use GPT-3's analogical "few-shot-learning" capability, where we provide a task description and some examples of solved tasks. We then subsequently ask the model to explain new fact checks. Besides providing an intuitive and compressed summary for downstream users, we show that the machine-generated explanations can themselves serve as evidence for automatically making true/false determinations. Along the way, we report new competitive fact-checking results for the FEVER dataset. Finally, we make the explanations corpus publicly accessible, providing the first large-scale resource for explainable automated fact checking.

1 Introduction

Automated fact checking (AFC) refers to the task of automatically assigning a truth value to a factual statement, i.e., a claim (Vlachos and Riedel, 2014; Thorne et al., 2018a; Augenstein et al., 2019). The standard approach is to frame AFC as a four-step pipeline: (1) extract (or "spot") claims from a corpus, (2) retrieve evidence for the extracted claims, (3) filter the evidence for each claim, and (4) perform recognition of textual entailment (RTE) between each claim and the filtered evidence (Thorne et al., 2018b). If the evidence entails the claim, the claim is *supported* or *true*, otherwise *refuted* or *false* (Thorne et al., 2018a).

The output of traditional fact checking – that is, by news organizations such as Politifact – looks quite different. Journalistic fact checks are news articles, which mainly consist of a discussion and explanation of the associated events, leading up to a verdict about a claim. In general, the fact-checking writers support their verdict by adding information about the claim and putting the claim in context of the evidence. These explanations help readers understand how the verdict has been reached. That said, the veracity of any given claim is often subject to uncertainty, in particular due to vagueness and multiple interpretations. In response, journalistic fact checks usually have multiple levels of "true" or "false". Fact-check scores are designed to help readers contextualize differences in the reliability of different claims. While not perfect, these fact checks often work to make complex events, and (dis)honest statements about those events, easier to understand.

One can see how contextualizing fact checks and explaining them for readers is a complex task. Given the difficulty for humans, some scholars in NLP have argued that explanations should not be attempted by automated fact-checking (AFC) systems (Vlachos and Riedel, 2014). A less fraught approach, taken by the FEVER agenda (Thorne et al., 2018a), is to focus on evidence retrieval for any given claim. This approach evaluates AFC systems not only by the correctness of the veracity prediction, but also by whether all necessary evidence to fact-check a claim is retrieved. These efforts are designed to make AFC systems more interpretable based on the evidence used for the fact check. Encouraging the use of the correct evidence is likely to prevent systems from making the right predictions for the wrong reasons.

Going beyond evidence retrieval, recent work has begun to explore how fact-checking explanations could be generated automatically. The leading paper is (Atanasova et al., 2020), who jointly learn to generate an explanation together with predicting the veracity of a claim. By allow-

Proceedings of the 2020 Truth and Trust Online (TTO 2020), pages 32–43,
Virtual, October 16-17, 2020.

ing downstream users to inspect whether a prediction has been made for the right reasons, automated explanation has the potential to increase system interpretability and to eventually bridge the gap between manual and automated fact checking. The current lack of adequate resources, e.g., fact-checking datasets with explanations, has impeded research efforts in this direction.

Meanwhile, recent breakthroughs in language modeling have revealed that high-capacity transformer language models trained on huge English-language corpora are intrinsic multi-task learners (Radford et al., 2019). GPT-3, in particular, has produced remarkable results across almost all NLP tasks (Brown et al., 2020). The new paradigm is "few-shot learning," where the system is prompted with a task description (e.g., "Translate to French:") and a handful of examples ("hello→bonjour,goodbye→au revoir"). After that, it can solve analogous tasks prompted via natural language ("good night→").

In this work, we show that GPT-3 can also generate coherent and satisfying summaries of relevant evidence w.r.t. a claim, thereby providing an explanation for any given fact check. We show that a claim and its GPT-produced summary are sufficient to predict the majority of the examples in the FEVER development set using a new state-of-the-art fact-checking pipeline based on the DOMLIN system (Stammbach and Neumann, 2019), validating the quality of such summaries. Furthermore, we can show that systems trained on the concatenation of the summary and retrieved evidence outperform systems trained only on the retrieved evidence. We release our explanations as a new dataset, e-FEVER, for use by the fact-checking community.[1]

2 Related Work

2.1 Explainable Fact Checking

This paper adds to the recent literature on generating explanations for AFC systems. The closest paper is Atanasova et al. (2020), who produce abstractive fact-checking explanations on the LIAR-PLUS dataset (Alhindi et al., 2018). That dataset contains 12,836 training claims with a corresponding veracity label and a justification for that label. Atanasova et al find that joint training on veracity prediction and explanation generation improves system performance.

[1]To obtain the dataset, please contact the authors.

Atanasova et al's abstractive summary approach is related to that taken by Camburu et al. (2018) for natural language inference. They produce abstractive explanations for the Natural Language Inference task on the SNLI dataset (Bowman et al., 2015). Ground-truth explanations are generated by Amazon Mechanical Turk workers. Camburu et al find that generating explanations during training helps the model to learn better sentence representations for several downstream tasks (relative to the vanilla InferSent model).

In a different approach to system interpretability, Ahmadi et al. (2019) retrieve evidence from a knowledge base and present it in a rule-based manner. Portelli et al. (2020) use a span extractor to distill the most informative spans in the retrieved evidence. They find that training only on the distilled span increases performance. Moreover, in a manual evaluation they find that this span provides a sufficient explanation in about three-fourths of the verifiable claims. This approach can be understood as extractive summarization of the filtered evidence to explain the fact check.

An abiding challenge to AFC research is that the large-scale AFC datasets – e.g. FEVER (Thorne et al., 2018b) and MultiFC (Augenstein et al., 2019) – do not include any explanations or justifications. Other such resources so far do not exist. The lack of an explanation corpus is an impediment to further research on explainable fact checking.

2.2 Autoregressive Transformer Language Models

Language model pre-training has transformed NLP (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2018; Radford et al., 2019; Brown et al., 2020), in particular through deep contextualized word and document embeddings using transformers. Radford et al. (2019) and Brown et al. (2020) show that autoregressive language models trained on general corpora are multi-task learners, given sufficient data and parameter capacity. An interpretation of these findings is that with enough pre-training data, all or most NLP tasks will be observed and learned. Previous work shows that pre-trained language models can perform translation, entailment, summarization, and classification tasks.

Specifically, the new state-of-the-art GPT-3 breaks new ground through an expanded training

corpus and (especially) an expanded model parameter space. The trained model excels in a variety of NLP tasks in a zero- or few-shot setting. We add to these applications by using GPT-3 to produce explanations of fact checks w.r.t. retrieved evidence in the FEVER dataset.

3 DOMLIN++ Fact Checking System

We work with FEVER, a large-scale fact-checking dataset comprising 185K annotated claims generated by human alteration of sentences from Wikipedia. Verifiable claims in the dataset are associated with annotated evidence. The task is to predict the veracity of a claim and to retrieve the relevant evidence (Thorne et al., 2018a).

Our starting point is the DOMLIN system, which follows the approach outlined above: 1) retrieve evidence, 2) filter evidence, 3) determine entailment (Stammbach and Neumann, 2019). Stammbach and Neumann include a two-stage evidence retrieval strategy to increase evidence recall. The first retrieval stage is standard, in that sentences identified by the document retrieval module are treated as potential evidence. In the second retrieval stage, the first-stage evidence is scanned for hyperlinks, and the content of hyperlinked pages is explored for additional potential evidence. The explored content is conditioned on both the claim and the previously retrieved evidence. Negative evidence sentences for all verifiable claims are randomly sampled out of all retrieved sentences yielded by the document retrieval module.

Our new system, DOMLIN++, differs from DOMLIN in four ways. First, we add a second document retrieval module to further increase evidence recall. This is the same module from the FEVER baseline, ranking sentences from Wikipedia according to their TF-IDF similarity to a given claim and returning the five highest-ranked sentences (Thorne et al., 2018a). Second, we replace the $\text{BERT}_{\text{BASE}}$ model for evidence filtering with a cased $\text{RoBERTa}_{\text{BASE}}$ model (Liu et al., 2019). Third, we collapse DOMLIN's dual evidence retrieval modules (direct and hyperlinked) into a single module. To make this work, we format the input such that the claim and the first evidence both appear before the *SEP* token.[2] Fourth, we replace the $\text{BERT}_{\text{BASE}}$ RTE module with a $\text{RoBERTa}_{\text{LARGE}}$ checkpoint that has been fine-tuned on the MultiNLI dataset (Williams et al., 2018), containing 433K examples for natural language inference.

We predict evidence for every example in the training set using our document and sentence retrieval modules. We then train the RTE module on the claim and the concatenation of the retrieved evidence. Instead of training on annotated gold evidence and retrieved evidence for non-verifiable claims, this procedure minimizes the difference between in- and out-of-sample evidence distributions, i.e., the input at training and testing time. This approach improves performance on the development set. Our pipeline retrieves at least one evidence sentence for 17'687 examples (out of 19'988) in the development set. We label the remaining 2'311 examples heuristically as "NOT ENOUGH INFO" because no information about the claim could be retrieved.

System	Evidence F1 (%)	Label Accuracy (%)	FEVER score (%)
DOMLIN++ (dev set)	37.33	77.48	74.98
UNC-NLP	**52.96**	68.21	64.21
DOMLIN	36.26	71.54	68.46
DeSePtion	37.65	72.47	68.80
DREAM	39.45	**76.85**	70.60
DOMLIN++ (test set)	36.69	76.60	**74.27**

Table 1: Results for the enhanced DOMLIN system

Table 1 presents the main results for the shared FEVER tasks (see Thorne et al., 2018b): correct evidence retrieved (column 1), correct label (true/false) predicted (column 2), and both (FEVER score, column 3). We assess the DOMLIN and DOMLIN++ systems and compare them briefly to other published FEVER systems: UNC-NLP and DREAM. UNC-NLP implement a similar pipeline to ours, including exploring hyperlinks (without conditioning additional evidence), but with Neural Semantic Matching Networks to tackle the subtasks in the pipeline (Nie et al., 2018). DeSePtion uses multiple pointer networks for document selection and jointly models evidence sentences and veracity predictions (Hidey et al., 2020). DREAM takes a different approach, transforming the retrieved evidence into a graph on which graph convolutional networks and graph attention networks propagate and aggregate information (Zhong et al., 2019).

In the evidence task, UNC-NLP (Nie et al., 2018) is the clear winner. In the label task, DREAM does very well and slightly better than DOMLIN++. In the combined (FEVER score) task, DOMLIN++ comes out on top. With some

[2] Specifically, we generate the following input for the claim: *claim evidence_1 SEP evidence_2*.

additional modifications, the FEVER baseline approach (document retrieval, evidence filtering, textual entailment) is still enough to achieve very strong results for FEVER. These modifications include a high-recall document retrieval module (in our case, we combine two different and complementary modules), conditioning evidence as shown in (Stammbach and Neumann, 2019), and using as a classifier a deep transformer decoder pre-trained on natural language inference tasks (in our case an NLI-tuned RoBERTa).

4 Teaching GPT-3 to Explain Fact Checks

4.1 Approach

The FEVER task consists of evidence retrieval and veracity prediction, as just reviewed. In the standard task, however, no explanation of the fact check is required. To make progress on this front, we use the evidentiary text retrieved by DOMLIN++ as the basis for generating summaries. We leverage the few-shot capabilities of GPT-3 to generate abstractive summaries of evidence w.r.t. claims in the FEVER task. As discussed in 2.2, most NLP tasks will be observed during language modeling pre-training, among them the task of summarizing evidence w.r.t. a claim.

[A] Summarize the context with respect to the claim. The summary should be such that we can fact-check the claim using this summary. Claims can be true, false or not verifiable. It is very possible that no relevant information about the claim can be found in the context. In this case, the summary should be: "The relevant information about the claim is lacking in the context."

[B] context: Fox Broadcasting Company + The Fox Broadcasting Company (often shortened to Fox and stylized as FOX) is an American English language commercial broadcast television network that is owned by the Fox Entertainment Group subsidiary of 21st Century Fox . + Nikolaj Coster-Waldau + He then played Detective John Amsterdam in the short-lived Fox television series New Amsterdam (2008) , as well as appearing as Frank Pike in the 2009 Fox television film Virtuality , originally intended as a pilot .

[C] claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.

[D] summary: Nikolaj Coster-Waldau played Detective John Amsterdam in the series New Amsterdam (2008) and appears as Frank Pike in the film Virtuality. Both are produced by the Fox Broadcasting Company, so he worked with the Fox Broadcasting Company.

Figure 1: GPT-3 Summarization Prompt

We teach GPT-3 to explain a fact check using the seed text input. First, we directly prompt the language model to summarize an evidence context w.r.t. a claim. We provide a relatively detailed task formulation, shown in Figure 1 Part A.[3] This is the general prompt that goes above any examples to learn from or to complete.[4]

Part B shows an example context, which is the set of evidence snippets retrieved by DOMLIN++. We prepend the page title (surrounded by "+") to each sentence for coreference resolution. Note that this list of evidence snippets can already be seen as an extractive summary of the fact check in the context of all retrieved evidence by the document retrieval modules. However, as this example demonstrates, these snippets do not work well as summaries. They tend to be lengthy, incoherent, disconnected, and full of irrelevant information. Thus the need for an abstracted representation.

In Part C, we provide the claim to be fact-checked. In Part D, we provide the desired output, i.e. the summary we created for this example.

To evaluate zero-shot learning, GPT-3 needs A, B, and C from Figure 1. We would then append "summary:" and ask the model to finish the prompt. In experiments, we found that zero-shot summaries did not perform well. Some examples are provided in Appendix A.2.

Summary quality increases by providing GPT-3 some examples to learn from – that is, few-shot learning. After some experimentation, and accounting for the space constraints of the model's seed text input, we settled on a specification with eight example fact-check explanations, with our own manually written summaries. An example provided summary is shown in Figure 1 Part D.[5] We would then go on to repeat parts B through D with the seven other examples. Then, the ninth item would be the new fact check to explain.

Note that this task is quite different from a standard abstractive summary. The evidence consists of multiple disconnected statements, some of which are irrelevant. They must be connected analytically to form the basis of a fact check. Also, claims can contain false information which has to be substituted with the correct information found in the evidence. Finally, GPT-3 has to determine whether a summary should be generated in the first place or whether it should fall back to the null summary.

4.2 Demonstration

Table 2 reports a hand-picked sample of GPT-3-produced explanations for fact-checks from the FEVER development set. While hand-picked,

[3] In the figure, items in brackets are for annotation purposes and were not included in the language model seed text.

[4] For a discussion about the statement *The relevant information about the claim is lacking in the context*, see Appendix A.3.

[5] The additional prompt examples are reported in Appendix A.1.

Claim	Extracted Evidence	Generated Explanation
Marnie was directed by someone who was "The Master of Nothing".	+ Alfred Hitchcock + Sir Alfred Joseph Hitchcock (13 August 1899 - 29 April 1980) was an English film director and producer , at times referred to as " The Master of Suspense " . + Marnie (film) + Marnie is a 1964 American psychological thriller film directed by Alfred Hitchcock ."	Marnie was directed by Alfred Hitchcock, who was "The Master of Suspense".
Colin Kaepernick is a poker player.	+ Colin Kaepernick + Colin Rand Kaepernick (born November 3 , 1987) is an American football quarterback who is currently a free agent . + American football + American football , referred to as football in the United States and Canada , and also known as " gridiron football " or simply " gridiron " , is a sport played by two teams of eleven players on a rectangular field with goalposts at each end .	Colin Kaepernick is a football player, not a poker player.
The Mighty Ducks was produced by Avnet-Kerner Productions in 2017.	+ The Mighty Ducks + It was produced by The Kerner Entertainment Company and Avnet - Kerner Productions and distributed by Walt Disney Pictures . + The Mighty Ducks + The Mighty Ducks is a 1992 American sports comedy-drama film directed by Stephen Herek , starring Emilio Estevez ."	The Mighty Ducks was produced by The Kerner Entertainment Company and Avnet–Kerner Productions in 1992.

Table 2: Examples of abstractive summaries produced by GPT-3

they are nevertheless impressive examples of GPT-3's analytical summarization capacities. In the first row, retrieving the full evidence set is only possible via conditioning multi-hop evidence (Stammbach and Neumann, 2019). GPT-3 correctly realizes that Marnie was directed by Hitchcock, and that Hitchcock was referred to as the "Master of Suspense". In the second row, the correct and complete fact check is already entailed in the generated summary, which contains all the relevant information, i.e., Colin Kaepernick being a football player and not a poker player.

The last example is interesting – it again highlights the problems in multi-document extractive summarization. In such cases, extractive summaries tend to be overly verbose and/or biased towards some parts of the evidence (Carenini et al., 2006). Abstraction works to circumvent that problem.

To highlight the information compression achieved by abstract summarization, Figure 2 plots the distributions of sequence lengths (number of words) for the retrieved evidence (in blue) and the generated abstractive summaries (in red). The retrieved evidence snippets are much longer and have a flat tail, with an average length of 84.7 words. The GPT-3 summaries, in contrast, are much shorter (averaging 13.5 words long), with a sharp peak at around 10-12 words, partially because the length of the null summary in case of no information found is 11 words.

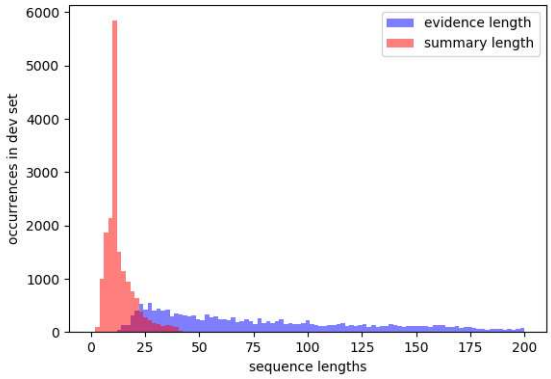

Figure 2: Distribution of sequence lengths (in words) for retrieved evidence and generated summaries

4.3 Evidence Summaries are Sufficient for Fact Checking

Now we investigate the use of the abstractive evidence summaries to perform the fact-checking task. To this end, we revisit the RTE module from the DOMLIN++ system to classify veracity in the development set. The new ingredient is to use as input the abstractive summaries produced by GPT-3, rather than directly using the retrieved evidence.

Input to the RTE classifier	Label Accuracy (%)
(1) claim + retrieved evidence	80.6
(2) claim + oracle deciding on label	86.0
(3) claim + abstractive summary	65.5
(4) claim + abstractive summary (cleaned devset)	72.9
(5) claim + abstractive and retrieved evidence	73.2

Table 3: Veracity Accuracy on Reduced Development Set

The results of this exercise are reported in Table 3. Our baseline is DOMLIN++ with the standard input (the claim and the retrieved evidence), achieving over 80% label accuracy on the examples for which we retrieve evidence (row 1). The upper-bound baseline, based on a perfect oracle deciding the veracity of a given claim given the verdict about the extractive and abstractive summary, we would achieve a label accuracy of 86% (row 2).

If we use as input the abstractive summaries generated by GPT-3, we still achieve an overall 65% accuracy on the development set (row 3). If we filter out those claims for which GPT-3 incorrectly generates the null summary,[6], the accuracy with using summaries as input jumps up to 73%

[6]*The relevant information about the claim is lacking in the context.*

(row 4). Overall, these numbers are promising and provide indirect evidence that most summaries are high-quality in terms of containing the relevant information.

The summaries do worse overall than the baseline, but an important question is whether they make the same mistakes. We compared their errors and found that the model from row 2 (using just the summary) predicts veracity correctly in about one-third of the cases where the baseline (row 1, using retrieved evidence) predicts incorrectly. Thus, we have some evidence that the two approaches are complementary. It could be that the baseline model errs in response to redundant and/or repeated information, which the compressed summary input addresses.

Motivated by this potential complementarity, we let DOMLIN++ predict the development set with both the retrieved evidence and the abstractive summary as input. However, we see in Table 3 row 5 that this model produces intermediate performance of 73.2% accuracy – better than the setting using just the summaries, but still worse than the baseline.

Class	Pr (%)	Rc (%)	F1 (%)
SUPPORTS	72 (79)	77 (90)	74 (84)
REFUTES	73 (82)	69 (74)	71 (78)
NOT ENOUGH INFO	49 (71)	47 (69)	48 (70)
MACRO	65 (77)	65 (77)	(77)

Table 4: Pr, Rc and F1 for different classes

In Table 4, we report precision, recall and F1 for the different classes in the development set using the summaries as input. In brackets, we report results for the DOMLIN++ pipeline (using the retrieved evidence as input). Performance for the verifiable classes remains rather stable, dropping around 10 F1 points for both classes. We still achieve a remarkable performance for both classes of above 70% F1. Given the difficulties in the non-verifiable cases, we expected to see a decline in performance and observe less than 50% F1, losing 22 F1 points in this setting. We attribute some of these declines over all classes to the faulty generation of the null summaries.

4.4 e-FEVER Dataset

Inspired by these promising results, we used the GPT-3-based system to generate summaries for the 17'687 claims in the FEVER development set (for which we can find evidence), as well as an additional 50K examples from the training set. The evidence for all these examples is retrieved by the pipeline described above. This process has resulted in a new dataset, which we call *e-FEVER* (explained FEVER), consisting of 67'687 examples. Each example contains the evidence retrieved by our fact-checking pipeline as well as the abstractive summary.

To the best of our knowledge, this is the first large-scale resource available for explainable fact checking. We make it publicly accessible in the hope that this further stimulates research in this area. Although the summaries are machine-generated, we find them to be of sufficient quality to successfully train new systems on, as shown in subsequent experiments.

We hypothesize that the summaries where DOMLIN++ predicts the correct veracity given the claim and the summary should be the most reliable and useful ones to perform further research on. However, this still has to be validated in future experiments.

5 Further Experiments on Fact Checking with Summaries

This section provides additional experiments to investigate the quality and effectiveness of the GPT-3-produced summaries. Specifically, we fine-tune several RoBERTa$_{\text{BASE}}$ models on the 50K training examples in e-FEVER to predict the veracity of claims. We vary the model training along two margins – first, by the model input: (1) claim only, (2) claim and retrieved evidence (baseline), (3) claim and generated summary, and (4) claim, generated summary, and retrieved evidence. Second, we vary the sample size in the training set. All models are trained for 2 epochs using a learning rate of 2e-5 and a batch size of 24.

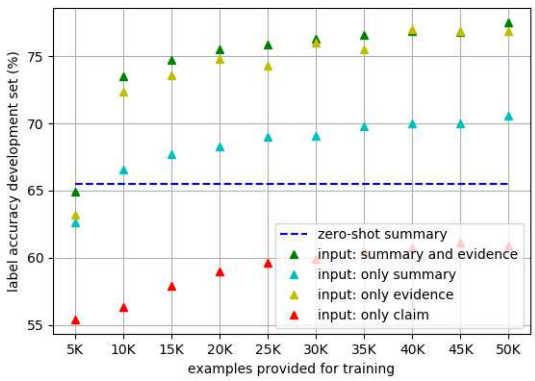

Figure 3: Sample efficiency for additional fine-tuning

The results of these experiments are reported in Figure 3. The vertical axis corresponds to classification accuracy in the FEVER development set. The horizontal axis corresponds to the training set size. The different series (distinguished by colors) correspond to the different combinations of inputs. The horizontal dashed blue line at 65.5% indicates DOMLIN++'s performance using the summaries as input (from Table 3).

First, consider the red series, representing performance using only the claims as input. This specification performs relatively poorly on veracity prediction and never beats the dashed blue line. This low accuracy supports the view that the GPT-3 summaries are providing useful information for fact-checking, rather than just exploiting spurious patterns in the dataset.

Next take the turquoise series. Fine-tuning on the summaries as input increases accuracy significantly, by 8 percent with the full 50K-example training set compared to only fine-tuning on 5K examples. Performance of DOMLIN++ using these summaries as input (dashed blue line) is surpassed by fine-tuning on 10K examples. Note that the achieved 70% accuracy with the full 50K-example should be interpreted relative to a maximal upper bound of 86%, imposed by the 14% error rate from mistakenly generating the null summary.

Lastly, in the top part of the figure we have the model that uses retrieved evidence (light green) as well as the model that uses both the summaries and the retrieved evidence (dark green). Training jointly on summaries and retrieved evidence slightly outperforms the vanilla setting of training on retrieved evidence only. It could be that the summaries catch some nuances in the data which would have been missed by fine-tuning on evidence only.

Meanwhile, in all models we observe convergence in performance at around 40K claims in the training set. This is perhaps a helpful indication that e-FEVER's training set of 50K examples is sufficiently large for further research.

These results could be improved in a number of directions. For one, we speculate that training on $\text{RoBERTa}_{\text{LARGE\_NLI}}$ would make a difference. It would also help to have summaries for the whole training set by mitigating the faulty generation of *The relevant information about the claim is lacking in the context.*. These improvements would likely generate summaries that outperform the DOMLIN++ system in terms of label accuracy.

6 Concluding Discussion

6.1 Recap

This paper has made several contributions to the literature on automated fact checking (AFC). First, we present a new AFC pipeline achieving competitive results on FEVER. We use this pipeline to retrieve evidence serving as input for summaries and to validate summaries.

Second, we generate and assess abstractive summaries of retrieved evidence w.r.t. claims using GPT-3's few-shot capabilities. We present a hand-picked sample of these summaries for qualitative validation, and quantitatively show that a majority of them should be useful in terms of down-stream fact-checking and perhaps other tasks. We further show that training on summaries and retrieved evidence outperforms training on retrieved evidence only. Still, a key limitation at this stage is the generation of null summaries indicating insufficient information.

Third, we produce a new large-scale dataset of fact-checking explanations. We release the dataset for the research community and show that training on the machine explanations yields reasonable results. We hope that this resource stimulates further research in explainable fact checking, being the first large-scale resource available for the task.

6.2 Summaries as Explanations

A final note is in order on summarization and explanation in the context of fact checking. Summarization is the task of compressing a sequence – that is, to condense its most salient and informative bits in a more compact sequence. In extractive summarization, the most important words, phrases or sentences from a document are extracted and presented as a new document, usually in chronological order. Filtering evidence in the AFC pipeline can be seen as an extractive summary of all retrieved evidence for a given claim. In abstractive summarization, meanwhile, the input document is encoded and a summary is generated that contains the most salient and important information of the original document. Good abstractive summaries preserve the meaning of the input document. In both cases, it is intuitive to see how good summaries can also serve as good explanations.

There is notable work that further distinguishes summaries as *indicative*, *informative*, or *critical* (Hahn and Mani, 2000; Radev et al., 2002). *Indicative summaries* provide enough content to alert users to the source through classical information retrieval approaches, while *informative summaries* substitute the source by providing the most relevant information. *Critical summaries* add value by applying expertise not available in the source texts.

How to relate these ideas to the special concerns of fact-checking? We argue that automatic explanations generated by fact-checking systems should satisfy the following properties:

- Summarize relevant information about a claim from the retrieved evidence;
- Ensure that decisions are made for the right reasons;
- Make the fact check more useful to downstream users.

These properties apply quite well to the abstractive *informative summaries* of evidence w.r.t. a claim we produced with GPT-3, as exemplified by those in Table 2. Thus, the problem of explaining fact checks – perhaps an intimidating task in principle – can in practice be reduced to abstractive summarization – a task with which the NLP community is familiar.

Taking this view opens the door to further exploration and applications of the newest NLP tools for abstractive summarization. Generating fact-check explanations will qualitatively benefit users by providing clues about how AFC systems interpret evidence. As we have previewed here, these explanations might also improve the quantitative performance of these systems.

Returning to our original motivation, it is worthwhile to consider how our work relates to journalistic fact-checking. Unlike our explanation of Wikipedia statements, journalistic fact checking consists of *critical summaries*, which include not just parsing of given text evidence but also elicitation of additional expertise not available in the text. This additional criticism element has important implications for how to interpret automated fact-checking systems based on news articles, such as the LIAR-PLUS dataset (Alhindi et al., 2018).

We can see the different character of journalistic fact-checking clearly in a LIAR-PLUS example from Atanasova et al. (2020), reproduced in the appendix as Figure 9. In that example, the Politifact journalists checked the following claim: ”The last major oil spill from a drilling accident in America happened over 40 years ago in 1969”. This statement is labeled ”Half-True” because, while the 1969 spill had been the largest in 40 years, interviews with three scientists established that the volume of a spill is not the most important factor in its environmental impact. The journalists assessed that the speaker was trying to minimize the recent environmental impact of recent oil spills. Therefore the claim is misleading or ”Half-True.”

In that case, as is common to these articles, the journalistic fact-checkers had to do something much more subtle than checking entailment of lists of statements on Wikipedia. In particular, they had to assess the intentions of the speaker and perform some original data collection (interviews of scientists). These are not tasks that automated fact-checking systems can reliably perform because they require information that is not available in the source text. This additional expert knowledge, in terms of questioning motives and comparing oil spills, would not be part of the input in any conceivable AFC system.

The extra information available to journalists is what distinguishes critical summaries like those published in Politifact. The explanation process itself involves production of new facts. In contrast, when automated abstractive summarization systems add facts – even systems as advanced as GPT-3 – those facts are rarely faithful enough to reality to support effective fact checking. More likely, these language model ”hallucinations” will add errors and reduce interpretability (Tian et al., 2019; Maynez et al., 2020).

These points voice a cautionary note about (explainable) AFC using LIAR-PLUS and similar datasets. While scraping and dissecting these corpora seems tempting, there is an insurmountable problem that critical summarization requires the production of new information unavailable to the classifier. It is therefore unlikely that models trained on fact-checking organization corpora will generalize to the task of performing fact checks on raw text data.

More generally, we should be skeptical about

efforts to build AFC systems for statements about complex events characterized by high uncertainty and strong incentives for deception. Looking forward, we do not see these challenges (complexity, uncertainty, and strategic deception) going away any time soon. Instead, we propose to allocate research attention to the more realistic task of generating informative summaries for more prosaic tasks like those in FEVER.

References

Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable fact checking with probabilistic answer set programming.

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims.

Samuel R. 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 (EMNLP)*. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9539–9549. Curran Associates, Inc.

Giuseppe Carenini, Raymond Ng, and Adam Pauls. 2006. Multi-document summarization of evaluative text. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

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*.

Udo Hahn and Inderjeet Mani. 2000. The challenges of automatic summarization. *IEEE Computer*, 33:29–36.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8593–8606, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2018. Combining fact extraction and verification with neural semantic matching networks. *CoRR*, abs/1811.07039.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

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.

Dragomir R. Radev, Eduard Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Dominik Stammbach and Guenter Neumann. 2019. Team DOMLIN: Exploiting evidence enhancement for the FEVER shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 105–109, Hong Kong, China. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and verification (fever) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9. Association for Computational Linguistics.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P. Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22. Association for Computational Linguistics.

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 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2019. Reasoning over semantic-level graph for fact checking.

A Details of GPT-3 Prompt Design

A.1 Few-shot System Prompt

In Figure 4, we provide the full system prompt with all few-shot learning examples. We replaced the word ”evidence” with the word ”context”, which increased the quality of the produced explanations. The prompt was followed by the retrieved evidence for a given claim, its claim, and the token ”summary:”.

Summarize the context with respect to the claim. The summary should be such that we can fact check the claim using this summary. Claims can be true, false or not verifiable. It is very possible that no relevant information about the claim can be found in the context. In this case, the summary should be: ”The relevant information about the claim is lacking in the context.”

context: + Fox Broadcasting Company + The Fox Broadcasting Company (often shortened to Fox and stylized as FOX) is an American English language commercial broadcast television network that is owned by the Fox Entertainment Group subsidiary of 21st Century Fox . + Nikolaj Coster-Waldau + He then played Detective John Amsterdam in the short-lived Fox television series New Amsterdam (2008) , as well as appearing as Frank Pike in the 2009 Fox television film Virtuality , originally intended as a pilot .
claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
summary: Nikolaj Coster-Waldau played Detective John Amsterdam in the series New Amsterdam (2008) and appears as Frank Pike in the film Virtuality. Both are produced by the Fox Broadcasting Company, so he worked with the Fox Broadcasting Company.

context: + Sense and Sensibility + Sense and Sensibility is a novel by Jane Austen , published in 1811 . + Jane Austen + Jane Austen (16 December 1775 - 18 July 1817) was an English novelist known primarily for her six major novels , which interpret , critique and comment upon the British landed gentry at the end of the 18th century .
claim: Sense and Sensibility was published in the summer of 1811.
summary: Sense and Sensibility was published in 1811, however it does not say whether it has been published in summer.

context: + Filmfare + Filmfare is an English-language , tabloid-sized magazine about Hindi-language cinema , popularly known as Bollywood . + Bollywood + Bollywood is the sobriquet for India ’s Hindi language film industry , based in the city of Mumbai , Maharashtra .
claim: Filmfare is about cheese.
summary: Filmfare is about Hindi-language cinema, not about cheese.

context: + Tiger Woods + After winning the Arnold Palmer Invitational on March 25 , 2013 , he ascended to the No. 1 ranking once again , holding the top spot until May 2014 . + Arnold Palmer Invitational + The Arnold Palmer Invitational is a professional golf tournament in Florida on the PGA Tour .
claim: Tiger Woods had an uncontested victory at the Arnold Palmer Invitational.
summary: Tiger Woods has won the Arnold Palmer Invitational on March 25, 2013. It does not say whether his victory was uncontested.

context: + Liverpool F.C. + Liverpool was the ninth highest-earning football club in the world for 2014 - 15 , with an annual revenue of $ 391 million , and the world ’s eighth most valuable football club in 2016 , valued at $ 1.55 billion .
claim: Liverpool F.C. was valued at $1.55 billion at one point.
summary: In 2016, Liverpool F.C. was valued at $ 1.55 billion.

context: + 19th G7 summit + The Group of Seven (G7) was an unofficial forum which brought together the heads of the richest industrialized countries : France , Germany , Italy , Japan , the United Kingdom , the United States , Canada (since 1976) and the President of the European Commission (starting officially in 1981) .
claim: The 19th G7 summit only included Russia.
summary: The 19th G7 summit did not only include Russia, but also the heads of the six other richest industrialized countries and the President of the European Commission.

context: + 2013 NBA draft + The 2013 NBA draft was held on June 27 , 2013 , at Barclays Center in Brooklyn , New York .
claim: The 2013 NBA draft was not held on June 27, 2013.
summary: The 2013 NBA draft was held on June 27, 2013 in Brooklyn , New York.

context: + Artemis + She was the Hellenic goddess of the hunt , wild animals , wilderness , childbirth , virginity and protector of young girls , bringing and relieving disease in women ; she often was depicted as a huntress carrying a bow and arrows .
claim: Artemis was the protector of Nazgul.
summary: ”The relevant information about the claim is lacking in the context.”

Figure 4: GPT-3 Summarization Prompt

A.2 Zero-Shot Generation of Summaries

In this section, we show some examples of providing GPT-3 only with the task description of summarizing a claim w.r.t. a claim in a zero-shot setting, the claim in question, and its context. The sequence **in bold** is generated by GPT-3 and the text before is the corresponding system prompt. The task description provided in Figure 1 part A is prepended to all of the following examples.

context: + Fox Broadcasting Company + The Fox Broadcasting Company (often shortened to Fox and stylized as FOX) is an American English language commercial broadcast television network that is owned by the Fox Entertainment Group subsidiary of 21st Century Fox . + Nikolaj Coster-Waldau + He then played Detective John Amsterdam in the short-lived Fox television series New Amsterdam (2008) , as well as appearing as Frank Pike in the 2009 Fox television film Virtuality , originally intended as a pilot .
claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
summary: **This claim is false.**
fact check: + https://www.imdb.com/name/nm0900897/

Figure 5: GPT-3 Zero-Shot Summarization (temperature = 0.9

context: + Fox Broadcasting Company + The Fox Broadcasting Company (often shortened to Fox and stylized as FOX) is an American English language commercial broadcast television network that is owned by the Fox Entertainment Group subsidiary of 21st Century Fox . + Nikolaj Coster-Waldau + He then played Detective John Amsterdam in the short-lived Fox television series New Amsterdam (2008) , as well as appearing as Frank Pike in the 2009 Fox television film Virtuality , originally intended as a pilot .
claim: Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.
summary: **The relevant information about the claim is lacking in the context.**

Figure 6: GPT-3 Zero-Shot Summarization (temperature = 0

context: + Artemis + She was the Hellenic goddess of the hunt , wild animals , wilderness , childbirth , virginity and protector of young girls , bringing and relieving disease in women ; she often was depicted as a huntress carrying a bow and arrows .
claim: Artemis was the protector of Nazgul.
summary: **”Artemis (or Diana) was the Hellenic goddess of the hunt, wild animals, wilderness, childbirth, virginity and protector of young girls, bringing and relieving disease in women; she often was depicted as a huntress carrying a bow and arrows. http://www.theoi.com/Daimon**

Figure 7: GPT-3 Zero-Shot Summarization (temperature = 0.9

context: + Artemis + She was the Hellenic goddess of the hunt , wild animals , wilderness , childbirth , virginity and protector of young girls , bringing and relieving disease in women ; she often was depicted as a huntress carrying a bow and arrows .
claim: Artemis was the protector of Nazgul.
summary: **+ The Nazgul were the nine servants of Sauron in the Lord of the Rings .**

Figure 8: GPT-3 Zero-Shot Summarization (temperature = 0

A.3 What to do about ”NOT ENOUGH INFO” examples

Designing a prompt for verifiable examples in FEVER is straight-forward. The challenge is to design a prompt priming GPT-3 to recognize whether there exists such a summary w.r.t. the claim in the first place. This includes most examples from the *NOT ENOUGH INFO* class. Consider the following claim: *Artemis was the protector of Nazgul.*. We retrieve the following evidence for this claim: *+ Artemis + She was the Hellenic goddess of the hunt , wild animals , wilderness , childbirth , virginity and protector of young girls , bringing and relieving disease in women ; she*

often was depicted as a huntress carrying a bow and arrows. There clearly does not exist such a summary, so we should not aim to construct one. We primed GPT-3 to generate *The relevant information about the claim is lacking in the context* in these cases. We have included the *Artemis* example as one of the few-shot examples in the system prompt, but additionally pointed to this peculiarity in the task description of the system prompt. If we do not include both, GPT-3 often copy-pastes the claim and returns the claim as the summary. Although this does not work perfectly, we did not manage to outperform this approach. We leave the problem for future research.

A.4 Journalistic Fact-Checking as Critical Summarization

Claim: The last major oil spill from a drilling accident in America happened over 40 years ago in 1969.

Ruling Comments: (...) The last major oil spill from a drilling accident in America happened over 40 years ago in 1969.
(...) The largest in volume was the Santa Barbara spill of 1969 referenced by Murdock and Johnson, in which an estimated 100,000 barrels of oil spilled into the Pacific Ocean, according to the API. The Santa Barbara spill was so big it ranked seventh among the 10 largest oil spills caused by marine well blowouts in the world, the report states. Two other U.S. spills, both in 1970, rank eighth and 10th. Fourteen marine blowouts have taken place in the U.S. between 1969 and 2007. Six of them took place after 1990 and spilled a total of nearly 13,700 barrels.
(...) We interviewed three scientists who said that the impact of a spill has little to do with its volume. Scientists have proven that spills far smaller than Santa Barbara's have been devastating.

Justification: While the nation's largest oil well blowout did take place in 1969, it's not factually correct to call it the "last major oil spill". First of all, two of the largest blowouts in the world took place in the U. S. the following year. More importantly, experts agree that spills far smaller in volume to the 1969 disaster have been devastating. From a scientific perspective, Johnson's decision to single out the 1969 blowout as the last "major" one makes no sense.

Ruling: Half-True

Figure 9: Example from the LIAR-PLUS dataset (figure taken from (Atanasova et al., 2020))

Existing resources for explainable fact checking contain claims and *extracted justifications* for the claims, e.g., the LIAR-PLUS dataset (Alhindi et al., 2018). An example is shown in Figure 9, where the oracle sentences for generating the explanation are highlighted in various colors.