Weakly Supervised Claim Localization in Scientific Abstracts

Marc Brinner[1](✉), Sina Zarrieß[1], and Tina Heger[2]

[1] Computational Linguistics, Department of Linguistics, Bielefeld University, Bielefeld, Germany
{marc.brinner,sina.zarriess}@uni-bielefeld.de
[2] Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany
t.heger@tum.de

Abstract. We explore the possibility of leveraging model explainability methods for weakly supervised claim localization in scientific abstracts. The resulting approaches require only abstract-level supervision, i.e., information about the general presence of a claim in a given abstract, to extract spans of text that indicate this specific claim. We evaluate our methods on the SciFact claim verification dataset, as well as on a newly created dataset that contains expert-annotated evidence for scientific hypotheses in paper abstracts from the field of invasion biology. Our results suggest that significant performance in the claim localization task can be achieved without any explicit supervision, which increases the transferability to new domains with limited data availability. In the course of our experiments, we additionally find that injecting information from human evidence annotations into the training of a neural network classifier can lead to a significant increase in classification performance.

Keywords: Explainability · Evidence localization · Claim verification

1 Introduction

A claim lies at the center of most scientific publications, as it constitutes the core proposition that is put forth for consideration and is targeted by the presented evidence [19]. Detailed knowledge about these claims addressed in scientific publications is essential for tasks like literature search and scientific claim verification [40], leading to a variety of research targeted at the annotation, recognition and localization of claims in scientific abstracts and full texts (see Sect. 2.1). Despite significant progress being made, the reliance on direct supervision (e.g., [23,41]) often limits the potential of these approaches, since large and high-quality datasets are uncommon in general and not present at all in many specific domains, and since existing models struggle to generalize to different domains [36]. Especially for the task of localizing evidence for claims within a text, the annotation process for creating the dataset is very time-intensive

P. Cimiano et al. (Eds.): RATIO 2024, LNAI 14638, pp. 20–38, 2024.
https://doi.org/10.1007/978-3-031-63536-6_2

and thus more costly, which naturally raises the question of whether a weaker supervision signal, that could be quicker, easier and more consistent to annotate, could be sufficient for solving this complex task.

In this study, we explore the possibility of using weak supervision for the task of claim localization in scientific abstracts. The supervision signal is the information about the general presence of a specific claim in a given abstract (i.e., a textual formulation of that claim or a discrete claim label). This information is used to train a standard neural network classifier that is able to verify the presence of such a semantically distinct claim in a given abstract. We then use model explainability approaches to create a rationale for the classification, which therefore selects spans or sentences from the input that constitute the evidence for the given claim. This is, to our knowledge, the first study that explores the sole use of weak supervision for solving this task.

To test our methods, we evaluate them on two datasets of scientific abstracts with annotated evidence. The first one is the INAS dataset [3], a dataset consisting of scientific abstracts from the field of invasion biology, annotated with information about which hypothesis from the field is addressed. Since no evidence annotations are provided by [3], we perform our own annotation study and annotate 750 abstracts with span-level hypothesis evidence. The second dataset is the SciFact dataset [35], which consists of hand-written claims for a set of scientific abstracts, in combination with sentence-level evidence annotations.

To explore the limits of using explainability approaches for evidence localization, we perform an experiment on injecting the information from evidence annotations into the training process of neural network classifiers. A similar approach has been explored by [38], but we are not aware of such techniques being used for claim verification. In our testing, we find that our method is able to drastically increase the classification performance of the resulting classifier.

The rest of our work is structured as follows: In Sect. 2 we provide background knowledge about scientific claim detection as well as the concept of using input optimization for model interpretability, while Sect. 3 will describe the datasets used in this study. Section 4 then explains our approach for localizing claim evidence as well as a method for injecting evidence information into a standard neural network classifier, while Sects. 5 and 6 will detail the corresponding experiments and results. Section 7 concludes this work with final thoughts and remarks.

The code for the experiments conducted in this study is available at github.com/inas-argumentation/claim_localization.

2 Background

2.1 Scientific Claim Detection

Scientific claim detection has its root in the field of general argument mining, which was formally introduced by [22] and is concerned with locating, classifying and linking argumentative components (so-called argumentative discourse units) in a given argumentative text. Based on the general theory of argumentation

[8,22,34] determined the *claim* to be the center of an argument, as it is the core proposition that is put forth for consideration. A claim, by its nature, is not inherently true and requires further substantiation, which is provided by *premises*, i.e., statements that are generally accepted to be true and do not require further support [29].

As scientific texts are argumentative in nature, the field of argument mining naturally extends to the scientific domain. Recognizing the argumentative structure in a scientific text, as well as the main claim in particular, is essential in tasks like literature search and scientific claim verification [40], leading to the creation of a variety of annotation schemes and datasets [1,10,31,32], many of which focus specifically on the scientific claim: [2] creates a detailed annotation scheme that captures the variety of ways a claim can be formulated in a scientific abstract, [35] create the scientific claim verification task by creating a dataset of hand-written scientific claims and by annotating which sentences in a corpus of scientific abstracts supports or refutes them, and [3] focus on a precise semantic categorization of scientific claims by annotating and classifying claims according to a domain-specific hypothesis network.

Given a specific claim, our study addresses the precise localization of evidence for this exact claim in a given scientific abstract. While many approaches have been proposed to solve similar tasks [2,13,23,41], these methods leverage data annotated on sentence level for supervised learning, which can limit their potential due to the rather small available datasets and the unavailability of any annotated data in many domains. Reasons for this lack of data include the need for expert annotators caused by the focus on the scientific domain, the time-intensive annotation process, as well as the complexity of the annotation task even for domain experts [11].

To our knowledge, no method exists that can reliably detect and locate claims in scientific texts without access to a dataset of samples with explicit sentence-level claim annotations, which can be a problem if a model shall be adapted to a new domain without an existing dataset, as performance has been shown to significantly decrease on out-of-domain samples [36]. Our study aims at closing this gap by creating an approach that only requires weak supervision in the form of abstract-level labels, thus drastically reducing the time and cost needed to create a training set for a new domain.

2.2 Input Optimization for Model Interpretability

For many datasets, evidence annotations for specific claims constitute a rationale for a corresponding classification (e.g., for the claim verification task [35], claim evidence is an explanation for an abstract-level validity label). This characterization of claim evidence creates a natural connection to the field of model interpretability, which is concerned with creating explanations for decisions (e.g., classifications) of black-box machine learning models like neural networks. In the field of natural language processing, explanations for classifications often take the form of individual scores assigned to each input token, with a higher score indicating an increased significance of that token for the predicted score. While a variety of methods have been proposed [20], we will focus on a recent study by [4], as their method called MaRC (Mask-based Rationale Creation) is specifically designed to extract longer, consecutive spans of text as explanations, thus making the explanations better aligned with human reasoning and annotations.

The MaRC approach relies on the concept of input optimization: As neural networks are differentiable, it is possible to calculate the gradient of an objective function with respect to the input features. The MaRC approach uses this concept to remove words from the input by gradually replacing them by *PAD* tokens (in the case of BERT) in a way that maximizes the likelihood of the class that is to be explained, meaning that the words that remain are highly indicative of the respective class.

The MaRC approach assigns parameters w_i and σ_i to each input word x_i, to calculate a mask λ in the following way:

$$w_{i \to j} = w_i \cdot \exp\big(-\frac{d(i,j)^2}{\sigma_i}\big) \tag{1}$$

$$\lambda_j = \text{sigmoid}(\sum_i w_{i \to j}) \tag{2}$$

The weight w_i of a word x_i is mainly responsible for its mask value λ_i, but each weight w_i also influences the mask values of the words around it: $d(i,j)$ denotes the distance between two words while $w_{i \to j}$ denotes the influence of weight w_i towards λ_j. This mask value λ_j is simply calculated by applying the sigmoid to the sum of all influences onto this mask value. This parameterization of the mask, together with an objective function that encourages large values of σ, leads to smooth masks with long consecutive spans of texts being selected. Using this mask, two altered inputs are created:

$$\tilde{x} = \lambda \cdot x + (1 - \lambda) \cdot b \tag{3}$$

$$\tilde{x}^{\mathsf{c}} = (1 - \lambda) \cdot x + \lambda \cdot b \tag{4}$$

b is here an uninformative background (e.g., *PAD* tokens for BERT), meaning that $\tilde{x}$ is created by applying the mask to input x which removes low-scoring words from the input, while $\tilde{x}^{\mathsf{c}}$ applies the reverse mask. The actual objective function that is optimized is the following:

$$\underset{w,\sigma \in \mathbb{R}^n}{\arg\min} \quad -\mathcal{L}(\tilde{x}, c) + \mathcal{L}(\tilde{x}^{\mathsf{c}}, c) + \varOmega_\lambda + \varOmega_\sigma \tag{5}$$

where we optimize our mask to maximize our class probability of desired class c (given by $\mathcal{L}(\tilde{x}, c)$), meaning that we select words that indicate this class, while minimizing this likelihood for the reverse mask, meaning that words indicative of c will not be masked. The additional regularizers enforce sparsity ($\varOmega_\lambda$) and smoothness ($\varOmega_\sigma$) of the mask. For a more detailed description and derivation, see [4].

3 Datasets

3.1 The INAS Dataset

We evaluate our claim localization approach on the INAS dataset [3]. The dataset consists of 954 paper titles and abstracts from the field of invasion biology, a field concerned with the study of human-induced spread of species outside of their native ranges, caused by factors like global transport and trade. The samples are annotated with labels indicating which of the ten main hypotheses in the field are addressed in a given paper, in combination with an even more fine-grained categorization about specific sub-hypotheses addressed in them, based on a hypothesis network created by [14]. We perform our own annotation study and asked three experts in the field of invasion biology to annotate 750 samples with span-level evidence. The task was to annotate all spans that, to the trained eye, indicate which hypothesis is addressed in the given paper, even if the hypothesis is not explicitly named or stated.

50 samples were annotated by all annotators and we achieved a rather low F1 score of 0.389, indicating that this is a generally challenging annotation task even for domain experts. This is in part caused by one annotator having much lower agreement with the other two, indicating that annotation guidelines were interpreted slightly differently, which, for such a complex task, can quickly reduce agreement scores. The higher F1 score of 0.579 between the other two annotators shows that the general task is well-defined and thus suitable to be tackled by neural networks.

3.2 The SciFact Dataset

We also evaluate our approach on the SciFact dataset [35]. It consists of 5,183 abstracts from a collection of well-regarded journals, in combination with a set of 1,409 hand-written claims that are supported or rejected by papers from the corpus. The papers that verify or reject a claim are annotated on sentence-level with evidence for the respective classification, so that, in contrast to the span-level annotations for the INAS dataset, each sentence completely belongs to the evidence or not.

4 Method

4.1 Span-Level Claim Evidence Localization

We propose a method to perform weakly supervised span-level claim evidence localization. In this setting, we assume the availability of a training set of texts labeled with information indicating which claim (from a fixed set of known claims) is addressed in each of them. Given a text consisting of words $x_1, ..., x_n$, the task of weakly supervised claim localization is now to predict a set $I \subset \{1, ..., n\}$ of indices of words that are part of the ground truth claim evidence annotated by a human annotator. We propose to utilize the MaRC approach (see Sect. 2.2) to solve this task by first training a classifier to perform the claim identification task using only abstract-level labels, which is a standard text classification problem. Afterward, MaRC can be used to create an explanation for the classification of a given sample to produce importance scores for each word in the abstract.

For improved rationale predictions, we propose to perform the optimization from Eq. 5 with respect to several models, but for a single set of mask values. Input optimization is known to overly adapt to the particularities of a given model, which we hypothesize to be mitigated by optimizing with respect to multiple models at once.

4.2 Sentence-Level Claim Evidence Localization

We also propose an approach for sentence-level claim evidence localization. The precise task we consider slightly differs from the one described in the previous section, as here we assume claims to be present in textual form, and to not originate from a fixed set of known claims. Given a claim and an abstract, the task is to predict one of the three labels $\{$*Supports*, *Refutes*, *Not Enough Info*$\}$.

We again start by training a standard text classification model, which now receives the claim and abstract as inputs and predicts one of the three given labels as output. While it would be possible to employ the same procedure as described in Sect. 4.1 and compute sentence scores from the scores for the individual words, this could lead to uncertainties in the case of only very few words in a sentence being selected, as these could be highly important (thus making the whole sentence important) or simple artifacts caused by important words from a neighboring sentence exerting influence.

For this reason, we directly optimize mask weights $w_1, ..., w_n$, with one value being assigned to each input sentence $s_i \in \{s_1, ..., s_n\}$, and define $\lambda_i = \sigma(w_i)$ as the mask value for the sentence. We also alter the interpretation of the mask values λ: Before, each input embedding was linearly blended towards an uninformative embedding, as the input embedding $\tilde{x}_i$ of token i was defined to be $\tilde{x}_i = \lambda_i \cdot x_i + (1 - \lambda_i) \cdot b_i$. Despite good performance of this approach [4], these shifted embeddings constitute out-of-domain inputs as they are not encountered during training, therefore potentially leading to unpredictable behavior of the network. Therefore, we explore the possibility of treating λ as a set of probability distributions, with each λ_i being the parameter of a Bernoulli distribution indicating the probability of sentence s_i belonging to the input. This allows sampling of inputs from this distribution, with each sentence being either completely present or completely removed (replaced by *[PAD]* tokens) in a given sample. We then optimize this distribution to increase the likelihood of samples with high scores according to our objective, leading to the following optimization problem:

$$\underset{w \in \mathbb{R}^n}{\arg\min} \;\; \mathbb{E}_{m \sim \lambda} \left[-\mathcal{L}(\tilde{x}, c) + \mathcal{L}(\tilde{x}^{\mathsf{c}}, c) \right] + \Omega_\lambda \tag{6}$$

where $\tilde{x}$ and $\tilde{x}^{\mathsf{c}}$ are computed using the mask m sampled from λ similarly to Eq. 3 and Eq. 4, but on sentence-level. This equation can not be optimized using standard gradient-descent, as it contains an expectation over a probability distribution. We therefore use the score function estimator [9]:

$$\frac{\partial}{\partial \lambda} \mathbb{E}_{m \sim p(\cdot;\lambda)} \left[f(m) \right] = \mathbb{E}_{m \sim p(\cdot;\lambda)} \left[f(m) \frac{\partial}{\partial \lambda} \log p(m;\lambda) \right] \tag{7}$$

The expectation on the right side can now be approximated by sampling a batch of masks from λ, with $f(m)$ being our likelihood scores for mask m as defined in Eq. 6.

For our specific task, only the sentences from the abstract are masked, while the claim does not receive a mask value to be optimized. Again, we perform the optimization with respect to multiple trained classifiers as further regularization.

4.3 Evidence Injection

While our general methods aim at using weak supervision only, we also explore how far the results can be improved by using evidence annotations in the course of the base classifier training. To do this, we develop a method to inject evidence annotation information into the standard classifier training process. To our knowledge, something remotely similar has only been explored for the case of Support Vector Machines [38]. We test this method on the SciFact dataset and therefore assume the presence of sentence-level evidence annotations.

The altered training paradigm works as follows: Given a training sample x, this sample will be fed three times into the network (all in the same batch). Once in its normal form, once with all evidence sentences removed, and once with all evidence sentences present, but with some other sentences removed. We then train the model to predict the correct label (*Supports* or *Refutes*) for the first and third versions of the sample, but train it to predict the *Not Enough Info* label for the second version. In this way, the classifier learns to differentiate sentences that actually support the claim from sentences that only address the same topic.

5 Experiments

5.1 Span-Level Claim Localization

Experimental Setup. We perform experiments on weakly-supervised span-level evidence localization on the INAS dataset. Given a sample x consisting of words $x_1, ..., x_n$, the task is to predict a score s_i for every word x_i, such that the words belonging to the ground truth evidence annotated by a human annotator are assigned the highest scores. We perform our experiments in a weakly supervised setting, meaning that no method will have access to samples with actual evidence annotation. Instead, the supervision signal will solely be the label indicating which hypothesis (from a set of 10 possible hypotheses) is addressed in a given abstract. This information will be available during training and testing, as we explore the setting of localizing evidence for a claim that is known in advance.

Our proposed method works by training a standard text classification model to predict the correct hypothesis label for a given sample and to use the MaRC method to extract an explanation for the given label of interest post hoc (see Sect. 4.1 for a detailed description). We hypothesize that this method will outperform other interpretability methods, as it is explicitly designed to generate human-like rationales in the form of consecutive spans of text. As we are not aware of other methods for weakly supervised claim localization, we evaluate this method against other explainability methods (see Appendix A for an overview) as well as against a supervised baseline to allow for a relative performance comparison. For model and training details, see Appendix A.

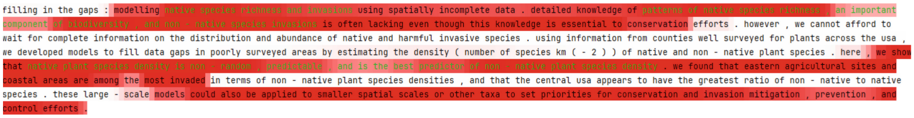
We additionally employ a post-processing step in our prediction pipeline: We split the abstract into individual sentences using ScispaCy [21] and set the predicted scores of the last token of each sentence to 0. This additional step improves span-matching performance, since claim evidence annotations in our particular task do not cross sentence boundaries and do not include punctuation.

Evaluation. We evaluate different measures for the quality of the predicted scores. To assess the quality of the scores assigned to the individual words (independent of their belonging to a longer span of text) we evaluate the area under the precision-recall-curve (*AUC-PR*).

We also evaluate the F1 score, which requires a binary prediction (i.e., each word is either predicted to belong to the evidence or not). Since many methods do not have an obvious way of determining a score threshold, we select the $p \cdot n$ highest-scoring words and average over 19 values of p (0.05, 0.10, 0.15, ..., 0.95).

The same technique is used for the *IoU-F1* score, which we propose as a measure for determining the quality of predicted spans of text. Given a binary prediction for each token, we determine predicted spans as continuous spans of words that were selected as evidence and calculate the IoU between all pairs of predicted and ground truth spans. As perfect matches are unlikely for this challenging task, we define generalized versions of precision and recall that allow for partial matches. To do so, we determine the highest IoU value of each span (ground truth and predicted) with anyone from the other set, and define the precision as the average of these highest values for the ground truth spans, which, analogous to the usual precision, is a measure for how well the ground truth spans have been recognized. Similarly, we define the recall as the average over the highest values for the predicted spans, thus measuring how likely a predicted span matches any of the ground truth spans. The F1 score is calculated from these values as usual and is again averaged over all values of p.

The three scores described so far are well-suited for comparing different methods with each other. To give a better feeling for the absolute quality of the predictions, we again use the F1 and IoU-F1 scores (now denoted as *D-F1* and *D-IoU-F1*), but for a single selection of words: We select a threshold t as the score of the k-th highest-scoring word, with k being the number of words in the ground truth evidence. As ground truth information is used, this is not an objec-



filling in the gaps : modelling native species richness and invasions using spatially incomplete data . detailed knowledge of patterns of native species richness , an important
component of biodiversity , and non - native species invasions is often lacking even though this knowledge is essential to conservation efforts . however , we cannot afford to
wait for complete information on the distribution and abundance of native and harmful invasive species . using information from counties well surveyed for plants across the usa ,
we developed models to fill data gaps in poorly surveyed areas by estimating the density (number of species km (- 2)) of native and non - native plant species . here , we show
that native plant species density is non - random , predictable , and is the best predictor of non - native plant species density . we found that eastern agricultural sites and
coastal areas are among the most invaded in terms of non - native plant species densities , and that the central usa appears to have the greatest ratio of non - native to native
species . these large - scale models could also be applied to smaller spatial scales or other taxa to set priorities for conservation and invasion mitigation , prevention , and
control efforts .

Fig. 1. Exemplary prediction of the MaRC method for an abstract from the INAS dataset for the *Biotic Resistance Hypothesis* label. Green text marks ground truth annotations, red spans indicate predicted scores.

tive measure of quality, but it nevertheless provides a more interpretable score. We additionally alter the IoU-F1 from the generalized, continuous version to a discrete one used in [6]: A ground truth span is counted as correctly recognized if any predicted span has an IoU of over 0.5, which allows for the calculation of standard precision and recall scores.

5.2 Sentence-Level Claim Localization

Experimental Setup. We perform experiments on weakly-supervised sentence-level evidence localization on the SciFact dataset, which is analogous to the task defined in Sect. 5.1, with the difference that each sentence receives only a single score. Since most explainability methods do not focus on complete sentences, we instead focus on testing different versions of the approach described in Sect. 4.2 and compare them to a supervised baseline, which is a RoBERTa-large classifier [18] that receives a textual claim and a sentence from the abstract and predicts the likelihood of this sentence belonging to the evidence.

We explore different versions of our approach, which differ in the way the base-classifier is trained: As a baseline, we test a classifier that is trained as usual on the SciFact dataset only. We also test a version that is trained with added spans of *PAD* tokens between sentences to align the input spaces present during training and optimization. We also explore the effect of pretraining on five other datasets (Fever [33], EvidenceInference [7,17], PubmedQA [15], HealthVer [26], COVIDFact [25]), which has been shown to improve the classifier performance [37]. Lastly, we also try a supervised version of our approach by employing the procedure described in Sect. 4.3 during classifier training. For more details on the training and evaluation, see Appendix A.

Evaluation. We again evaluate the AUC-PR as a holistic measure of the assigned ranking between the sentences. As for more interpretable measures, we provide the *precision@k* with $k \in \{1, 2, 3\}$, which is defined as the number of ground truth sentences correctly placed among the top-k scoring sentences by the classifier, divided by the maximum number possible (the minimum of the number of available ground truth sentences and k).

For all trained base classifiers, we also provide the F1 score of the abstract-level classification task (*Clf-F1*) to display the effect the different training paradigms have on the classifier performance.

6 Results

6.1 Span-Level Evidence Localization

The results for the span-level evidence localization are displayed in Table 1, while an exemplary output for the MaRC method is displayed in Fig. 1.

The MaRC method outperforms all other methods tested, both for scores measuring token-level performance (AUC-PR, F1, D-F1) as well as for scores evaluating span predictions (IoU-F1, D-IoU-F1). Especially with regards to the span predictions, we see that the MaRC approach significantly outperforms all other methods, which can be explained by it being explicitly designed to produce rationales that mirror human reasoning. The difference to other methods is here, that complete spans are selected as evidence, including words like “the”, “and”, etc., if they are directly part of an important span. Other methods, in comparison, mainly select the few rare words that are a more direct hint towards the hypothesis label, but do therefore not match human-annotated spans. This phenomenon also negatively affects token-level scores for other methods, since only few words per span are recognized as important. For the occlusion method, we produce a similar behavior by occluding longer spans of text at a time, leading to smoothly varying scores and thus to the only method that remotely rivals the MaRC method.

Notably, some methods barely outperform a random baseline (especially for span prediction evaluations), thus making them unusable for claim localization. As a possible explanation, [3] analyzed that classifiers for this task can make use of individual words like species names or locations as hints for the hypothesis if these names only occur in the context of this specific hypothesis. These will not be annotated by the human annotators, though, as hypothesis evidence (according to our definition) needs to clearly reference parts of the respective hypothesis. Overall, this shows a limitation of the proposed approach of using explainability methods for claim localization, as this approach relies on a high overlap between spans considered by humans as hypothesis evidence and words actually used by the classifier as the basis for the prediction, which is not always given.

As is to be expected, though, all methods are outperformed significantly by the supervised baseline. It is the only method that is explicitly trained to predict spans of the desired form, and the only method that has knowledge about the type of information that is to be selected. For weakly supervised methods, that do not have any of this information, predicting the precise span boundaries is extremely difficult. This result suggests, that for a smaller prediction space results could be improved, which we analyzed for the case of sentence-level evidence localization.

6.2 Sentence-Level Evidence Localization

The results for the sentence-level evidence localization are displayed in Table 2. Even though we altered the existing MaRC approach due to the differences between the tasks, our proposed method is still denoted as “MaRC”.

Table 1. Results for the span-level claim localization task on the INAS dataset.

Method	AUC-PR	F1	IoU-F1	D-F1	D-IoU-F1
MaRC	**0.357**	**0.331**	**0.210**	**0.350**	**0.151**
Occlusion	0.310	0.288	0.148	0.310	0.074
Saliency$_{L2}$	0.295	0.311	0.094	0.313	0.019
Saliency$_{Sum}$	0.241	0.265	0.070	0.259	0.002
InXGrad$_{L2}$	0.267	0.304	0.087	0.301	0.013
InXGrad$_{Sum}$	0.240	0.258	0.070	0.248	0.002
Int. Grads$_{L2}$	0.317	0.311	0.091	0.319	0.020
Int. Grads$_{Sum}$	0.320	0.305	0.090	0.322	0.017
LIME	0.271	0.281	0.072	0.273	0.004
Shapley	0.322	0.305	0.086	0.329	0.016
Random	0.221	0.256	0.067	0.223	0.003
Supervised	0.574	0.409	0.231	0.521	0.288

Table 2. Results for the sentence-level claim localization task on the SciFact dataset.

gt	pad	pre	sup	Method	Clf-F1	AUC-PR	Prec@1	Prec@2	Prec@3
X				MaRC	0.859	0.546	0.524	0.578	0.659
				MaRC	0.859	0.581	0.534	0.617	0.741
X	X			MaRC	0.842	0.632	0.612	0.675	0.710
	X			MaRC	0.842	0.655	0.641	0.689	0.736
X	X	X		MaRC	0.877	0.696	0.718	0.738	0.786
	X	X		MaRC	0.877	0.650	0.650	0.699	0.754
X	X	X	X	MaRC	0.936	0.720	0.757	0.772	0.780
	X	X	X	MaRC	0.936	0.718	0.757	0.777	0.780
			X	Sent-clf		0.882	0.883	0.893	0.905
		X	X	Sent-clf		0.902	0.883	0.898	0.951
		X		Sent-clf		0.664	0.650	0.655	0.778

The first four columns in Table 2 provide information about whether the model had access to the ground truth label during optimization (column *gt*), whether the base classifier was trained with added *PAD* tokens (column *pad*), whether the classifier was pretrained (column *pre*) and whether the classifier was trained using evidence supervision (column *sup*).

As, again, no previous study addressed our specific task of weakly supervised claim localization, and since none of the standard explainability methods tested on the INAS dataset proved particularly well-suited for the task at hand, we focus in this section on a comparison of our method with a supervised baseline, and analyze the challenges and solutions for mitigating the gap in performance.

Our most basic version of the MaRC approach (rows 1 and 2) uses a classifier trained without any changes to the standard training procedure. Even for this case, we already see reasonable performance, as it ranks an evidence sentence at the top in 52.4% of cases. Notably, if the optimization is done with respect to the ground truth label (row 1), the performance decreases compared to optimizing with respect to the predicted label (row 2), which is the case for the two lowest-performing base classifiers (up to row 4). This suggests, that without pretraining, the classifier is able to correctly identify the important sentences, but does not have the necessary capabilities to correctly infer the correct label from them.

Our second base classifier (rows 3 and 4) is trained in the same way as before, but receives samples with added *PAD* tokens during training, as these will be common during optimization, leading to otherwise misaligned input spaces. We see a significant improvement for the ground truth and the predicted label cases, so that we train all upcoming classifiers in this way. For this setting, only with access to the weak supervision labels on the SciFact dataset, the MaRC method manages to identify an evidence sentence as the most important sentence in 64.1% of cases, which we already consider quite good performance.

For our next classifier, we added additional pretraining on five similar datasets to the training procedure. This significantly improved the classifier performance and also led to improved results for evidence localization. Notably, from this point onward, having access to the ground truth label during optimization does improve evidence localization performance, indicating that pretraining increased the model's capability of inferring the correct label from the given sentences. Here, we also see the highest performance that we achieved using only weak supervision, with an evidence sentence being correctly identified as most important in 71.8% of cases.

Finally, we experiment with incorporating evidence supervision into the classifier training (as described in Sect. 4.3), to see how far the performance of our method can be pushed in a supervised setting.

At first, we note a significant improvement in the model's general classification performance, which even surpasses the improvement achieved by pretraining. This shows that the evidence injection strategy helped the model with actually understanding the rationale behind specific classifications, which seems to drastically boost the generalization performance.

On the other hand, we also see a significant improvement in the evidence localization results, which could be explained by the better general understanding of the model. We also hypothesize, that this is caused by the general setting of this task: Given an abstract and a claim, the model is supposed to predict one of three labels: *Supports*, *Refutes* or *Not Enough Info*. This means, that sentences that indicate that the general topic of the given abstract aligns with the given claim are considered important (even if they do not directly support or refute the claim) as they affect the likelihood of the *Not Enough Info* label. This leads to these sentences being selected by the MaRC approach as well, as it aims at maximizing the *Supports* or *Refutes* label. Our supervision approach mitigates this behavior, as it explicitly teaches the model to only take actual evidence sentences into account for the classification.

As is to be expected, the supervised baseline models with access to supervision on the SciFact dataset (rows 9 and 10) significantly outperform the weakly supervised models. For a more fair comparison, we also trained a supervised model only on the pretraining datasets and applied it to the SciFact dataset without any supervised training. In this case, the performance of the supervised classifier actually lags behind the MaRC approach in a similar setting (row 5), indicating that, if only abstract-level labels are present, the approach proposed in this work is a valid choice.

In summary, we managed to highlight several problems for our method, ranging from misaligned input spaces and insufficient understanding of the evidence sentences to the selection of non-evidence sentences due to the particular setup of the given task. Many problems can be mitigated by altering the training paradigm of the base classifier, but closing the gap to supervised models still proves to be a significant challenge.

7 Conclusion

In this work, we explored the possibility of using abstract-level labels about the general presence of a claim in this abstract to localize corresponding claim evidence. We proved that this is possible in both the span-level and sentence-level localization settings, but found that the complexity of precise span prediction makes achieving good performance challenging. For the sentence-level task, we found that weakly supervised methods can achieve reasonable performance and even be competitive in settings with only abstract-level labels available.

Since annotating a large number of samples with evidence annotations is very time-intensive and costly, we believe this to be an interesting direction for future research. Especially the fact that evidence supervision during classifier training can improve the performance of explainability methods on this task indicates, that creative changes to the training procedure of neural networks might lead to a substantial improvement of weakly supervised methods, which provides interesting possibilities for future research.

A Experimental Details

Model Details. We use PubMedBERT [12] and RoBERTa large [18] as the classification models for the INAS dataset and SciFact dataset, respectively. We train seven models, and keep the three best performing models with the highest validation F1 score.

The pretraining for the SciFact model is done on five datasets: Fever [33], EvidenceInference [7,17], PubmedQA [15], HealthVer [26], COVIDFact [25]. **MaRC Details.** The optimization for the MaRC method is done with respect to all three trained models. The parameters are set as described in [4], but we employ a new sparsity regularizer that actively forces a maximum average mask value. Similar to [4], we use the following weight regularizer:

$$\Omega_\lambda = \alpha_\lambda \left[\frac{1}{n} \sum_{i=1}^{n} \lambda_i \right]^2$$

but dynamically update α_λ at each gradient descent step i using to following formulas, to reach a maximum average mask value t (set to 0.35):

$$m_i = \frac{1}{n} \sum_{i=1}^{n} \lambda_i$$

$$\Delta_i = m_{i-1} - m_i$$

$$\Delta_{target} = (m_i - t)/150$$

Here, m_i is the current mask mean, Δ_i is the difference in mask means from the current optimization step to the last, and Δ_{target} is the desired value for Δ_i, which (if it is always optimal) ensures a steady but decelerating trajectory towards the optimal mask value. We define

$$\Delta_{\Delta_i,target} = \Delta_i - \Delta_{target}$$

to be the difference between our current single-step mask mean difference and the desired one, which we want to bring as close to 0 as possible. We then define our update for α_λ at iteration i as follows:

$$\alpha_\lambda^i = \alpha_\lambda^{i-1} \cdot (0.8 + 0.2 \cdot \gamma)$$

$$\gamma = \max\left(0.7, 1 - 0.9 \cdot \tanh\left(\frac{1}{0.002}\left(\frac{\Delta_{\Delta_i,target}}{2} - (\Delta_{i-1} - \Delta_i)\right)\right)\right)$$

so that $\gamma > 1$ leads to an increase in α_λ whereas $\gamma < 1$ leads to a decrease. The max operator prevents an overly steep decrease of α_λ, while the tanh is used to keep positive updates limited. The updates are mainly determined by $\Delta_{\Delta_i,target}$, so that α_λ increases when Δ_i is smaller than Δ_{target} and vice versa. The term $(\Delta_{i-1} - \Delta_i)$ is a second-order statistic to prevent "overshooting" in the form of changing α_λ further if Δ_i is already approaching Δ_{target} (which might take a while due to the momentum-based optimizer).

To give the optimization process the freedom to determine the optimal average mask value on its own after falling below $t + 0.1$, we alter the process of determining α_λ in the following way:

$$\alpha_\lambda^i = \alpha_\lambda^{i-1} \cdot (0.8 + 0.2 \cdot \gamma_{pred} \cdot \gamma_{weight})$$

$$\gamma_{pred} = \min\left(\frac{\mathcal{L}(\tilde{x}, c)_i}{\mathcal{L}(x, c)_0}, 0.5 \cdot \frac{\mathcal{L}(x, c)_0}{\mathcal{L}(\tilde{x}^{\mathsf{c}}, c)_i}, 1.1\right)$$

$$\gamma_{weight} = 1 + (m_i - 0.3)$$

Here, γ_{pred} pushes mask values further down if the model prediction for the current masked input is more confident than the initial unmasked prediction and the prediction for complement mask input is sufficiently less confident than the initial unmasked prediction, which indicates that more information can be removed. γ_{weight} pushes the average mask value to a value of 0.3, since values far below that lead to most words having scores close to 0, and thus to no clear ranking existing among them.

Comparison Methods. The other explainability methods are all used for each of the three models individually, and the scores are averaged afterward. We make use of the following methods and hyperparameter settings:

- *Occlusion* [39]: We chose to mask slightly larger spans of 5 tokens as this produced smoother masks which resulted in higher IoU F1 scores. Occluded parts were replaced by *PAD*-tokens.
- *Saliency* [28]: No hyperparameter settings required.
- *InXGrad* (Input times gradient [27]): No hyperparameter settings required.
- *Int. Grads* (Integrated Gradients [30]): We use a sequence of *PAD*-tokens as background and do 50 gradient evaluation steps per sample.
- *LIME* [24]: We do 50 function evaluations per sample. In each evaluation, we randomly select $5 - 13\%$ of tokens and replace them as well as the next three tokens with *PAD*-tokens. We train a linear classifier and use the resulting weights as rationale.
- *Shapley* (Shapley value sampling [5]): We evaluate the token contributions for 15 feature permutations per sample. Removed tokens are replaced by *PAD*-tokens.

We use the implementations provided by [16] for all methods. All methods have access to the ground truth label. The InXGrad, Saliency and Int. Grads methods all predict one score for each element of the embedding vector of a given word, which is reduced to a single score by using the L2-norm or the sum.

We also compare against a supervised baseline. It is trained on 517 samples from the INAS dataset annotated with span-level evidence, as well as on 204 samples without annotated evidence. To make use of the samples without evidence annotations we train in a multi-task setting by also training to predict the general hypothesis labels for the whole abstract.

INAS Evaluation. We evaluate all methods on a test set consisting of 141 samples that cover all ten possible classes. The test set contains all 50 samples that were annotated by all three annotators, as well as 91 further samples that were annotated by only one of the three annotators, with samples and annotators being assigned randomly. For the samples that were annotated by all annotators, we create a single ground truth by taking the intersection of the set of annotated tokens between each pair of annotators, followed by the union between the three resulting annotations of each pair.

SciFact Evaluation. As the test labels for the SciFact dataset are not publicly available, we create new splits with 50 claims for validation, 150 claims for testing and the remaining claims for training. The actual samples for the splits can then be created from the given claims and linked documents.

For evaluating the *AUC-PR* and *Precision@k* scores, we only take samples from the *Supports* and *Refutes* classes into account, as they are the only classes with corresponding evidence annotations.

References

1. Accuosto, P., Neves, M.L., Saggion, H.: Argumentation mining in scientific literature: from computational linguistics to biomedicine. In: BIR@ECIR (2021)
2. Blake, C.: Beyond genes, proteins, and abstracts: identifying scientific claims from full-text biomedical articles. J. Biomed. Inform. **43**(2), 173–189 (2010). https://doi.org/10.1016/j.jbi.2009.11.001
3. Brinner, M., Heger, T., Zarriess, S.: Linking a hypothesis network from the domain of invasion biology to a corpus of scientific abstracts: the INAS dataset. In: Proceedings of the first Workshop on Information Extraction from Scientific Publications, pp. 32–42. Association for Computational Linguistics (2022)
4. Brinner, M., Zarrieß, S.: Model interpretability and rationale extraction by input mask optimization. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 13722–13744. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.findings-acl.867
5. Castro, J., Gómez, D., Tejada, J.: Polynomial calculation of the Shapley value based on sampling. Comput. Oper. Res. **36**(5), 1726–1730 (2009). https://doi.org/10.1016/j.cor.2008.04.004
6. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: ERASER: a benchmark to evaluate rationalized NLP models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4443–4458. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.408
7. DeYoung, J., Lehman, E.P., Nye, B.E., Marshall, I.J., Wallace, B.C.: Evidence inference 2.0: more data, better models. arXiv abs/2005.04177 (2020)
8. Freeman, J.B.: Dialectics and the macrostructure of arguments. De Gruyter Mouton (1991). https://doi.org/10.1515/9783110875843
9. Fu, M.C.: Chapter 19 gradient estimation. In: Simulation, Handbooks in Operations Research and Management Science, vol. 13, pp. 575–616. Elsevier (2006). https://doi.org/10.1016/S0927-0507(06)13019-4
10. Green, N.: Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In: Proceedings of the First Workshop on Argumentation Mining, pp. 11–18. Association for Computational Linguistics (2014). https://doi.org/10.3115/v1/W14-2102
11. Green, N.: Identifying argumentation schemes in genetics research articles. In: Proceedings of the 2nd Workshop on Argumentation Mining, pp. 12–21. Association for Computational Linguistics (2015). https://doi.org/10.3115/v1/W15-0502
12. Gu, Y., et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthc. **3**(1), 1–23 (2022). https://doi.org/10.1145/3458754, arXiv:2007.15779 [cs]

13. Jansen, T., Kuhn, T.: Extracting core claims from scientific articles. In: Bosse, T., Bredeweg, B. (eds.) BNAIC 2016. CCIS, vol. 765, pp. 32–46. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67468-1_3
14. Jeschke, J.M., Heger, T.: Invasion biology: hypotheses and evidence (2018). https://doi.org/10.1079/9781780647647.0000
15. Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X.: PubMedQA: a dataset for biomedical research question answering. 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), pp. 2567–2577. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1259
16. Kokhlikyan, N., et al.: Captum: a unified and generic model interpretability library for PyTorch (2020)
17. Lehman, E., DeYoung, J., Barzilay, R., Wallace, B.C.: Inferring which medical treatments work from reports of clinical trials. 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), pp. 3705–3717. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1371
18. Liu, Y., et al..: RoBERTa: a robustly optimized BERT pretraining approach (2019)
19. Lloyd, E.A.: Confirmation of ecological and evolutionary models. Biol. Philos. **2**(3), 277–293 (1987). https://doi.org/10.1007/BF00128834
20. Madsen, A., Reddy, S., Chandar, S.: Post-hoc interpretability for neural NLP: a survey. ACM Comput. Surv. **55**(8) (2022). https://doi.org/10.1145/3546577
21. Neumann, M., King, D., Beltagy, I., Ammar, W.: ScispaCy: fast and robust models for biomedical natural language processing. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 319–327. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/W19-5034
22. Peldszus, A., Stede, M.: From argument diagrams to argumentation mining in texts: a survey. Int. J. Cogn. Inform. Nat. Intell. **7**(1), 1–31 (2013). https://doi.org/10.4018/jcini.2013010101
23. Pradeep, R., Ma, X., Nogueira, R., Lin, J.: Scientific claim verification with VerT5erini. In: Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pp. 94–103. Association for Computational Linguistics (2021)
24. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 1135–1144. Association for Computing Machinery (2016). https://doi.org/10.1145/2939672.2939778
25. Saakyan, A., Chakrabarty, T., Muresan, S.: COVID-fact: fact extraction and verification of real-world claims on COVID-19 pandemic. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 2116–2129. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.165
26. Sarrouti, M., Ben Abacha, A., Mrabet, Y., Demner-Fushman, D.: Evidence-based fact-checking of health-related claims. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 3499–3512. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.findings-emnlp.297

27. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3145–3153. PMLR (2017)
28. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. CoRR (2013)
29. Stab, C., Gurevych, I.: Parsing argumentation structures in persuasive essays. Comput. Linguist. **43**(3), 619–659 (2017). https://doi.org/10.1162/COLI_a_00295
30. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. ICML'17, pp. 3319–3328. JMLR.org (2017)
31. Teufel, S., Carletta, J., Moens, M.: An annotation scheme for discourse-level argumentation in research articles. In: Ninth Conference of the European Chapter of the Association for Computational Linguistics, pp. 110–117. Association for Computational Linguistics (1999)
32. Teufel, S., Siddharthan, A., Batchelor, C.: Towards domain-independent argumentative zoning: evidence from chemistry and computational linguistics. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pp. 1493–1502. Association for Computational Linguistics (2009)
33. Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: 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, pp. 809–819. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/N18-1074
34. Toulmin, S.: The Uses of Argument. Cambridge University Press, Cambridge (1958)
35. Wadden, D., et al.: Fact or fiction: verifying scientific claims. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 7534–7550. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.609
36. Wadden, D., et al.: SciFact-open: towards open-domain scientific claim verification. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 4719–4734. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.findings-emnlp.347
37. Wadden, D., Lo, K., Wang, L.L., Cohan, A., Beltagy, I., Hajishirzi, H.: MultiVerS: improving scientific claim verification with weak supervision and full-document context. In: Findings of the Association for Computational Linguistics: NAACL 2022, pp. 61–76. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.findings-naacl.6
38. Zaidan, O., Eisner, J., Piatko, C.: Using "annotator rationales" to improve machine learning for text categorization. In: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pp. 260–267. Association for Computational Linguistics (2007)
39. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53

40. Zeng, X., Abumansour, A.S., Zubiaga, A.: Automated fact-checking: a survey. Lang. Linguist. Compass **15**(10), e12438 (2021). https://doi.org/10.1111/lnc3.12438
41. Zhang, Z., Li, J., Fukumoto, F., Ye, Y.: Abstract, rationale, stance: a joint model for scientific claim verification. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3580–3586. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.emnlp-main.290