Extracting Fine-Grained Knowledge Graphs of Scientific Claims: Dataset and Transformer-Based Results

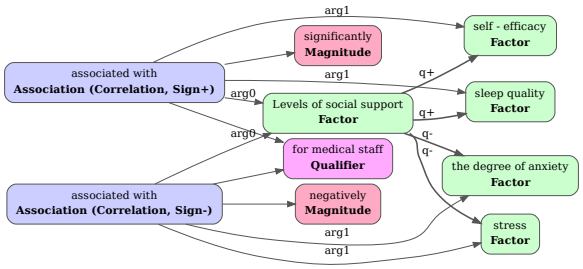
Ian H. Magnusson[♠♡] and **Scott E. Friedman**[♠]
[♠]SIFT, Minneapolis, MN, USA
[♡]Northeastern University, Boston, MA, USA
magnusson.i@northeastern.edu
friedman@sift.net

Abstract

Recent transformer-based approaches demonstrate promising results on relational scientific information extraction. Existing datasets focus on high-level description of how research is carried out. Instead we focus on the subtleties of how experimental associations are presented by building SciClaim, a dataset of scientific claims drawn from Social and Behavior Science (SBS), PubMed, and CORD-19 papers. Our novel graph annotation schema incorporates not only coarse-grained entity spans as nodes and relations as edges between them, but also fine-grained attributes that modify entities and their relations, for a total of 12,738 labels in the corpus. By including more label types and more than twice the label density of previous datasets, SciClaim captures causal, comparative, predictive, statistical, and proportional associations over experimental variables along with their qualifications, subtypes, and evidence. We extend work in transformer-based joint entity and relation extraction to effectively infer our schema, showing the promise of fine-grained knowledge graphs in scientific claims and beyond.

1 Introduction

Using relations as edges to connect nodes consisting of extracted entity mention spans produces expressive and unambiguous knowledge graphs from unstructured text. This approach has been applied to diverse domains from moral reasoning in social media (Friedman et al., 2021b) to qualitative structure in ethnographic texts (Friedman et al., 2021a), and is particularly useful for reasoning about scientific claims, where several experimental variables in a sentence may have differing relations. Scientific information extraction datasets such as SciERC (Luan et al., 2018) use relations for labeling general scientific language. Utilizing the advances of SciBERT (Beltagy et al., 2019) in scientific language modeling, SpERT (Eberts and Ulges, 2020)—a transformer-based joint entity and relation extraction model—advanced the state of the art on SciERC.



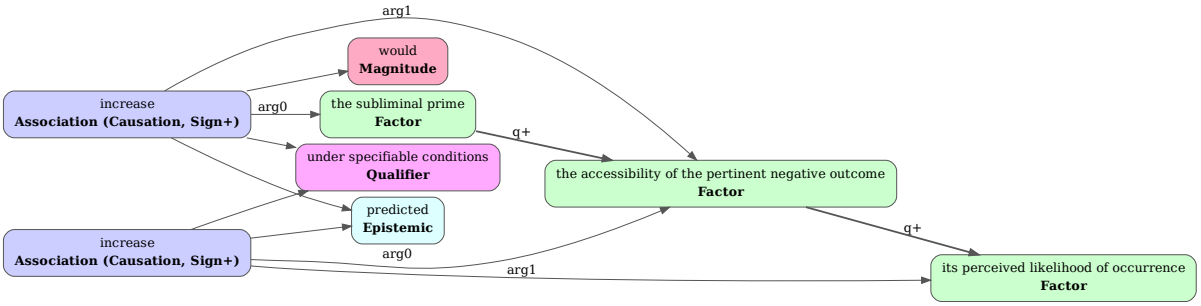
Input: *"Levels of social support for medical staff were significantly associated with self-efficacy and sleep quality and negatively associated with the degree of anxiety and stress."*

Figure 1: SciClaim knowledge graph with entities (nodes), relations (edges), and attributes (parentheticals) connecting an independent variable via *arg0* to distinct correlations with dependent variables via *arg1*.

To extend relational scientific information extraction to specifically target scientific claims, we annotate SciClaim,[1] a dataset of 12,738 annotations on 901 sentences from expert identified claims in Social and Behavior Science (SBS) papers (Alipourfard et al., 2021), detected causal language in PubMed papers (Yu et al., 2019), and claims and causal language heuristically identified from CORD-19 abstracts (Wang et al., 2020).

For annotation, we developed a novel graph schema that reifies claimed associations as entity spans with fine-grained attributes and extracts factors as additional entities connected with relations to one or more associations in which they are involved. In Figure 1, two association entities relate two pairs of dependent factors to an independent factor, while attributes and additional relations delimit the scope and qualitative proportionalities of the claim. Inspired by semantic role labeling, attributes modify associations and the roles of their arguments, allowing us to represent claims of causal, comparative, predictive, statistical, and proportional associations along with their qualifi-

[1]Dataset available at https://github.com/siftech/SciClaim.



Input: *"We predicted that the subliminal prime would, under specifiable conditions, increase the accessibility of the pertinent negative outcome and thereby increase its perceived likelihood of occurrence."*

Figure 2: This SciClaim graph captures the chaining together of associations and uncovers a mediating factor in the qualitative proportionality ($q+$) between the "subliminal prime" and "perceived likelihood of occurrence."

cations, subtypes, and evidence.

We adapt SpERT to model this additional multi-label attribute task and demonstrate that extraction of our highly expressive knowledge graphs is within reach of present methods.

2 Related Work

Many previous datasets for relational scientific information extraction—such as SemEval 2017 task 10 and 2018 task 7, SciERC, and SciREX (Augenstein et al., 2017; Gábor et al., 2018; Luan et al., 2018; Jain et al., 2020)—have annotated corpora from NLP, computer science, or similar engineering-oriented fields. As such their annotation schemas have emphasized the description of how research was carried out, by extracting categories of entities such as methods, tasks, metrics, and datasets as well as relations mostly describing their intrinsic properties such as uses, composition, and hyponymy. Two of these datasets (Luan et al., 2018; Gábor et al., 2018) contain associative relations that directly link entities being compared or producing a result. Our work extends further in this direction by examining not only which entities are associated, but also how the presentation of the associations is nuanced by the assertion of fine-grained attributes such as causality or proportionality.

SciClaim provides the largest number of fine-grained label types among comparable datasets. Table 1 shows SciClaim's remarkable label densities per word. SciClaim also contains 81.88% as many total labels as SciERC and more total labels than SemEval 2017 task 10 and 2018 task 7. On the other hand, SciREX utilizes distant supervision from an existing knowledge base and noisy automatic labeling trained on SciERC to provide an order of magnitude more labels and annotate complete documents. This is one example of how smaller yet more densely and directly labeled datasets like SciERC and SciClaim can enable and compliment larger, higher-level corpora.

Meanwhile, our dataset also focuses on scientific claims. Some previous work *identifies* claims within scientific texts (Wadden et al., 2020; Gelman et al., 2021), but does not extract the relations and factors within the claims themselves. Other recent symbolic semantic NLP systems do model relational representations of scientific claims (e.g., Friedman et al., 2017), but these approaches rely on rule-based engines with hand tuning, which require NLP experts to maintain and adapt to new domains. Instead we modify SpERT (Eberts and Ulges, 2020), a transformer-based method that has been shown to effectively extract relational scientific information on SciERC (Luan et al., 2018). We extend this model to accommodate our additional multi-label attributes and apply it to our claim graph extraction task.

3 Approach

3.1 Problem Definitions

SciClaim defines the multi-attribute knowledge graph extraction task as follows: for a sentence $\mathcal{S}$ of n tokens $s_1, ..., s_n$, and sets of entity types $\mathcal{T}_e$, attribute types $\mathcal{T}_a$, and relation types $\mathcal{T}_r$, predict the set of entities $\langle s_j, s_k, t \in \mathcal{T}_e \rangle \in \mathcal{E}$ ranging from tokens s_j to s_k, where $1 \leq j \leq k \leq n$, the set of relations over entities $\langle e_{head} \in \mathcal{E}, e_{tail} \in \mathcal{E}, t \in \mathcal{T}_r \rangle \in \mathcal{R}$ where $e_{head} \neq e_{tail}$, and the set of attributes over entities $\langle e \in \mathcal{E}, t \in \mathcal{T}_a \rangle \in \mathcal{A}$. This defines a directed multi-graph without self-cycles, where each unique span can be represented by at

Dataset	Words	Entities		Relations		Attributes/Corefs		Total Labels	
		Count	Per Word	Count	Per Word	Count	Per Word	Count	Per Word
SciREX	**2512806**	**157680**	6.27%	**9198**	0.37%	-	-	**166878**	06.64%
SemEval2017	84010	9946	11.84%	672	0.79%	-	-	10618	12.64%
SemEval2018	58144	7483	12.87%	1595	2.74%	-	-	9078	15.61%
SciERC	65334	8089	12.38%	4716	7.21%	**2752**	4.21%	15557	23.81%
SciClaim	20070	5548	**27.64%**	5346	**26.64%**	1844	**9.19%**	12738	**63.47%**

Table 1: Our SciClaim dataset contains the highest label densities per word and comparable label counts to other scientific information extraction datasets except SciREX, which uses distant supervision and noisy automatic labeling. Our dataset contains fine-grained attributes as additional labels, while SciERC contains coreference links.

most one entity node with zero to $|\mathcal{T}_a|$ attributes.

3.2 Dataset Construction

To create SciClaim, two NLP researchers annotated 901 sentences from several sources: 336 from papers in Social and Behavior Science (SBS) with expert identified claims (Alipourfard et al., 2021), 411 filtered for causal language in PubMed papers (Yu et al., 2019), 135 containing claims and causal language identified from CORD-19 abstracts (Wang et al., 2020) with heuristic keywords, and 19 manual perturbations included only in training data.

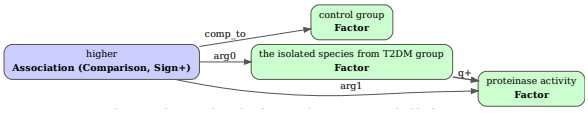
To aid in the labeling of these densely annotated sentences, we iteratively trained on already collected data and utilized the predictions of the partially trained model on new training examples as suggestions in our labeling interface. We disabled these model suggestions on our 100 example test set to ensure that this did not bias our evaluation.

Due to the dense and potentially overlapping span annotations, small decisions about what tokens to include in a span frequently influence the span boundaries of several other entities in a sentence. However, most of these decisions have negligible impact on the meaningfulness of the annotation (e.g. the decision to include a determiner in span), rendering exact match agreement ineffective. Instead to promote consistency and domain relevance we employed iterative schema design sessions in consultation with a subject matter expert in reproducibility of SBS experiments and a process of consensus, schema re-development, and re-annotation on 250 examples where annotators overlapped.

Table 1 contrasts SciClaim's label counts and density with the other relational scientific information extraction datasets discussed in Section 2, and precise counts for each label type are provided in Table 3. Further details are in Appendix A.

3.3 Graph Schema

The SciClaim graph schema is designed to capture associations between factors (e.g., causation, comparison, prediction, statistics, proportionality), monotonicity constraints across factors, epistemic status, subtypes, and high-level qualifiers.



Input: *"Compared to control group, the isolated species from T2DM group had higher proteinase activity."*

Figure 3: *Comparison* attributes modify arguments to account for a (sometimes implicit) frame of reference.

Entities are labeled text spans. The same exact span cannot correspond to more than one entity type, but two entity spans can overlap. Entities comprise the nodes of SciClaim graphs upon which attributes and relations are asserted. Our schema includes six entity types: **Factors** are variables that are tested or asserted within a claim (e.g., "*sleep quality*" in Figure 1). **Associations** are explicit phrases associating one or more factors (e.g., "*higher*" Figure 3). **Magnitudes** are modifiers of an association indicating its likelihood, strength, or direction (e.g., "*significantly*" and "*negatively*" in Figure 1). **Evidence** is an explicit mention of a study, theory, or methodology supporting an association (e.g., "*our SEIR model*"). **Epistemics** express the belief status of an association, often indicating whether something is hypothesized, assumed, or observed (e.g., "*predicted*" in Figure 2). **Qualifiers** constrain the applicability or scope of an assertion (e.g., "*for medical staff*" in Figure 1).

Attributes are multi-label fine-grained annotations (visualized in parentheses), where zero or more may apply to any given entity. Our schema includes the following attributes, all of which apply solely to *Association* entities: **Causation** ex-

presses cause-and-effect over its constituent factors (e.g., both "*increase*" spans in Figure 2). **Correlation** expresses interdependence over its constituent factors (e.g., both "*associated with*" spans in Figure 1). **Comparison** expresses an association with a frame of reference (as in the "*higher*" statement of Figure 3). **Sign+** and **Sign-** expresses high/low or increased/decreased factor value (e.g., "*correlates more closely with*" or "*shortened*" respectively). **Test** expresses statistical measurements (e.g., "*ANOVA*"). **Indicates** expresses a predictive relationship (e.g., "*prognostic factors for*").

Relations are directed edges between labeled entities in SciClaim graphs. They are critical for expressing what-goes-with-what over the set of entities. Note that in Figures 1 and 2 the unlabeled arrows are all *modifier* relations, left blank to avoid clutter. We encode six relations: **arg0** relates an association to its cause, antecedent, subject, or independent variable (e.g., "*levels of social support*" in Figure 1). **arg1** relates an association to its result or dependent variable (e.g., "*self-efficacy*" and "*stress*" in Figure 1). **comp_to** is an explicit frame of reference in a comparative association (e.g., "*control group*" in Figure 3). **subtype** relates a head entity to a subtype tail (e.g., "*stillbirth*" as a subtype of "*pregnancy outcome*"). **modifier** relates associations to qualifiers, magnitudes, epistemics, and evidence (e.g., all unlabeled arrows in Figure 1 and Figure 2). **q+** and **q-** indicate positive and negative qualitative proportionality, respectively, where increasing the head factor increases or decreases the tail factor, respectively (e.g., "*levels of social support*" is *q+* to "*sleep quality*" and *q-* to "*stress*" in Figure 1).

3.4 Model Architecture

In order to model the additional multi-label task in SciClaim, we extend SpERT (Eberts and Ulges, 2020) with an attribute classifier. SpERT provides components (Figure 4 a–c) for joint entity and relation extraction and permits the overlapping spans in our data. These classifiers utilize a span representation that combines the SciBERT (Beltagy et al., 2019) contextual embeddings of all tokens in the span through maxpooling, along with a context representation and learned width embedding. SpERT classifies entities first and only infers relations on pairs of identified entities.

Instead of maxpool we adopt an attention-based span representation (Figure 4 e) inspired by Lee et al. (2017). This produces scalars $\alpha_{i,t}$ for each SciBERT token vector $\mathbf{h}_t$ in a span i using learned parameters $\mathbf{w}$ and b:

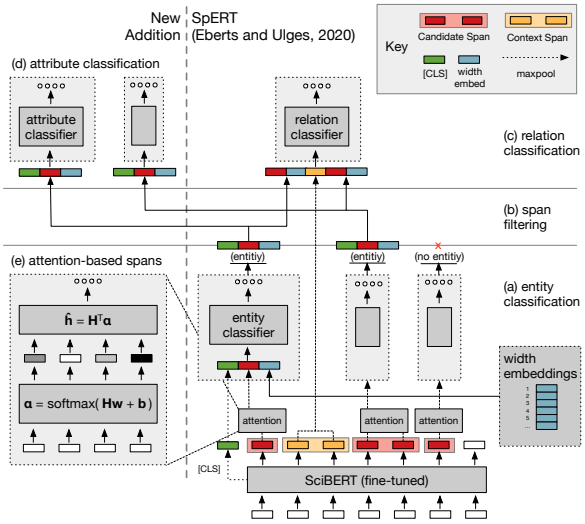


Figure 4: Our extension of SpERT components (a, b, and c) with multi-label attributes (d) and attention-based entity span representations (e).

$$\alpha_{i,t} = \frac{\exp(\mathbf{w} \cdot \mathbf{h}_t + b)}{\sum_{k=START(1)}^{END(i)} \exp(\mathbf{w} \cdot \mathbf{h}_k + b)} \quad (1)$$

These attention weights are used to make a span representation $\hat{\mathbf{h}}_i$ with the following weighted sum:

$$\hat{\mathbf{h}}_i = \sum_{t=START(1)}^{END(i)} \alpha_{i,t} \mathbf{h}_t \quad (2)$$

We use the same cascaded inference strategy and input the span representations of identified entities $\mathbf{x}^a$ to an attribute classifier (Figure 4 d) with weights $\mathbf{W}^a$ and bias $\mathbf{b}^a$. A pointwise sigmoid σ yields seperate confidence scores $\hat{\mathbf{y}}^a$ for each attribute:

$$\hat{\mathbf{y}}^a = \sigma(\mathbf{W}^a \mathbf{x}^a + \mathbf{b}^a) \quad (3)$$

We train the attribute classifier with a binary cross entropy loss $\mathcal{L}_a$ summed with the SpERT entity and relation losses, $\mathcal{L}_e$ and $\mathcal{L}_r$, for a joint loss:

$$\mathcal{L} = \mathcal{L}_e + \mathcal{L}_r + \mathcal{L}_a \quad (4)$$

4 Evaluation

In Table 2 we report micro performance metrics on the SciClaim test set averaged over 5 runs.

In addition to the **modified** SpERT (detailed in Section 3.4), we also test a variant **attrs-as-ents**

Data	Model	Entities			Attributes			Relations		
		P	R	F1	P	R	F1	P	R	F1
SciERC	SpERT	**70.87**	**69.79**	**70.33**	-	-	-	**53.40**	**48.54**	**50.84**
SciClaim	SpERT-attrs-as-ents	**90.13**	**88.63**	**89.37**	**92.35**	82.13	86.94	**77.59**	**74.34**	**75.92**
SciClaim	SpERT-modified	89.81	87.87	88.83	91.89	**82.62**	**87.01**	76.43	73.72	75.05

Table 2: Micro **P**recision, **R**ecall, and **F1** averaged over 5 runs on SciClaim with SciERC for comparison.

	Label	P	R	F1	S
Entities	factor	91.28	89.97	90.62	2756
	evidence	88.80	93.33	90.96	230
	epistemic	91.21	72.17	80.52	299
	association	92.45	88.20	90.27	1290
	magnitude	87.71	88.38	88.02	613
	qualifier	75.86	78.33	77.02	360
Attributes	causation	38.15	60.00	46.20	342
	comparison	86.69	80.00	83.19	329
	indicates	84.79	76.25	80.20	84
	sign+	97.27	88.31	92.58	542
	sign-	91.97	72.86	81.28	202
	correlation	98.42	84.41	90.88	320
Relations	arg0	79.53	75.03	77.19	1325
	arg1	79.92	77.57	78.71	1384
	comp_to	65.86	60.00	62.78	187
	modifier	77.71	76.21	76.95	1582
	subtype	40.00	33.33	36.00	156
	q+	65.53	67.61	66.50	504
	q-	70.70	53.00	60.09	208

Table 3: High **P**recision, **R**ecall, and **F1** across most types relative to total **S**upport in SciClaim, using SpERT-modified averaged over 5 runs.

where all attribute labels on an entity span are collapsed into a single combined annotation, allowing unmodified SpERT to process attributes. Precisely, we collapse $\mathcal{T}_e$ entity types with all combinations of $\mathcal{T}_a$ attribute types into $\{\mathcal{T}_e \times \binom{\mathcal{T}_a}{k} : 0 \leq k \leq |\mathcal{T}_a|\}$ multi-class entity labels. We hypothesized that the combinatorially larger number of labels required by *attrs-as-ents* would lower performance on rarely occurring combinations. Surprisingly the variants get almost identical results, suggesting that—at least for our data—a single layer classifier can infer the attributes of a span simultaneously just as well as doing so independently. We tested other model variants that also produced changes $\sim$1% F1 and thus are relegated to Appendix B.

To our knowledge no previous models exists that can run directly on all three tasks in our dataset due to the presence of both overlapped entity spans and multi-label attributes. For comparison we include SpERT's state-of-the-art performance on SciERC, the dataset closest to ours in terms of label density. The high performance of our adapted SpERT on SciClaim demonstrates the practicality of extracting our novel graph schema with present methods despite its fine-grained approach.

The per-class evaluations for our main model are reported in Table 3. With few exceptions performance is good, and generally follows support for the label in the dataset. The *Causation* attribute metrics may be influenced by noise from anomalously low representation in the test set (only 5 instances compared to 59 instances of *Correlation*). Likewise the *Test* attribute unfortunately does not appear in the test set at all, but receives validation F1 of 95.95% despite only appearing 25 times in the corpus. Another outlier, the *subtype* relation, is particularly challenging, especially with its low rate of occurrence, due to it being one of the few relation types occurring directly between factors rather than mediated through a reified association span. The *q+*/*q-* relations are likewise expressed as direct links between factors. Although these require complex reasoning about the qualitative proportionalities of factors (e.g., Figure 2), they nonetheless receive promising results. The attributes *Sign+*/*Sign-* serve a similar role and provide partial redundancy for *q+*/*q-* labels, allowing downstream reasoning to back off to these less precise, more robust attributes when the qualitative proportionalities are not extracted.

5 Conclusion

Previous scientific information extraction crafts useful high-level representation of papers, going as far as document level relations spanning thousands of words in Jain et al. (2020). Complimentary to these efforts, we propose fine-grained and densely annotated scientific information extraction that captures not just what is said but how it is presented and argued. SciClaim applies this approach to associative claims and demonstrates that existing models such as SpERT (Eberts and Ulges, 2020) can be modified to accurately extract fine-grained knowledge graphs ripe for downstream reasoning.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO) under Contract No. W911NF-20-C-0002. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO). We thank the reviewers for their helpful feedback.

References

Nazanin Alipourfard, Beatrix Arendt, Daniel M Benjamin, Noam Benkler, Michael M Bishop, Mark Burstein, Martin Bush, James Caverlee, Yiling Chen, Chae Clark, and et al. 2021. Systematizing confidence in open research and evidence (score).

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. *24th European Conference on Artificial Intelligence*.

Scott Friedman, Mark Burstein, David McDonald, Alex Plotnick, Laurel Bobrow, Rusty Bobrow, Brent Cochran, and J Pustejovsky. 2017. Learning by reading: Extending and localizing against a model. *Advances in Cognitive Systems*, 5:77–96.

Scott E. Friedman, Ian H. Magnusson, and Sonja M. Schmer-Galunder. 2021a. Extracting qualitative causal structure with transformer-based nlp. In *QR2021 @ IJCAI*.

Scott E. Friedman, Ian H. Magnusson, Sonja M. Schmer-Galunder, Ruta Wheelock, Jeremy Gottlieb, Pooja Patel, and Christopher Miller. 2021b. Toward Transformer-Based NLP for Extracting Psychosocial Indicators of Moral Disengagement. In *Annual Meeting of the Cognitive Science Community (CogSci)*.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Ben Gelman, Chae Clark, Scott Friedman, Ugur Kuter, and James Gentile. 2021. Toward a robust method for understanding the reproducibility and replicability of research. *AAAI Workshop on Scientific Document Understanding*.

Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A challenge dataset for document-level information extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7506–7516, Online. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. Cord-19: The covid-19 open research dataset.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4664–4674, Hong Kong, China. Association for Computational Linguistics.

		Entities			Attributes			Relations		
Model	**Avg Val F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
SpERT-attrs-as-ents	80.45	90.13	**88.63**	89.37	**92.35**	82.13	86.94	**77.59**	74.34	75.92
SpERT-modified	**80.66**	89.81	87.87	88.83	91.89	**82.62**	**87.01**	76.43	73.72	75.05
SpERT-modified-maxpool	80.22	**90.32**	88.54	**89.42**	92.00	80.90	86.09	76.11	**75.92**	**75.99**
SpERT-modified-unfiltered	79.99	89.28	88.03	88.64	91.65	80.74	85.84	75.62	73.98	74.78

Table 4: Micro **P**recision, **R**ecall, and **F1** averaged over 5 runs on the SciClaim test set as well as F1 averaged over the 3 tasks on 5 runs of SciClaim validation data (**Avg Val F1**).

A Claims Dataset

Our English language dataset SciClaim consists of 901 examples sentences divided into a training set of 721 sentences, a validation set of 80 sentences, and a test set of 100 sentences. The training and validation data contain examples that were labeled from corrected suggestions from a partially trained model, while the test set was labeled from scratch without any model suggestions as a starting point.

Our data from CORD-19 (Wang et al., 2020) is sampled with the following keywords as a heuristic identification of claims and causal language similar to our expert identified data from PubMed and Social and Behavioral Science (SBS) papers: associated with, reduce, increase, leads to, led to, our result, greater, less, more, cause, demonstrate, show, improve.

200 of our sentences (50 from PubMed and 150 from SBS) were selected to intentionally minimize the likelihood of claims and causal language. This includes sentences that discuss factors and other entities present in our schema but either do not contain associations or frame associations in unusual ways such as rhetorical questions. We intend for this data to encourage robustness that maintains correct labels for partial graph extractions rather than simply hallucinating associations in all sentences. We employ the following heuristics to identify this data: We sample 50 PubMed sentences from Yu et al. (2019) that are identified as having low causal content. We sample 100 titles from SBS paper present in Alipourfard et al. (2021), as titles contain factors but rarely contain explicit associations and may be present in input data from automatically extracted text from PDFs. Finally we sample 50 first lines of SBS papers from Alipourfard et al. (2021), as these lines frequently introduce topics or rhetorical questions which either lack associations or present highly hypothetical associations unlike those in our main corpus.

Each filtered data source was sampled chronologically.

We utilized the following procedure for labeling: The annotators undertook extensive, iterative schema design sessions in consultation with a subject matter expert in reproducibility of SBS experiments. After the schema was settled on pilot examples, a lead annotator established the annotation standards on several hundred examples through a process of relabeling and retraining the suggestion model. Once the suggestion model became effective, the lead annotator and model suggestions guided the other annotator in adopting the annotation standards. The lead annotator reviewed and corrected the 250 overlapping examples in a consensus process with the other annotator.

B Variants and Hyperparameters

B.1 Variants

We experiment with several variants none of which substantially outperformed the others. **SpERT-modified-maxpool** contains our modifications but simply uses SpERT's original maxpooling span representation instead of the attention-based representations inspired by Lee et al. (2017). **SpERT-modified-unfiltered** forgoes cascading inferences and instead classifies all possible spans for attributes. Full test and averaged validation results are presented in Table 4.

B.2 Hyperparameters

The following hyperparameters and settings were selected using manual tuning of 10-fold cross validation on the training set and optimizing for average micro-f1 performance over all 3 tasks: language model SciBERT uncased, mini batch size 8, epochs 40, optimizer AdamW, linear scheduling, warm up 0.05, learning rate 5e-5, learning rate warm up 0.1, weight decay 0.01, max grad norm 1.0, size embedding dimension 25, dropout probability 0.1, maximum span size 20, attribute filter threshold 0.55, relation filter threshold 0.4.

We ran 32 trials on 5 RTX 2080 ti GPUs, where each trial takes roughly 20 minutes. Our model contains 110 million parameters.

We explored the following hyperparameter bounds: language model $\in$ {BERT, SciBERT, SpanBERT, SciBERT tuned on SciERC }, epochs $\in$ {20, 40, 80}, batch size $\in$ {4, 8, 16}, learning rate $\in$ {1e-5, 5e-5, 1e-4}, scheduling $\in$ {linear, cyclic }, warm up $\in$ {0.0, 0.05, 0.1, 0.15. 0.2, 0.25, 0.3}, attribute filter threshold $\in$ {0.4, 0.5, 0.55, 0.6}, relation filter threshold $\in$ {0.35, 0.4, 0.5, 0.6}. The remaining settings we inherit from SpERT as initial experimentation on early datasets revealed little impact.