GOTTFRIED WILHELM LEIBNIZ UNIVERSITÄT HANNOVER
FAKULTÄT FÜR ELEKTROTECHNIK UND INFORMATIK

Semantic Parsing and Validation of Scholarly Contributions

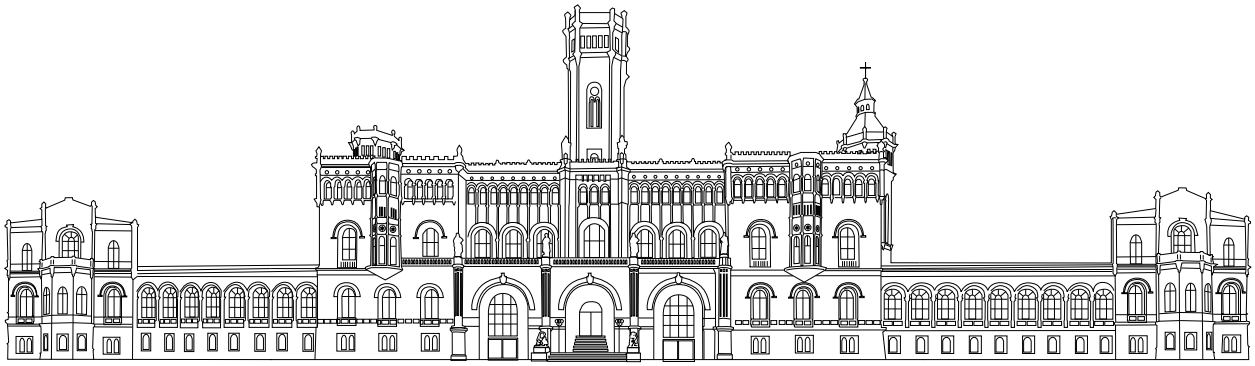
Artifical Intelligence Lab

BY

Meike Liedtke
E-mail: meike.liedtke@stud.uni-hannover.de

Supervisor: Ildar Baimuratov

28.01.2024



1 Introduction

With the increasing number of scientific publications, the problem arises, that an extensive peer review of newly published papers becomes increasingly difficult, due to the sheer amount of new publications [15]. In light of this development, automatic validation of scientific papers has become increasingly interesting.

In this Project, we propose an approach to semantic modeling of scientific claims allowing their automatic validation against evidence. For this, a Knowledge Graph is created from the evidence described in each Paper and generalized SPARQL queries are used to validate the claims.

The basic steps are as follows:

1. The claims at the beginning of the paper are identified and extracted.
2. For each claim, the experimental results in the paper need to be identified that prove or contradict the authors' claims. These also need to be extracted and, where appropriate, tables need to be included.
3. The results are modeled as triples in the ontology tool Protégé[19]. The resulting triples form a machine-readable knowledge graph that can be verified.
4. The appropriate queries are selected for each claim to validate the results.

The queries are divided into 2 classes, one to validate the quantitative results of the paper. And another one to validate the qualitative results or properties that can not be measured. Generalized queries allow the reuse of the two types of queries over all ontologies and papers.

All artifacts mentioned in this paper can be accessed at the following repository: Validation-of-Scholarly-Contributions[1]. This includes papers and artifacts from the implementation section and a tabular overview of the approaches from the related works section and their language expressiveness.

2 Related Work

There is a wide variety of semantic modeling of scientific papers. Some approaches focus on the metadata of papers, like controlled keywords, cited papers, and other bibliographic metadata [18, 14, 1].

Research also varies in what parts of the paper are used for semantification. Some research focuses mainly on the abstract [10, 4, 2], which allows for fast modeling of the content of a paper. Others include more sections of a paper but still omit the conclusion section[9]. However, to implement compliance checking it is important to model specific claims, as well as the results of a paper. This information can usually only be found in the second half of the paper, for example in the 'experimental results' or 'conclusion' sections.

Additionally, the granularity of semantification can differ. Annotation on a sentence-wise level [3, 16] allows for easy and efficient annotation. However, it does not allow for validating claims with the results stated in the paper.

Lastly, the expressivity of the annotation model is crucial for its use in validating claims. Most papers utilize ontologies or vocabularies on the OWL Lite level or equal[17, 11, 12, 20, 5]. Even though they can be used to model class hierarchy, this is insufficient to verify claims.
See also Table A.1 in the Appendix

[1] https://github.com/Ninniachwen/Validation-of-scholarly-contributions

3 Approach

This project shows the conversion of claims and experimental proof from scientific papers into a machine-actionable format. The evidence from the paper is modeled as a knowledge graph directly in an ontology tool and validated by SPARQL Queries, selected by the type and content of the claims. Generally, only those results are modeled that are needed to verify or contradict a claim. Experimental results which have no connection to the identified claims are not modeled.

3.1 Modeling Challenges

The conversion of claims and evidence from papers in the course of this project has shown several challenges concerning precise modeling. Firstly, claims and results need to be Identified. The first part is usually easy because most claims already appear in the Abstract. Sometimes they are repeated in more detail or there are additional claims in the 'Introduction' or 'Contribution' sections. But after this, there are usually no new claims. It proved more difficult to locate results or experimental proof for these claims. One difficulty lies in the wording. When a claim is formulated with "attains state-of-the-art performance on character-level language modeling"[13] the topic of this claim is "character-level language modeling". In the results sections, however, the proof of this claim is in the chapter "character prediction". Because of this, a precise study of each paper is required, to identify this hidden evidence of a claim. An Annotation framework like DOMEO [7] can help in such cases, by highlighting paragraphs of interest. This can either be done by using domain and paper-specific vocabulary, detected in the sentences annotated for claims or the title. Or a search for more generalized vocabulary like "results", "benchmark", and "show". However, this is not in all cases helpful, which is why a good understanding of the paper including a longer reading time is often required.

Another challenge in modeling the evidence in a paper is in the partial tabular expression. Very few papers mention both new and old scores in a sentence. Mostly no more than one new score is mentioned and a reference to the respective results table is given. Additionally, the Units of result tables are not always clearly stated as in Table 3 of one of the sampled papers [13]. This puts the annotator in the position to search the experimental section, to find a unit for the given results. After that, they still need to be able to interpret the score if not given by the author. Either domain knowledge or research time is required to fulfill this task satisfactorily. Also, the type of the result can make the modeling more difficult. A SPARQL query can easily compare quantitative results that appear as scores in a certain unit. A result formulated as "closer to actual or literal translations"[6] is harder to interpret. Since the authors didn't supply a formalization or measurement, this can only be modeled as a qualitative property. The same is true for statements like "linear runtime"[13]. It too can only be modeled as qualitative property which the new models have and the old ones don't - always assuming the claim proves true.

4 Implementation

The Dataset used for the experiments is the trial-data [8] from the ncg-task [10]. We chose this corpus because it is a collection of well-known papers from different NLP domains. Also, in addition to the original pdf, two text extractions in txt files are included in the corpus. These can later be used for the annotation step. For this implementation, pdf versions of the papers were extracted and several papers were randomly chosen. The proof was modeled as a single knowledge graph per paper in protégé [19] version 5.5. The queries were designed using the SPARQL Query plugin.

For each claim in a paper, a matching SPARQL query is chosen. The queries are divided into 2 types, each can be modified slightly to fit the purpose.

The first type of query covers quantitative claims. These can be verified using greater-than (>) or smaller-than (<) operators, as shown in Listing A.3 of the Appendix. Here, experimental results from the paper are modeled as a data property. The query extracts all data properties from the ontology and divides them into data properties of `baseline_models` and new (all other) models. The filter ensures that each entity only belongs to one of the model groups. Also, only those data properties that belong to this query type are selected (> or < accordingly). The values of the data properties are then compared, using the comparison operator chosen by the user. If the claim is valid, the query result contains exactly as many lines as there were new approaches that were modeled. The Query can be used as is, if there are both larger_than and smaller_than properties to compare (Listing A.3 of the Appendix). If only one type of property types exits, the query needs to be edited as can be seen in Listing A.4 of the Appendix.

```
"owl:Thing"
├── "experimental_results"
│   ├── new_property_1
│   ├── new_property_2
├── "model"
    ├── new_model_class
    ├── "baseline_model"

"owl:topObjectProperty"
├──experiment_op
    ├── interpreteble

"owl:topDataProperty"
├── "experiment_larger_than"
│   ├── BLEU_score
├── "experiment_smaller_than"
    ├── runtime
```

Figure 1: Class Hirarchy

The second type of query covers qualitative claims (A.2 of the Appendix). Again, the query separates the two groups of models and then keeps only those with object properties. If only the new models have the claimed property, the number of results resembles the number of new models. This also works for several claimed object properties. Again, if the claim is valid, the query result contains exactly as many lines as there were new approaches that were modeled.

Lastly, the count query is needed to answer the question if the number of returned lines equals the number of new approaches, as shown in A.1 of the Appendix. For these queries to work, the ontology must follow the schema shown in Figure 1. Classes and properties in quotation marks must be named exactly like this, all other names are flexible. Also, the hierarchy of sub-classes and -properties must be followed, for the queries to work without being changed for each paper. This means the triple `:experiment_larger_than rdfs:subPropertyOf owl:topDataProperty` must exist. The other subclass and subproperty relations should be noted down accordingly. Generally, before executing a query, the bottommost prefix must be changed according to the current ontologies prefix.

5 Conclusions and Future Work

In this Project, we propose an approach to semantic modeling of scientific claims allowing their automatic validation against evidence. For this, a Knowledge Graph is created from the evidence described in each Paper and generalized SPARQL queries are used to validate the claims.
Parsing and annotating of claims and contributions in a paper turned out to be a difficult task, which requires much human interpretation. Tools to support this task like DOMEO [7] can be very helpful and speed up the process.

A new Task that was recently published is related in its methodology to annotate and extract contributions of scientific papers. The SimpleText Task4 @ CLEF'24 is called SOTA [**sota**] and asks to report state-of-the-art performance from scientific publications. For this, it proposes to identify tasks, datasets, metrics, and scores from papers and to extract all relevant tuples. The resulting annotated text, or even the global knowledge graph can be used as input for our modeling of experimental proof from scientific papers.

Appendix A

1 Queries

Listing A.1: **SPARQL Query 1: Count** Count all entities belonging to baseline_models and new_models respectively

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX : <http://www.semanticweb.org/meike/ontologies/2024/0/m_0#>

SELECT ?model_class (COUNT(distinct ?model) AS ?count)
WHERE {
        ?model_class rdfs:subClassOf :model.
        ?model rdf:type ?model_class.
} Group by ?model_class
```

Listing A.2: **SPARQL Query 2: Quality** Find all models, that have the new quality property, or properties

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX : <http://www.semanticweb.org/meike/ontologies/2024/0/m_1#>

SELECT ?models ?model_classes
WHERE {
        ?model_classes rdfs:subClassOf :model.
        ?models rdf:type ?model_classes.

        ?prop_class rdfs:subPropertyOf owl:topObjectProperty.
        ?new_prop rdfs:subPropertyOf ?prop_class.       # define new object property
        ?models ?new_prop ?value.                # only models that contain new property

    ## If there are 2 qulaity properties (or more).
    ## Duplicate this block for each additional one
        ?new_prop2 rdfs:subPropertyOf ?prop_class.
        ?models ?new_prop2 ?value2.      # only models that contain both new properties
        Filter( ?new_prop != ?new_prop2 )        # two different new properties

} Group by ?models ?model_classes
```

Listing A.3: **SPARQL Query 3: Quantity** Find all new models, which satisfy the claim to have all data properties larger and smaller than the baseline models

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX : <http://www.semanticweb.org/meike/ontologies/2024/1/n_0#>

SELECT ?models_n (GROUP_CONCAT(DISTINCT ?prop; SEPARATOR=", ") AS ?properties) (GROUP_CONCAT(DISTINCT ?prop2; SEPARATOR=", ") AS
WHERE {
        ?model_b rdf:type :baseline_model.       # baseline models
        ?new_model_class rdfs:subClassOf :model.
        ?models_n rdf:type ?new_model_class.            # models other than baseline
        FILTER NOT EXISTS { ?models_n rdf:type :baseline_model }        # compare baseline with non baseline

        ## quantitative property tl to compare (>)
        ?prop rdfs:subPropertyOf :experiment_larger_than.
        ?model_b ?prop ?baseline_results_lt.            # baseline values
        ?models_n ?prop ?new_results_lt.                        # new values
        FILTER(?new_results_lt > ?baseline_results_lt)

        ## quantitative property st to compare (<)
        ?prop2 rdfs:subPropertyOf :experiment_smaller_than.
        ?model_b ?prop2 ?baseline_results_st.           # baseline values
        ?models_n ?prop2 ?new_results_st.                       # new values
        FILTER(?new_results_st < ?baseline_results_st)

} Group by  ?models_n
```

Listing A.4: **SPARQL Query 3: Quantity larger than** Find all new models, which have a data property larger than its counterpart of the baseline models (lt version of Quantity

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX : <http://www.semanticweb.org/meike/ontologies/2024/1/n_4#

SELECT ?models_n (GROUP_CONCAT(DISTINCT ?prop; SEPARATOR=", ") AS ?properties) (GROUP_CONCAT(DISTINCT ?prop2; SEPARATOR=", ") AS
WHERE {
        ?model_b rdf:type :baseline_model.       # baseline models
        ?new_model_class rdfs:subClassOf :model.
        ?models_n rdf:type ?new_model_class.            # models other than baseline
        FILTER NOT EXISTS { ?models_n rdf:type :baseline_model }        # compare baseline with non baseline

    ## quantitative property lt to compare (>)
        ?prop rdfs:subPropertyOf :experiment_larger_than.
        ?model_b ?prop ?baseline_results_lt.            # baseline values
        ?models_n ?prop ?new_results_lt.                        # new values
        FILTER(?new_results_lt > ?baseline_results_lt)

    ## quantitative property st to compare (<)
        #?prop2 rdfs:subPropertyOf :experiment_smaller_than.
        #?model_b ?prop2 ?baseline_results_st.          # baseline values
        #?models_n ?prop2 ?new_results_st.                      # new values
        #FILTER(?new_results_st < ?baseline_results_st)

} Group by  ?models_n
```

Table A.1: Related Works

	Paper	Language Expressivity	Granularity Level
[18]	Enhancing Knowledge Graph Extraction and Validation From Scholarly Publications Using Bibliographic Metadata	rdf owl lite level	sentence level
[17]	Semantic representation of scientific literature: bringing claims, contributions and named entities onto the Linked Open Data cloud	owl lite level, has hierarchy (rdfs) no equivalence	sub-sentence level
[11]	The Digitalization of Bioassays in the Open Research Knowledge Graph	owl lite level (no union, disjoint, etc)	sub-sentence level
[12]	ORKG Templates	owl lite level	sub-sentence level
[20]	Toward Representing Research Contributions in Scholarly Knowledge Graphs Using Knowledge Graph Cells	owl lite level	sub-sentence level
[5]	SciData: a data model and ontology for semantic representation of scientific data	JSON-LD = owl lite level	sub-sentence level
[14]	Understanding and using the medical subject headings (Mesh) Vocabulary to perform literature searches	pre rdf (1994)	headers only
[1]	Scalable, Semi-Supervised Extraction of Structured Information from Scientific Literature	triple structure owl lite level	phrase level
[10]	SemEval-2021 Task 11: NLPContributionGraph - Structuring Scholarly NLP Contributions for a Research Knowledge Graph	owl lite level	sub-sentence level
[4]	The contribution of cause-effect link to representing the core of scientific paper—The role of Semantic Link Network	owl lite level	sub-sentence level
[2]	Domain-Independent Extraction of Scientific Concepts from Research Articles	annotation only no owl or rdf	phrase level
[9]	NLPContributions: An Annotation Scheme for Machine Reading of Scholarly Contributions in Natural Language Processing Literature	owl lite level, orkg	sub-sentence level
[3]	SEPIO: A Semantic Model for the Integration and Analysis of Scientific Evidence	owl lite level	sentence level
[16]	Crowdsourcing Scholarly Discourse Annotations	rdf level = owl lite level	class annotation on sentence level
[7]	Open semantic annotation of scientific publications using DOMEO	OA: owl lite level SWAN: owl DL level	sub-sentence level

Bibliography

[1] Kritika Agrawal, Aakash Mittal, and Vikram Pudi. "Scalable, semi-supervised extraction of structured information from scientific literature". In: *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*. 2019, pp. 11–20.

[2] Arthur Brack et al. "Domain-independent extraction of scientific concepts from research articles". In: *European Conference on Information Retrieval*. Springer. 2020, pp. 251–266.

[3] Matthew H Brush, Kent A Shefchek, and Melissa A Haendel. "SEPIO: A Semantic Model for the Integration and Analysis of Scientific Evidence." In: *ICBO/BioCreative*. 2016.

[4] Mengyun Cao, Xiaoping Sun, and Hai Zhuge. "The contribution of cause-effect link to representing the core of scientific paper—The role of Semantic Link Network". In: *PloS one* 13.6 (2018), e0199303.

[5] Stuart J Chalk. "SciData: a data model and ontology for semantic representation of scientific data". In: *Journal of cheminformatics* 8 (2016), pp. 1–24.

[6] Kyunghyun Cho et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *arXiv preprint arXiv:1406.1078* (2014).

[7] Paolo Ciccarese, Marco Ocana, and Tim Clark. "Open semantic annotation of scientific publications using DOMEO". In: *Journal of biomedical semantics*. Vol. 3. 1. BioMed Central. 2012, pp. 1–14.

[8] D'Souza. *Trial data for the NLPContributionGraph Shared Task 11 at SemEval-2021*. GitHub. Feb. 2024. URL: https://github.com/ncg-task/trial-data.

[9] Jennifer D'Souza and Sören Auer. "Nlpcontributions: An annotation scheme for machine reading of scholarly contributions in natural language processing literature". In: *arXiv preprint arXiv:2006.12870* (2020).

[10] Jennifer D'Souza, Sören Auer, and Ted Pedersen. "SemEval-2021 Task 11: NLPContributionGraph–Structuring Scholarly NLP Contributions for a Research Knowledge Graph". In: *arXiv preprint arXiv:2106.07385* (2021).

[11] Jennifer D'Souza et al. "The Digitalization of Bioassays in the Open Research Knowledge Graph". In: *International Conference on Database and Expert Systems Applications*. Springer. 2022, pp. 63–68.

[12] Leibinz Universität Hannover. *ORKG Templates*. ORKG Templates. Jan. 2024. URL: https://orkg.org/about/19/Templates.

[13] Nal Kalchbrenner et al. "Neural machine translation in linear time". In: *arXiv preprint arXiv:1610.10099* (2016).

[14] Henry J Lowe and G Octo Barnett. "Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches". In: *Jama* 271.14 (1994), pp. 1103–1108.

[15] ncses. *Publications Output: U.S. Trends and International Comparisons*. Feb. 2024. URL: https://ncses.nsf.gov/pubs/nsb20206/executive-summary.

[16] Allard Oelen, Markus Stocker, and Sören Auer. "Crowdsourcing scholarly discourse annotations". In: *26th International Conference on Intelligent User Interfaces*. 2021, pp. 464–474.

[17] Bahar Sateli and René Witte. "Semantic representation of scientific literature: bringing claims, contributions and named entities onto the Linked Open Data cloud". In: *PeerJ Computer Science* 1 (2015), e37.

[18] Houcemeddine Turki et al. "Enhancing knowledge graph extraction and validation from scholarly publications using bibliographic metadata". In: *Frontiers in research metrics and analytics* 6 (2021), p. 694307.

[19] Stanford University. *Protégé, V5.5*. https://protege.stanford.edu/. Feb. 2024. URL: https://protegewiki.stanford.edu/wiki/Protege_Desktop_Old_Versions.

[20] Lars Vogt et al. "Toward representing research contributions in scholarly knowledge graphs using knowledge graph cells". In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. 2020, pp. 107–116.