# The Effect of Knowledge Graph Schema on Classifying Future Research Suggestions

Dimitrios Alivanistos[1,2][0000−0002−0090−2069], Seth van der Bijl[1][0009−0005−5234−8265], Michael Cochez[1,2][0000−0001−5726−4638], and Frank van Harmelen[1][0000−0002−7913−0048]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
[2] Discovery Lab. Elsevier, Radarweg 29a, 1043 NX Amsterdam, The Netherlands
https://discoverylab.ai/
{d.alivanistos, m.cochez, frank.van.harmelen}@vu.nl
seth.vanderbijl@icloud.com

**Abstract.** The output of research doubles at least every 20 years and in most research fields the number of research papers has become overwhelming. A critical task for researchers is to find promising future directions and interesting scientific challenges in the literature. To tackle this problem, we hypothesize that structured representations of information in the literature can be used to identify these elements. Specifically, we look at structured representations in the form of Knowledge Graphs (KGs) and we investigate how using different input schemas for extraction impacts the performance on the tasks of classifying sentences as future directions. Our results show that the MECHANIC-GRANULAR schema yields the best performance across different settings and achieves state of the art performance when combined with pretrained embeddings. Overall, we observe that schemas with limited variation in the resulting node degrees and significant interconnectedness lead to the best downstream classification performance.

**Keywords:** Information Extraction · Scientific Knowledge Graphs · Scientific Discourse · Classification

## 1 Introduction

Scientific papers often discuss future research directions and challenges, suggesting potential areas for further exploration. These are commonly found in the *future work* and *conclusion* sections and for multiple venues, they are a requirement for publication. Given a potentially infinite number of research trajectories, future research statements in papers attempt to guide researchers into promising directions. Knowledge gaps or unresolved questions may be identified and recommended as the most impactful directions.[3]

[3] Code, documentation and demo for exploring extracted triples, located: github repository, demo exporer

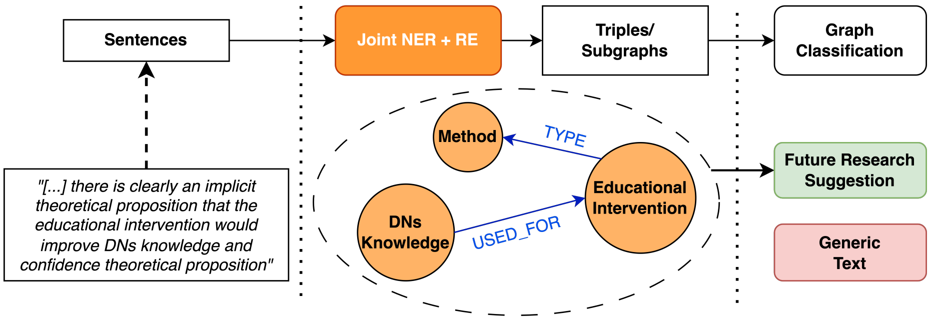
Sentences
Joint NER + RE
Triples/ Subgraphs
Graph Classification
"[...] there is clearly an implicit theoretical proposition that the educational intervention would improve DNs knowledge and confidence theoretical proposition"
Method
TYPE
Educational Intervention
DNs Knowledge
USED_FOR
Future Research Suggestion
Generic Text

**Fig. 1.** Illustration of the pipeline with an example. First, a sentence (either research suggestion or random) is fed to a joint NER and RE model. The generated triples form graph(s). Lastly, these graphs form the input to the graph classification which classifies them as containing a *future research suggestion* or not.

Multiple approaches have attempted to utilise Machine Learning (ML) and Natural Language Processing (NLP) to automatically identify such suggestions. In this work, we look at the problem of classifying future research suggestions when modelling the discourse not as text, but as a knowledge graph.

### 1.1 Problem Statement

Utilizing the future research directions to guide research becomes increasingly difficult due to the volume of published research. Several studies indicate that the growth of the size of research papers in terms of references, statistics, participants and tables [49] and especially the total volume of published studies in a given period increases continuously, leading to exponential growth [63,17,42,7,35]. Some studies suggest that research output currently doubles at least every 20 years with some periods of the 20th century seeing research output double every 7 years [34]. Meta-analysis can help condense information in certain disciplines but meta-analyses are subject to an even stronger volume increase [17], with specific fields seeing explosive growth [29]. The increasing volume in academic publications is an inspiring indicator of progress. However, the information overload is time-consuming [1] and can lead to a diminished quality in the interaction between researcher and research papers [44].

With the volume in research output increasing, the amount of potential future research directions becomes intractable. Thus, efficiently collecting and comparing future research indications becomes overly tedious for the average researcher. Furthermore, important directions might be overlooked, and researchers may fail to combine research directions that appear in separate sources [16]. Groundbreaking discoveries might be missed as the right questions or directions remain hidden in the large amount of data [24]. With knowledge being mostly published in natural language, distributed among research papers in multiple journals and via diverse media, the question of effective communication in academic research is of paramount importance [27].

Although modelling scientific discourse has received increased attention, there have been few attempts focused on structuring future research suggestions, and summarize and communicate these efficiently.

In this work, we introduce an architecture that takes as input scientific sentences that contain future research suggestions (or not), transforms them into graphs via triple extraction, and evaluates the suitability of the extracted graphs for the downstream task of predicting whether a (sub)graph contains a research challenge or direction recommendation. The produced graphs are further analysed based on topological metrics, allowing for a better comparison of the schemas used to generate them.

### 1.2 Research Questions

Our main research questions (RQs) examine how schema choice influences triple extraction and the resulting graph topology (RQ1) and its subsequent impact on downstream graph classification (RQ2).

For RQ1, we explore and analyze the effect of schema choice on both local and global KGs in terms of topological features. In RQ2, we investigate how different schema characteristics impact the classification of (sub)graphs containing research recommendations. We also study the relationship between classification performance and graph topology. Lastly, we assess the effectiveness of graph classifiers for research suggestion classification.

## 2 Related work

***Scientific Paper Segmentation & Argumentation Extraction.*** Substantial research has been performed in extracting (semi-)structured data from research papers, such as scholarly argumentation mining (SAM) [5,41,38], research paper segmentation of text and figures [37,8], and parsing the figures of research papers [55]. Scientific metadata extraction focuses on extracting title, authorship and other metadata from articles [47]. Other approaches attempt to link the extracted extracted concepts to articles or sentences from articles [32], also known as entity linking. Related research proposes KG based systems for recommending a scientific method or technique for a scientific problem [39]. Automated hypotheses generation generates promising hypotheses from research papers [56,65]. Furthermore, the shared task 11 of SemEval 2022 [43] of recognizing contributions of a paper brought increased attention to the problem of extracting knowledge from publications. Identifying the contributions of a paper can be a valuable task for researchers looking to build on existing work as it can lead to better recommendations. Lastly, subjectivity analysis has received attention, since filtering out the subjective sentences from a paper appears to improve Information Extraction (IE) [50,64].

***KG Construction for Scientific Discourse.*** Knowledge Graphs are graph structures (networks of labelled nodes and edges) [15] with additional semantics [9]. Entities and relations in a KG are expressed in the form of triples of *subject, predicate, object* that correspond to an edge (predicate) between two nodes (subject and object). Entities and relations are often typed with a class.

KGs can be constructed using several methods. Automated approaches attempt to model research content into a KG by employing pipelines involving Natural Language Processing (NLP) methods, rather than hand-engineering. These automated Information Extraction (IE) approaches can model scientific knowledge in the form of a graph of (scientific) triples. Initially, IE methods employed domain-specific, rule-based systems while recent years have seen approaches employing Machine Learning (ML) [52] and Deep Learning (DL) take over [24]. For these approaches, a ML model is trained based on a "high-quality" dataset, often manually annotated. The trained model is then applied to a larger text dataset to automatically extract entities and relations to build a KG. Examples of such KGs are the COVID-19 themed CORD-KG (generated with DyGIE++ trained on MECHANIC data, [25]), the material sciences themed MatKG (build with an LLM transformer, [60]), or the AI-based Intelligence Task Ontology (ITO) KG [6]. Employing AI/ML for the generation of KGs allows for automatic generation on a larger scale. However, automatic generation can be less robust to erroneous input data and the resulting graph can be of low quality. Different disciplines tend to exhibit large discrepancies in adopting IE [24].

Alternatively, KGs can be constructed by hand or by extracting data from semi-structured datasets (e.g. patient records), without employing ML approaches. These KGs are often high-quality but small-scale as human annotations are time-consuming and depend on the expertise of the annotator. Some examples include the chemical protein interactions of ChemProt [58] and adverse drug events (ADE) [19].

## 3 Methodology

The goal of this research is to construct a graph from sentences suggesting future research, observe and analyse the topological features of said graph, and test the influence of the schema on the downstream task of graph classification. To achieve this we designed the following pipeline. Starting from the dataset created by [25], we take future research suggestion sentences (either a *challenge* or a *direction*) from papers and extract their corresponding triples using a joint Named Entity Recognition (NER) and Relation Extraction (RE) model pre-trained on different selected schemas. Although we understand the importance of NER/RE model performance on the resulting graphs and hence downstream task, in this work we are focusing on the effect of their underlying schema on downstream performance, since more generic schemas can for example result in higher recall but lower precision. We then analyse the resulting graphs both locally (subgraph-level, i.e. collection of triples) and globally (full KG) in terms of topological features. Finally, we employ Relational Graph Convolutional Networks (R-GCNs) to perform graph classification on the resulting graphs.

The rest of this section describes the input data and schema choices and the graph classification using R-GCNs, including details around the architecture and experimental setup. The pipeline is illustrated by figure 1.

### 3.1 Schema and Dataset

There are multiple options when choosing a schema for modelling research content. An extensive selection can be found in table 5 of the Appendix A. Amongst the listed options, six candidates were considered fit for this work. These are the MECHANIC (Coarse-Granular), SciERC, ACE05, ACE-Event and GENIA. The criteria for the selection were based on granularity and generalizability. For example SciERC has several entity and relationship types, whereas MECHANIC-Coarse only has a single entity type and two relationship types. Concerning generalizability, MECHANIC-Coarse is about general science, while GENIA is specifically about biology. Another motivation for the chosen schemas was their accessibility in the pretrained DyGIE++ joint NER and RE model, which allowed for a straight-forward comparison. The performance of DyGIE++ for NER/RE can be found in the original DyGIE++ paper [61]. The results of our analysis focused on the schemas found in table 1. A comparison of the different entity and relation types defined by each schema can be found in the Appendix A in table 7.

**Table 1.** The selected datasets and schemas investigated in this work. The difference between 0 and N/A for the number of entities or relations is that for the former no entities or relations were defined (like the ACE-Event schema) while for the latter it was OpenIE.

| Dataset + Schema | Domain | Size | Entities | Relations | Characteristics |
|---|---|---|---|---|---|
| SciERC [40] | CS | 500 abstracts | 6 | 7 | Coarse-grained, domain-specific |
| MECHANIC-Coarse[25] | Bio | 1000 sentences | 1 | 2 | Coarse-grained, domain-specific |
| MECHANIC-Granular[25] | Bio | 1000 sentences | N/A | N/A | Fine-grained, domain-specific |
| ACE05 [62] | Various | 511 documents | 7 | 6 | Complex, general |
| ACE-Event [61] | Various | 599 documents | 6 | 18 | Complex, general |
| GENIA [30] | Bio | 2,000 abstracts | 6 | 5 | Fine-grained, domain-specific |

SciERC is a dataset accompanied by a simple schema aimed at extracting methods, tasks and metrics from Computer Science and Artificial Intelligence (CS/AI) abstracts. Nevertheless, it shows good generalisability when detecting concepts and relations in future research sentences outside of CS/AI. GENIA was meant for DNA, RNA and proteins, hence it performs well when dealing with biomedical data. Since it lacks generalisability, it would not be a perfect fit as the basis of a scientific modelling schema, but could be beneficial as an addition to another, more general schema. The ACE05 dataset was available in the ACE-Event and ACE05 format, where the ACE-Event is a filtered version of the

ACE05 dataset focused on identifying events. As such, the ACE-Event dataset had different entity types and extended (sub)relation types. ACE05 relations and entities were designed for capturing news events involving people, organizations, locations, movements, and concepts that are physical in nature. This makes it interesting to experiment with for scientific content. Finally, the two MECHANIC schemas were created by [25] in the context of designing a knowledge base of mechanisms extracted from COVID-19 papers. A coarse-grained and fine-grained version were defined, with the coarse-grained version detecting entities of type *Entity* and relations of type *mechanism* (direct mechanisms) and *effect* (indirect mechanism) in sentences. The fine-grained version was closer to Open Information Extraction (OpenIE), where the verb of the sentence would denote the relation (e.g. "COVID-19 influences diabetes" results in "influences" as the relation type [25]).

### 3.2 Graph Classification of Challenges & Directions

***Motivation.*** Graph Convolutional Networks (GCNs) [31], are a class of neural architectures that operate on graph-structured data and leverage its topological features. GCNs can be applied to graph classification tasks, i.e. classifying a graph given its nodes and edges. An advantage of GCN architectures for graph classification lies in their ability to integrate both node attributes and graph topology into the node representations, which can in turn encode global and local characteristics of the graphs.

GCNs can learn node representations that are more expressive and discriminative than random walk-based methods, which solely rely on the graph structure and ignore features like node attributes. GCNs can handle graphs with varying sizes and structures. Furthermore, they can parse different types of graphs, such as directed, weighted, or heterogeneous graphs, by using the appropriate convolution operators. Therefore, for this work we considered GCNs for graph classification.

**Architecture and Implementation.** Initially we pretrained the GCN with a Link Prediction (LP) objective to obtain meaningful node representations. In absence of pre-existing node features, GCN models use a random initialisation. With the use of a pretraining routine involving a separate GCN model trained on a Link Prediction (LP) objective we obtain meaningful node representations, which are subsequently used to initialise the graph classification models. We investigate the effect of this initialisation on the performance in the graph classification task. Below we describe the architecture in further detail.

The architecture for pretraining consists of 2 GCN layers. The first layer takes randomly initialized node embeddings (5 channels) while outputting 128 channel encodings. The second GCN layer takes the 128 hidden channels and learns embeddings with a dimension of 64 channels from the 128 hidden channels. In the decoding phase, the decoder layer operates on the node embeddings created by the preceding GCN layers. For each edge $e = (i, j)$, the decoder computes the score for that edge as the dot product of the embeddings of nodes $i$ and $j$. The score is a scalar which is interpreted as the likelihood of the existence of an edge between nodes $i$ and $j$. For each positive edge $e$ we sample one negative edge $\hat{e}$ where either $i$ or $j$ is replaced by a node sampled uniformly at random, i.e. corrupted. After training on LP the encoder part of the model was employed to get the initial embeddings for all nodes in the (sub)graphs of the graph classification task.

For graph classification we experimented with both a GCN and a Relational-GCN (R-GCN, [54]). At each layer the GCN model applies a linear transformation to the node features and aggregates the features from the neighboring nodes using a first-order approximation of spectral graph convolutions. In contrast, the R-GCN model extends the conventional convolution operation introducing relation-specific weights. Here, at each layer, for each node $j$ the representations of neighboring nodes $i$ (those with an incoming edge $(i, j)$ ) are aggregated after passing them through a relation specific linear layer. In our experiments we investigate the influence of this model choice.

For both GCN and R-GCN, the architecture as illustrated in fig. 2 is equivalent by swapping the GCN layer(s) with R-GCN layer(s). Each (sub)graph is initially passed to three (relational) graph convolution layers. Then, we obtain a single embedding for each graph by performing a mean pooling over the node representations. Finally, we apply a dropout layer for regularization and a final linear layer to perform the graph classification.

### 3.3 Experimental Setup

In this section we describe our experimental setup, including the dataset, the extraction and finally the classification, including training and pretraining of the model.

***The Dataset.*** The gold labeled sentences are the ones provided by [33]. In the original dataset these sentences were classified as either a research challenge, a research direction, both, or neither. In that work, the focus was on building a search engine that would distinguish between the two classes and allow for their retrieval separately. However, in the current work we do not make the same distinction, because we are interested in discovering all future research recommendations, which include both challenges and directions. Hence, we cast the problem into a binary classification problem, where each sentence is labeled positive if it contains either a research challenge, a research direction, or both, and negative otherwise.

***Joint NER and RE.*** Entities and relations forming subgraphs were extracted from the sentences employing the DYGIE++ model. The pretrained models hosted in the DYGIE++ repository[4] [61] were used. To improve performance over scientific text, we replaced BERT with SciBERT [4] as our encoder, which

[4] DyGIE++

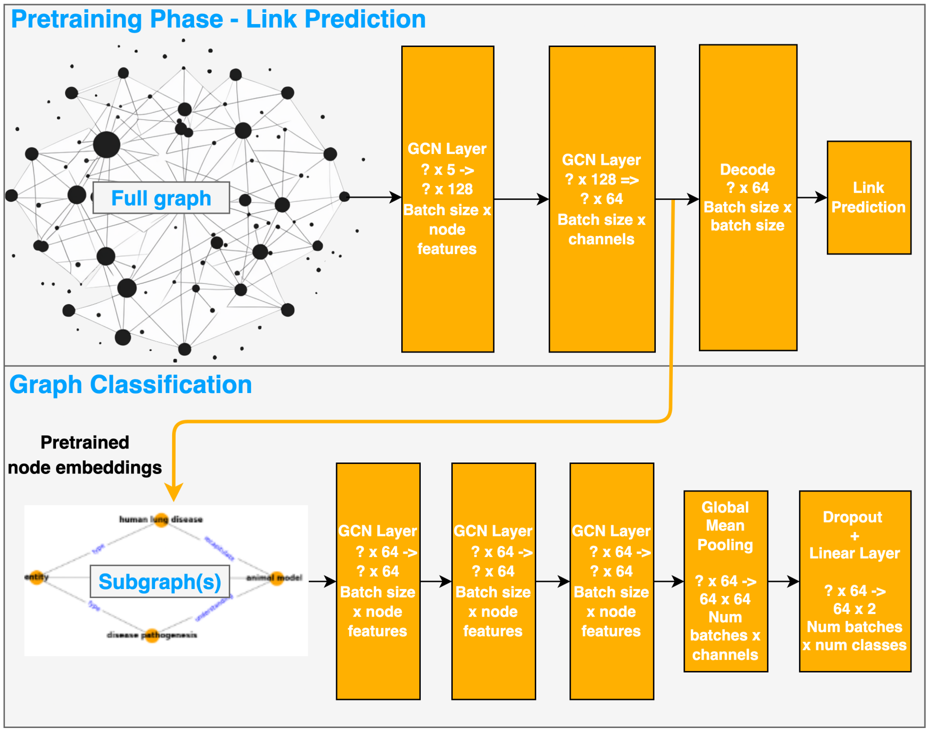
Pretraining Phase - Link Prediction
Full graph
GCN Layer ? x 5 -> ? x 128 Batch size x node features
GCN Layer ? x 128 => ? x 64 Batch size x channels
Decode ? x 64 Batch size x batch size
Link Prediction
Graph Classification
Pretrained node embeddings
human lung disease
type
entity
animal model
type
understanding
disease pathogenesis
Subgraph(s)
GCN Layer ? x 64 -> ? x 64 Batch size x node features
GCN Layer ? x 64 -> ? x 64 Batch size x node features
GCN Layer ? x 64 -> ? x 64 Batch size x node features
Global Mean Pooling ? x 64 -> 64 x 64 Num batches x channels
Dropout + Linear Layer ? x 64 -> 64 x 2 Num batches x num classes

**Fig. 2.** Illustration of using pretrained embeddings as initial node embeddings for the local subgraphs per sentence. Embeddings are learned by LP on the full graph using entities and relations from all sentences. The learned embeddings are used as initial node embeddings for the entities of the smaller graphs for each sentence which are classified using the (R)-GCN.

**Table 2.** Local graph topology metrics under different schemas. Relatively high values are in bold and relatively low values are in italics. Some statistics are first aggregated per subgraph and then averaged.

| | ACE-Event | ACE05 | GENIA | MECHANIC-Coarse | MECHANIC-Granular | SciERC |
|---|---|---|---|---|---|---|
| Entities | *4.730* ± 3.240 | 5.131 ± 3.551 | *3.386* ± 2.223 | **9.118** ± 9.513 | 6.206 ± 3.535 | **7.992** ± 6.389 |
| Relations | *2.243* ± 1.517 | 2.430 ± 1.617 | *1.625* ± 1.033 | **4.230** ± 3.542 | 3.062 ± 1.727 | **3.951** ± 3.143 |
| Degrees | 0.948 ± 0.729 | *0.947* ± 0.784 | 0.960 ± 0.669 | *0.928* ± 1.346 | **0.987** ± 1.099 | **0.989** ± 1.032 |
| Clusterings | 0.012 ± 0.103 | *0.011* ± 0.100 | *0.000* ± 0.000 | **0.204** ± 0.374 | **0.275** ± 0.434 | 0.080 ± 0.251 |
| Modularities | 0.185 ± 0.236 | **0.201** ± 0.243 | 0.071 ± 0.174 | *0.029* ± 0.075 | *0.023* ± 0.060 | **0.290** ± 0.229 |

**Table 3.** Global graph topology metrics under different schemas. Relatively high values are in bold while relatively low values are in italics.

| | ACE-Event | ACE05 | GENIA | MECHANIC-Coarse | MECHANIC-Granular | SciERC |
|---|---|---|---|---|---|---|
| Entities | *1568* | 1645 | *1065* | **7119** | 4240 | **6918** |
| Relations | *2677* | 2827 | *1873* | **11583** | 6367 | **9282** |
| Degrees | 3.415 ± 21.106 | **3.437** ± 21.986 | **3.517** ± 20.209 | 3.254 ± 72.537 | *3.003* ± 61.855 | *2.683* ± 47.125 |
| Clusterings | *0.054* ± 0.193 | 0.057 ± 0.199 | *0.039* ± 0.154 | **0.508** ± 0.444 | **0.557** ± 0.482 | 0.175 ± 0.348 |
| Modularities | **0.708** | **0.699** | 0.669 | *0.510* | *0.397* | 0.635 |

is fine-tuned on scientific data. Apart from this, default settings were used. For each schema, the DYGIE++ model extracted entities and relations. From the resulting entities and relations a (global) graph representing the full body of sentences was constructed.

***Graph Classification & Pretraining.*** The pipeline for the LP task begins by loading and preprocessing each extracted graph dataset. These graph datasets are then randomly divided into training, validation, and test sets, with 5% of edges set aside for validation and 10% for testing. The models are parameterized based on the feature dimensions of the nodes in the graph and are optimized using the Adam optimizer with a learning rate of 0.001. The loss function used for the pretraining task is the Binary Cross-Entropy (BCE) with Logits Loss. The models are then trained and subsequently evaluated on the test data, with the Area Under the ROC Curve (AUC) score being computed as a measure of model performance. The subsequent graph classification divided the subgraphs in 75% training graphs, 12.5% validation graphs and 12.5% test graphs. The classes were roughly balanced. The models were similarly optimized with the Adam optimizer with a learning rate of 0.001 using Binary Cross-Entropy (BCE) with Logits Loss. In the graph classification case precision, recall and F1 were used as measure of performance.

## 4 Results

The following section analyses the results as obtained with quantifiable metrics. Main patterns are noted, interpretation and contextualization of these results follows in the discussion section. We chose to focus on the resulting entities and relations, the node degree, clustering coefficient and modularity as they characterize numerically the topology of the generated graphs. The analysis is performed on both local and global level, i.e. over each subgraph and over the entire extracted graph.

***Joint NER and RE.*** As expected, the joint NER and RE for the different schemas resulted in different graphs. These are described and summarized through network graph metrics. We observe that the resulting graphs differ greatly in terms of topologies both visually and using metrics. The global full interconnected graph statistics are summarized in table 3. In parallel, the local graph statistics for the subgraph of each sentence are presented in table 2.

***Entities and Relations.*** The MECHANIC and SciERC schemas yield a higher number of detected entities and relations for both the global graph and local subgraphs. From our experiments we observe that a higher number of entities and relations is associated with better downstream task performance.

***Node Degree.*** Certain schemas result in high variation in terms of node degrees in the KG. Most schemas result in the majority of nodes having a low degree and a long tail of extremely connected nodes. GENIA has a mean degree on the higher side of the spectrum whereas MECHANIC-Granular and SciERC are positioned in the lower end of the spectrum. In terms of local subgraphs, SciERC and MECHANIC-Granular appear more interconnected, with higher degrees than other schemas, although the differences in terms of local degrees are subtle. In general, the standard deviation of these degrees increases with the average degree for local subgraphs. A higher standard deviation in the degree appears to be related to better performance in the downstream task in most of the settings. This phenomenon appears sensible since more densely connected subgraphs provide more insight in the relations between concepts in a subgraph. On a global level, during pretraining, without using types, degree centrality is the only measure that correlates with graph classification performance.

***Clustering Coefficients.*** Initially we expected a relation between degree standard deviation and clustering coefficient. However, this was not universally the case. MECHANIC-Granular and MECHANIC-Coarse exhibited very high clustering coefficients, while GENIA and the ACE schemas were on the low-side. This indicates presence of cliques was rather average in the SciERC graph. This positive influence of clustering coefficients was noted on both the global graph and local subgraph levels of the metrics.

***Modularity.*** In terms of modularity, SciERC, GENIA, ACE05 and ACE-Event resulted in rather tightly knit communities with few edges connected to other communities; MECHANIC-Granular and MECHANIC-Coarse produced graphs with communities that were less connected internally but more connected to other communities in the graph. This modularity is more intuitively illustrated in the network graph plots found in the demo. On a local level, modularity does not always relate to average degree. Some schemas are very interconnected but do not exhibit clear subcommunities, such as MECHANIC-Granular. On a global level higher modularity appears to lead to worse performance with high or moderate modularity schemas showing limited performance. The same holds on the local (subgraph) level, with the exception of SciERC which performs well even when having a high subgraph modularity.

### 4.1 Graph Classification

Results are aggregated based on whether the model was pretrained (+Pretrained), whether the model captured different relation types (+Relations) and for both. Table 4 denotes the average performance over 20 runs for the different combinations of these settings. In general, we observed that pretrained models (over the global extracted graph), exhibit a strong ability in the detection of future research challenges and directions. In table 4 we observe that the best performing run utilises both relation type information and the pretraining of node embeddings. In this setup, our model gives results comparable to the state-of-the-art in detecting future research suggestions.

## 5 Discussion

Overall, depending on the schema, we observed diverse graph topologies. For example, some resulted in more modular graphs compared to others. In specific cases, we observed that there were no entities or relations extracted at all. This could hint at the schema being unsuitable for the domain of choice. In parallel, the lack of extractions could imply that the sentence does not contain a research suggestion. It is noteworthy that schemas with a lower average degree generally also have much higher standard deviations in their degree. This indicates some very connected nodes and many sparsely connected nodes. In the current set-up it appears that the standard deviation and mean of the clustering coefficients on a local level are influential factors on downstream task performance. Another observation is that for most schemas/datasets the effect of the inclusion of relation types appears to have a stronger impact on performance in comparison to pretraining on the entire graph. Different schemas produce different graphs, and so differences emerge in the pretraining of embeddings, relation type usage and performance. When pretrained embeddings or relation types are not being utilised for classification, SciERC and MECHANIC-Granular consistently outperform the rest. Performing future research suggestion classification without pretraining and relation types proves difficult for any schema, resulting in poor performance on this setting, with the single exception being MECHANIC-Granular.

### 5.1 Limitations and Future Research

Future research can extend the present results in several directions. More schemas and input data, different joint NER and RE models and different downstream

**Table 4.** F1 scores when predicting the future research suggestion label $[0, 1]$ (generic text vs. future research suggestion). (R-)GCNs were applied to the local subgraphs per sentence. The *+Pretrained* indicates whether pretrained embeddings were used. *+Relations* indicates whether an R-GCN was used instead of a GCN for the local subgraph classification, to incorporate relation-type information. *+Pretrained + Relations* indicates the use of both. Best scores per mode are in bold and best combination of configuration (pretrained embeddings and typed relations) is underlined.

| | F1 | P | R |
|---|---|---|---|
| **ACE-Event** | 0.494 | 0.489 | 0.506 |
| + Pretrained | 0.511 | 0.462 | 0.580 |
| + Relations | 0.737 | 0.723 | 0.757 |
| + Pretrained + Relations | 0.964 | 0.976 | 0.953 |
| **ACE05** | 0.469 | 0.452 | 0.497 |
| + Pretrained | 0.527 | 0.486 | 0.580 |
| + Relations | 0.864 | 0.888 | 0.842 |
| + Pretrained + Relations | 0.978 | 0.971 | 0.985 |
| **GENIA** | 0.403 | 0.423 | 0.402 |
| + Pretrained | 0.369 | 0.354 | 0.394 |
| + Relations | 0.375 | 0.387 | 0.376 |
| + Pretrained + Relations | 0.579 | 0.510 | 0.682 |
| **MECHANIC-Coarse** | 0.544 | 0.481 | 0.629 |
| + Pretrained | 0.602 | 0.536 | 0.690 |
| + Relations | 0.878 | 0.871 | 0.887 |
| + Pretrained + Relations | 0.990 | 0.990 | 0.991 |
| **MECHANIC-Granular** | **0.594** | 0.507 | 0.723 |
| + Pretrained | 0.620 | 0.518 | 0.775 |
| + Relations | **0.906** | 0.903 | 0.911 |
| + Pretrained + Relations | **0.994** | 1.000 | 0.988 |
| **SciERC** | 0.560 | 0.478 | 0.684 |
| + Pretrained | **0.644** | 0.556 | 0.771 |
| + Relations | 0.884 | 0.877 | 0.893 |
| + Pretrained + Relations | 0.991 | 0.991 | 0.991 |

tasks (e.g entity linking) to list just a few potential extensions. While graph classification fits well to our task and purpose for testing the influence of the schemas, other downstream tasks, such as link prediction might be less sensitive to the graph topology resulting from a choice of schema given the setting of predicting the likelihood of a subject node being connected to an object node. Additionally, we tested a single joint NER and RE model for several different graphs. While DyGIE++ provides a baseline model for joint NER and RE, other more powerful extraction models may influence the resulting graph topology (e.g. detect more entities). DyGIE++ however has long been the state of the art and provides easily accessible models. The present research characterizes the fitness of a schema for scientific modelling by how it influences our graph topology and downstream classification task. However, downstream graph ML tasks can be influenced by multiple different parameters (i.e. hyper-parameter tuning), and isolating the effect of the schema might be challenging. Additionally, a limitation is observed over the gold set of [33] since the produced dataset is focused on COVID-19 research and hence is a domain-specific dataset. An intriguing direction of future research could employ OpenIE to dynamically construct schemas, with their usefulness evaluated by their performance on several downstream tasks in parallel (LP, graph classification etc.). This would yield schemas optimised for several downstream tasks at once, hence increasing robustness.

## 6 Conclusion

In this work we analysed the effect of the choice of schema when extracting knowledge from text in the form of KGs, to be further used for scientific knowledge discovery and recommendation. Specifically, we experimented with extracting graphs from sentences containing a scientific research suggestion or not, by employing pretrained models of DyGIE++ with different underlying schemas. We observed that the choice of schema can have a significant influence on both graph topology and downstream graph classification performance. Moreover, we observe that there is a correlation between several topology metrics of the resulting graphs and downstream task performance. The MECHANIC-Granular schema leads to solid downstream task performance with state of the art detection of future research suggestions when combined with pretrained embeddings and typed relations.

**Acknowledgments** We would like to thank Paul Groth and Thom Pijnenburg for their valuable comments and insights to this work.

## References

1. Achike, F.I., Ogle, C.W.: Information Overload in the Teaching of Pharmacology. The Journal of Clinical Pharmacology **40**(2), 177–183 (2000). https://doi.org/10.1177/00912700022008838, https://

onlinelibrary.wiley.com/doi/abs/10.1177/00912700022008838, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1177/00912700022008838

2. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 546–555. Association for Computational Linguistics, Vancouver, Canada (Aug 2017). https://doi.org/10.18653/v1/S17-2091, https://aclanthology.org/S17-2091
3. Becker, K.G., Barnes, K.C., Bright, T.J., Wang, S.A.: The genetic association database. Nature Genetics **36**(5), 431–432 (May 2004). https://doi.org/10.1038/ng0504-431
4. Beltagy, I., Lo, K., Cohan, A.: Scibert: Pretrained language model for scientific text. In: EMNLP (2019)
5. Binder, A., Verma, B., Hennig, L.: Full-Text Argumentation Mining on Scientific Publications. None (2022). https://doi.org/10.48550/ARXIV.2210.13084, https://arxiv.org/abs/2210.13084, publisher: arXiv Version Number: 1
6. Blagec, K., Barbosa-Silva, A., Ott, S., Samwald, M.: A curated, ontology-based, large-scale knowledge graph of artificial intelligence tasks and benchmarks. Scientific Data **9**(1), 322 (Jun 2022). https://doi.org/10.1038/s41597-022-01435-x, https://www.nature.com/articles/s41597-022-01435-x, number: 1 Publisher: Nature Publishing Group
7. Boschen, M.J.: Publication trends in individual anxiety disorders: 1980–2015. Journal of Anxiety Disorders **22**(3), 570–575 (Apr 2008). https://doi.org/10.1016/j.janxdis.2007.04.004, https://www.sciencedirect.com/science/article/pii/S0887618507001016
8. Bui, D.D.A., Del Fiol, G., Jonnalagadda, S.: PDF text classification to leverage information extraction from publication reports. Journal of Biomedical Informatics **61**, 141–148 (Jun 2016). https://doi.org/10.1016/j.jbi.2016.03.026, https://www.sciencedirect.com/science/article/pii/S153204641630017X
9. Davies, J., Fensel, D., Harmelen, F.v.: Towards the Semantic Web: Ontology-driven Knowledge Management. John Wiley & Sons (Jun 2003), google-Books-ID: kREOBAAAQBAJ
10. Dessì, D., Osborne, F., Reforgiato Recupero, D., Buscaldi, D., Motta, E., Sack, H.: AI-KG: An Automatically Generated Knowledge Graph of Artificial Intelligence. In: Pan, J.Z., Tamma, V., d'Amato, C., Janowicz, K., Fu, B., Polleres, A., Seneviratne, O., Kagal, L. (eds.) The Semantic Web – ISWC 2020. pp. 127–143. Lecture Notes in Computer Science, Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-62466-8_9
11. Deußer, T., Ali, S.M., Hillebrand, L., Nurchalifah, D., Jacob, B., Bauckhage, C., Sifa, R.: KPI-EDGAR: A Novel Dataset and Accompanying Metric for Relation Extraction from Financial Documents. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 1654–1659 (Dec 2022). https://doi.org/10.1109/ICMLA55696.2022.00254, http://arxiv.org/abs/2210.09163, arXiv:2210.09163 [cs]
12. D'Souza, J., Auer, S., Pedersen, T.: SemEval-2021 Task 11: NLPContributionGraph - Structuring Scholarly NLP Contributions for a Research Knowledge Graph. In: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021). pp. 364–376. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.semeval-1.44, https://aclanthology.org/2021.semeval-1.44

13. D'Souza, J., Hoppe, A., Brack, A., Jaradeh, M.Y., Auer, S., Ewerth, R.: The STEM-ECR Dataset: Grounding Scientific Entity References in STEM Scholarly Content to Authoritative Encyclopedic and Lexicographic Sources. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 2192–2203. European Language Resources Association, Marseille, France (May 2020), https://aclanthology.org/2020.lrec-1.268
14. Elsahar, H., Vougiouklis, P., Remaci, A., Gravier, C., Hare, J., Laforest, F., Simperl, E.: T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018), https://aclanthology.org/L18-1544
15. Euler, L.: Solutio problematis ad geometriam situs pertinentis. Commentarii academiae scientiarum Petropolitanae pp. 128–140 (Jan 1741), https://scholarlycommons.pacific.edu/euler-works/53
16. Feroz, H.M.B., Zulfiqar, S., Noor, S., Huo, C.: Examining multiple engagements and their impact on students' knowledge acquisition: the moderating role of information overload. Journal of Applied Research in Higher Education **14**(1), 366–393 (Jan 2021). https://doi.org/10.1108/JARHE-11-2020-0422, https://doi.org/10.1108/JARHE-11-2020-0422, publisher: Emerald Publishing Limited
17. Fontelo, P., Liu, F.: A review of recent publication trends from top publishing countries. Systematic Reviews **7**(1), 147 (Sep 2018). https://doi.org/10.1186/s13643-018-0819-1, https://doi.org/10.1186/s13643-018-0819-1
18. Gardent, C., Shimorina, A., Narayan, S., Perez-Beltrachini, L.: The WebNLG Challenge: Generating Text from RDF Data. In: Proceedings of the 10th International Conference on Natural Language Generation. pp. 124–133. Association for Computational Linguistics, Santiago de Compostela, Spain (Sep 2017). https://doi.org/10.18653/v1/W17-3518, https://aclanthology.org/W17-3518
19. Gurulingappa, H., Rajput, A.M., Roberts, A., Fluck, J., Hofmann-Apitius, M., Toldo, L.: Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics **45**(5), 885–892 (Oct 2012). https://doi.org/10.1016/j.jbi.2012.04.008
20. Gábor, K., Buscaldi, D., Schumann, A.K., QasemiZadeh, B., Zargayouna, H., Charnois, T.: SemEval-2018 Task 7: Semantic Relation Extraction and Classification in Scientific Papers. In: Proceedings of the 12th International Workshop on Semantic Evaluation. pp. 679–688. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/S18-1111, https://aclanthology.org/S18-1111
21. Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., Sun, M.: FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4803–4809. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/D18-1514, https://aclanthology.org/D18-1514
22. Hendrickx, I., Kim, S.N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., Szpakowicz, S.: SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In: Proceedings of the 5th International Workshop on Semantic Evaluation. pp. 33–38. Association for Computational Linguistics, Uppsala, Sweden (Jul 2010), https://aclanthology.org/S10-1006

23. Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., Declerck, T.: The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. Journal of Biomedical Informatics **46**(5), 914–920 (Oct 2013). https://doi.org/10.1016/j.jbi.2013.07.011
24. Hong, Z., Ward, L., Chard, K., Blaiszik, B., Foster, I.: Challenges and Advances in Information Extraction from Scientific Literature: a Review. JOM **73**(11), 3383–3400 (Nov 2021). https://doi.org/10.1007/s11837-021-04902-9, https://doi.org/10.1007/s11837-021-04902-9
25. Hope, T., Amini, A., Wadden, D., van Zuylen, M., Parasa, S., Horvitz, E., Weld, D., Schwartz, R., Hajishirzi, H.: Extracting a Knowledge Base of Mechanisms from COVID-19 Papers (Apr 2021). https://doi.org/10.48550/arXiv.2010.03824, http://arxiv.org/abs/2010.03824, arXiv:2010.03824 [cs]
26. Hou, Y., Jochim, C., Gleize, M., Bonin, F., Ganguly, D.: Identification of Tasks, Datasets, Evaluation Metrics, and Numeric Scores for Scientific Leaderboards Construction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5203–5213. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1513, https://aclanthology.org/P19-1513
27. Ifidon, E.I., Ugwuanyi, R.N.: Effective communication in academic libraries: An imperative for knowledge delivery. International Journal of Library and Information Science **5**(7), 203–207 (2013)
28. Jain, S., van Zuylen, M., Hajishirzi, H., Beltagy, I.: SciREX: A Challenge Dataset for Document-Level Information Extraction. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7506–7516. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.670, https://aclanthology.org/2020.acl-main.670
29. Kalantari, A., Kamsin, A., Kamaruddin, H.S., Ale Ebrahim, N., Gani, A., Ebrahimi, A., Shamshirband, S.: A bibliometric approach to tracking big data research trends. Journal of Big Data **4**(1), 30 (Sep 2017). https://doi.org/10.1186/s40537-017-0088-1, https://doi.org/10.1186/s40537-017-0088-1
30. Kim, J.D., Wang, Y., Yasunori, Y.: The Genia Event Extraction Shared Task, 2013 Edition - Overview. In: Proceedings of the BioNLP Shared Task 2013 Workshop. pp. 8–15. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013), https://aclanthology.org/W13-2002
31. Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks (Feb 2017). https://doi.org/10.48550/arXiv.1609.02907, http://arxiv.org/abs/1609.02907, arXiv:1609.02907 [cs, stat]
32. Krallinger, M., Valencia, A., Hirschman, L.: Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biology **9**(2), S8 (Sep 2008). https://doi.org/10.1186/gb-2008-9-s2-s8, https://doi.org/10.1186/gb-2008-9-s2-s8
33. Lahav, D., Falcon, J.S., Kuehl, B., Johnson, S., Parasa, S., Shomron, N., Chau, D.H., Yang, D., Horvitz, E., Weld, D.S., Hope, T.: A Search Engine for Discovery of Scientific Challenges and Directions (Jan 2022), http://arxiv.org/abs/2108.13751, arXiv:2108.13751 [cs]
34. Larsen, P.O., von Ins, M.: The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics **84**(3), 575–603 (2010). https://doi.org/10.1007/s11192-010-0202-z, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2909426/

35. Li, F., Li, M., Guan, P., Ma, S., Cui, L.: Mapping Publication Trends and Identifying Hot Spots of Research on Internet Health Information Seeking Behavior: A Quantitative and Co-Word Biclustering Analysis. Journal of Medical Internet Research **17**(3), e3326 (Mar 2015). https://doi.org/10.2196/jmir.3326, https://www.jmir.org/2015/3/e81, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada
36. Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database **2016**, baw068 (Jan 2016). https://doi.org/10.1093/database/baw068, https://doi.org/10.1093/database/baw068
37. Ling, M., Chen, J.: DeepPaperComposer: A Simple Solution for Training Data Preparation for Parsing Research Papers. In: Proceedings of the First Workshop on Scholarly Document Processing. pp. 91–96. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.sdp-1.10, https://www.aclweb.org/anthology/2020.sdp-1.10
38. Lippi, M., Torroni, P.: Argumentation Mining: State of the Art and Emerging Trends. ACM Transactions on Internet Technology **16**(2), 10:1–10:25 (Mar 2016). https://doi.org/10.1145/2850417, https://dl.acm.org/doi/10.1145/2850417
39. Luan, Y.: Information Extraction from Scientific Literature for Method Recommendation (Dec 2018). https://doi.org/10.48550/arXiv.1901.00401, http://arxiv.org/abs/1901.00401, arXiv:1901.00401 [cs]
40. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction (Aug 2018), http://arxiv.org/abs/1808.09602, arXiv:1808.09602 [cs] version: 1
41. Lytos, A., Lagkas, T., Sarigiannidis, P., Bontcheva, K.: The evolution of argumentation mining: From models to social media and emerging tools. Information Processing & Management **56**(6), 102055 (Nov 2019). https://doi.org/10.1016/j.ipm.2019.102055, https://www.sciencedirect.com/science/article/pii/S030645731930024X
42. Ma, Y., Dong, M., Zhou, K., Mita, C., Liu, J., Wayne, P.M.: Publication Trends in Acupuncture Research: A 20-Year Bibliometric Analysis Based on PubMed. PLOS ONE **11**(12), e0168123 (Dec 2016). https://doi.org/10.1371/journal.pone.0168123, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0168123, publisher: Public Library of Science
43. Malmasi, S., Fang, A., Fetahu, B., Kar, S., Rokhlenko, O.: SemEval-2022 Task 11: Multilingual Complex Named Entity Recognition (MultiCoNER). In: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). pp. 1412–1437. Association for Computational Linguistics, Seattle, United States (Jul 2022). https://doi.org/10.18653/v1/2022.semeval-1.196, https://aclanthology.org/2022.semeval-1.196
44. Melgoza, P., Mennel, P.A., Gyeszly, S.D.: Information overload. Collection Building **21**(1), 32–43 (Jan 2002). https://doi.org/10.1108/01604950210414706, https://doi.org/10.1108/01604950210414706, publisher: MCB UP Ltd
45. Mitchell, A., Strassel, S., Huang, S., Zakhary, R.: ACE 2004 Multilingual Training Corpus (Mar 2005). https://doi.org/10.35111/8M4R-V312, https://catalog.ldc.upenn.edu/LDC2005T09, artwork Size: 366008 KB Pages: 366008 KB

46. Mondal, I., Hou, Y., Jochim, C.: End-to-End Construction of NLP Knowledge Graph. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 1885–1895. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.165, https://aclanthology.org/2021.findings-acl.165
47. Nasar, Z., Jaffry, S.W., Malik, M.K.: Information extraction from scientific articles: a survey. Scientometrics **117**(3), 1931–1990 (Dec 2018). https://doi.org/10.1007/s11192-018-2921-5, https://doi.org/10.1007/s11192-018-2921-5
48. QasemiZadeh, B., Schumann, A.K.: The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1862–1868. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1294
49. Reis, H.T., Stiller, J.: Publication Trends in JPSP: A Three-Decade Review. Personality and Social Psychology Bulletin **18**(4), 465–472 (Aug 1992). https://doi.org/10.1177/0146167292184011, https://doi.org/10.1177/0146167292184011, publisher: SAGE Publications Inc
50. Riloff, E., Wiebe, J., Phillips, W.: Exploiting subjectivity classification to improve information extraction. In: Proceedings of the 20th national conference on Artificial intelligence - Volume 3. pp. 1106–1111. AAAI'05, AAAI Press, Pittsburgh, Pennsylvania (Jul 2005)
51. Roth, D., Yih, W.t.: A Linear Programming Formulation for Global Inference in Natural Language Tasks. In: Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004. pp. 1–8. Association for Computational Linguistics, Boston, Massachusetts, USA (May 2004), https://aclanthology.org/W04-2401
52. Samuel, A.L.: Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development **3**(3), 210–229 (Jul 1959). https://doi.org/10.1147/rd.33.0210, conference Name: IBM Journal of Research and Development
53. Sandhaus, E.: The New York Times Annotated Corpus (Oct 2008). https://doi.org/10.35111/77BA-9X74, https://catalog.ldc.upenn.edu/LDC2008T19, artwork Size: 3250585 KB Pages: 3250585 KB
54. Schlichtkrull, M., Kipf, T.N., Bloem, P., Van Den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15. pp. 593–607. Springer (2018)
55. Siegel, N., Horvitz, Z., Levin, R., Divvala, S., Farhadi, A.: FigureSeer: Parsing Result-Figures in Research Papers. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 664–680. Lecture Notes in Computer Science, Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_41
56. Spangler, S., Wilkins, A.D., Bachman, B.J., Nagarajan, M., Dayaram, T., Haas, P., Regenbogen, S., Pickering, C.R., Comer, A., Myers, J.N., Stanoi, I., Kato, L., Lelescu, A., Labrie, J.J., Parikh, N., Lisewski, A.M., Donehower, L., Chen, Y., Lichtarge, O.: Automated hypothesis generation based on mining scientific literature. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1877–1886. KDD '14, Association for Computing Machinery, New York, NY, USA (Aug

2014). https://doi.org/10.1145/2623330.2623667, https://dl.acm.org/doi/10.1145/2623330.2623667

57. Stoica, G., Platanios, E.A., Póczos, B.: Re-TACRED: Addressing Shortcomings of the TACRED Dataset (Apr 2021). https://doi.org/10.48550/arXiv.2104.08398, http://arxiv.org/abs/2104.08398, arXiv:2104.08398 [cs]
58. Taboureau, O., Nielsen, S.K., Audouze, K., Weinhold, N., Edsgärd, D., Roque, F.S., Kouskoumvekaki, I., Bora, A., Curpan, R., Jensen, T.S., Brunak, S., Oprea, T.I.: ChemProt: a disease chemical biology database. Nucleic Acids Research **39**(Database issue), D367–372 (Jan 2011). https://doi.org/10.1093/nar/gkq906
59. Tan, Q., Xu, L., Bing, L., Ng, H.T., Aljunied, S.M.: Revisiting DocRED – Addressing the False Negative Problem in Relation Extraction (Jun 2023), http://arxiv.org/abs/2205.12696, arXiv:2205.12696 [cs] version: 3
60. Venugopal, V., Pai, S., Olivetti, E.: MatKG: The Largest Knowledge Graph in Materials Science – Entities, Relations, and Link Prediction through Graph Representation Learning (Oct 2022). https://doi.org/10.48550/arXiv.2210.17340, http://arxiv.org/abs/2210.17340, arXiv:2210.17340 [cond-mat]
61. Wadden, D., Wennberg, U., Luan, Y., Hajishirzi, H.: Entity, Relation, and Event Extraction with Contextualized Span Representations. 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. 5784–5789. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1585, https://aclanthology.org/D19-1585
62. Walker, C., Strassel, S., Medero, J., Maeda, K.: ACE 2005 Multilingual Training Corpus (Feb 2006). https://doi.org/10.35111/MWXC-VH88, https://catalog.ldc.upenn.edu/LDC2006T06, artwork Size: 1572864 KB Pages: 1572864 KB
63. Ware, M., Mabe, M.: The STM Report: An overview of scientific and scholarly journal publishing. Copyright, Fair Use, Scholarly Communication, etc. (Mar 2015), https://digitalcommons.unl.edu/scholcom/9
64. Wiebe, J., Riloff, E.: Finding Mutual Benefit between Subjectivity Analysis and Information Extraction. IEEE Transactions on Affective Computing **2**(4), 175–191 (Oct 2011). https://doi.org/10.1109/T-AFFC.2011.19, conference Name: IEEE Transactions on Affective Computing
65. Wilson, S.J., Wilkins, A.D., Holt, M.V., Choi, B.K., Konecki, D., Lin, C.H., Koire, A., Chen, Y., Kim, S.Y., Wang, Y., Wastuwidyaningtyas, B.D., Qin, J., Donehower, L.A., Lichtarge, O.: Automated literature mining and hypothesis generation through a network of Medical Subject Headings (Aug 2018). https://doi.org/10.1101/403667, https://www.biorxiv.org/content/10.1101/403667v1, pages: 403667 Section: New Results
66. Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., Sun, M.: DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 764–777. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1074, https://aclanthology.org/P19-1074
67. Zhang, Y., Zhong, V., Chen, D., Angeli, G., Manning, C.D.: Position-aware Attention and Supervised Data Improve Slot Filling. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 35–45. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1004, https://aclanthology.org/D17-1004

# A Appendix: Tables & Figures

**Table 5.** Summary of existing schemas and datasets for joint NER and RE from research papers. Along with indications of the number of different entities and relations they have. The size is the number of annotated items.

| Dataset | Domain | Size | Entities | Relations | Limitations |
|---|---|---|---|---|---|
| SciERC [40] | CS | 500 abstracts | 6 | 7 | Coarse-grained, domain-specific |
| MECHANIC-Coarse[25] | Bio | 1000 sentences | 1 | 2 | na |
| MECHANIC-Granular[25] | Bio | 1000 sentences | N/A$^{??}$ | N/A$^{??}$ | na |
| ChemProt [58] | Bio | 1,820 abstracts | 2 | 12 | Imbalanced, domain-specific |
| ADE [19] | Bio | 6,821 sentences | 2 | 1 | Sparse, noisy, narrow |
| DDI [23] | Bio | 1017 abstract | 4 | 4 | Imbalanced, domain-specific |
| CoNLL04 [51] | News | 1,441 sentences | 4 | 5 | Low agreement, general |
| ACE05 [62] | Various | 511 documents | 7 | 6 | Complex, general |
| ACE-Event [61] | Various | 599 documents | 6 | 18 | Complex, general |
| GENIA [30] | Bio | 2,000 abstracts | 6 | 5 | Fine-grained, domain-specific |
| ACE04 [45] | Various | 451 documents | 7 | 6 | Complex, general |
| DocRED [66] | Various | 5,053 documents / 101k | 9 | 96 | Document-level |
| TACRED [67] | Various | 106k sentences | N/A | 41 | Sentence-level |
| SemEval-2010 Task 8 [22] | Various | 10,717 sentences | N/A | 8 | Sentence-level |
| WebNLG [18] | Various | 21,855 data/text pairs | N/A | N/A | RDF-to-text |
| NYT [53] | Various | 2,15M annotated docs | N/A | N/A | Summarization |
| GDA [3] | Bio | 30k sentences | 3 | 1 | Gene-disease association |
| BC5CDR [36] | Bio | 1,500 articles | 2 | 1 | Chemical-disease relation |
| Retacred [57] | Various | 106K sentences | N/A | 41 | Biomedical relation |
| Redocred [59] | Various | 5K documents | 9 | 96 | Document-level biomedical relation |
| FewREL [21] | Various | 70,000 sentences | N/A | 100 | Few-shot relation classification |
| KPI-EDGAR [11] | Financial | 1355 sentences | 12 | 4 | Joint NER and RE |
| T-REx [14] | Various | 6.2m sentences | N/A | 642 | Semantic Relation Classification |
| ACL RD-TEC 2.0 [48] | CL | 300 abstracts | 7 | 0 | 0 |
| SemEval 2017 Task 10: ScienceIE [2] | Various | 500 abstracts | 3 | 2 | 0 |
| SemEval-2018 Task 7 [20] | CL | 500 abstracts | 0 | 6 | 0 |
| ARC-PDN [26] | Various | 4k docs | 4 | 1 | 0 |
| STEM-ECR v1.0 [13] | ML | 332 abstracts | 4 | 0 | 0 |
| NLP TDMS [46] | AI | 30k docs | 3 | 4 | - |
| SciREX [28] | CL, Computer Vision | 438 docs | 4 | 1 | - |
| AI-KG [10] | CS, ML | 333k | 5 | 9 | - |
| SemEval-2021 Task 11 [12] | CL | 442 docs | 1 | 0 | - |

**Table 6.** Local graph topology metrics. For OR mode. Relatively high values are marked blue with relatively low values being marked red. Density is reported with 5 decimal precision and other factors with 3 decimals unless they are natural numbers. Some statistics for these small subgraphs are on node-level. This means that the presented statistic is an aggregation of an aggregation (for example the mean of the mean centralities of all subgraphs). This statistic could have been represented by listing the properties for every node and obtaining the statistic from that, but it is intentionally represented by aggregating the statistic on the subgraph level and obtaining the statistic over the different subgraphs to give an improved impression of the subgraph properties rather than the properties of the nodes in them. As a consequence, not all node statistics are weighed equally since some subgraphs consist of more nodes and other subgraphs consist of less nodes. Relations are after the addition of entity type relations.

| | ACE-Event | ACE05 | GENIA | MECHANIC-Coarse | MECHANIC-Granular | SciERC |
|---|---|---|---|---|---|---|
| Entities | *4.730* ± 3.240 | 5.131 ± 3.551 | *3.386* ± 2.223 | **9.118** ± 9.513 | 6.206 ± 3.535 | **7.992** ± 6.389 |
| Relations | *2.243* ± 1.517 | 2.430 ± 1.617 | *1.625* ± 1.033 | **4.230** ± 3.542 | 3.062 ± 1.727 | **3.951** ± 3.143 |
| Degrees | 0.948 ± 0.729 | *0.947* ± 0.784 | 0.960 ± 0.669 | *0.928* ± 1.346 | **0.987** ± 1.099 | **0.989** ± 1.032 |
| Degree centralities | **0.311** ± 0.329 | 0.278 ± 0.308 | **0.494** ± 0.403 | *0.146* ± 0.245 | 0.221 ± 0.286 | *0.164* ± 0.222 |
| Closeness centralities | **0.336** ± 0.324 | 0.307 ± 0.303 | **0.522** ± 0.388 | *0.162* ± 0.250 | 0.243 ± 0.287 | *0.187* ± 0.225 |
| Clusterings | 0.012 ± 0.103 | *0.011* ± 0.100 | *0.000* ± 0.000 | **0.204** ± 0.374 | **0.275** ± 0.434 | 0.080 ± 0.251 |
| Modularities | 0.185 ± 0.236 | **0.201** ± 0.243 | 0.071 ± 0.174 | *0.029* ± 0.075 | *0.023* ± 0.060 | **0.290** ± 0.229 |
| Densities | **0.505** ± 0.380 | 0.461 ± 0.368 | **0.697** ± 0.371 | *0.345* ± 0.362 | 0.379 ± 0.359 | *0.321* ± 0.328 |

**Table 7.** The different entity and relation types defined by each underlying schema. We can observe that some schemas employ much more generic entity types than others e.g. ”Entity” vs ”Cell type”

| Schema | SciERC | M-Coarse | M-Granular | ACE05 | ACE-Event | GENIA |
|---|---|---|---|---|---|---|
| Entity Types | Task<br>Method<br>Metric<br>Material<br>Other-ScientificTerm<br>Generic | Entity | Entity | Person<br>Organization<br>Geo-Political Entity<br>Location<br>Facility<br>Vehicle<br>Weapon | Person<br>Organization<br>Geo-Political Entity<br>Location<br>Facility<br>Time<br>Weapon | Protein<br>DNA<br>RNA<br>Cell line<br>Cell type<br>Other |
| Relation Types | Compare<br>Part-of<br>Conjunction<br>Evaluate-for<br>Feature-of<br>Used-for<br>Hyponym-of | Mechanisms<br>Effect | - (OpenIE) | Physical (PHYS)<br>Part-Whole(PART-WHOLE)<br>Artifact(ART)<br>General Affiliation(GEN-AFF)<br>Organization Affiliation(ORG-AFF)<br>Personal Social(PER-SOC) | ORG-AFF.Employment<br>PHYS.Located<br>PART-WHOLE.Geographical<br>ART.User-Owner-Inventor-Manufacturer<br>GEN-AFF.Citizen-Resident-Religion-Ethnicity<br>ORG-AFF.Membership<br>PART-WHOLE.Subsidiary<br>GEN-AFF.Org-Location<br>PHYS.Near<br>PER-SOC.Family<br>PER-SOC.Business<br>ORG-AFF.Founder<br>ORG-AFF.Sports-Affiliation<br>ORG-AFF.Investor-Shareholder<br>ORG-AFF.Student-Alum<br>ORG-AFF.Ownership<br>PER-SOC.Lasting-Personal<br>PART-WHOLE.Artifact | Component-of<br>Subunit-of<br>Site-of<br>Product-of<br>Theme-of |

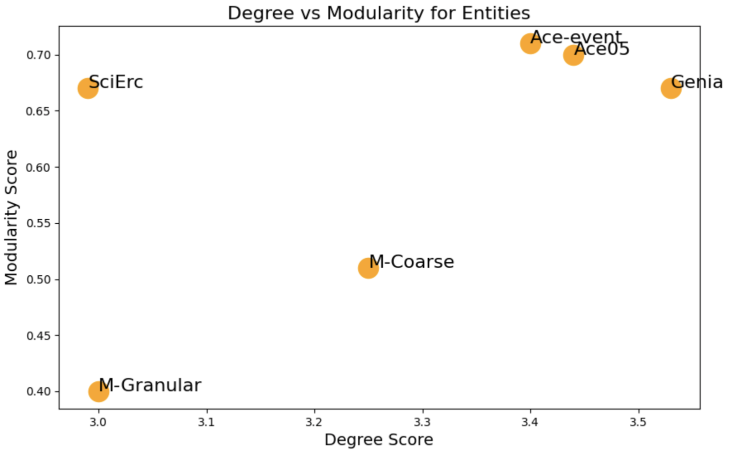

**Fig. 3.** Global spectrum of major graph topology metric

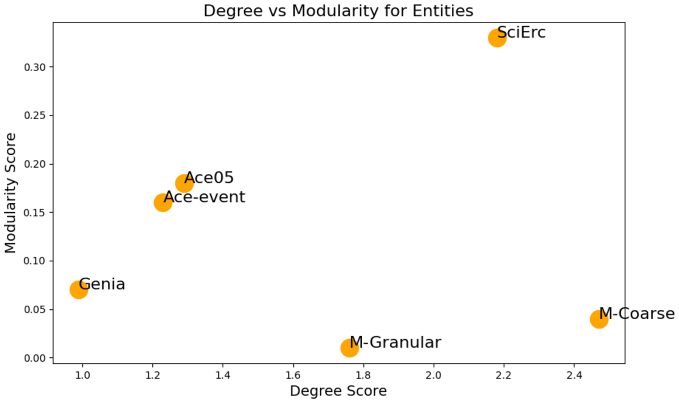

**Fig. 4.** Local spectrum of major graph topology metrics