



# Domain-Independent Extraction of Scientific Concepts from Research Articles

Arthur Brack<sup>1</sup>(✉)(iD), Jennifer D'Souza<sup>1</sup>(✉)(iD), Anett Hoppe<sup>1</sup>(✉)(iD),  
Sören Auer<sup>1</sup>(✉)(iD), and Ralph Ewerth<sup>1,2</sup>(✉)(iD)

<sup>1</sup> TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany  
{arthur.brack,jennifer.dsouza,anett.hoppe,soeren.auer  
ralph.ewerth}@tib.eu

<sup>2</sup> L3S Research Center, Leibniz University, Hannover, Germany

**Abstract.** We examine the novel task of *domain-independent scientific concept extraction from abstracts of scholarly articles* and present two contributions. First, we suggest a set of generic scientific concepts that have been identified in a systematic annotation process. This set of concepts is utilised to annotate a corpus of scientific abstracts from 10 domains of Science, Technology and Medicine at the phrasal level in a joint effort with domain experts. The resulting dataset is used in a set of benchmark experiments to (a) provide baseline performance for this task, (b) examine the transferability of concepts between domains. Second, we present a state-of-the-art deep learning baseline. Further, we propose the active learning strategy for an optimal selection of instances from among the various domains in our data. The experimental results show that (1) a substantial agreement is achievable by non-experts after consultation with domain experts, (2) the baseline system achieves a fairly high F1 score, (3) active learning enables us to nearly halve the amount of required training data.

**Keywords:** Sequence labelling · Information extraction · Scientific articles · Active learning · Scholarly communication · Research knowledge graph

## 1 Introduction

Scholarly communication as of today is a document-centric process. Research results are usually conveyed in written articles, as a PDF file with text, tables and figures. Automatic indexing of these texts is limited and generally does not access their semantic content. There are thus severe limitations how current research infrastructures can support scientists in their work: finding relevant research works, comparing them, and compiling summaries is still a tedious and error-prone manual work. The heightened increase in the number of published research papers aggravates this situation [7].

Knowledge graphs are recognised as an effective approach to facilitate semantic search [3]. For academic search engines, Xiong et al. [47] have shown that

exploiting knowledge bases like Freebase can improve search results. However, the introduction of new scientific concepts occurs at a faster pace than knowledge base curation, resulting in a large gap in knowledge base coverage of scientific entities [1], e.g. the task *geolocation estimation of photos* from the Computer Vision field is neither present in Wikipedia nor in more specialised knowledge bases like Computer Science Ontology (CSO) [39] or “Papers with code” [36]. Information extraction from text helps to identify emerging entities and to populate knowledge graphs [3]. It then is a first vital step towards a fine-grained research knowledge graph in which research articles are described and interconnected through entities like tasks, materials, and methods. Our work is motivated by the idea of the automatic construction of a research knowledge graph.

Information extraction from scientific texts, obviously, differs from its general domain counterpart: Understanding a research paper and determining its most important statements demands certain expertise in the article’s domain. Every domain is characterised by its specific terminology and phrasing which is hard to grasp for a non-expert reader. In consequence, extraction of scientific concepts from text would entail the involvement of domain experts and a specific design of an extraction methodology for each scientific discipline – both requirements are rather time-consuming and costly.

At present, a systematic study of these assumptions is missing. We thus present the task of *domain-independent scientific concept extraction*. We examine the intuition that most research papers share certain core concepts such as the mentions of research tasks or methods. If so, these would allow a domain-independent information extraction system to support populating a research knowledge graph, which does not reach all semantic depths of the analysed article, but still provides some science-specific structure.

In this paper, we introduce a set of common scientific concepts that we find are relevant over a set of 10 examined domains from Science, Technology, and Medicine (STM). These generic concepts have been identified in a systematic, joint effort of domain experts and non-domain experts. The inter-coder agreement is measured to ensure the adequacy and quality of concepts. A set of research abstracts has been annotated using these concepts and the results are discussed with experts from the corresponding fields. The resulting dataset serves as a basis to train two baseline deep learning classifiers. In particular, we present an active learning approach to reduce the number of required training data. The systems are evaluated in different experimental setups.

Our main contributions can be summarised as follows: (1) We introduce the novel task *domain-independent scientific concept extraction*, which aims at automatically extracting scientific entities in a domain-independent manner. (2) We release a new corpus that comprises 110 abstracts of 10 STM domains annotated at the phrasal level. (3) We present and evaluate a state-of-the-art deep learning approach for this task. Additionally, we employ active learning for an optimal selection of instances, which to our knowledge, is demonstrated for the first time on scholarly text. We find that strategic instance selection gives us the same performance with only about half of the training data. (4) We release a silver-labelled corpus

with 62K automatically annotated abstracts of Elsevier with CCBY license and 1.2 Mio. extracted unique concepts comprising 24 domains. (5) We make our corpora and source code publicly available to facilitate further research.

## 2 Related Work

This section gives a brief overview of existing annotated datasets for scientific information extraction, followed by related work on some exemplary applications for domain-independent information extraction from scientific papers.

### 2.1 Scientific Corpora

**Sentence Level Annotation.** Early approaches for semantic structuring of research papers focused on sentences as the basic unit of analysis. This enables, for instance, automatic highlighting of relevant paper passages to enable efficient assessment regarding quality and relevance. Several ontologies have been created that focus on the rhetorical [11, 19], argumentative [31, 46] or activity-based [37] structure of research papers.

Annotated datasets exist for several domains, e.g. PubMed200k [12] from biomedical randomized controlled trials, NICTA-PIBOSO [26] from evidence-based medicine, Dr. Inventor [15] from Computer Graphics, Core Scientific Concepts (CoreSC) [31] from Chemistry and Biochemistry, and Argumentative Zoning (AZ) [46] from Chemistry and Computational Linguistics, Sentence Corpus [8] from Biology, Machine Learning and Psychology. Most datasets cover only a single domain, while few other datasets cover three domains. Several machine learning methods have been proposed for scientific sentence classification [12, 15, 24, 30].

**Phrase Level Annotation.** More recent corpora have been annotated at phrasal level (e.g. noun phrases). SciCite [9] and ACL ARC [25] are datasets for citation intent classification from Computer Science, Medicine, and Computational Linguistics. ACL RD-TEC [20] from Computational Linguistics aims at extracting scientific technology and non-technology terms. ScienceIE-17 [2] from Computer Science, Material Sciences, and Physics contains three concepts PROCESS, TASK and MATERIAL. SciERC [32] from the machine learning domain contains six concepts TASK, METHOD, METRIC, MATERIAL, OTHER-SCIENTIFIC-TERM and GENERIC. Each corpus covers at most three domains.

**Experts vs. Non-experts.** The aforementioned datasets were usually annotated by domain experts [2, 12, 20, 26, 31, 32]. In contrast, Teufel et al. [46] explicitly use non-experts in their annotation tasks, arguing that text understanding systems can use general, rhetorical and logical aspects also when qualifying scientific text. According to this line of thought, more researchers used (presumably cheaper) non-expert annotation as an alternative [8, 15].

Snow et al. [43] provide a study on expert versus non-expert performance for general, non-scientific annotation tasks. They state that about four non-experts (Mechanical Turk workers, in their case) were needed to rival the experts' annotation quality. However, systems trained on data generated by non-experts showed to benefit from annotation diversity and to suffer less from annotator bias. A recent study [38] examines the agreement between experts and non-experts for visual concept classification and person recognition in historical video data. For the task of face recognition, training with expert annotations lead to an increase of only 1.5% in classification accuracy.

**Active Learning in Natural Language Processing (NLP).** To the best of our knowledge, active learning has not been utilised in classification approaches for scientific text yet. Recent publications demonstrate the effectiveness of active learning for NLP tasks such as *Named Entity Recognition* (NER) [41] and sentence classification [49]. Siddhant and Lipton [42] and Shen et. al. [41] compare several sampling strategies on NLP tasks and show that *Maximum Normalized Log-Probability* (MNLP) based on uncertainty sampling performs well in NER.

## 2.2 Applications for Domain-Independent Scientific Information Extraction

**Academic Search Engines.** Academic search engines such as Google Scholar [18], Microsoft Academic [34] and Semantic Scholar [40] specialise in search of scholarly literature. They exploit graph structures such as the Microsoft Academic Knowledge Graph [35], SciGraph [45], or the Semantic Scholar Corpus [1]. These graphs interlink the papers through meta-data such as citations, authors, venues, and keywords, but not through deep semantic representation of the articles' content.

However, first attempts towards a more semantic representation of article content exist: Ammar et al. [1] interlink the Semantic Scholar Corpus with DBpedia [29] and Unified Medical Language System (UMLS) [6] using entity linking techniques. Yaman et al. [48] connect SciGraph with DBpedia person entities. Xiong et al. [47] demonstrate that academic search engines can greatly benefit from exploiting general-purpose knowledge bases. However, the coverage of science-specific concepts is rather low [1].

**Research Paper Recommendation Systems.** Beel et al. [4] provide a comprehensive survey about research paper recommendation systems. Such systems usually employ different strategies (e.g. content-based and collaborative filtering) and several data sources (e.g. text in the documents, ratings, feedback, stereotyping). Graph-based systems, in particular, exploit citation graphs and genes mentioned in the papers [27]. Beel et al. conclude that it is not possible to determine the most effective recommendation approach at the moment. However, we believe that a fine-grained research knowledge graph can improve such systems. Although "Papers with code" [36] is not a typical recommendation system, it allows researchers to browse easily for papers from the field of machine learning that address a certain task.

### 3 Corpus for Domain-Independent Scientific Concept Extraction

In this section, we introduce the novel task of *domain-independent extraction of scientific concepts* and present an annotated corpus. As the discussion of related work reveals, the annotation of scientific resources is not a novel task. However, most researchers focus on at most three scientific disciplines and on expert-level annotations. In this work, we explore the domain-independent annotation of lexical phrasal units indicating scientific knowledge, i.e. scientific concepts, in abstracts from ten different science domains. Since other studies have also shown that non-expert annotations are feasible for the scientific domain, we go for a cost-efficient middle course: annotations by non-experts with scientific proficiency, and consultation with domain-experts. Finally, we explore how well a state-of-the-art deep learning model performs on this novel information extraction task and whether active learning can help to reduce the amount of required training data. Our novel corpus and the annotation process are described below.

#### 3.1 OA-STM Corpus

The OA-STM corpus [14] is a set of open access (OA) articles from various domains in Science, Technology and Medicine (STM). It was published in 2017 as a platform for benchmarking methods in scholarly article processing, amongst other scientific information extraction. The dataset contains a selection of 110 articles from 10 domains, namely Agriculture (*Agr*), Astronomy (*Ast*), Biology (*Bio*), Chemistry (*Che*), Computer Science (*CS*), Earth Science (*ES*), Engineering (*Eng*), Materials Science (*MS*), Mathematics (*Mat*), and Medicine (*Med*). This first annotation cycle focuses on the articles' abstracts as they contain a condensed summary of the article.

#### 3.2 Annotation Process

The OA-STM Corpus is used as a base for (a) the identification of potential domain-independent concepts; (b) a first annotated corpus for baseline classification experiments. The annotation task was mainly performed by two post-doctoral researchers with a background in Computer Science (acting as non-expert annotators); their basic annotation assumptions were checked by domain experts.

**Pre-annotation.** A literature review of annotation schemes [2, 11, 30, 31] provided a seed set of potential candidate concepts. Both non-experts independently annotated a subset of the STM abstracts with these concepts (non-overlapping) and discussed the outcome. In a three-step process, the concept set was pruned to only contain those which seemed suitably transferable between domains. Our set of *generic* scientific concepts consists of PROCESS, METHOD, MATERIAL, and DATA (see Table 1 for their definitions). We also identified TASK [2], OBJECT [30],

**Table 1.** The four core scientific concepts that were derived in this study

<b>PROCESS</b>	Natural phenomenon or activities, e.g. growing ( <i>Bio</i> ), reduction ( <i>Mat</i> ), flooding ( <i>ES</i> )
<b>METHOD</b>	A commonly used procedure that acts on entities, e.g. powder X-ray ( <i>Che</i> ), the PRAM analysis ( <i>CS</i> ), magnetoencephalography ( <i>Med</i> )
<b>MATERIAL</b>	A physical or abstract entity used in scientific experiments or proofs, e.g. soil ( <i>Agr</i> ), the moon ( <i>Ast</i> ), the carbonator ( <i>Che</i> )
<b>DATA</b>	The data themselves, measurements, or quantitative or qualitative characteristics of entities, e.g. rotational energy ( <i>Eng</i> ), tensile strength ( <i>MS</i> ), 3D time-lapse seismic data ( <i>ES</i> )

and RESULTS [11], however, in this study we do not consider nested span concepts, hence we leave them out since they were almost always nested with the other scientific entities (e.g. a RESULT may be nested with DATA).

**Phase I.** Five abstracts per domain (i.e. 50 abstracts) were annotated by both annotators and the inter-annotator agreement was computed using Cohen’s  $\kappa$  [10] at exact annotated spans. Results showed a moderate inter-annotator agreement of 0.52  $\kappa$ .

**Phase II.** The annotations were then presented to subject specialists who each reviewed (a) the choice of concepts and (b) annotation decisions on the respective domain corpus. The interviews mostly confirmed the concept candidates as generally applicable. The experts’ feedback on the annotation was even more valuable: The comments allowed for a more precise reformulation of the annotation guidelines, including illustrating examples from the corpus.

**Consolidation.** Finally, the 50 abstracts from phase I were reannotated by the non-experts. Based on the revised annotation guidelines, a substantial agreement of 0.76  $\kappa$  could be reached (see Table 2). Similar annotation tasks for scientific entities, i.e. SciERC [32] considering one domain and ScienceIE-17 [2] considering three domains achieved agreements of 0.76  $\kappa$  and 0.6  $\kappa$ , respectively. Subsequently, the remaining 60 abstracts (six per domain) were annotated by one annotator. This phase also involved reconciliation of the previously annotated 50 abstracts to obtain a gold standard corpus.

**Table 2.** Per-domain and overall inter-annotator agreement (Cohen’s Kappa  $\kappa$ ) for PROCESS, METHOD, MATERIAL, and METHOD scientific concept annotation

	<i>Med</i>	<i>MS</i>	<i>CS</i>	<i>ES</i>	<i>Eng</i>	<i>Che</i>	<i>Bio</i>	<i>Agr</i>	<i>Mat</i>	<i>Ast</i>	<i>Overall</i>
$\kappa$	0.94	0.90	0.85	0.81	0.79	0.77	0.75	0.60	0.58	0.57	0.76

### 3.3 Corpus Characteristics

Table 3 shows some characteristics of the resulting corpus. The corpus has a total of 6,127 scientific entities, including 2,112 PROCESS, 258 METHOD, 2,099 MATERIAL, and 1,658 DATA concept entities. The number of entities per abstract in our corpus directly correlates with the length of the abstracts (Pearson’s  $R$  0.97). Among the concepts, PROCESS and MATERIAL directly correlate with abstract length ( $R$  0.8 and 0.83, respectively), while DATA has only a slight correlation ( $R$  0.35) and METHOD has no correlation ( $R$  0.02). The domains *Bio*, *CS*, *Ast*, and *Eng* contain the most of PROCESS, METHOD, MATERIAL, and DATA concepts, respectively.

**Table 3.** The annotated corpus characteristics containing 11 abstracts per domain in terms of size and the number of scientific concept phrases

	<i>Ast</i>	<i>Agr</i>	<i>Eng</i>	<i>ES</i>	<i>Bio</i>	<i>Med</i>	<i>MS</i>	<i>CS</i>	<i>Che</i>	<i>Mat</i>
Avg. # Tokens/Abstract	382	333	303	321	273	274	282	253	217	140
# PROCESS	241	252	248	243	281	244	178	220	149	56
# METHOD	19	28	27	9	15	33	27	66	27	7
# MATERIAL	296	292	208	249	291	191	231	102	188	51
# DATA	235	169	258	197	62	132	138	165	119	183
# Gold scientific concept phrases	791	741	741	698	649	600	574	553	483	297
# Unique gold scientific concept phrases	663	631	618	633	511	518	493	482	444	287

## 4 Automatic Domain-Independent Scientific Concept Extraction

The current state-of-the-art for scientific entity extraction is Beltagy et al.’s deep learning system with SciBERT word embeddings [5], which were pre-trained on scientific texts using the BERT [13] architecture. It consists of three components: (a) a token embedding layer comprising a per-sentence sequence of tokens, where each token is represented as a concatenation of SciBERT word embedding and CNN-based character embeddings [33], (b) a token-level encoder with two stacked bidirectional LSTMs [21], and (c) a Conditional Random Field (CRF) based tag decoder [33] with BILOU (beginning, inside, last, outside, unit) tagging scheme. This deep learning architecture is implemented in AllenNLP [17] and uses spaCy [44] for text preprocessing, i.e. for tokenisation and sentence-splitting.

### 4.1 Supervised Learning with Full Training Dataset

Using the above mentioned architecture, we train one model with data from all domains combined. We refer to this model as the *domain-independent* classifier.

Similarly, we train 10 models for each domain in our corpus – the *domain-specific* classifier.

To obtain a robust evaluation of models, we perform five-fold cross-validation experiments. In each fold experiment, we train a model on 8 abstracts per domain (i.e. 80 abstracts), tune hyperparameters on 1 abstract per domain (i.e. 10 abstracts), and test on the remaining 2 abstracts per domain (i.e. 20 abstracts) ensuring that the data splits are not identical between the folds. All results reported in the paper are averaged over the five folds. We still obtain reliably trained domain-specific classifiers since on average they are trained on 400 concepts.

## 4.2 Active Learning with Training Data Subset

In this setting, we employ an active learning strategy [42,49] to train a new *domain-independent* classifier. Active learning is usually applied to determine the optimal set of sufficiently distinct instances to minimise annotation costs. With our application of active learning we find which proportion of our annotations suffice for training a robust classifier. We decide to use the MNLP [41] sampling strategy. We prefer it over its contemporary, Bayesian Active Learning by Disagreement (BALD) [22], since it has less computational requirements. The MNLP objective involves greedy sampling of sentences preferring those with the least logarithmic likelihood of the predicted tag sequence output by the CRF tag decoder, normalised by the number of tokens to avoid preferring longer sentences. In our experiments, we found that adding 4% of the data to be the most discriminative selection of classifier performance. Therefore, we run 25 iterations of active learning in each stage adding 4% training data. We perform five-fold cross validation as before and the per-fold models are retrained after data resampling.

# 5 Experimental Results and Discussion

In this section, we discuss the results obtained with our trained classifiers and the correlation analysis between inter-annotator agreement and performance of the classifiers.

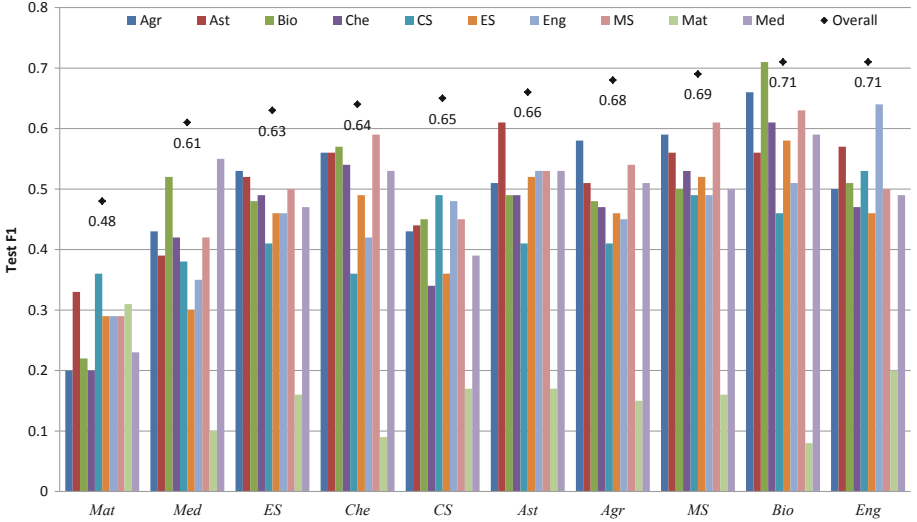
## 5.1 Domain-Independent and Domain-Specific Classifiers: Full Training Dataset

Table 4 shows an overview of the *domain-independent* classifier results. The system achieves an overall  $F1$  of 65.5% and has low standard deviation 1.26 across the five folds. For this classifier, MATERIAL was the easiest concept with an  $F1$  of 71% ( $\pm 1.88$ ), whereas METHOD was the hardest concept with an  $F1$  of 43% ( $\pm 6.30$ ). METHOD is also the most underrepresented in our corpus, which partly explains the poor extraction performance. Best reported results for similar datasets, ScienceIE17 [2] and SciERC [32] (both have 500 abstracts), have



**Table 4.** The *domain-independent* classifier results in terms of Precision ( $P$ ), Recall ( $R$ ), and F1-score on scientific concepts, respectively, and *Overall*

	PROCESS	METHOD	MATERIAL	DATA	<i>Overall</i>
$P$	65.5 ( $\pm 4.22$ )	45.8 ( $\pm 13.50$ )	69.2 ( $\pm 3.55$ )	60.3 ( $\pm 4.14$ )	64.3 ( $\pm 1.73$ )
$R$	68.3 ( $\pm 1.93$ )	44.1 ( $\pm 8.73$ )	73.2 ( $\pm 4.27$ )	60.0 ( $\pm 4.84$ )	66.7 ( $\pm 0.92$ )
$F1$	66.8 ( $\pm 2.07$ )	43.0 ( $\pm 6.30$ )	71.0 ( $\pm 1.88$ )	59.8 ( $\pm 1.75$ )	65.5 ( $\pm 1.26$ )

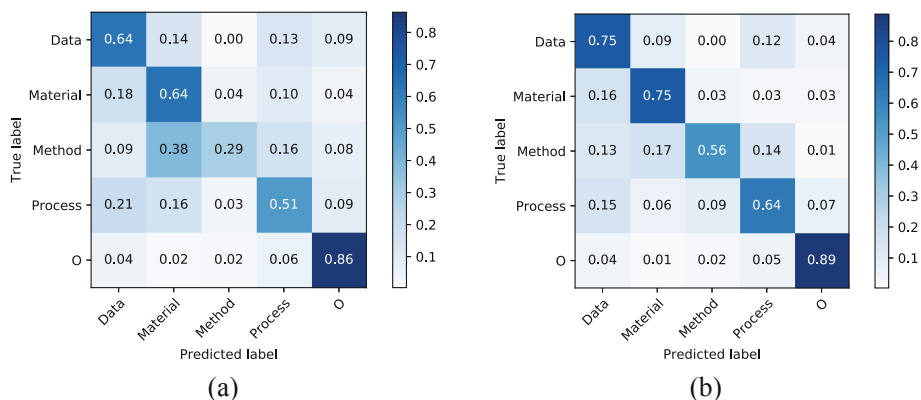
**Fig. 1.**  $F1$  per domain of the 10 *domain-specific* classifiers (as bar plots) and of the *domain-independent* classifier (as scatter plots) for scientific concept extraction; the x-axis represents the 10 test domains

an  $F1$  score of 65.6% [5] and 44.7% [32], respectively, indicating that the size of our dataset with only 110 abstracts is sufficient.

Next, we compare and contrast the 10 *domain-specific* classifiers (see Fig. 1) by their capability to extract the concepts from their own domains and in other domains.

**Most Robust Domain.** *Bio* (third bar in each domain in Fig. 1) extracts scientific concepts from its own domain at the same performance as the *domain-independent* classifier with an  $F1$  score of 71% ( $\pm 9.0$ ) demonstrating a robust domain. It comprises only 11% of the overall data, yet the *domain-independent* classifier trained on all data does not outperform it.

**Most Generic Domain.** *MS* (the third last bar in each domain in Fig. 1) exhibits a high degree of domain independence since it is among the top 3 classifiers for seven of the 10 domains (viz. *ES*, *Che*, *CS*, *Ast*, *Agr*, *MS*, and *Bio*).



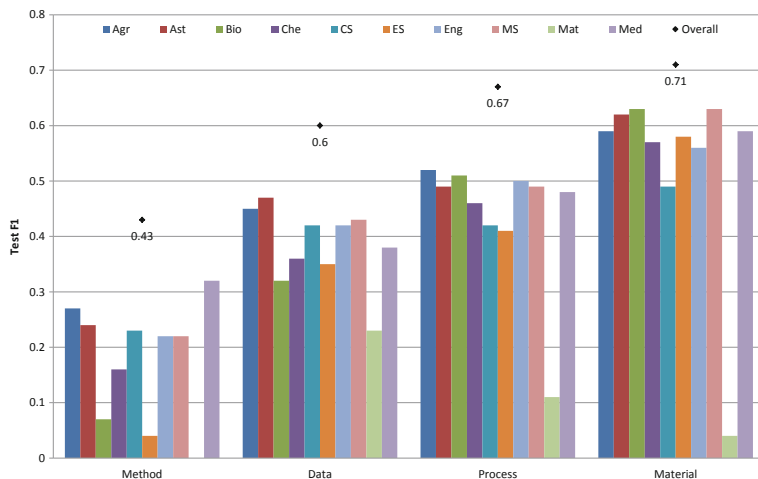
**Fig. 2.** Confusion matrix for (a) the *CS* classifier and (b) *domain-independent classifier* on *CS* domain predicting concept-type of tokens

**Most Specialised Domain.** *Mat* (the second last bar in each domain in Fig. 1) shows the lowest performance in extracting scientific concepts from all domains except itself. Hence it shows to be the most specialised domain in our corpus. Notably, a characteristic feature of this domain is that it has short abstracts (nearly a third of the size of the longest abstracts), so it is also the most under-represented in our corpus. Also, distinct from the other domains, *Mat* has triple the number of DATA entities compared to each of its other concepts, where in the other domains PROCESS and MATERIAL are consistently predominant.

**Medical and Life Science Domains.** The *Med*, *Agr*, and *Bio* domains show strong domain relatedness. Their respective *domain-specific* classifiers show top five system performances among the three domains, when applied to another domain. For instance, the *Med* domain shows the strongest domain relatedness and is classified best by *Med* (last bar), followed by *Bio* (third bar) and *Agr* (first bar).

**Domain-Independent vs. Domain-Specific Classifier.** Except for *Bio* the *domain-independent* classifier clearly outperforms the *domain-specific* one in extracting concepts from their respective domains. We attribute this, in part, to the improved span-detection performance. Span-detection merely relies on syntactic regularity, thus the *domain-independent* classifier can benefit from more training data of other domains. E.g., the *CS* classifier shows a relative improvement of 49.5% *domain-specific* *F1* score to 65.9% in the *domain-independent* setting, which is supported by the enhanced span-detection performance from 73.4% to 82.0% in *F1*. Accuracy on token-level also improves from 67.7% to 77.5% *F1* for *CS*, that is correct labelling of the tokens also benefits from other domains. This is also supported by the results in the confusion matrix depicted in Fig. 2 for the *CS* and the *domain-independent* classifier on token-level.

**Scientific Concept Extraction.** Figure 3 depicts the 10 *domain-specific* classifier results for extracting each of the four scientific concepts. It can be observed that *Agr*, *Med*, *Bio*, and *Ast* classifiers are the best in extracting their respective PROCESS, METHOD, MATERIAL, and DATA concepts.



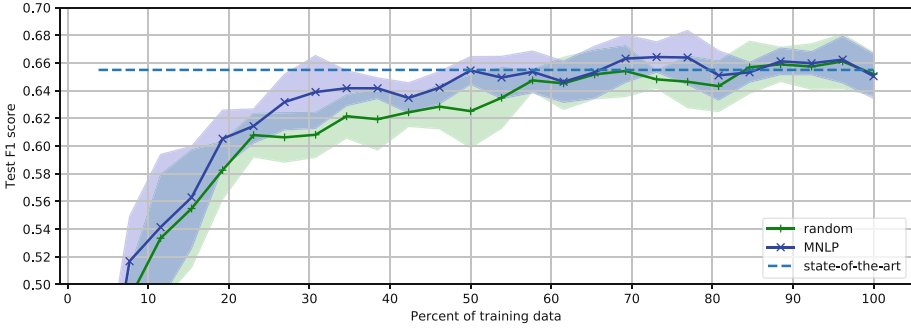
**Fig. 3.** *F1* scores of the 10 *domain-specific* classifiers (bar plots) and the *domain-independent* classifiers (scatter plots) for extracting each scientific concept; the x-axis represents the evaluated concepts

## 5.2 Domain-Independent Classifier with Active Learning

The results of the active learning experiment over the full dataset plotted over the 25 iterations are depicted in Fig. 4, showing that MNLP clearly outperforms the random baseline. While using only 52% of the training data, the best result of the *domain-independent* classifier trained with all training data is surpassed with an *F1* score of 65.5% ( $\pm 1.0$ ). The random baseline achieves an *F1* score of only 62.5% ( $\pm 2.6$ ) with the same proportion of training data. When 76% of the data are sampled by MNLP, the best active learning performance across all steps is achieved with an *F1* score of 69.0% on the validation set, having the best *F1* of 66.4% ( $\pm 2.0$ ) on the test set. Thus, 76% of our annotated sentences suffice to train an optimal performing model.

Analysing the distribution of sentences in the training data sampled by MNLP, shows (*Math*, *CS*) as the most preferred domains and (*Eng*, *MS*) the least preferred ones. Nonetheless, all domains are represented, that is a non-uniformly mix of sentences sampled by MNLP yields the most generic model with less training data. In contrast, the random sampling strategy uniformly samples sentences from all domains.

Further, we show in Table 5 the proportion of training data for MNLP when the performance using the entire training dataset is achieved for related SciERC [32] and ScienceIE-17 [2] datasets. The results indicate, that also for related datasets on scientific texts MNLP can significantly reduce the amount of labelled training data.



**Fig. 4.** Progress of active learning with MNLP and random sampling strategy; the areas represent the standard deviation (std) of the F1 score across 5 folds for MNLP and random sampling strategy, respectively

**Table 5.** Performance of active learning with MNLP and random sampling strategy for the fraction of training data when the performance with entire training dataset is achieved; for SciERC and ScienceIE-17 results are reported across 5 random restarts

	Training data	F1 (MNLP)	F1 (random)	F1 (full data)
STM (our corpus)	52%	65.5 ( $\pm 1.0$ )	62.5 ( $\pm 2.6$ )	65.5 ( $\pm 1.3$ )
SciERC [32]	62%	65.3 ( $\pm 1.5$ )	62.3 ( $\pm 1.5$ )	65.6 ( $\pm 1.0$ )
ScienceIE17 [2]	38%	43.9 ( $\pm 1.2$ )	42.2 ( $\pm 1.8$ )	43.8 ( $\pm 1.0$ )

### 5.3 Correlations Between Inter-annotator Agreement and Performance

In this section, we analyse the correlations (Pearson’s  $R$ ) of inter-coder agreement  $\kappa$  and the number of annotated concepts per domain (#) on (1) the performance  $F1$  and (2) variance resp. standard deviation (std) of the classifiers across five-fold cross validation.

Table 6 summarises the results of our correlation analysis. The active learning classifier (AL-trained) has been trained with 52% training data sampled by MNLP since it is the point at which the performance of the full data trained

model is surpassed (see Table 5). For the domain-specific, domain-independent and AL-trained classifier we observe a strong correlation between F1 and number of concepts per domain ( $R$  0.70, 0.76, 0.68) and a weak correlation between  $\kappa$  and F1 ( $R$  0.20, 0.28, 0.23). Thus, we surmise that the number of annotated concepts in a particular domain has more influence on the performance than the inter-annotator agreement.

**Table 6.** Inter-annotator agreement ( $\kappa$ ) and the number of concept phrases (#) per domain; F1 and std of domain-specific classifiers on their domains; F1 and std of domain-independent and AL-trained classifier on each domain; the right side depicts correlation coefficients ( $R$ ) of each row with  $\kappa$  and the number of concept phrases

	<i>Agr</i>	<i>Ast</i>	<i>Bio</i>	<i>Che</i>	<i>CS</i>	<i>ES</i>	<i>Eng</i>	<i>MS</i>	<i>Mat</i>	<i>Med</i>	$R \kappa$	$R \#$
Inter-annotator agreement ( $\kappa$ )	0.6	0.57	0.75	0.77	0.85	0.81	0.79	0.9	0.58	0.94	1.00	-0.02
# concept phrases (#)	741	791	649	483	553	698	741	574	297	600	-0.02	1.00
Domain-specific (F1)	0.58	0.61	0.71	0.54	0.49	0.46	0.64	0.61	0.31	0.55	0.20	0.70
Domain-independent (F1)	0.68	0.66	0.71	0.64	0.65	0.63	0.71	0.69	0.48	0.61	0.28	0.76
AL-trained (F1)	0.65	0.67	0.74	0.65	0.62	0.63	0.72	0.69	0.50	0.60	0.23	0.68
Domain-specific (std)	0.06	0.06	0.09	0.08	0.05	0.06	0.04	0.11	0.06	0.07	0.29	0.28
Domain-independent (std)	0.04	0.04	0.11	0.08	0.07	0.05	0.03	0.04	0.06	0.03	-0.11	-0.05
AL-trained (std)	0.04	0.04	0.09	0.08	0.07	0.04	0.07	0.05	0.15	0.02	-0.41	-0.72

The correlation values for the variance are different between the classifier types. For the domain-specific classifier the correlation between  $\kappa$  and std, and the number of concepts per domain and std are slightly positive ( $R$  0.29, 0.28), i.e. the higher the agreement and the size of the domain, the higher the variance of the domain-specific classifier. For the domain-independent classifier, there is no correlation ( $R$  0.11, -0.05) and for the AL-trained classifier, the correlations become negative ( $R$  -0.41, -0.72), i.e. higher agreement and more annotated concepts per domain lead to less variance for the AL-trained classifier. In summary, we hypothesise that more diverse training data from several domains lead to better performance and lower variance by introducing an inductive bias.

## 6 Conclusions

In this paper, we have introduced the novel task of *domain-independent concept extraction* from scientific texts. During a systematic annotation procedure involving domain experts, we have identified four general core concepts that are relevant across the domains of Science, Technology and Medicine. To enable and foster research on these topics, we have annotated a corpus for the domains. We have verified the adequacy of the concepts by evaluating the human annotator agreement for our broad STM domain corpus. The results indicate that the identification of the *generic* concepts in a corpus covering 10 different scholarly domains is feasible by non-experts with moderate agreement and after consultation of domain experts with substantial agreement (0.76  $\kappa$ ).

We evaluated a state-of-the-art system on our annotated corpus which achieved a fairly high F1 score (65.5% overall). The domain-independent system noticeably outperforms the domain-specific systems, which indicates that the model can generalise well across domains. We also observed a strong correlation between the number of annotated concepts per domain and classifier performance, and only a weak correlation between inter-annotator agreement per domain and the performance. It is assumed that more annotated data positively influence the performance in the respective domain.

Furthermore, we have suggested active learning for our novel task. We have shown that only approx. 5 annotated abstracts per domain serving as training data are sufficient to build a performant model. Our active learning results for SciERC [32] and ScienceIE17 [2] datasets were similar. The promising results suggest that we do not need a large annotated dataset for scientific information extraction. Active learning can significantly save annotation costs and enable fast adaptation to new domains.

We make our annotated corpus, a silver-labelled corpus with 62K abstracts comprising 24 domains, and source code publicly available.<sup>1</sup> Thereby, we hope to facilitate research on the task of scientific information extraction and its several applications, e.g. academic search engines or research paper recommendation systems.

In the future, we plan to extend and refine the concepts for certain domains. We also intend to apply and evaluate our automatic scientific concept extraction system to expand an open research knowledge graph [23]. For this purpose, we plan to extend the corpus with additional relevant annotation layers such as with coreference links [28] and relations [16,32].

## References

1. Ammar, W., et al.: Construction of the literature graph in semantic scholar. In: NAACL-HLT (2018)
2. Augenstein, I., Das, M., Riedel, S., Vikraman, L., McCallum, A.: Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. In: SemEval@ACL (2017)
3. Balog, K.: Entity-oriented search. The Information Retrieval Series. Springer, Heidelberg (2018). <https://doi.org/10.1007/978-3-319-93935-3>
4. Beel, J., Gipp, B., Langer, S., Breiting, C.: Research-paper recommender systems: a literature survey. *Int. J. Digit. Libr.* **17**(4), 305–338 (2015). <https://doi.org/10.1007/s00799-015-0156-0>
5. Beltagy, I., Lo, K., Cohan, A.: SciBERT: pretrained language model for scientific text. In: EMNLP (2019)
6. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**(Database issue), D267-70 (2004)
7. Bornmann, L., Mutz, R.: Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. *J. Assoc. Inf. Sci. Technol.* **66**(11), 2215–2222 (2015)

<sup>1</sup> <https://gitlab.com/TIBHannover/orkg/orkg-nlp/tree/master/STM-corpus>.

8. Chambers, A.: Statistical models for text classification and clustering: applications and analysis. Ph.D. thesis, University of California, Irvine (2013)
9. Cohan, A., Ammar, W., van Zuylen, M., Cady, F.: Structural scaffolds for citation intent classification in scientific publications. In: NAACL-HLT (2019)
10. Cohen, J.: A coefficient of agreement for nominal scales. *Educ. Psychol. Measur.* **20**(1), 37–46 (1960)
11. Constantin, A., Peroni, S., Pettifer, S., Shotton, D.M., Vitali, F.: The document components ontology (DoCO). *Semant. Web* **7**, 167–181 (2016)
12. Dernoncourt, F., Lee, J.Y.: Pubmed 200k RCT: a dataset for sequential sentence classification in medical abstracts. In: IJCNLP (2017)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805 (2018)
14. Elsevier OA STM Corpus. <https://github.com/elsevierlabs/OA-STM-Corpus>. Accessed 12 Apr 2019
15. Fisas, B., Saggion, H., Ronzano, F.: On the discursive structure of computer graphics research papers. In: LAW@NAACL-HLT (2015)
16. 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 (2018)
17. Gardner, M., et al.: AllenNLP: a deep semantic natural language processing platform. *arXiv preprint* [arXiv:1803.07640](https://arxiv.org/abs/1803.07640) (2018)
18. Google scholar. <https://scholar.google.com/>. Accessed 12 Sept 2019
19. Groza, T., Kim, H., Handschuh, S.: Salt: semantically annotated latex. In: SAAW@ISWC (2006)
20. Handschuh, S., Zadeh, B.Q.: The ACL RD-TEC: a dataset for benchmarking terminology extraction and classification in computational linguistics. In: COLING 2014: 4th International Workshop on Computational Terminology (2014)
21. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997)
22. Hounsby, N., Huszar, F., Ghahramani, Z., Lengyel, M.: Bayesian active learning for classification and preference learning. *CoRR* abs/1112.5745 (2011)
23. Jaradeh, M.Y., et al.: Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In: K-CAP 2019 (2019)
24. Jin, D., Szolovits, P.: Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In: EMNLP (2018)
25. Jurgens, D., Kumar, S., Hoover, R., McFarland, D.A., Jurafsky, D.: Measuring the evolution of a scientific field through citation frames. *Trans. Assoc. Comput. Linguist.* **6**, 391–406 (2018)
26. Kim, S., Martínez, D., Cavedon, L., Yencken, L.: Automatic classification of sentences to support evidence based medicine. In: BMC Bioinformatics (2011)
27. Lao, N., Cohen, W.W.: Relational retrieval using a combination of path-constrained random walks. *Mach. Learn.* **81**, 53–67 (2010)
28. Lee, K., He, L., Lewis, M., Zettlemoyer, L.S.: End-to-end neural coreference resolution. In: EMNLP (2017)
29. Lehmann, J., et al.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semant. Web* **6**, 167–195 (2015)
30. Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D.: Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* **28**(7), 991–1000 (2012)

31. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.R.: Corpora for the conceptualisation and zoning of scientific papers. In: LREC (2010)
32. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: EMNLP (2018)
33. Ma, X., Hovy, E.H.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. CoRR abs/1603.01354 (2016)
34. Microsoft Academic. <https://academic.microsoft.com/home>. Accessed 12 Sept 2019
35. Microsoft Academic Knowledge Graph. <http://ma-graph.org/>. Accessed 12 Sept 2019
36. Papers with code. <https://paperswithcode.com/>. Accessed 12 Sept 2019
37. Pertsas, V., Constantopoulos, P.: Scholarly ontology: modelling scholarly practices. *Int. J. Digit. Libr.* **18**(3), 173–190 (2017)
38. Pustu-Iren, K., et al.: Investigating correlations of inter-coder agreement and machine annotation performance for historical video data. In: TPDL (2019)
39. Salatino, A.A., Thanapalasingam, T., Mannocci, A., Osborne, F., Motta, E.: The computer science ontology: a large-scale taxonomy of research areas. In: International Semantic Web Conference (2018)
40. Semantic scholar. <https://www.semanticscholar.org/>. Accessed 12 Sept 2019
41. Shen, Y., Yun, H., Lipton, Z.C., Kronrod, Y., Anandkumar, A.: Deep active learning for named entity recognition. In: ICLR (2017)
42. Siddhant, A., Lipton, Z.C.: Deep Bayesian active learning for natural language processing: results of a large-scale empirical study. In: EMNLP (2018)
43. Snow, R., O'Connor, B.T., Jurafsky, D., Ng, A.Y.: Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In: EMNLP (2008)
44. spaCy: Industrial-strength natural language processing. <http://www.spacy.io>. Accessed 02 Sep 2019
45. Springer Nature SciGraph. <https://www.springernature.com/gp/researchers/scigraph>. Accessed 12 Sept 2019
46. Teufel, S., Siddharthan, A., Batchelor, C.: Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3, vol. 3, pp. 1493–1502. Association for Computational Linguistics (2009)
47. Xiong, C., Power, R., Callan, J.P.: Explicit semantic ranking for academic search via knowledge graph embedding. In: WWW (2017)
48. Yaman, B., Pasin, M., Freudenberg, M.: Interlinking SciGraph and DBpedia datasets using link discovery and named entity recognition techniques. In: LDK (2019)
49. Zhang, Y., Lease, M., Wallace, B.C.: Active discriminative text representation learning. In: AAAI (2016)