Domain-independent Extraction of Scientific Concepts from Research Articles

Arthur Brack[0000−0002−1428−5348], Jennifer D'Souza[0000−0002−6616−9509], Anett Hoppe[0000−0002−1452−9509], Sören Auer[0000−0002−0698−2864], and Ralph Ewerth[0000−0003−0918−6297]

Leibniz Information Centre for Science and Technology (TIB), Hanover, Germany
{arthur.brack|jennifer.dsouza|anett.hoppe|soeren.auer|ralph.ewerth}@tib.eu

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 two deep learning systems as baselines. In particular, we propose active learning to deal with different domains in our task. 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 strong 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. [42] 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) [35]

or "Papers with code" [32]. Information extraction from text helps to identify emerging entities and to populate knowledge graphs [3]. Thus, information extraction from scientific texts 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 structured study of these assumptions is missing. We thus present the task of *domain-independent scientific concept extraction*. This article examines the intuition that most domain-specific articles share certain core concepts such as the mentions of research tasks, used materials, or data. If so, these would allow a domain-independent information extraction system, 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 science concepts that generalise well over the set of examined domains (10 disciplines from Science, Technology and Medicine (STM)). These 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. Additionally, 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. (3) We present two baseline deep learning systems for this task, including an active learning approach. To the best of our knowledge, this is the first approach that applies active learning to scholarly texts. We demonstrate that about half of the training data are sufficient to maintain the performance when using the entire training set. (4) 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 scientific corpora before some exemplary applications for domain-independent information extraction from scientific papers and the respective state of the art are introduced.

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 [17,11], argumentative [41,27] or activity-based [33] structure of research papers.

Annotated datasets exist for several domains, e.g. PubMed200k [12] from biomedical randomized controlled trials, NICTA-PIBOSO [22] from evidence-based medicine, Dr. Inventor [14] from Computer Graphics, Core Scientific Concepts (CoreSC) [27] from Chemistry and Biochemistry, and Argumentative Zoning (AZ) [41] 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 [20,12,14,26].

Phrase level annotation. More recent corpora have been annotated at phrasal level. SciCite[9] and ACL ARC [21] are datasets for citation intent classification from Computer Science, Medicine, and Computational Linguistics. ACL RD-TEC [18] from Computational Linguistics aims at extracting scientific technology and non-technology terms. ScienceIE17 [2] from Computer Science, Material Sciences, and Physics contains three concepts PROCESS, TASK and MATERIAL. SciERC [28] from the machine learning domain contains six concepts TASK, METHOD, METRIC, MATERIAL, OTHER-SCIENTIFICTERM and GENERIC. Each corpus covers at most three domains.

Experts vs. non-experts. The aforementioned datasets were usually annotated by domain experts [12,22,2,28,18,27]. In contrast, Teufel et al. [41] 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 [14,8].

Snow et. al. [39] 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 [34] 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 applied to classification tasks for scientific text yet. Recent publications demonstrate the effectiveness of active learning for NLP tasks such as *Named Entity Recognition* (NER) [37] and sentence classification [44]. Siddhant and Lipton [38] and Shen et. al. [37] 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 [16], Microsoft Academic [30] and Semantic Scholar [36] specialise in search of scholarly literature. They exploit graph structures such as the Microsoft Academic Knowledge Graph [31], SciGraph [40], 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 [25] and Unified Medical Language System (UMLS) [6] using entity linking techniques. Yaman et al. [43] connect SciGraph with DBpedia person entities. Xiong et al. [42] 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 [23]. 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" [32] 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 Domain-independent scientific concept extraction: A corpus

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 scientific concepts based on abstracts from ten different science domains. Since other studies have also shown that non-expert annotations are feasible for the general and scientific domain, we go for a cost-efficient middle course: annotations of non-experts experienced in the annotation task and consultation with domain-experts. Finally, we explore how well state-of-the-art machine learning approaches do perform on this novel, domain-independent information extraction task and whether active learning can save annotation costs. The base corpus, which we make publicly available, and the annotation process are described below.

3.1 OA STM Corpus

The OA STM corpus [13] 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*). While the original corpus contains full articles, this first annotation cycle focuses on the articles' abstracts.

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. Main actors in the annotation process were two post-doctoral researchers with a background in computer science (acting as non-expert annotators); their basic annotation assumptions were checked by experts from the respective domains.

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

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

Pre-annotation. A literature review of annotation schemes [27,2,26,11] provided a seed set of potential candidate concepts. Both non-experts independently annotated a subset of the STM abstracts with these concepts 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 [26], 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 κ [10]. Results showed a moderate inter-annotator agreement of 0.52 κ.

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 κ could be reached (see Table 2). Subsequently, the remaining 60 abstracts (six per do-

main) were annotated by one annotator. This last 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 κ) for PROCESS, METHOD, MATERIAL, and METHOD scientific concept annotation

	Med	*MS*	*CS*	*ES*	*Eng*	*Che*	*Bio*	*Agr*	*Mat*	*Ast*	*Overall*
κ	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 in terms of size and the number of scientific concept phrases

	Ast	*Agr*	*Eng*	*ES*	*Bio*	*Med*	*MS*	*CS*	*Che*	*Mat*
Avg. # Tokens/Abstract	382	333	303	321	273	274	282	253	217	140
# 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
# 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

4 Experimental setup: Two baseline classifiers

The current state-of-the-art for scientific entity extraction is Beltagy et al.'s system [5]. We use their NER task-specific deep learning architecture atop SciBERT embeddings with a Conditional Random Field (CRF) based sequence tag decoder [29] and BILOU (beginning, inside, last, outside, unit) tagging scheme. The following classifiers are implemented in AllenNLP [15]. We report span-based micro-averaged F1 scores and use the ScienceIE17 [2] evaluation script.

4.1 Traditionally trained classifiers

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. Please note that 8 abstracts have about 445 concepts so that the training data should be sufficient for the domain-dependent classifier.

4.2 Active learning trained classifier

Based on the results of the aforementioned comparison studies [38,44], we decide to use MNLP [37] as the sampling strategy in the active learning setting. It is chosen over other possibly suitable candidates such as *Bayesian Active Learning by Disagreement* (BALD) [19], which is another powerful strategy, but has higher computational requirements. The objective involves strategically selecting sentences from the overall dataset in each iteration of the algorithm greedily, aiming at getting greater performance with a minimum number of 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. To obtain a robust evaluation of models, we repeat the experiment for five folds and average the results. The models use the same hyperparameters as for the domain-independent classifier. We retrain the model within each iteration and fold.

5 Experimental results and discussion

This section describes the results of the experimental setup and the correlation analysis between inter-annotator agreement and performance of the several classifiers.

5.1 Traditionally trained classifiers

Table 4 shows an overview of the *domain-independent* classifier results. The system achieves an $F1$ score of 65.5 ($\pm$ 1.26) in the overall task. 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). The concept METHOD is also the most underrepresented one in our corpus, which partly explains the poor extraction performance.

Next, we compare and contrast the 10 *domain-specific* classifiers according to their capability to extract the concepts from their own domains and in other domains. The results are shown as $F1$ scores in Figure 1 where the x-axis represents the 10 test domains. We discuss some observations in the sequel.

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	*Overall*
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)

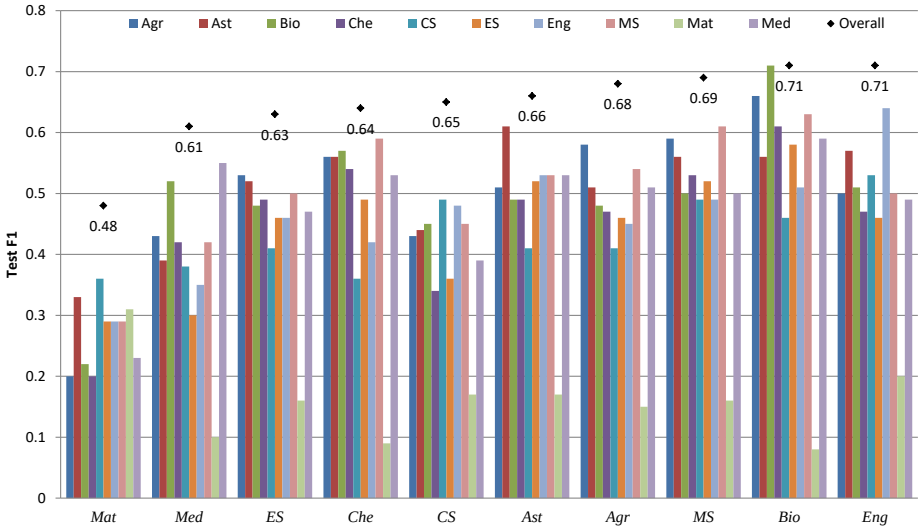
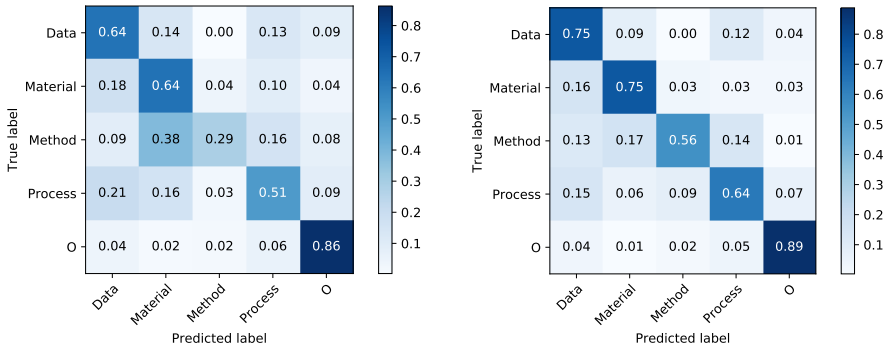


Fig. 1: $F1$ per domain of the 10 *domain-specific* classifiers (as bar plots) and of the *domain-independent* classifier (as scatter plots) for scientific concept extraction; the x-axis represents the 10 test domains

Most robust domain. *Bio* (third bar in each domain in Figure 1) extracts scientific entities 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 Figure 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*).

Most specialised domain. *Mat* (the second last bar in each domain in Figure 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 underrepresented 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.



(a) (b)

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

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-dependent classifier. Except for *Bio* the *domain-independent* classifier clearly outperforms the *domain-dependent* one extracting concepts from their respective domains. To analyse the reason, we investigate the improvements in *CS* domain. We have chosen *CS* exemplary as the size of the domain is slightly below the average and this domain strongly benefits from the *domain-independent* classifier and improves the $F1$ score for the *CS* classifier from 49.5 ($\pm$ 4.22) to 65.9 ($\pm$ 1.21). The $F1$ score for span-detection is improved from 73.4 ($\pm$ 3.45) to 82.0 ($\pm$ 3.98). Span-detection usually requires less domain-dependent signals, thus the *domain-independent* classifier can benefit from other domains. Accuracy on token-level also improves from 67.7 ($\pm$ 5.35) to 77.5 ($\pm$ 4.42) $F1$, 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 Figure 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 PROCESS, METHOD, MATERIAL, and DATA, respectively.

5.2 Active learning trained classifier

Figure 4 shows the results of the active learning experiment. Table 5 depicts the results for the fraction of training data when the performance using the entire training dataset is achieved. MNLP clearly outperforms the random baseline. While using only 52 %

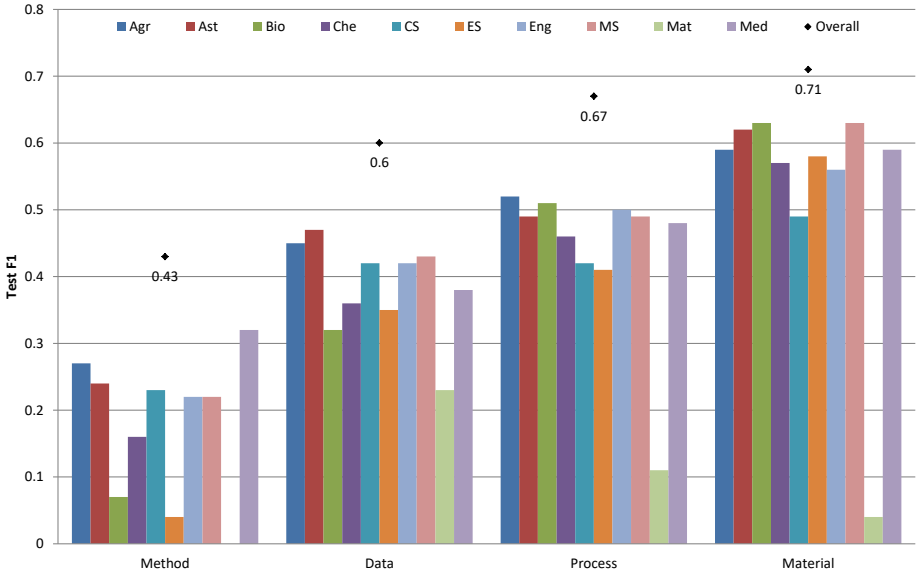


Fig. 3: $F1$ of the 10 *domain-specific* classifier (as bar plots) and the *domain-independent* classifier (as scatter plots) for extracting each scientific concept; the x-axis represents the evaluated concept

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). For comparison: the random baseline achieves an $F1$ score of 62.5 ($\pm$ 2.6) with 52 % of the 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. For SciERC [28] and ScienceIE17 [2] similar results are demonstrating that MNLP can significantly reduce the amount of labelled data.

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 ScienceIE17 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 [28]	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)

To find out which mix of training data produces the most generic model, we analyse the distribution of sentences in the training data sampled by MNLP. As expected, the random sampling strategy uniformly samples sentences from all domains in each iteration. However, (*Math*, *CS*) are the most and (*Eng*, *MS*) the least preferred domains

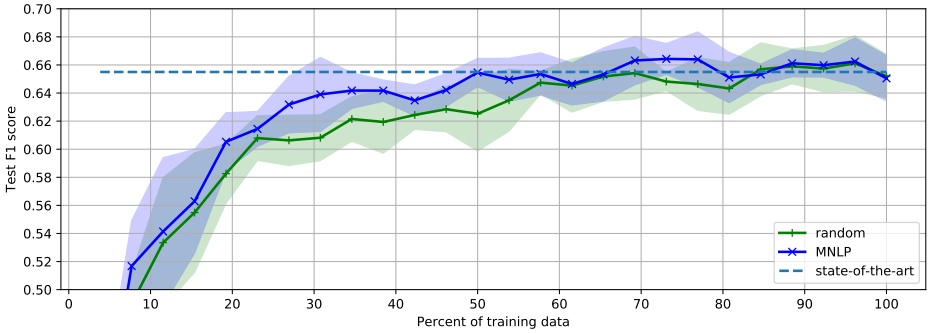


Fig. 4: Progress of active learning with MNLP and random sampling strategy; the areas represent the standard deviation (std) of the F1 score across 5 folds for MNLP and random sampling strategy, respectively

by MNLP. When using 52 % of the training data, 65.4% of *Math*, 66.2% of *CS* sentences were sampled, but only 41.6% of *Eng* and 37.3% of *MS*. Thereby all domains are present, that is a heterogeneous mix of sentences sampled by MNLP yields the most generic model with less training data.

5.3 Correlations between inter-annotator agreement and performance

In this section, we analyse the correlations of inter-annotator agreement κ and the number of annotated concepts per domain (#) on the performance and variance of the classifiers employing Pearson's correlation coefficient (Pearson's R).

Table 6: Inter-annotator agreement (κ) and the number of concept phrases (#) per domain; F1 and std of domain-dependent 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 κ and the number of concept phrases

	Agr	*Ast*	*Bio*	*Che*	*CS*	*ES*	*Eng*	*MS*	*Mat*	*Med*	$R\,\kappa$	$R\,\#$
inter-annotator agreement (κ)	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-dependent (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-dependent (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

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. For the domain-dependent, 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 κ and F1 (R 0.20, 0.28, 0.23). Thus, we can hypothesise that the number of annotated concepts in a particular domain has more influence on the performance than the inter-annotator agreement.

The correlation values for std is different between the classifier types. For the domain-dependent classifier the correlation between κ and std (R 0.29), and the number of concepts per domain and std (R 0.28) is slightly positive. In other words: the higher the agreement and the size of the domain, the higher the variance of the domain-dependent classifier. This is different for the domain-independent classifier as there is no correlation anymore. For the AL-trained classifier there is, on the other hand, a moderate negative correlation between κ and std (R -0.41), and a strong negative correlation between number of concepts per domain and std (R -0.72), i.e. higher agreement and larger amount of training data in a domain lead to less variance for the AL-trained classifier. We hypothesise that more diversity through several domains in the domain-independent and the AL-trained classifier leads 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 κ).

We have presented two deep learning systems which achieved a fairly high F1 score (65.5% overall). The domain-independent system noticeably outperforms the domain-dependent 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. We can hypothesise 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 [28] 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[1]. We hope to facilitate the research

[1] https://gitlab.com/TIBHannover/orkg/orkg-nlp/tree/master/STM-corpus

on that task and 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. Besides, we want to apply and evaluate the information extraction system to populate a research knowledge graph. For that we plan to extend the corpus with co-reference annotations [24] so that mentions referring to the same concept can be collapsed.

References

1. Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., Kinney, R., Kohlmeier, S., Lo, K., Murray, T.C., Ooi, H.H., Peters, M.E., Power, J., Skjonsberg, S., Wang, L.L., Wilhelm, C., Yuan, Z., van Zuylen, M., Etzioni, O.: 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. In: The Information Retrieval Series (2018)
4. Beel, J., Gipp, B., Langer, S., Breitinger, C.: Research-paper recommender systems: a literature survey. International Journal on Digital Libraries **17**, 305–338 (2015)
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 research **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. Journal of the Association for Information Science and Technology **66**(11) (2015)
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. Educational and psychological measurement **20**(1), 37–46 (1960)
11. Constantin, A., Peroni, S., Pettifer, S., Shotton, D.M., Vitali, F.: The document components ontology (doco). Semantic 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. Elsevier oa stm corpus. `https://github.com/elsevierlabs/OA-STM-Corpus`, accessed: 2019-04-12
14. Fisas, B., Saggion, H., Ronzano, F.: On the discoursive structure of computer graphics research papers. In: LAW@NAACL-HLT (2015)
15. Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., Zettlemoyer, L.: Allennlp: A deep semantic natural language processing platform. arXiv preprint arXiv:1803.07640 (2018)
16. Google scholar. `https://scholar.google.com/`, accessed: 2019-09-12
17. Groza, T., Kim, H., Handschuh, S.: Salt: Semantically annotated latex. In: SAAW@ISWC (2006)
18. Handschuh, S., QasemiZadeh, B.: 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)

19. Houlsby, N., Huszar, F., Ghahramani, Z., Lengyel, M.: Bayesian active learning for classification and preference learning. CoRR **abs/1112.5745** (2011)
20. Jin, D., Szolovits, P.: Hierarchical neural networks for sequential sentence classification in medical scientific abstracts. In: EMNLP (2018)
21. Jurgens, D., Kumar, S., Hoover, R., McFarland, D.A., Jurafsky, D.: Measuring the evolution of a scientific field through citation frames. Transactions of the Association for Computational Linguistics **6**, 391–406 (2018)
22. Kim, S., Martínez, D., Cavedon, L., Yencken, L.: Automatic classification of sentences to support evidence based medicine. In: BMC Bioinformatics (2011)
23. Lao, N., Cohen, W.W.: Relational retrieval using a combination of path-constrained random walks. Machine Learning **81**, 53–67 (2010)
24. Lee, K., He, L., Lewis, M., Zettlemoyer, L.S.: End-to-end neural coreference resolution. In: EMNLP (2017)
25. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web **6**, 167–195 (2015)
26. 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)
27. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.R.: Corpora for the conceptualisation and zoning of scientific papers. In: LREC (2010)
28. Luan, Y., He, L., Ostendorf, M., Hajishirzi, H.: Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In: EMNLP (2018)
29. Ma, X., Hovy, E.H.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. CoRR **abs/1603.01354** (2016)
30. Microsoft academic. `https://academic.microsoft.com/home`, accessed: 2019-09-12
31. Microsoft academic knowledge graph. `http://ma-graph.org/`, accessed: 2019-09-12
32. Papers with code. `https://paperswithcode.com/`, accessed: 2019-09-12
33. Pertsas, V., Constantopoulos, P.: Scholarly ontology: modelling scholarly practices. International Journal on Digital Libraries **18**(3), 173–190 (2017)
34. Pustu-Iren, K., Mühling, M., Korfhage, N., Bars, J., Bernhöft, S., Hörth, A., Freisleben, B., Ewerth, R.: Investigating correlations of inter-coder agreement and machine annotation performance for historical video data. In: TPDL (2019)
35. 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)
36. Semantic scholar. `https://www.semanticscholar.org/`, accessed: 2019-09-12
37. Shen, Y., Yun, H., Lipton, Z.C., Kronrod, Y., Anandkumar, A.: Deep active learning for named entity recognition. In: ICLR (2017)
38. Siddhant, A., Lipton, Z.C.: Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In: EMNLP (2018)
39. 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)
40. Springer nature scigraph. `https://www.springernature.com/gp/researchers/scigraph`, accessed: 2019-09-12
41. 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-Volume 3. pp. 1493–1502. Association for Computational Linguistics (2009)

42. Xiong, C., Power, R., Callan, J.P.: Explicit semantic ranking for academic search via knowledge graph embedding. In: WWW (2017)
43. Yaman, B., Pasin, M., Freudenberg, M.: Interlinking scigraph and dbpedia datasets using link discovery and named entity recognition techniques. In: LDK (2019)
44. Zhang, Y., Lease, M., Wallace, B.C.: Active discriminative text representation learning. In: AAAI (2016)