Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2192–2203
Marseille, 11–16 May 2020

The STEM-ECR Dataset: Grounding Scientific Entity References in STEM Scholarly Content to Authoritative Encyclopedic and Lexicographic Sources

Jennifer D'Souza, Anett Hoppe, Arthur Brack, Mohamad Yaser Jaradeh, Sören Auer, Ralph Ewerth
TIB Leibniz Information Centre for Science and Technology,
Hannover, Germany
{jennifer.dsouza,anett.hoppe,arthur.brack,yaser.jaradeh,auer,ralph.ewerth}@tib.eu

Abstract

We introduce the STEM (Science, Technology, Engineering, and Medicine) Dataset for Scientific Entity Extraction, Classification, and Resolution, version 1.0 (STEM-ECR v1.0). The STEM-ECR v1.0 dataset has been developed to provide a benchmark for the evaluation of scientific entity extraction, classification, and resolution tasks in a domain-independent fashion. It comprises abstracts in 10 STEM disciplines that were found to be the most prolific ones on a major publishing platform. We describe the creation of such a multidisciplinary corpus and highlight the obtained findings in terms of the following features: 1) a generic conceptual formalism for scientific entities in a multidisciplinary scientific context; 2) the feasibility of the domain-independent human annotation of scientific entities under such a generic formalism; 3) a performance benchmark obtainable for automatic extraction of multidisciplinary scientific entities using BERT-based neural models; 4) a delineated 3-step entity resolution procedure for human annotation of the scientific entities via encyclopedic entity linking and lexicographic word sense disambiguation; and 5) human evaluations of Babelfy returned encyclopedic links and lexicographic senses for our entities. Our findings cumulatively indicate that human annotation and automatic learning of multidisciplinary scientific concepts as well as their semantic disambiguation in a wide-ranging setting as STEM is reasonable.

Keywords: Entity Recognition, Entity Classification, Entity Resolution, Entity Linking, Word Sense Disambiguation, Evaluation Corpus, Language Resource

1. Introduction

Recently, an increasing number of research efforts are geared toward adopting Knowledge Graphs (KG) for modeling scholarly publications (Ammar et al., 2018; Jaradeh et al., 2019). They advocate for such advanced semantic machine-interpretability, via KGs, of the publications to enable their more intelligent automated processing (e.g., in search applications). To this end, KGs leveraged in academic and industrial settings fostering better search already serve as case-in-point, albeit so far only over commonsense world knowledge (consider the DBpedia (Auer et al., 2007) and Google (Singhal, Amit, 2012) KGs as two prominent examples among others).

To represent scholarly publications as KGs, from an Information Extraction (IE) perspective, and scientific IE, in particular, extracting scientific entities from scholarly publications becomes a vital task to address since entities lie at the core of KGs. While these scientific entities are essentially scientific terms, they are referred to with the broader notion of *entities* in the rest of the paper to posit them as valid knowledge graph node candidates. As an IE task, the extraction of scientific entities is being increasingly studied in the Natural Language Processing (NLP) community– the SemEval series itself has so far seen four tasks organized (Kim et al., 2010; Moro and Navigli, 2015; Augenstein et al., 2017; Gábor et al., 2018). However, very little of this work (Handschuh and QasemiZadeh, 2014) has been done in the tradition of corpus linguistics and none, so far, in the broad multidisciplinary setting of Science.

In this vein, with the aim of providing a platform for benchmarking methods of scholarly article processing, in 2017 Elsevier Labs released an open access corpus of publications across Science, Technology, Engineering, and Medicine (STEM),[1] thus providing a shared test bed for multidisciplinary scientific IE research. The corpus is based on the 10 most prolific STEM disciplines, viz. Agriculture (*Agr*), Astronomy (*Ast*), Biology (*Bio*), Chemistry (*Che*), Computer Science (*CS*), Earth Science (*ES*), Engineering (*Eng*), Materials Science (*MS*), and Mathematics (*Mat*). The study presented subsequently in this paper was performed on scientific abstracts in this STEM corpus.

In general, the challenges associated with scientific IE are seen to be greater than for, say newswire articles, in part because of the expected domain expertise required to annotate resources for machine learning making the annotation task costly, in turn limiting resources. Since, at present, a study of human agreement of annotating scientific entities across STEM is lacking, *consequently, we have little understanding about at least a certain set of scientific concepts that are generically applicable and the extent to which they can be decided without domain knowledge.* Quantitatively evaluating this based on a set of *generic* scientific concepts is the first goal of this paper. We refer to this problem as the *generic scientific concept extraction* task, for which our research question is: Can a generic formalism of scientific concepts offer human annotators who possess sufficient scientific competence, the ability to reliably annotate a multidisciplinary scholarly corpus with scientific entities? Examining this question, leads us in turn to make the observation that the problem of such otherwise costly data creation can be made accessible to a wider group of people with the minimal involvement of domain specialists.

In the second part of this study, with the goal of retaining accurate entities, we disambiguate our annotated entities via encyclopedic links and lexicographic senses. Specif-

[1] https://github.com/elsevierlabs/OA-STM-Corpus

ically, we carry out Entity Resolution (ER) (Getoor and Machanavajjhala, 2012) via Entity Linking (EL) and Word Sense Disambiguation (WSD) annotations. We adopt their respective standard task definitions for our scientific entities, where EL (Rao et al., 2013; Usbeck et al., 2015) entails linking entities to the most suitable entry in a knowledge base, and WSD (Navigli, 2009) involves explicitly assigning meanings to single-word and multi-word occurrences within text. Performing joint EL and WSD keeps us on course with the current "best of two worlds" trend in ER research considering their seamless integration as complementary semantic information units for disambiguating the entities (Hovy et al., 2013) to obtain boosted performance in their automatic disambiguation (Moro et al., 2014b). In addition, we note that without ER, our entity annotations can seem somewhat subjective on the one hand. While on the other hand, they might be limiting for real-world usage. With ER, we make the scientific entities locatable in the real world thus removing to a certain degree their subjective impression, while facilitating their categorization (conceptualization) according to the metadata in the knowledge base, without restricting users to our four concepts.

Further, the introduction of this STEM corpus facilitates the testing of the ER approaches in a unique setting: *to that of being restricted within the single broad domain of Science, while still being multidisciplinary*, thus, in a sense, validating the automatic systems for their semantic adaptability. Training algorithms on such a corpus entails switching senses of the the same word. E.g., "the Cloud" in *CS* should be resolved to a technological sense, versus in *Ast* where it takes the common interpretation of the mass of water vapor we see in the sky; or "neural networks" which in *CS* are an algorithm versus in *Bio* or *Med* which refer to the brain. It would require the seemingly evident shift from common interpretations of terms, though not always. E.g., "power" in *Mat* refers to exponentiation, which, otherwise in a common sense, takes on a human social interpretation; or even a common word as "ring" in *Mat* is an algebraic structure versus the jewelry common interpretation; or "subject" in *Med* as research participant rather than an academic discipline. However, we note that our corpus not only enables evaluating systems for their semantic adaptability in a joint fashion for both EL and WSD tasks, it can also be leveraged for designing approaches attempting either one of the tasks, on the premise of their dichotomy, offering still novel insights in this regard owing to its multidisciplinary nature.

Finally, in this study, for both parts of the STEM data, we present benchmark results with insights on respective model performances: 1) for entity recognition, we train a BERT-based neural model (Devlin et al., 2019) on our data of scientific entities; and 2) for entity resolution, we manually evaluate Babelfy (Moro et al., 2014b) for EL and WSD of the scientific entities in our corpus.

In summary of our contributions, we: 1) release a novel multidisciplinary STEM corpus of scientific entities under a *generic* conceptual formalism bridging STEM scientific domains; where the entities are further enriched with EL and WSD annotations, thus, both disambiguating their scientific sense and grounding them in the real world, thereby enabling additional semantic extensions of the entities; and 2) provide benchmark performances from state-of-the-art systems on our STEM data. For further research, the STEM-ECR v1.0 corpus can be downloaded at the following link: https://doi.org/10.25835/0017546 (ISLRN 749-555-840-571-2).

The remaining paper is organized in two main parts: 1) annotating scientific entities in a STEM setting; and 2) linking and sense disambiguation of the selected entities. We begin with a discussion on related work to ours.

2. Related Work

Early initiatives in semantically structuring scholarly publications focused on sentences as the basic unit of analysis using data from one or two scientific domains. To this end, ontologies and vocabularies were created (Teufel et al., 1999; Soldatova and King, 2006; Constantin et al., 2016; Pertsas and Constantopoulos, 2017), corpora were annotated (Liakata et al., 2010; Fisas et al., 2016), and machine learning methods were applied (Liakata et al., 2012). Recently, scientific IE has targeted search technology, thus newer corpora have been annotated at the phrasal unit of information with three or six types of scientific concepts (Handschuh and QasemiZadeh, 2014; Augenstein et al., 2017; Luan et al., 2018) facilitating machine learning system development (Ammar et al., 2017; Luan et al., 2017; Beltagy et al., 2019). But, again these corpora are from a single or at most three domains.

In this work, while we continue to address scientific IE at the phrasal unit, unlike existing work, by explicitly situating our task in the wide-ranging STEM scholarly communication setting, we do not impose a domain restriction on the task. Additionally, we enrich our entities with semantic EL and WSD annotations.

On this related note, the SemEval 2015 Task 13 (Moro and Navigli, 2015) corpus exists with joint EL and WSD annotations for entities. However, theirs isn't a scientific IE task since the genre of text they annotate does not comprise scholarly publications. Further, their annotations differs from ours based on the knowledge source they use, where ours is identical to theirs on the encyclopedic front but a subset of theirs on the lexical knowledge since their lexical knowledge is derived from at least three different sources. Nevertheless, restricting ourselves to just one lexicographic source, lets us test its coverage uniformly.

It was recently noted that the guidelines clarifying the inclusion or exclusion of entities take on an implicit inclusion criteria which only becomes apparent when noting the differences between the entities in datasets that share a common creation objective (Rosales-Méndez et al., 2019). In this work, we adopt a fine-grained annotation strategy following a safer maximal inclusion route.

3. Scientific Entity Annotations

By starting with a STEM corpus of scholarly abstracts for annotating with scientific entities, we differ from existing work addressing this task since we are going beyond the domain restriction that so far seems to encompass scientific IE. For entity annotations, we rely on existing scientific concept formalisms (Liakata et al., 2010; Constantin

et al., 2016; Augenstein et al., 2017) that appear to propose generic scientific concept types that can bridge the domains we consider, thereby offering a uniform entity selection framework. In the following subsections, we describe our annotation task in detail, after which we conclude with benchmark results.

PROCESS Natural phenomenon, or independent/dependent activities. E.g., growing (*Bio*), cured (*MS*), 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 digital entity used for scientific experiments. E.g., soil (*Agr*), the moon (*Ast*), Doppler lidars (*Eng*).
DATA The data themselves, or quantitative or qualitative characteristics of entities. E.g., rotational energy (*Eng*), tensile strength (*MS*), the Sylow p-groups (*Mat*).

Table 1: The four scientific concepts studied

3.1. Our Annotation Process

The corpus for computing inter-annotator agreement was annotated by two postdoctoral researchers in Computer Science.[2] To develop annotation guidelines, a small pilot annotation exercise was performed on 10 abstracts (one per domain) with a set of surmised generically applicable scientific concepts such as TASK, PROCESS, MATERIAL, OBJECT, METHOD, DATA, MODEL, RESULTS, etc., taken from existing work. Over the course of three annotation trials, these concepts were iteratively pruned where concepts that did not cover all domains were dropped, resulting in four finalized concepts, viz. PROCESS, METHOD, MATERIAL, and DATA as our resultant set of *generic* scientific concepts (see Table 1 for their definitions).[3] The subsequent annotation task entailed linguistic considerations for the precise selection of entities as one of the four scientific concepts based on their part-of-speech tag or phrase type. PROCESS entities were verbs (e.g., "prune" in *Agr*), verb phrases (e.g., "integrating results" in *Mat*), or noun phrases (e.g. "this transport process" in *Bio*); METHOD entities comprised noun phrases containing phrase endings such as simulation, method, algorithm, scheme, technique, system, etc.; MATERIAL were nouns or noun phrases (e.g., "forest trees" in *Agr*, "electrons" in *Ast* or *Che*, "tephra" in *ES*); and majority of the DATA entities were numbers otherwise noun phrases (e.g., "$(2.5\pm1.5)\text{kms}^{-1}$" representing a velocity value in *Ast*, "plant available P status" in *Agr*). Summarily, the resulting annotation guidelines hinged upon the following five considerations:

1. To ensure consistent scientific entity spans, entities were annotated as definite noun phrases where possible. In later stages, the extraneous determiners and articles could be dropped as deemed appropriate.
2. Coreferring lexical units for scientific entities in the context of a single abstract were annotated with the same concept type.
3. Quantifiable lexical units such as numbers (e.g., years 1999, measurements 4km) or even as phrases (e.g., vascular risk) were annotated as DATA.
4. Where possible, the most precise text reference (i.e., phrases with qualifiers) regarding materials used in the experiment were annotated. For instance, "carbon atoms in graphene" was annotated as a single MATERIAL entity and not separately as "carbon atoms," "graphene."
5. Any confusion in classifying scientific entities as one of four types was resolved using the following concept precedence: METHOD > PROCESS > DATA > MATERIAL, where the concept appearing earlier in the list was preferred.

After finalizing the concepts and updating the guidelines,[4] the final annotation task proceeded in two phases

1. In 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 (Cohen, 1960). Results showed a moderate inter-annotator agreement at 0.52 κ.
2. Next, in phase II, one of the annotators interviewed subject specialists in each of the ten domains about the choice of concepts and her annotation decisions on their respective domain corpus.[5] The feedback from the interviews were systematically categorized into error types and these errors were discussed by both annotators. Following these discussions, the 50 abstracts from phase I were independently reannotated. The annotators could obtain substantial overall agreement of 0.76 κ after phase II.

In Table 2, we report the IAA scores obtained per domain and overall. The scores show that the annotators had a substantial agreement in seven domains, while only a moderate agreement was reached in three domains, viz. *Agr*, *Mat*, and *Ast*.

	κ
Med	0.94
MS	0.90
CS	0.85
ES	0.81

	κ
Eng	0.79
Che	0.77
Bio	0.75
Agr	0.60

	κ
Mat	0.58
Ast	0.57
Overall 0.76	

Table 2: Per-domain and Overall inter-annotator agreement (Cohen's Kappa κ) for PROCESS, METHOD, MATERIAL, and METHOD scientific concept annotation

[2] We use BRAT (Stenetorp et al., 2012) for annotating entity spans and their categories.

[3] We do not consider nested span concepts in this study, hence we leave out TASK, OBJECT, and RESULTS since they were almost always nested with the other scientific entities. However, we found them as valid *generic* categories as well.

[4] The annotation guidelines are released with the corpus.

[5] The concepts were generally well received as generically applicable, except in Engineering where the subject specialist mentioned the need for an additional category TOOL for phrases which our scheme captures as MATERIAL.

	Ast	*Agr*	*Eng*	*ES*	*Bio*	*Med*	*MS*	*CS*	*Che*	*Mat*
# Tokens overall	3711	3134	2917	3065	2579	2558	2597	2383	1991	1334
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

Table 3: The annotated corpus characteristics in terms of size and the number of scientific concept phrases

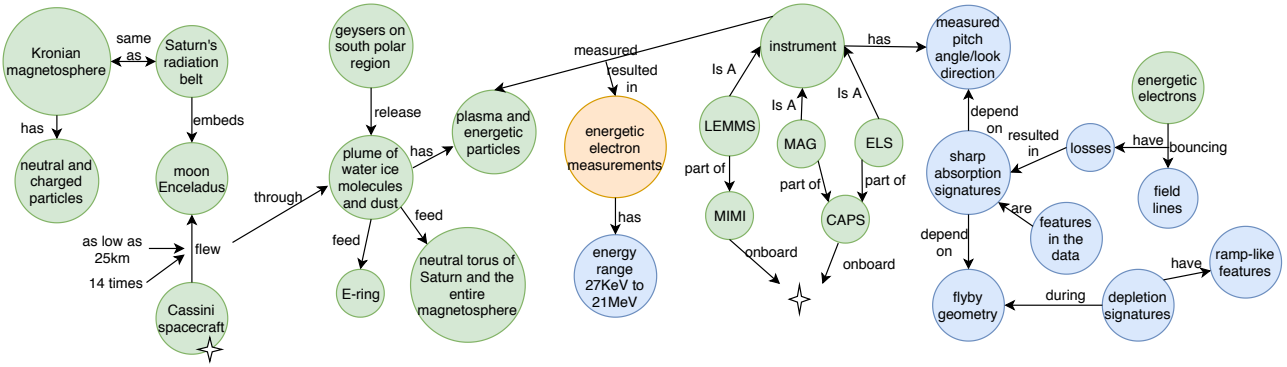

Figure 1: A text graph of the abstract of the article 'The Cassini Enceladus encounters 2005–2010 in the view of energetic electron measurements' (Krupp et al., 2012). Nodes are color-coded by concept type: orange corresponds to PROCESS, green to MATERIAL, and blue to DATA. No METHOD concepts were identified in this abstract.

Annotation Error Analysis We discuss some of the changes the interviewer annotator made in phase II after consultation with the subject experts.

In total, 21% of the phase I annotations were changed: PROCESS accounted for a major proportion (nearly 54%) of the changes. Considerable inconsistency was found in annotating verbs like "increasing", "decreasing", "enhancing", etc., as PROCESS or not. Interviews with subject experts confirmed that they were a relevant detail to the research investigation and hence should be annotated. So 61% of the PROCESS changes came from additionally annotating these verbs. MATERIAL was the second predominantly changed concept in phase II, accounting for 23% of the overall changes. Nearly 32% of the changes under MATERIAL came from consistently reannotating phrases about models, tools, and systems; accounting for another 22% of its changes, where spatial locations were an essential part of the investigation such as in the *Ast* and *ES* domains, they were decided to be included in the phase II set as MATERIAL. Finally, there were some changes that emerged from lack of domain expertise. This was mainly in the medical domain (4.3% of the overall changes) in resolving confusion in annotating PROCESS and METHOD concept types. Most of the remaining changes were based on the treatment of conjunctive spans or lists.

Subsequently, the remaining 60 abstracts (six per domain) were annotated by one annotator. This last phase also involved reconciliation of the earlier annotated 50 abstracts to obtain a gold standard corpus.

Annotated Corpus Characteristics Table 3 shows our annotated corpus characteristics. Our corpus comprises a total of 6,127 scientific entities, including 2,112 PROCESS, 258 METHOD, 2,099 MATERIAL, and 1,658 DATA entities. The number of entities per abstract 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).

In Figure 1, we show an example instance of a manually created text graph from the scientific entities in one abstract. The graph highlights that linguistic relations such as synonymy, hypernymy, meronymy, as well as OpenIE relations are poignant even between scientific entities.

3.2. Performance Benchmark

In this section, we report the performance of a state-of-the-art neural system for extracting scientific entities from our STEM corpus. Specifically, we leverage SciBERT (Beltagy et al., 2019), a pretrained language model based on BERT (Devlin et al., 2019) but trained on a large corpus of scientific text, as frozen embedding features in a NER task-specific neural model comprising BiLSTM layers and a CRF-based sequence tag decoder (Ma and Hovy, 2016).

Models. Using the above architecture implemented in the AllenNLP framework (Gardner et al., 2018), we train one model with data from all domains combined. We call this the *domain-independent* scientific entity extraction system. To test robust models, we performed five-fold cross validation experiments. In each fold experiment, we trained a model on 8 abstracts per domain (i.e. 80 abstracts; >4000 entities), tuned hyperparameters on 1 abstract per domain (i.e. 10 abstracts; >500 entities), and tested on the remain-

ing 2 abstracts (i.e. 20 abstracts; >1000 entities) ensuring that data splits were not identical between folds.

Results and Discussion. This system achieves 64.3% precision, 66.7% recall, and 65.5% overall task $F1$ at a stable 0.0140 standard deviation per fold. For it, MATERIAL was the easiest concept to extract at 71% $F1$, followed by PROCESS at 66.8%, followed by DATA at 59.8%, and METHOD the hardest at 43% $F1$. Of note, METHOD is also the most underrepresented in our corpus which could in part account for its poor extraction performance.

Further, in Table 4, we report the performance of this system per domain. It demonstrates highest performance on the *Eng* and *Bio* domains at 0.71 $F1$ and shows significantly lower performance on *Mat* at 0.48 $F1$. Again, compared to the remaining 9 domains in our corpus, *Mat* has the least annotated entities, which we attribute as the cause for the low extraction performance. We hypothesize that more training instances could improve performance. For more detailed machine learning results, we refer the reader to our related work (Brack et al., 2020).

	$F1$
Eng	0.71
Bio	0.71
MS	0.69
Agr	0.68

	$F1$
Ast	0.66
CS	0.65
Che	0.64
ES	0.63

	$F1$
Med	0.61
Mat	0.48
Overall 0.65	

Table 4: The *domain-independent* scientific entity extraction system results per-domain

In the second stage of the study, we perform word sense disambiguation and link our entities to authoritative sources.

4. Scientific Entity Resolution

Aside from the four scientific concepts facilitating a common understanding of scientific entities in a multidisciplinary setting, the fact that they are just four made the human annotation task feasible. Utilizing additional concepts would have resulted in a prohibitively expensive human annotation task. Nevertheless, there are existing datasets (particularly in the biomedical domain, e.g., GENIA (Kim et al., 2003)) that have adopted the conceptual framework in rich domain-specific semantic ontologies. Our work, while related, is different since we target the annotation of multidisciplinary scientific entities that facilitates a low annotation entrance barrier to producing such data. This is beneficial since it enables the task to be performed in a domain-independent manner by researchers, but perhaps not crowdworkers, unless screening tests for a certain level of scientific expertise are created.

Nonetheless, we recognize that the four categories might be too limiting for real-world usage. Further, the scientific entities from stage 1 remain susceptible to subjective interpretation without additional information. Therefore, in a similar vein to adopting domain-specific ontologies, we now perform entity linking (EL) to the Wikipedia and word sense disambiguation (WSD) to Wiktionary.

4.1. Our Annotation Process

The same pair of annotators as before were involved in this stage of the study to determine the annotation agreement.

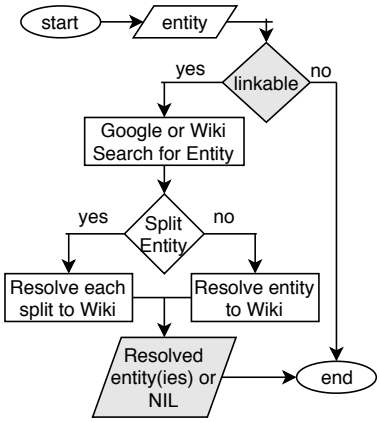

Figure 2: Flowchart depicting our Entity Resolution annotation steps. The boxes shaded in grey represent the steps where annotator agreement scores are computed.

4.1.1. Annotation Task Tools

During the annotation procedure, each annotator was shown the entities, grouped by domain and file name, in Google Excel Sheet columns alongside a view of the current abstract of entities being annotated in the BRAT interface (2012) for context information about the entities.[6] For entity resolution, i.e. linking and disambiguation, the annotators had local installations of specific time-stamped Wikipedia[7] and Wiktionary[8] dumps to enable future persistent references to the links since the Wiki sources are actively revised. They queried the local dumps using the DKPro JWPL tool (Zesch et al., 2008) for Wikipedia and the DKPro JWKTL tool (Meyer and Gurevych, 2012) for Wiktionary, where both tools enable optimized search through the large Wiki data volume.

4.1.2. Annotation Procedure for Entity Resolution

Through iterative pilot annotation trials on the same pilot dataset as before, the annotators delineated an ordered annotation procedure depicted in the flowchart in Figure 2. There are two main annotation phases, viz. a preprocessing phase (determining linkability, determining whether an entity is decomposable into shorter collocations), and the entity resolution phase.

The actual annotation task then proceeded, in which to compute agreement scores, the annotators worked on the same set of 50 scholarly abstracts that they had used earlier to compute the scores for the scientific entity annotations.

Linkability. In this first step, *entities that conveyed a sense of scientific jargon were deemed linkable.*

A natural question that arises, in the context of the *Linkability* criteria, is: Which stage 1 annotated scientific entities were now deemed unlinkable? They were 1) DATA entities that are numbers; 2) entities that are coreference mentions which, as isolated units, lost their precise sense

[6] Excel sheets were flexible to incorporate our various ER task annotations with separate columns reserved for a unique annotation. Existing tools (e.g., INCEpTION (De Castilho et al., 2018)) could not be adapted to our full task.

[7] https://dumps.wikimedia.org/enwiki/20190920/

[8] https://dumps.wikimedia.org/enwiktionary/20190920/

(e.g., "development"); and 3) PROCESS verbs (e.g., "decreasing", "reconstruct", etc.). Still, having identified these cases, a caveat remained: except for entities of type DATA, the remaining decisions made in this step involved a certain degree of subjectivity because, for instance, not all PROCESS verbs were unlinkable (e.g., "flooding"). Nonetheless, at the end of this step, the annotators obtained a high IAA score at 0.89 κ. From the agreement scores, we found that the *Linkability* decisions could be made reliably and consistently on the data.

Splitting phrases into shorter collocations. While preference was given to annotating non-compositional noun phrases as scientific entities in stage 1, consecutive occurrences of entities of the same concept type separated only by prepositions or conjunctions were merged into longer spans. As examples, consider the phrases "geysers on south polar region," and "plume of water ice molecules and dust" in Figure 1. These phrases, respectively, can be meaningfully split as "geysers" and "south polar region" for the first example, and "plume", "water ice molecules", and "dust" for the second. As demonstrated in these examples, the stage 1 entities we split in this step are syntactically-flexible multi-word expressions which did not have a strict constraint on composition (Sag et al., 2002). For such expressions, we query Wikipedia or Google to identify their splits judging from the number of results returned and whether, in the results, the phrases appeared in authoritative sources (e.g., as overview topics in publishing platforms such as ScienceDirect). Since search engines operate on a vast amount of data, they are a reliable source for determining phrases with a strong statistical regularity, i.e. determining collocations.

With a focus on obtaining agreement scores for entity resolution, the annotators bypass this stage for computing independent agreement and attempted it mutually as follows. One annotator determined all splits, wherever required, first. The second annotator acted as judge by going through all the splits and proposed new splits in case of disagreement. The disagreements were discussed by both annotators and the previous steps were repeated iteratively until the dataset was uniformly split. After this stage, both annotators have the same set of entities for resolution.

Entity Resolution (ER) Annotation. In this stage, the annotators resolved each entity from the previous step to encyclopedic and lexicographic knowledge bases. While, in principle, multiple knowledge sources can be leveraged, this study only examines scientific entities in the context of their Wiki-linkability.

Wikipedia, as the largest online encyclopedia (with nearly 5.9 million English articles) offers a wide coverage of real-world entities, and based on its vast community of editors with editing patterns at the rate of 1.8 edits per second, is considered a reliable source of information.[9] It is pervasively adopted in automatic EL tasks (Bunescu and Paşca, 2006; Mihalcea and Csomai, 2007; Rao et al., 2013) to disambiguate the names of people, places, organizations, etc., to their real-world identities. We shift from this focus on proper names as the traditional Wikification EL purpose has been, to its, thus far, seemingly less tapped-in conceptual encyclopedic knowledge of nominal scientific entities.

Wiktionary is the largest freely available dictionary resource. Owing to its vast community of curators, it rivals the traditional expert-curated lexicographic resource WordNet (Fellbaum, 1998) in terms of coverage and updates, where the latter evolves more slowly. For English, Wiktionary has nine times as many entries and at least five times as many senses compared to WordNet.[10] As a more pertinent neologism in the context of our STEM data, consider the sense of term "dropout" as a method for regularizing the neural network algorithms which is already present in Wiktionary.[11] While WSD has been traditionally used WordNet for its high-quality semantic network and longer prevalence in the linguistics community (c.f Navigli (2009) for a comprehensive survey), we adopt Wiktionary thus maintaining our focus on collaboratively curated resources.

In WSD, entities from all parts-of-speech are enriched w.r.t. language and wordsmithing. But it excludes in-depth factual and encyclopedic information, which otherwise is contained in Wikipedia. Thus, Wikipedia and Wiktionary are viewed as largely complementary.

ER Annotation Task formalism. Given a scholarly abstract A comprising a set of entities $E = \{e_1, ..., e_N\}$, the annotation goal is to produce a mapping from E to a set of Wikipedia pages $(p_1, ..., p_N)$ and Wiktionary senses $(s_1, ..., s_N)$ as $R = \{(p_1, s_1), ..., (p_N, s_N)\}$. For entities without a mapping, the corresponding p or s refers to NIL. The annotators followed comprehensive guidelines for ER including exceptions. E.g., the conjunctive phrase "acid/alkaline phosphatase activity" was semantically treated as the following two phrases "acid phosphatase activity" or "alkaline phosphatase activity" for EL, however, in the text it was retained as "acid" and "alkaline phosphatase activity." Since WSD is performed over exact word-forms without assuming any semantic extension, it was not performed for "acid." Annotations were also made for complex forms of reference such as meronymy (e.g., space instrument "CAPS" to spacecraft "wiki:Cassini Huygens" of which it is a part), or hypernymy (e.g., "parents" in "genepool parents" to "wiki:Ancestor").

As a result of the annotation task, the annotators obtained 82.87% rate of agreement in the EL task and a κ score of 0.86 in the WSD task. Contrary to WSD expectations as a challenging linguistics task (Ng et al., 1999), we show high agreement; this we attribute to the entities' direct scientific sense and availability in Wiktionary (e.g., "dropout").

Subsequently, the ER annotation for the remaining 60 abstracts (six per domain) were performed by one annotator. This last phase also involved reconciliation of the earlier annotated 50 abstracts to obtain a gold standard corpus.

4.1.3. Annotated Corpus Characteristics

In this stage 2 corpus, *linkability* of the scientific entities was determined at 74.6%. Of these, 61.7% were split into shorter collocations, at 1.74 splits per split entity. Detailed

[9] https://en.wikipedia.org/wiki/Wikipedia:Statistics

[10] https://en.wiktionary.org/wiki/Wiktionary:Statistics versus https://wordnet.princeton.edu/documentation/wnstats7wn

[11] https://en.wiktionary.org/wiki/dropout

	EL: total, %indomain, %overall	**WSD:** total, %indomain, %overall	**Total** (%)
ES	559 (80.3 / 9.9)	380 (54.6 / 6.8)	**696 (12.4)**
Ast	**583 (87.3 / 10.4)**	**477 (71.4 / 8.5)**	668 (11.9)
Agr	556 (83.3 / 9.9)	475 (71.2 / 8.4)	667 (11.9)
Bio	538 (88.2 / 9.6)	319 (52.3 / 5.7)	610 (10.8)
Eng	473 (80 / 8.4)	406 (68.7 / 7.2)	591 (10.5)
Med	476 (82.5 / 8.5)	320 (55.5 / 5.7)	577 (10.3)
MS	484 (85.1 / 8.6)	377 (66.3 / 6.7)	569 (10.1)
Che	381 (76.2 / 6.8)	304 (60.8 / 5.4)	500 (8.9)
CS	391 (81 / 6.9)	285 (59 / 5.1)	483 (8.6)
Mat	226 (85.3 / 4)	150 (56.6 / 2.7)	265 (4.7)
Overall	4667 (82.9 / -)	3489 (62.0 / -)	5627 (-)

Table 5: The annotated corpus characteristics in terms of its Entity Linking (EL) and Word Sense Disambiguation (WSD) annotations

Wikipedia		**Wiktionary**	
POS	%	POS	%
N	49.6	N	75.3
MWE	23.2	R	46.2
SW	11.8	ADJ	16.4
ADJ	9.7	SYM	3.8
SYM	2.8	V	2.3
V	1.3	NNP	1.6
NNP	1.2	INIT	0.2
R	0.3	PREP	0.03
INIT	0.2	PHRASE	0.03

Table 6: Part-of-speech tag distribution in our corpus. N - Noun; MWE - multiword expression; SW - single word; ADJ - adjective; SYM - symbol; V - verb; NNP - proper noun; R - adverb; INIT - initialism; PREP - preposition

statistics are presented in Table 5. In the table, the domains are ranked by the total number of their linkable entities (fourth column). *Ast* has the highest proportion of linked entities at 87.3% which comprises 10.4% of all the linked entities and disambiguated entities at 71.4% forming 8.5% of the overall disambiguated entities. From an EL perspective, we surmize that articles on space topics are well represented in Wikipedia. For WSD, *Bio*, *ES*, and *Med* predictably have the least proportion of disambiguated entities at 52.3%, 54.6%, and 55.5%, respectively, since of all our domains these especially rely on high degree scientific jargon, while WSD generally tends to be linguistically oriented in a generic sense. As a summary, linked and disambiguated entities had a high correlation with the total linkable entities (R 0.98 and 0.89, respectively).

In Table 6, the ER annotation results are shown as POS tag distributions. The POS tags were obtained from Wiktionary, where entities that couldn't be disambiguated are tagged as SW (Single Word) or MWE (Multi-Word Expression). These tags have a coarser granularity compared to the traditionally followed Penn Treebank tags with some unconventional tagging patterns (e.g., "North Sea" as NNP, "in vivo" as ADJ). From the distributions, except for nouns being the most EL and WSD instances, the rest of the table differs significantly between the two tasks in a sense reflecting the nature of the tasks. While MWE are the second highest EL instances, its corresponding PHRASE type is least represented in WSD. In contrast, while adverbs are the second highest in WSD, they are least in EL.

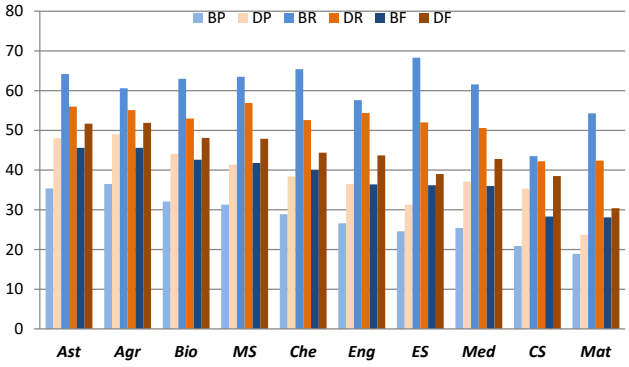

Figure 3: Percentage P, R, and $F1$ per domain of the Babelfy system for disambiguating to BabelNet (as shaded blue bars BP, BR, and BF) and for linking to DBpedia (as shaded orange bars DP, DR, DF)

4.2. Evaluation

We evaluate Babelfy (Moro et al., 2014a), a domain-agnostic, graph-based unified approach to EL and WSD. Since it is an unsupervised system, we do not retrain it on our data. We evaluate it for joint EL and WSD performance on *multidisciplinary* scientific entities.

Evaluation 1.1: Splitting phrases into substrings The first step of an end-to-end ER system like Babelfy is spotting the candidate text fragments to resolve. Babelfy domain-agnostically follows a substring matching heuristic which either at sentence-unit or phrase-unit has the same operating principle, i.e. fragmenting the input into non-overlapping longest matching linkable spans and recursively splitting each span to smaller content-word parts. Babelfy splits 96.2% of entities from stage 1 as opposed to 61.7% of them split by humans. Further, the annotators split entities at 1.74 parts per entity. On the other hand, Babelfy split them at 2.46 parts per entity (2.16 for non-overlapping fragments). Thus, evidently our human annotators have taken a conservative stance since they focused only on scientific entities, differing from the Babelfy approximately-all linkable domain-agnostic fragmentations (e.g., given "pearl millet", the annotators link it as it is, whereas Babelfy links it for "pearl" and "pearl millet").

In addition, Babelfy has a recall of 63.2% at precision 37.8% of the human-split entities. Thus, despite Babelfy's approximately-all linkable fragmentation, it shows a relatively low recall for our entities, since we note that Babelfy in most cases operates on at least partial lexical matches whereas several of our entities required inference decisions.

Evaluation 1.2: Entity Resolution In Figure 3, we depict the Babelfy ER performance in terms of the classical precision (P), recall (R) and $F1$ scores.[12]

EL versus WSD For all domains, EL is better than WSD in line with the claim that WSD is generally deemed a harder linguistics task than EL.[13] The highest EL score is obtained

[12] We provide details about our calculations in the Appendix.

[13] For EL, DBpedia and Wikipedia are equivalent resources; for

on *Agr* (at 51.9% $F1$), and the highest WSD score is on *Ast* and *Agr* (at 45.6% $F1$).

Most ambiguous resolution domain(s) *CS* and *Mat* were the most ambiguous resolution domains for Babelfy. This finding is consistent with that reported in the Semeval 2013 shared task (Moro and Navigli, 2015) for these two domains which are common in our datasets, albeit on different genre of text.

Finally, we did not find any consistent trend in the automatic resolution results to infer the influence of domain-specific terms on the performance.

Evaluation 2: EL and WSD baseline system accuracies for only the linked split entities

For EL, we select multiple baselines including the exact title match heuristic and four popular online systems that demonstrate state-of-the-art performance on non-scientific, open domain data. The baseline scores are: exact title match heuristic at 37.8% accuracy, and from the four automated systems, viz. Babelfy (2014a) to DBpedia at 52.6%, TagMe (Ferragina and Scaiella, 2010) at 40.5%, DBpedia Spotlight (Mendes et al., 2011) at 38.1%, and Falcon (Sakor et al., 2019) at 33.8%. Further in Figure 4, we show the performances of these systems obtained per domain. Of all four systems, Babelfy consistently has highest recall of the linked entities owing to its maximal linking objective, i.e. linking longest phrase matches and their sub-parts. Per-domain, Babelfy has highest recall on *MS*, TagME on *Agr*, DBpedia Spotlight on *Bio*, and Falcon on *Ast*, respectively.

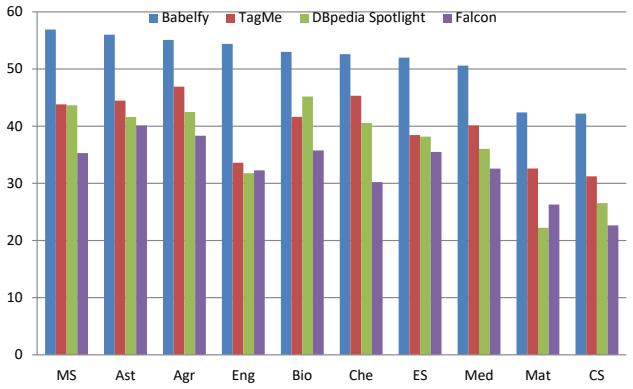

Figure 4: Percentage recall per domain of four popular online Entity Linking systems, viz. Babelfy (2014a), TagMe (Ferragina and Scaiella, 2010), DBpedia Spotlight (Mendes et al., 2011), and Falcon (Sakor et al., 2019)

For WSD, our baseline is just the first sense heuristic, since it is seen hard-to-beat for this task. It demonstrates 68.8% disambiguation accuracy.

We do not observe a significant impact of the long-tailed list phenomenon of unresolved entities in our data (c.f Table 5 only 17% did not have EL annotations). Results on more recent publications should perhaps serve more conclusive in this respect for new concepts introduced–the abstracts in our dataset were published between 2012 and 2014.

WSD, BabelNet draws from more than one source including Wiktionary which we consider.

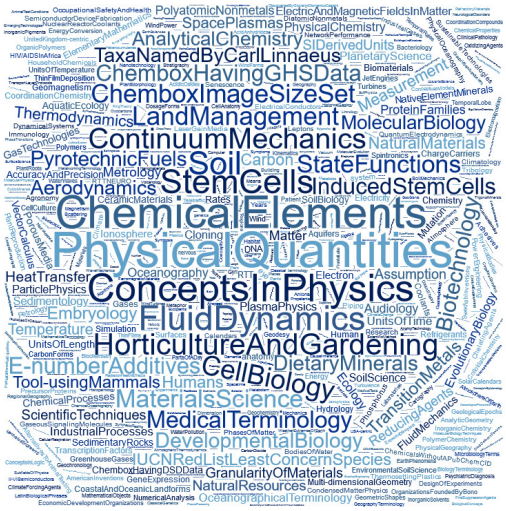

Figure 5: Wikipedia categories cloud on scientific entities

5. Conclusion

The STEM-ECR v1.0 corpus of scientific abstracts offers multidisciplinary PROCESS, METHOD, MATERIAL, and DATA entities that are disambiguated using Wiki-based encyclopedic and lexicographic sources thus facilitating links between scientific publications and real-world knowledge (see the concepts enrichment we obtain from Wikipedia for our entities in Figure 5). We have found that these Wikipedia categories do enable a semantic enrichment of our entities over our *generic* four concept formalism as PROCESS, MATERIAL, METHOD, and DATA (as an illustration, the top 30 Wiki categories for each of our four *generic* concept types are shown in the Appendix). Further, considering the various domains in our multidisciplinary STEM corpus, notably, the inclusion of understudied domains like Mathematics, Astronomy, Earth Science, and Material Science makes our corpus particularly unique w.r.t. the investigation of their scientific entities. This is a step toward exploring domain independence in scientific IE. Our corpus can be leveraged for machine learning experiments in several settings: as a vital active-learning test-bed for curating more varied entity representations (Brack et al., 2020); to explore domain-independence versus domain-dependence aspects in scientific IE; for EL and WSD extensions to other ontologies or lexicographic sources; and as a knowledge resource to train a reading machine (such as PIKES (Corcoglioniti et al., 2016) or FRED (Gangemi et al., 2017)) that generate more knowledge from massive streams of interdisciplinary scientific articles. We plan to extend this corpus with relations to enable building knowledge representation models such as knowledge graphs in a domain-independent manner.

6. Acknowledgements

We thank the anonymous reviewers for their comments and suggestions. We also thank the subject specialists at TIB for their helpful feedback in the first part of this study. This work was co-funded by the European Research Council for the project ScienceGRAPH (Grant agreement ID: 819536) and by the TIB Leibniz Information Centre for Science and Technology.

Appendix: Supplemental Material

A.1. Proportion of the *Generic* Scientific Entities

To offer better insights to our STEM corpus for its scientific entity annotations made in part 1, in Figure 6 below, we visually depict the proportion of PROCESS, METHOD, MATERIAL, and DATA entities per domain.

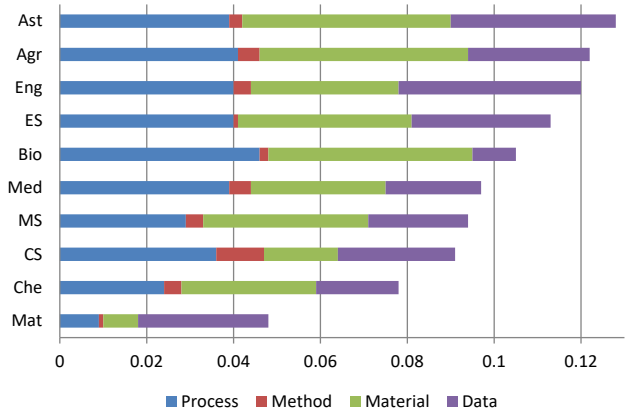

Figure 6: Percentage proportion of scientific entities in our STEM corpus per domain and per *generic* concept type

The Figure serves a complementary view to our corpus compared with the dataset statistics shown in Table 3. It shows that the *Ast* domain has the highest proportion of scientific entities overall. On the other hand, per *generic* type, *Bio* has the most PROCESS entities, *CS* has the most METHOD entities, *Ast* has the most MATERIAL closely followed by *Agr*, and *Eng* has the most DATA.

A.2. Cohen's κ Computation[14] Setup in Section 4.1.2

Linkability. Given the stage 1 scientific entities, the annotators could make one of two decisions: a) an entity is linkable; or b) an entity is unlinkable. These decisions were assigned numeric indexes, i.e. 1 for decision (a) and -1 for decision (b) and can take on one of four possible combinations based on the two annotators decisions: (1,1), (1,-1), (-1,1), and (-1,-1). The κ scores were then computed on this data representation.

WSD Agreement. In order to compute the WSD agreement, the Wiktionary structure for organizing the words needed to be taken into account. It's structure is as follows. Each word in the Wiktionary lexicographic resource is categorized based on etymology, and within each etymological category, by the various part-of-speech tags the word can take. Finally, within each POS type, is a gloss list where each gloss corresponds to a unique word sense.

Given the above-mentioned Wiktionary structure, the initial setup for the blind WSD annotation task entailed that the annotators were given the same reference POS tags within an etymology for the split single-word entities in the corpus.[15] Next, as data to compute κ scores, each annotator-assigned gloss sense was given a numeric index and agreement was computed based on matches or non-matches between indexes.

A.3. Per-domain Inter-annotator Agreement for Entity Resolution

To supplement the overall Inter-Annotator Agreement (IAA) scores reported in Section 4.1.2 'Entity Resolution (ER) Annotation' for the EL and WSD tasks, in Table 7 below, we additionally report the IAA scores for our ER tasks (i.e., EL and WSD) per domain in the STEM-ECR corpus. First, considering the domains where the highest ER agreement scores were obtained. For EL, the IAA score was highest in the *MS* domain. While for WSD, the IAA score was highest in the *Bio* domain. Next, considering the domains where the agreement was least for the two tasks. We found the the EL agreement was least for *CS* and the WSD agreement was least for *Mat*. In the case of low EL agreement, it can be attributed to two main cases: only one of the annotators found a link; or the annotators linked to related pages on the same theme as the entity (e.g., wiki:Rule-based_modeling versus wiki:Rule-based_machine_learning for "rule-based system"). And in the case of the low WSD agreement obtained on *Mat*, we see that owing to broad terms like "set," "matrix," "groups," etc., in the domain which could be disambiguated to more than one Wiktionary sense correctly, the IAA agreement was low.

	Wikipedia roa	Wiktionary κ
Agr	83.88	0.88
Ast	79.8	0.85
Bio	84.73	**0.93**
Che	86.83	0.85
CS	72.58	0.86
ES	82.17	0.83
Eng	78.83	0.84
MS	**88.24**	0.83
Mat	87.37	0.81
Med	87.89	0.87

Table 7: Per-domain inter-annotator agreement for Entity Linking to Wikipedia in terms of the percentage rate of agreement score (roa) and of Word Sense Disambiguation to Wiktionary in terms of the Cohen's kappa score (κ)

A.4. Babelfy's Precision (P) and Recall (R) Computation for Entity Resolution in Figure 3

For the P and R scores reported in Figure 3, the true positives (*TP*), false negatives (*FN*), true negatives (*TN*), and false positives (*FP*) were computed as follows:

TP = human-annotated entities that have a EL/WSD match with Babelfy results (for NIL, a match is considered as no result from the automatic system);

FN = human-annotated entities that have no EL/WSD match with Babelfy results;

TN = spurious Babelfy-created strings as entities that do not have a EL/WSD result; and

FP = spurious Babelfy-created entities that have a EL/WSD result.

[14] Cohen's κ scores are computed using the following script: https://github.com/jorgearanda/kappa-stats

[15] In separate experiments for determining etymological and POS agreement between the annotators, we found they had an insignificant disagreement (i.e. in less than 15 instances in approximately over the 2000 they annotated.)

PROCESS		METHOD	
PhysicalQuantities	ElectricAndMagneticFields-InMatter	NumericalDifferentialEquations	SystemsBiology
Mutation	Mechanics	ComputationalFluidDynamics	Petrology
ConceptsInPhysics	GenesOnHumanChromosome	ScientificTechniques	FiniteDifferences
MathematicalAnd-QuantitativeMethods	Software	Measurement	Spectroscopy
Electron	ChemicalElements	LandManagement	Teletraffic
MolecularEvolution	QuantumElectrodynamics	TransportLayerProtocols	DataTransmission
DevelopmentalBiology	Aerodynamics	Metabolism	ProteinStructure
VectorCalculus	Spintronics	EnzymeInhibitors	Optics
RadiationHealthEffects	AllAccuracyDisputes	NetworkPerformance	MolecularBiology
EvolutionaryBiology	Oceanography	Research	OperationsResearch
Rates	Dementia	MathematicalAnd-QuantitativeMethods	AllAccuracyDisputes
ChargeCarriers	Audiology	TechnologicalFailures	Stratigraphy
Leptons	CognitiveDisorders	ModelingAndSimulation	SoilImprovers
PsychiatricDiagnosis	WaterPollution	MathematicalOptimization	Solid-stateChemistry
FluidDynamics	StructuralProteins	MedicinalChemistry	Polarization(Waves)

Table 8: Top 30 Wikipedia categories pertaining to PROCESS and METHOD scientific entities in the STEM-ECR corpus

MATERIAL		DATA	
StemCells	NaturalResources	PhysicalQuantities	OrdersOfMagnitude
CellBiology	FluidDynamics	ConceptsInPhysics	UnitsOfTime
InducedStemCells	Matter	SIBaseQuantities	FluidDynamics
Biotechnology	GranularityOfMaterials	ContinuumMechanics	Geochronology
DevelopmentalBiology	ChemboxHavingGHSData	SIDerivedUnits	Density
Cloning	Polymers	PyrotechnicFuels	PartsOfADay
HorticultureAndGardening	CeramicMaterials	ChemicalElements	UnitsOfLength
BiologyAndPharmacologyOf-ChemicalElements	Adhesives	StateFunctions	MathematicalConstants
LandManagement	SpacePlasmas	ChemicalProperties	Length
HumanEyeAnatomy	TaxaNamedByCarlLinnaeus	Research	ImperialUnits
ChemicalElements	ChelatingAgents	UnitsOfTemperature	Symptoms
Soil	ChemboxHavingDSDData	Metrology	Calendars
MedicalTerminology	Humans	Size	UnitsOfLength
ChemboxImageSizeSet	Neurons	Measurement	UnitsOfPlaneAngle
NaturalMaterials	E-numberAdditives	ChemboxImageSizeSet	SolarCalendars

Table 9: Top 30 Wikipedia categories pertaining to MATERIAL and DATA scientific entities in the STEM-ECR corpus

A.5. Top 30 Wikipedia Categories for PROCESS, METHOD, MATERIAL, and DATA

In part 1 of the study, we categorized the scientific entities by our four *generic* concept formalism, comprising PROCESS, METHOD, MATERIAL, and DATA. Linking the entities to Wikipedia further enables their broadened categorization. While in Figure 5 is depicted the rich set of Wikipedia categories obtained overall, here, in Tables 8 and 9, we show the top 30 Wikipedia categories for the scientific entities by their four concept types. we observe the most of the Wikipedia categories pertinently broaden the semantic expressivity of each of our four concepts. Further that in each type, they are diverse reflecting the underlying data domains in our corpus. As examples, consider the Wikipedia categories for the DATA scientific entities: "SIBaseQuantities" category over the entity "Kelvin" in *Che*; "FluidDynamics" in *Eng* and *MS* domains; and "SolarCalendars" in the *Ast* domain.

7. Bibliographical References

Ammar, W., Peters, M. E., Bhagavatula, C., and Power, R. (2017). The ai2 system at semeval-2017 task 10 (scienceie): semi-supervised end-to-end entity and relation extraction. In *SemEval@ACL*.

Ammar, W., Groeneveld, D., Bhagavatula, C., Beltagy, I., Crawford, M., Downey, D., Dunkelberger, J., Elgohary, A., Feldman, S., Ha, V., et al. (2018). Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Augenstein, I., Das, M., Riedel, S., Vikraman, L., and McCallum, A. (2017). Semeval 2017 task 10: Scienceie - extracting keyphrases and relations from scientific publications. In *SemEval@ACL*.

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611.

Brack, A., D'Souza, J., Hoppe, A., Auer, S., and Ewerth, R. (2020). Domain-independent extraction of scientific concepts from research articles. *arXiv preprint arXiv:2001.03067*.

Bunescu, R. and Paşca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *11th conference of the European Chapter of the Association for Computational Linguistics*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Constantin, A., Peroni, S., Pettifer, S., Shotton, D., and Vitali, F. (2016). The document components ontology (doco). *Semantic web*, 7(2):167–181.

Corcoglioniti, F., Rospocher, M., and Aprosio, A. P. (2016). Frame-based ontology population with pikes. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3261–3275.

De Castilho, R. E., Klie, J.-C., Kumar, N., Boullosa, B., and Gurevych, I. (2018). Linking text and knowledge using the inception annotation platform. In *2018 IEEE 14th International Conference on e-Science (e-Science)*, pages 327–328. IEEE.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Christiane Fellbaum, editor. (1998). *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press.

Ferragina, P. and Scaiella, U. (2010). Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM.

Fisas, B., Ronzano, F., and Saggion, H. (2016). A multi-layered annotated corpus of scientific papers. In *LREC*.

Gábor, K., Buscaldi, D., Schumann, A.-K., QasemiZadeh, B., Zargayouna, H., and Charnois, T. (2018). Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688.

Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A. G., Draicchio, F., and Mongiovì, M. (2017). Semantic web machine reading with fred. *Semantic Web*, 8(6):873–893.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., Peters, M., Schmitz, M., and Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Getoor, L. and Machanavajjhala, A. (2012). Entity resolution: theory, practice & open challenges. *Proceedings of the VLDB Endowment*, 5(12):2018–2019.

Handschuh, S. and QasemiZadeh, B. (2014). The acl rd-tec: a dataset for benchmarking terminology extraction and classification in computational linguistics. In *COLING 2014: 4th international workshop on computational terminology*.

Hovy, E., Navigli, R., and Ponzetto, S. P. (2013). Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

Jaradeh, M. Y., Oelen, A., Farfar, K. E., Prinz, M., D'Souza, J., Kismihók, G., Stocker, M., and Auer, S. (2019). Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 243–246, New York, NY, USA. ACM.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). Genia corpus–a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Kim, S. N., Medelyan, O., Kan, M.-Y., and Baldwin, T. (2010). Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 21–26.

Krupp, N., Roussos, E., Kollmann, P., Paranicas, C., Mitchell, D., Krimigis, S., Rymer, A., Jones, G., Arridge, C., Armstrong, T., et al. (2012). The cassini enceladus encounters 2005–2010 in the view of energetic electron measurements. *Icarus*, 218(1):433–447.

Liakata, M., Teufel, S., Siddharthan, A., and Batchelor, C. R. (2010). Corpora for the conceptualisation and zoning of scientific papers. In *LREC*.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C., and Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Luan, Y., Ostendorf, M., and Hajishirzi, H. (2017). Scientific information extraction with semi-supervised neural tagging. *arXiv preprint arXiv:1708.06075*.

Luan, Y., He, L., Ostendorf, M., and Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *EMNLP*.

Ma, X. and Hovy, E. H. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8. ACM.

Meyer, C. M. and Gurevych, I. (2012). Wiktionary: A new rival for expert-built lexicons? exploring the possibilities

of collaborative lexicography. In *Electronic Lexicography*. Oxford Scholarship Online.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM.

Moro, A. and Navigli, R. (2015). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 288–297.

Moro, A., Cecconi, F., and Navigli, R. (2014a). Multilingual Word Sense Disambiguation and Entity Linking for Everybody. In *Proceedings of the 13th International Semantic Web Conference, Posters and Demonstrations (ISWC 2014)*, pages 25–28, Riva del Garda, Italy.

Moro, A., Raganato, A., and Navigli, R. (2014b). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):10.

Ng, H. T., Lim, C. Y., and Foo, S. K. (1999). A case study on inter-annotator agreement for word sense disambiguation. In *SIGLEX99: Standardizing Lexical Resources*.

Pertsas, V. and Constantopoulos, P. (2017). Scholarly ontology: modelling scholarly practices. *International Journal on Digital Libraries*, 18(3):173–190.

Rao, D., McNamee, P., and Dredze, M. (2013). Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, multilingual information extraction and summarization*, pages 93–115. Springer.

Rosales-Méndez, H., Hogan, A., and Poblete, B. (2019). Fine-grained evaluation for entity linking. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 718–727, Hong Kong, China, November. Association for Computational Linguistics.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer.

Sakor, A., Mulang, I. O., Singh, K., Shekarpour, S., Vidal, M. E., Lehmann, J., and Auer, S. (2019). Old is gold: linguistic driven approach for entity and relation linking of short text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2336–2346.

Singhal, Amit. (2012). Introducing the knowledge graph: Things, not strings. [Online; accessed 22-November-2019].

Soldatova, L. N. and King, R. D. (2006). An ontology of scientific experiments. *Journal of the Royal Society, Interface*, 3 11:795–803.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107. Association for Computational Linguistics.

Teufel, S., Carletta, J., and Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 110–117. Association for Computational Linguistics.

Usbeck, R., Röder, M., Ngonga Ngomo, A.-C., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., et al. (2015). Gerbil: general entity annotator benchmarking framework. In *Proceedings of the 24th international conference on World Wide Web*, pages 1133–1143. International World Wide Web Conferences Steering Committee.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *LREC*, volume 8, pages 1646–1652.