MS-MENTIONS: Consistently Annotating Entity Mentions in Materials Science Procedural Text

Tim O'Gorman[3]*, Zach Jensen[2], Sheshera Mysore[1], Rubayyat Mahbub[2], Kevin Huang[2], Elsa Olivetti[2], Andrew McCallum[1]
[1]UMass Amherst [2]MIT [3]Thorn
{smysore,mccallum}.cs.umass.edu
{zjensen,kjhuang,rmahbub,elsao}@mit.edu

Abstract

Material science synthesis procedures are a promising domain for scientific NLP, as proper modeling of these recipes could provide insight into new ways of creating materials. However, a fundamental challenge in building information extraction models for material science synthesis procedures is getting accurate labels for the materials, operations, and other entities of those procedures. We present a new corpus of entity mention annotations over 595 Material Science synthesis procedural texts (157,488 tokens), which greatly expands the training data available for the Named Entity Recognition task. We outline a new label inventory designed to provide consistent annotations and a new annotation approach intended to maximize the consistency and annotation speed of domain experts. Inter-annotator agreement studies and baseline models trained upon the data suggest that the corpus provides high-quality annotations of these mention types. This corpus helps lay a foundation for future high-quality modeling of synthesis procedures.

1 Introduction

The Material Science literature contains millions of *synthesis procedures*: descriptions that outline the specific steps required to create a particular material, such as the text in Figure 1. Large-scale analysis of these procedures could enable tasks such as automatic planning of new synthesis procedures (Kim et al., 2020). However, such tasks require extraction of high-quality representations of those synthesis procedures from raw text (Kim et al., 2017), and thus are limited by the size and quality of annotations.

An important component for information extraction of synthesis procedures involves predicting the entities, actions and attributes of the synthesis, including the steps followed in the synthesis procedures, the materials used and created in the synthesis, and all the quantitative attributes (including operation conditions, material amounts, and properties) necessary to replicate or understand the procedure. Such annotations require the judgment of domain experts, as not all verbs in a text are actual steps in the synthesis, and not all materials mentioned are actual inputs or outputs of the synthesis process. We outline an inventory of 15 mention types which can be consistently annotated while still making important distinctions regarding the roles that these operations and mentions play in the overall synthesis. Figure 1 illustrates how such an example synthesis procedure would be annotated with entity mentions.

We annotate these labels over a new corpus of 595 synthesis procedures. To improve the consistency of this annotation pipeline, we separate the annotation of operation-type mentions, material-type mentions, and quantitative mentions (including conditions, amounts and properties), allowing annotators to specialize upon a subset of the task and reducing inter-annotator variation in how each

P2- Na2/3Ni1/4TixMn3/4-xO2 was prepared through a simple solid state method. The precursor solution was prepared by mixing desirable amount of Ni(CH3COO)2*4H2O, Mn(CH3COO)2*4H2O and CH3COONa and titanium citrate solution. The obtained mixture was heated at 400 degC for 12 h. The ground powder was ball-milled for 1 h and was subsequently calcinated at 900 degC in air for 12 h to synthesize Na2/3Ni1/4TixMn3/4-xO2 (x=0, 0.05, 0.10, 0.15, 0.20, 0.30).

Figure 1: Part of an example synthesis procedure included in the dataset with entity annotations from Zhao et al. (2015). Colors represent entity types and underlines represent span boundaries. Colors: Target, Nonrecipe-operation, Unspecified-Material, Operation, Material, Condition-Unit, Number.

* Work done while at UMass Amherst

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1337–1352
November 7–11, 2021. 2021 Association for Computational Linguistics

phenomenon is annotated. We present baseline models trained on this dataset which demonstrate that models can be trained upon this dataset to achieve mention extraction performance exceeding 90 F1, providing a reliable base for future work.

The contributions of the work are as follows. First, we establish a large new dataset of Material Science mention annotations*, which includes both texts sampled from a wide range of different synthesis procedures. Secondly, we provide improvements to synthesis procedure annotation, both by providing a new label set that can be consistently annotated and by providing new approaches to expert mention annotation. Thirdly, we provide simple baseline models to be made public with the paper, and use those baseline models to illustrate both the remaining challenges with the data, as well as the promise this dataset holds for high-performing mention extraction over synthesis procedures.

2 Dataset Description

We annotate three different kinds of mentions: the individual steps or operations constituting the synthesis procedure, the different kinds of materials mentioned in a procedure, and a third class of quantitative mentions such as measurable conditions, quantities and apparatus mentions. These three broad classes are each annotated in a different stage, as outlined in §3. Note that the current dataset does not annotate relations between mentions as in Mysore et al. (2019), but focuses on annotating the much larger set of operation, material and condition mentions to be linked together.

2.1 Structures Annotated

2.1.1 Operation Annotation

A major component of annotating procedural scientific text is the annotation of operations — specific mentions of the steps taken during a synthesis procedure — since these form the primary structure of the synthesis procedure. We define three operation types:

`Operation`: Discrete actions physically performed by the researcher or discrete process steps taken to synthesize the target.

`Nonrecipe-Operation`: Verbs or action words that were not directly carried out by the researcher, or a reference to an operation with more descriptive wording.

`Meta`: A canonical name to specify a particular overall synthesis method used for synthesis. For example: "Graphite oxide was prepared by oxidation of graphite powder according to the modified Hummers method."

Although making this distinction between different types of events based upon their role in the procedure is novel in the context of synthesis procedure annotation (Mysore et al., 2019), domains such as newswire have generally established other tasks where one needs to go beyond a general notion of nouns and verbs being eventualities (Pustejovsky et al., 2003) and focus on events relevant to a specific task (Doddington et al., 2004). In this domain, many events describe processes which are not being enacted by the researchers, and therefore are not part of the list of "steps" that define a procedure, as seen in example 1:

(1) "After this, the autoclave was cooled to room temperature naturally."

A related issue occurs in which multiple words or phrases collectively refer to the same step in a synthesis process. In some of these situations one mention does not describe the step solely in terms of the researcher action. For example, in "prepared by mixing" or "heated to evaporate", "mixing" or "heated" are annotated as the `Operation` and semantically light non-recipe processes such as "prepared" and "evaporate" are annotated as `Nonrecipe-Operation`. In some other cases where both words, by themselves, would serve as descriptive and desirable operations for downstream analysis, both are labeled as operations, for example: "separated by filtration", "mixed by ultrasonication", or "mixed by stirring".

All operation types were annotated using minimal spans, i.e. annotating only the predicative trigger without any larger verbal nor nominal spans. An exception applies to manners and instruments hyphenated with a predicate, as in "heat-treated" or "ball-milled"; these were included in the operation mention spans.

2.1.2 Materials Annotation

Our dataset annotates a variety of material types, the distinctions of which were based primarily on usefulness to downstream materials science tasks. The set of types we define are as follows:

`Target`: Indicates the chemical entity made in the context of the synthesis procedure.

*Available at https://github.com/olivettigroup/

Dataset	Type of Text	Genre	Texts	Sentences	Tokens	Mentions	Relations
SOFC-Exp	Articles	Mat.Sci.	45	853	32428	5095	5095
SC-CoMIcs	Abstracts	Mat.Sci.	1000	6639	204884	42337	-
MSPT	Procedural	Mat.Sci.	230	2112	56,510	20849	18402
MS-MENTIONS	Procedural	Mat.Sci.	595	7980	157488	44295	-
WLP	Procedural	Organic	622	13679	177770	60721	42425
WLP 2020 test	Procedural	Organic	+111	3562	51688	104654	70591

Table 1: Corpus statistics for our dataset (MS-MENTIONS) and a range of related corpora for materials science (SOFC-Exp, SC-CoMIcs, MSPT) and biomedical (WLP) procedural texts.

Type	Train	Dev	Test
Operation	10744	1364	1252
Nonrecipe-Operation	2608	301	334
Meta	560	71	68
Material	6132	769	786
Target	1529	196	176
Unspecified-Material	2685	352	364
Nonrecipe-Material	1071	115	106
Sample	280	67	39
Number	7444	946	903
Condition-Unit	3628	426	403
Amount-Unit	2484	325	313
Synthesis-Apparatus	1206	163	159
Property-Unit	288	39	41
Apparatus-Unit	185	27	31

Table 2: Counts for each mention type

`Material`: A physically used, chemically defined object used in the synthesis but not the end result of the synthesis.

`Unspecified-Material`: A material stated without sufficient chemical specificity, often referring to intermediate states such as "the mixture","a dilute solution" or "the disk".

`Nonrecipe-Material`: A material mentioned that is not contributing to the target material, such as impurities filtered out of a solution or reference to alternative materials that were not used.

`Sample`: A material or minor variant of the target referenced with a potentially arbitrary signifier such as "Sample A", "S1", or "undoped TiO2".

These distinctions allow us to understand the role of these materials within the synthesis procedure. While some of these distinctions represent information that would be relations in other datasets (such as `Target`, which would be a `recipe-target` relation in Mysore et al. (2019)), others introduce new important distinctions. In particular, the new `Sample` label would be important for researchers attempting to link the outcomes of a particular synthesis procedure to other paper components, such as results tables, where such sample names are used.

In annotating materials, all materials are labeled with only the base chemical composition of the material, omitting modifiers and relative clauses that might describe the chemical, structural, or morphological modifications to that chemical. However, `Target` materials, because of their nature as the goal of a synthesis procedure and as a complex material, often end up with a more complex base description than other material type mentions, but are annotated using the same criterion.

2.1.3 Quantity and Instrument Annotation

In addition to operation and material entities our dataset annotates a range of other entity types. We utilized a range of labels for conditions of synthesis operations (`Condition-Unit`), instruments used during the synthesis (`Synthesis-Apparatus`), measurements of the apparatus (`Apparatus-Unit`), properties asserted about a material or material quantities (`Property-Unit`, `Amount-Unit`) and numbers to be linked to these units (`Number`). Appendix C.3 expands on these entity types.

2.2 Dataset Statistics

Table 1 outlines the resultant size of our corpus, MS-MENTIONS. We also list the corpus statistics for the Materials Science Procedural Text (MSPT) corpus described in Mysore et al. (2019), the SOFC-Exp Corpus of Friedrich et al. (2020), the SC-

CoMIcs corpus of Yamaguchi et al. (2020) and the Wet Lab Protocols (WLP) corpus of Kulkarni et al. (2018), which has also been augmented for a recent WNUT shared task (Tabassum et al., 2020). Of these, MSPT is the most comparable annotation in structures annotated and domain of text. WLP bears resemblance in most of its annotated structures, but is in a different domain, while both SOFC-Exp and SC-CoMIcs are in the same Material Science domain, but do not focus upon procedural text. The current corpus is therefore the largest corpus we are aware of for materials science procedural text. Furthermore, our corpus also spans a range of subdomains within materials science whereas SOFC-Exp and SC-CoMIcs span a more focused set of sub-domains. We elaborate on this in the following section.

3 Annotation Pipeline

The present dataset was annotated by 3 domain experts using the BRAT annotation tool[†], using non-nested mentions. In building the dataset, papers for annotation were picked to contain a mix of randomly selected papers and those from a more focused sub-domain of papers. Furthermore, to improve consistency and speed, the different classes of entity types were annotated by the same annotator. Each of these processes is elaborated on next.

3.1 Data Selection and Filtering

The 595 synthesis procedures annotated were selected from a database of over 3 million publications describing materials synthesis. This collection was obtained from journals containing material science and chemistry content through API access, web scraping, or direct contact with publishers. Legal agreements with publishers allowed access to closed access journals. Given this collection, the 595 synthesis procedures were picked using two approaches, an ALL DOMAINS approach in which 338 papers were selected randomly to provide a broad characterization of the field, and a BATTERY subset, in which 257 papers were sampled using keywords such as "Li battery" or "Li10GeP2S12". By providing this split, one can measure both the ability of models to learn a single focused domain, and whether models generalize across the broader domain.

Given the papers, synthesis procedures were extracted from each publication using a paragraph classifier (Mahbub et al., 2020) leveraging a pipeline of a high recall rule-based approach relying on section header information to find possible synthesis paragraphs, followed by a trained neural network classifier to make a final prediction. This paragraph classifier, which reports F1 of 0.96 for all paragraphs, and F1 of 0.90 for synthesis paragraphs, was used to select synthesis paragraphs, and only those were annotated; each synthesis paragraph was also manually checked before annotation. Synthesis procedures were ruled out if they did not describe the synthesis as a series of operations, as with procedures which only described a single high-level operation name with a variety of conditions. A more detailed account of preprocessing is included in Appendix B.3.

3.2 Stages of Annotation

Situations in which domain experts are needed for annotation add considerably to the cost and difficulty of a project, and we focus on allowing targeted use of expert annotation, by having annotators focus upon restricted components of the annotation pipeline. Our annotation procedure adopted a division-of-labor approach to annotation, in which a single annotator is employed for a given type of annotation across the entire corpus. In the case of this dataset, one domain expert annotated all operations entities (§2.1.1), a second expert then annotated all materials (§2.1.2), and a third expert added all apparatuses, conditions, numbers, amounts and properties (§2.1.3). After these passes, each document was checked for overall quality and consistency by one of the domain experts.

While the more consistent annotations achieved by this approach do not guarantee higher quality, we still observe high inter-annotator agreement scores (§7), and suggest that it is a valuable approach for expert annotation contexts. Having each phenomenon in the dataset annotated by a single expert allows us to minimize inter-annotator inconsistencies between annotators, and means that if a particular aspect of the annotation requires specialized knowledge, an appropriate expert can focus on it, in a manner similar to allocating different tasks to different expertise levels during crowdsourcing (Wang et al., 2017). This approach of dividing the mention annotation tasks also increases the speed and efficiency of annotation, by reducing the number of different phenomena each annotator must pay attention to. Because minimizing inter-

[†]BRAT: https://brat.nlplab.org/

annotator inconsistencies does not guarantee high quality by itself, different annotators trained for each task and distinctions were iterated over during annotation, in order to improve quality.

4 Baselines

4.1 Baseline Models

We report baselines using pre-trained transformer models, fine-tuned on the NER task. Each sentence after tokenization (see below) is encoded with a pre-trained transformer model (Vaswani et al., 2017) using the Transformers library(Wolf et al., 2020), and the final hidden layer is regularized with dropout and passed through a final linear prediction layer. The output is optimized with cross-entropy, using AdamW(Loshchilov and Hutter, 2019).

4.2 Data preprocessing and task evaluation

We split each raw text into sentences using SciSpacy(Neumann et al., 2019), tokenize using Huggingface Tokenizers models corresponding each encoder, and predict IOB2 tags. In order to establish maximally comparable baseline scores, we do not hard-code any given tokenization scheme for evaluation, but instead convert IOB2 tags into start and end offsets and evaluate against the standoff annotations themselves, reporting exact-match micro-F1 for all scores.

4.3 Additional Tests

We explore one preliminary options for improving performance on the dataset, as it also serves to give insight into the data: pre-training upon one of the related datasets before fine-tuning on the dataset. For the transfer from related datasets, we use simple hard parameter sharing (Caruana, 1997), keeping the same encoder but using a different final linear prediction layer for each task. These result are showns at the bottom of Table 4, designating each dataset (+SOFC or +MSPT).

5 Results and Experiments

To contextualize the provided baseline scores on our new task, Table 3 provides scores on three comparable datasets, MSPT (Mysore et al., 2019), SOFC (Friedrich et al., 2020), and the 2020 WNUT Wet Labs Protocols data (Tabassum et al., 2020), along with the current state of the art for each.

We then show the performance of this baseline model on the MS-MENTIONS task, using BERT (Devlin et al., 2019), SciBERT (Beltagy et al., 2019), and Electra (Clark et al., 2020). SciBERT shows the best performance at the task, a trend noted in other recent procedural text works and shared tasks Tabassum et al. (2020); Friedrich et al. (2020).

Model/Corpus	Dev	Test
WLP (WNUT-2020)		
Baseline - SciBERT	77.96	73.22
Singh and Wadhawan (2020)		77.99
Knafou et al. (2020)		77.57
SOFC - mention		
Baseline - SciBERT	73	78.57
Friedrich et al. (2020)		81.5
MSPT(Mysore et al., 2019)		
Baseline - SciBERT	82.8	78.15
Friedrich et al. (2020)		92.2
Mysore et al. (2017)		77.6

Table 3: Similar procedural text datasets with mention spans, along with current top scores for each task.

5.1 Results on Subsets

We also report against the two subsets of the data, the ALL DOMAINS subset randomly sampled from all material science, and the BATTERY subset. We suggest that evaluating against these splits provides interesting ways of testing the robustness of models, and provides an interesting test bed for studying the role of narrow-domain data. In particular, it has been argued that model evaluation should also focus upon the performance of models in different kinds of minor domain shifts (Søgaard et al., 2021). To study this more closely, we experiment with different training set sizes: for sets of 25 documents, 50 documents, 100 documents and 200 documents, we train upon both domains and evaluate those trained models. Figure 3 shows F1-micro across four conditions (both within and across each subset) against the development splits. Curiously, we can see that while the BATTERY subset is unambiguously more difficult, training upon those battery papers does not seem to provide a consistent improvement, and training upon such a narrow subset, unsurprisingly, hurts general performance.

5.2 Results Across Corpora

Similarly, we briefly examine how this corpus compares to a similar dataset, MSPT (Mysore et al., 2019), when one controls for the much larger size

Corpus	Dev.	Test
BERT	89.4	89.7
ELECTRA-BASE	88.8	88.6
SCIBERT	91.55	91.47
SCIBERT+MSPT	91.72	91.53
SCIBERT+SOFC	91.80	91.85
SCIBERT+MSPT+SOFC	91.8	91.2

Table 4: Baseline model results for the current dataset, showing different transformers (top), pre-training with other datasets (middle), and scores training and testing on subsets of the data (bottom)

of the current corpus. As with analysis of subdomain, Figure 2 shows the results of training upon increasing sizes of training data from each corpus, evaluated against development sets.

We can see that even after controlling for training data size, the predictions of a model trained on MS-MENTIONS are more consistent than one trained on MSPT. To exclude higher performance on easier categories such as Number, we also show performance on `Operation` and `Material` types.

6 Analysis of Annotator Agreement

As the pipeline uses a fixed annotator for each part of the mention annotation, the inter-annotator agreement could not simply be measured by "double-annotating" documents. To measure the inter-annotator agreement we instead annotated a set of five documents three times each, with different assignments of annotators for each stage of our division of labor approach (§3.2). We then score IAA using F1, selecting pairs of annotations, and taking turns treating one as the gold annotation and the next as a prediction, these individual pairs are then micro-averaged. The overall F1 between annotation sets on the span prediction task is 85.7. Further, when annotators both agree that a given span exists, we measure a Fleiss's Kappa of 0.928. As an alternative measure of chance-corrected agreement, we use Mathet's γ (Mathet et al., 2015; Titeux and Riad, 2021), which allows for chance-corrected measures for span-based tasks such as NER. We report a γ of 0.76 (wherein 0 is chance and 1 is perfect agreement), showing substantial agreement.

Next we also denote labels for which annotators disagree in Figure 5. One can see that the primary sources of label disagreements is the judgments about whether a mention is integral to the synthesis procedure, seen in the distinction between `Operation` and `Nonrecipe-Operation` labels, and between `Materials` and `Nonrecipe-Material` and `Unspecified-Material` labels. Examples of such disagreements for operations (ex. 2) and materials (ex. 3) are shown below

(2) After evaporation of the solvent, a brown tetraethylammoniumtricyanoimidazolate was obtained

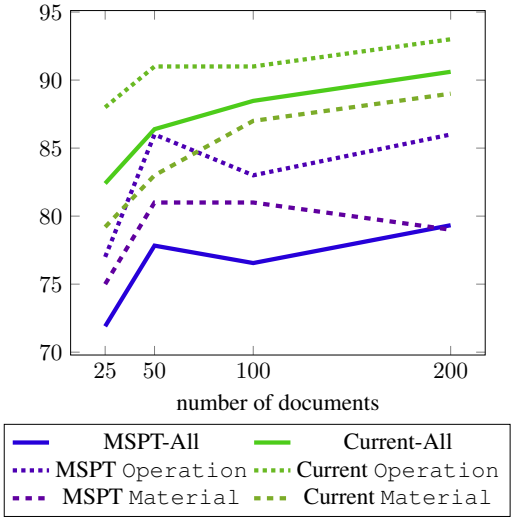

Figure 2: F1 against development set as training set size is increased, showing that the current work produces higher performance independent of the larger dataset size. MSPT refers to the similar, but not quite comparable, dataset of Mysore et al. (2019).

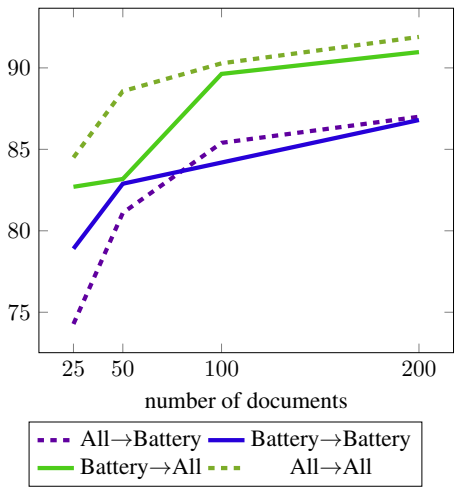

Figure 3: Peerformance within and across the two MS-Mentions subdomains on the development data, as number of training documents increases.

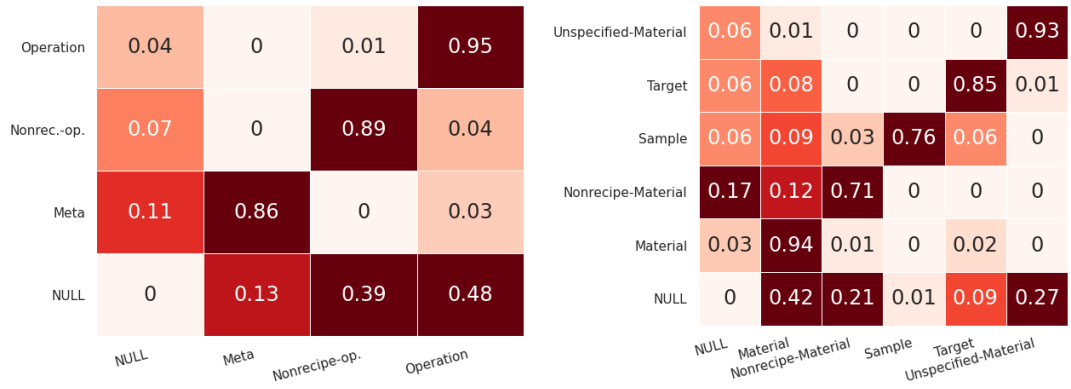

Figure 4: Confusion matrix of baseline prediction model against the development set. As with humans, Targets and non-recipe operations are the most challenging.

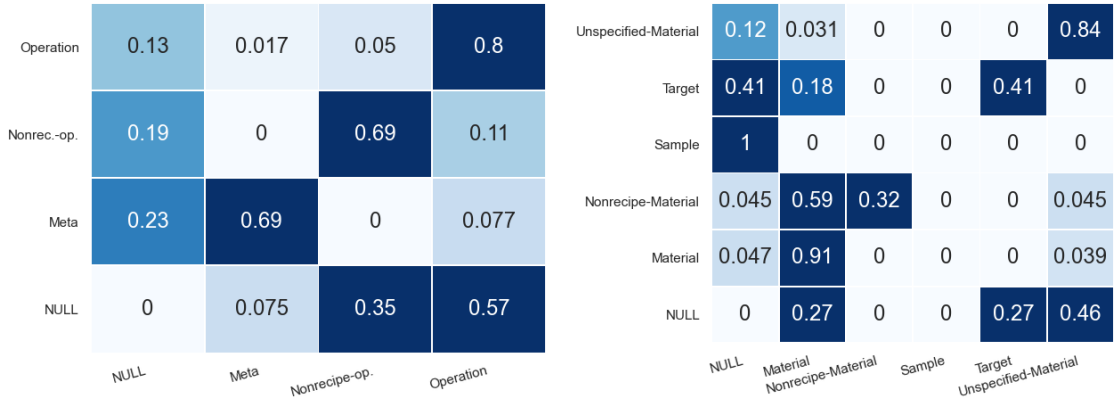

Figure 5: Confusion matrices for inter-annotator agreement regaarding for operations (left) and materials (right). Rows denotes an arbitrary annotation picked as gold to be compared to the other annotation. The primary source of disagreement for both is the challenging non-recipe vs recipe distinction.

(3) The crude product was yield-purified by column chromatography silica gel, gradient mixtures of acetonitrile-toluene-1:3,1:2, 1:1 (v/v))

Next, following prior analyses in parsing and SRL (Kummerfeld et al., 2012; He et al., 2017), we show the impact of correcting specific types disagreements with the circularly picked gold annotations for our annotations. This analysis first converts overlapping spans to exactly matching spans (fix span), then making label disagreements match the gold (label), then adding mentions that are missing (add), and finally removing spurious mentions (remove). Figure 6 (left panel) explores which kinds of errors cause disagreement between annotators looking at three sets of labels; all types of operation labels, all types of material labels, and all other quantities and conditions. While the additional details such as conditions, amounts, and properties have high agreement, we note that for material labels, fixing slight disagreements in span boundaries has a meaningful impact. We suggest that this is due to the difficulty of consistently defining the mention spans for complex chemical mentions; an example disagreement about `target` span boundaries is shown in ex. 4 for an instance of `Target`, where "[]" denote span boundaries.

(4) F-containing [[MIL-100] (Fe)(MIL-100(Fe)_F)] was prepared from hydrothermal reaction oof trimesic acid with metallic iron, HF, nitric acid and H2O at 160 decG for 8h as reported elsewhere [10].

7 Analysis of Model Errors

Here, we can contrast inter-annotator disagreements with the performance of models trained on MS-MENTIONS. Figure 6 (right) illustrates the F1 of the baseline model evaluated upon the develop-

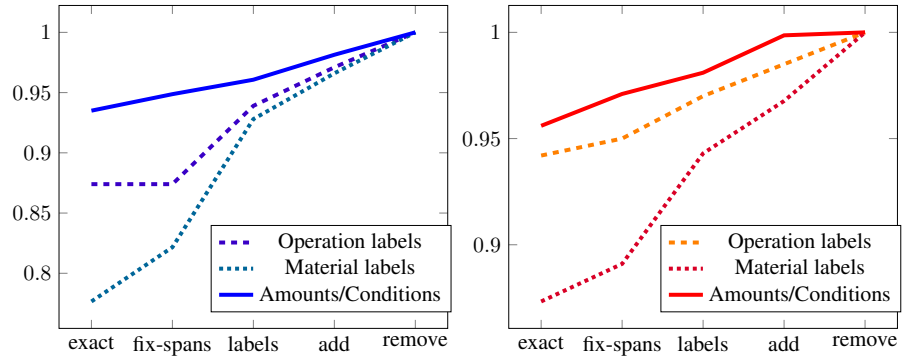

Figure 6: F1 of inter-annotator agreement (left) and performance of trained models on the dev set (right) as one progressively corrects span mismatches, incorrect labels, missing spans and spurious spans

ment set, as one automatically fixes specific errors as in §6.

One can see that the general contour of errors is similar between these model errors and the human error pattern. We suggest that the higher model performance (when compared to IAA F1) may be due to the approach of using a division of labor, so that each phenomenon in the main dataset is annotated by a single annotator. This means that a model is learning and evaluated upon a single annotator's annotation practices.

Finally to form an understanding of label disagreements, Figure 4 shows a confusion matrix for the predictions of the baseline model. We note that trained models also suffer errors in distinguishing between `Target` and `Material`, but are able to distinguish between `Material` and `Unspecified-Material`.

Qualitatively examining model errors, a common source of the span disagreements is seen in ex. 5 with "NaOH/urea", where a domain expert in context infers that this refers to a single material, but model predictions split up "NaOH" and "urea":

(5) 7 g NaOH and 12 g urea were added into the deionized water (100 mL) under vigorous stirring to form NaOH/urea solution.

A similar form of this kind of model error occurs with complex measurement units where the CONDITION-UNIT mention would refer to a ratio such as "mol/mol" or a rate such as a "degC/min.".

8 Related Work

The presented corpus fits into a larger body of work on scientific information extraction (Kim et al., 2003; Garg et al., 2016; Augenstein et al., 2017). While a majority of the work in scientific IE developed resources for biomedical text, more recent work has seen growing interest in information extraction from materials science text, as in the corpora outlined in Table 1, those of MSPT (Mysore et al., 2019), SoFC (Friedrich et al., 2020), and in organic procedural text (Kulkarni et al., 2018; Tamari et al., 2021). In addition, Yamaguchi et al. (2020) annotated a corpus of 1000 abstracts of papers about superconductive materials, Fang et al. (2021) annotated a corpus of chemical reaction snippets for coreference, Vaucher et al. (2020) weakly supervise a transformer model for chemical synthesis descriptions using rule-based methods. Beyond the material science domain, Luan et al. (2018) annotated a dataset for general scientific IE, and Mori et al. (2014) and Kiddon (2016) annotate cooking recipes with semantic structures.

This recent thrust to develop information extraction tools may be explained by the consensus in the materials science community that extracting the knowledge contained within natural language descriptions of inorganic materials syntheses will be a key step towards reducing the overall discovery and development time for novel materials (Butler et al., 2018).

9 Conclusion

The MS-MENTIONS dataset established a high-quality annotation of entity mentions for the Material Science domain. It achieves this in part through a novel approach to the division of annotation into multiple stages, reducing the inter-annotator variation of the annotation of each phenomenon. Our baselines show that entity recognition systems trained on this corpus can achieve high performance on this consistently defined task.

We hope that this can be utilized to improve downstream tasks relying upon mention detection performance.

10 Acknowledgements

We are grateful to Haihao Liu for contributions in annotation and to anonymous reviewers for feedback on the paper. This work was funded by IBM Research AI through the AI Horizons Network, the Chan Zuckerberg Initiative under the project Scientific Knowledge Base Construction, the NSF under grants IIS-1955567 and IIS-1763618. It was also funded through the NSF through DMREF Awards 1922311, 1922372, and 1922090, DARPA via Contract No. FA8750-17-C-0106 (subaward 89341790) from the University of Southern California, the Office of Naval Research (ONR) under contract N00014-20-1-2280 and 2114_2021, and ONR via Contract No. N660011924032 under Subaward No. 123875727 from the University of Southern California.

References

Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. 2017. SemEval 2017 task 10: ScienceIE - extracting keyphrases and relations from scientific publications. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 546–555, Vancouver, Canada. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6.

Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. 2018. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, pages 837–840. Lisbon.

Biaoyan Fang, Christian Druckenbrodt, Saber A Akhondi, Jiayuan He, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU-ref: A corpus for modeling anaphora resolution in the chemical domain. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1362–1375, Online. Association for Computational Linguistics.

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

Sahil Garg, Aram Galstyan, Ulf Hermjakob, and Daniel Marcu. 2016. Extracting biomolecular interactions using semantic parsing of biomedical text. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483.

Chloé Kiddon. 2016. *Learning to interpret and generate instructional recipes*. Ph.D. thesis, University of Washington.

Edward Kim, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti. 2017. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chemistry of Materials*, 29(21):9436–9444.

Edward Kim, Zach Jensen, Alexander van Grootel, Kevin Huang, Matthew Staib, Sheshera Mysore, Haw-Shiuan Chang, Emma Strubell, Andrew McCallum, Stefanie Jegelka, et al. 2020. Inorganic materials synthesis planning with literature-trained neural networks. *Journal of Chemical Information and Modeling*, 60(3):1194–1201.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Julien Knafou, Nona Naderi, Jenny Copara, Douglas Teodoro, and Patrick Ruch. 2020. Bitem at wnut 2020 shared task-1: Named entity recognition over wet lab protocols using an ensemble of contextual language models. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 305–313.

Chaitanya Kulkarni, Wei Xu, Alan Ritter, and Raghu Machiraju. 2018. An annotated corpus for machine reading of instructions in wet lab protocols. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 97–106.

Jonathan K Kummerfeld, David Hall, James R Curran, and Dan Klein. 2012. Parser showdown at the wall street corral: An empirical investigation of error types in parser output. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1048–1059.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Rubayyat Mahbub, Kevin Huang, Zach Jensen, Zachary D Hood, Jennifer LM Rupp, and Elsa A Olivetti. 2020. Text mining for processing conditions of solid-state battery electrolyte. *Electrochemistry Communications*, 121:106860.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada. 2014. Flow graph corpus from recipe texts. In *LREC*, pages 2370–2377.

Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 56–64.

Sheshera Mysore, Edward Kim, Emma Strubell, Ao Liu, Haw-Shiuan Chang, Srikrishna Kompella, Kevin Huang, Andrew McCallum, and Elsa Olivetti. 2017. Automatically extracting action graphs from materials science synthesis procedures. In *Workshop on Machine Learning for Molecules and Materials at NIPS*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

Janvijay Singh and Anshul Wadhawan. 2020. PublishInCovid19 at WNUT 2020 shared task-1: Entity recognition in wet lab protocols using structured learning ensemble and contextualised embeddings. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 273–280, Online. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online. Association for Computational Linguistics.

Jeniya Tabassum, Sydney Lee, Wei Xu, and Alan Ritter. 2020. Wnut-2020 task 1 overview: Extracting entities and relations from wet lab protocols. *arXiv preprint arXiv:2010.14576*.

Ronen Tamari, Fan Bai, Alan Ritter, and Gabriel Stanovsky. 2021. Process-level representation of scientific protocols with interactive annotation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2190–2202, Online. Association for Computational Linguistics.

Hadrien Titeux and Rachid Riad. 2021. pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in python. *Journal of Open Source Software*, 6(6).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Alain C Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H Nair, Philippe Schwaller, and Teodoro Laino. 2020. Automated extraction of chemical synthesis actions from experimental procedures. *Nature communications*, 11(1):1–11.

Chenguang Wang, Alan Akbik, Laura Chiticariu, Yunyao Li, Fei Xia, and Anbang Xu. 2017. Crowd-in-the-loop: A hybrid approach for annotating semantic roles. In *Proceedings of the 2017 Conference on*

Empirical Methods in Natural Language Processing, pages 1913–1922.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kyosuke Yamaguchi, Ryoji Asahi, and Yutaka Sasaki. 2020. Sc-comics: A superconductivity corpus for materials informatics. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6753–6760.

Wenwen Zhao, Akinobu Tanaka, Kyoko Momosaki, Shinji Yamamoto, Fabi Zhang, Qixin Guo, and Hideyuki Noguchi. 2015. Enhanced electrochemical performance of ti substituted p2-na2/3ni1/4mn3/4o2 cathode material for sodium ion batteries. *Electrochimica Acta*, 170:171–181.

A Baseline Model Details

Models on both the current dataset and Mysore et al. (2019) used the same overall parameters, for comparability. Hyperparameters were adopted based upon those used in (Friedrich et al., 2020) and participants of the WNUT shared tasks using similar approaches (Tabassum et al., 2020).

Models were trained with dropout between 0.1-0.3, and learning rates ranging from 5e-6 to 2e-5. Each were trained for 50 epochs of until no improvement over validation set performance for 10 epochs. During task-adaptive pre-training on other datasets, models were trained for 20 epochs on the prior task, the model saved, and then training restarted with the saved model (meaning that the optimizer did re-initialize).

B MS Mentions Datasheet

Following recommended practice to report finer details of the a dataset, we provide a Datasheet following prompts provided in Bender and Friedman (2018) and Gebru et al. (2018).

B.1 Motivation

For what purpose was the dataset created?: The broad goal of the annotations is to facilitate development of information extraction models. We expect the structured extractions from the scientific papers to in turn will facilitate tools for accelerated development of materials.

Who created the dataset and on behalf of which entity?: Anonymized.

Who funded the creation of the dataset?: Anonymized.

B.2 Composition

What do the instances that comprise the dataset represent?: We view the dataset primarily as consisting of materials synthesis procedures. Synthesis procedures consist of sentences. Our dataset annotates the sentences in the context of the whole synthesis procedure.

How many instances are there in total (of each type, if appropriate)? There are 595 synthesis procedures.

Does the dataset contain all possible instances or is it a sample of instances from a larger set?: The instances were selected from a collection of materials science and chemistry papers of more than 3 million papers obtained from a variety of scientific publishers through legal permissions.

Is there a label or target associated with each instance?: Each word in the synthesis section is labeled with the entity type which describes the role it plays in the context of the synthesis procedure.

Is any information missing from individual instances?: The dataset does not contain gold tokenization or gold sentence boundaries. More broadly, the dataset only releases information most relevant to the target tasks we envision. For each synthesis procedure it would have been possible to release bibliographic information and other metadata, this isn't released. Individual instances do come with a DOI which can be used to retrieve more detailed metadata. The sentences are also missing inline citation information and full text paper context. There is likely a range of other information which could have been included that isn't directly conceivable to us in writing this Datasheet.

Are relationships between individual instances made explicit (e.g., users movie ratings, social network links)?: Relations between instances could take the form of citation relations, authorship links etc, this information isn't part of the dataset.

Are there recommended data splits (e.g., training, development/validation, testing)?: Yes, splits are released along side the data. The splits are made at the level of the synthesis procedure.

Are there any errors, sources of noise, or redundancies in the dataset?: Annotation errors are likely to be present. Sources of noise might be introduced during parsing of the raw HTML and XML files.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?: The dataset is self-contained.

Does the dataset contain data that might be considered confidential?: None.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?: No.

Does the dataset relate to people?: The dataset is about people to the extent that it annotates text authored by researchers.

Does the dataset identify any subpopulations (e.g., by age, gender)?: No

Is it possible to identify individuals (i.e., one or more natural per- sons), either directly or indirectly (i.e., in combination with other data) from the dataset?: The dataset comes with DOIs which link to biblographic information which makes the authors of the authored synthesis procedures explicit.

The Mechanical Alloying (M.A.) was performed in a planetary mill (Fritsch P5). The elements (antimony Alfa Aesar 99.999% and zinc Sigma Aldrich 99.99%) were introduced in the required amounts in 80 mL tungsten carbide vials under argon. The used Ball to Powder Ratio (BPR) was fixed to 20:1 with 10 mm tungsten carbide balls. The milling speed was 400 rpm and for each milling duration the total time was divided in 30 min steps and stopped for another 30 min to prevent excessive heating of the sample. Several milling times were used: 2 h 30, 3 h, 5 h, 10 h and 20 h. For each experiment, the total weight of material was about 10 g.

Figure 7: An example synthesis procedure which wasn't included in the current dataset since it did not have a clear sequence of Operations. Colors represent entity types and underlines represent span boundaries. Colors: **Operation**, **Nonrecipe-operation** and **Meta**.

A simple web search of the synthesis procedure text will also reveal the authors.

Does the dataset contain data that might be considered sensitive in any way?: No.

Speaker demographic and Language Variety following (Bender and Friedman, 2018): Demographics of paper authors were not collected. All papers gathered are written in English but the specific language dialect of the papers wasn't selected for. The style of writing is scientific given that all text came from scientific papers.

B.3 Collection Process

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?: We used webscraping and HTML/XML parsing to gather and format the data. This was done using standard python libraries on a Linux server.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?: Synthesis procedures were selected by a mixture of keyword searches and random selection from a corpus of over 3 million papers: 13 papers were obtained by searching Engineering Village with "LGPS" and "Li10GeP2S12", 244 papers were obtained from Elsevier Scopus using "Li battery", 338 papers were picked randomly. Each paper was manually inspected to ensure consistency with our annotation schema. Examples of procedures that do not fit the schema are synthesis procedures consisting of one major operation and a description of the setting used for the synthesis. Such as example is shown in Figure 7.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?: One materials science PhD student and two materials science post-doctoral researchers were involved in the data collection and annotation. Annotator Guidelines were developed in conjunction with a computer science PhD student and a Linguistics post-doctoral researcher.

Over what timeframe was the data collected?: The selection and annotation of this data took place over the course of approximately 3 months.

Were any ethical review processes conducted (e.g., by an institutional review board)?: Since the dataset did not involve human subjects this wasn't conducted.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?: Most of the data is acquired through a website either API or direct web scraping. However, a smaller portion of the data is delivered directly to our servers by the publishers at their request.

Were the individuals in question notified about the data collection?: Individuals weren't notified about the data annotation.

Did the individuals in question consent to the collection and use of their data?: NA.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?: NA.

Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?: No.

Was any preprocessing/cleaning/labeling of the data done?: Each article is parsed into plaintext format. We run each paper through a section classifier to identify the synthesis sections of the article. Each section is then manually examined to ensure classification accuracy.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?: Yes, all preprocessed HTML or XML files for each article have been saved.

Is the software used to preprocess/clean/label the instances available?: Yes. Software will be linked upon acceptance.

B.4 Uses

Has the dataset been used for any tasks already?: No.

Is there a repository that links to any or all papers or systems that use the dataset?: NA.

What (other) tasks could the dataset be used for?: The primary task we envision this as supervised training data for is that of Named Entity Recognition. Given our careful distinctions of generic events into Operation, Nonrecipe-operation and Meta this dataset might be of value in training models for event detection in this domain. Further, given that most synthesis procedures contain a single `Target` material and our annotation of `Sample` mentions which typically reference `Target`s, this data can be used as an evaluation set (`Sample` mentions are somewhat infrequent in our dataset) for target mention co-reference.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?: This would depend on the future uses. But we expect it to serve as robust training and evaluation data for NER for the domain the data represents and if the definitions we use match those of the future use.

Are there tasks for which the dataset should not be used?: None. If future users attempt to use models trained on this data for a domain different from that of this data they should make sure to thoroughly analyze errors it makes before using models trained on this data.

B.5 Description

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?: Yes.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?: Will be released on GitHub.

When will the dataset be distributed?: On publication.

Will the dataset be distributed under a copyright or other intel- lectual property (IP) license, and/or under applicable terms ofuse (ToU)?: MIT License.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?: No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?: None that we are aware of.

B.6 Maintenance

Who is supporting/hosting/maintaining the dataset?: Anonymized.

How can the owner/curator/manager of the dataset be contacted?: Anonymized.

Is there an erratum?: There is not one at the moment. Our dataset release will be updated if non-trivial annotation or other errors are found.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?: Given the ongoing nature of this project non-trivial errors which can work within the current assumptions and annotation framework may be added. If this isnt possible and if the needs of the project change in future, newer datasets may be released.

C Annotation Guidelines

C.1 Operations Annotation

`Operation`: Actions, events, and verbs that denote distinct, individual process steps, actions that we would want to extract during data mining.

`Nonrecipe Operation` Actions, events, and verbs that do NOT denote a distinct, individual process steps, actions that we would NOT want to extract during data mining. These are often, but not always, less descriptive, general actions that apply to the recipe as a whole, not to individual steps in a process.

TYPICAL, BUT NOT UNIVERSAL, PATTERNS:

- "NONRECIPE OPERATION by OPERATION", for example, *Prepared by mixing*, *Obtained by filtration* , *Removed by filtration*. *Concentrated by centrifuging*
- "OPERATION to NONRECIPE OPERATION", for example, *Heated to remove, Filtered to obtain, Centrifuged to give*
- "NONRECIPE OPERATION to OPERATION", for example, *Allowed to cool*
- "OPERATION was NONRECIPE OPERATION", for example, *Heating was performed*, *Sintering was carried out*
- "OPERATION by OPERATION" and "OPERATION to OPERATION" are possible if both words are actions we want to extract during data mining, for example, *Mixed by ultrasonication Mixed by stir-*

ring Collected by filtration Stirred to mix Mixed to disperse

C.2 Material Annotation

`Materials` A material entity is a physically used, chemically defined object used in the synthesis but not the end result of the synthesis. Only the words required to define the base material chemical composition should be labelled. All other words describing chemical, structural, or morphological modification to the base material will not be labelled as a material.

- Chemical Formulas — *H20, C2H5OH, La(NO3)3*6H2O, NaNO3, Al(NO3)3[?]9H2O*
- Chemical Names — *Ethanol, water, pluronic F127, titanium (IV) isopropoxide, 1-cysteine*
- Abbreviations — *GO, CTAB, MTMS*

`Targets` A target is a physically present, chemically defined material that is made within the context of the paragraph in which the target is found. Only the words required to define the base material chemical composition should be labelled. All other words describing chemical, structural, or morphological modification to the base material will not be labelled as a target. Abbreviations denoting authors' differentiation between samples should not be labeled as targets. Writing defining the variable parts of the composition, either numerical or elemental, should not be included as part of the target although numbers and units should be labeled accordingly. An undefined material should never be labeled as a target even if filling that role linguistically.

- Chemical Formulas — *ZnAl2O4, TiO2, La3NiO7, LiMn0.98Zn0.02PO4, Li1+xV3O8, B-Ni(OH)2*
- Chemical Names — *Manganese dioxide, carbon, titania, manganese-cobalt oxysulfide*
- Abbreviations — *CNTs, TKF*
- Composites — *TiO2/Cu2O, Li1.95FeSiO4/C, Li2FeSiO4/MWCNTs, Copper sulfide—reduced graphene oxide, CuS—rGO, MoS2/polyaniline, P2-Na2/3Ni1/4Mn3/4O2*

`Unspecified Materials` An unspecified material is a physically present, chemically undefined material. It can play the role of target linguistically. Typically, they are found as the intermediate materials in the synthesis. Pronouns can be intermediate materials. Phase information, ie solution, mixture, product etc, are only labeled as unspecified material if there is no other, more chemically descriptive word describing the object. Words that describe material families are also labeled as unspecified materials. If a chemically undefined material is not physically present it is labeled as non-recipe.

- Material Families — *Metal oxides, metallic nitrates, metal salts*
- Words — *Solution, mixture, product, precursor, dispersion, reagents, sample, powders*
- Pronouns — *It, they*

`Non-recipe Materials` Non-recipe materials are both chemically defined or undefined materials that are not physically present in the synthesis but appear linguistically. These often describe ratios of previously mixed precursor elements or chemical species that are removed at various points of the synthesis. Elements defined as part of the noun phrases of targets are not considered non-recipe.

`Samples` Samples are abbreviations used for multiple target materials denoted by the authors for the purpose of distinguishing between different synthesized materials. Not all abbreviated targets are samples.

C.3 Quantitative Mentions

`Number` Numbers are either digits or the number words. All numbers should be labeled unless they are part of a chemical composition or if nothing else is labeled in the sentence. Hyphens and ratio characters (: and /) should not be labeled as numbers. Variables should never be labeled as numbers. Numbers describing a multi-dimensional amount/property should be labeled as all the numbers plus the connections (2x2).

`Units` Any unit of measurement for material amounts, operation conditions, materials properties, and apparatus properties.

`AMOUNT` units describe absolute amounts, con-

centrations, purities, ratios, and flow rates

- Mg, mL, M, % , mol %
- Ratio, weight ratio, mg/min, mL/min, sccm

`CONDITION units` describe intangible conditions under which operations are performed

- °C, K, Sec, RPM, mW, MPa, pH, times (as for repeated operations)

`PROPERTY units` describe measured materials properties

- mm, %, MPa, nF

`Apparatus units` describe values associated with apparatuses

- mm, mL

C.4 Apparatus

Tangible equipment used to perform an operation (synthesis apparatus) or to characterize a material's properties (characterization apparatus)

C.5 General Notes

Brackets and parentheses are only included when necessary to capture the entire chemical information Materials joined by joining token (/ or : or +) are labeled individually (joining token excluded) if the materials are not a single entity. Examples of single entities are composite materials or solutions that have multiple components specified.