Multi-Source (Pre-)Training for Cross-Domain Measurement, Unit and Context Extraction

Yueling Li[1], **Sebastian Martschat**[2], and **Simone Paolo Ponzetto**[3]

[1] Knowledge Architecture, BASF Digital Solutions GmbH, 67061 Ludwigshafen, Germany
[2] Knowledge Architecture, BASF SE, 67056 Ludwigshafen, Germany
[3] Data and Web Science Group, University of Mannheim, 68131 Mannheim, Germany
{yueling.li|sebastian.martschat}@basf.com
simone@informatik.uni-mannheim.de

Abstract

We present a cross-domain approach for automated measurement and context extraction based on pre-trained language models. We construct a multi-source, multi-domain corpus and train an end-to-end extraction pipeline. We then apply multi-source task-adaptive pre-training and fine-tuning to benchmark the cross-domain generalization capability of our model. Further, we conceptualize and apply a task-specific error analysis and derive insights for future work. Our results suggest that multi-source training leads to the best overall results, while single-source training yields the best results for the respective individual domain. While our setup is successful at extracting quantity values and units, more research is needed to improve the extraction of contextual entities. We make the cross-domain corpus used in this work available online[1].

1 Introduction

Numeric components such as counts, measurements and are crucial information for researchers across various disciplines. An automatic system for the extraction of numerical measurements (e.g. 10 %) and related *"contextual"* information such as measurands (e.g., concentration) and measured entities (e.g., chemical solution) for scientific literature can aid in the efficient construction of knowledge bases (Court and Cole, 2018) for downstream research tasks.

Ideally, the system should be able to handle multiple subject domains or even unseen domains, as relying on multiple specialized systems is inefficient and sometimes infeasible: For instance, each specialized model requires dedicated training and deployment resources. Further, the target-domain cannot always be known at inference time.

[1] https://github.com/liy140/multidomain-measextract-corpus

Related Work. Most existing work tackle domain-specific measurement extraction problems. In the bio-medical field alone, there exist several publications that specialize in one particular subfield and text genre, for instance food laboratory tests (Kang and Kayaalp, 2013), narrative radiology reports (Sevenster et al., 2015) or diabetes test trials (Hao et al., 2016). Many topic-specific systems are not designed as sole measurement extractors, additionally offering comprehensive knowledge base construction or text processing functionalities (Swain and Cole, 2016; Dieb et al., 2015; Epp et al., 2021; Lentschat et al., 2020; Friedrich et al., 2020). Due to their specialization, it is difficult to apply them to novel domains without considerable adaption effort.

A few domain-independent systems have been developed, but they either offer limited context extraction capabilities (Soumia Lilia Berrahou et al., 2013; Mündler, 2021) or lack concrete definitions of the extracted contextual entity types (Foppiano et al., 2019; Hundman and Mattmann, 2017). Moreover, for these systems, no study of cross-domain generalization capabilities has been performed.

Harper et al. (2021)'s SemEval Task represents a key milestone for the progress of domain-independent measurement extraction research. 25 teams participated in the competition, thereby introducing the task, which has before mostly been solved through traditional NLP methods, to a wide range of recent architectures, among others also the now dominantly used pre-trained language model architecture (Devlin et al., 2019). Moreover, Harper et al. (2021) define the task in a domain-agnostic manner and provide an annotated multi-domain measurement extraction corpus with data sourced from the OA-STM Corpus (Elsevier Labs, 2015). However, due to its small data size (295 paragraphs), the corpus is not sufficient on its own for studying cross-generalization effects.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 1–25

Contributions. To address the research gaps mentioned above, i.e. the need for cross-domain generalization as an evaluation criterion for the measurement and context extraction task, we aim to build a *cross-domain* measurement, unit *and* context extraction system. We make the following contributions:

- To facilitate multi-domain training, we expand the corpus published by Harper et al. (2021), creating a multi-domain, multi-source corpus for measurement, unit, and context extraction including two additional source domains.
- We construct an end-to-end model pipeline based on pre-trained language models (Devlin et al., 2019) and achieve state-of-the-art performance comparable to the first placed MeasEval team (Davletov et al., 2021).
- We study the effect of (a) adaptive intermediate *pre-training* (Gururangan et al., 2020) and (b) multi-source *fine-tuning* (Zhao et al., 2020) on cross-domain generalization. For (a) we apply full intermediate pre-training and adapter-based pre-training (Hung et al., 2021; Houlsby et al., 2019) using a curated multi-domain task-adaptive pre-training corpus (Gururangan et al., 2020). For (b), we experiment with different pooled combinations of *fine-tuning* domains.
- Finally, we carry out a task-specialized error analysis using entity-level analysis methods adapted from Fu et al. (2020) to determine concrete error sources for well-grounded model improvement.

In the following sections, we explicate our corpus construction approach (§2), the model architecture (§3), and domain adaption methods (§4). Finally, we present our experimental results (§5) coupled with the error analysis (§6) and a concluding discussion (§7).

2 A New Multi-Domain Corpus for Measurement Extraction

In this section we describe the creation of a multi-domain corpus for measurement and context extraction. This will enable the investigation of cross-domain prediction performance.

2.1 Data Model and Source Corpora

The first step in corpus creation for measurement extraction is to decide on a data model that relates objects to be measured, values, and their context. We adapt the data model and terminology as proposed by Harper et al. (2021), excluding the "Qualifier" and "Modifier" classes to increase the candidate pool for corpus expansion.

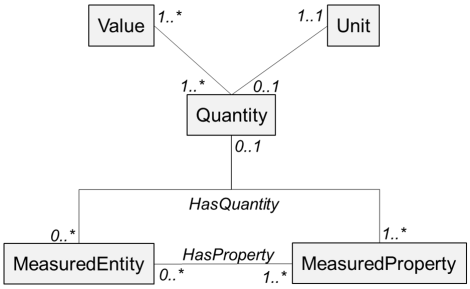

Figure 1: Data model based on the MeasEval task definition (Harper et al., 2021). Multiplicities between entities show the upper and lower bounds of entities for each relationship, i.e. 0..1 = zero or at most one, 0..* = zero or more, 1..1 = exactly one, 1..* = one or more.

Figure 1 presents the resulting adapted data model for our multi-source corpus: A **Quantity** (Q) is made up of one or more numeric Values (V) and optionally a **Unit** (U). A **MeasuredEntity** (ME) is the object, event or phenomenon, whose quantifiable property is measured. This would be the **MeasuredProperty** (MP), i.e., the measurand that can be attributed to the measured object. Figure 2 shows the annotation of an example sentence according to the presented data model. The extended data model definition can be found in Appendix A.

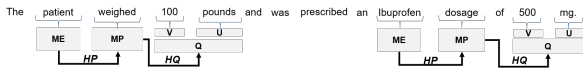

Figure 2: Annotated example sentence. Additional abbreviations: HP - HasProperty, HQ - HasQuantity

For selecting the source corpora, we compiled a candidate pool consisting of three datasets from related work and one additional patent dataset created by the chemical company BASF SE. We then evaluated candidate datasets with respect to compatibility with the data model. Table 1 presents the evaluation summary. As such the resultant corpus is comprised of the MeasEval corpus (Harper et al., 2021), the Battery Material Patents (BM) dataset as well as the Material Science Procedural (MSP) corpus (Mysore et al., 2019).

2.2 Data Processing and Annotation

We apply several processing steps to normalize each source corpus with respect to the measurement extraction data model of Figure 1.

Corpus	Q	U	ME	MP	R
MeasEval Corpus (Harper et al., 2021)	x	x	x	x	x
Battery Materials Patents Corpus (originally BASF internal dataset)	(x)	x	x	x	x
Material Science Procedural Corpus (Mysore et al., 2019)	x	x	(x)	(x)	(x)
ChemDataExtractor Evaluation Corpus (Swain and Cole, 2016)	(x)	x	x	x	x
SOFC-Exp Corpus (Friedrich et al., 2020)	x	x	(x)	(x)	/

Table 1: Candidate datasets evaluated by data model components. Selected corpora are bolded. A full fit to the evaluation criterion is denoted with "x", a partial fit is indicated with "(x)" and a unrepresented concept is marked as "/". Q = Quantity Value, U = Unit, ME = MeasuredEntity, MP = MeasuredProperties, R = Relations.

MeasEval Corpus. For the MeasEval corpus, in order to accommodate the limited input length of most pre-trained models, we split the paragraphs into sentences using spaCy[2]. We deal with particularities of scientific language, e.g., bibliographic references and abbreviations by applying custom segmentation rules.

BM Corpus. The BM dataset describes and classifies information regarding entities and properties of battery materials from patent claims. The original annotations specify the patent type of a claim (e.g., material claim vs. process claim) as well as phrase-level entity and relation information across 15 entity types and 13 relation types (e.g., stirrer elements, complexants, main metals). The entity types *Value*, *Unit* and *Property* can be directly mapped to entities defined in our data model, i.e. Quantity value (Q), Unit (U) and MeasuredProperty (MP) respectively. By contrast, there are multiple source entity types that can be mapped to the MeasuredEntity (ME) class. These are parsed through graph traversal: we follow the relations that are connected to *Value* entities, we thereby find their respective U, MP and ME. We save each claim separately and do not apply additional segmentation measures to preserve the unique structure of the patent style.

MSP Corpus. The MSP Corpus comprises 230 articles describing material synthesis procedures (MIT Open Source License, Mysore et al. 2019). Although the annotation scheme is comparable to the Battery Materials dataset, including entity types such as *Material*, *Operation*, *Amount-Unit*, *Synthesis-Apparatus*, *Number* etc., a more complex mapping would be required to cover the various semantic structures present in this dataset. For this reason, we opted to manually re-annotate the data instead of relying on entirely automatic re-labeling. To reduce the annotation effort, automatic labels using the original annotation data were created. As with the BM corpus, we follow all entities related to the entity *Number* and map them to our defined labels (see B).

The annotation process involved four non-native annotators from different scientific backgrounds and mixed genders.

We drafted a separate annotation guideline that (i) explained the task according to the MeasEval annotation guidelines and (ii) introduced dataset-specific instructions (Appendix C). We re-annotate all samples of the validation articles (89 sentences) and test articles (129 sentences), and a subset of the training articles (860 sentences) to limit the annotation effort. We use the evaluation split for NER provided by Mysore et al. (2019).

An inter-annotator-agreement (IAA) study validated the reproducibility of our guidelines, producing substantial agreement scores. This is described further in Appendix D.

Final Corpus. The final multi-domain, multi-source corpus consists of the normalized version of the three corpora described above, totaling 2,389 sentences from scientific articles (MeasEval), 1,077 sentences from material synthesis procedures and 352 battery material patent claims (not segmented). An overview of the final corpus is given in Appendix E.

3 Extraction Architecture

We now describe our model setup which is designed to extract the entity and relation types described in the previous section.

We model the extraction as a two-step pipeline made up of two token-classification models, which we coin as **Task 1** and **Task 2**. We first extract all Quantities (Task 1) and then simultaneously predict U, ME and MP (Task 2) based on each extracted Quantity. This cascading setup resolves the relation extraction problem of assigning the

[2]https://spacy.io

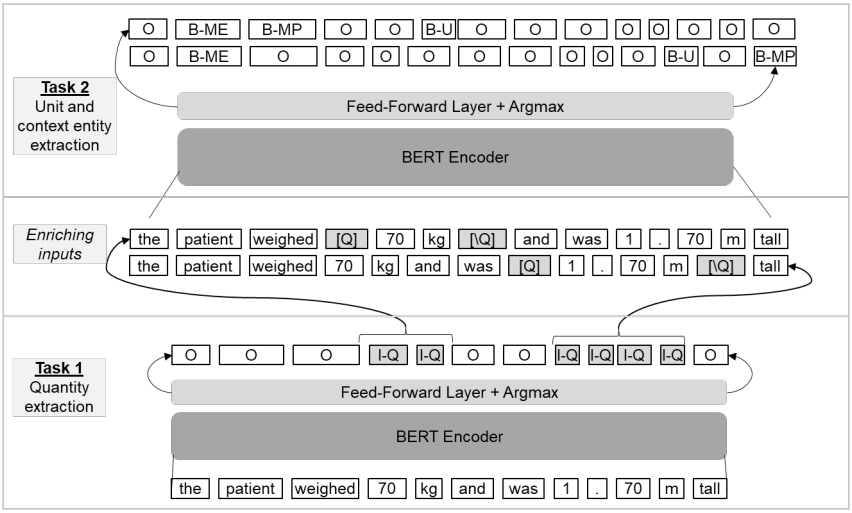

Figure 3: Extraction flow, Q = Quantity, U = Unit, ME = MeasuredEntity, MP = MeasuredProperty

context entities to the correct quantity span, as the data model allows for a deterministic, rule-based assignment of the relations between U, ME and MP (see Gangwar et al. (2021); Davletov et al. (2021)).

Figure 3 shows the extraction flow based on an example sentence: The information from the first task is input into the second task through special tokens [Q] and [/Q] which we wrap around the identified Q spans (see also Gangwar et al. (2021); Davletov et al. (2021)). For each identified Q, an enriched prediction sample is created, thereby allowing for overlapping entities and conditioning the unit and context entity extractor on one Q at a time. For Q extraction we use binary IO-tags (Liu et al., 2021). For Task 2 we use the BIO-tagging scheme. To accommodate the tokens [Q] and [/Q] which signal the identified Q spans from Task 1, we add them to the models' vocabulary as special tokens, extending the embedding size by two. For training, we use cross-entropy loss over all classes and train Task 1 and Task 2 separately.

A drawback of this simple architecture is the fact that it cannot enforce the 1:1 relationships prescribed by the data model, since it is possible to predict more than one ME or MP. Further, we set the input sequence to the size of a single sentence to account for the one-sentence annotation window of the MSP and Battery Materials dataset.

4 Domain Adaption and Generalization

We experiment with a) adaptive pre-training and b) multi-source fine-tuning. Figure 4 summarizes the applied methods and resulting model configurations. With the exception of the training setting with all sources, all shown configurations are applied to the models of both tasks.

Adaptive Pre-Training. This setup comprises a combination of pre-trained base models and intermediate pre-training: We use $\text{BERT}_{\text{BASE}}$[3] (Devlin et al., 2019) as the baseline model representing the canonical text domain, and SciBERT [4] (Beltagy et al., 2019), which we expect to be more closely related to the domains of our measurement extraction corpus, because it was pre-trained from scratch on scientific articles.

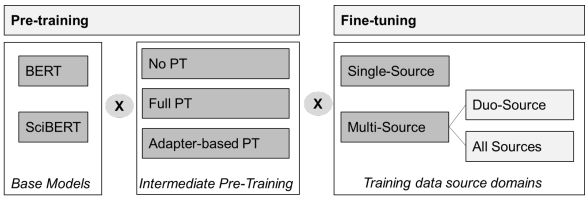

Figure 4: Model configurations for domain adaptation and domain generalization by training phase.

We create intermediately pre-trained variants for each of the two BERT-models using task-adaptive pre-training (TAPT) (Gururangan et al., 2020): we continue pre-training of the models on the unlabeled (training) data of our measurement extraction corpus. Thereby we aim to bring the models closer to the target domains of the task and induce increased task performance compared to the base models. We also apply adapter-based (Houlsby et al., 2019; Pfeiffer et al., 2020) intermediate pre-training to compare full TAPT against to a more

[3] https://huggingface.co/bert-base-uncased
[4] https://huggingface.co/allenai/scibert_scivocab_uncased

parameter efficient approach (Kim et al., 2021).

As pre-training data, we create a "curated" (Gururangan et al., 2020) multi-source corpus comprised of the pooled data from all three training datasets. We enhance it with unlabled data from the same datasets to further increase the corpus size. As such we add all articles of the OA-STM dataset (Elsevier Labs, 2015), the source on which the MeasEval annotations are based. We remove paragraphs which appear in the test and validation splits of the MeasEval data through fuzzy string matching [5]. Further, we add 1,128 Battery Materials claims which were excluded from the measurement extraction corpus due to lack of Quantity spans, and include the rest of the MSP data that was not re-annotated. This resulted in a pre-training corpus of approximately 630k words.

Multi-Source Fine-Tuning. To investigate the impact of multi-domain training, we also employ three experimental setups that are applied in the *fine-tuning* stage of model training. In each setup a model is trained on one, two, or all domains and evaluated on all applicable target domains. The first set-up **single-source** uses only a single data set. To build our multi-source corpora we pool multiple data sources from related tasks (Aue and Gamon, 2005; Zhao et al., 2020): As such, the second set-up **duo-source** uses the concatenation of two source datasets, e.g., BM + MeasEval, and the third set-up **all sources** uses all corpora. Due to considerable discrepancies between the Quantity annotation logic of the BM dataset and the other two datasets, no all sources setup was applied to Task 1.

Training Mode	Source Domain	Model	PT Setup	Task 1 Q	Task 2 U	Task 2 ME	Task 2 MP
Single-source	Meas Eval	BERT	No PT	0.671	0.96	0.448	0.473
			Full PT	0.702	0.963	0.424	0.446
			Adpt. PT	0.688	0.932	0.388	0.403
		SciBERT	No PT	0.721	0.972	0.501	0.522
			Full PT	0.719	0.961	0.491	0.508
			Adpt. PT	0.715	0.952	0.431	0.425
	MSP	BERT	No PT	0.632	0.968	0.452	0.445
			Full PT	0.634	0.956	0.437	0.461
			Adpt. PT	0.607	0.946	0.392	0.421
		SciBERT	No PT	0.667	0.964	0.456	0.502
			Full PT	0.667	0.965	0.45	0.5
			Adpt. PT	0.665	0.952	0.393	0.491
	BM	BERT	No PT	0.29	0.865	0.125	0.216
			Full PT	0.285	0.828	0.148	0.297
			Adpt. PT	0.315	0.81	0.119	0.238
		SciBERT	No PT	0.386	0.75	0.235	0.355
			Full PT	0.357	0.755	0.21	0.329
			Adpt. PT	0.222	0.559	0.143	0.263
Duo-source	MSP+ Meas Eval	BERT	No PT	0.71	0.968	0.477	0.538
			Full PT	0.726	**0.974**	0.508	0.496
			Adpt. PT	**0.732**	/	/	/
		SciBERT	No PT	**0.739**	0.969	**0.534**	**0.589**
			Full PT	0.721	**0.975**	**0.523**	0.557
			Adpt. PT	**0.727**	/	/	/
	BM + Meas Eval	BERT	No PT	/	0.968	0.401	0.441
			Full PT	/	0.967	0.439	0.474
		SciBERT	No PT	/	0.97	0.472	0.54
			Full PT	/	0.965	0.482	0.537
	BM + MSP	BERT	No PT	/	0.972	0.455	0.489
			Full PT	/	0.967	0.417	0.496
		SciBERT	No PT	/	0.967	0.435	0.506
			Full PT	/	0.963	0.448	0.502
All sources	MSP+ Meas Eval+ BM	BERT	No PT	/	**0.975**	0.479	0.506
			Full PT	/	0.969	0.46	0.536
			Adpt. PT	/	0.969	0.411	0.484
		SciBERT	No PT	/	0.971	0.512	**0.569**
			Full PT	/	0.972	**0.524**	**0.593**
			Adpt. PT	/	0.956	0.454	0.535

Table 2: Summary of the experiment results by task and extraction class. F1 scores are calculated based on the entire corpus. All setups were evaluated on the pooled test data from all domains.

5 Experiments

We perform various experiments investigating the generalization capabilities of our system depending on data selection and domain adaption techniques. The implementation details can be found in Appendix G.

5.1 Evaluation and Scoring.

For comparing the predicted outputs to gold spans, we use the competition evaluation script provided by the MeasEval authors (Harper et al., 2021), which is designed to jointly evaluate all sub-tasks by matching predicted Quantities to gold Quantities whilst taking into account the relationships between their respective contextual entities. To benchmark against the MeasEval competition results we report the competition metric *Overlap F1* (see Harper et al. 2021). For all other results, we report the traditional token-based strict F1, which is frequently used to evaluate NER and sequence-tagging tasks (cf. Fried et al. 2019; Swain and Cole 2016). To this end, we adapt the MeasEval evaluation script by including nervaluate's [6] strict F1 implementation.

5.2 Results.

Table 2 shows a summary of the experiment scores evaluated on the entire multi-source corpus for Task 1 and Task 2. We observe that the models are rather

[5] https://github.com/seatgeek/thefuzz

[6] https://github.com/MantisAI/nervaluate

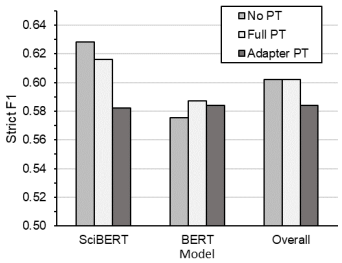

Figure 5: Task 1 – Avg. F1 score by model and pre-training setup.

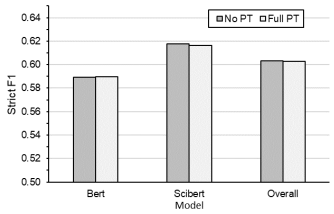

Figure 6: Task 2 – Avg. F1 score by model and pre-training setup.

accurate for the Q and U classes, while extraction performance for the contextual entities MP and ME is much lower. Below we study the results in more detail with regards to the influence of cross-domain fine-tuning and adaptive pre-training measures, and perform a dedicated analysis for the end-to-end performance of two pipeline compositions. The full result tables can be found in Appendix H.

Adaptive Pre-training Experiments. To study the effect of adaptive pre-training, we analyse the average F1 score by model type and task-adaptive pre-training setup (No PT vs. Full PT vs. Adapter PT):

SciBERT vs. BERT: For both Task 1 and Task 2, we observe that both pre-trained and base SciBERT models achieve higher scores than their BERT counterparts (Figure 5, Figure 6).

Full PT vs. No PT: When comparing different intermediate pre-training setups, we see no systematic patterns comparing Full PT and No PT performance. This is especially true for Task 2, where the average performance is quasi identical. Thus, we hypothesize that TAPT with limited pre-training data such as in our case is not sufficient for the more complex token classification task.

Adapter PT: From Table 2 we see that adapter PT is often comparable or worse to Full PT and No PT. The performance drop is especially pronounced for the more challenging ME and MP class, while it remains comparable for the Q and U class.

Multi-Source Fine-Tuning Experiments. For the investigation of multi-source fine-tuning, we analyse the average F1 score by source domain and task (Table 4, Table 5):

Cross-domain vs. in-domain: We observe that cross-domain fine-tuning is beneficial for overall generalization, while in-domain fine-tuning results in better in-domain performance. An exception to that is the rather specialized, low-resource BM domain. Its best Task 2 performance is achieved in the MSP+MeasEval →BM cross-domain setup. Similarly, leaving BM out from multi-source training setups results in higher overall scores, suggesting that overly specialized training signals negatively impact generalization.

End-To-End Evaluation. Table 3 shows the resulting E2E performance for selected model configurations, which allows us to assess the error propagation of the the cascading task flow.

We apply two separate Task 1 models for the multi-source setup, as we have not trained a model using all three datasets for Task 1 due to diverging Quantity annotation styles between BM and the other two data sources. We observe that the best overall end-to-end performance is achieved in the multi-source scenario, caused by superior performance on the MSP and BM domains. For MeasEval, we notice that the Unit extraction scores remain relatively high given the 0.2 drop in Quantity extraction. Remarkably, we see that Unit extraction works better in the MeasEval→MSP setup than in the in-domain setup.

5.3 Comparison with MeasEval Leaderboard

In Table 6 we compare a single-source and a multi-source end-to-end setup against the highest-ranking team of the MeasEval competition. All models were selected based on the best strict F1 MeasEval target domain performance (as opposed to the overall performance) of the development data.

Our single-source setup performs on par with the winning team from Davletov et al. (2021), showing superior scores for the Q and U classes, comparable scores for the ME class, and inferior scores for the MP class and the relation classes. As such we achieve competitive results with an arguably simpler model setup: Davletov et al. (2021)'s quantity extraction model is based on an ensemble of multiple LUKE models and entity-aware self-attention (Yamada et al., 2020). Further, they use XLM-RoBERTa-large for unit and context span extrac-

Source domains	Model Configuration by Task (all SciBERT)	MeasEval					MSP					BM					Overall				
		All	Q	U	ME	MP	All	Q	U	ME	MP	All	Q	U	ME	MP	All	Q	U	ME	MP
MeasEval only	T1: SciBERT Full PT; T2: SciBERT Full PT	**0.657**	**0.792**	**0.909**	**0.461**	0.442	0.734	0.915	**0.971**	0.472	0.550	0.353	0.379	0.474	0.251	0.384	0.602	0.722	0.844	0.408	0.449
MSP only	T1: SciBERT No PT; T2: SciBERT No PT	0.569	0.662	0.870	0.337	0.370	0.766	0.915	0.939	0.538	0.662	0.411	0.460	0.468	0.346	0.388	0.580	0.675	0.815	0.386	0.443
BM only	T1: SciBERT Full PT; T2: SciBERT Full PT	0.310	0.328	0.602	0.128	0.198	0.249	0.304	0.391	0.136	0.145	0.443	**0.505**	**0.524**	0.263	**0.508**	0.328	0.361	0.538	0.159	0.283
Multi	T1: BM SciBERT Full PT & MSP+MeasEval Full PT; T2: All Sources SciBERT No PT	0.647	0.782	0.893	0.445	**0.456**	**0.776**	**0.930**	0.957	**0.532**	**0.688**	**0.450**	**0.505**	0.519	**0.354**	0.440	**0.641**	**0.767**	**0.848**	**0.450**	**0.505**

Table 3: End-to-end results using (strict) F1 measure.

T1 classes: Q	Target domain			
Source domain	MeasEval	MSP	BM	O
MeasEval	**0.773**	0.847	0.386	0.703
MSP	0.632	0.916	0.408	0.645
BM	0.278	0.215	**0.467**	0.309
MSP+ MeasEval	0.765	**0.919**	0.424	**0.726**

Table 4: Task 1 – Avg. F1 score by source domain. Grey cells indicate cross-domain prediction setups.

T2 classes: U, MP, ME	Target domain			
Source domain	MeasEval	MSP	BM	Overall
MeasEval	0.615	0.655	0.656	0.631
MSP	0.551	**0.748**	0.661	0.619
BM	0.362	0.365	0.623	0.400
BM+MeasEval	0.612	0.642	0.656	0.626
BM+MSP	0.556	0.743	0.665	0.621
MSP+MeasEval	0.621	0.744	**0.700**	**0.664**
MSP+MeasEval+BM	**0.627**	0.741	0.657	0.661

Table 5: Task 2 – Average F1 score aggregated by source domain. Grey cells indicate cross-domain prediction setups.

tion and apply multi-task learning with parallel task-specific layers for each entity type. Moreover, we work with a smaller input context of one single sentence, while the winning team applies a data augmentation technique, increasing the available context (Davletov et al., 2021). However, we point out that their model learns far more entity types at the same time (seven in total), as we only work with a subset of the MeasEval task definition.

6 Error Analysis

To better understand the challenges of the task and deficiencies of our system, we perform an in-depth error analysis. We analyze error sources on a fine-grained entity level. To prevent the leakage of test data knowledge, we apply all error analysis methods on the development portion

Model	Q	U	ME	MP	HQ	HP	O
1st place MeasEval Davletov et al. (2021)	0.861	0.722	**0.437**	**0.467**	**0.482**	**0.318**	**0.551**
Single-source setup (T1: MeasEval+SciBERT+Full PT; T2: MeasEval+SciBERT+No PT)	**0.877**	**0.885**	0.432	0.437	0.465	0.307	0.550
Multi-source setup (T1: MSP+MeasEval+SciBERT+Full PT; T2: All sources+SciBERT+No PT)	0.876	0.864	0.404	0.440	0.46	0.27	0.533

Table 6: Benchmarking against MeasEval leaderboard's top team, scores correspond to MeasEval's competition scoring overlap F1.

of the corpus using our best development model setup (Task 1: BM+SciBERT+Full PT & Duo-Source(MSP+MeasEval)+Full PT; Task 2: All Sources+SciBERT+No PT). Due to the relatively high scores for the Unit class, we focus the analysis on the Q, ME and MP classes.

Entity Data Attributes. To detect model weaknesses related to the properties of entity spans, we draw on the notion of *data attributes* as defined by Fu et al. (2020): These are *"[...] values which characterize the properties of an entity that may be correlated with the NER performance."* (p. 6059). These values can be related to characteristics of the entity's surface string (e.g., entity length) or its surrounding context (e.g., sentence length). We analyse the following attributes:[7]:

- Entity length ($eLen$): The number of tokens in an entity.
- Sentence entity density ($eDen$): The number of entities in a sentence divided by the sentence length. Thus, paragraphs with multiple measurements and associated contextual entities will have a higher entity density than paragraphs with a single measurement.
- Gold quantity distance ($qDist$): The character-level span distance of the gold entity to its associated gold Q. $qDist$ only applies to

[7]We use Huggingface's BertTokenizerFast based on SciBERT vocabulary for tokenization based attributes https://huggingface.co/docs/transformers/model_doc/bert

		eLen			*eDen*			*qDist*			Match Type Count		
Class	Match Type	MeasEval	MSP	BM	MeasEval	MSP	BM	MeasEval	MSP	BM	MeasEval	MSP	BM
Quantity	match	3.9	3.0	8.9	0.130	0.202	0.073				223	179	86
	partial	5.5	8.9	8.6	0.153	0.220	0.065				57	14	34
	spurious	1.6	2.8	3.7	0.184	0.215	0.108				5	6	3
	missing	1.8	2.5	2.2	0.030	0.147	0.073				24	6	19
Measured Entity	match	2.8	3.7	2.7	0.126	0.204	0.070	18	22	73	85	119	44
	partial	5.4	15.1	2.8	0.125	0.198	0.093	18	55	39	45	59	40
	missing	2.9	14.0	2.6	0.169	0.184	0.056	45	57	127	149	29	38
	spurious	2.2	1.6	1.9	0.132	0.173	0.057	167	46	93	119	38	44
Measured Property	match	1.9	1.5	3.3	0.142	0.215	0.077	10	22	17	71	102	90
	partial	3.5	3.6	5.0	0.139	0.195	0.051	22	28	30	38	17	25
	missing	1.9	2.2	16.9	0.177	0.220	0.069	40	23	37	70	29	8
	spurious	1.7	1.4	4.2	0.124	0.187	0.076	70	28	10	112	37	21

Table 7: Count and average entity attributes $sLen$, $eDen$, $qDist$ by domain, class and match type. Bar dimensions are scaled to each domain.

the classes ME and MP. We have filtered out all cross-sentence entities for this attribute.

Analysis. In our approach, we first calculate the described data attributes. For (partial) matches and missing predictions we base the calculation on the gold entity span, for spurious predictions we base the it on the predicted span. Then, we average the attributes by match type, i.e., match, partial (match), missing and spurious, to allow the comparison of attribute averages between matches and errors. The last three columns of Table 7 show the distribution of match types by domain and entity class. We see that the number of matches is especially high for the Q class, while the number of errors is especially high in the ME class of the MeasEval and BM domains. The remaining columns show the grouped attribute averages by domain, entity class and match type. The bar charts indicate the relative magnitude of an attribute mean within one domain. We make the following observations:

- **eLen**: Partial matches occur particularly for longer entities. For instance, we observe that tokens such as brackets or dashes within a longer entity are often broken up and predicted with different labels (e.g., "deionized (DI) water" is broken into "deionized", "DI) water"). Further, spurious predictions are always relatively short, often shorter than the average $eLen$ of matches. For MSP the missing MEs have a high $eLen$, suggesting that the model has difficulties extracting longer phrases. The same phenomenon holds for the MPs of the BM domain.
- **eDen**: For both MeasEval and MSP spurious Qs are predicted for sentences with a lower entity density. Further, we observe that both missing MPs and MEs of the MeasEval data and missing Qs of the BM data appear in sentences with higher entity density. This implies that the model may 'overlook' entities when many potential entities are in one area.
- **qDist**: The model mainly struggles with long range dependencies for the ME class: $qDist$ shows the distance of the supposed gold ME to its root Q by match type. We see that our setup is good at predicting MEs that are close to their root Q, as $qDist$ is rather small for matches. However, for higher $qDist$ MEs the match type is often missing or spurious. This means that the model does not predict an ME at all (missing), or predicts a spurious one, probably closer to the root Quantity. This issue does not apply to MPs, as their $qDist$ is much lower on average.

7 Conclusion

We have applied pre-trained language models to end-to-end measurement, unit and context extraction. While our setup exhibits good extraction performance for Quantities and Units, more research has to be done to improve the extraction of contextual entities. We have identified long-range dependencies of MEs as a particular error source.

In terms of cross-domain generalization and multi-source training, multi-source training produced the best overall results, while single-source training often yielded the best results for the respective target domain. An exception to this was the small BM dataset, for which we observed the best unit and context extraction performance in the cross-domain prediction setting. This is an indicator for domain generalization, especially for low-resource domains. However, this needs to be confirmed in additional experiments with a dataset comprising even more domains.

When comparing adaptive pre-training methods, the most consistent performance driver was the use of the SciBERT base model instead of the BERT base model. Further, we found adapter-based intermediate pre-training to be worse in most cases for both model types and tasks, which may be due to the task complexity. This theory is affirmed by the fact that we saw better adapter-based pre-training results for the simpler Quantity extraction.

Finally, we found non-conclusive results for the comparison of no pre-training versus full pre-training. The instability of results may be due to the limited size of our pre-training data, or the effect of catastrophic forgetting. Future work with a larger pre-training corpus may give clearer insights into this case.

Limitations and Ethical Considerations

We would like to discuss the following limitations and ethical considerations:

In this paper, we investigated the cross-domain extraction performance based on a multi-source corpus. Our working assumption is that this corpus represents enough variety to support such a claim. However, we point out that the corpus is biased towards English scientific and patent language, as well as the chemical / material science subject domain. Further, we remark that the subjects distribution itself is biased towards the BM and MSP datasets as the the more varied MeasEval dataset only contains few examples for each of its 10 subjects. Consequently, a balanced corpus should have a more even distribution of both subject domains and language domains by increasing the size of the currently underrepresented domains and ideally including data from more than only the English language.

Further, despite having substantial IAA scores for the re-annotation of the MSP corpus, we often perceived the task as difficult and ambiguous and felt the limitations of only having two contextual entities, instead of the three as proposed by Harper et al. (2021). Yet, the low IAA score (0.334) for the excluded Qualifier entity suggests that including it may not have eased the task. Hence, it may be valuable to further the study of how the measurement extraction problem can be modelled to resolve some of the ambiguities for context extraction.

Finally, while we tried to stay as closely to the original annotation guidelines as proposed by Harper et al. (2021) as possible (with the exception of the two cases explicated in Appendix C, there is a high likelihood of annotation drift. The re-annotators of the MSP corpus were not involved in the original MeasEval annotation procedure and it is possible that the interpretation of the annotation guidelines was slightly different at places than the authors have originally intended. Our adaption of the annotation guidelines can be found at the end of this paper (Appendix J)

References

Anthony Aue and Michael Gamon. 2005. Customizing Sentiment Classifiers to New Domains: a Case Study. In *Submitted to RANLP-05, the International Conference on Recent Advances in Natural Language Processing*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Callum J. Court and Jacqueline M. Cole. 2018. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Scientific data*, 5:180111.

Adis Davletov, Denis Gordeev, Nikolay Arefyev, and Emil Davletov. 2021. LIORI at SemEval-2021 Task 8: Ask Transformer for measurements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1249–1254, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thaer M. Dieb, Masaharu Yoshioka, Shinjiro Hara, and Marcus C. Newton. 2015. Framework for automatic information extraction from research papers on nano crystal devices. *Beilstein journal of nanotechnology*, 6:1872–1882.

Elsevier Labs. 2015. OA STM Corpus: A corpus, and small treebank, of Open Access journal articles from multiple disciplines in Science, Technology, and Medicine.

Steffen Epp, Marcel Hoffmann, Nicolas Lell, Michael Mohr, and Ansgar Scherp. 2021. A Machine Learning Pipeline for Automatic Extraction of Statistic

Reports and Experimental Conditions from Scientific Papers.

Luca Foppiano, Laurent Romary, Masashi Ishii, and Mikiko Tanifuji. 2019. Automatic Identification and Normalisation of Physical Measurements in Scientific Literature. In *Proceedings of the ACM Symposium on Document Engineering 2019*, pages 1–4, New York, NY, USA. ACM.

Daniel Fried, Nikita Kitaev, and Dan Klein. 2019. Cross-Domain Generalization of Neural Constituency Parsers.

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.

Jinlan Fu, Pengfei Liu, and Graham Neubig. 2020. Interpretable Multi-dataset Evaluation for Named Entity Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6058–6069, Online. Association for Computational Linguistics.

Akash Gangwar, Sabhay Jain, Shubham Sourav, and Ashutosh Modi. 2021. Counts @IITK at SemEval-2021 Task 8: SciBERT Based Entity And Semantic Relation Extraction For Scientific Data.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Tianyong Hao, Hongfang Liu, and Chunhua Weng. 2016. Valx: A System for Extracting and Structuring Numeric Lab Test Comparison Statements from Text. *Methods of information in medicine*, 55(3):266–275.

Corey Harper, Jessica Cox, Curt Kohler, Antony Scerri, Ron Daniel Jr., and Paul Groth. 2021. SemEval-2021 Task 8: MeasEval – Extracting Counts and Measurements and their Related Contexts. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 306–316, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Kyle Hundman and Chris A. Mattmann. 2017. Marve: A measurement relation extractor.

Chia-Chien Hung, Anne Lauscher, Simone Paolo Ponzetto, and Goran Glavaš. 2021. DS-TOD: Efficient Domain Specialization for Task Oriented Dialog.

Yanna Shen Kang and Mehmet Kayaalp. 2013. Extracting laboratory test information from biomedical text. *Journal of pathology informatics*, 4:23.

Seungwon Kim, Alex Shum, Nathan Susanj, and Jonathan Hilgart. 2021. Revisiting Pretraining with Adapters. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 90–99, Online. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition): Chapter 11*. Sage Publications.

Martin Lentschat, Patrice Buche, Juliette Dibie-Barthelemy, and Mathieu Roche. 2020. SciPuRe: a new Representation of textual data for entity identification from scientific publications. In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics*, pages 220–226, New York, NY, USA. ACM.

Patrick Liu, Niveditha Iyer, Erik Rozi, and Ethan A. Chi. 2021. Stanford MLab at SemEval-2021 Task 8: 48 Hours Is All You Need. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1245–1248, Online. Association for Computational Linguistics.

Niels Mündler. 2021. quantulum3: Python library for information extraction of quantities, measurements and their units from unstructured text.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

M. Sevenster, J. Buurman, P. Liu, J. F. Peters, and P. J. Chang. 2015. Natural Language Processing Techniques for Extracting and Categorizing Finding Measurements in Narrative Radiology Reports. *Applied clinical informatics*, 6(3):600–110.

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie-Barthelemy, and Mathieu Roche. 2013. How to Extract Unit of Measure in Scientific Documents? In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval and the International Conference on Knowledge Management and Information Sharing*, pages 249–256. SCITEPRESS - Science and and Technology Publications.

Matthew C. Swain and Jacqueline M. Cole. 2016. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of chemical information and modeling*, 56(10):1894–1904.

Anthony J. Viera and Joanne M. Garrett. 2005. Understanding interobserver agreement: the kappa statistic. *Family medicine*, 37(5):360–363.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Sicheng Zhao, Bo Li, Colorado Reed, Pengfei Xu, and Kurt Keutzer. 2020. Multi-source Domain Adaptation in the Deep Learning Era: A Systematic Survey.

Appendix

A Extended Description of the Data Model

Below we explicate each of the entity types, relations and their associated cardinalities of our data model which are largely based on Harper et al. (2021)'s definitions.

1. **Entities** define *what* is to be extracted:
    - **Quantity (Q)**: A quantity is made up of a) one or more numeric **values** or counts signifying amounts or measurements and optionally b) a **Unit** (U) indicating the magnitude of the values. According to the MeasEval annotation guidelines, values and units are annotated in one span where possible. Contiguous values of a range or a list belong to the same Quantity span (e.g. "Possible beverage sizes are *200, 300 or 400 ml*").
    - **MeasuredEntity (ME**: A measured entity is the *object*, event or phenomenon, whose quantifiable property is measured.
    - **MeasuredProperty (MP)**: A measured property is a quantifiable property of the MeasuredEntity, i.e., the *measurand* that can be attributed to the measured *object*.
2. **Relations** define *how many* entities can be extracted and how they *relate* to each other:
    - **HasQuantity (HQ)**: This relationship links the context entities to their respective quantities. This relation can be drawn from a MeasuredEntity to the Quantity, if no associated MeasuredProperty exists. Otherwise, it is drawn from the MeasuredProperty to the Quantity. The cardinalities of this relationship show that there can be at most one HasQuantity relation for any Quantity span, whereas any MeasuredProperty or MeasuredEntity can be linked to multiple Quantity spans. Consequently, there can be at most one MeasuredProperty and one MeasuredEntity linked to any Quantity span. Consider the sentence *"The book was 600 pages long and weighed 0.5 kg."*. Here, the MeasuredEntity *"book"* can be linked to two Quantity spans.
    - **HasProperty (HP)**: This relation shows which MeasuredProperties can be attributed to a MeasuredEntity. While there can be MeasuredEntities without associated MeasuredProperties, the MeasEval data scheme prescribes that there must be a MeasuredEntity for any MeasuredProperty.

B MSP automatic label suggestions for re-annotation

Unit	MeasuredEntity	MeasuredProperty
Amount-Unit	Material	Apparatus-Property-Type
Property-Unit	Material-Descriptor	Amount-Misc
Apparatus-Unit	Nonrecipe-Material	Property-Type
Condition-Unit	Synthesis-Apparatus	Condition-Misc
	Apparatus-Descriptor	Condition-Type
	Characterization-Apparatus	
	Property-Misc	

Table 8: Mapping of MSP entities to measurement extraction entities

C Annotation Guidelines for the Re-Annotation of the MSP dataset

We provide the complete annotation guidelines for the re-annotation of the MSP dataset as at the end of this document (see Appendix (J).

Below we explicate specific re-annotation guidelines, diverging from the original measurement extraction guidelines provided by Harper et al. (2021).

Specific re-annotation guidelines Over the course of the annotation procedure, the annotators have agreed on additional guiding principles to better capture the relationships between measurements and entities in the context of experiments as described by the synthesis procedures:

Material-centered annotation As a general rule, we prioritize the annotation of experimental participants over other attributes of the experimental procedure. In the sentence "The mixture of elements were heated in evacuated quartz ampoules at 1220 K [...]", we would annotate the "mixture of elements" as the MeasuredEntity of the Quantity "1220 K" as opposed to the "evacuated quartz ampoules".

Experimental conditions Temperatures, times or rates specify the conditions under which experimental operations are performed. We annotate the activity for which the conditions apply as the MeasuredProperty and experiment participants which are worked on under

these conditions as the MeasuredEntity (Table 9).

Sentence	Cleaned sponge and diatom opal was dissolved via wet alkaline digestion at 100 °C for 40 min.		
	Quantity	**MeasuredProperty**	**MeasuredEntity**
Our guideline	100 °C	wet alkaline digestion	Cleaned sponge and diatom opal
MeasEval guideline	100 °C		wet alkaline digestion

Table 9: Example for the annotation of experimental conditions. MeasEval annotations taken from Harper et al. (2021)'s corpus.

Transformations Experimental procedures often describe transformations of the MeasuredEntities before a measurable operation occurs. It is often not possible to pin-point one particular noun phrase that represents the entity to which the operation is being applied. Thus, we annotate all prior steps that are relevant for the operation as the MeasuredEntity. (Fig. 10)

Sentence	To prepare C3N4-Pd composites, the as-prepared g-C3N4 was added into 100 mL ethanol and was sonicated for 2 h to obtain thin g-C3N4 nanosheets.		
	Quantity	**MeasuredProperty**	**MeasuredEntity**
Our guideline	2 h	sonicated	as-prepared g-C3N4 was added into 100 mL ethanol
MeasEval guideline	2 h		sonicated

Table 10: MSP example for the annotation of experimentally transformed MeasuredEntities

Although these guidelines deviate from the original MeasEval annotation guidelines, we believe that these rules are appropriate exceptions to accommodate the nature of experimental procedures, as these rules promote more information regarding measurements to be extracted. This goes in the direction of the "multiple hypothesis hypothesis" proposed by the authors of the MeasEval task, wherein they postulate that different interpretations of contextual information can be useful in different downstream applications (Harper et al., 2021).

D Inter-annotator-agreement study

We conducted an IAA study for the re-annotation of the MSP dataset which spanned five rounds. For the annotation procedure we used the annotation tool prodigy [8]. After each round, the IAA was analyzed both through comparing the agreement score and the annotations themselves. The final annotation was chosen by selecting the annotation on which most annotators agreed. When there was no agreement, a discussion with all annotators decided either on the solution that adhered most closely to the existing guidelines or an amendment to the guidelines. As agreement measures, we calculate Krippendorff's Alpha coefficient (Krippendorff, 2004). To ensure comparability, we follow the same implementation steps as the MeasEval authors and calculate the disagreement on the char-level using the python package simpledorff [9]. Under that assumption, each character in an annotation sample is treated as a "markable" entity with its own label. Figure 7 shows the development of

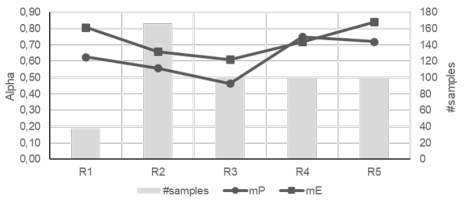

Figure 7: Krippendorff's Alpha over five annotation rounds on mapping MSP to the MeasEval data model.

agreement for the annotation of MeasuredEntity and MeasuredProperty over the five rounds The scores for the Unit and Value entities agreement were always near 1.0 and thus excluded from the analysis. The dip in agreement in round three was mainly due to a conflicting understanding of the supposed span length of MeasuredProperties. Having resolved this conflict, a "substantial" (Viera and Garrett, 2005) agreement > 0.67 could be achieved in round four and reproduced in round five.

Although there were a number of ambiguous cases, the structure and content of experimental descriptions is mostly simple and formulaic. This is reflected in the moderate to high IAA scores compared to the MeasEval IAA, where the scores for both ME (0.55) and MP (0.64) are lower.

E Corpus Overview

Table 11 details the main characteristics of each dataset.

F Domain similarity

Following the approach of Gururangan et al. (2020), we investigate the domain similarity of our datasets

[8] https://prodi.gy/

[9] https://github.com/LightTag/simpledorff

		MeasEval	MSP	Battery Materials
# Sentences/claims		1250, 415, 724	860, 89, 128	194, 86, 72
#Unigrams		38,897	17,062	16,788
#Unique unigrams		9,029	4,024	1,382
Ratio		23%	24%	8%
Quantity	Total ents	882, 281, 499	1671, 195, 201	278, 118, 102
	Unique/total ents	0.83, 0.89, 0.81	0.5, 0.72, 0.71	0.62, 0.69, 0.78
Measured Entity	Total ents	875, 273, 499	1669, 193, 199	278, 118, 102
	Unique/total ents	0.7, 0.6, 0.7	0.42, 0.61, 0.67	0.36, 0.36, 0.55
	Example ents	*'cells', 'electrons', 'samples', 'soil'*	*mixture', 'solution', 'reaction', 'V2O5'*	*'secondary particles', 'lithium metal oxide powder', 'precursor'*
Measured Property	Total entities	563, 179, 330	1379, 145, 157	263, 118, 99
	Unique/total ents	0.7, 0.61, 0.71	0.28, 0.41, 0.48	0.2, 0.34, 0.31
	Example ents	*'n', 'depth', 'p', 'odds ratios', 'ratio',*	*dissolved', 'dried', 'calcined', 'heated'*	*particle size distribution', 'tap density', 'sodium level', 'average particle size',*

Table 11: Main characteristics of the datasets by data split (train, val, test)

by studying their vocabulary overlap. The vocabulary overlap is based on the ratio of shared unigrams which we gather by tokenizing the texts with scispacy [10].

a)

500 most common	Meas Eval	MSP	Battery Materials
MeasEval			
MSP	49%		
Battery Materials	43%	89%	

b)

1000 most common	Meas Eval	MSP	Battery Materials
MeasEval			
MSP	25%		
Battery Materials	22%	45%	

Figure 8: Vocabulary overlap between datasets: a) Overlap over 500 most common unigrams, b) Overlap over 1000 most common unigrams

The matrices in Figure 8 show the resulting vocabulary overlap of the three datasets. They highlight the similarity between the MSP and BM dataset, which is especially pronounced in the comparison of the 500 most common unigrams with an overlap of 89%. All in all, we assume that the MSP and BM corpus share the most similarity, followed by MeasEval and MSP and MeasEval and BM.

G Implementation Details

Below we lay out our implementation details for pre-training, and fine-tuning of the model setup.

Computing infrastructure	Ubuntu 18.04.6 LTS (GNU/Linux 5.4.0)
CUDA Version	11.6
GPU Type	Tesla V100-SXM2-32GB
Available GPUs	8
Python version	3.8

Table 12: Environment details

Table 12 describes our computational infrastructure. We intermediately pre-train and fine-tune our base-models BERT-base-uncased (bert-base-uncased) and SciBERT-uncased (allenai/scibert_scivocab_uncased) which both have a 12 hidden layers with a hidden size of 768.

Adaptive Pre-training Full intermediate pre-training was carried out using the masked language modeling script provided by the Huggingface Transformers Library [11]. For adapter-based pre-training we use AdapterHub (Pfeiffer et al., 2020) as well as their script for masked language modeling [12]. The hyperparameters are given in Table 13. Adapter config and reduction factor are adapter pre-training exclusive parameters. Except for the parameters in the table we use the default values provided by the script. For adaptive pre-training we did performed no systematic hyperparameter search, so there might be more optimal parameter settings.

Implementation framework	huggingface/ AdapterHub run-mlm.py script
optimizer	Adam
adam betas	0.9, 0.98
adam epsilon	1e-06
adapter config	pfeiffer+inv
reduction factor	12
learning rate	0.0001
bs	64
lr scheduler type	linear
lr scheduler warmup steps	100
num epochs	40
evaluation strategy	epoch
seed	42

Table 13: Pre-training hyperparameters for (adapter) TAPT. Pre-training was implemented based on the run-mlm.py script provided by huggingface / AdapterHub.

Hyperparameter search for fine-tuning Hyperparameter tuning was performed using the ray.tune optimization framework for scalable hyperparameter tuning [13]. The tuning details are shown in Tables 14. The training and validation loops are implemented with pytorch-lightning [14], a research framework built on pytorch [15]. We train the models for Task 1 and Task 2 independently from each other, meaning that we train and tune our Task 2 models based on gold Quantities instead of prediction outputs from a Task 1 model. This is done by pre-enriching the Task 2 training sequences with

[10] https://allenai.github.io/scispacy/; en_core_sci_lg model

[11] https://github.com/huggingface/transformers/blob/main/examples/pytorch/language-modeling/run_mlm.py

[12] https://github.com/adapter-hub/adapter-transformers/blob/master/examples/pytorch/language-modeling/run_mlm.py

[13] https://docs.ray.io/en/latest/tune/index.html

[14] https://pytorch-lightning.readthedocs.io/en/stable/

[15] https://pytorch.org/docs/stable/index.html

special tokens ([Q] and [Q]) based on gold Quantity spans which simulates a perfect Task 1 performance. For future work, it might be also interesting to train and tune on the end-to-end pipeline. We optimize the models based the development strict F1 score, which is calculated by comparing predicted and gold BIO-tag sequences. We find that most models have their best parameter setting at a learning rate of 1e-05 or 5e-5 and a batch size of 16 or 32. Only the adapter-based models benefit from larger training rates of 1e-4 or 2e-4.

Implementation framework	pytorch-lightning + ray tune
scheduler	ASHA scheduler
optimizer	Adam
max length	512
max epochs	15
patience	5
gradient clipping	max norm 1.0
lr	[1e-05, 5e-5, 1e-4, 2e-4]
bs	[8, 16, 32, 64]
weight decay	0.01
stochastic weight averaging	yes
seed	1

Table 14: Fine-tuning hyperparameters for Task 1 and Task 2

H Test Results Tables

Table 15 shows the complete results table on the test data of Task 1. Table 16 shows the complete results table on the test data of Task 2. All scores refer to the strict F1, not overlap F1.

I Development Result Tables

Tables 17, 18, and 19 show the Task 1, Task 2 and end-to-end results on the development portion of the corpus. All scores refer to the strict F1, not overlap F1.

Training Mode	Source Domain	Model	PT Setup	Target Domain			
				Meas Eval	MSP	BM	Overall
Single-Source	MeasEval	BERT	No PT	0.777	0.697	0.385	0.671
			Full PT	0.759	0.849	*0.409*	0.702
			Adapter PT	0.749	0.840	0.369	0.688
		SciBERT	No PT	***0.786***	0.897	0.396	*0.721*
			Full PT	***0.786***	*0.913*	0.378	0.719
			Adapter PT	0.783	0.889	0.377	0.715
	MSP	BERT	No PT	0.627	0.924	0.369	0.632
			Full PT	0.637	0.923	0.344	0.634
			Adapter PT	0.558	0.893	0.424	0.607
		SciBERT	No PT	0.653	0.908	*0.455*	*0.667*
			Full PT	0.656	0.913	0.439	0.667
			Adapter PT	*0.659*	*0.935*	0.418	0.665
	BM	BERT	No PT	0.256	0.136	0.500	0.290
			Full PT	0.255	0.182	0.451	0.285
			Adapter PT	0.271	0.282	0.442	0.315
		SciBERT	No PT	*0.354*	*0.339*	***0.521***	*0.386*
			Full PT	0.324	0.301	0.505	0.357
			Adapter PT	0.208	0.049	0.387	0.222
Duo-Source	MSP+ MeasEval	BERT	No PT	0.744	0.915	0.402	0.710
			Full PT	0.761	0.893	0.448	0.726
			Adapter PT	0.763	***0.938***	0.422	0.732
		SciBERT	No PT	*0.778*	0.908	*0.457*	***0.739***
			Full PT	0.776	0.925	0.398	0.721
			Adapter PT	0.767	0.935	0.417	0.727

Table 15: Test F1 scores of Task 1: Quantity Extraction. **Bold** scores indicate the highest score across an entire target domain. *Italic* scores indicate the highest score within one source domain.

Training Mode	Source Domain	Model	PT Setup	Target Domain: MeasEval O	MeasEval U	MeasEval ME	MeasEval MP	MSP O	MSP U	MSP ME	MSP MP	BM O	BM U	BM ME	BM MP	Overall O	Overall U	Overall ME	Overall MP
Single-Source	Meas Eval	BERT	No PT	0.602	0.949	0.436	0.436	0.663	0.984	0.463	0.502	0.627	*0.957*	0.470	0.563	0.621	0.960	0.448	0.473
			Full PT	0.591	0.963	0.416	0.411	0.624	0.978	0.436	0.434	0.604	0.917	0.436	0.574	0.602	0.963	0.424	0.446
			Adapter PT	0.540	0.926	0.366	0.385	0.598	0.984	0.393	0.400	0.561	0.829	0.480	0.469	0.557	0.932	0.388	0.403
		SciBERT	No PT	*0.638*	***0.970***	*0.494*	*0.460*	*0.686*	*0.995*	*0.494*	*0.563*	0.676	0.929	0.545	0.667	*0.655*	*0.972*	*0.501*	*0.522*
			Full PT	0.628	0.958	0.481	0.450	0.648	0.989	0.468	0.468	***0.718***	0.900	*0.584*	*0.726*	0.646	0.961	0.491	0.508
			Adapter PT	0.574	0.951	0.424	0.369	0.625	0.973	0.405	0.506	0.604	0.905	0.519	0.502	0.591	0.952	0.431	0.425
	MSP	BERT	No PT	0.547	0.953	*0.385*	0.316	***0.750***	***1.000***	***0.589***	0.673	0.627	***0.971***	0.505	0.514	0.612	*0.968*	0.452	0.445
			Full PT	0.530	0.943	0.361	0.316	***0.751***	0.997	0.571	0.697	0.661	0.922	0.510	*0.633*	0.606	0.956	0.437	0.461
			Adapter PT	0.509	0.944	0.332	0.291	0.730	0.995	0.538	0.677	0.536	0.829	0.395	0.498	0.570	0.946	0.392	0.421
		SciBERT	No PT	*0.567*	0.954	0.372	0.371	0.744	***1.000***	0.556	0.687	*0.690*	0.922	***0.622***	0.615	*0.633*	0.964	*0.456*	*0.502*
			Full PT	0.561	*0.961*	0.368	*0.379*	0.748	***1.000***	***0.580***	0.684	0.668	0.901	0.558	0.625	0.626	0.965	0.450	0.500
			Adapter PT	0.531	0.948	0.310	0.365	0.737	0.995	0.532	***0.711***	0.608	0.863	0.479	0.588	0.597	0.952	0.393	0.491
	BM	BERT	No PT	0.353	*0.898*	0.094	0.141	0.382	0.755	0.157	0.102	0.575	***0.971***	0.229	0.628	0.389	*0.865*	0.125	0.216
			Full PT	0.352	0.801	0.130	0.202	0.379	0.826	0.124	0.180	0.618	0.971	0.294	0.694	0.396	0.828	0.148	0.297
			Adapter PT	0.305	0.782	0.076	0.151	*0.410*	*0.828*	0.189	0.121	0.524	0.906	0.187	0.634	0.363	0.810	0.119	0.238
		SciBERT	No PT	0.366	0.759	*0.195*	*0.264*	0.356	0.646	*0.251*	*0.195*	*0.650*	0.929	*0.398*	0.738	0.405	0.750	*0.235*	*0.355*
			Full PT	*0.375*	0.779	0.166	0.220	0.344	0.634	0.214	0.185	0.648	0.900	0.387	***0.757***	*0.409*	0.755	0.210	0.329
			Adapter PT	0.249	0.548	0.108	0.153	0.218	0.417	0.135	0.099	0.570	0.882	0.292	0.680	0.296	0.559	0.143	0.263
Duo-Source	BM + Meas Eval	BERT	No PT	0.558	0.959	0.380	0.389	0.618	0.992	0.435	0.428	0.633	0.950	0.439	0.631	0.584	0.968	0.401	0.441
			Full PT	0.604	0.957	0.440	0.431	0.627	0.989	0.421	0.440	0.658	***0.971***	*0.466*	0.650	0.617	0.967	0.439	0.474
		SciBERT	No PT	0.639	***0.968***	0.489	0.471	0.660	*0.995*	0.450	*0.532*	*0.678*	0.914	0.443	***0.769***	0.650	*0.970*	0.472	*0.540*
			Full PT	***0.647***	0.962	***0.503***	***0.490***	*0.662*	0.989	*0.464*	0.522	0.653	0.914	0.423	0.709	*0.652*	0.965	*0.482*	0.537
	BM + MSP	BERT	No PT	0.560	*0.965*	0.394	0.340	0.746	0.992	0.574	0.694	0.675	*0.957*	0.500	0.667	0.626	*0.972*	*0.455*	0.489
			Full PT	0.536	0.958	0.340	0.343	0.736	0.989	0.545	*0.696*	*0.695*	*0.957*	*0.534*	*0.689*	0.610	0.967	0.417	0.496
		SciBERT	No PT	0.559	0.961	0.373	*0.379*	***0.750***	***1.000***	*0.579*	0.688	0.641	0.914	0.434	0.676	0.621	0.967	0.435	*0.506*
			Full PT	*0.570*	0.956	*0.404*	0.367	0.738	***1.000***	0.542	0.695	0.650	0.908	0.459	0.667	*0.626*	0.963	0.448	0.502
	MSP + Meas Eval	BERT	No PT	0.599	0.962	0.437	0.420	0.745	0.995	0.551	***0.716***	0.669	0.929	0.511	0.660	0.648	0.968	0.477	0.538
			Full PT	0.611	0.966	0.471	0.396	0.736	0.997	0.560	0.662	0.665	*0.957*	0.585	0.554	0.651	***0.974***	0.508	0.496
		SciBERT	No PT	***0.651***	0.962	***0.499***	***0.510***	*0.749*	***1.000***	0.568	0.704	***0.721***	0.929	***0.622***	0.682	***0.687***	0.969	***0.534***	***0.589***
			Full PT	0.622	***0.972***	0.478	0.446	0.745	***1.000***	*0.570*	0.687	***0.746***	0.930	***0.652***	*0.721*	0.671	***0.975***	***0.523***	0.557
All Sources	MSP+ Meas Eval + BM	BERT	No PT	0.610	0.962	0.464	0.410	0.730	***1.000***	0.560	0.629	0.631	***0.986***	0.384	0.634	0.645	***0.975***	0.479	0.506
			Full PT	0.609	0.960	0.445	0.444	0.747	0.989	***0.588***	0.669	0.595	***0.971***	0.276	0.657	0.644	0.969	0.460	0.536
			Adapter PT	0.556	0.959	0.370	0.375	0.705	0.995	0.520	0.627	0.618	0.957	0.377	0.634	0.604	0.969	0.411	0.484
		SciBERT	No PT	0.634	*0.965*	0.482	0.476	*0.748*	***1.000***	0.562	0.700	0.698	0.929	*0.546*	0.692	***0.673***	0.971	0.512	***0.569***
			Full PT	***0.654***	0.963	***0.515***	***0.499***	0.741	***1.000***	0.556	0.691	*0.702*	0.943	0.500	***0.756***	***0.684***	0.972	***0.524***	***0.593***
			Adapter PT	0.575	0.952	0.413	0.398	0.737	0.986	0.547	***0.719***	0.667	0.900	0.466	0.725	0.630	0.956	0.454	0.535

Table 16: Test F1 scores of Task 2: Context extraction. **Bold** scores indicate the highest score across an entire target domain. *Italic* scores indicate the highest score within one source domain. O = Overall, U = Unit, ME = MeasuredEntity, MP = MeasuredProperty.

Training mode	Source domain	Model	PT Setup	MeasEval	MSP	BM	Overall
Single-Source	Meas Eval	BERT	No PT	0.749	0.756	0.281	0.629
			Full PT	0.757	0.856	0.285	0.661
			Adapter PT	0.705	0.805	0.304	0.624
		SciBERT	No PT	**0.768**	0.845	0.284	0.653
			Full PT	**0.774**	0.855	0.302	0.663
			Adapter PT	0.745	0.840	0.308	0.649
	MSP	BERT	No PT	0.625	0.922	0.195	0.584
			Full PT	0.584	**0.927**	0.187	0.581
			Adapter PT	0.522	0.912	0.222	0.563
		SciBERT	No PT	0.653	**0.931**	0.275	0.627
			Full PT	0.643	0.924	0.268	0.621
			Adapter PT	0.610	0.919	0.234	0.583
	BM	BERT	No PT	0.296	0.103	0.612	0.329
			Full PT	0.363	0.083	**0.621**	0.356
			Adapter PT	0.293	0.171	0.586	0.337
		SciBERT	No PT	0.356	0.179	**0.628**	0.373
			Full PT	0.361	0.228	**0.669**	0.399
			Adapter PT	0.224	0.065	0.580	0.278
Duo-Source	MSP+ Meas Eval	BERT	No PT	0.751	0.922	0.354	**0.704**
			Full PT	0.753	0.909	0.351	**0.700**
			Adapter PT	0.721	0.899	0.338	0.677
		SciBERT	No PT	0.756	**0.937**	0.324	**0.687**
			Full PT	**0.762**	0.909	0.313	0.681
			Adapter PT	0.756	0.897	0.320	0.676

Table 17: Development F1 scores of Task 1: Quantity Extraction. **Bold** scores indicate the highest score across an entire target domain. *Italic* scores indicate the highest score within one source domain.

				MeasEval				MSP				BM				Overall			
Training mode	Source domain	Model	PT Setup	All	U	ME	MP	All	U	ME	MP	All	U	ME	MP	All	U	ME	MP
Single-Source	Meas Eval	BERT	No PT	0.558	0.946	0.370	0.322	0.614	0.986	0.356	0.421	0.451	0.890	0.133	0.455	0.553	0.951	0.315	0.387
			Full PT	0.556	0.946	0.386	0.287	0.630	0.992	0.413	0.456	0.440	0.838	0.165	0.436	0.556	0.943	0.348	0.381
			Adapter PT	0.508	0.937	0.322	0.290	0.576	0.970	0.361	0.353	0.380	0.744	0.136	0.389	0.503	0.916	0.295	0.333
		SciBERT	No PT	0.582	0.950	**0.412**	0.357	0.636	0.984	0.407	0.504	0.504	0.849	0.205	0.596	0.583	0.944	0.365	0.469
			Full PT	0.579	**0.954**	0.385	**0.385**	0.667	0.975	0.468	0.517	0.547	0.883	0.239	0.579	0.602	0.949	0.384	0.482
			Adapter PT	0.531	**0.957**	0.340	0.297	0.634	0.978	0.408	0.482	0.451	0.766	0.192	0.482	0.548	0.931	0.333	0.402
	MSP	BERT	No PT	0.516	0.943	0.307	0.302	0.768	0.992	0.588	**0.728**	0.522	0.914	0.216	0.531	0.604	0.956	0.384	0.499
			Full PT	0.473	0.946	0.239	0.253	0.774	**0.997**	0.612	0.703	0.508	0.909	0.189	0.525	0.581	0.958	0.355	0.452
			Adapter PT	0.461	0.915	0.258	0.240	0.732	0.989	0.569	0.644	0.474	0.836	0.167	0.542	0.554	0.928	0.342	0.440
		SciBERT	No PT	0.532	0.948	0.324	0.291	**0.796**	0.992	**0.643**	**0.744**	0.477	0.818	0.167	0.537	0.610	0.941	0.395	0.509
			Full PT	0.527	0.950	0.308	0.329	0.770	0.986	0.608	0.724	0.534	0.807	0.247	0.624	0.611	0.938	0.398	0.531
			Adapter PT	0.475	0.940	0.227	0.262	0.744	0.989	0.549	0.693	0.456	0.701	0.175	0.567	0.561	0.916	0.323	0.475
	BM	BERT	No PT	0.372	0.855	0.076	0.216	0.398	0.827	0.094	0.074	0.651	0.895	0.385	0.709	0.439	0.852	0.145	0.316
			Full PT	0.408	0.859	0.119	0.262	0.355	0.784	0.102	0.098	0.659	**0.955**	**0.417**	0.672	0.450	0.852	0.175	0.354
			Adapter PT	0.347	0.810	0.052	0.234	0.385	0.840	0.114	0.079	0.595	0.881	0.337	0.650	0.415	0.834	0.134	0.325
		SciBERT	No PT	0.377	0.791	0.149	0.323	0.268	0.508	0.188	0.126	0.659	0.909	**0.452**	0.688	0.407	0.729	0.216	0.396
			Full PT	0.398	0.824	0.112	0.292	0.304	0.568	0.205	0.104	0.665	0.933	0.370	**0.763**	0.432	0.764	0.198	0.403
			Adapter PT	0.300	0.611	0.119	0.199	0.236	0.405	0.176	0.099	0.618	0.807	0.389	0.727	0.365	0.588	0.203	0.359
Duo-Source	BM + Meas Eval	BERT	No PT	0.532	0.937	0.338	0.354	0.587	0.984	0.347	0.439	0.648	0.939	0.411	0.686	0.575	0.954	0.355	0.470
			Full PT	0.556	0.936	0.369	0.353	0.627	0.989	0.410	0.454	0.650	**0.961**	0.369	0.721	0.601	0.960	0.382	0.491
		SciBERT	No PT	0.579	**0.952**	0.400	0.371	0.663	0.978	0.458	0.522	**0.668**	0.905	0.405	**0.769**	0.626	0.953	0.420	0.529
			Full PT	**0.587**	0.951	**0.423**	0.377	0.675	0.964	0.508	0.504	0.645	0.910	0.381	0.724	0.628	0.948	0.441	0.510
	BM + MSP	BERT	No PT	0.522	0.946	0.310	0.308	0.754	0.989	0.586	0.694	0.650	0.939	0.388	0.690	0.629	0.960	0.420	0.532
			Full PT	0.488	0.944	0.246	0.305	0.756	**0.997**	0.567	0.713	0.654	0.950	0.349	0.739	0.613	**0.964**	0.374	0.543
		SciBERT	No PT	0.527	0.948	0.314	0.323	**0.778**	0.992	**0.628**	0.719	0.655	0.843	**0.444**	0.728	0.639	0.945	0.448	0.554
			Full PT	0.538	**0.952**	0.338	0.317	0.767	0.992	0.625	0.669	0.653	0.899	0.414	0.691	0.640	0.957	0.451	0.526
	MSP + Meas Eval	BERT	No PT	0.556	0.948	0.374	0.337	0.746	0.986	0.561	0.696	0.523	0.893	0.187	0.596	0.614	0.952	0.398	0.522
			Full PT	0.554	0.939	0.361	0.362	0.751	**0.997**	0.576	0.692	0.500	0.939	0.185	0.515	0.610	0.960	0.396	0.515
		SciBERT	No PT	0.569	0.948	0.391	0.360	0.771	**0.997**	0.609	0.720	0.562	0.872	0.241	0.620	0.639	0.952	0.441	0.546
			Full PT	**0.594**	0.952	**0.441**	**0.380**	0.759	**0.997**	0.595	0.691	0.528	0.872	0.203	0.635	0.635	0.954	0.442	0.549
All Sources	MSP+ Meas Eval + BM	BERT	No PT	0.568	0.950	0.391	0.346	0.762	**0.997**	0.586	0.703	0.624	0.939	0.373	0.637	**0.646**	**0.965**	**0.454**	0.538
			Full PT	0.545	0.938	0.336	0.351	0.763	**0.997**	0.576	0.724	0.661	**0.961**	0.395	0.695	0.644	**0.963**	0.430	0.565
			Adapter PT	0.542	0.946	0.346	0.342	0.693	0.989	0.474	0.644	0.627	0.927	0.367	0.675	0.613	0.958	0.395	0.527
		SciBERT	No PT	**0.586**	0.952	0.389	**0.412**	0.764	0.992	0.594	0.716	**0.676**	0.915	0.417	**0.754**	**0.667**	0.960	**0.467**	**0.598**
			Full PT	0.577	0.943	0.405	0.365	**0.779**	0.992	**0.644**	0.687	**0.670**	0.916	0.412	0.744	**0.666**	0.956	**0.485**	**0.577**
			Adapter PT	0.550	0.948	0.363	0.340	0.751	0.986	0.559	**0.733**	0.653	0.883	0.403	0.733	0.641	0.950	0.438	**0.571**

Table 18: Development F1 scores of Task 2: Context extraction. **Bold** scores indicate the highest score across an entire target domain. *Italic* scores indicate the highest score within one source domain. O = Overall, U = Unit, ME = MeasuredEntity, MP = MeasuredProperty.

		MeasEval					MSP					BM					Overall				
Source domains	Model configuration by Task	All	Q	U	ME	MP	All	Q	U	ME	MP	All	Q	U	ME	MP	All	Q	U	ME	MP
MeasEval only	T1: SciBERT Full PT; T2: SciBERT Full PT	0.555	0.840	0.847	0.317	0.312	0.663	0.872	0.909	0.472	0.427	0.394	0.375	0.534	0.285	0.416	0.539	0.687	0.785	0.354	0.380
MSP only	T1: SciBERT No PT; T2: SciBERT No PT	0.502	0.777	0.846	0.262	0.218	0.803	0.929	0.958	0.668	0.649	0.340	0.377	0.446	0.203	0.374	0.529	0.681	0.769	0.341	0.378
BM only	T1: SciBERT Full PT; T2: SciBERT Full PT	0.342	0.543	0.582	0.115	0.209	0.188	0.298	0.215	0.140	0.076	0.628	0.775	0.718	0.416	0.677	0.357	0.512	0.476	0.185	0.302
Multi	T1: BM SciBERT Full PT & MSP+MeasEval Full PT; T2: All Sources SciBERT No PT	0.647	0.782	0.893	0.445	0.456	0.776	0.930	0.957	0.532	0.688	0.450	0.505	0.519	0.354	0.440	0.641	0.767	0.848	0.450	0.505

Table 19: Development E2E (strict) F1

J Re-Annotation guidelines for the Material Synthesis Procedural Text Corpus

Scientific knowledge is published and achieved in the form of unstructured texts. Numeric components in the form of counts, measurements and units (e.g. 500mg) and their contexts (e.g. Ibuprofen, dosage) are often crucial information for researchers across all domains. The goal of this annotation task is to prepare data for an end-to-end pipeline, which is able to extract quantities, units, measured objects and properties from texts, as well as the semantic relationships between each other. Section A) of this document describes how the entity and relation labels can be defined in a general setting. Section B) will provide guidance for the annotation of a specific dataset, the Materials Science Procedural Text Corpus by Mysore et al..

A) General task guidelines

The entity and relation labels are described in the following table. They are a subset of the SemEval 2021 Task 8, MeasEval Basic Annotation Set.

- **Number (N)**
  - **Definition**A numeric value or a count signifying an amount or measurement and contiguous specifiers (e.g. >, ~). This is the root entity in each sample, i.e. other entities must always be able to directly refer to a number. Numeric values which do not signify a quantifiable amount (e.g. page numbers, citations, mathematical formulas) are not annotated.
  - **Example** The patient weighted **~100** pounds and was prescribed an Ibuprofen dosage of **500** mg.
- **Unit (U)**
  - **Definition** The unit linked to the Number. To be annotated if available.
  - **Example** The sick patient weighted 100 **pounds** and was prescribed an Ibuprofen dosage of 500 **mg**.
- **measuredEntity (mE)**
  - **Definition** "A required (if possible) span that has a given [Number + Unit] either as its direct value or indirectly via a MeasuredProperty. Every Quantity should ideally be associated with a MeasuredEntity. If no relevant information appears in the text, the Number can be standalone, but can have no other relationships. A MeasuredEntity can be related to either a MeasuredProperty by a HasProperty relationship, or to a Quantity by a HasQuantity relationship." (cited from SemEval 2021 Task 8 annotation guidelines, "Quantity" reference replaced with "Number"). This label describes the concept that is being quantified by the number (and the unit). In most cases the measuredEntity consists of one or more noun phrases (and their specifiers if they are in a contiguous span).
  - **Example** The **sick patient** weighted ~100 pounds and was prescribed an **Ibuprofen** dosage of 500 mg.
- **measuredProperty (mP)**
  - **Definition** "An optional span associated with both a MeasuredEntity and a [Number]. Not every [Number] will be associated with a MeasuredProperty. A MeasuredProperty must be related from a MeasuredEntity by a HasProperty relationship, and must be related to a Quantity through the HasQuantity relationship." (cited from SemEval 2021 Task 8 annotation guidelines, "Quantity" reference replaced with "Number"). The measuredProperty can be interpreted as the "quantity-denoting target-word" of the number (definition from FrameNet). As such is it often a quantifiable specifier or attribute of the measuredEntity (e.g. volume, concentration, temperature etc.), but can also encompass longer target phrases.
  - **Example** The patient **weighted** ~100 pounds and was prescribed an Ibuprofen **dosage** of 500 mg.

Graph representation: The entity labels their relations can be depicted in a graph. This can be especially helpful when identifying the measured entity and measured property or verifiying one's annotations.

Case 1: (N, U) <- hasQuantity <- (mP) <- hasProperty <- (mE)

Case 2: (N, U) <- hasQuantity <- (mE)

Each data sample must contain at least a Number. The other labels are only to be annotated if they are contained in the text. Below are a few hints and rules for the general annotation task. Quantity and the Number label will be used synonymously.

A.1) Multi-class classification

Multi-entity classifications are possible, i.e. a measuredEntity for one Number can be a measuredProperty for another. This can be the case, because classification is always performed from the perspective of the root Number. For the same reason there can be measuredEntities or measuredProperties containing numbers (that are not the root number of the annotation sample).

Examples

> "The lowest input of odd nitrogen corresponds to 3.5-6.1 (x10-4) wt.% N accumulated over 3 byr and mixed into 1.5-2.6 m, of soil."

N	U	mE	mP
3.5-6.1 (x10-4)	wt.%	N	lowest input of odd nitrogen

N	U	mE	mP
3	byr	lowest input of odd nitrogen	accumulated

N	U	mE	mP
1.5-2.6	m	soil	

A.2) Span extent

We annotate measuredEntities and measuredProperties as completely as possible, i.e. using the longest coherent and informative text span. However, we do not annotate copula (e.g. were, have been etc.) prepositions or articles at the beginning or end of a span.

Examples

> "The earth surface temperatures have risen by 0.5 °C compared to baseline levels."

N	U	mE	mP
0.5	°C	earth surface temperatures	risen

A.3) Duplicate measuredEntities mentions

Some sentences will have multiple mentions of the same measuredEntity. We annotate the span that is closest to its root Number.

Example

> "The O2/N ratio was measured with the aforementioned machinery (O2/N = 2.8)."

N	U	mE	mP
2.8		O2/N	ratio

A.4) Part-whole relationships

Fractions and percentages often describe part-whole relationships, where the fraction or percentage describe a partial characteristic of a bigger whole. For annotation, we mark the whole as the measuredEntity and the part as the measuredProperty.

Examples

> "The hamburger consisted of 30% patty and 10% cheese."

N	U	mE	mP
30	%	hamburger	patty

N	U	mE	mP
10	%	hamburger	cheese

> "Steam activation was carried out by heating an amount of sample in a flow of 10% water vapor."

N	U	mE	mP
10	%	flow	water vapor

Part-whole relationships can also be described without the use of fractions or percentages:

> "The patty of the hamburger was 200g."

N	U	mE	mP
200	g	hamburger	patty

Graph representation: hamburger -> hasProperty -> patty -> hasQuantity -> 200g

A.5) Non-noun measuredProperties

MeasuredProperties can also be verbs or adjectives. To test whether a verb can be a measuredProperty, one can reformulate the sentence using the nominalized verb form valditate with the graph representation to check if all relations can be applied correctly.

Examples

> "The earth surface temperatures have risen by 0.5 °C compared to baseline levels."

Reformulated: There has been a rise of earth surface temperature by 0.5 °C compared to baseline levels.

> "The patient weighed 100 pounds."

Reformulated: The weight of the patient is 100 pounds.

A.6) Hints for the Unit entity

Ratios (e.g. weight ratio) and pH values are not considered units. Instead, they are labeled as measuredProperties.

B) Specific guidelines for Mysore et al.'s Materials Science Procedural (MSP) Text Corpus

Originally, the MSP Corpus contains annotations regarding the materials, operations and conditions of experiments in materials science.
To expand the existing MeasEval Dataset, we need to adapt these annotations to the above-introduced entities and labels.
For this, the dataset was automatically processed beforehand using mapping rules for each pre-existing label (e.g. all materials where labeled as measuredEntities). However, these automatically created labels are often incorrect and must be adjusted which is the main annotation task here.

Characteristics of the data:

Each data sample is pre-labeled with at least a number and in most cases suggestions for the Unit, measuredProperty and measuredEntity are given.
One data sample is created for each Number and its related measuredEntities and measuredProperties. Hence, sentences with multiple quantities and related contexts will yield as many data samples as there are Numbers in the text. Due to the specificity of this corpus, some additional rules apply. They are listed below.

B.1) Number specifiers

Symbols and textual specifiers of Numbers are often not included in the label suggestion. Therefore, we must expand the Number-span to also contain these specifiers. Example: In the first example the suggested number would be "The patient weighted ~ **[100]** pounds...", we would then extend the span to "The patient weighted **[~100]** pounds...".

B.2) Removing irrelevant entities and adjusting spans

Sometimes the there will be suggested entities, that are not related to the root Number. These false suggestions must be removed.
Further, to adhere to the rule for maximum span annotation we adjust spans for measuredEntities and measuredProperties which can be extended.

Examples

> "The gel was ground to powders and then calcined at 400 °C in a muffle furnace under air atmosphere."

Suggested:

N	U	mE	mP
400	°C	muffle furnace, air	calcinated

Corrected:

N	U	mE	mP
400	°C	powders	calcinated

> "The as-synthesized zeolites were calcined at 580 degC for 4 h under a flow of air."

Suggested measuredEntity = zeolites
Corrected measuredEntity = as-synthesized zeolites

B.3) Experiment procedures

A large fraction of this corpus' Numbers describe experimental conditions, e.g. how long a solution was stirred. As a result, the quantities can often not be linked to an explicitly measured object, but only to the object that is being experimented on. Therefore, we mark these objects as measuredEntities and the experimental circumstances as measuredProperties. If the measuredEntity is explicitly given, we mark that as measuredEntity instead of the object that is impacted by the experiment.

Examples

"The obtained sample was washed with absolute ethanol, and then dried at 60 °C for 10h."

N	U	mE	mP
60	°C	obtained sample	dried

N	U	mE	mP
10	h	obtained sample	dried

"The solution was modified by dissolving it in 10 wt% ethanol."

N	U	mE	mP
10	wt%	ethanol	

B.3.1) Experiment operations

We only mark procedural operations (e.g. added or dissolved) as the measuredProperty of a Number and a measuredEntity, if the measuredEntity is the main participant of the operation.

Examples

"The solution was modified by dissolving it in 10 wt% ethanol."

In the example above we do not mark *dissolving* as the mP, because ethanol is not the component that is being dissolved.

"500 g of the sample was dissolved in 10 ml NaCl solution."

N	U	mE	mP
500	g	sample	dissolved

N	U	mE	mP
10	ml	NaCL solution	

Here the sample is the entity that is being dissolved, thus we can mark 'dissolved' as its measuredProperty. Be careful that the operation marked as the measuredProperty has a proper relation to the Number span.

"The composite was ground, pressed and sintered at 300 °C."

N	U	mE	mP
300	°C	composite	sintered

In this example, sintered is the operation which directly related to the temperature measure, whereas the other operations **do not** have a measuredProperty (e.g. what the pressure of the pressing was or how granular the grounding was).

"Copper (99,99%) was purchased from Sigma-Aldrich."

In this case 'purchased' is not the target-word of the 99,99%, as it represents a purity measure and not an amount that was purchased.
Thus, there is no measuredProperty in this sentence.

B.3.2) MeasuredProperty operations span

Operations are often specified by additional descriptors, that are contiguous to the operation or in a separate span. In most cases we only annotate the operation as the measuredProperty, because the descriptors are semantically dependent (the so-called 'oblique nominal') on the operation phrase, which is difficult to express within our annotation scheme.

Examples

"The chemical was heated at 300 °C under constant airflow."

N	U	mE	mP
300	°C	chemical	heated

Here, "under constant airflow" is dependent on "heated". We would need additional labels to capture these kind of multi-level relations, which exceeds the scope of this annotation scheme. In the case, that a contiguous span with multiple properties could be annotated, we proceed in the same manner and only annotate the highest level to stay consistent.

"The chemical was dried in air at 300 °C."

N	U	mE	mP
300	°C	chemical	dried

Exception: Preceding adverbial and adjectival modifiers

We can add such descriptors which occur prior to the operation to the operation measuredProperty span, as they almost exclusively occur together contiguously, thus ensuring consistent annotations.

> "The chemical was under magnetic stirring for 2 h."

N	U	mE	mP
2	h	chemical	magnetic stirring

B.3.3) Properties of operations

When attributes or specifiers of an operation are given we try to mark the main experiment participant as the measuredProperty as opposed to the operation itself.

Example

> "The chemical was calcinated at 300 °C with a heating rate of 10 °C per minute."

N	U	mE	mP
10	°C per minute	chemical	heating rate

B.4) Ambivalent relations

Some experimental descriptions do not explicitly name the measuredEntity, but e.g. only the result of the experimental operation. If this is the case and an experimental operation is also in the sentence, we can mark the experimental operation as the measuredProperty.

Examples

> "The enhanced form was obtained by calcination at 220 °C under a flow of air."

N	U	mE	mP
220	°C	calcination	under a flow of air

This sentence does not mention the object that is being calcinated. Therefore, we annotate the operation as the measuredEntity.

Note that we can also annotate "under a flow of air" as a measuredProperty here, because it can be directly linked to "calcination" and "220 °C" without being dependent on another measuredProperty.

> "NH4OH solution was slowly added until the pH was 10."

N	U	mE	mP
10		pH	

This sentence does not mention, whose pH becomes 10. Because pH is not a unit, we can mark it as the measuredEntity.

B.4.1) Coreferences

If the measuredEntity is mentioned as a coreference, but not explicitly, we annotate the coreference as measuredEntity.

Example

> "Finally, it was filtered, washed with water and ethanol, and vacuum-dried at 70 °C."

N	U	mE	mP
70	°C	it	vacuum-dried

B.4.2) Transformation of the measuredEntity

Synthesis procedures often describe transformations of the measuredEntities before a measurable operation occurs. It is often not possible to pinpoint one particular noun phrase that represents the entity to which the operation is being applied. Instead, we annotate all prior steps that are relevant for the operation as the measuredEntity.

Examples

> "To prepare C3N4-Pd composites, the as-prepared g-C3N4 was added into 100 mL ethanol and was sonicated for 2 h to obtain thin g-C3N4 nanosheets."

N	U	mE	mP
2	h	as-prepared g-C3N4 was added into 100 mL ethanol	sonicated

N	U	mE	mP
100	mL	ethanol	

B.5) Dealing with nested information in brackets

Sometimes additional information about a measuredEntity is given in brackets. We only annotate measuredProperties that are directly related to both the Number and the measuredEntity.

Examples

"20g of gold (99.99% purity) were ground."

N	U	mE	mP
20	g	gold	ground

N	U	mE	mP
99.99	%	gold	purity

"In a typical process, NiCl2*6H2O (0.173 g) was dissolved in a solution."

N	U	mE	mP
0.173	g	NiCl2*6H2O	dissolved

B.6) MeasuredEntity for Ratios

Ratios explain "how many times one number contains another" (Wiki). This should also be expressed in the measuredEntity of a ratio. If the two concepts described by the ratio are explicitly mentioned, annotate them (either in a contiguous span if possible, and separately if not).

Examples

"At a weight ratio of 1:1, the MWCNT@MPC composite was mixed with sublimed sulfur."

N	U	mE	mP
1:1		MWCNT@MPC composite was mixed with sublimed sulfur	weight ratio

B.7) Abbreviations

N	U	mE	mP
100	ml	Hydrochloric acid (HCl	added

We try to include abbreviations into the entity span, if possible.

Example

"Hydrochloric acid (HCl, 100 ml) was added to the mixture."