The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain

Annemarie Friedrich[1] **Heike Adel**[1] **Federico Tomazic**[2] **Johannes Hingerl**[1]
Renou Benteau[1] **Anika Maruscyk**[2] **Lukas Lange**[1]

[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Corporate Research, Robert Bosch GmbH, Renningen, Germany
firstname.lastname@de.bosch.com

Abstract

This paper presents a new challenging information extraction task in the domain of materials science. We develop an annotation scheme for marking information on experiments related to solid oxide fuel cells in scientific publications, such as involved materials and measurement conditions. With this paper, we publish our annotation guidelines, as well as our SOFC-Exp corpus consisting of 45 open-access scholarly articles annotated by domain experts. A corpus and an inter-annotator agreement study demonstrate the complexity of the suggested named entity recognition and slot filling tasks as well as high annotation quality. We also present strong neural-network based models for a variety of tasks that can be addressed on the basis of our new data set. On all tasks, using BERT embeddings leads to large performance gains, but with increasing task complexity, adding a recurrent neural network on top seems beneficial. Our models will serve as competitive baselines in future work, and analysis of their performance highlights difficult cases when modeling the data and suggests promising research directions.

1 Introduction

The design of new experiments in scientific domains heavily depends on domain knowledge as well as on previous studies and their findings. However, the amount of publications available is typically very large, making it hard or even impossible to keep track of all experiments conducted for a particular research question. Since scientific experiments are often time-consuming and expensive, effective knowledge base population methods for finding promising settings based on the published research would be of great value (e.g., Auer et al., 2018; Manica et al., 2019; Strötgen et al., 2019; Mrdjenovich et al., 2020). While such real-life information extraction tasks have received considerable attention in the biomedical domain (e.g., Cohen et al., 2017; Demner-Fushman et al., 2018, 2019), there has been little work in other domains (Nastase et al., 2019), including materials science (with the notable exception of the work by Mysore et al., 2017, 2019).

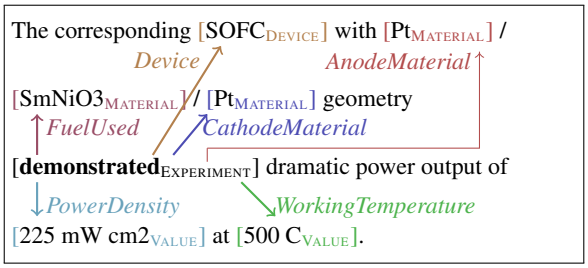

Figure 1: Sentence describing a fuel-cell related experiment, annotated with Experiment frame information.

In this paper, we introduce a new information extraction use case from the materials science domain and propose a series of new challenging information extraction tasks. We target publications about solid oxide fuel cells (SOFCs) in which the interdependence between chosen materials, measurement conditions and performance is complex (see Figure 1). For making progress within natural language processing (NLP), the genre-domain combination presents interesting challenges and characteristics, e.g., domain-specific tokens such as material names and chemical formulas.

We provide a new corpus of open-access scientific publications annotated with semantic frame information on experiments mentioned in the text. The annotation scheme has been developed jointly with materials science domain experts, who subsequently carried out the high-quality annotation. We define an "Experiment"-frame and annotate sentences that evoke this frame with a set of 16 possible slots, including among others *AnodeMaterial*, *FuelUsed* and *WorkingTemperature*, reflecting the

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1255–1268

role the referent of a mention plays in an experiment. Frame information is annotated on top of the text as graphs rooted in the experiment-evoking element (see Figure 1). In addition, slot-filling phrases are assigned one of the types MATERIAL, VALUE, and DEVICE.

The task of finding experiment-specific information can be modeled as a retrieval task (i.e., finding relevant information in documents) and at the same time as a semantic-role-labeling task (i.e., identifying the slot fillers). We identify three sub-tasks: (1) identifying sentences describing relevant experiments, (2) identifying mentions of materials, values, and devices, and (3) recognizing mentions of slots and their values related to these experiments. We propose and compare several machine learning methods for the different sub-tasks, including bidirectional long-short term memory (BiLSTM) networks and BERT-based models. In our results, BERT-based models show superior performance. However, with increasing complexity of the task, it is beneficial to combine the two approaches.

With the aim of fostering research on challenging information extraction tasks in the scientific domain, we target the domain of SOFC-related experiments as a starting point. Our findings based on this sample use case are transferable to similar experimental domains, which we illustrate by applying our best model configurations to a previously existing related corpus (Mysore et al., 2019), achieving state-of-the-art results.

We sum up our contributions as follows:

- We develop an annotation scheme for marking information on materials-science experiments on scientific publications (Section 3).
- We provide a new corpus of 45 materials-science publications in the research area of SOFCs, manually annotated by domain experts for information on experimental settings and results (Section 4). Our corpus is publicly available.[1] Our inter-annotator agreement study provides evidence for high annotation quality (Section 5).
- We identify three sub-tasks of extracting experiment information and provide competitive baselines with state-of-the-art neural network approaches for them (Sections 4, 6, 7).
- We show the applicability of our findings to modeling the annotations of another materials-science corpus (Mysore et al., 2019, Section 7).

[1] Resources related to this paper can be found at: https://github.com/boschresearch/sofc-exp_textmining_resources

2 Related work

Information extraction for scientific publications. Recently, several studies addressed information extraction and knowledge base construction in the scientific domain (Augenstein et al., 2017; Luan et al., 2018; Jiang et al., 2019; Buscaldi et al., 2019). We also aim at knowledge base construction but target publications about materials science experiments, a domain understudied in NLP to date.

Information extraction for materials science. The work closest to ours is the one of Mysore et al. (2019) who annotate a corpus of 230 paragraphs describing synthesis procedures with operations and their arguments, e.g., "The resulting [solid products$_{Material}$] were ... [dried$_{Operation}$] at [120$_{Number}$][celsius$_{ConditionUnit}$] for [8$_{Number}$] [h$_{ConditionUnit}$]." Operation-evoking elements ("dried") are connected to their arguments via links, and with each other to indicate temporal sequence, thus resulting in graph structures similar to ours. Their annotation scheme comprises 21 entity types and 14 relation types such as *Participant-material*, *Apparatus-of* and *Descriptor-of*. Kononova et al. (2019) also retrieve synthesis procedures and extract recipes, though with a coarser-grained label set, focusing on different synthesis operation types. Weston et al. (2019) create a dataset for named entity recognition on abstracts of materials science publications. In contrast to our work, their label set (e.g., *Material*, *Application*, *Property*) is targeted to document indexing rather than information extraction. A notable difference to our work is that we perform full-text annotation while the aforementioned approaches annotate a pre-selected set of paragraphs (see also Kim et al., 2017).

Mysore et al. (2017) apply the generative model of Kiddon et al. (2015) to induce *action graphs* for synthesis procedures of materials from text. In Section 7.1, we implement a similar entity extraction system and also apply our algorithms to the dataset of Mysore et al. (2019). Tshitoyan et al. (2019) train word2vec (Mikolov et al., 2013) embeddings on materials science publications and show that they can be used for recommending materials for functional applications. Other works adapt the BERT model to clinical and biomedical domains

(Alsentzer et al., 2019; Sun and Yang, 2019), or generally to scientific text (Beltagy et al., 2019).

Neural entity tagging and slot filling. The neural-network based models we use for entity tagging and slot filling bear similarity to state-of-the-art models for named entity recognition (e.g., Huang et al., 2015; Lample et al., 2016; Panchendrarajan and Amaresan, 2018; Lange et al., 2019). Other related work exists in the area of semantic role labeling (e.g., Roth and Lapata, 2015; Kshirsagar et al., 2015; Hartmann et al., 2017; Adel et al., 2018; Swayamdipta et al., 2018).

3 Annotation Scheme

In this section, we describe our annotation scheme and guidelines for marking information on SOFC-related experiments in scientific publications.

3.1 Experiment-Describing Sentences

We treat the annotation task as identifying instances of a semantic frame (Fillmore, 1976) that represents SOFC-related experiments. We include (1) cases that introduce novel content; (2) descriptions of specific previous work; (3) general knowledge that one could find in a textbook or survey; and also (4) suggestions for future work.

We assume that a frame is introduced to the discourse by words that *evoke* the frame. While we allow any part-of-speech for such frame-evoking elements, in practice, our annotators marked almost only verbs, such as "test," "perform," and "report" with the type EXPERIMENT. In the remainder of this paper, we treat all sentences containing at least one such annotation as experiment-describing.

3.2 Entity Mention Types

In a second annotation layer, annotators mark spans with one of the following entity types. The annotations are marked only on experiment-describing sentences as well as several additional sentences selected by the annotator.

MATERIAL. We use the type MATERIAL to annotate text spans referring to materials or elements. They may be specified by a particular composition formula (e.g., "$La_{0.75}Sr_{0.25}Cr_{0.5}Mn_{0.5}O_3$") or just by a mention of the general class of materials, such as "oxides" or "hydrocarbons."[2]

[2]If the material is referenced by a common noun or by a pronoun and a more specific mention occurs earlier in the text, we indicate this coreference with the aim of facilitating oracle information extraction experiments in future work.

VALUE. We annotate numerical values and their respective units with the type VALUE.

In addition, we include specifications like "more than" or "between" in the annotation span (e.g., "above 750 °C," "1.0 W cm^{-2}").

DEVICE. This label is used to mark mentions of the type of device used in the fuel cell experiment (e.g., "IT-SOFC").

3.3 Experiment Slot Types

The above two steps of recognizing relevant sentences and marking coarse-grained entity types are in general applicable to a wide range of experiment types within the materials science domain. We now define a set of slot types particular to experiments on SOFCs. During annotation, we mark these slot types as links between the experiment-evoking phrase and the respective slot filler (entity mention), see Figure 1. As a result, experiment frames are represented by graphs rooted in the node corresponding to the frame-evoking element.

Our annotation scheme comprises 16 slot types relevant for SOFC experiments. Here we explain a few of these types for illustration. A full list of these slot types can be found in Supplementary Material Table 11; detailed explanations are given in the annotation guidelines published along with our corpus.

***AnodeMaterial*, *CathodeMaterial*:** These slots are used to mark the fuel cell's anode and cathode, respectively. Both are entity mentions of type MATERIAL. In some cases, simple surface information indicates that a material fulfills such a role. Other cases require specific domain knowledge and close attention to the context.

***FuelUsed*:** This slot type indicates the chemical composition or the class of a fuel or the oxidant species (indicated as a MATERIAL).

***PowerDensity*, *Resistance*, *WorkingTemperature*:** These slots are generally filled by mentions of type VALUE, i.e., a numerical value plus a unit. Our annotation guidelines give examples for relevant units and describe special cases. This enables any materials scientist, even if he/she is not an expert on SOFCs, to easily understand and apply our annotation guidelines.

Difficult cases. We also found sentences that include enumerations of experimental settings such

as in the following example: "It can be seen that the electrode polarization resistances in air are $0.027\,\Omega cm^2$, $0.11\,\Omega cm^2$, and $0.88\,\Omega cm^2$ at 800 °C, 700 °C and 600 °C, respectively."[3] We decided to simply link all slot fillers (the various resistance and temperature values) to the same frame-evoking element, leaving disentangling and grouping of this set of parameters to future work.

3.4 Links between Experiments

We instruct our annotators to always link slot fillers to the syntactically closest EXPERIMENT mention. If the description of an experiment spans more than one clause, we link the two relevant EXPERIMENTs using the relation *same_exp*. We use *exp_variation* to link experiments done on the same cell, but with slightly different operating conditions. The link type *exp_variation* can also relate two frame-evoking elements that refer to two measurements performed on different materials/cells, but in the same experimental conditions. In this case, the frame-evoking elements usually convey an idea of comparison, e.g., "increase" or "reach from ... to."

4 Corpus Statistics and Task Definitions

In this section, we describe our new corpus and propose a set of information extraction tasks that can be trained and evaluated using this dataset.

SOFC-Exp Corpus. Our corpus consists of 45 open-access scientific publications about SOFCs and related research, annotated by domain experts. For manual annotation, we use the InCeption annotation tool (Klie et al., 2018). Table 1 shows the key statistics for our corpus. Sentence segmentation was performed automatically.[4] As a preparation for experimenting with the data, we manually remove all sentences belonging to the Acknowledgment and References sections. We propose the experimental setting of using the training data in a 5-fold cross validation setting for development and tuning, and finally applying the model(s) to the independent test set.

Task definitions. Our rich graph-based annotation scheme allows for a number of information extraction tasks. In the scope of this paper, we address the following steps of (1) identifying sentences that describe SOFC-related experiments, (2) recognizing and typing relevant named entities, and (3) extracting slot fillers from these sentences. The originally annotated graph structures would also allow for modeling as relations or dependency structures. We leave this to future work.

	train	**test**
documents	34	11
sentences	7,630	1,836
avg. token/sentence	29.4	35.0
experiment-describing sentences	703	173
in %	9.2	9.4
sentences with entity mention annotations	853	210
entity mention annotations	4,037	1058
MATERIAL	1,530	329
VALUE	1,177	370
DEVICE	468	130
EXPERIMENT	862	229

Table 1: **SOFC-Exp corpus annotation statistics.**

The setup of our tasks is based on the assumption that in most cases, one sentence describes a single experiment. The validity of this assumption is supported by the observation that in almost all sentences containing more than one EXPERIMENT, experiment-evoking verbs actually describe variations of the same experiment. (For details on our analysis of links between experiments, see Supplementary Material Section B.) In our automatic modeling, we treat slot types as entity-types-in-context, which is a valid approximation for information extraction purposes. We leave the tasks of deciding whether two experiments are the same (*same_exp*) or whether they constitute a variation (*exp_variation*) to future work. While our dataset provides a good starting point, tackling these tasks will likely require collecting additional data.

5 Inter-annotator Agreement Study

We here present the results of our inter-annotator agreement study, which we perform in order to estimate the degree of reproducibility of our corpus and to put automatic modeling performance into perspective. Six documents (973 sentences) have been annotated independently both by our primary annotator, a graduate student of materials science, and a second annotator, who holds a Ph.D. in physics and is active in the field of materials science. The label distribution in this subset is similar to the one of our overall corpus, with each annotator choosing EXPERIMENT about 11.8% of the time.

[3] See [PMC4673446].

[4] InCeption uses Java's built-in sentence segmentation algorithm with US locale.

	P	R	F1	count
Experiment	81.1	75.6	78.3	119
No-Experiment	96.6	97.5	97.1	854

Table 2: **Inter-annotator agreement study.** Precision, recall and F1 for the subset of doubly-annotated documents. **count** refers to the number of mentions labeled with the respective type by our primary annotator.

Identification of experiment-describing sentences. Agreement on our first task, judging whether a sentence contains relevant experimental information, is 0.75 in terms of Cohen's κ (Cohen, 1968), indicating substantial agreement according to Landis and Koch (1977). The observed agreement, corresponding to accuracy, is 94.9%; expected agreement amounts to 79.2%. Table 2 shows precision, recall and F1 for the doubly-annotated subset, treating one annotator as the gold standard and the other one's labels as predicted. Our primary annotator identifies 119 out of 973 sentences as experiment-describing, our secondary annotator 111 sentences, with an overlap of 90 sentences. These statistics are helpful to gain further intuition of how well a human can reproduce another annotator's labels and can also be considered an upper bound for system performance.

Entity mention detection and type assignment. As mentioned above, relevant entity mentions and their types are only annotated for sentences containing experiment information and neighboring sentences. Therefore, we here compute agreement on the detection of entity mention and type assignment on the subset of 90 sentences that both annotators considered as containing experimental information. We again look at precision and recall of the annotators versus each other, see Table 3. The high precision indicates that our secondary annotator marks essentially the same mentions as our primary annotator, but recall suggests a few missing cases. The difference in marking EXPERIMENT can be explained by the fact that the primary annotator sometimes marks several verbs per sentence as experiment-evoking elements, connecting them with *same_exp* or *exp_variation*, while the secondary annotator links the mentions of relevant slots to the first experiment-evoking element (see also Supplementary Material Section B). Overall, the high agreement between domain expert annotators indicates high data quality.

	P	R	F1	count
EXPERIMENT	100.0	89.3	94.3	112
MATERIAL	100.0	92.1	95.9	190
VALUE	100.0	91.5	95.5	211
DEVICE	96.3	98.7	97.5	78

Table 3: **Inter-annotator agreement study.** Precision, recall and F1 for labeling entity types. **count** refers to the number of mentions labeled with the respective type by our primary annotator.

	F1	IAA count	train count
AnodeMaterial	72.0	13	280
CathodeMaterial	86.7	44	259
Device	95.0	71	381
ElectrolyteMaterial	85.7	48	219
FuelUsed	85.7	11	159
InterlayerMaterial	71.8	25	51
OpenCircuitVoltage	90.0	10	44
PowerDensity	92.0	47	175
Resistance	100.0	26	136
Thickness	92.6	27	83
WorkingTemperature	96.5	73	414

Table 4: **Inter-annotator agreement study.** F1 was computed for the two annotators vs. each other on the set of **experiment slots**; **IAA count** refers to the number of mentions labeled with the respective type by our primary annotator in the inter-annotator agreement study (IAA).

Identifying experiment slot fillers. We compute agreement on the task of identifying the slots of an experiment frame filled by the mentions in a sentence on the subset of sentences that both annotators marked as experiment-describing. Slot fillers are the dependents of the respective edges starting at the experiment-evoking element. Table 4 shows F1 scores for the most frequent ones among those categories. See Supplementary Material Section C for all slot types. Overall, our agreement study provides support for the high quality of our annotation scheme and validates the annotated dataset.

6 Modeling

In this section, we describe a set of neural-network based model architectures for tackling the various information extraction tasks described in Section 4.

Experiment detection. The task of experiment detection can be modeled as a binary sentence classification problem. It can also be conceived as a retrieval task, selecting sentences as candidates for experiment frame extraction. We implement a bidirectional long short-term memory (**BiLSTM**)

model with attention for the task of experiment sentence detection. Each input token is represented by a concatenation of several pretrained word embeddings, each of which is fine-tuned during training. We use the Google News word2vec embeddings (Mikolov et al., 2013), domain-specific word2vec embeddings (mat2vec, Tshitoyan et al., 2019, see also Section 2), subword embeddings based on byte-pair encoding (bpe, Heinzerling and Strube, 2018), BERT (Devlin et al., 2019), and SciBERT (Beltagy et al., 2019) embeddings. For BERT and SciBERT, we take the embeddings of the first word piece as token representation. The embeddings are fed into a BiLSTM model followed by an attention layer that computes a vector for the whole sentence. Finally, a softmax layer decides whether the sentence contains an experiment.

In addition, we fine-tune the original (uncased) **BERT** (Devlin et al., 2019) as well as **SciBERT** (Beltagy et al., 2019) models on our dataset. SciBERT was trained on a large corpus of scientific text. We use the implementation of the BERT sentence classifier by Wolf et al. (2019) that uses the CLS token of BERT as input to the classification layer.[5]

Finally, we compare the neural network models with traditional classification models, namely a support vector machine (**SVM**) and a **logistic regression** classifier. For both models, we use the following set of input features: bag-of-words vectors indicating which 1- to 4-grams and part-of-speech tags occur in the sentence.[6]

Entity mention extraction. For entity and concept extraction, we use a sequence-tagging approach similar to (Huang et al., 2015; Lample et al., 2016), namely a **BiLSTM model**. We use the same input representation (stacked embeddings) as above, which are fed into a BiLSTM. The subsequent conditional random field (CRF, Lafferty et al., 2001) output layer extracts the most probable label sequence. To cope with multi-token entities, we convert the labels into BIO format.

We also fine-tune the original **BERT** and **SciBERT** sequence tagging models on this task. Since we use BIO labels, we extend it with a CRF output layer to enable it to correctly label multi-token mentions and to enable it to learn transition scores between labels. As a non-neural baseline, we train a **CRF** model using the token, its lemma, part-of-speech tag and mat2vec embedding as features.[7]

Slot filling. As described in Section 4, we approach the slot filler extraction task as fine-grained entity-typing-in-context, assuming that each sentence represents a single experiment frame. We use the same sequence tagging architectures as above for tagging the tokens of each experiment-describing sentence with the set of slot types (see Table 11). Future work may contrast this sequence tagging baseline with graph-induction based frame extraction.

7 Experiments

In this section, we present the experimental results for detecting experiment-describing sentences, entity mention extraction and experiment slot identification. For tokenization, we employ ChemDataExtractor,[8] which is optimized for dealing with chemical formulas and unit mentions.

We tune our models in a 5-fold cross-validation setting. We also report the mean and standard deviation across those folds as development results. For the test set, we report the macro-average of the scores obtained when applying each of the five models to the test set. To put model performance in relation to human agreement, we report the corresponding statistics obtained from our inter-annotator agreement study (Section 5). Note that these numbers are based on a subset of the data and are hence not directly comparable.

Hyperparameters and training. The BiLSTM models are trained with the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1e-3. For fine-tuning the original BERT models, we follow the configuration published by Wolf et al. (2019) and use AdamW (Loshchilov and Hutter, 2019) as optimizer and a learning rate of 4e-7 for sentence classification and 1e-5 for sequence tagging. When adding BERT tokens to the BiLSTM, we also use the AdamW optimizer for the whole model and learning rates of 4e-7 or 1e-5 for the BERT part and 1e-3 for the remainder. For regularization, we employ early stopping on the development set. We use a stacked BiLSTM with two hidden layers and 500 hidden units for all tasks with the exception of the experiment sentence de-

[5] https://github.com/huggingface/transformers

[6] We use sklearn, https://scikit-learn.org.

[7] We use sklearn-pycrfsuite, https://pypi.org/project/sklearn-pycrfsuite.

[8] http://chemdataextractor.org

Model	dev F1	test P	test R	test F1
RBF SVM	54.2+/-3.7	64.6	54.9	59.4
Logistic Regression	53.0+/-4.2	**68.2**	50.9	58.3
BiLSTM mat2vec	49.9+/-3.1	49.6	69.4	57.8
BiLSTM word2vec	52.3+/-4.6	51.1	65.3	57.4
+ mat2vec	55.9+/-4.2	52.0	59.0	55.3
+ bpe	58.6+/-3.0	58.9	64.7	61.7
+ BERT-base	66.8+/-4.9	60.2	71.7	65.4
+ SciBERT	67.9+/-4.0	58.6	74.6	65.6
BiLSTM BERT-base	64.7+/-4.6	63.7	69.9	66.7
BiLSTM SciBERT	**68.1**+/-3.7	60.2	73.4	66.1
BERT-base	66.0+/-4.6	58.6	71.1	64.2
SciBERT	67.9+/-4.0	60.8	74.6	67.0
BERT-large	64.3+/-4.3	63.1	**75.1**	**68.6**
humans	*78.3*	*81.1*	*75.6*	*78.3*

Table 5: **Experiments: identifying experiment-describing sentences.** P, R and F1 for experiment-describing sentences. With the exception of SVM, we downsample the non-experiment-describing sentences of the training set by 0.3.

tection task, where we found one BiLSTM layer to work best. The attention layer of the sentence detection model has a hidden size of 100.

Experiment sentence detection. Table 5 shows our results on the detection of experiment-describing sentences. The neural models with byte-pair encoding embeddings or BERT clearly outperform the SVM and logistic regression models. Within the neural models, BERT and SciBERT add the most value, both when using their embeddings as another input to the BiLSTM and when fine-tuning the original BERT models. Note that even the general-domain BERT is strong enough to cope with non-standard domains. Nevertheless, models based on SciBERT outperform BERT-based models, indicating that in-domain information is indeed beneficial. For performance reasons, we use BERT-base in our experiments, but for the sake of completeness, we also run BERT-large for the task of detecting experiment sentences. Because it did not outperform BERT-base in our cross-validation based development setting, we did not further experiment with BERT-large. However, we found that it resulted in the best F1-score achieved on our test set. In general, SciBERT-based models provide very good performance and seem most robust across dev and test sets. Overall, achieving F1-scores around 67.0-68.6, such a retrieval model may already be useful in production. However, there certainly is room for improvement.

Model	EXP.	MAT.	VAL.	DEV.	**avg.**
CRF	61.4	42.3	73.6	64.1	60.3
BiLSTM mat2vec	47.1	52.4	60.9	46.1	51.6
BiLSTM word2vec	55.8	58.6	59.1	51.7	56.3
+mat2vec	57.9	75.2	64.3	61.5	64.7
+bpe	63.3	81.6	68.0	68.1	70.2
+BERT-base	76.0	88.1	72.9	81.5	79.7
+SciBERT	76.9	89.8	74.1	85.2	**81.5**
BiLSTM BERT-base	75.4	87.6	72.6	80.8	79.1
BiLSTM SciBERT	77.1	**89.9**	72.1	**85.7**	81.2
BERT-base	81.8	70.6	88.2	73.1	78.4
SciBERT	**84.5**	77.0	**91.6**	72.7	**81.5**
humans	*94.3*	*95.9*	*95.5*	*97.5*	*95.8*

Table 6: **Experiments: entity mention detection and typing.** Results on test set (experiment-describing sentences only) in terms of F1, rightmost column shows the macro-average.

Entity mention extraction. Table 6 provides our results on entity mention detection and typing. Models are trained and results are reported on the subset of sentences marked as experiment-describing in the gold standard, amounting to 4,590 entity mentions in total.[9] The CRF baseline achieves comparable or better results than the BiLSTM with word2vec and/or mat2vec embeddings. However, adding subword-based embeddings (bpe and/or BERT) significantly increases performance of the BiLSTM, indicating that there are many rare words. Again, the best results are obtained when using BERT or SciBERT embeddings or when using the original SciBERT model. It is relatively easy for all model variants to recognize VALUE as these mentions usually consist of a number and unit which the model can easily memorize. Recognizing the types MATERIAL and DEVICE, in contrast, is harder and may profit from using gazetteer-based extensions.

Experiment slot filling. Table 7 shows the macro-average F1 scores for our different models on the slot identification task.[10] As for entity typing, we train and evaluate our model on the subset of sentences marked as experiment-describing, which contain 4,263 slot instances. Again, the CRF baseline outperforms the BiLSTM when using only

[9]The SOFC-Exp gold standard marks all entity mentions that correspond to one of the four relevant types occurring in these sentences, regardless of whether the mention fills a slot in an experiment or not.

[10]We evaluate on the 16 slot types as listed in Table 11. When training our model, we use the additional types *experiment_evoking_word* and *Thickness*, which are not frame slots but related annotations present in our data, see guidelines.

Model	dev	test
CRF	45.3+/-5.6	41.3
BiLSTM mat2vec	25.9+/-11.2	22.5
BiLSTM word2vec	27.5+/-9.0	27.0
+ mat2vec	43.0+/-11.5	34.9
+ bpe	50.2+/-11.8	38.9
+ BERT-base	64.6+/-12.8	54.2
+ SciBERT	67.1+/-13.3	59.7
BiLSTM BERT-base	63.3+/-12.9	57.4
BiLSTM SciBERT	**67.8**+/-12.9	**62.6**
BERT-base	63.4+/-13.8	54.9
SciBERT	65.6+/-13.2	56.4
humans	*83.4*	

Table 7: **Experiments: slot identification**. Model comparison in terms of macro F1.

mat2vec and/or word2vec embeddings. The addition of BERT or SciBERT embeddings improves performance. However, on this task, the BiLSTM model with (Sci)BERT embeddings outperforms the fine-tuned original (Sci)BERT model. Compared to the other two tasks, this task requires more complex reasoning and has a larger number of possible output classes. We assume that in such a setting, adding more abstraction power to the model (in the form of a BiLSTM) leads to better results.

For a more detailed analysis, Table 8 shows the slot-wise results for the non-neural CRF baseline and the model that performs best on the development set: BiLSTM with SciBERT embeddings. As in the case of entity mention detection, the models do well for the categories that consist of numeric mentions plus particular units. In general, model performance is also tied to the frequency of the slot types in the dataset. Recognizing the role a material plays in an experiment (e.g., *AnodeMaterial* vs. *CathodeMaterial*) remains challenging, possibly requiring background domain knowledge. This type of information is often not stated explicitly in the sentence, but introduced earlier in the discourse and would hence require document-level modeling.

7.1 Entity Extraction Evaluation on the Synthesis Procedures Dataset

As described in Section 2, the data set curated by Mysore et al. (2019) contains 230 synthesis procedures annotated with entity type information.[11] We apply our models to this entity extraction task in order to estimate the degree of transferability of our findings to similar data sets. To the best of our knowledge, there have not yet been any publications on the automatic modeling of this data set. We hence compare to the previous work of Mysore et al. (2017), who perform action graph induction on a similar data set.[12] Our implementation of BiLSTM-CRF mat2vec+word2vec roughly corresponds to their BiLSTM-CRF system.

Table 9 shows the performance of our models when trained and evaluated on the synthesis procedures dataset. Detailed scores by entity type can be found in the Supplementary Material. We chose to use the data split suggested by the authors for the NER task, using 200 documents for training, and 15 documents for each dev and test set. Among the non-BERT-based systems, the BiLSTM variant using both mat2vec and word2vec performs best, indicating that the two pre-trained embeddings contain complementary information with regard to this

	CRF	BiLSTM SciBERT	count
AnodeMaterial	25.0	19.0	280
CathodeMaterial	11.8	28.9	259
Device	59.3	67.6	381
ElectrolyteMaterial	20.0	47.2	219
FuelUsed	45.9	55.5	159
InterlayerMaterial	0.0	10.7	51
OpenCircuitVoltage	43.5	84.3	44
PowerDensity	69.0	97.6	175
Resistance	64.5	93.9	136
WorkingTemperature	72.5	90.3	414

Table 8: **Experiments: slot identification.** Results in terms of F1 on the test set, BiLSTM results averaged across 5 models.

Model	micro-avg. F1
DCNN (Mysore et al., 2017)	77.5
BiLSTM-CRF (Mysore et al., 2017)	77.6
BiLSTM mat2vec	73.9
BiLSTM word2vec	76.4
+ mat2vec	83.5
BERT-base	85.5
SciBERT	87.2
BiLSTM BERT-base	89.3
BiLSTM SciBERT	90.7
BiLSTM + all (with BERT-base)	89.3
BiLSTM + all (with SciBERT)	**92.2**

Table 9: **Experiments: modeling mention types in synthesis procedure data set.** Results from Mysore et al. (2017) are not directly comparable to ours as they are based on a slightly different data set; our BiLSTM mat2vec+word2vec roughly corresponds to their BiLSTM-CRF model.

[11] See https://github.com/olivettigroup/annotated-materials-syntheses

[12] According to correspondence with authors.

task. The best performance is reached by the BiLSTM model including word2vec, mat2vec, bpe and SciBERT embeddings, with 92.2 micro-average F1 providing a strong baseline for future work.

8 Conclusion

We have presented a new dataset for information extraction in the materials science domain consisting of 45 open-access scientific articles related to solid oxide fuel cells. Our detailed corpus and inter-annotator agreement studies highlight the complexity of the task and verify the high annotation quality. Based on the annotated structures, we suggest three information extraction tasks: the detection of experiment-describing sentences, entity mention recognition and typing, and experiment slot filling. We have presented various strong baselines for them, generally finding that BERT-based models outperform other model variants. While some categories remain challenging, overall, our models show solid performance and thus prove that this type of data modeling is feasible and can lead to systems that are applicable in production settings. Along with this paper, we make the annotation guidelines and the annotated data freely available.

Outlook. In Section 7.1, we have shown that our findings generalize well by applying model architectures developed on our corpus to another dataset. A natural next step is to combine the datasets in a multi-task setting to investigate to what extent models can profit from combining the information annotated in the respective datasets. Further research will investigate the joint modeling of entity extraction, typing and experiment frame recognition. In addition, there are also further natural language processing tasks that can be researched using our dataset. They include the detection of events and sub-events when regarding the experiment-descriptions as events, and a more linguistically motivated evaluation of the frame-semantic approach to experiment descriptions in text, e.g., moving away from the one-experiment-per-sentence and one-sentence-per-experiment assumptions and modeling the graph-based structures as annotated.

Acknowledgments

We thank Jannik Strötgen, Felix Hildebrand, Dragan Milchevski and everyone else involved in the Bosch MatKB project for their support of this research. We also thank Stefan Grünewald, Sherry Tan, and the anonymous reviewers for their insightful comments related to this paper.

References

Heike Adel, Laura Ana Maria Bostan, Sean Papay, Sebastian Padó, and Roman Klinger. 2018. DERE: A task and domain-independent slot filling framework for declarative relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–47, Brussels, Belgium. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sören Auer, Viktor Kovtun, Manuel Prinz, Anna Kasprzik, Markus Stocker, and Maria Esther Vidal. 2018. Towards a knowledge graph for science. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, WIMS 18, New York, NY, USA. Association for Computing Machinery.

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. 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.

Davide Buscaldi, Danilo Dess, Enrico Motta, Francesco Osborne, and Diego Reforgiato Recupero. 2019. Mining scholarly data for fine-grained knowledge graph construction. In *Proceedings of the Workshop on Deep Learning for Knowledge Graphs*, pages 21–30.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, and Junichi Tsujii, editors. 2017. *BioNLP 2017*. Association for Computational Linguistics, Vancouver, Canada,.

Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. 2018. *Proceedings of the BioNLP 2018 workshop*. Association for Computational Linguistics, Melbourne, Australia.

Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. 2019. *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy.

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*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Charles J. Fillmore. 1976. Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 471–482, Valencia, Spain. Association for Computational Linguistics.

Benjamin Heinzerling and Michael Strube. 2018. BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. In *CorRR*, volume abs/1508.01991.

Tianwen Jiang, Tong Zhao, Bing Qin, Ting Liu, Nitesh V. Chawla, and Meng Jiang. 2019. The role of "condition": A novel scientific knowledge graph representation and construction model. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1634–1642, Anchorage, AK, USA. ACM.

Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 982–992, Lisbon, Portugal. Association for Computational Linguistics.

Edward Kim, Kevin Huang, Alex Tomala, Sara Matthews, Emma Strubell, Adam Saunders, Andrew McCallum, and Elsa Olivetti. 2017. Machine-learned and codified synthesis parameters of oxide materials. *Scientific data*, 4:170127.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data*, 6(1):1–11.

Meghana Kshirsagar, Sam Thomson, Nathan Schneider, Jaime Carbonell, Noah A. Smith, and Chris Dyer. 2015. Frame-semantic role labeling with heterogeneous annotations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 218–224, Beijing, China. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lukas Lange, Heike Adel, and Jannik Strötgen. 2019. NLNDE: The neither-language-nor-domain-experts way of spanish medical document de-identification. In *Proceedings of the Iberian Languages Evaluation Forum*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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.

Matteo Manica, Christoph Auer, Valéry Weber, Federico Zipoli, Michele Dolfi, Peter W. J. Staar, Teodoro Laino, Costas Bekas, Akihiro Fujita, Hiroki Toda, Shuichi Hirose, and Yasumitsu Orii. 2019. An information extraction and knowledge graph platform for accelerating biochemical discoveries. *CoRR*, abs/1907.08400.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

David Mrdjenovich, Matthew K. Horton, Joseph H. Montoya, Christian M. Legaspi, Shyam Dwaraknath, Vahe Tshitoyan, Anubhav Jain, and Kristin A. Persson. 2020. propnet: A knowledge graph for materials science. *Matter*, 2(2):464 – 480.

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, Florence, Italy. Association for Computational Linguistics.

Sheshera Mysore, Edward H 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 *NIPS Workshop on Machine Learning for Molecules and Materials*.

Vivi Nastase, Benjamin Roth, Laura Dietz, and Andrew McCallum, editors. 2019. *Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications*. Association for Computational Linguistics, Minneapolis, Minnesota.

Rrubaa Panchendrarajan and Aravindh Amaresan. 2018. Bidirectional LSTM-CRF for named entity recognition. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Michael Roth and Mirella Lapata. 2015. Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Jannik Strötgen, Trung-Kien Tran, Annemarie Friedrich, Dragan Milchevski, Federico Tomazic, Anika Marusczyk, Heike Adel, Daria Stepanova, Felix Hildebrand, and Evgeny Kharlamov. 2019. Towards the bosch materials science knowledge base. In *Proceedings of the ISWC 2019 Satellite Tracks (Posters & Demonstrations, Industry, and Outrageous Ideas)*, volume 2456 of *CEUR Workshop Proceedings*, pages 323–324, Auckland, New Zealand. CEUR-WS.org.

Cong Sun and Zhihao Yang. 2019. Transfer learning in biomedical named entity recognition: An evaluation of BERT in the PharmaCoNER task. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 100–104, Hong Kong, China. Association for Computational Linguistics.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782, Brussels, Belgium. Association for Computational Linguistics.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571:95 – 98.

L. Weston, V. Tshitoyan, J. Dagdelen, O. Kononova, A. Trewartha, K. A. Persson, G. Ceder, and A. Jain. 2019. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of Chemical Information and Modeling*, 59(9):3692–3702. PMID: 31361962.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Supplementary Material

A Background on Solid Oxide Fuel Cells

A fuel cell is an electrochemical device that generates electricity exploiting the chemical reaction of a fuel (usually hydrogen) with an oxidant (usually air). The reactions take place on two electrodes, the cathode and the anode, while the circuit is closed by an electrolyte material that only allows the transfer of charged atoms (see Figure 2). Fuel

cells that use a solid oxide as electrolyte (Solid Oxide Fuel Cells or SOFCs) are very efficient and cost-effective, but can only operate at high temperatures (500-1000C), which can cause long start-up times and fast degradation. SOFCs can be used as stationary stand-alone devices, to produce clean power for residential or industrial purposes, or integrated with other power generation systems to increase the overall efficiency.

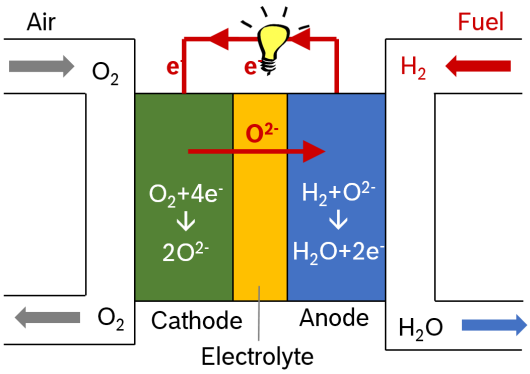

Figure 2: Solid Oxide Fuel Cell schema.

B Data Analysis: Between-Experiment Links

As stated in Section 3, we instructed annotators to mark the closest experiment-evoking word as EXPERIMENT and link the respective slot arguments to this mention. In addition, the EXPERIMENT annotations could then be linked either by *same_exp* or *exp_variation* links. Table 10 shows some statistics on the number of EXPERIMENT annotations per sentence and how often the primary annotator actually made use of the possibility to link experiments. In the training data, out of 703 sentences describing experiments, 135 contain more than one experiment-evoking word, with 114 sentences containing two, 18 sentences containing three, and 3 sentences containing four EXPERIMENT annotations (see Table 10). In the 114 sentences containing two experiment annotations, only in 2 sentences, the EXPERIMENTs were not linked to any others. Upon being shown these cases, our primary annotator judged that one of them should actually have been linked.

Next, we analyze the number of cross-sentence links. In the training data, there are 256 *same_exp* and 93 *exp_variation* links, of which 138 and 57 cross sentence-boundaries respectively. Cross-sentence links between experiment-evoking words and slot fillers rarely occur in our dataset (only 13 out of 2,540 times).

# EXPERIMENT per sentence	**1**	**2**	**3**	**4**
# sentences	568	114	18	3
# *same_exp*	0	82	28	7
# *exp_variation*	0	27	8	1
# sent. with 'unlinked' exp.	-	2	1	0

Table 10: **Data analysis.** Number of EXPERIMENT annotations per sentence, and counts of links between them (within sentence). Training set: 703 experiment-describing sentences.

C Inter-annotator Agrement Study: further statistics

Table 11 shows the full set of statistics for the experiment slot agreement.

D Additional Experimental Results

In the following tables, we give detailed statistics for the experiments described in the main paper.

Table 12 reports full statistics for the task of identifying experiment-describing sentences, including precision and recall in the dev setting.

Table 13 reports F1 per entity type for the dev setting including standard deviations.

Table 14 reports F1 per entity type/slot for the synthesis procedures dataset (Mysore et al., 2019).

	agreement study			IAA	train
	P	**R**	**F1**	**count**	**count**
AnodeMaterial	75.0	69.2	72.0	13	280
CathodeMaterial	84.8	88.6	86.7	44	259
Conductivity	-	-	-	-	55
CurrentDensity	100.0	60.0	75.0	5	65
DegradationRate	100.0	100.0	100.0	2	19
Device	97.1	93.0	95.0	71	381
ElectrolyteMaterial	78.9	93.8	85.7	48	219
FuelUsed	90.0	81.8	85.7	11	159
InterlayerMaterial	100.0	56.0	71.8	25	51
OpenCircuitVoltage	90.0	90.0	90.0	10	44
PowerDensity	100.0	85.1	92.0	47	175
Resistance	100.0	100.0	100.0	26	136
SupportMaterial	75.0	37.5	50.0	8	106
TimeOfOperation	83.3	100.0	90.9	5	47
Voltage	100.0	33.3	50.0	6	35
WorkingTemperature	98.6	94.5	96.5	73	414

Table 11: **Inter-annotator agreement study.** Precision, recall and F1 scores of the two annotators vs. each other on the set of **slots**. **IAA count** refers to the number of mentions labeled with the respective type by our primary annotator in the 6 documents of the inter-annotator agreement study. **train count** refers to the number of instances in the training set. (*Conductivity* has been added to the set of slots only after conducting the inter-annotator agreement study.)

	dev (5-fold cv)			test		
Model	P	R	F1	P	R	F1
RBF SVM	66.4	46.1	54.2$_{+/-3.7}$	64.6	54.9	59.4
Logistic Regression	**72.7**	41.9	53.0$_{+/-4.2}$	**68.2**	50.9	58.3
BiLSTM mat2vec	46.3	55.6	49.9$_{+/-3.1}$	49.6	69.4	57.8
BiLSTM word2vec	50.0	56.1	52.3$_{+/-4.6}$	51.1	65.3	57.4
+ mat2vec	59.8	53.6	55.9$_{+/-4.2}$	52.0	59.0	55.3
+ bpe	62.2	56.4	58.6$_{+/-3.0}$	58.9	64.7	61.7
+ BERT	66.1	67.8	66.8$_{+/-4.9}$	60.2	71.7	65.4
+SciBERT	68.6	68.0	**68.1**$_{+/-3.7}$	60.2	73.4	66.1
BiLSTM BERT	65.5	64.2	64.7$_{+/-4.6}$	63.7	69.9	66.7
BiLSTM SciBERT	67.1	69.1	67.9$_{+/-4.0}$	58.6	74.6	65.6
BERT-base	64.0	68.2	66.0$_{+/-4.6}$	58.6	71.1	64.2
BERT-large	61.8	68.9	64.3$_{+/-4.6}$	63.1	**75.1**	**68.6**
SciBERT	66.0	**70.2**	67.9$_{+/-4.0}$	60.8	74.6	67.0
humans (on agreement data)	*80.4*	*77.6*	*78.9*	*80.4*	*77.6*	*78.9*

Table 12: **Experiments: Identifying experiment sentences.** P, R and F1 for experiment-describing sentences. With the exception of SVM, we downsample the non-experiment-describing sentences by 0.3.

Model	EXPERIMENT	MATERIAL	VALUE	DEVICE	**macro-avg.**	EXPERIMENT	MATERIAL	VALUE	DEVICE	**macro-avg.**
CRF	66.5+/-3.5	47.0+/-9.1	73.0+/-6.4	56.2+/-10.0	60.7+/-4.5	61.4	42.3	73.6	64.1	60.3
BiLSTM mat2vec	52.9+/-3.4	55.3+/-2.0	47.9+/-6.3	53.2+/-1.9	52.3+/-3.4	47.1	52.4	60.9	46.1	51.6
+ BERT	80.3+/-3.2	87.7+/-3.3	76.8+/-5.3	81.9+/-5.5	**81.7**+/-4.3	74.3	87.9	71.0	80.7	78.5
BiLSTM word2vec	62.3+/-3.0	61.6+/-2.1	52.1+/-5.2	59.5+/-1.0	58.9+/-2.8	55.8	58.6	59.1	51.7	56.3
+mat2vec	65.8+/-4.2	78.4+/-1.6	61.9+/-8.2	69.6+/-4.0	68.9+/-4.5	57.9	75.2	64.3	61.5	64.7
+bpe	69.2+/-5.8	82.3+/-1.9	60.1+/-11.2	73.4+/-4.7	71.2+/-5.9	63.3	81.6	68.0	68.1	70.2
+BERT	80.0+/-3.4	87.9+/-2.8	74.4+/-5.6	80.7+/-3.9	80.8+/-3.9	76.0	88.1	72.9	81.5	79.7
+SciBERT	81.4+/-1.6	**89.4**+/-2.4	73.8+/-8.7	82.0+/-4.3	**81.7**+/-4.2	76.9	89.8	74.1	85.2	**81.5**
BiLSTM BERT	79.6+/-2.4	87.6+/-2.4	72.0+/-7.5	80.5+/-5.1	79.9+/-4.3	75.4	87.6	72.6	80.8	79.1
BiLSTM SciBERT	80.5+/-1.2	**89.4**+/-2.8	73.0+/-9.4	**82.3**+/-3.5	81.3+/-4.2	77.1	**89.9**	72.1	**85.7**	81.2
BERT-base	**85.4**+/-2.8	73.7+/-7.2	90.0+/-2.1	68.3+/-3.7	79.3+/-3.9	**81.8**	70.6	88.2	73.1	78.4
SciBERT	84.5+/-3.0	77.0+/-7.4	**91.6**+/-2.8	72.7+/-2.1	81.5+/-3.8	81.2	75.3	**91.9**	73.2	80.4
humans	*94.3*	*95.9*	*95.5*	*97.5*	*95.8*	*94.3*	*95.9*	*95.5*	*97.5*	*95.8*

Table 13: **Experiments: entity mention extraction and labeling.** Results on 5-fold cross validation for dev and test set (experiment-describing sentences only) in terms of F1.

Entity Types	Mysore et al. (2017)	BiLSTM w2v+m2v	BiLSTM + all (SciBERT)
Amount-Unit	83.5	93.5	**95.8**
Brand	-	67.9	**83.3**
Condition-Misc	74.6	85.1	**88.9**
Condition-Unit	94.5	**97.2**	95.0
Material	80.2	84.0	**92.3**
Material-Descriptor*	62.0	65.5	**88.5**
Nonrecipe-Material	-	45.8	**80.0**
Number	91.9	93.4	**98.4**
Operation	82.8	93.5	**98.1**
Synthesis-Apparatus	-	63.9	**81.3**

Table 14: **Experiments: Modeling mention types in synthesis procedure data, most frequent entity types.** Results in terms of F1. Results from Mysore et al. (2017) are not directly comparable. *Type called Descriptor in their paper.