Extracting Material Property Measurement Data from Scientific Articles

Gihan Panapitiya*, **Fred Parks, Jonathan Sepulveda** and **Emily Saldanha***
Pacific Northwest National Laboratory, Richland, WA, 99354, USA
{gihan.panapitiya, fred.parks, jonathan.sepulveda, emily.saldanha}@pnnl.gov

Abstract

Machine learning-based prediction of material properties is often hampered by the lack of sufficiently large training datasets. The majority of such measurement data is embedded in scientific literature and the ability to automatically extract these data is essential to support the development of reliable property prediction methods. In this work, we describe a methodology for an automatic property extraction framework using material solubility as the target property. We create an annotated dataset containing tags for solubility-related entities using a combination of regular expressions and manual tagging. We then compare five entity recognition models leveraging both token-level and span-level architectures on the task of classifying solute names, solubility values, and solubility units. Additionally, we explore a novel pretraining approach that leverages automated chemical name and quantity extraction tools to generate large datasets that do not rely on intensive manual effort. Finally, we perform an analysis to identify the causes of classification errors.

1 Introduction

While the application of machine learning methods for material property prediction holds great promise for material discovery and design across a broad range of applications, such methods are often hampered by a lack of sufficiently large and diverse training datasets. Typically, relevant measurements and information exist only in unstructured formats such as the published scientific literature and are not available in aggregated and standardized databases. The ability to automatically extract, process, and analyze large sets of material property data would represent a significant capability for the advancement of material design efforts.

The development of predictive models for the solubility of organic molecules is one such use case that would support a wide range of application areas including pharmaceutical, environmental, and energy storage applications. For example, molecular solubility is a key performance driver for redox flow battery (RFB) technologies which rely on energy-bearing redox active molecules that are dissolved in a liquid electrolyte. The solubility of a molecule determines its maximal concentration in the electrolyte and the resulting energy density of the system. To support the development of predictive models for the discovery and design of new materials for these batteries, a comprehensive solubility measurement database is required.

The ability to automatically parse the scientific literature for existing and newly published solubility measurements would provide a significant acceleration to our ability to enlarge existing datasets, as well as keep predictive models up-to-date with the newest data. However, the extraction of numerical data from the materials science and chemistry literature presents several key challenges. Material property measurements are often highly sensitive to the specifics of the experimental conditions. Additionally, in comparison with tasks for general information extraction, this task requires the ability of models to target specific measurement types and distinguish them from other measurements which may be expressed using similar language.

In this paper, we make several key contributions. First, we collect a corpus of solubility-related sentences from the scientific literature and perform manual tagging to annotate key components of the solubility measurements including the solute, the solubility value, and the solubility unit. Secondly, we apply and compare several entity-extraction deep learning models to the problem of automatic extraction of solubility data. Additionally, we develop and explore several possible methods for the pretraining of quantitative measurement extraction models. Overall, our best performing models achieves F1 scores of 0.75, 0.79, and 0.9 on the extraction of solutes, values, and units. Finally, we perform detailed performance and error analysis to

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5393–5402
November 7–11, 2021.

provide insights into the strengths and weaknesses of the current models and to identify directions for future improvements to the methodology.

2 Related Work

The majority of existing solubility datasets which have been employed for solubility prediction are small, with typically a few thousand data points (Boobier et al., 2017; Llinàs et al., 2008; Delaney, 2004; Huuskonen et al., 1997; Tang et al., 2020), with recent datasets starting to reach the level of 10,000 data points (Cui et al., 2020). There are no previously existing datasets linking the solubility measurement values to their source text in publications to support the development of information extraction efforts.

Several annotated datasets have been developed to support efforts for more general scientific information extraction including the ScienceIE dataset with task, process, and material annotations in the material science, physics, and computer science domains (Augenstein et al., 2017), the SciERC dataset with entity and relation annotations in the computer science domain (Luan et al., 2018), annotations of conditional facts from the life science and biomedical domain (Jiang et al., 2019), and a dataset with annotations of process, method, material, and data across ten scientific domains (Brack et al., 2020).

Prior work on information extraction in the materials science and chemistry domains has focused both the development of general tools such as ChemDataExtractor (Swain and Cole, 2016), which leverages conditional random field (CRF) models, custom dictionaries, and rule-based grammers for information extraction, and the development of models for specific extraction tasks. Some examples include extraction of zeolite synthesis information using regular expressions (Jensen et al., 2019), extraction of nanomaterial composition and morphology using unsupervised methodology on TF-IDF features and synthesis protocols using a sentence-level logistic regression classifier (Hiszpanski et al., 2020), extraction of synthesis conditions for metal oxides using supervised classification of parsed noun phrases (Kim et al., 2017), tagging of material science-related named entities using bidirectional LSTMs (Weston et al., 2019), and construction of a knowledge base related to solid oxide fuel cells through the development of entity and relation extraction models (Friedrich et al., 2020).

The majority of prior work has focused on the extraction of fixed scientific entities, such as material names or methodologies, and their relationships. However, less work has been performed to develop methods targeted toward the extraction of numerical and quantitative data as is needed for the extraction of experimental measurements like solubility. Several prior task-specific models include extraction of Curie and Néel magnetic phase transition temperatures using ChemDataExtractor combined with semi-supervised relation extraction (Cole et al., 2018) and the use of a rule-based approach to extract numerical data from electronic medical records (Cai et al., 2019).

3 Approach

3.1 Data with solubility-related tags

We aim to develop models which can extract quantities related to solubility measurements from scientific text. To support this task we develop a dataset of annotated sentences which contain solubility information including the solute, the solubility value, and the solubility unit. To extract sentences for annotation, we rely on the PubMed Central (PMC) open access text mining dataset[1] with the full text of 2.75 million articles and the S2ORC database (Lo et al., 2020) with the full text of 8.1 million papers. We first perform sentence segmentation of the articles using the *sent_tokenize* method from NLTK[2].

We next filter the sentences to those that may contain a solubility measurement by filtering those sentences that contain the word "solubility of" and at least one digit. Upon manual inspection of the sentences, we found occasional failure modes of the sentence segmentation leading to extractions which actually span multiple sentences. To minimize the presence of such sentences, we filter out extremely long sentence extractions containing 77 tokens or more. This results in 19,963 and 76,333 sentences from the PMC and S2ORC databases respectively.

We pursue two strategies for tagging these sentences: manual tagging and automated tagging using ChemDataExtractor (Swain and Cole, 2016). The manual tagging generates higher fidelity annotations while the automated tagging generates potentially noisier annotations at the benefit of increasing the data size with reduced manual anno-

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/textmining/
[2] https://www.nltk.org/

```
p = (solute + (I('has')|I('exhibits')|I('exhibited')) + Optional(I('the')|I('a'))\
    + Optional(I('notable')|I('high')|(I('relatively')+I('low')).add_action(join) )+Optional(I('water'))\
    + I('solubility')\
    + Optional(I('of')) )(u'prefix')
```

Sentence => In water, AgSCN has a solubility of 1.68 × 10 −4 g L −1 which is the lowest of the suitable silver salts.
Extracted entities => `[{'value': '1.68 × 10 − 4', 'units': 'g L − 1', 'solute': 'AgSCN'}]`

```
p = (I('solubility').hide() + I('of').hide() + solute + I('is').hide() + R('\S+').hide()\
     + R('\S+').hide() + R('\S+').hide() )(u'prefix')
```

Sentence => However, the water solubility of galangin is extremely low (~14.3 µg/mL), thus limiting its application.
Extracted entities => `[{'value': '~ 14.3', 'units': 'µg / mL', 'solute': 'galangin'}]`

Figure 1: Examples of ChemDataExtractor-based regular expressions and resulting extractions.

tation workload. For the automated tagging, custom regular expression were developed for ChemDataExtractor in order to extract solute names (CHEM), solubility values (VALUE) and solubility units (UNIT). Figure 1 shows several examples of the regular expressions and the entities extracted using them.

To capture more complex sentence structures and to establish a higher-fidelity labelled dataset, we rely on manual tagging. Some of the labelled data was initially collected using Excel spreadsheets before transitioning to the use of the Prodigy software[3] to generate additional labels. The tags we used for manual tagging are solute name (CHEM), solvent name (SLVN), solubility value (VALUE), solubility unit (UNIT), temperature unit (TEMPU), temperature value (TEMPV), pressure unit (PRSU), and pressure value (PRSV). In this paper, we focus on the extraction of the solute name, solubility value, and solubility unit, but future work will target the extraction of more complete experimental details captured in the additional annotations.

The dataset collected using manual and regular expression based taggers contains a total of 5,337 sentences with 4,478 sentences from manual tagging and 859 from extractions using regular expressions. This shows that relatively fewer sentences use the types of formulaic language that can be captured by regular expression extractions and that more sophisticated extraction techniques are needed to capture the variability in language used for expressing these measurements. Overall, 21% of the sentences in the dataset were found to contain at least one solubility value and 20% were found to contain both a solubility value and a solute name. The distribution of the number of entities of each type per sentence can be seen in Figure 2. We find that the majority of sentences contain a single measurement result only, with only a small proportion containing more than two such measurements in the same sentence. The sentences have a total of 1,652, 1,437, and 1,876 tags of solute names, solubility values, and solubility units respectively. The number of unique solute names is 1,243 and the number of unique units is 141. The most common units are "%", "mg/mL", "mg/ml", "mM", and "μg/mL". The tagged sentences were split into train, validation and testing sets containing 80%, 10% and 10% of the data respectively.

3.2 Pre-training data

Due to the time and effort intensiveness of the manual labelling of the candidate sentences, generating a large set of labelled sentences for training is a challenge. To support the training of measurement extraction tasks, we explore several transfer learning strategies that rely on pretraining tasks which leverage only automatically generated data labels. Specifically, we perform pretraining on the tasks of detecting general chemical names, measurement values, and measurement units, rather than tagging only those related to solubility. After this pre-training, we fine-tune the models on the solubility extraction task. Because the pretrained model has learned to identify chemical names, measurement values, and measurement units in general, the task of the solubility tagging fine-tuning will be to learn to select and tag which chemical names, measurement values, and measurement units are related to the desired quantity of interest.

To generate the pretraining labels, we rely on several previously existing information extraction tools, namely grobid-quantities (gro, 2015–2021) and ChemDataExtractor (Swain and Cole, 2016). The grobid-quantities package[4] leverages a linear CRF model to identify all expressions of measurements within the text including both values and units. Meanwhile, ChemDataExtractor leverages

[3] https://prodi.gy/

[4] https://github.com/kermitt2/grobid-quantities

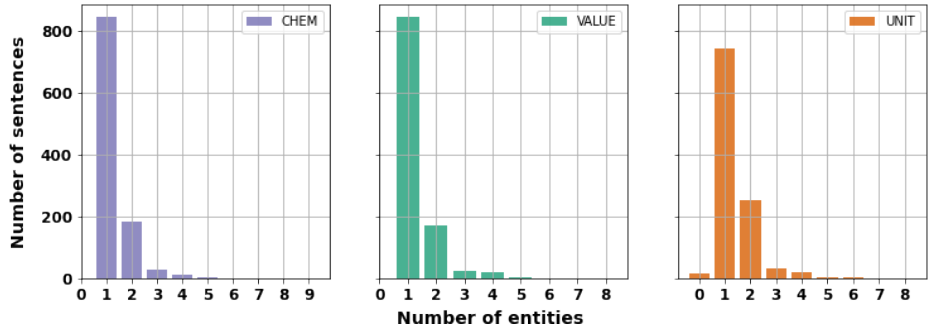

Figure 2: Distributions of entity counts within sentences in the solubility dataset for those sentences that contain at least one solute names and solubility value.

Chitneni et al reported that sulpiride was reasonably stable in Krebs - Ringer buffer solutions throughout the duration of their experiment (2 hours). 1 Meanwhile , Kohri et al20 reported that the aqueous solubility of sulpiride was 800 µg / mL , so a lower concentration of sulpiride (200 µg / 0.5 mL) was used in their intestinal permeability studies .

The molecular weight and solubility of BPA is 228 g / mol and 120 mg / l , respectively [10,11] .

Figure 3: Examples of entities tagged by the pretraining tagging (general chemical names and measurements) and the regex (upper) and manual (lower) solubility taggers (solutes and solubility measurements). Entities detected by the pretraining tagging are indicated by grey horizontal lines. Entities detected by both the pretraining and solubility taggers are colored boxes (blue = solute, green = value, brown = units).

CRF-based named entity recognizer in combination with a dictionary approach to identify chemical names. We use these tools to annotate a dataset with potentially noisy, low-fidelity annotations of chemical names, values, and units. We explore two variants of the pretraining task, one which uses all three tag types (chemical names, values, and units) and one which just uses values and units. In Figure 3, we show examples of entities tagged by the pretraining taggers and the solubility taggers.

To generate the pretraining dataset, we start from the PMC dataset sentences which contain the word "solubility". Next, we select the sentences (composed of more than 10 and less than 350 characters) that contain a chemical name and a value based on the extractions from grobid-quantities and ChemDataExtractor. The sentences that contain less than 5 and greater than 76 tokens were removed. The resulting number of sentences were 2,737,620 from which 100,000 random sentences were selected for the final pretraining dataset. This dataset contains a total of 184,129, 165,136, and 110,395 entities tagged as chemical names, values, and units respectively.

3.3 Models

To extract the relevant solubility data from the sentences, we explore several modeling architectures for token-level and span-level tagging of the CHEM, VALUE, and UNIT tags. For token-level classification, we explore several variations of the HuggingFace implementations of the BERT architecture (Devlin et al., 2019), leveraging the pretrained weights from *bert-base-cased* and SciBERT (Beltagy et al., 2019) (*scibert-scivocab-cased*). For these models, we perform simple token-level annotation prediction by sending the output of the BERT model through a linear layer with dropout to predict the CHEM, VALUE, UNIT, and O (other) tags. Secondly, we experiment with the addition of a conditional random field (CRF) layer to the BERT models to enable joint predictions of the tags across the sentence. In addition to the BERT-based approaches, we experiment with the span-based SpERT architecture (Eberts and Ulges, 2020), which has shown good performance on entity and relation extraction from scientific text. SpERT is a relation classifier which also contains a span classification layer. For this work, we removed the final relation classification layer and used the span classification layer to predict token class. All five models contain $\sim 110M$ parameters.

Table 1: Solubility tag prediction performance. Results in bold are best across all model configurations.

	Model	CHEM			VALUE			UNIT		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
No Pretrain	BERT	0.76	0.68	0.72	0.67	0.82	0.74	0.82	0.92	0.87
	SciBERT	**0.8**	0.66	0.73	0.77	0.55	0.64	0.76	0.97	0.85
	BERT+CRF	0.78	0.72	**0.75**	0.7	0.81	0.75	0.82	0.94	0.88
	SciBERT+CRF	0.79	0.72	**0.75**	0.75	**0.82**	**0.79**	**0.87**	0.92	**0.9**
	SPERT	**0.8**	0.63	0.71	0.67	0.65	0.66	0.78	0.88	0.83
C+V+U Pretrain	BERT	0.77	0.59	0.66	0.72	0.77	0.74	0.8	0.93	0.86
	SciBERT	0.65	0.7	0.68	0.76	0.46	0.57	0.68	**0.97**	0.8
	BERT+CRF	0.76	0.59	0.66	0.7	0.75	0.72	0.79	0.93	0.85
	SciBERT+CRF	0.6	0.63	0.62	**0.83**	0.4	0.54	0.67	**0.97**	0.79
	SPERT	0.5	**0.74**	0.59	0.53	0.77	0.63	0.64	0.96	0.77
V+U Pretrain	BERT	0.75	0.64	0.69	0.72	0.75	0.74	0.79	0.89	0.84
	SciBERT	0.77	0.64	0.7	0.78	0.76	0.77	0.8	0.9	0.85
	BERT+CRF	0.76	0.59	0.66	0.69	0.79	0.74	0.78	0.89	0.83
	SciBERT+CRF	0.68	0.62	0.65	0.8	0.36	0.5	0.67	**0.97**	0.79
	SPERT	0.53	0.53	0.53	0.55	0.63	0.59	0.64	0.83	0.72

For the BERT/SciBERT models, a maximum sequence length of 128 tokens was considered. A dropout rate of 0.3 is applied to the BERT/SciBERT output before linearly transforming it to predict the token labels. We leverage a learning rate scheduler that linearly decreases the learning rate from $3e{-}5$ to 0 and compare performance for 10, 12, 15, 20 or 25 epochs with 12 epochs performing best according to the validation F1 score. Models were trained using training and validation batch sizes of 32 and 8. All the hyperparameters used in the SpERT model are the default parameters used by the authors in the code in the GitHub repository[5].

4 Results

We summarize the results of our experiments in Table 1. For the models that are trained on the solubility data from scratch without leveraging pretraining, we find that the SciBERT+CRF model performs the best overall, achieving the highest F1 score across all three tag types. For both the BERT and SciBERT weights, the addition of the CRF layer improves extraction performance across all three tag types. UNIT tags are easiest to identify, which can be expected as the number of unique units in the dataset is small (141 unique values) allowing the model to memorize the types of unit tags to expect.

In Table 2 we show the confusion matrix for predictions made by the best-performing SciBERT+CRF model. Typically, the model does not confuse tags of different types but instead makes errors regarding whether or not a tag exists for the token. This makes sense as the three tag types are likely to be quite different from each other.

4.1 Impact of Pretraining

We find that the performance on the pretraining task is very high across all five model architectures, with F1 scores of around 0.89, 0.93, and 0.96 for CHEM, VALUE, and UNIT tags respectively. This shows that the models are successfully able to reproduce the automated labels for general measurement data and chemical names within scientific text. However, after fine-tuning these models on the solubility extraction task, we find that our two pretraining strategies have not been effective in improving the accuracy. Instead, we find that the pretrained models typically have reduced performance on the detection of all three tag types compared with the models trained from scratch on solubility data alone. Of the two pretraining strategies, we find that pretraining using only VALUE and UNIT tags seems to be favorable over using CHEM, VALUE and UNIT tags. When CHEM, VALUE and UNIT tags are used in the pretraining, BERT models without the CRF layer produce best fine tuning results. For the VALUE and UNIT only pretraining, the SciBERT model without the CRF layer produces the best results.

The failure of the pretraining task to improve the ability to extract solubility measurements, may be because the model struggles to adapt from the task of general value extraction to the task related to the measurement of interest. Since we find that inclusion of the chemical name tags in the pretraining task is more detrimental to performance, we hypothesize that the mismatch between traditional

[5]https://github.com/lavis-nlp/spert

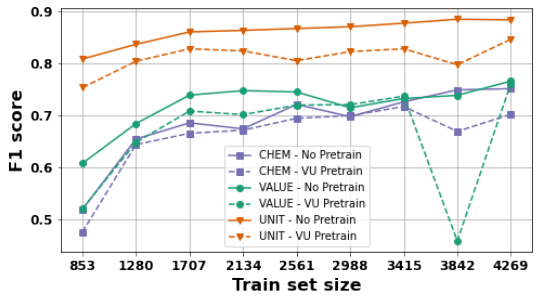

Figure 4: Effect of the training set size on model performance.

chemical names and the tokens which are actually tagged as solutes may be causing part of the issue. The automated taggers used for the pretraining data will tend to identify "official" chemical names, such as o-chloroacetophenone, 6-Gingerol, and Atrazine, while the actual phrasing of solubility measurement data often incorporates solute names that include abbreviations or shortened versions, such as DETC, M5bG7, and Se-L-M.

Manual inspection of the sentences which the SciBERT+CRF model predicted correctly but SciBERT with V+U pretraining predicted incorrectly revealed that the pretrained model seems to struggle when the solubility value is not immediately preceded by 'solubility of'. Examples of such alternate prefixing terms that lead to errors are "(<", "is extremely low (" and "is up to".

4.2 Impact of Data Size

We next evaluate the dependence of the model performance on the availability of training data. In Figure 4, we show the F1 scores using samples of different sizes from the full training set. We find that performance significantly increases as the dataset size increases from 853 sentences to 1707 sentences. After this point, performance improvements are slower with the addition of new data. These trends show that substantial amounts of additional data would likely be need to achieve significant improvements in the extraction performance. We also find that the use of the pretraining strategies are not effective compared with training from scratch, even when very small amounts of training data are available for the target task.

4.3 Error Analysis

We perform extensive analysis of the errors made by the model to understand the types of inputs on which the model performs well and the types on which it performs poorly. The results in this section are derived from the best performing SciBERT + CRF model.

Table 2: Confusion matrix for SciBERT+CRF model output.

		Predicted			
		CHEM	**VALUE**	**UNIT**	**O**
True	**CHEM**	406	3	0	154
	VALUE	1	381	1	79
	UNIT	0	0	437	36
	O	110	121	65	21870

4.3.1 Qualitative Error Exploration

We first perform qualitative observation of the prediction errors made by the model by exploring the properties of individual tokens for which the model made incorrect predictions. We observed the 154 cases where our model confused CHEM tags for O tags. 13 of these tags are "-" characters and 34 are associated with words that contain a "-" character as part of the chemical name (e.g., "DNM-2", "n-octanol"). 19 of the 154 tags contain one or more digits.

Out of the 79 tags where the model confused a VALUE tag for an O tag, 16 are "." characters. These 79 also contain the words, "greater","least","less","than","to","up", and "water". This shows that the model has difficulty recognizing non-digit tokens, which are relatively less common within the training data. Only 10% of labelled VALUE tokens are something other than a digit or a ".".

Out of the 36 unit tags predicted as O tags 13 are "%" characters. This is not surprising as "%", despite being a common solubility unit, is often used for non-solubility values. Units "mm" and "wt" have been confused for O tags four and three times respectively. Even though in solubility measurements "mm" corresponds to milli-moles, "mm" is also the abbreviation for millimeters, which is not a solubility unit.

Of the 110 tokens wrongly predicted to be CHEM tags, 6 of them are "-" characters which are part of chemical names. In one of the cases, the chemical name is not a solute name and in three of the cases, even though the chemical names are solutes, the regex tagger has not been able to correctly tag them as such. In the other two cases, the manual tagger has missed to tag the solute name. This shows that the model has picked up on correct solute name extractions despite some level of

Table 3: Extraction performance based on subsets of sentences that contain annotated tags of each type with certain characteristics.

Description	Recall	F1	Sentences
CHEM			
All upper case	0.70	0.79	43
Some lower case	0.69	0.76	87
Has a digit	0.64	0.75	25
No digit	0.69	0.76	105
Has "-"	0.68	0.77	20
No "-"	0.71	0.78	110
VALUE			
All digits or "."	0.79	0.83	95
Some non digits/"."	0.88	0.92	18
Has "."	0.83	0.87	65
No "."	0.74	0.77	50
UNIT			
Common units (>90 in train)	0.91	0.93	64
Less common units	0.93	0.93	67

noise in the labelled dataset. The words "raw" and "pure" are also among the tags that got confused for CHEM tags. These words may be inconsistently tagged as part of the description of a solute material.

Of the 121 tokens that were wrongly predicted to be VALUE tags, 70% are digits and 21% are "." characters. Among the other characters that got wrongly classified as VALUE are "<", "±", "×", "–", and "10^{-}". This shows that the model is correctly identifying numerical quantities as the VALUE tags, but sometimes identifies values that are not solubility measurements.

Out of the 65 tokens wrongly classified as UNIT tags, the most frequently confused tokens are "/", "ml", "mg", "%", and "mm". All of these are commonly found in solubility units. However, they can also be part of the concentration measurements that often occur in the same sentence containing solubility measurements.

4.3.2 Quantitative Error Exploration

Based on the common patterns observed in the qualitative analysis of errors made by the model, we next compare the model performance on different subsets of the test data based on the characteristics of the ground truth tags. In Table 3 we compare the model performance on sentences in which the tags have certain properties. For CHEM tags, we find that different properties of the tags have only small impact on the performance. We find that inclusion of digit characters slightly decreases performance, use of upper-casing slightly increases performance, and inclusion of a "-" character slightly reduces

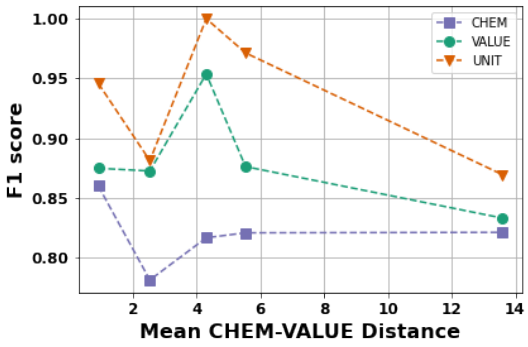

Figure 5: F1 score versus the distance between CHEM and VALUE tags in the sentence.

performance. For the extraction of VALUE tags, we find that the models perform well even when the values contain characters that are not strictly numeric. We also find that the model performs better for floating point numbers that contain a "." character. While we might expect the UNIT extraction to rely on memorization of commonly occurring units, we find that the performance is not significantly reduced for the less commonly occurring units.

We next observe how the larger structure of the sentences beyond the properties of the tagged tokens themselves affect the accuracy of the model. For example, how does the proximity of the solute name to the solubility value in the sentence affect accuracy? In Figure 5, we show the F1 score for different groups of sentences binned by the number of tokens separating the CHEM and VALUE tags for sentences which have exactly one CHEM and one VALUE entity. We find that model is able to achieve similar performance even when the solute name is separated from the corresponding solubility value by many tokens.

We also explore how the phrasing of the sentence affects the model performance. We analysed whether the extraction of solute names depends on whether or not the name is preceded by the phrase "solubility of" (Table 4). We find that employing this standard phrasing significantly improves the ability of the model to correctly detect the solute name. To further explore the impact of phrasing, we list the model performance on sentences tagged using regular expressions versus those tagged manually in Table 5. We expect that sentences tagged using regular expressions will employ more standard and formulaic phrasing than the manually tagged sentences which required human expertise to parse. However, we also expect that the regex

Table 4: Impact of preceding text on CHEM prediction

CHEM preceded by	Recall	F1	Sentences
"solubility of"	0.78	0.83	101
some other text	0.58	0.67	26

Table 5: F1 scores of manually versus regex tagged sentences

Tagger	CHEM	VALUE	UNIT	Sentences
Regex	0.82	0.72	0.93	72
Manual	0.68	0.82	0.86	462

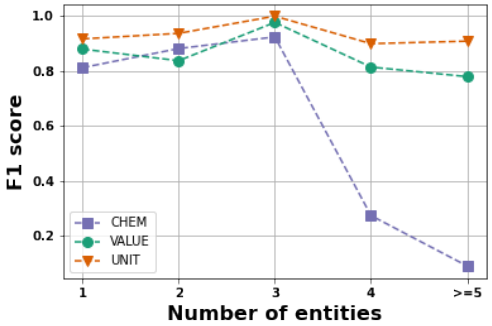

Figure 6: How the prediction accuracy depends on the number of chemical names, solubility values and units within the sentence.

tagging will have higher susceptibility to tagging errors. Interestingly, we find that the relative performance difference depends on the tag type. The model performs better on the manually tagged sentences for VALUE tags but performs better on the regex tagged sentences for CHEM and UNIT tags.

Finally, we observe the effect of the number of unique entities within a sentence on the success of the model extractions. While we showed in Figure 2 that the majority of sentences contain only a single solubility measurement, there are more complex sentences within the dataset that express multiple solubility measurements under different conditions or for different solutes. In Figure 6 we show the performance of the model as the number of tags of each type changes. We find that there is not a strong effect here, but that performance on solute name extraction significantly declines when the number of solutes in the sentence reaches 4.

5 Conclusions

We demonstrate the performance of several different token-level and span-level extraction architectures on the novel task of solubility measurement data extraction from scientific literature. We find that a token-level extraction model leveraging pretraining SciBERT weights combined with a CRF layer achieves the best performance on this task. We explore the impact of pretraining on a large set of general value extraction data but find that it reduces the model performance compared with training from scratch. We also find that the model performance has a weak sensitivity to increases in the training data size. This points to the necessity of developing alternative modeling approaches or more effective pretraining tasks to improve performance in this domain due to the challenge of collecting sufficiently large labelled datasets.

We perform error analysis to understand the strengths and weaknesses of the extraction model. We find that the model performance is robust to several variations that had the potential to make the extraction task more challenging. The model performs well at unit extraction even when less commonly used units are employed. The model also appears to perform well even when solute names and values are separated by many tokens in the sentence. However, we do observe that the model performs poorly on solute name extraction when the sentence does not employ the standard phrasing "solubility of". This points to the need for continuing to supplement the dataset with sentences that contain more diverse and complex phrasing.

While we have demonstrated promising performance on this novel scientific measurement extraction task, there are several key directions for future development of these efforts. First, while the pretraining strategies employed in this work did not provide a benefit to model performance, further exploration is needed to understand the types of pretraining tasks that may best support measurement extraction efforts. Next, to fully contextualize extracted solubility measurements it will be necessary to extract the relevant experimental details (solvent, temperature, pH, etc.). We have collected manual annotation data to support this expansion of the task. Additionally, the current approach can detect the existence of several solubility measurements within a sentence but cannot determine which solubility values are associated with which solutes or which experimental conditions. Relation extraction capabilities will be needed to associate solubility values with the correct solutes, solvents, and conditions.

Acknowledgements

This work was supported by Energy Storage Materials Initiative (ESMI), which is a Laboratory Directed Research and Development Project at Pacific Northwest National Laboratory (PNNL). PNNL is a multiprogram national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract no. DE- AC05-76RL01830

References

2015–2021. grobid-quantities. https://github.com/kermitt2/grobid-quantities.

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.

Samuel Boobier, Anne Osbourn, and John B. O. Mitchell. 2017. Can human experts predict solubility better than computers? *Journal of Cheminformatics*, 9(1):63.

Arthur Brack, Jennifer D'Souza, Anett Hoppe, Sören Auer, and Ralph Ewerth. 2020. Domain-independent extraction of scientific concepts from research articles. In *European Conference on Information Retrieval*, pages 251–266. Springer.

Tianrun Cai, Luwan Zhang, Nicole Yang, Kanako K Kumamaru, Frank J Rybicki, Tianxi Cai, and Katherine P Liao. 2019. Extraction of emr numerical data: an efficient and generalizable tool to extend clinical research. *BMC medical informatics and decision making*, 19(1):1–7.

Jacqueline M Cole et al. 2018. Auto-generated materials database of curie and néel temperatures via semi-supervised relationship extraction. *Scientific data*, 5.

Qiuji Cui, Shuai Lu, Bingwei Ni, Xian Zeng, Ying Tan, Ya Dong Chen, and Hongping Zhao. 2020. Improved prediction of aqueous solubility of novel compounds by going deeper with deep learning. *Frontiers in Oncology*, 10:121.

John S. Delaney. 2004. ESOL: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005. PMID: 15154768.

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.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2006–2013. IOS Press.

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.

Anna M Hiszpanski, Brian Gallagher, Karthik Chellappan, Peggy Li, Shusen Liu, Hyojin Kim, Jinkyu Han, Bhavya Kailkhura, David J Buttler, and Thomas Yong-Jin Han. 2020. Nanomaterial synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge. *Journal of chemical information and modeling*, 60(6):2876–2887.

Jarmo Huuskonen, Marja Salo, and Jyrki Taskinen. 1997. Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *Journal of Pharmaceutical Sciences*, 86(4):450 – 454.

Zach Jensen, Edward Kim, Soonhyoung Kwon, Terry ZH Gani, Yuriy Roman-Leshkov, Manuel Moliner, Avelino Corma, and Elsa Olivetti. 2019. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS central science*, 5(5):892–899.

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.

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

Antonio Llinàs, Robert C. Glen, and Jonathan M. Goodman. 2008. Solubility challenge: Can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *Journal of Chemical Information and Modeling*, 48(7):1289–1303. PMID: 18624401.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

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.

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. PMID: 27669338.

Bowen Tang, Skyler T Kramer, Meijuan Fang, Yingkun Qiu, Zhen Wu, and Dong Xu. 2020. A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility. *Journal of Cheminformatics*, 12(1):1–9.

Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav 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.