Advances in scientific literature mining for interpreting materials characterization

IOP Publishing Mach. Learn.: Sci. Technol. 2 (2021) 045007 https://doi.org/10.1088/2632-2153/abf751

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
31 December 2020

REVISED
25 March 2021

ACCEPTED FOR PUBLICATION
13 April 2021

PUBLISHED
15 July 2021

PAPER

Advances in scientific literature mining for interpreting materials characterization

Gilchan Park* **and Line Pouchard***

Computational Science Initiative, Brookhaven National Laboratory, Upton, NY 11973, United States of America
* Authors to whom any correspondence should be addressed.

E-mail: gpark@bnl.gov **and** pouchard@bnl.gov

Keywords: XAS, x-ray spectroscopy, figure explanatory text, scientific literature mining, text extraction, deep learning, natural language processing

Abstract

Using synchrotron light sources, such as the National Synchrotron Light Source II at Brookhaven National Laboratory, scientists in fields as diverse as physics, biology, and materials science, identify the atomic structure, chemical composition, or other important properties of varied specimens. x-ray spectroscopy from light sources is particularly valuable for materials research with vast information available about reference spectra in the scientific literature. However, as the technique is applicable to many science domains, searching for information about select x-ray spectroscopy spectra is impeded by the sheer number of publications. Moreover, useful information about the context of an experiment or figures presented in papers can be buried among the details, which takes time to assess. This work presents a scientific literature mining system that supports data acquisition, information extraction, and user interaction for referencing x-ray spectra identification and spectral interpretation. The goal is to provide efficient access to useful spectral data to researchers who may spend only a few days at a synchrotron light source. With this system, users browse a classification tree for papers arranged according to x-ray spectroscopic methods, chemical elements, and x-ray absorption spectroscopy edges. Relevant figures are extracted with sentences from the paper that explain them, known as 'figure explanatory text.' Notably, this system focuses on semantic aspects (logical analysis) to find figure explanatory text using deep contextualized word embeddings techniques and contains an interface to obtain labeled data from domain experts that is used to evaluate and improve the model.

1. Introduction

Text mining is the process of automatically extracting meaningful information from large volumes of unstructured text data, e.g. information that can be directly presented to users or put into structured formats for populating databases. The National Synchrotron Light Source II (NSLS-II) [1] at Brookhaven National Laboratory (BNL) offers spectroscopy capabilities amongst other techniques that enable scientific discoveries in e.g. clean and affordable energy, high-temperature superconductivity, and macromolecular crystallography. Users from academia and industry come to a beamline at NSLS-II with samples for characterization of chemical bonding and electron energy band structure with the guidance of beamline scientists. During the short period of time users spend at the beamline for their experiments (typically several 4 h sessions over 48 h), they compare their sample spectra to those of well-characterized reference samples and adjust their measurement parameters. In addition, users often need to look up additional figures of spectra published in the scientific literature, this is usually conducted manually. While reference spectra are known prior to an experiment, a new sample characterized at the beamline often displays characteristics that users did not anticipate and for which finding comparable spectra in the literature provides additional insights. NSLS-II users have complex information needs that include finding figures in the scientific literature representing the characteristics of chemical elements and compounds measured by techniques similar to the ones used in a particular experiment. These needs cannot be answered by popular search

engines such as Google Scholar or Web of Science that can retrieve thousands of papers for a single query—most irrelevant to the specific needs of these users. Furthermore, the specific details of the characterization technique used, the different frequency bands of a measurement, the energies, and numerous other details that make the sample and experiment comparable are buried in the text of a relevant paper. The difficulties for users to find pertinent information in the literature are compounded by two facts: (a) x-ray spectroscopy is a technique widely used for many different purposes in multiple disciplines, and (b) older scientific literature can still be relevant to the experiment at hand. As a result, finding comparable spectra in the vast literature at users' disposal to collect evidence and develop methods for experiments at the beamline is inefficient and haphazard.

This article describes a pilot system for scientific literature mining that uses deep learning techniques adapted to natural language processing (NLP) and provides direction for answering users' complex information needs. The system targets x-ray absorption spectroscopy (XAS) techniques as XAS is one of the most commonly performed x-ray experiments at NSLS-II. XAS is widely used for investigating structures of samples from various angles in many scientific disciplines, including chemistry, materials science, physics, biology, and biomedicine. When searching through published papers, experimenters first look for XAS spectra and textual information in the paper that explains the figure. While search systems that specialize in scientific literature mining, such as Semantic Scholar[1], present users with figures and captions extracted from papers, the captions typically lack enough information for users to compare with their experimental results [2, 3]. There is a growing need for a service where practitioners can find XAS spectra and relevant text from articles during the short period of time they spend at the beamline. Our pilot Text Mining system aims to address this need. This paper makes the following contributions:

- We present a curated, pre-processed collection of articles relevant to XAS techniques and make it available to facility users.
- We developed a portal and user interface that allows users to perform deep searches through these papers and extract figures of spectra.
- Users can also browse through the classified collection by transition metals and XAS edges.
- We present users with 'figure explanatory text'—additional details from the text of the paper relevant to figures, a new feature in scientific literature mining. Results are these sentences in the text of the paper that provide details about figures in addition to captions.
- We trained a model to find relevant information to XAS spectra in articles based on semantic analysis of texts that goes beyond literal comparison and measures sentence similarities by meaning to obtain these results.
- We integrate domain expertise in the trained model at several levels, (a) to guide the original data acquisition, (b) to provide informal feedback on the usability of the interface, and (c) next
- Most importantly, we incorporated into our user-facing system a mechanism for domain experts to rank results. The collected label data will improve accuracy of the resulting sentences related to the figures, when its size is sufficient.
- As there is no direct way to compare with other search engines since our system adds new features (the figure-explanatory text), we compare the method to random results.

Our research hypothesis is that (a) there is a lot of useful details that can explain a spectra contained in the text of a paper but do not appear in a caption; and (b) that text mining for this purpose must focus on the semantic aspects of language rather than keyword comparison and sentence similarity. This has been done before in materials characterization.

The text mining system deployed at NSLS-II includes a classification system for XAS-related publications and a text extraction model for explaining figures using NLP techniques. Both are deployed in a user-friendly search interface. Our system builds a data collection, extracts pertinent information from the scientific publications related to XAS, and presents it to users in a web portal. XAS spectra (figures) contained in relevant papers are extracted and presented with captions and snippets of text from the paper that are relevant to each figure. One of the obstacles in using machine learning techniques for science is the absence of labeled data for model evaluation. Our system affords users the ability to quickly rate the relevance of each text segment extracted by the model to the displayed figure. This helps with the production of labeled data and model fine-tuning. The system supports expandability, adaptability, and data integration for the development of future applications to other topics in scientific literature mining. This article is divided into the following sections: section 2 presents an overview of the system's architecture, and section 3 details the

[1] www.semanticscholar.org

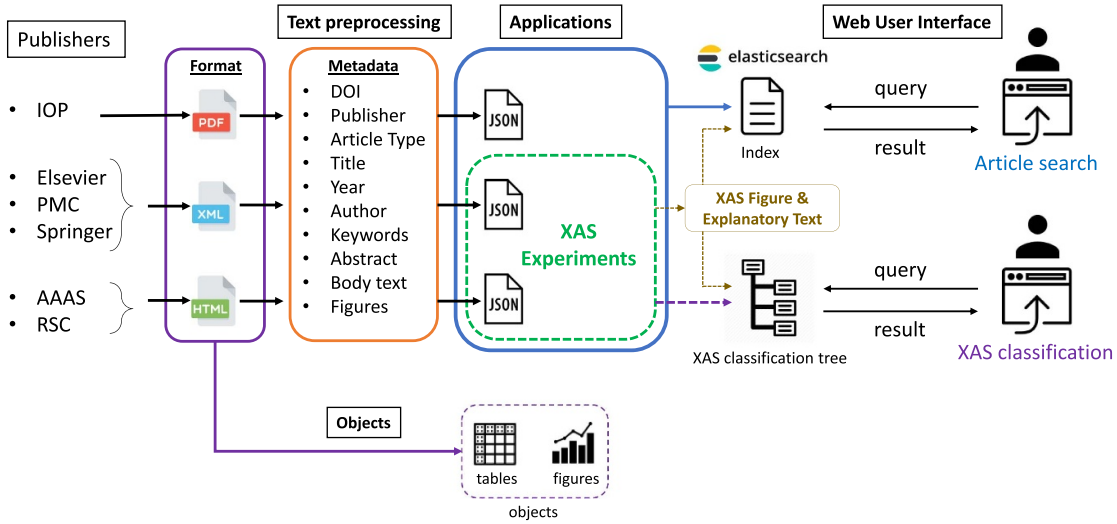

Figure 1. Architecture design of the text mining system.

deep learning methods used to extract explanatory text from each article, the datasets, and the feedback mechanisms.

2. Overview and related work

2.1. System architecture

The scientific literature mining system is composed of three major parts: article database construction, NLP model creation, and a web user interface (figure 1). The article database is built by collecting articles and supplementary materials such as figures from major publishers and open-source digital repositories (described in section 3). Text preprocessing includes extracting values using the metadata provided by publishers. Heterogeneous formats are unified using an XML parser and a PDF parsing tool [4] that provides JSON outputs. While parsing articles in semi-structured formats (XML; HTML) generates consistent outputs, parsing PDF articles is much more complicated. Moreover, PDF parsing results are not consistent or reliable [5]. Until reliable PDF parsing results are obtainable, we depend on XML/HTML articles for article collection. Unlike other PDF articles, PDF articles from IOP (Institute of Physics) have additional metadata files that can be used for searching. Thus, we include IOP PDF articles in the article search, but not in XAS figure analysis. With the structured JSON corpus, NLP models are designed for the desired purpose. Our system adopts several deep learning methods that recently have made significant progress in big data analysis and NLP [6, 7] and exploits existing domain-specific text mining tools, such as ChemDataExtractor [8, 9]. For the XAS corpus, users interact with developed NLP models via a web interface in two ways: they can find figures using the article classification and text from the article explaining these figures, hereafter noted as 'figure explanatory text.' We also have deployed Elasticsearch [10] to improve search functionality and performance over the entire collection with filtering that allows users to display figures and explanatory text without browsing the classification. The NLP models are shown in figure 1 in the XAS Experiments box and described in section 3.

2.2. Related work for XAS data and materials informatics

Out of many methods for materials design and discovery, XAS is a popular technique for materials characterization, and spectra yielded by XAS experiments provide the fingerprint of the specific chemical environment for elements including the local atomic structure, electronic properties, and coordination environment information [11]. When analyzing XAS images, if a library of representative reference spectra is available, researchers can perform spectra comparison analyses that help identify an unknown and gain in-depth structural insights [12]. However, it is challenging to acquire reliable reference spectra. XAS experiments are conducted at synchrotron radiation facilities, and the experimental data is stored at facility proprietary databases with various retention policies and usually not open to the public. Although many efforts are ongoing to create an open access XAS data from facilities [13] and standardize hydrogenous XAS data formats across facilities [14], open reference databases such as the electron energy-loss spectroscopy (EELS) database [15] still contain limited volume of spectra (e.g. only 17 number of K-edge spectra in EELS),

and more agreements on unifying data formats are needed to create a central framework for an international XAS database [16]. To enlarge the size of XAS reference data, the researchers [17, 18] created the computed reference XAS spectra based on theoretical calculations, called XASdb, that contains more than 800k K-edge XANES for over 40k materials. XAS experiments are expensive in terms of time and labor, and thus the previous experimental data is invaluable. Notwithstanding the main resource of XAS information is publications, no significant progress has been made in extracting XAS information from literature as it is an arduous task. To the best of our knowledge, our proposed work is the first attempt to mine literature for XAS information, which is supplied via a user-friendly search interface. Unlike the works that analyzed XAS images themselves for spectra similarity measures [18], materials parameter prediction [19], structural characterization [20], our work focuses on XAS information residing in textual data in the literature.

In recent years, materials informatics has been rapidly expedited along with the advance in high-throughput screening and big data analysis, and literature mining has been receiving more attention than ever in materials science and chemistry [21, 22]. Jensen *et al* [23] extracted Zeolite synthetic information from literature and trained a random forest regression model to predict Zeolite trends. Tshitoyan *et al* [24] has adopted a simple context-free word embedding method to identify relations between materials science literature. Kim *et al* [25] developed an automated pipeline to process materials synthesis parameters in scientific papers and applied machine leaning algorithms to predict parameters for titania nanotubes via hydrothermal methods. Kononova *et al* [26] generated a database for solid-state synthesis recipes extracted from literature using heuristic and neural network models. While the previous works adopted simple word representation methods and somewhat focused on syntactic patterns of texts, the proposed work will leverage contextual information of text to perform deeper semantic analysis in order to build a model to identify text segments that explain XAS figures.

2.3. Related work for figure explanatory text extraction

There have been previous research attempts to extract information relevant to figures from the scientific literature. A weight propagation approach [27] calculated word and sentence weights by an iterative weight update chain process of positional distance between sentences and word importance to rank figure-related sentences. Bhatia and Mitra [3] created a list of cue words for figures by manual inspection and used a traditional count-based method for similarity measure between caption and sentences to find information related to figures. A figure summarization system [2, 28] selected figure-relevant sentences by analyzing word similarity between sentences and figure captions, the relation to articles' thematic terms, and locational information of a sentence. The previous works have mainly focused on superficial characteristics to identify figure explanatory text, and a shallow semantic analysis has been conducted, such as sentence similarity based on the co-occurring words. To improve the machine's ability to distinguish sentences, extracting figure explanatory text in a paper requires deeper text understanding that goes beyond structural and literal patterns as presented in earlier work. Instead, our method will focus more on the process of semantic aspects of natural language. In our earlier work [29], we constructed an ontology to find figure-descriptive concepts in scientific papers. Although the ontological semantics-based approach performs well on capturing concepts for figure descriptions (e.g. SHOW-INFORMATION concept: 'show,' 'illustrate,' 'demonstrate;' GRAPHICAL-REPRESENTATION concept: 'shape,' 'color,' 'line'), the scope of the model is limited and does not take into account the entire article. Thus, it is not suitable for finding figure explanatory text from the entire body text because it is impractical to human engineer concepts to the extent of covering the corpus in the ontology.

Transfer learning has become a dominant paradigm in deep-learning-based NLP as it substantially enhances performance [30]. In general, a large amount of a labeled training dataset is required to create an effective deep learning model. However, domain-specific labeled data are scarce, resulting in a lack of knowledge for a model to learn. Transfer learning alleviates this problem by learning universal linguistic knowledge from unlabeled text data in an unsupervised way. This is called 'pre-training.' Subsequently, the knowledge is delivered to a model for downstream supervised tasks that can rely on a relatively small amount of labeled data. Models using transfer learning have proven successful at capturing more linguistic properties than models trained on small amounts of labeled data from scratch [31, 32]. With regard to pre-trained models, pre-training a language model [33] has shown better performance than other pre-training tasks, such as autoencoding or translation [34, 35]. The Bidirectional Encoder Representations from Transformers (BERT) is the most popular technique for transfer learning via pre-trained language models, and a number of pre-trained BERT models for general or special purpose are available [36]. BERT generates word embeddings by language modeling. BERT analyses the context of a word in a bidirectional way rather than unidirectionally (e.g. left to right and right to left). BERT achieves bidirectional language modeling by adapting encoder in the Transformer architecture [37]. Encoder is an input processor in a sequence-to-sequence model for language understanding. The bidirectional language models based on Transformer encoder have outperformed standard unidirectional language modeling methods in reading

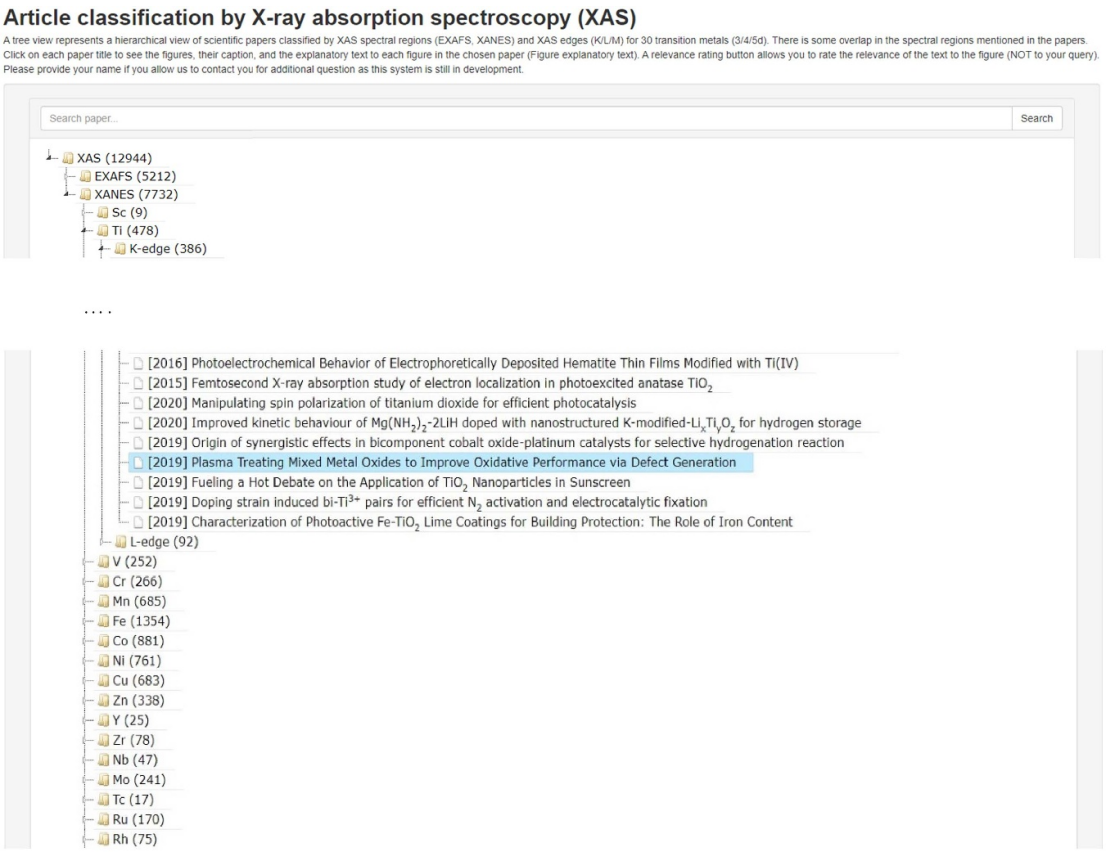

Figure 2. Classification tree for XAS articles.

comprehension tasks, including classification, question answering, and named entity recognition [36, 38–40]. Recently, a figure summarization task in a biomedical domain has adopted a domain-specific BERT model, called 'BioBERT,' and achieved improved results compared to rule-based and literal feature-based approaches [41].

3. Figure and text extraction for XAS with semantic methods

3.1. Classification of articles related to XAS

For XAS classification, articles are classified by XAS techniques, chemical elements, and types of XAS edges, and the classification result is presented as a tree structure (figure 2). The XAS technique is the top level of the tree, which are the two main XAS techniques, extended x-ray absorption fine structure (EXAFS) and x-ray absorption near edge structure (XANES). Chemical elements (e.g. transition metals) are the second level of the tree. We currently include 30 transition metals because these compounds are crucial to both fundamental studies and a wide range of technological applications. The last level of the tree presents three XAS K-, L-, and M-edges. The expression of those three features in scientific articles is relatively well standardized in the following way: the name of an element followed by a type of edge and the XAS technique used (e.g. Ti K-edge EXAFS, Au L2 edge XANES). We have established simple heuristic rules for capturing those three features from text with domain experts and applied the rules to figure captions for the classification. The details of the implementation and performance evaluation of the classification can be found in [42]. The XAS classification tree provides researchers with a filtered list of search results by XAS-specific features. Selecting an article in these results display XAS spectra images, figure captions, and figure explanatory text.

3.2. Extracting figure explanatory text for XAS with semantic methods and obtaining labeled data

To make search results more informative, the system provides text information that describes figures in addition to figure captions. Figure captions usually have a short description or a brief introductory comment and do not contain sufficient information to explain figures to readers or allow them to decide if a figure is useful for understanding their experimental results. Full information about figures resides in body text, and figure explanatory text from body text should be supplied to improve users' understanding of figures.

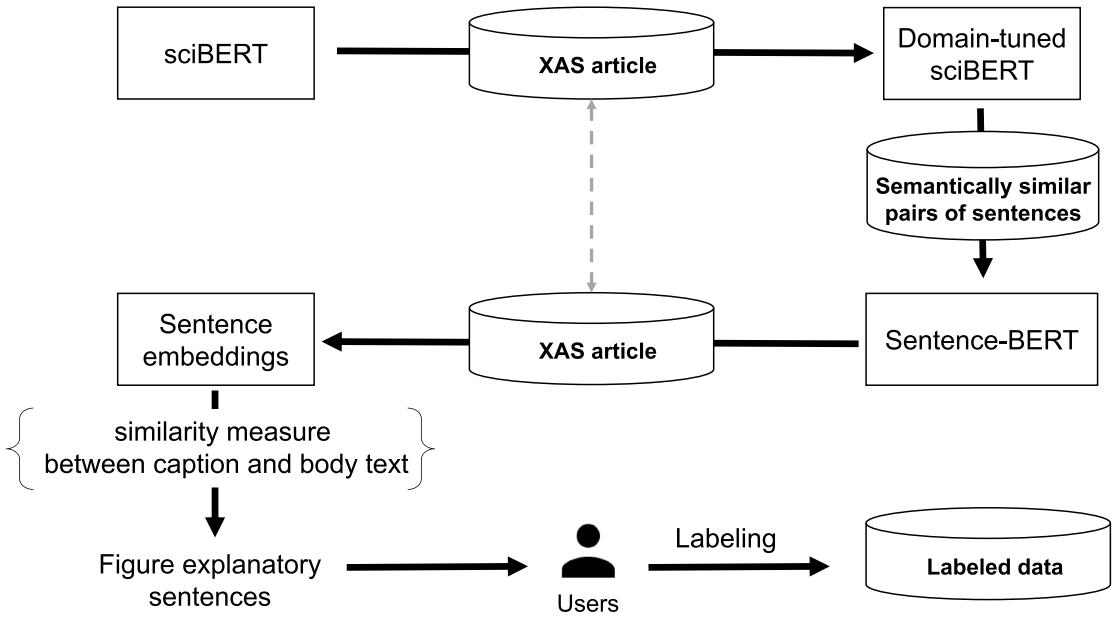

Figure 3. The process of figure explanatory text extraction for XAS figures.

Finding text to explain a specific XAS spectrum requires considerable domain knowledge. The knowledge can be embedded into a model if a large collection of labeled datasets exists. The absence of labeled data and practical difficulties in such data acquisition have led us to come up with a solution to train a model without labeled data. This method produces meaningful preliminary outcomes that can be improved by requesting user feedback. We built a model to find the most similar sentences to figure captions based on the assumption that sentences with high semantic similarity to captions would contain meaningful information about the figures.

For the task of text recognition of figure explanatory text centered on semantics, this project adopted the BERT deep learning-based language representation technique. In particular, we used SciBERT, a pre-trained model on one million full text scientific articles [43], and Sentence-BERT, a BERT fine-tuning method for sentence embeddings [44]. To make a model learn domain-specific knowledge, we further pre-trained the model on our collection of XAS articles using Sentence-BERT (figure 3).

To measure similarities, sentences are transformed into vector representations. Sentence-BERT uses a Siamese neural network (twin BERT), where each BERT model takes a sentence and generates a sentence embedding (pooling, e.g. a mean of output word embeddings) then compares the two sentence embeddings using a pair of similar sentence datasets (e.g. Natural Language Inference (NLI) corpus). Sentence-BERT updates model weights considering semantic similarity at a sentence level, and the performance showed it was superior to sentence representation approaches based on simple combinations of embeddings at a word level. We trained a Sentence-BERT model on the Stanford NLI (SNLI) dataset [45], Multi-Genre NLI (MultiNLI) dataset [46], and Semantic Textual Similarity dataset [47]. This is illustrated in figure 3 by the box labeled, 'Semantically similar pairs of sentences.' The trained Sentence-BERT model is deployed to calculate similarity scores between figure captions and sentences in body text. Before the top N sentences are selected as figure explanatory sentences by similarity scores, sentences are first filtered by NLI results to account for textual entailment relations. NLI is a classification task to determine whether a hypothesis is true (entailment label), false (contradiction label), or undetermined (neutral label) given a premise. We set a caption as premise and selected entailed sentences for each figure using a softmax classifier on Sentence-BERT trained on SNLI and MultiNLI.

Figure explanatory sentences selected by the model are presented along with XAS figures on the web interface, and users can rate the relevance of the sentences. The rating system aims to evaluate the model's accuracy and obtain a human-labeled dataset that can be used to fine tune the model.

3.3. Datasets

The selection of publications has been determined by domain experts in chemistry and materials science, which includes Elsevier, Springer Nature, Royal Society of Chemistry (RSC), American Association for the Advancement of Science (AAAS/Science), PubMed Central (PMC). Article scraping has been conducted under the publishers' text and data mining agreements. To respect these agreements, the articles in our

Table 1. The number of articles and XAS figures for publishers.

Data source	Number of articles	Number of XAS spectra
Elsevier	29167	8336
RSC	7922	2913
Springer Nature	2021	499
PMC	1408	252
AAAS	147	15
Total	40665	12015

collection and results are only available when logged into the BNL domain. While these publishers are providing bulk download access at no additional cost through the BNL library licenses, the American Chemical Society (ACS), a major publisher in chemistry and materials science of interest to users, has chosen to charge additional fees for text mining purposes. As a result of these prohibitive fees, ACS articles are not included in this collection. The data sources provide proprietary RESTful (Representational State Transfer) APIs to download articles. Objects such as figures are also downloaded if publishers provide object files.

We have collated 40665 XAS articles using four keywords: 'XAFS' (x-ray absorption fine structure) and 'EXAFS', 'XANES' (aka 'NEXAFS' (near edge x-ray absorption fine structure)) that are two regimes in XAFS [48]. Our scraper first searches and downloads all journals and manuscripts available for each source, and the article parser discards unnecessary content type articles. For instance, in case of Elsevier, articles about handbook series and reference works are excluded. The article parser unifies heterogeneous article formats collected across the data sources and generates JSON files that contains 'uid', 'publisher', 'article type', 'title', 'year', 'author', 'keywords', 'abstract', 'body_text', and 'figures'. Some aritlces can appear in multiple sources (e.g. PMC repository contains articles from other publishers such as Elsevier, Springer, RSC), and uids (unique identifiers) that are mostly DOIs are used to remove duplicates. The date range of the collected articles is from 1997 to 2020. Table 1 shows the number of articles and XAS figures. The XAS corpus is used to domain tune a SciBERT model and create sentence embeddings for similarity measurement.

3.4. Results

The web portal provides two services for a figure search: (a) XAS classification tree and (b) article search. As discussed in section 3.1, XAS classification tree displays classified articles by XAS technique (top level), 30 transition metals (second level), and types of edges (third level) in a tree structure. Article search supports a conventional, faceted search, where users can search specific areas of articles (e.g. title, abstract, body text, or figure caption) with the additional option to only display articles containing XAS spectrum images. The difference between search functionality in the XAS classification service and the article search service is that search is limited to titles and captions in the XAS classification, while users can search any part of articles in article search. When a user needs to find XAS spectra by XAS features, such as XAS method and edge, the classification tree is a better choice. The article search interface can be used if a user needs to retrieve articles by specific key terms and phrases. Both services provide a link to display a XAS information web page that contains a XAS image and its caption and figure explanatory text explaining the spectrum and extracted using the deep learning model results. Figures 4 and 5 show an example of a XAS information web page with the interaction buttons for feedback. Figure 6 depicts an example of search results with articles having XAS spectra in the article search interface. In the example, the figure and text are from the original paper [49], and our contribution is the list of explanatory sentences from the paper for users to rate. In its development phase, the system also includes a way to collect labeled data to improve the model by providing selected users with a method to report feedback on the explanatory power of the figure explanatory text.

3.5. Evaluation and discussion

We have conducted an evaluation on the model performance for figure explanatory text extraction using ratings from domain experts. For each figure, the service provides the five most similar sentences to the caption as determined by the model and five other sentences randomly chosen from the article's body text. During the sentence selection, sentences containing obvious clues such as 'Fig. X…', 'in figure Y' were excluded. The selected ten sentences were randomly shuffled before displayed to users. The rationale behind mixing model selections and random selections and shuffling sentences is that users may be biased in their responses if all sentences are regarded as relevant (e.g. acquiescence bias or dissent bias) which is based on the assumption that some of random choices will be highly irrelevant to figures and sentences are in regular order (question order bias). With mixing and shuffling, the model performance can be compared with some baseline (randomly chosen sentences).

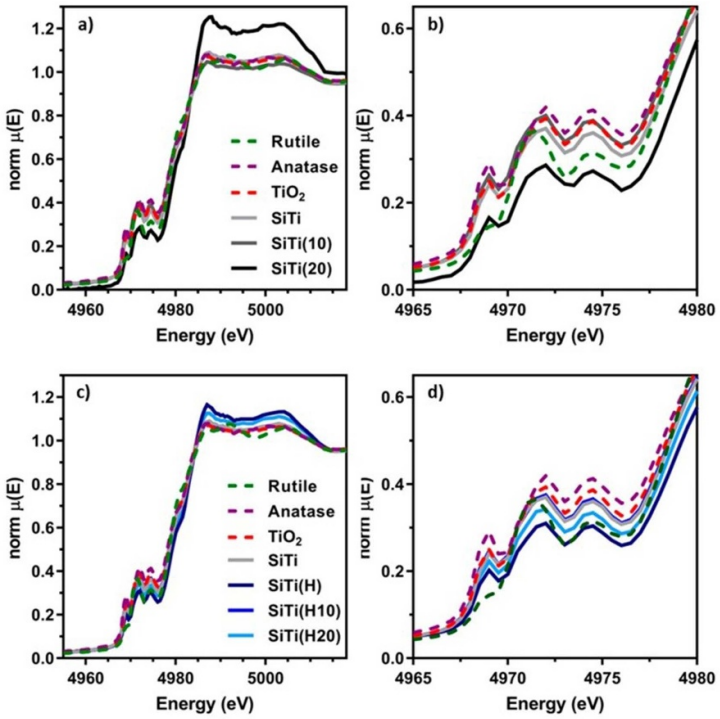

Figure 4. XAS spectra after the article selection in figure 2. Reproduced from [49]. CC BY 4.0.

[Figure caption]
(a) Ti–K edge and (b) pre-edge XANES for FSP-synthesized SiTi before and after 10 and 20 min of plasma treatment. (c) Ti–K edge and (d) pre-edge XANES for FSP-synthesized SiTi after hydrogenation and/or 10–20 min of plasma treatment. Pure anatase, rutile, and FSP-prepared TiO2 controls are also included.

Sentences from text that explain the above figure – Please rate the relevance of the sentence to the figures.

1. It is a low energy method, requiring only 27.6 kJ of energy over the course of the 20 min treatment. **Is this sentence explanatory for the figure?** * ○ more relevant ○ less relevant
2. Similarly, the plasma treatment increased the Ti3+ content of the hydrogenated SiTi(H) from 3.9% to 6.8% after a 20 min plasma treatment. **Is this sentence explanatory for the figure?** * ○ more relevant ○ less relevant
3. Upon plasma treatment, the Ti–Ti NNDs shift in the SiTi(10) sample, while a new feature appears in the SiTi(20) material. **Is this sentence explanatory for the figure?** * ○ more relevant ○ less relevant
4. A similar observation was made for SiTi(H20), but this trend was partially reversed for SiTi(H10), which was plasma treated for 10 min. **Is this sentence explanatory for the figure?** * ○ more relevant ○ less relevant
5. While neat SiTi shows no sign of Ti3+ content being present, as confirmed by EPR and XPS analyses, hydrogenation resulted in a 3.9% content of Ti3+ in SiTi(H). **Is this sentence explanatory for the figure?** * ○ more relevant ○ less relevant
6. Raman spectroscopy was carried out using a Renishaw inVia Raman microscope (Wotton-under-Edge, UK) in backscattering configuration, using an Ar laser to yield an excitation line of 514.5 nm. **Is this sentence explanatory for the figure?** * ○ more relevant ○ less relevant
7. Plasma treatments of 5, 10, 20, and 30 min were used on the neat and hydrogenated SiTi samples, which are denoted as SiTi(x) or SiTi(Hx), respectively, where x represents the plasma treatment time. **Is this sentence explanatory for the figure?** * ○ more relevant ○ less relevant
8. The net power used to generate the plasma was 15.0 + 0.1 W. **Is this sentence explanatory for the figure?** * ○ more relevant ○ less relevant
9. While this is often associated with macroporous materials, it has also been reported to result from mesopores forming between particle aggregates . **Is this sentence explanatory for the figure?** * ○ more relevant ○ less relevant
10. Modest white line intensity increases are observed for the SiTi(H) and SiTi(H20) materials, while a much larger white line intensity is observed for SiTi(20). **Is this sentence explanatory for the figure?** * ○ more relevant ○ less relevant

Your name * : **Comment:** Send Feedback

Figure 5. XAS explanatory text displayed after the article selection in figure 2. Reproduced from [49]. CC BY 4.0.

Users search for XAS spectra of interest using either the XAS classification tree or the article search and rate if a sentence is explanatory or not to the figure. Thus far, we have collected feedback from domain experts from several NSLS-II beamlines, e.g. inner-shell spectroscopy (ISS) and x-ray fluorescence microprobe (XFM), as well as the Center for Functional Nanomaterials (CFN), a Department of Energy Nanoscale

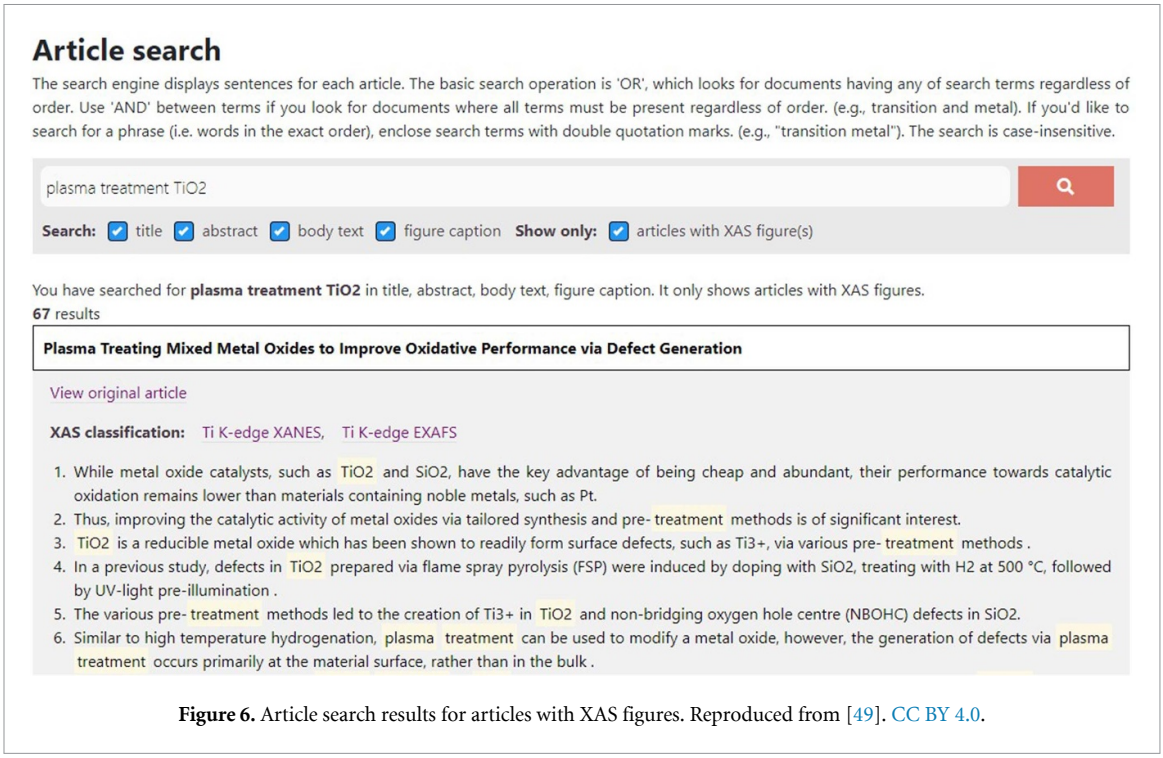

Figure 6. Article search results for articles with XAS figures. Reproduced from [49]. CC BY 4.0.

Table 2. Accuracy and F1 score of whether selected sentences by the trained model and random sentences are relevant to figures or not.

	ACC (%)	F1-Score
Sentence by model	50.32	0.6695
Random sentence	17.42	0.2967

Science Research Center located at BNL. We have collected 31 sets of feedback on figure explanatory text for figures from each facility, and the total labeled sentences were 930 sentences (465 sentences chosen by the model and 465 random sentences from body text) for 93 figures. Table 2 displays the evaluation results.

Both the model selections and random sentences were set as 'relevant' (predicted labels) for comparative evaluation, and the values were compared with the user decisions (true labels). The model's selection showed much higher relevance to figures than random choice. The result explains that 50% of sentences by the model were explanatory to the XAS spectra, whereas only 17% of the random sentences were explanatory. This result is in line with the findings of a recent study by Saini *et al* [41], where the model using BERT achieved much higher precisions compared to random sentences for biomedical figure summary datasets. Some of the disagreement between human and machine decisions can be due to the lack of context for given sentences. The domain experts mentioned that they were unable to make a clear judgement for some of the sentences unless more context was provided, including acronyms and abbreviation interpretations, which led them to decide that such sentences were not explanatory. Another factor can be that we adopted the binary rating scale for simplicity. However, this might not properly reflect users' decision that may reside in the range scale (e.g. 1–5). Additionally, when selecting figure explanatory sentences, we excluded sentences directly referring to figures. These sentences contain obvious clues (e.g. 'Fig. X describes' or 'as shown in figure Y'), it is a trivial task to find them, and it is reasonable to assume that they are highly associated with the corresponding figure. In fact, the clues referring to figures (e.g. figure X) contribute more to figure explanatory information than semantic similarity and textual entailment features [41]. When the model evaluation phase is finished, these sentences with obvious clues will be presented on the web portal. It is possible that finding the five best explanatory sentences, besides the obvious ones, may result in presenting less than ten explanatory sentences. In a human-annotated dataset of figure summary for biomedical journals [50], sentences referring to figures amount to 38% of summaries. After removing these sentences, the average number of sentences summarizing figures decreases from 7.5 sentences to less than five sentences. As an initial phase of model training, we have focused on simplifying the user feedback interface to obtain as many labeled samples as possible. We will continue to improve model results as we collect more labeled data from users.

4. Conclusion and future work

Scientific literature is a vast source of knowledge that remains relatively unexplored with automated approaches. The literature contains massive amounts of unstructured data and presents numerous opportunities for systematically extracting valuable knowledge. We have described a system for mining information about experimental spectra from scientific papers that performs article scraping, model creation, and information delivery. The system is designed for materials science and chemistry disciplines from experiments that are the most commonly performed at the scientific user facilities at BNL (NSLS-II and CFN). The system's primary purpose is to provide scientists at those facilities with supplementary information for experiment preparation, comparison, and evaluation. The system application to XAS experiments has supplied domain scientists with the needed functionality to retrieve relevant x-ray spectra information during their time at the beamlines.

We plan to continue improving the information delivery based on users' feedback. Many users at BNL have positively commented on our pilot system and made suggestions to further improve its usability. The rating feature will be refined by adding a more precise rating scale. An additional feature of the future system will be to extend the classification and model to the full periodic table, covering a broader materials space and further facilitating the application of data-driven methods in spectral interpretation. One interesting topic can be a XAS trend analysis on which the beamline scientists at NSLSII showed interests. This can be achieved when we obtain sufficient training data and domain expertise, and we will consider cooperation with domain experts to design and build a predictive model. The system supports expandability, adaptability, and data integration for the development of future applications to other topics in scientific literature mining, for instance, synthesis recipe prediction, a challenging task in materials science due to its complexity and copious density of pathways.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Acknowledgments

The authors would like to thank Huub Van Dam (CSI), Denis Leshchev (NSLS-II, ISS), Sarah Nicholas (NSLS-II, XFM), Deyu Lu (CFN), Wei Chen (CFN), and Mingzhao Liu (CFN) for providing feedback. The authors gratefully acknowledge the funding support from the U.S. Department of Energy Office of Science. This manuscript has been authored by employees of Brookhaven Science Associates, LLC under Contract No. DESC0012704. This research was funded in part by support from the Laboratory Director's Research and Development (LDRD 18- 005).

ORCID iDs

Gilchan Park https://orcid.org/0000-0002-0153-6646
Line Pouchard https://orcid.org/0000-0002-2120-6521

References

[1] National Synchrotron Light Source II (NSLS-II) (available at: www.bnl.gov/ps/)
[2] Agarwal S and Yu H 2009 FigSum: automatically generating structured text summaries for figures in biomedical literature *AMIA Annual Symp. Proc.* vol 2009 (American Medical Informatics Association) p 6
[3] Bhatia S and Mitra P 2012 Summarizing figures, tables, and algorithms in scientific publications to augment search results *ACM Trans. Inf. Syst.* **30** 1–24
[4] Liu R and J X McKie *PyMuPDF* (available at: http://pymupdf.readthedocs.io/en/latest/)
[5] Bast H and Korzen C 2017 A benchmark and evaluation for text extraction from pdf *2017 ACM/IEEE Joint Conf. on Digital Libraries (JCDL)* (IEEE) pp 1–10
[6] Young T, Hazarika D, Poria S and Cambria E 2018 Recent trends in deep learning based natural language processing *IEEE Comput. Intell. Mag.* **13** 55–75
[7] Tenney I, Das D and Pavlick E 2019 BERT rediscovers the classical NLP pipeline (arXiv:1905.05950)
[8] Court C J and Cole J M 2018 Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction *Sci. Data* **5** 180111
[9] Swain M C and Cole J M 2016 ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature *J. Chem. Inf. Model.* **56** 1894–904

[10] Gormley C and Tong Z 2015 *Elasticsearch: The Definitive Guide: A Distributed Real-time Search and Analytics Engine* (Newton, MA: O'Reilly Media, Inc.)
[11] Koningsberger D C and Prins R 1988 X-ray absorption: principles, applications, techniques of EXAFS, SEXAFS, and XANES
[12] Yano J and Yachandra V K 2009 X-ray absorption spectroscopy *Photosyn. Res.* **102** 241–54
[13] Cibin G, Gianolio D, Parry S A, Schoonjans T, Moore O, Draper R, Miller L A, Thoma A, Doswell C L and Graham A 2020 An open access, integrated XAS data repository at diamond light source *Radiat. Phys. Chem.* **175** 108479
[14] Ravel B, Hester J R, Solé V A and Newville M 2012 Towards data format standardization for x-ray absorption spectroscopy *J. Synchrotron Radiat.* **19** 869–74
[15] Ewels P, Sikora T, Serin V, Ewels C P and Lajaunie L 2016 A complete overhaul of the electron energy-loss spectroscopy and x-ray absorption spectroscopy database: eelsdb.eu *Microsc. Microanal.* **22** 717–24
[16] Asakura K, Abe H and Kimura M 2018 The challenge of constructing an international XAFS database *J. Synchrotron Radiat.* **25** 967–71
[17] Mathew K, Zheng C, Winston D, Chen C, Dozier A, Rehr J J, Ong S P and Persson K A 2018 High-throughput computational x-ray absorption spectroscopy *Sci. Data* **5** 1–8
[18] Zheng C, Mathew K, Chen C, Chen Y, Tang H, Dozier A, Kas J J, Vila F D, Rehr J J and Piper L F 2018 Automated generation and ensemble-learned matching of x-ray absorption spectra *npj Comput. Mater.* **4** 1–9
[19] Suzuki Y, Hino H, Kotsugi M and Ono K 2019 Automated estimation of materials parameter from x-ray absorption and electron energy-loss spectra with similarity measures *npj Computat. Mater.* **5** 1–7
[20] Timoshenko J and Frenkel A I 2019 'Inverting' x-ray absorption spectra of catalysts by machine learning in search for activity descriptors *ACS Catal.* **9** 10192–211
[21] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A and Kim C 2017 Machine learning in materials informatics: recent applications and prospects *npj Computat. Mater.* **3** 1–13
[22] Hakimi O, Krallinger M and Ginebra M-P 2020 Time to kick-start text mining for biomaterials *Nat. Rev. Mater.* **5** 553–6
[23] Jensen Z, Kim E, Kwon S, Gani T Z, Román-Leshkov Y, Moliner M, Corma A and Olivetti E 2019 A machine learning approach to zeolite synthesis enabled by automatic literature data extraction *ACS Central Sci.* **5** 892–9
[24] Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, Persson K A, Ceder G and Jain A 2019 Unsupervised word embeddings capture latent knowledge from materials science literature *Nature* **571** 95–8
[25] Kim E, Huang K, Saunders A, McCallum A, Ceder G and Olivetti E 2017 Materials synthesis insights from scientific literature via text extraction and machine learning *Chem. Mater.* **29** 9436–44
[26] Kononova O, Huo H, He T, Rong Z, Botari T, Sun W, Tshitoyan V and Ceder G 2019 Text-mined dataset of inorganic materials synthesis recipes *Sci. Data* **6** 1–11
[27] Takeshima R and Watanabe T 2012 The extraction of figure-related sentences to effectively understand figures *Innovations in Intelligent Machines–2* (Berlin: Springer) pp 19–31
[28] Ramesh B P, Sethi R J and Yu H 2015 Figure-associated text summarization and evaluation *PLoS One* **10** e0115671
[29] Park G, Rayz J T and Pouchard L 2020 Figure descriptive text extraction using ontological representation *The Thirty-Third Int. Flairs Conf.*
[30] Pan S J and Yang Q 2009 A survey on transfer learning *IEEE Trans. Knowl. Data Eng.* **22** 1345–59
[31] Kuncoro A, Dyer C, Rimell L, Clark S and Blunsom P 2019 Scalable syntax-aware language models using knowledge distillation (arXiv:1906.06438)
[32] Liu N F, Gardner M, Belinkov Y, Peters M E and Smith N A 2019 Linguistic knowledge and transferability of contextual representations (arXiv:1903.08855)
[33] Dai A M and Le Q V 2015 Semi-supervised sequence learning *Advances in Neural Information Processing Systems* pp 3079–87 (arXiv:1511.01432)
[34] Wang A, Hula J, Xia P, Pappagari R, McCoy R T, Patel R, Kim N, Tenney I, Huang Y and Yu K 2018 Can you tell me how to get past sesame street? Sentence-level pretraining beyond language modeling (arXiv:1812.10860)
[35] Zhang K and Bowman S 2018 Language modeling teaches you more than translation does: lessons learned through auxiliary syntactic task analysis *Proc. 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* pp 359–61
[36] Devlin J, Chang M-W, Lee K and Toutanova K 2018 Bert: pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805)
[37] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* pp 5998–6008 (arXiv:1706.03762)
[38] Sun Y, Wang S, Li Y-K, Feng S, Tian H, Wu H and Wang H 2020 ERNIE 2.0: a continual pre-training framework for language understanding *AAAI* pp 8968–75
[39] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R R and Le Q V 2019 Xlnet: generalized autoregressive pretraining for language understanding *Advances in Neural Information Processing Systems* pp 5753–63 (arXiv:1906.08237)
[40] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L and Stoyanov V 2019 Roberta: a robustly optimized bert pretraining approach (arXiv:1907.11692)
[41] Saini N, Saha S, Bhattacharyya P and Tuteja H 2020 Textual entailment–based figure summarization for biomedical articles *ACM Trans. Multimedia Comput. Commun. Appl.* **16** 1–24
[42] Park G and Pouchard L 2019 Scientific literature mining for experiment information in materials design *2019 New York Scientific Data Summit (NYSDS)* (IEEE) pp 1–4
[43] Beltagy I, Lo K and Cohan A 2019 SciBERT: a pretrained language model for scientific text (arXiv:1903.10676)
[44] Reimers N and Gurevych I 2019 Sentence-bert: sentence embeddings using siamese bert-networks (arXiv:1908.10084)
[45] Bowman S R, Angeli G, Potts C and Manning C D 2015 A large annotated corpus for learning natural language inference *Proc. 2015 Conf. on Empirical Methods in Natural Language Processing (EMNLP) 2015* (Lisbon: Association for Computational Linguistics) pp 632–42
[46] Williams A, Nangia N and Bowman S 2018 A broad-coverage challenge corpus for sentence understanding through inference *Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Long Papers) (NAACL-HLT) 2018* vol 1(New Orleans, LA: Association for Computational Linguistics) pp 1112–22

[47] Cer D, Diab M, Agirre E, Lopez-Gazpio I and Specia L 2017 SemEval-2017 task 1: semantic textual similarity multilingual and crosslingual focused evaluation *Proc. 11th Int. Workshop on Semantic Evaluation (Semeval-2017) SemEval 2017* (Vancouver: Association for Computational Linguistics) pp 1–14
[48] Newville M 2014 Fundamentals of XAFS *Rev. Mineral. Geochem.* **78** 33–74
[49] Horlyck J, Nashira A, Lovell E, Daiyan R, Bedford N, Wei Y, Amal R and Scott J 2019 Plasma treating mixed metal oxides to improve oxidative performance via defect generation *Materials* **12** 2756
[50] Ramesh B P 2013 Figure associated text summarization and evaluation (available at: https://figshare.com/articles/dataset/Figure_Associated_Text_Summarization_and_Evaluation/858903/1)