Opportunities and Challenges of Text Mining in Materials Research

Olga Kononova[1,2], Tanjin He[1,2], Haoyan Huo[1,2], Amalie Trewartha[2], Elsa A. Olivetti[3], and Gerbrand Ceder[1,2,*]

[1]Department of Materials Science & Engineering, University of California, Berkeley, CA 94720, USA

[2] Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[3] Department of Materials Science & Engineering, MIT, Cambridge, MA 02139, USA

[*]Lead contact and Corresponding author: gceder@berkeley.edu

January 30, 2021

Research publications are the major repository of scientific knowledge. However, their unstructured and highly heterogenuous format creates a significant obstacle to large-scale analysis of the information contained within. Recent progress in natural language processing (NLP) has provided a variety of tools for high-quality information extraction from unstructured text. These tools are primarily trained on non-technical text and struggle to produce accurate results when applied to scientific text, involving specific technical terminology. During the last years, significant efforts in information retrieval have been made for biomedical and biochemical publications. For materials science, text mining (TM) methodology is still at the dawn of its development. In this review, we survey the recent progress in creating and applying TM and NLP approaches to materials science field. This review is directed at the broad class of researchers aiming to learn the fundamentals of TM as applied to the materials science publications.

---

1 Introduction & Background

The first example of statistical analysis of publications dates back to 1887 when Thomas C. Mendenhall suggested a quantitative metric to characterize authors' writing styles (Mendenhall, 1887). At that time, the analysis of the literature was widely used to resolve authorship disputes, and, of course, was entirely manual. In the 1940-1960s, the development of computers gave a significant boost to the growth of linguistic analysis. The work of Stephen C. Kleene on regular expressions and finite automata (Kleene, 1956), subsequent formal language theory described by Noam Chomsky (1956), and the important fundamental work on information theory by Claude Shannon (1951) became the foundation for what is now known as natural language processing (NLP). The following decades brought diverse research results along different aspects of text mining (TM) and NLP: automated generation of article abstracts (Luhn, 1958), regular expressions compilers (Thompson, 1968), automated dialog assistant (Weizenbaum, 1983), the first structured text collection – the Brown University Standard Corpus of American English (www.korpus.uib.no/icame/manuals), and many others (Miner et al., 2012).

In the 1990s, technological progress permitted storage and access to large amounts of data. This shifted NLP and machine learning (ML) from a knowledge-based methodology towards data-driven approaches (Kurgan and Musilek, 2006). The accelerated development of the Internet and the Web during this decade facilitated information sharing and exchange. This is also reflected in the rapid growth of scientific publications (Bornmann and Mutz, 2015) over this period. Our analysis of the papers indexed in the Web of Science repository shows that since the beginning of 2000s, the number of publications in different fields of materials science has increased exponentially (Figure 1).

There are significant opportunities in leveraging data to guide materials research, which is driven by such aspects as property prediction, the search for novel materials, identifying synthesis routes, or determining device parameters. Data is central to the materials informatics enterprise as the availability of large quantities of machine-readable data is a prerequisite to leverage statistical approaches to accelerate materials research (Ramprasad et al., 2017). Not surprisingly, early work on data-driven learning approaches therefore focused on the few highly curated datasets in the materials field, such as crystal structure data (Fischer et al., 2006; Hautier et al., 2011) or on computed property data which can be generated homogeneously and at high rate (Jain et al., 2013; de Jong et al., 2015; Ricci et al., 2017).

However, knowledge acquisition in materials science must generally be performed across insufficient, diverse, and heterogeneous data. These data range across disparate materials systems and a multitude of characterization approaches to comprehend thermomechanical, electromagnetic and chemical properties (Morgan and Jacobs, 2020). Publications are still the primary way to communicate within the scientific discipline. Therefore, there is substantial potential in capturing unstructured information from the vast and ever-growing number of scientific publications.

Textual information exists in an unstructured or highly heterogeneous format. Manual data extraction is expensive, labor-intensive, and error-prone (although some powerful examples exist in the materials

community (Blokhin and Villars, 2020; Gallego et al., 2016b; Gallego et al., 2016a)). As a result, there are tremendous opportunities for large-scale automated data extraction to transform materials science into a more quantitative and data-rich field.

This review discusses recent advances in automated text processing and information extraction from a large corpus of chemical, physical and materials science publications. We first discuss the methods and approaches widely used in TM and NLP (Section 2). Then we survey some prominent case studies that are focused on data collection and data mining (Section 3). We highlight some major challenges and obstacles in scientific TM (Section 4). Lastly, we discuss potential future research developments for NLP in its application to materials science (Section 5).

2 Text Mining of Scientific Literature

Modern computers encode text as a monotonic sequence of bits representing each character, but without reflecting its internal structure or other high-order organization (e.g. words, sentences, paragraphs). Building algorithms to interpret the sequences of characters and to derive logical information from them is the primary purpose of TM and NLP. Unlike standard texts on general topics, such as newswire or popular press, scientific documents are written in specific language requiring sufficient domain knowledge to follow the ideas. Application of general-purpose TM and NLP approaches to the chemical or materials science domain requires adaptation of both methods and models, including development of an adequate training sets that comply with the goals of the TM project.

Generally, a scientific TM pipeline breaks down into the following steps (Figure 2): i) retrieval of documents and conversion from markup languages or PDF into plain text; ii) text pre-processing, i.e. segmentation into sentences and tokens, text normalization and morphological parsing; iii) text analysis and information extraction; iv) data normalization and database structuring. The resulting collection either serves as a final product of the TM or provides a source of data for further mining and analysis.

While a comprehensive discussion of the algorithms and methods used to accomplish each task of the pipeline is beyond the scope of this review, we cover in this Section those methods that are widely applied in scientific TM. We also revise state-of-the-art NLP parsing tools needed to handle chemical and materials science texts. We emphasize the challenges arising along the way and discuss possible solutions. For details and theoretical background on TM and NLP models in general, we refer the reader to the following books: (Miner et al., 2012) and (Jurafsky and Martin, 2009).

2.1 Obtaining the text corpus

In computational linguistics, a large organized set of human-created documents is referred to as a *text corpus*. Scientific discourse generally occurs across a wide variety of document formats and types: abstracts in proceedings, research articles, technical reports and pre-prints, patents, e-encyclopedias, and many more.

There are two primary ways to obtain the text corpus: i) by using existing indexed repositories with the available text-mining application programming interfaces (APIs) and search tools; or ii) by having access to an individual publisher's content.

Text databases. A comprehensive overview of scientific text resources can be found in review of Kolářik et al. (2008). Table 1 lists some common repositories for scientific texts in the domain of chemistry and material science, their document types, and access options. The main advantage of using established databases for text mining is the uniform format of their metadata, a convenient API, and sometimes analysis tools. However, the majority of the publications in these repositories are heavily biased toward biomedical and biochemical subjects with a smaller fraction belonging to physics, (in)organic chemistry, and materials science. Moreover, the access to the content is limited: it either requires having a subscription or provides a search over open-access publications only.

Individual publisher access. Implementation of a customized scraping routine to screen the publisher's web-pages and download the content requires more effort. However, this approach allows for accessing content from those resources that are not providing an API, for example, e-print repositories. In most cases, downloading and accessing significant publisher content require text and data mining (TDM) agreements. We note that this TDM agreement differs from a standard academic subscription granted to the libraries of the institutions, because scraping and downloading large volumes, affect the operation of the publishers' server.

Web-scraping not only requires a substantial amount of work, but it also has to respond to dynamic web pages in which content is generated by a client browser. In our recent work, we implemented such a solution for Elsevier, RSC, ECS, and AIP publishers (Kononova et al., 2019). Similarly, ChemDataExtractor (Swain and Cole, 2016) provides the web-scrapers for Elsevier, RSC and Springer. In the research fields where most of the literature has an open access repository, e.g. physics, mathematics or the rapidly growing literature collection on COVID-19 (Trewartha et al., 2020), the corpus acquisition step will be considerably easier.

2.2 Conversion into raw text

In general, the retrieved content includes the targeted text and other metadata, such as journal name, title, authors, keywords, and others. Querying text databases, as those in Table 1, provide a structured output with raw text ready for processing and analysis. In contrast, web-scraped content usually consists of a complete paper files requiring the additional step to convert it into a raw text. Nowadays, most of the text sources provide as HTML/XML/JSON documents, whereas older papers are usually available as embedded or image PDFs (Figure 1).

While parsing of HTML/XML markups can be performed with various programming tools, extraction of the plain text from PDF files is more laborious. Embedded PDFs usually have a block structure with the

text arranged in columns and intermixed with tables, figures, and equations. This affects the accuracy of conversion and text sequence. Some work has been done attempting to recover a logical text structure from PDF-formatted scientific articles by utilizing rule-based (Constantin et al., 2013) and ML (Tkaczyk et al., 2015; Luong et al., 2010) approaches. However, the accuracy of these models measured as F1-score is still below ~80%. The authors' experience demonstrates that this can dramatically impact the final output of the extraction pipeline (Fig. 2). Hence, the decision on whether to include PDF text strongly depends on the tasks that are being solved.

A great number of documents, in particular, those published before the 1990s, are only available as an image PDF (Figure 1). Conversion of these files into a raw text requires advanced optical character recognition (OCR), and, to the best of our knowledge, the currently available solutions still fail to provide high enough accuracy to reliably extract chemistry (Mouchère et al., 2016; Mahdavi et al., 2019). Often, interpretation errors in PDFs originate from subscripts in chemical formulas and equations, and from confusion between symbols and digits. Creating a rigorous parser for PDF articles, and especially an OCR for scientific text is an area of active research in the computer science and text mining community (Memon et al., 2020; Ramakrishnan et al., 2012).

2.3 Text pre-processing, grammatical and morphological parsing

The raw documents proceed through normalization, segmentation, and grammar parsing. During this step, the text is split into logical constitutes (e.g. sentences) and *tokens* (e.g. words and phrases), that are used to build a grammatical structure of the text. Depending on the final TDM goal, the text tokens may be normalized by *stemming* or *lemmatization* and processed through the *part of speech tagging* (POS tagging), and *dependencies parsing* to build the sentences structure. These are explained below.

Paragraph segmentation and sentence tokenization identify, respectively, the boundaries of the sentences and word phrases (tokens) in a text. In general, finding the start/end of a sentence segment requires recognition of certain symbolic markers, such as period (“.”), question mark (“?”), and exclamation mark (“!”), which is usually performed with (un)supervised ML models (Read et al., 2012). State-of-the-art implementations attain ~95-98% accuracy (measured as F1-score). However, applying these models to scientific text requires modification. Commonly used expressions such as “Fig. X”, “et al.” and a period in chemical formulas often result in over-segmentation of a paragraph. Conversely, citation numbers at the end of a sentence promote the merging of two sentences together. There is no generally accepted solution to this problem, and it is usually approached by hard-coding a set of rules that capture particular cases (Leaman et al., 2015).

Sentence tokenization, i.e. splitting a sentence into logical constituents, is a crucial step on the way to information extraction, because the errors produced in this step tend to propagate down the pipeline (Figure 2) and affect the accuracy of the final results. Tokenization requires both unambiguous definition of

grammatical tokens and robust algorithms for identification of the token boundaries. For general-purpose text, tokenization has been the subject of extensive research resulting in the development of various advanced methods and techniques (Jurafsky and Martin, 2009). However, for chemical and materials science text, accurate tokenization still requires substantial workarounds and revision of the standard approaches. Table 2 displays some typical examples of sentence tokenization produced by general-purpose tokenizers such as NLTK (Bird et al., 2009) and SpaCy (Honnibal and Johnson, 2015). As in the case of sentence segmentation, the major source of errors is the arbitrary usage of punctuation symbols within chemical formulas and other domain-specific terms. The chemical NLP toolkits such as OSCAR4 (Jessop et al., 2011), ChemicalTagger (Hawizy et al., 2011), and ChemDataExtractor (Swain and Cole, 2016) solve this problem by implementing their own rules- and dictionaries-based approaches to solve the over-tokenization problem. The advantage of chemical NLP toolkits is that they provide good performance on chemical terms, even if the rest of the text may have lower tokenization accuracy.

However, another prominent reason for tokenization errors is the lack of generally accepted rules regarding tokenization of chemical terms consisting of multiple words For instance, complex terms such as "lithium battery" or "yttria-doped zirconium oxide" or "$(Na_{0.5}K_{0.5})NbO_3$ + x wt% CuF_2" often become split into separate tokens "lithium" and "battery", "yttria-doped" and "zirconium" and "oxide", "$(Na_{0.5}K_{0.5})NbO_3$" and "+" and "x wt% CuF_2". This significantly modifies the meaning of the tokens and usually results in lowered accuracy of the named entity recognition (see below). Currently, this problem is solved case-by-case by creating task-specific wrappers for existing tokenizers and named entity recognition models (Huang and Ling, 2019; Alperin et al., 2016; He et al., 2020). Building a robust approach for chemistry-specific sentence tokenization and data extraction requires a thorough development of standard nomenclature for complex chemical terms and materials names. We discuss this challenge in detail in Section 4 below.

Text normalization, part-of-speech tagging, and dependency parsing are often used to reduce the overall document lexicon and to design words' morphological and grammatical features used as an input for entity extraction and other TM tasks (Leaman et al., 2015). Text normalization usually consists of lemmatization and/or its simpler version – stemming. While during the stemming the inflected word is cut to its stem (e.g. "changed" becomes "chang"), lemmatization aims to identify a word's lemma, i.e. a word's dictionary (canonical) form (e.g. "changed" becomes "change") (Jurafsky and Martin, 2009). Stemming and/or lemmatization help to reduce the variability of the language, but the decision whether to apply it or not, depends on the task and expected outcome. For instance, recognition of chemical terms will benefit less from stemming or lemmatization (Corbett and Copestake, 2008) as it may truncate a word's ending resulting in a change of meaning (compare "methylation" vs. "methyl"). But when a word identifies, for example, a synthesis action, lemmatization helps to obtain the infinitive form of the verb and avoids redundancy in the document vocabulary (Kononova et al., 2019).

Part-of-speech (POS) tagging identifies grammatical properties of the words and labels them with the

corresponding tags, i.e. noun, verb, article, adjective, and others. This procedure does not modify the text corpus but rather provides linguistic and grammar-based features of the words that are used as input for ML models. A challenge in identifying the POS tags in scientific text often arises due to the ambiguity introduced by the word's context. As an example, compare two phrases: "the chemical tube is on the ground" and "the chemical was finely ground". In first case, the general-purpose POS tagger will work correctly, while in the second example, it will likely misidentify "chemical" and "ground" as adjective and noun, respectively. Therefore, using a standard POS tagger often requires re-training of the underlying NLP model, or post-processing and correction of the obtained results.

Dependency parsing creates a mapping of a linear sequence of sentence tokens into a hierarchical structure by resolving the internal grammatical dependencies between the words. This hierarchy is usually represented as a *dependency tree*, starting from the *root* token and going down to the terminal nodes. Parsing grammatical dependencies helps to deal with the arbitrary order of the words in the sentence and establishes semantic relationships between words and parts of the sentence (Jurafsky and Martin, 2009). Grammatical dependency parsing is a rapidly developing area of NLP research providing a wealth of algorithms and models for general-purpose corpus (see www.nlpprogress.com for specific examples and evaluation).

Application of the currently existing dependency parsing models to scientific text comes with some challenges. First, sentences in science are often depersonalized, with excessive usage of passive and past verbs tense, and limited usage of pronouns. These features of the sentence are not well captured by general-purpose models. Secondly, the accuracy of the dependency tree construction is highly sensitive to punctuation and correct word forms, particularly verb tenses. As the scientific articles do not always exhibit perfect language grammar, the standard dependency parsing models can produce highly unpredictable results. To the best of our knowledge, these specific challenges of dependency parsing for scientific text have not yet been addressed or explored in detail.

2.4 Text representation modeling

The application of ML models requires mapping the document into a linear (vector) space. A common approach is to represent a text as a collection of multidimensional (and finite) numerical vectors that preserve the text features, e.g. synonymous words and phrases should have a similar vector representation, and phrases having an opposite meaning should be mapped into dissimilar vectors (Harris, 1954). Modeling of the vectorized text representation is a broad and rapidly developing area of research (Liu et al., 2020). In this section, we highlight only some of the approaches applied to scientific TM, whereas a more detailed discussion of the methods can be found elsewhere (Jurafsky and Martin, 2009).

The *bag-of-words model* is one of the simplest models of text representation. It maps a document into a vector by counting how many times every word from a pre-defined vocabulary occurs in that document. While this model works well for recognizing specific topics defined by keywords, it does not reflect word context and cannot identify the importance of a particular word in the text. The latter can be solved by

introducing a normalization factor and applying it to every word count. An example of such normalization is the *tf-idf model* (*term frequency-inverse document frequency*) which combines two metrics: the frequency of a word in a document and the fraction of the documents containing the word. The method can thereby identify the terms specific to a particular document. Bag-of-words and tf-idf are the most commonly used models to classify scientific documents or to identify parts of text with relevant information (Court and Cole, 2018; Kim et al., 2017c; Hiszpanski et al., 2020).

While bag-of-words and tf-idf are relatively versatile, they do not identify similarity between words across documents. This can be done through *topic modeling* approaches (Blei, 2012). Topic modeling is a statistical model that examines the documents corpus and produces a set of abstract topics – clusters of the keywords that characterize a particular text. Then, every document is assigned with a probability distribution over topical clusters. Latent Dirichlet Allocation, a specific topic modeling approach, (Blei et al., 2003) has been applied to analyze the topic distribution over materials science papers on oxide synthesis (Kim et al., 2017c) and to classify these papers based by synthesis method used in the paper (Huo et al., 2019).

Significant progress in TM and NLP has been achieved with the introduction of *word embedding* models which construct a vectorized representation of a single word rather than of the entire document. These approaches use the distributional hypothesis (Harris, 1954) and are based on neural networks trained to predict word context in a self-supervised fashion. Multiple variations of word embeddings models include GloVe (Pennington et al., 2014), ELMo (Peters et al., 2018), word2vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017). Besides being intuitively simple, the main advantage of word embedding models is their ability to capture similarity and relations between words based on mutual associations. Word embeddings are applied ubiquitously in materials science TM and NLP to engineer words features that are used as an input in various named entity recognition tasks (Kononova et al., 2019; Kim et al., 2020a; Huang and Ling, 2019; Weston et al., 2019). Moreover, they also seem to be a promising tool to discover properties of materials through words association (Tshitoyan et al., 2019).

Recently, research on text representation has shifted toward context-aware models. A breakthrough was achieved with the development of *sequence-to-sequence models* (Bahdanau et al., 2016) and, later, an *attention mechanism* (Vaswani et al., 2017) for the purpose of neural machine translation. The most recent models such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) and Generative Pre-trained Transformer (GPT) (Radford et al., 2019; Brown et al., 2020) are multi-layered deep neural networks trained on very large unlabeled text corpora, and demonstrate state-of-the-art NLP performance. These models offer fascinating opportunities for the future NLP development in domain of materials science (Kuniyoshi et al., 2020; Vaucher et al., 2020). We discuss them in greater details in the Section 5.

2.5 Retrieval of information from the text

Information retrieval (IR) represents a broad spectrum of NLP tasks that extract various types of data from the pre-processed corpus (Figure 3). The most ubiquitous IR task is *named entities recognition* (NER) which classifies text tokens in a specific category. In general-purpose text, these categories are usually names of locations, persons, etc., but in scientific literature the named entities can include chemical terms as well as physical parameters and properties. Extraction of action graphs of chemical synthesis and materials fabrication is another class of IR task that is closely related to NER. This task requires identification of action keywords, linking of them into a graph structure, and, if necessary, augmenting with the corresponding attributes characterizing the action (e.g. the action "material mixing" can be augmented with the attribute "mixing media" or "mixing time"). Lastly, data extraction from figures and tables represents another class of information that can be retrieved from scientific literature. This requires not only TM methods but also image recognition approaches. In this section we will mainly review the recent progress for chemical and materials NER and action graphs extraction, and will provide a brief survey of the efforts spent on mining of scientific tables and figures.

Chemical named entities recognition (NER) is a broadly defined IR task. It usually includes identification of chemical and materials terms in the text, but can also involve extraction of properties, physical characteristics and synthesis actions. The early applications of chemical NER were mainly focused on extraction of drugs and biochemical information to perform more effective document searches (Corbett and Copestake, 2008; Jessop et al., 2011; Rocktäschel et al., 2012; García-Remesal et al., 2013). Recently, chemical NER has shifted towards (in)organic materials and their characteristics (Swain and Cole, 2016; He et al., 2020; Weston et al., 2019; Shah et al., 2018), polymers (Tchoua et al., 2019), nanoparticles (Hiszpanski et al., 2020), synthesis actions and conditions (Vaucher et al., 2020; Hawizy et al., 2011; Kim et al., 2017c; Kononova et al., 2019). The methods used for NER vary from traditional rule-based and dictionary look-up approaches to modern methodology built around advanced ML and NLP techniques, including conditional random field (CRF) (Lafferty et al., 2001), long short-term memory (LSTM) neural networks (Hochreiter and Schmidhuber, 1997), and others. A detailed survey on the chemical NER and its methods can be found in recent reviews (Krallinger et al., 2017; Gurulingappa et al., 2013; Olivetti et al., 2020).

Extraction of chemical and materials terms has been a direction of intensive development in the past decade (Krallinger et al., 2017; Eltyeb and Salim, 2014). The publicly available toolkits use rules- and dictionaries-based approaches (e.g LeadMine (Lowe and Sayle, 2015)), statistical models (e.g OSCAR4 (Jessop et al., 2011)), and, predominantly, the CRF model (e.g. ChemDataExtractor (Swain and Cole, 2016), ChemSpot (Rocktäschel et al., 2012), tmChem (Leaman et al., 2015)) to assign labels to chemical terms. Some recent works implemented advanced ML models such as bi-directional LSTM models (He et al., 2020; Weston et al., 2019; Kuniyoshi et al., 2020) as well as a combination of deep convolutional and recurrent neural networks (Korvigo et al., 2018) to identify chemical and material terms in the text and use

context information to assign their roles. Table 3 shows a few examples of the NER output obtained using some of these tools and compares it to non-scientific NER models implemented in NLTK (Bird et al., 2009) and SpaCy (Honnibal and Johnson, 2015) libraries.

Often, the objective of scientific NER task is not limited to the identification of chemicals and materials, but also includes recognition of their associated attributes: structure and properties, amounts, roles and actions performed on them. Assigning attributes to the entities is usually accomplished by constructing a graph-like structure that links together all the entities and build relations between them. A commonly used graph structure is the grammatical dependency tree for a sentence (see Section 2.3). Traversing the sentence trees allows for resolving relations between tokens, hence, link the entities with attributes. ChemicalTagger (Hawizy et al., 2011) is one of the most robust frameworks that extends the OSCAR4 (Jessop et al., 2011) functionality and provides tools for grammatical parsing of chemical text to find the relation between entities and the corresponding action verbs. Similarly, ChemDataExtractor (Swain and Cole, 2016) can identify the chemical and physical characteristics (e.g. melting temperature) in the text and assign it to a material entity. A rules- and dictionaries-based relation-aware chemical NER model has been proposed by Shah et al. (2018) to build a search engine for publications. Weston et al. (2019) used the random forest decision model to resolve synonyms between chemical entities and materials-related terms. He et al. (2020) applied a two-step LSTM model to resolve the role of materials in a synthesis procedure. Onishi et al. (2018) used convolutional neural network model to build relations between materials, their mechanical properties and processing conditions which were extracted from publications by keywords search. Lastly, a combination of advanced NLP models has been recently used to extract the materials synthesis steps and link them into an action graph of synthesis procedures for solid-state battery materials (Kuniyoshi et al., 2020) and inorganic materials in general (Mysore et al., 2017).

Despite significant effort, the accuracy of the NER for chemical names and formulas is still relatively low compared to the general state-of-the-art NER models (Baevski et al., 2019; Li et al., 2020). Figure 4a displays the overall precision and recall for different chemical NER models reported in the corresponding publications. Both, precision and recall of the models vary from 60% to 98% (Figure 4a), whereas for the general-purpose NER, these values are >91% (see www.nlpprogress.com). There are two major challenges that obstruct training of high-accuracy chemical NER models: i) the lack of unambiguous definitions of the chemical tokens and their boundaries, and ii) the lack of the robust annotation schema as well as comprehensive labeled training sets for the supervised ML algorithms. Oftentimes, researchers manually create their own training set for specific tasks but with limited use for more general goals. Therefore, the success of chemical NER becomes a trade-off between the size of the annotated set and model complexity: either using simple model with limited capabilities on a small set of labeled data, or investing effort into annotation of a large dataset and using it with advanced models providing a higher accuracy of data extraction.

An early attempt in creating a labeled dataset for the chemical NER task was done by Kim et al. (2003) and Krallinger et al. (2015). The GENIA and CHEMDNER sets provide annotation schema and labeled

data of chemicals and drugs extracted from MEDLINE and PubMed abstracts, respectively. However, these corpora are heavily biased toward biomedicine and biochemical terms with only a small fraction of organic materials names present. The progress of the past few years brought a variety of annotated corpora to the materials science domain. Among the publicly available labeled dataset, there is the NaDev corpus consisting of 392 sentences and 2,870 terms on nanocrystal device development (Dieb et al., 2015), the dataset of 622 wet lab protocols of biochemical experiments and solution syntheses (Kulkarni et al., 2018), a set of 9,499 labeled sentences on solid oxide fuel cells (Friedrich et al., 2020), and an annotated set of 230 materials synthesis procedures (Mysore et al., 2019).

Extraction of information from tables and figures is another branch of scientific IR that has been rapidly developing in the past few years. The specific format of the figures and tables in scientific papers imposes substantial challenges for the data retrieval process. First, it is common that images (and sometimes the tables) are not directly embedded in the HTML/XML text but instead contain a link to an external resource. Second, connecting tables/images to the specific part of the paper text is an advanced task that does not have a robust solution to date. Third, both tables and images can be very complex: images can include multiple panels and inserts that require segmentation, while tables may have combined several rows and columns imposing additional dependencies on the data. To the best of our knowledge, only a few publications have attempted to parse tables from the scientific literature using heuristics and machine learning approaches (Jensen et al., 2019; Milosevic et al., 2019).

Image recognition methods have been broadly used in materials science but have so far been primarily focused on extracting information about the size, morphology, and the structure of materials from microscopy images. To date, the existing solutions for interpretation of microscopy images use variations of convolutional neural networks, and address diverse spectra of materials science problems (Azimi et al., 2018; Matson et al., 2019; Maksov et al., 2019; Roberts et al., 2019). While these models demonstrate a remarkable accuracy when applied directly to microscopy output, they are not intended to separate and process the images embedded in scientific articles. Steps toward parsing of article's images were reported recently. Mukaddem et al. (2020) developed the ImageDataExtractor tool that uses a combination of OCR and CNN to extract the size and shape of the particles from microscopy images. Kim et al. (2020b) used Google Inception-V3 network (Szegedy et al., 2016) to create the Livermore SEM Image Tools for electron microscopy images. This tool was later applied by Hiszpanski et al. (2020) to ~35,000 publications to obtain information about the variability of nanoparticles sizes and morphologies.

3 Using Text Mining in Materials Science: Case Studies

Data-driven materials discovery usually relies either on computational methods to calculate the structure and properties of materials and collect them in databases (Jain et al., 2013), or on experimental datasets that have been painstakingly collected and curated. Development of advanced approaches for scientific TM creates

broad opportunities to augment such data with a large amount of reported but uncollected experimental results. A few large-scale datasets extracted from the scientific publications have become available over the last few years (Court and Cole, 2018; Huang and Cole, 2020; Kim et al., 2017c; Jensen et al., 2019; Kononova et al., 2019). In this Section, we survey the publicly available datasets created by retrieval of information from chemistry, physics, and materials science publications, and discuss the most interesting results obtained from them.

3.1 Publicly available collections of text-mined data

While recently several data collections have been obtained by automated TM and NLP-based pipelines, there are a few large-scale datasets that have been manually extracted from scientific publications and are worth mentioning here.

The Pauling File Project (Blokhin and Villars, 2020) is one of the biggest manually curated collections of data for inorganic crystalline substances, covering crystallographic data, physical properties, and phase diagrams. The Pauling File Project provides data for the Materials Platform for Data Science (www.mpds.io), Pearson's Crystal Data (www.crystalimpact.com), and Springer Materials (www.materials.springer.com). Together, it contains more than 350,000 crystalline structures, 150,000 physical properties and 50,000 phase diagrams extracted from the scientific literature in materials science, engineering, physics, and inorganic chemistry from 1891 to present. The quality and accuracy of the extracted records are high, and they include expert interpretation and a summary of the original text. Nonetheless, significant human labor is required to maintain and update this database. Moreover, due to the human interpretation of the data, the records are highly heterogeneous and may require additional processing and normalization.

The Dark Reactions Project (www.darkreactions.haverford.edu) is another prominent dataset extracted manually from laboratory journals containing 3,955 parameters of failed hydrothermal synthesis experiments (Raccuglia et al., 2016). So-called "negative" sampling data are critical for ML applications that need to predict a "yes/no" answer. Unfortunately, the "no" results, i.e. unsuccessful experimental outcomes, are rarely published or made available to the broad research community. The Dark Reaction Project represents the first attempt to demonstrate the importance of sharing negative-result data within the chemistry and materials science domain.

A substantial effort in the automated extraction of materials properties from scientific publications has been done by the research group of J. Cole (University of Cambridge, UK). They developed ChemDataExtractor (Swain and Cole, 2016), an NLP toolkit for chemical text, and used it to build a large collection of phase transition temperatures of magnetic materials (Court and Cole, 2018), and a dataset of electrochemical properties of battery materials (Huang and Cole, 2020). The first set contains 39,822 records of Curie and Néel temperatures for various chemical compounds retrieved from 68,078 research articles (Court and Cole, 2018). These data augment the MAGNDATA database – a collection of ~1,000 magnetic structures manually extracted from publications by Gallego et al. (Gallego et al., 2016a; Gallego et al., 2016b). The battery

dataset includes 292,313 records collected from 229,061 papers covering electrochemical properties of battery materials such as capacity, voltage, conductivity, Coulombic efficiency, and energy density. It enhances by more than an order of magnitude the manually constructed dataset of Ghadbeigi et al. (2015) containing 16,000 property entries for Li-ion battery materials extracted from 200 publications.

A large-scale text-mined data collection of materials synthesis parameters has been developed by our team during the past few years. Kim et al. (2017c) generated a dataset of synthesis operations and temperatures for 30 different oxide systems mined from 640,000 full-text publications. Later on, this set was extended by 1,214 sol-gel-synthesis conditions for germanium-based zeolites (Jensen et al., 2019). A collection of 19,488 solid-state ceramics synthesis reactions containing precursors chemicals, synthesis steps and their attributes was generated from 53,538 materials science papers by Kononova et al. (2019).

It is important to highlight that although the TM and NLP methods help to generate large-scale datasets, the output can suffer from lower accuracy of extraction as compared to any manually curated dataset. For instance, the extraction precision of the Curie and Néel temperatures are ~82% (Court and Cole, 2018), and that of the electrochemical properties – ~80% (Huang and Cole, 2020), meaning that up to ~20% of the obtained records have one or more attributes incorrectly extracted. The dataset of oxides synthesis parameters shows categorical accuracy (i.e. the fraction of the predicted labels of the text tokens that match the true labels) for the chemical NER task of ~81% (Kim et al., 2017c). For the dataset of solid-state synthesis reactions, precision (i.e. fraction of correctly extracted entities) of extracted synthesis parameters varies from ~62% for fully accurate retrieval of synthesis conditions, to ~97-99% for extraction of precursor materials and final products (Kononova et al., 2019).

3.2 Text-mining-driven materials discoveries

Research exploring TM-based data-driven approaches to provide insights on materials emerged well before any progress in the development of robust NLP tools had been made. Several groups have attempted manual information extraction from a narrow set of publications with a specific scope.

The group of T. Sparks (University of Utah, US) explored the correlation between materials performance and the elemental availability for high-temperature thermoelectric materials (Gaultois et al., 2013) and Li-ion battery materials (Ghadbeigi et al., 2015). In both of these publications, the sets of physical parameters for materials classes were manually retrieved from queried materials science literature, and augmented with data on market concentration and Earth abundance for chemical elements. Based on this data the importance of considering global market state and geopolitical factors when designing materials was discussed.

An analysis of cellular toxicity of cadmium-containing semiconductor quantum dots was performed by applying random forest models to the 1,741 data samples manually collected from 307 relevant publications (Oh et al., 2016). The authors found that the toxicity induced by quantum dots strongly correlates with their intrinsic properties, such as diameter, surface ligand, shell, and surface modification.

The dataset of failed hydrothermal synthesis reactions collected in the course of the Dark Reactions

Project (see above) was used to explore synthesis routes for organically templated vanadium selenites and molybdates (Raccuglia et al., 2016). In particular, the authors applied support vector machine and decision tree models to define the upper/lower boundaries of the synthesis parameters that lead to formation of crystals from solution. The suggested synthesis routes were tested against real experiments and showed 89% success rate exceeding human intuition by 11%.

Although the manual approach to abstract a large text corpus is very laborious, it allows for obtaining high-quality data from the tables and figures as well as from the text, thus justifying the small size of these datasets. Nonetheless, a growing amount of research uses the automated TM pipelines to obtain a collection from which to initiate data-driven materials discoveries.

Young et al. (2018) developed a semi-automated TM pipeline to extract and analyze the growth conditions for four different oxide materials synthesized with pulsed laser deposition technique. They were able to obtain the range of growth temperatures and pressures, and predict the relative values of critical temperatures by applying a decision tree classifier.

Cooper et al. (2019) applied a TM pipeline to effectively screen and sort organic dyes for panchromatic solar cells. Their approach identified 9,431 dye candidates which were then narrowed down to five prospective molecules for experimental validation. This work is an important step toward a so-called "design-to-device" approach to fabrication of advanced materials (Cole, 2020). The approach consists of the four steps of i) data extraction from literature, ii) data augmentation with computations, iii) AI-guided materials design, and iv) experimental validation.

In other work, Court and Cole (2020) used the records of Curie and Néel temperatures text-mined from the scientific literature (Court and Cole, 2018) (see previous section) to reconstruct the phase diagrams of magnetic and superconducting materials. They used the materials bulk and structural properties as descriptors in ML models to predict the critical temperature for a magnetic phase transition. The trained models are formulated into a web application that provides multiple options for predicting and exploring magnetic and superconducting properties of arbitrary materials (www.magneticmaterials.org).

Our team has extensively used TM aiming to uncover insights about materials synthesis from scientific publications. Kim et al. (2017b) explored the parameters of hydrothermal and calcination reactions for metal oxides by analyzing the data extracted from 22,065 scientific publications. They found a strong correlation between the complexity of the target material and the choice of reaction temperature. A decision tree model applied to predict synthesis routes for titania nanotubes identified the concentration of NaOH and synthesis temperature as the most important factors that lead to nanotube formation. A similar approach was used to predict the density of germanium-containing zeolite frameworks and to uncover their synthesis parameters (Jensen et al., 2019).

In other work, Kim et al. (2017a) applied a variational autoencoder to learn the latent representation of synthesis parameters and to explore the conditions for the synthesis of TiO_2 brookite and for polymorph selection in the synthesis of MnO_2. Their results showed that the use of ethanol as a reaction medium

is sufficient but not necessary condition to form the brookite phase of TiO_2. Their latent representation of synthesis parameters also captures the requirement of alkalai ions for the generation of certain MnO_2 polymorph, consistent with *ab initio* findings (Kitchaev et al., 2017). A conditional variational autoencoder was also used to generate a precursors list for some perovskite materials (Kim et al., 2020a).

Building relations between materials, their properties and applications and combining them into a so-called *knowledge graph* structure is an emerging area of research in materials science that became enabled by the development of scientific TM. Onishi et al. (2018) implemented the Computer-Aided Material Design (CAMaD) system which is an elegant TM framework that reconstructs and visualizes a knowledge graph in the form of a process-structure-property-performance chart for desired materials. While the presented performance of the CAMaD system is still limited, it demonstrates the capabilities of TM to create a comprehensive knowledge-based structure that can be used for optimization of materials design.

The relation between materials reported in the different application areas of materials science was explored by Tshitoyan et al. (2019). They applied the word2vec model (Mikolov et al., 2013) to 3 million abstracts to learn a vectorized representation of words, and materials specifically. Interestingly, the model was able to not only learn some aspects of the chemistry underlying the relations between materials but also to draw a similarity between materials for different applications. In particular, it was demonstrated that such a cross-field correlation between the material properties required in different application could be used to predict novel thermoelectric materials. This work highlights an important aspect of scientific TM and NLP: its capability to uncover latent knowledge about a subject by comprehending a large amount of unstructured data – a task that is not possible for a human.

The question of materials similarity was also studied in recent work by He et al. (2020). In their work, a measure of similarity for synthesis precursors was defined by two parameters: i) the probability to substitute one precursor with another in the synthesis reaction for a common target material, and ii) the area of overlap of synthesis temperature distributions for two precursors. The results demonstrate that some of the empirical rules widely used by researchers when choosing the precursors for materials synthesis can be learned from text data.

4 Challenges and Caveats of the Text-Mining-Driven Research

While TM and NLP are tremendously promising tools to extract the enormous amount of information locked up in published research, several challenges for the approach remain. We categorize these below.

Lack of annotated data. The lack of a large dataset corresponding to a "gold standard" of annotated data significantly slows down the development of robust high-precision methods for chemical NER. The majority of the existing annotated sets have been created to serve a specific purpose or subfield of materials science and their broad application is not straightforward. Current attempts to create standardization for annotated data in materials science are limited to chemical named entities with emphasis on organic chemistry (Corbett

et al., 2007; Krallinger et al., 2015; Kim et al., 2003). Building more structured databases of experimental data that can be related to the papers from which the data are sourced, could potentially help to test the performance of NLP methods. One can even conceive creating machine-annotated data based on an existing relation between data and publications. We are however not hopeful that the scientific community can come together around central data deposition without an incentive structure from publishers or government agencies, which further stresses the important role that TM will have in generating large amounts of materials data.

Ambiguity and lack of standard nomenclature to describe and categorize complex materials. An engineering material is not merely a compound that requires a chemical description. It can be a doped system, inhomogeneous, a multi-phase system, or a composite. Each of these complexities comes with its morphology and length scale. While for common chemical terms, IUPAC provides nomenclature recommendations, writers usually prefer to simplify them or use arbitrary notations for materials names if no standard terminology is established. For instance, even for a basic concept such as a doped material, various nomenclatures are used e.g. "$Sc_2(MoO_4)_3$:Eu^{3+}", "$Sc_2(MoO_4)_3$ + x% Eu^{3+}" or "Eu^{3+}-doped $Sc_2(MoO_4)$". Composites and mixtures can be written in various ways (e.g. (1-x)$Pb(Zr_{0.52}Ti_{0.48})O_3$-x$BaTiO_3$ or $Pb(Zr_{0.52}Ti_{0.48})O_3$ + x wt% $BaTiO_3$). The abbreviated names of chemicals and materials (e.g. EDTA, BNT-BT-KNN, LMO) are also ubiquitous. Even within one journal or publisher no standards are applied. This complicates comparison and aggregation of extracted data across papers and requires substantial data post-processing in order to normalize and unify the results. In some cases it creates ambiguity that cannot be resolved, or whose resolution leads to errors.

Positive bias. Authors often "cherry-pick" data in the main body of a paper, either leaving out less successful data or moving it to supplementary information (which is often only available as PDF and with too low information content to do meaningful automated data extraction). This positive bias introduces substantial problems for ML models trained on these data, and requires caution when choosing the questions which one asks from ML models. In recent work, Jia et al. (2019) explored the effect of human bias in the choice of starting materials for the synthesis of metal organic crystals. They found a strong preference in the literature for selecting some reagents over others which was attributed to historically established rule-of-thumbs. In their explicit experimental testing they found no value of the implied precursor selection bias, something that a ML based on the published data would not have been able to resolve without additional data. In our own work on the prediction of novel compounds (Fischer et al., 2006; Hautier et al., 2011) or their synthesis methods (Kim et al., 2017b), the lack of negative information is severely limiting. For example, the lack of a known compound at a given composition in a complex phase diagram may mean that no compound exists at that composition, or, that nobody has looked carefully for it. These are very different pieces of input information for a ML model that tries to predict which compositions are compound forming or not. One can imagine that some researchers may have investigated the specific composition, but

because they did not find anything, the investigation was not reported. In a similar problem, failed synthesis experiments are rarely reported. This lack of negative data prevents one from capturing the boundaries on the space of possible ML outcomes. The effect of human bias on the quality of ML model predictions has not been investigated in detail and remains a challenging aspect of NLP-based data collections.

Language complexity and error accumulation. The narrative of a research paper is known to have a very specific style and language. It was shown for the corpus of newspapers of various subjects that the texts covering a scientific topic have the lowest readability score as compared to other topics, such as sports or weather (Flaounas et al., 2013). To explore the dependence between complexity of a scientific paragraph and the quality of the data extraction, we computed the categorical accuracy (fraction of predicted values that match with actual values) of data extraction for ~100 manually annotated paragraphs on materials synthesis and their corresponding Flesch-Kincaid grade level (FKGL) score (Kincaid et al., 1975). Figure 4b shows the obtained extraction accuracy of synthesis steps and material entities calculated using the NLP models developed by our team previously (Kononova et al., 2019; He et al., 2020) plotted against the corresponding FKGL score. Although the data are highly scattered, the negative correlation trend between the extraction accuracy and the FKGL score can be noticed. The computed Pearson correlation coefficients between the value of the FKGL score and the extraction accuracy of synthesis steps and materials entities are -0.42 and -0.38, respectively. It is worth noting that the correlation is stronger when the NLP model is applied to extract synthesis steps rather than materials entities. This can be explained with the fact that the context of a sentence defining a synthesis action is more ambiguous than that for materials terms (Kim et al., 2019). This complexity stresses the need to improve the general NLP tools to deal with scientific text. The accuracy of the text processing along the TM pipeline is crucial as errors usually accumulate from step to step, leading to a strong reduction in quality and size of the output (Kononova et al., 2019). As was noted before, the problem with sentence tokenization significantly affect the outcome of information extraction, in particular, chemical NER. Overcoming this problem may be possible by developing a hybrid NLP methods that introduces domain knowledge.

The accuracy of scientific NLP imposes constraints on the potential range of questions that the extracted data can address. Kauwe et al. (2019) have investigated the viability and fidelity of ML modeling based on a text-mined dataset. They used various ML algorithms and material structure models to predict the discharge capacity of battery materials after 25 cycles based on a dataset extracted from the literature and found inconclusive results. While one can speculate on the origin of this outcome, it is clear that the high level of uncertainty of the predictions can arise from invalid descriptors or models, as well as from the human bias and imperfectness of the experimental measurements (Kauwe et al., 2019). As the "no-free-lunch" theorem states, there is no any particular ML model that will work best for a given task. Therefore, interpretation of results obtained by application of ML algorithms to text mined data should always be treated with caution, and keeping the limitations of the input data in mind. In general, limitations of ML predictions are much

more likely to be caused by limitations of input data than by problem with the ML method.

5 Future Directions

Data is considered the 4$^{\text{th}}$ paradigm of science (Tolle et al., 2011). Access to a large amount of data allows the quantification and more accurate testing of hypothesis, and even potentially the machine learning of the relation between composition, structure, processing and properties of materials. The Materials Genome Initiative (MGI) (Holden, 2011) led to some highly successful data-driven research projects (e.g. www.mgi.gov, www.nsf.gov/funding/pgm_summ.jsp?pims_id=505073 and Jain et al. (2013) and Jain et al. (2016)). But the personal experience of one of the authors in helping launch MGI is that experimental data is unlikely to be collected one piece at a time, by having scientists enter it in databases, the way it was envisioned by some when MGI started. While ML is an exciting new direction for materials research, it is telling that much of published ML work is either on computed datasets (which can be generated with high-throughput computing) (Jain et al., 2011), or on very small experimental datasets, often containing no more than 50-100 data items. Because of this failure to collect experimental data in more organized ways, TM and NLP are likely to play a critical role in enabling more data-driven materials research. The willingness of publishers to share access to their large corpus for TM and several new developments in the NLP field are likely to lead to increased volume and quality of extracted information from scientific text.

The most notable advance in NLP in recent years has been the advent of *transformer models*, which have dramatically improved state-of-the-art performance on almost all benchmark tasks. The transformer utilizes an idea of sequence encoding-decoding (Bahdanau et al., 2016), and creates a latent vectorized representation of a text. The advantage of the model is its attention functionality (Vaswani et al., 2017) that allows for the model to recognize the key parts of a sequence that are crucial for understanding the meaning of text. The transformers have ushered in a new paradigm in NLP, whereby very large general-purpose models (with typically hundreds of millions of parameters) are pre-trained on publicly available corpora with unsupervised objective, before being fine-tuned to individual tasks. This so-called *transfer learning* approach allows the transformer to have high performance on supervised-training tasks with only a small number of training examples, significantly reducing the burden on human annotation.

From a materials science perspective, the transfer learning still meets some difficulties. The publicly available transformer models are pre-trained on general-purpose corpora, thus performing poorly on tasks involving scientific language. Moreover, the computational cost to train them "from scratch" is also significant: training $\text{BERT}_{\text{Large}}$ on a corpus of 3.3 billion words with 64 TPU cores took 4 days (Devlin et al., 2019). There have been a number of recent efforts to pre-train domain-specific transformer models on scientific text, including SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2019), and MedBERT (Rasmy et al., 2020). Although the corpus of available materials science publications (Figure 1) is of comparable size to the corpora used to train the original BERT models, no materials science-specific pre-trained BERT-style

model is publicly available to date. Training and release of such a model would be of tremendous impact for the materials science community.

Prominent progress has been also achieved for Neural Machine Translation (NMT), providing an opportunity to apply TM on scientific literature written in non-English languages. While NMT has reached parity with human translation in a number of languages (Hassan et al., 2018), the dominant methodology relies on supervised training on a large bilingual corpus with parallel texts in source and target languages. However, there are significant difficulties in implementing the parallel-translation approach tailored specifically to the peculiarities of the scientific text. The domain-specific vocabulary of scientific texts requires a significant bilingual corpora for training the parallel-translation model (Tehseen et al., 2018). The latest development in unsupervised NMT models (Lample et al., 2018; Artetxe et al., 2017; Lample et al., 2017) utilizes monolingual corpora, escaping the need for parallel texts. This opens possibilities for domain-specific training of the NMT and its application to the non-English scientific text.

As mentioned previously, the lack of large-scale annotated datasets often obstructs application of advanced NLP techniques for scientific TM. Crowd-sourcing for data collection may be a solution to this problem. Diverse approaches to collaborative data management have been widely used in projects such as OpenEI (www.openei.org), Folding@home (www.foldingathome.org) and others (Zhai et al., 2013; Doan et al., 2011), and have proven to be highly efficient for gathering a large amount of data. To date, only a few projects have utilized crowd-sourcing in materials science TM research (Young et al., 2018; Tchoua et al., 2016). But development of a collaborative data collection platform for application of NLP in materials science meets several challenges. First, building and maintenance of the software part requires a substantial labor investment one for which government science agencies do not seem quite ready for. Second, efficient data collection and annotation requires well established standards for labeling of scientific texts that can be unambiguously applied to a wide variety of research tasks.

The accelerated development of high-throughput computations and emergence of "big data" in materials science in the past few years has shifted focus towards data management and curation. This has resulted in engineering and production of high-quality databases with flexible graphical interfaces and programming APIs that provide facile and convenient access to the data for their mining and analysis (Alberi et al., 2018). Rapidly growing sets of the data extracted from scientific publications call for development of a similar advanced infrastructure for representations, maintenance and distribution of these data.

Prevalent, broad and accurate data is a pillar of science. It inspires, negates, or validates theories. In society and business, data has become a highly valued commodity from which to take strategic decision, construct more effective marketing campaigns, or to improve products. For materials science to fully benefit from the new data paradigm significantly more effort will need to be directed towards data collection. TM and NLP are clearly a tool to make the results of hundred years of materials research available toward the realization of this paradigm.

Acknowledgement

Funding to support this work was provided by the National Science Foundation under grant numbers 1922311, 1922372, and 1922090, the Office of Naval Research (ONR) Award #N00014-16-1-2432, the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under contract no. DE-AC02-05-CH11231 (D2S2 program KCD2S2), the Assistant Secretary of Energy Efficiency and Renewable Energy, Vehicle Technologies Office, U.S. Department of Energy under contract no. DE-AC02-05CH11231, and the Energy & Biosciences Institute through the EBI-Shell program (award nos. PT74140 and PT78473).

Author Contributions

Conceptualization, O.K., T.H., and H.H.; Investigation, O.K., T.H., H.H. and A.T.; Writing – Original Draft, O.K.; Writing – Review & Editing, O.K., A.T. and E.A.O; Funding Acquisition, G.C. and E.A.O; Supervision, G.C. All authors participated in the discussion and modification of the manuscript structure and text.

Declaration of Interests

The authors declare no competing interests.

Inclusion and Diversity

One or more of the authors of this paper self-identifies as a member of the LGBTQ+ community.

Lead Contact

Further information and requests for resources and data should be directed to the Lead Contact, Gerbrand Ceder (gceder@berkeley.edu).

References

Alberi, K. et al. (2018). "The 2019 materials by design roadmap". In: *J. Phys. D: Appl. Phys* 52.1, p. 013001. DOI: 10.1088/1361-6463/aad926.

Alperin, B.L., A.O. Kuzmin, L.Y. Ilina, V.D. Gusev, N.V. Salomatina, and V.N. Parmon (2016). "Terminology spectrum analysis of natural-language chemical documents: term-like phrases retrieval routine". In: *J. Cheminform.* 8, p. 22. DOI: 10.1186/s13321-016-0136-4.

Artetxe, M., G. Labaka, and E. Agirre (2017). "Learning bilingual word embeddings with (almost) no bilingual data". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 451–462. DOI: 10.18653/v1/P17-1042.

Azimi, S. M., D. Britz, M. Engstler, M. Fritz, and F. Mücklich (2018). "Advanced Steel Microstructural Classification by Deep Learning Methods". In: *Sci. Rep.* 8, p. 2128. DOI: 10.1038/s41598-018-20037-5.

Baevski, A., S. Edunov, Y. Liu, L. Zettlemoyer, and M. Auli (2019). "Cloze-driven Pretraining of Self-attention Networks". 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 5360–5369. DOI: 10.18653/v1/D19-1539.

Bahdanau, D., K. Cho, and Y. Bengio (2016). "Neural Machine Translation by Jointly Learning to Align and Translate". In: arXiv: 1409.0473 [cs.CL].

Beltagy, I, K. Lo, and A. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 3615–3620. DOI: 10.18653/v1/D19-1371.

Bird, S., E. Loper, and E. Klein (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Blei, D.M. (2012). "Probabilistic topic models". In: *Commun. ACM* 55, pp. 77–84. DOI: 10.1145/2133806.2133826.

Blei, D.M., A.Y. Ng, and M.I. Jordan (2003). "Latent dirichlet allocation". In: *J. Mach. Learn. Res.* 3, pp. 993–1022.

Blokhin, E. and P. Villars (2020). "The PAULING FILE Project and Materials Platform for Data Science: From big data toward materials genome". In: *Handbook of Materials Modeling: Methods: Theory and Modeling*, pp. 1837–1861. DOI: 10.1007/978-3-319-42913-7_62-1.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146. DOI: 10.1162/tacl_a_00051.

Bornmann, L. and R. Mutz (2015). "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references". In: *J. Assn. Inf. Sci. Tec.* 66, p. 2215. DOI: 10.1002/asi.23329.

Brown, T.B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). "Language models are few-shot learners". In: arXiv: 2005.14165 [cs.CL].

Chomsky, N. (1956). "Three models for the description of language". In: *IRE Trans. Inf. Theory* 2, pp. 113–124. DOI: 10.1109/TIT.1956.1056813.

Cole, J. M. (2020). "A Design-to-Device Pipeline for Data-Driven Materials Discovery". In: *Acc. Chem. Res.* 53, pp. 599–610. DOI: 10.1021/acs.accounts.9b00470.

Constantin, A., S. Pettifer, and A. Voronkov (2013). "PDFX: Fully-Automated PDF-to-XML Conversion of Scientific Literature". In: *Proceedings of the 2013 ACM Symposium on Document Engineering*. DocEng '13. Association for Computing Machinery, pp. 177–180. DOI: 10.1145/2494266.2494271.

Cooper, C. B. et al. (2019). "Design-to-Device Approach Affords Panchromatic Co-Sensitized Solar Cells". In: *Adv. Energy Mater.* 9, p. 1802820. DOI: 10.1002/aenm.201802820.

Corbett, P., C. Batchelor, and S. Teufel (2007). *Annotation of Chemical Named Entities*. Tech. rep., pp. 57–64.

Corbett, P. and A. Copestake (2008). "Cascaded classifiers for confidence-based chemical named entity recognition". In: *BMC Bioinformatics* 9.Suppl 11, S4. DOI: 10.1186/1471-2105-9-S11-S4.

Court, C. and J. M. Cole (2018). "Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction". In: *Sci. Data* 5, p. 180111. DOI: 10.1038/sdata.2018.111.

Court, C.J. and J.M. Cole (2020). "Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning". In: *npj Comput. Mater* 6, pp. 1–9. DOI: 10.1038/s41524-020-0287-8.

de Jong, M. et al. (2015). "Charting the complete elastic properties of inorganic crystalline compounds". In: *Sci. Data* 2, p. 150009. DOI: 10.1038/sdata.2015.9.

Devlin, J., M.-W Chang, K. Lee, and K. Toutanova (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: arXiv: 1810.04805 [cs.CL].

Dieb, T.M., M. Yoshioka, S. Hara, and M.C. Newton (2015). "Framework for automatic information extraction from research papers on nanocrystal devices". In: *Beilstein J. Nanotechnol.* 6, pp. 1872–1882. DOI: 10.3762/bjnano.6.190.

Doan, A., R. Ramakrishnan, and A. Y. Halevy (2011). "Crowdsourcing Systems on the World-Wide Web". In: *Commun. ACM* 54, pp. 86–96. DOI: 10.1145/1924421.1924442.

Eltyeb, S. and N. Salim (2014). "Chemical named entities recognition: A review on approaches and applications". In: *J. Cheminform.* 6, pp. 1–12. DOI: 10.1186/1758-2946-6-17.

Fischer, C. C., K J. Tibbetts, D. Morgan, and G. Ceder (2006). "Predicting crystal structure by merging data mining with quantum mechanics". In: *Nat. Mater.* 5, pp. 641–646. DOI: 10.1038/nmat1691.

Flaounas, I., O. Ali, T. Lansdall-Welfare, T. De Bie, N. Mosdell, J. Lewis, and N. Cristianini (2013). "Research Methods in the Age of Digital Journalism". In: *Digital Journalism* 1, pp. 102–116. DOI: 10.1080/21670811.2012.714928.

Friedrich, A., H. Adel, F. Tomazic, J. Hingerl, R. Benteau, A. Marusczyk, and L. 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*. Association for Computational Linguistics, pp. 1255–1268. DOI: 10.18653/v1/2020.acl-main.116.

Gallego, S.V., J.M. Perez-Mato, L. Elcoro, E.S. Tasci, R.M. Hanson, M.I. Aroyo, and G. Madariaga (2016a). "MAGNDATA: Towards a database of magnetic structures. II. The incommensurate case". In: *J. Appl. Cryst.* 49, pp. 1941–1956. DOI: 10.1107/S1600576716015491.

Gallego, S.V., J.M. Perez-Mato, L. Elcoro, E.S. Tasci, R.M. Hanson, K. Momma, M.I. Aroyo, and G. Madariaga (2016b). "MAGNDATA: towards a database of magnetic structures. I. The commensurate case". In: *J. Appl. Cryst.* 49, pp. 1750–1776. DOI: 10.1107/S1600576716012863.

García-Remesal, M., A. García-Ruiz, D. Pérez-Rey, D. De La Iglesia, and V. Maojo (2013). "Using nanoinformatics methods for automatically identifying relevant nanotoxicology entities from the literature". In: *Biomed. Res. Int.* 2013. DOI: 10.1155/2013/410294.

Gaultois, M.W., T.D. Sparks, C.K.H. Borg, R. Seshadri, W.D. Bonificio, and D.R. Clarke (2013). "Data-Driven Review of Thermoelectric Materials: Performance and Resource Considerations". In: *Chem. Mater.* 25, pp. 2911–2920. DOI: 10.1021/cm400893e.

Ghadbeigi, L., J.K. Harada, B.R. Lettiere, and T.D. Sparks (2015). "Performance and resource considerations of Li-ion battery electrode materials". In: *Energy Environ. Sci.* 8.6, pp. 1640–1650. DOI: 10.1039/C5EE00685F.

Gurulingappa, H., A. Mudi, L. Toldo, M. Hofmann-Apitius, and J. Bhate (2013). "Challenges in mining the literature for chemical information". In: *RSC Advances* 3, p. 16194. DOI: 10.1039/c3ra40787j.

Harris, Z.S. (1954). "Distributional structure". In: *Word* 10, pp. 146–162. DOI: 10.1080/00437956.1954.11659520.

Hassan, H. et al. (2018). *Achieving Human Parity on Automatic Chinese to English News Translation.* arXiv: 1803.05567 [cs.CL].

Hautier, G., C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder (2011). "Data Mined Ionic Substitutions for the Discovery of New Compounds". In: *Inorg. Chem.* 50, pp. 656–663. DOI: 10.1021/ic102031h.

Hawizy, L., D. M. Jessop, N. Adams, and P. Murray-Rust (2011). "ChemicalTagger: A tool for semantic text-mining in chemistry". In: *J. Cheminform.* 3, pp. 1–13. DOI: 10.1186/1758-2946-3-17.

He, T., W. Sun, H. Huo, O. Kononova, Z. Rong, V. Tshitoyan, T. Botari, and G. Ceder (2020). "Similarity of Precursors in Solid-state Synthesis as Text-Mined from Scientific Literature". In: *Chem. Mater.* 32, pp. 7861–7873. DOI: 10.1021/acs.chemmater.0c02553.

Hiszpanski, A.M., B. Gallagher, K. Chellappan, P. Li, S. Liu, H. Kim, B. Kailkhura, J. Han, D. Buttler, and T.Y.-J. Han (2020). "Nanomaterials Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge". In: *J. Chem. Inf. Model.* 60, pp. 2876–2887. DOI: 10.1021/acs.jcim.0c00199.

Hochreiter, S. and J. Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Comput.* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Holden, J. (2011). *Materials Genome Initiative for global competitiveness.* Tech. rep. National Science and Technology Council.

Honnibal, M. and M. Johnson (2015). "An Improved Non-monotonic Transition System for Dependency Parsing". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 1373–1378. DOI: 10.18653/v1/D15-1162.

Huang, L. and C. Ling (2019). "Representing Multiword Chemical Terms through Phrase-Level Preprocessing and Word Embedding". In: *ACS Omega* 4, pp. 18510–18519. DOI: 10.1021/acsomega.9b02060.

Huang, S. and J. M. Cole (2020). "A database of battery materials auto-generated using ChemDataExtractor". In: *Sci. Data* 7, pp. 1–13. DOI: 10.1038/s41597-020-00602-2.

Huo, H., Z. Rong, O. Kononova, W. Sun, T. Botari, T. He, V. Tshitoyan, and G. Ceder (2019). "Semi-supervised machine-learning classification of materials synthesis procedures". In: *npj Comput. Mater* 5, pp. 1–7. DOI: 10.1038/s41524-019-0204-1.

Jain, A., G. Hautier, C.J. Moore, S.-P. Ong, C.C. Fischer, T. Mueller, K.A. Persson, and G. Ceder (2011). "A high-throughput infrastructure for density functional theory calculations". In: *Computational Materials Science* 50, pp. 2295–2310. DOI: 10.1016/j.commatsci.2011.02.023.

Jain, A., K. A. Persson, and G. Ceder (2016). "Research Update: The materials genome initiative: Data sharing and the impact of collaborative ab initio databases". In: *APL Materials* 4, p. 053102. DOI: 10.1063/1.4944683.

Jain, A. et al. (2013). "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation". In: *APL Materials* 1, p. 011002. DOI: 10.1063/1.4812323.

Jensen, Z., E. Kim, S. Kwon, T.Z.H. Gani, Y. Roman-Leshkov, M. Moliner, A. Corma, and E. Olivetti (2019). "A machine learning approach to zeolite synthesis enabled by automatic literature data extraction". In: *ACS Cent. Sci.* 5, pp. 892–899. DOI: 10.1021/acscentsci.9b00193.

Jessop, D. M., S. E. Adams, E. L. Willighagen, L. Hawizy, and P. Murray-Rust (2011). "OSCAR4: a flexible architecture for chemical text-mining". In: *J. Cheminform.* 3, p. 41. DOI: 10.1186/1758-2946-3-41.

Jia, X., A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A.E. Ruby, H. Wang, S.A. Friedler, A.J. Norquist, et al. (2019). "Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis". In: *Nature* 573, pp. 251–255. DOI: 10.1038/s41586-019-1540-5.

Jurafsky, D. and J.H. Martin (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall Series in Artificial Intelligence. Pearson Prentice Hall. ISBN: 9780131873216.

Kauwe, S., T. Rhone, and T. Sparks (2019). "Data-Driven Studies of Li-Ion-Battery Materials". In: *Crystals* 9, p. 54. DOI: 10.3390/cryst9010054.

Kim, E., K. Huang, S. Jegelka, and E. Olivetti (2017a). "Virtual screening of inorganic materials synthesis parameters with deep learning". In: *npj Comput. Mater* 3, p. 53. DOI: 10.1038/s41524-017-0055-6.

Kim, E., K. Huang, O. Kononova, G. Ceder, and E. Olivetti (2019). "Distilling a Materials Synthesis Ontology". In: *Matter* 1, pp. 8–12. DOI: 10.1016/j.matt.2019.05.011.

Kim, E., K. Huang, A. Saunders, A. McCallum, G. Ceder, and E. Olivetti (2017b). "Materials synthesis insights from scientific literature via text extraction and machine learning". In: *Chem. Mater.* 29, pp. 9436–9444. DOI: 10.1021/acs.chemmater.7b03500.

Kim, E., K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, and E. Olivetti (2017c). "Machine-learned and codified synthesis parameters of oxide materials". In: *Sci. Data* 4, p. 170127. DOI: 10.1038/sdata.2017.127.

Kim, E. et al. (2020a). "Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks". In: *J. Chem. Inf. Model.* 60, pp. 1194–1201. DOI: 10.1021/acs.jcim.9b00995.

Kim, H., J. Han, and T.Y.-J. Han (2020b). "Machine Vision-Driven Automatic Recognition of Particle Size and Morphology in SEM Images". In: *Nanoscale* 12, pp. 19461–19469. DOI: 10.1039/D0NR04140H.

Kim, J.-D., T. Ohta, Y. Tateisi, and J. Tsujii (2003). "GENIA corpus – a semantically annotated corpus for bio-textmining". In: *Bioinformatics* 19.suppl_1, pp. i180–i182. DOI: 10.1093/bioinformatics/btg1023.

Kincaid, J. P., R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Tech. rep. Institute for Simulation and Training, University of Central Florida.

Kitchaev, Daniil A., Stephen T. Dacek, Wenhao Sun, and Gerbrand Ceder (2017). "Thermodynamics of Phase Selection in MnO2 Framework Structures through Alkali Intercalation and Hydration". In: *J. Am. Chem. Soc.* 139, pp. 2672–2681. DOI: 10.1021/jacs.6b11301.

Kleene, S. C. (1956). "Representation of Events in Nerve Nets and Finite Automata". In: Princeton: Princeton University Press, pp. 3–42. DOI: doi.org/10.1515/9781400882618-002.

Kolářik, C., R. Klinger, C. M. Friedrich, M. Hofmann-Apitius, and J. Fluck (2008). "Chemical Names: Terminological Resources and Corpora Annotation". In: *Workshop on Building and evaluating resources for biomedical text mining*, pp. 51–58.

Kononova, O., H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, and G. Ceder (2019). "Text-mined dataset of inorganic materials synthesis recipes". In: *Sci. Data* 6, pp. 1–11. DOI: 10.1038/s41597-019-0224-1.

Korvigo, I., M. Holmatov, A. Zaikovskii, and M. Skoblov (2018). "Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules". In: *J. Cheminform.* 10, p. 28. DOI: 10.1186/s13321-018-0280-0.

Krallinger, M., O. Rabal, F. Leitner, M. Vazquez, D. Salgado, et al. (2015). "The CHEMDNER corpus of chemicals and drugs and its annotation principles". In: *J. Cheminform.* 7, S2. DOI: 10.1186/1758-2946-7-S1-S2.

Krallinger, M., O. Rabal, A. Lourenço, J. Oyarzabal, and A. Valencia (2017). "Information Retrieval and Text Mining Technologies for Chemistry". In: *Chem. Rev.* 117, pp. 7673–7761. DOI: 10.1021/acs.chemrev.6b00851.

Kulkarni, C., W. Xu, A. Ritter, and R. Machiraju (2018). "An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 97–106. DOI: 10.18653/v1/N18-2016.

Kuniyoshi, Fusataka, Kohei Makino, Jun Ozawa, and Makoto Miwa (2020). *Annotating and Extracting Synthesis Process of All-Solid-State Batteries from Scientific Literature*. arXiv: 2002.07339 [cs.CL].

Kurgan, L.A. and P. Musilek (2006). "A survey of Knowledge Discovery and Data Mining process models". In: *The Knowledge Engineering Review* 21, pp. 1–24. DOI: 10.1017/S0269888906000737.

Lafferty, J., A. McCallum, and F. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282–289. ISBN: 1558607781.

Lample, G., A. Conneau, L. Denoyer, and M.A. Ranzato (2017). *Unsupervised Machine Translation Using Monolingual Corpora Only*. arXiv: 1711.00043 [cs.CL].

Lample, G., M. Ott, A. Conneau, L. Denoyer, and M. Ranzato (2018). "Phrase-Based & Neural Unsupervised Machine Translation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 5039–5049. DOI: 10.18653/v1/D18-1549.

Leaman, R., C.-H. Wei, and Z. Lu (2015). "tmChem: a high performance approach for chemical named entity recognition and normalization". In: *J. Cheminform.* 7, S3. DOI: 10.1186/1758-2946-7-S1-S3.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C.-H. So, and J. Kang (2019). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* 36, pp. 1234–1240. DOI: 10.1093/bioinformatics/btz682.

Li, X., X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li (2020). "Dice Loss for Data-imbalanced NLP Tasks". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 465–476. DOI: 10.18653/v1/2020.acl-main.45.

Liu, Z., Y. Lin, and M. Sun (2020). *Representation Learning for Natural Language Processing*. 1st ed. Springer Singapore. DOI: 10.1007/978-981-15-5573-2.

Lowe, D.M. and R.A. Sayle (2015). "LeadMine: a grammar and dictionary driven approach to entity recognition". In: *J. Cheminform.* 7, S5. DOI: 10.1186/1758-2946-7-S1-S5.

Luhn, H. P. (1958). "The Automatic Creation of Literature Abstracts". In: *IBM Journal of Research and Development* 2, pp. 159–165. DOI: 10.1147/rd.22.0159.

Luong, M.-T, T.D. Nguyen, and M.-Y. Kan (2010). "Logical Structure Recovery in Scholarly Articles with Rich Document Features". In: *Int. J. Digit. Libr. Syst.* 1, pp. 1–23. DOI: 10.4018/jdls.2010100101.

Mahdavi, M., R. Zanibbi, H. Mouchère, C. Viard-Gaudin, and U. Garain (2019). "ICDAR 2019 CROHME+ TFD: Competition on recognition of handwritten mathematical expressions and typeset formula detection". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, pp. 1533–1538. DOI: 10.1109/ICDAR.2019.00247.

Maksov, A., O. Dyck, K. Wang, K. Xiao, D.B. Geohegan, B.G. Sumpter, R.K. Vasudevan, S. Jesse, S.V. Kalinin, and M. Ziatdinov (2019). "Deep learning analysis of defect and phase evolution during electron beam-induced transformations in WS2". In: *npj Comput. Mater* 5, p. 12. DOI: 10.1038/s41524-019-0152-9.

Matson, T., M. Farfel, N. Levin, E. Holm, and C. Wang (2019). "Machine Learning and Computer Vision for the Classification of Carbon Nanotube and Nanofiber Structures from Transmission Electron Microscopy Data". In: *Microscopy and Microanalysis* 25, pp. 198–199. DOI: 10.1017/S1431927619001727.

Memon, J., M. Sami, R. A. Khan, and M. Uddin (2020). "Handwritten optical character recognition (OCR): a comprehensive systematic literature review (SLR)". In: *IEEE Access* 8, pp. 142642–142668. DOI: 10.1109/ACCESS.2020.3012542.

Mendenhall, T. C. (1887). "The Characteristic Curves of Composition". In: *Science* ns-9.214S, pp. 237–246. DOI: 10.1126/science.ns-9.214S.237.

Mikolov, T., I. Sutskever, K. Chen, G.S. Corrado, and J. Dean (2013). "Distributed representations of words and phrases and their compositionality". In: arXiv: 1310.4546 [cs.CL].

Milosevic, N., C. Gregson, R. Hernandez, and G. Nenadic (2019). "A framework for information extraction from tables in biomedical literature". In: *IJDAR* 22, pp. 55–78. DOI: 10.1007/s10032-019-00317-0.

Miner, G., J. Elder, A. Fast, T. Hill, R. Nisbet, and D. Delen (2012). *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Elsevier Science. ISBN: 9780123870117.

Morgan, D. and R. Jacobs (2020). "Opportunities and Challenges for Machine Learning in Materials Science". In: *Annu. Rev. Mater. Res.* 50. DOI: 10.1146/annurev-matsci-070218-010015.

Mouchère, H., R. Zanibbi, U. Garain, and C. Viard-Gaudin (2016). "Advancing the state of the art for handwritten math recognition: the CROHME competitions, 2011–2014". In: *IJDAR* 19, pp. 173–189. DOI: 10.1007/s10032-016-0263-5.

Mukaddem, K.T., E.J. Beard, B. Yildirim, and J.M. Cole (2020). "ImageDataExtractor: A Tool To Extract and Quantify Data from Microscopy Images". In: *Journal of Chemical Information and Modeling* 60, pp. 2492–2509. DOI: 10.1021/acs.jcim.9b00734.

Mysore, S., Z. Jensen, E. Kim, K. Huang, H.-S. Chang, E. Strubell, J. Flanigan, A. McCallum, and E. Olivetti (2019). "The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures". In: *LAW 2019 - 13th Linguistic Annotation Workshop, Proceedings of the Workshop*, pp. 56–64. arXiv: 1905.06939.

Mysore, S., E. Kim, E. Strubell, A. Liu, H.-S. Chang, S. Kompella, K. Huang, A. McCallum, and E. Olivetti (2017). "Automatically extracting action graphs from materials science synthesis procedures". In: arXiv: 1711.06872 [cs.CL].

Oh, E., R. Liu, A. Nel, K. B. Gemill, M. Bilal, Y. Cohen, and I.L. Medintz (2016). "Meta-analysis of cellular toxicity for cadmium-containing quantum dots". In: *Nature Nanotech.* 11, p. 479. DOI: 10.1038/nnano.2015.338.

Olivetti, E., J. Cole, E. Kim, O. Kononova, G. Ceder, T. Han, and A. Hiszpanksi (2020). "Data-driven Materials Research Enabled by Natural Language Processing". In: *Appl. Phys. Rev.* 7, p. 041317. DOI: 10.1063/5.0021106.

Onishi, T., T. Kadohira, and I. Watanabe (2018). "Relation extraction with weakly supervised learning based on process-structure-property-performance reciprocity". In: *Sci. Technol. Adv. Mater.* 19, pp. 649–659. DOI: 10.1080/14686996.2018.1500852.

Pennington, J., R. Socher, and C. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.

Peters, M., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202.

Raccuglia, P., K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist (2016). "Machine-learning-assisted materials discovery using failed experiments". In: *Nature* 533, pp. 73–76. DOI: 10.1038/nature17439.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). "Language models are unsupervised multitask learners". In: *OpenAI Blog* 1.8, p. 9.

Ramakrishnan, C., A. Patnia, E. Hovy, and G. Burns (2012). "Layout-aware text extraction from full-text PDF of scientific articles". In: *Source Code Biol. Med.* 7, p. 7. DOI: 10.1186/1751-0473-7-7.

Ramprasad, R., R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim (2017). "Machine learning in materials informatics: recent applications and prospects". In: *npj Comput. Mater* 3.1, pp. 1–13. DOI: 10.1038/s41524-017-0056-5.

Rasmy, L., Y. Xiang, Z. Xie, C. Tao, and D. Zhi (2020). *Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction.* arXiv: 2005.12833 [cs.CL].

Read, J., R. Dridan, S. Oepen, and L. J. Solberg (2012). "Sentence boundary detection: A long solved problem?" In: *Proceedings of COLING 2012: Posters*, pp. 985–994.

Ricci, F., W. Chen, U. Aydemir, G. J. Snyder, G.-M. Rignanese, A. Jain, and G. Hautier (2017). "An ab initio electronic transport database for inorganic materials". In: *Sci. Data* 4, p. 170085. DOI: 10.1038/sdata.2017.85.

Roberts, G., S.Y. Haile, R. Sainju, D.J. Edwards, B. Hutchinson, and Y. Zhu (2019). "Deep Learning for Semantic Segmentation of Defects in Advanced STEM Images of Steels". In: *Sci. Rep.* 9, p. 12744. DOI: 10.1038/s41598-019-49105-0.

Rocktäschel, T., M. Weidlich, and U. Leser (2012). "Chemspot: A hybrid system for chemical named entity recognition". In: *Bioinformatics* 28, pp. 1633–1640. DOI: 10.1093/bioinformatics/bts183.

Shah, S., D. Vora, B.P. Gautham, and S. Reddy (2018). "A Relation Aware Search Engine for Materials Science". In: *Integr. Mater. Manuf. Innov.* 7, pp. 1–11. DOI: 10.1007/s40192-017-0105-4.

Shannon, C. E. (1951). "Prediction and Entropy of Printed English". In: *Bell System Technical Journal* 30, pp. 50–64. DOI: 10.1002/j.1538-7305.1951.tb01366.x.

Swain, M. C. and J. M. Cole (2016). "ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature". In: *J. Chem. Inf. Model.* 56, pp. 1894–1904. DOI: 10.1021/acs.jcim.6b00207.

Szegedy, C., V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2016). "Rethinking the Inception Architecture for Computer Vision". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308.

Tchoua, R. B., J. Qin, D. J. Audus, K. Chard, I. T. Foster, and J. de Pablo (2016). "Blending Education and Polymer Science: Semiautomated Creation of a Thermodynamic Property Database". In: *J. Chem. Educ.* 93, pp. 1561–1568. DOI: 10.1021/acs.jchemed.5b01032.

Tchoua, R.B., A. Ajith, Z. Hong, L.T. Ward, K. Chard, A. Belikov, D.J. Audus, S. Patel, J.J. de Pablo, and I.T. Foster (2019). "Creating Training Data for Scientific Named Entity Recognition with Minimal Human Effort". In: *LNCS*. Ed. by J.M.F. Rodrigues, P.J.S. Cardoso, J. Monteiro, R. Lam, V.V. Krzhizhanovskaya, M.H. Lees, J.J. Dongarra, and P.M.A. Sloot. Vol. 11536. Springer International Publishing, pp. 398–411. DOI: 10.1007/978-3-030-22734-0_29.

Tehseen, I., G.R. Tahir, K. Shakeel, and M. Ali (2018). "Corpus Based Machine Translation for Scientific Text". In: *Artificial Intelligence Applications and Innovations*. Ed. by L. Iliadis, I. Maglogiannis, and V. Plagianakos. Cham: Springer International Publishing, pp. 196–206. DOI: 10.1007/978-3-319-92007-8_17.

Thompson, K. (1968). "Programming Techniques: Regular expression search algorithm". In: *Commun. ACM* 11, pp. 419–422. DOI: 10.1145/363347.363387.

Tkaczyk, D., P. Szostek, M. Fedoryszak, P.J. Dendek, and Ł. Bolikowski (2015). "CERMINE: automatic extraction of structured metadata from scientific literature". In: *International Journal on Document Analysis and Recognition (IJDAR)* 18, pp. 317–335. DOI: 10.1007/s10032-015-0249-8.

Tolle, K.M., D.S.W. Tansley, and A.J.G. Hey (2011). "The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]". In: *Proceedings of the IEEE* 99, pp. 1334–1337. DOI: 10.1109/JPROC.2011.2155130.

Trewartha, Amalie et al. (2020). *COVIDScholar: An automated COVID-19 research aggregation and analysis platform*. arXiv: 2012.03891 [cs.DL].

Tshitoyan, V., J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A Persson, G. Ceder, and A. Jain (2019). "Unsupervised word embeddings capture latent knowledge from materials science literature". In: *Nature* 571, pp. 95–98. DOI: 10.1038/s41586-019-1335-8.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L Kaiser, and I. Polosukhin (2017). "Attention is All You Need". In: arXiv: 1706.03762 [cs.CL].

Vaucher, A.C., F. Zipoli, J. Geluykens, V.H. Nair, P. Schwaller, and T. Laino (2020). "Automated extraction of chemical synthesis actions from experimental procedures". In: *Nat. Commun.* 11, p. 3601. DOI: 10.1038/s41467-020-17266-6.

Weizenbaum, J. (1983). "ELIZA – a computer program for the study of natural language communication between man and machine". In: *Commun. ACM* 26, pp. 23–28. DOI: 10.1145/357980.357991.

Weston, L., 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". In: *J. Chem. Inf. Model.* 59, pp. 3692–3702. DOI: 10.1021/acs.jcim.9b00470.

Young, S. R., A. Maksov, M. Ziatdinov, Y. Cao, M. Burch, J. Balachandran, L. Li, S. Somnath, R. M. Patton, S. V. Kalinin, et al. (2018). "Data mining for better material synthesis: The case of pulsed laser deposition of complex oxides". In: *J. Appl. Phys.* 123, p. 115303. DOI: 10.1063/1.5009942.

Zhai, H., T. Lingren, L. Deleger, Q. Li, M. Kaiser, L. Stoutenborough, and I. Solti (2013). "Web 2.0-Based Crowdsourcing for High-Quality Gold Standard Development in Clinical Natural Language Processing". In: *J. Med. Internet Res.* 15, e73. DOI: 10.2196/jmir.2426.

Figures

Figure 1. Publication trend over the past 14 years. *Top panel:* Number of publications appearing every year in different fields of materials science. All data were obtained by manually querying Web of Science publications resource. The analysis includes only research articles, communications, letters, and conference proceedings. The number of publications is on the order of 10^3. *Bottom panel:* Relative comparison of the fraction of scientific papers available on-line as image PDF or embedded PDF versus articles in HTML/XML format. The grey arrow marks time intervals for both top and bottom panels.

Figure 2. Schematic representation of the standard text mining pipeline for information extraction from the scientific publications.

Figure 3. Schematic representation of various information types that can be extracted from a typical materials science paper.

Figure 4. *Panel a:* Precision and recall of the published models for chemical NER manually extracted from the reports. Color denotes the primary algorithm underlying the model. *Panel b:* Accuracy of the data extracted from materials synthesis paragraphs plotted against the complexity of the paragraphs. The accuracy is computed using chemical NER models developed by our team (Kononova et al., 2019; He et al., 2020) to the manually annotated paragraphs. The text complexity is calculated as a Flesch-Kincaid grade level (FKGL) score indicating the education level required to understand the paragraph (Kincaid et al., 1975). ρ is a Pearson correlation coefficient between the accuracy of NER model and the FKGL score.

Tables

Data Repository	Documents Types	Access	Reference
CAplus	research articles, patents, reports	subscription	www.cas.org/support/documentation/references
DOAJ	research articles (open-access only)	public	doaj.org
PubMed Central	research articles	public	www.ncbi.nlm.nih.gov/pmc
Science Direct (Elsevier)	research articles	subscription	dev.elsevier.com/api_docs.html
Scopus (Elsevier)	abstracts	public	dev.elsevier.com/api_docs.html
Springer Nature	research articles, books chapters	subscription	dev.springernature.com/

Table 1: List of the some common text repositories in chemistry and material science subjects that provide an API for querying. *Note 1*: Elsevier provides API for both Science Direct (collection of Elsevier published full-text) and Scopus (collection of abstracts from various publishers). *Note 2*: Springer Nature provides access only to its own published full texts.

Reagents (NH4)2HPO4 and Sm2O3 were mixed	
NLTK	Reagents | **(** | **NH4** | **)** | **2HPO4** | and | Sm2O3 | were | mixed
SpaCy	Reagents | **(** | **NH4)2HPO4** | and | Sm2O3 | were | mixed
OSCAR4	Reagents | **(NH4)2HPO4** | and | Sm2O3 | were | mixed
ChemicalTagger	Reagents | **(NH4)2HPO4** | and | Sm2O3 | were | mixed
ChemDataExtractor	Reagents | **(NH4)2HPO4** | and | Sm2O3 | were | mixed
We made Eu2+-doped Ba3Ce(PO4)3 at 1200 °C for 2 h	
NLTK	We | made | **Eu2+-doped** | **Ba3Ce** | **(** | **PO4** | **)** | **3** | at | 1200 | **°C** | for | 2 | h
SpaCy	We | made | **Eu2** | **+** | **-doped** | **Ba3Ce(PO4)3** | at | 1200 | **°** | **C** | for | 2 | h
OSCAR4	We | made | **Eu2+** | **-** | **doped** | **Ba3Ce(PO4)3**| at | 1200 | **°C** | for | 2 | h
ChemicalTagger	We | made | **Eu2+-doped** | **Ba3Ce(PO4)3** | at | 1200 | **°C** | for | 2 | h
ChemDataExtractor	We | made | **Eu2+** | **-** | **doped** | **Ba3Ce(PO4)3** | at | 1200 | **°** | **C** | for | 2 | h
Lead-free a(Bi0.5Na0.5)TiO3-bBaTiO3-c(Bi0.5K0.5)TiO3 ceramics were investigated	
NLTK	**Lead-free** | **a** | **(** | **Bi0.5Na0.5** | **)** | **TiO3-bBaTiO3-c** | **(** | **Bi0.5K0.5** | **)** | **TiO3** | ceramics | was | investigated
SpaCy	**Lead** | **-** | **free** | **a(Bi0.5Na0.5)TiO3-bBaTiO3-c(Bi0.5K0.5)TiO3** | ceramics | was | investigated
OSCAR4	**Lead** | **-** | **free** | **a(Bi0.5Na0.5)TiO3-bBaTiO3-c(Bi0.5K0.5)TiO3** | ceramics | was | investigated
ChemicalTagger	**Lead-free** | **a(Bi0.5Na0.5)TiO3-bBaTiO3-c(Bi0.5K0.5)TiO3** | ceramics | was | investigated
ChemDataExtractor	**Lead-free** | **a(Bi0.5Na0.5)TiO3-bBaTiO3-c(Bi0.5K0.5)TiO3** | ceramics | was | investigated

Table 2: Examples of how different tokenizers split sentences into tokens. NLTK (Bird et al., 2009) and SpaCy (Honnibal and Johnson, 2015) are general-purpose tokenizing tools, whereas ChemDataExtractor (Swain and Cole, 2016), OSCAR4 (Jessop et al., 2011), ChemicalTagger (Hawizy et al., 2011) are the tools trained for a scientific corpus. Tokens are bound by "|" symbol.

An aqueous solution was prepared by dissolving lithium, cobalt, and manganese nitrates in de-ionized water	
NLTK	–
SpaCy	'manganese' (*nationalities or religious or political groups*)
OSCAR4	'aqueous', 'lithium', 'cobalt', 'manganese', 'nitrates', 'water'
tmChem	'lithium', 'cobalt', 'manganese nitrates'
ChemDataExtractor	'lithium', 'cobalt', 'manganese nitrates'
ChemSpot	'lithium', 'cobalt', 'manganese nitrates', 'water'
Bi-LSTM ChNER	'lithium, cobalt, and manganese nitrates', 'water'
A series of Ce3+-Eu2+ co-doped Ca2Si5N8 phosphors were successfully synthesized	
NLTK	–
SpaCy	–
OSCAR4	'Ce3+', 'Eu2+', 'Ca2Si5N8'
tmChem	'Ce3+-Eu2+', 'Ca2Si5N8'
ChemDataExtractor	'Ce3+-Eu2+', 'Ca2Si5N8'
ChemSpot	'Ce3+-Eu2', 'co', 'Ca2Si5N8'
Bi-LSTM ChNER	'Ce3+-Eu2+ co-doped Ca2Si5N8'
High-purity Bi(NO3)3 · 5H2O, Ni(NO3)2 · 6H2O and Cu(CH3COO)2 · H2O were used as starting materials for Bi2Cu1-xNixO4 powders	
NLTK	'NO3', 'NO3', 'CH3COO' (*organizations*); 'Ni', 'Cu' (*countries, cities, states*)
SpaCy	'Bi2Cu1-xNixO4' (*person*)
OSCAR4	'Bi(NO3)3 · 5H2O', 'Ni(NO3)2 · 6H2O', 'Cu(CH3COO)2 · H2O'
tmChem	'Bi(NO3)3 · 5H2O', 'Ni(NO3)2 · 6H2O', 'Cu(CH3COO)2 · H2O', 'Bi2Cu1-xNixO4'
ChemDataExtractor	'Bi(NO3)3 · 5H2O', 'Ni(NO3)2 · 6H2O', 'Cu(CH3COO)2 · H2O', 'Bi2Cu1-xNixO4'
ChemSpot	'Bi(NO3)3 · 5H2O', 'Ni(NO3)2 · 6H2O', 'Cu(CH3COO)2 · H2O', 'Bi2Cu1-xNixO4'
Bi-LSTM ChNER	'Bi(NO3)3 · 5H2O', 'Ni(NO3)2 · 6H2O', 'Cu(CH3COO)2 · H2O', 'Bi2Cu1-xNixO4'

Table 3: Examples of the chemical named entities extracted by the general-purpose NER tools NLTK (Bird et al., 2009) and SpaCy (Honnibal and Johnson, 2015), and the tools trained on chemical corpus OSCAR4 (Jessop et al., 2011), tmChem (Leaman et al., 2015), ChemSpot (Rocktäschel et al., 2012), ChemDataExtractor (Swain and Cole, 2016), Bi-LSTM chemical NER (He et al., 2020). For the general-purpose tools, the assigned labels are given in parenthesis. For the chemical NERs, only entities labeled as chemical compounds are shown.