An exploratory analysis: Extracting materials science knowledge from unstructured scholarly data

Journal:	*The Electronic Library*
Manuscript ID	EL-11-2020-0320.R2
Manuscript Type:	Article
Keywords:	Information science, Ontology, Knowledge, Digital scholarship, Materials science, Knowledge extraction

An exploratory analysis: Extracting materials science knowledge from unstructured scholarly data

Abstract

Purpose - The output of academic literature has increased significantly due to digital technology, presenting researchers with a challenge across every discipline, including materials science, as it is impossible to manually read and extract knowledge from millions of published literature. The research reported in this paper addresses this challenge by exploring knowledge extraction in materials science, as applied to digital scholarship. An overriding goal is to help inform readers about the status of knowledge extraction in materials science.

Design/methodology/approach – A two-part analysis was conducted, comparing knowledge extraction methods applied to materials science scholarship, across a sample of 22 articles; followed by a comparison of HIVE-4-MAT, an ontology-based knowledge extraction and MatScholar, a named entity recognition (NER) application. The paper covers contextual background, and a review of three tiers of knowledge extraction (ontology-based, NER, and relation extraction), followed by the research goals and approach. The discussion considers current knowledge extraction challenges in materials science.

Findings - The results indicate three key needs for researchers to consider for advancing knowledge extraction: 1) materials science-focused corpora, 2) for researchers to define the scope of the research being pursued, and 3) understand the trade-offs among different knowledge extraction methods. The paper also points to future materials science research potential with relation to extraction and increased availability of ontologies.

Originality - There are very few studies examining knowledge extraction in materials science. This work makes an important contribution to this underexplored research area.

Keywords: Information science, Ontology, Knowledge, Digital scholarship, Materials science, Knowledge extraction

Article classification: Research paper

1. Introduction

Scientific output has accelerated at an exponential pace due to digital technology. This trend is particularly noticeable through the significant growth in digital scholarly publications across every discipline (Khabsa and Giles, 2014; Ponte *et al*, 2017; Taubert and Weingart, 2017). The increase in digitally accessible scholarship is exciting, although this growth presents researchers with a daunting challenge as they seek to grapple with and benefit from the expanding corpus of information. This growing challenge is particularly prevalent in materials science (Mysore *et al*., 2017; Weston *et al*., 2019), where researchers seek to discover new recipes for improving materials performance.

To further explain the problem, digital scholarly publications, as a form of scholarly big data (Tuarob *et al*., 2016), generally describe the research design (method, sample, and treatment), report finding, and state conclusions. Specific to the case of materials science, results buried in a digital publication may help a researcher predict a material's performance for future work. For example, published results may report that "a specific recipe of magnesium, copper, and yttrium (Mg-Cu-Y), followed by heating and cooling these combined alloys at a designated temperature, results in a certain grade of metallic glass". The common approach for extracting this knowledge from digital text is extremely time-consuming. At a high level, a researcher needs to identify,

locate, and access relevant scholarly resources, such as peer reviewed journal publications, conferences papers, or patent documents, from the ever-expanding body of digital scholarship. Next, the researcher needs to read and manually extract key knowledge. Weston *et al*. (2019) emphasized that it is impossible for a researcher to read and extract knowledge contained in the vast, expanding store of published research, and referred to this problem as a bottleneck in materials discovery. This challenge underscores the need to explore machine driven approaches for extracting expert knowledge recorded in published research.

The research presented in this paper is motivated by both the *need* and the *opportunity* to advance knowledge extraction techniques for materials science scholarly resources. Interconnected to this goal is the need for researchers to understand levels of knowledge extraction, as well as the strengths, limitations, and potential application of these varied approaches for materials science. The research presented here aims to also contribute to this encompassing need. Specifically, the paper reports on a two-part comparative analysis: (1) an examination of knowledge extraction methods applied to scholarly materials science and reported on in journals, and (2) a comparison of ontology-based knowledge extraction and named entity recognition (NER), which are two of the most frequently applied knowledge extraction approaches across domains; both have high potential in helping to accelerate materials knowledge extraction. This work integrates with the larger goal of the NSF's *Harnessing the Data Revolution* (HDR) initiative, "Accelerating the discovery of electronic materials through human-computer active search" (www.nsf.gov/cise/harnessingdata/), where researchers across multiple disciplines are collaborating to harness knowledge buried in unstructured scholarly big data to assist with materials discovery.

The remainder of this paper is organized as follows. Section 2 defines materials science and provides an overview of the status of computational knowledge extraction in materials science. Then the paper presents a more in-depth examination of three tiers of knowledge extraction, followed by the research goals, approach, results, and a discussion. The paper wraps up a conclusion highlighting key findings as well as future directions.

2. Materials science and the status of computational knowledge extraction

Materials have a profound impact on our everyday lives. Consider the metals and plastics that comprise smartphones or medical implants, such as an artificial heart valve or pacemaker, both of which can both save and extend a human life. The history of materials science is founded in meeting society's needs for tools, drawing from metallurgy and mineralogy. Today, the discipline is defined as an interdisciplinary field, bringing together chemistry, engineering, and physics. Materials science researchers are particularly concerned with the performance of materials and study the relation between the structure, processing, and properties of materials. Materials science researchers are motivated by the overarching goal to discover and design materials with higher performance at lower cost than existing materials (Bao, 2017; Ding, 2014).

Computational approaches, including machine learning, in materials science have had impactful results for both inorganic (Ren *et al*., 2018) and organic materials design (Segler *et al*., 2018). The increase in computational approaches is motivated by the exponential growth in big data. Another motivating factor is the launch of the *Materials Genome Initiative* (MGI), with a mission to advance the materials science data infrastructure for computational design. Additionally, there is research supporting successful computationally designed materials in fields such as catalysis and thermoelectric (Kim *et al*., 2017a?). However, more comprehensive data is still needed for high-throughput computational materials design.

Knowledge extraction is one research area that connects to infrastructure needs, and researchers have also explored the use of word representation and sequence-to-sequence learning with neural networks. For example, researchers have explored different computational methods for extracting a synthesis procedure (Mysore *et al*., 2017); and results indicate that the performance of features from word embeddings can be even higher than features manually created by humans. To pave a way for further knowledge extraction approaches with deep learning, Tshitoyan *et al*. (2019) introduced their word embeddings generated the Word2Vec algorithm and trained on 3.3 million scientific abstracts of inorganic studies published between 1922 and 2018. Another example is Weston *et al*.'s (2019) recurrent neural network supporting the extraction of important information, such as synthesis methods, characterization methods, and material properties. Their work, MatScholar (https://matscholar.com), is a web interface that allows accurate knowledge retrieval. The model was built on Word2Vec word representation combined with a manually annotated dataset consisting of 800 abstracts.

Overall, although knowledge extraction in materials science is a recent development, research in this area shows promising results in materials design (Huang and Cole, 2020; Ren *et al*., 2018). Furthermore, these successes indicate there is likely a significant potential for computational approaches, including knowledge extraction, to advance materials discovery. The next section includes a review of three knowledge extraction approaches to contextualize the present work.

3. Knowledge extraction from textual data

Knowledge extraction, broadly speaking, involves the use of natural language processing (NLP) techniques to extract and correctly contextualize data or information, so that knowledge can be inferred. Three key knowledge extraction approaches include: 1) automatic indexing using ontological knowledge structures, 2) named entity recognition, and 3) relation extraction. These three approaches are reviewed below.

3.1 Ontology-based knowledge extraction

One of the simplistic forms of knowledge extraction is automatic indexing, which can include the application of a knowledge structure, such as an ontology. The process involves automatic indexing to extract key terms from a document; followed by matching these initial results to terms encoded with a selected knowledge structure, such as an ontology. Ontologies vary in their topical coverage by both breadth and depth, and they are considered valuable as they support semantic standardization and interoperability. Additionally, formalized ontologies encoded in the web ontology language (OWL) can support deductive reasoning and decision making.

Ontologies can be used to assist in automatic indexing. The process generally involves basic IR approaches, such as inverse document frequency (tf-idf), to determine the significance of a term or phrase in a document, and significant terms are mapped to a knowledge structure, such as an ontology, with similarity measures. Ontology-supported entity extraction has been popular in bioinformatics. One example of this is the well-known software MetaMap (Aronson, 2001), which maps keywords to ontologies, such as the National Library of Medicine's Unified Medical Language System (UMLS); a recent study using MetaMap shows that the majority of disease terms can still be correctly mapped (Senarath *et al*., 2020). Important to note for the purpose of the research presented in this paper, the advances noted above provide a model for other disciplines, such as materials science, where there is increased interest in ontologies. The HIVE-4-MAT prototype (https://hive4mat.cci.drexel.edu), explained further below, supports ontology-based

extraction and serves as a baseline tool for understanding additional knowledge extraction models, including approaches that support deep learning.

3.2 Named entity recognition

Named entity extraction (NER) is a more involved form of knowledge extraction, compared to the ontology-based approach. The term "named entity" first appeared in the sixth *Message Understanding Conference* (MUC-6) (Grishman and Sundheim 1996), where researchers gathered to share results on extracting structured information (entities) from unstructured text data. The focus was on entities involved detecting people's names, organizations, and geographical locations in text documents. In general, the term *named entity* in NLP refers to words that belong to designated semantic types, such as the organization mentioned above, and the task of NER is to recognize mentions of those rigid designators from text belongings.

Since the mid-1990s, early NER studies have mainly focused on extracting generic information from journal articles. Common datasets, including MUC-6 (Grishman and Sundheim, 1996), MUC-7 (Chinchor, 1998), CoNLL03 (Sang and De Meulder, 2003), ACE (Doddington *et al*., 2004), were widely used to test existing NER techniques. Later on, not restricted to generic semantic types, the interest in NER expanded to various domains, such as biology (Kim *et al*., 2003) and medicine (Doğan *et al*., 2014; Li *et al*., 2016). Weston *et al*. (2019) created an annotated dataset for inorganic materials which contains 800 manually annotated paper abstracts.

Named entity recognition is an important component of information extraction research, as well as an aspect of knowledge extraction. NER has a key role in a variety of applications, such as text summarization (Aone *et al*., 1999), information retrieval (Petkova and Croft, 2007), question-answering systems (Mollá *et al*., 2006), and knowledge base construction (Etzioni *et al*., 2005). As researchers seek to grapple with the massive volume of data, NER can also help in discovering new knowledge.

3.3 Relation extraction

A third area of knowledge extraction is relation extraction (RE). Similar to classic entities (person or organization name, place, and so forth), relation types are also an important component of human knowledge. Relational facts frequently appear in text along with their corresponding entities, forming a triple. As an example, the "University of Pennsylvania is located at Philadelphia" indicates the fact (*University of Pennsylvania*, organization-location, *Philadelphia*). Similarly, the fact (*Tim Cook*, person-title, *CEO*) can be inferred from the sentence "Tim Cook is the CEO of Apple, Inc". By extracting relational facts from unstructured texts, many applications, such as question-answering systems (Bordes *et al*., 2014) and knowledge base construction (Wang *et al*., 2014), are benefited; in a domain-specific area, relation extraction is widely used in drug discovery studies (Sang *et al*., 2018) to find drug-drug interactions (Liu *et al*., 2016; Quan *et al*., 2016), adverse drug events (Li *et al*., 2017), and protein-protein interactions (Hua and Quan, 2016).

The study of entity relation has gained researchers' attention since the 1990s. The most effective approaches have progressed from pattern recognition based on local syntax (Huffman, 1995), to feature-based approaches (Kambhatla, 2004), and most recently to deep learning (Jia *et al*., 2019; Miwa and Bansal, 2016). Advances in deep learning neural networks have been successful in extracting relations in various domains and have become state-of-the-art (Han *et al*., 2020). However, with the exception of a few studies (e.g., Court and Cole, 2018; Huang and Cole, 2020), relation extraction research in materials discovery is still quite limited.

The methods reviewed above, and the growing amount of digital scholarship in materials science help to motivate and shape the goals and objective of this research, presented in the next section.

4. Research goals and objectives

The overriding goal of this study is to gain a better understanding of the emerging research area of knowledge extraction in materials science. The research conducted is motivated by this goal and the need to address both the challenge and opportunity presented by growing amounts of scholarly big data.

Research objectives guiding this study are:

- Objective 1: To determine the status of knowledge extraction as applied to scholarly big data for materials science.
    - To address this objective, the researchers will identify the frequent and most effective knowledge extraction methods used in materials science, as well as current challenges.
- Objective 2: To compare two of the most frequently used knowledge extraction methods.

The work reported on here may also assist researchers in selecting appropriate knowledge extraction methods and inform future exploration and potential research directions.

5. Methodology

In order to address the above goals and objectives, a two-part comparative analysis was performed. The researchers first compared knowledge extraction methods reported in materials science literature. Then, they assessed the outputs of the following two knowledge extraction approaches: ontology-based knowledge extraction and named entity recognition (NER). Comparative analyses are useful for gaining insight into strengths and limitations of different approaches. The research design and procedures for this work is reported in the next section.

6. Research design and procedures

6.1 Knowledge extraction methods reported in materials literature

Part one of the study involved two steps. First, the researchers selected published journal articles reporting on knowledge extraction methods in materials science, and, second, they analysed the outcomes of the method. The journal article selection was based on chain-referral sampling, starting with the key article of Weston *et al*. (2019), and supplemented by conducting a keyword search for "knowledge extraction" in Google Scholar. Articles selected needed to satisfy two criteria: (1) it was required to report on the application of knowledge extraction methods specifically to scholarly big data in materials science and (2) it must appear in a key journal, such as *Nature*, *Physical Review Materials*, or *ACS Central Science*.

Although knowledge extraction has been around since the 1990s, the topic is still quite innovative in materials science, with articles appearing only within the last ten years. The selected sample reflected this status, and included 22 articles, published between 2017 and 2020. The sample allowed the researchers to gather data on the following three aspects of knowledge extraction as applied to materials science scholarly big data:

- The materials science themes, or subdomains, in which the knowledge extraction methods are being applied.
- The most frequently used knowledge extraction methods.

- The knowledge extraction approach that produced the best results.

As part of the data gathering task, the researchers also identified what seem to be current challenges and needs of materials researchers.

6.2 Knowledge extraction outcomes of the two main approaches

For this part of the study, the researchers focused on two of the main knowledge extraction approaches used across all domains. The following two applications were selected for comparison, given that they focus on materials science: 1) HIVE-4-MAT, which supports ontology-based knowledge extraction and 2) MatScholar, which supports NER-based knowledge extraction.

HIVE-4-MAT: HIVE uses the RAKE (rapid automatic keyword extraction) algorithm (Rose *et al*., 2010) for text processing, after which candidate keywords are mapped to ontologies. HIVE-4-MAT builds off the HIVE system incorporated into the DataNet Federation Consortium's iRODS system (Conway *et al*., 2013).

For the assessment, 60 abstracts were drawn from MatScholar and they were processed through the HIVE-4-MAT system using the following four ontologies for knowledge extraction: 1) BioAssay Ontology, 2) Library of Congress Subject Headings (LCSH), 3) Smart Appliances REFerence ontology (SAREF), and 4) US Geological Survey (USGS) terminology. The sample of 60 abstracts was sufficient for two reasons: (1) the keyword selection algorithm in HIVE-4-MAT is rule-based, which only uses the current input data (in the present case, the individual abstract) to make a decision, so sample size has limited impact on the performance; and (2) the empirical results from a pilot study reported on at the *Joint Conference on Digital Libraries* (Zhao *et al*., 2020) also showed no significant difference in performance based on ten abstracts, when compared to 60 abstracts. Knowledge extraction results were evaluated by information science and materials science experts using a three-tier scale of relevant (R), partially relevant (PR), and non-relevant (NR).

MatScholar: MatScholar is a web-accessible application that applies NER techniques to support knowledge extraction. The NER model uses the RNN-LSTM (recurrent neural network-long short term memory) structure. RNN-LSTM is a classic type of neural network that is widely applied in NLP tasks. MATScholar's NER algorithm is supported by a training set of 800 hand-annotated abstracts and uses colour and codes to identify seven entity classes: 1) inorganic material (MAT), 2) symmetry phase label (SPL), 3) sample descriptor (DSC), 4) material property (PRO), 5) material application (APL), 6) synthesis method (SMT), and 7) characterization method (CMT). A general assessment was conducted by processing sample abstracts provided in the system, through the NER feature in MatScholar.

7. Results

The results of the two-part comparative analysis are presented below. First, the researchers report on the assessment of existing methods reported in the materials science literature, followed by the assessment of the following two knowledge extraction approaches: ontology-based knowledge extraction and named entity recognition (NER).

7.1 Part I: Knowledge extraction methods reported in the materials science literature

Methods researchers are using: Knowledge extraction specifically targeting materials discovery, instead of overlapping fields like Chemistry or Biomedicine, is clearly an emerging field, with materials science-focused articles first appearing in 2017. As shown in Figure 1, of the 22

publications, 14 reported extraction structured materials knowledge from scholarly publications. Half of these 14 articles focus on how to extract needed entities; three studies covered both entity and relation extraction; two of them discussed how materials-related knowledge should be structured; the remaining two studies pointed out that more annotated corpora are needed to further advance knowledge extraction from the materials science literature.

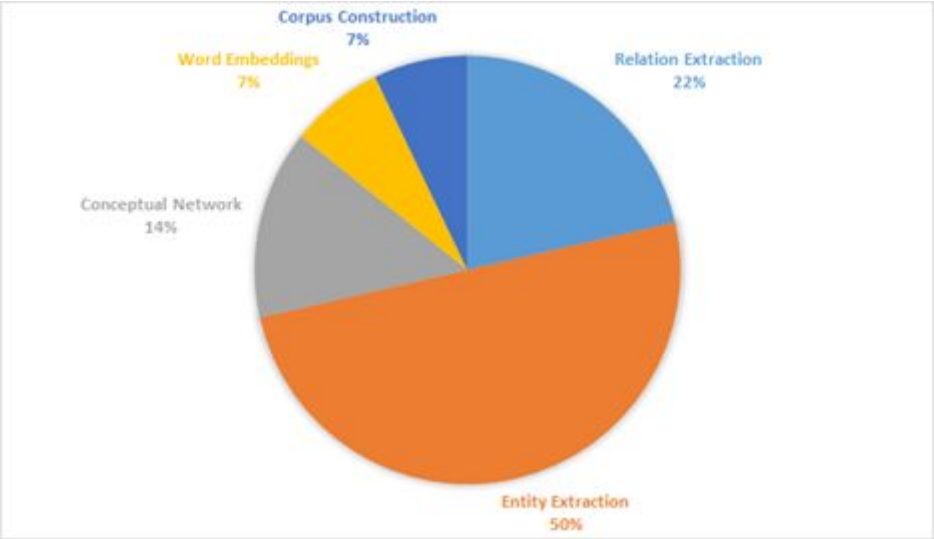

Figure 1. Distribution of reported knowledge extraction methods

Results verify that knowledge extraction has been effective across various topics (Figure 2). Results also confirm that different knowledge extraction methods support different goals. For example, entity extraction methods show high potential in synthesis design: six out of seven studies that involve entity extraction are on this topic. By extracting entities related to synthesis parameters and synthesis outcomes, the researchers were able to build a larger dataset to train machine learning models for synthesis outcome prediction (Kim *et al*., 2017).

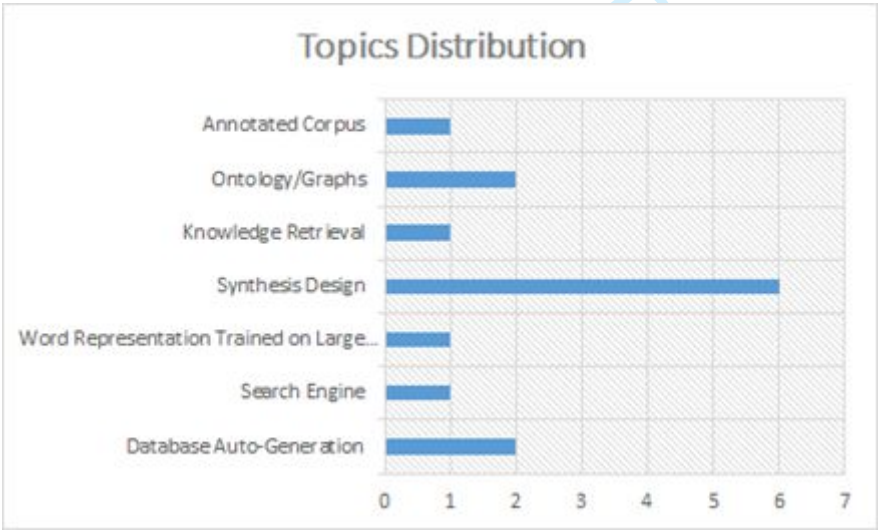

Figure 2. Topics distribution of knowledge extraction studies in materials science

Research applying relation extraction methods appears to have very different goals. One goal of relation extraction studies is database auto-generation. Materials science researchers commonly deal with quaternary, even quinary or more, relationships between materials and its corresponding properties. Two studies used relation extraction to auto-build databases for specific materials. Huang and Cole (2020) applied relation extraction methods to link battery materials with its five corresponding properties, such as conductivity and voltage. Similarly, Court and Cole (2018) presented their auto-generated database for magnetic materials, which also extracts materials and properties from large amounts of publications. This approach differs from manual database generation. Additionally, auto-generated databases usually have much larger data volume, which greatly helps in the computational materials design process.

Effective techniques to extract knowledge from texts: Results show a wide and varied use of pattern matching, external knowledge sources, and machine learning, with techniques combined at times. Based on the present analysis, the vocabulary needs of materials science are very specialized. For example, results showed that some properties and material names have unique patterns. Given this observation, many studies use techniques, such as regular expression, text matching to a self-defined vocabulary list, or databases, to identify/validate their candidate extracted terms. In addition to the above, results show that word representation is another way researchers extract key information from text, which is applied to accurate knowledge retrieval (Weston *et al*., 2019) and similar materials search (Kim *et al*., 2017).

7.2 Knowledge extraction assessment and comparison of the two main approaches

In this section, the researchers report their assessment and comparison of knowledge extraction approaches supported by HIVE-4-MAT and MatScholar.

7.2.1 HIVE-4-MAT assessment

The HIVE assessment, processing the 60 abstracts resulted in 987 extracted terms, of which 392 terms are evaluated as “Relevant”, 261 are marked as “Partially Relevant” and the rest 334 terms are evaluated as “Not Relevant” (Figure 3).

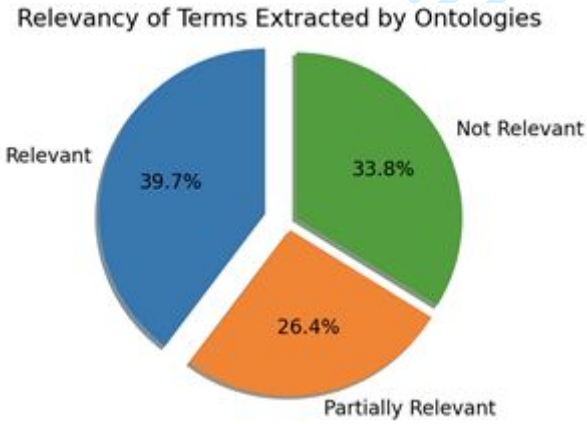

Figure 3. Relevance evaluation of terms extracted by HIVE-4-MAT

Figure 4. Relevant terms are in the upper portion and Partially Relevant in the lower

As shown in Figure 4, the upper portion of the figure and the lower portion show frequent terms that are marked as "R" and "PR", respectively. Since it is hard to determine whether a term is fully or partially relevant in specific abstracts, it can be observed that some terms appear in both categories. This way the "Relevant" and "Partially Relevant" categories were combined while measuring the percentage of relevant terms among all of the extracted terms. The results of the manual evaluation are summarized in Table I.

Table I. Results of manual evaluation of extracted terms

Evaluation results of HIVE-4-MAT	
Sample size (number of abstracts)	60
Number of extracted terms	987
Relevant terms	392
Partially Relevant terms	261
Not Relevant terms	334
Average terms extracted per abstract (range)	16.45 (5-30)
Percentage relevancy	66.16%

As shown in the Table I, 66.16 percent of the extracted terms are relevant, meaning that they accurately represent the abstract's content. The Not Relevant terms (e.g. terms extracted by the application but not relevant to the input abstracts) were also examined in Table II.

Table II. Analysis for Not Relevant (NR) terms

Failure analysis of false-positive extractions		
Error type	**Original text**	**Term extracted**
1. Extracted term is too broad	The deposition behaviour and physical properties of the films were investigated for use as an electrode in a metal-insulator-metal capacitor in future generation dynamic random access memory devices.	Investigation

2. Wrong terms identified	The Ru think films had a negligible oxygen content.	Contentment

Among all terms marked as "Not Relevant", most of them fall into two types: (1) the extracted terms are too broad to reflect the main idea of input abstracts; and (2) the HIVE-4-MAT identifies word roots, partial term matches for bound terms, and homographs. As shown in Table II, HIVE-4-MAT maps "content" to "contentment". This challenge is a result of stemming. While using ontologies to extract keywords from text documents, it is possible that an input term is mapped to homonyms (identical terms with the same spelling but with a different meaning).

7.2.2 Comparative analysis between ontology-based and NER-based extraction

To further analyse two different approaches, the sample abstract (Cho *et al*., 2002) described in Figure 5 was used as an input document to demonstrate different outputs from HIVE-4-MAT and MatScholar.

Sample Abstract

We investigated the effects of post-gate anneal and WN sputtering power on the gate dielectric integrity of W/WN/TaOxNy/SiO2/Si metal oxide semiconductor (MOS) capacitors. The process damage induced by physical vapor deposited metal gates in the high-permittivity (k) gate dielectric was partially relieved by a post-gate anneal. This is manifested by reduced leakage current, higher wear-out breakdown voltage, reduced charge trapping, and improved interface characteristics such as reduced hysteresis and interface state density (Dit). We observed a noticeable increase of charge trapping and interracial roughness at the WN/TaOxNy interface with WN power density while the Dit level remained similar. Degradation in the reliability characteristics with sputtering power density might be attributed to irrecoverable damage in the TaOxNy film.

Figure 5. Sample abstract

By mapping the sample abstract data to LCSH ontology, HIVE-4-MAT obtained the extraction output shown in Figure 6. All extracted terms are displayed at the left side of the output; by clicking on the terms at left, its corresponding information in the ontology will be listed at the right side. In Figure 6, taking the term "Dielectrics" as an example, the application is supposed to show its definition, broader concepts, narrower concepts, and related concepts at the right side. Based on the information stored in LCSH ontology, the term "Dielectrics" is a type of material related to electrical engineering (broader concept), and it is also related to narrower concepts, such as "Dielectric devices" and "Dielectric relaxation", which are bound terms (two words put together to form a term and represent a concept). The LCSH ontology does not contain the definition of the term "Dielectrics", so it is not displayed in the output.

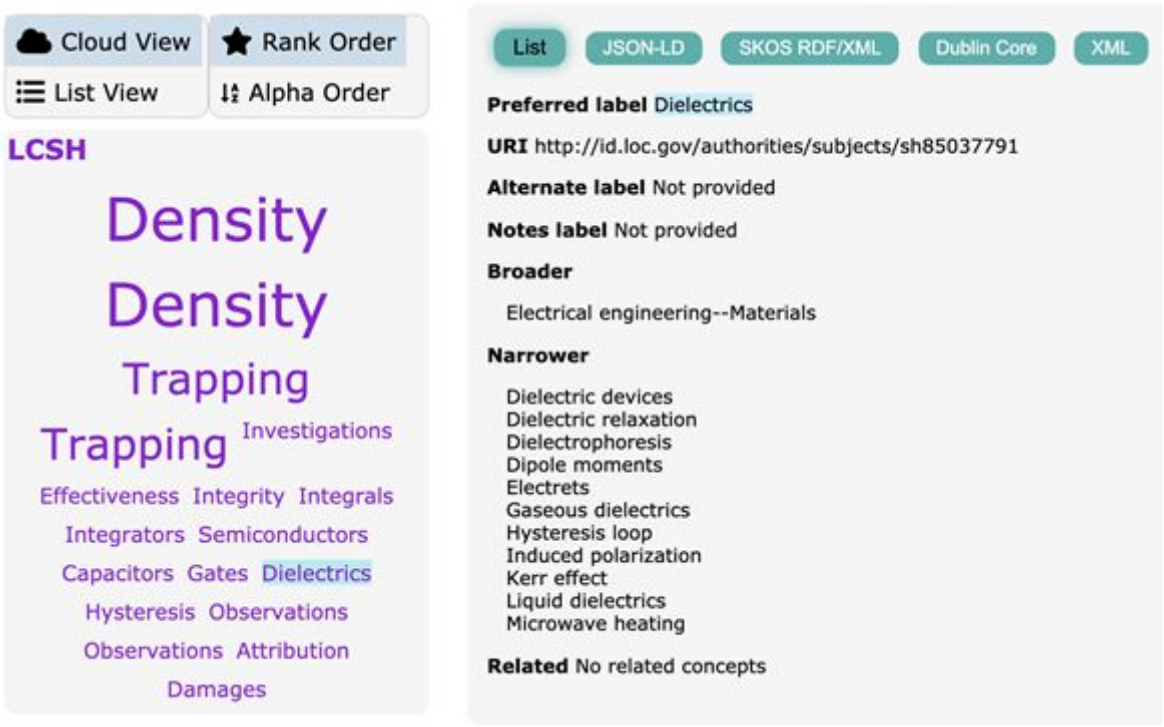

Figure 6. Sample extraction result from HIVE-4-MAT

The output example of HIVE-4-MAT shows that ontology-based knowledge extraction is good at providing further knowledge of extracted terms by capturing the full knowledge structure, which can help researchers to gain better understanding of the input literature.

Figure 7 shows the output of the same abstract from MatScholar (Weston *et al*., 2019).

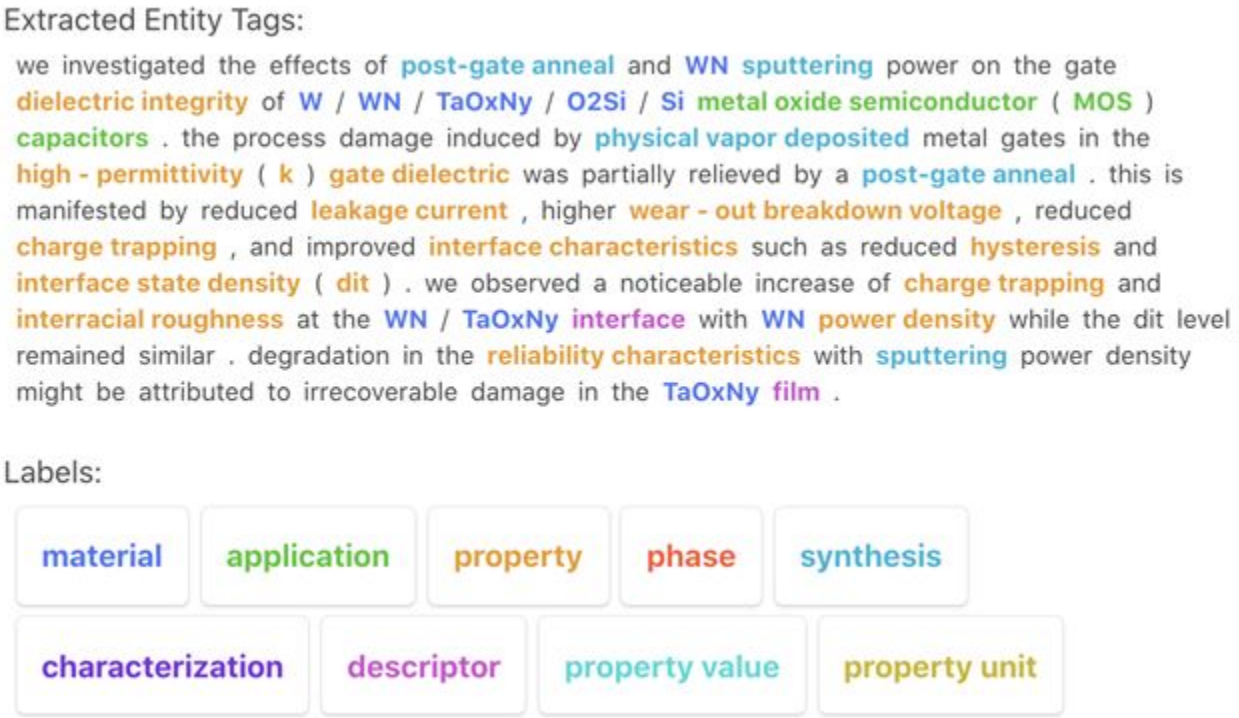

Figure 7. Sample extraction result from MatScholar

The MatScholar output in Figure 7 shows that their NER approach extracts terms related to the seven types (e.g. descriptor, synthesis method, etc.) described in the earlier section.

The HIVE-4-MAT and MatScholar comparison clearly presented four key differences that reflect pros and cons (see Table III). (1) **Terms in the extraction output:** The HIVE-4-MAT extraction output terms that are not always in the original input text (in the present case, the abstracts). Essentially, if there are terms in ontologies that are similar to input texts, the term will appear in the output, regardless of meaning. The case with a false match in meaning is demonstrated in the failure analysis above. MatScholar will only highlight exact words in original input texts. (2) **Semantic type classification:** MatScholar's NER model is designed to classify extracted key entities in seven different categories, whereas HIVE-4-MAT's ontology-based knowledge extraction only identifies knowledge terms by text matching, and does not further classify the semantic types. (3) **Related concepts and knowledge structure:** The HIVE-4-MAT embedded knowledge structure provides further information, such as concept relationships (e.g. broader concept, narrower concept, synonym, related subject, and so forth), and a user can traverse this structure and learn. MatScholar does not yet provide such information. (4) **Overall model differences:** While extracting key entities from input texts, HIVE-4-MAT relies on the content of its ontologies, whereas MatScholar relies on word representation and a deep learning model.

The reasons for the differences are addressed below in the second part of the discussion.

Table III. Comparison of different knowledge extraction approaches

	Pros	**Cons**
Ontology-based extraction	• Provides further knowledge of extracted terms by ontology structure • Large dataset for training model is not required	• Extraction performance heavily relies on the quality of ontologies • Provides less detailed extraction result
Named entity recognition	• Provides more accurate extraction result • Does not rely on external knowledge source as ontologies	• Requires large dataset for training model – some corpus has over billions of tokens

8. Discussion

The comparative analysis presented above gives insight into knowledge extraction methods used in materials scholarly data, as well as the difference of two widely-applied approaches. By conducting this research, the researchers hope their work can inspire more efforts to advance the knowledge extraction studies for materials scholarly data.

What are the current challenges? Part one of comparative results reveal two key challenges: (1) The absence of sufficient materials science-focused annotated/unannotated corpus necessary for training neural networks for knowledge extraction; (2) Limited examples of how knowledge should be structured in order to facilitate reproducibility and machine-readability. Knowledge extraction research using deep neural networks has achieved great success in many domains, particularly biomedicine (Li *et al*., 2017), given the availability of large, shared annotated

corpora. Training a neural network model requires millions of text documents, and a significant amount of time for annotation and fine tuning, and likely the limited availability of such resources is tied to the newness of this topic in materials science.

Materials science can learn from bioinformatics, which has a rich network of well-established ontologies supported by general agreements on the structure of knowledge. Ontologies covering materials science topics and reflecting key relationships are extremely limited. Researchers need to carefully identify and agree on the information that should be extracted and how to link related knowledge together, before starting the extraction process.

The second part of the present comparative analysis confirms the above findings. Ontology-based extraction, which is a straightforward rule-based approach, being used over decades, provides a general classified view without requiring a machine learning process and its performance is not affected by data size, whereas MatScholar's NER model using deep learning provides results with higher granularity and accuracy. As already stated, the availability of materials-based ontologies is currently limited and remains under-explored, and this very likely impacted HIVE-4-MAT's performance. In other words, more ontologies specific to the materials science domain would likely improve the results. Moreover, the quality of ontology can directly impact the performance and output of ontology-based extraction. Overall, there is a need for further work on materials science ontologies to assist with knowledge extraction.

These results prompt the question: *Should we only use one knowledge extraction approach instead of another?* Overall, the answer really depends on the circumstance of the study.

The effectiveness of rule-based approaches actually make a lot of sense in materials science. Because of the domain specificity of materials science, important information, such as chemical formulae, composites, synthesis methods, and material's names, usually have strong writing patterns or can be mapped to researcher-defined vocabulary lists (Court and Cole, 2018; Huang and Cole, 2020; Kim *et al*., 2017). In this case, using techniques such as text matching (e.g. vocabulary matching, suffix matching) and regular expression can achieve high accuracy scores without requiring a large dataset and training process. However, rule-based approaches are not able to perform more sophisticated extraction, such as finding materials occurring under similar semantic context, which can be performed by approaches involving word embeddings.

As a recently appeared method, word embeddings have also been proven to be effective for materials knowledge extraction (Kim *et al*., 2017b; Kim *et al*., 2017c; Weston *et al*., 2019). Although some studies have suggested that word embeddings could help further advance the accuracy of knowledge extraction (Kim *et al*., 2019), however, a large materials-specified corpus is required for training a word vector space. Furthermore, a deep learning model using word embeddings as features should be trained on annotated datasets for both NER and relation extraction tasks – data annotation for domain-specific tasks can be both computationally expensive and time prohibitive.

9. Conclusion

This study assessed the existing knowledge extraction methods in materials science, specifically applied to scholarly data. The research involved a two-part comparative analysis.

The conclusion highlights several important factors for researchers to consider when pursuing knowledge extraction in materials science. Research needs to consider the following:

- **The corpus/data availability:** Data size is a key factor that determines the performance of deep learning models. If there are sufficient volumes of corpus and annotated dataset

available, building a deep learning model to perform NER and even, and with a greater understanding of knowledge structures, RE may even be promising.

- **The scope of the research being pursued**: If the task is simply to extract entities that have a strong pattern and no further analysis is needed, then running rule-based approaches are both effective and can save a lot of time.
- **Trade-offs among methods:** There is a trade-off between the two examined approaches: NER-based extraction can extract more detailed terms but need extra training time, ontology-based extraction requires less preparation time and provides fairly acceptable output. If researchers seek higher accuracy with more details and can bear the longer preparation time, then recent NER techniques may be preferred; although as more materials science ontologies are produced, results from ontology-based extraction are likely to improve. Finally, if the dataset is small and researchers prefer high-level keywords, then the ontology-based approach is more ideal.

Overall, the results indicate that researchers should pick knowledge extraction methods based on their research goal and data availability. The results also suggest that ontologies can be helpful for knowledge discovery; although there is a need for more materials science-focused ontologies.

In terms of future research, this paper provides insight to more use of relation extraction in future materials science research. This can help with the creation of ontologies and support knowledge graph construction, leveraging digital scholarly literature. In conclusion, linking new materials science experiments and historical knowledge has high potential to accelerate the discovery of new materials.

Acknowledgements

The research reported on in this paper is supported, in part, by the U.S. National Science Foundation, Office of Advanced Cyberinfrastructure (NSF/OAC: #1940239 and #1940199). We also acknowledge the support of Cyra Gallano and Evan Dubrunfaut, Drexel University, for their role as data evaluators.

References

Aone, C., Okurowski, M.E., Gorlinsky, J. and Larsen, B. (1999), "A trainable summarizer with knowledge acquired from robust NLP techniques", in Mani, I. and Maybury, M. (Eds.), *Advances in Automatic Text Summarization*, MIT Press, Cambridge, MA, pp. 71-80.

Aronson, A.R. (2001), "Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program", in *Proceedings of the AMIA Symposium*, American Medical Informatics Association, p. 17.

Bao, J. (2017), "High-performance oxygen reduction and evolution carbon catalysis: From mechanistic studies to device integration", *Nano Research*, Vol. 10 No. 4, pp. 1163-1177, https://doi.org/10.1007/s12274-016-1347-8

Bordes, A., Chopra, S. and Weston, J. (2014), *Question Answering with Subgraph Embeddings*, arXiv preprint arXiv:1406.3676.

Chinchor, N.A. (1998), "*Overview of MUC-7/MET-2*", Science Applications International Corp., San Diego, CA.

Cho, H.J., Cha, T.H., Lim, K.Y., Park, D.G., Kim, J.Y., Kim, J.J., Heo, S., Yeo, I.S. and Park, J.W. (2002), "Reliability characteristics of W/WN/TaO x Ny/SiO2/Si metal oxide

semiconductor capacitors", *Journal of the Electrochemical Society*, Vol. 149 No. 7, p. G403.

Conway, M.C., Greenberg, J., Moore, R., Whitton, M. and Zhang, L. (2013, November), "Advancing the DFC semantic technology platform via HIVE innovation", in *Research Conference on Metadata and Semantic Research*, Springer, Cham, pp. 14-21.

Court, C.J. and Cole, J.M. (2018), "Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction", *Scientific Data*, Vol. 5 No. 1, pp. 1-12.

Ding, L. (2014), "Enhancing SOFC cathode performance by surface modification through infiltration", *Energy & Environmental Science*, Vol. 7 No. 2, pp. 552-575, https://doi.org/10.1039/c3ee42926a

Doddington, G.R., Mitchell, A., Przybocki, M.A., Ramshaw, L.A., Strassel, S.M. and Weischedel, R.M. (2004, May), "The automatic content extraction (ace) program-tasks, data, and evaluation", in *LREC*, Vol. 2 No. 1, pp. 837-840.

Doğan, R.I., Leaman, R. and Lu, Z. (2014), "NCBI disease corpus: A resource for disease name recognition and concept normalization", *Journal of Biomedical Informatics*, Vol. 47, pp. 1-10.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S. and Yates, A. (2005), "Unsupervised named-entity extraction from the web: An experimental study", *Artificial Intelligence*, Vol. 165 No. 1, pp. 91-134.

Greenberg, J., Zhao, X., Adair, J., Boone, J. and Hu, X.T. (2021), *HIVE-4-MAT: Advancing the Ontology Infrastructure for Materials Science*, arXiv preprint arXiv:2101.07960.

Grishman, R. and Sundheim, B.M. (1996), "Message Understanding Conference-6: A brief history", in *COLING '96 Volume 1: The 16th International Conference on Computational Linguistics*, pp. 466-471.

Han, X., Gao, T., Lin, Y., Peng, H., Yang, Y., Xiao, C., Liu, Z., Li, P., Sun, M. and Zhou, J. (2020), *More Data, More Relations, More Context and More Openness: A Review and Outlook for Relation Extraction*, arXiv preprint arXiv:2004.03186.

Hua, L. and Quan, C. (2016), "A shortest dependency path based convolutional neural network for protein-protein relation extraction", *BioMed Research International*, Vol. 2016, Article ID 8479587, https://doi.org/10.1155/2016/8479587

Huang, S. and Cole, J.M. (2020), "A database of battery materials auto-generated using ChemDataExtractor", *Scientific Data*, Vol. 7 No. 1, pp. 1-13.

Huffman, S.B. (1995, August), "Learning information extraction patterns from examples", in *International Joint Conference on Artificial Intelligence*, Springer, Berlin, Heidelberg, pp. 246-260.

Jia, R., Wong, C. and Poon, H. (2019), *Document-Level N-ary Relation Extraction with Multiscale Representation Learning*, arXiv preprint arXiv:1904.02347.

Kambhatla, N. (2004, July), "Combining lexical, syntactic, and semantic features with maximum entropy models for information extraction", in *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 178-181.

Khabsa, M. and Giles, C.L. (2014), "The number of scholarly documents on the public web", *PLoS One*, Vol. 9 No. 5, p. e93949.

Kim, E., Huang, K., Jegelka, S. and Olivetti, E. (2017b), "Virtual screening of inorganic materials synthesis parameters with deep learning", *NPJ Computational Materials*, Vol. 3, pp. 1-9.

Kim, E., Huang, K., Kononova, O., Ceder, G. and Olivetti, E. (2019), "Distilling a materials synthesis ontology", *Matter*, Vol. 1, pp. 8-12.

Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G. and Olivetti, E. (2017a), "Materials synthesis insights from scientific literature via text extraction and machine learning", *Chemistry of Materials*, Vol. 29 No. 21, pp. 9436-9444.

Kim, E., Huang, K., Tomala, A., Matthews, S., Strubell, E., Saunders, A., McCallum, A. and Olivetti, E. (2017c), "Machine-learned and codified synthesis parameters of oxide materials", *Scientific Data*, Vol. 4, pp. 1-9.

Kim, J.D., Ohta, T., Tateisi, Y. and Tsujii, J.I. (2003), "GENIA corpus – A semantically annotated corpus for bio-text mining", *Bioinformatics*, Vol. 19 Suppl. 1, pp. i180-i182.

Li, F., Zhang, M., Fu, G. and Ji, D. (2017), "A neural joint model for entity and relation extraction from biomedical text", *BMC Bioinformatics*, Vol. 18 No. 1, pp. 1-11.

Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C.H., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C. and Lu, Z. (2016), "BioCreative V CDR task corpus: A resource for chemical disease relation extraction", *Database*, 2016.

Liu, S., Tang, B., Chen, Q. and Wang, X. (2016), "Drug-drug interaction extraction via convolutional neural networks", in *Computational and Mathematical Methods in Medicine*.

Miwa, M. and Bansal, M. (2016), *End-to-end Relation Extraction Using LSTMs on Sequences and Tree Structures*, arXiv preprint arXiv:1601.00770.

Mollá, D., van Zaanen, M. and Smith, D. (2006), "Named entity recognition for question answering", in Cavedon, L. and Zukerman, I. (Eds.), *Proceedings of the Australasian Language Technology Workshop (ALTA '06)*, pp. 51-58.

Mysore, S., Kim, E., Strubell, E., Liu, A., Chang, H.S., Kompella, S., Huang, K., McCallum, A. and Olivetti, E. (2017), *Automatically Extracting Action Graphs from Materials Science Synthesis Procedures*, arXiv preprint arXiv:1711.06872.

Petkova, D. and Croft, W.B. (2007, November), "Proximity-based document representation for named entity retrieval", in *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, pp. 731-740).

Ponte, D., Mierzejewska, B.I. and Klein, S. (2017), "The transformation of the academic publishing market: multiple perspectives on innovation", *Electronic Markets*, Vol. 27 No. 2, pp. 97-100.

Quan, C., Hua, L., Sun, X. and Bai, W. (2016), "Multichannel convolutional neural network for biological relation extraction", *BioMed Research International*, Vol. 2016.

Ren, F., Ward, L., Williams, T., Laws, K.J., Wolverton, C., Hattrick-Simpers, J. and Mehta, A. (2018), "Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments", *Science Advances*, Vol. 4 No. 4, p. eaaq1566.

Rose, S., Engel, D., Cramer, N. and Cowley, W. (2010), "Automatic keyword extraction from individual documents", in *Text Mining: Applications and Theory*, pp. 1-20.

Sang, E.F. and De Meulder, F. (2003), *Introduction to the CoNLL-2003 Shared Task: Language-independent Named Entity Recognition*, arXiv preprint cs/0306050.

Sang, S., Yang, Z., Wang, L., Liu, X., Lin, H. and Wang, J. (2018), "SemaTyP: A knowledge graph based literature mining method for drug discovery", *BMC Bioinformatics*, Vol. 19 No. 1, pp. 1-11.

Segler, M.H., Preuss, M. and Waller, M.P. (2018), "Planning chemical syntheses with deep neural networks and symbolic AI", *Nature*, Vol. 555 No. 7698, pp. 604-610.

Senarath, Y., Chan, J., Purohit, H. and Uzuner, O. (2020), *Evaluating the Relevance of UMLS Concepts for Public Health Informatics during Disasters using MetaMap*.

Smith, B. and Ceusters, W. (2010), "Ontological realism: A methodology for coordinated evolution of scientific ontologies", *Applied Ontology*, Vol. 5 Nos. 3-4, pp. 139-188.

Taubert, N.C. and Weingart, P. (2017), "Changes in scientific publishing: A heuristic for analysis", in *The Future of Scholarly Publishing: Open Access and the Economics of Digitisation*, African Books Collective, pp. 1-34.

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K.A., Ceder, G. and Jain, A. (2019), "Unsupervised word embeddings capture latent knowledge from materials science literature", *Nature*, Vol. 571 No. 7763, pp. 95-98.

Tuarob, S., Bhatia, S., Mitra, P. and Giles, C.L. (2016), "AlgorithmSeer: A system for extracting and searching for algorithms in scholarly big data", *IEEE Transactions on Big Data*, Vol. 2 No. 1, pp. 3-17.

Wang, Z., Zhang, J., Feng, J. and Chen, Z. (2014, June), "Knowledge graph embedding by translating on hyperplanes", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 28, No. 1, pp. 1112-1119.

Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K., Ceder, G., Jain, A. (2019), "Named entity recognition and normalization applied to large-scale information extraction from the materials science literature", *Journal of Chemical Information and Modeling*, Vol. 59 No. 9, pp. 3692-3702.

Zhao, X., Greenberg, J., Hu, X., Nilsen, V. and Toberer, E. (2020), "Scholarly big data: Computational approaches to semantic labeling in materials science", presented at *ACM/IEEE Joint Conference on Digital Libraries Workshop 4: Organizing Big Data, Information, and Knowledge*.