2022

Band gap information extraction from materials science literature – a pilot study

Satanu Ghosh
University of New Hampshire, satanu.ghosh@unh.edu

Kun Lu
University of Oklahoma

Comments

This is an accepted manuscript published by Emerald Publishing in 2022 in Aslib Journal of Information Management, available online: https://doi.org/10.1108/AJIM-03-2022-0141

Recommended Citation

Ghosh, S. and Lu, K. (2022), "Band gap information extraction from materials science literature – a pilot study", Aslib Journal of Information Management, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/AJIM-03-2022-0141

Band Gap Information Extraction from Materials Science Literature - A Pilot Study

Journal:	*Aslib Journal of Information Management*
Manuscript ID	AJIM-03-2022-0141.R3
Manuscript Type:	Research Paper
Keywords:	band gap information extraction, photovoltaic cell, solar cell, renewable energy, text mining, academic text

Band Gap Information Extraction from Materials Science Literature - A Pilot Study

Abstract

Purpose — The purpose of this paper is to present our preliminary work on extracting band gap information of materials from academic papers. With increasing demand for renewable energy, band gap information will help material scientists design and implement novel photovoltaic (PV) cells.

Design/methodology/approach — We collected 1.44 million titles and abstracts of scholarly articles related to materials science, and then filtered the collection to 11,939 articles that potentially contain relevant information about materials and their band gap values. ChemDataExtractor was extended to extract information about PV materials and their band gap information. Evaluation was performed on randomly sampled information records of 415 papers.

Findings — Our findings show that the current system is able to correctly extract information for 51.32% articles, with partially correct extraction for 36.62% articles and incorrect for 12.04%. We have also identified the errors belonging to three main categories pertaining to chemical entity identification, band gap information, and interdependency resolution. Future work will focus on addressing these errors to improve the performance of the system.

Originality — We did not find any literature to date on band gap information extraction from academic text using automated methods. This work is unique and original. Band gap information is of importance to materials scientists in applications such as solar cells, light emitting diodes (LED), and laser diodes.

Keywords — band gap information extraction; photovoltaic cell; solar cell; renewable energy; text mining; academic text; ChemDataExtractor.

Introduction

Global energy consumption is expected to increase nearly 50% by 2050 as a result of economic and population growth according to the U.S. Energy Information Administration (Nalley & LaRose, 2021). Renewable and clean energy needs to play a bigger role in meeting the demand due to the growing evidence of climate change that has been associated with recent catastrophic events across the world (Pidcock & McSweeney, 2021). Photovoltaic (PV) materials can convert light into electricity, which allows us to harness the abundant clean solar energy. Existing PV materials have significant drawbacks in efficiencies, containing toxic metal, and/or relying on scarce elements (Todorov *et al.*, 2010; Mitzi, *et al.*, 2011; Saparov & Mitzi, 2016; Correa-Baena, *et al*., 2017). To address the problems, novel PV materials are needed. However, traditional avenues for the discovery and implementation of energy materials are inefficient, partially due to the reliance on trial-and-error methods. Recent advances in data-driven approaches offer new opportunities for more efficient materials design and discovery.

Photovoltaic (PV) materials can convert light energy to electric current due to a physical phenomenon called "photovolatic effect." For a PV material to convert light into electricity, photons in the light need to carry enough energy to excite electrons in the material into a free state to create electric current. Band gap is the minimum amount of energy required to excite an electron in a material into such a free state. Band gap is an intrinsic property of materials. Materials with too high band gaps are not suited for PV cells because photons will not have enough energy to excite the electrons in these materials. On the other hand, materials with too low band gaps are not ideal for PV cells either, because photons will carry excessive energy for exciting the electrons and the extra energy will be converted to heat, which is undesired. Knowing the band gap information is very important for material scientists to determine candidate materials for PV cells. This information has been widely reported in scientific literature from experimental and computational studies, and continues to appear in upcoming publications, but the volume of the literature prevents scientists from gaining a complete view of the band gaps of various materials. Manually collecting this information has been attempted (e.g. Kasap, 2006), but is inefficient and unable to keep up with the ever-increasing volume of scientific literature. As a result, most scientific decisions are made based on partial information, which can lead to missed opportunities for discovering novel solar materials.

This study develops an automated method to extract band gap information from materials science literature. The method is evaluated based on the extraction results on a random sample of 415 articles from a collection of 11,939 materials science articles potentially containing band gap information. Text mining for materials science is still in its early stage (Kononova, *et al.*, 2021). The closest tool available for extracting such chemical information from scientific literature is ChemDataExtractor, which was developed to extract spectroscopic attributes and experimental properties (Swain & Cole, 2016), and recently extended to extract material properties relevant to battery materials (Huang & Cole, 2020). We extend the ChemDataExtractor tool to extract band gap information. Machine-learning-based approaches could also be used to extract information from text. However, training data is very scarce in the materials science domain, and no specific training data can be found for the band gap information extraction task.

As far as we know, no existing study has developed automated methods to extract band gap information from materials science literature. This study aims to fill this gap. In addition to solar cells, the band gap information is also useful for other applications, such as Light emitting diodes and laser diodes.

Related Work

Text Mining for Scientific Literature

Generation of new knowledge is of utmost importance for scientific progress[1]. Scientific publications remain to be the primary channel for scientists to communicate new ideas and discoveries. As the volume of scientific publication continues to grow rapidly, it has become increasingly challenging for scientists to keep up with the latest development in the field. This can lead to suboptimal decisions based on incomplete information. Text mining relies on natural language processing techniques and/or manually curated ontologies to analyze large amounts of text automatically in order to offer more efficient ways for scientists to harness the existing knowledge in scientific literature. This may involve extracting (Mooney & Bunescu, 2005), summarizing (Nenkova & McKeown, 2012), aggregating (Serrano *et al.*, 2013), categorizing (Brindha, Prabha & Sukumaran, 2016) and inferring (Erraguntla *et al.*, 2012) information from text. In addition, by analyzing and synthesizing what has been reported in the literature, literature-based discoveries may also be achieved (Gordan & Dumais, 1998). Information extraction plays an important role in transforming the unstructured text into structured information that is easy to query and access. While the influx of data and expanding volume of scientific literature can lead to data-intensive scientific discovery (Tolle *et al.*, 2011), the manual extraction of information from this unstructured or semi-structured data is infeasible due to its volume. Rule-based and machine-learning-based information extraction have been proposed to automatically extract relevant information that scatters in different articles (Aggarwal & Zhai, 2013), and enable subsequently aggregating and organizing this information for more efficient use. The rule-based approach defines textual patterns based on how the relevant information is reported in the literature, and uses the defined patterns to extract relevant information from text. This usually involves the use of regular expressions and/or grammar parsing (e.g. Xiao, *et al.*, 2013; Torii *et al.*, 2015; Wu *et al.*, 2022). The machine-learning-based approach considers the information extraction task as a classification problem, that is, to classify whether a token belongs to the category of interest or not. A number of supervised machine learning models have been used for this purpose, including maximum entropy (Chieu & Ng, 2003), support vector machines (Isozaki & Kazawa, 2002), decision trees (Szarvas *et al.*, 2006), conditional random fields (McCallum & Li, 2003) etc. Conditional random field (CRF) shows an advantage in considering the interdependencies in the sequence of tokens (Peng & McCallum, 2006). More recently, deep learning models have been integrated with CRF for information extraction purposes (Huang, Xu & Yu, 2015).

The progress of text mining for scientific literature varies by different domains. Biomedical domain spearheads in this development. It has developed large scale ontologies, such as MeSH (Lipscomb, 2000) and UMLS (Bodenreider, 2004), annotated biomedical literature with MeSH in the PubMed database, cumulated many training corpuses (e.g. Kim, *et al.*, 2003; Nédellec *et al.*, 2013), created tools for information extraction (Wei, *et al.*, 2019), and provided open access to literature databases (e.g. PubMed and PubMed Central). This has contributed to much more organized knowledge in the biomedical domain than that in other domains. On the other hand, text mining for the materials science domain is still in its early stage (Olivetti, *et al.*, 2020), although it is starting to gather more attention. The next section will review the development of text mining for the materials science domain.

1 [1] Niiniluoto, I. (2002). Scientific progress. https://plato.stanford.edu/entries/scientific-progress/

Text Mining for Materials Science

Scientific literature on materials science is available in abundance, but text mining in the domain is still in its infancy. The bottleneck is the limited number of toolkits or libraries available at present that can be readily used to extract useful information. While some text mining has been done in biomedical (Shen *et al.*, 2020) and the intersection of chemistry and biomedical domain (Krallinger *et al.*, 2017, Tarasova *et al.*, 2019), there has been only a handful of research related to text mining in materials science. A recent research by Swain and Cole (2016) has presented a toolkit called ChemDataExtractor that can be used to mine information from materials science literature. They created a pipeline of NLP functions including tokenizer, named-entity recognizer, and parts-of-speech tagger that are chemistry aware. Additionally, rule-based grammars can be added to parse phrases from documents, thereby extracting relevant information. The capability of this toolkit to perform information extraction from structured and unstructured texts is also another excellent attribute. The ChemDataExtractor has been used recently by a few research groups since its advent to perform text mining for material synthesis (Kim *et al.*, 2017) and database creation of battery materials (Huang & Cole, 2020). Other than the ChemDataExtractor, HIVE (Greenberg *et al.*, 2021) is another tool that was created to advance ontology related to materials science. The HIVE provides a GUI interface through which material scientists can search vocabularies and concepts. It also provides a quick and automated way to index entity-related textual information, but HIVE is still in the developmental stages.

Chemical Named Entity Recognition

The literature of the materials science domain has unique language characteristics. There is a need for the development of specific resources (datasets, tools, and libraries) for this domain. The first major challenge for chemical information extraction is the detection of chemical entities (compounds, elements, formula) as information extraction is impossible without detecting the entities to which the information is related. While there are many well-known methods for named entity detection in NLP (Manning *et al.*, 2014) the task of chemical entity detection is more complicated. In general, a chemical named entity (CNE) can span over multiple words or may be a scientific formula (e.g. Zinc Oxide, ZnSnO3) containing a mix of characters, numbers, and special characters. Due to the lack of training data, supervised machine learning models have not always been possible. Rule-based and machine-learning-based methods have been attempted for this task.

Rule-based methods

Some earlier approaches have been lexicon-based (e.g. dictionaries) or rule-based. In the lexicon/dictionary-based CNE recognition (e.g. Hettne *et al.*, 2009; Rebholz-Schuhmann *et al.*, 2007; Klein, 2011), there is a predefined collection of entities like the Jochem (Hettne *et al.*, 2009) or the DrugBank dictionary (Wishart *et al.*, 2017), and each token in the text is matched fuzzily to the entries in the collection to find exact and/or partial matches. While these systems can achieve high precision, the recall is often very low. This is mostly due to the fact that various expressions in natural language do not necessarily match with a predefined collection of entities. In addition, maintaining a lexicon requires a periodic systematic update which is time-intensive and costly. Rule-based CNE recognition (Humphreys *et al.*, 1998; Budi and Bressan, 2003; Narayanaswamy *et al.*, 2003) is a little more generalizable approach than lexicon-based CNE recognition. Pattern matching rules curated from some common chemical naming conventions like IUPAC (Eaborn, 1988) can be used by systems to detect root forms. These rules are often a combination of orthographic and morphological rules that can help a

system understand different elements of a CNE. The drawback of a rule-based CNE recognition system is the lack of portability, high maintenance, and increasing complexity. With a slight change of naming convention, the time and effort to comprehend and revise the existing rules would not be cost-effective.

Machine-learning-based methods

For machine-learning-based methods, it is essential to develop field-specific corpora that could be used to train machine learning models. While there are several corpora like GENIA (Kim *et al.*, 2003) and CRAFT (Bada *et al.*, 2010) containing chemical mentions, they are not primarily concerned about chemical compounds. In Krallinger *et al.* (2015), the authors created a corpus (called CHEMDNER) including 84,355 chemical named entities from 10,000 PubMed abstracts for a workshop task in BioCreative IV of identifying chemical entities from literature abstracts. CHEMDNER is one of the most well-known corpora in material informatics as it contains manually created gold standard annotations for chemical entities. Several researchers used supervised machine learning approaches over the years to detect CNE. Some of the commonly used machine learning algorithms for CNE detection were statistical models (Bikel *et al.*, 1992), conditional random field (CRF) (Luo *et al*., 2018; Leaman *et al.*, 2015), support vector machine (SVM) (Tang *et al.*, 2015; Azari, 2013), Naive Bayes (NB) (Townsend *et al.*, 2005) and Maximum Entropy Markov Model (MEMM) (Borthwick, 1999). One of the best results using the CHEMDNER dataset was achieved by a system called tmChem (Leaman *et al.*, 2015) that used supervised machine learning using conditional random field (CRF). More recently, some different configurations of neural network based machine learning have also been employed for chemical named entity recognition (Luo *et al.*, 2018; Zhai *et al.*, 2019; Hemati and Mehler, 2019).

Feature selection for CNE recognition is of vital importance because most tokenizers that achieve state-of-the-art performance in other NLP tasks cannot be readily used for chemical entities (Leaman *et al.* 2015; Corbett *et al.*, 2007). For example, using normal tokenizers on a chemical name like K4(Fe(CN)6) will lead to the removal of all the "(" and ")" which would lead to three different chemical elements. Therefore, understanding the boundary of CNE is one problem that needs to be addressed. For a better detection of CNE boundary, most systems try to include a variety of features like presence of capitalization, presence of roman numerals, presence of numeral, word shape, orthographic features, and morphological features (Wang *et al.*, 2008). Some other NLP techniques like lemmatization and stemming have been reported to improve performance (Huber *et al.*, 2013). Some researchers also explored a hybrid approach (machine learning + dictionary or rule-based + dictionary) to improve the performance of their system. For example, ChemSpot (Rocktäschel *et al.*, 2012) uses a hybrid CRF plus lexicon-based CNE recognition system. The tagging is generated by both the taggers independently, and then merged using a union operation. In another study (Lana-Serrano *et al.*, 2013), it was found that semantic features do not yield a better result for CNE recognition.

In recent years, large transformer-based language models, like BERT (Devlin *et al.*, 2018), have been used to achieve state-of-the-art performance in several NLP tasks like named entity recognition. These language models can be used with limited training data for any downstream task and the process is known as transfer learning. Leveraging transfer learning, many recent studies have performed biomedical named entity recognition which is partially relevant to a chemical named entity recognition. Some of the more prominent among them can be found in Peng *et al.* (2019), Sun *et. al* (2021), and Naseem *et al.* (2021). Interestingly, Naseem *et al.* (2021) achieved an accuracy of 99.99% on combined tasks for chemical entities and drugs (BC5CDR and BC4CHEMD) using an ALBERT large model fine-tuned with PubMed and PMC corpus. In terms of chemical entity recognition for material science, no specific research has been reported that uses transfer learning with some large transformer-based language models.

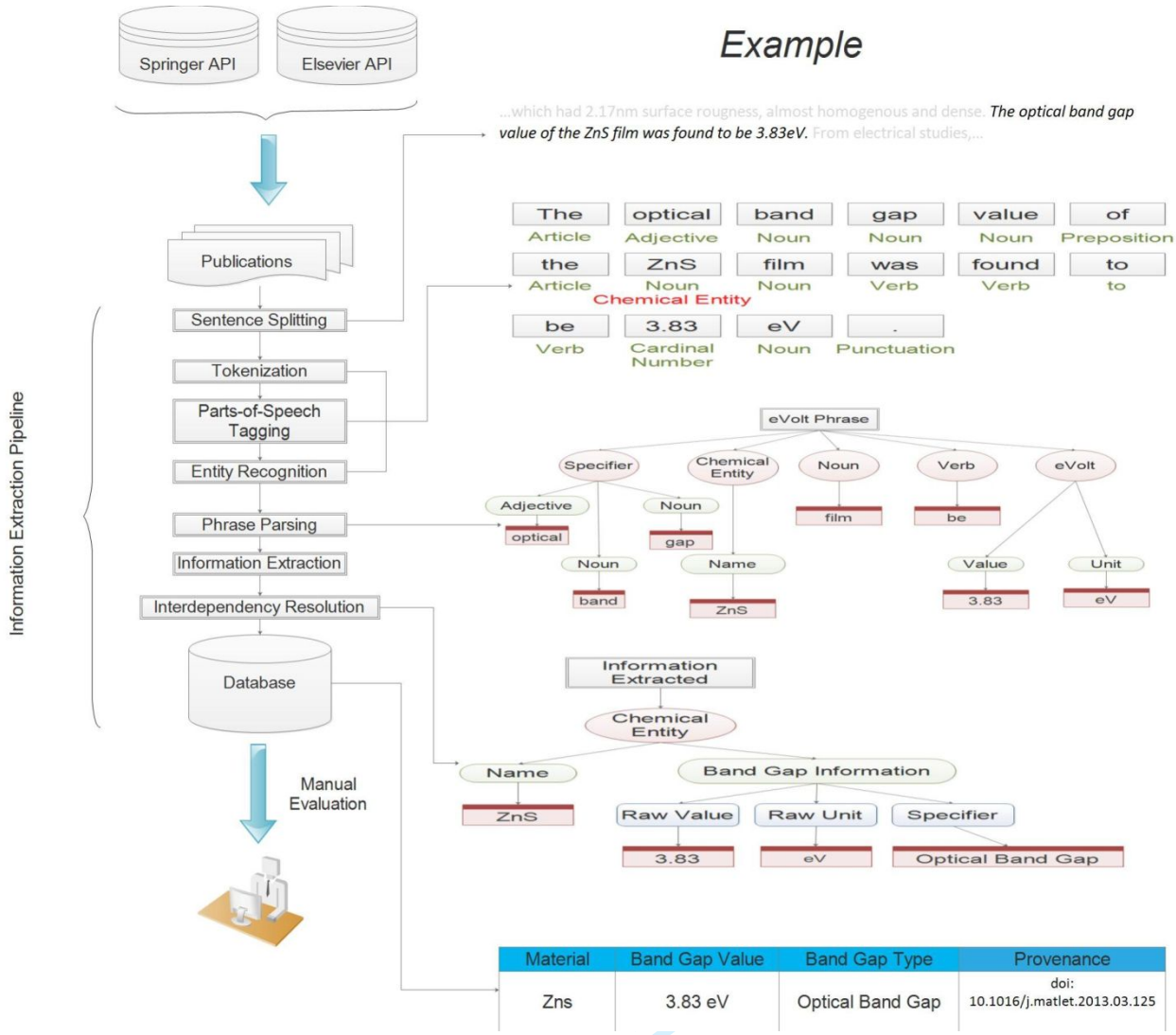

Figure 1: Data collection and system overview

Data Collection

To build our collection of articles related to materials science, we used the list of DOIs (Digital Object Identifiers) curated by Tshitoyan *et al.* (2019) of articles on inorganic materials. Using the application programming interfaces (APIs) of Springer Nature and Elsevier, we downloaded the title, abstract, date of publication, and journal name of all the articles. We collated them in a single database. The total number of research articles that we collected is about 1.44 million.

Method

The primary objective of this study was to extract information about chemical compounds and their associated band gap values. In addition, we also wanted to extract band gap specifiers (like direct, indirect, electronic, etc) that specify the types of band gap if available. A graphical overview of the system can be found in Figure 1. Our method consists of four different parts, and they are:

- Dataset filtering
- Information Extraction
- Automated screening
- Evaluation

The first step was to clean the dataset and only include titles and abstracts of literature that report band gap values of some chemical compounds. Our second step was directed towards extracting band gap information from the selected literatures by extending ChemDataExtractor. In this step, we created our own rule-based Band gap parser and combined it with existing Compound parser from ChemDataExtractor to create a record of information. The third stage was to clean extracted information as some records did not contain band gap information. The final step of our method was to come up with an evaluation strategy to assess the quality of information extracted by our system. We will discuss each of four parts in detail in the following subsections.

Dataset filtering

Not every article we collected reported band gap information. We investigated the database during the development phase of our system and found that many abstracts did not contain any information about band gap. Based on this initial investigation and consultation of domain experts, we decided to filter the dataset and include papers that mentioned any of the words “bandgap,” “band-gap,” “band gap,” “bandgaps,” “band-gaps,” or “band gaps” in their title or abstract. In addition, we performed a second level screening by searching for the token “eV,” which is the only unit for band gap values. The screening reduced the collection to 11,939 papers. The information extraction procedure was conducted on these 11,939 titles and abstracts.

Information Extraction

We used ChemDataExtractor (Swain & Cole, 2016) as the basic framework for our information extraction process. This is a convenient toolkit that enables building customized modules on a pre-existing information extraction framework. It has a pipeline that consists of different natural language processing techniques (like tokenization, parts of speech tagging, and named entity recognition) for information extraction. The challenging aspect of this research was to create a parser that can accurately identify and extract band gap value and associate it with the correct chemical compound, which has not been found in literature. ChemDataExtractor implements multiple chemical entity taggers including a case-sensitive lexicon-based tagger, cased-insensitive lexicon-based tagger, and CRF tagger that was trained on CHEMDNER. The chemical entities mentioned in a document are returned as a union of these taggers. We used the chemical entity tagger from ChemDataExtractor to identify chemical entities mentioned in the titles and abstracts of the literature. In addition, we designed a parser (BandGapParser) to extract information about the band gap value and band gap type of different compounds. Combining information extracted from both parsers, we created a chemical record containing the chemical entity name, its associated band gap value, and the band gap specifier (i.e. the type of the band gap reported).

Unit and Specifier

Similar to Huang and Cole (2020), we formulated some specific grammars to define our parser. While other material properties can be represented by units from different metric systems (like S.I or CGS), the band gap value of a compound usually is represented by only one type of unit, and that is “eV.” Several other properties that indirectly influence the band gap can have units similar to “eV.” For example, “eV/K” or “eVK^-1” or “eVK^{-1}” are very frequently encountered in materials science literature as it is the unit of Boltzman’s constant. This observation was made during the development phase when we tried about 1000 samples of unfiltered abstracts for information extraction. We needed to ensure that our parser disregarded information pertaining to units closely related to “eV” but not “eV.” It was also noted that band gap value is often preceded or followed by the mention of a band gap specifier in literature. Hence, we included the different types of specifier tokens that are commonly encountered to extract the type of band gap the literature mentions. The regular expressions for the unit and the specifier can be found in Table 1.

Table 1: Property description of band gap parser using regular expressions

Band gap properties	**Grammar**
Unit	(R('^eV$') + Not(I('/') | I('K-1') | I('K')))
Specifier	(Optional(I('direct') | I('indirect')) + Optional(I('electronic') | I('optical')) + Optional(I('tunable')) + Optional(I('energy')) + Optional(I('band')) + Optional(I('-')) + (I('Eg') | I('gap') | I('gaps') | I('bandgap') | I('bandgaps')))

Value Representation

In addition to this, we had to ensure that the parser is capable of extracting information pertaining to different types of value representation. Values in scientific literature often have a complicated representation. In particular, there are two elements related to value representation: expression style and range. Expression style means the different signs and symbols (like “+”, “-”, “~”, “±”, and “^”) in addition to the numbers that are used to represent a value. It is often difficult for the parser to identify the start and end range of a value because of the symbols. For example: “2±0.5” can be often misinterpreted two independent values of “2” and “0.5”. We curated grammar to extract value-related information that is expressed using different combinations of symbols and numbers. Exponential numbers being one of the most complex representations can also be captured successfully by our parser. For example, a sentence like “...the band gap value was found to be 2.46X10^-3 eV…” will determine band gap value to be “2.46x10^-3.” Range can be attributed to the different types of symbols, conjunctions or adjectives (like “to”, “and”, “up to”, “–”) used to express a range of values. It is important to capture the range as they make the information more contextually rich. To capture different variations of the ranges, we defined our own rule-based grammar namely and_range, to_range, up_to_range, joined_range, and spaced_range. For example, the value extracted from a phrase like “...the band bandgap value ranges from 4.6eV to 4.2eV…” would be “4.6 to 4.2.” The ranges that we used can be found in Table 2.

Table 2: Description and examples of different value range types

Range Type	Description	Examples taken from extracted data
and range	Two numeric values are mentioned with an “and” conjunction	...Band gap energies of the AP sample and the HT sample are estimated to be 2.02 and 2.37eV, respectively,...
to range	Two numeric values are mentioned with a “to” in between them	…optical band gap of the films were found to be ranging from 3.27 to 3.19 eV with Cd content…
spaced range	When there are two different band gap values mentioned but they are not joined by any conjunction	….These exhibited an optical band gap of ≈ 3.2 eV, estimated…… mass approximation by taking into account a fundamental energy band gap of 1.2 eV….
joined range	Multiple values are found clustered together in a single sentence and may end with another range type	…It is established that the band gap E g is 0.83–0.85 eV…
upto range	When the band gap range is described as a maximum limit then sometimes “upto” is used	…while SiH4/NH3, produced nitrogen-rich alloys (x~0.59) with E g upto 4.9eV…

Phrase Parsing

Phrase parsing is an essential step to understand the constituents of a sentence to enable information extraction. In general, the tags generated by parts-of-speech tagger are used in this stage to formulate a plausible segmentation of the noun phrase (NP) and verb phrase (VP) in a sentence. We customized phrase parsing to extract chemical entities, band gap values, and specifiers. Depending on the structure of phrases, the information related to band gap (value and specifier) can be located in different positions of a sentence. We had to define the various contexts where we could find the relevant information. Using a single grammar would have been too generalized and may have missed a lot of essential information. Therefore, we defined a few context-dependent rules that can capture the context of a sentence where the band gap information is usually presented in scientific literature. Before we describe the contextual grammar, we need to establish the basic elements that are merged together to form complex rules. There are three basic elements: cem (or chemical entity mentions), affix (words that are usually used before and after the band gap value to describe it), and value_and_specifier. The cem element is the name of the one or more chemical compounds or elements commonly preceded by an optional determiner (like “the”) and

followed by optional keywords (like “doped”, “thin films”) in different arrangements that may be present in the sentence containing the band gap information. The second element is affix which is composed of different frequently occurring words observed before or after discussing band gap information (like “with”, “measure”, “calculated”, “ranging”, etc.). The third or final element is the value_and_specifier, which is a combination of evolt value and evolt specifier (which is essentially the unit eV). In addition to this, we have another element called mcem which represents a structure where more than one cem element is joined using the word “and”. Combining these basic elements in various orders we created the grammar for phrase parsing. The orders are:

- cem+affix+value_and_specifier,
- cem+value_and_specifier+affix,
- affix+value_and_specifier+cem,
- affix+cem+value_and_specifier,
- value_and_specifier+cem+affix, and
- Value_and_specifier+affix+cem.

Our objective was to increase the scope of the grammar to accommodate varying scientific writing styles.

Interdependency Resolution

ChemDataExtractor implements an interdependency resolution process that takes the output from phrase parsing to merge the records that refer to the same chemical entity and/or fill in missing chemical entity information from context. The output of the phrase parsing is a list of records containing all the information extracted following the grammar from single sentences. Sometimes, a sentence reports a band gap value without explicitly mentioning the chemical entity associated with it. This is handled by Interdependency Resolution using Global Contextual Information (Swain & Cole, 2016). Following some predefined rules, the name of the chemical entity of this type of record is assigned from the preceding sentence or the header of the text which in our case is the title of the paper.

The other function of interdependency resolution is to merge the related records about the same chemical entity into one record. This is usually done through the recognition of variations of expressions, abbreviations or labels. This is called Chemical Identifier Disambiguation in Swain and Cole (2016).

These functions of interdependency resolution offer some capabilities of extracting and synthesizing information across sentences.

Automated Screening

We obtained a list of records from the information extraction procedure described above. Each record consists of information extracted using the CompoundParser (in-built parser of the ChemDataExtractor) and the BandGapParser. As our objective is to only extract information about the compounds or elements with a band gap value, we had to do some filtering since not all the elements or compounds extracted had a band gap value. In this post-processing step, we filtered out the records without band gap values and only the records with compound name and band gap value were retained. In some cases where we did not have a record with both these informations, we just returned a blank list which essentially means no information has been extracted.

Evaluation

To evaluate the quality of information extracted from the scientific literature, we randomly sampled 500 articles from the total collection of 11,939 articles. In spite of filtering out most of the irrelevant literature, there is still some literature which had no mention of band gap value. We had to manually remove them and annotate the rest with the labels described in Table 3. These labels represent three different levels of correctness. The number of valid literature was found to be 415 articles out of the set of 500. Two annotators independently went through the title and abstract of these articles to verify if the extracted information is correct, partially correct, or incorrect using the definitions listed in Table 3. In the 415 articles, we found 417 relevant materials and 422 associated band gap information.

In several material science articles, the authors report how they increased or decreased the band gap of material through different processes like doping or annealing. As our interest was to detect the primary material and its corresponding band gap value, we focused on the material used to make the solar films rather than the catalysts or doping agents. When evaluating, we primarily check the correctness of a record on the basis that the compound name of the film has been extracted along with the reported band gap value. In many cases, we can see that the name of the doping material or the substrate material has been extracted instead of the primary material for the film, and these are deemed incorrect material for evaluation purposes. To evaluate the correctness of the band gap value, we wanted to make sure that all the values reported in the paper are extracted, including the various ranges. If a record contains only a partial number of band gap values from the article, the record is annotated as partially correct. In Table 4, we have listed some examples of the records with different types of annotation labels.

Table 3: Description of annotation labels

Annotation labels	**Definition**
Correct	The Title and Abstract have information about a film material and its associated band gap value. In this scenario, both are extracted correctly and fully.
Partially Correct	The Title and Abstract have information about a film material and its associated band gap value. In this case, the extracted information includes correct material or band gap value, but not both. This category also includes the cases where both material and band gap value are extracted, but the value is not associated with the correct material. In addition, when multiple band gap values are reported, only a subset of them are extracted.
Incorrect	The Title and Abstract have information about a material and its associated band gap value. In this case, neither information is correctly extracted.

Table 4: Annotation examples

Example	Information available	Information extracted	Label
…donor-acceptor pairs located in the band gap. Radiative transitions from shallow donor levels located at 0.029 and 0.040 eV below the bottom of the conduction band to deep acceptor levels located 0.185 and 0.356 eV…	None	None	NA
…The optical band gap (Eg) of FNS is 3.27eV with direct transition…	Compound Name: FNS Band-gap value: 3.27eV Band-gap type: Optical band-gap	Compound Name: FNS Band-gap value: 3.27eV Band-gap type: Optical band-gap	Correct
…Besides, the calculated band gap of Sn1/32Bi30/32F3 with V Bi 1 2 - is 2.70 eV, which is smaller than that of pure BiF3…	Compound name: Sn1/32Bi30/32F3 Band-gap value: 2.70 eV Band-gap type: None	Compound name: Li Band-gap value: 2.70 eV Band-gap type: None	Partially correct
…CuInxGa1-xSe2 with the bandgaps 1.14-1.16 and 1.36-1.38eV have been evaluated…	Compound name: CuInxGa1-xSe2 Band-gap value: 1.14-1.16 and 1.36-1.38eV Band-gap type: None	None	Incorrect

Results and Findings

In the preliminary phase of annotation, two annotators independently classified the extracted records in three major groups: correct, partially correct and incorrect. The inter-annotator agreement was measured using Cohen's Kappa and the value was found to be 0.815, indicating a strong agreement. The confusion matrix of the final labels of annotations can be found in Table 5.

Table 5: Confusion matrix of results from two annotators after preliminary evaluation.

	Incorrect	**Partially Correct**	**Correct**
Incorrect	50	0	0
Partially Correct	1	126	18
Correct	0	26	194

From Table 5, we can see that the agreement between the two annotators is high. The disagreements mostly happen to the cases of partial correct versus correct. There were 45 conflicts out of 415 instances, and these conflicts were discussed and mutually resolved in the second phase of annotation by the annotators. The final results are presented in Table 6. We can see that the information extracted from 51.32 percent of the literature is correct. Information extracted from another 36.62 percent of literature is partially correct which means some portion of the information available in the literature is identified and extracted correctly. The system failed to correctly extract information from 12.04 percent of the literature. The errors related to partially correct and incorrect can be identified in some major classes and they are explained in detail in the next subsection.

Table 6: Performance of the information extraction.

Labels	**Frequency**	**Percentage**
Correct	213	51.32%
Partially Correct	152	36.62%
Incorrect	50	12.04%

In addition, we also calculated the Precision and Recall of the system for Correct extractions and (Correct + Partially Correct) extractions (Table 7). In the context of this study, if we analogize extraction to retrieval, the Precision is the fraction of correct extractions among all extractions and the Recall is the fraction of correct extractions out of all relevant articles. The Precision and Recall were calculated using the following equations:

$$Precision = \frac{of\,articles\,with\,correct\,extraction}{of\,articles\,with\,extraction}$$

$$Recall = \frac{of\,articles\,with\,correct\,extraction}{of\,articles\,with\,relevant\,information}$$

We have two different gradients of correctness. Considering extractions that were labeled as Correct, we found that the number of articles with correct extractions is 213. The total number of articles with extractions is 365. Therefore, the Precision was *0.58*. If we consider the Partially Correct labels in addition to the Correct labels, then the Precision was *1*. This suggests that for articles with any extracted records, the results are either partially correct or correct.

The numerator remains the same for Recall but the total number of articles with the relevant information was 415. Therefore, for Correct extractions, the recall was *0.513*. If we consider Partially Correct extractions in addition to Correct extractions then the recall was *0.88*.

Table 7: Precision and Recall of the extracted records.

	Precision	**Recall**
Correct	0.58	0.513
Correct + Partially Correct	1.00	0.88

The extraction system was also applied to the 11,939 articles that date from 1962 to 2020. No extraction record was obtained for 2,573 articles (21.56%). We suspect some of these articles did not report band gap values for compounds, like the 85 articles in the randomly sampled 500 articles for our manual evaluation. For the remaining 9,366 articles (78.45%), 10,608 band gap values were extracted for 10,292 compound names. Due to the sheer volume of the articles, we were unable to evaluate the correctness of the extraction. The performance metrics on the randomly sampled 415 articles should offer some insights.

Error analysis

On analyzing the partially correct and incorrect information records extracted by the system, we found that the errors are due to three main reasons: 1) failure to identify and extract the correct material of the film, 2) failure to extract the correct band gap information pertaining to value or specifier, and, 3) failure to relate the correct compound with the band gap value due to interdependency resolution issues. The three types of errors are related to each other because if name or value is not correctly recognized and extracted, then there is no question of interdependency error. Therefore, in a way the first two errors preclude the third type of error.

The error analysis includes two steps: In the first step, we checked to see if the system had extracted the compound name and the band gap information. If both information is extracted correctly but not related to each other, then we concluded that there are some interdependency resolution issues. We analyzed the 202 records individually to find the categories of error and discuss their variations below:

- ### Compound name error

  In many instances, the name of the compound is not recognized by the system. We identified that sometimes in these articles compound names are represented in different ways like using special characters (x) to denote the different concentration of elements in a compound (e.g. Ge1−x−ySixSny), using special characters (/) to show different layers of a solar film (e.g. AgBr/Ag4P2O7), and sometimes the names are extremely long and too complex for the system to recognize it as a single entity (e.g. Two soluble poly(1,4-phenylene vinylene-4,4′-biphenylene vinylene)s). In total, we

found 50 instances of compound name error from the 202 instances of partially correct and incorrect records.

- ## Band gap information error

  There were mainly three types of errors pertaining to band gap information and they were either inability to extract multiple band gap values represented by range, failure to extract the band gap specifier when that information is present in the article, and extraction of information that is not band gap value but represented by the unit of "eV." For example, the first type of error was encountered when there were two band gap value ranges in the abstract "The band gap values of the films annealed at 500 and 300°C were 3.3—4.0 eV and 3.4—4.2 eV, respectively." The first range was extracted but the second range was missed. The second type of error was found when two different band gap specifiers are mentioned in the same sentence, for example, "Optical transmission data of CuCrO2 films indicate a direct band gap and an indirect-gap of about 3.15eV and 2.66eV, respectively." Only the first band gap specifier was extracted. The third type of band gap information error occurred when other type of gap energies are mentioned in the abstract like: "HOMO LUMO gap energies of the clusters (CeO2)13 and (CeO1.5)13 are calculated to be almost 0 and 3.05eV, respectively." The system failed to comprehend that "HOMO LUMO gap energies" is not the same as band gap energy. While this error is a critical error, the number of records that had the error was the least and only 39 instances had this type of error.

- ## Interdependency error:

  This was the most frequent form of error with 114 instances and contributed to 56.43% of the total error. We found that on several instances even after identifying the compound and the band gap information correctly, the system failed to relate them properly. It failed to understand the context of the article and erroneously attributed the band gap value to a wrong chemical entity. For example, in an abstract there are two different chemical entities like Copper indium diselenide (CIS) and Indium gallium diselenide (CIGS) and direct band gap values of CIGS is reported as "between 1.02 eV and 1.68eV." The system extracted both the information correctly but failed to link CIGS and instead linked CIS with the band gap value.

Table 8 summarizes the error categories, their frequencies and percentages. One record was identified to have both compound name error and band gap information error, therefore the sum of the errors exceeds the total number of records by one.

Table 8: Error categories and their occurrences.

Error Category	Frequency	Percentage
Compound name error	50	24.63%
Band gap information error	39	19.21%
Interdependency error	114	56.16%

Conclusion

In this paper we describe our preliminary work on the extraction of band gap information related to PV material from the titles and abstracts of scholarly articles related to materials science. We built our corpus from a set of 1.44 million literature and filtered it to 11,939 articles that are potentially relevant to the task. To evaluate the performance of our approach, we randomly sampled information extracted from 415 articles. We found that our system can correctly extract information from the majority of articles (51.32%) and can extract partially correct information from 36.62% of articles. We analyzed the errors and found three primary reasons that contributed to the error. Our study shows that it is possible to obtain a large part of usable information by our approach and interdependency resolution between the material name and the band gap information is of utmost importance because it contributed to the majority of the errors.

Limitations and Future Work

Overall, the objective of the study was to understand the challenges of extracting information pertaining to PV material. From this pilot study, we have found the reasons that contributed to the errors. In the future, we will try to address these issues and develop a system that would be more robust with the capability to better comprehend the context of the article. While it will be challenging to disambiguate and resolve the interdependency between chemical entities and their band gap values, our initial results are promising and we hope to build on them. It should be noted that one limitation of the study is when we evaluated the performance of the information extraction, we directly rated the extracted results to be correct, partially correct, and incorrect by two independent annotators. A more rigorous procedure is to obtain the ground truth first, and then compare the extracted results with the ground truth. We acknowledge that there may be biases arising from our procedure. We have since reviewed the articles and obtained the ground truth for future studies. There are also other limitations of the study. One is the full-text availability. We had to remove a majority of the abstracts because they do not contain band gap values of materials. However, some of these values may be reported in the full-text rather than in abstracts. The availability of full-text is still an obstacle for text mining.

Reference

Aggarwal, C. C., & Zhai, C. X. (2013). Mining text data. In Mining Text Data (Vol. 9781461432234). https://doi.org/10.1007/978-1-4614-3223-4

Azari, S., 2013. A Comparative Analysis of Chemical Named Entity Recognition Using Support Vector Machines (Master's thesis, Eastern Mediterranean University (EMU)-Doğu Akdeniz Üniversitesi (DAÜ)).

Bada, M., Eckert, M., Palmer, M. and Hunter, L., 2010, July. An overview of the CRAFT concept annotation guidelines. In Proceedings of the Fourth Linguistic Annotation Workshop (pp. 207-211).

Bikel, D.M., Schwartz, R. & Weischedel, R.M., 1999. An Algorithm that Learns What's in a Name. Machine learning, 34(1), pp.211–231.

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270.

Borthwick, A.E., 1999. A maximum entropy approach to named entity recognition. PhD thesis, New York University.

Brindha, S., Prabha, K., & Sukumaran, S. (2016, January). A survey on classification techniques for text mining. In 2016 3rd International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1-5). IEEE.

Budi, I. and Bressan, S., 2003, December. Association rules mining for name entity recognition. In Proceedings of the Fourth International Conference on Web Information Systems Engineering, 2003. WISE 2003. (pp. 325-328). IEEE.

Chieu, H. L., & Ng, H. T. (2003). Named entity recognition with a maximum entropy approach. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 (pp. 160-163).

Corbett, P., Batchelor, C. and Teufel, S., 2007, June. Annotation of chemical named entities. In Biological, translational, and clinical language processing (pp. 57-64).

Correa-Baena, J. P., Abate, A., Saliba, M., Tress, W., Jacobsson, T. J., Grätzel, M., & Hagfeldt, A. (2017). The rapid evolution of highly efficient perovskite solar cells. Energy & Environmental Science, 10(3), 710-727.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Eaborn, C., 1988. Compendium of chemical Terminology: IUPAC Recommendations : compiled by V. Gold, K.L. Loening, A.D. McNaught, and P. Sehmi, Blackwell, Oxford, etc., 1987, viii 456 pages. £45.00 (hard cover) ISBN 0-632-01765-1; £29.50 (soft cover) ISBN 0-632-01767-3. Journal of Organometallic Chemistry, 356(2), pp.C76–C77.

Erraguntla, M., Gopal, B., Ramachandran, S., & Mayer, R. (2012, January). Inference of missing ICD 9 codes using text mining and nearest neighbor techniques. In 2012 45th hawaii international conference on system sciences (pp. 1060-1069). IEEE.

Gordon, M. D., & Dumais, S. (1998). Using latent semantic indexing for literature based discovery. Journal of the American Society for Information Science, 49(8), 674-685.

Hemati, W. and Mehler, A., 2019. LSTMVoter: chemical named entity recognition using a conglomerate of sequence labeling tools. Journal of cheminformatics, 11(1), pp.1-7.

Hettne, K.M., Stierum, R.H., Schuemie, M.J., Hendriksen, P.J., Schijvenaars, B.J., Mulligen, E.M.V., Kleinjans, J. and Kors, J.A., 2009. A dictionary to identify small molecules and drugs in free text. Bioinformatics, 25(22), pp.2983-2991.

Huang, S., & Cole, J. M. (2020). A database of battery materials auto-generated using ChemDataExtractor. Scientific Data 2020 7:1, 7(1), 1–13. https://doi.org/10.1038/s41597-020-00602-2

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991.

Huber, T., Rocktäschel, T., Weidlich, M., Thomas, P. and Leser, U., 2013, October. Extended feature set for chemical named entity recognition and indexing. In BioCreative Challenge Evaluation Workshop (Vol. 2, p. 88).

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H. and Wilks, Y., 1998. University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998.

Isozaki, H., & Kazawa, H. (2002). Efficient support vector classifiers for named entity recognition. In COLING 2002: The 19th International Conference on Computational Linguistics.

Kasap, S. (2006). Springer handbook of electronic and photonic materials. Springer Science & Business Media.

Kim, J.D., Ohta, T., Tateisi, Y. and Tsujii, J.I., 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics, 19(suppl_1), pp.i180-i182.

Klein, C., 2011. Information Extraction from Text for Improving Research on Small Molecules and Histone Modifications (Doctoral dissertation, Universitäts-und Landesbibliothek Bonn).

Kononova, O., He, T., Huo, H., Trewartha, A., Olivetti, E. A., & Ceder, G. (2021). Opportunities and challenges of text mining in materials research. IScience, 24(3). https://doi.org/10.1016/J.ISCI.2021.102155

Lana-Serrano, S., Sanchez-Cisneros, D., Campillos, L. and Segura-Bedmar, I., 2013, October. Recognizing chemical compounds and drugs: a rule-based approach using semantic information. In BioCreative Challenge Evaluation Workshop (Vol. 2, p. 121).

Lipscomb, C. E. (2000). Medical subject headings (MeSH). Bulletin of the Medical Library Association, 88(3), 265.

Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H. and Wang, J., 2018. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 34(8), pp.1381-1388.

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S. and McClosky, D., 2014, June. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60).

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons.

Mitzi, D. B., Gunawan, O., Todorov, T. K., Wang, K., & Guha, S. (2011). The path towards a high-performance solution-processed kesterite solar cell. Solar Energy Materials and Solar Cells, 95(6), 1421-1436.

Mooney, R. J., & Bunescu, R. (2005). Mining knowledge from text using information extraction. ACM SIGKDD explorations newsletter, 7(1), 3-10.

Nalley, S., & LaRose, A. (2021). International Energy Outlook 2021. U.S. Energy Information Administration.. https://www.eia.gov/outlooks/ieo/pdf/IEO2021_ReleasePresentation.pdf

Narayanaswamy, M., Ravikumar, K.E. and Vijay-Shanker, K., 2002. A biological named entity recognizer. In Biocomputing 2003 (pp. 427-438).

Naseem, U., Khushi, M., Reddy, V., Rajendran, S., Razzak, I. and Kim, J., 2021, July. Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.

Nédellec, C., Bossy, R., Kim, J. D., Kim, J. J., Ohta, T., Pyysalo, S., & Zweigenbaum, P. (2013, August). Overview of BioNLP shared task 2013. In Proceedings of the BioNLP shared task 2013 workshop (pp. 1-7).

Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. In Mining text data (pp. 43-76). Springer, Boston, MA.

Olivetti, E. A., Cole, J. M., Kim, E., Kononova, O., Ceder, G., Han, T. Y. J., & Hiszpanski, A. M. (2020). Data-driven materials research enabled by natural language processing and information extraction. In Applied Physics Reviews (Vol. 7, Issue 4). https://doi.org/10.1063/5.0021106

Peng, F., & McCallum, A. (2006). Information extraction from research papers using conditional random fields. Information processing & management, 42(4), 963-979.

Peng, Y., Yan, S. and Lu, Z., 2019, August. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In Proceedings of the 18th BioNLP Workshop and Shared Task (pp. 58-65).

Pidcock, R. & McSweeney, R. (2021). Mapped: how climate change affects extreme weather around the world. Available at: https://www.carbonbrief.org/mapped-how-climate-change-affects-extreme-weather-around-the-world

Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M. and Stoehr, P., 2007. EBIMed—text crunching to gather facts for proteins from Medline. Bioinformatics, 23(2), pp.e237-e244.

Rocktäschel, T., Weidlich, M. and Leser, U., 2012. ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics, 28(12), pp.1633-1640.

Saparov, B., & Mitzi, D. B. (2016). Organic–inorganic perovskites: structural versatility for functional materials design. Chemical reviews, 116(7), 4558-4596.

Serrano, L., Bouzid, M., Charnois, T., Brunessaux, S., & Grilheres, B. (2013, November). Events extraction and aggregation for open source intelligence: From text to knowledge. In 2013 IEEE 25th International Conference on Tools with Artificial Intelligence (pp. 518-523). IEEE.

Shen, S., Liu, X., Sun, H. and Wang, D., 2020. Biomedical knowledge discovery based on Sentence-BERT. Proceedings of the Association for Information Science and Technology, 57(1), p.e362.

Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H. and Wang, J., 2021. Biomedical named entity recognition using BERT in the machine reading comprehension framework. Journal of Biomedical Informatics, 118, p.103799.

Swain, M. C., & Cole, J. M. (2016). ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. Journal of Chemical Information and Modeling, 56(10), 1894–1904. https://doi.org/10.1021/acs.jcim.6b00207

Szarvas, G., Farkas, R., & Kocsor, A. (2006, October). A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms. In International Conference on Discovery Science (pp. 267-278). Springer, Berlin, Heidelberg.

Tang, B., Feng, Y., Wang, X., Wu, Y., Zhang, Y., Jiang, M., Wang, J. and Xu, H., 2015. A comparison of conditional random fields and structured support vector machines for chemical entity recognition in biomedical literature. Journal of cheminformatics, 7(1), pp.1-6.

Todorov, T. K., Reuter, K. B., & Mitzi, D. B. (2010). High-efficiency solar cell with earth-abundant liquid-processed absorber. Advanced materials, 22(20), E156-E159.

Torii, M., Arighi, C. N., Li, G., Wang, Q., Wu, C. H., & Vijay-Shanker, K. (2015). RLIMS-P 2.0: A generalizable rule-based information extraction system for literature mining of protein phosphorylation information. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 12(1), 17–29. https://doi.org/10.1109/TCBB.2014.2372765

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., … Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 571(7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8

Wang, H., Zhao, T., Tan, H. and Zhang, S., 2008. Biomedical Named Entity Recognition Based on Classifiers Ensemble. Int. J. Comput. Sci. Appl., 5(2), pp.1-11.

Wei, C. H., Allot, A., Leaman, R., & Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. Nucleic acids research, 47(W1), W587-W593.

Wishart, D., Feunang, Y., Guo, A., Lo, E., Marcu, A., Grant, J., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C. and Wilson, M., 2017. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research, 46(D1), pp.D1074-D1082.

Wu, L. T., Lin, J. R., Leng, S., Li, J. L., & Hu, Z. Z. (2022). Rule-based information extraction for mechanical-electrical-plumbing-specific semantic web. Automation in Construction, 135, 104108.

Xiao, L., Tang, K., Liu, X., Yang, H., Chen, Z., & Xu, R. (2013). Information extraction from nanotoxicity related publications. Proceedings - 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013, 25–30. https://doi.org/10.1109/BIBM.2013.6732723

Zhai, Z., Nguyen, D.Q., Akhondi, S.A., Thorne, C., Druckenbrodt, C., Cohn, T., Gregory, M. and Verspoor, K., 2019. Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. BioNLP 2019, p.328.