A Text Mining Pipeline for Mining the Quantum Cascade Laser Properties

Deperias Kerre, Anne Laurent, Kenneth Maussang, Dickson Owuor

A Text Mining Pipeline for Mining the Quantum Cascade Laser Properties

Deperias Kerre[1,2][0000−0002−7437−6735], Anne Laurent[2][0000−0003−3708−6429], Kenneth Maussang[3][0000−0002−8086−8461], and Dickson Owuor[1][0000−0002−0968−5742]

[1] SCES, Strathmore University, Nairobi, Kenya
{dkerre, dowuor} @strathmore.edu
[2] LIRMM, Univ Montpellier, CNRS, Montpellier, France
anne.laurent@umontpellier.fr
[3] IES, Univ Montpellier, CNRS, Montpellier, France
Kenneth.Maussang@umontpellier.fr

Abstract. The development of the Terahertz laser technology in quantum cascade lasers (qcl) has brought about great potential for industrial applications. These lasers are based on the Terahertz electromagnetic waves, in the frequency range from about 100GHz to 10THz. There is need to understand the structure of the laser and its influence on the performance in order to optimize the design process. One way of collating this information is by having ontologies and knowledge bases capturing the various qcl designs and their performance characteristics. Majority of the laser design data is usually contained in scientific literature. The main drawback of such textual data sources is their unstructured nature. The complex nature of the laser design and the varying author language styles poses some level of difficulty in retrieving this information. Owing to this, the existing methods needs improvement in order retrieve the laser information at a high precision(with minimal number of incorrect records extracted) and minimized number of correct records not extracted. In this paper, we tackle this initial challenge by proposing a text mining pipeline for mining the qcl properties by extending the grammar rules of a conditional random field (CRF) based model using a rule-based approach. The properties of interest include: hetero-structure (laser stacking properties), working temperature, lasing frequency, laser thickness and the optical power. We evaluate the pipeline on sample open access journal papers from AIP, OPTICA and IOP Publishers.

Keywords: CRF model· Information Extraction · Knowledge Bases · Ontologies · Property Models· Quantum Cascade Lasers · Text Mining

1 Introduction

There exists a lot of information in scientific literature published daily on quantum cascade laser technologies. The literature documents the various laser designs and their performance properties [1]. The terahertz quantum cascade lasers have varying industrial application potential ranging from the biomedical field, where the radiation can be used in detection of abnormal tissues, including cancers [2] and in the pharmaceutical field, where the lasers have been used in detecting organic compounds in drugs and identification of two or three dimensional distributions of molecules [3]. In electronics, the lasers can be used to pre-configure high speed telecommunication networks [4].

Quantum cascade lasers are complex hetero-structures. Most of the properties of the laser are defined by its growth sheet, i.e. the description of the different stacked layers: their thickness, the nature of the material, the order etc. The hetero-structural design of the laser constitutes the stacking properties i.e the different materials stacked together to form the laser while the opto-electronic characteristics of the laser entails the laser performance behaviour such as working temperature, power, frequency which is as a result of current injection in the laser.

Information regarding the description of the quantum cascade laser structures and performance is highly desired to give crucial insights for several purposes such as optimization of scientific design processes/implementations. The quantum cascade laser properties of interest to our study include: Working temperature, Optical Power, Lasing Frequency, Material design(Hetero-structure) and the Barrier thickness. These properties are crucial in evaluating the performance of the laser on various tasks. The quantum cascade laser working temperature, power, and frequency properties consists of value and a unit. For instance,"the quantum cascade laser lases at 9.7 THz, at a working temperature of 186 K with a maximum output power of 9 mW". The material design consist of a combination of chemical material names, in some cases with digits together with forward slashes. A sample statement containing this information is "We present two different terahertz quantum cascade laser designs based on GaAs/Al0.3Ga0.7As heterostructures". On the other hand, the quantum cascade laser barrier thickness consists of the thickness of the barriers in the hetero-structure. The property definition consist of the value(which in most cases consist of a sequence of numbers and forward slashes/commas)and a unit. A sample expression of this unit may be as follows: "The improved structure has layer sequence 31/93/14/73.4/23/155.4/11/110.2/14/84.7/20/155.4/17/110.1 Å".

One of the ways to capture the quantum cascade laser design and performance information from scientific literature is by designing ontologies and knowledge bases from the unstructured textual data. The main limitation of such textual data is their unstructured nature owing to the domain specific terminologies and different language styles by the authors. Some of the quantum cascade laser properties such as the barrier thickness poses difficulty in extraction due to the presence of special characters such as the forward slash(/) and the comma(,). In some cases, the properties are expressed in different ways such that there is need for contextualized rules to identify the property. The initial step of achieving structured ontologies and knowledge bases of quantum cascade laser design and performance properties is therefore implementing a text mining pipeline for extracting the quantum cascade laser properties from scientific literature.

There has been advances in the field of Information Extraction to structure the unstructured textual data in order to extract meaningful information from them. This has been accelerated by the adoption of the TDM Exception, a policy framework that advocates for the use of published resources for text and data mining purposes [5]. Examples include the use of machine learning algorithms in accelerated materials discovery [6]. With this breakthrough, there is enormous potential for applicability to other domains.

In this paper, we propose a text mining pipeline for mining the quantum cascade laser properties based on an extension of the ChemDataExtractor pipeline, a chemistry aware toolkit based on the CRF model [7]. We propose this as the first step in developing ontologies and knowledge bases for the quantum cascade laser domain. Our main contribution in this paper constitutes proposed efficient qcl property mining rules with improved precision and minimized number of correct records that are not extracted. This is achieved by defin-

ing new property parsing rules in form of property parsers using the rule based grammar approach [8]. We also extend the extraction capabilities by defining new property models for the qcl properties to be used along with the defined rules.

The rest of the paper is organized as follows: we first review the related works in Section 2, then we propose the workflow in the methodology in section 3, we present the experimental and evaluation results in section 4, and finally we conclude in section 5.

2 Literature Review

Several works have been reported in the field of information retrieval in the materials science domain. The methodologies used can be broadly categorized into machine learning approaches and those that use a combination of machine learning and natural language processing principles. One of the crucial tasks under IR in materials science is the Chemical Name Entity Recognition(CNER).

CNER usually involves the identification of chemical and materials terms in the text. It can also be used to extract properties, physical characteristics, and synthesis actions. Early works on CNER focused on the on extraction of drugs and biochemical information [9,10]. Recently, CNER has gained alot of interest in the extraction of chemical and materials terms. The methods used in the CNER vary from traditional rule-based and dictionary look-up approaches to modern methodology built based on advanced machine learning(ML) and NLP techniques [11,12].

Examples of publicly available toolkits for extracting material terms include: those using rules and dictionaries-based approaches e.g LeadMine [13], ChemicalTagger [14], statistical models e.g OSCAR4 [15] and predominantly, the CRF model e.g ChemDataExtractor [7], ChemSpot [16], tmChem [17]. ChemDataExtractor has been extended/modified to extract several material terms and properties: semi-conductor bandgaps [18] , thermo-electric materials [19], battery materials [20], refractive indices and dielectric constants [21], transition temperatures of magnetic materials [22] and an auto-populated ontology of material sciences [23]. Machine Learning techniques have also been utilized in CNER to identify chemical materials and their roles based on context information. Examples include bidirectional LSTM models [24,25] and a combination of deep convolutional and recurrent neural networks [26]. Others studies have also proposed mined datasets of inorganic materials synthesis recipes [27] and gold nanoparticle synthesis procedures, morphologies, and size entities [28]. Pre-trained BERT models have also been utilized in the extraction of battery materials [29] and for optical materials research [30].

Material science information has also been extracted from tables and figures. There has been attempts to parse tables from the scientific literature using heuristics and machine learning approaches [31]. Attempts have been reported on parsing article images , for instance ImageDataExtractor tool that uses a combination of OCR and CNN to extract the size and shape of the particles from microscopy images [32] and the Livermore SEM Image Tools for electron microscopy images using Google Inception-V3 network [33].

As noted from the literature review, several works have been reported on the applications of machine learning and NLP to materials discovery. In this study, the interest is more on "wafer fabrication" or hetero-structure properties, which is a critical step in the quantum cascade lasers development. Despite the great advancements reported in the literature, there is still a great potential for research in the materials science domain in order to achieve structured information regarding the quantum cascade lasers. The existing

methodologies cannot be readily applied to mining these structures and the corresponding performance without modification/extension. Most of the natural language toolkits perform well in chemical terms, but when generalized from chemistry to the wider materials science, the grammar-based parsing rules used become less efficient. The BERT based models also need a lot of training data which involves manual annotation of the various properties by an expert. This may be cumbersome for large collections of articles. There is therefore need to extend the parsing capabilities of these techniques in order to adapt to the problem of mining the quantum cascade laser properties.

3 Methodology

In this section, we provide a detailed description of the workflow of the text mining pipeline for mining the qcl properties. The pipeline is based on an extension of the ChemDataExtractor, a chemical aware software toolkit [7]. We define new rules and make targeted extensions in order to fit to our domain of interest. The steps are as follows:

3.1 Document Retrieval and Processing

The first step in the text mining workflow is to acquire the scientific articles documenting the design of quantum cascade lasers. The study targets open access journals published by AIP, OPTICA and IOP publishers. The papers are retrieved using the keyword "quantum cascade lasers" and manually downloaded in the HTML format for further processing. The downloaded documents are then fed into ChemDataExtractor which uses the bespoke to process their information one document at a time. The downloaded HTML documents have a hierarchical structure with semantic markup tags. An example of such tags is the <head> tag which contains the metadata about the document such as title of the paper, the doi, authors etc. These tags are utilized by ChemDataExtractor to identify the key information about the papers such as the abstract, paragraphs, sentences etc. These files are then converted into plain text using the "reader" package in ChemDataExtractor which is then stored in the Document object of the toolkit for further processing.

3.2 Natural Language Processing

In this step, state-of-the-art Natural Language Processing techniques are applied to the document text. These capabilities are provided by the ChemDataExtractor toolkit. The techniques, which are tailored to the materials science domain include Sentence splitting, Tokenization, Part-of-Speech Tagging and Chemical-Named Entity Recognition(CNER). Sentence splitting, Tokenization and Part-of-Speech-Tagging were adopted from ChemDataExtractor without modification. The CNER rules are extended and adapted to the quantum cascade laser domain as described in the information extraction section.

3.3 Information Extraction

The ChemDataExtractor toolkit provides three ways of extracting information from text. These include:(i) Rule-based approach-which involves explicit crafting of statements that utilize regular expressions patterns and POS tags, (ii) automatic parsing and (iii) the

modified snowball algorithm that can be trained in a semi-supervised manner on documents dataset and probabistically used to extract information. In this paper, we adopt an extended rule-based approach by defining new property models and grammar logic for the qcl properties of interest. The property models are defined based on the user model concept [22].

The user model concept in ChemDataExtractor consists of a collection of defined property models for extracting different information. In general, a property model specifies the information to be extracted and the extraction rules to be used in retrieving the information. The information can be in form of physical quantities or chemical names. The user model consists of three models i.e the quantity models, general base model and the compound model.

The quantity model defines physical quantities such as time, electric charge, volume and the compound model defines chemical names together with the corresponding chemical name labels and roles. The general base model on the other hand contains user-defined fields, such as words, regular expressions, or other models. Every quantity model has the respective fields that will be populated upon data extraction from the document. The fields include the value, units, error, the standardized value and the specifier used to extract the data. For our text mining pipeline, we define five new property model to capture each of the properties of interest.

The property models for the working temperature, lasing frequency, power and laser thickness constitute quantity models while the hetero-structure model constitutes the compound model as the hetero-structure consist of material names. The Working Temperature property model inherits/nests the existing Temperature model in ChemDataExtractor. This handles the unit standardization process for the extracted temperatures. This is also the case for the OpticalPower model which inherits from the Power Model and the Heterostructure model which inherits from the Base model. In the quantum cascade laser literature, power readings are expressed in milliwatts(mW). We include this as an additional unit in the Power model. For the Lasing Frequency property model and the Barrier thickness, we define the Frequency model and the Barrier thickness model to handle the units. The fields to be populated from are also defined from scratch.

For all the property models, one of the important attributes is the parser. A property model can have one or more parsers. The parsers includes the defined grammar rules(logic) for relationship extraction. More information on parsers is given in relationship extraction section.

Phrase Parsing and Relationship Extraction: This is a key step that entails the extraction of suitable relationships. The relationship can be in the form of (i) a specifier expression/keyword and a chemical name or (ii) a specifier expression, a value and a unit. These relations are the ones that populate the specific records of the various qcl properties. ChemDataExtractor makes use of a hybrid approach to Chemical Named Entity Recognition (CNER); machine-learnt, dictionary-based and rule-based methods are all used.

The default parser of ChemDataExtractor , AutoSentenceParser, uses multiple specialized grammar rules that have been designed to extract more specific types of chemical information. In order to use the autosentence parser, a specifier expression is defined to capture the property relationship extraction rule. The rules are formed by combining the different keywords(table 1) and parser elements (table 2) in form of tokens.

Table 1. Quantum Cascade Laser Properties and the Keywords

Target Property	Keyword/Sentence	Unit
Working Temperature	heat-sink, Tmax, Maximum Temperature, Working Temperature	K
Optical Power	Optical power, Output Power, Peak Power	W
Hetero-structure (Material)	Growth, Grown in, Wafer, MBE, Laser-structure	N/A
Frequency	Laser Frequency, Lasing at, output Frequency	THz
Barrier Thickness	Layer thicknesses	Å

Table 2. The Parser Elements

Elements	Description	Elements	Description
R(Regex)	Match text with regular expression	T (Tag)	Match tags
W(Word)	Match case-sensitive token text	I(IWord)	Match case-insensitive token text
Any	Match any single token	H(Hide)	Ignore the matched tokens
Not	Match only if not followed by some text	FollowedBy	Match only if followed by some text
ZeroOrMore	Match zero or more of the expressions	OneOrMore	Match one or more of the expressions
Optional	Match if it exists	SkipTo	Skips to the next occurrence of text

The keywords adopted for each of the properties shown in Table 1 were settled upon based on consultation and advise by experts in the quantum cascade laser domain.

The default Autosentence parser however fails and under performes on some properties due to the high level of ambiguity and implicit knowledge carried within natural language. This has an implication on the precision of the extraction process and also leads to various correct records not being extracted. For instance, where several temperatures are mentioned in an article, there is need to define more precise rules for extracting the temperature of interest.

The Autosentence parser also requires a chemical compound in order to merge a complete record for extraction. This causes it to fail in cases where properties are mentioned without an associated chemical compound as this is the case with most of the qcl properties. The requirement to display a compound also leads to many false positives as the many records with characters in form of compounds are extracted.

Some of the properties such as the barrier/layer sequence have special characters such as the forward slash (/). A sample property of this is as follows: "42/67.8/23/96/34/73/40/206.2 nm". In some cases, the unit is also is put in brackets immediately after the value. Experimental analysis indicates failure by the AutoSentence parser in extracting these properties due to the unique combination of the special characters. In cases where the unit is mentioned after the property, the parser only extracts the last digits close to the unit and the unit(i.e 206.2 nm in the example property given).

In order to extract the material design (hetero-structure), working temperature, the lasing frequency, barrier thickness and the optical power, we define five efficient grammar parsing rules for the respective defined property models for these properties in form of property parsers. The parsers capture the phrase extraction rules expressed in form of regular expressions. These expressions are based on the selected keywords describing the various qcl parameters. The grammar rules of the parsers are defined based on a set of parser elements indicated in table 2 as defined in Chemdataextractor.

A parser typically consist of a prefix, value and the unit for the properties capturing physical quantities. The prefix contains the combined tokens of the various keywords used in identifying a property, the value contains the rules(in form of regular expressions) for matching the property value and the unit captures the units for the property.

For the qcl material parser, we define the 'heterostructure' as the main field to capture the value of the material design. We also only have the prefix(key phrases contextualizing the qcl material property) and the material attributes for the material design parser. The material consists of a series of regular expressions to match the qcl material names. The prefix and the material properties are combined to populate a complete heterostructure record.

Figure 1 shows a sample material design(hetero-structure) tree structure(upper) and extracted record output(lower) for a sample sentence from a journal paper:"We present two different terahertz quantum cascade laser designs based on GaAs/Al0.3Ga0.7As heterostructures that feature a depopulation mechanism of two longitudinal-optical phonon scattering events.".

```
b'<material>GaAs / Al0.15Ga0.85As</material>'

[{'QclMaterialDesign': {'heterostructure': 'GaAs / Al0.3Ga0.7As'}}]
```

Fig. 1. Sample Extracted Material Design(Heterostructure) Record.

For the barrier/layer thickness property parser, the rules are defined in such a way that the records are extracted with the special characters matched. The defined parsers interpret the manually defined grammar rules into an xpath parse tree from which the data model is constructed. The different parser elements, are combined with the "+" or "—" operators making the grammar rule flexible for update. The nested grammatical rules constitutes the specifier expression. The defined property rules are run for each document containing the qcl properties of interest. Algorithm 1 shows the workflow of the pipeline for extracting the qcl properties.

Algorithm 1 Mining the QCL properties

Input: D-Union{Ei}, input document object.
Output: : R-Union{Ri}

1 $S \leftarrow Union\{Si\}$ /* Prefix for the various keywords describing qcl properties */
2 $M \leftarrow PropertyModel$ /* specifies the fields to be captured for a particular property. */
3 $P \leftarrow Parser$ /* grammar logic. */
4 Set $D.model \leftarrow M$ and $parser \leftarrow P$ /* Defining the parser and property model. */
5 **for** *each document element Ei in D* **do**
6 | scan(Ts) **if** $Ti \leftarrow Si \subseteq S$ **then**
7 | | match Ri.
8 | **else**
9 | | skip to the next Ei and repeat step 6 and 7.
10 | **end**
11 **end**
12 **repeat** line 5-11 until all Ris are merged.
13 **return** *record R with the matched property relationship(s).*

We consider a document object D, capturing the various qcl properties of interest. D contains different document elements E1. . . En such as the title, paragraphs, sentences etc. A defined prefix S captures the possible expressions Si for a particular property. The expressions consists of keywords/a combination of the keywords used in context of a particular qcl property of interest. R consists of set of the property records which may consist of several of the individual record elements Ri. Ri captures the contextual property name, value and units. M and P consists of the defined property models and the parsers respectively. Before scanning though the document tokens, the property models and parsers have to be specified as shown in step 3 of the algorithm. The subsequent steps now involves searching for the matching token expressions for the prefix, property values, units and names which are merged into complete records.

4 Results and Discussions

4.1 Evaluation Metrics

In order to evaluate the performance of the proposed pipeline, we use the precision and recall as the evaluation metrics. In this context, the precision is the fraction of correct (relevant) records among all extracted records and the recall is the fraction of successfully extracted records among all correct (relevant) records in the articles. The word "correct" implies that the relationship of that record can be identified by a human when reading the corresponding sentence. In contrast, an "incorrect" (false) record suggests that a human expert cannot deduce the relationship of that record from the corresponding sentence. The metrics are determined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where TP is the true positive count (the number of correct records extracted), FP is the false positive count (the number incorrect records extracted), and FN is the false negative count (the number of correct records that are not extracted). The metrics are used to assess the chances of the pipeline leaving correct records unextracted and the chances of getting only the correct records in a given number of records.

4.2 Discussions

We use a sample of 43 open access articles as the evaluation dataset. The articles are randomly sampled from AIP, OPTICA and IOP publishers using the keyword "terahertz quantum cascade lasers". We restrict the sample to the articles describing proposed qcl designs and the corresponding performance characteristics. The articles consist of a total of 192 records manually extracted. The distribution of records is as follows: Optical power(33), Working Temperature (32), Lasing Frequency(44), Hetero-structure(66) and Barrier thickness 15 records. The records are manually extracted by an expert in quantum cascade lasers. The records are compared with those extracted by the pipeline in order to come up with the evaluation metrics. We compare the performance of our defined parsers and the default autosentence parser in chemdataextractor except for the sequence layer thickness

property. The performance evaluation of the pipeline on the sequence layer thickness is done separately as the autosentence is not used in extracting this property. This is owed to the inability of the autosentence parser in extracting these records. A correct extracted record is one that can be identified to correspond to the one manually extracted by an expert. Table 3 shows the evaluation metrics for the default autosentence parser and table 4 shows the performance evaluation metrics for our defined parsers.

Table 3. Performance Evaluation Metrics for the Default Autosentence Parser.

Property	**Records Extracted**	**TP**	**FP**	**FN**	**Precision(%)**	**Recall(%)**
Optical Power	16	11	5	22	68.75	33.33
Working Temperature	128	31	97	1	24.22	91.17
Lasing Frequency	16	10	6	34	62.50	22.73
Hetero-structure	140	50	90	16	35.71	75.76
Total	300	102	198	73	**47.80**	**55.75**

Table 4. Performance Evaluation Metrics for the Defined Parsers.

Property	**Records Extracted**	**TP**	**FP**	**FN**	**Precision(%)**	**Recall(%)**
Optical Power	25	24	3	9	96.00	72.73
Working Temperature	32	25	7	7	78.13	78.13
Lasing Frequency	37	29	8	15	78.34	65.91
Hetero-structure	60	59	1	7	98.33	89.39
Total	154	137	19	38	**87.70**	**76.55**

The autosentence parser achieves a precision of 68.75% and recall of 33.33% for the optical power property. This implies a higher number of correct records that are not extracted. Most the unextracted records are expressed in different contexts with varying keywords hence posing a chellenge to the default parser. The defined parser achieves a higher precision of 96.00% and has a higher recall of 72.73% indicating a lesser number of correct records that are not extracted. This is as indicated by the false negative(FN) values in tables 3 and 4. The defined rules take into consideration the various contexts in which the power values are expressed.

For the working temperature, the autosentence parser has a higher false positive rate hence resulting to a lower precision of 24.22%. One the other hand, the working temperature parser achieves minimal incorrect and unextracted records hence attaining a higher precision and recall of 78.13% and 78.13% respectively. This clearly indicates that as more temperatures are mentioned in literature, there is increased difficulty in retrieving the temperature of interest. The autosentence parser extracts most of the temperatures mentioned including the working temperature but has increased number of incorrect records extracted.

The default parser results to a recall of 22.73% for the lasing frequency. The defined parser on the other hand has a recall of 65.91% . The defined logic has therefore higher chances of extracting the correct records due to the specialized grammar rules. This is also pointed out by the higher precision of 78.34% for the defined parser. The lasing frequency

property is however expressed in many forms. This needs a wider definition of the rules hence the lower recall for both the autosentence parser and the defined parser. For the hetero-structure/material property, the default autosentence parser exhibits a higher false positive rate hence resulting to a precision of 35.71% and a recall of 75.76%. The higher false positive rate is attributed to the records having compound like names but are not necessarily qcl material names. The defined parser on the other hand achieves a precision of 98.33% and recall of 89.39%. The higher precision in attributed to the more specialized rule combination for the material design.

Overally, the default autosentence parser achieves a precision of 47.80% and a recall of 55.75%. The defined parsers on the other hand achieves a precision of 87.70% and a recall of 76.55%. This indicates a better performance of the defined parsers on the power, frequency, working temperature and the hetero-structure properties as shown in table 3. For the barrier thickness grammar logic, 2 records are left unextracted. The unextracted records consists of a combination of values with units in different positions and not after the reading. This results to a precision of 72.22% and a recall of 86.67%. The performance of the defined parsing rules in extracting the qcl barrier thickness indicates a great potential of their applicability on this property as they constitute the initial attempt to extract such properties with special characters.

5 Conclusion

In this paper, we propose a text mining pipeline for mining the qcl hetero-structure and the opto-electronic properties based on efficient rule based grammar logic. This is achieved by defining new parsing rules for the properties of interest in order to minimize the number of incorrect records extracted and the number of correct records not extracted. The rules are also able to match readings with special characters such as the qcl barrier thickness. Experimental analysis of comparative performance indicates better performance by the proposed rules. The work is however limited on open access articles for the specified publishers and more articles will be needed in future for extensive experimentation. The grammar logic is also limited to descriptions where the unit immediately follows the readings. We aim to extend this in future work to capture situations where the barrier thickness values are separated by commas and no unit mentioned after the value. We also aim to explore the integration of the named entity recognition with ontology population techniques in order to generate ontologies for the extracted properties.

Acknowledgement: This work was supported by the CNRS(French Centre National de la Recherche Scientifique) through the founding of a project within the Programme "Dispositif de Soutien aux Collaborations avec l'Afrique sub-saharienne". The authors would also like to thank the Strathmore University, School of Computing and Engineering Sciences and the Strathmore University Doctoral Academy for their involvement in creating the opportunity for this work to be produced and lastly, Qingyang Dong (University of Cambridge, Cavendish laboratory-molecular engineering group) for the insightful discussions.

Availability of Materials: The source code and the materials used for the production of this work are publicly available at our GitHub repository: `https://github.com/DeperiasKerre/qclProperties`.

References

1. Kumar, S., Hu, Q., Reno, J. L. (2009). 186 K operation of terahertz quantum-cascade lasers based on a diagonal design. Applied Physics Letters, 94(13), 131105.https://doi.org/10.1063/1.3114418
2. Vafapour, Z., Keshavarz, A., Ghahraloud, H. (2020). The potential of terahertz sensing for cancer diagnosis. Heliyon, 6(12), e05623.https://doi.org/10.1016/j.heliyon.2020.e05623
3. Shur, M., Liu, X. (2022, March). Biomedical applications of terahertz technology. In Advances in Terahertz Biomedical Imaging and Spectroscopy (Vol. 11975, p. 1197502). SPIE.https://doi.org/10.1117/12.2604800
4. Kanno, A., Dat, P. T., Sekine, N., Hosako, I., Kawanishi, T., Yoshida, Y., Kitayama, K. I. (2015). High-speed coherent transmission using advanced photonics in terahertz bands. IEICE Transactions on Electronics, 98(12), 1071-1080.https://doi.org/10.1103/PhysRevMaterials.4.123802
5. Rosati, E. (2018). The exception for text and data mining (TDM) in the proposed Directive on copyright in the Digital Single Market-technical aspects. Briefing Requested by the Juri Committee, European Parliament.https://doi.org/10.1093/jiplp/jpy063
6. Liang, H., Stanev, V., Kusne, A. G., Takeuchi, I. (2020). CRYSPNet: Crystal structure predictions via neural networks. Physical Review Materials, 4(12), 123802.https://doi.org/10.1103/PhysRevMaterials.4.123802
7. 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. DOI: 10.1021/acs.jcim.6b00207https://doi.org/10.1021/acs.jcim.6b00207
8. Hawizy, L., Jessop, D. M., Adams, N., Murray-Rust, P. (2011). ChemicalTagger: A tool for semantic text-mining in chemistry. Journal of cheminformatics, 3, 1-13.https://doi.org/10.1186/1758-2946-3-17
9. Corbett, P., Copestake, A. (2008). Cascaded classifiers for confidence-based chemical named entity recognition. BMC bioinformatics, 9(11), 1-10.https://doi.org/10.1186/1471-2105-9-S11-S4
10. García-Remesal, M., García-Ruiz, A., Pérez-Rey, D., De La Iglesia, D., Maojo, V. (2013). Using nanoinformatics methods for automatically identifying relevant nanotoxicology entities from the literature. BioMed research international, 2013.https://doi.org/10.1155/2013/410294
11. Lafferty, J., McCallum, A., Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
12. Hochreiter, S., Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.https://doi.org/10.1162/neco.1997.9.8.1735
13. Lowe, D. M., Sayle, R. A. (2015). LeadMine: a grammar and dictionary driven approach to entity recognition. Journal of cheminformatics, 7(1), 1-9.https://doi.org/10.1186/1758-2946-7-S1-S5
14. Hawizy, L., Jessop, D. M., Adams, N., Murray-Rust, P. (2011). ChemicalTagger: A tool for semantic text-mining in chemistry. Journal of cheminformatics, 3, 1-13.https://doi.org/10.1186/1758-2946-3-17
15. Jessop, D. M., Adams, S. E., Willighagen, E. L., Hawizy, L., Murray-Rust, P. (2011). OSCAR4: a flexible architecture for chemical text-mining. Journal of cheminformatics, 3(1), 1-12.https://doi.org/10.1186/1758-2946-3-41
16. Rocktäschel, T., Weidlich, M., Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. Bioinformatics, 28(12), 1633-1640.https://doi.org/10.1093/bioinformatics/bts183
17. Leaman, R., Wei, C. H., Lu, Z. (2015). tmChem: a high performance approach for chemical named entity recognition and normalization. Journal of cheminformatics, 7(1), 1-10.https://doi.org/10.1186/1758-2946-7-S1-S3

18. Dong, Q., Cole, J. M. (2022). Auto-generated database of semiconductor band gaps using chemdataextractor. Scientific Data, 9(1), 193.https://doi.org/10.1038/s41597-022-01294-6
19. Sierepeklis, O., Cole, J. M. (2022). A thermoelectric materials database auto-generated from the scientific literature using ChemDataExtractor. Scientific Data, 9(1), 648.https://doi.org/10.1038/s41597-022-01752-1
20. Huang, S., Cole, J. M. (2020). A database of battery materials auto-generated using ChemDataExtractor. Scientific Data, 7(1), 260.https://doi.org/10.1038/s41597-020-00602-2
21. Zhao, J., Cole, J. M. (2022). A database of refractive indices and dielectric constants auto-generated using chemdataextractor. Scientific data, 9(1), 192.https://doi.org/10.1038/s41597-022-01295-5
22. Court, C. J., Cole, J. M. (2018). Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. Scientific data, 5(1), 1-12.https://doi.org/10.1038/sdata.2018.111
23. Mavracic, J., Court, C. J., Isazawa, T., Elliott, S. R., Cole, J. M. (2021). ChemDataExtractor 2.0: Autopopulated ontologies for materials science. Journal of Chemical Information and Modeling, 61(9), 4280-4289.https://doi.org/10.1021/acs.jcim.1c00446
24. He, T., Sun, W., Huo, H., Kononova, O., Rong, Z., Tshitoyan, V., ... Ceder, G. (2020). Similarity of precursors in solid-state synthesis as text-mined from scientific literature. Chemistry of Materials, 32(18), 7861-7873.https://doi.org/10.1021/acs.chemmater.0c02553
25. Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K. A., ... 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, 59(9), 3692-3702.https://doi.org/10.1021/acs.jcim.9b00470
26. Korvigo, I., Holmatov, M., Zaikovskii, A., Skoblov, M. (2018). Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules. Journal of cheminformatics, 10(1), 1-10.https://doi.org/10.1186/s13321-018-0280-0
27. Kononova, O., Huo, H., He, T., Rong, Z., Botari, T., Sun, W., ... Ceder, G. (2019). Text-mined dataset of inorganic materials synthesis recipes. Scientific data, 6(1), 203.https://doi.org/10.1038/s41597-019-0224-1
28. Cruse, K., Trewartha, A., Lee, S., Wang, Z., Huo, H., He, T., ... Ceder, G. (2022). Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities. Scientific Data, 9(1), 234.https://doi.org/10.1038/s41597-022-01321-6
29. Huang, S., Cole, J. M. (2022). BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. Journal of Chemical Information and Modeling, 62(24), 6365-6377.https://doi.org/10.1021/acs.jcim.2c00035
30. Zhao, J., Huang, S., Cole, J. M. (2023). OpticalBERT and OpticalTable-SQA: Text-and Table-Based Language Models for the Optical-Materials Domain. Journal of Chemical Information and Modeling.https://doi.org/10.1021/acs.jcim.2c01259
31. Milosevic, N., Gregson, C., Hernandez, R., Nenadic, G. (2019). A framework for information extraction from tables in biomedical literature. International Journal on Document Analysis and Recognition (IJDAR), 22, 55-78.https://doi.org/10.1007/s10032-019-00317-0
32. Mukaddem, K. T., Beard, E. J., Yildirim, B., Cole, J. M. (2019). ImageDataExtractor: a tool to extract and quantify data from microscopy images. Journal of chemical information and modeling, 60(5), 2492-2509.https://doi.org/10.1021/acs.jcim.9b00734
33. Kim, H., Han, J., Han, T. Y. J. (2020). Machine vision-driven automatic recognition of particle size and morphology in SEM images. Nanoscale, 12(37), 19461-19469.https://doi.org/10.1039/D0NR04140H