processes

MDPI

Article

IP Analytics and Machine Learning Applied to Create Process Visualization Graphs for Chemical Utility Patents

Amy J. C. Trappey [1,*], Charles V. Trappey [2], Chih-Ping Liang [1] and Hsin-Jung Lin [1]

[1] Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu 300, Taiwan; s109034534@m109.nthu.edu.tw (C.-P.L.); tinairislin@gmail.com (H.-J.L.)
[2] Department of Management Science, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan; trappey@nycu.edu.tw
* Correspondence: trappey@ie.nthu.edu.tw; Tel.: +886-3-5742651

Abstract: Researchers must read and understand a large volume of technical papers, including patent documents, to fully grasp the state-of-the-art technological progress in a given domain. Chemical research is particularly challenging with the fast growth of newly registered utility patents (also known as intellectual property or IP) that provide detailed descriptions of the processes used to create a new chemical or a new process to manufacture a known chemical. The researcher must be able to understand the latest patents and literature in order to develop new chemicals and processes that do not infringe on existing claims and processes. This research uses text mining, integrated machine learning, and knowledge visualization techniques to effectively and accurately support the extraction and graphical presentation of chemical processes disclosed in patent documents. The computer framework trains a machine learning model called ALBERT for automatic paragraph text classification. ALBERT separates chemical and non-chemical descriptive paragraphs from a patent for effective chemical term extraction. The ChemDataExtractor is used to classify chemical terms, such as inputs, units, and reactions from the chemical paragraphs. A computer-supported graph-based knowledge representation interface is developed to plot the extracted chemical terms and their chemical process links as a network of nodes with connecting arcs. The computer-supported chemical knowledge visualization approach helps researchers to quickly understand the innovative and unique chemical or processes of any chemical patent of interest.

Keywords: chemical utility patents; knowledge graph visualization; text mining; machine learning; bidirectional encoder representations (ALBERT); chemical manufacturing process visualization; IP analytics

check for updates

Citation: Trappey, A.J.C.; Trappey, C.V.; Liang, C.-P.; Lin, H.-J. IP Analytics and Machine Learning Applied to Create Process Visualization Graphs for Chemical Utility Patents. *Processes* **2021**, *9*, 1342. https://doi.org/10.3390/pr9081342

Academic Editor: Jae-Yoon Jung

Received: 9 July 2021
Accepted: 29 July 2021
Published: 30 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Researchers studying new materials or novel chemical processes rely on prior art in published technical documents and patented processes registered with patent agencies, such as the United States Patent and Trademark Office (USPTO) or the European Patent Office (EPO). Reading lengthy technical documents to distinguish between the non-obvious and original chemical processes from the obvious and nonoriginal processes is a significant challenge. The researcher must be capable of organizing the experimental steps, understanding patent claims, and recognizing if processes used in-house may be infringing on the rights of others. If there is infringement, then the manufacturer must approach the owner to either buy the IP or acquire a license for the chemical process to continue the development of a new product for the company. Both the product and the process to make the product may not be new, and the ownership rights may have expired after a period of about twenty years. The cost and time delays increase if development money is spent re-inventing outdated processes or failing to discover that the process or rights to make the new chemical must be purchased directly by buying ownership of the patent or negotiating a license agreement to manufacture with the patent owner's consent and terms. Expired

patents and prior art literature may be useful for research but may not represent the most cost-effective and currently accepted safety practices and standards of manufacturing. The currently registered chemical utility patent documents must be reviewed in a timely manner for effective research and development (R&D). The descriptions of the chemical processes are not always clearly described in the research literature and patents. The major challenge for researchers in any given field is the required reading and reviewing of huge numbers of documents (both patent and non-patent literature), which dilutes and delays the researchers' inventive R&D progress [1,2]. In a competitive market, the sooner the product is developed, the better it will perform [3]. This research focuses on developing a computer-supported patent knowledge graph generation system to effectively depict chemical process knowledge. The creation of a visual and logical knowledge-based process graph helps untangle the descriptive text of a chemical utility patent. The knowledge graph shortens the time spent by researchers on prior-art reviews and enables researchers to devote more time to pursue novel and patentable research endeavors.

IP analytics includes text mining techniques, such as high term frequency (TF) word extraction and stop word removal (words with low knowledge relevance, such as pronouns and special characters), to speed up the classification, ranking, and clustering of the information under topics within a knowledge domain. IP analytic tools are effective in assisting researchers to summarize documents and provide an abstract of keywords that represent the essential content. The text is processed and analyzed by computer algorithms, and researchers read the abstract output, which saves time and creates a logical structure of the domain knowledge. There are limitations when using traditional text mining and IP analytic techniques to process chemical literature [4,5]. The traditional method of word tokenization tends to break down specific chemical terms incorrectly. When chemical noun phrases, such as sodium chloride are divided into sodium and chloride, the meaning is completely different, resulting in an incorrect representation of the chemical formula for salt. The traditional method treats numbers as a stop word and removes numbers that do not have a relationship to the domain, such as the numbering of text sections. However, this approach is not applicable to chemical nouns or noun phrases. If the stop word algorithm removes the numbers originally used to describe the atomic structure and molecular weights, this will omit important information identifying the chemicals and lower the effectiveness of text mining the processes. Therefore, it is necessary to process chemical terms in a non-traditional way to correctly represent the information in this domain.

In order to correctly extract chemical-related content from text and to accelerate computer execution, our research integrates IP text analytics with machine learning for text paragraph classification. The approach identifies chemical-related and non-chemical-related paragraphs and pre-selects the chemical-related text for extracting chemical terms and processes. By extracting chemical-related paragraphs using machine learning, the system can act autonomously without relying on researchers to read and classify the text and paragraphs. In addition, classifying the important paragraphs using a machine learning model reduces the time spent on text mining. The system only analyzes the paragraphs classified by the model as a chemical process instead of the whole patent, reducing the computation time with improved results. A tool specifically designed to extract chemical terms and related verbs for chemical nouns and noun phrases are abstracted from the paragraphs describing the processes, which yields the chemical knowledge domain keywords to create a corpus. For example, ChemDataExtractor had been used by researchers for article retrieval, data extraction, data cleaning, data post-processing, and evaluation [6]. Traditionally, using a computer system for IP analytics and document abstraction is often limited to the knowledge classification and clustering of important keywords within the text. The graphical representation of patent/IP knowledge, particularly for chemical processes, is a valuable computer-supported function. Some chemical process research [7] has started to show the graph-based modeling potential for the visualization of chemical process knowledge, although using the machine learning approach to help effectively and

accurately generate the graph-based knowledge representation for chemical-related patents needs much more research and development. For example, NetworkX is a python package for the study of complex networks [8]. Therefore, this study emphasizes graphically presenting the chemical process described in patent text in order to increase the intuitive understanding of a chemical patent selected by the user.

The paper is organized into several sections. In Section 2, the literature related to machine learning (ML) for natural language processing (NLP), classification, and graph-based knowledge representation (GKR) is reviewed. In Section 3, the framework for the methodology and step-by-step ML and GKR techniques are explained. Section 4 presents the case example of computer-supported generation of graph-based knowledge representation disclosed in polymer patents. Finally, the research outcomes, contributions, conclusions, and future research are discussed. Our purpose is to better enable researchers to understand the chemical parts of a patent. Chemical parts are represented by graphs with nodes and lines. With the assistance of graphs, researchers can quickly comprehend the process. Moreover, graphs are also comparable, which enables researchers to discuss the similarity and differences of patents.

2. Literature Review

This research emphasizes three key methodologies as essential modules for computer-supported chemical patent knowledge graph generation. These methodologies are machine learning modeling for natural language processing (NLP), chemical process-related text mining, and knowledge graph creation. NLP machine learning is used by the ALBERT model for semantic text classification. The training and testing procedure of a specific ALBERT model for chemical-related and non-chemical-related paragraph classification are described in Section 3. The case example of a polymer patents' chemical and non-chemical paragraph classification is provided in Section 4. For chemical-related paragraph chemical text mining, the ChemDataExtractor is reviewed and added to the system development framework with a case demonstration in Sections 3 and 4. The knowledge graph approaches are discussed in Section 2.3 and describe the NetworkX spring layout graph for chemical process visualization.

2.1. Chemical Text Mining

The objective of preprocessing for chemical text mining is to retrieve tokenized chemical names and text so the algorithm can train the word vectors [9–13]. The first step is to download papers from a website within the domain of interest. The ALBERT algorithm parses the HTML/XML to clean up the text and treat chemical names as a single unit. Finally, the project trains the word vectors [14]. The process of chemical text mining is to determine the paragraphs within the patent that describe the manufacturing process steps and then analyze the labels. ALBERT segments the paragraphs in the document and creates a database. Next, the ChemDataExtractor codifies the words extracted by the natural language process sequence as shown in Figure 1.

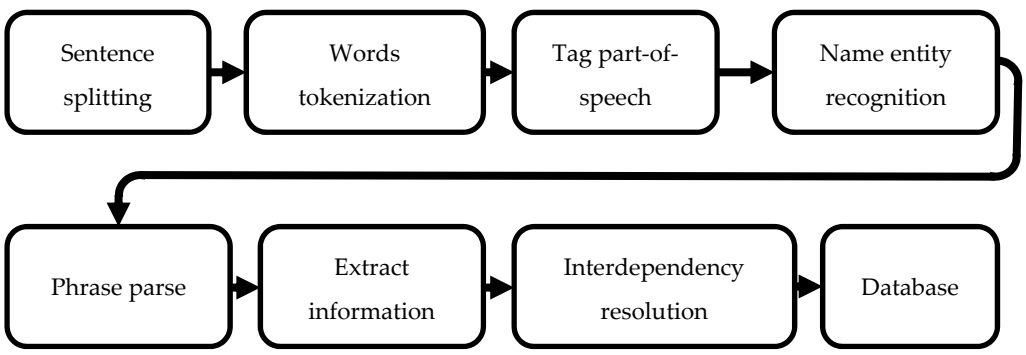

Figure 1. Standard natural language preprocess for text mining.

Machine learning methods are used in the chemical domain and directly extract relationships among text in the raw data [15–18]. The tool Chemical Tagger is frequently used for chemical text mining [19] and parses text while classifying the chemical information (e.g., reactant, solvent) in the text files [20,21]. Chemical Tagger can be used to define relationships with actions, such as heated and stirred. The common tools include OSCAR4, SpaCy, and ChemDataExtractor [22]. In this article, ChemDataExtractor is the text mining algorithm used in the case examples given its extensive use in peer-reviewed academic articles.

ChemDataExtractor is a tool for automatically extracting chemical information from scientific documents and provides a wide range of state-of-the-art natural language processing algorithms [23,24] to interpret the English language text of scientific documents. The state-of-the-art methods include a chemistry-aware part-of-speech (POS) tagger, a named entity recognizer that combines conditional random fields and dictionaries, a rule-based grammar for phrase parsing [25], and a word clustering algorithm to improve the performance of unsupervised machine learning [26,27]. The ChemDataExtractor extracts information from semi-structured tabulated data using a table parser and resolves data interdependencies between information extracted from different parts of a document by using document-level post-processing algorithms.

The algorithm uses tokenization, word clustering, part of speech tagging, a chemical named entity recognition, phrase parsing, and chemical identifier disambiguation. During the tokenization procedure, text data passages are transformed into sentences and then split into individual words and punctuation, allowing a flow of tokens that are suitable for NLP [28]. The Brown word clustering algorithm generates a pattern of the hierarchical grouping of words based on the contexts in which they occur and improves the performance of part-of-speech tagging and named entity recognition [26]. POS tagging nominates a tag to each token that represents its syntactic function. Chemical named entity recognition [29,30] uses a CRF (conditional random field)-based recognizer for chemical names and is combined with a dictionary-based recognizer to improve performance by removing trivial terms and trade names [31]. The parsing process prevents a single sentence from being parsed in different ways to produce different meanings. Finally, chemical identifier disambiguation shows a series of mappings between their matching full, unabbreviated names and abbreviations. The mapping results combine information that is defined with regard to dissimilar identifiers into single nodes for each particular chemical entity. This tool can extract chemical data from documents automatically, taking less effort and time to simplify the creation of large chemical databases. The ChemDataExtractor consists of a modular document procedure channel for POS analysis [26].

2.2. Paragraph Classification

BERT (bidirectional encoder representations from transformers) is a two-way encoder representation transformer [32,33]. This is the language model that Google uses in unsupervised learning with a large number of unannotated texts. The model makes use of transformers, which is an encoder that reads the text input and a decoder that produces a prediction for the task. The results depend on many model parameters, so the time and resource costs required to train a set of BERT models are very large, and even the final complex model may not be effective.

ALBERT (a lite bidirectional encoder representations from transformers) is a refined BERT model proposed by Google in 2020 [34]. It uses several optimization strategies to obtain a much smaller model than BERT. ALBERT achieves the effect of reducing the parameters by reducing the dimensionality of the embedding part through matrix decomposition. The formula is described as follows: V is for vocabulary size, E is for embedding size, and H is for hidden size:

$$\text{Parameter size } O(V \times H) \rightarrow \text{Parameter size } O(V \times E + E \times H) \quad (1)$$

Figure 2 shows the one-hot vector input of BERT. The first projection indicates that there is no interaction between words. Therefore, the first projection does not require high-dimension vectors.

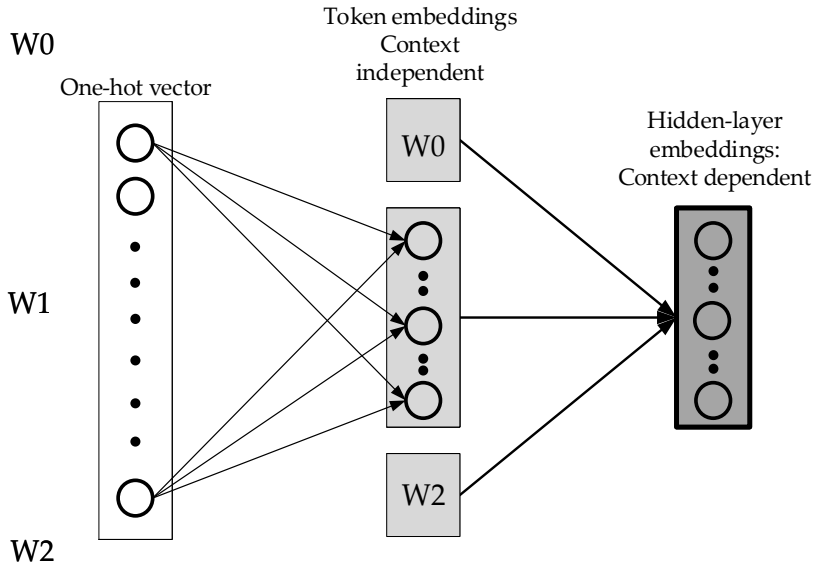

Figure 2. One-hot vector input of BERT.

Figure 3 below is an improvement of ALBERT, which breaks the one-hot vector into two different matrices, and fewer parameters are needed. The advantage is that the context-independent expression of the word and the context-dependent expressions are unlocked. Therefore, the context-dependent expressions can be increased, which means the network becomes wider. Using a one-hot vector requires fewer parameters to be set for the first mapping.

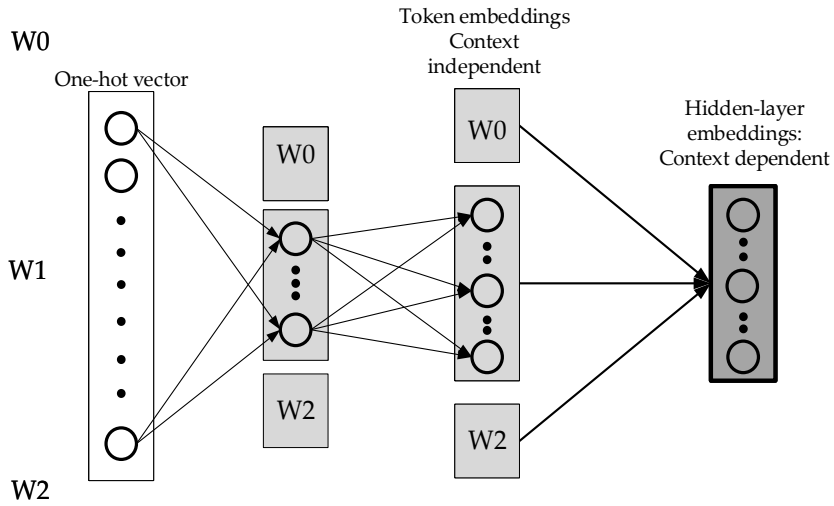

Figure 3. One-hot vector input of ALBERT.

Another way for ALBERT to reduce the number of parameters is to share parameter settings between multiple layers. There are three ways of parameter sharing: sharing only the parameters of the feed-forward network, sharing only the parameters selected and sharing all parameters. ALBERT shares all parameters by default. ALBERT improves parameter setting efficiency, self-supervised learning tasks, and model capacity. To reduce the complexity of the word vector space, ALBERT shares weights and creates a composite matrix to reduce parameters.

2.3. Graph Visualization

The research uses graph visualization to present the experimental process using NetworkX, a python program that enables an improved layout to visualize the chemical process flow.

2.3.1. NetworkX

NetworkX is a network construction tool based on graph theory. For easy analysis of complex network data and simulation models, NetworkX has algorithms for complex network analysis and graph visualization. By using NetworkX, the project can perform network visualizing, analyze the network architecture, save networks by using normalized and non-normalized data formats, create network calculations, construct a wide variety of classic networks and random networks, and develop network models [35]. Simple undirected graphs, multigraphs, and directed graphs are supported by NetworkX. In addition, it contains many graph algorithms. The boundary value dimension is unlimited, and nodes in the graph can be created by any data. The tool is easy to use and is rich in functions. NetworkX uses graphs as the basic data structure. Graphs can be generated by programs, online data sources, files, and databases [36].

2.3.2. Spring Layout

Force-directed algorithms are also known as spring embedders. They are flexible methods for calculating graph layouts. A graph layout is calculated by using the information contained within the graph's structure. A spring layout can be drawn with these algorithms (Figure 4). Applying force-directed algorithms, graphs are aesthetically more pleasing and symmetric for visualization [37]. The nature of the spring embedding model has three main points. First, edges are springs with a constant ideal length. Second, nodes repel each other (repulsive force). Third, edges cannot be arbitrarily long (attractive force) [38]. In terms of Hooke's law in physics, to achieve the ideal state of a spring, the spring force is linear in length and convergence. That is, a state of force balance is reached and is calculated using coordinates p_u, p_v for each node u and v, and lengths l_e for each edge e:

$$\text{repulsive force}: \ f_{rep}(u,v) = \frac{l_{uv}^2}{||p_u - p_v||} \cdot \vec{p_u p_v} \tag{2}$$

$$\text{attractive force}: \ f_{attr}(u,v) = \frac{||p_u - p_v||^2}{l_e} \cdot \vec{p_u p_v} \text{ for } (u,v) = e \in E \tag{3}$$

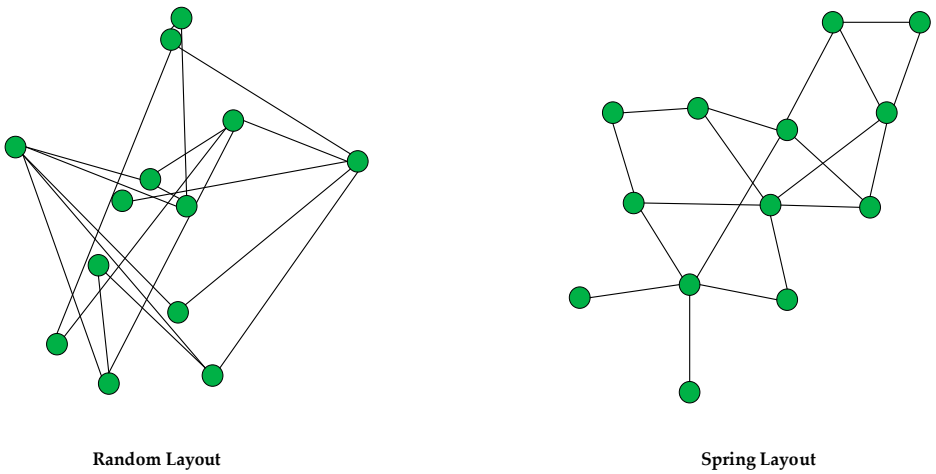

Figure 4. Random Layout and Spring Layout.

For the iterate calculation, calculate the force vector $F(v)$ of each node v and then move each node v according to the force vector:

$$p_v + \delta \cdot F(v) \rightarrow p_v \quad (4)$$

Using the spring layout, the graph is presented as a symmetrical structure, which helps visualize chemical processes that require reading graphical details consisting of chemical names and processes in a logical order.

3. Methodology

The structure and framework for the computer-supported chemical patent knowledge graph generator are shown in Figure 5. The first step is to download the hypertext markup language (HTML) of a chemical patent of interest from Google Patent. Using the HTML file, the irrelevant HTML instructions for markup are excluded leaving the source text. After extracting the chemical patent text designating the paragraphs that describe the chemical steps, ChemDataExtractor is used to extract keywords and label the text in each sentence. Four process labels are used, including target, precursor, operations, and condition. The labeled words are stored in the database. The system places the words stored in each grid in the database as a node, displays different colors for each label, and then connects to the nodes. The system sets target, precursor, and operations as main nodes and conditions as sub-nodes. The extractor algorithm places each node with a solid line according to the sequence of the steps, while the sub-nodes are connected to the main nodes with a dotted line. Finally, the entire graph ends in the final chemical output.

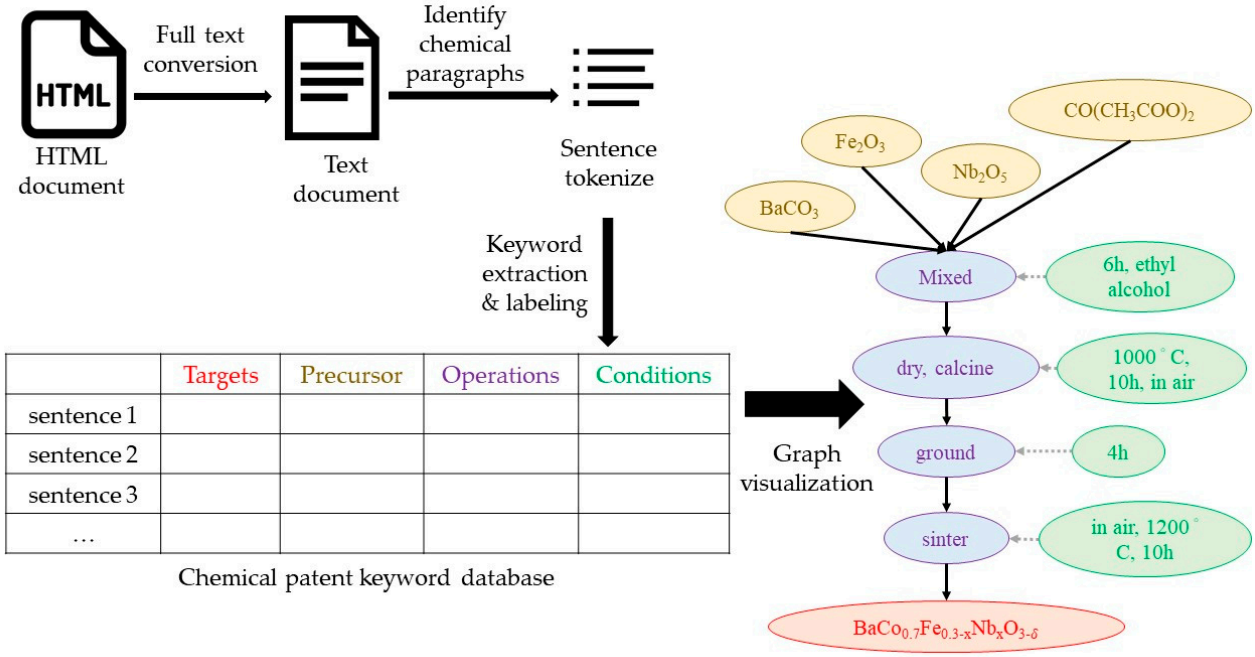

Figure 5. The system framework of computer-supported chemical patent knowledge graph generation.

As shown in Figure 6, the system workflow begins with retrieving the HTML file of any patent on the web for the patent's text extraction. The sentence tokenization splits the sentences and analyzes whether the sentences belong to the chemical process description classification. The sentences are labeled as part of chemical paragraphs, and the sentences that are part of non-chemical paragraphs are omitted. After finishing the identification and recognition of chemical-related sentences and paragraphs, the information is labeled and saved in a database. In the last step, the nodes and links are created using the labeled chemical terms to create the chemical process patent knowledge graph. The system modules for the entire workflow are implemented using Python programs. The

main Python package versions are listed in Appendix B. Further, Appendix A depicts the pseudo-code of the proposed system.

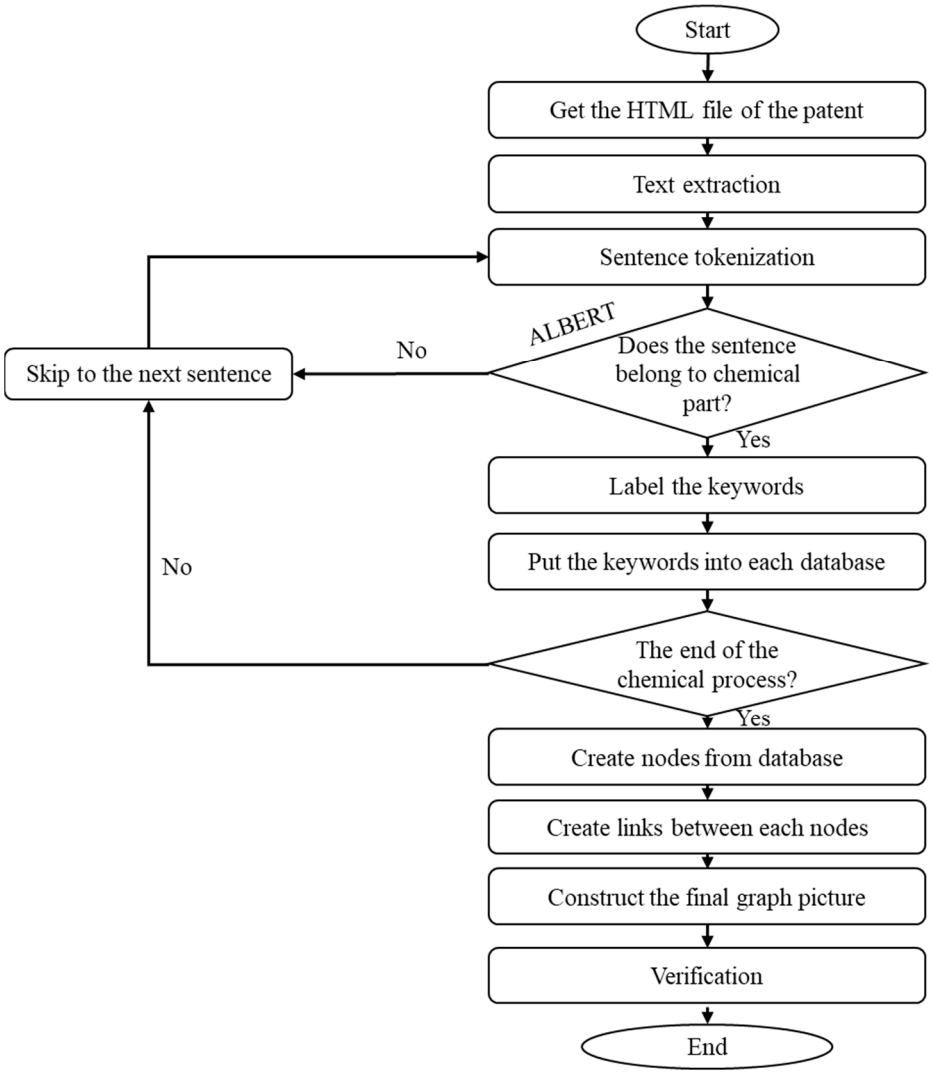

Figure 6. The workflow of three key modules for generating chem-process graphs through patent text mining.

4. Case Implementation—Polymer Patent Graphs

This research chooses polymer patents for the two case examples. Polymers are hydrogen chain molecules with large structures and weight, which are the most common materials used in our daily lives, such as plastics and synthetic fibers [39]. From the search of the global patent publications, the numbers of polymer patents have steadily increased over the past decade, with more than 6000 polymer-related patents published in 2020 [40]. For our case example, patent number US5132253A [41], the Sol-gel method for making ceramic materials, is a classic polymer patent invented by Corning researchers in early 1990. This patent describes a method of synthesizing alkaline earth-containing materials by using solvent and metal processing. For the second case example, patent number US8455097B2 [42], the coating liquid invention for covering glass fiber and rubber-reinforcing glass fiber, is a chemical (polymer) patent, granted to Central Glass in 2013. This patent describes a glass-fiber coating liquid for enhanced adhesion between the rubber-reinforcing glass fiber and the base rubber. In the following sub-sections, the step-by-step computer-supported chemical patent knowledge extraction and visualization, following

Figure 6's workflow, are described, and the results (of both cases' process graphs) are shown using the proposed methodology.

4.1. Document Preprocess

First, the HTML file is retrieved because the chemical names, amounts, and process steps are easier to abstract in HTML than in excel or pdf file format. After downloading the HTML file, a Python program extracts the information without HTML markups. In our system, the description part of the patent is used as the analysis input. A program commonly used to interpret the structure of chemicals called SMILE (simplified molecular-input line-entry specification) is applied. SMILE is a standard specification that uses ASCII strings to describe the molecular structure [43]. The chemical structure is converted into characters. Some examples of conversion rules are listed in Table 1.

Table 1. Some conversion rules of SMILE.

Rule	Description
1	Atoms are represented by their respective atomic symbols.
2	Omit simple H connections.
3	Adjacent atoms mean connected to each other.
4	Double bond and triple bond are represented by "=" and "#", respectively (single bond and the aromatic bonds can be omitted).
5	The branch is represented by "()".
6	Use assigned numbers to indicate the atoms connected in the ring.
7	The ring structure is cleaved to form a chain structure, and the cleavage site is numbered.

Different classes of text can be found directly from HTML to obtain a specific type of paragraphs. Google Patent's patents have a structure that is divided into the title, abstract, claim, description, citations, similar documents, priority and related applications, and concepts. The specific text paragraphs are retrieved by specifying the class in HTML, eliminating the need for the computer to analyze the complete patent. The program extracts the description class from the HTML patent. The text is separated with line breaks as segments and is stored in the chemical paragraph as a list. The next step is to train the paragraph recognition model.

4.2. Chemical Paragraph Recognition Model

Training the ALBERT model and testing use 1050 USPTO patents as the dataset for system prototyping (additional patents can be added for future system refinement). These patents are searched and found under the relevant International Patent Classifications (IPCs) as listed in Table 2. IPC C08, organic macromolecular compounds, are known as polymers. The project randomly chooses 150 patents to form each sub C08 category. Choosing chemical paragraphs and non-chemical paragraphs from these domain patents, 2294 paragraphs are retrieved for the training dataset. After splitting 90% of the dataset as the training set and 10% for the testing dataset, the results show that the accuracy of the model is 93.7%. Figure 7 demonstrates that the model test accuracy is 90% on average. The loss function demonstrates that the model is not overfitting, which means that the model is effective.

The research places tokenized paragraphs into the ALBERT model, which will recognize whether the paragraph is classified as a chemical paragraph or not. The model will mark each paragraph as yes or no. The identifier yes represents paragraphs that belong to the chemical paragraphs set. Table 3 shows the ALBERT classification outputs of a chemical patent [44]. Some descriptive paragraphs are recognized as non-chemical paragraphs. Afterward, the recognized chemical paragraphs are split into sentences for the following extraction of chemical terms.

Table 2. List of International Patent Classifications (IPCs).

IPC	Description
C08B	Polysaccharides substance and its derivatives thereof.
C08C	Rubbers' treatment or the chemical modification.
C08F	Macromolecular compounds are made by involving carbon-to-carbon unsaturated bonds reactions.
C08G	Macromolecular compounds are made otherwise by involving carbon-to-carbon unsaturated bonds reactions.
C08H	Derivatives of natural macromolecular compounds.
C08J	Working-up; general processes of compounding; after-treatment.
C08K	Use of inorganic or non-macromolecular organic substances as compounding ingredients.
C08L	Compositions of macromolecular compounds.

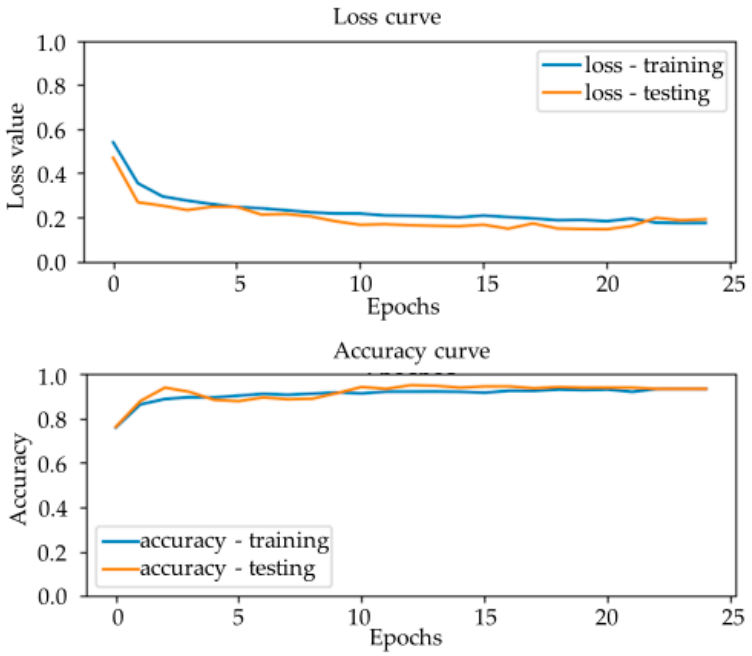

Figure 7. Test results of model loss values and accuracy.

Table 3. Sample outputs of chemical and non-chemical paragraph recognition.

Paragraph Type	Paragraph Example
Non-chemical	"Recently, there is a growing demand for high performance, high stability secondary batteries as electric, electronic, communication, and computer industries have rapidly developed. Particularly, in line with miniaturization and lightweight trends of . . . "
Chemical	"44.0 g of acrylic acid, 55.0 g of acrylonitrile, 1.0 g of 2-hydroxyethyl acrylate, 0.25 g of N,N-methylenebisacrylamide, 122.1 mL of a 4 mol/L sodium hydroxide aqueous solution, and 757.9 g of ion exchange water were put in a 2000 mL solution . . . "

The paragraphs selected for the case demonstrations were extracted using ALBERT.

Case 1: "Under nitrogen gas, 0.80 g Ba metal is first reacted with 20 mL of methanol, with the evolution of H_2. In a separate flask, 1.99 g Mg $(OC_2\ H_5)_2$, 2.36 g Al $(OC_3\ H_7)_3$, and 2.41 g Si $(OC_2\ H_5)_4$ are dissolved and refluxed in 142 mL 2-methoxyethanol and 8 mL HNO_3. The resulting solution is added to the barium solution to produce a cloudy white sol." [41] The cloudy white sol is hydrolyzed and heated after adding other solutions to obtain a crystallized product, which is one of the important raw materials for the final ceramic matrix material.

Case 2: "More specifically, 100 parts by weight of the chlorosulfonated polyethylene (C) available as the halogen-containing polymer (G) under the trade name of TS-430 from Tosoh Corporation, 40 parts by weight of p-dinitrosobenzene, 0.3 parts by weight of N-N≡-hexamethylene diallylnagiimide available under the trade name of BANI-H from Maruzen Petrochemical Co., Ltd. (Tokyo, Japan) and 30 parts by weight of carbon black were mixed together, followed by dispersing the resultant mixture into 1315 parts by weight of xylene to obtain the secondary glass-fiber coating liquid. The contents percentages of N-N≡-hexamethylene diallylnagiimide as the bis-allylnagiimide (H), p-dinitrosobenzene as the vulcanization agent (L), and carbon black as the inorganic filler (N) in the secondary glass-fiber coating liquid were H/G = 0.3 wt %, L/G = 40 wt % and N/G = 30 wt %, respectively, based on the weight of the chlorosulfonated polyethylene (C)." [42] The secondary glass-fiber coating liquid is used for enhancing adhesion between the glass fiber and the base rubber.

4.3. Label

Each sentence that contains operation verbs represents a process step. After breaking down the steps, the program labels each word in the sentence. The system uses ChemDataExtractor to do the word tokenization, which correctly identifies the material or chemical names and labeling each word to identify different word properties. Words will be identified in different part-of-speech processes. This project will use number, noun, and verb to construct unit, substance, and operation verbs.

Labeling is a straight-forward rule. First, if the number is followed by a noun, combine these two and label it as a unit. Second, if the noun is followed by adjectives or adverbs, combine them as a noun. Third, if the noun is followed by a noun, combine these two, and label it as a substance. Fourth, if the noun does not have a noun before and after, it will be labeled separately as a substance. Fifth, remove the operation verb separately and label it as an action. As shown in Tables 4 and 5, the example terms in both patents' chem-paragraphs are properly extracted and labeled.

Table 4. Example list of chemical terms after labeling for patent US5132253A.

Term	Label
"1.99 g"	unit
"$Mg(OC_2H_5)_2$"	substance
"2.36 g"	unit
"$Al(OC_3H_7)_3$"	substance
"2.41 g"	unit
"$Si(OC_2H_5)_4$"	substance
"dissolved"	action
"refluxed"	action
"142 mL"	unit
"2-methoxyethanol"	substance
"8 mL"	unit
"HNO_3"	substance

Table 5. Example list of chemical terms after labeling for patent US8455097B2.

Term	Label
"100 parts"	unit
"chlorosulfonated polyethylene"	substance
"40 parts"	unit
"p-dinitrosobenzen"	substance
"0.3 parts"	unit
"N-N≡-hexamethylene diallylnagiimide"	substance
"30 parts"	unit
"carbon black"	substance
"mixed"	action
"dispersing"	action
"1315 parts"	unit
"xylene"	substance
"secondary glass-fiber coating liquid"	substance

4.4. Graph Visualization

The NetworkX program, described in Section 2.3, is adopted as the graph visualization module. The graph generation module assists users in defining the relation-links and nodes between chemical terms that were extracted and labeled in the previous step (Tables 4 and 5). The elements in the From list will connect to the elements in the To list. All elements are categorized based on their properties. After completing the link definition, the chemical process depicting the essential chemical patent knowledge is presented as a graph-based network.

The graph visualization prototype contains three node types: the orange node represents the unit, the gray node is a substance, and the blue node is an action. Since all elements are labeled in categories, the labels will become the properties of the elements. The NetworkX elements' properties are defined as ID and type. ID is the name of element, and the word is shown on the node. Type represents the properties of the nodes shown in the type's color (orange, gray, or blue). The example patent [41] knowledge graph is shown in Figure 8. The spring layout enables the users reviewing the chemical process graph to intuitively understand the critical processes, ingredients, amounts, and actions. The chemical process starts from the middle of Figure 8 and ends in the upper left corner's "cloudy white sol" node, which is an important solution to produce ceramic matrix material. For the second patent case [42], a knowledge graph for the given chem-paragraph is generated and shown in Figure 9. Figures 8 and 9 show how the spring layout of a process graph better enables the users to review and understand the chemical process by depicting the ingredients, quantities, units, and actions. In Figure 9, the chemical process ends in the upper right corner's "secondary glass-fiber coating liquid," applicable as a reinforcement layer for enhancing adhesion between glass-fiber and its base rubber.

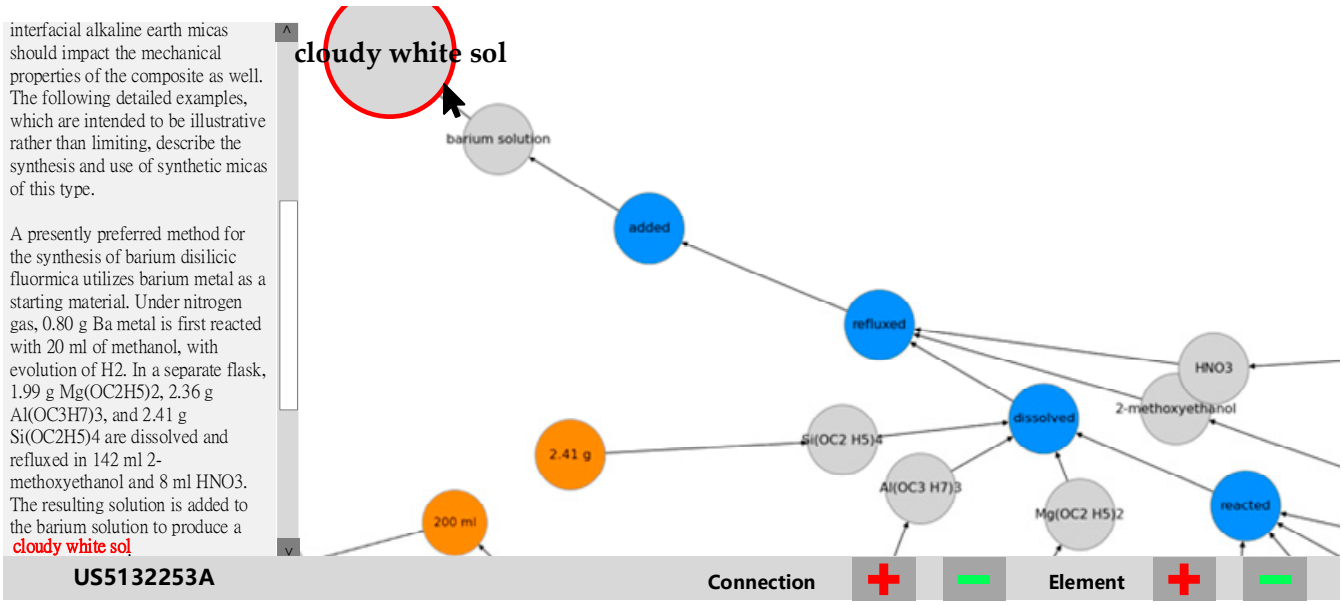

Figure 8. A schematic process graph sample created for chemical utility patent US5132253A.

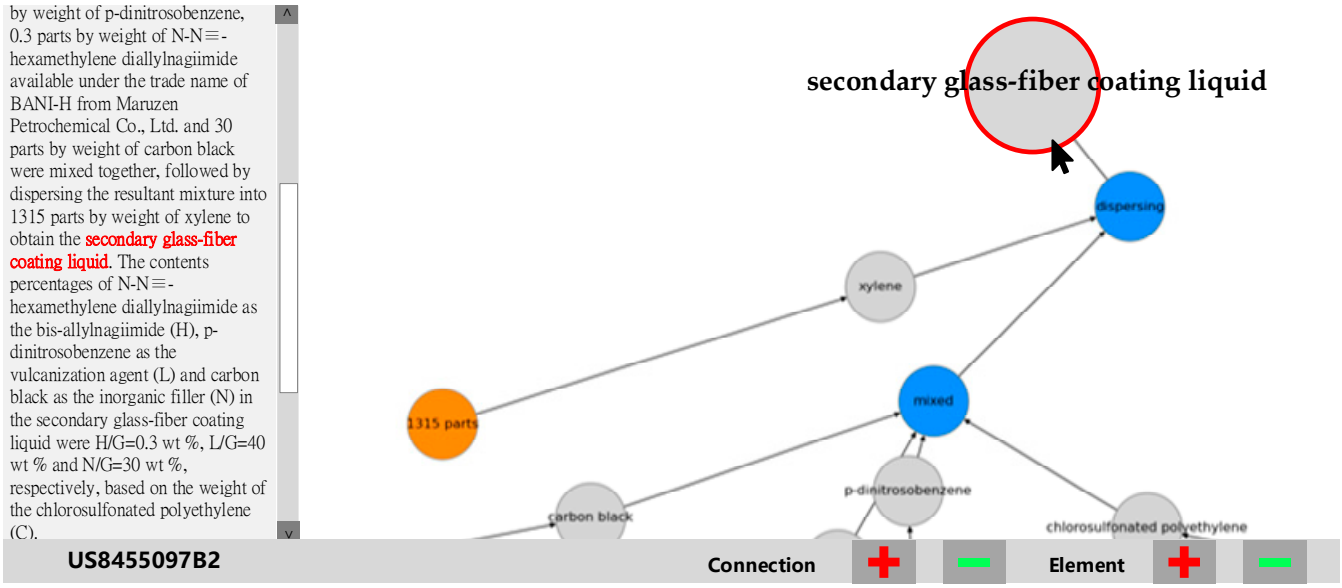

Figure 9. A schematic process graph sample created for chemical utility patent US8455097B2.

5. Conclusions

The computer-supported chemical patent knowledge graph generation is a valuable and promising research direction for visualizing chemical processes. The graph visualization solutions demonstrate the initial research for offering a complete visual solution to visualizing chemical process patents or other types of process patents. The results show a new field of process research and investigation, especially adopting and developing machine and deep learning approaches for technical document understanding, analysis, and synthesis. The proposed method uses the ALBERT model to pre-identify the chemical paragraphs for effective and accurate text mining of chemical text and sentences. The Network X model helps define the relationships between chemical terms and their chemical actions. The ALBERT model is trained using more than a thousand pre-labeled chemical patent documents. This research demonstrates the integration of three key modules for developing the computer-supported chemical patent IP knowledge graph generation system. For future research, a user interface is required for adding multiple functions, such as versatile visualization with drilled-down patent details, collaborative R&D chatrooms during patent reviews, multiple patent knowledge graph comparisons, similarity analysis, and even extrapolating information for new novel ideas. In addition, the functions of graph editing can be enhanced to extend from a patent knowledge graph visualization to a graph-based chemical innovation platform for collaborative R&D.

Author Contributions: Conceptualization, A.J.C.T. and C.V.T.; methodology, A.J.C.T. and C.-P.L.; software, C.-P.L. and H.-J.L.; validation, A.J.C.T. and C.V.T.; formal analysis, C.-P.L. and H.-J.L.; investigation, A.J.C.T. and C.-P.L.; resources, A.J.C.T. and C.V.T.; data curation, C.-P.L. and H.-J.L.; writing—original draft preparation, A.J.C.T. and C.-P.L.; writing—review and editing, C.V.T.; visualization, C.-P.L. and H.-J.L.; supervision, A.J.C.T.; project administration, A.J.C.T.; funding acquisition, A.J.C.T. and C.V.T. All authors have read and agreed to the published version of the manuscript.

Funding: The research is partially supported by the research grants of Ministry of Science and Technology, Taiwan (Grant numbers: MOST-108-2221-E-007-075-MY3 and MOST-108-2410-H-009-025-MY2) and National Tsing Hua University multi-disciplinary research grant.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. List of Python Packages Versions

Python and Packages	Version
Python	3.7.6
chemdataextractor	1.3.0
keras	2.4.3
matplotlib	3.3.4
networkx	2.5
numpy	1.19.2
pandas	1.2.4
tensorflow	1.13.1

Appendix B. Pseudo Code of the Proposed System

Algorithm A1. Pseudo code of the proposed system

```
for paragraph in paragraphList do
   convert each paragraph to vectors
   load ALBERT model    //predict the paragraph
   if paragraph ∈ chemical paragraph
      paragraph. label ('Yes')
   else
      paragraph. label('No')
for each paragraph and label in paragraphList do
   if paragraph. label = 'Yes'
      paragraph do words tokenization then
      PartOfSpeechList = paragraph do part-of-speech tagging
    for each word in PartOfSpeechList do    //label
      if word. label ∈ number
         if next word. label ∈ noun
            NodesList. append (combine word and next word)
            PropertyList. append('unit')
      else if word. label ∈ noun
         if next word. label ∈ (adjective or adverb)
               NodesList. append (combine word and next word)
            PropertyList. append('noun')
      else if word. label ∈ operation verb
               NodesList. append (word)
               PropertyList. append('action')
counter = 0
for counter in range (0, length (NodesList)) do
      if PropertyList[counter] = 'noun'
         if PropertyList[counter+1] = 'noun'
            NodesList. combine NodesList[counter] and NodesList[counter+1]
            pop PropertyList[counter+1]
         else
            PropertyList[counter] = 'substance'
```

References

1. Akhondi, S.A.; Klenner, A.G.; Tyrchan, C.; Manchala, A.K.; Boppana, K.; Lowe, D.; Muresan, S. Annotated chemical patent corpus: A gold standard for text mining. *PLoS ONE* **2014**, *9*, e107477. [CrossRef]
2. Zhang, T.; Sahinidis, N.V.; Rosé, C.P.; Amaran, S.; Shu, B. Forty years of Computers and Chemical Engineering: Analysis of the field via text mining techniques. *Comput. Chem. Eng.* **2019**, *129*, 106511. [CrossRef]
3. Schneider, N.; Fechner, N.; Landrum, G.A.; Stiefl, N. Chemical topic modeling: Exploring molecular data sets using a common text-mining approach. *J. Chem. Inf. Modeling* **2017**, *57*, 1816–1831. [CrossRef] [PubMed]
4. Hettne, K.M.; Williams, A.J.; van Mulligen, E.M.; Kleinjans, J.; Tkachenko, V.; Kors, J.A. Automatic vs. manual curation of a multi-source chemical dictionary: The impact on text mining. *J. Cheminform.* **2010**, *2*, 1–7. [CrossRef]
5. Himanen, L.; Geurts, A.; Foster, A.S.; Rinke, P. Data-driven materials science: Status, challenges, and perspectives. *Adv. Sci.* **2019**, *6*, 1900808. [CrossRef]
6. Huang, S.; Cole, J.M. A database of battery materials auto-generated using ChemDataExtractor. *Sci. Data* **2020**, *7*, 1–13. [CrossRef] [PubMed]
7. Ashaari, A.; Ahmad, T.; Awang, S.R.; Shukor, N.A. A Graph-Based Dynamic Modeling for Palm Oil Refining Process. *Processes* **2021**, *9*, 523. [CrossRef]

8. Kaur, M.; Kaur, H. Implementation of Enhanced Graph Layout Algorithm for Visualizing Social Network Data using NetworkX Library. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 287–292. [CrossRef]
9. Kim, E.; Huang, K.; Kononova, O.; Ceder, G.; Olivetti, E. Distilling a materials synthesis ontology. *Matter* **2019**, *5*, 8–12. [CrossRef]
10. Mehr, S.H.M.; Craven, M.; Leonov, A.I.; Keenan, G.; Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **2020**, *370*, 101–108. [CrossRef]
11. Vaucher, A.C.; Schwaller, P.; Geluykens, J.; Nair, V.H.; Iuliano, A.; Laino, T. Inferring experimental procedures from text-based representations of chemical reactions. *Nat. Commun.* **2021**, *12*, 1–11. [CrossRef] [PubMed]
12. Vaucher, A.C.; Zipoli, F.; Geluykens, J.; Nair, V.H.; Schwaller, P.; Laino, T. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **2020**, *11*, 1–11. [CrossRef]
13. Kononova, O.; Huo, H.; He, T.; Rong, Z.; Botari, T.; Sun, W.; Ceder, G. Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **2019**, *6*, 1–11. [CrossRef]
14. Shetty, P.; Ramprasad, R. Automated knowledge extraction from polymer literature using natural language processing. *Iscience* **2021**, *24*, 101922. [CrossRef]
15. George, J.; Hautier, G. Chemist versus Machine: Traditional Knowledge versus Machine Learning Techniques. *Trends Chem.* **2021**, *3*, 86–95. [CrossRef]
16. Johansson, S.; Thakkar, A.; Kogej, T.; Bjerrum, E.; Genheden, S.; Bastys, T. AI-assisted synthesis prediction. *Drug Discov. Today Technol.* **2019**, *32*, 65–72. [CrossRef]
17. Ai, Q.; Williams, D.M.; Danielson, M.; Spooner, L.G.; Engler, J.A.; Ding, Z.; Schrier, J. Predicting inorganic dimensionality in templated metal oxides. *J. Chem. Phys.* **2021**, *154*, 184708. [CrossRef] [PubMed]
18. Li, H.; Armiento, R.; Lambrix, P. An Ontology for the Materials Design Domain. In *International Semantic Web Conference*; Springer: Cham, Switzerland, November 2020; pp. 212–227. [CrossRef]
19. Hawizy, L.; Jessop, D.M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminform.* **2011**, *3*, 1–13. [CrossRef] [PubMed]
20. Jessop, D.M.; Adams, S.E.; Willighagen, E.L.; Hawizy, L.; Murray-Rust, P. OSCAR4: A flexible architecture for chemical text-mining. *J. Cheminform.* **2011**, *3*, 1–12. [CrossRef] [PubMed]
21. Ashino, T. Materials ontology: An infrastructure for exchanging materials information and knowledge. *Data Sci. J.* **2010**, *9*, 54–61. [CrossRef]
22. Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E.A.; Ceder, G. Opportunities and challenges of text mining in materials research. *Iscience* **2021**, *24*, 102155. [CrossRef]
23. Gao, X.; Tan, R.; Li, G. Research on text mining of material science based on natural language processing. In *IOP Conference Series: Materials Science and Engineering*; IOP Publishing: Bristol, England, March 2020; Volume 768, p. 072094.
24. Elton, D.C.; Turakhia, D.; Reddy, N.; Boukouvalas, Z.; Fuge, M.D.; Doherty, R.M.; Chung, P.W. Using natural language processing techniques to extract information on the properties and functionalities of energetic materials from large text corpora. *arXiv* **2019**, arXiv:1903.00415v1.
25. Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **2017**, *29*, 9436–9444. [CrossRef]
26. Swain, M.C.; Cole, J.M. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Cheminform.* **2016**, *56*, 1894–1904. [CrossRef] [PubMed]
27. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95–98. [CrossRef] [PubMed]
28. Tao, J.; Brayton, K.A.; Broschat, S.L. Automated Confirmation of Protein Annotation Using NLP and the UniProtKB Database. *Appl. Sci.* **2021**, *11*, 24. [CrossRef]
29. Campos, D.; Matos, S.; Oliveira, J.L. A document processing pipeline for annotating chemical entities in scientific documents. *J. Cheminform.* **2015**, *7*, 1–10. [CrossRef] [PubMed]
30. Akhondi, S.A.; Hettne, K.M.; Van Der Horst, E.; Van Mulligen, E.M.; Kors, J.A. Recognition of chemical entities: Combining dictionary-based and grammar-based approaches. *J. Cheminform.* **2015**, *7*, 1–11. [CrossRef] [PubMed]
31. Das, A.; Ganguly, D.; Garain, U. Named entity recognition with word embeddings and wikipedia categories for a low-resource language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP)* **2017**, *16*, 1–19. [CrossRef]
32. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
33. Gong, L.; He, D.; Li, Z.; Qin, T.; Wang, L.; Liu, T. Efficient training of bert by progressively stacking. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 2337–2346. Available online: http://proceedings.mLr.press/v97/gong19a.htmL (accessed on 10 May 2020).
34. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
35. Xiang, S.; Wang, L.; Yan, Z.C.; Qiao, H. A program for simplifying summation of Wigner 3j-symbols. *Comput. Phys. Commun.* **2021**, *264*, 107880. [CrossRef]
36. Hagberg, A.; Swart, P.; Chult, D.S. Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference; Los Alamos National Lab, Los Alamos, NM, USA, 21 August 2008.

37. Fruchterman, T.M.; Reingold, E.M. Graph drawing by force-directed placement. *Softw. Pract. Exp.* **1991**, *21*, 1129–1164. [CrossRef]
38. Kobourov, S.G. Spring Embedders and Force Directed Graph Drawing Algorithms. *arXiv* **2012**, arXiv:1201.3011v1.
39. Charles, E., Jr. *Polymer Chemistry*, 7th ed.; CRC Press Taylor & Francis Group: New York, NY, USA, 2008; pp. 1–18, ISBN 978-1-4200-5102-5.
40. WIPO. *World Intellectual Property Indicators 2020*; World Intellectual Property Organization: Geneva, Switzerland, 2020; pp. 40–45, ISBN 978-92-805-3201-2.
41. Dawes, S.B. Sol-Gel Method for Making Ceramic Materials. U.S. Patent 628,413, 21 July 1992.
42. Monden, T.; Hyakutake, H.; Ogaku, K. Coating Liquid for Covering Glass Fiber and Rubber-Reinforcing Glass Fiber Using Same. U.S. Patent 11/664,114, 4 June 2013.
43. O'Boyle, N.M. Towards a Universal SMILES representation-A standard method to generate canonical SMILES based on the InChI. *J. Cheminform.* **2012**, *4*, 1–14. [CrossRef] [PubMed]
44. Fukatani, T.; Hoshiba, K.; Fukuchi, I. Binder for Non-Aqueous Electrolyte Rechargeable Battery, Negative Electrode Slurry for Rechargeable Battery Including the Same, Negative Electrode for Rechargeable Battery Including the Same, and Rechargeable Battery Including the Same. U.S. Patent 16/857,809, 29 October 2020.