npj Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Automated pipeline for superalloy data by text mining

Weiren Wang[1,6], Xue Jiang[1,2,3,6], Shaohan Tian[1], Pei Liu[1], Depeng Dang[4], Yanjing Su[1]✉, Turab Lookman[5]✉ and Jianxin Xie[1]✉

Data provides a foundation for machine learning, which has accelerated data-driven materials design. The scientific literature contains a large amount of high-quality, reliable data, and automatically extracting data from the literature continues to be a challenge. We propose a natural language processing pipeline to capture both chemical composition and property data that allows analysis and prediction of superalloys. Within 3 h, 2531 records with both composition and property are extracted from 14,425 articles, covering γ′ solvus temperature, density, solidus, and liquidus temperatures. A data-driven model for γ′ solvus temperature is built to predict unexplored Co-based superalloys with high γ′ solvus temperatures within a relative error of 0.81%. We test the predictions via synthesis and characterization of three alloys. A web-based toolkit as an online open-source platform is provided and expected to serve as the basis for a general method to search for targeted materials using data extracted from the literature.

npj Computational Materials (2022)8:9 ; https://doi.org/10.1038/s41524-021-00687-2

INTRODUCTION

Artificial intelligence (AI)/machine learning (ML) is transforming materials research by changing the paradigm from "trial-and-error" to a data-driven methodology, thereby accelerating the discovery of new materials[1–16]. Well-characterized data remains a prerequisite for the success of AI/ML. Currently, there are two main sources of scientific data: (1) experimental and calculated results from a researcher's own laboratory and (2) data collected from papers published by other researchers. The scientific literature contains a vast amount of peer-reviewed, and largely high-quality, reliable data. Yet, manual data extraction with expert knowledge is time-consuming and labor-intensive for the tens of thousands of articles communicated using free-flowing natural language[17]. With an ever-increasing number of new publications, maintaining and updating a database manually becomes increasingly difficult for the individual researcher. Therefore, developing methods for automatically extracting data rapidly and accurately has increasingly become a necessity.

Recently, pipelines for automatic data extraction of organic and inorganic chemical substances from articles in the fields of chemistry and materials science have been introduced[18–22] using natural language processing (NLP) techniques. The named entity recognition (NER) and relation extraction tasks are considered critical components of data extraction from articles. The general methods of NER range from dictionary look-ups, rule-based, and machine-learned approaches. Cases that cannot be handled by dictionaries or rules are investigated using ML approaches which require substantial expert-annotated data for training along with detailed annotation guidelines[23]. Kim et al. used neural network- and parse-based methods to recognize and extract synthesis parameters with an F1 score of 81% from over 640000 journal articles[24]. "ChemDataExtractor" has been developed to recognize chemically named entities to extract relations of organic and inorganic compounds from a massive article corpus (hundreds of thousands) using a dictionary with ML and multiple grammar rules[18]. Court et al. used "ChemDataExtractor" with a modified "Snowball" algorithm to extract Curie and Néel temperatures for magnetic materials with an estimated overall precision of 73% from a corpus of 68078 articles[22]. Although this database is for magnetic materials, ferroelectrics and antiferroelectrics also share the Curie and Néel terminology for the transition temperatures. The terms are often not necessarily used in the same manner as the corresponding magnetic systems; hence, the database also includes those materials which are not "magnetic".

Superalloys are widely used in turbine blades and vanes of the most advanced aero engines and industrial gas turbines. Knowledge of their properties, including those associated with transition temperatures in the multicomponent alloy phase diagrams together with their chemical composition and synthesis conditions, are required information for alloy design. Moreover, there are ~20,000 articles on superalloys; hence, to accelerate data-driven superalloy design[25–30], extraction and assimilation of existing data from the literature is crucial. The direct application of supervised deep learning methods for NER or relation extraction requires adequate and effective large hand-labeled datasets for training. Even certain semi-supervised methods, such as "Snowball", require a given number of labeled samples as seeds to start learning, and this presents difficulty in achieving high precision and recall simultaneously[31].

In this paper, we propose an automated NLP pipeline to capture both chemical composition and property data of a superalloy into a single dataset, which subsequently allows us to perform a global analysis on superalloys using the data extracted from 14425 journal articles from the literature. In particular, a rule-based NER method and a heuristic text multiple-relation extraction distance-based algorithm, which requires no labeled samples, are developed for a small corpus. In addition, a common table parse and relation extraction algorithm is also developed catering to table processing needs. The F1 score of NER for alloy named entity reaches 92.07%, much higher than the 42.91% and 24.86% achieved using the bidirectional long short-term memory (BiLSTM) network with a conditional random field (CRF) layer (BiLSTM-CRF) model and "ChemDataExtractor" tool, respectively. The F1 score of text relation extraction for the γ′ solvus temperature was 79.37%,

[1]Beijing Advanced Innovation Center for Materials Genome Engineering, Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China. [2]Collaborative Innovation Center of Steel Technology, University of Science and Technology Beijing, Beijing 100083, China. [3]Beijing Key Laboratory of Materials Genome Engineering, University of Science and Technology Beijing, Beijing 100083, China. [4]School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China. [5]AiMaterials Research LLC, Santa Fe, NM 87501, USA. [6]These authors contributed equally: Weiren Wang, Xue Jiang. ✉email: yjsu@ustb.edu.cn; turablookman@gmail.com; jxxie@ustb.edu.cn

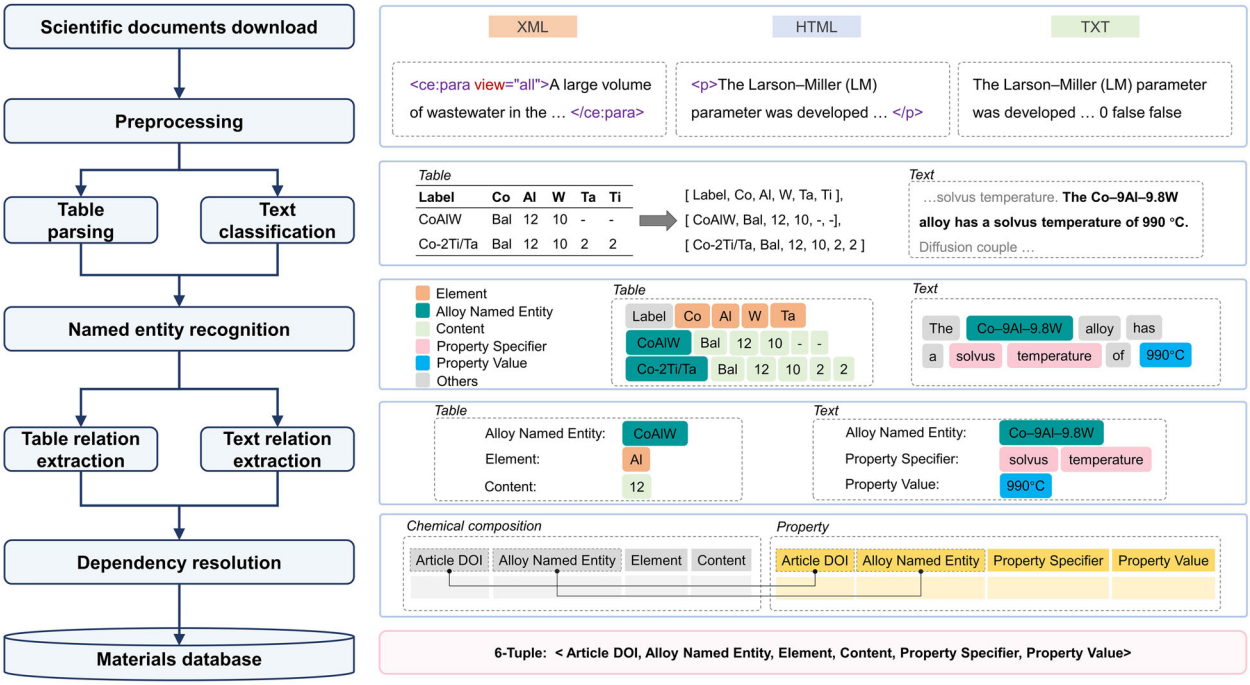


Fig. 1 Schematic workflow of the automated text mining pipeline. The workflow involves several stages of scientific documents download, preprocessing, table parsing, text classification, named entity recognition, table and text relation extraction, and interdependency resolution. A corpus of scientific articles is scraped and the irrelevant information in raw corpus is then filtered during preprocessing. According to the table parsing and text classification, the tables and sentences with target information are determined for named entity recognition and relation extraction. The alloy named entity, property specifier and property value are recognized by named entity recognition, and relation extraction of text and table gives the specific tuple relations. Interdependency resolution resolves the linkage to chemical composition and property data fragments for one specific material, and finally outputs a complete record into materials database.

higher than the 33.21% and 43.28% obtained by the well-known "Snowball" and modified "Snowball" semi-supervised algorithms, respectively. Our distance-based algorithm, without the need for labeled samples to handle multiple-relation extraction, thus performs better under small corpus conditions. The method leads to a higher recall than "Snowball", which does not fare well as the seed tuples used for starting the learning are too few, making the sentence clustering process in "Snowball" ineffective to cover all sentence forms, and only a few tuple relations can be extracted with low recall. The table parsing and relation extraction tool performed well with an F1 score of 95.23%. In total, a dataset with 2531 instances covering chemical compositions and physical properties, such as the γ' solvus temperature, density, solidus temperature, and liquidus temperature, were automatically extracted from a corpus of 14,425 articles from Elsevier and other publishers.

We interrogate the database to distill trends, which are consistent with the known behavior of superalloys. Our database does not incorporate synthesis or processing conditions and other experimental aspects, including measurement uncertainties, all important in superalloy development. Therefore, to gauge how predictive the extracted data is, we built a data-driven ML model to predict and compare with the γ' solvus temperatures of 15 superalloys not part of our extracted data as they were reported subsequently in 2020 and 2021. The predictions are within a relative error of 2.27%. The model was further used to predict the three unexplored Co-based superalloys Co-36Ni-12Al-2Ti-1W-4Ta-4Cr, Co-36Ni-12Al-2Ti-1W-4Ta-6Cr and Co-12Al-4.5Ta-35Ni-2Ti with γ' solvus temperature >1250 °C. By synthesizing and characterizing the alloys, we show that the temperatures are in agreement within a mean relative error of 0.81%. Hence, our ML studies show the potential of the pipeline, and the accuracy of the extracted database by text mining provides a valuable resource for superalloy development.

All the source code used in this work is available at https://github.com/MGEdata/SuperalloyDigger. Furthermore, a web-based toolkit has been developed; further examples of how to use and adapt the toolkit can be found at http://SuperalloyDigger.mgedata.cn. This extraction strategy and source code can be used for other alloy materials by re-designing regular expressions. It presents a practical and effective means of data extraction from articles to accelerate the development of data-driven materials design.

RESULTS

Extraction strategy

Our automated text mining pipeline for superalloys involves several stages of scientific documents download, preprocessing, table parsing, text classification, named entity recognition, table and text relation extraction, and interdependency resolution, which are schematically shown in Fig. 1. Starting with a corpus of scientific articles scraped in extensible markup language (XML), hypertext markup language (HTML) or plain text format, we preprocess the raw archived corpus to produce a complete document record and filter out irrelevant information (see Retrieval of articles and preprocessing in "Methods"). The idea underlying text classification is to determine which sentence contains the target property information to be extracted (see Text classification in "Methods"). Table parsing transforms the complete table caption and body into a structural format and then classifies which table contains the chemical composition and target property information to be extracted (see Table parsing in "Methods"). NER methods are designed to recognize the alloy named entity, property specifier, and property value from the English-language text and table, and these are followed by relation extraction. Relation extraction of text and table gives the specific tuple relations for the element content and property, and interdependency resolution

resolves the linkage to chemical composition and property data fragments for one specific material. Finally, the extracted tuple entities containing the article digital object identifier (DOI), alloy named entity, chemical element, content, property specifier, and property value are automatically compiled into a highly structured format to form a materials database.

Named entity recognition

The problem of chemical composition and property extraction from the superalloy literature can be summarized as a 6-tuple relation extraction, where the 6-tuple consists of article DOI, alloy named entity, chemical element, content, property specifier, and property value. The alloy named entity is usually described in the form of elemental composition (e.g., Co-9Al-9.8W and 8Al1W2Mo), superalloy designation (e.g., ERBOCo-0 and U720Li), or a pronoun (e.g., this alloy). The chemical element can be identified according to the periodic table, and its composition is expressed as a numeric value with units in the form at.% or wt.%. Property specifier refers to the target property name, such as γ' solvus temperature or density. Property value gives the value and unit of each property. NER technology to recognize alloy named entity from English-language text and table is essential for subsequent relation extraction[32]. In this work, based on the automatically archived ~14,000 superalloy documents, a rule-based method for NER was explored, following text and table classification. Here, we take γ' solvus temperature as an example for the property specifier to illustrate the NER procedure for superalloys (Fig. 2), which provides entity sequences for the subsequent relation extraction.

Multiple specialized grammar rules are tailored to recognize specific types of superalloy information. Table 1 summarizes nine patterns of common writing forms for superalloys. Among them, three patterns for nickel-based superalloys, such as IN738LC, Udimet 720Li, and Hastelloy C276, and six patterns that adapt to all superalloys, such as ERBOCo-0, alloy 718, CoWAlloy2, and 8Al1W2Mo. Different patterns correspond to different grammar rules of natural language, and we tailored nine types of regular expressions in Python to define the grammar patterns. For example, "8Al1W2Mo" and "2Nb2Re" are composed of the elements and corresponding percentages that can be recognized by the pattern "^[0–9]+\.?[0–9]{0, 2}[A-JL-Z]". "FGH98" is the designation composed by capital letters and numbers in the pattern of "^[A-Z]\S + [0–9]$". If a word or word chunk belonging to an alloy named entity or property value is successfully

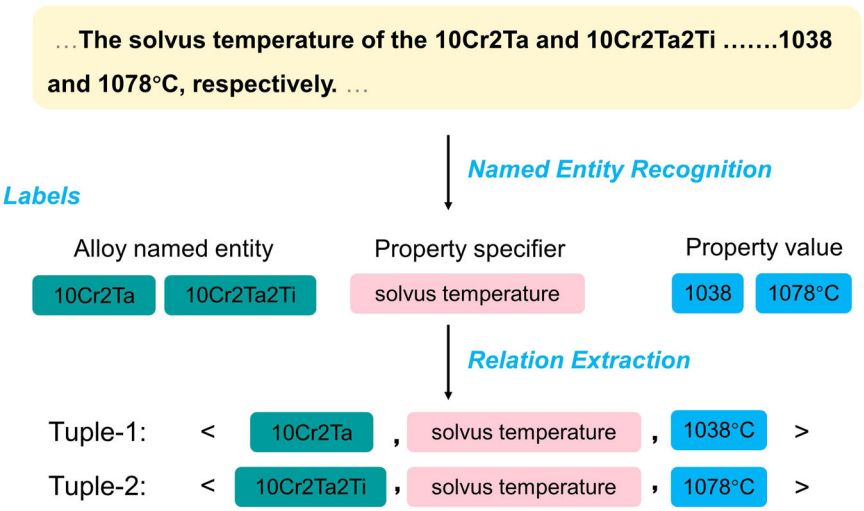


Fig. 2 Named entity recognition technology to provide entity sequence for subsequent relation extraction. The alloy named entity, property specifier, and property value are recognized as entity sequence from English-language text and table.

Table 1. In-domain word patterns, examples, and rules for writing alloy named entities and property value and unit.

Category	Pattern	Examples	Rules for writing (Regular expression)
Alloy named entities	1	9W-0.08B, Co-9Al-7.5W	[0–9]{0, 2}(\.[0–9]{1, 2})?[A-Z][a-z]?\-[0–9]{0, 2}(\.[0–9]{1, 2})?[A-Z][a-z]?
	2	SLM-1170 °C, TMS-138A, ERBOCo-0	^[A-Z]+[a-z]*\-[0–9]\w*
	3	alloy 718, 282 alloy, 9.5W alloy	\s([0–9A-Z]+\w*)s+\S*alloys?\s+\S*alloys?\s([0–9A-Z]+\w*)\s
	4	CoWAlloy2, RRHT3, IN718	^[A-Z]\S+[0–9]$
	5	8Al1W2Mo, 8W, 718Plus	^[0–9]+\.?[0–9]{0, 2}[A-JL-Z]
	6	Hayness 188, IN738LC, Incoloy 800HT	[A-Z]+[a-z]*\s+[A-Z]*[0–9]{2,}[A-Z]*
	7	IN738LC, U720Li	^[A-Z]+[0–9]+[A-z]+
	8	Udimet 720Li, Incoloy 25-6Mo	[A-Z]+[a-z]*\s+\d+\-?\d*[A-Za-z]+
	9	Hastelloy C276, Inconel X-750	[A-Z]+[a-z]*\s+[A-Z]+\-?[0–9]+
Property value entities (with unit)	1	1050 °C, 850–950 °C, >950 °C	^\W{0, 1}[7–9][0–9]{2}(\.[0–9]{1, 2})?\S*°C$
			^\W{0, 1}1[0–9]{3}\S*°C$
	2	1050 K, 850–1180 K, >1620K	^\W{0, 1}[7–9][0–9]{2}(\.[0–9]{1, 2})?\S*K$
			^\W{0, 1}1[0–9]{3}\S*K$

Table 2. Precision, recall, and F1 score of the NER.

Category		Precision	Recall	F1 score	Test set
Alloy named entity	This work	90.58%	93.60%	92.07%	545 sentences
	BiLSTM-CRF	51.99%	36.53%	42.91%	545 sentences
	ChemDataExtractor	36.52%	18.84%	24.86%	545 sentences
Property value (with unit)		85.71%	99.25%	91.98%	845 sentences

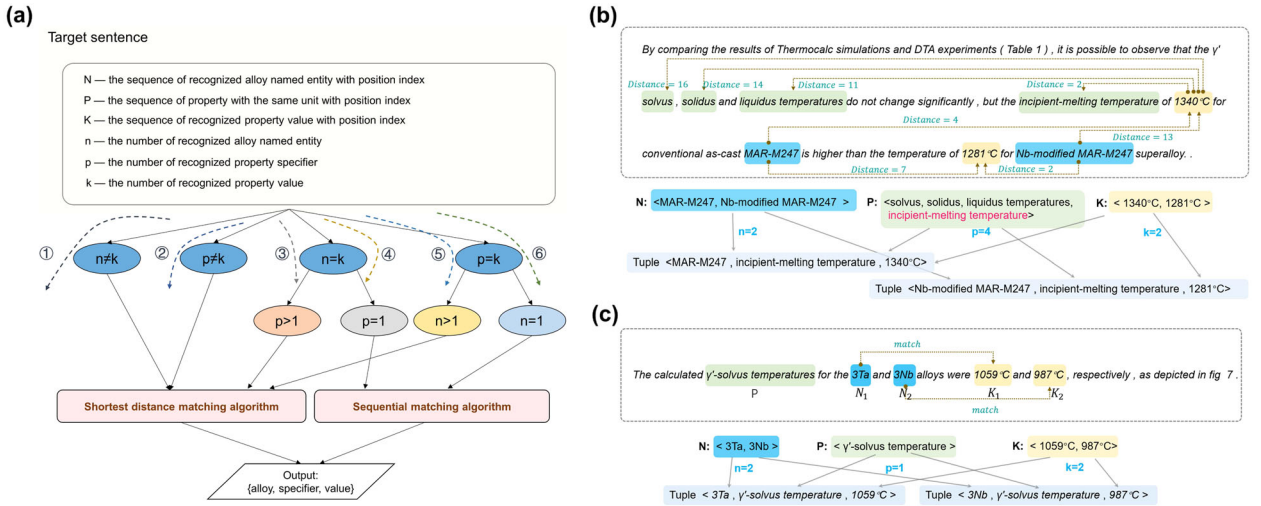


Fig. 3 The algorithm of text relation extraction. a Algorithm flow chart of relation extraction by shortest distance matching and sequential matching. The number of words and word sequences between entities can act as the main syntactic features in feature-based relationship extraction methods. **b** Schematic of shortest distance matching algorithm. **c** Schematic of sequential matching algorithm. The phrases in blue, green, and yellow rectangle are recognized as alloy named entity, property specifier, and property value by NER, respectively.

recognized, the word or word chunk can be regarded as positive. The rules were run on 545 sentences (~19,000 words) from 283 articles, and the obtained precision, recall, and F1 score are listed in Table 2. The recognition procedure for property values during text classification is illustrated by the rule in Table 1, and its precision, recall, and F1 score on 845 sentences are also shown in Table 2.

We also used the BiLSTM-CRF model for NER tasks[33] (see BiLSTM-CRF model in "Methods"). Moreover, the NER tool in "ChemDataExtractor" was also used to perform superalloy NER. Compared with the BiLSTM-CRF model and "ChemDataExtractor", our proposed rule-based method performs better for alloy name entity (Table 2). As for the BiLSTM-CRF model, its vast space of model parameters causes over-fitting for model training on the small labeled corpus. "ChemDataExtractor" uses CRF-based, rule-based and dictionary-based methods to recognize chemical substances. As the rules and dictionaries are different, it does not perform well for superalloys.

Text relation extraction

Relation extraction identifies and resolves ambiguities in semantic relationships between two entities in unstructured text data[34]. For property extraction from superalloy articles, the relation can be treated as a quaternary-tuple <article DOI, alloy named entity, property specifier, property value > . Article DOI can be archived during the retrieval of articles and preprocessing. The most challenging task for superalloy property extraction is multiple-relation extraction from a single sentence[35]. In particular, it is common for several alloy named entities (≥1) to be reported with their corresponding property values for a specified property in one sentence, or a specified alloy named entity may be reported with several properties (≥1) and corresponding values (≥1). This results in several obstacles to relation extraction based on the limited superalloy corpus. A supervised relation extraction algorithm requires a large number of labeled samples above ~70,000[36], and even semi-supervised methods require a certain number of labeled samples as seeds to start learning. Here, we propose a distance-based algorithm without the requirement for labeled samples to handle multiple-relation extraction; the workflow of the relation extraction is shown in Fig. 3a. In feature-based relationship extraction methods, the number of words and word sequences between entities can act as the main syntactic features[37]. Therefore, the number of entities and distance between entities provide a basis for evaluating the relationship dependence.

After NER, target sentences are organized by the form of the entity sequences with position index for alloy named entity, property, and value. The shortest distance matching algorithm is applied where (i) the number of alloy named entities is not equal to that of property values ($n \neq k$ as flow 1 in Fig. 3a) and (ii) number of property specifiers is not equal to that of property values ($p \neq k$ as flow 2 in Fig. 3a), $n = k$, $p > 1$ as flow 3, and $p = k$, $n > 1$ as flow 5. The shortest distance matching algorithm adopts a greedy strategy, and the pseudocode is given in Supplementary Fig. 1. Taking the conditional branch $n \leq k$ in Supplementary Fig. 1 as an example, for each alloy named entity, N_i, the distance between each property value entity K_j and N_i is calculated from Eq. (1) to find the closest property value entity K_m to the current N_i.

$$\mathrm{Distance}(x_1, x_2) = |\mathrm{pi}(x_1) - \mathrm{pi}(x_2)| \quad (1)$$

where x_1 and x_2 are two entities and $pi(x)$ is the position index of entity x in the target sentence. K_m is treated as the anchor to search the closest property entity P_m among all property entities.

Table 3. Precision, recall, and F1 score of the "Snowball", modified "Snowball" algorithm and method employed in this work.

Category		Precision	Recall	F1 score	Validated on
Snowball	Seeds = 50	76.78%	20.47%	32.33%	329 sentences
	Seeds = 100	73.77%	21.42%	33.21%	329 sentences
Modified Snowball	Seed = 50	60.71%	18.78%	28.69%	329 sentences
	Seed = 100	66.67%	32.04%	43.28%	329 sentences
This work		75.86%	83.22%	79.37%	329 sentences

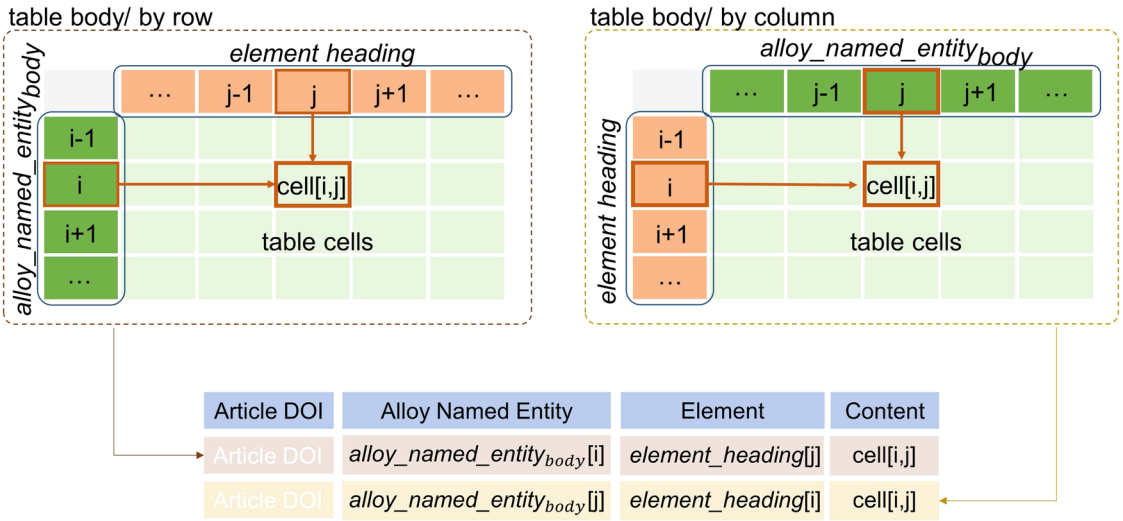


Fig. 4 Schematic of table relation extraction algorithm. Table direction ("by row" or "by column") is first detected by estimating the row or column location of the target information in the table body, e.g., chemical elements. Alloy named entities in green cells and elements in orange cells are jointed by the row and column index of each table cell from corresponding sequences respectively, and finally written into a quaternary-tuple <article DOI, alloy named entity, property specifier(element), property value(content)>.

Therefore, a set of triple-tuples $<N_i, P_m, K_m>$ is successfully extracted and added to the relation set. A schematic diagram of the shortest distance matching algorithm is shown in Fig. 3b; the phrases in blue, green, and yellow rectangle are recognized as alloy named entity, property specifier entity, and property value, respectively, during the former NER process.

For the situation wherein $n = k$ and where there is only one property entity in a sentence ($p = 1$ as flow 4), the sequential matching algorithm is performed to match the alloy named entity and property value for the specified property in order; the same is true for $p = k$, $n = 1$. Supplementary Fig. 2 shows the pseudocode of the sequential matching algorithm, and a schematic diagram of it is shown in Fig. 3c.

If the relation amongst alloy named entity, property specifier, and property value in a sentence is correctly captured, the extracted quaternary-tuple <article DOI, alloy named entity, property specifier, property value> is regarded as a positive sample. The above relation extraction algorithm was applied to the 458 target sentences classified from ~14,000 articles, and 680 γ' solvus temperature instances in total were extracted automatically. After manual inspection on randomly selected 329 sentences, the precision, recall, and F1 score of the relation extraction algorithm for γ' solvus temperature were 75.86%, 83.22%, and 79.37%, respectively.

We also used the original "Snowball" algorithm[31] and modified "Snowball" algorithm[22] with seeds of 50 and 100 to extract property-tuple relations (see Snowball algorithm in "Methods"). Our method presented a higher recall and F1 score than the "Snowball" algorithm, as shown in Table 3. The recall of "Snowball" was worse than our method because the seed tuples used for starting learning were too few, so each cluster of sentence forms contained fewer training tuples. This made the sentence clustering process in "Snowball" ineffective as it could not cover all sentence forms, and only a few tuple relations could be extracted with very low recall. Therefore, our distance-based algorithm without the need for labeled samples to handle multiple-relation extraction, performed better under such small corpus conditions.

Table relation extraction

Tables are attractive targets for information extraction due to their high data density and semi-structured nature. Compared to completely unstructured natural language, tables under XML and HTML format are more interpretable. Table parsing transforms complete table information, including table caption and body, into the structural format of a nested table cell list, and then classifies which table contains the chemical composition and target property information to be extracted. After table parsing, 5327 composition tables and 114 tables with solvus temperature were obtained. Table relation extraction gives the specific tuple relations for concrete element content and property. Taking composition extraction as an example, during table relation extraction, table direction ("by row" or "by column") is first detected by estimating the row or column location of the chemical elements in the table body. The table caption is then checked to see whether there exists an alloy named entity. Figure 4 depicts the schematic diagram under the above scenario. Taking the case "by row" as an example, if the recognized alloy named entities are more than one or there is none specified in the table caption, the alloy named entities and elements are addressed by the row and column index of each table cell from *alloy_named_entity*$_{body}$ and *element_heading* sequences respectively. If there

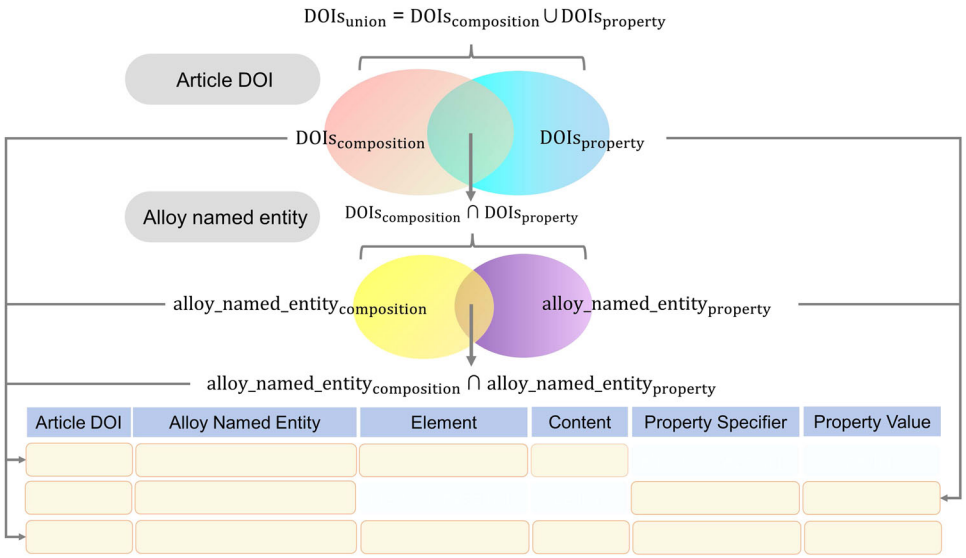


Fig. 5 Schematic of interdependency resolution algorithm by divide and conquer strategy. To merge the fragments of chemical composition and property, the instances of the same alloy named entity with the same article DOI are joined into a complete record in a 6-Tuple: <article DOI, alloy named entity, element, content, property_specifier, property_value>.

is only one alloy named entity in table caption, elements are addressed by the column index of each table cell from the *element_heading* sequence. Table relation extraction finally outputs composition tuples. The pseudocode of the table relation extraction algorithm can be found in Supplementary Fig. 3. The process of property extraction from the table is the same as composition, except that the objects from NER from table headings are changed.

If the relation amongst alloy named entity, property specifier (element) and property value(content) in table cells is correctly captured, the extracted quaternary-tuple <article DOI, alloy named entity, property specifier(element), property value(content)> is regarded as a positive sample. The above relation extraction algorithm was applied to 5441 tables on chemical composition (5327 tables) and property γ′ solvus temperature (114 tables) from ~14,000 articles after table parsing, and in total 12703 composition-tuple relation instances and 579 property-tuple relation instances from tables were extracted automatically. After manual inspection of 45 articles by random resampling, the precision, recall, and F1 score were 90.89%, 100%, and 95.23%, respectively.

Data interdependency resolution

In many cases of alloy structural materials, property extraction of superalloys departing from specific chemical compositions is not permissible. Hence, interdependency resolution aims to resolve the linkage to chemical composition and property data fragments for a specific material. After text and table relation extraction, chemical composition tuples of <article_DOI, alloy_named_entity, element, content> are obtained from tables and property tuples of <article_DOI, alloy_named_entity, property_specifier, property_value> are from text and tables. Of the ~14000 articles, we automatically extracted in total 12703 chemical composition instances and 1259 γ′ solvus temperature instances (680 instances from text and 579 from tables). To merge these data fragments, the chemical composition and property instances of the same alloy named entity with the same article DOI are joined into a complete record in a 6-Tuple: <article DOI, alloy named entity, element, content, property_specifier, property_value>. A divide and conquer strategy is adopted during composition and property data fragments linkage by our algorithm, as shown in Fig. 5. The details of the algorithm are depicted in Supplementary Fig. 4.

For $Tuple_{composition}$ and $Tuple_{property}$ from text and table relation extraction, all the tuples are divided into three sets ($DOIs_{intersection}$, $DOIs_{composition}$–$DOIs_{intersection}$, and $DOIs_{property}$–$DOIs_{intersection}$) according to whether they possess the same article DOIs. For the tuples in the set $DOIs_{composition}$–$DOIs_{intersection}$, no corresponding property information is found. For the tuples in the set $DOIs_{property}$–$DOIs_{intersection}$, no chemical composition information is extracted. For the tuples in $DOIs_{intersection}$, chemical composition and property information are both extracted from one article at the same time. These tuples, under the same article DOI, will continue to be divided into three sets according to whether they have the same alloy named entity. Finally, tuples in set $alloy\_named\_entity_{intersection}$ are joined into the complete record with both chemical composition and property for one alloy named entity of each article DOI, while tuples in $alloy\_named\_entity_{composition}$–$alloy\_named\_entity_{intersection}$ or $alloy\_named\_entity_{property}$–$alloy\_named\_entity_{intersection}$ can only obtain chemical composition or property information for one alloy named entity. In total, for γ′ solvus temperature, we obtained 743 complete records from 12703 chemical composition instances and 1259 property instances from ~14000 articles.

It is worth mentioning that during the course of concatenating on the basis of the same alloy named entity, the pronoun (e.g., this alloy) and abbreviation (e.g., Cr-5) as alloy named entities in the property extraction result cause difficulty in matching the chemical composition. This is because the alloy named entity usually appears in the form of the full name in the extraction result of the chemical composition table. We, therefore, looked for the full name corresponding to the abbreviation in the previous context under the pattern "full name(abbreviation)". There are 743 complete records with concrete chemical composition amongst 1259 property instances, and the difference is mainly associated with pronouns acting as alloy named entities in the property extraction result. Certain chemical compositions are expressed in text rather than in the table so that they can't be extracted.

DISCUSSION

The methodology and pipeline presented above demonstrate the ability to extract chemical composition and properties, such as γ′ solvus temperature, accurately from the superalloy scientific literature, even for a limited size corpus. Furthermore, we applied

our automated text mining pipeline to other physical properties of superalloys, including density, solidus, and liquidus temperatures, by regenerating the synonym dictionary of the property specifier based on the pre-trained word embedding model and adjusting the writing rules for the value and unit. In total, we obtain 5136 property instances of γ′ solvus temperature (1259), density (2296), solidus temperature (793), liquidus temperature (788), and 12703 chemical compositions from 14425 superalloy articles. The precision, recall, and F1 scores are given in Supplementary Table 1. The average precision, recall, and F1 score for γ′ solvus temperature, density, solidus, and liquidus temperatures are 83.67%, 93.08%, and 88.13%, respectively. Among these, 2531 instances successfully matched with their chemical compositions, and the precision, recall, and F1 score were validated on 30 randomly sampled articles for chemical composition and each property. We have obtained similar performance by applying the pipeline to extract hardness information for high entropy alloys (https://github.com/MGEdata/Superalloydigger_HEAs_use_case).

For a relatively small corpus such as superalloys, we have presented a rule-based NER method and an effective distance-based heuristic multiple-relation extraction algorithm for the pipeline to overcome the drawback of limited training corpus labels. We achieved an F1 score of 92.07% for alloy named entity, and an average F1 score of 77.92% for relation extraction for γ′ solvus temperature, density, solidus and liquidus temperatures. Our pipeline does not require any labeled corpus to achieve high precision and recall, avoiding the over-fitting problem of supervised and semi-supervised learning with low recall caused by insufficient labeled corpus. Moreover, our common table processing tool with table parsing and relation extraction algorithm performed well with an F1 score of 95.23%. Therefore, the methodology presented here is expected to perform well for subject-oriented information extraction, even for small corpus as lack of adequate labeled data often creates problems in the use of supervised or semi-supervised learning methods.

From the perspective of superalloy development, Fig. 6a shows the γ′ solvus temperature trends of cobalt-based and nickel-based superalloys arranged by years. The reported highest γ′ solvus temperature of a nickel-based superalloy is 1308 °C in the year 2012 by Pang[38], while that for the highest cobalt-based superalloy is 1269 °C in the year 2017 by Lass EA[39], which was referenced by Li in the year 2019[40]. An Ashby chart showing the superalloys as a function of γ′ solvus temperature and density is plotted in Fig. 6b. The superalloys in blue circles with high Ni content and Ta content have a higher γ′ solvus temperature than other superalloys, whereas superalloys in orange and pink circles without Ni content exhibit relatively low γ′ solvus temperatures. This is consistent with the reported behavior[41,42]. Fig. 6c shows that the addition of W in the ternary superalloy Co-9Al-xW ternary promotes an increase in the γ′ solvus temperature. This is because W tends to accumulate in the γ′ phase and occupy B-sites of the A_3B ordered phase. Also, we see from the error bars the variation in the measured values of γ′ solvus temperature for the same superalloy in different articles. Thus, for Co-9Al-10W there is a distribution of values ranging from 980 to 1060 °C in three articles. In Fig. 6d, for the Co-Ni-Al-Mo-based superalloy, the γ′ solvus temperature and density are significantly increased by the addition of Ta compared with Nb, which is consistent with the results reported by Lass EA[42]. After adding Ti, the value of γ′ solvus temperature is further increased because Ti is a forming element of γ′ phase, and its promotion effect on the γ′ solvus temperature is greater than that of Co. Therefore, the data we have extracted support the known behavior of superalloys.

As latent knowledge regarding future discoveries can lie in past publications, we next examined the value of the extracted data to provide actionable insights for materials discovery. We, therefore, constructed a data-driven model from the extracted 743 records with chemical composition and γ′ solvus temperature. Of the 743 records, we focused on 259 cobalt-based and 73 nickel-based compounds after data screening and cleaning by removing duplicates and errors (see Data preprocessing for machine learning in "Methods"). Figure 7a presents the distribution of γ′ solvus temperature for the Co-9Al-9W, Mar-M247, U720Li, IN738LC, CMSX-4 and CMSX-10 superalloys with γ′ solvus temperature of 993 ± 9 °C, 1206 ± 28 °C, 1160 ± 20 °C, 1168 ± 31 °C, 1286 ± 30 °C, 1343 ± 9 °C respectively, with mean and standard deviation. The entire composition space consists of Co, Al, W, Ni, Ti, Cr, Ta, B, Mo, Re, Nb, Si, V, Fe, Hf, Ru, Ir, Cu, Pt and C, and we built a prediction model via support vector regression with a radial basis function kernel for γ′ solvus temperature of superalloys (see Prediction model for γ′ solvus temperature in Methods). The model selection and evaluation process are shown in Fig. 7b and Fig. 7c. The model was used to predict γ′ solvus temperatures of the latest reported 15 superalloys from 12 different published articles in the years 2020 and 2021, which are not in the dataset extracted by our pipeline (Supplementary Table 2). The mean relative error between the reported and predicted γ′ solvus temperatures by the SVR.rbf model is 2.27%. Figure 7d shows the reported and predicted γ′ solvus temperatures of the 15 superalloys reported, and the relative error of the temperatures in orange box is <1%. Furthermore, the trained SVR.rbf model was used to design Co-based superalloys with the target of high γ′ solvus temperatures (>1250 °C). We considered $Co_{1-a-b-c-d-e-f}Al_aW_bNi_cTi_dTa_eCr_f$ alloys with compositions a, b, c, d, e and f, where each element varies in steps of 0.5% with constraints 11%≤a ≤ 12%, 0%≤b ≤ 1%, 35%≤c ≤ 37%, 1%≤d ≤ 2%, 4%≤e ≤ 5% and 0%≤f ≤ 6%. Three alloys Co-36Ni-12Al-2Ti-1W-4Ta-4Cr, Co-36Ni-12Al-2Ti-1W-4Ta-6Cr and Co-12Al-4.5Ta-35Ni-2Ti with predicted γ′ solvus temperatures >1250 °C, not reported previously, were selected out of 15,795 possibilities for experimental synthesis. These were considered to precipitate γ′ phase from expert knowledge. The measured γ′ solvus temperatures are 1251 °C, 1239.3 °C, and 1263 °C respectively, determined by differential scanning calorimetry (DSC) (see Synthesis and characterization in "Methods" section). The microstructures and DSC heating curves for Co-36Ni-12Al-2Ti-1W-4Ta-4Cr, Co-36Ni-12Al-2Ti-1W-4Ta-6Cr and Co-12Al-4.5Ta-35Ni-2Ti are shown in Fig. 7e–g respectively. The relative errors between experimental values and predicted values are 0.56%, 1.41%, and 0.48% respectively and details are given in Supplementary Table 3.

Finally, we discuss aspects that we have not incorporated and where further progress is necessary. Among the 2531 records the pipeline automatically extracted, errors and duplicates are inevitable and not easy to automatically eliminate. The use of records still requires manual intervention for cleaning. Also, the pipeline proposed does not accurately capture property values described as a range, such as "between…and…". Additionally, the scenario in which complete property-tuple information is distributed across two or more separate sentences needs to be resolved. We have not incorporated synthesis or processing conditions or other experimental parameters, as well as measurement uncertainties. These aspects are important for alloy development and need to be augmented to enrich the existing database. As the database continues to grow with more properties, experimental parameters, and compositions, the models will tend to be more predictive. As the amount of scientific literature grows, NLP provides a mean to make the vast scientific information accessible to enable a new paradigm of machine-assisted discovery.

In summary, we have proposed an automated data extraction pipeline for superalloys to generate a structural database by NLP, including scientific document download, preprocessing, table parsing and text classification, NER, relation extraction of text and table respectively, and interdependency resolution automatically. The extracted entities with a total of 2531 instances covering the physical properties of γ′ solvus temperature, density, solidus

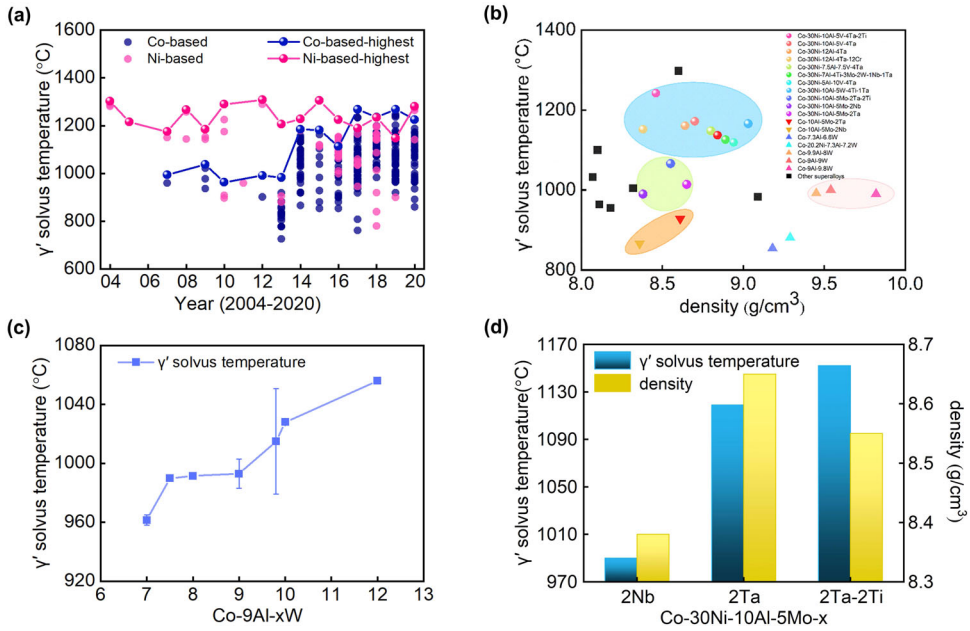
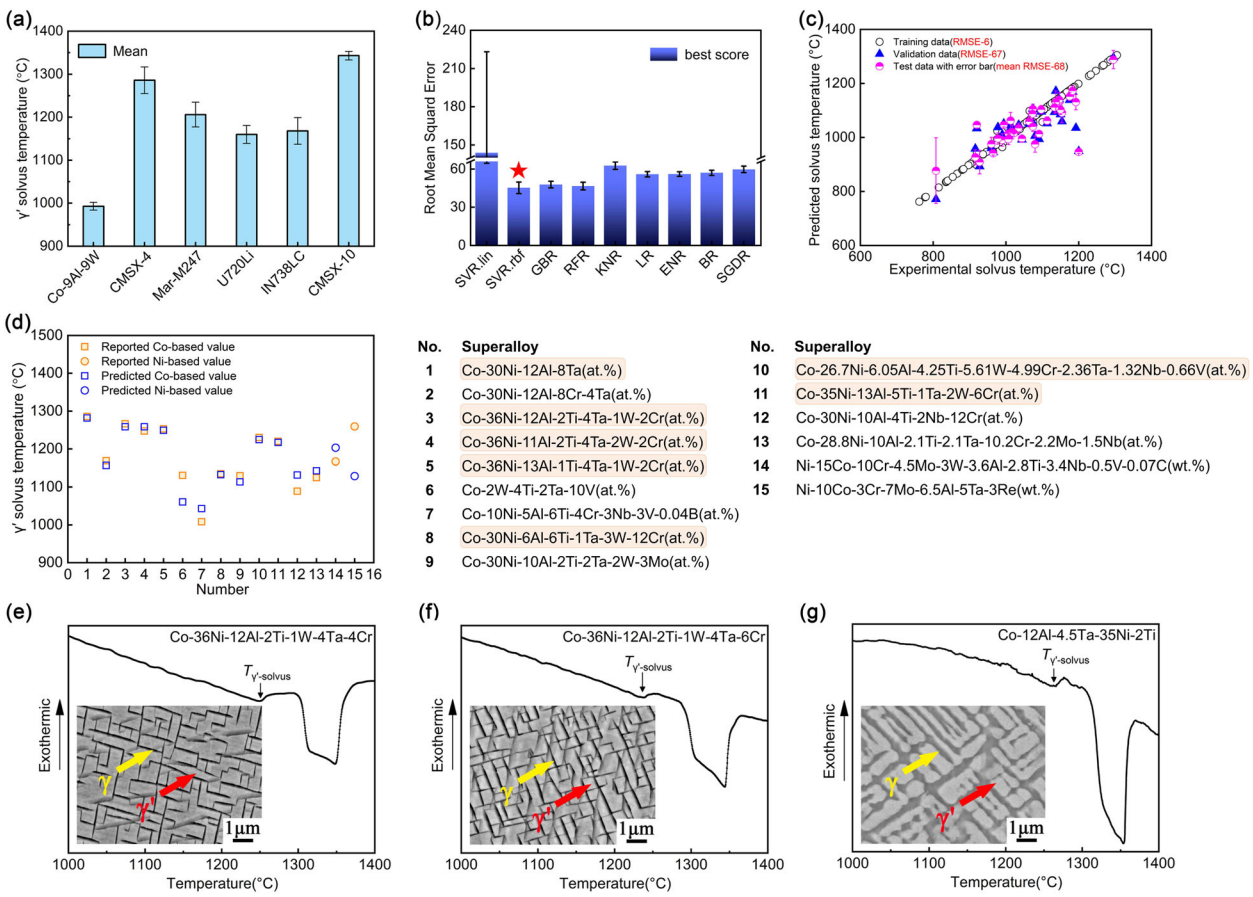


Fig. 6 Materials insights from the extracted data. a Extracted dataset of γ′ solvus temperature published from year 2004–2020. **b** Ashby chart for γ′ solvus temperature and density data. **c** Effect of element W on γ′ solvus temperature of Co-9Al-xW alloy. **d** Effect of different elements (Nb, Ta, and Ti) on γ′ solvus temperature and density of Co-30Ni-10Al-5Mo-x alloy.



No.	Superalloy
1	Co-30Ni-12Al-8Ta(at.%)
2	Co-30Ni-12Al-8Cr-4Ta(at.%)
3	Co-36Ni-12Al-2Ti-4Ta-1W-2Cr(at.%)
4	Co-36Ni-11Al-2Ti-4Ta-2W-2Cr(at.%)
5	Co-36Ni-13Al-1Ti-4Ta-1W-2Cr(at.%)
6	Co-2W-4Ti-2Ta-10V(at.%)
7	Co-10Ni-5Al-6Ti-4Cr-3Nb-3V-0.04B(at.%)
8	Co-30Ni-6Al-6Ti-1Ta-3W-12Cr(at.%)
9	Co-30Ni-10Al-2Ti-2Ta-2W-3Mo(at.%)
10	Co-26.7Ni-6.05Al-4.25Ti-5.61W-4.99Cr-2.36Ta-1.32Nb-0.66V(at.%)
11	Co-35Ni-13Al-5Ti-1Ta-2W-6Cr(at.%)
12	Co-30Ni-10Al-4Ti-2Nb-12Cr(at.%)
13	Co-28.8Ni-10Al-2.1Ti-2.1Ta-10.2Cr-2.2Mo-1.5Nb(at.%)
14	Ni-15Co-10Cr-4.5Mo-3W-3.6Al-2.8Ti-3.4Nb-0.5V-0.07C(wt.%)
15	Ni-10Co-3Cr-7Mo-6.5Al-5Ta-3Re(wt.%)

Fig. 7 Results from a machine-learned model of γ′ solvus temperature, and predictions for unexplored Co-based superalloys with high γ′ solvus temperature. a γ′ solvus temperatures distribution for certain superalloys. **b** Root mean square error for model selection by 5-fold cross validation. **c** The predicted values compared to values extracted from our pipeline for the SVR.rbf model showing the behavior of the training and test datasets. **d** The measured and predicted γ′ solvus temperatures of 15 superalloys recently reported in 2020 and 2021 that were not part of our database. **e** The microstructure and DSC heating curve for alloy Co-36Ni-12Al-2Ti-1W-4Ta-4Cr. **f** The microstructure and DSC heating curve for alloy Co-36Ni-12Al-2Ti-1W-4Ta-6Cr. **g** The microstructure and DSC heating curve for alloy Co-12Al-4.5Ta-35Ni-2Ti.

temperature, and liquidus temperature, were compiled into a highly structured material database containing article DOI, alloy named entity, chemical element, content, property specifier and property value. For subject-oriented text mining task of a small corpus like superalloys, a practical rule-based NER method and an effective heuristic multiple-relation extraction algorithm were proposed for the pipeline to overcome the obstacle of limited training corpus labels, and we achieved an F1 score of 92.07% for alloy named entity, and an average F1 score of 77.92% for relation extraction of γ′ solvus temperature, density, solidus temperature, and liquidus temperature. We also developed a common table processing tool with table parsing and relation extraction algorithm, which performs well with an F1 score 95.23%. Finally, we used the database to build a data-driven model for the γ′ solvus temperature to predict solvus temperatures of 15 new superalloys reported in 2020 and 2021, which were not part of our corpus. We obtain good agreement with a relative error of 2.27 %. The model was further employed to design unexplored Co-based superalloys with high γ′ solvus temperature (>1250 °C). Thus, our work emphasizes how knowledge represented in past publications can provide actionable insights for materials discovery by text mining. The code for the pipeline is available at https://github.com/MGEdata/SuperalloyDigger. A web-based toolkit is also available at http://SuperalloyDigger.mgedata.cn for online use. Automated text mining methods and tools to extract literature data for superalloys (and other alloy materials) have not previously been reported. Our extraction strategy and source code are not customized merely for superalloys; they present a general method for text extraction for alloy materials.

METHODS

Metrics for classification tasks

Precision, recall, and F1 score based on the confusion matrix were used as the metrics for classification tasks including text classification, table parsing, named entity recognition, text and table relation extraction. Precision evaluates the proportion of correctly classified instances among those classified as positive[43]. Recall quantifies the number of correct positive predictions made out of all actual positive cases[44]. The F1 score, which weights precision and recall equally, as calculated from Eq. (2), is the form most often used when learning from imbalanced data[45].

$$\text{F1 score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

Retrieval of articles and preprocessing

The scientific articles used in this work were taken largely from Elsevier publishers. Combined with the CrossRef search application programming interface (API)[46] and Web of Science search engine, a list of DOIs for superalloy articles was constructed. Then, a total of 14425 journal articles in plain text, XML, and HTML format corresponding to these DOIs were programmatically archived using Elsevier's Scopus and Science Direct APIs (https://dev.elsevier.com/) and the extended scrape package of "ChemDataExtractor".

The first stage of preprocessing for HTML and XML files was to isolate the relevant document domains, extract the raw text, and merge potentially fragmented data to produce a complete document record. For text from HTML and XML sources, semantic markups of paragraphs were parsed and merged into a plain text document as a list of paragraphs. For tables, individual cells are treated as separate text domains and stored in nested lists that represent the original table structure in the following table parser process. In particular, for Elsevier publications, its full-text API supports access to XML and plain text formats of one specific article. Here we used XML and plain text files from Elsevier for its table and text content extraction, respectively. The next stage is the metadata filtration of the programmatically downloaded articles, such as DOI, article ID, article title, journal, publication information, and many URLs. These metadata introduce obstacles to subsequent NER and relation extraction. Therefore, we programmatically filtered out the metadata information from the raw downloaded article documents by keywords and retained only the abstract and body of each full text. Furthermore, there were many inconsistent styles of designating values and units, such as "1039°C" and "1039□°C" (here "□" represents the space between 1039 and °C). The latter notation with a space was split into "1039" and "°C" after word tokenization by the Natural Language Toolkit (NLTK), an open source Python library for NLP[47]. We used regular expressions to locate all values followed by a unit in the full-text corpus and unified them by eliminating spaces.

Text classification

Amongst the hundreds of sentences in a document, classification allows us to determine which sentence contains the target property information to be extracted. It is a common practice with paragraph classification to train a binary classifier with positive samples representing related paragraphs and negative samples representing all other paragraphs[24]. This requires an adequate number of binary labels to be manually assigned to paragraphs. For sentence classification, the number of positive sentence samples to be labeled are limited, the overwhelming majority of sentences in a superalloy document are negative samples. Therefore, for such an imbalanced dataset, a supervised binary classifier does not perform well with a high degree of precision and recall. Therefore, we employed a rule-based method assisted by a semi-automatically generated dictionary to distinguish target sentences. Whether a sentence is relevant or not was determined by recognizing the superalloy name, target property specifier together with the specified value and unit.

A property specifier, such as the γ′ solvus temperature for a superalloy, can be written in various forms in a sentence, e.g. γ′-solvus temperature, and this needs to be captured in an appropriate way. To generate a synonym dictionary, we pre-train a word embedding model for the superalloy corpus on ~14,000 unlabeled full-text superalloy articles by using Word2Vec continuous bag of words (CBOW) in the code gensim (https://radimrehurek.com/gensim/)[48]. This uses information about the co-occurrences of words by assigning high-dimensional vectors (embeddings) to words in a text corpus to preserve their syntactic and semantic relationships. Assuming we have $V = 360{,}000$ unique words in the vocabulary of the whole superalloy corpus, Word2vec CBOW loops through all words in the training text and uses its one-hot encoding as an input for a neural network with window size 10. The weights of the hidden layer are given by a $V \times N$ dimensional matrix, where N is the space size (100 in our case) to "embed" the words in.

The order of magnitude of ~14,000 articles is smaller than the typical corpora used to train word embedding models, and this may cause the word embeddings to not be learned properly. So after training Word2Vec, we performed an intrinsic evaluation for the word embeddings by word correlation and word analogies. We tested top 100 most similar words to "superalloy", and amongst them are 38 words that are alternative forms for "superalloys", such as {superalloys, 0.884}, {superalloy, 0.815} and {superalloys, 0.810}, or misspelling forms {superlloy, 0.805}, {supperalloy, 0.729}, {superallys, 0.719}, {superlloys, 0.705}, etc. (The number after the comma is the cosine similarity.) The word "superalloys" and their similar spellings are highly relevant according to the word embedding model. Besides, a vector of "cobalt-based" – "nickel-based" + "IN-792" is closest to the vector "Haynes-188" by the word embedding model. It could be represented as "cobalt-based" − "nickel-based" + "IN-792" = "Haynes-188" (similarity = 0.508027). Similarly, we also obtained some relationships as "nickel-based"-"cobalt-based"+"Co-9Al-10W" = "GH690"(similarity = 0.570067), "nickel-based"-"cobalt-based"+"Co-9Al-9W" = "GH4742"(similarity = 0.540617), "nickel-based"-"cobalt-based"+"Co-9Al-9W-2Zr" = "GH690"(similarity = 0.588166), "nickel-based"-"cobalt-based"+"Co-30Ni-10Al-5Mo-2Nb-2Re" = "GH4169"(similarity = 0.556948). It illustrates that the word embeddings still captured useful relationships to a degree.

After training, we screened the top 100 words most similar to the target property "solvus" by calculating the cosine similarity in Word2Vec based on the obtained word embedding model. Supplementary Fig. 5 shows the Word2Vec CBOW and top 100 words with similarity in descending order of syntactic and semantic relationship in corpus. We manually selected the most likely synonym of solvus from these 100 words by expert knowledge (shown as the words in pink in the box in Supplementary Fig. 5), forming a synonym dictionary of the target "solvus". The property value may be a single number or a range, and the unit may be °C or K (Kelvin), which can be recognized by regular expression, as illustrated in Table 1. This then allows us to determine that a sentence is the target sentence (positive sample for sentence classification) in one document when a word from the synonym dictionary and a property value with a

specific unit appears simultaneously in a sentence. Our sentence classification method exhibited a precision, recall, and F1 score of 88.46%, 97.87%, and 92.93% respectively, evaluated by randomly sampling 30 articles (~3000 sentences).

Table parsing

Table parsing transforms complete table information, including table caption and body, into the structural format of a nested table cell list, and then classifies which table contains the chemical composition and target property information to be extracted. Initially, we performed table parsing on XML and HTML documents, given that plain text possesses no structural information of the table. For Elsevier publications, we modified an open source "table_extractor" tool to extract tables into a list format from XML files[49]; whereas for publications containing HTML files, pandas, an open source easy-to-use data structure and data analysis tool for the Python programming language, was used for HTML markup processing. Finally, the table under XML or HTML format was converted into cell list format by row with table caption. Subsequently, we performed table classification to screen tables containing chemical composition and target properties. Similar to text classification, the keyword "composition" and target property specifier, e.g., "γ′ solvus temperature", were matched in table captions; the workflow is shown in Supplementary Fig. 6. If the content and position of a table cell are converted correctly, then the table cell is regarded as a positive sample. In total, 9158 tables from ~14,000 articles were successfully extracted with an F1 of 98.8% by manual inspection from 4593 table cells of 20 articles. Table classification yielded 5327 composition tables and 114 tables with solvus temperature, respectively.

BiLSTM-CRF model

Usually, a bidirectional long short-term memory (BiLSTM) network with a conditional random field (CRF) layer, namely the BiLSTM-CRF model, can be used for NER tasks[33]. Supplementary Fig. 7 shows the neural architecture of our BiLSTM-CRF model. BiLSTM is a bidirectional recurrent neural network with an LSTM cell to solve the problem of long-term dependency in text data, capturing more semantic context dependence of sentences[50]. The input of BiLSTM is a layer of the word embedding (pre-trained in text classification) to yield a transformation function that accepts a plain text word and outputs a dense, real-valued, fixed-length vector. The outputs of BiLSTM are the corresponding probabilities under all labels of each word in a sequence, which are input into the CRF layer afterward to consider the correlations between labels in neighborhoods and jointly decode the best chain of labels for a given input sentence[51].

To train such a BiLSTM-CRF model, phrase-level labels were applied using the "BIO" sequence labeling method on 47777 words of 1090 sentences from 507 articles by humans[52]. "B" is used for the beginning of a named entity, "I" is for the middle part of a named entity, and "O" is for unrelated words. For example, the alloy named entity "Inconel 718" can be labeled as "B I". The 1090 annotated samples were split into train set and test set, where the ratio of train and test set was 1:1. Parameter tuning was employed by 5-fold cross validation with randomly selected hyper parameters, and the BiLSTM-CRF model was then trained with the best parameters. The final parameters of BiLSTM-CRF were set as: embedding_dim = 100, num_layers = 1, hidden_size = 16, lr = 0.01, dropout = 0.9. It achieved a categorical precision of 81.87%, recall of 66.97%, and F1 score of 73.67%. Then the model was applied on test set with 545 sentences (the same test set with ChemDataExtractor and our NER method) and the precision, recall and F1 were 51.99%, 36.53% and 42.91% respectively.

Snowball algorithm

Snowball system is a semi-supervised algorithm for generating patterns and extracting tuples from text documents, especially for limited labeled samples[31]. Snowball introduces a strategy for evaluating the quality of pattern and tuple extracted based on DIPRE[53]. Modified Snowball algorithm[22] can deal with quaternary relations, and performs clustering based on the ordering and number of entities. It can achieve high precision with fewer seeds than the original Snowball algorithm. In this work, we manually labeled 329 sentences containing γ′ solvus temperature information, and separately obtained 467 tuples in binary and quaternary form to evaluate Snowball and modified Snowball algorithm. The tuples in binary form include superalloy named entity, property value and their context information between different categories of entities, the tuples in quaternary form include superalloy named entity, property specifier, property value, property unit and their context information between different categories of entities. The number of initial seeds affects the algorithm performance greatly, so 50 and 100 manually labeled tuples were used as seeds to start the Snowball system and modified Snowball system for training, respectively. Finally, the trained Snowball system and modified Snowball system were used to extract relations from the remaining corpus, and precision, recall, and F1 score of the Snowball algorithm and modified Snowball algorithm were calculated through manual inspection.

Parameters after manual tuning for evaluating Snowball and modified Snowball on the test set are given in Supplementary Table 4. Table 3 shows the precision, recall, and F1 score of Snowball and modified Snowball algorithm on different seeds.

Data preprocessing for machine learning

After the automatic data extraction by the pipeline, 743 instances with both chemical composition and γ′ solvus temperature were obtained. For some superalloys, the extracted γ′ solvus temperatures for the same superalloy present differences. On one hand, we do not incorporate synthesis or processing conditions and other experimental aspects, including measurement uncertainties; on the other hand, for the same superalloy, some temperatures are calculated but some are by experiment, and some are reported in a range. Figure 7a presents the distribution of γ′ solvus temperature for the Co-9Al-9W, Mar-M247, U720Li, IN738LC, CMSX-4 and CMSX-10 superalloys with γ′ solvus temperature of 993 ± 9 °C, 1206 ± 28 °C, 1160 ± 20 °C, 1168 ± 31 °C, 1286 ± 30 °C, 1343 ± 9 °C respectively, with mean and standard deviation. In order to use these data for further analysis, some data preprocessing steps were manually performed as follows:

1. When different γ′ solvus temperatures were extracted for the same superalloy in the tables and text, data from the tables were retained and other data were excluded.
2. When simultaneously extracting the experimental and calculated γ′ solvus temperatures of a specific superalloy, the property value from the experiment was retained and other data excluded.
3. When multiple different γ′ solvus temperatures of a specific superalloy are obtained from different articles at the same time, the value with the highest occurrence rate was retained and other data excluded.
4. When the γ′ solvus temperature or composition value extracted for a specific superalloy is a given as a range (e.g. 1140–1150 °C), the mean value (1145 °C) of this range is retained.
5. The units of composition and γ′ solvus temperature were unified as atomic percentages and degrees Celsius, respectively.

Prediction model for γ′ solvus temperature

After data preprocessing, the extracted 743 records with both chemical composition and γ′ solvus temperature were down selected to 340 records including 262 cobalt-based records and 78 Ni-based records. We used 20 elements with 332 instances separated out from 340 instances to train machine learning models. Several well-known machine learning algorithms were used for model selection and parameter optimization by grid search, including support vector regression (SVR) with linear kernel (SVR.lin) and radial basis function kernel (SVR.rbf), Bayesian linear regression (BR), stochastic gradient descend regression (SGDR), k-nearest neighbor regression (KNR), random forest regression (RFR), gradient boosting regression (GBR), lasso regression (LR), and elastic net regression (ENR), under 100 times of 5-fold cross-validation. The SVR.rbf model performs the best with the lowest root mean square error (RMSE) on test sets. The model selection process is shown in Fig. 7b.

We divided the 332 data points into 298 data (90%) for training and validation, and the remaining 34 data (10%) for testing, and re-trained the SVR.rbf model with optimized parameters. During training, we employed 1000 bootstrap samples chosen randomly with replacement (238 data as training set from 298 each time, and the remainder as validation set) and trained 1000 different SVR.rbf models. The models were applied on the test set to yield 1000 corresponding predictions. The RMSE with mean and standard deviation on the test set is shown in Fig. 7c (for the training set, the uncertainties are from the 1000 bootstrap samples).

Synthesis and characterization

Raw metals with purity >99.95% were used, and the oxides and impurities on the surface of the raw metals were removed before processing the alloy. In order to ensure the homogeneity of the alloy composition and facilitate comparison, the alloy button ingots were prepared by vacuum arc melting, where each 30 g alloy was melted at least six times. After ultrasonic cleaning, the as-cast ingot was sealed in a quartz tube filled with high purity argon, and subjected to solution heat treatment at 1245–1260 °C for 12 h followed by air cooling. All samples were cut and subsequently aged at 1000 °C for 50 h followed by water cooling. The γ′ solvus temperatures were determined by DSC (NETZSCH STA 449 C) with high purity Ar flow. The samples for DSC of size φ 3 mm × 1 mm were tested in a temperature range 800–1400 °C at a heating rate of 5 °C min^{-1}. The line intercept method was used to measure the transformation temperatures based on the DSC heating curves.

The codes of our pipeline and machine learning model were run on Intel (R) core (TM) i7-9700U CPU with 3.00 GHz frequency and 8GB RAM, and a Graphics Processing Unit (GPU) from NVIDIA GeForce RTX 2080 Ti.

DATA AVAILABILITY

The original extracted dataset containing both composition and property information is open access with DOI of 10.12110/mater10.121.NKRDP.20211126.ds.61a0931d3b352a2169065520. The data used for machine learning after manual cleaning can also be found by DOI of 10.12110/mater10.121.NKRDP.20211126.ds.61a0931f3b352a2169065523.

CODE AVAILABILITY

The codes that support the findings of this study are available from https://github.com/MGEdata/SuperalloyDigger.

Received: 29 August 2021; Accepted: 10 December 2021;
Published online: 19 January 2022

REFERENCES

1. Zhang, H., Fu, H., Zhu, S., Yong, W. & Xie, J. Machine learning assisted composition effective design for precipitation strengthened copper alloys. *Acta Mater.* **215**, 117118 (2021).
2. Zhang, H. et al. Dramatically enhanced combination of ultimate tensile strength and electric conductivity of alloys via machine learning screening. *Acta Mater.* **200**, 803–810 (2020).
3. Granda, J. M., Donina, L., Dragone, V., Long, D. L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
4. Gesmundo, N. J. et al. Nanoscale synthesis and affinity ranking. *Nature* **557**, 228–232 (2018).
5. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
6. Wang, C., Fu, H., Jiang, L., Xue, D. & Xie, J. A property-oriented design strategy for high performance copper alloys via machine learning. *npj Comput. Mater* **5**, 1–8 (2019).
7. Rickman, J. M., Lookman, T. & Kalinin, S. V. Materials informatics: from the atomic-level to the continuum. *Acta Mater.* **168**, 473–510 (2019).
8. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput. Mater.* **5**, 1–17 (2019).
9. Xue, D. et al. An informatics approach to transformation temperatures of NiTi-based shape memory alloys. *Acta Mater.* **125**, 532–541 (2017).
10. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 1–9 (2016).
11. Wen, C. et al. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater.* **170**, 109–117 (2019).
12. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
13. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
14. Wen, C. et al. Modeling solid solution strengthening in high entropy alloys using machine learning. *Acta Mater.* **212**, 116917 (2021).
15. Zhang, Y. et al. Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models. *Acta Mater.* **185**, 528–539 (2020).
16. Jiang, X. et al. A strategy combining machine learning and multiscale calculation to predict tensile strength for pearlitic steel wires with industrial data. *Scr. Mater.* **186**, 272–277 (2020).
17. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
18. Swain, M. C. & Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
19. Krallinger, M., Rabal, O., Lourenço, A., Oyarzabal, J. & Valencia, A. Information retrieval and text mining technologies for chemistry. *Chem. Rev.* **117**, 7673–7761 (2017).
20. Kim, E. et al. Inorganic materials synthesis planning with literature-trained neural networks. *J. Chem. Inf. Model.* **60**, 1194–1201 (2020).
21. Kim, E., Huang, K., Jegelka, S. & Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Comput. Mater.* **3**, 1–9 (2017).
22. Court, C. J. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. data* **5**, 1–12 (2018).
23. Olivetti, E. A. et al. Data-driven materials research enabled by natural language processing and information extraction. *Appl. Phys. Rev.* **7**, 41317 (2020).
24. Kim, E. et al. Machine-learned and codified synthesis parameters of oxide materials. *Sci. data* **4**, 170127 (2017).
25. Ruan, J. et al. Accelerated design of novel W-free high-strength Co-base superalloys with extremely wide γ/γ′ region by machine learning and CALPHAD methods. *Acta Mater.* **186**, 425–433 (2020).
26. Liu, Y. et al. Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning. *Acta Mater.* **195**, 454–467 (2020).
27. Liu, P. et al. Machine learning assisted design of γ′-strengthened Co-base superalloys with multi-performance optimization. *npj Comput. Mater.* **6**, 1–9 (2020).
28. Jiang, X. et al. An materials informatics approach to Ni-based single crystal superalloys lattice misfit prediction. *Comput. Mater. Sci.* **143**, 295–300 (2018).
29. Su, Y., Fu, H., Bai, Y., Jiang, X. & Xie, J. Progress in materials genome engineering in China. *Acta Met. Sin.* **56**, 1313–1323 (2020).
30. Xie, J. et al. Machine learning for materials research and development. *Acta Met. Sin.* **57**, 1343–1361 (2021).
31. Agichtein, E. & Gravano, L. Snowball: extracting relations from large plain-text collections. In *Proc. 5th ACM Conference on Digital libraries* 85–94 (ACM, 2000).
32. Nadeau, D. & Sekine, S. A survey of named entity recognition and classification. *Lingvisticae Investig.* **30**, 3–26 (2007).
33. Huang, Z., Xu, W. & Yu, K. *Bidirectional LSTM-CRF models for sequence tagging*. Preprint at https://arxiv.org/abs/1508.01991 (2015).
34. Zhou, G., Su, J., Zhang, J. & Zhang, M. *Proc. 43rd annual meeting of the association for computational linguistics* 427–434 (ACL, 2005).
35. Sorokin, D. & Gurevych, I. Context-aware representations for knowledge base relation extraction. In *Proc. 2017 Conference on Empirical Methods in Natural Language Processing*(ed. Palmer, M. et al.) 1784–1789 (ACL, 2017).
36. Takanobu, R., Zhang, T., Liu, J. & Huang, M. A hierarchical framework for relation extraction with reinforcement learning. *Proc. AAAI Conf. Artif. Intell.* **33**, 7072–7079 (2019).
37. Bach, N. & Badaskar, S. A review of relation extraction. *Lit. Rev. Lang. Stat. II* **2**, 1–15 (2007).
38. Pang, H. T., Zhang, L., Hobbs, R. A., Stone, H. J. & Rae, C. M. F. Solution heat treatment optimization of fourth-generation single-crystal nickel-base superalloys. *Metall. Mater. Trans. A* **43**, 3264–3282 (2012).
39. Lass, E. A. Application of computational thermodynamics to the design of a Co-Ni-based γ′-strengthened superalloy. *Metall. Mater. Trans. A* **48**, 2443–2459 (2017).
40. Li, W., Li, L., Antonov, S. & Feng, Q. Effective design of a Co-Ni-Al-W-Ta-Ti alloy with high γ′ solvus temperature and microstructural stability using combined CALPHAD and experimental approaches. *Mater. Des.* **180**, 107912 (2019).
41. Ooshima, M., Tanaka, K., Okamoto, N. L., Kishida, K. & Inui, H. Effects of quaternary alloying elements on the γ′ solvus temperature of Co–Al–W based alloys with fcc/L12 two-phase microstructures. *J. Alloy. Compd.* **508**, 71–78 (2010).
42. Lass, E. A., Sauza, D. J., Dunand, D. C. & Seidman, D. N. Multicomponent γ′-strengthened Co-based superalloys with increased solvus temperatures and reduced mass densities. *Acta Mater.* **147**, 284–295 (2018).
43. Sniegula, A., Poniszewska-Marańda, A. & Chomatek, L. Study of named entity recognition methods in biomedical field. Procedia Comput. *Sci* **160**, 260–265 (2019).
44. Goutte, C. & Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *European Conference on Information Retrieval* (Losada, D. E. & Fernández-Luna, J. M.) 345–359 (Springer, 2005).
45. Japkowicz, N. Why question machine learning evaluation methods. In *AAAI workshop on evaluation methods for machine learning* (2006).
46. Lammey, R. CrossRef's text and data mining services. *Learn. Publ.* **27**, 245–250 (2014).
47. Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. ('O'Reilly Media, Inc.', 2009).

48. Rehurek, R. & Sojka, P. Software framework for topic modelling with large corpora. In *Proc. of the LREC 2010 workshop on new challenges for NLP frameworks* 45–50 (Citeseer, 2010).
49. Jensen, Z. et al. A machine learning approach to zeolite synthesis enabled by automatic literature data extraction. *ACS Cent. Sci.* **5**, 892–899 (2019).
50. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: continual prediction with LSTM. *Neural Comput.* **12**, 2451–2471 (2000).
51. Lafferty, J., McCallum, A. & Pereira, F. C. N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conference on Machine Learning* (ed. Brodley, C. E. & Danyluk, A. P.) 282–289 (ICML, 2001).
52. Reimers, N. & Gurevych, I. Optimal hyperparameters for deep LSTM-networks for sequence labeling tasks. Preprint at https://arxiv.org/abs/1707.06799 (2017).
53. Brin, S. In *International Workshop on the World Wide Web and Databases* (eds Atzeni, P. et al.) 172–183 (Springer, 1998).

ACKNOWLEDGEMENTS

This work is financially supported by the National Key Research and Development Program of China (2020YFB0704503, 2016YFB0700500), Guangdong Province Key Area R&D Program (2019B010940001), 111 Project (B170003), and USTB MatCom of Beijing Advanced Innovation Center for Materials Genome Engineering. We also acknowledge Science Direct APIs (https://dev.elsevier.com/) in term of the open access to subscription content for text mining provided to subscribers for noncommercial research purposes.

AUTHOR CONTRIBUTIONS

J.X. conceived the project. The study was designed by J.X., Y.S., and T.L., text mining programs were performed by W.W., X.J., S.T., and D.D., the manuscript prepared by W.W., X.J., S.T., D.D., Y.S., T.L., and J.X.. All authors discussed the results and commented on the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00687-2.

Correspondence and requests for materials should be addressed to Yanjing Su, Turab Lookman or Jianxin Xie.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.