特許文書の材料科学技術表現抽出

Named Entity Recognition from Materials Science Patent Documents

酒井　敏彦*
Toshihiko Sakai
千綿　伸彦**
Nobuhiko Chiwata
峯　恒憲***
Tsunenori Mine

*
九州大学大学院システム情報科学府
Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University

**
株式会社プロテリアル
研究開発本部グローバル技術革新センター
兼　知的財産部IPソリューショングループ
Digital Technology Development Dept., Global Research & Innovative Technology Center, IP Solution Group, Intellectual Property Dept., R&D Division, Proterial, Ltd.

***
九州大学大学院システム情報科学研究院
Faculty of Information Science and Electrical Engineering, Kyushu University

材料科学分野では，マテリアルズ・インフォマティクスと呼ばれるデータ駆動型の材料設計が盛んに検討されている。文書から抽出した技術データの活用は，その一環として重要視されている。特に，材料科学分野の研究者視点で整理された材料組成等の情報，その中でも，材料を構成するための基本的情報である元素比率を技術文書から機械的に抽出し，構造的に把握する方法が求められている。文書中の特定情報（固有表現）を抽出する方法として，抽出対象の手掛かりとなる「手掛かり表現」を利用する方法が提案されている。しかし，この方法は新たな手掛かり語を獲得するためのコストが問題であった。そこで，本研究では，特許明細書を対象として，材料に関する情報を固有表現として特定するための新たな手法に，手掛かり語タグを用いた固有表現抽出手法を提案する。手掛かり語タグは，抽出対象の固有表現と手掛かり語とを同時に抽出することを可能とするだけでなく，抽出対象の固有表現との意味的まとまりを持つ構造の獲得に有効である。この手掛かり語タグを導入することで，固有表現の抽出精度が向上し，全体の抽出性能は劣化しないこと，さらに，新たな手掛かり語の獲得も可能であることを実験により確認した。

In the field of materials science, data-driven materials design, called materials informatics, has been considered. The use of technical data extracted from documents is an important part of this process. In particular, there is a need for a method to automatically extract from technical documents information such as material compositions organized from the point of view of materials science researchers, especially element ratios, which provide fundamental information on the composition and structure of materials. As an approach to extracting specific information (named entities) from documents, methods using “clue expressions”, which are clues to the target named entities for extraction, have been proposed. However, the cost of acquiring new clue words is a problem in these methods. In this study, we propose a new method for extracting named entities in materials science from patent specifications using clue word tags; the proposed method identifies information about materials as named entities. The clue word tag not only makes it possible to extract the target named entities and the clue word simultaneously, but is also effective in obtaining a structure that has semantic cohesion with the named entities to be extracted. Experiments confirmed that the introduction of clue word tags improves the accuracy of named entity extraction without degrading the overall extraction performance, and also enables the acquisition of new clue words.

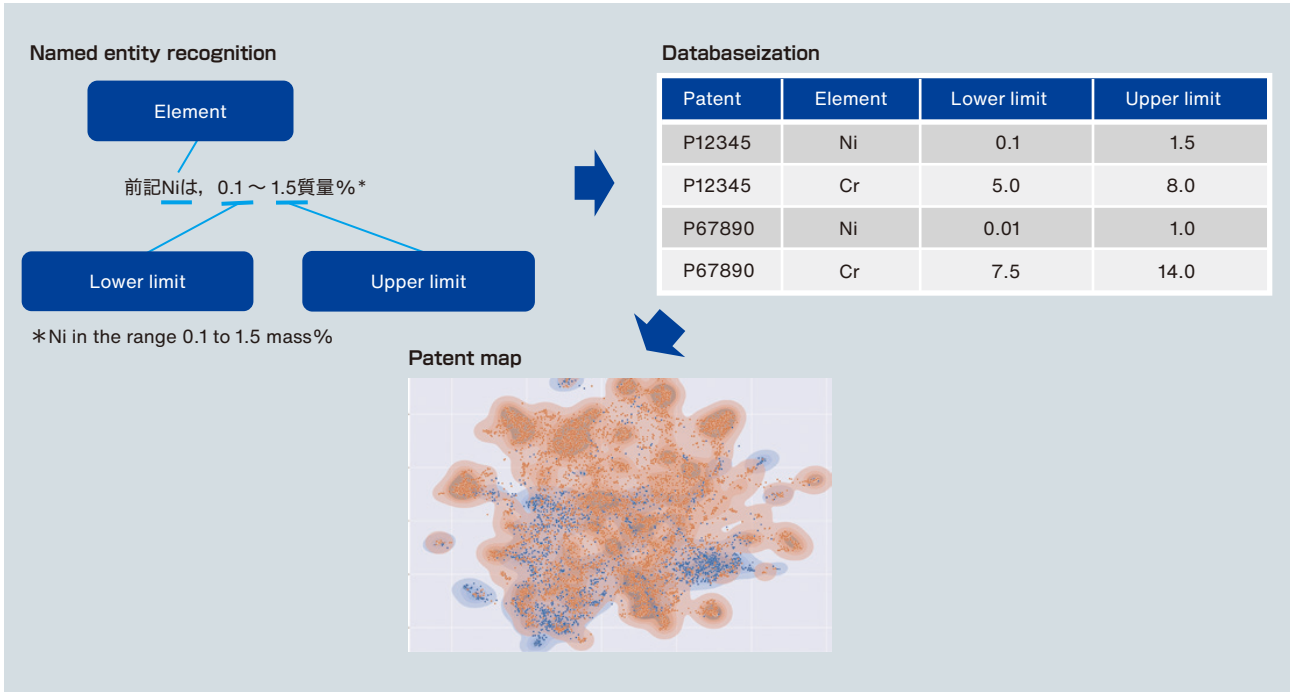
■ **Key Words**：固有表現抽出，材料科学，特許文書
■ **R&D Stage**：Research

1. 諸言

材料科学分野では，データ科学の知識と技術を用いて効果的に材料開発を行う「マテリアルズ・インフォマティクス（MI）」[1),2)] による新材料開発手法の活用が進んでいる。MIは，プロセス科学基盤の拡充を目的とした「プロセス・インフォマティクス」[3)] と組み合わせることが期待されている。そのためには，実際に材料を生成して得られた実験の値や論文情報だけでなく，特許の情報も活用していく必要がある[4)-6)]。

テキストから，地名や日付，人名などの固有表現を抽出する技術として，固有表現抽出[7)-12)] がある。固有表現抽出を用いることで，例えば，テキストにおける文脈に応じて，さまざまな形で用いられている「Fe」や「鉄」という文字列のうち，元素記号として用いられている「Fe」や「鉄」を特定することができる。このように，固有表現抽出により，文字列に対して，文字種や品詞ではなく，情報の性質ごとに関連付けることができる。したがって，固有表現抽出を用いれば，非構造データから関連付けられた情報の性質ごとに文字列を抽出し，情報の性質を考慮して得られた文字列を使ってデータベースを構築することができる。一旦，データベースを構築してしまえば，検索や要約，質問応答などへの情報活用が容易となる。しかしながら，一般的な地名や人名といった固有表現ではなく，特定の技術分野における固有表現抽出の応用は，当該技術特有の情報を扱う専門性を考慮する必要がある。本研究で着目したのは，材料科学分野における，材料を構成する構成物や比率である。材料技術文書には，「ニッケルが0.1%～1.5%，クロムが0.01%～1.0%」のように構成物と量の比率が記載されており，新たな材料開発を行う際の貴重な参考情報となる。一般に材料技術文書には，比率だけでなく物質の特性値やプロセス制御範囲といった数値範囲との関連付けが求められている情報が存在し，本技術の展開が可能である。

本研究では，材料科学技術文書として特許明細書を対象とする。特許調査においても，構成物と量の比率は，調査の重要な対象であるとともに，貴重な技術情報でもある。専門性を反映した特許データベースが構築できれば，特許調査の効率化に繋げることができるのみならず，大量に存在する特許明細書に含まれる技術情報をマテリアルズ・インフォマティクスへ応用できる。**図1**に特許明細書からデータ可視化までの一連のデータ活用の概要を示す。特許文書から自動で構成物の量や比率の抽出を行い，データベースを構築することで，図示化による比較が容易に可能となる。この時，情報を抽出するためには，専門的な知識が必要である。例えば，「ニッケルが0.1%～1.5%」というテキストには，「ニッケル」という構成物が「0.1%～1.5%」含まれる情報を有しているが，類似したテキストである「変形量が5%～



Patent	Element	Lower limit	Upper limit
P12345	Ni	0.1	1.5
P12345	Cr	5.0	8.0
P67890	Ni	0.01	1.0
P67890	Cr	7.5	14.0

図1 技術文書からの情報抽出
Fig.1 Conceptual diagram of information extraction from technical documents

10%」は，構成物ではなく，状態の変化量を示している。さらに，関連する情報を構造化するためには，「ニッケル」と下限を表す数値「0.1%」，上限を表す数値「1.5%」とを関連付けた意味的なグループとして関連付けて抽出する必要がある。実際の抽出においては，専門家でなければ判別ができない記述も多く，得られた情報の関連付けには専門家の多大な工数が必要になる。そこで，本研究では，後述する手掛かり語タグを導入することにより，意味的関連性を含めた技術情報抽出を可能とした。さらに，手掛かり語タグ自体を新たな固有表現として設計し，新たな手掛かり語を獲得しながら情報抽出を行う可能性を見出した。

なお，本報告は，株式会社プロテリアルにおけるマテリアルズ・インフォマティクス技術「D2Materi™」の一環として実施したものである。

2．関連研究

2.1 材料科学技術文書からの情報抽出

材料科学分野においても，データ駆動型のマイニング手法が求められており[13]，材料開発や材料設計のための情報抽出に，機械学習を用いるアプローチが盛んに研究されている[5),14)]。材料科学技術文書から情報抽出を行う研究として，論文から材料名等を固有表現抽出により抽出する研究[15),16)]や材料の特性値を予測する研究[17),18)]，教師なしで単語の埋め込み表現を獲得し，材料科学分野の潜在的な知識を得る研究がある[19)]。また，材料科学技術文書からキーフレーズとその間の関係性を抽出する研究がある[20)]。特許から情報抽出を行う研究として，中国語の医学特許において，ルールベースに基づく手法により固有表現抽出を行う研究がある[21)]。しかしながら，従来の研究では，構成物とその比率を関連付けた状態で抽出するアプローチは取られていない。

2.2 手掛かり語に関する研究

抽出する対象と共起する語を手掛かり語として定義することで，情報抽出を行う研究が取り組まれてきた。この手法では，抽出する対象と共起する語群を手掛かり語として，事前に定義する必要がある。しかし，最初に定義した手掛かり語のみでは，目的とする情報を抽出するための手掛かりとして不十分である可能性がある。そのため，定義された手掛かり語と共起する表現パターンから新たな手掛かり語を自動的に獲得するブートストラップ型の手法[21)-25)]が提案されている。しかしながら，ブートストラップ型の手法では，手掛かり語を元に共起する表現パターンの獲得，獲得した表現パターンから新たな手掛かり語の獲得，という2つの手順を繰り返し行う必要があり，計算コストがかかる。また，情報抽出と新たな手掛かり語を同時並行的に獲得することができない。

2.3 大規模言語モデルを用いた研究

従来は，Transformer[26)]のエンコーダを用いたBERT[27)]による教師有り学習が主流であったが，2023年に入り，OpenAIが発表した生成AIであるChatGPT[28)]に代表される大規模言語モデル（LLM: Large Language Model）の導入が，さまざまな分野で進んでいる。材料科学分野などの専門文書での情報抽出においても，LLMを用いた研究が行われている。Huら[29)]は，医学文書から症状や診断，医学的問題などの固有表現を固有表現抽出する際にChatGPTやGPT-3.5を用いた手法に取り組んでいる。しかし，BERTを医学文書に特化させたBioClinicalBERT[30)]をベースとした既存の教師有り学習の手法が性能が高かったと報告している。また，ChatGPTでは，医療的知識が不足していたこと，および，予測する単語を勝手に言い換えてしまうことが要因となり抽出精度が低下した，と報告している。Polakら[31)]は，材料科学技術文書から材料名や材料の特性値を抽出するために，GPTを用いた手法に取り組んでいる。抽出目的の文書から，最初にGPT-3.5を用いて，学習を行わず関係する文の分類を行い，教師データでファインチューニングしたBARTやDeBERTaを用いた文分類を実施した後，人手で材料の特性値を抽出していた。また，Polakら[32)]は，プロンプトを繰り返し試すフローチャート方式を提案した。このようにLLMを用いることで材料科学分野の情報を抽出できる可能性はあるが，現時点では，既存の教師有り学習の手法を利用した専用モデルによる抽出性能が高いこと，さらには，本研究のように固有表現抽出対象の設計から行う際には，大規模な計算機資源を必要とせず，抽出性能の上でも最先端モデルと遜色のない高速で軽量な教師有り学習モデルが用いやすい。

3. 提案手法

3.1 固有表現抽出

本節では，本研究で扱う固有表現抽出について説明する。固有表現抽出とは，文書内の固有表現を識別し，抽出することである。このとき，文のいくつかの要素が連なったものを系列，系列内のそれぞれの要素にラベルを付与することを系列ラベリングと呼ぶと，n単語からなる入力$\boldsymbol{x}=(x_1, x_2, \ldots, x_n)$に対して，出力$\boldsymbol{y}=(y_1, y_2, \ldots, y_m) \in Y$を返す系列ラベリング問題として定式化できる。ここで，Yはm個のラベルの集合である。本研究では，固有表現抽出を解くため，$m = n$と固定し，Yを固有表現の集合とする。入力$\boldsymbol{x}$に対する出力$\boldsymbol{y}$を条件付き確率$p(\boldsymbol{y}|\boldsymbol{x})$でモデル化し，

$$\boldsymbol{y}^* = \underset{\boldsymbol{y}}{arg\,max}\ p(\boldsymbol{y}|\boldsymbol{x})$$

となる出力$\boldsymbol{y}^*$を探索する。

3.2 固有表現と手掛かり語タグの設計

本節では，本研究で扱う固有表現について説明する。情報抽出や文の構造化を協働で行う場として，日本では主に，IREX*1や森羅プロジェクト*2で固有表現抽出タスクが取り組まれている。IREXで定義された固有表現の種類として，人名や組織名，地名などがある。本研究では，材料科学分野の構成物の量や比率を抽出するため，材料を構成する構成物や比率の情報を表す独自の固有表現を設計した。なお，実験に用いた固有表現表法には，IOB2（Inside, Outside, Beginning）方式[33]を採用した。

固有表現抽出は，ある意味を表すラベル（タグ）を，そのタグと対応する語句に付与する（ラベリングする）ことで実施する。加えて，本研究では，固有表現抽出と同時に，意味的なまとまりを持つグループ化の手掛かりを与える。具体的には，抽出対象である語句へのラベリングを行うタグと，意味的なグループ化のための手掛かり語となる手掛かり語タグの2つを定義した。主な固有表現の定義を**表1**に示す。抽出対象と手掛かり語タグの固有表現を設計している。また，表に記載したタグ以外に計15種類の固有表現を設計した。**表1**では，本研究の目的である構成物と量の比率に議論を絞るため，構成物の量や比率に関与しない手掛かり語タグと抽出対象以外の固有表現の記載は省略する。抽出対象のタグとしては，構成物である元素（element）や化合物（compound）を示す固有表現として「atom」を用意した。比率を表す下限（lower limit）・上限（upper limit）の数値を示す固有表現には，下限値「fig_LL」，上限値「fig_UL」の2つのタグを用いた。さらに，数値の関係性および意味を付与する「手掛かり」に相当する手掛かり語タグとして「limitation」のタグを用いる。手掛かり語タグが付与される語句自体は，意味ある情報とはならないが，抽出された他の固有表現との位置関係から，文字通り，意味的なグループ化のための手掛かり語となる。

3.2.1 手掛かり語タグによる意味的関連性の抽出

手掛かり語タグを導入することで，抽出対象の固有表現との意味的関連性を抽出することが可能となる。元素と元素の含有量の下限・上限の例文を**表2**に示

*1 https://nlp.cs.nyu.edu/irex/index-j.html
*2 http://shinra-project.info/

表1 本研究で扱う固有表現の定義
Table 1 Definition of named entities used in this study

Named entity	Description	Tag species
atom	Elements or compositional components of a substance 物質の元素や組成成分を表す語句	Target tag
fig_LL	Lower limit of elemental content 元素の含有量の下限	Target tag
fig_UL	Upper limit of elemental content 元素の含有量の上限	Target tag
limitation	Limits and ranges 以上，以下，未満など制限・範囲を表現	Clue word tag
selection	Expressions of selectivity, such as the number of types of elements 元素の種類数など選択性を表す表現	Clue word tag

す。この例では，元素の含有量の下限・上限の条件部分「～」が元素の含有量の下限・上限を抽出するための手掛かりとなる。「～」より前が下限であり，後が上限であると推定され，「fig_LL」，「limitation」，「fig_UL」の順にタグが付与され，手掛かり語タグの位置により，数量を表す「fig_LL」，「fig_UL」が関連した情報であることが明確となる。同様の手掛かり語としては，「以下」や「以上」などもあるが，必ずしも下限・上限の順で出現しない。手掛かり語タグを考慮しない場合は，構成物の含有量の下限・上限の数値のみを抽出していたが，含有量が複数回続いた場合，単なる数値の大小による判別だけでは，数値の順序を正しく識別できない。一方で，手掛かり語タグを導入することで，手掛かり語タグの出現位置と抽出対象の位置関係から数値の順序を正しく識別することができる。さらに，得られた比率範囲の関係性から構成物との関係性を推定し，意味的なグループを抽出する。**図2**に本節で議論した手掛かり語タグと抽出対象の関係性の概念図を示す。

構成物の含有量比率だけでなく，複数の構成物からなる，という意味を持つ情報も存在する。そこで，手掛かり語タグ「selection」を設計した。例えば，「(構成物である)元素が少なくとも一種類含有される」，もしくは「二種類以上含有される」という情報は，構成物と共起する文脈で用いられる。例文を**表3**に示す。元素「Mg, TI, Al, Mg」のいずれかから，「1種又は2種以上を含有」という複数の構成物の含有を表す手掛かり語タグ「selection」を示した例である。

以上のことから，意味的なグルーピングを実施するために，抽出対象以外のタグである手掛かり語タグを導入した。しかしながら，新たな概念の導入による抽出対象そのものの抽出精度や設計したタグ全体へのタグ付与精度の評価が必要である。本研究では，手掛かり語タグ導入による精度への影響を議論する。

表2 limitation のタグ付与例

Table 2 Example of limitation tag

Word	Sn	:	1	.	5	～	2
Tag	atom	O	fig_LL	fig_LL	fig_LL	limitation	fig_UL

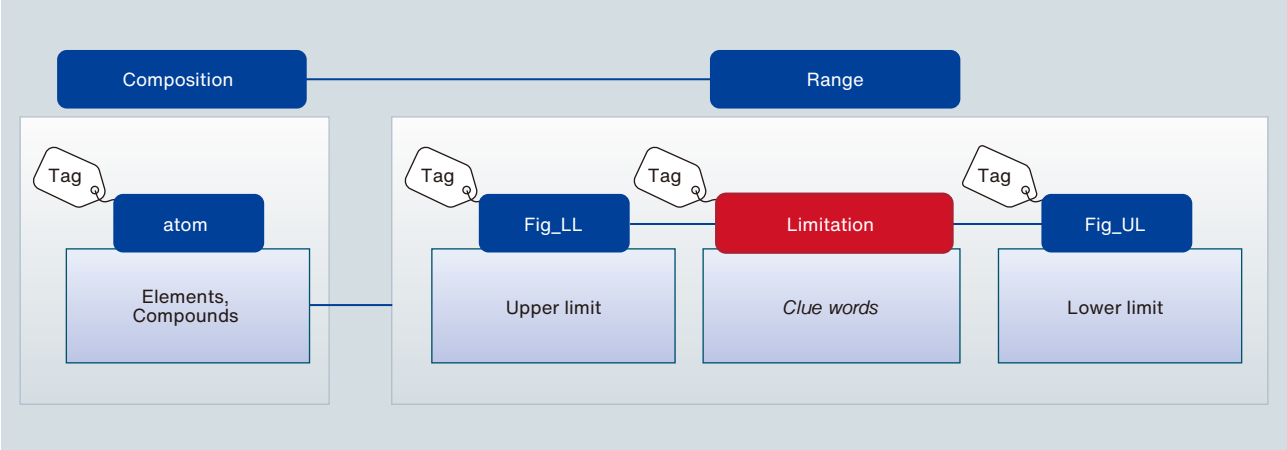


図2 手掛かり語タグと抽出対象の関係性の概念図

Fig.2 Schematic diagram of relationship between clue words tag and extracted words

表3 selection のタグ付与例

Table 3 Example of selection tag

Word	Mg	…	の	1	種	又は	2	種	以上	を	含有
Tag	atom	…	O	selection	selection	selection	selection	selection	selection	selection	selection

3.2.2 手掛かり語タグによる新たな手掛かり語の獲得

従来手法では，手掛かり語を抽出規則として予め定義することで，手掛かり語と共起する表現を新たな手掛かり語として獲得していた。本手法では，手掛かり語自体を抽出するため手掛かり語タグを固有表現として定義する。これにより，予め定義された語句以外も手掛かり語として，抽出対象と同時に得ることができる。したがって，情報抽出と新たな手掛かり語の獲得とを同時に実施可能である。本研究では，新たな手掛かり語の獲得の可能性についても議論を行う。

3.2.3 手掛かり語タグ導入によるリサーチクエスチョンの設定

手掛かり語タグを導入することにより，比率範囲の関係性から構成物との関係性を推定し，意味的なグループを抽出できる。また，固有表現抽出過程で手掛かり語タグとして付与された単語は，新たな手掛かり語の獲得と捉えることができる。しかしながら，手掛かり語タグの導入は，新たな固有表現を追加することとなるため，全体性能や個別タグの精度低下が懸念される。そこで，本研究では，以下の三点のリサーチクエスチョン（RQ）を設定し，条件付き確率場[34]（CRF: Conditional Random Fields）の手法を用いた実験により検証を行うこととした。

RQ1）手掛かり語タグにより抽出対象タグの抽出精度は向上するか
RQ2）手掛かり語タグの追加は全体性能に影響するか
RQ3）手掛かり語タグにより有益な手掛かり語が獲得できるか

4. 実験

4.1 固有表現抽出モデル

本節では，固有表現抽出モデルについて述べる。前節で提案した手掛かり語タグが抽出対象とするタグの精度向上および精度向上の際に寄与する素性を検証するため，CRFを用いた固有表現抽出モデルを構築する。CRFは，タグ間の遷移を用いる手法であり，軽量なモデルでもある。そこで，今回の設計の有効性を議論するのに適していると考えた。

CRFでは，入力系列 $\boldsymbol{x}=(x_1, x_2, ..., x_n)$ に対して，出力 $\boldsymbol{y}=(y_1, y_2, ..., y_n)$ を求める。

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z_{\boldsymbol{x},\boldsymbol{w}}} \exp(\boldsymbol{w}\cdot\phi(\boldsymbol{x},\boldsymbol{y}))$$

このとき，$\boldsymbol{w}$は学習するパラメータ，$Z_{\boldsymbol{x},\boldsymbol{w}}$は，$\Sigma_{\boldsymbol{y}} P(\boldsymbol{y}|\boldsymbol{x})=1$を保証する係数で，$Z_{\boldsymbol{x},\boldsymbol{w}}=\Sigma_{\boldsymbol{y}} \exp(\boldsymbol{w}\cdot\phi(\boldsymbol{x},\boldsymbol{y}))$と定義する。$\phi(\boldsymbol{x},\boldsymbol{y})$は素性関数である。CRFを用いて事例 $\boldsymbol{x}$ を分類する場合は，以下の最大化問題を解く。

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} \frac{1}{Z_{\boldsymbol{x},\boldsymbol{w}}} \exp(\boldsymbol{w}\cdot\phi(\boldsymbol{x},\boldsymbol{y})) = \arg\max_{\boldsymbol{y}} \boldsymbol{w}\cdot\phi(\boldsymbol{x},\boldsymbol{y})$$

しかし，上記の場合，計算に時間がかかるため，CRFでは，以下のような仮定を置き，最大化問題を解く。このとき，tは拡張可能であり，$\sum_t \phi_k(\boldsymbol{x}, y_t, y_{t-1}, y_{t-2})$などの仮定を置くこともできる。

$$\phi_k(\boldsymbol{x},\boldsymbol{y}) = \sum_t \phi_k(\boldsymbol{x}, y_t, y_{t-1})$$

したがって，次の最大化問題を解くこととなる[35]。

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} \sum_t \boldsymbol{w}\cdot\phi(\boldsymbol{x}, y_t, y_{t-1})$$

4.2 データセット

特許明細書にはさまざまな項目の記述があるが，本研究では「特許請求の範囲」の項目を対象とした。「特許請求の範囲」には，発明を特定するための要件が記載されているため，これにより保護すべき発明の範囲が明示される。したがって，特許明細書にとって非常に重要な項目である。また，出願される特許には，さまざまな技術分野の特許明細書が含まれる。そこで，本研究では，今回取り扱う検討技術分野に近い「合金」に関する国際特許分類カテゴリ（IPC）を利用することで，構成物の量や比率の記述をより多く含む母集団を設定した。具体的には，日本特許庁に出願されたIPC分類C22C（合金）に属する登録特許のうち，2000年1月から2021年8月までの期間に公開された特許明細書の「請求の範囲」の項目に記載された文章中に「組成」を含む15,053件を抽出した。専門家の助言の下，文章中に「組成」を含む文章には「元素や化合物の含有量の下限・上限の数値を示す」表記が含まれると判断したためである。得られたデータから，「請求の範囲」に含まれる99,585文から978文を無作為に取り出し，専門家によるアノテーションを実施し，教師データとした。教師データのうち，一部で手掛かり語タグを含まない文が含まれていた。そこで，今回作成した教師データすべてを含むデータセット（All）と，手掛かり語タグを文中に含むデータセット（OnlyClueTag）の2種類のデータセットを作成した。

すなわち，

・データセット（All）：すべてのタグを含む978文

・データセット（OnlyClueTag）：手掛かり語タグを含む922文

の2種類である。**表4**に各データセットのタグ数を示す。

表4 データセットのタグ数

Table 4 The number of tags in dataset

	Dataset (All)	Dataset (OnlyClueTag)
atom	6,096	6,056
fig_LL	2,939	2,938
fig_UL	4,162	4,162
limitation	4,935	4,935
selection	1,022	1,022
balance	382	382
fig	20	12
unit	4,693	4,677
formula	341	314
sum	88	88
use	990	937
variable	589	567

4.3 実験条件

CRFのフレームワークにはcrfsuite*3を用いた。CRFでは，入力される単語の持つ属性により素性関数を学習する。本研究では，前後の1単語と2単語，単語の品詞細分類の情報，単語の文字種および正解のタグを素性として利用し，その組み合わせごとにF1-scoreにより，評価した。なお，入力した文は，MeCab*4を用いて，IPA辞書に従って形態素解析を実施し，品詞細分類を求めた。また，文字種は，空白，アラビア文字，英字小文字，英字大文字，ひらがな，カタカナ，その他を用いた。本研究で用いた素性の組み合わせを**表5**に示す。本条件において，学習データに対し，グリッドサーチにより，最適なハイパーパラメータを求め，5分割交差検証による評価を行った*5。

*3 https://github.com/scrapinghub/python-crfsuite

*4 https://taku910.github.io/mecab/

*5 最大イテレーション数を100とし，アルゴリズムはL-BGFS法を用いた勾配降下法のlbfgsを用いた。L1正則化およびL2正則化のパラメータはそれぞれ，L1: (0.1, 0.5, 1.0)，L2: (0.05, 0.1, 1.0)の中から探索を行った。また，all possible transitionsはTrueとした。その他は，デフォルトのパラメータを用いた。

表5 CRF で用いた素性の組み合わせ

Table 5 Combination of features used in CRF

	Word count before and after	PoS type	Type of character
Pattern A	2	○	○
Pattern B	2	–	○
Pattern C	2	○	–
Pattern D	2	–	–
Pattern E	1	○	○
Pattern F	1	–	○
Pattern G	1	○	–
Pattern H	1	–	–

4.4 評価方法

データセットをテストデータと学習データに分割する比率を決定する。まず，テストデータを10%と固定し，学習データの比率は，90%，80%，...10%とした。このとき，学習比率が増加するほど，F1-scoreが高い結果が得られたため，学習・テストの比率を90%，10%とした。データセット（OnlyClueTag）から，手掛かり語タグ（limitation, selection）を含むデータセットと含まないデータセットを作成した。手掛かり語タグを含むデータセットをCW: ClueWords，含まない場合をNCW: NotClueWordsとし，以後，このように表記する。この2つのデータセットによる評価結果を比較することで，手掛かり語タグの有効性を評価する。なお，手掛かり語タグ自体は，抽出対象タグを抽出することが目的であるため，「CW: ClueWords」では，手掛かり語タグ自体の評価値は除いて，評価を実施した。

4.4.1 RQ1），RQ2）の検定方法

性能評価結果において，統計的検定を行うため，ウィルコクソンの順位和検定を用いた。有意水準は $p < 0.05$ とし，両側検定を実施した。CW: ClueWordsとNCW: NotClueWords各群の抽出性能はMicro-F1-scoreを用い，個別のタグの抽出性能は，各タグのF1-scoreを用いて，検定を行う。

・検定（素性群）：学習比率を固定した場合における，パターンA～パターンH（8種類）のF1-scoreで比較

・検定（学習比率群）：素性を固定した場合における，学習比率90%～10%（9種類）のF1-scoreで比較

以降は，各検定について，検定（素性群），検定（学習比率群）と記載する。

4.4.2 RQ3)の手掛かり語獲得評価

RQ3)の手掛かり語獲得の評価では，学習比率90%から10%の9通り，CRFで用いるパターンAからパターンHの8通り，2種類のデータセットであるCW: ClueWordsとNCW: NotClueWordsの組み合わせ合計144通りの予測結果中の手掛かり語タグlimitationおよびselectionが付与された語句を個別に評価した。

5. 実験結果と考察

5.1 RQ1:手掛かり語タグにより抽出対象タグの抽出精度は向上するか

5.1.1 抽出対象タグの結果

結果を**表6**に示す。値はF1-scoreを表す。個別の素性集合において，CW: ClueWordsとNCW: NotClueWordsを比較し，性能が高い項目に下線を付す。また，各固有表現およびMicro-F1-scoreでの最良値を太字とした。パターンCにおいて，Micro-F1-scoreが一番高い性能を得た。パターンCにおけるatom, fig_LL, fig_ULはいずれもCW: ClueWordsがNCW: NotClueWordsを上回る結果を得た。

F1-scoreにおいて検定(素性群)を行った結果，atomおよびfig_ULは有意差がみられなかった。一方，fig_LLはp値が0.0027となり，有意差が確認できた。続いて，F1-scoreの値を比較すると，fig_LLとfig_ULでは，いずれの素性においても，CW: ClueWordsはNCW: NotClueWordsと比較し，同等以上の結果となっている。したがって，手掛かり語タグlimitationの導入の影響は，性能を低下させず，むしろ上昇させる結果となった。本結果を，手掛かり語タグlimitationと意味的関連性を持つfig_LLとfig_ULとの文中の隣接割合と比較する。タグの隣接割合を**表7**に示す。この表は，データセット(All)における手掛かり語タグと抽出対象のタグが隣接する割合を調査したものである。手掛かり語タグ「limitation」の前後に抽出対象「fig_LL」，「fig_UL」のいずれかが含まれる割合および手掛かり語タグ「selection」の前後に抽出対象「atom」が含まれる割合を示す。limitationおよびfig_LLとfig_ULと隣接する割合が61.96%と高い。すなわち，CRFはタグの遷移を利用するため，手掛かり語タグlimitationが抽出対象のfig_LLおよびfig_ULの精度向上に直接的に寄与している結果が得られたと考えられる。

表6 抽出対象タグの結果
Table 6 Results for target tags

Feature of pattern	Dataset	atom	fig_LL	fig_UL	Micro-F1-score
A	CW	0.9658	**0.9865**	**0.9935**	0.9727
	NCW	0.9630	0.9712	0.9909	0.9690
B	CW	0.9658	0.9846	0.9922	0.9719
	NCW	**0.9688**	0.9729	0.9909	0.9730
C	CW	0.9668	**0.9865**	**0.9935**	**0.9741**
	NCW	0.9610	0.9730	0.9922	0.9705
D	CW	0.9608	0.9807	0.9922	0.9700
	NCW	0.9620	0.9767	0.9922	0.9706
E	CW	0.9530	0.9733	0.9870	0.9664
	NCW	0.9522	0.9572	0.9832	0.9635
F	CW	0.9539	0.9733	0.9870	0.9683
	NCW	0.9558	0.9572	0.9858	0.9659
G	CW	0.9546	0.9771	0.9896	0.9677
	NCW	0.9568	0.9592	0.9858	0.9653
H	CW	0.9556	0.9731	0.9844	0.9671
	NCW	0.9595	0.9613	0.9819	0.9647

表7 手掛かり語タグと抽出対象の隣接割合

Table 7 Percentage of adjacencies between clue word tags and extracted targets

Clue word tag	Number of tags	Number of target tag neighbors	Ratio
limitation	4,935	3,058	61.96%
selection	1,022	22	2.15%

5.1.2 抽出対象タグ以外の結果

前節で用いたMicro-F1-scoreが一番高かったパターンCの組み合わせについて抽出対象タグ以外のF1-scoreを比較した(**表8**)。前節で述べたように, 手掛かり語タグを導入することで, 抽出対象のatom, fig_LL, fig_ULの抽出性能は向上し, 特にfig_LLは有意差があったが, formulaタグとsumタグは, 大きくF1-scoreを落とす結果となった。手掛かり語タグを導入することで, 手掛かり語タグの抽出対象でない他のタグの抽出精度に影響を及ぼすことがわかる。Formulaタグは, 構成物を含んだ複合語であり, sumタグは, 抽出対象の語と距離のある手掛かり語タグである。一方, 手掛かり語タグを含めたデータセットにおいて抽出精度が上がったのは, variableとbalanceである。variableは, 構成物を変数として表記したもので, atomと混同しやすい。また, balanceは, 構成物の比率のうち, 明記された比率以外の構成物であるという事象を示すタグである。いずれも, 構成物と構成比率の記述近傍における表記であり, 意味的関連性も多少有している。構成物と比率を得るという当初の目的の範囲では, 手掛かり語タグが高い性能を確保できた。しかしながら, 直接的に手掛かり語タグに関与しないタグでは, 性能が低くなるタグがあることを確認できた。

手掛かり語タグに関与しないタグの抽出性能低下は, 本研究におけるCRFでは, 前後1つもしくは2つの語句により, 関係性のみを記述していること, また, 複合語については, 形態素解析の影響も考慮する必要がある。また, 他のタグとの関連性が低いタグの抽出については, 近傍の素性の影響だけではなく, より文脈に近い特徴量を採用できるBERT-CRFのような手法の方が有利である可能性も考えられる。

表8 パターンCの結果

Table 8 Results for pattern C

	NCW: NotClueWords	CW: ClueWords
atom	0.9610	0.9668
balance	0.8539	0.8764
fig_LL	0.9730	0.9865
fig_UL	0.9922	0.9935
formula	0.9411	0.8235
sum	0.8750	0.7500
unit	0.9904	0.9904
use	0.9890	0.9890
variable	0.7346	0.8076

5.2 RQ2:手掛かり語タグの追加は全体性能に影響するか

結果を**表6**に示す。全体性能では, パターンCのCW: ClueWordsが, Micro-F1-scoreが一番高かった。一方で, 検定(素性群)および検定(学習比率群)では, 有意差は見られなかった。以上の結果から, 全体性能では, 手掛かり語タグを加えたとしても, 大きな性能劣化はなく, 検定による結果からも有意差はみられなかった。したがって, 手掛かり語タグを新たに固有表現として加えたとしても, 特に大きな性能劣化を起こす可能性は低いと考えられる。

5.3 RQ3:手掛かり語タグにより有益な手掛かり語が獲得できるか

5.3.1 手掛かり語タグlimitationの手掛かり語獲得の結果

データセット(OnlyClueTag)では, 有益な手掛かり語は獲得できなかったが, データセット(All)では, 有益な手掛かり語が獲得できた。新たに手掛かり語として認識された表現は「並びに」と「002」であった。単語「002」は誤抽出であったため, 結果として有益な手掛かり語として認識された表現は1種類であった。素性Dの学習比

率が90%から40%において,「並びに」がlimitationの手掛かり語として獲得できた。

予測結果から,「並びに」は「12質量%のCr」と「35質量%以下のAl, Si, Y」の間にあり,元素の含有量と単位を含む表現の間にlimitationとして,付与されていた。この結果は,当初想定していた,元素の含有量以外にも,元素の含有量と単位を含む表現も手掛かり語タグlimitationにより,抽出可能なことを示唆している。

5.3.2 手掛かり語タグselectionの手掛かり語獲得の結果

表9では,二つのデータセット(OnlyClueTag, All)で使われる学習データの中で「selection」のタグが付与されておらず,テストデータで「selection」のタグが付与された単語の一覧を示す。新たに手掛かり語として認識された表現の数は18種類であった。そのうち,selectionの定義を踏まえ,有益であると思われる単語は太字とした。したがって,有益な手掛かり語として,selectionでは,16種類の手掛かり語を新たに獲得することができた。

続いて,データセット(OnlyClueTag)において,新たに獲得した「下記」,「示す」の獲得事例を述べる。予測結果から,「下記に示す群から選択される1種以上の元素を含有」がselectionとして付与された。これまで,予測は請求項の一文単位で実施していた。一方で,予測結果では,selectionが付与された文章の直後の文章に記載されている元素名を関連付けていた。この結果,次の文章に記載された抽出対象に対する関係性もselectionにより識別可能であり,グループとして抽出可能なことを確認した。文章間における抽出対象を関連づけることが可能となれば,特許調査の更なる効率化が期待できる。文書を跨ぐ関連付けは,今後の評価課題としたい。

6. 結 言

本研究では,材料科学技術に関する特許文書からの固有表現抽出を行った。具体的には,材料組成に関する情報抽出に着目した。材料組成を示す,材料を構成する元素とその比率範囲を示す数字を抽出するためには,元素名や数字を単独で抽出するのではなく,特定の元素に対する比率範囲を一つのグループとして抽出する必要がある。そこで,本研究では,元素やその比率範囲を表す数字の他に,手掛かり語タグを導入し,元素とその比率範囲を一つのグループとして,まとめて抽出する手法を提案した。しかしながら,抽出精度向上のためには,手掛かり語タグの導入が抽出対象以外の抽出精度を低下させる影響を抑制する必要がある。そこで,リサーチクエスチョン(RQ)を設定し,その影響評価を実施した。その結果,手掛かり語タグが抽出対象の精度向上に寄与(RQ1)し,特に,比率下限を示す数値であるfig_LLの抽出精度が有意に向上することを確認した。また,手掛かり語タグとして,固有表現のクラスを新たに追加したとしても,大きな性能劣化は示さなかった(RQ2)。さらに,予め設定した手掛かり語タグが付与された手掛かり語だけでなく,新たな有益な手掛かり語が獲得(RQ3)できた。

以上の結果から,本研究では,手掛かり語タグの導入により,効率的に元素とその比率範囲を抽出することが可能であることがわかった。さらに,手掛かり語タグにより,新たな手掛かり語が獲得可能であることを示した。今後は,獲得した手掛かり語の自動精査方法についても検討を行う予定である。本研究により,材料科学技術における特許文書からの固有表現抽出の精度と効率性が向上し,より正確な情報抽出となると考えられる。また,元素とその比率範囲だけでなく,各種特性とその数値範囲といった技術情報の抽出に広く応用可能であり,蓄積された技術文書の活用への貢献が期待できる。

※D2Materiは株式会社プロテリアルの商標です。

表9 新たにselectionが付与された単語
Table 9 Newly assigned word selection

示す(indicates)	満たさ(fulfillment)	若しくは(or)	任意(any)
組み合わせ(combination)	必ず(always)	置換(substitution)	下記(below)
場合(case)	単独(independence)	一つ(one)	成る(consist of)
含ま(include)	それぞれ(respectively)	含み(tone)	もしくは(or)
並びに(as well as)	過剰(excess)		

引用文献

1）特許庁：令和元年度特許出願技術動向調査結果概要マテリアルズ・インフォマティクス, 特許庁（オンライン）, 入手先<https://www.jpo.go.jp/resources/report/gidou-houkoku/tokkyo/document/index/201907.pdf>（参照 2023-05-24）.
2）黒川,他：マテリアルズインフォマティクス技術を活用した材料開発：環境経営に寄与するChemicals Informatics, 日立評論, Vol.104（2022）, No.2, p.249-254.
3）研究開発戦略センター（CRDS）：材料創製技術を革新するプロセス科学基盤 ～プロセス・インフォマティクス～, 研究開発戦略センター（CRDS）（オンライン）, 入手先<https://www.jst.go.jp/crds/report/CRDS-FY2021-SP-01.html>（参照 2023-05-25）.
4）Kononova, et al.: Opportunities and challenges of text mining in materials research, iScience, Vol.24 (2021), No.3, p.102155.
5）Wei, et al.: Machine learning in materials science, InfoMat, Vol.1 (2019), No.3, p.338-358.
6）Olivetti, et al.: Data-driven materials research enabled by natural language processing and information extraction, Applied Physics Reviews, Vol.7 (2020), No.4, p.41317.
7）Li, et al.: A Survey on Deep Learning for Named Entity Recognition, IEEE Transactions on Knowledge and Data Engineering, Vol.34 (2022), No.1, p.50-70.
8）Yadav, et al.: A Survey on Recent Advances in Named Entity Recognition from Deep Learning models, Proceedings of the 27th International Conference on Computational Linguistics, (2018), p.2145-2158.
9）Ratinov, et al.: Design Challenges and Misconceptions in Named Entity Recognition, Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009), (2009), p.147-155.
10) Sekine, et al.: Extended Named Entity Hierarchy, Proceedings of the Third International Conference on Language Resources and Evaluation（LREC'02), (2002), p.1818-1824.
11) Seon, et al.：Named Entity Recognition using Machine Learning Methods and PatternSelection Rules., Proceedings of the NLPRS, (2001), p. 229-236.
12) Jie, et al.: Efficient Dependency-Guided Named Entity Recognition, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, (2017), p.3457-3465.
13) Draxl, et al.: NOMAD: The FAIR concept for big data-driven materials science, MRS Bulletin, Vol.43 (2018), No.9, p.676-682.
14) Chen, et al.: A Critical Review of Machine Learning of Energy Materials, Advanced Energy Materials, Vol.10 (2020), No.8, p.1903242.
15) Weston, et al.：Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature, Journal of Chemical Information and Modeling, Vol.59 (2019), No.9, p.3692-3702.
16) Kuniyoshi, et al.: Analyzing research trends in inorganic materials literature using nlp, Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, (2021), p.319-334.
17) Jensen, et al.: A machine learning approach to zeolite synthesis enabled by automatic literature data extraction, ACS Central Science, (2019), p.892-899.
18) Court, et al.: Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning, npj Computational Materials, Vol.6 (2020), No.1, p.18.
19) Tshitoyan, et al.: Unsupervised word embeddings capture latent knowledge from materials science literature, Nature, Vol.571 (2019), p.95-98.
20) Augenstein, et al.: SemEval 2017 Task 10: ScienceIE - Extracting Keyphrases and Relations from Scientific Publications, Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), (2017), p.546-555.
21) Chen, et al.: Application of NER and Association Rules to Traditional Chinese Medicine Patent Mining, 2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom)and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), (2020), p.767-772.
22) Pantel, et al.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, (2006), p.113-120.
23) Sumida, et al.: Concept-Instance Relation Extraction from Simple Noun Sequences Using a FullText Search Engine, Proceedings of the ISWC 2006 workshop on Web Content Mining with Human Language Technologies (WebConMine), (2006), p.442-447.
24) Etzioni, et al.：Unsupervised named-entity extraction from the Web: An experimental study, Artificial Intelligence, Vol.165 (2005), No.1, p.91-134.
25) 坂地, 他：Cross-Bootstrapping:特許文書からの課題・効果表現対の自動抽出手法, 情報爆発論文, Vol.93 (2010), No.6, p.742-755.
26) Vaswani, et al.: Attention is all you need, Advances in neural information processing systems, Vol.30 (2017).
27) Devlin, et al.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), (2019), p.4171-4186.
28) OpenAI: GPT-4 Technical Report, arXiv preprint

arXiv: 2303.08774, (2023).
29) Hu, et al.: Zero-shot clinical entity recognition using chatgpt, arXiv preprint arXiv: 2303. 16416, (2023).
30) Alsentzer, et al.: Publicly Available Clinical BERT Embeddings, Proceedings of the 2nd Clinical Natural Language Processing Workshop, (2019), p.72-78.
31) Polak, et al.: Flexible, Model-Agnostic Method for Materials Data Extraction from Text Using General Purpose Language Models, arXiv preprint arXiv: 2302. 04914, (2023).
32) Polak, et al.: Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering-Example of ChatGPT, arXiv preprint arXiv: 2303. 05352, (2023).
33) Tjong, et al.: Representing Text Chunks, Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics, (1999), p.173-179.
34) Lafferty, et al.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, Proceedings of the Eighteenth International Conference on Machine Learning, ICML'01, (2001), p.282-289.
35) 奥村, 他：言語処理のための機械学習入門, コロナ社, (2010).

酒井　敏彦
Toshihiko Sakai
九州大学大学院システム情報科学府

千綿　伸彦
Nobuhiko Chiwata
株式会社プロテリアル
研究開発本部グローバル技術革新センター
兼　知的財産部 IP ソリューショングループ
博士（工学）

峯　恒憲
Tsunenori Mine
九州大学大学院システム情報科学研究院
博士（工学）