様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ６ 月 １０ 日現在

機関番号：１０１０１
研究種目：挑戦的萌芽研究
研究期間：2014～2015
課題番号：２６５４０１６５

研究課題名（和文）ナノ結晶デバイス開発論文からの情報抽出とその活用

研究課題名（英文）Information Extraction from Nanocrystal device development papers and its utilization

研究代表者

吉岡　真治（Yoshioka, Masaharu）

北海道大学・情報科学研究科・准教授

研究者番号：４０２９０８７９

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：本研究では、ナノ結晶デバイス開発論文を対象として、主に論文中の実験に関する条件や結果に関わる情報(材料や関係するパラメータなど)を抽出するためのコーパスの作成と、そのコーパスを利用した情報抽出システムNaDevの開発を行った。本システムを用いることにより、論文、アブストラクト、図表のキャプションから、単純なキーワードではない実験に関係するキーワードを抽出することができ、論文の類似性判定、抽出した情報を利用した検索の絞り込み支援システムなどを作成することが可能となる。本研究では、プロトタイプシステムとしての図表検索システムの提案を行った。

研究成果の概要（英文）：In this project, we have constructed a corpus for extracting useful information related to the experiment of nanocrystal device development (e.g., Source material and parameters) and developed an automatic information extraction system NaDev that tries to extract such information from papers. By using NaDev, we can extract characteristic keywords related to the experiment (e.g., source material, experimental parameter) and these extracted keywords are used for calculating similarity among different papers and information retrieval system that have query construction support. In this project, we proposed a figure retrieval system that uses such parameter information as query.

研究分野：知識科学

キーワード：知識工学　知識マネージメント　コーパス　ナノデバイス

１．研究開始当初の背景

ナノインフォマティックスとは、ナノスケールの科学やデバイス開発などに役立つ情報を有効に活用することを目指した分野である。2011 年に米国の National Nanomanufacturing Network がまとめた Nanoinformatics 2020 Roadmap [1]では、実験データの標準化による共有・データの可視化やマイニングなどに加えて、 論文からのテキストマイニング・情報抽出が大きな研究テーマとして提案されている。しかし、このロードマップで提案されている研究テーマとしては、ナノ創薬の分野で、バイオインフォマティックスと同様の手法を用いるものがあるだけで、ナノ結晶デバイス開発に関する情報抽出の研究は行われていない。

２．研究の目的

本研究では、ナノ結晶デバイス開発論文から、 デバイス開発に関する背景情報(最終製品、評価基準や実験の詳細(材料や重要なパラメータ)を情報抽出)することにより、パラメータ間の依存関係の分析や事例の類似性評価に役立てることをその目標とする。

３．研究の方法

本研究では、 これまでに開発してきたナノ結晶デバイス開発論文のための情報抽出の研究を発展させ、情報知識処理の研究者(吉岡)によるシステムの開発と、ナノ結晶デバイスの研究者(原,Newton（研究協力者：英 Southampton 大))による評価のフィードバックを繰り返すことにより、実際のナノ結晶デバイス開発者にとって有用な論文の活用方法を目指した。

具体的には、論文中から抽出すべき情報とそのタイプを定義したコーパス作成のためのガイドラインを作成するとともに、そのガイドラインを適用したコーパスをナノ結晶デバイスの研究者の協力を受けて作成し、そのコーパスを用いた論文からの情報抽出システムの作成を行う。このシステムでは、化学物質名やパラメータに用いられる物理量のリストなどの情報を利用することにより、コーパスの小ささを補う方法を用いる。

また、抽出した結果の活用方法として、 論文の類似性判定とグラフを利用したパラメータ間の関係の抽出と表示の手法を開発し、論文のサーベイ支援の枠組を構築する。

４．研究成果

（１） ナノ結晶デバイス論文からの情報抽出用のコーパスの作成（雑誌論文１）

本研究では、ナノ結晶デバイスの研究者から、適切なフィードバックを得ながらコーパスの作成を行うために、ナノ結晶デバイスの研究者である原を研究分担者に迎えて、コーパスの作成を行った。コーパスの対象とする文章としては、対象のバリエーションを重視して、アブストラクトを多数使う方法も考えられるが、本研究では、まず、論文中の表現のバリエーションに関する情報を重視して、論文の全文情報を利用することとした。

このコーパスでは、下記のタイプの情報について、アノテーションを行うこととした。

- 材料(SMaterial)：実験に用いる素材や化合物(As や InGaAs など)
- 物の特性(MChar)： 素材や化合物が持つ特性(結晶の異方性に関する情報など)
- 実験パラメータ(ExP)：実験でデバイス作成の制御のために用いるパラメータ(圧力や流量など)
- 実験パラメータの値(ExPVal)：上記の実験パラメータに対応する値
- 評価パラメータ(EvP)：デバイスの性質を評価するためのパラメータ(ピークエネルギーや工学特性における半値幅など)
- 評価パラメータの値(EvPVal)：上記の評価パラメータに対応する値
- デバイスの作成手法(MMethod)：作成手法の名前 (SA-MOVPE など
- 最終製品(TArtifact)： 最終製品の名前(半導体ナノワイヤなど)

このコーパス構築の際には、ナノ結晶デバイスの研究室の修士課程の学生複数名によるコーパスへのアノテーション（文書中の抽出すべき単語とそのタイプの情報をメタデータとして追加）の結果を比較することで、コーパスへのアノテーションのガイドラインを事例の情報とともに精緻化するという作業を進めてきた。

本研究期間では、これまでの研究の結果として得られている学生が作ったコーパスの情報をベースとして、シニアの研究者（研究分担者の原）によるチェックを行い、論文 5 編の全文情報からなる最終版のコーパス NaDev コーパスを作成した。

具体的には、これまでの学生によるアノテーション結果を、複数名の学生が一貫してアノテーションをした部分と、議論があった部分の 2 種類に整理した。共通したアノテーションが行われた部分については、そのアノテーションが正しいかどうかを確認するとともに、問題があればアノテーションを変更する。議論があった部分については、お互いのアノテーション結果を示し、適切なものを選ぶ、あるいは、新たにアノテーションをしなおすという形で作業を行った。

この結果、幾つかのガイドライン上の修正が行われるとともに、コーパスの最終版を確定した。この過程で、コーパス中の論文に、主に、デバイスの生成を重視した論文と、出来上がったデバイスの性質の分析を重視した論文があり、この論文のタイプの違いが、表現のバリエーションの違いを生み、学生によるアノテーションの品質が違うという事象が観察された。具体的には、前者の主にデバイスの生成を目的とした論文では、ガイド

ラインの修正による影響を除くと、2 名の学生によるアノテーションが一貫している場合のアノテーションの性能は、精度が 0.99 で、再現率が 0.96 と非常に高いレベルであったのに対し、分析を重視した論文では、精度が 0.97 であったものの、再現率が 0.81 と十分なものではなかった。これは、ガイドライン作成の際のディスカッションが、主に、デバイスの生成を重視した論文の事例で行われており、分析を重視した論文に関する議論が不十分であったためと考えられる。シニアの研究者からの指摘を受けて、ガイドラインを修正や迷いやすいアノテーション事例の収集を行ったこともあり、次回以降の学生によるアノテーションでは、もう少し、学生によるアノテーションの性能がよくなることが期待される。

現在、このコーパスのガイドラインは、テクニカルリポートとして公開するとともに、問い合わせに応じて、コーパスを提供する準備をしている。

（2） ナノ結晶デバイス論文からの自動情報抽出システム（雑誌論文２）

本研究では、（1）で作成した NaDev コーパスにおける情報抽出基準に基づいて、論文などのナノ結晶デバイス開発のための文書から情報を抽出するシステム NaDevEx を作成した。このシステムでは、固有名抽出やバイオインフォマティックスなどの分野で用いられるテキストからの系列ラベリングの問題として、情報抽出の問題を定式化した。

系列ラベリングによる情報抽出とは、入力として与えられたテキストを、形態素解析などのツールを用いて、単語（形態素）の系列として分割し、その系列の中のどの部分が抽出すべき情報に対応するかを抽出する。単語に関する情報としては、一般的には、文字列、標準形、品詞、形態素といった文法情報、形態（英語の場合、大文字から始まるか、数字、記号など）といった言語特有の情報、領域知識として与えられた辞書とのマッチングの結果などが用いられる。

NaDevEx では、標準形、品詞、形態素を文法情報として、大文字、数字・記号などの分類を形態の情報として利用するとともに、主に物理量を中心としたパラメータ辞書とのマッチング結果、材料としてよく用いられる化学物質の情報抽出システムの適用結果を単語に関する情報とした系列ラベリングを CRF（Conditional Random Field）を用いて行った。また、抽出する情報間に依存関係があること（たとえば、材料に対応する単語を含む形でパラメータの情報が定義される）を考慮して、全てのタイプの情報を、一つの系列ラベリングにより決定するのではなく、先に述べた単語の依存関係を考慮した逐次的な情報抽出を行う。具体的には、他のタイプの単語を単語中に含むタイプの情報については、先に、依存するタイプの情報抽出を行う。他のタイプの単語を含む情報のタイプについては、依存するタイプの情報抽出の結果も系列ラベリングのための情報として利用する。

NaDevEx の情報抽出の性能をはかるために、NaDev コーパスを用いた 5 分割交差検定（4 本の論文をトレーニングデータとして、残りの 1 本のデータの評価を行う）を行ったところ、精度 0.95。再現率 0.74 という結果となった。特に、材料については、精度 0.98,再現率 0.97 と非常に良い性能であったが、一方、評価パラメータについては、精度 0.86、再現率 0.60 とあまり良くない性能となった。一つの原因は、コーパスの作成時にも述べたデバイスの生成を重視した論文と分析を重視した論文の違いにあり、分析を重視した論文では、全体の精度が 0.8、再現率が 0.51 と非常に低い結果となり、特に、評価パラメータは、精度が 0.77、再現率が 0.47 と非常に低い値となっている。これは、４つの生成を重視した論文をトレーニングデータとして、タイプの違う評価を重視した論文をテストデータとしているため、評価を重視した論文にのみ存在するような表現をうまく抽出できなかったためと考えられる。今後は、コーパスをバランス良く作成していくとともに、パラメータリストの拡充を図ることで、全体の抽出性能の向上を図りたいと考えている。

（3） NaDevEx の抽出結果に基づく図表検索システム（学会発表 3）

ナノ結晶デバイスの作成のためには、デバイスの持つ構造を設計するだけではなく、そのデバイスの構造を製造するための適切な条件を決定する必要がある。しかし、どの様なパラメータが最終的なデバイスの性能に大きく影響を与えるのかについては、実験の初期段階では必ずしも確定しておらず、実際の実験中の試行錯誤により、有用なパラメータ間の関係について検討を行い、その結果が、論文の中にまとめられるという状況が存在している。そのため、論文中に存在するグラフに現れるパラメータ間の関係に関する情報は、類似した実験を行う際にどの様なパラメータについて考慮した方が良いのかといった有用な知識となることが期待される。

また、ある程度、関係しそうなパラメータが思い浮かぶ場合においても、これらの関係を表現したグラフの多くを閲覧し、通常と異なった関係を持つようなグラフを発見することは、パラメータ間の関係をコントロールするような別の実験条件の発見へとつながることが期待される。

本研究では、このようなナノ結晶デバイスの開発者に対して、論文のグラフの検索を支援すると共に、関連するパラメータの情報などを提供することができる図表検索システムのプロトタイプシステムを作成した。

論文中の図表を検索するシステムとしては、Elsevier の Science Direct

(http://www.sciencedirect.com)などの論文検索システムにおいても、キャプションを対象とした検索システムが提供されている。ここでは Science Direct の図表の検索システムを例にとり、その一般的な機能について説明する。この検索システムでは、キャプション中の文字列に加え、論文タイトル、著者といった書誌情報に関するメタデータを利用した検索が行えると共に、検索結果から掲載されている論文誌や発表年度といった主に書誌情報から得られるようなメタデータの集約結果が表示され、さらなる図表の絞り混みを支援することが可能になっている。

本研究では、このような図表の検索システムとは異なり、ナノ結晶デバイス開発論文に代表されるような物性系の実験などに関する図表を対象とした検索システムの構築を目的としている。具体的には、グラフのキャプションに記述されることが多い、パラメータ間の関係や、実験条件に関連するような材料や手法の名前といったメタデータを抽出し、Science Direct などで提供する書誌情報に関するメタデータではなく、実験条件に関するメタデータの集約結果を表示することにより、関係するパラメータの発見や実験条件に関する知見を支援するグラフイメージ検索システムの提案を行う。

その実現方法としては、図表のキャプションを対象として、NaDevEx を用いることにより、実験条件に関連する情報のタグ付けを行い、その結果をメタデータとして利用する。ユーザは、メタデータのタイプを考慮した検索を行う事ができるだけでなく、検索結果に表示される他のメタデータの集約結果を見ることにより、目的のパラメータと同時に現れることの多い他のパラメータのリストなどを見ることが可能となり、ナノ結晶デバイス開発における利用方法に即した検索が実現できると考えている。

我々は、このプロトタイプを Creative commons のライセンスで公開されているナノ分野のオンラインジャーナルである Beilstein Journal of Nanotechnology から、集めた図 6,133 枚に対してカラムストア型データベースであり、全文検索の機能を持つ groonga を用いてデータベースを利用して作成した。本来は、全ての図ではなく、グラフイメージのみに限定してデータベースを構築することを考えていたが、現時点では、グラフイメージとそれ以外を区別するための処理が準備できていなかったため、全ての図を用いてデータベースを作っている。

各々の図に対しては、以下のデータが付与されている。

- キャプション：図のキャプション(テキスト)
- 要旨：論文の要旨(テキスト)
- タイトル：論文のタイトル(テキスト)
- メタデータ：(1) で述べた 8 種類のタイプの単語の各々について、その抽出された語のリスト

また、図１にプロトタイプシステムのスクリーンショットを示す。

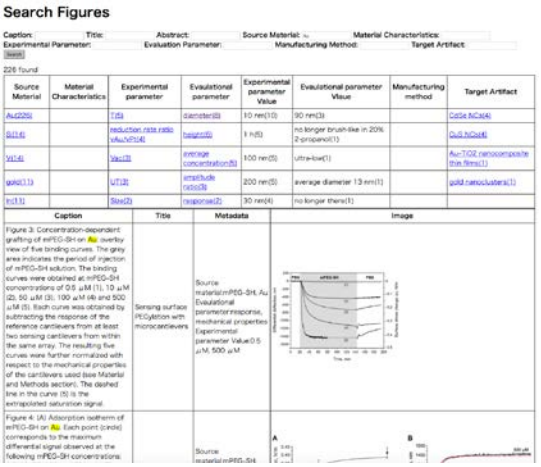

図１：グラフイメージ検索システム

本システムでは、上部の検索条件の設定部分においてキャプション、タイトル、アブストラクトに加え、メタデータに対応する条件を設定して、図のデータの検索を行うことができる。また、検索結果について、全てのメタデータについて集約操作を行うことにより、検索結果群に含まれるパラメータ名のリストがサマリーとして表示される。また、ユーザは、これらのキーワードをクリックすることで、検索条件に追加することにより、検索結果の絞り込みを行うことができる。

検索結果としては、図を表示するだけでなく、キャプション、タイトル、メタデータの情報をあわせて表示すると共に、検索キーワードのハイライトを行っている。また、グラフをクリックすることで、アブストラクトの情報も含む全ての情報を閲覧することが可能となる。

本システムを、共著者の一人であるナノ結晶デバイス研究者(原)に提供し、簡単な検索を行ってもらい、システムの利点ならびに問題点に関する議論を行った。この評価の際には、より、実際の利用状況に近づけるため、共著者の興味を反映して集められた論文 916 件から作成した 3142 枚の図から構成されるデータベースを作成した。このようなデータの構築については、著作権的な問題があるため、研究グループが持っている論文データなどから作る方法についても並行して検討している。

以下に、指摘された主な利点と問題点を列挙する。

- 利点
  - 目的や構造などのキーワードから出発して、表示される関連するパラメータなどを用いてグラフを絞り混んでいくという考え方は面白い。
  - パラメータを2つキーワードに与えて、それらの関係を示したグラフをまとめてみるのは面白い。

- 問題点
  - 適切なメタデータの付与が必要
    検索結果をみたところ、本来パラメータなどのメタデータが付与されるべき単語にメタデータが付与されていない事があるため、絞り混みの情報として不十分である。
  - 同義語や複合語の取扱い
    量子ドットのように、Quantum dots, QD, QDs などで表される概念について、取りまとめをして欲しい。また、メタデータを付与する単語の単位の問題から、the gas temperature と、temperature of the gas などをまとめることができないので、これについても検討が必要である。

（4） 複数の化学物質名抽出システムを組み合わせたアンサンブルラーニングによる化学物質名抽出システムの構築（雑誌論文３）

化学物質名の抽出は、バイオインフォマティックスや他の分野にも応用可能なこともあり、ChemSpot や OSCAR4 といった様々なツールが公開されている。これらのツールは、初期段階においては、単純な化学物質を抽出することを目的としたコーパスである SCAI コーパスをベースに開発されてきたが、近年では、化学物質の一種である薬の名前も含めて認識するような、より複雑な CHEMDNER Task（Chemical compound and drug name recognition task: 化学物質と薬の名前の認識タスク）への応用も考えられている。

しかし、初期のツールにおいては、必ずしもこれらの名前を認識するのに十分なルールや機械学習のためのトレーニングデータなどが提供されていないため、CHEMDNER タスクにおいては、十分な性能を実現することができなかった。また、機械学習をベースにしたシステムとルールベースのシステムにおいては、認識が得意なデータが異なり、それぞれのツールの特性をうまく認識して組み合わせることにより、全体のシステムの性能を向上させることができると考えた。

具体的には、化学物質の名前の抽出システムを NaDevEx と同様に、系列ラベリングのタスクとして定式化し、主にパターンとのマッチングに基づいたシステムである OSCAR4 と機械学習を組み合わせた ChemSpot の二つの特徴の違うシステムの認識結果を単語に関する情報として利用するアンサンブルラーニングアプローチを用いた化学物質名抽出システムを構築した。本システムでは、主にタスクの違いなどに起因するような各々のシステムの一貫した間違いの情報なども利用した学習を行うことが可能となるため、単純な複数システムの多数決などとは異なる形でのシステムの出力を取りまとめることが可能となる。このような考え方は、一種のドメイン適合とも考えられる。

この複数システムの結果を取りまとめる際の問題としては、テキスト分割の一貫性をうまく扱う必要があることが確認された。本研究では、通常の文書解析システムによる単語分割の結果を ChemSport や OSCAR4 の化学物質の認識結果に矛盾しないように、さらに詳細に分解することにより、この問題を解決することとした。

適切なテキスト分割を行った結果、アンサンブルラーニングに基づくシステムは、精度 0.87、再現率が 0.79 という性能となり、一つのシステムの結果のみを用いた場合（Chemspot の出力のみ：精度 0.85、再現率 0.77、OSCAR4 の出力のみ：精度 0.86、再現率 0.76）と比較して、精度、再現率が共に向上することを確認した。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

1. Thaer M. Dieb, Masaharu Yoshioka, and Shinjiroh Hara : NaDev: An Annotated Corpus to Support Information Extraction from Research Papers on Nanocrystal Devices. Journal of Information Processing, Vol. 24, No. 3, pp. 554-564, 2016. （査読有）
2. Thaer M. Dieb, Masaharu Yoshioka, Shinjiroh Hara, and Marcus C. Newton : Framework for automatic information extraction from research papers on nanocrystal devices. Beilstein Journal of Nanotechnology, Vol. 6, pp. 1872-1882, 2015. （査読有）
3. Thaer M. Dieb and Masaharu Yoshioka: Extraction of Chemical and Drug Named Entities by Ensemble Learning Using Chemical NER Tools Based on Different Extraction Guidelines. Transactions on Machine Learning and Data Mining, Vol. 8, No. 2,2015. （査読有）
4. Martin Krallinger 他 Masaharu Yoshioka (53 名中 34 番目): The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of Cheminformatics, Vol. 7, No. Suppl 1, pp. S2, 2015.
5. Thaer M. Dieb, Masaharu Yoshioka, Shinjiroh Hara, and Marcus C. Newton : Automatic Annotation of Parameters from Nanodevice Development Research Papers. In Proceedings of the 4th International Workshop on Computational Terminology (Computerm), pp. 77-85, 2014. （査読有）

〔学会発表〕（計 3 件）

1. 朱 濤、Thaer M. Dieb、吉岡 真治、原 真二郎：ナノ知識探索プロジェクト：実験記録からの知識発見(第4報) -キャプションからのメタデータの自動抽出によるグラフイメージ検索システム-, 2016年度人工知能学会全国大会, 1J2-4, 北九州国際会議場, 北九州市, 2016 年 6 月 6-9 日.
2. Masaharu Yoshioka, Thaer Dieb, Shinjiroh Hara: Automatic Information Extraction of Experiments from Nanocrystal devices Development Papers, Nanoinformatics 2015 Workshop, Arlington, VA, USA, Jan 26-28, 2015.
3. Masaharu Yoshioka: Supporting design and engineering by physical concept ontologies, TMCE2014 Academic Workshop (Invited), Budapest, Hungary. May 19-23, 2014.

〔図書〕(計 0 件)
〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）
〔その他〕
ホームページ等
Knowledge Exploratory Project for Nanocrystal Device Development
http://nanoinfo.ist.hokudai.ac.jp

６．研究組織

(1)研究代表者

吉岡 真治 (YOSHIOKA, Masaharu)
北海道大学大学院・情報科学研究科・准教授
研究者番号：40290879

(2)研究分担者

原 真二郎 (HARA, Shinjiroh)
北海道大学大学院・量子集積エレクトロニクスセンター・准教授
研究者番号： 50374616

(3)連携研究者