HOKKAIDO UNIVERSITY

Title	Framework for Experimental Information Extraction from Research Papers to Support Nanocrystal Device Development [an abstract of dissertation and a summary of dissertation review]
Author(s)	Moustafa Dieb, Thaer
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第12046号
Issue Date	2015-12-25
Doc URL	https://hdl.handle.net/2115/60482
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	doctoral thesis
File Information	Moustafa_Dieb,_Thaer_review.pdf, 審査の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　氏名　Moustafa Dieb, Thaer

審査担当者　主　査　准教授　吉岡 真治
副　査　教　授　原口 誠
副　査　教　授　有村 博紀

学位論文題名

Framework for Experimental Information Extraction from Research Papers to Support Nanocrystal Device Development
（ナノ結晶デバイス開発支援のための論文からの実験情報抽出フレームワーク）

現在、ナノスケールのデバイスの開発を情報科学の技術を用いる事により支援するナノインフォマティックスの研究が盛んに行われている。この研究分野では、第一原理などに基づく計算機シミュレーション、ポータルサイトなどを用いたデータの共有などの研究に加え、大量に存在する論文からの情報獲得の研究がその柱として考えられている。論文からの情報獲得の研究としては、2000 年代の初頭頃から、バイオインフォマティックスの分野などを中心に、トレーニングデータであるコーパスを用いた機械学習を利用した方法が提案され、多くの成果を出しつつある。しかし、このような機械学習の枠組みを利用するためには、論文からどのような情報を抽出するのかを定義し、基盤となるトレーニングデータであるコーパスを作成することが不可欠である。そのため、ナノインフォマティックスの分野においては、その重要性が認識されているにも関わらず、バイオインフォマティックスに近いナノ製薬や毒物の分野を除いて、目立った成果がないという状況であった。

本研究では、ナノ結晶デバイス開発という分野を対象に、論文中から抽出すべき情報の定義とその定義に基づいた信頼性の高いコーパスの作成を行うとともに、そのコーパスを用いた情報の自動抽出システムの提案を行っている。この研究は、異分野融合の研究であり、ナノ結晶デバイス開発の研究者が行っている研究活動を、情報科学・知識工学の観点から分析し、その結果として、どのような情報を論文から抽出すべきなのかを決定すると共に、抽出すべき情報を一貫して抽出するためのコーパス作成のためのガイドラインの整備を行い、コーパス NaDev を作成した。本コーパスについては、ナノ結晶デバイス開発の研究者により、その妥当性を検証されている。また、さらに、このコーパスを利用した情報抽出システム NaDevEx を提案し、NaDev コーパスでの評価により、その有用性を確認した。

これにより、本研究が作成した NaDev コーパスにより、様々な情報抽出システムの作成ならびに評価が可能になるとともに、情報抽出システム NaDevEx を用いた論文からのナノ結晶デバイス開発論文からの情報抽出を実現することが可能になった。この情報抽出結果を利用することにより、材料、実験パラメータなどの観点の重要性を考慮した柔軟な類似度の定義を行うことにより、新しい、論文検索システムが作成できるという可能性を示している。その他にも、ポータルサイトなどがデータを収集する際の前処理としての利用など、様々な応用が期待される。本研究における成果は、以下の 3 点である。

1. ナノ結晶デバイス開発論文からの情報抽出に有用なコーパス NaDev の作成
2. NaDev コーパスに基づく論文からの情報抽出システム NaDevEx の作成
3. 論文からの化学物質名、抽出における複数の化学物質名抽出ツールのアンサンブル学習による利用方法の提案

本論文の構成は以下の通りである。第 1 章では、本研究の背景や動機について述べ、第 2 章では、ナノインフォマティックスにおける論文からの情報抽出の研究の動向とバイオインフォマティックスなどにおけるその情報抽出に関する研究の動向を紹介し、本研究の位置付けを述べている。第 3 章では、NaDev コーパスを作成するために行った、コーパス作成実験とその評価について述べ、ドメインの学生による基本作業とドメインのエキスパートの研究者による評価という 2 段階のコーパス作成の枠組みを提案している。本コーパスは、エキスパートの研究者の評価を受けた公開版を作成し、Web を通じて公開中である。第 4 章では、NaDev コーパスを基盤となるトレーニングデータとして用いることで、論文から新しい情報を抽出することが可能なシステム NaDevEx を作成した。また、その有用性を NaDev コーパスを用いて、実験的に検証した。さらに、化学物質名の抽出においては、様々なツールが公開されているが、これらのシステムを個別に使うのではなく、その結果を機械学習におけるアンサンブル学習の枠組みでまとめあげることが、特定ドメインの化学物質の抽出には有用であることも、実験的に確認した。5 章では、抽出した情報の利用方法に関する予備的検討結果を報告し、6 章で、全体のまとめを行っている。

これを要するに、筆者は、近年ますますその有用性が認知されてきている論文からの知識発見という分野において、具体的なナノ結晶デバイス開発論文からの情報抽出を実現するための枠組を異分野の研究者と協力して作成し、実際にシステムとして実装することによって、その有効性を示したものであり、知識工学ならびにナノインフォマティックスに対して貢献するところ大なるものがある。よって、筆者は、北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。