専門用語の知識保全エコシステムを有するインハウス論文・図表データベースの構築

吉岡 真治　尹 磊
北海道大学院情報科学研究科
yoshioka@ist.hokudai.ac.jp

原 真二郎
北海道大学量子集積
エレクトロニクス研究センター
hara@rciqe.hokudai.ac.jp

鈴木 晃　高山 英紀　石井 真史
物質・材料研究機構
{SUZUKI.Akira3, TAKAYAMA.Eiki, ISHII.Masashi}@nims.go.jp

1　はじめに

近年、データ駆動型科学などの実現を目指して、論文からの情報抽出、知識発見の研究が盛んに行われているが、専門用語に関する大規模な概念辞書や情報抽出のための訓練コーパスといった情報抽出用知識の作成コストが高いといった問題があり、限定された分野での利用にとどまっている。

本研究では、この問題を解決するために、ユーザである特定分野の研究者に有用な最新の論文を継続的に追加していくと共に、その分析を行うという日々の研究活動に役立つ情報を提供する基本システムを提供した上で、そのシステムで用いる専門用語の拡充や、パターンベースで付与されたタグ付きコーパスの修正を行う枠組を持つ論文データベースを提案する。

本システムでは、ユーザが、専門用語辞書の修正などを行った結果が、直接的に、システムの分析結果に反映させる枠組をつくることで、ユーザが積極的にデータを入力するモチベーションにつなげるという形で、知識保全のサイクル (知識の修正→知識の利用→知識の修正→…) をまわすことが出来る知識保全エコシステム (生態系) の形成に寄与すると考えている。

2　論文・図表データベースの構築

2.1　予備的検討

我々は、これまでに、ナノ結晶デバイス開発の分野を対象とした、論文からの実験情報の抽出を目標としたコーパスの作成 [1] や、そのコーパスを用いた自動情報抽出システムの構築 [2] を行ってきた。また、その結果を用いて、論文中の図表データを多観点分析するシステム [3] の提案を行った。このシステムにおける多観点分析とは、用語を材料、パラメータなどに分類し、分類毎の専門用語の出現頻度などの情報を表示し、検索式の詳細化を支援するための分析手法である。

しかし、コーパスのサイズが十分でなかったため、コーパスに含まれていないような語から構成される用語を見つけられないだけでなく、同じ用語についても、コンテキストに応じて、用語として判断されたりされなかったりするという状況になってしまった。一方で、同義語として集約して取り扱ってもらいたい正式表記、略語、複数形などが、表記のみを手がかりとして集約すると、別々の単語として扱われることに対する問題などが提起された。

この問題を、コーパスを増やす形で対応するためには、より大規模なコーパスを作成する必要が出てくるが、そのアノテーションには、専門的な知識が必要な場合もあり、簡単にはコーパスサイズが増やせないという問題があった。

2.2　用語集を用いたパターンマッチングによる用語抽出

そこで、本研究では、コーパスサイズを増やさずに、様々な用語のバリエーションを扱うための用語辞書を作成し、そのパターンマッチングによる用語抽出を行うことで、検索もれを減らすという方針でシステム開発をすることとした。

このパターンマッチングを行う手法の利点は、次の通りである。

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

- コーパスを作成する場合には、同じタイプの用語について、ある程度、複数の用例を含む形でデータを集める必要があるが、用語辞書の作成は、単純にリストをつくるだけで良い。
- 多くの専門用語は、一般語として使われることが少ないので、文脈を考慮しないパターンマッチングでも、誤抽出の可能性は低い。また、誤抽出の可能性のある語については、ユーザもある程度、気をつけて、結果を解釈することが可能である。
- 用語辞書に同義語のリストを登録することで、同義語の名寄せなどが簡単に行える。

用語辞書が必要とはなるが、機械可読な用語集などを保有している場合には、それが利用可能であるし、一から作成するとしても、コーパスの作成に比べると、圧倒的にコストが低い。

逆に、欠点は、次の通りである。

- 用語リストに登録されていないものが抽出されない。特に、研究の発展と共に、新しい専門用語が生まれた場合などに、対応できない。
- 想定している意味と異なる形で用いられた場合でも、表記に合わせて抽出してしまう。

本研究では、前者の問題に対し、専門用語抽出 [4] の研究成果を用いた用語辞書の拡充支援を行い、後者に対して、ユーザが誤った抽出結果を見つけた場合に、それを修正するインタフェースを提供することにより、抽出結果をより良いものにできるようにする。

ただ、後者の作業は、コーパス作成に近い作業となるため、作業コストは高いと考えられるが、十分な量がたまった場合には、機械学習のためのコーパスとして利用可能となる。ユーザが興味を持つ検索結果については、多く、ユーザの目に触れる可能性が期待できるため、ユーザが興味を持つ用語を含む文については、一定サイズのコーパスが作成されることが期待できる。

2.3 知識保全エコシステムの構築

上記のような枠組を用意しても、ユーザである分野の研究者が利用したいと思うような情報が提供されれない場合には、どのように良いユーザーインターフェースを作成しても、システム自体が使われないと分野の研究者からの知識提供が望まれないこととなる。そのため、分野の研究者が入力したいと思わせるモティベーションを与えるだけでなく、入力した結果として、分野の研究者が得られる情報が改善されるという知識保全エコシステム (生態系) を実現する必要がある。

図 1 に、本研究で実現を目指すエコシステムの枠組を示す。本エコシステムにとって、最も重要な観点は、分野の研究者に、日常の研究活動で利用してもらえることであり、そのためには、ある時点で構築した固定の論文データベースを用いるのでは不十分である。そのため、本研究では、分野の研究者が興味を持つ最新の論文についても継続的に追加し、検索を行えるだけでなく、研究動向の分析にも利用可能な機能を提供することで、日々の研究活動を直接支援する方法を提案する。

具体的には、研究室などを単位とした研究グループを想定し、その研究グループで購読している雑誌、定期的に参加・調査している国際会議などの論文を入手し次第、追加する事が可能なインハウスデータベースとして運用する。この様な特定のグループの興味に応じて網羅的に所蔵することにより、出版者毎にサービスが分かれているような商用の電子図書館とは異なった性格のデータベースとなり、ユーザが利用する動機の一つとなりうると考えている。

また、特定の分野に限って論文を収集することで、最新の論文を元にした専門用語抽出の結果は、分野の新しいキーワードを発見することにつながるといったことが期待できる。また、その新しいキーワードを登録し、再度、用語抽出を行うことにより、その用語の出現頻度の時間遷移などを見ることが可能となるため、研究動向分析にも資すると考えている。

このように直接的に利用者に有用な情報が提供できることが理解されることにより、ユーザが専門用語辞書の拡充などの操作を行う動機付けを行い、その結果、ユーザもより良い分析結果が得られるというサイクルをもつエコシステムが構築できると考えている。

3 プロトタイプシステムの構築

現在、前節で紹介した機能を持つインハウス論文・図表データベースのプロトタイプシステムを構築している。

本システムでは、最新論文の登録を行うために、PDF からのテキスト・図表データの抽出システムを

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

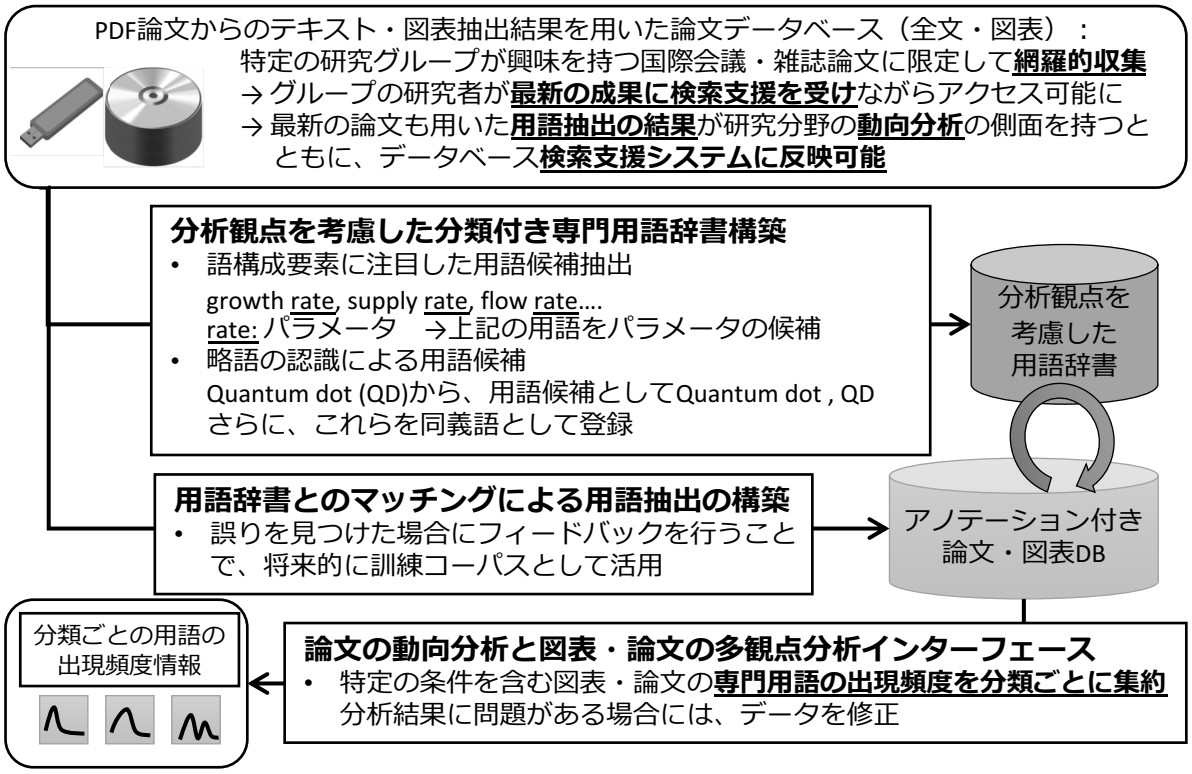

図 1: 本研究でのエコシステムの枠組

利用する。現在、PDFNLT[5] を候補として、試用実験を行っている。

論文抽出後の専門用語抽出のシステムとしては、[4] の考え方に基づく用語抽出システムである termextract[1] を利用する。それ以外にも、パターンベースで作成した略称 (Acronym) の抽出システムなどを組み合わせることで、同義語辞書などの拡張に利用する。

本システムによる用語抽出の結果を用いるアプリケーションとしては、先に紹介した論文中の図表データを多観点分析するシステム [3] を利用する。この際、キャプションについての用語抽出結果については、brat[2] を用いて、アノテーションを編集可能とする。

4 おわりに

現在、プロトタイプシステムを作成すると共に、特定分野の研究者である共著者の研究室である北大の量子集積エレクトロニクス研究センター量子結晶フォトニクス分野と物質・材料研究機構において、本システムのセットアップと予備的な運用を始める予定である。

[1] http://gensen.dl.itc.u-tokyo.ac.jp/pytermextract
[2] http://brat.nlplab.org/

謝辞

国立情報学研究所の相澤教授には、論文 PDF の活用について議論をさせて頂きました。ここに記して、謝意をあらわしたいと思います。また、本研究は、2018 年度国立情報学研究所公募型研究 (自由 24) の助成を受けています。

参考文献

[1] Thaer Moustafa Dieb, Masaharu Yoshioka, and Shinjiroh Hara. 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.

[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.

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.

[3] 朱濤, Moustafa Dieb Thaer, 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 (第 30 回) 論文集, 2016. CD-ROM 1J2-4.

[4] Hiroshi Nakagawa and Tatsunori Mori. A simple but powerful automatic term extraction method. In *COLING-02 on COMPUTERM 2002: Second International Workshop on Computational Terminology - Volume 14*, COMPUTERM '02, pp. 1–7, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[5] Takeshi Abekawa and Akiko Aizawa. Sidenoter: Scholarly paper browsing system based on pdf restructuring and text annotation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pp. 136–140. The COLING 2016 Organizing Committee, 2016.

Copyright(C) 2019 The Association for Natural Language Processing.
All Rights Reserved.