非構造化ドキュメントからの情報抽出のための アノテーション・マニュアルキュレーションシステム

中津井 雅彦　小島 諒介　岩田 浩明　奥野 恭史
京都大学 大学院医学研究科

{nakatsui.masahiko.5n, kojima.ryoshuke.8e,
iwata.hiroaki.3r, okuno.yasushi.4c}@kyoto-u.ac.jp

1　はじめに

日々生産され続け, 共有される情報には, 計算機で容易に活用可能なように構造化され, 公共・商用データベース等に格納されているものがある一方, 自然言語のみで記載されているものや, 図表に含まれる情報など, 構造化されていないデータも数多い. 医薬品に関連する情報を例にとってみると, 薬剤のうち生理活性を示す有効成分の物理的プロパティ（溶解度や酸解離定数など）ですら十分にデータベース化されていない現状がある. また, 創薬・医療における研究開発では, ヒトの生体に対する有効性・安全性や ADMET（吸収, 分布, 代謝, 排泄, 毒性までの, 薬物が生体内に取り込まれてから体外に排出されるまでの過程）などの情報が重要となるが, モデル細胞や動物を対象とした実験結果は比較的容易に得られるものの, ヒトにおけるデータを得ることは比較的ハードルが高く, またデータベース化されている情報も一部に留まっているのが現状である.

一方で, 独立行政法人 医薬品医療総合開発機構（以下, PMDA）等では, CTD（Common Technical Document）など医薬品の承認審査に関わる情報や, 添付文書・インタビューフォームといった医薬品関連情報をオープンアクセスにて公開している [1]. CTD には臨床試験の結果や動物実験などの非臨床の情報の他, ヒトを対象とした臨床試験の結果や ADMET パラメータを含む様々な情報が, また添付文書には医薬品のヒトにおける有効性・安全性・使用に関する情報や医薬品そのものの物性に関わる情報が記載されている. これらの情報は, 情報科学的アプローチによって創薬・医療を効率化するために有効と考えられる. しかしこれらの文書は基本的に人が読み, 理解することを第一の目的として作成されており, 計算機で直接情報を利活用できる構造化されたデータにはなっていない. 例えば, PMDA からは平成 16 年以降に承認されたすべての薬剤の CTD がダウンロード可能であるが, 計算機での利活用には以下の課題がある.

- PDF ファイルに自然言語及び図・表によって記載されている
- 文書の構造（章立て）は規定があるが, 書き方については規定がなく, 記載方法がまちまちである
- 表のフォーマットが統一されていない
- テキスト情報を持つ PDF ばかりではなく, 画像のみで構成された, またはテキスト情報を持つページと画像のみのページが混在する PDF が存在する

これらの課題は, CTD などの製薬・医療関係文書に限ったものではない. たとえば, がんに関連する論文は毎年 18 万件のペースで増加している. これらをすべて人が読み, 理解する事は不可能な状況であり, 非構造化データを構造化するための汎用的なワークフローおよびシステムが必要とされている. この困難を解決するために, 筆者らは非構造データから機械学習により高精度に情報抽出を行うためのフローを開発している（図 1）. 具体的には, 以下の要素を開発中である.

- PDF ファイルに記載された自然言語・表を対象とするアノテーションモデル・ガイドライン
- アノテーション及びマニュアルキュレーションシステム
- 機械学習による自然言語からの関係性抽出モデル
- 機械学習による表のアノテーションモデル

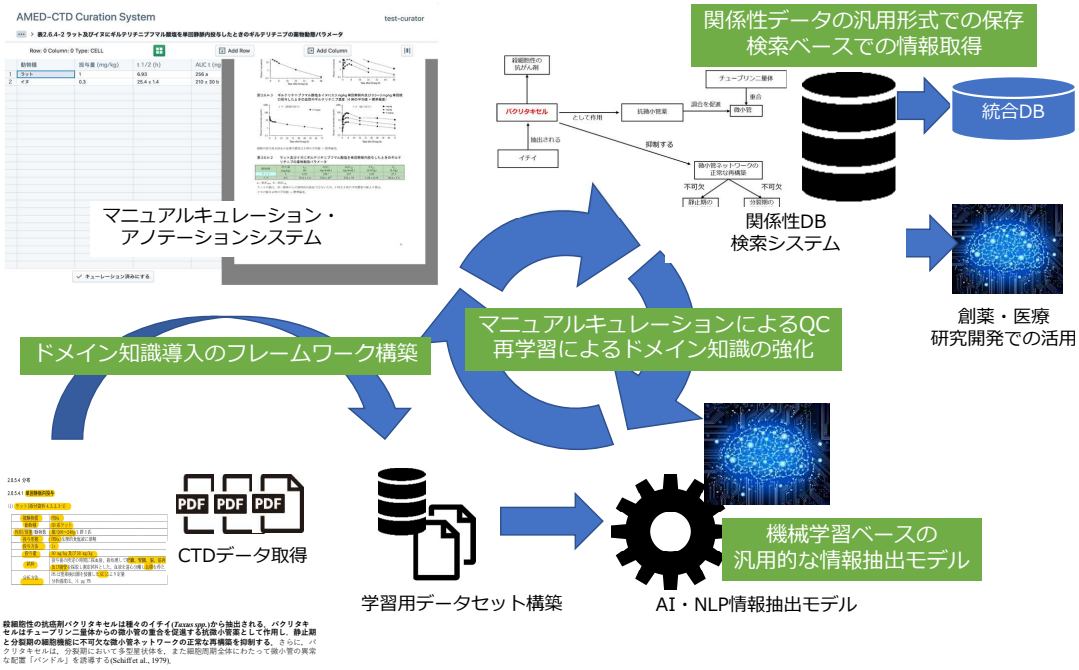


図 1: 非構造化データからの高精度情報抽出システムの概要

一方で, 非構造データからの情報抽出においては, 有用な情報の記載箇所や方式がその非構造データが属する専門分野においてまちまちであるという現状があり, 専門家の持つドメイン知識を情報抽出へ反映させることが重要である. 本研究では, 機械学習による自然言語からの情報抽出を前提として, そのための専門家のドメイン知識を学習用データセットに反映可能なアノテーションシステム, 及び機械学習による抽出結果を人手で確認・修正し, その情報を機械学習モデルにフィードバックすることでデータの精度の担保と情報抽出の高精度化を行うためのマニュアルキュレーションシステムの構築を行った.

2 アノテーションによる学習用データセットの構築

本研究では, まずは CTD を対象として, 以下の情報抽出を実施することとした.

- 薬物動態パラメータ
- 肝毒性・心毒性パラメータ
- ADMET パラメータ

情報抽出を行う機械学習モデルの構築に使用する学習セットを構築するために, CTD に含まれる文字列や数値に対しての意味付けを行い, 単語間の関係を記述するアノテーションを行うためのガイドラインを策定した. 情報抽出を行う機械学習モデルへ, 専門家が持つドメイン知識を反映するために, 薬物動態パラメータ, 肝毒性・心毒性パラメータについては, 医薬基盤・健康・栄養研究所への, また ADMET パラメータについてはライフインテリジェンスコンソーシアム（LINC）へのヒアリング結果をもとに, アノテーションガイドラインを策定した. このアノテーションガイドラインを活用し, 後述のアノテーションシステムを用いて 50 医薬品の CTD を対象として, アノテーションによる学習用データセットの構築を実施中である.

3　アノテーション・マニュアルキュレーションシステムの開発

アノテーションガイドラインによって規定されるアノテーション手法を GUI 上で容易に実現するための手法を開発し, それをもとに使いやすいユーザインタフェースを持つアノテーションシステムを構築した. 本アノテーションシステムでは, 入力された PDF ファイルの自然言語での記載に対して, OCR 技術によるテキスト化を行い, 該当テキスト中での単語間の関係性をアノテーション可能である. また, 表に対するアノテーションについても手法の開発及び実装を実施中である（図 2）.

本アノテーションシステムは, 以下の特徴を持つ.

- PDF ファイルを入力とする
- CTD 以外のドキュメントにも対応する汎用的な実装である
- OCR 技術により, 画像 PDF にも対応し, PDF ファイル全体のテキスト・表抽出を実施する
- PDF 画面を確認しながら, アノテーション対象箇所を選択できる
- 自然言語文の中の, 単語間の二項関係をアノテーションできる
- アノテーションの項目・値のカテゴリ（文字列・数値など）は, 対象文書に合わせてユーザが定義できる
- マルチユーザの Web インターフェースを持ち, 複数人で共同のアノテーション作業ができる

オープンソース実装である Tesseract[2] と AWS の Textract サービス [3, 4] を組み合わせ, PDF 文書の構造化を行う. また, kuromoji[5] による形態素解析により, 名詞, 助詞などを分割している. さらに, PDFBox[6] を用いて, PDF 埋め込みのテキスト情報がある場合はそちらも利用する.

アノテーションの手順としては, まずアノテーション対象の文書を選択する（図 2 a.）. アノテーション対象の文書の PDF 表示から, アノテーション対象の箇所をマウス操作で選択すると OCR によってあらかじめテキスト化された文書が取り込まれ（図 2 b.), その文章に含まれる単語と, 単語間の関係（述語）についてグラフィカルな操作でアノテーションできる（図 2 c.）. 表記の揺らぎについては, 専用辞書, ユーザ辞書およびユーザ定義により代替表現をタグ付けできる（図 2 d.）. アノテーション結果は, JSON 形式にて出力可能である.

また, 本研究で開発したアノテーションシステムを拡張し, 機械学習による情報抽出によって得られたデータのマニュアルキュレーションを行うためのシステムを開発中である. 抽出されたデータについて, ドメイン知識を持つ専門家によるクオリティチェック及び修正を行い, その結果を情報抽出のための機械学習モデルへフィードバックすることで, 情報抽出の精度向上を目指す.

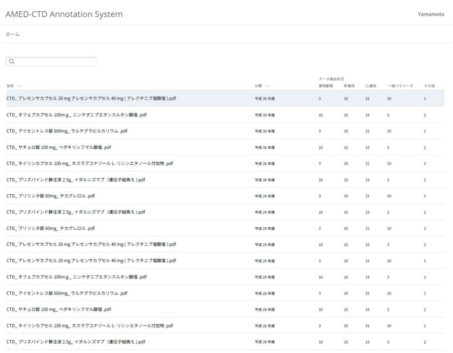
4　おわりに

本研究では, PDF ファイルに記載された自然言語及び表から高精度に情報抽出を行うため, 専門家のドメイン知識を活用するアノテーション（学習用データセットの構築）, 機械学習モデルによる情報抽出, マニュアルキュレーションによる抽出された情報の確認及び修正, 及びマニュアルキュレーション結果の機械学習モデルへのフィードバックによる精度向上を行うためのシステム開発を実施している.

製薬・医療関係文書のみならず, 自然言語によって記載され, 計算機で容易に活用することができない情報は多くある. 本研究では, まずは CTD からの情報抽出を目的としたが, 構築した自然言語・表からのアノテーションによる学習用データセットの構築, 機械学習による情報抽出, 及びマニュアルキュレーションのフロー・システムを活用することで, 様々な分野において, 各分野のドメイン知識を反映した非構造化データの構造化が可能になると期待される.

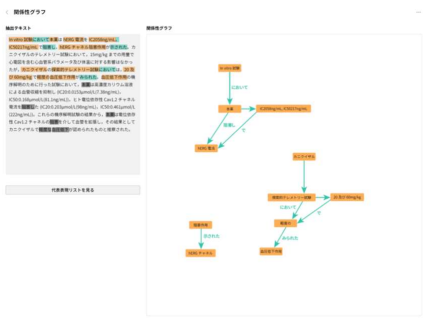
5　謝辞

本研究は, 日本医療研究開発機構 AMED の課題番号 JP19nk0101101h「製薬・医療関係文書からの情報抽出と基礎データベースの拡充」の支援を受けた. 成果の一部は, 国立研究開発法人新エネルギー・産業技術総合開発機構（ＮＥＤＯ）の助成事業の結果得られたものである. 情報抽出元の CTD ファイルは, 医薬品医療機器総合機構 PMDA から提供を受けた.

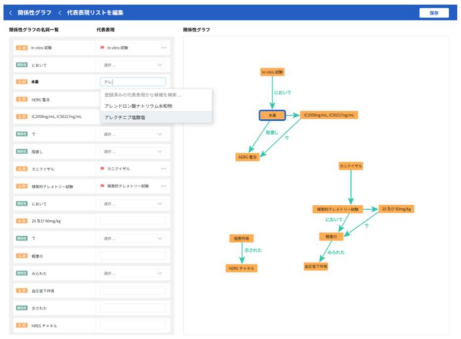


a. アノテーションサマリ

b. PDF文書ビュー



c. 関係性アノテーション



d. 代表表現によるタグ付け

図 2: アノテーションシステム

参考文献

[1] 独立行政法人　医薬品医療機器総合機構（PMDA）　医療用医薬品　情報検索, https://www.pmda.go.jp/PmdaSearch/iyakuSearch/

[2] ”An Overview of the Tesseract OCR Engine”, R. Smith, *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*.

[3] ”Introducing Amazon Textract: Now in Preview—easily extract text and data from virtually any document”. *AWS*. 2018-11-28. Retrieved 2019-06-09.

[4] ”Amazon’s newest machine learning product makes sense of unstructured medical text”. Dave Muoio, *Mobi Health News*. Retrieved 2019-06-09.

[5] ”kuromoji”, https://www.atilika.org/

[6] ”Apache PDFBox - A Java PDF Library”, https://pdfbox.apache.org/