

厚生労働科学研究費補助金
(政策科学総合研究事業 (臨床研究等 ICT 基盤構築・人工知能実装研究事業))
分担研究報告書

課題名 : 新薬創出を加速する症例データベースの構築・拡充/創薬ターゲット推定アルゴリズムの開発に関する研究
研究分担者名 : 高村 大也 国立研究開発法人産業技術総合研究所
人工知能研究センター・研究チーム長

研究要旨

基礎生物学系文献からの分子、病態の知識抽出については、言語リソースの構築を行い、エンティティ、関係及びイベント抽出モデルの開発および学習と評価を行った。また、論文から情報を抽出するキュレーション技術の開発を行った。

さらに、基礎生物学情報と抽出情報との融合を目指し、当研究所で開発された PoSSuM 等の蛋白質に関するデータベースを利用し、検索・推論基盤の構築に向けた準備を実施した。

A. 研究目的

本研究プロジェクトは、論文などの技術文献から、生命系のエンティティ(蛋白質、疾患など)やイベント、およびそれらの間の関係を、自動的に抽出する技術を開発することを目的とする。さらに、それを用いて実際に知識抽出を行い、抽出した知識を、エンティティやイベントをノードとし、関係を持つノード同士を結合した、巨大な高次ネットワーク構造(疾患ネットワーク)で表す。また、既存の蛋白質データベースの情報などを合わせて、蛋白質の配列比較や化合物の構造比較に関するバイオインフォマティクス/ケモインフォマティクスの技術と自然言語処理の技術を統合して、蛋白質などのエンティティ間の類似度も測定可能にする。疾患ネットワークと、エンティティ間の類似度を利用することで、柔軟かつ広範囲にエンティティ間の相互作用を推論する技術を開発する。

より具体的には、次の3つの課題を解く。

- 1 基礎生物学系文献からの分子、病態の知識抽出
- 2 基礎生物学情報と抽出情報との融合
- 3 知識の表現方法および柔軟な推論技術

B. 研究方法

① 基礎生物学系文献に関する言語リソース構築 :

論文などの技術文献から、生命系のエンティティ(蛋白質、疾患など)やイベント、およびそれらの間の関係を、自動的に抽出する技術を開発する。そのために、アノテーションガイドラインの作成や実際のアノテーションを含むコーパス開発を行う。

エンティティのアノテーション ; 特発性肺線維症(IPF)を中心とする肺疾患に関する文献要旨データ 120 件に対して、Disorder (疾患名)、GGPs (遺伝子・遺伝子産物)、Cell (細胞)、Anatomical entity (組織・器官など解剖学上のエンティティ)、Pharmacological substance (医薬品) など主要なエンティティのアノテーションとノーマリゼーションを済ませているが、更に追加の 30 件のデータに対して、エンティティとノーマリゼーションを行う。

関係性、イベントのアノテーション ; 医薬基盤研の肺疾患の専門家と議論・検討しながら、アノテーションのガイドラインとスキーム作成を進める。作成されたガイドライン、スキームに従って、エンティティ、ノーマリゼーションのアノテーションが完了した文献データに対して、アノテーションできる関係性やイベントのアノテーションを進めていく。

② 基礎生物学系文献からの分子レベルの知識抽出：

上の基礎生物学系文献に関する言語リソース構築で開発したアノテーションコーパスを基に、機械学習手法(特にニューラルネットワークに基づく手法)を開発し、実際に学習させることでエンティティ抽出技術の開発を行う。また、アノテーションコーパスにより、精度評価を行う。同様に、エンティティリンキング(ノーマライゼーション)の技術、またイベント抽出技術を開発する。

③ 知識の表現方法及び柔軟な推論技術の開発：

抽出されたエンティティや関係で構築される疾患ネットワークにおいて、そこに含まれる知識の表現技術や、ネットワーク上の柔軟な推論技術を開発する。ネットワーク上の推論技術として、知識グラフの埋め込み表現などを通じたイベント列を抽出する方法を開発する。特に、不確かな情報も有効に利用できるようなロバストなネットワーク上の推論手法の開発を進める。

④ 基礎生物学情報と抽出情報との融合：

新規標的候補等の同定に向けた文献情報と生物学情報との融合を目指し、当研究所で開発された PoSSuM 等の蛋白質に関するデータベースを利用し、知識抽出および推論技術により構築された疾患ネットワークの検証および拡充(新規情報の追加等)を可能とする検索・推論基盤の構築に資する。近年、蛋白質立体構造決定技術の向上により利用可能な構造情報が急増している。これらの全ての情報を利用するため新規サーバーを導入し、昨年度より PoSSuM データベースの更新作業を行っている。本年度は PoSSuMds データベースの更新にも着手する。これと同時に、検索・推論基盤構築に向けたケーススタディとして、肺線維症などの発症や治療に関わる免疫システムの論文調査を行い、抽出データを基に検索・推論基盤の構築を目指す。

(倫理面への配慮)

人及び動物を研究対象としていない研究であるため、倫理面の問題はないと判断した。

C. 研究結果

基礎生物学系文献に関する言語リソース構築のため、アノテーションスキームの設計、ノーマライゼーションのためのデータベースの調査、アノテーションガイドラインの作成、および実際のマニュアルアノテーションを行った。

また、上で開発された言語リソースを用い、エンティティ抽出モデル、エンティティリンキングモデル(ノーマライゼーションモデル)、関係抽出モデル、イベント抽出モデルの開発および学習と評価を行った。また、上記モデルで抽出した情報から、疾患ネットワークの構築し視覚化するデモシステムを構築した。また、論文から情報を抽出するキュレーション技術の開発を行った。

さらに、当研究所で開発された蛋白質の基質結合(候補)部位に関するデータベース PoSSuM を例として、検索・推論基盤の構築に向けた準備(PoSSuM の更新)を実施した。PDB データベースに 2020 年 9 月時点で登録されていた約 16 万件の蛋白質の立体構造およびアノテーションデータに対する、基質結合(候補)部位の類似性検索等の計算およびウェブサイトの更新作業を完了した。また、本プロジェクトで対象疾患の発症や治療に関わると予測された蛋白質の学術論文調査について、調査結果の再抽出および抽出結果の再フォーマットを実施した。

D. 考察

言語リソース構築については、アノテーションの質を測るためにアノテータ間一致度の算出が必要である。また、情報抽出については、実用性を高めるための技術開発が次の段階になる。また、PoSSuM と TargetMine の接続について今後さらなる検討を要すると考えられる。

E. 結論

コーパス構築、情報抽出技術の開発、検索・推論基盤の構築のそれぞれについて、予定していた開発目標を達成した。

G. 研究発表

1. 論文発表

なし

2. 学会発表

- 1) Mohammad Golam Sohrab (産総研), Anh-Khoa Duong Nguyen (産総研), Makoto Miwa (産総研, 豊田工業大学), and Hiroya Takamura (産総研), “Neural Exhaustive Approach for Entity and Relation Recognition Over Wet Lab Protocols”, The 6th workshop on Noisy User-generated Text (W-UT), 2020.
- 2) Mohammad Golam Sohrab (産総研), Anh-Khoa Duong Nguyen (産総研), Makoto Miwa (産総研, 豊田工業大学), Goran Topic (産総研), Masami Ikeda (産総研), and Hiroya Takamura (産総研), “BENNERD: A Neural Named Entity Linking System for COVID-19”, The 28th International Conference on Empirical Methods in Natural Language Processing (EMNLP), System Demonstration. 2020.

F. 知的財産権の出願・登録状況（予定を含む。）

1. 特許取得
なし
2. 実用新案登録
なし
3. その他
なし