提出日：2020 年 5 月 11 日

2019 年度研究実施報告書

(対象期間 2019 年 11 月 13 日～2020 年 3 月 31 日)

担当課題　　：新薬創出を加速する人工知能の開発

研究機関名　：国立研究開発法人産業技術総合研究所
研究責任者　：高村　大也　　takamura.hiroya@aist.go.jp

【１．実施内容】

実施項目１　（言語リソース構築）

１－１　研究目的

文献からの分子、病態の知識抽出を行うために、AI に学習させるための言語リソースを構築する。

１－２　研究概要・要旨

文献からの分子や病態に関する知識、特にプロテイン名や疾患名などのエンティティとそれらの間の関係を自動的に抽出するために､抽出を行う AI を学習するための訓練データとしてアノテーション付きコーパスを構築する。コーパス構築は、アノテーションガイドラインの作成および、実際のアノテーションを含む。

１－３　実施内容

文献としては IPF に関する基礎分子系の論文を選択し、そのアノテーションガイドラインを作成する。また、作成したアノテーションガイドラインに基づき、文献要旨に対してアノテーションを行う。

１－４　結果、成果等

IPF 基礎分子系ガイドライン作成に関しては、Regulation（制御）などの複雑な生命現象イベントについて、テキストマイニングの専門家や肺疾患の専門家にアノテーション例を複数提案して、意見を取りまとめながら、アノテーション用ガイドラインの作成を継続した。

IPF の基礎分子系要旨データのアノテーションに関しては､121 件の要旨データに対して､Gene expression(遺伝子発現)など比較的シンプルなイベントのアノテーションを開始した。また、アノテーション用の文献要旨データ 30 件を新たに選抜し､アノテーション用のデータセットとして準備した。この際に、主に Genes and gene products(GGPs；遺伝子や蛋白質など遺伝子産物)に関する記述の多い要旨を選抜した。

実施項目２（文献情報と生物学情報の融合）

２－１　研究目的

新規標的候補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データベース等を用いて検索・推論基盤を構築する。

２－２　研究概要・要旨

文献から抽出された情報と生物学情報との融合のために、抽出情報との関連探索技術を開発し、またデータベースの開発・更新・維持を行う。

２－３　実施内容

タンパク質の基質結合(候補)部位に関するデータベース PoSSuM に関し、検索・推論基盤の構築を行う。また、肺疾患に関し、学術論文や既存データベースなどからタンパク質や相互作用に関する情報を抽出する。

２－４　結果、成果等

当研究所で開発されたタンパク質の基質結合(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を開始(PoSSuM の更新に着手)している。データの劇的な増加に対処するため導入した新規計算機を用い、更新作業を行っている。これと並行し、ケーススタディとして、対象疾患の発症や治療に関わるとされる免疫システムについて、特に肺線維症に着目して学術論文や既存データベースなどから関連するタンパク質や相互作用等に関する情報を抽出した。またデータベース化に向けデータ編集方法を検討した。

実施項目３（文献からのエンティティおよび関係の抽出)

３－１　研究目的

基礎生物科学系文献、および臨床系文献から、蛋白質、疾患などのエンティティ、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。

３－２　研究概要・要旨

文献から、蛋白質、疾患などのエンティティを抽出する技術、エンティティをデータベースにリンキングする技術、エンティティ間の関係を推定する技術を開発する。また、文献のキュレーションデータを直接推定する技術の開発も行う。

３－３　実施内容

研究期間内では、エンティティリンキング技術の開発を進め、項目 1 で開発されたコーパスおよび、一般公開されている別ドメインのコーパスで性能評価を行う。また、文献からのキュレーションデータの推定については、一部の属性について、実際にプロトタイプツールを構築する。

３－４　結果、成果等

エンティティリンキングについて、モデル構築、実験などをさらに進めた。開発手法は、mention detection, candidate generation, candidate ranking の三段階から成る。このうち、candidate generation の段階で、高精度に候補を挙げられるように手法を改良した。また、開発した手法は、性能評価実験において良好な結果が得られている。文献からのキュレーションデータの獲得は、一部の属性については自動的に高精度に抽出できるようになった。また、簡易的なツールという形に実装した。

【２．外部発表、論文投稿等】

該当なし

【３．知財化について】

該当なし