日独仏 AI 研究
2020 年度採択研究代表者

2020 年度
年次報告書

松本 裕治

理化学研究所 革新知能統合研究センター
チームリーダ

医薬品安全性監視のための言語を越えた知識強化情報抽出

§1. 研究成果の概要

日本側研究チームは、研究計画書の WP3 および WP6 の 2 つのワークパッケージを主として担当する。WP3 の主なテーマは、固有表現認識と関係認識を機械学習によって行う言語非依存の手法とツール化であり、対象として科学技術論文とソーシャルメディアの文書データを扱う。医薬品安全性監視に関する情報抽出を行うための固有表現のセットやアノテーションスキーマの設定、その設定に基づくデータアノテーションを行い、教師付機械学習を適用する。固有表現のセットを決定するための基盤となる知識ベース（UMLS などを想定）へのエンティティリンキングを行う手法や表記揺れに対応するための用語の正規化に関する研究を行う。

初年度は、対象とする学術論文およびアノテーション対象の調査を行い、アノテーションスキーマの基本設計に着手した。アノテーションツールの選定を行い、データへのアノテーション作業を実行するための環境の整備を行った。少数のアノテーションデータを有効活用するための固有表現抽出手法の研究、および、文書全体から特定の役割を果たすエンティティや関係の抽出、複数段落またがって記述される関係を認識する手法の研究開発に取り組んだ。また、乳がん患者の発話、Tweet2000 件をサンプリングし、症状や医薬品の表現と、有害事象の有無をアノテーションしたデータを構築した。日本語症例報告 200 文書に、医薬品安全に関する8つの固有表現カテゴリにアノテーションを行った。このデータを用いたシェアードタスク NTCIR16 Real-MedNLP について、本研究の副作用抽出タスクを含んだタスクが採択され、2022 年の開催に向けての準備を開始した。

§2. 研究実施体制

(1) 松本グループ

① 研究代表者：松本　裕治　（理化学研究所　革新知能統合研究センター　チームリーダ）

② 研究項目　医薬品安全性監視のための言語を越えた知識強化情報抽出

・医薬品関係情報のアノテーションスキーマ設計とアノテーションデータ構築

・知識ベース構築のための関係情報抽出技術

(2) 相澤グループ

① 主たる共同研究者：相澤　彰子　（国立情報学研究所　コンテンツ科学研究系　教授）

② 研究項目　医薬品安全監視のための知識統合・検索技術

・知識統合のための固有表現の認識・正規化技術

・知識検索を強化する文書レベルでの情報抽出技術

(3) 荒牧グループ

① 主たる共同研究者：荒牧　英治　（奈良先端科学技術大学院大学　先端科学技術研究科　教授）

② 研究項目　医薬品安全に関する SNS データ解析基盤とテストベッドの構築

・SNS データ解析基盤のためのデータ仕様策定と予備的データ構築

・医薬品安全性監視のための多言語テストベッドの開発