

日独仏 AI 研究
2020 年度採択研究代表者

2021 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
チームリーダー

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(文書からの非言語依存の構造化知識獲得) および WP6(普及活動)の2つのワークパッケージを主として担当している。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。

初年度から昨年度にかけて、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を実施した。毎月1回の日独仏定例ミーティングにて国際的な副作用の記述仕様を議論し、詳細化を行っている。同時に日本語部分については、独自に国内でアノテーション仕様を議論し、パイロットデータを構築した。また、一部の仕様については、NTCIR Real-MedNLP タスクにおける ADE サブタスクに導入し、本年度、シェアードタスクを実施した。結果、国内のみならず海外も含め、9チームが結果を提出し、次年度6月にワークショップを開催予定である。言語処理タスクに落とし込めるような仕様化には慎重な検討が必要であり、日本チームとして以上のように先行して検討・実践を進め積極的にフィードバックしたことで、副作用記述仕様が概ね固まった。固有表現抽出に関しては、少量のアノテーションデータを有効活用するためのドメイン適応の手法を検討するとともに、医薬品に関する日本語論文のコーパスを整備して、固有表現や関係抽出に必須となる深層学習言語モデル構築の準備を進めた。

§ 2. 研究実施体制

(1) 松本グループ(研究機関別)

- 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 チームリーダー)
- 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(研究機関別)

- 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(研究機関別)

- 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

【代表的な原著論文情報】

- 1) Noriki Nishida and Yuji Matsumoto: “Out-of-Domain Discourse Dependency Parsing via Bootstrapping: An Empirical Analysis on Its Effectiveness and Limitation,” Transaction of the Association for Computational Linguistics, Vol.10, pp.127-144, February, 2022.
- 2) An Dao and Akiko Aizawa: “Domain Adaptation for Named Entity Recognition: An Analysis on Letter-case.” Fifth International Workshop on SCientific DOcument Analysis (SCIDOCA2021) associated with JSAI International Symposia on AI 2021 (IsAI-2021). November, 2021.
- 3) Masaya Ohagi and Akiko Aizawa: “Citation Context-Aware Citation Network Embeddings Based on Pre-trained Transformer” SciNLP 2021: the 2nd Workshop on Natural Language Processing for Scientific Text, hosted at the 3rd Conference on Automated Knowledge Base Construction (AKBC 2021). Online, October 8, 2021.
- 4) Kaito Sugimoto and Akiko Aizawa: “Context-aware Citation Recommendation Based on BERT-based Bi-Ranker.” SciNLP 2021: the 2nd Workshop on Natural Language Processing for Scientific Text, hosted at the 3rd Conference on Automated Knowledge Base Construction (AKBC 2021). Online, October 8, 2021.
- 5) 大萩雅也, 相澤彰子: “Pre-trained Transformer による引用文脈を考慮した引用ネット

ワーク埋め込み.” 言語処理学会第 28 回年次大会(NLP2022), オンライン開催,
2022.03.(若手奨励賞)