

Knowledge Acquisition Team

Yuji Matsumoto

知識獲得チーム

松本 裕治



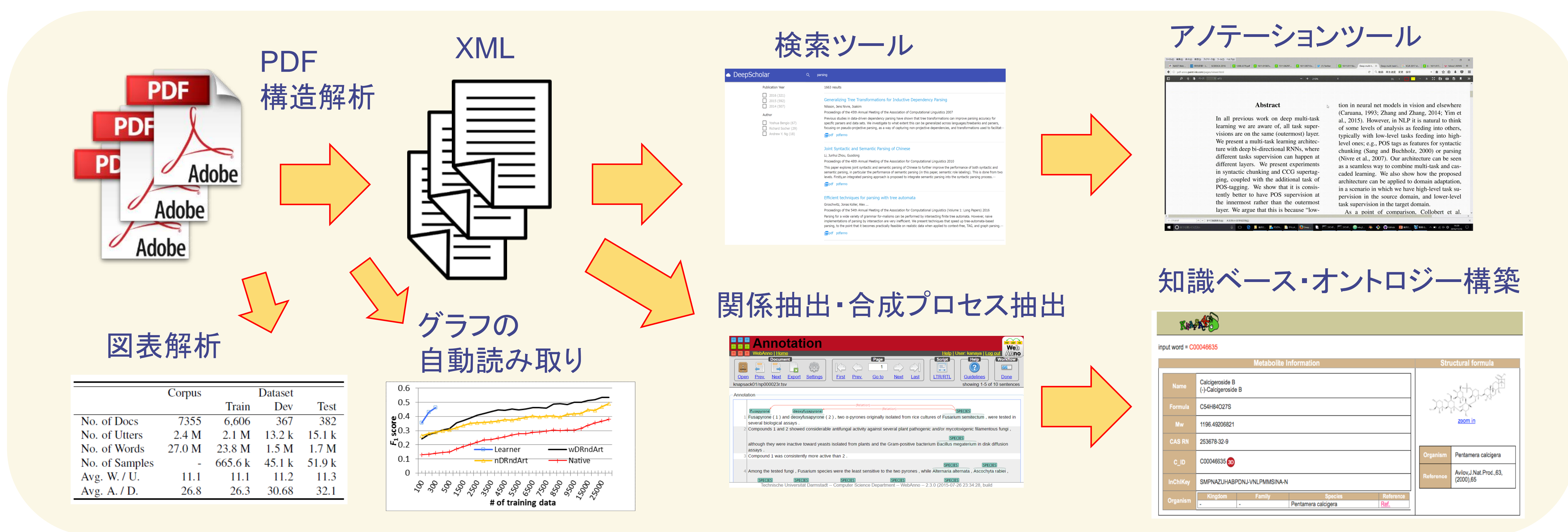
Center for
Advanced Intelligence Project

メンバー:

- チームリーダー: 松本裕治
- 特別研究員: 西田典起, 寺西裕紀
- テクニカルスタッフ: 徳永なるみ
- 客員研究員: 新保仁、重藤優太郎(千葉工大), 進藤裕之, 大内啓樹 (NAIST), 奥村貴史(北見工大), 古崎晃司(大阪電通大) 相澤彰子, 森純一郎, 濱口拓男(東大), 山田育矢(Studio Ousia), 辰巳守祐(NTTドコモ)
- パートタイマー研究員: 博士学生2名

概要・研究の意義

- 文書データ、主に専門分野の文書からの知識獲得技術の開発、および、専門分野オントロジーの自動構築を目的とする研究開発を行う
- 文書解析のための自然言語処理に関する基盤技術の研究開発を行う
- 専門分野の研究者との協働により実応用システムの研究開発を行う



研究項目:

- 専門文書からの知識獲得に関する基盤技術の開発および応用システムの研究開発を行う
- 材料分野における特性情報やプロセス情報の抽出手法の研究開発を行、材料データベースの半自動構築を実現する
- 専門用語、複雑な文、文章構造などの言語解析、高度な文書検索などの基礎研究を行う

これまでの代表的な研究成果:

- COVID-19関連論文の意味に基づく検索システムCovRelexをCRESTの共同研究として構築
- NEDO研究の成果として材料論文からのデータ構造化AIツール群を構築
- 論文からの知識獲得に関する手法の研究開発
 - PDF論文の構造解析 / 図表の内部構造解析
 - カスタマイズ可能な固有表現認識
 - 並列句を伴う固有表現の認識
 - 文書構造解析(文書構造を利用した共参照解析)
 - 関係情報の抽出(生成モデルによる関係認識)

共同研究・外部資金など(企業との共同研究を除く):

- NIMS: ポリマーデータベースPolyInfoのための材料特性情報・プロセス情報の自動獲得
- NISTEP: 文献間の引用文脈解析
- NHO: 退院サマリの自動生成
- JST日独仏AI研究: 「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」
- NEDO: 「材料データ構造化AIツール開発」

今後の予定・目標:

- 汎用かつ分野適応可能な知識獲得: 専門文書からの用語認識、物性情報や合成プロセス抽出のための汎用かつ分野適応可能なツールを構築し、公開する
- 基盤的言語解析: 複雑な文構造の解析の精度向上を達成し、表面的な言語解析では抽出が難しい知識の獲得手法を達成する
- マルチモーダル文書解析: 文書の本文、図表などのマルチモーダルな解析技術を開発する
- 潜在的構造に基づく知識獲得: 分野オントロジーや課題のための構造獲得と並行した知識獲得
- 主な研究発表:
 - N. Nishida, Y. 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, 2022.
 - Y. Wang, H. Shindo, Y. Matsumoto, T. Watanabe, "Nested Named Entity Recognition via Explicitly Excluding the Influence of the Best Path," ACL-IJCNLP-2021.
 - T. Wada, T. Iwata, Y. Matsumoto, T. Baldwin, Jey Han Lau, "Learning Contextualised Cross-lingual Word Embeddings and Alignments for Extremely Low-Resource Languages Using Parallel Corpora," Workshop on Multilingual Representation Learning, 2021. (Best Paper Award受賞)
 - U. Honda, Y. Ushiku, A. Hashimoto, T. Watanabe, Y. Matsumoto, "Denoising Pseudo-Captions in Unsupervised Image Captioning," EACL-2021.
 - V. Tran, V.-H. Tran, P. M. Nguyen, C. M. Nguyen, K. Satoh, Y. Matsumoto, M. L. Nguyen, "CovRelex: A COVID-19 Retrieval System with Relation Extraction," EACL, System Demonstrations, 2021