知識獲得チーム　松本 裕治

Knowledge Acquisition Team

Yuji Matsumoto

- メンバー:
  - チームリーダ:　松本裕治
  - テクニカルスタッフ:近藤修平,
  - 客員研究員:新保仁、重藤優太郎(千葉工大), 進藤裕之(NAIST), 奥村貴史(北見工大), 林克彦(東大), 濱口拓男(統数研), 山田育矢(Studio Ousia)
  - パートタイマー研究員:博士学生4名

- 概要・研究の意義
  - 専門文書からの知識獲得に関する統合的なツールおよびシステムの研究開発を進める
  - テキスト部分以外の図(グラフ)や表の自動読み取りによる物性情報抽出法の研究開発を行う
  - 専門用語、複雑な文、文章構造などの言語解析の研究を進め、多言語データから専門分野辞書やオントロジー構築技術に関する研究を行う。

	Corpus	Dataset		
		Train	Dev	Test
No. of Docs	7355	6,606	367	382
No. of Utters	2.4 M	2.1 M	13.2 k	15.1 k
No. of Words	27.0 M	23.8 M	1.5 M	1.7 M
No. of Samples	-	665.6 k	45.1 k	51.9 k
Avg. W. / U.	11.1	11.1	11.2	11.3
Avg. A. / D.	26.8	26.3	30.68	32.1

- 研究項目:
  - 論文構造解析
    - 論文PDFの構造解析, テキスト, 数式, 表, 図(グラフ)の領域の認識, キャプションと図表の対応などの自動化
    - テキスト部の文章としての構造解析, 文書検索ツールの開発
  - 多言語用語辞書・オントロジー構築支援
    - 専門分野の用語辞書やオントロジー構築のための情報抽出の半自動化. 主要な抽出項目については, 作業者(アノテータ)とのインタラクティブなツールを構築, ドメイン適用可能な用語抽出, 関係抽出, イベント系列抽出のツール化
  - 専門文書の言語解析技術
    - 並列構造など, 専門性の高い文書に現れる複雑な言語表現を解析するための技術

- これまでの代表的な研究成果:
  - 文書からの関係情報抽出に関する新しい手法を提案し、共通のデータセットに置いて最高性能を達成
  - 文書解析技術:複雑な言語表現の解析、とくに、専門分野の文章に頻出する並列句の解析について最高精度を達成、複数の語によってまとまった意味を表す複単語表現解析のための基本リソースと解析手法を提案
  - 教師無しの多言語間の単語対応において従来法を大きく上回る方法を提案
  - PDFからの図表領域の特定、グラフと表の自動読み取りに関して、従来法を上回る性能を達成
  - 材料名称の日英自動翻訳、PDF論文のアノテーションなどの支援ツールを開発

- 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、関係認識、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

- 共同研究・外部資金など:
  - PolyInfo(NIMS):ポリマー特性情報の獲得
  - 化合物データベース(富士通連携研究):日本語版材料データベースの半自動構築
  - 胸部画像からの読影レポート生成(富士フイルム)
  - JST-CREST「構造理解に基づく大規模文献情報からの知識発見」
  - NEDO材料データ構造化AIツール開発」

- 主な研究発表:

  - J. Liu, H. Shindo, Y. Matsumoto, “Development of a computer-assisted Japanese functional expression learning system for Chinese-speaking learners,” Educational Technology Research and Development, Springer, Vol.67, No.5, 2019. (Outstanding ETR&D Journal Article Award受賞
  - H. Teranishi, H. Shindo, Y. Matsumoto, “Decomposed Local Models for Coordinate Strcture Parsing,” NAACL-HLT 2019.
  - T. Wada, T. Iwata, Y. Matsumoto, “Unsupervised Multilingual Word Embedding with Limited Resources using Neural Language Models,” ACL-2019.
  - V-T. Phi, J. Santoso, V-H. Tran, H. Shindo, M. Shimbo, Y. Matsumoto, “Distant Supervision for Relation Extraction via Piecewise Attention and Bag-Level Contextual Inference,” IEEE Access, Volume:7, Issue:1, July 2019.