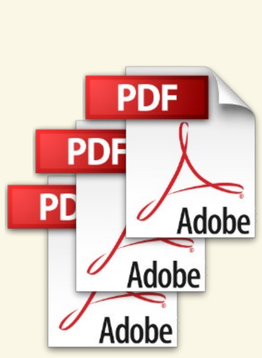
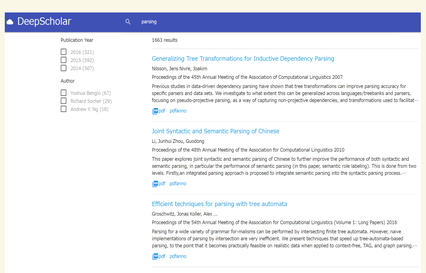
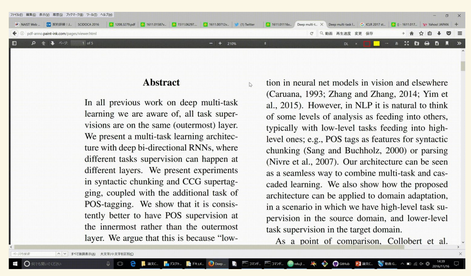
知識獲得チーム　松本 裕治
Knowledge Acquisition Team
Yuji Matsumoto

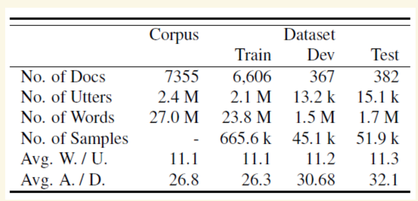
- メンバー:
  - チームリーダ:　松本裕治
  - 研究員:後藤啓介，近藤修平, 濱口拓男
  - 客員研究員:新保仁, 進藤裕之(NAIST),
    奥村貴史(北見工大)，重藤優太郎(千葉工大),
    武田浩一（名古屋大）, 林克彦（大阪大）,
    山田育矢(Studio Ousia)
  - パートタイマー研究員:博士学生5名，修士2名

- 研究背景:
  - 論文出版数の飛躍的な増加
  - 専門分野の知識ベース構築コストの増加
- 研究目標:
  - 類似論文検索・論文間関係解析
  - 専門文書からの知識抽出
  - 知識ベース構築・知識ベース補完

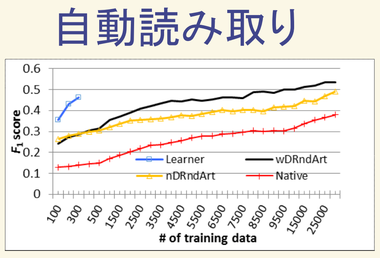


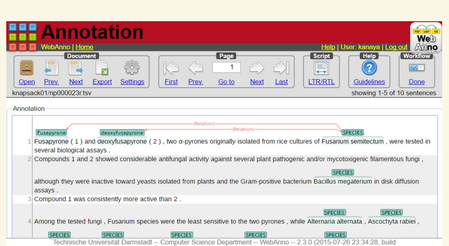


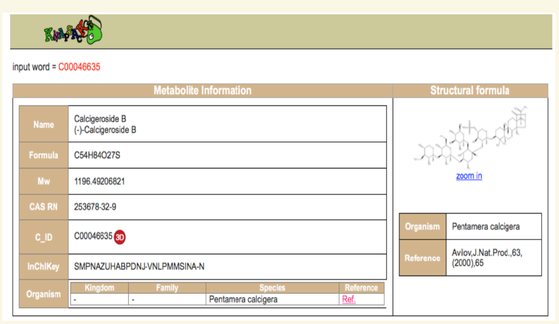




	Corpus	Dataset		
		Train	Dev	Test
No. of Docs	7355	6,606	367	382
No. of Utters	2.4 M	2.1 M	13.2 k	15.1 k
No. of Words	27.0 M	23.8 M	1.5 M	1.7 M
No. of Samples	-	665.6 k	45.1 k	51.9 k
Avg. W. / U.	11.1	11.1	11.2	11.3
Avg. A. / D.	26.8	26.3	30.68	32.1







- 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
  - 知識ベース構築支援
    - 専門分野の知識データベース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
  - 専門知識の要約，生成
    - 要約:単一・複数論文の要約
    - 生成:医療画像からの読影レポート生成，病院患者カルテからの退院サマリ生成

- 専門分野のグループとの協働:
  - PolyInfo（物質・材料研究機構）:ポリマーに関する特性情報の獲得
  - 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）

- 今後の予定:
  - 材料系論文からの知識ベース構築支援環境
  - 論文解析および論文からの知識獲得
    - グラフ・図表と本文との関係解析
    - 用語認識・関係抽出・合成プロセス抽出
    - 知識ベース補完
  - 主な研究発表:

- Hiroyuki Shindo and Yuji Matsumoto, “PDFAnno: a Web-based Linguistic Annotation Tool for PDF Documents,” LREC-2018, May 2018.
- Motoki Sato, Jun Suzuki, Hiroyuki Shindo, Yuji Matsumoto, “Interpretable Adversarial Perturbation in Input Embedding Space for Text,” IJCAI, July 2018.
- Van-Thuy Phi, Joan Santoso, Masashi Shimbo and Yuji Matsumoto, “Ranking-Based Automatic Seed Selection and Noise Reduction for Weakly Supervised Relation Extraction,” ACL-2018, July 2018.
- Hiroki Ouchi, Hiroyuki Shindo, Yuji Matsumoto, “A Span Selection Model for Semantic Role Labeling,” EMNLP 2018, November 2018.

- 具体的な成果の例:
  - グラフの自動読み取り
    - CNNによる特徴抽出＋ピクセル分類

Method	Accuracy	F1
FigureSeer	0.264	-
Rule-color	0.152	0.598
Rule-template	0.095	0.386
Our method	**0.291**	**0.679**

    - x、y軸の目盛り、凡例のラベルと形の読み取り

x軸と目盛	y軸と目盛	凡例のラベル	凡例の図形
0.901	0.897	0.865	0.884

  - 化学物質名の英日翻訳
    - 日化辞の300万語（日英対訳）を訓練データに利用
    - 統計翻訳（SMT）、ニューラル翻訳（NMT）の併用
    - 両者の訳が完全一致した際の精度は、99.76%

完全一致率(test)

	第1分割	第2分割	第3分割	全体
SMT	98.74%	98.92%	98.68%	98.78%
NMT	98.39%	98.49%	98.35%	98.41%