PDFAnno: PDFドキュメントのための言語アノテーションツール

進藤 裕之　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{shindo, matsu}@is.naist.jp

1 はじめに

自然言語処理において，テキストデータに対する言語アノテーション作業は，統計モデルの訓練や評価のために必要不可欠である．ここで言語アノテーションとは，テキストデータに対して品詞，依存構造，固有表現や共参照関係などの情報を付与することを指す．人手による言語アノテーション作業は時間的・経済的に非常にコストがかかるため，誰でも簡単に使用でき，管理が容易な言語アノテーションツールは，高品質な注釈付きデータを素早く構築するために非常に重要である．

誰でも無料で使える汎用的な言語アノテーションツールとして，brat [5] や WebAnno [7] が挙げられる．これらのツールは，Web ブラウザ上に構築されたユーザーインターフェースによって，アノテーターが文章に対して注釈作業を行うことのできる環境を提供する．ただし，これらのアノテーションツールは，テキスト形式のドキュメントのみを対象としている．

一方，科学技術論文をはじめとする様々な出版物は，PDF 形式で提供されることが多い．PDF は，構造化されたテキストや画像などのマルチメディア情報を埋め込むことが可能で，特定の PC 環境に左右されず，同じレイアウトでドキュメントを表示できるという利点がある．Adobe Acrobat Reader のような一般的な PDF リーダーには，PDF ドキュメントに対してテキストのハイライト機能や，コメントを注釈できる機能が標準で付属しているが，これらは言語処理タスクで想定される「アノテーション」とは根本的に異なり，例えば単語間の依存関係やエンティティ間の共参照関係を注釈することは想定されていない．

PDF は様々な出版物の標準的なフォーマットであるため，PDF に含まれるテキストに対して効率的にアノテーションを行うツールがあれば，例えば科学技術論文から情報抽出を行うタスクの正解データ作成などに役立つことが期待できる．実際に，ACL Anthology[1] に含まれる学術論文のデータセットに対して，共参照関係のアノテーションを行った研究が存在する [3, 4, 6]．彼らの研究では，PDF 形式の論文を OCR ソフトウェアを用いてプレインテキスト形式へ変換し，テキスト化された文章に対して共参照アノテーションを行っている．しかしながら，これには２つの大きな問題がある．１つ目の問題点は，OCR ソフトウェアの出力は多くの認識誤りを含むという点である．例えば，現在の OCR ソフトウェアでは，論文に含まれる表，数式，脚注などと本文との区別が難しく，本文テキストのみを上手く抽出できない場合が多い．実際，上記の既存研究では，アノテーション作業を行う前にオリジナルの PDF ファイルと変換後のテキストファイルとを比較して，認識誤りやノイズを除去する作業が必要であったと報告している．２つ目の問題点は，PDF を一旦プレインテキストへ変換してしまうと，段落などの構造化情報が失われてしまうという点である．そのため，アノテーターが文章の意味を素早く正確に理解して注釈作業を行うことが困難になり，アノテーションの品質低下の原因となり得る．

そこで本研究では，直接 PDF ドキュメントに対して言語アノテーションを行い，その後にアノテーション情報をテキスト形式で取り出すことを提案する．図 1 に，既存研究における PDF ドキュメントのアノテーション作業と，本研究におけるアノテーション作業の処理プロセスを示す．PDF に直接アノテーションを行うことによって，アノテーション作業の効率化が期待できる．また，アノテーション情報が OCR ソフトウェアの出力に依存しなくなるため，より高性能な OCR が今後開発されたときに，アノテーション情報を修正する必要がないという利点がある．

我々の開発した PDF 言語アノテーションツールである PDFAnno は，Web ブラウザ上で動作するアプリケーションであり，PDF 上に直接アノテーションを行うことができる．また，複数人でのアノテーション作業をサポートするため，複数のアノテーターが作業

[1] http://aclweb.org/anthology/

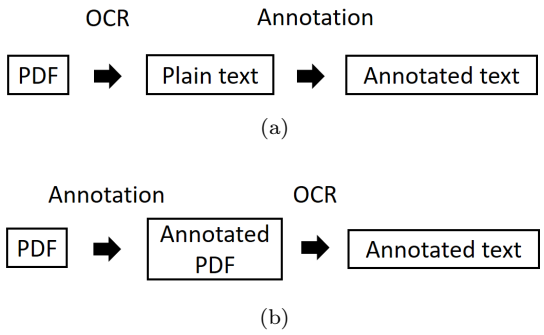

図 1: PDF ドキュメントに対するアノテーション作業の処理プロセス．(a) はじめに PDF をプレインテキストへ変換し，その後アノテーション作業を行う．(b) PDF 上に直接アノテーションを行い，その後にテキスト形式へ変換する．

を行った結果を，同一の PDF 上に重ねて描画することができる．これにより，アノテーター間の一致率を計算することや，複数人のアノテーション結果を確認しながら，不一致部分を修正することが可能である．PDFAnno は Github 上 [2] で MIT ライセンスに基づいて公開している．

2 関連研究

自然言語処理分野では，多くの言語アノテーションツールが構築されてきた [1, 2, 5, 7]．ここでは，特定のタスクに特化せず，汎用的な言語アノテーション機能を提供するものについて記述する．

brat (brat rapid annotation tool) [5] は，Python で実装されたテキストのアノテーション・可視化のための汎用的なツールである．brat は品詞タグ，固有表現，依存構造など様々なアノテーションに使用することができるが，テキスト形式のドキュメントのみを対象としており，PDF は使用できない．また，複数人でのアノテーションに対する機能はあまり充実していない．

WebAnno [7] は，brat と同様に幅広い種類のアノテーションに利用可能な Web ベースの言語アノテーションツールである．WebAnno は，プロジェクト管理やユーザー管理のツールが充実しているという特徴がある．しかし，アノテーションの描画部分は brat を利用しているため，PDF 上に直接アノテーションを行うことはできない．

PDF に対するアノテーションソフトウェアとして，Adobe Acrobat，PDF Annotator[3] や A.nnotate[4] などの多くの商用ソフトウェアが存在する．前述のように，それらのソフトウェアでは，PDF に含まれるテキストのハイライトやノートを付与することができるが，言語アノテーションに使われることは想定していないため，単語の依存関係や共参照関係などの言語処理に必要なアノテーションをサポートしていない．

3 PDFAnno

3.1 ユーザーインターフェース

図 2 に PDFAnno のスクリーンショットを示す．PDFAnno はブラウザ上で動作するアプリケーションであり，PDF やアノテーションの描画には標準的な Web 技術を用いている．また，brat や WebAnno とはとは異なり，サーバーとの通信が発生しないため，一度ブラウザでページを開いておけばオフラインで使用することが可能である．

PDF の描画には，PDF.js[5] を用いている．PDF.js は，Firefox ブラウザの標準的な PDF ビューアーであり，ユーザーが使い慣れた環境で PDF の拡大・縮小や印刷，検索などの機能を使用することができる．我々は，PDF.js の上に JavaScript を用いてアノテーションレイヤーを実装した．現在のところ，PDFAnno は Span，Region，Relation の 3 種類のアノテーションをサポートしている．これらを組み合わせることによって，品詞，固有表現，エンティティ間の関係や共参照などのアノテーションを行うことができる．

3.1.1 Span

Span は，テキストの連続した範囲に対するアノテーションである．注釈したテキスト範囲は，PDF 上にハイライトされて表示される．また，注釈を行ったテキスト範囲に対して，ラベルを付与することができる．これは，単語の品詞や固有表現の種類を記述するために使用する．

PDFAnno では，ラベルの入力テキストにオートコンプリーション機能を実装している．そのため，アノテーターが過去に入力した履歴から，ラベル情報を選択することによって入力を行うことが可能である．

[2] https://github.com/paperai/pdfanno

[3] https://www.pdfannotator.com/
[4] http://a.nnotate.com/
[5] https://mozilla.github.io/pdf.js/

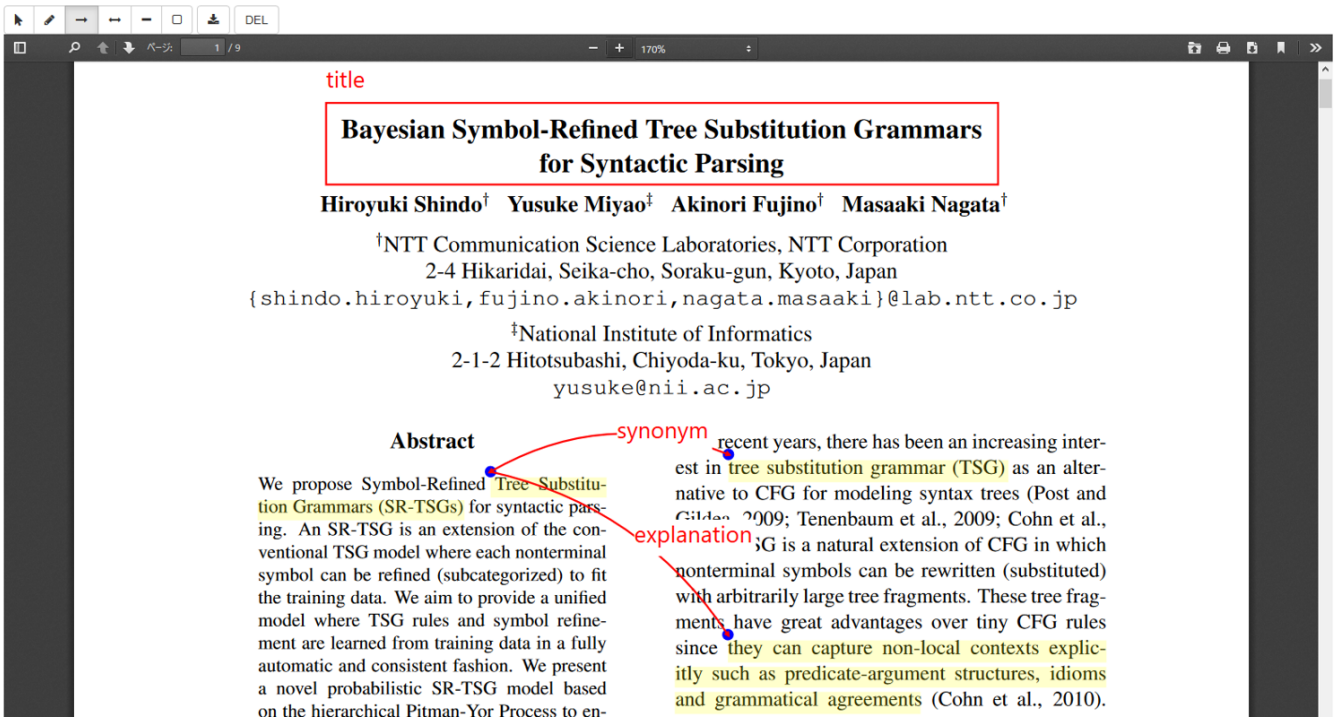

図 2: Screenshot of the PaperAnno user-interface, showing example annotations of text span, region, and relation.

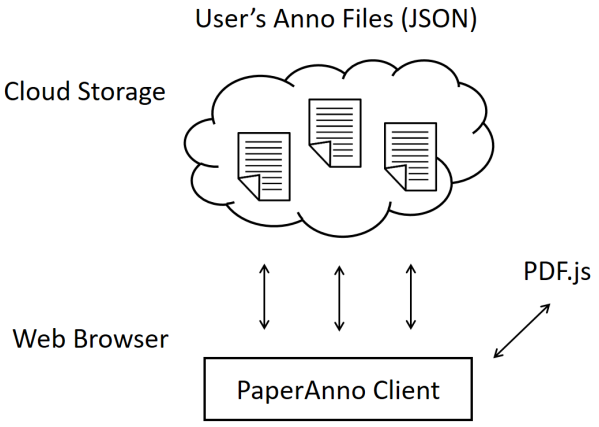

図 3: System architecture of PaperAnno.

3.1.2 Region

Region は，PDF 上で長方形の領域を指定するアノテーションである．これは，PDF に含まれる画像などの領域に対して，何らかの注釈を与えるために使用することができる．Span と同様に，Region にもラベルを割り当てることができる．これは厳密には言語アノテーションに含まれないが，PDF 中の図に含まれる情報に対して注釈を行いたい場合などに有用である．

3.1.3 Relation

Relation は，Span や Region 間に何らかの関係を注釈したい場合に用いるアノテーションである．PDFAnno では，片方向，双方向，無方向の 3 種類の関係矢印を注釈することができる．片方向の矢印は，主に依存関係を表し，双方向は，例えばエンテティ間に双方向の関係性が存在する場合に用いる．方向性の無い関係矢印は，例えば複数のエンテティをグループ化したり，共参照関係などのアノテーションに用いることができる．また，Span と Span との関係だけでなく，Span と Region や Region と Region との関係も同様に注釈することが可能である．

```
{
  "version": "0.0.1",
  "P12-1046.pdf": {
    "span-1": [
      [1, 95.818, 252.977, 181.761, 10.909],
      [1, 95.818, 264.806, 107.136, 10.909],
      "label1"
    ],
    "span-2": [
      [1, 323.863, 230.715, 213.988, 11.590],
      [1, 313.125, 244.522, 224.829, 10.795],
      "label2"
    ],
    "rel-1": [
      1, "two-way", "span-1", "span-2", "label3"
    ],
  },
}
```

図 4: アノテーションファイルの例.

3.2 アノテーションファイル

PDFAnno では，ユーザーのアノテーション情報は，オリジナルの PDF とは独立しており，ユーザーがいつでも JSON 形式のテキストファイルとしてダウンロード可能である．図 4 にアノテーションファイル（.anno）の例を示す．基本的には，PDF 単位ではなく，ユーザーごとにアノテーションファイルを作成することを想定しており，一つのアノテーションファイルに複数の PDF に関する注釈情報をまとめて保存することが可能な設計となっている．図 4 では，２つの Span と１つの Relation が“ P12-1046.pdf ”というファイルに対して注釈されたことを表している．Span は，ページ数と，(x, y, width, height) の座標として表現する．Region も同様である．Relation は，Span や Region の ID と，矢印の種別で表現される．

3.3 複数ユーザーによるアノテーション

PDFAnno のシステム全体像を図 3 に示す．PDFAnno は，サーバーとの通信を行わないクライアントアプリケーションであり，一度 Web ページにアクセスすれば，後はオフラインでも使用可能である．また，複数人でアノテーションを行う場合には，各アノテーターのアノテーションファイルをクラウドストレージなどで共有することによって，他人のアノテーション情報を参照することができる．

我々は，複数ユーザーによるアノテーションをサポートするため，複数のアノテーションデータを一つの PDF へ重ねて表示させる機能を実装した．このとき，各アノテーションファイルごとに表示色を選択することができ，アノテーターの違いは色の違いとして判別できる．また，複数のアノテーション情報を同時に表示して確認しながら，どれか一つのアノテーションを編集することもできる．そのため，アノテーター間のアグリーメントを確認したり，アノテーター間の不一致を確認しながらアノテーションの修正を行うことができる．

4 おわりに

我々は，PDF ドキュメントに直接言語アノテーションを行うことのできるツール：PDFAnno を開発した．PDFAnno を用いることによって，学術論文などの PDF ドキュメントに，固有表現の範囲や，エンティティ間の共参照関係を注釈することができる．今後は，PDF 以外の形式，例えば HTML や XML 形式のサポートや，OCR ソフトウェアとの統合などを行っていく予定である．

参考文献

[1] Kalina Bontcheva, Hamish Cunningham, Ian Roberts, and Valentin Tablan. Web-based collaborative corpus annotation: Requirements and a framework implementation. In *Proc. of New Challenges for NLP Frameworks workshop at LREC*, 2010.

[2] Christoph Muller and Michael Strube. Multi-level annotation of linguistic data with {MMAX2}. In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, pages 197–214, 2006.

[3] Chaimongkol Panot, Aizawa Akiko, and Tateisi Yuka. Corpus for Coreference Resolution on Scientific Papers. In *Procs of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 26–31, 2014.

[4] Ulrich Schafer, Christian Spurk, and Jorg Steffen. A Fully Coreference-annotated Corpus of Scholarly Papers from the ACL Anthology. In *Proceedings of COLING 2012: Posters*, pages 1059–1070, 2012.

[5] Pontus Stenetorp, Sampo Pyysalo, Goran Topic, Sophia Ohta, Tomoko Ananiadou, and Jun'ichi Tsujii. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proc. of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 102–107, 2012.

[6] Bird Steven, Dale Robert, Dorr Bonnie, Gibson Bryan, Joseph Mark, Kan Min-Yen, Lee Dongwon, Powley Brett, Radev Dragomir, and Tan Yee Fan. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proc of the Sixth International Conference on Language Resources and Evaluation (LREC)*, pages 1755–1759, 2008.

[7] Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. In *Proc. of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*, pages 1–6, 2013.