1H4-NFC-01a-2

# PubAnnotation - テキスト 注釈リ ポジト リ

PubAnnotation - Text Annotation Repository

金 進東 *1
Jin-Dong Kim

*1情報・ システム研究機構/ライフサイエンス統合データベースセンター
Rsearch Organization of Information and Systems / Database Center for Life Science

Text annotation is expensive but indispensable for development of text mining technology. While there are a few projects developing text annotation data, their reusability is becoming more and more a serious issue. PubAnnotation is developed to address the reusability issue. Developed primarily as an annotation storage system that can easily scale to tera-bytes, PubAnnotation features, among others, text alignment, which enables comparative and integrative use of text annotation. Together with its satellite tools, including PubDictionaries and TextAE, it also serves as a platform of crowd-sourcing annotation.

## 1. はじ めに

テキスト マイニング (text mining) に関する研究開発においては、人がテキストを 読むさいに取得すべき情報を、機械が抽出しやすいように構造化された形でテキストに追加する作業が行われており、テキスト 注釈 (text annotation) と呼ばれる。テキスト 注釈データ (注釈付きコーパス) は手動あるいは自動で作られるが、自動でテキスト 注釈 (以下、自動注釈) を 高精度で行うことができれば、テキスト マイニングの性能向上が期待される。手動で作られた注釈データは一般的に品質が高く、ベンチマーク (benchmark) 用のデータとして自動注釈の性能を 測るため使われたり、自動注釈のモデルとして活用されたりと 重要な資源である。しかし、高品質のテキスト 注釈データは生産コスト が高いことから、既存データを 再利用することが期待される。

世界的にテキスト 注釈データを 構築するプロジェクト は多数あり、成果は貴重な 資源として広く 活用されている。しかし、注釈データは各プロジェクト 独自のやり 方で作られ、形式 (format) やテキスト 前処理 (pre-processing) 方法等がばらばらになってしまい互換性が一般的に低く、再利用を 困難にする大きな 原因になっている。これらのテキスト データを 統一し、標準的な手順でアクセスできるようにすることでデータの再利用性が高まると 期待される。

このような状況を 踏まえ、テキスト 注釈の統合的な管理のためのスト ーリッジシステム (storage system) として PubAnnotation を 開発した。本論文では注釈レ ポジトリ (repository) とクラウドソーシングプラット フォーム (crowd-sourcing platform) としての PubAnnotation の機能に関して紹介する。

## 2. PubAnnotation

### 2.1 注釈レ ポジト リ

PubAnnotation は標準化されたテキスト 注釈データのレ ポジトリとして機能できるよう 開発している。そのため 複数の研究グループもしくは 個人によって 作られた様々な 注釈データに対応する 必要がある。そのため、以下の機能を 実装した。

連絡先: 金 進東，ライフサイエンス統合データベースセンター，千葉県柏市若柴 178-4-4，Tel:04-7135-5508，Fax:04-7135-5534，e-mail:jdkim@dbcls.rois.ac.jp

○ **テキスト アライメント**： 異なるグループやユーザが作成した注釈データを 集めると、その中には同一のテキスト に付けられた注釈もある。しかし、多くの注釈プロジェクト は注釈作業を 容易にする 目的でテキスト を 前処理するためにテキスト が変わる 場合が多い。例えば、ギリシャ 文字の展開 (「 $\alpha$ 」を「 alpha」に変換するなど) やト ークン化が行われる。その結果、同一テキスト に対して 作られた注釈データであっても、異なるプロジェクト で作られたもの同士では互換性のない場合が多い。この問題を 解決するため、PubAnnotation は Generalized LCS algorithm[Kim 2013] を 用いたテキスト アライメント (text alignment) 機能を 実装している。

○ **プロジェクト 管理**: 複数のグループ、ユーザが注釈データを 登録する 場を 提供するためにプロジェクト 管理機能は必須である。PubAnnotation では誰もがアカウント を 作り、注釈プロジェクト を 始めることができる。自身のプロジェクト に自身の注釈データを 格納することができるだけでなく、副管理者を 指定して共同作業することもできる。

○ **関係データベースの利用**: PubAnnotation は安定性とスケーラビリティの面で技術的に熟成されている関係データベースを 基盤として用いている。そのため大量の注釈データを 効率的に管理することが可能である。

○ **検索機能**: 大量の注釈データを 格納するため、必要な時に必要な部分にアクセスできる 仕組みが必要になる。PubAnnotation は関係データベースに格納しているため、SQL を 使った検索が行える。

○ **REST API**: 様々な 条件でデータにアクセス可能な REST API を 提供している。

○ **形式変換**： PubAnnotation は注釈データを 格納するために独自のテーブル形式を 用いているが、様々な 要求に対応するため形式変換機能を 持つ。現時点での出力可能な 形式は、プログラミングのための JSON、交換形式として 広く 使われる XML、セマンティックウェブのための RDF である。形式変換はより 多くの要求に対応するため Plug-in システムとして実装されている。

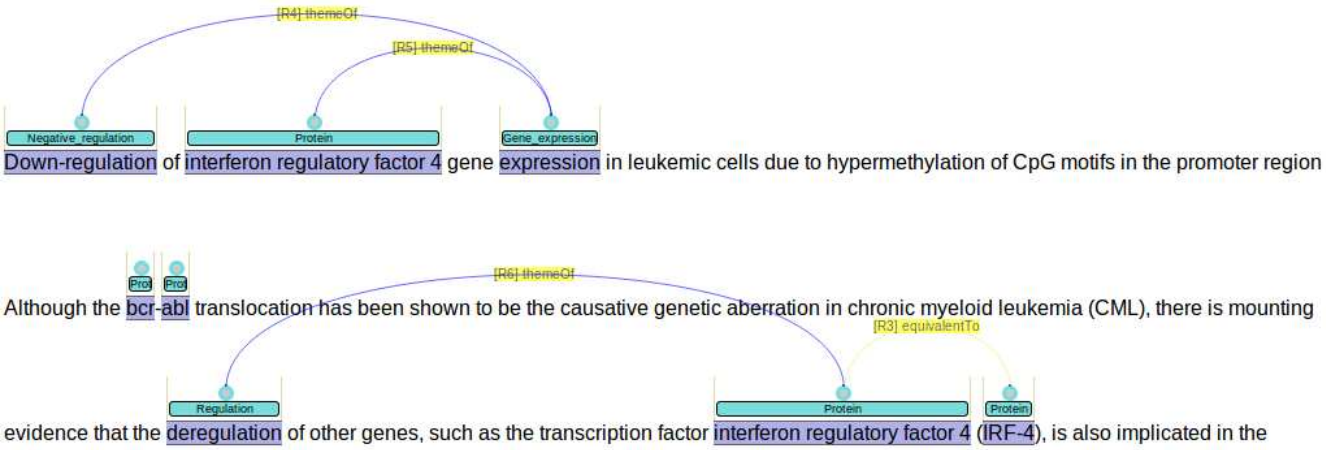


図 1: PubAnnotation が格納している注釈の例。TextAE の編集イメージ。

### 2.2 クラウドソーシング

PubAnnotation はクラウドソーシングによるテキスト注釈にも対応するため以下の機能を実装している。これらの機能を使って作られた注釈データはすぐに誰とでも共有できる。

- **手動注釈作業のためのエディター**: 人が論文を読みながら簡単に注釈できるよう PubAnnotation の連携ツールとして注釈エディターである TextAE を提供している。TextAE は Javascript で実装されたウェブベースのツールで、グラフィックユーザインタフェース (GUI) を使って注釈の編集ができる。

- **辞書による自動注釈**: 人手で注釈を付ける作業であっても、最初に自動的に付け、その結果の誤りを人手により修正することが最近の一般的な手順になっている。特に注釈のための用語辞書を用意し、それに基づいて自動的に注釈を付けたいという要求は多い。このため、PubAnnotation は連携システムとして辞書の管理と辞書による自動注釈ができる PubDictionaries というシステムを提供している。

- **オープンアーキテクチャ**: 自動注釈と注釈エディターは人手による注釈には必須なツールであり、PubAnnotation は連携ツールとして PubDictionaries と TextAE を提供しているが、同様のツールを開発するグループは他にもある。このため、外部のツールも必要に応じて PubAnnotation と一緒に使えるよう REST API を公開している。PubAnnotation と TextAE も当該 API を通じて PubAnnotation と繋がる仕組みになっている。

## 3. 終わりに

テキスト注釈データは構築にコストが掛かるものであるが、信頼できるテキストマイニングシステムの開発のためには必須の資源である。従って、構築されたデータを共有することでテキストマイニング研究開発コミュニティにおける開発コストが軽減できると思われる。このため、多様な注釈を統合的に管理でき、かつ、スケーラブルなレポジトリとして PubAnnotation を開発した。連携ツールとして辞書ベースの自動注釈システムである PubDictionaries やウェブベースの注釈エディターである TextAE も開発し、誰もが気軽に注釈データを作り、共有できるクラウド注釈にも対応している。図 1 は PubAnnotation に格納されている注釈の例である。PubAnnotation とその連携システムは以下の URL からアクセス可能である。

- PubAnnotation: `http://www.pubannotation.org`
- PubDictionaries: `http://www.pubdictionaries.org`
- TextAE: `http://textae.dbcls.jp`

## 参考文献

[Kim 2013] Kim, Jin-Dong: A Generalized LCS Algorithm and Its Application to Corpus Alignment (2013), Proceedings of the Sixth International Joint Conference on Natural Language Processing, 1112–1116.