E-007

YCAT: Web ベースのコーパスアノテーションツール

YCAT: a web-based corpus annotation tool

内海 慶 † 颯々野 学 † 山田 薫 † 谷尾 香里 † 前澤 敏之 †
Kei Uchiumi Manabu Sassano Kaoru Yamada Kaori Tanio Toshiyuki Maezawa

1 はじめに

我々は，Web 上のテキストを対象としたコーパス作成プロジェクトを進めている．このプロジェクトの目的は，各種言語処理に適用可能な学習コーパスを Web 上のテキストから作成することである．

大規模な学習コーパスの作成には，効率良くタグ付けを行うためのアノテーションツールが必要である．そこで，我々は学習コーパスの作成と並行して，内部でアノテーションツールの開発を行っている．

本稿では，我々が開発するアノテーションツール，YCAT について報告する．

2 Yahoo Corpus Anotation Tool

YCAT は，ヤフーで開発している係り受け情報付きコーパス（YWT:Yahoo Web Treebank）用のアノテーションツールである．

2.1 付与する情報

YCAT は，YWT コーパスに対する文節区切り情報と係り受け情報の修正，及びコーパスの内容に対する注釈を行うインターフェースとして機能する．

YWT コーパスには，形態素情報，文節区切り情報，係り受け情報が与えられる．初期の形態素情報，文節区切り情報，係り受け情報は，内製した形態素解析，文節区切り，係り受け解析によって付与する．YCAT では，付与された情報のうち，文節区切り情報と係り受け情報の修正を行う．この理由は，形態素解析の辞書や品詞体系の充実，チューニングは，本コーパス作成とは独立に行なわれているためである．将来，品詞体系の変更や辞書のエントリの変更があった場合には，形態素情報のみ入れ替える予定である．

図 1 に，YCAT でアノテーションを行った文の例を示す．形態素ごとに与えられている二種類のタグ {B, I} は，文節区切り情報を表している．B は文節の先頭を意味し，I は文節の途中を表す [3]．文節の開始位置ごとに与えられている情報は，係り受け情報を意味する．先頭の数字は，文節番号を表す．2 つ目の数字及びアルファベットは，係り先の文節番号と，係り受け関係を表す．係り受け関係は，常の係り受け関係，列関係，格関係などの情報を区別して与える．現在のアノテーションでは，{D, P, A} のタグを，それぞれ通常の係り受け関係，並列関係，同格関係を表すタグとして使用している．

2.2 ツールの特徴

我々の作成する YCAT は，ブラウザ経由で使用する，Web ベースのマルチユーザ型アノテーションツールである．実装は，Ajax による UI をベースに行っている．

ブラウザ経由で利用できるツールの利点は，プラットフォームへの依存の低さ，導入の迅速さ，開発やメンテナンスの容易さが挙げられる．

これに対し，従来の，係り受け関係のアノテーションを行うツールには，京大コーパス・プロジェクト [1] で作成されたアノテーションツールや，松本らが作成する「茶器」[2] などがある．これらを使用するためには，アノテーターが自分の PC にソフトウェアをインストールする必要がある．また，ソフトウェアに変更があった場合に，一括した更新は難しい．ツールの開発をするにあたり，我々はこれらの点を考慮した．

次に，YCAT の機能について説明する．

```
# written by 128.0.0.1 date Fri Jan 26 18:25:26 2007
# ID0000
0 5D
約 やく 約 接頭辞 冠数 * B
1 1 1 名詞 数詞 * I
年 ねん 年 接尾辞 助数 * I
前 まえ 前 名詞 名詞 * I
、 、 、 特殊 読点 * I
1 3D*2P
不安 ふあん 不安 名詞 名形 * B
と と と 助詞 助詞副詞化 * I
2 3D
期待 きたい 期待 名詞 名サ他 * B
を を を 助詞 格助詞 * I
3 5D
抱え かかえ 抱える 動詞 一段 連用形 B
ながら ながら ながら 助詞 接続助詞 * I
4 5D
ここ ここ ここ 名詞 名詞場所 * B
ハリウッド はりうっど ハリウッド 名詞 地名 * I
に に に 助詞 格助詞 * I
5 -1D
やって来 やってき やって来る 動詞 カ変 連用形 B
まし まし ます 助動詞 助動詞ます 連用形 I
た た た 助動詞 助動詞た 基本形 I
。 。 。 特殊 句点 * I
EOS
```

図 1 コーパスの例

2.3 機能概要

YCAT では，アノテーターの作業支援のために次のような機能を実装している．

1. 解析誤りの修正機能
   文節区切り，係り先情報の修正機能．
2. 表示機能
   係り受け解析の出力を見やすい形式に変換して表示する機能．
3. コメント挿入機能
   各文節に対するコメントや文全体に対して，アノテーターが気付いた点を記述する機能．
4. 検索機能
   コーパスに対する文字列検索や文 ID による検索を行う機能．
5. テキストの追加機能
   アノテーターが収集したテキストをコーパスに追加する機能．

† ヤフー株式会社，Yahoo Japan Corporation

6. 係り受け解析のデモ機能
　コーパスから学習したモデルを使って，係り受け解析を実行する機能.

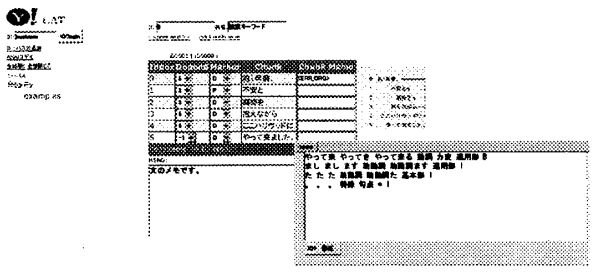

図 2　YCAT アノテーション画面

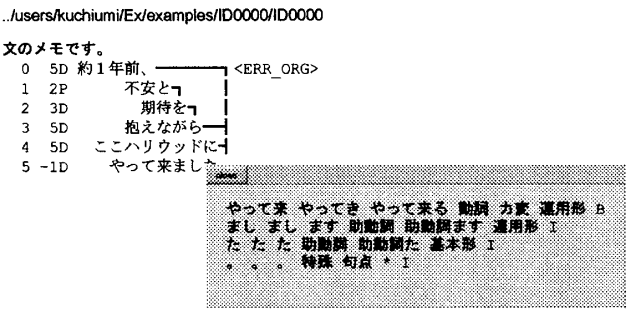

図 3　YCAT-アノテーション結果のスナップショット

2.4　機能詳細

解析誤りの修正機能　YCAT では，係り受け情報の誤りと文節区切りの誤りを想定しており，それらの修正機能を実装している.

　図 2 は YCAT のアノテーション画面である．プルダウンによる係り先と係り関係の修正，文節区切り情報のエディットが行える.

　係り先が修正された際には，非交差条件の自動判定と自動修正が行われる．このため，アノテーターが誤った修正を行った場合にも，非交差条件が破られることは無い.

　文節区切りの修正では，1 文節とされている区間を任意の位置で複数の文節に分割する，連続する文節を連結する，という処理を行う．YCAT では，エディットしたい文節をマウスでクリックすると，図 2 にあるようなエディットウィンドウが表示される．アノテーターは，ウィンドウ内の文節タグを編集することで，文節区切りの修正を行う.

　文節区切りの修正を行った場合，連結，あるいは分割された文節の係り先を決める必要がある．YCAT では，ある 文節 i の係り先の 文節 j が分割される場合は，文節 i の係り先は，分割された 文節 j の最後の 文節 j_{last} に係るとした．文節 j が分割される場合には，文節 j の係り先を 文節 j_{last} が継承し，分割された 文節 j の先頭から $last-1$ までの文節は，それぞれ直後に係るものとした．連続する文節を連結する場合は，一番後ろの文節の係り先を，連結して作られる文節の係り先とした.

表示機能　YCAT では，係り受け解析の出力を，直感的に理解しやすい形で出力するツリー表示と，アノテーションの様子をアノテーターに伝えるためのツリーのリアルタイム更新，アノテーター間でのアノテーション結果の共有をサポートするためのスナップショット生成，形態素情報の表示機能を実装している.

　アノテーション結果のスナップショットは，他のアノテーターのアノテーション結果を参照する場合に用いる．そのため，この機能はパーマリンクとして使用可能となっている．図 3 に，アノテーション結果のスナップショット画面を示す．スナップショット画面では，編集機能は除かれているが，形態素情報の表示も行える.

　形態素情報の表示は，図 2 のエディットウィンドウが兼ねており，文節区切りとあわせて修正も行えるようになっている．しかし，形態素解析のチューニングや辞書の整備は，独立した形で行っているため，現在は編集を行っていない.

コメント挿入機能　YCAT では，文節に対するコメントと，文全体に対するコメントの記述が行える．文節に対するコメントでは，本来の係り先とは異なる係り先候補が存在する場合や，顔文字などの記号が含まれている場合，原文に誤字脱字がある場合，文切りが失敗している場合などに，それらを示すタグを記述する.

　文全体に対するコメントでは，アノテーションの過程で見つけた言語現象などを自由に記述するために利用する.

検索機能　YCAT では，コーパスに対する文字列検索が行える．アノテーターは，検索フォームから文字列を入れることで，入力文字列を含む文を検索し，文 ID を得る事が出来る．検索結果で表示された文 ID をクリックすることで，その文を読み込んでアノテーションが行える.

　また，文 ID が予め分かっている場合は，ID を直接指定して文を読み込むことも可能である.

テキストの追加機能　YCAT には，テキストの追加機能が実装されており，アノテーターは収集したテキストを，コーパスに追加することが出来る．YCAT では，追加されたテキストに対して，形態素情報，文節区切り情報，係り受け情報の付与を行い，コーパスに追加する．追加されるテキストには，収集したアノテーターが分野を与えるようにしており，コーパスを共有する他のアノテーターは，分野の一覧を見ることで，どの部分が追加されたのかを確認することが出来る.

係り受け解析のデモ機能　YCAT では，YWT コーパスから学習したモデルを使用した，係り受け解析のデモを行うことが出来る．入力フォームから文字列を入力すると，入力された文字列に対して係り受け解析を実行し，解析結果をツリー形式で表示する．形態素情報は，表示されたツリーの文節をクリックすることで見ることが出来る.

3　おわりに

　本稿では，我々が作成するコーパスアノテーションツール（YCAT：Yahoo Corpus Annotation Tool）の紹介を行った.

参考文献

[1] 黒橋禎夫, 長尾眞. 京都大学テキストコーパス・プロジェクト. 言語処理学会第 3 回年次大会発表論文集, pp. 115 – 118, 1997.

[2] 松本裕治, 浅原正幸, 橋本喜代太, 投野由紀夫, 大谷朗, 森田敏生. タグ付きコーパス管理／検索ツール「茶器」. 言語処理学会第 12 回年次大会発表論文集, pp. 460 – 463, 2006.

[3] Lance Ramshaw and Mitch Marcus. Text chunking using transformation-based learning. In *Proc. of the Third Workshop on Very Large Corpora*, pp. 82 – 94, 1995.