アノテーションツール “Tagrin” の紹介*

高橋 哲朗† 乾 健太郎††
† 富士通研究所
†† 奈良先端科学技術大学院大学 情報科学研究科
takahashi.tet@jp.fujitsu.com
inui@is.naist.jp

1 はじめに

自然言語処理の研究において人手によるテキストへのアノテーションは重要な役割を持つ．たとえば機械学習における学習事例の作成や統計量の獲得，規則抽出，そして評価型ワークショップにおける正解データの作成などにおいて，人手によるアノテーションや付加情報の修正は不可欠である．

本稿では我々の作成したアノテーションツール Tagrin を紹介する[1]．Tagrin は情報抽出や照応解析など多様なタスクにおけるコーパス作成を目的としており，ユーザの定義したチャンクのラベルや関係の付与，ブラウズ，編集を可能としている．本稿ではこの実装例を元にテキストへのアノテーションの種類について整理する．

2 アノテーションの整理

本節ではテキストへのアノテーションにおいて付与する以下の 3 つの情報について議論する．

- チャンクの範囲
- チャンクのラベル
- チャンク間の関係

なお，これらの情報が付与されたチャンクをここではタグと呼ぶ．

2.1 チャンクの範囲とラベル

もっとも基本的なアノテーションは，チャンクの範囲の指定 (セグメンテーション) とそのチャンクへのラベル付与である．

アノテーションを行なうためにはテキストの特定の個所を指定する必要がある．アノテーションの対象となるテキストは一次元のデータであるので，チャンクの開始地点と終了地点を指定すればよい．

テキストの一部ではなく全体に対してアノテーションを行なう場合も考えられる．たとえば文書分類の問題では文書そのものに “政治” や “スポーツ” などの情報を付与する．この問題は，チャンクの範囲がテキスト中の一部分ではなくテキスト全体だと見ることで一般化できる．

次に，指定したチャンクを他と区別するために何らかのラベルを与える必要がある．ラベルは 1 つのチャンクに対して最低 1 つは付与しなければならない．

一般的にこのラベルは階層を持つことができる．たとえば，固有表現のラベルとして「人名」というラベルが考えられるが,「人名」の下位には「人名_ 女性」や「人名_男性」のようなラベルが考えられる．しかし，a) ラベルの階層はラベル間のオントロジーの問題でありアノテーションと直接の関係はない，b) ラベルの構成が複雑になるとアノテーションの作業効率が低下する，という二点を考慮し，Tagrin ではラベルの階層を扱う方法を提供していない．ラベルが階層を持つ場合はそれらを展開し列挙することで対応できる．

2.2 チャンク間の関係

複数のチャンクの間には，何らかの関係が存在し得る．関係には次節以降で述べる推移性と有向/無向という少なくとも 2 つの性質がある．

2.2.1 推移性

(1) において 吾輩[wagahai-1] と 猫[neko-1] は**同一指示**の関係にある．また，吾輩[wagahai-1] と 吾輩[wagahai-2] も**同一指示**の関係にある．このとき 猫[neko-1] と 吾輩[wagahai-2] も同様に**同一指示**の関係にあると言える．すなわちこれらの関係は推移的であると言える，

(1) 吾輩[wagahai-1] は 猫[neko-1] である。名前[namae-1] はまだ無い。どこで 生れ[umare-1] たかとんと見当がつかぬ。何でも薄暗いじめじめした所でニャーニャー 泣い[nai-1] ていた事だけは 記憶し[kiokushi-1] ている。吾輩[wagahai-2] は ここで[kokode-1] 始めて 人間というもの[ningen-to-iu-mono-1] を 見た[mita-1]。しかもあとで聞くとそれは 書生[shosei-1] という 人間[ningen-2] 中で一番獰悪な 種族[shuzoku-1] であったそうだ。

次に，見た[mita-1] と ここで[kokode-1] は**述語–項構造**の関係にあり，見た[mita-1] と 人間というもの[ningen-to-iu-mono-1] の間にも同様の関係がある．しかしこの二つの関係から ここで[kokode-1] と 人間というもの[ningen-to-iu-mono-1] に**述語–項構造**の関係は導き出せない．この例は先程の**同一指示**の例とは異なり非推移的であると言える．

*A multi-purpose corpus annotation tool: Tagrin.
TAKAHASHI Tetsuro†, INUI Kentaro††.
† Fujitsu Laboratories LTD.
†† Nara Institute of Science and Technology
[1]Tagrin は http://kagonma.org/tagrin/ にて公開している

このように，関係は推移的なものと非推移的なものとに分かれる．

2.2.2 有向/無向

2.2.1 節で挙げた**述語-項構造**の例において，見た[mita-1]とここで[kokode-1]は方向性を持っており有向な関係であると言える．たとえば「ここで→見た」という関係は成り立つが，そのとき「見た→ここで」に同様の関係は成り立たない．

一方**同一指示**の例において，吾輩[wagahai-1]と吾輩[wagahai-2]，猫[neko-1]の関係には方向性がない．つまりこれらの関係は無向であると言える．

2.2.3 4 種類の関係

2.2.1 節，2.2.2 節 をまとめると，チャンク間の関係には少なくとも推移性と有向/無向という 2 つの性質があり，これらの組み合わせにより表 1 に示す 4 種類の関係が考えられる．

表 1: 4 種類の関係

	推移的	非推移的
有向	a	c
無向	b	d

それぞれについて例を見る．

a 推移的有向関係

(2) において,「車」,「ドア」,「ガラス」の間には**部分–全体**の関係があり，この関係は推移的有向関係である．

(2) この車はドアのガラスに UV 加工が施してある。

また (1) における，猫[neco-1]と名前[namae-1]の間の**ブリッジング (「A の B」)** の関係もこの種類に属する．

b 推移的無向関係

推移的無向関係は集合を表わす関係である．

(3) において三四郎[sanshirou1]と こころ[kokoro1] は**並列**の関係にあり，この種類に属する．

(3) 彼は最近三四郎[sanshirou1]と こころ[kokoro1]を読んだ。

また 2.2.1 節の (1) における**同一指示**の関係もこの種類の関係に属する．

c 非推移的有向関係

(1) において，人間[ningen-2]は属性として種族[shuzoku-1]を持っておりその値は書生[shosei-1]である．このような**属性–値**の関係は非推移的かつ有向的である．

また 2.2.1 節で挙げた **述語–項構造**もこの種類に属する．

d 非推移的無向関係

非推移的無向関係は対を表わす関係である．

たとえば試合の対戦相手の関係や，企業合併における 2 つの企業の間の関係などが挙げられる．

このように上記の 4 種類の関係によりチャンク間の多様な関係を表現でき，様々なタスクへの応用が可能となる．

3 アノテーションツール Tagrin

本節では，我々の作成したアノテーションツール “Tagrin” の概要を紹介する．Tagrin はこれまでに議論してきたチャンクの範囲，ラベル，そして 2.2 節で挙げた 4 種類の関係を編集・可視化する機能を持つ．図 1 に Tagrin の概観を示す．

タグの付与は，マウスでテキストを選択し，あらかじめ設定したキーまたはボタンを押すことにより行なえる．

関係を扱うために，Tagrin では フォーカスという概念を取り入れている．たとえば図 1 では，(03) 行の「里谷多英選手」がフォーカスされており，この状態で (07) 行の「金メダリスト」にタグを付与することで，これらの間に関係を付与することができる．

Tagrin においてチャンクの可視化は，範囲とその種類を色で表わすことにより行なっている．また関係の可視化には上記のフォーカスを用い，フォーカスしたタグとそのタグに張られている関係をハイライトする．たとえば，図 1 では，(03) 行の「里谷多英選手」がフォーカスされており，(07) 行の「金メダリスト」，(09)，(14)，(17) 行の「里谷選手」が関係付けられている．

関係のハイライトは，2.2 節で挙げた 4 種類の関係を考慮したハイライトとなっている．たとえばこの例において (14)，(17) 行の「里谷選手」はそれぞれ (03) 行の「里谷多英選手」と直接は関係付けられていない．しかし以下の関係が付与されているので，(03) 行の「里谷多英選手」がフォーカスされたときに (14)，(17) の「里谷選手」も推移的にハイライトされる．

- (03) 行の「里谷多英選手」 – (09) 行の「里谷選手」
- (09) 行の「里谷選手」 – (14) 行の「里谷選手」
- (09) 行の「里谷選手」 – (17) 行の「里谷選手」

「里谷選手」の関係は 2.2 節で述べた b 推移的無向関係の関係である．有向関係については，関係の向きが考慮されハイライトが推移する．また非推移的関係については，直接関係を持つタグだけがハイライトされる．

Tagrin ではタグの数や種類，関係，キーバインド，画面上の色などは設定ファイルにより自由に設定できるようになっており，様々な用途に用いることができる．

Tagrin はこれまでに，ゼロ照応のタグ付与 [3]，名詞句照応のタグ付与 [4]，因果関係の出現特性の調査 [5]，評価値表現へのタグ付与 [8], 意見タグ付きコーパスの作成 [7], 意見情報抽出 [2], 動作性名詞の項構造解析 [9]

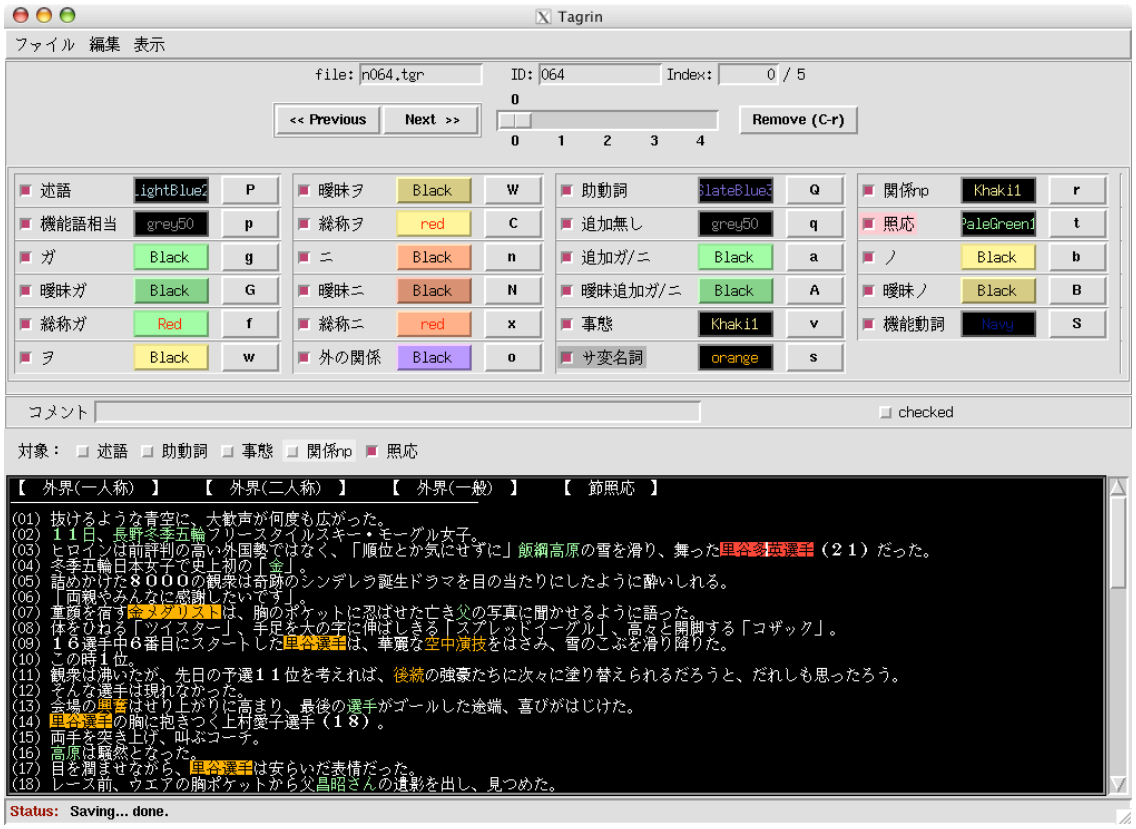

図 1: Tagrin のスクリーンショット

などのタスクに用いられてきた[2].

Tagrin はその他に以下の特徴を持っている.

- **SGML 形式のファイルのインポート／エクスポートが可能**
  すでに付与されたタグの情報をインポートし，それらのタグ間への関係の付与や，新たなタグの付与が可能となっている．たとえば NE チャンキングを機械的に行ない，その結果に対して関係を付与するといった使い方が可能である．また関係もインポート可能であるので，係り受け解析の結果をインポートし，格構造のアノテーションを行なうといった使い方もできる.
- **他のプログラムとの連結**
  アノテーションすべきタグが複雑になると，テキスト中の情報だけでは判断できない状況が起こりうる．そこで Tagrin ではアノテーションに応じた情報を標準出力へ出力する機能を持っており，他のプログラムと連結させることが可能である．たとえば UNIX においては次のようなコマンドにより，入力ファイル `input.tgr` への編集が逐次 `dic_lookup.prl` に渡され，この外部プログラムにより辞書引きを行ない，その結果を見ながらアノテーションするといった使い方が可能である.

  ```
  % ./tagrin.tcl input.tgr | dic_lookup.prl
  ```

- **テキストの編集**
  インポートしたコーパス中のタグの情報だけでなく，テキストそのものに誤りがあり，アノテーションの作業中にそれを発見する場合がある．そのような際には必要に応じてテキストを編集できる.
- **タグの表示/非表示切り替え**
  図 1 のようにタグの種類が増えた場合，すべてを表示するとアノテーションがやりづらくなる．また，複数の異なる種類のタグを同時に付与する場合などにはタグが重複し可読性が下がる．そこで Tagrin では特定のタグのみを表示する機能を持たせている.
- **タグのコピー**
  タグを追加する方法には，a) 選択する，b) コピーする，の二種類の方法がある．a) は本節の前半で説明した方法である．b) はすでに付与されたタグの範囲をコピーすることにより，チャンクの範囲を指定することなくタグを付与できるというものである．この機能を使って，たとえばすべての名詞句に対してチャンクの候補としてのタグを付与しておくことにより，名詞句の上でクリックしカー

[2] これらのコーパスの詳細は `http://cl.naist.jp/\%7Einui/opinion-mining/` を参照されたい

ソルを置き，あらかじめ設定されたキーを押すという短い操作でのアノテーションが可能となる．

- **Undo が可能**
- **Tcl/Tk で実装しておりプラットフォームフリー**

4 既存のアノテーションツール

橋田らの大域文書修飾 (GDA[3]) [1] の作成には，Emacs 上で動くタギングエディタが使われていた．このタギングエディタは GDA に基づいたタグ付きコーパスを作成することを目的とし，その目的に特化しているため他の目的に用いることが難しい．たとえば GDA ではタグの交差が許されていないが，タスクによってはタグの交差を必要とする場合も考えられる．たとえば (4) からは「*target*:統括部, *attribute*:部長, *value*:彼」というような情報が抽出され得るが，これらのチャンクは交差している．

(4) 彼は統括部長だ。

Tagrin と比較して GDA タギングエディタの優れた点には，木構造とテキストを同時に見ながらタグの編集をできる点や，タグ毎にその属性を編集できる点などが挙げられる[4]．

河原ら [6] は京大コーパス作成に用いたツールを拡張し，格関係，名詞間の関係，共参照などの様々な関係を付与できるようにしている．このツールは GUI を備えておりまた構文木を表示できるためタグのブラウズや修正に優れている．しかし GDA のタギングエディタと同様に使用目的が定まっているため，汎用的ではない．

Tagrin と同様の機能を持つアノテーションツールとしては，現在以下に挙げるようなツールが公開されている．

- *LDC's ACE 2005 Annotation Toolkit* [5]
  LDC (Linguistic Data Consortium)[6]が開発したアノテーションツールであり，ACE (Automatic Content Extraction) で使われている．特にチャンクの属性や関係のアノテーションにおいて多機能であるが，ACE に特化しているため汎用的ではない．
- *Callisto*[7]
  ACE プロジェクトの "Event Task" におけるアノテーションに使われていた．設定ファイル (DTD) によりチャンクやそれらの間の関係を含めたタスクの定義が可能となっており，汎用に用いることができる．
- *PALinkA (Perspicuous and Adjustable Links Annotator)*[10][8]
  特定のタスクに非依存であり汎用的なツールである．ユーザの設定したチャンクのラベリングやリンクの付与が可能であり，これまでに，同一指示のアノテーションやセンタリング理論に基づくアノテーション，自動要約のためのコーパス作成などに用いられている．このツールの欠点の一つに単語の区切をあらかじめ与えておく必要があることが挙げられる．これは未知語などにより単語の区切がはっきりしない場合や，(4) の例のように単語内に区切が存在し得る場合に問題となる．

5 まとめ

本ツールは，タグ付きコーパス作成の必要があったときに汎用的なツールを見付けることができなかったため作成したものである．汎用性を重視したため特定のタスクに特化した深いアノテーションには向いていないが，3 節で示したように現在の機能だけでもこれまでに多数のタスクに活用できている．

今後，言語処理の研究において本ツールが少しでも役に立つことができれば幸いである．

参 考 文 献

[1] 橋田浩一, 長尾確, 内山将夫, Christoph J. Neumann, 高橋直人. Gda タグ集合の設計と応用. 言語処理学会第 5 回年次大会発表論文集, pp. 132–135, 1999.

[2] 廣瀬峰史, 乾健太郎, 松本裕治. レストランドメインにおける意見情報抽出. 言語処理学会第 12 回年次大会発表論文集, 2006.

[3] 飯田龍, 乾健太郎, 松本裕治. 文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定. 情報処理学会論文誌, Vol. 45, No. 3, 2004.

[4] 飯田龍, 乾健太郎, 松本裕治, 関根聡. 最尤先行詞候補を用いた日本語名詞句同一指示解析. 情報処理学会論文誌, Vol. 46, No. 3, pp. 831–844, 2005.

[5] 乾孝司, 奥村学. 文書内に現れる因果関係の出現特性調査. 計量国語学, Vol. 25, No. 3, 2005.

[6] 河原大輔, 黒橋禎夫, 橋田浩一. 「関係」タグ付きコーパスの作成. 言語処理学会第 8 回年次大会発表論文集, pp. 495–498, 2002.

[7] 小林のぞみ, 乾健太郎, 松本裕治. 意見情報の抽出/構造化のタスク仕様に関する考察. 情報処理学会研究報告 NL-171, pp. 111–118, 2006.

[8] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. 意見抽出のための評価表現の収集. 自然言語処理, Vol. 12, No. 2, pp. 203–222, 2005.

[9] 小町守, 飯田龍, 乾健太郎, 松本裕治. 共起用例と名詞の出現パターンを用いた動作性名詞の項構造解析. 言語処理学会第 12 回年次大会発表論文集, 2006.

[10] Constantin Orasan. Palinka: a highly customizable tool for discourse annotation. In *the 4th SIGdial Workshop on Discourse and Dialog*, pp. 39 – 43, 2003.

[3] http://i-content.org/gda/

[4] Tagrin はタグ毎にコメントを付与する仕組みは持っており，そこに属性を記述できるようにはなっているが，明示的に属性を与える仕組みはない．

[5] http://projects.ldc.upenn.edu/ace/tools/

[6] http://www.ldc.upenn.edu/

[7] http://callisto.mitre.org/

[8] http://clg.wlv.ac.uk/projects/PALinkA/