階層構造を持つタグセットのための タグ選択インターフェース

松井 信太朗　　野口 正樹　　飯田 龍　　徳永 健伸

東京工業大学 大学院情報理工学研究科
{smatsui,mnoguchi,ryu-i,take}@cl.cs.titech.ac.jp

1　はじめに

近年，自然言語処理の研究分野において，対象とする課題に関する情報 (タグ) を付与した結果を用いた統計的手法が，形態素解析のような基盤処理から情報抽出のような応用処理までさまざまな分野で成果をあげている．統計手法の品質はタグ付けされたコーパスの品質に依存しているため，コーパスへの網羅的な揺れのないタグ付け作業を実現すること自体にも研究者の関心が高まっている [3].

高品質なタグ付けを実現するためには，(1) 漏れなくタグ付けする対象を探索し，(2) 付与するタグを正しく選択し，(3) 付与した結果から誤りを検出する，などの作業が必要になる．本研究では (2) の付与するタグを選択する問題に着目し，特に階層構造を持つタグセットから適切なタグを選択する問題について考える．階層構造を持つ言語資源の例としては，WordNet [1]，分類語彙表 [5]，関根の拡張固有表現階層 [7] などがある．たとえば，これらの言語資源で定義されている意味クラスや固有名をタグとして付与することを考えると，階層の深さや同一階層の要素数が増えるほど全体を表示することが困難になる．このため，タグ付けの作業者があらかじめ付与したいタグ名を把握していた場合でも，効率的にそのタグを選択することが難しくなる．BOEMIE (Bootstrapping Ontology Evolution with Multimedia Information Extraction) プロジェクト [2] で利用されているオントロジーアノテーションツールや関根の拡張固有表現階層 [7] のタグ付けに利用されている Fuu Tag [1]では，Windows のエクスプローラで採用されている Folder Tree 形式のインタフェースを利用している．近年，情報視覚化の研究も進められており [4, 6]，Folder Tree 形式が階層構造の情報を付与するために必ずしも最適な選択とは言えない．

そこで，本研究では階層構造のタグセットから必要なタグを選択するためのインタフェースとして Folder Tree 形式の表示と WordNet の構造を可視化するために採用されている Hyperbolic Tree 形式の表示の 2 種類を用いて，実際に階層構造を持つタグセットについてタグ付け作業を行うことで，それぞれの特性について調査した．

本稿では，まず 2 節で階層構造を持つタグセットからタグを選択するための 2 種類の表現形式を説明し，次に 3 節でその表示形式を含めたタグ付けツールの実装について述べる．4 節で階層構造を持つ固有名タグ付けの課題を対象に，2 つの表示形式でどのように作業者の振舞いが異なるかを調査する．最後に 5 節でまとめる．

2　階層構造の表示形式

本研究で採用した Folder Tree 形式と Hyperbolic Tree 形式の 2 つの表示形式について，それらの特徴を以下でまとめる．いずれの表示形式においてもタグはグラフのノードとして表現される．

2.1　Folder Tree 形式

この表示形式は，フォルダの開閉により着目しているノードの子ノードの表示/非表示を切り替える機能を持つ表示形式であり，階層の深さがインデントの深さに対応するという特徴を持つ．このため，階層構造の深さが同一であるノードを把握しやすいという利点を持つ．また，Windows のエクスプローラなどでも導入されており，一般作業者が操作する機会が多く，利用開始時の負荷が小さいと考えられる．階層構造を持つタグセットを用いてタグ付けを行うことができる FuuTag や BOEMIE のツールなどはこの形式を採用している．フォルダの開閉によりすべてのタグへアクセスできるが，タグの種類が多い場合は画面内にすべて表示できないため，作業に必要となるタグが頻繁に変わると選択が困難になるという欠点を持つ．

2.2　Hyperbolic Tree 形式

この表示形式では，親ノードの周りを子ノードが放射状に取り巻く形で配置され，画面の中央ほど大きく，周辺に行くほど小さく表示し，中心から一定以上離れた位置に配置されたノードは表示されない．2 つのノー

[1] http://nlp.cs.nyu.edu/ene/

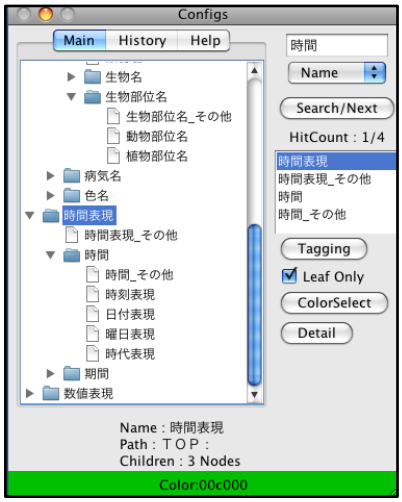

図 1: Folder Tree 表示

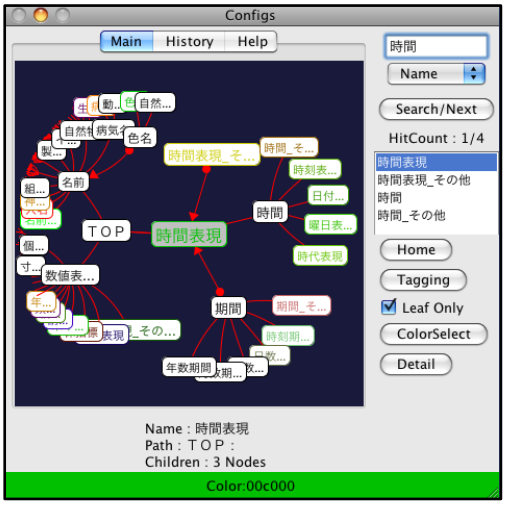

図 2: Hyperbolic Tree 表示

ド間の親子関係は矢印で表現され，矢印の先が親ノード，矢印の元が子ノードを表す．表示されている任意のノードを選択することで，そのノードを中心に再表示されるため，ノードの選択を繰り返すことによりアクセスしたいタグへ移動できる．また，画面上の任意の点を選択しドラッグすることで，見たい方向への視点の切り替えをより簡単に行うことができる．あるノードに対して多くの子ノードが存在する場合，Folder Tree 形式では子ノードを表示するとそれ以外の階層関係がわかりにくくなるのに対し，Hyperbolic Tree 形式では階層の上下が把握しやすいという利点を持つ．ただし，Folder Tree 形式と比較して一般のユーザが操作する機会はほとんど無く，作業開始時にはタグ選択の操作に慣れる必要がある．また，着目しているタグを中心に表示するため，タグ名を表示しただけでは今着目しているタグが全体のどの階層に位置しているのかを把握しづらいという欠点がある．

3　タグ付けツールの実装

インターフェースの違いによる作業効率，精度の違いを見るため Folder Tree 形式と Hyperbolic Tree 形式でタグ選択ができるタグ付けツールを実装した．今回実装したツールでは，タグ選択のインターフェース (Folder Tree 形式か Hyperbolic Tree 形式のいずれか) を表示するウインドウ，タグ付け対象となる文書を表示するウインドウの 2 つで構成されている．タグ選択ウインドウでは，タグ選択のインターフェースに加え，3.1 で後述するように，選択しているタグの補足情報の表示やタグ選択の支援となる操作を行うことができる．

3.1　タグ選択ウィンドウ

今回実装したツールのタグ選択ウィンドウのスナップショットを図 1 と図 2 に示す．

図 1 の Folder Tree 形式では，葉ノードはファイルのアイコン，中間ノードはフォルダのアイコンで表現される．また，画面内に表示できないノードについては，スクロールバーで移動することによりアクセス可能となる．

図 2 の Hyperbolic Tree 形式の実装には TREEBOLIC2 ライブラリ[2]を利用した．選択しているノードは背景を暗くすることで示され，画面右側の “Home” ボタンによりルートノードを選択できる．

また，このウィンドウでは以下のような情報を表示もしくは操作できる．

- **情報表示** : ウィンドウ下部に，選択しているタグに関して，葉となるタグでは親情報と使用例を，中間ノードのタグでは親情報と子ノードの個数を表示する．
- **タグ付け履歴**: “History” タブを選択することで，これまでに付与したタグのうち最大 20 個の履歴がリスト表示される．タグ選択インターフェースの代わりに，このリストからもタグを選択できる．
- **タグ検索**: 画面右上部に入力した文字列と，タグセット全体から日本語名，英語名，使用例のいずれかの部分文字列が一致するタグを検索できる．検索結果はタグの一覧リストとして表示し，このリストからもタグを選択できる．
- **タグ表示色の設定**: “ColorSelect” ボタンにより，タグ付け作業ウィンドウの表示色を設定できる．

[2]http://treebolic.sourceforge.net/

これによりテキストのどの位置にどのタグを付与したのかを視覚的に提示する．図 2 に示すように，HyperbolicTree 形式の場合はノードの色にも反映される．

- **タグの仕様の詳細**: “Detail” ボタンにより，選択しているタグの情報 (仕様や作業例など) を表示する．また，“Help” タブを選択すると個別のタグでなく，複数のタグ，あるいはタグセット全体に関連する注意事項が提示される．

3.2 タグ付け作業ウィンドウ

このウィンドウでは，タグ付け対象となる文章とその文章に付与されたタグのリストを表示する．文章中のタグを付与する範囲をマウスで選択し，タグ選択ウィンドウでタグを選択した後に “Tagging” ボタンを押すことで，タグを付与できる．タグのリスト表示部には既に付与されたタグの一覧を表示する．このリストはタグ名，付与された範囲の文字列などでソートができるため，誤ったタグを付与していないかのチェックを行うことができる．

4 比較実験

Folder Tree 形式と Hyperbolic Tree 形式の作業による影響を調査するために比較実験を行った．

4.1 実験設定

関根の拡張固有表現階層 [7] の version7.1.0 のタグセットについて，熟練の作業者二名が二つのタグ選択インターフェースを使って作業を行う．タグ付け作業中に作業者が行ったすべての動作は，その時刻と共に作業ログに記録する．また，本作業前に，タグ付けツールに慣れてもらうために，両作業者は両方のツールを用いて 10 文書ずつの練習作業を行った．

使用したタグセットは固有表現を詳細に分類したもので,「時間表現」,「人名」など 243 の固有名ノードからなる．そのうち葉ノードの数は 196 である．階層の深さは，最も深いところで 5 階層である．また，個別のタグの詳細情報やタグ付けの際の注意事項は 3.1 で述べた “Detail” ボタンを押すことで参照可能にし，HTML で記述されたオリジナルの仕様書の閲覧は禁止した．

実験に使う文書として特定研究「日本語コーパス」[3]で構築中の日本語コーパスのコア・データから 20 文書を選択した．これを 5 文書ずつ 4 つのグループに分け，それぞれの作業者が Folder Tree 形式を使用する場合と Hyperbolic Tree 形式を使用する場合の 4 つの組み合わせの試行で用いた．また，作業対象となる文章には，予め蓄積された過去の固有名タグ付けの履歴からパターンマッチによってある程度のタグを付与した状況で作業を開始する．

この比較実験では，タグ付け作業の作業時間，付与したタグの総数とその異なり数 (文書ごとに使用された異なりタグ数の平均)，以下に示す作業のタグ選択の一致率，不一致率，タグ情報の参照回数 (“Detail” ボタンが押された総数)，履歴からタグが選択された回数を比較する．タグ選択の一致率と不一致率は以下の式に基づいて算出する．

$$\text{一致率}_i = \frac{\text{両作業者が同一箇所に同一タグを付与した数}}{\text{作業者}_i\text{が付与したタグの数}}$$

$$\text{不一致率} = \frac{\text{両作業者が同一箇所に異なるタグを付与した数}}{\text{両作業者が同一箇所にタグを付与した数}}$$

4.2 結果と考察

表 1 に作業者 2 名の作業の比較結果をまとめる．FT_i は作業者 $_i$ が Folder Tree 形式で作業した場合を指し，同様に HT_i は作業者 $_i$ が Hyperbolic Tree 形式で作業した場合を表す．たとえば，FT_1-FT_2 は同じ 5 文書に対して両作業者とも Folder Tree 形式で作業をした試行を表わす．なお，FT_1-FT_2 と HT_1-HT_2 については，システムのバグのために両作業者の挙動が変わってしまった文書が 1 つずつあったので結果から除いた．これら二つのグループは残り 4 文書の結果を示している．

表 1 にまとめた内容とログの情報から以下に示す内容がわかった．

- 表 1 の FT_1-HT_2 と HT_1-FT_2 より，Hyprebolic Tree 形式を使用した方が付与されたタグの総数が多いことがわかる．これは，Hyperbolic Tree でタグを概観した方が付与すべきタグを把握しやすかったためだと推測される．

- タグ付けの総時間は一貫して作業者 1 の方が時間がかかっており，作業効率について表示形式の影響は観察できなかった．作業者 2 の方が作業時間が少ない理由としては，表 1 の履歴使用回数の差から分かるように，作業者 2 の方が作業を効率的に行うために履歴を多用していることがあげられる．今回の作業では，1 文書で使用するタグの異なりは多くてもタグセット全体の 1 割程度なので，どのような種類のタグが使用されるかが把握できればそれまでの履歴を用いてタグを選択する方が作業が効率的に進められることを示している．

[3] http://www.tokuteicorpus.jp/

表 1: 実験の結果

		FT_1-FT_2	FT_1-HT_2	HT_1-FT_2	HT_1-HT_2
総タグ数	作業者 1	374	798	871	361
(個)	作業者 2	361	830	805	331
総作業時間	作業者 1	3.01	4.94	7.42	3.89
(時間)	作業者 2	1.83	2.55	3.19	1.39
一致率	作業者 1	73.8%	81.2%	78.4%	83.1%
	作業者 2	76.5%	78.1%	84.8%	90.6%
タグの種類	作業者 1	19.5	22.0	24.0	19.3
(種類)	作業者 2	17.3	19.4	21.0	16.8
不一致率		10.97%	5.81%	5.40%	3.23%
タグ情報の参照回数	作業者 1	64	43	29	56
(回)	作業者 2	23	38	46	24
履歴使用回数	作業者 1	1	10	25	73
(回)	作業者 2	29	162	151	61

- 表 1 の一致率を見ると，作業者 2 名が Folder Tree 形式を使用した場合に比べ，共に Hyperbolic Tree 形式を使用した場合の方が高くなっており，また不一致率も減少している．これは，Hyperbolic Tree 形式で表示した場合には，例えばルートノードを表示している状況で視界に入ったタグから次に付与するタグを選択するといったバイアスがかかるのに対し，Folder Tree 形式の場合は画面に表示できるタグが隣接する兄弟ノードの個数によって制限され，より貪欲にタグを探索する必要があるために，作業者 2 人でタグ選択の差がでたのだと考えられる．

5 まとめ

本稿では階層的な構造を持つタグセットを用いてタグ付けをおこなう場合に，タグセットの表現形式の違いによってタグ付けの効率や作業の一致率にどのような影響が出るかを調査した．まず，階層構造を表示する形式として Folder Tree 形式と Hyperbolic Tree 形式をとりあげ，それぞれを用いてタグ付けをする 2 つのツールを実装した．これらのツールを使い，関根の拡張固有表現階層のタグセットを用いて 2 人の作業者が作業を行った結果を比較した結果，Hyperbolic Tree 形式のほうがタグ付けの揺れが少なくなるったことがわかった．ただし，今回行った比較実験では，2 種類のタグ選択インターフェースに加え，履歴や文字列検索を利用したタグの選択も可能であり，また実装の都合上，Hyperbolic Tree 形式に関してのみタグ名の表示を色分けして表示したため厳密な比較とは言えず，今後さらに調査を続ける必要がある．また，今回は階層構造のタグ付与について Folder Tree 形式と Hyperbolic Tree 形式の 2 種類の比較に着目したが，今後は他の表示形式についても吟味したい．

参考文献

[1] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech and Communication)*. MIT Press, 1998.

[2] Pavlina Fragkou, Georgios Petasis, Aris Theodorakos, Vangelis Karkaletsis, and Constantine Spyropoulos. Boemie ontology-based text annotation tool. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

[3] Kazuaki Maeda, Haejoong Lee, Shawn Medero, Julie Medero, Robert Parker, and Stephanie Strassel. Annotation tool development for large-scale corpus creation projects at the linguistic data consortium. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

[4] Ramana Rao, Jan O. Pedersen, Marti A. Hearst, Jock D. Mackinlay, Stuart K. Card, Larry Masinter, Per-Kristian Halvorsen, and george G. Robertson. Rich interaction in the digital library. *COMMUNICATIONS of the ACM*, Vol. 38, No. 4, pp. 29–39, 1995.

[5] 国立国語研究所（編）. 分類語彙表 (国立国語研究所資料集). 大日本印刷, 2004.

[6] 菊池司, 伊藤貴之, 岡崎章. Web ナビゲーション技術にみる情報デザイン・情報視覚化の最近の動向. 芸術科学会論文誌, Vol. 4, No. 1, pp. 1–12, 2005.

[7] 関根聡, 竹内康介. 拡張固有表現オントロジー. 言語処理学会第 13 回年次大会ワークショップ「言語的オントロジーの構築・連携・利用」, pp. 23–26, 2007.