タグ付きコーパスの格納/検索ツール「茶器」*

松本 裕治, 高岡 一馬, 浅原 正幸, 乾 健太郎, 橋本 喜代太
奈良先端科学技術大学院大学 大阪女子大学
投野 由紀夫, 大谷 朗, Edson T Miyamoto, 森田 敏生
明海大学 大阪学院大学 筑波大学 総和技研

1 はじめに

コーパスに基づく自然言語処理の進展に伴い，品詞情報や統語情報，さらに詳細なタグ付きテキストの蓄積が進んでいる．また，種々の統計的言語解析システムの進歩により，テキストに対する自動タグ付与がかなりの精度で行えるようになってきた．

タグ付きコーパスは，言語学/言語処理研究の基本データとして資するだけでなく，統計的機械学習に基づく言語処理の高性能化のためにも貴重な資源である．前者のためには，柔軟な検索機能および統計解析を行うための加工機能を持ったタグ付きコーパス支援システムの存在が重要である．コーパスへのタグ付与はもちろん正確であることが要求されるが，特に，後者の機械学習にとっては，コーパスのタグ付与誤りは，システムの性能に直接影響を及ぼす．人手により長い時間と手間をかけて構築されたタグ付きコーパスであっても，一定のエラーや不整合が残されていることが知られている．また，辞書や品詞体系の変更など，タグ付けの定義そのものの変更により，コーパスへの修正が必要な状況が生じることもあり，継続的な修正作業や不整合の除去は，タグ付きコーパスの利用者からの切実な要望である．

本稿では，タグ付きコーパスの格納および検索を支援する目的で今年度から開始した「茶器」というツールの開発目標と現状について説明する．

2 背景および関連研究

テキストコーパス (タグ付き/タグなし) 内の文字列や単語を検索し，KWIC(keywords in context) 形式で表示するシステムは，既に数多く存在する．例えば，Corpus Wizard[1] や KWIC Concordance[2]は，英語文書の単語検索を行い，KWIC 表示するシェアウェアあるいはフリーウェアである．WordSmith[3] は，品詞タグ付きコーパスの検索/KWIC 表示を行うことができる．British National Corpus[4]には SARA-32 という品詞情報を利用した専用の検索システムが添付され，検索結果の KWIC 表示や簡単な統計処理が可能である．しかし，品詞以上の情報を考慮したタグ付きコーパス検索システムは，特定のコーパス，あるいは，特定の品詞体系専用であることが多い．日本語については，Text Finder, KWIC Center, KWIC Finder[5]なども文字列や単語の検索/KWIC 表示を行うことができる．しかし，日本語のタグ付きコーパスを柔軟に検索できるフリーのツールはまだないと言ってよい．

以前，我々は，品詞，文節，係り受けといった統語情報をタグ付けされたコーパスを柔軟に検索するシステムを関係データベースを利用した検索システムとして実装した [1]．また，ドイツ語の Treebank を対象にした同様の提案もある [2]．我々の今回のプロジェクトでは，品詞，文節等のチャンク，係り受け等の統語情報を含むタグ付けコーパスに対して，柔軟な検索機能を備えるだけでなく，各種統計解析機能や辞書とコーパスの連携，タグ付けエラーの修正などの機能を持ったタグ付きコーパスの検索/管理システム [3] を目指している．現在は，文字列と単語列を対象とする検索機能と簡単な単語統計機能を実装している．

*ChaKi:Storing/Retrieving Tool for Annotated Corpora, Yuji Matsumoto, Kazuma Takaoka, Masayuki Asahara, Kentaro Inui, Kiyota Hashimoto, Yukio Tono, Akira Ohtani, Edson T Miyamoto, Toshio Morita

[1] rd.vector.co.jp/soft/dl/win95/util/se027330.html
[2] www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic.html
[3] www.lexically.net/wordsmith/
[4] www.natcorp.ox.ac.uk/
[5] これらのツールは，Web site が今後変わることもあり，Web 検索ツールで見つけることが容易であるので，ここでは表示しない．

3 タグ付きコーパスの格納/検索ツール「茶器」

3.1 システムに求める機能

タグ付きコーパスの利用と作成には様々な要件が考えられる．利用については，種々の視点からコーパスの検索が可能であること，および，検索結果の加工や統計処理を行う必要が生じる．検索要求は，単語の綴りそのもの，単語の一部，単語の原形，品詞などの文法情報，および，それらを組み合わせたものを扱うことが好ましい．

タグ付きコーパス作成については，コーパス全般に渡っての整合性と，辞書や想定する品詞体系に対する整合性を維持することが重要である．例えば，同じ単語の同じ用法には同じ品詞やその他の統語情報が付与されなければならないし，コーパスに存在する単語は辞書に存在する語であるべきである．実は，このような一見当たり前のことが成り立っていないことが多い．ベースとなる辞書や品詞体系等の文法情報の見直しが，コーパスにはただちに反映されなければならない．また，システムは日本語や英語を始めとする様々な言語コーパスに対して利用可能であり，さらに，異なるタグ体系についても対応できるものであることが好ましい．

本稿で紹介するのは，我々が３年間のプロジェクトとして開発しているシステムの初年度の成果である．プロジェクトとして目指しているシステムにどのような機能を持たせようとしているか，その全体像について以下に示し，次節で現状について述べる．

タグ情報の詳細度

I. 文字列 ：言語の構成要素となる任意の文字列，および，基本的な正規表現による検索

II. 単語： 単語の全体または一部，読み，品詞，活用情報，原形など，単語が持つ様々な文法情報を指定した検索

III. 複合語，固有表現，文節，基本句： 複合語等チャンクの全体あるいはその部分構成要素による検索．日本語の文節や英語の基本句などを指定した検索．日本語の品詞タグ付きコーパスでは，例えば,「教科書」が一語で単語と扱われているのか,「教科」と「書」に分かれているのか，利用者には事前にわからないかも知れない.「教科書」をこれら２つの要素からなる複合語と定義することにより，複合語全体と構成語の一部のいずれからも検索可能であることが好ましい．

IV. 統語： 統語構造の部分構造を指定した検索．統語解析の結果は文法に依存して異なるので，本プロジェクトでは，単語あるいは文節 (基本句) からなる係り受け解析木を対象とし，その部分係り受け構造を指定した検索機能を対象とする．

システムに求める機能

a. 検索： 上記タグ情報を検索要求として用いることができ，検索結果を KWIC のような形で表示する

b. 統計： 検索結果に対し，語形や品詞などの視点を指定して頻度や比率などの統計情報の提示する機能．検索結果と周辺の他の単語やパターンとの共起に関する統計を取る機能

c. 辞書との連携： 指定した辞書に含まれない語の検出や辞書内容の変更時にタグ付きコーパスの修正をうながすための情報を提示する機能

d. コーパスの整合性維持 コーパス中の不整合を検出したり，不整合部分を一括修正する機能．

その他の要件

多言語対応： 特定の言語に限定せず，様々な言語のコーパスにも対応可能であること．現在は，日本語と英語を対象にしている．

文法情報のカスタマイズ： 品詞体系の違いなど文法情報を利用者が変更することができること．

配布に対する制限： 配布が自由なソフトであることが好ましい．現システムでは，フリーの関係データベースシステム MySQL[6]をコーパスの格納と検索に利用しているが，その他の部分はプロジェクト内で構築しており，システム全体としてもフリーソフトウェアとして配布する予定である．また，プラットホームとしては，利用人口の多い Windows 上で稼動するシステムとして開発している．

言語処理ツール： 残念ながら，共通に利用可能なタグ付きコーパスの数は多くない．現実の利用としては，利用者が個別に持っているタグ付きコーパス，あるいは，利用者が持っている生テキストに自動タグ付与したコーパスを対象に検索や統計を取ることが多いと考えられる．現時点では，茶筌 [4] と南瓜 [5] を日本語の品詞タグ付けと係り受け解析に用いる予定であり，これらのシステムの解析出力をそのまま取り込めるように考えている．また，英語データのタグ付けの自動化のために，茶筌の英語対応と英語辞書の作成に着手している．

[6] www.mysql.com/

図 1: 文字列検索結果の KWIC 表示

3.2 システムの現状

前節で挙げた諸機能のうち，現在完成しているのは,「タグ情報の詳細度」では，I.，II. の文字列および単語列による検索要求の指定,「システムの機能」としては，a. の検索結果の KWIC による表示および b. の簡単な統計処理である．現在は，日本語と英語の品詞タグ付きコーパスに対応しており，日本語は茶筌の出力，英語は Penn Treebank を対象に検索が可能である．検索結果のスナップショットを図 1，図 2 に示した．これらの図を参照しつつ，現在のシステムの概要を説明する．

1. 利用者は，まずコーパス名を指定し，次に String Search(文字列検索) か Advanced Search(現在は，品詞タグ付きコーパスに対する検索) を選択する．
2. String Search(図 1) では，利用者は任意の文字列を入力して検索を行うことができる．英語に対しては大文字と小文字を区別するかどうかを選択することができる．また，文字を用いたなる簡単な正規表現 (否定はサポートしていない) による検索も可能である．図 1 は,「通常国会」という文字列で RWCP コーパスを検索した場面を示しており，結果は KWIC 表示される．
3. Advanced Search(図 2) では，任意の単語列を検索対象にすることができ，各単語は,「表層綴り」「読み」「発音」「原形」「品詞」「活用型」「活用形」(すべて正規表現を使用可) を自由に用いて，検索したい単語列を指定できる．図 2 の上半分に見える箱が個々の単語を表しており，箱の中の各行が上に列挙した情報に対応している．箱はいくつでも表示させることができ，前後の文脈を指定することができる．図 2 では,「国会」という単語の直前に「名詞」[7]を品詞の最上位分類として持つ単語が現れるパターンを検索している．結果は図のように KWIC 表示されるが，2 つの単語のいずれを KWIC の中心に持ってくるかは，自由に決めることができる．KWIC には，各単語の下に品詞情報が表示されているが，この例では，品詞の最上位の分類のみ示されている (茶筌が用いている IPADIC では，最大 4 階層の品詞細分類が定義されている)．この表示は off することもできるし，詳細な品詞表示を行うこともできる．上記に示した情報を KWIC に表示するかは，利用者が指定することができる．多くの情報を表示させると，各単語がが数行を占めてしまい，多くの文を表示することができない．そのような場合には，少ない情報のみを常時表示させるようにし，各単語をマウスでポイントする毎にその単語の情報だけをポップアップ表示するよう選択することもできる．

両検索モードで共通の機能として次のようなものがある．

- いずれの検索モードでも，検索ヒット数が表示さ

[7]品詞の分類は「名詞-固有名詞-人名-…」のようにハイフンによって階層定義される．図では最上位の品詞が「名詞」であり，それ以降は任意であることが正規表現によって記述されている．品詞名はポップアップメニューによって指定可能なので，品詞入力の誤りは生じにくくなっている．

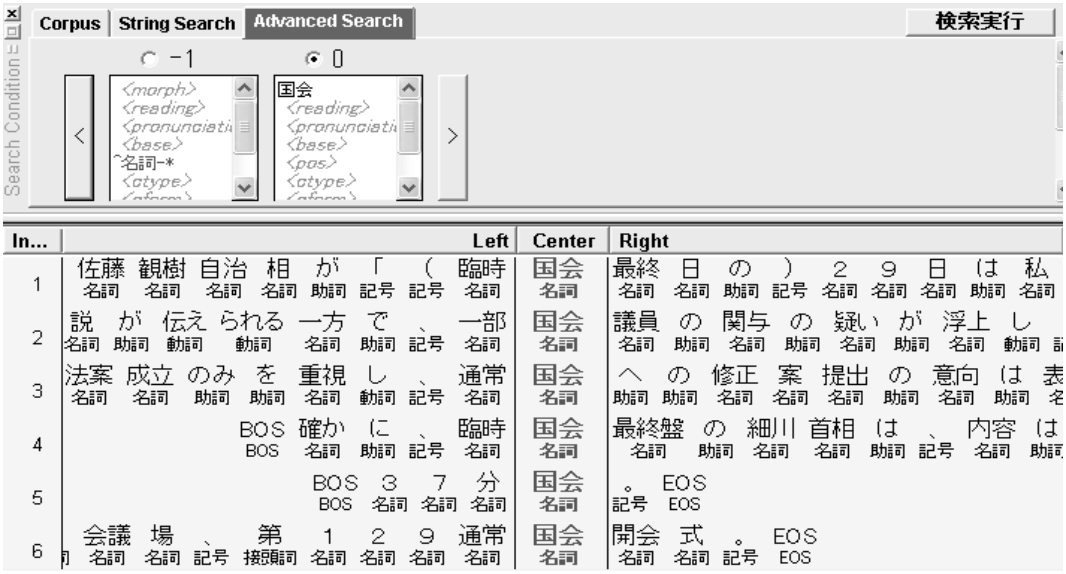

図 2: 文脈を指定した単語検索結果の KWIC 表示

れた後でデータベースから検索インタフェースへの転送が開始されるので，ヒット数が大量の場合は，転送を途中で中止させて，より絞り込んだ検索要求を記述し直すことができる．

- KWIC の Center の位置で文字列ソートを行うことができる．
- 検索結果を 1 単語ずつ左右へシフトすることができ，最初の検索要求の中心の単語から左右へ任意の数だけ離れた単語を KWIC 表示の Center 位置へ移動することができ，そこでソートすることができる．(String Search では単語の区切りがないので，この機能はほとんど意味がない)
- KWIC の Center 要素について，何を同一視するかを指定した上で頻度統計を取ることができる．

現在の実装は，MySQL によってタグ付きコーパスを格納し，ユーザインタフェース部は VisualC++，MySQL への検索要求の生成と結果の表示等の処理は Ruby を用いている．

4 おわりに

現在開発中の，タグ付きコーパスの管理と検索を行う汎用のツール「茶器」の開発指針と現状について紹介した．本プロジェクトは 3 年計画の初年度であるが，今後，フリーソフトウェアとして適宜公開し，利用者からのフィードバックを得る予定である．「茶器」のダウンロードと今後の予定につては，`http://cl.aist-nara.ac.jp/`の「自然言語処理のためのツール」から情報が入手可能である．

謝辞: 本研究は，文部科学省科学技術研究補助金 基盤研究 B「言語研究のためのコーパスの作成と利用に関する研究」(研究期間:平成 15 年度〜17 年度, 課題番号: 15300046) の支援を受けて行ったものである．

参考文献

[1] 工藤拓、松本裕治, “RDB を利用したタグ付きコーパス検索支援環境の構築,” 情報処理学会自然言語処理研究会 2001-NL-144, pp.135-142, 2001.

[2] Kallmeyer, L., “A Query Tool for Syntactically Annotated Corpora,” EMNLP/VLC-2000, pp.190-198, 2000.

[3] 浅原正幸，他,「語長変換を考慮したコーパス管理システム」, 情報処理学会論文誌 Vol.43,No.7, pp.2091-2097, 2002.

[4] 松本裕治,「形態素解析システム『茶筌』」, 情報処理, Vol.41, No.11, pp.1208-1214, 2000.

[5] 工藤拓, 松本裕治,「チャンキングの段階適用による日本語係り受け解析」, 情報処理学会論文誌 Vol.43,No.6, pp.1834-1842, 2002.