タグ付きコーパス検索ツール「茶杓」

谷口 雄作, 新保 仁, 浅原 正幸, 松本 裕治

奈良先端科学技術大学院大学 情報科学研究科

{yuusaku-t, shimbo, masayu-a, matsu }@is.naist.jp

1. はじめに

現在，約 1 億語の収録を目指した大規模な日本語コーパス構築プロジェクト[1] が国立国語研を中心に行われており，我々の研究室は種々の文法情報のタグ付けや処理ツールの開発に関わっている．コーパスの構築には多分野に渡る多くのチームが携わっており，コーパスを用いた言語研究を行うための支援環境の構築が重要課題となっている．また将来コーパスの公開に伴い，言語学的な要求を満たす検索や統計処理といった，有用な情報を一般利用者に提供するインターフェースが必要になる．

本稿では，現在開発を行っているタグ付きコーパス検索ツール「茶杓」について述べる．本システムは，このプロジェクトに関わる研究者，並びに言語処理及び言語学と関連分野の研究者を対象に，柔軟な検索と容易な操作を実現することを目的として構築されたものである．

同様のツールとしては，我々のグループで既に開発が進んでいるコーパス管理・検索ツール「茶器」[2] がある．このツールは個々の利用者が独自のコーパスを持ち，その検索と修正を行うことを目指して実現されたものであり，各ユーザがそれぞれコーパスを操作できる環境を事前に用意する必要があった．本システムはWebブラウザを用いてコーパスを操作するWebアプリケーションであり，このような事前の用意を必要とせず，操作を行うことが出来る．また，不特定多数のユーザへのコーパス公開にも有用である．

本稿はまず本システムの機能の紹介とその詳細について述べ，500万語規模の日本語コーパスに対して検索速度の測定実験を行った結果について報告する．最後に現在の課題について考察し，今後の方針について述べる．

2. 機能紹介

本節ではタグ付きコーパス検索ツール「茶杓」が備えている機能を紹介する．現在本システムが行える検索機能は主に文検索・単語列の頻度統計・文集合内の非連結頻出系列パターンの獲得である．これらの機能について，インターフェースの機能も交えて紹介する．

本システムは，文字列または単語列を入力として与えることでコーパスに含まれる文を検索する事が出来る．検索された文はKWIC（KeyWord In Context）形式にユーザに表示するもので，文だけでなく，品詞や活用型などのコーパスにタグ付けられた情報を同時に閲覧することができる．

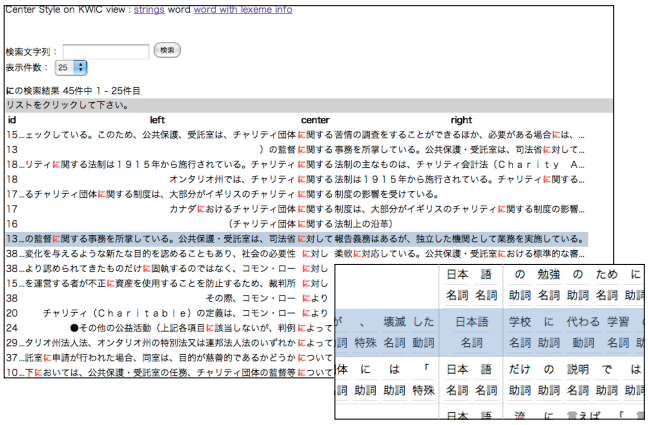

図 1　検索結果のKWIC表示

文字列による検索では，KWIC表示の中央に表示される語句を，入力文字列とその該当する単語の単位に切り替えることができ，また単語ごとにソートすることも可能である．図 1 は単語区切りに表示を行った例である．また，検索された各文の前後の文章の表示，入力文字列に該当する単語の検索，同種の品詞を持った単語に対して色別表示なども行うことができる．

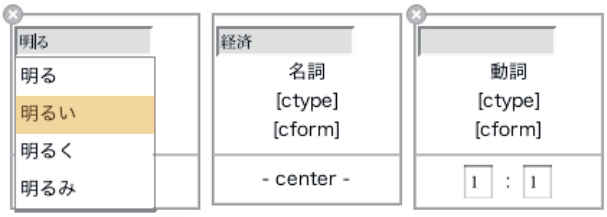

図２　単語列入力インターフェース

単語列による検索では，各語がもつ任意の情報を指定して連続あるいは非連続の単語列を検索することができる．例えば図２は表層形に「経済」を持った名詞を中心に，前に形容詞の「明るい」後に動詞を含む文を検索する際の条件指定の様子を示している．また，このインターフェースには表層形の入力補助としてオートコンプリート機能を備えている．

文の検索では結果件数が多い場合にどのような単語がどの程度出現しているかを瞬時には把握しにくい．本システムは単語列に対する頻度統計を取るができる． 図３は入力として「名詞＋銀行」を与えた場合の例である．最左列（Count）に出現頻度が表示されている．

	Center Lexeme				Right side Lexeme 1			
Count	Surface	POS	C type	C from	Surface	POS	C type	C from
39	の	名詞-固有名詞-地域-一般			銀行	名詞-固有名詞		
26	中央	名詞-固有名詞			銀行	名詞-固有名詞		
23	大和	名詞-数			銀行	名詞-固有名詞		
19	民間	名詞-固有名詞			銀行	名詞-固有名詞		
16	都市	名詞-固有名詞			銀行	名詞-固有名詞		
14	共同	名詞-一般			銀行	名詞-固有名詞		
14	開発	名詞-一般			銀行	名詞-固有名詞		
13	信託	名詞-一般			銀行	名詞-固有名詞		
12	三菱	名詞-数			銀行	名詞-固有名詞		
8	信用	名詞-一般			銀行	名詞-固有名詞		
7	輸出入	名詞-固有名詞			銀行	名詞-固有名詞		
7	や	名詞-固有名詞-地域-一般			銀行	名詞-固有名詞		
7	大手	名詞-固有名詞			銀行	名詞-固有名詞		
5	住友	名詞-数			銀行	名詞-固有名詞		
5	世界	名詞-固有名詞			銀行	名詞-固有名詞		
5	全国	名詞-固有名詞			銀行	名詞-固有名詞		
4	地方	名詞-固有名詞			銀行	名詞-固有名詞		
3	救済	名詞-一般			銀行	名詞-固有名詞		
3	する	名詞-サ変接続	特殊・ダ	連用形	銀行	名詞-固有名詞		
3	省	名詞-固有名詞			銀行	名詞-固有名詞		
3	投資	名詞-一般			銀行	名詞-固有名詞		

図３　単語列に対する頻度統計

これらの機能によって抽出された単語列は，文中のひとつの固まりとして，より大きなパターンの中で使われることがある．例えば「名詞＋を＋動詞」の周辺に出現するパターンには図４のようなものも起こると考えられる．

最後に作業を終えたら・・・しましょう
危険を犯したら・・・になるだろう
最新の機器を使えば
また結論を言えば・・・もするだろう

図４　非連結頻出系列パターン例

本システムは，コーパスに含まれる特定の文集合から，このような非連結頻出系列パターンを探し出すことが出来る．

これらの検索はタグ付きコーパスを関係データベースに格納して行っている．また本システムは言語に依存しておらず，英語・日本語・中国語のコーパスが利用できる．

3. システムの詳細

本システムの詳細について述べる．まずタグ付けコーパスを格納するデータベースについて検索に関連するテーブルの概説を行い，各検索機能の基本的な検索の流れについて述べる．

3.1.データベース

本システムが利用するデータベースの各テーブルはコーパス管理ツール「茶器」が使用するテーブルの仕様に準拠している．図５に本システムが利用する各テーブルの関係を示す．

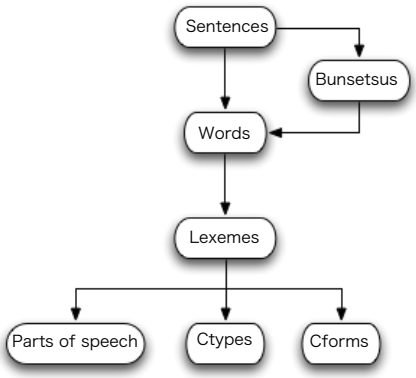

図５　検索対象となるテーブルの関係図

各テーブルはテーブル間の関係を示すIDを持っている．例えば文内の単語情報を受け持つWords テーブルには各文のテキストを持つ Sentences ・文節の情報を受け持つ Bunsetsus ・語彙の情報を持つ Lexemes の各テーブルのIDを持っており，Lexemes は各単語の情報として Part of speech（品詞）・Ctypes（活用型）・ Cforms（活用形）を持つ．

Words テーブルに登録されたあるアイテムを指定すると，そのアイテムがどの文のどの文節のどの単語を持ったアイテムであるかを特定することができるといった仕様になっている．

3.2.文字列による文の検索

本システムがデータベースに対して行う文検索は，基本的に結果件数の獲得と表示に必要な情報の獲得の２つのステップから成る．

文字列による文検索の際に生成する検索要求の概略を図６に示す．文字列を入力として受け取った場合，Sentences テーブルに登録されたテキストに対して全文検索を行い，文字列を含む文章を検索する．また，単語情報や文節情報の表示といったユーザの要求に合わせて表示に必要な情報を持ったテーブルを左結合し，同時に検索を行っている．

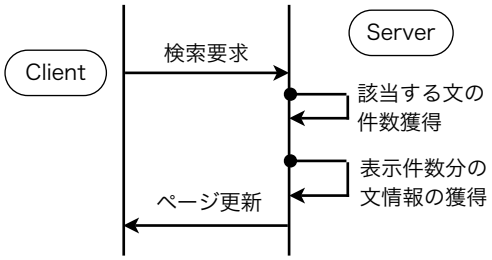

図６　文字列による文の検索

3.3.単語列による文の検索

単語列による検索では，Wordsテーブルを中心に検索を行う．また品詞や活用型などの情報は検索を行う前に事前獲得しておき，検索の範囲を狭くしている．

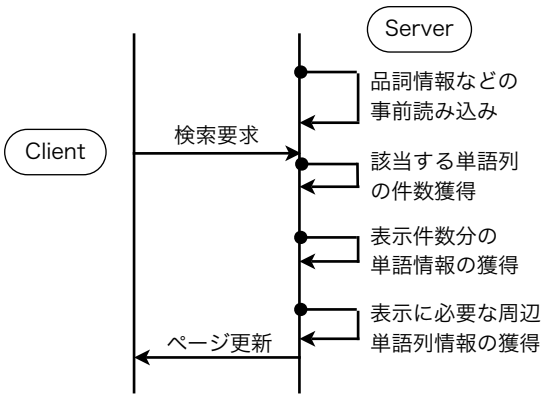

図７　単語列による文章の検索

まず入力された各単語の検索結果を内部結合した結果から件数とKWIC表示の中心語となる単語の情報を取得した後，表示に必要な前後の単語列の情報を取得している．

3.4.非同期通信による詳細情報の獲得

前後の文章の獲得や入力に該当する単語情報の獲得，表層形のオートコンプリート入力機能はサーバーとの非同期通信によって獲得している．

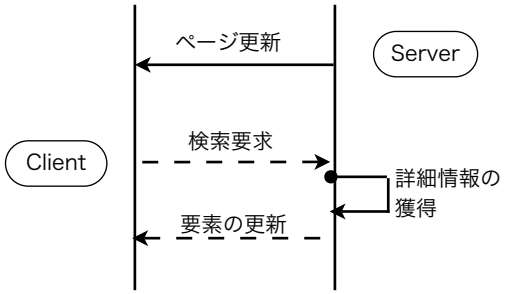

図８　非同期通信による詳細情報の獲得

この通信方法を利用することによって，ページ遷移によるページ全体の更新を行わず，変更したい要素のみを書き換えることができる．これにより，必要以上の描画処理を行わずに情報を提供できる．

3.5.単語列の頻度統計

この機能はユーザが入力した単語列の出現頻度を計算するものである．頻度統計を行うアルゴリズムは基本的に文検索と同じアルゴリズムによって行われる．

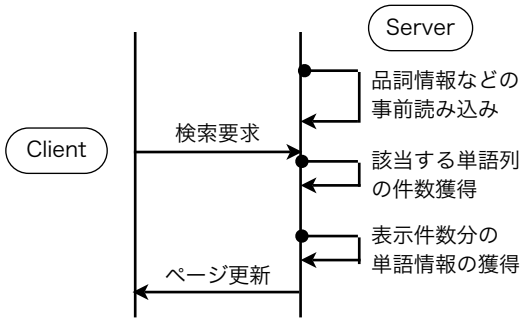

図８　単語列による頻度統計

KWIC表示の中心語となる単語を中心に各単語の検索結果を内部結合し，各異なり単語にグルーピングを行ってそれぞれの頻度を計算する．

3.6.文内の非連結頻出系列パターンの獲得

本システムではマイニングアルゴリズムBIDE[3] を用いて文内に含まれる非連続系列パターンを獲得することができる．文中の単語列の集合から頻出する単語列を抽出する．

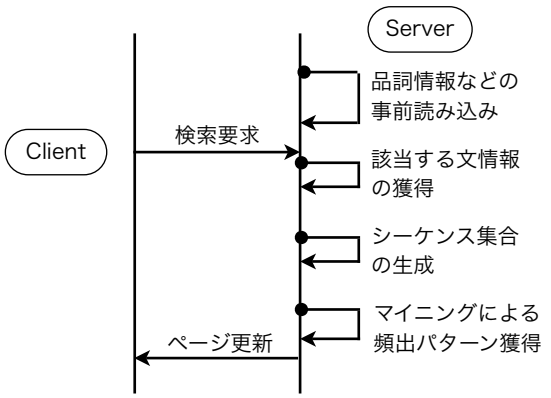

図９　非連結頻出系列パターン獲得

4. 検索速度の測定実験

文検索の機能について検索速度の測定を行う．実験に利用するコーパスは現代日本語書き言葉コーパスに収録予定の白書のデータを用いる．コーパスの概要を表１に示す．

表１ コーパスの概略

文字数	857万字
文数	127647文
単語数	5455385語
異なり単語数	37977語
単語の平均頻度	143.65

本実験では，このコーパス含まれる頻度145程度の単語を無作為抽出し，長さ１～３の単語列情報を生成，本システムに入力として与え，検索時間測定する．文検索機能に対する実験結果を表２に示す．

表２　文検索における平均検索速度

単語列長	平均検索時間（秒）	平均描画時間（秒）
1	10.42	0.0032
2	10.89	0.0028
3	10.6	0.0031

この表はデータベースに対する平均検索時間，検索結果の描画に必要とした平均時間を，検索を行った単語列の各長さごとに示している．この結果から，検索時間のほどんどはデータベースの検索に費やしている．また，現在の検索アルゴリズムは単語列の長さが長くなるにつれ，テーブル結合を行う数が増え，データベースに対する質問も複雑になってしまうが，結果をみると単語列の長さに関わらず，一定の検索速度を保っていることが分かる．

5. おわりに

本稿ではインターネットを介してタグ付きコーパスの検索を行うことができるツール「茶杓」について説明した．

今回行った実験から，データベースに対する検索速度の向上が必要であることが分かった．より多くの実験を重ね，速度低下の原因究明と，検索アルゴリズムの改善が課題として挙げられる．

今後の予定として，依存構造の検索や統計機能の充実，プログラムから検索結果を呼び出すWeb API及びユーザインターフェースの開発に力を入れて行きたい．

参考文献

[1] 代表性を有する大規模日本語書き言葉コーパスの構築
http://www.tokuteicorpus.jp/

[2] コーパス管理/検索ツール「茶器」
http://chasen.naist.jp/hiki/ChaKi

[3] J. Wang and J. Han, BIDE : efficient mining of frequent closed sequences.In Proc. 2004 Intl. Conf. on Data Engineering (ICDE'04)