R 言語を用いた言語情報処理教育*

堀内泰輔*1

Education of Language Information Processing using R Language

HORIUCHI Taisuke

In our school , for few years, linguistic information processing has been done from the aspect of general studies as an optional subject for 4th grade students. For the programming language for this practice, we selected the “R language” in place of the Perl language, last year. This new language have been developed aiming at the statistic calculation. But, this language has excellent respect more than other languages besides the scientific computation for the graph making and the character manipulation. In this thesis, our course content is described at first. Next, the utility is discussed about R language that seems not to be so familiar it.

キーワード：言語情報処理教育，R 言語，形態素解析，N-gram，MeCab

１．まえがき

本校では，４年生に一般科目の選択科目を設けており，筆者は「情報処理応用 A」ならびに「情報処理応用 B」という科目名で，言語情報処理と音楽情報処理関連の教育を行ってきてきた．なお，前者は前期，後者は後期の開講科目である [1)].

高専での情報処理関連の科目では，コンピュータやインターネットのリテラシー教育や，数値計算のような数値データをターゲットとする教育内容が多い．しかし，現在のコンピュータにおいては，インターネットの進化とともに，文字データや音楽データなどの，これまで高専での情報処理教育であまり日の目を見なかった種類のデータ処理が重要となってきている．このような状況から，数年前より上記の選択科目を開設した．

このように「情報処理応用 A」においては，言語をターゲットとした言葉の情報処理を教育するものであるが，実習用の言語としては，従来は Perl 言語を用いてきた．これは，Perl 言語が文字列処理に最適なことを理由に選択したものであるが，短期間の授業では暗号めいた Perl プログラムの作成は困難が伴うことを経験してきた．

そこで，昨年度からは，「R 言語」と呼ばれる統計計算用と一般に称される言語を採用することとした．

本論文においては，本科目の授業内容を述べるとともに，一般にはあまり馴染みがないと思われる R 言語について，その有用性を論じる．

２．授業の構成

表１に，昨年度における「情報処理応用 A」の授業内容を示す．中間試験前の前半では言語情報処理の重要なデータ源であるコーパス（巨大テキストファイル）を構築することを目標に，インターネットからの諸情報を効率的にダウンロードする方法を実習するとともに，言語情報の処理に欠かせない正規表現や KWIC 検索などの理解を重要視している．

コンピュータの OS 環境は，後述のように Windows 環境であるが，コマンドプロンプトでのコマンド実行の方が効率的であるため，MS-DOS や UNIX 系のコマンドの使い方も実習させている．

後半においては，テキストマイニングの理解を目標としている．形態素解析のツールとしては，MeCab を採用した．R 言語には，このツールを直接リンクして利用できる RMeCab というツールが用意されているので，R 言語での形態素解析はこれにより非常

---

* 2009 年 8 月 27 日 第 29 回高等専門学校情報処理教育研究発表会にて一部を発表

*1 一般科教授

原稿受付 2010 年 5 月 20 日

に容易となる.

最初に R 言語の基本的な使い方を履修させ,MeCab 自体をコマンドプロンプトから利用させる．次に，RMeCab による形態素解析の実習を行う．さらには，言語処理での有力なアルゴリズムである N-gram を理解させる.このための関数も RMeCab に用意されているので,学生は R 言語という一つのツールのみで，すべての言語処理が行えるという，操作上非常に有効な環境が得られることになる．最後には，R 言語によるテキストマイニングの実習を，前半に作成したコーパスを用いて行う.

表１　「情報処理応用 A」の授業内容

授業項目	時間	内容
1.情報収集の技術	2	インターネットから文字情報を収集する技術が理解できる.
2.情報検索の技術	2	文字情報を効率的に検索する技術が理解できる.
3.パーソナルデータベースの構築	2	インターネットから収集した文字情報をコーパス化できる.
4.情報発見の技術（正規表現）	2	正規表現が使え情報発見に利用できる.
5.テキストエディタの応用	2	エディタにおいて，正規表現が利用できる.
6.KWIC 検索	2	KWIC 検索が理解でき，情報発見に応用できる.
7.コーパス作成実習	4	大規模なテキストであるコーパスを作成することができる.
前期中間試験		
8.テキストマイニングとは	2	テキストマイニングの意味と意義が理解できる.
9.R 言語実習	2	R 言語の基本（特に文字列処理）が理解できる.
10.形態素解析と MeCab	2	形態素解析が理解でき，MeCab の利用ができる.
11.RMeCab によるテキスト解析	2	RMeCab が利用でき，形態素分析に応用できる.
12.N-gram について	2	N-gram の意味と意義が理解できる.
13.テキストマイニング実習	4	前半に作成したコーパスを用いて基本的なテキストマイニングができる.
前期期末試験		

３．R 言語について

R 言語は，オープンソースでフリーソフトウェアの統計解析向けプログラミング言語，及びその開発実行環境である [3].

統計解析向けといっても，一般的な工学計算やグラフ作成機能，強力な文字列処理機能なども併せ持っているため，本科目のような言語情報処理のためのプログラミング言語としてふさわしいと考えられる [2].

R 言語の実行例として,前述の RMeCab を利用した形態素解析例を以下に示す.

```
> res <- RMeCabC("R 言語を用いた言語情報処理教育")
> res
[[1]]
名詞
 "R"

[[2]]
名詞
言語

[[3]]
助詞
を

[[4]]
動詞
用い

[[5]]
助動詞
  "た"

[[6]]
名詞
言語

[[7]]
名詞
情報処理

[[8]]
名詞
教育
```

４．実習内容の詳細

ここでは，前後 2 回設けてある実習の内容詳細について述べる.

４－１　コーパス作成実習

巨大なテキストファイル群を意味するコーパスの作成は言語情報処理に欠かせないものであるが，インターネットは大規模なコーパスの宝庫と捉えることができる．何をインターネットからダウンロードするかは，言語処理の目的によって異なるが，前半最後に行うコーパス作成実習では，以下の３つを対象とした.

(1) 聖書　(2)日本文学作品　(3)百科事典

(1)の聖書は旧約・新約合わせて 66 巻からなり，2500 ページ以上におよぶ．このことから，単一の書籍としてはコーパスと呼ぶにふさわしいデータ量であろう.

日本語以外の聖書は多くの言語のものがインターネット上に公開されているが，日本国内の著名な聖書は著作権の関係で全文をダウンロードできるサイトは存在しない．その中で,「新改訳聖書」に関して

は，1 回 200 節の制限があるものの，繰り返してダウンロードすれば，聖書全文が入手できることがわかった[4].

しかし，全体が 3 万節以上に上る聖書では，ダウンロードを 150 回以上も手動で繰り返す必要がある．そこで，マウスやキーボードの操作内容を記憶しておいて，自動的に複数回繰り返すことができるユーティリティソフト（今回は「FreeMacro」[5]を用いた．）を使わせて，ダウンロードの効率化を図った．この結果，31215 節（行）からなるテキストが作成できた．

次にこのファイルを元にして，カタカナ用語の頻度や１-gram を，コマンドベースで作成させる実習を行った．図１には，結果例を示す．

次に(2)の日本文学作品については，著作権を失った作品を集めたサイトとして有名な「青空文庫」[6]を用いた．

ここでは，wget という，URL を指定するとそのサイトのファイル群が連続的にダウンロードできるコマンドを利用する．ただし，すべてのファイルを再帰的にダウンロードすることはネチケットからも許されないため，必要な URL のみをダウンロードできるようなバッチファイルを，青空文庫のインデックスのページの情報から自動作成できるようなプログラムを作らせた．

実際の実習ですべての作家のデータをダウンロードすることは現実的ではないので，サンプルとして夏目漱石の全作品を練習用にダウンロードさせ，次に学生の好む作家の全作品を対象とさせた．

これらのコーパス化されたテキストファイルを用いて，聖書と同様に N-gram を中心に実習を行い，聖書との結果比較をさせた．

最後に(3)の百科事典であるが，2008 年 11 月に無料公開された Yahoo 百科[7]を対象とした．これは，小学館の日本大百科全書全 26 巻を自由に検索できるサイトであるが，URL の解析を行わせて，約 10 万項目におよぶ全項目の自動ダウンロードのプログラムを作成させた．ただ，ダウンロード時間がかなり長時間となるので，15 名の学生にジャンル別に分割させることで効率化を図った．

このほか，青空文庫の外国版ともいうべき gutenberg や，インターネット版の百科事典として定評のある wikipedia なども同様の手法でコーパス化が可能であるが，これは学生の応用課題として自習させた．

以上の実習により，ある程度巨大なコーパスが，身近な PC を用いて作成できることを教育できた．

用語	度数	用語	度数	用語	度数
イスラエル	2630	バビロン	305	カナン	164
イエス	2054	ヨセフ	305	ヨハネ	161
ダビデ	1220	ペリシテ	300	エレミヤ	160
モーセ	895	キュビト	298	イサク	157
ユダ	893	パロ	298	マナセ	157
エルサレム	822	アブラハム	265	サムエル	156
エジプト	751	ヨシュア	260	ヨアブ	153
キリスト	690	パウロ	236	アモン	150
ヤコブ	462	モアブ	206	アッシリヤ	149
サウル	458	ヨルダン	198	ギルアデ	146
パン	420	ペテロ	192	シェケル	139
レビ	374	エフライム	187	エドム	137
アロン	361	ベニヤミン	185	サマリヤ	135
ユダヤ	351	アラム	173	バアル	134
ソロモン	316	シオン	172	ヒゼキヤ	133

図１　聖書コーパスの分析例（カタカナ用語 ベスト 45）

４－２　R 言語による N-gram 作成実習

N-gram の実習は，前半のコマンドベースのプログラミングでも一部実習させたが，R 言語を用いると極めて簡単に作業できるので，テキスト[8]の例題を中心に実習させた．

以下には，処理例として，本論文の 2-gram の一部を示す．下線部分がその命令である．

```
> res <- Ngram("ronbun.txt", type = 1, pos = "名詞")
file = ronbun.txt Ngram = 2
length = 1615
>
```

この結果を csv ファイルに収めたものを，Excel で整形したものを図２に示す．ここでは，形態素解析結果のうち，名詞のみを対象に 2-gram の度数の多い順に度数３以上のものを示した．

５．実習コンピュータ環境について

本科目の実習にあたっては，用いるソフトについて OS 以外はフリーソフトで行えるように配慮した．したがって，これまでに紹介した R 言語，MeCab，RMeCab などはすべてフリーソフトである．

このようなフリーソフトを，学校でも自分の PC でも，どこでも利用しやすいように，すべてのフリーソフトを USB フラッシュメモリに納めた．さらに，これらソフトを使う場合に，いちいち実行ファイルのアイコンを画面に表示させるのでは非常に煩

2-gram	度数	2-gram	度数	2-gram	度数
R-言語	23	フリー-ソフト	5	RMeCab-利用	3
--gram	12	作成-実習	5	jp-/	3
.-1	10	情報処理-応用	5	www-.	3
.-2	10	正規-表現	5	インターネット-上	3
3-.	10	.-6	4	カリキュラム-改訂	3
7-.	10	.-jp	4	クラ-スター	3
テキスト-マイニング	10	1-)	4	コーパス-化	3
形態素-解析	10	2-)	4	スター-分析	3
.-3	9	3-)	4	ドライブ-名	3
.-4	9	応用-A	4	解析-5	3
5-.	9	授業-内容	4	解析-結果	3
N--	8	青空-文庫	4	記述-文	3
9-.	7	文字-列	4	検索-技術	3
言語-情報処理	7	列-処理	4	言語-処理	3
情報処理-教育	7	.-7	3	効率-的	3
8-.	6	.-8	3	自由-記述	3
.-5	5	.-情報	3	章-テキスト	3
1-.	5	2--	3	情報-収集	3
10-.	5	2-.	3	情報-発見	3
4-.	5	8-%	3	的-ダウンロード	3
6-.	5	8-)	3	百科-事典	3
-	5	://-www	3	文字-情報	3
http-://	5	>-res	3	用語-度数	3
コーパス-作成	5	KWIC-検索	3		

図２　本論文の名詞に関する 2-gram（度数３以上のもの）

雑なため，USB メモリに対応したランチャー（今回は，Portable Start Menu）を用いた．

以上により，学生は場所を選ばずに本授業の実習ができる点は特筆できよう．

６．本年度におけるカリキュラム改訂

以上は，昨年度の授業内容に関するものであるが，R 言語を採用した 2 年目である本年度については，以下の点を念頭に，カリキュラムの改訂を行い，現在，これによる授業が進行中である．

- 情報収集・検索に関する技術は，学生にも浸透してきたので，これを削除する．
- 正規表現の説明は必要最小限とする．
- テキストエディタの実習は不要．
- テキストマイニングの実例説明と実習の時間を豊富に取る．

この方針により，参考文献 [8]のみをテキストとして選定し，ほぼこれに沿って説明・実習を行うカリキュラムに改訂を行った．本テキストの目次を以下に示す．

第 1 章 テキストマイニングとは何か
- 1.1 マイニングとは
- 1.2 応用事例
- 1.3 日本語処理
- 1.4 ツール

第 2 章 テキストマイニングの準備
- 2.1 R の導入
- 2.2 R の実行

第 3 章 R に慣れる
- 3.1 基本操作
- 3.2 ベクトル
- 3.3 行列
- 3.4 データフレーム
- 3.5 リスト
- 3.6 ファイルの入出力
- 3.7 グラフィックス作成
- 3.8 R でのプログラミング
- 3.9 R による文字列処理

第 4 章 MeCab と RMeCab
- 4.1 形態素解析とは
- 4.2 MeCab の導入
- 4.3 MeCab の実行
- 4.4 RMeCab の導入

第 5 章 RMeCab によるテキスト解析
- 5.1 RMeCab の利用
- 5.2 RMeCab による形態素解析
- 5.3 MeCab の辞書整備
- 5.4 データファイルの解析
- 5.5 ターム・文書行列
- 5.6 行列の重み付け
- 5.7 N–gram
- 5.8 語の共起関係

第 6 章 インターネット上のクチコミ情報の分析
- 6.1 インターネット上のクチコミ情報
- 6.2 携帯電話の評価
- 6.3 解析結果の整理
- 6.4 解析結果の評価

第 7 章 アンケートの自由記述文の分析
- 7.1 アンケートの特徴
- 7.2 アンケートの内容
- 7.3 分析の目的
- 7.4 気をつかう程度の差の検定
- 7.5 自由記述文の解析

7．今後の課題

上述のフリーソフトのうち，R 言語と RMeCab は USB フラッシュメモリのドライブ名が変わっても問題なく動作した（いわゆる，ポータブルなソフトウェア）が，MeCab については，インストール時にドライブ名が記憶されてしまうため，他の PC 環境において，当該ドライブ名が異なる場合にはエラーを生じてしまい，実行ができないことが判明した．

この対策としては，今のところは，新しい環境で MeCab を再インストールして，その環境設定ファイルを RMeCab に通知する設定を行うしか方法がないようである．今後は，この辺が柔軟になるよう，本ソフトの作者に要望を出していきたい．

8．あとがき

R 言語の試用は始めて日が浅いため，今後アンケート等により，学生の受け入れの度合いを調査し，よりよい情報処理教育につなげていきたい．

また高専においては，情報関連以外の科目でも R 言語はかなり教育に役立てられると思われるので，筆者が担当している応用数学などの科目でも，今後試用していく予定である．

参　考　文　献

1) 国立長野高専：「平成 21 年度シラバス」，国立長野高専（2009）
2) 高階知巳：「プログラミング R（基礎からグラフィックスまで)」，オーム社（2008）
3) Wikipedia「R 言語」の項，http://ja.wikipedia.org/wiki/R%E8%A8%80%E8%AA%9E
4) 「Bible Word Search　聖書用語検索」，http://www.tuins.ac.jp/~takao/biblesearch.html
5) 「FreeMacro」，http://www.vector.co.jp/soft/win95/util/se070172.html
6) 「青空文庫」，http://www.aozora.gr.jp/
7) 「Yahoo 百科」，http://100.yahoo.co.jp/
8) 石田基広：「テキストマイニング入門」，森北出版（2008）