テキストマイニングシステム Simpleminer の開発

村田 真樹 *　金丸 敏幸 *　一井 康二 **　白土 保 *　馬 青 ***,*　井佐原 均 *

* 独立行政法人 情報通信研究機構　({murata,kanamaru,shirado,qma,isahara} @nict.go.jp)

** 広島大学 大学院工学研究科　(ichiikoji @hiroshima-u.ac.jp)

*** 龍谷大学 理工学部　(qma @math.ryukoku.ac.jp)

1　はじめに

本稿では，われわれが開発した簡易テキストマイニングシステム Simpleminer について紹介する．このシステムは，Windows 上で簡便に動作する．自由記述のアンケートデータの分析や，論文書誌情報・論文タイトル情報からの動向分析に用いることができる．一般的なテキストマイニングシステム [1] が持つ，単語の頻度分析，クロス分析が可能である．そのうえ，情報抽出表とソートグラフと呼ぶ，他のシステムにない新規な技術を利用した分析も可能である．情報抽出表は，データを分かりやすい表の形で整理して表示する機能である．ソートグラフは，昔多かった傾向，最近多くなった傾向を簡便につかむことができる機能である．

図 1: インストール画面

2　インストール

Simpleminer は，Windows に簡便にインストールができる．インストーラーが用意されてあり，それをクリックすることで，図 1に示すようなインストール画面が立ち上がる．ここで，「次へ」をクリックしていくことで簡単にインストールできる．Simpleminer を起動すると，図 2に示すような画面が立ち上がる．

3　入力

入出力ファイルは csv 形式 (カンマ区切りのデータ形式) である．入力ファイルの例を図 3に示す．図のような入力ファイルを準備し，そのファイルを Simpleminer にセットして，図 2の各種ボタンを押すことで様々な処理ができるようになっている．図 2中の「表示」または「Excel」ボタンを押すと，それに該当するファイルを，Note Pad または，MS Word で開いて表示できる．

入力データとして，言語処理学会論文誌「自然言語

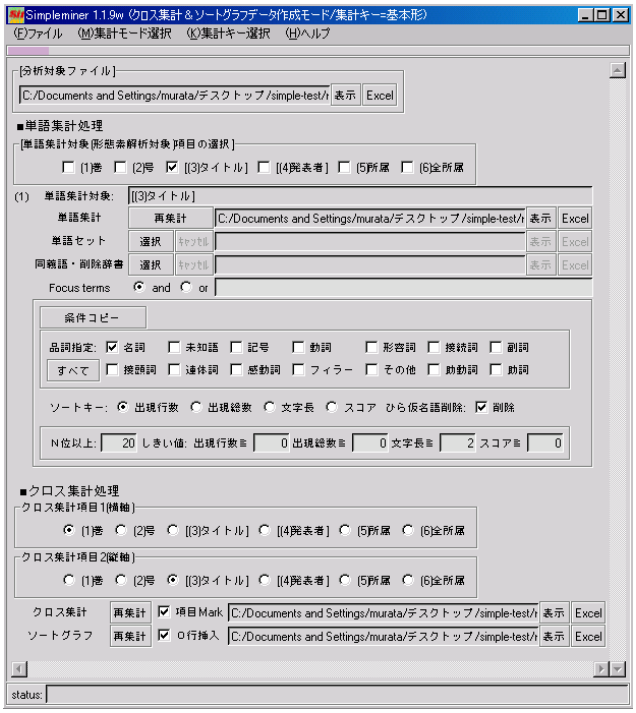

図 2: Simpleminer の画面

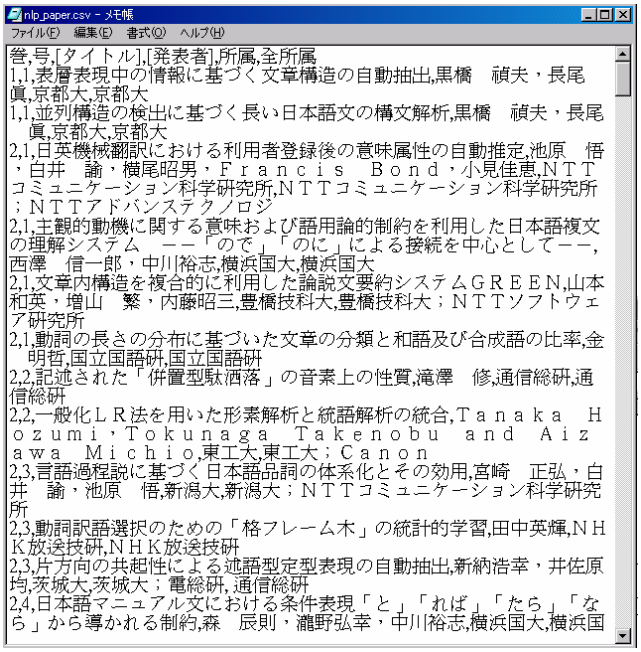

図 3: 入力の csv 形式のファイルの例

処理」の 1 巻から 10 巻までの書誌情報を与えた．毎年 1 巻ずつ出るので 10 年分のデータである [2]．タイトルの部分を対象としてテキストマイニング処理を行った．

4　単語集計

単語集計機能を使うと，図 4の結果を得る．どの単語が何個の論文のタイトルに出現したかを示している．図 4ではすべての品詞の単語を示した．これは，図 2で，すべての品詞をチェックし，文字長の設定値も 0 とし，ひら仮名語削除の機能のチェックを外して単語集計の「再集計」ボタンを押して実行したものである．図 4では分析に役立たない単語が多く出現している． Simpleminer では分析に用いる単語を簡便に指定することができる．ここでは，名詞のみを対象とし，文字長も 2 文字以上のもののみを対象とし，ひら仮名語も不要として対象外として単語集計をしてみる．これは，図 2のように設定させて，単語集計の「再集計」ボタンを押して単語集計を実行する．そうすると，図 5の結果を得る．図 5は不要な単語が削除され見やすくなる．単語集計機能により得られた図 5を見ることで，データの大雑把な傾向をつかめる．日本語を対象とし

	A	B	C	D
1	出現形	基本形	品詞	出現行数
2	の	の	助詞	155
3	を	を	助詞	63
4	た	た	助動詞	59
5	と	と	助詞	49
6	日本語	日本語	名詞	44
7	に	に	助詞	43
8	的	的	名詞	38
9	基づく/基づ	基づく	動詞	34
10	用い	用いる	動詞	34
11	による	による	助詞	33
12	し/さ/する	する	動詞	30
13	解析	解析	名詞	29
14	文	文	名詞	28
15	における	における	助詞	27
16	情報	情報	名詞	23
17	表現	表現	名詞	22
18			記号	22
19	翻訳	翻訳	名詞	21
20	自動	自動	名詞	21
21	ため	ため	名詞	21

図 4: 単語集計の例

表記	見出し	品詞	出現行数
日本語	日本語	名詞	44
解析	解析	名詞	29
情報	情報	名詞	23
表現	表現	名詞	22
翻訳	翻訳	名詞	21
自動	自動	名詞	21
抽出	抽出	名詞	19
システム	システム	名詞	18
モデル	モデル	名詞	17
機械	機械	名詞	17
手法	手法	名詞	17
検索	検索	名詞	14
コーパス	コーパス	名詞	14
要約	要約	名詞	14
意味	意味	名詞	14
名詞	名詞	名詞	13
言語	言語	名詞	13
学習	学習	名詞	12
構造	構造	名詞	12
単語	単語	名詞	11

図 5: 名詞のみによる単語集計の例

た研究が多いことがわかる．また，翻訳の研究も比較的多いことがわかる．ここでは名詞のみを取り出して分析したが Simpleminer では対象とする品詞を変更することもできる．また，同義語辞書により，異なる単語を同じ単語として扱ったり，削除辞書により集計対象から単語を強制的に排除することもできる．単語への分割と品詞の推定には ChaSen を利用している．

5　情報抽出表

次に情報抽出表の機能を示す (図 2の画面にはないが，集計モードを切り替えると情報抽出表の処理画面

[タイトル]	翻訳	システム	意味	日本語
スコア	42	20	8	12
出現行数	21	5	4	4
出現総数	26	18	16	44
開発者の視点からの機械翻訳システムの	翻訳	システム		
点字翻訳ボランティアのための対話型分	翻訳	システム		
日韓機械翻訳システムの現状分析及び	翻訳	システム		
頑健な英日機械翻訳システム実現のた	翻訳	システム		
日英機械翻訳システムTWINTRANの言	翻訳	システム		
日英機械翻訳のための日本語抽象名詞	翻訳		意味	日本語
日英機械翻訳における利用者登録後の	翻訳		意味	
意味的類似性を用いた音声認識正解部	翻訳		意味	
ターム間の意味的関連性に基づくターム	翻訳		意味	
派生文法に基づく日本語動詞句のウルク	翻訳			日本語
日本語－ウルグアイ語機械翻訳のための	翻訳			日本語
EMアルゴリズムを用いた教師なし学習の	翻訳			日本語

図 6: 情報抽出表の例

が表示される．)．ここでは，「翻訳」という単語を含む論文だけを対象に実行した．図 2の Focus terms の欄に「翻訳」という単語を入力すると「翻訳」という単語を含む論文だけを使った処理を実行できる．さらに，「機械」という単語をこの分析では不要であるのでそれは事前に分析から取り除く設定を行った．この設定は，単語集計の「単語セット」に分析に利用する単語だけをセットするか，「同義語・削除辞書」に「機械」を削除する単語として登録すると行える．このようにして，情報抽出表の処理を行った．その結果を図 6に示す．「翻訳」という単語を含む論文の中で出現が大きかった単語の順に左から右に表示している．また各論文タイトルの右側の欄にはその列の単語をタイトルに含んでいればその単語を表示している．論文タイトルもソートしており，なるべく左側の単語を含むタイトルの順に表示している．この表では各論文タイトルがどのような単語を含んでいるかを簡便に把握することができる．図 6から，翻訳の研究には，システムを対象とするもの，意味を特に扱っているもの，日本語を対象とするものの三種類の研究アプローチが主にあることがわかる．情報抽出表は，各データがどういった単語を含んでいるかを簡便な表の形で見るのに役立つ．

6 クロス分析

次にクロス分析を実行した．クロス分析では二つの事柄を指定して分析を行う．図 2で，「タイトル」と「所属」でクロス分析を行う．クロス集計項目 1(横軸)

	A	B	C	D	E	F	G	H	I
1		■日本語	■解析	■情報	■表現	■翻訳	■自動	■抽出	■システム
2	通信総研	4	1	3	4	1	2	3	2
3	京都大	6	3	2	4	0	2	1	1
4	SHARP	0	3	0	0	5	2	0	2
5	徳島大	0	0	3	0	0	2	1	1
6	東京工大	2	3	2	0	0	3	1	1
7	豊橋技科大	1	0	0	5	0	0	1	2
8	ATR	2	2	0	0	2	0	0	0
9	慶応大	0	1	3	1	0	1	4	0
10	横浜国大	4	1	1	1	0	0	0	1
11	NTTコミュ	1	1	0	0	1	1	1	0
12	鳥取大	2	0	0	0	1	0	0	0

図 7: クロス分析の例

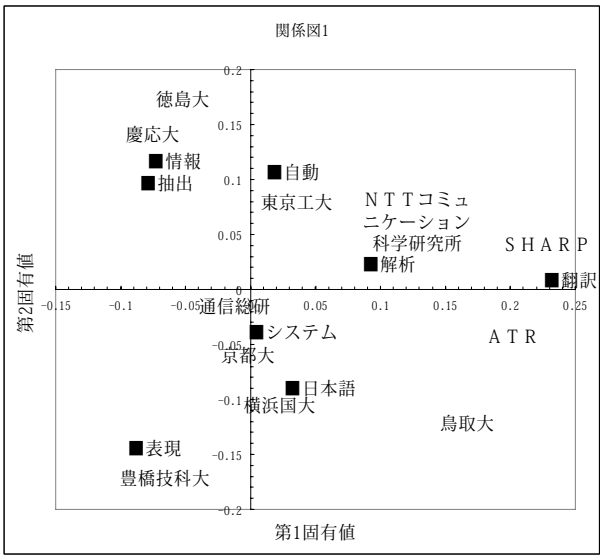

図 8: 双対尺度法の結果

に「タイトル」，クロス集計項目 2(縦軸) に「所属」を選択して，クロス集計の「再集計」ボタンを押す．そうすると，図 7に示す結果が得られる．この図は，どの組織がどういう単語を含む論文を何件発表したかを示す．図 7の結果に対して，双対尺度法 [3] を実行すると，図 8の結果が得られる． Simpleminer には，双対尺度法の機能は付いていない．双対尺度法の実行は，上田データマイニング塾 (http://www.datamining.jp/) から別途購入したツールを利用した．図 8から， ATR, SHARP が翻訳の研究が多く，徳島大，慶応大が情報抽出の研究が多いことがわかる．

テキストデータから数値データに落すことができれば，上述の双対尺度法など，種々の数値解析手法が利用できる． Simpleminer は，テキストデータから数値データに変換するところで役立っていることになる．

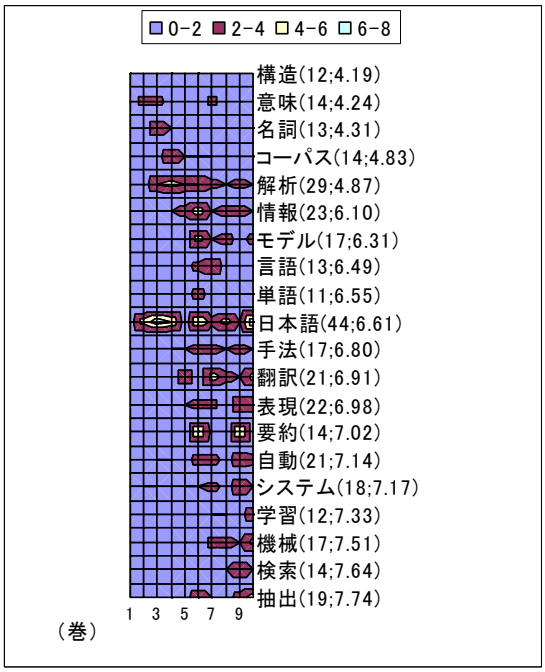

図 9: ソートグラフの例 (等高線の高さは件数を示す．図中の各単語に付与した二つの数字は左が合計件数を右が発表件数の巻数の平均を意味する (厳密な定義は本文を参照のこと))

7 ソートグラフ

最後にソートグラフの機能を示す．これは片方が数値である場合のクロス表のデータの分析に利用できる．図 2で，クロス集計項目 1(横軸) に「巻」，クロス集計項目 2(縦軸) に「タイトル」を選択して，ソートグラフの「再集計」ボタンを押して，ソートグラフを作成する．作成したソートグラフの例を図 9に示す．図で等高線の高さが論文の数を意味する．ソートグラフの横軸は巻数で，右側の単語は，タイトルに出現した単語である．この単語につけている一つ目の数字は，その単語を含む論文の合計で，二つ目の数字はその単語が多く出現している巻数の平均を示す．二つ目の数字は厳密には，発表される巻の平均値と最頻値と中央値の平均である．この値の小さいものから順に上から表示している．システムは等高線グラフを描きやすい csv 形式のファイルを出力する．ユーザはそのファイルから Excel を使って簡単に等高線グラフを描くことができる．等高線の高いところを見ることで，どの巻でどの単語を含む論文が多かったかがわかる．昔は「意味」「名詞」といった文法的な研究が多かったが，最近は「学習」「検索」「抽出」という処理的な研究が多いことがわかる．ソートグラフでは文献 [4] を参考にして等高線表示を利用している．

8 おわりに

本稿では，われわれが開発した簡易テキストマイニングシステム Simpleminer について紹介した．一般的なテキストマイニングシステム [1] が持つ，単語の頻度分析，クロス分析が可能である．そのうえ，情報抽出表とソートグラフと呼ぶ，他のシステムにない新規な技術を利用した分析も可能である．

具体例として自然言語処理学会の論文書誌情報を対象にテキストマイニングを行った結果を示したが，本システムを利用することでもっと多くの学会動向を簡便に調べることができる．また，本システムは，動向調査のみならず，自由記述のアンケートデータの分析にも利用できる．

参考文献

[1] 上田太一郎, 村田真樹, 小木しのぶ, 高山泰博, 末吉正成, 今村誠, 渕上美喜, 事例で学ぶテキストマイニング, (共立出版, 2008).

[2] 村田真樹, 一井康二, 馬青, 白土保, 金丸敏幸, 井佐原均, 過去 10 年間の言語処理学会論文誌・年次大会発表における研究動向調査, (2007), 言語処理学会ホームページ (http://www.nak.ics.keio.ac.jp/NLP/trend-survey.html).

[3] 上田太一郎, 刈田正雄, 本田和恵, 実践ワークショップ Excel 徹底活用多変量解析, (秀和システム, 2003).

[4] 谷口敏夫, 『人工知能と人間／長尾真』のテキスト可視化 —KT システムによるテキスト分析 —, (http://www.koka.ac.jp/taniguti96M/0/30/2000/Note2Nagao/Note20000409.htm, 2000).