簡易テキストマイニングシステム Simpleminer

A Simple Text Mining System: Simpleminer

村田 真樹　金丸 敏幸　一井 康二　馬 青　白土 保　井佐原 均*

Summary. We have developed a simple text mining system called “Simpleminer.” The system (cf. Fig. 1) works on Windows machines. It can be used, for example, to analyze questionnaires that include text and trends in journal titles. It can determine the usage frequencies of words (cf. Fig. 2) and make multiway tables, which are fundamental functions of text mining systems. In addition, Simpleminer has two unique functions: it can generate information extraction tables and sort graphs. An information extraction table (cf. Fig. 3) shows whether a text data includes words with high usage frequencies. A sort graph (cf. Fig. 4) shows the changes in the frequencies of word usage over a certain period of time, such as the 10 years shown in the figure. A useful feature of this graph is that words appearing more frequently in the past are displayed higher on the list. This enables users to recognize such words more effectively.

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．

2　Simpleminer

簡易テキストマイニングシステム Simpleminer の画面を図 1 に示す．入出力ファイルは csv 形式である．入力ファイルをセットして，図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．

入力データとして，言語処理学会論文誌「自然言語処理」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 [1]．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．単語集計機能を使うと，図 2 の結果を得る．何個の論文のタイトルに出現したかを示している．単語集計機能によりデータの大雑把な傾向をつかめる．日本語を対象とし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

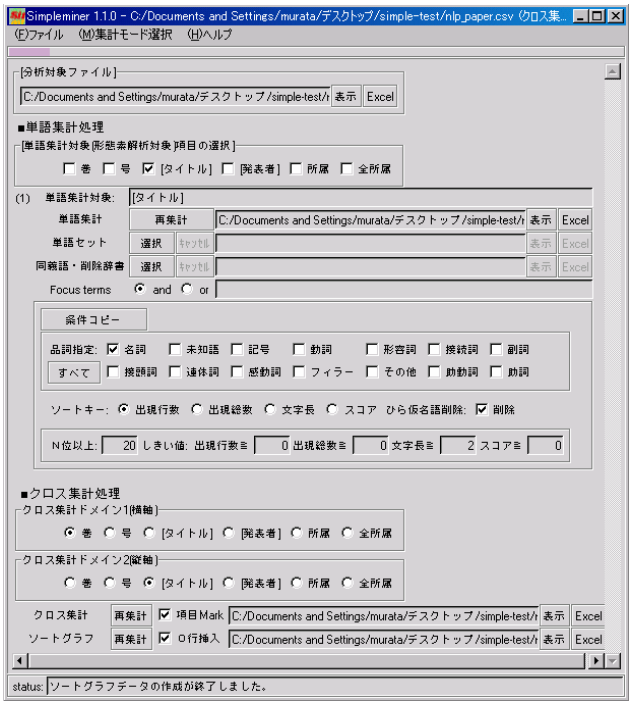

図 1. 画面の例

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．

次に情報抽出表の機能を示す (図 1 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．その結果を図 3 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

* Masaki Murata, Toshiyuki Kanamaru, Tamotsu Shirado, Qing Ma and Hitoshi Isahara, 独立行政法人情報通信研究機構, Toshiyuki Kanamaru, 京都大学, Koji Ichii, 広島大学, Qing Ma, 龍谷大学

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 2. 単語集計の例

[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め	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 3. 情報抽出表の例

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 3 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．

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．巻とタイトルに出現した単語のクロス表で，ソートグラフを作成した例を図 4 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

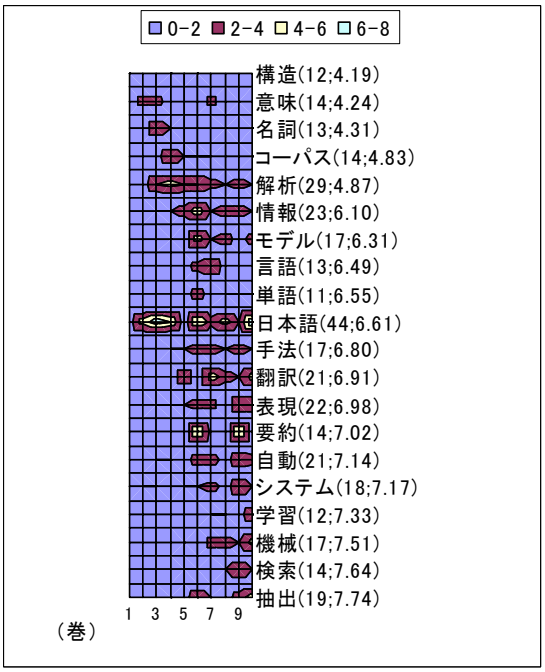

図 4. ソートグラフの例 (等高線の高さは件数を示す．図中の各単語に付与した二つの数字は左が合計件数を右が発表件数の巻数の平均を意味する (厳密な定義は本文を参照のこと))

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．

3 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．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] 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).