文章ジャンルに基づくテキストマイニング結果の比較考察

落合由治[1] 曾秋桂[1] 王嘉臨[1] 葉麥[1]
[1]台湾・淡江大学日本語文学科 {098194,ochiai,137176, 152790}@mail.tku.edu.tw

概要

2018年からAI情報処理の進歩に合せ,台湾での人文社会系日本語関係学科および日本語教育での,研究と教育内容への接続を目指す試みを種々,行っている.その中で,研究面ではテキストマイニング技術を人文社会系資料の読解また要点や論点抽出に応用する取り組みを各種のテキストを用いて試行してきた.その結果,テキストの文章ジャンルの相違によって抽出できる特徴に大きな違いがあることが分かってきた.また,その文章ジャンルによる特徴は言語の相違を超えて共通性があり,言語類型論などによる文レベルの言語の差異とは異なっていることが分かった.しかし,これらは対象とする言語資料に関する質的理解と相即的な結果で,対象資料の質的特徴が分からないままテキストマイニングの結果のみを見ても,有意義な特徴抽出は得られない.なぜ,こういう結果のかについて,新しい言語表現の特徴を捉える必要があると考えられる.

1 はじめに

テキストマイニングは,言語資料の特徴抽出手法のひとつで,計量言語学や言語をデータとして扱う各種の分野では,言語の特徴量を見出す手法として応用されてきた [1].一方,社会科学分野では,資料から有意義な表現特徴を数理的に取り出す手法として広く応用され,現在では,BI ツールとしてビジネスでも広く活用されるようになっている [2].しかし,質的方法を中心にしている人文社会系での応用は限定的であり,人文社会系の質的言語資料読解との接続は難しかった.論者たちは,2018 年から AI 情報処理の進歩に合せ,台湾での人文社会系日本語関係学科および日本語教育での研究と教育内容への接続を目指す試みを実施する中で,研究面でテキストマイニング技術を人文社会系資料の読解また要点や論点抽出に応用する取り組みを各種のテキストを用いて試行してきた.本発表では,その結果明らかになったテキストの文章ジャンルの相違によって抽出できる特徴に大きな違いがあること,その文章ジャンルによる特徴は言語の相違を超えて共通性があること,対象とする言語資料に関する質的理解と相即的な結果であることの三点について述べ,テキストマイニングによる言語資料の特徴抽出の課題について考察していきたい.

2 日本語の文章ジャンルによる結果

まず,テキストマイニングはテキストの文章ジャンルの相違によって抽出できる特徴に大きな違いがある点について述べる. 現代社会で一般に広く読まれている日本語の文章ジャンル例として,小説,論説,韻文を選び,現代小説は村上春樹「ドライブマイカー」,論説は読売新聞社説「データ流通　世界の成長加速するルールに 2021/12/26」,詩にはアニメソング「planetarian―ちいさなほしのゆめ」を取り上げた.テキストマイングには樋口耕一(2020)の KH Coder を使用し,分析方法は共起ネットワークを用いた [3].以下,結果の要点を紹介する.

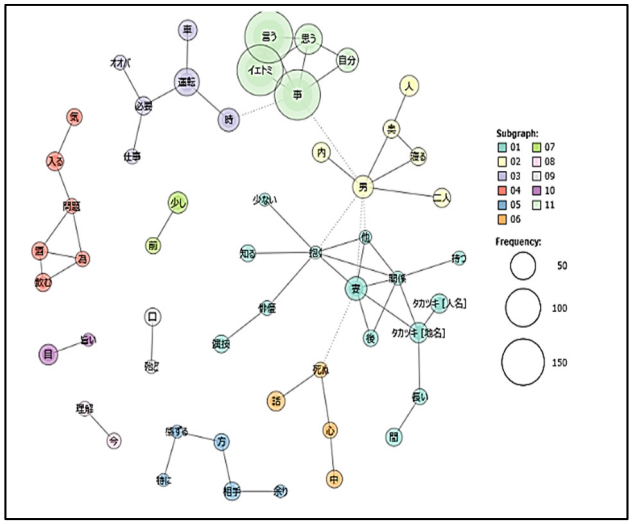

図1　日本語現代小説の結果

まず,現代小説では語の出現頻度に左右されて人物名とその物語中での動きに関する語が作品の部分ごとに多数出て,読解で重要な出現頻度の低い語彙は質的読解を事前にしない限り十分取り出せない.

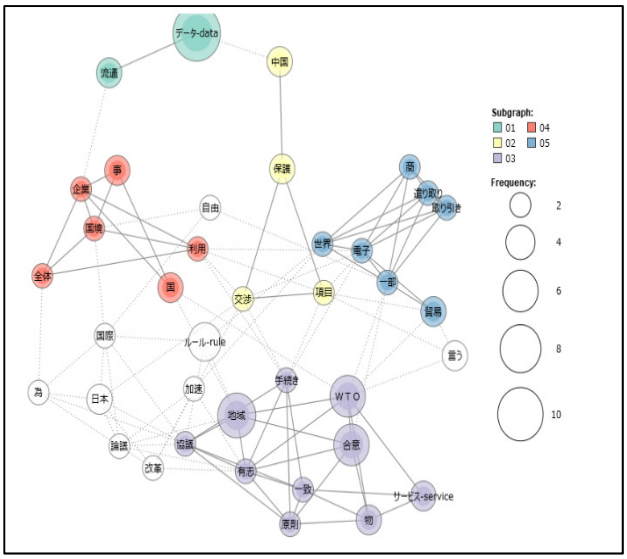

図 2　日本語論説の結果

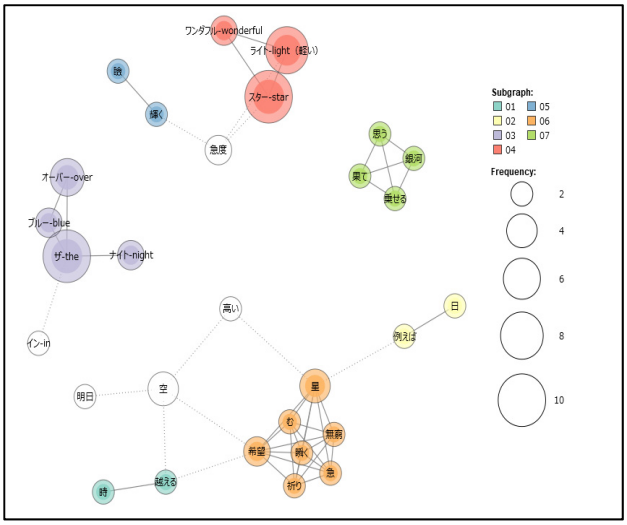

図 3　日本語歌詞の結果

一方，論説である社説の場合は，図 2 のように内容の要点を示す語彙群を関連性のあるクラスターとして容易に抽出でき，各クラスターを見るだけで，内容をほぼ推測できる．最後に，歌詞の場合では，韻文中で反復される，いわゆるリフレーン部分にあたるような語彙群が主に抽出され，歌詞の重要なモチーフを理解する手掛かりになるが，それ以外の部分は読解しない限り抽出しにくい．分析結果の理解容易性で見ると（易）論説＞韻文＞小説（難）の順になり，言語表現のジャンルという文章構成的差異に応じて，抽出結果に質的差異が存在すると考えられる．

3 外国語文章との分析結果比較

次に，日本語文章の分析結果と，外国語文章の分析結果を比較して考察する．外国語文章の例は，華語の例に，現代小説は金庸「書剣恩仇録」，論説は自由時報社説「不要見不得台灣好 2021/12/27」，詩には五月天「温柔」，英語の例に，現代小説は Philip K. Dick「Second Variety」，論説は朝日新聞「Draft budget shows lack of focus, no fiscal discipline 2021/12／25」，詩には『アナと雪の女王』「Let it go」を取り上げた．以下，順に結果を見ていく．

3.1 華語と英語の小説の場合

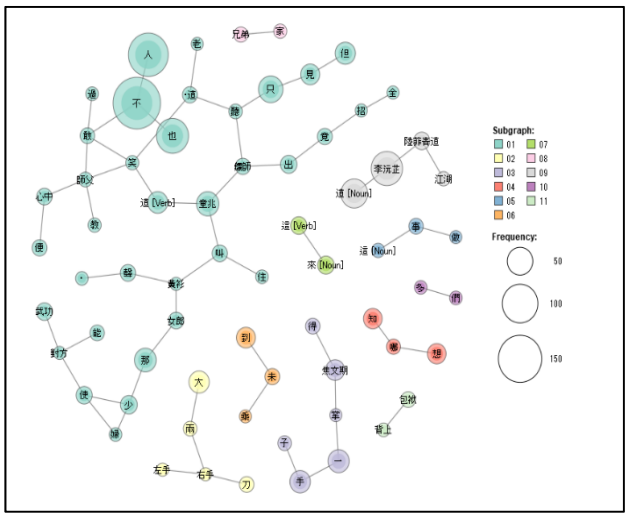

図 4　華語現代小説の結果

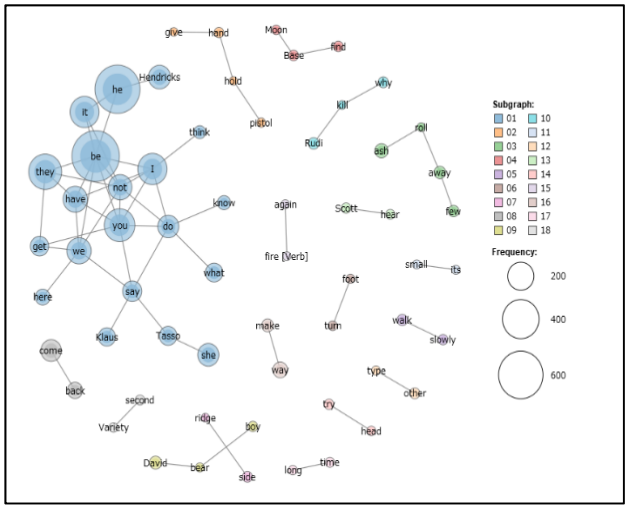

図 5　英語現代小説の結果

華語の結果でも日本語の場合と同じく，現代小説では語の出現頻度により人物名とその物語中での動きに関する語が作品の部分ごとに多数出て，読解で重要な出現頻度の低い語彙は質的読解を事前にしない限り十分取り出せない．テキストマイニングだけでは，ストーリーや背景を理解する手掛かりは得られない．英語の場合も同じで，主な登場者とその動きを描く語彙が中心になり，部分で出ている小さなク

ラスターのとの関係は明確にはならず, ストーリーを把握し, 作品読解の重要なキーワードを決めるには, 事前の十分な読解が必要である. 小説のような文章構成は, 特定の時の持続の中で登場者の動きを描写する部分が中心になり, その前後や内部に, さまざまな説明を行う要素が入った枠構造をなしている場合が多く [4], 質的に異なる表現で構成された文章はテキストマイニングの方法では十分に要点を抽出できないと考えられる.

3.2 華語と英語の論説の場合

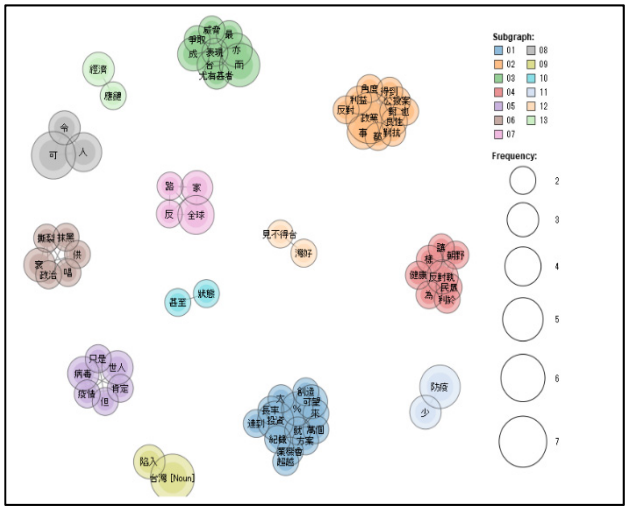

図 6　華語論説の場合

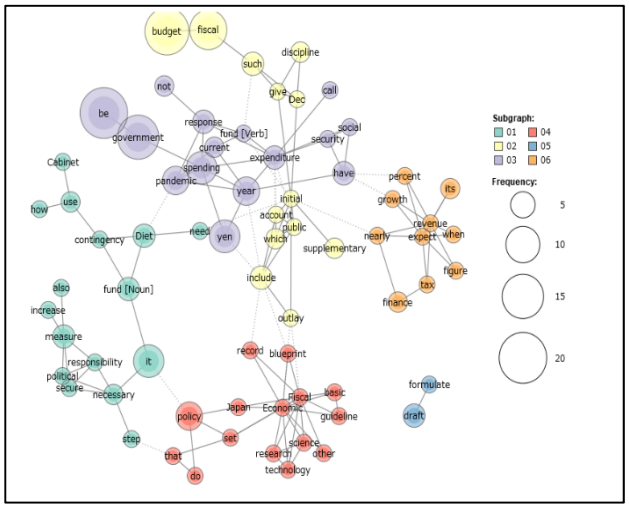

図 7　英語論説の結果

次に, 華語と英語の論説の結果を比較してみると, 日本語の論説の場合と同じく, 論説中の要点を示す語群がそれぞれクラスターとして大きなまとまりで抽出されて, 各部分の内容を容易に理解できる. 論説は, 語の階層化構造による各概念のネットワークがあり, こうした構造を用いて各部分で複数の話題の焦点を作ることにより論点が生まれる文章構成と考えられるので, [5]テキストマイニングによって, 各部分の要点を容易に取り出すことができると言える. 論説ジャンルの文章は, 最もテキストマイニングに向いている資料となる. 論説ジャンルは、作者の意見を書いたもので、インターネットの消費者の意見、各種のエッセイ、論文、解説、説明など、作者の意見を述べた類の、このジャンルに含まれる文章はテキストマイニングなど数理的手法で容易に特徴を抽出できると言える。

3.3 華語と英語の詩の場合

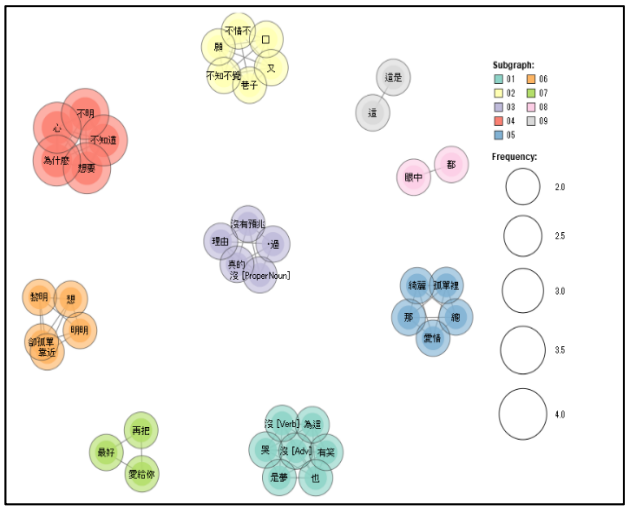

図 8　華語歌詞の結果

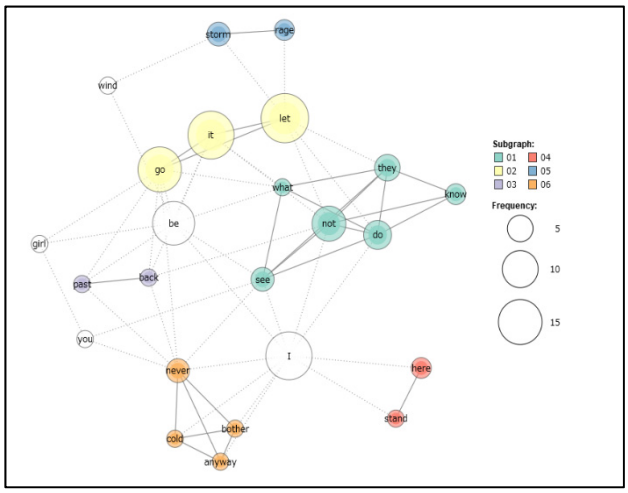

図 9　英語歌詞の結果

最後に, 華語と英語の歌詞の結果を比較してみると, これも日本語の場合と同じく, 詩の中で反復して用いられている語彙がクラスターのまとまりになって抽出された. いわゆるリフレーンに当たる部分に出てくる語彙が多く, その部分を詩の中心モチーフ

と理解する手掛かりになる.

しかし,それ以外の部分の語彙は断片的で,そのままでは内容理解の手掛かりにはならない.韻文は,現代詩の場合,各部分の各連がリフレーンに当たる反復される部分に収斂し,どの部分も基本的に同じ収斂部分を持つことで構成されている文章構成と考えられるので [5],反復される収斂部分はテキストマイニングの手法には適している.しかし,その収斂部分に到る各部分の内容は,抽出しがたい.

3.4 各言語を超えた文章ジャンルの共通性

以上の結果を分析結果の理解容易性で見ると,日本語と同じく華語,英語でも(易)論説>韻文>小説(難)の順になる.このことから,文章表現のジャンルという文章構成の差異に応じて,抽出結果に質的差異が各言語の形式的文法的差異を超えて,存在すると考えられる.つまり,各言語の形式的文法的差異を超えて,文章構成の質的差異による文章ジャンルがそれぞれあり,それは今回取り上げた三言語では共通した構成であると推測できる.テキストマイニングの有効性の相違は,各言語の差異には拠らず,文章構成の差異によって決まっている可能性があると考えられる.

今回は,特に人文社会系でテキストマイニングを応用して質的量的研究を行うことを念頭に [6],文章ジャンルごとで結果の比較をおこなった.人文社会系でテキストマイニングを応用する場合は,元々の基本的研究方法である質的研究を前提に応用をおこなえば,質的量的研究に寄与できる可能性が十分にあることが分かった.

一方,広汎な言語資料を対象にした自然言語処理においては,言語一般という前提が成り立つのは,文章ジャンルが関係しないレベルでの言語の規則性についてではないかという問題提起ができるかもしれない.そして,意味的処理やさらに高度の多義的処理を扱う場合には,語レベルでの問題を超えた文章レベルの文章ジャンルごとの問題が存在していることを視野に入れて,処理をデザインしていくことで,今まで解決できなかった問題を乗り越えられる可能性があると考えられる. [7]

また,言語の問題として,各言語の形式的文法的差異を超えて,文章ジャンルが具体的な言語運用を規定している可能性があることを考えられる.はたして文章レベルの言語規則とは何であるのか,今まで意識されてこなかった言語表現の問題が存在しているのではないか,こうした課題も新たに浮上してくると思われる.

どのように新しい問題の圏域を扱っていくか,今後の探究をおこなっていきたい.

4 おわりに

人文社会系分野で AI 技術に接続できる部分を探りながら,手掛かりとしてテキストマイニングの有効性と限界という視点を得たことで,文章ジャンルによる言語表現の規定という,今までの研究では,主題化しにくかったテーマが見えてきた.文章ジャンルの問題は、1950－70 年代までは関心を集めた分野であったが、現在は自明視されて扱われることはほとんどない。こうした問題をどう考えるか、言語研究などの分野で新しい視点を提出していかなくてはならない。同時に、こうした言語の特徴を自然言語処理の特徴抽出の問題として考えるなど,自然言語処理の面からもアプローチできるか探求しながら,新しい問題の探索を続けて行きたい.

謝辞

本研究は台湾科技部研究案 109-2410-H-032 -061 -MY3 の助成を受けたものです.

参考文献

1. **計量言語学会篇.** データで学ぶ日本語学入門. 東京 : 朝倉書店, 2017.
2. **IT トレンド.** 【図解】BI ツールとは？機能や目的、種類などわかりやすく徹底解説！ IT トレンド. (オンライン) 2021 年 9 月 22 日. (引用日: 2021 年 12 月 27 日.) https://it-trend.jp/bi/article/explain.
3. **樋口耕一.** 社会調査のための計量テキスト分析 第 2 版. 京都 : ナカニシヤ出版, 2020.
4. **ジェラルド・プリンス.** 物語論辞典. 東京 : 松柏社, 1997.
5. **永尾章曹.** 国語表現法研究. 東京 : 三弥井書店, 1975.
6. **ウド・クカーツ.** 質的テキスト分析法. 東京 : 新曜社, 2018.
7. **北原保雄（監修）佐久間まゆみ（編）.** 朝倉日本語講座　新装版 7　文章・談話. 東京 : 朝倉書店, 2018.