P14-5　日本語書籍タイトルの形式的構造の分析

UTLiS 東京大学図書館情報学研究室

矢田 竣太郎　東京大学大学院 教育学研究科
影浦 峡
岩井 美樹　東京大学大学院 学際情報学府

背景・動機

書籍のタイトル (書名) は固有表現の一種

書名抽出の困難

- 名詞句とは限らず、文ですらありうる
- それ自身として表記上の規則性に乏しい
- 他の固有表現を内包することがある

書名抽出の手がかり

- 書名を表すための社会的なルールがある
- 包括的な書名のリストが組織的に作成されている

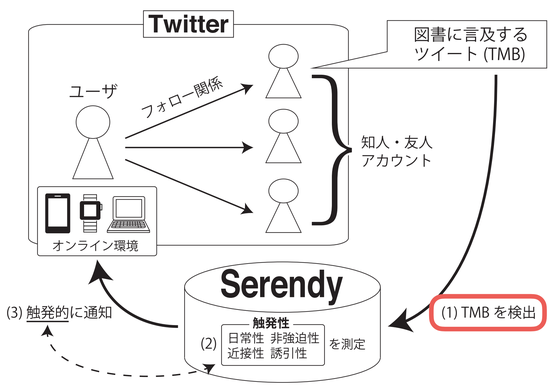
とはいえ、書名抽出に役立つような、書名それ自体の規則性や特徴はあるのではないか？

応用

図書推薦システム開発のサブタスク

書名と合致する文字列を含むツイートから、実際に図書に言及しているツイートを識別する

(Yada & Kageura, 2015-2016)



ツイート分類器に用いる素性として「検出した書名文字列の**書名らしさ**」の有効性が示唆されたが、表現方法には改善の余地がある　(Yada & Kageura, 2016)

目的

書名の形式的特徴を分析するための予備的調査として、書名の形式的構造の記述を試みる

データ

日本語書籍のタイトル **1,477,278** 件

- 国立情報学研究所運営の Webcat Plus に 2016 年 6 月までに採録された日本語書籍タイトルのうち、ISBN を付与されたもの
- うまく正規化されていない書名は除外した (2000件程度)
- サブタイトルは取り除いた

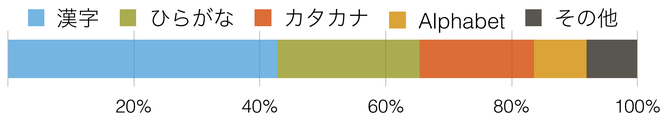
分析観点

言語表現の文法的単位を基準とし、大きく次の3つの観点から書名の形式的特徴を計測する

1. **文字レベル**
   - (a) 文字種の割合
   - (b) 書名の文字数
2. **単語レベル**
   - (a) 1単語ごとの頻度
   - (b) 書名の単語数
   - (c) 品詞の割合
   - (d) 1書名あたりの品詞の平均頻度
3. **句・節・文レベル**
   - (a) 書名の文節数
   - (b) 句・節・文の割合
   - (c) 頻出構文の抽出

結果

1(a) 文字種の割合



1(b), 2(b), 3(a) 書名の長さ

要約統計量	文字数	単語数	文節数
平均	11.7	5.7	2.2
標準偏差	7.3	3.1	1.2
最小値	1	1	1
1/4分位点	7	4	1
中央値	10	5	2
3/4分位点	14	7	3
最大値	218	78	29

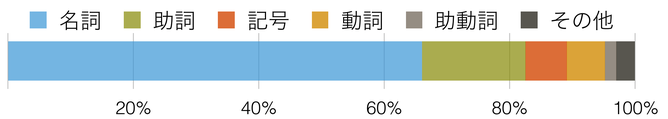
12文字の書名の例

『有機農業と消費者のくらし』
『知ってるつもりの親知らず』
『5分でできる早ワザおかず』
『日本のオートバイの歴史。』
『フリードリヒ・エンゲルス』

29文節からなる書名

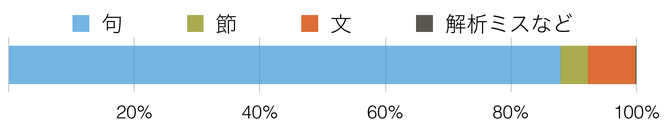
『悩みも迷いも若者の特技だと思えば気にすることないですよ。皆そうして大人になっていくわけだから。ぼくなんかも悩みと迷いの天才だったですよ。悩みも迷いもないところには進歩もないと思って好きな仕事なら何でもいい。見つけてやって下さい。』

2(d) 品詞の割合



2(d) 1書名あたりの品詞の平均頻度

3(b) 句・節・文の割合



展望

書名抽出への貢献に向け、書名に性質の似た他の固有表現（映画・音楽のタイトルなど）と同様の比較をすることで、書名固有の特徴を見出したい

本研究では全ての書名を等価の固有表現として分析したが、ジャンルや出版形態、価格や年代といった他の書誌的事項に照らすことで、書名はいかなるかたちをしているか/してきたかという図書館情報学的な基礎研究にも発展させられる

2(a) 1単語ごとの頻度

全体の上位10件

単語 [品詞]	割合
の [助詞-連体化]	7.88%
・ [記号-一般]	1.79%
と [助詞-並立助詞]	1.21%
を [助詞-格助詞]	1.20%
！ [記号-一般]	0.97%
に [助詞-格助詞]	0.95%
で [助詞-格助詞]	0.77%
が [助詞-格助詞]	0.73%
「 [記号-括弧開]	0.67%
」 [記号-括弧閉]	0.67%

内容語のみの上位10件

単語 [品詞]	割合
する [動詞-自立]	0.57%
日本 [名詞-固有名詞]	0.53%
集 [名詞-接尾]	0.48%
学 [名詞-接尾]	0.47%
ため [名詞-非自立]	0.43%
法 [名詞-接尾]	0.40%
わかる [動詞-自立]	0.37%
入門 [名詞-サ変接続]	0.37%
問題 [名詞-ナイ形容詞語幹]	0.33%
本 [名詞-一般]	0.31%

3(b) 頻出構文

構文パターン	頻度	例
[複合名詞]	**353,166**	羅生門
[複合名詞]集	8,733	万葉集
[複合名詞]入門	6,062	言語学入門
[複合名詞]論	4,963	自由論
[複合名詞]ガイド	4,295	国内学会誌ガイド
[複合名詞]の[複合名詞]	**228,736**	種の起源
[複合名詞]の[複合名詞]学	3,021	言語の脳科学
[複合名詞]の研究	2,606	善の研究
日本の[複合名詞]	2,257	日本の図書館
[複合名詞]の[複合名詞]たち	1,979	海の生き物たち
[複合名詞]と[複合名詞]	**34,026**	罪と罰
[複合名詞]社会と[複合名詞]	377	現代社会と著作権
[複合名詞]と[複合名詞]社会	241	暗号と情報社会
[複合名詞]と[複合名詞]たち	236	現代物理とわたしたち
[複合名詞]・[複合名詞]	**29,431**	ネルソン・マンデラ
[複合名詞]・[複合名詞]集	725	ピアノ弾き語り・バラード集
[複合名詞]・[複合名詞]ガイド	427	名水・わき水ガイド
[複合名詞]・[複合名詞]入門	365	集合・位相入門
[複合名詞]の[複合名詞]の[複合名詞]	**21,326**	人生の意味の心理学
[複合名詞]のための[複合名詞]	4,734	エンジニアのための物理化学
[複合名詞]のための[複合名詞]入門	716	臆病者のための億万長者入門
[複合名詞]の[複合名詞]と[複合名詞]	**20,235**	中世の内乱と社会
日本の[複合名詞]と[複合名詞]	307	日本の民謡と舞踊
日本[複合名詞]の[複合名詞]と[複合	299	日本各地の自然とくらし
[複合名詞][接頭詞][複合名詞]	**13,483**	園芸植物大事典
[複合名詞]大[複合名詞]	2,904	ニンテンドウ64大百科
[複合名詞]新[複合名詞]	862	路面電車新時代
[接頭詞][複合名詞]	**13,011**	名探偵コナン
新[複合名詞]	2,182	新明解国語辞典
大[複合名詞]	894	大脱出
[複合名詞]と[複合名詞]の[複合名詞]	**12,862**	ジャックと豆の木
[複合名詞]と[複合名詞]の[複合名詞]	427	アロマと月の占星学
[複合名詞]・[複合名詞]の[複合名詞]	**9,795**	保育・幼児教育の原理
[複合名詞]・[複合名詞]市の[複合名	399	愛知県・名古屋市の理科
[複合名詞]で[動詞][複合名詞]	**8,468**	1日で読める徒然草
[複合名詞]でわかる[複合名詞]	1,041	マンガでわかる量子力学

言語処理学会 第23回年次大会 @筑波, 茨城県　P14: ポスター (3) 3月15日(水) 13:50-15:20 7A206

本研究は科研費『オンラインを介して「前読書家」の読書を触発する方式・環境の開発』(16K12542) の支援を受けたものである