社会学へのNDL全文テキストデータの活用

第7回東京大学学術資産アーカイブ化推進室主催セミナー
2023.12.13

東京大学大学院人文社会系研究科社会学研究室
瀧川裕貴

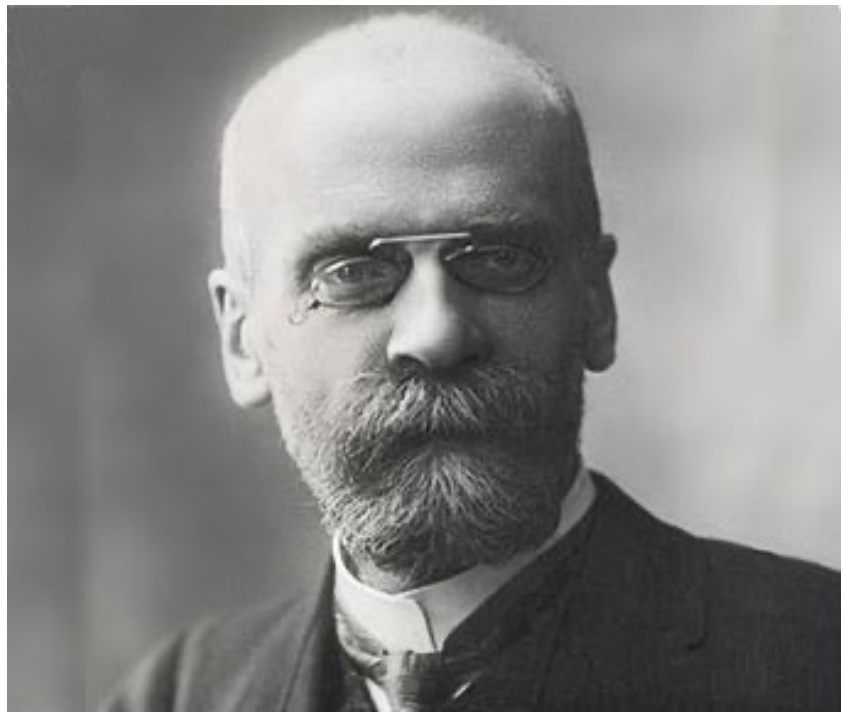
自己紹介

- 専門は数理社会学、計算社会科学
- ビッグデータや大規模テクストデータを用いた文化社会学、社会意識の研究

E.Durkheimの社会学の方法規準

- 社会を「モノ」のように
- 社会が現れている一つの有力な対象は全文テキスト

→集合意識・集合表象



E.Durkheim

社会学にとってのテキスト

- 書かれたテキストの読解はあらゆる人文学、社会科学の基礎であり、社会学も同様
- 人間の社会的行為の多くの部分は言語を媒介
  - 書籍、論文、新聞、雑誌、手紙、パンフレット、広告文、ウェブ上の投稿…
- 従来は質的アプローチが主。
- しかし、近年、テキストのデジタル化を背景に、大規模テキストに対する計算的テキスト分析、Text as Dataアプローチが勃興

質的アプローチの限界

前提：専門家による質的読解は（現在のところ）機械より優れている。ゴールドスタンダード。その上で…

- 大規模なテキストに対する質的アプローチの限界
  ①読みの選択性
  ②選択基準の恣意性
  ③再現不可能性
- とくに、長期にわたる、広範な範囲の、大量のテキストの読解には不向き

➢ ○○の考古学、○○の比較etc…

計算テクスト分析の長所

1. 網羅的「読解」による選択バイアスの回避
   - 「資料を選択的に読んでいるだけでは」疑惑の回避

2. 統計モデルによる再現可能性の確保
   - 職人芸は真似できないので。。。

3. アブダクティブな発見を可能にする柔軟性
   - 人間の読みは意外と思い込みが激しい

注意：従来型に取って代わるわけではない！
- 結果の解釈には質的読解との組み合わせは不可欠
- Validate, validate, validate! (Grimmer)

NDL全文データ

https://dl.ndl.go.jp/ja/

NDL全文データ

資料群	年代等
図書	明治期以降、1968年までに受け入れた図書
	震災・災害関係資料の一部（1968年以降に受け入れたものを含む。）
雑誌	明治期以降に刊行された雑誌（刊行後5年以上経過したもの）
古典籍	貴重書・準貴重書、江戸期以前の和漢書等
博士論文	1990～2000年度に送付を受けた論文
官報	1883（明治16）年7月2日（創刊）～1952（昭和27）年4月30日に発行された官報
憲政資料	幕末から昭和までの日本の政治家・官僚・軍人などが所蔵していた書簡・書類・日記等
録音・映像関係資料	カセットテープ、ソノシートなどの録音資料（付属する冊子等を含む）、レーザーディスクなどの映像資料（付属する冊子等を含む）、日本脚本アーカイブズ推進コンソーシアムから寄贈された1980年以前の放送脚本（テレビ・ラジオ番組の脚本・台本）の一部、明治期以降の日本人作曲家の手稿譜及びその関連資料の一部
地図	大正後期から昭和前期までに国内で刊行された地図資料
その他	他機関が所蔵するアナログ資料をデジタル化したもの。 • 日本占領関係資料：米国の国立公文書館が所蔵する戦後の日本占領に関する公文書のうち、米国戦略爆撃調査団文書、極東軍文書等の一部 • プランゲ文庫：プランゲ文庫（戦後GHQが検閲のために集めた日本国内出版物）のうち図書等の一部 • 歴史的音源：1900年初めから1950年頃までに国内で製造されたSP盤等に収録された音楽・演説等 • 他機関デジタル化資料：科学映像、東京大学附属図書館デジタル化資料、愛・地球博、内務省検閲発禁図書など

資料種別	デジタル化資料提供数（概数）			
	インターネット公開資料	図書館送信対象資料[1]	国立国会図書館館内提供資料	合計
図書	35万点	55万点	7万点	97万点
雑誌	1万点	80万点	53万点	134万点
古典籍	7万点	2万点	-	9万点
博士論文	1万点	12万点	2万点	15万点
官報	2万点	-	-	2万点
憲政資料	0.5万点	-	-	0.5万点
録音・映像関係資料	-	0.3万点	0.6万点	0.9万点
地図	-	-	0.1万点	0.1万点
その他	6万点	1万点	9万点	16万点
合計	**55万点**	**150万点**	**71万点**	**275万点**

NDL全文データの社会学的活用

- 巨大なテキストデータから、社会的に共有された意識や態度（集合表象）を取り出す。
- 幸福（ウェルビーング）をめぐる日本社会の人々の集合表象

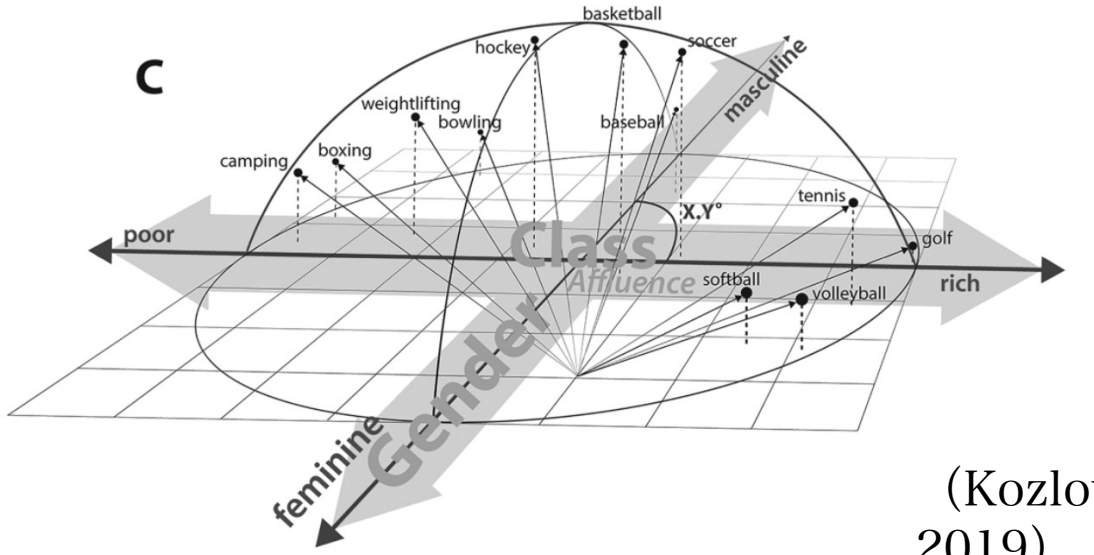
NDL全文データに関する注意

- 生成AIの開発は行っていない。
- 社会学の学術研究のためのモデル構築であり、その他の目的への転用はしない。

全文データからどのように集合表象を取り出すか？

- 文化の幾何学アプローチ（Kozlowski et al. 2019）
  - 「単語埋め込みモデルを文化の社会学的分析に応用する方法」(Kozlowski 2019).
  - 意味空間上に文化次元を設定し、文化次元間の関係や構造を検討

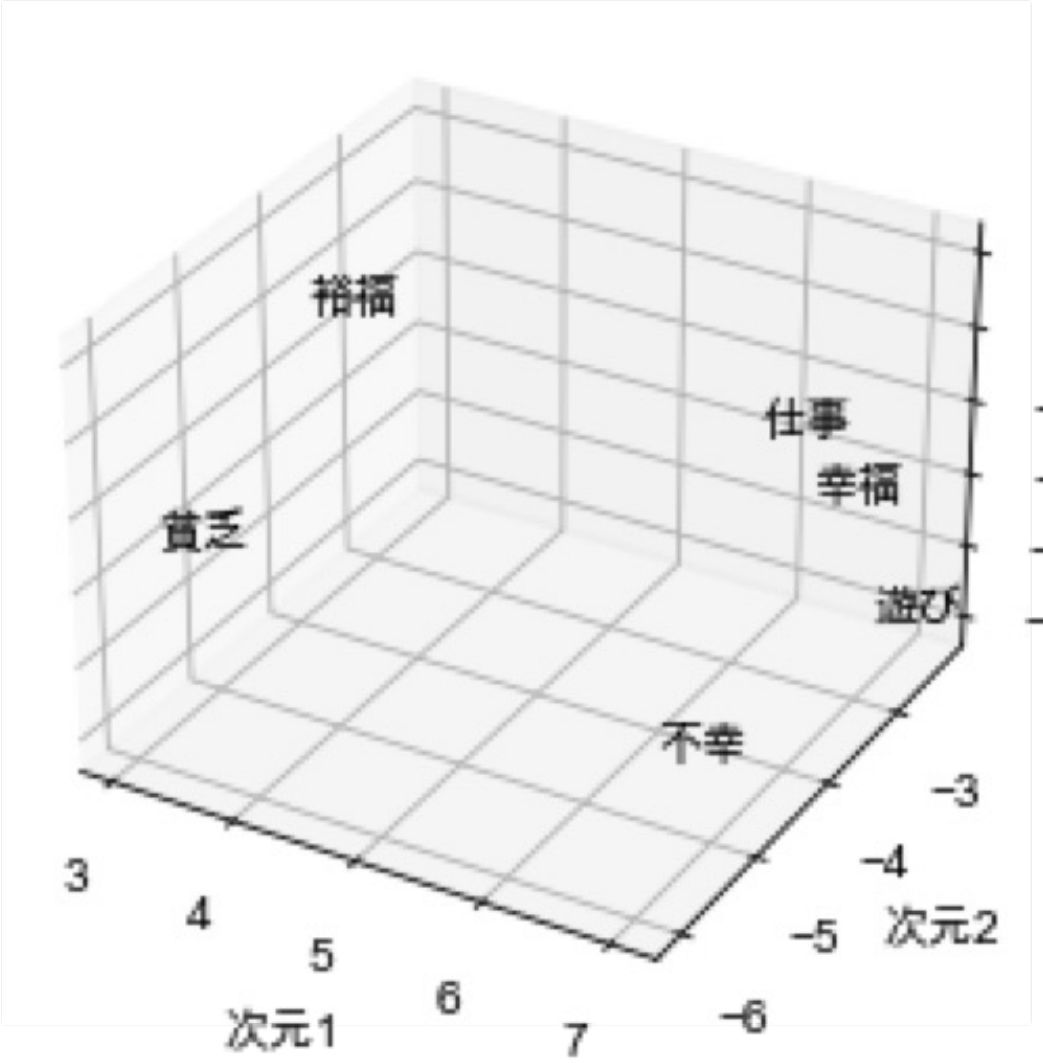
  →文化次元が集合的表象に相当



(Kozlowski et al. 2019)

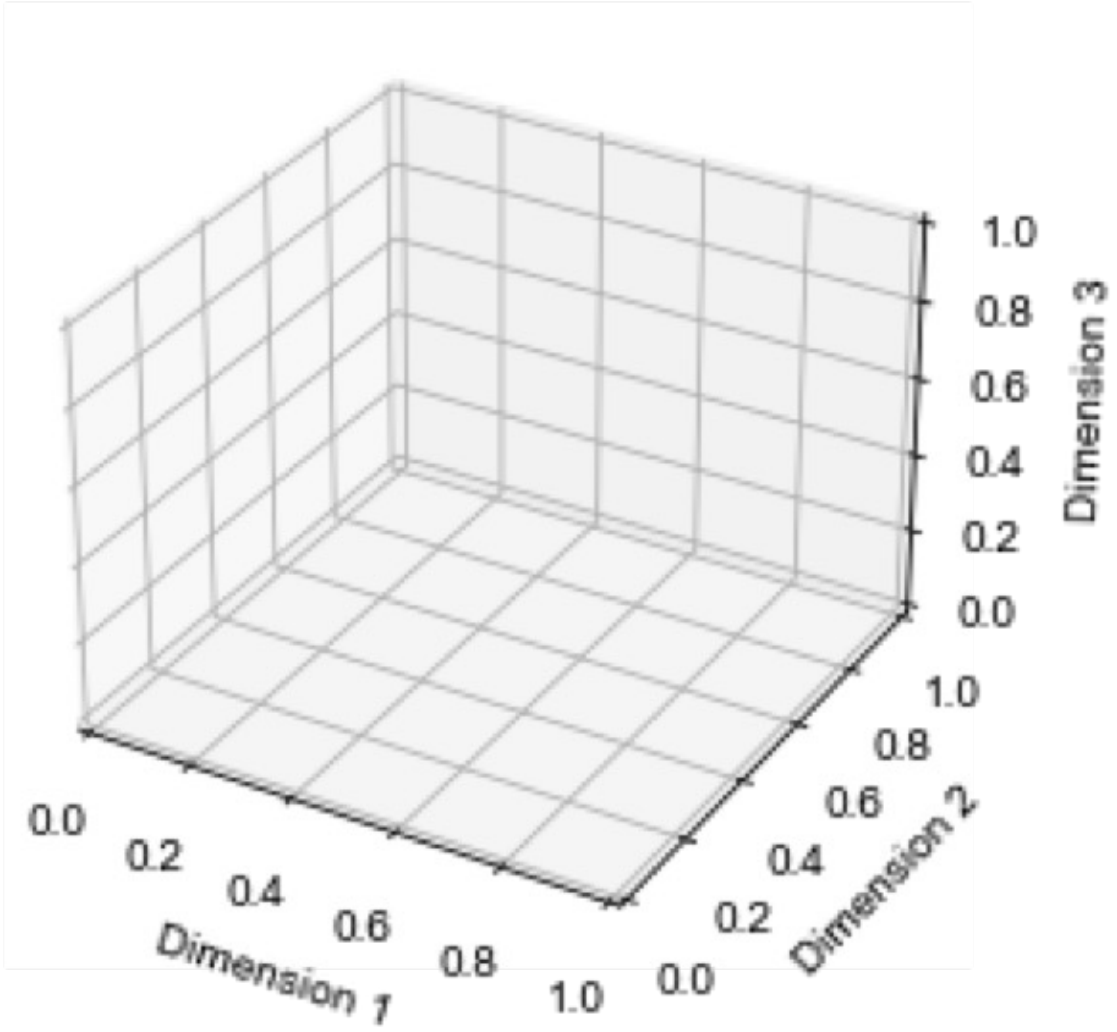
単語埋め込みモデルとは何か？

- テキスト内の単語の意味をベクトル表現し、数百次元の意味空間にマップする手法



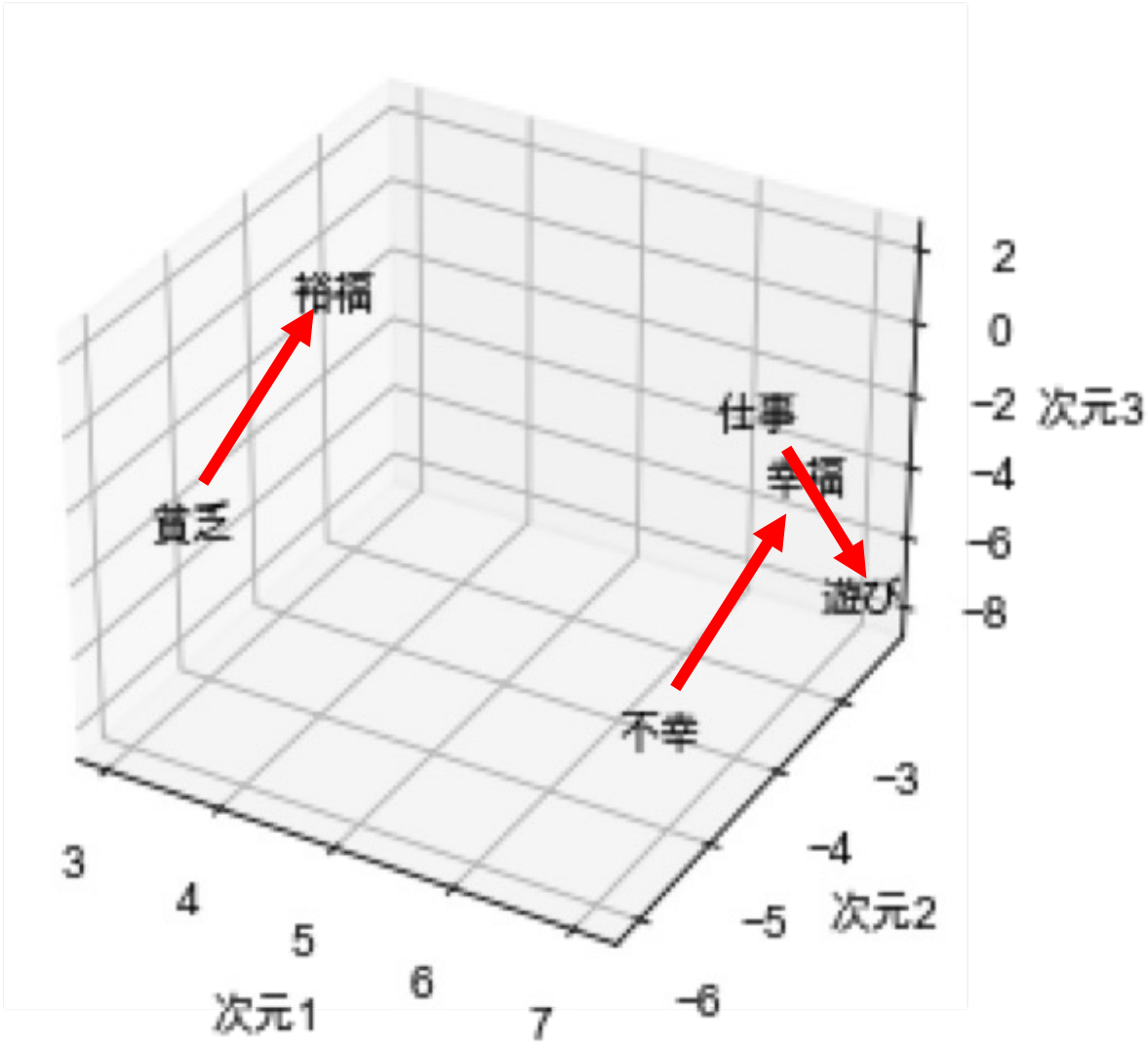
素の意味空間は解釈できない

?



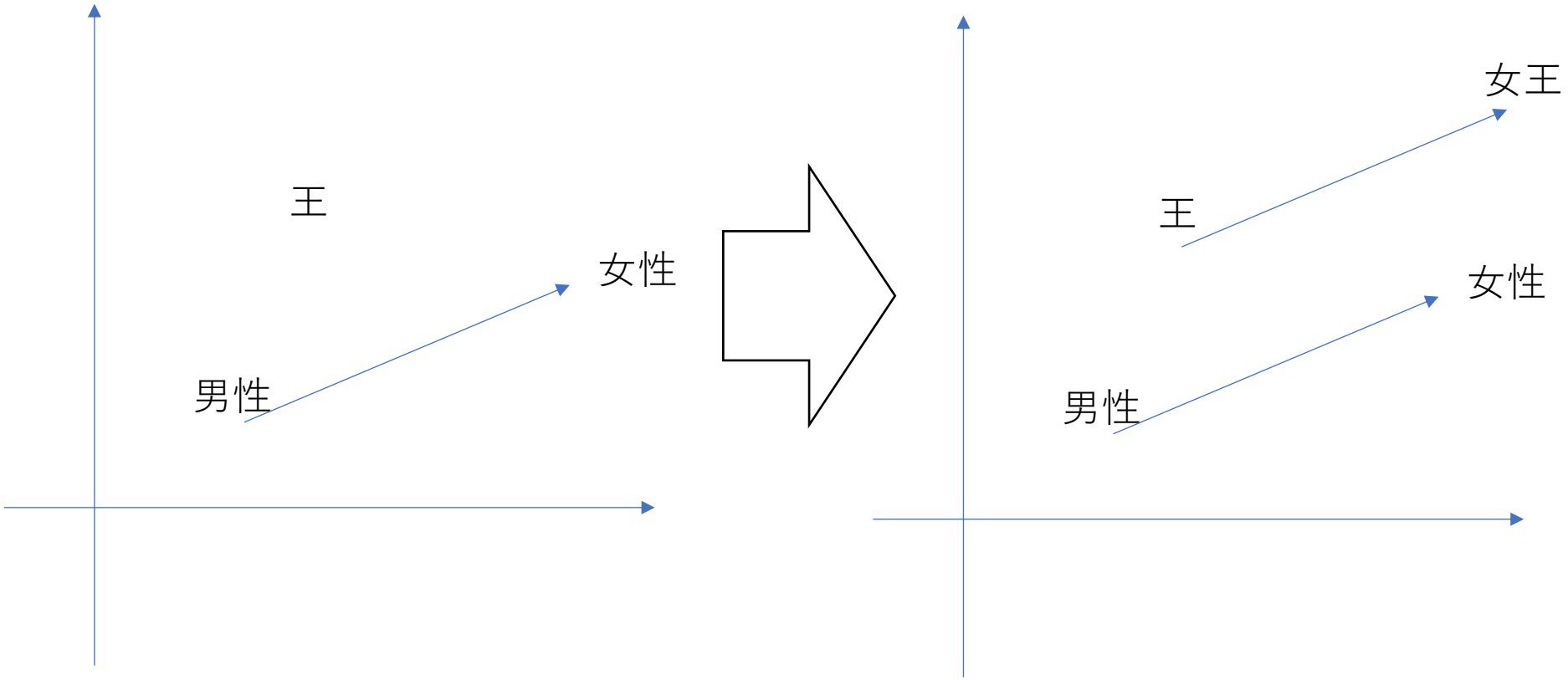
文化の幾何学アプローチ

- 解釈不可能な広大な意味空間に解釈可能な文化次元を創出する方法



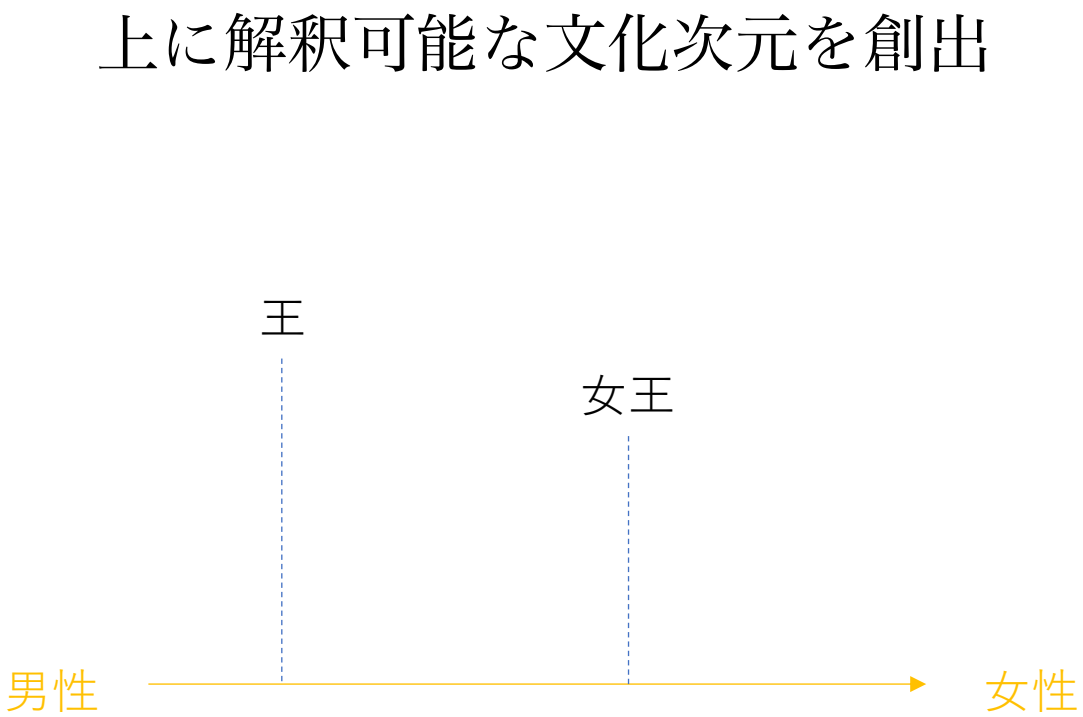
単語埋め込みモデルとアナロジー計算

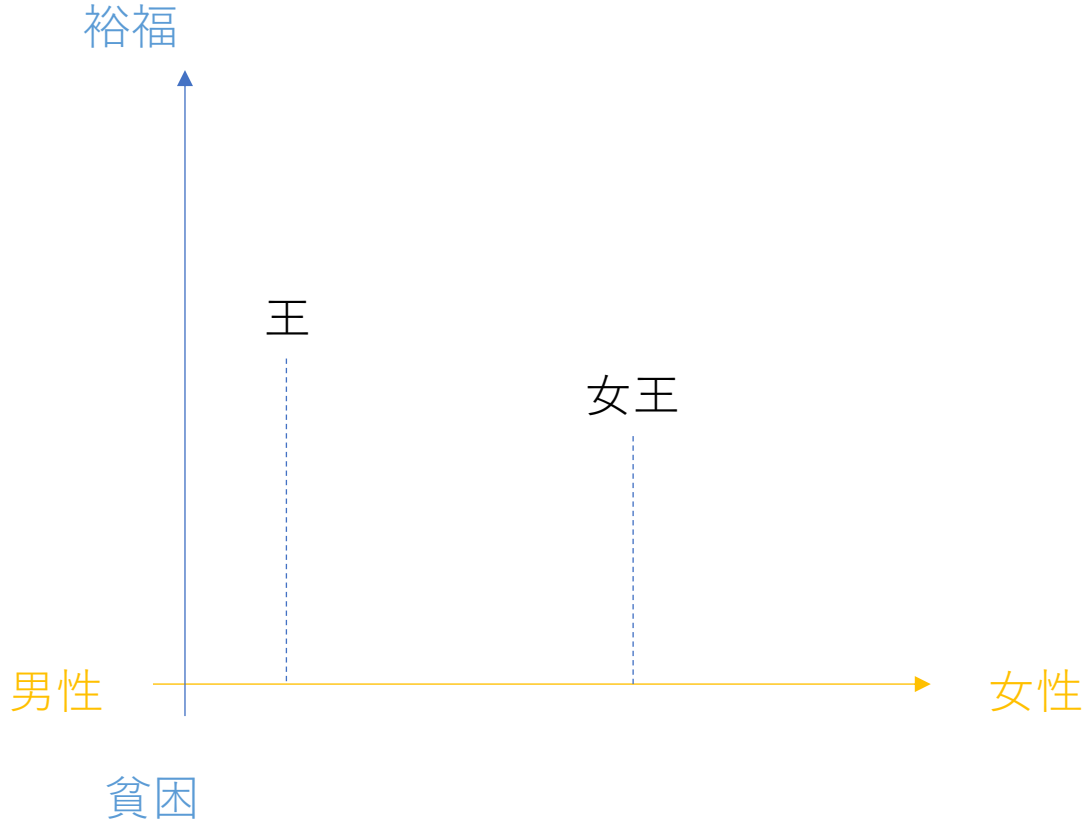
王　－　男性 + 女性　＝　女王



単語埋め込みモデルとアナロジー計算

- 対義語ベクトルのペア集合から意味空間上に解釈可能な文化次元を創出





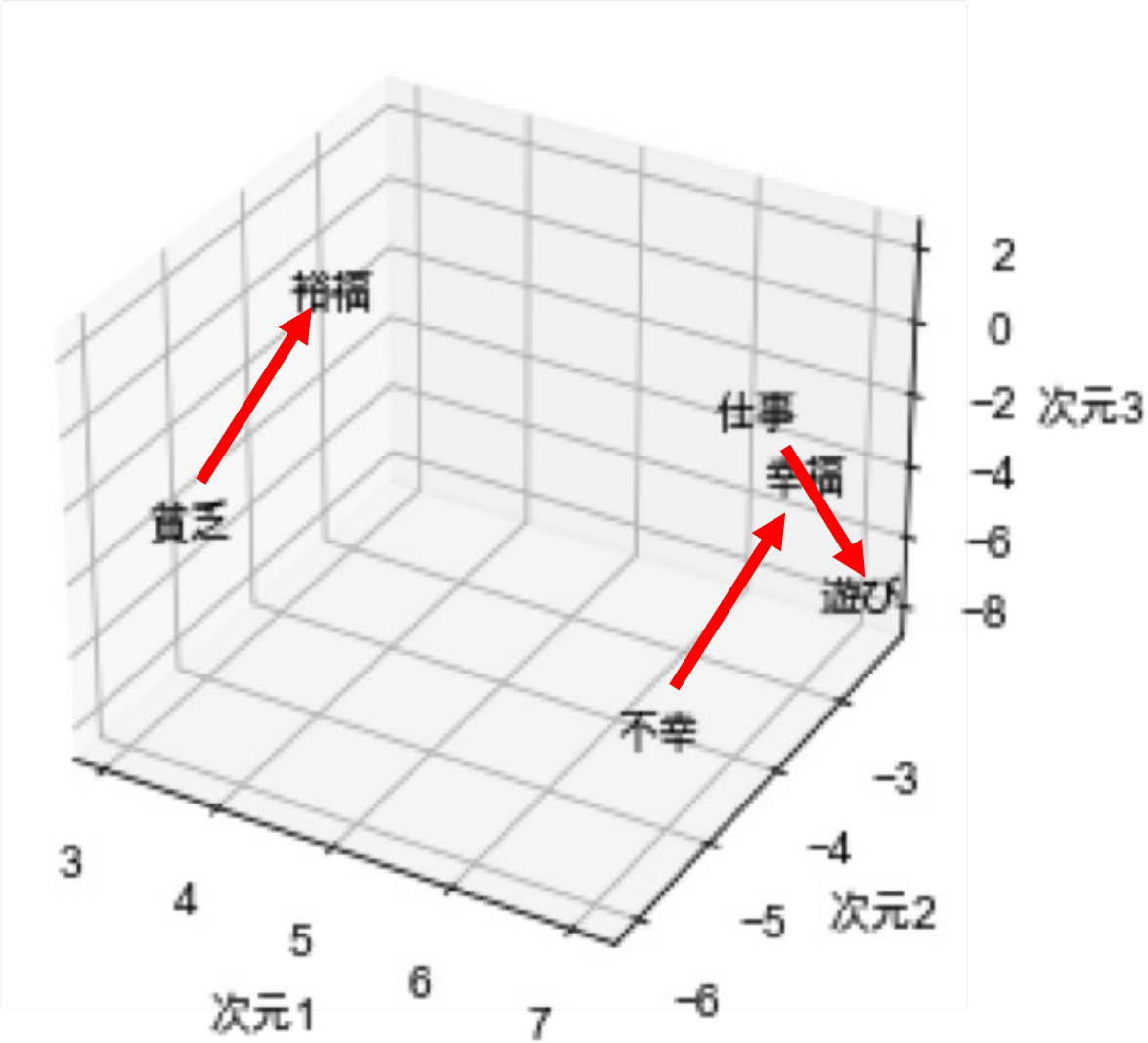
文化の幾何学アプローチ

幸福に関わる文化次元を対義語ベクトルから構築

不幸 → 幸福

貧乏 → 裕福

仕事 → 遊び



文化の幾何学アプローチによる幸福の多次元的意味の分析

- 全文データを用いることで、幸福をめぐる日本社会の複合的な集合表象を明らかにする。
- とくに、アンケートなどの存在しない戦前の幸福意識を明らかにし、歴史的変化を追跡する

この研究は、ムーンショット型研究開発事業：目標９研究開発プロジェクト「脳指標の個人間比較に基づく福祉と主体性の最大化」の一環として行われたものです。

国会図書館全文データ

https://dl.ndl.go.jp/ja/

国会図書館所蔵刊行物のデジタル化されたデータ
図書1910-1968
雑誌1910-1999
のテキストデータを利用

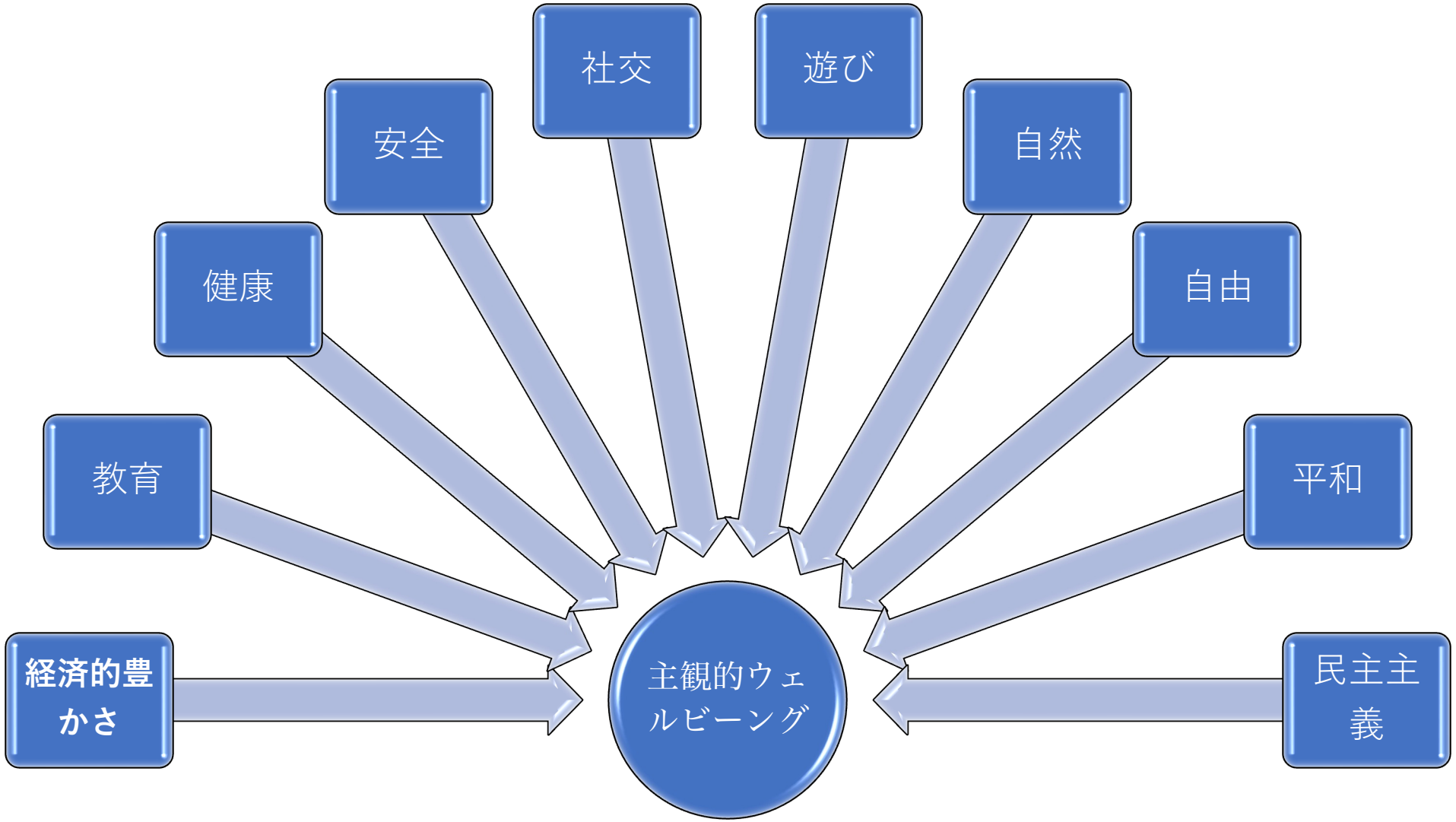
幸福の構成次元

- 哲学で提案されてきた理論（とくにM.Nussbaumの潜在能力理論）をもとに、社会学的な考察も加味して、右の幸福リストを提案
- 各概念を構成する対義語ベクトルのペア集合から幸福に関わる文化次元を構築

幸福	不幸
喜ぶ	不快
嬉しい	不愉快
楽しい	悲しい

幸福の構成次元リスト
主観的ウェルビーイング
経済的豊かさ
教育
健康
安全
社交
遊び
自然
自由
平和
民主主義

多次元幸福モデル



まとめ

- 国会図書館全文データから、幸福と関連すると想定される社交や経済的豊かさ、遊びなど10の意味次元と主観的ウェルビーングとの関連を分析した。
- ほとんどの意味次元は現時点で主観的ウェルビーングと関連していたが、遊びや民主主義など歴史的に大きく位置づけを変えた意味次元もあった。
- 全文の歴史的テキストを用いることで、アンケートなどでは分からない戦前の集合表象も明らかにできた。