Vol.2017-CH-113 No.2
2017/2/4

地域研究における論文と史料からの用語文脈の抽出

亀田尭宙[†1]

概要：ある地域、ある歴史上のできごとや、宗教的概念を論文や史料群の中から研究者が捉えることを情報技術で支援しようとしたとき、最も基本的なアプローチは検索を可能にすることである。しかし、そこから全体像を描くには、検索結果を逐一確認して検討する必要がある。そこで、対象となる用語の文脈を抽出し分類することで、全体像の把握をも支援することを試みた。本論文では、複数の事例に対して自然言語処理の基礎的な技術を適用しその結果を検討することで、実践的な支援の可能性と課題を探った。

1. 背景

東南アジア地域研究研究所（旧、地域研究統合情報センター。2017 年 1 月に東南アジア研究所と合併）では、多くの地域研究者を抱え、筆者は情報技術でその研究の支援に当たっている。その活動の中で「資料の全体像を把握したい」「資料の中における○○というものの位置づけが知りたい」といった要求は非常に多い。本稿では３つの事例を元に、特に後者の要求に対して、特定の用語の文脈を抽出するための方法論について検討する。

2. 事例 1: マレーシアを対象とした研究論文における地域間関係を可視化する

マレーシアを対象とした地域研究を行っている光成歩氏から頂いた手作業で整理された 1690 件のマレーシア研究の書誌情報（2004 年～2015 年）のうち、本文が利用可能な 323 ファイルを元に、その研究の全体像を描くための試みを行った。

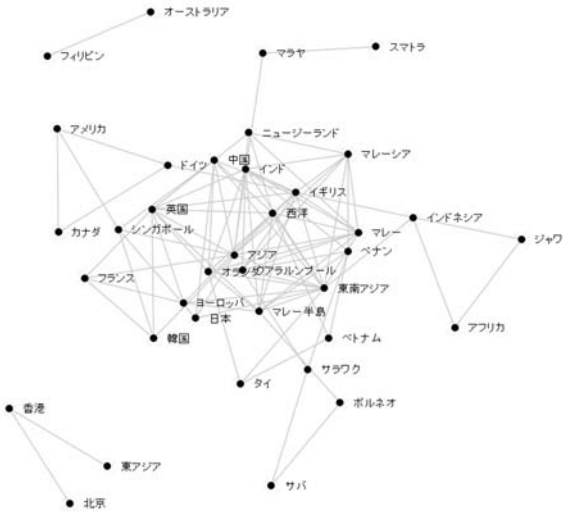

図 1　地名間の関係のネットワーク可視化

まずは、Latent Dirichlet Allocation [1] （トピック数 15、出現数 10 回以下の単語とストップワードの除去を行った）を用いて、各論文をトピックのベクトルで表現することで類似性を算出し、頻出キーワードに紐づいた論文のリストを提示する検索システムを構築した。この機能自体は、LDA に基づく分散表現で語を表現することで多義語や類似語や同義語の検索に対応しているため、適切な論文集合を提示しやすいという利点はあるが、特に地域研究者に発見をもたらすような効果は得られなかった。

そこで、論文の部分集合を任意の検索語を通してみるのではなく、論文の全体集合を通して特定の種類の用語の関連を見るアプローチに切り替えてみた。具体的には、論文に出現する地名を前述の LDA のトピックによる分散表現で表現し、それらのコサイン距離を元に閾値（0.84）以下の関係を無視し d3.js [a] を用いてネットワークとして可視化した（図１）。マレーシア研究の論文集合を用いているので、マレーシアが様々な地域と繋がっているさまが見て取れ、また一般的に地理的に遠い国が遠くに配置される一方で植民地時代の宗主国であるイギリスは近くに配置されるなど、納得の得られる結果が示された一方で、「中心部は全体的に関係が密でより細やかな関係を知ることができない」「意外な関係について、その文脈が知りたい」というフィードバックが得られた。それに関しては前述の検索システムに「ジャワ アフリカ」のように２地域名を入れることでそれらを強く結びつけている論文を知ることができるが、例えば「アジア・アフリカ言語文化研究」という雑誌の論文が「ジャワ」地域についての論文を出していて、各ページに記された雑誌名の「アフリカ」と文章中に頻出する「ジャワ」が共起してしまった影響が考えられる。論文 PDF を OCR にかけて用語抽出しているため、このような問題が生じている。こういった元データの問題による影響を調べるため、特定の雑誌（『東南アジア研究』）を対象に、ページ中に現れる雑誌名やページ番号などを省き、言語処理に適したように成形したデータセットを作成して比較する試みも行っている（図２、データセット作成のための支援シス

†1 京都大学 東南アジア地域研究研究所
Center for Southeast Asian Studies, Kyoto University

a) https://d3js.org/

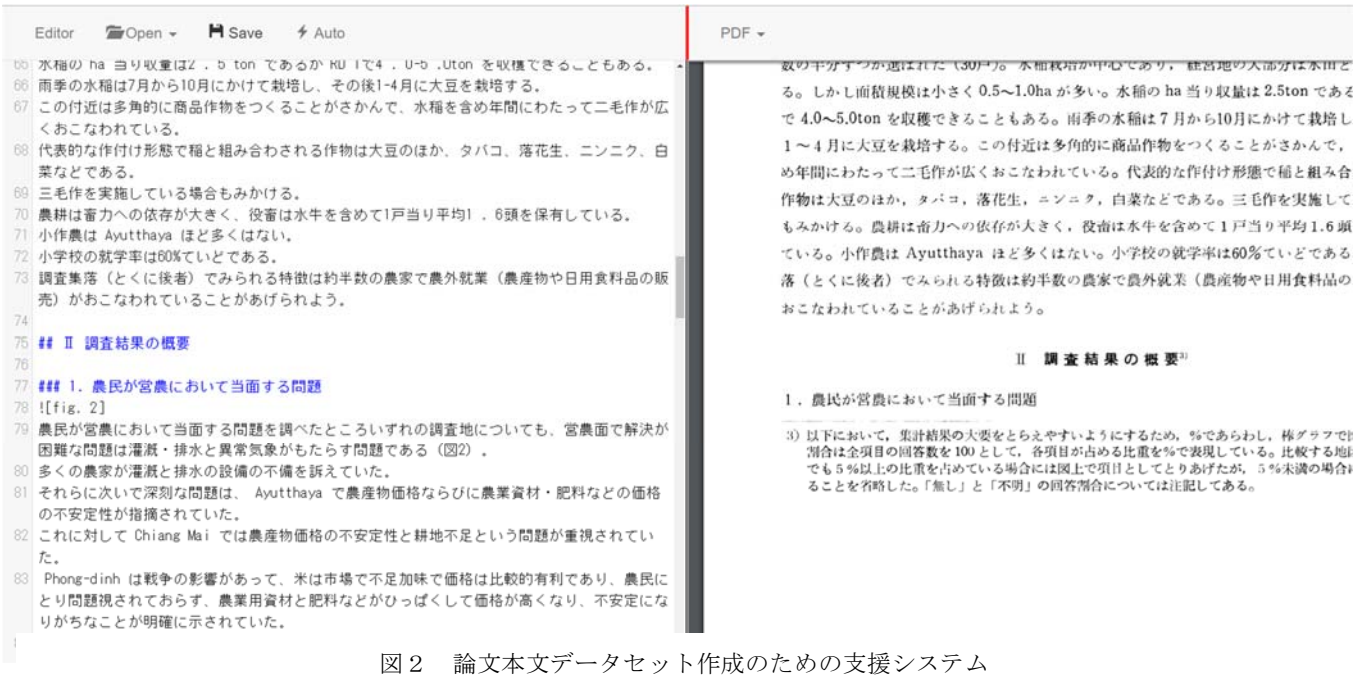

図２　論文本文データセット作成のための支援システム

テム)。

今後は、そういったノイズの除去に加え、論文単位のみではなく近い文で出現するかやその周りに他にどういった単語が出現しているかなども踏まえて適切に用語間の関係理解を支援する仕組みを作りたいと考えている。

3. 事例 2: 雑誌のＱＡコーナーから質問の型を抽出する

QALAM 雑誌記事データベース[b] には多数の QA が掲載されており、質問の多くは、何かの行為がイスラム法的に合法か違法かを問うている（例：「ネクタイ、帽子や膝の見える半ズボンを着用することは違法ですか。」[Qalam 1951.3]）[c]。他には事実知識を問うものもあれば（例：「国連会議に参加する代表は各国何人いますか。」[Qalam 1951.2])、人生相談もある（例：「女性はいつ結婚するのが最も良いですか。」[Qalam 1951.2])。

この中から、「ある行為がどのように捉えられてきたか」という問いに答えるには、まず、想定している行為を指し示す表現を同定し、それがどのように評価されているかといった評価表現を抽出することが必要になる。ここでは、評価表現の抽出のために、まずは表現の型の抽出に取り組んだ。

主な流れは以下のとおりである:

- 語の並びに着目するため、前処理として質問文を正規化した語の並びに変換する
- 各語の出現頻度など統計値を取り、それに基づいて、型の骨格を成す頻出語の列を得る
- 頻出語列の部分列に対し「型らしさ」を測る指標を作り、適用することで型を探す

データとしてはローマ字化されたマレー語（元はアラビア文字で書かれた Jawi と呼ばれる表記であるが、ほぼすべて Rumi と呼ばれるローマ字表記に翻刻されている）を対象とし、Apache Lucene の Stemmer [d] で正規化し、次のように語幹の列として質問文を表現するe

例[Qalam 1954.8:]

Meminta sedikit penjelasan tentang binatang sembelihan – qurbān yang biasa dikerjakan oleh orang Islam pada Hari Raya Haji.

→[“inta”,“sedikit”, ”jelas”, ”tentang”, ”binatang”, ”sembelih”, ”qurban”, ”yang”,”biasa”, ”kerja”, ”oleh”, ”orang”, ”islam”, ”pada”, ”hari”, ”raya”, ”haji”]

その後、今回は 40 回以上出現している 42 語を頻出語とみなし、各質問文の列から相当する部分列を抜き出した。その後、その出現頻度を加味して文に当該の頻出語列が出現したか／しないかについての二項分布の下側累積確率を全文について求めることで、頻出語列の型らしさを測定した。

その結果、２語だと [“apa”, “hukum”], [“agama”, “islam”] といった頻出語列がほぼ 100% [f] 型として共起しているという結果が得られ、４語の ["bagaimana", "hukum", "orang", "yang"] （4 件）もほぼ 100% という結果になった。一方で、[“yang”, “saya”]（12 件）は 13.9% 、[“hukum”, “ada”]

b http://majalahqalam.kyoto.jp/
c QALAM からの引用は[QALAM 年.月]で示す
d http://lucene.apache.org/core/5_4_1/analyzers-common/org/apache/lucene/analysis/id/package-summary.html
e はじめの Meminta は minta が語幹として正しいので Stemmer の処理が誤っているがそのまま記した。
f) プログラムの精度から、小数点以下 10 桁までしか見ておらず、その精度では 100% とみなされた

（11 件）は 64.1% となり、["bagaimana", "hukum"]（12 件）がほぼ 100% になったことを考えても、出現件数に比して型らしさは低いと考えられる。

既に課題として見つかっているのが、[“agama”, “islam”]のように、確かにこの質問文に特徴的な共起であるが、質問の型とは言えないペアも型として高く評価されてしまう点があり、少なくとも 1 つ以上の機能語を含むことを条件とすることを検討している。

4. 事例 3: 絵葉書のデータから地域を描写する

現在ラファイエット大学と京都大学で共同の絵葉書データベースを構築しており、3442 件の絵葉書のデータから例えば、大連もしくは大连（中国語簡体字）で得られる絵葉書は 140 件存在する。「大連とはどのような地域か」に答えるには、この検索結果を他の地域と比べて描写する、また検索結果の概要を描写する必要がある。

既に簡体字表記について触れたように、多言語のデータを扱っている場合に言語間の横断検索を可能にすることが必要な場合がある。その他にも表記ゆれや時代による地名の変遷などをどの程度考慮するかについても考える必要がある。しかし、異なる時代の地名は指し示す範囲も異なることが多いので、何をアイデンティティとするかは難しい問題である。

このデータはすべてメタデータが整備されているため、個々の地域と共起するメタデータの傾向抽出を試みている。

5. おわりに

地域研究における論文と史料からの用語文脈の抽出のために、

- 表記が多様であったり同綴異義があったりする用語の同定
- 文脈をカテゴライズするためのレトリックの同定
- 用語間の関係抽出
- それらを支えるためのデータセット作成

といったことが課題となっており、それぞれ上述のようなアプローチで取り組んでいる。今後、各研究の進捗と共に、それぞれのタスクのためのツールの公開や方法論の構築を進めていきたい。

謝辞

本研究は科研費 16K21124「情報抽出技術と LOD を用いた地域研究論文の構造化と分析」、国立情報学研究所公募型共同研究「地域研究における論文と史料からの知識抽出」の助成を受けたものです。また、「ジャウィ文献と社会」研究会の皆様には多くのフィードバックを頂きました、ありがとうございます。

参考文献

[1] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022, 2003.