N-015

地域研究史資料を対象とした時空間的特徴の抽出と場面の構造化
A Method of Spatio-Temporal Feature Extraction and Scene Structuring from Historical Materials of Area Study

山田 太造†
Taizo Yamada

1. はじめに

本研究では，地域研究に関する史資料，特にフィールドノートを対象とし，そこに記述されている場面の時空間的特徴を，テキストマイニング手法を用いることで，定性的ではなく定量的に表現・解析する手法の確立を目指している．特に，テキストから各場面を特徴付ける用語とともに地名・日付を抽出し，各場面に対する定量的解析可能なデータ表現を行った．フィールドノート内に出現する用語のうち，文章の文脈に依存した形式でのルールを作成し，このルールが適用される用語が地名であるかどうかの判定について SVM を用いて行った．

2. フィールドノート

フィールドで得られる観察記録やメモ，スケッチ，撮影した写真などは，現地のことを知る重要な研究資料[1]である．フィールドノートはいわば，これらを研究者が自分なりの情報収集方法を模索・整理し，記したものである．フィールドノートへの記載は，調査者が現地の調査の最中もしくはその直後に 行われ，しかも多くの場合，調査者が個人的に把握するために作成されるため，その利用においては，調査者以外が利用することは極めて困難である．そのため，フィールドノートを他人が断片化したときに，個々の情報のみから調査者が意図した情報を別の読者が十分に読み取ることができるとは必ずしも限らない[1]．フィールドノートを他者が利用可能とするためには，調査者が再整理を行う必要があると考えられる．

高谷好一著『地域研究アーカイブズ フィールドノート集成』（京都大学地域研究統合情報センター CIAS Discussion Paper Series）は，高谷氏の協力のもと，フィールドノートのテキスト化を行い，本人によりイラスト・写真の整理，各種資源の検証を行うことで作成された．これまでに『フィールドノート集成』として 8 冊（合計 4405 ページ）刊行され，その対象地域としては，東南アジア，インド，ヨーロッパ，アフリカなど様々である．

本研究では『フィールドノート集成』のうち「フィールドノート集成 2」にあるスマトラ（1984.10-19-1985.1.18）を用いた．該当のフィールドノートは，分量としては A4 サイズ 198 ページ程度であるが，スマトラ島全域をカバーしている．

3. 時空間的特徴の抽出

本研究で用いたフィールドノートは，基本的には調査者である高谷氏が調査・移動した場面の風景を時系列的に記している．本研究では，フィールドノートに記載された場面を単位とし，その場面の特徴を見出した．フィールドノート内にある記載としては，観察した場面の日時，場所，観察内容などである．そこで，時間情報・空間情報により識別される場面を，観察内容から場面の特徴を検出することにした．

3.1. 潜在トピックの検出

フィールドノートには，観察内容に何らかの話題がある．この話題はフィールドノートに明記されておらず，潜在しており，意味的には読解することで把握することになる．ここではこの潜在する話題を潜在トピックと呼ぶ．潜在トピックを検出し，このトピックに応じて場面を分類する．潜在トピック検出のため次式で表現される LDA[2]を用いる．

$$p(d|\alpha,\beta) = \int Dir(\theta|\alpha)\left(\prod_{n=1}^{|d|}\sum_{k=1}^{C} p(w_n|z_{k,\beta})p(z_k|\theta)\right)d\theta \quad (1)$$

ここでα, βはパラメータ，$z = z_1, z_2, \ldots, z_C$は潜在トピック，$\theta = \theta_1, \theta_2, \ldots, \theta_C$は潜在トピックの生成確率，$Dir(\theta|\alpha)$はディリクレ分布，$d = (w_1, w_2, \ldots, w_{|d|})$は場面，$w_n$は単語，|d|は場面 d の総用語数を示す．LDA は潜在トピックの生成確率がディリクレ分布に従うと仮定した文書生成モデルといえる．

6	unk
中国人	名詞,一般,*,*,*,*,中国人,チュウゴクジン,チューゴクジン
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
日本	名詞,固有名詞,地域,国,*,*,日本,ニッポン,ニッポン
軍	名詞,接尾,一般,*,*,*,軍,グン,グン
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ
来る	動詞,自立,*,*,カ変・来ル,基本形,来る,クル,クル
と	助詞,接続助詞,*,*,*,*,と,ト,ト
Selat	unk
panjang	unk
から	助詞,格助詞,一般,*,*,*,から,カラ,カラ
逃げ	動詞,自立,*,*,一段,連用形,逃げる,ニゲ,ニゲ
て	助詞,接続助詞,*,*,*,*,て,テ,テ
いっ	動詞,非自立,*,*,五段・カ行促音便,連用タ接続,いく,イッ,イッ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。	記号,句点,*,*,*,*,。,。,。
EOS	

unk : 未知語

図 1 ：形態素解析結果の例

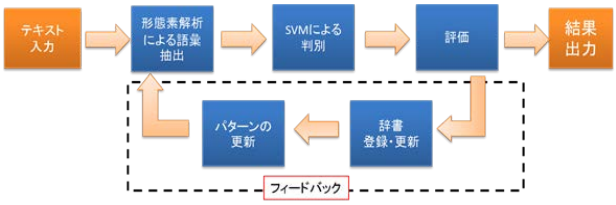

図 2 ：地名抽出フロー

†東京大学史料編纂所

パターン1：	ここ	（より|で）	＜地名＞	
パターン2：		ここ	＜地名＞	
パターン3：			＜地名＞	（出発|着|泊|川|湖|…|中心）
パターン4：			＜地名＞	（より|から|まで|へ）
パターン5：			＜地名＞	＜記号，（句点｜読点）＞
パターン6：			＜地名＞	（の|に|を） （町|帰る|東|西|南|北|港|旧港|移動|到着|向かう|方向|入る|泊まる|泊|行く|経|川岸）
パターン7：		＜行頭＞	＜地名＞	（＜終点＞|＜記号，（句点|読点）＞）
パターン8：		（小村）	＜地名＞	

図 3：地名抽出パターン

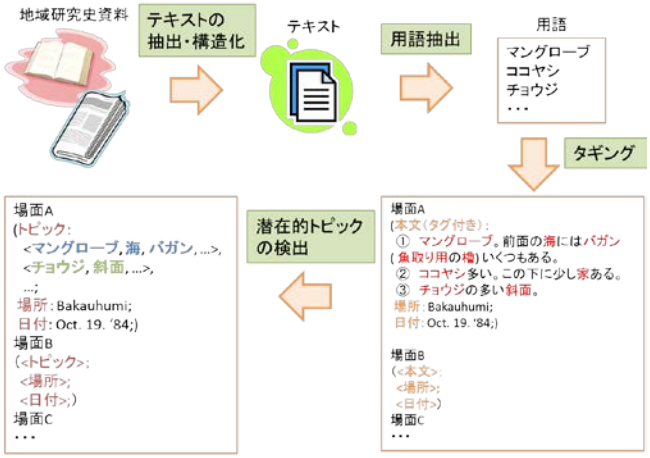

図 4：場面の構造化

本研究に当てはめた場合，LDA は，1 場面におけるトピックは複数あり，トピックはそれごとに複数の用語を生成することをモデル化している．（1）式をそのまま計算することはかなり困難であるが，崩壊形ギブスサンプラを用いた解法が知られており[3]，本研究ではこれを用いて潜在トピックを算出する．

3.2. 用語の抽出

(1)式を用いるためには用語が必要であるため，観察等による記述内容から用語を抽出する必要がある．そこで，本研究では形態素解析による抽出を行った．ここで，形態素解析器として mecab，形態素解析用辞書として IPADic を用いた．形態素解析によるすべてを用語とはせず，名詞を対象とし，代名詞・数・接尾・副詞可能・形容動詞語幹・ナイ形容詞語幹・接続詞的・非自立は対象外とした．また，連続する名詞はチャンクした．結果として，用語の異なり数は 5,666 だった．

3.3. 時空間語彙の検出

フィールドノートには記載した日付があるため，これを時間語彙として用いた．空間語彙としては，3.2.により抽出した用語のうち，地名であるもの，および未知語であるが地名であると思われるものを用いた．図 1 は例として形態素解析の結果を示す．この例では，”日本”は地名ではあるが，チャンクの結果，地名と適せず一般名詞として扱うことができる．また”Selat panjang”のように形態素解析の結果としては未知語ではあるが実際は地名を示す場合がある．未知語が地名であるかどうかを判定するため，地名が出現する文章のシーケンスパターンをもとに，SVM（Support Vector Machine[4]）を用いて地名かどうかの評価を行い，地名として評価されたものを地名として用いた．このフローを図 1 に示す．また図 2 はシーケンスパターンを示す．

本研究における地名抽出の精度としては約 0.633 だった．しかしながら，図 1 に示すように，SVM での評価に対してフィードバックし，改めてシーケンスパターンを追加することでこの精度は向上していく．1 回あたり 10 件の結果に対して再評価を行い，これを 5 回繰り返したところ，約 0.76 まで精度が向上した．

4. 場面の構造化

図 3 は本研究におけるフィールドノートにおける場面の構造化のフローを示す．

テキストの抽出・構造化では 2 節で述べたように，『フィールドノート集成』作成時のテキストを用いた．これをベースに，記述内容を日付単位に区切り，さらに場面ごとに区切った．

用語抽出・地名抽出・タギング・潜在トピックの検出は 3 節で述べた手法により行った．

5. おわりに

本研究では，これまで複数人で利用するなど，広く利用されることがほとんど無かったフィールドノートに対し，時間語彙・空間語彙とともにトピックモデルを用いてテキスト特徴づける手法について述べた．今後は，フィールドノートという資料を地域研究推進において重要な素材として更に高めていく．フィールドノートを利活用していくうえで，単なる全文検索可能な公開方法ではなく，テキスト分析の成果を広く利活用可能な形式でのデータ構造を探求していくことも重要だろうと考えている．

謝辞

本研究の一部は，JSPS 科研費 15H01723，26730167 の助成を受けたものです．

参考文献

[1]柳澤雅之．フィールドノート・プロジェクト，Seeder11 号，pp.14-22，2014.
[2]D. M. Blei, A. Y. Ng, and M. I. Jordan: “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol.3, pp.993-1022, 2003.
[3]T. L. Griffiths and M. Steyvers: “Finding scientific topics,” Proc. of the National Academy of Sciences of the United States of America, vol.101, pp.5228-5235, 2004.
[4] Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, Information Processing and Management: an International Journal, Vol. 24, No. 5, pp. 513–523 (1988).