

時間的推移を考慮した報道トピックの抽出と再構成

松本直彰¹ 湯本高行¹ 山本岳洋¹ 大島裕明¹

概要: 大規模な震災など影響が長期間にわたる報道では、トピックは少しずつ変化しているが、その推移を把握することができれば、理解をより深めることができる。時系列テキストデータに対するトピックモデルとしては DTM があるが、トピックの消失や途中での発生を適切に表現することはできない。そこで本研究では、期間ごとに LDA を用いて求めたトピックに対して、同一期間内のトピックの集約と次期間のトピックへの推移関係の発見を行う手法を提案する。提案手法では、トピックの特徴ベクトルをそのトピックを構成する主要語の生成確率を基に作成し、それらのコサイン類似度によってトピック間の類似性を表現する。本研究では阪神・淡路大震災に関する記事を用いて解析を行った。その結果、「住宅」が主要語となったトピックなどでは集約や推移が適切に行えていたが、しきい値付近のトピックで過剰に集約を行っているものがあつた。

1. はじめに

生活の中で世の中に大きな影響を与える事件や出来事が起こった場合、新聞記事などで長期的な報道が行われることがある。例えば、災害の発生などでは生活インフラの復旧、街の復興、被災者の心のケア、というように長期的な報道が行われる。その長期的な報道において報道期間内にトピック内容は少しずつ変化していき、長期報道全体の理解のためにはその推移を把握することは必須である。

しかし、長期的な報道を人の手で処理することは困難である。例えば、阪神・淡路大震災は 2020 年で発生から 25 年となる。よって人手で報道内容の変化をとらえるには 25 年間分の報道を把握する必要があるが、現実的ではなく人の手に頼らない処理システムの構築が必要になってくる。

本研究では、トピックモデルを用いて、ある長期的な報道を全く知らない人物を対象に報道トピックの推移を視覚的にわかりやすく提示することを目的とする。そのためには報道トピックの展開と収束をとらえることが必要である。

まず先行手法として時系列を考慮したトピックモデルである DTM での解析を行ったところ、途中出現、途中消滅するトピックを表現することができないという問題があつた。

そこで、本研究では、トピックの類似性から集約と推移関係の発見を行う手法を提案する。提案手法では、一定期間で分割されたデータセットごとに LDA で解析を行い、トピックの語の生成確率で語の分散表現を重みづけしたベ

クトルをトピックの特徴ベクトルとして用いる。トピックの特徴ベクトル間の類似度から、同期間で類似しているトピックは集約し、次期間で類似しているトピック間には推移関係があるとみなす。トピックの集約を行うことで途中出現・消滅するトピックにも対応することが可能になる。実験では、提案手法を用いて阪神・淡路大震災に関する 25 年分の新聞記事を解析し、定性的および定量的評価を行う。

2. 先行手法での解析

時系列データを扱うトピックモデルは DTM[1] などがある。そこで長期報道に含まれるトピックの時間的な推移をとらえるために、先行手法として DTM を用いた解析を行う。

2.1 新聞記事データの前処理

2.1.1 データセット

神戸新聞社より提供された新聞記事データを用いた。集計期間は 1995 年 1 月から 2019 年 8 月まで、記事総数は 1,361,665 件である。

本研究では、提供された新聞記事の中でも阪神淡路大震災の記事に着目して解析を行う。従って、新聞記事データから阪神淡路大震災に関連する記事を抽出してデータセットを作成しなければならない。政府が 1995 年 2 月 14 日に「阪神・淡路大震災」と名称を閣議決定するまでの期間においては「震災」「被災」「地震」の三つのキーワードがタイトルまたは本文に含まれる記事を抽出した。そして、2 月 14 日以降は表 1 で示す「阪神・淡路大震災」の表記揺れをキーワードとして、タイトルまたは本文に含まれる記

¹ 兵庫県立大学

表 1 震災記事抽出のキーワード
1995 年 1 月 17 日から 1995 年 2 月 14 日まで
震災, 被災, 地震

1995 年 2 月 14 日以降
兵庫県南部地震, 阪神大震災, 関西大震災,
神戸大震災, 神戸地震, 関西大地震, 阪神淡路大震災,
阪神・淡路大震災, 南兵庫大地震

表 2 記事の分割期間

期間名	期間	備考
EO1-w1.1	1995 年 01 月 17 日 -1995 年 01 月 23 日	阪神淡路大震災 発生一週目
w1.1-w1.2	1995 年 01 月 24 日 -1995 年 02 月 16 日	阪神淡路大震災 発生二週目
w1.2-y1.1	1995 年 02 月 17 日 -1996 年 01 月 16 日	二週目から一年
y1.1-y1.5	1996 年 01 月 17 日 -2000 年 01 月 16 日	一年から五年
y1.5-y1.10	2000 年 01 月 17 日 -2005 年 01 月 16 日	五年から十年
y1.10-EO2	2005 年 01 月 17 日 -2011 年 03 月 10 日	十年から 東日本大震災発生前日
EO2-y2.3	2011 年 03 月 11 日 -2014 年 01 月 16 日	東日本大震災発生直後 から三年目
y2.3-y2.6	2014 年 01 月 17 日 -2017 年 01 月 16 日	東日本大震災発生 三年目から六年目
y2.6-now	2017 年 01 月 17 日 -2019 年 08 月 17 日	東日本大震災発生 六年目から現在

表 3 命名ルール

期間名	詳細
EO1	阪神淡路大震災発生
EO2	東日本大震災発生
w(a)-(b)	(a)[EO1 or EO2] から (b) 週目
y(a)-(b)	(a)[EO1 or EO2] から (b) 年目

事を出した。作成したデータセットの記事総数は 57,741 件となった。

2.1.2 期間区分と主要語抽出

DTM で解析を行うために、事前にデータセットを適切な期間に区切る必要がある。そこで、データセットとともに神戸新聞社から提供された「新聞記者が節目と感じた期間」を参考に期間区分を設定した。その期間区分を表 2 に、期間名の命名ルールを表 3 に示す。「EO1」を阪神・淡路大震災発生時、「EO2」を東日本大震災発生時とし、「w」を週目、「y」を年目とした。例えば、「w1.2」なら阪神・淡路大震災発生から 2 週目であり、「y2.3」なら東日本大震災発生から 3 年目である。

また、記事の中で主要な役割を果たしている単語のみを DTM で解析に用いる。本実験では記事中の「一般名詞」「固有名詞」「サ変接続名詞」を主要語とする。ただし、その中でも「人名」「地名」に属する語は具体性が強く、個人

名などは重要でないにもかかわらず多数出現するので、解析に用いないこととした。

2.2 結果と考察

トピック数を 10 個として DTM を用いて解析を行った。各トピックの単語の生成確率を縦軸、時間経過を横軸として最大生成確率が 0.005 以上の単語の生成確率の時間的推移を表した。図 1 はあるトピックの生成確率の推移について示したものである。

図 1 で示すように、DTM を用いた解析では単語の重要度の変化が確認できる。時間が経過するにつれて「調査」や「復旧」の生成確率が下がっているのに対し、「防災」や「訓練」の生成確率は上がっている。したがって、人々は震災からの「復旧」が完了し「防災」に目を向けていることがうかがえる。

しかし、解析で分かるのは単語の重要度の変化のみであり、トピックの出現や消滅は表現できない。DTM では時系列の初めから終わりまでを通してトピック数が変化しないものとして扱われる。最初の期間に存在していたトピックが途中で消滅してしまっても、存在するものとして解析が進んでしまう。これでは、長期間の報道において途中出現、途中消滅する話題を正確に追うことができない。

3. 提案手法

先行手法の DTM では長期報道の途中で出現したり、消失したりするトピックを表現することができない。従って、本研究では LDA[2] を基にした新しい手法を提案する。

図 2 は本研究が提案する手法のイメージである。本研究では、トピックの特徴を表現するためにトピックごとにベクトルを作成することにした。トピックを特徴ベクトルで表現すれば、それを用いてトピック同士の類似性を求めることができる。例えば期間 2 の Topic1 と Topic2 のように、同期間内で類似しているトピックの集約を行ったり、期間 1 の Topic1 と期間 2 の Topic1,2 のように、次期間のトピックへの推移などを確認できると考えた。

3.1 トピックの特徴ベクトル表現

LDA を用いて解析を行うと、各トピックの語の生成確率を得ることができる。そして、生成確率にしきい値を設け、しきい値以上の語と生成確率を用いてベクトルを作成する。

2.1.1 で作成したデータセットを用いて Word2Vec[3] で学習を行い、語の分散表現を求める。この分散表現を用いてそれぞれのトピックの特徴ベクトルを式 (1) のように表現する。

$$\mathbf{T} = \sum_{i=0}^W \mathbf{w}_i \cdot g_i \quad (1)$$

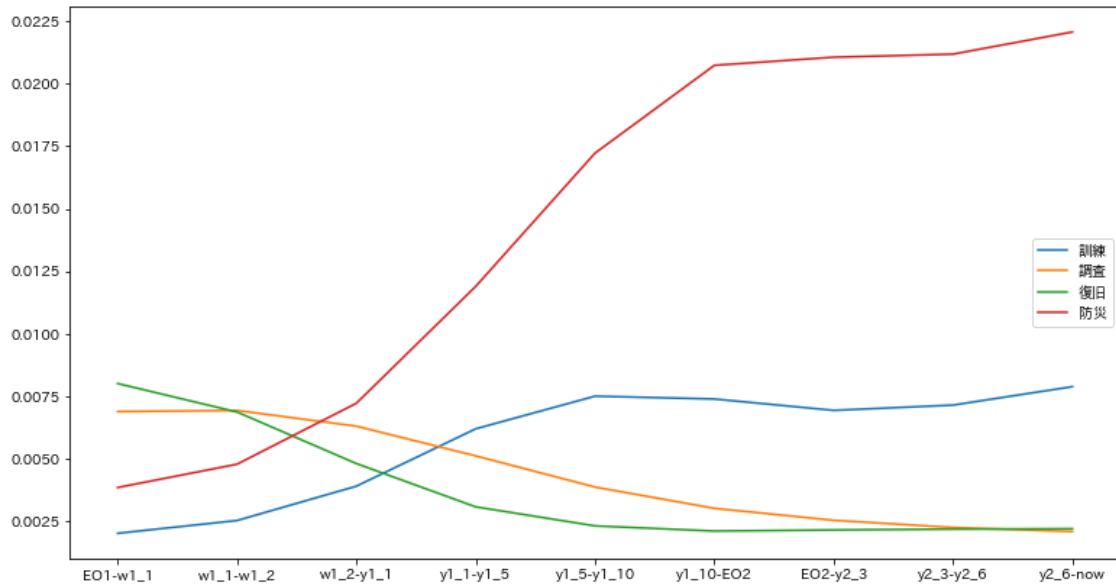
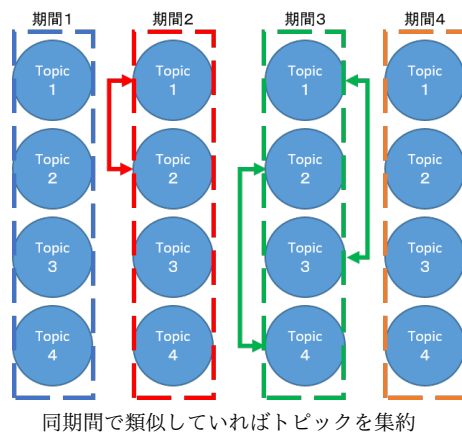
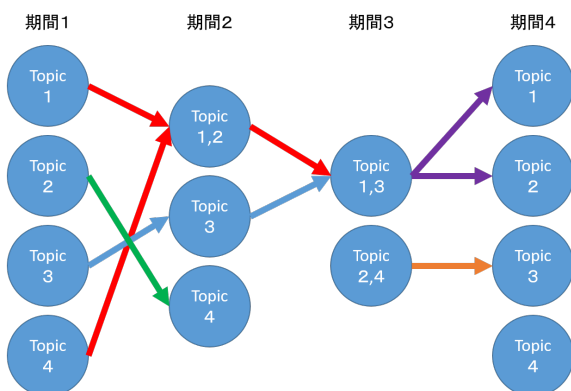


図 1 あるトピック内の単語の重要度の変化



同期間で類似していればトピックを集約



次期間に類似してトピックがあればトピックを推移

図 2 各期間で LDA を行う解析のイメージ

ここで、 $\mathbf{T}$ はトピックの特徴ベクトル、 $\mathbf{W}$ はそのトピックにおいて生成確率がしきい値 θ_f 以上の語の集合、 $\mathbf{w}_i$ は語を分散表現で表したベクトル、 g_i はその語の生成確率である。ただし、生成確率がしきい値以上の語かつ、同期間

の全てのトピックに共通して出現している語はベクトル生成に用いなかった。これは、その期間に広く話題となっている語であり、各トピックの固有の特徴を表現するには不必要だと判断したためである。

3.2 同期間内での類似トピックの集約

予想されるトピック数が事前に予想できない場合、トピック数を多く設定して LDA の解析を行うことがある。これは、適切なトピック数以上のトピック数で解析することで類似したトピックを後から集約することができるためである。

特徴ベクトル同士のコサイン類似度が θ_{sim} 以上となるトピックを類似しているトピックとみなし、集約する。手順は次の通りである。

1) LDA の解析の結果からトピックごとの特徴ベクトルを作成し、それぞれのトピック同士のコサイン類似度を算出する。

2) 類似度が θ_{sim} 以上のトピック同士を類似しているトピックと見なす。ここで、複数のトピックで類似性が確認されるトピック同士は集約を行う。二つのトピック間のみで類似している場合は図 3(a) のように単純に集約する。複数のトピックで類似している場合は、すべてのトピックにおいてそれぞれ自分以外とすべて類似しているのであれば図 3(b) のように単純に集約する。ただし複数のトピックの類似でも図 3(c) のような間接的な類似は認めず、それぞれ類似しているトピック同士のみで集約をするものとする。

3) 類似しているトピックの特徴ベクトルの総和をとり、集約後の新たなトピックとする。集約前のトピックは破棄する。

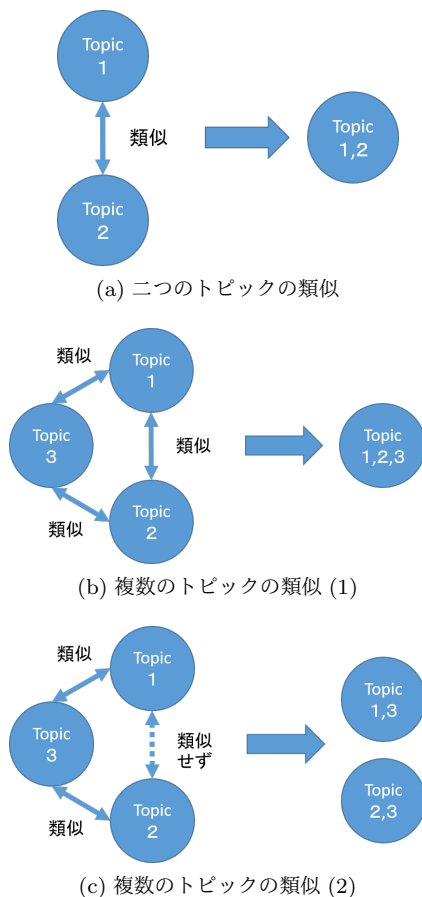


図 3 トピックの集約

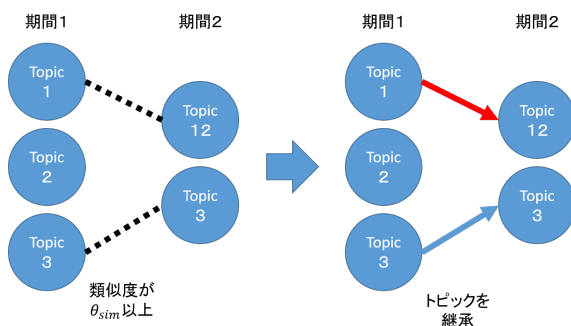


図 4 次期間のトピックの推移

3.3 次期間へのトピックの推移

現在のトピックが次の期間のトピックにどのように推移するかを、コサイン類似度によるトピック同士の類似性の評価で求める。図 4 は次期間のトピックへトピックが推移する様子を示したものである。これも、コサイン類似度の値が θ_{sim} 以上のトピック同士は前の期間から後の期間へと話題を推移しているものとする。

4. 実験

それぞれの期間において LDA を用いた解析を行い、トピックごとの特徴ベクトルを作成した。なお、トピック数は 10 個、解析時に表 4 のストップワードを用い、 $\theta_f = 0.003$ 、 $\theta_{sim} = 0.8$ とした。

表 4 解析時のストップワード

兵庫県南部地震、阪神大震災、関西大震災、
神戸大震災、神戸地震、関西大地震、阪神淡路大震災、
阪神・淡路大震災、南兵庫大地震、神戸、神戸市、
兵庫県、写真、説明、地域

表 5 「避難」の出現数の変化

期間名	総出現回数	1 記事あたりの平均出現回数
EO1-w1.1	178	0.4395
w1.1-w1.2	738	0.3284
w1.2-y1.1	922	0.1779

実験結果を直感的に分かりやすくするために、ワードクラウドを用いて結果を表現する。

4.1 可視化結果の評価

図 5 にワードクラウドを用いたトピックの遷移の様子の一部を示す。図 5 より提案手法で期間ごとにトピックの移り変わりが追えていることがうかがえる。例えば、「w1.2-y1.1」のトピック 5.8 や「y1.1-y1.5」のトピック 1 などは復興「住宅」を主軸にしたトピックだと推測できる。また、「w1.1-w1.2」のトピック 0.4 や「w1.2-y1.1」のトピック 1、「y1.1-y1.5」のトピック 3.5 は「復興」を主軸にしたトピックだと推測できる。どちらも、提案手法で次の期間への推移が表現できた。また「住宅」の二つのトピックを見比べてみると、「w1.2-y1.1」のトピック 5.8 では「事業」の「計画」を立てていることがうかがえ、「y1.1-y1.5」のトピック 1 では住宅の「再建」や「支援」が実際に行われていることがうかがえる。従って、本研究の目的の一つである、一つの長期的な報道の中で少しずつ変化するトピックが提案手法でとらえることができていると考える。

次に「w1.1-w1.2」のトピック 1.8 を見ると、「被災者」の「避難」「生活」についてのトピックだと推測できる。表 5 より、「避難」は「w1.2-y1.1」で 1 記事あたりの平均出現回数が前の期間と比べて大きく減少している。このトピックは「w1.2-y1.1」のどのトピックにも推移しなかったことから、震災発生 1 週目から 2 週目で「避難」のトピックは途中消失したと考えられる。また「w1.2-y1.1」のトピック 3.7 は、「防災」「対策」についてのトピックだと推測できる。表 6 より、「防災」は「w1.2-y1.1」で 1 記事あたりの平均出現回数が前の期間と比べて大きく上昇している。このトピックは「w1.1-w1.2」のどのトピックからも推移してこなかったことから、震災発生 2 週目から 1 年目で途中出現したトピックだと考えられる。これらのことから、途中発生したトピックや途中消滅したトピックの発見についても提案手法は対応していると言える。

しかし、本実験においてトピック集約が適切に表現できていないものはいくつか見受けられた。例えば、「w1.2-y1.1」のトピック 6.9 などが挙げられる。図 6 はトピック 6.9 の

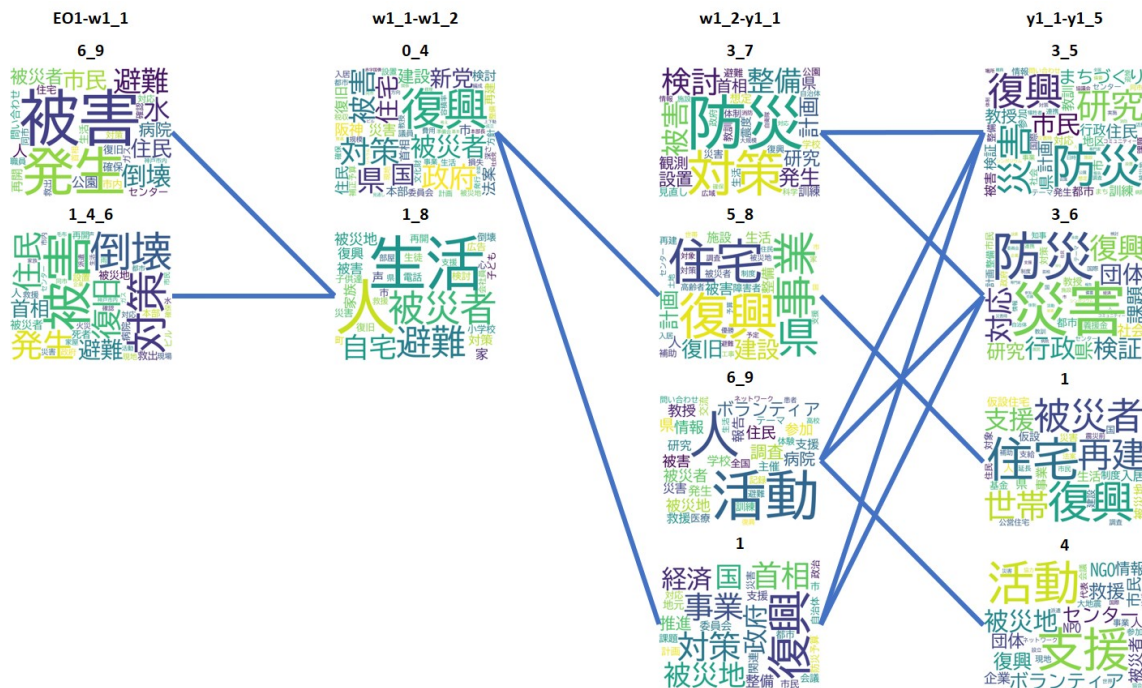


図 5 「EO1-w1.1」から「y1.1-y1.5」におけるトピックの遷移の例

表 6 「防災」の出現数の変化

期間名	総出現回数	1 記事あたりの平均出現回数
EO1-w1.1	39	0.0962
w1.1-w1.2	351	0.1562
w1.2-y1.1	1145	0.2209

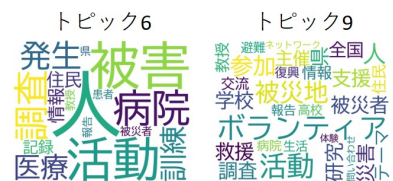


図 6 不適切な集約（「w1.2-y1.1」のトピック 6.9）

集約前のトピックをワードクラウドで表したものである。トピック 6 は「被害」「病院」のトピックであるのに対し、トピック 5 は「ボランティア」のトピックとなっている。これらを類似したトピックとして集約するのは適切でないと考える。このような集約が起こった原因として、類似とみなすしきい値の設定が挙げられる。このトピック 6 とトピック 9 のコサイン類似度は 0.81 であり、しきい値に近い値だった。このような不適切な集約は、類似とみなすしきい値を適切に決めることができなければなくすことができると考える。

また、類似度が高いが集約が適切に行えていないものもあった。例えば、図 7 に示す「y1.1-y1.5」のトピック 3.5 などが挙げられる。トピック 3 は「防災」「被害」のトピックであるのに対し、トピック 5 は「まちづくり」のトピックとなっている。トピック 3 とトピック 5 のコサイン類似度は 0.81 であり、しきい値を大きく上回る類似度であるが、これらも類似したトピックとして集約するのは適切でないように思える。これは、トピック 5 が様々な要素を一樣に含むトピックになっていることが原因だと考えられる。よって、ベクトル作成時に局所的に使われている語を注視し、広域的に使われている語の影響を少なくする idf のような処理をすれば改善すると考える。

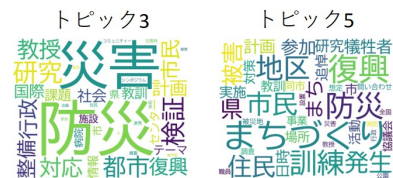


図 7 不適切な集約（「y1.1-y1.5」のトピック 3.5）

4.2 集約結果の評価

トピックの一貫性を定量的に評価する指標としてコヒーレンス [4][5][6][7] がよく用いられる。トピックの集約において、集約前のトピック同士が類似したトピックならば集約後も一貫性が保たれるはずである。従って、集約後のトピックのコヒーレンスは集約前のトピックのコヒーレンスと比較して大きくは悪化しないと考える。

本実験では式 (2) で定義される u_mass コヒーレンス [5] を用いて同期間のトピックの集約を評価した。 u_mass コヒーレンスではトピックを代表する特徴語として生成確率の上位数件の語を選ぶ。

$$C(t; V^{(t)}) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \quad (2)$$

t はあるトピック、 $D(v)$ は単語出現文書数、 $D(v_1, v_2)$ は単

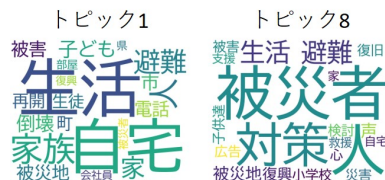


図 8 適切な集約（「w1.1-w1.2」のトピック 1.8）

表 7 コヒーレンス計算結果（「w1.1-w1.2」のトピック 1.8）

トピック名	コヒーレンス	集約後との差
1	-2.822	-0.071
8	-2.612	0.139
1.8	-2.683	-

表 8 コヒーレンス計算結果（「w1.2-y1.1」のトピック 6.9）

トピック名	コヒーレンス	集約後との差
6	-3.149	-0.049
9	-2.834	0.267
6.9	-2.882	-

表 9 コヒーレンス計算結果（「y1.1-y1.5」のトピック 3.5）

トピック名	コヒーレンス	集約後との差
3	-2.691	-0.284
5	-2.971	-0.003
3.5	-2.975	-

語共起文書数である。従って、トピックを代表する特徴語が解析文章中で多く共起すれば、一貫性があるとみなされコヒーレンスのスコアが上がる。

本実験では集約するいくつかのトピックにおいて、それぞれのトピックの特徴語はベクトル作成時に用いた語とした。そして、それらのトピックの中で最も特徴語が少ないものの語数を N_{min} として、集約前のトピックと集約後のトピックの生成確率上位 N_{min} 件をコヒーレンスの計算に用いた。

計算の結果、コヒーレンスが改善したと考えられるものは全 21 個の集約のうち 12 個となった。図 8 は適切な集約が行われていると考えられるトピックの例、表 7、表 8、表 9 はコヒーレンス計算の結果の例である。集約が適切であると考えられる「w1.1-w1.2」のトピック 1.8 では、集約後のコヒーレンスが集約前のコヒーレンスに比べてほぼ変わらないか、良くなっていた。また、集約が不適切であると考えられる「y1.1-y1.5」のトピック 3.5 では、集約後のコヒーレンスが集約前のコヒーレンスに比べてほぼ変わらないか、悪くなっていた。しかし、「w1.2-y1.1」のトピック 6.9 の様に、集約が不適切だと考えられるものでも集約後のコヒーレンスが集約前のコヒーレンスよりも大幅に良くなっているものもあった。

そして、21 個の集約のうち妥当性について著者のうちの 1 名が目視で判断した結果、コヒーレンスと定性的な評価が合致したものは 14 個であった。従ってコヒーレンスで

の集約の評価は定性的な評価に従い、ある程度定量的に評価できると考える。

5. おわりに

本研究では、一つの長期的な報道の中で少しずつ変化するトピックをとらえるためにトピックモデルを用いた解析を行った。

第 2 章では先行手法である DTM を用いて解析を行い、先行手法の問題点を指摘した。第 3 章では LDA を用いて時系列データを解析する方法として、トピックの特徴ベクトルを生成し類似性を確認することで、同期間トピックの集約や次期間トピックへの推移を表現することを提案した。第 4 章では実験結果を定量的および、定性的な評価を行った。その結果、提案手法を用いてトピックの推移、途中発生、途中消滅を表現でき、全 21 個の集約のうち 12 個のコヒーレンスを改善させることができた。

今後の課題としては、トピックの特徴ベクトルの改善などがあげられる。ベクトル作成時に、局所的に使われている語に注視し、広域的に使われている語の影響を少なくする idf のような処理が必要だと考えられる。

謝辞 本研究に取り組むにあたって、データをご提供いただいた神戸新聞社、特にメディアビジネス局企画推進部長の豊川聡氏、峯大二郎氏に深く感謝致します。また、本研究は JSPS 科研費 JP18H03243, JP18H03494 の助成を受けたものです。加えて謝意を表します。

参考文献

- [1] D.M.Blei, J.D.Lafferty, "Dynamic Topic Models" Proceedings of the 23rd international conference on Machine learning, pp.113-120, 2006
- [2] D.M.Blei, A.Y.Ng, M.I.Jordan, "Latent Dirichlet Allocation" Journal of machine Learning research, pp.993-1022, 2003
- [3] T.Mikolov, I.Sutskever, K.Chen, G.S.Corrado, J.Dean, "Distributed Representations of Words and Phrases and their Compositionality" Advances in Neural Information Processing Systems 26, pp.1-9, 2013
- [4] D.Newman, J.H.Lau, K.Grieser, T.Baldwin, "Automatic Evaluation of Topic Coherence", Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp.100-108, 2010
- [5] D.Mimno, H.M.Wallach, E.Talley, M.Leenders, A.McCallum, "Optimizing Semantic Coherence in Topic Models", Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp.262-272, 2011
- [6] N.Aletras, M.Stevenson, "Evaluating Topic Coherence Using Distributional Semantics", Proceedings of the 10th International Conference on Computational Semantics, pp.13-22, 2013
- [7] M.Röder, A.Both, A.Hinneburg, "Exploring the Space of Topic Coherence Measures", WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp.399-408, 2015