トピックモデルによるアンケート自由記述回答の潜在意味解析

Latent Semantic Analysis of Free Description Answers Utilizing Topic Model

西田　豊（Yutaka Nishida）*1

要約　アンケート調査において，選択式の設問とともに，自由記述式の設問が設定されることが多い．自由記述回答には選択式の設問では捉えきれなかった意見が含まれていることも多く，非常に情報量が多いデータといえる．しかしながら，自由記述回答は質的なデータであり統計的に扱いにくかったため，分析において量的なアプローチがとられることは多くなかった．本研究では自然言語処理分野において使用されるトピックモデルを用いて，自由記述データに潜在するトピックを抽出した．また抽出したトピックの中には選択式の設問で得られた総合得点と相関するものが見いだされた．これらの結果は安全文化醸成の程度と自由記述の内容が関連することを示していると考えられる．

キーワード　安全風土，安全文化，自由記述，潜在意味解析

Abstract　In questionnaire surveys, free-form questions are often set along with multiple-choice questions. Free-description answers often include opinions that cannot be captured by the multiple-choice questions, and are considered to be very informative data. However, the free-description answers are qualitative data and they are difficult to handle statistically, so a quantitative approach is not often taken in their analysis. In this study, we extracted the latent topics in the free description data by using the topic model used in the field of natural language processing. We found some topics correlated with the total score obtained in the multiple choice questions. We consider these results to be related to the degree of fostering safety culture and the content of free description.

Keywords　safety climate, safety culture, free description answer, Latent Semantic Analysis

1. はじめに

職場における業務を安全に遂行するためには，技術的要因のみならず人的要因，組織的要因の重要であることが指摘されている．安全を向上させるためには，これらに加え文化の観点から安全を考えることが必要である（原子力安全システム研究所，2019）．各組織においては状況把握のため，安全文化醸成の程度の評価が試みられている．安全文化の評価方法にはいくつかのアプローチがあるが(竹内，2012)，アンケート調査での評価が用いられることも多い（西田，2017）．安全風土という概念により職場の安全性について評価が行われてきた（Zohar，1980）．

原子力産業分野での需要の高まりを受け，アンケート調査を用いた研究も増えてきている．例えば，近年では回答者の属性に関する回答傾向の違いや（藤田，2017; 福井，2012; 西田，2018），属性による影響の程度が事業所間で異なること（藤田，2018）が報告されている．また，調査項目の改良を目的とした研究も実施されている（西田，2019）．

このようなアンケート調査が実施される場合，自由記述回答の設問が設定されることも多く，回答者は選択式の回答方法では表現できなかった「思い」が書き込まれることがある．したがって，アンケート実施者が想定していなかった回答が含まれていることも多く，非常に情報量の多いデータであるといえる．

テキストデータの扱いにくさによると考えらえるが，自由記述に書かれている内容がどのような話題

＊1　(株)原子力安全システム研究所　社会システム研究所

であるのかや，自由記述の内容が回答者の属性によってどのように異なるのかといった自由記述内容についての把握や，自由記述の内容とアンケートで実施している量的調査部分とがどのような関連にあるのかといった分析についてはあまり実施されてこなかったといえる．

近年では計量的なテキストデータ解析の方法とツールが整備されつつあり，テキストデータ解析を実行しやすい環境ができつつある（樋口，2014; 石田，2017）．組織研究におけるテキストデータを扱った計量的研究として，工藤（2015）は組織内の縦コミュニケーションについて自由記述と評定値から考察をおこなっている．また, 藤田（2019）では，職場に対する評定値と文字数や使用される語に関連があることが指摘されている．

本研究ではアンケートにおける自由記述から得られたテキストデータを用いて，自由記述の内容を把握するとともに，同時に得られた量的な評定値との関連を検討する．

2. テキストデータ解析

2.1 質から量へ

一般的にテキストデータ解析においては，まず得られたテキストデータを文法や単語の品詞等の情報をもとに形態素と呼ばれる意味を持つ最小単位にまで分割する．そしてどの文書にどの形態素が何回含まれていたかを示す集計表を作成する．この集計表を文書と語彙からなる共起データ行列と呼び，質的なテキストデータから量的な頻度データへと変換したことにより統計的なアプローチを可能にする．

2.2 テキストデータに潜在するトピック

テキストデータ解析の目的の一つに，ある文書の内容がどのような話題について記述されているものであるかを半自動的に判定するというものがある．文書中に含まれる話題を指してトピックと呼ぶことにする．例えば，「リズム，演奏，歌」などの単語が文書中に含まれていたとすると，その文書は音楽について記述されていることと理解できる. つまり，文書のトピックはその文書に含まれる単語によって推定可能である．文書中にどのようなトピックが含まれているかの推定にはトピックモデルと呼ばれる方法が用いられてきた（岩田，2015；佐藤，2015）．

自然言語処理研究におけるトピックの推定は当初，情報検索を目的としており，文書・語彙行列を特異値分解により低ランク近似し，潜在的共起性を抽出した．この方法は自然言語処理の研究分野ではLatent Semantic Analysis（LSA; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990）と呼ばれるが，主成分分析による次元削減と等しい．LSAは得られたパラメータに負の値を含むため解釈の困難さが残る点や，モデルの制約の強さから扱いにくいという問題があった．これらの問題を解決したのがLSAを確率モデルとして再定式化したprobabilistic LSA（pLSA; Hofmann, 1999）である．pLSAを階層ベイズモデルとしてさらに拡張したLatent Dirichlet Allocation（LDA; Blei, Ng, & Jordan, 2003）が現在におけるトピックモデルの標準的な方法となっている．

2.3 LDA

LDAは文書中の単語がどのトピックによって生成されたかをモデル化したものである．このトピックは，観測されていない潜在的な変数である．文書数をM，文書dに含まれる単語数をn_dとし，文書dのi番目の単語を$w_{d,i}$，対応する潜在トピックを$z_{d,i}$とする．文書はK個のトピックから構成され，その比率は離散分布として表現される．文書dでトピックkが出現する確率を$\theta_{d,k}$とし，$\boldsymbol{\theta}_d$をトピックの出現分布とする．トピックkにおける単語vの出現確率を$\varphi_{k,v}$とし，$\boldsymbol{\varphi}_k$を単語の出現分布とする．$\boldsymbol{\theta}_d$と$\boldsymbol{\varphi}_k$は以下のようにDirichlet分布により生成されると仮定する．

$$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\boldsymbol{\alpha}) \ (d=1, \cdots, M),$$
$$\boldsymbol{\varphi}_k \sim \mathrm{Dir}(\boldsymbol{\beta}) \ (k=1, \cdots, K).$$

ここで$\boldsymbol{\alpha}$および$\boldsymbol{\beta}$はDirichlet分布のパラメータである．潜在トピック$z_{d,i}$と単語$w_{d,i}$は以下のように多項分布からの生成を仮定する．

$$z_{d,i} \sim \mathrm{Multi}(\boldsymbol{\theta}_d) \ (i = 1, \cdots, n_d),$$
$$w_{d,i} \sim \mathrm{Multi}(\boldsymbol{\varphi}_{z_{d,i}}) \ (i = 1, \cdots, n_d).$$

LDAにおける確率変数間の依存関係は図１のように表現される．

本研究では，LDAを用いて，自由記述データに内在する潜在的なトピックを抽出することを試みる．また，アンケートでは５段階評定による定量的なデータも取得しており，自由記述におけるトピックが定量的データとどのような関係性にあるかも検討する．

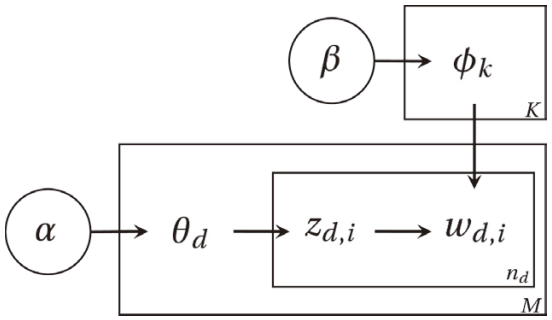

図１　LDAのグラフィカルモデル

3. 方法

3.1 自由記述データ

2018年５月から９月にかけて原子力産業従事者に対して実施された調査票調査で得られた24,503名のデータのうち電力会社に所属し，自由記述設問に回答していた2,955名のデータを解析対象とした．自由記述の設問は「安全文化について，日頃感じていることがありましたら，500文字以内で自由にご記入ください」というものであった．

3.2 安全文化総合得点

量的データについては全78項目の設問があり，「そう思う」～「そう思わない」の５段階で評定をしてもらい，５点から１点を割り当てた．逆転項目は安全に関する状態が良いほど得点が高くなるよう反転し処理を行った．全78項目の全平均を算出し，総合得点とした．総合得点の記述統計量を表１に示す．

表１　総合得点の記述統計量

最小値	中央値	最大値	平均値	標準偏差
1.24	3.92	4.99	3.89	0.63

3.3 自由記述データの分類

総合得点と自由記述との関連を検討するため，総合得点をもとに自由記述を10グループに分類する．各グループに含まれる回答者数がおよそ等しくなるよう，総合得点の値の順に自由記述を10グループに分類した．総合得点の値が元も低いグループをG0とし，値が高くなるにつれてグループ番号が大きくなり，総合得点が最も高いグループをG9とした．LDAにおける解析ではこれらのグループを１つの文書として扱った．

3.4 分析方法

自由記述データから解析用データ行列の作成にあたり，日本語形態素解析エンジンMeCab Ver. 0.996（工藤，2013）およびR言語のパッケージRMeCab ver.1.04（Ishida, 2019）を使用した．品詞は名詞のみを指定し，最小出現数は５としたところ639語が抽出された．自由記述全体で出現した頻度の高い上位10語を表２に示す．

LDAによる解析においてはトピック数を分析者が決定する必要があるが，今回の分析ではトピック数を増加させながら推定したときの尤度の増加程度を参考に解釈可能なトピック数を検討したところ６トピックが適当であると判断した．実行にはR言語のパッケージlda ver.1.4.2（Chang, 2015）を使用した．

表２　頻出語上位10語

	単語	頻度
1	安全	3993
2	文化	1854
3	業務	1068
4	必要	659
5	原子力	594
6	職場	417
7	重要	404
8	現場	295
9	会社	280
10	人	278

4. 結果と考察

4.1 LDAによるトピック推定

各トピックに含まれる頻出語上位10語は表３のようになった．またトピックと語の関連性の指標として，ある語が当該トピックに出現した回数を当該トピックに含まれる全語の出現回数で除した出現確率を算出した．

表３　各トピックに含まれる頻出語上位10語とその出現確率

	topic 1		topic 2	
1	業務	.29	原子力	.17
2	社員	.05	人	.08
3	ルール	.05	状況	.06
4	工程	.04	コスト	.05
5	状態	.03	部門	.04
6	部署	.02	環境	.04
7	レベル	.02	現状	.03
8	予算	.02	考え方	.03
9	上司	.02	考え	.03
10	余裕	.02	アンケート	.02

	topic 3		topic 4	
1	必要	.21	会社	.13
2	現場	.09	コミュニケーション	.05
3	大切	.05	企業	.04
4	姿勢	.04	言葉	.04
5	技術	.03	人員	.04
6	当社	.03	日頃	.04
7	不安	.03	効率	.03
8	グループ	.03	目標	.03
9	機会	.02	傾向	.03
10	事業	.02	幹部	.03

	topic 5		topic 6	
1	安全	.47	重要	.11
2	文化	.22	事故	.06
3	職場	.05	自分	.05
4	個人	.02	リスク	.05
5	取り組み	.02	プラント	.03
6	情報	.02	雰囲気	.03
7	十分	.01	大事	.02
8	疑問	.01	内容	.02
9	危険	.01	事例	.02
10	課題	.01	基本	.02

トピック１には「業務」が最も多く出現し,「ルール」,「工程」といった語も多く出現するトピックになっていることから，業務についてのルールや工程についてのトピックと解釈できる．

トピック２は「原子力」が最も多く出現しており,「状況」,「コスト」といった語も多く出現していることから，原子力を取り巻く現状についてのトピックであると解釈できる．

トピック３には「必要」が最も多く出現しており「現場」,「姿勢」という語も多く出現していることから現場における安全に対する姿勢についてのトピックであると解釈できる．

トピック４は「会社」が最も多く出現し,「コミュニケーション」や「言葉」といった語多く出現していることから社内におけるコミュニケーションについてのトピックであると解釈できる．

トピック５は「安全」が最も多く出現しており,「文化」,「取り組み」といった語が多く出現していることから安全文化への取り組みについてのトピックであると解釈できる．

トピック６は「重要」が最も多く出現しており,「事故」,「リスク」といった語も多く出現していることからプラントにおける事故やリスクについてのトピックであると解釈できる．

4.2 トピック比率

各得点グループにおいて各トピックがどのような比率で記述されていたかを推定したものが図２である．得点グループによって若干のトピック比率の高低があるが，おおむねの傾向は一致していると考えられる．すべての得点グループにおいて比率が高

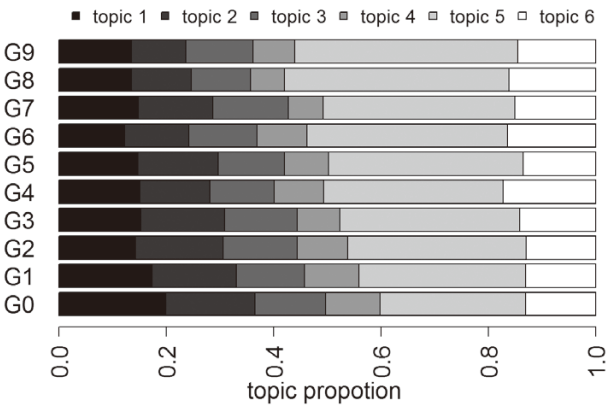

図２　各得点グループのトピック比率

かったのはトピック５の安全文化への取り組みについてのトピックであった．最も記述の比率が少なかったのはトピック４のコミュニケーションについてのトピックであった．トピック３は得点グループ間で大きな差はなかった．

4.3　総合得点とトピック比率との関連

図３にトピック別に得点グループごとのトピック比率を示す．総合得点が上昇すると増加するトピックと，減少するトピックがあることがわかる．G0～G9の順位とトピック比率との相関係数を計算したところ表４のようになった．

いずれのトピックも総合得点と比較的強い相関があることが確認された．特にトピック５は総合得点と非常に強い正の相関関係にある（r = .95）．これは総合得点が高いグループほど安全文化への取り組みに関するトピックが多いことを示している．またトピック２とは負の相関関係にあり（r = −.89），総合得点が高いほど原子力の現状に関するトピックが少なくなるといえる．

表４　各トピックにおける総合得点のグループ順位とトピック比率との相関係数

topic	相関係数
1	−.77
2	−.89
3	−.37
4	−.79
5	.95
6	.58

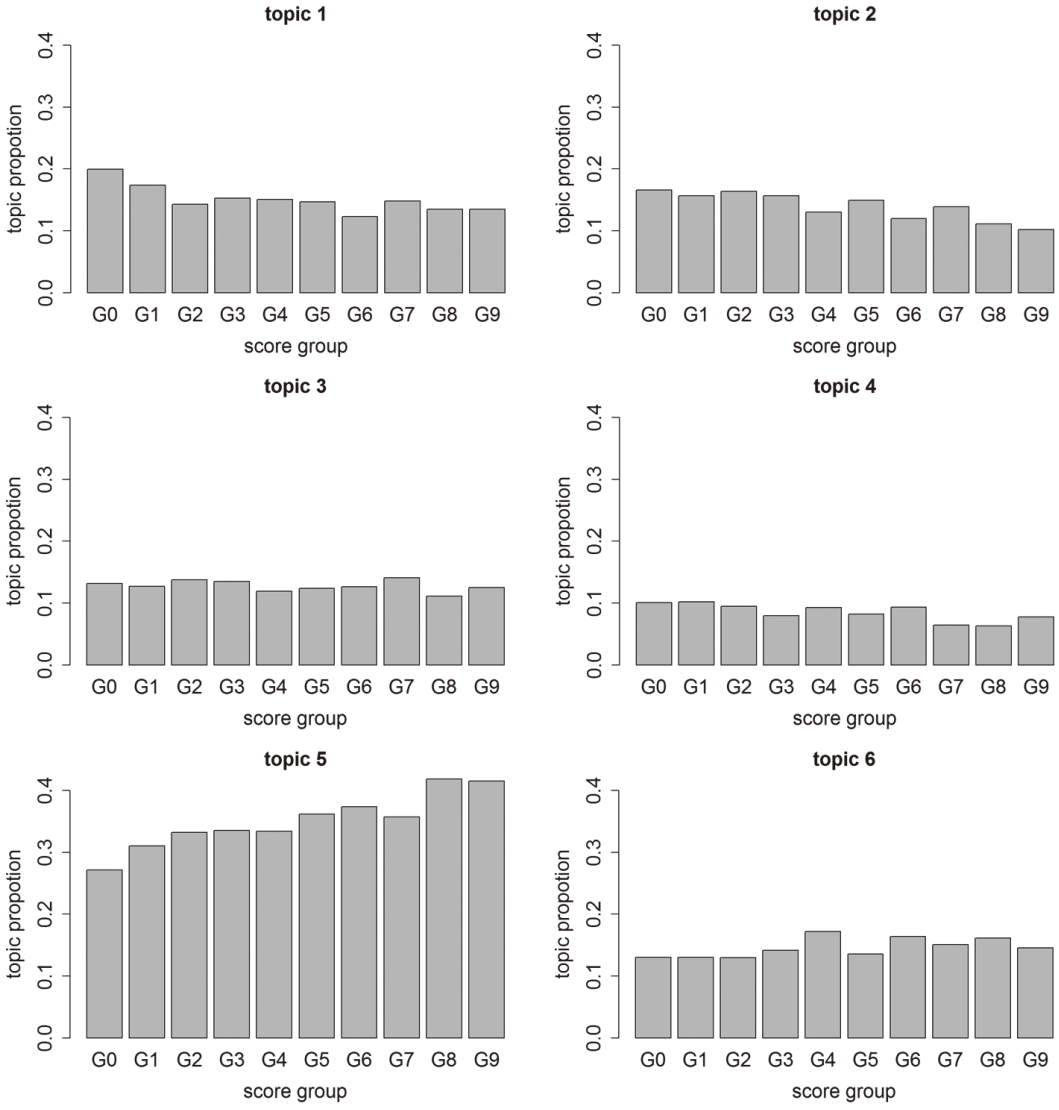

図３　トピック別の得点グループごとのトピック比率

5. まとめ

本研究では，LDAによりアンケートで得られた自由記述データにおいてどのようなトピックが記述されていたのかを推定した．６つのトピックはいずれも原子力安全に関連して解釈が可能なものであった．得点グループにかかわらず，最も大きな比率でトピックになっていたのは，安全文化についてであった．また，安全文化についてのトピックは安全文化総合得点が高いグループほど，記述が増えるということが示唆された．そのほか５つのトピックについても，安全文化総合得点との高い相関関係が見いだされた．これは，アンケートの得点は所属する組織の安全文化状態を反映した数値と考えられるため，安全文化の醸成の程度と自由記述の内容が関連することを意味している．藤田 (2019) においても，組織に対する評価と自由記述内容が関連することが指摘されたが，本研究では異なるアプローチにより確認されたといえる．

原子力産業にかかわるデータは社会的・経済的動向の影響を受け時系列的に変化することが予想される．例えば，藤田（2019）や河合（2020）では，原子力産業従事者の意識が東日本大震災の後に変化していることが指摘されている．本研究では一時点の自由記述データを使用したが，自由記述においても時系列変化が予想される．時系列自由記述データに対してDynamic Topic Models（DTM: Blei & Lafferty, 2006）などの時系列トピックモデルを用いることで，社会的・経済的動向を踏まえたトピックの変化を検討することが可能になると考えられる．

謝辞

本研究は一般社団法人原子力安全推進協会（JANSI）と各事業所のご協力のもとに実施できたものである．厚く感謝申し上げる．

引用文献

Blei, D. M. and Lafferty, J. D.（2006）. Dynamic topic models. Proceedings of the 23rd International Conference on Machine Learning, 113-120.

Blei, D. M., Ng, A. and Jordan, M.（2003）. Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.

Chang, J.（2015）. lda: Collapsed Gibbs Sampling Methods for Topic Models. R package version 1.4.2.

Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K., and Harshman, R.（1990）. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41（6）, 391-407.

藤田 智博（2017）. 安全確認を抑制するメカニズム －知識・技能への自信に注目して－ INSS JOURNAL, 24, 48-57.

藤田 智博（2018）. 原子力産業の安全風土調査へのマルチレベル分析の適用 INSS JOURNAL, 25, 17-24.

藤田 智博（2019）. テキストデータが映し出す「安全」－自由記述の活用－ INSS JOURNAL, 26, 2-9.

藤田 智博（2019）. 2010年代の原子力産業従事者の仕事への意識 －世代差に着目して－ INSS JOURNAL, 26, 10-17.

福井 宏和（2012）. 原子力発電所の安全風土に関する質問紙調査 集団力学, 29, 69-86.

原子力安全システム研究所 社会システム研究所（編著）（2019）. 安全文化を作る 日本電気協会新聞部.

樋口 耕一（2014）. 社会調査のための計量テキスト分析 －内容分析の継承と発展を目指して－ ナカニシヤ出版.

Hofmann, T.（1999）. Probabilistic Latent Semantic Analysis. Uncertainty in Artificial Intelligence. 289-296.

石田 基広（2017）. Rによるテキストマイニング入門（第２版）森北出版株式会社.

Ishida, M.（2019）. RMeCab: interface to MeCab. R package version 1.04.

岩田 具治（2015）. トピックモデル 講談社.

河合 学（2020）. 原子力産業に従事する組織成員意識の変化に関する探索的検討 INSS JOURNAL, 27.

工藤 直志（2015）. 自由回答を用いた組織内コミュニケーションの分析 INSS JOURNAL, 22, 2-12.

工藤 拓（2013）. MeCab ver. 0.996

(https://taku910.github.io/mecab/).
西田 豊（2017）．安全風土と安全文化 －概念，測定と理論，醸成について－ INSS JOURNAL, 24, 21–31.
西田 豊（2018）．スパース判別分析による属性別安全風土の特徴抽出 INSS JOURNAL, 25, 25–30.
西田 豊（2019）．項目反応理論による安全風土調査の項目分析 INSS JOURNAL, 26, 18–26.
佐藤 一誠（2015）．トピックモデルによる統計的潜在意味解析 コロナ社.
竹内 みちる（2012）．組織の安全文化（安全風土）評価・測定の手法に関する試論 INSS JOURNAL, 21, 10–19.
Zohar, D.（1980）．Safety climate in industrial organizations: Theoretical and applied implications. Journal of Applied Psychology, 65, 96–102.