様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 ２３ 日現在

機関番号：１７１０２
研究種目：基盤研究(B)（一般）
研究期間：2018～2021
課題番号：１８Ｈ０３４９５

研究課題名（和文）テキストからわかる価値観を対象にした内容分析とその半自動化手法に関する総合的研究

研究課題名（英文）Comprehensive research on content analysis and semi-automated content analysis for human values in texts

研究代表者

石田　栄美（Ishita, Emi）

九州大学・データ駆動イノベーション推進本部・教授

研究者番号：５０３６４８１５

交付決定額（研究期間全体）：（直接経費）　　15,640,000円

研究成果の概要（和文）：原子力発電所事故関連の新聞記事の社説に表れる人の価値観の内容分析を対象に、これを半自動化するため、内容分析を3ステージに分け、それぞれに分類器を用いるアプローチを提案した。内容分析に必要な分類器を学習するためには人手によるコーディング結果を用いるが、この質を効率的に向上させる手法と、分類器の学習に効果的な学習用データの選択手法を開発した。最後に、新たな価値観カテゴリの発見を支援するための方法を検討した。

研究成果の学術的意義や社会的意義
質的な研究手法である内容分析に対して、内容分析の質を極力落とさずに、できるだけ自動的な手法を取り入れ、人の労力を減らすためのアプローチ、手法の開発・提案に取り組んだ。主に、内容分析を3ステージに分け、それぞれに分類器を用いる半自動化の方法の提案、分類器に必要な質の高いデータを効率的に構築する手法、分類器が効率的に学習するための学習用データの選択手法等を開発することで、内容分析の効率化に貢献したといえる。

研究成果の概要（英文）：To develop a semi-automatic content analysis of human values appearing in newspaper editorials related to the nuclear power plant disaster, we divided the content analysis into three stages and proposed to use a classifier for each stage. Manual coding results were used to train the classifiers for the content analysis. We developed a method for efficiently improving the quality of the manual coding results. We also developed a method for selecting training data for classifiers efficiently. Finally, we studied methods to find new human value categories.

研究分野：図書館情報学

キーワード：内容分析　自動分類　人の価値観　原発事故　コーディングの質向上　バンデットアルゴリズム　効果的な学習用集合の分析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

社会的に大きく複雑な問題の場合、それに対して賛成や反対といった様々な立場が存在し、その背後にはその立場のもととなる多様な価値観が存在する。テキストから人が持つ価値観を分析することによって、世論や一般の人々がある事柄に対して持つ価値観を把握することができ、政府や行政にとっては問題解決のための政策や意思決定の際の助けとなる。 本研究課題では、東日本大震災による福島第一原子力発電所の事故に起因する原発政策や エネルギー政策に関する新聞記事に着目した。この事故以降、日本では原子力発電所の存続に支持と反対の立場で活発な議論が行われている。また、それだけでなく、台湾、ドイツ、イタリアをはじめとする国々が自国の原子力発電やエネルギー政策に関する議論をしている。原発政策のような賛成、反対等様々な立場の意見があり、かつ、単純に方針や政策を判断できない複雑な事柄に対して意見の根底にある価値観の分析をすることは大きな意義があるといえる。また、一般の新聞記事は世論が重要だと考える価値観が間接的に、社説では新聞社の持つ価値観が直接的に反映されている。そのため、新聞記事を分析することで、世論の価値観、マスメディアの価値観を知ることができる。

原発に関する議論は、福島第一原子力発電所の廃炉を考えても数十年に渡って議論されると考えられる。このような議論を対象に分析する場合、長いスパンでのデータ収集や大量テキストの分析を行う必要があるが、すべて手作業で行うには限界がある。内容分析には、コーディングフレームの構築、フレームに基づくテキストへのコーディングなどのステップを含むが、工学的観点からの興味としてはフレームの自動構築やコーディングを 100%の精度で目指すことになる。しかし、社会科学者は機械的な手法を用いて示された結果をすぐに信じることはできない。そこで、内容分析の完全自動化は期待しないが、内容分析の質を担保しながら、研究が効率的にできるような機械学習等を用いた支援方法を検討することが必要である。それらを開発することにより、内容分析における人の労力を軽減することが可能となる。

２．研究の目的

原発政策やエネルギー政策に関する新聞記事に人の価値観がどのように表れているかという内容分析を対象にし、その過程に機械的な手法を取り入れることにより、手作業で行う内容分析の質を担保したまま、内容分析がより効率的に行える手法を開発することが目的である。これには様々なアプローチの提案や手法の開発が考えられるが、本研究では、多面的に検討、開発を行った。主に、以下を具体的な研究目的として設定した。(1)3 ステージから構成される内容分析の半自動化手法の提案、(2)構築コストが異なる学習用データが分類器の性能に与える影響の分析、(3)学習用データのためのコーディングの質向上のための手法の開発、(4)分類器のための効率的な学習用データ選択手法の開発、(5) 新たな価値観カテゴリ発見支援のための分析方法の検討などを行った。同時に、人の価値観カテゴリ以外の内容分析の可能性も探るため、(6)原発関連の新聞記事に対して人の価値観カテゴリ以外の内容分析も実施した。

３．研究の方法

以下では、上記の研究目的の各項目に関する研究方法を説明する。

(1) 3 ステージから構成される内容分析の半自動化手法の提案

原子力発電等の議論を扱う新聞社社説に表れる人の価値観に関する内容分析を対象に、3 ステージから成る内容分析手法のアプローチを提案した。提案した 3 ステージの各ステージは、①社説が分析対象であるかどうかを判定し（on/off topic identification）、次に②分析対象記事の中の各文に価値観が含まれる文(価値観文)か事実文であるかどうかの判定を行い、最後に③価値観文に対して具体的な価値観カテゴリを付与する、であり、それぞれのステージにおいて分類器を導入することにより、半自動化を試みる。実際の新聞社社説を用いて、それぞれのステージで、どの程度の性能が得られるかを調べ、実現可能性を検証した。

(2) 構築コストが異なる学習用データが分類器の性能に与える影響の分析

コストと質という 2 つの点で構築コストが異なる学習用データが、分類器の性能にどのような影響を与えるかを分析した。半自動化手法の第 3 ステージである文に人の価値観カテゴリを付与するタスクに対して、A.それぞれのコーダーが判定した結果を混ぜた学習用データ(hybrid)と、B.2 名のコーダーの合議結果である adjudicated の学習用データをそれぞれ構築した。B の判定コストは 2 名のコーダーの合議がはいるため、A の構築コストの 2 倍であるが、判定の質は A より高い。これらの構築コストの異なる学習用データを用いて、実際に分類器を用いて価値観カテゴリの付与を行うことで、分類器の性能から構築コストが異なる学習用データの影響を分析した。

(3) 学習用データのためのコーディングの質向上のための手法の開発

分類器を用いたコーディングの質を上げるための方法を提案した。人によるコーディングは一貫性が保証されないことが指摘されている。一度、コーディングしたものを、再度、全て見直

せば質は改善すると考えられるが、作業が 2 倍になってしまい効率的ではない。この問題を解消するために、分類器を用いて見直すべきコーディング結果を特定する方法を提案した。コーディングタスクは、社説が分析対象であるかどうかの 2 値の判定である(on/off トピック判定)。最初に、コーダーが対象社説を全てコーディングし(第 1 ラウンド)、そのコーディング結果を用いて分類器(SVM)を学習した。次に、学習した分類器が同じ対象社説を判定し、分類器の結果と第 1 ラウンドの結果を比べた。この中で、分類器と第 1 ラウンドの結果が異なるものを選択し、改めて人がコーディングし直した(第 2 ラウンド)。これらの結果と、経験を積んだコーダーとの結果を比べることにより、見直しをするコーディング結果の選定を分類器によって行うことが有効かを検証した。

(4) 分類器のための効率的な学習用データ選択手法の開発

分類器のための学習用データ作成に関して、強化学習の観点から限られた予算の中で逐次的になるべく高性能な分類器を構築するためのデータセットのコーディング戦略を提案した。2 名のコーダーを雇うという条件において、1)コーダーA が未コーディングのテキストにコーディング、2)コーダーB が未コーディングのテキストにコーディング、3)コーダーB がコーディングしたテキストにコーダーA がコーディングを行い、その後コーダー間で合議、4)は 3)の逆のという 4 種類のコーディング方法を設定した。各方法を「アーム」とみなし、ε-greedy アルゴリズムに基づくバンディットアルゴリズムにより、次のコーディングにおいて分類性能が向上する可能性が高いアームを選択する手法を提案し、実験により有効性を検証した。

(5) 新たな価値観カテゴリ発見支援のための分析方法の検討

価値観は時代の流れとともに変化することを想定し、新たな価値観ラベルの発見を支援するための方法を検討した。以下の 2 種の分析を行った。(5.1)先行研究で構築したデータセットを用いて分類器を構築し、価値観カテゴリが付与されていない文集合に、Bertopic を用いてトピック分析を行った。先行研究でラベル付け対象とした年以降の社説集合に対して上記のアプローチを適用し、獲得した分析結果が、新たな価値観カテゴリの発見に有効かどうか検討した。さらに、(5.2)トピック分析技術を用いて、価値観カテゴリに内在するトピックを解析した。

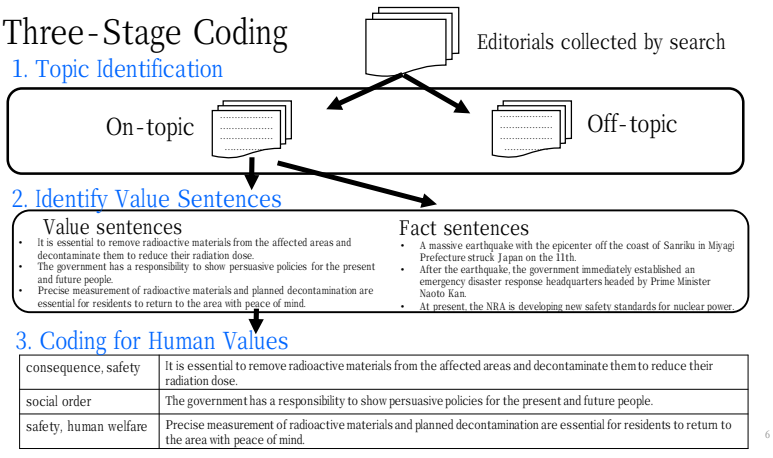
(6) 原発関連の新聞記事に対して人の価値観カテゴリ以外の内容分析

毎日新聞と読売新聞の記事に対して、人手により内容分析をおこなった。2011 年～2013 年までの原発関連記事に対して、イベント種別、責任主体、トピックごとに内容の分類、時系列ごとの推移、責任についての言説の変遷等の分析を行った。

４．研究成果

(1) 3 ステージから構成される内容分析の半自動化手法の提案

図に示すような 3 ステージによる内容分析の半自動化手法を提案した。①分析対象記事の判定は、原発や原子力の語が含まれている社説を検索し収集した上で原子力発電やエネルギー政策が中心的な話題である社説を判定する。検索で収集した社説に対し、2 名のコーダーがそれぞれコーディングを行い、その結果を持ち寄り判定結果が異なる場合は合議し最終的な判定結果(adjudicated)を決めることで 2 種の分類器のための学習用データを構築した(448 社説)。SVM と深層学習を用いた fastText の 2 分類器を用いて、A.それぞれのコーダーが判定した結果を混ぜた学習用データと B.adjudicated の学習用データ(判定コストは A の 2 倍、判定の質は A より良い)を用いた場合の分類器の判定性能を比べたところ、判定コストに制限がある場合(小さい場合)には A を学習用データに用いたほうが性能が高くなることがわかった。また、価値観文の判定については、量が少ない学習用データを用いても、SVM においてある程度の判定性能を得られることがわかった。



(2) 構築コストが異なる学習用データが分類器の性能に与える影響の分析

2 名のコーダーにより、120 社説の各文に対して人の価値観カテゴリのコーディングを行い、その後、合議を行った。これにより、A.それぞれのコーダーが判定した結果を混ぜた学習用データ(hybrid)と B.2 名のコーダーの合議結果である adjudicated の学習用データを構築した。SVM 分類器を用いて判定性能を比べたところ、常に A または B のほうが性能が高い価値観カテゴリ

と第 1 ステージと同様に途中で性能が逆転するカテゴリがあった。これは第 1 ステージの分類器の性能とは異なる傾向であった。いくつかの分析を行ったところ、トレーニングデータの量だけでなく、コーディング結果の一致率、正例と負例の数のバランス、コーダーが感じた各タスクのコーディングの難しさ等が分類器の性能に影響を与える可能性があることがわかった。

(3) 学習用データのためのコーディングの質向上のための手法の開発

原発に関連する 780 社説を対象に、3 人のコーダーが第 1 ステージの on/off トピック判定を行った。3 人のコーダーの平均では、第 2 ラウンドでは、第 1 ラウンドの結果の 11%が再検討され、そのうちの 46%が第 1 ラウンドの結果から変更された。また、その変更された結果を含めた結果を、経験を積んだコーダーの結果と比べたところ、71%が一致していた。最終的な結果として、コーディングコストを平均 11%だけ増加させるだけで、経験を積んだコーダーとの Kappa による一致率が 0.69 から 0.73 に向上した。この提案手法を用いることで、少ないコストで質の高い学習用データを構築できることがわかった。

(4) 分類器のための効率的な学習用データ選択手法の開発

実験では、既定の 6 種類の各価値観を付与したデータセット(120 社説)に対するデータセット作成タスクを設定し、各価値観カテゴリにおける各コーディング戦略により構築されたデータセットを用いて、マルチアームドバンディットを用いたラーニングカーブを示した。いずれの価値観カテゴリにおいても、概ね逐次的に品質の高いデータセットを作成できることを示した。

(5) 新たな価値観カテゴリ発見支援のための分析方法の検討

(5.1) 価値観ラベルが付与された 1,659 件の文集合に対して、Bertopic によるトピック分類をしたところ、30 種類のトピックが得られ、トピック分布からこれらを 5 種類のカテゴリに統合し、価値観ラベルに基づいて各カテゴリを調べたところ、全体的には、必ずしも価値観カテゴリに基づいてトピックが生成されるとは限らないとわかった。分類器により、いずれの価値観ラベルも付与されなかった 203 文に対する分析結果を調べたところ、生成されたトピックは 4 種類であり、各トピックは独立している傾向にあると考えられた。いずれの価値観も付与されなかった文に対してトピック分析を実行することで、何らかのグループ化が行われ、また、そのトピックを表す語を確認することができた。新しい価値観ラベルを発見したい場合には、これらの語を参考にできる可能性がある。

(5.2) Bertopic を用いて価値観カテゴリが付与されている文がどのようなトピックを持っているのかを調べた。また、各トピックにおける別の価値観カテゴリとの関係性を明らかにするための分析を行った。たとえば、価値観カテゴリ Human Welfare では、避難者または被災者の生活や現状を取り上げているトピックが主な話題を占めているが、2012 年には電力供給に関するトピックが増加していること、価値観カテゴリ Consequence、Social Order、Wealth などとの関係性も高いことなどがわかった。これらの分析は、中長期間にわたる議論の中において、価値観ラベルを形成するトピックの性質を明らかにし、そのラベルが持つ意味を詳細に分析するための情報として活用できると考えられる。

(6) 原発関連の新聞記事に対して人の価値観カテゴリ以外の内容分析

考察の結果、当初は脱原発を評価し、政府の説明責任のまずさや不手際を批判、それが時間的な推移とともに原発存置派・再稼働派への対抗言説の形成し、政治的求心力の重要性、政治的指導力の重要性を訴え、政府を支持するような言説 が作られていくことが分かった。

(7)総括

本研究では、質的な研究手法である内容分析に対して、主に、内容分析の質を極力落とさずに、できるだけ自動的な手法を取り入れる半自動的化手法の開発・提案に取り組んだ。半自動化手法の提案、効率的にコーディングの質を向上させる手法、分類器が学習する際に効果的に学習ができるための学習用データの選択手法、新たな価値観カテゴリの発見を支援するための分析手法等、多方面から、内容分析に自動的な手法を取り入れることを検討し、それぞれに効果があることがわかった。これらの開発は、内容分析の半自動化手法や、内容分析において人の労力を削減することに貢献できると考えられる。

分類器を用いたカテゴリの自動付与という形で内容分析の自動化に取り組んでいる研究例はあるが、内容分析として社会科学者が実際に利用できるアプローチや手法を総合的に開発している事例はあまり見られない。今後も、実質的な貢献をするための手法の開発を続けていく必要がある。

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　3件／うちオープンアクセス　2件）

１．著者名 Ishita, Emi; Fukuda, Satoshi; Tomiura, Yoichi; Oard, Douglas W.;	４．巻 57
２．論文標題 Using text classification to improve annotation quality by improving annotator consistency	５．発行年 2020年
３．雑誌名 Proceedings of the Association for Information Science and Technology	６．最初と最後の頁 6 pages
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/pra2.301	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Ishita, Emi; Fukuda, Satoshi; Oga, Toru; Tomiura, Yoichi; Oard, Douglas W.; Fleischmann, Kenneth R.	４．巻 2020
２．論文標題 Cost-effective learning for classifying human values	５．発行年 2020年
３．雑誌名 Proceedings of iConference 2020	６．最初と最後の頁 9 pages
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Kodama, Mei; Abe, Kaori; Fukushima, Kana; Hayashi, Eri; Hua, Zhiyi; Jiang, Min; Kang, Ping; Nishida, Emi; Sakai, Shinji; Tomiura, Yoichi; Watanabe, Yukiko; Ishita, Emi;	４．巻 2019
２．論文標題 Content Analysis of Library Use on Microblog: Pre-coding Results.	５．発行年 2019年
３．雑誌名 Proceedings of the 9th Asia-Pacific Conference on Library & Library Education and Practice (A-LIEP 2019)	６．最初と最後の頁 423-428
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Ishita, Emi; Fukuda, Satoshi; Oga, Toru; Oard, Douglas W.; Fleischmann, Kenneth R.; Tomiura, Yoichi; Cheng, An-Shou;	４．巻 11420
２．論文標題 Toward Three-Stage Automation of Annotation for Human Values	５．発行年 2019年
３．雑誌名 Proceedings of 14th iConference 2019 (Lecture Notes in Computer Science)	６．最初と最後の頁 188-199
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-15742-5_18	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　1件）

１．発表者名
大賀哲

２．発表標題
原発事故における「責任」言説の変容―毎日新聞社説を事例として

３．学会等名
政治社会学会研究大会

４．発表年
2019年

１．発表者名
山腰修三、三谷文榮

２．発表標題
福島原発事故をめぐる「危機」と「責任」に関するメディア言説の分析―『朝日新聞』の社説を事例として

３．学会等名
政治社会学会研究大会

４．発表年
2019年

１．発表者名
加藤朋江

２．発表標題
福島原発事故をめぐる「危機」と「責任」をめぐるメディア言説の分析―読売新聞社説を事例として

３．学会等名
政治社会学会研究大会

４．発表年
2019年

１．発表者名
Ishita, Emi; Oga, Toru; Fukuda, Satoshi; Tomiura, Yoichi;

２．発表標題
Developing Semi-Automatic Content Analysis for Studying Human Values in the Nuclear Power Debate

３．学会等名
10th Asia Library and Information Research Group Workshop (ALIRG2018)（国際学会）

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	冨浦　洋一 (Tomiura Yoichi) (10217523)	九州大学・システム情報科学研究院・教授 (17102)	
研究分担者	福田　悟志 (Fukuda Satoshi) (10817555)	中央大学・理工学部・助教 (32641)	
研究分担者	大賀　哲 (Oga Toru) (90445718)	九州大学・法学研究院・准教授 (17102)	

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	山腰　修三 (Yamakoshi Shuzo)		
研究協力者	三谷　文榮 (Mitani Fumie)		
研究協力者	加藤　朋江 (Kato Tomoe)		

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
米国	University of Maryland, College Park	University of Texas, Austin		
台湾	National Sun Yat-sen University			