2ZM-01

記事の意味的構造に着目したこたつ記事の類型化

杉原 昂紀 † 藤川 雄翔 † 藤代 裕之 †† 松下 光範 †
† 関西大学大学院総合情報学研究科 †† 法政大学社会学部

1 はじめに

インターネットの普及とスマートフォンの発展により，ウェブメディアでニュースを閲覧することが一般的になっている．閲覧されるウェブメディア（e.g., Yahooニュース）は誰でも即時的に情報を発信できる一方で，ページビューに基づいた収益構造のため衆目を集めるように誇張された記事が配信され，情報の正確性が担保されないことがある．そのような特徴を持つ記事として，記者が現地調査や直接取材を行わず，インターネット上の情報やテレビ番組などの他媒体をもとに作成される「こたつ記事」[1] が増加しており，記事内容の品質および信頼性の観点から社会問題となっている．こたつ記事は記事内でこたつ記事であるとは明記されず，ウェブメディアで新聞記事と混交する形で配信される点も問題を複雑化させている．ただし，こたつ記事はその記事全体が信用できないわけではなく，記事中に信頼性の高い情報と低い情報が混在しているため，信頼性の低い情報を除去することができれば，情報源として一定の価値を持つ可能性がある．

新聞記事は重要な情報から詳細な情報を書く形式が多い [2] のに対し，こたつ記事はネット上の意見や感想の記述が多く見られる．また，こたつ記事には類型が存在することが知られている†．事実に関する記述についてはどちらの記事も同様の形式を取るという仮説に基づき，こたつ記事を類型化して新聞記事との意味的構造の違いを分析することで，こたつ記事特有であるネット情報由来の箇所を除去し，信頼性の高い情報のみを抽出できるのではないかと考えた．そこで，本研究では，機械学習を用いて記事の意味的構造に基づくこたつ記事の類型化を行い，信頼性の低い部分を検出し，閲覧者の注意を喚起することを目的とする．その端緒として，本稿ではこたつ記事と新聞記事の両記事において各文の意味的構造を推定し，その割合のパタンや表現技法に違いが見られるか分析を行う．

Classification of Churnalistic Articles Based on Semantic Structures
† Koki SUGIHARA
† Taketo FUJIKAWA
†† Hiroyuki FUJISHIRO
† Mitsunori MATSUSHITA
Graduate School of Informatics, Kansai University (†)
Faculty of Social Sciences, Hosei University (††)
† 「もはや、これ、ライターの仕事じゃない」NHK ねほりんぱほりん「こたつ記事」特集に反響，https://www.j-cast.com/2021/01/14402965.html?p=all(2026/1/8 確認)

2 関連研究

狩野らは国内の主要新聞 5 紙（朝日，読売，日経，毎日，産経）の社説を対象に，新聞社ごとの記事の違いを内容と文体の両面から分析した [3]．その結果，新聞社間の違いは文体に表れやすく，話題の違いは内容語に影響を与えることが確認された．

本研究では，こたつ記事の検出のため，こたつ記事と新聞記事間の違いを明らかにすることを目指す．こたつ記事は二次メディアが，新聞記事は一次メディアが作成しているため，メディア間の特徴分析と同じ手法を用いることが可能であると考えたことから，本研究でこたつ記事の類型化を行う際も，構造と文章の内容から分析を行うこととした．

3 実験

3.1 手続き

本研究では，こたつ記事における信頼性の低い情報を特定するため，

(1) 記事構造を把握することを企図した記事の意味的構造の推定

(2) 新聞記事とこたつ記事の識別可能性

(3) こたつ記事内の各文が信頼性の低い情報を含む箇所の同定可能性

について検証する．

(1) では，記事各文に意味ラベルを付与したデータを人手で作成し，それらを教師データとした機械学習を用いて，記事への意味的構造の自動付与を試みる．実験データについて，新聞記事は 2021 年度毎日新聞記事コーパスから「コロナ」「COVID-19」いずれかの語彙を含む本文が 5-8 行の記事を，こたつ記事は 2024 年 1 月に発生した能登半島地震に関連のある記事に対して専門家がこたつ記事か否かのラベルを付与した記事を 310 記事ずつ用いた．各文に付与する意味的構造のラベルは，主観が入っていない文章である「事実」，主観が入っていてかつその主観に対する理由づけが同文章内で明示された文章である「推測」，主観が入っていてかつその主観に対する理由づけが同文章内で明示されていない文章である「感想」の 3 種類に定めた．

(2) では，意味ラベルの出現頻度や順序を特徴量とし，コタツ記事と新聞記事のの識別を試みる．こたつ

記事は特有の文体的・構造的パターンを有すると考えられるため，こたつ記事と新聞記事における意味的構造の出現割合や順序のパタンを比較することで，こたつ記事の識別が可能になると考えた．

(3) では，こたつ記事における信頼性の低い情報を含む可能性のある文の推定を試みる．1 章で述べたように，こたつ記事と新聞記事では情報源が異なり，それが記事中の表現に表出している可能性が考えられるため，両者の表現技法の違いに着目した．

3.2 実験結果

(1) については，クラウドソーシングによってラベルを付与した新聞記事を教師データとして条件付き確率場 (Conditional Random Field; CRF) [4] を用いて，こたつ記事の各文に意味的構造ラベルの推定を行った．CRF に学習させる特徴量は，文頭 2 文字，文頭 3 文字，文末 2 文字，文末 3 文字，最初の助詞，前後の文の意味的構造ラベルの 6 種類とした．その結果，accuracy は 0.830 であり，前述の 6 種類の特徴量を用いることで意味的構造の推定が可能であることが示唆された．

(2) については，ラベルを付与したこたつ記事と新聞記事に対して，記事のラベルの出現頻度を特徴量としたクラスタリングを行った．クラスタリングには X-means を用い，こたつ記事の割合が高いクラスタ（以下，こたつクラスタ）と，こたつ記事の割合が低いクラスタ（以下，新聞クラスタ）からこたつ記事に頻出する意味的構造ラベルの違いを確認した．こたつクラスタの上位 3 クラスタと，新聞クラスタの上位 3 クラスタの意味的構造ラベルの出現頻度を表 1 に示す．χ 二乗検定の結果，有意差は認められたものの（$p = 3.6 \times 10^{-7} < 0.05$），その効果量は小さく（Cramér's V=0.069），両クラスタのラベル割合の違いはわずかであった．また，こたつクラスタと新聞クラスタの意味的構造ラベルの構造をもとに，記事をこたつ記事と新聞記事に分類可能であるかを，SVM を用いて検証した．特徴量には，記事の意味的構造ラベルを scikit-learn の Countvectorizer を用いたものと，隣接する文の意味的構造ラベルの順序の特徴量として n-gram（n=3）を使用した．5 分割交差検証によって精度を評価した結果（表 2 参照），accuracy の平均は 0.491 となり，countvectorizer と n-gram によって作成した特徴量に違いが見られなかった．

(3) については，こたつクラスタと新聞クラスタの表現技法の違いを確認するため，推量を意味する文末表現（e.g.,「だろう」）が利用される頻度を調査した．その結果，こたつ記事では 120 回，新聞記事では 81 回，推量表現が用いられていることが確認できた．加えて，こたつ記事は最初に概要を書き，ネット上の賛否の声を書く「型」があると言われているため，この「ネットでは」,「X では」,「声が」,「反応が」,「SNS では」といったようなネット上の賛否の声を聞く表現が含まれている文章の意味的構造ラベルを確認したところ，事実が 124 件，推測が 84 件，感想が 3 件であった．これにより，ネットの声を聞く文章内に事実や推測が混じっているため，単にこれらの表現を手がかりとし，信頼性の低い情報を含む箇所であると同定することは難しいことが明らかになった．

表 1: 各上位 3 クラスタのこたつクラスタと新聞クラスタの意味的構造ラベル出現頻度とその割合

	事実	推測	感想
こたつ記事	1496(56.54%)	1096(41.42%)	54(2.04%)
新聞記事	2296(62.87%)	1267(34.70%)	88(2.41%)

表 2: SVM の精度検証結果

	precision	recall	F 値	accuracy
こたつ記事	0.514	0.913	0.640	0.491
新聞記事	0.810	0.071	0.129	

4 おわりに

本稿では，意味的構造を用いたこたつ記事の類型化手法を提案した．分析の結果，こたつ記事と新聞記事間の意味的構造の割合パターンには違いは見られなかった．表現技法の違いとして，こたつ記事は断定的表現を避けていることが示唆された．今後は断定表現や内容の情報源を特徴量として用いることによるこたつ記事分類の精度向上を目指す．

謝辞

本研究は JST RISTEX（課題番号 JPMJRS23L2）の支援を受け実施された．記して謝意を表す．

参考文献

[1] 藤代裕之: 「こたつ」記事を定義する, 情報処理学会第 85 回全国大会, Vol. 4, pp. 453–454 (2023).

[2] van Dijk, T. A.: Structures of news in the press, *Discourse and communication*, Walter de Gruyter, pp. 69–93 (1985).

[3] 狩野恵里奈, 荒川唯, 鈴木崇史: 内容と文体による五大全国紙の比較分析, 言語処理学会第 19 回年次大会発表論文集, pp. 334–337 (2013).

[4] Lafferty, J. D., McCallum, A. and Pereira, F. C. N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proc 18th International Conference on Machine Learning*, pp. 282–289 (2001).