社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

PMIを用いた新聞記事におけるカテゴリ・テーマ推定

姫野　拓未† 　嶋田　和孝† 　村重　剛弘††

† 九州工業大学大学院 情報工学府 〒820-8502 福岡県飯塚市川津 680-4
†† 西日本新聞社経営企画局新メディア戦略室 〒810-8721 福岡県福岡市中央区天神 1-4-1
E-mail: †{t_himeno,shimada}@pluto.ai.kyutech,ac.jp, ††takehiro.murashige@nishinippon-np.jp

あらまし　近年，新聞社は多くの人に情報を伝えるために新聞を紙で発行するだけでなく，電子データで発行することが増えている．その中で電子データの新聞記事には関連記事を検索しやすくするためのタグが付与されている．しかし，タグの数は膨大な数であり，新聞記者は手動で記事に対して適切なタグを付与するため多大なコストを費やさなければならない．そこで本研究では，タグの付与の自動化に向け，自己相互情報量 (PMI) を用いた新聞記事のタグ推定を行い，分析結果について報告する．まず，推定のために PMI を新聞記事コーパスより算出し，定量評価と定性評価によってタグの推定結果について考察する．
キーワード　タグ推定，自己相互情報量 (PMI)，新聞記事

Predicting categories and themes in newspaper articles using PMI

Takumi HIMENO†, Kazutaka SHIMADA†, and Takehiro MURASHIGE††

† Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
†† The Nishinippon Shimbun Co. Ltd.
1-4-1 Tenjin, Fukuoka-shi Chuo-ku, Fukuoka, 810-8721, JAPAN
E-mail: †{t_himeno,shimada}@pluto.ai.kyutech,ac.jp, ††takehiro.murashige@nishinippon-np.jp

Abstract　In recent years, newspaper companies publish newspapers by not only paper but also electronic data to spread information. Electronic newspaper on the Web contains various tags easily to find related articles. However, assigning appropriate tags to each article is costly for writers because of many types of tags. In this paper, we propose a score-based model for tag prediction. The score is based on Point-wise Mutual Information (PMI). First, we calculate PMI from the newspaper article corpus. Next, we confirm the results of predicting tags through quantitative evaluation and qualitative evaluation.
Key words　Tagging, Point-wise Mutual Information (PMI), Newspaper Article

1. はじめに

近年，新聞社は多くの人に情報を伝えるために新聞を紙で発行するだけでなく，電子データで発行することが増えている．その中で電子データの新聞記事には関連記事を検索しやすくするためにタグが付与されている．しかし，そのタグの種類は膨大であり，新聞記者は手動で記事に対して適切なタグを付与するため多大なコストを費やさなければならない．

自然言語処理のタスクにおいてタグ推定を行う場合，教師あり学習が一般的な手法として用いられる [1]〜[3]．しかし，新聞記事のタグの種類は膨大な数存在しており，教師あり学習を用いて膨大なタグの種類を推定するタスクは難しい．また，タグは流行によって増減することが考えられ，タグの推定する種類が変わった際に再学習を行うため多くの時間が必要となり，実際の業務には向いていない．

本研究では，タグの付与の自動化に向け自己相互情報量 (Point-wise Mutual Information; PMI) を用いた新聞記事のタグ推定を行い，その分析結果について報告する．ここでのタグには，「福岡」や「社会」などの大まかな情報を表すカテゴリと「新型コロナウイルス」や「福岡速報」などの細かい情報を表すテーマが存在する．PMI を利用すれば，教師あり学習とは異なりタグの推定する種類が変わった際も PMI の更新のみで実現でき，実際の業務に適応すると考えられる．

本研究の貢献は以下の通りである．

- 教師あり学習を用いることなくタスクに適した手法でタグを推定すること
- 膨大な数存在するタグ候補の中から適切なタグがランキングの上位に位置していること

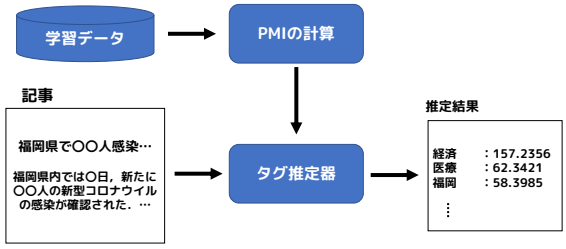

図 1　提案手法の概要

2. 提 案 手 法

提案手法の概要を図 1 に示す．2.1 節では，各カテゴリ・テーマに対応する単語の PMI の算出について説明する．次に，2.2 節では算出した PMI の新聞記事に付与されているカテゴリとテーマそれぞれの推定手法について説明する．

2.1 自己相互情報量

カテゴリ・テーマを推定するために用いる PMI について説明する．まず，学習に用いる新聞記事全体のタイトル，本文を MeCab で形態素解析 [4] を行い，名詞である単語 w のみを抽出する．このとき，数字は名詞ではあるが，カテゴリ・テーマらしい単語として判断しづらいため対象外とする．次に，単語 w が新聞記事に表れる確率を $P(w)$，カテゴリ・テーマ c が新聞記事に付与されている確率を $P(c)$ とする．また，新聞記事中に単語 w が存在し，その新聞記事にカテゴリ・テーマ c が付与されている確率を $P(w,c)$ とする．このとき PMI は以下のように算出される．

$$PMI(w,c) = \log_2 \frac{P(w,c)}{P(w) \cdot P(c)}$$

2.2 カテゴリ・テーマ推定器

2.1 節で説明した PMI を用いたカテゴリ・テーマ推定について説明する．まず，2.1 節と同様にカテゴリ・テーマ推定を行う新聞記事のタイトル，本文を MeCab で形態素解析を行い，名詞である単語のみを抽出する．ここでも同様の理由で数字は対象外とする．次に新聞記事から抽出した名詞を用いて以下の計算式でカテゴリ・テーマらしさの値を算出する．

$$Score(c,d) = \sum_{w \in d} PMI(w,c)$$

ここでの c はタグ候補となるカテゴリ・テーマを表し，d はカテゴリ・テーマ推定を行う新聞記事，w はカテゴリ・テーマ推定を行う新聞記事に含まれる名詞を表している．タグ候補となるカテゴリまたはテーマから $Score$ を算出し，$Score$ の値が大きい方からカテゴリやテーマがその新聞記事を表しているタグであると推定する．

3. 事 例 分 析

提案手法によって得られた推定結果について考察する．まず，3.1 節では本研究に用いたデータについて説明する．次に，3.2 節では提案手法を用いて得られた結果について確認し，3.3 節で定量評価，3.4 節で定性評価の両方から分析結果について考察する．

図 2　西日本新聞記事

3.1 事例データ

事例データは，西日本新聞社の 2019 年 4 月 28 日から 2020 年 4 月 27 日までの 94765 記事を学習データとして PMI の算出に使用した．また，2020 年 7 月に掲載された 3 記事をタグの推定に使用した[1]．西日本新聞の電子データの記事の例を図 2 に示す．図 2 の例では，上部にタイトル，中部に本文と写真で新聞記事が構成されている．さらに，タイトルの上部にカテゴリ，本文の下部にテーマが新聞記者によって手動で付与されている．カテゴリは 73 個，テーマは 396 個存在する中から新聞記者が適切なタグを少なくとも一つは選択している．本研究では，西日本新聞の電子データの新聞記事を対象にカテゴリとテーマをそれぞれ推定した．また，カテゴリ・テーマの推定に使用した 3 記事を表 1〜3 に示す．全ての記事例において，新型コロナウイルスに関連する記事を選択しカテゴリ・テーマを推定した．

3.2 分 析 結 果

各 3 記事の PMI を用いたカテゴリ・テーマの推定結果の上位 10 件を表 4〜9 に示す．各表の太字で表しているものは，実際の記事に付与されたタグを表している．表 1 に示した記事の推定結果では表 4，5 より，記事に付与されたタグが上位 10 件に全て存在することが確認できた．また，表 2 に示した記事の推定結果では表 6，7 より，テーマの「新型コロナウイルス」以外の記事に付与されたタグが上位 10 件に存在することが確認できた．しかし，表 3 に示した記事の推定結果では表 8，9 より，記事に付与されたタグが一つも推定できていないことが確認できた．以上の結果から，タグが上手く推定できる記事とできない記事が存在することが確認できた．

3.3 定量評価による考察

本節では，3.2 節の分析結果から定量評価による考察を行う．表 4〜9 の推定結果の上位 10 件をタグ推定器の示す正解タグと仮定する．カテゴリ・テーマ推定器が示す正解タグの数を A，

(注1)：https://www.nishinippon.co.jp/

表 1　タグ推定に利用した記事例 1

タイトル	新型コロナ，九州 7 県で累計 1000 人に　福岡市で新たに 1 人感染
本文	福岡市は 10 日，新型コロナウイルスの感染者を新たに 1 人確認したと発表した．同市で感染者が判明するのは 3 日以来，7 日ぶり．市によると，感染したのは，同市中央区の 30 代自営業男性．これで九州 7 県の感染者は，長崎県のクルーズ船で確認された感染者や再陽性を除いて累計 1 千人となった．このうち約 8 割を福岡県が占めている．
カテゴリ	福岡，速報，社会
テーマ	新型コロナウイルス，福岡速報

表 2　タグ推定に利用した記事例 2

タイトル	日韓交流「今こそ」奮闘　知られざる観光情報配信やイベント企画
本文	日韓関係の悪化や新型コロナウイルスの世界的な感染拡大に伴い，国境を越えた人の往来が激減する中，韓国釜山市を拠点に日韓交流に取り組む人々が奮闘している．知られざる釜山の観光情報を動画投稿サイトで発信したり，体験型イベントを企画したり…．両国間を覆う重苦しい空気を民間の力で吹き飛ばそうと「自分たちにできる何か」に挑戦し続けている． 釜山拠点の団体「できることを」 1 日午後，釜山市中区の事務所．日韓交流に取り組む社団法人「ぷさんさらん」のメンバー 8 人が釜山，ソウル，大阪，京都を結ぶテレビ電話で，今後の活動について意見を交わした． 会社社長，留学生，韓国語講師など，メンバーの肩書はさまざま．新型コロナウイルスの影響を考慮して，4 月に福岡市で予定していた行事や 6 月の釜山バスツアーの延期を決めたが，8 人は「マイナスをプラスに変え，新しいものを生み出す思考が大事だ」「視聴者に何を提供できるのか考えよう」と前を向く．会議では視聴者とテレビ電話でつないで開く「リモート飲み会」などの案も飛び出した． ぷさんさらんは 2015 年 6 月，釜山市で貿易会社を営む昆雅之代表（46）や韓国語講師の斉藤優さん（29）が中心になって始めた日韓学生交流会をきっかけに，非営利団体としてスタートした．名称は韓国語の「愛（サラン）」にちなみ，国際交流，釜山の観光情報発信，ボランティアを活動の柱に据える． 同年 8 月に動画投稿サイト「ユーチューブ」に「わぼいそ釜山」のチャンネルを開設．ガイド本に載っていないような飲食店や観光名所などを楽しく紹介する動画を配信している．配信数は 270 本を超え，チャンネル登録者数は 6 万人を突破した．メンバーの一人で，筑豊なまりの日本語を操る金賢珍（キムヒョンジン）さん（36）は「動画を通じて地元の釜山を好きになってくれる人が増えるのはうれしい」と行きつけの店に案内する． 視聴者参加型のイベントも人気だ．昨年 6 月の釜山バスツアーには日本人 37 人が参加した．郊外の島や市場を巡り，貸し切りヨットのクルージングを満喫．夕食には韓国人の視聴者も駆けつけ，新たな交流が生まれた．日韓関係の悪化が際立った 8 月には「自分たちがやれることをしよう」と観光スポットの民楽水辺公園でのごみ拾いを呼びかけ，日韓の 25 人が一緒に汗を流した． 活動の功績を認めた釜山市は昨年 3 月，昆代表に名誉市民証を授与．同 11 月には自治体の補助金などが活用できる社団法人の認可も受けた．今後は言葉の違いを超えて楽しめる音楽やスポーツを通じた交流も企画したいと考えている． 昆代表は「日韓は地理的に近いがゆえに歴史的な摩擦もあるが，近いからこそ気軽に付き合える関係を築けるよう，国際交流のハードルを下げていきたい」と話す． グルメ紹介の動画人気　ユーチューバー韓国人カップル 釜山市内の飲食店などを日本語で紹介する動画を配信し，ネットを通じて日韓交流を楽しむ韓国人カップルがいる．ウギさん（29）＝本名チョヒョンウク，同市＝とチョイさん（33）＝本名崔敬善（チェギョンソン），光州市＝だ．2 人は動画投稿サイト「ユーチューブ」に「たび男女」というチャンネルを開設．ほのぼのとした会話を交わしながら，ご当地グルメなどを紹介している． 大学で日本語を専攻したチョイさんと，独学で磨いたウギさんは「旅行好きな日本人と友達になりたい」と 2018 年 10 月に動画配信をスタート．飲食店の紹介だけでなく，注文時に使える韓国語レッスン，ホテルで出前を頼む方法など，旅行者視点の動画も人気だ．チャンネル登録者数は 1 万人未満だが，日本人旅行者に街で声を掛けられることも増えてきたという． 昨夏以降，日韓関係の悪化は続くが，2 人は「国対国ではなく人対人の関係が大切だ．日本が好きな韓国人と，韓国が好きな日本人との間で良い関係を維持したい」と関係改善を待ち望んでいる．（釜山・前田絵）
カテゴリ	国際
テーマ	海外特派員リポート，国際面，新型コロナウイルス，海外支局発

推定する記事に付与されているカテゴリやテーマのそれぞれのタグ数を B，記事に付与されているそれぞれのタグが推定器の示す正解タグ中に含まれる数を C とするとき，以下の計算指標によって定量評価を行う．

$$Precision = \frac{C}{A}$$

$$Recall = \frac{C}{B}$$

$$Fscore = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

また，定量評価を行った結果を表 10 に示す．

表 10 より，$Fscore$ は全体的に低い値であることが確認できた．しかし，カテゴリ・テーマともに記事例 3 以外の $Recall$ の値が高く，記事に付与されていたタグが PMI を用いたカテゴリ・テーマ推定器によって推定できていることが確認できる．また，カテゴリ・テーマともに $Precision$ の値は全体的に低い

表 3　タグ推定に利用した記事例 3

タイトル	サッカーの先輩へ「ラストマッチを」　中 2 が企画，感染対策も奔走
本文	先輩たちにラストマッチをー．新型コロナウイルス感染拡大の影響で引退試合がなくなった中学 3 年生のために，福岡市早良区の中学 2 年松本京之介さん（14）がサッカー大会を企画した．会場の確保から感染対策まで，ほぼ 1 人で奔走．大会本番の 11 日を前に「大切な思い出づくりのため，ぜひ出場してほしい」と参加を呼び掛けている． 松本さんは同区の西南学院中サッカー部に所属．5 月 26 日の練習中，3 年生の最後の大会になるはずだった今夏の市中学校総合体育大会が中止になったと顧問の教諭に告げられた．3 年生はうつむいたまま，黙って聞いていた．小学校の頃のクラブチーム，学外で所属するサッカーアカデミーの先輩たちが悔しがる姿も脳裏に浮かんだ． 「先輩たちはどんなにショックだろう．自分の力でお世話になった全ての 3 年生に恩返しをしよう」．そう思い立ち，大会開催を計画．父親に相談すると「何かあったときは責任を取るから，好きにやりなさい」と背中を押してくれた． 最も気を使ったのは感染予防策だった．体育の授業の注意点をまとめたスポーツ庁の文書を参考に，参加者は手指の消毒を徹底して必ず検温し，体調が悪い人は参加できないルールにした．熱中症対策のため開催時間は夕方に設定した． 大会名は「ぼくらの中学ラストマッチ in 福岡」．試合会場の確保や消毒液，横断幕の購入などにかかった約 12 万円は，自分の貯金から出した． 会員制交流サイト（SNS）を通じて参加を呼び掛けると，小学校時代の先輩たちから連絡があった．「ありがとう．楽しみにしてるよ」．賛同した地元の人や同級生，サッカー関係者らが撮影や救護の担当を引き受けてくれた．松本さんは「準備は大変だったけど，諦めたくなかった．1，2 年生にも参加してもらい，3 年生の思いを受け継ぐ場にしたい」と意気込む．　（玉置采也加） 大会は 11 日午後 6 時から，福岡市東区の福岡フットボールセンターで．学校のサッカー部やクラブチームに所属している全国の中学生が対象で，定員は 33 人．当日 3 チームを編成し，1 試合 10 分で計 6 試合行う．応募締め切りは 7 日．大会のホームページはこちら．
カテゴリ	福岡，社会
テーマ	あなたの取材特命班，今日の 3 本，＃最後の夏残したい，社会面，新型コロナウイルス

表 4　記事例 1 におけるカテゴリの推定結果

カテゴリ	PMI
医療・健康	91.4171
速報	**56.7389**
宮崎	48.8991
熊本	37.1338
社会	**34.2861**
福岡	**27.0891**
ロアッソ熊本	21.6843
大分	20.0109
剣道	13.6141
全国・海外ニュース	11.6774

表 5　記事例 1 におけるテーマの推定結果

テーマ	PMI
重大速報	97.0087
福岡速報	**83.3768**
新型コロナウイルス	**72.0947**
超速報	71.6274
情報まとめ	68.4151
HTLV1	50.1121
病気のこと（コロナニュース）	43.6232
高校入試	40.1975
三菱重工	39.2138
福岡の新型コロナニュース	37.6369

表 6　記事例 2 におけるカテゴリの推定結果

カテゴリ	PMI
ロアッソ熊本	121.9331
長崎	83.9473
福岡	71.3308
国際	**66.0588**
宮崎	48.3409
剣道	43.2534
V・ファーレン長崎	41.8894
デスク日記	41.4529
全国・海外ニュース	32.4102
大分	32.4047

表 7　記事例 2 におけるテーマの推定結果

テーマ	PMI
海外特派員リポート	**559.1997**
めんたいぴりり	294.8626
セウォル号	290.7503
経済面	275.0274
国際面	**239.8927**
特派員オンライン	226.5607
家で過ごす（コロナニュース）	181.9863
海外支局発	**175.8988**
カピバラ	174.8647
生きる働く	159.0849

が，記事に付与されているタグの数に対して仮定した正解タグの数が多いためこのような結果になっているため，問題はないように考えられる．

3.4　定性評価による考察

本節では，3.2 節の分析結果から定性評価による考察を行う．表 1 のような記事ではタイトル，本文ともに新型コロナウイルスの感染情報を読者に伝える内容であり，新型コロナウイルスに関連したテーマタグが上位に表れていることが確認できた．しかし，表 2 や表 3 のような記事では新型コロナウイルスに関連する記事内容ではあるが，主題は新型コロナウイルスによっ

表 8　記事例 3 におけるカテゴリの推定結果

カテゴリ	PMI
高校サッカー	159.7015
剣道	153.1726
ロアッソ熊本	134.1184
玉竜旗	98.7630
玉鷲旗	87.2406
教育	85.6355
バレー	71.6860
ギラヴァンツ北九州	67.2993
野球	60.8791
V・ファーレン長崎	56.6058

表 9　記事例 3 におけるテーマの推定結果

テーマ	PMI
＃みんなの卒業式	189.5683
卒業式	146.8256
モヤモヤまとめ	144.4863
福岡マラソン	143.5264
アンプティサッカー	137.8670
TOKYO2020 地方からの挑戦	120.2688
かたろう九州 2020 東京五輪・パラリンピック	108.7193
お知らせ	106.0538
スポーツ面	106.0278
平和台を創った男〜岡部平太伝〜	103.0118

表 10　各記事における計算指標の計算結果

	カテゴリ			テーマ		
	Precision	Recall	Fscore	Precision	Recall	Fscore
記事例 1	0.30	1.00	0.46	0.20	1.00	0.33
記事例 2	0.10	1.00	0.18	0.30	0.75	0.43
記事例 3	0.00	0.00	0.00	0.00	0.00	0.00
macro ave.	0.10	0.83	0.22	0.17	0.58	0.25

て生活様式が変化したことに対して述べている．このことから，記事全体から新型コロナウイルスに関連する単語が表れづらいため，テーマの推定結果において「新型コロナウイルス」が上位 10 件に表れていない．また，表 3 の記事では全体的な内容として，サッカーに関連する内容であるためカテゴリの推定結果においてサッカーに関連するタグが上位に表れている．PMI の特徴として，カテゴリやテーマに依存した単語の値を計算する．そのため，本文にサッカー関連の単語が多くなるとサッカー関連のカテゴリやテーマが上位となり，新型コロナウイルス関連のタグが存在していない．さらに，「今日の 3 本」のような新聞のコラムに掲載されたときに付与されるテーマは，タイトル，本文のみで付与されるか判断することは難しい．そのため，現在のカテゴリ・テーマ推定器を用いて推定することも難しく，このようなタグの推定方法は今後の課題となる．

表 4〜9 の推定結果を確認すると，記事に付与されたカテゴリやテーマではないが付与されていても誤りではないタグが上位に表れていることが確認できる．例えば，表 4 の「医療・健康」は表 1 の新型コロナウイルスに関連する記事において適切なタグであると考えられる．このように，新聞記者が適切なタグを付与することができていない場合に，付与されるべきタグを推薦することで適切なタグが付与されていないといった問題を解決されることが考えられる．

4. おわりに

本研究では，電子データの新聞記事を対象に付与されたカテゴリとテーマについて PMI を用いて推定した．西日本新聞社の新聞記事のタイトル，本文の名詞を用いて PMI を算出した．また，算出した PMI を用いてカテゴリ・テーマらしさを算出し，PMI の算出に用いなかった西日本新聞社の 3 記事を用いて推定を行い分析結果について報告した．その結果，1 記事を除いてカテゴリ・テーマともに推定結果の上位 10 件に付与されたタグが多く存在していることが確認できた．さらに，分析結果に対して定量評価と定性評価から提案したカテゴリ・テーマ推定器から推定されたタグの妥当性などを考察した．

今後の課題としては 2 点挙げられる．1 つ目としては表 3 のような記事である場合に上手く推定できていないことである．付与されているカテゴリは「福岡」と「社会」であるがタイトル，本文を確認するとサッカー関連の単語が多いことが確認でき，「福岡」や「社会」らしさを表す単語が少ない．しかし，本文全体で考えたときに文脈から福岡でサッカーの大会が行われたことや新型コロナウイルスによって記事のような大会形式になったことが書かれている．本文から付与されたカテゴリついて読み解くことができるため，PMI を単語の位置を考慮した計算式に変更することなどに取り組む必要がある．2 つ目として，表 3 に付与されているテーマのタグにおいて「今日の 3 本」のようなタイトルや本文から判断できないタグが存在することである．このようなタグは現在の PMI を用いた手法で推定することは難しい．そこでタイトル，本文だけでなく，紙面における掲載位置やページ数などの外部情報を用いることで問題点の解決に取り組んでいく．

文　献

[1] Y. Liu, K. Han, Z. Tan, and Y. Lei, “Using context information for dialog act classification in DNN framework,” Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp.2170–2178, Association for Computational Linguistics, Copenhagen, Denmark, Sept. 2017.

[2] T. Jung, D. Kang, H. Cheng, L. Mentch, and T. Schaaf, “Posterior calibrated training on sentence classification tasks,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.2723–2730, Association for Computational Linguistics, Online, July 2020.

[3] X. Chen, C. Sun, J. Wang, S. Li, L. Si, M. Zhang, and G. Zhou, “Aspect sentiment classification with document-level sentiment preference modeling,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.3667–3677, Association for Computational Linguistics, Online, July 2020.

[4] T. Kudo, K. Yamamoto, and Y. Matsumoto, “Applying conditional random fields to japanese morphological analysis,” Proceedings of the 2004 conference on empirical methods in natural language processing, pp.230–237, 2004.