4B4-3

流行表現の原形推定に基づく由来検索システム

An Origin Search System for Trend Expression with Estimating Base Form

大部達也 *1 Tatsuya Ohbe　白松俊 *1 Shun Shiramatsu　大囿忠親 *1 Tadachika Ozono　新谷虎松 *1 Toramatsu Shintani

*1名古屋工業大学大学院情報工学専攻
Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

There are numerous trend expressions on the Web, which often include additional meaning from their origin. Understanding the expressions in conversation is important in terms of using social networking services. Trend expressions are often used with transforming, therefore it is difficult to identify their base form. In this paper, we implemented an origin search system trend expression with estimating its base form. We use Levenshtein distance, Jaccard distance and Dice distance for estimating the base form. We collected trend expressions and conducted an evaluate experiment the performance of each method.

1. はじめに

Web 上では様々な流行表現が使用されるが，流行表現には単純に文面からは推測できない何らかのニュアンスが込められている場合がある．流行表現の多くはその表現の由来に関する背景知識を前提としている．単純に文面通り意味だけでなく，その表現の由来に基づく意味を含む場合が多い．もしくは，文面が持つ意味と全く異なる意味を持つ場合もある．表現の由来を知っていて背景知識を持つ人は発言者が示唆していることを理解することができる．しかし，表現の由来を知らないと発言者の意図を正しく理解できないという問題がある．また，このような流行表現はしばしば一部を変形させて使用される．本研究では，変形された流行表現の元の表現を原形と呼ぶ．変形された表現を見た人はその原形を知らない場合，表現の原形を知らないと由来についての Web 検索が困難という問題がある．これには，変形された表現をそのまま Web 検索しても検索結果に原形が現れない，原形からどこが変化したのかわからないという理由が挙げられる．近年利用が増えている Twiter などの SNS 上では特にこのような表現が使用されることが多い．人間による文の理解や文の機械可読のために流行表現の理解は重要である．

本稿では，Web ブラウジング中に発見した流行表現に対して，流行表現の原形を推定する手法およびそれを利用した流行表現の由来を検索するシステムの実装について述べる．本研究では，レーベンシュタイン距離と Jaccard 係数および Dice 係数により表現同士の類似度を計算することで，その表現の原形を推定する．本稿では，人手で抽出した流行表現と思われる表現について本システムを使用して実験を行い，その結果と考察について述べる．

2. 関連研究

文書中に現れた表現が指す実体を推定することは，文書の理解や機械可読の上で重要である．本研究に関連する分野として表記ゆれの処理やあいまい検索などが挙げられる．Lu らはマイクロブログの投稿について指示語が指す実体の推定を行っている．[Li 15] 一般的な文書では，指示語が指す内容はその後より前に出現する．しかし，Twitter のような一つの投稿が短い SNS においては，文中に登場した指示語が指す語が前に出現しない場合が多い．Lu らは Wikipedia の記事をを用いて条件付き確率場 (CRF) により指示語が指す元の語を推定している．Lu らの実体検索では，異なる表現が同じ意味を表す場合を扱っている．本研究では，流行表現に特化したオンライン百科事典を利用する．また，同じ表現が変形することで，元の意味を継承しつつ，異なった意味を表す場合を扱う，

文中の文字の変化の修正として文字認識における誤り文字訂正が挙げられる．河田らは両方向 N-gram を利用して文字列中の誤りを検出している．[河田 05] 文字認識における誤りは 1 文字程度であることが多く，誤り箇所の前後の文字列を利用することができる．流行表現は時には表現の一部が大きく変化することが考えられる．また，表現が単体で使用されるために前後の単語が存在しない場合や，文脈を無視して唐突に使用される場合などが考えられ，周辺の単語を利用した推定が困難であると考えられる．そのため，本研究ではレーベンシュタイン距離を利用して，変化した流行表現に対してその原形を推定する．

日本語の自然言語処理において，単語を分割して品詞を推定する形態素解析は基本的な処理であり重要である．形態素解析の際に固有表現などの未知語が存在すると文が正しく分割されない問題がある．そのため，未知語の抽出が重要となる．Web における匿名掲示板や Twitter のような SNS では一般的な文書に比べ，新語や造語が頻繁に利用される．造語は既存の語を元に，規則的な変化をすることで作られる場合が多い．鈴木らはカタカナの名詞などから派生して生まれた動詞について，派生元の単語とその動詞が指す意味の推定を行っている．[鈴木 12] 鈴木らは Jaccard 係数とコサイン類似度により単語動詞の類似度計算を行っている．本研究では，同様に流行表現の類似度計算に Jaccard 係数と Dice 係数を用いる．河合らは伏せ語に対して有効な検索手法を提案している [河合 06]．河合らは，人手で作成したルールを用いて元の検索語から複数の伏せ語を生成することで検索語としている．また，生成した伏せ語の意味の違いを考慮して検索結果をクラスタリングしている．

笹野らは既存の語から派生した語と未知のオノマトペに対して有効な形態素解析の手法を提案している．[笹野 14] 笹野

連絡先: 大部達也，名古屋工業大学大学院工学研究科情報工学専攻，愛知県名古屋市昭和区御器所町，052(735)5584，tobe@toralab.org

らは派生した語や未知のオノマトペの生成ルールを複数併用している．これらの関連研究の例に対して，流行表現の変形の予測は困難であるため，本研究では，予め流行表現を Web 上から収集してデータベースを作成する．工藤らはひらがな混じりの文に対して頑健な形態素解析の手法を提案している．[工藤 12] 工藤らは EM アルゴリズムを用いて，ひらがな混じり文の生成モデルを推定している．

既存の語を組み合わせてできた固有表現は，一般的な語に分割されてしまうため，これらの手法による抽出は困難である．日本語の新語・造語辞書として mecab-ipadic-neologd[*1] がある．neologd は日々新たに生まれる新語や造語に対応するために，月に数回更新されている．neologd を利用することで，最新の語に対応した形態素解析を行うことができる．しかし，neologd を利用すると，新語や造語はひとまとまりの固有名詞として扱われる．本研究では，流行表現中の個々の単語に注目する必要があるため，通常の ipadic を利用する．

3. 流行表現の原形推定

本研究では，流行表現に関するデータベースを作成し，ユーザにより入力された文とデータベース中のエントリとの類似度を計算することで，原形を推定する．データベースの作成は，流行表現や新語などに特化したオンライン百科事典中の記事を利用する．本研究では，以下の仮定に基づいて流行表現の原形を推定する．

- 流行表現を構成する単語は由来に関する記事中に出現する頻度が高い
- 入力の表現と原形のレーベンシュタイン距離は小さくなる

一つ目の仮定から，本研究では，オンライン百科事典の記事中の単語と入力となる文に含まれる単語に対して集合間類似度を用いることで原形を推定する．二つ目の仮定から，記事のタイトルと入力となる文の距離を求めることで原形を推定する．

3.1 流行表現の分類

本研究では，流行表現を一般表現とローカル表現の 2 種に分類する．一般表現とは，SNS やテレビ掲示板など発祥の広く一般に用いられる表現である．ローカル表現は，特定のグループ間のみで限定的に使用されている表現である．例えば，主に実世界で使用されている場合や Line 等の非公開な SNS 上で使用されている場合がある．ローカル表現は Web 上から由来を取得することが困難であるため，本研究では一般表現のみを扱う．

本研究では，流行表現を表 1 のようなグループに分ける．「変化なし」に分類される表現は表現の変化が起こらず，そのままの形式で使用されると考えられる表現である．この内，(a) は個人名や団体名などの固有名詞である．(b) は既存の語の略語を指す．(b) は既存の語の言い換えに当たる表現である．(b) は既存の名詞やオノマトペなどから変化してできた動詞を指す．(c) は一般名詞が特定のトピックに関して用いられる場合に何らかの違った意味を持つ場合である．(d) は感嘆やオノマトペなどのフレーズである．

「変化あり」に分類される表現は表現中の一部の文字や単語を置き換えて使用される可能性があると考えられる表現である．(e) は名詞句のみからなる表現で，(f) は動詞句を含む表現で，(g) は記号やカタカナが特定の使用の仕方がされる場合である．(h) は語尾に特定の表現が使用される場合である．(i) は顔文字やアスキーアートである．これらの内，本研究では特に，(e) (h) の変化の起こる流行表現に焦点を当てる．

また，本研究では，流行表現の変化を大きく単語変化型，パターン型，省略型の 3 つに分ける．

- 単語変化形
  流行表現中の単語が変化して使用される場合．(g)，(h) に多いと考えられる．
- パターン型
  記号など特定の文字種によるパターンとして使用される場合．流行表現中の大部分の文字が変化する場合がある．(h)，(i)，(j)，(k) に多いと考えられる．
- 省略型
  流行表現の一部が省略されて使用される場合．(h) に多いと考えられる．

単語変化型の表現は，表現中の一部の単語が変化している場合である．単語変化型では，(g) や (h) において表現中の名詞や動詞が別の語に置き換わって使用されるといったことが考えられる．この場合，置き換わった単語以外の文字列は元の文字列と一致するため，レーベンシュタイン距離を用いた原形推定が可能であると考えられる．パターン型の表現は，記号や半角カタカナが特定のパターンで使用される場合であり，元の表現と使用される単語が大きく異る変化が考えられる．そのため，BoW を用いた原形推定よりもレーベンシュタイン距離による原形推定が適していると考えられる．省略型の表現は，(h) において句点を含む表現の場合に，句点より前のみ使用されるといった場合が考えられる．このような場合，入力となる表現は元の表現の一部に一致するため，BoW による原形推定が適していると考えられる．

表 1: 流行表現の分類

変化	特徴	例
変化なし	(a) 固有名詞	ＳＥＡＬＤｓ
	(b) 略語	ステマ
	(c) 言い換え	タヒる
	(d) 動詞	もにょる
	(e) 一般名詞	エンブレム
	(f) 感嘆	じぇじぇじぇ
変化あり	(g) 名詞句のみ	倍返し
	(h) 動詞句あり	安心して下さい、穿いてますよ
	(i) 記号	お・も・て・な・し
	(j) 語尾	～マン
	(k) 顔文字	(’ ω’)

3.2 原形推定

事典中の全エントリからタイトルと本文を切り出し，データベースに保存する．全エントリの本文に対して，形態素解析を行い，辞書を作成し，各エントリの BoW ベクトルを計算してデータベースに保存する．

文字列の類似度計算に用いられるレーベンシュタイン距離を利用している．レーベンシュタイン距離は，文字の挿入，削除および置換により元の文字列から対象となる文字列に変形するために必要な操作の回数である．レーベンシュタイン距離を用いることで，流行表現に変形が起こった場合でも元の表現を

*1 https://github.com/neologd/mecab-ipadic-neologd/

適切に選択することが可能であると考えられる．本研究では，入力された文字列に対して，データベース中の各エントリのタイトルとのレーベンシュタイン距離を計算し，最も距離の近いエントリを由来候補とする．

また，入力された文字列とデータベース中のエントリとの本文との類似度計算には Jaccard 係数および Dice 係数を用いる．これは，本文中から流行表現に当たる部分のみを適切に抽出することが困難であることから，レーベンシュタイン距離による類似度計算が行えないためである．また，記号や長音・拗音を含むような表現は，それらの使用の有無や回数が変わる表記揺れが起こりうる．レーベンシュタイン距離を利用する場合は表記ゆれのような場合の影響を受けるが，BoW の場合はその影響を受けないと考えられる．Jaccard 係数および Dice 係数は集合間類似度の計算に用いられる類似度である．集合 X, Y の Jaccard 係数は以下のように表される．

$$Jaccard(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

集合 X, Y の Dice 係数は以下のように表される．

$$Dice(X,Y) = \frac{2 * |X \cap Y|}{|X| + |Y|}$$

本研究では，検索語とデータベース内の各エントリの BoW ベクトルを Jaccard 係数および Dice 係数を用いて比較する．

4. 由来検索システム

本システムは，Google Chrome[*2] のエクステンションとして実装されている．本システムのシステム構成を図 1 に示す．本システムはサーバ・クライアント型 Web アプリケーションである．サーバでは事前にオンライン百科事典から流行表現の記事を取得する．Web ページ上のテキストをユーザが選択してクリックすることでテキストがサーバに送られる．取得した記事の本文から BoW ベクトルを生成し，データベースに保存する．データベース中の各ドキュメントには，記事のタイトル，本文，単語群および BoW ベクトルが保存される．原形推定部では，入力されたテキストとデータベース中の本文の類似度を計算し，類似度が閾値以上のエントリの概要をクライアントに送る．また，入力されたテキストからデータベース中の記事のタイトルとレーベンシュタイン距離の計算を行う．原形推定部で取得した由来候補のタイトルは由来検索部に送られる．由来検索部では，入力されたテキストから BoW ベクトルを生成し，データベース中の本文との類似度計算を行い，類似度の高い記事をクライアントに送る．また，原形推定部から取得した原形候補の本文中から概要に当たる部分をデータベースから取得してクライアントに送る．クライアントではサーバから受け取った検索結果をツールチップとして表示する．

本システムでは，ユーザが調べたいテキストを選択し，クリックすることでツールチップとして検索結果を表示する．本システムの実行例を図 2 に示す．実行例では Web ページ上で「芋戦車」という単語を選択したことで，「芋○○」という表現の由来である「芋スナイパー」に関する記事が検索結果としてツールチップに表示されている．

5. 実験

本研究では，匿名掲示板から流行表現と思われる文を人手で抽出し，本システムの手法で由来の取得を行った．今回，対

*2 https://www.google.co.jp/chrome/

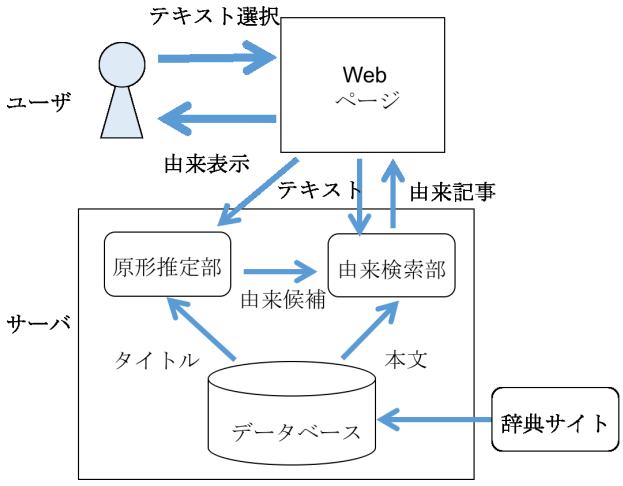

図 1: システム構成

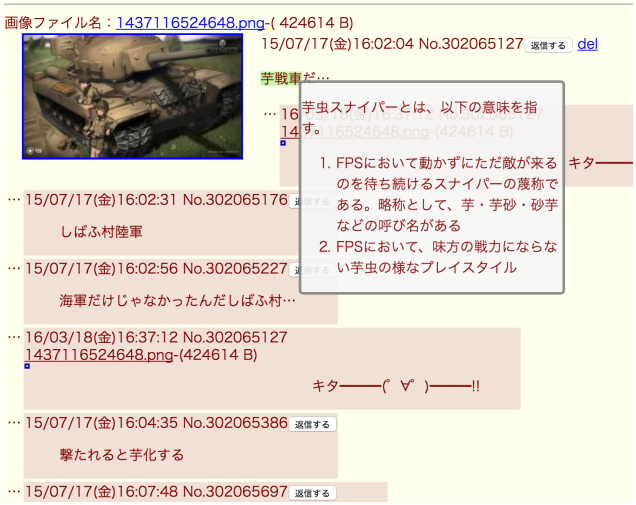

図 2: システム実行例

象としたのはニュース速報に関する匿名掲示板であり，掲示板中から人手で抽出した流行表現と思われる文は 70 件であった．本実験では，オンライン百科事典としてニコニコ大百科を利用し，146,686 件の記事のタイトルと本文を収集した．辞書を作成する際には，ベクトルの次元を減らすために，サイズを 10000 語とした．また，1 記事のみに出現する語および全記事中の 10%以上に出現する語を除外した．形態素解析には，形態素解析器 MeCab とその辞書 ipadic を利用し，BoW ベクトルを生成には，名詞，動詞および形容詞を使用し，その他の品詞を除外した．

実験の結果を表 2 に示す．レーベンシュタイン距離を用いた類似度計算では，人手で抽出した表現 70 件に対して 24 件が正しい由来を取得することができた．Jaccard 係数による類似度計算では 4 件が，Dice 係数による類似度計算では 2 件が正しい由来を取得することができた．

表 2: 実験結果

	Leven	Jaccard	Dice
正解件数	19	6	4

6. 考察

今回の実験では人手で抽出した流行表現の内の一部しかシステムでは抽出できなかった．今回の実験で抽出された流行表現は 3 章の (g) 名詞句のみ，(h) 動詞句あり，(j) 語尾，(k) 顔文字に該当する表現であった．今回，システムで抽出されなかった流行表現は以下の 4 種類に分けられる．

1. 人手で抽出した際に流行表現であると思われたが，実際は流行表現ではなかった
2. 対象とした事典中に該当する記事が存在しない
3. 対象とした事典中に該当する記事は含まれているが，タイトルに含まれていない
4. 対象とした事典中の記事のタイトルに含まれているが，抽出されなかった

1. は Web 上で検索した場合にその流行表現と由来についての記事が出てこないものであり，70 件中に 4 件あった．これらは流行表現ではないか，もしくは特定の狭いコミュニティで使用されているローカル表現である可能性がある．また，生まれたばかりの非常に新しい流行表現である可能性もある．2. は対象としたオンライン百科事典中に該当する記事が見当たらないが，Web 上で検索した場合にその流行表現と由来が出てくるものであり，70 件中に 13 件あった．これらを抽出するには，他のオンライン百科事典も利用するといった対策が考えられる．3. は対象としたオンライン百科事典中に該当する記事が確認されたが，タイトルには流行表現が含まれていないものであり，70 件中に 28 件あった．レーベンシュタイン距離による類似度計算では，記事のタイトルのみを利用しているため，これらの表現について由来を取得することができなかった．これらに対処するために，今後はタイトルのみではなく，本文も利用する必要がある．また，今回対象とした事典では流行表現がタイトルになっていないが，別の事典では流行表現がタイトルとなっている場合がある．そのため，(2) 同様に他のオンライン百科事典を利用することで対処できる場合もあると考えられる．4. は，事典中の記事のタイトルに該当する表現があるにもかかわらず，抽出されなかった表現で，70 件中に 25 件あった．これは，似た文字列の表現が他にあったことと，表現の変形により似た文字列の表現に近くなったことが原因と考えられる．また，省略型に当たる変化において正しく由来を取得できない場合があった．省略型の変形は，元の表現の一部であるにも関わらず文字数が大きく減ってしまうために，編集距離が大きくなることが原因である．

テストデータ中の 70 件の流行表現中，その由来がデータベース内の記事のタイトルとなっているものは 28 件であった．レーベンシュタイン距離による原形推定はタイトルに流行表現が含まれる場合については 28 件中の 24 件と，ほとんど正しく由来を取得することができた．このため，記事中から流行表現と考えられる部分を抽出することができれば，より多くの表現に対して由来を取得できると考えられる．BoW と Dice 係数や Jaccard 係数を用いた場合の結果が悪かった原因として，形態素解析の段階で文が正しく分割されていなかったことが挙げられる．これは，流行表現には未知語が含まれやすかったり，表記ゆれがあることによると考えられる．記号や顔文字が表れやすいことも正しい分割が行われなかったことの一因と考えられる．そのため，単語を正しく分割するか単語に分割しない手法を利用する必要がある．単語の分割に関しては別の辞書や形態素解析器を利用することが考えられるが，形態素解析においては一般的に複合名詞を 1 語として分割することが望ましい．多くの形態素解析器も複合名詞を 1 語として分割する．しかし，流行表現の変化は表 1(g) の場合のように複合名詞の一部が置き換わる例が想定されるため，本研究ではあえて複合名詞を各名詞に分割することが望ましい．よって，形態素解析を利用しない手法が適していると考えられる．形態素解析を利用しない手法として N-gram のように文字列を 1 文字ずつ分割する方法の利用が考えられる．本実験では，BoW を用いた場合ではレーベンシュタイン距離による原形推定では正しく由来を取得できなかった表現について由来の取得ができた．そのため，レーベンシュタイン距離と BoW を併用することで性能の改善が見込まれる．具体的には，BoW を利用した類似度によって由来候補の記事を事前に絞り込んでおき，その後レーベンシュタイン距離を用いて類似度を求めることで改良することができると考えられる．

今後は，複数の由来を持つ流行表現や由来となる表現にさらに由来がある場合についても扱えるように改良することが考えられる．また，ユーザがテキストを選択するのではなく，Web ページ中の流行表現と思われる部分を自動で抽出する改良が考えられる．

7. おわりに

本稿では，Web ブラウジング中に発見した流行表現に対して，流行表現の原形を推定する手法およびそれを利用した流行表現の由来を検索するシステムの実装について述べた．また，匿名掲示板から流行表現と思われる表現を人手で抽出し，システムによる流行表現抽出の実験を行った．本システムでは，事典のタイトルに含まれれる流行表現について由来を取得することができた．これにより Web 上の文書の理解の支援が可能になると考えられる．実験では，記事が本文に含まれる流行表現がうまく抽出されないことがわかった．今後は，本文中の流行表現に対応するための改良が必要である．

参考文献

[Li 15] Jun-Li Lu et.al. : “Entity Identification on Microblogs by CRF Model with Adaptive Dependency, ” Proceedings of the IEEE/WIC/ACM Web Intelligence Conference (WI2015)

[河田 05] 河田岳大 他 : “両方向 N-gram 確率を用いた誤り文字検出法”, 電子情報通信学会論文誌 D vol. J88-D2 No. 3 pp.629-635.

[鈴木 12] 鈴木 雄登 他 : “リムる・ドヤる・ポジる・パフェる—Web を用いたカタカナ動詞の言い換え・語源の獲得—” 情報処理学会研究報告 研究報告自然言語処理 vol 8 pp.1-7.

[河合 06] 河合利政 他 : “検索質問拡張に基づく伏せ語検索システムの試作” 第 5 回情報科学技術フォーラム FIT2006

[笹野 14] 笹野 遼平 他 : “日本語形態素解析における未知語処理の一手法—既知語から派生した表記と未知オノマトペの処理—” 自然言語処理 Vol. 21 No. 6 pp.1183-1205.

[工藤 12] 工藤拓 他 : “Web 上のひらがな交じり文に頑健な形態素解析” 言語処理学会 第 18 回年次大会発表論文集, pp.1272-1275.