The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

ソーシャルメディアストリームからの新固有表現の発見

Finding Emerging Named Entities from Social Media Streams

赤崎　智 *1　　吉永 直樹 *2　　豊田　正史 *2
Satoshi Akasaki　Naoki Yoshinaga　Masashi Toyoda

*1東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

*2東京大学 生産技術研究所
Institute of Industrial Science, The University of Tokyo

New things and concepts are born everyday in the real world. A database covering such information becomes indispensable in various tasks such as document understanding, information retrieval, entity monitoring. However, it is no longer impossible to manage the database manually, so it is desirable to collect these entities automatically and quickly. We aim to discover new named entities pointing to new things and concepts from social media streams using distant supervision. In the experiment, we use Twitter as a social media stream, and apply the proposed method to the weekly tweets. We confirmed that the performance of proposed method outperformed the baseline in terms of precision@k.

1. はじめに

我々の住む現実世界では製品や映画，音楽などの作品，イベント，人物などといった新しい事物が日々次々と誕生している．例えば，事物に関する情報を蓄積する有名な知識ベースである Wikipedia*1 では，2015 年 4 月から 2015 年 5 月の間に 41,579 件の新しい記事が登録されている [1]．Wikipedia に登録されるもの以外でも，スポーツ選手や店舗名，飲食物などは把握できないほど事物の数が年々増加している．これらを集積する知識ベースや辞書など（以降，まとめて辞書と呼ぶ）の事物のデータベースは事物の情報を利用するさまざまなタスクで必要不可欠となっている．例えば，昨今のウェブに存在する膨大な量のテキストを理解するためには，情報を事物単位にまとめ人間にわかりやすい形で整理して提示することが有効である．また，製品の不具合情報や，災害などのイベント等の特定の事物の動向をソーシャルメディアストリームなどで監視するソーシャルモニタリングも政府や自治体，企業にとって大変重要である [2]．

ここで，実世界に真に新しく出現する新事物は，その事物に関する情報の新規性から特に重要であるが，辞書に未登録であるためにその情報を活用することができず問題となる．また，冒頭で述べたような把握できないほど事物が多い種や，そもそも出現数が低頻度であるロングテールなものについては現状無視せざるを得ない状況となっている．

そのため，新しい事物が出現次第辞書を更新し，かつ事物をできるだけ網羅し最新の状態に保つことが必要となる．一般に辞書の整備は人手により行われるが，日々増え続ける新しい事物を人手で収集し辞書に追加することはもはや不可能であるため，計算機を用いて自動化することが望ましい．

そこで本研究では，事物のデータベース構築の支援のため，リアルタイムに次々と新しい情報が発信されるソーシャルメディアストリームである Twitter を対象とし，そこから新しい事物や概念を指す新固有表現を精度良く発見することを目指す．提案手法は新固有表現が最初に出現する時にその新規性を示唆するような文脈で出現することに着目し，そのような文脈を含むツイートを Wikipedia を用いた Distant Supervision で自動的かつ大量に獲得し，それらのツイートから新固有表現を発見する系列ラベリングモデルを学習する．

実験では，一週間分のツイートに対して，辞書に未登録のものを新固有表現と判定するベースラインと提案手法を適用し，後者がより精度良く新固有表現を発見できることを示す．

連絡先: {akasaki,ynaga,toyoda}@tkl.iis.u-tokyo.ac.jp

*1　https://en.wikipedia.org/wiki/

2. 関連研究

マイクロブログ等のソーシャルメディアストリームを対象とした固有表現抽出に関する研究がこれまで盛んに行われている．ソーシャルメディアストリームを対象に固有表現抽出を行う際の問題として，投稿が一般の新聞やニュース記事などの媒体と比べて文量が短くノイジーになりがちであるため，予測のための適切かつ十分な手がかりが得られないことが挙げられる．Ritter ら [3] はマイクロブログの文の短さや単語の表記ゆれ，不正確な Capitalization が品詞タグ付けやチャンキング，固有表現抽出の精度に影響を与えていることを指摘し，マイクロブログ投稿へのラベル付けによる学習データの拡張，ブラウンクラスタリングによる表記ゆれ語に関する素性の作成などを行い，各タスクの精度を改善した．また，固有表現に人名や場所といったカテゴリを割り当てる固有表現分類のタスクも行っており，そこでは抽出した固有表現が出現する全投稿からトピックを推定することで適切なカテゴリの割当を行っている．Ritter らの研究は一般の固有表現を対象としているが，本研究では新固有表現を対象としている点が異なる．

現実世界の事物（エンティティ）を指す固有表現を認識し，知識ベース中の項目と対応付けるエンティティリンキングというタスクにおいて，未知エンティティと呼ばれる対応付け先が知識ベースに存在しないようなエンティティが精度に影響を与えることが報告されている．Wu[4] はこの問題に対し，未知エンティティの判定を単一のタスクとして解くことを提案した．具体的には，固有表現に対し対応付けを行う前に，それが未知エンティティであるかどうかを文脈，分散表現，トピック，検索エンジンでの注目度，統語的要素の 5 つの要素から総合的に

判定するモデルを提案した．実際に彼らの手法をエンティティリンキングの前処理として実行したところ，対応付けの精度が有意に向上した．Wu らの手法は予め対応付けの対象となる固有表現が与えられており，固有表現部分の同定を問題としていない点が本研究と異なる．

3. 新固有表現

新固有表現を発見するにあたって，本研究における新固有表現が何であるかを定義する．

Graus ら [5] は，ニュースストリームとソーシャルメディアストリームにおいて新しく出現する事物が Wikipedia に記事として登録されるまでにどのような振る舞いをするか分析し，それらの事物のほとんどが「ニュースやソーシャルメディアストリームなどのメディアで散発的に言及されている状態」から「参考文献などの二次情報の充実により一記事として確立された状態」へと推移することを報告している．

これはあくまで Wikipedia に登録された事物を対象とした報告であるが，実際にはこのような状態の推移は Wikipedia に新規に登録されるような事物のみに限らず，実世界のあらゆる事物において確認できる普遍的なものである．Wikipedia は記事に内容の特筆性 [*2] を要求するため，いわゆる後者の状態へと推移した時点で記事として登録される場合がほとんどであるが，ソーシャルモニタリングなどの応用を考慮すると新しい事物は前者の状態のうちに早期発見することが重要である．また，ユーザはソーシャルメディアストリームにおいて新しい事物についての投稿を発信するとき，以下の例 [*3] のようにその事物の新規性を示唆するような表現で発信を行うことが多い．

> 『ポケモン』新作映画のタイトルが『劇場版ポケットモンスター みんなの物語』に決定！　予告編映像も公開
> https://www.famitsu.com/news/201802/27152683.html

そのため，このような表現が現れる投稿を収集することができれば，そこに現れる固有名詞，すなわち新固有表現を発見することが可能となる．これらを踏まえ，本研究では発見を目指す新固有表現の定義を「公共性を有するメディアにおいて，新規出現に関する情報が提供・共有されるような固有名詞」とする．

4. 提案手法

3 節で述べた新固有表現をソーシャルメディアストリームから発見するための提案手法について述べる．提案手法は新固有表現が出現する投稿を Distant Supervision により獲得し，それらを用いて新固有表現にラベル付けを行う系列ラベリングモデルを構築する．構築したモデルをソーシャルメディアストリームである Twitter の投稿（ツイート）へと適用し，実際に新固有表現の発見を試みる．

4.1 Wikipedia を利用した Distant Supervision

教師あり学習は大量の学習データを要求するが，データへのラベル付与は基本的に人手で行う必要があるためコストが大きい．また，人手で付与を行ったとしても，将来的に辞書が更新されたり発見対象の性質や種類が変化した場合には学習データの再構築をする必要があり現実的でないため，学習データを自動生成する仕組みが求められる．

Mintz ら [6] は，知識ベースなどを手がかりに擬似的な学習データを生成する Distant Supervision という手法を提案した．新固有表現のタグ付けがなされたソーシャルメディアストリームのデータセットは存在しないため，本研究でも Distant Supervision を用いて擬似的な学習データを大量に作成し，それらのデータを用いて新固有表現を発見するモデルを学習する．

3 節で述べた通り，新事物は出現の初期の段階では，その新規性を示唆するような表現で発信される．我々はこれに着目し，既知の事物の時系列における初期のツイートを正例として収集した．具体的な手順は以下の通りである．

1. Wikipedia に登録された記事のタイトルを既知の事物の固有表現として収集する
2. 得られた各固有表現を含むツイートを固有表現毎に時系列の古い順に収集する．

負例については，正例として集めた事物について，それぞれ収集した時刻の一ヶ月後のツイートを正例と同数収集する．

Distant Supervision は大量の疑似学習データを生成できる一方で，誤ったラベルのデータが混入してしまうことが多い．そのため，収集の際に以下のような制約を設ける．

- 収集する事物について，Wikipedia に記事として登録された半年前以前に 5 回以上出現しているものは候補から除く．これは，語義の曖昧さが存在する事物や，既に既知であるものとして普及してしまっている事物が正例及び負例に混入しないようにする目的がある．
- 上記手順 2 で正例のツイートとして収集する対象はリツイートされたものとする．また，1 日に最低 10 回以上リツイートされない事物については候補から除く．これは出現の初期の段階で一定頻度リツイートされるような事物は，ユーザが注目するような新規性の高いものであるという仮定を置いている．
- 上記手順 2 で事物毎に収集するツイートの時系列の範囲は，事物の出現が最も古い日から，1 日に 10 回以上リツイートされる最初の日までとした．これは，事物毎に最初にバーストするツイートとそれ以前のツイートではその事物の新規出現に関する情報が提供されやすいという仮定を置いている．

4.2 CRF を用いた新固有表現の発見

4.1 節において収集した疑似的な学習データから新固有表現を発見するための教師あり学習モデルを構築する．提案手法では新固有表現の発見を，固有表現抽出のタスクにおける標準的な枠組みである系列ラベリング問題として解く．教師あり学習のモデルとしては条件付き確率場（CRF）を採用する．

CRF はトークン列 $\boldsymbol{x}$ を入力とし，トークンのラベル列である $\boldsymbol{y}$ を出力する．具体的には以下の式のような条件付き確率で定義される．

$$Pr(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Lambda}) = \frac{1}{Z_{\boldsymbol{x}}} \exp\left(\sum_{i=1}^{|\boldsymbol{x}|-1}\sum_{k}\lambda_k f_k(y_i, y_{i+1}, \boldsymbol{x}, i)\right)$$

$Z_{\boldsymbol{x}}$ は正規化係数である．$f_k()$ は素性関数でトークンとラベルの依存関係を表現しており，形態素解析の結果や文字種などが用いられることが多い．$\boldsymbol{\Lambda}$ はモデルのパラメータであり，学習により調整される．学習したモデルは，任意のトークン列

∗2 https://ja.wikipedia.org/wiki/Wikipedia:独立記事作成の目安
∗3 https://twitter.com/famitsu/status/968399069189087235

$\boldsymbol{x}'$ に対して最適なラベル列 $\hat{\boldsymbol{y}}' = \arg\max_{\boldsymbol{y}} Pr(\boldsymbol{y}|\boldsymbol{x}'; \hat{\boldsymbol{\Lambda}})$ を出力する．

CRF の学習データは 4.1 節で収集したツイートを分かち書きし新固有表現に対応するトークンに B，I，O，E，S の 5 種のラベルを割り当てたものを用いる．ラベルはそれぞれ，新固有表現の開始位置，新固有表現の 2 トークン目以降，新固有表現の外側，新固有表現の終了位置，単一のトークンからなる新固有表現を意味する．

素性については，現在位置のトークンとその前後 2 トークンの表層，品詞細分類，文字種，前トークンの出力ラベルを用いる．

5. 実験

提案手法をツイートに適用し，実際に新固有表現を発見することで有効性を確認した．以下ではまず詳細な実験設定について述べ，その後実験結果を定量的・定性的に確認する．

5.1 実験設定

学習データ

学習・評価に用いるソーシャルメディアストリームのデータセットを我々の研究室で 2011 年 3 月より継続して収集 *4 している Twitter のデータをもとに構築した．まず 4.1 節の手順 1 を実行し，2012 年 3 月 11 日から 2015 年 12 月 30 日までに Wikipedia に新しく登録された記事タイトルを既知の事物の固有表現として収集（このとき，リダイレクトページや曖昧さ回避の記事については除外した）し，これらの事物について 2011 年 3 月 11 日から 2012 年 3 月 11 日までの期間でツイート頻度を集計し，頻度が 5 を超えるものは除外した．これにより残ったそれぞれの事物について，手順 2 を実行し時系列の古い順にリツイートを収集し，1 日に 10 回以上リツイートされた事物 11,698 件を残し，それぞれの事物の出現が最も古い日から，1 日に 10 回以上リツイートされる最初の日までのリツイートを最終的に抽出し，結果 207,508 件を最終的な正例として採用した．負例については 4.1 節に述べた方法で 11,698 件の事物それぞれについて収集し，結果 155,254 件を負例として採用した．これら正例と負例を合わせたものに ipadic を用いた MeCab*5 で分かち書きを行ったものを学習データとした．

評価手順

2016 年 6 月 10 日から 2016 年 6 月 16 日までの 1 週間を対象に，各日付の全日本語ツイート（リツイートとリプライ，クオートは除く）に各手法を適用し，各日付毎に出力された新固有表現の候補を出力された回数が多い順に並び替え，Precision@k(k=50) を算出した．この時，短期間で同名の新固有表現はほとんど出現しないという仮定を置き，対象としている日の各出力が前日までの各日付の出力上位 50 件に入っていた場合にはそれを評価から除外した．新固有表現の発見の評価には本来 Recall も重要であるが，数千万ツイートを対象に出現した新固有表現の総数を算出することは現状不可能と判断したため，便宜的にこのような手法を取った．正解ラベルについては各日付の出力である新固有表現候補の上位 50 件に対して，第一著者が発見元のツイートを確認し新固有表現であるか否かを判断し付与した．

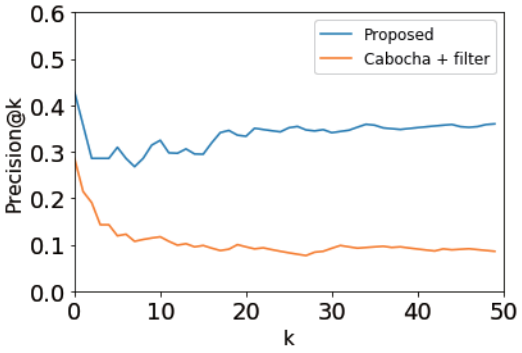

図 1: 各手法の Precision-Curve

比較手法

新聞記事のコーパスから学習された一般固有表現抽出器をツイートに対して適用し，出力された固有表現が Wikipedia に未登録であれば新固有表現とみなす手法をベースラインとして採用し提案手法との比較を行った．提案手法の CRF の実装物として CRFsuite*6 を，ベースラインの一般固有表現抽出器として Cabocha*7 を用いた．CRF のパラメータ最適化には L-BFGS 法に基づく L2 正則化を適用し，L2 正則化の係数については学習データの 1 割を開発データとして用いて調整した．

5.2 実験結果

5.1 節の評価手順における各日付の Precision@k を平均し，各 k の値を手法毎にプロットしたものを図 1 に示した．どの k においても提案手法がベースラインの数値を上回っており，特に k が大きくなるにつれ数値の差も顕著になっていることが確認できる．これは出力の数 k を増やしても提案手法の性能が比較的安定していることを示している．提案手法はツイート中の新規性を示唆するような表現を捉えたうえで新固有表現を認識しているが，ベースラインは辞書に未登録なものをすべて新固有表現として判定してしまうため，偽陽性の割合が大きく性能低下の原因となっている．

提案手法で発見できた新固有表現の例を発見元のツイートとともに表 1 に例示する．「ゼルダの伝説ブレスオブザワイルド」などの有名な事物だけでなく，「MUFG コイン」や「GeForce GTX1070」などのロングテールな事物の固有表現を適切に発見できていることが確認できる．また，「ソニーネットワークコミュニケーションズ」などの事物の名称変更や，「アクセル・ワールド」などの表層としては既知だが新しい語義などを発見できているのは興味深い．とくに新語義については，ベースラインのように既存の辞書でフィルタリングを行う手法では発見できないものであるため，この点でも提案手法は有効であることがわかる．

最後に提案手法で誤って出力した例を発見元のツイートとともに表 2 に例示する．1 つめの例の「ソードアート・オンライン」のように，実際は既知の事物の固有表現だが，ツイート中に新規性を示唆するような表現が存在したために誤って出力してしまったものが多かった．2 つ目の例の「舛添要一」も同様だが，こちらはツイート中に新規性を示唆するような表現が存在しないのにもかかわらず誤って出力してしまっている．

*4 2011 年 3 月から 2017 年 1 月までの期間で約 380 億ツイートを保有

*5 http://taku910.github.io/mecab/

*6 http://www.chokkan.org/software/crfsuite/

*7 https://taku910.github.io/cabocha/

表 1: 提案手法で正しく出力した例

出力	発見元のツイート
ゼルダの伝説ブレスオブザワイルド	最新作 『ゼルダの伝説ブレスオブザワイルド』 映像が公開! _URL_
	【速報】 ゼルダ新作の正式タイトルが『ゼルダの伝説ブレスオブザワイルド』に決定!!! _URL_
MUFG コイン	三菱東京 UFJ が独自の仮想通過『MUFG コイン』を発行! 本格的に紙幣を持ち歩かない時代に _URL_
	【仮想通貨】 三菱東京 UFJ 銀行、「MUFG コイン」を来秋発行へ _URL_ 利用者同士が手軽にやり取りしたり、割安な手数料で外貨に交換したりできる。スマホのアプリに取り込むなどして使う。
GeForce GTX1070	「GeForce GTX1070」が 6 月 10 日 (金)22 時に解禁 一部ショップでは深夜販売を実施 -アキバ総研 _HASH_ _URL_
	【超悲報】 本日 22 時解禁予定の NVIDIA 新型 GPU「GeForce GTX1070」搭載ビデオカードですが、ASUS 製は本日発売ではありません。近日中発売ですのでご注意ください。 _URL_
ソニーネットワークコミュニケーションズ	ソネット、7 月より新社名 「ソニーネットワークコミュニケーションズ株式会社」 に _URL_ _HASH_
	ソネットが社名変更ー「ソニーネットワークコミュニケーションズ」に、また長く ーインターネットコム _URL_
アクセル・ワールド	【特報】 劇場版 『アクセル・ワールド』 新作アニメの新キービジュアル & 新キャラ初公開! 来場者プレゼントで川原礫書き下ろし「AW」特別短編小説が貰えることも判明 _URL_ _HASH_
	劇場版「アクセル・ワールド」 第 2 弾キービジュアル公開 新キャラクターも明らかに _URL_

表 2: 提案手法で誤って出力した例

出力	発見元のツイート
ソードアート・オンライン	TV アニメ 『ソードアート・オンライン』 第 1 期の再放送が決定! 松岡禎丞さん・日高里奈さんら出演のニコ生特番も _URL_
	【朗報】 アニメ『ソードアート・オンライン』1 期が再放送決定キター (゜∀゜) ー!! _URL_
舛添要一	血税を納めてる庶民には、生活が苦しい人もいっぱいいるのに。リオに行きたいからって居座るな 【炎上】 舛添要一が政治資金で「高級すきやき牛肉」と「高級ワイン」を購入した事が判明/公用車でプロ野球観戦の疑惑も _URL_
	「舛添要一語録」の裏側をコンサル目線で解読する - 今週もナナメに考えた鈴木貴博ダイヤモンド _URL_

これについては学習データを分析したところ，政治家の人物がニュースなどで話題になっている状況のツイートが多数含まれていた．これは Distant Supervision で学習データを構築したのが原因であり，政治家などは Twitter 上での新規出現がスキャンダルなどなんらかのニュースになったタイミングである場合が多く，提案手法の Distant Supervision でそれらのツイートを学習データとして収集してしまった結果といえる．

6. おわりに

本研究では，ソーシャルメディアストリームから新固有表現を発見するタスクに取り組んだ．提案手法は Distant Supervision により自動的かつ大量に収集した新固有表現とその新規性を示唆するような表現が現れるツイートを，系列ラベリングモデルである CRF で学習し，実際に 1 週間分のツイートからの新固有表現の発見を行った．結果としては提案手法はベースラインの手法を大きく上回ることとなり，1 節で述べたようなロングテールな事物や表層としては既知だが新しい事物なども発見することが出来た．

今後の課題としては，Distant Supervision の工夫や CRF の素性の拡張による精度改善がまず挙げられる．また，評価手法として今回 Precision@k のみを用いたが，実世界の新しい事物をどれだけカバーできたかを示す指標である Recall を評価する手段も検討する必要がある．

7. 謝辞

本研究の一部は JSPS 科研費 16K16109 と 16H02905 の助成を受けたものです．

参考文献

[1] Michael Färber, Achim Rettinger, and Boulos El Asmar. On emerging entity detection. *European Knowledge Acquisition Workshop*, pp. 223–238, 2016.

[2] Yue Sui and Xuecheng Yang. The potential marketing power of microblog. *Communication Systems, Networks and Applications (ICCSNA), 2010 Second International Conference on*, Vol. 1, pp. 164–167, 2010.

[3] Alan Ritter, Sam Clark, and Oren Etzioni. Named Entity Recognition in Tweets : An Experimental Study. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1524–1534, 2011.

[4] Zhaohui Wu, Yang Song, and C Lee Giles. Exploring Multiple Feature Spaces for Novel Entity Discovery. *Proceedings of the Conference on Artificial Intelligence (AAAI)*, pp. 3073–3079, 2016.

[5] David Graus, Daan Odijk, and Maarten de Rijke. The birth of collective memories: Analyzing emerging entities in text streams. *arXiv preprint arXiv:1701.04039*, 2017.

[6] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011, 2009.