マイクロブログからの消失エンティティの検知

赤崎 智 [1] 吉永 直樹 [2] 豊田 正史 [2]
[1] 東京大学大学院 情報理工学系研究科 [2] 東京大学 生産技術研究所
{akasaki，ynaga，toyoda}@tkl.iis.u-tokyo.ac.jp

概要

人物や作品，イベント等のエンティティに関する行動や意思決定を補助するため，閉店する施設や終了するサービス等の現実世界から消失するエンティティに着目し，それらをマイクロブログから検知する新しいタスクに取り組む．我々は消失エンティティが特有な文脈を伴うことに着目し，それらを高精度に収集できるように distant supervision の手法を改良する．また，収集したデータを使ってエンティティ認識モデルを学習する際に，単語分散表現をマイクロブログ投稿で追加学習し与えることで，大域的な特徴を考慮する．実験により提案手法が精度良く消失エンティティを検知できることを確認する．

1 はじめに

我々は，人物や作品，イベントといった現実世界を構成するエンティティの変化を意識して，日々の行動や意思決定を行っている．これらのエンティティの中でも特に重要なのが，その消失[1)]に関する情報である．例えば，一般人であれば興味関心のある施設や店舗等の閉店情報や近日中に終了するイベントをいち早く知る必要があり，企業であれば終了するサービス等を網羅することが業界内外のトレンド分析等に繋がる．また，エンティティを集積する知識ベースにおいても，消失したというエンティティの状態を更新することが必要である．

エンティティの消失を捉える方法として，知識ベース等のエンティティ辞書を用意し，それらの出現頻度を監視する手法が考えられる．しかし，エンティティは消失しても言及され続ける場合があったり，逆に長期間言及されていなくてもそれが消失ではない他の要因であったりするため，何をもってエンティティが消失したかを判断するかが難しい．

これを踏まえ，本研究は公式情報やニュース，個人の体験に至るあらゆる情報が発信されるマイクロブログを対象に，消失エンティティを検知するタスクに取り組む．我々はユーザがマイクロブログで消失エンティティに言及する際，消失を示唆する特有の文脈 (表 1) を使うことに着目する．これらの文脈を効率的に収集するため，知識ベースと時系列テキストから半教師あり学習でラベル付きデータを生成する time-sensitive distant supervision (TDS) [1] を改良して用いる．具体的には，知識ベースに記述されたエンティティの消失タイミングを文脈収集時に考慮し，なるべくノイズが混入しないようにする．

収集したデータを用いて，NER による消失エンティティ認識モデルを構築する．この際，消失エンティティがマイクロブログ上でバーストして出現しやすいことに着目し，これらの複数投稿を用いて単語分散表現を追加学習し，NER モデルの追加特徴として与えより頑健なエンティティ認識を行う．

実験ではマイクロブログである Twitter から，改良した TDS で訓練データを収集し，提案する NER モデルを学習する．人手で構築したテストデータに対しモデルを適用することで消失エンティティの認識を評価し，提案手法の有効性を確認する．

2 消失エンティティ

本節では消失エンティティを，赤崎ら [1] の新エンティティの定義を参考に定義する．赤崎らは，現実世界に新しく現れたエンティティは最初に現れてから世間に周知されるまでの過程があり [2]，その過程において特有な表現を伴って言及されることを報告している．我々は消失エンティティも同様に消失した時点だけでなくそれまでの過程で，予定や前兆を示す特有な表現が文脈に現れる (表 1) ことに着目し，以下のように消失エンティティを定義した:

消失文脈 読み手がエンティティの消失を認知していないことを，書き手が仮定している文脈

消失エンティティ 消失文脈で現れるエンティティ

1) 本研究では簡便のため、エンティティの存在が実世界からなくなること (例: 人の死没，店の閉店) を，消失とみなす．

タイプ	エンティティ数	投稿数	消失文脈の例 (**太字**は消失エンティティ、下線は消失文脈特有の表現)
PERSON	780	49378	**Roger Ailes** died of complications of a subdural hematoma after he fell at home, hit his head.
CREATIVE WORK	975	48444	RT @USER: See what's coming up in the **Diggnation** Finale airing next week⋯
LOCATION	240	3545	Sad news, crime sleuths. The **National Museum of Crime and Punishment** in D.C. will close⋯
GROUP	1187	60768	⋯ anymore. Adam's not in Three Days Grace anymore. **My Chemical Romance** broke up⋯
EVENT	186	9880	Live music blow. MT @TomTilley: C3 confirms **Big Day Out** will NOT go ahead in 2015.
SERVICE&PRODUCT	709	35148	was what inspired me to write for mags. RT @USER: Future closes **Nintendo Gamer magazine**
計	**3213**	**163700**	

表 1: TDS で Twitter から収集した消失エンティティデータセットの統計 (英語) と消失文脈の例

この消失文脈を捉えることで，死没した人物等の既に消失したエンティティだけでなく，終了前のイベントや施設，サービス等の事前にその消失を知ることが重要なエンティティも早期に認識できる．

3 関連研究

我々が定義する消失エンティティを認識する研究は存在しないが，本節では関連研究を概観する．

エンティティについての知識を抽出するために，テキスト中のエンティティを知識ベースのエントリと対応付ける Entity Linking [3] タスクがある．消失の情報も対応付けた先から抽出できるが，知識ベースの更新は遅いため早期に知識が取り出せず [4]，また低頻度なエンティティについては対応付けるエントリがそもそも存在しないという問題がある．

テキスト中に出現するエンティティの様々なイベント (例: 人物の出生や結婚，死没) を認識するイベント抽出 [5, 6, 7, 8] や，その期間 (例: 人物の存命期間) を同定する時間的穴埋めタスク [9, 10] がある．しかし，これらが扱うイベントの種類はごく僅かであり，多様なエンティティの消失に対応できない．

我々とは対照的に，赤崎ら [1, 11] は現実世界に新しく現れる新エンティティを発見するタスクに取り組んだ．彼らは新エンティティが現れる際に特有の文脈を伴うことに着目し，それらをマイクロブログ投稿の時系列を利用し収集する TDS を提案した．

4 提案手法

4.1 改良 TDS に基づく消失文脈の収集

赤崎ら [1] の TDS はマイクロブログ等の時系列テキストのタイムスタンプを利用し，知識ベースのエンティティの特定の文脈を収集する手法である．彼らは新エンティティに特有の文脈に着目し，Wikipedia の各エンティティについて，それがマイクロブログの時系列で最初に現れた投稿を新エンティティの文脈とみなし収集した．これを消失エンティティの収集に当てはめると，エンティティについて時系列で最後の投稿を集めることになるが，消失エンティティは新エンティティと異なり実際に消失した後も言及が続くため，消失とは関係のない文脈が集まってしまう．そこで我々は，消失の大まかなタイミングを考慮し精度良く消失文脈を収集するよう TDS を改良する．具体的な手順は以下である:

Step 1: 消失エンティティ候補の収集　まず，Wikipedia の記事タイトルをエンティティとして収集する．この時，実際に消失しているエンティティのみを集めるため，Wikipedia の「廃止された事物一覧[2)]」からエンティティとその廃止年を収集する．ここから，廃止年と Twitter 上で最初に出現した年が同一であるエンティティ及び，曖昧さ回避の記事が存在するエンティティを除去する．これは，訓練データに新エンティティや同名異義のエンティティが混ざることで文脈が汚染されることを防ぐ．

Step 2: 消失文脈の収集　収集した各エンティティについて，廃止年の Twitter 上で最も出現頻度が高かった日の投稿を上限 100 件まで収集する．これは，廃止年に最も関心を得たタイミングの投稿は消失文脈を含むという仮定に基づいている．

以上で収集した文脈を正例としそのまま NER モデルを訓練してしまうと，モデルは訓練事例のエンティティ文字列を検出するよう学習してしまい [12]，文脈を識別する能力が欠如する．そのため，赤崎らは各エンティティについて，収集した文脈とは異なる文脈の投稿を負例としてサンプルし，モデルが文脈を識別しエンティティを認識できるよう工夫した．我々も同様に正例として集めた各エンティティと消失文脈について，それの廃止年よりも前年の投稿，すなわち確実に消失していない時点の投稿を非消失文脈として同数サンプルし負例とする．

2) https://en.wikipedia.org/w/index.php?title=Category:Disestablishments_by_year

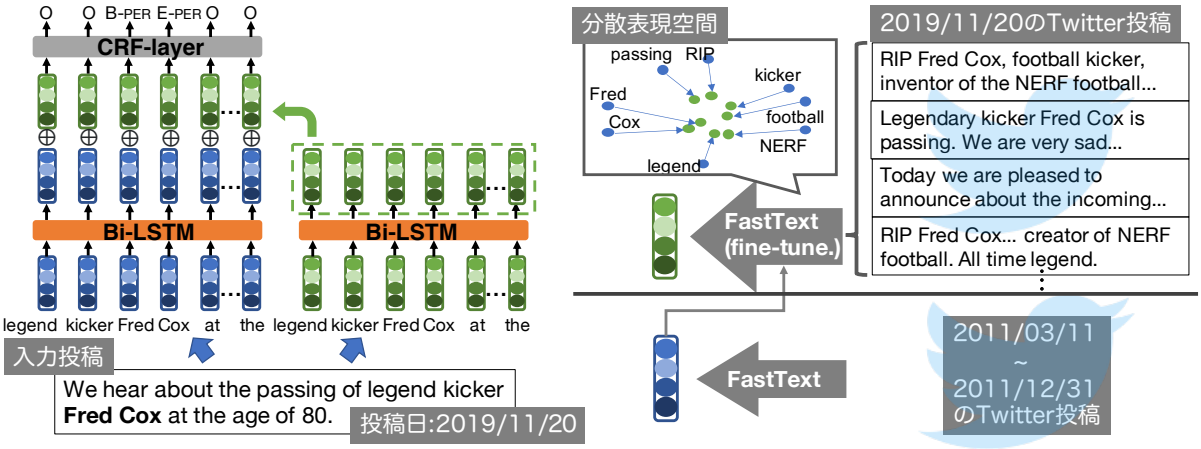

図 1: マイクロブログを利用した単語分散表現の追加学習 (右) とそれを用いた消失エンティティの認識 (左)

4.2 消失エンティティの検知

収集したデータを用いて消失エンティティを認識する NER モデルを訓練する．消失エンティティはその数自体が少なく，収集したデータもある程度ノイズを含むためモデルの学習が不安定になる．そこで我々はマイクロブログにおいて消失エンティティが観測される時にバーストしやすい現象に着目し，それらの投稿から得た特徴をモデルに入力し頑健な学習を行う．以下で単語分散表現を利用した特徴抽出と，それを組み込む提案モデルについて説明する:

マイクロブログを用いた単語分散表現の追加学習

前述した通り複数投稿から特徴を抽出したいが，モデルのテスト時に認識対象のエンティティは不明であるため，関連する投稿を大規模な Twitter 投稿から選択することは難しい．そこで我々は，多くの深層学習モデルが単語分散表現をモデルに入力することに着目し，これを消失エンティティを検出する日付の Twitter 投稿で追加学習することを提案する (図 1 右)．追加学習は投稿群を入力とする教師なし学習であり関連投稿も抽出する必要がない上，得られる分散表現は検出日の Twitter 投稿によく現れる単語や共起を反映しており有用な特徴となっている．

具体的には我々はまず，前節でデータを収集した期間以前の投稿を用いてベースとなる単語分散表現 v_{base} の事前学習を行う．次に，前節で収集したデータの正例負例の全投稿の日付を列挙し，各日付 d の投稿を用いて v_{base} を追加学習し，v_d を得る．

追加学習済単語分散表現を用いた系列ラベリング

我々は，系列ラベリングで NER を行う Akbik ら [13] の手法を認識モデルとして用いる．このモデルは，入力文の各単語を文字ベース Bi-LSTM 言語モデルでエンコードし，事前学習済単語分散表現 v_{base} と合わせて Bi-LSTM に入力し CRF 層で予測を行う．

我々はもう一つ Bi-LSTM を準備し，入力投稿の投稿日 d における追加学習済の v_d を各単語毎に入力し，得られた各時刻の出力を元の Bi-LSTM の出力と連結し CRF 層に入力する (図 1 左)．これにより，モデルは各単語に関する Twitter 上の大域的な情報も考慮することができ，頑健な予測が可能となる．

5 実験

5.1 データ

我々は 4.1 節の手法を研究室で 2011 年から収集している 50 億投稿規模の Twitter アーカイブに適用し，英日の言語のデータセットを構築した．まず，Step 1 を実施し，Wikipedia の廃止された事物一覧から 2012 年から 2019 年の記事を収集した．この際，廃止ページに記載された記事のカテゴリを人手で粗いタイプに分類し (英語 6 種, 日本語 5 種)，それらをエンティティに付与した．この時，PERSON と CREATIVE WORK タイプのエンティティ数が他のタイプのものより極端に多かったため，それぞれ 1000 件にダウンサンプルした．次に Step 2 を実施し，各エンティティについて正例と負例を収集した．我々はこの内の 2018 年までのデータを訓練データ，2019 年のデータを後述のテストデータとして用いた．表 1, 4 に構築した訓練データの内訳をタイプ別に示す．訓練データの 10%を開発データとして用いた．

我々は適切な評価を行うため，テストデータを人手で以下の手順で作成した．英日両言語において，

2019 年の各エンティティの正例からランダムに 3 投稿を抽出し，主著者を含めた 3 人のアノテータに各投稿が消失文脈を含むか判定させ，2 人以上が含むと判定した投稿をテストデータの正例として採用した．また負例を用いて同手順で各投稿が非消失文脈を含むか判定させ，正例と同数テストデータの負例として採用した．これにより，英語の消失エンティティ 357 件と文脈 1,922 件，日本語の消失エンティティ 235 件と文脈 1,326 件を得た (表 5, 6)．Fleiss's Kappa による判定者内一致は英語が 0.722，日本語が 0.786 で両者ともに十分に高く，我々の消失エンティティの定義が適切であることを示している．

5.2 モデル

我々は以下の 3 つのモデルを比較に用いる:

改良 TDS + 追加学習: 提案手法で構築したデータを用いて，提案モデルを訓練した．

改良 TDS: 提案手法で構築したデータを用いて Akbik らのモデルを訓練した．本モデルは追加学習した単語分散表現を用いない．

ベースライン: TDS を 4.1 節の改良なしで行いデータを構築し (手順の詳細は A.2 参照)，それを用いて Akbik らのモデルを訓練した．

5.3 実験設定

我々は構築したデータの各投稿から URL，ハッシュタグ，ユーザ名を除去し，日本語については MeCab (ipa 辞書) を，英語については spaCy を用いてトークナイズした．Akbik らの NER モデルは Keras (ver.2.3.1) を用いて実装し，パラメータは Akbik らが推奨するものと同じとした．文字ベース Bi-LSTM 言語モデルは 2011 年 3 月 11 日から 2011 年 12 月 31 日の英語 20 億投稿，日本語 8 億投稿を用いて訓練した．同様の投稿を用いて，300 次元の単語分散表現 v_{base} を fastText [14] を用いて訓練した．訓練及びテストデータの投稿の日付 d の投稿 (平均 200 万投稿) を用いて v_{base} を追加学習し，v_d を獲得した．我々は開発データで F 値が最大になるエポックの NER モデルをテストデータに適用した．

5.4 結果

我々は各モデルを英日のテストデータの各投稿に適用し，NER の最もシンプルな評価手法である CoNLL-2003 [15] のスキーマを用いて評価した．

	F1-value		F1-value	
	micro	macro	micro	macro
改良 TDS + 追加学習	**0.699**	**0.592**	**0.708**	**0.567**
改良 TDS	0.665	0.522	0.648	0.513
ベースライン	0.271	0.284	0.240	0.196

表 2: 各モデルの評価 (英日)

	Precision	Recall	F1-value	#NE
Person	0.865	0.901	0.883	426
Creative work	0.480	0.500	0.490	24
Location	0.727	0.491	0.586	114
Group	0.570	0.526	0.547	272
Service&Product	0.566	0.409	0.475	115
Event	0.800	0.444	0.571	18

(a) 英語

	Precision	Recall	F1-value	#NE
Person	0.948	0.858	0.901	233
Location	0.814	0.569	0.670	123
Group	0.818	0.418	0.553	194
Service&Product	0.777	0.552	0.645	145
Event	0.250	0.042	0.071	24

(b) 日本語

表 3: **改良 TDS + 追加学習**のタイプ別評価 (英日)

表 2 に各モデルの F 値を示す．これより，英日いずれのデータにおいて二つの提案手法がベースラインの性能を大きく上回っていることがわかる．ベースラインは文脈を収集する際に消失タイミングを考慮しない結果，ノイズとなる文脈が訓練データに混入し性能が大幅に下がったと考えられる．二つの提案手法を比べると，追加学習した単語分散表現を用いた方が数値が向上しており，これは単一の入力投稿からは判定が難しいエンティティでも，大域的な情報を考慮して検知できていることを示している．

表 2 に，**改良 TDS + 追加学習**のタイプ別の結果を示す．英日ともに Person タイプの精度が良く，これは人物自体がエンティティ表層から容易に認識可能であるからと考えられる．一方で英語の Creative work や Service&Product，日本語の Event タイプは精度が低く，これらのタイプは TDS で構築した学習データにノイズが多いことを示唆している．

6 おわりに

本研究はマイクロブログから消失エンティティを検知するタスクに取り組んだ．エンティティの消失のタイミングが不明瞭であることに対応するため，TDS を改良し精度良く消失文脈を収集した．データ数が少なく認識モデルの学習が安定しないことに対応するため，単語分散表現をマイクロブログ投稿で追加学習する手法を提案した．実験の結果，提案手法は高精度に消失エンティティを検知した．

謝辞

本研究は JSPS 科研費 JP21H03494 の助成を受けたものです.

参考文献

[1] Satoshi Akasaki, Naoki Yoshinaga, and Masashi Toyoda. Early discovery of emerging entities in microblogs. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4882–4889, 2019.

[2] David Graus, Daan Odijk, and Maarten de Rijke. The birth of collective memories: Analyzing emerging entities in text streams. *Journal of the Association for Information Science and Technology*, Vol. 69, No. 6, pp. 773–786, 2018.

[3] Wei Shen, Jianyong Wang, and Jiawei Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 2, pp. 443–460, 2014.

[4] Zhaohui Wu, Yang Song, and C Lee Giles. Exploring multiple feature spaces for novel entity discovery. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3073–3079, 2016.

[5] Alan Ritter, Oren Etzioni, and Sam Clark. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 1104–1112, 2012.

[6] Thien Huu Nguyen and Ralph Grishman. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 365–371, 2015.

[7] Jing Lu and Vincent Ng. Joint learning for event coreference resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 90–101, 2017.

[8] Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1641–1651, 2020.

[9] Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. Overview of the tac 2010 knowledge base population track. In *Third text analysis conference (TAC 2010)*, Vol. 3, pp. 3–3, 2010.

[10] David McClosky and Christopher D Manning. Learning constraints for consistent timeline extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 873–882, 2012.

[11] Satoshi Akasaki, Naoki Yoshinaga, and Masashi Toyoda. Fine-grained typing of emerging entities in microblogs. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4667–4679, 2021.

[12] Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text (WNUT)*, pp. 140–147, 2017.

[13] Alan Akbik, Duncan Blythe, and Roland Vollgraf. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pp. 1638–1649, 2018.

[14] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.

[15] Erik Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT)*, pp. 142–147, 2003.

タイプ	エンティティ数	投稿数	消失文脈の例 (**太字**は消失エンティティ、下線は消失文脈特有の表現)
PERSON	416	56098	アンパンマンの**やなせたかし**さんがお亡くなりになりました．94 歳でした． 本当にたくさんの夢と勇気を…
LOCATION	341	18868	**TOHO シネマズ有楽座**が閉館 58 年の歴史に幕(写真 全 2 枚) HASH HASH URL
GROUP	723	44808	**スタジオコクピット**解散．最初の 5 年ほどしか在籍してなかったけとれど，ほんとうに色んな勉強をせて…
EVENT	90	7170	【京都の宇治川花火 再開を断念】今夏まで 4 年連続で中止となっている「**宇治川花火大会**」について…
SERVICE&PRODUCT	672	11630	既報の通り，**月刊 IKKI** は 9 月 25 日発売 11 月号をもって休刊いたします．支えてくださった皆様，本当に…
計	**1906**	**150204**	

表 4: TDS で Twitter から収集した消失エンティティデータセットの統計 (日本語) と消失文脈の例

タイプ	エンティティ数	投稿数
PERSON	147	844
CREATIVE WORK	10	46
LOCATION	46	126
GROUP	103	540
EVENT	8	38
SERVICE&PRODUCT	43	128
計	**357**	**1922**

表 5: TDS で Twitter から収集した消失エンティティデータセットのテストデータ統計 (英語)

タイプ	エンティティ数	投稿数
PERSON	73	440
LOCATION	42	228
GROUP	64	346
EVENT	9	50
SERVICE&PRODUCT	47	262
計	**235**	**1326**

表 6: TDS で Twitter から収集した消失エンティティデータセットのテストデータ統計 (日本語)

A 付録

A.1 データセット統計

表 4, 5, 6 に紙面の都合で記載しきれなかったデータセットの統計情報を示した．

A.2 ベースライン (赤崎らの TDS)

紙面の都合で記載しきれなかったベースラインの TDS 手法について説明する．赤崎らは新エンティティのデータセットを作成するため，Wikipedia の各エンティティについて，Twitter で最初に 10 回以上リツイートされた日の時系列で最初の 100 件の投稿を正例として収集した．負例については各エンティティについて，正例を集めた日から一年後の投稿を新規文脈でないとみなし，負例として正例と同数サンプルした．

我々はこれを消失エンティティに当てはめ，時系列を逆にして収集した．すなわち，Wikipedia の廃止された事物一覧の各エンティティについて，Twitter で最後に 10 回以上リツイートされた日の時系列で最後の 100 件の投稿を正例として収集した．負例については各エンティティについて，正例を集めた日から一年前の投稿を消失文脈でないとみなし，負例として正例と同数サンプルした．