新旧のタグなしコーパスを用いた新エンティティ収集の評価

井上 裁都　　粟村 誉　　長田 誠也　　立石 健二　　宮崎 林太郎　　山下 達雄
ヤフー株式会社
{tatinoue, tawamura, sosada, ktateish, rimiyaza, tayamash}@yahoo-corp.jp

1. はじめに

筆者らは，ヤフーが進めるマルチビッグデータの利活用に向けて，Entity Linking ライブラリの開発と知識ベースの構築を進めている．Entity Linking とは，テキスト中に出現するエンティティ(人物，組織，場所などの実存する概念)に言及した表記(メンション)を特定し，知識ベース内の対応するエンティティと結び付けるタスクである．知識ベースとしては Wikipedia が用いられることが多く、その場合は特に Wikification[1]と呼ばれる．

本論文では以後，知識ベースのことを辞書と呼ぶ．一般に Entity Linking の研究は，固定の辞書とタグ付きコーパスを使った評価を前提とすることが多い[1][2][4][5]．しかし，実世界への応用においては，辞書を常に新しい状態にすることはもちろん，更新した辞書の品質を評価できることが重要になる．例えば，常に最新の Wikipedia を辞書に用いるとき，そうでない場合とのカバレッジの差が，時間経過によりどの程度広がるか評価できる意義は大きい．

とはいえ，更新した辞書の品質評価は容易ではない．新しい辞書を定量的に評価するには，辞書中の新しいエンティティを含む新しいタグ付きコーパスが必要になる．タグ付きコーパスの作成には人手を要しコストが大きい．

本論文では，コスト削減のため，新しい辞書をタグなしコーパスを用いて評価する手法を提案する．提案手法は，新旧の辞書を使って新旧のコーパスを解析し，その結果の差分を分析することで評価する．提案手法の有効性を検証するため，筆者らが構築した辞書に手法を適用し，評価・分析した結果も併せて報告する．

2. 新エンティティの収集

筆者らが構築している辞書も Wikipedia をベースとする．各 Wikipedia ページには DBpedia のクラス体系†をベースとした独自体系の意味カテゴリ(人物，組織など)を付与しており，この付与ができたページをエンティティとして辞書に追加する．

ページへの意味カテゴリの付与には機械学習と Yahoo!クラウドソーシング‡の 2 つを併用している．意味カテゴリ付与のフローは，次の通りである．

(1) **学習データを作成**
人手でページに意味カテゴリを付与する．

(2) **意味カテゴリ分類器を構築**
Wikipedia が持つカテゴリ情報などを素性とし，学習データから分類器を構築する．分類手法はベクトル空間法を採用している．

(3) **意味カテゴリ候補を分類器で付与**
カテゴリ未付与のページに候補を付与する．

(4) **意味カテゴリ候補の正誤を判定**
クラウドソーシングにより，正誤を判定する．

(5) **学習データを追加**
正しくカテゴリ付与できたページを追加する．

(6) **(1)に戻る**
(4)のエラー分析により，データ不足の意味カテゴリを特定し，人手で補強する．

上記の(1)～(6)の手順を反復することで，最終的には，全ページに正しい意味カテゴリが付与されることが期待される．実際の運用では費用対効果なども考慮し，約 3 ヶ月に 1 回このフローを回している．意味カテゴリ付与の対象とするページは，作業時点で最新の Wikipedia ダンプデータから収集する．

本論文では，この Wikipedia ダンプデータからのページ収集とページへの意味カテゴリ付与を合わせて，新エンティティの収集と呼ぶ．単に新しいページをエンティティとして辞書に取り込むのではなく，ページに意味カテゴリを付与することで，エンティティ曖昧性解消の精度向上[2]や応用時の意味カテゴリ活用[3]が可能になる．4 章で述べる通り，新エンティティ収集の意味カテゴリ別評価も可能になる．

† http://mappings.dbpedia.org/server/ontology/classes/

‡ https://crowdsourcing.yahoo.co.jp/

3. 新エンティティの評価

一般に Entity Linking の評価にはタグ付きコーパスが用いられる．しかし，新エンティティの収集を評価するには，タグ付きコーパスに収集したエンティティが含まれる必要がある．新エンティティを収集する度に，新しいコーパスを用意するのは現実的と言えない．そこで本章では，新旧のタグなしコーパスを使って評価する手法を提案する．

用語の定義から始める．2017 年に話題になった人物の一人に「藤井聡太」がいる．2017 年は「藤井聡太」を話題にしたニュース記事が多数存在する．このような新しいエンティティを含むコーパスと辞書を，本論文ではそれぞれ新コーパス，新辞書と呼ぶ．逆に新エンティティを含まないコーパスと辞書は，それぞれ旧コーパス，旧辞書と呼ぶ．

筆者らが着目したのは，新エンティティの Entity Linking が可能なのはコーパスも辞書も新しい場合のみということである．逆にコーパスか辞書のどちらかが古い場合，コーパスから新エンティティは抽出できない．これを踏まえ，各辞書とコーパスの組合せに対し次の通り名前を付ける．

A： 新辞書 × 新コーパス
B： 新辞書 × 旧コーパス
C： 旧辞書 × 新コーパス
D： 旧辞書 × 旧コーパス

新辞書と旧辞書を使い，新コーパスに対し Entity Linking をした結果，つまり組合せ A と C を比べる．これは，新エンティティを辞書に加えた効果により，A の方が多くのエンティティを抽出できると考える．続いて，旧コーパスに対する新旧辞書の Entity Linking 結果，つまり組合せ B と D を比較する．こちらは，いずれも新エンティティを抽出できないので，結果に差は付かないと考える．

実際の新エンティティ収集では，新辞書に新エンティティ以外も追加されるのが普通である．これは旧辞書に既存エンティティ全ては登録されておらず，この未登録エンティティも新辞書に追加されるためである．すると，組合せ B と D のエンティティ抽出数は，既存エンティティの追加効果により，実際には差が出ると考えられる．そして，組合せ A と C のエンティティ抽出数は，新旧両方のエンティティの追加効果で更に大きな差になると考えられる．

以上を踏まえ，組合せ A, B, C, D におけるそれぞれのエンティティ抽出数を a, b, c, d と置けば，新エンティティ収集の効果 g は A と C の抽出数の差と B と D の抽出数の差との差，すなわち，

$$g = (a - c) - (b - d) \quad \cdots (1)$$

で求められる．

「差」の差に代わり，「増加率」の差も指標にできる．新コーパスでの増加率 a/c と旧コーパスでの増加率 b/d の差を新エンティティ収集の効果とする

$$g_r = (a/c) - (b/d) \quad \cdots (2)$$

も併せて提案する．式(2)は意味カテゴリ間の頻度差を考慮した比較評価ができるため，4.2 節の評価実験ではこちらの指標を用いる．また，既存エンティティの収集効果 h_r も次の通り定義し用いる．

$$h_r = b/d \quad \cdots (3)$$

4. 実験と考察

3 章で提案した評価手法の有効性を示すため，筆者らが構築した辞書の評価実験をした結果とそれに対する考察を述べる．なお，筆者らの Entity Linking 手法の詳細は石川ら[4]の報告を参照されたい．

4.1. データセット

提案手法で用いるタグなしコーパスは， 2013 年 4 月から 2017 年 9 月までのニュース記事より，掲載年月毎に 10,000 件ずつサンプリングし作成した．

評価対象の辞書は，2 章で述べた方法により作成した．今回は 2016 年 5 月から 2017 年 5 月までの間に構築した計 5 バージョンの辞書を評価対象とした．また，評価対象の意味カテゴリは，人物(Person)と組織(Organisation)の 2 種とした．

各バージョンで，収集した意味カテゴリ別の新エンティティ数，および収集の対象とした日本語 Wikipedia ダンプデータの日付を表 1 に示す．v0 は評価のベースラインとする辞書で，このエンティティ数は辞書中の総数を示す．v1～v4 のエンティティ数は，各バージョンのエンティティ総数とその前バージョンの総数の差に相当する．

表 1　辞書・カテゴリ別の収集エンティティ数

ver.	v0	v1	v2	v3	v4
人物	(270,740)	8,241	5,840	4,814	9,581
組織	(101,827)	2,680	2,127	3,423	1,580
ダンプ	2016/5/1	16/8/20	16/11/1	17/2/1	17/5/1

4.2. 抽出エンティティ増加率による評価

最初の実験では，3 章で提案した式(2)による辞書の評価結果を次の手順で視覚的に示す．

(1) **エンティティを抽出**
各年月のタグなしコーパスより，各バージョンの辞書を使ってエンティティを抽出する．

(2) **抽出エンティティ増加率を算出**
v0 辞書をベースラインとし，各辞書のエンティティ増加率を，コーパス年月毎に求める．

(3) **抽出エンティティ増加率をグラフ化**
辞書とコーパスの組合せ毎に，エンティティ増加率をグラフにプロットする．

意味カテゴリ別に抽出エンティティ増加率をプロットしたグラフを図 1, 2 に示す．各図の破線は，新エンティティ収集に用いた Wikipedia ダンプデータの日付(v0:緑, v4:青)を表す．

図 1 の人物カテゴリでは，各辞書の増加率は，コーパスの新しさと正の相関があることが分かる．また，新旧の辞書を比べると，新しい方が増加率の上昇幅が大きいことも分かる．新エンティティ効果は，例えばコーパスに 2014 年 4 月と 2017 年 4 月を，辞書に v0 と v4 を選び，式(2)を適用すると，図より $g_r = 8\% - 3\% = 5\%$ と求められる．既存エンティティ効果は式(3)より $h_r = 3\%$ で，$g_r > h_r$ である．

図 2 の組織カテゴリでは，各辞書の増加率は，コーパスの新しさとの相関が図 1 と比べて小さいことが分かる．一方，新しい辞書の方が増加率が大きい点は図 1 と共通している．g_r は最大で 2 ~ 3%，h_r は 8% 前後で，$h_r > g_r$ である．

4.3. 新出頻度上位エンティティの分析

図 1, 2 が示す各意味カテゴリの新エンティティ効果が有用な指標か確認するため，意味カテゴリ毎に新出頻度上位のエンティティを分析する．v4 の辞書に存在し，v0 には存在しないエンティティを新出条件とし，この頻度が 2017 年のコーパスにおいて上位のエンティティを表 2 に示す．

これも人物から見ていくと，「藤井聡太」など新出と見られるエンティティが比較的多いことが分かる．これは図 1 の $g_r > h_r$ という結果と適合する．「ネイマール」など新出ではないエンティティも含まれるが，これは辞書更新の際に Wikipedia ページの改名にも対応していることが影響している．「ネイマール」は v2 と v3 の間にページ改名があり，条件上新出となっている．

続いて組織を見ると，「学校法人森友学園」など新出のエンティティが見られる一方，「Twitter」など明らかに新出ではないエンティティが多数含まれる．これは新エンティティ収集とは別に，アドホックな

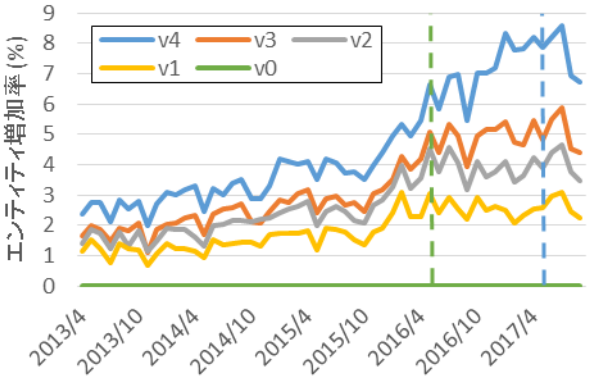

図1 人物エンティティの増加率

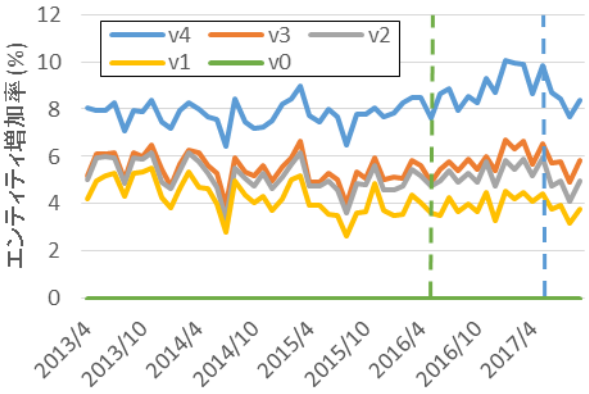

図2 組織エンティティの増加率

表 2　新出頻度ランキング

	人物エンティティ	組織エンティティ
1	ネイマール	Twitter
2	藤井聡太	フォーミュラ 1
3	籠池泰典	Facebook
4	崔順実	J2 リーグ
5	チェ・スンシル	アップル_(企業)
6	久保建英	衆議院
7	ブルゾンちえみ	学校法人森友学園
8	マックス・フェルスタッペン	参議院
9	経験値_(漫画家)	加計学園グループ
10	川谷絵音	AbemaTV

意味カテゴリ付与をしたことが理由の一つとなっている．「Twitter」はウェブサービス名と会社名の 2 つの意味を持つが，v3 と v4 の間でこのようなエンティティに組織カテゴリを付与したため，条件上新出となっている．以上より，図 2 の $h_r > g_r$ という結果も表 2 の内容と適合すると考える．

実験対象にしたコーパス，辞書，カテゴリを分析した結果が 4.2 節の g_r と h_r の比較結果と適合するため，提案手法の有用性は示せたと考える．

4.4. 辞書の経年劣化の分析

Entity Linking の実応用にて，辞書の経年劣化を検知する意義は大きい．しかし式(2)による評価は，ベース辞書からの性能差を測るものであり，ベース辞書自体の劣化を測れない．そこで，式(1)による評価も視覚的に確認する．コーパスと辞書(v0, v4 のみ)の組合せ毎に，人物エンティティの抽出頻度をプロットしたグラフを図 3 に示す．緑と青の破線の意味は図 1, 2 と同じである．

2 破線間の期間に着目すると，v0 と v4 の間で頻度差が徐々に開いており，v0 辞書が経年劣化の傾向にあることが確認できる．各辞書の頻度ピークの時期は，辞書と対応する破線付近にあり，辞書更新の前後から劣化が進行していると考えられる．

4.5. タグ付きコーパスによる評価

日本語 Entity Linking のタグ付きコーパスとしては，Davaajav らが公開した日本語 Wikification コーパス[5](以下 JWC)がある．JWC は BCCWJ の新聞記事サブコーパスをベースとし，その出版年は最も新しくて 2005 年である．よって，2017 年の新エンティティを辞書に追加しても，その評価は JWC では困難と考えられる．これを確かめるため，人物カテゴリに対する辞書別の Entity Linking 精度を JWC で評価した結果を表 3 に示す．

表 3　人物カテゴリの Entity Linking 精度

	適合率	再現率	F 値
v0	0.749	0.617	0.677
v1	0.749	0.621	0.679
v2	0.746	0.616	0.675
v3	0.748	0.622	0.679
v4	0.748	0.621	0.679

表 3 の再現率を見ると，v0 と v4 の差は 1%に満たず，新エンティティ収集による有意な改善は見られない．適合率も v0 と v4 の間に有意差はなく，解析精度に問題はないと言える．このようにタグ付きとタグなしのコーパスを併用することで，より有意な評価が可能になると考える．

5. 関連研究

本論文で評価対象としたのは Wikipedia に存在する新エンティティのみだが，Wikipedia に存在しない未知の新エンティティ検出の評価にも提案手法は

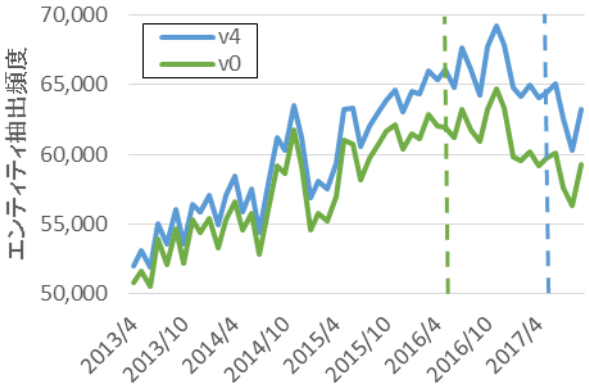

図3 人物エンティティの抽出頻度

応用できると考えられる．日本語の新エンティティ検出の研究としては，槇ら[6]の報告がある．槇らは独自のタグ付きコーパスによる評価をしているが，本論文の手法を併用した評価も可能と考える．

6. おわりに

本論文では，新旧辞書間の抽出エンティティの差を，新旧タグなしコーパス間で比較し，この「差の差」を新エンティティ収集の効果として評価する手法を提案した．実験と分析により，この差の差指標が有用であることを確かめた．更に差の差指標を用いて，辞書の経年劣化の存在も示した．

実験ではタグなしコーパスとしてニュース記事を扱ったが，今後は例えば Twitter などその他のコーパスを使った評価や分析も進めたい．

参考文献

[1] Rada Mihalcea and Andras Csomai. Wikify!: Linking Documents to Encyclopedic Knowledge. CIKM2007, 2007.

[2] Xiao Ling, Sameer Singh, and Daniel S. Weld. Design challenges for entity linking. TACL, Vol.3, 2015.

[3] 長田誠也，末永圭吾，善積正伍，庄司和正，吉田享晴，橋本恭明．エンティティリンキングを用いたドキュメントに対する地点情報の付与とその応用．言語処理学会第 21 回年次大会, 2015.

[4] 石川裕貴，小林健，長田誠也．ウェブ検索ログと Wikipedia 内部リンクを用いたエンティティの曖昧性解消．言語処理学会第 21 回年次大会, 2015.

[5] Davaajav Jargalsaikhan, 岡崎直観，松田耕史，乾健太郎．日本語 Wikification コーパスの構築に向けて．言語処理学会第 22 回年次大会, 2016.

[6] 槇佑馬，吉永直樹，鍛治伸裕，喜連川優．テキストストリームからの新エンティティの即時的検出．情報処理学会研究報告 2015-NL-220, 2015