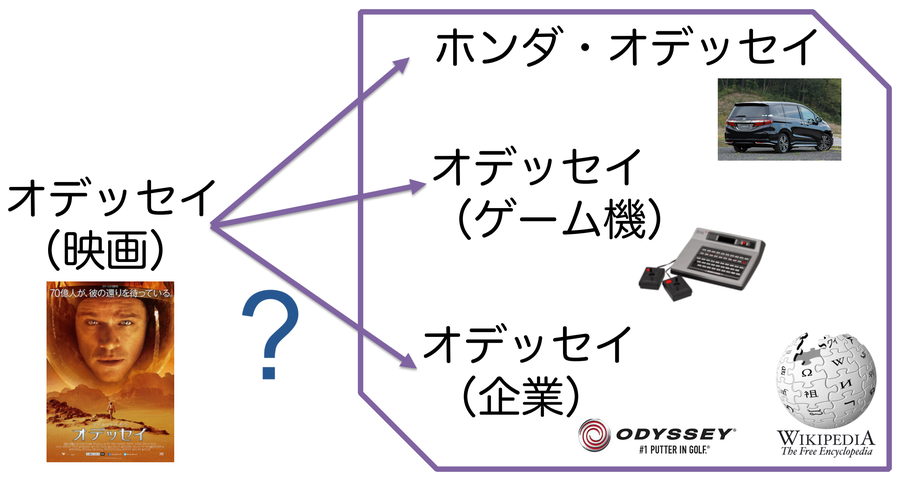
テキストストリームからの同名異義な未知エンティティの検出

赤崎智†，乾孝司‡，吉永直樹†，豊田正史†
†東京大学，‡筑波大学

背景

- Entity Linkingでは知識ベースに存在しないような未知のエンティティの扱いが難しく，既知のエンティティと同名で言及されるものも存在



本研究の目的：
このような例を**同名異義な未知エンティティ**と定義付け，Wikipediaに登録される前にTwitterなどのテキストストリームから検知する

関連研究

Exploring Multiple Feature Spaces for Novel Entity Discovery [Wu+, 16]

- 文脈，トピック，分散表現，検索クエリといった情報を用いて新聞記事から未知エンティティを発見する

同名異義な未知エンティティの存在を考慮していない

外れ値検出手法を利用した新語義の検出 [新納+, 12]

- WSDにおいて未知の語義は外れ値であるとし．外れ値検出の手法（LOF）を用いる

予めkを決めてLOF値のtop-k用例を検出するので，未知語義文の数がkより多い場合に対応できない

アプローチ

- ユーザは未知エンティティに対しSNSへの投稿やWebでの検索行動を取ると仮定
- そこで既存の素性に加え，以下のような様々なWeb上の手がかりを用いる
  - ✓ Twitter等の時系列データのバースト
  - ✓ 検索エンジンのクエリ頻度
  - ✓ WikipediaのPageView

- これらを用いてテキストストリーム上からリアルタイムな検出を行う（予定）

データセット

Wikipedia に追加されるエンティティには GO (言語) のような新たに出現・話題になったものだけでなく，駅名など単にマイナーなものも含まれるため，下記手順で検出対象のエンティティを収集

① 記事が登録されてから10日以内のPageViewを記事別に集計
② 集計値が高いものから人手で未知エンティティを抽出
③ 対象エンティティの語句を含むツイートを記事登録から10日前の間で100件取得

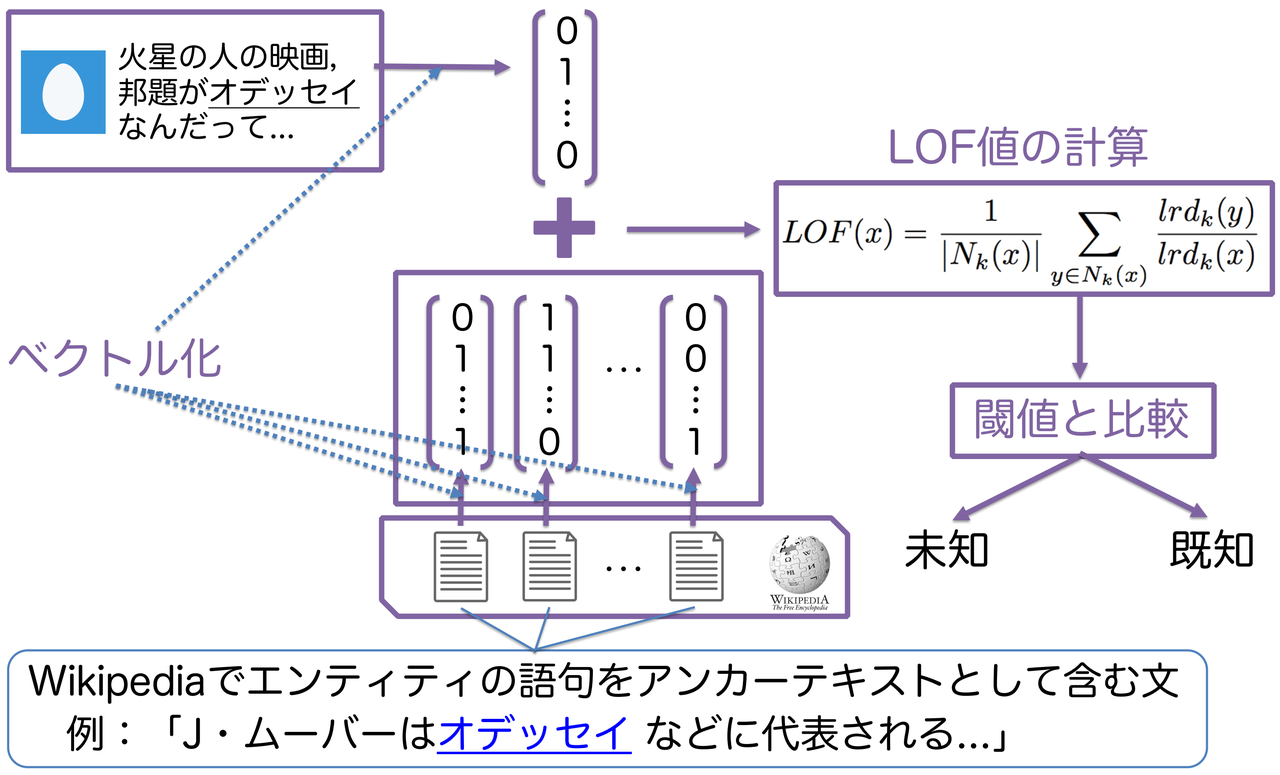
ツイートへのラベル付与の内訳

未知	対象未知	578
	その他	1,008
既知		514
計		2,100

対象エンティティは21件

提案手法

[新納+, 12]のLOFを用いて文ごとにLOF値を算出し閾値で判定



実験

- 評価データ
  - ➢ 21の語句に対応するWikipedia文を計1,171件用意した

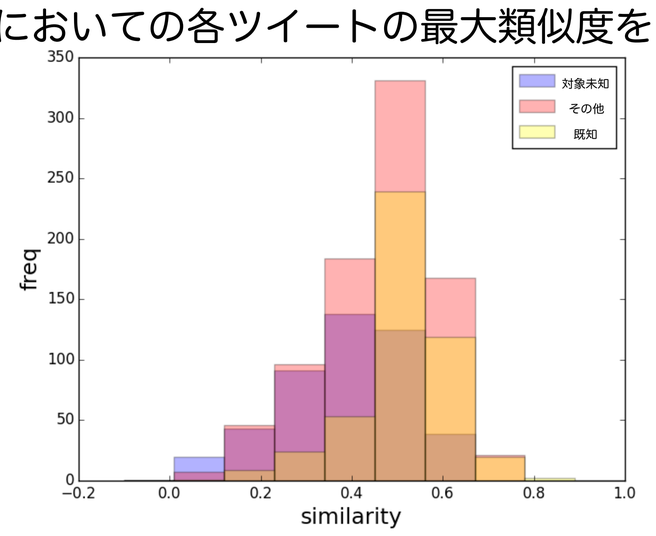
- 比較手法
  - ➢ 比較手法として，文同士のコサイン類似度を算出し，全ての比較対象の文との類似度が閾値を下回れば検出する手法を用いる [Shen+, 2012; Li+, 2013; and others]

また，各手法の文のベクトル表現については
内容語/名詞 BoW，内容語/名詞 word2vecの平均を用いる

ツイートからの新エンティティの検出結果

	cos (F,recall)	LOF (F,recall)
内容語BoW	(0.771, 0.945)	(0.770, 0.901)
名詞BoW	(0.812, 0.945)	(0.805, 0.927)
内容語w2v	(0.848, 0.985)	(0.854, 0.969)
名詞w2v	(0.843, 0.953)	(0.849, 0.961
全て正例	(0.861, 1.000)	

内容語w2vにおいての各ツイートの最大類似度をヒストグラム化



ツイート事例と最大類似度

ラベル	最大類似度	Twitter用例
対象未知	0.669	...監督のフィル・エイブラハムは４月放映開始の「デアデビル」演出してる...
その他	0.705	今泥酔すればキャロルちゃんの名前を口にしながら泣き出す不審者に...
既知	0.097	カーラ・デルビーニュとミシェル・ロドリゲスが熱愛！？
対象未知	0.028	あ，そういえばスティーラーズのエルヴィス（違）始めマクファディンさん...
その他	0.213	おシャンティ崖野さんおシャンティだ！って思ってたら画像欄がやば...
既知	0.845	【石鹸の勧め】髪の毛についた化学物質（シャンプーやリンス，ワックス...